Testing The Adaptive Efficiency Of U.S. Stock Markets: A Genetic Programming Approach

Stan Miles, Thompson Rivers University, Canada
Barry Smith, York University, Canada

ABSTRACT

Genetic programming is employed to develop trading rules, which are applied to test the efficient market hypothesis. Most previous tests of the efficient market hypothesis were limited to trading rules that returned simple buy-sell signals. The broader approach taken here, developed under a framework consistent with the standard portfolio model, allows use of trading rules that are defined as the proportion of an investor's total wealth invested into the risky asset (rather than being a simple buy-sell signal). The methodology uses average utility of terminal wealth as the fitness function, as a means of adjusting returns for risk. With data on daily stock prices from 1985 to 2005, the algorithm finds trading rules for 24 individual stocks. These rules then are applied to out-of-sample data to test adaptive efficiency of these markets. Applying more stringent thresholds to choose the trading rules to be applied out-of-sample (an extension of previous research) improves out-of-sample fitness; however, the rules still do not outperform the simple buy-and-hold strategy. These findings therefore imply that the 24 stock markets studied were adaptively efficient during the period under study.

Keywords: adaptive efficiency, trading rules, genetic programming

1. INTRODUCTION

When markets are efficient, investors cannot make profits by exploiting publicly available information. Daniel and Titman (1999) introduced a weaker concept of market efficiency called adaptive efficiency: A market is characterized by adaptive efficiency if profit opportunities disappear when they become obvious. The objective of this paper is to test adaptive efficiency of stock markets by conducting a broadly representative study (using data from 24 stocks across a wide spectrum of industries) of the efficacy of trading rules evolved using genetic programming methodology (a machine learning technique inspired by biological evolution). If the rules evolved by genetic programming using in-sample data have low fitness when applied to new (out-of-sample) data, this is interpreted as evidence of adaptive efficiency.

The efficient market hypothesis (EMH) was first introduced in the mid-1960s, and debate about the validity of EMH continues today. This proposition is of interest to everyone concerned with the workings of stock and commodities markets, from policy makers and regulators to investment professionals and small investors. EMH is of interest to both investment professionals and small investors because if EMH is true and the capital asset pricing model therefore does correctly predict returns on securities, then the only rational investment strategy is diversification. In order to choose a portfolio, each investor would decide on the level of risk he or she is comfortable with accepting and select an efficient portfolio that maximizes return given this level of risk.

A test of the validity of EMH involves more than figuring out whether financial markets can be beaten—whether a trading strategy can consistently generate a better rate of return than that seen in the market as a whole. One also is testing whether the financial markets are rational and whether financial markets result in prices that are "correct" in the sense that these prices reflect all available information. When that is the case, these prices provide

correct signals to economic agents for resource allocation. The more evidence that policy makers and regulators have regarding the validity of EMH, the easier it will be for them to determine the role that should be given to financial markets. In particular, this information will help to choose the optimal level of financial regulation.

EMH has been widely accepted as true by U.S. policy makers since the 1970's, and belief in the validity of EMH motivated the hands-off approach to financial regulation that began in that era. EMH made an impact on the prevailing doctrines and regulations. If EMH is true, financial markets regulate global affairs and allocate an economy's capital with a speed and decisiveness unmatched by individuals, firms, and governments. The belief in EMH implies policy responses that largely deregulate markets, since markets already operate at optimal efficiency, but favor regulations that require firms to publicly report to bodies such as the Securities and Exchange Commission (SEC), as information is the grease that makes the market work at its best.

This study extends the results of recent work that tested EMH, which focused on whether market participants can use historical data as input to identify trading rules that consistently produce abnormally high out-of-sample risk-adjusted returns (indicating that the markets are not efficient). Many of the previous studies were limited to trading rules that output simple buy–sell signals. These trading rules are "bang–bang" strategies, that is, strategies that alternate between investing all of one's wealth in a single risky asset and investing all of it in a single riskless asset. Past research has revealed that bang–bang strategies are dominated by strategies that can diversify between risky and riskless assets at every point in time (rather than only across time), implying that studies using bang–bang strategies are biased toward accepting market efficiency. This insight motivates a broader approach that allows the study of trading strategies developed under a framework consistent with the standard portfolio model, with a trading strategy defined as the proportion of an investor's total wealth allocated to the risky asset (that is, a strategy is a proportion rather than a simple "0–1" buy–sell signal).

Most previous studies that tested the efficiency of financial markets examined the performance of technical trading rules and adjusted the trading rule returns for risk. The distinctive feature of the fitness criterion used in our research is that it adjusts returns for risk in a manner consistent with the standard portfolio model, in that the criterion is based on the investor's expected utility and incorporates risk aversion.

Another contribution of this study is the use of complex thresholds to select the rules to be tested out-of-sample. Most commonly, previous studies used a simple procedure of applying, in the out-of-sample period, the rule with the highest fitness (during what is called the selection period), as long as that rule outperformed a certain threshold, namely, the buy-and-hold rule. It can be argued, however, that even a moderate (and reasonable) level of risk aversion would lead most investors to reject a complicated trading strategy with potentially large risks that barely outperforms the buy-and-hold rule in-sample. We allow for risk aversion by introducing additional money management criteria as thresholds for rule selection.

The study is thorough, examining the performance of trading rules found by the genetic programming algorithm in 21 out-of-sample periods for each of 24 stocks (for a total of 504 out-of-sample periods). The present findings show that using more stringent thresholds helps improve the out-of-sample fitness of the rules that are saved. In general, the trading rules that the current methodology generates do not outperform the simple buy-and-hold rule. Thus, our comprehensive study leads to the conclusion that the 24 stock markets examined were adaptively efficient between 1985 and 2005.

The remainder of the paper is organized as follows. A brief review of the relevant literature is presented in the next section. Section 3 describes genetic programming, the means by which trading rules are generated. The methodology employed in the paper is discussed in Section 4. The data set used in our study is presented in Section 5. The empirical results and their implications for market efficiency of 24 U.S. stock markets are detailed in Section 6. Finally, Section 7 provides some concluding remarks.

2. BACKGROUND

The most widely used theoretical framework for analyzing financial markets is based on the efficient market hypothesis (EMH). Fama (1991) defined an efficient market as one in which economic agents, when placing

values on financial securities, take all relevant information into account. Among the theories and models proposed as alternatives to EMH, it is disputable which corresponds most closely to reality. This dispute is significant because, as Fama (1976) noted, all empirical tests of EMH involve a joint hypothesis of market efficiency (rationality) and a particular equilibrium model. Empirical rejections of market efficiency may be due either to an anomaly (failure of rationality of economic agents) or to a misspecification of the equilibrium model (Fama, 1991).

Technical analysis, defined as the use of past prices as inputs to trading strategies, has a long history among investors but only relatively recently has found support within the academic community. Fama (1976) pointed out that if EMH is true, prices already reflect all public information, and one cannot expect to increase one's utility by switching back and forth between risky and riskless assets. According to EMH, technical trading rules that are inexpensive to implement should not yield excess profits if markets are efficient.

Technical analysis nevertheless is prevalent in many financial markets. Allen and Taylor (1990) stated that more than 90% of foreign exchange dealers in the London market reported using technical analysis to help them make forecasts. Gehrig and Menkhoff (2006) determined that 90% of the German and Austrian foreign exchange dealers and international fund managers they studied assigned a relative importance of 20% or more to technical analysis. Menkhoff and Schmidt (2005) conducted a survey of German equity fund managers and found that more than 90% of their respondents used momentum strategies to some extent, with 11% of respondents preferring these strategies to the buy-and-hold strategy and contrarian strategies. Menkhoff and Schmidt interpreted their findings as evidence of the absence of market efficiency.

Many researchers test EMH by assessing the returns that would be realized by traders who use various trading rules. These studies include fundamental analysis studies that consider trading rules that make use of macroeconomic variables as well as industry-specific and company-specific variables (Al-Debie and Walker, 1999; Lev and Thiagarajan, 1993). Other studies examined technical analysis trading strategies that make use of past prices (Brock et al., 1992; Gençay, 1999; Kwon and Kish, 2002; Skouras, 2001; Taylor, 2000). Some researchers used nonlinear methods such as artificial neural networks, pattern recognition algorithms, and fuzzy logic to identify patterns in prices and develop technical trading rules that exploit these patterns (Fernández-Rodríguez et al., 2000; Lo et al., 2000; Zhou and Dong, 2004). For a comprehensive review, see Park and Irwin (2007). In a survey of technical analysis studies, Park and Irwin stated that only 24 of those that had addressed the performance of technical trading rules found results consistent with EMH, compared with 58 studies that reported results that were inconsistent with EMH.

Daniel and Titman (1999) defined a new notion of market efficiency called adaptive efficiency. They argued that it takes time for rational arbitrageurs to gain knowledge of the strategies and the degree of irrationality of other traders. This knowledge is required to remove price patterns that stem from trading on the part of irrational market participants. This implies that risk-averse arbitrageurs with limited capital find it difficult to instantaneously remove these price patterns, contrary to the very premise of EMH. Adaptive efficiency is a weaker notion of market efficiency than EMH; it allows profit opportunities to appear in historical data but requires that they dissipate as soon as they become apparent. Daniel and Titman rejected this weaker form of market efficiency in an empirical study of U.S. equity markets.

2.1 EMH Tests and Bang–Bang Strategies

Bang–bang strategies are essentially an extreme form of market timing. Investors practicing these strategies invest their entire endowment either in a risky asset or in a riskless asset. That is, at any given time there is no diversification among assets; however, switching between assets means that there is diversification across time. As implemented, technical analysis is best suited to solving the problem of forecasting stock market returns. One fundamental difference between portfolio choice modeling and technical analysis is that the 0–1 signals based on technical analysis usually are interpreted as a bang–bang strategy, whereas portfolio choice models output a fraction of wealth to be invested in each asset, most often calling for diversification at any given point in time.

Samuelson (1997) proved that the expected utility of the investor who employs the constant diversification trading rule is necessarily higher than the expected utility of the investor who uses the bang–bang strategy. Gollier

(1997) demonstrated that the bang–bang strategy consisting of investing in one asset in the first period and in another asset in the second period is second-order stochastically dominated by the strategy of splitting one's wealth evenly between the two assets during the two periods. These two results illustrate that strategies that are allowed to diversify between the risky and the riskless assets at every point in time are likely to outperform bang–bang strategies. Most studies that test EMH by evaluating the performance of technical analysis trading rules nevertheless limit their scope to bang–bang strategies. The implication of the research performed by Samuelson (1997) and Gollier (1997) is that those studies are biased toward accepting EMH. Thus, a proper test of EMH would involve examining the performance of technical trading rules that are allowed to diversify between the risky and the riskless assets at every point in time.

2.2 EMH Tests and Genetic Programming

Genetic programming (GP) is an artificial intelligence technique that mimics the processes observed in natural evolution (i.e., survival of the fittest) to search for candidate solutions to problems. It has been applied to a diverse array of problems in econometrics, economics, and finance, as well as to problems in other fields that are beyond the scope of this paper. Kaboudan (1999) used GP to find the underlying data-generating process behind the time series data for stock prices and to measure time series' predictability. Álvarez-Díaz and Álvarez (2005) combined forecasts generated by GP and neural networks to forecast exchange rates. Wagner and Brauer (2007) employed a dynamic forecasting version of GP to forecast U.S. GDP. Jin et al. (2009) developed a constraint-guided method with GP and applied it to problems in bargaining and financial prediction. Lien et al. (2003) used GP to study lead–lag relationships of the structural changes in spot and futures markets. Studies by Lensberg (1999) and Östermark (1999) explored the usefulness of GP for solving highly irregular optimization problems and for generating hypotheses about rational behavior in situations where explicit maximization is not well defined. Álvarez-Díaz and Miguez (2008) used GP to investigate the functional relation between the quality of institutions and a set of historical, economic, geographic, religious, and social variables. Chen et al. (2008) proposed a dynamic proportion portfolio insurance strategy that introduced a risk multiplier (that is allowed to change according to market conditions) into the popular constant proportion portfolio insurance strategy. Chen et al. made use of GP to build the equation tree for the risk multiplier in their model. Lensberg and Schenk-Hoppé (2007) generalized an evolutionary finance model by using GP to maintain the diversity of investment strategies. McKee and Lensberg (2002) used GP together with rough sets theory to construct an efficient and effective bankruptcy prediction model.

One interesting application of GP is the automated discovery of financial asset trading strategies, with these strategies then used to test weak-form financial market efficiency. Sullivan et al. (1999) argued that studies that use technical analysis to test EMH are subject to so-called data-snooping bias. This bias arises because these studies use historical data to evaluate the performance of technical trading rules that are popular in practice. It seems likely, however, that the rules became popular in the first place because of their good ex post performance. That implies that these technical trading rules are expected a priori to perform well during some time periods, and therefore studying the performance of these rules using historical data is not an appropriate way to test EMH, since such tests will be biased toward rejecting EMH. Allen and Karjalainen (1999) and Neely et al. (1997) suggested that GP is not subject to data-snooping bias because GP uses simple arithmetic and logical operators as building blocks to construct its own new trading rules with a high fitness in-sample, then evaluates the fitness of these rules using a different, out-of-sample set of historical data.

Applications of GP in the S&P 500 market can be found in Allen and Karjalainen (1999), Fyfe et al. (2005), Neely (2003), Ready (2002), and Wang (2000). Allen and Karjalainen performed one of the first studies of out-of-sample returns resulting from ex ante optimal trading rules evolved by GP. They concluded that the rules generated in their study did not outperform the simple buy-and-hold strategy after transaction costs were figured in. Fyfe et al. and Neely extended the experiments presented by Allen and Karjalainen by using a fitness criterion that adjusts trading rule returns for risk to evolve ex ante optimal trading rules. Both studies, as well as a further extension of Allen and Karjalainen's work by Ready, found that when the out-of-sample trading rule returns are adjusted for risk, the rules cannot beat the buy-and-hold strategy.

A number of studies have employed GP to study the properties of prices in individual stock markets. Kaboudan (2000) used GP to produce one-day-ahead stock price forecasts of six individual stocks, and then

evaluated trading strategies based on these forecasts and concluded that these strategies yield a relatively high return on investment. Fyfe et al. (1999) employed GP to evolve trading rules for one UK stock and found that risk-adjusted returns were inferior to those based on the buy-and-hold rule. Potvin et al. (2004) applied GP to evolve trading strategies for 14 Canadian companies. Although Potvin et al. found the rules they had evolved by GP to be valuable when the market fell or when it was stable, these rules were dominated by the buy-and-hold approach during times when the market was rising.

Researchers have also applied GP to futures markets, with similar results. Roberts (2005) found rules that can be characterized as profitable out-of-sample for only 2 of 24 futures markets studied. Wang (2000) applied GP methods to examine the trading and hedging effectiveness of S&P 500 spot and futures markets, finding these markets to be efficient. In the spot market, GP rules duplicated the buy-and-hold rule. The rules constructed by GP did not have consistent out-of-sample performance on a risk-adjusted basis in the futures market. Wang also reported that more than 40% of the trading rules generated by GP had significant market-timing ability.

Another application of GP is using it to evolve trading strategies in foreign exchange markets. Neely and Weller (1999) and Neely et al. (1997) found that GP evolved trading rules with significant out-of-sample excess returns. Both Dempster and Jones (2001) and Neely and Weller (2003) applied GP to intra-daily data in foreign exchange markets. The former study found profitable rules even when realistic transaction costs were taken into account; the latter, in contrast, showed results that were consistent with market efficiency when reasonable transaction costs and trading hours were used.

Most studies that test EMH by evaluating the returns based on existing technical analysis strategies employed rules that output 0–1 signals and can be interpreted as bang–bang strategies. Studies such as those by Dempster and Jones (2001), Fyfe et al. (2005), and Potvin et al. (2004) used GP to search within the space of trading rules that output 0–1 signals and can be interpreted as bang–bang strategies. In contrast to that, some of the studies described in the preceding paragraphs arrived at their conclusions regarding market efficiency by studying the out-of-sample performance of portfolios formed by using GP trading rules, in addition to studying the out-of-sample performance of GP rules themselves. Allen and Karjalainen (1999), Neely et al. (1997), Neely and Weller (1999, 2003), Neely (2003), and Roberts (2005) all took the approach of using GP to evolve and save a number of simple bang–bang trading rules, and then evaluating the out-of-sample returns on portfolios formed by using either the signals output by the saved rules or the characteristics of the in-sample returns based on the saved rules.

Allen and Karjalainen (1999) evolved and saved 100 rules for the S&P 500 index, using daily data. Neely and Weller (2003) evolved and saved 25 rules for foreign exchange markets, using intra-day data. Both studies examined the out-of-sample returns based on the equally weighted portfolio rule that assigned equal weights in the portfolio to all the rules that had satisfied their selection criteria. After using GP to generate 100 trading rules for the foreign exchange market, Neely et al. (1997) and Neely and Weller (1999) examined returns based on the equally weighted portfolio rule (which they referred to as the uniform rule) during a testing (out-of-sample) period and compared those returns to those obtained from the median portfolio rule. For the median portfolio rule, a long position was taken if more than half of the GP trading rules output a buy signal, and a short position was entered into otherwise.

Neely (2003) used GP to evolve, during each of 10 training (in-sample) periods, 10 trading rules for the S&P 500 index, then examined the returns resulting from five different portfolios in each of the testing periods: the equally weighted portfolio rule, the median portfolio rule, and three other portfolios. The latter three portfolios were constructed using the following rules, respectively: Arbitrarily split the portfolio equally between the buy-and-hold strategy and the trading rule, select the portfolio weights to maximize the in-sample Sharpe ratio (Sharpe, 1966), and select the portfolio weights to maximize the out-of-sample Sharpe ratio. The last of these portfolios was used as a benchmark to gauge an upper bound on the fitness improvement due to using the technical rule constructed by GP.

Roberts (2005) first used GP to generate trading rules that output ternary signals (i.e., long, neutral, or short position). Roberts then compared the out-of-sample returns on three equity indices (as benchmarks) to the out-of-sample returns on a futures portfolio in which 30% of the assets were devoted to initial margin (this amount being split equally among the 24 commodities being traded in accordance with GP's trading rules) and the remaining

assets were held in U.S. T-bills.

Wang's (2000) use of GP to generate trading and hedging rules in S&P 500 spot and futures markets came closest to our approach of evolving the fraction of wealth to be allocated to the risky asset. In Wang's study, GP futures trading rules were limited to five trading signals—two contracts held long, one contract held long, no position, one contract held short, and two contracts held short.

Taking an indirect approach to obtaining portfolio rules, such as using GP to evolve rules that output one of five trading signals (or using GP to generate and save simple bang–bang rules) and then forming a portfolio using the signals derived from the saved rules, may not be the most efficient way to find a portfolio strategy with high out-of-sample fitness. Samuelson's (1997) results indicate that out-of-sample performance of rules could be improved by allowing rules that return a proportion of wealth (not restricted to 0% or 100%, as is the case in bang–bang strategies) to be invested into the risky asset. We employ that suggestion in our experiments. Thus, our study complements the studies described above.

3. GENETIC PROGRAMMING

Genetic programming, introduced by Koza (1992) as a modification of genetic algorithms, is a nonlinear procedure for searching for and refining candidate solutions to problems. This methodology is designed for problems in which the search space of possible solutions consists of entities—such as computer programs or analytical expressions—that can be expressed as decision trees. The main features of GP are its flexible representation of solutions and its use of operators inspired by the theory of natural selection to generate new candidate solutions. GP comprises five components: population, evaluation, selection, crossover, and mutation.

Initially, a population of random candidate solutions (the first generation) is produced. The only requirements for solutions are that they be well defined and produce output appropriate to the problem of interest. Most of these random solutions will do quite poorly in meeting the criteria set in the problem, but some, purely by chance, will be better than the rest. The population is allowed to "evolve" over a series of generations. Each generation is created by "mating" among the parent generation, using crossover and mutation operators, to create the subsequent generation of "children." Crossover mixes subtrees of the population, whereas mutation replaces subtrees with new, randomly generated subtrees. GP selects the parents that will "mate" through a randomization process with weighting by fitness; individuals with high fitness in solving the problem are more likely to be chosen than individuals with low fitness. GP runs until either a solution is found or a fixed number of generations (predetermined by the GP user) have been created. In this way, the genetic program searches promising areas of the solution space by evolving a population of decision trees, with the decision trees in each successive generation tending to become more adept at solving the problem.

GP is a robust optimization technique that can find good solutions to very complex problems. It cannot compete computationally with classical algorithms in the classical algorithm domain (i.e., linear programming problems). GP's advantage lies in the domain of problems that cannot be solved easily, or at all, using classical techniques.

4. METHODOLOGY

GP methodology is employed here to generate portfolio rules for each of 24 stocks and then study the out-of-sample performance of these rules. GP is used to evolve portfolio rules that determine a fraction of wealth to be allocated to a risky asset (i.e., one of the 24 stocks in this study), with the remaining wealth being invested into a riskless asset, as opposed to simple binary (0–1) bang–bang rules evolved in earlier studies (e.g., Dempster and Jones, 2001; Fyfe et al., 2005; Potvin et al., 2004). Allen and Karjalainen (1999) and Neely et al. (1997) suggested, as a means to avoid results being subject to the data-snooping bias, testing market efficiency by assessing the performance of rules generated by GP. We follow that suggestion here. Hence, our setup allows us to test adaptive efficiency of a financial market while avoiding both the data-snooping bias and the bias that plagues studies that analyze the out-of-sample performance of simple binary (bang–bang) trading strategies.

4.1 Genetic Programming Setup

The building blocks used in all of the experiments in the present study consist of numerical constants, arithmetic and logical operators, and simple functions of past price data. These building blocks are listed in Table 1. To guarantee that a trading rule is well defined, the root node of each GP decision tree (the node at the top of the tree, which is the location of the final output of the rule) must be constrained to be a probability function (a function that outputs a number in the range 0 to 1 inclusive). To this end, all root nodes are constrained to be one of the functions in the set {pconstant, And, Or, Not, If-Then, If-Then-Else, AND, OR, NOT, IF-THEN, IF-THEN-ELSE, GT, LT, SRatio} (see Table 1). We interpret the number output by each GP candidate solution decision tree as the fraction of wealth that the investor allocates to buying stock shares.

Table 1
Genetic Programming Building Blocks

Building Blocks	Input Data Type	Input	Output Data Type	Output
pconstant		No input	Probability	Real number between 0 and 1
P_t		No input	Variable	Current value of asset price
R_t		No input	Variable	Current value of riskless rate
W_t		No input	Variable	Current value of investor's wealth
days remaining		No input	Variable	Number of days remaining until end of subperiod
and	Boolean	*a*, *b*	Boolean	If *a* is true and *b* is true, output true; else, output false
or	Boolean	*a*, *b*	Boolean	If *a* is true or *b* is true, output true; else, output false
not	Boolean	*a*	Boolean	If *a* is true, output false; else output true
if-then	Boolean	*a*, *b*	Boolean	If *a* is true, output *b*; else, output false
if-then-else	Boolean	*a*, *b*, *c*	Boolean	If *a* is true, output *b*; else, output *c*
And	Boolean	*a*, *b*	Probability	If *a* is true and *b* is true, output 1; else, output 0
Or	Boolean	a, b	Probability	If *a* is true or *b* is true, output 1; else output 0
Not	Boolean	*a*	Probability	If *a* is true, output 1; else, output 0
If-Then	Boolean (*a*),Probability (*b*)	*a*, *b*	Probability	If *a* is true, output *b*; else, output 0
If-Then-Else	Boolean (*a*),Probability (*b*, *c*)	*a*, *b*, *c*	Probability	If *a* is true, output *b*; else, output *c*
AND	Probability	*a*, *b*	Probability	$a \times b$
OR	Probability	*a*, *b*	Probability	$(a + b) - (a \times b)$
NOT	Probability	*a*	Probability	$1 - a$
IF-THEN	Probability	*a*, *b*	Probability	If $a = 1$, output *b*; else, output 0
IF-THEN-ELSE	Probability	*a*, *b*, *c*	Probability	If $a = 1$, output *b*; else, output *c*
<	Real	*a*, *b*	Boolean	If $a < b$, output true; else, output false
>	Real	*a*, *b*	Boolean	If $a > b$, output true; else, output false
GT	Real	*a*, *b*	Probability	If $a > b$, output 1; else, output 0
LT	Real	*a*, *b*	Probability	If $a < b$, output 1; else, output 0
SRatio	Real	*a*, *b*	Probability	If b = 0, or if a and b are of opposite sign, output 0. Otherwise, output 1 if $\|a\| \geq \|b\|$; else, output a/b
+	Real	*a*, *b*	Real	$a + b$
–	Real	*a*, *b*	Real	$a - b$
/	Real	*a*, *b*	Real	If $\|b\| > 0$, output a/b; else, output 1
*	Real	*a*, *b*	Real	$a \times b$
absolute value	Real	*a*	Real	$\|a\|$
ln	Real	*a*	Real	If $a > 0$, output $\ln(a)$, else output 0
power	Real	*a*, *b*	Real	a^b
maximum	Real	*a*, *b*	Real	$\max(a, b)$
minimum	Real	*a*, *b*	Real	$\min(a, b)$
lag	Variable (*a*), Integer (*n*)	*a*, *n*	Real	Value of the variable *a*, *n* days ago
moving average	Variable (*a*), Integer (*n*)	*a*, *n*	Real	Average of the last *n* observations of variable *a*
return	Variable	P_t, P_{t-1}	Real	$\ln(P_t/P_{t-1})$

At the bottom of the tree are elements of the terminal set (the set of inputs to the rule). They consist of a numerical constant (pconstant) and the variables that represent the stock price, the stock return, the riskless interest

rate, the investor's wealth, and the number of days remaining in the investor's trading horizon (the number of days until the date on which trading ends and the utility of wealth is evaluated). The function set contains real-valued functions, Boolean functions, and probability functions. Not all logical functions return a Boolean output (true or false). Functions in the set {and, or, not, if–then, if–then–else} have Boolean inputs and outputs, while functions in the set {And, Or, Not, If–Then, If–Then–Else} convert Boolean inputs into a probability output of 1 or 0, and functions in the set {AND, OR, NOT, IF–THEN, IF–THEN–ELSE} take probability inputs and compute probability outputs. Lastly, the functions > and < convert real numbers into Boolean values, and the functions GT and LT convert real numbers into probabilities.

The real-valued functions in the function set include the arithmetic and mathematical operators +, –, /, *, absolute value, ln, and power. Also included are the functions maximum and minimum, a lag function that returns the value of the variable argument as it was *n* days ago (*n* is the second, integer-valued, argument), and a function ("moving average") that returns a moving average of the variable argument in a window defined by the second, integer-valued, argument.

The output type of every function below the top of the tree matches the input type of the function above it in the tree. The function and terminal sets enable GP to search for trading strategies in the space of complex and nonlinear decision trees. This setup gives GP the potential to identify factors that are important for successful trading strategies, as well as to combine these factors in ways that form decision rules that correspond to profitable trading strategies.

The Neely et al. (1997) and Neely and Weller (1999, 2003) studies all had one training, one selection, and one testing period. Potvin et al. (2004) evaluated performance of trading rules evolved using two sets of training periods—a short training period and a long training period. The rules found using both training periods were then applied to the same testing period. In the present study, we adopt an approach similar to that of Allen and Karjalainen (1999): To prevent possible "data snooping" in the choice of time periods, they used 10 sets of successive training, selection, and testing periods, whereas we used 21 such sets. To ensure that our results are not the artifact of the particular choice of the in-sample and out-of-sample periods (that is, are not the results of data snooping), we conduct experiments for multiple stock markets and also examine multiple training, selection, and testing periods for each market.

In our experiments, we employ a GP algorithm, with data from a given 5-year in-sample period as input, to evolve and select trading rules that are then applied (tested) in the following (sixth) year. The first half of each 5-year in-sample period is allotted to training, and the second half to selection. Associated with each calendar year is a set of four subperiods that are each split evenly between an observation phase (when the GP methodology examines data) and a trading phase (when trades are executed in the simulations). The trading phase of one subperiod coincides with the observation phase of the next subperiod. Ten subperiods are associated with every training or selection period (since each of those periods is comprised of 2.5 calendar years), while only four subperiods are associated with every testing period (since a testing period encompasses just one calendar year).

We chose the length of the trading phase of each subperiod to be equal to the investor's trading horizon. In our stock experiments, we assume the investor's trading horizon is 60 trading days (approximately 3 months). Thus the total length of each subperiod is 120 trading days. On the first day of the trading phase of any (training, selection, or testing) subperiod, GP has access to all of the stock prices during the corresponding observation phase. On each remaining day of the trading phase, GP can continue to use all of that stock price information and, in addition, all stock price information accumulated thus far during the trading phase, up to and including the data from the previous day.

We use a criterion of fitness that involves computing the utility of terminal wealth at the end of each subperiod within a given (training, selection, or testing) period and then averaging those values (see Section 4.2 for details). We note that a longer trading horizon (i.e., longer subperiods) would result in fewer terms being averaged, and thus a less meaningful average; a shorter trading horizon seems unrealistically shortsighted, though we recognize that some traders (e.g., day traders) have extremely short horizons.

For each of the 24 stocks used in this study, the data set spans the years 1980–2005—plus the last quarter of 1979 (the observation phase of the first subperiod associated with 1980), which is hereinafter implicitly included in reference to the data for 1980. Thus after allowing for the first 5-year time interval during which trading rules are evolved (1980–1984), we have 21 testing periods (namely, calendar years 1985 through 2005) and 84 testing subperiods (as there are 4 subperiods associated with each testing period).

For example, an investor at the beginning of 1985 has access to 5 years of historical data (1980–1984). The investor assigns the first half of this data set to the training period and the rest to the selection period. Ten GP trials (to be discussed later) are run using the periods specified in this way, and under certain conditions (if the threshold criteria defined in Section 4.3 are satisfied for any of the rules generated in these 10 trials) one rule is selected to be applied out-of-sample. The investor then uses this rule to trade in 1985, the testing period that corresponds to the in-sample (training plus selection) period 1980–1984. The time interval for the first testing subperiod associated with 1985 is October 1984 through March 1985 (observation phase in October through December of 1984, trading phase in January through March of 1985), the time interval for the second testing subperiod associated with 1985 is January through June of 1985 (observation phase in January through March, trading phase in April through June), and so on.

At the beginning of 1986, all periods are reassigned; in particular, the testing period is rolled forward 1 year, and the in-sample (training plus selection) period is redefined as the years 1981–1985. We run 10 GP trials, select one rule to be applied out-of-sample (again, if the threshold criteria defined in Section 4.3 are satisfied for any of the rules generated in these 10 trials), and trade according to this rule in the testing period, 1986. We continue rolling our window forward in this manner until we use the last year in our data set, 2005, as the testing period. Hence, we use 5 years of data at a time to evolve trading rules to employ for the year that follows, and we do this 21 times for every stock (once for each of the 21 testing periods).

Every GP experiment conducted as part of this study involved 10 trials, and each trial consisted of 50 generations. In every generation, a population size of 50,000 trading rules was used. The depth of each candidate solution decision tree was limited to 25 levels.

The process of running the 10 GP trials and selecting at most one rule to be tested in the out-of-sample period consists of the following steps:

1. Generate 50,000 random rules, evaluate their fitness in the training and selection periods, and identify and save all of the rules that satisfy the first six threshold criteria (see Section 4.3 and Table 2). If more than 50 rules satisfy the criteria, save only the 50 rules that have the highest fitness in the selection period.
2. For each rule, attach a probability of being chosen to be used in creating "offspring" rules in the next generation. The probability should correspond to each rule's fitness during the training and selection periods, so that the "more fit" rules will be more likely to mate. Choose rules from the current generation randomly, using the attached probabilities, and apply to these rules either the crossover operator (with probability 95%) or the mutation operator (with probability 5%), so as to generate 50,000 rules for the next generation. As above, evaluate the fitness of the rules in this population in the training and selection periods, then save all of the rules that satisfy the first six threshold criteria (up to a maximum of 50).
3. If this is not Generation 50, go back to Step 2 to create the population in the next generation. If this is Generation 50, begin the next trial by going back to Step 1, unless this is Trial 10.
4. If this is Generation 50 of Trial 10, take the rules that were saved during the 10 trials and discard those that do not satisfy the last two threshold criteria (see Section 4.3 and Table 2). If any rules remain, select the rule with the highest fitness in the selection period. Study the performance of this rule in the testing period.

Note that with a population size of 50,000 trading rules, our GP search is substantially more thorough than the GP searches of most earlier studies within this field, many of which have population sizes of 500 or fewer trading rules.[1] For example, Wang (2000) and Dempster and Jones (2001) used population sizes of 100 trading rules, whereas Ready (2002) and Potvin et al. (2004) followed the lead of Neely et al. (1997) and Allen and Karjalainen (1999) in using a population size of 500 trading rules. Roberts (2005) used a population size of 20,000 trading rules. It is well known in the GP literature (and inherently logical) that algorithm performance improves as

the population size (in other words, the number of candidate solutions created at every generation) increases. Naturally, and unfortunately, larger population sizes correspond to increased computing times.

4.2. Criterion of Fitness

We assume that investors' preferences are characterized by the logarithmic utility of terminal wealth (W_T) on the day that trading ends; that is, we assume that an investor's utility function is given by $U(W_T) = \ln(W_T)$. The goal of our experiments is to investigate whether investors whose trading horizon is 3 months, and whose preferences are characterized by this utility function, can increase their expected utility by switching from trading in accordance with the simple buy-and-hold rule to trading in accordance with a rule evolved by GP. (In these experiments, GP provides trading signals to rebalance the trading portfolio between the risky stock asset and the riskless 3-month T-bill asset at the end of every trading day, throughout the investment horizon.) Consistent with this goal, we set the fitness measure of a given GP rule for a given subperiod to be the utility of terminal wealth that would be realized at the end of that subperiod if that rule were followed throughout that subperiod. It then naturally follows that we evaluate a rule's fitness for a given training (or selection or testing) period by averaging the utilities of terminal wealth for the subperiods that constitute that training (or selection or testing) period. Thus the fitness criterion that we obtain after evaluating a given rule can be interpreted as an estimate of the expected utility of terminal wealth that corresponds to trading according to that rule. This fitness criterion allows us to take the raw returns based on the trading rules and adjust them for risk.

What follows is a sketch of how the trading process is simulated in our experiments. This trading setup is used to evaluate the fitness of one candidate solution trading strategy in one subperiod, be it a training, a selection, or a testing subperiod. On the first day of the trading phase, the value of the investor's cash account is set to an amount W_0 (we use \$100,000 for W_0). We make an important assumption, common in the literature, that the activities of our simulated trader do not have a major impact on the stock price.

We define the investor's wealth at the end of day t, W_t, as the value of the shares of stock in the portfolio on day t plus the value of the cash account. Suppose that on trading day t, the investor's wealth is W_t and the stock's price is P_t. This implies that the investor can have at most int(W_t/P_t) shares of stock on trading day t in his or her portfolio. (The "int" function truncates its input, to give the largest integer that does not exceed the input.)

The portfolio trading rule generated by GP determines the fraction of wealth that the investor will allocate to the purchase of stocks. The rule output for trading day t is computed using information available from the start of the observation phase up to trading day $t - 1$, inclusive of both the start and end dates. If on day t the rule output is α_t (with α_t constrained such that $0 \le \alpha_t \le 1$), the investor's interpretation is that he or she should have int[$\alpha_t (W_t / P_t)$] shares of stock in his or her portfolio. If the current number of shares of stock in the portfolio is different from this, the portfolio is rebalanced accordingly, at the end of (and at the closing price on) day t. The remaining amount, $[W_t - (P_t)(\text{int}[\alpha_t (W_t / P_t)])]$, is held in the cash account. We assume that this amount earns a rate of return equal to the 3-month T-bill rate.

When the investor buys shares of stock, both the cost of the shares and the transaction costs are subtracted from the cash account. Neely et al. (1997) have pointed out that adopting higher transaction costs in the training and selection periods would decrease the incidence of retaining rules that over-trade. Rules that over-trade are more likely to be overfitting the data. We follow Neely et al.'s lead and adopt unrealistically high transaction costs in training and selection periods, then use realistic transaction costs in the testing period. For the testing period, we choose to use the transaction cost structure used by Allen and Karjalainen (1999) for simulating trading in the S&P 500 index: a one-way transaction cost of 0.25%. This transaction cost structure was motivated by Sweeney (1988), who found that one-way transaction costs for institutional traders were in the range of 0.1–0.2%. Allen and Karjalainen argued that a one-way transaction cost of 0.25% incorporates all costs at realistic levels, including the cost of the market impact. Wang (2000) stated that the transaction cost structure for his S&P 500 index trading simulations corresponds to a one-way transaction cost of 0.12% and that this is a realistic assumption for institutional investors. For training and selection periods, we use the following (deliberately unrealistically high) transaction cost structure: a one-way transaction cost of 0.5% of the value of the transaction plus a two-way flat rate of \$5 per share of stock.

Table 2
Initial Threshold Parameter Settings for the Stock Experiments[a]

Threshold	Threshold Value
Minimum value of average subperiod utility of terminal wealth for training period[b]	11.537618 = ln(102,500)
Minimum value of average subperiod utility of terminal wealth for selection period[b]	11.537618
Minimum fraction of profitable subperiods in training period[b]	90%
Minimum fraction of profitable subperiods in selection period[b]	90%
Minimum wealth at end of every trading day in training period[b]	$90,000
Minimum wealth at end of every trading day in selection period[b]	$90,000
Minimum wealth at end of training period[c]	$125,000
Minimum wealth at end of selection period[d]	$125,000

Thus, we assume that the following activities take place on trading day *t*: First, we check whether the investor's wealth is negative. A bankruptcy constraint implies that the investor has to sell all of his or her shares of stock on the day the investor's wealth becomes negative. Second, the interest payment is added into the cash account. This interest payment is a function of the balance in the cash account on day $t - 1$, the 3-month T-bill rate, and the number of calendar days between trading days $t - 1$ and t. Third, the trading signal, α_t, is generated, and the portfolio is rebalanced at the close of day *t*. Fourth, the day's trading costs (which depend on whether day *t* is part of the in-sample period or a part of the testing period, as described above) are deducted from the cash account if the new position in stock is different from the old position. Lastly, the cash account is credited (debited) accordingly when stock is sold (bought).

On the last day of the trading phase of each subperiod, the investor sells the stock in the portfolio. We compute the terminal wealth, W_T, and the investor's utility of terminal wealth, as indicated above, with the equation $U(W_T) = \ln(W_T)$. To evaluate a rule's fitness in the training (selection) period, we average the utilities of terminal wealth for the subperiods that make up the training (selection) period for the particular stock. We use a similar method to evaluate the out-of-sample fitness of a specific rule evolved by GP.

In studies similar to ours, which evaluate the out-of-sample performance of trading rules evolved by GP as a means to test EMH, various trading rule performance criteria were used. We briefly describe several representative studies to give a sense of the variety of performance criteria being used. Allen and Karjalainen (1999) and Potvin et al. (2004) used as a fitness criterion the excess return due to trading by the rule (the return earned by applying the rule, less the return earned by using the buy-and-hold strategy). Fyfe et al. (2005) used the Sharpe ratio to adjust trading returns for risk; the measure employed by Dempster and Jones (2001) was a modified Stirling ratio, which is a function of the ratio of return to maximum drawdown. Relevant to this discussion, Neely (2003) used GP to generate three sets of rules that maximize three corresponding fitness criteria that adjust for risk: the Sharpe ratio, the X^* statistic (Sweeney and Lee, 1990), and the X_{eff} criterion (Dacorogna et al., 2001).

The risk-adjustment criterion introduced by Dacorogna et al. (2001), the X_{eff} criterion, is related to the criterion of fitness employed in our study to adjust for risk. Their criterion originates in utility theory but branches away from straightforward utility by measuring the utility derived from a trading strategy by an investor (whose preferences are characterized by constant absolute risk aversion) over a weighted average of return horizons, where the weights are based on the relative importance of those return horizons. The weighting function is chosen somewhat arbitrarily, may be arbitrarily changed for trading models with different trading frequencies, and does not originate in the standard portfolio model. In the present study, we are not considering multiple trading horizons; hence, we chose to use a fitness criterion that is related to expected utility maximization in a more straightforward fashion.

[a] The thresholds in this table had to be satisfied in the training and selection periods, as indicated, while the following (deliberately unrealistically high) transaction cost structure was used: a one-way transaction cost of 0.5% of the value of the transaction plus a two-way flat rate of $5 per share of stock.
[b] Wealth was reset to the initial value ($100,000) at the beginning of each subperiod.
[c] Wealth was not reset during the entire training period of 2.5 years (10 training subperiods).
[d] Wealth was not reset during the entire selection period of 2.5 years (10 selection subperiods).

According to the methodology followed in these experiments, a rule's performance in the in-sample subperiods determines whether the trading rule is saved in order to be applied out-of-sample. In Section 4.3 we clarify the process by which a rule's performance in-sample is used to select the rules to be applied out-of-sample in the testing period.

4.3 Thresholds for Saving Rules

This paper reports on tests of whether GP can use past price data to evolve trading rules that work well out-of-sample. In accordance with Daniel and Titman's (1999) definition of adaptive efficiency, we are testing whether rules that attain high fitness in-sample continue to have high fitness out-of-sample. The studies conducted by Neely et al. (1997), Allen and Karjalainen (1999), and Neely and Weller (1999) adopted an approach of saving one rule (the one with the highest fitness in the selection period) per trial and applying it out-of-sample if it outperforms the buy-and-hold rule in the selection period. These authors used a simple threshold criterion for determining which rule(s), if any, were selected to be tested out-of-sample. Their threshold criterion, however, doesn't seem to be the most intuitive one to follow. Few risk-averse real-world investors would be willing to accept only marginally improved performance in-sample as a reason to switch from investing in accordance with the buy-and-hold rule to investing in accordance with a rule evolved by GP. Consequently, in our experiments we investigate whether using more sophisticated threshold criteria tends to yield outcomes in which the algorithm saves rules with better out-of-sample fitness.

The idea is to find rules with the same performance (or better) in-sample as the performance the investor would like to see out-of-sample. In addition to looking for the rule that produces a high (on average) utility of terminal wealth at the end of the trading horizon, the investor might want to search for a rule that satisfies certain money management criteria (e.g., the rule is profitable a certain minimum fraction of the time; when a loss occurs, it doesn't exceed a certain maximum allowed amount; the rule yields a certain minimum return at longer horizons). The underlying goal is assumed to be to find a trading strategy that consistently produces high returns without the risk of extreme losses; each of the four threshold criteria above (high average utility and the three criteria related to money management) is a different way of specifying this goal.

Table 2 formally presents the set of threshold criteria that are to be satisfied in-sample (four criteria for each training period and each selection period, for a total of eight criteria). We use these criteria in each of the 504 individual experiments, as a means of homing in on the single rule, for each experiment, that we test out-of-sample. The set of 504 experiments comprises 21 experiments for each of the 24 individual stocks studied, corresponding to the 21 in-sample periods and the out-of-sample period associated with each of them. The procedure is as follows.

For each stock, and each of the 21 testing periods, we conduct one GP "run" comprising 10 trials. Each trial is made up of 50 generations. For every generation, we evaluate the fitness of each rule, and then discard the rules that do not satisfy the following six threshold criteria (three for the training period, and three for the selection period; see Table 2): (a) The average subperiod utility attained in the training period—which is obtained by computing the utility of terminal wealth at the end of the 3-month trading phase of each of the subperiods associated with that training period and then averaging over those subperiods—must have a certainty equivalent of at least 102.5% of the initial wealth, and likewise for the average subperiod utility attained in the selection period; (b) 90% of the training subperiods and 90% of the selection subperiods must be profitable; and (c) the minimum wealth observed at the end of every trading day in the training and selection periods must be greater than 90% of the initial wealth.

We run the algorithm described above and at the end of the 10th trial discard all the saved rules that do not satisfy the following two additional threshold criteria: The terminal wealth at the end of the training and selection periods must be at least 125% of the initial wealth. For purposes of these last two criteria, the wealth is not reset during the training period or the selection period, each of which is 2.5 years long and constitutes 10 subperiods. This is in stark contrast to the first six criteria, for purposes of which the wealth is reset at the beginning of every subperiod of the training period and every subperiod of the selection period. From the rules that are saved and satisfy all eight threshold criteria given in Table 2, we then select the rule with the highest fitness for the selection period and use that rule for out-of-sample testing. Thus, we are searching for a trading strategy that will maximize

the average utility of terminal wealth over the selection period and satisfy all eight threshold criteria given in Table 2.

In this paper, we also experiment with adopting more restrictive threshold criteria that rules must meet if they are to be saved (see Table 7). Because these thresholds are more restrictive, by definition fewer rules will satisfy them. We study how this change will influence the testing-period fitness of the rules that will be saved.

To summarize, the specific setup of our experiments is as follows: For the market in each individual stock, and each testing period, we perform 10 GP trials of 50 generations apiece. During each trial, at every generation we save at most 50 candidate solution decision trees: those that satisfy the first six threshold criteria referred to above and (if there are more than 50 candidates in that group) are among the 50 that have the highest selection-period fitness (highest utility of terminal wealth averaged over all selection subperiods). We thus save at most 25,000 rules during the 10 trials (i.e., at most 50 [rules per generation] × 50 [generations per trial] × 10 [trials]). At the end of 10 trials, we discard the rules that do not satisfy the final two thresholds (namely, the rules that do not result in terminal wealth of at least 125% of the initial wealth at the end of the training and selection periods). Of all the rules that remain, we select the one with the highest fitness in the selection period and then evaluate it in the testing period. The last procedure takes place, of course, only if at least one rule is found that satisfies all of the threshold criteria in the particular experiment; if no such rule is found, no rule is recorded (and used out-of-sample) in the corresponding testing period.

5. DATA

For these experiments, we chose 24 diverse companies traded on the NYSE, all well known and operating in a variety of industries. To ensure diversity among the companies studied, we picked two stocks each from the 12 industries in Fama and French's industry classification scheme.[2] Specifically, two companies were selected from each of the following industries: Consumer Durables, Consumer Non-Durables, Manufacturing, Energy, Chemicals, Business Equipment, Telecommunications, Utilities, Shops, Healthcare, Finance, and Other. Companies must have been active in the market for the time period that began at the start of the last quarter of 1979 and goes all the way through the end of 2005. Table 3 lists the companies used in this study, along with their corresponding industries. These data, along with 3-month T-bill rates, were provided by Datastream. The stock prices in these data sets are not adjusted for dividends. Because the trading rules evolved by GP sometimes result in not being fully invested in the stock, the decision not to include the dividends in the data set has the effect of underestimating returns to a greater extent for the buy-and-hold trading rule than for the GP trading rules. Bessembinder and Chan (1998) estimated the dividend yield to be 0.016% per day for the Dow Jones Industrial Average.

6. RESULTS

Our test of EMH involves using a GP algorithm to evolve trading rules that are increasingly fit, in-sample, for achieving high average utility of the investor's terminal wealth. For each testing period for which a trading rule was selected in-sample by the GP methodology (and satisfied all eight of the threshold criteria given in Table 2), Table 4 provides the average subperiod utility of terminal wealth (achieved by investing in accordance with that rule out-of-sample) for each individual stock, as a measure of how well the trading rules performed out-of-sample.

Blank spaces in Table 4 indicate testing periods corresponding to in-sample periods in which GP did not evolve any rules satisfying all the thresholds in Table 2. GP saved rules in 202 training periods (of the possible 504, corresponding to all 24 stocks and 21 training periods per stock).

We would like to compare the expected utility of using the GP methodology over the entire 21 years in which testing was done (1985–2005) to the expected utility of using the buy-and-hold methodology during that time period. For each methodology, we could compute the expected utility for each stock by applying that methodology to each of the 84 subperiods of the stated 21-year time period and then taking the average over the 84 subperiods. We could then average those expected utilities over the 24 stocks.

Such an approach would work well for the buy-and-hold rule, and we could interpret the final average as an estimate of the expected utility (for an investor whose preferences are characterized by the logarithmic utility function) of using the buy-and-hold rule. According to our GP methodology, however, we apply a rule evolved by GP to the out-of-sample period only if the rule meets all of the threshold criteria presented in Table 2. Thus, a comparison of the expected utilities for the two methodologies would be meaningful only if both methodologies were to produce a rule in each in-sample period for every stock. For some stocks, there are in-sample periods for which no rules were found that met all of the thresholds in Table 2; thus, the GP methodology and the buy-and-hold methodology cannot be compared in a completely consistent manner. To get around this problem and evaluate the expected utility of using GP to evolve trading rules to be used out-of-sample, we could compute the average (over all 21 testing periods and all 24 stocks) out-of-sample utility of terminal wealth using one of the two strategies described below.

Strategy 1 involves using the GP rule to trade in the out-of-sample periods corresponding to the in-sample periods for which GP is able to find a rule that satisfies the threshold criteria in Table 2, but investing in T-bills in out-of-sample periods corresponding to in-sample periods for which GP does not find a satisfactory rule. Strategy 2 is similar, but it uses the buy-and-hold strategy (rather than investing in T-bills) in out-of-sample periods corresponding to in-sample periods for which GP does not find a satisfactory rule. In order to ensure that path dependency is not a factor, these trading simulations assume that on the first day of the trading phase of every subperiod, the value of the investor's cash account, W_0, is reset to \$100,000. For each of the 24 stocks chosen for this study, Table 5 provides expected utilities of using Strategy 1, Strategy 2, and the buy-and-hold strategy to trade The results presented in Tables 5 and 6 indicate that the price series of the 24 markets in our study generally were characterized by adaptive efficiency between 1985 and 2005. In other words, according to our trading simulations, stock investors could not benefit from identifying trading strategies that were profitable and met various criteria in-sample, and then investing according to these strategies out-of-sample. Though the data presented in Table 6 show that the expected utility of Strategy 1 (11.5252, with a certainty equivalent \$101,235.02) is higher than that of investing all of one's wealth in T-bills (11.5247, certainty equivalent \$101,184.41), the expected utility of Strategy 1 is nevertheless lower than that of the buy-and-hold strategy (11.5306, certainty equivalent \$101,783.17). Strategy 2 is also dominated by the buy-and-hold strategy, despite having a higher expected utility (11.5289, certainty equivalent \$101,610.28) than Strategy 1. For the majority of the 24 individual stocks, moreover, Table 5 shows that we generally could not outperform a simple benchmark strategy by using GP to identify trading rules to be used (in either Strategy 1 or Strategy 2) in the out-of-sample period.

Seven stocks were exceptions. For GM (General Motors) and GT (Goodyear Tire & Rubber), employing Strategy 1 achieves a higher expected utility than either investing all of one's wealth in T-bills or trading in accordance with the buy-and-hold rule. For DD (DuPont), DUK (Duke Energy), S (Sprint), T (AT&T Inc.), and XRX (Xerox), employing Strategy 2 achieves a higher expected utility than either investing all of one's wealth in T-bills or trading in accordance with the buy-and-hold rule.

As stated above, rules that satisfied all of the thresholds presented in Table 2 were found for 202 of the 504 in-sample periods. It is possible that adopting a more restrictive set of thresholds would result in fewer rules being saved, and that these rules would have a higher average utility out-of-sample than the rules evolved and saved using the criteria listed in Table 2. To investigate this possibility, we established a more restrictive set of thresholds, presented in Table 7. With those criteria in mind, we took the rules saved during our original GP experiments (using the thresholds from Table 2), identified the rules that satisfied all of the more restrictive thresholds presented in Table 7, and applied these rules to the testing periods.

For each individual stock, Table 8 (the counterpart of Table 4) gives the average subperiod utility of terminal wealth for each testing period. The blank spaces indicate testing periods corresponding to the in-sample periods for which GP evolved no rules satisfying the thresholds in Table 7. When the thresholds presented in Table 7 were used to determine which rules get saved, GP saved rules in only 17 (3.37% of the possible 504) training periods, and those 17 training periods involve only 7 of the 24 stocks. When the thresholds presented in Table 2 were used, rules were saved for more than 40% of all of the training periods (202/504, as stated earlier).

Table 3
Individual Stocks Used for the GP Experiments

Industry	Company	Ticker Symbol
Consumer non-durables	Altria Group	MO
Consumer non-durables	Pepsico	PEP
Consumer durables	General Motors	GM
Consumer durables	Whirlpool	WHR
Manufacturing	Eastman Kodak	EK
Manufacturing	Goodyear Tire & Rubber	GT
Energy	Exxon Mobil	XOM
Energy	Halliburton	HAL
Chemicals	Dow Chemical	DOW
Chemicals	DuPont	DD
Business equipment	IBM	IBM
Business equipment	Xerox	XRX
Telecommunications	AT&T Inc.	T
Telecommunications	Sprint	S
Utilities	American Electric	AEP
Utilities	Duke Energy	DUK
Shops	Target	TGT
Shops	Wal-Mart Stores	WMT
Healthcare	Johnson and Johnson	JNJ
Healthcare	Pfizer	PFE
Finance	Bank of America	BAC
Finance	Merrill Lynch	MER
Other	Disney	DIS
Other	Hilton Hotels	HLN

We applied two additional strategies, Strategy 3 and Strategy 4, to compute the expected utilities of the individual stocks over the entire 21 years in which testing was done (1985–2005). Strategy 3 and Strategy 4 are analogous to Strategy 1 and Strategy 2, respectively, the only difference being that the new strategies are based on the thresholds given in Table 7 (rather than the ones from Table 2). For each of the 24 stocks chosen for this study, Table 9 provides expected utilities of using Strategy 3, Strategy 4, and the buy-and-hold strategy to trade out-of-sample. As before, the average utility corresponding to the riskless strategy is invariant across stocks; it is reported in Table 9 for comparison purposes. We also averaged (over all 24 stocks) the expected utilities corresponding to applying Strategy 3 to the individual stocks, and thus obtained the overall average utility of using Strategy 3 to trade out-of-sample. We did the same for Strategy 4. These results are presented in Table 6.

The fitness measure corresponding to using Strategy 3 (11.5250, certainty equivalent $101,214.77) is lower than that for Strategy 1 (11.5252, certainty equivalent $101,235.02). This indicates that either the rule saved using the more stringent thresholds is, on average, worse than the rule saved using less stringent thresholds (for the in-sample periods in which a rule gets saved for both sets of threshold criteria) or that the rule saved using less stringent thresholds performs, on average, much better than the Strategy 3 default option of holding T-bills (for the in-sample periods in which a rule is saved using the less stringent set of threshold criteria, but a rule is not saved using the more stringent set of threshold criteria), or possibly both. Similar reasoning would hold for Strategies 2 and 4, but in that case the more stringent thresholds produce higher expected utility: The fitness corresponding to using Strategy 4 (11.5306, certainty equivalent $101,783.17) exceeds that of Strategy 2 (11.5289, certainty equivalent $101,610.28). Furthermore, the two strategies that employed buy-and-hold when a GP rule was not saved (Strategies 2 and 4) outperformed the corresponding T-bill strategies (1 and 3, respectively). Looking at it another way, because Strategy 2 outperforms Strategy 1, and Strategy 4 outperforms Strategy 3, neither Strategy 1 nor Strategy 3 (the strategies with a default option of holding T-bills) would be chosen by a rational investor. Thus the data presented in Table 6 show that using more stringent criteria improves the out-of-sample fitness of the rules that get saved, because Strategy 4 outperforms Strategy 2. The fact that Strategy 1 outperforms Strategy 3 is irrelevant because a rational investor would never be interested in either of these two strategies, instead preferring buy-and-hold—rather than T-bills—as the default when a rule is not saved.

Table 4
Average Subperiod Utility of Terminal Wealth [a] for Individual Testing Periods, Using GP Rules Subject to Threshold Criteria from Table 2

Panel A:

Year	Altria Group (MO)	Pepsico (PEP)	General Motors (GM)	Whirlpool (WHR)	Eastman Kodak (EK)	Goodyear Tire & Rubber (GT)
1985					11.5294	
1986						
1987					11.4288	
1988						
1989	11.6314	11.5312			11.4711	
1990	11.5273	11.5152				
1991	11.5950	11.5260			11.5260	
1992	11.5221	11.5214				11.5805
1993	11.3682	11.4992				
1994						
1995	11.5251					
1996		11.5821				
1997	11.5248	11.5460				
1998	11.4834	11.5275			11.5523	
1999	11.5083	11.5178			11.5479	
2000	11.5204	11.4857			11.5194	
2001	11.5768	11.5271			11.3876	
2002	11.5091	11.5373	11.4973		11.4724	
2003	11.4571	11.5233	11.5587	11.5233		
2004	11.4872	11.5229		11.5232		
2005	11.5233	11.5330				
Average[b]	11.5173	11.5259	11.5280	11.5232	11.4928	11.5805

Panel B:

Year	Exxon Mobil (XOM)	Halliburton (HAL)	Dow Chemical (DOW)	DuPont (DD)	IBM (IBM)	Xerox (XRX)
1985	11.5272	11.4596			11.5584	
1986		11.4690			11.5262	
1987					11.5251	
1988	11.5298	11.5620			11.5154	
1989	11.5314	11.5312				
1990				11.5303		
1991				11.5260		
1992	11.5214	11.5214	11.5214	11.5214	11.5109	11.5214
1993						
1994						
1995		11.5251				
1996						11.5466
1997						
1998	11.4990	11.4780			11.5524	
1999		11.5060		11.5483	11.5241	11.5230
2000	11.5334	11.5526		11.4707	11.4486	11.4343
2001	11.5724	11.5163		11.5069		11.5231
2002	11.5109	11.3554		11.5234		11.5508
2003	11.5053	11.5233		11.5215		
2004	11.5351	11.5637	11.5232	11.5232		11.5189
2005	11.5104	11.5922	11.5232	11.5304		11.5232
Average	11.5251	11.5111	11.5226	11.5202	11.5201	11.5176

[a] $U(W_T) = \ln(W_T)$, initial wealth = \$100,000.
[b] Average over all testing periods for which a GP rule for the corresponding in-sample period was found.

Table 4 *(continued)*

Panel C:

Year	AT&T Inc. (T)	Sprint (S)	American Electric (AEP)	Duke Energy (DUK)	Target (TGT)	Wal-Mart Stores (WMT)
1985					11.4966	11.5688
1986					11.5262	11.5734
1987					11.5250	11.5406
1988					11.5271	11.5478
1989					11.5889	11.5791
1990						11.6283
1991	11.5085				11.5260	11.5513
1992					11.5075	
1993					11.5195	
1994					11.5147	
1995						
1996						
1997						
1998					11.6253	
1999					11.6049	
2000	11.4833				11.5698	
2001	11.5360	11.4901				
2002		11.5234				
2003	11.4856					
2004	11.4895	11.5208		11.5320		
2005	11.5232	11.5232				
Average	11.5044	11.5144		11.5320	11.5443	11.5699

Panel D:

Year	Johnson and Johnson (JNJ)	Pfizer (PFE)	Bank of America (BAC)	Merrill Lynch (MER)	Disney (DIS)	Hilton Hotels (HLN)
1985	11.5294			11.5294	11.5294	11.5294
1986		11.5262		11.5262		
1987	11.5535	11.5251			11.5251	
1988	11.5091	11.5671		11.5473	11.5476	
1989	11.5312	11.5227			11.5312	
1990	11.5332				11.4660	11.4624
1991	11.5388	11.5342			11.5623	11.5021
1992	11.5214	11.5210		11.5191	11.5613	11.5214
1993	11.4909	11.5116	11.5195		11.5045	11.5195
1994	11.5383	11.5775		11.4886	11.5209	
1995	11.5199	11.5756		11.5613		11.5251
1996	11.5800	11.6189		11.5456	11.5165	11.6062
1997	11.5499	11.6087		11.5244	11.5780	11.5538
1998		11.5396		11.5814	11.5245	11.4821
1999		11.4903	11.5191		11.5142	
2000		11.5404	11.4248		11.5837	
2001			11.5251		11.5007	11.5201
2002			11.5234		11.5165	
2003			11.5233		11.5233	
2004			11.5232		11.5232	11.5232
2005			11.5235		11.5332	11.5232
Average	11.5330	11.5471	11.5102	11.5359	11.5296	11.5224

The results presented in Tables 6 and 9 allow us to conclude that the price series of the 24 markets in the present study were generally characterized by adaptive efficiency between 1985 and 2005. Specifically, our trading simulations reveal that stock investors would not benefit by identifying trading strategies that worked in-sample and then trading according to these strategies out-of-sample. According to Table 6, the expected utility of Strategy 3

(11.5250, certainty equivalent $101,214.77) is higher than the expected utility of investing all of one's wealth in T-bills (11.5247, certainty equivalent $101,184.41) but lower than that of the buy-and-hold strategy (11.5306, certainty equivalent $101,783.17). Two of the strategies (Strategy 4 and buy-and-hold) have expected utilities that are identical when taken to four decimal places, but we can find small differences by expanding to five decimal places, and those results support the conclusion above: The expected utility of Strategy 4 (11.53061, certainty equivalent $101,784.18) is lower than that of the buy-and-hold strategy (11.53063, certainty equivalent $101,786.22). Moreover, according to Table 9, using GP to identify trading rules to be used to trade out-of-sample and then employing either Strategy 3 or Strategy 4 does not outperform the simple benchmark strategies for the majority of the 24 individual stocks. Table 9 demonstrates that for all 24 stocks, employing Strategy 3 does not achieve a higher expected utility than both alternative strategies of investing all of one's wealth in T-bills and trading in accordance with the buy-and-hold rule. For the stocks with the ticker symbols MER (Merrill Lynch) and TGT (Target), employing Strategy 4 achieves a higher expected utility than either investing all of one's wealth in T-bills or trading in accordance with the buy-and-hold rule.

Table 5
Average, over All Testing Subperiods, of Subperiod Utilities of Terminal Wealth,[a] Using GP Rules (Subject to Threshold Criteria from Table 2) and Simple Trading Strategies[b]

Panel A: Consumer Non-Durables, Consumer Durables, and Manufacturing			
Company	**Buy-and-Hold Strategy[c]**	**Strategy 1[d]**	**Strategy 2[e]**
Altria Group (MO)	11.5457	11.5197	11.5307
Pepsico (PEP)	11.5500	11.5262	11.5371
General Motors (GM)	11.5075	11.5251	11.5156
Whirlpool (WHR)	11.5187	11.5247	11.5209
Eastman Kodak (EK)	11.5063	11.5107	11.5060
Goodyear Tire & Rubber (GT)	11.5119	11.5275	11.5119
Panel B: Energy, Chemicals, and Business Equipment			
Company	**Buy-and-Hold Strategy**	**Strategy 1**	**Strategy 2**
Exxon Mobil (XOM)	11.5370	11.5248	11.5356
Halliburton (HAL)	11.5203	11.5156	11.5170
Dow Chemical (DOW)	11.5340	11.5247	11.5313
DuPont (DD)	11.5318	11.5229	11.5373
IBM (IBM)	11.5188	11.5228	11.5195
Xerox (XRX)	11.5178	11.5226	11.5348
Panel C: Telecommunications, Utilities, and Shops			
Company	**Buy-and-Hold Strategy**	**Strategy 1**	**Strategy 2**
AT&T Inc. (T)	11.5288	11.5192	11.5336
Sprint (S)	11.5274	11.5230	11.5399
American Electric (AEP)	11.5176	11.5247	11.5176
Duke Energy (DUK)	11.5279	11.5251	11.5283
Target (TGT)	11.5445	11.5358	11.5426
Wal-Mart Stores (WMT)	11.5526	11.5387	11.5473
Panel D: Healthcare, Finance, and Other			
Company	**Buy-and-Hold Strategy**	**Strategy 1**	**Strategy 2**
Johnson and Johnson (JNJ)	11.5505	11.5290	11.5347
Pfizer (PFE)	11.5416	11.5397	11.5316
Bank of America (BAC)	11.5355	11.5199	11.5311
Merrill Lynch (MER)	11.5407	11.5294	11.5344
Disney (DIS)	11.5535	11.5292	11.5353
Hilton Hotels (HLN)	11.5230	11.5234	11.5240

[a] $U(W_T) = \ln(W_T)$, initial wealth = $100,000.
[b] The riskless strategy consists of always investing in T-bills. The utility from this strategy, averaged across all subperiods, is 11.5247.
[c] Investing the entire wealth in the given stock, then holding these shares until the end of the trading horizon.
[d] Investing the initial wealth, during each testing period, in accordance with the GP rule for the corresponding in-sample period if such a rule was found, and investing the initial wealth in T-bills otherwise. When no rule is found, the value given for Strategy 1 is 11.5247, the utility of the strategy that allocates all wealth to T-bills.
[e] Investing the initial wealth, during each testing period, in accordance with the GP rule for the corresponding in-sample period if such a rule was found, and investing the initial wealth in accordance with a simple buy-and-hold rule otherwise. When no rule is found, the value given for Strategy 2 is the utility of the buy-and-hold strategy.

Table 6
Average, over All Testing Subperiods and Stocks, of Subperiod Utilities of Terminal Wealth,[a] and Associated Values of Certainty Equivalent Wealth and Annualized Rate of Return, Using GP Rules and Simple Trading Strategies

Strategy	Overall Average Subperiod Utility of Terminal Wealth	Certainty Equivalent Wealth	Annualized Rate of Return
Riskless[b]	11.5247	$101,184.41	4.82%
Buy-and-hold[c]	11.5306	$101,783.17	7.33%
Strategy 1[c]	11.5252	$101,235.02	5.03%
Strategy 2[c]	11.5289	$101,610.28	6.60%
Strategy 3[c]	11.5250	$101,214.77	4.95%
Strategy 4[c]	11.5306[d]	$101,783.17	7.33%

Table 7
Additional Threshold Parameter Settings for the Stock Experiments[a]

Threshold	Threshold Value
Minimum value of average subperiod utility of terminal wealth for training period[b]	11.573550= ln(106,250)
Minimum value of average subperiod utility of terminal wealth for selection period[b]	11.573550
Minimum fraction of profitable training subperiods in training period[b]	90%
Minimum fraction of profitable selection subperiods in selection period[b]	90%
Minimum wealth at end of every trading day in training period[b]	$90,000
Minimum wealth at end of every trading day in selection period[b]	$90,000
Minimum wealth at end of training period[c]	$150,000
Minimum wealth at end of selection period[d]	$150,000

Table 8
Average Subperiod Utility of Terminal Wealth[a] for Individual Testing Periods, Using GP Rules Subject to Threshold Criteria from Table 7

Panel A: Consumer Non-Durables, Consumer Durables, and Manufacturing

Year	Altria Group (MO)	Pepsico (PEP)	General Motors (GM)	Whirlpool (WHR)	Eastman Kodak (EK)	Goodyear Tire & Rubber (GT)
1985						
1986						
1987						
1988						
1989						
1990						
1991						
1992						
1993						
1994						
1995						
1996						
1997						
1998						

[a] $U(W_T) = \ln(W_T)$, initial wealth = $100,000.
[b] Averaged over 21 × 4 = 84 subperiods.
[c] Averaged over 84 subperiods and 24 stocks.
[d] When expanded to one more decimal place, this utility value (11.53061, certainty equivalent $101,784.18) is slightly lower than the expected utility of the buy-and-hold strategy (11.53063, certainty equivalent $101,786.22).
[a]The thresholds in this table had to be satisfied in the training and selection periods, as indicated, while the following (deliberately unrealistically high) transaction cost structure was used: a one-way transaction cost of 0.5% of the value of the transaction plus a two-way flat rate of $5 per share of stock.
[b] Wealth was reset to the initial value ($100,000) at the beginning of each subperiod.
[c] Wealth was not reset during the entire training period of 2.5 years (10 training subperiods).
[d] Wealth was not reset during the entire selection period of 2.5 years (10 selection subperiods).
[a] $U(W_T) = \ln(W_T)$, initial wealth = $100,000.

Table 8 ***(continued)***

Year	Altria Group (MO)	Pepsico (PEP)	General Motors (GM)	Whirlpool (WHR)	Eastman Kodak (EK)	Goodyear Tire & Rubber (GT)
Panel A:						
1999						
2000						
2001						
2002	11.5458					
2003						
2004	11.4872					
2005	11.5232					
Average[b]	11.5187					

Panel B:	**Energy, Chemicals, and Business Equipment**					
Year	**Exxon Mobil (XOM)**	**Halliburton (HAL)**	**Dow Chemical (DOW)**	**DuPont (DD)**	**IBM (IBM)**	**Xerox (XRX)**
1985						
1986						
1987						
1988						
1989						
1990						
1991						
1992						
1993						
1994						
1995						
1996						
1997						
1998						
1999						
2000					11.4521	
2001					11.5759	
2002						
2003						
2004						
2005						
Average					11.5140	

Panel C:	**Telecommunications, Utilities, and Shops**					
Year	**AT&T Inc. (T)**	**Sprint (S)**	**American Electric (AEP)**	**Duke Energy (DUK)**	**Target (TGT)**	**Wal-Mart Stores (WMT)**
1985						11.5764
1986						
1987						11.5406
1988						11.5102
1989						
1990						
1991						
1992						
1993						11.4821
1994						
1995						
1996						

[b] Average over all testing periods for which a GP rule for the corresponding in-sample period was found.

Panel C: Telecommunications, Utilities, and Shops *(continued)*

Year	AT&T Inc. (T)	Sprint (S)	American Electric (AEP)	Duke Energy (DUK)	Target (TGT)	Wal-Mart Stores (WMT)
1997						
1998						
1999						
2000					11.5840	
2001					11.5855	11.4972
2002						
2003						
2004						
2005						
Average					11.5847	11.5213

Panel D: Healthcare, Finance, and Other

Year	Johnson and Johnson (JNJ)	Pfizer (PFE)	Bank of America (BAC)	Merrill Lynch (MER)	Disney (DIS)	Hilton Hotels (HLN)
1985						
1986						
1987						
1988						
1989						
1990						
1991						
1992						
1993						
1994						
1995						
1996						
1997						
1998				11.5814		
1999		11.4903				
2000		11.5332				
2001						
2002						
2003				11.5942		
2004						
2005					11.5232	
Average		11.5117		11.5878	11.5232	

7. CONCLUSION

In the research presented here, we used GP to study whether stock markets are adaptively efficient. Specifically, we applied GP to the trading of 24 stocks and tested whether the algorithm could discover trading strategies that outperform simplistic trading rules in the out-of-sample period. The scope of the study was broad, encompassing 504 in-sample periods (24 stocks × 21 in-sample periods per stock). Our GP search was thorough, in that we examined the algorithm's performance in the 504 out-of-sample periods, using a population size of 50,000 trading rules in each generation, substantially larger than the 500 or fewer trading rules employed in most earlier studies.

In our trading simulations, a trading strategy is assumed to be the fraction of wealth allocated to the risky asset. As mentioned earlier, Samuelson (1997) and Gollier (1997) showed that trading strategies that diversify between the risky and the riskless assets at every point in time dominate "bang–bang" strategies. Our definition of a trading strategy allowed us to test adaptive efficiency without the testing being subject to the biases encountered by earlier studies (e.g., Dempster and Jones, 2001; Fyfe et al., 2005; Potvin et al., 2004), which limited GP strategies to simple buy–sell signals.

Table 9
Average, over All Testing Subperiods, of Subperiod Utilities of Terminal Wealth,[a] Using GP Rules (Subject to Threshold Criteria from Table 7) and Simple Trading Strategies[b]

Panel A: Consumer Non-Durables, Consumer Durables, and Manufacturing

Company	Buy-and-Hold Strategy[c]	Strategy 3[d]	Strategy 4[e]
Altria Group (MO)	11.5457	11.5240	11.5409
Pepsico (PEP)	11.5500	11.5247	11.5500
General Motors (GM)	11.5075	11.5247	11.5075
Whirlpool (WHR)	11.5187	11.5247	11.5187
Eastman Kodak (EK)	11.5063	11.5247	11.5063
Goodyear Tire & Rubber (GT)	11.5119	11.5247	11.5119

Panel B: Energy, Chemicals, and Business Equipment

Company	Buy-and-Hold Strategy	Strategy 3	Strategy 4
Exxon Mobil (XOM)	11.5370	11.5247	11.5370
Halliburton (HAL)	11.5203	11.5247	11.5203
Dow Chemical (DOW)	11.5340	11.5247	11.5340
DuPont (DD)	11.5318	11.5247	11.5318
IBM (IBM)	11.5188	11.5238	11.5205
Xerox (XRX)	11.5178	11.5247	11.5178

Panel C: Telecommunications, Utilities, and Shops

Company	Buy-and-Hold Strategy	Strategy 3	Strategy 4
AT&T Inc. (T)	11.5288	11.5247	11.5288
Sprint (S)	11.5274	11.5247	11.5274
American Electric (AEP)	11.5176	11.5247	11.5176
Duke Energy (DUK)	11.5279	11.5247	11.5279
Target (TGT)	11.5445	11.5305	11.5501
Wal-Mart Stores (WMT)	11.5526	11.5238	11.5510

Panel D: Healthcare, Finance, and Other

Company	Buy-and-Hold Strategy	Strategy 3	Strategy 4
Johnson and Johnson (JNJ)	11.5505	11.5247	11.5505
Pfizer (PFE)	11.5416	11.5236	11.5394
Bank of America (BAC)	11.5355	11.5247	11.5355
Merrill Lynch (MER)	11.5407	11.5308	11.5426
Disney (DIS)	11.5535	11.5247	11.5444
Hilton Hotels (HLN)	11.5230	11.5247	11.5230

Our study complements the research done by Allen and Karjalainen (1999), Neely et al. (1997), Neely and Weller (1999, 2003), and Neely (2003). Those studies saved a number of buy–sell rules, evolved using GP, and then considered the returns on portfolios formed by using either the signals derived from the saved rules or the in-sample return characteristics of those saved rules. Also, similar to other studies that used GP to find trading rules, our study avoided the data-snooping bias that plagues studies that test market efficiency by analyzing the out-of-sample performance of common technical trading rules.

Recent studies (e.g., Fyfe et al., 2005; Neely, 2003) used GP to evolve trading rules using a number of fitness criteria that adjusted trading rule returns for risk. In this study we adjusted trading rule returns for risk in a manner different from that used in the earlier studies in this field: We used average utility of terminal wealth as our

[a] $U(W_T) = \ln(W_T)$, initial wealth = \$100,000.
[b] The riskless strategy consists of always investing in T-bills. The utility from this strategy, averaged across all subperiods, is 11.5247.
[c] Investing the entire wealth in the given stock, then holding these shares until the end of the trading horizon.
[d] Investing the initial wealth, during each testing period, in accordance with the GP rule for the corresponding in-sample period if such a rule was found, and investing the initial wealth in T-bills otherwise. When no rule is found, the value given for Strategy 3 is 11.5247, the utility of the strategy that allocates all wealth to T-bills.
[e] Investing the initial wealth, during each testing period, in accordance with the GP rule for the corresponding in-sample period if such a rule was found, and investing the initial wealth in accordance with a simple buy-and-hold rule otherwise. When no rule is found, the value given for Strategy 4 is the utility of the buy-and-hold strategy.

fitness criterion. Out of all the criteria presented in the literature that we examined, our criterion is closest to, but distinct from, one of the risk-adjustment measures used by Neely, namely, the X_{eff} measure. That measure (introduced by Dacorogna et al., 2001, in a study that did not employ GP methodology) is the only one in the literature on the use of GP methodology to generate trading rules which employed the concept of utility.

Another contribution of the present paper is that, in our experiments, we extended the simple threshold criterion of earlier studies (e.g., save a rule only if it outperforms the buy-and-hold rule in the selection period) to more complex threshold criteria. Previous research tested EMH by simply retaining rules, for future analysis, that outperformed the buy-and-hold rule in-sample (e.g., Allen and Karjalainen, 1999; Fyfe et al., 2005; Neely and Weller, 1999; Neely et al., 1997; Ready, 2002). In contrast, we saved rules that satisfied various money management criteria in-sample, in addition to having the highest average utility of terminal wealth in the selection period. Our findings indicate that our use of more stringent thresholds to choose which rules to retain (and then apply in the testing period) improved the fitness (albeit slightly) of the rules that ended up being applied in the testing period, while at the same time (and as fully expected) reducing the number of rules retained.

Results indicate that the stock markets studied can, in general, be characterized by adaptive efficiency during the time period 1985–2005. An investor who bought and held stocks usually achieved a higher expected utility than an investor who used a rule saved by a GP algorithm (given that the rule was used when this algorithm did save a rule, and a buy-and-hold strategy was used otherwise). This result is not unexpected: Fyfe et al. (2005) and Neely (2003) found that when the returns based on rules generated by GP are adjusted for risk, the results are consistent with market efficiency.

It is important to note the possibility that the results might change if some of the study's parameters (e.g., the investor's trading horizon or the values of the threshold criteria) were changed or if some of the GP settings (e.g., the number of generations, the population size, or the number of "building blocks" that GP is allowed to use to create trading rules) were altered. Our experiments covered a wider range of stock price data than used in most studies, and the population size of our candidate solutions was larger than most, so we believe our results to be more generalizable than most that appear in the literature.

One of the more interesting extensions of this research would use a variety of other relevant time series (such as data from the companies' financial statements) as building blocks when applying GP to evolve trading strategies for individual stocks. To reduce the possibility that our results would be artifacts of the periods chosen, we adopted a "sliding window" approach that allowed us to run our GP experiments for many in-sample periods (and the corresponding out-of-sample periods) for a given data set. In the research presented here, we used 5 years as the length of an in-sample (training plus selection) period, and 1 year as the length of a testing period. We looked only at trading horizons of 3 months. A natural extension of our research would involve studying how the results would change when the trading horizon is changed, and when lengths of the various (training, selection, and testing) periods are changed. This extension would allow a study of the "term structure" of adaptive efficiency: It would be possible to learn how quickly information contained in past prices starts to lose its value to investors.

The crash of 2008 and earlier crashes led a number of observers to conclude that markets might not be as rational as previously believed, and that EMH might not be true. If a market is established to be inefficient, policy makers, investment professionals, and individual investors might be interested in the relative degree of the inefficiency. The methodology developed in this paper provides a means to measure the relative degree of inefficiency, by measuring the profits of traders who are trying to exploit market inefficiencies. It thus has potential value as a guide to market regulation.

ACKNOWLEDGEMENTS

We are grateful for the financial support provided by the Canadian Social Sciences and Humanities Research Council and for assistance with computer coding provided by Ian Davis.

AUTHOR INFORMATION

Stan Miles is an Assistant Professor of Economics at Thompson Rivers University, School of Business and Economics, 900 McGill Road, Kamloops, B.C. V2C 5N3, Canada. His research interests are Financial Economics and Computational Economics. He can be reached at stanmiles@tru.ca.

Barry Smith is Chair and Professor of Economics at York University, Faculty of Liberal Arts and Professional Studies, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada. His research interests include Econometrics and Labor Economics. He can be reached at jbsmith@yorku.ca.

REFERENCES

1. Al-Debie, M., & Walker, M. (1999). Fundamental information analysis: An extension and UK evidence. *Journal of Accounting Research, 31,* 261–280.
2. Allen, F., & Karjalainen, R. (1999). Using genetic algorithms to find technical trading rules. *Journal of Financial Economics, 51,* 245–271.
3. Allen, H. L., & Taylor, M. P. (1990). Charts, noise and fundamentals in the foreign exchange market. *Economic Journal, 100,* 49–59.
4. Álvarez-Díaz, M., & Álvarez, A. (2005). Genetic multi-model composite forecast for non-linear prediction of exchange rates. *Empirical Economics, 30,* 643–663.
5. Álvarez-Díaz, M., & Miguez, G. C. (2008). The quality of institutions: A genetic programming approach. *Economic Modelling, 25,* 161–169.
6. Bessembinder, H., & Chan, K. (1998). Market efficiency and the returns to technical analysis. *Financial Management, 27,* 5–17.
7. Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance, 47,* 1731–1764.
8. Chen, J. S., Chang, C. L., Hou, J. L., & Lin, Y. T. (2008). Dynamic proportion portfolio insurance using genetic programming with principal component analysis. *Expert Systems with Applications, 35,* 273–278.
9. Dacorogna, M. M., Gençay, R., Müller, U. A., & Pictet, O. V. (2001). Effective return, risk aversion and drawdowns. *Physica A, 289,* 229–248.
10. Daniel, K., & Titman, S. (1999). Market efficiency in an irrational world. *Financial Analysts Journal, 55,* 28–40.
11. Dempster, M. A. H., & Jones, C. (2001). A real-time adaptive trading system using genetic programming. *Quantitative Finance, 1,* 397–413.
12. Fama, E. F. (1976). *Foundations of finance.* New York, NY: Basic Books.
13. Fama, E. F. (1991). Efficient capital markets: II. *Journal of Finance, 46,* 1575–1617.
14. Fernández-Rodríguez, F., González-Martel, C., & Sosvilla-Rivero, S. (2000). On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market. *Economics Letters, 69,* 89–94.
15. Fyfe, C., Marney, J. P., & Tarbert, H. F. E. (1999). Technical analysis versus market efficiency—a genetic programming approach. *Applied Financial Economics, 9,* 183–191.
16. Fyfe, C., Marney, J. P., & Tarbert, H. F. E. (2005). Risk adjusted returns from technical trading: A genetic programming approach. *Applied Financial Economics, 15,* 1073–1077.
17. Gehrig, T., & Menkhoff, L. (2006). Extended evidence on the use of technical analysis in foreign exchange. *International Journal of Finance and Economics, 11,* 327–338.
18. Gençay, R. (1999). Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules. *Journal of International Economics, 47,* 91–107.
19. Gollier, C. (1997). On the inefficiency of bang–bang and stop-loss portfolio strategies. *Journal of Risk and Uncertainty, 14,* 143–154.
20. Jin, N., Tsang, E., & Li, J. (2009). A constraint-guided method with evolutionary algorithms for economic problems. *Applied Soft Computing, 9,* 924–935.
21. Kaboudan, M. A. (1999). A measure of time series predictability using genetic programming applied to stock returns. *Journal of Forecasting, 18,* 345–357.

22. Kaboudan, M. A. (2000). Genetic programming prediction of stock market prices. *Computational Economics, 16,* 207–236.
23. Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection.* Cambridge, MA: MIT Press.
24. Kwon, K., & Kish, R. J. (2002). Technical trading strategies and return predictability: NYSE. *Applied Financial Economics, 12,* 639–653.
25. Lev, B., & Thiagarajan, R. (1993). Fundamental information analysis. *Journal of Accounting Research, 31,* 190–215.
26. Lensberg, T. (1999). Investment behavior under Knightian uncertainty—an evolutionary approach. *Journal of Economic Dynamics & Control, 23,* 1587–1604.
27. Lensberg, T., & Schenk-Hoppé, K. R. (2007). On the evolution of investment strategies and the Kelly rule—a Darwinian approach. *Review of Finance, 11,* 25–50.
28. Lien, D., Tse, Y. K., & Zhang, X. (2003). Structural change and lead-lag relationship between the Nikkei spot index and futures price: A genetic programming approach. *Quantitative Finance, 3,* 136–144.
29. Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *Journal of Finance, 55,* 1705–1765.
30. McKee, T. E., & Lensberg, T. (2002). Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research, 138,* 436–451.
31. Menkhoff, L., & Schmidt, U. (2005). The use of trading strategies by fund managers: Some first survey evidence. *Applied Economics, 37,* 1719–1730.
32. Neely, C. J. (2003). Risk-adjusted, ex ante, optimal technical trading rules in equity markets. *International Review of Economics and Finance, 12,* 69–87.
33. Neely, C. J., & Weller, P. A. (1999). Technical trading rules in the European monetary system. *Journal of International Money and Finance, 18,* 429–458.
34. Neely, C. J., & Weller, P. A. (2003). Intraday technical trading in the foreign exchange market. *Journal of International Money and Finance, 22,* 223–237.
35. Neely, C. J., Weller, P., & Ditmar, R. (1997). Is technical analysis in the foreign exchange market profitable? A genetic programming approach. *Journal of Financial and Quantitative Analysi, 32,* 405–426.
36. Östermark, R. (1999). Solving irregular econometric and mathematical optimization problems with a genetic hybrid algorithm. *Computational Economics, 13,* 103–115.
37. Park, C. H., & Irwin, S. H. (2007). What do we know about the profitability of technical analysis? *Journal of Economic Surveys, 21,* 786–826.
38. Potvin J. Y., Soriano, P., & Vallee, M. (2004). Generating trading rules on the stock markets with genetic programming. *Computers & Operations Research,31,* 1033–1047.
39. Ready, M. J. (2002). Profits from technical trading rules. *Financial Management, 31,* 43–62.
40. Roberts, M. C. (2005). Technical analysis and genetic programming: Constructing and testing a commodity portfolio. *Journal of Futures Markets, 25,* 643–660.
41. Samuelson, P. A. (1997). Proof by certainty equivalents that diversification-across-time does worse, risk corrected, than diversification-throughout-time. *Journal of Risk and Uncertainty,14,* 129–142.
42. Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business, 39,* 119–138.
43. Skouras, S. (2001). Financial returns and efficiency as seen by an artificial technical analyst. *Journal of Economic Dynamics and Control, 25,* 213–244.
44. Sullivan, R., Timmermann, A., & White, H. (1999). Data snooping, technical trading rule performance, and the bootstrap. *Journal of Finance, 54,* 1647–1691.
45. Sweeney, R. J. (1988). Some new filter rule tests: Methods and results. *Journal of Financial and Quantitative Analysis, 23,* 285–300.
46. Sweeney, R. J., & Lee, E. J. Q. (1990). International dimensions of securities and currency markets. In R. Aggarwal and C. F. Lee (Eds.), *Advances in financial planning and forecasting series*(Vol. 4, Part A, pp. 59–79). Greenwich, CT: JAI Press.
47. Taylor, S. J. (2000). Stock index and price dynamics in the UK and the US: New evidence from trading rule and statistical analysis. *European Journal of Finance, 6,* 39–69.

48. Wagner, N., & Brauer, J. (2007). Using dynamic forecasting genetic programming (DFGP) to forecast United States Gross Domestic Product (US GDP) with military expenditure as an explanatory variable. *Defence and Peace Economics, 18,* 451–466.
49. Wang, J. (2000). Trading and hedging in S&P 500 spot and futures markets using genetic programming. *Journal of Futures Markets, 20,* 911–942.
50. Zhou, X., & Dong, M. (2004). Can fuzzy logic make technical analysis 20/20? *Financial Analysts Journal, 60,* 54–75.

Footnotes

1. A GP study with the scope of the present one demands significant amounts of computing time when a population size of 50,000 solution candidates is used. Each trial takes more than 8 CPU hours on a 3.0 GHz Intel Pentium 4 processor. Rough estimation suggests that the experiments performed as part of this study would take more than 40,000 hours to repeat using one computer [8 (hours per trial) × 10 (trials per subperiod) × 21 (subperiods per stock) × 24 (stocks)].
2. http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_12_ind_port.html.