Occasional Paper

2006-2

METHODOLOGY FOR THE SHORT-TERM

CONNECTICUT INDUSTRY EMPLOYMENT FORECASTS

Opportunity • Guidance • Support

CONNECTICUT DEPARTMENT OF LABOR

A Partner in CTWorks

Patricia H. Mayfield, Commissioner

www.ct.gov/dol

METHODOLOGY FOR THE SHORT-TERM

CONNECTICUT INDUSTRY EMPLOYMENT FORECASTS

Prepared By:
Daniel W. Kennedy, Ph.D., Senior Economist

August 2005
(Revised August 2006)

Connecticut Department of Labor
200 Folly Brook Boulevard
Wethersfield, CT 06109 – 1114

Patricia H. Mayfield, Commissioner

Roger Therrien, Director
Office of Research

TABLE OF CONTENTS

EXECUTIVE SUMMARY

A. Introduction i

B. Producing Short-Term Employment Forecasts: The Approach i

C. Forecasting Methods ii

D. Eight-Step Process to Producing the CT DOL Employment Forecasts vi

E. Concluding Remarks and Endnotes viii

I. Introduction 1

II. Producing Short-Term Employment Forecasts

- A. Three Levels of Forecasts **2**
- B. The Base-Line Forecast **3**
- C. Intercept Corrections: Adjusting the Forecast **4**
- D. The Final Forecast **5**

III. Forecasting Methods: A Detailed Discussion

- A. How Do Economist Forecast? **6**
- B. The Super-Control Forecasts of Connecticut Employment **7**
- C. The Control Forecasts **34**
- D. The Detailed-Level Forecasts **48**

IV. Concluding Remarks 54

APPENDIX A: Data Requirements 56

APPENDIX B: Bayes's Theorem 58

APPENDIX C: Seemingly Unrelated Regressions (SUR), or Near VAR 63

APPENDIX D: Stationarity 69

APPENDIX E: List of 'Getting Started' Forecast References 73

ENDNOTES 74

REFERENCES 76

EXECUTIVE SUMMARY

A. INTRODUCTION

In June 1998, the Office of Research, Connecticut Labor Department (CTDOL), began developing short-term forecasts (two years ahead) of industry employment to supplement the existing long-term projections (ten years out) program. This was a result of a decision by the Employment and Training Administration (ETA) of the U.S. Labor Department (USDOL) in 1995 to award grants to a consortium of states, lead by Illinois and Utah, to develop models for short-term industry forecasts. The Short-Term Industry and Occupational Forecasts support the One-Stop Delivery and Re-Employment Program. Through the One-Stop Centers, job-seekers faced with an occupational choice, change, or adjustments are provided with a primary place in the community to learn about employment opportunities. The Short-Term Employment and Occupational Forecasts are part of a service-delivery system that reflects customer demand for user-friendly information. The Short-Term Forecasts provide current Labor Market Information (LMI) on job opportunities, allow informed choices for short-term training with a goal of immediate re-employment, and they establish state-to-state comparability to facilitate job-match searches beyond the local labor market. The Short-Term Forecasts also serve as critical LMI for Workforce Investment Planning under the *Workforce Investment Act* (WIA). The employment forecasts are not only end products in themselves, but they also serve as the inputs to the Occupational Forecasts. In addition, the employment forecasts are an important source on Connecticut's economic outlook over the two-year forecast period. This provides valuable information on the State's near-term economic and labor-market prospects to decision-makers in both, business and government

B. PRODUCING SHORT-TERM EMPLOYMENT FORECASTS: The Approach

Connecticut short-term employment is forecasted at three different levels of detail: The Super-Control Forecast, the Control Forecasts, and the Detailed-Level Forecasts. After forecasts are produced, the three different levels of forecasts are reconciled. Each level of forecast produces progressively more detailed forecasts. The Super-Control Forecast is the top-line level of Connecticut, Non-Agricultural Employment, and it gives the least level of detail. The Control Forecasts provide a greater level of detail. The Control Forecasts are produced at the NAICS sector level, or two-digit level of detail. Forecasts are produced for the 20 NAICS sectors, including some of their major sub-aggregates, such as Durable Goods and Non-Durable Goods under the Manufacturing Sector. Finally, the detailed-Level forecasts, as would be expected, provide the most detail. The Detailed-Level Forecasts are produced at the NAICS three- and four-digit level of detail. Forecasts are produced for some 100 three- and four-digit sectors in Connecticut. The next section now turns to an overview of the methodology employed to produce each of the three levels of forecasts.

C. FORECASTING METHODS

Economists, and forecasters in general, use many different methods to make their forecasts. These include more formal methods such as Model-Based Statistical Analysis and Statistical Analysis not based on Parametric Models. Some other techniques that economists turn to for making their forecasts include Simple Extrapolations, Leading Indicators, and ‘Chartist’ approaches (also called Technical Analysis). But, such informal methods as ‘Back-of-the-Envelope’ calculations and Informed Judgment are also used. Some forecasters might even resort to some really informal methods such as Tossing a Coin, Guessing, or ‘Hunches’. However, the tools most frequently used are Econometric and Time-Series Models. They are the primary methods of forecasting in economics, but Judgment, Indicators, and even Guesses may modify the resulting forecasts[I].

Time-Series models, which describe the historical patterns of data, are popular forecasting methods and they forecast well compared to Econometric Systems of Equations. Particularly, in their multivariate forms, such as *Vector Autoregression* (VAR), Time-Series models do very well. However, Econometric Systems of Equations are the main

[I] This paragraph draws heavily on Hendry, David F., *How Economists Forecast* in UNDERSTANDING ECONOMIC FORECASTS, Edited by David F. Hendry and Neil R. Ericsson (2003) MIT Press: Cambridge, MA., pp. 21-22.

tool in economic forecasting. Econometric Forecasting Models are systems of relationships between variables such as GDP, Money, Employment, Inflation, etc. The relationships or ‘equations’ in these models are then estimated from the available data, which are mainly aggregate time-series.

The ***Super-Control Forecast*** is based on a single-equation regression model. A regression equation relates one or more *Independent* or *Explanatory* variables to a *Dependent* or *Explained* variable. Specifically, the regression model used to forecast top-line employment introduces dynamic effects into the model is by means of lagged values of the dependent variable: This is known as an *Autoregression*, or AR model. However, introducing lagged values of the independent, or exogenous variables, (as well as current values) introduces still another dynamic dimension to the model. Thus, the model used to forecast Connecticut’s top-line, super-control forecast is an *AR model with Exogenous variables*.

The next level of forecasting detail, moving from the Super-Control Total down to more detail, is the set of ***Control-Total Forecasts***. Many of the Control-Total Forecasts are produced using *multivariate time-series methods.* Particularly, *Vector Autoregressions* (VAR) are used in many instances. The VAR can be thought of as a generalization of the AR process, (see the discussion of the Super-Control Forecast, above), to two or more AR processes. Thus, a VAR is a system of two or more simultaneous equations expressing two or more interrelated AR processes. Central to the VAR is the concept of a *Recursive or, Feedback Relationship*. This allows forecasting models to draw on economic linkages and interconnectedness to construct feedback systems that tap into the direct and indirect effects of employment-changes in a given industry on other, related industries. An example of a grouping of industries for forecasting the Control Totals is the link or chain of Durable Goods sectors. A VAR constructed to capture this relationship would contain endogenous variables for each stage along the production chain from Durable Goods in Manufacturing, to Durable Goods in Wholesale Trade, to Consumer Durables in Retail Trade. Other relationships also exist, such as, firms interacting at the same stage of production, and interconnections at the same stage of production, and at different stages, simultaneously. Much more detail on inter-firm and inter-industry connections can be found in the literature on combining VAR's with Input-Output Analysis[II] and Industry Clusters[III].

The Vector Autoregression (VAR) has many advantages as a forecasting tool. However, one disadvantage is the 'one-size-fits-all' approach to specifying the equations in the VAR system. However, there is a more flexible approach. This approach is known as a *Seemingly Unrelated Regressions* (SUR) model, or *Near-VAR*, which was first suggested by Arnold Zellner (1962)[IV] in the early 1960's.

As discussed above, grouping industries according to similarities in the behavior of their employment dynamics can be captured by taking advantage of the Vector Autoregression (VAR) specification. Extensions of the VAR to the *Dynamic Simultaneous Equations Model* (SEM) framework allows the introduction of exogenous[V] variables into the model to account for seasonality, business cycles, industry-specific factors, and other influences external to the recursive relationship reflected in the endogenous variables of the VAR system. Nevertheless, since the VAR specification assumes that the endogenous variables all have the same number of lags in each equation in the system, and that the independent variables across all equations are the same and, that contemporaneous correlation among the error series across equations is minimal or nonexistent, it still constrains the system to a 'one-size-fits-all' specification. In some cases, gains in forecasting accuracy may be realized by allowing for differences in the lags of endogenous, as well as, exogenous, variables across equations, and for taking into account instances of significant

[II] Rickman, Dan S., "Generalizing the Bayesian Vector Autoregression Approach for Regional Interindustry Employment Forecasting", *Journal of Business and Economic Statistics* (1998) 16(1): pp. 62-72.

[III] See Nicholas Jolly, *Connecticut's Industry Clusters* (July 2005) OCCASIONAL PAPERS & REPORTS, Office of Research, Connecticut Labor Department: Wethersfield for a discussion on Connecticut's industry clusters. VARs could be specified such that, industries included in the system are grouped by industry clusters.

[IV] Zellner, Arnold, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", *Journal of the American Statistical Association* Vol.57 (June 1962): pp. 348-368.

[V] Within the VAR context, *Exogenous* variables are variables that do not have an equation within the VAR system. Their values, and forecasts, are determined outside the VAR model. Whereas, *Endogenous* variables are represented by an equation within the VAR system, and their forecasts are produced by the interaction of the equations within the VAR system.

contemporaneous correlation. This is especially important in regard to the set of exogenous variables. Under certain circumstances, the restriction to a ‘one-size-fits-all’ specification of the exogenous variables in the conventional VAR framework, in some instances, compromises the ability to produce more accurate forecasts.

The advantage offered by the SUR specification lies in its ability to capture structural breaks that frequently occur at different points, or may not even apply to some series in the system. Further, one equation may have statistically significant seasonality, while another may not. An example is the modeling and forecasting of the control totals (in this case, at the NAICS three-digit level), for Connecticut’s wholesale trade employment series. While the durable goods component displayed no discernible seasonality, there was a strong seasonal movement in the non-durable Goods employment series. Both employment series displayed structural breaks at the same point, and had similar trends.

The more flexible Near-VAR specification allows the forecaster to capture those factors common to both industries in the two-equation system, on the one hand, but it also allows the introduction of variables that represent factors effecting the level of employment that are unique to one industry’s employment-level in the system.

Both, VAR and Near-VAR models are used to produce the Control Forecasts of Connecticut employment.

Given the level of detail, the process for producing the ***Detailed-Level Employment forecast*** is necessarily the most mechanical. There are some 100 three- and four-digit level NAICS industries in Connecticut, which limits the amount of time and effort that can be devoted to developing and estimating a given forecasting model. There are two primary tools used for forecasting Connecticut Employment at the detailed level, the *Short-Term Industry Projections* (STIP) system developed by the consortium of states for ALMIS (America’s Labor Market Information System) to provide a tool for states’ LMI (Labor Market Information) units to develop timely, relatively uniform employment forecasts. SAS/ETS, the Econometric and Time-Series package is also used, particularly, the Forecasting Menu System, and PROC FORECAST, the multiple-series forecasting utility.

The forecaster using the STIP system has five models to choose from: Exponential Smoothing with Linear Trend and Random Walk options, OLS (single-equation, Linear Regression), ARMA (Autoregressive Moving Average), VAR, and BVAR[VI]. Mix gives a weighted average forecast based on the five models available in the STIP system. Most of the models used to forecast industry-employment at the Detailed-Level are multiple, time-series systems. The VAR and BVAR specifications are drawn on quite frequently. In addition to the specific employment-series being forecasted, other, related-industries included in a VAR or BVAR, as endogenous variables, are those suggested by the inter-industry relationships found in the 1997 Benchmarked, U.S. Input-Output Table. However, in some instances, there are no related industries. In such cases, univariate models are used to forecasts the employment series. There are two types of univariate models used in the Connecticut Forecasts: *Deterministic* and *Stochastic.* Section D, below, provides a brief overview of the eight-step process employed to produce the final set of industry-employment forecasts.

D. EIGHT-STEP PROCESS TO PRODUCING THE CT DOL EMPLOYMENT FORECASTS

The approach to producing the final Connecticut employment forecast can be summarized by the following eight-step process, in which the three different levels of forecasts, (1.) The Super-Control Forecast, (2.) the Control Forecasts, and (3.) the Detailed-Level Forecasts, are produced and reconciled:

1. ***The Super-Control Forecast*** is a single-equation, autoregression model relating Connecticut Non-Farm Employment with past values of itself, current-period and past values of Capacity Utilization in Manufacturing, and deterministic components to capture seasonality, long-run trend, and structural

[VI] BVAR is a Bayesian VAR. See Section II and Appendix B of the unabridged version of this paper to find an explanation of the BVAR.

breaks. Past versions of the model have included short-term interest rates, although the current version does not.

2. ***The Control Forecasts*** use different models to forecast employment at the NAICS sector-level and some sub-aggregates, such as Durable Goods under Manufacturing. Though different modeling approaches are used for different sectors and sub-aggregates, most rely on systems of equations, including multivariate time-series models. Further, some aggregates across NAICS sectors may be grouped together, where appropriate, to construct time-series models for forecasting employment. Forecasting models range from VARs and BVARs, to Near-VARs (SUR), to, in some instances, single-equation models, including ARIMAs and time-series regressions.

3. ***Detailed-Level Forecasts*** are produced in SAS/ETS, or the consortium of states' software called 'STIP' (Short-Term Industry Projections), or in some forecast rounds, both. This level of employment forecasts is necessarily the most mechanical of the three complementary approaches, as even for a small state like Connecticut, there are over 200 industries at the three- and four-digit NAICS levels of detail.

4. ***Pooling or Combining*** of forecasts is done after all three methods have been implemented. The Control-Total and Detailed-Level of forecasts present the opportunity for Pooling or Combining forecasts at the sector-level of employment. The Detailed-Level forecasts are added up to the NAICS sectors and their major sub-aggregates levels and then combined with the Control Forecasts to produce a set of simple average forecasts for each of the 19 non-Agricultural NAICS sectors. Then, the sum of the Control Totals, the Detailed-Level, and the Super Control Total are averaged together to produce the simple-average forecast for Top-Line, Non-Farm Employment. It should be noted that this is not a purely mechanical process. That is, the simple average forecast is not necessarily the final forecast for a given sector. Judgment does play a role, and one or the other forecast may be picked over the average forecast in more than one instance, especially when considering the top-line forecast.

5. ***Reconciliation*** of the three forecasts is done at both, the *top-down*, and *bottom-up* approaches. Three top-line forecasts are produced: (1.) the Super-Control Forecast, (2.) the Sum of the Control Forecasts, and (3.) the Sum of the Detailed-Level Forecasts. As discussed above, any pooling or combining of forecasts will be done before reconciliation.

6. ***The Base-Line Forecast*** is the product of steps 1 to 5 above. Once the Base-Line Forecast is in place, any Intercept Corrections are then implemented.

7. ***Intercept Corrections*** are done at four different stages: (1.) If any revisions to the employment data become available after the forecasts are produced (up to

a certain point), they will be used to make any necessary Intercept Corrections to put the forecasts *on track* with the historical series, (2.) Announced job eliminations and additions are used to make Intercept Corrections at the three- and four-digit NAICS industry-levels of detail, (3.) Reconciliation of the top-line forecasts with all three approaches, after announced job-changes have been incorporated, will invariably lead to further Intercept Corrections, (4.) The last of the Intercept Corrections are based on Macroeconomic considerations about any anticipated impacts of any policy-changes likely to be implemented over the forecast horizon, including variables to be affected, as well as the magnitude and duration of those affects, the stage of the business cycle the economy is believed to be in at the base period, and where it will be over the forecast horizon. A final consideration in adjusting the intercept is the forecaster's subjective probabilities about the economic and non-economic risks to the forecast.

8. ***The Final Forecast*** is the product of the above seven steps.

E. CONCLUDING REMARKS

It is hoped that this summary has succeeded in providing an informative, non-technical overview of the quantitative and qualitative methodologies used, as well as the process employed to produce Connecticut's Short-Term Employment Forecasts. The forecast horizon of two years, or eight quarters, for the short-term forecasts requires the forecaster to focus on analyzing the economy in the short- to intermediate-run. This means that forecasting methods must identify the expected seasonal, cyclical, and even some trend effects, as well as, regional and macroeconomic factors that influence the behavior of Connecticut's industry employment. It is this process of capturing these critical phenomena, in order to construct models that produce optimal forecasts, given time and resource constraints, that has guided the development of the methodologies applied to the short-term employment forecasts.

For more information, or any questions concerning the methodology used to produce the employment-forecasts, please contact:

Daniel W. Kennedy, Ph.D., Senior Economist
CT Department of Labor – Office of Research
(860) 263-6268
daniel.kennedy@ct.gov

I. INTRODUCTION

In June 1998, the Office of Research, Connecticut Labor Department (CTDOL) began developing short-term forecasts (two years ahead) of industry employment to supplement the existing long-term projections (ten years out) program. This was a result of a decision by the Employment and Training Administration (ETA) of the U.S. Labor Department (USDOL) in 1995 to award grants to a consortium of states, lead by Illinois and Utah, to develop models for short-term industry forecasts. The Short-Term Industry and Occupational Forecasts support the One-Stop Delivery and Re-Employment Program. Through the One-Stop Centers, job-seekers faced with an occupational choice, change, or adjustments are provided with a primary place in the community to learn about employment opportunities. The Short-Term Employment and Occupational Forecasts are part of a service-delivery system that reflects customer demand for user-friendly information. The Short-Term Forecasts provide current Labor Market Information (LMI) on job opportunities, allow informed choices for short-term training with a goal of immediate re-employment, and they establish state-to-state comparability to facilitate job-match searches beyond the local labor market. The Short-Term Forecasts also serve as critical LMI for Workforce Investment Planning under the *Workforce Investment Act* (WIA). The employment forecasts are not only end products in themselves, but they also serve as the inputs to the Occupational Forecasts. In addition, the employment forecasts are an important source on Connecticut's economic outlook over the two-year forecast period. This provides valuable information on the State's near-term economic and labor-market prospects to decision-makers in both, business and government

What follows, focuses on the methods used to produce the Connecticut Short-Term, Industry-Employment Forecasts. Statistical and econometric modeling is combined with pooling of forecasts and intercept corrections over the forecast horizon, based on statistical techniques, as well as expert judgment, to produce the final forecasts. The employment forecasts are produced and reconciled at three different levels: (1.) The Super-Control Forecast, (2.) the Control Forecasts, and (3.) the Detailed-Level Forecasts. Section III provides a detailed discussion of these three levels of forecasts. However, before getting into the details, Section II, below, provides an overview of the methodological steps that produce the base-line forecast, and the adjusted, final forecast of Connecticut Industry Employment. A detailed discussion of the data used in estimating the models and forecasting is presented in Appendix A.

II. PRODUCING SHORT-TERM EMPLOYMENT FORECASTS: An Overview

A. Three Levels of Forecasts

As introduced in Section I, above, there are three different levels of forecasts that are produced and reconciled: The Super-Control Forecast, the Control Forecasts, and the Detailed-Level Forecasts. Each level of forecast produces progressively more detailed forecasts. The Super-Control Forecast is the top-line level of Connecticut, Non-Agricultural Employment, and it gives the least level of detail. The Control Forecasts provide a greater level of detail. The Control Forecasts are produced at the NAICS sector level, or two-digit level of detail. Forecasts are produced for the 20 NAICS sectors, including some of their major sub-aggregates, such as Durable Goods and Non-Durable Goods under the Manufacturing Sector. Finally, the detailed-Level forecasts, as would be expected, provide the most detail. The Detailed-Level Forecasts are produced at the NAICS three- and four-digit level of detail. Forecasts are produced for some 100 three- and four-digit sectors in Connecticut. What follows below, is an overview of the methodology employed to produce each of the three levels of forecasts.

The Super-Control Forecast is a forecast of single series, the *Top-Line*. Connecticut, Non-Agricultural Employment is forecasted with a single-equation, autoregressive model, with exogenous variables, relating Connecticut Non-Agricultural Employment with past values of itself, current-period and past values of U.S. Non-Farm Employment, Capacity Utilization in Manufacturing, and deterministic components to capture seasonality, long-run trend, and structural breaks. Past versions of the model have included short-term interest rates, although the current version does not.

The Control Forecasts use different models to forecast employment at the NAICS sector-level and some sub-aggregates, such as Durable Goods under Manufacturing. Though different modeling approaches are used for different sectors and sub-aggregates, most rely on systems of equations, including multivariate time-series models. Further, some aggregates across NAICS sectors may be grouped together, where appropriate, to construct time-series models for forecasting employment. Forecasting models range from Vector Autoregressions (VAR) and Bayesian VARs (BVAR), to Near-VARs, or Seemingly Unrelated Regressions (SUR), and in some instances, single-equation models, including Autoregressive Integrated Moving Averages (ARIMA) and time-series regressions.

Detailed-Level Forecasts are at the most detailed level. These forecasts project employment at the NAICS three- and four-digit level of industry detail. They are produced in SAS/ETS, or the consortium of states' software called 'STIP' (Short-Term Industry Projections), or, in some forecast rounds, both. This level of employment forecasts is necessarily the most mechanical of the three complementary approaches, as, even for a small state like Connecticut, there are over 100 industries at the three- and four-digit NAICS levels of detail.

B. The Base-Line Forecast

The Base-Line Forecast is the product of two steps. First, forecasts are Pooled or Combined then, Reconciliation of the three forecasts is done using both*, top-down,* and *bottom-up* approaches. Once the Base-Line Forecast is in place, any Intercept Corrections are then implemented.

Pooling or Combining of forecasts is done after all three methods have been implemented. The Control-Total and Detailed-Level of forecasts present the opportunity for Pooling or Combining forecasts at the NAICS sector-level of employment. The Detailed-Level forecasts are added up to the NAICS sectors and their major sub-aggregates levels and then combined with the Control Forecasts to produce a set of simple average forecasts for each of the 19 non-Agricultural NAICS sectors. Then, the sum of the Control Totals, the Detailed-Level, and the Super Control Total are averaged together to produce the simple-average forecast for Top-Line, Non-Farm Employment. It should be noted that this is not a purely mechanical process. That is, the simple average forecast is not necessarily the final forecast for a given sector. Judgment does play a role, and one or the other forecast may be picked over the average forecast in more than one instance, especially when considering the top-line forecast.

Reconciliation of the three forecasts is done at both, the *top-down*, and *bottom-up* approaches. Three top-line forecasts are produced: (1.) the Super-Control Forecast, (2.) the Sum of the Control Forecasts, and (3.) the Sum of the Detailed-Level Forecasts. As discussed above, any pooling or combining of forecasts will be done before reconciliation.

C. Intercept Corrections: Adjusting the Forecast

Intercept Corrections are done at four different stages: (1.) If any revisions to the employment data become available after the forecasts are produced (up to a certain point), they will be used to make any necessary Intercept Corrections to put the forecasts *on track* with the historical series, (2.) Announced job eliminations and additions are used to make Intercept Corrections at the three- and four-digit NAICS industry-levels of detail, (3.) Reconciliation of the top-line forecasts with all three approaches, after announced job-changes have been incorporated, will invariably lead to further Intercept Corrections, (4.) Finally, the last of the Intercept Corrections is based on Macroeconomic considerations. This basis for adjusting the forecast is discussed in detail below.

Intercept Corrections based on Macroeconomic Considerations, more frequently than not, draws on the forecaster's expert judgment, as opposed to statistical methods. This approach is more likely to be employed if the shift or break is expected to occur over the forecast horizon (i.e., beyond the historical period). This expert judgment may be based on a number of factors, such as experience and overall belief about what drives the economy. Further, regardless of whether it is the model-based part, or the judgment-based part, every forecast is based on a set of assumptions. This set of assumptions is, in turn, guided by a theory of how the economy works, which is either incorporated into the construction of the model, or into the judgment the economist draws on to adjust the model-based forecasts, or both. Kennedy and Gunther[1] suggest that these factors will effect an economist's adjustments to his or her forecast in, at least, three different ways: (1.) How the economist views the anticipated impacts of any policy-changes that are likely to be implemented over the forecast horizon, including what variables will be affected, as well as the magnitude and duration of those affects; (2.) His or her belief about what stage of the business cycle the economy is in at the base period, and where it will be over the forecast horizon; and (3.) What his or her subjective probabilities are about the economic and non-economic risks to the forecast over the forecast horizon. And, it is these factors that play an important role in adjusting Connecticut's Short-Term Employment Forecast (i.e., correcting the Intercept) from the baseline forecast, based on Macroeconomic considerations about the current state, and likely outlook, for the U.S. and Connecticut economies.

D. The Final Forecast

The Final Forecast is the product of the process outlined above. Specifically, the above sequence of methodologies can be summarized as a four-step process. ***First***, the three levels of forecasts are produced: The Super-Control Forecast, the Control Forecasts, and the Detailed-Level Forecasts. ***The second step*** is to produce three top-line forecasts. The Super-Control, the sum of the Controls, and the sum of the Detailed Forecasts are used to produce three top-line forecasts. The simple average of the three forecasts is also considered. Then, the Controls are compared to the Detailed Forecasts' sums by NAICS sector. The two sector-level forecasts are also averaged to produce a third Control-Level Forecast. ***The third step*** is to perform both, top-down and bottom-up reconciliations of the forecasts. Upon completion of this step, the ***Base-Line Forecast*** is set.

The fourth, and final step, in producing the ***Final Forecast***, involves Intercept Corrections. ***Intercept Corrections*** are done at four different stages: (1.) If any revisions to the employment data become available after the forecasts are produced (up to a certain point), they will be used to make any necessary Intercept Corrections to put the forecasts *on track* with the historical series, (2.) Announced job eliminations and additions are used to make Intercept Corrections at the three- and four-digit NAICS industry-levels of detail, (3.) Reconciliation of the top-line forecasts with all three approaches, after announced job-changes have been incorporated, will invariably lead to further Intercept Corrections, (4.) Finally, the last of the Intercept Corrections is based on Macroeconomic considerations.

III. FORECASTING METHODS: A Detailed Discussion

A. How Do Economists Forecast?[2]

Economists, and forecasters in general, use many different methods to make their forecasts. These include more formal methods such as *Model-Based Statistical Analysis* and *Statistical Analysis* not *based on Parametric Models.* Some other techniques that economists turn to for making their forecasts include *Simple Extrapolations*, *Leading Indicators,* and ‘*Chartist*’ approaches (also called *Technical Analysis*). But, such informal methods as ‘*Back-of-the-Envelope’* calculations and *Informed Judgment* are also used. Some forecasters might even resort to some *really* informal methods such as *Tossing a Coin*, *Guessing*, or *‘Hunches’*. However, the tools most frequently used are *Econometric* and *Time-Series Models*. They are the primary methods of forecasting in economics, but *Judgment*, *Indicators*, and even *Guesses* may modify the resulting forecasts.

Time-Series models, which describe the historical patterns of data, are popular forecasting methods and they forecast well compared to Econometric Systems of Equations. Particularly, in their multivariate forms, such as *Vector Autoregression* (VAR), Time-Series models do very well. However, *Econometric Systems of Equations* are the main tool in economic forecasting. Econometric Forecasting Models are systems of relationships between variables such as GDP, Money, Employment, Inflation, etc. The relationships or ‘equations’ in these models are then estimated from the available data, which are mainly aggregate time-series. These models have three main components described in Table 1 below. Understanding and properly specifying the components depicted in Table 1 is the key to building a good forecasting model. Relationships involving any of the three components could be inappropriately formulated or inaccurately estimated, or could alter in time in unanticipated ways. Surprising results from research has shown that the critical factor to understanding forecast failure depends on *the behavior of the deterministic terms*—even though their future values are known—rather than on the behavior of variables with unknown future values (i.e., the *observed stochastic variables* in the model).

TABLE 1: **COMPONENTS OF ECONOMETRIC FORECASTING MODELS**

COMPONENT	DESCRIPTION	FUNCTION OR PURPOSE
Determinist Terms	Intercept, Trend	To capture averages and steady growth, and whose future values are known.
Observed Stochastic Variables	GDP, Prices, Employment	To capture systematic variation in movements among aggregate relationships in the economy. Their future values are unknown.
Unobserved Errors	These values are not directly observed in the economy.	To capture random influences not included in the Observed Stochastic Variables that tend to cancel each other out. All of the values (past, present, and Future) are unknown—although, perhaps estimable in the context of a model.

SOURCE: Hendry (2001), *How Economists Forecast* in UNDERSTANDING ECONOMIC FORECASTS, p. 21 and Kennedy, Peter, A GUIDE TO ECONOMETRICS, 5th Edition (2004), MIT Press: Cambridge, MA, p.p. 3-4 and 8-9.

B. The Super-Control Forecast of Connecticut Employment

The Super-Control Forecast is based on a single-equation regression model. A regression equation relates one or more *Independent* or *Explanatory* variables to a *Dependent* or *Explained* variable. That is, they explain the variation in the dependent variable. Put another way, knowing the values of the Independent variables allows one to improve on a guess of the value of the Dependent variable, over and above just using the mean to guess the Dependent variable's value. Before introducing regression, it will be helpful to look at an example of a linear relationship between two variables: X and Y:

$$Y = a + bX \tag{1.}$$

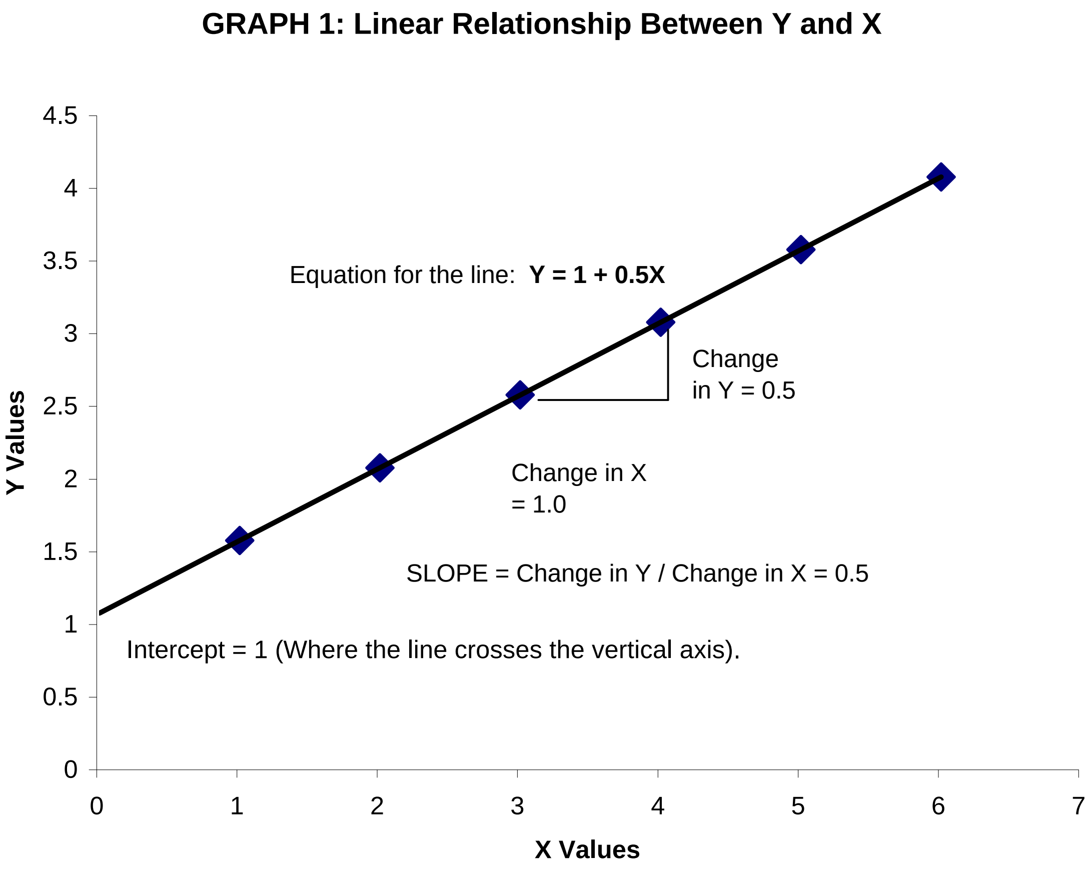
Where: a = intercept or constant

b = slope, or Change in Y, due to a 1-unit change in X

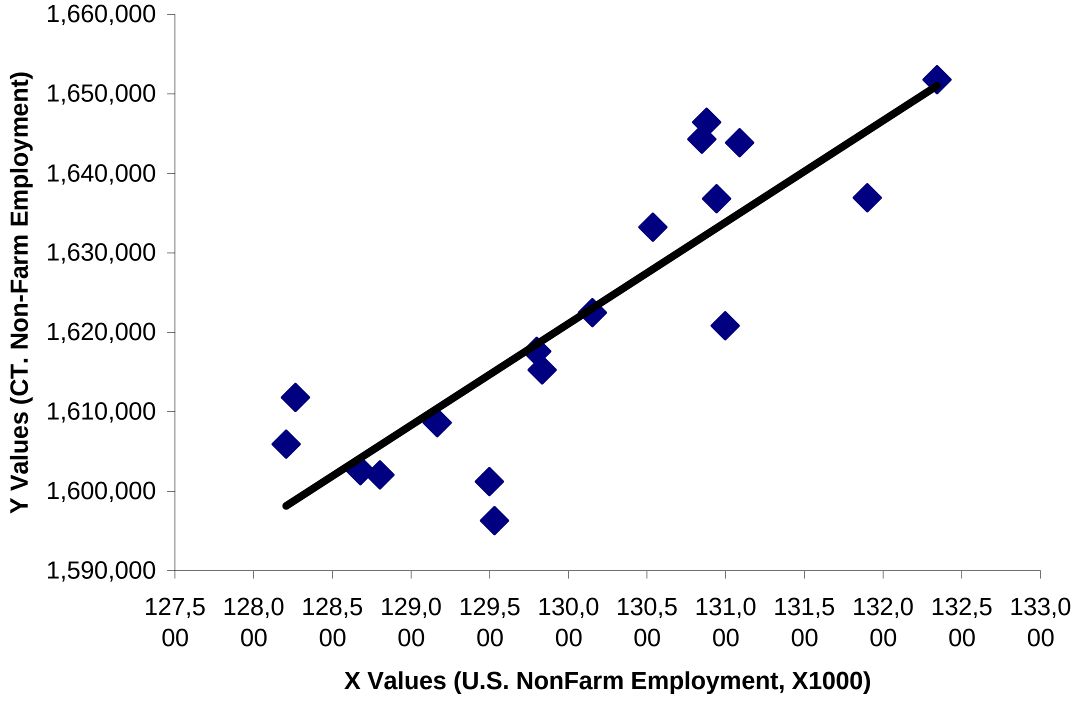
X = Independent or Explanatory variable

Y = Dependent or Explained variable

Graph 1, below, shows the relationship between X and Y in a linear model. In practice, a set of data will seldom provide the neat straight-line configuration depicted in Graph 1. This is because there are many other factors that influence the value of Y at any given point, and X will not capture them all. This situation is shown in Graph 2.



GRAPH 2: Regression Relationship Between Y (CT. Non-Farm Employment) and X (U.S. Non-Farm Employment)



In Graph 2, above, Connecticut Non-Farm Employment is on the Y-axis, playing the role of the dependent variable in this relationship. U.S. Non-Farm Employment, on the X-axis, is playing the role of independent, or explanatory variable. Though the points in the scatter seem to be grouped closely together, and seem to be sloping upward (i.e., the higher the level of U.S. Employment, the higher the level of Connecticut Employment), all of the points do not fall on the line. Why do the points in Graph 2 not line up perfectly such that they all fall on the line, as the set of points do in Graph 1? The answer lies in the variables and influences *not* included in the model. The scatter of points observed in Graph 2, as opposed to the perfectly lined-up configuration in Graph 1, reflect the many other influences, other than the level of U.S. Non-Farm Employment, that effect the level of Connecticut's Non-Farm Employment. These other influences are not captured by the independent variable in this model. But, these other influences cannot be accounted for in the deterministic formulation of the model in Equation (1.). In that formulation all points fall on the line, and X explains all the variation in Y. These other influences can only be introduced by adding an error term to the model. This results in a reformulation of Equation (1.) from a *Deterministic Model* to a *Statistical* (or *Probabilistic* or *Stochastic*) *Model*. This is illustrated in Equation (2.) below.

$$Y = a + bX + \mu \qquad (2.)$$

Equation (2.) adds a new component, μ, the ***Error*** or ***Disturbance Term***. It is the Disturbance Term that represents the influence of all variables excluded from the model, including those that are unobservable[3]. It is the factors excluded from the model that result in the scatter of points in Graph 2, rather than the perfectly lined up set of points depicted in Graph 1, which manifests a deterministic process. Now, it becomes necessary to estimate a line that runs through the scatter of points, such that it minimizes the distance between any point in the scatter, and its closest corresponding point on the line. This process is called ***Regression Analysis***. The equation expressing the relationship in Equation (2.) is called a ***Regression Equation***, or ***Regression Model***. Statistical methods are used to estimate the intercept (**a** in Equation (2.)), and the slope (**b** in Equation (2.)). Together the intercept and slope (in a *Multiple Regression* there will be more than one explanatory variable, and thus, more than one slope to estimate) to be estimated are called ***Parameters.*** The most frequently encountered method of estimating regression parameters is a method known as ***Ordinary Least Squares*** (OLS)[4].

In most fields of endeavor, it is *Stochastic,* or *Probabilistic* processes that are encountered. This is particularly true in economics and forecasting. There are three major contexts in which these relationships exist: ***Cross-Sectional***, ***Times-Series***, and ***Panel*** (which combines Cross-Sectional and Time-Series data).

In Cross-Sectional regression, the observations used to estimate the model are at a given point in time. For example, estimating a model whose data is from a survey of 1,000 households' consumption patterns, by income, across Labor Market Areas (LMA), conducted in March 2004 would be a cross-sectional regression. Observations taken at different time periods are used in a Time-Series regression. If households' consumption patterns and income, in a single LMA, were surveyed every March for ten years, say from 1994 to 2004, then the economist or researcher would be estimating a time-series regression model with data from this survey. In a ***Time-Series-Cross-Section*** (TSCS) regression, cross-section data is combined with time-series data, known as a ***Panel Study***. In a Panel Study, data is collected across several observational units through time. Continuing with the household-survey example, if households' consumption patterns and income, by LMA, were tracked every March over the 10-year period from 1994-2004, then a model estimated with this survey's data, which accounts for differences in behavior across LMA's, *and* through time, would be a TSCS, or *Panel*, regression. The most frequently encountered regression model in forecasting industry employment is *Time-Series Regression***.** Nevertheless, *TSCS Regression* may be employed in making sub-state industry-employment forecasts.

Time-Series regressions can either be ***Static*** or ***Dynamic***[5]. An example of a *Static Model* is:

$$Y_t = a + bX_t + \mu_t \qquad (3.)$$

Equation (3.) is static because if X changes, Y immediately responds and no further change takes place in Y if X then remains constant. This relationship is implied by the subscript 't'. That is, both X and Y, are in the same time period, t. This implies that the system is always observed in an *equilibrium* position. However, introducing lagged values of X change the nature of the relationship by introducing a *dynamic* element. This new relationship is expressed in Equation (4.) below:

$$Y_t = a + b_1X_t + b_2X_{t-1} + \mu_t \qquad (4.)$$

Equation (4.) is a *Dynamic Model*. Now if X increases by one unit, the expected value of Y increases immediately by b_1, but the full range of $(b_1 + b_2)$ units is only felt after one whole time period has passed. A system such as Equation (4.) is not in equilibrium. The system has been disturbed and is adjusting from one equilibrium state to another. Further, the adjustment is not instantaneous. In the case of Equation (4.), adjustment takes one whole period.

An alternative way of introducing dynamic effects into a model is by means of a lagged dependent variable. Equation (5.), below, is an example of a model with a lagged dependent variable:

$$Y_t = a + \alpha Y_{t-1} + \mu_t \qquad (5.)$$

Equation (5.) is also known as an ***Autoregression*** (AR). AR models will be discussed in more detail under Autoregressive Moving Average (ARMA) in the section on the methodology used in producing the Detailed-Level Forecasts. Equation (6.), below, combines the features of Equation (4.), a Dynamic Time-Series Regression, with Equation (5.), an AR model, to specify an ***AR model with Exogenous variables***:

$$Y_t = a + \alpha Y_{t-1} + bX_t + bX_{t-1} + \mu_t \qquad (6.)$$

.

Forecasting Connecticut's Short-Term, Top-Line, Non-Agricultural Employment: An AR Model with Exogenous Variables. The model presented as Equation (7.), below, was used in making the Super-Control Total for the second-quarter 2006, Short-Term Employment Forecast. The template for specifying the Connecticut model is Equation (6.), above.

$$\begin{aligned} \ln(E_t) = a + \Sigma^{11}{}_{i=1}\, \alpha_i S_i + b_1(\text{POST2000}) + b_2(\text{TREND}) + b_3(\text{SPLINE}) + \\ b_4\ln(Et_{-1}) + b_5\ln(\text{USNFEmp})_t + b_6\ln(\text{USNFEmp})_{t-1} + \\ b_7(\text{CURMDiff})_t + b_8(\text{CURMDiff})_{t-1} \qquad (7.) \end{aligned}$$

Where: $\ln(E_t)$ = the natural log of the level of Connecticut Non-Agricultural Employment for period t.

$\ln(E_{t-1})$ = Autoregressive term (Endogenous variable) representing the natural logs of past levels of Connecticut Non-Agricultural Employment in period t-1.

S_i = Seasonal Dummies to capture seasonal variation in Connecticut Non-Agricultural Employment.

POST2000 = A dummy variable coded ‘1’ for time periods after December 2000 to capture the effects of the bursting of the Stock-Market Bubble, collapse of Business Investment Spending, the 2001 Recession, the September 11th Attacks, and the Corporate scandals that all followed in the subsequent period. It is coded ‘0’ for the sample period preceding January 2001.

TREND = A linear time index representing the long-term economic and demographic forces effecting the secular growth rate in Connecticut’s Non-Agricultural Employment.

SPLINE = This represents the structural break in the Connecticut Non-Agricultural Employment series at July 2000. It is the point at which U.S. Business Investment spending collapsed, and U.S. Industrial Production began contracting. The effects on Connecticut’s Labor Market were almost immediate. The State’s employment-cycle expansion peaked at this point, and then began declining afterward.

CURMDiff_t, CURMDiff_{t-1} = Exogenous variables representing the current level of the Capacity Utilization Rate (CUR) Difference in Manufacturing, and the lagged level at period t-1.. It is formed by subtracting the CUR, for a given period for Manufacturing, from the long-run average CUR in Manufacturing.

$\ln(USNFEmp_t)$, $\ln(USNFEmp_{t-1})$ = Exogenous variables representing the level of U.S. Non-Farm Employment for the current period, and lagged one period.

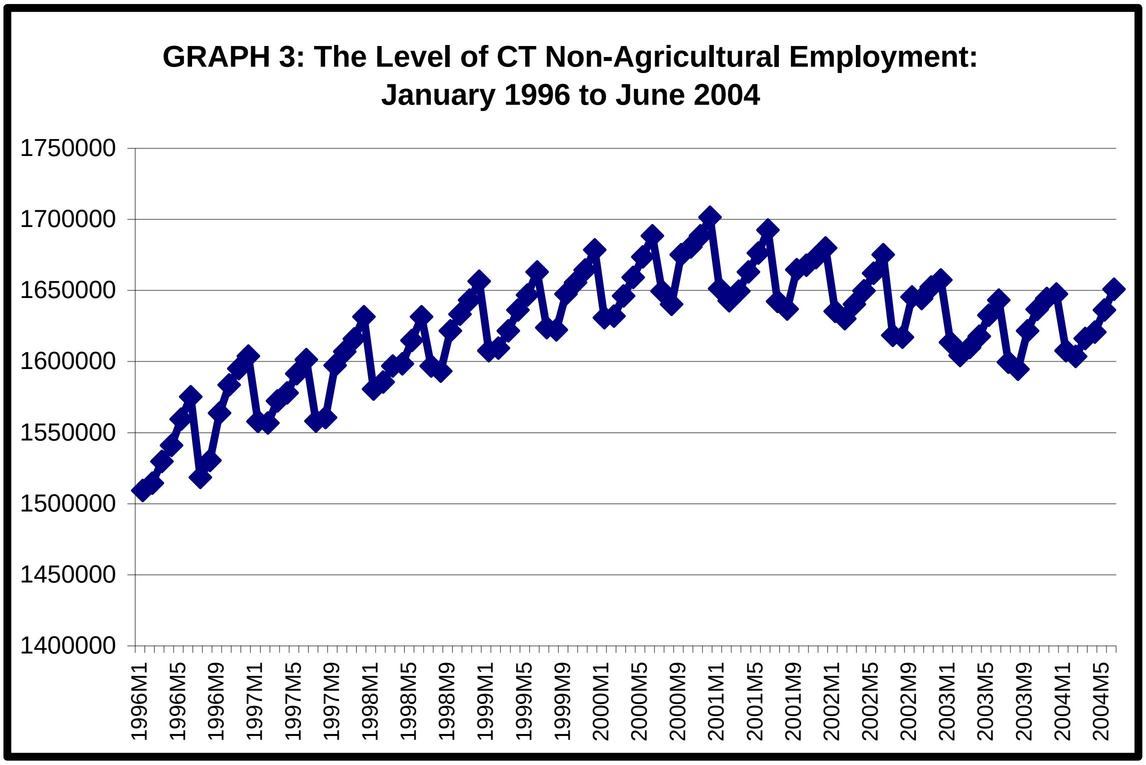
α, b_1- b_7 = The Model Parameters

t = Subscript ‘t’ indicates the time period.

The model was estimated using RATS Version 6. At this point, it is critical to briefly digress and discuss a few econometric software issues.

The software developed by the consortium of states, lead by Illinois and Utah, to provide states’ Labor Market Information units with a tool to produce relatively uniform short-term, industry employment forecasts is the **Short-Term Industry Projections** (STIP) system. The forecaster using the STIP system has five models to choose from: Exponential Smoothing (with several options), Single-Equation Linear Regression, univariate time-series models: Autoregressive Moving Average (ARMA); multivariate time-series models: Vector Autoregression (VAR) and Bayesian VAR (BVAR)[6]. Mix gives a weighted average forecast based on the five models available in the STIP system. The Connecticut Department of Labor (CTDOL) uses the STIP system to produce the Detailed-Level forecasts. An alternative set of Detailed-Level forecasts may also be made using the SAS/ETS Forecasting Procedure or the Interactive Forecasting Menu. However, the Super-Control and Control forecasts are produced using RATS, and possibly EViews in the future. Since the STIP system automatically internally forecasts any imported exogenous variables, previously forecasted exogenous variables are ignored in estimating the models. Consequently, to use outside vendors’ forecasts of exogenous variables (or, even if internally forecasted), in building and estimating forecasting models, RATS, EViews, or SAS must be used. For this reason, the Super-Control and Control forecasts cannot be produced using the STIP system.

With the above considerations in mind, the discussion now turns to the details of Equation (7.) above. The model was estimated using monthly data, though the final forecasts are published in quarterly form. In regard to the model, the first feature to note is the transformation of the dependent variable Connecticut Non-Agricultural Employment, E_t. It appears in the model as a natural log transformation. The actual form of the path through time of Connecticut Employment is non-linear. This is depicted in Graph 3, below.


GRAPH 3: The Level of CT Non-Agricultural Employment:
January 1996 to June 2004

Since the forecasting model is a ***linear*** multiple regression, it is necessary that the model be statistically estimated in linear form. For example, if a functional relationship between Connecticut employment, and its past value were linear, but, its relationship to a time trend were exponential, then this would be expressed as Equation (8.), below:

$$E_t = A^*e^{(TREND)} + E_{t-1} \qquad (8.)$$

Again, referring to Graph 3, above, the long-term trend in the level of Connecticut Non-Agricultural Employment is clearly non-linear. But, Equation (8.) must be transformed into a linear form in order to be estimated as a linear regression. By taking the natural log of both sides of Equation (8.), its linear version is obtained, and it is expressed as Equation (9.) below:

$$\ln(E_t) = \ln(A) + \text{TREND} + \ln(E_{t-1}) \qquad (9.)$$

The multiplicative expression, $A^*e^{(\text{TREND})}$, in Equation (8.) has been trans formed into the two separate linear and additive terms, ln(A) and TREND in Equation (9.) through a log transformation. This is known as the ***Log-Log Linear form.*** Equation (9.) is now in a form where it can be estimated as a linear regression model. This is the same transformation that was applied to Connecticut's Super-Control Forecast Model. The result was the form of the Super-Control Model presented as Equation (7.), above.

Turning to the specifics of the Connecticut Forecasting Model, Table 2, below, reproduces Table 1, but within the context of the Connecticut Model. As discussed in Section II, results from research have shown that the critical factor to understanding forecast failure depends on the behavior of the deterministic terms, and whether or not they have been accurately captured in the forecasting model. As shown in Table 2, the Connecticut model captures the ***Deterministic terms*** by including a constant, *Seasonal Dummies* (S_i, where i= 1, 2, …, 11)[*] to capture the seasonal cycle, since the industry employment data are not seasonally adjusted, the TREND variable to capture the long-run growth-rate in Connecticut's Non-Agricultural Employment, and two variables to capture structural change. The SPLINE variable captures the structural change in the growth-path of Connecticut after the collapse in Business Investment and the contraction in Industrial Production in July 2000.

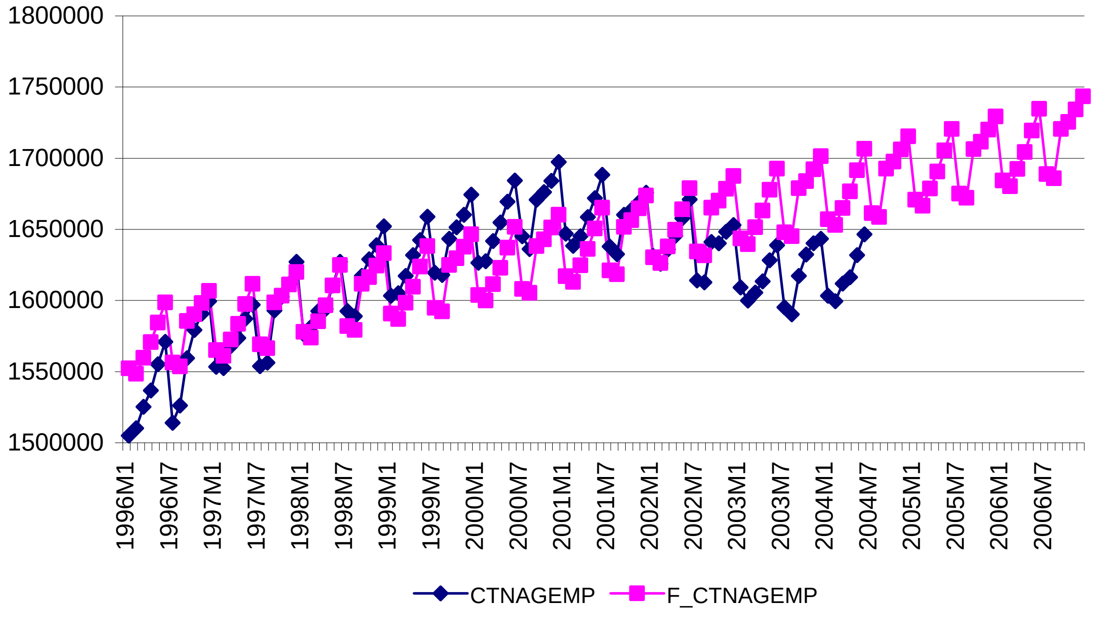
[*] Though there are twelve months in the year, only 11 seasonal dummies are included in the model because it includes an intercept. If there were no intercept in the model, then the forecaster would include 12 dummies to capture the seasonal variation within a year for monthly data.

TABLE 2: COMPONENTS OF CONNECTICUT'S FORECASTING MODEL

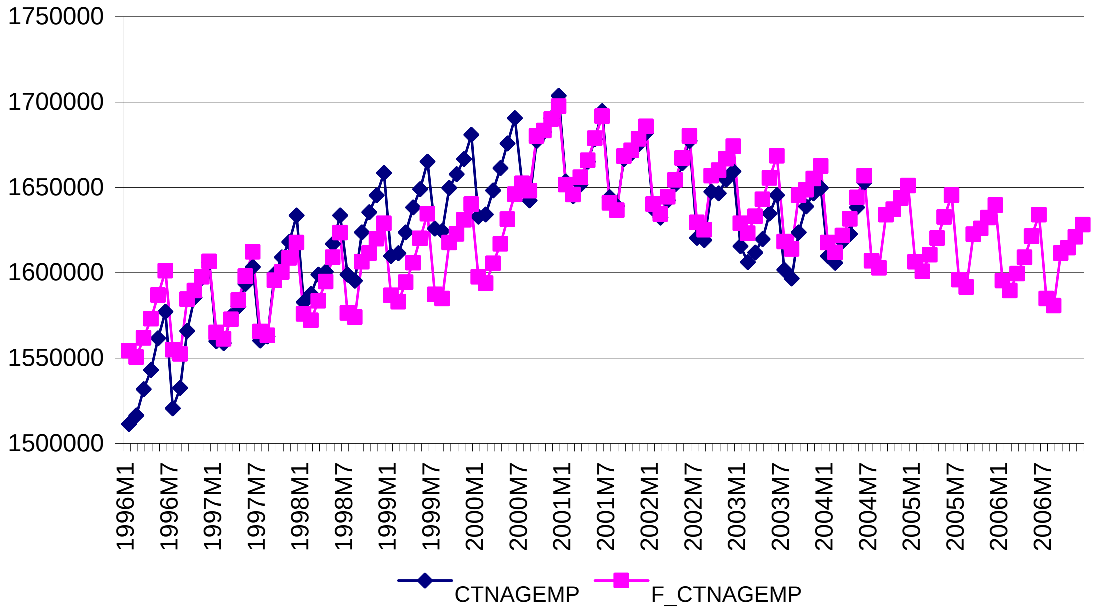
COMPONENT	VARIABLES	FUNCTION OR PURPOSE
Determinist Terms	INTERCEPT, S_i, TREND, SPLINE, POST2000	To capture seasonal cycles, averages and steady growth over the long-run, and structural shifts in the State's Economy. Their future values are known.
Observed Stochastic Variables	$Ln(E_{t-1})$, $CURMDiff_t$, $CURMDiff_{t-1}$, $ln(USNFEmp_t)$, $ln(USNFEmp_{t-1})$	To capture systematic variation in movements among aggregate relationships in the U.S, and Connecticut economies. Their future values are unknown.
Unobserved Errors	μ	To capture random influences not included in the Observed Stochastic Variables that tend to cancel each other out. All of the values (past, present, and future) are unknown—although, perhaps estimable in the context of a model.

This had an immediate impact on Connecticut, as employment peaked in July, and then declined. Also capturing the shift in the economy after 2000, is the dummy variable POST2000. To illustrate how critical the deterministic components of the model are to forecasting, two examples are provided in Graphs 4 and 5, below. Graph 4 illustrates the forecast failure that would occur if, in addition to the intercept and seasonal dummies, the model only included a linear trend (i.e., the variable TREND) to capture long-run factors effecting Connecticut's employment growth-rate. Without accounting for the structural break in the economy that occurred in the year 2000, the model predicts Connecticut's Non-Agricultural Employment to continue on its pre-2000 trajectory, and as a consequence, the forecast seriously misses the mark. Graph 5 depicts the forecast of employ after accounting for the structural breaks in the year 2000. Without even including the stochastic variables in the model, it tracks and forecasts well.

GRAPH 4: CT Short-Term Employments Forecasts: An Example of Forecast Failure



GRAPH 5: CT Short-Term Employments Forecasts: An Example of Successful Capture of the Structural Components



The critical nature of getting the deterministic components of the model right motivated an approach to forecast model-building that follows a modification of Francis X. Diebold's approach to forecasting in the second edition of his text, *Elements of Forecasting*, published in 2001 by Southwestern[7]. His approach is to first, account for seasonality and trend in the specified forecasting model. At this point, the forecaster should be left with an error series with a very nice cyclical pattern. To complete the process, Diebold then uses autoregressive terms to model the cyclical component of the example time series. Though he did not use this sequential approach to model the classical time-series components in his discussion of regression, it was literally followed in the approach to building and estimating the Super-Control model. Once the seasonal and trend components, as well as any structural breaks, have been isolated, the forecaster is then left with a cyclical pattern in the error series (as discussed above). At this point, the opportunity then exists to focus attention on those exogenous factors in the U.S., Connecticut, and regional economies that account for the observed movement in the level of employment over the business cycle.

To account for the systematic variation in movements among aggregate relationships in the U.S, and Connecticut economies, over the business cycle, the Connecticut model includes the following ***Observed Stochastic Variables***: $\ln(E_{t-1})$, CURMDiff_t, CURMDiff_{t-1}, $\ln(\text{USNFEmp}_t)$, and $\ln(\text{USNFEmp}_{t-1})$. The lagged value of Connecticut Employment, $\ln(E_{t-1})$, accounts for the influence the level of the immediate past period of employment has on the current-period's level of employment. However, the previous period's employment does not have a 100% influence on the current level of employment. Its influence is discounted by b_4, the regression coefficient for $\ln(E_{t-1})$, in Equation (7.), the Super-Control Model. This implies that the absolute value of b_4 must be less than one. In notation, this would be expressed as:$| b_4 | < 1$. Further implications of this requirement will be discussed below in Section III on ARMA. It should also be noted at this point, that since the dependent variable and the AR term are both in logs, rather than levels, the regression, or slope coefficient, b_4, is now interpreted as the ***percent*** job-change in the current period, due to a one-percent change in jobs at period t-1, *not* the change in the *number* of jobs. The variables, CURMDiff_t and CURMDiff_{t-1}, appear in the model as levels, not logs, thus, the percent change in Connecticut Employment is interpreted as the result of a *percentage-point* change in the CUR difference. CURMDiff_t and CURMDiff_{t-1} capture the influence of the current and lagged-period level of capacity utilization in U.S. Manufacturing, relative to its average utilization rate over the long-run (defined by the Federal Reserve Board as the 1972-2003 Period). As of December 2004, the long-run utilization rate for Manufacturing is 80.1%. Thus, for a given current period, if the actual capacity utilization exceeded its long-run average, then $\text{CURMDiff}_t > 0$. If the utilization rate at the current period (period t) were below the long-run average utilization, then $\text{CURMDiff}_t < 0$. The same results would apply to CURMDiff_{t-1}. If there is a one percentage-point increase in CUR Difference in U.S. Manufacturing, the percent change in Connecticut Employment increases immediately by b_7, (see Equation (7.) above), but the full range of the $(b_7 + b_8)$ percent change is only felt after one whole time period has passed, or in the case of the Connecticut Model, since it is monthly data, after one whole month has passed. The strong positive relationship between the level of U.S. Non-Farm Employment and Connecticut Non-Farm Employment has already been introduced in the scatter plot in Graph 2, above. The two exogenous variables, $\ln(\text{USNFEmp}_t)$, and $\ln(\text{USNFEmp}_{t-1})$, which are in log form, represent the influence of the level of U.S. Non-Farm Employment on the level of Connecticut Non-Farm Employment over the business cycle. Since the U.S. Employment variable is in log form, the regression coefficients, b_5 and b_6 , are interpreted as indicating the percent-change in Connecticut's Employment due to a one-percent change in U.S. Employment in the current month, or the immediate past month. Thus, if U.S. Non-Farm Employment increases by one percent, Connecticut Employment increases immediately by b_5, but the full range of $(b_5 + b_6)$ percent increase in jobs is only felt after one whole month. Having defined and discussed the variables in the model, the section below turns to the estimation results.

Model Estimation and Forecasting. The model-estimation results are presented in Tables 3 and 4. Though significance levels are asterisked for p-values below the 10%, 5%, and 1% probability of a Type I (α-level), their statistical significance is not critical when using a model for forecasting.

TABLE 3: PARAMETER ESTIMATES FOR SUPER-CONTROL MODEL

Variable		Coeff	StdErr	T-Stat	Signif
1	Constant	-1.0059	0.4306	-2.3359	0.0208**
2	FEBRUARY	-0.0031	0.0065	-0.4725	0.6372
3	MARCH	0.0043	0.0069	0.6279	0.5310
4	APRIL	0.0021	0.0072	0.2897	0.7724
5	MAY	0.0032	0.0074	0.4326	0.6659
6	JUNE	0.0063	0.0068	0.9265	0.3557
7	JULY	-0.0114	0.0034	-3.4126	0.0008***
8	AUGUST	-0.0005	0.0058	-0.0860	0.9316
9	SEPTEMBER	0.0142	0.0072	1.9774	0.0498**
10	OCTOBER	0.0006	0.0068	0.0954	0.9242
11	NOVEMBER	0.0056	0.0062	0.9016	0.3687
12	DECEMBER	0.0093	0.0055	1.6875	0.0935*
13	POST2000	-0.0260	0.0129	-2.0130	0.0459**
14	TREND	-0.0004	0.0001	-3.0050	0.0031***
15	SPLINE3	0.0002	0.0001	2.1437	0.0336**
16	LNCTNAGEM{1}	0.8525	0.0390	21.8838	0.0000***
17	LNUSNFEMP	1.2791	0.2559	4.9977	0.0000***
18	LNUSNFEMP{1}	-1.0111	0.2436	-4.1506	0.0001***
19	CURMDIFF	0.0010	0.0007	1.4081	0.1611
20	CURMDIFF{1}	-0.0015	0.0007	-2.1628	0.0321**

0.05 < p <= 010* 0.01 < p < 0.05** p <= 0.01***

TABLE 4: ESTIMATION STATISTICS FOR SUPER-CONTROL MODEL

Linear Regression Estimation by Least Squares

Dependent Variable LNCTNAGEM

Monthly Data From 1990:02 To 2004:06
Usable Observations 173
Degrees of Freedom 153

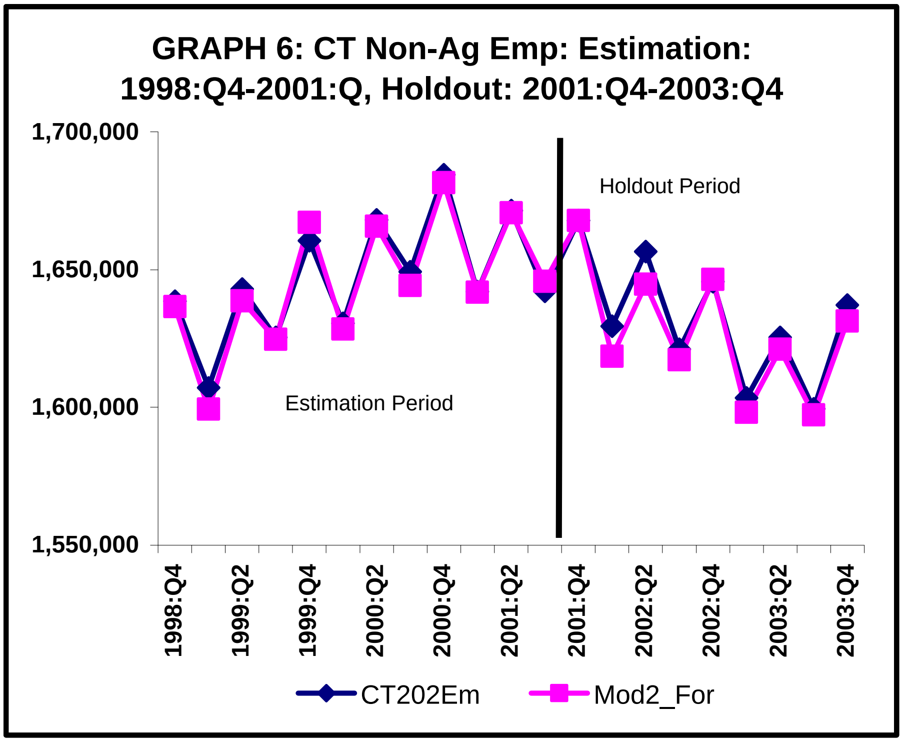
Centered R**2 = 0.99409
R Bar **2 = 0.993357

Mean of Dependent Variable = 14.27078
Std Error of Dependent Variable = 0.0399579
Standard Error of Estimate = 0.003257
Sum of Squared Residuals = 0.001623
Durbin-Watson Statistic = 1.98219

Nevertheless, save CURMDiff$_t$, all the coefficients for most of the structural components, the AR term, and exogenous variables were significant at an α-level of 10%, or lower. The Adjusted R^2, 0.99 is to be expected for a time-series regression model. In fact, an

Adjusted R^2 less than 0.90, in the time-series context, may suggest a problem with the model. The Durbin-Watson statistic of 1.98 indicates that there is no problem with first-order autocorrelation in the errors. Although there appears to be autocorrelation in the error series five lags back, and the Ljung-Box Q-Statistic of 47.61, with a p-value < 0.001 (not shown in the tables), indicates that the error series is not *white noise* (the concept of White Noise is discussed below in Section III). Whether or not this presents a problem with the model depends on how it forecasts.

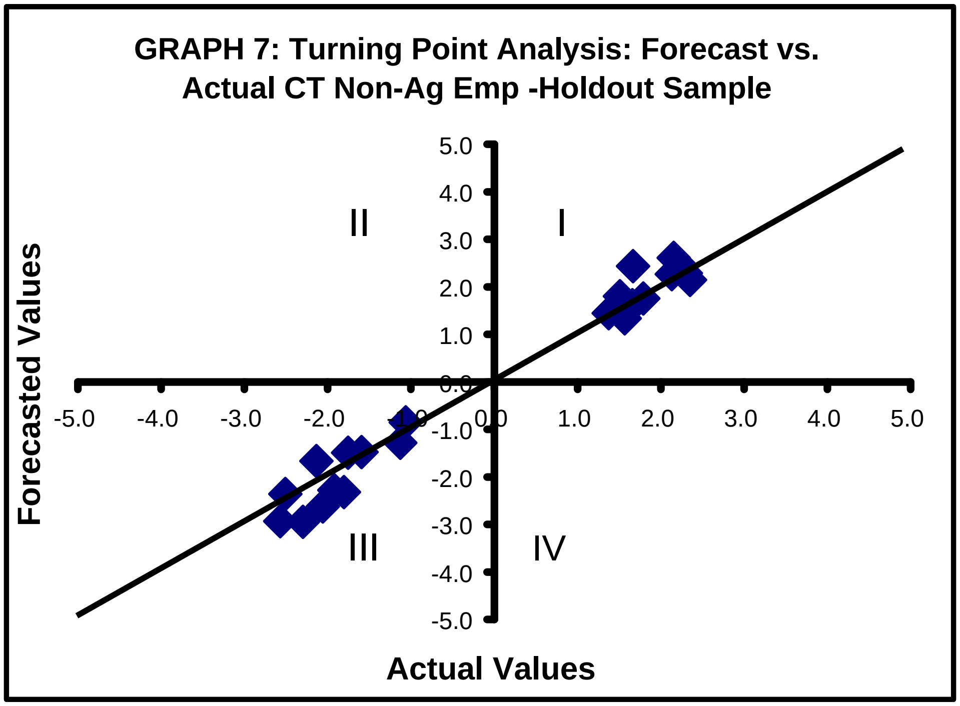
To test the ability of the model to forecast, the model was re-estimated, but only using data from January 1990 to December 2001. Data over the period January 2002-December 2003 were held out to test how well the model forecasts out of sample. The results appear in Graph 6. As indicated above, the forecasts themselves are expressed quarterly. Therefore, in Graph 6, the estimation period is 1998:Q1-2001:Q4, and the holdout period is 2001:Q4 to 2003:Q4.



The forecast over the holdout period is known as an ***Ex Post*** forecast since both the endogenous variables and the exogenous, explanatory variables are known with certainty.

Thus, the *ex post* forecast can be checked against the existing data and provide a means of evaluating the forecasting model[8]. Whereas, an ***Ex Ante*** forecast predicts values of the independent variable, in this case, the level of Connecticut Employment, beyond the estimation period using explanatory variables that may or may not be known with certainty. Also, the *ex post* forecast, based on the holdout sample, is an ***Unconditional Forecast*** since the values of the explanatory variables are known with certainty. When the actual forecast is made, forecasts of the exogenous variables are obtained from an outside vendor (see Appendix A). This means that the actual forecast is a ***Conditional Forecast*** since the values of the explanatory variables are not known with certainty.

In addition to the ***Time-Series Forecast*** shown in Graph 6, which projects future values of a time-series, the forecaster may also want to evaluate the model's ability to make ***Event-Timing Forecasts.*** In this case, there is an event that is certain to happen, but its timing is unknown[9]. Within the context of economic and labor-market forecasting, this would involve forecasting turning points in the business and employment cycles. One of the ways to evaluate an Event-Timing forecast is with a ***Turning-Point Error Diagram***[10]. The Turning-Point Error Diagram for the Time-Series Forecast presented in Graph 6 is depicted in Graph 7. In Graph 7, the actual turning points are measured along the horizontal axis, and the forecasted turning points are measured along the vertical axis. The 45-degree line represents the locus of points that would obtain from making perfect forecasts. The quadrants are numbered I to IV in a counterclockwise direction. Points in Quadrants I and III represent correctly identifying turning points. Points in Quadrant II represent instances where the model predicts false turning points. Points in Quadrant IV are instances where the model missed actual turning points. Graph 7 indicates that the model, at least based on graphical analysis, did fairly well in predicting actual turning points, and it avoided predicting false turning points.



GRAPH 7: Turning Point Analysis: Forecast vs. Actual CT Non-Ag Emp -Holdout Sample

As helpful as the above graphical evaluation has been for evaluating the forecasting model, the next step should be to a more precise evaluation using quantitative measures of forecast performance. In fact both, graphical and quantitative methods should be used to evaluate forecast performance. Some of the more frequently used quantitative tools for evaluating the performance of a forecast are presented in Table 5, below. The criteria used most often to evaluate the Connecticut models are the BIAS (or Mean Error), MAE (Mean Absolute Error), MPE (Mean Percent Error), MAPE (Mean Absolute Percent Error), and the RMSFE (Root Mean Square Forecast Error). Particularly, the MAPE, and comparing the difference between the RMSFE and the MAE are used to evaluate a model's forecasts. Since, the Mean Square Forecast Error (MSFE) penalizes large errors, a RMSFE much larger than the MAE would signal large errors at some observations.

Two aspects of the Connecticut model were tested using the above evaluation criteria: the within-sample model fit, and the model's out-of-sample forecasting ability. The results are presented in Tables 6 and 7 below.

***TABLE 5:* Quantitative Forecast Evaluation Criteria**

Evaluation Statistic	Description	Comments
BIAS (Mean Error)	$1/n\ \Sigma^{n}{}_{i=1} (F_t - Y_t)$	Measure the average forecast error expressed in the units of the forecasted variable. Also referred to as the Bias, because its sign will indicate whether, on average, the model is overforecasting (-) or underforecasting (+).
MPE (Mean Percent Error)	$1/n\ \Sigma^{n}{}_{i=1} [(F_t - Y_t)/ Y_t]$	The Bias expressed in relative terms.
MAE (Mean Absolute Error)	$1/n\ \Sigma^{n}{}_{i=1} \lvert F_t - Y_t \rvert$	This is the mean of the absolute value of the forecast errors. It measures the absolute size of the average forecast error in the units of the forecasted variable.
MAPE (Mean Absolute Percent Error)	$1/n\ \Sigma^{n}{}_{i=1} [\lvert F_t - Y_t \rvert / Y_t]$	This is the MAE expressed in relative terms.
RMSFE (Root Mean Square Forecast Error)	$[1/n\ \Sigma^{n}{}_{i=1} (F_t - Y_t)^2]^{1/2}$	This takes the square root of the Mean Square Forecast Error, thereby translates it back into the original units of the forecasted variable. This is the one forecast-evaluation statistic that will not be used to compare forecast performance across models. It is compared with the MAE within a given model. If the RMSFE is much larger than the MAE then there are some large forecast errors.

SOURCE: Kennedy, Daniel W, PROPOSED SEEMINGLY UNRELATED REGRESSIONS (SUR) OPTION FOR THE STIP FORECASTING SYSTEM (October 2002), Office of Research, Connecticut Labor Department: Wethersfield, CT.

***TABLE 6:* Within-Sample Model Fit**

PERIOD	Bias	MPE	MAE	MAPE	RMSFE
First Seven	1,944	0.12	3,804	0.23	4,503
Middle Seven	757	0.05	3,894	0.24	4,444
Last Eight	-302	-0.02	2,437	0.15	3,546
Full Period	750	0.05	3,335	0.20	4,160

***TABLE 7:* Out-of-Sample Model Performance**

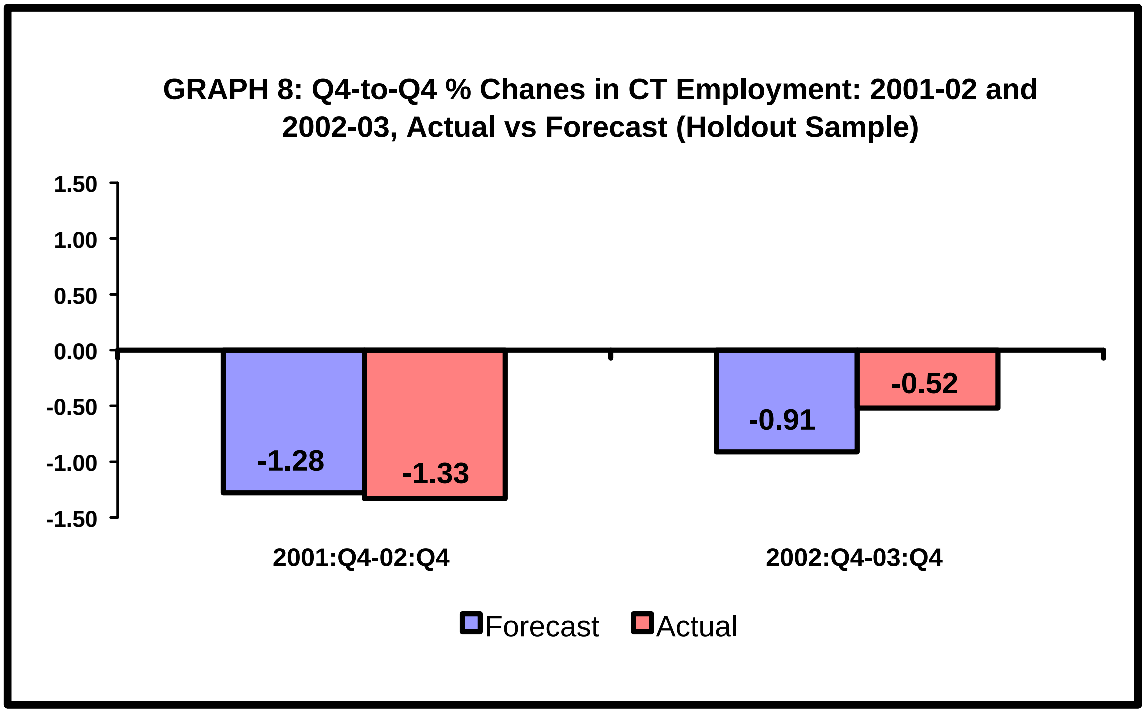
PERIOD	Bias	MPE	MAE	MAPE	RMSFE
First Seven	1,796	0.11	3,697	0.23	4,441
Middle Seven	2,316	0.14	3,352	0.20	4,857
Last Seven	4,638	0.29	4,861	0.30	5,854
Full Period	2,917	0.18	3,970	0.24	5,085
Holdout Period	**5,421**	**0.33**	**5,615**	**0.35**	**6,695**

Table 6 presents the results of full-sample estimation of the Connecticut, Super-Control Forecasting Model. Historical data from January 1990 to December 2003 were used to estimate the parameters. As discussed above, the final results are expressed quarterly by taking the monthly average for the three months comprising each quarter. The last 21 observations were evaluated. They were broken up into three equal periods. The full range was also evaluated. The model seems to fit the data fairly well. The MAPE is 0.20%. Though the relative Bias is very small at 0.05%, the larger RMSFE relative to the MAE indicates that, though they tend to cancel each other out, there are some relatively larger errors at a few observations. A good sign is that the MAPE declines over the last eight evaluation quarters indicating that the model is *learning* as it gets closer to the end of the historical data, a critical region for a forecasting model. In summary, the model seems to fit the historical series quite well, but can it forecast? Table 7 turns to addressing that question.

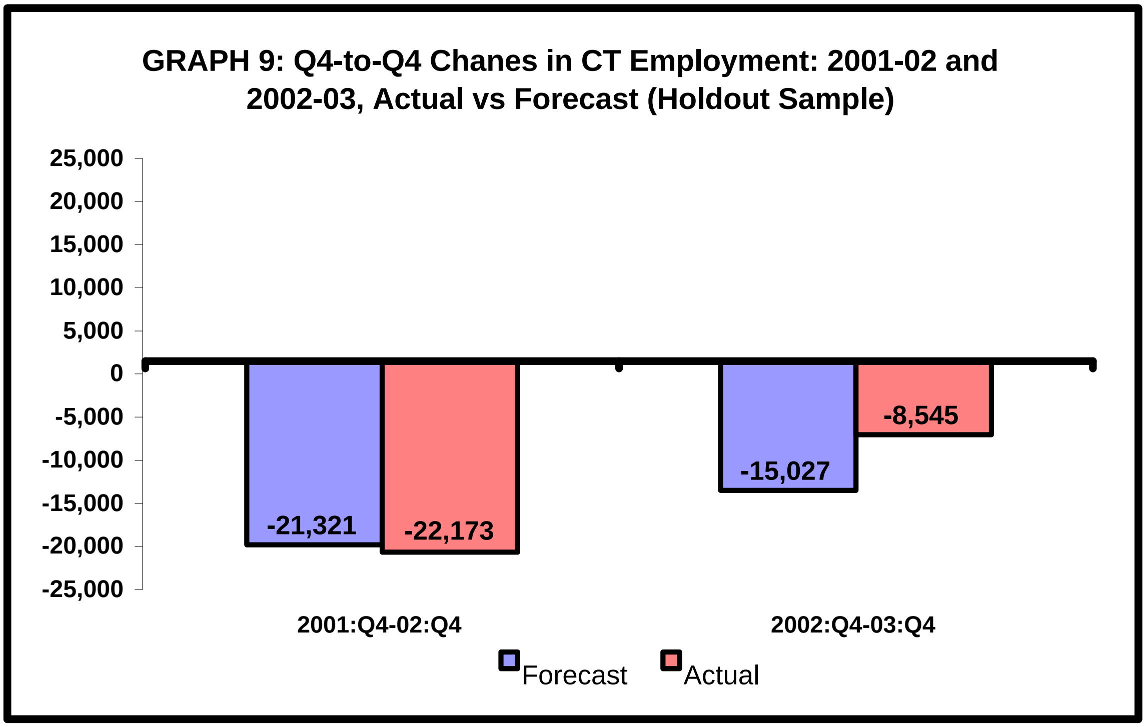
After evaluating the fit, the model was re-estimated with a holdout sample. That is, only the data from January 1990 to December 2001 were used to estimate the parameters. The data from January 2002 to December 2003 were held out to test the model's ability to forecast out-of-sample. Again, the forecasts are presented as quarterly data. Analysis of the historical fit of the holdout is in black font. Out-of-sample results are in blue boldface. The overall MAPE at 0.24% is still a good performance, though it deteriorated slightly over the last part of the historical range. The Bias increased slightly to 0.18%, and the difference between the MAE and the RMSFE increased slightly. These results are not surprising since information is lost when estimating with a holdout sample. Critical to the model's ability to forecast are the out-of-sample evaluation criteria, in blue boldface, in the last row of Table 7. The period covered in the last row of Table 7 is the January 2002-December 2003 Holdout Period, which, again, are expressed in, and evaluated in, quarterly terms. The MAPE for the *ex post* forecast period is 0.33%. The relative Bias (MPE) is 0.33%. The absolute difference between the MAE and RMSFE is about the same. And, since their magnitudes are slightly larger over the holdout period, the relative difference declined. Based on the results in Table 7, the model seems to be forecasting fairly well. Graph 6, above, illustrated the time-series forecast for the holdout model. Graphs 7 and 8, show the forecasts versus the actual for the percent change in jobs and the change in jobs, measured fourth-quarter-to-fourth-quarter over the 2001-2003 *Ex Post* Forecast Period for 2002 and 2003.

From Graphs 8 and 9, it is apparent that the model did a much better job forecasting over the first year of the forecast period, relative to the second year. This is to be expected. As the forecasted period becomes further from the mean of the series, which contains the most sample information,[11] it moves out of the range of experience used to estimate the model. Thus, the further out the forecast goes the larger the error. Forecasting too far beyond the historical range is perilous.

Finally, the actual values and the forecasts, over the holdout period are presented in Table 8, below. The next section turns to the methodology used in producing the Control Forecasts.



GRAPH 8: Q4-to-Q4 % Chanes in CT Employment: 2001-02 and 2002-03, Actual vs Forecast (Holdout Sample)



GRAPH 9: Q4-to-Q4 Chanes in CT Employment: 2001-02 and 2002-03, Actual vs Forecast (Holdout Sample)

TABLE 8: Actual Values vs. Forecasts of CT Employment, Holdout Period-2001: Q4-2003:Q4

Date	CT Employ	Forecast	Fore Diff Act-Fore	%Diff Act-Fore
2001:Q4*	1,667,862	1,667,788	74	0.0044
2002:Q1	1,629,535	1,618,639	10,896	0.6732
2002:Q2	1,656,557	1,644,696	11,862	0.7212
2002:Q3	1,621,203	1,617,321	3,882	0.2400
2002:Q4	1,645,689	1,646,467	-779	-0.0473
2003:Q1	1,603,407	1,598,128	5,279	0.3303
2003:Q2	1,625,446	1,621,223	4,222	0.2604
2003:Q3	1,599,515	1,597,217	2,298	0.1439
2003:Q4	1,637,144	1,631,440	5,704	0.3496

*Last Historical Data Point, which serves as the Base Period for the Eight-Quarter Holdout (*Ex Post*) Forecast.

Beware the Ides of March (Minus Three Days)

The Ides of March, minus three days, is March 12th. Why is this date important? It is important, because, March 12th is the date of the annual Benchmarking of the Establishment Survey data. The Establishment Survey is based on a statistical sample of employers drawn from the Unemployment Insurance (UI) Tax database called the Quarterly Census of Employment and Wages (QCEW), formally known as the ES-202. Benchmarking is done each year to account for any changes in information on the birth and death of establishments, and in employment and wages, that may have occurred between the time the sample was drawn and administered, and when the Non-Farm Employment series is benchmarked. This brings up the issue of *revisions to the data* and their effects on the forecast. In the following passage, Hendry and Clements relate data revisions to Intercept Corrections to address Forecast Failure:

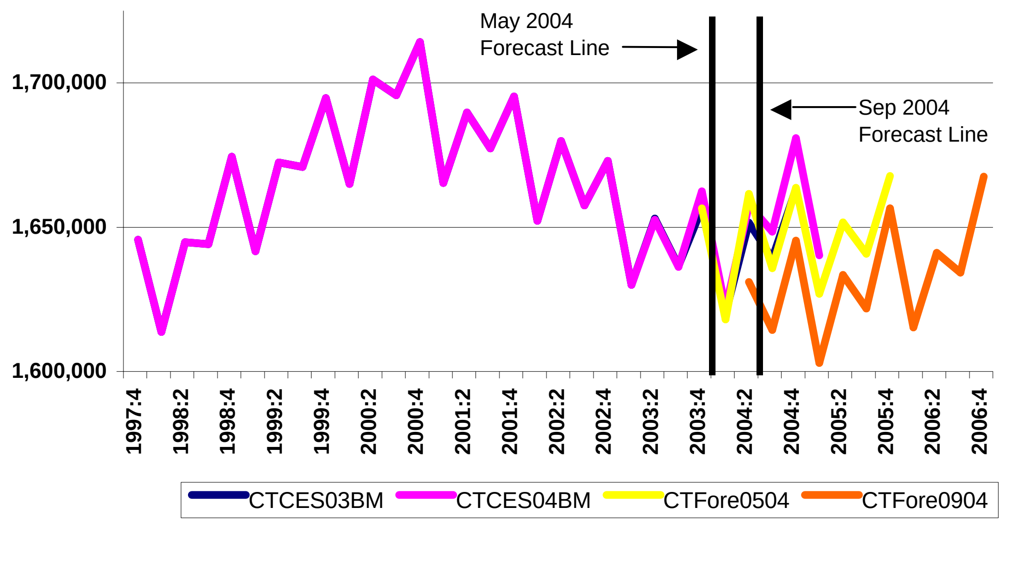
> Revisions to ‘first-release’ data are often substantial relative to the growth of the variables being forecast, confirming the benefits of appraising all available information about the forecast origin, and suggesting ‘smoothing’ IC’s, but a formal analysis is not yet available. (Hendry, David F. and Michael P. Clements, *Economic Forecasting: Some Lessons from Recent Research*, October 22, 2001, ECB Conference on Forecasting Techniques, p. 19.)

The March 2005 Benchmarking of the Establishment Survey resulted in some significant revisions to the preliminary 2004 employment estimates for a number of states, including Connecticut. Significant revisions from benchmarking are more likely to occur when the economy enters a turning point in the business cycle. In 2004 and 2005, the Economy experienced more than one turn, as it went from the 'Soft Patch' of the first half of 2004, to the expansion of the last half of 2004, to 'Soft Patch II' in the middle of 2005. These are the very conditions that are conducive to producing larger revisions during the benchmarking process. And, it affected the Short-Term forecast of Connecticut Employment, especially for the fourth quarter of 2004 (2004:Q4). Thus, even if, as in the last section, the forecast evaluation indicates that the forecast is on track, it does not necessarily mean that forecast failure has been avoided. If what is called '*Optimality Theory*' held in forecasting practice, then economists' forecasts would probably all pretty much be on the mark This theory of forecasting relies on two key assumptions: (1.) The model is a good representation of the economy, and (2.) The structure of the economy will remain relatively unchanged. But, as Clements and Henry observe:

> Unfortunately, empirical experience in economic forecasting has highlighted the poverty of these two assumptions. Such an outcome should not be a surprise: all econometric models are mis-specified, and all economies have been subject to important unanticipated shifts…(p.4, *Economic Forecasting: Some Lessons from Recent Research*, 2001.)

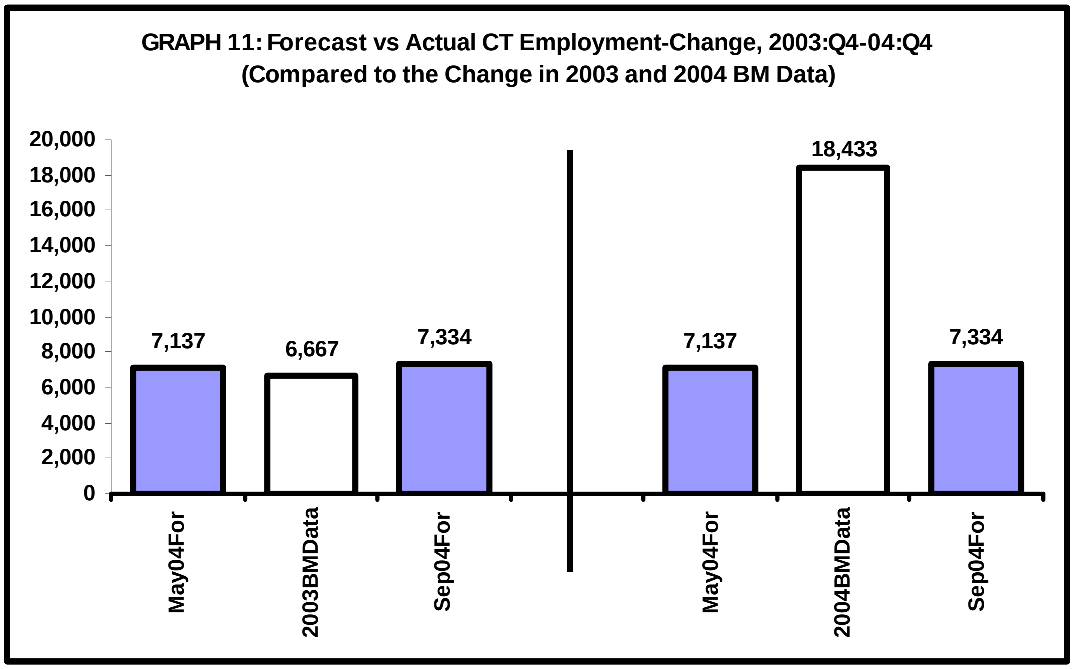
The Connecticut Short-Term Employment forecasts covering the four quarters for the 2004 calendar year illustrate the point. Graph 10, below, presents the quarterly time-series of the two forecasts that include forecasts for fourth-quarter 2004 Connecticut Employment: the May 2004 Short-Term Forecast, and the September 2004 Short-Term forecast. In addition, both forecasts are compared to the initial 2004 estimates based on the 2003 Benchmarked (BM) time-series data, and the final values (through the third quarter of 2004), based on the 2004 BM time-series data. As depicted in Graph 10, below, it is in the second half of 2004 where the 2003 BM and 2004 BM data part company, especially in the fourth quarter.

GRAPH 10: May and Sep 2004 Forecasts - 2003 vs 2004 BM Data



As it turned out, based on quarterly time-series data, the May 2004 Time-Series, or Extrapolation, Forecast actually tracked the 2004 BM data better than the September 2004 Time-Series Forecast. And, it tacked the 2003 BM data very well. Both forecasts expected growth to accelerate in the last half of 2004, however, both underestimated the magnitude of that acceleration, especially for the fourth quarter. In addition to comparing the forecasted levels to the actual levels, it is also instructive to compare the forecasted growth in employment to the actual growth in employment.

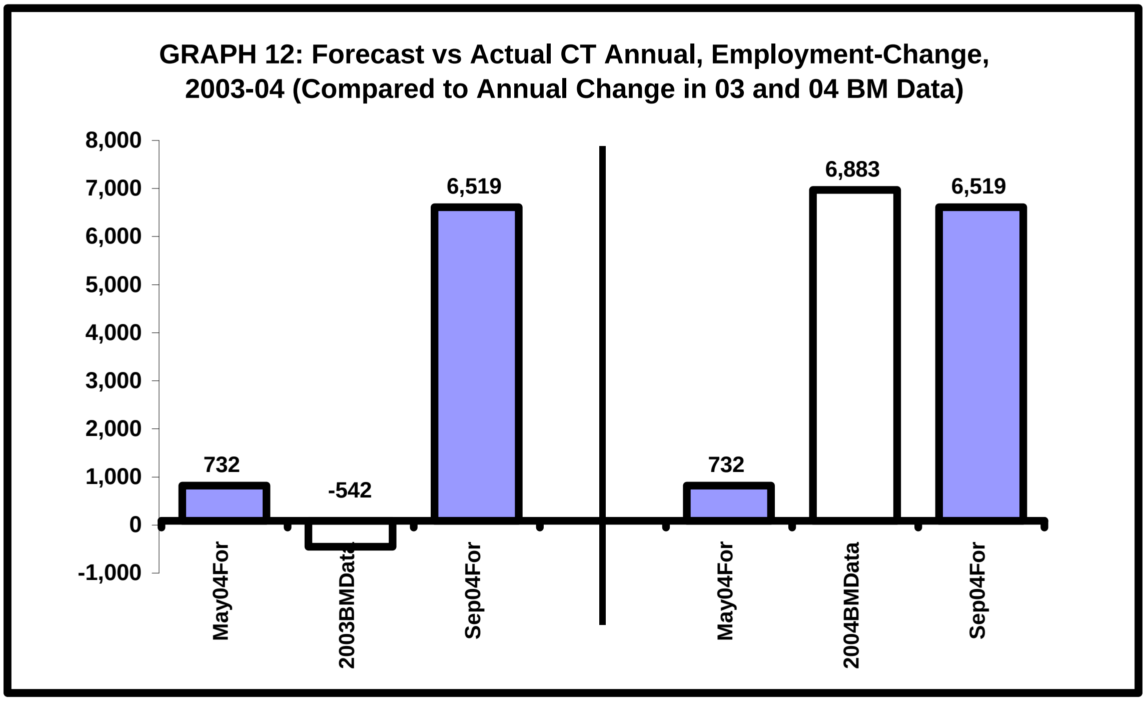
Graph 11 below, compares Connecticut employment-growth between the fourth quarter of 2003 and the fourth quarter of 2004, as forecasted by the May 2004 Forecast, and the September 2004. Both fourth-quarter-to-fourth-quarter forecasts are also compared to the growth based on the 2003 BM and 2004 BM data. Graph 12, below, makes the same comparisons, except it is based on the forecasts and benchmarked data for annual-average employment-growth between 2003 and 2004.



It is clear from Graph 11 that, both, the May 2004 and September 2004 forecasts were very close to predicting the 2003 to 2004, fourth-quarter employment-growth, based on the 2003 BM data. However, the 2004 BM data show that both forecasts significantly underestimated Connecticut's employment growth in the fourth quarter of 2004, on a Year-to-Year (YTY) basis. However, the results are different when looking at the predicted change in annual employment.

Turning to Graph 12, the 2003 BM data showed a slight decline in Connecticut's Annual Employment between 2003 and 2004. The May 2004 Forecast predicted a slight gain. On the other hand, the September 2004 Forecast expected an annual increase of 6,500 between 2003 and 2004. Though it is not a particularly impressive gain, it was a significant over-estimate of 2003-04 annual employment-growth, based on the 2003 BM data, however, it was right in line with the annual employment-growth based on the 2004 BM data. To sum up, based on the 2004 BM, it appears that both the May and the September 2004 forecasts significantly underestimated the surge in job-growth over the fourth quarter of 2004. Nevertheless, the September 2004 Forecast did closely predict

Connecticut's growth in the average, annual level of employment between 2003 and 2004.



The above results make it clear, that in addition to mis-specifying, or omitting deterministic components from a forecasting model, significant revisions to the data can also result in forecast failure. Such an outcome should not be a surprise, as all econometric models are mis-specified, and all economies are subject to important unanticipated shifts. Particularly, turning points in the business cycle can result in structural changes in the time-series, and subsequent large revisions to the data.

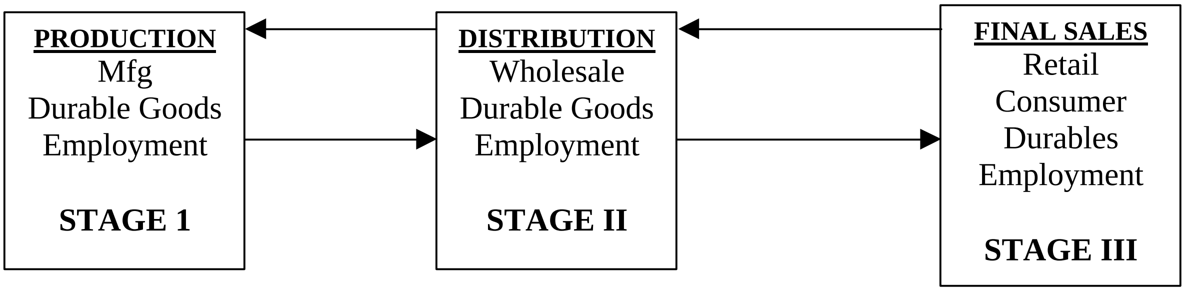
One final note on this issue: now that the QCEW data is available sooner, the effects of BM revisions should be reduced. For instance, the March 2005 Establishment Survey BM had available QCEW data up to September 2004. This compares with June for the year before, and March for all previous years. However, this still leaves the last quarter of the year's data as estimates until the next year's benchmarking.

C. The Control-Total Forecasts of Connecticut Employment

The next level of forecasting detail, moving from the Super-Control Total down to more detail, is the set of Control-Total Forecasts. The control-total level of detail requires the estimation of several different forecasting models. In most instances, these models draw on the interrelationships between and among related industries to produce several industry-employment forecasts simultaneously. As would be expected, considerably more effort is put into building and estimating, both, the super-control total and the Control-Total Forecasts than for the Detailed-Level Forecasts.

In order to capture inter-industry relationships, many of the Control-Total Forecasts are produced using *multivariate time-series methods*. Particularly, *Vector Autoregressions* (VAR) are used in many instances. This allows forecasting models to draw on economic linkages and interconnectedness to construct feedback systems that tap into the direct and indirect effects of employment-changes in a given industry on other, related industries. An example of a grouping of industries for forecasting the Control Totals is the link or chain of Durable Goods sectors. A VAR constructed to capture this relationship would contain endogenous variables for each stage along the production chain. This idea is depicted in Diagram 1, below.

***DIAGRAM 1*: Durable Goods Industry Chain: Recursive (Feedback) Mechanism**



In Diagram 1, say flat-screen TVs are produced at the plant in Stage I of the *Production-Distribution-Final-Sales Chain*. They are then shipped to the warehouse-distribution center in Stage II. At Stage III in the chain, the TVs are delivered to the retail outlets and purchased by consumers. However, there is also a feedback, because, if sales fall, then the retailer will reduce his or her inventory demand from the warehouse. This, in turn, will result in the warehouse-distribution center reducing its shipments from the plant. Finally, the plant will cut back on production and produce fewer runs. If, instead, the retailer increased his or her orders for flat-screen TVs, then the opposite set of signals and adjustments would be transmitted through the chain. Changes in the signals sent by firms, at a given stage of production-distribution-sales, ripple back and forth through the chain, which causes firms at each stage to adjust their employment and output to meet each new set of business conditions. It is precisely this kind of feedback mechanism that is well suited to a VAR formulation. Other relationships also exist, such as, firms interacting at the same stage of production, and interconnections at the same stage of production, and at different stages, simultaneously. Much more detail on inter-firm and inter-industry connections can be found in the literature on combining VAR's with Input-Output Analysis[12] and Industry Clusters[13]. The next section turns to a detailed discussion of VAR models and their role in the Connecticut Control Forecasts.

Vector Autoregression[14] The *Vector Autoregression* (VAR) model is a widely used tool in econometrics today, especially for forecasting. Developed by Sims (1980)[15], it was motivated as an answer to the large number of, what he called, 'haphazard' restrictions imposed on the equations that make up large multi-equation Macroeconomic models. He proposed a new, and, what was then, radical alternative. He advocated an approach that would estimate large-scale macroeconomic models as unrestricted reduced forms, treating all variables as endogenous. Since then, the VAR approach has been widely adapted as an econometric tool used for hypothesis-testing, impact analysis, and forecasting in the areas of Finance, Macroeconomics, Regional Economics, and many others. The following discussion presents the basics of the VAR as a forecasting tool.

The VAR can be thought of as a generalization of the AR process, (see the discussion of the Super-Control Forecast, above), to two or more AR processes. Thus, a VAR is a system of two or more simultaneous equations expressing two or more interrelated AR processes. Central to the VAR, as introduced above, is the concept of a ***Recursive or, Feedback Relationship***. For example, say there are two time-series, y_t and z_t, and both are AR(1) processes like the one encountered in Section B above. But now, y_t is not only dependent on its own past value, y_{t-1}, but also on the current and past values of z_t. Likewise, z_t is dependent, not only on z_{t-1}, but also on y_t and y_{t-1}. This *recursive* relationship can be expressed as follows:

$$y_t = a_{10} + a_{11}\, y_{t-1} + a_{12}\, z_{t-1} + e_{1t} \qquad (10.)$$

$$z_t = a_{20} + a_{21}\, z_{t-1} + a_{22}\, y_{t-1} + e_{2t}$$

It is assumed that e_{1t} and e_{2t} are serially uncorrelated but the Covariance, Cov(e_{1t} e_{2t}), need not be zero. If the variances and covariance are time invariant, then the Variance-Covariance matrix can be written as:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Where: Var(e_{it}) = σ_{ii} and Cov(e_{1t} e_{2t}) = $\sigma_{12} = \sigma_{21}$

The right-hand side of the VAR equations contains only pre-determined variables. Since the error terms are serially uncorrelated with constant variances, *each equation in the system can be estimated using OLS*. Moreover, OLS estimates are consistent and asymptotically efficient. Even though the errors are correlated across equations, estimation using Seemingly Unrelated Regressions (SUR) does not add to the efficiency of the estimation procedure since both regressions have identical right-hand side variables. (This issue will be taken up at length in Part C, below).

Example

To give an example from the industry short-term, employment forecasting context, and drawing on the production-chain concept discussed above, the model below contains two endogenous variables: the NAICS Industry 452, General Merchandise Stores in the Retail Trade Sector (44-45), and, its associated industry is NAICS Industry 424, Merchant Wholesalers-Non-Durable Goods in the Wholesale Trade Sector (42). A VAR forecasting model reflecting the relationship between employment in the General Merchandise Stores industry and employment in Merchant Wholesalers, assuming one effects the other with a one-month lag, could be expressed as follows:

$$E^{GM}_{t} = a_{10} + a_{11} E^{GM}_{t-1} + a_{12} E^{MW}_{t-1} + e_{1t} \qquad (11.)$$

$$E^{MW}_{t} = a_{20} + a_{21} E^{MW}_{t-1} + a_{22} E^{MW}_{t-1} + e_{2t}$$

Where: E^{GM}_{t} = Employment in General Merchandise stores at time period (month) t.

E^{MW}_{t} = Employment in Merchandise Wholesalers at time period (month) t.

a_{10}, a_{10} = Intercepts (Constants) of the two regression equations.

$a_{11}, a_{12}, a_{21}, a_{22}$ = Regression (Slope) Coefficients of the two regression equations.

As discussed above, the VAR is composed of AR processes in a recursive system in which all variables are endogenous to the process. That is, all variables are determined within the specified system of equations that make up the VAR. Variables determined within the system are called ***Endogenous.*** In the above example, both General Merchandise Employment (E^{GM}_{t}) and Merchandise Wholesalers Employment (E^{MW}_{t}) are *endogenous* variables. However, there may be instances in which the forecast of employment may be improved by including variables that have been determined outside the system. These variables are known as ***Exogenous*** Variables. Exogenous variables may be stochastic (e.g., Income) or Non-Stochastic (e.g., Seasonal Dummy Variables). For example, to improve on forecasts of General Merchandise and Merchant Wholesale employment, exogenous variables representing certain, common, or shared, economic phenomena that may effect employment in both industries, such as Interest Rates, Income, and a set of dummy variables to capture seasonal effects on these industries activities. The new model would be expressed as:

$$E^{GM}_t = a_{10} + a_{11}\, E^{GM}_{t-1} + a_{12}\, E^{MW}_{t-1} + \Sigma^{11}_{i=1}\, \alpha D_i + \gamma Y + \varphi R + e_{1t} \qquad (12.)$$

$$E^{MW}_t = a_{20} + a_{21}\, E^{MW}_{t-1} + a_{22}\, E^{MW}_{t-1} + \Sigma^{11}_{i=1}\, \alpha D_i + \gamma Y + \varphi R + e_{2t}$$

Where: E^{GM}_t = Employment in General Merchandise stores at time period (month) t.

E^{MW}_t = Employment in Merchandise Wholesalers at time period (month) t.

a_{10}, a_{20} = Intercepts (Constants) of the two regression equations.

a_{11}, a_{12}, a_{21}, a_{22} = Regression (Slope) Coefficients for the endogenous variables in the two regression equations.

α, γ, φ = Regression (Slope) Coefficients for the exogenous variables in the two regression equations.

$\Sigma^{11}_{i=1} D_i$ = Eleven Seasonal Dummies representing any seasonal effects on employment in the two industries. (Why only 11? This follows the rule that, with an intercept in the model, the number of dummies equals one minus the number of categories.)

Y= Income

R= the appropriate interest rate.

The above extension of the VAR model is known as a *Dynamic Simultaneous Equations Model*[16], or Dynamic SEM. Notice, in the above, expanded model, that the two employment series are determined within the system, while Income and Interest Rates are determined from outside the model, and not by the model itself. That is, there are no equations in the system for Income and Interest Rates. The dummies for seasonal effects are deterministic (i.e., non-stochastic) variables.

The VAR models, and extensions, discussed above, are predicated on the use of *Classical Statistical* methods to estimate their parameters. Within the Classical framework, an *unbiased* estimator is considered desirable because, as more and more samples are taken, the average value of the sample estimates tends toward the value of the unknown population parameter. In the class of unbiased estimators, a *minimum variance* estimator is preferred, because, *on average*, it yields values that are closer to the real parameter than those obtained from any other unbiased estimator. Basically, evaluation takes place within a repeated sampling context because classical analysis prefers techniques with a high *probability* of giving the correct result, and probability is defined in terms of the *limit of a relative frequency*.

Bayesian Vector Autoregression[17] In a *Bayesian* framework, probability is *defined in terms of a degree of belief*. And, although the properties of estimators and tests in repeated samples are of some interest, they do not provide the main basis for inference and estimator choice. The probability of an event is given by an individual's belief in how likely or unlikely the event is to occur. This belief may depend on qualitative or quantitative information, or both, but it does not necessarily depend on the relative frequency of the event in a large number of future hypothetical experiments. Further, in a Bayesian framework, *parameters are treated as random variables.* However, this is not to be construed as the notion that different values of the parameter are obtained as a result of different outcomes of an experiment, but, instead, as the idea that there is a subjective probability distribution associated with a parameter that describes the state of knowledge about that parameter. In the classical framework, because a parameter is fixed in repeated samples, a probability distribution cannot be assigned to the parameter.

The Bayesian subjective probability distribution on a parameter summarizes an individual's knowledge about that parameter. The knowledge may exist before observing any sample information. This is known as the ***Prior Distribution***. If knowledge is derived from both prior and sample information, then it is reflected in the ***Posterior Distribution***. A posterior distribution in relation to some past sample can be regarded as a prior distribution in relation to a future sample. In either case, the subjective distribution is the source of all inferences about the unknown parameter. In contrast to the Classical approach that concentrates on point estimates, the final objective in a Bayesian investigation is often attainment of the Posterior Distribution. The procedure that combines a Prior distribution with sample information to form a Posterior distribution is known as *Bayes's Theorem*[18]. Bayes's Theorem is discussed in detail in Appendix B. The BVAR approach provides an objective and reproducible procedure for combining a forecaster's beliefs and data. For the reasons mentioned in Appendix B, this particular system of Bayesian priors is known as the *Minnesota System of Prior Beliefs,* or the ***Minnesota Prior***.

After completing the usual process of choosing the variables to be included in the VAR, the prior beliefs about the values of each of the coefficients in the equations in the VAR system can be expressed in the form of probabilities about which set of values will give the best forecasts. In the Minnesota Prior, these probabilities can be described by assigning a best guess and a measure of confidence to each coefficient in the model. Both of these guesses would be quantitative (i.e., a number). The best guess is set according to the *Random Walk Hypothesis*. This hypothesis states that variables behave in such a way, that changes in their values are unpredictable. For such a variable, the best forecast of its one-step ahead value is equal to its current value. To implement the random walk hypothesis, the best guesses of the Minnesota Prior are that all coefficients in the equation, save the most recent value, are zero. The coefficient for the most recent value is guessed to be 1. In addition, the forecaster must supply a quantitative measure of confidence in each best guess. This is expressed as the *Prior Variance of the Coefficient*. The smaller the prior variance, the more confidence the forecaster has that his or her best guess will be close to the forecast. With one exception, the system then proceeds in two stages. First, the forecaster selects a few restrictions that group the prior variances and mainly determine the relative sizes of the prior variances within each group. Second, the forecaster selects a range of possible values for a scale factor that completes the determination of the prior variances. The one exception to the two-stage process is the procedure for determining the prior variance of the constant terms (intercepts) in each equation. These variances are simply set to vary large numbers, which amounts to saying that, at least over a very large range, the forecaster regards all possible values of the constant term as almost equally likely. In other words, the forecaster is willing to allow the constant term to be determined by the data alone.

A specific example will help in understanding how this procedure works. To begin with, the two-industry equation system from the VAR(1) example above is reproduced below, except another lag has been added. Now, it is a VAR(2). Further, the perspective is from the one-step-ahead forecast. Thus, the dependent variable now becomes E^{GM}_{t+1} rather than E^{GM}_{t}.

$$E^{GM}_{t+1} = a_{10} + a_{11} E^{GM}_{t} + a_{12} E^{GM}_{t-1} + a_{13} E^{GM}_{t-2} + a_{14} E^{MW}_{t} + a_{15} E^{MW}_{t-1} + a_{16} E^{MW}_{t-2} + e_{1t} \quad (13.)$$

$$E^{MW}_{t+1} = a_{20} + a_{21} E^{MW}_{t} + a_{12} E^{MW}_{t-1} + a_{23} E^{MW}_{t-2} + a_{24} E^{GM}_{t} + a_{25} E^{GM}_{t-1} + a_{26} E^{GM}_{t-2} + e_{1t}$$

Where: E^{GM}_{t} = Employment in General Merchandise stores at time period (month) t.

E^{MW}_{t} = Employment in Merchandise Wholesalers at time period (month) t.

a_{10}, a_{10} = Intercepts (Constants) of the two regression equations.

a_{11} - a_{16} and a_{21}- a_{26} = Regression (Slope) Coefficients of the two regression equations.

In the above equation, the first restriction takes the form of weights that shape the prior variances of the coefficients of current and past values of the given variable. These values are known as *Direct*, or *Own* lags (of the variable that the given equation forecasts). For the General Merchandise Employment equation they are: E^{GM}_{t}, E^{GM}_{t-1}, and E^{GM}_{t-2}. For the Merchant Wholesalers Employment equation they are: E^{MW}_{t}, E^{MW}_{t-1}, and E^{MW}_{t-2}. The Minnesota Prior asserts that the less important a variable is believed to be for forecasting, the greater the forecaster's confidence in his, or her, best guess of its coefficient (i.e., that its value is zero). Since more recent values of the variable are considered to be more important for forecasting future values than those further into the past, the prior variances of the direct lags should get smaller, or tighter, around the best guess, as the number of lags increases. As the lag length increases, this feature of the Minnesota Prior, which combines the random-walk best guess with increasing confidence that the coefficients are zero for the direct-lag variables, will lead to good forecasts. The restriction is imposed by weighting each direct-lag variance by 1/(k + 1), where k equal to the number of lags. In the E^{GM} Equation, this means that the prior variances of the coefficients for E^{GM}_{t-1} and E^{GM}_{t-2} are one-half and one-third as large as the prior variance of the coefficient of E^{GM}_{t}.

In the equation that forecasts a given variable, again using the E^{GM} Equation as an example, the second restriction takes the form of weights that shape the prior variances of the current and past values of all variables, besides the given variable. In the E^{GM} Equation, these variables are: E^{MW}_{t}, E^{MW}_{t-1}, and E^{MW}_{t-2}. These values are known as *Cross-Lags.* The prior variances of the coefficients of the cross-lags have the same relative sizes as the coefficients of the direct lags. In addition, the coefficients of the cross-lags are each weighted by a *direct-versus-cross* variance factor, which gives the cross prior variances units that are comparable to the direct prior variances.

The first stage of the determination (i.e., the combined effect of the random walk and best guess of the confidence levels), results in a wide probability distribution for the current value, which puts a high probability on the chance that the parameter value could be far from the best guess. The distributions for the lagged values of the variables become tighter and more peaked as the lag-length increases. This reflects the low probability assigned to the parameter value being vary far from the best guess. Further, it reflects the forecasters belief that as lag-length increases, he or she is increasingly confident that a zero coefficient will be consistent with a model that forecasts well.

Once the relationships between the parameters for the direct- and cross-lag variables has been specified, the next step to complete the specification of the prior variances, is to pick a number, a *scale factor*, called a ***Hyperparameter***, for each group of parameters. That Hyperparameter would simultaneously multiply all the weights assigned to the coefficients in the group and convert these weights from *relative* to *absolute* prior variances.

To complete the second stage (assuming the forecaster was certain of the absolute size of at least one of the variances within each group of relative variances), the appropriate Hyperparameter would be assigned to each group, completing the specification of the prior probabilities (i.e., best guesses and variances) of the model's coefficients. However, instead of picking a single probability distribution for the model's coefficients, the forecaster specifies a group of similar probability distributions, one for each setting of the Hyperparameters, and treats all the distributions within the group as equally likely. Standard Bayesian statistical procedures would then be applied to the data to compute revised (***Posterior***) coefficient probabilities for each possible setting of the Hyperparameters. The final coefficient probabilities, and hence, the final forecast, would be formed as a weighted average of these, with the weight attached to each proportional to the probability that the setting of the Hyperparameters that generated it is consistent with the historical data.

In the labor-market forecasting environment, sets of labor-market and economic data can consist of the same cross-sectional sample and reflect outcomes of economic and labor-market relations that exist at different points in time. In addition to time, geographical areas (e.g., labor market areas) and employment in related industries are two examples that give rise to the need for partitioning sample observations and thus defining a set of economic relations. Because these economic relations may have parameters that vary over time (e.g., months, quarters, years) and space (e.g., states, regions, labor market areas), these properties need to be recognized when specifying and estimating forecasting models. Particularly, if separate, single-equation regression models were used to forecast employment for the two industries (General Merchandise Retailers and Merchant Wholesalers) in the VAR and BVAR examples above, then common economic circumstances faced by these related industries (income, consumer sentiment, interest rates, etc.) would be implicitly reflected in their error terms. If estimated by Ordinary Least Squares (OLS) as a single-equation regression, these factors and their common economic circumstances, reflected in the error terms, cannot be captured, and the OLS estimates are inefficient. And, of course, OLS assumptions have been violated. Also, not captured, is the recursive relationship between the two employment series in the above VAR and BVAR example, since there is only one equation. Because the rest of the equations belonging to this system are 'hidden' when only one equation is estimated, Zellner[19] referred to this phenomenon as *Seemingly Unrelated Regressions.* The next section turns to this system for forecasting multiple time-series.

Seemingly Unrelated Regressions (SUR) or Near-VAR. As was made clear in the previous section, the Vector Autoregression (VAR) has many advantages as a forecasting tool. However, one disadvantage is the 'one-size-fits-all' approach. However, there is a more flexible approach. The SUR approach was first suggested by Arnold Zellner (1962)[20] in the early 1960's.

As discussed above, grouping industries according to similarities in the behavior of their employment dynamics can be captured by taking advantage of the Vector Autoregression (VAR) specification. Extensions of the VAR to the Dynamic SEM framework allows the introduction of exogenous variables into the model to account for seasonality, business cycles, industry-specific factors, and other influences external to the recursive relationship reflected in the endogenous variables of the VAR system. However, the VAR specification assumes that the matrices of independent variables across all equations are the same and, that contemporaneous correlation among the error series across equations is minimal or nonexistent.

However, in some cases, gains in forecasting accuracy may be realized by allowing for differences in the size of the matrices of independent variables across equations, and for taking into account instances of significant contemporaneous correlation. This is especially important in regard to the set of exogenous variables. Under certain circumstances, the restriction to a ‘one-size-fits-all’ specification of the exogenous variables in the conventional VAR framework, compromises the ability to produce more accurate forecasts.

The motivation for this approach arises from the adoption of a modification of Francis X. Diebold’s approach to forecasting in the second edition of his text, *Elements of Forecasting*, published in 2001 by Southwestern. As detailed in Section II, above, his approach is to first, account for seasonality and trend in the specified forecasting model[**]. At this point, the forecaster should be left with an error series with a very nice cyclical pattern. To complete the process, Diebold then uses autoregressive terms to model the cyclical component of the example time series. Though he did not use this sequential approach to model the classical time-series components in his discussion of regression or VAR, his approach is adapted to specifying regressions and, within the multi-equation framework, Seemingly Unrelated Regressions (SUR). Following the Equation (6.) template, in Section II, specifications are autoregressive models with exogenous variables. (This type of specification, technically, recasts the model from a static SUR system to a dynamic ***Near-VAR system***). Once the seasonal and trend components have been isolated, the forecaster is then left with a cyclical pattern in the error series (as discussed above). At this point, the opportunity then exists to focus attention on those exogenous factors in the U.S. and Connecticut economies, and those specific to that industry, that account for the observed behavior of employment over the business cycle.

The advantage offered by the SUR specification lies in its ability to capture structural breaks that frequently occur at different points, or may not even apply to some series in the system. Further, one equation may have statistically significant seasonality, while

[**] And, as also detailed in Section II, getting the deterministic components of the model ‘right’ is critical to avoiding catastrophic forecast failure.

another may not. An example is the modeling and forecasting of the control totals (in this case, at the NAICS three-digit level), for Connecticut's wholesale trade employment series. While the durable goods component displayed no discernible seasonality, there was a strong seasonal movement in the non-durable Goods employment series. Both employment series displayed structural breaks at the same point, and had similar trends.

$$E^{GM}_t = a_{10} + a_{11} E^{GM}_{t-1} + a_{12} E^{MW}_{t-1} + \Sigma^{11}_{i=1} \alpha D_i + \gamma Y_t + \gamma Y_{t-1} + e_{1t} \quad (14.)$$

$$E^{MW}_t = a_{20} + a_{21} E^{MW}_{t-1} + a_{22} E^{MW}_{t-1} + \Sigma^{11}_{i=1} \alpha D_i + \varphi R_t + \varphi R_{t-1} + e_{2t}$$

Where: E^{GM}_t = Employment in General Merchandise stores at time period (month) t.

E^{MW}_t = Employment in Merchandise Wholesalers at time period (month) t.

a_{10}, a_{20} = Intercepts (Constants) of the two regression equations.

$a_{11}, a_{12}, a_{21}, a_{22}$ = Regression (Slope) Coefficients for the endogenous variables in the two regression equations.

α, γ, φ = Regression (Slope) Coefficients for the exogenous variables in the two regression equations.

$\Sigma^{11}_{i=1} D_i$ = Eleven Seasonal Dummies representing any seasonal effects on employment in the two industries. (Why only 11? This follows the rule that, with an intercept in the model, the number of dummies equals one minus the number of categories.)

Y_t, Y_{t-1} = Income at periods t and t-1.

R_t, R_{t-1} = Short-Term interest rates at periods t and t-1.

Equations (14.) reproduces and modifies Equations (13.) used in the BVAR example. But, now the General Merchandise stores has current-period and one-period lagged, exogenous variables for Income, but there are no exogenous variables for Short-Term Interest Rates. Also, the Merchandise Wholesalers' equation has current-period and one-period lagged, exogenous variables for Short-Term Interest Rates but there are no exogenous variables for Income. Since, current and lagged levels of income may play a greater role in determining the level of employment in General Merchandise stores, and since short-term credit plays a large role in Merchandise Wholesalers' fortunes, its level of employment may be more dependent on the level of Short-Term Interest Rates. Such a specification would not be amenable to estimation as a Classical or Bayesian VAR. Since the variables are not all the same, the problem of *contemporaneous correlation arises* (see Appendix C). Further, both the endogenous variables might appear in one model, but only one in the other, in the two-equation system of Equations (14.). Again, the SUR problem would arise. The more flexible Near-VAR specification in Equations (14.) allows the forecaster to capture those factors common to both industries in the two-equation system, on the one hand, but it also allows the introduction of variables that represent factors effecting the level of employment that are unique to one industry's employment behavior in the system.

Clearly, in many instances, the VAR specification will produce the best results in obtaining optimal forecasts. However, there are enough instances where circumstances are such that a SUR or Near-VAR specification will clearly offer superior forecasts. Further, tests such as the LM Test (Breusch-Pagan) and the Likelihood Ratio Test can be applied to determine whether a SUR specification should be explored, or whether OLS in the form of a VAR, or even separate regressions would produce the optimal forecast results.

Due to the number of models used to forecast the Control-Totals, the discussion will not include an assessment of the empirical estimation of the control-total models, as did the discussion of the Super-Control Forecast model. With that, the next section turns to the Detailed-Level Employment Forecasts.

D. The Detailed-Level Forecasts of Connecticut Employment

Given the level of detail, the process for producing the Detailed-Level Employment forecasts is necessarily the most mechanical. There are some 100 three- and four-digit level NAICS industries in Connecticut, which limits the amount of time and effort that can be devoted to developing and estimating a given forecasting model. There are two primary tools used for forecasting Connecticut Employment at the detailed level, the Short-Term Industry Projections (STIP) system developed by the consortium of states for ALMIS (America's Labor Market Information System) to provide a tool for states' LMI (Labor Market Information) units to develop timely, relatively uniform employment forecasts (see Section I, Introduction, to this paper). SAS/ETS, the Econometric and Time-Series package is also used, particularly, the Forecasting Menu System, and PROC FORECAST, the multiple-series forecasting utility.

The forecaster using the STIP system has five models to choose from: Exponential Smoothing with Linear Trend and Random Walk options, OLS (single-equation, Linear Regression), ARMA (Autoregressive Moving Average), VAR, and BVAR[21]. Mix gives a weighted average forecast based on the five models available in the STIP system. Most of the models used to forecast industry-employment at the Detailed-Level are multiple, time-series systems. The VAR and BVAR specifications are drawn on quite frequently. In addition to the specific employment-series being forecasted, other, related-industries included in a VAR or BVAR, as endogenous variables, are those suggested by the inter-industry relationships found in the 1997 Benchmarked, U.S. Input-Output Table. However, in some instances, there are no related industries. In such cases, univariate models are used to forecasts the employment series. There are two types of univariate models used in the Connecticut Forecasts: *Deterministic* and *Stochastic.*

Exponential Smoothing is a *weighted moving average* making it an extension of the moving average method[22]. In the Exponential Smoothing class of models, past values are discounted such that those observations further in the past are assigned weights that give them less influence over forecasts than values in the more recent past. And, these weights decrease *exponentially* as the observations go further back in time. Additionally, in

Exponential Smoothing, parameters are determined explicitly, and the parameters chosen determine the weights assigned to observations.

The simplest exponential smoothing model is the *Single Exponential Smoothing* model presented below:

$$F_{t+1} = F_t + \alpha(Y_t - F_t)$$

Where: F_{t+1} = Forecast at period t + 1

F_t = Forecast at period t

Y_t = Actual value at period t

α = A constant between 0 and 1

To forecast a value of the time-series Y_t, where F_t is the forecast for period t, the first step after observation Y_t becomes available, is to find the forecast error, $Y_t - F_t$. Next, the Single Exponential Smoothing model takes the forecast for the previous period and adjusts it using the forecast error. The result, shown in the equation above, is the forecast for time period F_{t+1}. Thus, the new forecast is the old forecast plus an adjustment for the error in the previous forecast. When α has a value close to 1, the new forecast will have been substantially adjusted for the error in the previous forecast. Conversely, an α-value close to 0 implies that the new forecast required very little adjustment. A *large* or *small* value of α implies (in the *opposite* direction) a *small* or *large* number of observations when computing the moving average.

Exponential Smoothing involves a basic principal of *negative feedback* since it works much like the control process employed by thermostats and automatic pilots. That is, the past forecast error is used to correct the next forecast in a direction opposite to that of the error. If properly applied, this procedure can be used to develop a self-adjusting process that corrects for forecasting error automatically.

The general form, used in expressing exponential smoothing methods, can be stated by re-writing the equation above as:

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t$$

Where: F_{t+1} = Forecast at period t + 1

F_t = Forecast at period t

Y_t = Actual value at period t

α = The weight for the most recent observation (a constant between 0 and 1)

$1 - \alpha$ = The weight for the most recent forecast.

This form requires only the most recent observation to make a forecast for the next period.

In actual practice, the movements in the data the forecaster is confronted with will be too complex to be adequately captured by the above Single Exponential Smoothing model. Industry employment time-series are likely to contain trend and seasonal fluctuation components. To adequately take into account these components of time-series, some complexity must be added to the model. The Holt-Winters method expands on the Single Exponential Smoothing model to include the ability to model *Trend* and *Seasonal* components, as well as the level of an employment time-series. The Holt-Winters method is based on three smoothing equations: one for the level, one for the trend, and one for seasonality. This is the type of exponential smoothing model estimated by the STIP system. Equation (4.4) from Chapter 4 (p.54) of *A Primer for ALMIS Forecasting* is re-stated below:

$$E_{t+1} = S_t + I_{t+I-12}$$

$$S_t = S_{t-1} + T_{t-1} + \alpha \varepsilon_t$$

$$T_t = T_{t-1} + \alpha \gamma \varepsilon_t$$

$$I_t = I_{t-12} + (1 - \alpha)\delta \varepsilon_t$$

Where: E_{t-1} = Previous month’s employment level

S_t = Current month’s smoothed employment level

T_t = Trend component

I_t = Seasonal component

$\alpha, \gamma, \varepsilon_t, \delta$ = Model parameters

The ***Autoregressive Moving Average*** (ARMA), model is the other univariate model used in producing the detailed forecasts. Further, the ARMA is a stochastic or statistical, model[23]. Introducing the concept of stochastic, processes brings up the issue of ***Stationarity***. The term 'ARMA' implies that the time-series modeled and forecasted is Stationary**.** The issue is a critical one, and therefore, it is discussed in detail in Appendix D. In fact, the reader is urged to read Appendix D before proceeding to the next section.

Stationary Stochastic Processes[24]

Assume that a particular set of observations, ordered through time, are *realizations* of random variables. And, moreover, assume that these random variables are only part of an infinite sequence of variables. If these assumptions hold, then this sequence is called a ***Stochastic Process***. More precisely, it is a *Discrete* Stochastic Process, because the time index t assumes only integer values. If in addition to meeting these requirements, the stochastic process also adheres to the conditions set down (see Appendix D and the discussion on stationarity), then the process is a ***Stationary Stochastic Process***. The following discussion turns to three stationary stochastic processes encountered in building, estimating, and forecasting with univariate, statistic models.

Autoregressive Processes

The AR process was first introduced in Section II in the discussion of the Connecticut Super-Control Forecast Model. It is now re-visited in more detail. In an ***Autoregressive (AR) Process,*** a given observation, y_t, of a stochastic process, is dependent on its past values. This dependence is important for forecasting. Information on past values of the time-series can be used to predict future values of the variable. A simple example of a process for which such a dependence exists is the AR Process:

$$y_t = \rho y_{t-1} + e_t$$

This is an *AR Process of Order 1* denoted: AR(1). That is, the current value, y_t, is dependent on its immediate past value, or *first lag*, only. Further, it is important that

$|\rho| < 1$, for this process to be stationary. Otherwise, the process would not converge. From an intuitive perspective, if, when going back to earlier and earlier observations in the process (e.g., y_{t-1}, $y_{t-2,\ldots,}$ y_{t-n}) each past value has progressively less influence on the value of the current observation, then each past value should be discounted at progressively steeper rates. This is guaranteed if the absolute value of ρ is less than one. Finally, it is assumed that e_t is a ***White Noise*** process. This means that the e_t are assumed to be *Normally, Identically, and Independently Distributed* with a mean of zero and a constant variance. This implies that a White Noise series is stationary.

Usually, the generating process of a time-series will be unknown, and, if the process is stationary, it can have a process that is more complicated than the simple AR(1) given above. In general, an Autoregreesive Process is of order p is indicated by the notation: AR(p). This indicates that there can be more than one lag, thus, 'p' can be '1', '2', etc. It indicates the number of past values of the process, y_t, needed to determine the value of the current observation.

Moving Average Processes

If a process cannot be represented by a low-order AR process, then it can be re-stated as a Moving Average (MA) process. In fact, it can be shown that any stationary AR process can be written as an MA process. A ***Moving Average*** process is a process where the current value, y_t, is a weighted sum of the past values of the White-Noise series, e_t (also known as ***Innovations*** or, ***Random Shocks***). The following expression is an example of an MA of order 1:

$$y_t = e_t + \theta e_{t-1}$$

As for the AR process, the MA process can be more complicated than the MA(1) above. A higher order MA of order q is denoted by MA(q). An MA(q) process that can be written as an infinite, stationary AR process is said to be ***Invertible.***

ARMA Models

The task faced by the forecaster is to identify a *parsimonious* representation of the data-generating process. Under certain conditions, the AR process will provide the best representation, while other conditions may suggest an MA representation. However, there are many instances when the best, parsimonious representation is one that includes both AR and MA terms. The simplest process would be that which combined an AR(1) with an MA(1), and is presented below:

$$y_t = \rho y_{t-1} + e_t + \theta e_{t-1}$$

This process is called an ***Autoregressive Moving-Average Process*** of order (1,1). It is denoted by: ARMA(1,1). As for the AR and MA processes, the ARMA process can be generalized to a longer-lag representation by: ARMA(p,q). If the data were differenced before being modeled, and have to be integrated (see discussion above) then the process would be an ***Autoregressive Integrated Moving Average Process***. If the data were differenced once to make it stationary then it is ***Integrated of order 1***, denoted by I(1). In ARIMA notation, the above order-one process would be denoted by ARIMA(1,1,1). Again, this can be generalized by the notation ARIMA(p,d,q). A series that is stationary after differencing it d times is sometimes said to be *Homogeneous Non-Stationary of Degree d*, or Integrated of order d denoted, I(d).

In practice, it can be difficult to adequately identify the orders of p and q. This is where the acf and pacf become important tools in identifying ARMA models (see Appendix D). In fact, the acf and pacf are critical to the Box-Jenkins Approach to time-series model building and forecasting.[25] Rather than using the Box-Jenkins Approach, the STIP software runs a tournament of nine different ARMA models for the forecaster to choose from. In addition, the forecaster can specify a user-defined model. In which case, the appropriate statistics can be used to identify a model. However, any exploration or identification of the proper model order requiring the use of the acf or pacf must be done outside the STIP system. The SAS Forecasting System does allow the forecaster to apply the Box-Jenkins approach, or to default to SAS picking the model. The PROC

FORECAST procedure in SAS/ETS will automatically pick AR models to forecast large numbers of series simultaneously.

IV. CONCLUDING REMARKS

To summarize, the Connecticut Short-Term Employment Forecasts are the product of several different steps, procedures, and modeling frameworks. Three different levels of forecasts are produced and reconciled: The Super-Control Forecast, the Control Forecasts, and the Detailed-Level Forecasts. Each level of forecast produces progressively more detailed forecasts. The Super-Control Forecast is the top-line level of Connecticut, Non-Agricultural Employment, and it gives the least level of detail. The Control Forecasts provide a greater level of detail. The Control Forecasts are produced at the NAICS sector level, or two-digit level of detail. Forecasts are produced for the 20 sectors, including some of their major sub-aggregates, such as Durable Goods and Non-Durable Goods under the Manufacturing Sector. Finally, the detailed-Level forecasts, as would be expected, provide the most detail. The Detailed-Level Forecasts are produced at the NAICS three- and four-digit level of detail. Forecasts are produced for 100 three- and four-digit level industries in Connecticut. Once forecasts have been completed at all three levels of detail, the Base-Line Forecast is then produced. The Base-Line Forecast is the product of two steps. First, forecasts are Pooled or Combined then, Reconciliation of the three forecasts is done using both, *top-down*, and *bottom-up* approaches. Once the Base-Line Forecast is in place, any macroeconomic-based Intercept Corrections are then applied.

The Final Forecast is the product of the process outlined above. Specifically, the above sequence of methodologies can be summarized as a four-step process. ***First***, the three levels of forecasts are produced: The Super-Control Forecast, the Control Forecasts, and the Detailed-Level Forecasts. The ***Second*** step is to produce three top-line forecasts. The Super-Control, the sum of the Controls, and the sum of the Detailed Forecasts are used to produce three top-line forecasts. The simple average of the three forecasts is also considered. Then, the Controls are compared to the Detailed Forecasts' sums by NAICS sector. The two sector-level forecasts are also averaged to produce a third Control-Level Forecast. The ***Third*** step is to perform both, top-down and bottom-up reconciliations of the forecasts. Upon completion of this step, the ***Base-Line Forecast*** is set. The ***Fourth*** and final step in producing the ***Final Forecast***, involves macroeconomic-based Intercept Corrections.

It is hoped that this paper has succeeded in providing an informative presentation of the quantitative and qualitative methodologies used in producing Connecticut's Short-Term Employment Forecasts. The forecast horizon of two years, or eight quarters, for the short-term forecasts requires the forecaster to focus on analyzing the economy in the short- to intermediate-run. This means that forecasting methods must identify the expected seasonal, cyclical, and even some trend effects in industry employment. It is the process of capturing these critical phenomena, in order to construct models that produce optimal forecasts, given time and resource constraints, that has guided the development of the methodologies applied to the short-term employment forecasts.

Finally, a list of '*getting-started*', introductory, to intermediate, forecasting references is provided in Appendix E. These works should provide the novice with a solid foundation for practicing the art and science of forecasting. For more information, or any questions concerning the methodology used to produce the employment-forecasts, please contact:

Daniel W. Kennedy, Ph.D., Senior Economist
Connecticut Department of Labor – Office of Research
(860) 263-6268
daniel.kennedy@ct.gov

APPENDIX A: Data Requirements

The employment time-series used for the Connecticut Employment Forecasts are from the data reported under the Unemployment Insurance (UI) tax program, formally known as the Covered Employment and Wages Series (ES-202), which is now known as the ***Quarterly Census of Employment and Wages*** (QCEW). It is the QCEW employment series that are used for the industry-employment side of the labor-market forecasts produced by America's Labor Market Information System (ALMIS) program's statistical software. The QCEW data is used for both, the long-term and short-term forecasts. The Long-Term Industry Projections (LTIP) system uses annual employment series to produce the long-term industry, employment forecasts (10 years ahead), and the Short-term Industry Projections (STIP) system uses monthly employment series to produce quarterly forecasts two years, or eight quarters ahead. All employment data, at both, the national and state levels, have been converted from the 1987 Standard Industrial Classification (SIC) scheme to the ***North American Industry Classification System*** (NAICS). Connecticut's employment series were constructed from the pushback files on CD ROM's produced by the U.S. Bureau of Labor Statistics, which provided monthly employment series, at the six-digit NAICS industry level of detail, covering the period: 1990:M01-2001:M12. Appended to that was the Connecticut QCEW employment series, produced by the Office of Research of the Connecticut Labor Department, from 2002:M01 to 2003:M12. This resulted in an uninterrupted monthly, time-series of Connecticut employment data for the 14 years covering the period: 1990:M01-2003:M12. This provided the frame, which served as the basis, for the sample of employment series drawn for the study.

The exogenous or indicator variables representing economic and industry factors at the National, State, and Region levels were those series provided by, and forecasted by, the STIP software for those parts of the study done in the STIP software. Those portions of the study done outside of the STIP forecasting system used an algorithm in EViews written by Roy Pearson, Professor of Management at The College of William and Mary. The series for the exogenous variables used in the super-control and control-forecast models were obtained from the U.S. Bureau of Economic Analysis's website, the U.S. Census Bureau's website, the Conference Board website, the Connecticut Labor Department, and the Boston Federal Reserve Bank's New England Economic Indicators website. Forecasts of the exogenous variables are obtained from the New England Economic Partnership Forecast, Ray C. Fair's website, GlobalInsight, and the Blue Chip Economic Indicators.

APPENDIX B: Bayes's Theorem

Before introducing Bayes's Theorem, it will be helpful to first introduce the concept of *Conditional Probability*[1]. If A and B are events in a *Sample Space* (which contains all possible events) denoted as 'S', then the Conditional Probability of event A, given that event B has occurred, is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Provided P(B) > 0.

Where: $P(A \cap B)$ = The *Intersection* of A and B, which is the set of all points in both A and B. It is the probability of the occurrence of this set of points.

$P(A \mid B)$ = Probability of A, given B.

Thus, the probability of A is conditional upon B occurring. ***Bayes's Theorem*** is a more detailed re-statement of the above conditional-probability expression. It is stated as follows:

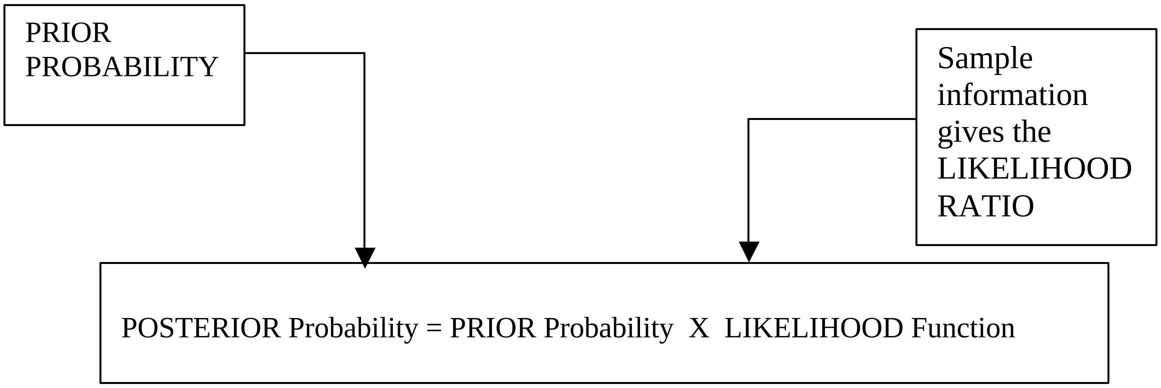
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

The probability **P(A)** is the ***Prior Probability*** discussed above, and the conditional probability, **P(A | B)**, is the ***Posterior Probability*** that represents the *revised* assignment of probabilities *after* obtaining the updating information (e.g., from sample evidence). Bayes's Theorem can be re-stated verbally as:

> …the posterior probability of an event A, is proportional to the probability of the sample evidence after A, times the prior probability of A.[2]

The logic of Bayes's Theorem is demonstrated in the figure below.

FIGURE 1: The Logic of Bayes's Theorem


PRIOR PROBABILITY
Sample information gives the LIKELIHOOD RATIO
POSTERIOR Probability = PRIOR Probability X LIKELIHOOD Function

SOURCE: Wonnacott and Wonnacott (1990), p. 587.

To calculate the Posterior Probability distribution of β, the regression slope for a model of interest, given the observed sample data, X, would be:

$$p(\beta \mid X) = \frac{P(\beta, X)}{p(X)}$$

Re-expressing the numerator, gives the following expression[3]:

$$p(\beta \mid X) = [1/p(X)]p(\beta)p(X \mid \beta)$$

Since the sample data, X, has been observed, it is given and fixed. Therefore, 1/p(X) is a fixed constant. Next, p(β) is the *Prior Distribution* incorporating all prior knowledge about β. Finally, the last term, $p(X \mid \beta)$, with fixed X, while β varies, is called the ***Likelihood Function***. By removing the constant, 1/p(X), from the expression, it can be simplified as:

$$p(\beta \mid X) \propto p(\beta)p(X \mid \beta)$$

Where: $\propto$ = ‘is proportional to’. It indicates that, save a constant, the expression is an equality. It allows unnecessary clutter to be removed from an expression.

The above relationship can be written out in words as:

Posterior Distribution $\propto$ Prior Distribution x Likelihood Function.

With the above introduction to the Bayesian framework, the discussion now turns to the specification of a Bayesian Vector Autoregression (BVAR)[4], and the role of the Minnesota Prior.

The Bayesian approach to building and estimating VAR’s was motivated by critics of a common practice among forecasters. Namely, it is the ‘*art*’ part of the ‘Art and Science’ of forecasting that they were uncomfortable with. Critics were bothered by the common practice, in which the forecaster incorporates his or her personal beliefs, or judgment, into the forecast. Particularly, since there is no documentation of this subjective process, it cannot be reproduced by other forecasters. Their answer to this state of affairs was to advocate an approach that was based on an *objective* procedure for combining beliefs and data in building economic forecasting models. That objective procedure is the *Bayesian* approach to building econometric models for forecasting. Within the VAR context, this approach yields the ***Bayesian Vector Autoregression*** or **BVAR.** Researchers at the University of Minnesota and the Federal Reserve Bank of Minneapolis developed BVAR procedures to give modelers and forecasters more flexibility in expressing their beliefs about the economy, and its direction, as well as, an objective way to combine those beliefs with the historical record.

All statistical forecasting models combine information from the historical data with information supplied by the modeler-forecaster. Modelers introduce their own information into the process because they believe it will improve the model's forecasting ability. This information that they supply is known as their ***Prior Beliefs***, which is the *Prior* in the Expression, above.

The BVAR approach discussed here provides an objective and reproducible procedure for combining a forecaster's beliefs and data. For the reasons mentioned above, this particular system of Bayesian priors is known as the *Minnesota System of Prior Beliefs*, or the ***Minnesota Prior***.

After completing the usual process of choosing the variables to be included in the VAR, the prior beliefs about the values of each of the coefficients in the equations in the VAR system can be expressed in the form of probabilities about which set of values will give the best forecasts. In the Minnesota Prior, these probabilities can be described by assigning a best guess and a measure of confidence to each coefficient in the model. Both of these guesses would be quantitative (i.e., a number). The best guess is set according to the *Random Walk Hypothesis*. This hypothesis states that variables behave in such a way, that changes in their values are unpredictable. For such a variable, the best forecast of its one-step ahead value is equal to its current value. To implement the random walk hypothesis, the best guesses of the Minnesota Prior are that all coefficients in the equation, save the most recent value, are zero. The coefficient for the most recent value is guessed to be 1. In addition, the forecaster must supply a quantitative measure of confidence in each best guess. This is expressed as the *Prior Variance of the Coefficient*. The smaller the prior variance, the more confidence the forecaster has that his or her best guess will be close to the forecast. With one exception, the system then proceeds in two stages. First, the forecaster selects a few restrictions that group the prior variances and mainly determine the relative sizes of the prior variances within each group. Second, the forecaster selects a range of possible values for a scale factor that completes the determination of the prior variances. The one exception to the two-stage process is the procedure for determining the prior variance of the constant terms (intercepts) in each equation. These variances are simply set to vary large numbers, which amounts to saying that, at least over a very large range, the forecaster regards all possible values of the constant term as almost equally likely. In other words, the forecaster is willing to allow the constant term to be determined by the data alone.

APPENDIX B ENDNOTES

[1] Judge, Et al, pp. 15-18.

[2] Ibid. p.18.

[3] To re-express the numerator, which produces the new expression (Wonnacott and Wonnacott, Equation 3-18)

[4] This section draws heavily upon Todd (1984).

APPENDIX C: Seemingly Unrelated Regressions (SUR), or Near-VAR[1]

To understand the SUR approach, it is best to start with some of the assumptions about the *Classical Normal Linear Regression*. Particularly important, are the assumptions about the disturbance term. Specifically:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

Which is supposed to satisfy the following requirements:

$$E(\varepsilon_i) = \sigma^2 \text{ for all i,}$$

$$E(\varepsilon_i \, \varepsilon_j) = 0 \text{ for all } i \neq j.$$

In compact matrix form, the above assumptions can be expressed as:

$$E(\varepsilon \, \varepsilon^{\mathbf{T}}) = \sigma^2 \mathbf{I_n}$$

In words, the above requirements state that the error variance, σ^2, is constant for all observations. And, $\sigma^2 \mathbf{I_n}$ is known as the ***Variance-Covariance Matrix***. The superscript symbol, 'T', indicates the transpose. This condition is called ***Homoskedasticity***. When this requirement is violated, and there are unequal variances over observations, it is called ***Heteroskadasticity***. Further, it is also assumed that there is no correlation between the errors across the observations. That is, their *Covariances* are zero. $\mathbf{I_n}$ is an identity matrix of order (n x n). For instance, in a regression model with three observations (i.e., n=3), the Variance-Covariance Matrix would be written out as:

$$E(\varepsilon \, \varepsilon^{\mathbf{T}}) = \sigma^2 \mathbf{I_n} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

The constant variance is demonstrated by writing the same value, σ^2, for each element on the principal diagonal, and the zero covariance is indicated with zeros for all off-diagonal elements.

If these two assumptions fail, but all the other assumptions of the Classical Linear Regression model hold, then such a model is called a ***Generalized Linear Regression Model.*** Stated in matrix notation, the model now is:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

Where: **Y** = (n x 1) vector of observations on the *Dependent Variable*.
X = (n x K) matrix of n observations on K *Independent Variables*.
β = (K x 1) vector of *Regression Coefficients*.
ε = (n x 1) vector of *Errors* on the n observations.

The Variance-Covariance is now denoted as Ω, and would be written as follows:

$$E(\varepsilon\,\varepsilon^{\mathbf{T}}) = \Omega$$

Using the three-observation example above, the new variance-covariance matrix is written as:

$$\Omega \quad = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

Notice that now the diagonal elements are no longer all σ^2, they now have subscripts indicating that they are now no longer necessarily the same. This reflects the possibility of unequal variances at different observations, referred to above as *Heteroskadasticity*. Also, the off-diagonal elements are no longer zeros, and they too now have subscripts. This indicates that they can now have values other than zero.

That is, $E(\varepsilon_i\ \varepsilon_j) = 0$ for all $i \neq j$, is no longer assumed to necessarily be the case. This can occur especially when observations are taken over time. This correlation between the error terms over time is known as ***Autocorrelation*** (see the ARMA discussion above, and Appendix D). This model is called '*generalized*' because it includes other models as special cases. One such special case is the Classical Normal Linear Regression.

The data encountered by the forecaster in specifying and estimating models to project industry employment is, of course, time-series data, where each observation is a point in time (e.g., a month, a quarter, a year, etc.). Also, it should be noted at this point, that some differences do exist between the classical assumptions for cross-sectional and time-series regression. As mentioned above, Autocorrelation is likely to be encountered in time-series data. Estimation methods that correct for the problems discussed in this section (i.e., Heteroskadasticity and Autocorrelation) and result in the Generalized Linear Regression Model introduced above are called Linear Statistical Models with a ***General Error Variance-Covariance Matrix.*** The estimation procedure, which corrects for the above violations of the Classical assumptions, is called ***Generalized*** or ***Weighted Least Squares***. The details of the estimation procedure can be found in the references in Endnote 1 of this appendix. The Weighted Least Squares (WLS) procedure picks weights for the estimation process, before estimating the parameters. Thus, it is a two-stage process. The result is that the Generalized Least-Squares (GLS) Estimator is the minimum variance linear unbiased estimator under any general error covariance specification with the presence of Heteroskadasticity, Autocorrelation, or both (Actually, in practice, since the *true* Variance-Covariance matrix will not be known, the *Estimated or Feasible GLS* estimator is used).

Sets of Error-Related Economic Relations

If in the VAR and BVAR example, each equation were estimated separately, without regard to the related behavior of Merchandise Retail and Merchant Wholesale employment, the result would have been imprecisely estimated parameters. That is, they would have had large standard errors. Why should this be? Because, estimating the two equations separately, would leave out information about the interrelationship of employment between these two industries, which would not be explicitly incorporated into the model as independent variables. Further, economic factors affecting both industries simultaneously are also not accounted for in each of the stand-alone equations. The result is a violation of Classical assumptions. Specifically, in regard to the error term, in Classical linear regression, the Independent or Explanatory variables explain *all* the systematic variation in the Dependent variable. All random, or non-systematic influences are captured in the error term. When all systematic influences cannot be captured in a single-equation specification, then any unaccounted-for systematic variation in the Dependent variable (in this case, Industry Employment) will show up in the error term, which violates the Classical assumptions. The consequence is an error series that is not *White Noise* (see the ARMA discussion above). That is, the error term is no longer random, and the explanatory variables do not account for all the systematic variation in the Dependent variable.

To see how this problem arises, and how the SUR estimation addresses it, a slightly more complicated version of the Variance-Covariance matrix is presented below:

$$E(\boldsymbol{\varepsilon}_1\, \boldsymbol{\varepsilon}_2^{\mathbf{T}}) = \Sigma = \begin{bmatrix} \sigma_1^2\, \mathbf{I_n} & 0 \\ 0 & \sigma_2^2\, \mathbf{I_n} \end{bmatrix}$$

The above expression combines the Variance-Covariance matrices of the Merchandise Retail Employment equation ($\sigma_1^2\, \mathbf{I_n}$) and the Merchant Wholesalers Employment equation ($\sigma_2^2\, \mathbf{I_n}$) into one composite, or joint, Variance-Covariance matrix expression. The above expression assumes that there is no relation through the error terms between these two equations. This is indicated by the zero values for the off-diagonal elements. This indicates that there is no correlation in the error terms across equations for the same time period. That is, there is no ***Contemporaneous Correlation***. In this case, there is no set of error-related relations and the two equations would be estimated separately. And, the separate estimations would be efficient and adhere to the Classical assumptions. All information required to explain the systematic variation in the two employment series would be captured by the independent variables in each, separately estimated equation, and their error series, at least in regard to this assumption, would be white noise.

However, there is, in fact, based on the previous discussion, an error-related relation between Merchandise Retail and Merchant Wholesale employment. Now, the joint Variance-Covariance matrix is written as:

$$E(\boldsymbol{\varepsilon_1}\ \boldsymbol{\varepsilon_2}^{\mathbf{T}}) = \Sigma = \begin{bmatrix} \sigma_{11}{}^{2}\,\mathbf{I_n} & \sigma_{12}{}^{2}\,\mathbf{I_n} \\ \sigma_{21}{}^{2}\,\mathbf{I_n} & \sigma_{22}{}^{2}\,\mathbf{I_n} \end{bmatrix}$$

Notice that in the new expression above, the off-diagonal values are no longer zero. This reflects the presence of *Contemporaneous Correlation*. That is, there is correlation across equation errors in the same time period. To account for this set of error-related economic and labor-market relations, the *Generalized Least Squares* estimation procedure (used to account for the problems of Heteroskadasticity and Autocorrelation in estimating single-equation models, discussed above), can be generalized to produce efficient parameter estimates for systems of equations (i.e., two or more equations) in the presence of *Contemporaneous Correlation*. That is, the method used to ‘fix’ the problem of non-zero, off-diagonal elements in the single-equation Variance-Covariance matrix, is now extended, to ‘fix’ the presence of non-zero, off-diagonal values in the multiple-equation, joint Variance-Covariance matrix. This is the type of statistical model (or set, or system, of equations) that Zellener called ***Seemingly Unrelated Regressions*** (SUR), *or Error-Related Regression Equations.* Further, it is an extension, or another form of, the *General Error Variance-Covariance* statistical model.

The higher the Contemporaneous Correlation of the cross-equation errors, the more efficient the SUR estimation is over Ordinary Least Squares (OLS). If there is a high degree of collinearity among the independent variables in each equation in the system, then the efficiency-gain is offset somewhat. An interesting result obtains if each equation has the same variables and, there are the same number of variables in each equation. In this case, the Estimated GLS estimator becomes the OLS estimator, and there is no gain from estimating the equations as a SUR system. Further, in such a case, there is no gain in efficiency by estimating the equations' parameters simultaneously. In such a case, estimation of each equation separately, using OLS, yields efficient parameter estimates. This, of course, is the VAR system discussed in the previous section of this report. In fact, the VAR can be thought of as a special case of the SUR.

APPENDIX C ENDNOTES

[1] This section draws on Kmenta, Jan, *Elements of Econometrics*, MacMillen Publishing (1971): New York, Chapter 12; Griffiths, Williams E., R. Carter Hill, and George G. Judge, *Learning and Practicing Econometrics*, John Wiley & Sons (1993): New York, Chapter 17; Judge, El al, Chapter 12; Zellner, pp. 240-246

APPENDIX D: Stationarity[1]

Stationary Time-Series

A ***Stationary*** time-series has a mean, variance, and autocorrelation function (acf)[2] that are essentially constant through time. More precisely, consider the ***realization***[3] $y_{1,}\ldots, y_n$. Suppose these observations are drawn from a joint probability distribution:

$$P(y_{1,}\ldots, y_n)$$

Where: P = A Joint Probability Density Function that assigns a probability to each possible combination of values for the random variables $y_{1,}\ldots, y_{n.}$

The goal in forecasting is to make statements about the likely values of future y's. If the joint density function $P(y_{1,}\ldots, y_{n+1})$, including the relevant marginal probabilities, is known, then the following conditional distribution could be formed:

$$P(y_{n+1} \mid y_{1,}\ldots, y_n)$$

Then, from the knowledge of past values, $(y_{1,}\ldots, y_n)$, the above expression could be used to make a probability statement about the future value, y_{n+1}. For a process to be ***Stationary***, the joint distribution function describing that process must be *invariant with respect to time.* That is, if each random variable, $(y_{1,}\ldots, y_n)$, is displaced by m time periods, then the following stationarity condition holds:

$$P(y_{t + m,}\ldots, y_{t + k + m}) = P(y_{t,}\ldots, y_{t+ k})$$

The above condition is sometimes referred to as ***Strong*** or ***Strict Stationarity*** .It shows that the entire probability structure of the joint function is *constant* through time. ***Weak Stationarity*** requires only that certain characteristics of the joint function be time invariant. But, there is a simplification that can be made if the joint function is *Normally* distributed. If the joint function is a Normal distribution, then it is *strongly* stationary if its mean (first moment) and variances and covariances (second moment) are constant over time. In the discussion below, it will be assumed that the random shocks, a_t, are Normally distributed. This is equivalent to the assumption that the joint distribution for the y's is a joint Normal distribution.

If the joint distribution for the y's is Normal, then the following holds:

1. There is a constant Mean, $\mu = E(y_t) = E(y_{t+m})$.
2. There is a constant Variance, $s^2{}_y = \gamma_0 = E(y_t - \mu)^2 = E(y_{t+m} - \mu)^2$, for all y's.
3. And, constant Covariances, $\gamma_k = E[(y_t - \mu)(y_{t+k} - \mu)]$, for any two y's separated by k time periods.

Why go through this exercise? Because, it is important to establish stationarity in order to apply classical statistics to test hypotheses or, within the current context, to make inferences about the forecast and the forecast error, based on information about the given realization. If the data are *non-stationary*, then the first and second moments are not constant over time, and classical statistical statements cannot be made about 'moving targets'. Such statements are based on inferring sample (i.e., realization) statistics to fixed population (i.e., underlying stochastic process) parameters.

For those who have worked with economic, demographic, or labor-market data, it is apparent that most time-series encountered in these situations are not stationary. Particularly, long-run phenomena such as long-run growth in GDP, population, and the labor force produce time-series that are trended. In these cases the mean is not constant over time—it is either increasing or decreasing over time. Fortunately, many non-stationary series can be transformed into stationary series through a simple transformation: differencing. ***Differencing*** is the process of calculating successive changes, from one time-period to the next, in the values of a time-series. Thus, the *First Difference* of y_t is:

$$\Delta y = y_t - y_{t-1} \quad t = 2, 3, \ldots, n$$

Successive differencing can be carried out if the data are not stationary after the first difference. Accept in rare instances, differencing more than twice may ‘overdifference’ the data. That is, new, artificial patterns can be introduced into the data. Seasonal differencing can also be done if the seasonal pattern is not stationary.

If the data are differenced, and a univariate, statistical model is estimated for forecasting, then the data must be re-transformed back into the original levels, after making forecasts. This process of re-transforming, or ‘backing into’, or reversing the process to get back to the original levels of the data is called ***Integration***. This is known as an *Autoregressive Integrated Moving Average* (ARIMA) model. In Section 4.3, of Chapter 4, *A Primer for ALMIS Forecasting*, the ARMA model is detailed. Notice that this is an ***Autoregressive Moving Average*** model, not an ARIMA. An ARMA assumes that the data is stationary. Since the STIP software automatically differences all data that is input into its statistical models, it is assumed that stationary data is being used to estimate the models. Thus, the ‘ARMA’ convention is used, rather than ‘ARIMA’.

Forecasting Employment and Non-Stationary Economic Time-Series[4]

Clements and Hendry (1999)[5] argue that differencing and intercept corrections have no rationale when models are correctly specified. Further, Sims has criticized the differencing of series in a VAR as throwing away valuable information. Thus, differencing all series before estimating forecasting models would result in the deterministic influences on the behavior of the series, such as trend and structural breaks being discarded in specifying a model for forecasting. This is especially a concern since deterministic shifts in the model, relative to the *Data Generating Process* (DGP), are a dominant source of forecast failure[24]. With the flexibility of the more general SUR specification, the deterministic peculiarities specific to a given series in the SUR system can be specified without applying them to every other series in the system. On the other hand, phenomena common to all series in the system can still be captured in the SUR specification.

Finally, the more generalized SUR specification would allow the forecaster to construct a more flexible multi-equation forecasting model. By allowing for different lags of the endogenous and exogenous variables in each equation in the system to account for and tailor the different structural (i.e., deterministic) features specific to a given time-series, while still capturing their shared common economic and labor-market characteristics.

APPENDIX D ENDNOTES

[1] This section draws heavily on Pankratz, Alan, *Forecasting with Univariate Box-Jenkins Models*, John Wiley & Sons (1983): New York, p. 11 and Chapters 2 and 3.

[2] The ***Autocorrelation Function*** (acf) is the time-series counterpart to the *Correlation Coefficient* (r) in the cross-sectional context.

[3] The ***Realization*** from an underlying ***Stochastic Process*** in the time-series context is the counterpart to a *sample* drawn from a *population* for cross-sectional data.

[4] The section draws heavily on Clements, Michael P., and David F. Hendry, *Forecasting Non-Stationary Economic Time-Series*, The MIT Press (1999): Cambridge, MA.

[5] Ibid. p. 6.

APPENDIX E: List of '*Getting-Started*' Forecasting References

LIST OF FORECASTING AND ECONOMETRICS BOOKS		
AUTHOR	**TITLE**	**PUBLISHER (YEAR)**
Bails, Dale G. G. and Larry C. Peppers	Business Fluctuations*: Forecasting Techniques and Applications*, 2nd Ed.	Prentice Hall (1997)
Diebold, Francis X.	Elements of Forecasting, 2nd Edition	South-Western (2001)
Griffiths, William E., R. Carter Hill, and George G. Judge	Learning and Practicing Econometrics	John Wiley & Sons (1993)
Hall, Stephen, Editor	Applied Economic Forecasting Techniques	Harvester Wheatsheaf (1994)
Hendry, David F. and Neil R. Ericsson, Editors	Understanding Economic Forecasts	MIT Press (Paperback-2003)
Holden K, D.A. Peel, and J.L. Thompson	Economic Forecasting: An Introduction	Cambridge U. Press (1994)
Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee	**Introduction to the Theory and Practice of Econometrics**	**John Wiley & Sons (1988)**
Lutkepohl, Helmut	***Introduction to Multiple Time Series Analysis, 2nd Edition***	***Springer-Verlag (1993)***
Makridakis, Spiro, Steven C. Wheelwright, and Rob J. Hyndman	Forecasting: Methods and Applications, 3rd Edition	John Wiley & Sons (1998)
Wooldridge, Jeffrey M.	Introductory Econometrics	South-Western (2000)

Regular Font = Introductory to Intermediate
Boldface = Intermediate
Boldface and Italicized = Advanced

ENDNOTES

[1] Kennedy, Daniel W. and Peter E. Gunther, *Why do Economists' Forecasts Differ?* THE CONNECTICUT ECONOMY, (March 2005), Connecticut Center for Economic Analysis, University of Connecticut: Storrs, CT.

[2] This section draws heavily on Hendry, David F., *How Economists Forecast* in UNDERSTANDING ECONOMIC FORECASTS, Edited by David F. Hendry and Neil R. Ericsson (2003) MIT Press: Cambridge, MA., pp. 21-22.

[3] Harvey, Andrew, THE ECONOMETRIC ANALYSIS OF TIME SERIES (1990, 1993), The MIT Press: Cambridge, MA., p.2.

[4] For an explanation of the details of the OLS Method, see Pindyck, Robert S. and Daniel L. Rubinfeld, ECONOMETRIC MODELS AND ECONOMIC FORECASTS, 2nd Edition (1991), McGraw-Hill: New York, Chapter 1.

[5] This section draws on Harvey (1990, 1993), pp. 7-8.

[6] LaSage, Jim, *A Primer for ALMIS Forecasting*, University of Toledo for the Illinois Department of Employment Security: (1997): Ch. 3, Industry Forecasting Module.

[7] Diebold, Francis X., ELEMENTS OF FORECASTING, 2nd. Edition (2001), South-Western: Cincinnati.

[8] See Pindyck and Rubinfeld, (1991), p. 204.

[9] See Diebold (2001), p. 37.

[10] See Bails, Dale G. G. and Larry C. Peppers, BUSINESS FLUCTUATIONS*: Forecasting Techniques and Applications*, 2nd Ed., (1997), Prentice Hall, Chapter 4.

[11] See Pindyck and Rubinfeld, (1991), p. 210.

[12] Rickman, Dan S., "Generalizing the Bayesian Vector Autoregression Approach for Regional Interindustry Employment Forecasting", *Journal of Business and Economic Statistics* (1998) 16(1): pp. 62-72.

[13] See Nicholas Jolly, *Connecticut's Industry Clusters* (July 2005) OCCATIONAL PAPERS & REPORTS, Office of Research, Connecticut Labor Department: Wethersfield - for a discussion on Connecticut's industry clusters. VARs could be specified such that, industries included in the system are grouped by industry clusters.

[14] This section draws on Judge, Et al, Chapter 18; Lutkepohl, Helmut, *Introduction to Multivariate Time Series Analysis,* Springer-Verlag (1993): New York, Chapters 2, 5, and 10; Enders, Walter, *RATS Programming Language,* Walter Enders (2003): Distributed by Estima, Chapter 2; and *Applied Econometric Time Series*, John Wiley & Sons (1995): New York, Chapter 5, Section 5.

[15] Sims, Christopher, "Macroeconomics and Reality", *Econometrica* 48 (Jan. 1980): 1-49.

[16] See Lutkepohl, Chapter 10.

[17] This section draws on Wonnacott, Thomas H. and Ronald J. Wonnacott, *Introductory Statistics for Business and Economics, 4th Edition,* (1990) John Wiley & Sons: New York, Chapter 19; Zellner, Arnold, *Bayesian Methods for Econometrics*, (1971) John Wiley & Sons: New York; Lutkepohl, Chapter 5; Judge, Et al, Chapters 4 and 7; Todd, Richard M., "Improving Economic Forecasting with Bayesian Vector Autoregression", *Quarterly Review,* Federal Reserve Bank of Minneapolis, (8): 4 (Fall 1984).

[18] Bayes's Theorem is named after Thomas Bayes, an English Presbyterian minister, who lived from 1702 to 1761. The ideas now called 'Bayes's Theorem' appeared in his paper "*An Essay in Solving a Problem in the Doctrine of Chances*" It was published posthumously.

[19] Zellner (1971), pp. 240-246.

[20] Zellner, Arnold, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", *Journal of the American Statistical Association* Vol.57 (June 1962): pp. 348-368.

[21] LaSage (1997), Ch. 3.

[22] This discussion draws heavily on Makridakis, Spiro, Steven C. Wheelwright, and Rob J. Hyndman, *Forecasting: Methods and Applications*, 3rd Edition, John Wiley & Sons (1998): New York, Ch. 4-Section 4/3.

[23] This section draws heavily on Pankratz, Alan, *Forecasting with Univariate Box-Jenkins Models,* John Wiley & Sons (1983): New York, p. 11 and Chapters 2 and 3.

[24] This section draws heavily from Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics*, 2nd Edition, John Wiley & Sons (1988): New York, Chapter 16.

[25] See Pankratz (Endnote 23 above) or Judge, Et al (Endnote 24 above) for discussions of the Box-Jenkins Approach. There are many other statistics, econometrics, and forecasting textbooks that cover the Box-Jenkins Approach to time-series model building and forecasting.

REFERENCES

Bails, Dale G. G. and Larry C. Peppers, *BUSINESS FLUCTUATIONS: Forecasting Techniques and Applications*, 2nd Ed., (1997) Prentice Hall: New York

Diebold, Francis X., *ELEMENTS OF FORECASTING*, 2nd Edition (2001) South-Western: Cincinnati

Enders, Walter, *APPLIED ECONOMETRIC TIME SERIES*, (1995) John Wiley & Sons: New York

---------------, *RATS Programming Language*, Walter Enders (2003): Distributed by Estima: Evanston, IL.

Griffiths, William E., R. Carter Hill, and George G. Judge, *LEARNING AND PRACTICING ECONOMETRICS*, (1993) John Wiley & Sons: New York

Hall, Stephen, Editor, *APPLIED ECONOMIC FORECASTING TECHNIQUES*, (1994) Wheatsheaf: London

Harvey, Andrew, *THE ECONOMETRIC ANALYSIS OF TIME SERIES* (1990, 1993), The MIT Press: Cambridge

Hendry, David F. and Neil R. Ericsson, Editors, *UNDERSTANDING ECONOMIC FORECASTS*, (Paperback-2003) MIT Press: Cambridge

Hendry, David F., *How Economists Forecast* in *UNDERSTANDING ECONOMIC FORECASTS*, Edited by David F. Hendry and Neil R. Ericsson (2003) MIT Press: Cambridge, MA., pp. 21-22

Holden K, D.A. Peel, and J.L. Thompson, *ECONOMIC FORECASTING: An Introduction*, (1994) Cambridge U. Press:

Jolly, Nicholas A., *Connecticut's Industry Clusters* (July 2005) *OCCASIONAL PAPERS & REPORTS*, Office of Research, Connecticut Labor Department: Wethersfield

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee, *INTRODUCTION TO THE THEORY AND PRACTICE OF ECONOMETRICS*, 2nd Edition, John Wiley & Sons (1988): New York

Kennedy, Daniel W. and Peter E. Gunther, *Why do Economists' Forecasts Differ? THE CONNECTICUT ECONOMY*, (March 2005), Connecticut Center for Economic Analysis, University of Connecticut: Storrs, CT

Kmenta, Jan, *ELEMENTS OF ECONOMETRICS*, MacMillen Publishing (1971): New York, LaSage, Jim, *A Primer for ALMIS Forecasting*, (1997) University of Toledo for the Illinois Department of Employment Security: Chicago

Lutkepohl, Helmut, *INTRODUCTION TO MULTIPLE TIME-SERIES ANALYSIS*, 2nd Edition, (1992) Springer-Verlag: New York

Makridakis, Spiro, Steven C. Wheelwright, and Rob J. Hyndman, *FORECASTING METHODS AND APPLICATIONS*, 3rd Edition, (1998) John Wiley & Sons: New York

Pankratz, Alan, *FORECASTING WITH UNIVARIATE BOX-JENKINS MODELS*, (1983) John Wiley & Sons: New York

Pindyck, Robert S. and Daniel L. Rubinfeld, *ECONOMETRIC MODELS AND ECONOMIC FORECASTS*, 2nd Edition (1991), McGraw-Hill: New York

Rickman, Dan S., *Generalizing the Bayesian Vector Autoregression Approach for Regional Interindustry Employment Forecasting, JOURNAL OF BUSINESS AND ECONOMIC STATISTICS*, (1998) 16(1): pp. 62-72

Sims, Christopher, *Macroeconomics and Reality*, *ECONOMETRICA*, (January 1980) 48: 1-49

Todd, Richard M., *Improving Economic Forecasting with Bayesian Vector Autoregression*, *QUARTERLY REVIEW*, Federal Reserve Bank of Minneapolis, (Fall 1984) (8): 4

Wonnacott, Thomas H. and Ronald J. Wonnacott, *INTRODUCTORY STATISTICS FOR BUSINESS AND ECONOMICS*, 4th Edition, (1990) John Wiley & Sons: New York

Wooldridge, Jeffrey M., *INTRODUCTORY ECONOMETRICS*, (2000) South-Western: Cincinnati

Zellner, Arnold, *An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias*, *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, (June 1962) Vol.57: pp. 348-368

------------------, *BAYESIAN METHODS FOR ECONOMETRICS*, (1971) John Wiley & Sons: New York