2002 FIFA Worldcup KoreaJapanの参加各国のデータにおける主成分分析およびクラスター分析両面での分類

東京理科大学
綾部英明

平成16年11月8日

目 次

1　はじめに

データを分析する手法は様々あるが、その中の代表的な手法のひとつに主成分分析がある。主成分分析はデータの中で数ある変数をより少ない変数で説明できるようにする方法である。各々のサンプルの分類を考えるとき、主成分得点を見るだけでは容易に特徴づけを行うことは難しいが、各主成分の得点ごとに散布図を描くことで容易に理解ができる。しかし、サンプル数が多くなると分類が困難なデータが多く現れてしまうという問題点がある。

また同様にデータを分類するといった手法の中にクラスター分析がある。クラスター分析はサンプルの距離を定義しその距離の近さにより分類する方法である。この手法は樹形図（デンドログラム）を描くことで視覚的に容易に分類を行えるが、どの距離において分類すべきか判断が難しい。また、データ数が多いと任意のグループで特徴を探そうとしても見つけるのに非常に手間がかかるといった問題点がある。

今回はこれら二つの分析方法を用い、互いの欠点を補いマクロな視点とミクロな視点から分析を行えること、分類の精度を高めることができることを示し、また S-Plus を用いることで特に容易に実行・理解ができることを示すことを目的とした。

2　データの概要

今回使用するデータは

2002FIFA worldcup Japan/Korea (http://fifaworldcup.yahoo.com/02/jp//index.html) から集めたサッカーに関するデータである。出場 32 チーム分の、大会中の一試合平均ゴール数、シュート総数、ゴール枠内シュート総数、アシスト総数、コーナーキック総数、フリーキック総数、オフサイド総数、ショートパス総数、ロングパス総数、一試合平均被ゴール数、ファウル総数、タックル総数の 12 個の変数で作られている。

3　分析目的

2002FIFA worldcup の出場国 32 チームはどのように試合結果から分類できるのかを調べようと考えた。以上のデータから自分は強いチームと弱いチームという分け方の他、戦術の取り方などで分類ができるのであろうかという予想をした。しかし他にもどのような分類基準が背後に隠されているのかを知り、さらなるサッカーに対する、見るべき点の発見することを目的とした。

4　分析方法

まず、12 ある変数での分類は困難であるため、特徴ある変数を選択するため主成分分析を用い、そこで特徴づけができるチームを選択し、データのチーム数を少なくする。そして、主成分分析では分類が困難であるチームに限りクラスター分析を行い、細分化し、各々の特徴づけをし、チームを分類することとした。

5　分析結果と考察

5.1　主成分分析

5.1.1　累積寄与率を求め、考察を行う

まず、S-Plus にデータを取り入れたところからはじめる。(本研究のデータセット名は ” データ ” とする。)

```
pr1 <- princomp(データ,cor=T)
```

#今回のデータは各変数の数値の大きさにばらつきが見られるため、相関係数を用いた。

```
plot(pr1)
```

このコマンドより累積寄与率が示される。

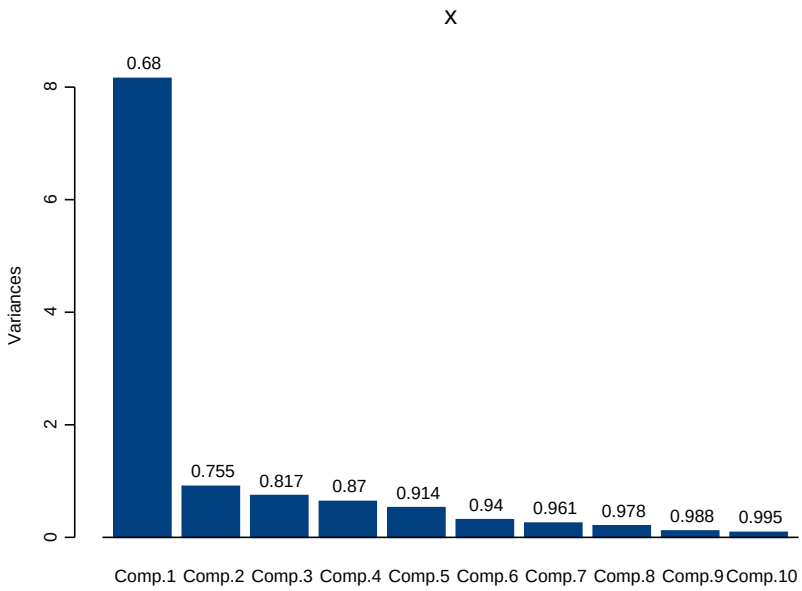

図 1: 累積寄与率

図 1 より累積寄与率を見ていくと、第三主成分までで全体の 80 %を超えるために今回は第三主成分まで取り上げればいいということがわかる。

ここで図 1 にて、累積寄与率を求めたが、次のコマンドでも参照可能である

```
vv_sum(pr1$sdev^2)

cumsum (pr1$sdev^2/vv)
```

表 1: 累積寄与率

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
0.679552	0.75521	0.81701	0.870243	0.914189	0.94021

Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12
0.961176	0.978388	0.987839	0.995125	0.99849	1

5.1.2 因子負荷量を求め、各主成分の意味について考察を行う

```
loadings(pr1)[,1:3]
#このコマンドより第一主成分から第三主成分の因子負荷量を求める
```

表 2: 因子負荷量

	Comp.1	Comp.2	Comp.3
平均ゴール数	0.235313	-0.60541	0.163022
シュート総数	0.334754	0.007624	0.091102
ゴール枠内シュート総数	0.324994	-0.0063	0.15843
アシスト数	0.303015	-0.26922	0.075421
コーナーキック数	0.299684	0.267252	-0.21752
フリーキック数	0.22944	-0.37938	-0.27688
オフサイド数	0.251029	-0.23117	0.396102
ショートパス数	0.300888	0.32084	0.201652
ロングパス数	0.304692	0.24609	-0.12709
平均被ゴール数	-0.21561	0.135476	0.759005
ファウル総数	0.315169	0.190106	-0.04781
タックル数	0.316973	0.272238	0.135171

表 5.1.2 より各主成分の意味を考えてみると、第一主成分は因子負荷量の値から攻撃の要素を構成しているものが高く、失点だけが低いため「総合的な強さ」を表すものと考えられる。

第二主成分はパスをつないでいくチームであるかゴールに突き進んでいくチームということを表すものと考えられる。

第三主成分はロングキックなどキックの精度が高い選手がいて、それを戦術としているチームと逆にそういう選手がいなくチーム全体でパスなどを使い苦し紛れになったチームということを表すものと考えられる。

5.1.3　各チームの主成分得点を求め、チームの特徴づけや分類を行う

```
pr1$scores
```

このコマンドより主成分得点が示される。

表 2 より各チームを特徴づけが行うことはできるが、得点を眺めるだけでは容易ではないため次に各主成分を組み合わせて散布図を載せ各チームを分類する。

表 3: 各チームの主成分得点

	Comp.1	Comp.2	Comp.3
ブラジル	6.72089845	-1.56916457	1.04203551
ドイツ	7.45040099	-0.31484755	-0.72814171
スペイン	3.47649853	-0.61724112	1.226101
トルコ	5.06259026	1.43711356	0.5899325
韓国	5.45402404	3.03480202	-0.21908562
セネガル	1.37448315	0.30116154	0.86032621
アメリカ	1.49127643	-0.01817338	0.6306671
ベルギー	0.60799426	-0.60986351	-0.42754872
パラグアイ	0.05287533	-0.81984399	0.37513963
ポルトガル	-1.7421939	-1.551542	0.01885072
アイルランド	0.34783606	-0.35559267	-1.03130318
イングランド	0.82862535	0.21733388	-0.63564973
コスタリカ	-1.02437167	-1.13965169	1.38813483
デンマーク	-0.97398388	0.17763762	0.10985066
日本	0.18461065	0.34210766	-1.36780559
スウェーデン	-0.45094221	0.1187888	-0.22652824
南アフリカ	-1.92664274	-0.6893657	0.23541203
イタリア	1.58252696	-1.01717119	-0.07658171
メキシコ	0.1347485	0.27874651	-0.24592963
ウルグアイ	-1.77740406	-0.88763197	-0.33580324
ロシア	-1.45434943	-0.68359532	-0.08959561
ポーランド	-2.16581203	-0.33116725	0.98296748
アルゼンチン	-0.52284584	-0.04996203	-1.16599132
エクアドル	-2.01984687	0.67055143	-0.40271973
スロベニア	-2.71657107	0.71732031	0.6609584
クロアチア	-1.60708716	-0.24202475	-0.84815079
カメルーン	-1.25853781	-0.88549767	-1.30543142
ナイジェリア	-2.53488102	0.14912895	-1.03688234
チュニジア	-3.16727474	0.77219222	-0.55205428
中国	-3.91954303	1.26179812	1.01401891
サウジアラビア	-4.01049241	1.02641282	2.23093897
フランス	-1.49660908	1.27724092	-0.67013108

```
biplot(pr1)
```

このコマンドより散布図が示される。

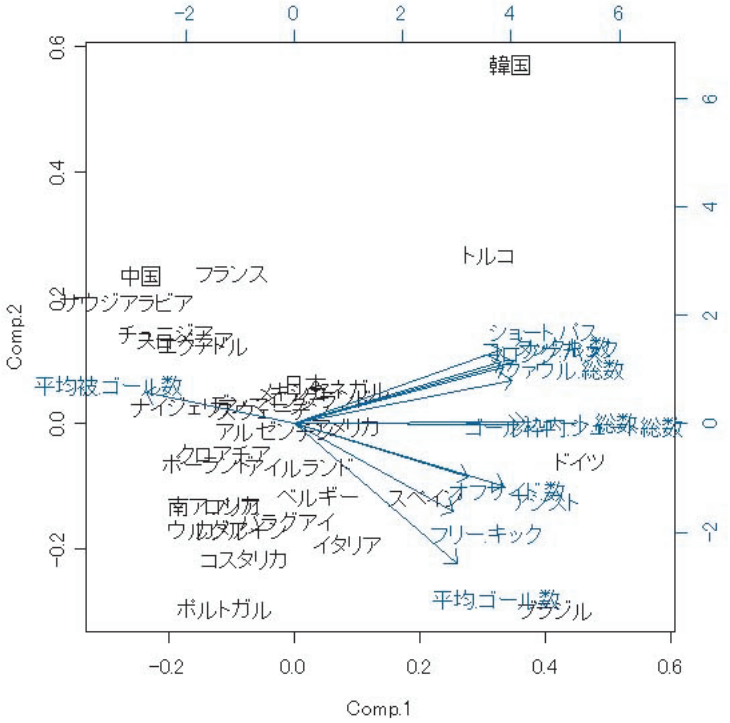

図 2: 第一主成分と第二主成分の散布図

またこれ以外の散布図は GUI で作る。

図 2・3・4 より、第一主成分はベスト 4 に入ったチームや予選リーグ全勝のスペインなどが得点が大きく、予選リーグで敗退したチームが小さいことが見て取れる。とくに予選リーグで大敗を喫したチームは特に小さいことがわかる。

また第二主成分は個人技が魅力のブラジルや、西洋のブラジルといわれるポルトガルなどといったパスワークが自慢であるチームの得点が小さく、韓国のように蹴って走るカウンターサッカーを展開するチームの得点が大きいことがみて取れる。

第三主成分は、フリーキックやコーナーキックなどを多用するチームや背の高いターゲットとなる選手がいて接戦に持ち込むチームの得点が小さく、点を取られても取り返しにいくといった傾向のあるチームの得点が大きい。

以上から、主成分得点により各チーム大体の特徴づけはできたように思われる。

しかし、散布図を見てわかるとおり、主成分分析ではデータ数が多くなればなるほど分類が明確にならないデータが発生する。散布図上で言うならば、原点の近くに集まるものである。これらは主成分得点が 0 に近いため特徴づけが困難なものである。主成分分析の欠点であるこれらのデータの分類を、今回はクラスター分析を行い細かくみていく。

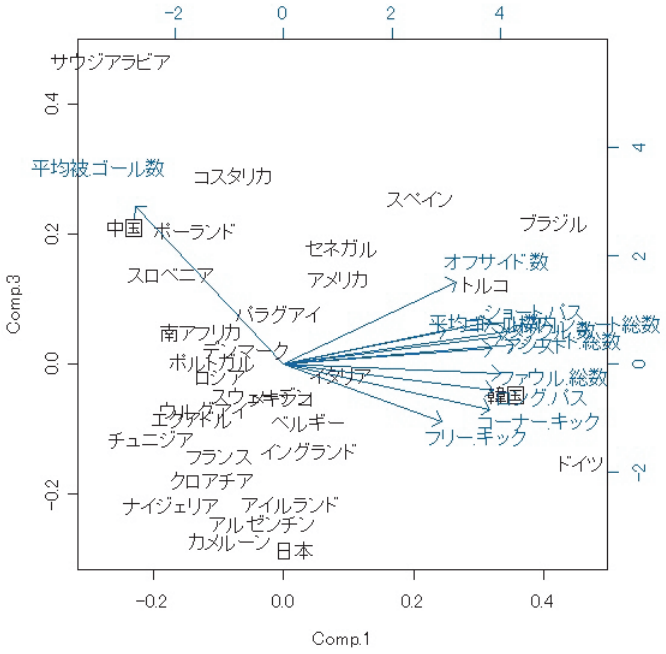

図 3: 第一主成分と第三主成分の散布図

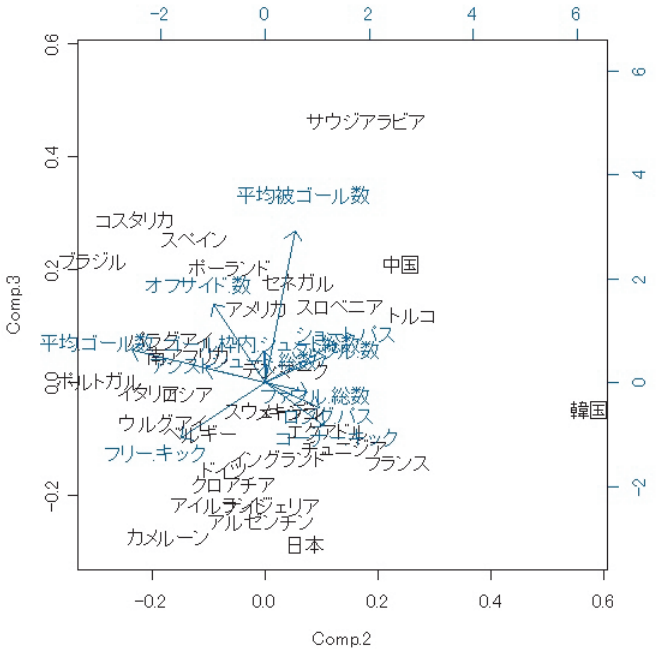

図 4: 第二主成分と第三主成分の散布図

またGUIを用いることにより三次元散布図を作ることも可能である。

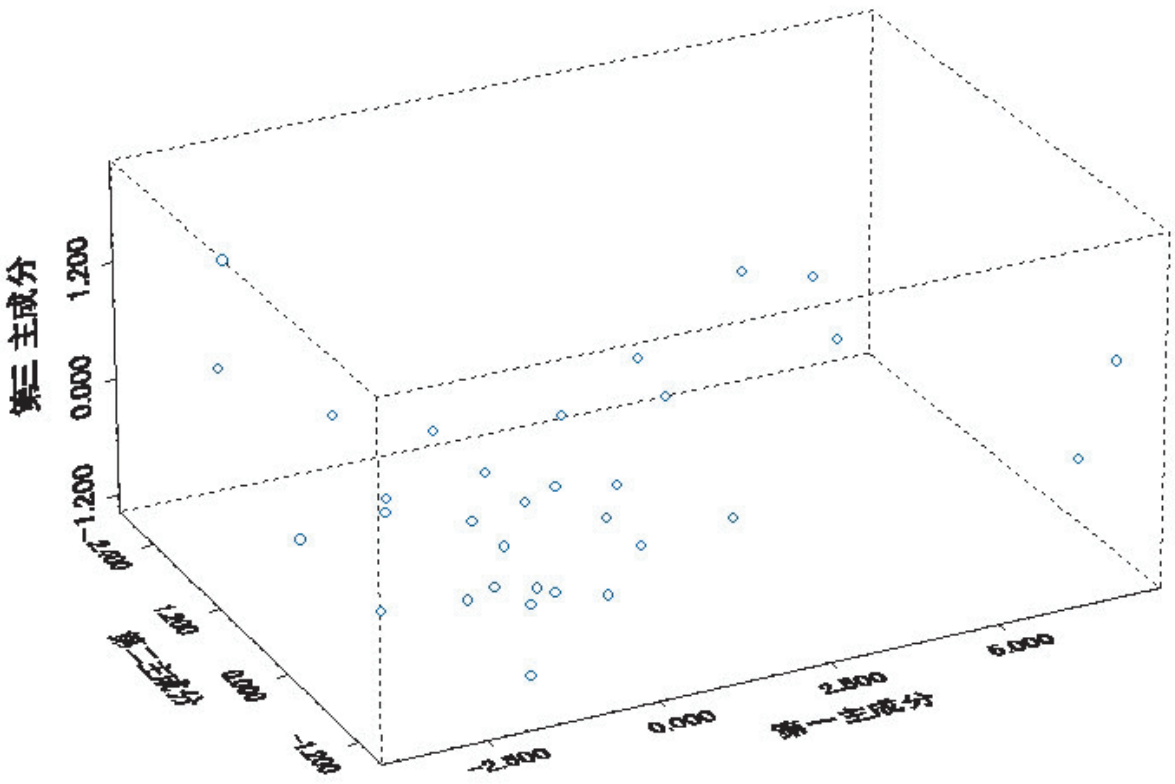

図 5: 第一主成分、第二主成分および第三主成分の三次元散布図

5.2 クラスター分析

5.2.1 分析対象

クラスター分析を行うが、すべてのチームを分析してしまっては前述の主成分分析が意味をなさないので、主成分分析で特徴付けられたチームのデータを除くことにする。ここで S-Plus の特徴を活かすことにする。

まず散布図を作るため表 5.1.3 のデータをデータフレームに保存し、また列の名前をわかりやすく変更する。

```
d <- data.frame(pr1$scores[,1:3])
#新たに作るデータフレームの名前を d とした。
rename.col( obj=d, col=1, newColName="第一主成分" )
rename.col( obj=d, col=2, newColName="第二主成分" )
rename.col( obj=d, col=3, newColName="第三主成分" )

#次に散布図を作る
plot(d$第一主成分,d$第二主成分)

#散布図を見やすくするために補助線を引く
abline(h=0,v=0,lty=2)

#特徴ある点を選択する
e <- identify(d$第一主成分,d$第二主成分,label=row.names(データ))
#このコマンドにより点を選択でき、選択したデータの番号が e に保存される

#e の中を確認する
> c(e)
 [1]  5  4  2  1  3 30 31 29 32 10 13
```

同様に第一主成分と第三主成分、第二主成分と第三主成分においても行う

```
plot(d$第一主成分,d$第三主成分)
abline(h=0,v=0,lty=2)
f <- identify(d$第一主成分,d$第三主成分,label=row.names(データ))

> c(f)
 [1]  1  2  5  4  3 31 30 13 15 27 23 28 11 29 25

plot(d$第二主成分,d$第三主成分)
abline(h=0,v=0,lty=2)
g <- identify(d$第二主成分,d$第三主成分,label=row.names(データ))
```

```
 c(g)
> [1]  5  4 30 32 31  1 10 13  3 27 11 23 28 15
```

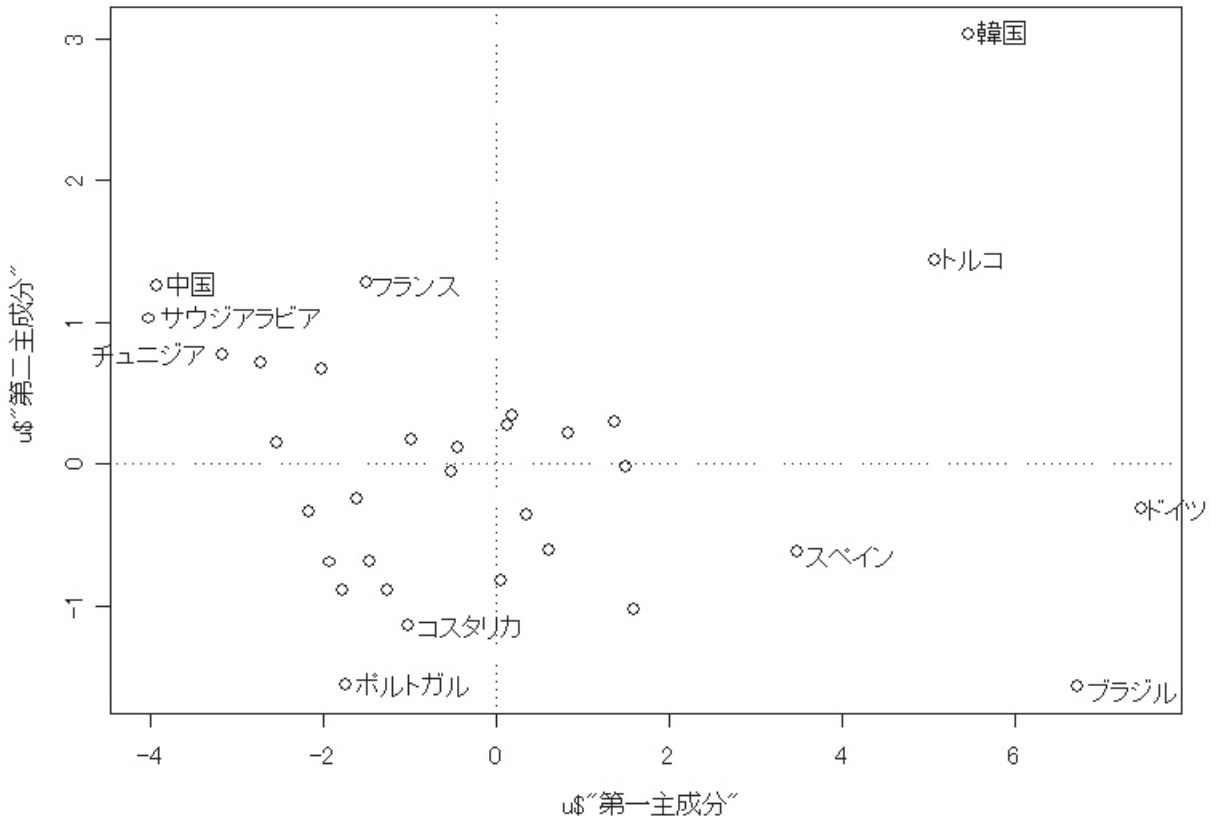

図 6: 第一主成分と第二主成分の散布図

（注）選んだ点のみにチーム名が付される。

以上のことを行うと e,f,g にそれぞれ特徴づけが可能であると考えられるチームの番号が示されている。

```
c(e,f,g)
> [1]  1  2  5  4  3 31 30 13 15 27 23 28 11 29 25  1  4  5  2  3 31 30 29
> [24] 25 28 27 15 23 13 11  5 30  4 32 31  1 13 10 27 11 23 28 15  3

#以上であげられた番号のデータを元データから除いたデータフレームを新たに作成する
h<- data.frame(データ [c(-e,-f,-g),])
```

以上より今回の分析対象を作成できた。

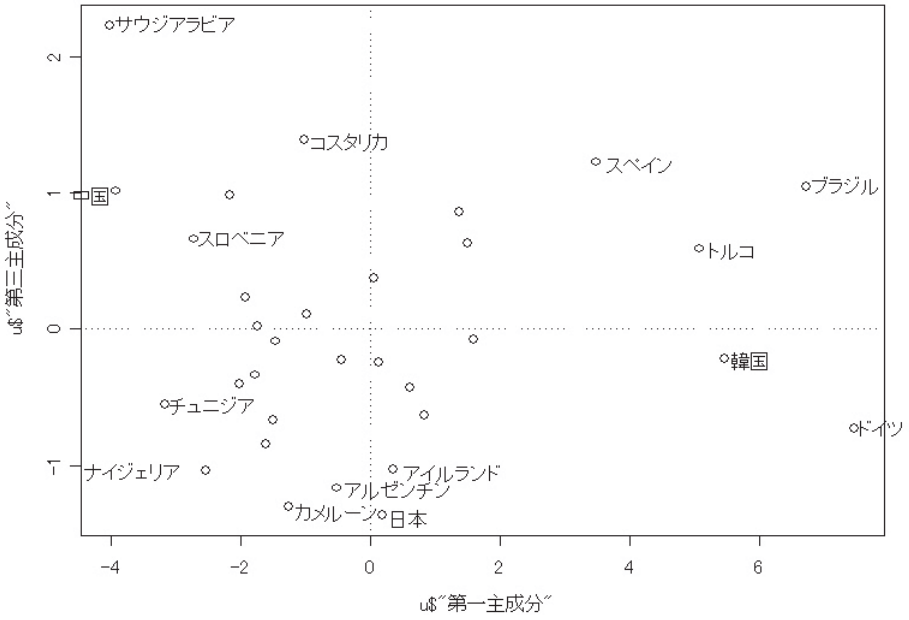

図 7: 第一主成分と第三主成分の散布図

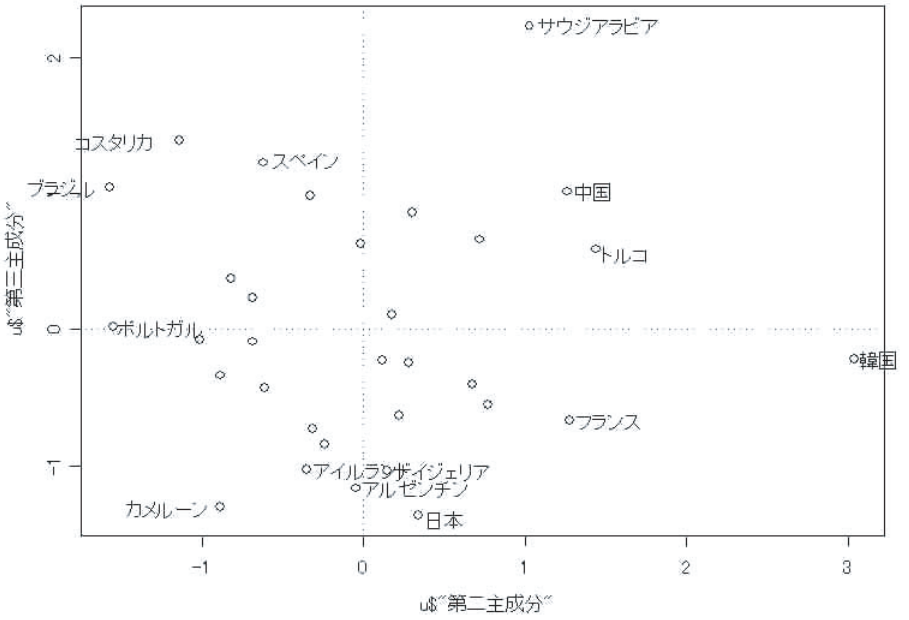

図 8: 第二主成分と第三主成分の散布図

5.2.2　クラスター分析を行いチームをクラスターにわける

前節で作成されたデータセットを用いて分析を行う

- コマンドを用いた方法

```
i <- hclust(dist(h,metric="euclidean"))
plclust(i,label=row.names(h))
#このコマンドでクラスター分析を実行
（今回はユークリッド距離を用いた最長距離法）
```

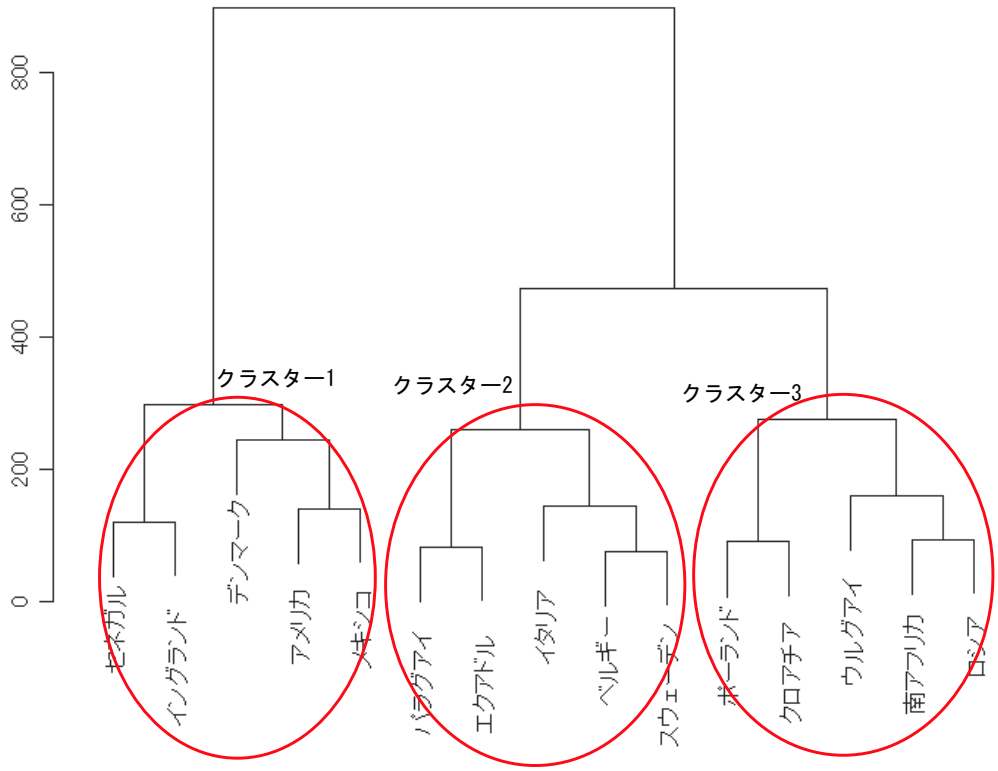

図 9: デンドログラム（最長距離法）

また距離を見るには、

```
i$height
> [1]  75.36785  81.98698  90.53109  93.97463 119.68459 141.06496 144.56595
> [8] 160.07846 245.47734 260.71307 277.24878 297.88660 473.66190 899.69702
```

で求められる。

（注）クラスター間の距離が短い順になっている。

- GUI を用いた方法

  GUI を用いて分析を行うとさらに距離に関するグラフを見ることができる

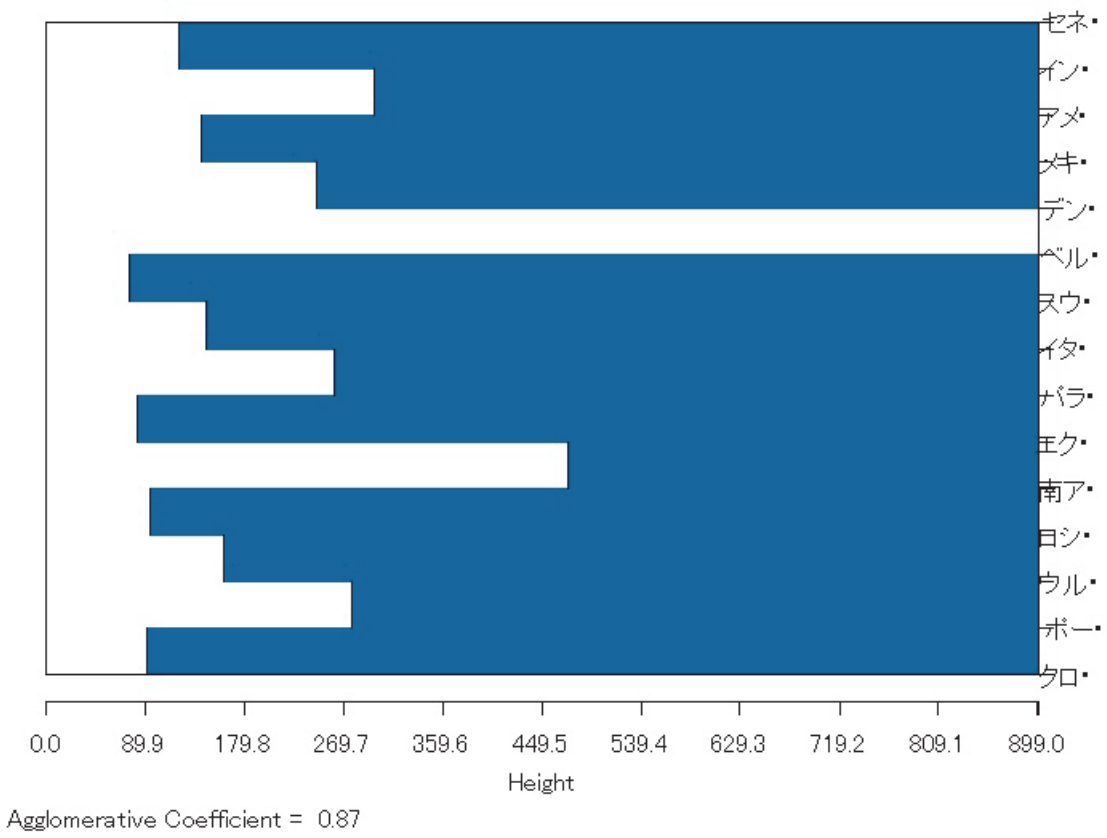

図 10: 各クラスター間の距離

5.2.3 クラスターよりチームの特徴づけを行う

図 9・10 により視覚的にいくつかのクラスターに分けることが可能であるように思える。

クラスター 1 の各チームは、前述の主成分分析で取り上げられた数チーム（ブラジル・ドイツなど）ほどではないにせよ成績が決勝トーナメントに進出したチームであることが見て取れる。

クラスター 2 の各チームは、イタリアをはじめとして堅守を特色としているチームであることが見て取れる。

クラスター 3 の各チームはこれも先ほどと同様に主成分分析ではわかりにくかったが成績が芳しくないチームであることが見て取れる。また、パス数が少ないチームであることが見て取れた。

6　まとめ

以上よりW杯出場国32チームを分類することができた。主成分分析では分類をするのが困難であったチームをクラスター分析において分類ができることを示した。今回はデータ数が少ないものであったがこの方法を使うことにより膨大な量のデータにおいても分類が容易に行うことが可能になると思われる。ここで最初からクラスター分析を行わない理由として、クラスター分析では分類した後にデータを参照して類似点を見つけなければいけないためデータ数が多い場合はとても大変な作業となる。一方主成分分析では主成分得点を見ることにより類似点などを見つけるのはたやすい。見つけたデータを減らすことでクラスター分析の際の分類の手数を減らすという点で有効な手法であるのではないかと考えられる。

今回はS-Plusを用いたが主成分分析での分類において特徴ある点を選び、そのデータを抜いてクラスター分析に行くという流れのときにS-Plusのコマンドは非常に有益であった。また三次元プロットは他のソフトでも行えるがビジュアル的にS-Plusは非常にわかりやすいという長所がある。そして解析で一番助かった点は、時間がかからないという点であった。この処理の早さがS-Plusの特徴であることを実感した。

参考文献

[1] R.A. ベッカー・J.M. チェンバーズ・A.R. ウィルクス,「S 言語（I）」, 共立出版 (2000)

[2] R.A. ベッカー・J.M. チェンバーズ・A.R. ウィルクス,「S 言語（II）」, 共立出版 (2002)

[3] J.M. チェンバーズ・T.J. ヘイスティ,「S と統計モデル」, 共立出版 (2002)

[4] 数理システム,「S-PLUS for windows 入門」, 数理システム (2001)

[5] W.N. ヴェナブルズ・B.D. リプリー,「S-PLUS による統計解析」, シュプリンガー・フェアラーク東京 (2001)

[6] 永田　靖・棟近　雅彦,「多変量解析法入門」, サイエンス社 (2003)

[7] 「FIFAworldcup.com」http://fifaworldcup.yahoo.com/02/jp//index.html

[8] 「TSP21.com」http://www.tsp21.com/

表 4: 今回使用したデータ (1)

チーム	平均ゴール数	シュート総数	ゴール枠内シュート総数	アシスト数	コーナーキック数	フリーキック数
ブラジル	2.57	93	54	9	34	17
ドイツ	2	100	40	12	49	20
スペイン	2	70	36	5	28	9
トルコ	1.43	62	32	7	39	9
韓国	1.14	89	44	3	53	8
セネガル	1.4	54	23	3	22	3
アメリカ	1.4	54	30	5	21	7
ベルギー	1.5	48	22	4	27	14
パラグアイ	1.5	50	24	2	17	11
ポルトガル	2	32	16	3	12	4
アイルランド	1.5	45	20	4	21	8
イングランド	1.2	52	24	4	20	5
コスタリカ	1.67	43	18	4	19	3
デンマーク	1.25	29	15	4	19	1
日本	1.25	37	17	3	20	8
スウェーデン	1.25	39	22	1	17	6
南アフリカ	1.67	32	17	2	12	3
イタリア	1.25	48	26	4	25	16
メキシコ	1	38	19	2	16	8
ウルグアイ	1.33	34	18	1	19	10
ロシア	1.33	41	17	1	10	6
ポーランド	1	32	14	2	15	4
アルゼンチン	0.67	45	18	1	33	9
エクアドル	0.67	25	13	2	15	2
スロベニア	0.67	33	17	1	12	2
クロアチア	0.67	28	16	2	23	7
カメルーン	0.67	30	14	2	20	15
ナイジェリア	0.33	31	11	1	11	6
チュニジア	0.33	22	6	0	17	3
中国	0	19	10	0	16	3
サウジアラビア	0	26	10	0	4	5
フランス	0	42	26	0	24	4

表 5: 今回使用したデータ (2)

チーム	オフサイド数	ショートパス数	ロングパス数	平均被ゴール数	ファウル総数	タックル数
ブラジル	26	2082	710	0.57	106	299
ドイツ	16	1848	903	0.43	133	338
スペイン	25	2014	708	1	74	214
トルコ	21	2247	846	0.86	119	355
韓国	12	2348	894	0.86	133	319
セネガル	21	1197	611	1.2	101	236
アメリカ	14	1380	589	1.4	90	217
ベルギー	7	951	587	1.75	84	174
パラグアイ	13	1061	528	1.75	80	142
ポルトガル	7	874	370	1.33	54	102
アイルランド	7	1254	710	0.75	67	137
イングランド	11	1223	713	0.6	74	187
コスタリカ	16	943	355	2	48	144
デンマーク	9	1386	469	1.25	66	143
日本	4	1095	646	0.75	99	189
スウェーデン	10	888	611	1.25	70	200
南アフリカ	5	897	395	1.67	59	131
イタリア	20	823	572	1.25	85	238
メキシコ	14	1450	702	1	63	184
ウルグアイ	6	826	315	1.67	57	123
ロシア	10	846	473	1.33	64	130
ポーランド	12	625	345	2.33	64	148
アルゼンチン	15	971	461	0.67	52	123
エクアドル	6	995	498	1.33	59	154
スロベニア	3	874	338	2.33	57	166
クロアチア	11	712	366	1	55	145
カメルーン	11	871	420	1	55	118
ナイジェリア	7	727	440	1	46	137
チュニジア	6	609	408	1.67	57	138
中国	5	949	274	3	41	135
サウジアラビア	8	938	461	4	39	114
フランス	9	1223	405	1	49	129