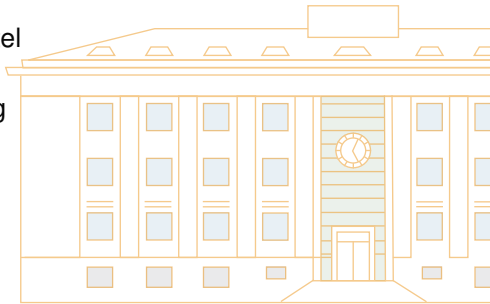
unine
Université de Neuchâtel
Institut de recherches économiques

Swiss Economic Sentiments for the 19th, 20th, and 21st centuries

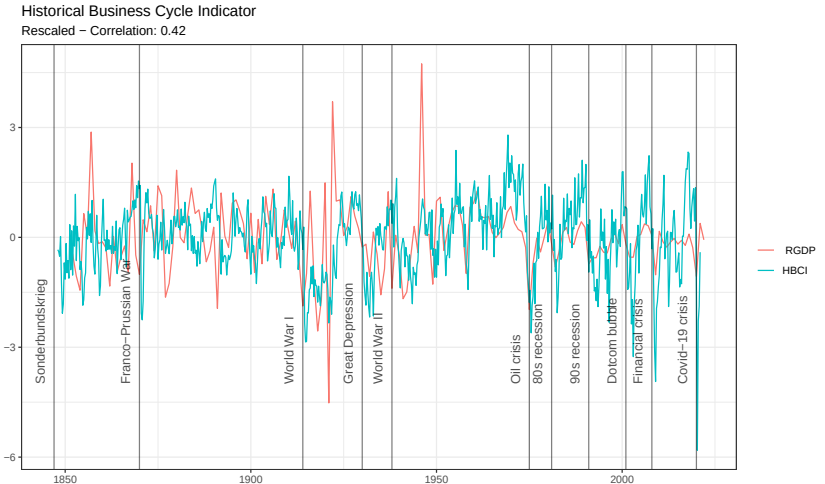
Marc Burri
University of Neuchâtel

IRENE PhD Meeting
September 8, 2023



What I do and what I find

- I construct business cycle indicators for several hand-selected economic concepts based textual data such as newspapers, business reports and business association reports.
- A composite indicator is highly correlated with existing economic activity data for the 20th century.



Why you should care

- Real activity data in the 19th century Switzerland is measured inaccurately and at low frequency.
- Earliest business cycle indicator for Switzerland based on survey data starts in 1966.
- Use of alternative data sources is on the rise in many areas of economics.

Literature and contribution

Textual data and news sentiment indicators receive more and more attention

- for measuring economic activity (See e.g. Buckman et al. (2020), Thorsrud (2020), Kalamara et al. (2022) or Barbaglia et al. (2022))
- to measure recession perception/economic uncertainty (See e.g. The Economist (2011), Iselin and Siliverstovs (2013), Baker et al. (2016) or Larsen (2021)).

Contribution:

- Longest indicator of Swiss economic activity (1848-2021).
- High frequency (quarterly).
- Collection and digitizing of valuable archival material.
- Propose method for signal extraction from noisy historical text data.

Text data

- Language: German and French
- Number of publications: 107
- Raw data size: approx. 7 TB

Data sources

Source	German	French	Availability
AWP	1.86		2000 - 2021
e-newspapersarchives.ch	17.79	34.56	1848 - 2020
SWA	0.16		1848 - 2014
Abbyy	10.63	4.05	1848 - 2021
Tamedia	4.57	11.45	1995 - 2021
Scriptorium		22.27	1848 - 2021
Total	35.03	62.04	

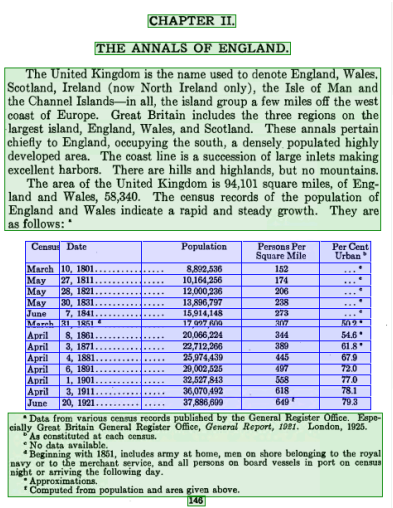
Notes: In Millions, SWA only includes publications scanned by them, Abbyy includes own scans and PDFs from SWA, CS, ZKB, SNB, Bund and KfK

Challenges

- Very heterogeneous pile of text types and formats.
- Some individual scans are of poor quality.
- Changing language over time.
- Texts in different languages.
- Huge amount of data but little computing power.
- Texts are getting longer over time due to technological progress.
- Great variation of optical character recognition (OCR) quality within the same publication.
- Publications differ in time coverage as well as publication frequency.
- Validation of the indicator difficult, especially in early sample.

1. Image preprocessing, OCR and layout parsing

- Preprocessing of scanned images: Crop, black and white, increase sharpness and contrast, remove speckles and curvature.
- Use Abbyy Finereader for OCR and layout parsing.
- Assess quality of OCR and filters:
  - Count words appearing in a German/French lexicon (dict.cc).
  - 20% of words in text must be contained in lexicon.
  - And more German than French words must be identified (or vice-versa).
  - Filter out tables, advertisements and pages with lots of numbers.



CHAPTER II.

THE ANNALS OF ENGLAND.

The United Kingdom is the name used to denote England, Wales, Scotland, Ireland (now North Ireland only), the Isle of Man and the Channel Islands—in all, the island group a few miles off the west coast of Europe. Great Britain includes the three regions on the largest island, England, Wales, and Scotland. These annals pertain chiefly to England, occupying the south, a densely populated highly developed area. The coast line is a succession of large inlets making excellent harbors. There are hills and highlands, but no mountains.

The area of the United Kingdom is 94,101 square miles, of England and Wales, 58,340. The census records of the population of England and Wales indicate a rapid and steady growth. They are as follows: [a]

Census	Date	Population	Persons Per Square Mile	Per Cent Urban [b]
March	10, 1801	8,892,536	152	... [c]
May	27, 1811	10,164,256	174	... [c]
May	28, 1821	12,000,236	206	... [c]
May	30, 1831	13,896,797	238	... [c]
June	7, 1841	15,914,148	273	... [c]
March	31, 1851 [d]	17,927,609	307	50.2 [c]
April	8, 1861	20,066,224	344	54.6 [c]
April	3, 1871	22,712,266	389	61.8 [c]
April	4, 1881	25,974,439	445	67.9
April	6, 1891	29,002,525	497	72.0
April	1, 1901	32,527,843	558	77.0
April	3, 1911	36,070,492	618	78.1
June	20, 1921	37,886,699	649 [e]	79.3

[a] Data from various census records published by the General Register Office. Especially Great Britain General Register Office, *General Report, 1921*. London, 1925.
[b] As constituted at each census.
[c] No data available.
[d] Beginning with 1851, includes army at home, men on shore belonging to the royal navy or to the merchant service, and all persons on board vessels in port on census night or arriving the following day.
[e] Approximations.
[f] Computed from population and area given above.

146

Challenges

- ~~Very heterogeneous pile of text types and formats.~~ ✓
- ~~Some individual scans are of poor quality.~~ ✓
- Changing language over time.
- Texts in different languages.
- Huge amount of data but little computing power.
- Texts are getting longer over time due to technological progress.
- Great variation of OCR quality within the same publication.
- Publications differ in time coverage as well as publication frequency.
- Validation of the indicator difficult, especially in early sample.

2. Creation of text indicators for each publication

- Select keywords defining 12 economic topics (e.g real activity, recession/crisis etc.) by reading through historical texts (Burri, 2023). Keywords
- Use ChatGPT to translate selected keywords to French.

Sentiment-based indicators

- Extract keywords, the 15 preceding words, and the 15 following terms (T).
- Define a list of positive (P) and negative (N) terms (Remus et al., 2010, Abdaoui et al., 2017).
- Calculate a sentiment score: $S_{t,s} = (\sum T \in P - \sum T \in N)/\sum T$.
- Sentiment indicators for a given topic and publication are calculated as a simple average of the sentiment scores.

Count-based indicators

- Count number of appearances of these keywords.

Challenges

- ~~Very heterogeneous pile of text types and formats.~~ ✓
- ~~Some individual scans are of poor quality.~~ ✓
- ~~Changing language over time.~~ ✓
- ~~Texts in different languages.~~ ✓
- ~~Huge amount of data but little computing power.~~ ✓
- Texts are getting longer over time due to technological progress.
- Great variation of OCR quality within the same publication.
- Publications differ in time coverage as well as publication frequency.
- Validation of the indicator difficult, especially in early sample.

3. Indicator anomalies correction

Sentiment-based indicators Example

- Outlier removal: Remove sentiment scores that are more than 3 standard deviations away from mean.
- Detrending: Substract trend (LOESS) from calculated sentiments
- Breaks: Remove structural changes in mean and variance using a binary segmentation method allowing for a maximum of 2 breaks and specifying a minimum sequnce length of 5 years.

Count-based indicators Example

- Dynamic normalization (Ardia et al., 2021): Normalize $S_{t,s}$ by its m-past years observations (rolling-window).

Challenges

- ~~Very heterogeneous pile of text types and formats.~~ ✓
- ~~Some individual scans are of poor quality.~~ ✓
- ~~Changing language over time.~~ ✓
- ~~Texts in different languages.~~ ✓
- ~~Huge amount of data but little computing power.~~ ✓
- ~~Texts are getting longer over time due to technological progress.~~ ✓
- ~~Great variation of OCR quality within the same publication.~~ ✓
- Publications differ in time coverage as well as publication frequency.
- Validation of the indicator difficult, especially in early sample.

4. Aggregation

Indicators for 12 economic topics:

- Aggregate high frequency publications to quarterly.
- Spread low frequency indicators to quarters.
- Weight publications by number identified keywords.

Composite indicators:

- Calculate Dynamic Factor Model (DFM) with all topics Details

Challenges

- ~~Very heterogeneous pile of text types and formats.~~ ✓
- ~~Some individual scans are of poor quality.~~ ✓
- ~~Changing language over time.~~ ✓
- ~~Texts in different languages.~~ ✓
- ~~Huge amount of data but little computing power.~~ ✓
- ~~Texts are getting longer over time due to technological progress.~~ ✓
- ~~Great variation of OCR quality within the same publication.~~ ✓
- ~~Publications differ in time coverage as well as publication frequency.~~ ✓
- Validation of the indicator difficult, especially in early sample.

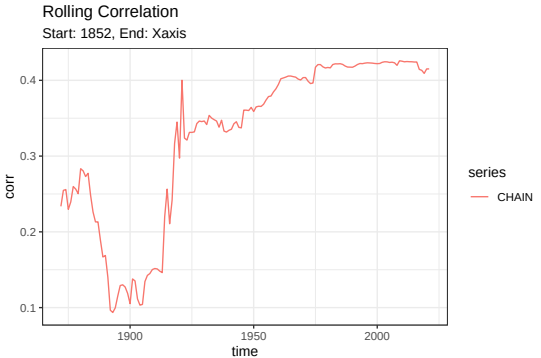
5. Validation

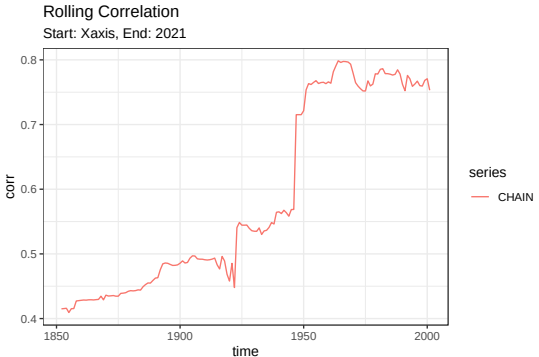
- Annual real GDP growth chained from several sources:
  - 1848 - 1948: Maddison database
  - 1949 - 1979: FSO
  - 1980 - now: SECO

  Graph
- International GDP: Maddison databse (1848-2019) Graph
- Labor market data: HSSO (1890-2005) Graph
- Trade volumes: HSSO (1852-2000) Graph
- Thorp GPT Sentiment (1800-1925) Graph Details

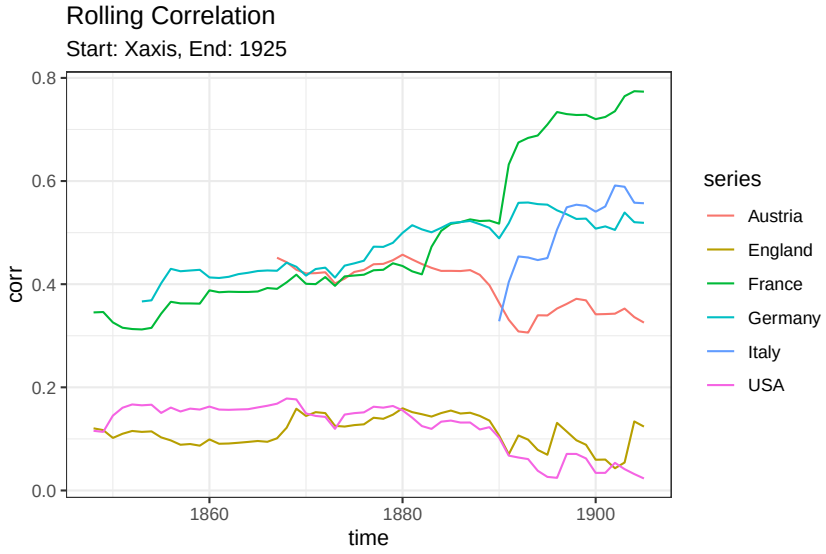
Composite indicator cross correlations

Rolling correlation with Real GDP growth





Rolling correlation with Thorp sentiment



Challenges

- ~~Very heterogeneous pile of text types and formats.~~ ✓
- ~~Some individual scans are of poor quality.~~ ✓
- ~~Changing language over time.~~ ✓
- ~~Texts in different languages.~~ ✓
- ~~Huge amount of data but little computing power.~~ ✓
- ~~Texts are getting longer over time due to technological progress.~~ ✓
- ~~Great variation of OCR quality within the same publication.~~ ✓
- ~~Publications differ in time coverage as well as publication frequency.~~ ✓
- ~~Validation of the indicator difficult, especially in early sample.~~ ✓

Concluding remarks

Using textual analysis of historical data sources I create a quarterly business cycle indicator:

- The indicator is highly correlated with existing data for economic activity in the 20th century.
- Lower correlation with existing data for 19th century.
- Strong correlation in 19th century with a created sentiment indicator for neighboring countries.
- My results indicate that in the 19th century:
  - Swiss economic activity is indeed inaccurately measured.
  - Business cycles were more regional in nature.

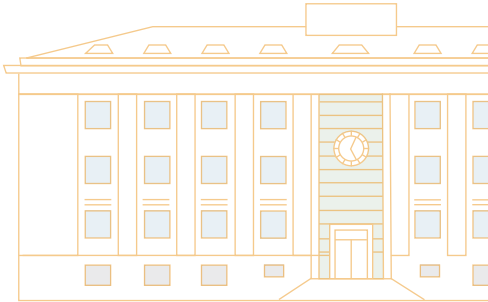
Next steps

1. Rethink handling of different frequencies (use temporal disaggregation).
2. Fit the factors to quarterly GDP growth from 1980 - 2020 (or different measure for business cycle, HP filtered GDP?) and backcast to 1848.

unine
Université de Neuchâtel
Institut de recherches économiques

Thank you!

marc.burri@unine.ch
marcburri.github.io



Appendix

References

Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2017). FEEL: a French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51(3):833–855, DOI: 10.1007/s10579-016-9364-5, Retrieved from https://hal-lirmm.ccsd.cnrs.fr/lirmm-01348016. Publisher: Springer Verlag.

Ardia, D., Bluteau, K., and Kassem, A. (2021). A century of Economic Policy Uncertainty through the French–Canadian lens. *Economics Letters*, 205:109938, ISSN: 01651765, DOI: 10.1016/j.econlet.2021.109938, Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0165176521002159.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, ISSN: 1531-4650, 0033-5533, DOI: 10.1093/qje/qjw024, Retrieved from https://academic.oup.com/qje/article/131/4/1593/2468873.

Barbaglia, L., Consoli, S., and Manzan, S. (2022). Forecasting with Economic News. *Journal of Business & Economic Statistics*, pages 1–12, ISSN: 0735-0015, 1537-2707, DOI: 10.1080/07350015.2022.2060988, Retrieved from https://www.tandfonline.com/doi/full/10.1080/07350015.2022.2060988.

References II

Buckman, S. R., Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). News sentiment in the time of COVID-19. *FRBSF Economic Letter*, 2020(08):1–05, Retrieved from https://www.frbsf.org/economic-research/publications/economic-letter/2020/april/news-sentiment-time-of-covid-19.

Burri, M. (2023). Do daily lead texts help nowcasting GDP growth? IRENE Working Papers 23-02, IRENE Institute of Economic Research, Retrieved from https://ideas.repec.org/p/irn/wpaper/23-02.html.

Doz, C., Giannone, D., and Reichlin, L. (2012). A QUASI—MAXIMUM LIKELIHOOD APPROACH FOR LARGE, APPROXIMATE DYNAMIC FACTOR MODELS. *The Review of Economics and Statistics*, 94(4):1014–1024, ISSN: 00346535, 15309142, Retrieved from http://www.jstor.org/stable/23355337. Publisher: The MIT Press.

Iselin, D. and Siliverstovs, B. (2013). The R-word index for Switzerland. *Applied Economics Letters*, 20(11):1032–1035, ISSN: 1350-4851, 1466-4291, DOI: 10.1080/13504851.2013.772290, Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13504851.2013.772290.

References III

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., and Kapadia, S. (2022). Making text count: economic forecasting using newspaper text*. *Journal of Applied Econometrics*, ISSN: 0883-7252, 1099-1255, DOI: 10.1002/jae.2907, Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/jae.2907.

Larsen, V. H. (2021). Components of Uncertainty. *International Economic Review*, 62(2):769–788, ISSN: 0020-6598, 1468-2354, DOI: 10.1111/iere.12499, Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/iere.12499.

Mariano, R. S. and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, 18(4):427–443, ISSN: 0883-7252, 1099-1255, DOI: 10.1002/jae.695, Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/jae.695.

References IV

Remus, R., Quasthoff, U., and Heyer, G. (2010). SentiWS - a publicly available German-language resource for sentiment analysis. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA), Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/490⟨sub⟩P⟨/sub⟩aper.pdf.

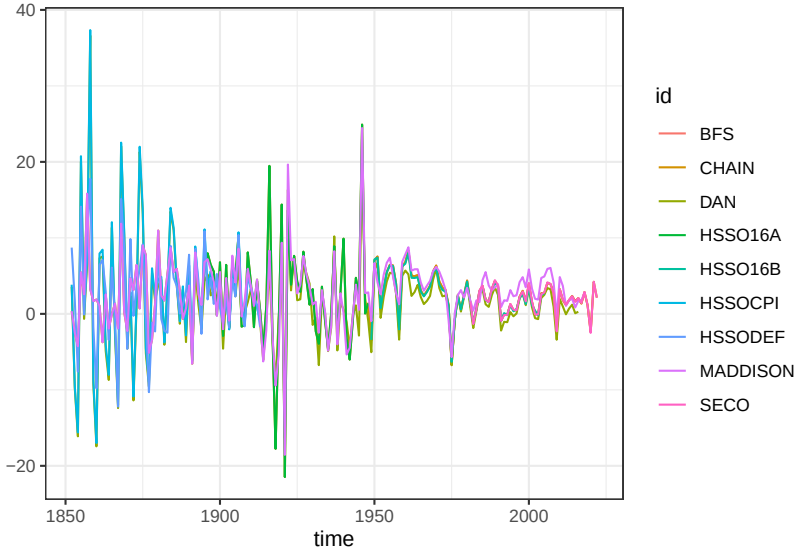
The Economist (2011). Up means down. The economist's gauge of gloom. Retrieved from http://www.economist.com/node/21529079.

Thorp, W. L. (1926). *Business annals*. National Bureau of Economic Research, Inc, Retrieved from https://EconPapers.repec.org/RePEc:nbr:nberbk:thor26-1.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409, DOI: 10.1080/07350015.2018.1506344.

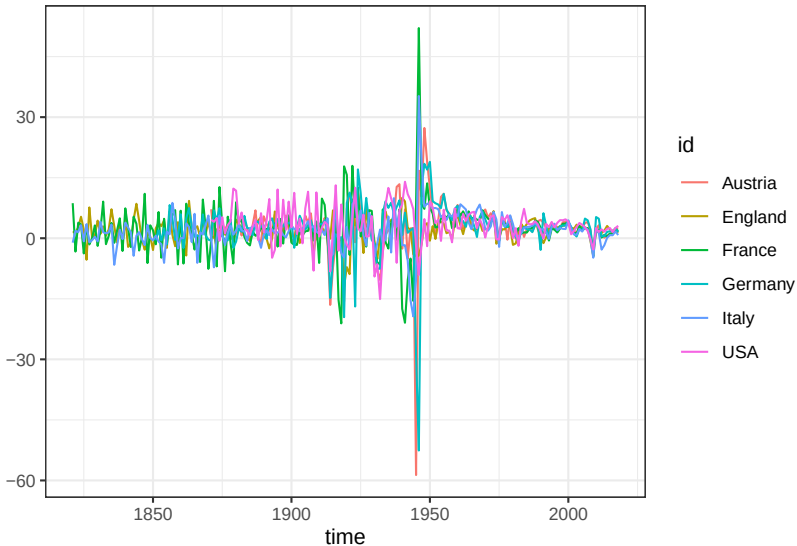
Validation data: Real GDP

Back



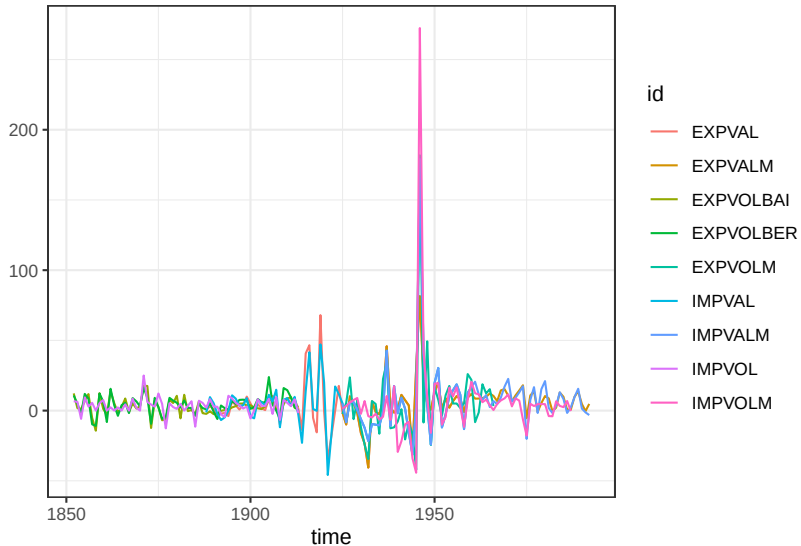
Validation data: International GDP

Back



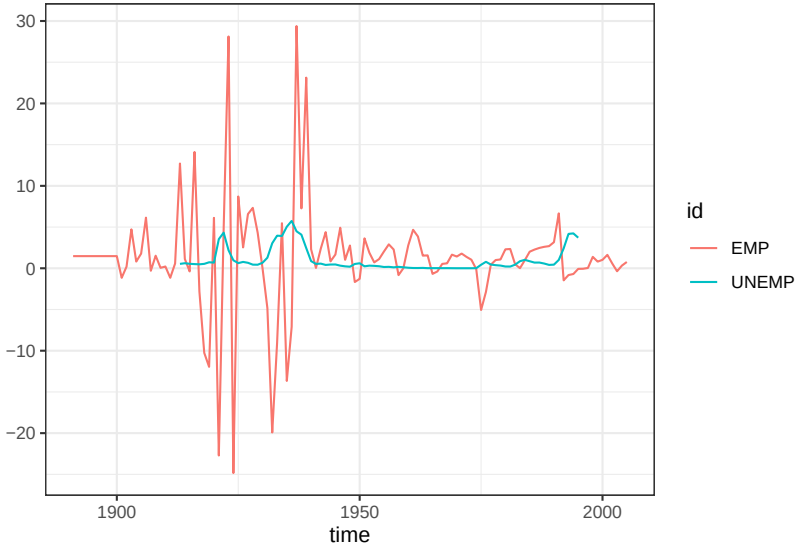
Validation data: Trade

Back



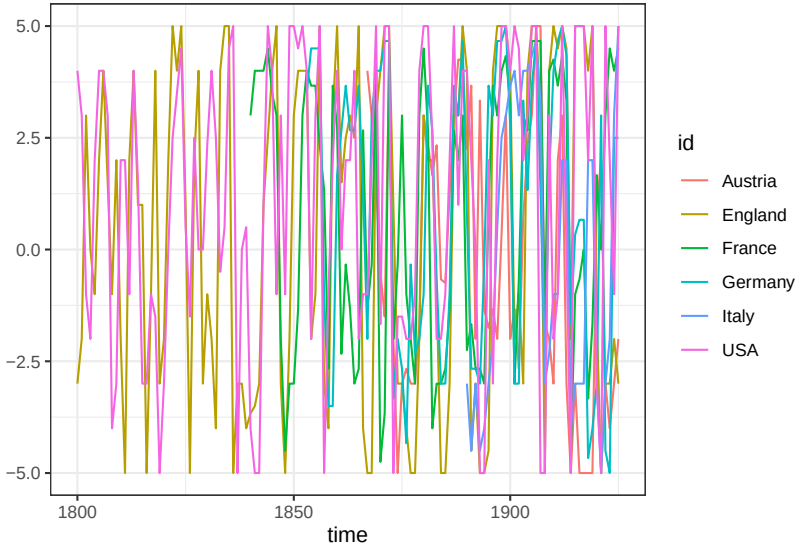
Validation data: Labor Market

Back



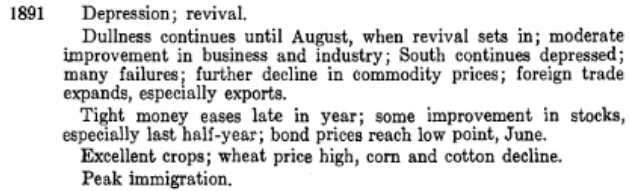
Validation data: Thorp Sentiment

Back



Thorp GPT Sentiment

Fig.: Excerpt from Business Annal for 1891. Source: Thorp (1926, p.136)



1891 Depression; revival.

Dullness continues until August, when revival sets in; moderate improvement in business and industry; South continues depressed; many failures; further decline in commodity prices; foreign trade expands, especially exports.

Tight money eases late in year; some improvement in stocks, especially last half-year; bond prices reach low point, June.

Excellent crops; wheat price high, corn and cotton decline.

Peak immigration.

I ask GPT-3.5: The following text describes how the economy in the USA was doing in the years from 1905 to 1925. Please rate the state of the economy based on these texts from -5 to 5, with -5 being the worst and 5 the best. Basically you should judge based on the text and also by comparing between the different years in which state the economy was in that given year. In your answer only provide the year and the assigned value in table format. Back

Topic defining keywords

Back

Topic	Keywords (based on readings)	Method
Recession	crisis krise rezession recession finanzcrisis finanzkrisis spekulationskrise krach krisis zahlungsfähigkeit fallimente bankerottir liegenschaftenkrisis konkurse schaden konjunkturrückgang weltbrand notstand valutasturz depression valutaschwierigkeit wirtschaftskatastrophe schäden liquidation liquidier hemmnisse zusammenbruch notlage katastrophe baisse	Count
Real activity	wirtschaft ware absatz nachfrage geschäft konsum waare fabrikant erlös umsatz markt industrie branche käufer unternehmer ernte ergebniss konjunktur kundschaft verkauf produktion dienstleistung verarbeitung gewerbe ertrag einnahmen ausgaben fabrikation bestellung versorgung materialbeschaffung einkäufer verlust konkurrenten fabrizieren fabrikat besteller werth neugründung materialien betriebsmittel materialeinkäufe jahresresultat geschäftsperiode werkstätten erfolg bestellungen eigenkosten produkte rohmaterialien einbusse fabric fabrik herstellung geschäftsgang wirtschaftsleben wirtschaftlich nachfrage geschäftslage marge angebot erträgnis rendite produzent vertrieb volkswirtschaft konjunkturrückgang verkaufsziffern kauflust kaufunlust geschäftsjahr kleinbetrieb bautätigkeit verbraucher konkurrenz erzeugnis konsum profit fremdenverkehr dienste dienstleistungen investition versorgungsmöglichkeit versorgungslage versorgungsschwierig bruttoinlandprodukt bruttosozialprodukt realwachstum wertschöpfung	KWIC

Topic defining keywords (cont) Back

Topic	Keywords (based on readings)	Method
Trade	eingangszölle eingangszoll konkurrenzverhältnisse konkurrenzverhältnis einfuhr ausfuhr export import sendungen aussendung importeure exporteure handelsstatistik absatzfeld wettbewerb absatzgebiet zwischenhandel handel importhaus zollverhältnisse zollverhältnis handelsbilanz waarenverkehr warenverkehr waarenausfuhr warenausfuhr waareneinfuhr wareneinfuhr importhandel handelsverkehr zoll zolleinnahmen zölle weltbedarf fracht exportziffern gesammtexport gesamtexport taxen verkehrserleichterung bezugsquelle ausland generaltarif tarif einfuhrverbote zufuhren zufuhr grosshändler seefracht wasserweg welthandel weltverkehr güterstrom güteraustausch güterumschlag umschlagverkehr wagenverkehr	KWIC
Capacity	lager kornspeicher speicher ueberproduktion überproduktion vorräthe vorräte liefertermine lieferfrist vorrat vorrath aufträge lieferfristen lagerware lager depots bestellungen lieferungen wagenmangel	KWIC
Labor	arbeit erwerb beruf erwerbende arbeiter aufsichtspersonal arbeiterin angestellte arbeitskräfte beschäftigung arbeitszeit arbeitgeber arbeiterschaft ueberzeit überzeit arbeitstag arbeitsleistung ausbildung lehrlinge ueberzeitarbeit überzeitarbeit streik arbeitseinstellung arbeitsfeld thätigkeit tätigkeit arbeiterinnen personal	KWIC

Topic defining keywords (cont) Back

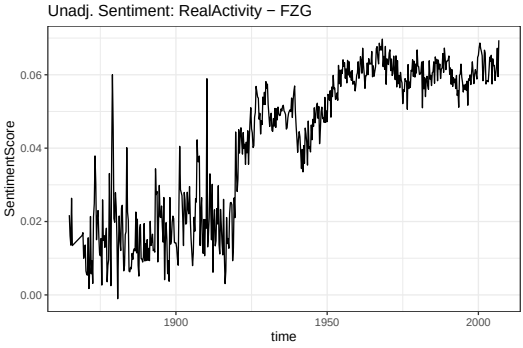
Topic	Keywords (based on readings)	Method
Inflation	preis getreidepreise theuer teuer preisfall kostenpreise wechselkurs silberpreis preisaufschlag silberkurs preisbasis kurs preissturz baumwollpreise preisnotierung preisnotirung waarenpreise warenpreise kursschwankungen kostenpreis preissteigerung schleuderpreise abschlag materialpreise maschinenpreise unterbietung preisvorteile preisvortheile verkaufspreise rohpreise preiserhöhung entwerthung entwertung goldkurs geldvertheuerung geldverteuerung geldwerth geldwert vertheuerung verteuerung wechselkurse pari preise teuerung kaufkraft aufschlag inflation inflatorisch	KWIC
Wages	entgelt lohnerhöhung arbeitslöhne löhne arbeitslohn kaufkraft lohnverhältnisse lohnsätze lohnansätze stundenlöhne stundenlohn akkordlöhne akkordlohn akkord tagesverdienst löhnungen löhnung lohnverhältnis verdienstverhältnisse verdienstveräl̈tnis einkommen lohnniveau lohnstopp lohnpolitik	KWIC
Credit	kreditverhältnisse banknoten münzen einleger guthaben einzahlung rückzahlung prämien renten kapitalien verzinsung amortisation wechsel geldmarkt zinsen zins rendite diskontsätze disconto discontsätze diskontosätze diskontosatz diskontsatz geldkraft zinsfuss wechselgeschäft wechselkverkehr zinssätze zinssatz diskontopolitik geldinstitute lombardvorschüsse notenemission vorschüsse kontokorrent kreditwirtschaft kredit schuldbriefe kreditwesen leihmarkt emissionsbanken diskonto geldbedarf diskonti diskontoverkehr geldsuchenden emissionen geldverteuerung liquidität geldknappheit geldstand lombardsatz geldleihpreis geldmärkte kapitalmarkt anleihen hypothek depositen darlehen pfandbrief geldpolitik geldmenge	KWIC

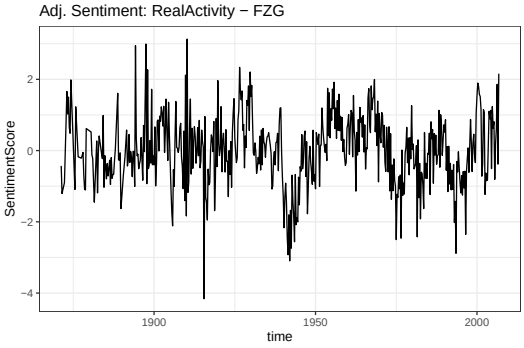
Topic defining keywords (cont) Back

Topic	Keywords (based on readings)	Method
Financial	kapitalvermehrung emissionskurs tageskurs agio kurse dividenden emittirt emittiert obligationen rentabilität konversion börsengeschäft kapital werthpapiere wertpapiere titelverkäufe krach krisis crisis papiere kurssteigerung portefeuille tratten ueberspekulation überspekulation entwerthung entwertung finanzcrisis finanzkrisis börse emission gründung aktie actie aktien kurs effekten märkte wechselkurs devisen valoren dividende wertschriften	KWIC
Uncertainty	spekulation unsicherheit unsichern vorsicht riskirt riskiert misstrauen reserven risiko schwankungen unsicher riskant riskieren risiken gefährlichkeit gefahr sorgen erschütterung furcht unruhe unbehagen	Count
Boom	hausse aufschwung hochkonjunktur prosperität erholungsperiode boom	Count
War	krieg konflikt	Count

Example sentiment-based indicator correction

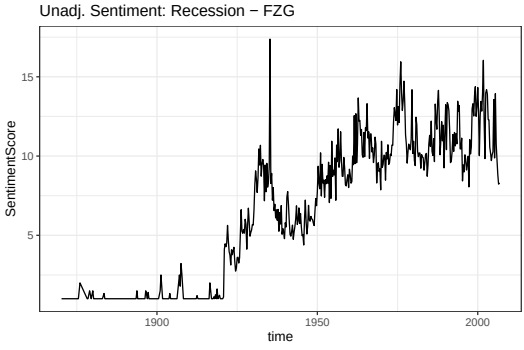
Back

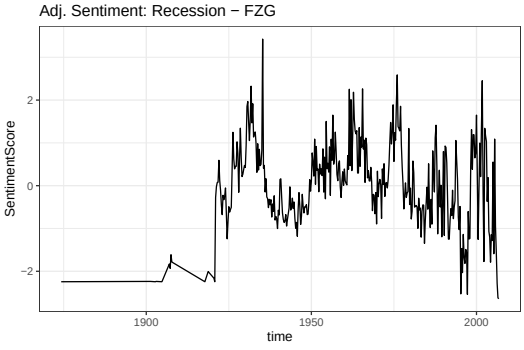




Example count-based indicator correction

Back





Methodology: Dynamic Factor Model

Back

$$x_t = Cf_t + e_t, \quad e_t \sim N(0, R)$$

$$f_t = \sum_{j=1}^{p} A_j f_{t-j} + u_t, \quad u_t \sim N(0, Q)$$

- First factor can be interpreted as a coincident business cycle indicator (Mariano and Murasawa, 2003, Doz et al., 2012).

- x_t = data matrix
- f_t = common factors
- C = factor loadings matrix
- A_j = state transition matrix
- e_t = the unexplained error term
- u_t = shocks to factors