Volume Title: Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches

Volume Author/Editor: Ernst R. Berndt and Charles R. Hulten, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-04449-1; 978-0-226-04449-1

Volume URL: http://www.nber.org/books/bern07-1

Conference Date: September 19-20, 2003

Publication Date: October 2007

Title: A Consistent Accounting of U.S. Productivity Growth

Author: Eric J. Bartelsman, J. Joseph Beaulieu

URL: http://www.nber.org/chapters/c0886

15

A Consistent Accounting of U.S. Productivity Growth

Eric J. Bartelsman and J. Joseph Beaulieu

15.1 Introduction

Zvi Griliches thought so much of the difficulty inherent in building data sets suitable for analysis that he devoted a chapter to the problem in the *Handbook of Econometrics* (1986). With respect to available data, he wrote:

> There are at least three interrelated and overlapping causes of our difficulties: (1) the theory (model) is incomplete or incorrect; (2) the units are wrong, either at too high a level of aggregation or with no way of allowing for the heterogeneity of responses; and, (3) the data are inaccurate on their own terms, incorrect relative to what they purport to measure. The average applied study has to struggle with all three possibilities. (1468–69)

The problems are especially acute in the study of productivity, where researchers usually have to "find" their data from different sources. These disparate data sources are produced by different government agencies or private outfits and are designed to answer different questions. As such,

Eric J. Bartelsman is a professor of economics at the Vrije Universiteit Amsterdam, and a research fellow of the Tinbergen Institute. J. Joseph Beaulieu is an economist at Brevan Howard, Inc., and was a staff member of the Board of Governors of the Federal Reserve System when this chapter was written.

The authors would like to thank Carol Corrado for comments at various stages of the project; Jonathan Eller, Suzanne Polatz, Marcin Przybyla, Brian Rowe, Koen Vermelyen, and Matt Wilson for research assistance; the Bureau of Economic Analysis (BEA) for providing data; comments from Barbara Fraumeni, Edward Prescott, and Daniele Coen-Pirani; participants at the NBER Conference on Research in Income and Wealth Summer Institute; the Federal Reserve Bank of Minneapolis; and the winter meetings of the Econometric Society. This paper was prepared while Beaulieu was an economist at the Board of Governors of the Federal Reserve System.

changes in the ratio of real outputs to inputs may reflect inconsistencies in the data set, rather than movements in productivity.

A basic measure of total factor productivity (TFP) requires data on real output, labor input, capital services, real intermediate inputs, and the distribution of income to factors of production. These data often are assembled from different sources and thus may not measure the activities of the same set of producers—even if the descriptions of the producers' activities are the same. Across different variables, the underlying microdata of producing units may have been collected in an inconsistent manner. Data sets may be organized using disparate classifications to describe the activities of producers. Even if the data come from one source, the classification system can vary over time. Sometimes different classification schemes reflect more fundamental differences, but the data still may contain exploitable information.

Another problem is the estimation of missing data. Data are often published at higher levels of aggregation than desired. Sometimes data are available at a fine level of detail in one dimension but only at a very high level of aggregation in another dimension when detailed data are needed in both dimensions. A practical problem occurs when new data releases first provide totals, followed with detailed data at a considerable lag. In order to conduct research with all the desired detail and to make use of the latest data, procedures are needed to best use all the available information.

The purpose of this paper is twofold. First, it describes how our systematic approach to data organization and manipulation helped us in constructing a consistent data set to study productivity. Many of the data hurdles just described had to be overcome in order to line up information on outputs, inputs, and prices from multiple sources in a way that minimized inconsistencies in definitions and coverage of the data. Second, the paper presents some simple applications. It reports estimates of TFP growth by industry and by legal form of organization, and it reconsiders these estimates assuming that firms scrapped an unusual amount of capital addressing potential Y2K bugs.

Productivity in U.S. industries and sectors is the focus of the paper and is prominent in the discussion of the data problems. However, the data issues are ubiquitous, and the systematic approach can be applied to other areas of empirical study; can be scaled down for use with microdata; or can be used to handle additional dimensions, such as regions or countries. At present, the systematic approach is applied in a system using statistical and relational database facilities of the software package SAS.[1] Versions of the system presently are in use at the Board of Governors of the Federal Reserve System and by the European Union (EU) 6th Framework research program EUKLEMS.[2]

1. The SAS Institute is in Cary, NC.
2. See http://www.euklems.org.

There are distinct advantages to the systematic approach to data organization and standardization of manipulation techniques. First, much tedious work in mapping relations between data sources is simplified. Next, documentation of any particular application can refer to standardized data manipulation routines. Further, the data organization scheme simplifies management of complex projects and the sharing of data between users.

For researchers interested in data quality, the systematic approach provides a laboratory: researchers can easily vary the particular assumptions that they used to create estimates in order to test for their sensitivity. A researcher could go farther and produce confidence bands around such estimates. One could also apply this methodology to already published data that are produced by the statistical agencies. Indeed, some data sets that have been made available by government agencies rely on the same techniques available in our system to produce their estimates. As such, a reconsideration of the assumptions that these agencies employ to produce their estimates may be useful.[3] Finally, this approach allows one to consider rigorously counterfactual exercises or to explore the implications of mismeasurement, such as in Jorgenson and Stiroh (2000).

15.2 Constructing a Consistent Data Set

The systematic approach to data organization and manipulation that we have developed provides a practical method to cope with the data problems. Before describing our approach, we give some brief examples of the types of hurdles faced in building a consistent productivity data set.

15.2.1 Data Hurdles

A potentially difficult problem arises when two aggregates that share a common description or title in statistical publications in fact are defined differently. For example, before the 2003 comprehensive benchmark revision to the National Income and Product Accounts (NIPAs), the Bureau of Labor Statistics (BLS) and the BEA had different definitions of *nonfarm business.* The BLS (and since 2003, the BEA) excludes two imputations (owner-occupied housing and the rental value of nonprofits' capital equipment and structures) that the BEA makes to estimate gross domestic product (GDP). Although a careful reading of underlying documentation can trap such differences, only the detailed reconstruction and reaggregation

3. The exception of the recent literature considering mismeasured or biased prices, we are not aware of a lot of papers that directly explore the idea that published data are partially built on assumptions and models where alternatives can be considered. Exceptions include Wilcox (1992) and Miron and Zeldes (1989). There is, however, a developed literature studying the effects of measurement error; see Bell and Wilcox (1993) and references therein. See Weale (1985) for an approach similar to the strategy that could be contemplated with our system.

Table 15.1 **Comparison of 2001 compensation and profits data in the GPO and the NIPA data sets**

	Compensation		Profits with inventory valuation adjustment	
	GPO	NIPA	GPO	NIPA
Manufacturing	939.2	939.2	52.1	83.4
Transportation and utilities	382.1	382.1	23.8	27.7
Wholesale trade	379.8	379.8	48.5	44.8
Retail trade	531.1	531.1	79.3	79.1
Remaining domestic private industries	2586.9	2586.9	320.8	524.4

Note: GPO and NIPA compensation data are collected on an establishment basis. NIPA profits data are collected by firms; the GPO converts these data to an establishment basis.

of the underlying data will allow one to reconcile the differences in outcomes of analysis based on the two output definitions.

A more fundamental problem is related to differences in data collection underlying the aggregates. One well-known example is the firm-establishment problem. United States business data are usually collected at one of two levels: at the establishment level, such as at individual plants, stores, or comparable places of work; or at the firm level (Postner 1984). A problem arises, however, when a firm has multiple establishments that are engaged in different lines of work. General Electric (GE) has extensive operations in manufacturing, finance, and services. Data collected at the establishment level will effectively split GE data among different industries along the different lines of work of the individual establishments. Data collected at the firm level will classify all of GE in one industry based on its major line of work. Currently Compustat assigns GE to the catchall category "miscellaneous," although a few years ago GE was designated an electrical equipment manufacturer.

Researchers manipulating the data need to know how the data were collected. In putting together the Gross Product Originating (GPO) data set, economists in the Industry Division at the BEA have converted all of the data to an establishment-basis concept.[4] The NIPA, on the other hand, also present some industry data, but the definition of industry is not consistent across different types of income. The NIPA compensation data are collected at the establishment level, and as table 15.1 illustrates, the two

4. In this paper we refer to these industry data as the GPO data set. The BEA has published versions of this data set under different names. In 2003, the data set was called the *GDP-by-Industry data.* In 2004, the BEA significantly changed the method by which it calculates these data, yet it continues to refer to them as *GDP-by-Industry data.* In order to clearly identify the data that we are using as the data consistent with the 2003 and prior methods, we use the older name, *GPO data.*

sources match. The NIPA profit data, however, are collected at the firm level from administrative sources, and, therefore, the two databases do not agree on the mix of profits across industries although they do match in the aggregate.

A problem that is particularly annoying to researchers is when classification schemes vary over time. Researchers often need long time series, but changes in classification schemes cause breaks in the series. Usually, industry data before 1987 are based on the 1972 Standard Industrial Classification (SIC) System, while later data are organized on the 1987 SIC. Recently statistical agencies have switched to the North American Industry Classification System (NAICS). Input-output tables use their own classification systems, which also have changed over time. Reclassifying the data so that they are all on one system—a procedure called *concording*—can be difficult when only published data sources are available.

Sometimes, two classification systems may be motivated by entirely different concepts. Nevertheless, incorporating information from both systems may be useful. A good deal of NIPA data is presented by legal form of organization (LFO). While these data cannot be simply linked to data split by industry, we know from the economic censuses that the mix of corporate versus noncorporate businesses varies across industries. Manufacturing, mining, and utilities are predominately corporate, while some service industries, such as membership organizations, personal, and legal services have a large fraction of unincorporated firms. As such, the LFO data contain exploitable information on the mix of an aggregate across industries.

Another way data can be mismatched to the needs of the researcher is when some data are incomplete or missing. One data set may present manufacturing industry data split at the two-digit SIC level, while another may include only durable and nondurable subaggregates. A different example can be found in the NIPA where, at the national level, taxes on production and imports, business transfers, and the surplus of government enterprises less subsidies are presented separately, but for corporations only the sum is listed.

A second example of the problems presented by missing data arises when a researcher has data of aggregates in different dimensions but does not have detailed estimates broken out in each dimension. For instance, the GPO contains information on noncorporate net interest paid by various industries. The NIPA provide national totals for net interest paid by partnerships and proprietorships and by other private businesses. No published data, however, exist on net interest paid split both by industry and by this level of legal form of organization.

A final way in which data can be incomplete is when aggregate data are updated, but updated disaggregated data are not yet available. For example, the BEA publishes initial data on all of the components of gross domestic income (GDI) for a particular year at the end of March of the following year. Typically, it publishes benchmarked data at the end of July, but the industry data are not finalized until well after. One could imagine that it would be possible to develop initial estimates for the recently completed year and incorporate the revised national data to update quickly industry estimates in prior years. Indeed, the BEA has developed a program to produce such "accelerated current-dollar estimates" (see Yuskavage 2002); even so, revised data at the more detailed level are only available with the release of the full data set.

15.2.2 Overview of Data Organization and Manipulation

The system we use consists of four interrelated components that provide practical tools to deal with these problems. First, we store economic data, such as the NIPA, GPO, and input-output data in a relational database, using an appropriate data model. The data model in our application is set up to reflect the conceptual structure of the system of national accounts, with flows between actors tagged by information on their sector, activity (industry), commodity and transaction type, among others.

The second aspect of the data organization is to code information on data definitions and on ways in which the data interrelate, the so-called metadata. In particular, these metadata describe how the detailed data in one dimension, for example, industry, aggregate in a hierarchy, or how detailed industries in two different industry classifications map into each other. The metadata imply linear restrictions across observations that ensure overall consistency of the data set.

Third, the relational database and the metadata make it possible to write standardized routines, or tools, to manipulate the data. These generalized tools fall in one of four categories: aggregating, disaggregating, balancing, and concording data. These four operations help to overcome many of the hurdles that researchers face when using data from different sources.

Finally, the system contains some specialized tools necessary for the study of productivity. These specialized tools allow users to estimate capital stocks, capital services, and TFP employing a variety of assumptions.

15.2.3 Relational Structure of the Productivity Data Set

The main component of the consistent productivity database we have put together is the GPO data set published by the BEA. The GPO data set includes annual data on price deflators, real and nominal measures of gross output, intermediate inputs, and value added. Industries are defined roughly at the two-digit SIC level. The data set also includes nominal income components of value added, such as capital consumption allowances, compensation, and capital income. The data are consistent with the income-side measure of domestic product in the NIPA; the sum across all industries equals gross domestic income.[5]

5. Gross domestic income equals gross domestic output less the statistical discrepancy (see Parker and Seskin 1997).

Table 15.2 **Structure of GPO data set**

Date	Sector	Industry	Imputed	Transaction	Units	Value
1987	Total	Total	Total	Value added	Bil. Ch-96$	6,113.2
1995	Total	Farms	Total	Compensation	Bil. $	15.7
1996	Total	Retail trade	Total	Gross output	Deflator	100.0

To bring the GPO data into our system, we defined a conceptual data model to code the information. In a relational database, a particular piece of datum is not simply the particular numerical value an observation takes, but a set of all the relevant characteristics that identifies the value within the data model. The model we chose for the GPO data set looks like the one in table 15.2.

Five columns, or dimensions, of data characterize the observations in this data set. *Industry* indicates the specific industry (organized on the 1987 SIC); the first observation in the table is for real GDI, and so industry equals the total over all industries. *Transaction* describes where the product or input relates in the chain of production. There are two types of transactions, distributive and productive. Distributive transactions are typically income, or incomelike items such as compensation, profits, capital consumption, and so on. Production transactions relate to goods and services produced or consumed as inputs to the production process, such as gross output, intermediate inputs, labor hours, capital services, investment, consumption, and so on. *Date* is the particular date of the observation. In this case, the date simply records the year because the GPO data set contains annual data. In other cases, one could code values in this dimension to incorporate information on the frequency of the data (monthly, quarterly, etc.) and other timing attributes, such as average over period, end of period, and so on. One could imagine that in some applications, frequencies and timing attributes would be coded in different dimensions when the data set contains a variety of combinations. *Unit* describes how the variable is measured and whether it is a nominal variable, price deflators, or real variable. *Value* reports the numerical value of the data of interest.

Because we augment the GPO data set with other information, we have added two additional dimensions to describe the data. *Sector* represents the NIPA institutional sectors (business, general government, households, and nonprofit institutions). The business sector is further refined by legal form of organization (corporate, noncorporate, etc.).

Imputed accounts for whether the data apply to the two imputed sectors in the NIPA, owner-occupied housing and the rental value of nonprofits' capital equipment and structures, or not. In the NIPA, a large component of consumption is owner-occupied housing. The BEA accounts for the production of this service by assuming that there is an entity that owns the stock of owner-occupied housing and rents it back to its owners. The rental value of owner-occupied housing is treated as consumption. To preserve the identity that GDP equals GDI (up to the statistical discrepancy), the BEA imputes income to this entity. This accounting convention also makes GDP and GDI invariant to whether a house is rented or occupied by its owners. The second imputation involves the rental value of nonprofits' capital equipment and structures. As for owner-occupied housing, the BEA pretends that there is an entity that owns this capital and rents it to nonprofit organizations.[6]

In some cases, the GPO data provide information on sectors. Capital consumption allowances, inventory valuation adjustments, and net interest paid by industry are split between corporate and noncorporate sectors. Other income items are known to accrue to only one type of sector. Profits only go to corporations; proprietors' income and rental income involve only noncorporate businesses, and government surpluses accrue only to government enterprises. Other items, such as compensation, indirect taxes, and gross output have no information on sector. We know totals across industries from the NIPA; we allocate these domestic totals to different industries using the techniques described below. Likewise, we appeal to the NIPA table on imputations to calculate estimates of income for owner-occupied housing and the rental value of nonprofits' capital. These imputations involve one industry (real estate) and therefore require no additional estimation.

A complete accounting of the circular flow of goods, services, and income would include a few other dimensions that identify not only who produces the good or service or who pays the income, but also who purchases the good or service or receives the income. In such a way one could fully integrate all of the NIPA data into the system (for example, tables of personal income and government receipts and expenditures). Such an analysis would be necessary when studying income dynamics or general equilibrium, but these dimensions are not needed for the present study and are excluded.

The presence of various industrial classification schemes presents a small dilemma. One could imagine having separate dimensions to describe each classification scheme: one for SIC 1972, another for SIC 1987, and a third for NAICS. Under this strategy, observations using one system would be concorded to all of the other relevant classification schemes before they were stored in the database. We do not follow this strategy. Usually one is

6. To be exact, these are only nonprofit institutions that primarily serve individuals. Nonprofits that primarily serve businesses, such as trade associations, are treated like any other business in the NIPA in that their consumption of nondurable goods are counted as intermediate usage, and their purchases of equipment and structures are counted as investment. The income paid by these institutions to various factors of production is included in the aggregates for corporations.

not particularly interested in seeing how the different classification systems compare; instead, one just wants to convert all of the data to one particular system. Maintaining new data on an old classification scheme could become burdensome, and the new classification should have some advantages in representing the current structure of the economy. Nonetheless, it would be possible to implement this strategy, and in some cases, such as building a concordance from microdata, it would be the way to go.

15.2.4 Metadata for the Productivity Data Set

The second part of the system involves coding various linear constraints in two types of metadata, hierarchies, and concordances. A hierarchy describes how the data add to their total. Knowing the hierarchy is useful for several reasons. It makes the calculation of interesting subaggregates possible, and it makes matching data sets that differ on their level of aggregation easier. One can keep some subtotals in a database and use the hierarchy to then exclude those subaggregates when calculating other subaggregates or totals. It may be important to carry these subaggregates in the database, especially when they are read directly from a data source. Rounding issues make the subaggregates read directly from the data source more accurate than anything that can be calculated, especially for chain-weighted aggregates.

Finally, and perhaps most important, the hierarchies code the myriad of linear constraints that exist in economic theory as well as various data sets. In our data model, we need hierarchies in four main dimensions: industries, transactions, sectors, and imputed indicators. (See table 15.3.)

Note that these hierarchies apply at any level of another hierarchy. Value added for the whole economy equals the consumption of fixed capital and income. Likewise, value added for corporate farms equals the consumption of fixed capital and income of corporate farms.

While there are four relevant conceptual dimensions in our data model, in practice the various data sets we work with each have their own classifications for each dimension; for example, some data sources use the 1972 SIC, while others use the 1987 SIC to describe the industry hierarchy.

The second type of metadata, a *concordance,* describes how two classification schemes relate. The concordance can be as simple as a list of which components in one system map to the components of a second system and vice versa, or it can provide more detail on the relative magnitudes of how much of a component of one system is split among the components of the other system. What distinguishes a concordance with detailed information on relative magnitudes from simply a detailed data set is that the information on magnitudes in a concordance is typically available for only one year. The concordance tool ensures that these relative magnitudes are applied across years, and the discussion of the concordance tool describes concordances in more detail. In particular, it explains how we constructed

Table 15.3 Hierarchies for four main dimensions

Industries	Transactions
Domestic total	Gross output
Farms	Intermediate inputs
Nonfarm	Value added
Agricultural services, forestry, fishing	Consumption of fixed capital
Mining	Income
Metal mining	Compensation
Coal mining	Taxes on production and imports
Oil and gas extraction	Net operating surplus
Mining services	Current transfers
Construction	Proprietors’ income
Manufacturing	Rental income
Durable manufacturing	Profits
Lumber	Inventory valuation adjustment
Furniture and fixtures	Surplus of government enterprises

Sectors	Imputed
Domestic total	Domestic total
Business	Owner-occupied housing
Corporate	Rental value of nonprofits’ capital
Financial corporations	Not imputed
Nonfinancial corporations	
Noncorporate business	
Sole proprietors and partnerships	
Other private business	
Households and institutions	
Households	
Institutions	
Government	
General government	
Federal	
State and local	
Government enterprises	
Federal	
State and local	

the concordance to convert GPO data from the 1972 SIC for the years 1977 to 1986 to the 1987 SIC.

15.2.5 Standardized Operations

The third component of the system uses the metadata along with the organization of the data in a relational database to automate regular data operations. For example, if a data set contains information in the dimension industry using the 1987 SIC classification, the aggregation routine refers to the 1987 SIC metadata to find out which aggregates need to be created by summing over which detailed industries.

Aggregating

The most straightforward operation is aggregation. Nominal dollar and count data, such as hours and employment, are simply added up a defined hierarchy to calculate aggregates at various levels of detail. Other types of aggregation, such as Laspeyres, Paasche, Fisher Ideal, or Divisia chained indexes involve more complex operations that require additional data on weights. In our particular application, we aggregate over all nonfarm industries and over all business sectors for observations where the imputed dimension has value "not imputed" in order to calculate the aggregate consistent with the definition employed by the BLS in its multifactor productivity program.

Disaggregating

A second operation that is often required is disaggregation, which is the inverse operation of aggregation. Given an aggregate, one can estimate the constituent pieces. For instance, in the GPO data before 1987, industries 36 (electrical machinery) and 38 (instruments) are aggregated together; however, we would like to have historical data for both separately. The difference between aggregation and disaggregation, however, is that the former is a many-to-one operation. No other information besides the constituent pieces, and perhaps corresponding weights in the case of fancier aggregates, is required to calculate the aggregate. On the other hand, disaggregation is usually a one-to-many operation, and, thus, one needs additional information to choose among the infinite possible ways to split a total.[7] We refer to this additional information as a *pattern.* The pattern data need not be consistent with the original data of interest. After all, if the pattern data were to aggregate to the target, one would already have consistent estimates of the pieces.

For simple data that add, the procedure scales up or down the pattern data to yield disaggregated pieces that sum to the known total. Let i index the component of an aggregate T. Denote the observed aggregate d^T, and suppose that there are pattern data on the pieces, p^i. Then the estimate of the disaggregated pieces is given by $d^i = d^T p^i / \Sigma_{j=1}^{I} p^j$. In the case of Fisher-ideal indexes, the procedure does this separately for Paasche and Laspeyres indexes, which do add, and then takes the geometric average of the two.

The quality of the result depends on how well the initial pattern reflects the true distribution of the aggregate. Sometimes, only a few scraps of information may be available that provide rough guidance; in the limit, the fall back could be simply to split the aggregate evenly. Other times, some

7. In cases when the aggregate and all but one component are known, the procedure is exact, and no pattern data are needed. This case arises when one wants to exclude one component from a large aggregate; typically, all of the data on both the aggregate and the piece to be excluded are known.

market conditions or other reasonable assumptions may be used to justify a particular pattern. In building our data set, we augmented the GPO database with an estimate of hours by industries. The NIPA contain estimates of hours paid by industry at a fairly aggregated level. We disaggregated this information down to the industry level in our data set using full-time equivalent employees as an indicator. By using employees as an indicator, we are implicitly assuming that average hours per full-time employee are the same across different detailed industries that are part of the aggregate observed in the NIPA data. We then used these estimates to disaggregate the BLS hours measures, which adjust for the difference between hours worked and hours paid, to get a measure of employee hours worked by industry.

The GPO data set also includes data on all persons engaged in production, which equals the number of employees in an industry plus the number of people working alone. The BLS publishes aggregate estimates of the labor hours of the self-employed and an estimate of self-employed compensation. This last measure represents the fraction of proprietors' income that could be considered labor compensation, as if the proprietor pays a salary to him or herself. The BLS makes this calculation in order to correctly weight the contribution of labor and capital in production function estimates. We make this same adjustment at a more detailed level; we estimate self-employed hours and compensation by industry controlled to the BLS's aggregates using the disaggregation procedure. For self-employed hours, we use an estimate of self-employed workers as a pattern indicator; for self-employed compensation, we use employees' compensation per hour times our estimate of self-employed hours as a pattern indicator.

The automated nature of the tool provides an additional advantage. By varying the pattern data, such as by adding random noise, one can measure how sensitive the results are to the original pattern. Indeed, with a set of statistical assumptions, one could estimate standard errors around these estimates.

Balancing

A third operation, balancing, allows one to estimate data subject to linear constraints in multiple dimensions. An example of a balancing problem shows up when trying to calculate capital services. To do this, one needs investment by type of equipment and by type of industry, while only data on economywide investment by type of equipment and total investment by industry are typically available.

As with disaggregation, there are multiple solutions to the linear constraints; several solutions to the problem of finding one, best set of estimates have been proposed in the literature (Schneider and Zenios 1990). (See table 15.4.) The first is directly applicable when, as in the preceding investment example, there are linear constraints in two dimensions. In this particular example, one can think of the unknowns as a matrix, where the rows repre-

Table 15.4 **Investment flow matrix**

		Asset types			Row controls
		T_1	T_2	…	Totals
Industries	I_1	a_{11}	a_{12}	…	$\sum_{j=1}^{J} a_{1j}$
	I_2	a_{21}	a_{22}	…	$\sum_{j=1}^{J} a_{2j}$
	⋮	⋮	⋮	⋱	
Column controls	Totals	$\sum_{i=1}^{I} a_{i1}$	$\sum_{i=1}^{I} a_{i2}$		$\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}$

sent different values in one dimension, and the columns represent different values in the second dimension. For instance, the rows can represent different industries, while the columns could represent different asset types.

The constraints are represented as restrictions on the sums across the rows and columns.

Suppose one has an initial guess of the matrix, $\mathbf{A}_{k-1}$, which is not consistent with the row and column controls. The first technique, the so-called RAS procedure, estimates $\mathbf{A}$ through the following algorithm. One multiplies $\mathbf{A}_{k-1}$ by R_k so that $R_k\mathbf{A}_{k-1}$ satisfies the column controls. Then one multiplies $R_k\mathbf{A}_{k-1}$ by S_k so that $R_k\mathbf{A}_{k-1}S_k$ satisfies the row controls. Let $\mathbf{A}_k = R_k \mathbf{A}_{k-1}S_k$. Repeating the procedure leads to a series of matrices that, under certain conditions, converges, so that $\mathbf{A} = R\mathbf{A}S$, where $\mathbf{A}$ satisfies both row controls and column controls.[8] The limiting condition, $\mathbf{A} = R\mathbf{A}S$, also explains the moniker "RAS" algorithm that has been attributed to Stone (Stone and Brown 1962). The restriction implied by the procedure that the final matrix is a function of only a series of row and column-scaling factors is also known as the biproportional constraint, and this algorithm is also known as biproportional matrix balancing.

A different strategy is to stack the columns of the matrix into a vector and write $\mathbf{a}_i^0 = a_i + \varepsilon_i$ or $\mathbf{a}_i^0 = \varepsilon_i a_i$ where $\mathbf{a}^0$ is the vector of initial guesses of the true value a and the error term ε_i has a known distribution. Two commonly used distributions are the normal and log normal distributions. The advantage of this approach is that it can handle multiple dimensions and more general restrictions. We further generalize the problem by allowing the constraints also to be measured with error.

8. Bacharach (1965) provides uniqueness and convergence results. Schneider and Zenios (1990) credit Sinkhorn (1964) for an early result that if the entries of $\mathbf{A}$ are strictly positive, then the RAS procedure will converge.

The unknown values are estimated via a maximum likelihood procedure:

$$\min_{a_i, v_k} \sum_{i=1}^{N} \frac{1}{\sigma_i}[\log(a_i) - \log(\mathbf{a}_i^0)]^2 + \sum_{k=1}^{K} \frac{1}{\sigma_k}\left(v_k - \sum_{i=1}^{N} \phi_i^k a_i\right)^2$$

$$\text{or} \min_{a_i, v_k} \sum_{i=1}^{N} \frac{1}{\sigma_i}(a_i - \mathbf{a}_i^0)^2 + \sum_{k=1}^{K} \frac{1}{\sigma_k}\left(v_k - \sum_{i=1}^{N} \phi_i^k a_i\right)^2$$

both subject to the k linear constraints $\sum_{i=1}^{N} \phi_i^k a_i = v_k$.

If $\sigma_k = 0$, the control is measured exactly, and λ_k replaces $1/\sigma_k$ in the minimization problem where λ_k is now an unknown Lagrangian multiplier to be solved for. Stone, Champernowne, and Meade (1942) first proposed a least squares model. In their application, they weighted the observations according to how precise the estimates of the pattern were, but they assumed the controls were measured exactly.

Each method has its own advantages. The advantages of the RAS model is that it is easy to calculate, and under certain circumstances, the biproportional constraint has been given an economic interpretation. In the case of calculating an input-output matrix in year t based on a known input-output matrix in year $t - 1$, Parikh (1979) interprets the two scaling factors, R and S, as follows:

- A *substitution effect* that measures the extent to which the output of a sector substitutes or has been substituted by the output of the product of other sectors as an intermediate input
- A *fabrication effect* that measures the extent to which the ratio of intermediate goods to total gross output has changed in a sector

The benefit from the statistical approach is that it allows one to test a subset of restrictions using either a likelihood ratio test or a Wald test. Weale (1985) uses this insight to test the hypothesis that the U.S. current account was positive in 1982 to 1983 instead of negative, as measured by the BEA.[9] Modeling the distribution of the errors as a normal distribution, perhaps with a standard deviation proportional to the observed values of $\mathbf{a}^0$, also allows the procedure to choose negative values. In cases where several values are known to be zero, a solution to the problem may require a switch in the signs of the initial guess, and in such a case, the RAS procedure will not converge.[10]

9. Golan, Judge, and Robinson (1994) develop a generalized version of the RAS model whereby the probabilities over a discretized space of values are estimated via something like the RAS procedure. These estimates also allow one to conduct statistical tests.

10. The RAS procedure can be adapted to allow for negative values (Günlük-Şenesen and Bates 1988), but the procedure will not switch the signs of the initial guesses.

In creating our data set, we employ the balancing procedure several times. We used it to build a consistent concordance between the 1972 and 1987 SICs (described in the next subsection). We also used the procedure to estimate income components by industry and by legal form of organization for those transactions that the BEA did not already publish such splits.[11] For example, we have information on compensation by industry and total compensation by legal form of organization from the NIPA that serve as our controls. As an initial pattern, we use some unpublished out-of-date information on these splits from the BEA, augmented with observations from the 1987 and 1992 censuses.

Splitting the industry output by legal form is useful because it better matches the sources of at least some of the income components. Many of the income data are collected through tax records, and corporations and other businesses file different forms. The data also have to be adjusted for misreporting; the dollar adjustment to proprietors' income was more than twice as large as to corporate profits in 1996, even though proprietors' income is a much smaller fraction of national income (Parker and Seskin 1997). This suggests that the measurement of output for the noncorporate sector is subject to larger errors than for the corporate sector.

Corrado and Slifman (1999) showed that productivity in the noncorporate business sector was measured to have been declining for over two decades, even though capital income as a share of output was relatively high. They pointed to mismeasured prices as one likely explanation for the confluence of these observations. To the extent that prices are biased upward in industries that have a disproportionate share of noncorporate business, the real output of noncorporate business would be biased down more than for corporate business. Splitting individual industries by legal form—where presumably the output and input prices to the sectors within an industry are similar—and comparing their relative performances may shed some additional light on the issue.

Concording

The last basic tool concords two data sets whose dimensions are organized on different classification schemes. For example, the GPO data from 1949 to 1987 are organized along the 1972 SIC; from 1987 to 2000 they are organized along the 1987 SIC. Some of these industries map to more than one industry.

As suggested by figure 15.1, the problem of concording data organized by the hierarchy on the left to the hierarchy on the right is simply to split the pieces of the left-hand side into parts so that they can be allocated to the different categories on the right-hand side and then added back up.

11. See Seskin and Parker (1998) for definitions of corporations, sole proprietorships and partnerships, and other forms of legal organization as used in the NIPAs.

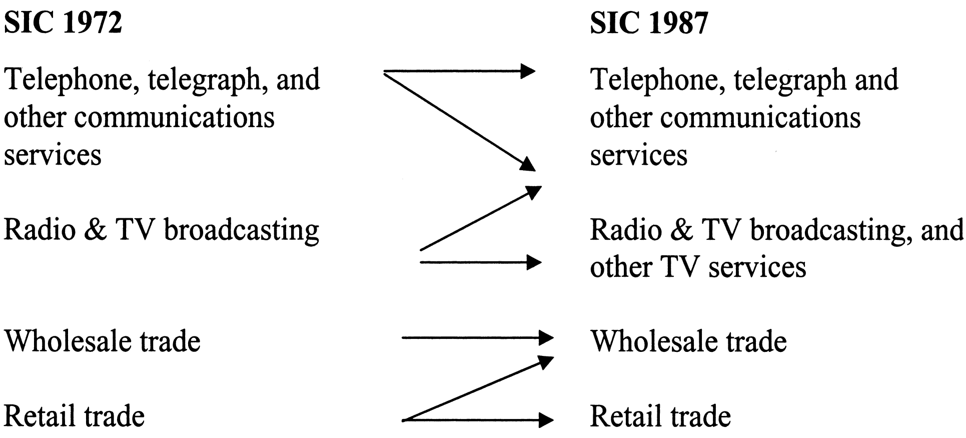

Fig. 15.1 Concordance mapping (example)

Concording the right-hand side to the left-hand side is the mirror image of this operation.

Thus, for the most part, the problem of concording is simply the organized use of aggregating and disaggregating operations. As such, the important part of the implementation is developing weights for the disaggregation. In most cases, information on the relative weights is limited because no data are reported on both bases. As a result, the weights have to be developed using whatever information is available. In concording the input-output tables to the GPO data, a few input-output industries had to be split; to do this, we used a variety of data, such as detailed employment shares and census shipments data (see appendix).

In one important case, data are reported on two bases in a reference year, allowing for a richer concordance: the GPO data for 1987 are available using the 1972 SIC and the 1987 SIC. For example, industries 481,2,9 (telephone, telegraph, and other communications services) and 483–4 (radio and TV broadcasting) on the 1972 basis map to industries 481,2,9 (telephone, telegraph, and other communications services) and 483–4 (radio and TV broadcasting and other TV services) on the 1987 basis. One can think of the problem of developing concordance weights as a balancing problem where the 1972 and 1987 totals are controls. As initial guesses for the pattern for all of the industries, we used the concordance in the NBER Productivity Database (Bartelsman and Gray 1996) for manufacturing, and simply used $1/N$ for other industries for cells that are nonzero according to an available mapping. See table 15.5 for this simple example.

The cells of the matrix are the concordance weights. The advantage of balancing a matrix of weights is that one can concord data both ways in a consistent manner. Concording data from the 1972 SIC to the 1987 SIC and then back again yields the original 1972 data.

Table 15.5 **Gross output of communications, 1987**

			SIC 87	
			481,2,9	483–4
			157.8	42.1
SIC 72	481,2,9	170.1	157.8	12.3
	483–4	29.7	0.0	29.7

Concording provides a means for moving the data between two classification schemes in the same conceptual dimension. Technically analogous is the problem of cross-classification, such as moving data collected at the firm level and published by industry, to match data by industry collected from establishments. The cross-classification table would contain data akin to that in a concordance, showing the amount in a firm-based industry that splits into various establishment-based industries.

15.2.6 Specialized Productivity Tools

We have developed several tools needed specifically to study productivity. One tool accumulates weighted levels of past investment using the so-called perpetual inventory method to estimate stocks of particular assets. The weights are modeled in the same manner as the BLS and Federal Reserve Board (FRB; Mohr and Gilbert 1996) use to account for wear and tear, the average rate of discards, and the effects of obsolescence. A second tool weights these stocks using the standard user cost model of Hall and Jorgenson (1967) in order to estimate capital services. The rate of return can be an ex-ante rate, such as a corporate bond rate, or an ex-post rate, such as property-type income divided by an estimate of the value of the capital stock. A third tool estimates TFP growth by calculating a Divisia index of the inputs using different approaches; the implementation in this paper uses cost shares to weight the inputs.

15.3 Completing the Data Set for the Study of Productivity

15.3.1 Basic Industry Data

Besides the various steps described in the preceding, we had to fill out some of the price data for 1977 to 1986. We concorded the 1982, 1987, and 1992 input-output tables to the GPO data (see appendix) and used the implicit weights in these tables to calculate price deflators for intermediate inputs. Along with available gross product deflators, these gave us gross output deflators. All told, we have information on nominal and real gross output, intermediate inputs, and value added by industry and by legal form of organization. We use these data, along with estimates from the input-output tables to estimate the amount of nominal and real intermediate inputs produced and consumed by the same industry. We exclude the consumption of these inputs from other intermediate inputs; at the same time we exclude them for gross output. The resulting measure is known as sectoral output, and it is the suitable measure for the study of productivity (see Domar 1961; Gollop 1979). In addition, we have information on various components of income paid and employee and nonemployee hours worked by industry and by legal form of organization. To complete the information needed to study productivity, we developed estimates of capital services and labor quality.

15.3.2 Investment and Capital Stocks

The investment series that we use are the detailed industry estimates of industry investment by asset type that the BEA made available on its Web site. We refine these data by splitting industry investment between corporate and noncorporate investment for each type of equipment and structure, controlling the total for each legal form to equal the data available in tables 4.7 of the Standard Fixed Asset Tables and the residential investment tables of the Detailed Fixed Asset Tables. The nonresidential investment tables report investment in equipment and in structures by legal form, divided among three activity groups (farm, manufacturing, and other). To refine these data by industry and by asset type, we used total industry investment by industry and by asset type as an initial pattern in our balancing routine. A practical problem in working with the data was that the investment figures were rounded to integers. In early years, or for activity-type combinations with low levels of investment, dividing nominal values by reals provided a poor estimate of the deflator. To rectify this, we assumed that these asset prices did not vary by activity and used the deflator calculated from aggregate data.

Capital stocks are calculated by accumulating the investment data using the standard BLS stochastic mean-service life and beta-decay parameters. We estimate capital services using the Hall-Jorgenson formula using ex-ante returns, and to analyze trends, we separate capital services into three categories, high-tech equipment and software (ICT), other equipment, and structures.[12]

15.3.3 Labor Services

Analogous to capital, a unit of labor of a certain type may provide a different level of service than a unit of labor of another type. The measure of labor input appropriate for productivity analysis, labor services, is computed as a quality-weighted aggregate of labor hours by type. The weights used to aggregate labor are expenditures shares for each type.

12. *ICT capital* is defined as computers and peripheral equipment, communications equipment, photocopy and related equipment, instruments, and software.

For each industry and sector, information is thus needed on hours worked and compensation for workers by type. These data are not directly available from firm- or establishment-based data sources. However, the Current Population Survey (CPS) March Supplement from the U.S. Bureau of the Census has information on wages of workers, along with other worker characteristics such as age, gender, occupation, education, and industry.

To calculate labor services, we first estimated Mincer's wage equation of the following form:

$$\log[w(s, x)] = \text{const} + \alpha \cdot s + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3 + \beta_4 \cdot x^4 + Z\mathbf{\Gamma} + \varepsilon,$$

where $w(s, x)$ represents wage earnings of someone with s years of schooling and x years of work experience. In the regression we also included gender, part-time/full-time, and ICT occupation dummies, summarized in Z, with coefficient vector $\mathbf{\Gamma}$. The wage equation was estimated using U.S. Census Bureau CPS March survey data for years 1977–2001. We used the fitted values of this equation to impute wages to all workers in the data set. Using estimated wages and hours of individual workers, hours and imputed compensation are computed by industry and by four types of workers. The four worker types that we use are technology workers and three other worker types based on education attained (high school dropout, high school graduate, and college plus).[13]

With these estimates from the CPS, we disaggregated the GPO employee hours and compensation paid to obtain these variables by worker type consistent with the aggregates we observe in our augmented GPO data set. We then aggregated hours of the four worker types by industry using Törnqvist compensation weights to obtain labor services. The labor quality index is defined as labor services divided by hours, and so labor services are defined as labor quality times hours.

15.4 Applications

15.4.1 Productivity Growth of Nonfarm Business

As an initial exercise, we estimated TFP by industry and by legal form of organization, aggregated to private nonfarm business. At the individual industry level, we model the growth rate of TFP as the growth rate of real sectoral output less the share-weighted growth rates of real intermediate in-

13. For the years 1977–1981, *ICT workers* are defined as compute programmers, computer systems analysts, computer specialists, not elsewhere classified (n.e.c.), electrical and electronic engineers, and computer and peripheral equipment operators. For the years 1983–2000, *ICT workers* are defined as electrical and electronic engineers; computer systems analysts and scientists; operations and systems researchers and analysts; supervisors, computer equipment operators; chief communications operators; computer operators; and peripheral equipment operators.

Table 15.6 Growth accounting, nonfarm business

	Capital services				Labor services				
	Materials	Information and communications technologies	Equipment	Structures	Hours	Reallocation effects	Quality	Total factor productivity	Sectoral output
1978–2001	0.16	0.70	0.24	0.35	0.66	0.37	0.17	0.45	3.10
1978–1989	–0.08	0.52	0.22	0.46	0.81	0.54	0.20	0.20	2.88
1990–1995	0.71	0.59	0.15	0.24	0.37	–0.05	0.17	0.54	2.72
1996–2001	0.09	1.17	0.36	0.24	0.63	0.46	0.10	0.88	3.93

puts, labor input, and capital services. We use data from the input-output tables on the ratio of sectoral output to gross output to estimate own-industry inputs. The data on real gross output, intermediate inputs, and cost-weighted expenditure shares come from our modified GPO data set.

To calculate aggregate TFP growth, we take a weighted sum of the individual components, where the weights are calculated as sketched in Domar (1961).[14] We estimate the ratio of sectoral output to gross output in each industry times the ratio of sectoral output to gross output of all private industries excluding farm and owner-occupied housing as measured in the 1982, 1987 and 1992 input-output tables. We interpolate these ratios between years and then multiply them by the ratio of gross output in our data set for each industry to gross output of all private nonfarm industries to obtain annual Domar weights. The contribution of inputs (excluding materials) and TFP to nonfarm business sectoral output growth equals the weighted sum of the contributions to growth of the inputs and TFP to individual industry sectoral output growth, where the weights are the annual Domar weights. The contribution from materials is calculated as the growth rate of sectoral output less the sum of the contributions from the other inputs and TFP. As noted by Domar, the weighted sums of TFP growth rates measures the increase in aggregate output holding the factors of production constant, which is the closest thing to the concept of technical progress that we have.

Table 15.6 reports the growth rate of aggregate sectoral output for private nonfarm businesses over each of the time periods considered, as well as an estimate of the contributions to growth from the use of materials, capital, labor, and TFP. As described in the table, sectoral output grew, on average, 3.1 percent per year. Capital services contributed 1.3 percentage points per years, and labor hours added 2/3 percentage point. We estimate that increases in the quality of labor contributed a little over .5 per-

14. See also Gollop (1987) and Hulten (1978) for a more detailed discussion of the derivation and interpretation of the Domar weights.

centage points to sectoral output growth. The Domar weighted average across industries of labor quality contributed 0.17 percentage points, while the Domar-weighted average of the contribution of hours less the simple sum of nonfarm-business hours times labor share, which we refer to as *reallocation effects,* added 0.37 percentage points. These estimates, including the reallocation effects, are a little higher than is implied by the results in Jorgenson, Ho, and Stiroh (2002) and in Aaronson and Sullivan (2001). We estimate that TFP rose on average 0.45 percent per year. Over the 1996 to 2001 period, sectoral output climbed 3.9 percent per year. TFP accelerated to 0.9 percent per year, and the average contribution of high-tech capital services increased to 1.2 percentage points.

Tables presenting estimates of output growth and contributions of the various input factors and TFP for the sixty industries are available upon request. As noted elsewhere, important contributors to the TFP acceleration in the late 1990s were machinery manufacturing (which includes computers) and electrical machinery manufacturing industries (which includes communication equipment and semiconductors). Technical progress also picked up in the trade industries, as did the growth rate of their stock of high-tech equipment. Some other industries, such as depository institutions and business services, also pushed up their rates of investment in high-tech equipment. But TFP growth increased only 0.3 percentage points in depository institutions and fell sharply in business services.

Table 15.7 reports TFP growth split between corporate and noncorporate private businesses. At the aggregate level, the acceleration noted in table 15.6 in nonfarm business TFP is due to the sharp improvement among noncorporate business. Indeed, among all major components, TFP rose more rapidly among noncorporate businesses than corporations. This could be an artifact of mismeasured capital services. As shown in the bottom half of the table, the contribution to growth from capital services was more rapid among corporations than noncorporate businesses.

15.4.2 Y2K

In the late 1990s, businesses spent a large amount of money working to fix potential Y2K bugs. Software that could not recognize that the year represented by the two-digit number "00" was one year larger than the year "99" had to be modified or replaced. Industry reports indicate that some firms regarded the purchase of whole new systems, including hardware, as preventive maintenance.

These stories suggest that the rate of depreciation and discards of computers and software was unusually high in advance of the century data change. The models that we employ to estimate capital stocks do not directly measure this rate. Unless augmented, these models assume that the rate is a function of the stock and age of equipment of each vintage. As a small experiment with our system, we adjust the stocks of computers and

Table 15.7 Output contribution from capital services and total factor productivity by legal form of organization

	1977–1989		1990–1995		1996–2001	
	Corporate	Noncorporate	Corporate	Noncorporate	Corporate	Noncorporate
			Total factor productivity			
Nonfarm private business	0.30	–0.06	0.78	–0.11	0.64	1.77
Agricultural services	3.11	1.35	–0.92	–0.44	–0.52	0.95
Mining	–1.43	–2.64	0.87	4.14	–0.42	0.73
Manufacturing	0.89	2.24	0.93	–0.41	0.72	2.71
Transportation and utilities	0.18	1.12	0.88	3.52	–0.41	0.83
Wholesale trade	1.52	–0.07	1.19	–1.07	3.97	3.29
Retail trade	–0.91	–0.32	–0.87	–1.60	1.76	2.15
Finance, insurance, and real estate	–2.21	1.65	0.25	0.50	1.06	2.28
Services	0.76	–1.33	0.64	–0.72	–1.99	–0.14
			Capital services			
Nonfarm private business	1.18	1.27	0.99	0.43	1.83	0.85
Agricultural services	–4.15	–0.89	2.02	0.26	2.58	0.36
Mining	1.13	2.00	–0.29	–1.06	0.67	–0.97
Manufacturing	0.38	0.06	0.40	–0.11	0.59	0.12
Transportation and utilities	1.39	1.31	1.07	0.52	2.06	1.23
Wholesale trade	0.79	–2.16	0.68	0.81	0.93	0.98
Retail trade	0.95	–0.29	0.86	0.44	1.72	0.96
Finance, insurance, and real estate	2.24	1.63	1.61	0.53	2.59	0.75
Services	1.01	0.44	0.71	0.16	1.39	0.41

software assuming that some share of Y2K spending represented additional scrappage.

To parameterize the experiment, we used figures reported by the U.S. Department of Commerce (1999). That report cites a study from International Data Corporation (IDC) that public and private spending from 1995 to 2001 to fix the Y2K problem was roughly $114 billion. It also cites an Office of Management and Budget (OMB) report that the federal government was spending a little over $8 billion and a Federal Reserve study that suggests spending by state and local governments was roughly half of federal spending. The Commerce report also provides some figures developed by Edward Yardeni of the distribution of spending across industries. We used the aggregate estimates to calculate baseline spending on Y2K by the private sector over 1995 to 2001, and we used the Yardeni estimates to split them across broad industry aggregates. We assume that Y2K spending across different types of computer equipment and software was the same as total spending, except that we goosed up the fraction on software by 50 percent based on some IDC figures on the split on spending between hardware and software to redress the Y2K bug.

Two considerations suggest these figures are not precise. The IDC indicates that a lower and upper range for spending was plus or minus 50 percent. In addition, all of this Y2K spending does not necessarily reflect additional spending *on investment.* Estimates from the IDC indicate that only 27 percent of worldwide spending was on "hardware or software," whereas the rest was on "internal or external" spending, which may not have been counted as investment. As a lower bound, we assume none of the "internal or external" spending was investment; as an upper bound, we assume all of it was. This leaves a wide range of investment of $14 to $152 billion, which we assume also represents the additional scrappage of older stocks of hardware and software.

Table 15.8 reports the change in estimates of TFP by broad aggregates

Table 15.8 Effect of Y2K spending on total factor productivity growth

	$150 billion								$50 billion
	1995	1996	1997	1998	1999	2000	2001	Cumulative	Cumulative
Nonfarm private business	0.05	0.17	0.26	0.22	0.10	–0.12	–0.13	0.56	0.14
Forestry, fishing, agricultural services	0.00	0.00	0.00	0.00	0.00	–0.00	–0.00	0.01	0.00
Mining and construction	0.02	0.05	0.09	0.08	0.04	–0.03	–0.04	0.21	0.05
Manufacturing	0.02	0.06	0.09	0.08	0.03	–0.04	–0.04	0.21	0.05
Durable goods	0.02	0.08	0.11	0.10	0.04	–0.05	–0.05	0.25	0.06
Electronic equipment and instruments	0.04	0.14	0.17	0.13	0.05	–0.07	–0.08	0.39	0.10
Motor vehicles and equipment	0.01	0.02	0.02	0.02	0.01	–0.01	–0.01	0.06	0.01
Other manufacturing durables	0.02	0.06	0.10	0.09	0.04	–0.04	–0.04	0.21	0.05
Nondurable goods	0.01	0.04	0.06	0.05	0.02	–0.02	–0.03	0.13	0.03
Chemical, petroleum, coal	0.01	0.03	0.04	0.04	0.02	–0.02	–0.02	0.10	0.02
Excluding petrochemicals	0.01	0.04	0.06	0.05	0.02	–0.02	–0.03	0.13	0.03
Transportation and utilities	0.05	0.17	0.25	0.21	0.11	–0.11	–0.11	0.57	0.14
Communications	0.11	0.36	0.52	0.41	0.21	–0.23	–0.21	1.16	0.29
Excluding communications	0.03	0.08	0.13	0.12	0.06	–0.05	–0.06	0.31	0.08
Wholesale and retail trade	0.04	0.12	0.18	0.16	0.08	–0.08	–0.09	0.41	0.10
Finance, insurance, and real estate	0.05	0.14	0.21	0.17	0.07	–0.09	–0.10	0.45	0.11
Depository and nondepository	0.09	0.32	0.44	0.35	0.14	–0.19	–0.21	0.93	0.23
Other finance and insurance	0.05	0.09	0.14	0.12	0.05	–0.06	–0.07	0.31	0.08
Real estate	0.01	0.04	0.07	0.06	0.03	–0.02	–0.03	0.17	0.04
Services	0.06	0.20	0.29	0.24	0.09	–0.15	–0.14	0.59	0.15
Business and other services	0.08	0.25	0.36	0.29	0.09	–0.20	–0.18	0.69	0.17
Recreation and motion pictures	0.03	0.09	0.13	0.10	0.00	–0.06	–0.06	0.24	0.06
Other services	0.05	0.16	0.24	0.20	0.09	–0.10	–0.11	0.52	0.13

when one assumes that the upper bound of Y2K spending ($150 billion) went to replacing high-tech equipment and software that was scrapped and replaced.[15] The largest effect on any aggregate in any year is +0.52 percent in the communications industry in 1997. The extra scrappage reduces the growth rate of capital services. Because real output is not changed, the lower contribution from capital services means that TFP must have been higher, in this case by 0.52 percentage points. In a few industries, such as communications, depository and nondepository institutions, and business and miscellaneous professional services, the effect of Y2K scrappage could be important. For the rest, the effect appears to have been small relative to the average year-to-year variation in TFP. In total, if capital services are adjusted along the lines suggested in the preceding, the rate of growth in TFP would be 16 basis points higher in the second half of the 1990s and 13 points lower in 2000 and 2001. Assuming a more moderate level of Y2K spending that represents replacement investment ($50 billion) reduces the cumulative effect to one-quarter of the upper-bound effect.

15.5 Conclusion

This paper explicates a general approach to the problem of building a consistent data set for the study of economic issues. Coding observations in a relational database allows us to easily manipulate economic data, while the metadata help us to preserve the numerous linear relations across variables. The tools that we have developed take advantage of the standardized data and metadata in order to build a consistent data set.

The system was originally conceived to aid in the study of productivity. To that end, we started with the BEA's GPO data. We concorded the GPO data before 1987, which are organized using the 1972 SIC, to the more recent data, which use the 1987 SIC to classify industries. We then supplemented the data set by including estimates of employee and all persons hours from the NIPA and the BLS, as well as estimating some missing pieces of data, such as gross output for some industries before 1987 and some price deflators. We also concorded the BEA's estimates of investment by industry and by type to the GPO data. To study productivity, we linked data from the input-output tables to calculate Domar weights; we incorporated data from the Current Population Survey, March Supplement 1977–2001, to estimate labor services; and we employed some specialized tools that we developed to estimate capital stocks, capital services, and TFP. Finally, we decomposed all of the data by legal form of organization,

15. For the exercise, the underlying database was first aggregated to the level of detail available for the Y2K spending. For each of these activity groups, TFP was computed using sectoral output and net intermediate use concepts. For higher level aggregates, the TFP was aggregated using appropriate Domar weighting.

controlling the estimates to be consistent with industry totals and aggregate legal-form totals in the NIPA.

Our overall estimates of TFP growth by industry generate the same qualitative results seen elsewhere. Total factor productivity accelerated in the latter part of the 1990s and was particularly high in most industries outside the service sector. The contribution to output growth from increased investment in high-tech capital equipment also increased. We also demonstrated how the system could be employed to reconsider assumptions made in the construction of data and counterfactual exercises. In this small experiment, we took estimates of the amount of spending to remedy the Y2K problem and assumed that some fraction of this estimate was not an increment to the capital stock but instead purely replaced an unusually high amount of capital that was scrapped because it was potentially infected with the Y2K bug. Except for a few industries, the effects on TFP were likely small unless one were to assume that the scrappage associated with the century date change was very large.

A few obvious extensions are possible. Fully incorporating the input-output data, including making them fully consistent with the value added data in the GPO data set, would open up several research avenues. Immediately, it would allow us to have a fully consistent application of Domar weighting. It would allow us to study various price-markup models and to perform various counterfactuals, such as the effects of different productivity growth rates among intermediate producers on prices and aggregate productivity. If, at the same time, separate estimates of input-output tables at the same level of aggregation controlled to the current expenditure-side estimates of GDP were available, we could study the statistical discrepancy. Extending the input-output tables further back and incorporating auxiliary information on prices will enable us to estimate industry price deflators before 1977.

In putting together our preliminary estimates of capital services, we simply used the BEA's estimates of investment by industry and by type that it employs to estimate capital consumption and wealth. However, these estimates are based on limited data of investment by industries outside of census years and are not based on any systematic information on investment by both industry and by type in any year (see, for instance, Bonds and Aylor 1998). Indeed, even though the BEA has made these data available to the public on their Web site, they consider them unpublished because they do not rise to their usual standards for statistical reliability. In the future, we plan to examine how sensitive the capital services estimates are to other plausible distributions of investment. Based in part on conversations with our colleagues, we suspect that the distribution of computer investment could matter importantly, but for other types of equipment, the effects may be small. At the same time, we plan to examine how important the depreciation estimates are for estimates of capital services.

Finally, the system has the tools necessary to start with the most micro-level data sets. Many of the problems of switching classifications and cross-classification would be better approached by working with plant- and firm-level data. For example, a better concordance between the SIC and NAICS could be developed by attaching SIC and NAICS codes to each firm or establishment in a particular year (based on the same logic used to apply the original activity code to a respondent in the survey) and then tabulating a concordance for each relevant variable. Indeed, a joint Federal Reserve–Census project is currently under way to develop such a concordance for manufacturing using the Longitudinal Research Database (Klimek and Bayard 2002). The same method could be used in making a firm-establishment cross-classification by linking enterprise, firm, and establishment codes at the micro level, and then merging and aggregating different data sources to create a cross-classification table.

Appendix

Concording the Input-Output Tables to the GPO Data

A handful of input-output industries had to be split among two or more GPO industries. The following tables describe how the weights for the concordance were calculated in order to allocate the outputs and inputs of these IO commodities and industries among the GPO industries. The 1982 table was mapped to 1972 GPO industries and then concorded to 1987 industries using the same concordance that was used for gross output in the GPO. In calculating price deflators, the reverse was done, and the 1987 table was concorded to the 1972 SIC.

After the concordance, the IO tables were adjusted to account for the new treatment of software in the NIPA. All three published tables (1982, 1987, 1992) treat prepackaged and custom software as intermediate inputs and do not count own-account software as an output. As of the 2000 revision, the BEA began to count software as investment (Parker and Grimm 2000). To adjust the IO tables, we reduced the amount of the use of the commodity "computer and data processing services" by the amount of investment in prepackaged and custom software, and we raised the make of the same commodity by the amount of own-account software investment.[16]

The first columns of tables 15A.1–15A.3 report the IO code, and the

16. We did not adjust manufacturing in 1992 for custom software because Moylan (2001) indicates that the 1992 and 1997 censuses did not collect information on purchases of services by manufacturers, which we take to mean what is now known as custom software investment.

Table 15A.1 **Splitting IO industries to different GPO industries, 1982**

IO code	GPO industry	Indicator	GPO indicator	No.	SIC code	Comment
04.0001	01–2	Dir		.5	0254, 0279pt	No information; split 04.0001 evenly between 01–2 and 07–09
	07–9	Dir		.5	071–2, 5–6, 085, 092	
11.0101	15–17	GP	15–17	.796	15, 16	Ratio of employees of 15 and 17 to 15–17 in 1982
	65re	GP	65re	.122	6552	Ratio of employees in 655 to 65 times 1/2 to split 655 between 6552 and 6551
11.0103	15–17	GO	15–17	1.0	15–17	
	65re	GO	65re	.122	6552	Ratio of employees in 655 to 65 in 1982 times 1/2 to split 655 between 6552 and 6551
11.0602	10	GO	1081	1.0	1081	
	11–12	GO	1112	1.0	1112	
	13	GO	138	1.0	138	
	14	GO	1481	1.0	1481	
11.0603	10	GO	1081	1.0	1081	
	11–12	GO	1112	1.0	1112	
	14	GO	1481	1.0	1481	
14.1801	20	Sh	2051	1.0	2051	
	52–59	GO	542–9	.434	5462	Ratio of employees in 5462 (Bakeries) to employees in all food stores excluding grocery stores in 1982
18.0400	23	Sh	231–8	1.0	231–8	
	39	Dir		.1	3999pt	Shipments of 39996 (Furs dressed and dyed) in 1982 Census
38.0400	28	Sh	2819	1.0	2819	
	33		3334	1.0	3334	

(*continued*)

Table 15A.1 (continued)

IO code	GPO industry	Indicator	GPO indicator	No.	SIC code	Comment
65.0100	40	GP	40	1.0	40	Have to use value added because no gross output data are available for 47
	47	GP	47	.05	474, 4789pt	Assume 4741, 4738, 4785, and 4789 are same size and 4789 split evenly between 65.0100 and 65.0300
65.0300	42	GP	42	1.0	42	Have to use value added because no gross output data available for 47
	47		47	.025	4789pt	Same as with 65.0100
69.0200	52–59	Clc	Mixed		52–7,9 excl. 5462	Calculated as GP(52 – 9) · [1 – GO(548)]/GO(52 – 9)
	73	GP	73	.01	7396	Assumed to be small
	80	Dir		14.5	8042	Revenue of 8042 from 1982 Census
70.0200	61	GP	61	1.0	61	
	67	GP	67	.888	67, excl. 6732	One minus ratio of employees of 673 in 2000 (from occupation by industry data) to employees in 67 times 1/2 to split between 6732 and 6733
77.0302	07–09	GO	07	.140	074	Ratio of employees in 074 to 07 in 1982
	80	Dir			8049, 807–9	Revenue of 8049 and 807–9 from 1982 Census
77.0504	67	GP	67	.1125	6732	Ratio of employees of 673 in 2000 (from occupation by industry data) to employees in 67 times 1/2 to split between 6732 and 6733
	84, 89	GP	84, 89	.073	84, 8922	1/4 of employees in 873 + employees in 84 in 1999 (from occupation data) divided by employees in 84, 87, and 89
	86		86	.083	865, 9	Ratio of employees in political organizations and membership organizations, n.e.c. to all employees in 86 (from occupation data)

Table 15A.2 Splitting IO industries to different GPO industries, 1987

IO code	GPO industry	Indicator	GPO indicator	No.	SIC code	Comment
04.0001	01–2	Dir		.5	0254, 0279pt	No information so split 04.0001 evenly between 01–2 and 07–09
	07–09			.5	071–2, 075–6, 085, 092	
11.0000	15–17	GO	15T7	1.0	15–17	
	65re	GO	653	.149	6552	Ratio of employees in 655 to 653 in 1987 times 1/2 to split 655 between 6552 and 6551
11.0602	10	GO	1081	1.0	1081	
	12	GO	1241	1.0	1241	
	13	GO	138	1.0	138	
	14	GO	1481	1.0	1481	
11.0603	10	GO	1081	1.0	1081	
	12	GO	1241	1.0	1241	
	14	GO	1481	1.0	1481	
14.1801	18	Sh	2051	1.0	2051	
	52–59	GO	542–9	.485	5461	Ratio of employees in 5461 (Bakeries) to employees in all food stores excl. grocery stores in 1987
38.0400	28	Sh	2819	1.0	2819	
	33		3334	1.0	3334	
65.0100	40	G0	40	1.0	40	
	47	G0	474–8	.375	4741, 4789pt	Assume 4741, 4738, 4785, and 4789 are same size and 4789 split evenly between 65.0100 and 65.0300
65.0300	42	GO	42	1.0	42	
	47		474–8	.125	4789pt	Like 65.0100

(*continued*)

Table 15A.2 (continued)

IO code	GPO industry	Indicator	GPO indicator	No.	SIC code	Comment
69.0200	52–59	GO	527–9	1.0	52–7,9	
	52–59	GO	542–9	–.485	excl. 5462	Have to exclude 5462, as calculated above (14.1801)
	73	Dir		.3	7396	Revenue of 7396 from 1987 Census
	80	Dir		3.5	8042	Revenue of 8042 from 1982 Census
70.0200	61	GO	61	1.0	61	
	67	go	67	.888	67 excl. 6732	One minus ratio of employees of 673 in 2000 (from occupation by industry data) to employees in 67 times 1/2 to split between 6732 and 6733
77.0302	07–09	GO	074	1.0	074	
	80	Dir		3.6	8043, 8049	Revenue of 8043 and 8049 in 1987 Census
	80	GO	807–9	1.0	807–9	
77.0504	67	GO	67	.112	6732	Ratio of employees of 673 in 2000 (from occupation by industry data) to employees in 67 times 1/2 to split between 6732 and 6733
	84	GO	84	1.0	84	
	86	GO	865	1.0	865	
	86	GO	869	1.0	869	

Table 15A.3 Splitting IO industries to different GPO industries, 1992

IO code	GPO industry	Indicator	GPO indicator	No.	SIC code	Comment
04.0001	01–02	Dir		.5	0254, 0279pt	No information so split 04.0001 evenly between 01–2 and 07–09
	07–09			.5	071–2, 075–6, 085, 092	
11.0101	15–17	GO	110101	1.0	15, 17	
	63re	Dir		4.6	6552	Half of revenue of 6552 in 1992 Census
11.0108	15–17	GO	110108	1.0	15, 17	
	65re	Dir		4.6	6552	Half of revenue of 6552 in 1992 Census
11.0602	10	GO	1081	1.0	1081	
	12	GO	1241	1.0	1241	
	13	GO	138	1.0	138	
	14	GO	1481	1.0	1481	
11.0603	10	GO	1081	1.0	1081	
	12	GO	1241	1.0	1241	
	14	GO	1481	1.0	1481	
65.0100	40	GO	40	1.0	40	
	47	Dir		1.9	474	Revenue of 474 in 1992 Census
70.0200	61	GO	61	1.0	61	
	67	GO	67	.888	67, excl. 6732	One minus ratio of employees of 673 in 2000 (from occupation by industry data) to employees in 67 times 1/2 to split between 6732 and 6733.
77.0504	67	GO	67	.112	6732	Ratio of employees of 673 in 2000 (from occupation by industry data) to employees in 67 times 1/2 to split between 6732 and 6733.
	84	GO	84	1.0	84	
	86	GO	865	1.0	865	
	86	GO	869	1.0	869	

second columns indicate to which GPO industries these IO codes map. The next three columns show how, in one of two ways, the weights were calculated. Either the weight was written down directly, or it was set as some fraction of a particular indicator. If the weights were entered directly, the column "Indicator" equals "Directly"; the column "No." reports the value of the weight in billions of dollars; and the last column reports the source for the weight. Otherwise, the weight equals the value in "No." times the indicator noted in the columns "Indicator" and "GPO indicator." The values in the "Indicator" column can equal GO (gross output), GP (gross product), or Sh (manufacturing shipments). The column "GPO indicator" reports the particular industry that is used as an indicator. If "No." does not equal one, the "Comment" column describes how the fraction was calculated.

For instance, the 1982 IO industry 11.0101 had to be split in two. The weight used to calculate the fraction that is part of GPO industry 15–17 was set to 0.796 times the gross product of GPO industry 15–17; the weight used to allocate the rest of 11.0101 to GPO industry 65re (real estate excluding owner-occupied housing) was set equal to 0.122 times the gross product of industry 65re.

References

Aaronson, Daniel, and Daniel Sullivan. 2001. Growth in worker quality. *Federal Reserve Bank of Chicago Economic Perspectives* 25 (4): 53–74.

Bacharach, Michael. 1965. Estimating nonnegative matrices from marginal data. *International Economic Review* 6 (3): 294–310.

Bartelsman, Eric J., and Wayne Gray. 1996. The NBER manufacturing productivity database. NBER Technical Working Paper no. T0205. Cambridge, MA: National Bureau of Economic Research, October.

Bell, William R., and David W. Wilcox. 1993. The effect of sampling error on the time series behavior of consumption. *Journal of Econometrics* 55 (1–2): 235–65.

Bonds, Belinda, and Tim Aylor. 1998. Investments in new structures and equipment in 1997 by using industries. *Survey of Current Business* 78 (12): 26–51.

Corrado, Carol, and Lawrence Slifman. 1999. Decomposition of productivity and unit costs. *American Economic Review: Papers and Proceedings* 89 (2): 328–32.

Domar, Evsey D. 1961. On the measurement of technological change. *Economic Journal* 71 (284): 709–29.

Fraumeni, Barbara M. 1997. The measurement of depreciation in the U.S. national income and product accounts. *Survey of Current Business* 77 (7): 7–23.

Golan, Amos, George Judge, and Sherman Robinson. 1994. Recovering information from incomplete or partial multisectoral economic data. *Review of Economics and Statistics* 76 (3): 541–49.

Gollop, Frank M. 1987. Modeling aggregate productivity growth: The importance of intersectoral transfer prices and international trade. *Review of Income and Wealth* 33 (2): 211–27.

———. 1979. Accounting for intermediate input: The link between sectoral and aggregate measures of productivity. In *Measurement and interpretation of productivity,* Washington, DC: National Academy of Sciences.

———. 1979. Accounting for intermediate input: The link between sectoral and aggregate measures of productivity. In *Measurement and interpretation of productivity,* ed. Albert Rees and John Kendrick, 318–33. Washington, DC: National Academy of Sciences.

Griliches, Zvi. 1986. Economic data issues. In *Handbook of econometrics.* Vol. 3, ed. Zvi Griliches and Michael D. Intriligator, 1465–1514. Oxford, UK: North-Holland.

Günlük-Şenesen, Gulay, and John M. Bates. 1988. Some experiments with methods of adjusting unbalanced data matrices. *Journal of Royal Statistical Society* A 151 (3): 473–90.

Hall, Robert E., and Dale W. Jorgenson. 1967. Tax policy and investment behavior. *American Economic Review* 57 (3): 391–414.

Hulten, Charles R. 1978. Growth accounting with intermediate inputs. *Review of Economic Studies* 45 (3): 511–18.

Jorgenson, Dale W., Mun S. Ho, and Kevin J. Stiroh. 2002. Information technology, education, and the sources of economic growth across U.S. industries. Harvard University. Mimeograph.

Jorgenson, Dale W., and Kevin J. Stiroh. 2000. Raising the speed limit: U.S. economic growth in the information age. *Brookings Papers on Economic Activity,* Issue no. 1:125–211. Washington, DC: Brookings Institution.

Klimek, Shawn D., and Kimberly N. Bayard. 2002. Classifying the Census of Manufactures from the standard industry classification system, 1963 to 1992. U.S. Census Bureau, Center for Economic Studies. Mimeograph.

Miron, Jeffrey A., and Stephen P. Zeldes. 1989. Production, sales, and the change in inventories: An identity that doesn't add up. *Journal of Monetary Economics* 24 (1): 31–51.

Mohr, Michael F., and Charles E. Gilbert. 1996. Capital stock estimates for manufacturing industries: Methods and data. Board of Governors of the Federal Reserve System, March. http://www.federalreserve.gov/releases/G17/capital_stock_doc-latest.pdf.

Moylan, Carol. 2001. Estimation of software in the U.S. national accounts: New developments. OECD Statistics Directorate Working Paper no. STD/NA(2001) 25. Paris: Organization for Economic Cooperation and Development, September.

Parikh, Ashok. 1979. Forecasts of input-output matrices using the R.A.S. method. *Review of Economics and Statistics* 61 (3): 477–81.

Parker, Robert, and Bruce Grimm. 2000. Recognition of business and government expenditures for software as investment: Methodology and quantitative impacts, 1959–98. Bureau of Economic Analysis. Unpublished Manuscript. http://www.bea.gov/bea/papers/software.pdf.

Parker, Robert P., and Eugene P. Seskin. 1997. The statistical discrepancy. *Survey of Current Business* 77 (8): 19.

Postner, Harry H. 1984. New developments towards resolving the company-establishment problem. *Review of Income and Wealth* 30 (4): 429–59.

Schneider, Michael H. and Stavros A. Zenios. 1990. A comparative study of algorithms for matrix balancing. *Operations Research* 38:439–55.

Seskin, Eugene P., and Robert P. Parker. 1998. A guide to the NIPAs. *Survey of Current Business* 78 (3): 26–68.

Sinkhorn, Richard. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics* 35:876–79.

Stone, Richard and Alan Brown. 1962. *A compatible model of economic growth.* London: Chapman and Hall.

Stone, Richard, David G. Champernowne, and James E. Meade. 1942. The precision of national income estimates. *The Review of Economic Studies* 9 (2): 111–25.

Weale, Martin. 1985. Testing linear hypothesis on national account data. *Review of Economics and Statistics* 67 (4): 685–89.

Wilcox, David W. 1992. The construction of U.S. consumption data: Some facts and their implications for empirical work. *American Economic Review* 82 (4): 922–41.

U.S. Department of Commerce, Economics and Statistics Administration. 1999. *The economics of Y2K and the impact on the United States.* November 17. Washington, DC: U.S. Department of Commerce.

Yuskavage, Robert E. 2002. Gross domestic product by industry: A progress report on accelerated estimates. *Survey of Current Business* 82 (6): 19–27.