II. Attempt of Quantitative Analysis

II-1. Materials (Excel dataset)

The dataset analyzed here is from Berg-Schlosser D. and De Meur (1994), “Conditions of Democracy in Interwar Europe. A Boolean test of major hypotheses.” This dataset was also used in Benoit Rihoux and Charles Ragin (2009) “CONFIGURATIONAL COMPARATIVE METHOD: Qualitative Comparative Analysis (QCA) and Related Techniques” to introduce csQCA, mvQCA, and fsQCA. An additional item, the stability of the regime, has been added to the data. The purpose of the analysis by Berg-Schlosser and De Meur (1994) was to test the hypothesis proposed by Lipset (1960) that “modernization fosters democracy.” This commentary also discusses the effectiveness of the analysis method with this hypothesis in mind.

Table 1 shows the dataset to be analyzed. The subjects of analysis are the 18 countries in Europe that existed from the end of World War I to the beginning of World War II (1919-1939): Austria, Belgium, Czechoslovakia, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, the Netherlands, Poland, Portugal, Romania, Spain, Sweden, and the United Kingdom. The data items include GNP: per capita GNP in 1930, Urbanization: the percentage of the population living in cities with more than 20,000 people, Education: literacy rate (%), Industrial population: the percentage of factory workers in the total labor force, Stability (instability): the number of regime changes over the 20-year interwar period, and the degree of democracy preservation (dependent variable) with an evaluation score ranging from -10 to 10 (Berg-Schlosser and Mitchell (2000, 2003)).

Table 1 Analyzed dataset (Excel dataset)

		GNP/	Urbani	Litera	INDLAB	GOVST	SURVIV
country	Case Id	A	B	C	D	E	R
Austria	AUT	720	33.4	98	33.4	10	-9
Belgim	BEL	1098	60.5	94.4	48.9	4	10
Czech Slovakia	CZE	586	69	95.9	37.4	6	7
Estonia	EST	468	28.5	95	14	6	-6
Finland	FIN	590	22	99.1	22	9	4
France	FRA	983	21.2	96.2	34.8	5	10
Germany	GER	795	56.5	98	40.4	11	-9
Greece	GRC	390	31.1	59.2	28.1	10	-8
Hungary	HUN	424	36.3	85	21.6	13	-1
Ireland	IRL	662	25	95	14.5	5	8
Italy	ITA	517	31.4	72.1	29.6	9	-9
Netherland	NLD	1008	78.8	99.9	39.3	2	10
Poland	POL	350	37	76.9	11.2	21	-6
Portugal	PRT	320	15.3	38	23.1	19	-9
Romania	ROU	331	21.9	61.8	12.2	7	-4
Spain	ESP	367	43	55.6	25.5	12	-8
Sweden	SWE	897	34	99.9	32.3	6	10
United Kingdom	UK	1034	74	99.9	49.9	4	10

The analysis data (Table 2) was standardized by subtracting the mean from each data item and then dividing by the standard deviation to make the impact of individual data items on the results easier to understand. For the political stability of E, the sign of the government was reversed (Excel standardize).

Using this data, correlation analysis, multidimensional scaling, cluster analysis, principal component analysis, factor analysis, and regression analysis were performed.

Table 2 Standardized dataset (Excel standardized)

Case Id	A	B	C	D	E	R
AUT	0.30	-0.35	0.73	0.40	-0.24	-1.12
BEL	1.75	1.09	0.54	1.75	0.98	1.24
CZE	-0.21	1.54	0.62	0.75	0.58	0.87
EST	-0.66	-0.61	0.57	-1.28	0.58	-0.74
FIN	-0.20	-0.95	0.79	-0.59	-0.03	0.50
FRA	1.31	-0.99	0.63	0.52	0.78	1.24
GER	0.59	0.88	0.73	1.01	-0.44	-1.12
GRC	-0.96	-0.47	-1.36	-0.06	-0.24	-0.99
HUN	-0.83	-0.19	0.03	-0.62	-0.85	-0.12
IRL	0.08	-0.79	0.57	-1.24	0.78	0.99
ITA	-0.47	-0.45	-0.66	0.07	-0.03	-1.12
NLD	1.40	2.06	0.83	0.91	1.39	1.24
POL	-1.11	-0.16	-0.40	-1.53	-2.47	-0.74
PRT	-1.23	-1.30	-2.49	-0.49	-2.06	-1.12
ROU	-1.19	-0.95	-1.22	-1.44	0.37	-0.50
ESP	-1.05	0.16	-1.55	-0.29	-0.64	-0.99
SWE	0.98	-0.31	0.83	0.30	0.58	1.24
UK	1.50	1.80	0.83	1.83	-0.98	1.24

II-2. Comprehensive Correlation Analysis (R: line 20-25)

Table 3 shows the results of the comprehensive correlation analysis . The calculations were performed using R (line 20-25). The correlation coefficients with the outcome variable R are significant for all data items. In the case of comprehensive correlation, the correlation coefficients are calculated for all combinations (6x5), so when testing for statistical significance, it is necessary to correct for multiple comparisons (e.g., Bonferroni correction). As a result, the significance of the correlations between B, D, and R is statistically denied, but we are not trying to use this result for any prediction. In this case, statistical testing is not very meaningful. Rather, the fact that there was a non-negligible correlation between B, D, and R, but that correlation was lower compared to the correlations between A, C, E, and R, might be more important. Keeping this in mind, we proceed to the next analysis.

Table 3. Correlations between items

Correlation

	A	B	C	D	E	R
A	1.00	0.56	0.73	0.80	0.69	0.74
B	0.56	1.00	0.44	0.72	0.41	0.41
C	0.73	0.44	1.00	0.41	0.62	0.63
D	0.80	0.72	0.41	1.00	0.47	0.44
E	0.69	0.41	0.62	0.47	1.00	0.69
R	0.74	0.41	0.63	0.44	0.69	1.00

Prob(t≧1)

	A	B	C	D	E	R
A	0.00	0.11	0.01	0.00	0.02	0.01
B	0.02	0.00	0.35	0.01	0.35	0.35
C	0.00	0.07	0.00	0.35	0.05	0.05
D	0.00	0.00	0.09	0.00	0.28	0.35
E	0.00	0.10	0.01	0.05	0.00	0.02
R	0.00	0.09	0.01	0.07	0.00	0.00

Entries above the diagonal are adjusted for multiple tests.

II-3. Distance Matrix (Excel dist) and Spatial Relationships of Data (MDS: R line65-74)

When we directly apply this data to regression analysis, issues with multicollinearity will likely prevent successful analysis. In such cases, categorizing samples (in this case, countries) based on their similarities or differences can help simplify the discussion. There are various methods for categorization, but one straightforward approach is to represent the differences between each country as distances and create a distance matrix. The issues here are the differences in variance between items (differences in the scales of the data represented) and the presence of correlations. Since we are already using standardized data, there is no need to correct for differences in variance between items, but we must consider the impact of correlations between items. Ignoring this would result in evaluating changes in a certain direction redundantly. Therefore, we need to perform a transformation that makes the correlations between data zero. In linear algebra terms, this is called diagonalizing the correlation matrix. When diagonalized, the off-diagonal elements (correlations) of the correlation matrix become zero, meaning the transformed axes are all orthogonal. The distance obtained by transforming the data matrix into an orthogonal matrix is called the Mahalanobis distance (Figure 1). In middle school, we learned that the square of the hypotenuse of a right triangle is equal to the sum of the squares of the other two sides. Extending this to multidimensional space, the square of the distance between two points in multidimensional space is:

$$d^{2} = \boldsymbol{uu^{t}} = u_1{}^2 + u_2{}^2 + \cdots$$

The distance is

$$d = \sqrt{\boldsymbol{uu^{t}}}$$

$$\boldsymbol{u} \text{ : } vector\ of\ diffeerenc\ between\ \ same\ item(u_1 \quad u_2 \quad \cdots)$$

After all, it is the Pythagorean theorem, so this calculation implicitly assumes that each item is orthogonal and uncorrelated. This is called the Euclidean distance. Since actual data has correlations between items, we must transform the axes between items to orthogonalize them before calculating the distance. The formula for Mahalanobis distance is…

$$md = \sqrt{\boldsymbol{ur}^{-1}\boldsymbol{u}^{t}}$$

$$\boldsymbol{r}\text{: } \boldsymbol{corelation\ matrix}$$

$$\boldsymbol{u} \text{ : } vector\ of\ diffeerenc\ between\ \ same\ item(u_1 \quad u_2 \quad \cdots)$$

How to derive the formula for Mahalanobis distance is shown in Figure 1. In summary, the formula for finding the square of the Euclidean distance involves multiplying the vector of differences between items by its transpose. However, if we insert the inverse of the correlation matrix between the vector and its transpose, and perform matrix operations, we get the square of the Mahalanobis distance.

The formula itself is simple. Calculating distances exhaustively is cumbersome even for Euclidean distance. When I tried to calculate it in R and looked up the function "dist" that creates a distance matrix, for some reason, Mahalanobis distance was not an

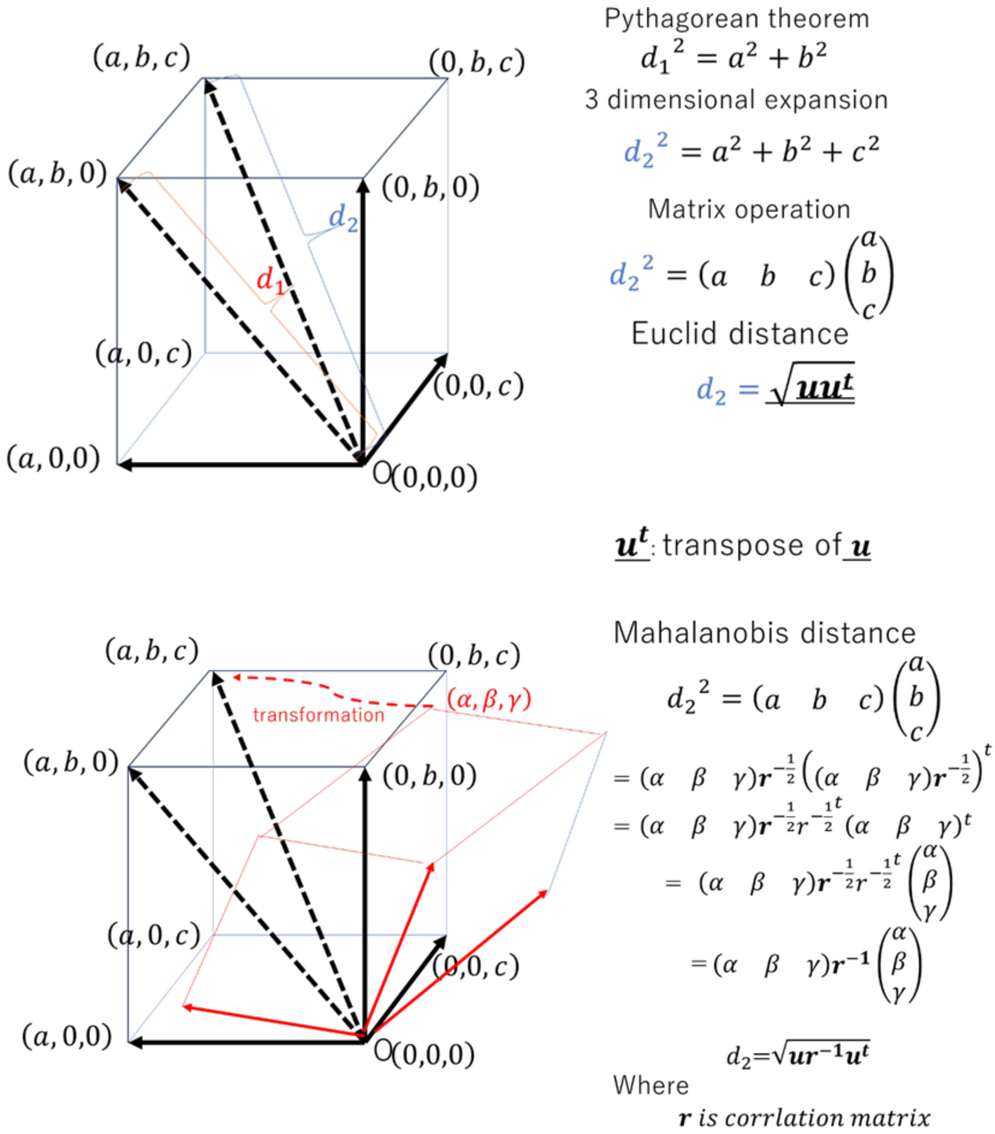

Fig.1 Derivation of the formula of Mahalanobis distance

option. There is a function called "mahalanobis" to calculate Mahalanobis distance, but it does not create a distance matrix exhaustively. If there are 18 countries, we have to repeat the calculation 18×17 times. It might be easier to do it in Excel. So, I created a distance matrix in Excel. Since I calculated the data matrix all at once, the number of calculations was reduced to 18. I left the process in Excel dist. It can be used as a template for calculating Mahalanobis distance and also helps in understanding Mahalanobis distance.

The distance matrix is shown in Table 4. Even when looking at the distance matrix, it is not clear how to interpret it. However, by organizing the combinations with close

Table 4. Distant matrix in Mahalanobis distance

Distant matrix

	AUT	BEL	CZE	EST	FIN	FRA	GER	GRC	HUN	IRL	ITA	NRL	POL	PRT	ROU	ESP	SWE
BEL	2.86																
CZE	3.07	3.77															
EST	2.85	3.83	2.99														
FIN	1.26	3.44	3.23	1.88													
FRA	2.50	2.37	4.78	3.54	2.62												
GER	1.52	2.47	2.69	3.36	2.42	3.28											
GRC	2.78	3.05	2.88	3.01	2.93	3.42	3.04										
HUN	1.75	3.34	2.44	2.09	1.66	3.53	1.77	2.51									
IRL	3.39	3.38	4.27	1.80	2.62	2.86	3.77	3.61	2.97								
ITA	2.01	2.61	2.71	2.56	2.19	2.76	2.46	0.83	2.02	3.14							
NLD	4.32	2.57	3.86	3.67	4.35	4.18	3.59	4.09	3.78	3.10	3.85						
POL	4.09	4.61	4.95	4.21	4.07	4.96	3.48	4.83	3.03	3.98	4.51	4.28					
PRT	3.65	3.51	4.98	4.59	4.05	3.64	3.63	2.74	3.48	4.34	2.83	4.79	4.03				
ROU	3.94	3.70	3.91	2.29	3.39	3.72	4.24	2.39	3.18	2.25	2.47	3.53	4.70	3.68			
ESP	3.34	2.93	3.02	3.13	3.47	3.91	2.99	1.56	2.44	3.40	1.86	3.18	3.88	2.53	2.22		
SWE	1.85	1.94	3.79	2.60	1.88	1.30	2.34	3.00	2.47	2.10	2.27	3.23	3.99	3.53	3.18	3.18	
UK	2.76	1.27	2.73	3.56	3.35	3.25	1.95	2.97	2.90	3.60	2.58	2.27	4.44	3.99	3.80	2.72	2.41

Summary

Asending order				Swscening order		
From	To	dist		From	To	dist
ITA	GRC	0.83		PRT	CZE	4.98
AUT	FIN	1.26		POL	FRA	4.96
BEL	UK	1.27		POL	CZE	4.95
SWE	FRA	1.30		POL	GRC	4.83
GER	AUT	1.52		PRT	NLD	4.79
ESP	GRC	1.56		FRA	CZE	4.78
HUN	FIN	1.66		ROU	POL	4.70
HUN	AUT	1.75		POL	BEL	4.61
HUN	GER	1.77		PRT	EST	4.59
IRL	EST	1.80		POL	ITA	4.51

distances and those with far distances, as in the summary, the results become easier to understand. In such cases, Excel's "Sort" function is useful. If the conditions are correctly selected and the conditions are the same, it is expected that the results will be the same. Among the combinations with similar conditions, specifically, within the top 10 combinations (3.3% of the total) in terms of proximity, the combinations that resulted in the same outcome were Italy-Greece, Germany-Austria, Spain-Greece, Hungary-Austria, Hungary-Germany, Belgium- British Empire, and Sweden-France. The combinations that resulted in different outcomes were Austria-Finland, Hungary-Finland, and Ireland-Estonia. The fact that 7 out of the 10 combinations resulted in the same outcome suggests that the conditions A, B, C, D, and E, selected as factors leading to the maintenance or collapse of democracy, have a certain degree of validity overall. Upon closer examination, within the countries where democracy collapsed, a strong similarity can be seen in the cluster of Germany, Austria, and Hungary. These countries are geographically close to Finland, yet the outcomes were different. In other words, despite Finland having conditions that could have led to the collapse of democracy, some factors allowed democracy to be preserved. Finland, situated between the Soviet Union and Germany, managed to maintain its independence through various strategic and diplomatic skills despite being invaded. This might also be related to the maintenance of democracy.

Similarly, Ireland and Estonia, which had different outcomes, share the commonality of having gained independence from foreign rule during the interwar period. Therefore, it is believed that both countries were delayed in urbanization and industrialization. Numerically, it can be said that their governments were relatively stable, but there must have been various destabilizing factors shortly after gaining independence. In this context, a comparative study is needed to determine what factors contributed to the preservation of democracy in Ireland.

When looking at combinations of countries with distant conditions, 5 out of the 10 combinations with the greatest distance resulted in the same outcome. Among these, 4 combinations were of countries where democracy collapsed. If democracy collapses even when all five conditions are not similarly positioned in the 10 combinations with different conditions, it suggests that the outcome might have been influenced by fewer conditions or by differences in conditions not shown in the data. Particularly, since 4 out of the 5 combinations resulted in the collapse of democracy, it indicates that the collapse of democracy might have been caused by a few specific conditions rather than a combination of many conditions.

This type of analysis is called MDSO/MSDO. MDSO/MSDO stands for Most Different, Similar Outcome/Most Similar, Different Outcome. MDSO/MSDO is used in the analysis of survey data, and distances such as Hamming distance, which is used in information theory, are employed. To understand MDSO/MSDO, which is considered one of the methods of QCA, I have demonstrated that similar analysis is possible even with continuous numerical data.

Next, to recognize the overall differences and similarities in the conditions, each country which placed in the five-dimensional distance matrix was visualized on a two-dimensional plane (Figure 2). The method used was Multi-dimensional Scaling (MDS) (R script lines 65-74). MDS is a technique that represent multi-dimensional positional relationship in a two-dimensional or three-dimensional positioning. It is often used in the analysis of questionnaire survey data and ecological studies.

The vertical and horizontal axes of MDS have no inherent meaning, so when viewing MDS, one rotates the figure to interpret it. When slightly rotated unticlockwise, it appears that countries where democracy collapsed and those where it was maintained are biased towards the upper and lower parts, respectively. In the upper part, countries where democracy collapsed, such as Poland, Romania, Hungary, Spain, Italy, Greece, and Portugal, are located. In the lower part, countries where democracy was maintained, such as the Netherlands, Sweden, France, the United Kingdom, and

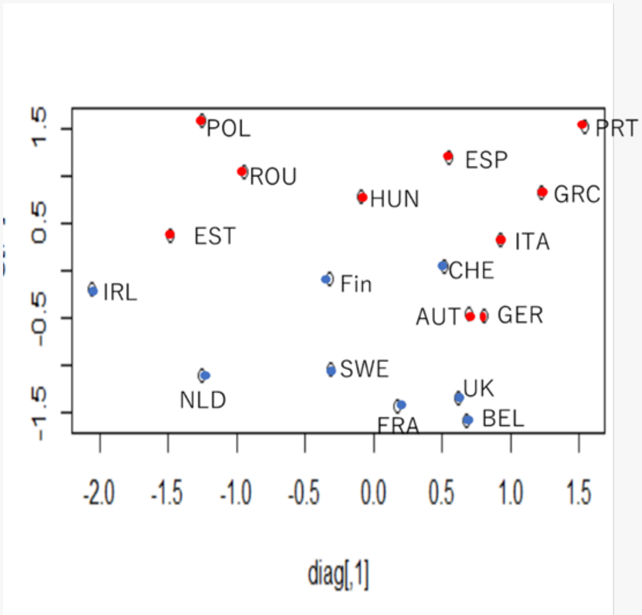

Fig.2 Positional relationship among countries drown by MDS

Belgium, are located. In the middle zone, Ireland, Estonia, Finland, the Czechoslovakia, Germany, and Austria are located. This middle zone includes three countries where democracy was maintained and three where it collapsed.

Excluding the countries in the middle zone, it is easy to analyze the factors for the collapse and maintenance of democracy in those countries, and it is likely that the validity of Lipset's (1960) hypothesis will be verified. The issue here is why democracy was maintained in three countries in the middle zone and collapsed in the other three. The writer of this practical guide is neither a sociologist nor a political scientist, so this analysis cannot be performed. However, a suitable sociologist might be able to draw some conclusions from a detailed comparison among these six countries. The appropriate direction for analysis would be to find differences between Czechoslovakia, Finland, and Germany, Austria in dimensions other than wealth, education, and political stability, and to reanalyze by adding those differences.

II-4. Principal Component Analysis (R line32-62)

Although Mahalanobis distance is also a form of orthogonalization, calculating the distance matrix is cumbersome, and subsequent cluster analysis did not work very well (perhaps because the commentator did not know how to write the R script). Therefore, as a quick method of orthogonalization, principal component analysis (PCA) is performed. PCA involves the diagonalization of the variance-covariance matrix or the correlation matrix, and can be used for clustering based on similarity.

Table 5 summarizes the results of the PCA. Since the data is standardized by standard deviation, the sum of the eigenvalues (total variance) is 5. This is because there are five items with a variance of 1, making the total variance 5. Of this, the first principal component accounts for 66.8% of the variance, and the second principal component accounts for 17.4%. Cumulatively, these two principal components account for over 80% of the total, with the other principal components being minor. Looking at PC1, the loadings for each item are all negative and the lengths of the arrows are similar. The variations in all measurement items are associated with some specific directional variation.

Table 5. Summary of PCA (R script 24-28)

Eigenvalue

	PC1	PC2	PC3	PC4	PC5
variance	3.338282	0.870646	0.396127	0.330964	0.063982
porportion	0.667656	0.174129	0.079225	0.066193	0.012796
cum Prop.	0.667656	0.841785	0.921011	0.987204	1

Nearly 70%

Loading

Same direction

	PC1	PC2	PC3	PC4	PC5
A	-0.51346	0.140817	-0.16802	-0.42142	0.714636
B	-0.39561	-0.62646	0.12747	0.616873	0.232942
C	-0.43968	0.406381	-0.62819	0.380992	-0.319
D	-0.46505	-0.43521	0.006496	-0.53027	-0.55955
E	-0.41263	0.482873	0.748904	0.124637	-0.14205

Table 6. Interpretation of the first and second principle components

PC1

A(-0.51): not rich →poor
B(−0.40) not urbanized
C(-0.44) low literacy→ not educated
D(−0.47) not industrialized
E(−0..41) unstable government

→ **Underdevelopment** ?

PC2

B(−0.63) not urbanized
C(0.41) high literacy→ educated
D(−0.44) not industrialized
E(−0..48) stable government

→ **Advanced Agricultural country** ?

Using the items with large loadings for the first and second principal components, Table 6 presents an analysis of what these principal components represent. For the first principal component, the items are poor, non-urbanized, low education level, non-industrialized, and political instability. In the correlation analysis, A (wealth—poverty: opposite signs) showed a high correlation with the other four items. Therefore, while the first principal component can be considered as related to poverty, the fifth principal component is clearly related to wealth. Given its small variance, the fifth principal component can be ignored. However, if the fifth principal component represents wealth, the first principal component can be seen as representing a more comprehensive social structure. Considering Lipset's (1960) hypothesis, the first principal component was named backwardness (anti-modernity).The second principal component, similar to the first, has negative loadings for B:urbanization and D:industrialization, but positive loadings for C: literacy rate, and E:political stability. This suggests an image of a wealthy and stable agricultural country. Therefore, the second principal component was named agricultural. The third principal component has a large loading for E and is considered to represent political stability. The fourth principal component represents urbanization.

Figure 3 is a scatter plot of various countries using these principal component scors. In this figure, red arrows indicate the direction of the vectors represented by each data item. In the PC1-PC2 plot, the wealth of country A is almost 180 degrees opposite to the positive direction of PC1. Other items also point in the opposite direction to PC1. The group of Ireland, Finland, Czechoslovakia, Estonia, Austria, and Germany is enclosed by a yellow line in the PC1-PC2 scatter plot. To the right of this group, all

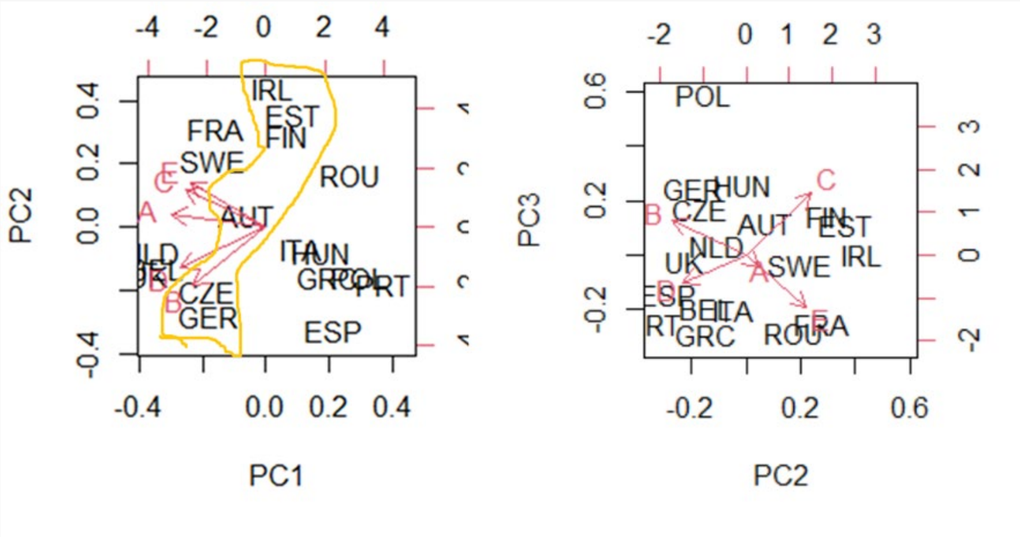

Fig. 3.Scatter diagram PC-PC2, PC2-PC3

countries are those where democracy has collapsed, while to the left, all countries have maintained democracy. Here, the fact that this intermediate zone is slightly tilted clockwise and diagonal will have important implications in later analysis. In any case, this analysis partially supports Lipset's hypothesis that something (modernization?) including wealth is a factor related to the maintenance of democracy. In principal component analysis, it is often the case that the first principal component extracts an unclear component that comprehensively represents the multifaceted distribution characteristics of the data set. How to interpret this often troubles researchers. At present, the commentator can only say that "the first principal component is something that comprehensively relates to wealth, education level, political stability, etc." Whether to describe the first principal component as anti-modernization following Lipset (1960) or to use other terms should be determined by the analyst based on past research cases, experience, etc. In any case, it is not possible to clarify the factors that caused the differences between Ireland, Finland, Czechoslovakia, Estonia, Austria, and Germany with the first and second principal components.

Figure 4 shows the results of cluster analysis using principal component scores. Since the principal components are orthogonal, the distance was measured using Euclidean distance, and the clustering method used was Ward's method. In the large cluster on the right, mainly composed of countries where democracy has collapsed, one sub-cluster, enclosed by a red line, is formed by Finland and Ireland together with Estonia. Similarly, in the large cluster on the left, mainly composed of countries that have maintained democracy, one sub-cluster, enclosed by a blue line, is formed by Austria

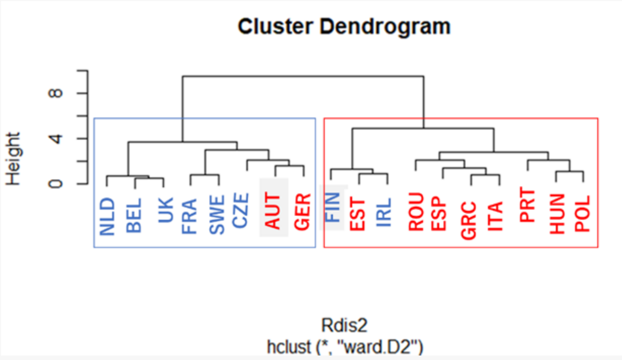

Fig.4. Cluster dendrogram using principle component scores

and Germany together with the Czech Republic. This dendrogram represents the perspective that among the countries with a high possibility of democratic collapse, Finland and Ireland are unique, and among the countries with a possibility of maintaining democracy, Germany and Austria are unique.

II-5. Regression Analysis (R line 76-121)

Based on the characteristics of the data distribution as described above, regression analyses are conducted. The dependent variable is set as R. Two types of regression analyses are performed: one (Regression Analysis 1) using standardized data items A, B, C, D, and E as independent variables, and another (Regression Analysis 2) using the principal component scores of each principal component as independent variables. The data list targeted by Regression Analysis 1 is shown in Table 2, while the data list analyzed in Regression Analysis 2 is shown in Table 7. The results of Regression Analysis 1 are as follows:

Results of regression analyses are as follows

Simple linear regression

	PR(>t)	variance ratio
R=0.7399A	0.000448**	0.547
R=0.4072B	0.0935	0.166
R=0.6269C	0.00536**	0.393
R=0.4360D	0.0765	0.190
R=0.7399E	0.00158**	0.474

Multiple regression

R=0.4240A+0.1077C+0.371E 0.610

PR(>t) 0.128 0.672 0.200

$$variance\ ratio = 1 - \frac{variance\ of\ error}{total\ variance}$$

Table 7. Principle component scores

PC score

Standardized R

	score					
ID	PC1	PC2	PC3	PC4	PC5	R
AUT	-0.46959	0.147189	-0.91418	-0.42496	-0.24998	-1.11631
BEL	-2.64316	-0.50358	0.387021	-0.27646	0.187727	1.240347
CZE	-1.46529	-0.80074	0.326466	0.887345	-0.58806	0.868243
EST	0.686033	1.414189	0.152967	0.884029	-0.22566	-0.74421
FIN	0.515894	1.14744	-0.72665	0.14239	-0.20949	0.496139
FRA	-1.22524	1.225289	-0.09471	-1.1583	-0.00669	1.240347
GER	-1.38023	-1.1128	-0.9795	-0.01534	0.017861	-1.11631
GRC	1.562003	-0.62098	0.711752	-0.47955	-0.23846	-0.99228
HUN	1.369426	-0.30518	-0.69783	0.487664	-0.04962	-0.12403
IRL	0.214302	1.752321	0.153952	0.419581	0.298637	0.992278
ITA	0.881597	-0.28057	0.502764	-0.5062	-0.21354	-1.11631
NLD	-2.77554	-0.30509	0.472418	0.407505	0.303656	1.240347
POL	2.291442	-0.63537	-0.81251	0.657366	0.446106	-0.74421
PRT	2.87268	-0.72202	-0.20397	-0.84476	0.03812	-1.11631
ROU	2.055549	0.665163	1.245056	0.103768	0.112798	-0.49614
ESP	1.663234	-1.2918	0.49379	0.0945	0.16194	-0.99228
SWE	-1.27245	0.836212	-0.26645	-0.44917	0.117729	1.240347
UK	-2.88066	-0.60968	0.249597	0.070591	0.096918	1.240347

Since standardized data is used, there is no constant term (intercept = 0). To compare the explanatory power of each explanatory variable, the variance ratio was shown. The variance ratio is the ratio of the variance that can be explained by the regression equation to the total variance. Using Excel, it was calculated with the formula 1 - (error variance / total variance). This calculation process is left in the Excel sheet “V ratio”. The variable with the largest absolute coefficient is the richness of A, which explains more than 50% of the total variance including error variance and is statistically highly significant. The next largest absolute coefficient is the political stability of E, which explains more than 45% and is statistically significant. The next largest coefficient is the literacy rate of C, which explains nearly 40% of the total. Adding these together, the three items explain 140% of the total. This is because the explanatory power is counted redundantly due to the correlation between factors. When performing multiple regression analysis with A, C, and E as explanatory variables, only the coefficient of A is significant, and the others are not significant. This variance ratio is 0.610, and the explanatory power of this regression equation is 60% of the total variance. Compared to the sum of the explanatory power calculated by simple regression, which is 140%, it is extremely small. By multiple regression, the redundant count due to correlation is reduced by 80%. This indicates that in simple regression, the explanatory power was calculated redundantly two or three times.

To prevent double counting of explanatory power due to redundancy, the results of Regression Analysis 2 (R script line 78-100) using orthogonalized principal components as explanatory variables through principal component analysis are presented.

Simple linear regression

	PR(>t)	variance ratio
$R=-0.4040PC1$	0.000805***	0.514
$R=0.3806PC2$	0.161	0.119
$R=0.2297PC3$	0.578	0.018
$R=0.1131PC4$	0.803	0.004
$R=0.7964PC5$	0.436	0.038

Multiple regression

$R=-0.404PC1+0.3906PC2+0.7963PC5$ 0.672

PR(>t) 0.00035 *** 0.04065* 0.22164

$$variance\ ratio\ = 1 - \frac{variance\ of\ error}{total\ variance}$$

The coefficient of the first principal component is negative. This is because the first principal component positively represents negative trends such as backwardness or poverty. This equation alone can explain more than half (0.514) of the variance. The second principal component has a smaller explanatory power, with a variance ratio of 0.119. The explanatory power of the components below the second is even smaller, but despite its small variance, the fifth principal component has a higher explanatory power than the third and fourth components. When performing multiple regression analysis using the first, second, and fifth principal components, the first principal component is extremely significant, and the second principal component is also significant. In this analysis, the original sample size is small, so the degrees of freedom are insufficient, and it is judged to be not significant. However, as already mentioned, discussing statistical significance in this explanation is meaningless. The variance ratio of the equation that includes the fifth principal component is 0.672, which is greater than the equation that uses the three data items A, C, and E as explanatory variables. This variance ratio is the same as the sum of the variance ratios of the simple regression analyses (0.514 + 0.119 + 0.038). Also, the regression coefficients are the same as those of the simple regression. This is because the principal components are orthogonal and uncorrelated. By the way, the total variance ratio of the simple regression is 0.693, which is the ratio of the explained variance when the orthogonalized explanatory variables explain the data variance, and 1 - 0.693 = 0.307 is the ratio of the total error variance.

By performing multiple regression with principal components, the explanatory power of individual variables can be calculated additively, increasing the overall explanatory power. However, increasing the explanatory power of the equation does not directly relate to linguistic explanatory power. PC1 is an axis that eliminates correlation, and it is clear that it is related to factors such as A: wealth, B: urbanization, C: literacy rate, D: industrialization, and E: political stability. If this is what Lipset (1960) referred to as modernization, it appears that Lipset’s (1960) hypothesis is supported. However, since all the data adopted as explanatory variables point in the same direction and are positively correlated with the results, it is unclear what aspect of modernization contributes to the maintenance of democracy. There may be something not used in this analysis that contributes to the maintenance of democracy. The question is what that something is. The commentator, being of limited knowledge, does not know the details of the famous political scientist Lipset’s theory. It may be detailed in his book. If it is written, adding some data indicating the degree of that factor and performing partial correlation analysis between it and other data items and R, can deny the correlation between the parts that do not mediate modernization, such as wealth, literacy rate, and political stability, and R, thus verifying Lipset’s (1960) hypothesis. At present, since the fifth principal component clearly involves a principal component related to wealth, it is only considered that the first principal component is not related to something solely involving wealth. At present, Lipset's (1960) hypothesis cannot be denied, but it is also impossible to deny the possibility of other more fundamental causes.

II-6. Factor analysis (R script 124-147)

Finally, factor analysis was conducted. The dataset used was the one in Table 2. In the factor analysis, factors were extracted using the maximum likelihood method. Promax rotation was used to rotate the axes. Factor analysis limits the number of factors to fewer than the data items, maximizing the variance that can be explained by that number of factors. The goal is to limit the number of factors, concentrating and maximizing the variance in a smaller number of factors. Factor analysis inherently ignores the constraint of zero correlation, so the axes are not orthogonal like in principal component analysis. Furthermore, Promax rotation ignores orthogonality and performs rotation, increasing the correlation between factors. The purpose of the rotation is to concentrate factor loadings on a few data items, making it easier to interpret the meaning of the factors. This operation is expected to separate the content of PC1 in principal component analysis into several parts, making the interpretation of the factors easier.

Table 8 shows the results of the factor analysis. As expected, the content of the first principal component was divided into two groups: A: Wealth, C: Literacy Rate, E: Political Stability, and B: Urbanization, D: Industrialization. In other words, industrialization/urbanization and modernization were separated. This was quite predictable from the PC1-PC2 plot in Figure 3, where A, C, and E were grouped upwards, and B and E were grouped downwards. However, the fact that “modernization” was divided into two factors by Promax rotation is a significant achievement. Looking more closely, in the principal component analysis, the principal component loading of A (Wealth) was higher than that of C (Literacy Rate) and E (Political Stability). However, in the factor analysis, the loadings of C and E on Factor 1 exceeded that of A, with C (Literacy Rate) having a loading approximately 1.5 times higher than A (Wealth). If this factor corresponds to what Lipset (1960) referred to as

Table 8. Result of factor analysis

summary						loading	
	FA1	FA2				FA1	FA2
variance	1.858	1.701			A	0.665364	0.393248
propotion	0.372	0.34			B	0.040678	0.697954
cum prop	0.372	0.712			C	0.946746	-0.16137
					D	-0.02962	1.015819
					E	0.718941	0.038406

"modernization," it can be said that modernization has a greater impact on the general public's literacy rate (spread of education) than on economic development.

It should be noted that the cumulative variance ratio (0.712) is shown as a summary of the results, but this value is meaningless. The correlation coefficient between FA1 and FA2, calculated from the factor scores, was 0.599. This part is overlapping. It is incorrect to interpret the sum of the variances of the two factors as explanatory power.

To evaluate explanatory power, a regression equation was created to calculate the error variance between the predicted values and the actual values, and the explanatory power was assessed. The dataset is presented in Table 9.

Table 9. Data set for regression analysis using factor score

	FA1	FA2	R
AUT	0.132457	0.299984	-1.11631
BEL	0.30431	1.513244	1.240347
CZE	-0.75465	1.175544	0.868243
EST	1.204714	-1.97787	-0.74421
FIN	0.851779	-1.09541	0.496139
FRA	1.301073	-0.28362	1.240347
GER	-0.32362	1.170895	-1.11631
GRC	-1.70417	0.969633	-0.99228
HUN	-0.34578	-0.39953	-0.12403
IRL	2.045005	-2.43715	0.992278
ITA	-0.94714	0.636869	-1.11631
NLD	1.168683	0.190251	1.240347
POL	-0.09063	-1.41671	-0.74421
PRT	-2.38368	0.967647	-1.11631
ROU	-0.06602	-1.3536	-0.49614
ESP	-1.69413	0.752486	-0.99228
SWE	1.245663	-0.45887	1.240347
UK	0.056134	1.746197	1.240347

The results are shown below.

Simpe linear regression

	PR(>t)	variance ratio
R=0.5529FA1	0.0041**	0.412
R=-0.01809FA2	0.931	0.00048

Multiple regression

R=0.8543FA1+0.4674FA2 0.617

PR(>t) 0.000187 *** 0.01253*

$$variance\ ratio = 1 - \frac{variance\ of\ error}{total\ variance}$$

As a result of the simple regression analysis, the coefficient of FA1 was 0.553, with a p-value of 0.0041, indicating that this coefficient is statistically significant and explains 41% of the total variance. The coefficient of FA2 was -0.0018 and was not statistically significant at all. Additionally, the predicted values of R were almost entirely different from the actual values of R, explaining only 0.05% of the total variance. However, when performing multiple regression analysis with the two factors, the coefficient of FA1 was 0.854, and the coefficient of FA2 was 0.467, both of which were statistically significant. These coefficient values are not reliable due to the suspected multicollinearity caused by the correlation between the two factors. In fact, the variance ratio calculated using the equation was 0.617, which is lower than the variance ratio of 0.633 obtained from the multiple regression of PC1 and PC2 in the principal component analysis, indicating a decrease in explanatory power. The purpose of factor analysis, especially the purpose of Promax rotation, was to further analytically decompose the content of "modernization" extracted by principal component analysis, sacrificing the accuracy of the coefficient estimates. This is a significant achievement. The reliability of the coefficients divided into two in this analysis is another matter.

Figure 5 shows a scatter plot with FA1 and FA2 as the axes. In the figure, countries maintaining democracy are distributed in the upper right, and countries where democracy has collapsed are distributed in the lower left, separated by the blue line. This is a rearrangement of the scatter plot of PC1 and PC2 from the principal

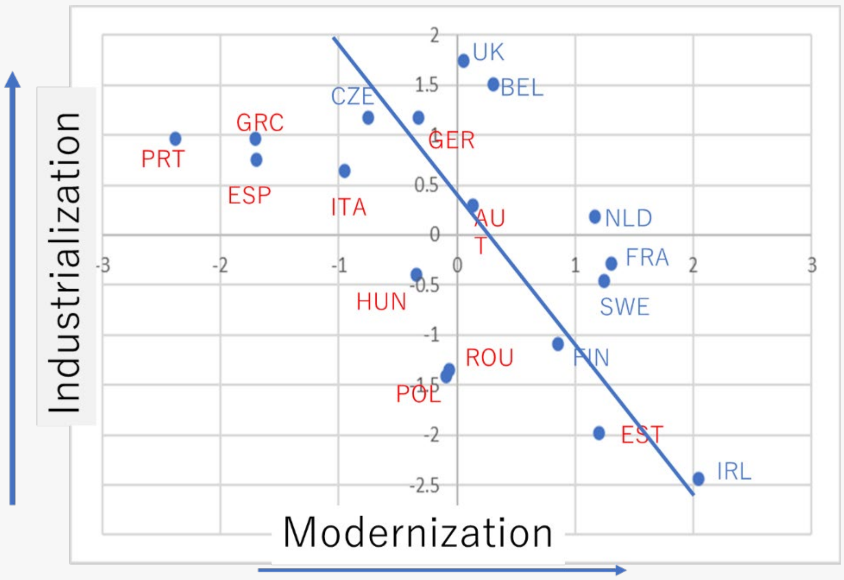

Fig. 5 FA1-FA2 scatter plot

component analysis, with the left and right sides swapped. On the boundary line, there are Czechoslovakia, Finland, Ireland, Germany, Austria, and Estonia. This result is similar to those of the MDS (Figure 2) and the principal component analysis (Figure 3), but the boundary line is clearer in this figure. In other words, the possibility of maintaining democracy increases in the upper right direction of this figure. In other words, modernization and industrialization both result in modernization. However, the question remains whether it was meaningful to separate modernization and industrialization in this discussion.

II-7. Organization of Numerical Analysis Results

Summary of Achievements in Numerical Analysis:

1. The dataset used for the analysis showed relatively high correlations among all analysis items.
2. In all analyses (MDS, principal component analysis, and factor analysis), it was shown that six countries—Austria, Czechoslovakia, Estonia, Finland, Germany, and Ireland—were in the boundary zone in the distribution of national characteristic values.
3. Clustering using principal component scores supported the view that the above six countries were in the boundary zone.
4. In the principal component analysis, the first principal component, which accounted for 67% of the total variance, and the second principal component, which accounted for 17%, were extracted. The first principal component had a certain load on all analysis items and was ambiguous in content, but the second principal component had a positive correlation with the level of education (literacy rate) and a negative correlation with urbanization and industrialization, indicating a modernized agricultural country.
5. Factor analysis with promax rotation allowed the separation of "modernization" into "modernization" and "industrialization."
6. Regression analysis was conducted using standardized original data, principal component scores, and factor scores, with the result (maintenance of democracy) as the dependent variable. As a result, in the original data, the coefficients for wealth, education, and political stability were large, and all three were statistically significant. However, the total variance of the predicted values, obtained by subtracting the variance of the difference between the predicted and actual data from the total variance, exceeded 100%. This was due to the high correlation among data items. When performing multiple regression with these three, none of the coefficients were significant. This result was thought to be due to the high correlation among items and the small data size. The explanatory power accounted for 61% of the total variance.
7. In the regression analysis using principal component scores, the coefficient of the first principal component was significant in simple regression, with an explanatory power of 51%. The coefficient of the second principal component was not significant, but its explanatory power was 12%. Additionally, despite the small variance of the fifth principal component, it had a larger coefficient

than the third and fourth principal components, with an explanatory power of 4%. The result of multiple regression analysis using these three principal components showed that the coefficients were significant up to the second principal component, with an overall explanatory power of 67%. The total explanatory power of all five principal components was 69%, which was higher than the regression analysis using the original data items and factor scores (theoretically expected).

8. In the regression analysis using factor scores, the coefficient of the first factor was significant in simple regression, with an explanatory power of 41%. The coefficient of the second factor was not significant, and its explanatory power was minimal. In the multiple regression analysis using these two factors, both coefficients were significant, and the explanatory power of the equation was 62%, suggesting that the second factor complemented the first factor in influencing the results.