

News Event Modeling

Capturing the impact of news on a stock price



Yvo Schoemaker

Bachelor Thesis Informatics & Economics

Erasmus University Rotterdam

February 2005

Supervisor

dr.ir. J. van den Berg

Yvo Schoemaker
Erasmus University Rotterdam
Bachelor Thesis Informatics & Economics
Telephone: +31647264267
Email: yvoschoemaker@yahoo.com

'It seems a waste of time to try to isolate a trend in data such as these. The Stock Exchange, it would appear, has a memory lasting less than a week.' ^[1]
- Maurice Kendall -

'Judgment and intuition will proceed more soundly if not hindered by an unnecessary grappling with market "patterns".'
- Harry Roberts - ^[2]

CONTENTS

1. Introduction	
1.1 General Background	2
1.2 Goal	2
1.3 Methodology	3
1.4 Structure Of The Thesis	3
2. Stock pricing	
2.1 News	4
2.2 Stock Pricing Methods	4
2.3 Limitations And Improvements	7
3. Research approach	
3.1 News Events	8
3.2 Research Problem	9
3.3 Implementation	10
3.4 A Graphical View	14
4. Experimental setup	
4.1 ANN Development	16
4.2 ANN Input Factors	16
4.3 Stock Data	17
4.4 Preprocessing	17
4.5 ANN Architecture	18
4.6 Categorizing Events	19
4.7 News Events	21
5. Results	
5.1 ANN Test Results	22
5.2 ANN Type A Deviations	23
5.3 Network Validation	26
5.4 News Events	27
6. Conclusions and further research	
6.1 Conclusions	30
6.2 Further Research	30
7. References	31

1. INTRODUCTION

1.1 General Background

In the past few years the interest in the influences and psychology around stock exchanges has grown to great heights. Started during the worldwide internet hype, stock investment became a significant means of individual finance. After the bubble exploded, many investors won but most lost a lot of money. The importance of trying to understand and predict future stock prices in order to lower investment risks and maximize returns seems clear.

Predicting future values of stock prices can be considered as the 'alchemy' of the twenty-first century. The question then is how to predict future stock prices. Traditional methods are often statistical analysis of historical stock prices in order to extract a linear system. However there are two major drawbacks. First, it seems obvious that due to the many factors influencing the stock price, it probably does not follow a linear system. Second, an important input factor is left out during analysis, which today also is partly responsible for the nonlinear system in the stock price formation; that is news. What is necessary is a more sophisticated analytical tool to understand stock prices and to research the influence of news on stock prices.

One current approach that takes the influence of news into account is that of 'event studies'. Event studies focus on the existence of a relation between stock price developments and all kinds of business specific events ^[3]. Its research is about how information of an event is valued by the market. If an 'event' occurs, an abnormal return should be discovered at the stock price where the event has its influence on. Such an abnormal return can be considered as a deviation on an expected pattern.

A commonly used tool for extracting patterns out of historical data is the Artificial Neural Network (ANN). ANNs provide a practical method for learning functions from examples and are among the most efficient learning methods currently known ^[4].

1.2 Goal

The presumption made for the research is the existence of a two-part formation of the stock price. It will be assumed that every stock has an expected pattern based on past news influences and a random residue, which together describe the current stock price. The random residue should represent the influence of current news events. If price developments of a common stock are based upon an expected pattern and a random residue, it should be possible to form a company specific expected pattern and relate the deviation of historical stock prices from the predicted stock

price to reported news events. The goal of the research described in this thesis is to capture the influence of news events on stock prices.

1.3 Methodology

An ANN is built up, used to perform a type of regression analysis and predict the expected pattern of a stock price using historical data. Assuming the stock price consists of an expected pattern and its residue, the questions rise if firstly an ANN can extract the expected pattern of an individual stock price and secondly if the deviation of historical stock prices on predicted stock prices corresponds to news events. In other words, will the poor performance of the ANN on particular records correspond to just the news events we want to find.

Deviations of the historic stock prices and deviations of current-day-deviations from previous-day-deviations will be scaled into impact categories including a threshold area explained further on in this thesis. Checking if data records lie in the same category using different ANN models will validate the performance and usage of the ANN models.

To capture the effect of news events, the extreme positive and negative deviations above a threshold must be related to reported news events. The performance and deviation of the ANN will be investigated to make conclusions about the influence of news on stock prices.

1.4 Structure Of The Thesis

The structure of this thesis is as follows. First the financial background, existing methods for understanding and predicting stock prices and their limitations will be discussed in chapter 2. Next the research approach with the goal of this thesis and the assumptions will be further explained in chapter 3. The experimental setup with the development of the ANN and categorizing of the news events will be described in chapter 4. In chapter 5 the results of the research will be shown while conclusions and thoughts about further research will be discussed in chapter 6. References can be found at the end of the thesis.

2. STOCK PRICING

2.1 News

It is important for all further research described in this thesis, first to define news and a news event. "News: a report of a recent event; intelligence; information" ^[5]. Relating this definition to the financial markets, *news* will be assumed to be a report on new information published by media, an institution or an individual, which has or had influence on the financial markets. The term *news events* will be used for all new information, which has influence on the *current* stock price. News events are always unknown and unpredictable in advance, because otherwise at least part of the information was not new.

2.2 Stock Pricing Methods

According to the Efficient Market Hypothesis (EMH), the present values of stocks are determined by a discounting process in which stock values equal the discounted value of expected future cash flows ^[6]. In other words, the EMH states that all known business specific information is reflected in the present value of a stock. A present value S_t of a stock is therefore supposed to be accurate. Future news events are random and unknown. A news event that will happen in the future is therefore unpredictable and cannot be part of the expected value of the stock price.

According to the EMH there are three forms of efficiency in which the market is commonly expressed. A weak form of efficiency: historical information has no predicting value, a semi strong form of efficiency: all 'public' information is instantly incorporated into the stock price, and a strong form of efficiency: all 'known' information is instantly incorporated into the stock price ^[7]. For all forms applies, new information must be instantly and completely incorporated into the present value of the stock.

To understand and predict future stock prices, an analytical tool is needed. There are several existing methods of which the Capital Asset Pricing Model (CAPM) lies on the basis of many others and is still widely used by investors ^[8]. The CAPM was setup in the mid-1960s by Sharpe, Lintner and Treynor. The CAPM states that the expected risk premium on each investment is proportional to its beta ^[9]. So the expected risk premium on stock equals beta times the expected risk premium on the market.

$$r - r_f = \beta(r_m - r_f)$$

Where r_m represents the return on the market index and r_f represents the risk free rate. This means that the return of a stock investment should lie on the Security Market Line (SML), which is a linear

line between the risk free investment and the market portfolio. An efficient portfolio offers the highest expected return for a given standard deviation. An investment which lies under the SML is 'under' efficient, will not be bought so the price will go down until it reaches the SML in equilibrium. The same applies for an investment that lies above the SML; it is 'over' efficient, traders want the high return with low risk so the price will go up until it reaches the SML.

I. Geometric Brownian Motion

As discussed by Hull, the process for stock prices can be described by an adjusted generalized Wiener process called *geometric Brownian motion*^[10]. The *Wiener process* is described by a current value with a future change expressed in terms of a probability distribution.

$$\delta S = \gamma \sqrt{\delta t}$$

Where δS stands for the change in the stock price during a small period of time δt and γ is a random drawing from a standardized normal distribution. This Wiener process has a drift term of zero, which means that the expected value of S at any future time is equal to its current value. The *generalized Wiener process* adds a constant drift term and variance rate.

$$\delta S = a \delta t + b \gamma \sqrt{\delta t}$$

Where a stands for the expected drift rate and b for the variance rate. The *geometric Brownian motion* tries to solve the problem that the generalized Wiener process fails to capture an important aspect of stock prices, which is that the expected percentage return required by investors is independent of the stock's price. The constant expected drift rate is therefore replaced by a constant expected return (drift divided by the stock price).

$$\delta S = \mu S \delta t + \sigma S \gamma \sqrt{\delta t}$$

Where μS stands for the expected drift rate in S and σS for the volatility in S . So the return in period δt can be described as follows:

$$r_{\delta t} = \frac{\delta S}{S} = \mu \delta t + \sigma \gamma \sqrt{\delta t}$$

II. Arbitrage Pricing Theory

Another analytical tool is the Arbitrage Pricing Theory (APT). It uses a sum of different influencing factors r with their sensitivities b plus a noise term to determine the price of an investment. If the portfolio has no sensitivity to any macroeconomic factor it can be called risk free and will be priced equal to the risk free investment ^[11].

$$r = a + b_1(r_{factor1}) + b_2(r_{factor2}) + \dots + noise$$

So the expected risk premium on a stock equals the sum of the macroeconomic factors times their expected risk premium, plus a noise term.

$$r - r_f = b_1(r_{factor1} - r_f) + b_2(r_{factor2} - r_f) + \dots + noise$$

III. Event Studies

A totally different tool for understanding stock prices is the research in event studies. Event studies investigate the possible existence of a relation between investment return developments and all kind of business specific events ^[12]. The importance lies on the impact a specific event has on the valuation of a stock price by the market. An abnormal return a is obtained if the realized return R of a stock over period t differs from the expected return R^* without any event.

$$a_t = R_t - R^*_t$$

$$R_t = R^*_t + a_t$$

Using this formula the effect of an announcement can be isolated. The expected return can represent the return on the market index or a prediction of a model. All relevant observations of period t grouped together create the average abnormal return AR . The assumption is made that all abnormal returns within AR are related to the event or else are coincidence. Grouping the abnormal returns let the coincidental effect diminish. If AR significantly differs from 0, the realized return during period t systematically differs from the normal expected return; the event had its influence on the stock price.

Another tool to extract a model out of training examples is the Artificial Neural Network (ANN). The ANN model is based on the physical construction of the human brain. It consists of a number of layers with each a number of neurons. Every neuron is connected to every neuron in the next layer with certain weights. Training with examples with input values for the first layer and a validation value for the output neuron cause the weights to be adjusted towards the wanted model. ANNs are very useful to learn complex patterns out of training data.

2.3 Limitations and Improvements

All described tools have their limitations when applied in research. The CAPM is very limited in considering different factors that form the stock price. An improvement on this comment exists in the 4-factor model. Three known systematical defects of the CAPM are the size effect, the B/M effect and the momentum effect ^[13]. These are proven defects and will not be discussed further here. The main focus of the CAPM lies on whether or not portfolios are efficiently priced in relation to the market. Future market values need to be known to give a prediction of a future stock price given its Beta. The same applies for the APT, without any known future factor values, no prediction can be made. Without any exact expected values, the model only can check if the current stock price is correctly priced. Event studies specifically look at the influence of an event on the stock price. It provides a practical method to statistically prove that an event had influence on the stock price, which caused an abnormal return. The main difficulty lies on the unknown normal expected value. If something has to be said about whether an abnormal return is realized, a model is needed to tell what the normal return should have been without any event influence.

These methodological limitations are not present with the use of the ANN. The ANN can try to learn in order to predict future stock prices with the use of historical data. The use of an ANN has however a few disadvantages. Firstly the model created during training remains a black box; it is very difficult, if not impossible, to understand and hence control the function created by the ANN. Only performance tests should tell if the model is correct. Secondly there is the problem of overfitting. If the ANN overfits the training examples it loses the ability to generalize to a given example of the remaining hypothesis space. The ANN also does not consider any error term or event influence, which is not already modeled into the network.

To capture the influence of a news event, the normal return is needed, or in other words the expected return. The next chapter describes how the ANN is chosen to learn this return.

3. RESEARCH APPROACH

3.1 News Events

The goal of the research described in this thesis is to capture the influence of news events on stock prices. The presumption is made that the stock price S exists of a two-part structure. The first part covers all news influences until time t , the second part is the random news event influence ε in period δt , which also can be called the residue or error term. The next function shows the *geometric Brownian motion* discussed in the previous chapter.

$$GBM: \quad \delta S = \mu S \delta t + \sigma S \gamma \sqrt{\delta t}$$

Writing this as a function in terms of stock price S :

$$S_{t+\delta} = S_t + \mu S_t \delta t + \sigma S_t \gamma \sqrt{\delta t}$$

The term $\mu S_t \delta t$ represents the expected value of the return during period δt of S or, in other words, the trend term of S . $S_t + \mu S_t \delta t$ together represent an independent term IT , which can be described as the result of all news influences until t ; the first part of the stock price. The term $\sigma S_t \gamma \sqrt{\delta t}$ is the stochastic component ε of the return, or all random news events E in the period δt ; the second part of the stock price.

$$S_{t+\delta} = \underbrace{S_t + \text{trendterm}}_{[\text{news influence until } t]} + \underbrace{\varepsilon(E_{\delta})}_{[\text{news influence in period } \delta]}$$

$$S_{t+\delta} = IT + \varepsilon(E_{\delta})$$

The two-part structure of the stock price can also be found in *event studies*. It states that the actual stock return equals the expected stock return plus the abnormal stock return. The abnormal stock return represents the random residue or news event influence.

$$ES: \quad R_t = R_t^* + a_t$$

The independent term IT is approximated by a number of factors where past news had its influence on. The term simulates the effect of the expected behavior of all traders, influenced by all past news. ε is the random residue; this should be the result of news events in period δt .

This model can be described in the same way as the *Arbitrage pricing theory* discussed in the previous chapter.

$$APT: \quad r = a + b_1(r_{factor1}) + b_2(r_{factor2}) + \dots + noise$$

$$IT = S_t + b_1(r_{factor1}) + b_2(r_{factor2}) + \dots + noise$$

The model is described by a number of factors where past news had its influence on times their sensitivities b . An example of such a factor is beta, a company's risk, which can be presented as market risk Mr + unique risk Ur .

$$\beta = Mr \cdot \beta_m + Ur \cdot \beta_u$$

The same type of formula can be used to describe a news event E_t . The influence of a news event on the stock price at time t : market event Me + unique event Ue with their sensitivities λ .

$$E_t = Me_t \cdot \lambda_m + Ue_t \cdot \lambda_u$$

According to the EMH the total influence of all news events $\varepsilon(E_t)$, must be incorporated in the stock price at time t .

3.2 Research Problem

Predicting future stock prices seems a hopeless task. Most sophisticated algorithms can make a regression model with a high training performance on historical stock data, however testing these models to make future stock price predictions often show that they cannot beat the market. (EMH of weak form – historical information has no predicting value.) A model based on historical data has overfitted the training examples what means the model performs excellent on the training data but generalizes poorly.

The main problem is that part of the stock price is assumed to be the influence of random news events. When overfitting a regression on historical stock data, the news events are 'forced into' a model. However the news events are random and unpredictable so the model will have little predicting value. Historical data cannot be used to train a model that predicts news events. If stock prices consist of the earlier described two-part structure, also the historical data needs to be split in an independent term and a random residue related to news. How can this split be found? One way is to track all possibly important news for an individual stock price and investigate the stock price at the times of news. The problem here is that it is generally unknown which exact development the stock price would have had at the time of news without the impact of the news. Therefore it is impossible to find out how big the impact of the news was or if the news had a positive or negative effect, if there is no model created to predict the stock price without news influence. Even to statistically prove there was any effect at all is here considered as a too difficult problem. The independent term is needed to know the normal development of the stock price without its random residue related to news. In this thesis it is chosen to first find that expected pattern and capture the abnormal stock return, or in other words to try to isolate the effect of the announcement of news.

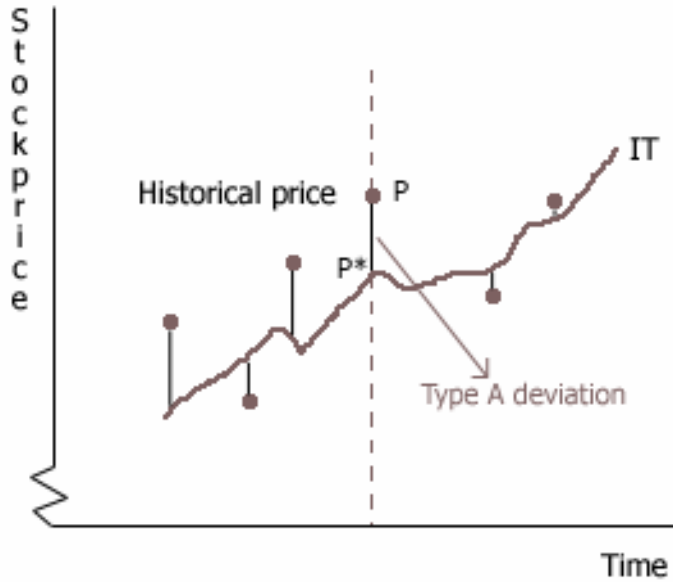
Not the total stock price must be trained, but only the independent term IT of it. On the one hand the problem exist of overfitting and forcing part of the random news events into the model, on the other hand making a too general model leaves an large noise term, which is actually part of the IT . Because the IT is approximated and it is assumed that the IT will not be found exactly, a noise term is added to the IT to represent the current stock price plus trend term.

$$S_{t+\delta t} = IT + noise + \varepsilon(E_{t+\delta t})$$

3.3 Implementation

The deviation of the stock price at $t + \delta t$ from the created model's IT estimation represents the random news event influence $\varepsilon(E_{\delta t})$. Deviations of the historical stock prices on the IT estimation by the created model will be linked to historical news events. There are two types of deviations defined.

Type A deviation

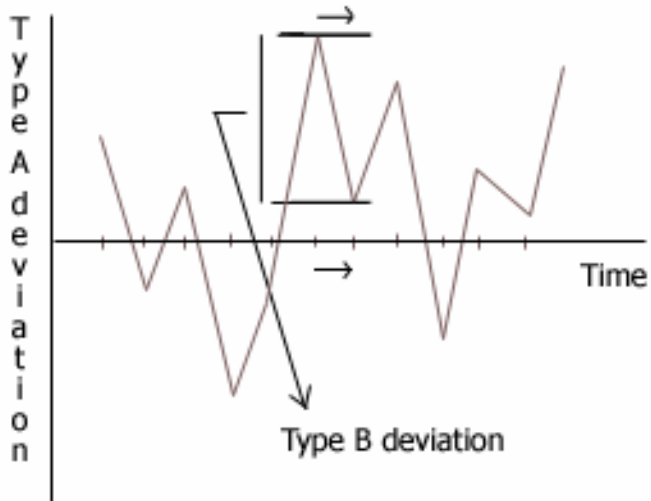


Type A deviation

The Type A deviation represents the relative deviation of the historic stock price P_t on the IT estimation P^*_t .

$$TypeA_t = \frac{P_t - P^*_t}{P^*_t} \%$$

Type B deviation



Type B deviation

The deviation of the current-day Type A deviation on the previous-day Type A deviation.

$$TypeB_t = TypeA_t - TypeA_{t-1} \%$$

If there has been a Type A deviation of 10% for 2 days long (so Type B on day 2 is zero), it is assumed there is no news effect on day 2. There are two reasons for making this assumption: first, day 2 could still carry the news effect of day 1; second and more important, after some analysis it appeared that the noise term added to the approximation of the independent term causes systematic errors; periods of 'over valuation' and 'under valuation', in those periods the stock price will not have a news influence on every day. So Type B deviations, or in other words movements in relative deviations, must be linked to historic news events.

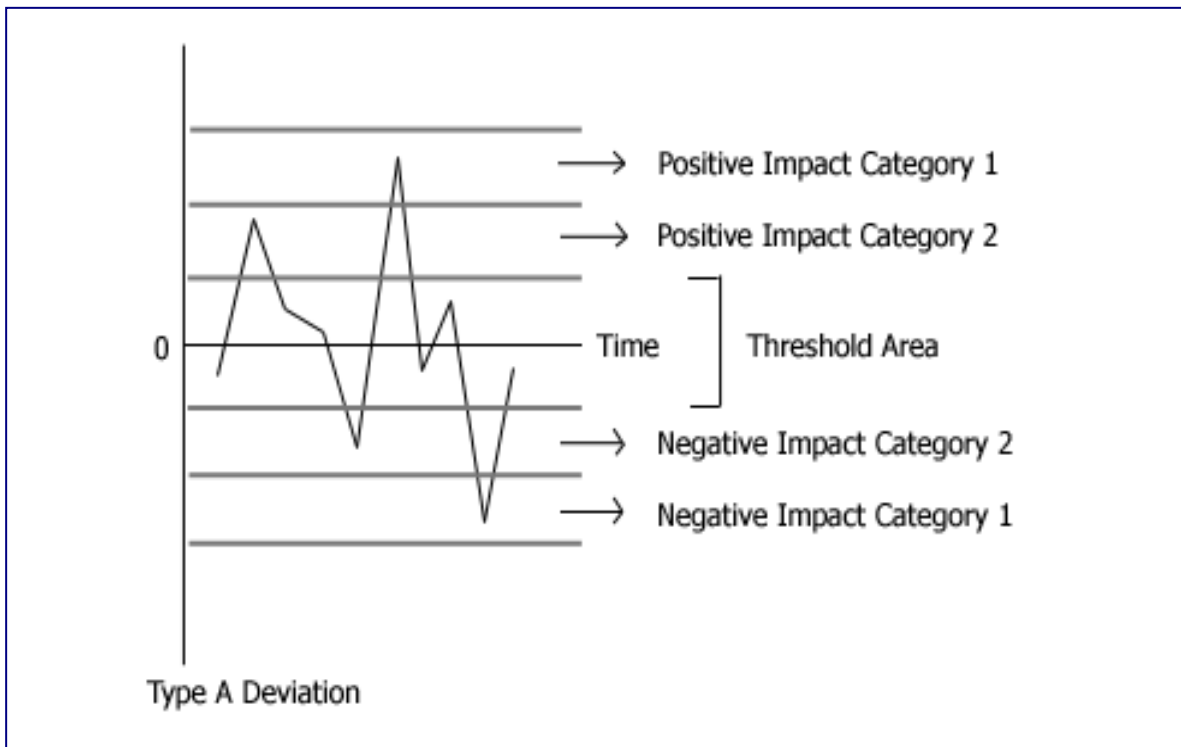
The Artificial Neural Network

The analytical tool chosen to learn the independent term or expected pattern of a stock price is the Artificial Neural Network (ANN). The advantages of the ANN for this research on the one hand are that they are easy to use and in a few steps a simple ANN model can be compared to a much more complicated ANN model. The great disadvantage of the ANN discussed earlier on the other hand is that a created model is like a black box. It is very hard, and with complicated models almost impossible, to find out how the model functions. In this thesis the main focus lies on finding the news events and not the understanding of the independent term, which is why this disadvantage is not a problem here.

The ANN will be trained on historical stock data. As mentioned earlier, the predicting value of the model will be limited. The ANN however will not be used for predicting because it will only represent the independent term of the historical data. There are two reasons why it chosen not to use a test or validation set. First, using a separate part of the dataset as test set or validation set does not test nor validate the ANN model. The model must be an approximation of the independent term, testing or validating must therefore also be done on the independent term, which is not available. Second, even when using a test set, after some analysis it appeared that the model became far too general, deviations became huge. Because of these reasons the use of a separate test or validation set will only be an abuse of valuable training data. When the ANN model is trained it will be tested on the existing training data. The idea is that records with a Type B deviation with a value above a certain threshold (an abnormal return) could have had a news cause. The threshold is needed because of the noise term added to the independent term; it is almost certain that the independent term found during research will not be perfect. Note that by using a threshold area, not only the noise is filtered out, but also small news events will be absorbed by the threshold area. The size of the Type A and Type B deviations depends on the complexity of the ANN model.

Categorizing Events

All deviations will be categorized in positive and negative impact categories and the threshold area. For every date or record in the dataset, a deviation is computed. This deviation lies in a positive impact category, within the threshold or in a negative impact category. To validate the usage of the ANN model as consequent regression model and therefore an approximation of the independent term, deviations of records computed must lay in the same impact category or threshold area using different ANN models. For example if the deviation of record 347 lies in the highest positive impact category using ANN model 1, and in the threshold or even a negative impact category using ANN model 2, the models are inconsequent.



The percentage in which the deviations lie in the same categories using different ANN models is assumed to be the consistency of the ANN models. If the consistency is high enough it can be assumed that the noise of the *IT* is not large enough to affect the news event categorization. Records with a high positive or negative Type B deviation had an unexpected effect; there could be an influence of news. To prove this is the case, the dates of those records need be linked to historic news events.

3.4 A Graphical View

Figure 3.1 shows how deviations of an ANN can represent the news event. Having trained the ANN, it will be tested on the training data. This will give the approximation of the independent term IT ; it is assumed to represent the expected pattern of the stock price. Around it, the threshold is build up with an upper resistance level and a lower support level. The width of the threshold is assumed to be twice the average absolute relative deviation between the historical stock data and the IT (the total width from support to resistance will be 4 times the average absolute relative deviation). This average deviation will change between the different complexities of ANN models that will be applied. The historical stock data, represented in figure 3.1 by the dots, will either be inside the threshold or outside it. A historical stock price, which falls outside the threshold, will be called a breakout. If also the Type B deviation of that breakout is above a certain Type B threshold, a possible news event influence could have occurred at that record.

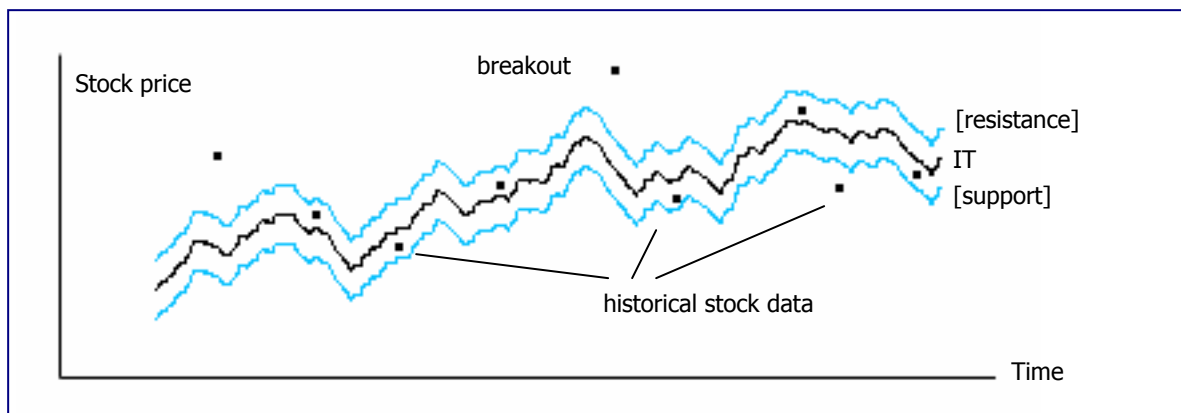
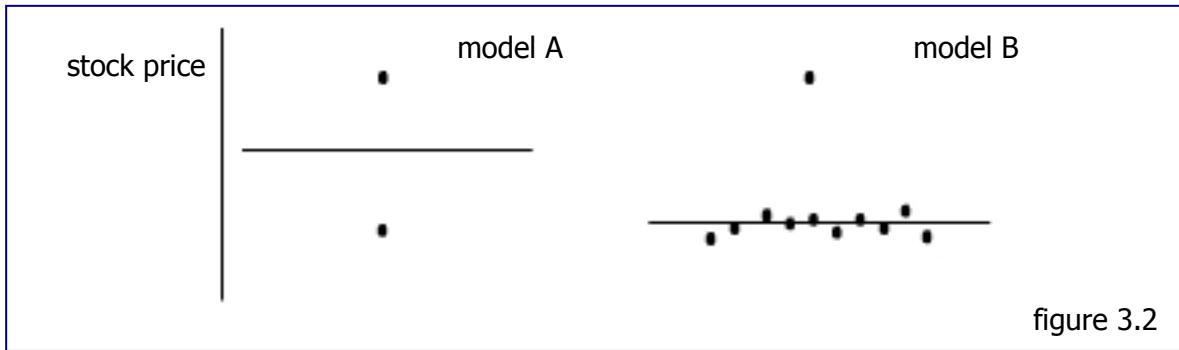


figure 3.1

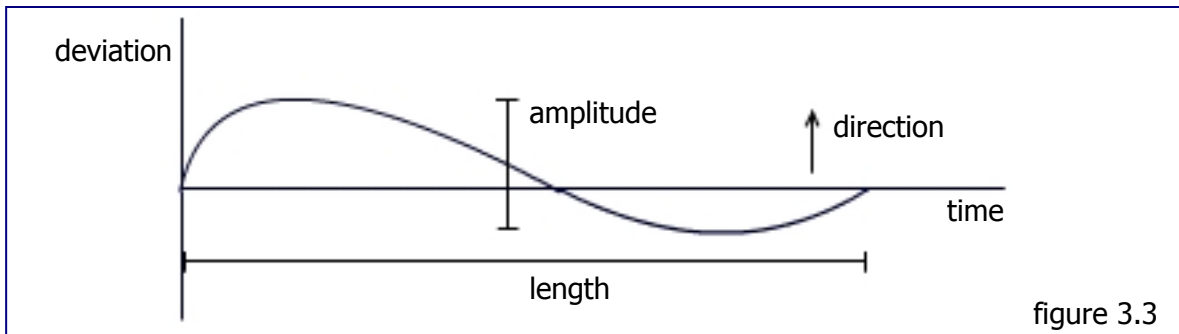
All breakouts will be categorized relative to its deviation. If breakouts occur on the same records using ANN models with different complexity (preferably falling in the same impact category), it is assumed there were abnormal returns at those records. If also the dates of those records correspond to historical news facts, the goal to capture news events is reached.

Figure 3.2 on the next page shows why a breakout can be called an abnormal return of the ANN model. Model A represents an ANN model with two training records within a certain area. The ANN model will learn a compromise between the two records, which is the line in between. The two historical data points have the same deviation to the ANN model while it is possible that only the upper point has an abnormal return.



Training with a massive dataset, which is shown in model B, solves this problem and let deviations only occur to records that are indeed different to the majority, the ANN model.

What should the influence of a news event on an individual stock price look like graphically? Imagining the deviation between an individual stock price and its expected pattern, we should find a kind of pulse if an abnormal return has occurred. Figure 3.3 shows a hypothetically event pulse. An event pulse has its amplitude, direction and length. A high amplitude of an event pulse could occur when a news event creates distrust or chaos. The nature of the news events - positive or negative - is (yet) unclear. If there is either a positive pulse of a negative pulse, the direction of the event pulse states the nature of the news event.





Another important property of the event pulse is length. The length of an event pulse depends on the impact of the news event but also the efficiency of the market. The EMH states that new information must be instantly and completely incorporated into the present value of the stock. Dann, Mayers and Raab (1977) concluded that at the New York Stock Exchange large block trades are reflected in the share price within 15 minutes. If therefore the deviation between a stock price and its expected pattern over time caused by a news event can be fitted into a function – with considerably length – the EMH must be rejected.

4. EXPERIMENTAL SETUP

4.1 ANN Development

Calculations during research have been done with Microsoft Excel, while the creation of the ANNs has been done with NeuroSolutions. The development of the artificial neural network needed to learn the expected pattern of the stock price can be divided in the following steps.

1. Decide which stocks will be studied
2. Determine timeframe and input factors of the ANN
3. Gather the needed input data
4. Preprocess raw data to be correct input data for the ANN
5. Determine ANN architecture (# layers, # nodes)
6. Cycle through training and testing

Two stocks are chosen to study the impact of news events. The first one is **Hagemeyer**,  a Dutch value added business-to-business distribution services group which has a highly fluctuating stock price. The second one is **Royal Dutch** or Shell Group of companies, a Dutch-British global group of energy and petrochemical companies with a relatively stable stock price. Artificial neural networks will be trained to learn the expected pattern of these two stocks.  The timeframe for the Hagemeyer input data is from January 1st 1999 until April 16th 2004, in other words from the start of the Euro time period until the start of this thesis research. Roughly a five-year period which stands for 1314 trading days. This amount of trading days is also the amount of records available to train the ANN. Royal Dutch exists far longer than Hagemeyer so a bigger timeframe can be used for researching its stock price. The timeframe for the Royal Dutch input data is from January 1st 1994 until April 16th 2004. This ten-year period stands for 2605 trading days, which will be used for training.

4.2 ANN Input Factors

1. Excel numeric value of the current date
2. Last day high and low stock price
3. Last day trading volume and interest rate
4. Business specific Alpha, Beta, R^2 and volatility
5. Last 3 days and last 10 days volatility

6. Current day open stock price
7. Timeframe of the last 5 end-of-day AEX closing values
8. Timeframe of the last 25 end-of-day closing values

These 42 factors above represent the input factors that will be used to train both the Hagemeyer ANN and Royal Dutch ANN. The first one is the current date of the input record. It is for example often said that January has systemically higher returns than other months ^[14] and therefore the date can prove to be a useful input factor for the ANN. Last day information like high stock price, low stock price, trading volume and interest rate will also be used as input factors. The statistical data Alpha, Beta, R^2 and volatility are constant over time. They have no influence on the difference between predicted output values but they can be considered as bias nodes in the ANN and therefore could improve training. Other input factors that will be used are the last 3 days and last 10 days volatility, which represent the stability of the time period that the current record is in. Current day open stock price, maybe a matter of argument but the goal is to find the effect of news during the trade day, not any overnight effects. The last factors are two timeframes, the first timeframe consists of the last 5 end-of-day AEX closing values and the second timeframe consists of the last 25 end-of-day closing values of the stock in question. The target for the ANN that will be using the input factors is the current end-of-day closing value of the stock price.

These 42 input factors (of which 4 constant) are assumed to form the stock price. They will be used to determine the expected pattern without any additional news influences. The weights in the network will converge to the independent term during training.

4.3 Stock Data

The historical stock data needed for training and testing the neural network will be obtained from the Finance department of the Yahoo website ^[15] where stock quotes are provided by Reuters. Yahoo provides all kind of data of the most derivatives of the major markets in the world. It also provides the business specific data Alpha, Beta, R^2 and volatility that will be used as input for the ANN.

4.4 Preprocessing

All financial data must be arranged in Excel in order to be used with NeuroSolutions. Before the historical stock data can be used for training, the data needs preprocessing to solve three encountered problems. The first one is the introduction of the Euro; all stock prices before January 1st 1999 are in Dutch guilders. This would give problems for the Royal Dutch ANN because the research timeframe for Royal Dutch overlaps this date. Therefore all stock prices before the Euro

introduction date will be converted to Euros. The second problem is the usage of dates as input factor for the ANN. To be able to use the dates it is necessary take the (Excel) numeric value of each date. Only numeric values can be used by NeuroSolutions to train its model. Theoretically it would be possible to use non-numerical input values as well but the available NeuroSolutions software could not give this option. The last data problem is stock splits. Both Hagemeyer and Royal Dutch splitted their stocks during the timeframes. For Hagemeyer that was on January 16, 2004 at 64 : 67 and Royal Dutch splitted its stocks on July 30, 1997 at 4 : 1. A stock split changes the expected pattern of a certain stock, which is why the stock splits will be corrected backwards in time. The latest stock price in the timeframe sets the standard, walking back through the dataset; all records with a date older then stock split date will be corrected. Trade volume is also an input factor for the ANN so this will also be corrected for stock splits at the same way.

Fortunately there are no relevant missing values in the historical data. The only missing values are on days when the stock exchange is closed. When a news event occurs, its effect will lie on the current day or the next day when the stock exchange is open. When finding an abnormal return it necessary not only to research the news of that day but also of the preceding days if those where closed days.

Hagemeyer and Royal Dutch paid out cash dividend to its stockholders several times during the timeframes investigated. Cash dividend can be corrected the same way as stock splits but a payment of cash dividend does not change the expected pattern of a certain stock so it is chosen to consider a cash dividend as an event. When an abnormal return is found at the moment of the payment of cash dividend this will be ignored if the deviation is limited to the size of the cash dividend. (The announcement of payment of cash dividend will be considered as news event indeed.)

4.5 ANN Architecture

Using the input factors mentioned earlier, the ANN would have 42 input nodes. At the beginning of the research, very large ANNs will be used to generate the expected pattern of a stock price. Afterwards smaller scaled ANNs will be created with less layers and hidden units to validate if large deviations fall on the same records using models with different complexity. Theoretically seen, the larger the scales of the ANN, the better it is capable to extract and fit the pattern. The disadvantage of this feature is that the dimensions and complexity of the ANN becomes incomprehensible and overfitting may occur. The black box character an ANN already has will only become worse, but investigating how the pattern functionally works is not in the scope of this thesis. The ANN must perform excellent on the training data and if the ANNs deviation corresponds to news events, the ANN model is indeed the needed pattern.

For all Neurosolutions based neural networks during research applies that weight update will be performed online and the used transfer function will be the tangens hyperbolicus (tanh). The momentum is 0.7, stepsize is set to 1.0 and the maximum number of epochs during a training cycle is 1000. These properties will be constant during research; only number of layers, number of hidden units and number of input units will be alternated.

The first ANN architecture that will be used for both Hagemeyer and Royal Dutch is a six-layer model. The number of units per layer from input layer to output layer will be as follows:

$$\text{ANN-1: } 42 - 50 - 40 - 10 - 4 - 1$$

The second less complex four-layer architecture:

$$\text{ANN-2: } 42 - 10 - 4 - 1$$

The last architecture will be used to try to take the validation of the largest deviations being on the same records to the limit. A two-layer model with 42 input nodes and 1 output.

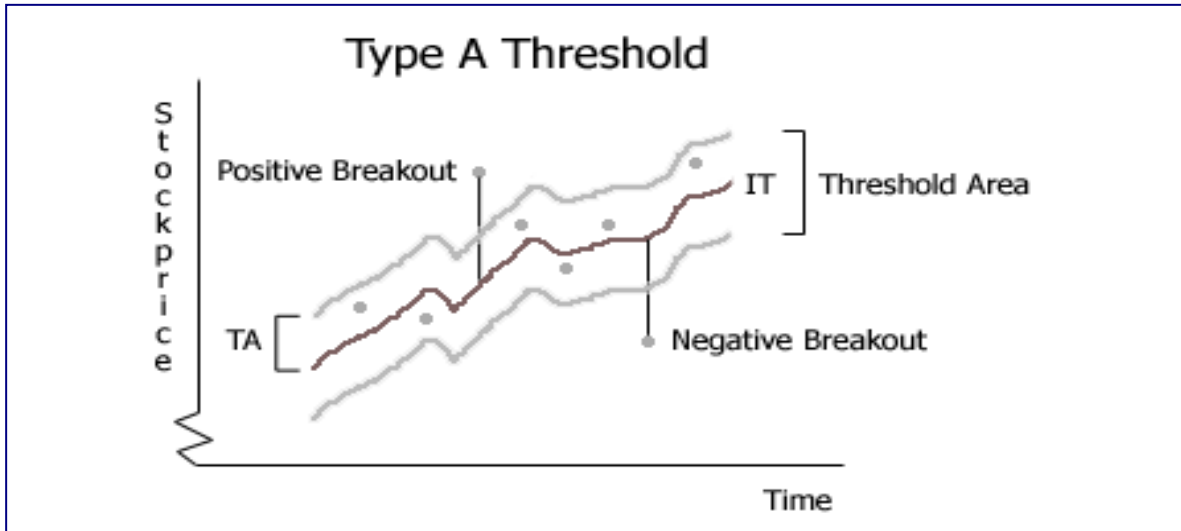
The Hagemeyer and Royal Dutch data sets will be totally used as training sets, to get the best estimate of their expected patterns

4.6 Categorizing Events

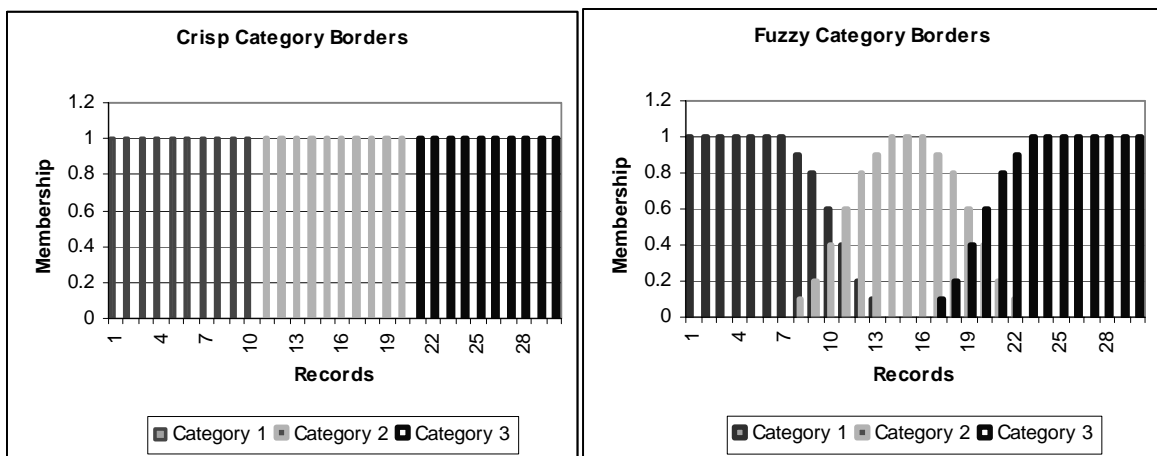
To determine the breakouts of the ANNs, the threshold will be used. The Type A threshold TA for relative Type A deviations of the ANN model is assumed to have a one-sided width of 2 times the average absolute relative Type A deviation of the historical stock price P on the predicted stock price P^* (the independent term). The word absolute means 'positive' in this description, not the real value. N stands for the number of records.

$$TA = 2 \cdot \frac{\sum_{i=1}^N \frac{|P_i - P_i^*|}{P_i^*}}{N}$$

A breakout is either a historical stock price higher than the predicted stock price plus the threshold, or a historical stock price lower than the predicted stock price minus the threshold.



The ANN models will be validated on consistency by scaling the relative Type A deviations. All breakouts will be categorized into negative or positive impact categories 1 and 2. The difference between the positive threshold barrier and the highest positive Type A deviation represents the area of the two equally large positive impact categories. The same applies for the negative impact categories. The crossings of different categories now have crisp type borders. The chance that all records lie in exactly the same categories for both ANN-1 and ANN-2 is almost zero in any experiment. A better approach is to assume categories to have fuzzy type borders. Now crossing borders are not explicitly defined but are gradual transitions. The next picture shows this approach graphically. In practice for this problem it means that if a record lies in different categories using ANN-1 and ANN-2, but the difference between the two Type A deviations coming out of ANN-1 and ANN-2 is below a certain value X , it is not considered as classified wrong. For value X the average of the two average absolute relative deviations of ANN-1 and ANN-2 will be used.



The percentage of records being in same category or threshold areas using different complexity type of ANN models is defined as the consistency of the ANN models.

4.7 News Events

The next step is to link the dates of breakouts with a relative Type B deviation above a Type B threshold TB to news reports to prove there was a news event. TB will have a one-sided width of twice the average absolute Type B deviation.

$$TB = 2 \cdot \frac{\sum_{i=2}^N |TA_i - TA_{i-1}|}{N}$$

If a breakout is situated above the threshold TB , a positive news report should be found. Also if a breakout is situated below the threshold, a negative news report should be found. There are however a few Type B breakouts which will be filtered before they will be linked to past news; these are 'correction' Type B breakouts. After some research it appeared that a day after a real Type B breakout occurred, the neural network obviously adapts its expected value incorporating this new event, but by doing this, the Type A deviation reduces and if the correction is big enough, a second Type B breakout is produced. Correction Type B breakouts are assumed not to have a news event influence.

News reports will be collected using multiple online news sites, which were found searching on a particular date in different formats and a stock name with Google. It is chosen not to use news reports of only one news provider, because different news providers sometimes report news on a different date. If the consistency of the ANN models is higher than 80%, the predicted stock prices are qualified to represent the underlying pattern of the stock price. The match percentage is the percentage classified Type B breakouts that can be related to past news reports. The match percentage determines in which degree the deviations are caused by news events. There will also be a quick view on the different types of matched news reports.

5. RESULTS

5.1 ANN Test Results

Hagemeyer test results on the six-layer ANN-1:

	Full training
Mean Squared Error	0.105731584
Normalized MSE	0.00109249
Mean Absolute Error	0.246272244
Min Absolute Error	0.000173793
Max Absolute Error	1.343017197

With a MSE of 0.11 and a mean absolute error of 0.25 for a stock price which fluctuates between 30 and 3 euros, it is an accurate model for a 5-year period. This ANN model will be used to further investigate news events for Hagemeyer.

To validate the usage of the ANN model as a 'benchmark', the four-layer network ANN-2 is also tested using the full training method with the Hagemeyer data set. The following table shows the test results:

	ANN-2
Mean Squared Error	1.098684891
Normalized MSE	0.011352352
Mean Absolute Error	0.782819671
Min Absolute Error	0.00105958
Max Absolute Error	4.134496765

The smaller model ANN-2 performs tolerable. In the next paragraph it will be used as validation for the ANN models.

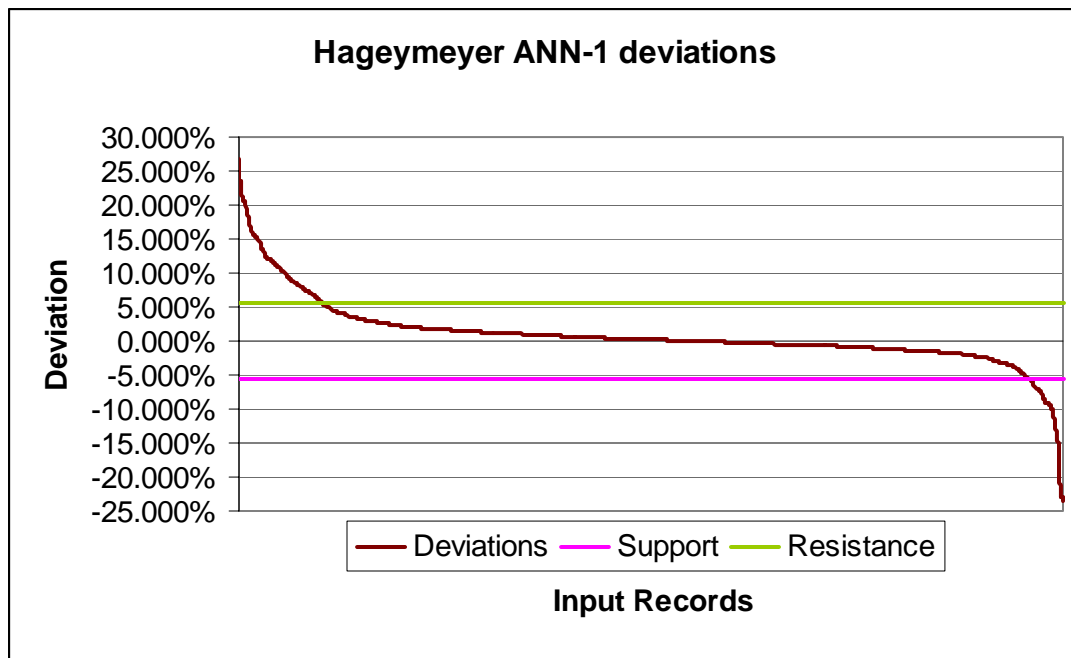
Training results for the **Royal Dutch** data set. The next table shows the test results.

	Full training
Mean Squared Error	0.469061908
Normalized MSE	0.002066917
Mean Absolute Error	0.522896586
Min Absolute Error	3.60107E-05
Max Absolute Error	3.513774109

5.2 ANN Type A Deviations

For **Hagemeyer**, using the six-layer ANN-1 with full training method, the average absolute relative Type A deviation of the historic stock price on the predicted ANN stock price is 2.76%. A one-sided Type A threshold is defined as twice the average absolute relative deviation as mentioned earlier. The resistance level of the threshold TA is 5.52% and the support level -5.52%. The next graph shows the Hagemeyer ANN-1 ordered Type A deviations.

$$TypeA_t = \frac{P_t - P^*_t}{P^*_t}$$

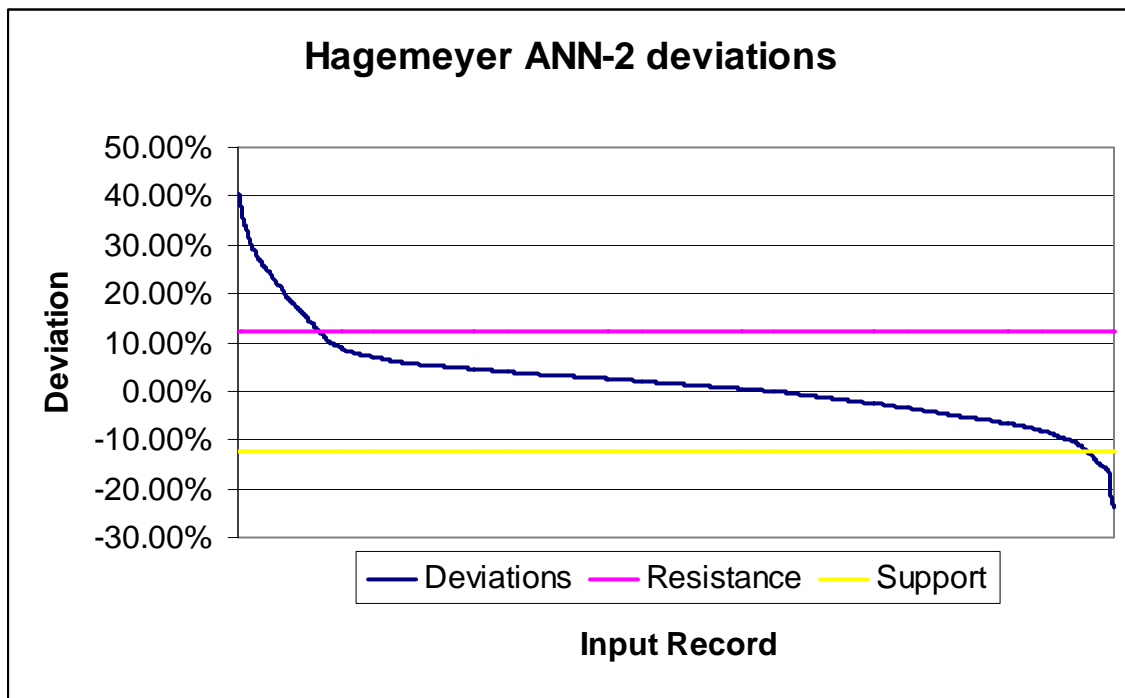


Of all 1314 historic stock prices of the Hagemeyer dataset, 10.2% lies above the resistance level of the threshold; 4.4% lies below the support level of the threshold. This means that 85.4% of the Type A deviations lies inside the threshold and is assumed 'news event free'. The highest positive Type A deviation is on September 29th 2003 and had a value of 26.63%. The highest negative Type

A deviation is on January 23rd 2004 with a value of -23.46%. The breakouts above and below the threshold TA are divided into 2 positive impact categories and 2 negative impact categories. This gives the next proportions:

Highest Positive Deviation	Point: 26.63%	0% Deviation	Point: 0%
Positive Impact Cat. 1	Volume: 1.6%	Lower Threshold Area	Volume: 40.0%
Category Border	Point: 16.08%	Support Level	Point: -5.52%
Positive Impact Cat. 2	Volume: 8.6%	Negative Impact Cat. 2	Volume: 3.8%
Resistance Level	Point: 5.52%	Category Border	Point: -14.49%
Upper Threshold Area	Volume: 45.4%	Negative Impact Cat. 1	Volume: 0.6%
0% Deviation	Point: 0%	Highest Negative Deviation	Point: -23.46%

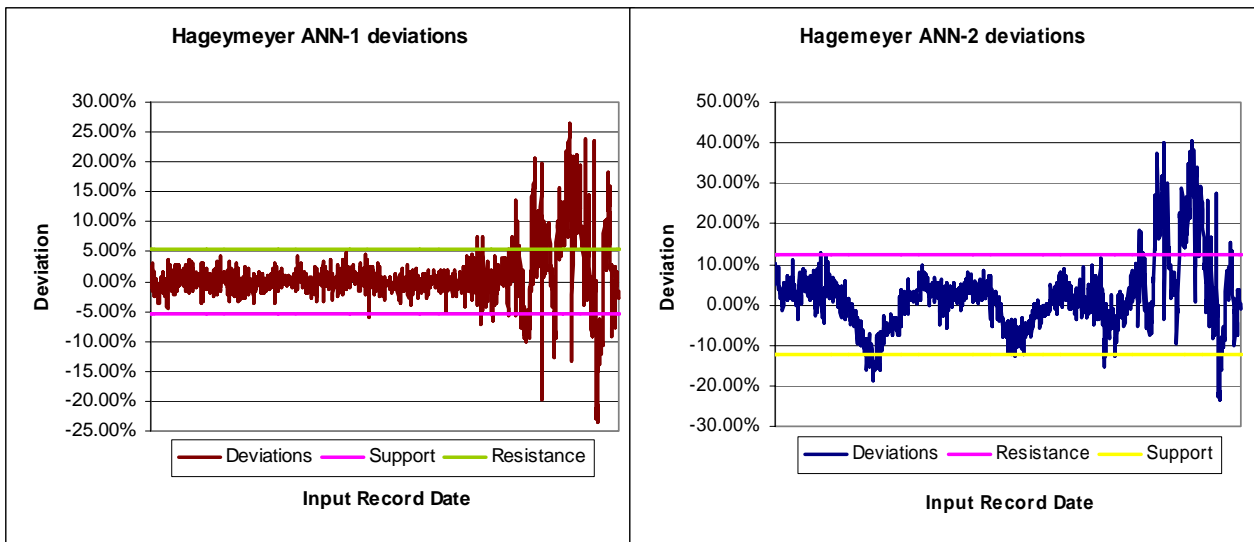
The same calculations are done for the four-layer network ANN-2. The average absolute relative Type A deviation on this network is 6.15%, this gives a resistance level of 12.31% and a support level of -12.31%. The threshold of ANN-2 is much larger then the more accurate six-layer network ANN-1. The following graph show the ANN-2 ordered Type A deviations.



87.7% of the Type A deviations lie inside the threshold. Proportions:

Highest Positive Deviation	Point: 40.70%	0% Deviation	Point: 0%
Positive Impact Cat. 1	Volume: 2.7%	Lower Threshold Area	Volume: 36.1%
Category Border	Point: 26.51%	Support Level	Point: -12.31%
Positive Impact Cat. 2	Volume: 6.5%	Negative Impact Cat. 2	Volume: 2.6%
Resistance Level	Point: 12.31%	Category Border	Point: -17.89%
Upper Threshold Area	Volume: 51.6%	Negative Impact Cat. 1	Volume: 0.5%
0% Deviation	Point: 0%	Highest Negative Deviation	Point: -23.68%

The highest positive Type A deviation also is on September 29th 2003 and had a value of 40.70%. The highest negative Type A deviation is on January 20th 2004 with a value of -23.68%. The next highest negative Type A deviation is on January 23rd 2004, which was the highest negative Type A deviation for the six-layer network ANN-1. The following graphs show the Type A deviations of ANN-1 and ANN-2 in historic order. Clearly it can be seen that the largest fluctuations occur in the last 2 years.



5.3 Network Consistency

The question is if the same records appear in the same categories or threshold areas for both Hagemeyer ANN-1 and ANN-2. If this is not the case, the ANN models are highly fitted approximations of the stock price development patterns, but are not consequent and therefore not representative.

The average absolute relative Type A deviation of ANN-1 was 2.76%; the average absolute relative Type B deviation of ANN-2 was 6.15%. Using the fuzzy border approach, the fuzzy category crossing area is 4.46%. As stated before, it is assumed that if a record lies in different categories using ANN-1 and ANN-2, but the difference between the two Type A deviations coming out of ANN-1 and ANN-2 is below 4.46%, it is not considered as classified wrong. The percentage of records lying in the same category or threshold area considering that assumption is the consistency of the ANN models.

Validation statistics:

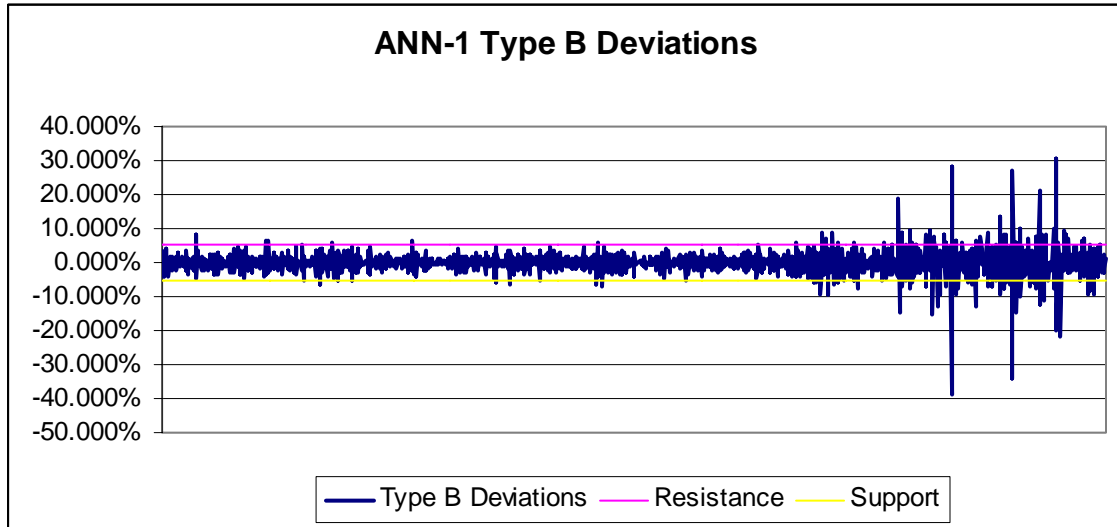
Crisp Threshold Mismatch	9.7%	
Fuzzy Threshold Mismatch	4.7%	
2 nd Degree Crisp Category Mismatch	0.6%	
2 nd Degree Fuzzy Category Mismatch	0.6%	
Above Crisp Threshold Mismatch	4.0%	
Above Fuzzy Threshold Mismatch	1.7%	
Below Crisp Threshold Mismatch	5.7%	
Below Fuzzy Threshold Mismatch	3.0%	
1 st Degree Crisp Category Mismatch	11.8%	
1st Degree Fuzzy Category Mismatch	6.4%	consistency: 93.6%

2nd Degree mismatch means that the predicted category by ANN-1 was 2 categories remote from the predicted category by ANN-2.

With a consistency of 93.6% the ANN models are very representative. The next step is to link the breakouts to historic news events, taking into account that in any case around 6.4% of the breakouts will have no news event because they are classified wrong.

5.4 News Events

The next graph shows the relative Type B deviations for **Hagemeyer** in historic order.



Type A breakouts with a relative Type B deviation outside the Type B threshold TB , must be linked to historic news events. The average absolute relative Type B deviation for Hagemeyer ANN-1 is 2.29%, this makes the width of the one-sided threshold TB 4.59%. Of all 1314 records, 148 lie outside the threshold area. Of these 148 Type B breakouts, 64 are also Type A breakouts and 10 of these 64 records are filtered out because they are 'correction' Type B breakouts described earlier, which leaves 54 classified Type B breakouts.

The next table shows a part of the records with an expected news influence and their short version news search results.

Deviation	Date	News search result
-38.89%	13-jun-03	Profit warning - Net loss exceeds 25 million - Daughter company sold to reduce burden of debt.
-33.94%	10-okt-03	Banks negative - Financing not complete.
-21.85%	16-jan-04	Multiple claims - Board member Hol resigns - New weight in AEX Costs 2003 higher than expected - Stock splits
-14.89%	20-okt-03	No agreement with banks
-13.06%	31-jul-03	Disappointing H1 figures says Kempen & Co - Results worse than expected
-12.70%	19-mei-03	ABN AMRO lowers advice to reduce
-10.22%	28-okt-03	Demand claim emission - Revenue drops
-9.57%	7-okt-02	Profit warning Buhrmann
-9.39%	11-mrt-04	Madrid bomb attacks
-9.35%	22-mrt-04	Biggest loser stock exchange; reason unknown

9.01%	6-mrt-03	Hagemeyer does not expect any big disinvestments in 2003
9.14%	26-jan-04	Rudi de Becker new CEO Hagemeyer starting march 1 2004
9.26%	5-mei-03	AEX gains 10%
10.18%	27-okt-03	Hagemeyer proposes emission of stocks to solve financial problems
13.38%	17-sep-03	Good performance results in higher valuation by traders
18.97%	27-feb-03	Europe's stock exchanges in lift
21.31%	3-dec-03	ABN Amro takes part debts
27.02%	13-okt-03	300 million through emission
30.32%	8-jan-04	ABN Amro expands share in Hagemeyer from 5,6 to 15,1 percent

The classified Hagemeyer Type B breakouts have a match percentage of 90.7%; for 5 classified Type B breakouts no news was found. 6.4% of the missing news could be caused by the 93.6% accuracy of the used ANN model. Of the 64 non-classified Type B breakouts, which had no Type A breakout and were no correction Type B breakout, the match percentage is 68.3%. There are 2 cases when a record is a Type B breakout but not a Type A: the deviation moves from outside the Type A threshold, inside the Type A threshold; this could still be a news influence but does produce a Type A breakout, or the deviation moves only within the Type A threshold; there was probably no news influence but if the jump was high enough a Type B breakout is produced. Looking at 68.3% in relation to 90.7%, around 75% of the Type B breakouts, which were no Type A breakout, can be explained by case 1 and the remaining 25% by case 2. No news event matched to a Type B breakout had an opposite, and therefore impossible, news influence, so no positive news caused a negative effect.

It appeared that 3 types of news scope could be discerned. News which is directly related to Hagemeyer; this represents approximately 60% of the investigated news events. News which influences the whole AEX stock index, this represents about 30% of the news events and news which is related to Hagemeyer's competitor in the same line of business: Buhrmann, this takes up less than 10%. Bad news for Buhrmann affects also its sector; traders will also react towards Hagemeyer.

The same calculations are done for **Royal Dutch**. The average absolute relative Type B deviation for Royal Dutch ANN-1 is 0.94%, this makes the width of the one-sided threshold TB 1.88%. Note that this is much less than the Hagemeyer Type B threshold, which indicates that Royal Dutch is less volatile. Of all 2425 records, 311 lie outside the threshold area. Of these 311 Type B breakouts, a tiny 47 are also Type A breakouts and 4 of these 47 records are filtered out because they are 'correction' Type B breakouts.

The 43 classified Type B breakouts have a match percentage of 76.6%. Part of the match error is again probably due to inaccuracy of the ANN model, another part is probably caused by not finding many search results on the period before the year 2000. Also the training period could have been too long to represent the expected pattern over the whole period. There are 227 non-classified Type B breakouts; this is 264 minus the correction Type B breakouts. The non-classified Type B breakouts had a low 46.3% match percentage, estimated with an 80 records random survey.

The next table shows a part of the classified Royal Dutch Type B breakouts with their short version news search result.

Deviation	Date	News search result
-4.66%	12-mrt-03	Lowest AEX in 2 years
-3.55%	29-apr-03	Oil price drops - Pressure on results
-3.31%	18-mrt-04	Negative advice ABN AMRO and Friesland Bank Securities
-3.23%	31-mrt-95	No news found
-3.18%	14-feb-94	No news found
-3.17%	23-jan-01	Shareholders meeting - Thousands of litres of bunker fuel spill in Durban bay
-2.98%	11-mrt-04	Madrid Bomb Attacks
3.42%	5-feb-03	Annual returns - 41% extra EBIT - Takeover
3.51%	1-okt-97	No news found
3.75%	2-jan-03	Strikes in Venezuela cause drop American oil supply - Higher oil price
3.82%	6-okt-98	AEX stock index 5.8% higher
3.87%	16-dec-02	Highest oil price in 2 months - Disorder in Venezuela - Iraq
4.04%	11-sep-00	No news found
4.46%	3-nov-00	High profit not caused by high fuel prices but reorganisations
4.84%	13-nov-01	OPEC-countries decided on restrictions of oil production per day

Just as with the Hagemeyer news results the news scope could be discerned into 3 main types. About 30% of the news results consisted of news influencing the whole AEX stock index. 50% consisted of news concerning the oil price and the last 20% consisted of news directly in relation to Royal Dutch.

News types could be scaled into 2 categories: the first one are facts & figures, the second one are advices & views of banks and analysts. The first category is easily traced because almost all facts and figures are publicly reported. The second category is harder to find. Some views of influencing analysts are only open to professional traders. For both categories apply that it could be that a part of the news events was not found due to only insider accessibility.

6. CONCLUSIONS AND FURTHER RESEARCH

6.1 Conclusions

The consistency of the ANN models is 93.6%, therefore it can be concluded that the ANN model is consequent enough to represent the expected traders behavior. The goal to capture news events is reached with a match percentage of 90.7% for the classified Hagemeyer Type B breakouts and 76.6% for the classified Royal Dutch Type breakouts. News here has a great influence in the determination of the new stock price. Looking at the non-classified Type B breakouts the match percentage drops to respectively 68.3% and 46.3%. From these results it can be concluded that indeed the price of the investigated stocks can be modeled by a two-part formation. The deviation between the ANN prediction and historic stock price has been successfully matched to news. It also appeared to be very important which decisions are made towards deviations and threshold areas. During research the threshold used to describe the inaccuracy of the expected pattern is quit large, so for tracking down smaller news events, the expected pattern must be increasingly accurate and also be assumed to be more accurate using the threshold. Comparing the match percentages of Hagemeyer and Royal Dutch, one can conclude that the more volatile Hagemeyer stock price is more influenced by news events than the more stable Royal Dutch stock price. Given the assumptions made during research, the research methods have proven to be accurate to capture the impact of news.

6.2 Further Research

The research and results of this thesis can form a method or a basis for further research. Further research can be done about the news event types. When news events are found, which types can be discerned and do different types have different influences on the stock price. What are the differences in influence between the 2 news type categories found; facts & figures and advices & views. Research can also be done in finding the optimal threshold parameters. When shortening the training period, the expected pattern could become more accurate for that period. Using the different deviations, indicators can also be created; for example the spread of type B breakouts during a period. When different news reports can be linked to influence types, it might be possible to predict the impact of news events.

7. REFERENCES

Reference Notes

- ¹ Maurice Kendall, *The Analysis of Time Series: Prices*, 1953. Quoted by Lars O Sødahl, *Systematic Elements in the Price Formation in Speculative Markets*, 2000, p. 3
- ² Harry Roberts, *Stock-Market "Patterns" and Financial Analysis: Methodological Suggestions*, *Journal of Finance*, Vol. 12, no. 1, 1959. Quoted by Lars O Sødahl, *Systematic Elements in the Price Formation in Speculative Markets*, 2000, p. 3
- ³ R. Ball & P. Brown, *An Empirical Evaluation of Accounting Income Numbers*, *Journal of Accounting Research*, Vol. 6, No. 2, 1968, p. 159-178
- ⁴ T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997
- ⁵ Definition from The College Dictionary, The Random House
- ⁶ R. Brealey & S. Myers, *Principles of Corporate Finance*, McGraw-Hill, 2003
- ⁷ P.C. Van Aalst et al., *Financiering en Belegging 2*, Rhobeta, 1997
- ⁸ J. Graham & C. Harvey, *The theory and Practice of Corporate Finance: Evidence from the Field*, *Journal of Financial Economics* 60, p. 187 – 244, June 2001. According to the research of Graham and Harvey 74% of firms always used the CAPM to estimate cost of capital.
- ⁹ R. Brealey & S. Myers, *Principles of Corporate Finance*, McGraw-Hill, 2003
- ¹⁰ J.C. Hull, *Options, Futures, and Other Derivatives*, Prentice Hall, 2003, p. 222
- ¹¹ R. Brealey & S. Myers, *Principles of Corporate Finance*, McGraw-Hill, 2003
- ¹² N. L. van der Sar, *Event-Studies: Methodologische Aspecten*, *Bundel Financiële methoden en technieken*, 1997, p. 119 - 142
- ¹³ G.T. Post & P. Van Vliet, *Conditional Downside Risk and the CAPM*, 2004
- ¹⁴ Haugen & Lakonishok, *The Incredible January Effect: The Stock Market's Unsolved Mystery*, 1988
- ¹⁵ <http://finance.yahoo.com>

Literature

- P.C. Van Aalst et al., *Financiering en Belegging 1*, Rhobeta, 2002
P.C. Van Aalst et al., *Financiering en Belegging 2*, Rhobeta, 1997
R.C. Brealey and S.C. Myers, *Principles of Corporate Finance*, McGraw-Hill, 2003
E. Frank and I. Witten, *Data Mining*, Morgan Kaufmann, 2000
J.C. Hull, *Options, Futures, and Other Derivatives*, Prentice Hall, 2003
P. Kennedy, *A Guide To Econometrics*, Blackwell, 2003
T.M. Mitchell, *Machine Learning*, McGraw-Hill, 1997
Jess Stein, *College Dictionary*, The Random House, 1975

Articles

- S. Bouman and B. Jacobsen, *Een maandelijks patroon in aandelenrendementen*, MAB, 1996
R. Haugen & J. Lakonishok, *The incredible January Effect: The stock market's unsolved mystery*, Dow-Jones-Irwin, 1988
M. Kendall, *The Analysis of Time Series: Prices*, 1953
G.T. Post & P. Van Vliet, *Conditional Downside Risk and the CAPM*, ERIM Working paper, 2004
H. Roberts, *Stock-Market "Patterns" and Financial Analysis: Methodological Suggestions*, Journal of Finance, Vol. 12, no. 1, 1959
N. L. van der Sar, *Event-Studies: Methodologische Aspecten*, Bundel FMT, 1997
L. O Sødahl, *Systematic Elements in the Price Formation in Speculative Markets*, 2000

Internet Resources

- <http://finance.yahoo.com/>
<http://finance.yahoo.com/q/cp?s=^aex>
<http://www.hagemeyer.com>
<http://www.shell.com>
<http://www.rusnet.nl>
<http://www.dft.nl>
<http://www.reuters.com>
<http://www.ta.nl>
<http://www.google.com>