Correlates of War Project

National Material Capabilities Data Documentation

Version 3.0

Last update: May 2005

** Many individuals have contributed to the collection of national capabilities data and this documentation/coding manual over many years. Particular contributors to the v3.0 coding manual include Reşat Bayer, Diane Dutka, Faten Ghosn, and Christopher Housenick. Important contributors to the 1990 version of the manual included Paul Williamson, C. Bradley, Dan Jones, and M. Coyne.

Table of Contents

Introduction

"Power" - here defined as the ability of a nation to exercise and resist influence - is a function of many factors, among them the nation's material capabilities. Power and material capabilities are not identical; but given their association it is essential that we try to define the latter in operational terms so as to understand the former.

This manual examines some of the more crucial issues in the selection and construction of indicators of these capabilities, discusses the implications of the various options, and indicates the decisions made by the Correlates of War project for the Composite Index of National Capacity. It presents with detail the terminology and definitions of each indicator, data collection techniques, problems and irregularities, and data manipulation procedures. Additionally, it functions as a guideline for reading the data set and provides a bibliography. Not all of the decisions undertaken were optimal, and often the trade-offs are difficult. Nor did the enterprise start from scratch. Historians, social and physical scientists, military analysts, and operations researchers have examined the ideas of power base, national strength, and material capabilities. As the bibliography makes clear, about two dozen authors have tried to develop - and generate data for - indicators of national attributes. We profited greatly from these prior efforts, be they speculative, empirical, or both. This literature has been of great assistance, especially in illuminating the difficulties and highlighting those myriad strategies we have avoided.

General Considerations

There are certain general considerations we must note before turning to the specific dimensions in any detail. First and foremost is that of comparability across a long time period (1816 to the present) of a staggering variety of territorial states, peoples, cultures, and institutions at radically different stages of their economic, social, and political development at any given moment. An indicator that might validly compare a group of European states in 1960 may very well be useless in comparing one of them to a North African state in the same year, let alone 1930 or 1870. We selected our indicators from among those that were both meaningful at any given time and that had roughly the same meaning across this broad time range. This requirement limited our choices, even in the statistically better endowed post-World War I years.

Various caveats must be made concerning the validity of the indicators the project selected. The first of these is comparison, which relies on the sometimes questionable assumption that equal values of the same indicator make equal contributions to capability. To differentially weight the contributions of individual nations entails questions that the project was not ready to address. Certain indices where this caution especially applies are noted later.

A second caveat concerns the choice of coding rules given several equally plausible alternatives. Here, the purpose is that the value assigned to the underlying concept not be highly sensitive to this choice. In some cases, we estimated this sensitivity by recollecting data for a sample subset, applying alternative choices, and determining their distribution of data values around those previously gathered.

A third caveat is information sources. We consulted several sources. We were particularly interested in series having long runs of data from multiple sources overlapping the same time period because this allowed better discrimination of reliable figures. Given different volumes of the same series, we used the most recent data reported, although alert to the possibility that revisions reflected manipulation by the reporting nation or changes in the methods of reporting, rather than improvements in accuracy.

A fourth caveat is the role of estimation. It is not surprising that we could not find all the requisite information. We did not expend considerable time and effort to produce a series complete save for some small remaining bit of ignorance. Rather, we filled in the gaps through interpolation, where it was reasonable to assume that the difference in values of the endpoints of the gap were accurate and that the change rate between them was uniform. We discuss this further under particular sections. In the case of missing data or lack of comparability among sources, we often resorted to bivariate regression of the known values on time, using the latter to estimate all the data in the series. A contrast between the two methods is that estimates obtained by interpolation are assumed correct even if they depart from the long-run trend. Estimates obtained by regression assume that the true change rate is constant over a longer sequence of several known data points, of which the endpoints and all other reported values may be in error. The approach we used depended on the context of all that was known about each individual case.

A fifth caveat data availability and the inevitability of error. Most of the indicators used in the Correlates of War project are generated by the application of operational criteria and coding rules to the often ambiguous "traces" of history. In some cases we can be quite confident about the reliability of this approach because we ourselves developed the data. In other cases, we rely on apparently precise numerical traces recorded by others at earlier times with coding and scaling criteria ranging from unknown to inconsistent. For instance, given that our definitions of national territories sometimes differ from source definitions, and given the imprecision of the latter, the figures we obtained may have reflected these incorrect boundaries. Likewise, errors could have been introduced through efforts to correct for boundary changes.

Error could also arise from inappropriate uses of estimation. The assumption (in the case of interpolation) of accurate endpoints or (in the case of regression) that transient residuals in documented values do not represent historically real fluctuations may be wrong. In either case, the assumption of constant change rates may have been mistaken. While we sought to leave no

stone unturned, the reporting of national statistics is a recent practice. As one moves further back toward 1816, statistical availability and quality deteriorates. Given the paucity of documentation, figures and estimates of inferior reliability often were the only kind available. In those cases, and despite the possibility of error, we had no choice but to identify, select, and combine numerical estimates of evidence, hoping that we have recognized and taken account of differing criteria.

Given the multiplicity of interpretations as well as the difficulty of validation, we expect alternative national capability indicators to be put forth with some regularity well into the future. This leads us to a brief consideration of the dimensions and indicators of capability we adopted and why. We intended to tap the scholarly consensus on the major components of general capabilities, and not the development of the most comprehensive predictor of success in diplomacy, crisis, or war. The extent to which these capabilities do account for such success is an empirical question and there is mounting evidence that the two differ in important ways.

Basic Dimensions

The project selected demographic, industrial, and military indicators as the most effective measures of a nation's material capabilities. These three indicators reflect the breadth and depth of the resources that a nation could bring to bear in instances of militarized disputes.

Why have we treated only demographic, industrial, and military indicators of national capabilities? Why have not geography of location, or terrain, or natural resources (all of which clearly affect material capabilities) been addressed? Location, for example, could be important in several senses: island and peninsular states are often more able to trade with a larger number of others, are somewhat more defensible against invasion, emphasize sea power over land power (thus appearing less able to threaten another with invasion), and have fewer close neighbors with whom to quarrel. Landlocked states are typically more restricted in their choice of trading partners, are more vulnerable to invasion, occupation, or annexation, have more immediate neighbors, and "require" greater land forces that often appear threatening. All these facets could detract from or enhance a state's capabilities. However, they are too dyad-specific to permit valid cross-national comparison because they pertain to the *relationship between* nations rather than to the *characteristics* of a given nation. As to natural resources such as arable land, climate, and resource availability, these factors are already reflected to a considerable extent in the indicators we employed.

There is, of course, the question of effective political institutions, citizen competence, regime legitimacy, and the professional competence of the national security elites. While these are far from negligible, they contribute to national power and the efficiency with which the basic material capabilities are utilized, but they are not a component of such capabilities.

A final and major point is that while most researchers grant that the demographic, industrial, and military dimensions are three of the most central components of material strength,

nevertheless they may quarrel either with (1) the specific subcomponents or (2) the decision to stay with them over nearly two centuries. These issues are dealt with later in their specific contexts. The value of uniform series throughout the period is a question that must be subject to further inquiry, and by empirical means based on datasets such as this one.

Next we address the procedures and problems of the individual indicators. Where there are important departures from core procedures, we note them in this document and in the data set itself. For each of the three indicators, we begin with an introductory section and follow it, for each of the two subdimensions on which the indicators rest, with discussions of data acquisition and generation, and data problems and potential errors.

Overview of Version 3.0

Version 3.0 of the National Material Capabilities data set is the result of several years of effort undertaken at the Pennsylvania State University by the COW2 Project. Two major updates have taken place. First, additional detail about the source for and quality of data points was added to some component sets. We hope to continue this practice in the future. Second, each component series was extended and each series was examined and in some cases was revised. A brief overview of these changes is outlined below, starting with the universal updates and moving then to individual component updates. Once those two discussions are complete, this manual then goes into greater detail about each of the six indicators of national capabilities.

Sub-Component Data

Along with overall data, the COW 2 project is releasing additional information about each separate sub-component. Each sub-component has its own separate data set (saved in Microsoft Access format) which contains new detail about the particular variable. Information in these sub-data sets includes in particular source data identification, Quality Codes, and Anomaly Codes, along with the values for the variables in each state-year. The final values for each state year are then placed in the final overall 6-component data set typically used by analysts.

Discontinuities and Source/Quality Codes

It is important to document the source of and confidence we have in our data points. Therefore, coding schemes for source and quality codes have been developed during the collection of v3.0, and included as was possible and practical during the update. For instance, the sub-component data sets include the source of the value in the data series. In many cases, we were unable to track data value to a particular source. In such cases, we have left original values, which did come from specific sources, but which we simply do not know.

In any data set, there are data points that must be interpolated, extrapolated, or estimated. Previously, COW data sets have not listed which data points are interpolated and

which come from solid data sources.[1] In this version of the national capabilities data set, we made these estimations transparent to users when possible by creating a quality code variable as a separate column in four of the national capability indicators. These 4 indicators are iron and steel production, primary energy consumption, total population, and urban population. It is important to note that each component has its own quality coding scheme. Because of very different coding rules and potential fluctuations, each component needed its own coding approach. For instance, total population changes very slowly, and a census every ten years is the norm. Basic growth can easily be calculated for each country, and anything that can radically alter a state's population will most often be well documented. Examining a concept like primary energy consumption, however, it is quite possible for there to be quite rapid fluctuations in energy usage. Oil embargoes, new technologies, and wars can make energy consumption values fluctuate greatly. Therefore, this commodity has a higher standard for its data point quality, and that higher standard is reflected in its quality codes.

Ideally, these data quality codes would be a temporary element of this data set. The long-term goal of this project should be to eventually find specific data for each data point that falls short of the standard for receiving an "A" (the universal designation for a well-documented data point). As this research advances, once all data points in a series receive an "A", the quality codes for that series would then be irrelevant and could be dropped from the data set.

A second new element added to these data sets are the identification of anomalies. One of the most routine questions that arise over any data set is the major fluctuations in data values. Oftentimes, these fluctuations reflect true changes in the data. In other cases, however, they can be created by the coders themselves. Changing data sources, differing conversion factors, or introducing new components can create an apparent disconnect in a data series.

In a proactive approach to these discontinuities that appear in many data sets, each component now has an anomaly code column included in the data set. When a potential discontinuity was found in a data series, it was noted and supplemental research was done attempting to identify the cause of the anomaly. In some cases, a specific cause was easy to identify and document, such as changes in population after wars or losses of territory. In such cases, the fluctuation is real, and understandable. In other cases, anomalies were created because of changes in the data structure itself, such as when switching indicators from iron to steel production. In other cases a new source introduces a jump in a series. In these cases, the apparent increase or decrease in an indicator is artificial, and the jump must be accounted for in time-series analysis of the component series. Unfortunately, there were cases where no discernable reason could be found for the anomaly between previous and subsequent data points. These points were documented and it should be the future goal of this project to fully document all the reasons for anomalies in these data sets.

Individual Data Set Updates

Each of the six indicators of national capabilities underwent revisions and updates over the course of this project. While there is more detail in the sections that follow, it is important to note at least briefly what some of the major modifications and improvements are.

The Military Personnel Data Set was both updated and modified. It was modified from previous versions by replacing previous data with data from the U.S Arms Control and Disarmament Agency (ACDA) for all data points from 1961 until 1993. The data were also extended from 1993 forward using ACDA data where possible, supplemented with data from the Military Balance.

The Military Expenditure Data Set was updated from 1993 to 2001[2].

The Iron and Steel Data Set was first updated to 2001. Then researchers went back through the data set and re-confirmed the entire series, re-documenting the sources for all data points in the series.

The Primary Energy Consumption Data Set was completely re-constructed for version 3.0 of the data set. All energy values were re-calculated from raw data sources, and compiled into a total energy consumption data value for each state in a given year. The data were also extended to 2001.

The Total Population Data Set was first updated from 1993 until 2001 using United Nations data. Then researchers went back through the data set, re-documenting the data points; some data series were replaced, and some interpolations were re-calculated.

The Urban Population Data Set was updated from 1990 until 2001.

Notes on the format of data file "NMC_3.0.csv"

The file NMC_3.0.csv contains version 3.0 of the Correlates of War National Material Capabilities Data Set (1816-2001). The file is in "comma-separated-variable" (comma-delimited) form, a flat text format which may also be read automatically into computer software packages such as Microsoft Excel, or read using specific commands into other programs (e.g. using the "insheet using nmc_3.0.csv" command in Stata). The first line of the data set contains the variable names. The data set contains the following 11 variables, in order:

Position	Variable	Description
1	"stateabb"	3 letter country Abbreviation
2	"ccode"	COW Country code
3	"year"	Year of observation
4	"irst"	Iron and steel production (thousands of tons)

5	“milex”	Military Expenditures (For 1816-1913: thousands of current year British Pounds. For 1914+: thousands of current year US Dollars.)
6	“milper”	Military Personnel (thousands)
7	“energy”	Energy consumption (thousands of coal-ton equivalents)
8	“tpop”	Total Population (thousands)
9	“upop”	Urban population (population living in cities with population greater than 100,000; in thousands)
10	“cinc”	Composite Index of National Capability (CINC) score
11	“version”	Version number of the data set

Missing values are indicated by the value “-9”. Users must ensure that their statistical analysis software takes this coding into account.

Military Personnel

The Military Personnel Data Set contains data on the size of state armies (defined below) from 1816 until 2001.

What's New in Version 3.0

This version of the data set has undergone three important modifications. First, whenever possible, researchers have re-documented the source for the data points. Second, the data were extended from 1991 until 2001. Third, the data between 1961 and 1999 now comes from the U.S Arms Control and Disarmament Agency (ACDA). Previous versions used a combination of both ACDA data and data from the International Institute for Strategic Studies (IISS). Version 3.0 uses ACDA data for all data points where it was available and only supplements with IISS data in cases where ACDA data were not available.

Data Acquisition and Generation

Military personnel are defined as troops under the command of the national government, intended for use against foreign adversaries, and held ready for combat as of January 1 of the referent year. It is important to note that any date besides January 1st would have been appropriate for the majority of cases because the data values change slowly. On occasion, however, there are instances where there are rapid changes in troop strength, such as mobilizations for conflicts and wars. Short-term variations in strength are not reflected in the project's data unless the changes remained in effect until the following January 1. With this definition in place, there are five important aspects of quantifying military personnel that need elaboration.

First, the project counted only those troops under the command of the national government. These troop strengths include active, regular military units of the land, naval, and air components. Troops in the reserves such as those found in the United States were not included in the state's annual total. Colonial troops (such as Indian troops under British command during India's colonial period) were usually not included in this total if they were a separately administered force.

Second, the military personnel data exclude the military forces of foreign military forces, semi-autonomous states and protectorates, and insurgent troops. Such units were not part of a regular national armed force under a military chain of command. Their inclusion would distort the number of personnel that could be summoned when deemed necessary.

Third, these figures reflect the project's best judgment on which forces were intended for combat with foreign parties. Irregular forces such as civil defense units, frontier guards, gendarmerie, carabineri, and other quasi-military units were nominally responsible for defending

outlying districts or for internal security and could be mobilized in time of war. We usually excluded them, however, because they were not integral to the regular armed forces (e.g. Cossack troops of nineteenth century Russia). When these forces were the only military a nation had they were still excluded (e.g. Costa Rica and Switzerland).

A fourth aspect concerns armed forces in several semi-feudal nations, including the warlord armies in pre-modern Japan and China, and Jannissary troops in the Ottoman Empire. Not all nations were quick to adopt Western military organization. We counted only those forces that were acting at the behest of the central government. For example, we included only the Imperial troops and those armies of feudal lords operating on the behalf of the throne in the case of pre-modern Japan.

A final aspect concerns national police forces organized for both foreign and domestic purposes and found in several developing nations in the twentieth century. Such units come directly under the military chain of command and are fully a part of the armed forces at the immediate disposal of a national government. Examples include the old National Guard of Nicaragua and the national police forces of many African states. When such forces provided dual functions of foreign combat and internal security, we included them in its military personnel figures; otherwise, they were excluded.

Usually it was only after 1960 that we found ready-made data (including army, navy, and air force totals) meeting our coding criteria and aggregated into the desired totals. Elsewhere, we assembled the data from bits and pieces. Given a figure that did not fully meet our inclusion/exclusion criteria, we used it only after locating supplementary information that could be used to adjust it. Confronted with conflicting figures, we adopted those that best matched the contemporary data, and only if they seemed historically plausible. In practice, frequently it was impossible to find documentation reflecting the January 1 criterion. In most such cases, however, the figures were changing sufficiently slowly to afford an acceptable approximation. In cases of rapid military change, such as the onset of war, we took note of the fact in arriving at a plausible estimate. Because of the relatively great sensitivity of personnel levels to transitory circumstances such as war involvement, we used estimates to fill missing entries only when they did not occur in such circumstances.

Problems and Potential Errors

The precise numbers of active forces remains uncertain in a conceptual basis. It is easy to see that during the course of their foreign policy, states often have an incentive to exaggerate their troop strengths when deterring a potential opponent or understate their troop strength when attempting to avoid notice by other powerful states or a potential target of hostilities. These potential motivations to misrepresent troop strengths can create difficulties with this project's data collection efforts. However, because we use sources that themselves often use multiple sources

and channels of estimation, we believe that these differences in opinions are ironed out of the data and the numbers presented here are reflective of the military personnel of these states.

Inadequate source documentation is another potential source of difficulty in assembling this data. There is some possibility that personnel which were never counted in a general source total have been missed. We were not aware of such flaws in our research, however, and do not consider this a major potential for error.

Similarly, our criteria for including or excluding various "irregular" types of forces may have led us to exclude forces which did indeed contribute to national totals. Equally plausible is that we classed as active some military units that should have been excluded by its criteria, such as those performing internal security functions. Source limitation frequently precluded the requisite distinctions.

Quality/Anomaly Codes

There are no quality or anomaly codes for this component.

Component Data Set Layout

The layout of the military personnel Access data set is found in Table MILPER 1 below. The data set contains seven columns. The first and second columns correspond to the COW state abbreviation and COW state number. The third column is the year of observation. The fourth column contains the value for that year (in thousands), unless the value is missing. Missing values are indicated by -9. The fifth column provides the source of the data point or “See note.” If the column contains “See note,” the note column should be consulted to see how that data point was calculated. The next (sixth) column, “Note,” explains how that data point was obtained (estimation, or whether the value was verified as coming from a particular source). All data points that have been verified are so indicated. The seventh column is entitle “Source Code”, but has not been used and is blank.

Table MILPER 1: Data Set Layout

Military Personnel							
StateAbb	**CCode**	**Year**	**MilPer**	**Source**	**Note**	**Source code**	**Version**
USA	2	1816	17	"Historical Statistics of the U.S., Colonial Times to 1957" (U.S. Department of Commerce and Bureau of the Census) p737	verified 10/22/2001. DLD		3.01

Bibliography

Historical Statistics of the U.S., Colonial Times to 1957. U.S. Department of Commerce (1960). Page 730.

Statistical Abstract of the U.S. U.S. Department of Commerce (1970). Page 257.

The Statesman's Year-Book. New York. .

Almanach de Gotha.

Annual Abstract of Statistics. Great Britain.

Clode, Charles M. *The Military Forces of the Crown*. London: John Murray, 1869. Vol. I, pp. 399-400.

Whiting, Kenneth R. *The Development of the Soviet Armed Forces. 1917-1966*. Maxwell Air Force Base, Alabama: Air University, 1966.

Erickson, John. *The Soviet High Command. A Military Political History, 1918-1941*. London: Macmillan & Co, 1962.

The Institute for Strategic Studies. *The Military Balance, 1965-1966*. 18 Adam Street, London, WC 2. December, 1965.

The Europa Year Book. London: Europa Publishing Co., Ltd.

National Attributes Data.

Upton, John. *The Armies of Asia and Europe*. New York: D. Appleton & Company, 1878.

Schnitzler, J. H. *Essai D'une Statistique Generale de L'Empire De Russie*. Paris, 1829.

The Japan Year Book. 1906-1952. Tokyo: The Japan Year Book Office.

League of Nations. *Armaments Year Book*. Geneva, 1935 = vol. XI.

Wolowski, M. Louis. *Les Finances La Russie*. Paris: Librairie de Guillaumin, 1864.

Great Britain, Central Statistical Office. *Annual Abstract of Statistics.* London: Her Majesty's Stationery Office.

Great Britain, House of Commons, Sessional Papers, 1816-1965.

Cobden, Richard. *The Three Panics: An-Historical Episode*. London: Cassell and Company, Ltd., 1884.

Ravay, Espagnac du. *Vingt Ans Politique Navale, 1919-1939*. Grenoble: Arthaud, 1942.

Mulhall, Michael G. *The Dictionary of Statistics*. London: George Routledge & Sons, Ltd., 1899.

The Encyclopedia Britanica, 11th Edition. Cambridge, England, 1910. Published by University Press, New York..

Monteilhet, Joseph. *Les Institutions Militaires de la France, 1814-1932, De la Paix Armee a la Paix Desarmee*. Paris: Librairie Felix Alcan, 1932.

U.S., Office of Naval Intelligence. *Information Concerning Some of the Principal Navies of the World*. Washington.: Government Printing Office, December, 1911.

Annuaire de L`Economie Politique. Paris: Rue Richelieu, No. 14.

Annuario Statistico Italiano.

Lobell, V*. Jahresberichte uber die Veranderungen and Fortschritte im Militarwesen. (L.C. Title Rustung and Albrustung) 1874-1913*. (Vol . 1= 1874) E. S. Mittler and Sohn, Berlin.

League of Nations. *Armaments Yearbook.* (Yearly 1924-1939.)

The Military Balance: 19_/ _. London, Institute for Strategic Studies.(1959/60 = vol. 1)

U.N. Office of the Secretary-General. *Economic and Social Consequences of Disarmament: Replies of the various Governments anti Communications from International Organizations.* (U.N., New York; 1962)

World-wide Military Expenditures and Related Data: 1965. U.S- Arms Control and Disarmament Agency, Washington; 1967. Research Report 67-6.

*Statistical Abstract of Latin America 19-. (*1954 = Vol. 1) Latin American Center, Univ. of California, Los Angeles; 1965.

M.D.R. Foot. *Men in Uniform*. London, Institute for Strategic Studies; 1961.

New York Times. *Military Survey*. May 12, 1947: p. 14.

New York Times, March 27, 1949. p. E-5.

New York Times, May 16, 1948, p. 4E.

Hanson, Harry (ed.). *The World Almanac and Book of Facts: 19-.* New York, Scripps-Howard.

Information Please Almanac. Doubleday, New York.

Whitakers Almanac 1801. (1961 = 93rd annual). London.

Republic of Ghana, *Statistical Yearbook for 1863*. Accra.

Central Statistical Office. *Statistical Abstract for the U.K.*

Bureau of Statistics, Office of the P.M. *Japan Statistical Yearbook*. (1906 = vol. 1)

Census and Statistics Department of N.Z. *New Zealand Official Yearbook*.

Commonwealth Bureau of Census and Statistics, *Official Yearbook of the Commonwealth of Australia.*

Dominion Bureau of Statistics, Dept. of Trade and Commerce. *The Canada Yearbook.*

Vaterlandischer Pilger. (1809-1848; published irregularly in the early pre-1817) years. Annual thereafter.) Publisher not given. 804.J95

Genealogisch-historisch-statistischer Almanac fur das Jahr-. Verlag von Landes-Industrie-Comptoir, Weimar. (1823= Vol.I) CS27.G32

Australia. *The Parliament of the Commonwealth of Australia. Naval Expenditures of the Principal Naval Powers*. CMD Paper. Nov. 27. 1914.

Readers Digest Almanac. (1966= vol. 1)

Mitteilungen aus dem Gebeit des Seewesens. Seidel und Sohn, Jura.

H. Roberts Coward. *Military Technology in Developing Countries.*

Beer, Francis A., ed. *Alliances: Latent War Communities in the Contemporary World*. New York; Holt, Rinehart and Winston, Inc., 1970.

Annuaire Statistique de la France.

Hinsley. *Catalogue of German Naval Archives*.

Das koenigliche Prussische Kriegsministerium, 1809-1909 (Berlin, 1909).

Guichi, Ono. *War and Armament Expenditure of Japan*. (Concord, Carnegie Endowment for International peace,]922)

Rathgen, Karl. *Japans.Volkwirtschaft and Staatshaushalt*. Leipzig, Verlag von Duncker and Humboldt, 1891)

Australia, Parliament. *Naval Expenditures of the Principle Naval Powers*. (Cmd Paper, Nov. 27, 1914.)

Kolb, Georg Friedrich. *Handbuch der Vergleichenden Statistik*. (Leipzig, A. Forstnersche Buchhandlang, published irregularly beginning in 1862)

Schnabel, *General Statistik des Europaischen Kaiserstaats*. NA1107.536 1841

Almanach de Paris. (Paris, annual, 1865=vol. 1)

Genealogisches and Statistisches Almanach fur Zeitunpysleser. (Weimar, only one volume in L.C., no information as to dates of series publication)

Block, Maurice. *Statistique de la France*. (2 vol.) Paris 1860.

France. Ministre du Commerce. *Documents Statistiques sur la France*. (Paris, Imprimerie Royale. 1835)

Great Britain, House of Commons. *East India Reports. Synopsis of the Evidence Taken Before the Select Committee in Relation to the Army of India*. 1832.

U. S. Arms Control and Disarmament Agency. *World-Wide Military Expenditure and Related Data 1971*

Stein, Friedrich von. *Geschichte des russischen Heer*.(leipzig, 1895).

U.S. War Dept., Military Commission to Europe. *The Armies of Europe*. (Philadelphia, 1861.)

Ruestow, Wilhelm von, *Die Russische Armee*. Vienna, 1876.

Fadieev. *Russias Kriegsmacht Und Kriegpolitik*. Leipzig, 1870.

Larroque, Patrice, *De La Guerre Et Des Armees Permanents*. Third ed., Paris, 1870.

Jerome Chen, 1972 (no title given)

Fijiwara Akiva, 1961. *Gunjishi (The History of Military Affairs in Japan)*

China Yearbook

China H.B.

J. G. Godaire "The Claim of the Russian Military Establishment" in *Dimensions of Soviet Economic Power* (GPO, Washington, 1962)

A.J.A. Gerlach. *Fastes Militares des Indes-Orientales Neerlandisches* (?) Paris, 1859)

Bodart, Gaston. *Losses of Life in Modern Wars* (Oxford, Clarendon; 1916)

O'balance, Edgar. *The Sinai Campaign of 1956* (New York, Praeger, 1960)

Clowes, William Laird. *The Royal Navy; A History from Earliest Times to the Present.* (London, 1901) 7 vol.

Schnabel, JamesF. Schnabel. *The United States Army in the Korean War. Policy and Direction: The First Year*. (Office of the Chief of Military History, United States Army, Washington, 972)

Naikoku Tokeikyoku (Cabinet Statistical Bureau), *Meiji Hyakunen Shiryo* (Statistical Data for the Hundred Years since Meiji) (Tokyo, 1967)

Appleman, Roy E. *The United States Arm in the Korean War: South to the Nakton , North to the Yalu*(?) Office of the chief of Military History, Department of the Army, Washington, 1961)

Naikoku Tokei Kyoku: *Nihon Teikoku Tokeineukau* (check spelling)

0'balance, Edgar. *Korea 1950-1953* (Faber and Faber, London, 1959)

Walter G. Hermes. *The United States Army in the Korean War; Truce Tent and Fighting Front.* (Office of the Chief of Military History, United States Army, Washington, 1966).

Howard, Michael. *The Franco-Prussian War* (ICY. Collier, 1969; first published 1961)

Hammer, Kenneth M. "Huks in the Philippines." in Franklin M. Osaka. *Modern Guerilla Warfare* Free Press of Glencue, 1962, pp. 177-183.

Wood, David G. "Twentieth Century Conflicts" Adelphi Papers #1968 (From English (?) Foreign Policy Institute.)

Ralph L. Powell. *The Rise of Chinese Military Power 1895-1912.* Princeton U. Press, 1955

Statistisk Arsbok fur Sverige (Statistical Abstract for Sweden)

John Gittings. *The Role of the Chinese Army.* London, Oxford U. Press, 1967

Dernberger, Robert "Evaluation of Existing Estimates For China's Military Costs and Preliminary Illustration of the 'Best' Available Method for Making New Estimates." (unpublished paper)

U.S. Arms Control and Disarmament Agency. *World Expenditures and Arms Transfers, 1966-1975*. Washington, 1976.

Rothenberg, Gunther. *The Army of Franz Joseph*., Purdue Univ. Press, W.Lafayette, 1976.

Alvin D. Coox, "Effects of Attrition on National Wav Effort: The Japanese Army Experience in China, 1937-38," *Military Affairs* v. XXXII, no. 2. (1968), pp.57-62.

Military Expenditures

The second indicator of military capabilities is military expenditures. Military expenditure is defined as the total military budget for a given state for a given year.

What's New in Version 3.0

The data were updated through 2001.

Data Acquisition and Generation

Since our primary interest was to index all financial resources available to the military in time of war, we coded all resources devoted to military forces that could be deployed, irrespective of their active or reserve status.

Appropriations for all the types of units mentioned earlier were included when the units were under the authority of officials of the national government, even if the units did not contribute to the personnel variable. Such units typically were excluded from published budgets, in any case. It is important to note that in our assessments the sources of military expenditure data often provided gross (rather than net) expenditure figures.

We sought to identify and exclude all appropriations of a non-military character because some nations have civil ministries under military control (national police forces is the most prevalent example). The use of such unadjusted budgets would substantially over-estimate the military capability of those nations. If there was a clear bureaucratic division between the execution of civil and military functions, this task was easily accomplished. For instance, if there were separate accounting and authorization procedures for merchant- and military-marine, expenditures of the former were excluded. On the other hand, merchant marine expenditures charged to the same administrative units which carried out military marine functions were included in the project's tabulations. Likewise, the budget figures were adjusted upward where we determined that outlays in other parts of the budget served to enhance military capacity.

Having made the above distinction concerning money spent on military forces, we delimited part of the latter directly related to a country's war fighting capacity; that is, we had to distinguish which figures going for military purposes were destined to enhance capability. We deemed that expenditures on pensions, superannuation pay, relief, and subsidies to widows and orphans do not contribute to military power and excluded them where possible. For most statistically developed countries, these items were found to be readily identified in a separate section of the military budget, or charged outright to the finance ministry.

We decided to identify gross rather than net expenditures, so as to sidestep problems of accounting for the yearly variations in stockpile buildup, depreciation, and liquidation. As with the

accounting of energy stocks, little was found that would have allowed us to determine net expenditures.

We closely attended to allocations, usually found in supplemental budgets, special accounts, and war credits and loans, over and above regular appropriations. Examples include the special funds and credits voted during the mobilizations prior to and during the two world wars, and the loans contracted by Prussia prior to the Franco-Prussian War.

With regard to these special appropriations, some ambiguity exists as to which year the expenditures should be assigned. Since our objective was that each unit of currency spent on military capabilities should be counted only in the year that it directly enhanced military capability, it counted surpluses and credits transferred from past years (when known) among the expenditures of the referent year.

For example, expenditures from special accounts (such as the construction of fortifications or the purchase of armaments) were included in the expenditure totals for that year. If the special account was composed of transfers from the general budget, expenditures on that account were included in the year in which they were spent or projected to be spent. If the special account was composed of credits budgeted to a war ministry in previous years, but unspent in those previous years, we included only actual expenditures from that account in the project's totals for the appropriate years. Outlays for the amortization of debts incurred were excluded, since the project had already counted them in the year in which the military items were acquired. Thus, if a naval ship was acquired in 1923 but not paid for until 1926, we counted the corresponding expenditure in 1923. Surplus military appropriations from previous years were counted as military expenditures only for those years when the funds were actually spent.

The customary difficulties in Soviet statistics were resolved by period. For the years prior to the Second World War, the fragmentation of the evidence precludes an appraisal of real expenditures. Rather than engage in speculation, the project reported the official figures published in the League of Nations Armaments Year-Book from 1924 to 1940. From 1955 to 1963 we utilized SIPRI estimates and from 1963 on have used ACDA figures.

Currency Conversion

In most cases, expenditures were originally collected in national currency. The data were then converted into a standard unit - British pounds sterling prior to 1914, U.S. dollars thereafter - using the COW currency conversion dataset (which uses current exchange rates). We entered beginning of the year market rates wherever available, except for periods of marked inflation in the twentieth century, where we entered black market rates, if available. This was the case for most Western nations throughout the data period, and for most nations since 1945. Otherwise, we used government rates, except for Eastern European states in the period after 1945, for which we used dollar amounts. In all remaining cases - most in the first half of the nineteenth century,

for which documentation is particularly scarce - we used project estimates. Principal sources were the Times (London) for the years prior to 1914, League of Nations *Statistical Yearbook* for 1919-1939, and International Monetary Fund from 1948 onward. Supplementary sources included *de Gotha* and *Statesman's Yearbook,* as well as economic and historical monographs.

To moderate short-term fluctuations, we sometimes revised the resulting series by a smoothing process that used a seven-year moving average. A prime example of its application is the smoothing of rate changes during the wholesale suspension of the gold standard in the 1930s. In the event of introduction of a new currency, we omitted this process. Occasional interpolations were performed to fill small intervals in a series, but only when currency conditions seemed stable. Data during extremely inflationary times (e.g. Germany during the early Weimar Republic) should be viewed with special care.

Problems and Possible Errors

It was often difficult to identify and exclude civil expenditures from reported budgets of less developed nations. For many countries, including some major powers, published military budgets are a catch-all category for a variety of developmental and administrative expenses - public works, colonial administration, development of the merchant marine, construction, and improvement of harbor and navigational facilities, transportation of civilian personnel, and the delivery of mail - of dubious military relevance. Except when we were able to obtain finance ministry reports, it is impossible to make detailed breakdowns. Even when such reports were available, it proved difficult to delineate "purely" military outlays. For example, consider the case in which the military builds a road that facilitates troops movements, but which is used primarily by civilians. A related problem concerns those instances in which the reported military budget does not reflect all of the resources devoted to that sector. This usually happens when a nation tries to hide such expenditures from scrutiny; for instance, most Western scholars and military experts agree that officially reported post-1945 Soviet-bloc totals are unrealistically low, although they disagree on the appropriate adjustments.

We also encountered difficulty concerning lack of sufficient information about local currencies. Nineteenth century sources frequently shift from one name to another, for the same currency. Thus, *Almanac de Gotha* uses the "thaler", the "thaler en espece," and the "riksdaler" as currency unit names. After consulting several sources dealing with currencies, we determined all three to be the same unit. Occasionally, the sources report a budget, particularly of states newly independent in the nineteenth century, in *different* units from one referent year to the next. Thus, *Statesman's Yearbook* and *Almanac de Gotha* report Guatemalan expenditures first in silver pesos and later in paper pesos. Although we encountered situations in which currencies of the same name but of different values were in circulation, usually the values were sufficiently different to distinguish by comparison the units in question. Not surprisingly, these difficulties

were less prevalent in later years. Thus, SIPRI informed us that their series are always represented in the most recent currency unit, to which prior data are adjusted. Again, usually the scale of the reported figures is indicative of the referent unit.

A final problem concerning currency conversion is conceptual in nature. When comparing economic magnitudes across time or space, there is a choice to be made concerning what price weights apply to what quantities of each good or service under consideration. Our particular choice of standard units (sterling and dollars) implies a decision to assign these weights to each nation's military program according to British or U.S. opportunity costs for the referent year. This choice is implicitly made when the project converts local currency units to sterling or dollars, for it is then computing what Britain or the U.S. would have given up in order to make the same purchases. Given the relatively free international monetary and trade movements that obtained during much of nineteenth century, in which pounds sterling, dollars, francs; deutschmarks, lira, etc., were readily convertible into each other, there was arguably a single world economy. These opportunity costs would then have been approximately the same for any standard unit, since each nation was drawing on this single economy. In the most autarkical situations that occasionally arose in the twentieth century, the opportunity costs were no longer roughly equivalent; the relative monetary costs often depended on the currency in which they were expressed. This was the situation during the world wars, when normal monetary and commercial exchange was disrupted.

The most extreme cases, however, are the economies of the Soviet Union, China since 1949, and the centrally directed economies of Eastern European states since 1945, for which there has been relatively little freedom of movement. Here, one might find Soviet military expenditures exceeding U.S. expenditures, when they are valued in U.S. dollars, but the reverse, when they are both valued in rubles. Moreover, because Soviet prices were set by fiat rather than by market bidding, the prices of military goods and services, compared among themselves or to civilian items, are not necessarily reflective of their relative value in the sense that we normally ascribe, even as measured in the local currency. These difficulties compound the problem we noted earlier, that the military accounts in question have often been distorted or partially hidden to outside eyes. Like others before it, we found no way around these inherent ambiguities. In the cases noted, we simply stuck with them.

The Future of Military Expenditures

Two tasks exist relevant to the future of the military expenditure data. First, the military expenditure data set requires that there be a consistent, accurate, and well documented currency conversion data set. Raw military expenditure data often come in a variety of different monetary units, such as rubles, dollars, pounds, franks, or marks. Because of these differing units, it is

quite important to have a universally accepted and accurate key for converting all those raw data values into one common metric.

Unfortunately, the original COW Currency Conversion data set appears to have been lost to time, and so although we have converted expenditure variables, we do not have the conversion series.

If a new version of the currency conversion data set could be completed, a second major endeavor of the military expenditure data set could begin: the re-documentation of the data points before approximately 1960.

Quality/Anomaly Codes

There are no quality or anomaly codes for this component.

Component Data Set Layout

The layout of the data set is found in Table MILEX 1 below. The data set contains six columns. The first and second columns correspond to the COW state abbreviation and COW state number, respectively. The third column is the year of observation. The fourth column contains the value for that year (from 1816 to 1913, in thousands of current year British pounds and from 1914 onwards, in thousands of current year U.S. dollars), unless the value is missing. Missing values are indicated by -9. The fifth and sixth columns contain any information that was available from the original COW project. Since we did not attempt to verify this data, these columns are often left blank, in cases where we could not find any information about sources from the original project.

Table MILEX 1: Data Set Layout

Military Expenditure						
StateAbb	**CCode**	**Year**	**MilEx**	**Source**	**Note**	**Version**
USA	2	1816	3823			3.01

Bibliography

See Source Notes in sub-component data set.

Total Population

The total population of a state has been theorized to be one of the major factors in determining the relative strength. A state with a large population can have a larger army, maintain its home industries during times of war, and absorb losses in wartime easier than a state with a smaller population.

What's New in Version 3.0

The series was updated through 2001. The original data were verified and in some cases replaced.

Data Acquisition and Generation

While the most reliable total population figures usually appear in national government tallies, modern census-taking was rare before 1850 in Europe and countries of European settlement, and rare before the First World War elsewhere. In all periods, the accuracy and reliability of national census data seem to vary with the level of economic development. As a result, data from the developing world require particular scrutiny.

A census may be of the *de facto* population, comprising all residents within the national boundaries, or of the *de jure* population, comprising only those who are legal residents. We used the former, where possible, to which totals of military personnel abroad were added. Since the differences between *de jure* and de facto (between "total" and "total home") population are typically small, we did not analyze this data for sensitivity to these coding distinctions.

The United Nations Statistical Office has an estimated yearly total population series, corrected for over- and under-enumeration to the extent possible, for most nations since 1919. We relied on those series where possible.

For prior years and nations where we found one or more plausible time series, we took data from the sources presenting the greatest continuity with the U.N. data. We uncovered most of the general censuses taken since 1816 and used alternative sources for the numerous remaining gaps. For example, Japan maintained a system of population registration through a rough running tally. Other countries took sample surveys from which they constructed estimates of the total population. We judged these sources the most reliable.

For the occasional nation maintaining reasonably complete registers of vital events (e.g. the United Kingdom), we estimated missing data utilizing Formula TPOP One:

Formula TPOP 1: Missing Total Population Data Estimations

$$p(t) = p(t_o) + b(t) - d(t) + i(t) - e(t),$$

where:
p(t) is the known or estimated population at time t,
p (t_o) is the population recorded at time to, and
b(t), d(t), i(t), and **e(t)** are the respective numbers of births, deaths, immigrants, and emigrants recorded since t_0.

Net migration is usually small enough to be safely disregarded. For nations maintaining registers of births and deaths but not of migration, we estimated i(t) - e(t) to be zero.

In lieu of complete demographic records, we resorted to estimation either (1) by interpolation, or (2) by least-squares linear regression with time as the independent variable. The choice between them was based on the availability and quality of information, and on whether the period in question was marked by major wars or territorial boundary changes. First, however, we must note four types of situations in which such change did not take place.

The first concerns the many cases for which censuses had been taken regularly. In these cases, the only missing records were for the intervening years. In such instances we interpolated between census records using Formula TPOP Two Below:

Formula TPOP Two: Interpolation Between Known Data Points

$$\log p(t) = \frac{(\log p(t_2) - \log p(t_x)}{(t_2 - t_1)(t - t_1) + \log p(t)}$$

where:
$p(t_1)$ and $p(t_2)$ are the known population figures at time t_1 and t_2.

This method entailed the assumption of a constant growth rate over the period delineated by t_1, t_2, and t_x from which the formula is derived.
In a second type of situation, taking account of the country's demographic history, the manner and quality of its census-taking, the later population trends, and the opinion of demographers, we were able to discern a plausibly reliable population series even though regular censuses were not available. Again, we resorted to interpolation as here it seemed appropriate.

A second type of concern arose where data for the final years in a series were missing, either because of loss of national identity (e.g. Estonia, Latvia, and Lithuania in 1940, or Austria-Hungary during World War One). In these cases, we resorted to extrapolation using the above interpolation formula.

A third concern was a problem of having no uniform data series at our disposal and what sources were available providing only a patchwork of spotty and conflicting coverage. In this type of situation, we estimated population by regression performed on the logarithms of the known data. A prime consideration in our willingness to use this technique was that data for well-documented nations indicates that growth rates usually change quite slowly. Distortions were thereby introduced but not to as great a degree as would arise from applying interpolation and

extrapolation to these highly erratic data. Where necessary, to bring the estimates into agreement with the uniform (typically post-1919 United Nations) series of an adjoining epoch, we raised or lowered the regression line - while maintaining the same slope - such that the line passed through the adjacent values of the series.

Finally, in situations of marked discontinuity in population trends associated with wars and exchanges of territory, we applied the above methods as appropriate, but only to the separate intervals on either side of the discontinuity, and only where it could document its demographic magnitude. Interpolation, for example, could be used only if total population before and after the break, or one of them plus the magnitude of the change, was known. We treated all cases in which the nation gained or lost at least 1% of its total home population in this manner. For territorial exchanges, we were able to document many of the gains and losses. We were, however, usually unsuccessful in documenting war losses. Its method was to adjust for the affect of territorial changes and then to extrapolate forward from pre-war and backward from post-war data. Unless otherwise noted, population losses due to war were prorated over its duration. The most pronounced instance was over estimate of Chinese population during and immediately after the Taiping rebellion.

Problems and Potential Errors

There are two difficulties with territorial boundary changes and with our estimation assumptions. Concerning the former, the United Nations series occasionally fails to adjust the base to reflect them; estimates for prior years may measure the population living within the present national boundaries even though territorial changes occurred in the interim. We attempted to determine where this was the case and make adjustments to reflect the actual boundaries at the time.

The second difficulty is that we assumed a constant growth rate in regression. We regard any observed deviations from constant growth as due to under or over-enumeration. This procedure would cause us to miss the effects of, for example, a famine. The population growth rate for a particular state may have been much higher than our estimate; when a famine stuck, however, for a few years between those censuses the population dropped severely, but then resumed a higher growth rate than our estimation procedure captured. The result is that our estimation procedures would be the actual trend of the population, and not an accurate reflection of the year-to-year population fluctuations. We assume that these circumstances are rather rare and generally not of a magnitude large enough to cause the data to be distorted in any significant manner in the aggregate, particularly when combined in the aggregate CINC score.

Quality Codes

In version 3.0 of this data set, a measure has been included to capture the source of each data point, reflecting the confidence we have in each point. The quality codes for total Population are listed below in Table TPOP 1 below.

Table TPOP 1: Total Population Quality Codes

Code	Interpretive Meaning
A	Value from identified source
B	Linear Interpolation from identified sources
C	Linear Interpolation from at least one unidentified source
D	Regression from identified sources
E	Regression from at least one unidentified source
F	Extrapolation from identified sources
G	Extrapolation from at least one unidentified source
M	Missing value.

In documenting and revising the entire population data set for version 3.0, we first identified data points in the 2.1 data for which we have or do not have an identified source. States with the most accurate data were given a rating of "A." Data points generated from linear interpolation were given a rating of "B" if they were produced using two known data points and a "C" if they came from one or more unidentified sources (including a value from the version 2.1 capabilities data set if the source was unknown). Data generated utilizing regression techniques received quality codes of "D" and "E", again based on the number of known data points that were utilized in generating them. Extrapolated data points received quality codes of "F" and "G." Any missing data points received a quality code of "M." It is important to note two things about this quality code scheme. First, it is *NOT* meant to be an ordinal scale for all data points; while all "A" data points are of the highest quality and standards, a point with a "C" quality code should not be taken as being of superior quality than a "G" valued-point. The second important point to note is that the vast majority (over 85%) of the data points have a quality code of "A." The eventual goal of this project should be to gather data on the points where data is less available (codes "B" through "M") and convert them into "A" values.

Anomaly Codes

Version 3.0 of this data set also identifies points where the time series of total population makes radical changes. Identifying these inconsistencies will make future versions of this data set more robust, as it will be easier to identify where there are difficulties and concerns with particular data points.

Each indicator of the CINC possesses differing standards for what constitutes an anomaly. For total population, the standard change is a two percent increase or decrease in one

year's time. For some scale, this would be the equivalent of the United States losing the total population of the state of Washington in just one year. If the population changed by more than +/- 2%, we investigated the data point further to try to determine the source of the rapid growth. The lists of anomaly codes for total population are listed in Table TPOP 2 below.

Table TPOP 2: Total Population Anomaly Codes

Code	*Substantive Meaning*
A	No Anomaly (< 2% change)
B	Explained Inconsistency (e.g. change in territory, loss in wartime)
C	Change of Sources (between 2 non-UN sources or 1 non-UN to UN source)
D	Change of UN Sources
E	UN Internal Inconsistency within same UN source
F	Internal inconsistency within non-UN source
G	Unexplained Anomaly

Over 95% of data points for total population are "A"s. A second point worth mentioning about these codes is that "C", "D", "E", and "F" are constrained by time. "F" values are almost always before 1919 when the League of Nations began collecting data, while the other three potential anomalies are always found during the times when UN data is utilized (1919 to the present).

Component Data Set Layout

The layout of the Access sub-component data set is found in Table TOT POP 3 below. The data set contains nine columns. The first and second columns correspond to the COW state abbreviation and COW state number, respectively. The third column is the year of observation. The fourth column contains the value for that year (in thousands), unless the value is missing. Missing values are indicated by -9/ The fifth column provides the source of the data point or "See note." If the column contains "See note," the note column should be consulted to see how that data point was calculated. The sixth and seventh columns, respectively, list the data anomaly and quality codes for that value. The eighth column, "Note," explains how that data point was obtained (i.e. linear interpolation, extrapolation, etc.). This column is usually empty for data points with a quality code of A. The ninth and final column lists the version number of this data set.

Table TOT POP 3: Data Set Layout

Total Population 3-0								
StateAbb	**CCode**	**Year**	**TPop**	**Source**	**AnomCode**	**QualCode**	**Note**	**Version**
USA	2	1816	8659	Historical Statistics of the United States: Colonial Times to 1970 Part 1, Page 8	A	A		3.01

Bibliography

Bunle, Henri. 1954. *Le Mouvement Naturel de la Population dans le Monde de 1906 a 1936.* Paris: L'Instit National d'Etudes Demographiques.

Flora, Peter. 1983. *State, Economy, and Society in Western Europe 1815-1975 'A Data Handbook in two Volumes.* Frankfurt: Campus Verlag; London: Macmillan Press; Chicago: St. James Press.

France, Bureau de la Statistique Generale de la France. 1907. *Statistique internationale du movement de la population d'apres les registres d'etat civil.* Paris: Imprimere Nationale.

Maddison, Angus. 1995. *Monitoring the World Economy, 1820-1992.* Paris: OECD.

Mitchell, Brian R. 1988. *British Historical Statistics.* Cambridge: Cambridge University Press.

Mitchell, Brian R. 1998. *International Historical Statistics: Europe, 1750-1993.* London: Macmillan Reference; New York: Stockton Press.

Mitchell, Brian R. 1998. *International Historical Statistics: Africa, Asia & Oceania, 1750-1993.* London: Macmillan Reference; New York: Stockton Press.

Mitchell, Brian R. 1998. *International Historical Statistics: the Americas, 1750-1993.* London: Macmillan Reference; New York: Stockton Press.

Statistics Division of the Department of Economic and Social Affairs of the United Nations Secretariat, United Nations *Demographic Yearbook Historical Supplement 1948-1997.* 49th issue. CD-Rom.

Tir, Jaroslav, Philip Schafer, Paul F. Diehl, and Gary Goertz. 1998. "Territorial Changes, 1816-1996." *Conflict Management and Peace Science* 16: 89-97.

Wilkie, James. 2000. *Statistical Abstract of Latin America.* V36. Los Angeles: University of California at Los Angeles, Latin American Center Publications.

Urban Population

Besides sheer numbers of people, it is important to capture other elements of a state's population. Factors such as education, societal organization, and social services are not captured by the measure of total population. In order to capture the net effect of these more abstract and amorphous ideas, this project includes a measure of urban population. Urbanization is associated with higher education standards and life expectancies, with industrialization and industrial capacity, and with the concentrated availability of citizens who may be mobilized during times of conflict.

What's New in Version 3.0

The series was updated through 2001. Some series were recomputed when new data suggested that reinterpretation or extrapolation was necessary.

Data Acquisition and Generation

"Urban population" is a difficult concept to specify and operationalize for a professional demographer, let alone an international relations researcher. What criterion best captures the meaning of the term? A common approach is to include all cities that exceed a size threshold. Many such thresholds, ranging from 5,000 to 100,000 inhabitants, have been advanced. By virtue of its simplicity, we adopted the threshold criterion using the upper value of 100,000.

This choice has the advantage of facilitating data completeness, which is problematic at lower values. It has the corresponding liability that, in the early 1800s, many areas that one might consider "urban" did not contain 100,000 people. Moreover, the approach appears less well suited for the contemporary period, when build-up areas frequently are comprised, in large part, of many smaller cities and unincorporated places.

While the best data came from national censuses, several of them do not tabulate urban population. Some developed nations take sample surveys to construct reasonable estimates of urban population while multinational sources and demographic experts also publish data based on their own estimation procedures. We used such estimates whenever they did not contradict formal census figures.

The data reflect varying national definitions of what constitutes an incorporated city or urban area; we used these figures where alternatives were unavailable. Occasionally, a source changed its city definition, thus creating a discontinuity in the time series. In instances before 1945 where more than one alternative was offered as to the boundaries of a city, we adopted the one more closely reflecting the built-up area. Otherwise, we entered the data as it was reported.

Occasionally, the data reflect a mix of and *de jure*[3] information. In some states, it was the case that there would be *de facto* data for one urban area while there would only be *de jure*

data for another urban area of within a state. For instance, looking to Russian urban data, it is rather easy to find recorded urban population data for the Moscow urban area; finding recorded data on St. Petersburg or Vladivostok is much more challenging. Usually we found only one or the other; secondary sources offered scant clarification in order to present a series with as much documented data as possible. Faced with this ambiguity, we averaged across *de facto* and *de jure* totals. For the occasional country that mixes data from different years in the same report, the project used interpolation and extrapolation to estimate the referent year.

Often, the value of the same urban population datum is revised from one demographic yearbook to the next. Presuming that revised data are more accurate, we used them. When, as often was the case, this introduced a discontinuity between the first year appearing in the revised series and the previous year appearing in the old, we performed log-linear regression on all the old data in our pooled series and adjusted the regression line to match the revised data points.

When we encountered numbers from other sources significantly different from the United Nations series, we used the U.N. figures unless they were irregular. In the latter cases, we used the log-linear regression method on available data points, the United Nations and otherwise. For cases of recently declining urbanization (e.g. Belgium and the Netherlands in the 1970s), we filled the data gaps in the same way using a constant *negative* growth rate.

We conceive of urbanization as a continuous process, for which the growth rate should vary smoothly. On the other hand, the inclusion of additional cities, as they exceed the population threshold, introduces discontinuities in the census totals. Moreover, some cities appear in one enumeration, but are absent from the next. Cities also occasionally make first-time appearances bearing totals well over the threshold population value. Secondary sources remedied the situation to a limited extent. Since interpolated and extrapolated values can be dominated by such irregularities, we frequently used log-linear regression as a means of smoothing the data obtained by the above methods to obtain a final estimate.

Problems and Potential Errors

In the contemporary period, there is some debate over whether urban population or urban agglomeration is a better measure of a country's level of development. Urban agglomeration includes both the population of people living within the city proper and its suburbs. Since there has been a population shift away from large cities and toward suburban areas in most industrialized countries, we thought that shifting to urban agglomeration data in the years after 1945 would provide a better indicator of each country's level of development.

Unfortunately, logistical problems prevented us from making the shift from urban population to urban agglomeration. The United Nations statistical yearbook reports figures for both urban population and urban agglomeration, but the numbers for urban agglomeration are much less complete. Many countries do not report urban agglomeration at all, and those that do

generally report it less frequently than they report urban population. Furthermore, there is not one year in which a critical mass of countries began reporting urban agglomeration. While some developed countries began releasing data on urban agglomeration shortly after the end of World War II, other developed countries did not release any data on urban agglomeration until the 1980's.

A small number of countries released only urban agglomeration data instead of urban population data. In those cases, we included the agglomeration figures in the data set with a note indicating that they were agglomeration rather than population figures.

We also investigated a number of other sources for data on urban population, most notably the U.N. World Urbanization Prospects. While these sources provided data at regular intervals, they did not provide a clear definition of urban population, and so we did not use these sources.

Quality Codes

Urban population employs a system of alphabetical codes to identify the relative strength or confidence a particular data point, as listed in Table UPOP 1.

Table UPOP 1: Urban Population Quality Codes

Code	*Substantive Interpretation*
A	Value from UN Demographic
B	Assumed 0 (Ex.: Vanuatu)
C	Linear Interpolation from identified sources
D	Linear Interpolation from at least one unidentified source
E	Extrapolation from identified sources
F	Extrapolation from at least one unidentified source

It is important to note that there is a unique quality code for urban population—the assumption of zero, coded as a "B." While it is a rare data value, there are states where there are no cities that reach the standards of 100,000 set above. It is also important to note that these quality code numbers are not meant to be an ordinal scale, except for the value of "A," which should be taken to be the most reliable and best quality data points in the data set.

Anomaly Codes

There are no anomaly codes for this component.

Component Data Set Layout

The layout of the Access sub-component data set is found in Table UPOP 2 below. The data set contains eleven columns. The first and second columns correspond to the COW state abbreviation and COW state number, respectively. The third column is the year of observation.

The fourth column contains the value for that year (in thousands), unless the value is missing. Missing values are indicated by -9. The fifth column provides the source of the data point, when this information is available. The next two columns deal with cases where figures were estimated using the growth rates. The column “Growth Rate” gives the number of the growth rate for that particular year if needed/used, while the “Growth Rate Source” column indicates the source for that rate. The “Note” column contains any other pertinent information. The ninth and tenth columns, respectively, list the data quality and anomaly codes for that value. The eleventh and final column lists the version number of this data set.

Table UPOP 2: Data Set Layout

Urban Population										
StateAbb	**CCode**	**Year**	**UPop**	**Population Source**	**Growth Rate**	**Growth Rate Source**	**Note**	**Quality**	**Anomaly**	**Version**
USA	2	1816	101				For 1810, HS US 1975 gives 0.			3.01

Bibliography

See Source Notes in sub-component dataset.

Iron and Steel Consumption

Iron and Steel Production is one of the two components of the industrial dimension, and one of the six indicators of national power. It reflects all domestically produced pig iron before 1899 and steel after 1900.

What's New in Version 3.0

In addition to cleaning and updating the data set through 2001, this version contains two main new features: data quality codes and anomaly codes.

Data Acquisition and Generation[4]

Iron and steel production trends since 1816 involve transitions concerning the categories of iron produced and the types of fuels used in making iron and steel. In general, "cast iron" means all iron, including "pig iron" that has at least 0.3% carbon. Specifically, cast iron includes all iron that has been molded into functional shapes. "Wrought iron" ("puddle iron" or "bar iron") is made from pig iron (except in a small percentage prior to 1850, when it was made directly from ore) in a puddling furnace. It is very pure (containing less than 0.04% carbon) and relatively malleable. Steel has an intermediate carbon content between 0.04 and 2.25%.

Until around 1870, cast iron and wrought iron were the principal products. The proportion of the former as a final product steadily decreased until castings, as a proportion of total blast furnace production, amounted to less than 0.1% and wrought iron became the primary metal of construction.

By 1880, the Bessemer invention and improvements in coking made wrought iron production obsolete. The use of coke as an inexpensive, non-volatile, and structurally solid fuel allowed the construction of larger blast furnaces. The use of coke combined with the rapid steel production in the Bessemer invention, made steel the primary commercial metal.

While wrought iron was of primary importance as a finished good prior to 1870, we did not use it as an indicator because: 1) pig iron data is more readily available; 2) in our judgment, use of the former would underestimate industrial activity in some states, notably the United States; and 3) such use would downplay the importance of cast iron production, especially prior to 1850. Steel production totals were too low in many states to reflect accurately industrial activity in the nineteenth century. Instead, for the years 1816-1899, we estimated iron production from pig iron output. When direct castings output was reported separately from pig iron, we added these totals to the reported pig iron. This reflects our judgment that direct castings are nothing more than "cast" pig iron. Our selection of crude pig iron plus separately reported direct castings is plausible because this output was part of every activity in the iron and steel sectors of the economy.

Where iron production appeared in disaggregated form, we summed the appropriate raw figures to form the total pig iron output. This was done most often for Prussian and Austrian data when we had to transform the old Prussian and Austrian centners into tons.

By 1900, the preferred product of this economic sector was clearly steel, hence our use of steel output as an indicator. This date is somewhat arbitrary since any year from 1890 to around 1910 could have been chosen for the same reason. It is, however, a reasonable midpoint for our analysis. By 1910, virtually every nation that produced iron in the nineteenth century had matched in the output of steel its previous rank as measured in pig iron. We are confident that the two indicators are roughly equivalent measures of industrial activity at the point of transition.

Data Sources

The approach for refining and updating the data were similar to detective work. In many cases we had the data and a source list but did not know which sources corresponded to which values. As a result, we had to rely on memos from the original COW project at the University of Michigan to put the puzzle together. Hence, most of the sources used were the same as those used by the COW. However, some minor changes were made when extending the data set to 2001. In some cases, the original COW data set had estimates for which we were unable to identify a source. Since there is no reason to doubt those numbers we retained those estimates. A list of states where we used COW1 estimates is found in Table IRST 1.

Table IRST 1: States Utilizing Original COW 1 Data Points

State	Years	State	Years
Mexico	1874-1899	Saxony	1837-1867
Netherlands	1945	Wurttemberg	1834-1870
Switzerland	1850-1899	Austria-Hungary	1816-1820
Bavaria	1816-1833 1851-1852	Greece	1979
	1859-1871	Sweden	1816-1820
Germany	1945	China	1860-1899

In many cases the values from the original data set matched the values in B.R. Mitchell's volumes of International Historical Statistics. A second important source which allowed us to update the data set until 2001 was the Steel Statistical Yearbook published by the International Iron and Steel Institution. This not only allowed us to update the data but also to discover that there were many states which according to COW1 had a value of zero, but which actually had production. Changes were made to replace those zero values with the values provided by the Steel Statistical Yearbook. All the states in which we found no evidence of production were given a value of zero and a note was placed in the note column indicating that there was no known production capability for these states.

Table IRST 2: Interpolations and Extrapolations

Log Linear Interpolations		Log-Linear Extrapolations	
State	**Years**	**State**	**Years**
United States	1816-1927	Cuba	1956-62
France	1816-1819 1821-1823	Mexico	1919
Spain	1831-1845	Belgium	1830
Poland	1919	Switzerland	1905-14
Albania	1981-1987	Germany	1816-20
Rumania	1919, 1941, 1942, 1944	Italy	1816-45
Soviet Union	1941-1944	Yugoslavia	1919
Denmark	1939-1940	Bulgaria	1908-1936
Morocco	1976-1979	China	1935
Egypt	1957		
Israel	1954-1958		
Pakistan	1983-1989		
Dem. Republic of Vietnam	1984-1989		

For some states where a complete series did not exist the value was estimated using log linear interpolation, or in some cases log linear extrapolation. States where log linear interpolation or log linear extrapolation was performed are listed in Table IRST 2 above.

Table IRST 3: Special Estimation Cases (See Data File for Specific Details)

State	Years	State	Years
Netherlands	1816-1830, 1831-1841	Greece	1951-1952, 1954-1956
Spain	1846-60	Sweden	1821-1835
Germany	1821-60	Angola	1976-1979, 1981-1982, 1986
Austria-Hungary	1821-1840, 1841-1899, 1900-1909	Morocco	1963-64
Austria	1919	Iran	1974, 1977-1979, 1980
Italy	1846-60	China	1936-37

There are also a few data points where specialized estimation procedures were utilized. For example, to calculate the values for Sweden 1821-35, COW1 used the number provided by Woytinsky's five year interval and divided it by 5 to get the average over the 5 years. The States and the years where specialized estimation techniques were utilized appear in Table IRST 3 above.

Problems and Potential Errors

Some might question the project's retention of steel to the present. Steel production is currently declining for some highly developed states, and many scholars argue that it is no longer

a valid indicator of industrial activity. This decline, though, reflects the trend in virtually every industrialized sector of states. The decline of steel production in the United States, for example, closely parallels the decline in automobile production. We think it fair to say that the downward trend primarily characterizes the manufacture of such durable goods and represents the passage from one stage of development (heavy industrialization and consumer durables) to another (computers, information processing, and other "high technology"). Therefore, we are not troubled by our use of steel production as an indicator since it mirrors the more general trend. Our choice of pig iron and steel as indicators of industrial strength is plausible since these materials are both the primary product of the blast furnace and hence the closest thing we can find to raw industrialization. The project has considered shifting to (or adding) materials such as aluminum, or semiconductors, or PCs, but each indicator brings with it its own problems, and such discussions have not been finalized.

Quality Codes

The quality coding scheme is listed in Table IRST 4 below. A data point received the quality code A if the value came from an identified data source such as Mitchell or the Steel Statistical Yearbook. A quality code of a B was given for those data points where a state had no known production capability. If a data point was interpolated by COW2 it received a quality code of C. The quality code D refers to data from the earlier COW data set. This quality code includes values in the earlier data set from which we could not confirm the source of the value, as well as interpolation, extrapolation or other estimation techniques performed by COW1. Finally, a data point receives the quality code M if the value is missing.

Table IRST 4: Iron and Steel Quality Codes

Quality Code	Interpretation
A	Value from identified source
B	No known production—assumed to be zero
C	COW2 interpolation
D	Data from earlier COW data set, but with missing or unidentified source
M	Missing value

Anomaly Codes

In the data set there are places where there is a large increase in iron or steel production from one year to the next. We identified these large increases and created a coding scheme to alert users of the discontinuities in the time series. An anomaly was defined by the project as an increase or decrease in a value from the previous year by at least 100%[5]. We have identified 263

data points where the difference from the previous year was at least 100%. These data points encompass 2% of all data points (13002 total data points).

When there was a difference from year t to t+1 of less than 100% the value at year t+1 was coded as A. A data point was coded as B when the increase occurred as a result of initial industrialization, that is, moving from having no production capability to production. There are 59 data points with this type of anomaly code. A value is coded as C if the difference from the previous year was a result of changing sources. There are 29 of this type of anomaly, most of which occur when we moved from using UN data to Mitchell Data. A value is coded as D if the increase occurred within the same source. There are 175 data points which had internal source inconsistencies. For example, if Mitchell reports a value at year t and there is at least a 100% increase at year t+1, the value at t+1 would be coded as D. Finally, a value is coded as E if we could not find an explanation for the increase. No values received this code. The second year of the anomaly is given the code. For example, if there is an anomaly from year t to year t+1, year t+1 is given the anomaly code.

Table IRST 5: Iron and Steel Anomaly Codes

Anomaly Code	Interpretation
A	No anomaly
B	No known production capability to production (Ex.: Brazil 1924-1925)
C	Changes of sources (Ex.: China 1911-1912)
D	Internal source inconsistency (Ex.: Algeria 1963-64)
E	Unexplained anomaly

Component Data Set Layout

The layout of the Access sub-component data set is found in Table IRST 7 below. The data set contains nine columns. The first and second columns correspond to the COW state abbreviation and COW state number, respectively. The third column is the year of observation. The fourth column contains the value for that year (in thousands of tons), unless the value is missing. Missing values are indicated by -9. The fifth column provides the source of the data point or “See note.” If the column contains “See note,” the note column should be consulted to see how that data point was calculated. The next (sixth) column, “Note,” explains how that data point was obtained (i.e. linear interpolation or COW1 memo). This column is usually empty for data points with a quality code of A. The seventh and eighth columns, respectively, list the data quality and anomaly codes for that value. The ninth and final column lists the version number of this data set.

Table IRST 6: Data Set Layout

Iron & Steel Production								
StateAbb	CCode	Year	IrSt	Source	Note	QCode	Anomaly Code	Version
USA	2	1816	80	Mulhall, Michael. "The Dictionary of Statistics," George Routledge and Sons, Limited, 1892, p. 332.	COW1 memo states that they interpolated 1816-1819 using Mulhall's 1810 (55,000 metric tons) and 1820 (110,000 metric tons) figures.	D	A	3.01

Bibliography

Europa Yearbook. 1986. London: Europa Publications Limited.

Hartmann, Carl. 1861. Der Heutige Standpunkt Des Deutschen Eisenhuttengewerbes, Leipzig: Verlag von Veit und Comp.

Hood, Christopher. 1911. Iron and Steel: Their Production and Manufacture. London: Sir Isaac Pitman & Sons, Ltd.

Hsia, Ronald. 1971. Steel in China: Its Output Behavior, Productivity & Growth Pattern. Wiesbaden: O. Harrassowitz.

Imperial Institute. 1938. The Mineral Industry of the British Empire and Foreign Countries: Statistical Summary. His Majesty's Stationery Office.

International Iron and Steel Institute. 2000. Steel Statistical Yearbook (CD-Rom). Belgium: International Iron and Steel Institute.

International Iron and Steel Institute. Steel Statistical Yearbook 2002. Belgium: International Iron and Steel Institute.

League of Nations. Various years. Statistical Yearbook of the League of Nations. Geneva: Series of League of Nations Publications.

Mitchell, B.R. 1998. International Historical Statistics: Africa, Asia & Oceania 1750-1993. Third Edition. New York, NY: Stockton Press.

Mitchell, B.R. 1998. International Historical Statistics: Europe 1750-1993. Fourth Edition. New York, NY: Stockton Press.

Mitchell, B.R. 1998. International Historical Statistics: The Americas 1750-1993. Fourth Edition. New York, NY: Stockton Press.

Mulhall, Michael. 1892. The Dictionary of Statistics. London: Routledge and Sons, Limited.

Oechelhauser, Wilhelm. 1852. Vergleichende Statistik der eisen-industrie aller lander, und erorterung ihrer okonomischen lage im Zollverein. Berlin: Derlag bon Veit und Comp.

Pakistan Statistics Division. 1990 & 1994. Pakistan Statistical Yearbook. Karachi: The Manager of Publications.

Ruess, Conrad, Emile Koutny, and Leon Tychon. 1960. Le Progrès Economique En Siderurgie: Belgique, Luxembourg, Pays-Bas, 1830-1955. Louvain.

Seiichi, Tohata (ed.). 1966. The Modernization of Japan. Tokyo: Institute of Asian Economic Affairs.

Singer, J. David, with P. Williamson, C. Bradley, D. Jones, and M. Coyne. 1990. National Material Capabilities Dataset: User's Manual. University of Michigan: Correlates of War Project.

Strumlin, S.G. 1955. Istoria Chernoi Metalurgii v SSSR (The History of Ferrous Metallurgy in the USSR). Moscow:

Svennilson, Ingvar. 1954. Growth and Stagnation in the European Economy. Geneva: United Nations Economic Commission for Europe.

Temin, Peter. 1964. Iron and Steel in Nineteenth-Century America: An Economic Inquiry. Cambridge, MA: The M.I.T. Press.

The Middle East and North Africa Yearbook. Various years. London: Europa Publications Limited.

United Nations. Various years. Statistical Yearbook. New York: United Nations.

Woytinsky, Wladimir S. 1925-28. Die Welt in Zahlen. 7 Bd. Berlin: Rudolf Mosse.

Renmei, Nihon Tekko. 1970. "Tokei kara mita Nihon tekkogyo hyakunenkan no ayumi,"

U.S. Bureau of Mines. 1954.

Collection of Modern China's Economic Statistics (translated), 1955

Statistics of China's Steel and Iron Industry (translated), 1985.

Primary Energy Consumption

This section deals with the similarities and differences between this new Primary Energy Consumption (abbreviated PEC) data set and previous versions of this data set[6]. Of the six indicators of national capabilities, this data series underwent the most extensive reconstruction and re-evaluation of previous coding rules. Therefore, this documentation supercedes all previous versions.

What's New in Version 3.0

The energy values contained in the Version 3.0 data set have been recomputed from raw figures. There are seven areas where changes or additions have been made in the basic coding rules as compared to previous versions: 1) assumptions of zero values; 2) conversions of energy commodities into one-thousand metric coal ton equivalents; 3) interpolation; 4) bringing technology to bear; 5) data quality codes; 6) data merging methods; and 7) identifying anomalies. Quality and anomaly codes will be discussed in a separate section.

Assumption of Zero Values. One major difference between previous data sets and version three presented here is a change in coding of developing states. Previous versions of this data set have almost no values of zero. If a state had no PEC, it was always assumed to be missing; for instance, Colombia (COL, 100)[7] has missing data values from its founding in 1831 until 1925. While the data may be missing, it is very possible that there was no industry (and therefore no commercial energy consumption) in this state at that time. Most Central and South American states were almost exclusively agrarian societies well into the Twentieth Century. It is quite possible that they did not experience industrialization until very late in the data presented here. Looking at version 2.1 of this data set as a whole, the extent of this assumption becomes readily apparent. There are only eight data points out of a possible 11,323 that have a value equal to zero. On the other hand, there are 2,815 missing data points.

Assuming that these data points are all missing does not account for pre-industrial periods that most states would seem to possess. It is possible that many states that did not have data available simply did not have industrial energy consumption of any kind. Therefore, it was deemed necessary to change the coding rule and code a 0 in order to reflect pre-industrial societies. A list of states where this applies appears in Table One.

The coding rule used to determine if a state was pre-industrial is as follows: If the first data entry for a given state is 10 or less, then it is assumed that all values before this point are zero. This threshold was chosen because of the data values contained in the Mitchell (1998) volumes. For many states, this is the lowest possible value that a state could have and still be provided data. The states that fell into this category are listed in the Table One, and make up half

of the states in the international system (twenty-six out of fifty-three states) that went through a pre-industrial period in this data set.

Table ENER 1: States with Pre-Industrial Periods

State	Years With Zero Values	State	Years With Zero Values
Afghanistan	1920-1949	Laos	1954-1959
Albania	1914-1925	Liberia	1920-1942
Bolivia	1848-1928	Mauritania	1960-1964
Burundi	1962-1965	Nepal	1920-1953
Dominican Republic	1894-1945	Nicaragua	1900-1948
El Salvador	1875-1950	Panama	1920-1945
Ethiopia	1898-1929	Paraguay	1846-1945
Guatemala	1868-1945	Peru	1839-1898
Haiti	1859-1950	Spain	1816-1830
Honduras	1899-1950	Sri Lanka	1948-1950
Japan	1860-1868	Thailand	1887-1934
Jordan	1946-1956	Venezuela	1841-1884
Korea	1887-1905	Yemen Arab Republic	1926-1948

If a given state had a first data value of more than 10, however, then this assumption is violated and therefore does not apply. It was necessary to apply some other coding rule. If a state's first available data value was more than 10, an industrializing period was computed for that state. We assumed that the state without data developed at a similar rate (in terms of energy consumption per capita) to another state with full data. Using the PEC data for the similar state, in conjunction with the population data for the two states and the first measured data point of the state in question, it was possible to compute a reasonable approximation of the PEC for the state with the missing data. A list of all the states where this technique was utilized, as well as the similar states and the extrapolated periods, appear in Table Two.

Table ENER 2: States with Computed Pre-Industrial Periods

State	Similar State	Extrapolated Years	First Year With Mitchell Data
Argentina	Spain	1841-1886	1887
Brazil	Spain	1836-1900	1901
Chile	Spain	1839-1894	1895
Colombia	Mexico	1891-1921	1922
Costa Rica	Mexico	1924-1949	1950
Cuba	Mexico	1902-1927	1928
Denmark	Germany	1816-1842	1843
Ecuador	Mexico	1900-1924	1925
Greece	Austria-Hungary	1828-1866	1867
Iran	Turkey	1898-1910	1911
Italy	Spain	1833-1860	1861

Mexico	Spain	1838-1890	1891
Mongolia	China	1921-1956	1957
Portugal	Spain	1836-1871	1872
Romania	Austria-Hungary	1878-1881	1882
Russia	Austria-Hungary	1816-1859	1860
Saudi Arabia	Iraq	1933-1936	1937
Sweden	Germany	1816-1839	1840
Switzerland	Germany	1816-1857	1858
Turkey	Austria-Hungary	1816-1897	1898
Uruguay	Mexico	1910-1945	1946
Yugoslavia	Austria-Hungary	1878-1909	1910

There are four exceptions exist to the rules listed above. Morocco (MOR, 600), Tunisia (TUN, 616), and Egypt (EGY, 651) were states early in the time span covered by this data set (1847-1911, 1825-1881, and 1855-1882, respectively) with no available data during their initial existence in the international system. These states were eventually all subsumed by other states for extended periods of time. When the colonial system in Africa broke down, however, these states re-entered the international system with PEC values that were greater than zero. Because these were occupied states, however, it appears safe to assume that their industrialization periods were during occupations, and not during these early independent times. Therefore, we assume that the PEC for these three states is zero during their independence in the 1800s.

The final exception, the Netherlands (NTH, 210), was somewhat more complicated. From 1830 (the first available data point) to 1846 (the secession and independence of Belgium), the Netherlands was assumed to have the same yearly change in PEC as Belgium. Using this yearly change, the data series for the Netherlands was extrapolated backwards. For 1829, the values for Belgium and the Netherlands were added together for 1830, and then an annual growth rate of five percent was assumed. From 1816 to 1828, an annual growth rate of five percent was assumed, and the PEC values were extrapolated over this span. Using the above method produced logically consistent data values for this series.

Conversion into One Thousand Metric Coal-Ton Equivalents. One element that particularly complicated this research was the validation of the conversion formulas used to turn quantities of energy-producing substances into the “coin of the realm.” There appear to be two major methods for converting various energy commodities into thousands of metric coal-ton equivalents—Darmstadter and the UN. In previous versions of this data set, this project relied primarily on Darmstadter for the conversion formulas. The reasoning behind this was that Darmstadter was the primary source for a majority of data points. As this project continues to grow and evolve by adding more data points computed by UN techniques, however, this reasoning becomes less valid.

In order to correct for this, version three of the data set adheres to UN standards. For this reason, there have been some small changes to the conversion factors (which will be discussed

in greater detail in the respective commodity sections) that may alter the final computed energy consumption from version 2.1 to the version presented here.

Interpolation. In the original version of the data set, most interpolation was done using the total energy consumption of a given state. This stemmed from the notion that Darmstadter (the original raw data source for earlier versions of this data set) would often report total energy consumption for a state, already converted into one thousand metric coal-ton equivalents. However, this source would only list data points intermittently, leaving out certain spans of data values. For instance, data for the United States (USA, 002) was available for 1925, 1929, 1933, 1937, 1938, 1950, 1953, 1955, 1957, 1960, 1961, 1962, 1963, 1964, and 1965 (Darmstadter et al, p. 225). All other data points (particularly the war years between 1941 and 1945) were not available from this source. In order to obtain data points for the missing years, previous researchers would have to interpolate the total energy consumption.

The Mitchell data, however, made it possible to avoid doing dramatic interpolations. In data points assembled using Mitchell data, any necessary interpolations were calculated using individual commodity data (i.e., coal, petroleum, etc.).

Whenever an interpolation was performed in the Mitchell data period, it was computed using Log-Linear Interpolation (abbreviated LLI). These interpolations assume a logarithmic growth rate, and are computed using the following formula:

Equation ENER 2: Logarithmic Growth Rate Computation Formula

$$Rate = \exp\left(\frac{\ln X_{n+t} - \ln X_{n}}{t}\right),$$

where X_{n+t} and X_n are the known starting and finishing points of the range of values to be interpolated, and *t* is the number of data points to be interpolated.

This rate is then multiplied through for all points as shown in Equation ENER 3 below:

Equation ENER 3: Interpolation of Data Points Using Logarithmic Growth Rate

$$X_{n+1} = Rate * X_n \; ; \; X_{n+2} = Rate * X_{n+1} \; ; \; \ldots\ldots \; ; \; X_{n+t} = Rate * X_{n+t-1}$$

Bringing Technology to Bear. The original Correlates of War energy consumption data set relied on individual paper computation sheets for assimilating much of the data. Each data point originally consisted of a computation page that listed raw data values, sources, conversion calculations, and total computed energy consumption. Overall, there should have been almost 12,000 of these computation sheets. Unfortunately, over time these sheets were lost and only a few dozen remain. Therefore, it was necessary to re-compute every data point from scratch, including documentation and computation.

To recreate this data set using these previous technologies would be impossible. However, due to advances in computer technology and computing strength (particularly in spreadsheet and scanning technology) completely re-creating this data set was possible. These new technologies were fully utilized for this project. Raw data sources were scanned in from their source books into computer-readable tables. From these raw data sources, a Microsoft Excel worksheet was constructed for each state from 1816 or its inception until 1970. These workbooks contained five pages—one for each of the four energy commodities, and one for the total energy consumption of the state in question. The data cells in each of these workbooks are fully linked together, in order to make updating data simpler. Each workbook page contains data points, source listings, conversion factors, any necessary interpolations, and documentation and discussion of individual problems. After 1970, the data came from the UN and was already converted into one thousand metric coal ton equivalents.

Data Merging Methods. One potential problem area in version 2.1 was where previous researchers merged the UN data together with data from other various sources. In version 2.1, UN data were used for every state only after 1970. Literally, every state in the international system changed conversion formulas and data sources at exactly the same point. The authors of the User's Manual wrote: "The slight difference in conversion methods introduced discontinuities from one year to the next in coal-ton energy values" (Singer et al, p. 30). Having every state in the international system change data source and conversion methods at once created a potentially large discontinuity in the data, making examinations over time much more difficult.

This version of the energy consumption data attempts to correct for this potential bias. The most recent UN data covered every state in the international system with very little missing data starting in either 1968 or 1970. Some states (particularly major ones such as the United States, Soviet Union, Western Europe, and Japan) had UN energy consumption data starting in 1950. With this in mind, energy consumption data were computed from the Mitchell volumes for all states from either 1816 or their inception until 1970. These two data sources were then merged. If there were both UN and Mitchell data values, the UN data values were utilized for the data point. This merging method will hopefully smooth out some of this discontinuity contained in previous versions.

Data Acquisition and Generation

Primary Energy Consumption measures one element of the industrial capacity of states in the international system. Simply put, the greater the energy consumption, the larger the potential manufacturing base of an economy, the larger the potential economy of the state in question, and the more wealth and potential influence that state could or should have. PEC is a derived indicator, computed using Equation One below:

Equation ENER 1: Primary Energy Consumption Formula

Consumption = Production + Imports – Exports – Δ in Domestic Stocks

This formula is quite similar to the one utilized in the original coding manual, except for one change—the inclusion of domestic stocks into the equation (Singer et al, p. 21). This reflects that states will maintain supplies of energy-producing commodities in the event that there are disruptions of import or export flows.

Primary Energy Consumption comes from (and is computed using data about) four broad categories of sources—coal, petroleum, electricity, and natural gas. Each of these elements is broken into a variety of different elements. It is important to note that these forms of energy are all types of commercial energy. Many other forms (such as animal waste, peat, and wood-burning) exist, however these other energy sources are of such small amounts that they do not qualify as industrial energy sources. The raw data for each commodity is converted into a common unit (in this case, one thousand metric ton coal equivalents) and then summed to produce the energy consumption for a given state in a particular year.

The data series runs from 1816 (when the Correlates of War project begins to track the international system) until 1998 (the last year the United Nations publishes comparable, cross-national data on energy consumption). Data on these commodities comes primarily from two sources. For the pre-1970 portion of the data, much of the data necessary to compute PEC comes from the Mitchell International Historical Statistics series. After 1970, the data come from the Energy Statistics Yearbook published by the United Nations. This is a change from previous data sets. Older versions of the data set obtained much of the PEC data during the pre-1970 period through state-specific sources, and not a single, common source. This made tracing the source of many of the original data points impossible. In this version, however, there are far more points that come from only a few sources instead of an amalgamation.

United Nations Data. This data source was utilized for all states whenever possible. Overall, the UN began collecting PEC data for some states (particularly the United States, Western Europe, Soviet Union, China, Japan, and Australia) starting in 1950. Comprehensive data on all the states in the international system only began between 1968 and 19708.

The United Nations data arrives already converted into one thousand metric coal-ton equivalents. However, Mitchell data were disaggregated into four major commodities (coal, petroleum, electricity, and natural gas); UN data is aggregated into four major categories: production, imports, exports, and changes in domestic stocks (in accordance with Equation One above). This required a different combination scheme. Simply put, Equation One was applied to the UN data to calculate PEC. However, there were a number of blank cells contained within the data that had to be addressed in order to make the calculations. The assumption that was used

for the UN data only was that if the data were missing, the value was zero. These entries are contained in Table ENER 3 below.

Table ENER 3: UN Data Codes

UN Code	Present Data	Missing Data (Assumed to be 0)	*N*
1	None	All Data	0
2	Stock Change	Production, Imports, Exports	0
3	Exports	Production, Imports, Stock Change	0
4	Exports, Stock Changes	Production, Imports	0
5	Imports	Production, Exports, Stock Change	0
6	Imports	Production, Exports	0
7	Imports, Exports	Production, Stock Change	0
8	Imports, Exports, Stock Change	Production	0
9	Production	Imports, Exports, Stock Change	543
10	Production, Stock Change	Imports, Exports	58
11	Production, Exports	Imports, Stock Change	141
12	Production, Exports, Stock Change	Imports	191
13	Production, Imports	Exports, Stock Change	587
14	Production, Imports, Stock Change	Exports	506
15	Production, Imports, Exports	Stock Change	1030
16	All Data	None	2771

Note: UN data codes 1-8 are included here in order to account for the future possibility that certain state's data values may be missing.

Using this technique, there were no negative data values produced. As is also apparent, there are no data cells where all the information is missing; there are some values that can be calculated for each state in the international system.

Mitchell Data. Primary Energy Consumption computed using Mitchell data is comprised of four energy-producing commodities: 1) Coal; 2) Petroleum; 3) Electricity; and 4) Natural Gas. This section will discuss each of these commodities, looking at a brief history, conversion formulas, and potential problems found within each commodity.

Coal. Of all the industrial indicators, coal is the only indicator that covers the entire time span from 1816 to the present. Coal is the primary energy consumption element for all states prior to World War One. It is also the metric standard by which all this energy consumption data is measured.

In this data collection effort, three types of coal were identified: Anthracite, Bituminous, and Brown. Anthracite and Bituminous are very similar; they are the hard, black coal found in most mines throughout the world9. These two types of coal are the standard by which all other energy consumption elements are measured.

Brown coal, on the other hand, is softer, quicker burning, and less efficient as an industrial fuel. There are a variety of different types of brown coals (a type called lignite is mentioned most often), and their quality is often dependent on where a state is located in the

world. In order to account for these differences, this data set utilized a state-by-state brown coal conversion table. These conversion values appear as Table ENER 4.

Some similar conversion values appeared in previous versions of the coder's manual (Singer et al, Table Three, p. 28). There are some differences between that table and the one presented by Darmstadter. We choose to utilize the table as presented by Darmstadter. One potential problem arose in these brown coal conversions. There were three cases where there was no brown coal conversion presented for a given state, even though the Mitchell data documented that the state in question produced brown coal. These states are Hungary (HUN, 310), Iran (IRN, 630), and Mongolia (MON, 712). For these three states, this computation utilized the conversion factor for a state on the list that is geographically proximate to the state in question. These proximate states were Austria (AUS, 305), Turkey (TUR, 640), and North Korea (PRK, 731), respectively.

Table ENER 4: Brown Coal Conversion Values for Given States[10]

State	Conversion	State	Conversion
Thailand	0.7	Netherlands	0.33
Canada	0.65	Tunisia	0.33
Czechoslovakia	0.6	Turkey	0.33
France	0.6	United States	0.33
Hungary	0.6	Germany, West	0.31
Romania	0.6	Bulgaria	0.3
Albania	0.5	Germany, East	0.3
Austria	0.5	India	0.3
Greece	0.5	Indo-China	0.3
Japan	0.5	Italy	0.3
New Zealand	0.5	Korea, North	0.3
Portugal	0.5	Korea, South	0.3
Spain	0.5	Poland	0.3
Yugoslavia	0.5	Denmark	0.29
Chile	0.33	Australia	0.25

Petroleum. Petroleum is the second most prevalent source of industrial energy consumption. Relatively speaking, petroleum products were a minor source of commercial energy until the advent of the automobile after the turn of the century. Since then, however, petroleum has become a highly important industrial energy source.

In generating usable data from the raw petroleum figures presented, it is often necessary to perform two types of conversions. First, it is necessary to convert the raw data into metric tons. Second, it is then necessary to convert data from the metric ton of petroleum into metric tons of coal equivalency. I will look at each of these in turn.

There were two alternate measures used throughout the Mitchell Data. They were "One Million US Gallons" and "One Thousand Barrels." The conversions from their respective units into metric ton equivalencies for each of these measures are listed in Equations ENER 4 and ENER 5 below[11]:

Equation ENER 4: Millions of US Gallons to Thousands of Metric Tons of Oil[12]
____M US Gallons * 3.2468 = ____K Metric Tons of Oil

Equation ENER 5: Thousands of Barrels to Thousands of Metric Tons of Oil[13]
____K Barrels * 0.1366 = ____K Metric Tons of Oil

In converting petroleum into coal-ton equivalents, Mitchell distinguished between two major forms of petroleum: 1) Crude, and 2) Refined. The conversions for each of these two types of oil are listed below:

Equation ENER 6: Crude Petroleum to Coal Ton Equivalents[14]
____K Metric Tons Crude Petroleum * 1.429 = ____K Coal Ton Equivalent

Equation ENER 7: Refined Petroleum to Coal Ton Equivalents[15]
____K Metric Tons Refined Petroleum * 1.474 = ____K Coal-Ton Equivalent

One of the difficulties of converting petroleum products into coal equivalents is that petroleum products come in a variety of different weights and types, each of which has its own conversion value. Unfortunately, Mitchell does not distinguish between the many different weights and types of petroleum products—this source only utilizes the crude and refined categories listed above. In order to overcome this problem, it was necessary to make some assumptions about the conversion formulas utilized here. For crude oil, the conversion factor utilized here is the conversion for crude oil of average viscosity[16]. For refined products, the conversion stems from two considerations. First, this value is the conversion for kerosene, the major refined petroleum product prior to World War One. Second, it is also the approximate mean value for all refined petroleum products that have been produced following World War One. For instance, gasoline and liquefied petroleum gases both have higher conversion factors to coal-ton equivalents than kerosene (1.500 and 1.554, respectively, as compared to 1.474 for kerosene; taken from Energy Statistics Yearbook, p. xlv). However, gas-diesel oils and residual fuel oil have lower conversion formulas than kerosene (1.450 and 1.416, respectively; taken from Energy Statistics Yearbook, p. xlv). Because some types of refined petroleum products have greater conversion factors and

others have smaller conversion factors, this conversion value appears to be a good approximation for all types of refined petroleum products.

It is vital to note, however, that in the UN data this assumption is unnecessary. The United Nations collects data on each individual commodity, in its original form, and makes its calculations based on the values of these individual characteristics instead of on an assumption about homogeneity. Therefore, these assumed values above are only concerned with the Mitchell Data calculated before 1970.

Electricity. After 1900, electrical production enters the world energy picture. It is important to note, however, that the electrical production values listed here DO NOT include electricity produced from fossil fuels; these types of energy production are included in the coal, oil, and natural gas components of the data. Electrical production here includes three types of electrical production: 1) Hydroelectric, 2) Nuclear, and 3) Geothermal.

Conversion from electrical energy production into coal-ton equivalents utilizes the following formula:

Equation ENER 8: Electrical Energy Conversion

____ Giggawatts * 0.123 = ____ K Coal-Ton Equivalents

Mitchell's aggregation of raw data again makes assuming necessary, which is central to this conversion factor. Mitchell does not distinguish between hydroelectric, nuclear, or geothermal energy. He aggregates these three vastly different categories into one category. Therefore, it was again necessary to assume about the type of energy that was produced prior to 1970.

The assumption utilized here (and in the conversion above) is that all electricity before 1970 is hydroelectric power. Prior to 1970, this assumption is quite tenable—before World War Two, nuclear and geothermic electricity did not exist. After World War Two, the nuclear reactor was only becoming commercially available and viable in the early 1960s[17], and only by 1970 were there enough nuclear plants to make any measurable contribution to energy production. Only after the oil shocks of the 1970s (when this data set utilizes UN data that separates conversion rates for each type of electricity) did research and utilization of these alternate forms of electrical generation step into high gear.

The potential biasing impact of this assumption is diminished again because of the prevalence of UN data. The UN again distinguishes between all three of the aforementioned electricity types, converting each according to its own conversion factor. Therefore, in states where nuclear power is prevalent, such as the United States, Western Europe, Soviet Union, China, and Japan, their data come from the UN beginning in 1950, making this assumption not apply to these states in question.

One difference between the version presented here and version 2.1 is a change in the conversion rates utilized. Version 2.1 utilized a conversion rate that evolved from 1.0 in 1919 to 0.3 in 1971 (Singer et al, 1990, p. 28; originally published in Darmstadter 1971, p. 830[18]). The original researchers believed that there was an evolution of electric-producing technology, making more efficient electrical production possible over time, necessitating a moving scale. The UN, however, rejects this conversion and utilizes fixed conversion factors. Because the UN data is utilized for computing far more data points than anything in the Darmstadter era, version three utilizes UN data conversion techniques, which for electrical consumption utilizes a constant conversion rate.

Natural Gas. Natural gas production was the last of the four energy commodities to appear on the industrial scene. Present for as long as there were petroleum production facilities, natural gas was often burned off at the site, instead of being used for more commercial purposes. Only in the last fifty years has the condensation, refrigeration, and storage technology been available to harness this source of energy for commercial purposes.

The conversion formulas for computing natural gas production into coal ton equivalents appear as Formulas ENER 9 and ENER 10 below.

Formula ENER 9: Cubic Meters of Natural Gas to Coal Ton Equivalents
____M Cubic Meters Natural Gas * 1.33 = ____K Coal-Ton Equivalents

Formula ENER 10: Petajoules of Natural Gas to Coal Ton Equivalents
____Petajoules * 34.121 = ____K Coal-Ton Equivalents

For most of the data points, the “million cubic meters” is the standard unit for natural gas production. The Petajoule became the basic unit of natural gas production in 1966, necessitating a different conversion formula.

Problems and Potential Errors

In calculating these revised energy consumption data values, certain problems arose. This section will describe and list four of the more important problems: 1) Negative Values, 2) Multiple Data Values, 3) Missing Data, and 4) Un-Documented Data Points.

Negative Values. One of the problems with generating historical energy consumption data were that in a small number of cases (twelve to be precise) the computations of energy consumption produced a negative value for energy consumption. These data points are listed in Table ENER 6.

Two issues surrounding the data available on one particular commodity—petroleum—contributes to this problem of negative PEC. Looking at Equation ENER 1 again the formula for calculating PEC below can also help illuminate these potential biases.

PEC = Production + Imports – Exports – Change in Domestic Stocks

The "Change in Domestic Stocks" portion of the equation above is the first avenue that can create this problem of apparent negative PEC. Simply put, there are not historical records of domestic stockpiles of energy commodities, such as coal and petroleum. Oil-producing countries often have domestic stockpiles of petroleum. Much like the United States' "Strategic Oil Reserve," most states keep some sort of stockpiles of petroleum in case of shortages, embargoes or other possible disruptions that can stop or reduce the flow of oil into or out of a state. In petroleum-producing states, however, these stockpiles can be massive. In lower-production (or higher demand) years, these states would often export from these domestic stocks, while keeping production low. The UN was the first source to begin gathering compete domestic stock data in the 1970s, and without being able to account for these stocks, a state could easily appear to export more oil than it produced, creating the problem of apparent negative energy consumption. Without some sort of entry into this variable, an omitted variable bias is created in the above equation, making it appear that a state had negative PEC.

The production portion of the equation above is the second issue that drives this problem of apparent negative PEC. As a policy, OPEC monitors and attempts to manage petroleum production by looking at production values (www.opec.org) and setting quotas based on these production values, while not appearing to examine import or export amounts. Therefore, a state that would want to break its quota would simply falsify its production amounts by reporting less oil production than they truly produced while maintaining their accurate import and export figures. This would again result in a negative value for energy consumption.

Negative energy consumption data values were corrected by altering the production of crude oil. These corrections appear in the "Correction" column of Table Five above. For the most part, the domestic crude oil productions were inflated between one and ten percent. The exact amount was determined by looking at the data points surrounding the negative value, using a common-sense approach. The only exception to this was Iran, where some data values were shifted from one year to another to account for what appeared to be oil produced in one year and exported in another.

Table ENER 5: Negative PEC Data Points and Their Corrections

State	Year	Original PEC	Adjusted PEC	Adjustments Made
Mexico	1922	-263	1603	Petrol. Production Increased by 5%

Venezuela	1930	-568	869	Petrol. Production Increased by 5%
Venezuela	1931	-370	861	Petrol. Production Increased by 5%
Venezuela	1936	-180	1436	Petrol. Production Increased by 5%
Gabon	1963	-78	49	Petrol. Production Increased by 10%
Gabon	1964	-10	66	Petrol. Production Increased by 5%
Gabon	1965	-19	71	Petrol. Production Increased by 5%
Iran	1911	-1	11	No Petrol. Production Value--LLI from 1910 (0) to 1912 (80)
Iran	1919	-278	151	Moved 300 1K MT Oil Production from 1915 to 1919 in order to smooth curve
Iran	1920	-1011	319	No Mathematical Correction Possible Assumed Petrol. Production Missing LLI Production from 1919 to 1921
Iran	1933	-545	375	Moved 644 1K MT Petrol. Production from 1931 to 1933 in order to smooth curve
Iraq	1948	-34	210	Petrol. Production Increased by 5%

Multiple Data Values. Throughout the Mitchell (1998) data, there are a number of points where this source lists two data points for a state in a given year[19]. Often, these come from some sort of change in reporting, which can take place in a variety of different ways. There could have been some change in accounting procedure that generates two data points; for instance, many states changed their accounting procedures from using calendar years to fiscal years or vice versa. Two data points could be generated if a new region became included into a state. There could be changes of measurement units, for instance moving from Millions of Cubic Meters to Petajoules of Natural Gas Production. Generally, the procedure for handling this potential problem was to average the two values and assign a data quality value of "B." This tended to create a smoother time series picture of the change in a given commodity over time.

Missing Data. There was a problem of missing data upon the completion of the major data recreation utilizing Mitchell (1998) as a source. Nineteen states, often due to brief or early existences in the international system, had data values in version 2.1 of the data set but did not have data values available through Mitchell (1998). Therefore, it was necessary to perform some sort of estimation of these phantom data points.

The technique utilized to estimate fourteen of these problematic data series is called population-based energy consumption estimation. This technique involves three steps. First, a state that is geographically proximate and industrially similar to the state with missing data is identified[20]. Second, energy consumption per capita (that is, energy consumption divided by total population) was computed for the neighboring state with documented data. Third, the yearly energy consumption per capita values from this similar state was multiplied by the population data for the state with missing energy consumption data. This produces an estimate of what the energy consumption would be for that state.

Table ENER 6 contains a list of states whose industrial energy consumptions were computed in this manner. It also lists the proxy states whose energy consumptions were utilized in the calculations presented above, as well as the years that these calculations were performed and the number of data points generated in this manner.

Table ENER 6: States with Population-Based PEC Estimations

States With Missing Data Points	Similar States Utilized for Estimations	Estimation Span	Number of Data Points
Luxemburg	Belgium	All Data Points before 1970	48
Estonia	Poland	All Data Points before 1970	23
Latvia	Poland	All Data Points before 1970	23
Lithuania	Poland	All Data Points before 1970	23
Saxony	Germany	1850 to 1867	18
Hanover	Germany	1838 to 1866	29
Bavaria	Germany	1816 to 1871	56
Hesse Electoral	Germany	1816 to 1866	51
Cyprus	Greece	All Data Points before 1970	11
Malta	Italy	All Data Points before 1970	7
Equatorial Guinea	Cameroon	All Data Points before 1970	3
Gambia	Senegal	All Data Points before 1970	6
Zanzibar	Tanzania	All Data Points before 1970	2
Maldive Islands	Sri Lanka	All Data Points before 1970	6

The final three undocumentable states—Hanover (HAN, 240), Bavaria (BAV, 245), and Hesse Electoral (HSE, 273)—all possessed a very unusual data pattern. These three states only had one data point each in the original COW PEC data set—1853. Every other data point for each of these three states both before and after 1853 was missing. Somehow, some researcher found that one value for these three states. Unfortunately, that one data source for that one point cannot now be identified. First attempts to use population-based energy consumption estimates produced data figures that were far too high to be realistic, especially once these states amalgamated into Germany proper. Therefore, this technique was dismissed. It was apparent that some other technique was necessary for estimating these unusual data points.

The following equation was utilized in order to make an educated estimation about the data values for these three states:

Equation ENER 11: PEC Estimates for Hanover, Bavaria, and Hesse Electoral

$$PEC_X = \frac{\frac{PEC_{X-1}}{Population_{X-1}}}{\frac{PEC_{x-1}}{German}} * \frac{German\ PEC_x}{German\ Population_x} * Population_X$$

This formula rests on three thoughts. First, it follows the industrial growth rate of Germany for the same time span. Second, it anchors these three data points to a value that some researcher was able to find and document (1853, even though it is undocumented as of this writing). Third, it also centers on the population growth rate of these three states (which is fully documented in version three of the data set. This technique has produced data values for these three states than seem fairly reasonable. Future research should focus on finding more exact measures for these states.

Quality Codes

One of the realizations in creating this version of the PEC data set was that there were numerous methods used to calculate data points in both previous and current versions of the primary energy consumption data set. Some data points are compiled using very precise data points, gathered for a state in a given year. Sometimes it was necessary to extrapolate or interpolate particular commodities. In other instances, it was necessary to make estimations about the energy consumption of a state with little available data. The quality codes for this data series reflect these situations.

Table ENER 7: Primary Energy Consumption Quality Codes

Quality Code	Substantive Interpretation
A	All Components Present; or, only electricity interpolated from 1900-1945
B	All Components known, but averaged. Often happens when a state changes reporting units (for example, moving from calendar years to fiscal years or vice-versa).
C	Some (but not all) component data points interpolated
D	All component data points interpolated (Example: China during the Boxer Rebellion
E	Log Linear Extrapolation based on growth rates (Example: Mexico before 1981)
M	Missing Data Values (Example: Lesotho, Papal States)

Anomaly Codes

In many data sets, there are often discontinuities that exist in the data. A state's international trade will suddenly increase by 400% in one year.

In version 3.0, this project identified data points that appear to create discontinuities in the data values. These discontinuities can wreak havoc during analysis; if a researcher is using time series analyses, running an analysis across these anomalies will create estimation problems that can lead to Type I or Type II error during analysis.

For energy consumption, we considered an anomaly was defined as an increase or decrease in total primary energy consumption that was as least 100% from its previous value[21].

Table Energy 8: Total Population Anomaly Codes

Code	*Substantive Meaning*
A	No Anomaly (< 2% change)
B	Explained Inconsistency (e.g. change in territory, loss in wartime)
C	Change of Sources (between 2 non-UN sources or 1 non-UN to UN source)
D	Change of UN Sources
E	UN Internal Inconsistency within same UN source
F	Internal inconsistency within non-UN source
G	Unexplained Anomaly

Component Data Set Layout

The layout of the Access sub-component data set is found in Table ENER 9 below. The data set contains eight columns. The first and second columns correspond to the COW state number and COW state abbreviation, respectively. The third column is the year of observation. The fourth column contains the value for that year (in thousands of coal-ton equivalents), unless the value is missing. Missing values are indicated by -9. The fifth column provides the source of the data point or "See note." If the column contains "See note," the note column should be consulted to see how that data point was calculated. The next (sixth) column, "Note," explains how that data point was obtained (i.e. linear interpolation or extrapolation). This column is usually empty for data points with a quality code of A. The seventh and eighth columns, respectively, list the data anomaly and quality codes for that value.

Table ENER 9: Data Set Layout

PE Consumption								
CCode	**State**	**Year**	**Energy**	**Source**	**Note**	**Anomaly Code**	**QCode**	**Version**
2	USA	1816	254	B.R. Mitchel, International Historical Statistics: the Americas, 1750-1993.	Derived from production, export, and import values of coal, petroleum, natural gas, and electricity.	A	A	3.01

Bibliography

One of the improvements of version three of the industrial energy consumption data set and previous research efforts is that there are far fewer sources that were utilized for gathering data. The sources are listed below in annotated bibliographic form.

"Australia's HDR Resources." 2000. The University of New South Wales School of Petroleum Engineering Website.
URL: http://www.petrol.unsw.edu.au/research/resource.html

This web site was invaluable in determining what a "petajoule" both is theoretically and mathematically. The UN conversions and documentation deal in terajoules, not petajoules. Without this source, the conversion could not have been performed.

CIA World Factbook. 2002. Published by the Central Intelligence Agency.
URL: http://www.cia.gov/cia/publications/factbook/index.html

This web-based source was important in extending some of the data until 1997, because it provided a measure of total energy consumption when other sources did not.

Darmstadter, Joel, Perry D. Teiterbaum and Jaroslav G. Polach. 1971. *Energy in the World Economy: A Statistical Review of Trends in Output, Trade, and Consumption Since 1925.* Washington, DC: Johns Hopkins Press.

This source was utilized primarily for the brown coal conversion values. However, for two states (Albania and Iceland) this source was also utilized for primary energy consumption data.

Energy Statistics Yearbook (United Nations. Statistical Office). 1997. New York: United Nations Press.

This volume was the source for all the conversion formulas utilized throughout this research. It was also the data source for some states from 1950 until 1970 and for all states starting around 1970. The UN only publishes data on energy consumption with a four-year lag, therefore their data collection (and the scope of this project) ends in 1997.

Mitchell, B.R. 1998. International Historical Statistics: The Americas 1750-1993. Fourth Edition. New York, New York: Stockton Press.

Mitchell, B.R. 1998. International Historical Statistics: Africa, Asia, & Oceania 1750-1993. Third Edition. New York, New York: Stockton Press.

Mitchell, B.R. 1998. International Historical Statistics: Europe 1750-1993. Fourth Edition. New York, New York: Stockton Press.

These three volumes contain international historical statistics on most states in the international system from 1816 until approximately 1993. They were the major source of raw energy commodity data for all states in the international system.

Singer, J. David, with Contributions from P. Williamson, C. Bradley, D. Jones, and M. Coyne. May 1, 1990. "National Material Capabilities Dataset: User's Manual." Correlates of War Project: The University of Michigan.

---

[1] Some data sets, such as alliances or contiguity, do not have this sort of consideration. Others, such as interstate trade or foreign direct investment, are the type that is being addressed in this discussion.

[2] It is important to note that 1993 was part of version 2.1 of the data set. It was necessary to update this data value as well. Because the last revision was in 1992-1993, many of these data points were either estimates or missing. Therefore, we were able to go back and enter non-estimated data values for many of these previously troublesome data points.

[3] *De facto* information or data is information that is taken with surveys, censuses, and other forms of direct counting. *De jure* data is best described as data that comes from historian's impressions or estimates of the population of an urban center at the time of their writing.

[4] The coding rules and procedures are largely taken from the 1990 coding manual. This manual also includes a discussion of the theoretical relationship between iron and steel production and national power.

[5] In order to determine whether 100% was the appropriate threshold, a stratified random sample was conducted with ten states using thresholds of 25 and 50%. There was not a large difference in the number of anomalies identified at these thresholds.

[6] Unless otherwise stated, anything not discussed in this coding manual should be assumed to remain the same as in the original coding manual (Singer et al, 1990).

[7] The letters and numbers appearing in parentheses are the COW abbreviation and country numbers for the states in question throughout the rest of this article.

[8] There appears to be no reasoning or pattern as to when the UN began collecting this data.

[9] The only state where there was a distinction made between anthracite and bituminous coal was the United States (USA, 002). Assuming that these two types were the same yielded an interesting result—43 out of 44 data points for this state from 1816 to 1859 were *identical* to those contained in version 2.1 of the PEC data set.

[10] Reproduced from Darmstadter, p. 828.

[11] Notation used from here on is K=1,000 and M=1,000,000. Cross-cancellations supporting these conversion factors appear in Appendix Two at the end of this document.

[12] Energy Statistics Yearbook, p. xlix.

[13] Ibid.

[14] Energy Statistics Yearbook, p. xlv.

[15] Ibid.

[16] Much like brown coal, crude petroleum is not the same everywhere on the planet. However, these differences have not been documented or mathematically differentiated as well as in the brown coal case.

[17] http://geocities.com/RainForest/Andes/6180/history.html#top

[18] It would appear that the original project researchers interpolated conversion values from 1965 to 1971.

[19] Specific numbers are not available, however as a "best guess," probably every state in the international system between 1816 and 1970 has at least one individual commodity data point with two values for the same year.

[20] Needless to say, the state selected must have energy consumption data available.
[21] We also performed these tests at thresholds of 50% and 20% for a sample of 10 states spread through all the regions of the world, and found that there was very little change.