Timing Small versus Large Stocks

Using artificial intelligence to decide when to be long or short.

Jean-François L'Her, Tammam Mouakhar, and Mathieu Roberge

Ever since Fama and French [1992] published their seminal work, it has been widely accepted that small stocks should yield a higher return on average than large ones. One key point to remember is that this conclusion holds only on average, and not all the time. A small-size tilt is profitable in the long run, but can be detrimental for performance over medium-term periods. The Fama-French small-minus-big (SMB hereafter) U.S. premium over July 1926-June 2005 was 1) negative for 45 of the 75 five-year rolling periods, and 2) positive just a little over half of the time (50.84%). Such observations pave the way for an investigation into when it is preferable to hold large rather than small stocks, and vice versa.

Using ex post data, Reinganum [1999] and Ahmed, Lockwood, and Nanda [2002] document potential paper gains to be derived from size-timing strategies. Kester [1990] estimates the transaction costs from such size-timing strategies, and shows that profitability would hold even after reasonable transaction costs.

Most studies examining size timing on an ex ante basis are based on parametric estimation methods. In the U.K., Levis and Liodakis [1999] propose a size-timing approach for the SMB premium that relies on multivariate ordinary least squares (OLS) and logit regressions. Both recursive out-of-sample forecast models (repeated annually) use lagged macroeconomic data. They outperform a buy-and-hold benchmark over the 1974-1997 period even after accounting for transaction costs.

More recently, in the U.S., Cooper, Gulen, and Vassalou [2001] test recursive OLS models (repeated monthly) using filter rules and long-short size deciles over the out-of-sample 1963-1998 period. They find evidence of size-timing strategies.

JEAN-FRANÇOIS L'HER
is a senior vice president, Investment Policy Research, at the Caisse de Dépôt et Placement du Québec in Montréal.
jlher@lacaisse.com

TAMMAM MOUAKHAR
is a research advisor at the Caisse de Dépôt et Placement du Québec.
tmouakhar@lacaisse.com

MATHIEU ROBERGE
is a research advisor at the Caisse de Dépôt et Placement du Québec.
mroberge@lacaisse.com

In the same vein, Amenc et al. [2003] test recursive OLS models using real indexes over an out-of-sample 2000-2002 period. They also document profitable size-timing strategies.

These parametric methods have the advantage of using parsimonious models that clearly identify functional forms and the marginal contribution of each variable, but they also suffer from several flaws—restrictive distribution assumptions, linear functional forms, and sensitivity to outliers. Size-timing is a good candidate for non-parametric methods because even if there were evidence of the predictive power of lagged macroeconomic variables, we do not have a priori hypotheses on the variables and on the functional form of the relation.

To our knowledge, there is only one application of a non-parametric style-timing method—Kao and Shumaker [1999] build a successful value-growth timing model based on a recursive partitioning algorithm.

We focus mainly on comparing the predictive power of different approaches based on artificial intelligence (AI) with respect to the size-timing problem. We compare the ability of recursive partitioning (RP), neural networks (NNs), and genetic algorithm (GA) approaches to correctly time the size premium, and examine the resulting performance. We also examine the predictive power as well as the performance of a strategy based on a consensus of the three approaches.

METHODS

Three popular AI models, recursive partitioning, neural networks, and genetic algorithms, are initially trained during an in-sample period from January 1975 through December 1989. Out-of-sample predictive power and investment performance are then assessed for January 1990-December 2004.

Advantages and Disadvantages of AI Methods

While we do not conduct an extended review of the benefits and drawbacks associated with AI methods, it is appropriate that we discuss some of the main advantages and disadvantages of these methods. First, all these approaches are non-parametric, so they rely on much less restrictive distribution assumptions. This could be viewed as an advantage, considering that such hypotheses are often not valid when we deal with financial and economic data. AI approaches do not require the relation between inputs and outputs to be linear to follow a specific functional form or that the explanatory variables be independent.

Another advantage of AI approaches is that they are not based on formalized knowledge; they rely on their own learning processes. This is particularly useful as we deal with problems where we have skimpy knowledge.

Such methods also have the advantage that they accept a large number of candidate explanatory variables simultaneously. Moreover, AI methods, particularly RP, can provide a hierarchy of independent variables.

AI techniques are not without pitfalls. The main risk is overfitting, which occurs when the model fits the in-sample data too closely, and later fails to generalize the results of the model on out-of-sample data. Overfitting is related to, and is often a consequence of, data snooping, which arises "when a researcher chooses what to do and how to do it in the light of what others have done using similar data" (Black [1995]).

Campbell, Lo, and MacKinlay [1997] warn us of the dangers of non-experimental inference in financial economics. With the scrutiny and reuse of the same set of data, especially in the analysis of time series, there is always the risk that one will not distinguish the spurious from the substantive (see White [2000]). As Leinweber [2003] correctly points out: "If you look hard enough, you will always find something that looks great statistically, but makes no sense."

To account for this problem, one solution is to feed the model only with variables that are theoretically sound. Still, the AI methods might yield relations that are contrary to theory or intuition.

AI approaches also present some operational difficulties—they are very data-consuming in their learning phases. Therefore, they cannot easily be applied to problems where data are limited. Except in the case of recursive partitioning, interpretation of the results is very challenging. Many consider these approaches to be black boxes.

Recursive Partitioning

RP (also known as binary tree, decision tree, or classification tree) is a method that repeatedly splits a subset of observations into two descendant subsets. The idea is to grow a tree so that descendant subsets are "purer" than the parent subset.[1]

Two key elements govern the construction of the tree: selection of the best variable to use at each split point,

and choice of a criterion that lets us determine when to stop splitting (hence creating a terminal node) or when to split further. Criteria such as a minimal number of observations in each terminal node or a minimum level of purity are typically used to determine whether a node should be split further. Once the tree has been built, all terminal nodes are assumed to classify an observation according to the outcome most frequently encountered at that node.

A more detailed description can be found in Breiman et al. [1984]. Additional information on specific applications to finance appears in Kao and Shumaker [1999] and Sorensen, Miller, and Ooi [2000].

For our particular application, numerous tree configurations are tested in an attempt to maximize the in-sample hit rate (measured as the number of correct predictions divided by the total number of predictions made). A constraint forces all trees to grow in such a fashion that at least ten observations end up in any single terminal node. This constraint is enforced in an effort to minimize the risk of overfitting.

The Gini diversity index is used to determine the best split at each step. No prior probabilities are used.

Neural Networks

An artificial neural network uses a set of concepts based on the biological neural system. It is composed of artificial neural nodes (or processing elements) interconnected in a way that allows parallel processing. Each node receives an input signal, processes the input through a transfer function, and transforms it into an output.

The NN architecture is flexible, and is characterized by the number of layers, the number of neurons in each layer, the interconnection between the neurons, and the transfer function of each neuron. The multiple-layer structure is composed of an input layer (the first layer where external inputs are received), the hidden layer (which can consist of one or more intermediate layers), and the output layer (the last layer, which delivers the output). All nodes in each preceding layer are connected to the next layer by arcs.

In order to process the output, NNs need to be trained. The process of training (or learning) consists of adjusting the parameters (the arc weights) of the neural network model by minimizing the difference between the NN output and the known in-sample target according to a learning algorithm. Once the NN has been trained on the in-sample set, it is applied to an out-of-sample set in order to evaluate the predictive capability of the model.

A more complete description of this method can be found in Medsker, Trippi, and Turban [1993]. Kryzanowski, Galler, and Wright [1993] and Kingdon and Feldman [1995b] present interesting applications to finance.

To determine the structure of our NN, the learning period is subdivided into two subperiods. Different structures are trained in the first subperiod (1975-1983). The training phase is conducted in order to minimize the difference between the NN monthly outputs and the associated actual in-sample return. Models are then tested on the second subperiod (1984-1989).

The model that is selected to be applied to the out-of-sample set is the one that minimizes the mean squared errors between the NN outputs and the 1984-1989 in-sample returns.

Genetic Algorithms

A genetic algorithm is a stochastic search technique that is based on the theories of natural selection and genetics, and is designed to optimize a fitness function. The approach randomly generates a set of potential solutions. The set is known as the *population*, and each potential solution is called an *individual*. Each individual, also called a *chromosome* by some, is represented by a codified binary vector.

The GA first evaluates the fitness function for each individual in the population, and then the processes of natural selection, reproduction (cross-over), and mutation operate to make the population evolve to the next generation. The process goes on until a predetermined stopping criterion is reached. Typical stopping criteria include reaching a fixed number of generations or a targeted homogeneity in the population.

Natural selection is the process through which individuals with the highest fitness are automatically retained in the newly generated population (these individuals are cloned). Cross-over is the operation through which two individuals are mixed in an attempt to obtain a new individual with better fitness than that of both its parents. Finally, mutation represents a rare and random innovation in the binary sequence of an individual.

The main challenge in GA is defining an appropriate fitness function. GAs are very flexible, and provide many parameters that can be calibrated to better fit the problem under study.

For further details, see Bauer [1994]. Kingdon and Feldman [1995a] provide a good example of how GA could be applied to finance.

For our size-timing application, we specify the problem as a linear function of 20 variables such that:

$$\text{If } \beta_1 V_1 + \beta_2 V_2 + \beta_3 V_3 + \ldots + \beta_{20} V_{20} > \pi:$$
Model favors small - cap
$$\text{If } \beta_1 V_1 + \beta_2 V_2 + \beta_3 V_3 + \ldots + \beta_{20} V_{20} \leq \pi:$$
Model favors large - cap

V_1 through V_{20} correspond to the values taken by the different predictive values at any date, and β_1 through β_{20} as well as π are values optimized by the genetic operators such as reproduction and mutation. β_1 through β_{20} and π are optimized to maximize the in-sample return. We use a population size of 100 individuals and a fixed number of 100 generations as stopping criteria. The chromosomes are long binary strings representing the 20 parameters to be optimized in the equation.

The translation into and back from the binary representation is conducted automatically by Matlab. The fitness function evaluated with each chromosome is maximization of the return obtained when applying the rule depicted in the equation to the in-sample period. Reproduction is performed with an elite count of ten and a cross-over fraction of 0.8.[2]

APPLICATION

Many authors suggest that the positive small-minus-big premium is related to the fundamental risk in the economy (see Fama and French [1993, 1996] and Jensen, Johnson, and Mercer [1998]). They shed light on the links between the fluctuation of the SMB premium and the economic cycle. Fama and French [1993, 1996] suggest that the earnings of small firms could be more sensitive to economic conditions. Jensen, Johnson, and Mercer [1998] note that the size of the SMB premium is linked to a tightening of monetary policy. When the Fed raises interest rates, monetary policy becomes more restrictive and the premium narrows. Conversely, they find that the SMB premium widens with a more accommodating monetary policy.[3]

Other authors use lagged macroeconomic variables to design a style-timing model focusing on size.[4] The macroeconomic variables most frequently used to predict the SMB premium are related to interest rates. Variables such as term spreads and the level of the short-term rate are used in Levis and Liodakis [1999], Cooper, Gulen, and Vassalou [2001], and Amenc et al. [2003].

Variables related to the stock market are also common. Levis and Liodakis [1999], Cooper, Gulen, and Vassalou [2001], and Amenc et al. [2003] use the dividend yield, while Kao and Shumaker [1999] consider the differential between the earnings-to-price ratio of the S&P 500 and the long-term bond yield. Other authors document the predictive power of New York Stock Exchange volume, the market premium, and corporate credit.

Some also use classic macroeconomic variables such as GDP growth and inflation. Finally, Amenc et al. [2003] examine more original variables, including commodity prices, exchange rates, and the level of consumer confidence.

We use the 20 variables described in Exhibit 1 to feed the three AI approaches.

Models for the Three Approaches

We first train the three AI approaches over January 1975-December 1989. The recursive programming results are the easiest to interpret. The RP tree has 28 terminal nodes. The shortest path to a terminal node is two splits, while the longest requires ten splits. Thirteen variables are used as split criteria at least once within the tree structure: CREDIT, TBILL, COIN, LEAD, MOM, DIV, CAP, IND, CSI, ISM, SAV, PPI, and TRAD. The most important variable (the first split criterion) is the U.S. Conference Board Coincident Economic Indicators Index (COIN). The next split on the left uses the variable DIV, and the next on the right the MOM.[5]

The NN structure is composed of one hidden layer with 20 nodes. Its architecture corresponds to the standard neural network architecture for non-linear regression (see Haykin [1994]). The hidden units use hyperbolic tangent activation functions, and the output unit has a linear activation function. We use the Levenberg-Marquardt back-propagation algorithm to train the network. Bayesian regularization is implemented in this algorithm to prevent overfitting and to produce a network that can be fully generalized.

Interpretation of the NN model is more complex, because the multi-layer structure involves a high degree of black box. We can only roughly measure the importance of the different input variables. The most suitable method for comparing the relative importance of the variables is

EXHIBIT 1
Variables Used For Size-Timing

Variable	Definition	Source
TERM	Yield Differential (10Y Treasury Bond – 1Y Treasury Bond)	Calculated using series from Federal Reserve's FRED database
CREDIT	Yield Differential (Baa Corporate Bond – Aaa Treasury Bond)	Calculated using series from Federal Reserve's FRED database
TBILL	3-Month Treasury Bill secondary market rate	Federal Reserve's FRED database
INFL	One-month change in the CPI All Item	Calculated using series from Federal Reserve's FRED database
COIN	US Conference Board Coincident Economic Indicators Index	Datastream
LEAD	US Conference Board Leading Economic Indicators Index	Datastream
MOM	Six-Month momentum of the S&P 500	Calculated using series from RIMES database
EARN	Price Earning of the S&P 500	Calculated using series from RIMES database
DIV	Dividend Yield of the S&P 500	DRI Financial Markets Indexes
GSCI	Monthly return of the Goldman Sachs Commodity Index	Calculated using series from RIMES database
CAP	Capacity Utilization Rate: Total Industry	Federal Reserve's FRED database
IND	One-Month change in the Industrial Production Index	Calculated using series from Federal Reserve's FRED database
CSI	University of Michigan Consumer Sentiment Index	Federal Reserve's FRED database
ISM	ISM Manufacturing : PMI composite Index	Federal Reserve's FRED database
SAV	Personal Saving Rate	Federal Reserve's FRED database
M2	One-Month Change in M2 Money Stock	Calculated using series from Federal Reserve's FRED database
CEXP	Monthly Real growth of Personal Consumption Expenditure	Calculated using series from Federal Reserve's FRED database
PPI	One-Month change in the PPI Industrial Commodities	Calculated using series from Federal Reserve's FRED database
NYSE	One-Month change in the NYSE Volume	Datastream
TRAD	One-Month change in the Trade-Weighted exchange rate (Board)	Calculated using series from Federal Reserve's FRED database

the connection weight approach (see Olden, Joy, and Death [2004]). The results show that INFL, TERM, COIN, EARN, and MOM are the factors that contribute the most.

In the GA approach, the variables TERM, COIN, LEAD, MOM, EARN, DIV, GSCI, CAP, ISM, M2, PPI, NYSE, and TRAD are all associated with a positive coefficient.[6] The seven other variables, CREDIT, TBILL, INFL, IND, CSI, SAV, and CEXP, have negative coefficients. This implies that, ceteris paribus, a positive change in the first set of variables will tend to favor small-caps and a positive change in the second set will favor large-caps.

Predictive Ability and Investment Performance

Panel A of Exhibit 2 presents the performance of the three size-timing strategies over January 1990-December 2004. The performance of the classic small-minus-big strategy is also shown for comparison. Two of the three style-timing strategies (RP and NN) dominate SMB in terms of return per unit of risk, but the third method (GA) fails to time the size premium.

The performance of the two successful timing strategies in terms of return per unit of risk is driven by average excess returns that are more than 400 basis points higher (6.25% for RP and 6.91% for NN versus 1.52% for SMB). The three AI timing strategies and the SMB are roughly as volatile, with a difference of under 20 basis points between the highest and the lowest.

The relative performance of timing strategies is greatly influenced by extreme observations. Indeed, our timing strategies imply binary decisions: short or long positions in the small- or large-caps. Therefore, making the wrong decision in a given month can convert a wonderful month into a nightmarish one, and vice versa. The highest absolute value of the SMB is seen in February 2000 (21.49%). This atypical observation translates into a maximum monthly gain for the RP and NN timing strategies, but turns into a maximum loss for the GA.

Another way to analyze the performance of the three AI approaches is to compare their hit rates, which show if a predictive model is right on average. A high rate with a disappointing return implies that a few wrong decisions hurt much more than the more frequent good decisions helped. In this case, a poor return might simply be due to misfortune. This could be the case for the GA timing model, which posts a satisfying hit rate of 54.4%, higher than the SMB hit rate of 52.8%. The RP and NN approaches reach higher hit rates of 56.1% and 57.0%.

All results so far are obtained by training the different AI approaches on a static learning sample. That is, as we move along the test sample (out-of-sample), new predictions do not take into account the newest information available, and are not based on recent available data. It seems logical to examine whether expandable learning samples could improve prediction. We thus reexamine timing strategies trained on expandable learning samples, or *dynamic* strategies. This idea corresponds to the concept of an

EXHIBIT 2

Size-Timing Strategies Trained in a Static Learning Sample (1975-1989)

	SMB	RP	NN	GA	Consensus
Panel A: Total period					
			1990–2004		
Annualized geometric return	1.52%	6.25%	6.91%	0.18%	7.77%
Annualized standard deviation	13.21%	13.08%	13.05%	13.23%	13.01%
Return per unit of risk	0.12	0.48	0.53	0.01	0.60
Biggest monthly loss	-16.69%	-16.69%	-7.71%	-21.49%	-16.69%
Biggest monthly gain	21.49%	21.49%	21.49%	8.49%	21.49%
Hit ratio	52.78%	56.11%	56.98%	54.44%	57.78%
Average return when correct	2.74%	2.91%	2.93%	2.56%	2.93%
Average return when wrong	-2.64%	-2.41%	-2.38%	-2.86%	-2.37%
Recommended switches	1	32	55	62	56
Max trading cost (=Return$_{SMB}$) (bp)	*N/A*	111.45	54.48	N/A ‡	145.96
Max trading cost (=0%) (bp)	*N/A*	281.01	81.75	4.42	198.93
H-M test p-stat	*N/A*	1.10	1.13	1.11	1.15
H-M test p-value	*N/A*	0.095*	0.051*	0.091*	0.029**
Panel B: Subperiods					
			1990–1994		
Annualized geometric return	1.28%	9.23%	-0.69%	5.71%	10.32%
Annualized standard deviation	8.35%	7.92%	8.36%	8.18%	68.83%
Return per unit of risk	0.15	1.17	-0.08	0.70	-0.07
Hit ratio	55.00%	58.33%	47.46%	56.67%	58.33%
			1995–1999		
Annualized geometric return	-4.50%	2.73%	2.70%	6.07%	5.51%
Annualized standard deviation	12.56%	12.51%	12.57%	12.50%	12.46%
Return per unit of risk	-0.36	0.22	0.21	0.49	0.44
Hit ratio	43.33%	50.00%	56.67%	58.33%	55.00%
			2000–2004		
Annualized geometric return	8.19%	6.89%	19.81%	-10.32%	8.89%
Annualized standard deviation	11.04%	11.25%	11.32%	11.56%	11.41%
Return per unit of risk	0.74	0.61	1.75	-0.89	0.78
Hit ratio	60.00%	60.00%	66.67%	48.33%	60.00%

** indicates statistical significance at the 10% level.*

‡We do not report the maximum trading cost, because the GA excess return over the SMB is negative.

evolving tree, as discussed in Sorensen, Miller, and Ooi [2000].

In our case, we let the AI models evolve at the beginning of each year rather than each month. Therefore, each predictive model is used to make 12 monthly predictions before it is updated. The results for evolving predictive models are presented in Exhibit 3.

Only one method, the genetic algorithm, yields better results with an expandable learning sample. The return associated with neural networks worsens only slightly, but the recursive partitioning return drops from 6.3% to 3.8%. Even if two of the individual timing strategies perform more poorly, all three methods now post returns per unit of risk that are more than twice as high as SMB strategy returns.

An important point to note is that the weakening of the two methods could well be due to the same kind of misfortune that affected the GA earlier; that is, a few wrong decisions cancel part of the value added in the more frequent good decisions. After all, the hit rate increases over the static method for all three AI approaches.

Subperiod Analysis and Nature of the Consensus

As we look at the full-period out-of-sample results, we might not see important differences over subperiods. Panel B of Exhibits 2 and 3, presents results for the three timing approaches over three subperiods: 1990-1994, 1995-1999, and 2000-2004.

Note three interesting observations for static models (Exhibit 2, Panel B): 1) the classic SMB strategy is dominated by at least one timing strategy in every subperiod; 2) all three AI approaches beat the SMB in two of the three subperiods; and 3) each AI approach has days of glory, when it beats the other two methods by a wide margin.

In the first five-year period, RP outperforms the classic SMB strategy by roughly 800 basis points. The second-best timing model for the period, the GA, falls short of RP by more than 3.50 percentage points. In this subperiod, the NN is dominated by the classic SMB strategy.

In the second subperiod, the classic SMB has a significant negative return of 4.5% per year, while the RP, NN, and GA approaches post returns of 2.7%, 2.7%, and 6.1%, respectively. The GA is clearly the best strategy here.

Finally, in the third period, the NN posts a spectacular average annualized return of almost 20%. This return is more than twice the return of SMB or RP. While the NN shines, the GA plunges completely into darkness, with an average annualized return of −10.3%, almost 20 percentage points short of the SMB benchmark. This sole five-year subperiod is responsible for the overall disappointing performance of the GA, despite the other two subperiods when it performs very well.

The dynamic models support similar observations for the subperiod results (Panel B of Exhibit 3). As in the

EXHIBIT 3

Size-Timing Strategies Trained on an Expandable Learning Sample (1975 to Prediction)

	SMB	RP	NN	GA	Consensus
Panel A: Total period					
			1990–2004		
Annualized geometric return	1.52%	3.80%	6.37%	4.14%	8.60%
Annualized standard deviation	13.21%	13.16%	13.07%	13.15%	12.96%
Return per unit of risk	0.12	0.29	0.49	0.31	0.66
Biggest monthly loss	-16.69%	-21.49%	-21.49%	-21.49%	-21.49%
Biggest monthly gain	21.49%	16.69%	16.69%	16.69%	16.69%
Hit ratio	52.78%	57.78%	59.44%	58.33%	64.44%
Average return when correct	2.74%	2.66%	2.76%	2.66%	2.68%
Average return when wrong	-2.64%	-2.73%	-2.60%	-2.74%	-2.72%
Recommended switches	1	68	54	61	46
Max trading cost (=Return_{SMB}) (bp)	*N/A*	54.48	114.21	62.64	165.25
Max trading cost (=0%) (bp)	*N/A*	81.75	170.82	99.22	267.45
H-M test p-stat	*N/A*	1.16	1.20	1.18	1.30
H-M test p-value	*N/A*	0.024**	0.006***	0.008***	0.000***
Panel B: Subperiods					
			1990–1994		
Annualized geometric return	1.28%	3.15%	5.10%	6.63%	9.45%
Annualized standard deviation	8.35%	8.30%	8.21%	8.12%	7.90%
Return per unit of risk	0.15	0.38	0.62	0.82	1.20
Hit ratio	55.00%	56.67%	61.67%	61.67%	68.33%
			1995–1999		
Annualized geometric return	-4.50%	4.70%	1.78%	3.25%	5.42%
Annualized standard deviation	12.56%	12.50%	12.54%	12.55%	12.51%
Return per unit of risk	-0.36	0.38	0.14	0.26	0.43
Hit ratio	43.33%	58.33%	56.67%	53.33%	60.00%
			2000–2004		
Annualized geometric return	8.19%	3.55%	12.49%	2.59%	11.02%
Annualized standard deviation	11.04%	11.55%	11.23%	11.49%	10.89%
Return per unit of risk	0.74	0.31	1.11	0.23	1.01
Hit ratio	60.00%	58.33%	60.00%	60.00%	65.00%

, * *indicate statistical significance at the 5% and 1% levels.*

static case, note that each of the three approaches beats the other two in one of the three subperiods. In the case of the dynamic models, the SMB strategy cannot outperform any of the three AI approaches, except in the third subperiod, when it does better than RP and the GA.

The subperiod analysis clearly indicates that each AI method tends to encounter periods of very good performance and periods of weaker performance, and that success and failure do not happen simultaneously for the three methods. Such non-simultaneous periods of success and failure obviously translate into low correlations: −0.37 between the NN and the GA, −0.10 between RP and the GA, and 0.26 between RP and the NN. Not surprisingly, the three timing models yield the same decisions in only 49 of the 180 months (27.2%).

This non-simultaneity calls for an investigation into the power of consensus to predict which of large or small stocks will perform best in a given month. The idea is to tilt the capitalization favored by at least two of the three AI approaches each month (a consensus strategy following a two-tier majority rule).

The last columns in Exhibits 2 and 3 present the performance of the consensus strategy over the full period and the three subperiods. Over the full period, the Exhibit 2 static model consensus outperforms the SMB by more than 600 basis points annually. Interestingly, the consensus beats the SMB in every subperiod, by margins of 9.0, 10.0, and 0.7 percentage points in the first, second, and third subperiods, respectively. At the same time, the consensus is beaten by a single timing strategy in each subperiod. Overall, the return, the hit rate, and the return per unit of risk of the consensus are higher than those of any of the individual approaches. For the dynamic consensus in Exhibit 3, we see an increase in return overall of 83 basis points (8.60% annualized return versus 7.77%) and an impressive effect on the hit rate, which increases from 57.8% to 64.4%.

Formal Test of Market Timing Ability

While some of the timing models might have suffered from misfortune, others may have simply been gifted by luck. To distinguish luck from real timing ability, we apply the non-parametric market-timing test of Henriksson and Merton [1981].[7]

The last two lines in Panel A in Exhibits 2 and 3 report the p-stat and p-value (level of confidence) associated with this test. All the eight cases (AI approaches and consensus) have p-stat values significantly higher than 1 at a level of confidence of at least 10%. The NN and the consensus both reach the 1% level in their dynamic form.[8]

Practical Considerations

There are at least three practical issues to take into account: outliers, transaction costs, and benchmarks and implementation.

Impact of outliers: We should note that the 2000-2004 subperiod is unique, marked by exceptionally good performance of the classic small-minus-big strategy. A closer examination of this subperiod quickly shows that most unusual return events (positive or negative) happened in the year 2000. Indeed, we see in that year three monthly returns higher than 10% in absolute value. This is highly abnormal, given that such extreme monthly returns have occurred only five times between 1926 and 1999. We thus examine what happens with our timing models under such extreme conditions.

We are also interested in determining the behavior of the models under more normal conditions, so we reevaluate the performance of the AI approaches and the consensus and the SMB strategies in a normalized out-of-sample data set that excludes the year 2000. The results indicate all AI approaches have a return per unit of risk higher than the return of the SMB strategy. Most notably, the return of the GA strategy increases from a disappointing average annualized return of 0.2% to a more acceptable return of 2.9%. The average return of the consensus falls slightly, from 7.7% to 7.1%. For the dynamic models, the average increases from 8.6% to 9.3% when we exclude the year 2000.

Transaction cost considerations: One element in Exhibits 2 and 3 that we have not yet explained is the number of required switches. To calculate this metric, we count one "required switch" every time the portfolio switches from large-cap to small-cap, and vice versa.

If minimizing the number of switches and associated transaction costs is the primary objective, recursive partitioning (in its static form) is clearly the winner; it has only 32 switches in 180 months, compared to 55 and 62 for the NN and the GA (Panel A of Exhibit 2). In its dynamic form (Panel A of Exhibit 3), RP turns in the highest number of switches. The other two approaches require about the same number of switches in their static or dynamic forms.

Interestingly, the consensus requires 10 fewer switches in the dynamic version (46 versus 56). It thus has two main advantages: The majority rule produces stronger predictive signals and hence better performance and fewer switches for the dynamic strategies, which are the ones practitioners are most likely to follow.

We also calculate break-even transaction costs so that a method yields either the same net return as the classic SMB strategy or a null return. With the exception of the static GA, whose return is already lower than the return of the classic SMB strategy, all other five cases remain relatively profitable compared to the classic SMB. The consensus strategy, in its dynamic form, can handle transaction costs of up to 165 basis points (i.e., break-even transaction costs), while the RP strategy, in its dynamic form, can support transaction costs of up to 54 basis points.

We believe it is reasonable to assume that real transaction costs are far lower than these values.

Benchmark and implementation: All results so far use the SMB factor as the source of return. We decided to use the longest possible history of data in training our approaches, even though the factor could be hard to replicate in practice. Some sensitivity analyses confirm that our results are not sensitive to the size-premium benchmark used—(S&P 600–S&P 500), (Russell 2,000–Russell 1,000), or (Wilshire 1,750–Wilshire 750).

Some of these alternative benchmarks could be more manageable for implementation purposes, thanks to the availability of futures, swaps, or exchange-traded funds. For the Russell indexes in the dynamic case, the return per unit of risk (hit rate) becomes 0.19 (55.0%) for the RP, 0.66 (62.2%) for NN, 0.49 (56.7%) for the GA, and 0.9 (65.0%) for the consensus. This implementation provides some comfort with respect to the results obtained.[9]

CONCLUSION

U.S. equity managers who might consider an alpha-generating strategy using a small-size bet would earn, on average, a positive expected alpha in the long run. They could also experience long periods of underperformance. The classic small-minus-big strategy, which systematically favors small-caps, might well be too naive, and size timing, even if risky, can present an opportunity to add further value.

We show that strategies based on three artificial intelligence approaches—recursive partitioning, neural networks, and genetic algorithms—could successfully time the U.S. size premium over the 1990-2004 period. Of the six individual timing strategies examined—three artificial intelligence approaches conditioned on historical data (1975-1989) and on recent data (1975-month preceding the prediction)—five outperform the SMB premium.

None of the six timing strategies systematically outperforms the SMB strategy during the three five-year subperiods examined. Yet a strategy based on the majority rule (that is, a strategy favored by at least two of the three artificial intelligence approaches) outperformed the SMB strategy in each subperiod. Not only does the consensus strategy benefit from stronger predictive signals, but it also allows the number of bets (transaction costs) to be reduced.

Five of the six timing strategies, as well as the consensus strategies, remain profitable even after transaction costs.

Although all methods have their merits, recursive partitioning could be favored for its much greater transparency and ease of interpretation. In the case of NN and GA, we deal mostly with black boxes. For investors favoring results over understanding, the black box syndrome is not a serious issue, but when the model does fail (as each method does in one subperiod), the investor will find it quite complex to see what went wrong, given the opaque nature of the model.

In this presentation, we consider only extreme bets, 100% long in small-caps and 100% short in large-caps, and vice versa. This follows Fox [1999], who stresses that managers with superior forecasting skills (60% hit rate or higher) should favor more extreme tilts because they improve the entire range of possible returns. Still, we can reasonably conceive of less extreme strategies that would allow for neutral allocation when choices are less clear-cut. Therefore, considering three states of the world—small-cap tilt, large-cap tilt, and no tilt—could be more interesting, as it could yield stronger predictive signals and fewer switches (lower transaction costs).

ENDNOTES

The views expressed in this article are those of the authors and do not necessarily reflect the position of the Caisse de Dépôt et Placement du Québec.

[1]Suppose that after all observations are allocated down the tree, 20 observations end up in a given node. If 9 of these observations are associated with binary case A, and 11 are associated with binary case B, then the level of purity at this node is 0.55. If a further split is possible so that 8 of the 9 A cases go to the right along with two of the B cases, the resulting right and left descendant nodes will have purity levels of 0.8 and 0.9. Such an additional split would obviously be welcomed.

[2]The elite count refers to the number of individuals that are directly cloned in the next generation. A cross-over fraction of 0.8 implies that 80% of the non-cloned individuals in the next generation are obtained via reproduction rather than mutation.

[3]Liew and Vassalou [2000] go the other way and demonstrate that the return on the SMB portfolio conveys information on future GDP growth and that it could be useful in predicting economic cycles.

[4]Coggin [1998] suggests timing models on macroeconomic factors to forecast style index returns.

[5]The full tree is available upon request.

[6]The GA equation is: If [$0.1977\ TERM_{t-1} - 0.4317\ CREDIT_{t-1} - 0.5819\ TBILL_{t-1} - 1.2883\ INFL_{t-1} + 0.7378\ COIN_{t-1} + 0.4094\ LEAD_{t-0} + 0.3359\ MOM_{t-1} + 0.1166\ EARN_{t-1} + 0.2360\ DIV_{t-1} + 0.4717\ GSCI_{t-1} + 0.1183\ CAP_{t-1} - 0.6987\ IND_{t-1} - 0.4827\ CSI_{t-1} + 1.8991\ ISM_{t-1} + 0.0332\ SAV_{t-1} + 0.1905\ M2_{t-1} - 0.8035\ CEXP_{t-1} + 1.7319\ PPI_{t-1} + 0.9418\ NYSE_{t-1} + 0.6027\ TRAD_{t-1}] > 0.0171$, the model favors small-cap. Otherwise, the model favors large-cap.

[7]Like the hit rate, the non-parametric market-timing test of Henriksson-Merton [1981] focuses on the number of correct predictions rather than on the level of the returns associated with each decision. The output of this test, the p-statistic, requires first the calculation of the number of correctly predicted positive months and the number of correctly predicted negative months. A significant p-stat (higher than 1) indicates that the model has genuine predictive ability.

$$p - stat = \frac{n_1}{N_1} + \frac{n_2}{N_2}$$

where

n_1: number of correctly predicted months SMB is positive;
N_1: number of months SMB is positive;
n_2: number of correctly predicted months SMB is negative; and
N_2: number of months SMB is negative.

The p-value is calculated according to Park and Switzer [1996]:

$$p - value = \sum_{n_1}^{\min(N_1, n)} \left[\binom{N_1}{x} \binom{N_2}{n-x} \Big/ \binom{N}{n} \right]_3$$

where $N = N_1 + N_2$ and $n = n_1 + n_2$.

[8]We also use the parametric market-timing test of Henriksson-Merton [1981]. The results are similar to results for the non-parametric test (available on request).

[9]More detailed results for the static and dynamic cases using Russell indexes are available from the authors.

REFERENCES

Ahmed, P., L.J. Lockwood, and S. Nanda. "Multistyle Rotation Strategies." *The Journal of Portfolio Management*, 28 (Spring 2002), pp. 17-29.

Amenc, N., P. Malaise, L. Martellini, and D. Sfeir. "Tactical Style Allocation—A New Form of Market Neutral Strategy." *Journal of Alternative Investments*, Summer 2003, pp. 8-22.

Bauer, R.J., Jr. *Genetic Algorithms and Investment Strategies*. New York: John Wiley & Sons, 1994.

Black, F. "Estimating Expected Return." *Financial Analysts Journal*, January/February 1995, pp. 168-171.

Breiman, L., J.H. Friedman, R.A. Olsen, and C.J. Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.

Campbell, J.Y., A.W. Lo, and A. C. MacKinlay. *The Econometrics of Financial Markets*. Princeton: Princeton University Press, 1997.

Coggin, T.D. "Long-Term Memory in Equity Style Indexes." *The Journal of Portfolio Management*, Winter 1998, pp. 37-46.

Cooper, M., H. Gulen, and M. Vassalou. "Investing in Size and Book-to-Market Portfolios Using Information about the Macroeconomy: Some New Trading Rules." Working paper, Graduate School of Business, Columbia University, 2001.

Fama, E., and K.R. French. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33 (1993), pp. 3-56.

————. "The Cross-Section of Expected Stock Returns." *Journal of Finance*, 47 (1992), pp. 427-465.

————. "Multifactor Explanations of Asset Pricing Anomalies." *Journal of Finance*, 51 (1996), pp. 55-84.

Fox, S. "Assessing TAA Manager Performance." *The Journal of Portfolio Management*, 26 (Fall 1999), pp. 40-49.

Haykin, S. *Neural Networks*. New York: Macmillan, 1994.

Henriksson, R.D., and R.C. Merton. "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecast Skills." *Journal of Business*, 54 (1981), pp. 513-517.

Jensen, G.R., R.R. Johnson, and J.M. Mercer. "The Inconsistency of Small-Firm and Value Stock Premiums." *The Journal of Portfolio Management*, 24 (Winter 1998), pp. 27-35.

Kao, D.L., and R.D. Shumaker. "Equity Style Timing." *Financial Analysts Journal*, January/February 1999, pp. 37-48.

Kester, G.W. "Market Timing with Small vs. Large Firm Stocks: Potential Gains and Required Predictive Ability." *Financial Analysts Journal*, September/October 1990, pp. 63-69.

Kingdon, J., and K. Feldman. "Genetic Algorithm and Applications to Finance." *Applied Mathematical Finance*, 2 (1995a), pp. 89-116.

————. "Neural Networks and Some Applications to Finance." *Applied Mathematical Finance*, 2 (1995b), pp. 17-42.

Kryzanowski, L., M. Galler, and D.W. Wright. "Using Artificial Neural Networks to Pick Stocks." *Financial Analysts Journal*, July/August 1993, pp. 21-27.

Leinweber, D. "The Perils and Promise of Evolutionary Computation on Wall Street." *The Journal of Investing*, 12 (Fall 2003), pp. 21-28.

Levis, M., and M. Liodakis. "The Profitability of Style Rotation Strategies in the United Kingdom." *The Journal of Portfolio Management*, 26 (Fall 1999), pp. 73-86.

Liew, J., and M. Vassalou. "Can Book-to-Market, Size and Momentum be Risk Factors that Predict Economic Growth?" *Journal of Financial Economics,* 57 (2000), pp. 221-245.

Medsker, L., R.R. Trippi, and E. Turban. *Neural Networks in Finance and Investing*. Chicago: Probus Publishing Company, 1993.

Olden, J.D., M.K. Joy, and R.G. Death. "An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks Using Simulated Data." *Ecological Modelling*, 178 (2004), pp. 389-397.

Park, T.H., and L.N. Switzer. "Mean Reversion of Interest-Rate Term Premiums and Profits from Trading Strategies with Treasury Futures Spreads." *Journal of Futures Markets*, 16(3) (1996), pp. 331-352.

Reinganum, M.R. "The Significance of Market Capitalization in Portfolio Management over Time." *The Journal of Portfolio Management*, 25 (Summer 1999), pp. 39-50.

Sorensen, E.H., K.L. Miller, and C.K. Ooi. "The Decision Tree Approach to Stock Selection." *The Journal of Portfolio Management*, 27 (Fall 2000), pp. 42-52.

White, H. "A Reality Check for Data Snooping." *Econometrica*, 68 (September 2000), pp. 1097-1126.