DISCUSSION PAPER SERIES

No. 3060

FORECAST EVALUATION WITH SHARED DATA SETS

Ryan Sullivan, Allan G Timmermann
and Halbert White

FINANCIAL ECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP3060.asp**

ISSN 0265-8003

FORECAST EVALUATION WITH SHARED DATA SETS

Ryan Sullivan, Economic Analysis LLC
Allan G Timmermann, University of California, San Diego and CEPR
Halbert White, QuantMetrics R&D Associates LLC

Discussion Paper No. 3060
November 2001

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **Financial Economics**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as a private educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions. Institutional (core) finance for the Centre has been provided through major grants from the Economic and Social Research Council, under which an ESRC Resource Centre operates within CEPR; the Esmée Fairbairn Charitable Trust; and the Bank of England. These organizations do not give prior review to the Centre's publications, nor do they necessarily endorse the views expressed therein.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Ryan Sullivan, Allan G Timmermann and Halbert White

CEPR Discussion Paper No. 3060

November 2001

ABSTRACT

Forecast Evaluation with Shared Data Sets*

Data sharing is common practice in forecasting experiments in situations where fresh data samples are difficult or expensive to generate. This means that forecasters often analyse the same data set using a host of different models and sets of explanatory variables. This practice introduces statistical dependencies across forecasting studies that can severely distort statistical inference. Here we examine a new and inexpensive recursive bootstrap procedure that allows forecasters to account explicitly for these dependencies. The procedure allows forecasters to merge empirical evidence and draw inference in the light of previously accumulated results. In an empirical example, we merge results from predictions of daily stock prices based on (1) technical trading rules and (2) calendar rules, demonstrating both the significance of problems arising from data sharing and the simplicity of accounting for data sharing using these new methods.

JEL Classification: C10
Keywords: bootstrap, calendar effects, data sharing, forecast evaluation and technical trading

Ryan Sullivan
Economic Analysis LLC
2049 Century Park East
Suite 2310
La Jolla
Los Angeles, CA 90067
USA
Tel: (1 310) 556 0709
Fax: (1 310) 556 0766
Email: rmsullivan@econanalysis.com

Allan G Timmermann
UCSD Department of Economics
9500 Gilman Drive
La Jolla
CA 92093-0508
USA
Tel: (1 858) 534 4860
Fax: (1 858) 534 7040
Email: atimmerm@weber.ucsd.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new~dps/dplist.asp?authorid=144119

For further Discussion Papers by this author see:
www.cepr.org/pubs/new~dps/dplist.asp?authorid=116464

Halbert White
Quantmetrics Corporation
6540 Lusk Boulevard, c-157
San Diego, CA 92121
USA
Tel: (1 888) 777 7226
Fax: (1 619) 457 2473
Email: halwhite@earthlink.net

For further Discussion Papers by this author see:
www.cepr.org/pubs/new~dps/dplist.asp?authorid=102814

* The authors acknowledge research support provided by QuantMetrics R&D Associates, LLC. The methods described in this paper are covered by US Patent 5,893,069.

Submitted 24 October 2001

I. Introduction

Data sharing is common practice in forecasting experiments. This practice is often inevitable. There is only one time series for the US gross domestic product and only one history of stock market prices for US firms. Investigators intending to predict a given economic time series will typically not be the first or only persons to have looked at the data. Studies may generate conflicting results although they use the same data set. Such an outcome could be a result of the use of different functional forms, explanatory variables or methodology, and one research design may not entirely dominate others. In this situation it is important to have available a procedure that allows outsiders to merge the reported forecasting results and draw conclusions across studies.

Closely related to the problems arising from data sharing are biases in statistical inference arising from data mining. Data mining is the practice of re-using a given set of data for purposes of inference or model selection. It is widely regarded as an important procedure for extracting information from data sets in the engineering, physical and biological sciences. The reason for this is obvious: important regularities in available data that may not have been predicted by theory or appear as a result of a simple preliminary statistical analysis may nevertheless emerge from systematic analysis of the data. However, Chatfield (1995) demonstrates forcefully that if the same data set is used to formulate, estimate and test a model, serious biases in inference are likely to result. As increasingly advanced statistical fitting procedures are brought to bear on particular data sets, the dangers of over-fitting induced by data mining become ever greater and a procedure that controls for the resulting biases is urgently needed.

If forecasters subsequently use fresh data samples to test propositions based on close inspection of an initial data sample, then the standard assumptions underlying statistical inference need not be violated. This ideal situation exists whenever new data can easily be generated or the initial data set is sufficiently large that a hold-out sample can be reserved for subsequent prediction analysis. However, in the common situation where new data are unavailable and the same data are used to form an hypothesis and test it, then standard assumptions underlying classical statistical inference will be violated. Moreover, if sufficiently many models are fitted to a finite data sample, then by pure chance some of these models are bound to detect patterns, even if these are truly spurious. This phenomenon has been

well documented in the literature on subset selection, c.f. Miller (1990).

Distortions in the research community's inference as a result of data mining is widely recognized. Economics Nobel Laureate Robert Merton puts it this way: "Is it reasonable to use the standard t-statistic as a valid measure of significance when the test is conducted on the same data used by many earlier studies whose results influenced the choice of theory to be tested?" (Merton (1987) page 107).

The plan for the paper is as follows. Section II describes the existing procedures for handling data mining. Section III outlines the distributional theory underlying our proposal and compares alternative Monte Carlo and bootstrap procedures. Section IV provides an empirical forecasting application and Section V concludes.

II. Existing Procedures for Dealing with Data-snooping Biases

Because it constitutes such a widespread problem, a variety of procedures have been developed to account for the biases in statistical inference resulting from data sharing and data mining in forecast evaluation. We describe some of the more common of these practices in turn and explain their potential shortcomings.

Report the full set of models investigated in a particular study. This would seem to be the best research strategy available to researchers who want to account for possible contamination of their analysis due to the inspection of a large number of competing models. Nevertheless, even if a researcher only entertains a small number of models and honestly reports the full set of test results, he is still likely to be strongly influenced by his peers, who may previously have inspected the same or closely related data sets. This is a real problem as science often progresses by cumulating evidence from earlier studies conducted by different groups of researchers, each group building on previously reported work. Even if a particular researcher controls for his own data mining, it may be impossible to control for other researchers' efforts. This contaminates the data for purposes of hypothesis testing. How important this source of contamination is depends on how many models previous researchers have studied (c.f. Lo & MacKinlay (1990)), the correlation between the combined set of (new and old) models, the sample size and the stochastic structure of the shared data set. Unless all these effects are acounted for and quantified, correct inference becomes extremely difficult.

Bonferroni bounds provide a statistical procedure for bounding the probability that the best available model does not outperform the benchmark against which it

is compared. Let p_i be the probability (p-value) associated with the null hypothesis that the ith model does not outperform the benchmark and suppose that l forecasting models have been considered by the research community. Further suppose that interest lies in testing the joint null hypothesis that none of the models is superior to some given benchmark. The Bonferroni bound simply states that, for arbitrary correlations between the performance statistics used to evaluate the models, the p-value for the joint test is given by

$$p \leq Min\left(l \bullet Min(p_{1,}, ..., p_l), 1\right). \tag{1}$$

Hence the procedure scales the smallest p-value, representing the strongest evidence against the null hypothesis, by the number of models under consideration to obtain an upper bound on the probability that the best model does not reject the null. Unfortunately, this practice can be extremely conservative, particularly when there are strong positive correlations between performance measures as is often the case. In applications where many studies have been conducted and l is consequently very large, the bound simply states that p is less than one.

Wait for new data to become available. While this is sometimes feasible, it also can be a very costly and slow procedure. For example, economic theory may predict dynamic structures in economic variables at the business cycle frequency. With only eight post-war recessions, researchers would have to wait for an extraordinarily long period of time before getting sufficient new data on which to test the validity of such hypotheses. Furthermore, even if time is not of the essence, structural breaks or regime switches of the type modeled by Hamilton (1989) could render new data useless for purposes of testing the original effect.

Use similar data from other sources. This is a common practice in the social sciences. For example, suppose that a researcher has discovered that a certain variable predicts US stock prices. In the absence of new data from the US, the researcher may use international data from other stock markets to see if the finding holds only for US stock prices or holds more generally in other markets. Presence of the pattern in other markets is then considered strong corroborative evidence for the hypothesis, while absence of the pattern is interpreted as evidence against the hypothesis. There are two problems with this approach. First, the time-series of the dependent and explanatory variables are often strongly correlated and are far from representing independent samples. Second, because institutional structures

differ across markets, the hypothesized effect could well be present in the US data but not in other markets. In this case failure to find evidence of the effect from other markets does not necessarily lead to a revision in the $p-$value computed from the US study.

III. Accounting for Dependencies Across Forecasting Studies

We argued in the introduction that dependencies across studies potentially contaminate statistical inference about forecasting performance and in the previous section we suggested that common current practices to deal with this contamination are not satistisfactory. In this section we set up a classical statistical framework for quantifying such dependencies and discuss procedures that can quantify the biases. Our interest lies in conducting inference using many models on the *same* data set and thus is distinct from meta analysis which is concerned with the combination of data from multiple sources, cf. Mosteller & Chalmers (1992). The principle is to account for dependencies across models by evaluating the probability distribution of the performance measure of interest in the context of the full universe of models leading to the best-performing model. This is the logical way to handle data-mining distortions induced by conducting statistical inference *after* (and hence conditional on) model selection. By considering the full set of models entertained by the community of researchers we effectively undo the conditioning on the model selection. A comparison of the distribution of the performance statistic of the best model when standing alone (and hence conditional on the model selection process) to the distribution of the performance statistic in the context of the full universe of models quantifies the size of the model selection bias.

Here we evaluate model performance by predictive accuracy but we emphasize that the approach is readily applicable in other contexts. Suppose then that each model produces a sequence of forecasts that depend on a set of predictor variables and recursively updated parameter estimates. Suppose further that l models of a time series process have been (jointly) considered by the research community. To account for dependencies across models, the test procedure considers the distribution of the lx1 performance statistic

$$\overline{f} = n^{-1} \sum_{t=R}^{T} \widehat{f}_{t+1}, \tag{2}$$

where n is the number of prediction periods indexed from R through T so that $T = R + n - 1$, $\widehat{f}_{t+1} = f(Z_{t+1}, \widehat{\beta}_t)$ is the observed performance measure for period $t+1$, and $\widehat{\beta}_t$ is a vector of recursively updated parameter estimates. The first R observations are used to obtain the initial estimate, $\widehat{\beta}_R$. The estimation window is then expanded such that $\widehat{\beta}_{R+1}$ is estimated from the first $R+1$ observations, and so on. Z_{t+1} is a vector of variables containing both target and predictor variables. Target variables are observed for the first time in period $t+1$, whereas predictor variables are available at time t. The setup is assumed to satisfy the conditions of Diebold & Mariano (1995) or West (1996). The elements $f_{k,t+1}$ of f_{t+1} measure the performance of the individual models $k = 1, ..., l$ relative to some given benchmark. The relevant null hypothesis is that the best model is not capable of outperforming the benchmark:

$$H_0 : \max_{k=1,...l} \{E(f_k^*)\} \preceq 0, \tag{3}$$

where $E[f_k^*] \equiv E[f_k(Z_t, \beta^*)]$, $\beta^* = p\lim(\widehat{\beta}_t)$.

The seminal papers by Diebold & Mariano (1995) and West (1996) develop methods for assessing the out-of-sample predictions generated by economic models. Corradi, Swanson & Olivetti (2001) compare the assumptions underlying these two papers, the key being that West (1996) accounts for parameter estimation errors. Building on this work, White (2000) shows that, under the element of H_0 least favorable to the alternative,

$$\max_{k=1,..l} n^{1/2} \overline{f}_k \xrightarrow{d} \max_{k=1,...,l} \{\Psi_k\}, \tag{4}$$

where $\Psi \sim N(0, \Omega)$ is an lx1 multivariate normally distributed random vector with elements Ψ_k having a specific covariance matrix Ω, and $\xrightarrow{d}$ denotes convergence in distribution. This result is intuitively satisfying. If only a single model is considered, then by standard results its performance measure would follow an asymptotic normal distribution. Suppose now that the best model has been selected from a universe of l candidate models, all of which may be correlated. Then its p-value should be evaluated based on the maximum value drawn from an l-dimensional normal distribution with mean zero and a covariance matrix reflecting the correlations across the included models' performances.

A. Monte Carlo Methods

The distribution of $\max_{k=1,...l}\{\Psi_k\}$ for general Ω is analytically intractable except when $l = 1$ or $l = 2$ as White (2000) discusses. Hence one has to resort to numerical evaluation of this statistic. One possibility is to use Monte Carlo simulation methods. These can proceed as follows. First consider the case where $l = 1$. Here the problem is simply to find the distribution of

$$\Psi_1 = \max \Psi_1 \sim N(0, \sigma_1^2). \tag{5}$$

In this case Monte Carlo methods are unnecessary (although easily carried out) because standardizing yields a statistic $n^{1/2}\overline{f}/\sigma_1$ that is standard normal. σ_1^2 is unknown but can be replaced by a consistent estimator that converges in probability to σ_1^2 : $\widehat{\sigma}_1^2 \xrightarrow{p} \sigma_1^2$. Estimation of $\widehat{\sigma}_1^2$ is complicated by its dependency on the full set of autocovariances of the underlying performance measure and requires estimation of a spectral density at frequency zero. Nevertheless, there are a variety of procedures appropriate for doing this. For example, we can use the stationary bootstrap estimates of Politis & Romano (1994) or the HAC estimates of Newey & West (1987).

Next consider the case in which $l = 2$. This case requires computation of the distribution of $\max_{k=1,2}\{\Psi_k\}$. The $2x1$ vector $\Psi \sim N(0, \Omega)$ has covariance matrix

$$\Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

and estimates $\widehat{\sigma}_1^2$, $\widehat{\sigma}_{12}$, and $\widehat{\sigma}_2^2$ must be obtained. To estimate these consistently, we need to have available $\widehat{f}_{k,t}$ and $k = 1, 2$, $t = 1, ..., T$. Once the estimates have been obtained, the Monte Carlo proceeds as follows. Draw

$$\begin{aligned} &\Psi_{1i} \sim N(0, \widehat{\sigma}_1^2) \\ &\Psi_{2i} = \widehat{\beta}_{21}\Psi_{1i} + \widehat{\gamma}_2\varepsilon_{2i}, \quad \varepsilon_{2i} \sim iidN(0, 1) \end{aligned} \tag{6}$$

where $\widehat{\beta}_{21}$, $\widehat{\gamma}_2$ are chosen such that $E[\Psi_{2i}^2] = \widehat{\sigma}_2^2$ and $E[\Psi_{1i}\Psi_{2i}] = \widehat{\sigma}_{12}$. This implies that $\widehat{\beta}_{21} = \widehat{\sigma}_{12}/\widehat{\sigma}_1^2$ and $\widehat{\gamma}_2^2 = \widehat{\sigma}_2^2 - \widehat{\sigma}_{21}(\widehat{\sigma}_1^2)^{-1}\widehat{\sigma}_{12}$. From the simulated series we can evaluate the performance measure $\max\{\Psi_{1i}, \Psi_{2i}\}$. Repeating this procedure B times yields a histogram for the maximum of the mean performance measure against which the mean of the best-performing model from the actual sample can be compared.

In the general case with an arbitrarily large value of l, we have to evaluate the distribution of $\max_{k=1,..,l}\{\Psi_k\}$. For the lth term this amounts to estimating σ_{l1}, σ_{l2}, ..., σ_l^2 and then simulating Ψ_{li} from the equation

$$\Psi_{li} = \widehat{\beta}_{l1}\Psi_{li} + \widehat{\beta}_{l2}\Psi_{2i} + ... + \widehat{\beta}_{l,l-1}\Psi_{l-1,i} + \widehat{\gamma}_l\varepsilon_{li}, \tag{7}$$

for suitable choices of $\widehat{\beta}_{lj}$ and $\widehat{\gamma}_l$. Again a histogram for $\max\{\Psi_{1i}, \Psi_{2i}, ..., \Psi_{li}\}$ can be formed using B separate Monte Carlo simulations, and this can be compared to the actual performance of the best model to form a p-value.

We conclude from this description of the Monte Carlo procedure that, in order to evaluate the data-snooping bias, first, one would need to keep all $l \cdot n$ values of $\widehat{f}_{k,t}$ in memory. Secondly, one would need to estimate $l(l+1)/2$ covariance terms $\widehat{\sigma}_{jk}$ explicitly. Third, one would effectively have to perform regressions based on $\widehat{\sigma}_{jk}$ in order to get the estimates $\widehat{\beta}_{jk}$ used in the simulations. Finally, every time an extra model is added to the set, one would have to either keep all previous Monte Carlo draws in memory or redraw. See White (2000) for further discussion.

B. The Bootstrap Methods

As White (2000) shows, an alternative to the Monte Carlo approach is to simulate the needed distribution by means of the stationary bootstrap of Politis & Romano (1994) applied to the observed values of $\widehat{f}_{k,t}$. Resampling the performance measures from the forecasting rules yields B bootstrapped values of $\overline{f}_k$, which we denote by $\overline{f}^*_{k,i}$, where i indexes the B bootstrap samples. Then we can construct the following statistics:

$$\overline{V}_l = \max_{k=1,...,l}\{\sqrt{n}\,\overline{f}_k\}, \tag{8}$$

$$\overline{V}^*_{l,i} = \max_{k=1,...,l}\{\sqrt{n}(\overline{f}^*_{k,i} - \overline{f}_k)\}, i = 1, ..., B. \tag{9}$$

Comparing the performance measure from the actual data ($\overline{V}_l$) to the quantiles from the bootstrap experiment ($\overline{V}^*_{l,i}$) one obtains White's bootstrap reality check p-value for the null hypothesis. Using the maximum value of the performance measure across all l trading rules ensures that the effects of data-mining are accounted for.

The sample statistic can also be based on functions g of sample moments $\overline{h}_k$, $k = 0, ..., l$:

$$f_k = g(\overline{h}_k) - g(\overline{h}_0), \tag{10}$$

where $\overline{h}_k$ and $\overline{h}_0$ are averages computed over the prediction sample for the kth model and the benchmark, respectively:

$$\overline{h}_k = n^{-1} \sum_{t=R}^{T} h_{k,t+1}. \tag{11}$$

In this case the bootstrap procedure is applied to yield B bootstrapped values of $\overline{f}_k$, denoted as $\overline{f}^*_{k,i}$, where

$$\overline{f}^*_{k,i} = g(\overline{h}^*_{k,i}) - g(\overline{h}^*_{0,i}), i = 1, ..., B, \tag{12}$$

$$\overline{h}^*_{k,i} = n^{-1} \sum_{t=R}^{T} h^*_{k,t+1,i}, i = 1, ..B. \tag{13}$$

The distribution needed to evaluate the p-value is that of $n^{1/2}(\overline{f} - E[f^*])$. The bootstrap computes $n^{1/2}(\overline{f}^* - \overline{f})$, where $\overline{f}^* = n^{-1} \sum_{t=R}^{T} f^*_{\theta_t}$, and $t = R, ..., T$. The random indexes θ_t are chosen from $R, ..., T$ according to the stationary bootstrap of Politis & Romano (1994):

for $t = R$, θ_t is drawn uniformly on [0,1]
for $t > R$, U is drawn uniformly on [0,1]
if $U > q$ pick θ_t uniformly on $\{R, ..., T\}$
if $U \preceq q$ pick $\theta_t = \theta_{t-1} + 1$, and reset to R if $\theta_t = T + 1$,

where q is a smoothing parameter chosen to accomodate dependence in the data. White's Theorem 2.3, which we reproduce below, states that the distribution of the performance measure can be evaluated by means of the stationary bootstrap:

Theorem 2.3 (White (2000)): Suppose West (1996)'s conditions and Politis & Romano (1994)'s conditions hold for the elements of $f(Z_t, \beta^*)$. Also suppose that, for all $k-$vectors λ, $\lambda'\lambda = 1$

$$\Pr(\lim sup_t \frac{t^{1/2}|\lambda'(\widehat{\beta}_t - \beta^*)|}{(\lambda' G \lambda \log(\log(\lambda' G \lambda) t))^{1/2}} = 1) = 1$$

where $\widehat{\beta}_t$ is such that $t^{1/2}(\widehat{\beta}_t - \beta^*) \xrightarrow{d} N(0, G)$. Let $F = 0$ or $(n/R)\log(\log(R)) \to 0$. Then

$$\rho(\pounds(n^{1/2}(\overline{f}^* - \overline{f})|Z_1, ..., Z_{T+1}), \pounds(n^{1/2}(\overline{f} - E(f^*))|Z_1, ..., Z_{T+1})) \xrightarrow{p} 0,$$

where £ denotes the probability law of its argument and ρ is any metric metrizing convergence in distribution.

An immediate consequence of this result, which justifies using the bootstrap procedure, is that

$$\rho(\pounds(\max_{k=1,...,l} n^{1/2}(\overline{f}_k^* - \overline{f}_k)|Z_1, ..., Z_{T+1}), \pounds(\max_{k=1,...,l} n^{1/2}(\overline{f}_k - E(\overline{f}_k^*))|Z_1, ..., Z_{T+1})) \xrightarrow{p} 0. \tag{14}$$

Implementation of the bootstrap is simple and involves comparing $\overline{V}_l = \max_{k=1,...l} \{n^{1/2}\overline{f}_k\}$ to the distribution of $\overline{V}_l^* = \max_{k=1,...,l}\{n^{1/2}(\overline{f}_k^* - \overline{f}_k)\}$. Here $\overline{f}_k^*$ is the kth component of $\overline{f}$ obtained by resampling $\widehat{f}_{k,t+1}$ according to the stationary bootstrap. A very attractive property of the bootstrap is its recursive structure, which greatly economizes on the information required to update the p-value for the null hypothesis that no model outperforms the benchmark. This can be seen from the expression

$$\max_{k=1,...,l} n^{1/2}(\overline{f}_{k,i}^* - \overline{f}_k) = \max(n^{1/2}(\overline{f}_{l,i}^* - \overline{f}_l), \max_{k=1,...,l-1} n^{1/2}(\overline{f}_{k,i}^* - \overline{f}_k)), i = 1, .., B. \tag{15}$$

For a given bootstrap iteration (i) the p-value for the null hypothesis can be updated recursively using a simple spreadsheet:

value of k	performance statistic	compare to the distribution of
$k = 1$	$\overline{V}_1 = n^{1/2}\overline{f}_1$	$\overline{V}_1^* = n^{1/2}(\overline{f}_1^* - \overline{f}_1) \xrightarrow{d} N(0, \Omega_{11})$
$k = 2$	$\overline{V}_2 = \max(n^{1/2}\overline{f}_2, \overline{V}_1)$	$\overline{V}_{2,i}^* = \max(n^{1/2}(\overline{f}_{2,i}^* - \overline{f}_2), \overline{V}_{1,i}^*)$
...	...	...
$k = l$	$\overline{V}_l = \max(n^{1/2}\overline{f}_l, \overline{V}_{l-1})$	$\overline{V}_{l,i}^* = \max(n^{1/2}(\overline{f}_{l,i}^* - \overline{f}_l), \overline{V}_{l-1,i}^*)$

An important consequence of this economy of information is that it facilitates the communication of research results across forecasting experiments. When additional models are fitted to the same data set, updating the bootstrap simply requires a knowledge of the histogram of $\max_{k=1,...l-1} n^{1/2}(\overline{f}_{k,i}^* - \overline{f}_k)$ $i = 1, ..., B$. The recursions (15) make it clear that to continue a specification search using the bootstrap,

it suffices to know V_{l-1}, $V^*_{l-1,i}$ and the indexes θ_{it}, $i = 1, .., B$. For the latter, knowledge of R, T, q, B, the random number generator (RNG) and the (RNG) seed suffice. Knowing or storing $\widehat{f}_{k,t+1}$ is unnecessary, nor do we need to compute or store $\widehat{\Omega}$ or the previous Monte Carlo draws. This demonstrates not only a significant computational advantage for the bootstrap method over the Monte Carlo version but also the possibility for researchers at different locations or at different times to further understanding of the phenomenon modeled without needing to know the specifications tested by their collaborators or competitors. Some cooperation is nevertheless required, as R, T, q, B, the RNG, the RNG seed and V_{l-1}, $V^*_{l-1,i}$, $i = 1, .., B$ must still be shared, along with the data and the specification and estimation method for the benchmark model.

IV. Empirical Application: Evaluating the Efficient Market Hypothesis using Daily Stock Price Data

To demonstrate how easily data can be shared for purposes of conducting inference that accounts for dependencies across studies, we demonstrate in this section how to merge two separate forecasting experiments applied to daily stock prices on the Dow Jones Industrial Average and reported in more detail in Sullivan, Timmermann & White (1999) and Sullivan, Timmermann & White (2001). The sample spans the period January 1, 1897 to June 30, 1998 and has 27,447 observations. To the best of our knowledge, this is the first time a statistical experiment of this nature has been conducted.

The first experiment considers forecasting daily stock prices by means of technical trading rules. These rules attempt to discover repeated patterns in stock prices, for example by comparing short-term movements to long-term trends. We consider 497 filter rules, 2049 moving averages, 1220 support and resistance rules, 2040 channel break-out rules and 2040 on-balance volume average rules. Each rule uses a different parameterization and the total number of technical trading rules, which we refer to as l_1, is 7846.

We also consider separately the extent to which there are calendar effects in stock returns. An extensive literature in finance reports evidence of day-of-the-week, week-of-the-month, month-of-the-year and holiday effects in mean returns on stocks. See Lakonishok & Smidt (1988) for a summary of the literature. The best known calendar effect is probably the Monday effect: stocks have been found

to have lower mean returns on Mondays than on other days of the week. We consider a total of $l_2 = 9452$ calendar rules, all of which reasonably could have been explored by an investor in search of a calendar pattern. These include 60 day of the week rules, 60 week of the month rules, 8,188 month of the year rules, 100 semi-month rules, 8 holiday rules, 12 end of December rules, and 1024 turn of the month rules. The combined experiment is comprised of 7846 + 9452 = 17,298 ($l_1 + l_2 = l$) trading rules. Since the prediction methods used in both studies do not require estimation of any parameters, the parameterizations of the resulting decision rules ($\beta_k,\ k = 1, ..., l$) directly generate returns that are then used to measure performance.

Performance is measured using the continuously compounded returns on a forecasting rule relative to returns from following a passive strategy of always being in the market:

$$f_{k,t+1} = \ln(1 + y_{t+1} S_k(\chi_t, \beta_k)) - \ln(1 + y_{t+1} S_0(\chi_t, \beta_0)), k = 1, ..., l \qquad (16)$$

where χ_t is the information that forms the basis for deciding the position to hold from the end of period t to the end of period $t + 1$, $y_{t+1} = (X_{t+1} - X_t)/X_t$ is the holding period return, and X_t is the stock price series under investigation. S_k and S_0 are signal functions that convert χ_t into trading positions for given system parameters β_k. We consider the following range of signal functions: 1 represents a long position, 0 a neutral position (i.e., being out of the market), and -1 a short position (selling the asset for future delivery).

The efficient market hypothesis is usually taken to imply that there does not exist a trading system that can outperform the market index. Hence the relevant null hypothesis is that there does not exist a forecasting rule that generates performance superior to always being in the market:

$$H_0 : \max_{k=1,...l} \{E(f_k)\} \preceq 0. \qquad (17)$$

Rejection of this null hypothesis implies that the best model is capable of outperforming the benchmark according to the performance measures embodied in f. Although we follow the finance literature in measuring performance through mean returns, we notice that performance measures that account for risk, such as the Sharpe ratio, can easily be computed.

Table 1 shows that the best technical trading rule (mean return of 17.07%) appears to be capable of producing genuine outperformance relative to the benchmark of always being invested in the market portfolio (mean return of 4.80%). The p-value corrected for data mining is zero for practical purposes even when the search for the best technical trading rule across 7846 models is considered. In stark contrast there is no evidence that the best calendar model (mean return of 8.55%) outperforms the benchmark, once data-mining is taken into account within the set of calendar rules. Notice that the data-mining correction makes a big difference in this second case: When the best calendar rule (the Monday rule) is considered in isolation it appears to be highly significant with a p-value close to zero. However, when the data-mining effects are taken into consideration, the p-value rises to 0.29 and the first conclusion is completely overturned. Further details of these separate experiments are reported in Sullivan et al. (1999) and Sullivan et al. (2001).

Separate consideration of these studies thus leaves the issue of testing the efficient markets hypothesis unsettled. One study leads to a rejection of the null while the other study does not reject the null. Suppose that the calendar effect study was first conducted and showed no evidence of genuine predictability. Subsequently the technical trading experiment is conducted and shows evidence of predictability. Ultimately the purpose of the two studies is the same, namely to test the efficient market hypothesis. How should a researcher revise his beliefs about the validity of the (shared) null hypothesis that stock prices cannot be predicted? Inspection of the separate analyses alone cannot decide the issue. Clearly there is a need for a procedure that allows us to merge the results from the two studies.

In Table 2 we present the results from the bootstrap experiment that uses the same random number seed to combine the two studies. The total number of models under consideration is now $l = l_1 + l_2 = 17{,}298$. No study in economics has come even close to considering simultaneously such a large number of models. The results show that the mean return performance of the best technical trading rule in fact is so good over the full sample period that it continues to be significant at conventional critical levels even when considered in the full universe comprising both calendar rules and technical trading rules.

Figure 1 shows each model's mean return performance. Two separate lines track the performance of the best model updated across all models as the universe of models expands as well as the p-value adjusted for data-mining in this expanding

universe. The first 9452 models comprise the calendar rules while the subsequent 7846 models comprise the technical trading rules. Since the ordering of the models is arbitrary, only the terminal p-value ultimately matters. Nevertheless, the picture provides fascinating insights into data-mining effects. Among the set of calendar rules, the best-performing forecasting model is identified early on and since no improvement occurs within this subset of rules, the p-value slowly drifts upward, reflecting the effect of having drawn the best model from a distribution with a wider support. The picture changes dramatically once the technical trading rules are introduced. After approximately 9,700 models have been considered, a technical trading rule with a very significant outperformance reduces the bootstrap p-value to a number close to zero. Since the performance of this model is very strong, adding further technical trading rules does not lead to any visible increase in the bootstrap p-value.

The above results are valid for the lengthy sample January 1897 to June 1998. However, the quality of the data at the beginning of the sample is unlikely to be sufficiently high to represent prices at which investors could in fact have carried out transactions. The practical implications for market efficiency of these full-sample tests may thus be limited. To deal with this, we consider separately data on the very liquid S&P500 futures price index. Our sample starts one year after the contract started public trading, namely in 1984, and finishes in 1996. This recent sample is likely to provide a more realistic test of market efficiency. Results analogous to those just described are presented in Table 3. We see that there is no longer any evidence of predictable patterns in stock returns. Even the best model does not outperform the benchmark on a mean return basis. As a further confirmation of the lack of evidence against the null hypothesis, Figure 2 show that, as the universe of models expands, the p-value never goes below 0.60.

These results suggest the following. First, a scientist who ignores the effects of data-mining would be led to believe that both calendar rules and technical trading rules that outperform the market benchmark do indeed exist. However, our investigation suggests that this conclusion is premature. While there is evidence in the full historical sample from 1897 to 1998 that the bootstrap p-value is significant at standard critical levels, the more recent data sample is likely to produce the most reliable results since they are based on futures contracts that actually traded in public markets.

V. Conclusion

This paper has proposed a simple recursive bootstrap methodology that allows researchers to update their p-values of a shared null hypothesis once new forecasting models are fitted to an existing data set. Given that the data are shared along with the specification and the estimation method for the benchmark model, only (i) the start and end observation indexes for the best out-of-sample window (R and T); (ii) the number of bootstrap resamples (B) and the bootstrap smoothing parameter (q); (iii) the random number generator and its seed; and (iv) the best sample and resample performance values $\overline{V}_{l-1}$ and $\overline{V}^*_{l-1,i}$, $i = 1, ..., B$ need to be communicated between researchers. Remarkably, details of the specific models investigated do not need to be shared, so that scientific progress is possible using shared data without full disclosure. Thus, progress can be made even when details of earlier forecasting experiments became lost or when commercial interest dictates that specific forecasting models remain proprietary or confidential. The procedure is thus ideally suited to accumulating empirical evidence on issues of common interest to an entire community of reserachers separated in space and time.

References

Chatfield, C. (1995), 'Model uncertainty, data mining and statistical inference', *Journal of the Royal Statistical Society (A)* **158**, 419–466.

Corradi, V., Swanson, N. & Olivetti, C. (2001), 'Predictive ability with cointegrated variables', *Journal of Econometrics* **104**, 315–358.

Diebold, F. & Mariano, R. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–265.

Hamilton, J. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica* **57**, 357–384.

Lakonishok, J. & Smidt, S. (1988), 'Are seasonal anomalies real? a ninety-year perspective', *Review of Financial Studies* **1**, 403–425.

Lo, A. W. & MacKinlay, A. C. (1990), 'Data-snooping biases in tests of financial asset pricing models', *The Review of Financial Studies* **3**, 431–467.

Merton, R. (1987), 'On the state of the efficient market hypothesis in financial economics', *In R. Dornbusch, S. Fischer, and J. Bossons (eds.) Macroeconomics and Finance: Essays in Honor of Franco Modigliani. MIT Press, Cambridge, Mass.* pp. 93–124.

Miller, A. (1990), 'Subset selection in regression', *Chapman and Hall. London* .

Mosteller, F. & Chalmers, T. (1992), 'Some progress and problems in meta-analysis of clinical trials', *Statistical Science* **7**, 227–236.

Newey, W. & West, K. (1987), 'A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**, 703–708.

Politis, D. & Romano, J. (1994), 'The stationary bootstrap', *Journal of the American Statistical Association* **89**, 1303–1313.

Sullivan, R., Timmermann, A. & White, H. (1999), 'Data-snooping, technical trading rule performance and the bootstrap', *Journal of Finance* **54**, 1647–1692.

Sullivan, R., Timmermann, A. & White, H. (2001), 'Dangers of data-mining: The case of calendar effects in stock returns', *Journal of Econometrics (forthcoming)* .

West, K. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **64**, 1067–1084.

White, H. (2000), 'A reality check for data snooping', *Econometrica* **68**, 1097–1126.

TABLE 1

Dow Jones Industrial Average: Full Sample (January 1897 – June 1998)
Calendar Frequency and Technical Trading Rules

Calendar Frequency Trading rules (9,452 rules)	
	Mean Return Criterion
Benchmark	4.80%
Best Rule	Monday Effect (neutral on Mondays, long otherwise
Performance	8.55%
Nominal *p*-value	0.00
White's *p*-value	0.29
Technical Trading Rules (7,846 rules)	
	Mean Return Criterion
Benchmark	4.80%
Best rule	2-day-on-balance volume
Performance	17.07%
Nominal *p*-value	0.00
White's *p*-value	0.00

The performance on the full sample of the Dow Junes Industrial Average, according to the mean return criterion, is provided for two sets of rules: (1) the calendar frequency trading rules, and (2) the technical trading rules. The performance measure is displayed for the benchmark (buy-and-hold) trading strategy and the best performing rule, along with the nominal and data-snooping adjusted *p*-values for the best performing rule. The type of rule that exhibits the best performance is also described.

TABLE 2

Dow Jones Industrial Average: Full Sample (January 1897 – June 1998)
Combined Universe of Trading Rules

Combined Universe of Trading Rules (17,298 rules)	
	Mean Return Criterion
Benchmark	4.80%
Best Rule	2-day-on-balance volume
Performance	17.07%
Nominal *p*-value	0.00
White's *p*-value	0.00

The performance on the full sample of the Dow Jones Industrial Average, according to the mean return criterion, is provided for the combined universe of trading rules which includes both calendar frequency trading rules and technical trading rules. The performance measure is displayed for the benchmark (buy-and-hold) trading strategy and the best performing rule, along with the nominal and data-snooping adjusted *p*-values for the best performing rule. The type of rule that exhibits the best performance is also described.

TABLE 3

DJIA and S&P Futures: Out-of-Sample
Combined Universe of Trading Rules

DJIA, 1987-1996: Combined Universe of Trading Rules (17,298 rules)	
	Mean Return Criterion
Benchmark	13.55%
Best Rule	Week of the Month (1,2,3,4,5 = 1,1,1,0,1)
Performance	17.27%
Nominal *p*-value	0.10
White's *p*-value	0.98
S&P 500, 1984-1996: Combined Universe of Trading Rules (17,298 rules)	
	Mean Return Criterion
Benchmark	8.01%
Best Rule	Week of the Month (1,2,3,4,5 = 1,1,1,0,1)
Performance	10.69%
Nominal *p*-value	0.22
White's *p*-value	0.99

The performance on both the out-of-sample Dow Jones Industrial Average and the S&P 500 futures, according to the mean return criterion, is provided for the combined universe of trading rules which includes both calendar frequency trading rules and technical trading rules. The performance measure is displayed for the benchmark (buy-and hold) trading strategy and the best performing rule, along with the nominal and data-snooping adjusted *p*-values for the best performing rule. The type of rule that exhibits the best performance is also described.

Figure 1: Mean Return Criterion

Dow Jones Industrial Average (1897 - 1998)

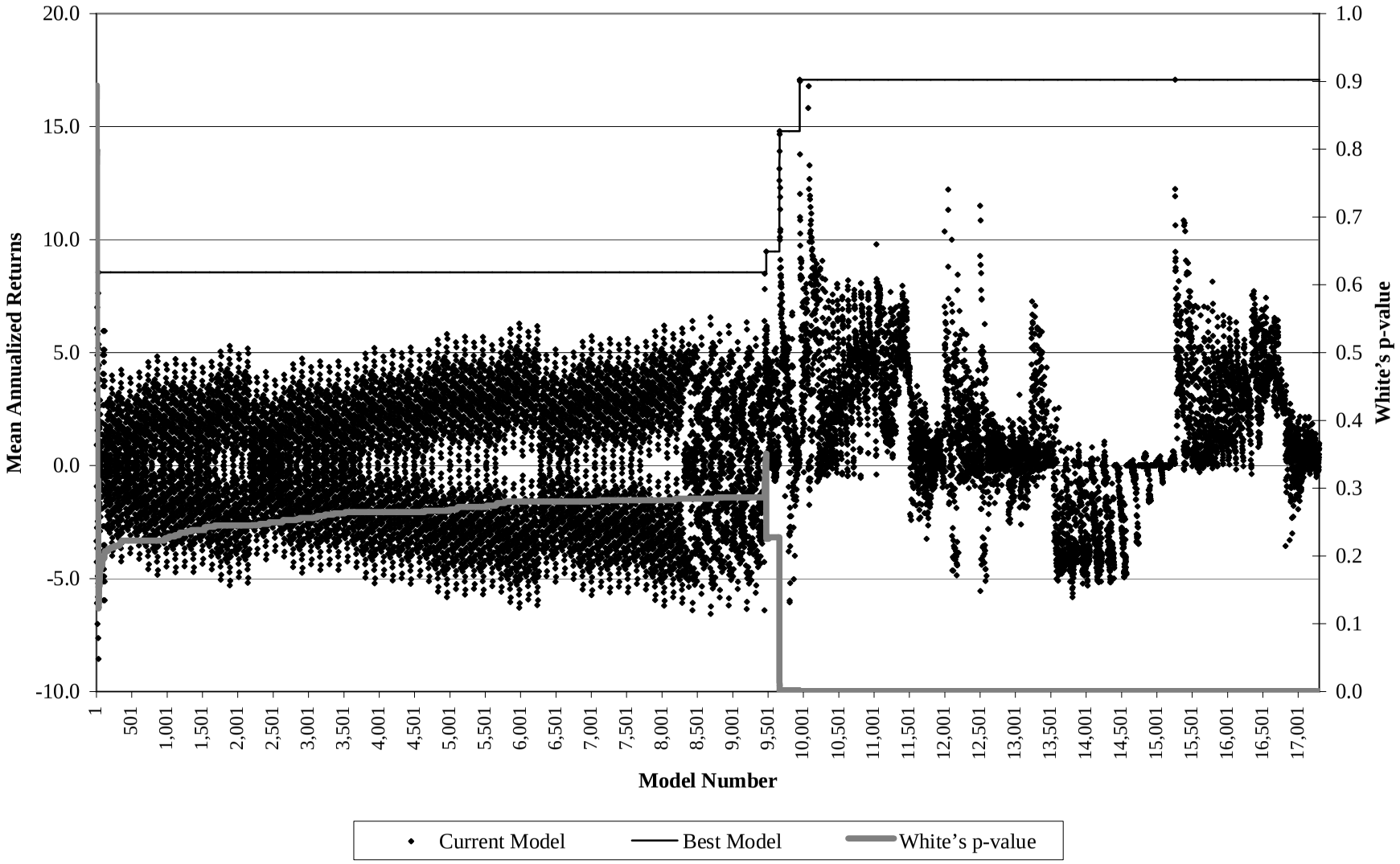

Figure 2: Mean Return Criterion

S&P 500 Futures (1984 - 1996)