Secular Mean Reversion and Long-Run Predictability of the Stock Market*

Valeriy Zakamulin

This revision: December 18, 2015

Abstract

Empirical financial literature documents the evidence of mean reversion in stock prices and the absence of out-of-sample return predictability over periods shorter than 10 years. The goal of this paper is to test the random walk hypothesis in stock prices and return predictability over periods longer than 10 years. Specifically, using 141 years of data, this paper begins by performing formal tests of the random walk hypothesis in the prices of the real S&P Composite Index over increasing time horizons up to 40 years. Even though our results cannot support the conventional wisdom which says that the stock market is safer for long-term investors, our findings speak in favor of the mean reversion hypothesis. In particular, we find statistically significant in-sample evidence that past 15-17 year returns are able to predict future 15-17 year returns. This finding is robust to the choice of data source, deflator, and test statistic. The paper continues by investigating the out-of-sample performance of long-horizon return forecast based on the mean-reverting model. These latter tests demonstrate that the forecast accuracy provided by the mean-reverting model is statistically significantly better than the forecast accuracy provided by the naive historical-mean model. Moreover, we show that the predictive ability of the mean-reverting model is economically significant and translates into substantial performance gains.

Key words: predictability, stock returns, long-run, random walk, mean reversion, bootstrap simulation

JEL classification: C12, C14, C22, G12, G14, G17.

The author is grateful to the following individuals for their helpful comments and suggestions regarding previous drafts of this paper: two anonymous referees; Brian Berry; Ed Easterling; Jochen Jungeilges; Ole Gjelberg; Steen Koekebakker; and the participants at the 2011 Computational and Financial Econometrics Conference (London, UK), the 2013 Fagkonferansen i Bedriftsøkonomiske Emner (FIBE Conference, Bergen, Norway), and the 2013 Forecasting Financial Markets Conference (Hannover, Germany). Any remaining errors in the manuscript are the author’s responsibility.

a.k.a. Valeri Zakamouline, School of Business and Law, University of Agder, Service Box 422, 4604 Kristiansand, Norway, Tel.: (+47) 38 14 10 39, Fax: (+47) 38 14 10 27, Valeri.Zakamouline@uia.no

1
1 Introduction

Until the late 1980s there was a widespread agreement in the academic community that stock prices follow a random walk. Indeed, a large body of empirical literature seemed to support this point of view (see Fama (1970) and Leroy (1982) for surveys). The efficient market hypothesis is strongly associated with the idea of a random walk in stock prices and loosely says that stock returns are unpredictable. However, during the late 1980s there appeared a series of papers where the authors challenged the random walk hypothesis (see, for example, Summers (1986), Campbell and Mankiw (1987), Fama and French (1988b), Lo and MacKinlay (1988), and Poterba and Summers (1988)). In particular, these authors considered the time series properties of stock returns over increasing time horizons up to 10 years and found the indications of mean reversion\(^1\) and return predictability. For example, Fama and French (1988b) discovered a substantial negative autocorrelation in returns over periods of 3-5 years and concluded that past 3-5 year returns are able to predict future 3-5 year returns. Poterba and Summers (1988) found that stock returns exhibit positive and statistically significant autocorrelation in returns over periods shorter than one year and negative, though not statistically significant at conventional levels (1\% or 5\%), autocorrelations over longer periods.

However, the conclusions reached in these earlier papers were strongly criticized on statistical grounds. For example, Kim, Nelson, and Startz (1991) demonstrated that due to the small-sample bias the statistical significance of the test statistics in Fama and French (1988b) and Poterba and Summers (1988) was overstated and there was no predictability of future 3-5 year returns on the basis of past 3-5 year returns. Similarly, Richardson and Stock (1989) and Richardson (1993) showed that correcting for the small-sample bias may reverse the results obtained by Fama and French (1988b) and Poterba and Summers (1988).

\(^1\)Mean reversion is an ambiguous concept and exists in several different forms. Most often, the concept of mean reversion can be expressed by the common investment wisdom which says that “over time markets tend to return to the mean”. For example, when stocks go too far in one direction, they will eventually come back. Another type of mean reversion, which is studied in this paper, implies that the reversion is much more than just returning back to the mean. In reality the movement is far greater. This type of mean reversion incorporates another common investment wisdom which says that “an excess in one direction will lead to an excess in the opposite direction”. That is, when stocks go too far in one direction, they will not just come back to the mean, but overshoot in the opposite direction. For example, a period of above average returns tends to be followed by a period of below average returns and vice versa. Throughout the paper, the term “period” is used to denote the period of mean reversion. The term “horizon” is mainly used to denote the average length of a complete cycle of reversion which consists of two periods: a period of higher than average returns and a period of lower than average returns (or vice versa).
Apparently, the statistical power of earlier tests was insufficient to reject the random walk hypothesis. Jegadeesh (1991) suggested a new more powerful test and detected statistically significant evidence of mean reversion in stock prices (over periods of 4-8 years). In addition, Jegadeesh found evidence of mean reversion not only for the US stock market, but also for the UK stock market. Later on, based on a panel approach, Balvers, Yangru, and Gilliland (2000) found statistically significant evidence of mean-reverting behavior (over periods of 3-3.5 years) in many international stock indices. Thus, mean reversion in stock prices seems to be an international phenomenon. Using the same technique as in Balvers et al. (2000), Gropp (2003) and Gropp (2004) found statistically significant evidence of mean reversion in the prices of portfolios of small cap stocks (over periods of 3.5 years) and industry-sorted portfolios (over periods of 4.5-8 years). Moreover, Balvers et al. (2000), Gropp (2003), and Gropp (2004) showed that parametric contrarian investment strategies that exploit mean reversion outperform buy-and-hold and standard contrarian strategies. This provides further support for the mean reversion findings in these papers.

Thus, nowadays the evidence of mean reversion in the prices of some stock portfolios over periods of 3-8 years seems to have been manifested. In contrast, the predictability of stock returns is still a source of heated debate within the academic community. Earlier papers, that demonstrated the existence of in-sample stock return predictability, include, among others, Fama (1981), Campbell (1987), Fama and French (1988b), Fama and French (1988a), Campbell and Shiller (1988), and Fama and French (1989). Again, the conclusions reached in these earlier papers were strongly criticized on statistical grounds. For example, Richardson and Stock (1989) and Nelson and Kim (1993) pointed to the small-sample bias problem, whereas Cavanagh, Elliott, and Stock (1995), Stambaugh (1999), and Lanne (2002) pointed to a neglected near unit root problem. Responding to the critique, Torous, Balkanov, and Yan (2004), Lewellen (2004), Rapach and Wohar (2005), and Campbell and Yogo (2006) developed new tests, that are free from the discovered flaws in the earlier tests, and again found some evidence of in-sample predictability. Yet, Bossaerts and Hillion (1999), Goyal and Welch (2003), and Welch and Goyal (2008) demonstrated that, despite evidence of in-sample predictability, the predictive models have no out-of-sample forecasting power. These authors therefore argued that in-sample predictability appears as a result of data mining. It should be noted, however, that in all these tests the longest forecast horizon
was 10 years. Consequently, the results of these tests imply that the predictive models fail to demonstrate statistically significant predictive ability over short-term and medium-term horizons.

To the best knowledge of the author, no one has ever tested the random walk hypothesis in stock prices over periods longer than 10 years. Yet, anecdotal evidence suggests the presence of mean reversion in stock prices over very long horizons. Probably the best known evidence is presented by Siegel (2002) in his famous book “Stocks for the Long Run”. In particular, using a historical sample that covers nearly 200 years, Siegel computed the standard deviation of average real annual returns on a broad US stock market index over increasing horizons up to 30 years. Siegel found that the standard deviation declines far faster than predicted by the random walk hypothesis. This led many to conclude that stocks are less risky in the long run. However, so far there have been no studies conducted on whether the decline in the standard deviation over very long horizons is statistically significant.

Another well-known anecdotal evidence, explicitly related to the mean reversion in stock prices over very long horizons, suggests the existence of long-lasting alternating periods of bull and bear markets. These long-lasting bull and bear markets are often termed as “secular” bull and bear markets. Alexander (2000), Easterling (2005), Rogers (2005), Katsenelson (2007), and Hirsch (2012), among others, analyzed the dynamics of the real S&P Composite Index since 1870 and found the indications of existence of secular stock market trends that last from 5 to 25 years, with average duration of about 15 years. Motivated by the seeming regularity in the reversion of secular trends, some authors made quite successful forecasts for the long-run US stock market outlook. For example, Alexander (2000) predicted that during the period from 2000 to 2020 the stock market will not beat the money market. So far, this forecast seems to come true. This anecdotal evidence suggests, among other things, that a price change over a given long-run period may be able to predict the price change over the subsequent long-run period. This idea motivates to re-examine the predictive performance of the model introduced by Fama and French (1988b). Even though Kim et al. (1991) demonstrated that this model has no predictive power on increasing periods up to 10 years, as far as the author knows, no one has ever tested this model on periods longer than 10 years. This paper aims to fill these gaps in scientific knowledge about the stock
market dynamics over very long horizons.

The first contribution of this paper is to provide, for the first time, statistically significant evidence against the random walk hypothesis over periods longer than 10 years. Even though our results cannot support the anecdotal evidence which says that the stock market is safer for long-term investors, our findings do speak in favor of mean reversion in stock prices over periods of 15-17 years. In particular, using the whole sample of data, we find statistically significant evidence that a given change in price over 15-17 years tends to be reversed over the next 15-17 years by a predictable change in the opposite direction. This implies the existence of in-sample long-horizon predictability. Since the conventional wisdom says that in-sample evidence of stock return predictability might be a result of data mining, we investigate the performance of out-of-sample long-horizon return forecast. Besides the mean-reverting model, we investigate the out-of-sample forecast accuracy of a few other competing models which employ, as a predictor for long-horizon returns, the cyclically adjusted price-to-earnings ratio, the price-to-dividends ratio, and the long-term bond yield.

The second contribution of this paper is to demonstrate that the out-of-sample long-horizon forecasts provided by the mean-reverting model and the models that employ the price-to-earnings and price-to-dividends ratios are statistically significantly better than the forecast provided by the historical-mean model. It is worth emphasizing that Welch and Goyal (2008) also used the price-to-earnings and price-to-dividends ratios in their study and found that these models have no predictive ability over forecast horizons up to 5 years. Our results therefore advocate that these models do have predictive ability, but over forecast horizons longer than 10 years. We also demonstrate that the advantages of the models, that show the predictive ability, translate into significant performance gains. For example, we estimate that risk-averse investors would be willing to pay from 30 to 77 basis points fees per year to switch from the historical-mean model to a model with a superior forecast accuracy. Moreover, our tests suggest that over the recent past the out-of-sample forecast accuracy provided by the mean-reverting model was substantially better than that provided by the competing models. In addition, we find that the mean-reverting model delivers the highest performance gains when investors have to make long-term allocation decisions.

The rest of the paper is organized as follows. Section 2 presents the data for our study,
namely, the returns on the real Standard and Poor’s Composite Stock Price Index over the period from 1871 to 2011. In Section 3 we perform the tests of the random walk hypothesis using the S&P Composite Index. In Section 4 we study the out-of-sample predictability of multi-year returns on the S&P Composite Index. Finally, Section 5 summarizes and concludes the paper.

2 The Data

The data for the study in this paper are the annual log real returns on a broad US stock market index for the period from 1871 to 2011. The returns are adjusted for dividends and computed using the real (i.e., inflation-corrected) Standard and Poor’s Composite Stock Price Index data and corresponding dividend data. The inflation adjustment is done using the Consumer Price Index (CPI) for the US. All the data are provided by Robert Shiller.2 The Standard and Poor’s Composite Stock Price index is a value-weighted stock index. The index for the period from 1871 to 1925 is constructed using the Cowles Commission Common Stock Index series. From 1926 to the present, the index data come from various reports of the Standard and Poor’s. From 1957 this index is identical to the Standard and Poor’s 500 Index which is intended to be a representative sample of leading companies in leading industries within the US economy. Stocks in the index are chosen for market size, liquidity, and industry group representation. For more details about the construction of the index and its dividend series see Shiller (1989), Chapter 26. Formally, let $(p_0, p_1, \ldots, p_n)$ be observations of the natural log of an inflation-corrected stock index price over $n + 1$ years. Denote the one-year log return during year $t$, $1 \leq t \leq n$, by

$$r_t = p_t - p_{t-1}.$$  

The resulting sample of $n$ return observations is $(r_1, r_2, \ldots, r_n)$. The probability distribution of $r_t$ is unknown, yet it is well-documented that stock returns are non-normal and heteroscedastic.

In order to check the robustness of findings, in particular, to see whether the results of  

---

2See http://www.econ.yale.edu/~shiller/data.htm. The real dividend adjusted annual return series on the index are readily available in the file chapt26.xls. Robert Shiller stopped maintaining his database in 2012.
the testing the random walk hypothesis depend on a specific historical period, we divide the
total sample period from 1871 to 2011 (141 annual observations) in two equal overlapping
sub-samples, the first one is from 1871 to 1956 and the second one is from 1926 to 2011.\footnote{The reasons for using overlapping sub-samples are as follows. First, in order to perform statistical tests on the presence of long-run mean reversion we need longer time series. Second, the starting point of our second sub-sample coincides with the starting point of the database of historical stock market data provided by the Center for Research in Security Prices. Therefore the data on the stock market returns over the second sub-sample is much more accurate than that over the first sub-sample.}
Both of these sub-samples cover a span of 86 years. Table 1 presents the descriptive statistics
for the annual stock index returns, \( r_t \), for the total sample and both sub-samples. Table 2
reports the results of the \( t \)-test on difference in mean returns and \( F \)-test on difference in
standard deviations between the first and the second sub-sample. The descriptive statistics
and the results of the tests suggest that the mean and variance of returns on the index were
more or less stable during the total sample. Specifically, using a \( t \)-test for equal means we
cannot reject the hypotheses that the mean returns are alike in both sub-samples. Similarly,
using an \( F \)-test for equal variances we cannot reject the hypotheses that the variances are
alike in both sub-samples. All the series exhibit negative skewness and positive excess
kurtosis which indicates a deviation from normality. Observe also that the return series
during the overall sample period exhibits a statistically significant negative autocorrelation
at lag 2 (at the 5\% level). There are no other indications of serial dependence in the return
series.

3 Testing the Random Walk Hypothesis

3.1 Methodology

One of the main questions we want to study in this paper is whether the log of the real
S&P Composite Stock Price Index follows a random walk. To answer this question we
perform two well-known tests. The first test is based on the examination of the first-order
autocorrelation function of \( k \)-year returns. This test is used by, for example, Fama and
French (1988b), Fama and French (1989), and Fama (1990) and based on the computation
of the following test statistic

\[
AC1(k) = \frac{\text{Cov}(r_{t, t+k}, r_{t-k, t})}{\sqrt{\text{Var}(r_{t, t+k})\text{Var}(r_{t-k, t})}}.
\]
| Statistics     | Sample period          |       |       |       |
|---------------|------------------------|-------|-------|-------|
|               | 1871-2011              | 1871-1956 | 1926-2011 |
| Mean, %       | 6.28                   | 6.91   | 6.24  |
| Std. dev., %  | 17.14                  | 17.76  | 18.77 |
| Skewness      | -0.57                  | -0.48  | -0.59 |
| Kurtosis      | 3.41                   | 3.32   | 3.24  |
| \( \rho_1 \)  | 0.02                   | 0.04   | 0.04  |
| \( \rho_2 \)  | \textbf{-0.19}         | -0.20  | -0.18 |
| \( \rho_3 \)  | 0.09                   | 0.07   | 0.02  |
| \( \rho_4 \)  | -0.08                  | -0.18  | -0.14 |
| \( \rho_5 \)  | -0.11                  | -0.10  | -0.07 |
| \( \rho_6 \)  | 0.10                   | 0.12   | 0.11  |
| \( \rho_7 \)  | 0.10                   | 0.06   | 0.16  |
| \( \rho_8 \)  | -0.08                  | -0.15  | -0.02 |
| \( \rho_9 \)  | -0.06                  | -0.04  | 0.04  |
| \( \rho_{10} \)| 0.02                   | 0.06   | 0.06  |
| \( \rho_{11} \)| 0.02                   | 0.06   | -0.07 |
| \( \rho_{12} \)| -0.08                  | -0.04  | -0.10 |
| \( \rho_{13} \)| -0.09                  | -0.15  | -0.19 |
| \( \rho_{14} \)| 0.03                   | 0.03   | -0.14 |
| \( \rho_{15} \)| -0.09                  | -0.07  | -0.02 |
| \( \rho_{16} \)| -0.09                  | -0.09  | 0.06  |
| \( \rho_{17} \)| 0.06                   | 0.16   | -0.02 |
| \( \rho_{18} \)| -0.08                  | -0.06  | -0.13 |
| \( \rho_{19} \)| -0.17                  | -0.11  | -0.21 |
| \( \rho_{20} \)| 0.06                   | 0.09   | -0.07 |

Table 1: Descriptive statistics of the annual log real returns on the Standard and Poor’s Composite Stock Price Index. \( \rho_k \) denotes the autocorrelation between \( r_t \) and \( r_{t+k} \). For each \( \rho_k \) we test the hypothesis \( H_0 : \rho_k = 0 \). Bold text indicates values that are statistically significant at the 5% level.

| Test statistic                     | P-value |
|-----------------------------------|---------|
| \( t \)-test on difference in mean returns | 0.24    | 0.81    |
| \( F \)-test on difference in standard deviations | 0.89    | 0.61    |

Table 2: Results of the \( t \)-test on difference in mean returns and \( F \)-test on difference in standard deviations between the first and the second sub-sample.
where $r_{i,j}$ is the compounded return from year $i$ to year $j$, $r_{i,j} = p_j - p_i$, $Cov(\cdot, \cdot)$ and $Var(\cdot)$ denote the covariance and variance respectively, and $AC1(k)$ stands for the first-order autocorrelation function of $k$-year returns. The second test is based on the examination of the variance ratio. This test is very popular and used by Cochrane (1988), Lo and MacKinlay (1988), Poterba and Summers (1988), and many other afterwards. The test is based on the computation of the following test statistic

$$VR(k) = \frac{Var(r_{t,t+k})}{k \times Var(r_t)},$$

(2)

Both the tests are motivated by the notion that if the stock returns are independent and identically distributed, then the first-order autocorrelation function is zero and the variance ratio is unity irrespective of the number of years $k$. In other words, without serial dependence in data, the variance of $k$-year returns equals $k$ times the variance of one-year returns and there is no correlation between two successive non-overlapping $k$-year returns. The null hypothesis of a random walk is rejected if the first-order autocorrelation is significantly different from zero or the variance ratio is significantly different from unity.

We want to compute the variance ratio $VR(k)$ for return horizons $k$ from 20 to 40 years and the first-order autocorrelation $AC1(k)$ for periods from 10 to 20 years (note that in the latter case we also study serial dependence in data over time horizons from 20 to 40 years). The fundamental problem with these computations is that we have only a few non-overlapping intervals of length 20-40 years. Therefore in the computations of the two test statistics we employ overlapping intervals (rolling $k$-year periods). To compute $AC1(k)$ we regress $k$-year returns $r_{t,t+k}$ on lagged $k$-year returns $r_{t-k,t}$. That is, we run the following regression

$$r_{t,t+k} = a(k) + b(k) r_{t-k,t} + \varepsilon_{t,t+k}.$$  

(3)

Observe that the slopes of the regression, $b(k)$, $k \in [10, 20]$, are the estimated autocorrelations of $k$-year returns, $AC1(k)$. The variance of $k$-year returns is computed as

$$Var(r_{t,t+k}) = E \left[ (r_{t,t+k} - E[r_{t,t+k}])^2 \right].$$

The use of overlapping returns leads to some potentially very serious econometric issues
which are commonly termed as “small-sample bias”. In particular, when it comes to the estimation of regression (3), there are two econometric problems. First, the estimates for the slope coefficients are biased. The sources of this bias in the estimation of autocorrelation are described in details by Orcutt and Irwin (1948) and Marriott and Pope (1954). More specifically, these authors show that an estimate of autocorrelation obtained using overlapping blocks of data is downward biased. Therefore, the estimates must be corrected for the bias. The second problem is that the standard errors of estimation using overlapping blocks of data are also downward biased, see, for example, Nelson and Kim (1993). Both biases work in the direction of making the values of t-statistic too large so that standard inference may indicate dependence in return series even if none is present.4

Similarly, the estimate for the variance of multi-year returns, \( \text{Var}(r_{t,t+k}) \), is downward biased when one uses overlapping blocks of data.5 As an immediate consequence, the estimate for the variance ratio \( VR(k) \) becomes also downward biased. Therefore, the estimates for \( VR(k) \) must be corrected for the bias. In addition, since the estimate for \( VR(k) \) is a random variable, for the purpose of statistical inference we need to know the probability distribution of \( VR(k) \). This is necessary in order to be able to estimate standard errors and confidence intervals for \( VR(k) \). This is also necessary for performing hypothesis tests about the value of \( VR(k) \).

When the nature of the data generating process is unknown, it is generally not possible to tackle the econometric problems described above. However, in the context of the null hypothesis our goal is primarily to test whether or not stock returns are distributed independently of their ordering in time. Since under the null there is no dependence in return series, in order to estimate the significance level and perform the bias correction of the test statistics, we follow closely Kim et al. (1991) and Nelson and Kim (1993) where the authors employ the randomization method. The randomization method is introduced by Fisher

\footnote{Specifically, in case where returns are independent, using overlapping blocks of data produces a negative value of the estimated slope coefficient in regression (3). In addition, the standard error of estimation of the slope coefficient using overlapping blocks of data is downward biased. That is, the estimated standard error is smaller than it is in reality. The higher the overlap, the more negative the slope coefficient and the smaller the estimated standard error. As a result, the values of t-statistic may falsely indicate the presence of dependence in return series when none is present.}

\footnote{Note that this is also related to the second econometric problem in the estimation of regression (3). That is, the standard errors of estimation of slope coefficients using overlapping blocks of data are downward biased, because the estimates for variance using overlapping blocks of data are downward biased. For the sake of motivation, consider what happens to the estimate for \( \text{Var}(r_{t,t+k}) \) when \( k \to n \). Obviously in the limit, when the length \( k \) converges to the sample length, there is only one available block of data to estimate \( \text{Var}(r_{t,t+k}) \). Therefore, regardless of the nature of the data generating process, \( \text{Var}(r_{t,t+k}) \to 0 \) as \( k \to n \).}
(1935) and provides a very general and robust approach for computing the probability of obtaining some specific value for an estimator under the null hypothesis of no dependence. We refer the interested readers to Noreen (1989) and Manly (1997) for extensive discussion of the randomization tests. In a nutshell, randomization consists of reshuffling the data to destroy any dependence and then recalculating the test statistics for each reshuffling in order to estimate its distribution under the null hypothesis of no dependence. The great advantage of the randomization method is that it is very simple and no assumptions are made about the actual distribution of stock returns.

To be more specific, consider the estimation of the significance level and the bias correction of the estimate for the autocorrelation of $k$-year returns $AC1(k)$. First, we run regression (3) using the original series $(r_1, r_2, \ldots, r_n)$ to obtain the actual historical estimates for $AC1(k)$. Then we randomize the original series to get a permutation $(r_1^*, r_2^*, \ldots, r_n^*)$. This is repeated 10,000 times, each time running regression (3) and obtaining an estimate for $AC1^*(k)$. In this manner we estimate the sampling distribution of $AC1(k)$ under the null hypothesis. Finally, to estimate the significance level for some particular $k$, we count how many times the computed value for $AC1^*(k)$ after randomization falls below the value of the actual historical estimate for $AC1(k)$. In other words, under the null hypothesis we compute the probability of obtaining a more extreme value for the autocorrelation of $k$-year returns than the actual historical estimate. Note that in this manner we compute p-values of one-tailed test. The estimation bias is defined as the difference between the expected and the true value of $AC1^*(k)$. Since the true value is zero under the null hypothesis, the bias correction is done by subtracting the expected value of $AC1^*(k)$ from the actual historical estimate for $AC1(k)$. That is, the bias adjusted values of the first-order autocorrelation of $k$-year returns are computed as $AC1(k) - E[AC1^*(k)]$.

The estimation of the significance level and the bias correction of the estimate for the variance ratio $VR(k)$ is done in a similar manner. First, we use the original series to obtain the actual historical estimates for $VR(k)$. Then we randomize the series and compute $VR^*(k)$ to obtain the sample distribution under the null hypothesis. Finally, to estimate the significance level for some particular $k$, we count how many times the computed value for $VR^*(k)$ after randomization falls below the value of the actual historical estimate of $VR(k)$. The estimation bias in this case is given by $E[VR^*(k)] - 1$ since the true value is
unity under the null hypothesis. Finally, the bias adjusted values of the variance ratio are computed as $V_R(k) - E[V_R^*(k)] + 1$.

There is ample evidence that the series of stock returns is heteroscedastic, see, for example, Officer (1973) and Schwert (1989). In particular, many researchers document that the variance of stock returns is not constant, but time-varying. To see whether a change in the variance of returns might affect the sampling distribution of a test statistic, we follow closely Kim et al. (1991) and Nelson and Kim (1993) and use the stratified randomization. In the stratified randomization method the total sample (or a sub-sample) is divided into several separate bins (urns) and the randomization is performed within each bin. Such a stratified randomization allows us to see whether the sampling distribution of a test statistic is sensitive to the particular pattern of heteroscedasticity that occurred historically.

### 3.2 Empirical Results

Figure 1 plots the sample first-order autocorrelations and variance ratios of the $k$-year returns on the Standard and Poor’s Composite Stock Price Index. The first-order autocorrelations and variance ratios are computed according to formulas (1) and (2) respectively using overlapping blocks of data. Apparently, for the total sample and both the sub-samples the first-order autocorrelations and variance ratios generally decline with increasing $k$. The indications against the null hypothesis on very long horizons are stronger (i.e., the declines in the first-order autocorrelations and variance ratios are larger) for the second sub-sample (1926 to 2011) than for the total sample or the first sub-sample (1871 to 1956).

Recall, however, that the estimates for both the first-order autocorrelations and the variance ratios presented in Figure 1 are downward biased. As a matter of fact, under the null hypothesis of no serial dependence in return series we expect to see declining first-order autocorrelations and variance ratios with increasing $k$. In order to find out whether the observed declines are statistically significantly different from the expected declines under the null hypothesis, and in order to correct for the estimation bias under the null, we perform the randomization method with and without the stratification. These results are reported in Tables 3 and 4 which show the estimates for the bias-adjusted first-order autocorrelations and variance ratios, respectively, with corresponding p-values. The estimates are based on 10,000 reshuffles and computed using different numbers of bins in the stratification. The
Figure 1: The sample first-order autocorrelations (top panel) and variance ratios (bottom panel) for the $k$-year log real returns on the Standard and Poor's Composite Stock Price Index. Neither the first-order autocorrelations nor the variance ratios are adjusted for the estimation bias.
number of bins varies from 1 (no stratification) to 5.

Without the stratification (that is, when the number of bins equals to one) both the test statistics suggest that the return series over the total sample (1871 to 2011) and the second sub-sample (1926 to 2011) exhibit clear evidence against the random walk on horizons of about 30-40 years. In particular, for the overall sample the values of the first-order autocorrelation are statistically significantly negative at the 5% level at periods 12-20 years (which indicates dependence over 24-40 year horizons). In addition, the values of the variance ratio are statistically significantly below unity at the 5% level at horizons 30-34 years. Thus, both the test statistics present evidence against the null hypothesis over horizons of 30-34 years. For the second sub-sample the values of the first-order autocorrelation are statistically significantly negative at the 5% level at periods 15-18 years (which indicates dependence over 30-36 year horizons), and the values of the variance ratio are statistically significantly below unity at horizons 34-36 years. For the first sub-sample the evidence against the random walk is weaker. Yet, if we use the 10% significance level, then we can reject the null hypothesis of no dependence in return series at several horizons.

Further, our results suggest that accounting for heteroscedasticity in stock returns does not influence the outcomes of the randomization tests on the first-order autocorrelations of \(k\)-year returns. Regardless of the number of bins in the stratified randomization, the first-order autocorrelation of \(k\)-year returns remains statistically significantly different from zero at the 5% level over periods of 15-17 years for the total sample and the second sub-sample. In contrast, stratification of the sample weakens the evidence against the null hypothesis for the value of the variance ratio. In particular, for the total sample and the stratification with either 2, 4, or 5 bins, the variance ratio is not statistically significantly below unity at conventional levels. Similarly, for the second sub-sample and the stratification with either 3 or 5 bins the variance ratio is not significantly below unity at conventional levels. For the first sub-sample the variance ratio is not significantly below unity regardless of the number of bins in the stratified randomization.

Consequently, we do not have strong enough evidence to claim that the variance ratio decreases with increasing investment horizon. Even though without stratification the variance ratio over horizons of 30-34 years is statistically significantly below unity, stratification of the sample suggests that this effect can be attributed to the historical pattern.
| Period, years | 1       | 2       | 3       | 4       | 5       |
|--------------|---------|---------|---------|---------|---------|
| 10           | -0.23 (0.14) | -0.15 (0.22) | -0.19 (0.16) | -0.10 (0.29) | -0.06 (0.36) |
| 11           | -0.29 (0.09) | -0.21 (0.15) | -0.25 (0.10) | -0.15 (0.20) | -0.11 (0.26) |
| 12           | **-0.38 (0.04)** | -0.29 (0.08) | **-0.34 (0.05)** | -0.23 (0.10) | -0.18 (0.13) |
| 13           | **-0.43 (0.03)** | -0.34 (0.04) | **-0.39 (0.03)** | -0.28 (0.06) | -0.23 (0.07) |
| 14           | **-0.48 (0.02)** | -0.38 (0.03) | **-0.44 (0.02)** | -0.32 (0.03) | -0.27 (0.04) |
| 15           | **-0.50 (0.01)** | -0.40 (0.02) | **-0.46 (0.01)** | -0.34 (0.03) | -0.29 (0.03) |
| 16           | **-0.51 (0.01)** | -0.40 (0.02) | **-0.47 (0.01)** | -0.34 (0.02) | -0.28 (0.03) |
| 17           | **-0.51 (0.02)** | -0.40 (0.02) | **-0.46 (0.01)** | -0.34 (0.02) | -0.27 (0.02) |
| 18           | **-0.50 (0.02)** | -0.38 (0.03) | **-0.45 (0.02)** | -0.33 (0.03) | -0.25 (0.03) |
| 19           | **-0.47 (0.03)** | -0.34 (0.04) | **-0.41 (0.03)** | -0.29 (0.04) | -0.21 (0.05) |
| 20           | **-0.43 (0.05)** | -0.30 (0.07) | **-0.37 (0.05)** | -0.26 (0.06) | -0.17 (0.09) |

Panel B: First sub-sample 1871 to 1956

| Period, years | 1       | 2       | 3       | 4       | 5       |
|--------------|---------|---------|---------|---------|---------|
| 10           | -0.25 (0.17) | -0.11 (0.32) | -0.10 (0.33) | -0.06 (0.39) | -0.26 (0.18) |
| 11           | -0.28 (0.15) | -0.13 (0.29) | -0.12 (0.31) | -0.10 (0.33) | -0.31 (0.14) |
| 12           | -0.36 (0.09) | -0.20 (0.21) | -0.20 (0.22) | -0.19 (0.20) | -0.41 (0.08) |
| 13           | -0.31 (0.13) | -0.13 (0.29) | -0.15 (0.29) | -0.15 (0.26) | -0.37 (0.09) |
| 14           | -0.22 (0.23) | -0.03 (0.46) | -0.06 (0.41) | -0.08 (0.37) | -0.27 (0.12) |
| 15           | -0.13 (0.32) | 0.07 (0.60) | 0.01 (0.52) | -0.01 (0.48) | -0.16 (0.20) |
| 16           | -0.03 (0.47) | 0.19 (0.77) | 0.11 (0.68) | 0.10 (0.66) | -0.01 (0.47) |
| 17           | 0.02 (0.52) | 0.25 (0.83) | 0.15 (0.74) | 0.14 (0.72) | 0.09 (0.74) |
| 18           | -0.07 (0.41) | 0.19 (0.75) | 0.06 (0.60) | 0.06 (0.59) | 0.07 (0.68) |
| 19           | -0.16 (0.32) | 0.14 (0.66) | -0.03 (0.46) | -0.03 (0.46) | 0.05 (0.62) |
| 20           | -0.17 (0.32) | 0.16 (0.66) | -0.03 (0.45) | -0.03 (0.46) | 0.11 (0.70) |

Panel C: Second sub-sample 1926 to 2011

| Period, years | 1       | 2       | 3       | 4       | 5       |
|--------------|---------|---------|---------|---------|---------|
| 10           | -0.06 (0.41) | -0.05 (0.41) | 0.09 (0.67) | -0.03 (0.45) | 0.06 (0.64) |
| 11           | -0.16 (0.28) | -0.15 (0.27) | -0.00 (0.49) | -0.14 (0.27) | -0.02 (0.45) |
| 12           | -0.23 (0.20) | -0.23 (0.17) | -0.07 (0.35) | -0.22 (0.17) | -0.08 (0.33) |
| 13           | -0.32 (0.11) | -0.33 (0.09) | -0.17 (0.19) | -0.33 (0.08) | -0.16 (0.18) |
| 14           | -0.41 (0.06) | -0.42 (0.05) | -0.27 (0.08) | -0.42 (0.03) | -0.23 (0.07) |
| 15           | **-0.46 (0.04)** | **-0.47 (0.03)** | **-0.34 (0.05)** | **-0.48 (0.02)** | **-0.29 (0.03)** |
| 16           | **-0.53 (0.03)** | **-0.54 (0.01)** | **-0.42 (0.03)** | **-0.54 (0.01)** | **-0.37 (0.01)** |
| 17           | **-0.53 (0.04)** | **-0.54 (0.02)** | **-0.45 (0.03)** | **-0.55 (0.01)** | **-0.41 (0.01)** |
| 18           | **-0.50 (0.06)** | **-0.51 (0.03)** | **-0.43 (0.04)** | **-0.51 (0.01)** | **-0.42 (0.01)** |
| 19           | **-0.45 (0.09)** | **-0.46 (0.05)** | -0.40 (0.06) | **-0.46 (0.01)** | **-0.43 (0.01)** |
| 20           | **-0.40 (0.13)** | -0.43 (0.08) | -0.36 (0.09) | **-0.41 (0.03)** | **-0.42 (0.02)** |

Table 3: First-order autocorrelations of the k-year log real returns on the Standard and Poor’s Composite Stock Price Index (AC1(k)). These estimates are obtained using the randomization method with stratification (when the number of bins is greater than one). The estimates are corrected for the bias under the null hypothesis. The values in the brackets report the p-values of one-tailed test for the hypothesis $H_0: AC1(k) = 0$. Bold text indicates values that are statistically significant at the 5% level.
| Horizon, years | Panel A: Total sample 1871 to 2011 | Panel B: First sub-sample 1871 to 1956 | Panel C: Second sub-sample 1926 to 2011 |
|---------------|-----------------------------------|--------------------------------------|--------------------------------------|
| 20            | 0.75 (0.21) 0.87 (0.32) 0.80 (0.22) 0.93 (0.39) 1.00 (0.50)  | 0.66 (0.06) 0.73 (0.12) **0.64** (0.04) 0.80 (0.12) 0.87 (0.19) | 0.98 (0.48) 1.01 (0.51) 1.20 (0.79) 1.10 (0.67) 1.18 (0.81) |
| 22            | 0.71 (0.17) 0.84 (0.28) 0.77 (0.17) 0.90 (0.34) 0.97 (0.45)  | 0.67 (0.10) 0.82 (0.20) 0.84 (0.16) 0.90 (0.27) **0.69** (0.05) | 0.94 (0.38) 0.91 (0.38) 1.12 (0.71) 0.98 (0.47) 1.10 (0.69) |
| 24            | 0.66 (0.12) 0.79 (0.21) 0.71 (0.11) 0.86 (0.25) 0.93 (0.35)  | 0.65 (0.06) 0.81 (0.13) 0.82 (0.09) 0.87 (0.17) **0.66** (0.02) | 0.83 (0.26) 0.99 (0.46) 0.97 (0.43) 1.01 (0.53) 0.84 (0.12) |
| 26            | 0.62 (0.09) 0.76 (0.16) 0.67 (0.07) 0.83 (0.18) 0.90 (0.26)  | 0.70 (0.09) 0.85 (0.18) 0.86 (0.14) 0.90 (0.25) **0.69** (0.03) | 0.79 (0.18) 0.94 (0.34) 0.93 (0.31) 0.97 (0.40) 0.77 (0.06) |
| 28            | 0.58 (0.06) 0.73 (0.12) **0.64** (0.04) 0.80 (0.12) 0.87 (0.19) | 0.74 (0.13) 0.89 (0.25) 0.90 (0.23) 0.94 (0.33) **0.73** (0.04) | 0.84 (0.26) 0.99 (0.48) 0.98 (0.43) 1.01 (0.53) 0.87 (0.13) |
| 30            | 0.58 (0.05) 0.73 (0.11) **0.63** (0.03) 0.80 (0.10) 0.87 (0.16) | 0.79 (0.18) 0.94 (0.34) 0.93 (0.31) 0.97 (0.40) 0.77 (0.06) | 0.84 (0.26) 0.99 (0.48) 0.98 (0.43) 1.01 (0.53) 0.87 (0.13) |
| 32            | 0.59 (0.05) 0.75 (0.11) **0.64** (0.03) 0.81 (0.11) 0.88 (0.17) | 0.84 (0.06) 0.81 (0.16) **0.70** (0.04) 0.86 (0.18) 0.93 (0.28) | 0.84 (0.26) 0.99 (0.48) 0.98 (0.43) 1.01 (0.53) 0.87 (0.13) |
| 34            | 0.60 (0.05) 0.76 (0.11) **0.65** (0.03) 0.82 (0.11) 0.89 (0.17) | 0.74 (0.13) 0.89 (0.25) 0.90 (0.23) 0.94 (0.33) **0.73** (0.04) | 0.85 (0.21) 0.97 (0.40) 0.95 (0.34) 0.98 (0.42) 0.87 (0.08) |
| 36            | 0.62 (0.06) 0.78 (0.13) **0.67** (0.03) 0.83 (0.13) 0.91 (0.21) | 0.84 (0.16) 0.96 (0.33) 0.94 (0.27) 0.97 (0.33) 0.89 (0.08) | 0.84 (0.26) 0.99 (0.48) 0.98 (0.43) 1.01 (0.53) 0.87 (0.13) |
| 38            | 0.65 (0.08) 0.81 (0.16) **0.70** (0.04) 0.86 (0.18) 0.93 (0.28) | 0.85 (0.09) 0.84 (0.18) **0.72** (0.05) 0.87 (0.20) 0.95 (0.32) | 0.84 (0.26) 0.99 (0.48) 0.98 (0.43) 1.01 (0.53) 0.87 (0.13) |
| 40            | 0.67 (0.09) 0.84 (0.18) **0.72** (0.05) 0.87 (0.20) 0.95 (0.32) | 0.85 (0.14) 0.96 (0.31) 0.94 (0.24) 0.97 (0.30) 0.92 (0.11) | 0.84 (0.26) 0.99 (0.48) 0.98 (0.43) 1.01 (0.53) 0.87 (0.13) |

Table 4: Variance ratios of the $k$-year log real returns on the Standard and Poor’s Composite Stock Price Index ($VR(k)$). These estimates are obtained using the randomization method with stratification (when the number of bins is greater than one). The estimates are corrected for the bias under the null hypothesis. The values in the brackets report the $p$-values of one-tailed test for the hypothesis $H_0: VR(k) = 1$. Bold text indicates values that are statistically significant at the 5% level.
of heteroscedasticity (that is, existence of periods of high and low variance). Thus, our results cannot support the anecdotal evidence which says that the stock market is safer for long-term investors. Nevertheless, we do have strong enough evidence that allows us to reject the random walk hypothesis in stock prices over periods of about 15-17 years. This evidence is based on the first-order autocorrelation of multi-year returns. Yet, our results suggest that the departure from the random walk on very long horizons has been primarily a phenomenon of the post-1926 period.

### 3.3 Robustness Tests

In order to check the robustness of our findings regarding the statistical significance of the secular mean reversion, we conducted a series of robustness checks which results are not reported in this paper in order to save the space. These additional robustness tests are described below.

First, the results reported in this section are obtained using the annual data provided by Robert Shiller. More specifically, these data are annual series of (average) January values of the real Standard and Poor Composite Stock Price Index. Hence, the results obtained in this section might be affected by seasonality.\(^6\) To test the seasonality problem, we used the monthly data instead and obtained virtually the same levels of statistical significance of the mean-reverting behavior over very long horizons.

Second, Robert Shiller uses the CPI to adjust the nominal returns for inflation. We tested whether our evidence of mean reversion depends on the choice of deflator used to construct real stock returns.\(^7\) For this purpose we constructed the real stock returns using the GDP deflator and value of the Consumer bundle.\(^8\) We found that regardless of the choice of a deflator the evidence on mean reversion remains intact.

Third, since Kim et al. (1991) demonstrated that the mean-reversion in the study by Fama and French (1988b) is primarily a phenomenon of pre World War II period which is presented in both our sub-samples, we tested whether there is evidence of mean-reversion in the post 1940 period.\(^9\) We found that the evidence is weaker (which is naturally to expect

---

\(^6\)We thank Ole Gjølberg for pointing this.

\(^7\)We thank an anonymous referee for pointing this.

\(^8\)The data on the GDP deflator and the Consumer bundle are downloaded from [www.measuringworth.com](http://www.measuringworth.com).

The value of the consumer bundle is defined as the average annual expenditures of consumer units.

\(^9\)We thank an anonymous referee for pointing this.
since the sample length becomes shorter), but is still statistically significant at the 10% level.

Fourth, instead of the first-order autocorrelation of multi-year returns test statistic, suggested by Fama and French (1988b), we used the test statistic suggested by Jegadeesh (1991). In particular, instead of regression (3), we used the following regression

\[ r_t = a(k) + b(k) r_{t-k,t} + \varepsilon_{t,t+k}. \]  

(4)

Note that in this regression the stock market return at year \( t \) is predicted using the aggregated return over the preceding \( k \) years. Using this regression we could also reject the random walk hypothesis in stock prices over very long horizons in the post-1926 period.

Finally, instead of using the data provided by Robert Shiller, we used the real annual returns on the large cap stocks provided by Kenneth French\(^ {10} \) over the period from 1927 to 2012. Again we found that the values of the first-order autocorrelation of multi-year returns are statistically significantly negative over periods of 15-18 years.

Thus, on the basis of the results from numerous robustness tests, we conclude that our evidence on the secular mean reversion is robust to the choice of data, deflator, sample period, and test statistics.

4 Testing the Long-Horizon Return Predictability

4.1 Motivation

The results of the tests performed in the preceding section allow us to reject the hypothesis that the S&P Composite Stock Price Index follows a random walk. Rather surprisingly, considering a seemingly insufficient span of available historical observations of the returns on the stock index, convincing evidence against the random walk is present over long-lasting periods of about 15-17 years. That is, our tests support the alternative hypothesis that there is serial dependence in stock returns. The question arises: what kind of serial dependence? In other words, what is the alternative hypothesis? Usually a statistically significant decrease in the variance ratio with increasing investment horizon (this effect is sometimes

\(^{10}\text{See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. We use the large-cap stocks because the S&P Composite is a large-cap index.}\)
termed as the “variance compression”) is interpreted as evidence of mean reversion. Unfortunately, the evidence of mean reversion based on the variance ratio test appears to be not strong enough under stratified randomization of data. However, variance compression seems to be the sufficient, but probably not necessary condition for mean reversion. Luckily, besides the variance ratio we have another test statistic, namely, the first-order autocorrelation of multi-year returns. The significance of this test statistic is unaffected by the choice of a randomization method. The presence of the values of the autocorrelation of \( k \)-year returns that are statistically significantly below zero suggests mean reversion behavior in stock prices. Specifically, a given change in price over first \( k \) years tends to be reversed over the next \( k \) years by a predictable change in the opposite direction. For the full sample period, evidence for mean reversion comes from the negative and statistically significant values of the first-order autocorrelations at periods of 15, 16, and 17 years particularly.

Considering the above mentioned, the results reported in the previous section suggest the presence of long-term mean reversion over periods of about 15-17 years in the real Standard and Poor’s Composite Stock Price Index. In this case, if the pattern of the first-order autocorrelation of multi-year returns suggests the presence of mean reversion over the horizon of \( 2k \) years, there should be some degree of predictability of multi-year returns over a half-part of this horizon, that is, over a period of \( k \) years. Indeed, regression (3) is a predictive regression. To demonstrate the predictability of multi-year returns, Figure 2 presents a scatter plot of \( r_{t,t+15} \) versus \( r_{t-15,t} \) for the returns on the real Standard and Poor’s Composite Stock Price Index for the total sample period from 1871 to 2011. In addition, a regression line is fitted through these data points. The scatter plot clearly suggests a tendency for the past 15-year returns to predict future 15-year returns. The regression line has a strongly negative slope, and \( R^2 \) statistic is 42%.

However, if we use the full sample period to estimate the first-order autocorrelation of multi-year returns, our estimate measures the degree of in-sample (IS) predictability. Yet it is known that in-sample predictability might be spurious (for example, it appears as a result of data mining) and not hold out-of-sample (OOS) (see, for example, Bossaerts and Hillion (1999), Goyal and Welch (2003), and Welch and Goyal (2008)). In order to guard against data mining, in this section we assess the performance of the OOS forecast based on the mean-revering model given by regression (3). Besides the mean-reverting model, we use
several other competing predictive models. We demonstrate that in the OOS tests the mean-reverting model and a few other predictive models perform statistically significantly better than the naive historical-mean model. In addition, we demonstrate that the advantages of the predictive models translate into significant utility gains.

4.2 Methodology of Assessing the Performance of OOS Forecasts

Our OOS recursive forecasting procedure is as follows. The initial IS period \([1, m], m < n\), is used to estimate regression (3) for different period lengths \(k \in [10, 20]\) years. In this manner we estimate a number of autocorrelations of \(k\)-year returns, \(AC_1(k)\). Then we perform the bias adjustment of \(AC_1(k)\). Next we select the value of \(k = k_1\) which produces the lowest estimate of the bias-adjusted autocorrelation. That is,

\[
k_1 = \arg \min_{k \in [10, 20]} AC_1(k).
\]

Presumably, over the initial IS period the evidence of mean reversion is strongest over the period of \(k_1\) years. Subsequently, the estimated coefficients from regression (3) with \(k_1\) are used to compute the first \(k_1\)-year ahead return forecast for the period \([m + 1, m + k_1]\). We then expand our IS period by one year (it becomes \([1, m + 1]\)), perform the selection of \(k_2\) at which the evidence of mean reversion is strongest over the second IS period, and compute
the OOS forecast for the period \([m + 2, m + k_2 + 1]\). We repeat the procedure, increasing every time our IS window by one year, until we compute the last \(k_l\)-year ahead return for the period \([n - k_l + 1, n]\).

Observe that our OOS forecasting procedure is free from look-ahead bias, since to forecast the return for the period \([m + j, m + k_j + j - 1]\), \(j \geq 1\), we use only information that is available at time \(m + j - 1\). It is worth noting that since we are dealing with a long-horizon forecast, in performing the recursive forecasting procedure we need not just to update the estimates for the coefficients of regression (3), but first of all we need to update the optimal length of the prediction period \(k\). Observe that, in order to avoid the look-ahead bias, the optimal length of the prediction period \(k\) is determined using only information that is available at the end of each IS period as well. Thus, our OOS recursive forecasting procedure updates all the values of the model parameters and is able to adapt to changing conditions in the time series. For example, it can accommodate the possibility that the period of mean reversion is monotonically changing over time.\(^\text{11}\)

To assess the performance of OOS forecast, a common approach in the empirical literature is to run a “horse-race” among several competing predictive models. A standard criterion by which to compare two alternative predictive models is to compare their mean squared prediction errors (MSPE). As a matter of fact, the comparison of the mean squared prediction errors of two alternative models has a long tradition in evaluating which of the two models has a better ability to forecast, see McCracken (2007) and references therein. In our study, we run OOS horse races involving the mean-reverting model (MR), the historical-mean model (HM), Robert Shiller’s model (PE10) that uses the cyclically adjusted price-to-earnings ratio as a predictor for long-horizon returns, the model that uses the price-to-dividends ratio (PD) as a predictor, and the model that uses the long-term bond yield

\(^{11}\)Recall that the results presented in the previous section indicate that the period of the long-term mean reversion seems to have been increasing over time. In particular, during the first sub-sample the evidence of mean reversion is strongest over horizons of about 24-26 years (judging by the values of the most statistically significant first-order autocorrelation and variance ratio). In contrast, during the second sub-sample the evidence of mean reversion is strongest over horizons of about 34-36 years. Apparently this results in the fact that over the total sample period the evidence of mean reversion is strongest over horizons of about 30-34 years.
(LTY) as a predictor. These models are given by

\begin{align*}
\text{MR} & : r_{t,t+k} = a(k) + b(k) r_{t-k,t} + \varepsilon_{t,t+k}, \\
\text{PE10} & : r_{t,t+k} = a(k) + b(k) \text{pe}_{10t} + \varepsilon_{t,t+k}, \\
\text{PD} & : r_{t,t+k} = a(k) + b(k) \text{pd}_t + \varepsilon_{t,t+k}, \\
\text{LTY} & : r_{t,t+k} = a(k) + b(k) \text{lty}_t + \varepsilon_{t,t+k}, \\
\text{HM} & : r_{t,t+k} = a(k) + \varepsilon_{t,t+k},
\end{align*}

where $\text{pe}_{10}$ is the natural log of the ratio of price to 10-year moving average of earnings (this ratio is usually denoted as CAPE or PE10), $\text{pd}$ is the natural log of the price-to-dividends ratio, and $\text{lty}$ is the natural log of the long-term bond yield. The data for the price-to-earnings ratio, price-to-dividends ratio, and the long-term bond yield are also provided by Robert Shiller.

Robert Shiller’s model was introduced by Campbell and Shiller (1998) and further popularized and developed by Shiller (2000) and Campbell and Shiller (2001). Shiller’s model is based on a simple mean reversion theory which says that when stock prices are very high relative to recent earnings, then prices will eventually fall in the future to bring the price-to-earnings ratio back to a more normal historical level. Using this model Campbell and Shiller (1998) predicted the stock market crash of 2000 on the basis of an unreasonably high PE10 ratio. Since that time, Shiller’s model has been extremely popular among practitioners. Originally, Campbell and Shiller (1998), Shiller (2000), and Campbell and Shiller (2001) used this model to forecast future 10-year returns. Yet, Asness (2003) demonstrated that the PE10 ratio is a good predictor of the future returns over periods from 10 to 20 years.\footnote{This conclusion is made on the basis of studying $R^2$ of the predictive regression for different forecasting horizons. It should be noted, however, that in estimating the coefficient in front of the predictor and its significance level, Asness (2003) does not account for the estimation biases discovered by Cavanagh et al. (1995) and Stambaugh (1999).} Thus, Shiller’s model represents a natural competitor to our long-term mean-reverting model.

The model that uses the price-to-dividends ratio as a predictor for future returns was presented by Fama and French (1988a). This model is also based on a simple mean reversion theory which says that if the price-to-dividends ratio is unusually high or low, then this ratio tends to return to its long-run historical mean. The motivation for the model that
uses the long-term bond yield as a predictor is based on a simple idea that stocks and long-term bonds are two major competing assets. Therefore simple logic suggests that the changes in the long-term bond yield must be highly correlated with the changes in the stock market earnings yield (earnings-to-price ratio). If, for example, the bond yield increases, stock prices should decrease and the stock market earnings yield increase. The so-called “Fed model” postulates that the stock’s earnings yield should be approximately equal to the long-term bond yield. Empirical support for this model is found in the studies by Lander, Orphanides, and Douvogiannis (1997), Koivu, Pennanen, and Ziemba (2005), Berge, Consigli, and Ziemba (2008), and Maio (2013).

The historical-mean model can be interpreted as a reduced version of any other predictive model. This model uses the historical average of $k$-year returns to predict the return for the next $k$ years. It is worth emphasizing that Welch and Goyal (2008) also employed in their study the predictive models that use the price-to-earnings ratio, price-to-dividends ratio, and the long-term bond yield. They found that in out-of-sample tests these models perform worse than the historical-mean model. However, these authors used an increasing forecast horizon up to 5 years only. In our study the goal is to compare the out-of-sample forecast accuracy from these models on horizons longer than 10 years.

Now we turn to the formal presentation of our test statistic that is employed to assess the performance of OOS forecasts provided by two competing models. Let $r^AC_{t,t+k}$, $t > m$, be the actual $k$-year returns and $r^mod_1_{t,t+k}$ and $r^mod_2_{t,t+k}$ be the OOS forecast of the $k$-year returns provided by models 1 and 2. To compute the test statistic, we first compute the OOS prediction errors of the two competing models

$$
\varepsilon^mod_1_{t,t+k} = r^mod_1_{t,t+k} - r^AC_{t,t+k}, \quad \varepsilon^mod_2_{t,t+k} = r^mod_2_{t,t+k} - r^AC_{t,t+k}.
$$

Our test statistic is the ratio of the MSPE of model 1 to the MSPE of model 2

$$
\text{MSPE-R} = \frac{1}{T-m} \sum_{t=m+1}^{T} \left( \varepsilon^mod_1_{t,t+k} \right)^2 \div \frac{1}{T-m} \sum_{t=m+1}^{T} \left( \varepsilon^mod_2_{t,t+k} \right)^2,
$$

where $T - m$ is the number of OOS forecasted $k$-year returns.\textsuperscript{13} The null hypothesis in this

\textsuperscript{13}Note that $k$ is not constant, but a variable which is exogenously determined by our recursive forecasting procedure. We suppress its dependence on time in order to simplify the notation.
test is that the forecast provided by model 2 is not better than the forecast provided by model 1. Formally, under the null hypothesis the MSPE of model 1 is less than or equal to the MSPE of model 2. Formally, \( H_0 : \text{MSPE-R} \leq 1 \). Consequently, we reject the null hypothesis when the actual estimate for the MSPE ratio is significantly above unity. In our tests, the model 1 is always the historical-mean model. Therefore the outcome of our tests is whether a predictive model can “beat” the historical-mean model (a similar approach is used by Goyal and Welch (2003), Welch and Goyal (2008), and many others).

If two alternative prediction errors are assumed to be Gaussian, serially uncorrelated, and contemporaneously uncorrelated, then an MSPE-R statistic under the null hypothesis has the usual \( F \)-distribution.\(^{14}\) However, in our case the assumptions listed above are not met. First, because of using overlapping multi-year returns, the prediction errors of all our models are serially correlated. Second, since the historical-mean model is the reduced version of any other predictive model, the prediction errors of the historical-mean models and any other predictive model are contemporaneously correlated. Finally, the assumption of Gaussian errors also seems to be unpalatable. One potential possibility to obtain correct statistical inference in this case is to perform asymptotically valid tests in the spirit of the seminal tests by Diebold and Mariano (1995). However, because we use relatively small samples, and because of the variable length \( k \) of the prediction horizon in our forecasting procedure, in order to compute the p-value of the MSPE ratio we employ a bootstrap method.

Our bootstrap method follows closely Welch and Goyal (2008). In this method we assume that the returns are serially independent, whereas the log of the PE10, the log of the PD, and the log of LTY follow the first-order autoregressive (AR(1)) process. Therefore the data generating process is assumed to be

\[
\begin{align*}
   r_t & = \mu + u_t, \\
   \text{pe}10_t & = \alpha_1 + \beta_1 \text{pe}10_{t-1} + w_t, \\
   \text{pd}_t & = \alpha_2 + \beta_2 \text{pd}_{t-1} + z_t, \\
   \text{LTY}_t & = \alpha_3 + \beta_3 \text{LTY}_{t-1} + e_t.
\end{align*}
\]

\(^{14}\)In this case testing the null hypothesis largely corresponds to the standard \( F \)-test of equal forecast error variances.
In this case the return series \( r_t \) follows the random walk\(^{15} \) and a bootstrapped resample is generated using the nonparametric bootstrap method. In particular, a random resample \((r_1^*, r_2^*, \ldots, r_n^*)\) is generated by drawing with replacement from the original series \((r_1, r_2, \ldots, r_n)\). In contrast, a bootstrapped resample of any other predictive variable is generated using the semi-parametric bootstrap method. The construction of a bootstrapped resample for the log of the PE10 series, \( pe_{10t} \), is performed as follows. First of all, the parameters \( \alpha_1 \) and \( \beta_1 \) are estimated by OLS using the full sample of observations, with the residuals stored for resampling. Afterwards, to generate a random resample \((pe_{101}^*, pe_{102}^*, \ldots, pe_{10n}^*)\) we pick up an initial observation \( pe_{101}^* \) from the actual data at random. Then a series is generated using the AR(1) model and by drawing \( w_t^* \) with replacement from the residuals.\(^{16} \) The construction of a bootstrapped resample for the log of the PD and the LTY series is done in a similar manner.

Now we turn to the description of how we compute the MSPE-R statistic and its p-value. First, using the original series \((r_1, r_2, \ldots, r_n)\) we employ the recursive forecasting procedure described above to obtain the OOS forecasts of the mean-reverting model. Note that one of the outcomes of our recursive forecasting procedure is a sequence of lengths of prediction periods \((k_1, k_2, \ldots, k_l)\). Second, using the same sequence of lengths of prediction periods we obtain the OOS forecasts of all the other models. Afterwards we compute the mean squared prediction errors, and after that the MSPE-R statistic. Then we bootstrap the original series to get random resamples. The next crucial step is to generate a sequence of lengths of prediction periods \((k_1^*, k_2^*, \ldots, k_l^*)\). All this is repeated 10,000 times, each time running the recursive forecasting procedures\(^{17} \) and obtaining an estimate for MSPE-R*. In this manner we estimate the sampling distribution of the MSPE-R statistic under the null hypothesis. Finally, to estimate the significance level, we count how many times the computed value for the MSPE-R* after bootstrapping happens to be above the value of the actual estimate for the MSPE-R. In other words, under the null hypothesis we compute

\(^{15} \text{Note that in this case the historical-mean model is a version of the random walk hypothesis.} \)

\(^{16} \text{It should be noted, however, that our data generating process assumes no contemporaneous correlation between the stock return and a predictive variable. In the actual data there is a small but statistically significant correlation between the returns and the price-to-earnings (as well as the price-to-dividends) ratio. To check the robustness of our findings, we also implemented another bootstrap method which retains the historical correlations between the data series. We found that both the bootstrap methods deliver similar p-values of our test statistic.} \)

\(^{17} \text{Note that this time the recursive forecasting procedures for all the models use the exogenously determined sequence of lengths of prediction periods.} \)
the probability of obtaining a more extreme value for the MSPE ratio than the actual estimate.\(^{18}\)

It is not clear what method should be used to generate a sequence of lengths of prediction periods for each bootstrap simulation. To the best of the author’s knowledge, there are no similar forecasting procedures in the relevant scientific literature. Therefore we entertain four different methods listed below. In the first method we always use the original sequence of lengths of prediction periods \((k_1, k_2, \ldots, k_l)\). In the second and third methods a generated sequence \((k_1^*, k_2^*, \ldots, k_l^*)\) is a bootstrapped version of the original sequence. Whereas in the second method we use the nonparametric bootstrap, in the third method we use the semiparametric bootstrap. In the semi-parametric bootstrap we assume that the length of a prediction period is a linear function of time.\(^{19}\) In the fourth method a sequence of lengths of prediction periods is endogenously determined by the recursive forecasting procedure on the basis of the bootstrapped series \((r_1^*, r_2^*, \ldots, r_n^*)\). We find that the first three methods produce virtually similar p-values, whereas the fourth method produces notably lower p-values. Therefore when we report the p-values of the MSPE-R statistic we use the highest p-values. Thus, our statistical inference is based on the “worst case scenario” for the rejection of the null hypothesis. In other words, if we can reject the null in the “worst case scenario”, we would reject it for any other case.

### 4.3 Empirical Results on Performance of OOS Forecasts

Our OOS forecast begins 50 years after the data are available, that is, in 1921, and ends in 1997 with the last forecast for the 15-year period from 1997 to 2011. To check the robustness of findings, we split the total OOS period in two equal OOS subperiods, the first one from 1921 to 1959, and the second one from 1959 to 1997. As in Goyal and Welch (2003), we employ a simple graphical diagnostic tool that makes it easy to understand the relative performance of two competing forecasting models. In particular, in order to monitor the predictive power of the unrestricted model relative to the predictive power of

\(^{18}\)Note again that in this manner we compute p-values of one-tailed test.  

\(^{19}\)Indeed, for our OOS period from 1921 to 1997 the length of a prediction period is almost monotonically stepwise increasing from 10 to 15 years. The goodness of fit to the linear function, as measured by \(R^2\), amounts to 73%. To perform the semi-parametric bootstrap, first of all we estimate the simple linear trend model for the original sequence of lengths of prediction periods \((k_1, k_2, \ldots, k_l)\) with the residuals stored for resampling. Afterwards, to generate a random resample of the sequence of lengths of prediction periods, we pick up the original initial prediction period \(k_1\). The rest of the sequence is generated using the estimated linear trend model by drawing the error terms from the residuals with replacement.
the restricted model, Goyal and Welch (2003) suggested using the cumulative difference between the MSPE of the restricted model (the HM model in our case) and the MSPE of the unrestricted model:

$$CUDIF_t = \sum_{i=m}^{t} \left( \varepsilon_{i,i+k}^{mod1} \right)^2 - \left( \varepsilon_{i,i+k}^{mod2} \right)^2 .$$

By visual examination of the graph of $CUDIF_t$ it is easy to understand in which periods the unrestricted model predicts better than the restricted model. Specifically, in periods when the cumulative MSPE difference increases, the unrestricted model predicts better, in periods when it decreases, the unrestricted model predicts worse than the restricted model.

Figure 3 shows the performance of the unrestricted models versus the performance of the restricted (historical-mean) model. Specifically, left panels in the figure plot the actual $k$-year ahead returns versus the OOS forecasted $k$-year ahead returns produced by the unrestricted and restricted models. Right panels in the figure plot the cumulative difference between the MSPE of the restricted model and the MSPE of the unrestricted model. The results of the estimations of the MSPE-R test statistic with corresponding p-values are reported in Table 5.

The p-values of the MSPE-R statistic demonstrate that over the total OOS period 3 out of 4 unrestricted models performed statistically significantly better (at the 5% level) than the restricted model. These unrestricted models are: the mean-reverting model, the price-earnings model, and the price-dividends model. However, over the first OOS subperiod only the price-dividends model performed statistically significantly better than the historical-mean model. In contrast, over the second OOS subperiod only the mean-revering and the price-earnings models showed the evidence of superior forecasting accuracy as compared to that of the historical-mean model. Our results advocate that the model, which uses the long-term bond yield as predictor, performed substantially worse than all the other competing models. Our results on the predictive ability of the long-term bond yield support the conclusions reached in the studies by Estrada (2006) and Estrada (2009). Specifically, Estrada argued that the predictive ability of the long-term bond yield is supported by data in the post 1960 period only.\(^{20}\) Prior to 1960, there is no empirical support for the model.

\(^{20}\)In all empirical studies that demonstrate the predictive ability of the long-term bond yield the sample period starts after 1960. In this case if, for example, the initial IS period is chosen to be 1960-1980, then over the OOS period 1980-2010 one finds the evidence of OOS predictability of stock return using the long-term bond yield.
Figure 3: Performance of the unrestricted models versus the performance of the restricted (historical-mean) model. Left panels plot the actual $k$-year ahead returns (black line) versus the $k$-year ahead returns forecasted OOS by the unrestricted (red line) and restricted (green line) models. The initial IS period is from 1871 to 1920 which covers a span of 50 years. The OOS forecast begins in 1921 and ends in 1997 with the last forecast for the 15-year period from 1997 to 2011. Right panels plot the cumulative difference between the MSPE of the restricted model and the MSPE of the unrestricted model.
that uses the long-term bond yield as a predictor for stock returns.

The graphs of the cumulative difference between the MSPE of the restricted (historical-mean) model and the unrestricted model allow us to see in which historical periods one model performed better than the other. Visual monitoring of these graphs reveals the following observations. The price-dividends model performed relatively well until about 1970 only. After that, the accuracy of the forecast provided by the price-dividends model was substantially worse than that of the historical-mean model. Both the mean-reverting and price-earnings models performed significantly better than the historical-mean model over 1960-1990. From about 1990 the price-earnings model lost its advantage over the historical-mean model. Starting from about 1980 the mean-reverting model performed substantially better than all the other competing models. Only over the decade of 1950s the mean-reverting model performed notably worse than the historical-mean model.

4.4 Economic Significance of Return Predictability

In the preceding subsection we found a statistically significant evidence of long-term predictability of stock returns. This evidence was obtained by comparing the MSPE of the predictive model with the MSPE of the historical-mean model. However, over the total OOS period the ratios of the MSPE of the restricted model to the MSPE of the unrestricted model are not substantially above unity. This raises the important question of whether they are economically meaningful. Put it differently, statistical significance is not the same thing as economic significance.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{OOS period} & \text{HM to MR} & \text{HM to PE10} & \text{HM to PD} & \text{HM to LTY} \\
\hline
1921-1997 & 1.35 (0.01) & 1.44 (0.01) & 1.33 (0.03) & 0.44 (0.98) \\
1921-1959 & 0.86 (0.38) & 1.14 (0.10) & 1.38 (0.04) & 0.37 (0.97) \\
1959-1997 & 2.14 (0.01) & 1.78 (0.01) & 1.28 (0.09) & 0.50 (0.92) \\
\hline
\end{array}
\]

Table 5: The values of the MSPE-R statistic with corresponding p-values (in brackets). A MSPE-R statistic is a ratio of the mean squared prediction error of the restricted (historical-mean) model to the mean squared prediction error of an unrestricted model. The four competing unrestricted models are: the mean-reverting (MR) model, the price to 10-year moving average of earnings (PE10) model, the price-dividends (PD) model, and the long-term bond yield model (LTY). For example, the column HM to MR reports the values of the ratio of the MSPE of the historical-mean model to the MSPE of the mean-reverting model. The estimated p-values of the MSPE ratios are based on performing 10,000 bootstraps. Bold text indicates values that are statistically significantly above unity at the 5% level.
To estimate the economic significance of return predictability, we follow closely the methodology employed in the studies by Fleming, Kirby, and Ostdiek (2001), Campbell and Thompson (2008), and Kirby and Ostdiek (2012). We consider an investor who, at time $t$, allocates the proportion $y_t$ of his wealth to the stock market index and the proportion $(1 - y_t)$ to the risk-free asset. The investor revises the composition of his portfolio at time $t + q$; that is, after $q$ years, $q \geq 1$. The investor’s return over period $(t, t + q)$ is given by

$$R_{t,t+q} = y_t r_{t,t+q} + (1 - y_t) r_{t,t+q}^{free},$$

where $r_{t,t+q}$ and $r_{t,t+q}^{free}$ are the stock market return and the risk-free rate of return over period $(t, t + q)$.

We assume that the investor is equipped with the mean-variance utility function which can be considered as a second-order approximation to the investor’s true utility function. As a result, the investor’s realized utility over period $(t, t + q)$ can be written as

$$u(R_{t,t+q}) = y_t \left( r_{t,t+q} - r_{t,t+q}^{free} \right) - \frac{1}{2} \gamma y_t^2 \sigma_{t,t+q}^2,$$

where $\sigma_{t,t+k}$ is the volatility of the stock market index over period $(t, t + q)$ and $\gamma$ is the investor’s coefficient of risk aversion. The total investor’s realized utility is found as the sum of single-period utilities

$$U(R) = \sum_{i=1}^{n} u(R_{t,t+q}), \quad t = (i - 1) \times q,$$

where $n = \frac{T}{q}$ is the number of periods of length $q$ from time $0$ to time $T$ (the end of the investment horizon).

The investor’s optimal proportion $y_t$, which maximizes the expected utility, is given by (see Bodie, Kane, and Marcus (2007), Chapter 7)

$$y_t = \frac{1}{\gamma} \left( \frac{E[r_{t,t+q}] - r_{t,t+q}^{free}}{\sigma_{t,t+q}^2} \right),$$

where $E[r_{t,t+q}]$ and $\sigma_{t,t+q}$ are the expected return and volatility over $(t, t + q)$ that need to be forecasted at time $t$. The forecasting of expected returns is done using two competing models, 1 and 2. Specifically, $r_{t,t+q}^{mod1}$ and $r_{t,t+q}^{mod2}$ denote the return forecasts provided by
models 1 and 2 respectively. Since we do not have a specific predictive model to forecast the volatility, the volatility over \((t, t + q)\) is forecasted using the historical-mean model for volatility. Formally,
\[
y_{t}^{\text{mod1}} = \frac{1}{\gamma} \left( \hat{\sigma}_{t,t+q}^{2} - \hat{r}_{t,t+q}^{\text{free}} \right), \quad y_{t}^{\text{mod2}} = \frac{1}{\gamma} \left( \hat{\sigma}_{t,t+q}^{2} - \hat{r}_{t,t+q}^{\text{free}} \right),
\]
where \(\hat{\sigma}_{t,t+q}\) denotes the forecasted volatility.

It is important to observe that our predictive models forecast the stock market returns for a period of \(k \geq 10\) years. Since generally \(q \neq k\) (most often \(q < k\)), the \(q\)-year forecasted returns for model \(i \in \{1, 2\}\) are computed as
\[
\hat{r}_{t,t+k}^{\text{modi}} = \hat{r}_{t,t+k} \times \frac{q}{k},
\]
where \(\hat{r}_{t,t+k}^{\text{modi}}\) is the \(k\)-year return forecast provided by model \(i\).

As before, the model 1 in our study is the historical-mean model. The economic significance of return predictability is measured by equating to total realized utilities associated with two alternative forecasting models
\[
\sum_{i=1}^{n} u \left( R_{t,t+q}^{\text{modi}} \right) = \sum_{i=1}^{n} u \left( R_{t,t+q}^{\text{modi}} - q \times \Delta \right),
\]
where \(\Delta\) denotes the annual fees the investor is willing to pay to switch from predictive model 1 to predictive model 2. Whereas Fleming et al. (2001) and Kirby and Ostdiek (2012) used the equation above to compute the annual fees, Campbell and Thompson (2008) demonstrated that the total realized investor’s mean-variance utility can alternatively be measured by means of the Sharpe ratio. That is, the computation of the annual fees can be done using
\[
SR \left( R_{t,t+q}^{\text{mod1}} \right) = SR \left( R_{t,t+q}^{\text{mod2}} - q \times \Delta \right),
\]
where \(SR(\cdot)\) denotes the Sharpe ratio.

In our computations we assume that the investor’s risk aversion \(\gamma = 5\) (as in Kirby and Ostdiek (2012)). Since we do not have data for the real risk-free rate of return, to perform the computations we assume that the nominal annual risk-free rate of return equals the annual inflation rate. Therefore, in real terms, \(\hat{r}_{t,t+p}^{\text{free}} = 0\). We measure the annual
performance fees over our total OOS period 1921-2011. Table 6 reports the Sharpe ratios associated with each predictive model and the estimated annual fees measured in basis points. The results are reported for two values of $q$: $q = 1$ and $q = 15$. In the first case the investor rebalances his portfolio once a year, in the second case the investor rebalances his portfolio once in 15 years.

First we consider the case where the investor rebalances his portfolio once a year. In this case the Sharpe ratios of all predictive models, which perform statistically significantly better than the historical-mean model, are higher than the Sharpe ratio of the historical-mean model. The advantages of these predictive models translate into significant utility gains. Specifically, risk-averse investors would be willing to pay from 30 to 77 basis points fees per year to switch from the historical-mean model to a model with a superior forecast accuracy. In contrast to these models, our results indicate that the model that uses the long-term bond yield as a predictor demonstrates an inferior forecast accuracy as compared with that of the historical-mean model. As a result, not only the Sharpe ratio of this model is lower than that of the historical-mean model, but also the investor would require to be paid 20 basis points fees per year to switch from the historical-mean model to the bond

| Forecasting model       | Sharpe ratio | Basis point fees |
|-------------------------|--------------|------------------|
| **Panel A : Portfolio rebalancing once a year** |              |                  |
| Historical-Mean         | 0.35         | 0                |
| Mean-Reverting           | 0.42         | 46               |
| Price-Earnings           | 0.45         | 77               |
| Price-Dividends          | 0.39         | 30               |
| Bond Yield               | 0.32         | -20              |
| **Panel B : Portfolio rebalancing once in 15 years** |              |                  |
| Historical-Mean         | 0.35         | 0                |
| Mean-Reverting           | 0.40         | 47               |
| Price-Earnings           | 0.35         | 1                |
| Price-Dividends          | 0.39         | 37               |
| Bond Yield               | 0.21         | -129             |

Table 6: The table reports the performance of alternative predictive models and the annual fees the investor is willing to pay to switch from the historical-mean model to another predictive model. The performance is measured by means of the Sharpe ratio. The annual fees are measured in basis points.
When the investor can rebalance his portfolio once a year, the price-earnings model performs best while the mean-reverting model performs second best. However, when the investor decreases the portfolio revision frequency, the performance gains delivered by the price-earnings model diminish whereas the performance gains provided by the mean-reverting model remains rather stable. When the investor rebalances his portfolio once in 15 years, the performance gains of the price-earnings model virtually disappear. In contrast, the performance gains of the mean-reverting model (as measured in annual fees) remain virtually intact. Therefore in cases where the investor has to make long-term allocation decisions, the mean-reverting model delivers the highest performance gains.

5 Summary and Conclusions

We started the paper by performing two tests of the random walk hypothesis using the real Standard and Poor’s Composite Stock Price Index data for the period from 1871 to 2011. In particular, we investigated the time series properties of the index returns at increasing horizons up to 40 years. In our tests of the random walk hypothesis we used two well-known test statistics: the autocorrelation of multi-year returns and the variance ratio. In the context of the null hypothesis our goal was to test whether the index returns are distributed independently of their ordering in time. In order to estimate the significance level of the test statistics under the null hypothesis, we employed the randomization methods which are free of distributional assumptions.

Rather surprisingly, considering a seemingly insufficient span of available historical observations of the returns on the stock index, either of the test statistic allowed us to reject the random walk hypothesis at conventional statistical levels over very long horizons of about 30-34 years. By studying the impact of sample period on the test statistics we concluded that mean reversion seems to be an extraordinary strong phenomenon of the post-1926 period. Having performed the same randomization tests with stratification we found that the results based on the use of the variance ratio are sensitive to the particular pattern of heteroscedasticity that occurred historically,\textsuperscript{21} while the results based on the use of the autocorrelation of multi-year returns are not.

\textsuperscript{21}A similar conclusion is drawn by Nelson and Kim (1993).
Consequently, we do not have strong enough evidence to claim that the variance ratio decreases with increasing investment horizons. In other words, our results cannot support the conventional belief that the stock market is safer for long-term investors. In contrast, we do have convincing evidence that suggests that a given change in price over 15-17 years tends to be reversed over the next 15-17 years by a predictable change in the opposite direction. Overall, our findings support the mean reversion hypothesis as the alternative to the random walk hypothesis. Our evidence of secular mean reversion in stock prices is robust to the choice of data source, deflator used to compute the real prices and returns, sample period, and test statistic.

The results of our tests demonstrated the evidence of in-sample predictability. However, conventional wisdom says that in-sample evidence of stock return predictability might be a result of data mining. In order to guard against data mining, we investigated the performance of out-of-sample forecast of multi-year returns. We demonstrated that the out-of-sample forecast provided by the mean-reverting model is statistically significantly better than the forecast provided by the historical-mean model. Moreover, the out-of-sample forecast accuracy of the mean-reverting model is comparable to that of very popular (among practitioners) Robert Shiller’s model that uses the cyclically adjusted price-earnings ratio as a predictor for long-horizon returns, and of the model that uses the price-dividends ratio as a predictor for long-horizon returns. In addition, we demonstrated that the advantages of these three predictive models translate into significant utility gains. We found that in cases where the investor has to make long-term allocation decisions, the mean-reverting model delivers the highest performance gains. Besides, in the post-1960 period the mean-reverting model showed the best forecast accuracy among all competing model.

Given the main result of our study, it is natural to ask the following question. What causes this long-lasting mean reversion in the stock market prices? Put it differently, what is the economic intuition behind this result? One possible answer is suggested by previous research on the link between the demography and stock market returns and on the long-term variations in the birth rates and population growth in the US. In particular, on the one hand, Bakshi and Chen (1994), Dent (1998), Geanakoplos, Magill, and Quinzii (2004), and Arnott and Chaves (2012) observe the interrelationship between the demography and the US stock market returns and argue that the demography determines the stock market
returns. On the other hand, the evidence presented by Kuznets (1958), Dent (1998), Berry (1999), and Geanakoplos et al. (2004) suggests the presence of secular trends in birth rates in the US that last from 10 to 20 years. Thus, if the population growth goes through long-term alternating periods of above-average and below-average rates, and it is the demography that determines the stock market returns, then it is naturally to expect that the stock market also goes through long-term alternating periods of above-average and below-average returns.

A more elaborate model of cyclical dynamics of economic activity, interrelated with similar movements in other elements, is presented by Schlesinger (1949), Schlesinger (1986), Berry (1991), Berry, Elliot, Harpham, and Kim (1998), and Alexander (2004). These authors argue that the dynamics of economic activity in the US has a long-term rhythm (with a period of 12-18 years) of accelerated and retarded secular growth. This cyclical fluctuation in economic activity, in particular the alternation of long-term periods of good and bad economic times, gives rise to similar long-term fluctuations in social and political activities. In brief, a long-term period of rapid economic growth and technological development coincides with a conservative political wave (era). The conservative politics reduces the scope and the role of government in the life of the nation and frees up business and capital. Such a period is also characterized by a higher population growth, increase in inequality, and deflationary conditions. Yet inevitably a long-term period of economic growth comes to a long-term stagflationary crisis. During such a crisis conservative leaders are replaced by liberal leaders committed to business regulation, social innovation, equity, and redistribution via an enhanced role of government. A liberal era is usually characterized by a lower population growth, decrease in inequality, and inflationary conditions. In our opinion, the secular mean-reverting behavior of the stock market fits nicely into this model of socioeconomic dynamics. It seems to be possible to demonstrate that the conservative political waves are usually associated with above average stock market returns, whereas during the liberal political waves the stock market returns are below average.
References

Alexander, M. A. (2000). *Stock Cycles: Why Stocks Won’t Beat Money Market Over the Next Twenty Years*. iUniverse Star.

Alexander, M. A. (2004). *Cycles in American Politics*. iUniverse Star.

Arnott, R. D. and Chaves, D. B. (2012). “Demographic Changes, Financial Markets, and the Economy”, *Financial Analysts Journal, 68*(1), 23–46.

Asness, C. S. (2003). “Fight the Fed Model”, *Journal of Portfolio Management, 30*(1), 11–24.

Bakshi, G. S. and Chen, Z. (1994). “Baby Boom, Population Aging, and Capital Markets”, *Journal of Business, 67*(2), 165–202.

Balvers, R., Yangru, W., and Gilliland, E. (2000). “Mean Reversion Across National Stock Markets and Parametric Contrarian Investment Strategies”, *Journal of Finance, 55*(2), 745–772.

Berge, K., Consigli, G., and Ziemba, W. T. (2008). “The Predictive Ability of the Bond-Stock Earnings Yield Differential Model”, *Journal of Portfolio Management, 34*(3), 53–80.

Bodie, Z., Kane, A., and Marcus, A. J. (2007). *Investments*. McGraw Hill.

Bossaerts, P. and Hillion, P. (1999). “Implementing Statistical Criteria to Select Return Forecasting Models: What Do We Learn?”, *Review of Financial Studies, 12*(2), 405–428.

Campbell, J. Y. (1987). “Stock Returns and the Term Structure”, *Journal of Financial Economics, 18*(2), 373–399.

Campbell, J. Y. and Mankiw, N. G. (1987). “Are Output Fluctuations Transitory?”, *Quarterly Journal of Economics, 102*(4), 857–880.

Campbell, J. Y. and Shiller, R. J. (1988). “The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors”, *Review of Financial Studies, 1*(3), 195–227.
Campbell, J. Y. and Shiller, R. J. (1998). “Valuation Ratios and the Long-Run Stock Market Outlook”, *Journal of Portfolio Management, 24*(4), 11–26.

Campbell, J. Y. and Shiller, R. J. (2001). “Valuation Ratios and the Long-Run Stock Market Outlook: An Update”, Working paper, Harvard University and Yale University.

Campbell, J. Y. and Thompson, S. B. (2008). “Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?”, *The Review of Financial Studies, 21*(4), 1509–1531.

Campbell, J. Y. and Yogo, M. (2006). “Efficient Tests of Stock Return Predictability”, *Journal of Financial Economics, 81*(1), 27–60.

Cavanagh, C. L., Elliott, G., and Stock, J. H. (1995). “Inference in Models with Nearly Integrated Regressors”, *Econometric Theory, 11*(5), 1131–1147.

Cochrane, J. H. (1988). “How Big Is the Random Walk in GNP?”, *Journal of Political Economy, 96*(5), 893–920.

Dent, H. (1998). *The Roaring 2000s*. Simon and Shuster, New York.

Diebold, F. X. and Mariano, R. S. (1995). “Comparing Predictive Accuracy”, *Journal of Business and Economic Statistics, 13*(3), 253–263.

Easterling, E. (2005). *Unexpected Returns: Understanding Secular Stock Market Cycles*. Cypress House.

Estrada, J. (2006). “The Fed Model: A Note”, *Finance Research Letters, 3*(1), 14 – 22.

Estrada, J. (2009). “The Fed Model: The Bad, the Worse, and the Ugly”, *The Quarterly Review of Economics and Finance, 49*(2), 214 – 238.

Fama, E. F. (1970). “Efficient Capital Markets: A Review of Theory and Empirical Work”, *Journal of Finance, 25*(2), 383–416.

Fama, E. F. (1981). “Stock Returns, Real Activity, Inflation, and Money”, *American Economic Review, 71*(4), 545–565.

Fama, E. F. (1990). “Term-Structure Forecasts of Interest Rates, Inflation, and Real Returns”, *Journal of Monetary Economics, 25*(1), 59–76.

Fama, E. F. and French, K. R. (1988a). “Dividend Yields and Expected Stock Returns”, *Journal of Financial Economy, 22*, 3–25.

Fama, E. F. and French, K. R. (1988b). “Permanent and Temporary Components of Stock Prices”, *Journal of Political Economy, 96*(2), 246–273.

Fama, E. F. and French, K. R. (1989). “Business Conditions and Expected Returns on Stocks and Bonds”, *Journal of Financial Economics, 25*(1), 23–49.
Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Fleming, J., Kirby, C., and Ostdiek, B. (2001). “The Economic Value of Volatility Timing”, *Journal of Finance*, 56(1), 329–352.

Geanakoplos, J., Magill, M., and Quinzii, M. (2004). “Demography and the Long-Run Predictability of the Stock Market”, *Brookings Papers on Economic Activity*, 1, 241–325.

Goyal, A. and Welch, I. (2003). “Predicting the Equity Premium with Dividend Ratios”, *Management Science*, 49(5), 639–654.

Gropp, J. (2003). “Mean Reversion of Size-Sorted Portfolios and Parametric Contrarian Strategies”, *Managerial Finance*, 29, 5–21.

Gropp, J. (2004). “Mean Reversion of Industry Stock Returns in the U.S., 1926-1998”, *Journal of Empirical Finance*, 11, 537–551.

Hirsch, J. A. (2012). *The Little Book of Stock Market Cycles*. John Wiley and Sons, New Jersey.

Jegadeesh, N. (1991). “Seasonality in Stock Price Mean Reversion: Evidence from the U.S. and the U.K.”, *Journal of Finance*, 46(4), 1427–1444.

Katsenelson, V. N. (2007). *Active Value Investing: Making Money in Range-Bound Markets*. Wiley Finance.

Kim, M. J., Nelson, C. R., and Startz, R. (1991). “Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence”, *Review of Economic Studies*, 58(3), 515–528.

Kirby, C. and Ostdiek, B. (2012). “It’s All in the Timing: Simple Active Portfolio Strategies that Outperform Naive Diversification”, *Journal of Financial and Quantitative Analysis*, 47, 437–467.

Koivu, M., Pennanen, T., and Ziemba, W. T. (2005). “Cointegration Analysis of the Fed Model”, *Finance Research Letters*, 2(4), 248 – 259.

Kuznets, S. (1958). “Long Swings in the Growth of Population and in Related Economic Variables”, *Proceedings of the American Philosophical Society*, 102(1), 25–52.

Lander, J., Orphanides, A., and Douvogiannis, M. (1997). “Earnings Forecasts and the Predictability of Stock Returns: Evidence from Trading the S&P”, *Journal of Portfolio Management*, 23(4), 24 – 35.

Lanne, M. (2002). “Testing the Predictability of Stock Returns”, *Review of Economics and Statistics*, 84(3), 407–415.

Leroy, S. F. (1982). “Expectations Models of Asset Prices: A Survey of Theory”, *Journal of Finance*, 37(1), 185–217.
Lewellen, J. (2004). “Predicting Returns with Financial Ratios”, *Journal of Financial Economics, 74*(2), 209–235.

Lo, A. W. and MacKinlay, A. G. (1988). “Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test”, *Review of Financial Studies, 1*(1), 41–66.

Maio, P. (2013). “The Fed Model and the Predictability of Stock Returns”, *Review of Finance, 17*(4), 1489–1533.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall/CRC.

Marriott, F. H. C. and Pope, J. A. (1954). “Bias in the Estimation of Autocorrelations”, *Biometrika, 41*(3/4), 390–402.

McCracken, M. W. (2007). “Asymptotics for Out of Sample Tests of Granger Causality”, *Journal of Econometrics, 140*(2), 719–752.

Nelson, C. R. and Kim, M. J. (1993). “Predictable Stock Returns: The Role of Small Sample Bias”, *Journal of Finance, 48*(2), 641–661.

Noreen, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. John Wiley and Sons, New York.

Officer, R. R. (1973). “The Variability of the Market Factor of the New York Stock Exchange”, *Journal of Business, 46*(3), 434–453.

Orcutt, G. H. and Irwin, J. O. (1948). “A Study of the Autoregressive Nature of the Time Series Used for Tinbergen’s Model of the Economic System of the United States”, *Journal of the Royal Statistical Society. Series B, 10*(1), 1–53.

Poterba, J. M. and Summers, L. H. (1988). “Mean Reversion in Stock Prices: Evidence and Implications”, *Journal of Financial Economics, 22*, 27–59.

Rapach, D. E. and Wohar, M. E. (2005). “Valuation Ratios and Long-Horizon Stock Price Predictability”, *Journal of Applied Econometrics, 20*(3), 327–344.

Richardson, M. (1993). “Temporary Component of Stock Prices: A Skeptic’s View”, *Journal of Business and Economic Statistics, 11*(2), 199–207.

Richardson, M. and Stock, J. (1989). “Drawing Inferences From Statistics Based on Multi-Year Asset Returns”, *Journal of Financial Economics, 25*, 323–348.

Rogers, J. (2005). *Hot Commodities: How Anyone Can Invest Profitably in the World’s Best Market*. Random House Publishing Group.

Schlesinger, Jr., A. M. (1986). *The Cycles of American History*. Boston: Houghton Mifflin.
Schlesinger, Sr., A. M. (1949). *Paths to the Present*. New York: Macmillian Co.

Schwert, G. W. (1989). “Why Does Stock Market Volatility Change Over Time?”, *Journal of Finance, 44*(5), 1115–1153.

Shiller, R. J. (1989). *Market Volatility*. The MIT Press.

Shiller, R. J. (2000). *Irrational Exuberance*. Princeton University Press.

Siegel, J. (2002). *Stocks for the Long Run*. McGraw-Hill Companies.

Stambaugh, R. F. (1999). “Predictive Regressions”, *Journal of Financial Economics, 54*(3), 375–421.

Summers, L. H. (1986). “Does the Stock Market Rationally Reflect Fundamental Values?”, *Journal of Finance, 41*(3), 591–601.

Torous, W., Valkanov, R., and Yan, S. (2004). “On Predicting Stock Returns with Nearly Integrated Explanatory Variables”, *Journal of Business, 77*(4), 937–966.

Welch, I. and Goyal, A. (2008). “A Comprehensive Look at The Empirical Performance of Equity Premium Prediction”, *Review of Financial Studies, 21*(4), 1455–1508.