Historical International Macro-Finance Data Sources

Chenzi Xu, Stanford

December 2021

Agenda

Readily available data

1. Trade
2. Capital Flows
3. Crises, disasters, major events
4. Macro-data across countries

Creating your own datasets

Historical (international) papers

These slides with active **links** available on my website: **chenzi-xu.com**

Historical perspective

"GOOD DATA"

1850-1914: First age of Globalization	1914-1970: WWI, Interwar, WWII, Bretton Woods	1970-present: Second age of Globalization

- **First globalization:** huge increases in goods & capital integration around the world
  - Led by UK on pound sterling/gold standard
  - Technologies like steam ships, railroads, telegraph helped to integrate information & real markets
- **WWI–Bretton Woods:** collapse in international trade & private capital flows
  - Transition from £ to $
- **Second globalization:** broadly very similar to first globalization
  - Led by US $ with (mostly) freely floating exchange rates
  - ICT revolution, GATT, WTO, regional trade agreements

"Good data" period: **25%** of years since First globalization—we have a lot to learn from history!

Readily available data

Types of data

1. Trade

1. Trade data overview

A. **Bilateral country-level aggregate trade panels:** most complete

B. **Bilateral country-industry trade panels:** a lot of work in progress

C. **Other measures:**

1A. Country-level bilateral panels

Ultimate sources: government reports, either compiled into Statistical Abstracts (done by the UK, US, France, etc) or by individual contry. These often contain lots of undigitized, unexplored data (!)

1. **TRADHIST database 1827–2014:** (link)
   - Data structure: contemporary borders/entities, values standardized to nominal £, other country characteristics such as GDP and bilateral connections like colonial ties

2. **RICardo project 1800–1938:** (link)
   - Data structure: imports and exports; nominal values in different currencies; thousands of geographical entities (directly copied from publications)

3. **Pascali (AER 2017) 1850–1900:** (link)
   - Data structure: imports and exports; standardized to nominal £; different measures of distance by transport technology

Constructing a comprehensive country-level panel

No single source has comprehensive coverage & each source has unique data

- Overlapping source material but often differences in exchange rates chosen, borders, observations included, other covariates, etc.
- Europe & North America generally well-covered; big differences are in coverage of the ROW

Researcher decisions:

- Country entities/borders:
  - Disaggregated entities example: provinces of Australia reported separately
  - Aggregated entities example: Norway-Sweden combined
- Conflicts:
  - Known issue where often exports from A $\rightarrow$ B $\neq$ imports in B $\leftarrow$ A
  - Sources do not always agree, even before standardizing borders
- Constructing data:
  - Missing bilateral resistance measures need to be created
  - Inverting the imports data to create additional exports data

Xu (2019) deals with these issues forglobal trade from 1850–1914 at pre-WWI borders (link)

1B. Within-country bilateral panels: industry composition

Various researchers have constructed industry-level bilateral panels of trade by country. There has been no systematic effort to collect these data for the world.

- **Italy** 1862–1950 (imports & exports at 4-digit SITC, annual): Bank of Italy (link)
- **UK** 1700–1899 (exports at 2-digit SITC, annual): Jacks O'Rourke Taylor (2020) (link)
- **USA** 1866–1914 (imports & exports at 5-digit SITC, annual): completed by Xu & Meissner
- **Belgium** 1870–1910 (manufactured imports & exports at 3-digit SITC, every 5 years): Huberman Meissner Oosterlinck (2017) (link)
- **Germany** 1880–1913 (imports & exports at 5-digit SITC, annual): Hungerland Wolf (2021) (link)
- **Japan** 1880–1910 (exports at 5-digit SITC, every 5 years): Meissner Tang (2018) (link)

Norm is to use SITC revision 2 for better coverage of historical products (link)

1C. Other measures

- **Federico-Tena database 1800–1938:** (link)
  - Data: aggregate country-level trade NOT bilateral
  - Unique for: product composition (manufactured vs commodities) for total exports
- **Lloyd's List 1700s–today**
  - Data: Juhasz (2018) and Xu (2019) digitize parts. Otherwise available as scans of original newspapers:
    - link to scans from 1741–1800s
    - link to scans from 1800–1912
  - Unique for: daily, within-country trade flows + weather conditions, marine casualties

Major changes in trade costs:

- **Sail → Steam:** Pascali (2017) estimates sailing times using weather & current patterns; validates with historical log books
- **Canals:** historical maritime guides print port-to-port matrices of travel times using routes including/excluding canals
- **Telegraph:** Steinwender (2019) & Juhasz Steinwender (2019) provide historical telegraph connection dates

Types of data

1. Trade
2. Capital flows

2A. Bonds & equities prices

Main historical international financial centers: London, Paris, New York

All databases from Yale's International Center for Finance unless otherwise noted

- **London 1869–1930:** *Investor's Monthly Manual* (IMM) monthly records of prices, dividends for all bonds/stocks (link)
  - Source for sovereign & international corporate debt/equity (and domestic UK securities)
  - Before 1869: *Course of the Exchange* (CoE)–similar source to IMM that is not digitized
- **Paris 1795–1976**: DFIH database from PSE (link)
  - Work in progress; data access available by request + some downloadable–much more comprehensive than IMM (spot, forward, option, repo prices as well)
- **New York, 1815–1915**: monthly NYSE prices; daily afterward from CRSP (link)
- **Shanghai, 1870–1940**: annual, abolished afterward (link)
- **St. Petersburg, 1865–1914**: monthly, abolished afterward (link)
- **Amsterdam, 1796–1980**: in progress (link)

Non-digitized sources: for higher frequencies, need to return to original sources. Subsets have been digitized by different scholars but nothing systematic because of the scale

- Scans accessible through Gale newspaper databases; bulk purchases from libraries also possible and not too expensive

2B. Sovereign & bank debt

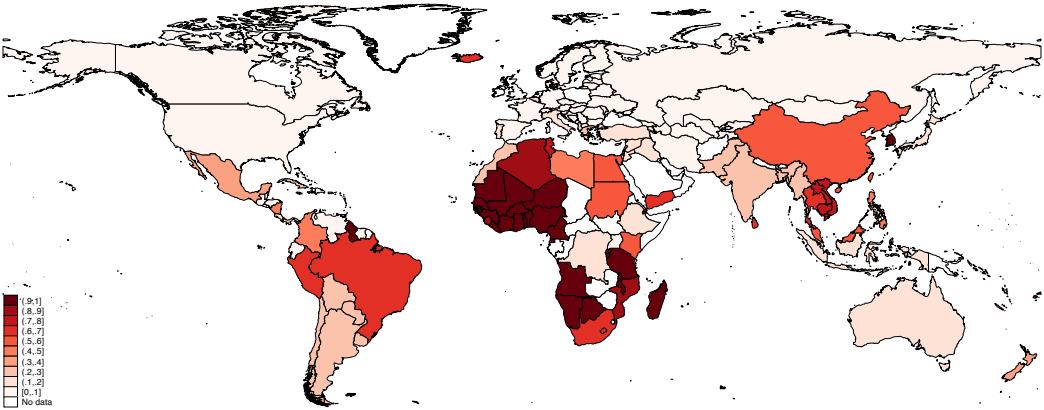
Sovereign debt:

- Meyer Reinhart Trebesch (2019) database (link)
  - **1815–2016:** All foreign-currency bonds traded in London & NY; 91 sovereigns
  - Unique for having monthly prices (note: a large subset of price data come from the IMM & tracking bonds through defaults/restructuring)
- Xu ⓡ Indarte (2021): in progress
  - **1815–1939**: bonds traded in London
  - Text-based measures of issuance & default characteristics/severeity from 3 major newspapers (Economist, FT, London Times)

International banks:

- Kisling Meissner Xu (2021) (link)
  - **1850–1914:** bank-city-year operations around the world, linked to bank nationality
  - Bank size and operations inferred from annual aggregate bank-level balance sheet (1890 onward)

ROW reliance on banks from advanced economies



Data: Kisling Meissner Xu (2021)

Types of data

1. Trade
2. Capital flows
3. Crises, disasters, major events

3. Major events are rare and only history provides data

Financial crises databases: varying definitions of crises leads to different dates

- Reinhart Rogoff: Banking, sovereign debt, and currency crises, 1800-2008, all countries (link)
- MACROHISTORY: financial crises based on narrative evidence, 1870–2017, 18 AEs (link)
- Baron Dieckelman: bank equity index losses, 1870–2016, 43 countries (link)
  - Nice summary of differences between existing databases: (link)
- Metrick Schmelzing (2021): banking crisis interventions, 1257–2019, 138 countries (link)
- Bordo Eichengreen Klingebiel Martinez-Peria (2001): banking and crises, 1880–1997, 21 countries
  - Bordo Meissner (2006) adds 9 EM countries to pre-1914 period + debt crises (link)

Many papers study **particular crises**:

- US crises: Kelly Grada (2000) for 1857; Benmelech Bordo (2007) for 1873; Carlson (2005) for 1893; Frydman Hilt (2015) for 1907
- UK crises: Xu (2019) for 1866; Paolera Taylor (2001) for 1890
- Latin American crises: Kaminsky Vega-Garcia (2014)

3. Major events are rare and only history provides data (cont'd)

Wars and military conflicts:

- Correlates of War (COW) project: (link)
  - Formal alliances, territorial changes, land borders (useful for bilateral trade), colonial dependencies, intergovernmental organizations
  - Bilateral trade (1870–2014) subsumed by previously mentioned datasets

Types of data

1. Trade
2. Capital flows
3. Crises, disasters, major events
4. Macro-data across countries

4. List of resources

- Maddison Project Database: (link)
  - More detailed national accounts (industry-level): (link)
- Global Financial Database: financial indicators (market cap, market rates, central bank indicators)
  - Pre-WWI variables much more limited and mostly for UK, France, Germany, US (big overlap with availability on NBER)
  - Indices often constructed from a few stocks, not the market
- MACROHISTORY database: 1870 onwards, 18 advanced economies (link)
  - Housing, equities, other macro measurements
- Central bank websites: Bank of England/Italy/France/FRED
  - Often have the best/most detailed historical series for that country
  - Example: Bank of England's "Millennium of macrodata" (link)
- Center for Financial Stability (link)
  - Annual data series covers many countries, but better for unusual series rather than a panel

Creating your own datasets

Types of historical data & sources

Types of data:

- Quantitative
- Text/narrative

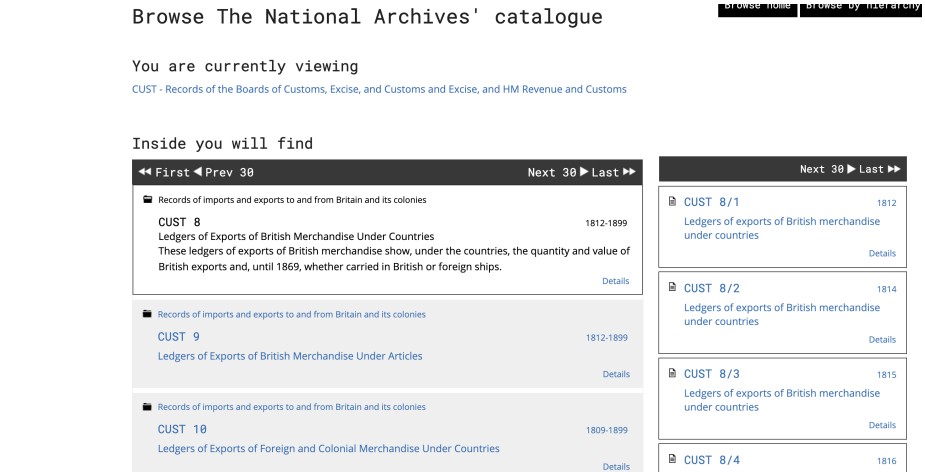
Ultimate sources:

- Archives & libraries
- Digitial respositores: Hathitrust, google books, library databases
- Huge volumes of untapped historical textual data: newspapers, government reports, declassified files, etc.

Examples of ultimate sources

Archival source: CUST 8 ledgers for the United Kingdom are the source for bilateral-industry imports/exports

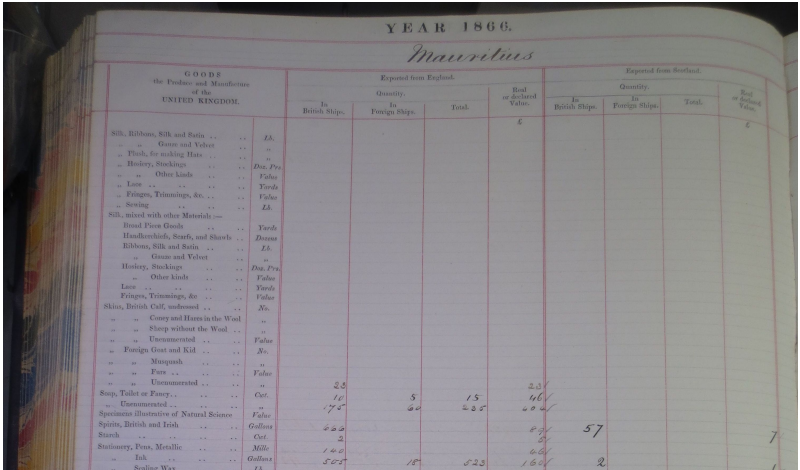
National Archives documentation: (link)



Examples of ultimate sources

Archival source: CUST 8 ledgers for the United Kingdom are the source for bilateral-industry imports/exports

National Archives documentation: (link)



YEAR 1866.

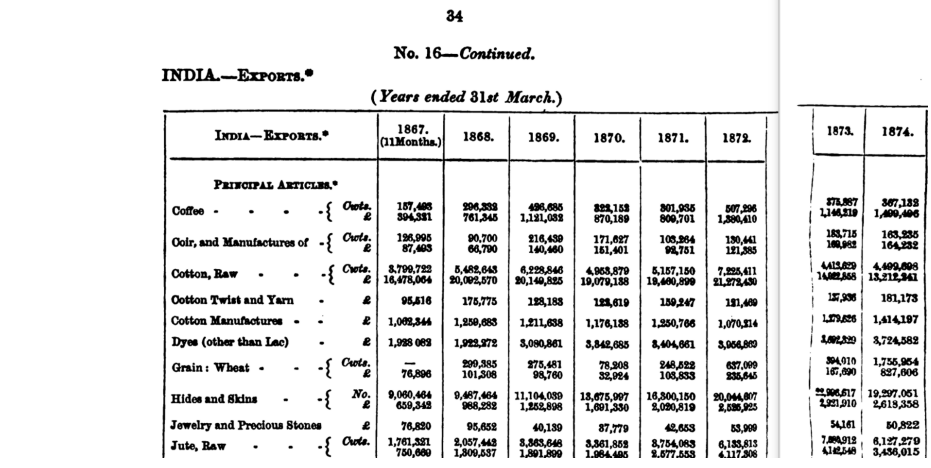
Mauritius

GOODS the Produce and Manufacture of the UNITED KINGDOM.		Exported from England. Quantity. In British Ships.	In Foreign Ships.	Total.	Real or declared Value.	Exported from Scotland. Quantity. In British Ships.	In Foreign Ships.	Total.	Real or declared Value.
					£				£
Silk, Ribbons, Silk and Satin	*Lb.*								
" " Gauze and Velvet	"								
" Plush, for making Hats	"								
" Hosiery, Stockings	*Doz. Prs.*								
" " Other kinds	*Value*								
" Lace	*Yards*								
" Fringes, Trimmings, &c.	*Value*								
" Sewing	*Lb.*								
Silk, mixed with other Materials :—									
Broad Piece Goods	*Yards*								
Handkerchiefs, Scarfs, and Shawls	*Dozens*								
Ribbons, Silk and Satin	*Lb.*								
" Gauze and Velvet	"								
Hosiery, Stockings	*Doz. Prs.*								
" " Other kinds	*Value*								
Lace	*Yards*								
Fringes, Trimmings, &c	*Value*								
Skins, British Calf, undressed	*No.*								
" " Coney and Hares in the Wool	"								
" " Sheep without the Wool	"								
" " Unenumerated	*Value*								
" Foreign Goat and Kid	*No.*								
" " Mosquish	"								
" " Furs	*Value*								
" " Unenumerated	"	23			23				
Soap, Toilet or Fancy	*Cwt.*	10	5	15	16				
" Unenumerated	"	175	60	235	204				
Specimens illustrative of Natural Science	*Value*								
Spirits, British and Irish	*Gallons*	600			84	57			
Starch	*Cwt.*	2			2				7
Stationery, Pens, Metallic	*Mille*	140			60				
" Ink	*Gallons*	505	18	523	160	2			
" Sealing Wax	*Lb.*								

Examples of ultimate sources

Best references for sources (as a starting point) are in the documentation of existing databases. Many have been digitized and are on Hathitrust

Below: *Statistical Abstract for the several colonial and other possessions of the UK, 1864–1883*; (link)



34

No. 16—*Continued.*

INDIA.—EXPORTS.*

(*Years ended 31st March.*)

INDIA—EXPORTS.*		1867. (11 Months.)	1868.	1869.	1870.	1871.	1872.	1873.	1874.
PRINCIPAL ARTICLES.*									
Coffee	Cwts.	157,493	296,333	496,685	823,153	801,935	597,296	575,887	567,133
	£	394,331	761,345	1,121,032	870,189	809,701	1,386,410	1,146,219	1,499,496
Coir, and Manufactures of	Cwts.	126,995	90,700	216,439	171,627	103,264	130,441	183,715	163,235
	£	87,493	66,790	140,460	161,401	92,751	121,585	169,863	164,232
Cotton, Raw	Cwts.	3,799,722	5,482,643	6,228,846	4,953,879	5,157,150	7,225,411	4,413,629	4,499,698
	£	16,478,064	20,092,570	20,149,825	19,079,138	19,460,299	21,272,430	14,822,555	13,212,241
Cotton Twist and Yarn	£	95,516	175,775	123,183	123,619	159,247	121,469	17,936	181,173
Cotton Manufactures	£	1,062,344	1,259,685	1,211,638	1,176,138	1,250,766	1,070,214	1,279,626	1,414,197
Dyes (other than Lac)	£	1,928,082	1,922,972	3,080,861	3,542,685	3,404,661	3,956,869	3,692,229	3,724,582
Grain: Wheat	Cwts.	—	299,385	275,481	78,208	248,522	637,099	394,010	1,755,954
	£	76,896	101,508	98,760	32,924	103,533	235,645	167,590	627,606
Hides and Skins	No.	9,060,464	9,467,464	11,104,039	13,675,997	16,300,150	20,044,807	22,266,517	19,297,051
	£	659,342	968,282	1,252,898	1,691,330	2,020,819	2,535,925	2,221,910	2,618,358
Jewelry and Precious Stones	£	76,820	95,652	40,139	37,779	42,553	53,999	54,161	50,822
Jute, Raw	Cwts.	1,761,321	2,057,443	3,363,646	3,361,852	3,754,063	6,133,613	7,899,912	6,127,279
	£	750,669	1,309,537	1,891,899	1,984,495	2,577,553	4,117,308	4,162,548	3,456,015

Processing

Quantitative data:

- Manual digitization (outsource outsource outsource! lots of firms)
  - Create a template → check the sample output!
  - Consider double-entry for handwriting or low quality
  - Figure out the necessary data precision beforehand
- OCR is improving significantly, but table structure can make this untenable
  - `pytesseract` and `layout parser` are tools to explore

Text data:

- OCR: manual digitization not feasible + textual data errors are observable
- Post-OCR processing

Post-OCR processing

The problem:

- Most OCR'd text of historical sources have significant noise
- Need 80%+ accuracy for topic models & language models (van Strien et al 2020)
- Luckily, most issues are with *segmentation* and *misspelling*

The steps:

1. Find a relevant corpus (Google Web Trillion Word Corpus, or custom ones for the historical period)
2. Segmentation: `wordsegment`
3. Spelling correction: Norvig algorithm
4. Fix incorrectly segmented words: `symspell`

More detail about these steps in the appendix of these slides

Historical (international) papers

Papers written by PhD students

- Hanlon (2015), "Necessity is the mother of invention: Input supplies and directed technological change" (Econometrica)
- Koudijs (2016): "The boats that did not sail: Asset price volatility in a natural experiment" (JF)
- Juhasz (2018), "Temporary protection and technology adoption: Evidence from the Napoleonic Blockade" (AER)
- Steinwender (2018), "Real effects of information frictions: When the States and the Kingdom became United" (AER)
- Xu, G (2018), "The costs of patronage: Evidence from the British Empire" (AER)
- Giorcelli (2019), "The long-term effects of management and technology transfers" (AER)
- Xu, C (2019): "Reshaping global trade: The immediate & long-run effects of bank failures" (R&R QJE)
- Van Patten & Mendez-Chacon (2021) "Multinationals, monopsony, and local development: Evidence from the United Fruit Company" (R&R Econometrica)
- Indarte (2019): "Bad news bankers: Underwriter reputation & contagion in pre-1914 sovereign debt markets"
- Olmstead-Rumsey (2019) "Country banks and the Panic of 1825"

Questions/comments welcome!

chenzixu@stanford.edu

APPENDIX

Steps

0. Clean with regular expressions
1. Segment words
2. Correct words
3. Further refine spelling
4. Append words that should not be split

Step 0: Regular Expressions

- Should there be known consistent errors in the text, these can be easily dealt with in the pre-processing phase via regular expressions

Step 1: Word Segmentation

- Segment using `wordsegment` Python package (Segaran and Hammerbacher, 2009)

$$P(W_{1:n}) = \Pi_{k=1:n} P(W_k)$$

- If the product is higher than any other candidate c's product, that is the best answer, which satisfies:

$$best = \arg\max_{c \in candidates} P(c)$$

- Default Corpus: Google Web Trillion Word Corpus.
- Other corpuses can also be used to improve the quality of segmentation (i.e. Hansard Corpus)

Step 2: Word Correction

- Norvig Algorithm

$$\begin{aligned}\arg\max_c P(c|w) &= \arg\max_c \frac{P(w|c)P(c)}{P(w)} \\ &= \arg\max_c P(w|c)P(c)\end{aligned}$$

- Uses Levenshtein Distance to find permutations of edit distance n from the original word
  - i.e. $n = 3$
- Compares permutation to words in a word frequency list.
  - Words that are found more often in the frequency list are more likely the correct results.
  - Corpus examples:
    - Word2Vec model from Hosseini et al. (2021)
    - Hansard Corpus

Step 3: Further refine corrections

- Depending on the necessity, we further refined words detected initially incorrect and refine them with continued iterations of Step 0, 1 and 2 until a steady state is achieved.

Step 4: Append missplit words

- `symspell` algorithm (Garbe, 2018)
  - `symspell` only requires deletes, no transposes, replacements, inserts, etc – reducing the complexity of edit candidate generation and dictionary lookup
  - Levenshtein distance (i.e. of 2)
  - Properly segments text as well.
  - Default (bigram) dictionary is sourced from Google Books Ngram data and SCOWL.
    - User-generated corpuses are also fine

Examples

	Original	Segment 1	Correction 1	Segment 2	Correction 2	Document Number
0	beengreat	been great	been great	been great	been great	0
1	strongo	strong o	strong	strong	strong	0
2	porfeuille	por feuille	portfeuille	port feuille	portefeuille	0
3	aspart	as part	apart	apart	apart	0
4	mcotom	mco tom	cottom	cot tom	cottom	0
5	runperson	run person	run person	run person	run person	0
6	openedcredit	opened credit	opened credit	opened credit	opened credit	0
7	receiveconsiderable	receive considerable	receive considerable	receive considerable	receive considerable	0
8	counteractedby	counteracted by	counteracted by	counteracted by	counteracted by	0