Multidimensional data analysis on French budget data from 1872 until 1971

Soütrik BANERJEE

Data ScienceTech Institute,

4 rue de la Collegiale,

75005 Paris.

soutrik.banerjee@edu.dsti.institute

Introduction

In the current analyses, the components of the French budget from 1872 until 1971, were analysed. There were 11 different components in which the budget was divided into, namely, accommodation, agriculture, authorities, debt, trade and companies, defence, education, social, veterans, work and miscellaneous expenditures. It was hypothesised that there would be principally 3 groups, namely, pre-world war I period, pre-world war II period, and the post-world war II period, where each of these would be significantly distinct in characteristics from the other.

Methods

The year variable was grouped into 3 epochs for descriptive purposes: the first category from 1872 till 1912 representing the 'pre-world war I era', the second category from 1920 till 1938 representing the 'pre-world war II era', and the third category from 1947 till 1971 representing the 'post-world war II era'.

Principal components analysis (PCA) was undertaken to analyse the data considering its multivariate nature. The individuals or observations here were the years, n = 24, and the 11 indicators were the variables. There were periods of gap years in the budget during the war periods, but for a given year, there was no missing data for all the variables.

Clustering approach – Agglomerative Hierarchical Clustering (AHC) using Ward's method on PCA was used to determine similar groups of individuals. Finally, Classification and Regression Tree (CART) algorithm also was explored with clusters as labels in the supervised learning mode.

R software (CRAN) version 3.3.1 was used to analyse the data.

Results

Principal components analysis

Table 1. Summary statistics of the 11 indicator variables (N = 24).

Variable	Mean	SD	Median	IQR	Min	Max
accom	3.96	4.27	1.85	5.53	0.5	15.8
agri	2.00	1.68	1.40	1.85	0.3	6.0
auth	12.21	2.24	12.60	2.85	7.6	18.0
comp	3.94	4.59	1.30	6.95	0.1	16.5
debt	19.14	12.46	19.30	20.10	3.5	41.6
defns	30.26	7.47	29.15	11.10	18.8	42.4
educ	9.94	5.34	8.70	3.28	2.1	23.8
misc	1.18	1.05	1.40	2.03	0.0	3.0
social	4.82	3.48	4.55	5.00	0.5	11.3

veterns	4.28	4.24	3.80	5.45	0.0	13.4
work	8.32	2.52	8.00	2.48	4.5	15.3

SD = standard deviation; IQR = inter−quartile range.

accom = accommodation; agri = agriculture; auth = authorities; comp = trade and companies; defns = defence; educ = education; misc = miscellaneous; veterns = veterans.

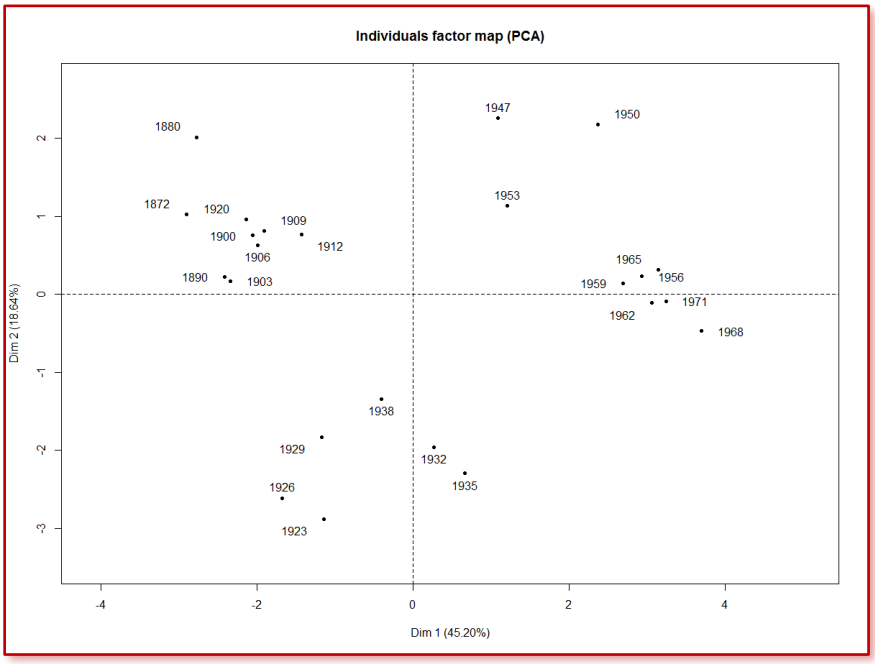

Fig 1. Loadings plot of the years (“individuals”) for the first two dimensions.

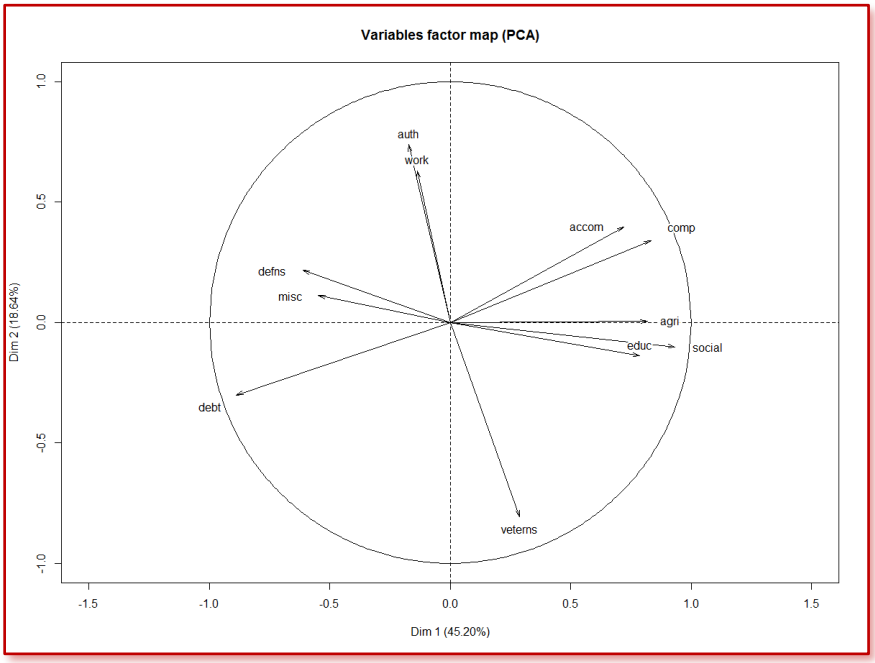

Fig 2. Correlation circle of the indicators (“variables”) for the first two dimensions.

Table 2. Correlation coefficients (cos^2) between the indicators ("variables") and the first two dimensions.

Variable	Dim 1	Dim 2
auth	−0.17	**0.74**
agri	**0.82**	0.01
comp	**0.83**	0.34
work	−0.14	**0.63**
accom	**0.72**	0.40
educ	**0.79**	−0.14
social	**0.93**	−0.10
veterns	0.29	**−0.81**
defns	**−0.61**	0.22
debt	**−0.89**	−0.30
misc	**−0.55**	0.11

Table 3. Variability decomposition (standard deviation) for the principal components.

Principal components	Eigenvalue	% of variance	Cumulative % of variance
Comp 1	4.97	45.20	**45.20**
Comp 2	2.05	18.64	**63.85**
Comp 3	1.29	11.73	**75.57**
Comp 4	0.99	9.03	84.60
Comp 5	0.71	6.44	91.04
Comp 6	0.56	5.07	96.12
Comp 7	0.20	1.86	97.97
Comp 8	0.13	1.14	99.11

Comp 9	0.06	0.57	99.68
Comp 10	0.04	0.32	100.00
Comp 11	0.00	0.00	100.00

Table 4. Contribution (in relative %) of the years ("individuals") to the first two dimensions.

Year	Dim 1	Dim 2
1872	7.05	2.13
1880	6.42	8.23
1890	4.89	0.10
1900	3.54	1.16
1903	4.58	0.06
1906	3.30	0.80
1909	3.05	1.34
1912	1.72	1.20
1920	3.83	1.86
1923	1.10	16.89
1926	2.35	13.85
1929	1.15	6.81
1932	0.06	7.80
1935	0.36	10.71
1938	0.14	3.66
1947	0.99	10.35
1950	4.71	9.62
1953	1.21	2.62
1956	7.18	0.11
1959	6.04	0.04
1962	7.82	0.03
1965	8.27	0.19
1968	11.44	0.45
1971	8.79	0.02

Table 5. Contribution (in relative %) of the indicators ("variables") to the first two dimensions.

Variable	Dim 1	Dim 2
auth	0.60	26.69

agri	13.47	0.00
comp	13.96	5.68
work	0.38	19.40
accom	10.47	7.71
educ	12.45	0.91
social	17.52	0.50
veterns	1.68	31.78
defns	7.54	2.28
debt	15.89	4.43
misc	6.05	0.61

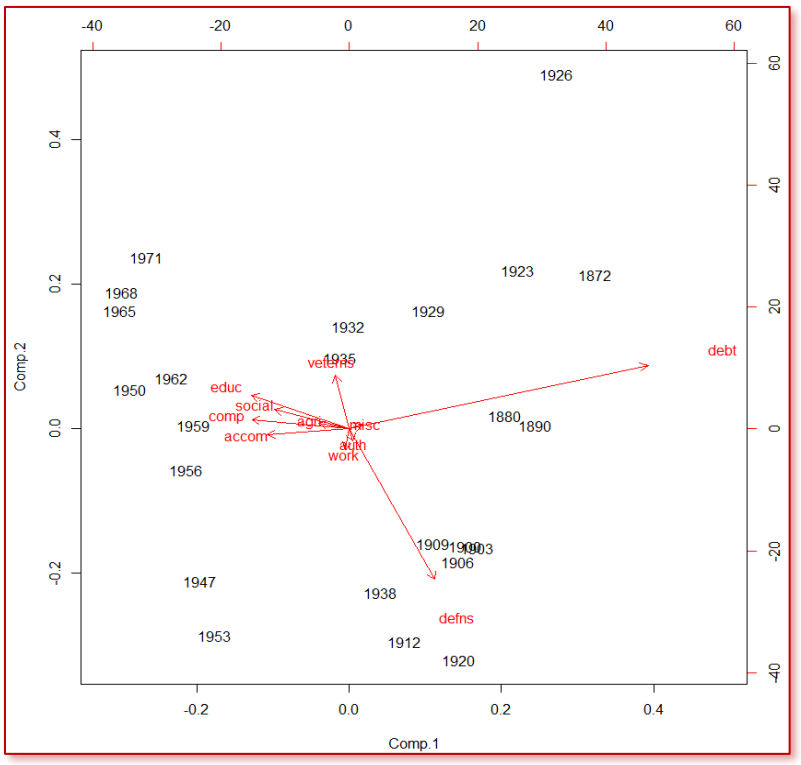

Fig 3. Biplot of "individuals" and "variables" on the same factorial map (a different R-function princomp() was used only for 'visually overlaying' purposes) for the first two dimensions.

Agglomerative Hierarchical Classification (clustering)

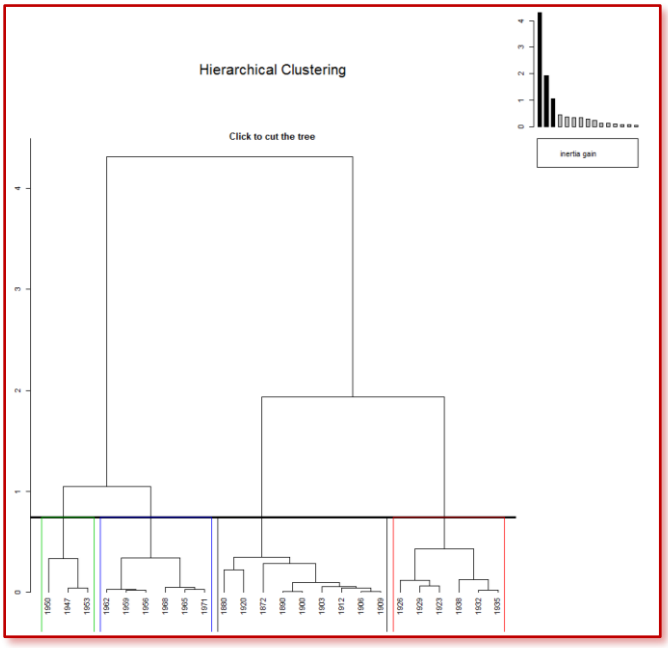

Fig. 4. Hierarchical clustering showing a dendrogram using Ward's approach with 4 clusters selected (= 3 splits).

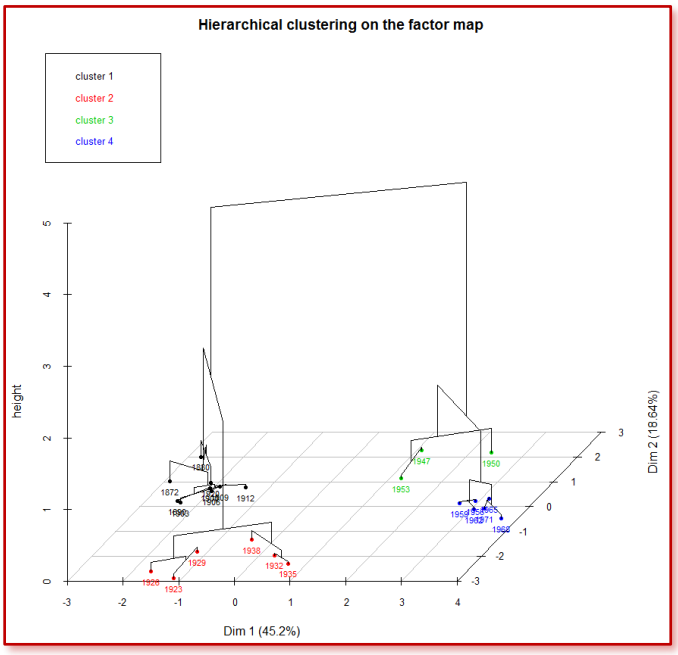

Fig. 5. Hierarchical clustering showing a 3-D dendrogram on the factor map for the first two dimensions.

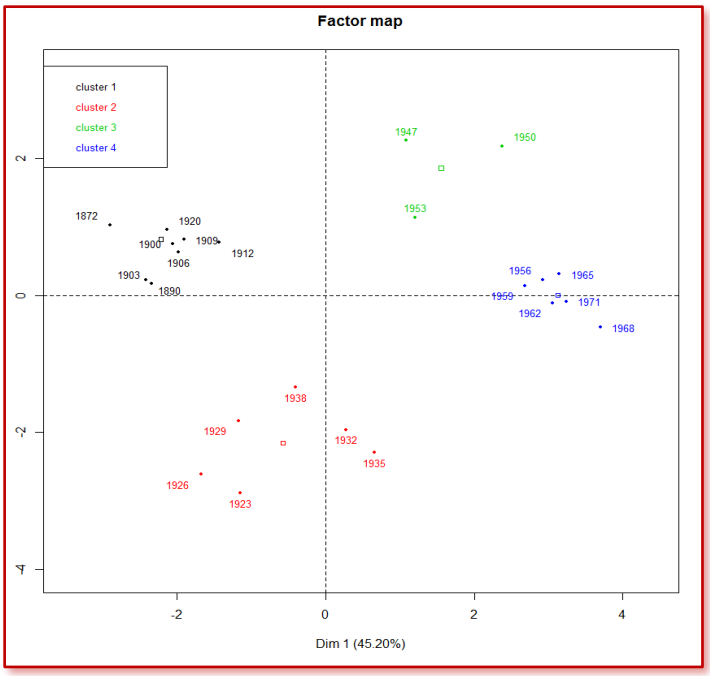

Fig. 6. Hierarchical clustering showing the "individual" years in 4 different colours projected into the first two dimensions.

Classification and Regression Tree (CART)

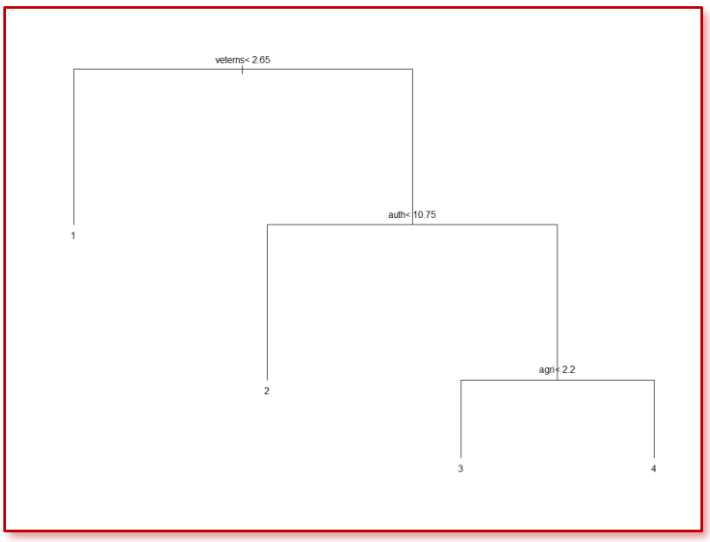

Fig. 7. CART algorithm using clusters as labels.

Table 6. Variable importance ranking based on Gini's criteria.

Variable	Gini's criteria
veterns	13.250000

accom	12.516667
comp	12.516667
debt	10.700000
auth	9.850000
social	9.100000
agri	8.666667
educ	4.000000
work	2.666667

Discussion

In the analysis of the French budget data spanning from 1872 until 1971, consisting of 11 different indicators or variables, showed 3 major axes or dimensions, capturing approximately three quarters of the total variability. It was hypothesised that the years ranging from 1872 until 1912 would form the group comprising the pre-world war I era, then the period between the world war I and II would be explained by the years ranging from 1920 until 1938, and finally the post-world war II era would be explained by the years ranging from 1947 until 1971, which is known as the period of 'glorious expansion' in France.

However, in the current analysis, we observe two exceptions to our hypothesis: first, the year 1920, after the end of the world war I, was 'taken up' by the group prior to the world war I; and second, the post-world war II era was further divided into two groups, although, they were quite close on the factor map of the first two dimensions – the first smaller group comprised of the years – 1947, 1950 and 1953, and the second smaller group comprised of the years 1956 until 1971.

About 75% of the variability was captured in the first three dimensions, and another 9%, meaning 84% of the variability was capture in the first four dimensions. The first dimension, which is often referred to as the *size* variable, derived from a linear combination of the 11 original variables, gives an inherent idea of the major data inertia in this direction. It showed good correlation with agriculture, trade and companies, accommodation, education and social welfare expenditures, and inversely correlated with defence, debt and miscellaneous expenditures. The second dimension, and the following ones, are often known as *shape* variables, is/are conditional on the first/previous dimensions; they represent contrasting characteristics between different components adding refinement to the mean inertia direction. This dimension correlated with authorities and services, work, and inversely correlated with veterans' expenditures. The third dimension correlated weakly, but statistically significantly with miscellaneous and education expenditures.

Interestingly, we can observe that the veterans' expenditure rose in the late post-world war I era from 1932 till 1938, which could partly be due to the increased post-world war I pension invalidity. Secondly, although 4 principal components were needed to explain 84% of the variability in the data, the 4 clusters appear to be well demarcated in the first two-dimensional factorial map, and again in a supervised classification using CART algorithm, it was possible to perfectly separate the 4 clusters (*cf.* below).

In the current analysis PCA and AHC using Ward's method were employed for unsupervised analysis. As future perspectives, *k*-means cluster analysis (*k* determined by cross-validation, compare with AHC), exploratory factor analysis (EFA) (compare with PCA, as factor analysis uses Singular Value Decomposition (SVD) instead of eigenvalue-eigenvector decomposition), and Fisher's linear discriminant analysis using the clusters could be explored to get more insight.

Finally, a supervised approach was also explored, and additionally to rank the variables importance based on the AHC-derived clusters. Interestingly, defence and miscellaneous expenditures were excluded in the variables importance list. The most important factor was found to be veterans, followed by accommodation, trade and companies, and debt. The CART algorithm perfectly classified the observations into the clusters with the help of only 3 splits and 3 variables, which were involved to build the CART model.

References

1. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis, 6th ed., 2008.
2. Lê S, Josse J, Husson F. An R package for multivariate analysis. *J Stat Software*, 2008;**25**(1):1-18.
3. Comprehensive R Archive Network (CRAN); https://www.r-project.org/.

Appendix

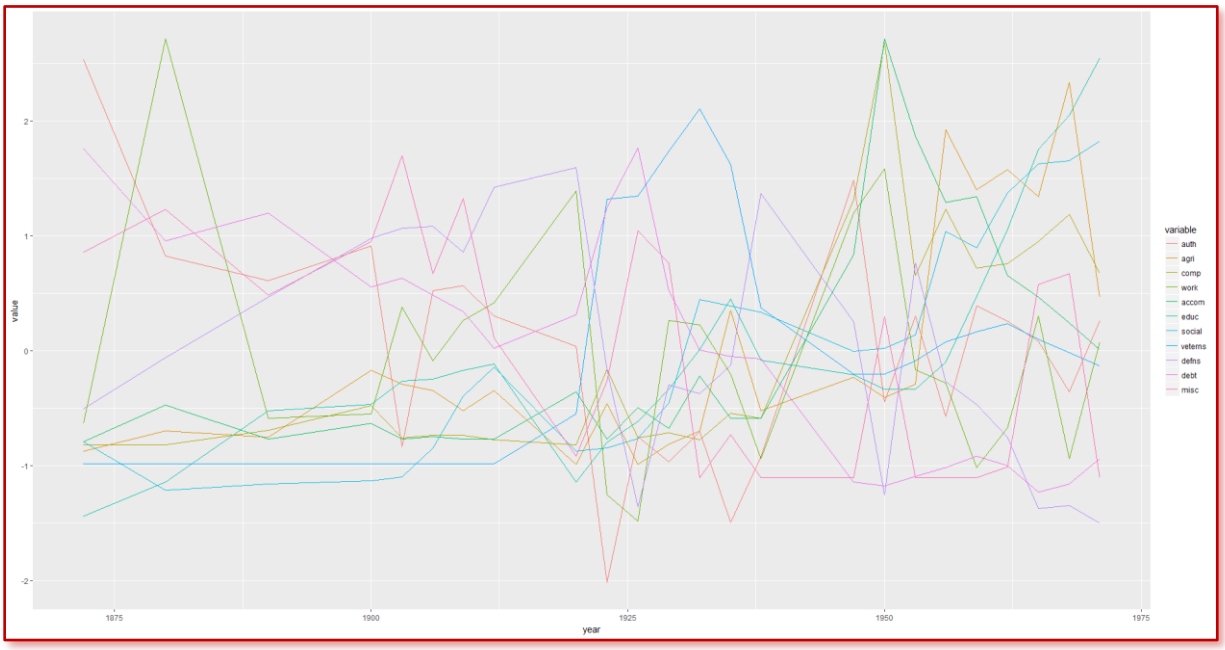

Fig. 8. Evolution of the budget indicators (corrected z−score) over the years.

Table 7. Summary statistics by 'clusters' of years of the 11 indicator variables.

Variable	Year group	N	Mean	SD	Median	IQR	Min	Max
auth	**1872–1912**	8	13.76	2.12	13.55	0.87	10.3	18.0
	1920–1938	7	9.99	1.48	10.10	1.15	7.6	12.3
	1947–1953	3	13.23	2.22	12.90	2.20	11.2	15.6
	1956–1971	6	12.23	0.88	12.60	1.15	10.9	13.1
agri	**1872–1912**	8	1.14	0.43	1.25	0.65	0.5	1.7
	1920–1938	7	0.99	0.79	0.80	0.70	0.3	2.6
	1947–1953	3	1.47	0.15	1.50	0.15	1.3	1.6
	1956–1971	6	4.58	1.08	4.55	0.83	2.8	6.0
comp	**1872–1912**	8	0.54	0.51	0.45	0.30	0.1	1.7
	1920–1938	7	1.03	1.07	0.60	0.95	0.1	3.2
	1947–1953	3	11.20	4.84	10.10	4.75	7.0	16.5
	1956–1971	6	8.25	1.14	7.95	1.88	7.1	9.7
work	**1872–1912**	8	8.94	2.81	8.55	2.45	6.7	15.3
	1920–1938	7	7.59	2.61	7.80	3.45	4.5	11.9
	1947–1953	3	10.57	2.36	11.40	2.25	7.9	12.4
	1956–1971	6	7.23	1.40	7.10	2.20	5.7	9.1
accom	**1872–1912**	8	0.84	0.48	0.60	0.23	0.5	1.9
	1920–1938	7	1.66	0.82	1.40	0.90	0.6	3.0
	1947–1953	3	11.83	4.11	12.10	4.10	7.6	15.8
	1956–1971	6	6.87	2.39	6.40	3.65	4.0	9.8
educ	**1872–1912**	8	6.96	2.65	7.95	2.45	2.1	9.3
	1920–1938	7	7.99	2.94	8.10	3.65	3.7	12.4
	1947–1953	3	8.33	0.40	8.10	0.35	8.1	8.8
	1956–1971	6	17.00	5.46	17.60	7.40	9.4	23.8
social	**1872–1912**	8	1.80	1.39	1.35	1.58	0.5	4.3
	1920–1938	7	3.91	2.20	3.20	4.15	1.7	6.4
	1947–1953	3	5.00	0.26	4.90	0.25	4.8	5.3
	1956–1971	6	9.80	1.31	10.15	1.88	8.0	11.3
veterns	**1872–1912**	8	0.00	0.00	0.00	0.00	0.0	0.0
	1920–1938	7	9.20	3.97	10.10	3.60	1.9	13.4
	1947–1953	3	3.57	0.29	3.40	0.25	3.4	3.9
	1956–1971	6	4.58	0.57	4.65	0.62	3.7	5.3
defns	**1872–1912**	8	35.31	4.97	37.25	5.63	26.4	41.1
	1920–1938	7	30.96	7.92	29.00	7.30	19.9	42.4
	1947–1953	3	29.67	8.01	32.20	7.70	20.7	36.1
	1956–1971	6	23.00	4.00	22.25	6.30	18.8	28.2
debt	**1872–1912**	8	28.60	6.93	26.70	7.23	19.4	41.5

	1920–1938	7	25.91	9.09	23.10	11.55	18.2	41.6
	1947–1953	3	4.67	0.50	4.60	0.50	4.2	5.2
	1956–1971	6	5.87	1.59	6.30	2.15	3.5	7.5
misc	**1872–1912**	8	2.16	0.54	2.15	0.68	1.3	3.0
	1920–1938	7	0.83	0.96	0.40	1.35	0.0	2.3
	1947–1953	3	0.50	0.87	0.00	0.75	0.0	1.5
	1956–1971	6	0.63	0.94	0.05	1.38	0.0	1.9

SD = standard deviation; IQR = inter−quartile range.