Historical Analysis of National Wellbeing Using Millions of Digitized Books

Daniel Sgroi
University of Warwick, CAGE & IZA
daniel.sgroi@warwick.ac.uk
-
Joint work with Thomas Hills (Warwick) and Eugenio Proto (Bristol)

Subjective Wellbeing and Gross Domestic Product

- Subjective wellbeing (or "happiness") has played a minor role in the development and application of economic policy in the past
- Growing literature on international patterns of subjective wellbeing.
- Several nations including the UK, Australia, China, France and Canada now collect subjective wellbeing data to use alongside GDP in national measurement exercises. OECD & UN also active since 2011.

Our Approach

- Our primary objective is to produce a workable proxy for subjective wellbeing going back to 1800, which would enable direct comparisons with GDP over that period.
- Our methods rely on the digitization of books, available in the Google Books corpus.
- We elected to start in 1800 because the number of digitalized books is too small before.

Word Norms or *Valence*

- The approach we take here is a common approach among the studies Inferring public mood and relies on affective word norms to derive sentiment from text
- In a study of 17 million blog posts, (Nguyen et al, 2010) found that a simple calculation based on the weighted affective ratings of words was highly effective (70% accuracy) at predicting the mood of blogs compared against the groundtruth provided by the bloggers

Language Corpus Data

- The language corpora we used is the *Google Books Ngram* Corpus `https://books.google.com/ngrams`
- The corpus is based on a digitalised database of several million published books, which was developed as part of the Google Books programme.
- We analysed data for 6 languages, English (British), English (American), German, Italian, Spanish, French.
- There are no word norms available for Chinese, Hebrew and Russian

Affective Norms for Different Languages

- For English, ANEW contains about 10,000 words, all rated on a 1 to 9 valence scale by a group of subjects.
- For German, we used the Affective norms for German sentiment terms. This is a list of 1003 words, a German translations of the ANEW list. The valence ratings were collected on a -3 to +3 scale. The mean values were adjusted to reflect a 1 to 9 scale.
- the French and Spanish norms were also adaptations of the ANEW. These contained 1031 and 1034 words respectively. Both used a 1 to 9 points scale.
- For Italian, we used an adaptation of the ANEW norms containing 1121 Italian words, based on the ANEW material on a 1 to 9 scale.

Valence and Words in different languages

- High end: Happiness 8.53, Enjoyment 8.37, Vacation 8.53, Joy 8.21, Relaxing 8.19, Peaceful 8, Lovemaking 7.95, Celebrate 7.84.
- Low end: Murder 1.48, Abuse 1.53, Die 1.67, Disease 1.68, Starvation 1.72, Stress 1.79, Unhappy 1.84, Hateful 1.9.
- Middle: Neutral 5.5, Converse 5.37, Eight 5.37, Century 5.36, Machinery 4.65, Platoon 4.65.

Language Average Valence Computation

- For each language we compute the weighted valence score, $Valence_t$, for each year, t, using the valence, v for each word, j, as follows,

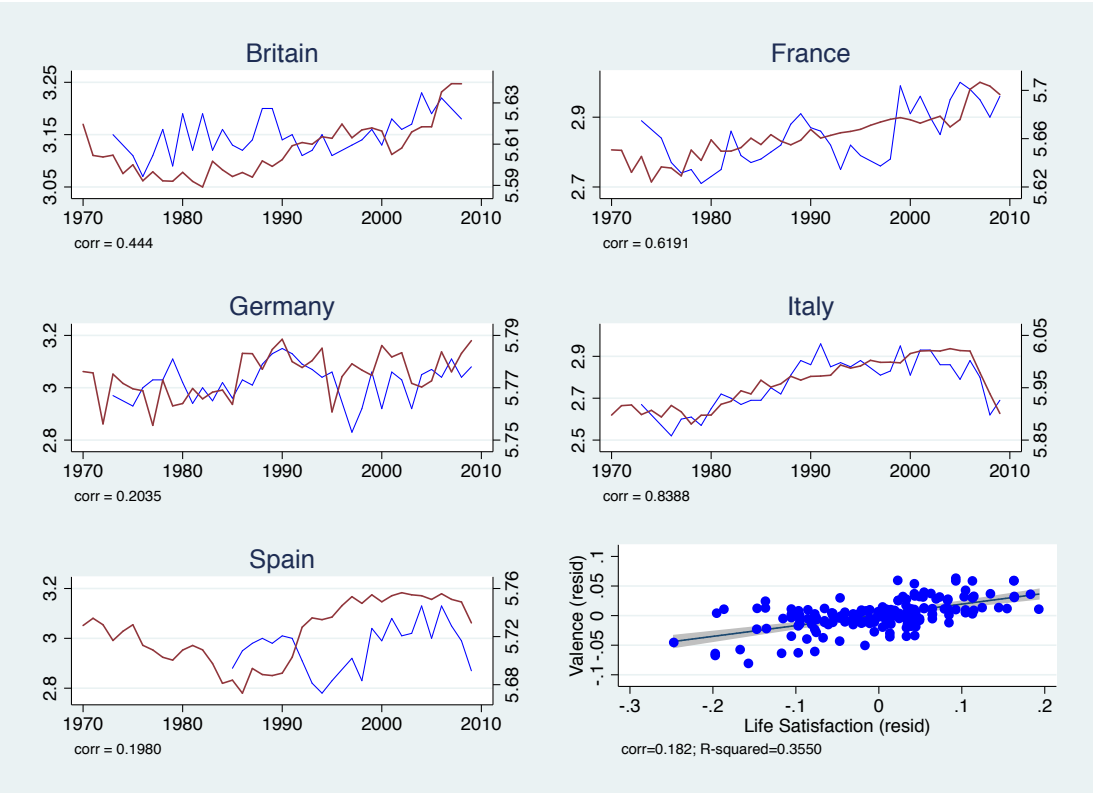
$$Val_{i,t} = \sum_{j=1}^{n} v_{j,i} p_{j,i,t};$$

- Note that $v_{j,i}$ is the valence for word j as found in the appropriate valence norms for language i, and $p_{j,i,t}$ is the proportion of word j in year t for the language i.

How to Interpret the Index

- Think about the book market as highly competitive (lots of potential writers and publishers): publishers "match" books to demand.
- It could be that publishers match happy people to happy books and the opposite?
- It could be that writers are inspired by periods and happy period inspires happy books and the opposite?
- We will try to answer this question by comparing the available data on SWB with word-valence based index

Valence and Existing data of Life Satisfaction

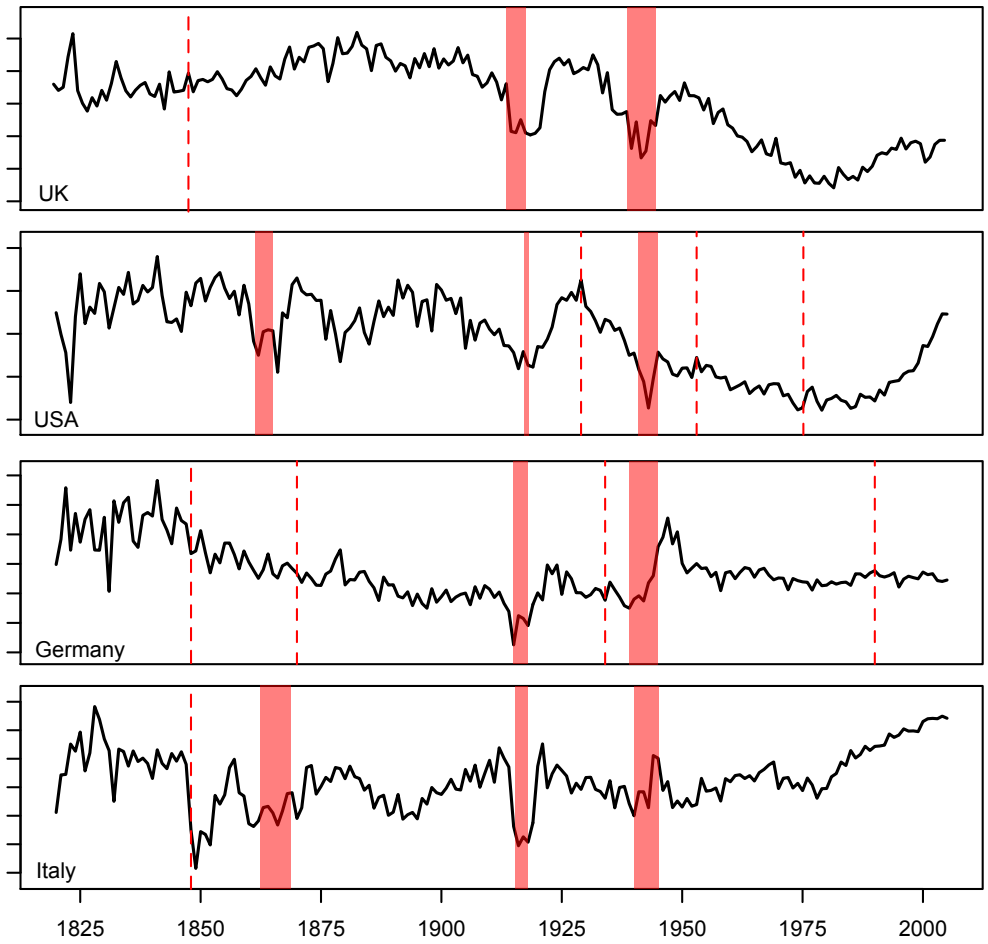


Valence Predicts Aggregate Life Satisfaction

Table: Average life satisfaction per country and year is the dependent variable.

	1 Baseline b/se	2 Until 2000 b/se	3 w/o Sp.and Fr. b/se	4 Trends b/se	5 +GDP b/se	6 Year FE b/se
Valence	1.9554***	1.6941***	2.1696***	1.5549***	0.7180**	1.6107***
	(0.2221)	(0.3093)	(0.2339)	(0.3408)	(0.3499)	(0.2784)
Log GDP					0.8243***	0.1452
					(0.1537)	(0.1300)
Words Covered	0.9816	–0.0037	–0.4491	8.7147	–0.1693	–15.6331*
	(6.2645)	(9.1248)	(5.7111)	(15.0425)	(13.9245)	(8.5557)
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Country Specific Trend	No	No	No	Yes	Yes	No
Year FE	No	No	No	No	No	Yes
r2	0.358	0.227	0.501	0.387	0.485	0.645
N	163	119	104	163	163	163

A Time-Series Plot of the Valence Index Over the Period 1820-2009



Data Concerns

- Long-run biases might emerge from country-specific factors such as culture, language, religion and demographics (immigration, population age structure). We can control these to some extent through country fixed effects.
- Literacy was lower in the past, Language different. We control for education, trends, year fixed effect.
- Freedom of the press, we control for democracy.

Historical Determinants of the Valence Index from 1820 to 2009.

Table: The countries included are Germany, Italy, the UK and the United States

	1 Year FE b/se	2 CS Trends b/se
(log) GDP(t-3)	0.0821**	0.0517*
	(0.0174)	(0.0213)
Life Expectancy(t)	0.0036**	0.0016
	(0.0008)	(0.0014)
Internal Conflict(t)		–0.0190**
		(0.0049)
World Covered(t)	Yes	Yes
Democracy(t)	Yes	Yes
Education Inequality(t)	Yes	Yes
Year FE	Yes	No
Country-Specific Trends	No	Yes
r2	0.736	0.494
N	412	412

Summary

- Average Word Valence of a language predicts country aggregate Subjective Wellbeing of the corresponding country
- Valence Index positively correlates with Life Expectancy, GDP and negatively with conflict