Cleaning up the Errors in the Monthly *Employment Situation* Report: A Multivariate State-Space Approach

Mark W. French

December 1997

JEL # C32, J21
Keywords: signal extraction, Kalman filter, Employment Situation

Board of Governors of the Federal Reserve System
Division of Research and Statistics
20th and Constitution Ave., NW
Washington, DC 20551

I would like to thank David Reifschneider and workshop participants at the Federal Reserve Board for their valuable comments.

ABSTRACT

This paper estimates the underlying state of the labor market, assuming data in the monthly Employment Situation are contaminated by measurement error and other transient noise. To better filter out unobserved noise, the methodology exploits correlations among labor-market series. Household employment and labor force have cross-correlated sampling errors; establishment employment and hours-worked may, also. The Kalman filtering procedure also exploits fundamental economic relationships among these series. Error cross-correlations and economic relationships shape a multivariate labor-market model where observed variables embody unobserved components: trend, cycle and noise. Maximum-likelihood estimation enables construction of labor series from which noise components have been removed.

Summary

This paper presents a methodology to estimate the underlying state of the labor market, under the assumption that the data in the monthly Employment Situation are contaminated by measurement error and other sources of transient noise. That release is based on two surveys, one of establishments and another of households. To better filter out unobserved noise, the methodology exploits certain correlations among the various labor-market series. For example, the aggregate employment and labor force series from the household survey have sampling errors which are correlated; likewise, the employment and hours-worked series from the establishment survey have sampling errors which may be correlated. The filtering procedure also takes advantage of the fact that all four of these series are tied together by fundamental economic relationships. Specifically, the error cross-correlations and the economic relationships are incorporated into a multivariate labor-market model in which all variables are assumed to be composed of three unobserved components: trend, cycle and noise. Maximum-likelihood estimation of the system enables the construction of filtered labor series from which the noise component has been removed. Appendix 1 discusses the solution procedure for the state space model.

An earlier study by Bryan and Cecchetti (1993) made use of a multivariate state-space model to filter out noise from components of the consumer price index and extract a measure of overall core inflation. This study uses similar solution procedures, described in Appendix 1, to jointly filter labor market variables. In contrast to Bryan and Cecchetti, however, this study uses several fundamental economic relationships to help in the signal extraction process, as discussed below.

The filtered labor-market series are potentially valuable aids to understanding the recent and near-future course of the economy. For example, filtered hours series tell a rather different story than the raw data about current-quarter activity, in the context of an "hours-to-output" methodology.

Linking the observed and "true" state of the labor market

Each of four data series taken from the labor release[1] is assumed to have three unobserved components: trend, cycle, and one-period "sampling" noise. (In addition, an exogenous proxy for bad weather is assumed to affect hours worked.) Specifically, the observation equations take the form (all variables are in logs):

1. Household employment (HHE)

HHE = trend component (HHET) + cyclical component (HHEC) + noise component (eHHE)

2. Establishment employment (EE)

EE = trend component (EET) + cyclical component (EEC) + noise component (eEE)

3. Labor force (LF)

LF = trend component (LFT) + cyclical component (LFC) + noise component (eLF)

4. Production-worker hours (HRS)

HRS =trend components (HRST) +cyclical component (HRSC) -.036 * bad weather proxy (BW)
(4.4)
+ noise component (eHRS)

Modeling the relationships among the "sampling errors"

Sampling error and other non-serially-correlated shocks to the observation equations are lumped together in the "noise component" of the four observation equations. In a typical filtering exercise, the covariance of these noise components is assumed to be zero. In this case, however, sampling errors for the two series from the household

---

[1] Production-worker hours, establishment employment, household employment and civilian labor force.

survey (eHHE and eLF) are likely to be cross-correlated. *A priori* the sampling errors for the two series from the establishment survey might also be expected to be correlated.

As we will see, this noise in the observation equations is in practice very important for labor force, household employment and hours, though not so important for establishment employment.

The underlying state of the labor market

The trend and cyclical components of the four published labor-market aggregates represent the state of the labor market in our model. Like the sampling errors, these eight variables are not observable. However, it is possible to form estimates of the trend and cyclical components by linking them structurally. Specifically, this is done by assuming that these state variables follow a restricted vector autoregressive process, as outlined below:

Trend components

Labor supply considerations are assumed to dominate long-run trends in the model. Several variables -- growth of working-age population (N16), trends in participation rates, a "natural rate" of unemployment, and trends in the workweek -- work together to determine trend levels of production-worker hours. Trend[2]growth of labor force equals the exogenous growth of working-age population, plus an adjustment for trend growth in labor-force participation. The latter is estimated to be more rapid in the 1990s than in the first part of the sample (January 1984 through December 1989).

---

[2] Initially, each of the four trend variables was assumed to be affected by a permanent stochastic shock, as well. But the variance of the stochastic shocks ended up being insignificant: there was no need to go beyond the simple deterministic (split) trends used for this sample period. Since the unconditional mean of all the cycle and noise components is zero, the Kalman filter and smoother select an unbiased estimate of the trend level even when this differenced formulation is used.

Trend growth of household employment depends one-for-one on trend growth of the labor force: a stationary "natural rate" of unemployment was not rejected for the sample period. Trend growth of establishment employment equals trend growth in household employment, adjusted for an increasing discrepancy between the two series over the sample period. Trend growth of production-worker hours depends one-for-one on trend growth in establishment employment, with a correction for trend movements in the workweek.

Cyclical components

In the specification of the model, the cyclical state of the labor-market aggregates is linked to the lagged cyclical state of labor hours , as well[3] as to contemporaneous shocks to each of the aggregates (uHRSC, uEEC, uHHEC and uLFC). The model specification also incorporates the significant covariance among the contemporaneous shocks to the cyclical components of hours and employment.

The final version of these cyclical equations has a fairly simple form. The cyclical component of production-worker hours depends on two lags of itself. The cyclical component of establishment employment depends on two lags of itself, and two lags of the cyclical part of production-worker hours. The cyclical component of household employment depends on two lags of itself, and two lags of the cyclical part of production-worker hours. Finally, the cyclical component of labor force depends on one lag of itself, and two lags of the cyclical part of household employment.

Early versions of the model were less restrictive; for example, three lags were allowed for all variables in the above specification. However, the coefficients of those additional terms were not significantly different from zero, and were omitted in the final specification.

---

[3] Production-worker hours clearly Granger-cause employment, for this sample period and frequency--in the sense that lags of the cycle in hours were clearly significant in the state equations for the cycle in establishment and household employment. On the other hand, it's not clear whether employment Granger-causes hours: the first two lags of the cycle in establishment employment were not clearly significant in the state equation for production-worker hours.

The estimated state equations**, **January 1984 - October 1997

The system has behavioral equations for eight state variables. Estimation[4] results for these state equations are laid out below, with t-statistics in parentheses.

5. Log of trend labor force (LFT)

$$LFT = LFT_{-1} + \Delta N16 + \underset{(2.3)}{.0043/100} + \underset{(12.0)}{.0364/100} * DUM90FORWARD$$

6. Log of trend in household employment (HHET)

$$HHET = HHET_{-1} + LFT_{-1} - LFT_{-2}$$

7. Log of trend in establishment employment (EET)

$$EET = EET_{-1} + HHET_{-1} - HHET_{-2} + \underset{(8.6)}{.0325/100}$$

8. Log of trend in production-worker hours (HRST)

$$HRST = HRST_{-1} + (EET_{-1} - EET_{-2}) - \underset{(18.4)}{.0158/100}$$

9. Log of cyclical component of production-worker hours (HRSC)

$$HRSC = \underset{(182.4)}{.99} * HRSC_{-1} + \underset{(12.9)}{.82} * \Delta HRSC_{-1} + uHRSC$$

standard deviation of uHRSC = .090x12=1.1%, annual rate

---

[4] Beyond these eight, there are several state identities, which allow for a transformation of a higher order vector autoregression into the standard AR(1) form for the state equations.

10. Log of cyclical component of establishment employment (EEC)

$$EEC = .88*EEC_{-1} - .28*\Delta EEC_{-1} + .10*HRSC_{-1} + .53*\Delta HRSC_{-1} + (uEEC).$$

(40.9) (3.0) (5.7) (4.6)

standard deviation of uEEC = .093x12=1.1%, annual rate

11. Log of cyclical component of household employment (HHEC)

$$HHEC = .98*HHEC_{-1} + .13*\Delta HHEC_{-1} + .011*HRSC_{-1} + .34*\Delta HRSC_{-1} + uHHEC$$

(104.5) (0.7) (1.3) (3.6)

standard deviation of uHHEC = .088x12=1.0%, annual rate

12. Log of cyclical component of labor force (LFC)

$$LFC = .78*LFC_{-1} + .045*HHEC_{-1} + .35*\Delta HHEC_{-1} + uLFC$$

(10.0) (2.7) (3.2)

standard deviation of uLFC = .093x12=1.1%, annual rate

Covariance of errors to the state equations

The errors of the state equations (5) through (12) are assumed independent, with three exceptions. We estimate the covariance between cyclical shocks to (1) household and establishment employment; (2) household employment and hours; and (3) establishment employment and hours.[5] All three of these covariances end up being nontrivial in practice. The standardized error covariances (i.e. error cross-correlations) between the cyclical components are as follows:

- establishment employment and production-worker hours, .78
- household employment and production-worker hours, .83
- household employment and establishment employment, .33

---

[5] In fact, I estimate the nonzero terms of the Choleski decomposition of the covariance matrix of the errors in the state equations--to avoid any possibility of negative estimated variances. From the estimates in the decomposition I calculate the covariance matrix itself.

This last cross-correlation seems implausibly low, given that the other two are so high. However, the low value doesn't seem to have caused problems with the estimates of the state itself: this is a loose end to tie up in future work.

Variance/covariance matrix of the errors in the observation equations

Estimation results suggest that production-worker hours are measured with a great deal of error ,[6] with somewhat smaller but still very noticeable errors in measured labor force and household employment. Error in measurement of establishment employment was apparently not significant.

- standard deviation of eHHE = .15%, monthly rate
- standard deviation of eEE = .0%, monthly rate
- standard deviation of eLF = .13%, monthly rate
- standard deviation of eHRS = 0.29%, monthly rate
- cross-correlation between the errors eHHE and eLF = .954

For comparison, the standard deviation of observed hours growth over the sample period from 1984 to 1997 was 0.53% at a monthly rate. The standard deviation of observed labor force growth over the sample period was .20% at a monthly rate. It appears that half or even a bit more of the monthly movement in observed hours and labor force is sampling error and other non-serially correlated noise.

<u>Estimation of the state vector</u>

The key output of the state-space estimation process is not the set of state-space model parameters. Rather, it is the estimate of the unobserved state of the labor market, each month — an estimate generated by the Kalman filter. One objective is to calculate and display the observed labor force, employment and hours series net of sampling error, to better estimate the underlying month-to-month changes in these

---

[6] Part of the apparent error in measuring hours might plausibly be shifts in strike activity. However, a strike proxy (along the lines of the weather proxy in the observation equation for hours) turned out to be insignificant.

series. For labor force and employment, this filtered series is just the sum of the estimated cyclical and trend components — displayed in figures 1 and 2. For production-worker hours, the model also distinguishes an additional weather-related component. Thus for hours, the model's underlying series can be defined in two alternative ways. Figure 3a shows underlying hours including weather shocks, while figure 3b nets out both sampling error and weather shocks.

The state of the labor market can be redefined in terms of participation rate and workweek, rather than labor force, and aggregate hours. The underlying or filtered participation rate is defined as the filtered labor force data divided by the published figures for working-age population. A proxy for the underlying or filtered workweek is defined as the filtered hours series divided by the filtered series for establishment employment.[7] The underlying or filtered versions of participation and workweek are compared with the observed counterparts in figures 4 and 5.

Applications

The filtered hours data have the potential to yield more accurate estimates of current-quarter growth in output, in the context of an hours-to-output framework (Braun, 1990). Table 1 compares the published data with the corrected data for production-worker hours and for civilian labor force. Growth rates differ dramatically in January, February April, and July of 1997. (Currently the filtered estimate of hours is about 0.2 percent above the published level). For the third quarter of 1997 overall, corrected hours growth was about 0.6 percent faster at an annual rate than the published data (1.9 percent versus 1.3 percent).[8] Such measurement problems may

---

[7] Strictly speaking, one would use total hours, rather than production-worker hours, in the numerator — yielding an exact counterpart to the published workweek series. Unfortunately, this would add to the size and complexity of the model: the "curse of dimensionality" in going from four to five observed series would slow solution times substantially. The two hours series are not cointegrated, but they do move together fairly closely from month to month.

[8] This corrected growth rate goes well beyond what one would get by simply removing the effects of the United Parcel Service strike. The strike might have taken about 0.2 percent off the

explain part of the spectacular surge in productivity recorded in the third quarter. For labor force growth, the difference between published and corrected growth rates are not as striking on a quarterly basis, but can be quite significant from month to month. The less noisy monthly pattern for both corrected series is clear in figures 1 and 3.

Conclusions

The results presented above suggest several conclusions. First, observed production-worker hours and labor force are very noisy series. Judging from the standard errors, the month-to-month movements in these series result from sampling error and other non-serially-correlated shocks at least as much as they represent changes in the actual state of hours or labor force.

Second, there is a substantial correlation between the noise component of labor force and that of household employment. The multivariate signal-extraction methodology presented here exploits this correlation, as well as the dynamics of workweek and labor-force participation implicit in the state equations. It thus offers a more precise estimate of the underlying growth of labor-market aggregates than is likely using univariate methods.

This paper's estimates of underlying growth in labor force and hours worked differ substantially from the published monthly data. At times, the published figures for hours growth appear to be distorted substantially even on a **quarterly** basis. Given this result, it is probably unwise to accept published monthly changes in hours and labor force at face value. Filtering is needed, and this paper offers a means of reading through observation noise to get a better measure of underlying movements in hours worked and labor force.

---

annualized growth rate of hours in the third quarter of 1997.

.

Table 1: Published versus corrected growth of hours and labor force

	Production-Worker Hours Growth		Civilian Labor Force Growth	
	Published	Corrected	Published	Corrected
Jan 97	-0.6%	0.3%	0.4%	0.2%
Feb 97	1.3%	0.3%	-0.2%	0.1%
Mar 97	0.1%	0.3%	0.5%	0.1%
Apr 97	-0.4%	0.3%	-0.2%	-0.1%
May 97	0.3%	0.2%	0.1%	0.0%
Jun 97	0.4%	0.1%	0.0%	0.2%
Jul 97	-0.3%	0.2%	0.1%	-0.0%
Aug 97	0.4%	0.1%	0.1%	0.1%
Sep 97	-0.1%	0.1%	-0.0%	0.1%
Oct 97	0.2%	0.2%	-0.1%	-0.0%
Q1 96 (car)	0.4%	2.4%	2.0%	1.6%
Q2 96 (car)	5.3%	3.7%	1.5%	1.6%
Q3 96 (car)	2.6%	3.1%	1.5%	1.5%
Q4 96 (car)	3.0%	2.5%	2.2%	2.2%
Q1 97 (car)	4.0%	3.2%	2.4%	2.3%
Q2 97 (car)	1.7%	2.8%	0.7%	0.5%
Q3 97 (car)	1.3%	1.9%	0.7%	0.9%

Figure 1
Civilian Labor Force

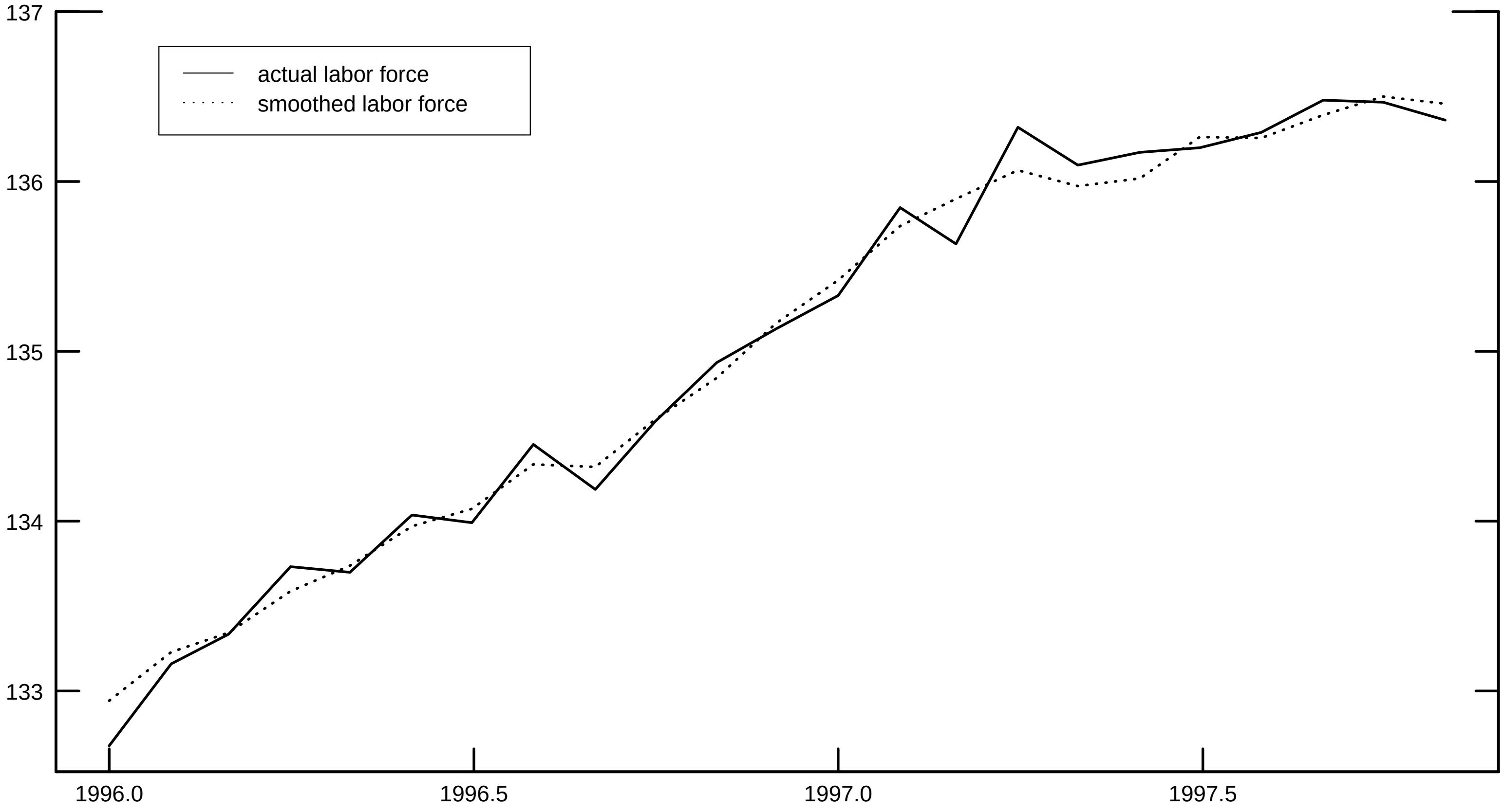


Figure 2
Household Employment

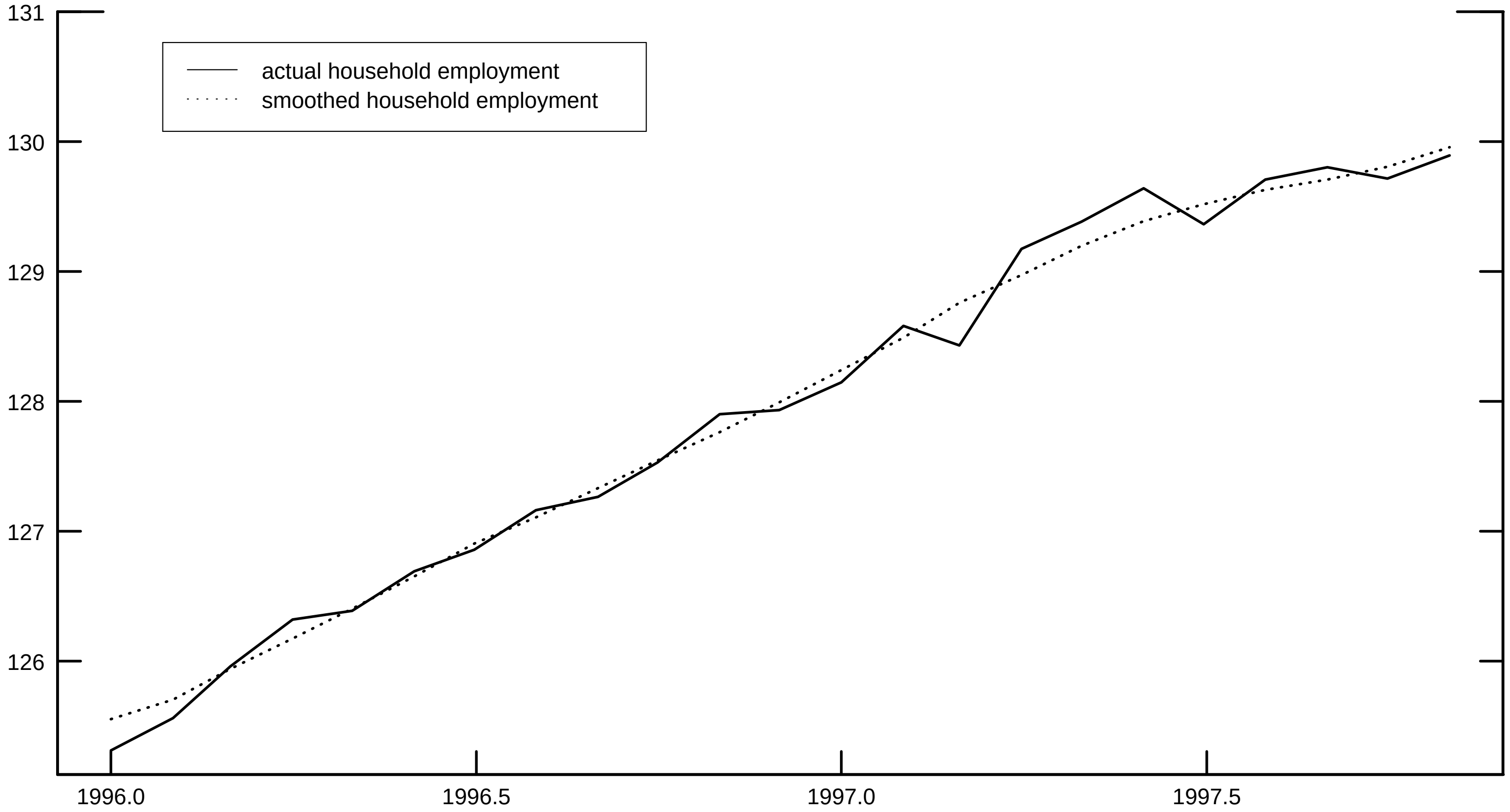


Figure 3A
Production-Worker Hours

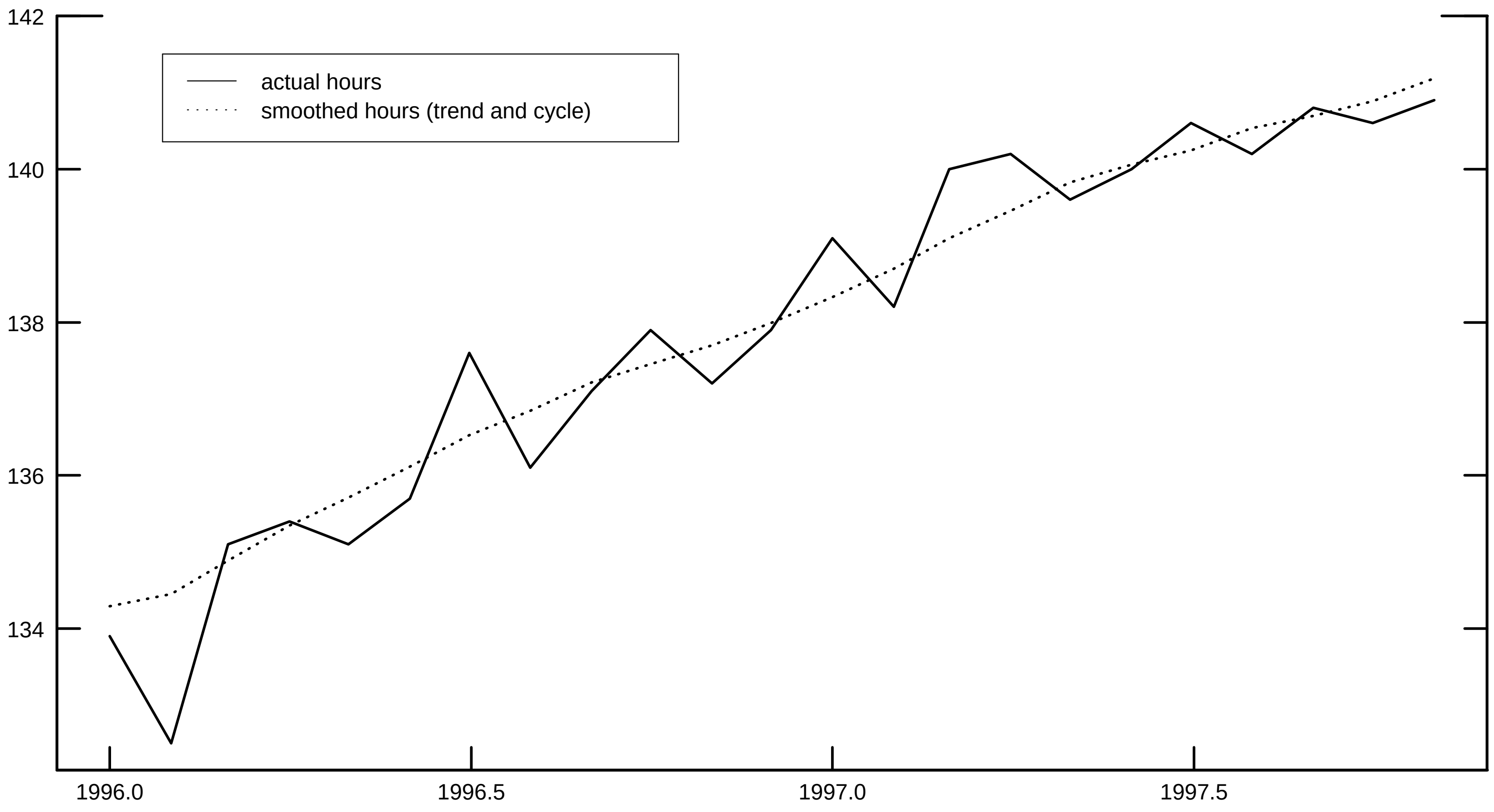


Figure 3B
Production-Worker Hours

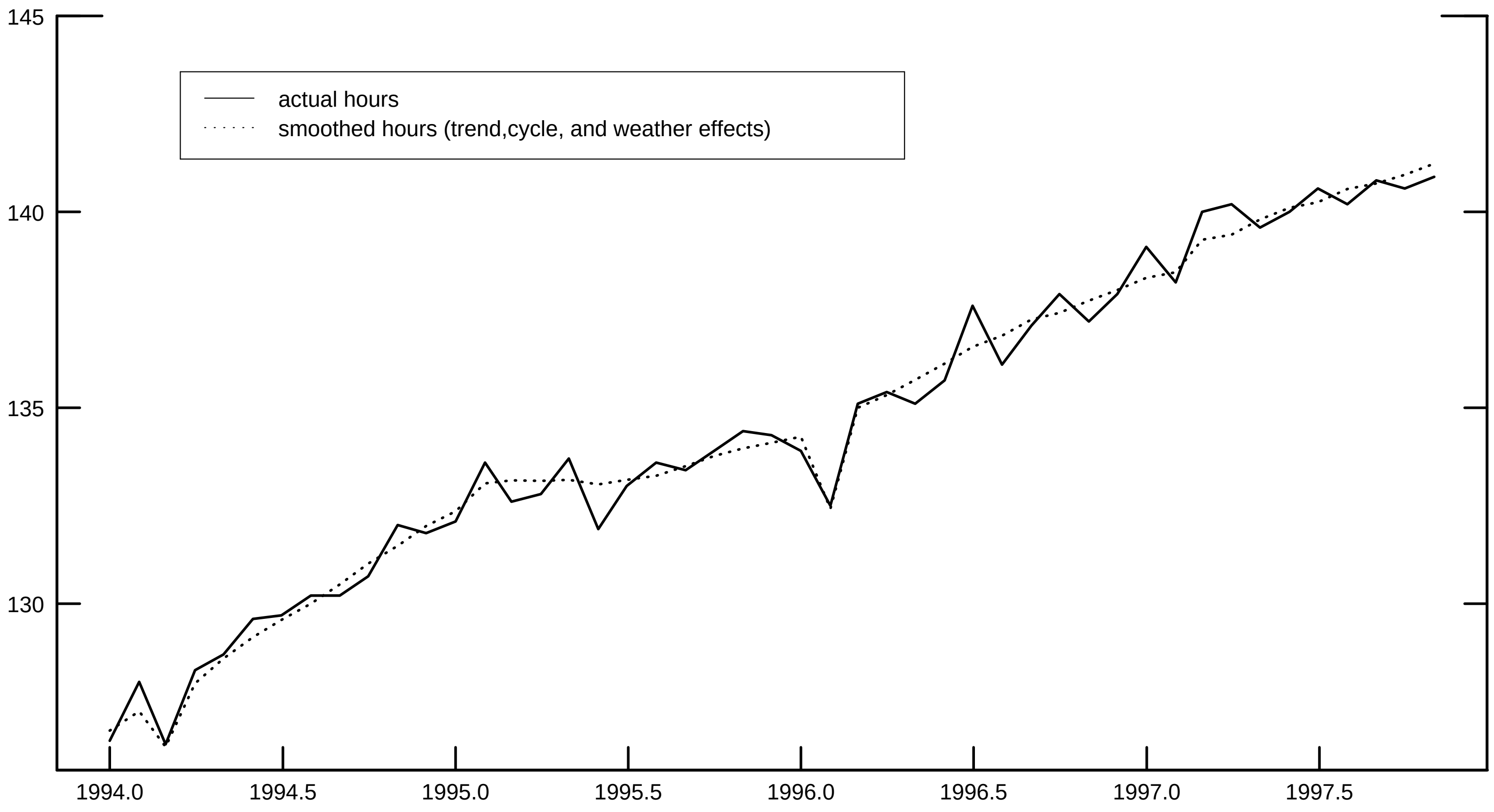


Figure 4
Labor-Force Participation Rate

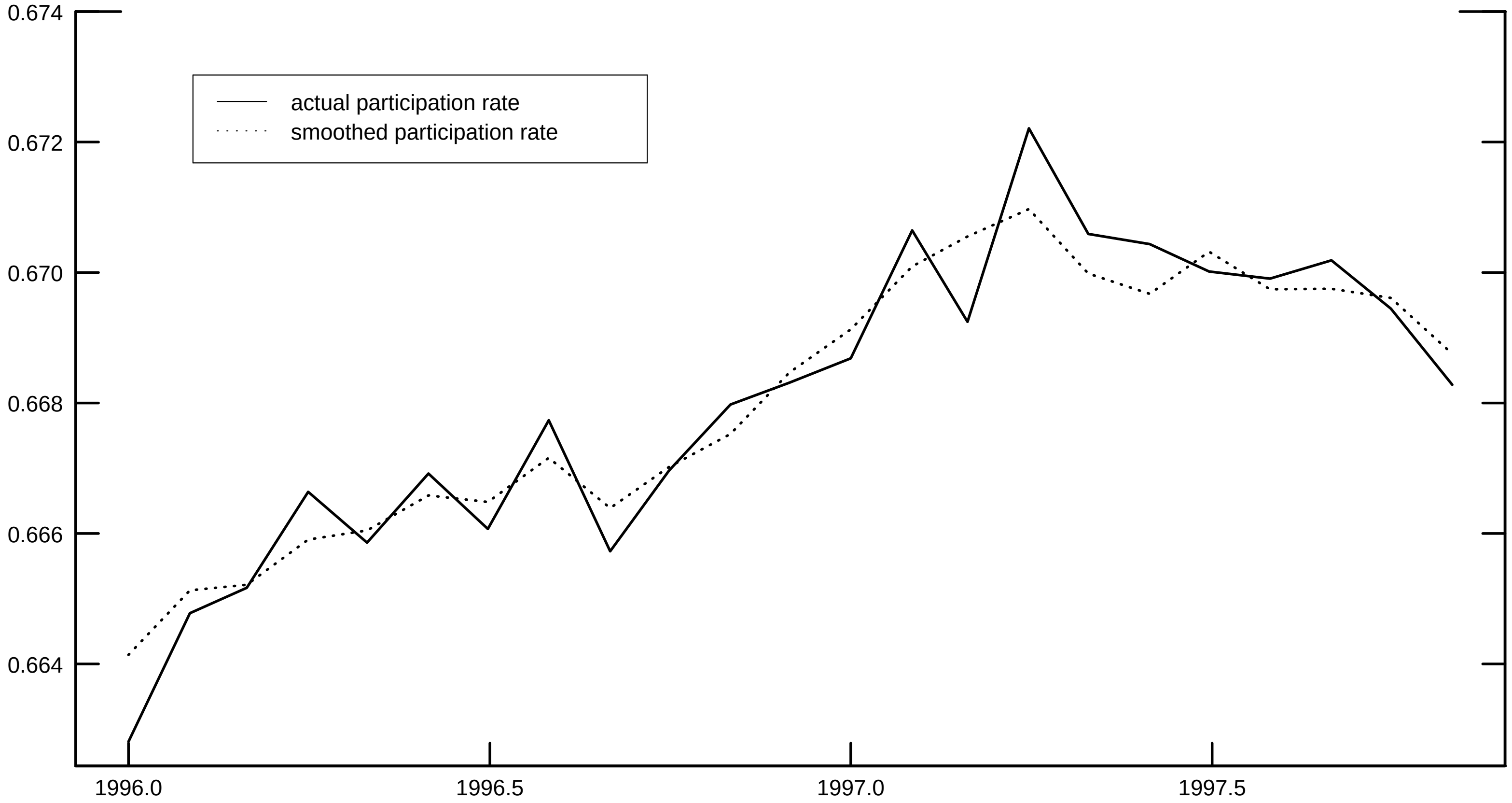


Figure 5
Workweek Proxy

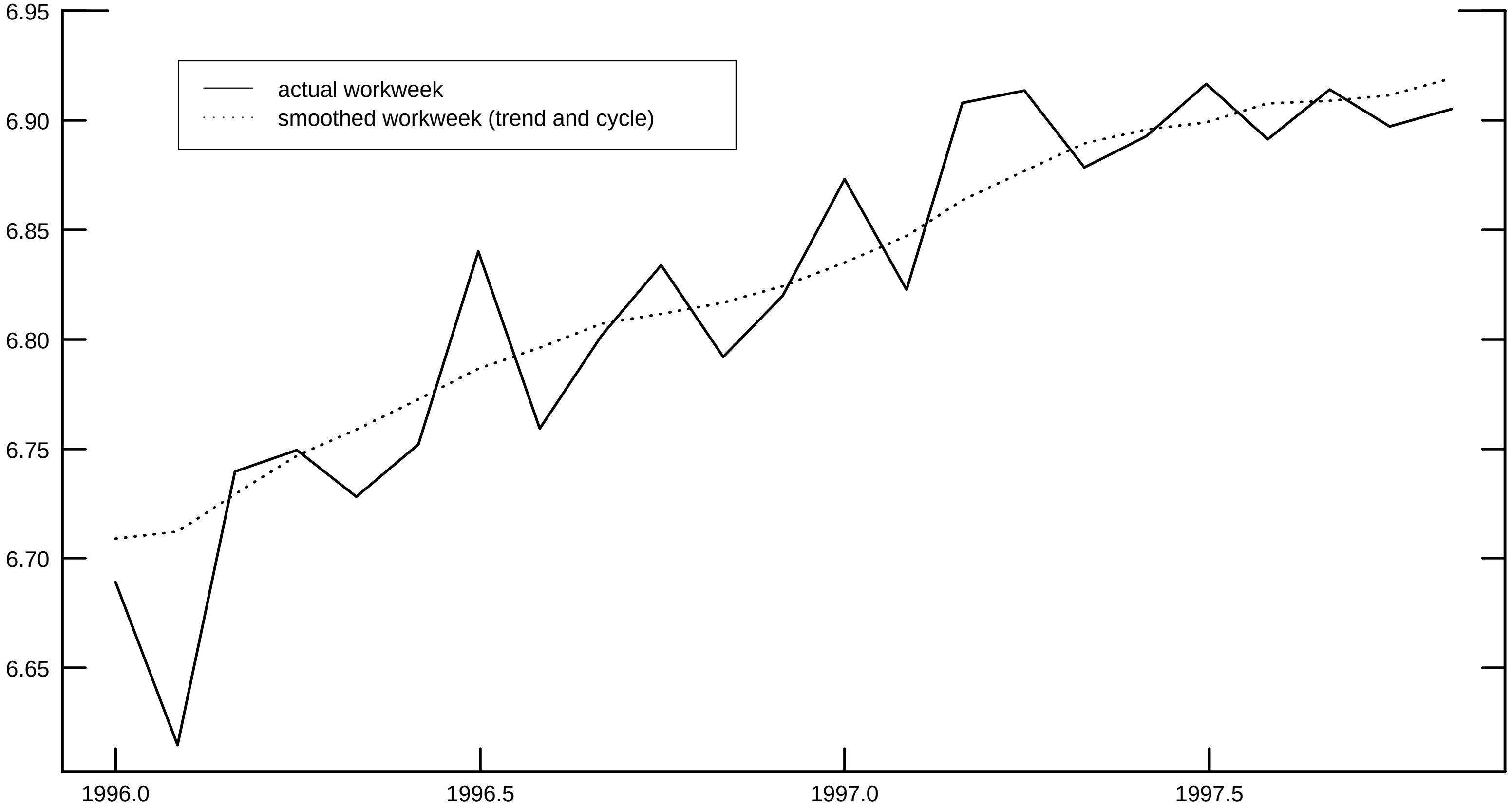
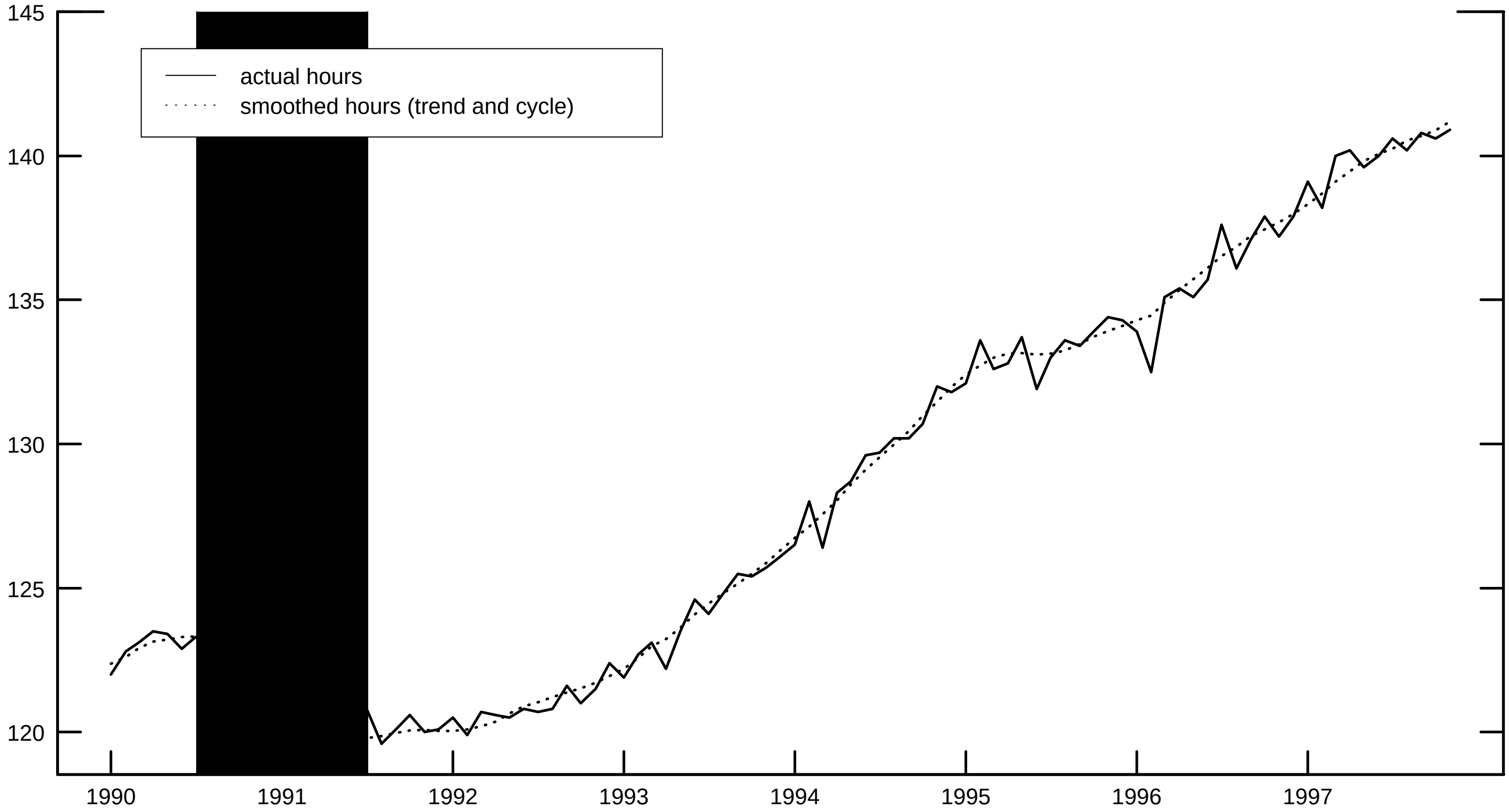


Figure 3A’, 1990-1997
Production-Worker Hours



Appendix 1

The Hamilton 1994 model

The state-space model presented in this paper is a variant of the model presented in Hamilton (1994, p. 399). Hamilton's model is as follows (changing notation slightly):

I. Let y_t be a vector of observed endogenous variables at time t.
Let x_t be a vector of observed exogenous or predetermined variables.
Let z_t be a vector of unobserved state variables.

Let A, H, F, Q and R be matrices whose elements are functions of x_t.
Let w_t and v_t be Gaussian error vectors.

The observation equation is:

1. $y_t = A(x_t) + [H(x_t)]'z_t + w_t$

The state equation is:

2. $z_t = F(x_t)z_{t-1} + v_t$

Conditional on x_t, the two error vectors v_t and w_t are independent of each other. Each of the two error vectors is mean-zero and Gaussian: however, their covariance matrices can change with x_t over time. That is,

3. $$\begin{bmatrix} v_t \\ w_t \end{bmatrix} \Bigg| \; x_t \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q(x_t) & 0 \\ 0 & R(x_t) \end{bmatrix} \right)$$

The six equations of the Kalman filter for Hamilton's model are as follows :[9]

First, the definition of the vector of prediction errors in forecasting the observed endogenous variables:

4. $e_t = y_t - A - H'z_{t|t-1}$

The MSE matrix of this prediction error vector is

5. $MSE_t = H'P_{t|t-1}H + R$

6. Updating equation for the estimated state --

$$z_{t|t} = z_{t|t-1} + P_{t|t-1} \ H(MSE_t) \ e_t$$

7. MSE of the estimate of the state --

$$P_{t|t} = P_{t|t-1} - P_{t|t-1} \ H(MSE_t) \ H'P_{t|t-1}$$

8. Forecast of the state --

$$z_{t+1|t} = F \, z_{t|t}$$

9. MSE of the forecast of the state --

$$P_{t+1|t} = F \, P_{t|t} \, F' + Q$$

---

[9] We omit the arguments x_t for compactness.

To calculate the parameters of Hamilton's state-space model,
one can maximize the sum of the log-likelihood of the prediction errors conditional on x_t. The likelihood function at time t (call it L_t), conditional on x_t is the multivariate normal density --

10. $L_t = N(e_t, MSE_t)$

REFERENCES

Braun, Steven (1990), "Estimation of Current-Quarter Gross National Product by Pooling Preliminary Labor-Market Data", Journal of Business and Economic Statistics, pp. 293-304, July.

Bryan, Michael and Stephen Cecchetti (1993), “The Consumer Price Index as a Measure of Inflation”, Federal Reserve Bank of Cleveland Economic Review, Volume 29 #4,
pp.15-24.

Hamilton, James D. (1994), Time Series Analysis, Princeton University Press, Princeton, NJ.

U.S. Department of Labor, Bureau of Labor Statistics, Employment Situation, various issues.

Appendix 2: **Data adjustments for labor force, population and establishment employment**

Labor force

The series for civilian labor force is the monthly aggregate series published by the Bureau of Labor Statistics. However, that monthly series is first adjusted for breaks as follows:

January 1990 to present --(published data)

January 1986 to December 1989--(published data)*1.0084

January 1978 to December 1985--(published data)*1.0084*1.0034

Working age population

Working age population (16 and over) is the sum of several monthly component series published by the Bureau of Labor Statistics, adjusted for breaks as follows:

January 1990 to present -- (published sum)

January 1986 to December 1989 -- (published sum)*1.0059

January 1972 to December 1985 -- (published sum)*1.0059*1.0023

Establishment employment

These data for April 1996 to April 1997 are corrected for the ”resizing” which occured at the end of that period.