Improving the Accessibility of the NBER's Historical Data

Daniel Feenberg
NBER

Jeffrey A. Miron
Department of Economics, Boston University and NBER

July 28, 1995

We are indebted to Geoffrey Moore for providing us with the NBER's handwritten data sheets, to Anna Schwartz for helpful comments, and to Tianlun Jian and Maria Borga for research assistance.

Improving the Accessibility of the NBER's Historical Data

Daniel Feenberg
NBER

Jeffrey A. Miron
Department of Economics, Boston University and NBER

Abstract

During the early years of its existence, the National Bureau of Economic Research (NBER) assembled an extensive data set on all aspects of the pre-WWII macroeconomy. Until 1978, this data set existed only on the handwritten sheets to which the early NBER researchers copied the data from original sources. In 1978, the Inter-University Consortium for Political and Social Research (ICPSR) transferred the data to magnetic tape.

A number of researchers have used the ICPSR tape, but two key problems discourage many from taking advantage of this unique data set. The first is that modern econometric software does not have the ability to read the obsolete ICPSR format. The second is that the process of transferring the data from the NBER's handwritten sheets to the ICPSR tape introduced a number of mistakes. We have eliminated these two impediments to use of the NBER data set by converting the ICPSR tape to a portable format and by verifying the accuracy of the data using the NBER's original handwritten sheets. The data set is now available on the Internet and can be accessed using standard gopher or web-browser software.

JEL Numbers: C82, E32, N12
URN=:IANA716::nberwo5186:::

Daniel Feenberg
NBER
1050 Massachusetts Avenue
Cambridge, MA 02138
(617) 868-3900
feenberg@nber.harvard.edu

Jeffrey A. Miron
270 Bay State Rd.
Dept. of Economics
Boston University
Boston, MA 02215
(617) 353-4442
jmiron@acs.bu.edu

Since its founding in 1920, one goal of the National Bureau of Economic Research (NBER) has been the accurate measurement of the aggregate economy. Especially during its early years, the NBER devoted considerable resources to data collection, culling this information from both government and private sources. These efforts proved extremely valuable to later researchers in both academia and government. For example, they led directly to development of the index of leading economic indicators and aided greatly in the production of Friedman and Schwartz's (1963) *A Monetary History of the United States*.

When the data were collected (mainly during the interwar period), a technology for storing them electronically was not available, so the NBER researchers copied them by hand from the original sources onto data sheets. In 1978, the Inter-University Consortium for Political and Social Research (ICPSR) in Ann Arbor, Michigan, transferred the data from the NBER's handwritten sheets to magnetic tape. This tape is available to the public from the ICPSR. A number of researchers have used the tape, but two key problems discourage many from taking advantage of this unique data set.

The first problem with the ICPSR tape is that most macroeconomic researchers now use PC's rather than mainframe computers. This means that use of data on the ICPSR tape requires the cumbersome task of reading the tape on a mainframe

and then downloading the data to a PC. Since the tape contains over 7,500 series, downloading the entire data set is time consuming. Additionally, the tape's structure makes it difficult to extract a single series or a small group of series. Thus, the data are accessible in the existing format, but they would be far more accessibile in a format compatible with common time series packages for PCs and workstations.

A second problem is that the process of transferring the data from the NBER's handwritten sheets to the ICPSR tape introduced a number of mistakes. These take several forms, which we detail below. The fact that these mistakes are known to exist exerts a chilling effect on most researchers' desire to use the ICPSR tape.

This paper describes our efforts to verify the data on the ICPSR tape and convert the data set to an accessible format. For reasons that we explain below, this task was less daunting than it may appear; in particular, it did not involve consulting the original sources from which the NBER obtained the data. Instead, we verified the data from the NBER handwritten sheets.

In the remainder of the paper, we first describe the data set contained on the ICPSR tape and discuss its value to empirical researchers in macroeconomics. The second section of the paper documents the severity of the mistakes present in the ICPSR tape, while the third explains the procedures we used to correct the data. The final section explains how to obtain the revised and verified data set.

1 The NBER's Macroeconomic Data Set

The NBER macroeconomic data set was compiled mainly during the interwar and early post-WWII periods. The NBER researchers collected data on all consistent time series related to economic activity, so the data set is extensive and includes measures of the economy at a high level of disaggregation. The data set covers all aspects of the economy, including production, construction, employment, money, prices, asset market transactions,

foreign trade, and government activity. In many cases the series exist at the monthly or quarterly frequency. The data set has some coverage of the United Kingdom, France and Germany, although it predominantly covers the United States.

The most important characteristic of the data set is that it provides extensive coverage of the pre-WWI and interwar economies. It thus provides for these years a data set analogous to those widely available for the post-WWII period (e.g., Citibase). No such machine-readable, macroeconomic data set is currently available for the pre-WWII period.

The availability of accurate data for the pre-WWII years is of considerable importance for empirical macroeconomics, which in recent years has witnessed an

enormous resurgence of interest in historical research. The reason for the importance of pre-WWII data is that the interwar and pre-WWI economies often provide natural laboratories in which to evaluate competing economic paradigms. Learning from these historical episodes requires accurate measures of the economy's performance, and the NBER data set provides such measures.

2 Problems with the Existing ICPSR Data Tape

As mentioned above, the ICPSR tape version of the NBER data set contains mistakes of various kinds. In this section we describe what is known about the nature and magnitude of these mistakes.

We discovered the mistakes described below by the following process. We first obtained xeroxes of all the original NBER handwritten data sheets from the offices of Geoffrey Moore's Center for International Business Cycle Research (CIBCR) at Columbia University.[1] We also produced a printout of all of the series on the ICPSR tape. We then used these two materials to conduct an initial, preliminary check of the contents of the tape. This check did not verify each entry in every series but instead checked whether each tape series appeared to roughly match the underlying NBER series. This first round inspection also checked a sample of

[1] The CIBCR became the repository for the data in 1978 after production of the ICPSR tape.

individual entries to determine the frequency of random typos and other problems. This investigation revealed the following kinds of errors in the tape data.

The tape data contain numerous typos. In some cases these are inconsequential (e.g., only in the last decimal place), but in many cases they are substantial. The distribution of these typos appears random, i.e., the kind that would result from typing in data from handwritten sheets.

In at least fifty series, the decimal point has been entered in the wrong place for all or part of a series. In some cases the units were simply entered incorrectly. In other cases, the units change part way through a series according to the NBER documentation (say from thousands to millions), but this change was ignored when the data were entered onto the tape.

The documentation on the tape contains numerous mistakes. For example, in one case the NBER documentation says the source of the data is the 1935 Annual Supplement to the Survey of Current Business but the tape documentation says it is the 1953 Supplement. The documentation on the tape accompanying one series say a strike affected output in 1922, while the NBER documentation says the strike occurred in 1912.

The documentation for many series refers to the documentation for other series in the data set using code numbers from the original NBER data sheets.

These numbers do not appear in the tape documentation, however, so these cross references provide no useful information to persons using the tape but without access to the original NBER documentation.

For some series many values that are coded as missing on the ICPSR tape are in fact present on the NBER data sheets. For other series the NBER data sheets report certain values as missing but non-missing values appear in the tape series.

When the tape contains both an annual and a monthly version of a given series (or an annual and a quarterly version) the annual version is always the sum of the monthly (quarterly) values that ICPSR entered onto the tape. This creates several problems. First, any errors made by ICPSR in entering the underlying monthly or quarterly series are carried over to the annual series. Second, in some cases the appropriate procedure for creating an annual series from an underlying monthly series is to divide the summed monthly values of the underlying series by twelve (as with an index) rather than simply to sum the monthly values. In many cases where appropriate this division by twelve (or four) has not been carried out, while in other cases where not appropriate it has been carried out.

A few series on the tape have been incorrectly titled. For example, imports of foodstuffs to France is titled as imports of manufactured goods to France. One purported U.S. series is actually a U.K. series.

Finally, several useful series for which NBER data sheets exist were not entered onto the tape.

In our view, the nature and extent of the mistakes documented above makes the ICPSR tape version of the NBER data set extremely unattractive to potential users. Anyone interested in using such data would of necessity check every entry against either the original source material or the NBER data sheets. In many cases the original sources are difficult to find, since pre-WWII materials are often not readily available at university libraries. The CIBCR is more than willing to provide xeroxes of the NBER documentation when asked, but it is not reasonable to expect their limited staff to accommodate frequent requests for such documentation. Even if it were feasible to compare the tape data to one of these two sources, the necessity of this verification means that use of the tape provides only a modest time saving over obtaining the data from original sources. The fact that the tape is time consuming and difficult to read exacerbates this problem.

We emphasize that the work performed by the ICPSR in producing the existing tape is extremely valuable, despite the limitations discussed above. The additional effort needed to make the data set readily accessible is modest in comparison to the original task of entering all of the data onto the tape.

3 Making the ICPSR Tape Data Available for the PC

We now explain the procedures we employed in converting the existing, ICPSR tape data set into a reliable, fully verified PC format data set.

The basic issue to be addressed is whether verification of the data set required checking every series against original sources. This would have been a time-consuming task, to say the least. With the exception of the typos introduced onto the NBER handwritten sheets by NBER researchers, however, all the mistakes in the ICPSR data tape were attributable to ICPSR and could be eliminated using only the NBER's handwritten sheets. Further, we believe that most if not all of the random typos in the tape version of the data set were introduced by ICPSR in producing the tape rather than by the NBER in producing the handwritten sheets. The reason for this conclusion is as follows.

In the course of earlier work (Miron and Romer, 1991), Christina Romer and Miron used the NBER tape as the source of several data series. Because of ambiguities in the tape's documentation of several series, Romer and Miron checked each of these tape series against the original sources. During this process they discovered numerous inconsistencies between the tape data and the original

sources. At this point they did not know whether the mistakes had arisen during the NBER's process of entering the data onto its handwritten sheets or during the ICPSR's process of entering the handwritten data onto the tape. Subsequent examination of the NBER sheets determined that while the NBER did introduce a small number of typos in producing the handwritten sheets, these were always inconsequential. The typos introduced by ICPSR, by contrast, were often large.

In addition, the NBER staff in New York is confident that most typos in the tape data reflect coding mistakes by the ICPSR, since the NBER researchers in the 1920's and 1930's checked the original data sheets many times. Miron and Romer's experience confirms the current NBER staff's confidence in the quality of the NBER's handwritten sheets. Indeed, in several cases the NBER data sheets are more reliable than original sources because the NBER detected and corrected typos that existed in these sources. For example, in one case the monthly numbers in an original sources did not add to the annual total, so the NBER used other sources to determine whether the monthly or annual data were correct.

We have therefore checked the tape data only against the handwritten NBER sheets. As mentioned above, xeroxes of these sheets have been provided by Geoffrey Moore's office at the CIBCR; these xeroxes were produced when the original sheets were sent to the ICPSR in Michigan and have been stored since

then at the CIBCR. While we cannot state that this approach eliminates all errors, the evidence we have suggests it is sufficiently accurate to make the endeavor worthwhile. In particular, it should eliminate all mistakes in the data set with the exception of any typos introduced by the NBER in creating the original sheets.

In addition to verifying the data series using the procedure discussed above, we have modified the original NBER data set in two ways. First, we eliminated all seasonally adjusted series for which a corresponding seasonally unadjusted series exists in the data set.[2] Second, we eliminated annual series for which an underlying monthly or quarterly series exists in the data series. These two modifications reduced the number of series to be checked from approximately 7,500 to approximately 3,500, with no loss of information.

4 How to Access the NBER Macro-Economic History Database

The revised and verified version of the NBER Macro-Economic History Database is available only via the internet. To access the data set, point your browser to one of

[2] In a few cases, the tape provides a seasonally adjusted series but no unadjusted equivalent. The NBER created such series in cases where it need to aggregate underlying series with different seasonal patterns.

gopher://nber.harvard.edu (Gopher client)

http://nber.harvard.edu (Web Browser)

and select "NBER Macro-Economic History Database" from the menu offered. This is a WAIS searchable database, and you will be prompted for search terms. Enter a word or words that would be part of the description of the series you want, for instance, "concrete production," or "auto sales." Do not enter the quotes, just the words. The search engine does not care about word order and does not insist on a perfect match, although better matches will be listed first in the response. You can use "and" between words to mean both; otherwise, "or" is implied. Parentheses are required for complex searches; for example, "(total and cost) and labor." Numbers are indexed too, but false matches are likely since the data items are indexed also. You can search for sources, units, or anything else that appears in a description. You should not, however, include the terms monthly, quarterly, or annual in your search since the description of every series contains each of these three terms. A catalogue of all series in the original data set can be purchased from the NBER's publication office.

We note that a certain amount of care is required to avoid matching an excessive number of series, given the size and complexity of the data set. A search for "money," for example, will turn up a huge number of matches since the data set

contains many series on “money wages.” To reduce this problem, one can search for “money and stock,” which produces a much smaller number of matches, all of them involving the money stock in some way. Better yet, one can search for “demand and deposits” or “currency,” which will return a very small number of matches.

If a series returned seems interesting, you can use your browser to save the file on your local hard disk. You should choose a file name with eight characters or less with a .DB suffix; this will allow .DB aware packages to recognize the data format.

The data are stored in the Micro-TSP .DB format. This particular format has been chosen because it is pure ASCII and therefore allows documentation to be stored and transmitted with the data. Along with Micro-TSP, both RATS and (real) TSP can read the .DB format. The package DBMS/Copy can translate .DB files to most other formats, and any user with a text editor can remove the header lines and make value code acceptable to any reasonable package. While the .DB format is not economical of storage, this is hardly an issue with economic time series.

The construction of a .DB file is fairly obvious on inspection. Each series is stored, with a description, in a single file. The description comes first, with each line enclosed in quotes. Then comes a line with only a single negative number on

it. This shows the periodicity of the data:

-1 annual,

-4 quarterly,

-12 monthly.

Then come two lines showing the starting and ending dates for the file. Each line is in x.y format, where x shows the year and y the period within the year. So 1945.10 refers to October of 1945. Then come the data, listed with 4 items per line (quarterly, monthly) or just one item per year (annual). The missing value code is 1.e-37. We realize this is a poor choice, since it too easily translates to zero when read by ordinary input conversion routines. It is part of Micro-TSP .DB specification, however, so we could not avoid this misfeature.

To obtain the entire database, download the Unix tar format file

ftp://nber.harvard.edu/pub/macrohist/macrohist.tar.Z

This tar file contains 3500 files, each with a single series, but users must provide their own search software.

5 Disclaimer

We believe the data set we have placed on the Web is in good shape. Nevertheless, mistakes undoubtedly remain. Neither we nor the NBER in any way, shape or

form guarantees the complete accuracy of the data set. Users must apply standard diagnostics to check data reliability. Users who find mistakes should inform us so we can incorporate these revisions.

References

[1] Friedman, Milton and Anna J. Schwartz (1963), *A Monetary History of the United States, 1867-1960*, Princeton, New Jersey: Princeton University Press.

[2] Mankiw, N. Gregory, Jeffrey A. Miron, and David N. Weil (1987), "The Adjustment of Expectations to a Change in Regime: A Study of the Founding of the Federal Reserve," *American Economic Review* 77, 3(June), 358-74.

[3] Miron, Jeffrey A. (1989), "The Founding of the Fed and the Destabilization of the Post-1914 U.S. Economy," in *A European Central Bank? Perspectives on Monetary Unification after Ten Years of the E.M.S.*, Marcello de Cecco and Alberto Giovannini, eds., Cambridge: Cambridge University Press.

[4] Miron, Jeffrey A. and Christina Romer (1990), "A New Index of Industrial Production, 1884-1940," *Journal of Economic History*, 50, 321-37.