Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 25 (2009) 716–733
international journal of forecasting
www.elsevier.com/locate/ijforecast

Introduction

# Decision making and planning under low levels of predictability

Spyros Makridakis[a,*], Nassim Taleb[b,1]

[a] *INSEAD, Boulevard de Constance, 77305 Fontainebleau, France*
[b] *Polytechnic Institute of NYU, Department of Finance and Risk Engineering, Six MetroTech Center, Rogers Hall 517, Brooklyn, NY 11201, USA*

**Abstract**

This special section aims to demonstrate the limited predictability and high level of uncertainty in practically all important areas of our lives, and the implications of this. It summarizes the huge body of solid empirical evidence accumulated over the past several decades that proves the disastrous consequences of inaccurate forecasts in areas ranging from the economy and business to floods and medicine. The big problem is, however, that the great majority of people, decision and policy makers alike, still believe not only that accurate forecasting is possible, but also that uncertainty can be reliably assessed. Reality, however, shows otherwise, as this special section proves. This paper discusses forecasting accuracy and uncertainty, and distinguishes three distinct types of predictions: those relying on patterns for forecasting, those utilizing relationships as their basis, and those for which human judgment is the major determinant of the forecast. In addition, the major problems and challenges facing forecasters and the reasons why uncertainty cannot be assessed reliably are discussed using four large data sets. There is also a summary of the eleven papers included in this special section, as well as some concluding remarks emphasizing the need to be rational and realistic about our expectations and avoid the common delusions related to forecasting.


*Keywords:* Forecasting; Accuracy; Uncertainty; Low level predictability; Non-normal forecasting errors; Judgmental predictions

## 1. Introduction

The unknown future is a source of anxiety, giving rise to a strong human need to predict it in order to reduce, or ideally eliminate, its inherent uncertainty. The demand for forecasts has created an ample supply of "experts" to fulfill it, from augurs and astrologists to economists and business gurus. Yet the track record of almost all forecasters is dismal. Worse, the accuracy of "scientific" forecasters is often no better than that of simple benchmarks (e.g. today's value, or some average). In addition, the basis of their predictions is often as doubtful as those of augurs and astrologists. In the area of economics, who predicted the subprime and credit crunch crises, the Internet bubble, the Asian contagion, the real estate and savings and loans crises, the Latin American lending calamity, and the other major disasters? In business, who "predicted" the collapse of Lehman Brothers, Bear Stearns, AIG, Enron or WorldCom (in the USA), and Northern Rock,

* Corresponding editor. Tel.: +30 6977661144.
*E-mail addresses:* smakrid@otenet.gr (S. Makridakis), nnt@fooledbyrandomness.com (N. Taleb).
[1] Tel.: +1 718 260 3599; fax: +1 718 260 3355.

0169-2070/$ - see front matter 
doi:10.1016/j.ijforecast.2009.05.013

Royal Bank of Scotland, Parmalat or Royal Ahold (in Europe); or the practical collapse of the entire Iceland economy? In finance, who predicted the demise of LTCM and Amaranth, or the hundreds of mutual and hedge funds that close down every year after incurring huge losses? And these are just the tip of the iceberg.

In the great majority of situations, predictions are never accurate. As is mentioned by Orrell and McSharry (2009), the exception is with mechanical systems in physics and engineering. The predictability of practically all complex systems affecting our lives is low, while the uncertainty surrounding our predictions cannot be reliably assessed. Perpetual calendars in handheld devices, including watches, can show the exact rise and set of the sun and the moon, as well as the phases of the moon, up to the year 2099 and beyond. It is impressive that such small devices can provide highly accurate forecasts. For instance, they predict that on April 23, 2013, in Greece:

> The sun will rise at 5:41 and set at 7:07
> The moon will rise at 4:44 and set at 3:55
> The phase of the moon will be more than 3/4 full, or 3 days from full moon.

These forecasts are remarkable, as they concern so many years into the future, and it is practically certain that they will be perfectly accurate so many years from now. The same feeling of awe is felt when a spaceship arrives at its exact destination after many years of traveling through space, when a missile hits its precise target thousands of kilometers away, or when a suspension bridge spanning 2000 m can withstand a strong earthquake, as predicted in its specifications.

Physics and engineering have achieved amazing successes in predicting future outcomes. By identifying exact patterns and precise relationships, they can extrapolate or interpolate them, to achieve perfect, error free forecasts. These patterns, like the orbits of celestial objects, or relationships like those involving gravity, can be expressed with exact mathematical models that can then be used for forecasting the positions of the sun and the moon on April 23, 2013, or firing a missile to hit a desired target thousands of kilometers away. The models used make no significant errors, even though they are simple and can often be programmed into hand-held devices.

Predictions involving celestial bodies and physical law type relationships that result in near-perfect, error free forecasts are the exception rather than the rule—and forecasting errors are of no serious consequence, thanks to the "thin-tailedness" of the deviations. Consider flipping a coin 10 times; how many heads will appear? In this game there is no certainty about the outcome, which can vary anywhere from 0 to 10. However, even with the most elementary knowledge of probability, the best forecast for the number of heads is 5, the most likely outcome, which is also the average of all possible ones. It is possible to work out that the chance of getting exactly five heads is 0.246, or to compute the corresponding probability for any other number.

The distribution of errors, when a coin is flipped 10 times and the forecast is 5 heads, is shown in Fig. 1, together with the actual results of 10,000 simulations. The fit between the theoretical and actual results is remarkable, signifying that uncertainty can be assessed correctly when flipping a coin 10 times.

Games of chance like flipping coins, tossing dice, or spinning roulette wheels have an extremely nice property: the events are independent, while the probability of success or failure is constant over all trials. These two conditions allow us to calculate both the best forecast and the uncertainty associated with various occurrences. Moreover, when $n$, the number of trials, is large, the central limit theorem applies, guaranteeing that the distribution around the mean, the most likely forecast, can be approximated by a normal curve, knowing that the larger the value of $n$ the better the approximation. Even when a coin is tossed 10 times ($n = 10$), the distribution of errors, with a forecast of 5, can be approximated pretty well with a normal distribution, as can be seen in Fig. 1.

With celestial bodies and physical law relationships, we can achieve near-perfect predictions. With games of chance, we know that there is no certainty, but we can figure out the most appropriate forecasts and estimate precisely the uncertainty involved. In the great majority of real life situations, however, there is always doubt as to which is the "best" forecast, and, even worse, the uncertainty surrounding a forecast cannot be assessed, for three reasons. First, in most cases, errors are not independent of one another; their variance is not constant, while their distribution cannot be assured to follow a normal curve—which means that the variance itself will be either intractable or a poor indicator of potential errors, what has been

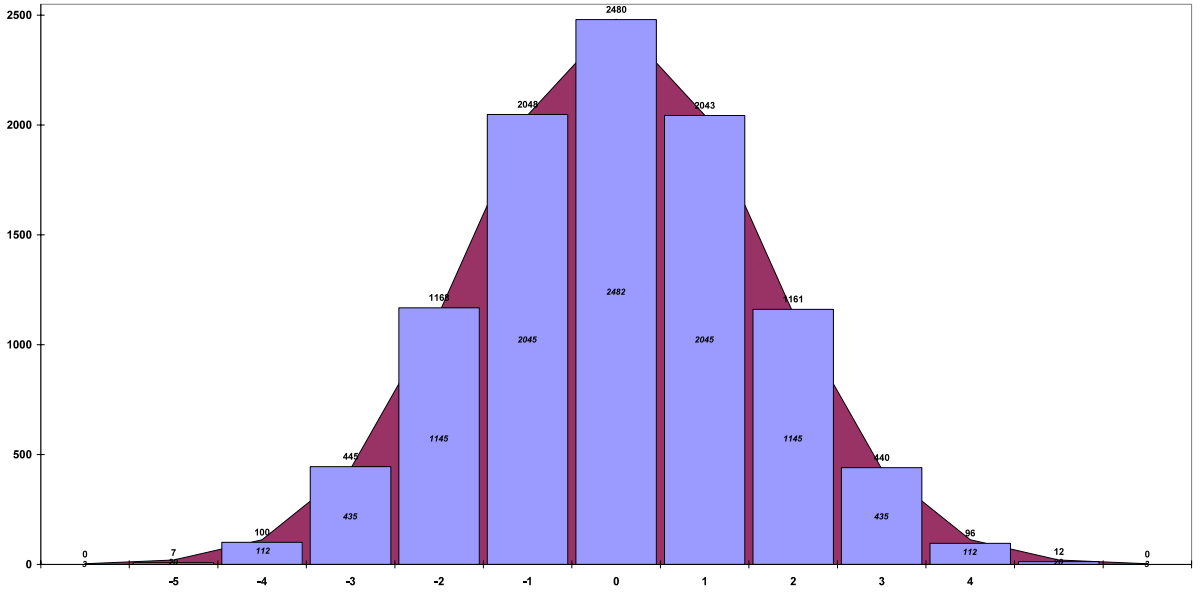



Fig. 1. The erros assuming 5 heads when a coin is flipped 10 times (10,000 replications).

called "wild randomness" by Mandelbrot (1963). Second, there is always the chance of highly unlikely or totally unexpected occurrences materializing — and these can play a large role (Taleb, 2007). Third, there is a severe problem outside of artificial setups, such as games: probability is not observable, and it is quite uncertain which probabilistic model to use.

In addition, we must remember that we do not forecast for the sake of forecasting, but for some specific purpose, so we must realize that some forecast errors can cause harm or missed opportunities, while others can be benign. So to us, any analysis of forecasting needs to take the practical dimension into account: both the consequences of forecast errors and the fragility and reliability of predictions. In the case of low reliability, we need to know what to do, depending on the potential losses and opportunities involved.

## 2. The accuracy and uncertainty in forecasting

This section examines each of two distinct issues associated with forecasting: the accuracy of predictions and the uncertainty surrounding them. In doing so, it distinguishes three types of predictions: (a) those involving patterns, (b) those utilizing relationships, and (c) those based primarily on human judgment. Each of these three will be covered using information from empirical studies and three concrete examples, where ample data are available.

### 2.1. The accuracy when forecasting patterns

The M-Competitions have provided abundant information about the accuracy of all major time series forecasting methods aimed at predicting patterns. Table 1 lists the overall average accuracies for all forecasting horizons for the 4004 series used in the M-Competition (Makridakis et al., 1982) and the M3-Competition (Makridakis & Hibon, 2000). The table includes five methods. Naïve 1 is a simple, readily available benchmark. Its forecasts for all horizons up to 18 are the latest available value. Naïve 2 is the same as Naïve 1 except that the forecasts are appropriately seasonalized for each forecasting horizon. Single exponential smoothing is a simple method that averages the most recent values, giving more weight to the latest ones, in order to eliminate randomness. Dampen exponential smoothing is similar to single, except that it first smooths the most recent trend in the data to remove randomness and then extrapolates and dampens, as its names implies, such a smoothed trend. Single smoothing was found to be highly accurate in the M- and M3-Competitions, while dampen was one

Table 1
MAPE[a] (average absolute percentage error) of various methods and percentage improvements.

| | MAPEs: Forecasting horizons | | | | Improvement (in Avg. MAPE) | % Improvement in Avg. MAPE: | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1st | 6th | 18th | Avg. MAPE (1–18 horizons) | over Naïve1 | Naïve2 over Naïve1 | Single over Naïve2 | Dampen over Single | Box–Jenkins over Dampen |
| Naïve1 | 11.7% | 18.9% | 24.6% | 17.9% | | | | | |
| Naïve2 | 10.2% | 16.9% | 22.1% | 16.0% | 1.9% | 11.6% | | | |
| Single exponential smoothing | 9.3% | 16.1% | 21.1% | 15.0% | 2.9% | | 6.4% | | |
| Dampen exponential smoothing | 8.7% | 15.0% | 19.2% | 13.6% | 4.3% | | | 8.1% | |
| The Box–Jenkins methodology to ARIMA models | 9.2% | 14.9% | 19.8% | 14.2% | 3.7% | | | | **−2.5%** |

[a] All MAPEs and % improvements are symmetric; that is, the divisor is: (Method1 − Method2)/(0.5∗Method1 + 0.5∗Method2).

of the best methods in each of these competitions. Finally, the Box–Jenkins methodology with ARIMA models, a statistically sophisticated method that identifies and fits the most appropriate autoregressive and/or moving average model to the data, was less accurate overall than dampen smoothing.

Table 1 shows the MAPEs of these five methods for forecasting horizons 1, 6 and 18, as well as the overall average of all 18 forecasting horizons. The forecasting errors start at around 10% for one period ahead forecasts, and almost double for 18 periods ahead. These huge errors are typical of what can be expected when predicting series similar to those of the M- and M3-Competitions (the majority consisting of economic, financial and business series). Table 1 also shows the improvements in MAPE of the four methods over Naïve 1, which was used as a benchmark. For instance, Naïve 2 is 1.9% more accurate than Naïve 1, a relative improvement of 11.6%, while dampen smoothing is 4.3% more accurate than Naïve 1, a relative improvement of 27.2%.

The right part of Table 1 provides information about the source of the improvements in MAPE. As the only difference between Naïve 1 and Naïve 2 is that the latter captures the seasonality in the data, this means that the 11.6% improvement (the biggest of all) brought by Naïve 2 is due to predicting the seasonality in the 4004 series. An additional improvement of 6.4% comes from single exponential smoothing, which averages the most recent values in order to eliminate random noise. The final improvement of 8.1%, on top of seasonality and randomness, is due to dampen smoothing, which eliminates the randomness in the most recent trend (we can call this trend the momentum of the series). Finally, the Box–Jenkins method is less accurate than dampen smoothing by 0.6%, or, in relative terms, has a decrease of 2.5% in overall forecasting accuracy.

As dampen smoothing cannot predict turning points, we can assume that the Box–Jenkins does not either, as it is less accurate than dampen. In addition, dampen smoothing is considerably more accurate than Holt's exponential smoothing (not shown in Table 1), which extrapolates the most recent smoothed trend, without dampening. This finding indicates that, on average, trends do not continue uninterrupted, and should not, therefore, be extrapolated. Cyclical turns, for instance, reverse established trends, with the consequence of huge errors if such trends are extrapolated assuming that they will continue uninterrupted.

### 2.2. *The uncertainty when forecasting patterns*

What is the uncertainty in the MAPEs shown in Table 1? Firstly, uncertainty increases together with the forecasting horizon. Secondly, such an increase is bigger than that postulated theoretically. However, it has been impossible to establish the distribution of forecasting errors in a fashion similar to that shown in Fig. 1 or Table 1, as the number of observations in the series in the M-Competitions is not large enough. For this reason, we will demonstrate the uncertainty in

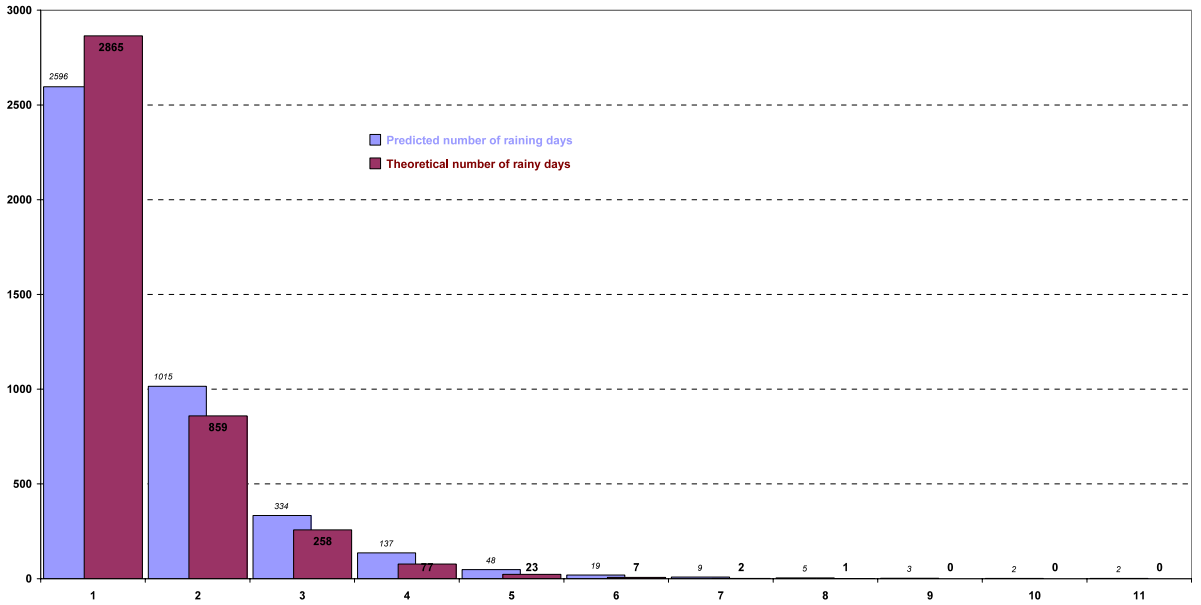



Fig. 2. Predicted and theoretical number of rainy days.

forecasting by using four long series, allowing us to look at the distributions of forecasting errors.

Rainfall data from January 1, 1971 to May 6, 2008 ($n = 13{,}648$) in Amsterdam show that the chance of rain on any given day is very close to that of flipping a coin (0.506, to be precise). Since it rains more during some periods than during others (i.e. events are not independent), we can use Naïve 1 to improve our ability to forecast. By doing so, we increase the probability of correctly predicting rain from 0.506, assuming that rainy days are independent of each other, to 0.694. Fig. 2 shows the theoretical and actual forecasting errors using Naïve 1. The fit between the theoretical and actual errors is remarkable, indicating that we can estimate the uncertainty of the Naïve 1 model with a high degree of reliability when using the theoretical estimates. It seems that in binary forecasting situations, such as rain or no rain, uncertainty can be estimated reliably.

Fig. 3 shows the average daily temperatures in Paris for each day of the year, using data from January 1, 1900 to December 31, 2007. Fig. 3 shows a smooth pattern, with winter days having the lowest temperatures and summer days the highest ones, as expected. Having identified and estimated this seasonal pattern, the best forecast suggested by meteorologists for, say, January 1, 2013, is the average of the temperatures for all 108 years of data, or 3.945 °C.

However, it is clear that the actual temperature on 1/1/2013 will, in all likelihood, be different from this average. An idea of the possible errors or uncertainty around this average prediction can be inferred from Fig. 4, which shows the 108 errors if we use 3.945, the average for January 1, as the forecast. These errors vary from −13 to 8 degrees, with most of them being between 7 and 11 °C. The problem with Fig. 4, however, is that the distribution of errors does not seem to be well behaved. This may be because we do not have enough data (a problem with most real life series) or because the actual distribution of errors is not normal or even symmetric. Thus, we can say that our most likely prediction is 3.945 degrees, but it is difficult to specify the range of uncertainty in this example with any degree of confidence.

The number of forecasting errors increases significantly when we make short term predictions, like the temperature tomorrow, and use Naïve 1 as the forecast (meteorologists can improve the accuracy of predicting the weather over that of Naïve 1 for up to three days ahead). If we use Naïve 1, the average error is zero, meaning that Naïve 1 is an unbiased forecasting model, with a standard deviation of 2.71 degrees and a range of errors from −11.2 to 11 degrees. The

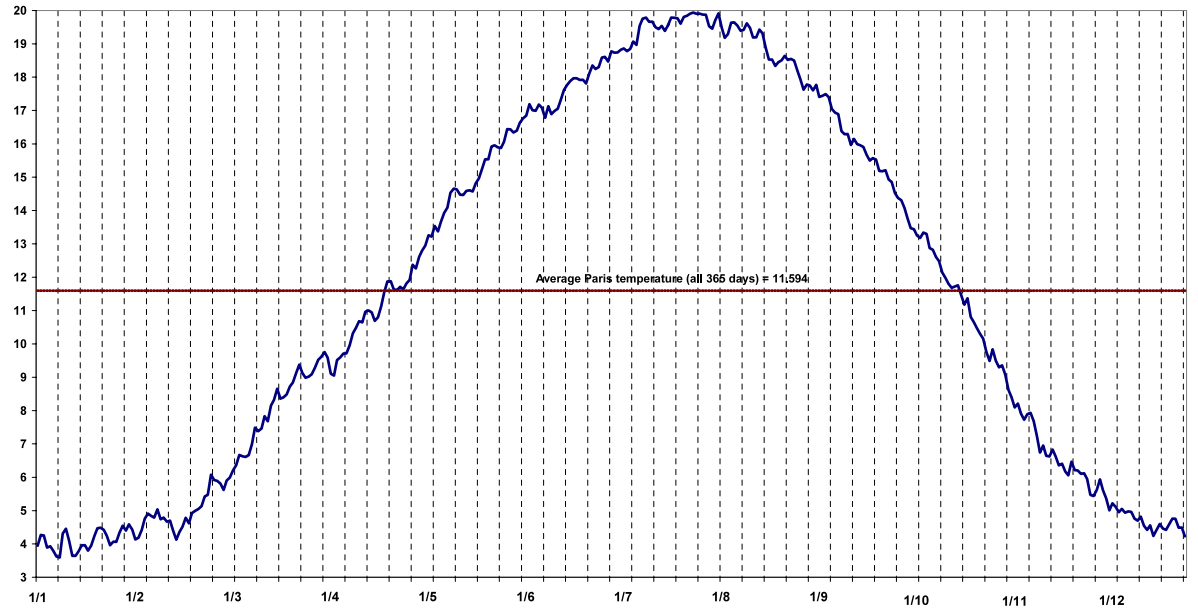



Fig. 3. Average daily temperatures in Paris: 1900 to 2007.

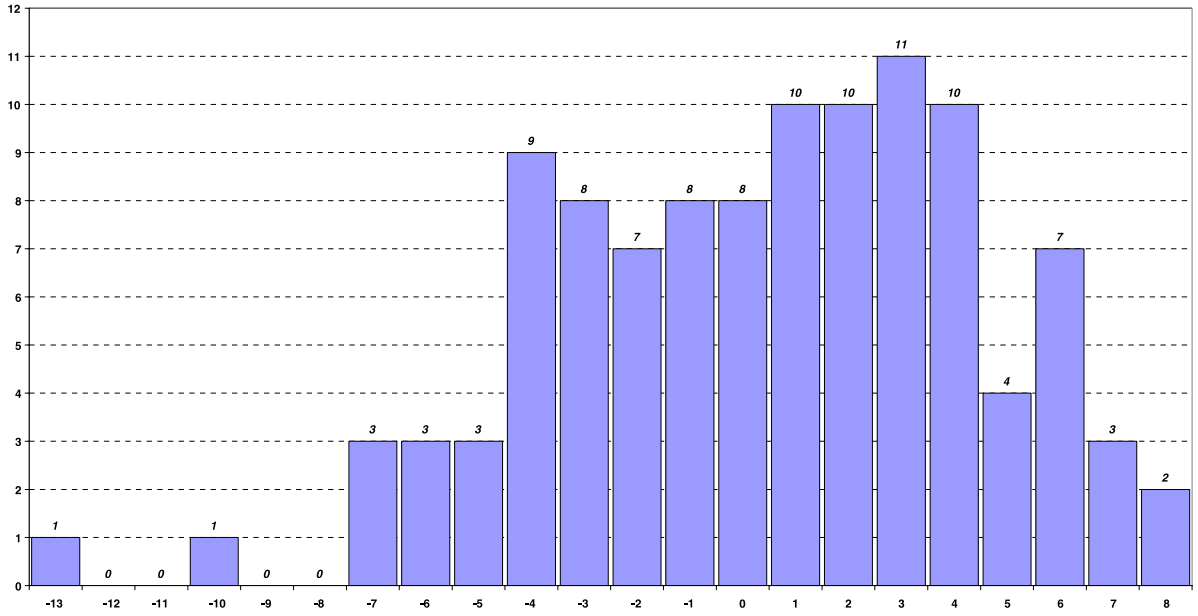



Fig. 4. Errors from the mean in daily temperatures (in Celsius) on January 1st: 1900–2007.

distribution of these errors is shown in Fig. 5, superimposed on a normal curve.

Two observations come from Fig. 5. First, there are more errors in the middle of the distribution than postulated by the normal curve. Second, the tails of the error distribution are much fatter than if they were following a normal curve. For example, there are 14 errors of temperature less than −8.67 degrees, corresponding to more than 4 standard deviations from the mean. This is a practical impossibility if the actual

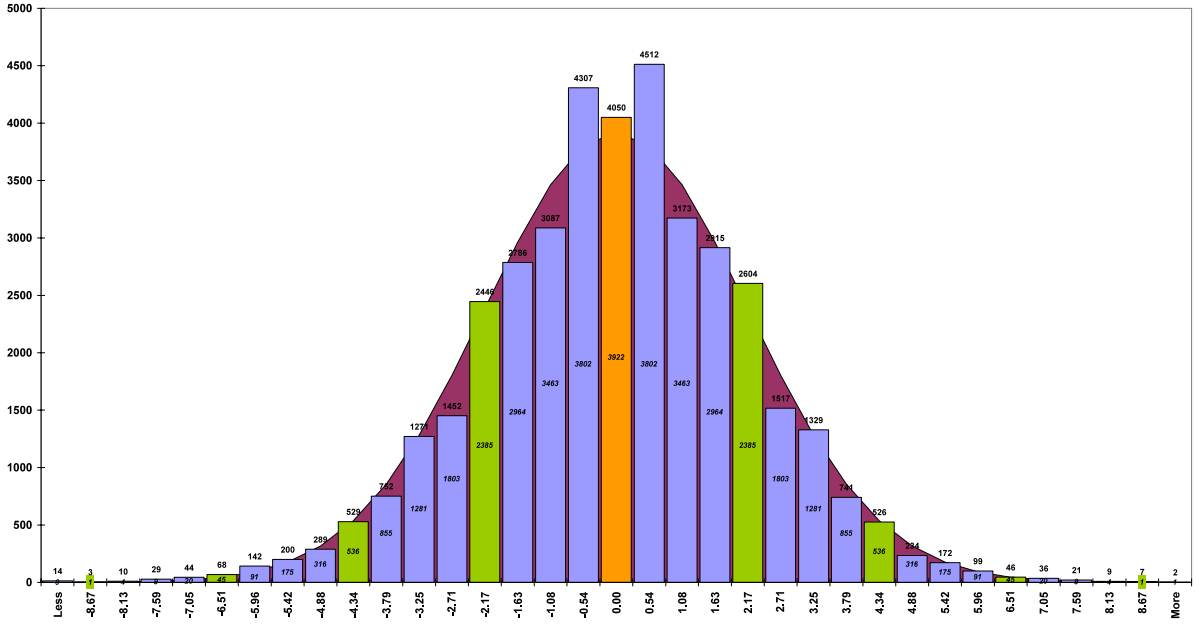



Fig. 5. Paris temperatures 1900–2007: Daily changes.

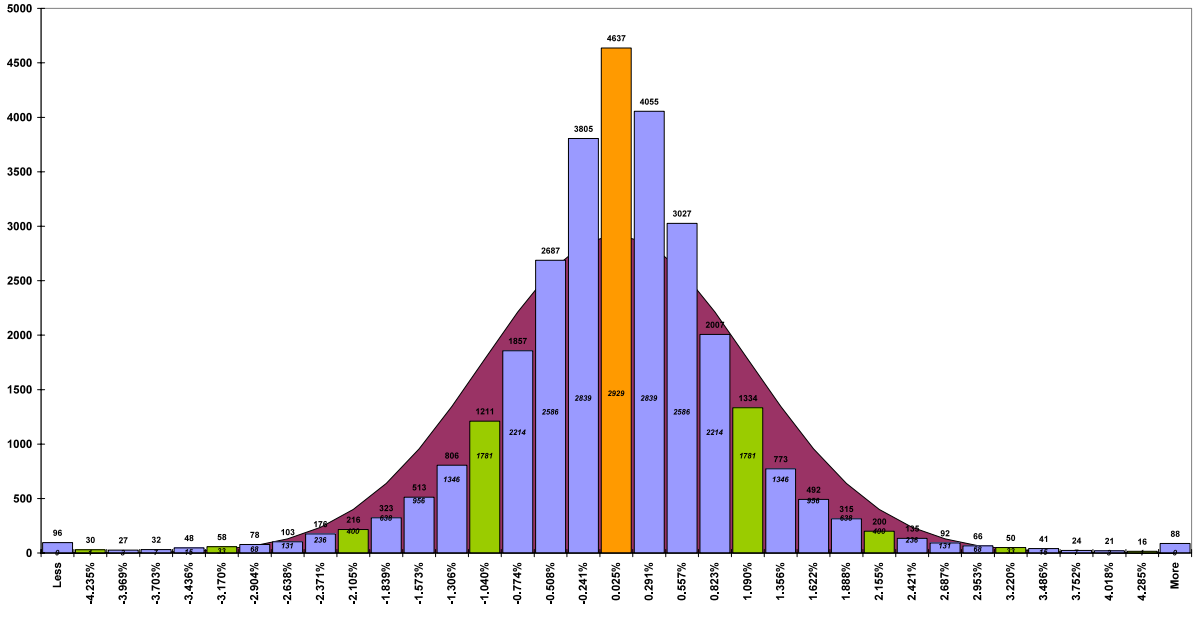



Fig. 6. The daily forecasting errors for the DJIA, 1900–2007.

distribution was a normal one. Similarly, there are 175 errors outside the limits of the mean ±3 standard deviations, versus 69 if the distribution was normal. Thus, can we say that the distribution of errors can be approximated by a normal curve? The answer is complicated, even though the differences are not as large as those of Fig. 6, describing the errors of the next example: the DJIA.

Table 2
DJIA 1900–2000: Worst-best daily returns.

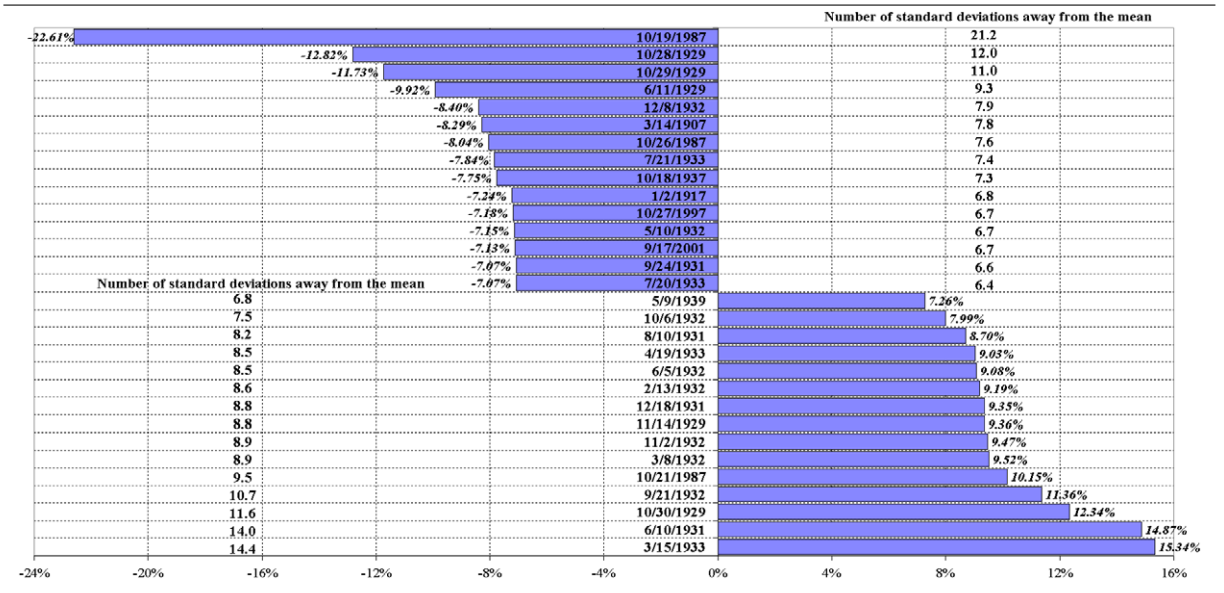



Fig. 6 shows the same information as Fig. 5, except that it refers to the values of the DJIA when Naïve 1 is used as the forecasting model. The data ($n = 29{,}339$) cover the same period as the Paris temperatures, January 1, 1900 to December 31, 2007 (there are fewer observations because the stock market is not open during weekends and holidays). The actual distribution of Fig. 6 also does not follow a normal curve. The middle values are much higher than those of Fig. 5, while there are many more values outside the limits of $\pm 4$ standard deviations from the mean. For instance, there are 184 values below and above 4 standard deviations, while there should not be any such values if the distribution was indeed normal.[2]

Table 2 further illustrates the long, fat tails of the errors of Fig. 6 by showing the 15 smallest and largest errors and the number of standard deviations away from the mean such errors correspond to (they range from 6.4 to 21.2 standard deviations). Such large errors could not have occurred in many billions of years if they were part of a normal distribution.

The fact that the distribution of errors in Fig. 6 is much more exaggerated than that of Fig. 5 is due to the human ability to influence the DJIA, which is not the case with temperatures. Such an ability, together with the fact that humans overreact to both good and bad news, increases the likelihood of large movements in the DJIA. There is no other way to explain the huge increases/decreases shown in Table 2, as it is not possible for the capitalization of all companies in the DJIA to lose or gain such huge amounts in a single day by real factors.

Another way to explain the differences between the two figures is that temperature is a physical random variable, subject to physical laws, while financial markets are informational random variables that can take any value without restriction—there are no physical impediments to the doubling of a price. Although physical random variables can be non-normal owing to nonlinearities and cascades, they still need to obey some structure, while informational random variables do not have any tangible constraint.

[2] Departure from normality is not accurately measured by counting the number of observations in excess of 4, 5, or 6 standard deviations (sigmas), but in looking at the contribution of large deviations to the total properties. For instance, the Argentine currency froze for a decade in the 1990s, then had a large jump. Its kurtosis was far more significant than the Paris weather, although we only had one single deviation in excess of 4 sigmas. This is the problem with financial measurements that discard the effect of a single jump.

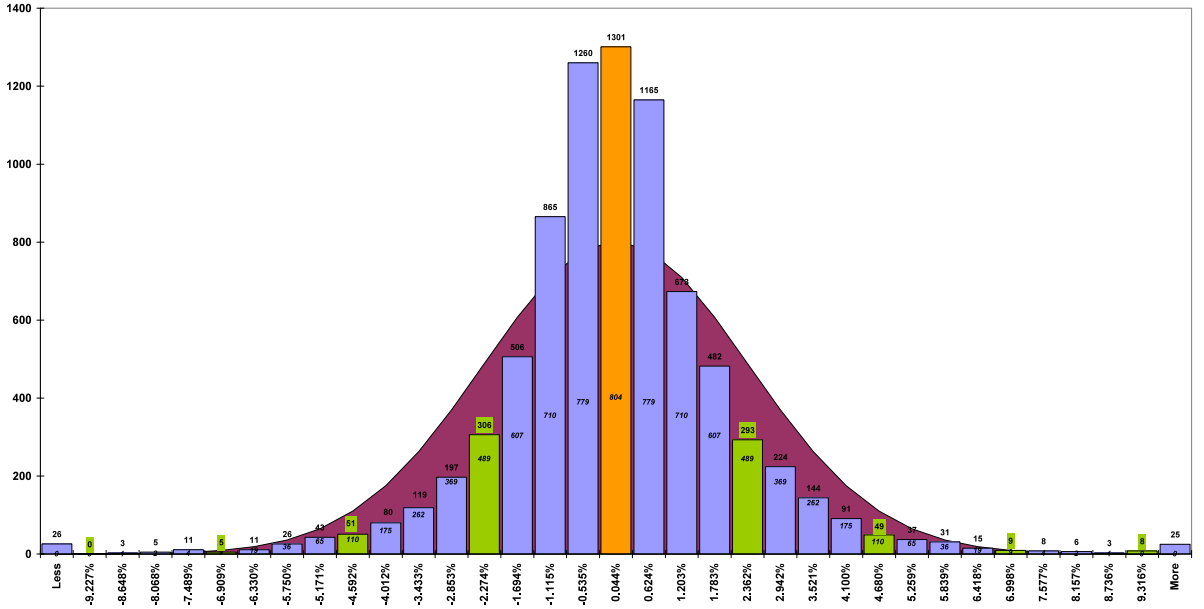



Fig. 7. The daily forecasting errors for Citigroup, 1977–2008.

Non-normality gets worse where individual stocks are concerned, as the recent experience with bank stocks has shown. For instance, the price of Citigroup dropped 34.7% between September 9 and 17, 2008, and then increased by 42.7% on the two days of September 18 and 19. These are huge fluctuations that are impossible to explain assuming independence and well behaved errors (the mean daily return of Citigroup is 0.044% and the standard deviation is 2.318%). Therefore, the uncertainty surrounding future returns of Citigroup cannot be also assessed either, as the distribution has long, fat tails (see Fig. 7), and its errors are both proportionally more concentrated in the middle, and have proportionally more extreme values in comparison to those of the DJIA shown in Fig. 6.

### *2.3. The accuracy and uncertainty when forecasting relationships*

There is no equivalent of the M-Competitions to provide us with information about the post-sample forecasting accuracy of relationships. Instead, econometricians use the $R^2$ value to determine the goodness of fit of how much better the average relationship is in comparison to the mean (used as a benchmark).

Estimating relationships, like patterns, requires "averaging" of the data to eliminate randomness. Fig. 8 shows the heights of 1078 fathers and sons,[3] as well as the average of such a relationship passing through the middle of the data.

The most likely prediction for the height of a son whose father's height is 180 cm is 178.59 cm, given that the average relationship is:

$$\begin{aligned}\text{Height Son} &= 86.07 + 0.514(\text{Height Father})\\ &= 178.59. \end{aligned} \quad (1)$$

Clearly, it is highly unlikely that the son's height will be exactly 178.59, the average postulated by the relationship, as the pairs of heights of fathers and sons fluctuate a great deal around the average shown in Fig. 8. The errors, or uncertainty, in the predictions depend upon the sizes of the errors and their distribution. These errors, shown in Fig. 9, fluctuate from about −22.5 to +22.8 cm, with the big majority being between −12.4 and +12.4. In addition, the distribution of forecast errors seems more like a normal curve, although there are more negative

[3] These are data introduced by Karl Pearson, a disciple of Sir Francis Galton.

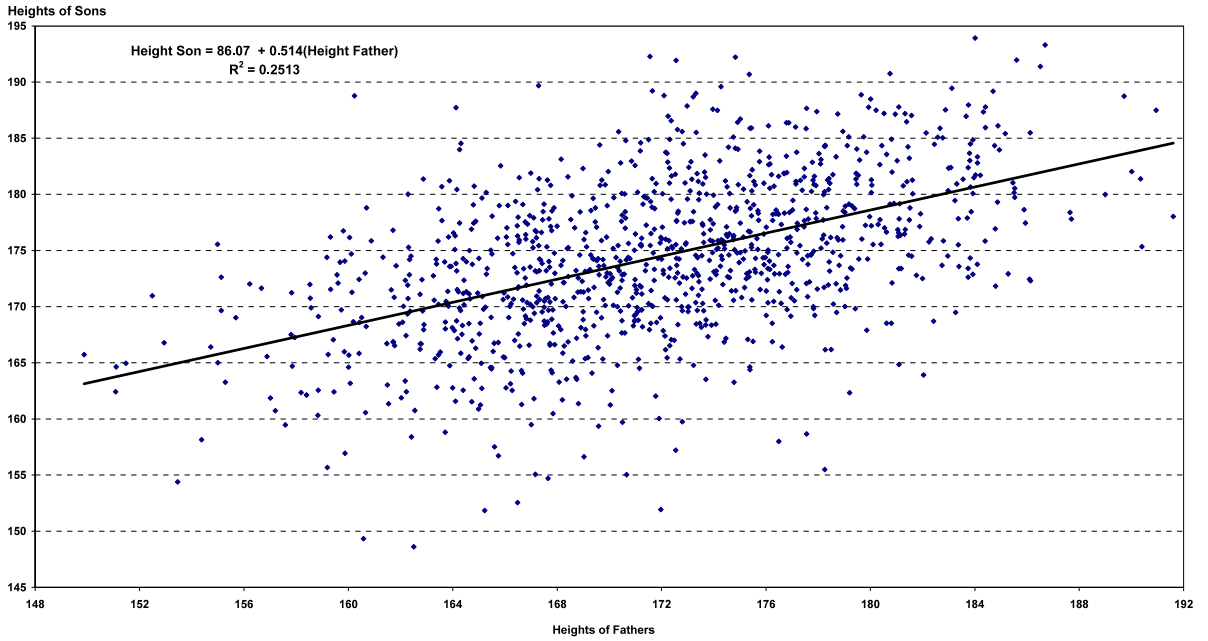



Fig. 8. Heights: Fathers and sons.

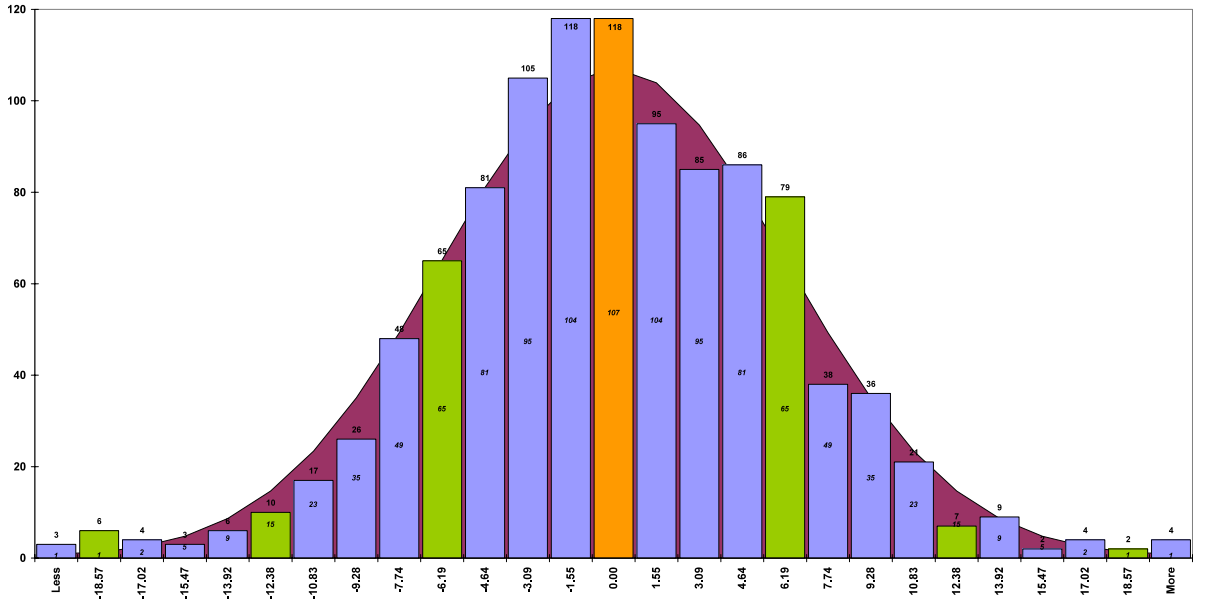



Fig. 9. The residual errors of the relationship height of fathers/sons.

errors close to the mean than postulated by the normal distribution, and more very small and very large ones. Given such differences, if we can assume that the distribution of errors is normal, we can then specify a 95% level of uncertainty as being:

$$\text{Height Son} = 86.07 + 0.514(\text{Height Father}) \pm 1.96(6.19) \quad (2)$$

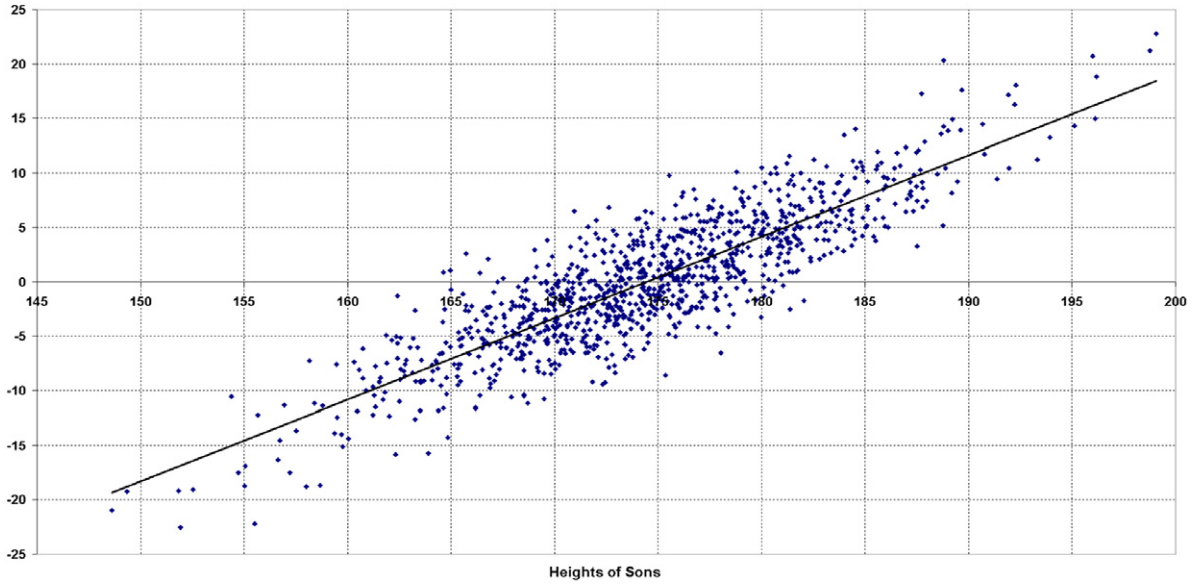



Fig. 10. Residual errors vs heights of sons.

(6.19 is the standard deviation of residuals).

Thus,

Height of Son $= 178.59 \pm 12.3$.

Even in this simple example, $\pm 12.3$ cm indicates a lot of uncertainty in the prediction, which also suffers from the fact that the distribution of errors is not entirely normal. In addition, there is another problem that seriously affects uncertainty. If the errors are plotted against the heights of the sons (Fig. 10), they show a strong correlation, implying that expression (1) underestimates short heights and overestimates tall ones. It is doubtful, therefore, that the forecast specified by expression (1) is the best available for the heights of sons, while the uncertainty shown in expression (2) cannot be estimated correctly, as the errors are highly correlated. Finally, there is an extra problem when forecasting using relationships: the values of the independent variables must, in the great majority of cases, be predicted (this is not the case with (1) as the height of the father is known), adding an extra level of uncertainty to the desired prediction.

Forecasts from econometric models used to be popular, giving rise to an industry with revenues in the hundreds of millions of dollars. Today, econometric models have somewhat fallen out of fashion, as empirical studies have showed that their predictions were less accurate than those of time series methods like Box–Jenkins. Today, they are only used by governmental agencies and international organizations for simulating policy issues and better understanding the consequences of these issues. Their predictive ability is not considered of value (see Orrell & McSharry, 2009), as their limitations have been accepted by even the econometricians themselves, who have concentrated their attention on developing more sophisticated models that can better fit the available data.

Taleb (2007) revisits the idea that such consequences need to be taken into account in decision making. He shows that forecasting has a purpose, and it is the purpose that may need to be modified when we are faced with large forecasting errors and huge levels of uncertainty that cannot be assessed reliably.

### 2.4. Judgmental forecasting and uncertainty

Empirical findings in the field of judgmental psychology have shown that human judgment is even less accurate at making predictions than simple statistical models. These findings go back to the fifties with the work of psychologist Meehl (1954), who reviewed some 20 studies in psychology and discovered that the "statistical" method of diagnosis was superior to the traditional "clinical" approach.

When Meehl published a small book about his research findings in 1954, it was greeted with outrage by clinical psychologists all over the world, who felt professionally diminished and dismissed his findings. Many subsequent studies, however, have confirmed Meehl's original findings. A meta-analysis by Grove, Zald, Lebow, Snitz, and Nelson (2000) summarized the results of 136 studies comparing clinical and statistical predictions across a wide range of environments. They concluded by stating:

"*We identified no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction. It holds in general medicine, in mental health, in personality, and in education and training settings. It holds for medically trained judges and for psychologists. It holds for inexperienced and seasoned judges*".

A large number of people can be wrong, and know that they can be wrong, brought about by the comfort of a system. They continue their activities "because other people do it". There have been no studies examining the notion of the diffusion of responsibility in such problems of group error.

As Goldstein and Gigerenzer (2009) and Wright and Goodwin (2009) point out, the biases and limitations of human judgment affect its ability to make sound decisions when optimism influences its forecasts. In addition, it seems that the forecasts of experts (Tetlock, 2005) are not more accurate than those of other knowledgeable people. Worse, Tetlock found out that experts are less likely to change their minds than non-experts, when new evidence appears disproving their beliefs.

The strongest evidence against the predictive value of human judgment comes from the field of investment, where a large number of empirical comparisons have proven, beyond the slightest doubt, that the returns of professional managers are not better than a random selection of stocks or bonds. As there are around 8500 investment funds in the USA, it is possible that a fund can beat, say, the S&P500, for 13 years in a row. Is this due to the ability of its managers or to chance? If we assume that the probability of beating the S&P 500 each year is 50%, then if there were 8192 funds, it would be possible for one of them to beat the S&P500 for 13 years in a row by pure chance. Thus, it is not obvious that the funds that outperform the market for many years in a row do so by the ability of their managers and rather than because they happen to be lucky. So far there is no empirical evidence that has conclusively proven that professional managers have consistently outperformed the broad market averages due to their own skills (and compensation). In addition to the field of investments, Makridakis, Hogarth, and Gaba (2009) have concluded that in the areas of medicine, as well as business, the predictive ability of doctors and business gurus is not better than simple benchmarks. These findings raise the question of the value of experts: why pay them to provide forecasts that are not better than chance, or than simple benchmarks like the average or the latest available value?

Another question is, how well can human judgment assess future uncertainty? Empirical evidence has shown that the ability of people to correctly assess uncertainty is even worse than that of accurately predicting future outcomes. Such evidence has shown that humans are overconfident of positive expectations, while ignoring or downgrading negative information. This means that when they are asked to specify confidence intervals, they make them too tight, while not considering threatening possibilities like the consequences of recessions, or those of the current subprime and credit crisis. This is a serious problem, as statistical methods also cannot predict recessions and major financial crises, creating a vacuum resulting in surprises and financial hardships for large numbers of people, as nobody has provided them with information to enable them to consider the full range of uncertainty associated with their investments or other decisions and actions.

## 3. A summary of the eight papers of this issue

This introductory paper by Makridakis and Taleb demonstrates the limited predictability and high level of uncertainty in practically all important areas of our lives, and the implications of this. It presents empirical evidence proving this limited predictability, as well as examples illustrating the major errors involved and the high levels of uncertainty that cannot be adequately assessed because the forecasting errors are not independent, normally distributed and constant. Finally, the paper emphasizes the need to be rational and realistic about our expectations from forecasting,

and avoid the common illusion that predictions can be accurate and that uncertainty can be assessed correctly.

The second paper, by Orrell and McSharry, states that complex systems cannot be reduced to simple mathematical laws and be modeled appropriately. The equations that attempt to represent them are only approximations to reality, and are often highly sensitive to external influences and small changes in parameterization. Most of the time they fit past data well, but are not good for predictions. Consequently, the paper offers suggestions for improving forecasting models by following what is done in systems biology, integrating information from disparate sources in order to achieve such improvements.

The third paper, by Taleb, provides evidence of the problems associated with econometric models, and proposes a methodology to deal with such problems by calibrating decisions, based on the nature of the forecast errors. Such a methodology classifies decision payoffs as simple or complex, and randomness as thin or fat tailed. Consequently, he concentrates on what he calls the fourth quadrant (complex payoffs and fat tail randomness), and proposes solutions to mitigate the effects of possibly inaccurate forecasts based on the nature of complex systems.

The fourth paper, by Goldstein and Gigerenzer, provides evidence that some of the fast and frugal heuristics that people use intuitively are able to make forecasts that are as good as or better than those of knowledge-intensive procedures. By using research on the adaptive toolbox and ecological rationality, they demonstrate the power of using intuitive heuristics for forecasting in various domains, including sports, business, and crime.

The fifth paper, by Ioannidis, provides a wealth of empirical evidence that while biomedical research is generating massive amounts of information about potential prognostic factors for health and disease, few prognostic factors have been robustly validated, and fewer still have made a convincing difference in health outcomes or in prolonging life expectancy. For most diseases and outcomes, a considerable component of the prognostic variance remains unknown, and may remain so in the foreseeable future. Ioannidis suggests that in order to improve medical predictions, a systematic approach to the design, conduct, reporting, replication, and clinical translation of prognostic research is needed. Finally, he suggests that we need to recognize that perfect individualized health forecasting is not a realistic target in the foreseeable future, and we have to live with a considerable degree of residual uncertainty.

The sixth paper, by Fink, Lipatov and Konitzer, examines the accuracy and reliability of the diagnoses made by general practitioners. They note that only 10% of the results of consultations in primary care can be assigned to a confirmed diagnosis, while 50% remain "symptoms", and 40% are classified as "named syndromes" ("picture of a disease"). In addition, they provide empirical evidence collected over the last fifty years showing that less than 20% of the most frequent diagnoses account for more than 80% of the results of consultations. Their results prove that primary care has a severe "black swan" element in the vast majority of consultations. Some critical cases involving "avoidable life-threatening dangerous developments" such as myocardial disturbance, brain bleeding and appendicitis may be masked by those often vague symptoms of health disorders ranked in the 20% of most frequent diagnoses. They conclude by proposing that (1) primary care should no longer be defined only by "low prevalence" properties, but also by its black-swan-incidence-problem; (2) at the level of everyday practice, diagnostic protocols are necessary to make diagnoses more reliable; and (3) at the level of epidemiology, a system of classifications is crucial for generating valid information by which predictions of risks can be improved.

The seventh paper, by Makridakis, Hogarth and Gaba, provides further empirical evidence that accurate forecasting in the economic and business world is usually not possible, due to the huge uncertainty, as practically all economic and business activities are subject to events which we are unable to predict. The fact that forecasts can be inaccurate creates a serious dilemma for decision and policy makers. On the one hand, accepting the limits of forecasting accuracy implies being unable to assess the correctness of decisions and the surrounding uncertainty. On the other hand, believing that accurate forecasts are possible means succumbing to the illusion of control and experiencing surprises, often with negative consequences. They suggest that the time has come for a new attitude towards dealing with the future that accepts our limited ability to make predictions in the economic and business environment, while also providing a framework

that allows decision and policy makers to face the future — despite the inherent limitations of forecasting and the huge uncertainty surrounding most future-oriented decisions.

The eighth paper, by Wright and Goodwin, looks at scenario planning as an aid to anticipation of the future under conditions of low predictability, and examines its success in mitigating issues to do with inappropriate framing, cognitive and motivational bias, and inappropriate attributions of causality. They consider the advantages and limitations of such planning and identify four potential principles for improvement: (1) challenging mental frames, (2) understanding human motivations, (3) augmenting scenario planning through adopting the approach of crisis management, and (4) assessing the flexibility, diversity, and insurability of strategic options in a structured option-against-scenario evaluation.

The ninth paper, by Green, Armstrong and Soon, proposes a no change, benchmark model for forecasting temperatures which they argue is the most appropriate one, as temperatures exhibit strong (cyclical) fluctuations and there is no obvious trend over the past 800,000 years that Antarctic temperature data from the ice-core record is available. These data also show that the temperature variations during the late 1900s were not unusual. Moreover, a comparison between the *ex ante* projections of the benchmark model and those made by the Intergovernmental Panel on Climate Change at 0.03 °C-per-year were practically indistinguishable from one another in the small sample of errors between 1992 through 2008. The authors argue that the accuracy of forecasts from the benchmark is such that even perfect prediction would be unlikely to help policymakers in getting forecasts that are substantively more accurate than those from a no change, benchmark model.

Because global warming is an emotional issue, the editors believe that whatever actions are taken to reverse environmental degradation cannot be justified on the accuracy of predictions of mathematical or statistical models. Instead, it must be accepted that accurate predictions are not possible and uncertainty cannot be reduced (a fact made obvious by the many and contradictory predictions concerning global warming), and whatever actions are taken to protect the environment must be justified based on other reasons than the accurate forecasting of future temperatures.

The tenth paper, by the late David Freedman, shows that model diagnostics have little power unless alternative hypotheses can be narrowly defined. For instance, independence of observations cannot be tested against general forms of dependence. This means that the basic assumptions in regression models cannot be inferred from the data. The same is true with the proportionality assumption, in proportional-hazards models, which is not testable. Specification error is a primary source of uncertainty in forecasting, and such uncertainty is difficult to resolve without external calibration, while model-based causal inference is even more problematic to test. These problems decrease the value of our models and increase the uncertainty of their predictions.

The final paper of this issue, written by the editors, is a summary of the major issues surrounding forecasting, and also puts forward a number of ideas aimed at a complex world where accurate predictions are not possible and where uncertainty reigns. However, once we accept the inaccuracy of forecasting, the critical question is, how can we plan, formulate strategies, invest our savings, manage our health, and in general make future-oriented decisions, accepting that there are no crystal balls? This is where the editors believe that much more effort and thinking is needed, and where they are advancing a number of proposals to avoid the negative consequences involved while also profiting from the low levels of predictability.

## 4. The problems facing forecasters

The forecasts of statistical models are "mechanical", unable to predict changes and turning points, and unable to make predictions for brand new situations, or when there are limited amounts of data. These tasks require intelligence, knowledge and an ability to learn which are possessed only by humans. Yet, as we saw, judgmental forecasts are less accurate than the brainless, mechanistic ones provided by statistical models. Forecasters find themselves between Carybdis and Scylla. On the one hand, they understand the limitations of the statistical models. On the other hand, their own judgment cannot be trusted. The biggest advantage of statistical predictions is their objectivity,

Table 3
Values of daily statistics for DJIA and Paris temperatures for each decade from 1900 to 2008.

| | Daily DJIA values | | | | | | | Daily Paris temperatures | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Decade | Mean | St. Dev | Min | Max | Skewness | Kurtosis | n | Mean | St. Dev | Min | Max | Skewness | Kurtosis | n |
| 1900 - 1910 | 0.019% | 1.03% | -8.29% | 6.69% | -0.36 | 4.75 | 2992 | -0.001 | 2.238 | -10.4 | 10.6 | 0.04 | 0.81 | 3650 |
| 1910 - 1920 | 0.018% | 0.98% | -7.24% | 5.47% | -0.50 | 4.74 | 2876 | 0.001 | 2.174 | -9.8 | 8.2 | -0.05 | 0.58 | 3650 |
| 1920 - 1930 | 0.034% | 1.11% | -12.82% | 12.34% | -0.94 | 21.41 | 2986 | 0.000 | 2.158 | -10.1 | 8.6 | 0.04 | 0.62 | 3650 |
| 1930 - 1940 | 0.000% | 1.85% | -8.40% | 15.34% | 0.63 | 6.35 | 2988 | -0.002 | 2.148 | -7.8 | 8.8 | 0.04 | 0.24 | 3650 |
| 1940 - 1950 | 0.013% | 0.74% | -6.80% | 4.73% | -1.16 | 10.84 | 2918 | 0.001 | 2.265 | -10.1 | 11.0 | 0.00 | 0.64 | 3650 |
| 1950 - 1960 | 0.049% | 0.66% | -6.54% | 4.13% | -0.84 | 6.76 | 2598 | 0.002 | 2.204 | -9.3 | 8.0 | -0.02 | 0.59 | 3650 |
| 1960 - 1970 | 0.009% | 0.65% | -5.71% | 4.69% | 0.02 | 5.45 | 2489 | -0.003 | 2.162 | -9.0 | 7.4 | -0.16 | 0.40 | 3650 |
| 1970 - 1980 | 0.006% | 0.93% | -3.50% | 5.08% | 0.33 | 1.89 | 2526 | 0.002 | 2.090 | -9.4 | 7.7 | -0.26 | 0.49 | 3650 |
| 1980 - 1990 | 0.054% | 1.13% | -22.61% | 10.15% | -3.08 | 68.81 | 2528 | -0.001 | 2.158 | -11.2 | 6.5 | -0.27 | 0.46 | 3650 |
| 1990 - 2000 | 0.061% | 0.89% | -7.18% | 4.98% | -0.31 | 4.88 | 2528 | 0.002 | 2.140 | -10.3 | 6.8 | -0.30 | 0.20 | 3650 |
| 2000 - 2008 | -0.004% | 1.30% | -7.87% | 11.08% | 0.26 | 9.12 | 2264 | 0.005 | 2.088 | -7.7 | 7.4 | -0.18 | 0.32 | 3163 |
| 1900 - 2008 | 0.023% | 1.08% | -22.61% | 15.34% | 0.18 | 18.89 | 29693 | 0.000 | 2.167 | -11.2 | 11.0 | -0.09 | 0.51 | 39663 |

which seems to be more important than the intelligence, knowledge and ability of humans to learn. The problem with humans is that they suffer from inconsistency, wishful thinking and all sorts of biases that diminish the accuracy of their predictions. The biggest challenge and only solution to the problem is for humans to find ways to exploit their intelligence, knowledge and ability to learn while avoiding their inconsistencies, wishful thinking and biases. We believe that much work can be done in this direction.

Below, we summarize the problem of limited predictability and high levels of uncertainty using the daily values of the DJIA and the Paris temperatures. The availability of fast computers and practically unlimited memory has allowed us to work with long series and study how well they can forecast and identify uncertainty. Table 3 shows various statistics for the daily % changes in the DJIA and the daily changes in Paris temperatures, for each decade from 1900 to 2008 (the 2000 to 2008 period does not cover the whole decade). Table 3 allows us to determine how well we can forecast and assess uncertainty for the decade 1910–1920, given the information for the decade 1900–1910, for the decade 1920–1930 given the information for 1910–1920, and so on.

### *4.1. The mean percentage change of the DJIA and the average change in Paris temperature*

The mean percentage change in the DJIA for the decade 1900–1910 is 0.019%. If such a change had been used as the forecast for the decade 1910–1920, the results would have been highly accurate. In addition, the volatility in the daily percentage changes from 1900–1910 would have been an excellent predictor for 1910–1920. The same is true with both the means and the standard deviations of the changes in daily temperatures, as they are very similar in the decades 1900–1910 and 1910–1920. Starting from the decade 1920–1930 onwards, however, both the means and the standard deviations of the percentage daily changes in the DJIA vary a great deal, from 0.001% in the 1930s to 0.059% in the 1990s (this means that $10,000 invested at the beginning of 1930 would have become $10,334 by the end of 1939, while the same amount invested at the beginning of 1990 would have grown to $44,307 by the end of 1999). The differences are equally large for the standard deviations, which range from 0.65% in the 1960s to 1.85% in the 1930s. On the other hand, the mean daily changes in temperatures are small, except possibly for the 2000–2008 period, when they increased to 0.005 of a degree. In addition, the standard deviations have remained pretty much constant throughout all eleven decades.

Table 3 conveys a clear message. Forecasting for some series, like the DJIA, cannot be accurate, as the assumption of constancy of their patterns, and possibly relationships, is violated. This means that predicting for the next decade, or any other forecasting horizon, cannot be based on historical information, as both the mean and the fluctuations around the mean vary too much from one decade to another. Does the increase to 0.005 in the changes in daily Paris temperature for the period of 2000–2008 indicate global warming? This is a question we will not attempt to answer, as it has been dealt with in the paper by Green et al. in this issue. However, the potential exists that even in series like temperature we have to worry about a possible change in the long term trend.

Another technique for looking at differences is departures from normality. Consider the kurtosis of the two variables. The 5 largest observations in the temperature represent 3.6% of the total kurtosis. For

the Dow Jones, the 5 largest observations represent 38% of the kurtosis (e.g., the kurtosis in the decade 1970–1980 is 1.89, while that of the following decade is an incredible 68.84—see Table 3). Furthermore, under aggregation (i.e., by taking longer observation intervals of 1 week, 1 fortnight, or 1 month), the kurtosis of the temperature drops, while that of the stock market does not change.

In real life, most series behave like the DJIA; in other words, humans can influence their patterns and affect the relationships involved by their actions and reactions. In such cases, forecasting is extremely difficult or even impossible, as it involves predicting human behavior, something which is practically impossible. However, even with series like the temperature human intervention is also possible, although there is no consensus in predicting its consequences.

### *4.2. The uncertainty in predicting changes in DJIA and Paris temperatures*

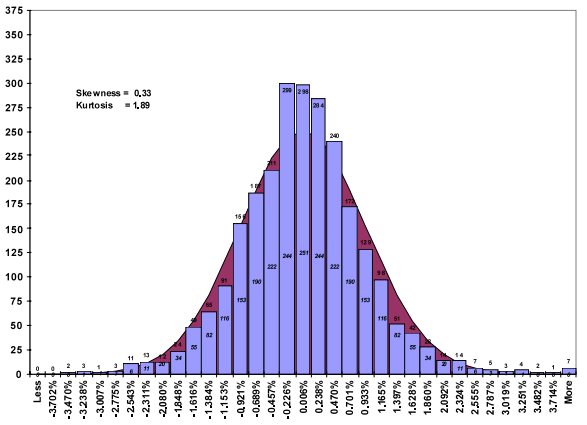
Having data since 1900 provides us with a unique opportunity to break it into sub-periods and obtain useful insights by examining their consistency (see Table 3), as we have already done for the mean, and we can now assess the uncertainty in these two series. The traditional approach to assessing uncertainty assumes normality and then constructs confidence intervals around the mean. Such an approach cannot work for the percentage changes in the DJIA for three reasons. First, the standard deviations are not constant; second, the means also change substantially from one decade to another (see Table 3); and finally, the distribution is not normal (see Fig. 6). Assessing the uncertainty in the changes in Paris temperatures does not suffer from the first or second problem, as the means and standard deviations are fairly constant. However, the distribution of changes is not quite normal (see Fig. 5), as there are a considerable number of extremely large and small changes, while there are more values around the mean than in a normal curve.

There is an additional problem when attempting to assess uncertainty. The distribution of changes also varies a great deal, as can be seen in Fig. 11. Worse, this is true not only in the DJIA data, but also in the temperature data. In the 1970s, for instance, the distribution of the DJIA percentage changes was close to normal with not too fat tails (the skewness and kurtosis of the distribution were 0.33 and 1.89 respectively), while that of the 1980s was too tall in the middle (the kurtosis was 68.84, versus 1.89 in the 1970s) with considerable fat tails on both ends. Given the substantial differences in the distributions of changes, or errors, is it possible to talk about assessing uncertainty in statistical models when (a) the distributions are not normal, even with series like temperatures; (b) the means and standard deviations change substantially; and (c) the distributions or errors are not constant? We believe that the answer is a strong no, which raises serious concerns about the realism of financial models that assume that uncertainty can be assessed assuming that errors are well behaved, with a zero mean, a constant variance, a stable distribution and independent errors.

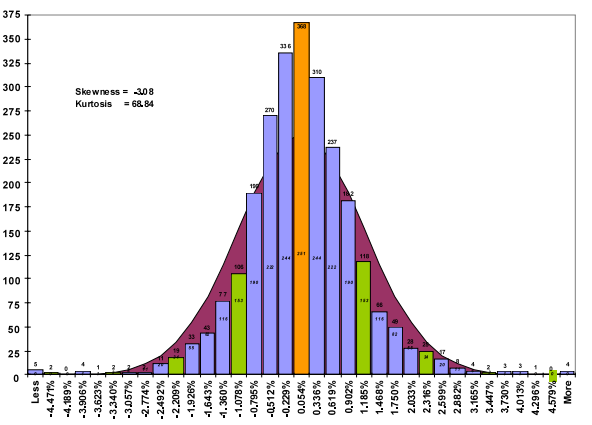
The big advantage of series like the DJIA and the Paris temperatures is the extremely large number of available data points that allows us to extract different types of information, such as that shown in Table 3, which is based on more than 2500 observations in the case of the DJIA, and 3650 for the temperatures. Real life series, however, seldom exceed a few hundred observations at most, making it impossible to construct distributions similar to those of Table 3. In such a case we are completely unable to verify the assumptions required to assure ourselves that there are not problems with the assessment of uncertainty. Finally, there is another even more important assumption, that of independence, that also fails to hold true, and negatively affects both the task of forecasting and that of assessing uncertainty. For instance, it is interesting to note that between September 15 and December 1, 2008, 52.7% of the daily fluctuations in the DJIA were greater than the mean $\pm 3$ (standard deviations). In the temperature changes there are fewer big concentrations of extreme values, but since 1977 we can observe that the great majority of such values are negative, again obliging us to question the independence of series like temperatures, which seem to be also influenced by non-random runs of higher and lower temperatures.

## 5. Conclusions

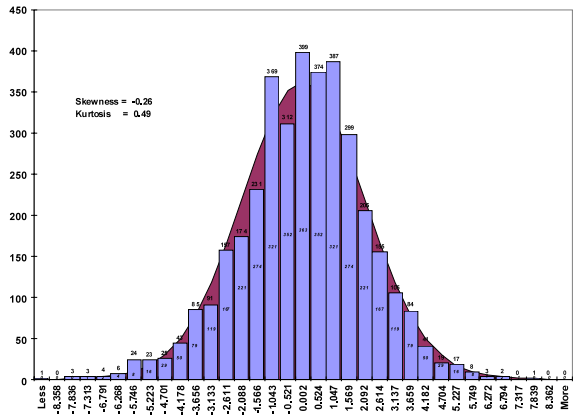
Forecasting the future is neither easy nor certain. At the same time, it may seem that we have no choice. But



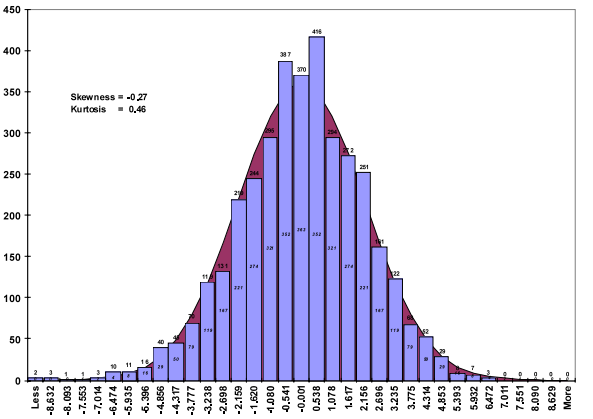

(a) The distribution of daily percentage changes in the DJIA in the 1970s.




(b) The distribution of daily percentage changes in the DJIA in the 1980s.




(c) The distribution of daily changes in the Paris temperatures in the 1970s.




(d) The distribution of daily changes in the Paris temperatures in the 1980s.

Fig. 11. The distribution of daily changes in the DJIA and Paris temperatures.

in reality we do have a choice: we can make decisions based on the potential sizes and consequences of forecasting errors, and we can also structure our lives to be robust to such errors. In a way, which is the motivation of this issue, we can make deep changes in the decision process affected by future predictions.

This paper has outlined the major theme of this special section of the *IJF*. Our ability to predict the future is limited, with the obvious consequence of high levels of uncertainty. It has proved such limited predictability using empirical evidence and four concrete data sets. Moreover, it has documented our inability to assess uncertainty correctly and reliably in real-life situations, and has discussed the major problems involved. Unfortunately, patterns and relationships are not constant, while in the great majority of cases: (a) errors are not well behaved, (b) their valiance is not constant, (c) the distribution of errors are not stable, and, worst of all, (d) the errors are not independent of each other.

## References

Goldstein, D., & Gigerenzer, G. (2009). Fast and frugal forecasting. *International Journal of Forecasting*, *25*(4), 760–772.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Makridakis, S., Hogarth, R., & Gaba, A. (2009). *Dance with chance: Making luck work for you*. Oxford: Oneworld.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., et al. (1982). The accuracy of extrapolative (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, *36*(4), 394–419.

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: The University of Minnesota Press.

Orrell, D., & McSharry, P. (2009). System economics: Overcoming the pitfalls of forecasting models via a multidisciplinary approach. *International Journal of Forecasting*, *25*(4), 734–743.

Taleb, N. (2007). *The black swan: The impact of the highly improbable*. Random House (US) and Penguin (UK).

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Wright, G., & Goodwin, P. (2009). Decision making and planning under low levels of predictability: Enhancing the scenario method. *International Journal of Forecasting*, *25*(4), 813–825.