2001.10.24（北海道大学）

制約条件付き変量の解析
－選挙結果に関するデータ解析－

成蹊大学　経済学部　中西寛子

１．制約条件付き変量（Compositional Variables）

ここで扱う制約条件付き変量についてはじめに説明する．i 番目に得られた j 番目の非負変量を x_{ij} $\left(i=1,\cdots,n,\quad j=1,\cdots,d\right)$ とする．変量間に一般性を失わず，制約条件

$$\sum_{j=1}^{d} x_{ij}=1 \quad \left(i=1,\cdots,n\right) \tag{1}$$

が成り立つ場合，Compositional Variables または Compositional Data と呼ばれる．つまり, 見かけ上 d 個の変量があるが, 実際は $d-1$ 個の変量でデータが一意に決まる．一般に，このようなデータの解析は容易ではなく，未だ有効な解析方法は見いだされていない．

Compositional Data に関する著書としては Aitchison（1986）がある．本書は 400 ページを超えるものであるが, データの取り扱いについての注意や解析の困難さなどが述べられ, 有用な手法はあまりみられないというのが個人的な感想である．Aitchison はその後も研究を進め，最近 Compositional Data に関する２次元配置のプロットについての Discussion Paper（Aitchison and Greenacre(2001.7.)）を発表した．ここに，最近の Compositional Data に関する研究の成果がまとめられているので，参考にされるとよい．また，Compositional Variables を説明変量とし，目的変量 y がある場合については岩崎（1994）が過去の研究成果をまとめている．

Compositional Data の理解を容易にするため，ここでは選挙における得票率を扱う．本データを解析することによって Compositional Data の扱いの困難さについて簡単にふれる（第３章）．第４章からは，この制約条件を逆に有効利用する手法を提示し，研究成果について述べる．

２．データの概要（平成１２年衆議院総選挙）

平成１２年６月２７日付け朝日新聞に, 衆議院選挙における小選挙区と比例代表区の都道府県別得票数および得票率が示された. 政党の支持や投票行動に関する背景についてはあまり考えず, ここでは, 各都道府県の比例代表区の得票率データが得られたというところから解析することにする. 政党は「自民党」,「民主党」,「共産党」,「公明党」,「社民党」,「自由党」,「その他の政党」の７党にまとめる．この得票率データは先に述べた Compositional

Data であり，変量 x_{ij} $\left(i=1,\cdots,47,\quad j=1,\cdots,7\right)$ に対し，式(1)は

$$\sum_{j=1}^{7} x_{ij}=1 \quad \left(i=1,\cdots,47\right) \tag{2}$$

となる．

はじめに，変量の正規性について考察する．４７都道府県別得票率（%）の統計量（平均，標準偏差，歪度，尖度）について示す（表１）．「自由党」と「その他の政党」において，歪度，尖度が非常に大きく，「共産党」と「社民党」についても若干，同様の傾向が見られる．そこで，都道府県別得票率に関する標準化変量の計算を行った．表２は政党の標準化変量の絶対値が３以上を示した都道府県を示す．表１において「共産党」，「社民党」，「自由党」，「その他の政党」の歪度，尖度が非常に大きくなった原因はそれぞれ「京都府」，「大分県」，「岩手県」，「鹿児島県」にあることがわかる．

「京都府」，「岩手県」，「大分県」，「鹿児島県」の値を外れ値と見なし，４３都道府県別得票率の統計量（平均，標準偏差，歪度，尖度）をあらためて表３に示す．歪度，尖度の値において「その他の政党」に関する値が若干大きいが，他の政党に関してはおおよそ正規分布に従っていると見なしてよいであろう．第３章では，上記４府県を除いた４３都道府県の得票率の相関係数について考察する．

表１　得票率（%）の統計量

	自民	民主	共産	公明	社民	自由	その他
平均	32.2	23.1	9.9	12.6	9.9	10.7	1.6
標準偏差	6.4	5.3	3.4	2.9	3.3	5.3	1.9
歪度	-0.1	-0.1	1.1	0.3	1.5	4.0	3.4
尖度	-0.7	1.0	1.4	0.7	3.9	23.9	15.1

表２　得票率の標準化変量

	自民	民主	共産	公明	社民	自由	その他
岩手	-1.40	-2.90	-0.85	-1.91	0.23	5.74	0.89
京都	-0.71	-0.14	3.28	-0.12	-0.69	-0.54	-0.20
大分	-0.06	-0.90	-0.99	0.56	3.87	-0.97	-0.48
鹿児島	1.23	-1.90	-1.36	-0.48	0.86	-0.90	5.14

表３　得票率（%）の統計量：４３都道府県

	自民	民主	共産	公明	社民	自由	その他
平均	32.3	23.8	9.9	12.7	9.6	10.3	1.4
標準偏差	6.4	4.6	3.0	2.8	2.8	2.7	1.3
歪度	-0.1	0.5	0.7	0.4	0.7	0.0	1.8
尖度	-0.6	0.1	-0.1	0.7	0.4	-0.8	4.3

３．都道府県別得票率の相関係数

各党の間の相関係数を表４に示す.「自民党」と他党の間の相関係数の絶対値は比較的大きいく，負であることがわかる．各党の得票率の和は 100%であるため,「自民党」に票をとられた都道府県は必然的に他党の票が少なくなり負の相関になる.いわゆる完全多重共線性が生じるため，これらの相関係数はこのままでは利用できない．また，相関行列については正則行列でないため,逆行列が存在しない.一般に,「どれか一つの変量を削除する」,「変量に重みをかける」など工夫して解析を行うが，それらの工夫を行う理由（背景）が十分でなければならない.

表４　各党の間の相関係数：４３都道府県

	自民	民主	共産	公明	社民	自由	その他
自民	1.00						
民主	**-0.54**	1.00					
共産	**-0.71**	0.27	1.00				
公明	**-0.38**	-0.14	0.35	1.00			
社民	**0.08**	-0.51	-0.11	-0.16	1.00		
自由	**-0.12**	-0.03	-0.28	-0.31	-0.08	1.00	
その他	**-0.52**	-0.05	0.32	0.34	0.03	0.12	1.00

本データに対しても「その他の党」の得票率の値を削除することが先ず考えられ，それを行った．しかしながら，それだけでは多重共線性が完全に回避できなかった．もう一ついずれかの党を削除し解析を続けることは非常に作為的であるため,

$$A=(\text{各党の得票率})\div(100-\text{自民党の得票率}) \tag{3}$$

を考える．これは,「自民党」への得票の残りを「自民党」以外の６つの政党がどのように分配したかというものである．表５に「宮城県」と「静岡県」の得票率を，また，表６に値Ａを例として計算し示す.

「自民党」以外の政党の値Ａを合計すると１となるため，多重共線性が完全に回避されたわけではない．ここでは「自民党」の影響を削除しただけである．値Ａに関する各党の間の相関係数を表７に示す.「自民党」以外の得票率のうち，第２党である「民主党」に票をとられた都道府県は必然的に他党の票が少なくなり，これもまた，一般に負の相関になるため考察において注意が必要である.

表５　得票率（%）

	自民	民主	共産	公明	社民	自由	その他
宮城	29.9	27.4	9.3	11.3	9.6	10.9	1.6
静岡	29.8	29.6	10.1	10.9	8.1	9.6	1.8

表６　値Ａの例

	自民	民主	共産	公明	社民	自由	その他
宮城	0.43	0.39	0.13	0.16	0.14	0.16	0.02
静岡	0.42	0.42	0.14	0.16	0.12	0.14	0.03

表７　値Ａの相関係数：４３都道府県

	自民	民主	共産	公明	社民	自由	その他
自民	1.00						
民主	**-0.02**	1.00					
共産	**-0.48**	-0.14	1.00				
公明	**0.02**	-0.39	0.09	1.00			
社民	**0.33**	-0.58	-0.23	-0.12	1.00		
自由	**0.24**	-0.09	-0.56	-0.42	0.01	1.00	
その他	**-0.45**	-0.38	0.14	0.11	-0.05	-0.07	1.00

表４では「民主党」,「共産党」,「公明党」が「自民党」の得票率に対し比較的強い負の相関を示したのに，表７ではその値が大きく減少し，逆に，表４では「社民党」,「自由党」が「自民党」の得票率に対し無相関であったのに対し，表７では正の相関が見られる.

先に述べたように「各党の得票率の和は 100%である」という値への制約条件は統計的解析を困難にさせる. ここでは, 値Ａを導入することによって「自民党」の影響を削除したが,「自民党」以外の政党については多重共線性が生じるので, 政党間の関係を見るには何らかの方法を見いださなければならない. 一案として，第２党である「民主党」の影響を「自民党」の影響を除いた方法と同様に続けて除くことができるが, これには選挙者の思考方法が関与する. つまり,「自民党」に投票するか否かを判断し，投票しないなら,「民主党」を投票するか否かを判断し，…といった時系列的な思考が仮定できなければ本案は適切でない.

このように多重共線性のような問題をはじめとする, 本データ特有の問題について様々な解決策は考えられるが, どれもそのデータの背景に依存することが多く, どのデータにも利用できる普遍的な手法はまだ見いだされていない.

４．Compositional Data における近似度と距離

Compositional Data の各ケースに対し，原点を始点とするベクトル $(\sqrt{x_{i1}}, \cdots, \sqrt{x_{id}})$ $(i=1,\cdots,n, \quad j=1,\cdots,d)$ を考える. 制約条件 $\sum_{j=1}^{d}\left(\sqrt{x_{ij}}\right)^2 = \sum_{j=1}^{d} x_{ij} = 1$ が成り立つため, これらのベクトルの終点は $d-1$ 次元超球面上の第１象限のどこかに布置する．さらに，２つのケースのベクトルのなす角を θ とすると，内積は $\cos\theta$ と表すことができるので２つのケースの近似度として考えることができる（Matusita(1956)）.

詳細に述べると，２つのベクトル$(\sqrt{p_1}, \cdots, \sqrt{p_d})$，$(\sqrt{q_1}, \cdots, \sqrt{q_d})$に対し内積は

$$\cos\theta = \sum_{i=1}^{d} \sqrt{p_i}\sqrt{q_i} \tag{4}$$

となる．

$$2(1-\cos\theta) = \sum_{i=1}^{d} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \tag{5}$$

と書き直すことができる．(5)は角谷の情報量ともよばれ，情報量の凸性を持ち，変量を細分化するほど情報量が大きくなることが知られている（河田(1987)）．これらのことから，距離として

Ⅰ　$\arccos\theta$＝角度θ（Domenges and Volle(1980)）

Ⅱ　$\left(2(1-\cos\theta)\right)^{\frac{1}{2}}$（Matusita(1951)）

Ⅲ　$1-\cos\theta$

などが考えられる．Ⅲは距離の三角不等式の性質が成り立たない．Ⅱは$d-1$次元超球面上の２点の弦の距離となるため，超球面上の距離より短くなる．ベクトルの終点が超球面上にあることから弦より弧の方が好ましい．Ⅰは$d-1$次元超球面上の２点の弧の距離と比例し，距離として採用するには十分である．そこで，角度θを距離とし今後の議論を進める．また，(5)の情報量の凸性から，Ⅰにおいても変量を細分化するほど距離が大きくなるとうい性質を持つ．このことは距離の概念から理想的である．

上記の議論を都道府県別得票率にあてはめると，原点を始点とするベクトル$(\sqrt{x_{i1}}, \cdots, \sqrt{x_{i7}})$の終点は６次元超球面上の第１象限のどこかに布置することになる．さらにθを利用することによって，都道府県の間の距離を測ることができる．表８および表９に「宮城県」と「静岡県」の各党に対する得票率とその平方根を示す．表５の値より$\cos\theta$が求まり，$\arccos\theta$を考えると$\theta = 2.4^{o}$となる．

表８　得票率（%）

	自民	民主	共産	公明	社民	自由	その他
宮城	29.9	27.4	9.3	11.3	9.6	10.9	1.6
静岡	29.8	29.6	10.1	10.9	8.1	9.6	1.8

表９　得票率の平方根

	自民	民主	共産	公明	社民	自由	その他
宮城	5.47	5.23	3.05	3.36	3.10	3.30	1.26
静岡	5.46	5.44	3.18	3.31	2.85	3.10	1.35

このような方法で求めた都道府県間の距離を付表Aに示す．各県に対し，付表Aに示された他県との間の距離について和を計算（付表Aの最終行を参照）することにより「岩手県」が大きく他の都道府県と離れていることがわかる．先にあげた「宮城県」と「静岡県」の距離は都道府県間の距離の中で一番近い.付表Aに示したような距離行列を考えると多次元尺度構成法，因子分析，クラスター分析などの分析が行える．次章では，多次元尺度構成法を用いて都道府県の関係を見ることにする．

５．多次元尺度構成法による都道府県の関係

付表Aのように都道府県間の距離を個々に見るのも興味深いことであるが,これらの距離より空間的に都道府県を布置し，視覚的に都道府県の関係を考察する．この考察に多次元尺度構成法を用いる.表１０は２次元布置における都道府県の次元１および次元２の結果である．

表１０　多次元尺度構成法（２次元布置）における
都道府県の次元１および次元２の結果

都道府県	次元１	次元２	都道府県	次元１	次元２
北海道	1.2312	-.3785	滋　賀	.8224	-.1714
青　森	-.7968	.3657	京　都	1.5777	.2727
岩　手	-2.1371	3.6751	大　阪	1.5837	.8796
宮　城	.3562	.0786	兵　庫	.9400	1.0231
秋　田	-.7337	.0945	奈　良	.8476	-.0039
山　形	-1.0237	-.7076	和歌山	.7354	1.9505
福　島	-.1669	.6865	鳥　取	-.4272	-1.1954
茨　城	-.1856	-.2553	島　根	-.1913	-1.4866
栃　木	-.3607	-.7770	岡　山	.1123	-1.2473
群　馬	-.7042	-.4476	広　島	-.3900	-.1853
埼　玉	1.0151	.2143	山　口	.4644	-.3613
千　葉	.7334	.4688	徳　島	.9117	-.9851
東　京	1.5071	1.3855	香　川	-.7875	-.7939
神奈川	.9439	.8251	愛　媛	-.4667	-.5354
新　潟	-1.1308	-.2691	高　知	.8506	-1.2398
富　山	-1.7821	-.4570	福　岡	.7751	.7000
石　川	-1.1822	-1.3277	佐　賀	-.9094	-.0790
福　井	-.6835	-1.2448	長　崎	-.6324	.6515
山　梨	.6607	.1781	熊　本	-.2276	-.1118
長　野	1.8459	-.6558	大　分	-1.5950	-1.0910
岐　阜	.1332	-.3678	宮　崎	-.9481	-.5717
静　岡	.7031	.0963	鹿児島	-2.2959	1.8361
愛　知	1.3297	.3219	沖　縄	-1.0010	1.3018
三　重	.6789	-.0583			

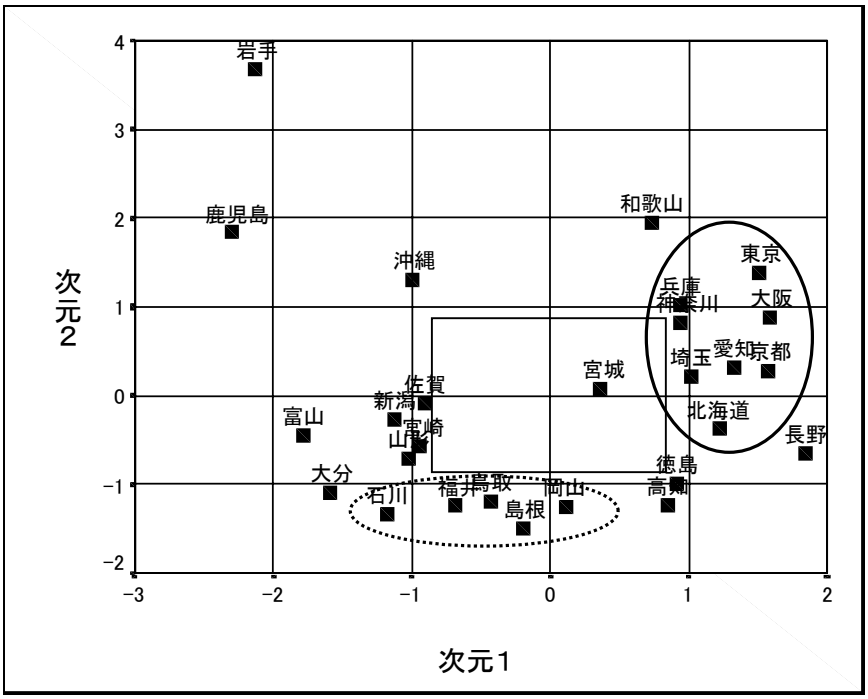

図１　都道府県の関係

これらの結果を２次元座標に布置すればよい. すべてを表示すると文字が重なり合い見苦しくなる．このため，次元１または次元２どちらかの値が 0.9 を越えた 26 都道府県およびほぼ中心に位置する「宮城県」のみを表示する（図１）．つまり，「宮城県」を中心とする□枠の中には 21 都道府県が存在する．

図１から考察されることを列挙する．

１．「岩手県」，「鹿児島県」が他の都道府県より離れた位置にある．表２の考察において見られた結果と同じである．

２．「北海道」，「埼玉県」，「東京都」，「神奈川県」，「愛知県」，「京都府」，「大阪府」といった都会型の都道府県が右側に位置する．外れ値であると思われた「京都府」がこれらの中にいることが興味深い．

３．「石川県」，「福井県」，「鳥取県」，「島根県」，「岡山県」が下部に位置する．これらの県は「自民党」への得票率が大きい．

６．多次元尺度構成法による都道府県の関係（「自民党」を除いた場合）

第３章で提案した値Ａのうち，「自民党」を除く残りの６党に対するものを用いて再度，都道府県間の距離を求める．つまり，「自民党」への得票の残りを「自民党」以外の６つの政党がどのように分配したかという割合Ａを用いる. ６政党の値Ａを合計すると１となるため，第６章で説明した方法で距離を算出できる（付表Ｂ）.

他県との間の距離について和を計算（付表Ｂの最終行を参照）することにより，やはり，

「岩手県」と「鹿児島県」が大きく他の都道府県と離れていることがわかる．これらの県の離れ方は付表Aの「自民党」が含まれている場合より大きく，「自民党」以外の票の取り方が他の都道府県と比較して大きく異なることがわかる．さらに，「大分県」，「和歌山県」，「沖縄県」... と続くが，これらも付表Aよりも大きな値を示している．一方，「神奈川県」，「宮城県」などが他の都道府県との距離が近い．

次に，多次元尺度構成法を用いて２次元布置における都道府県の次元１および次元２を計算する．表１１はその結果である．

表１１　多次元尺度構成法（２次元布置）における
都道府県の次元１および次元２の結果（自民党を除く）

都道府県	次元１	次元２	都道府県	次元１	次元２
北海道	1.1251	-.1948	滋　賀	.8646	-.2059
青　森	-1.0351	-.1458	京　都	1.5550	.4225
岩　手	-4.2739	-1.5868	大　阪	.6866	.7914
宮　城	.0212	-.3818	兵　庫	-.2001	.6649
秋　田	-.7752	-.3695	奈　良	.6626	.0037
山　形	-.1180	.8412	和歌山	-.8743	2.0244
福　島	-.7118	-.6024	鳥　取	.6984	1.2248
茨　城	-.1287	-1.0989	島　根	1.3072	.0403
栃　木	.0487	-1.0972	岡　山	1.2860	.2757
群　馬	-.1665	-.3144	広　島	-.3039	.0149
埼　玉	.4807	-.4057	山　口	.5509	-.4656
千　葉	.1024	-.3491	徳　島	1.4622	-.6235
東　京	-.1512	-1.0648	香　川	.1591	.9863
神奈川	-.0169	-.2186	愛　媛	.0872	-.1634
新　潟	-.7875	-.9821	高　知	1.4977	1.0895
富　山	-1.3853	-1.1585	福　岡	-.0250	.4743
石　川	.0033	-1.5708	佐　賀	-.7994	.4932
福　井	.5538	-.8801	長　崎	-1.1174	-.2969
山　梨	.1598	-.8692	熊　本	-.2212	.5541
長　野	1.7112	-.3075	大　分	-.2920	2.3314
岐　阜	.2408	-.7389	宮　崎	-.3524	.9997
静　岡	.4644	-.4395	鹿児島	-2.2652	2.6329
愛　知	.9501	-.5132	沖　縄	-1.1268	1.7526
三　重	.4488	-.5732			

これらの結果を２次元座標に布置する．図１と同様，次元１または次元２どちらかの値が 0.9 を越えた 24 都道府県およびほぼ中心に位置する「神奈川県」のみを表示する（図２）．「神奈川県」を中心とする□枠の中には 23 都道府県が存在する．

図１との比較をも含め，図２から考察されることを列挙する．

１．図１と同様「岩手県」，「鹿児島県」が他の都道府県より離れた位置にある．

２．「神奈川県」が中央に位置することから，「北海道」，「埼玉県」，「東京都」，「愛知県」，

「大阪府」が中央に寄る傾向が見える．「京都府」はこれらの動きと異なる．

3．「北海道」，「長野県」，「愛知県」，「島根県」，「徳島県」が右側に位置する．これらの県は「民主党」への得票率が大きい．

4．「岩手県」，「富山県」，「石川県」が下部に位置する．これらの県は「自由党」への得票率が大きい．

5．「和歌山県」，「鹿児島県」，「沖縄県」が上部に位置する．これらの県は「その他の党」への得票率が大きい．

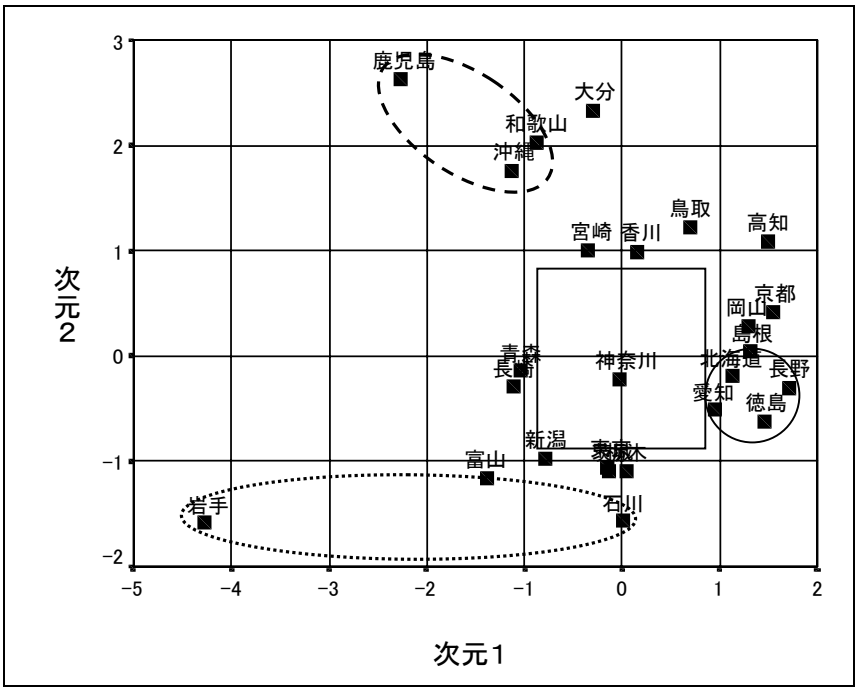

図２　都道府県の関係（自民党を除く）

７．終わりに

本稿で扱った「各党の得票率の和は 100%である」といった制約条件は，統計的解析においてデータの扱いを非常に困難にする．たとえば，表４の相関係数のように必要以上に強い負の相関があらわれ，解釈を困難にする．この制約条件を回避するには，数理モデルをデータに入れなければならない．第３章の値Ａは数理モデルの一例である．数値モデルは統計学を学ぶものが勝手に定義することは好ましくない．その分野の専門家，ここでは政治学者が国民の投票行動を数理モデル化した方がよい．その意味において，本データについては何らかの助言を求めるべきであろう．このように変数間の相関等については未だ解決されず，今後の課題である．

一方，第４章で提案したケース間の距離は先の制約条件を逆に利用したものである．一般のデータにこの距離を定義することはできない．この距離の解釈は三角関数の初歩的な理論を学んだ者ならば容易である．高等数学を用いることなく簡単に計算できるので利用価値が

高いと思われる. 距離が定義できれば, 様々な多変量解析を用いることができる. ここでは, 多次元尺度構成法を用いたが, クラスター分析や因子分析などへの利用も考えられる.

最後に, 多次元尺度構成法に関しては統計パッケージソフト SPSS 8.0J を用いたことを言及する.

参考文献

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, J. and Greenacre, M. (2001) Biplots of Compositional Data. In *Economics Working Papers* from Department of Economics and Business, Universitat Pompeu Fabra (http://www.econ.upf.es/deehome/what/wpapers/postscripts/557.pdf)

Domenges, D. and Volle, M. (1980) L'analyse Factorielle Spherique. In *Data Analysis and Informatics*, North-Holland, 253-257.

Matusita K. (1951) On the Theory of Statistical Decision Functions, *Ann. Inst. Stat. Math.*, **3**, 17-35.

Matusita K. (1956) Decision Rule, Based on the Distance, for Problems of Fit, Two Samples, and Estimation. *Ann. Math. Stat.*, **26**, 631-640.

岩崎学（1994） 混合実験の計画と解析. サイエンティスト社.

河田敬義(1987) 情報量と統計.「統計数理」, 35 巻, 第 1 号. 1-57.

参照した選挙に関するホームページの URL

http://www.asahi.com/senkyo2000/index.html
（2000 年総選挙(asahi.com)）

http://www.mainichi.co.jp/eye/sousenkyo/index.html
（Mainichi INTERACTIVE　総選挙 2000）

http://www.yomiuri.co.jp/election2000/main.htm
（2000 衆院選（読売））

追加

Ⅰ　平成 13 年 7 月の参議員選挙においても同じような距離を計算し，多次元尺度構成法を施した．次元 1 または次元 2 どちらかの値が 0.9 を越えた都道府県は 13 にとどまった．これは，どの都道府県においても自民党への票が大きくのび，都道府県の差（距離）が縮まったことに原因がある（表，図略）．

Ⅱ　各都道府県および全国に対し，平成 12 年衆議院選挙と平成 13 年参議員選挙の距離を計算した（下の表）．

都道府県	距離	都道府県	距離
北海道	13.64	滋　賀	11.45
青　森	14.19	京　都	11.42
岩　手	7.40	大　阪	12.83
宮　城	11.97	兵　庫	12.97
秋　田	12.91	奈　良	12.77
山　形	11.38	和歌山	8.81
福　島	10.32	鳥　取	13.86
茨　城	17.74	島　根	13.21
栃　木	13.64	岡　山	12.71
群　馬	12.21	広　島	16.70
埼　玉	15.18	山　口	15.04
千　葉	12.79	徳　島	15.33
東　京	14.03	香　川	15.28
神奈川	14.32	愛　媛	15.60
新　潟	18.04	高　知	15.18
富　山	14.10	福　岡	13.27
石　川	14.06	佐　賀	12.48
福　井	15.40	長　崎	10.41
山　梨	13.78	熊　本	11.81
長　野	14.20	大　分	11.86
岐　阜	14.51	宮　崎	10.01
静　岡	13.69	鹿児島	6.67
愛　知	11.70	沖　縄	18.00
三　重	12.16	全　国	12.27

Ⅲ　平成 10 年参議院選挙，平成 12 年衆議院選挙，平成 13 年参議院選挙の全国に対する距離を下の表に示す．

	参院 98	衆院 00	参院 01
参院 98	0	9.02	9.87
衆院 00	9.02	0	12.27
参院 01	9.87	12.27	0