1 Introduction

In September 1968, The Center for Comparative Political Research, State University of New York (Binghamton), undertook the assembly of a longitudinal aggregate data archive, with computer-facilitated procedures for the storage and retrieval of its content. The present volume is a preliminary presentation of a substantial portion of the CCPR file.

The CCPR archive was, in part, the outgrowth of an effort initiated in Washington, D.C., in the spring of 1967 to codify the aggregate data resources of *The Statesman's Yearbook,* an annual with a history of continuous publication since 1864, which had never been systematically mined for quantitative materials of potential utility for comparative social scientists. Many of the data extracted from this source proved to be of questionable reliability (particularly for the earlier years), and a large number of additional sources were ultimately consulted. The *Yearbook* nonetheless remained a principal source for the period prior to World War I, since other sources, such as the *Almanach de Gotha, Whitaker's Almanac,* and *The World Almanac,* proved to be even less reliable, more limited in range of subject matter, or both.

In establishing the CCPR archive, it was decided to assemble materials, insofar as possible, for the time-period 1815 (immediately after the Congress of Vienna and establishment of the modern international system) to 1966, excluding the two major wartime periods, 1914–1918 and 1940–1945. It was also decided that all commonly recognized members of the international community would be included, excluding a handful of quasi-states such as Andorra, Bhutan, Liechtenstein, Monaco, and Vatican City. Taiwan is also excluded, since its government structure reflects, in part, the claim of its leadership to represent Mainland China. A complete list of the political units regarded, for present purposes, as "nation-states," together with their dates of independence, is given in Appendix 1. It need hardly be observed that establishing precise "dates of independence" for certain nations is virtually impossible. Dates for nations such as Canada, Egypt, Italy (as a unified state), Korea, Sa'udi Arabia, and South Africa (to mention but a few) must always be to some extent arbitrary. In most such cases we have indicated in footnotes the reasons for the dates selected.

The CCPR archive currently contains data on some 200 variables, of which only 102 are included in the present volume. Reasons for the omissions are various. Some variables (such as area in square kilometers as well as in square miles) are essentially redundant. Others (such as composition of the work force) were not completely assembled at the time the present volume went to press, while still others (such as area and population of empire) involve relatively few cases. In addition, the temporal coverage for certain variables has been foreshortened when the substantive entries for earlier years encompass relatively few cases or are too scattered, longitudinally, for meaningful array estimation.

The variables themselves are listed, by segment, in the table of contents, and are discussed in the next chapter.

Assembly of the Archive

The data contained in this volume and in the CCPR archive may be categorized in a variety of ways. First, as regards scalor type, the overwhelming proportion of the data are *interval*-scaled, that is to say, are expressed in true numeric units, be they dollars, miles, or what have you. The only *ordinal*-scaled data (ranked on a "more" or "less" basis without the implication of true numeric units) are contained in Segments 1 and 10 of this volume, while only four variables (located in Fields D, G, H, and I of Segment 1) are *nominal*-scaled (ranked by qualitative category rather than on a "more-less" basis).

Second, a distinction may be made between *primary* and *secondary* (derived) data. The latter are derived by means of mathematical manipulation of the primary data (such as the calculation of population density figures from area and population figures). Approximately one-third of the variables included in the present volume are of this character.

Third, most of the interval-scaled longitudinal arrays contain both *original* and *estimated* data. Most of the original data are taken directly from an external source, such as *The Statesman's Yearbook.* The estimated data, on the other hand, are one of two types, depending on whether they were supplied by the compiler or are computer-generated. In the present volume, *each data field that contains both original and computer-estimated entries contains asterisks which are appended to the original data items.* All nonasterisked items in these fields are linear estimates obtained by "averaging" between original data points. For example, if, for a given variable, original data are available for the years 1936 and 1939 with values of 12 and 27, respectively, the calculated annual increment is 5 [(27-12)/3], and the estimated values for 1937 and 1938 are 17 and 22, respectively. If the arrays in question are essentially incremental in character and the original-source entries are relatively frequent and evenly distributed, it is felt that this quite simple method of missing data estimation yields results that, in most cases, are reliable within tolerable margins of error. Should the user disagree, he is, of course, free to disregard the estimates entirely. It goes without saying that more sophisticated estimation procedures, such as least-squares or polynomial regression techniques, are possible, and a number of such procedures were evaluated for possible use in place of the method described above. In most instances, however, the results were not significantly more impressive than those obtained by the above technique, or consumed an inordinate amount of computer time, or both. In addition, the use of polynomial regression for estimation *beyond* given data points has a tendency to yield rather unpredictable results, depending on the precise character of the curve to be fitted.

The overwhelming proportion of the estimated data are of the computer-generated type discussed above. In a limited number of cases, however, estimates have been supplied by the compiler, usually on the basis of indirect evidence contained in the literature or to remedy obvious discrepancies in reported figures due to typographical or

other error. Estimates of this character are identified by an "E" immediately following each datum entry.

One additional "flag" symbol is also employed. For a limited number of newly independent African nations, the only figures that could be obtained regarding free or black market currency exchange rates (Segment 9, Field H) were figures *averaged* over a number of years (usually seven or less). In all such cases, the data are identified by an "A" immediately following each entry.

Scaling Factors

Scaling factors, if any, are given in parentheses following each variable designation in the segment listings. For example, "Area in Square Miles (1000)" means the data are given in thousands, while "Population Density (.1)" means that multiplication by .1 is required for conversion to actual density figures.

Country Labels and Ordering of Cases

In the CCPR archive, each country (case) is, of course, provided with an I.D. I.D. numbers have been omitted in the main body of this volume but are included in the listing of states (Appendix 1) for the benefit of those who may wish to employ the data contained in the tape version.

In general, the ordering of cases is alphabetic, with the following exceptions: Austria and Hungary appear immediately after Austria-Hungary; the German states (Baden, Bavaria, etc.) that became united under Bismarck in the 1860s are listed after Germany, followed by the contemporary Democratic and Federal Republics of Germany; the component states of Italy (Modena, Parma, etc.) appear after Italy; the People's Republic and the Republic of Korea appear after Korea; Norway and Sweden appear after Sweden-Norway; Zanzibar appears after Tanzania; Ireland appears after the United Kingdom; and Montenegro and Serbia appear after Yugoslavia.

In addition, certain countries have, over time, undergone changes of name. In all such cases, the *ordering* exhibited in Appendix 1 is retained, but the more appropriate labels are used, as follows:[1]

Austrian Empire for Austria-Hungary, 1815–1866
Republic of China for China, 1919–1948
People's Republic of China for China, 1949–1966
Santo Domingo for Dominican Republic, 1844–1913
Abyssinia for Ethiopia, 1898–1936
Persia for Iran, 1815–1913
Siam for Thailand, 1815–1913
Ottoman Empire for Turkey, 1815–1913
Russia for USSR, 1815–1913
Egypt for United Arab Republic, 1951–1957

Missing Records

The occurrence of one or more missing records within a given data set is indicated by a blank line. There are three possible reasons for such omissions. First, as stated earlier, no data were assembled for the two wartime periods, 1914–1918 and 1940–1945. Second, no data were

[1] Relatively minor alternative labels, such as Salvador or San Salvador for El Salvador, are omitted. It should also be noted that certain dates selected for label change are necessarily somewhat arbitrary, such as those for Abyssinia/Ethiopia and Siam/Thailand.

assembled for countries losing their independence, however temporarily, such as Santo Domingo, which reverted to the status of a Spanish territory from 1861–1864, Syria, which was part of the United Arab Republic from 1958–1960, Austria, which was annexed by Germany in 1938, and so forth. Third, all record-segments for which no substantive data at all could be obtained are, for reasons of economy, omitted.[2]

Sources

Reference is made in Chapter 2 to some of the principal sources consulted in assembly of the CCPR archive. A list of the major *serial* publications consulted is contained in Appendix 2. The relative paucity of such citations is due not to the fact that these were the only sources consulted, but rather to the enormous variety of items, from five different libraries, examined in the course of a literature search extending over three years, and which still continues. Several hundred volumes, for example, were consulted in assembling the political data contained in Segment 1—a corpus of data that is still far from complete. Given both the horizontal and longitudinal dimensions of the present undertaking, a complete item-by-item citation of sources would require another volume of near-equal magnitude. Furthermore, in the assembly of certain data arrays (area and population for much of the nineteenth century, for example) a number of series were initially prepared from separate sources, components of each of which were utilized, to some extent, in construction of the final arrays.

The Problem of Reliability[3]

Given the volume of data processed, considerable variation in the calibre of the sources consulted, and the ever-present possibility of both human and machine error, some consideration of the overall reliability of the materials presented in this volume can scarcely be avoided.

Absolute reliability is, of course, an impossibility. However impeccable the source, one would be foolish indeed to insist that the precise number (in thousands) of items mailed in the United States in 1966 was 7,392,299 or that there were exactly 9 miles of railroad line in Brazil in 1854. A margin of error in the vicinity of 5% must probably be allowed for even the most rigorous of aggregate statistics, while it would appear that margins of from 10–20% will not completely destroy their utility for most comparative purposes, assuming the error to be random, rather than system-

[2] Loss of independence is indicated in Appendix 1. All other omissions (save for the wartime periods) may be attributed to missing data.

[3] Collateral problems of redundancy, comparability, and validity are frequently (and justifiably) raised in connection with aggregate data such as are contained in this volume. With the possible exception of comparability however, these matters are best dealt with in the context of utilization, rather than of presentation of a data file, and will be discussed in a forthcoming study by the present writer based on the materials contained herein.

An excellent example of the problem of comparability is provided by Guy Hunter, who suggests that figures for income per capita for many developing countries may have to be multiplied by as much as a factor of 3 in order to achieve true comparability with those of more developed countries. See Hunter's *Modernization in Peasant Societies* (New York and London: Oxford University Press, 1969), p. 260.

atic.[4] It is our conviction that the latter range is seldom exceeded with regard to interval data appearing in the present volume, with the possible exception of certain early and mid-nineteenth century entries,[5] of some of the financial and trade data for the interwar period, and of the conflict data contained in Segment 10. To make such an assertion is, of course, an invitation to our critics to prove us wrong. But efforts of precisely this sort would be most welcome. Since the compilers regard the present volume as a work in progress, rather than a finished product, they would be grateful for comments from users which might serve to strengthen the utility and reliability of its content.

As an example of one type of discrepancy that the informed user might well unearth, let us consider the case of railroad mileage figures for Pakistan for 1947 to 1966 (Segment 3, Field F). At the time our production tape was generated we had what were assumed to be reasonably reliable figures for only three years: 1951, 1955, and 1964. The remaining seventeen items in the array were estimated. Subsequently, we obtained official Pakistani figures for all twenty years.[6] The two sets of figures, together with error percentages are as follows:

1947	6321	6931	−8.80%
1948	6411	6935	−7.55%
1949	6502	6938	−6.28%
1950	6593	6938	−4.97%
1951	6684*	6998	−4.48%
1952	6774	6998	−3.20%
1953	6865	6998	−1.90%
1954	6956	7040	−1.19%
1955	7047*	7043	+0.05%
1956	7046	7042	+0.05%
1957	7045	7047	−0.02%
1958	7044	7045	−0.01%
1959	7043	7042	+0.01%
1960	7042	7039	+0.04%
1961	7041	7039	+0.02%
1962	7040	7039	+0.01%
1963	7039	7039	0.00%
1964	7039*	7046	−0.09%
1965	7038	7047	−0.12%
1966	7037	7047	−0.14%

It will be noted that the *range* of error is between −8.80% and +0.05%, the average being 1.94, or just under 2%. Needless to say, however, the official figures will

[4] Margins of error attributed to most of the data series appearing in Bruce M. Russett et al., *World Handbook of Political and Social Indicators* (New Haven: Yale University Press, 1964) range generally from 5-20%, with a scattering as high as 50%, and at least one (p. 188) as high as 100%.

[5] In certain cases prior to 1913 telegraph and postal data will prove to be somewhat inflated due to the fact that many of our original sources did not adequately distinguish between domestic and international items. In both cases, however, "in transit" items (telegrams or mail involving neither domestic senders or recipients) have been excluded.

[6] Government of Pakistan, Central Statistical Office, *20 Years of Pakistan in Statistics* (Karachi, 1968), p. 148.

be much more extensively utilized in any future edition of the present work.

Availability of the Data

In due course, a copy of the tape used in the production of this volume will be forwarded to the Inter-University Consortium for Political Research, Ann Arbor, Michigan, for dissemination to potential users. Personnel of the Center for Comparative Political Research, SUNY-Binghamton, would be quite happy to engage in correspondence regarding the file, but would prefer *not* becoming involved in the distribution of its contents.