A Forecasting Model for Stock Market Diversity

Francesco Audrino[a,*] Robert Fernholz[b,†] Roberto G. Ferretti[c,‡]

[a,c]University of Lugano, Switzerland

[b]INTECH

[c]BSI SA

Revised: March 2006

Abstract

We apply the recently introduced general tree-structured (GTS) model to the analysis and forecast of stock market diversity. Diversity is a concentration measure of capital across a market, which plays a central role in the search for arbitrage. The GTS model allows for different conditional mean and volatility regimes that are directly related to the behavior of macroeconomic fundamentals through a binary threshold construction. Testing on US market data, we collect empirical evidence of the strong potential of the model in estimating and forecasting diversity accurately, also in comparison with other standard approaches. Moreover, the GTS model allows the construction of very simple portfolio strategies that systematically beat the standard cap-weighted S&P500 index.

Keywords: Diversity; Generalized tree-structured threshold models; Maximum-likelihood estimation; Diversity-based portfolio strategies.

*Address: Institute of Finance, USI, Via Buffi 13, Centrocivico, CH-6900 Lugano, Switzerland. E-mail: francesco.audrino@lu.unisi.ch. Financial support by the Foundation for Research and Development of the University of Lugano and by the National Centre of Competence in Research "Financial Valuation and Risk Management" (NCCR FINRISK) is gratefully acknowledged. The authors thank four anonymous referees for helpful comments.

†Address: INTECH, One Palmer Square, Priceton, NJ 08542. E-mail: bob@enhanced.com.

‡Address: Institute of Finance, USI, Via Buffi 13, Centrocivico, CH-6900 Lugano, Switzerland. E-mail: roberto.ferretti@lu.unisi.ch.

1 Introduction

The concept of market diversity was introduced in the research carried out by R. Fernholz over the past few years (for an overview, see Fernholz, 1999, 2001, and 2002, and Fernholz, Garvy, and Hannon 1998 and Fernholz and Garvy 1999). Diversity is a concentration measure of capital across a market. Diversity is higher when capital is spread more evenly across the stocks in the market, and is lower when capital is concentrated over a few large stocks. Knowing the cap-weights of the market portfolio and its diversity, it is possible to explicitly define a diversity-weighted portfolio (see equation (6)). This portfolio is slightly less exposed to the larger stocks than the standard cap-weighted index. This means that the diversity-weighted portfolio is likely to under-perform the market in times of decreasing diversity, and is likely to outperform the market in times of increasing diversity. Diversity plays an important role in the search for arbitrage in financial markets. We say that the market is diverse if no single company is allowed to dominate the entire market, in terms of relative capitalization. Given this hypothesis the diversity-weighted portfolio outperforms the market over a sufficiently long time-horizon. In fact, the relative logarithmic performance of the diversity-weighted portfolio is given by the sum of the change of diversity of the market plus a positive drift. Since diversity is bounded, the drift will dominate in the long run. However, in the real world, portfolios strategies are assessed in time intervals ranging from three months to five years. A portfolio manager willing to exploit the arbitrage opportunities given by the diversity-weighted portfolio must therefore be able to predict the monthly changes in market diversity in some way.

The no-arbitrage hypothesis is central to modern mathematical finance. Examples of arbitrage in the literature do not seem to resemble actual equity markets, so perhaps it was thought that arbitrage could occur only in bizarre circumstances. From a normative point of view, the condition that the market is diverse looks like an innocuous enough assumption, and it would be imposed upon an actual equity market by any credible regulation. However, it implies the existence of an arbitrage opportunity. It is difficult, if not impossible, to prove the validity of the no-arbitrage hypothesis empirically. In contrast, statistical tests of various forms of the efficient market hypothesis have appeared (see Taylor, 1986, and Malchiel, 1990). However, none of these constitute a test of no-arbitrage.

Changes in diversity also explain more than half the annual variations in relative equity manager performance (see Fernholz and Garvy, 1999). It is well known that managers are reluctant to concentrate as much capital in the largest stocks as occurs in the cap-weighted indices, and this causes their returns to be strongly correlated with the change in diversity. If the relationship between managers' relative performance and change in diversity continues to hold in the future, then prediction of change in market diversity could be used to determine the allocation of assets between index managers and active managers. In periods of increasing diversity, more assets could be invested with active managers, and in periods of declining diversity, more assets could be indexed.

But how can market diversity changes be forecast? It is intuitively clear that diversity changes have somehow to be related to the economic activity of the market. Different studies in the literature provided empirical evidence that classic asset pricing factors can be useful in determining economic characteristics of the stock market. For instance, macro variables for inflation and real activity, financial variables (interest rates, spreads, volatilities) or Fama and French benchmark portfolios and risk factors (see, for example, Fama and French, 1993 or 1996) may may be reasonable variables to predict diversity changes. Indeed, when market diversity grows, capital concentrates. Hence big stocks become bigger and small stocks smaller. Since Fama and French portfolios are the classical explanatory factors of all market large/small cap effects, it is natural to consider them as closely related to diversity changes.

Moreover, empirical evidence suggests that diversity dynamics should be driven by a mean reverting process, and that points of high, respectively low, diversity can be described by different criteria, implying that different regimes may explain diversity changes.

For all the reasons presented so far, we analyze the diversity time series dynamics using the generalized tree-structured (GTS) model first introduced by Audrino and Bühlmann (2001) and Audrino and Trojani (2003). The GTS model allows for different conditional mean and volatility regimes constructed in a very simple way. Moreover, since GTS regimes are determined by multivariate thresholds on lagged values of some relevant endogenous (for example, diversity changes themselves) and exogenous (for example, macroeconomic indicators for real activity and inflation) variables, the GTS model can tie the behavior of the diversity dynamics to macroeconomic fundamentals.

In this study, we collect empirical evidence for the existence of more than one volatility regime driving the diversity process. In particular, we find that the most relevant predictors in the GTS model are the well-known Fama and French benchmark portfolios constructed using equities with high book-to-market. This result confirms our first impression that these kinds of macroeconomic variables are closely related to diversity.

Apart from a number of in-sample estimation contributions useful to better understand the dynamics of the diversity process, this study also investigates the accuracy of the GTS model in forecasting out-of-sample diversity. Note that accurate forecasts of diversity directions are particularly important to develop reliable portfolio strategies. First, we provide statistical empirical evidence for the strong predictive potential of the GTS model, also comparing it to alternative approaches. Second, we implement a portfolio strategy based on the diversity direction forecasts from the GTS model. We test our strategy on US market data and we compare its profitability to that of the cap-weighted and diversity-weighted S&P500 portfolios. We provide empirical evidence of the potential of our strategy in terms of gross and net cumulative returns, and in terms of statistical tests on differences among rolling one-year Sharpe ratios.

The rest of the paper is organized as follows. A review of the theory on stock market diversity is presented in Section 2. Section 3 presents our generalized tree-structured model and the corresponding estimation procedure. The empirical in-sample results for the time series of monthly US diversity data are summarized in Section 4. The out-of-sample performance of the models is tested in Section 5. Section 6 presents and applies on US market data our diversity based portfolio strategy. Section 7 concludes the paper.

2 Stock market diversity

Suppose that $X_1, \ldots, X_n$, with $n > 1$, are a family of continuous positive semimartingales representing the total capitalizations of each of the stocks in an equity market. These processes satisfy a number of conditions that do not concern us here, but which can be found, for example, in Fernholz (2002). We define the *covariance* processes σ_{ij} for the stocks to satisfy

$$\sigma_{ij}(t)\, dt = d\langle X_i, X_j \rangle_t, \qquad t \geq 0, \tag{1}$$

where $\langle X_i, X_j \rangle$ is the cross variation process of X_i and X_j.

The total capitalization of the market at time $t \geq 0$ is given by

$$Z_\mu(t) = X_1(t) + \cdots + X_n(t), \qquad t \geq 0, \tag{2}$$

so Z_μ is also a continuous positive semimartingale. The processes

$$\mu_i(t) = X_i(t)/Z_\mu(t), \qquad t \geq 0, \tag{3}$$

for $i = 1, \ldots, n$, represent the *weights* or *proportions* of each of the stocks in the market. These market weights are usually referred to as *capitalization weights* (or *cap weights*) since they represent the proportion of the total market capitalization that resides in each stock. It is not difficult to show that the dynamics of the market capitalization satisfy

$$\frac{dZ_\mu(t)}{Z_\mu(t)} = \sum_{i=1}^n \mu_i(t)\frac{dX_i(t)}{X_i(t)}, \qquad t \geq 0, \qquad \text{a.s.} \tag{4}$$

Process $\mu = (\mu_1, \ldots, \mu_n)$ is called the *market portfolio* and $Z_\mu(t)$ represents its value at time $t \geq 0$. In general, a bounded process $\pi q = (\pi_1, \ldots, \pi_n)$ is a *portfolio* process if for all $t \geq 0$, $\pi_1(t) + \cdots + \pi_n(t) = 1$. The processes π_i are called the *weight* processes for the portfolio, and represent the relative proportion of the portfolio invested in each of the stocks. The *portfolio value process* Z_π satisfies

$$\frac{dZ_\pi(t)}{Z_\pi(t)} = \sum_{i=1}^n \pi_i(t)\frac{dX_i(t)}{X_i(t)}, \qquad t \geq 0, \qquad \text{a.s.} \tag{5}$$

If the value of portfolio π at time $t = 0$ is $Z_\pi(0)$, then $Z_\pi(t)$ represents the value of the portfolio at time $t \geq 0$. For details on these processes, see Fernholz (2002).

Let Δ_n be the n-dimensional simplex

$$\Delta_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \; x_1 + \ldots + x_n = 1, 0 < x_i < 1, i = 1, \ldots, n\}.$$

Let $\mathbf{S}$ be a positive C^2 function defined on a neighborhood of Δ_n such that for all i, $x_i \partial_i \log \mathbf{S}(x)$ is bounded on Δ_n. Then $\mathbf{S}$ generates the portfolio π with weights

$$\pi_i(t) = \Big(\partial_i \log \mathbf{S}(\mu(t)) + 1 - \sum_{j=1}^n \mu_j(t)\partial_j \log \mathbf{S}(\mu(t))\Big)\mu_i(t)$$

for $t \in [0, T]$ and $i = 1, \ldots, n$. In this case we have

$$d\log\big(Z_\pi(t)/Z_\mu(t)\big) = d\log \mathbf{S}(\mu(t)) + d\Theta(t) \qquad t \geq 0, \qquad \text{a.s.}, \tag{6}$$

where Θ, the *drift process*, is locally of bounded variation. It takes a.s. the form (for $t \in [0, T]$)

$$d\Theta(t) = -\frac{1}{2\mathbf{S}(\mu(t))}\sum_{i,j=1}^n \partial_{i,j}\mathbf{S}(\mu(t))\mu_i(t)\mu_j(t)\tau_{ij}(t)dt,$$

where $\tau_{ij}(t)$ denotes the relative covariance term (see Fernholz, 2002, Theorem 3.1.5). We define a *measure of diversity* $\mathbf{D}_p$ by

$$\mathbf{D}_p(x_1, \ldots, x_n) = \Big(\sum_{i=1}^n x_i^p\Big)^{1/p}, \tag{7}$$

with p a real number satisfying $0 < p < 1$. This function generates the portfolio with weights

$$\pi_i(t) = \frac{\mu_i^p(t)}{(\mathbf{D}_p(\mu_1, \ldots, \mu_n))^p} \tag{8}$$

for $i = 1, \cdots, n$, and drift process

$$d\Theta(t) = \frac{1-p}{2}\Big(\sum_{i=1}^{n} \pi_i(t)\sigma_{ii}(t) - \sum_{i,j=1}^{n} \pi_i(t)\pi_j(t)\sigma_{ij}(t)\Big)dt \qquad t \geq 0, \qquad \text{a.s.} \tag{9}$$

Portfolio π is called a *diversity-weighted* portfolio.

Note that in information theory (see Renyi, 1961) the normalized logarithm of diversity is called *Renyi entropy.* When $p \to 0$ it converges towards the usual Shannon entropy measure

$$E(x_1, \cdots, x_n) = -\sum_{i=1}^{n} x_i \log(x_i).$$

All these concentration measures express different flavors of the information content hidden in portfolio weights.

The drift process Θ in (7) can be shown to be strictly increasing (see Fernholz, 2002). In this case, if the value of $\mathbf{D}_p(\mu(t))$ has not decreased after a given period of time has elapsed, then (6) shows that the diversity-weighted portfolio π will have higher returns than the market portfolio μ over that period. Hence, changes in the *market diversity*, $\mathbf{D}_p(\mu(t))$, are of vital importance to the investment performance of the diversity-weighted portfolio π. Here we shall show how these changes can be predicted.

3 Model description and estimation

This section outlines the model used to estimate and predict the log-diversity time series $\log \mathbf{D}_p(\mu_1(t), \ldots, \mu_n(t)),\ t \geq 0$, introduced in section 2 and briefly reviews the model selection procedure that can be applied to it.

3.1 Starting point

Let us briefly denote by D_t the series of log-diversity $\log \mathbf{D}_p(\mu_1(t), \ldots, \mu_n(t))$, where for the sake of simplicity, we drop the dependence of diversity from p and the portfolio μ. For estimation purposes, we discretize the stochastic differential of the diversity dD_t in (6)

$$\Delta D_t = D_t - D_{t-1} = \mu_t + \varepsilon_t, \tag{10}$$

where $E[\varepsilon_t \mid \Phi_{t-1}] = 0$ and Φ_{t-1} is the information set up to time $t-1$. Such a discretization serves to approximate the true diversity process. We assume that the dynamics of the diversity changes follow

$$\varepsilon_t = \sqrt{h_t} z_t, \quad \mu_t = g(\Phi_{t-1}), \quad h_t = f(\Phi_{t-1}), \tag{11}$$

for some functions $g(\cdot) \in \mathbb{R}$ and $f(\cdot) \in \mathbb{R}^+$. $(z_t)_{t=0,1,2,\ldots}$ is a sequence of independent identically distributed innovations with zero mean and unit variance. The model (10)-(11) is a general (nonparametric) model for the diversity (log-) changes ΔD_t.

As we saw in the Introduction, it may be important to allow the relevant conditioning information set Φ_{t-1} to be as wide as possible to obtain reliable conditional first and second moment estimates. For this reason, we set $\Phi_{t-1} = \{\tilde{D}_{t-1}, h_{t-1}, \mathbf{x}^{\text{ex}}_{t-1}\}$, where $\tilde{D}_{t-1} = \{D_{t-1}, D_{t-2}, \ldots\}$ is the whole past history of the (log-) diversity series and $\mathbf{x}^{\text{ex}}_{t-1}$ is a vector of all other relevant exogenous variables that are used for prediction. In particular, we consider as relevant exogenous factors past values of some macroeconomic variables such as indices for real activity and inflation. In addition, we also consider as predictors term structures variables, the well-known Fama and French risk factors and benchmark portfolios constructed considering different sizes and boot-to-market, and five Fama and French industry portfolios. More details on all such variables are given in Section 4.1. In particular, note that the dependence of μ_t on h_{t-1} allows for a (possibly nonlinear) conditional mean effect of volatility. Similarly, the dependence of h_t on $\tilde{D}_{t-1}, h_{t-1}$ and $\mathbf{x}^{\text{ex}}_{t-1}$ allows for a broad variety of asymmetric volatility patterns in reaction to past market and macroeconomic information.

The general model (10) proposed in this paper nests several classical models found in the literature. For instance, one quickly spots that Bollerslev's AR(1)-GARCH(1,1) model is encompassed by (10)-(11). In fact, the conditional mean function is parameterized by

$$\mu_t = g(\Phi_{t-1}) = a + b\Delta D_{t-1}, \tag{12}$$

where a and b are unknown parameters to be estimated. Similarly, the conditional variance is parameterized by

$$h_t = f(\Phi_{t-1}) = \alpha_0 + \alpha_1 \varepsilon^2_{t-1} + \beta h_{t-1}, \tag{13}$$

where α_0, α_1 and β are the unknown parameters. Clearly, the recursive definition of the GARCH(1,1) model implies that the conditional variance depends on the entire history of the data and $\Phi_{t-1} = \{\tilde{D}_{t-1}\}$.

A second classical model encompassed by (10) is the generalized regime-switching (GRS) model proposed by Gray (1996). In Gray's two-regime GRS model, assuming conditional normality within each regime, the conditional mean function is given by

$$\mu_t = g(S_t, \Phi_{t-1}) = p_{t,1}\mu_{t,1} + (1 - p_{t,1})\mu_{t,2} \tag{14}$$

and the variance of diversity changes by

$$h_t = f(S_t, \Phi_{t-1}) = p_{t,1}(\mu_{t,1}^2 + h_{t,1}) + (1 - p_{t,1})(\mu_{t,2}^2 + h_{t,2}) - [p_{t,1}\mu_{t,1} + (1 - p_{t,1})\mu_{t,2}]^2, \quad (15)$$

where $p_{t,1}$ denotes the conditional probability to be in regime 1 at time t given the past history $\tilde{D}_{t-1}$, i.e., $p_{t,1} = P[S_t = 1 \mid \Phi_{t-1}]$, and S_t is the unobserved regime at time t. $\mu_{t,i}$ and $h_{t,i}$, $i = 1, 2$, denote the regime-dependent conditional means and variances. $\Phi_{t-1} = \{\tilde{D}_{t-1}\}$ does not contain S_t or lagged values of S_t. In (14) and (15) the regime's dependent conditional mean functions are parameterized by AR(1) models and the regime's dependent conditional variances by GARCH(1,1) models. Note that Gray's model, despite being of the general form (10), is not encompassed in (11), since conditional means and variances are functions also of unobservable ex-ante probabilities (and not only of an observable information set ϕ_{t-1}).

In this study, we use the generalized tree-structured (GTS) GARCH model introduced by Audrino and Trojani (2003) to analyze diversity changes. The GTS model is a parametric model for (11) which allows for flexibility in the conditional mean and variance functions g and f and which is still computationally manageable when applied to real data examples. More details on the model are given in the next section.

3.2 The GTS model

Analogously to the models introduced in section 3.1, the generalized tree-structured (GTS) model parameterizes the conditional mean $\mu_t(\theta) = g_\theta(\Phi_{t-1})$ and conditional variance $h_t(\theta) = f_\theta(\Phi_{t-1})$ by means of some parametric threshold functions and an unknown parameter vector θ.

The basic idea of the GTS model introduced by Audrino and Bühlmann (2001) and Audrino and Trojani (2003) is in the spirit of a sieve approximation of g and f by means of piecewise linear functions. This can be accomplished by partitioning the domains of g and f in a finite sequence of regimes (or cells) using a binary tree construction. For any given regime we specify a regime's dependent AR-GARCH type structure for conditional means and volatilities. The additional information deriving from the exogenous variables is used to determine the optimal thresholds.

The parametric GTS model considered is given by

$$\Delta D_t = \mu_t(\theta) + \sqrt{h_t(\theta)} z_t = g_\theta(\Phi_{t-1}) + \sqrt{f_\theta(\Phi_{t-1})} z_t, \quad (16)$$

where the conditional mean and variance functional forms g_θ and f_θ are constructed using a binary tree-structured model that involves a partition $\mathcal{P}$ of the state space G of $\Phi_{t-1} =$

$\{D_{t-1}, D_{t-2}, \varepsilon_{t-1}, h_{t-1}, \mathbf{x}^{\text{ex}}_{t-1}\}$:

$$\mathcal{P} = \{\mathcal{R}_1, \ldots \mathcal{R}_k\},\ G = \cup_{j=1}^{k} \mathcal{R}_j,\ \mathcal{R}_i \cap \mathcal{R}_j = \emptyset\ (i \neq j)\ .$$

Given a partition cell $\mathcal{R}_j$, we describe the dynamics of D_t on this cell with a local AR(1)-GARCH(1,1) model. This leads to functions for $\mu_t(\theta)$ and $h_t(\theta)$ that depend on (i) the set of parameters of any local AR(1)-GARCH(1,1) model in the generalized tree-structured GARCH model and (ii) the structure of partition $\mathcal{P}$. More precisely, we have:

$$g_\theta(D, \varepsilon, h, x^{\text{ex}}) = g_\theta^{\mathcal{P}}(D, \varepsilon, h, x^{\text{ex}}) = \sum_{j=1}^{k} (a_j + b_j \Delta D)\, I_{[(D,\varepsilon,h,x^{\text{ex}}) \in \mathcal{R}_j]}, \tag{17}$$

$$f_\theta(D, \varepsilon, h, x^{\text{ex}}) = f_\theta^{\mathcal{P}}(D, \varepsilon, h, x^{\text{ex}}) = \sum_{j=1}^{k} \left(\alpha_{0,j} + \alpha_{1,j}\varepsilon^2 + \beta_j h\right) I_{[(D,\varepsilon,h,x^{\text{ex}}) \in \mathcal{R}_j]}, \tag{18}$$

where $\theta = (a_j, b_j, \alpha_{0,j}, \alpha_{1,j}, \beta_j;\ j = 1, \ldots, k)$. Clearly, $k = 1$ implies Bollerslev's AR(1)-GARCH(1,1) model specified in (12) and (13). For $k \geq 2$ we obtain a richer class of threshold models, where k also indicates the number of model regimes.[1] In (17)-(18), we model $h_t(\theta)$ by means of a threshold GARCH function f_θ and $\mu_t(\theta)$ by means of a threshold regime's dependent mean-reverting function g_θ. We incorporate in the threshold definitions behind f_θ and g_θ the joint impact of D_{t-1}, D_{t-2}, ε_{t-1}, h_{t-1} and all other relevant exogenous variables of interest.

Note that the GTS model (16)-(18) is for some important aspects different from the GRS model (14)-(15) proposed by Gray (1996). First of all, in our approach regimes are determined by multivariate tree-structured thresholds. Second, in contrast to regime-switching models the optimal number of regimes is estimated endogenously during the procedure and not given a priori. Third, the GTS model being of a threshold type, it allows for a perfect regime-classification of the observed data.

3.3 The estimation procedure

The negative log-likelihood[2] of the GTS model (16) is given by

$$-\ell(\theta; \Phi_2^n) = -\sum_{t=2}^{n} \log \left[\sqrt{h_t(\theta)}^{-1} p_Z \left((\Delta D_t - \mu_t(\theta)) / \sqrt{h_t(\theta)} \right) \right], \tag{19}$$

where $p_Z(\cdot)$ denotes the density function of the distribution of the standardized innovation z_t and $\Phi_2^n = \{\Phi_2, .., \Phi_n\}$. Therefore, for any given partition $\mathcal{P}$, the GTS model (16)-(18) can be estimated by means of (pseudo) maximum likelihood. The choice between different partition structures (i.e., the selection of the optimal threshold functions) involves a model choice procedure for non-nested hypotheses.

More precisely, the flexible procedure for the estimation of the GTS model described in Audrino and Bühlmann (2001) is based on the following two steps.

(i) For any given partition $\mathcal{P}$ the estimation of θ is performed by (pseudo) maximum likelihood based on a Gaussian (pseudo) log likelihood[3] and the parametric forms (17) and (18) for g_θ and f_θ.

(ii) Model selection of the optimal threshold function (i.e., the optimal partition $\mathcal{P}$) is performed via a tree-structured partial search in order to avoid a computationally impracticable exhaustive search. Within any data-determined tree structure the optimal model is finally selected according to the Bayesian-Schwarz Information Criterion (BIC) for predictive accuracy.

For all details about the estimation procedure applied to the GTS model we remand to Audrino and Bühlmann (2001) and Audrino and Trojani (2003).

4 Data and in-sample estimation results

4.1 Data

The data used in this study are the monthly diversity (log-) changes $\log \mathbf{D}_p(\mu_{(1)}(t), \ldots, \mu_{(m)}(t))$ of the S&P500 Index, already analyzed in Fernholz (2002), where parameter p was chosen to be 0.76. Bear in mind that parameter p can be selected anywhere between 0 and 1. The flexibility in the choice of this parameter is an issue that has to be further analyzed. The value $p = 0.76$ was chosen, since with this value of p, the diversity-weighted version of the S&P500 Index retains characteristics common to other well-known large-stock indices such as the given (capitalization weighted) S&P500 of Russell 1000. These characteristics are, first, that the Index holds a representative selection of large companies; second, that the selection and weighting of the securities in the Index are objectively established; and third, that the portfolio turnover is minimal.

The data spans the time period between January 1960 and December 2001, for a total of 504 observations. Figure 1 plots the data as well as the monthly changes of the diversity process. Table 1 presents some sample statistics.

FIGURE 1 AND TABLE 1 ABOUT HERE.

Figure 1 illustrates well the changes in the diversity process that occurred, for example, during the Nifty-Fifty period (first half of the Seventies) or the months after the October 1987 stock market crash and the months around the March 2000 stock market crash. The volatility of the monthly changes associated with such particular events is striking. Table 1 shows that the mean change and skewness in the diversity series are both close to zero. The AR(1) term is found to be significant and the correlation between ΔD_t and D_{t-1} is negative.

To exploit the possible additional information included in the yield curve, we downloaded the one-month US Treasury bill rates and the 60-month zero coupon bond rates from the Fama CRSP discount bond files. Some sample statistics for such yields, as well as for the spread between long- and short-term rates, are summarized in Table 2. As expected, the average yield curve is upward sloping. The yields are highly autocorrelated, with increasing autocorrelation for the long-term interest rates. The 60-month yield levels show mild excess kurtosis and positive skewness.

We also use additional macroeconomic variables as conditioning predictors in our generalized tree-structured model, since we believe that they can substantially improve estimation and forecast. We divide the macroeconomic variables into two main groups. The first group consists of two inflation measures which are based on the CPI and the PPI of finished goods. The second group contains variables that capture real activity: the index of Help Wanted Advertising in Newspapers (HELP), unemployment (UE) and the growth rate of industrial production (IP). In addition, we also consider monthly log-returns of the S&P500 Index. All the macroeconomic data have been downloaded from *Datastream International* for the time period under investigation. This list of variables includes most variables that have been used in the macro literature. Among these variables, CPI and HELP are traditionally thought of as leading indicators of inflation and real activity, respectively. We consider as a predictor also a binary variable equal to 1 in case of observations belonging to a contraction period according to the NBER cycles of expansions and contractions. As we have already seen in the Introduction, such kind of macroeconomic variables can be associated to particular behaviors of the diversity process. Summary statistics of these variables are reported in Table 1.

Finally, we also downloaded from K. French's web site the time series of monthly Fama and French

(i) six benchmark portfolios constructed considering different sizes and book-to-market (big (B) and small (S) size, and high (H), neutral (M) and low (L) book-to-market);

(ii) three risk factors: the excess return on the market minus the one-month treasury bill (Rm-Rf), the average return on three small portfolios minus the average return on the big portfolios (SMB) and the average return on two value portfolios minus the average returns on two growth portfolios (HML);

(iii) five industry portfolios.

We believe that such factors can be of great importance in predicting the diversity process. In fact, when diversity grows, market capitalization concentrates on fewer stocks. In this situation large caps have a tendency to earn more than small caps. Hence, factors that explain large/small cap effects, like Fama and French portfolios, may also be useful to explain diversity changes. As expected, regressing the diversity changes time series using as predictors the six contemporary Fama and French benchmark portfolios we obtain an R^2 of more than 85%. Similarly, regressing the diversity changes using the three contemporary Fama and French risk factors we get an R^2 of about 80%. Summary statistics of these variables are reported in Table 2.

TABLE 2 ABOUT HERE.

Portfolios characterized by high book-to-market show the highest means. This feature has already been observed in De Bondt and Thaler (1987) and may be related to the fact that a small book-to-market value corresponds to an excessive optimism regarding the future profitability of the firm. Moreover, as already observed in other empirical studies, small stocks have a tendency to have higher long-term returns than large stocks. Fama and French (1993) (1996) interpreted this result as a consequence of the fact that small firms, or firms with high book-to-market are riskier and, therefore, must obtain a higher long-term return. Some sort of positive autocorrelation (like an AR(1) term) is observed for the three small-size portfolios (and consequently also for the SMB and HML risk factors) similarly to Lo and MacKinlay (1988) and Conrad and Kaul (1988). All Fama and French factors and portfolios show excess kurtosis, and most are negatively skewed. Note also that the mean of the industry portfolios for the time period under investigation is about 1%.

4.2 Estimation results

In this section, we summarize the results of the analysis of the diversity changes from a standard AR(1)-GARCH(1,1) model fit (single regime model) and the GTS model fit described in section 3.2. The classical AR(1)-GARCH(1,1) model can be used as a benchmark model to test the accuracy of the GTS model. Note that the single-regime model is included in the generalized tree-structured GARCH construction as a simple special case. We estimate the GTS model using all the endogenous and exogenous (macroeconomic) variables introduced in the last section as predictors. The parameter estimates as well as some statistics for estimating conditional first and second moments appear in Table 3 and 4. Results are computed for the whole time period beginning January 1960 and ending December 2001, for a total of 504 monthly observations. The detailed specification of the models appears below each table. Standard errors for the parameter estimates are computed using a model-based bootstrap from the standardized residuals. See Efron and Tibshirani (1993) for more details.

TABLE 3 AND 4 ABOUT HERE.

Table 3 reports the estimates for the single regime AR(1)-GARCH(1,1) model. The AR(1) parameter is, as expected, significant at the 1% confidence level. The GARCH effect appears to be important and statistically significant in characterizing the conditional variance dynamics. Both recent volatility and shocks are important factors in determining volatility. The assumption of stationarity (i.e., $a+b<1$) is not violated, although we observe strong persistence in volatility ($a+b$ near to one). The AR(1)-GARCH(1,1) model does a good job in modeling the stochastic volatility of the diversity process. The Ljung-Box statistics relating to the squared standardized residuals indicate no significant serial correlation.

Table 4 reports the results for the GTS model. We find that the GTS model has three different regimes characterized by different past returns of both the big and small size, high book-to-market Fama and French benchmark portfolios[4]. Value Fama and French benchmark portfolios are clearly the variables, among all the predictors considered, with the highest predictive power. In each regime we observe different dynamics for the conditional mean and variance.

The first regime is characterized by low past returns of both the big and small size, high book-to-market Fama and French benchmark portfolios. The constant parameter in the conditional mean equation is negative and highly significant. The implied long-run mean ($-a_1/b_1$) is also negative and equals -7.43%. On average, first lags of diversity changes classified in regime one are about -0.713% (and a minimal value of -4.839%) with a speed of reversion to the implied long-run mean of about 37 basis points. Consequently, during periods characterized by low-value Fama and French benchmark portfolios, diversity tends to decrease and the stock market to concentrate with a relatively moderate speed of reversion. Individual shocks have a small immediate effect on the conditional variance, but are strongly persistent.

The second regime is characterized by low past returns of the big size, high book-to-market Fama and French benchmark portfolio and positive past returns of the small size Fama and French value portfolio. In this regime, the AR(1) parameter in the conditional mean equation is found to be statistically significant. The implied long-run mean is positive, equals 0.55% and acts as a reflecting barrier. The speed of the reflection is moderate: when the first lag of the diversity changes is -1% and 2% the conditional mean change is about -42 and 39 basis points. (note that first lag of diversity changes classified in the second regime are between -2.213 and 3.788, with a mean value of 0.319). The GARCH parameter is the significant effect in determining volatility. The impact of individual shocks is larger than in regime one, but is less persistent.

The third and last regime is characterized by high past returns of the big size, high book-to-market Fama and French benchmark portfolio. Similarly to regime one, the constant parameter in the conditional mean equation is found to be significant. Nevertheless, the conditional mean behavior is the opposite. The implied long-run mean is positive and extremely high and acts as a an external attractor. The speed of the attraction is high: for a typical value of 3% the conditional mean change is about 60 basis points. In this regime, conditional mean changes are always positive. Consequently, during periods characterized by high past returns of big size, high book-to-market Fama and French benchmark portfolio, diversity tends to increase rapidly. In this regime, individual shocks have no immediate impact and are also much less persistent than in the first regime.

As expected, the GTS model does a good job in modeling stochastic volatility of the diversity process. The Ljung-Box statistics relating to the squared standardized residuals indicate no significant serial correlation. We perform a classical likelihood ratio test to judge whether the improvements given by the estimation of the bigger number of parameters in the GTS model over the nested AR(1)-GARCH(1,1) model are statistically significant. The value of the test statistic is 70.554. As a consequence, the GTS model is strongly preferred to the nested AR(1)-GARCH(1,1) model, both at the 5% and 1% confidence levels (critical values are 21.026 and 26.217, respectively).

The GTS model also clearly outperforms the classical single regime AR(1)-GARCH(1,1) model with respect to all performance statistics for estimating conditional first and second moments. The gains range from 4% to more than 15% depending on the performance measure.

The top panel of Figure 2 contains a plot of the regime classification of the diversity changes using the GTS model. The bottom panel of Figure 2 contains a plot of the conditional standard deviation implied by the GTS model. The detailed specification of each regime appears below the figure. Periods of high and low volatility are particularly apparent.

FIGURE 2 ABOUT HERE.

5 Forecasting results

In this section, we investigate the forecasting power of the GTS model applied to the diversity process. Moreover, we are also interested in testing whether the introduction of multiple regimes leads to over-fitting. This can be determined (i) by performing a series of out-of-sample tests and (ii) by computing confidence bound predictions using the Monte Carlo simulation method.

We always compare goodness-of-fit results from a GTS fit with those from (i) a classical single-regime AR(1)-GARCH(1,1) model and (ii) an extended Gray's two-regime generalized regime switching (GRS) model introduced in (14)-(15), where the time-varying transition probabilities are also allowed to depend on further exogenous variables[5]. The second comparison is particularly useful since it allows us to relate the performances of the GTS model with those from an alternative approach for multiple regimes.

5.1 Two out-of-sample tests

In performing the out-of-sample tests, we estimate the parameters of each particular model over an in-sample period and compute the time series of conditional means and variances over a subsequent out-of-sample period, holding the estimated parameters and regime structure fixed.

We quantify the goodness-of-fit of the different models for estimating and predicting monthly conditional first and second moments by means of various measures. Since the conditional variance is an expectation of squared innovations to the diversity process, we compare volatility estimates from the different approaches to the actual squared innovations. The difference between volatility estimates and actual squared innovations is computed for the in-sample estimation period and for the out-of-sample backtesting period. This difference is then summarized in the form of root mean squared errors (RMSE), mean absolute errors (MAE) and the R^2 between actual volatility and estimated volatility. Mathematically speaking, we consider

$$\text{V-RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{h}_t - (\Delta D_t - \hat{\mu}_t)^2\right)^2}, \tag{20}$$

$$\text{V-MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{h}_t - (\Delta D_t - \hat{\mu}_t)^2| \quad \text{and} \tag{21}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}\left(\hat{h}_t - (\Delta D_t - \hat{\mu}_t)^2\right)^2}{\sum_{t=1}^{n}(\Delta D_t - \hat{\mu}_t)^4}. \tag{22}$$

The R^2 measure, in providing a direct measure of the goodness-of-fit of the estimate, differs from the R^2 measure, which would be obtained by projecting actual volatility on forecast volatility. It imposes an intercept at zero and a slope of one on such projection, permitting direct conclusions about a particular estimate rather than about some linear transformation of that estimate.

Since we are also interested in the accuracy of the different models in predicting conditional first moments, we also compute classical in-sample and out-of-sample MAE and RMSE statistics for the estimated innovations $\hat{\epsilon}_t = \Delta D_t - \hat{\mu}_t$. We denote these statistics for the conditional mean by M-MAE and M-RMSE. In addition to these performance measures, we also consider the negative log-likelihood computed for the out-of-sample period.

To verify whether differences in the above mentioned performances are statistically significant across the models, we perform a powerful size controlled comparison of multiple models by means of the so-called Model Confidence Set (MCS), first introduced in Hansen et al. (2003). Details of the construction of MCS are given in Appendix A.

TABLES 5-7 ABOUT HERE.

Table 5 shows performance results of two different out-of-sample tests for conditional mean and volatility predictions constructed using the classical single-regime AR(1)-GARCH(1,1) model, an extended two-regime GRS models where the exogenous variable used to estimate transition probabilities is the big size, high book-to-market Fama and French benchmark portfolio[6], and the GTS model.

In Tables 6-7, we report results for the construction of MCS. In particular, we report results for both out-of-sample M- and V-MAE statistics. Qualitatively similar results can be obtained using other performance statistics. We report values of two different equal predictive ability (EPA) test statistics (range and semi-quadratic) with corresponding p-values between parentheses. In addition, we also report results of the worst performing index computation used to identify the model that has to be eliminated in case of a rejection of the null-hypothesis of EPA.

In the first test, the models are estimated over the first half of the sample and predictions are made over the second half of the sample. The in-sample estimation period includes the first half of the Seventies (Nifty-Fifty period) and three contraction periods as determined by the NBER. The out-of-sample period begins with the short January 1980 - July 1980 contraction, includes three more recession periods and the stock market crashes of October 1987 and March 2000. In this experiment, the in-sample and out-of-sample periods are more or less equally volatile (sample variance of the diversity changes in the two periods is 1.6243 and 1.5133, respectively). The GTS model has two optimal regimes in response to past returns of the big size Fama and French value portfolio. The GTS model clearly outperforms both competitors for predicting conditional variances with respect to all performance measures. The relative gains range between 6% and 45%. Similar results also hold when considering the prediction of conditional first moments, although in this case differences are smaller (GTS model about 1%-2% better than alternatives).

The second test examines the short-term forecasting ability of the models in estimating the models over the entire sample except for the last 8 years. In this case, the out-of-sample period is significantly more volatile than the in-sample period, principally because of the period before the stock market crash of March 2000 (sample variance of the diversity changes in the two periods is 1.4281 and 2.1484, respectively). The optimal GTS model has three regimes characterized by past returns of the Fama and French Cnsmr and Manuf industry portfolios. Once again, the GTS model clearly outperforms both alternative approaches in predicting accurate conditional first and second moments, with relative gains ranging from 4% up to 15% depending on the performance measure.

Table 7 shows that such differences are statistically significant for forecasting future conditional variances. In most cases, only the GTS model belongs to the MCS at the 10% or the 5% confidence levels. Differences in M-type statistics are in general not significant (see Table 6).

In summary, in our two out-of-sample tests we collect empirical evidence of the forecasting power of the GTS model applied to the time series of diversity changes, also in comparison to standard alternative competitors.

5.2 Forecasting one-period-ahead confidence bounds and diversity direction

As a second check on the forecasting properties of the GTS model we compute one-period-ahead 90% and 95% confidence bound predictions for the diversity process. Such confidence bound estimates can be easily computed using the model-based predictions for one-step-ahead conditional means and variances and the quantiles of the standardized filtered innovations. We backtest one-month-ahead confidence bound predictions using the different models beginning on January 1975, where the model parameters are estimated on all the previous data available (i.e., we compute confidence bounds for the diversity process in January 1980 estimating the model parameters on the period January 1960 - December 1979). To save time, in our GTS estimations we keep the GTS regime-structure listed in Table 4 unchanged.

An illustrative example of 90% confidence bound predictions for the diversity process using the classical AR(1)-GARCH(1,1) model and the GTS model is shown in Figure 3.

FIGURE 3 ABOUT HERE.

All models considered in the paper work quite well in constructing one-month-ahead confidence bound predictions. In particular, when performing standard overall frequency tests for the total number of exceedances[7] (realized values which lie outside the estimated confidence intervals), we never got a rejection of the null hypothesis of unconditional unbiasedness of the estimates.

As is shown in the next section, computing accurate predictions for the diversity direction (increase or decrease) in the next period is crucial to developing a good trading strategy. For this reason, we search for good rules to predict reasonably well one-month-ahead diversity directions. The first rule is based only on the median of future possible scenarios. We say that if the median of the one-month-ahead diversity scenarios is bigger than today's diversity, diversity has a tendency to increase. On the contrary, if the median of the one-month-ahead diversity scenarios is lower than or equal to today's diversity, diversity tends to decrease. Based on this rule, we backtest the predicted diversity directions using the realized values. In particular, we compute the probability to forecast correct diversity directions using the GTS model in comparison to the classical AR(1)-GARCH(1,1) model and the extended GRS model. Results are summarized in Table 8.

TABLE 8 ABOUT HERE.

Table 8 clearly shows the better potential of the GTS model to predict diversity directions over the standard AR(1)-GARCH(1,1) model. The extended GRS model is even less accurate than the classical AR(1)-GARCH(1,1) model. In about 6 out of 10 cases using the GTS model we are able to correctly determine the one-period-ahead diversity direction.

We also investigate whether a rule that takes into account other quantiles of the distribution of the future possible diversity values can improve the probability to correctly determine one-month-ahead diversity directions. Thus, as a second rule we say that if the median of the one-month-ahead diversity scenarios is bigger than today's diversityand if the lower 5%-quantile of the one-month-ahead diversity scenarios is nearer to today's diversity than the upper 95%-quantile, diversity tends to increase. Conversely, if the median of the one-month-ahead diversity scenarios is lower than or equal to today's diversity and if the upper 95%-quantile of the one-month-ahead diversity scenarios is nearer to today's diversity than the lower 5%-quantile, diversity tends to decrease. The resulting probabilities for a correct prediction of one-month-ahead diversity directions using the GTS model, the AR(1)-GARCH(1,1) model and the extended GRS model are summarized in Table 8. Based on this rule, too, the GTS model clearly outperforms the alternative competitors. Probabilities of detecting the right diversity direction are higher than those from rule 1 (in about 7 out of 10 cases we predict the right direction using the GTS model). Note that rule 2 allows us to take a decision on the future diversity direction in about 85% of cases using the GTS model. In the remaining 15% of the cases, median and higher/lower quantiles of the one-month-ahead diversity distribution yield contrasting signals on the future diversity direction.

Probabilities of correctly predicting one-month-ahead diversity direction are fairly constant among the three optimal regimes estimated in the GTS model. An important finding is that almost all observations for the ones rule 2 yield no decision about the future diversity direction are classified in regime 2. In fact, regime 2 is characterized by contrasting past information from the market (i.e., past returns of the Fama and French B.H and S.H benchmark portfolios) that, in some cases, provides no clear signal as to the future behavior of the diversity process.

FIGURE 4 ABOUT HERE.

Results for one-month-ahead prediction of diversity directions are shown graphically in Figure 4 for a GTS estimation. Circles, triangles and squares indicate observations where rule 2 predicts an increase in diversity, a decrease in diversity and does not yield a decision, respectively. Figure 4 illustrates that, using rule 2 and the GTS model, we are able to predict accurately the direction of the diversity process, especially during periods when diversity continues to increase or decrease. Moreover, the direction of the diversity process is particularly well predicted during events like the stock market crash of October 1987, the months before the stock market crash of March 2000 and the contraction period between July 1990 and March 1991.

In this paper we do not provide multi-step-ahead forecasts for the diversity process. The GTS model does not directly yield transition dynamics for the diversity process and for all related exogenous predictor variables. However, multi-step-ahead forecasts can be computed in a standard way using simulation and model-based bootstrap. Note that in such a case we need explicit univariate time-series models for each predictor variable in addition to the GTS model for the diversity process or a general multivariate model for diversity and all the other predictors.

6 A diversity based trading strategy

In this section we implement a portfolio strategy based on the one-period-ahead diversity direction forecasts introduced above. Once we have diversity direction forecasts (from rule 2 and the GTS model as explained in the last section) the portfolio strategy is constructed using the following algorithm.

1. At time $t = 0$. Invest in the usual cap-weighted S&P500 index.
2. (a) At time $t+1$. If the forecasting rule predicts a growing diversity, invest in the diversity-weighted S&P500 portfolio described by formula (3) of Section 2.
   (b) If the forecasting rule predicts a decreasing diversity, invest in the cap-weighted S&P500 portfolio.
   (c) If the forecasting rule does not predict any change in diversity, stay with the portfolio chosen at time t.
3. Iterate point (2).

We test the profitability of this strategy on monthly US market data. The testing period goes from January 1975 to December 2001. Data for the diversity-weighted S&P500 portfolio are provided by INTECH. Consistently with INTECH portfolio management policies, we estimated the turnover for the diversity-weighted index to about 12% a year (if it is traded every month), but this combines buys and sells, so actually 24% of the value of the portfolio will be traded on average every year. This means that the total cost is $0.24 * 20 = 4.8$bp of trading cost a year. In order to take into account also management costs, we subtracted annually 4.8bp from the gross performance of the diversity-weighted S&P500 portfolio. This clearly affects the performance of our strategy, too.

To evaluate the performance of this portfolio strategy we compare its cumulated returns as well as annualized volatility and total Sharpe ratio to those of the cap-weighted and diversity-weighted S&P500 portfolios, used as benchmarks. The results are summarized in Table 9, Panel A.

TABLE 9 ABOUT HERE.

Surprisingly, a simple switching strategy between two highly correlated market portfolios can drastically increase the returns while showing similar volatility patterns. In particular, the yearly improvement of our strategy over the cap-weighted S&P500 portfolio is about 61 bp. Notice that the corresponding improvement for the diversity-weighted S&P500 portfolio is only about 37 bp.

To gain even more empirical evidence of the benefits of the proposed strategy, we compute rolling one-year Sharpe ratios beginning January 1976 to the end of the testing period. We perform the statistical t-type and sign-type tests firstly introduced by Diebold and Mariano (1995) and described extensively in Audrino and Bühlmann (2004) on differences of Sharpe ratios using our strategy and the diversity-weighted S&P500 against the standard cap-weighted S&P500[8]. The results are summarized in Table 9, Panel B.

The t-type tests yield significant differences at the 5% confidence level only in the comparison between our strategy and the cap-weighted S&P500, preferring the former over the latter. No significant difference in terms of Sharpe ratios is found between the cap-weighted and the diversity-weighted S&P500. This may be just a fact of low power due to non-Gaussian observations. However, the sign-type tests, which are robust against deviations from the Gaussian model, confirm and strengthen this result. In terms of Sharpe ratios, our strategy is significantly better than the cap-weighted S&P500 already at the 1% confidence level. As before, this is not the case when confronting cap- and diversity-weighted S&P500 portfolios.

To end this analysis we would like to make some remarks about the transaction costs involved in our strategy that have not been considered so far. From historical data we estimated at about 14% the turnover needed to switch from the diversity-weighted to the cap-weighted S&P500 portfolio, and vice versa (see, also, Fernholz, Garvy, and Hannon 1998, and Fernholz, 2002, p. 146). We multiply this number by the standard 20bp rate applied by brokerage houses and subtract the resulting cost from the monthly performance of our strategy. The yearly improvement over the cap-weighted S&P500 portfolio decreases to about 42bp. Hence, most gains over the cap-weighted S&P500 portfolio remain after considering switching costs.

7 Conclusions and suggestions for future research

We have shown that monthly changes in market diversity can be modeled by an AR(1)-GARCH(1,1) process, and that it is possible to forecast the one-month-ahead change in diversity using this AR-GARCH model along with the generalized tree-structured (GTS) model of Audrino and Trojani (2003). The GTS model used as input various macroeconomic variables as well as the Fama and French benchmark portfolios. This combined model correctly predicted the direction on the next month's change in diversity about 70% of the time. When these predictions were used to switch between cap-weighted and diversity-weighted S&P500 portfolios, the strategy produced about 65 bp. better performance than the cap-weighted S&P 500 alone, and it did so without increasing the risk. Even after reasonable transaction costs were subtracted, the strategy continued to perform significantly better than the standard S%P500 index.

These results show that it is possible to significantly improve the performance of a well-known index such as the S&P500 by exploiting diversity weighting enhanced by monthly forecasts of the change in diversity. There are other possible applications of diversity forecasting that we have not investigated, but that might be the concern of a future study. For example, the possible use of diversity forecasting for asset allocation between indexing and active management would appear to be an area of particular interest. Moreover, the forecasts we have produced here are relatively short term; it would be interesting to develop longer-term diversity forecasts.

A Appendix. Tests for equal predictive ability and model confidence set construction

Without loss of generality, let us denote by $\widehat{D}_{t,wk}$ the differences of each term in the out-of-sample M-MAE statistic:

$$\widehat{D}_{t,wk} = \widetilde{U}_{t;\text{model}_w} - \widetilde{U}_{t;\text{model}_k},\ t = 1,\ldots,n,\ w,k = 1,\ldots,3, w < k,$$

where

$$\sum_{t=1}^{n} \widetilde{U}_{t;\text{model}} = \text{out-of-sample M-MAE}.$$

We consider also the sign of $\hat{D}_{t,wk}$:

$$\widehat{W}_{t,wk} = \begin{cases} -1\ , & \text{if } \widehat{D}_{t,wk} \leq 0 \\ 1, & \text{else} \end{cases}\ ,\ t = 1,\ldots,n,\ w,k = 1,\ldots,3, w < k.$$

Statistics based on time averages $\overline{D}_{wk}$ and $\overline{W}_{wk}$ of $\widehat{D}_{t,wk}$ and $\widehat{W}_{t,wk}$ allow us to investigate whether there is a systematic difference in out-of-sample forecasting power between the different models. Tests based on $\widehat{D}_{t,wk}$ are t-type tests, while tests based on $\widehat{D}_{t,wk}$ are sign-type tests. In a similar way, one can proceed by using a different out-of-sample goodness of fit statistic, like V-MAE, M-MSE or V-MSE. In our application, we compute tests based on all such statistics and report those for the M- and V-MAE statistics.

The MCS is defined as the smallest set of models which, at a given confidence level α, cannot be significantly distinguished on the basis of their forecasting power. The MCS is determined after sequentially trimming the set of candidate models, which in our application consists of the simple AR(1)-GARCH(1,1) model, the two-regimes extended GRS model, and the GTS model. At each step of such a trimming procedure, the null-hypothesis of equal predictive ability (EPA) $\mathcal{H}_0 : \mathbb{E}[D_{t,wk}] = 0,\ \forall w,k \in \mathcal{M}$ (respectively, $\mathcal{H}_0 : \mathbb{E}[W_{t,wk}] = 0$) is tested for the relevant set of models $\mathcal{M}$ at a confidence level α. In a first step, $\mathcal{M}$ consists of all models under investigation. If, in the first step, $\mathcal{H}_0$ is rejected, then the worst performing model according to the relevant criterion is eliminated. The test procedure is then repeated for the new set $\mathcal{M}$ of surviving models, and it is iterated until the first non rejection of the EPA hypothesis occurs. The set of resulting models is called the model confidence set $\widehat{\mathcal{M}}_\alpha$ at the given confidence level α. In our application we work with $\alpha = 0.05, 0.10$.

Our tests of EPA are based on the range statistic T_R and the less conservative semi-quadratic statistic T_{SQ}:

$$T_R = \max_{k,w \in \mathcal{M}} \frac{|\overline{D}_{kw}|}{\sqrt{\widehat{\mathrm{var}}(\overline{D}_{kw})}} \quad \text{and} \quad T_{SQ} = \sum_{k<w} \frac{\overline{D}_{kw}^2}{\widehat{\mathrm{var}}(\overline{D}_{kw})},$$

where the sum in T_{SQ} is taken over the models in $\mathcal{M}$, $\overline{D}_{kw} = n^{-1} \sum_{t=1}^{n} \widehat{D}_{t,kw}$, and $\widehat{\mathrm{var}}(\overline{D}_{kw})$ is an estimate of $\mathrm{var}(\overline{D}_{kw})$ obtained from a block-bootstrap of the series $\widehat{D}_{t,kw},\ t = 1, \ldots, n$. Using statistics T_R or T_{SQ}, we test the null hypothesis EPA at confidence level α for model set $\mathcal{M}$. If hypothesis EPA is rejected for model set $\mathcal{M}$, we compute a worst performing index, in order to trim the worst performing model from $\mathcal{M}$.

The worst performing index for Model$_k$ is computed as the mean across models $w \neq k$ of statistic $\overline{D}_{wk}$. More specifically, it is defined as $\overline{D}_k / \sqrt{\mathrm{var}(\overline{D}_k)}$, where $\overline{D}_k = \mathrm{mean}_{w \neq k \in \mathcal{M}} \overline{D}_{kw}$. As above, our estimate of $\mathrm{var}(\overline{D}_k)$ is based on a block-bootstrap. The model with the highest worst performing index is finally trimmed from $\mathcal{M}$. Consistency of estimates of the (asymptotic) distributions of T_R and T_{SQ} can be proved under mild regularity conditions on the bootstrap. For details, see Hansen et al. (2003).

Notes

[1]Within this framework, the conditional mean and variance could have an even more general parameterization. However, the parameterization adopted here represents a good tradeoff between flexibility and computational feasibility.

[2]The log-likelihood is always considered conditionally on Φ_1 and on some reasonable starting value $h_1(\theta)$.

[3]Regularity conditions for the consistency of pseudo maximum likelihood estimators of tree structured GARCH models are given in Audrino and Bühlmann (2001).

[4]"Low" and "high" in this context means sufficiently below or above the estimated threshold.

[5]In particular, we consider the same Gray's two-regime GRS model in (14)-(15) but we extend the time-varying transition probabilities P_t and Q_t in

$$p_{t,1} = (1 - Q_t)\frac{g_{t-1,2}(1 - p_{t-1,1})}{g_{t-1,1}p_{t-1,1} + g_{t-1,2}(1 - p_{t-1,1})} + P_t\frac{g_{t-1,1}p_{t-1,1}}{g_{t-1,1}p_{t-1,1} + g_{t-1,2}(1 - p_{t-1,1})},$$

to be of the form

$$P_t = \Phi_N(c_1 + d_1\Delta D_{t-1} + s_1 x^{\mathrm{ex}}_{t-1})$$

and

$$Q_t = \Phi_N(c_2 + d_2\Delta D_{t-1} + s_2 x^{\mathrm{ex}}_{t-1}),$$

where Φ_N is the standard normal distribution and $g_{t,j}$ $(j = 1, 2)$ is the density of a Gaussian variable with conditional mean $\mu_{t,j}$ and conditional variance $h_{t,j}$. The chosen extended two-regime GRS specification is the one with the exogenous variable x^{ex} among all the possible candidates minimizing the negative conditional (pseudo) maximum likelihood.

[6]We also tried to estimate the extended two-regime GRS model using other exogenous (macroeconomic) variables to determine transition probabilities. Nevertheless, we found that the best performance results are reached when using past information from the B.H. benchmark portfolio.

[7]Under the hypothesis of unconditional unbiasedness of the confidence bound estimates, the number of exceedances are binomially distributed around their expected values. A two-standard deviation interval can be used as tolerance for testing the null hypothesis of unconditional unbiasedness.

[8]Similarly to what is presented in Appendix A, we consider differences $\widetilde{U}_t,\ t = 1, \ldots, T$, of one-year Sharpe ratios at time t

$$\widetilde{U}_t = \text{Sharpe}_{t;\text{strategy}_1} - \text{Sharpe}_{t;\text{strategy}_2},\ t = 1, \cdots, T\ ,$$

where the rolling historical Sharpe ratios are constructed over the last 12 months. We are now testing the null hypothesis that the differences $\widehat{U}_t$ have mean zero against the alternative of mean less than zero, i.e., the Sharpe ratios from portfolio strategy$_1$ are bigger than the ones from portfolio strategy $_2$. For this purpose, we use a version of the t-test adapted to the case of dependent observations. Analogously, the version of the sign test in the case of dependent observations is based on the number of positive differences

$$\widehat{W}_t = I_{\{\widehat{U}_t \geq 0\}},\ t = 1, \ldots, T\ ,$$

for the null hypothesis that the positive differences $\widehat{W}_t$ have mean $\frac{1}{2}$ against the alternative of mean bigger than $\frac{1}{2}$.

References

Audrino, F. and Bühlmann, P. (2001), Tree-structured GARCH models. *Journal of the Royal Statistical Society, Series B*, **63**, No. 4, 727-744.

Audrino, F. and Bühlmann, P. (2004). Synchronizing multivariate financial time series. *Journal of Risk*, **6**, No. 2, 81-106.

Audrino, F. and Trojani, F. (2003), Estimating and predicting multivariate volatility regimes in global stock markets. Forthcoming in the *Journal of Applied Econometrics.*

Conrad, J. and Kaul, G. (1988). Time-variation in expected returns. *Journal of Business* **61**, 409-425.

De Bondt, W.F.M. and Thaler, R.H. (1987). Further evidence of overreaction and stock market seasonality. *Journal of Finance* **42**, 557-581.

Diebold, F.X. and Mariano, R.S. (1995), Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253-263.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap.* Chapman & Hall, London.

Fama, E.F. and French, K.R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3-56.

Fama, E.F. and French, K.R. (1996). Multifactor explanations of asset-pricing anomalies. *Journal of Finance* **47**, 426-465.

Fernholz, R. (1999). On the diversity of equity markets. *Journal of Mathematical Economics* **31**, 393-417.

Fernholz, R. (2001). Equity portfolios generated by functions of ranked market weights. *Finance & Stochastics* **5**, 469-486.

Fernholz, R. (2002). *Stochastic Portfolio Theory.* Springer-Verlag, New York.

Fernholz, R. and R. Garvy (1999, May 17). Diversity changes affect relative performance. *Pensions & Investments*, 112.

Fernholz, R., R. Garvy, and J. Hannon (1998, Winter). Diversity-weighted indexing. *Journal of Portfolio Management* **24**, 74-82.

Gray, S.F. (1996), Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* **42**, 27-62.

Hansen, P.R., Lunde, A. and Nason, J.M. (2003). Choosing the best volatility models: The model confidence set approach. *Oxford Bulletin of Economics and Statistics* **65**, 839-861.

Lo, A.W. and MacKinlay, A.C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* **1**, 41-66.

Malchiel, B. (1990). *A Random Walk Down to Wall Street.* Norton, New York.

Renyi, A. (1961). *On measures of entropy and information.* In *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. **1**, 547-561, UC Berkeley.

Taylor, S. (1986). *Modelling Financial Time Series.* John Wiley & Sons, Chichester.

Summary statistics of data

	Central moments				Autocorrelations		
	Mean	Stdev	Skew	Kurt	Lag 1	Lag 2	Lag 3
Diversity rates	0.5927	11.147	-0.3289	2.5784	0.9919	0.9818	0.9704
Diversity changes	-0.0209	1.2528	-0.0333	4.6663	0.1283	0.0838	-0.0377
1 mth rates	4.6573	2.1204	1.2966	5.2241	0.9556	0.9203	0.8875
1 mth changes	-0.0038	0.6095	1.0215	15.199	-0.1045	-0.0332	-0.0653
60 mth rates	6.9773	2.4603	0.9309	3.5981	0.9860	0.9696	0.9541
Spread	2.3200	1.1502	0.6745	3.6596	0.8723	0.8064	0.7391
CPI	4.2779	2.8568	1.2667	4.1458	0.9909	0.9782	0.9629
PPI	3.4385	3.7274	1.3461	4.4884	0.9843	0.9630	0.9407
HELP	85.318	22.647	0.3225	2.2329	0.9880	0.9759	0.9601
IP	3.1998	4.6552	0.8555	3.8134	0.9622	0.9015	0.8288
UE	1.3244	10.117	0.5628	4.5274	0.9764	0.9509	0.8581
S&P500	0.6939	4.2662	0.3433	4.9306	0.0093	-0.0473	0.0120

Table 1: The diversity (log-) levels and changes are from the data used in Fernholz (2002) for the US stock market. 1 month yield is from the Fama CRSP Treasury bill files. The 60 month yield is annual zero coupon bond yields from the Fama CRSP bond files. Spread refers to the difference between long and short-term interest rates. The inflation measures CPI and PPI refer to CPI inflation and PPI (Finished Goods) inflation, respectively. We calculate the inflation measure at time t using $\log(P_t/P_{t-12})$ where P_t is the (seasonal adjusted) inflation index. The real activity measures HELP, IP and UE refer to the Index of Help Wanted Advertising in Newspapers, the (seasonal adjusted) growth rate in industrial production and the unemployment rate, respectively. The growth rate in industrial production is calculated using $\log(I_t/I_{t-12})$ where I_t is the (seasonal adjusted) industrial production index. S&P500 refers to S&P500 monthly log-returns. The sample period is January 1960 to December 2001, for a total of 504 observations.

Summary statistics of Fama and French data

	Central moments				Autocorrelations		
	Mean	Stdev	Skew	Kurt	Lag 1	Lag 2	Lag 3
B.L	0.9283	4.7649	-0.2887	4.8626	0.0517	-0.0373	0.0013
B.M	1.0027	4.2724	-0.1468	5.0532	0.0032	-0.0788	-0.0144
B.H	1.1736	4.4333	-0.0091	4.7869	0.0411	0.0620	0.0024
S.L	0.9954	6.9929	-0.3365	4.8311	0.1604	-0.0368	-0.0532
S.M	1.2661	5.3251	-0.4565	6.3114	0.1835	-0.0548	-0.0537
S.H	1.4363	5.5244	-0.0656	7.4401	0.1713	-0.0810	-0.0629
Rm-Rf	0.5055	4.4409	-0.4797	5.0061	0.0526	-0.0513	-0.0008
SMB	0.1976	3.0471	0.3279	4.8244	0.1176	0.0488	-0.0730
HML	0.3433	3.1262	-0.6541	10.017	0.1575	0.0533	0.0149
Cnsmr	1.0861	4.7542	-0.3129	5.6146	0.1167	-0.0345	-0.0381
Manuf	0.9017	4.1717	-0.2667	5.3515	-0.0019	-0.0748	0.0043
HiTec	0.9716	5.4709	-0.3533	4.9449	0.0398	-0.0312	0.0525
Hlth	1.1804	5.2491	0.0791	5.2543	0.0018	0.0072	-0.0573
Other	1.0614	5.10712	-0.3135	4.5576	0.1017	-0.0393	-0.0437

Table 2: The six Fama and French benchmark portfolios constructed considering different sizes and book-to-market (big (B) and small (S) size, and high (H), neutral (M) and low (L) book-to-market), the three Fama and French risk factors (the excess return on the market minus the one-month treasury bill (Rm-Rf), the average return on three small portfolios minus the average return on the big portfolios (SMB) and the average return on two value portfolios minus the average returns on two growth portfolios (HML)) and five industry portfolios are from the web-site of K. French. The sample period is January 1960 to December 2001, for a total of 504 observations.

AR(1)-GARCH(1,1) parameter estimates

Parameter	Estimate	t |(p-value)
a	-0.0182	0.3528
b	0.1536	3.2815*
α_0	0.0713	0.6309
α_1	0.0938	2.1024
β	0.8657	8.4847*
Log-likelihood	-803.5629	
AIC	1617.126	
BIC	1638.239	
LB_5	1.4562	(0.9180)
LB_{10}	8.8528	(0.5461)
LB_5^2	0.5509	(0.9901)
LB_{10}^2	2.4639	(0.9914)
M-MAE	0.9244	
M-RMSE	1.2416	
V-MAE	1.6254	
V-RMSE	3.0317	

Table 3: Parameter estimates and related statistics for the classical AR(1)-GARCH(1,1) model (single regime model). The sample period is January 1960 to December 2001, for a total of 504 monthly observations. t-statistics are based on heteroskedastic-consistent standard errors. Asterisks denote significance at the 1% level. AIC and BIC denote the Aikaike and the bayesian-Schwarz information criteria for predictive accuracy. MAE and RMSE are classical mean absolute error and root mean squared error statistics for conditional means (M-) and conditional variances (V-). LB_i and LB_i^2 denotes the Ljung-Box statistic for serial correlation of residuals and squared residuals out to i lags. p-values are in parentheses. In the AR(1)-GARCH(1,1) model:

$$\Delta D_t \mid \Phi_{t-1} \sim \mathrm{N}(a + b\Delta D_{t-1}, \alpha_0 + \alpha_1 \varepsilon_{t-1} + \beta h_{t-1})$$

where $\varepsilon_t = \Delta D_t - \mu_t$, μ_t and h_t the conditional mean and variance, respectively.

GTS parameter estimates

		Optimal: $k = 3$ regimes	
Regime Structure	Parameter	Estimate	t |(p-value)
	a_1	-0.4131	4.6973*
	b_1	-0.0556	0.6909
$\text{B.H}_{t-1} \leq 4.05$ and $\text{S.H}_{t-1} \leq 0.076$	$\alpha_{0,1}$	0	0
	$\alpha_{1,1}$	0.0812	1.2483
	β_1	0.9112	13.231*
	a_2	-0.1495	2.1072
	b_2	0.2705	3.8271*
$\text{B.H}_{t-1} \leq 4.05$ and $\text{S.H}_{t-1} > 0.076$	$\alpha_{0,2}$	0.1437	1.0515
	$\alpha_{1,2}$	0.1105	1.5296
	β_2	0.6939	4.6457*
	a_3	0.5854	6.7063*
	b_3	0.0054	0.0691
$\text{B.H}_{t-1} > 4.05$	$\alpha_{0,3}$	0.2153	1.2997
	$\alpha_{1,3}$	0.0059	0.1849
	β_3	0.7304	5.4990*
Log-likelihood		-768.2859	
AIC		1566.572	
BIC		1629.910	
LB_5		2.8678	(0.7204)
LB_{10}		14.270	(0.1610)
LB_5^2		0.7183	(0.9820)
LB_{10}^2		2.7433	(0.9868)
M-MAE		0.8835	
M-RMSE		1.1848	
V-MAE		1.4214	
V-RMSE		2.5644	

Table 4: Parameter estimates, regime's structures and related statistics for the generalized tree-structured (GTS) model based on the use of endogenous and exogenous variables as predictors. B.H and S.H denote the Fama and French benchmark portfolios constructed using equities with big and small sizes and high book-to-market. The sample period is January 1960 to December 2001, for a total of 504 monthly observations. Goodness-of-fit statistics are the same as in Table 3.
In the full GTS model: $\Delta D_t \mid \Phi_{t-1} \sim \text{N}(\mu_t, h_t)$,

$$\mu_t = \sum_{j=1}^{k} (a_j + b_j \Delta D_{t-1})\, I_{[(D_{t-1}, D_{t-2}, \varepsilon_{t-1}, h_{t-1}, \mathbf{x}_{t-1}^{\text{ex}}) \in \mathcal{R}_j]},$$

$$h_t = \sum_{j=1}^{k} \left(\alpha_{0,j} + \alpha_{1,j}\varepsilon_{t-1}^2 + \beta_j h_{t-1}\right) I_{[(D_{t-1}, D_{t-2}, \varepsilon_{t-1}, h_{t-1}, \mathbf{x}_{t-1}^{\text{ex}}) \in \mathcal{R}_j]},$$

for a partition $\mathcal{P}$ of the state space G of $\Phi_{t-1} = \{D_{t-1}, D_{t-2}, \varepsilon_{t-1}, h_{t-1}, \mathbf{x}_{t-1}^{\text{ex}}\}$

$$\mathcal{P} = \{\mathcal{R}_1, \dots \mathcal{R}_k\},\ G = \cup_{j=1}^{k} \mathcal{R}_j,\ \mathcal{R}_i \cap \mathcal{R}_j = \emptyset\ (i \neq j).$$

Out-of-sample specification tests

In-sample

Period (No. obs.)	Statistic	AR-GARCH	Two-regime ex. GRS	GTS
January 1960-	M-RMSE	1.266	1.273	1.235
December 1979	M-MAE	0.941	0.958	0.917
(240)	V-RMSE	3.366	3.047	3.038
	V-MAE	1.731	1.690	1.534
	R^2	0.195	0.279	0.314
January 1960-	M-RMSE	1.182	1.194	1.143
December 1993	M-MAE	0.880	0.891	0.837
(408)	V-RMSE	2.946	3.039	2.560
	V-MAE	1.452	1.606	1.298
	R^2	0.203	0.047	0.351

Out-of-sample

Period (No. obs.)	Statistic	AR-GARCH	Two-regime ex. GRS	GTS
January 1980-	M-RMSE	1.235	1.232	1.212
December 2001	M-MAE	0.930	0.921	0.912
(264)	V-RMSE	3.001	2.848	2.663
	V-MAE	1.649	1.697	1.498
	R^2	0.179	0.193	0.256
	ONL	429.4	431.7	426.9
January 1994-	M-RMSE	1.483	1.473	1.414
December 2001	M-MAE	1.139	1.139	1.098
(96)	V-RMSE	3.624	3.546	3.267
	V-MAE	2.119	2.047	1.743
	R^2	0.254	0.207	0.225
	ONL	174.3	186.0	173.9

Table 5: Predictive performance measures for a classical, single-regime AR(1)-GARCH(1,1) model, an extended two-regime generalized regime-switching (GRS) models and the generalized tree-structured (GTS) model. The table reports root mean squared errors (RMSE) and mean absolute errors (MAE) for conditional means (M-) and variances (V-), and R^2 between actual volatility ε_t^2, where $\varepsilon_t = \Delta D_t - \mu_t$, and the conditional variance h_t. In addition, we also consider as a measure of predictive performance the out-of-sample negative likelihood (ONL). Parameters are estimated over the in-sample period, and held fixed over the out-of-sample period. Low is better for all statistics, except for the R^2. Parameter estimates are based on the Fernholz (2002) data set of monthly diversity changes for the whole US stock market. The sample extends from January 1960 to December 2001, for a total of 504 observations.

Panel A. January 1980 - December 2001

	EPA t-type test statistic		EPA sign-type test statistic	
	range	semi-quadratic	range	semi-quadratic
1st step:	1.056 (0.534)	2.185 (0.485)	1.678 (0.072)	3.731 (0.290)
2nd step:	–	–	0.934 (0.306)	–

Model	Worst performing index			
	t-type test		sign-type test	
	1st step	2nd step	1st step	2nd step
AR(1)-GARCH(1,1)	0.3531	–	1.0638	–
extended GRS	0.7202	–	0.3592	0.9335
GTS	-1.1553	–	-1.5120	-0.9335

Panel B. January 1994 - December 2001

	EPA t-type test statistic		EPA sign-type test statistic	
	range	semi-quadratic	range	semi-quadratic
1st step:	1.156 (0.482)	1.985 (0.495)	1.065 (0.523)	1.486 (0.648)
2nd step:	–	–	–	–

Model	Worst performing index			
	t-type test		sign-type test	
	1st step	2nd step	1st step	2nd step
AR(1)-GARCH(1,1)	1.2567	–	1.0719	–
extended GRS	-0.2262	–	0	–
GTS	-0.6977	–	-0.8865	–

Table 6: Testing differences of out-of-sample M-MAE performance terms among the different models. Values of equal predictive ability (EPA) tests using the range statistic T_R and the semi-quadratic statistic T_{SQ} and worst performing index results for the construction of the confidence model sets for the first out-of-sample specification test from January 1980 to December 2001 (264 observations, Panel A) and for the second one from January 1994 to December 2001 (96 observations, Panel B). Corresponding P-values computed using a block-bootstrap procedure are given between parentheses. If the null hypothesis of EPA is rejected, the model with the largest worst performing index value is eliminated.

Panel A. January 1980 - December 2001

	EPA t-type test statistic		EPA sign-type test statistic	
	range	semi-quadratic	range	semi-quadratic
1st step:	2.251 (0.039)	8.631 (0.055)	2.257 (0.046)	6.614 (0.091)
2nd step:	1.673 (0.086)	2.801 (0.086)	0.729 (0.454)	0.532 (0.472)

Model	Worst performing index			
	t-type test		sign-type test	
	1st step	2nd step	1st step	2nd step
AR(1)-GARCH(1,1)	0.6458	1.6734	-0.4225	0.7293
extended GRS	1.9292	–	1.8488	–
GTS	-2.5266	-1.6734	-0.9343	-0.7293

Panel B. January 1994 - December 2001

	EPA t-type test statistic		EPA sign-type test statistic	
	range	semi-quadratic	range	semi-quadratic
1st step:	2.9747 (0.009)	11.76 (0.024)	3.935 (0)	18.84 (0)
2nd step:	1.670 (0.048)	2.787 (0.088)	1.825 (0.028)	3.329 (0.036)

Model	Worst performing index			
	t-type test		sign-type test	
	1st step	2nd step	1st step	2nd step
AR(1)-GARCH(1,1)	1.7786	–	2.5691	–
extended GRS	0.701	1.6695	0.9170	1.8247
GTS	-2.6077	-1.6695	-3.4218	-1.8247

Table 7: Testing differences of out-of-sample V-MAE performance terms among the different models. Values of equal predictive ability (EPA) tests using the range statistic T_R and the semi-quadratic statistic T_{SQ} and worst performing index results for the construction of the confidence model sets for the first out-of-sample specification test from January 1980 to December 2001 (264 observations, Panel A) and for the second one from January 1994 to December 2001 (96 observations, Panel B). Corresponding P-values computed using a block-bootstrap procedure are given between parentheses. If the null hypothesis of EPA is rejected, the model with the largest worst performing index value is eliminated.

Probability for predicting correct diversity directions

Rule 1: *if median of predicted* D_{t+1} *distribution* $> D_t$ $\Rightarrow$ *diversity increases*
else $\Rightarrow$ *diversity decreases.*

Diversity direction	Model		
	AR(1)-GARCH(1,1)	GTS	Ex. GRS
increase	0.568	0.636	0.482
decrease	0.552	0.599	0.498

Rule 2: *if median of predicted* D_{t+1} *distribution* $> D_t$ and
D_t- *5%-quantile of* $D_{t+1} <$ *95%-quantile of* $D_{t+1} - D_t$ $\Rightarrow$ *diversity increases*
else if median of predicted D_{t+1} *distribution* $\leq D_t$ and
D_t- *5%-quantile of* $D_{t+1} \geq$ *95%-quantile of* $D_{t+1} - D_t$ $\Rightarrow$ *diversity decreases*
else $\Rightarrow$ *no decision.*

Diversity direction	Model		
	AR(1)-GARCH(1,1)	GTS	Ex. GRS
increase	0.568	0.686	0.482
decrease	0.583	0.683	0.424
prob. to have a decision	0.777	0.861	0.438

Table 8: Probabilities for a correct prediction of diversity directions using the generalized tree-structured (GTS) model in comparison to the classical AR(1)-GARCH(1,1) model and the extended generalized regime-switching (Ex. GRS) model. Results are reported for two different rules based (i) only on the median of the predicted one-month-ahead diversity distribution and (ii) on different quantiles of the predicted one-month ahead diversity distribution. The backtesting period goes from January 1975 to December 2001, for a total of 324 monthly observations.

Panel A: performance statistics

Portfolios	Annualized Return	Annualized Volatility	Total Sharpe Ratio
Cap. S&P500	14.05%	15.26%	58.835%
Div. S&P500	14.42%	15.50%	60.336%
GTS Strategy	14.66%	15.23%	61.811%

Panel B: t-type and sign-type tests

Portfolio strategy 1	Portfolio strategy 2	One-year Sharpe Ratios	
		t-type test	sign-type test
GTS Strategy	Cap. S&P500	2.019 (0.022)	2.491 (0.006)
Div. S&P500	Cap. S&P500	1.293 (0.098)	0.505 (0.301)

Table 9: Panel A: Profitability results for the cap-weighted S&P500 portfolio (Cap. S&P500), the diversity-weighted S&P500 portfolio (Div. S&P500) and the portfolio strategy described in Section 6 based on Generalized Tree Structured diversity direction forecast (GTS Strategy). The backtesting period goes from January 1975 to December 2001, for a total of 324 monthly observations. Panel B: Testing differences of rolling one-year Sharpe Ratios between portfolio strategy 1 and portfolio strategy 2 beginning January 1996 to the end of the period. The values of t-type and sign-type test statistics adapted to the case of dependent observations are summarized. The corresponding P-values are given between parentheses.

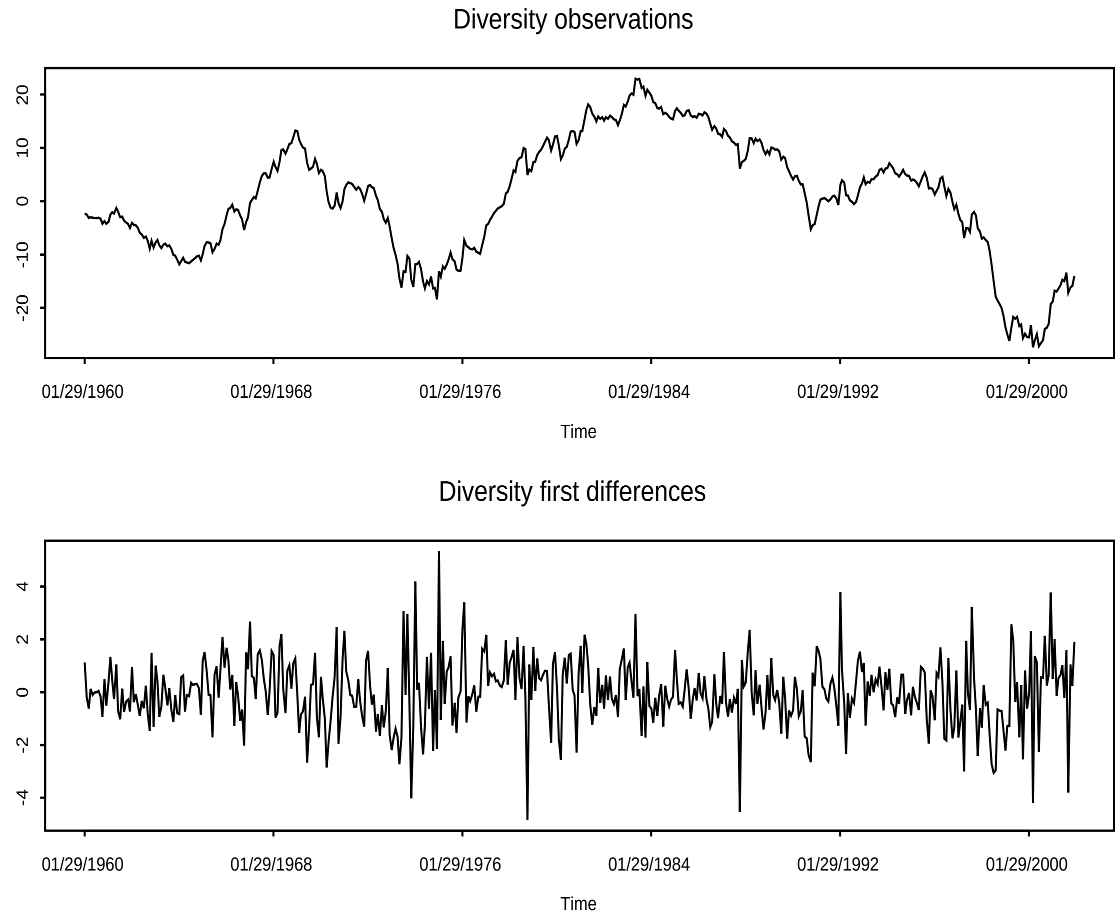
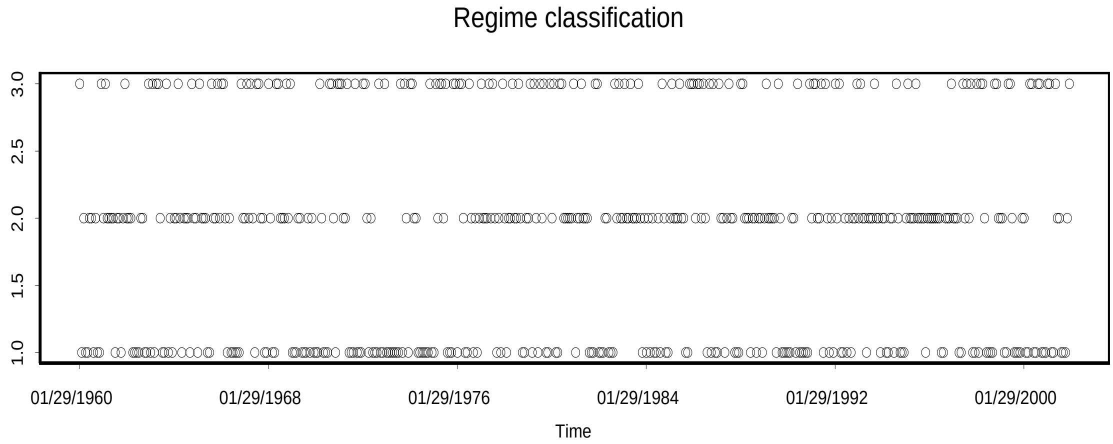


Figure 1: The top panel contains a time series of monthly diversity observations. The first differences of this series are shown in the bottom panel. The sample period is January 1960 to December 2001, for a total of 504 observations.



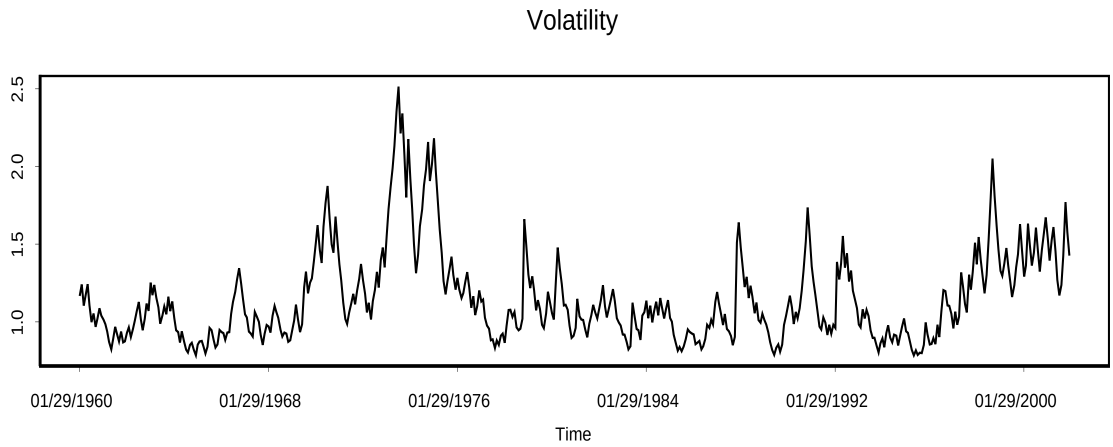


Figure 2: The top panel contains a time series plot of the regime classification of the observation at time t (i.e., $I_t = \{1, 2, 3\}$) according to the generalized tree-structured (GTS) model. The regime classification is based on all available information at time $t-1$, using as predictors all the variables introduced in Tables 1 and 2. The bottom panel contains a time series plot of the conditional standard deviation of changes in the diversity based on the generalized tree-structured (GTS) model. Parameter estimates are based on the Fernholz (2002) data set of diversity changes for the whole US stock market. The sample period is January 1960 to December 2001, for a total of 504 monthly observations.
Optimal regimes of the GTS model:

$$\begin{aligned}
&\text{Regime 1:} \quad \mathcal{R}_1 = \{\text{B.H}_{t-1} \leq 4.05 \text{ and } \text{S.H}_{t-1} \leq 0.076\} \\
&\text{Regime 2:} \quad \mathcal{R}_1 = \{\text{B.H}_{t-1} \leq 4.05 \text{ and } \text{S.H}_{t-1} > 0.076\} \\
&\text{Regime 3:} \quad \mathcal{R}_3 = \{\text{B.H}_{t-1} > 4.05\}
\end{aligned}$$

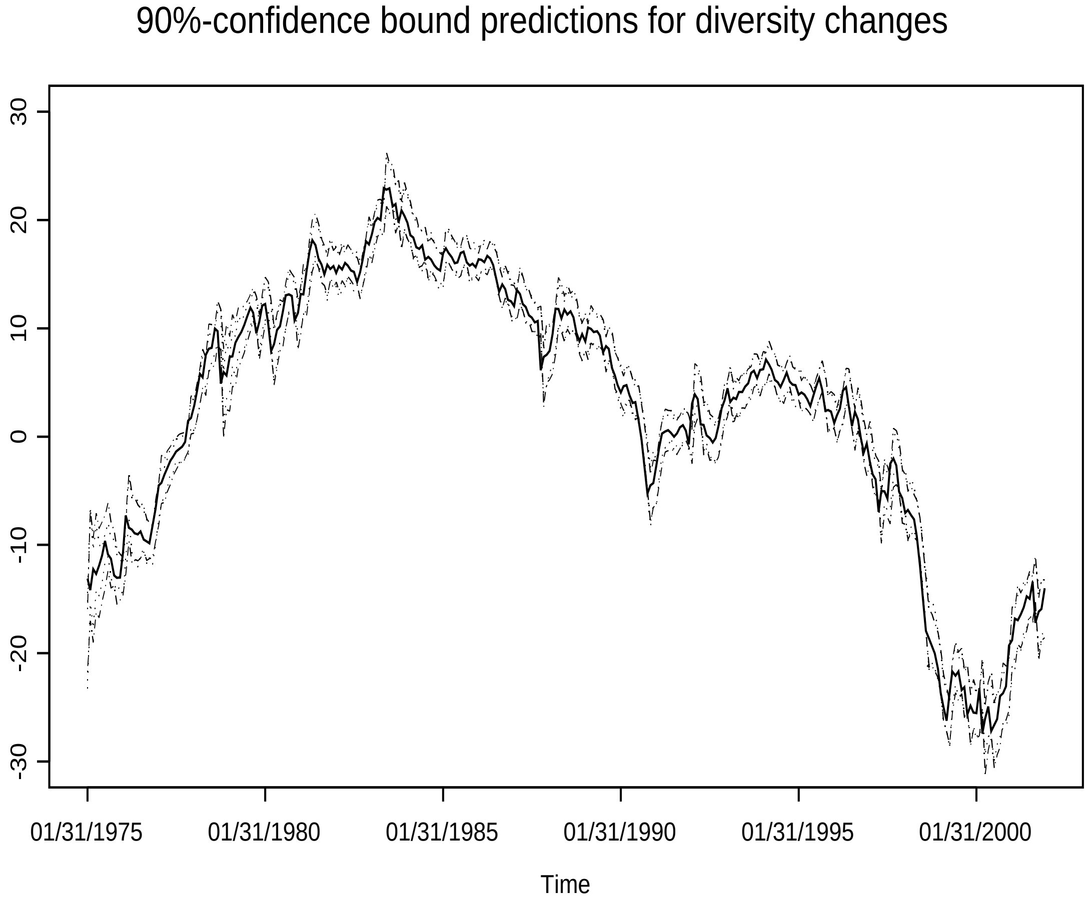


Figure 3: Out-of-sample one-month-ahead 90% diversity interval estimates. The straight line is the realized US equity market diversity process, based on the data used in Fernholz (2002). The two dotted lines are the estimated upper and lower diversity quantiles when using the GTS model with regime-structure specified in Table 4. The two dot-dashed lines are the estimated upper and lower diversity quantiles when using the classical AR(1)-GARCH(1,1) model. The backtesting period goes from January 1975 to December 2001, for a total of 324 monthly observations.

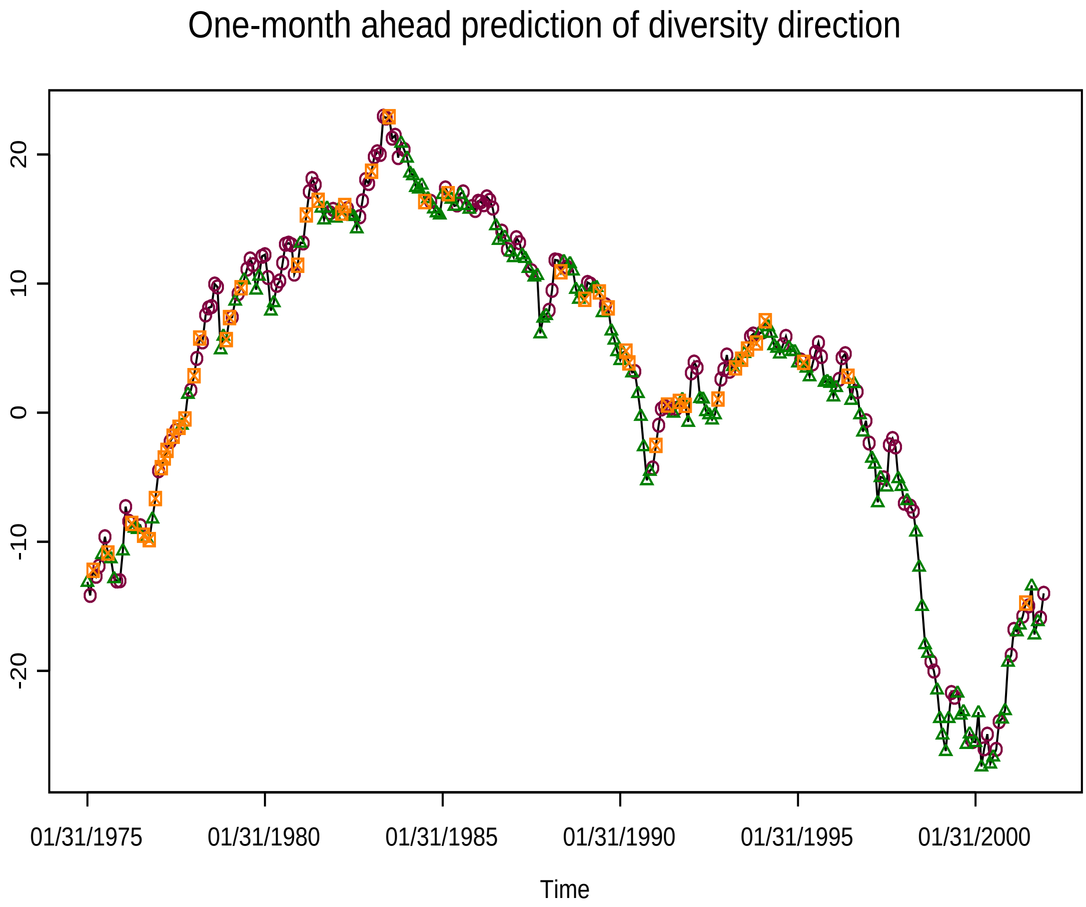


Figure 4: Out-of-sample one-month-ahead predictions of diversity directions using the generalized tree-structured (GTS) GARCH model. Decisions are taken according to rule 2. The straight line is the realized US equity market diversity process, based on the data used in Fernholz (2002). Circles, triangles and squares indicate observations where the rule predicts an increase in diversity, predicts a decrease in diversity and does not yield a decision, respectively. The backtesting period goes from January 1975 to December 2001, for a total of 324 monthly observations.