EST.04

A Practical Application of Monte Carlo Simulation in Forecasting

Mr. James D. Whiteside II, PE

This paper describes a practical application of the Brownian-walk Monte Carlo simulation in forecasting. By setting up a simple spreadsheet and time-dependent historical data, this simple Monte Carlo routine is useful in forecasting productivity, installation rates, labor trends, etc. This is especially useful when historical data can not be modeled using standard data regression techniques. The paper discusses basic forecasting for single commodities, and it outlines a more robust methodology to create a composite forecast by combining several single commodities.

While Monte Carlo simulation is very popular to calculate the "P50 value" for contingency planning, the true power of Monte Carlo simulation is in data extrapolation, or forecasting beyond the known data points and understanding the range of outcomes expected. The paper addresses some basic elements of Monte Carlo simulation, approach for its application via a Brownian-walk approach in understanding cost trends and assessments for hard to model situations. The approach can help companies develop unique near term market insights and trends in order to compete in marketplace on analytics.

A PRACTICAL APPLICATION USING MONTE CARLO SIMULATION IN FORECASTING

This paper describes a practical application of the Brownian-walk Monte Carlo simulation in forecasting. By setting up a simple spreadsheet and time-dependent historical data, this simple Monte Carlo routine is useful in forecasting productivity, installation rates, labor trends, etc. This is especially useful when historical data can not be modeled using standard data regression techniques. The paper discusses basic forecasting for single commodities, and it outlines a more robust methodology to create a composite forecast by combining several single commodities.

While Monte Carlo simulation is very popular to calculate the "P50 value" for contingency planning, the true power of Monte Carlo simulation is in data extrapolation, or forecasting beyond the known data points and understanding the range of outcomes expected. The paper addresses some basic elements of Monte Carlo simulation, approach for its application via a Brownian-walk approach in understanding cost trends and assessments for hard to model situations. The approach can help companies develop unique near term market insights and trends in order to compete in marketplace on analytics.

Model Forecasts are not statements of what will happen, but what might happen given specific data, assumptions, and analytical methodologies used. Model forecasts are abstractions of market, regulatory activities, and producer and consumer behavior. Trends depicted in forecasts are indicative of tendencies in the real world rather than representations of specific real-world outcomes. Even where trends are stable and well understood, the projections are subject to high uncertainty and a Monte Carlo approach helps in understanding the range of outcomes anticipated.

EXTRAPOLATION/FORECASTING

Extrapolating or forecasting beyond or outside the known data realm has challenges, unless the data follows a normal trend or pattern. It is safe to use a trend line to forecast the next point past the last known data point because it is still close to the known data. However, there is a need to forecast and make a reasonable prediction far beyond the last data point in many business decisions. Predicting a point that is well beyond the last data point requires a good extrapolation routine. This numerically-based routine should be combined with other parameters to produce a range of probable outcomes that can be individually evaluated to assist with the decision-making process.

It is a basic tenet of forecasting that forecast data is always wrong because it is not "actual" data. A forecast is only correct if the base assumptions continue to exist in the future; however, most people treat the forecasts as nearly factual without updating the basis. Worse yet, most people evaluate a single point of data rather than a range of possible outcomes. Many published forecasts are different than the forecasts based purely on the data, the science, and the available mathematical models. Published forecasts generally can not capture changing policies, regulatory regime and unintended consequences in market dynamics. In reality, a forecast needs to capture such non-technical factors in order for the forecasts to be realistic in business planning context. Everyone should to take time to read the details (i.e. footnotes and qualifications) for a particular commodity because published forecasts have differing data sources and intended applications. Additionally, nobody wants to assume responsibility for forecasts because of the general misapplication or intended use of published forecasts. To be clear, this paper is focused on the science of data forecasting and not forecasts that include non-data adjustments.

Forecasts generally should guide and educate decision makers about how the market reacts when certain assumptions and data (independent variables) are considered. A reliable forecast tool should include the following four things:

- established forecasting principles;
- use of parameters to characterize the historical data;
- near-term market conditions; and
- data-driven methods.

THREE FORECASTING MODELS

There are three prominent forecasting models but only one meets the above criteria. The models are causal, judgmental, and time series. A causal model is based on the premise that the forecast is associated with the changes in other variables. Causality means that a strong correlation suggests that the event is not a completely random event. Multivariate regression (MR) is frequently used to test for causality. MR is best used as a test, much like a T test or Chi Test is used to determine if two data series are correlated. Causal modeling loosely associates events as being more than random, but it is not a parametric forecast model.

The second model is a judgmental methodology. This is when experience and intuition outweighs the lack of hard data. The collective reasoning power of people still out-performs the most sophisticated computer. Unlike a computer, this judgmental method is affected by tactical experiences (i.e., what has worked in the past), emotions, personalities, lack of sleep, and these factors often cause people to make poor decisions.

This paper addresses time series model in detail. Time series is based a direct correlation of data to time, with a forecast that is able to mimic the pattern of past behavior. This can be further refined through data decomposition, which is breaking data into multiple related components. The decomposition helps to determine whether there is a root cause that demonstrates a significant impact to the parent data. Unlike most time series forecast models using trend lines and data regressions, the Brownian-walk method is more advanced since it is able to sense, learn, and calculate order in seemingly random data.

MONTE CARLO SIMULATION

Probability is a way to bracket the volatility of short-term forecasts (seemingly random data). Monte Carlo simulation is a specialized probability application that is no more than an equation where the variables have been replaced with a random number generator. In other words, Monte Carlo is another computer approximation routine or numerical method that replaces geometry, calculus, etc.

A Monte Carlo simulation is a method for iteratively evaluating a deterministic model using sets of random numbers as inputs. This method is often used when the model is complex, nonlinear, or involves more than just a couple uncertain parameters.

The power of Monte Carlo simulation is two-fold: it is simple and fast. First, it appeals to a broad audience of who can describe a system using a set of basic math components and yet do not understand how all the individual elements, when working simultaneously, can provide solution points to develop market insights. Second, the increasing speed of computers make approximations faster than working out a complex math solution, which still may need sophisticated and time-consuming computer programming in order to run multiple case studies.

The downside of Monte Carlo is that it is more trusted than historical data. This misplaced trust is rooted in the idea that if a person has no historical data, then the Monte Carlo forecast can be anything that is believed plausible. For example, the phrase: “Contingency was determined by running Monte Carlo” is equivalent to “PI was determined by using Monte Carlo.” Monte Carlo *could* render reasonable results for each value. They both will not be precise answers because Monte Carlo is an approximation, and the results for both could be completely wrong.

Before venturing into applying Monte Carlo to calculate probabilities, program users should be able to validate a known value with some degree of precision. Otherwise, they will not be able to discern when the application is correct, when to apply it, or if the answer is realistic. For example, the value of PI can be approximated and quickly validated. However, nobody seems to question the validity of the contingency value if the Monte Carlo application is flawed.

Monte Carlo is used anyway, and the results are often misinterpreted or incomplete. The top abuse in Monte Carlo simulations is that few companies have data to support the distributions applied to the contingency elements.

BROWNIAN-WALK

A Monte Carlo application is a model that calculates the expected outcome of a system. Monte Carlo (uniform distribution) will be used to regulate the volatility range in a time-series equation called a Brownian-walk.

In experiments that Robert Brown [1] conducted in 1828, he noticed that pollen in water moved irregularly [1]. This irregular walk was named “Brownian” after the 19th century botanist. An example of the path of the pollen in water would look something like figure 1.

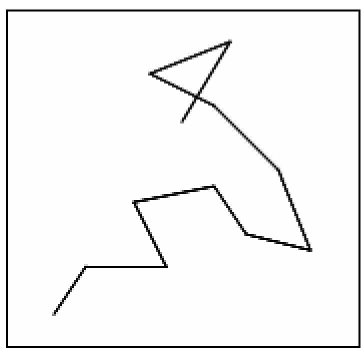

Figure 1—Brownian-walk Example

Although its path is irregular, the Brownian-walk can be calculated through an equation that incorporates the elements of Chaos Theory.

$$dS/S = \exp(\mu \times (dT) + std \times \epsilon \times (dT)^{1/2}) \qquad \text{Eq 1}$$

dS = change in the variable’s value from one step to the next
S = previous value
μ = the annualized growth or average increase between steps
dT = change in time from one step to the next
std = annualized volatility, or standard deviation
ϵ = value from a probability distribution (Monte Carlo)

The Brownian-walk is a stochastic process. A stochastic process helps us understand many outcomes in the future dependent upon probability distributions of inputs instead of dealing only with on possible ‘reality’ of how the process might evolve under time. This means that even if the initial condition (or starting point) is known, there are many possibilities the process might go to, but some paths are more probable. A stochastic process is like answering a doorbell. The pre-determined part is when the doorbell rings; you know (based on history) someone is at the door, so you go answer it. The undetermined part is the person at the door. Anyone (or anything) could be at the front door.

There are many articles about stochastic processes and most state that historical data is not necessarily required. The Brownian-walk equation only needs an annualized growth and annual volatility value. In the absence of historical data, it is a bad practice to arbitrarily adjust these two parameters until the forecast resembles the historical trend characteristics. The Brownian-walk can be used to produce stellar images by arbitrarily setting these parameters, but forecasting is not about the artwork. Forecasting is all about what has been learned from historical and current market conditions. Therefore, the implementation of the Brownian-walk should be constrained to real data.

There are two things that are important about the Brownian-walk. First, historical data is used to calculate the annualized growth and annual volatility values. Second, based on these values, a set of possible outcomes are generated until they represent a data regression with an acceptable "goodness of fit" value. That is to say, it would be best if 80 percent of the data was explained by the projected forecast and the root mean square error (RMSE) was low.

An inference engine differs from a stochastic process because there is a limited range of expected results based on the defined equation. Brownian-walk can also be considered an inference engine that is based on the nature of the process and the set parameters that mimic the process in order to forecast the most probable result. The range of calculated potential results should exhibit the same characteristics as the parent. For example, the doorbell rings when there is someone at the door, mail is usually delivered at 10 am, and the mailman normally brings packages to the door. If it is now 10:05 am and the doorbell rings, the inference engine (based on the parameters) would suggest that the mailman is at the door with a package.

The variables of the Brownian-walk equation that make it an inference engine are the parameters set for the standard deviation (std) and the annualized growth (μ). When they are set to values that are consistent with historical data, these "events" render results that are most probable. When the values are set arbitrarily in an attempt to mimic past behavior, it is simply "guessing". A person must understand how volatility and growth are related for the inference engine to forecast realistically. Otherwise, it is a stochastic process.

UNIFORM PROBABILITY DISTRIBUTION

The Brownian-walk equation needs a uniform probability distribution as shown in figure 2, "Uniform Probability Distribution." In this Monte Carlo application, it is important to have results with equal probability. It is also important to have a fairly wide bandwidth for the distribution. The most common mistake in using Monte Carlo is setting the most likely value and least likely value too close to the average value. After a few thousand iterations, the result is a data point with a narrow bandwidth rather than a range of possible outcomes around an average value. The Central Limit Theorem (CLT) states that the sampling distribution will have the same average as the data population, which is generally much narrower than the actual data population, and it narrows the bandwidth in the results. It is a better practice to set the upper and lower values of possibilities just outside of the highest and lowest historical data values. The best case scenario would be for the Monte Carlo result to produce most likely and least likely values that match the historical data.

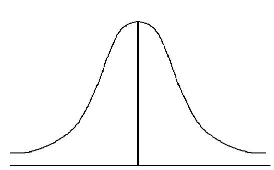

Figure 2—Uniform Probability Distribution

This is the uniform probability distribution equation:

$$\epsilon = \left(\sum_{i=1}^{n=12} \text{PRNG} \right) - 6$$

PRNG is the value returned from the portable random number generator (PRNG). Since most built-in random number generators have flaws, only validated PRNGs should be used in Monte Carlo simulations.

UNOBVIOUS ORDER

The most common argument used to argue against forecasting of near-term data is that it is highly volatile and highly random. Since it is "highly unpredictable," critics then argue forecasting is unachievable and revert to judgmental methods (probably the ultimate answer is somewhere in the middle). However, many people have never seen an application based on Chaos Theory. The reason that the Brownian-walk equation could be considered bordering chaos is that it has the same basic elements as in the following example of a Lorenz Attractor. A Lorenz Attractor is based on three differential equations. Figure 3 represents only two of the three variables required, but this paper is not going to address this further.

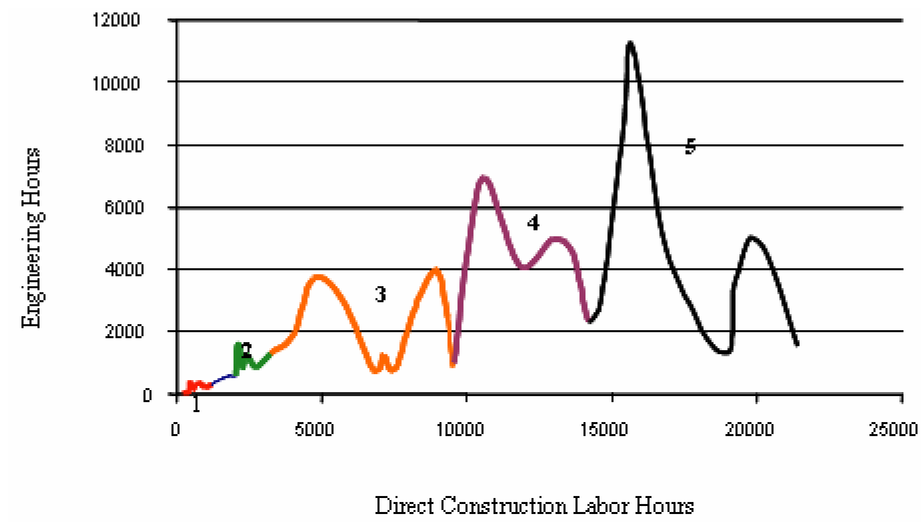

Figure 3—Engineering Hours vs Direct Construction Labor Hours

Figure 3—"Engineering Hours vs Direct Construction Labor Hours," appears to be random and without order. However, when re-plotting the data in figure 3 as shown in figure 4, "Lorenz Attractor" it appears ordered and not random at all. It serves as an important example of the ordered nature of some common "highly unpredictable" data sets. The point is that too many forecasts are ignored because the data appears too random to model and any forecasts are therefore assumed to be flawed and unusable.

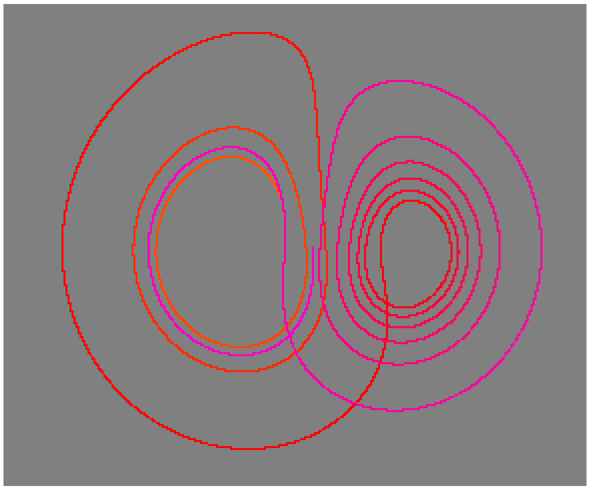

Figure 4—Lorenz Attractor
(Re-plotting of Figure 3)

FORECASTING RAW MODE

The Brownian-walk provides a slate of possible path outcomes as shown in figure 5, "Units vs Time (Raw Mode)." This will be referred to as a "raw mode" because there is no attempt to correct the forecasts. The raw mode is a pure Brownian-walk output. The data represented by the curves in figure 5 are easier to forecast than the Lorenz's curves in figure 3 because they do not have the magnitude of peaks and valleys.

The pathways in figure 5 resemble the curves in figure 3. Units can be a price, productivity value, rate of change, etc. Time can be weeks, months, years, etc. as long as it is a value that the unit data is regularly harvested for comparative purposes. For example, if stock price is analyzed for long-term forecasting, time could be expressed in months. If stock price is analyzed for the affects that day traders have on this particular stock, then using hours may be more appropriate.

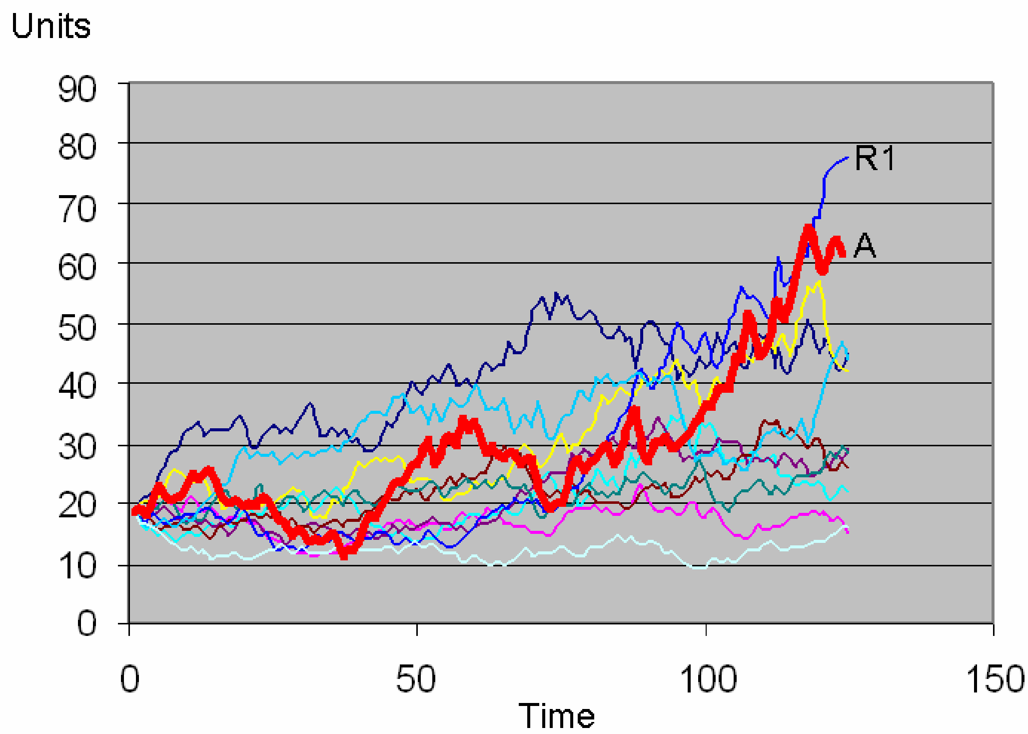

Figure 5—Units vs Time (Raw Mode)

Line “A” in figure 5 and figure 6, ”Units vs Time (Regression Mode),” is a representation of the actual data that needs to be forecasted. In both graphs, all the forecasts start at time 0 and end at time 125. The annual growth and annualized volatility parameters of Line “A” are used to generate the other forecast lines. While the other forecasts seem to mimic the characteristics of Line “A,” their margin of error is large. Nine of the lines each explain less than 10 percent of the data represented by Line “A.” Line “R1” explains better than 80 percent of the data in Line “A.” This line can be considered a regression by Monte Carlo.

FORECASTING REGRESSION MODE

A regression that explains more than 80 percent of the data will be an improvement over the raw mode. Normally, a regression is performed by a least-squares method to produce an equation (regression line). The equation represents a set of predicted data points with the least amount of error compared to the actual data points. In this case, this Monte Carlo approach replaces the least-squares method and will produce a “regression data set” rather than a regression line.

Therefore, forecast lines are generated until the margin of error falls inside a pre-determined limit set by the Root Mean Square Error (RMSE) of Line “R1” on Graph 1.

The Root Mean Square Error (RMSE) equation is:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\mathrm{Error}^2)}{n}} \qquad \text{Equation 3}$$

Error is the difference between the actual value and the predicted value. RMSE is the average of the forecast errors.

Once the margin of error is within acceptable limits this produces a data regression set between time 0 and time 125. Then a forecast line is generated from time 125 to time 150. The process of producing regression data sets is repeated until there are sufficient forecasts to be evaluated. In graph 2, seven regression-forecast lines were removed for clarity because there was significant overlap. The other lines were somewhat evenly spread between Line "F1" and Line "F3".

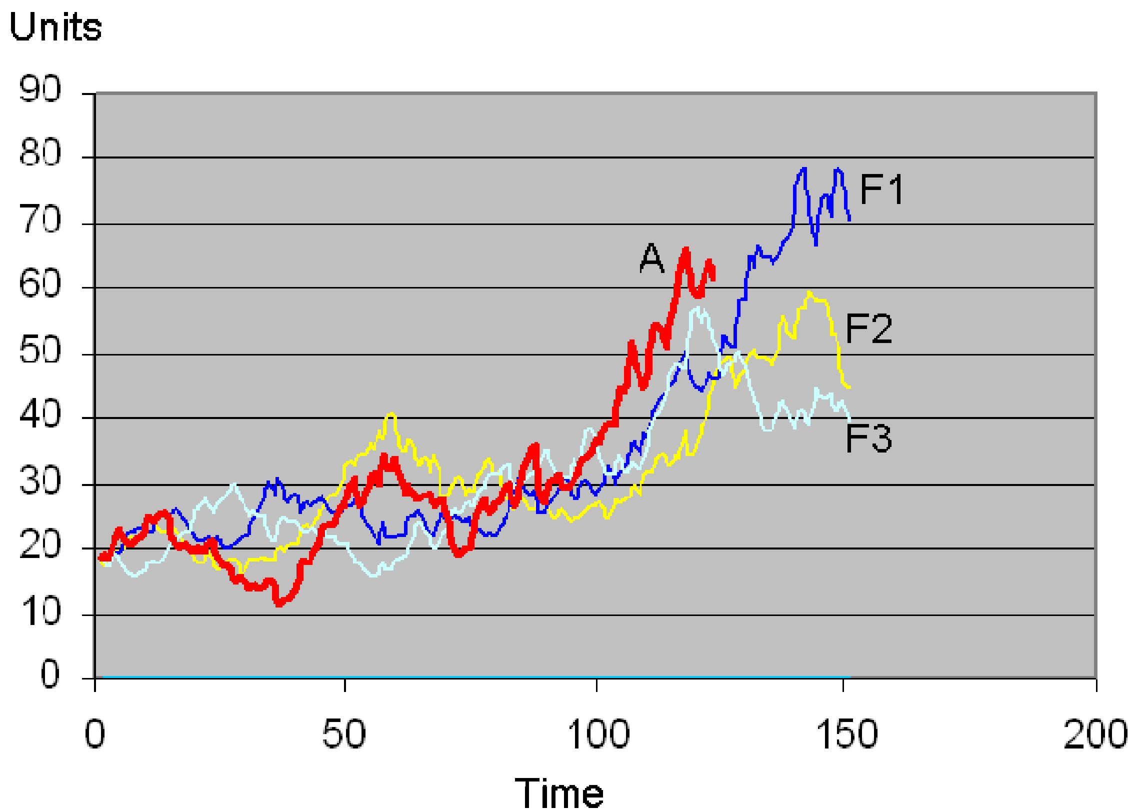

Figure 6—Units vs Time (Regression Mode)

INTERPRETING THE RESULTS

Decision makers get inputs from statisticians, macro economists, etc. and make decisions. The question is why decision makers do not consider inputs from statisticians etc., and we need to think about bringing market savvy, market drivers, etc., in our modeling so that decision makers value our overall assessments and consider them in their plans.

The idea behind using a forecasting tool like a Brownian-walk is to consider the number of possible pathways within the boundaries set by statistics and Monte Carlo probability results. All the predictions are equally probable, and these predictions provide planners with a range of considerations that may not be obvious at first glance. The best practice is to make a long range forecast and update it regularly with new data.

Interpretation 1: Simple Probability

Without consideration to the context of the data source, there are some basic mathematical conclusions that can be made about the forecasts depicted by lines "F1", "F2", and "F3". Line "F1" suggests that the units will continue to rise. Line "F2" suggests that the units will continue to rise until time 145 and then drop off. Given that "time now" is at 125, in order for the forecast Line "F3" to be correct, the units will start dropping precipitously in the next few time periods. It can be said that each of the forecasts ("F1", "F2", "F3") have a 30 percent chance of being correct. Since Line "F1" and Line "F2" each have a 30 percent chance of occurring, then there is a 2/3 chance that units will continue to increase between time 125 and time 150.

Interpretation 2: Weighted Data

Another interpretation can be made on a weighted basis. The weight could be the product of the percent data explained and the probability of occurrence. Line "F1" has a stronger likelihood followed by Line "F3" as shown by the ordinal number in table 1, "Weighted Line Values". Line "F1" is 35 percent (27/20) stronger than Line "F2", and it is 17 percent (27/23) stronger than Line "F3." Given no other data, the Line "F1" forecast seems to be the most likely.

Line	Percent Data Explained	Probability	Weight	Ordinal
F1	0.90	30.00	27.00	1
F2	0.67	30.00	20.10	3
F3	0.77	30.00	23.10	2

Table 1: Weighted Line Values

Interpretation 3: Simple Stats

Looking at time 150 there is a 2/3 chance that the units will remain between 40 and 50. There is only a 1/3 chance that the Units will remain above 60. Line "F2" and Line "F3" suggest that the units will flatten out or decline between time 125 and time 150.

RECONCILIATION OF ALTERNATIVE INTERPRETATIONS

Deciding which forecast is more appropriate rests within the context of the data. From a strictly data point of view, planning should be weighted heavier on scenarios addressing the forecast of increasing of units. However, a forecast must be considered in the context of the data. Just because a forecast does not immediately support a preconceived outcome, some time should be spent to determine what would have to occur to make the unpopular forecast become "true."

Following are three context examples. If the data represents a cyclical or "fad" commodity like toys and sales are strong, then forecast "F1" should be the near-term forecast. If sales have been sluggish for two time periods, then forecast "F2" should be considered and action plans made to correct sales or switch to another product line. "F3" suggests that a significant recession is on the brink of being triggered, signaling that there is no more market for this particular toy.

APPLICATION OF THE BROWNIAN-WALK APPROACH

A composite forecast can be produced by combining single commodities, which is represented by the near-term forecast in figure 7, "Composite Forecast." In this example, an asset is broken down into three categories: equipment, material, and labor. The asset's distribution of these categories is weighted 20 percent for equipment, 30 percent for material, and 50 percent for labor. The category weights, however, will be different depending on the type of asset. The line for the Material forecast is a composite of a similar forecast of its categories: carbon steel piping, alloy piping, instrumentation, electrical, steel, and civil materials (needless to say, individually each of the categories can have dramatically different price paths).

For illustration purposes only, consider the time axis to be in months for figure 7, "Composite Forecast." Today's month is represented by 90, ten months from now is represented by 100, and 10 months ago is represented by 80. The y-axis represents the percent increase per month. In this example, all commodities have been increasing significantly, and projects in the month 90 are now 104 percent (see table 2, "compounded interest": month 90) more expensive than in month 80.

Month	Annual Percentage	Annual Percentage (Decimal)	Monthly Percentage		Monthly Componded Percentage
		B/100	1+C/12		Prev Value X Current Value
80	2.43	0.0243	1.00203		1.0020
81	2.69	0.0269	1.00224		1.0043
82	3.05	0.0305	1.00254		1.0068
83	2.76	0.0276	1.00230		1.0091
84	3.27	0.0327	1.00272		1.0119
85	3.15	0.0315	1.00263		1.0145
86	4.93	0.0493	1.00411		1.0187
87	4.91	0.0491	1.00409		1.0229
88	6.03	0.0603	1.00502		1.0280
89	7.33	0.0733	1.00611		1.0343
90	9.68	0.0968	1.00806		1.0426
91	9.83	0.0983	1.00819		1.0512
92	9.41	0.0941	1.00784		1.0594
93	6.38	0.0638	1.00531		1.0650
94	8.07	0.0807	1.00673		1.0722
95	10.25	0.1025	1.00854		1.0814
96	9.01	0.0901	1.00751		1.0895
97	5.00	0.0500	1.00417		1.0940
98	4.36	0.0436	1.00363		1.0980
99	3.59	0.0359	1.00299		1.1013
100	4.11	0.0411	1.00343		1.1051

Table 2—Compounded Interest

The question is how much more they will be in 5 months and 10 months. This model forecasts labor as reaching its peak increase today (time 90) and labor will continue to decrease over the next 10 months. Equipment and material will continue to increase for the next eight months. The material forecast exceeds equipment because it is mostly alloy-based material which has seen very high cost escalation in recent months.

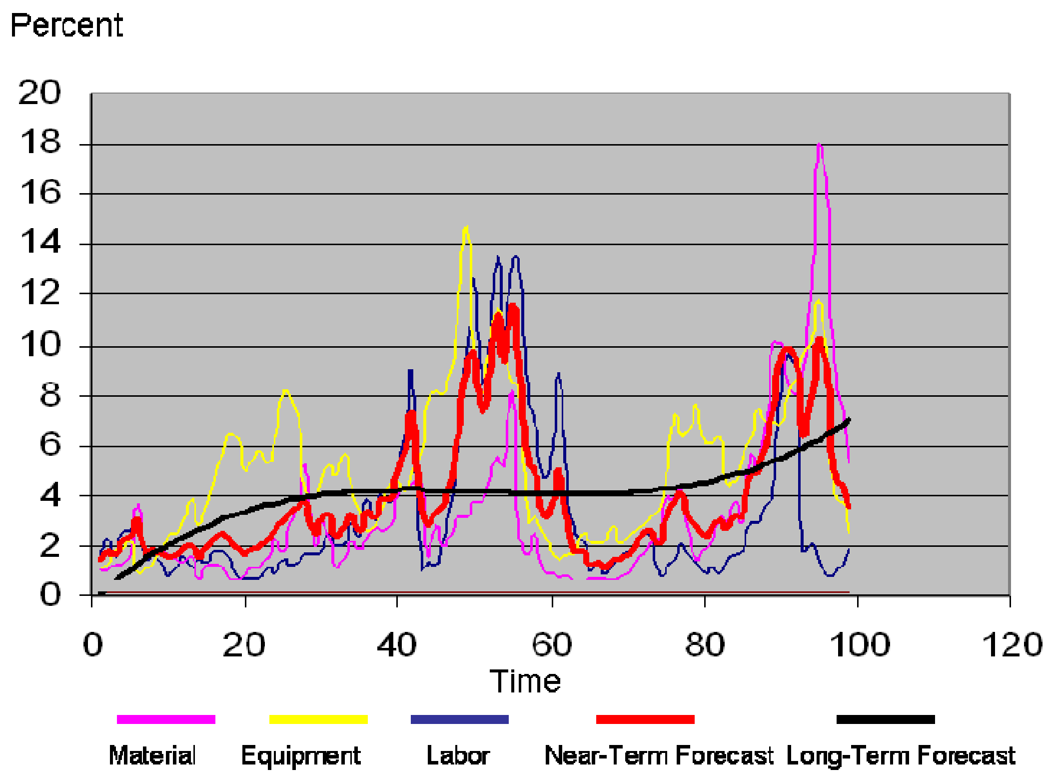

Figure 7—Composite Forecast

There are three basic applications for the forecasts represented in figure 7. The long-term forecast line is conservative and best suited for long range planning. The long-term forecast (black curve) is a simple trend line of the near-term forecast. The near-term forecast line (red curve) in figure 7 is the sum of the weighted values for material, equipment, and labor. The near-term forecast line can be best used for projects which will probably complete in the next five to ten months (possibly 1-2 years even). The data in the commodity forecasts can be placed into project risk economic models.

A CASE FOR MICROECONOMICS

Brownian-walk is a great tool for microeconomic forecasting. Microeconomics focuses on factors that affect the decisions made by individuals and businesses. In this case, the influences of the near-term market pressures on the pricing commodities. These factors are represented by the commodities in figure 7 and the near-term forecast line is the aggregation.

Macro economists study the movement and trends in the economy as a whole. They are generally in charge of setting the escalation long-range forecast. When macro economists evaluate a forecast, they project an aggregated long-term forecast as in figure 7.

It appears that their general theory is that there will be market corrections forcing the overall escalation to be less than half the value of the current market growth value. Therefore, they tend to default to an inflation value around 2.5 percent or 3 percent per year beyond year three because these values are useful in long-term forecasting and business planning. The disruption of large pricing swings of a few smaller commodities muted by the overall economy.

Macroeconomic forecasting does not fit near-term forecasting because projects have momentum. A project will not automatically get refunds because the market was in decline after the project was started. The market may be completely out of sync with the local or national economy.

Market pricing is a function of supply and demand. It is important to use macro economics to plan long-range business development. However, it is more important to use micro economics to manage near-term forecasting, and that is the value of this particular forecasting tool. This Brownian-walk forecasting tool brings order to seemingly random data and produces reliable near-term forecasting of data with volatile characteristics.

RECOMMENDED READING

1. Scholastic 2006 Encyclopedia Americana [US].

Mr. James D. Whiteside II, PE
ConocoPhillips
jim.d.whiteside@conocophillips.com