Data Vintages and Measuring Forecast Model Performance

JOHN C. ROBERTSON AND ELLIS W. TALLMAN

Robertson is a visiting scholar and Tallman is a senior economist in the macropolicy section of the Atlanta Fed's research department.

They thank Lucy Ackert and Mary Rosenbaum for comments, Robert Parker and Bruce Grimm of the BEA for the NIPA time series data and comments and insights into the bureau's data revision process, Glenn Rudebusch and James Hamilton for the CLI time series data used in their studies, and Amanda Nichols for research assistance.

THE DATA ON MOST ECONOMIC VARIABLES ARE ESTIMATES. THESE ESTIMATES ARE REVISED, SOMETIMES FREQUENTLY, AND OFTEN THEY CONTINUE TO BE REVISED MANY YEARS AFTER THE FIRST ESTIMATE APPEARS. FOR EXAMPLE, ON JULY 31, 1998, THE BUREAU OF ECONOMIC ANALYSIS (BEA) OF THE U.S. DEPARTMENT OF COMMERCE ANNOUNCED THAT THE SEASONALLY ADJUSTED ESTIMATE OF REAL GROSS DOMESTIC PRODUCT (GDP) GROWTH FOR THE SECOND QUARTER OF 1998 WAS AN ANNUALIZED 1.4 PERCENT. IN ADDITION, THE JULY PRESS RELEASE CONTAINED REVISED ESTIMATES FOR THE REAL GDP SERIES (AND COMPONENTS) FROM THE FIRST QUARTER OF 1995 UNTIL THE FIRST QUARTER OF 1998. THE REVISION SHOWED AN INCREASE IN THE ESTIMATED AVERAGE YEAR-OVER-YEAR REAL GDP GROWTH FROM 2.9 PERCENT TO 3.3 PERCENT FOR THE PERIOD FROM 1995 TO 1997.

Chart 1 presents the year-over-year GDP growth estimates for the period from 1995 to 1998 as of June 1998 (referred to as the June 1998 data vintage) together with the corresponding estimates as reported in July 1998 (the July 1998 data vintage).[1] As is apparent from the chart, the growth estimates from the older vintage are systematically lower than those from the more recent vintage. In mid-1996, for example, the year-over-year growth rate was nearly 1 percentage point lower in the June vintage of data than in the revised July vintage.

The timing of revisions to data usually follows a regular schedule, even if the size or the direction of the revisions do not. For example, the BEA usually publishes revisions of the National Income and Product Accounts (NIPA) for the three prior years each July. Moreover, by the time this article appears in print, the 1.4 percent growth estimate for the second quarter of 1998 will have been revised twice—in August and again in September.

Other estimates of economic activity also change over time as new vintages are constructed. For example, the historical data on the seasonally adjusted index of total industrial production (IP) were revised in January 1997 and then again in December 1997. Chart 2 plots the year-over-year percentage change in each of these two data vintages over the period from 1994 until the end of 1996. The difference between the vintages appears sizable; for example, growth during mid-1995 was more than 2 percent lower using the January 1997 version rather than the December 1997 revision.

In a policy context the distinction between vintages of data can be important. For example, Orphanides (1997) shows that a rule-based monetary policy performs dramatically worse when real-time data are used instead of subsequently revised versions of the data. The result—that revised data help make better policy—is an interesting finding; the more relevant issue, though, is that good rule-based policy performance requires the use of data unavailable to the policymaker in real time.

This article finds that the choice of data vintage can be important when comparing the performance of competing forecasting models of real output. Specifically, the research considers a choice between competing forecast models that is based on relative out-of-sample forecast performance. The study requires (1) using data available at the time the forecast would have been made to construct the forecast and (2) using data available not too long after the period being forecast to evaluate the model's performance. For the IP measure of output this approach leads to a quite different conclusion about relative model performance from that derived by using the latest available or most recent vintage of data throughout the analysis.

This result emphasizes the important distinction between an actual real-time forecast analysis and a pseudo real-time forecast analysis. In a pseudo real-time analysis a forecaster uses only the latest available vintage of historical data series in constructing and evaluating the forecasts. In contrast, an actual real-time analysis requires using the vintage of data actually available at the forecast date, together with forecast errors constructed using a vintage of data available soon after the period being forecast. To the extent that future data revisions will be similar to past ones, the results from simulating the past real-time performance of competing models should provide a better guide to a model's subsequent performance than would the results of simulations using only the current vintage of data.

The following section of the article discusses in more detail the distinction between simulating actual real-time forecasts and pseudo real-time forecasts. The position argued is that most results reported in the academic forecasting literature are from pseudo real-time forecast experiments. Of the few studies that have attempted to introduce a real-time aspect into the analysis, most have tended to use the notion either too loosely or too tightly to reflect accurately what a forecaster would have been able to do in real time. The discussion then presents the empirical results of the model comparison exercise using real-time data and contrasts these with the results of using only the most recent data vintage.

Using only the latest vintage of historical data may influence the measured forecast performance in misleading ways.

Real-Time Forecasting

The standard forecast estimation and evaluation strategy is to estimate or fit a model over some period, construct an out-of-sample forecast, and compare this forecast with the actual outcome. Then the forecaster makes a decision, based on the relative size of the resulting forecast errors, about the quality of the model's previous forecast performance. The forecaster hopes that a model that has performed well relative to previous alternatives will continue to do so in the future.

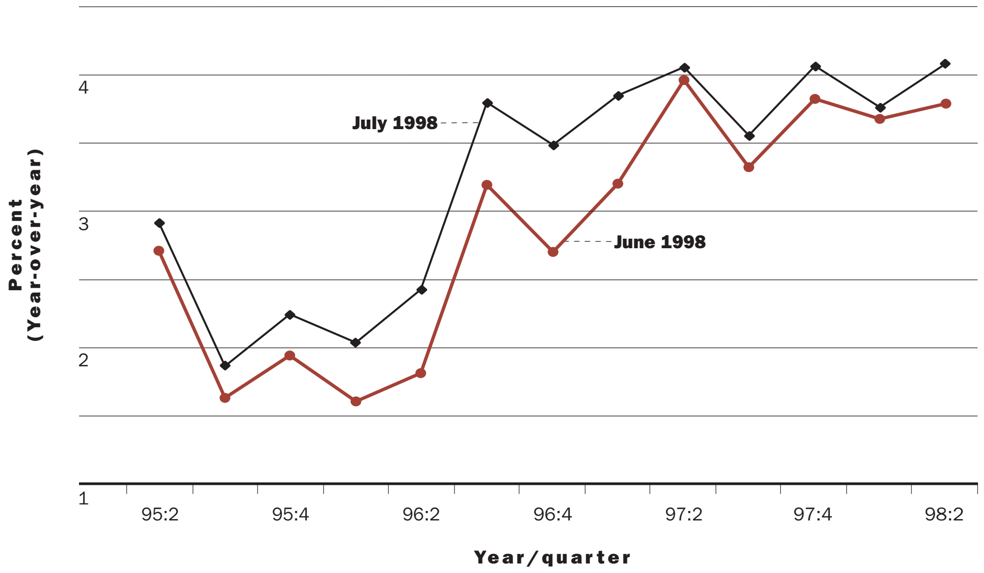
As described in most econometric textbooks as well as in the academic literature, forecast evaluations of a model typically employ the most recent vintage of the relevant time series at each stage of the process. It is possible, however, that using only the latest vintage of historical data may influence the measured forecast performance in misleading ways, and the result may not be a good approximation of forecasting accuracy in real time.

Two potential problems arise when forecast evaluations employ the latest vintage of historical data for both estimating and evaluating. First, in a realistic forecasting situation, one can use only the vintage of historical data available at the time the forecast is made. That even more refined measurements will become available is of little relevance.[2] Thus, forecasts with revised data are not realistic, real-time forecasts.

[1] A data series vintage or "age" is denoted by the month in which the entire data series existed—when that specific set of numbers was available as data.

[2] From today's perspective, it could be argued that the latest available vintage provides the most accurate historical record of series such as gross domestic product or industrial production. But an even more accurate record will likely be available in the future after further revisions have taken place. Consequently, the notion of an "ultimately revised" or "true" history for estimates is somewhat nebulous.

CHART 1 GDP Growth as of June and July 1998



Source: Data from Bureau of Economic Analysis, Department of Commerce

The second problem that arises from using the latest vintages of data centers on forecast evaluation. A forecaster typically wants to evaluate the model's forecast performance against an outcome that is measured not too long after the month or quarter being forecast. It is unlikely that forecasters or their clients would be prepared to wait for a more revised historical record.

In an important empirical study Fair and Shiller (1990) describe in detail the necessary conditions that a historical analysis of real-time out-of-sample or ex ante forecasts must satisfy. To be specific, suppose that the goal is to evaluate the accuracy of a particular forecasting model of GDP; the forecast is made using data available in some period, and forecast values are generated for subsequent periods. For these out-of-sample forecasts to be constructed in real time,

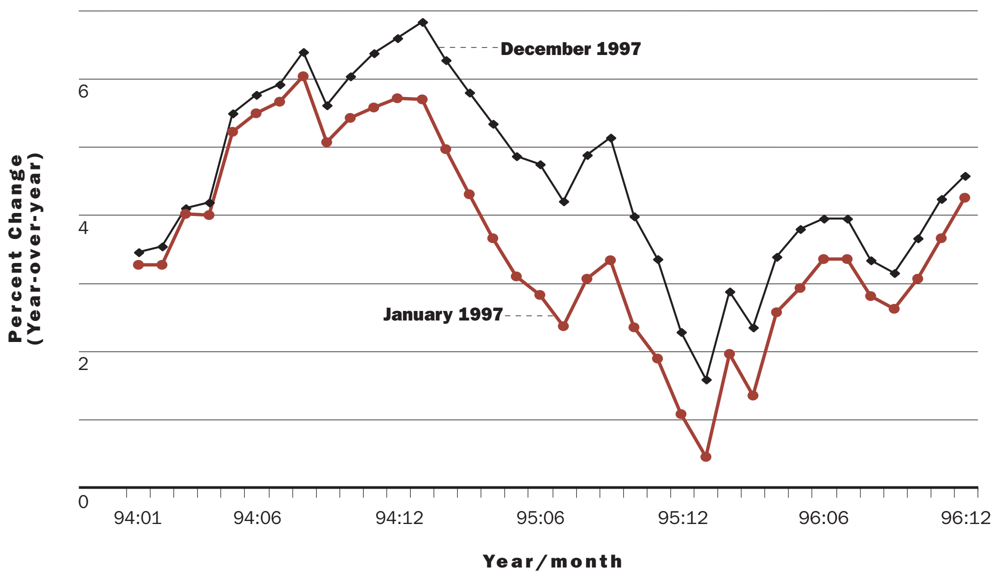
(1) *Future values of variables in the model must be only forecasts.* These forecasts are used in constructing the forecasts of the particular variables of interest. For example, suppose one is interested in forecasting real GDP growth and the federal funds rate (FFR) is believed to affect future real output growth. On any particular date on which a forecast is made, any future values of the FFR used in forming real GDP growth forecasts will themselves be forecasts. To allow actual values of the FFR into the forecasting model is to give the model an unfair advantage. Of course, if the future path for the FFR were known in advance, then it would make sense to use these values when constructing the forecast.

(2) *The coefficients of the model must be estimated only over the sample period up to the time the forecasts are being formed.* For example, suppose there are data from 1959 up until 1998. Estimating the coefficients of the model using all the data through 1998 and then forecasting from 1988 on would be giving the forecast model information from future data observations contained in coefficient estimates that were not actually available in 1988.

(3) *Only data for the period prior to the time the forecast is made can be used in determining the model specification.* Following from the previous example, suppose that the model specification (say, the number of lagged observations to use in the model) is chosen by a criterion that used all the data in the sample through 1998 and the chosen specification is then fitted and forecast from 1988 on. Again, the chosen model's forecasts would have been partially based on information from future data observations. Instead, the model specification should be chosen only on the basis of analysis of observations available through 1988.

(4) *The vintage of the data used to estimate the model and construct the forecasts must be actually available at the time the forecast is made.* This restriction is the focus of this article. Here, the forecasting model is limited to using only the data vintage available at the time the forecast would have been constructed, preventing future information in the form of data revisions from entering into the forecasts. Thus, for example, a forecast formed in

CHART 2 Industrial Production Index as of January and December 1997



Source: Board of Governors of the Federal Reserve System

July 1988 uses only the vintage of data actually available in July 1988.

It is unlikely that any published out-of-sample forecast model evaluation would have failed to satisfy the first two of Fair and Shiller's requirements. Yet it is surprising that a number of studies have ignored the third requirement. In these cases, researchers used the full sample, including the period to be forecast, when determining the model specification (that is, the variables included, the lag length, and so on). Notably, failure to satisfy the fourth requirement is almost universal in the literature. In some studies, like those of Staiger, Stock, and Watson (1997) and Stock and Watson (1998), the researchers are aware that they are only simulating pseudo real-time forecasts when they use the most recent data vintage.[3] In fact, even though Fair and Shiller explicitly address the first three issues in their paper, they admit that they use only the latest revised data in their pseudo real-time forecast evaluation.[4] Presumably the argument is that using real-time vintages of data simply does not matter for the results, but almost no work has been conducted to investigate whether there is evidence to support this proposition.

Moreover, because Fair and Shiller used only the latest vintage of data in their analysis, they did not have to deal with the equally important conceptual issue of which data vintage to evaluate the forecast against. Quantifying forecast accuracy requires a benchmark series against which to compare forecasts. The most recent vintage of data is often suggested as that benchmark because these data give a somewhat cleaner and more accurate measurement. But the frequent redefinition and rebenchmarking of the data series may alter the series properties in ways that a forecaster cannot be expected to predict. Moreover, while using the latest available estimates places the forecasts against measures with the least measurement error, forecasters are most likely to be held accountable for their ability to forecast, say, real GDP growth, using an estimate that is available not too long after the quarter being forecast. This article proposes that a decision about the data vintage that the forecasts are to be evaluated against should be considered prior to beginning a forecasting exercise.

Recent Research on Real-Time Forecast Evaluation

Recent research on the accuracy of forecasting models has moved closer to satisfying the necessary conditions of a real-time exercise as laid out in Fair and Shiller. The key criterion seems simple: if the forecasts cannot be reproduced using the available data

[3] See Staiger, Stock, and Watson (1997, note 8) and Stock and Watson (1998, note 1).

[4] Fair and Shiller, using the Fair model, would have faced a daunting task in compiling a real-time data set of the hundreds of data series involved. Similarly, Stock and Watson (1998) employed more than 200 different time series in their forecasting study.

BOX 1

The Composite Index of Leading Economic Indicators

The composite leading economic indicator (CLI) series was developed as a tool for business cycle analysis in the late 1960s. Prior to December 1995, the BEA produced the index of leading indicators data series. As of December 1995, the BEA stopped publishing the CLI and the Conference Board took over its production and publication. Detailed information regarding the construction of the CLI is available at the Conference Board Web site (www.tcbindicators.org).

The idea of the CLI is to summarize in one series the data on variables that typically move in a business cycle pattern prior to standard measures of economic output such as GDP. The primary objective is to help detect, ex ante, turning points in the business cycle—that is, whether the economy was likely to enter a recession (or to recover and grow out of a business cycle contraction). The business cycle research of Burns and Mitchell in the 1930s and 1940s helped motivate indicator analysis; however, the predominant researcher associated with the indicators (of which the leading indicators index is only one) was Geoffrey H. Moore (1990). While the CLI is primarily used as a turning point predictor, in recent publications the Conference Board has also suggested that it may be useful for forecasting the growth in economic output over time (Conference Board 1997, 1998).

The CLI is constructed as a weighted average of several publicly available data series. Currently there are ten component series in the index although both the number of series and the specific series used have changed over time. The weights applied to each series in forming the index are occasionally revised, and the index is usually recalculated every year to incorporate historical revisions to the component data.

Changes to the component series and the associated weights are in response to perceived changes in the empirical relationships between the components and the business cycle. The June 1997 issue of *Business Cycle Indicators* discusses in detail how the composition of the CLI has changed over time. The appendix in Beckman (1997) annotates the numerous revisions and improvements to the CLI historical data series.

It appears that changes to the composition of the CLI need not be substantial for them to be important in a real-time sense. Diebold and Rudebusch (1991) show that using revised historical CLI series in tests of the forecast value of the CLI generates spurious results supporting significant forecasting power for the CLI in predicting an index of industrial production. Because the results use revised data, the revised CLI reflects future information in both the choices of the component series as well as in the weights assigned to the component series in the index. The argument is that in the revision process the CLI is designed to maximize its correlation with the business cycle, so it would not be surprising that empirical results using revised data support the forecast power of the CLI more than do real-time vintages of the CLI. In contrast to those results, Hamilton and Perez-Quiros (1996) and the results in this article support a real-time role for the CLI in forecasting growth rates of real GNP and GDP, respectively.

set, then the exercise is not real-time and is unlikely to be a useful evaluation of real-time performance. Several studies that introduce aspects of real-time data construction fail to satisfy this criterion completely. Some other studies use real-time data sets but fail to use the most up-to-date versions available at the time the forecasts were made.

Research by Makridakis and others (1993) provides a good example of the few studies that undertake a real-time forecast analysis. In their research design, the authors provide real-time data sets to a group of forecasters and ask them to make forecasts several periods into the future. The research then evaluates the accuracy of the forecasts years later, after the actual data for the forecast observations have been released in a relatively final form. This type of research is a valuable contribution to the forecasting literature in that it evaluates forecasts in a true real-time framework.

One drawback to this approach is the long time lag needed to generate forecast accuracy results. The forecasters were given real-time data in 1987, 1988, and 1989, and forecasts were made for up to fifteen months ahead in each case. However, the authors performed the evaluation in 1991, using the vintage of historical data then available. For an academic exercise, the work is useful and informative. From a policy perspective, it is too slow in producing the information necessary to distinguish between good and bad forecasting models.

In related work, McNees (1992, 1995) investigates the GDP (or gross national product [GNP]) forecasting performance of several private- and public-sector macroeconomic forecasters. McNees evaluates the outcomes of the real-time forecasts over several historical periods, comparing the forecasts with revised GDP (or GNP) data. Although McNees discusses the issue of what vintage GDP the forecasts should be compared with, he bases his decision to use the most revised GDP (GNP) series on the idea that this vintage of data has the least measurement error.

McNees's studies are important checks on the forecast accuracy of many macroeconomic forecasters. By maintaining a real-time data set of the actual forecasts, he offers objective evaluation criteria for real-time macroeconomic forecasts. His studies, however, do not focus on constructing or evaluating forecasting models per se but on forecast outcomes. Thus, there is no analysis of the impact of real-time data on models or on model selection.

Recent studies by Swanson (1996) and Swanson and White (1997a, b) also provide a useful benchmark for research on real-time aspects of forecasting. Swanson (1996) collects the initial (or first-reported) estimates for a number of variables. For example, for GDP he creates a vector of all the advance GDP estimates for each quarter. These data clearly could be used in a real-time forecasting environment. However, these data are not what a forecaster would actually use in generating a forecast. For example, rather than using a vector of GDP advance estimates, a forecaster would use a vector defined by the available data vintage, containing the newly released observation (advance, preliminary, or final) along with revised values for the prior observations. Thus, at the end of July 1988, the GDP data through the end of 1982 would be obtained from the BEA's December 1985 benchmark (historical) revision. Those for 1983 are from the July 1986 annual revision, those for 1984 are from the July 1987 annual revision, and those for 1985 through the first quarter of 1988 are from the July 1988 annual revision. In essence, a real-time data set is the time sequence of vintages of data—each vintage is a vector of data values. A newer vintage data vector usually contains more observations than does an older vintage and has also usually been subjected to more revisions.[5]

Swanson (1996) and Swanson and White (1997a, b) have used the data set of initial estimates in a number of empirical analyses. For instance, Swanson (1996) compares a set of statistical tests formed using the initial estimate data with outcomes obtained using the most recent vintage of data. He finds a number of instances in which the test results are substantially different, suggesting that data revision matters. However, Swanson provides a test of a more extreme information restriction than would represent an actual real-time forecasting effort. If Swanson found no difference between using initial estimates and latest available data, then it is doubtful that the difference between real-time data and latest available data would be important for forecast accuracy results.

Diebold and Rudebusch (1991) and Hamilton and Perez-Quiros (1996) present empirical results of out-of-sample forecast analyses in which one of the two variables in the forecasting models was measured in a real-time context. Diebold and Rudebusch examine whether the composite index of leading economic indicators (CLI) is useful for forecasting real output. Specifically, they investigate how well the CLI can forecast the industrial production index relative to an autoregressive model that uses only current and past values of the IP index. They cite previous research suggesting the CLI series was a strong predictor of IP. However, Diebold and Rudebusch hypothesized that using only the latest revised vintage of historical CLI data might have inflated the significance of the CLI's forecasting potential for output. They argue that the construction of the CLI has been subject to change (see Box 1), and these changes might constitute ex post attempts to better correlate the CLI with output. Using a real-time CLI series, they find that the CLI does not add significant forecasting power to an autoregressive forecasting model for IP.

Recent research on the accuracy of forecasting models has moved closer to satisfying the necessary conditions of a real-time exercise.

Notably, Diebold and Rudebusch decided not to use real-time IP data in their empirical analysis. In particular, they used only the latest vintage when estimating the models and evaluating the forecast accuracy. By using the most recent vintage of IP data when forming their forecasts, they were in fact doing something impossible in real time. Diebold and Rudebusch justify this decision by arguing that they are searching for evidence on “the

[5] *Over the period covered by this data set, the BEA rebenchmarked the data series several times, making the level of the real GDP series discontinuous. To address this problem, Swanson converts all the initial estimates into a single series based in 1987. Doing so raises the issue of the influence of the benchmark on the behavior of the spliced data series.*

ability of the CLI to forecast truth, which is taken to be the final IP value" (1991, 609). They reason that the latest revision is the best estimate of real output. In other words, using real-time IP in fitting the model and constructing the forecasts could mask the fact that the CLI has little "intrinsic" forecasting ability for the "true" IP. Any forecasting ability when using real-time IP would reflect that the CLI was simply compensating for the inadequacies of the real-time IP measure. Of course, this feature is characteristic of any real-time forecasting problem; the best data estimates are not usually available at the time a forecast is made. Moreover, while one can debate which vintage of IP the forecasts should be evaluated against, waiting (up to twenty years in Diebold and Rudebusch's case) for a more refined IP estimate is hardly a realistic strategy for judging a professional forecaster or policy model.

Hamilton and Perez-Quiros (1996) also examine the real-time forecast performance of the CLI, but they focus on forecasting real GNP rather than industrial production. To translate the monthly CLI observations to a quarterly frequency, they use the first revised CLI estimate for the last month in the quarter. This number is usually released late in the second month of the next quarter. For example, they would use the revision of the March CLI that is released late in May, making the March CLI estimate roughly contemporaneous with the preliminary estimate for the first quarter's real GNP. Despite incorporating sufficient detail for making the vintage of CLI approximate real-time data, Hamilton and Perez-Quiros do not take into account the real-time availability of the GNP series in constructing the forecasts. Their justification is that they only "want to evaluate how close the forecast is to the value of GNP as ultimately revised" (1996, 42). However, as argued above, the choice of vintage to evaluate the forecasts against is somewhat different from the problem of constructing forecasts in real time. Hamilton and Perez-Quiros effectively make no distinction between data availability in constructing real-time forecasts and in evaluating the subsequent forecast.[6]

Real-Time Forecasting Experiments—Comparisons with Latest Vintages of Data

A simple experiment helps examine how the data revision process affects the selection and evaluation of economic forecasting models. The objective is to uncover whether the CLI can help forecast economic activity in real time. Separate forecasting models are estimated for two economic output measures—quarterly real GDP growth and the monthly growth rate of IP.

(1) Forecasts are constructed at the end of a month. The approach is to choose the specification of the forecasting model for each output series in real time and then estimate the coefficients of the chosen model using only the data actually available at the time the forecast is made.[7]

(2) Forecasts of GDP growth and IP growth were constructed for each of the next two time periods—quarters for GDP and months for IP.

(3) The forecasts were compared with subsequently announced values of the output series. For the real GDP series, the comparison is with the final estimate of the growth rate reported three months after the advance estimate. For the IP series, it is with the first release of the next month of data, as well as against the next two subsequent revisions of that initial estimate.

(4) Repeating the above steps each period through 1998 generates a sequence of real-time, one- and two-period-ahead forecasts for GDP and IP growth. The last step is to compute summary measures of forecast accuracy from these sets of forecast errors.

To be concrete, suppose the task is to examine a forecast of the growth rate of IP for September 1988 and the forecast is formed at the end of July 1988. This is a two-month-ahead forecast. At the end of July there is available an initial estimate of IP for June, a revised estimate for May, a second revised estimate for April, and a historical series constructed from annual revisions in 1986 and 1987 and the benchmark revision released in December 1985. At the end of July there is also an initial estimate of the CLI for June 1988. This estimate can be combined with a historical CLI series obtained from a major revision in February 1983, a revision to post-1983 data in March 1987, and a revision to the most recent twelve months' data in July 1988. This is the data set used to determine the model specification (lag length), estimate the model coefficients, and make a forecast of monthly IP growth for September 1988. The resulting forecast is compared with the initial estimate of IP for September released in October as well as with the revisions released in November and December.

The goal is to determine whether following the above procedure for replicating real-time forecasts produces results that compel inferences different from those of an analogous simulation using the latest available data vintage throughout. An experiment using the July 1998 vintage of historical time series investigates these differences, following the above steps and acting as if the numbers in this data set were actually available in real time. Thus, in updating the forecasting model, a new observation is added, but the historical observations do not change. Similarly, the accuracy of the resulting forecasting model is always evaluated against the latest, 1998 vintage of historical data.

The specification of the output models that include the CLI uses a bivariate vector autoregression (VAR) of the form

$$\begin{aligned}\Delta y_t &= m_1 + \sum_{i=1}^{p}(a_i \Delta y_{t-i} + b_i \Delta x_{t-i}) + u_t \\ \Delta x_t &= m_2 + \sum_{i=1}^{p}(c_i \Delta y_{t-i} + d_i \Delta x_{t-i}) + v_t\end{aligned} \qquad (1)$$

where Δ denotes the first-difference operation. When t represents a quarter, y_t is 400 times the natural logarithm of GDP for quarter t and x_t is 400 times the natural logarithm of the CLI for quarter t. When t represents months, y_t is 1,200 times the natural logarithm of IP for month t and x_t is 1,200 times the natural logarithm of the CLI for month t. The errors u_{t+h} and v_{t+h} for $h = 1$ and 2 are taken to be unforecastable relative to the current and past of Δy_t and Δx_t and so are set to zero in constructing the forecasts of Δy_{t+1} and Δy_{t+2}. The number of lagged observations to include, p, and the values of the coefficients are all unknowns estimated from the available historical data at the time the forecast is made. The lag length is chosen by selecting the p that minimizes the so-called Akaike information criterion (AIC). The AIC is a statistic that trades off the improved fit of the model to the data, gained by including more lags, with the cost of having to estimate more and more coefficients from a fixed number of observations.

An autoregressive (AR) model for output growth is obtained by imposing the restriction in the first equation of (1) that the $b_i = 0$ for $i = 1, \ldots, p$. The AR model ignores the CLI data completely and relies instead solely on current and past values of the output measure for forecasting future output growth. The AR model thus provides a benchmark against which the VAR's forecasts can be compared, although such a comparison is a rather low hurdle. Notice also that specifying the models in the first differences of the variables precludes the possibility that the levels of the CLI and output might provide additional forecasting ability to the models.[8]

Using the CLI to Forecast Real GDP Growth in Real Time. The experiment examines whether the CLI measured in real time helps forecast real GDP growth over one- and two-quarter forecast horizons. To construct a real-time forecasting test, the study uses only data for both the CLI and real GDP that would have been available at the time the forecasts were made. In essence the experiment is examining the same basic issue that Hamilton and Perez-Quiros (1996) explored, but altering the data set in a number of ways ensures that both the construction and evaluation of the forecasts are closer to real-time exercises. The results from the real-time simulation are then compared with those obtained using the most recent vintage of time series in a pseudo real-time analysis.

As explained in Box 2 on the construction of GDP and Box 1 on the CLI, the historical data on GDP and the CLI can change from month to month. Making the real-time GDP data coincide with the timing of the leading indicator series means considering a number of alternatives. One possibility is to use the second revision of the CLI estimate that corresponds to the last month of the quarter (as opposed to the first revision used in Hamilton and Perez-Quiros 1996). Thus, for example, a second revision of the March CLI is released in late June, and this revision could be aligned with the final estimate of the first-quarter GDP released in mid- to late June.[9]

Differences in the assessment of forecast performance arise primarily from the choice of series to evaluate against.

Deciding to use a real-time data set pins down one aspect of the forecast evaluation exercise, but there are many other choices to be made that may, in principle, qualitatively affect the results. Of course, this difficulty appears precisely because the test is replicating a real-time forecasting problem.

As noted above, the model specification employs growth rates of the CLI and real GDP. The VAR model is first fit to data covering the period from 1959:1 to 1977:1, with a maximum of $p = 4$ lags of each variable included in the VAR. Respecifying the lag structure and reestimating the model's coefficients for each quarter through 1997:4 provides a framework allowing the most flexibility

[6] *Another curious feature of the Hamilton and Perez-Quiros study is that the authors estimate the chosen model specification up to 1975:3 in order to generate the out-of-sample forecasts from 1975:4 to 1993:2. Thus, they ignore all the interim information in the data that may alter the coefficient estimates. In real time, a forecaster would likely attempt to incorporate more recent information by updating the coefficient estimates periodically.*

[7] *The model specification is described in equation 1.*

[8] *The growth rate forecast results are all qualitatively the same if the VAR and AR models are fitted in levels of GDP, the CLI, and IP. However, the forecast accuracy of each model is always greater when the growth rate model specification is used.*

[9] *By the time these data are collected the current quarter is virtually over, a fait accompli. In essence, then, a one-quarter-ahead forecast amounts to predicting how the BEA will measure that quarter's GDP growth. As an alternative, the experiment also matched, for example, the first revision of March's CLI with the advance GDP estimate for the first quarter, both of which are released late in April. Hence, the forecast could be made only one month into the second quarter. Using this earlier vintage of data did not materially affect the forecast accuracy results described in the next subsection. The issue of matching vintages of the CLI and GDP data does not arise if one uses only the latest available data vintage.*

BEA Revisions of NIPA Economic Data

The Bureau of Economic Analysis (BEA), a division of the U.S. Department of Commerce, puts a substantial amount of its resources into the production of the National Income and Product Accounts (NIPA). Its efforts include both source data gathering—that is, compiling data on the measures that are combined into GDP—and statistical refinement of the data (seasonal adjustment, redefinitions, and rebenchmarking of the price-deflated series). Data production must meet the demands of its users for timely release of data measures but must also take great care to produce accurate data measurement. In such an environment, the BEA tries to satisfy both needs—timeliness and accuracy—by producing three estimates of GDP for the prior quarter.[1]

The first estimate, referred to as the advance GDP estimate, is released toward the end of the month immediately following the quarter to which the data refer. This advance estimate of GDP is based on incomplete source data, but it usually provides a fairly good forecast of the value of future revisions to that quarter's GDP because of how much source data, mainly on consumption, is available. However, because the BEA lacks complete source data for some subcomponents it must make judgmental assumptions about the likely values taken by specific GDP components. This action is simply "forecasting" what the outcome might be.

The so-called preliminary GDP estimate is a revision to the advance estimate that is released toward the end of the second month after the quarter, and a final GDP estimate for the quarter is released toward the end of the third month. The main reason for the revisions of these numbers from the advance to preliminary (and to final) estimates is that the source data for these measures take time to arrive at the BEA to be compiled into the statistics.

The BEA schedules additional revisions that improve the accuracy of the GDP estimate after the release of the final GDP estimate. The revision occurs at this time to improve the estimate of seasonal adjustment of the data. These annual revisions of the data usually occur in July, revising the data in the prior three calendar years as well as the one quarter of data available for the given year.

The BEA then revises the entire history of the quarterly data series (currently back to 1959) approximately every five years in what is known as a benchmark revision. At these benchmark revisions, there are often redefinitions of component data series that revise the entire history of the series. Here, the BEA updates the base year for computing the real measure of GDP and the implicit deflator. Base-year changes often have substantial effects on the overall estimates of economic growth because the initial relative price conditions set in the benchmark year may change dramatically as time passes. In other words, base-year effects alter growth rates in real GDP because the relative prices at which the new real GDP estimate is calculated could differ from those of the previous base year.

The most obvious example of this phenomenon is the 1972 benchmark during the mid-1970s. There were substantial oil price increases in 1973 through 1974 that changed dramatically the relative prices between oil and other goods. Deflating nominal oil prices by using a deflator based on 1972 prices lowered the measured adverse impact of oil imports on real net exports. Looking at GDP measures based on alternative base years provides quite different views of the depth of the economic contraction during the 1974–75 recession. Using a 1977 base year, the relative size of the economic contraction appears larger because oil was a larger share of imports in that year (as a result of both higher prices and the larger quantity of imported oil). Benchmark revision data are generally more accurate because they use more revised source data from which to measure economic activity. For instance, the BEA uses more final information sources because there is more time to check the validity of initial reports.

In 1995 the BEA changed the definition of the real GDP measure by moving to a chain-weighted index, which allows for the effects of changes in relative prices and in the composition of output over time (see Landefeld and Parker 1997). This change in the definition of real GDP alters the behavior of the estimated real GDP series relative to the prior, fixed-weight constant-dollar estimates. If a forecaster uses this series as the series against which real-time forecasts are compared, then implicitly the forecaster is attempting to forecast the change in the definition of real GDP. Part of the motivation for this study is to investigate whether the change in definition is a sizable problem for real-time forecasters.

Research by the BEA (such as Young 1993) examines how each subsequent revision in the GDP series has changed the series and how good initial growth rate measures are as estimates of the more recent vintages of the series. Simply stated, the changes from one announcement to the next reflect the tension between the need for data that are both timely and accurate. One expects the later estimate to be more accurate than the advance estimate, but it is not always.

What is the source of the revisions? First, as time passes, the BEA can replace preliminary source data with more revised or comprehensive data. More complete monthly data may be one example of this data revision source. As mentioned above, the advance estimate of real GDP often contains BEA judgmental estimates of measures that the BEA does not possess only one month after the end of the quarter. In the subsequent revisions, these estimates are replaced with source data as they become available. This type of revision occurs with relatively high frequency. One would expect that the BEA would on average be relatively accurate so that the advance and preliminary real GDP data would not possess an obvious bias, for example, an average positive or negative error relative to the final measure for that quarter. Young (1993) offers evidence to support the accuracy of the estimates of real GDP growth rates; in the most recent sample, there appears to be no significant bias in the three announcements or even in the advance estimate relative to a temporally close "latest available" estimate.[2]

As argued by Mariano and Tanizaki (1995), the "true" real GDP measure for any particular quarter is effectively unobservable because the current measure will be subject to future revision. The position held here is that the latest available time series of real GDP is the best historical record currently available. However, no part of this currently available data set would have been available to a researcher in earlier time periods when making forecasts of the then-unknown future values of the series. In the same way, no researcher today has available the data set that will eventually exist when subsequent revisions are made in one or five or ten years.

[1] The BEA shifted its focus in November 1991 from reporting gross national product (production by U.S. nationals regardless of the location of the factors of production) to GDP (production within the borders of the United States regardless of production factor ownership). Real GNP was long recognized as harder to produce in a timely fashion than real GDP because there are little reliable, timely data on net income from foreign sources. For the United States the numerical difference between the two constructs is relatively small.

[2] Importantly, Young (1993) avoids comparing the advance estimate with the most revised, latest available estimate because the latest version is often a substantially revised measure of a possibly even redefined construct. Fleming, Jordan, and Lang (1996) examine accuracy of the measured level of real GDP over the limited sample 1985 to 1991 and find evidence of sizable and systematic measurement bias. Young's results, however, suggest that these findings do not translate into systematic biases for the growth rates.

for the statistical model to adjust to the new information that arrives with each additional observation. Each time the model was reestimated, new one- and two-quarter-ahead forecasts were generated, yielding a set of eighty-four one-quarter-ahead forecasts and eighty-three two-quarter-ahead forecasts that could be evaluated against final GDP numbers for 1977:2 to 1998:1. Forecasts constructed using real-time data were compared with forecasts from models estimated using the most recent vintage data.

For measuring the accuracy of the forecasts, the decision of which vintage data to compare the forecasts against is an important one. In real time, this choice is clearly important. For example, who knew in 1981 (or even 1991) that the BEA would change to the use of chain-weighted real GDP in 1996? It is perhaps unfair to burden the forecaster in 1981 with the problem of also forecasting definitional changes in the data series. There is a second issue: The most-revised data series takes many years to produce. It is doubtful that a forecaster's accuracy would not be evaluated until years later when the most-revised time series is determined. More likely, the forecasts will be compared with the final GDP growth estimate released approximately three months after the end of the quarter being forecast. Thus, the study compares the actual real-time forecasts to the growth rate implied by the initial final GDP estimate, whereas it compares the pseudo real-time forecasts with the most recent (1998) vintage of data.

Empirical Results for Real GDP. Relative forecast accuracy results are reported based on the root mean squared error (RMSE). This figure is simply the square root of the average of the squared forecast errors, with the square root taken to put the measure back into the units of the variable being forecast (that is, annualized percentage points). The same pattern of results appears using other standard forecast accuracy measures such as the average absolute forecast errors (the average of the sum of forecast errors with sign disregarded). One would expect the variability of the errors to be smaller the more accurate the model is.

Table 1 displays the summary forecast statistics comparing the one- and two-quarter-ahead forecasting performance for the VAR and autoregressive models. For

TABLE 1 RMSE of Pseudo and Real-Time GDP Growth Forecasts

		Forecast Model			
		VAR		AR	
Forecast Type and Timing	Evaluated against	One-Quarter	Two-Quarter	One-Quarter	Two-Quarter
Pseudo (July 1998)	July 1998 vintage	3.02	3.13	3.40	3.53
Real-time (end of quarter)	Initial final estimate	2.56	2.73	2.88	3.00

Source: GDP, Bureau of Economic Analysis; CLI, Conference Board

both forecast horizons and all the data sets, the RMSE of the VAR is considerably less than the RMSE of the AR model. Thus, the model that includes the CLI provides more accurate forecasts than the AR alternative. Consistent with this finding, the real-time forecasts are more highly correlated with the initial final vintage of GDP growth estimates, and the pseudo real-time forecasts are more highly correlated with the latest available vintage of estimates than are either of the corresponding forecasts from the AR model.

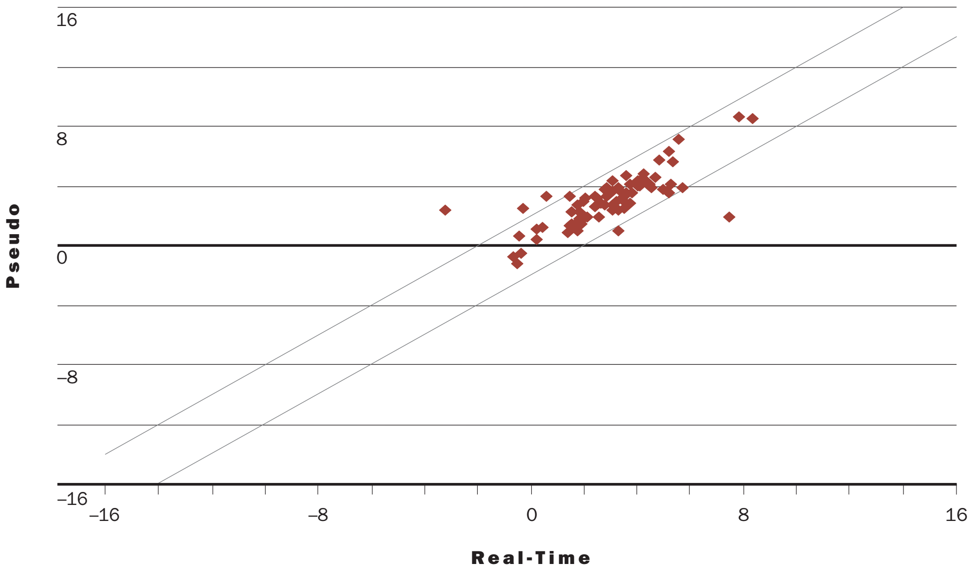
In discussing the results, it is more compact to report the ratio of the RMSE of the VAR model with CLI to the RMSE of the AR model, given that both models employ the same data set restrictions. Using the latest available data vintage, the ratio of RMSE is 0.89 for the one- and two-step horizons. The ratios using real-time data are 0.89 for the one-step horizon and 0.91 for the two-step horizon. The differences between the real-time and latest available data vintages do not seem to change the basic inference that the CLI has some marginal predictive power for GDP.

The next step examines the differences between the measured accuracy of the actual and pseudo real-time VAR forecasts in a little more detail. Chart 3 is a scatter diagram of the one-quarter-ahead forecasts from the VAR for the period from 1977:2 to 1998:1. In the chart, each point reflects the actual real-time forecast on the x-axis and the corresponding pseudo real-time forecast on the y-axis. Points along the 45-degree line indicate that the two forecasts are the same. Only six of the eighty-four forecasts (7 percent) differ by more than 2 percentage points, emphasizing that the forecasts are quite similar despite being based on different vintages of historical data.[10] Chart 4 is a scatter diagram of quarter-on-quarter GDP growth rate estimates for 1977:2 to 1998:1. The July 1998 vintage of GDP growth estimates is on the y-axis, and the initial final vintage is on the x-axis. Examining this chart reveals that thirteen observations out of the eighty-four, slightly more than 15 percent of the revisions, are more than 2 percentage points apart. Comparing Chart 3 with Chart 4 makes it clear that the difference in the measured forecast accuracy arises primarily from the variation between the versions of data being forecast rather than the forecasts themselves.[11] The mean and standard deviation of the most recent vintage are 2.8 and 3.5, whereas the same statistics for the initial final estimates are 2.6 and 3.0, respectively. It is notable that the later vintage of real GDP growth is also more variable.

This empirical evidence indicates that the CLI helps predict real GDP growth, but the forecast accuracy statistics themselves are rather unimpressive. To illustrate this fact consider what happens if one uses the advance estimate of real GDP growth as the forecast for that quarter. The advance estimate is available approximately three weeks after the real-time VAR/AR forecasts are formed at the end of the quarter. Thus, for example, instead of a forecast of second-quarter GDP growth formed at the end of June, the advance estimate can be thought of as a second-quarter forecast formed late in July. In the first case, the advance forecast generates a humbling 0.75 percent RMSE when evaluated against the resulting final estimate of real GDP. This percentage is substantially lower than the RMSE of 2.56 obtained from the real-time VAR model, as reported in Table 1. Moreover, the correlation of the advance estimate with the final estimate is approximately 0.97, as compared with only 0.54 for the real-time VAR model forecasts. The strong results are understandable since the advance GDP estimate uses the same BEA data measurement design and much of the same source data are used to generate the final estimate released only two months later.

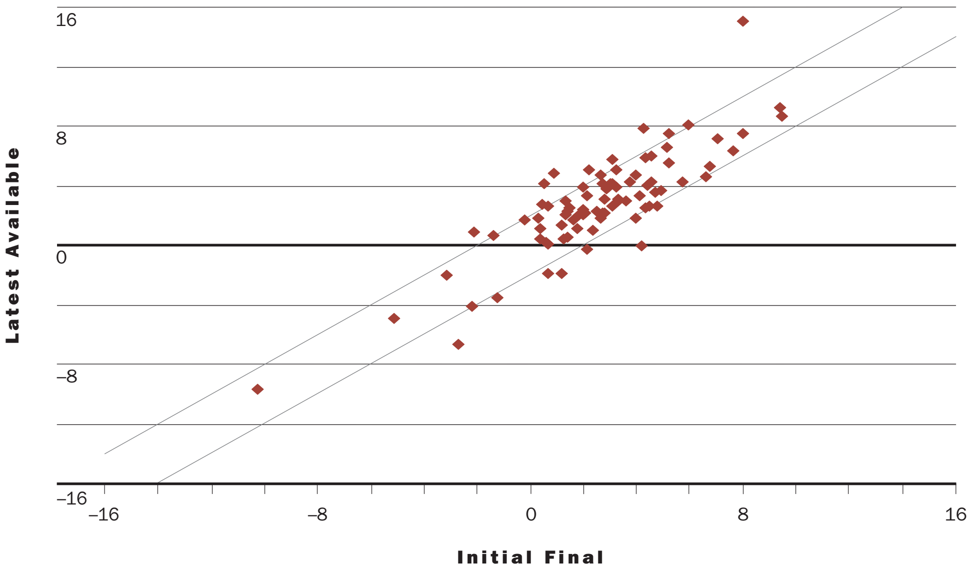
When the advance estimate is compared with the latest available vintage of GDP growth for the current quarter, the forecast error increases substantially—the RMSE is 1.9 percent—and the correlation with the latest vintage of estimated GDP growth drops to 0.84. The increase in the forecast error and the decline in the correlation emphasize the importance in the choice of the estimate against which forecasts will be evaluated. Revision and redefinition of the GDP series over time almost guarantees that real-time forecasts will worsen when forecast accuracy statistics are taken relative to the most recent vintage of data. Nonetheless, the advance

CHART 3 One-Quarter-Ahead VAR GDP Growth Forecasts, 1977:2–1998:1



Source: GDP, Bureau of Economic Analysis; CLI, Conference Board

CHART 4 GDP Growth Estimates, 1977:2–1998:1



Source: Bureau of Economic Analysis

10. The fact that the forecasts simply are not very different is emphasized in observing that comparing the real-time forecasts with the most recent vintage of data yields RMSE values that are virtually the same as those from the pseudo out-of-sample forecasts.

11. The finding that real-time GDP forecasts are more accurate for the penultimate data than the pseudo real-time forecasts are for the most recent data vintage is consistent with McNees (1988), who notes that forecast errors are generally smaller when real-time forecasts are compared with less-revised vintages of data.

BOX 3

The Index of Industrial Production

The Board of Governors (BOG) of the Federal Reserve System produces the index of industrial production (IP) on a monthly basis; it is one of the few measures of output produced monthly.[1] The index estimates output in a number of industrial sectors that, combined, currently account for approximately 25 percent of total output. That portion of total output (as measured by real GDP) has diminished over time but remains an oft-cited economic statistic. It measures the change in output in the following industrial sectors: manufacturing, mining, and electric and gas utilities. Output is measured in physical units, rather than by price and quantity. The index excludes output in other activities, such as agriculture and services. Thus, the index numbers that the BOG releases midmonth are estimates of the monthly level of total output of the nation's factories, mines, and gas and electric utilities.

The construction of the IP index involves a substantial degree of estimation. Typically, less reliable source data are available on a more timely basis than are the more accurate measures used to revise estimates of IP. But even in the composition of IP, there are three different types of input series for estimating the index: physical product, production-worker hours, and electrical power use by industry. For some industries the monthly estimates of production are based on measures of physical output, and that is the most desirable measure of output. Physical product counts the physical output in quantity. These data are not often available in a timely manner. For industries in which direct measurement of physical product is not readily available, the BOG estimates (infers) a measure of output from production-worker hours per industry from numbers produced by the Bureau of the Census or from electrical power use. These data are used in lieu of the physical product data that are not yet available.

The IP figures are available in the middle of the month following the month they measure. The BOG issues preliminary data for the preceding month, and these data are subsequently revised in the next three months. Annual revisions are made in the fall. The BOG revises the series to a greater degree on a periodic basis, linking or benchmarking the individual industrial production series to more comprehensive data sources. One of the major sources for benchmark revisions is the Census of Manufactures, which is released every five years. The IP index was built, for the most part, in five-year segments, each with value-added weights taken from the census year. Now, like real GDP, the IP index is a chain-weighted index.

The major revisions are in the IP series (1971, 1976, 1985, 1990, and in 1997). The BOG completed a revision of its measures of industrial production in January 1997. The primary feature of that particular revision was a new formulation for aggregating the index using weights that are updated annually instead of every five years. The revisions of the data series went back as far as 1977, but some additional changes were made to data from 1976 back to 1967 to improve their consistency with the new data formulation. In addition, the revision also involved the rebasing of the total IP series back to the initial observation (1919); the data are now expressed as percentages of output in 1992.

The IP index, despite covering only about 20 percent of total U.S. output, measures industries that may account for a large proportion of output volatility during a business cycle. Typically, the IP index rises more during economic expansions, and contracts more during economic downturns, than the aggregate real GDP series. Also, it is released more frequently than other output measures (like real GDP). However, the timeliness of the series must also be compared with its measurement error: relative to real GDP, IP estimates appear to have more substantial measurement error.

[1] *Frumkin (1994) and Rogers (1998) provide detailed information on the construction, release timing, and revision schedule of various economic indicators including the IP index, as well as their standard uses and interpretations.*

TABLE 2 RMSE of Pseudo and Real-Time IP Growth Forecasts

		Forecast Model			
		VAR		AR	
Forecast Type and Timing	Evaluated against	One-Month	Two-Month	One-Month	Two-Month
Pseudo (July 1998)	July 1998 vintage	5.50	5.39	6.10	6.00
Real-time (end of month)	First revised estimate	5.41	5.40	5.33	5.49

Source: IP, Board of Governors of the Federal Reserve System; CLI, Conference Board

estimate performs considerably better than either of the one-quarter-ahead real-time forecasting models.

Using the CLI to Forecast Growth of Industrial Production in Real Time. Another examination of the forecasting properties of the CLI looks at its marginal forecasting contribution for IP, a more frequently released output measure. As described in Box 3, IP is a less general output measure than real GDP—it measures only output in mining, manufacturing, and electric and gas utilities—comprising somewhere around 25 percent of total U.S. output. Still, IP is an oft-cited economic measure and was central to Diebold and Rudebusch's (1991) research on the predictive power of the CLI.

The real GDP forecasting example combines a more frequent activity indicator (CLI) with the quarterly real activity measure. In that application, it was necessary to choose the CLI measure for the month in the quarter that coincided with the relevant release of quarterly real GDP growth data. For the IP forecasting application, the same statistical framework is used, but the model is fitted to the CLI and IP on a monthly frequency. A monthly IP measure for a given month is released in the middle of the subsequent month. The corresponding CLI number is released at the end of the month following the month to which it refers. Because of this slight staggering in the releases within the month, it is assumed that the IP forecasts are formed at the end of the month so that both figures are available for the prior month. This decision rule, then, means that at the end of July 1988, say, there are measures of the CLI and IP up to June 1988, and IP growth is forecast for July and August. The Board of Governors of the Federal Reserve System releases an initial IP estimate for July in mid-August, or about fifteen days after the forecast is made. A first revision of this estimate is released one month later in mid-September, and a second revision is reported after another month has elapsed. As noted above, the real-time forecasts are compared with the initial estimate, the first revision, and the second revision. However, these different vintages have little impact on the results for the real-time forecasting models, so results report only what is based on comparison with the first revised data. In contrast, the pseudo real-time forecasts are always compared with the vintage of historical data available in July 1988.

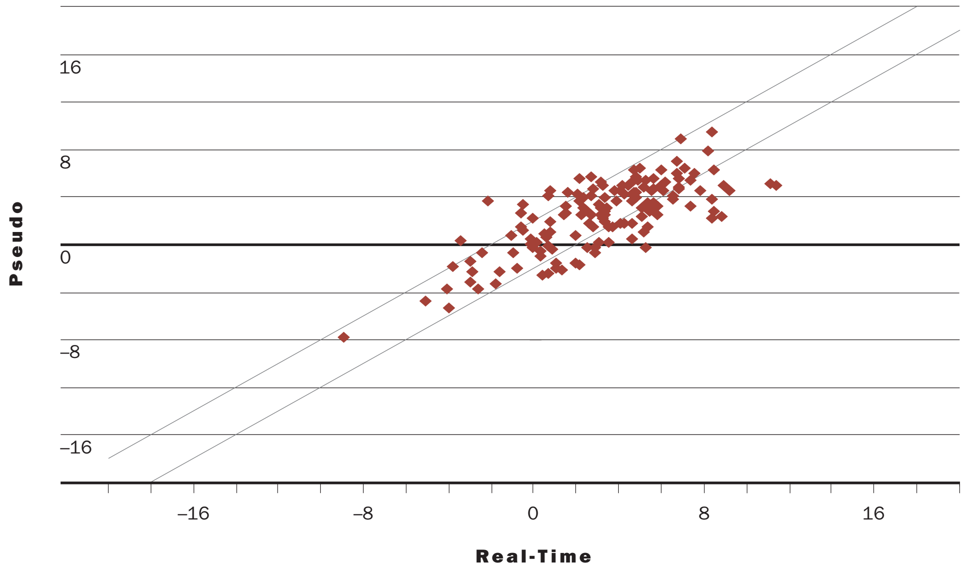
The statistical model shown in equation 1 is employed to relate the monthly growth rates of the IP and CLI series. The models are first fit to data covering the period from 1959:01 to 1985:12, with a maximum of $p = 12$ lags of each variable included in the VAR. The actual lag length is chosen via the AIC. The lag structure is then respecified and the model's coefficients are reestimated for each month through 1998:02. Each time the model was reestimated new one- and two-month-ahead forecasts were generated, yielding a set of 147 one-month-ahead forecasts and 146 two-month-ahead forecasts that could be evaluated against IP growth rates for 1986:01 to 1998:03.[12]

To analyze the contribution of the CLI to forecasts of IP, the forecast accuracy of models that use only past observations of IP (AR models) is compared with that of models that also exploit the CLI data (the VAR model). This simple criterion provides the same low hurdle for the CLI data that were examined with the real GDP data: if the CLI data contribute to the forecast accuracy of IP, then the VAR model that includes the CLI will generate forecasts with lower RMSE than those of the AR models.

Table 2 presents the summary forecast statistics comparing the one- and two-month-ahead forecasting performance for the VAR and autoregressive models. For the pseudo real-time forecasts the RMSE of the VAR is considerably less than that of the AR model at both forecast horizons. The RMSE ratio is 0.90, suggesting that the use of the CLI helps reduce forecast error by about 10 percent. For the two-step horizon, the ratio is 0.89. In a separate experiment combining real-time CLI

[12] *The study notes that Diebold and Rudebusch (1991) examine whether the level of the CLI improves the one-step-ahead forecasts for the level of IP. The model examines the ability of growth rates (percentage changes) in the CLI to help forecast the growth in IP so that the two sets of results are not directly comparable. The results with IP, however, produce inferences that are comparable to those made in their research.*

CHART 5 One-Month-Ahead VAR IP Growth Forecasts, 1986:01–1998:03



Source: IP, Board of Governors of the Federal Reserve System; CLI, Conference Board

data along with latest available IP data (similar to the work of Diebold and Rudebusch), the RMSE ratio moves to 0.95 and 0.93. Thus, using a hybrid data set that mixes data vintages reduces the measured forecast improvement, but it still suggests that the CLI provides a very marginal improvement in forecasting accuracy over the forecasting model that simply uses lags of IP.

The one-month-ahead forecasts from the real-time VAR and AR models evaluated against the first revised IP estimates produced a RMSE ratio of 1.03, suggesting that real-time CLI actually worsens the VAR model's forecast accuracy relative to a simple AR model. The ratio for the two-step-forecasting horizon is 0.98.[13] Thus, it seems that the CLI does not help forecast IP growth in a real-time setting. This result is noteworthy because the statistical results using the most recently revised series, and even those that combine revised and real-time data, favor including the CLI in a forecasting model of IP. Hence, forecast evaluation tests using the most recently revised data series for the CLI and IP will generate inferences that suggest a positive contribution of the CLI to forecasts of IP, and these inferences will not hold up in real-time applications.

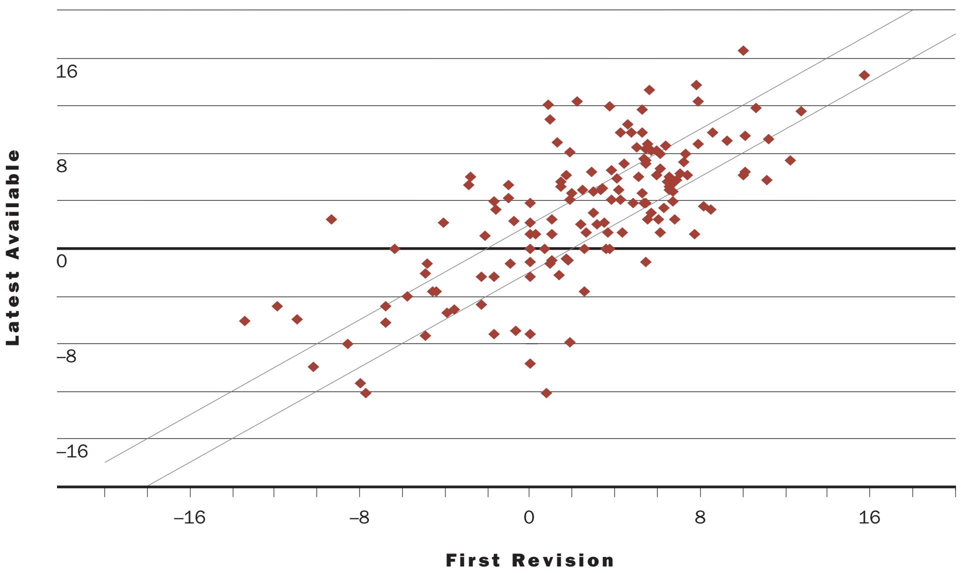
The differences between the real-time and pseudo real-time VAR forecasts are illustrated in Chart 5. This scatter diagram presents the one-month-ahead forecasts for the period 1986:01 to 1998:03, generated from the real-time VAR model and the VAR constructed using the most recent data vintage. There is considerably more variation between the forecasts due to data vintage than between the corresponding GDP forecasts; 46 out of 147 forecasts (31 percent) are different by more than 2 percentage points. The real-time one-month-ahead forecast has a correlation of 0.35 with the next month's revised estimate while the corresponding pseudo real-time forecasts have a correlation of 0.42 with the July 1998 vintage of estimates. Both are considerably lower than the corresponding correlation for the GDP forecasts.

Chart 6 is a scatter diagram of the latest available and the first-released monthly IP growth rates for the period from 1986:01 to 1998:03. Observe that 83 of the 147 observations (56 percent) differ by more than 2 percentage points, emphasizing that the latest available vintage of historical data over the forecast horizon differs substantially from the corresponding initial estimates. Comparing Chart 5 with Chart 6 makes it clear that the difference in the measured forecast accuracy arises primarily from the variation between the vintages of data being forecast rather than the forecasts themselves. The mean and standard deviation of the growth rates computed using the most recent data vintage are 2.95 and 6.04, whereas the same statistics for growth rates computed using the first revised estimates (and the second revision of the estimate for the preceding month) are 2.40 and 5.58, respectively. This pattern is the same one found for GDP data: the latest vintage of data is more variable and has a higher average than the less-revised estimates.

Conclusion

This article describes what historical real-time forecast evaluation should look like and how it is conceptually different from what is referred to here as a pseudo real-time forecast evaluation. The results suggest

CHART 6 IP Growth Estimates, 1986:01–1998:03



Source: Board of Governors of the Federal Reserve System

that using real-time vintages of data is a basic ingredient for generating valid out-of-sample forecast evaluations.

The practical question is whether a failure to use real-time data sets leads to inferences different from those made using only the latest available vintage of data. In principle, the specification and estimation of the forecasting model may differ due to the choice of the vintage of the data set, as may the evaluation of the model's forecasts. To shed some light on the practical importance of the issue, the article examines the ex ante forecast performance of two separate vector autoregressive (VAR) models. The discussion of both examples examines whether the CLI helps forecast measures of economic activity—real GDP growth and IP growth, respectively.

For real GDP, the results indicate (1) that the use of the latest vintage of historical time series on the CLI and real GDP does not cause the fitted VAR's forecasts to be much different from those of a VAR fitted using the actual historical data available at the time the forecasts were made and (2) that the choice of vintage of the real GDP data does alter the measured forecast accuracy of the VAR model but does not change the model's ranking. Relative to an autoregressive model for real GDP, knowing the value of the CLI within the current quarter leads to more accurate forecasts of GDP growth over each of the next two quarters.

For IP, the findings show (1) that using the latest vintage of the CLI and IP data does not cause the fitted VAR's forecasts to be much different from those from a VAR fitted using the actual historical data available at the time the forecasts were made and (2) that the pseudo real-time forecast results, when evaluated against the latest available data, suggest that the CLI can help predict the growth in IP. Using real-time data on the CLI, combined with latest available data vintage for IP (comparable to Diebold and Rudebusch 1991), generates weaker but still supportive results of the predictive power of the CLI for IP when compared with the latest available vintage of data. When the real-time forecasts are evaluated against the next available and nearby IP estimates, the results suggest that a VAR actually produces less accurate forecasts than does a simple AR model of IP. For the models considered here, failure to use real-time data in constructing and evaluating the forecasts was not too serious a problem for real GDP, but it produced an apparently misleading inference for the IP model.

Differences in the assessment of forecast performance arise primarily from the choice of series to evaluate against. The revisions to IP vintages of historical data are of a larger magnitude and are more extensive than those made to real GDP data. However, in both cases the differences among the data revisions are much larger than the differences among the forecasts. This insight reflects the fact that the models do not generate forecasts that vary greatly across vintages of historical data.

[13] The ratio of the VAR to AR RMSE is always slightly greater than 1 when the real-time IP forecasts are compared with the initial IP growth estimate.

In the present case this finding affects the magnitude of the measures of accuracy for a given model as well as across models. Of course, the models used here involve only two series. It remains to be seen whether these empirical results generalize to more realistic forecasting models that typically involve a larger number of variables. Still, the article highlights the finding that the accuracy of results clearly depends upon the target series chosen as a forecast accuracy criterion.

REFERENCES

Beckman, Barry A. 1997. "Reflections on BEA's Experience with Leading Economic Indicators." Bureau of Economic Analysis, unpublished manuscript.

The Conference Board, Inc. 1997. "Using the Individual Leading Indicators to Predict Growth." *Business Cycle Indicators* 2 (April): 3–4.

———. 1998. "Leading Indicators and the Prospects for Growth in 1998." *Business Cycle Indicators* 3 (May): 3–4.

Diebold, Francis X., and Glenn Rudebusch. 1991. "Forecasting Output with the Composite Leasing Index: A Real-Time Analysis." *Journal of the American Statistical Association* 86:603–10.

Fair, Ray C., and Robert J. Shiller. 1990. "Comparing Information in Forecasts from Econometric Models." *American Economic Review* 80, no. 3:375–89.

Fleming, Martin, John S. Jordan, and Kathleen M. Lang. 1996. "The Impact of Measurement Error in the U.S. National Income and Product Accounts on Forecasts of GNP and Its Components." *Journal of Economic and Social Measurement* 22:89–102.

Frumkin, Norman. 1994. *Guide to Economic Indicators.* Armonk, N.Y.: M.E. Sharpe.

Hamilton, James D., and Gabriel Perez-Quiros. 1996. "What Do the Leading Indicators Lead?" *Journal of Business* 69:27–49.

Landefeld, J. Steven, and Robert P. Parker. 1997. "BEA's Chain Indexes, Time Series, and Measures of Long-Term Economic Growth." *Survey of Current Business* 77 (May): 58–68.

Makridakis, Spyros, Chris Chatfield, Michele Hibon, Michael Lawrence, Terence Mills, Keith Ord, and LeRoy F. Simmons. 1993. "The M2-Competition: A Real-Time Judgmentally Based Forecasting Study." *International Journal of Forecasting* 9:5–22.

Mariano, Roberto S., and Hisashi Tanizaki. 1995. "Prediction of Final Data with Use of Preliminary and/or Revised Data." *Journal of Forecasting* 14:351–80.

McNees, Stephen K. 1988. "How Accurate Are Macroeconomic Forecasts?" Federal Reserve Bank of Boston *New England Economic Review* (July/August): 15–36.

———. 1992. "How Large Are Economic Forecast Errors?" Federal Reserve Bank of Boston *New England Economic Review* (July/August): 25–42.

———. 1995. "An Assessment of the 'Official' Economic Forecasts." Federal Reserve Bank of Boston *New England Economic Review* (July/August): 13–23.

Moore, Geoffrey H. 1990. *Leading Indicators for the 1990s.* Homewood, Ill.: Dow Jones-Irwin.

Orphanides, Athanasios. 1997. "Monetary Policy Rules Based on Real-Time Data." Board of Governors of the Federal Reserve System, unpublished manuscript.

Rogers, Ralph Mark. 1998. *Handbook of Key Economic Indicators.* New York: McGraw-Hill.

Staiger, Douglas, James H. Stock, and Mark W. Watson. 1997. "The NAIRU, Unemployment, and Monetary Policy." *Journal of Economic Perspectives* 11 (Winter): 33–50.

Stock, James H., and Mark W. Watson. 1998. "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." National Bureau of Economic Research Working Paper No. 6607, June.

Swanson, Norman. 1996. "Forecasting Using First-Available versus Fully Revised Economic Time-Series Data." *Studies in Nonlinear Dynamics and Economics* 1, no. 1:47–64.

Swanson, Norman, and Halbert White. 1997a. "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks." *Review of Economics and Statistics* 79:540–50.

———. 1997b. "Forecasting Economic Time-Series Using Flexible versus Fixed Specification and Linear versus Nonlinear Econometric Models." *International Journal of Forecasting* 13:439–61.

Young, Allan H. 1993. "Reliability and Accuracy of the Quarterly Estimates of GDP." *Survey of Current Business* 73 (October): 29–43.

Valuation Models for Default-Risky Securities: An Overview

SAIKAT NANDI

The author is a senior economist in the financial section of the Atlanta Fed's research department. He thanks Lucy Ackert, Jerry Dwyer, Will Roberds, and Larry Wall for helpful comments.

VALUING FINANCIAL SECURITIES OFTEN ASSUMES THAT THE CONTRACTUAL OBLIGATIONS OF THE SECURITY ARE GOING TO BE HONORED. HOWEVER, FREQUENTLY A PARTY TO A CONTRACT WILL DEFAULT ON ITS OBLIGATIONS. AN ISSUER OF A CORPORATE BOND MAY BE UNABLE TO MAKE ITS PROMISED COUPON AND PRINCIPAL PAYMENTS, AND A PARTY TO A DERIVATIVES CONTRACT SUCH AS AN INTEREST RATE SWAP MAY DEFAULT ON THE PERIODIC PAYMENTS OF THE SWAP CONTRACT.

Because the contractual features of defaultable securities are usually complex and it may be difficult to find comparable securities for which to observe prices, valuation based on simple rules of thumb is often infeasible. For example, one may have to value an interest rate swap subject to termination if one of the parties has its credit downgraded, and there are no comparable swaps that one can look up for a reference price (see the glossary for a definition of a swap and of other terms used throughout this discussion). Hence it becomes necessary to resort to formal models that can value a defaultable security on the basis of expected future cash flows, taking into account the contractual features of the defaultable security and the uncertainties surrounding the future cash flows. Many financial institutions hold large amounts of default-risky securities of various degrees of complexity in their portfolios, and it is important that these institutions have a reliable estimate of the resulting credit exposure. Estimating the credit exposure often involves knowing the possible values of the defaultable security at various times in the future. Therefore, understanding the different valuation models of default-risky securities and the strengths and drawbacks of various modeling approaches (see Table 1) is also important for implementing prudent risk-management policies to manage credit exposures.

This article discusses some of the models for valuing financial instruments that are subject to default risk, the implications of these models, and the difficulties that might be encountered in implementing them. The focus is on the valuation of default-risky bonds and swaps, although the general principles of valuation can be applied to other related instruments.[1] The first section explains the classic Merton (1974) model for valuing a default-risky bond. Subsequently, the text discusses some of the more recent models for valuing such bonds. These models differ in terms of how predictable the degree of default is and whether the firm's value is needed as an input in the valuation formula. Next, some of the valuation models for default-risky swaps are considered, including ones in which both parties to a swap can default. The discussion concludes with a review of the strengths and drawbacks of different valuation models and some thoughts for future research.

Merton's Model of Default-Risky Debt

One of the first models for valuing defaultable bonds was developed by Robert Merton (1974) using the principles of option pricing developed by Black and Scholes (1973) and Merton (1973). To understand the model, consider a firm that has equity and a single zero-coupon bond in its capital structure. Let the face value of the bond be $1,000, so that $1,000 has been promised to the bondholder at a future date when the bond matures. If the firm's value (the value of its assets) when the bond matures is less than $1,000, the firm cannot completely pay the bondholder even by liquidating all its assets. As a result, the firm will default on the debt.

How is a default-risky bond valued, given the knowledge that the firm may default at the maturity of the debt and that the claims of the bondholders are senior to those of the equityholders? Note that the payment to the bondholder at the maturity of the debt is the smaller of two quantities: the face value of the bond or the market value of the firm. If the firm's market value at maturity is greater than the face value of the bond, then the bondholder gets back the face value of the bond. However, if the value of the firm is less than the face value of the bond, the equityholders get nothing and the bondholder gets back the market value of the firm. The payoff from the default-risky bond at maturity resembles the payoff from an option, where the underlying asset is the value of the firm. Specifically, the payoff to the bondholder at maturity is the face value of the bond minus a put option on the firm's value with an exercise price equal to the face value of the bond.[2]

Using insights from option pricing theory, Merton (1974) derived an explicit valuation formula for default-risky bonds that default at maturity. The valuation formula requires knowing the following inputs: the value of the firm, the face amount of the debt, the volatility of the firm's value, the yield on a default-free bond that matures at the same time, and the time to maturity of the bond.

The yield on a defaultable bond should exceed the yield on an otherwise identical default-free bond because risk-averse investors must be compensated for the default risk in the form of additional yield or they will not hold a defaultable bond. The differential yield between a defaultable and a default-free bond (identical along all other dimensions) is known as the credit spread of the defaultable bond. In Merton's model, the credit spread increases as the leverage of the firm rises. The firm's leverage is measured by the ratio of the present value of the face amount of debt (discounted by the risk-free rate) to the value of the firm. This increase in the credit spread is natural because increased leverage heightens the probability that the firm may default. Higher default probability is reflected in an increase in the credit spread. Similarly, a rise in the volatility of the firm's value increases the probability that the firm may default, thus expanding the credit spread.

An important aspect of any model for valuing defaultable bonds is the term structure of credit spreads. As the term structure of interest rates in the Treasury market shows how the yields of zero-coupon bonds vary with their maturities, the term structure of credit spreads depicts the relationship between the credit spreads of defaultable bonds and their respective maturities. The term structure of credit spreads not only is important from a valuation perspective but also has some relevance for banking regulations related to allocating capital for credit-risky instruments. At the time of this writing, such regulations do not always recognize the term structure of credit spreads in defaultable bonds and other instruments. In other words, for allocating capital to cover potential defaults and credit downgrades, a one-year default-risky bond is treated the same as a ten-year default-risky bond although the two bonds may have quite different default and downgrade probabilities.

Many financial institutions hold large amounts of default-risky securities of various degrees of complexity in their portfolios, and it is important that these institutions have a reliable estimate of the resulting credit exposure.

In Merton's model, the term structure of credit spreads depends on the current credit quality of the issuer. Credit quality is measured by the ratio of debt to the firm's value; higher levels of debt lower credit quality by increasing the probability of default. In particular, the term structure of credit spreads is upward-sloping for high-credit-quality issuers, downward-sloping for low-credit-quality issuers, and hump-shaped for intermediate-quality issuers. Why do these patterns hold true? The value of the default-risky bond depends on the probability of default, which in turn depends on the value of the firm. If a firm is currently enjoying high credit quality, the impact on the bond value of further improvement in the credit quality through further increases in the firm value is limited because the payoff from the bond at maturity is capped at its face value. On the other hand, the firm's

[1] *An example would be instruments that help hedge default/credit risk, such as credit derivatives.*

[2] *Payoff from the bond at maturity is* min(V, B) = B – max(B – V, 0), *where* min( ) *and* max( ) *give the minimum and maximum, respectively, of two quantities,* V *is the value of the firm, and* B *is the face value of the bond.*

Glossary

Absolute Priority: The strict seniority order in which claims to the firm's assets are paid in the event of bankruptcy/default.

Call Option: An option that gives its owner the right (but not the obligation) to buy the underlying asset at a fixed price (called the strike or the exercise). This right can be exercised at some fixed date in the future (European option) when the option matures or at any time until the option matures (American option).

Full Two-Way Payments Scheme: A swap settlement scheme (upon default) in which the party for whom the swap is an asset, even though the party may be the defaulter, has a claim equal to the current value of the swap from the counterparty.

LIBOR: An acronym for the London Interbank Offer Rate, the interest rate on dollar-denominated deposits outside the United States, deposited by one bank with another bank.

Poisson Process: A type of statistical process often used to describe the random arrival (through time) of customers in a queue, such as the arrival of telephone calls in a switchboard. The number of arrivals until a certain point of time has a Poisson distribution, and the interarrival times (that are statistically independent) are exponentially distributed. Thus, the Poisson process jumps by a certain amount at each arrival time.

Put Option: An option that gives its owner the right (but not the obligation) to sell the underlying asset at a fixed price (called the strike or the exercise). This right can be exercised at some fixed date in the future (European option) when the option matures or at any time until the option matures (American option).

Swap: A contract involving periodic exchange of payments between two parties according to some prespecified terms. In an interest rate swap, one party pays a fixed rate (called the swap rate) and receives a floating rate, usually tied to the LIBOR, on a notional principal.

Swap Rate: The fixed rate paid in a fixed-for-floating swap.

Term Structure of Credit Spreads: A relationship that shows how the differential yields between default-risky and default-free zero-coupon bonds that are otherwise similar vary with maturity.

Transition Matrix: A matrix in which a typical element corresponds to the probability of transition (of the debt of a firm) to a different credit rating over a period of time, given that the debt has a particular credit rating as of today.

credit quality may deteriorate with the passage of time, thus increasing the risk of default. In other words, the upside potential is limited and the downside risk is substantial as time elapses. As a result, credit spreads widen as maturity increases for high-quality bonds, resulting in an upward-sloping term structure of credit spreads. On the other hand, for a firm that currently has low credit quality, the downside risk is limited and the upside potential is substantial as time elapses. Therefore, one would expect credit spreads to decrease with maturity, yielding a downward-sloping term structure of credit spreads.

The empirical evidence on term structure of credit spreads and credit quality is mixed. Sarig and Warga (1989) and Fons (1994) document evidence supporting the relation between the term structure of credit spreads and credit quality mentioned above. However, Helwege and Turner (1997) examine bonds of different maturity issued by the same firm and find that the term structure of credit spreads of some low-quality firms is upward-sloping.

Despite its simplicity and intuitive appeal, Merton's model has many limitations. First, in the model the firm defaults only at maturity of the debt, a scenario that is at odds with reality. Second, for the model to be used in valuing default-risky debts of a firm with more than one class of debt in its capital structure, the priority/seniority structures of various debts have to be specified. Also, this framework assumes that the absolute-priority rules are actually adhered to upon default in that debts are paid off in the order of their seniority. However, empirical evidence in Franks and Torous (1989, 1994) and from other researchers indicates that the absolute-priority rule is often violated. Yet another problem with the Merton model is that the value of the firm, an input to the valuation formula, is very difficult to ascertain. Unlike the stock price in the Black-Scholes-Merton formula for valuing equity options, the current market value of a firm is not easily observable. One can argue that, ideally, the firm's market value is equal to the market value of its

TABLE 1
Strengths and Drawbacks of Various Models for Default-Risky Bonds and Swaps

Model	Advantages	Drawbacks
Merton (1974)	Simple to implement.	(a) Requires inputs related to the value of the firm. (b) Default occurs only at the maturity of the debt. (c) Information in the history of defaults and credit-rating changes cannot be used.
Longstaff and Schwartz (1995)	(a) Simple to implement. (b) Allows for stochastic term structure and correlation between defaults and interest rates.	(a) Requires inputs related to the value of the firm. (b) Information in the history of defaults and credit-rating changes cannot be used.
Jarrow, Lando, and Turnbull (1997)	(a) Simple to implement. (b) Can exactly match the existing prices of default-risky bonds to infer risk-neutral probabilities of defaults and credit-rating changes. (c) Uses the information in the history of defaults and credit-rating changes.	(a) Correlation not allowed between default probabilities and the level of interest rates. (b) Credit spreads change only when credit ratings change.
Lando (1998)	(a) Allows correlation between default probabilities and interest rates. (b) Allows many existing term-structure models to be easily embedded in the valuation framework.	(a) Historical probabilities of defaults and credit-rating changes are used under the assumption that the risk premiums due to defaults and rating changes are zero.
Duffie and Singleton (1997a, b)	(a) Allows correlation between default probabilities and the level of interest rates. (b) Recovery ratio can be random and depend on the predefault value of the security. (c) Any default-free term structure model can be accommodated, and existing valuation results for default-free term structure models can be readily used.	(a) Information in the history of credit-rating changes and defaults cannot be used.
Duffie and Huang (1996) (swaps)	(a) Has all the advantages of Duffie and Singleton. (b) Asymmetry in credit qualities is easily accommodated. (c) ISDA guidelines for settlement upon swap default can be incorporated.	(a) Information in the history of credit-rating changes and defaults cannot be used. (b) Can be computationally burdensome to implement for some swaps, such as cross-currency swaps, if domestic and foreign interest rates are taken to be random.

equity plus the market value of the debt; however, the market values of the various classes of debt a firm issues are difficult to ascertain because many corporate debts are very thinly traded, and those arranged with a private lender are not traded at all. Further, values of many firms may embody intangible brand-name components that may simply be unobservable except perhaps during mergers or acquisitions.

The drawbacks of the Merton (1974) model have led researchers to develop other models for valuing defaultable debt. One class of models relaxes some of the questionable assumptions of the Merton model but still requires parameters related to the firm's value as inputs in the valuation formula. This approach is often called the structural approach.[3] The other approach to valuing defaultable debt, called the reduced-form approach, does not require any parameters related to the value of the firm.

Shortcomings of structural models make it necessary to look at other classes of models for valuing defaultable securities that are not predicated on the value of the firm and that take into account credit-rating changes.

Structural Models

The structural models for valuing defaultable debt relax one of the unrealistic assumptions of the Merton model, namely, that default can occur only at the maturity of the debt when the firm's assets are no longer sufficient to cover debt obligations. Instead it is assumed that default may occur any time between the issuance and maturity of the debt and that default is triggered when the value of the firm (that is, its assets) reaches a lower threshold level.[4] It is also often assumed that debtholders, upon default, get back a fraction of the face value of the debt, sometimes called the recovery ratio, and that the recovery ratio is known a priori. While this assumption is somewhat unrealistic, it circumvents the difficult issue of explicitly specifying the seniority structure of debt, a drawback of the Merton (1974) model. Some authors—for example, Longstaff and Schwartz (1995)—argue that, by looking at the history of defaults and the recovery ratios for various classes of debt of comparable firms, one can form a reliable estimate of the recovery ratio.

The structural models may be difficult to implement in terms of actually valuing defaultable debt. This difficulty lies in the fact that some of the inputs to the valuation formula require parameters related to the value of the firm and, as noted before, the value of the firm is difficult to quantify. However, one may argue that the model's parameters could always be backed out or inferred from the market prices of some traded bonds of the firm in question.[5] For example, if a valuation model has n unknown parameters, then observing the prices of the firm's n bonds (say, of n different maturities) at a given time will yield the values of the n parameters such that the model prices of the defaultable bonds equal their market prices.[6]

Inferring the model parameters by matching the model's bond prices with the market bond prices is conceptually similar to the widely used practice of inferring a volatility (an unobservable quantity) from the observed market price of an option using the popular Black-Scholes (1973) model for equity options. In turn, the inferred/implied parameters can be used to price other bonds of the firm. However, there are some differences. First of all, one can easily relate to the volatility inferred from an option using the Black-Scholes model because the implied volatility is simply the expected average volatility that is supposed to prevail until the option expires. Observing past stock prices provides a good estimate of historical volatility, so it is not difficult to judge whether the inferred volatility estimate is reasonable. It is difficult, though, to ascertain whether the implied parameters related to the firm's value are reasonable. As a result, using these implied parameters to price other classes of bonds issued by the firm with a sufficient degree of confidence is not easy.

In addition, there have been cases in which an otherwise solvent firm has sought the protection of the bankruptcy court for previously unanticipated future legal liabilities (see Franks and Torous 1989). The relevance of structural models for these cases is questionable given that the defaults are not directly related to the current firm values but are "sudden surprises" (Duffie and Lando 1998). Another drawback of the structural models is that they cannot incorporate credit-rating changes that occur quite frequently for default-risky corporate debts. Many corporate bonds undergo credit downgrades by credit-rating agencies before they actually default, and bond prices react to these rating changes either in anticipation or when they occur. Thus, any valuation model should take into account the uncertainty associated with credit-rating changes as well as the uncertainty surrounding default. The shortcomings of structural models make it necessary to look at other classes of models for valuing defaultable securities that are not predicated on the value of the firm and that take into account credit-rating changes.

Reduced-Form Models

Unlike structural models, reduced-form models do not condition default explicitly on the value of the firm, and parameters related to the firm's value need not be estimated to implement the model. Also, reduced-form models fundamentally differ from typical structural models in the degree of predictability of the default. In fact, they are more general than the structural models because they can easily accommodate defaults that are sudden surprises.

A typical reduced-form model assumes that an exogenous random variable drives default and that the probability of default over any time interval is nonzero. Many reduced-form models further assume that the probability of default could vary through time, possibly with variations in the level of interest rates.[7] Actual default occurs when the random variable undergoes a discrete shift in its level. The time at which the discrete shift will occur cannot be foretold on the basis of information available today. In other words, the time at which default might occur is a random variable. However, even in a structural model such as that of Longstaff and Schwartz (1995), the default time is not known in advance because the value of the firm is a random variable (the exact time the value of the firm will touch the lower threshold that results in default is not necessarily predictable on the basis of current information). Yet there are technical conditions that make a crucial distinction between the properties of the default time in most structural models and those in reduced-form models. Roughly speaking, the default time in reduced-form models is much more unpredictable than in structural models, where the time of bankruptcy can be foretold just before it occurs so that, as Duffie and Lando have remarked, in most structural models "bankruptcy occurs not with a bang but with a whimper" (1998).[8]

It may appear that reduced-form models are somewhat ad hoc in that default is modeled only implicitly through the discrete shift of an exogenous variable. However, the value of a firm used in structural models can only be imprecisely observed, and one of the motivations for building reduced-form models is to circumvent using parameters related to an imprecisely observed quantity. Further, by allowing the default probability to vary through time and to depend on the level of interest rates, a reduced-form model can be made rich enough to accommodate many stylized features of defaults, as in Lando (1998) and Duffie and Singleton (1997a).

The following example shows how one can value a defaultable zero-coupon bond in a simple reduced-form model, specifically the model of Jarrow, Lando, and Turnbull (1997). Suppose that the value of a default-free zero-coupon bond that will mature at time T and pay one dollar at maturity is known at time t. Denote the value of this bond by $p(t, T)$. If $v_i(t, T)$ denotes the value of a defaultable zero-coupon bond of a firm that currently has credit rating i (for example, AAA) at time t, will mature at time T, and has a promised payoff of $1 at maturity, then Jarrow, Lando, and Turnbull show that

$$v_i(t, T) = p(t, T)\,[\phi + (1 - \phi)q_i(t, T)], \qquad (1)$$

where ϕ is the recovery ratio—the fraction of the face value ($1) that is recovered at time T after default—and $q_i(t, T)$ denotes the probability of a default occurring after T given that the debt has credit rating i as of time t. The valuation formula indicates that the higher the probability of default not occurring before maturity, the higher the value of the defaultable bond is and therefore the lower the credit spread is. To arrive at the valuation formula, Jarrow, Lando, and Turnbull (1997) assume that default is independent of the level of interest rates.

The assumption of independence is not critical to reduced-form modeling, however. Lando (1998) relaxes the independence assumption and extends Jarrow, Lando, and Turnbull's model so that the default probability can depend on the level of interest rates. Since interest rates vary through time, Lando does not restrict changes in credit spreads in terms of when they can occur. Lando's model is discussed in Box 1. (Box 1 also illustrates the valuation principle of Duffie and Singleton 1997a for defaultable debt, which is related to but somewhat different from the principles of Jarrow, Lando, and Turnbull 1997 and Lando 1998 in that credit-rating changes are not taken into account.)

At this stage, it might seem that valuing a defaultable bond is relatively straightforward in a typical reduced-form model: one needs to know the recovery ratio, the price of an identical maturity default-free bond, and the probability of default. Recovery ratios can be estimated by looking at past recovery ratios of similar bonds, and

[3] *Technically, Merton's model is also a structural model.*

[4] *One of the earlier studies based on this framework is Black and Cox (1976). More recent ones include Longstaff and Schwartz (1995) and Nielsen, Saa-Requejo, and Santa-Clara (1993).*

[5] *The same argument can be made about the Merton model.*

[6] *The parameters are found by solving a system of* n *simultaneous equations (that are generally nonlinear) in* n *unknowns using a numerical routine for finding out the roots of equations.*

[7] *It is assumed that defaults are governed by what are known as Poisson processes. In a model without credit-rating changes, default occurs when the Poisson counter changes for the first time, for example, from zero to one. The intensity of the Poisson process, which determines the probability of default over a small time interval, could depend on the level of interest rates.*

[8] *Most structural models based on the value of a firm assume that, as a mathematical function, the value of the firm is continuous (in time). As a result, the time of the bankruptcy can actually be predicted just before it happens and hence there are no sudden surprises (Duffie and Lando 1998). One can, however, model the value of a firm as a function consisting of a continuous part and a jump part—the so-called jump diffusion process, as in Zhou (1997)—so that the default can be a sudden surprise.*

the price of an identical maturity default-free bond (that is, a Treasury bond/bill) can be observed in the market or, alternatively, can be estimated from observed prices of zero-coupon bonds of various maturities.[9] The problem lies in determining the default probability—not the historical probability of default but, roughly speaking, an artificial probability called the risk-neutral probability of default. The risk-neutral probability of default can be thought of as an adjusted probability that takes into account investors' compensation for default risk (see Box 2 for the principles behind risk-neutral valuation).

How does one determine the risk premium attributed to default for various bonds and thereby estimate the risk-neutral probabilities of default needed for valuation? All that can be estimated are the historical probabilities of credit-rating changes and defaults from historical data, probabilities that could be computed using data available from credit-rating agencies such as Moody's and Standard and Poor's. Of course, as with structural models, one can always infer the model parameters from a cross section of bond prices observed in the market and then use these inferred parameters to price similar bonds. However, as discussed earlier, a default-risky bond can undergo several credit-rating changes before it actually defaults, and the changes are priced by the market. The information from past credit-rating changes and defaults is useful and should not simply be ignored. Furthermore, there are instruments such as credit-sensitive notes and certain types of swaps (with credit triggers) whose payoffs explicitly depend on particular credit events occurring. A more complete framework for modeling defaultable instruments, therefore, has to take into account credit-rating changes.

Jarrow, Lando, and Turnbull (1997) have shown how one can use the probabilities of credit-rating changes and defaults computed from historical data to price defaultable bonds. The migration/transition probabilities of credit-rating changes and defaults can be estimated from the

BOX 1

Reduced-Form Models of Duffie and Singleton (1997a) and Lando (1998)

It is well known that the price (at time t) of a default-free zero-coupon bond (which is traded continuously) that matures at time T (and pays out \$$Y$) is given by

$$E_t[\exp(-\int_t^T r_u du)Y],$$

where r_t is the interest rate on a loan that matures at the next instant and $E_t(\ )$ is the conditional expectation under the risk-neutral distribution of r_t. Duffie and Singleton (1997a) assume that defaults are governed by a Poisson process with intensity λ_t. The probability of default over a small time interval could be thought of as proportional to λ_t. In particular, the probability of default could be time-varying and depend on the level of interest rates. Denote λ_u^* to be the intensity corresponding to the risk-neutral default probability. Under the assumption that upon default one recovers a fraction—say, ϕ—of the predefault value of the bond, Duffie and Singleton show that the price of a defaultable zero-coupon bond that matures at time T (and is supposed to pay \$$Y$) is

$$E_t\Big(\exp\{-\int_t^T [r_u+\lambda_u^*(1-\phi)]du\}Y\Big).$$

Thus, the valuation formula for risky bonds is the same as for risk-free bonds, except for an adjusted short rate given by $r_t + \lambda_t^*(1-\phi)$ that is used for discounting purposes. In other words, with the possibility of default, a default risk premium has to be added to the interest rate for discounting. The required risk premium increases with an increase in the probability (risk-neutral) of default and increases in the amount of value lost upon default.

Lando's (1998) model also gives rise to a valuation formula somewhat similar to that of Duffie and Singleton (1997a) under a similar set of assumptions. Additionally, the model can incorporate credit-rating changes, specifically the probabilities of a firm transitioning to a different credit rating or default as computed from historical data. However, the risk-neutral probabilities of default and credit-rating changes are taken to be the same as the ones computed from historical data, a somewhat questionable assumption.

information available in credit reports, such as *Moody's Special Report* (1992) or Standard and Poor's *Creditreview* (1993). Although the historical transition probabilities may not reflect the most current information because the probabilities may not be updated frequently, the information is useful. Jarrow, Lando, and Turnbull (1997) also show how one can use historical probabilities to estimate the risk-neutral probabilities (of defaults and credit-rating changes) at various future dates from a cross section of defaultable and default-free bonds of various maturities by exactly matching the observed market prices of defaultable and default-free bonds. These risk-neutral default probabilities can, in turn, be used to value other defaultable financial instruments of the firm, such as a swap. (See the appendix for Jarrow, Lando and Turnbull's procedure.)

There are other reduced-form models, such as those of Jarrow and Turnbull (1995) and Madan and Unal (1995). However, both of these models assume that defaults are not correlated with interest rates and cannot incorporate information in credit-rating changes.

Shortcomings of the Models

The structural and reduced-form models for valuing default-risky debt that have been discussed so far cannot readily incorporate financial restructuring that often occurs upon default, such as renegotiating of the terms of the debt contract by extending the maturity or lowering/postponing the promised payments, exchanging the debt for other forms of securities, or some combination of the above. Similarly, the institutional features of a reorganization under court supervision, such as Chapter 11 bankruptcy, cannot be incorporated in any of these valuation models without making them intractable. Debt restructurings anticipated by the market will be priced into the value of a defaultable bond in ways that none of these models captures. Another issue relevant to pricing default-risky securities is that, unlike Treasury securities, many defaultable securities are thinly traded. A liquidity premium may therefore be incorporated into the prices of these securities, another factor that is outside the realm of the valuation models discussed.

Empirical evidence regarding the validity of these models is rather limited. Duffee (1996) finds that reduced-form models based on the Duffie and Singleton (1997a) framework have difficulty explaining the observed term structure of credit spreads across firms of different credit qualities. Such problems could arise from incorrect statistical specifications of default probabilities and interest rates or from the model's inability to incorporate some of the features of default/bankruptcy mentioned in the previous paragraph.

Default-Risky Swaps

So far the discussion has described the different valuation models for default-risky bonds. Another class of instruments that is very heavily traded and could be subject to defaults is swaps—obligations created when parties swap streams of payments. Swaps come in various forms. The focus here is on the plain vanilla interest rate swap, which involves swapping interest payments at a fixed rate for payments at a floating rate, and simple currency swaps, in which principal and interest payments in one currency are swapped for payments in another currency. However, the general valuation principles discussed can be adapted to other types of swaps. Only reduced-form models for valuing default-risky swaps are discussed because models based on the value of the firm are usually difficult to implement.[10]

When a swap is initiated, by definition it has a market value of zero for both parties. However, with the passage of time, as interest rates and exchange rates change, a swap can become an asset to one party and a liability to the other. As a result, the expected loss profile from a swap is somewhat different from that of a straight loan such as a bond. For example, if default by the counterparty occurs when the swap is an asset (that is, has positive value) to a party, then the default represents a real loss as opposed to its occurring when the swap is a liability (has negative value).

A defaultable swap has option-like features embedded in it. In fact, valuation for defaultable swaps based on options theory has been advocated by Bollier and Sorensen (1994). This approach is somewhat ad hoc because the value of a defaultable swap is not derived using the fundamental principles of valuation, namely, by calculating the expectation (under the risk-neutral distribution) of the discounted future dividends/cash flows, including the cash flows that would accrue upon default. Instead, the value of a defaultable swap is arrived at by making some adjustments to the default-free value. Nevertheless, the approach is intuitive as it clearly illustrates the option-like features embedded in a defaultable swap.

To understand these adjustments, consider a swap in which only one of the parties—say, party B—can default at the next date while party A is default-free. Also assume that A does not recover anything upon default. If B defaults, then A's loss is either the current value of the swap (if the swap is an asset to A) or zero (if the swap is

9. See Bliss (1997) and Waggoner (1997) for various methods of computing the price of a zero-coupon bond of a given maturity from the observed prices of various Treasury bills, notes, and bonds.

10. For valuation of defaultable interest-rate swaps that is predicated on the value of the firm, see Cooper and Mello (1991) and Abken (1993).

Risk-Neutral Valuation

This box illustrates the basic principles behind risk-neutral valuation, which is often used to value risky assets such as derivatives and bonds. Consider a hypothetical economy with two assets, one risky and the other risk-free. Suppose the risky asset will pay \$3 and \$6, respectively, in the two states of nature, one of which will be realized tomorrow; the risk-free asset will pay \$1 irrespective of which state occurs. To find out the current values of these two assets, one needs to know how much the future dollars in the two states are worth as of today. Suppose one dollar in state 1 is worth p_1 and one dollar in state 2 is worth p_2 as of today. In fact, we may view the current price of an elementary security that pays off \$1 in state 1 (and nothing in state 2) as p_1 and the current price of an elementary security that pays off \$1 in state 2 (and nothing in state 1) as p_2. These elementary securities are often known as Arrow-Debreu securities, named after two famous economists, Kenneth Arrow and Gerard Debreu. In general, p_1 and p_2 may be different. If an additional dollar in state 2 is more valuable to the investor than in state 1, then the security that pays off in state 2 will be worth more to the investor than the security that pays off in state 1. The reason an additional dollar may be more valuable in state 2 than in state 1 could be a general market downturn in state 2.

A well-functioning economy should not admit free lunches or arbitrage in that there should not exist any portfolio of the two assets that does not cost anything today but pays off a positive quantity with certainty tomorrow.[1] It turns out that in the absence of arbitrage, a set of simple equations links p_1, p_2, and the future payoffs of the two assets to the present values of these assets. In general, one does not know p_1 and p_2. However, given the current values of the two assets, p_1 and p_2 can be inferred and, in turn, can be used to price a derivative asset such as an option on the risky asset. Assume that the current prices of the risky and risk-free asset are \$4.2 and \$0.9, respectively. Given the payoff structure, absence of arbitrage implies that the following system of equations has to hold:

$$3p_1 + 6p_2 = 4.2$$

$$p_1 + p_2 = 0.9.$$

One can solve the two equations to find $p_1 = 0.4$ and $p_2 = 0.5$. Thus, the market price of \$1 received in state 1 is 40 cents and the market price of \$1 received in state 2 is 50 cents. Also note that the above system of equations basically says that the two securities are linear combinations of the elementary Arrow-Debreu securities and that the prices of these Arrow-Debreu securities are positive. For example, the risky asset can be viewed as comprising three Arrow-Debreu securities that pay off in state 1 and six Arrow-Debreu securities that pay off in state 2.[2] Thus the pricing rule is linear and positive. A different valuation equation can be constructed from an algebraic rearrangement of the above set of equations to yield what is known as the risk-neutral valuation relationship. Under the risk-neutral valuation, the actual probabilities are adjusted so that the mean return on every asset (risky and risk-free) becomes the risk-free rate. The above system of equations can be rewritten as

$$(p_1 + p_2)[3p_1/(p_1 + p_2) + 6p_2/(p_1 + p_2)] = 4.2$$

$$(p_1 + p_2)[p_1/(p_1 + p_2) + p_2/(p_1 + p_2)] = 0.9.$$

Define two variables, q_1 and q_2, such that $q_1 = p_1/(p_1 + p_2)$ and $q_2 = p_2/(p_1 + p_2)$. These variables can be viewed as probabilities because each of them is nonnegative and $q_1 + q_2 = 1$. Also note that the gross return of the risk-free asset is $R = 1/(p_1 + p_2)$ because the current price of the risk-free security is $p_1 + p_2$ and it pays \$1 with certainty tomorrow. In terms of q_1, q_2, and R, the above system of equations can be written as

$$(1/R)(3q_1 + 6q_2) = 4.2$$

$$(1/R)(q_1 + q_2) = 0.9.$$

The equations above indicate that as of today (time t), the price of a security is $V_t = (1/R)E^q(V_{t+1})$, where $E^q(\ )$ denotes the statistical expectation over the probabilities q_1 and q_2, and V_{t+1} is tomorrow's (that is, $t + 1$) payoff from the asset. This expression is known as valuing the asset under risk-neutral probabilities. Note that under the artificial probabilities, q_1 and q_2, the expected price of the asset in the future is the current price multiplied by the risk-free return, R. In other words, the expected/mean return on every asset (risky and risk-free) is the risk-free rate. If investors are risk-neutral, the expected appreciation on every asset has to equal the risk-free rate. Hence, q_1 and q_2 are called the risk-neutral probabilities and are, in general, different from the actual probabilities of the occurrences of the two states. Calculating the value of a risky asset as being the discounted expectation under the risk-neutral probabilities/distribution is the very fundamental principle of valuation and is valid even in more general settings with multiple states and multiple trading periods.

One can value all the derivative assets, such as call and put options, using either q_1 and q_2 or p_1 and p_2. Often it is the case that valuing a primary asset, such as a bond, or a derivative asset, such as an option, is much easier (algebraically or computationally) using the risk-neutral probabilities. Researchers and practitioners therefore often use the risk-neutral approach to value derivative assets and bonds, including defaultable assets. However, the actual probabilities of default (that can be computed from historical data) have to be adjusted to arrive at the risk-neutral default probabilities used for valuation.

1. Also, there should not exist any portfolio of the two securities such that an investor (who owns the portfolio) gets a net cash inflow as of today and does not have to pay anything tomorrow.

2. The prices of Arrow-Debreu securities are unique because there are two possible states of nature and two securities to span these two states.

a liability to A). Thus the exposure of A to B's default is the maximum of the following two quantities: the value of the swap to A or zero. In other words, the exposure of A to B's default resembles the payoff from a European call option on the value of the swap with a strike price of zero.[11] Party A has implicitly written an option to B and should be compensated for the same. The compensation is equal to the probability that B will default at the next date times the value of the above option. Since default can also take place at other future dates, this add-on option approach suggests that the amortized value of these compensations over the remaining life of a swap should be taken into account to arrive at the default-adjusted value of the swap to A.

In general, the value of the option will depend on the current shape of the term structure. For example, if the swap in question is an interest rate swap and A is the fixed rate payer, then in an upward-sloping term structure environment, the option has value to A. It has value because the floating rate payments, and hence the value of the swap to A, are expected to increase. Therefore, the possibility of default by B will result mostly in a lower fixed rate (the swap rate) being paid by A. If A can also default, then, following the same arguments as before, it is easy to see that A has bought an option from B and written another option to B. If the valuation allows for default possibilities at various other dates during the life of the swap, then, as for swaps in which only one party can default, each of these options has to be valued separately to get the default-adjusted value of the swap or, equivalently, the swap rate.

Although intuitive, the add-on option approach does not explicitly take into account some of the settlement issues upon default per the guidelines of the International Swaps and Derivatives Association (ISDA) that are usually part of a swap agreement. For example, in a full two-way payments scheme between two parties who are engaged in only one swap, the party for whom the swap is an asset, even though it may be the defaulter, has a claim equal to the current value of the swap from the counterparty. Often it could be the case that only a fraction of the swap value can be recovered. In particular, Duffie and Huang (1996) show that if parties to a swap are not symmetric in terms of their credit characteristics, settlement issues do matter, and accounting for them explicitly could yield swap values quite different from those achieved by using the add-on option approach. Before turning to Duffie and Huang's model, the discussion examines a more complete model for pricing defaultable swaps in which the parties are symmetric in terms of their credit characteristics.

Swaps with Symmetric Credit Risk

As Box 1 illustrates, computing the price of a defaultable bond is similar to computing a default-free bond price. The difference lies in the factor used for discounting. In pricing a default-free bond, cash flows from the bond are discounted at the risk-free interest rates whereas for a defaultable bond the relevant discounting factor is the sum of two terms: the risk-free interest rate and the product of two factors—one proportional to the risk-neutral default probability and another

[11] *Formally, the exposure of A is* max(V_t, 0)*, where* V_t *is the value of the swap at time* t *to A without default.*

that equals the fraction of predefault value of the bond lost upon default. In other words, discounting is done with respect to an interest rate adjusted for default probability and the fraction of value lost upon default. Since valuing a swap amounts to computing the expected value of the stream of cash flows at an appropriate discount rate, it is reasonable to expect that the above methodology for valuing defaultable bonds can be extended to value defaultable swaps.

Note that the price of a default-risky instrument differs from an equivalent (in terms of promised cash flows and maturity) default-free instrument through the extra term used in discounting that captures the probability of default and loss in value upon default. Thus, this extra term could be thought of as representing the credit quality of a firm/institution that can default on its obligations. In fact, Duffie and Singleton (1997b) show that in a swap involving both parties of symmetric credit quality (that is, the extra term is the same for both parties), valuation of defaultable swaps, taking into account the settlement payments upon default, can be very similar to that of defaultable bonds discussed in Box 1. In other words, the value of the swap is simply the statistical expectation (under the risk-neutral distribution) of the future cash flows/dividends from the swap, discounted by an interest rate adjusted to reflect the risk-neutral default probability and the fraction of the predefault swap value lost upon default. Duffie and Singleton's model does implicitly take into account the previously discussed options embedded in a defaultable swap. However, it is more complete than the add-on option approach in that it values the swap using the fundamental valuation principles, instead of some adjustments to the default-free value of the swap. It is to be noted that the model requires risk-neutral probability of default in the valuation formula and therefore cannot readily be implemented solely on the basis of historical data. Instead, the extra term used in discounting (and other parameters) has to be inferred from the observed market prices of swaps. Also, because the model does not take into account credit-rating changes, the model cannot readily price swaps with embedded credit triggers that result in termination of the swap if a particular credit event occurs.

By allowing the default probability to vary through time and to depend on the level of interest rates, a reduced-form model can accommodate many stylized features of defaults.

To arrive at a swap value in Duffie and Singleton (1997b) for an interest rate swap, one needs to assume the type of statistical process that the adjusted interest rate follows through time. In their empirical work, the authors assume that the default-adjusted interest rate follows what is known as the square root process (see Cox, Ingersoll, and Ross 1985).[12] The model also permits other statistical specifications for the default-adjusted interest rate. Duffie and Singleton find that their model fits the interest rate swap data reasonably well. However, they do not describe how well their estimated model predicts future swap rates. Although a particular model may fit the data well during the course of statistical estimation in that the pricing errors could be quite small, it is not necessarily true that the model would be equally good at predicting future market values of the securities based on the parameters estimated.[13]

Swaps with Asymmetric Credit Risk

Duffie and Huang (1996) develop a model for valuing defaultable swaps in which the credit qualities of a swap's two parties can be asymmetric. The model also incorporates features of settlement payments upon default per some of the ISDA guidelines.

Consider a plain vanilla interest rate swap in which A is the floating rate payer (paying the LIBOR), B is the fixed rate payer, and settlements upon default will allow for full two-way payments. For A, predefault dividends are simply the net payments (fixed minus floating) that would accrue according to the swap agreement whereas the postdefault dividend is the payment made after default per the full two-way payments scheme. Thus, the postdefault dividend may be the value of the swap (or a fraction thereof) at the time of default or zero. The authors calculate the value of a defaultable swap to be the statistical expectation (under the risk-neutral distribution) of the discounted dividends that would be paid until default.[14] However, the discounting rate used is quite different from that of Duffie and Singleton because of the asymmetric credit quality.

According to Duffie and Huang's valuation procedure, if the swap's value is positive for a party, then it is the default characteristics of the other party that matter for arriving at the default-adjusted interest rate used for discounting. For example, if the value of the swap to A is positive, then the extra term added in the discounting factor is the product of a factor proportional to the risk-neutral default probability of B and a factor that equals the fraction of the swap value lost upon default through transacting with B. Since the swap's value can switch between positive and negative values for either of the parties, the discounting rate switches between the default characteristics of A and B. The switching discount rate makes the valuation model nonlinear in that multiplying the promised dividends/cash flows (to A) by a factor does not change the value of the swap to A by the same factor. This characteristic is in contrast to the valuation of a typical financial instrument without default in which, for example, doubling/halving the promised cash flows would result in the value of the instrument being doubled/halved.

One of Duffie and Huang's (1996) primary insights is that the degree of asymmetry in the default characteristics of the two parties in a plain vanilla interest rate swap (with no exchange of principals) is not very important in determining swap rates. For example, replacing the fixed rate payer with a riskier counterparty whose corporate bond yield is 100 basis points higher changes the swap rate (the fixed rate paid in the swap) by approximately a basis point. In contrast, the authors claim that the add-on option approach can substantially overstate the impact of asymmetry in credit qualities despite the netting of fixed and floating payments that in general make a swap much less credit sensitive than a straight loan, such as a bond. Consequently, the estimate of expected default losses in switching from a higher-credit-quality to a lower-credit-quality counterparty can be substantially higher under the add-on option approach, perhaps resulting in unnecessary capital being set aside to cover default losses during the course of managing credit risk. Duffie and Huang (1996) also show that in a currency swap involving fixed-for-fixed payments in which an exchange of principals takes place, the impact of the asymmetry in the credit risk is somewhat higher. The asymmetry matters more because, besides the periodic exchange of interest payments, principal payments are exchanged at the end of the swap, resulting in an additional exposure. The exposure increases if the volatility of the exchange rates is higher. A similar result for currency swaps has also been found by Hull and White (1995) but under more restrictive conditions.

Conclusion

Although valuation models for defaultable securities date back to Merton (1974) and researchers have improved considerably on the basic Merton framework, problems remain. Table 1 highlights some of the advantages and disadvantages of using the valuation models discussed in this article.

One class of models, sometimes referred to as structural models, requires the use of an imprecisely observed quantity (or quantities), such as a firm's value or variables related to it, in the valuation formula. In contrast, another class of models known as reduced-form models does not need firm-value-related variables and so holds more promise. It is often the case that the valuation formulas of reduced-form models are very similar to those used for valuing the corresponding default-free securities. The only difference is that the discounting factor is adjusted upward, taking into account the probability of default and the fraction of value lost upon default. Therefore, many of the existing results for valuing default-free securities such as default-free bonds can be readily extended to price default-risky securities. This advantage is significant as some of the models for valuing default-free securities are analytically and computationally tractable. Some of the reduced-form models can also incorporate the historical probabilities of credit-rating changes and defaults. These probabilities not only expand the information set used in valuation but also can be crucial for pricing instruments whose payoffs are explicitly linked to credit events, such as credit upgrades or downgrades.

However, because a limited amount of work has been done so far in validating the empirical efficacy of various reduced-form models, caution is warranted in using these models for pricing and hedging defaultable securities. Also, it is often the case that if one allows for realistic features, such as correlation between the interest rate level and default probabilities, historical probabilities of credit-rating changes and defaults can be used in a tractable fashion only under the questionable assumption that the risk premiums due to defaults and credit-rating changes are zero. Many of the institutional features of bankruptcy and defaults, such as renegotiation between the debtor and creditors and rescheduling of debts, cannot be readily incorporated in any of the valuation models discussed in this article as otherwise the models would be rendered intractable. It is hoped that the next generation of valuation models will be able to incorporate at least some institutional features and be able to use the historical probabilities of defaults and credit-rating changes without making unnecessarily strong assumptions.

[12] This type of interest rate has the advantage that negative interest rates are precluded.

[13] In technical jargon, good in-sample pricing by a model does not always lead to good out-of-sample pricing. This characteristic is especially true if a model is not parsimonious in the number of parameters that need to be estimated.

[14] Duffie and Huang (1996) show that in the presence of settlement payments, one needs to take into account a feedback effect from the value of the swap itself to the stream of dividends that needs to be valued to arrive at the swap value. The add-on option approach ignores this type of feedback effect.

APPENDIX

Estimating the Risk-Neutral Probabilities of Default and Credit-Rating Change

The purpose of this appendix is to show how one can estimate the risk-neutral probabilities of default and credit-rating changes from a cross section of default-free and defaultable bond prices of various maturities using the methodology of Jarrow, Lando, and Turnbull (1997). The authors suggest an algorithm such that one can exactly match the observed prices of default-free and defaultable bonds and at the same time use the historical probabilities of migration to other credit ratings, including default. Alternatively, if one assumes that the defaultable bond prices are observed with error, their procedure yields estimates of risk-neutral default probabilities by using the information in the history of credit-rating changes. The estimated risk-neutral default probabilities can then be used to price other defaultable securities of a firm, such as a plain vanilla swap the firm is party to or a swap with embedded credit triggers.

The risk-neutral probabilities of default or credit-rating changes are computed by multiplying the historical probabilities by a factor that can be interpreted as a default risk premium.[1] Table A presents a hypothetical transition matrix of credit-rating changes and defaults. There are three possible rating categories—AAA, BBB, and D. Assume that D represents default by a firm on its debt. The matrix has nine entries, and a typical entry shows the probability of moving from one rating category to another in one period. For example, the entry of 0.1 for row 2, column 3 indicates that the probability is 0.1 that a firm currently rated BBB will default in one period. Similarly, an entry of 0.06 for row 1, column 2 indicates that the probability of a currently AAA-rated debt becoming BBB-rated is 0.06 in the next time period. Note that all entries in the last row of the transition matrix are 0 except the last, which is 1. The probability of a debt migrating to a different credit-rating category once it enters default is 0, and therefore the probability of remaining in the default state once the debt enters default is 1.

Let $v_i(0, 1)$ and $p(0, 1)$ denote the prices of a defaultable and a default-free bond, respectively, at time 0 (that is, the current time); all bonds mature at time 1 and are supposed to pay $1 at maturity. Assume that the fraction of face value to be recovered upon default from the defaultable bond is ϕ. Given this information and the assumptions the authors make, two risk premiums have to be determined, namely, moving from AAA and BBB to other credit categories. With these risk premiums, the risk-neutral probabilities of credit-rating changes, including default, can be found by multiplying the historical probabilities and the risk premiums. Given this example, Jarrow, Lando, and Turnbull (1997) show that the risk premium attributed to a credit-rating category i, denoted by Π_i, is given by

TABLE A
Transition Matrix of Defaults and Rating Changes

	AAA	BBB	D
AAA	0.9	0.06	0.04
BBB	0.05	0.85	0.1
D	0	0	1

$$\Pi_i = [p(0, 1) - v_i(0, 1)]/[p(0, 1)(1 - \phi)q(\mathrm{i}, 3)],$$

where $q(\mathrm{i}, 3)$ is the entry in the row i and column 3 of the transition matrix and represents the probability of moving from rating category i to default. In the example, Π_i needs to be calculated only for the AAA and BBB firms. Once Π_i is computed the risk-neutral default probability for a firm currently rated AAA will be given as $\Pi_{\mathrm{AAA}} \times 0.04$ (the entry in row 1, column 3 of the transition matrix). Similarly, the risk-neutral default probability for a BBB-rated firm is $\Pi_{\mathrm{BBB}} \times 0.1$ (the entry in row 2, column 3 of the transition matrix). The risk-neutral probabilities calculated in the example are due to credit-rating changes and defaults occurring between time 0 and 1. For risk-neutral probabilities of credit-rating changes and defaults at other times in the future—say, between time 1 and time 2—one has to use defaultable and default-free bonds that mature at time 2. For details in calculating the risk-neutral probabilities at other times in the future, see Jarrow, Lando, and Turnbull (1997).

[1] *Jarrow, Lando, and Turnbull (1997) make the somewhat questionable assumption that the risk premium in moving from credit rating* i *to credit rating* j *is the same as moving from credit rating* i *to credit rating* k.

REFERENCES

ABKEN, PETER. 1993. "Valuation of Default-Risky Interest-Rate Swaps." *Advances in Futures and Options Research* 6:93–116.

BLACK, FISCHER, AND JOHN COX. 1976. "Valuing Corporate Securities: Some Effects of Bond Indenture Provisions." *Journal of Finance* 31:351–67.

BLACK, FISCHER, AND MYRON SCHOLES. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81:637–54.

BLISS, ROBERT. 1997. "Testing Term Structure Estimation Methods." *Advances in Futures and Options Research* 9:197–231.

BOLLIER, THIERRY F., AND ERIC H. SORENSEN. 1994. "Pricing Swap Default Risk." *Financial Analysts Journal* 50 (May/June): 23–33.

COOPER, IAN, AND ANTONIO MELLO. 1991. "The Default Risk of Swaps." *Journal of Finance* 46:597–620.

COX, JOHN, JONATHAN INGERSOLL, AND STEVEN ROSS. 1985. "A Theory of the Term Structure of Interest Rates." *Econometrica* 53:385–407.

DUFFEE, GREGORY. 1996. "Estimating the Price of Default Risk." Board of Governors of the Federal Reserve System Working Paper No. 96-29.

DUFFIE, DARRELL, AND MING HUANG. 1996. "Swap Rates and Credit Quality." *Journal of Finance* 51:921–49.

DUFFIE, DARRELL, AND DAVID LANDO. 1998. "Term Structure of Credit Spreads with Incomplete Accounting Information." Stanford University Working Paper.

DUFFIE, DARRELL, AND KENNETH SINGLETON. 1997a. "Modeling Term Structures of Defaultable Bonds." Stanford University Working Paper.

———. 1997b. "An Econometric Model of the Term Structure of Interest-Rate Swap Yields." *Journal of Finance* 52:1287–321.

FONS, JEROME S. 1994. "Using Default Rates to Model the Term Structure of Credit Risk." *Financial Analysts Journal* 50 (September/October): 25–32.

FRANKS, JULIAN, AND WALTER TOROUS. 1989. "An Empirical Investigation of U.S. Firms in Reorganization." *Journal of Finance* 44:747–69.

———. 1994. "A Comparison of Financial Recontracting in Distressed Exchanges and Chapter 11 Reorganizations." *Journal of Financial Economics* 35:349–70.

HELWEGE, JEAN, AND CHRISTOPHER TURNER. 1997. "The Slope of the Credit Yield Curve for Speculative-Grade Issuers." Federal Reserve Bank of New York Working Paper No. 97-25.

HULL, JOHN, AND ALAN WHITE. 1995. "The Impact of Default Risk on the Prices of Options and Other Derivative Securities." *Journal of Banking and Finance* 19:299–322.

JARROW, ROBERT, DAVID LANDO, AND STUART TURNBULL. 1997. "A Markov Model for the Term Structure of Credit Spreads." *Review of Financial Studies* 10:481–523.

JARROW, ROBERT, AND STUART TURNBULL. 1995. "Pricing Options on Financial Securities Subject to Default Risk." *Journal of Finance* 50:53–86.

LANDO, DAVID. 1998. "On Cox Processes and Credit Risky Securities." *Review of Derivatives Research*, forthcoming.

LONGSTAFF, FRANCIS, AND EDUARDO SCHWARTZ. 1995. "A Simple Approach to Valuing Risky Fixed and Floating Debt." *Journal of Finance* 50:789–819.

MADAN, DILIP, AND HALUK UNAL. 1995. "Pricing the Risks of Default." University of Maryland Working Paper.

MERTON, ROBERT. 1973. "The Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science* 4:141–83.

———. 1974. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance* 29:449–70.

MOODY'S INVESTORS SERVICE. 1992. "Corporate Bond Defaults and Default Rates." *Moody's Special Report*, December.

NIELSEN, LARS, JESUS SAA-REQUEJO, AND PEDRO SANTA-CLARA. 1993. "Default Risk and Interest Rate Risk: The Term Structure of Default Spreads." INSEAD Working Paper.

SARIG, ODEG, AND ARTHUR WARGA. 1989. "Some Empirical Estimates of the Risk Structure of Interest Rates." *Journal of Finance* 44:1351–60.

STANDARD AND POOR'S. 1993. "Corporate Default, Rating Transition Study Updated." *Creditreview* (supplement to *Creditweek*), January 25.

WAGGONER, DANIEL F. 1997. "Spline Methods for Extracting Interest Rate Curves from Coupon Bond Prices." Federal Reserve Bank of Atlanta Working Paper No. 97-10, November.

ZHOU, CHUNSHENG. 1997. "A Jump-Diffusion Approach to Modeling Credit Risk and Valuing Defaultable Securities." Board of Governors of the Federal Reserve System Working Paper No. 97-15.

Venture Capital Investment: Emerging Force in the Southeast

EDGAR PARKER AND PHILLIP TODD PARKER

Edgar Parker is an analyst in the regional section of the Atlanta Fed's research department. Phillip Todd Parker is a graduate student in the Johnson School of Management, Cornell University. They thank Tom Cunningham, Madeline Zavodny, and Steve Smith for helpful comments on earlier drafts.

VENTURE CAPITAL INVESTMENT THROUGHOUT THE UNITED STATES IN GENERAL AND THE SOUTHEAST IN PARTICULAR HAS GROWN DRAMATICALLY IN RECENT YEARS, BECOMING AN INTEGRAL PART OF OUR ECONOMY. IT HAS HELPED CREATE SUCH COMPANIES AS APPLE COMPUTER, INTEL, FEDERAL EXPRESS, DIGITAL EQUIPMENT, AND MICROSOFT (SAHLMAN 1990). PENSION FUNDS, BANK HOLDING COMPANIES, INSURANCE COMPANIES, INVESTMENT BANKS, AND NONFINANCIAL INSTITUTIONS ALL INVEST VENTURE CAPITAL IN ORDER TO PURSUE HIGH RETURNS AND DIVERSIFY INVESTMENT RISKS.

However, returns from venture capital investment have been mixed over the relatively short history of the industry. As more and more large institutional investors pour increasing amounts of assets into venture capital and as state and local governments seek to attract this capital and the industries it fosters, the potential benefits will grow, but not without raising public policy issues (Berlin 1998).

This article examines the history, structure, and evolution of the national venture capital industry. After providing that broad background, the authors focus on current development of venture capital in the Southeast and of states' promotion of such investment. This discussion includes a state-by-state analysis of local venture capital markets and state policies.[1]

What Is Venture Capital Investment?

Venture capital investing can be defined broadly as "investment by professional investors of long-term, risky equity finance where the primary reward is an eventual capital gain, rather than interest income or dividend yield" (Wright and Robbie 1997, xiii). This capital gain is realized when the venture capitalist or investing partners sell or otherwise liquidate their equity stake in the venture.

A diverse group of investors join venture capital partnerships. These investors include pension funds, endowments, foundations, bank holding companies, insurance companies, wealthy individuals, investment banks, and nonfinancial corporations. Table 1 shows the amounts of investment and distribution by each group of investors nationally from 1986 to 1992.

By investing in a particular entrepreneurial firm, the venture capitalist assumes a high level of risk. Sahlman (1990) found that 34.5 percent of venture capital investment results in a loss. The investor attempts to minimize these risks by controlling the stages and level of capital infusion, using built-in incentives to reward entrepreneurs' desirable behavior and often taking a very active role in managing the firm.

TABLE 1 U.S. Venture Capital Investors, 1986–92

	Investment ($ billions)	Percentage
Pension Funds	9.85	45
Corporate	5.91	27
Public	3.94	18
Endowments and Foundations	2.57	12
Bank Holding Companies and Insurance Companies	2.49	12
Wealthy Families and Individuals	2.33	11
Investment Banks and Nonfinancial Corporations	2.11	10
Other	2.33	11
Total	21.68	100

Source: Fenn, Liang, and Prowse (1995, 45)

Venture capitalists may be categorized by either the sources of investment capital—whether captive or independent—or the stage of business development on which they focus their investments. Captive venture capitalists are generally subsidiaries of banks or insurance companies and are funded through the mother institution; independent firms must seek funding through third parties.

Independent firms are primarily organized as limited partnerships. The venture capitalists are general partners, and the third-party investors are limited partners. As general partners, venture capitalists have considerable control over the firm and its management. Venture capitalists set certain developmental targets for enterprises and may release additional funds only as each goal is met. This sequential financing arrangement results in the release of enough capital to get the firm to the next level of maturity and no more.

Limited partners, on the other hand, use the venture capitalists as investment intermediaries and play a much more restricted role in management of the firm(s). Even though limited partners have little involvement in day-to-day management, the contractually specified relationship between general and limited partners helps ensure that the interests of the latter are not overlooked, as is discussed further below.

Venture capitalists pool investment funds from a variety of limited partners. These funds usually have a fixed life span (typically specified as ten years in the initial contract but extended if the fund is successful) and are used to invest in new ventures for their first three to five years of existence. After this initial stage, the funds are focused almost exclusively on moving the businesses they have financed up the development ladder toward eventual realization of investor returns.

Venture capitalists tend to set up new funds for different ventures before an existing fund's capital is exhausted, and then they repeat the process, often with the same limited partners. In this way they can preserve and leverage the knowledge and contacts associated with previous successful ventures.

Potential Conflicts of Interests between General and Limited Partners

While in principle all parties are interested in maximizing the value of the firms in the venture capital portfolios, venture capitalists may make decisions that run counter to outside investors' interests. These decisions include "spending too little time advising or monitoring the companies and entrepreneurs, charging excessive management fees, taking undue investment risks, and reserving the most attractive investment opportunities for themselves and their associates" (Fenn, Liang, and Prowse 1995, 35). A variety of contractual methods can minimize the potential misalignments of general and limited partners' interests. The contracts may include some or all of the following methods: limiting the life span of the venture fund, specifying limited partners' right to halt any further investments into the fund, tying most of the venture capitalists' ultimate profit to the "final" value of the firm, mandating distribution of the fund's proceeds, and outlawing other specific activities that would unfairly reward venture capitalists at limited partners' expense (Fenn, Liang, and Prowse 1995).[2]

[1] Southeast *refers to the six states that in whole or part make up the Sixth Federal Reserve District—Alabama, Florida, Georgia, Louisiana, Mississippi, and Tennessee.*

[2] *Such agency problems are far from unique to venture capital investment. Wall and Peterson (1998), for example, discuss this issue in the context of costs imposed on banks by the measurement and regulation of capital adequacy.*

The Investment Process

Stages. Sahlman (1990) presents eight stages of venture capital investing, described in the box on page 39. The primary goal, regardless of the stage at which venture capitalists enter the relationship, is to move the investment sequentially to a final, agreed-upon level of development, such as a public offering. After that level is reached the partners liquidate their equity and obtain their investment gains. If the venture has been successful, both the venture capitalists and the outside investors will realize most of their profits at this point.

Mechanics. Fenn, Liang, and Prowse (1995, 29) describe four investment activities undertaken by the general partners in a venture capital firm during the phases of a project. These activities include selecting, structuring, monitoring, and exiting investments. Proper execution of each responsibility is essential if general partners are to profit from a venture.

General partners contact entrepreneurs, investment bankers, brokers, consultants, lawyers, and accountants in their search for information on potential deals. Of the hundreds of business plans that they receive from firms seeking capital, only those with the highest probability of success are funded.

As large institutional investors pour assets into venture capital and as state and local governments seek to attract this capital, the potential benefits will grow, but not without raising public policy issues.

Once this process of screening potential firms is complete, general partners attempt to negotiate the terms of the investment agreement with the target firm(s). As mentioned above, because of potential agency problems, the structuring of the deal is extremely critical. The ideal contract aligns the interests of the general and limited partners with those of the target firm(s).

General partners closely monitor their portfolio companies. Board representation, management employment contracts, voting rights, consulting services, and control of access to additional funding are the primary means by which venture capitalists influence the enterprise. Limited partners have very little direct control except through their ability to refuse further funding.

Finally, general partners must exit the relationship with the firm. Various means of exiting include public offering, private sale, and share repurchase. Of these, "the public offering generally results in the highest valuation of a company" (Fenn, Liang, and Prowse 1995, 34). In a public offering the company issues stock and becomes a public enterprise, but the partners generally do not completely sever their relationship with the firm. For example, they are often legally required to hold shares of the firm for a specific period. During this period, venture capital investors often remain very active in the firm's management, reducing agency costs by forcing continued focus on the longer-term health of the firm. In a private sale the company is merged with or acquired by a larger company, and the general and limited partners are paid in cash or liquid securities. In the share repurchase option the firm is forced to buy back stock held by the general partners. The venture capital firm often uses this means of exiting when investments have been unsuccessful.

Common Characteristics of Projects

Finance theorists have developed classes of models that attempt to rigorously describe common characteristics of venture capital projects. The most prominent of the features modeled are the sequentiality of investments, the irreversibility of investments, and the option to postpone or terminate future investments in a project. While these features characterize new investments of most kinds, their rigorous treatment is more important in venture capital projects because of the higher risk involved.

Sequentiality. Venture capital investments tend to be made sequentially. Each dollar spent can be thought of as purchasing an option to make future investments in the firm. Even investments that appear to involve only a single decision can turn out to be sequential because many projects (especially large ones) take time to complete and can be halted in midstream (Dixit and Pindyck 1994, 320). For example, the construction of a large silicon chip manufacturing plant might involve the intermediate steps of building the physical infrastructure, purchasing and installing equipment, and training workers. Before such a project is completed market conditions could shift significantly and thus alter its final profitability.

Irreversibility. Another important feature of venture capital investing is that the investments made at each stage are largely irreversible. Once a factory is built or the initial research is completed, it is difficult if not impossible to recoup much of the investment if the project is unsuccessful. The potential for such sunk costs increases the total risk of the venture (Dixit and Pindyck 1994, 8).

Postponement or Termination. A third significant characteristic of venture capital investments is that projects can be postponed or terminated at each stage. The investors can evaluate whether to make further investments or delay or close down the operation altogether. Such a decision is sensitive to changes in the expected final value of the project or changes in the costs of completing the investments (Dixit and Pindyck 1994, 320).

BOX

Stages of Venture Capital Investment

Seed Investments

- Small amounts of capital are provided to an inventor or entrepreneur to determine whether an idea deserves further consideration.

Start-Up

- Companies are less than one year old.
- The company uses the money for product development, prototype testing, and test marketing.

First Stage—Early Development

- Investment continues through the first stage only if the prototypes look good enough for further technical risk to be considered minimal.
- First-stage companies are unlikely to be profitable.

Second Stage—Expansion

- A company in the second stage has shipped enough product to enough customers to have real feedback from the market.
- The firm is probably still unprofitable.
- The firm probably needs more capital for equipment purchases, inventory, and receivables financing.

Third Stage—Profitable but Cash Poor

- Sales growth is probably fast.
- New venture capital may be used for further expansion of manufacturing facilities, expanded marketing, or product enhancements.

Fourth Stage—Rapid Growth toward Liquidity

- A company may still need outside cash to sustain growth.
- The risk to outside investors is much reduced, and the cash-out point and method are underdetermined.

Bridge Stage—Mezzanine Investment

- Despite potentially knowing the approximate timing and form of exit of the venture capital from the company, the company still needs capital to continue growth.

Liquidity Stage—Cash-Out or Exit

- Investors can gain liquidity of a substantial portion of their holdings in a company.

Source: Sahlman (1990, 479)

Kinds of Risks

Venture capitalists face many risks in deciding to make initial investments and continue investing in portfolio firms. Berk, Green, and Naik (1997) develop a model to analyze three sources of risk, which they label technical, exogenous, and traditional. These various risks come to play at different points in the development of firms. Technical risks dominate the seed investment and start-up stages. Firms are susceptible to exogenous risks at all stages of development, and the full effects of traditional risks become apparent as the firm moves through the final stages of the process.

Technical Risks. In a start-up research and development venture, technical uncertainty refers to the "uncertainty associated with the success of the research itself" required to push the firm beyond the development stage (Berk, Green, and Naik 1997, 1). An example would be the difficulty entailed in developing a new software product, which may or may not work under actual programming conditions.

Exogenous Risks. Exogenous risks are those associated with the possible obsolescence of the firm's final output or product. This sort of risk is especially great in rapidly evolving markets such as the computer and software industries. For example, if Java becomes the industry standard, Sun's virtual machine and Java-based operating system could threaten the supremacy of software designed for a Microsoft Windows environment (Clark 1997).

Traditional Risks. Berk, Green, and Naik (1997, 2) define traditional risks as those related to the "uncertainty about the costs and [general] demand [conditions] that determine the ultimate cash flows" from the venture. In other words, fluctuations in the larger economy could affect supply and demand for the firm's final output. An unexpected economic recession (as opposed to the narrower threat posed by a competitor's product), for example, could cause a new venture to fail that in other times might succeed.

Early History of the U.S. Market

A multitude of factors converged to create the venture capital industry in the United States. This market has evolved over time in response to developments in technology, entrepreneurial need, capital availability, and the appropriate legal framework.

The first venture capital firm, American Research and Development, was established by Ralph E. Flanders, the former president of the Federal Reserve Bank of Boston, and General Georges Doriot of Harvard Business School, in 1946 (Pfirrmann, Wupperfeld, and Lerner 1997). One of the firm's first ventures was investment of seed capital in a company created by four MIT graduate students in 1957. American Research and Development provided $70,000 in exchange for 77 percent of common stock in the company. The company eventually evolved into Digital Equipment Corporation, and the original investment grew to $355 million by 1971.

The venture capital market in the United States has evolved over time in response to developments in technology, entrepreneurial need, capital availability, and the appropriate legal framework.

The next major step in the evolution of the U.S. venture capital industry was the Small Business Investment Company (SBIC) program of the Small Business Administration. Initiated in 1958, the aim of the program was to foster new company formation by augmenting more traditional sources with new sources of venture investment capital. The Small Business Investment Companies were allowed to borrow $4 from the Small Business Administration for each dollar of equity they raised, and "by 1965, the 700 licensed SBIC's dominated the domestic supply of venture capital" (Pfirrmann, Wupperfeld, and Lerner 1997, 22). Incompetence, fraud, and the resulting new regulatory environment in the industry led to the downfall of the program and to the eventual growing importance of private venture capital funds in this industry. As of 1997, Small Business Investment Companies made up 5 percent of the total capital pool (Pfirrmann, Wupperfeld, and Lerner 1997, 22).

While the recession of the 1970s contributed to a dampening of the venture capital market, "venture capitalists, entrepreneurs, and government joined in a combined effort to help revive the industry" (Pfirrmann, Wupperfeld, and Lerner 1997, 22). Several legislative changes helped. The first was the 1978 Employee Retirement Income Security Act's (ERISA) "Prudent Man" rule, which allowed pension funds to invest in higher-risk investments, including venture capital funds. Two more law and regulation changes in 1980 also contributed to the evolution of this market. First, the Small Business Investment Act of 1980 reduced the reporting requirements for venture capital firms by redefining them as business development companies as opposed to investment advisers. Moreover, the ERISA "Safe Harbour" regulation in 1980 reduced the legal oversight and potential liabilities of venture capitalists by legally defining pension funds as limited partners.

These regulatory changes opened up a large new source of venture capital funding. For example, pension funds, which supplied only 15 percent of the capital committed to venture funds in 1978, accounted for 46 percent by 1994 (see Table 2).

Later Evolution

With new funding sources and a conducive legal environment, the amount of capital raised by venture capital partnerships mushroomed. Over time institutional investors have come to dominate the market. As the industry has matured, these large investors have increased the size of the average venture capital fund from $18 million in 1979 to $68 million in 1993. As Table 3 illustrates, total capital commitment in the industry rose from $661 million in 1980 to $3.764 billion in 1994. At the same time the number of partnerships involved in later-stage deals grew from 4 percent to 26 percent of total partnerships, mostly at the expense of balanced partnerships. This shift reflects not only investors' ability to fund these more expensive investments but also their demand for earlier profit realization. Later-stage companies may show profits in a couple of years rather than the five or more needed for seed-level investments (Pfirrmann, Wupperfeld, and Lerner 1997).

As discussed earlier, most successful ventures are exited through a public offering or a private sale. Venture capitalists realize the highest returns from firms that go public (Pfirrmann, Wupperfeld, and Lerner 1997). The number of venture capital–related initial public offerings and acquisitions grew from 27 and 28, respectively, in 1980 to 136 and 97 in 1994. The companies that go public are the relatively rare successes and represent a small fraction—only about 10 to 30 percent of the total—of all firms that receive seed and early-stage financing (Fenn, Liang, and Prowse 1995, 21). However, their net effect can be relatively large. Sahlman found that "in aggregate, 579 venture-capital-backed companies went public during the 11 years ending in 1988. Their total market value exceeded 30% of the total market value of all comparable companies going public during the same period" (1990, 482).

The returns on venture capital have fluctuated over time. Sahlman (1990) reports that between 1965 and

TABLE 2
Sources of Capital Commitments to Private Independent Funds in the United States[a]

	Total Capital Commitments[b] ($ billions)	Corporations	Individuals and Families	Pension Funds	Foreign	Endowments and Foundations	Banks and Insurance Companies
1980	0.661	18	17	29	8	15	13
1981	0.867	17	23	23	10	12	15
1982	1.423	12	21	33	13	7	14
1983	3.408	12	21	31	16	8	12
1984	3.185	14	15	34	18	6	13
1985	2.327	12	13	33	23	8	11
1986	3.332	11	12	50	11	6	10
1987	4.184	11	12	39	13	10	15
1988	2.947	11	12	46	14	12	9
1989	2.399	20	6	36	13	12	13
1990	1.847	7	11	53	7	13	9
1991	1.271	4	12	42	12	24	6
1992	2.548	3	11	42	11	19	15
1993	2.545	8	7	59	4	11	11
1994	3.764	9	12	46	2	21	9

[a] Percentage of annual total
[b] Excludes funds of funds

Source: Pfirrmann, Wupperfeld, and Lerner (1997)

TABLE 3 Capital Raised by Venture Capital Partnerships by Stage of Investment

	Total Capital Commitments[a] ($ billions)	Number of Partnerships by Investment Stage (Percent)		
		Seed	Balanced	Later
1980	0.661	35	61	4
1981	0.867	43	57	0
1982	1.423	38	57	5
1983	3.408	32	59	9
1984	3.185	34	59	7
1985	2.327	37	49	14
1986	3.332	41	49	10
1987	4.184	32	60	8
1988	2.947	41	55	4
1989	2.399	50	45	5
1990	1.847	14	72	14
1991	1.271	48	47	5
1992	2.548	36	40	24
1993	2.545	22	66	12
1994	3.764	30	44	26

[a] Excludes funds of funds

Source: Pfirrmann, Wupperfeld, and Lerner (1997)

TABLE 4
Amount Raised by Venture Capital Funds in the Southeast[a]

	Venture Capital ($ millions)
1990	14.0
1991	12.2
1992	7.8
1993	32.0
1994	266.1
1995	99.4

[a] Includes Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee

Source: Brooks (1996)

1984 the median rate of return on venture capital firms exceeded 26 percent per year. For the 1991–94 period, the average rate of return realized by limited partners for each year was 24.0, 12.5, 19.7, and 16.2, respectively. Rates of return have declined since the early 1980s for a variety of reasons, including the rising valuation of deals caused by increased competition, greater focus on later-stage investments with lower risks and expected returns, and, possibly, a reduction in the quality of venture capitalists' decision making (Pfirrmann, Wupperfeld, and Lerner 1997).

Venture Capital Developments in the Southeast

As shown in Table 4, venture capital funds raised in the Southeast have grown from $14 million in 1990 to $99.4 million in 1995. While this amount is not massive (individual funds can grow to $100 million or more), the trend is still clear and follows the growth of venture capital funds nationwide. Many factors have contributed to this growth, including the region's overall economic development, the gradual emergence of regional venture capital firms, increasing competition in venture markets elsewhere, state policies favorable to venture capital, and the growth of high-technology and communications-related industries.

Venture capital investments in firms throughout the Southeast vary widely by industry type, stage of development, and magnitude. As Table 5 shows, 32 percent of funds invested by venture firms went to companies in healthcare, 27 percent to communications firms, 13 percent to software and information companies, and smaller percentages to other industries. Nationally, software and information companies received the largest share of funds at 25 percent, followed by communications, healthcare, and business services.

Table 6 presents the stage of development of the companies that received venture capital in the Southeast in 1997. Expansion-stage companies picked up the largest percentage of venture capital at 36 percent. Early-stage firms followed closely with 28 percent while start-up and late-stage companies received 13 percent.

Nationally, in 1997, 70 percent of venture capital funds went to expansion- and late-stage firms, with the largest amount—53 percent—going to expansion-stage firms. Early- and start-up-stage firms received 15 and 5 percent of the total, respectively. This distribution contrasts with the more even balance between the start-up/early-stage and expansion/late-stage sectors in the Southeast.

Alabama. The largest share of Alabama's venture investment went to the healthcare industry (Table 7). Electronics and instrumentation companies, consumer businesses, communications firms, and the software and information companies followed. Venture capitalists made sixteen distributions of funds, totaling $49,291,000, throughout the state in 1997 (Price Waterhouse 1998). Seed and early-stage companies each composed 31 percent of the companies given venture capital funding in Alabama in 1997. Expansion- and late-stage firms received 19 and 13 percent, respectively. As with the Southeast in general, the relatively high percentage of investment in earlier-stage firms probably reflects the relative nascence of venture capital investment in this state.

Florida. In Florida a total of $340 million was distributed in 1997 (Table 7). Communications companies received the largest portion of this total, with the healthcare industry collecting the next-largest portion. Software and information firms, business services, and electronics and instrumentation firms, each with single-digit shares, make up the rest of the top five companies. Venture capitalists made fifty-nine separate distributions to firms in the state in 1997 (Price Waterhouse 1998). Expansion capital was awarded to 38 percent of the companies receiving venture capital in 1997 (Table 6). Florida, with 18 percent of the total, had the highest share in the Southeast of late-stage companies receiving venture funding (Table 6).

Georgia. Last year Georgia's venture capital investments were not only the largest in the Southeast but also the most diversified. The majority of the funds, 73 percent, went to healthcare, software and information, and communications companies, with the consumer and the distribution/retailing sectors completing the top five recipients. Georgia firms received eighty-one distributions of venture capital (Price Waterhouse 1998) totaling $347,700,000 in 1997 (Table 7).

Georgia's success in 1997 in attracting ventures based on high technology is evident in a study by Price Waterhouse of Internet-related venture capital investments (see Table 8). Although still dwarfed by such

TABLE 5 Venture Capital Investment by Type of Company, 1997

	Southeast (Dollars)	Percentage	U.S. (Dollars)	Percentage
Biotechnology	3,039,000	0	670,014,500	5
Business Services	51,432,000	6	712,890,000	6
Communications	235,072,000	27	2,858,832,348	22
Computers and Peripherals	3,816,000	0	588,096,000	5
Consumer	58,420,000	7	693,311,000	5
Distribution/Retailing	24,620,000	3	700,364,000	6
Electronics and Instrumentation	40,085,000	5	408,940,000	3
Environmental	—	—	71,231,000	1
Healthcare	283,736,000	32	1,248,399,677	10
Industrial	52,745,000	6	693,679,500	5
Medical Instruments and Devices	6,305,000	1	612,919,000	5
Miscellaneous	—	—	21,450,000	0
Pharmaceuticals	7,475,000	1	233,282,000	2
Semiconductors/Equipment	—	—	101,251,000	1
Software and Information	116,371,000	13	3,176,119,248	25
Total	883,116,000	100	12,790,779,273	100

Source: Price Waterhouse (1998)

technology powerhouses as California's Silicon Valley and Boston's Route 128, Georgia ranked tenth among states in the number of Internet-related venture capital deals within the state, seventh in the total amount invested, and second in the amount invested from among those deals. In 1997 the $30 million invested in an Atlanta developer of multimedia and Web/Internet services for corporations was the third-largest deal in the nation in this sector. Georgia, New York (second), and Connecticut (ninth) were the only states besides California included in Price Waterhouse's listing of the 1997 top ten Internet deals in terms of the total amount invested (Price Waterhouse 1998).

Louisiana. Business services firms received the lion's share of Louisiana's $36,100,000 in venture capital funds in 1997 (Table 7). The industrial, healthcare, software and information, and communications sectors had much smaller shares. A total of twelve distributions of funds were made to firms in Louisiana in 1997 (Price Waterhouse 1998). Expansion-stage companies made up 67 percent of the venture capital recipients (Table 6). The relative number of expansion companies in the state receiving venture funding was the largest in the Southeast.

Mississippi. Most of Mississippi's venture capital investments—90 percent of $10,420,000—went to consumer firms, with the remaining 10 percent to the medical instruments and devices industry. Two distributions of venture capital were made to firms in Mississippi in 1997 (Price Waterhouse 1998). These two disbursements went to a start-up and a public company.

Tennessee. Healthcare firms picked up almost three-quarters of the state's total venture capital investments, while industrial and communications firms received most of the remainder. Twenty-one distributions of venture capital were made to firms in Tennessee in 1997 (Price Waterhouse 1998). Expansion-stage firms, with 43 percent of the total, received the largest share of funding.

State Policies

The states in the Southeast have followed a variety of policies in attempting to increase the amount of venture capital and the number of venture capital firms operating in their states. The states have focused to varying degrees on increasing the number of venture capital funds and the amount funds invest within their respective states, developing the high-technology and research sectors, and increasing interaction among the various actors involved in the venture capital industry. These efforts have been implemented only relatively recently, so few clear results have emerged to show their impacts on investment or, more importantly, employment or income in the states.

Alabama. The science, technology, and energy division of Alabama's Department of Economic and Community Affairs helps provide research grants to scholars and businesses and aids in technology transfers from the National Aeronautics and Space Administration to local businesses. The state's Small Business Innovation Research Program provides information on

TABLE 6 Venture Capital Investment by Stage of Development, 1997 (percent)

	Start-Up/Seed	Early	Expansion	Late	Public	Turnaround	Not Categorized
Alabama	31	31	19	13	0	0	6
Florida	7	26	38	18	2	0	10
Georgia	15	33	32	11	1	1	6
Louisiana	8	17	67	8	0	0	0
Mississippi	50	0	0	0	50	0	0
Tennessee	10	24	43	10	5	0	10
Southeast	13	28	36	13	2	1	7
United States	5	15	53	17	0	0	9

Source: Price Waterhouse (1998)

federal financial assistance for technology development to small businesses and entrepreneurs.

Florida. Focusing on the number of venture capital firms operating in the state, Florida lawmakers are concerned that most of Florida's venture capital comes from outside the state. As a result, legislation has been proposed that would give a 100 percent credit on the state's premium tax for insurance companies that invest in Florida-based venture capital firms (McKinnon 1997). By focusing on creating more Florida-based institutional investors (because, as discussed above, it is institutional investors who seek out larger and later-stage deals), the state may promote even further the venture funding of later-stage companies.

Georgia. Initial failures in luring high-technology firms helped motivate Georgia's business leaders to see what could be done to make the state more competitive in securing and developing homegrown high-technology companies ("20 Years…" 1998). This motivation led in 1990 to the creation of the Georgia Research Alliance, a partnership among universities, businesses, and state government. Its main areas of focus are Georgia's telecommunications, environmental technology, and biotechnology industries. The alliance seeks to spur the creation of high-technology start-ups by bringing together scientists and entrepreneurs and systematically investing in the state's research infrastructure. The organization has raised more than $200 million through a combination of state, federal, and private sources. The money is invested in eminent scholars, research facilities, and scientific equipment for Georgia's research universities.

Louisiana. Louisiana's Department of Economic Development has several programs and incentives to promote the flow of venture capital funds to businesses based in the state. For example, one program provides matching state funds of up to $5 million for private Louisiana-based venture capital funds worth at least $5 million. A related program provides for a coinvestment in a business located in the state of up to one-fourth of the funding for a given stage of investment, but not more than $500,000, for any qualified venture capital fund with at least $7.5 million in private capital, which may come from outside the state. Yet another program provides $1 for every $2 of private capital up to $5 million for minority venture capital funds that have at least $250,000 of private investment (Louisiana Department of Economic Development 1998). In addition, tax credit legislation similar to that proposed in Florida has led to the creation of seventeen in-state venture capital firms (McKinnon 1997). Despite these initiatives, Louisiana still lags behind most of the Southeast in venture capital investment.

Mississippi. The Mississippi legislature passed the Venture Capital Act of 1994 to foster industry in the state. This program was funded through the sale of a $20,000,000 general obligation bond guaranteed by the state. The act created the Magnolia Venture Capital Corporation and the Magnolia Venture Capital Limited Partnership to "increase the rate of capital formation, stimulate new growth-oriented business formations, create new jobs for Mississippi, develop new technology, enhance tax revenues for the state, and supplement conventional business financing" (Mississippi Legislature 1997, 4).

The Magnolia Venture Capital Corporation is intended to serve as the general partner primarily for potential high-growth businesses located in Mississippi. The corporation will invest in the companies, which in turn will be the limited partners. Seventy percent of the funds are to be invested in start-up businesses (less than thirty-six months old), and the remainder can go to older firms. From January 1, 1996, through January 31, 1997, the corporation received eighty business plans. Of those, sixty-two were determined to be eligible to apply for the program, ten are under review, five were referred to another venture capital firm, one was retracted, and one was approved (Mississippi Legislature 1997).

Tennessee. State Senate Joint Resolution 704, filed April 23, 1998, calls for Tennessee's Department of Treasury to study the feasibility of investing the assets of the Tennessee Consolidated Retirement System in

TABLE 7 Venture Capital Investment in Southeastern States by Type of Company, 1997[a]

	AL	%	FL	%	GA	%	LA	%	MS	%	TN	%
Biotechnology	500	1	—	—	2,539	1	—	—	—	—	—	—
Business Services	—	—	24,307	7	4,375	1	22,250	62	—	—	500	1
Communications	5,000	10	140,268	41	79,641	23	700	2	—	—	9,463	10
Computers and Peripherals	90	0	226	0	3,500	1	—	—	—	—	—	—
Consumer	7,500	15	12,000	4	28,300	8	—	—	9,420	90	1,200	1
Distribution/Retailing	—	—	3,800	1	20,320	6	500	1	—	—	—	—
Electronics and Instrumentation	10,600	22	17,785	5	11,700	3	—	—	—	—	—	—
Environmental	—	—	—	—	—	—	—	—	—	—	—	—
Healthcare	23,100	47	96,400	28	88,770	26	3,050	8	—	—	72,416	73
Industrial	—	—	10,965	3	18,680	5	8,100	22	—	—	15,000	15
Medical Instruments and Devices	—	—	—	—	5,305	2	—	—	1,000	10	—	—
Miscellaneous	—	—	—	—	—	—	—	—	—	—	—	—
Pharmaceuticals	—	—	6,450	2	—	—	—	—	—	—	1,025	1
Semiconductors/Equipment	—	—	—	—	—	—	—	—	—	—	—	—
Software and Information	2,501	5	27,800	8	84,570	24	1,500	4	—	—	—	—
Total	49,291	100	340,001	100	347,700	100	36,100	100	10,420	100	99,604	100

[a] Thousands of dollars

Source: Price Waterhouse (1998)

TABLE 8
Internet-Related Investments, 1997

	Number of Deals
California	220
Massachusetts	48
Colorado	20
New York	17
Pennsylvania	16
Minnesota	14
Texas	12
Virginia	12
Washington	12
Georgia	9

	($ millions)
California	1,088
Massachusetts	224
New York	96
Colorado	72
Texas	56
Pennsylvania	53
Georgia	45
Virginia	40
Minnesota	33
Connecticut	29

	$ per Deal (millions)
New York	5.6
Georgia	5.0
California	4.9
Massachusetts	4.7
Texas	4.7
Colorado	3.6
Virginia	3.3
Pennsylvania	3.3
Minnesota	2.4

Source: Price Waterhouse (1998)

alternative investments, including, but not limited to, venture capital, private equity, corporate restructuring, expansion capital, and energy and natural resources. In addition, Tennessee's Department of Economic and Community Development has new programs that target small and minority-owned telecommunications firms. These programs provide firms with education, training support, market development counseling, and loan guarantees of up to 80 percent for a $400,000 project.

Conclusion

In recent years venture capital investment throughout the United States and in the Southeast has shown dramatic gains. Nationally, venture capital has already played a major role in the formation of some of the most important and innovative firms in the U.S. economy and has provided an alternative outlet for investment funds from major institutional investors. As the venture capital industry matures in regions of the country where it has a longer history, it is seeking out new arenas for expansion, such as the Southeast.

The growth of venture capital investment in the Southeast has been supported by trends such as the region's economic development, the emergence of regional venture capital firms, increasing competition in venture markets in other parts of the country, state policies favorable to venture capital investing, and the growth of high-technology and communications-related industries in the Southeast. Venture capital investing has become an important alternative source of funds for less-developed, higher-risk entrepreneurial firms that may not have access to more traditional capital sources.

Starting from a small base just a few years ago, venture capital has become an integral part of new business formation in the Southeast. New technological advances, business opportunities, and entrepreneurial needs should continue to spur development of the region's venture capital industry.

State actions to spur venture capital investing in the region have been quite varied in nature. Government support of venture capital funds or projects has been active in some southeastern states and at least considered by the rest. So far, clear evidence on the impact of state venture capital support and, implicitly, funds on income and employment is not available. The role of public support for funds and projects therefore may still be questioned. Nonetheless, with or without state involvement, it seems likely that venture capital will become increasingly important to the emergence of new industries and technologies in the region.

REFERENCES

BERK, JONATHAN, RICHARD C. GREEN, AND VASANT NAIK. 1997. "Valuation and Return Dynamics of R&D Ventures." Unpublished manuscript.

BERLIN, MITCHELL. 1998. "The Thing Venture Capitalists Do." Federal Reserve Bank of Philadelphia *Business Review* (January-February): 15–26.

BROOKS, RICH. 1996. "Southeast Journal: Two Big Funds Infuse Region with Capital." *Wall Street Journal*, May 15.

CLARK, DON. 1997. "Technology: Sun Microsystems Pours Some New Java; Programming Language to Run on Servers, Appliances." *Wall Street Journal*, April 1.

DIXIT, AVINASH K., AND ROBERT S. PINDYCK. 1994. *Investment under Uncertainty.* Princeton, N.J.: Princeton University Press.

FENN, GEORGE W., NELLIE LIANG, AND STEPHEN PROWSE. 1995. "The Economics of the Private Equity Market." Board of Governors of the Federal Reserve System, Staff Study 168, December.

LOUISIANA DEPARTMENT OF ECONOMIC DEVELOPMENT. 1998. "Financing Assistance: Investment Programs." Available on-line at <http://www.lded.state.la.us/new/financing.htm> [July 29, 1998].

MCKINNON, JOHN D. 1997. "Florida Journal: Lawmakers Propose Plan to Spur Local Venture-Capital Investments." *Wall Street Journal*, April 9.

MISSISSIPPI LEGISLATURE. 1997. PEER Committee. "Review of Implementation of the Venture Capital Act of 1994 and the Operations of the Magnolia Venture Capital Corporation." March 11, 1997.

PFIRRMANN, OLIVER, UDO WUPPERFELD, AND JOSHUA LERNER. 1997. *Venture Capital and New Technology-Based Firms.* Heidelberg: Physica-Verlag.

PRICE WATERHOUSE. 1998. "National Venture Capital Survey, Full Year 1997."

SAHLMAN, WILLIAM A. 1990. "The Structure and Governance of Venture-Capital Organizations." *Journal of Financial Economics* 27 (October): 473–521.

"20 YEARS OF TELECOMMUNICATIONS." *Atlanta Business Chronicle*, February 2, 1998.

WALL, LARRY D., AND PAMELA P. PETERSON. 1998. "The Choice of Capital Instruments." Federal Reserve Bank of Atlanta *Economic Review* 83 (Second Quarter): 4–17.

WRIGHT, MIKE, AND KEN ROBBIE. 1997. *Venture Capital.* Aldershot, England: Dartmouth Publishing Company Limited.