# Evaluating the Performance of Nearest Neighbour Algorithms when Forecasting US Industry Returns[1]

C. S. Pedersen[2]      S. E. Satchell[3]

## Abstract

**Using both industry-specific data on 55 US industry sectors and an extensive range of macroeconomic variables, the authors compare the performance of nearest neighbour algorithms, OLS, and a number of two-stage models based on these two methods, when forecasting industry returns. As industry returns are a relatively under-researched area in the Finance literature, we also give a brief review of the existing theories as part motivation for our specific choice of variables, which are commonly employed by asset managers in practice. Performance is measured by the Information Coefficient (IC), which is defined as the average correlation between the 55 forecasted returns and the realised returns across industries over time. Due to transaction costs, investors and asset managers typically want a steady out-performance over time. Hence, the volatility of IC is taken into account through the application of "Sharpe Ratios". We find that two-stage procedures mixing industry-specific information with macroeconomic indicators generally outperform both the stand-alone nearest neighbour algorithms and time-series based OLS macroeconomic models.**

Keywords: *Nearest Neighbour Algorithm, US Industry Returns, Forecasting*

## 1. Introduction

The purpose of this paper is to build models of US industry returns and to compare the forecasting properties of these models. We will use both macroeconomic and industry-specific variables, and apply non-linear econometric techniques which are compared with more conventional models based on OLS. In general, nearest neighbour algorithms are an example of kernel/robust regression and applicable when the exact functional relationship between input and output is not known. This is certainly the case for the application, which forms the focus of this article, as how industry returns are affected by macroeconomic events is not known with any precision.

In all, we shall consider seven models, two of which apply nearest neighbour algorithms as a major part of the modelling process. To compare with these, we also present three benchmark OLS models, which, respectively, use industry-specific variables, macroeconomic variables, and both together. Finally, we compare all these

[1] We would like to thank Franklin Portfolio Associates for encouraging and supporting this research.

[2] Trinity College, Trinity Lane, Cambridge CB2 1TQ, England.

[3] Trinity College, Trinity Lane, Cambridge CB2 1TQ, England and Faculty of Economics and Politics, Cambridge University, Austin Robinson Building, Sidgewick Site, Sidgewick Avenue, Cambridge CB3 9DD, England.

to models, which apply OLS in two stages to the macroeconomic and industry-specific indicators.

As industry-specific variables, we use cross-sectional aggregate industry data on growth, book-to-price ratio and a measure defined as “success”, all of which are assumed to affect industry returns in much the same manner as their individual counterparts would affect the returns of a specific firm. In addition, some sectors of the economy may be particularly sensitive to certain macroeconomic fluctuations than others. Whilst little has been done in theory to support the choice of specific variables in forecasting industry-returns, we discuss the small available literature and offer an intuitive interpretation of a version of the production CAPM of Cochrane (1991) and Arroyo (1996) as part of our motivation. We include in our database a fair number of leading indicators together with the more conventional measures of unemployment, inflation and trailing market volatility as well as interest rate yields and credit spreads. From these, we construct four different information sets, all of which are applied to each of our models. Hence, the relative performance of the seven econometric models is examined across four different sets of macroeconomic variables, including relevant lags for several leading indicators.

To derive measures of forecasting performance of any given model, for each time period, we compute the correlation between the 55 industry returns forecasts and their realisation, known as the Information Coefficient. The average (through time) of this correlation is then our first measure of performance. We also recognise that, from a transaction cost and general asset management perspective, one wishes to have out-performance in every period, rather than some periods of extremely good performance followed by poor forecasts. Consequently, the volatility of this correlation will also be of interest, and we apply standard Sharpe Ratio arguments to produce a measure of risk-adjusted performance.

The paper proceeds as follows: in the next section, we introduce the underlying theoretical work partly motivating the use of the variables in our study. Section 3 present the data and Section 4 contains a description of the different estimation techniques and models. Section 5 summarises the forecasting results; Section 6 is reserved for our conclusions.

## 2. Modelling Industry Returns

In this section, we shall briefly review the sparse relevant academic literature on modelling industry returns. Recent papers on the modelling and forecasting of security returns almost exclusively deal with broad classes of financial assets, and so industry returns in isolation has received surprisingly little attention. King (1966) initially argued that stock returns could be broken into industry and market components. Reilly and Drzycimski (1974) additionally report differences in industry performance, betas and return volatility. More recently, studies have found further evidence for this. In particular, Breeden, Gibbons and Litzenberger (1989) report that industries producing goods with relatively high income-elasticity of demand have higher consumption betas. Industry-specific variables were also found to explain the variability in industry returns better than non-industry factors using alternative multi-factor models (see Rosenberg (1974) and Kale, Hakanson and Platt (1991)). Finally, various measures of growth or “success”, as well as common balance sheet measures (e.g. Book-to-Price or Price-to Earnings ratio), are frequently used by asset managers at both company and industry level.

The use of specific macroeconomic factors in predicting sector returns is a much less researched topic, and no theoretical model addressing this issue in particular is known to the authors. However, some studies have reported relationships between macroeconomic variables and sector returns. Boudoukh, Richardson and Whitelaw (1994) found that non-cyclical industries had a positive correlation with expected inflation whereas cyclical industries had negative correlation. Other studies have used larger industry groups as classification, and macroeconomic data combined with a variety of econometric techniques for prediction. (See, for instance, Sorensen and Burke (1986), Grauer, Hakanson and Shen (1990), Fama and French (1988), Ferson and Harvey (1991) and Lo and MacKinlay (1996).) These studies are surveyed in the introduction of Beller, Kling and Levison (1998), whose analysis focuses solely on the prediction of industry returns. In particular, they apply Bayesian techniques to forecast both cap- and equally- weighted indices of industry returns and examine the predictability of several macroeconomic variables adapted from the earlier studies cited, including term spread, default spread, T-bill spread, real interest rates, expected inflation and dividend yield. They conclude that industry returns are separately predictable both directly and when used in a more complicated model of portfolio optimisation.

In this paper we shall use both industry-specific and macroeconomic data, focusing solely on forecasting industry returns. Before proceeding to describe our econometric models and data, we will recall the production CAPM, which will act to partially fill the lack of theoretical motivation for using macroeconomic variables. Specifically, we give an explanation of the model in the light of our focus on industry return predictability and explain how one might proceed to customise these models for prediction. However, the main focus is empirical predictability rather than theoretical modelling, and so we do not pursue the challenge of deriving a theoretical model for industry prediction – rather, we hope our results will add to the motivation for future research in this direction.

We give a brief exposition of the production CAPM, following Arroyo (1991). Conventional CAPM models are based on the marginal rate of substitution of the representative agent, i.e.

$$P_0 = \frac{E(U'(\tilde{C}_1)\tilde{P}_1)}{U'(C_0)} \qquad (1)$$

in which $U'(C_0)$ denotes the marginal utility of consumption now, $U'(\tilde{C}_1)$ the marginal utility of consumption next period, $\tilde{P}_1$ the price of the asset next period, and $P_0$ the price of the asset now.

Estimating Equation 1 using aggregate consumption data leads to the equity premium puzzle. Consumption data are too smooth to give rise to the risk premia observed in the data unless the representative agent is spectacularly risk-averse. Attempts to rectify this based on the production CAPM (following Arroyo (1991)) work as follows. Consider a representative firm, which is usually assumed to have a Cobb-Douglas production function, and maximise profits subject to the constraints imposed by the production technology, an identity relating output to the sum of sales and investment, and a relationship relating capital next period to current capital and investment. The Cobb-Douglas assumption can be written as follows

$$Y_t = F(K_t,L_t) = A_tK_t^{\theta}L_t^{1-\theta} \qquad (2)$$

where $K_t$ denotes capital at time $t$, $L_t$ labour, $A_t$ technology, and $F$ the output/production function. The usual equation of motion of capital is

$$K_t = (1-\delta)K_{t-1} + I_t \qquad (3)$$

where $I_t$ is net investment and $\delta$ denotes the depreciation rate. Arroyo (1996) modifies Equation (3) to accommodate costs of adjustment of the capital stock $K_t$:

$$K_t = (1-\delta)K_{t-1} + I_t\Psi(I_t|K_{t-1})^{-\gamma}\,;\ \Psi > 0,\ 1 > \gamma > 0 \qquad (4)$$

The extra term $G(I_t|K_{t-1}) = \Psi(I_t|K_{t-1})^{-\gamma}$ reflects costs of adjustment, where $\Psi$ is a scale factor that ensures that $G$ will be less than or equal to one over the relevant range of $I_t|K_{t-1}$. One would then get the standard condition that the expected marginal product of capital equals the marginal cost of capital $h_t$:

$$E_t\theta A_t K_t^{\theta-1} L_t^{1-\theta} = E_t h_t. \qquad (5)$$

Now suppose that total investment consists of the proceeds from the issuance in capital markets of real bonds (units of which are denoted by $b_t$) and real equities (denoted $z_t$):

$$I_t = b_t + z_t \qquad (6)$$

If $r_t$ is the interest rate at time $t$ and $y_{t+1}$ the unknown yield on the security (counting both dividends and price changes), then the optimisation problem becomes

$$\max E_0 \sum_{t=0}^{\infty} \Delta^t ( A_t K_t^{\theta} L_t^{1-\theta} - (1+r_t)b_t - (1+y_{t+1})z_t - w_t L_t) \qquad (7)$$

subject to Equation (4) and Equation (6). Substituting Equation (4) into the production function and then deriving the first-order conditions on the optimal choice of $b_t$ and $z_t$ gives

$$E_t\theta\psi(1-\gamma)A_t K_t^{\theta-1} L_t^{1-\theta} I_t^{-\gamma} K_{t-1}^{\gamma} = (1+r_t) \qquad (8)$$

$$E_t\theta\psi(1-\gamma)A_t K_t^{\theta-1} L_t^{1-\theta} I_t^{-\gamma} K_{t-1}^{\gamma} = E_t(1+y_{t+1}) \qquad (9)$$

If each industry has a production function then Equations 8 and 9 tell us that the expected rate of return on corporate bonds and equity should be the same in each industry, a view we can reject immediately.

We can however extend the argument by replacing Equation 6 by Equation 10 below

$$I_t = \pi_t^b b_t + \pi_t^z z_t \qquad (10)$$

where $\pi_t^b$ and $\pi_t^z$ are time-varying functions that represent the costs of converting financial investment $I_t$ into real (i.e. plant) investment. Thus, for a given level of investment $I_t$, there will be a unique optimal level of bond and equity finance. To arrive at the optimum of this model, one solves Equation (7) subject to the cost of capital adjustment (Equation (4)) using Equation (10) in place of Equation (6). Conditions (8) and (9) then become

$$E_t\theta\psi(1-\gamma)A_tK_t^{\theta-1}L_t^{1-\theta}I_t^{-\gamma}K_{t-1}^{\gamma}\pi_t^{b} = (1+r_t) \qquad (11)$$

$$E_t\theta\psi(1-\gamma)A_tK_t^{\theta-1}L_t^{1-\theta}I_t^{-\gamma}K_{t-1}^{\gamma}\pi_t^{z} = E_t(1+y_{t+1}) \qquad (12)$$

or (in terms of stationary variables)

$$E_t\theta\psi(1-\gamma)(Y_t \mid K_t)(I_t \mid K_{t-1})^{-\gamma}\pi_t^{b} = (1+r_t\ ) \qquad (13)$$

$$E_t\theta\psi(1-\gamma)(Y_t \mid K_t)(I_t \mid K_{t-1})^{-\gamma}\pi_t^{z} = E_t(1+y_{t+1}\ ) \qquad (14)$$

The above equations could be adapted to build models, which relate industry rates of return to:

(i) Costs of financing investment; i.e. costs that transform financial capital into real investment specific to the industry

(ii) Capital adjustment costs

(iii) The average product of capital ($Y_t/K_t$)

(iv) The investment capital ratio ($I_t/K_t$)

(v) Labour markets ($L_t$ and $W_t$) will influence output and hence industry returns.

(vi) Other input factors and prices than capital or labour (e.g. oil, property).

Remarks (v) and (vi) suggest how macroeconomics variables will impact differently at the industry level. Also, output measured in units can be converted into value so that the price of output, and in particular the elasticity of the industry market, becomes a key factor. We recognise that Equations (2)-(10) make strong assumptions on functional forms, which need to be relaxed. However, to the extent that the variables in the model above vary across US industries, we might expect different returns across industries, which indeed are observed. To the extent that they vary in their predictability, we might also expect different degrees of predictability across industries. A paper by Basu and Vinod (1994) partly addresses this issue, by assuming a production CAPM and allowing for correlation via the production function; i.e. wealth not consumed is invested as capital, which becomes wealth (output) next period. They show that the equilibrium stock price will be mean-reverting in the case of log-utility (see Basu and Vinod (1994), Proposition 1). In particular, the more diminishing the returns to scale, the more mean-reverting the price, so that a high degree of predictability would be achieved if the industry was just below constant returns to scale. (The same results apply also with increasing returns to scale.) This suggests that mature industries are less predictable than growing ones. However, the

Basu and Vinod paper assumes a single firm, so that it would require further analysis to differentiate between industries if adapting such a framework. The authors are not aware of the existence of such generalisations.

Indeed, the general way macroeconomic variables affect industry-returns is not known with any precision. As stated earlier, we focus on the empirical predictability in this paper. Hence, whilst we take onboard the relationships suggested by the production CAPM explained above, our empirical analysis will make no restrictive assumptions about functional forms or properties of the returns to scale. Rather, we shall treat the above as a rough guide and consider a more extended set of macroeconomic variables than those suggested by any one particular model. Also, as the next section describes, our data will include variables commonly used by asset managers, who have been actively investing in industry sectors for decades. Hence, we hope that our description of the production CAPM lilterature – apart from being informative for the making of investment strategies – also could shed light on which variables could or should be included in future statistical models of industry returns.

# 3. Data

As mentioned earlier, our data set takes the form of a set of industry-specific variables and four permutations of certain macroeconomic factors. In this section, we introduce our data and discuss our choice of variables used in the analysis. We start with the basic properties of our industry return data, which constitute the dependent variables in each of our seven models.

## 3.1 Basic Properties and Predictability of Excess Industry Returns

All our industry-specific data, including returns, are monthly in the period January 1983 to September 1998. The industry returns were derived from individual company returns (in the form of a *$(c\times 1)$*-vector, *R*) using a *$(c\times 55)$* industry loadings matrix, *l*, which, based on US corporate tax returns, determines how large a fraction of the industry contributes to each company. Each row of *l* thus adds up to one; *c* denotes the total number of companies and 55 the number of industries.

Hence, in this framework, industry returns are given by the *$(55\times 1)$*-vector, *r*, determined by the regression

$$R = lr + e \qquad (15)$$

where *e* is some shock. Table 2 in the Appendix contains the time series summary statistics across the 55 industry returns; Min. of Quartile 1, for instance, is the minimum of the 1st Quartiles (the 25% cut-off points) in the 55 time series of industry returns. In Table 3, we give the cross-sectional summary statistics, where Min. of Quartile 1, for example, denotes the minimum of the 1st quartiles in the 176 cross-sectional industry returns series.

Comparing Tables 2 and 3, we see that the standard deviation of the 55 historical averages is 0.054. Each is calculated over 176 observations. Thus a measure of historical volatility is $\sqrt{176}\times 0.054 = 0.67$. Similarly, for the 176 cross-sectional means we have volatility 0.033, and since these are averages of 55 points, a measure of cross-sectional volatility is $\sqrt{55}\times 0.033 = 0.24$. This shows that industry returns have very *low* cross-sectional variability. Relative to their historical variability the ratio is about 0.24/0.67, or a little over a third, which implies that strategies based on exploiting cross-sectional variation (industry selection models) are unlikely to produce much excess return without massive gearing.

Table 4 sums up the serial correlation estimates (i.e. the beta in a regression of returns on a constant and lagged returns, AR(1) model) for the 55 industry time series, which gives a rough idea of predictability. We note that 41 out of 55 series have positive serial correlation, and only two are negative and significant at the 5% level. Overall, the low level of significance indicates in general, there is little value in predictions based on lagged variables only. This would hint at the need to use additional information, such as the macroeconomic and industry specific explanatory variables we duly turn to in the following sections.

When evaluating the forecasting performance of various econometric models introduced later, we shall use two criteria. Firstly, we will consider the average and cumulative correlation of forecasts and actual returns across the 55 industries over time, which gives a direct measure of forecasting ability. Our second measure concerns the practical advantages some forecast patterns may have over others. For asset managers and investors in general, steady out-performance is generally considered favourable to periods of large out-performance in between realising large losses. This is because (a) Transactions costs will be lower due to less rebalancing and (b) Targeted sponsors are less likely to change managers. Thus, we will also consider a “Sharpe Ratio” criterion function, which takes the volatility of correlation into account.

## 3.2. Derivation of Industry Characteristics: Growth, B/P, Success

The three industry-specific predictors we employ (defined as growth, book-to-price (B/P) and “success”) are derived by an aggregation technique identical with that for returns described above. The original data for these variables comes from Barra variables with the same numbers. (Barra Inc. being a well-known investment/asset management technology and consulting firm.) These Barra variables are widely used by US asset managers and considered traditional explanatory variables for individual company returns. In what follows, the symbol *l* still denotes the ($c \times 55$) industry loadings matrix for stocks l to *c*. Our variables are then defined by the following three regressions, in each of which *e* denotes the typical regression residual: For growth, we consider

$$G = lg + e \tag{16a}$$

where *G* is a growth factor loadings vector for stocks l to *n*. The resulting vector *g* gives monthly growth factor loadings by industry. For B/P,

$$B = lb + e \tag{16b}$$

where *B* is a *B/P* factor loadings vector for stocks l to *n*, is the defining regression. The vector *b* gives monthly B/P factor loadings by industry. Similarly, with “success”, we use

$$S = ls + e \tag{16c}$$

where *S* is a success (relative strength/momentum) factor loadings vector for stocks l to *n*. The vector *s* thus gives monthly success factor loadings by industry.

These regressions generate the elements of a time series of vectors for industries 1 to 55, with the elements [*r; g; b; s*]; i.e. the return in month *t* together with loadings on growth, *B/P*, and “success” at the start of month *t*. The vector [*g;b;s*] is called the *industry state vector*, i.e. the current values of the industry-specific predictors. In our analysis, we shall be using this vector and its first two lags to predict returns (so the industry-specific variables mentioned later consist of nine variables). Using lagged data implies that we shall allow some time for the information to filter through to market participants, some of whom may rebalance their portfolios less frequently than once a month.

### 3.3 Macroeconomic Data

The macroeconomic variables will be referred to as the Z-variables. We define four separate sets of variables, which are listed in the Appendix together with Datastream codes of the original data utilised. Whilst partly motivated by the previous studies surveyed in Section 2, our choice of variables is also affected by our reluctance to use those conventional macroeconomic factors subject to revisions and those not always available at monthly frequencies. Thus we prefer variables that are prices, interest rates or inflation rates. We do, however, keep a measure of unemployment and other leading indicators in our database.

In the few cases where missing variables where observed, we used a 24-month moving average to interpolate. The difference variables were then recomputed. Table 5 contains specific dates for each variable when this procedure was adapted. Next we removed the first observation from each variable (due to the lags defining the difference variables). We also note that as short-dated bond we used a two-year bond[4]. Also, since we were unable to find credit yield data for a basket of AAA or CCC-rated bond, we used the Credit Premium (differences in rates of return of the respective indices) rather than the Credit Spread as a measure of the current level of credit risk in the market. The rest of the variables are self-explanatory from the definitions given in Table 5 in the Appendix, and are selected – as stressed earlier – to reflect a broad set of current market and macroeconomic characteristics.

We shall run the models with 2- and 3-period lags of Z1, Z2, Z3 and Z5 in addition to the variables in levels. The decision to include/exclude specific variables in our “lag” set was based on definitional reasoning (e.g. Z4) or inspection of some preliminary correlation. It might be worth including larger lags of some of these variables, as some macroeconomic relationships may take more than three months to feed through to returns. However, larger lags also use up degrees of freedom, which mainly drives our decision to set our maximum lag at three.

## 4. Forecasting Models

Having introduced the data, we now introduce the models whose forecasting ability will be assessed. We will examine seven models, all of which are based on OLS and nearest neighbour algorithms. To recap, the standard OLS technique simply involves a linear regression of industry returns on the selected explanatory variables.

[4] Data was unavailable for bonds with a shorter term throughout our time period. However, we regard the difference in yield between a two-year and, say, one-year bond to be too small to seriously affect our main results and conclusions.

We now present the specific type of nearest neighbour algorithm we shall apply, before detailing the models of particular interest.

## 4.2. The Nearest Neighbour (NN) Algorithms

There are a large number of families of NN-algorithms in the literature, and many of these are described in detail in Campbell et al. (1997), Chapter 12. All NN-algorithms are based on the assumption that the past sets of explanatory variables, which are closest to (by some measure) the currently observed explanatory variables, initially should be identified. Then the corresponding values of the dependent variables in these time periods should be used to predict the current value of the dependent variable, possibly using a weighting scheme in which weights are functions of the distances. The particular model we shall apply is one where we pre-select the number of neighbours (*n*) in advance and then "optimise" over *n*. Such a procedure is known as an *n* - nearest neighbour regression model, commonly written as *n*-NN, and works as follows:

Firstly, if $y_t$ is the variable to be forecasted and $x_t$ the (*px1*) set of explanatory variables, then the general model

$$y_t = f(x_{t-1}) \qquad t = 1, \ldots\ldots.\mathrm{T} \qquad (17)$$

where *T* denotes the end of the sample, represents the basic framework. The model is then based on calculating the "distances" between $x_{t-1}$ and each of $x_1, x_2, \ldots, x_{t-2}, x_t$, and then selecting the *n* closest ones. In order to avoid inappropriate measurement of distance, all explanatory variables are normalised to have a mean of zero and standard deviation one; as distance measure, we employ Euclidian distance, which is the most commonly used distance measure in statistical analysis[5].

Let $k_t = (1, y_{1t}^*, y_{2t}^*, \ldots.., y_{nt}^*)$ be the $(1 \times (n+1))$ vector of values of the dependent variable for the *n* nearest neighbours of $x_{t-1}$. To fit (17) in-sample, we calculate a local regression for

$$\underset{T\times(n+1)}{K_t} = \begin{bmatrix} k_1 \\ k_2 \\ . \\ . \\ k_r \end{bmatrix} \quad \text{and} \quad \underset{T\times(n+1)}{Y_t} = \begin{bmatrix} k_1 \\ k_2 \\ . \\ . \\ k_r \end{bmatrix}$$

This gives us

$$\hat{y}_t = \hat{f}(x_{t-1}) = k_t (K'_T \Omega K_T)^{-1} (K'_T \Omega Y_T) \quad t = 1, \ldots\ldots..T \qquad (18)$$

where $\Omega$ is a $(T \times T)$ diagonal weight matrix ($\Omega = I_T$ is the default value). The above procedure fits $y_t$ for $t = 1, \ldots, T$. Defining $\hat{B}_T = (K'_T \Omega K_T)^{-1} (K'_T \Omega Y_T)$, we note that $\hat{B}_T$ can be interpreted as the "best" in-sample weights.

[5] We recognise that other distance measures could be employed rather than the Euclidian norm, such as for instance Mean Absolute Deviation or even Range. However, this we leave for future research, and stick with the more common approach presented.

By an appropriate choice of weights, $w_t$, we can choose to "downweigh" outliers in Equation (16) by constructing an $(n+1) \times 1$ weight vector $w_t$, such that $K_t^w = (1, w_{1t}y_{1t}, w_{2t}y_{2t}, \ldots\ldots, w_{nt}y_{nt})$ replaces $K_t$ in the above. Such a weighting regime ensures that the "closer" the past state of the economy and the industry-specific factors are to the present state, the more influence the returns observed then will have on the estimate of current industry returns. We choose an exponential patterns of weights[6], i.e. $w(x) = exp(-dx)$ $d > 0$, which we normalise to add up to 1 across the $n$ neighbours chosen.

Our forecasts are then simply computed by $\hat{y}_{T+1} = f(x_T) = k_T \hat{B}_T$ (or from Equation (16). To compute $\hat{y}_{T+2}$, repeat the calculation, except now recalculate $\hat{f}(x_{t-1})$ for $t$ = 1, T + 1, thus selecting from T + 1 rather than T neighbours. Hence, as our available data increases, we expand the window in order to incorporate as much information as possible.

### 4.2.1 Notes on the Nearest Neighbour Methodology

Before we present our results, we add a few remarks on our chosen $n$-NN methodology outlined above. Firstly, the procedure uses little information about $x_t$ itself; it only uses the information that some set of values $(x_t^1, x_t^2, \ldots\ldots, x_t^n)$ are the $n$ "closest" points to $x_t$, so we estimate an in-sample fit based on a weighted sum of the corresponding values of $y_t$. Consequently, it differs from conventional techniques in that the closest (in time) state of the world is not explicitly an argument of the forecasting function.

We also add that the choice of weights and $n$ is, of course, at our discretion. Searching over some time horizon, it is not clear how many points would be of interest. We shall investigate a selection of values of nearest neighbours (in particular, $n$ = 2,5,10,15,20,40,50,60,80 and 100) when evaluating forecasts, each time selecting the model with best apparent forecasting ability. Further, nearest neighbour procedures are known to be consistent (in the statistical sense) for different weighting schemes, as proven by Stone (1977). These proofs assume that the distance is fixed rather than the number of neighbours. The arguments in favour of the number of nearest neighbours being fixed is that it avoids, to some extent, the "curse of dimensionality", i.e. the need for an unrealistically high number of observations in order to get statistically valid results (see Campbell, Lo and MacKinlay (1997), page 504).

## 4.3 Seven Forecasting Models

The specific models we shall assess are defined in terms of the techniques and data utilised; we remind the reader that when referring to the macroeconomic data, we mean each of the four data-sets discussed in Section 3.3 and listed in the Appendix. Our Model I is defined as a simple OLS regression of the industry returns on the industry-specific variables only, Model II the same regression only on the macroeconomic data, and Model III an OLS regression on the combined data. These

[6] We also used "tri-cube weights", i.e. $w(x) = (1-|x|^3)^3$ when $|x| < 1$ and 0 when $|x| \geq 1$. This is discussed in Cleveland (1979) and gave results almost identical with those reported for exponential weighting.

models constitute the benchmark models against which more advanced models will be compared.

Models IV and V use a two-piece OLS technique. For Model IV, we rerun Model I, and then run a regression of the residuals in Model I on the macroeconomic variables. Forecasts are then simply formed as the sum of the forecasts in these two regressions. Model V mirrors this procedure, only running Model II first, and then regressing the residuals on the industry-specific variables.

Model IV is then compared with Model VI, which repeats the first step of Models IV but forecasts the residuals using the *n*-NN algorithm rather than OLS. As will be made apparent in Section 5, the industry-specific variables perform much better when entered linearly rather than non-linearly, and the macroeconomic variables perform poorly when using OLS in Model II. Thus, we do not pursue the option of firstly using OLS on the macroeconomic variables followed by an *n*-NN algorithm using the industry-specific variables on the residuals. Our final model, Model VII, uses the *n*-NN algorithm on all data to forecast returns directly. Table 1 in the Appendix summarises these models.

## 5. Results

In all models, the data are monthly, and our in-sample period is 03/83 – 09/92. The out-of-sample period, over which we evaluate the forecasting results, is 10/92 – 09/97, i.e. 56 time periods. A summary of the performance for all the models, using the number of nearest neighbours *n,* which gave the best results[7] for each run in Models VI and VII, are presented in Table 8. Before addressing this, we consider in detail the results of Model I, which threw additional light on what model/data combination works best. Hence, we regressed excess industry returns on the industry specific variables (growth, success and book/price) and their one- and two-period lags, and, in addition to our usual performance statistics, compute the $R^2$ of these regressions and the correlation between each pair of the residuals (1485 pairs in all).

The results are summarised in Tables 6 and 7 in the Appendix. The R-squares in Table 6 imply an average correlation of 0.276. The average correlation between predicted and actual returns[8] is 0.073 when applying an *n*-NN algorithm to the same data. Consequently, where we to restrict ourselves to industry-specific variables, simple OLS regressions dominate the *n*-NN approach. Also, the low cross-sectional residual correlation reported in Table 7 indicates a reduced role for common factors (such as for instance macroeconomic variables) in improving forecasts. However, this correlation is computed within a linear framework. Thus, it may still be possible that non-linear models perform well on the macroeconomic variables; in other words, Models III and IV could be inferior to Models V to VII.

Table 8 reports the results of our analysis in the form of the performance criteria spelled out earlier. In the forth column, we have highlighted the model with the highest average correlation (IC) between predicted and actual returns (calculated across the 55 industries and averaged over the out-of-sample periods). This measure is

[7] Recall that each run of the *n*-NN algorithm was done using *n*=2, 5, 10, 15, 20, 40, 60, 80 and 100. These results are available upon request.

[8] We also ran an n-NN neighbour version of the model to see if a non-linear approach was beneficial. The dependent variable is, of course, industry returns data, and the X-matrix in this case only has industry-specific variables and it's 2- and 3- period lags included; thus, X has nine inputs. The results were far inferior to the OLS results and so are omitted. However, they are available upon request.

commonly reported and used by asset managers. The variance of the IC is also calculated, and the “Sharpe Ratio” - which is the average IC divided by the standard deviation of IC – determines how evenly distributed correlation is over time. A good model would predict actual returns well in every period (high “Sharpe Ratio”) rather than predicting accurately in some periods in between periods of observing little or no correlation between estimated and predicted returns.

Summarising our results so far, it appears that when a nearest neighbour algorithm is used (i.e. in Models VI and VII), using few nearest neighbours gives better performance. Hence, superior predictive patterns in the explanatory variables are identified using relatively few observations, which are close in distance to the present state. Using more information, which is further away (by our definition, but not necessarily in time) from the present state, appears to add more noise than predictive power.

We also note changes when using the different macroeconomic data sets. For instance, when we replace the change in Risk Premium by the change in Excess Dividend Yield in our information set (when moving from Macroset 1 to Macroset 2 and from Macroset 3 to Macroset 4 - see Appendix), performance drops in almost every case. Also, on the whole, using first differences of the macroeconomic data appears bad at identifying winning and losing industries as compared with the data in levels. This suggests a reduced role for “macroeconomic surprises” when modelling industry returns.

In all, the best forecasting procedures for industry returns measured by IC is Model IV, which first modelled industry-specific effects using OLS and then regressed the residuals on the macroeconomic variables, also using OLS. In particular, Model IV has an average IC of 8.7% for Macroset 1, which only used one differenced variable, namely that of the change in the Risk Premium. However, is interesting to see that when one uses a Sharpe Ratio argument, which takes into account the volatility of the correlation between predicted and observed returns, Model IV is no longer optimal. In particular, it is beaten by Model I, which uses plain OLS on all variables (Sharpe Ratio 0.444 versus 0.407). More convincingly, Model IV, which applies a nearest neighbour algorithm to the macroeconomic factors, had a Sharpe Ratio of 0.493, and we regard this as the best overall model examined.

## 6. Conclusion

We have investigated forecasting industry returns using both macroeconomic and industry-specific information. There is other information than that we have used, which could be used to build an industry returns model. Theory suggests input and output prices, measures of economies of scale and intra-industrial competitiveness, and much other available information could be utilised. Although we do not derive a specific theoretical model for industry returns, but focus on the forecasting properties of the econometric models we examine, we have given an overview of how this may be done, so a formal theoretical rationale may be provided for the techniques utilised in this paper.

Motivated by preliminary investigations, we have looked at seven different models, which are defined from OLS and a non-linear technique known as an *n*-NN algorithm. Overall, the *n*-NN algorithms showed reasonable results without being entirely convincing. There seemed little stability in predictability when changing the number of nearest neighbours used and no clear improvement over OLS (or two-stage OLS techniques) was observed. Furthermore, results for all models varied depending on changes in variable definition as we moved from one macroeconomic data set to another. In particular, there was an observed deterioration of performance when replacing the change in Risk Premium with the change in the Excess Dividend Yield, and an even larger decrease in performance when the most common macroeconomic variables were replaced by their first differences.

However, a procedure which is based on regressing on industry factors, and using the nearest neighbour algorithms on the residuals, produced model with an IC of up to 8.1% based on one month ahead rolling forecast. This magnitude, corresponding to an $R^2$ of about 0.5%, may sound very small to statisticians working with non-financial data, but in fact this would be seen by practitioners as a reasonable result and one upon which profitable investment strategies could be based. Although other models narrowly beat this average, no model is superior when also taking into account the desire to have regular out-performance rather than more sporadic prediction success. In particular, the much lower variance of the IC means that the model will deliver good forecasting performance fairly frequently. Thus, an investment strategy based on the model can be kept in place without incurring the high transaction costs, which would result if the model became unreliable and the strategy had to be cancelled.

# Bibliography

**Arroyo, C. R., “**Testing a Production-Based Asset-Pricing Model”, *Economic Inquiry*, Vol. 34 (1996) 357–377.

**Basu, P., and H. D. Vinod,** “Mean Reversion in Stock Prices: Implications from a Production Based Asset Pricing Model”, *Scandinavian Journal of Economics*, 96(1), (1994) 51–65.

**Beller, K. R., Kling, J. L. and H. D. Vinod,** “Are Industry Returns Predictable”, *Financial Analysts Journal*, September/October 1998, 42–57.

**Breeden, D. T., Gibbons, M. T. and R. H. Litzenberger,** “Empirical Tests of the Consumption-oriented CAPM”, *Journal of Finance* 44(2) (June 1989), 231–262.

**Boudoukh, J., Richardson, M. and R. F. Whitelaw,** “Industry Returns and the Fisher Effect”, *Journal of Finance* 49(5) (December 1994), 1595-1616.

**Campbell, J., Lo, A. and A. McKinlay, “***The Econometrics of Financial Markets”*, Princeton University Press, (1997).

**Cleveland, W. S., “**Robust Locally Weighted Regression and Smoothing Scatterplots”, *Journal of the American Statistical Association*, Vol. 4, No. 368 (1979), 829–838.

**Cochrane, J. H.,** “Production-Based Asset Pricing and the Link between Stock Returns and Economic Fluctuations”, *Journal of Finance* 46(1) (1991), 209–237.

**Fama, E. F. and K. R. French, “**Business Conditions and Expected Returns on Stocks and Bonds”, *Journal of Financial Economics* 25 (1989), 23–49.

**Ferson, W. E. and C. R. Harvey,** “The Variation in Economic Risk Premiums”, *Journal of Political Economy* 99(2) (April 1991), 385-415.

**Grauer, C. W. J., Hakanson, N. H. and F. C. Shen,** “Industry Rotation in the US Stock Market: 1934-1986 Returns on Passive, Semi-Passive and Active Strategies”, *Journal of Banking and Finance* 14(2/3) (August 1991), 513-538.

**Kale, J. K., Hakanson, N. K. and G. W. Platt,** “Industry vs. Other Factors in Risk Prediction”, *Finance Working Paper 201, Haas Business School, University of California at Berkeley.*

**Lo, A. and A. C. MacKinlay,** “Maximising Predictability in the Stock and Bond Markets”, *MIT Laboratory for Financial Engineering Working Paper LFE-1019-96.*

**King, B. F.,** “Marker and Industry Factors in Stock price Behaviour”, *Journal of Business* 39(1), Part II (January 1966), 139–190.

**Manski, C. F.,** “Regression”, *Journal of Economic Literature,* 29 (March 1991), 34–50.

**Rosenberg, B.,** “Extra-market Components of Covariance in Securities Markets”, *Journal of Financial and Quantitative Analysis* 9(2) (March 1974), 263–274.

**Sorensen, E. H. and T. Burke,** “Portfolio Returns from Active Industry Group Rotation, *Financial Analysts Journal* 42(5) (September/October 1986), 43–50.

**Reilly, F. K. and E. F. Drzycimski,** “Alternative Industry Performance and Risk”, *Journal of Financial and Quantitative analysis* 9(3) (June 1974), 423–474.

**Stone, C. J.,** “Consistent Non-Parametric Regression”, *Annals of Statistics* 5 (1977), 595–645.

# APPENDIX

Table 1: Forecasting Models Examined

| Model | Description |
|---|---|
| **I** | OLS on industry variables |
| **II** | OLS on macroeconomic variables |
| **III** | OLS on all variables |
| **IV** | OLS on industry, then OLS of residuals on macroeconomic variables |
| **V** | OLS on macroeconomic, then OLS of residuals on industry variables |
| **VI** | OLS on industry, then *n*-NN of residuals on macroeconomic variables |
| **VII** | *n*-NN on all variables |

Table 2: Summary Statistics of Time Series of Industry Returns

| | Mean | Standard Deviation |
|---|---|---|
| **Mean** | 0.011 | 0.054 |
| **St. Dev** | 0.002 | 0.009 |
| **Minimum** | 0.005 | 0.040 |
| **Maximum** | 0.016 | 0.092 |
| | | |
| | **Minimum** | **Maximum** |
| **Mean** | -0.287 | 0.147 |
| **St. Dev** | 0.050 | 0.035 |
| **Minimum** | -0.428 | 0.086 |
| **Maximum** | -0.133 | 0.270 |
| | | |
| | **1st Quartile** | **3rd Quartile** |
| **Mean** | -0.019 | 0.043 |
| **St. Dev** | 0.007 | 0.006 |
| **Minimum** | -0.047 | 0.030 |
| **Maximum** | -0.009 | 0.057 |

Table 3: Summary Statistics of Cross-Sectional Industry Returns

| | Mean | Standard Deviation |
|---|---|---|
| **Mean** | 0.011 | 0.033 |
| **St. Dev** | 0.043 | 0.008 |
| **Minimum** | -0.282 | 0.020 |
| **Maximum** | 0.122 | 0.070 |
| | | |
| | **Minimum** | **Maximum** |
| **Mean** | -0.084 | 0.100 |
| **St. Dev** | 0.060 | 0.051 |
| **Minimum** | -0.428 | -0.133 |
| **Maximum** | 0.035 | 0.270 |
| | | |
| | **1st Quartile** | **3rd Quartile** |
| **Mean** | -0.006 | 0.028 |
| **St. Dev** | 0.045 | 0.043 |
| **Minimum** | -0.322 | -0.250 |
| **Maximum** | 0.101 | 0.147 |

Table 4: Autocorrelation Parameters

| **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
|---|---|---|---|
| 0.05 | 0.08 | -0.18 | 0.24 |
| | | | |
| | **Positive** | **Negative** | **Total** |
| **Significant Two-tailed Test** | 6 | 1 | 7 |
| **Not Significant** | 35 | 13 | 48 |
| **Total** | 41 | 14 | 55 |

Table 5: Interpolation Dates

| **Variable** | **Dates** |
|---|---|
| **Z1** | 2/82 – 12/82 |
| **Z2** | 2/82 – 1/85 |
| **Z3** | 7/98 – 8/98 |
| **Z6** | 8/98 |
| **Z7** | 8/98 |
| **Z8** | 2/82 |

Table 6: R-square Summary Table of Model I

| | |
|---|---|
| **Mean** | 0.076 |
| **Minimum** | 0.004 |
| **Maximum** | 0.211 |
| **St. Deviation** | 0.040 |

Table 7: Residual Correlation Summary Table

| | |
|---|---|
| **Mean** | -0.007 |
| **Minimum** | -0.595 |
| **Maximum** | 0.742 |
| **Standard Deviation** | 0.185 |

Table 8: Summary of Results of Models I-VII

| **Model** | **Macroset** | ***N*** | **Mean Correlation** | **Variance of Correlation** | **Sharpe Ratio** |
|---|---|---|---|---|---|
| **I** | **N/A** | **N/A** | 0.073 | 0.027 | 0.444 |
| **II** | **1** | **N/A** | 0.065 | 0.049 | 0.294 |
| | **2** | **N/A** | 0.047 | 0.038 | 0.241 |
| | **3** | **N/A** | 0.018 | 0.041 | 0.089 |
| | **4** | **N/A** | 0.008 | 0.041 | 0.040 |
| **III** | **1** | **N/A** | 0.072 | 0.050 | 0.322 |
| | **2** | **N/A** | 0.069 | 0.042 | 0.337 |
| | **3** | **N/A** | 0.055 | 0.029 | 0.333 |
| | **4** | **N/A** | 0.059 | 0.029 | 0.346 |
| **IV** | **1** | **N/A** | ***0.087*** | 0.046 | 0.407 |
| | **2** | **N/A** | 0.078 | 0.040 | 0.390 |
| | **3** | **N/A** | 0.064 | 0.034 | 0.347 |
| | **4** | **N/A** | 0.060 | 0.034 | 0.325 |
| **V** | **1** | **N/A** | 0.081 | 0.047 | 0.374 |
| | **2** | **N/A** | 0.076 | 0.038 | 0.390 |
| | **3** | **N/A** | 0.060 | 0.030 | 0.346 |
| | **4** | **N/A** | 0.063 | 0.029 | 0.370 |
| **VI** | **1** | **10** | 0.081 | 0.027 | ***0.493*** |
| | **2** | **2** | 0.063 | 0.032 | 0.352 |
| | **3** | **2** | 0.078 | 0.032 | 0.436 |
| | **4** | **6** | 0.066 | 0.026 | 0.409 |
| **VII** | **1** | **6** | 0.056 | ***0.025*** | 0.354 |
| | **2** | **15** | 0.059 | 0.036 | 0.311 |
| | **3** | **2** | 0.015 | 0.046 | 0.070 |
| | **4** | **6** | 0.001 | 0.036 | 0.001 |

## Macroset 1

| | ***Variable*** | ***Definition*** | ***Datastream Code*** |
|---|---|---|---|
| **Z1** | **Term Spread** | Yield on 2-year bond *minus*<br>Yield on 30-year bond<br>*(includes lags up to and including 3 periods)* | USBDS2Y(RY)<br>USBD30Y(RY) |
| **Z2** | **Credit Premium** | Return on CCC bond index *minus*<br>Return on AAA bond index<br>*(includes lags up to and including 3 periods)* | USBC2A710<br>79499650 |
| **Z3** | **Inflation** | Log-difference of the Price Index<br>*(includes lags up to and including 3 periods)* | USOCPCONF |
| **Z4** | **Trailing Market Vol.** | Level of Market Volatility over last 24 months | USS&PCOM |
| **Z5** | **Unemployment Rate** | Level of Unemployment<br>*(includes lags up to and including 3 periods)* | USUNRATEE |
| **Z6** | **Change in Leading Indicator** | Log-difference of the Leading Indicator[9] | USLEADIN |
| **Z7** | **Change in the Coincidence Indicator** | Log-difference of the Coincidence Indicator[10] | USCOINN |
| **Z8** | **Change in Risk Premium** | Change in (Earnings Yield *minus*<br>Yield on 30-year bond) | USS&PCOM<br>USBD30Y(RY) |

## Macroset 2

Macroset 2 is defined as in Macroset 1, except **Z8** is replaced by:

| | ***Variable*** | ***Definition*** | ***Datastream Code*** |
|---|---|---|---|
| **Z9** | **Change in Excess Dividend Yield** | Change in (Dividend Yield *minus*<br>Yield on 30-year bond) | S&PCOMP(DY)<br>USBD30Y(RY) |

[9] More detailed information on the Leading Indicator can be found in Datastream.

[10] More detailed information on the Coincidence Indicator can be found in Datastream.

## Macroset 3

Macroset 3 is defined as in Macroset 1, except we replace **Z1**, **Z2**, **Z4** and **Z5** by the changes in these variables; i.e. use **Z3**, **Z6**, **Z7** and **Z8** together with:

| | ***Variable*** | ***Definition*** | ***Datastream Code*** |
|---|---|---|---|
| **Z1'** | **Change in Term Spread** | Change (Yield on 2-year bond *minus* Yield on 20-year bond)<br>*(includes lags up to and including 3 periods)* | USBDS2Y(RY)<br>USBD30Y(RY) |
| **Z2'** | **Change in Credit Premium** | Change (Return on CCC bond index *minus* Return on AAA bond index)<br>*(includes lags up to and including 3 periods)* | USBC2A710<br>79499650 |
| **Z4'** | **Change in the Trailing Market Volatility** | Change (Level of Market Volatility over last 24 months) | USS&PCOM |
| **Z5'** | **Change in the Unemployment Rate** | Change in the Level of Unemployment<br>*(includes lags up to and including 3 periods)* | USUNRATEE |

## Macroset 4

**Macroset 4** is defined as MacroSet 3, but where **Z9** replaces **Z8**.