ELSEVIER Journal of Econometrics 105 (2001) 249–286

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Dangers of data mining: The case of calendar effects in stock returns

Ryan Sullivan[a], Allan Timmermann[b,*], Halbert White[b]

[a]Bates White & Ballentine, LLC, 2001 K Street, NW, 744 Floor, Washington, DC 20005, USA

[b]Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, USA

Abstract

Economics is primarily a non-experimental science. Typically, we cannot generate new data sets on which to test hypotheses independently of the data that may have led to a particular theory. The common practice of using the same data set to formulate and test hypotheses introduces data-mining biases that, if not accounted for, invalidate the assumptions underlying classical statistical inference. A striking example of a data-driven discovery is the presence of calendar effects in stock returns. There appears to be very substantial evidence of systematic abnormal stock returns related to the day of the week, the week of the month, the month of the year, the turn of the month, holidays, and so forth. However, this evidence has largely been considered without accounting for the intensive search preceding it. In this paper we use 100 years of daily data and a new bootstrap procedure that allows us to explicitly measure the distortions in statistical inference induced by data mining. We find that although nominal p-values for individual calendar rules are extremely significant, once evaluated in the context of the full universe from which such rules were drawn, calendar effects no longer remain significant. © 2001 Elsevier Science S.A. All rights reserved.

JEL classification: C120; C530; G140

Keywords: Data mining; Market efficiency; Bootstrap testing; Calendar effects

[*] Corresponding author. Tel.: +1-858-534-4860; fax: +1-858-534-7040.
E-mail address: atimmerm@weber.ucsd.edu (A. Timmermann).

> *October. This is one of the peculiarly dangerous months to speculate in stocks in. The others are July, January, September, April, November, May, March, June, December, August and February.*
>
> Mark Twain (1894)

1. Introduction

Economic theory often is vague about the relationship between economic variables. As a result, many economic relations have been initially established from apparent empirical regularities and had not been predicted ex ante by theory. Like many of the social sciences, economics predominantly studies non-experimental data and thus does not have the advantage of being able to test hypotheses independently of the data that gave rise to them in the first instance. If not accounted for, this practice, referred to as data mining, can generate serious biases in statistical inference.[1] In the limited sample sizes typically encountered in economic studies, systematic patterns and apparently significant relations are bound to occur if the data are analyzed with sufficient intensity.

One of the most striking examples of a data-driven finding that was not anticipated by theory is the apparently very strong evidence of seasonal regularities in stock returns. Calendar effects were the first to be analyzed in the "Anomalies" section of the inaugural issue of *Journal of Economic Perspectives* (Thaler, 1987a, b). Indeed, theoretical considerations would suggest that researchers should not even be looking for such patterns in the first instance. According to standard economic theory, stock prices should follow a martingale process and returns should not exhibit systematic patterns, thus ruling out seasonal components unless these can be related to systematic variations in risk premiums, cf. Samuelson (1965), Leroy (1973), and Lucas (1978).

As reflected in the initial Mark Twain quote, investors have nevertheless long been fascinated by the possibility of finding systematic patterns in stock prices that, once detected, promise easy profits when exploited by simple trading rules. Moreover, Merton (1987) points out that "economists place a premium on the discovery of puzzles, which in the context at hand amounts to finding apparent rejections of a widely accepted theory of stock market behavior" (p. 104). Consequently, there is a long tradition among investors and academics of searching through stock market data; published academic studies on calendar effects go back to at least the early 1930s, e.g. Fields (1931,1934). As a result, common stock market indexes such as the Dow Jones Industrial

[1] Leamer (1978) discusses such pretest biases in considerable detail.

Average and the Standard&Poor's (S&P) 500 Index are among the most heavily investigated data sets in the social sciences.

Lo and MacKinlay (1988) point out that the degree of data-mining bias in a given field can be expected to increase with the number of studies published on the topic. Since so many academics and investors have looked at common US stock price indexes in an attempt to detect regularities, the performance of the best calendar rules cannot be viewed in isolation. Data with important outliers, such as those observed in stock market returns, are particularly prone to data-mining biases. If enough economic models are studied, by pure chance some of them are likely to outperform a given benchmark by any economic or statistical criterion. For example, models that implied investors should have been short in the stock market during October 19, 1987 are likely to outperform the market index in a longer sample simply because of the paramount significance of this single observation.

As a result of these endeavors, there is now a very large literature reporting apparent "anomalies" in stock returns, see, e.g., Dimson (1988). Grouped by calendar frequency, researchers have reported evidence of abnormal returns related to *day of the week* effects (Ball and Bowers, 1988; Cross, 1973; Fields, 1931; French, 1980; Gibbons and Hess, 1981; Jaffe and Westerfield, 1985; Keim and Stambaugh, 1984; Lakonishok and Levi, 1982; Rogalski, 1984), *week of the month* effects (Ariel, 1987; Lakonishok and Smidt, 1988), *month of the year* effects (Haugen and Lakonishok, 1988; Keim, 1983; Roll, 1983; Rozeff and Kinney, 1976), *turn of the month* effects (Ariel, 1987; Hensel and Ziemba, 1996; Lakonishok and Smidt, 1988), *turn of the year* effects (Haugen and Lakonishok, 1988; Jones et al., 1987; Lakonishok and Smidt, 1984; Ritter and Chopra, 1989; Roll, 1983) and *holiday* effects (Fields, 1934; Haugen and Lakonishok, 1988; Jones et al., 1987; Lakonishok and Smidt, 1988). Interestingly, none of these calendar effects were preceded by a theoretical model predicting their existence. This is an important consideration; surveying the philosophy of science literature, Campbell and Vinci (1983) write that "Philosophers of science generally agree that when observational evidence supports a theory the confirmation is much stronger when the evidence is novel". Similarly, Kahn et al. (1996) present a Bayesian model that formalizes the idea that empirical evidence giving rise to a new theory does not support the resulting theory as strongly as when the evidence had been predicted ex ante by the theory.

Thus, the findings of systematic seasonal patterns in stock returns leave us with a conundrum: do the apparent regularities in stock returns really imply a rejection of simple notions of market efficiency, or are they just a result of a large, collective data-mining exercise? Many researchers express awareness of this problem. Lakonishok and Smidt (1988), for example, comment on the seasonal regularities this way: "However, it is at least possible that these new facts are really chimeras, the product of sampling error and data mining".

In this paper we conduct an analysis that addresses these concerns. Based on a survey of the types of calendar rules that have been studied by researchers, we construct a universe of calendar trading rules using permutational arguments that do not bias us in favor of, or against, particular calendar effects. The universe contains nearly 9500 different calendar effects and the best calendar rule is evaluated in the context of this set. We do not imagine that this large set of calendar rules was inspected by any one individual investor or researcher. Rather, the search for calendar rules has operated sequentially across the investment community as a whole with the results of individual investors being reported gradually through the survival of the "fittest" calendar rules. Viewed this way, the number of rules we inspect does not appear to be unrealistically large.

We find that although many different calendar rules produce abnormal returns that are highly statistically significant when considered in isolation, once the dependencies operating across different calendar rules are accounted for, then the best calendar rule no longer achieves a p-value that is significant at conventional critical levels. This conclusion is robust to whether a mean return criterion or a risk-adjusted Sharpe ratio criterion is used in the assessment. Consistent with this finding, the performance of the calendar rule that was best in-sample actually generates inferior performance in an out-of-sample experiment using either cash or futures market prices.

To alleviate the concern that even genuinely significant calendar rules may not appear to be significant if assessed jointly with a sufficiently large set of "irrelevant" rules, we also evaluate the best-known calendar effects in a much smaller universe of 244 calendar rules. Again we find that the apparent statistical significance of the best calendar effects is not robust to data-mining effects.

The paper proceeds as follows. Section 2 discusses why standard asset pricing theory precludes calendar effects and Section 3 explains the bootstrap procedure that we use in our analysis to account for the effects of data mining. Section 4 describes the empirical evidence supporting the existence of calendar effects, and Section 5 explains the design of our experiment. Sections 5 and 6 report the empirical results of our analysis, while Section 7 concludes.

2. Calendar effects and efficient markets

In a frictionless market with no arbitrage where agents agree on the possible, the efficient market theory implies the moment condition:

$$E_t[Q_{t+1}R_{t+1}] = 1, \tag{1}$$

where R_{t+1} is an asset's rate of return during period $t+1$, Q_{t+1} is a pricing kernel reflecting a representative agent's intertemporal rate of substitution between future $(t+1)$ and current (t) consumption, and E_t is the conditional expectation operator taken with respect to information set I_t. It follows from this equation that the conditionally expected return is

$$E_t[R_{t+1}] = \frac{1 - \mathrm{cov}_t(Q_{t+1}, R_{t+1})}{E_t[Q_{t+1}]}. \tag{2}$$

Thus, standard asset pricing models explain predictable patterns in expected returns through a time-varying conditional covariance between returns and the pricing kernel or through variations in the expected pricing kernel itself, cf. Campbell et al. (1997). Asset pricing models typically assume that the pricing kernel is some power function of aggregate consumption and naturally imply that expected asset returns should vary over the economic cycle. Although economic significance of trading rules must ultimately be evaluated in the context of an equilibrium model for asset prices, it seems difficult to explain a priori why expected returns should display calendar regularities. Unless calendar rules provide hedging opportunities for individuals' consumption plans and the marginal rate of substitution between future and current consumption is systematically different on, say, Tuesdays or during the fourth week of the month compared with other points in time, then there is no underlying economic reason to expect the presence of calendar effects in asset returns.

To test the implications of calendar effects for the efficient market hypothesis, we evaluate such effects through the predictive signals they imply when adopted in trading strategies. This is the right criterion to use when assessing the *economic* significance of calendar effects. Jensen (1978) states it this way: "A market is efficient with respect to information set θ_t if it is impossible to make economic profits by trading on the basis of information set θ_t" (p. 96). Hence, a calendar rule has economic value if it allows investors to exploit it in a simple trading rule that outperforms the benchmark on a mean return or risk-adjusted return basis, depending on the risk-adjustment procedure in place.[2]

3. The bootstrap snooper

Researchers have long been aware of the potential dangers of data-mining effects. For example, Merton (1987, p. 107) poses the question "Is it reasonable to use the standard t-statistic as a valid measure of significance

[2] Importantly, the information set at time t should be interpreted as including the best forecasting model and the best parameter estimates available at time t. This suggests that an assessment of the efficient market hypothesis must be based on the recursive prediction signals generated by the forecasting models under consideration. We adopt just such a procedure.

when the test is conducted on the same data used by many earlier studies whose results influenced the choice of theory to be tested?" In a parametric context, Foster et al. (1997) show convincing evidence that the practice of cross-validating the predictive power of standard instruments over security returns by separately considering the returns on portfolios from different industries or countries can be very misleading due to the high correlation between such portfolios' returns and the resulting dependence between their maximal R^2 statistics. Likewise, Savin (1984) and Lakonishok and Smidt (1988) point out that the size of the induced test following from a search for the largest possible t-statistic can be very different from its nominal value.

To deal with such problems, White (2000), building on work of Diebold and Mariano (1995) and West (1996), provides a procedure for testing whether a particular model (e.g., calendar trading rule) has predictive superiority over a benchmark model once the effects of data-mining have been accounted for. To accomplish this, the distribution of a suitable performance measure is evaluated in the context of the full set of models that led to the best-performing rule. Here we provide a brief summary of the approach and refer the reader to White (2000) for more details. The test procedure is based on the $l \times 1$ performance statistic:

$$\bar{f} = n^{-1} \sum_{t=R}^{T} f_{t+1}, \tag{3}$$

where l is the number of calendar rules, n is the number of prediction periods indexed from R through T so that $T = R + n - 1$, $f_{t+1} = f(Z_t, \hat{\beta}_t)$ is the observed performance measure for period $t+1$, and $\hat{\beta}_t$ is a vector of (recursively) estimated parameters. Generally, Z_t consists of a vector of dependent variables and predictor variables consistent with assumptions described in Diebold and Mariano (1995) or West (1996).

Our particular application involves no estimated parameters. Rather, the various parameterizations of the calendar rules (β_k, $k = 1, \dots, l$) directly generate returns that can be used to measure performance. In our full sample of the Dow Jones Industrial Average, n is set equal to 27,567, representing 100 years of daily predictions, and R is set equal to 1. The choice of performance metric is simple in our application: the finance literature assesses significance of calendar rules based on their return performance. Thus, a particular calendar rule's performance, where each of the calendar rules is indexed by a subscript k, is measured as follows:

$$f_{k,t+1}(\beta) = \ln[1 + y_{t+1} S_k(\mathscr{D}_{t+1}, \beta_k)] - \ln[1 + y_{t+1} S_0(\mathscr{D}_{t+1}, \beta_0)],$$

$$k = 1, \dots, l, \tag{4}$$

where $y_{t+1} = (X_{t+1} - X_t)/X_t$, X_t is the original price series (the Dow Jones Industrials Average and S&P 500 futures, in our case), and $S_k(\cdot)$ and $S_0(\cdot)$ are "signal" functions that, based on the system parameters β_k and β_0 convert the sequence of deterministic calendar indicators $\mathscr{D}_{t+1}$ into market positions. The signal functions take one of three values: 1 represents a long position, 0 a neutral position (i.e., out of the market), and -1 represents a short position. We also utilize an extension of this set-up to evaluate the calendar rules with the Sharpe ratio in addition to mean returns.

The null hypothesis to test when assessing whether there are abnormal calendar effects is that the expected performance of the best calendar rule is no better than the expected performance of the benchmark of always being in the market ($S_0(\mathscr{D}_{t+1}, \beta_0) = 1$). In other words,

$$H_0\colon \max_{k=1,\dots,l} \{\mathrm{E}(f_k)\} \leqslant 0. \tag{5}$$

Rejection of this null hypothesis would lead us to believe that the best calendar rule genuinely achieves performance superior to the benchmark.[3]

Since we have no estimated parameters in our forecasting models, the results in Diebold and Mariano (1995) immediately apply and the mean performance of the set of calendar rules follows an l-dimensional multivariate normal distribution. No closed-form expression exists that would permit us to assess the statistical significance of the maximum value of a general multivariate normal distribution. However, the null hypothesis can be tested by using White's (2000) method, which involves applying the stationary bootstrap of Politis and Romano (1994) to the observed values of $f_{k,t}$. Suppose B bootstrapped values of $\bar{f}_k$ are obtained from the calendar rules applied to the resampled time series of the returns and that we denote each of these resampled values by $\bar{f}^*_{k,i}$, where i indexes the B bootstrap samples. We set $B = 500$ and then construct the following statistics:[4]

$$\bar{V}_l = \max_{k=1,\dots,l} \{\sqrt{n}(\bar{f}_k)\}, \tag{6}$$

$$\bar{V}^*_{l,i} = \max_{k=1,\dots,l} \{\sqrt{n}(\bar{f}^*_{k,i} - \bar{f}_k)\}, \quad i = 1, \dots, B. \tag{7}$$

[3] Our procedure thus emphasizes deviations from the most general and perhaps strongest implication of the efficient market hypothesis, namely that risky assets never pay a negative risk premium—otherwise, in equilibrium, risk averse investors would not hold stocks but would switch into cash.

[4] The stationary bootstrap of Politis and Romano (1994) relies upon a "smoothing" parameter, q, to determine the average resampled block length. The value of q chosen in our experiments is 0.1, corresponding to a mean block length of 10 observations. This value appears to be reasonable given the weak correlation in daily stock returns. Furthermore, the bootstrap experiments of Sullivan et al. (1999) suggest that the results of White's reality check test procedure are rather robust to the choice of q for the data we are analyzing.

We compare $\bar{V}_l$ to the quantiles of $\bar{V}^*_{l,i}$ to obtain White's reality check p-value for the null hypothesis. For each resampled time series we employ the maximum value over all the l calendar rules to ensure that the reality check p-value incorporates the effects of data mining from the search over the l rules. Asymptotically, as n goes to infinity, with probability one the best-performing calendar rule will be identified and, if it truly outperforms the benchmark, then this will also be revealed in the sense that the performance measure exceeds zero, as White (2000) proves. In applying this procedure to asset market returns, it is important to consider the potential consequences of the well-known volatility persistence of asset market returns, and the possibility that returns are not necessarily stationary, but may be heterogeneously distributed. In fact, our procedures are constructed in a manner that endows them with a degree of robustness to both of these features of the data.

With regard to volatility persistence, the relevant effect is the introduction of neglected conditional heteroskedasticity into the estimation of the means of interest. As we are estimating a mean vector (as opposed to running a regression), this has no effect on the asymptotic variance–covariance matrix of the estimator of the mean, and the estimate of this matrix implicit in our bootstrap procedure is consistent for this variance–covariance matrix. (In fact, the implicit bootstrap variance–covariance matrix estimator is heteroskedasticity and autocorrelation consistent under mild conditions, so even in more complex situations, no difficulties arise on this score, asymptotically.)

With regard to possible heterogeneity of the data, we note that it is the bootstrap procedure that has the stationarity property and that this does not preclude the accuracy of the procedure when applied to heterogeneously distributed data. The key property required of the stationary bootstrap in this context is that it delivers a consistent estimate of the asymptotic variance–covariance matrix despite heterogeneity. That the stationary bootstrap is valid asymptotically for near epoch-dependent functions of heterogeneous mixing processes has recently been proven by Gonçalves and White (2000). Use of the stationary bootstrap thus will not have adverse consequences asymptotically in the face of volatility persistence or heterogeneity.

The bootstrap reality check is very general and can be applied to many traditional performance measures, including the R^2 goodness-of-fit statistic investigated by Foster et al. (1997). Here we also evaluate forecasts based on the Sharpe ratio, which measures the average excess return per unit of total risk. In this case we seek to test the null hypothesis

$$H_0\colon \max_{k=1,\ldots,l} \{g(\mathrm{E}(h_k))\} \leqslant g(\mathrm{E}(h_0)), \tag{8}$$

where h is a 3×1 vector with components given by

$$h^1_{k,t+1} = y_{t+1} S_k(\mathscr{D}_{t+1}, \beta_k), \tag{9}$$

$$h^2_{k,t+1} = (y_{t+1} S_k(\mathscr{D}_{t+1}, \beta_k))^2, \tag{10}$$

$$h^3_{k,t+1} = r^f_{t+1}, \tag{11}$$

the form of $g(\cdot)$ is given by

$$g(\mathrm{E}(h_{k,t+1})) = \frac{\mathrm{E}(h^1_{k,t+1}) - \mathrm{E}(h^3_{k,t+1})}{\sqrt{\mathrm{E}(h^2_{k,t+1}) - \mathrm{E}(h^1_{k,t+1})^2}} \tag{12}$$

and r^f_{t+1} is the risk-free interest rate. Expectations are estimated using arithmetic averages. Relevant sample statistics are

$$\bar{f}_k = g(\bar{h}_k) - g(\bar{h}_0), \tag{13}$$

where $\bar{h}_0$ and $\bar{h}_k$ are averages computed over the prediction sample for the benchmark model and the kth calendar rule, respectively. That is,

$$\bar{h}_k = n^{-1} \sum_{t=R}^{T} h_{k,t+1}, \quad k = 0, \ldots, l. \tag{14}$$

As in the previous case, the Politis and Romano (1994) bootstrap procedure is applied to yield B bootstrapped values of $\bar{f}_k$, denoted as $\bar{f}^*_{k,i}$, where

$$\bar{f}^*_{k,i} = g(\bar{h}^*_{k,i}) - g(\bar{h}^*_{0,i}), \quad i = 1, \ldots, B, \tag{15}$$

$$\bar{h}^*_{k,i} = n^{-1} \sum_{t=R}^{T} h^*_{k,t+1,i}, \quad i = 1, \ldots, B. \tag{16}$$

The bootstrap procedure can then be repeated to obtain p-values for the Sharpe ratio performance criterion.

A closely related idea of adjusting the critical value applied to assess the significance of the best model drawn from a larger set of competing models has also been explored in the literature on structural breaks. Searching across all (discrete) data points in order to find the best way of partitioning a data set is of course a type of data mining, and the appropriate adjustment of critical values in structural break tests can be very sizeable, as demonstrated in Christiano's (1992) analysis of Perron's (1989) trend-break model of US GNP. As in our setting, the critical values for the optimal break point must now be computed as the supremum of a set of standard test statistics; Andrews (1993) derives analytical results for the asymptotic case. Intriguingly, Diebold and Chen (1996) find that finite sample distortions of asymptotic critical values can be very substantial and they propose a bootstrap procedure whose size distortions appear to be small.

4. Calendar effects and data-mining biases

A vast empirical literature reports on the existence of calendar effects in stock returns. Best known are probably the low mean returns on Mondays and the high mean returns in January. French (1980) reports that returns on the S&P 500 tend to be negative from Friday's close to Monday's close and that this is not simply a result of the longer three-day period between these closing prices. Building on this discovery, Keim and Stambaugh (1984) find that the Monday effect is a weekend effect and that it is closely related to the January effect: during January, Monday returns are positive, while they become negative during the remaining part of the year.

Although the Monday and January effects are best known, a wealth of different calendar effects have also been reported. Ariel (1987) finds that mean stock returns are positive only for days immediately prior to or during the first half of calendar months. Consistent with this evidence, Lakonishok and Smidt (1988) report that, for the period 1897–1986, the null hypothesis that the difference between returns during the first and second half of the month is zero can be rejected at the 1 percent critical level. They also report very high abnormal returns for the period beginning on the last pre-Christmas trading day and ending on the last trading day of the year, and find a very strong turn-of-the-month effect, especially between the last day of the month and the first 3 days of the subsequent month.

Subsequent to their discovery, some of the calendar effects have been justified by theories relating to institutional arrangements in the markets. For example, the January effect has been linked to year-end tax-loss selling pressure that could suppress stock prices in December, only for them to bounce back in early January. Explanations offered for the strong Monday effect in stock returns data include delays between trading and settlements in stocks (Lakonishok and Levi, 1982), measurement error (Keim and Stambaugh, 1984), institutional factors (Flannery and Protopapadakis, 1988), and trading patterns (Lakonishok and Maberly, 1990). These factors appear to explain only a small portion of the Monday effect, however.

Most importantly, such theories are "after the fact" rationalizations of observed phenomena. Both Cross (1973) and French (1980) acknowledge that their study of weekend effects in stock returns were based on market participants' claim that prices tend to fall on Mondays. Clearly, the hypothesis of a Monday effect was not based on any theory and the same data were used both to formulate and to test the hypothesis, thus enhancing the dangers of data mining.[5]

[5] In fact, Levi (1988) points out that Cross's (1973) early study of Monday effects appeared right after a period over which a large Monday effect could be identified (namely, 1969–1972).

4.1. Universe of calendar effects

Our experiment replicates academics' and the investment community's search for successful calendar rules. At the heart of this experiment is the construction of a universe from which the calendar rules that have been reported could conceivably have been drawn. It is important not only to include the rules that were actually publicly reported, but also to include other rules that were considered. Lo and MacKinlay (1990) refer to the so-called "file drawer problem": many rules that do not appear to generate abnormal returns are never published and will be filtered out through the sequence of studies that focus on the successful ones. Nevertheless, the successful rules are still drawn from the larger universe that also includes unsuccessful, unreported rules and should thus be evaluated in this context.

Our strategy for constructing the universe of calendar rules is to identify certain types of calendar rules based on the calendar frequencies used in published studies. For a given frequency, the number of calendar effects is then constructed by permutational rules.

This ensures that all possible rules within a given frequency are considered and that we do not bias our study towards searching in a particular direction. The design of the universe is important since the correction for data mining is only done relative to the collection of models included in this set. For this reason we do not consider more recent calendar rules based on new technologies (e.g., genetic algorithms) since these have only become available more recently and it is very unlikely that the best calendar rule was drawn from a set that included such rules.

Two types of errors can occur in the statistical analysis. First, our universe may be too small and we may overlook some types of calendar rules that were investigated but never reported in published studies because they turned out not to be successful. This would tend to bias our estimated p-values towards zero since the mining-adjustment would not account for the full set of rules from which the best performer is drawn.

Second, we may include many rules in our universe that were not actually considered by investors. A natural concern then is whether considering the best calendar rules in conjunction with a very large number of possibly insignificant calendar rules will automatically lead to a loss of power so that even genuinely superior rules will not appear to do well if evaluated in a large enough universe. Fortunately, this is unlikely to be so. Since the bootstrap procedure corrects for the effects of data mining by way of the joint distribution of the returns generated by each rule, adding another trading rule will increase the reality check p-value only if the new rule increases the effective "span" of the universe from which the best rule was drawn, without simultaneously improving on the best performance. Adding uninformative rules whose returns are closely correlated with returns from rules previously included in the universe will not lead to a change in the p-value once the span of the universe is sufficiently large. Furthermore, if the additional calendar rule extends the span, then there is clearly the possibility that this rule might outperform the best-performing rule in the smaller universe of rules, so that the direction of the bias may be reversed. Although there does not yet exist a rigorous demonstration of the conclusions suggested by this heuristic argument, empirical evidence exists that reinforces the notion that genuinely superior rules can survive extensive data mining. Specifically, in a related study of technical trading rules (Sullivan et al., 1999), we found that the best technical trading rule for the period 1897–1987 maintained a p-value of effectively 0.000 despite the consideration of over 7800 other rules.

To answer both concerns, that of under-searching and that of over-searching we report the results for a large universe of almost 9500 calendar rules and for a much smaller universe that contains only the most basic types of calendar rules. The robustness of our finding effectively answers these competing concerns.

The full set of calendar effects comprises 9452 different rules. The basic frequencies of calendar effects are days, weeks, semi-months, months, and holidays. Some rules also combine two of these frequencies, such as in day-of-the-month effects. To keep the experiment feasible we do not use three or more layers of compounding such as in day-of-the-week-of-the-month rules, although such rules have in fact been studied (see, e.g., Wang et al., 1997). For similar reasons, and since the literature has focused on abnormal returns from following simple calendar rules relative to the benchmark of always being in the market, we do not investigate the compound rules resulting from combining long and short positions in the same trading strategy.[6]

We also consider a reduced universe of calendar effects in addition to the full universe. This permits us to examine the effects of data mining when only a core group of calendar rules is considered. The reduced universe contains a total of only 244 rules.

4.1.1. Day of the week

For the full universe of calendar effects, these rules explore the full combination of possibilities (ignoring Saturdays), except for long (short) all 5 days, and neutral all 5 days.[7] We consider both long and short positions. The reduced universe of calendar effects contains rules which are long (short) on each of the five days while being neutral otherwise, and rules which are neutral on each of the five days while being long (short) otherwise. Studies

[6] This decision is also based on what is computationally feasible. Combining short, neutral, and long positions in the individual trading rules would generate in excess of half a million candidate trading rules.

[7] We do not include Saturday effects since trading on Saturdays stopped in 1952.

that consider these effects include Ball and Bowers (1988), Cross (1973), Fields (1931), French (1980), Gibbons and Hess (1981), Jaffe and Westerfield (1985), Keim and Stambaugh (1984), Lakonishok and Levi (1982), Lakonishok and Smidt (1988), Rogalski (1984), and Smirlock and Starks (1986). The full universe number of rules is $2 \times (2^5 - 2) = 60$. The reduced universe number of rules is $10 + 10 = 20$.

4.1.2. Week of the month

For the full universe of calendar effects, again we use the full combination of possibilities, except for long (short) all 5 weeks, and neutral all 5 weeks. Both long and short positions are considered. The reduced universe of calendar effects contains rules which are long (short) on each of the 5 weeks while being neutral otherwise, and rules which are neutral on each of the 5 weeks while being long (short) otherwise. The "weeks" are constructed such that the first trading day of the month always occurs in the first week. If the first trading day of the month is a Friday (and there is no Saturday trading), then the first week will only contain 1 day; the following Monday will be part of week 2, and so forth. Ariel (1987), Lakonishok and Smidt (1988), and Wang et al. (1997) report on such effects. The full universe number of rules is $2 \times (2^5 - 2) = 60$. The reduced universe number of rules is $10 + 10 = 20$.

4.1.3. Month of the year

These rules generate by far the largest number of calendar effects for the full universe of calendar effects. We use the full combination of possibilities, except for long, short, or neutral all 12 months. Both long and short positions are considered. The reduced universe contains rules which are long (short) on each of the 12 months while being neutral otherwise, and rules which are neutral on each of the 12 months while being long (short) otherwise. Haugen and Lakonishok (1988), Keim (1983), Keim and Stambaugh (1984), Roll (1983), and Rozeff and Kinney (1976) investigate these calendar rules. The full universe number of rules is $2 \times (2^{12} - 2) = 8188$. The reduced universe number of rules is $24 + 24 = 48$.

4.1.4. Semi-month

There are two types of semi-month rules. The first type examines the halves of all months collectively. Thus, we consider four rules: long in the first half, neutral in the second; short in the first half, neutral in the second; neutral in the first half, long in the second; and neutral in the first half, short in the second. The second type of semi-month rule examines the halves of specific months. For instance, we look separately at the first half of March, or the second half of September. Then we apply one of four rules: long in the half of a specific month, neutral otherwise; short in that month's half, neutral otherwise; neutral in a specific month's half, long otherwise; and neutral in that month's half, and short otherwise. The semi-month rules included in both the full universe and reduced universe of calendar rules are identical. Ariel (1987) and Lakonishok and Smidt (1988) investigate these calendar rules.[8] The number of rules is $4+(4\times 24)=100$.

4.1.5. Holidays

We classify each day into one of three categories: pre-holiday, post-holiday, or normal. Pre-holiday days are those trading days which directly precede a day where the market is closed, but would normally be open for trading. Post-holiday days are those which directly follow pre-holiday days. All other days are considered normal. We examine pre-holiday and post-holiday days separately, with four rules for each: long on the pre(post)-holiday, neutral otherwise; short on the pre(post)-holiday, neutral otherwise; neutral on the pre(post)-holiday, long otherwise; and neutral on the pre(post)-holiday, short otherwise. The holiday rules included in both the full universe and reduced universe of calendar rules are identical. Fields (1934), French (1980), Haugen and Lakonishok (1988), Jones et al. (1987), and Lakonishok and Smidt (1988) investigate such holiday effects. The number of rules is $4+4=8$.

4.1.6. End of December

Following Lakonishok and Smidt (1988), we create three classifications of days at the end of December: from the middle of December up to, but not including, the last trading day prior to Christmas; from the first trading day after Christmas up to, but not including, the last trading day prior to New Year's day; the last trading day prior to Christmas and the last trading day prior to New Year's day. We then examine four rules for each of these classifications: long on the classified day, neutral otherwise; short on the classified day, neutral otherwise; neutral on the classified day, long otherwise; and neutral on the classified day, short otherwise. The end-of-December rules included in both the full universe and reduced universe of calendar rules are identical. Studies on these rules include Haugen and Lakonishok (1988), Jones et al. (1987), Lakonishok and Smidt (1984, 1988), Ritter and Chopra (1989) and Roll (1983). The number of rules is $4+4+4=12$.

4.1.7. Turn-of-the-month

For the full universe of calendar rules, we examine the first four trading days of the month and the last four trading days of the month and consider all possible combinations of these. In particular, we look at four strategies for each combination of these days: long on the turn-of-the-month day(s),

[8] Because of the different ways weeks and semi-months are defined, the semi-month rules are not spanned by the week of the month rules.

neutral otherwise; short on the turn-of-the-month day(s), neutral otherwise; neutral on the turn-of-the-month day(s), long otherwise; and neutral on the turn-of-the-month day(s), short otherwise. The reduced universe of calendar effects contains rules which are long (short) on each of the 8 days while being neutral otherwise, and rules which are neutral on each of the 8 days while being long (short) otherwise. Additionally, the 8 turn-of-the-month days are considered collectively in the same manner. Ariel (1987), Lakonishok and Smidt (1988) and Hensel and Ziemba (1996) inspect such trading rules. The full universe number of rules is $4 \times 2^8 = 1024$. The reduced universe number of rules is $16 + 16 + 4 = 36$.

The total number of calendar rules for the full universe is $60+60+8188+100+8+12+1024=9452$. The total number of calendar rules for the reduced universe, on the other hand, is $20 + 20 + 48 + 100 + 8 + 12 + 36 = 244$.

5. Empirical results

Some of the best-documented seasonal effects occur at either the daily or the monthly frequencies, so it is important to have a long data set with daily observations when testing for calendar effects. The longest continuously available daily data set on stock prices is the Dow Jones Industrial Average (DJIA) analyzed by Lakonishok and Smidt (1988) and Brock et al. (1992) (BLL). This data set goes back to 1897 and we have subsequently extended it from 1987 up to the end of 1996.[9] Two features of the data are worth noting. First, as pointed out by Lakonishok and Smidt (1988), the DJIA only includes large, actively traded firms and hence constitutes an ideal index from the perspective of identifying short-term market movements such as day of the week effects. Second, although the price index excludes dividends, Lakonishok and Smidt (1988) conduct a sensitivity analysis for this omission and find that excluding dividends does not affect their conclusions regarding the existence of seasonalities.[10]

We initially investigate calendar effects in the Dow Jones Industrial Average for the 90-year period from 1897 to 1986. Following Lakonishok and Smidt (1988), we split this sample into seven short sub-samples, each comprising approximately 13 years of data. Also, we include a long sub-sample covering the 90-year period from 1897 to 1986. This ensures that our results are directly comparable to those reported by Lakonishok and Smidt (1988),

[9] We thank Blake LeBaron for providing us with the data set used in the BLL study.

[10] Lakonishok and Smidt (1988) find that a quarterly seasonality is the largest regular calendar effect in the dividends of the firms included in the DJIA. Daily and other seasonal patterns are much smaller by comparison.

who conduct perhaps the most systematic study of calendar effects in the literature. The sub-samples are:

Sub-period 1: January 1897–December 1910
Sub-period 2: January 1911–December 1924
Sub-period 3: January 1925–December 1938
Sub-period 4: January 1939–May 1952
Sub-period 5: June 1952–December 1963
Sub-period 6: January 1964–December 1975
Sub-period 7: January 1976–May 1986
Sub-period 8: January 1897–May 1986,

while the full sample of the Dow Jones Industrial Average is

Full sample: January 1897–December 1996.

The stability of the best-performing trading rule across sub-samples will provide important information about calendar effects. For example, if the same calendar rule appears to be optimal in many different sub-samples, it would indicate that this rule is indeed capable of outperforming the benchmark. It also would suggest that investors could have adopted a recursive decision rule to identify the best performer and have used this to produce genuinely superior out-of-sample performance.[11]

5.1. Results for the mean return criterion (full universe)

For the reasons discussed above, we initially consider which types of rules were optimal in the various sample periods. This information is reported in Table 1. The most striking observation is that the optimal calendar rule changes between every single short sub-sample. Month of the year rules were chosen in three samples, a turn of the month rule in one sample, and three different day of the week rules were chosen in the remaining short sub-samples. Thus, there is no single calendar effect that clearly dominates in the sense that it almost always gets chosen as the optimal rule. This is exactly the kind of situation that is likely to lead researchers and investors to experiment with historical data since different rules, when studied in different periods, appear to be optimal. In the full sample, and in the long sub-sample, the optimal calendar rule is to be neutral on Mondays and be back in the market from Tuesdays to Fridays. Note that this corresponds to the well-publicized Monday (or Weekend) effect.

[11] See also Thaler (1987b) for an argument that the use of several sub-samples is a remedy against the effects of data mining.

Table 1
Best model from the full universe of calendar effects: mean return criterion[a]

Sample	Best model	Benchmark return	Model return	Nominal p-value	White's p-value
Jan 1897–Dec 1910	Month of the year: j, f, m, a, m, j, j, a, s, o, n, d = 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1	4.17	9.40	0.028	0.617
Jan 1911–Dec 1924	Month of the year: j, f, m, a, m, j, j, a, s, o, n, d = 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1	2.43	6.05	0.089	0.739
Jan 1925–Dec 1938	Day of the week: m, t, w, th, f = − 1, 0, 0, 0, 0	1.51	13.22	0.065	0.367
Jan 1939–May 1952	Turn of the month: −4, −3, −2, −1, 1, 2, 3, 4, otherwise = 0, 0, 0, 0, 0, 0, 0, 1, 1	3.42	8.01	0.048	0.481
June 1952–Dec 1963	Day of the week: m, t, w, th, f = 0, 1, 1, 1, 1	9.21	17.38	0.000	0.216
Jan 1964–Dec 1975	Day of the week: m, t, w, th, f = 0, 0, 1, 1, 1	0.93	10.88	0.000	0.241
Jan 1976–May 1986	Month of the year: j, f, m, a, m, j, j, a, s, o, n, d = 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1	7.56	10.61	0.090	0.915
Jan 1897–May 1986	Day of the week: m, t, w, th, f = 0, 1, 1, 1, 1	3.88	8.50	0.000	0.196
Jan 1897–Dec 1996	Day of the week: m, t, w, th, f = 0, 1, 1, 1, 1	4.63	8.66	0.000	0.243
Out-of-sample					
June 1986–Dec 1996	Week of the month: 1, 2, 3, 4, 5 = 1, 1, 1, 0, 1	11.61	15.23	0.117	0.874
S&P 500 Futures (Jan 1983–Dec 1996)	Month of the year: j, f, m, a, m, j, j, a, s, o, n, d = 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1	8.54	10.33	0.297	0.992

[a] This table presents the performance results of the best calendar rule, chosen with respect to the mean return criterion, in each of the sample periods, for the full universe of 9452 calendar effects. The table reports the type of the best-performing model, the annualized mean return for the benchmark model and the best-performing model, White's reality check p-value, and the nominal p-value (i.e., that which results from applying the reality check methodology to the best trading rule only, thereby ignoring the effects of the data snooping).

Table 1 also reports the performances of the benchmark market portfolio and of the calendar rule with the highest average return in any given sample period. Also included in the table are the nominal[12] and reality check p-values. In the full sample, 1897–1996, the market paid a mean annualized return of 4.63 percent, compared to 8.66 percent on the optimal rule that is neutral on Mondays. As a result, the nominal p-value of the best calendar rule is less than 0.002 in the full sample and in the long 90-year sub-sample.[13] In sharp contrast, the reality check p-value of the best trading rule is 0.20 in the 90-year sub-sample and 0.24 in the full sample.

In most of the shorter sub-samples, the superior performance of the best calendar rule relative to the market index is even larger, and this difference varies from 3.05 to a very substantial 11.71 percent per year. In these samples, the nominal p-values are significant at the 10 percent critical level in all of the seven periods, while the reality check p-value is always higher than 0.21. These numbers effectively illustrate the role of data mining in explaining findings of “abnormal” returns in the stock market related to calendar effects. Our results confirm the concern expressed in many empirical studies that the effect on statistical inference from data mining can be very serious.

For each of the calendar rules included in the universe and using again the full sample period 1897–1986, Fig. 1 plots the mean return as a function of the model number. Also shown in the figure are the sequentially updated highest mean return and the corresponding reality check p-value. Since the sequential ordering of calendar rules is arbitrary, only the terminal value of the highest mean return and the terminal reality check p-value matter to the final assessment. The clear emergence of patterns in the mean returns effectively demonstrates the complicated dependencies operating across the mean payoffs generated by the different calendar rules. The regularity with which these patterns occur reflects the ordering of the permutations used to generate the universe of trading rules. This graphical illustration of the complexity of the cross-sectional dependencies also indicates that it would be very difficult to deal with these other than through a bootstrap procedure.

Another interesting point that emerges from Fig. 1 is that the best model (i.e., the model that is neutral on Mondays) is selected early on, after only 29 models have been considered. From this point the “increased spanning” effect of the calendar rules leads to a gradual increase in the mining-adjusted p-value. That the p-value of the best model is not below 0.10 when the best model is only considered in conjunction with 28 other day of the week rules also demonstrates that the effects of data mining can be very large even in

[12] The nominal p-value is that which results from applying the bootstrap methodology to the best trading rule *only*, thereby ignoring the effects of the data mining.

[13] Our choice of utilizing 500 bootstrap re-samples in the reality check procedure results in an accuracy of $1/500 = 0.002$.

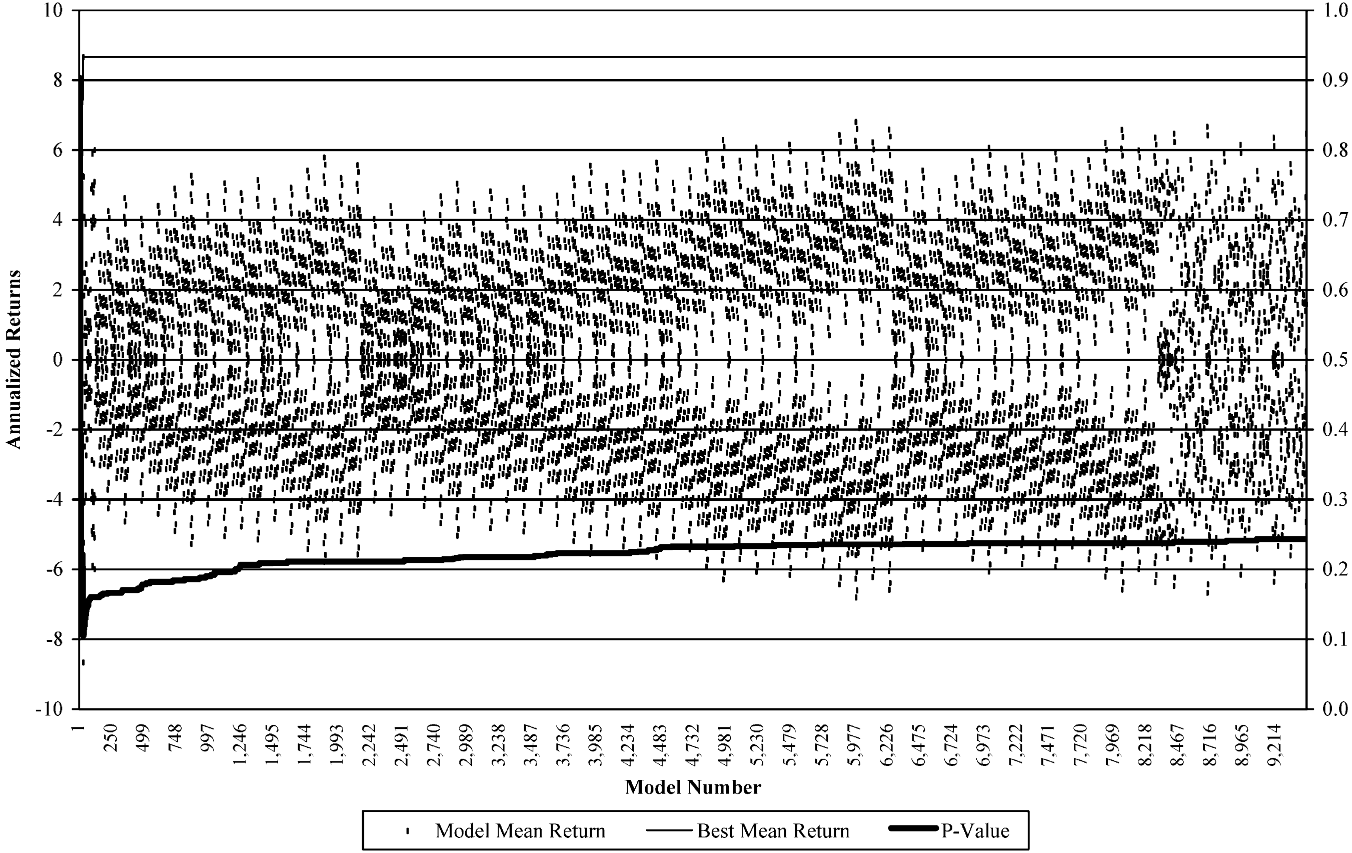

Fig. 1. This figure presents the economic and statistical performance of the best model chosen according to the mean return criterion for the DJIA, 1897–1996. For a given trading rule, n, indexed on the x-axis, the scattered points plot the mean annualized returns experienced during the sample period. The thin line measures the best mean annualized return among the set of trading rules $i = 1,\ldots,n$ and the thick line measures the associated data-snooping adjusted p-value.

a very small universe of forecasting models. Also note that the "dilution" or "erosion" of the p-value occurs at a rather moderate (and decreasing) rate as we expand the universe of rules.

5.2. Results for the Sharpe ratio criterion (full universe)

We construct the Sharpe ratio as excess returns divided by the standard deviation of returns. Excess returns are measured as the returns from a given calendar effect trading rule less the risk-free interest rate. Considering that data on daily risk-free interest rates is not readily available over the entire 100 year sample, we utilize data from three separate sources for three over-lapping periods. The first source is Banking and Monetary Statistics, 1914–1941 (1943), where we extract the interest rate for 90-day stock exchange time loans. Since these rates are reported on a monthly basis only, we convert them into a daily series by simply applying the interest rate reported for a given month to each day of that month. This source is used for our risk-free interest rate from 1897 to 1925. From 1926 to June 1954, we use the 1-month T-bill rates from the Fama/Bliss risk-free rates CRSP file. These are also reported on a monthly basis and converted into a daily series in the same way. Finally, from July 1954 to 1996, we are able to employ daily rates from the daily Federal funds rate.[14] The three sources of interest rates are concatenated to form a single data series. Each data source reports the daily interest rate in annualized form. We thus convert the annualized rates reported into daily rates using the following formula:

$$r_{\mathrm{d}} = \frac{\ln(1 + r_{\mathrm{ann}})}{252}, \tag{17}$$

where r_{d} is the daily interest rate, r_{ann} is the reported annualized rate, and 252 represents the average number of trading days in a year.[15]

The main effect of including the risk-free rate in the Sharpe ratio is that of a (time varying) drift-adjustment, since the volatility of daily interest rates is substantially smaller than that of daily stock returns. For this reason, our use of monthly interest rates in the earlier samples is unlikely to affect the results in any important way.

[14] The Federal funds rate is the cost of borrowing immediately available funds, primarily for one day. The effective rate is a weighted average of the reported rates at which different amounts of the day's trading through New York brokers occurs.

[15] Examining the behavior of our interest rates in the first overlapping period (1925–1941, 193 observations), we find that monthly values for the stock exchange 90-day time loans and the Fama/Bliss risk-free rates have a correlation of 0.964. To compare the Fama/Bliss risk-free rates (monthly) to the Federal funds rates (daily), we converted the risk-free rates to daily rates by applying the Fama/Bliss rate for a given month to all days in that month. The overlap period of 1954–1996 (15,525 observations) produced a correlation of 0.963.

Table 2 shows that the best calendar rules chosen according to the Sharpe ratio criterion tend to be different from those based on selecting calendar rules according to the highest mean return. However, for the longest sub-sample period, 1897–1986, the day of the week rule that invested in stocks Tuesday through Friday and was out of the market on Mondays produces the highest Sharpe ratio. The many changes in the type of optimal rule chosen in the different sub-samples is quite striking: three day of the week, two turn of the month, and two month of the year rules are selected.

The best calendar rule, chosen according to the Sharpe criterion, produces a Sharpe ratio of 0.33 in the full sample 1897–1996. This value is much higher than that of the market index (0.01). As a result, the nominal p-value of the best calendar rule is less than 0.002. This may seem to be fairly conclusive evidence that calendar effects matter. However, once data-mining effects are accounted for even the most successful calendar rule is no longer significant at the 10 percent critical level. In fact, the best calendar rules generate nominal p-values below 0.1 in every single sub-sample, but not a single one of the reality check p-values is significant at the 10 percent critical level.

Fig. 2 plots the estimated Sharpe ratio for each of the models in the universe of calendar trading rules. It also shows the sequence tracking the highest Sharpe ratio and the corresponding reality check p-value. As in Fig. 1, the systematic patterns in the sequence of estimates of the performance measure reveal the complicated cross-sectional dependencies operating in the simulation experiment. Note from the break-up in patterns that the effective span of models increases after 8300 models have been inspected. Since the maximum Sharpe ratio does not increase at this point, the reality check p-value jumps dramatically from about 0.33 to 0.52 subsequent to including just a handful of these new rules.[16] From this point onward, without an increase in the maximum Sharpe ratio, the p-value experiences a gradual increase throughout the remainder of the universe. This illustrates succinctly how the effective span of trading rules affects the reality check p-value.

5.3. Out-of-sample results

As pointed out by many researchers, new data provides an effective remedy against data mining. Use of new data ensures that the sample on which a hypothesis was originally based effectively is separated from the sample used to test the hypothesis. The two most prominent studies that use the DJIA

[16] This break-point is due to the change from month of the year rules to half of the month rules.

Table 2
Best model from the full universe of calendar effects: Sharpe ratio criterion[a]

Sample	Best model	Benchmark Sharpe ratio	Model Sharpe ratio	Nominal p-value	White's p-value
Jan 1897–Dec 1910	Turn of the month: $-4,-3,-2,-1,1,2,3,4$, otherwise $=1,0,1,1,1,1,1,0,1$	0.01	0.46	0.036	0.845
Jan 1911–Dec 1924	Month of the year: $j,f,m,a,m,j,j,a,s,o,n,d=1,0,1,1,1,0,0,1,1,1,0,1$	−0.13	0.12	0.055	0.976
Jan 1925–Dec 1938	Day of the week: $m,t,w,th,f=-1,0,0,0,0$	−0.01	0.99	0.006	0.130
Jan 1939–May 1952	Turn of the month: $-4,-3,-2,-1,1,2,3,4$, otherwise $=1,1,0,0,0,0,0,1,1$	0.23	1.27	0.000	0.125
June 1952–Dec 1963	Day of the week: $m,t,w,th,f=0,0,1,1,1$	0.63	1.66	0.000	0.274
Jan 1964–Dec 1975	Day of the week: $m,t,w,th,f=0,0,1,1,1$	−0.37	0.51	0.000	0.584
Jan 1976–May 1986	Month of the year: $j,f,m,a,m,j,j,a,s,o,n,d=1,0,1,1,0,1,1,1,0,0,1,1$	−0.13	0.12	0.081	1.000
Jan 1897–May 1986	Day of the week: $m,t,w,th,f=0,1,1,1,1$	0.01	0.33	0.000	0.467
Jan 1897–Dec 1996	Day of the week: $m,t,w,th,f=0,1,1,1,1$	0.04	0.33	0.000	0.554
Out-of-sample					
June 1986–Dec 1996	Week of the month: $1,2,3,4,5=1,0,1,0,1$	0.35	0.83	0.029	0.915
S&P 500 Futures (Jan 1983–Dec 1996)	Turn of the month: $-4,-3,-2,-1,1,2,3,4$, otherwise $=1,1,1,1,1,1,1,0,1$	0.11	0.33	0.212	0.997

[a] This table presents the performance results of the best calendar rule, chosen with respect to the Sharpe ratio criterion, in each of the sample periods, for the full universe of 9452 calendar effects. The table reports the type of the best-performing model, the Sharpe ratio for the benchmark model and the best-performing model, White's reality check p-value, and the nominal p-value (i.e., that which results from applying the reality check methodology to the best trading rule only, thereby ignoring the effects of the data snooping).

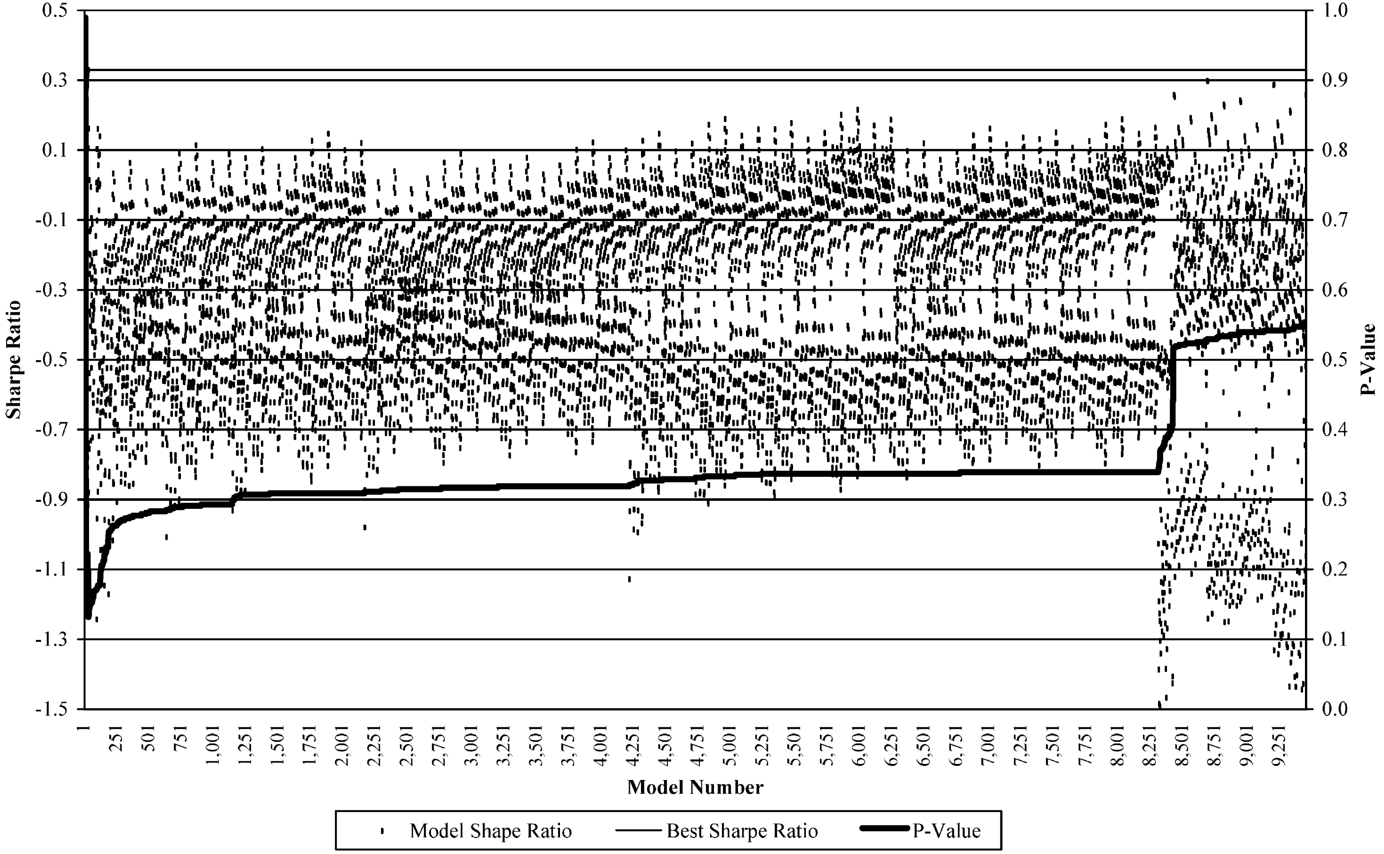

Fig. 2. This figure presents the economic and statistical performance of the best model chosen according to the Sharpe ratio criterion for the DJIA, 1897–1996. For a given trading rule, n, indexed on the x-axis, the scattered points plot the Sharpe ratio experienced during the sample period. The thin line measures the best Sharpe ratio among the set of trading rules $i = 1,\ldots,n$ and the thick line measures the associated data-snooping adjusted p-value.

index, namely Lakonishok and Smidt (1988) and Brock et al. (1992), both end in May 1986. Thus, we have access to a post-sample data set comprising daily observations from June 1986 to December 1996. This is a relatively short period to evaluate the monthly effects but it is a long sample for the daily calendar effects.

To address issues related to the absence of dividends from the DJIA price index and transaction costs incurred from implementing the calendar rules in a trading strategy based on the individual stocks traded in the cash market, we also consider the longest available futures price index, namely the Standard and Poor's (S&P) 500 Futures Price Index, established in 1983. Hence, we use the following out-of-sample periods: [17]

Sub-period 9: June 1986–December 1996
S&P 500 Futures: January 1983–December 1996.

Pinnacle Data Corporation is the provider of the S&P 500 futures data. We use the prices from the nearest futures contract and apply a rollover date of the 9th of the delivery month for the contract. In doing so, on the 9th of March, June, September, and December, any position in the current contract is closed out, and a new position is opened, in accordance with the calendar trading rule. We then develop a series of returns that is linked together at the rollover dates. A new price series is generated from this returns series by employing the price of the S&P 500 futures contract on the first available day of our data set.

According to the mean return criterion, in the sample June 1986–December 1996, a calendar rule that is in the market all the time, except during the fourth week of each month, turns out to be optimal. In the case of the S&P 500 Futures data covering the period 1983–1996, a month of the year rule that is out of the market in September and October is optimal. [18] The best calendar rule chosen for the DJIA data generated a mean return in excess of the market index of 3.6 percent per year. This yielded a nominal p-value of 0.12 and a reality check p-value of 0.87. An even weaker performance of the best calendar rule emerges from the Futures data set: here the abnormal performance is 1.8 percent per year, yielding nominal and reality check p-values of 0.30 and 0.99, respectively.

These findings also show that if an investor had chosen to use the calendar rule that fared best up to 1986, namely the Monday rule, in a trading strategy between 1987 and 1996, then this would not have produced statistically

[17] The futures price data effectively covers an out-of-sample period since it only has a short overlap with the 1897–1986 period considered by earlier studies and all other studies have looked at cash market prices.

[18] Interestingly, if we omit trading on October 19, 1987, it is no longer optimal to be neutral during the month of October.

significant superior performance relative to the market index.[19] Furthermore, analyzing the Monday effect from its initial date of publicity in Cross (1973) through to 1996, reveals that it did not continue to significantly outperform. The market during this period earned a mean annualized return of 8.78 percent, whereas the Monday rule earned just slightly more at 9.02 percent. The nominal p-value is not significant at 0.44. The outcome is similar when examining the Sharpe ratio, with a nominal p-value of 0.36. The Sharpe ratios for the benchmark and Monday rule were 0.078 and 0.112, respectively. This is additional evidence against the presence of calendar effects in stock returns.

Further extending this argument, it is worth noting that the literature had already identified a variety of calendar anomalies by the mid-1970s. One of our sub-samples, 1976–1986, closely corresponds to this period and it is reassuring to see that even the nominal p-values of the best rule, at 0.09 and 0.08 for the mean return and Sharpe ratio criteria, respectively, are rather high. This again suggests that no investor could have outperformed the market benchmark after reading the early literature even if this would have allowed the investor to correctly pick the best rule.

Fig. 3 provides a fascinating picture of complicated cross-dependencies in the payoffs produced by the universe of calendar rules selected according to the mean return criterion over the period 1986–1996. As in Fig. 2, the patterns break up after 8300 rules have been considered. In contrast to the earlier figure, the reality check p-value does not dramatically increase at this point. These new rules do not improve on the best mean return performance and thus experience a gradual increase in the p-value.

The earlier in-sample results on the Sharpe ratio carry over to the out-of-sample period. For these, the Sharpe ratios of the best calendar rules are more than twice as high as those of the benchmark, resulting in nominal p-values of 0.03 for the DJIA, and 0.21 for the S&P 500 futures. However, the reality check p-value is greater than 0.91 for both series.

6. Extensions of the empirical results

In this section, we investigate the robustness of the results on the significance of calendar rules with regard to data-mining effects. Initially, we analyze the impact of data mining on calendar effects in the context of the reduced universe comprising only 244 calendar rules. We also look at the sensitivity of the findings in the original studies on the Monday effect with

[19] Indeed, an investor engaging in the Monday rule during the out-of-sample period in the DJIA would have experienced an annualized mean return of 10.2%, which is less than the 11.6% provided by the buy-and-hold strategy. Note that this is true even with the “assistance” provided to the rule by the market break that occurred on October 19, 1987, which is a Monday.

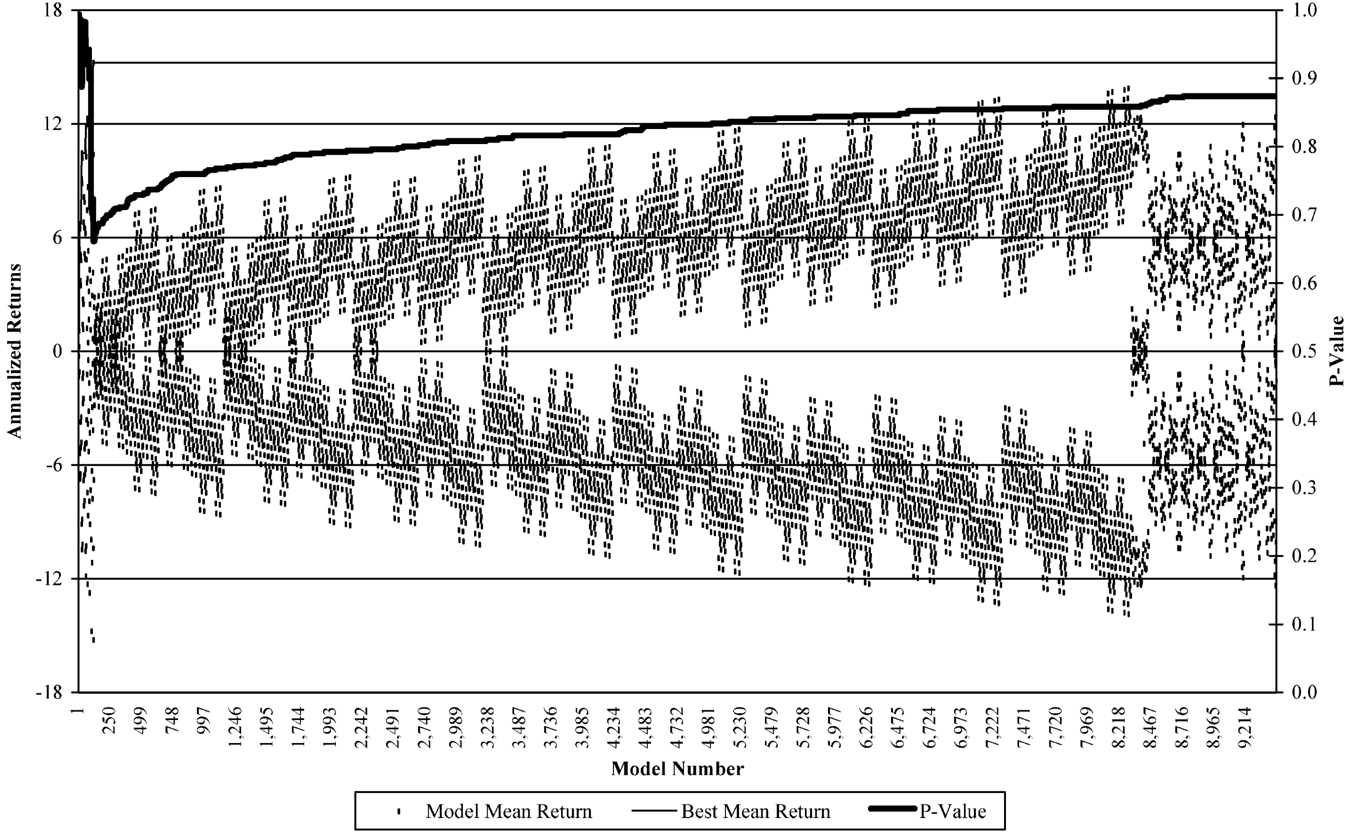

Fig. 3. This figure presents the economic and statistical performance of the best model chosen according to the mean return criterion for the DJIA, 1986–1996. For a given trading rule, n, indexed on the x-axis, the scattered points plot the mean annualized returns experienced during the sample period. The thin line measures the best mean annualized return among the set of trading rules $i = 1, \ldots, n$ and the thick line measures the associated data-snooping adjusted p-value.

regard to data mining, using only a handful of rules and the original sample periods adopted by those studies. We conclude this section by examining the possible presence of calendar effects in small-cap stocks.

6.1. Results from the reduced universe of calendar effects

In addition to the full universe of calendar rules, we consider a reduced universe that contains only the core set of calendar effects. This universe comprises only 244 rules rather than 9452. The entire experiment explained above is repeated for this smaller set of calendar effects. The reason for doing this is to answer the question, "How does the best rule fare, in terms of the data-mining adjusted p-value, when examined in the context of only a handful of rules?"[20]

Table 3 provides the best calendar effect trading rule, according to the mean return criterion, for each of the sub-samples, the full data set, and both the DJIA and S&P 500 futures "out-of-sample" periods. Similar to the results for the full universe, the optimal rule varies in nearly every sub-period, with the exception being the two sub-periods extending from June 1952 to December 1975. These two sub-periods find the Monday effect rule to be the best, which is quite interesting considering that it is these periods which were most heavily investigated to find the Monday effect in the first place. In two of the sub-periods, the best rule chosen from the full universe is also contained in the reduced universe. The reality check p-values, however, are still not significant, with values greater than 0.17. This same phenomenon also occurs for the original data set from 1897 to 1986, and the full sample from 1897 to 1996. We conclude that the results are strikingly similar to those for the full universe of calendar rules.

The performance of the best rule chosen from the reduced universe according to the Sharpe ratio criterion, is presented in Table 4 for each of the sample periods. Although the best rule varies from one sub-period to the next for all of the sub-periods, only three unique rules are found to be best when evaluated by the Sharpe ratio. Two of these three rules exploit the anomaly of negative returns on Mondays. Six out of the seven sub-periods have data-mining adjusted p-values greater than 0.24. In addition, both of the out-of-sample periods perform very poorly with reality check p-values above 0.93.

[20] Note that the best rule chosen from the full universe of calendar effects may not be included in the reduced universe.

Table 3
Best model from the reduced universe of calendar effects: mean return criterion[a]

Sample	Best model	Benchmark return	Model return	Nominal p-value	White's p-value
Jan 1897–Dec 1910	Month of the year: j, f, m, a, m, j, j, a, s, o, n, d = 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1	4.17	7.60	0.000	0.553
Jan 1911–Dec 1924	Month of the year: j, f, m, a, m, j, j, a, s, o, n, d = 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1	2.43	4.86	0.115	0.687
Jan 1925–Dec 1938	Day of the week: m, t, w, th, f = − 1, 0, 0, 0, 0	1.51	13.22	0.065	0.270
Jan 1939–May 1952	Turn of the month: $-4, -3, -2, -1, 1, 2, 3, 4$, otherwise = 1, 1, 1, 1, 1, 1, 1, 1, 0	3.42	7.60	0.062	0.349
June 1952–Dec 1963	Day of the week: m, t, w, th, f = 0, 1, 1, 1, 1	9.21	17.38	0.000	0.170
Jan 1964–Dec 1975	Day of the week: m, t, w, th, f = 0, 1, 1, 1, 1	0.93	9.79	0.000	0.211
Jan 1976–May 1986	Semi-month: second half of October, otherwise = 0, 1	7.56	9.32	0.048	0.906
Jan 1897–May 1986	Day of the week: m, t, w, th, f = 0, 1, 1, 1, 1	3.88	8.50	0.000	0.119
Jan 1897–Dec 1996	Day of the week: m, t, w, th, f = 0, 1, 1, 1, 1	4.63	8.66	0.000	0.167
Out-of-sample					
June 1986–Dec 1996	Week of the month: 1, 2, 3, 4, 5 = 1, 1, 1, 0, 1	11.61	15.23	0.117	0.600
S&P 500 Futures (Jan 1983–Dec 1996)	Week of the month: 1, 2, 3, 4, 5 = 1, 1, 1, 0, 1	8.54	10.20	0.299	0.913

[a] This table presents the performance results of the best calendar rule, chosen with respect to the mean return criterion, in each of the sample periods, for the reduced universe of 244 calendar effects. The table reports the type of the best-performing model, the annualized mean return for the benchmark model and the best-performing model, White's reality check p-value, and the nominal p-value (i.e., that which results from applying the reality check methodology to the best trading rule only, thereby ignoring the effects of the data snooping).

Table 4
Best model from the reduced universe of calendar effects: Sharpe ratio criterion[a]

Sample	Best model	Benchmark Sharpe ratio	Model Sharpe ratio	Nominal p-value	White's p-value
Jan 1897–Dec 1910	Turn of the month: $-4, -3, -2, -1, 1, 2, 3, 4$, otherwise $= 1, 1, 1, 1, 1, 1, 1, 1, 0$	0.01	0.23	0.133	0.956
Jan 1911–Dec 1924	Day of the week: m, t, w, th, f $= 0, 1, 1, 1, 1$	−0.13	0.02	0.157	0.978
Jan 1925–Dec 1938	Day of the week: m, t, w, th, f $= -1, 0, 0, 0, 0$	−0.01	0.99	0.006	0.069
Jan 1939–May 1952	Turn of the month: $-4, -3, -2, -1, 1, 2, 3, 4$, otherwise $= 1, 1, 1, 1, 1, 1, 1, 1, 0$	0.23	1.02	0.000	0.241
June 1952–Dec 1963	Day of the week: m, t, w, th, f $= 0, 1, 1, 1, 1$	0.63	1.60	0.000	0.276
Jan 1964–Dec 1975	Day of the week: m, t, w, th, f $= -1, 0, 0, 0, 0$	−0.37	0.48	0.056	0.579
Jan 1976–May 1986	Semi-month: second half of October, otherwise $= 0, 1$	−0.13	0.00	0.043	0.997
Jan 1897–May 1986	Day of the week: m, t, w, th, f $= 0, 1, 1, 1, 1$	0.01	0.33	0.000	0.435
Jan 1897–Dec 1996	Day of the week: m, t, w, th, f $= 0, 1, 1, 1, 1$	0.04	0.33	0.000	0.511
Out-of-sample					
June 1986–Dec 1996	Week of the month: $1, 2, 3, 4, 5 = 1, 1, 1, 0, 1$	0.35	0.73	0.031	0.935
S&P 500 Futures (Jan 1983–Dec 1996)	Week of the month: $1, 2, 3, 4, 5 = 1, 1, 1, 0, 1$	0.11	0.28	0.196	0.988

[a] This table presents the performance results of the best calendar rule, chosen with respect to the Sharpe ratio criterion, in each of the sample periods, for the reduced universe of 244 calendar effects. The table reports the type of the best-performing model, the Sharpe ratio for the benchmark model and the best-performing model, White's reality check p-value, and the nominal p-value (i.e., that which results from applying the reality check methodology to the best trading rule only, thereby ignoring the effects of the data snooping).

6.2. In-sample data-mining biases for the Monday effect

We also investigate whether the Monday effect was really present in the exact sample period for which it was originally reported. Cross (1973), French (1980), and Gibbons and Hess (1981) analyze the sample periods January 1953–December 1970, January 1953–December 1977, and July 1962–December 1978, respectively. Each of these studies uses the S&P 500 cash index, and Gibbons and Hess (1981) also include the CRSP value-weighted and equal-weighted portfolios.[21]

We examine each of these sample periods with the DJIA and examine the effects of very moderate data mining on the results. Interestingly, for each of the three sub-periods, the best model from the reduced universe of rules is indeed the Monday effect. We abridge the reduced universe of rules even further and consider only day-of-the-week trading rules. In particular, we consider the performance of the Monday effect when considered in conjunction with only 10 day-of-the-week rules (including long and neutral positions only), as well as after examining the 20 day-of-the-week rules which include both long and short positions.

Table 5 presents the results of this exercise. When considered in isolation, the Monday rule (neutral on Mondays, long otherwise) has nominal p-values of nearly zero for both the mean return and Sharpe ratio criteria. However, introducing only a handful of models tends to increase the reality check p-value substantially.

In both the 1953–1970 sample and the 1953–1977 sample, the reality check p-value does not experience any fluctuation while examining the first 10 day-of-the-week rules which include only long and neutral positions, and thus remains near zero. This is true for both criteria. The 1962–1978 sample, on the other hand, shows an increase in the p-value from zero to 0.034 for the mean return criterion, and an increase to 0.087 for the Sharpe ratio criterion. Note that this increase in the p-value occurs after examining only the 10 standard day-of-the-week rules.

[21] One could argue that the period following the study by Fields (1931) determines an out-of-sample period with regard to the Monday effect. We believe that there are two reasons why this argument is implausible. First, Fields (1931) was explicitly *not* looking for a Monday effect but instead considered a Saturday effect: “More specifically, it is often held that the unwillingness of traders to carry their holdings over the uncertainties of a week-end leads to a liquidation of long accounts and a consequent decline of security prices on Saturday” (p. 415). This would seem to predict a *positive* Monday effect as prices bounce back after the weekend. Furthermore, Fields refers to the “complete absence” of a Saturday effect, rather than pointing to the existence of a Monday effect. Second, other calendar effects were no doubt focused on historically but did not subsequently receive the same attention because they were not as large in subsequent samples as the Monday effect. This leaves us again with the task of assessing the best individual calendar rule in conjunction with a larger set of (historical) calendar rules.

Table 5
Performance of the Monday effect trading rule[a]

Results for the mean return criterion

Sample	Benchmark return	Model return	Nominal p-value	10 Model p-value	20 Model p-value
Jan 1953–Dec 1970					
Cross (1973)	5.90	14.72	0.000	0.000	0.062
Jan 1953–Dec 1977					
French (1980)	4.20	12.24	0.000	0.000	0.062
July 1962–Dec 1978					
Gibbons and Hess (1981)	2.20	8.99	0.000	0.034	0.212

Results for the Sharpe ratio criterion

Sample	Benchmark Sharpe ratio	Model Sharpe ratio	Nominal p-value	10 Model p-value	20 Model p-value
Jan 1953–Dec 1970					
Cross (1973)	0.23	1.20	0.000	0.000	0.032
Jan 1953–Dec 1977					
French (1980)	0.00	0.76	0.000	0.000	0.051
July 1962–Dec 1978					
Gibbons and Hess (1981)	−0.26	0.30	0.000	0.087	0.265

[a] This table presents the performance results of the Monday effect trading rule, evaluated with respect to both the mean return and Sharpe ratio criteria, in three sample periods corresponding to the original literature. The table reports the annualized mean return or Sharpe ratio for the benchmark model and the Monday rule, the nominal p-value, and White's reality check p-value after examination of the 10 long and neutral day-of-the-week rules, and the full 20 day-of-the-week rules including short positions.

In the context of an examination of all 20 day-of-the-week rules, all three sub-periods exhibit a substantial increase in the p-value. For the mean return criterion, the p-value increases to 0.062 for the two samples beginning in 1953 and jumps to over 21 percent in the 1962–1978 sample. The results are similar for the Sharpe ratio.

This suggests that if the original researchers had been able to formally account for data-mining biases, they may have been relatively skeptical of their results. Evaluating the Monday effect in the context of only the core day-of-the-week rules renders the statistical significance of the Monday effect doubtful, even in the period during which the Monday effect was discovered.

6.3. Calendar effects in small firms' returns

So far we have only considered calendar effects in the largest stocks. This choice of portfolio was made to ensure that the included stocks are very liquid and trade without large transaction costs so that findings of calendar effects would be more of a challenge to the efficient market hypothesis than if these were found for less liquid firms. However, the emphasis on large firms may be a concern since some calendar effects, such as the high returns in January, are largely associated with small firms, cf. Keim (1983) and Reinganum (1983). It is somewhat dangerous to extend our study to discriminate by firm size without accounting for data mining in this regard since market capitalization is likely to be one of many potential firm attributes that was itself detected through extensive data exploration. To genuinely control for data mining one would need to expand the study to include calendar effects in firms sorted by size, accounting characteristics, market sector, and so forth. This does not appear to be computationally feasible at this moment. Since calendar effects to many are synonymous with a firm size effect, we simply expand the bootstrap experiment to a portfolio of small firms. The reality check p-values reported in the context of this study thus constitute a lower bound for the value that would be found under a fuller accounting for the effects of data mining.

The very smallest firms' stocks are highly illiquid so calendar effects in their returns would be difficult to interpret economically. Therefore, we conduct the bootstrap experiment on the S&P Small Cap 600 index. This is a value-weighted portfolio that comprises small firms chosen for their liquidity and industry group representation. We have daily data from Datastream going back to 1973 and split the period 1973–1996 into two subsamples, following again the 1986 cutoff date. Using this more recent sample also has the advantage that the findings are easier to interpret since round-trip transaction costs have been coming down, making it more likely that existing calendar effects could be exploited to earn a profit.

The results, reported in Table 6, are again striking. Across sample periods, the best calendar rule is to be out of the market on Mondays and

Table 6

Results for the S&P600 Small-Cap Index

This table presents the performance results of the best calendar rule for the S&P600 Small-Cap index, chosen with respect to both the mean return and Sharpe ratio criteria, in three sample periods, for both the full and reduced universe of calendar effects. The table reports the type of the best performing model, the annualized mean return or Sharpe ratio for the benchmark model and the best performing model, White's Reality Check p - value, and the nominal p-value (i.e., that which results from applying the Reality Check methodology to the best trading rule only, thereby ignoring the effects of the data-snooping).

Results for the Mean Return Criterion

	Benchmark Return	Model Return	Nominal p-value	White's p-value
Full Universe of Calendar Effects				
Jan 1973 - May 1986 [1]	10.66	20.84	0.0000	0.1964
June 1986 - Dec 1996 [1]	9.40	16.24	0.0076	0.4078
Jan 1973 - Dec 1996 [1]	10.05	18.82	0.0000	0.1336
Reduced Universe of Calendar Effects				
Jan 1973 - May 1986 [2]	10.66	17.95	0.0000	0.2435
June 1986 - Dec 1996 [2]	9.40	14.95	0.0079	0.3715
Jan 1973 - Dec 1996 [2]	10.05	16.57	0.0000	0.1885

Results for the Sharpe Ratio Criterion

	Benchmark Sharpe Ratio	Model Sharpe Ratio	Nominal p-value	White's p-value
Full Universe of Calendar Effects				
Jan 1973 - May 1986 [1]	0.13	1.21	0.0000	0.7775
June 1986 - Dec 1996 [3]	0.28	1.26	0.0011	0.7043
Jan 1973 - Dec 1996 [1]	0.19	1.19	0.0000	0.5925
Reduced Universe of Calendar Effects				
Jan 1973 - May 1986 [2]	0.13	0.79	0.0000	0.9455
June 1986 - Dec 1996 [4]	0.28	1.16	0.0000	0.7451
Jan 1973 - Dec 1996 [2]	0.19	0.82	0.0000	0.8182

Best Models
1: Day of the Week: m, t, w, th, f = 0, 0, 1, 1, 1
2: Day of the Week: m, t, w, th, f = 0, 1, 1, 1, 1
3: Turn of the Month: -4, -3, -2, -1, 1, 2, 3, 4, otherwise = 0, 0, 1, 1, 1, 1, 1, 0, 1
4: Turn of the Month: -4, -3, -2, -1, 1, 2, 3, 4, otherwise = 1, 1, 1, 1, 1, 1, 1, 1, 0

sometimes also on Tuesdays. This rule comes close to doubling the mean return of the benchmark and consequently leads to nominal p-values below 1 percent. However, once again the reality check p-values are always higher than 10 percent. The corrected p-values are mostly higher for the reduced universe and they are higher still, and always exceed 0.50, when the Sharpe ratio criterion is adopted. These findings suggest that calendar effects are not present in a recent sample of returns on a portfolio of liquid small firm stocks.

7. Conclusion

In their systematic study of calendar effects in the DJIA index, Lakonishok and Smidt (1988) conclude “In summary, DJIA returns are persistently anomalous over a 90-year period around the turn of the week, around the turn of the month, around the turn of the year, and around holidays”. They explicitly acknowledge the potential dangers of data-mining effects and state that “The possibility that these particular anomalies could have occurred by chance cannot be excluded, but this is very unlikely”.

In this paper we have shown that, in fact, when assessed in the context of either a large universe or a restricted universe of calendar rules that could plausibly have been considered by investors and academics with access to our data set, the strength of the evidence on calendar anomalies looks much weaker. Using reality check p-values that adjust for the effects of data mining, no calendar rule appears to be capable of outperforming the benchmark market index. This is true in all of the individual sample periods, in the out-of-sample experiment with the DJIA and S&P 500 Futures data, and in the full sample using a century of daily data.

We find it suggestive that the single most significant calendar rule, namely the Monday effect, has indeed been identified in the empirical literature. This is probably not by chance and it indicates that very substantial search for calendar regularities has been carried out by the financial community.[22] It is particularly noteworthy that when the Monday effect is examined in the context of as few as 20 day-of-the-week trading rules, and during the sample period originally used to find the Monday effect, its statistical significance becomes questionable. Subsequent to its appearance, various theories have attempted to explain the Monday effect without much success. Thaler (1987b) lists a number of institutional and behavioral reasons for

[22] Brock et al. (1992) provide a very fitting quote from Merton (1987, p. 104) “All this fits well with what the cognitive psychologists tell us is our natural individual predilection to focus, often disproportionately so, on the unusual”.

calendar effects. Our study suggests that the solution to the puzzling abnormal Monday effect actually lies outside the specificity of Mondays and rather has to do with the very large number of rules considered besides the Monday rule.

Blame for data mining cannot and must not be laid on individual researchers. Data exploration is an inevitable part of the scientific discovery process in the social sciences and is indeed capable of revealing unsuspected regularities when these exist. The danger lies in confusing apparent with real effects. Many researchers go to great lengths in attempts to avoid this pitfall. Ultimately, however, it has been extremely difficult for a researcher to account for the effects the cumulated "collective knowledge" of the investment community may have had on a particular study. The methods used here provide a principled way to conduct research as a sequential process in which new studies build on evidence from earlier papers.

In evaluating a body of research it is important to assess the results not by treating the individual studies as independent observations but by explicitly accounting for their cross-dependencies. In doing this, one should not be overwhelmed by the sheer amount of empirical evidence. This is sometimes difficult because the dependencies between results in different studies are unknown. For example, Michael Jensen, in his introduction to the 1978 volume of the Journal of Financial Economics on market anomalies writes "Taken individually many scattered pieces of evidence ... don't amount to much. Yet viewed as a whole, these pieces of evidence begin to stack up in a manner which make a much stronger case for the necessity to carefully review both our acceptance of the efficient market theory and our methodological procedures" (p. 95). Our results show that even supposedly strongly supported empirical phenomena may not stand up to closer scrutiny. There may well be many other such surprises waiting for researchers trying to establish our degree of knowledge about economic phenomena.

Acknowledgements

In writing this paper we have benefited from the comments of Frank Diebold, Bruce Lehmann, Andy Lo, Richard Roll, William Ziemba and seminar participants at University of Bristol, University of California at Los Angeles and at San Diego, EC2 conference on Forecasting in Stockholm, London School of Economics, Monash University, Nuffield College, University of Technical Science in Sydney, University of Warwick, and Western Finance Association meetings in Santa Monica. We also thank an anonymous referee for many helpful suggestions. The authors are grateful to QuantMetrics

R&D Associates, LLC of San Diego, California for making available its patented Reality Check software, US Patent 5,893,069.

References

Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856.

Ariel, R.A., 1987. A monthly effect in stock returns. Journal of Financial Economics 17, 161–174.

Ball, R., Bowers, J., 1988. Daily seasonals in equity and fixed-interest returns: Australian evidence and tests of plausible hypotheses. In: Dimson, E. (Ed.), Stock Market Anomalies. Cambridge University Press, Cambridge.

Brock, W., Lakonishok, J., LeBaron, B., 1992. Simple technical trading rules and the stochastic properties of stock returns. Journal of Finance 47, 1731–1764.

Campbell, J.Y, Lo, A.W., MacKinlay, C.R., 1997. The Econometrics of Financial Markets. Princeton University Press, Princeton, NJ.

Campbell, R., Vinci, T., 1983. Novel confirmation. British Journal for the Philosophy of Science 34, 315–341.

Christiano, L.J., 1992. Searching for a break in GNP. Journal of Business and Economic Statistics 10, 237–249.

Cross, F., 1973. The behavior of stock prices on Fridays and Mondays. Financial Analysts Journal November–December, 67–69.

Diebold, F.X., Chen, C., 1996. Testing structural stability with endogenous break point: a size comparison of analytic and bootstrap procedures. Journal of Econometrics 70, 221–241.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–265.

Dimson, E., 1988. Stock Market Anomalies. Cambridge University Press, Cambridge.

Fields, M.J., 1931. Stock prices: A problem in verification. Journal of Business 4, 415–418.

Fields, M.J., 1934. Security prices and stock exchange holidays in relation to short selling. Journal of Business 7, 328–338.

Flannery, M.J., Protopapadakis, A.A., 1988. From T-bills to common stocks: Investigating the generality of intra-week return seasonality. Journal of Finance 33, 431–450.

Foster, F.D., Smith, T., Whaley, R.E., 1997. Assessing goodness-of-fit of asset pricing models: the distribution of the maximal R^2. Journal of Finance 52, 591–607.

French, K.R., 1980. Stock returns and the weekend effect. Journal of Financial Economics 8, 55–70.

Gibbons, M.R., Hess, P., 1981. Day of the week effects and asset returns. Journal of Business 54, 579–596.

Gonçalves, S., White, H., 2000. The bootstrap of the mean for heterogeneous dependent processes. Economic Theory, forthcoming.

Haugen, R.A., Lakonishok, J., 1988. The Incredible January Effect. Dow Jones-Irwin, Homewood, IL.

Hensel, C., Ziemba, W., 1996. Investment results from exploiting turn-of-the-month effects. Journal of Portfolio Management 22, 17–23.

Jaffe, J., Westerfield, R., 1985. The week-end effect in common stock returns: the international evidence. Journal of Finance 40, 433–454.

Jensen, M.C., 1978. Some anomalous evidence regarding market efficiency. Journal of Financial Economics 6, 95–101.

Jones, C.P., Pearce, D.K., Wilson, J.W., 1987. Can tax-loss selling explain the January effect? A note. Journal of Finance 42, 453–461.

Kahn, J.A., Landsburg, S.E., Stockman, A.C., 1996. The positive economics of methodology. Journal of Economic Theory 68, 64–76.

Keim, D.B., 1983. Size-related anomalies and stock return seasonality: further empirical evidence. Journal of Financial Economics 12, 13–32.

Keim, D.B., Stambaugh, R.F., 1984. A further investigation of the weekend effect in stock returns. Journal of Finance 39, 819–835.

Lakonishok, J., Levi, M., 1982. Weekend effects on stock returns. Journal of Finance 37, 883–889.

Lakonishok, J., Maberly, E., 1990. The weekend effect: trading patterns of individual and institutional investors. Journal of Finance 40, 231–243.

Lakonishok, J., Smidt, S., 1984. Volume and turn-of-the-year behavior. Journal of Financial Economics 13, 435–456.

Lakonishok, J., Smidt, S., 1988. Are seasonal anomalies real? A ninety-year perspective. Review of Financial Studies 1 (4), 403–425.

Leamer, E., 1978. Specification Searches. Wiley, New York.

Leroy, S.F., 1973. Risk aversion and the martingale property of stock prices. International Economic Review 14, 436–446.

Levi, M., 1988. Weekend effects in stock market returns: an overview. In: Elroy, D. (Ed.), Stock Market Anomalies. Cambridge University Press, Cambridge.

Lo, A.W., MacKinlay, A.C., 1990. Data-mining biases in tests of financial asset pricing models. The Review of Financial Studies 3, 431–467.

Lucas, R.E., 1978. Asset prices in an exchange economy. Econometrica 46, 1429–1445.

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

Perron, P., 1989. The great crash, the oil price shock, and the unit root hypothesis. Econometrica 57, 1361–1401.

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

Reinganum, M.R., 1983. The anomalous stock market behavior of small firms in January: empirical tests for tax-loss selling effects. Journal of Financial Economics 12, 89–104.

Ritter, J.R., Chopra, N., 1989. Portfolio rebalancing and the turn of the year effect. Journal of Finance 44, 149–166.

Rogalski, R.J., 1984. New findings regarding day-of-the-week returns over trading and nontrading periods: a note. Journal of Finance 39, 1603–1614.

Roll, R., 1983. Vas ist Das? The turn-of-the-year effects and the return premia of small firms. Journal of Portfolio Management Winter, 18–28.

Rozeff, M.S., Kinney Jr., W.R., 1976. Capital market seasonality: the case of stock returns. Journal of Financial Economics 3, 379–402.

Samuelson, P.A., 1965. Proof that properly anticipated prices fluctuate randomly. Industrial Management Review 6, 41–49.

Savin, N.E., 1984. Multiple hypothesis testing. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. 2. North-Holland, Amsterdam (Chapter 14).

Smirlock, M., Starks, L., 1986. Day of the week and intraday effects in stock returns. Journal of Financial Economics 17, 197–210.

Sullivan, R., Timmermann, A., White, H., 1999. Data-mining, technical trading rule performance, and the bootstrap. Journal of Finance 54, 1647–1691.

Thaler, R.H., 1987a. Anomalies: the January effect. Journal of Economic Perspectives 1 (1), 197–201.

Thaler, R.H., 1987b. Anomalies: weekend, holiday, turn of the month, and intraday effects. Journal of Economic Perspectives 1 (2), 169–178.

Twain, M., 1894. (Samuel Clemens). The Tragedy of Pudd'nhead Wilson, reprint 1996. Oxford University Press, New York (Chapter 13).

Wang, K., Li, Y., Erickson, J., 1997. A new look at the Monday effect. Journal of Finance 52, 2171–2187.

West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.

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