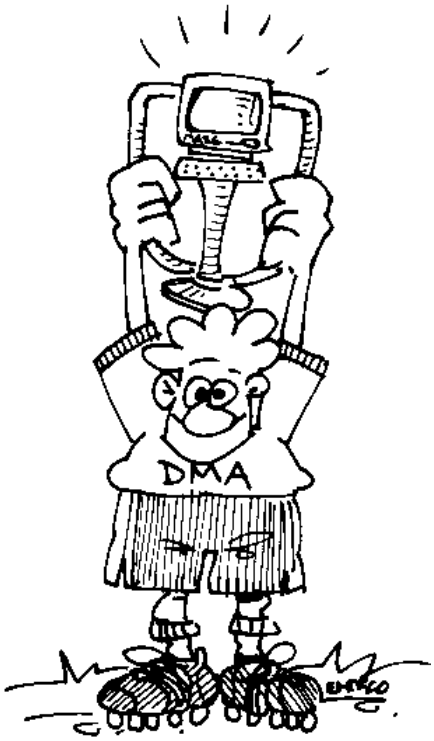
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Statistical Models for Knock-out Soccer Tournaments



Diego Kuonen
Department of Mathematics
Chair of Applied Statistics

Prof. S. Morgenthaler, DMA, EPF Lausanne
Assistant: E. Chavez, DMA, EPF Lausanne

Winter 1996/97

Table of Contents

1 Introduction **1**

1.1 Preliminary 1

1.2 European Cups 1

1.2.1 Champions Cup 1

1.2.2 Cup Winners Cup 2

1.2.3 UEFA Cup 2

1.2.4 UEFA Intertoto Cup 4

2 Descriptive Data Analysis **5**

2.1 Data Structure 5

2.2 Tree Structure 6

2.3 Seeding Coefficients 8

2.3.1 First Approach 9

2.3.2 Improvement 11

2.4 Tournament Analysis 13

2.4.1 Possible Sets of Winners 13

2.4.2 Probability of Winning 14

2.4.3 Potential Opponents 16

3 Statistical Modelisation **17**

3.1 Logistic Regression Model 17

3.2 Constancy of Team Strength 17

3.2.1 Model Chosen 18

3.2.2 The Program 20

3.2.3 Number of Correct Predictions 23

3.3 Variation of Team Strength 25

3.3.1 Model Chosen 25

3.3.2 Using Probable Opponents 26

3.3.3 Using All Potential Opponents 29
3.3.4 Number of Correct Predictions 32
3.3.5 Evaluation of the Coefficients 34

4 Comparison of the Methods **36**

5 Conclusion **39**

A Appendix **40**
A.1 The Bradley-Terry Model 40
A.2 Tournament Trees 41
A.2.1 Cup Winners Cup 41
A.2.2 UEFA Cup 48
A.3 Evaluation of the Coefficients 55
A.3.1 Using Probable Opponents 55
A.3.2 Using All Potential Opponents 59

B References **63**

1 Introduction

1.1 Preliminary

Sports events and tournament competitions provide excellent opportunities for model building and using basic statistical methodology in an interesting way. In this paper, a logistic regression model using seed positions (conceived through a seeding coefficient) is applied to European soccer Cups tournament data in order to predict the probability of winning the tournament for each one of the participating teams, and the predicted probabilities of each team reaching a certain leg such as the quarter final.

1.2 European Cups

This section was coined by M. Protzen (1996).

There are four different European Cups

- the Champions Cup (CC),
- the Cup Winners Cup (CWC),
- the UEFA Cup (UC), originally intended for League runners up,
- and the UEFA Intertoto Cup (UIC).

The qualification for these competitions depends on the performance in National Leagues Cup competitions respective. The Champion usually enters the CC, the Cup Winner (losing finalist in case of a double respective) enters the CWC, while a variable number of League runners up enter the UC (in some countries winners of a League Cup enter the UC as well). Teams finishing below those qualifying for the UC may enter the UIC.

1.2.1 Champions Cup

Originally this Cup was intended for the league winners of the member FAs of UEFA. A while ago the qualification process for the CC has been modified such that only the CC defender together with the top 23 champions enters the CC while all other champions will enter the UC. The CC defender and the 7 best champions advance directly to the league stage while the champions ranked 8th to 15th have to play a preliminary tie against one of the champions ranked 16th to 23rd. To determine the top champions (and the number of UC berths allocated to a country) UEFA maintains a ranking list.

New format for 1997/98

For the upcoming season UEFA has once again changed the format to allow the vice champions of the 8 top ranked countries as well as the champions of up to 48 countries into the competition.

The league stage will be played in six groups of four teams. From there the six group winners together with two best runners-up will advance to the quarter finals. The two best runners-up will be determined by points achieved, goal difference, goals scored and goals scored away.

To cut down the number of teams the champions of countries ranked 17 to 48 in the most recent ranking table play a preliminary round in late July. For the losers of this tie their European campaign is over. Winners play in the qualification round together with champions of countries ranked 9 to 16 and vice champions of countries ranked 1 to 8. Winners of this round advance to the league stage where they will be joined by the champions of the top ranked countries. Losers will play in the first round of the UEFA Cup.

The defending Cup Winner gets an automatic place in the league stage regardless where he finishes in his national competition. This may mean that the champions of the countries ranked 8th, 16th (and 48th) are 'downgraded'.

1.2.2 Cup Winners Cup

The Cup Winners of the FA Cups of the member FAs of UEFA get to play in this Cup unless they also win the domestic championship (in which case they will be replaced by the losing cup finalists). The CWC defender is also eligible to play (unless he becomes national champion).

UEFA is also considering to allow a second team from the eight top ranked countries to play the CWC. How the second team will be determined has not yet been fixed. This modification will not effect for 96/97 and 97/98 competitions.

1.2.3 UEFA Cup

The UEFA Cup started (under the name "Fairs Cup") as a competition between city teams from cities which host a trade fair. Later it was a competition for the runners up of the various leagues, some countries enter also their League Cup winner. The number of participants from each country is determined by the UEFA ranking list. Since the Champions League was introduced the champions not eligible for the Champions League are entered into the UEFA Cup as well, another three berths are reserved for the Intertoto Cup semi finalists and three are given to countries who top the UEFA Fair Play competition.

- 79 participants are determined by the UEFA coefficient table:
  - 3 countries with four berths
  - 5 countries with three berths
  - 13 countries with two berths
  - 26 countries with one berth
- 24 champions not entering the CC
- 3 Intertoto Cup participants
- 3 Fair Play competition winners

  The three countries which top the UEFA Fair Play competition each get one additional UEFA cup berth in the next competition. The 1995/96 berths were allocated to Norway, England and Luxembourg, the 1996/97 berths have been awarded to Sweden, Russia and Finland (one berth originally awarded to England has been revoked because of fielding youth and reserve teams in the 1995 UIC) and in 1997/98 Norway, England and Sweden will get the berths.
- 1 defending UEFA Cup Holder (only if they do not qualify for any European competition otherwise)
- 8 preliminary round losers from the Champions Cup (since 1996/97)

Altogether, there are 118 participants.

For the *96/97 competition* the following format will be used:

- Preliminary round in July, involving 54 teams from countries with the lowest country coefficient in the 95/96 ranking
- Qualifying round in August, involving 50 teams (27 winners from the qualifying round, 23 teams from the countries with low coefficients which did not play in the preliminary round) + 6 Intertoto teams.
- First round in September, involving 64 teams (25 winners from the preliminary round, 8 teams eliminated from the Champions Cup, 28 teams from countries with high rankings and 3 teams qualifying through the UIC).

Barring participation of teams from San Marino, Andorra and Bosnia the 97/98 competition will have 16 participants fewer: 24 champions of low ranked countries will not enter the UC but 16 instead of 8 CC qualification round losers will play in the first round.

From *1997/98 on* the final will be decided in a single game on neutral ground.

1.2.4 UEFA Intertoto Cup

The Intertoto Cup started as a way to guarantee income to the Toto-industry in the early sixties (in fact, this generated a significant sum of money for clubs in several countries, one of them the Netherlands). The first one or two seasons the group winners played out a final stage to determine an overall winner (the first being Ajax). After a few years the competition degenerated into summer practice for mid table teams of North, Central and East European countries. Teams from the British Islands, Spain or Italy rarely participated. No final stage was played after the initial (one or two) season(s) and nobody really cared for it. This new directive by UEFA seems like a final attempt of attracting some attention to it.

The 'new' UEFA Intertoto-Cup has started in the summer of 1995. The cup will be played during the summer, and the teams allowed to participate should have been the best finishing teams in the respective leagues the previous season not already qualified for any of the other three European Cups. The number of teams from each country is determined by the UEFA ranking of countries just as in the UC (with the exception that also the bottom clubs are entitled to one berth).

60 teams are divided into 12 groups with the winners of each group qualifying for a knock-out round, from where the best 3 teams qualify for the first round proper of the UEFA Cup the following autumn.

From the 1995 UIC two French teams, Racing Strasbourg and Girondins de Bordeaux, emerged successful and participated in the UC, Bordeaux even reached the UC final.

A number of changes have been announced for the 1996 UIC: The number of UC berths available through this competition has been raised to 3, Italian teams will join the competition and teams are urged to take the competition serious, i.e. play with their first team.

Some countries do not take part in the competition: England, Portugal, Scotland and Luxembourg. Italy sends only three teams, Spain and Greece only one. Italy and Spain later renounced their remaining berths for the UIC, they were heavily fined for that. Also Greece and Albania did not take their berths. Those six berths went to Estonia, Lithuania, Austria, Turkey, Denmark and Sweden.

Also the format of the cup has changed slightly: Again there will be 12 groups with 5 teams each. But fortunately the concept of "best seconds" has been cancelled: The 12 winner will play one elimination round with the 6 winners playing another round as part of the UC preliminary round.

In 1996 Silkeborg IF, Karlsruher SC and EA Guingamp qualified for the UC.

2 Descriptive Data Analysis

2.1 Data Structure

As said in the introduction European soccer Cups tournament data is used. We only considered the data from the Champions Cup (CC), from the Cup Winners Cup (CWC) and from the UEFA Cup (UC). The typical tournament structure does not begin until after the preliminary round(s), so we did not consider them. In this context the structure of an ordinary data looks like this:

```
leg    team.a  country.a  team.b  country.b  s.a.1  s.b.1  s.a.2  s.b.2
Round1 FKAustria(Wien)    Aut  DinamoMinsk            Bls    1  2  0  1
Round1 Milan              Ita  ZaglebieLubin          Pol    4  0  4  1
...
Round1 SpartakVladikavkaz Rus  Liverpool              Eng    1  2  0  0
Round1 ZimbruChisinau     Mol  RAFYelgava             Lat    1  0  2  1
Round1 RodaJC(Kerkrade)   Ned  SCTOlimpija(Ljubljana) Slo    5  0  0  2
Round2 BrondbyIF          Den  Liverpool              Eng    0  0  1  0
...
Round2 SpartaPraha        Tch  ZimbruChisinau         Mol    4  3  2  0
Round2 GirondinsBordeaux  Fra  RotorVolgograd         Rus    2  1  1  0
Round3 PSV(Eindhoven)     Ned  WerderBremen           Ger    2  1  0  0
Round3 BayernMuenchen     Ger  SLBenfica              Por    4  1  3  1
...
Round3 Sevilla            Esp  FCBarcelona            Esp    1  1  1  3
Round3 SlaviaPraha        Tch  RCLens                 Fra    0  0  1  0
Quarter FCBarcelona       Esp  PSV(Eindhoven)         Ned    2  2  3  2
Quarter SlaviaPraha       Tch  Roma                   Ita    2  0  1  3
Quarter Milan             Ita  GirondinsBordeaux      Fra    2  0  0  3
Quarter BayernMuenchen    Ger  NottinghamForest       Eng    2  1  5  1
Semi BayernMuenchen       Ger  FCBarcelona            Esp    2  2  2  1
Semi SlaviaPraha          Tch  GirondinsBordeaux      Fra    0  1  0  1
Final BayernMuenchen      Ger   GirondinsBordeaux     Fra    2  0  3  1
```

where `s.a.1:s.b.1` is for example the score of the 1st leg. This list represents the data from the UEFA Cup 1995/96.

2.2 Tree Structure

An example of a typical tournament structure is represented by figure 1. The tree structure is cognizable. In this section we want to describe how we obtained the tournament tree through the data. So let us have a look at figure 1. We always start with the winner of the tournament, in this case it is A. Then we will look for the team A played in the previous round, that would be in this case C. In the future C is marked. Now we are looking for the team A played against before playing against C, that would be B. We realize that the game A against B had be played in the first round of the tournament, so we are going back to the last marked team, that would be C, and so on.

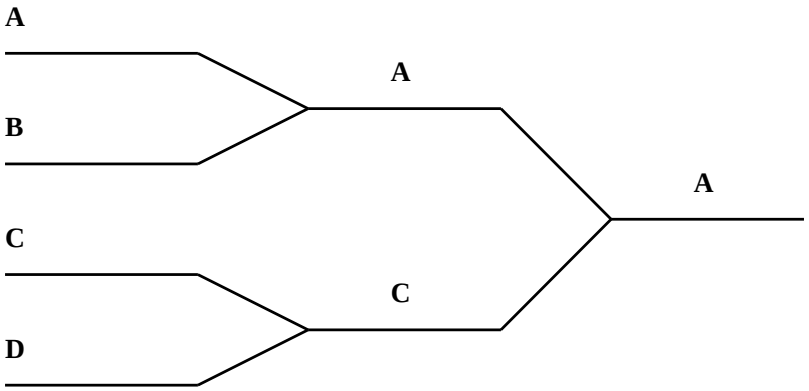


Figure 1 *Example of a typical tournament structure.*

For this reason we have programmed a Splus routine titled `pairings.tree`. In applying this routine to the Cup Winners Cup of the year 1989/90 we obtain figure 2.

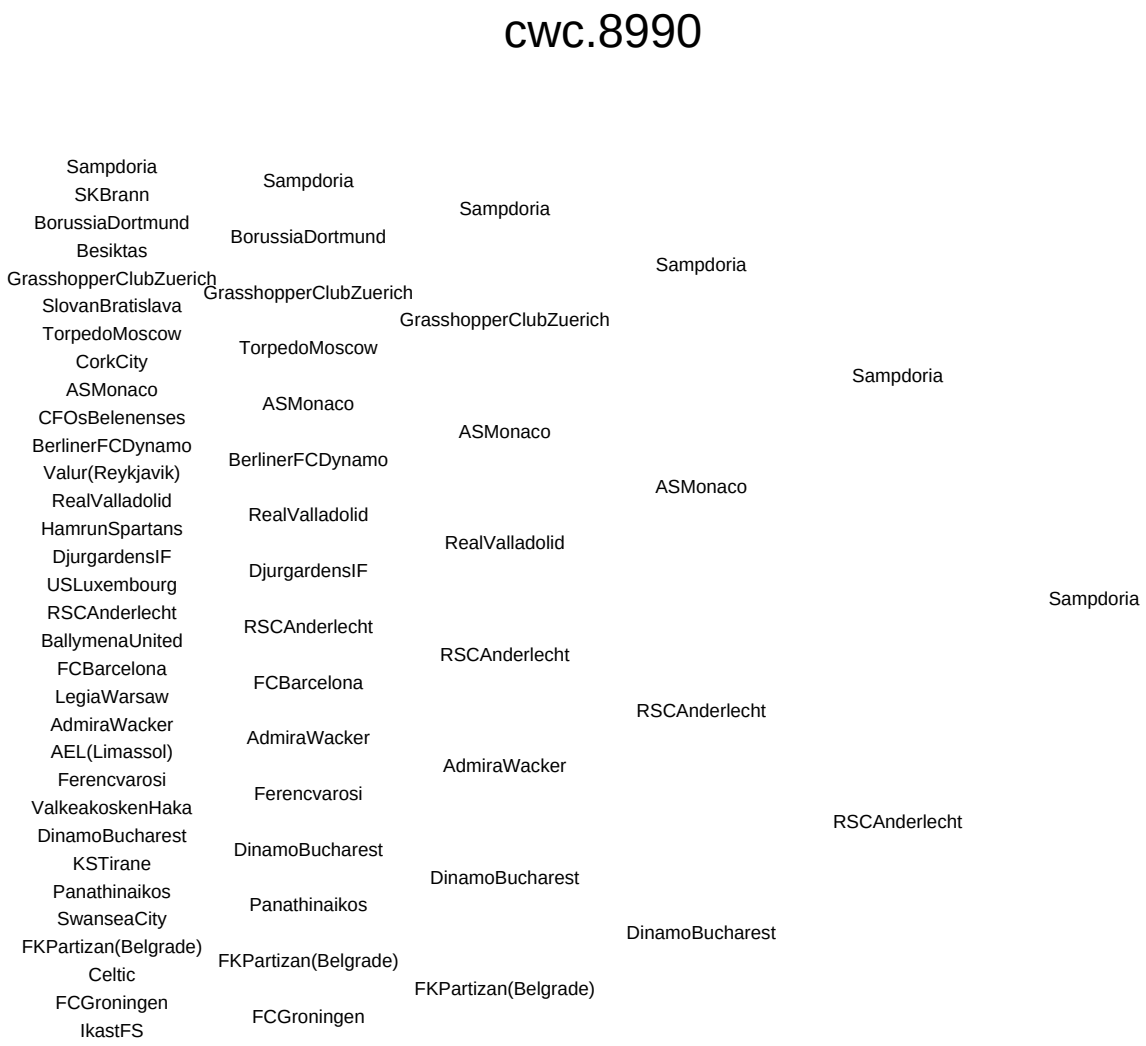


Figure 2 *The Cup Winners Cup 1989/90.*

In the appendix you will find the tournament trees of all European Cups considered.

2.3 Seeding Coefficients

To obtain the seed positions we calculated the seeding coefficients with the help of a ranking list. The basis for this ranking is the performance of teams in the three major European Cups during a three year period. Each team gets two points for a win and one point for a draw. One bonus point is allocated for reaching the quarter final, the semi final and the final. This procedure is similar to the one the UEFA uses for their ranking list. In the first time we calculated a coefficient which will be replaced later on, by a more appropriated coefficient: A weighted coefficient. Both coefficients are based on a list containing the performances during an European Cup. An example is the following list which shows the points achieved for the participants during the Champions Cup of the year 1995/96:

```
    THE TABLE FOR  cc.9596
    ----------------------------------
                      team country win draw lost scored  against points
 1                    Ajax     Ned   8    3    1     22        3     22
 2                Juventus     Ita   6    3    3     22        9     18
 3           Panathinaikos     Gre   5    3    2     11        6     15
 4           SpartakMoscow     Rus   6    1    1     17        8     14
 5                FCNantes     Fra   4    4    2     14       12     14
 6              RealMadrid     Esp   4    1    3     12        7     10
 7       BorussiaDortmund     Ger   2    3    3      8       11      8
 8             LegiaWarsaw     Pol   2    2    4      5       11      7
 9                 FCPorto     Por   1    4    1      6        5      6
10         SteauaBucharest     Rom   1    2    3      1        5      4
11             RosenborgBK     Nor   2    0    4     11       16      4
12                 Rangers     Sco   1    2    3      6       13      4
13            Ferencvarosi     Hun   1    2    3      9       19      4
14         BlackburnRovers     Eng   1    1    4      5        8      3
15            AaB(Aalborg)     Den   1    1    4      5       11      3
16 GrasshopperClubZuerich     Sui   0    2    4      3       13      2
```

2.3.1 First Approach

In this approach we calculated the performances of the teams over a three year period and summarized them into a list. The following list shows the points achieved for all teams who participated in an European Cup from 1992 to 1995.

	team	country	win	draw	lost	scored	against	points
1	Milan	Ita	23	8	4	55	10	60
2	Juventus	Ita	22	5	5	68	22	56
3	Parma	Ita	18	8	6	37	15	53
4	Ajax	Ned	17	6	3	47	14	45
5	ParisSaintGermain	Fra	15	6	7	39	19	42
6	BorussiaDortmund	Ger	16	4	10	43	30	42
7	SLBenfica	Por	13	7	3	43	23	37
8	FCPorto	Por	14	5	8	43	20	35
9	Arsenal	Eng	11	7	2	35	15	35
10	FCBarcelona	Esp	11	7	6	42	27	32
11	SpartakMoscow	Rus	12	6	5	46	31	32
12	RealMadrid	Esp	11	4	5	40	23	28
13	EintrachtFrankfurt	Ger	10	4	6	35	14	26
14	BayerLeverkusen	Ger	9	4	3	41	19	25
15	IFKGothenburg	Swe	10	3	5	26	18	24
16	AJAuxerre	Fra	8	5	5	37	21	24
17	OlympiqueMarseille	Fra	10	3	2	26	9	24
18	Internazionale	Ita	10	1	3	23	11	24
19	RSCAnderlecht	Bel	7	9	5	31	24	23
20	RealZaragoza	Esp	9	1	6	31	17	22
21	ClubBruggeKV	Bel	9	2	5	19	15	21
22	Boavista	Por	7	5	4	22	11	20
23	Galatasaray	Tur	6	7	9	17	29	19
24	Lazio	Ita	8	2	2	15	7	19
...								
211	KSVWaregem	Bel	0	0	2	1	6	0
212	DACDunajskaStreda	Tch	0	0	2	0	4	0
213	Motherwell	Sco	0	0	2	0	3	0

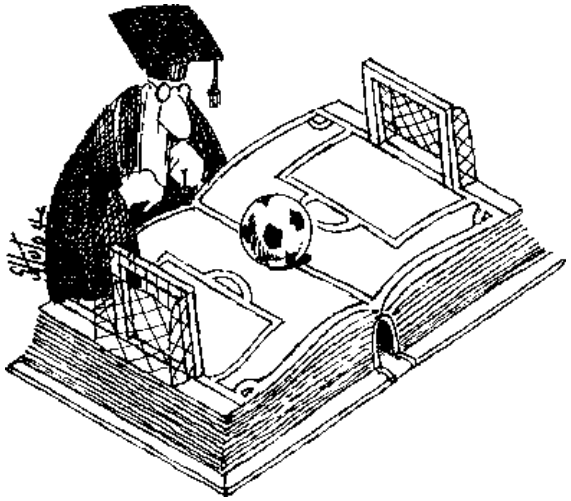
This ranking list over a three year period is now used to calculate the seeding coefficients. We defined the coefficient in the following way:

$$\text{Coefficient} = \frac{\text{Points achieved during the three year period}}{\text{Games played during the three year period}}$$

This is different from the UEFA coefficients, because the UEFA uses the sum of the ratio points achieved over games played for each of the five past years. An example of our calculations is the following list which shows the seeding coefficients for the UEFA Cup of the year 1993/4.

```
                  team country     coeff
 1             Juventus     Ita 2.4000000
 2       AtleticoMadrid     Esp 2.0625000
 3       BayernMuenchen     Ger 1.9166667
 4            AJAuxerre     Fra 1.8571429
 5     BorussiaDortmund     Ger 1.7222222
 6   SportingCP(Lisbon)     Por 1.6428571
 7       Internazionale     Ita 1.6428571
 8            BrondbyIF     Den 1.5000000
 9           AstonVilla     Eng 1.5000000
10         RoyalAntwerp     Bel 1.4166667
11   Olympiakos(Pireus)     Gre 1.4000000
12         AdmiraWacker     Aut 1.4000000
13      Vitesse(Arnhem)     Ned 1.2500000
14           Trabzonspor     Tur 1.1666667
15    GirondinsBordeaux     Fra 1.1666667
16   EintrachtFrankfurt     Ger 1.1000000
17           SlaviaPraha     Tch 1.0000000
18         DinamoMoscow     Rus 1.0000000
...
63         BSCYoungBoys     Sui 0.0000000
64         AaB(Aalborg)     Den 0.0000000
```

This approach was not satisfiable because there is a problem with teams which were playing in the Champions Cup. In the Champions Cup less games are played and so the teams will achieve less points. That is why their seeding coefficient would be smaller, even if they played in the most important European Cup. By calculating the probability for a team of winning the tournament for example (in using the program `prob.cte` (cf. section 3.2)) for the UEFA Cup 1995/96 we noticed that Girondins Bordeaux (Fra) had more chances to win the tournament than Barcelona (Esp). For every soccer fan this seems to be impossible, so for us. For this reason we did not use this approach to calculate the seeding coefficients.



2.3.2 Improvement

The improvement consists in calculating a weighted mean of the ratio points achieved over games played for each of the three past years, let be C_x this ratio for the year x. Hence we defined the coefficient for the year x as

$$\boxed{\text{Coefficient} = \tfrac{3}{6}C_{x-1} + \tfrac{2}{6}C_{x-2} + \tfrac{1}{6}C_{x-3}.}$$

With this choice of the weights we give most importance on the ratio of the previous year. If for example a team plays in a given year's UEFA Cup, but it did not play any European Cup for the last two years, the team will get a seeding coefficient (in using the previous method (see 2.3.1)) which seems to be too high because they played for the last time in the European Cup three years before - that is nonsense. This problem is absorbed by the improved coefficient because there will be the weight 1/6. That is why from now on this weighted seeding coefficient (briefly coefficient) is used. The following list shows the coefficient for the Cup Winners Cup of the year 1992/93:

	team	country	coeff
1	ASMonaco	Fra	1.44722222
2	AtleticoMadrid	Esp	1.33333333
3	SpartakMoscow	Rus	1.25000000
4	WerderBremen	Ger	1.21666667
5	Trabzonspor	Tur	1.00000000
6	SteauaBucharest	Rom	1.00000000
7	Feyenoord	Ned	0.91666667
8	AdmiraWacker	Aut	0.72222222
9	SpartaPraha	Tch	0.70833333
10	Glenavon	Nir	0.66666667
11	Boavista	Por	0.58333333
12	Liverpool	Eng	0.56250000
13	Parma	Ita	0.50000000
14	Olympiakos(Pireus)	Gre	0.50000000
15	RoyalAntwerp	Bel	0.37500000
16	FCLucerne	Sui	0.33333333
17	AvenirBeggen	Lux	0.33333333
18	Apollon(Limassol)	Cyp	0.33333333
19	Valur(Reykjavik)	Isl	0.25000000
20	ChernomoretsOdessa	Ukr	0.25000000
21	FCLevski(Sofia)	Bul	0.08333333
22	UjpestiTE	Hun	0.00000000
23	TPS(Turku)	Fin	0.00000000
24	MiedzLegnica	Pol	0.00000000
25	MariborBranik	Slo	0.00000000
26	HapoelPetachTikva	Isr	0.00000000
27	Hannover96	Ger	0.00000000

```
28        CardiffCity     Wal 0.00000000
29          Bohemians     Irl 0.00000000
30      Airdrieonians     Sco 0.00000000
31     AIK(Stockholm)     Swe 0.00000000
32         AGF(Arhus)     Den 0.00000000
```

Figure 3 shows the coefficient for different three year periods for some chosen teams. Let us have a look at the evolution of the coefficient of Juventus (Ita): Between the periods 1989/92 and 1990/93 the coefficient increased because they won the UEFA Cup in 1992/93 (yes, Juventus also wins the UEFA Cup not only the Champions Cup) in achieving 24 points. Between the periods 1990/93 and 1991/94 the coefficient decreased because in the 1993/94 edition of the UEFA Cup Juventus reached the quarter final in achieving in 8 games only 10 points (4 wins, 1 draw and 3 lost). And between the periods 1991/94 and 1992/95 he increased again because they won the UEFA Cup in 1994/95 again before becoming Italian champion. We also remark the constancy of the coefficient of Barcelona (Esp) who was the dominating team in European Cups during this periods.

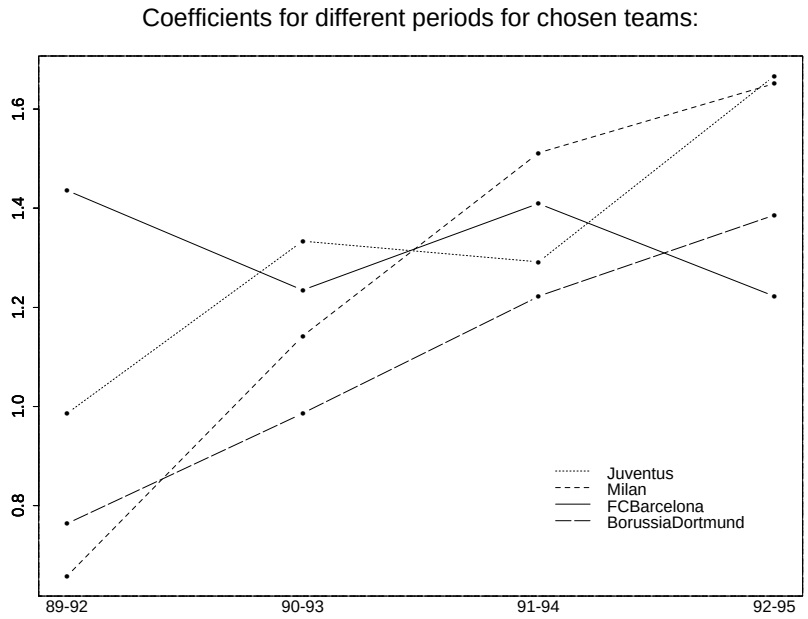


Figure 3 *The coefficient for different periods for some chosen teams.*

2.4 Tournament Analysis

2.4.1 Possible Sets of Winners

Predicting the probability of each seed winning the tournament requires the consideration of all possible paths and opponents. An example of a tournament is shown by figure 4.

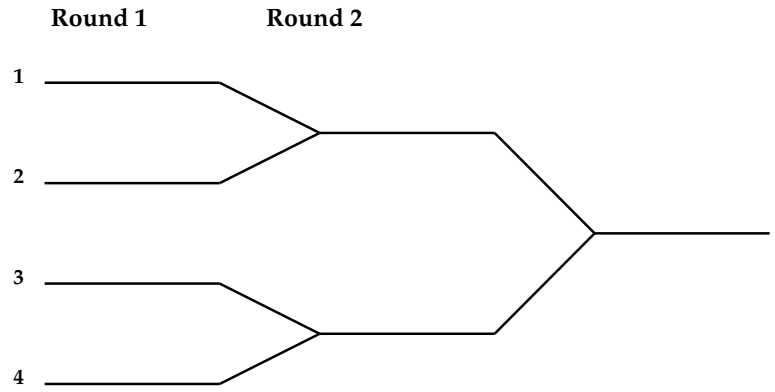


Figure 4 *Example of a tournament.*

Let us determine the teams that must be played for a team to become the champion. In Round 1 there are 2^2 possible sets of winners and in Round 2 there are 2 possible outcomes. Hence there are $2^2 2 = 2^3$ possible outcomes for the 3 games.

By generalizing, we obtain the following table:

Round number	Games per round	Possible sets of winners
1	$n/2$	$2^{n/2}$
2	$n/4$	$2^{n/4}$
$\vdots$	$\vdots$	$\vdots$
m	$n/(2^m)$	$2^{n/(2^m)}$

where n is the number of teams participating in the tournament. From this it follows that in a tournament with n teams (and so $n-1$ games) there are

$$\prod_{k=1}^{m} 2^{n/(2^k)}$$

possible sets of winners for the tournament, where m is the number of legs to play for reaching the final game.

Remark: In fact $m = \log_2 n$, because of the following relationship: $n = 2^m$.

2.4.2 Probability of Winning

Let $P_k(i,j)$ be the probability that team i wins against team j in the kth leg ($\forall i \neq j$). As in a tournament there is always a winner (i.e. there are any draws), it follows that $P_k(j,i) = 1 - P_k(i,j)$.

In the first round each team has only one possible opponent, but in the second round there are two possible opponents as shown in the previous section. That is why the probability analysis must include not only the probability of defeating each potential opponent, but also the probability of each potential opponent advancing to a particular game. To illustrate, let us have a look at figure 5.

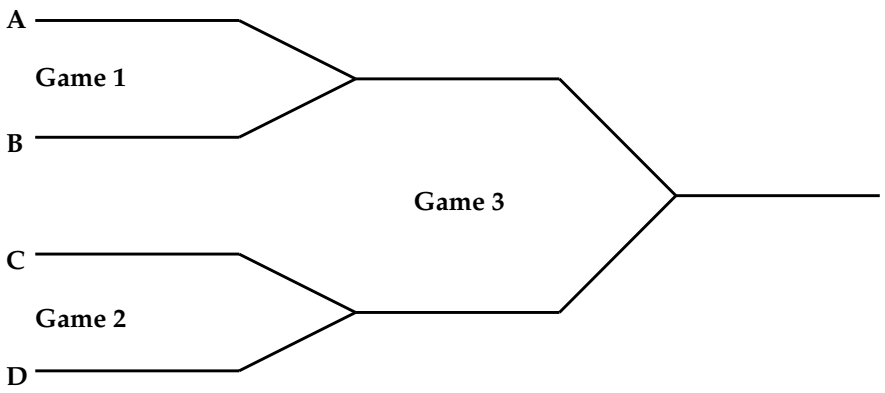


Figure 5 *Example of a tournament.*

Suppose that team A wins the tournament by defeating team B in game 1 and the winner of game 2 (C or D) in game 3, then the probability that team A wins the tournament would be :

$$\begin{aligned}
P(A \text{ wins the tournament}) &= P(A \text{ wins in game } 3) \\
&= P_1(A,B)P_2(A, \text{winner of game } 2) \\
&= P_1(A,B)\{P_1(C,D)P_2(A,C) + P_1(D,C)P_2(A,D)\} \\
&= P(A \text{ wins in leg } 1)\{P(C \text{ wins in leg } 1)P_2(A,C) + \\
&\quad P(D \text{ wins in leg } 1)P_2(A,D)\} \\
&= P(A \text{ wins in leg } 1) \sum_{j \in \{C,D\}} P(j \text{ wins in leg } 1)P_2(A,j)
\end{aligned}$$

This probability is the probability of A winning the tournament and as well the probability of A winning in game 3, *i.e.* in leg $m = \log_2 n = \log_2 4 = \frac{\log 4}{\log 2} = 2$.

By generalizing, we obtain:

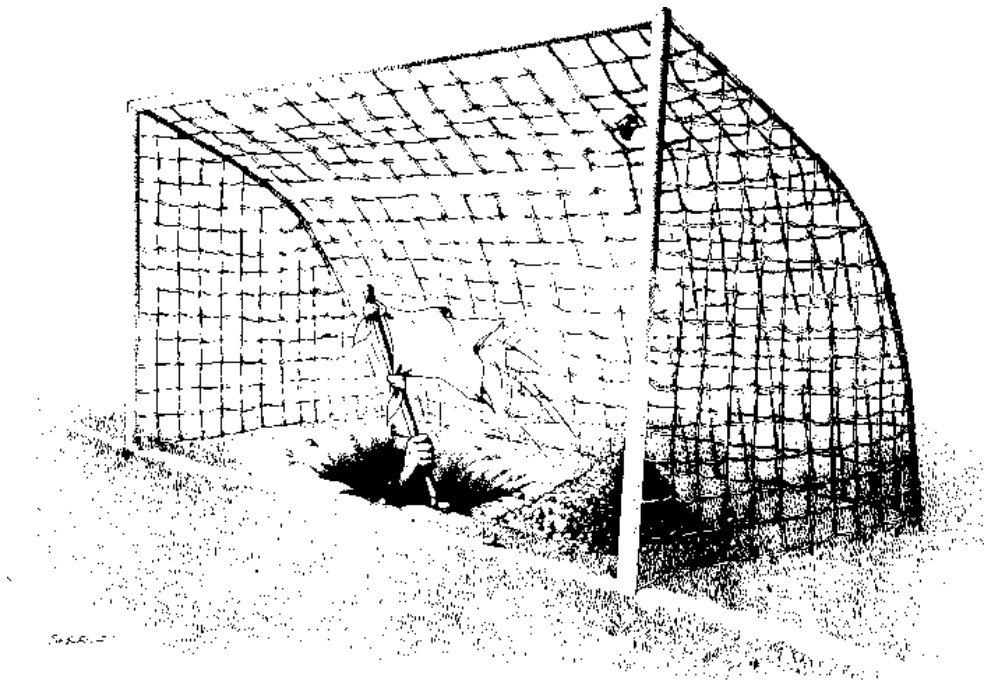
$$\boxed{P(i \text{ wins in leg } k) = P(i \text{ wins in leg } k-1) \sum_{j \in \mathcal{J}} P(j \text{ wins in leg } k-1) P_k(i,j)} \quad (1)$$

for $k = 2, \ldots, \log_2 n$. $\mathcal{J}$ is the set of all potential opponents of i for leg k.

Hence:

$$\boxed{P(i \text{ wins the tournament }) = P(i \text{ wins in leg } k = \log_2 n)} \quad (2)$$

We now must find probability models for determining $P_k(i,j)$.



2.4.3 Potential Opponents

By equation (1) we are forced to determinate the set of all potential opponents for a chosen team. For this reason we have programmed a function titled `potential.opponents`. The work of this function is illustrated by figure 6.

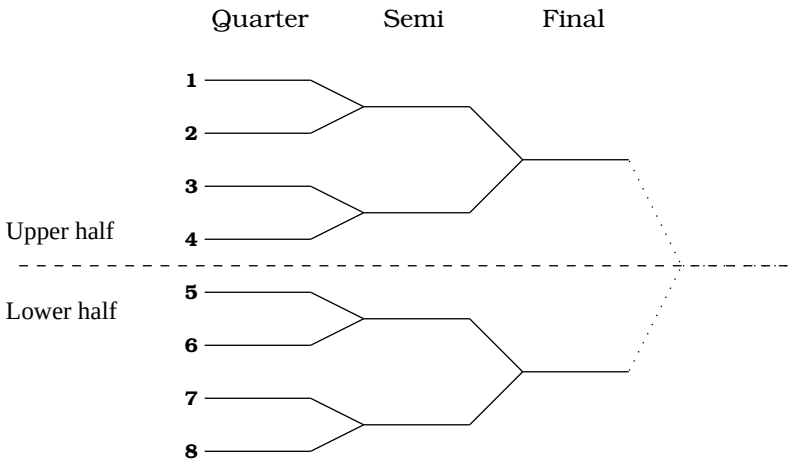
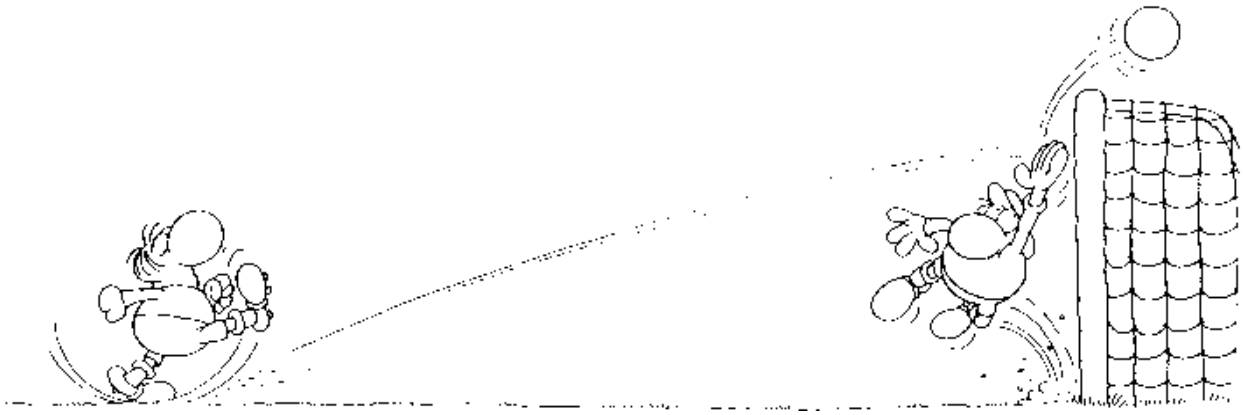


Figure 6 *Illustration of* `potential.opponents`.

In figure 6 there are eight teams participating in the tournament. If we want to determine all potential opponents for the final game of a team we only have to look if the team chosen is in the upper or in the lower half of the tournament tree. If the team is in the upper half the potential opponents will be all teams playing in the lower half, *i.e.* the following four teams: 5, 6, 7 and 8. In the same way it is possible to determine all potential opponents for the semi final: If the team is among the teams 3 and 4 the potential opponents are the teams playing in the upper quarter of the tournament tree, *i.e.* 1 and 2, and so on.



3 Statistical Modelisation

3.1 Logistic Regression Model

Schwertman *et al.* (1996) suggested in their work on probability models for the NCAA Regional Basketball Tournaments 11 different models for assigning probabilities of winning for each team in each individual game. These models were compared in three ways by using a chi-square statistic as a measure of the relative fit of the models. The chi-square values provided a measure of the relative accuracy of the various models. Hence the logistic model seemed to be the most satisfactory if the objective is to predict the winner of the tournament. That is the reason why we choose a logistic regression model. The model we assume is closely related to the Bradley-Terry model for paired comparisons (Bradley and Terry 1952). This relationship is shown in the appendix of this paper. Therefore the formula relating the seeding coefficients to winning probabilities is:

$$\boxed{P_k(i,j) = \frac{e^{\alpha + \beta(S_k(i) - S_k(j))}}{1 + e^{\alpha + \beta(S_k(i) - S_k(j))}}} \tag{3}$$

where $S_k(i)$ is the seeding coefficient of team i depending on leg k. Clearly we have $0 \leq P_k(i,j) \leq 1$ and $P_k(j,i) = 1 - P_k(i,j)$, $\forall k$ and $\forall i \neq j$.

These seeding coefficients represent the team strength.

3.2 Constancy of Team Strength

In the first time, we assume that the games are independent and that the seeding coefficients remain constant throughout the tournament. Hence $S_k(i) = S(i)$ and therefore $P_k(i,j) = P(i,j)$.

As each game has only two outcomes: A win or a loss, we are opposed to an independent Bernoulli trial. Let be Y_{ij} one Bernoulli trial, therefore:

$$Y_{ij} = \begin{cases} 1 & \text{if } i \text{ defeats } j \\ 0 & \text{else.} \end{cases}$$

It follows that $P(Y_{ij} = 1) = P(i,j)$. Hence the model (3) becomes:

$$E[Y_{ij}] = \frac{e^{\alpha + \beta(S(i) - S(j))}}{1 + e^{\alpha + \beta(S(i) - S(j))}}$$

where α and β are the parameters to estimate.

3.2.1 Model Chosen

The model is fit to a data set consisting of the game outcomes from season 1992/93 to 1995/6 in all three European Cups, *i.e.* the UEFA Cup, the Cup Winners Cup and the Champions Cup. It is important to notice that in the Champions Cup we took only the games of type knock-out, *i.e.* we did not take into consideration the league stage of the Champions Cup. In all we considered 442 game outcomes. In this regard it is important to say that even if two teams played twice against each other in a leg, we considered only the outcome of the leg, and not the outcomes of each game of the leg.

As response variable for the regression we took the following variable, let it be $\boldsymbol{y}$:

$$\boldsymbol{y} = \begin{bmatrix} y_{uc.9596} \\ y_{uc.9495} \\ y_{uc.9394} \\ y_{uc.9293} \\ y_{cwc.9596} \\ \dots \\ y_{cc.9293} \end{bmatrix}$$

where for example $y_{uc.9596}$ is a vector containing 1's and 0's. Let us have a look at two games of the UEFA Cup 1995/96:

	leg	team.a	country.a	team.b	country.b	s.a.1	s.b.1	s.a.2	s.b.2
20	Round1	Glenavon	Nil	WerderBremen	Ger	0	2	0	5
21	Round1	BayernMuenchen	Ger	LokomotivMoscow	Rus	0	1	5	0

The 20th element of $y_{uc.9596}$ would be 0 because WerderBremen (i.e. `team.b`) won the game, and the 21th element of $y_{uc.9596}$ would be 1 because BayernMuenchen (i.e. `team.a`) won the game.

By fitting the model (3) we obtained the following estimations:

Estimations	Value	Std. Error	t value
$\hat{\alpha}$	-0.1305	0.1039	-1.2559
$\hat{\beta}$	1.2610	0.1608	7.8420

It is a good fit. This statement is validated by figure 7. One notices that this regression rejects the necessity of the intercept α. Therefore (3) becomes:

$$P(i,j) = \frac{e^{\hat{\beta}(S(i)-S(j))}}{1+ e^{\hat{\beta}(S(i)-S(j))}} \tag{4}$$

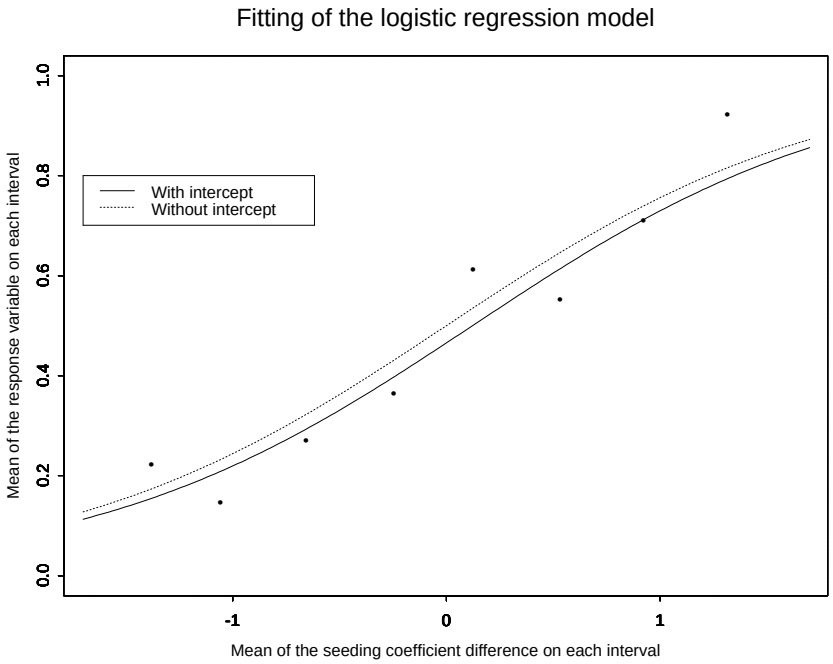
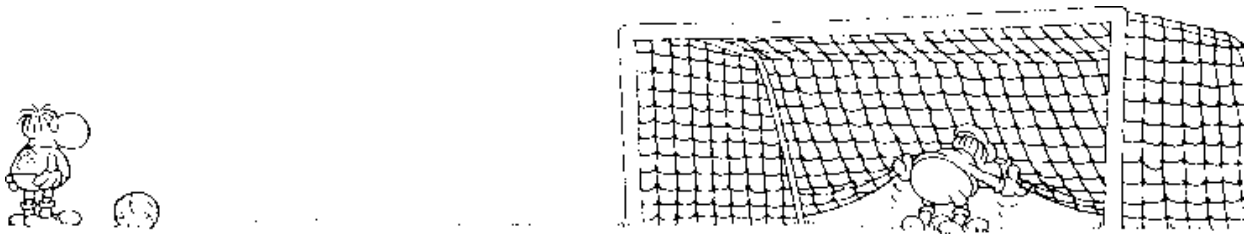


Figure 7 *Fitting of the logistic regression model. We noticed that the seeding coefficient differences lied between -1.8 and 1.8, therefore we divided this interval in 8 subintervals. In each of this subintervals we calculated the mean of the response variable and the mean of the seeding coefficient differences. One remarks that the curve without the intercept fits the observations best.*



3.2.2 The Program

With the help of equation (4) it is now possible to calculate the probability of each team winning against any other team (i.e. $P(i, j)$). In introducing these probabilities into equation (1) we are able to know the probability of each team winning any chosen leg and so, using equation (2), the tournament. For this reason we have programmed a function titled `prob.cte`. An issue of this function is shown by the following table who shows the probabilities for the Cup Winners Cup 1995/96:

	Round1	Round2	Quarter	Semi	Final
ParisSaintGermain	87.60	71.81	39.57	27.36	18.98
MoldeFK	12.40	4.97	0.95	0.24	0.06
Celtic	65.26	17.30	4.93	1.88	0.73
DinamoBatumi	34.74	5.93	1.13	0.28	0.07
Parma	89.01	79.22	49.19	35.38	25.51
KSTeuta(Durres)	10.99	5.50	1.12	0.28	0.07
HalmstadsBK	50.00	7.64	1.56	0.39	0.10
LokomotivSofia	50.00	7.64	1.56	0.39	0.10
RCDeportivoLaCoruna	68.37	32.21	15.36	4.69	2.21
Apoel(Lefkosia)	31.63	9.49	2.88	0.51	0.15
Trabzonspor	74.09	47.98	27.31	10.53	5.93
ZalgirisVilnius	25.91	10.33	3.36	0.65	0.21
RealZaragoza	76.19	40.65	21.63	7.39	3.81
InterZTSBratislava	23.81	6.63	1.80	0.27	0.07
ClubBruggeKV	78.28	46.19	25.88	9.50	5.16
ShakhtarDonetsk	21.72	6.53	1.77	0.27	0.07
SKRapid(Wien)	55.24	16.78	6.17	1.98	0.39
PetrolulPloiesti	44.76	11.76	3.79	1.06	0.18
SportingCP(Lisbon)	71.32	54.96	34.67	19.76	8.18
MaccabiHaifa	28.68	16.50	6.86	2.50	0.58
DinamoMoscow	75.09	48.26	27.06	14.01	5.09
AraratErevan	24.91	9.49	2.98	0.83	0.14
SKHradecKralove	32.40	10.09	3.17	0.89	0.15
FCKobenhavn	67.60	32.15	15.29	6.63	1.90
Feyenoord	82.37	66.36	47.22	30.42	14.08
DAGLiepaya	17.63	8.31	2.99	0.88	0.15
Everton	44.76	10.45	3.76	1.10	0.18
KR(Reykjavik)	55.24	14.89	6.06	2.04	0.40
BorussiaMonchengladbach	50.00	15.06	4.18	1.23	0.21
SileksKratovo	50.00	15.06	4.18	1.23	0.21
AEK(Athens)	40.91	26.70	10.75	4.64	1.24
FCSion	59.09	43.18	20.87	10.82	3.69

Let us take for example Paris Saint Germain (Fra): Their chance to win in round 1 against Molde FK (Nor) is 87.6%, to win in round 2 against Celtic (Sco) or Dinamo Batumi (Geo) their chance is 71.8%, and to win the final, *i.e.* the tournament, their chance is about 19%. The real outcomes are shown by figure 8.

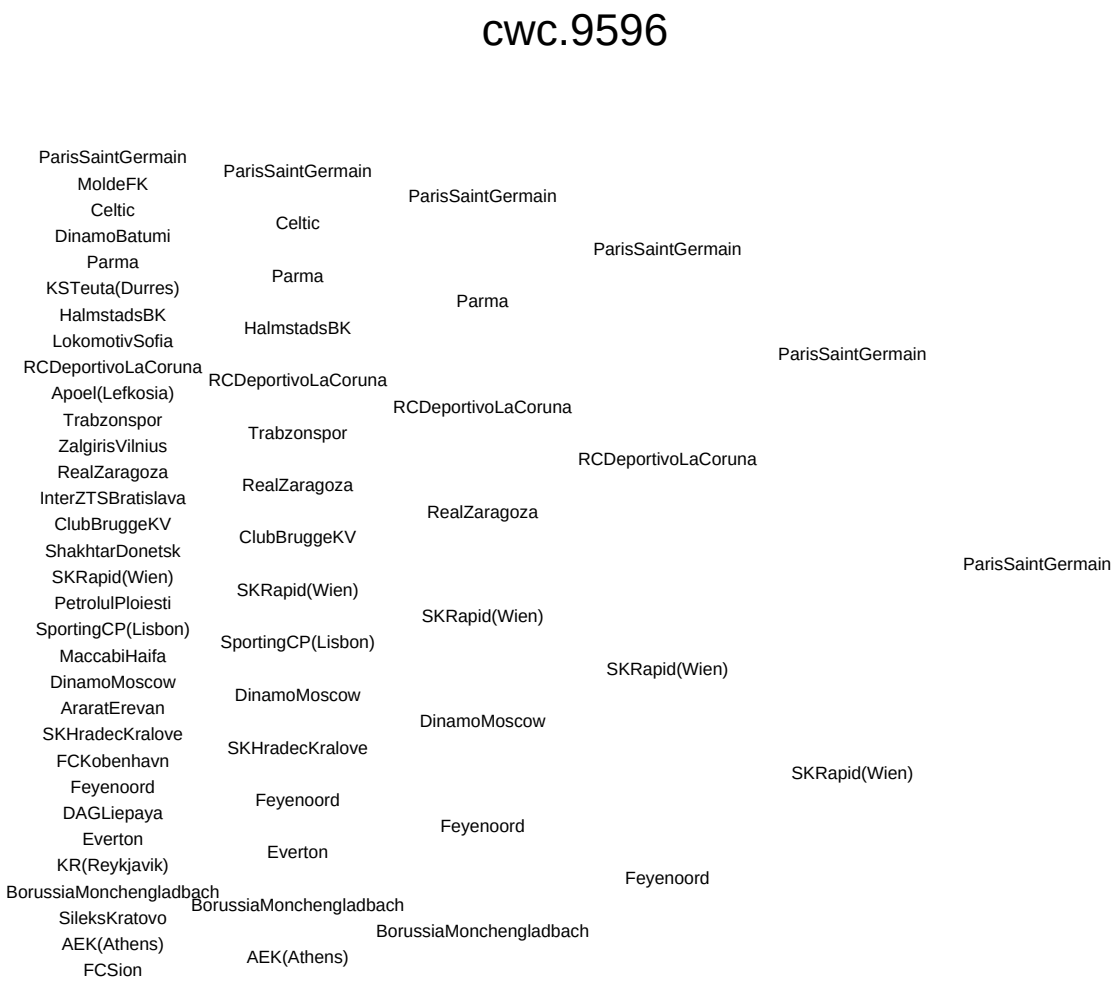


Figure 8 *Tournament tree of the Cup Winners Cup 1995/96.*

One remarks the surprising performance of SK Rapid (Wien) (Aut): Their chance to reach the quarter final was 16.7%, but they even reached the semi final where their chance was only 6.17%. Knowing this surprising fact the probabilities of the other teams remaining in the tournament will change. Therefore it is not only interesting to know the probability of each team winning in a chosen leg at the beginning of the tournament, but also predicting the probability of each team winning knowing the remaining teams. That is why we refined the program `prob.cte`. For example in the Cup Winners Cup 1995/96 the following teams remain after the quarter finals : Paris Saint Germain (Fra), RC Deportivo La Coruna (Esp) , SK Rapid (Wien) (Aut) and Feyenoord (Ned). We are able to calculate the new probabilities knowing this fact. It is important to notice that these new probabilities are calculated without care of the outcomes of the previous rounds.

A more advanced method will be discussed in section 3.3. The result of the function is shown by the following table:

	Semi	Final
ParisSaintGermain	72.59	47.47
RCDeportivoLaCoruna	27.41	11.80
SKRapid(Wien)	20.90	4.07
Feyenoord	79.10	36.66

We notice that this time the chance of Paris Saint Germain to win the tournament is about 50% - and they won it. And the chance of SK Rapid (Wien) (Aut) to defeat their opponent, Feyenoord (Ned), in the semi final is about 21%. At the beginning of the tournament the chance of SK Rapid (Wien) (Aut) to win the semi final was only about 2%. So this approach seems to be more realistic.

3.2.3 Number of Correct Predictions

Let us consider the following table:

Team	Probability of winning
A	X
B	1 - X

where $X = P(A, B)$. If $X > \frac{1}{2}$ the probable winner will be A and if the real winner is also A we will mark a point, if $X < \frac{1}{2}$ ($1 - X > \frac{1}{2}$) and the real winner is B we will mark a point as well and if $X = 1 - X = \frac{1}{2}$ we will mark half a point (this semi point is justified by the fact that the expectation of a point is $0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$).

The following example issues from the Cup Winners Cup 1992/93 and shows the probabilities of winning the quarter final for the teams remaining in the tournament.

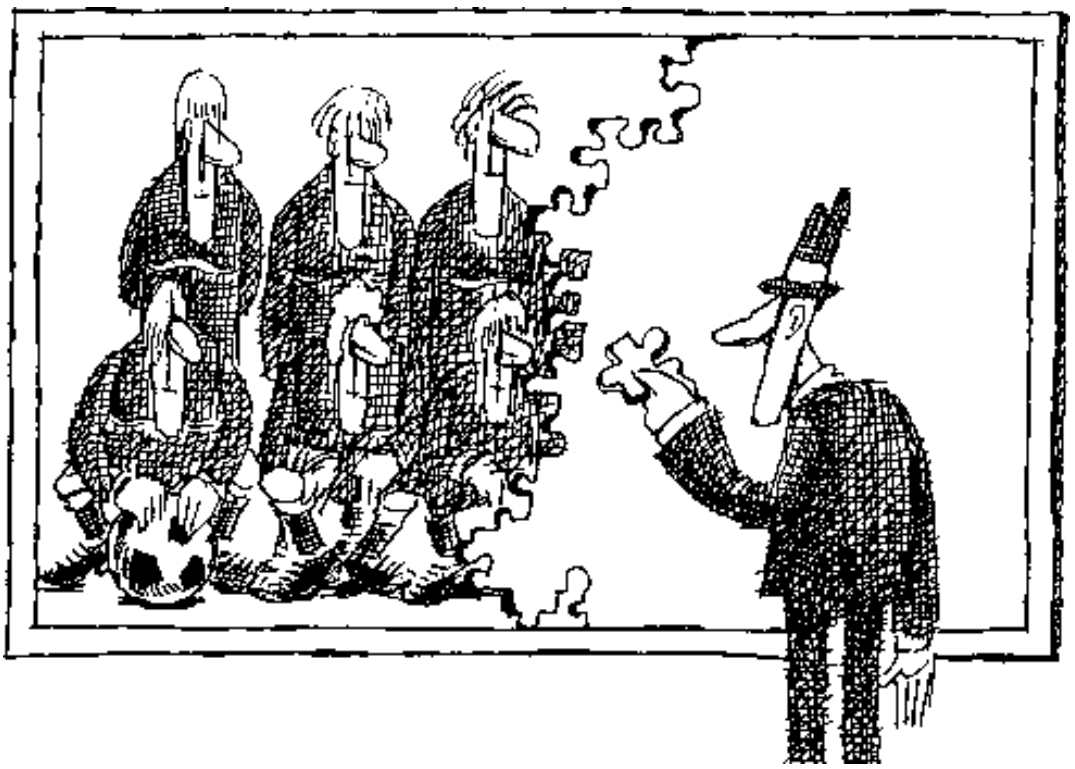
	Quarter
Parma	43.47
SpartaPraha	56.53
AtleticoMadrid	74.09
Olympiakos(Pireus)	25.91
RoyalAntwerp	31.26
SteauaBucharest	68.74
SpartakMoscow	60.36
Feyenoord	39.64

As we know that the real winners were Parma, Atletico Madrid, Royal Antwerp and Spartak Moscow, it is possible to calculate the number of correct predictions. That would be 2 because Atletico Madrid and Spartak Moscow are probable winners as well.

In applying this procedure to the entire tournament we will get a list of points whose sum represents the number of correct predictions. For this reason we have programmed a function titled `nb.pronos.ok`. In using this function for various data sets we obtained the following table:

YEAR	OK	GAMES	%OK
cwc.9293	21.5	31	69.35
cwc.9394	20	31	64.52
cwc.9495	20.5	31	66.13
cwc.9596	19	31	61.29
uc.9293	45	63	71.43
uc.9394	33.5	63	53.17
uc.9495	45.5	63	72.22
uc.9596	37.5	63	59.52

`OK` stands for the number of correct predictions, `GAMES` represents the number of games per tournament and `%OK` the percentage of games predicted correctly.



3.3 Variation of Team Strength

This time we assume that the team strength (the seeding coefficients) do not remain constant throughout the tournament - they are depending on the leg.

3.3.1 Model Chosen

The principal idea is to calculate for each leg a new coefficient for each team. The initial coefficients that are used to predict the issues of the first leg are the coefficients calculated in section 2.3.2. For the second round we have to calculate new coefficients. To do so we introduce two different methods (see sections 3.3.2 and 3.3.3). Once we have calculated these new coefficients we center and reduce them in order to obtain coefficients of equal mean and equal standard deviation as the initial ones.

Justification: Let $x_1, \ldots, x_n$ be the initial coefficients of mean $\bar{x}$ and standard deviation s_x and let $y_1, \ldots, y_n$ be the new coefficients of mean $\bar{y}$ and standard deviation s_y. We define the new coefficients as

$$y_t^* = \bar{x} + s_x \left(\frac{y_t - \bar{y}}{s_y} \right), \ t = 1, \ldots, n.$$

Hence $\bar{y^*} = \bar{x}$ and $s_{y^*} = s_x$. An example of this calculation is given by the following table:

Team	x_t	y_t	y_t^*
A	1.4	2.1	1.382
B	1.2	1.9	1.241
C	1	1.5	0.958
D	0.8	1.3	0.817
Mean	1.1	1.7	1.1
Standard Deviation	0.258	0.365	0.258

Therefore we are allowed to use the estimations made in section 3.2.1. Thus our model is given by equation (3):

$$\boxed{P_k(i,j) = \frac{e^{\hat{\beta}(S_k(i) - S_k(j))}}{1 + e^{\hat{\beta}(S_k(i) - S_k(j))}}} \tag{5}$$

where $k = 1, \ldots, \log_2 n$ and $\hat{\beta} = 1.2610$.

Remark on the effect of the centering and the reduction of the coefficients: If a strong team wins against a weak team their coefficient will decrease, but if a weak team defeats a strong team their coefficient will increase. That seems to be logical.

3.3.2 Using Probable Opponents

To illustrate this first method, let us have a look at figure 9.

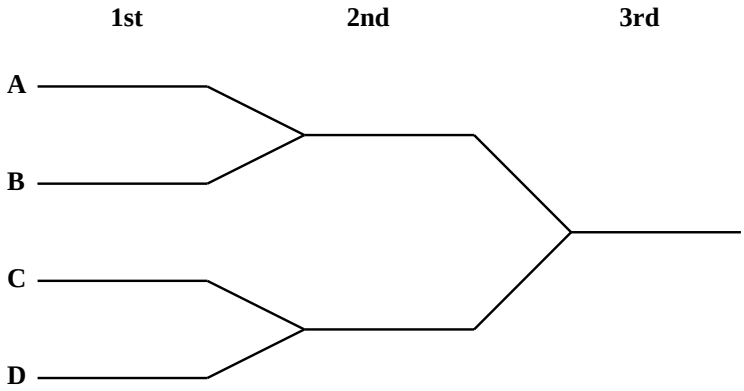


Figure 9 *Example of the upper part of the tournament tree.*

The initial coefficients are used to calculate the probability of each team winning against their potential opponent in the first round (for the first round only one potential opponent is possible). The new coefficients for the second round who will be used to calculate the probabilities of winning in the second round are defined in the following way, for example the new coefficient for a chosen team i would be:

$$C_{new,i} = C_{old,i}\left(1 + \frac{C_{old,O(i)} + \gamma}{C_{old,i} + \gamma}\right) \tag{6}$$

where γ is a constant ≥ 0 and $C_{old,O(i)}$ is the coefficient of the probable opponent $O(i)$ of team i in the previous round (if $i = A$ it would be B). Later on (see section 3.3.4) we will choose γ in order to maximize the number of correct predictions. It is important to notice that $C_{new,i}$ is $S_k(i)$ and $C_{old,i}$ is $S_{k-1}(i)$, $k = 1, \ldots, \log_2 n$.

In realizing that some initial coefficients are equal to 0 (see section 2.3.2) we allow us to add 0.1 to all initial coefficients. Let us consider equation (6): If we will not do this the coefficients would remain 0 throughout the tournament. In adding 0.1 the expectation of the new initial coefficients increases and the variance remains the same. But as our model (5) is depending on a difference of coefficients this will not change our estimations.

In reusing equation (6) we are able to calculate the coefficients for the third round. The only problem consists in the knowledge of the probable opponent $O(i)$: Let i be A, so for the third round there are two potential opponents for A: C and D. But we have calculated $P_2(C, D)$ yet, so if this probability is $> \frac{1}{2}$ the probable opponent for A would be C, if not the probable opponent would be D (if this probability is equal to $\frac{1}{2}$ it does not matter which team we will take for further considerations because both teams have the same strength).

In generalizing this example we only have to keep in mind that if a team was not a

probable opponent for a previous round it will not be one for the next round. Therefore we have programmed a function titled `prob.non.cte`. An outcome of this function is shown by the following table where the Cup Winners Cup of the year 1993/94 has been considered.

	Round1	Round2	Quarter	Semi	Final
Arsenal	57.82	22.70	8.95	2.95	0.89
OB(Odense)	42.18	13.21	3.89	0.84	0.14
StandardCL(Liege)	62.84	42.79	21.46	9.19	3.61
CardiffCity	37.16	21.30	8.69	2.98	0.94
Torino	75.09	50.36	32.17	15.04	6.40
LillestromSK	24.91	10.71	4.54	1.29	0.33
Aberdeen	43.47	15.51	7.46	2.50	0.77
Valur(Reykjavik)	56.53	23.42	12.85	5.05	1.83
ParisSaintGermain	69.42	41.50	14.14	7.92	3.11
Apoel(Nicosia)	30.58	12.21	2.25	0.69	0.12
UniversitateaCraiova	65.26	33.72	10.62	5.63	2.09
HB(Torshavn)	34.74	12.57	2.29	0.70	0.12
RealMadrid	87.37	58.21	43.54	29.06	13.78
FCLugano	12.63	3.40	1.30	0.44	0.09
FCTirolInnsbruck	46.50	17.26	11.42	6.83	2.88
Ferencvarosi	53.50	21.13	14.43	8.88	3.86
Parma	81.64	49.27	23.79	14.19	8.69
DegerforsIF	18.36	5.42	1.07	0.27	0.08
MaccabiHaifa	22.08	5.63	1.10	0.28	0.08
TorpedoMoscow	77.92	39.67	18.19	10.50	6.28
Ajax	82.74	64.41	40.11	25.35	16.15
HajdukSplit	17.26	9.28	3.84	1.79	0.92
Besiktas	72.03	22.03	10.90	5.84	3.32
Kosice.1.FC	27.97	4.28	1.00	0.25	0.07
SLBenfica	70.52	52.49	38.36	18.76	11.50
GKSKatowice	29.48	18.25	11.62	4.75	2.62
CSKA(Sofia)	72.03	24.48	15.60	6.38	3.52
FCBalzers	27.97	4.77	1.73	0.31	0.08
BayerLeverkusen	56.96	20.65	5.62	1.50	0.61
FCBobyBrno	43.04	12.76	2.52	0.41	0.09
Panathinaikos	76.25	55.45	22.27	9.01	4.94
Shelbourne	23.75	11.13	2.28	0.40	0.10

The real outcomes are shown by figure 10.

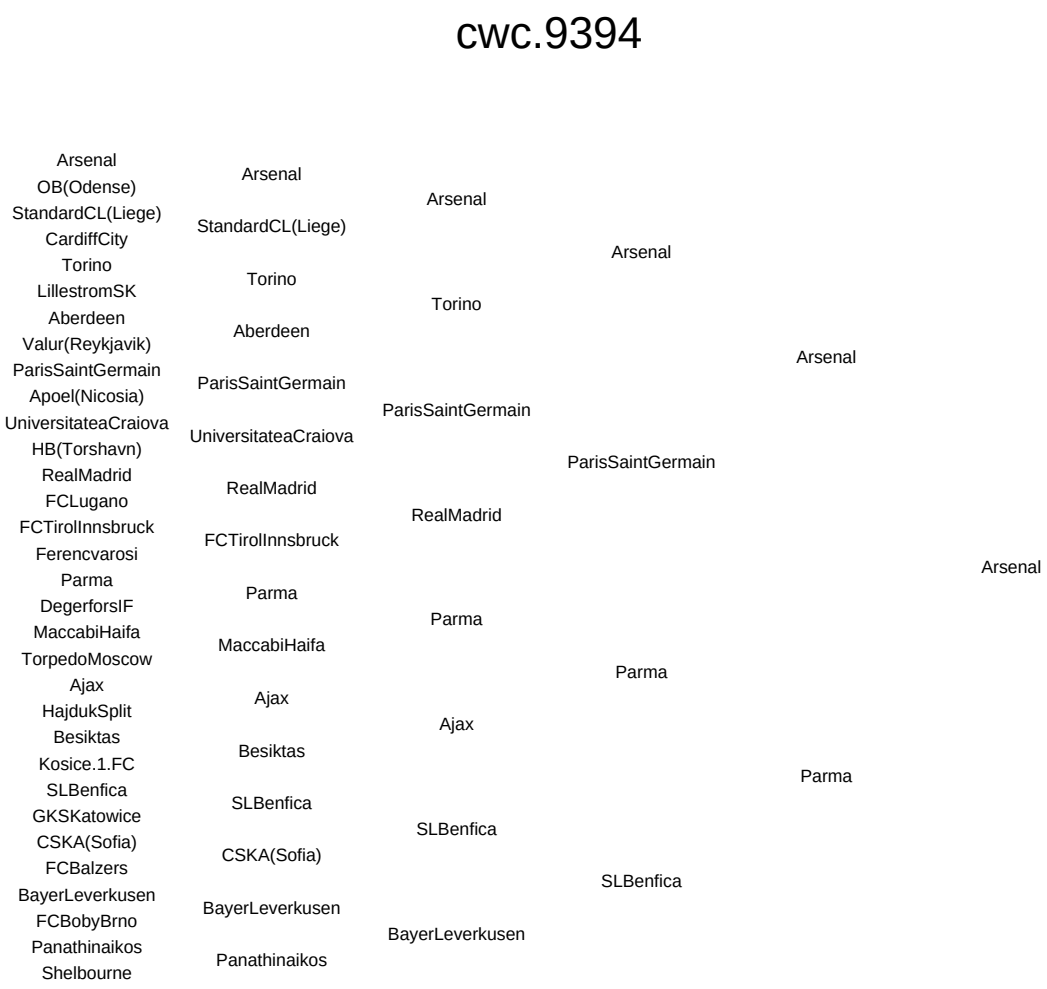


Figure 10 *Tournament tree of the Cup Winners Cup 1993/94.*

3.3.3 Using All Potential Opponents

The second method takes all potential opponents into consideration to calculate the new coefficients. This method seems to be more natural than the first one where we considered only the probable opponents. The new coefficient for a chosen team i is defined as:

$$C_{new,i} = C_{old,i} \sum_{j \in \mathcal{O}} (1 + r_j) P(j \text{ wins in the previous round}) \tag{7}$$

where γ is a constant ≥ 0, $r_j = \frac{C_{old,j}+\gamma}{C_{old,i}+\gamma}$ and $\mathcal{O}$ the set of all potential opponents for the previous round for the team i. One remarks that this new coefficient is a sort of weighted sum of old coefficients.

To illustrate this method let us have a look at the example shown by figure (9). For the second round the new coefficient is exactly the same as the one calculated for the second round in the previous section because there is only one potential opponent for the first round for each team. And for the third round the new coefficient of A would be

$$C_{new,A} = C_{old,A} \sum_{j \in \{C,D\}} (1 + r_j) P(j \text{ wins in the first round})$$

We have to notice that the procedure of calculating these new coefficients is the same as in the previous section. Hence we have reprogrammed the function `prob.non.cte`. An issue of this function is shown by the following table where the UEFA Cup of the year 1994/95 has been considered.

	Round1	Round2	Round3	Quarter	Semi	Final
Parma	73.00	47.03	34.40	19.37	12.02	6.50
Vitesse(Arnhem)	27.00	12.06	7.22	3.17	1.69	0.80
AIK(Stockholm)	57.82	25.47	16.84	8.37	4.82	2.45
SlaviaPraha	42.18	15.45	9.34	4.15	2.23	1.07
AthleticBilbao	50.00	17.92	3.66	0.75	0.21	0.05
AnorthosisofAmmochostos	50.00	17.92	3.66	0.75	0.21	0.05
NewcastleUnited	27.34	13.41	2.59	0.49	0.12	0.02
RoyalAntwerp	72.66	50.75	22.29	10.39	5.75	2.82
OB(Odense)	42.18	17.53	4.95	1.87	0.85	0.35
Linfield	57.82	28.68	10.59	4.89	2.59	1.22
Kaiserslautern	53.06	29.23	11.30	5.39	2.91	1.40
IA(Akranes)	46.94	24.56	8.79	3.97	2.07	0.96
RealMadrid	63.33	40.23	27.80	16.57	10.39	5.66
SportingCP(Lisbon)	36.67	19.79	12.19	6.58	3.86	1.99
DinamoMoscow	77.92	36.41	23.36	13.05	7.84	4.13
RFCSeraing	22.08	3.58	1.02	0.25	0.07	0.02
BayerLeverkusen	53.15	34.87	18.71	10.21	4.60	2.17
PSV(Eindhoven)	46.85	29.65	15.20	8.01	3.47	1.59
KispestHonved	56.53	21.99	9.02	3.95	1.37	0.52
FCTwente(Enschede)	43.47	13.49	4.17	1.34	0.28	0.06

	Round1	Round2	Round3	Quarter	Semi	Final
GKSKatowice	65.26	30.80	16.44	8.55	3.65	1.65
Aris(Thessaloniki)	34.74	10.37	3.46	1.11	0.23	0.05
GirondinsBordeaux	60.36	37.54	22.15	12.50	5.84	2.83
LillestromSK	39.64	21.29	10.85	5.42	2.21	0.97
FCNantes	57.82	33.38	11.83	4.94	1.68	0.63
RotorVolgograd	42.18	20.99	5.32	1.54	0.29	0.05
TekstilschikKamyshin	50.00	22.82	5.85	1.72	0.33	0.06
BekescsabaiElore	50.00	22.82	5.85	1.72	0.33	0.06
FCSion	55.24	23.84	16.49	8.70	3.86	1.80
Apollon(Limassol)	44.76	16.96	11.06	5.42	2.22	0.97
OlympiqueMarseille	41.18	22.70	16.11	8.77	4.02	1.92
Olympiakos(Pireus)	58.82	36.51	27.49	16.10	7.95	4.01
Juventus	61.61	44.04	30.39	16.95	8.86	4.72
CSKA(Sofia)	38.39	24.48	15.33	7.72	3.69	1.85
CSMaritimo	50.00	15.74	7.90	3.11	1.18	0.50
FCAarau	50.00	15.74	7.90	3.11	1.18	0.50
AdmiraWacker	60.36	32.52	13.86	6.01	2.50	1.13
GornikZabrze	39.64	17.08	5.14	1.54	0.42	0.13
ASCannes	40.91	18.38	5.86	1.89	0.56	0.19
Fenerbahce	59.09	32.02	13.62	5.89	2.45	1.11
EintrachtFrankfurt	82.49	63.83	36.47	21.52	11.13	5.89
SCTOlimpija(Ljubljana)	17.51	8.46	2.08	0.55	0.10	0.02
RapidBucharest	50.00	13.86	3.68	1.08	0.25	0.06
RSCCharleroi	50.00	13.86	3.68	1.08	0.25	0.06
Napoli	57.82	22.87	11.14	5.43	2.34	1.09
SkontoRiga	42.18	13.21	5.20	2.06	0.71	0.28
Boavista	82.87	57.78	35.94	21.56	11.32	6.05
MyPa(Anjalankoski)	17.13	6.15	1.82	0.49	0.10	0.02
BorussiaDortmund	79.10	50.16	30.32	17.40	9.95	5.44
Motherwell	20.90	7.04	2.25	0.74	0.26	0.10
SlovanBratislava	50.00	21.40	10.45	5.07	2.56	1.27
FCKobenhavn	50.00	21.40	10.45	5.07	2.56	1.27
RCDeportivoLaCoruna	37.16	16.69	6.47	2.78	1.27	0.59
RosenborgBK	62.84	35.43	17.91	9.43	5.06	2.64
FCTirolInnsbruck	73.76	40.35	20.50	10.83	5.83	3.04
DinamoTblisi	26.24	7.52	1.66	0.39	0.09	0.02
Lazio	65.26	45.23	19.68	9.32	4.74	2.38
DinamoMinsk	34.74	19.81	5.71	1.89	0.72	0.29
TrelleborgsFF	50.00	17.48	3.60	0.79	0.19	0.05
BlackburnRovers	50.00	17.48	3.60	0.79	0.19	0.05
Trabzonspor	63.25	37.75	26.93	15.07	8.63	4.73
DinamoBucharest	36.75	17.96	11.51	5.68	2.98	1.53
AstonVilla	34.74	12.05	6.90	2.98	1.42	0.67
Internazionale	65.26	32.25	22.08	11.77	6.52	3.48

The real outcomes are shown by figure 11.

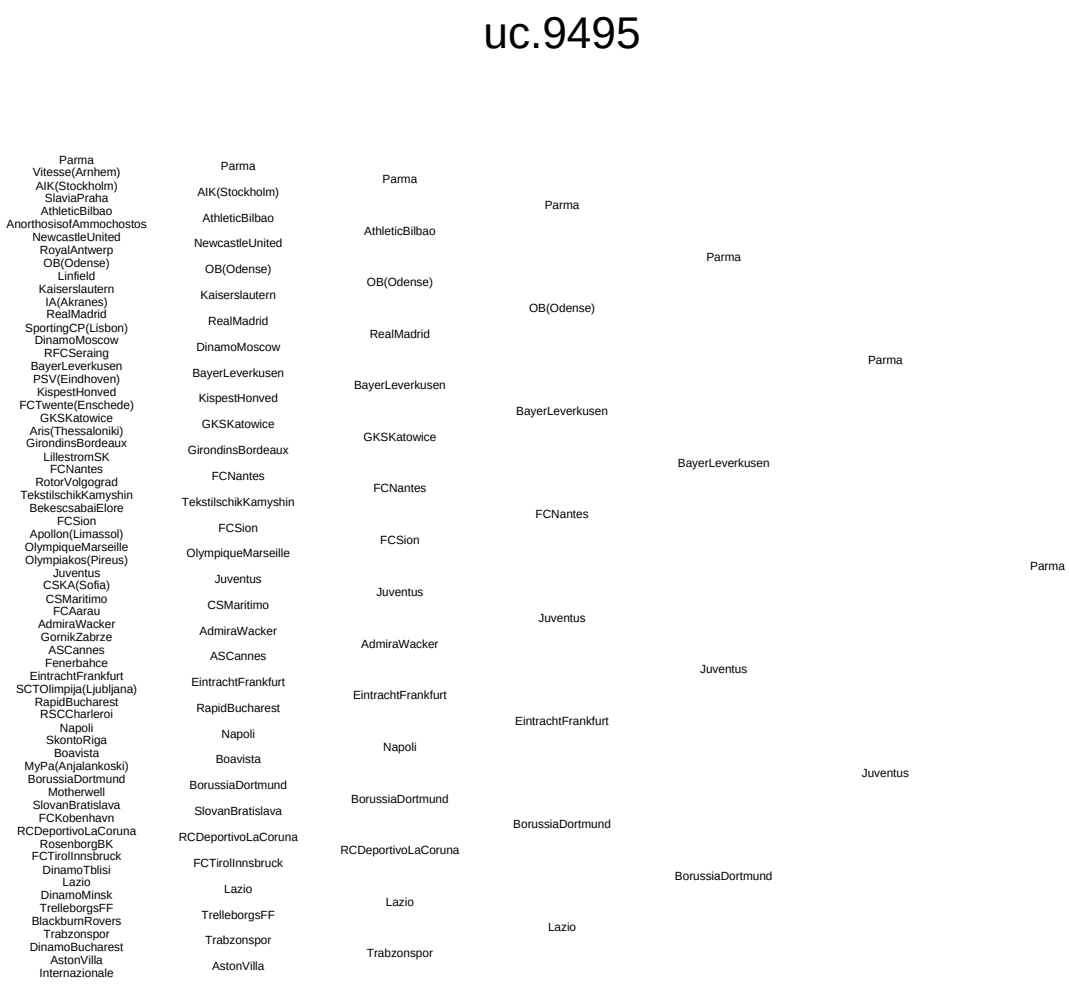


Figure 11 *Tournament tree of the UEFA Cup 1994/95.*

3.3.4 Number of Correct Predictions

In regarding the two methods presented in the two preceding sections our goal in this section is to choose γ in order to maximize the number of correct predictions. The procedure of obtaining this number has been discussed in section 3.2.3. The calculation of the new coefficients for a certain round is based on the real outcomes of the previous round so there is only one potential opponent (the real one). The new coefficients will be calculated in using the following equation:

$$C_{new,i} = C_{old,i}\left(1 + \frac{C_{old,R(i)} + \gamma}{C_{old,i} + \gamma}\right)$$

where $R(i)$ is the real opponent of team i for the previous round. If a team is not in a particular round, *i.e.* it lost in one of the previous rounds, their coefficient has no reason to be changed. This remark is important because we will center and reduce the new coefficients in using all initial coefficients.

With $\gamma = 1$ we have obtained the following table:

```
    YEAR   OK GAMES    %OK
-------------------------
cwc.9293 21.5    31 69.35
cwc.9394 21      31 67.74
cwc.9495 20.5    31 66.13
cwc.9596 20      31 64.52
uc.9293  45      63 71.43
uc.9394  34      63 53.97
uc.9495  44      63 69.84
uc.9596  37      63 58.73
```

We have executed the function `nb.pronos.ok` for several values of γ, but how to know which γ is optimal? For this reason we have to define a criterion of comparison.

A general outcome of the function looks like this:

YEAR	OK	GAMES	%OK
1	o_1	g_1	p_1
2	o_2	g_2	p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
8	o_8	g_8	p_8

Let W be $\sum_{i=1}^{8} w_i p_i$, the weighted sum of the percentages, where w_i are the weights fulfilling $\sum_{i=1}^{8} w_i = 1$ and $w_i \propto g_i$. Hence $w_i = g_i/g$ where $g = \sum_{i=1}^{8} g_i$. By construction of the table $p_i = o_i/g_i$ therefore $W = (\sum_{i=1}^{8} o_i)/g$.

So let us define our criterion of comparison:

Criterion: γ is optimal if $W = \frac{1}{g}\sum_{i=1}^{8} o_i$ is maximal.

In calculating W for several values of γ we obtain the following table:

γ	0	0.2	0.4	0.6	0.8	1	2	3	4	5	15
W	58.51	62.5	63.29	64.36	64.36	64.62	64.62	64.62	64.36	64.09	64.09

One remarks that in taking for example $\gamma = 1$ we obtain the biggest W, *i.e.* the highest weighted sum of games predicted correctly. Therefore let 1 be the optimal γ to choose.

We also tried another model to calculate the new coefficients. This model is given by the following equation:

$$C_{new,i} = C_{old,i}\left(1 + \lambda\frac{C_{old,R(i)}}{C_{old,i}}\right)$$

We obtain the following table:

λ	0	0.2	0.4	0.6	0.8	1	2	3	4
W	64.5	64.5	64.22	63.56	63.56	58.51	58.9	59.3	60.1

One remarks that this time W does not exceed 64.5% as before. That is the reason why we do not take into further considerations this model.

In this section we chose γ in order to maximize the number of correct predictions. Our next goal will be to compare the methods presented in this paper (see section 4).

3.3.5 Evaluation of the Coefficients

General Remark

The results of this section have been achieved in considering several years. So let us have a look for example at the UEFA Cup 1995/96 in using the method that considers all potential opponents for a team. The other method furnishes the same results. In figure 12 the evaluation of the seeding coefficients throughout the whole tournament for four chosen teams is shown. Milan was the best seeded team at the beginning of the tournament, Olympiakos and Spartak Vladikavkaz were medium seeded teams and FC Lugano was among the worst seeded teams.

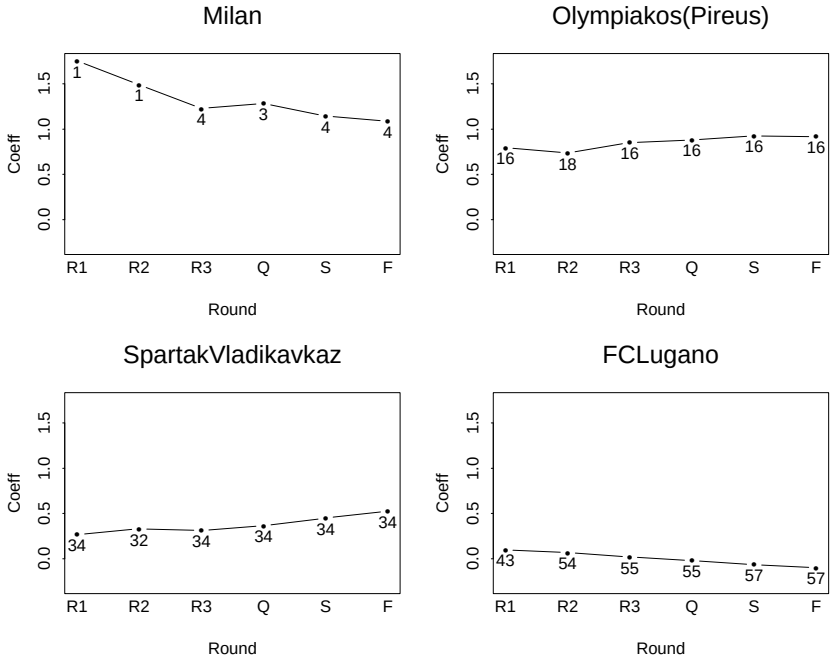


Figure 12 *Evaluation of the seeding coefficients for four chosen teams during the UEFA Cup 1995/96. The number represents the seeding position of the team.*

In general we noticed that the best teams, as Milan, will see their coefficient decrease, but they will still remain among the strongest teams. The medium teams, as Olympiakos and Spartak Vladikavkaz, will see their coefficients increase, and the worst teams will stay among the worst teams. The reason of this fact is that we centered and reduced the coefficients otherwise the better teams would always get higher coefficients and so become unbeatable.

In the appendix you will find the evaluation of the coefficients for all teams that have participated in the UEFA Cup 1995/96, in considering both methods.

Particular Examples

The two particular examples shown by figure 13 illustrate two things: First of all we remark that the two methods used are different, and in a second time we notice that in using the method considering only the probable opponent we have some strange evaluations: The coefficient of Milan decreases too much in the quarter final, they are losing six seed positions in falling from position 4 to position 10. This fact does not come up with the other method. The same can be said for Girondins Bordeaux in considering the coefficient for round 2. In recapitulating, the method using all potential opponents seems to be more natural.

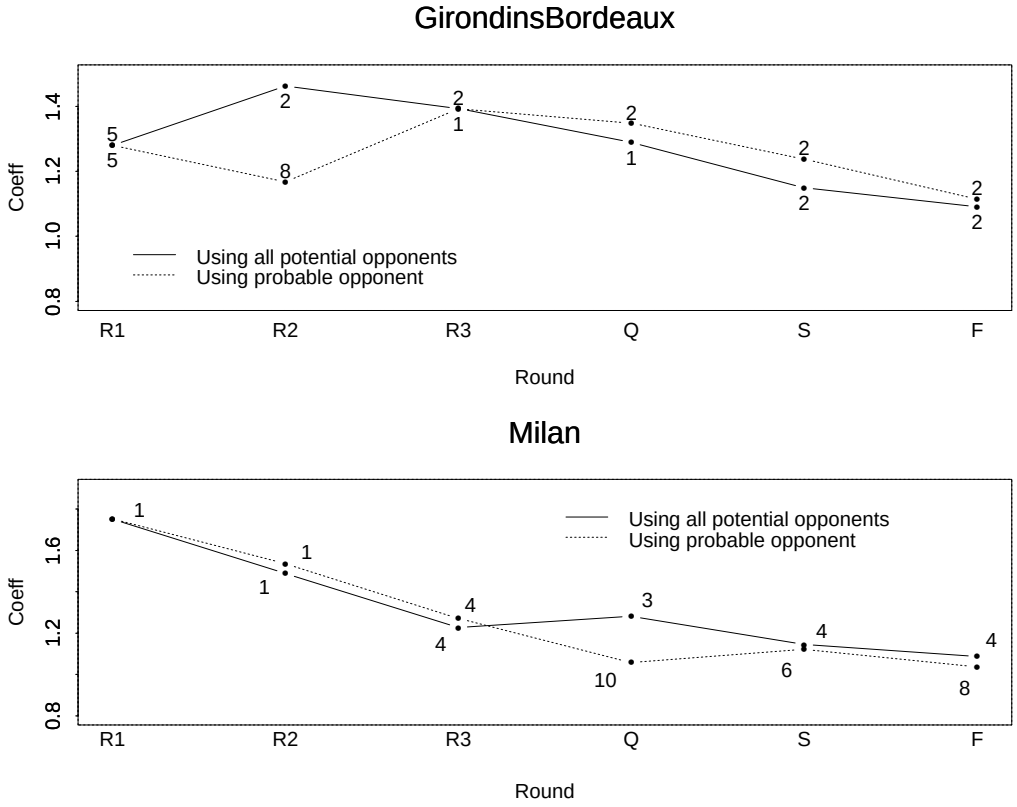
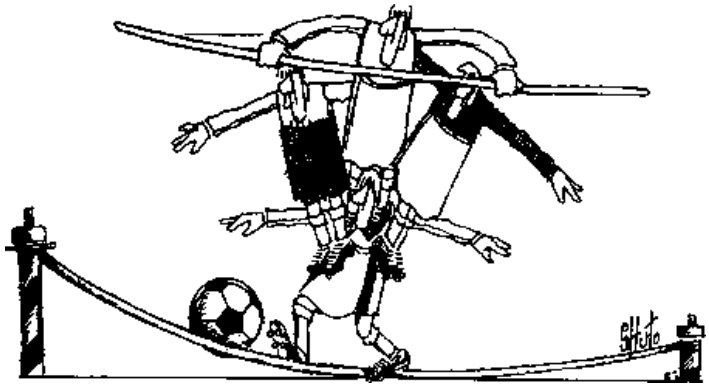


Figure 13 *Evaluation of the seeding coefficients for Milan and Girondins Bordeaux during the UEFA Cup 1995/96.*

4 Comparison of the Methods



In this section we want to compare the methods proposed in the previous sections:

M1 The first method calculates the probabilities in using the constancy of the team strength.

M2 The second one assumes the variation of the team strength in using only the probable opponents for a team.

M3 The third one assumes the variation of the team strength in using all potential opponents for a team.

To compare the three methods **M1**, **M2** and **M3** we will take as example the quarter final of the UEFA Cup 1995/96. The table 1 shows the eight quarter finalists, the probability of each quarter finalist to reach the quarter final in using one of the three methods and the rank of this probability. This rank is defined as the order of the sorted probabilities of the team and of all his potential opponents for the previous rounds. Here is an example how this rank is calculated: Let us have a look at the following table who shows the probabilities to reach the quarter final for Bayern Muenchen and all his potential opponents for the previous rounds:

	Round3
BayernMuenchen	27.87
LokomotivMoscow	2.10
RaithRovers	2.80
IA(Akranes)	9.86
SLBenfica	46.03
LierseSK	1.40
RodaJC(Kerkrade)	2.58
SCTOlimpija(Ljubljana)	7.35

The rank of Bayern Muenchen would be 2 because SL Benfica had bigger chance to win the third round (46.03% > 27.87%), *i.e.* to reach the quarter final.

	M1		M2		M3	
Team	Prob.	Rank	Prob.	Rank	Prob.	Rank
Bayern Muenchen	23.99	2	26.33	2	27.87	2
Nottingham Forest	3.62	3	3.98	4	4.43	3
FC Barcelona	43.70	1	39.15	1	38.13	1
PSV (Eindhoven)	13.80	4	13.20	4	14.70	4
Girondins Bordeaux	29.18	1	25.04	2	29.21	1
Milan	67.68	1	60.62	1	56.71	1
Slavia Praha	22.88	2	25.06	3	25.17	2
Roma	10.19	3	10.16	4	11.40	3

Table 1 *This table shows the eight quarter finalists, the probability of each one to reach the quarter final in using one of the three methods and the rank of this probability (UEFA Cup 1995/96).*

For a balanced tournament all teams would have 12.5% chance to reach the quarter final because there are in all 7 potential opponents for each team in the previous rounds and in considering the team as well we have 8 teams sharing 100%. For the teams with probability smaller than 12.5% we remark that **M3** gives more chance to reach the quarter final than **M2**. One notices as well that Nottingham Forest, Girondins Bordeaux, Slavia Praha and Roma had more chance to reach the quarter final in using **M3** than in using **M2** because their ranks are smaller in using **M3**. Hence **M3** seems to be better than **M2**. On the other hand **M1** and **M3** deliver equal ranks.
Moreover, in comparing the number of correct predictions made in using one of the three methods we obtained the following table:

	% games predicted correctly	
Year	**M1**	**M2** or **M3**
cwc.9293	69.35	69.35
cwc.9394	64.52	67.74
cwc.9495	66.13	66.13
cwc.9596	61.29	64.52
uc.9293	71.43	71.43
uc.9394	53.17	53.97
uc.9495	72.22	69.84
uc.9596	59.52	58.73

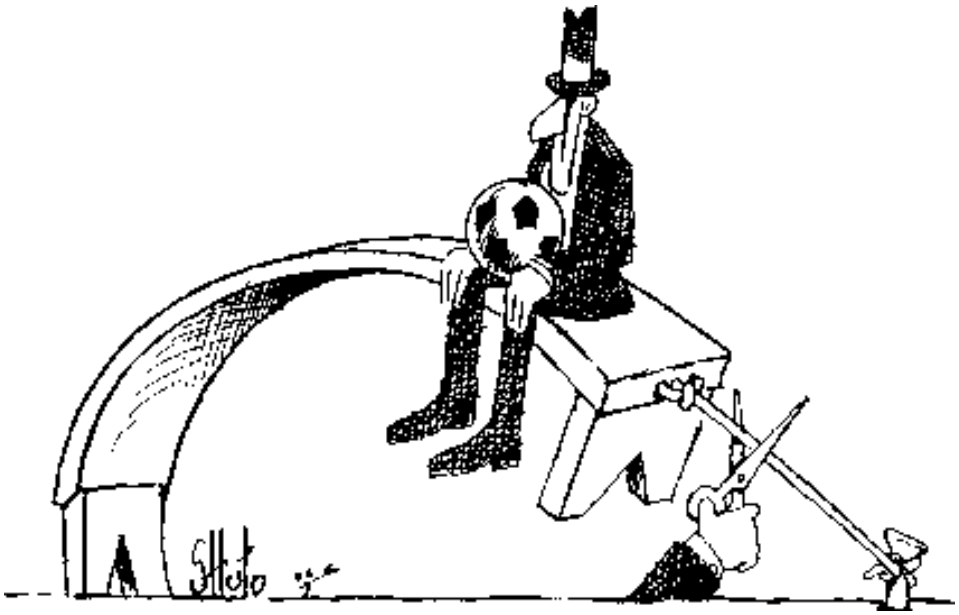
This table is a result from sections 3.2.3 and 3.3.4 and from the fact that for **M2** and **M3** we have the same number of correct predictions. Hence the feeling we had in the previous sections, *i.e.* that **M3** is the best method proposed, does not seem to be confirmed. **M3** does not predict more games correctly. Let us remind that for **M1** we obtained a weighted sum of games predicted correctly equal to 64.49% (see section 3.2.3) and for **M3** we obtained 64.62% (see section 3.3.4). At the end of section 3.3.4 we even tried another method to calculate the new coefficients where we obtained 64.5%.

Therefore we are able to conclude:

> **M1** which assumes the constancy of the team strength is the most appropriated method among the three.

This conclusion is based on the following points:

1. In assuming variation of the team strength we do not obtain better results. This statement has been verified in using 376 game outcomes, although the model was fitted on 442 game outcomes.
2. The practical application of **M1** is easier, faster in time and does not involve the search of an optimal γ.



5 Conclusion

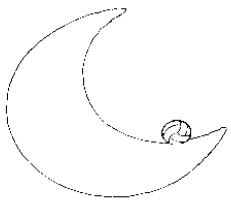
In this paper, the logistic regression model (3) using seed positions (conceived through a seeding coefficient) was applied to European soccer Cups tournament data in order to predict the probability of winning the tournament for each one of the participating teams, and the predicted probabilities of each team reaching a certain leg such as the quarter final. We proposed three different methods to do this: The first one assumed the constancy of the team strength and the two others assumed the variation of the team strength. In the previous section we compared the three methods and the conclusion was the following one: The best method is the one who assumes the constancy of the team strength. For every soccer fan this seems to be rather strange, but we obtained this fact in considering 376 game outcomes (!) of the European Cups from 1992 to 1996. So let us recapitulate this method which can be divided into two major parts:

1. Calculation of the weighted coefficients: One calculates a weighted mean of the ratio points achieved over games played for each of the three past years (see section 2.3.2).

2. Calculation of the probability of each team reaching a certain leg to obtain finally the probability of winning the tournament in using equations (1) and (2) (see section 2.4.2) with help of the logistic regression model (3).

Besides we remarked in section 3.2.3 that in using this method we had in average 64.49% of the games predicted correctly.

The attentive reader has certainly detected that we never took into consideration the scores. To calculate the new coefficient for a team we did not consider the goals for and against the teams in the tournament. This would surely be an interesting starting point to develop a new method for calculating new coefficients in hope to get a better method than the one who assumes the constancy of the team strength.



A Appendix

A.1 The Bradley-Terry Model

The Bradley-Terry model (Bradley and Terry 1952) is a model representing the results of experiments in which responses are pairwise rankings of treatments (so-called paired comparison experiments). Let us consider t treatments $T_1, \ldots, T_t$ in an experiment involving paired comparisons. We consider that these treatments have true ratings (or preferences) $\pi_1, \ldots, \pi_t$. These parameters represent relative selection probabilities for the treatments, subject to the constraints $\pi_i \geq 0$ $(i = 1, \ldots, t)$ and $\sum_{i=1}^{t} \pi_i = 1$. Bradley and Terry (1952) defined the probability that treatment T_i is preferred over the treatment T_j in a single comparison in the following way:

$$P(T_i \rightsquigarrow T_j) = \frac{\pi_i}{\pi_i + \pi_j} \tag{8}$$

where $i \neq j$ and $i, j = 1, \ldots, t$.

But what is the relationship between this model and our logistic model?

Let be $\pi_i = e^{\beta S_k(i)}$ and $\pi_j = e^{\beta S_k(j)}$ the ratings (or preferences) representing relative selection for the teams. Hence our logistic model (3) becomes:

$$\begin{aligned} P_k(i,j) &= \frac{e^{\alpha}\pi_i {\pi_j}^{-1}}{1 + e^{\alpha}\pi_i {\pi_j}^{-1}} \\ &= \frac{e^{\alpha}\pi_i}{\pi_j + e^{\alpha}\pi_i}. \end{aligned} \tag{9}$$

In section 3.2.1 we noticed that the regression rejects the necessity of the intercept α, hence equation (9) becomes:

$$P_k(i,j) = \frac{\pi_i}{\pi_j + \pi_i}$$

Therefore we can conclude that in rejecting the necessity of the intercept we obtain exactly the Bradley-Terry model (8) for a single comparison. The reject of $H_0 : \alpha = 0$ is also explicable by the following fact: α represents in a way the home advantage. But as for every round of the tournament two games are played because each team plays once at home and once on the road, hence the home advantage can be omitted.

A.2 Tournament Trees

A.2.1 Cup Winners Cup

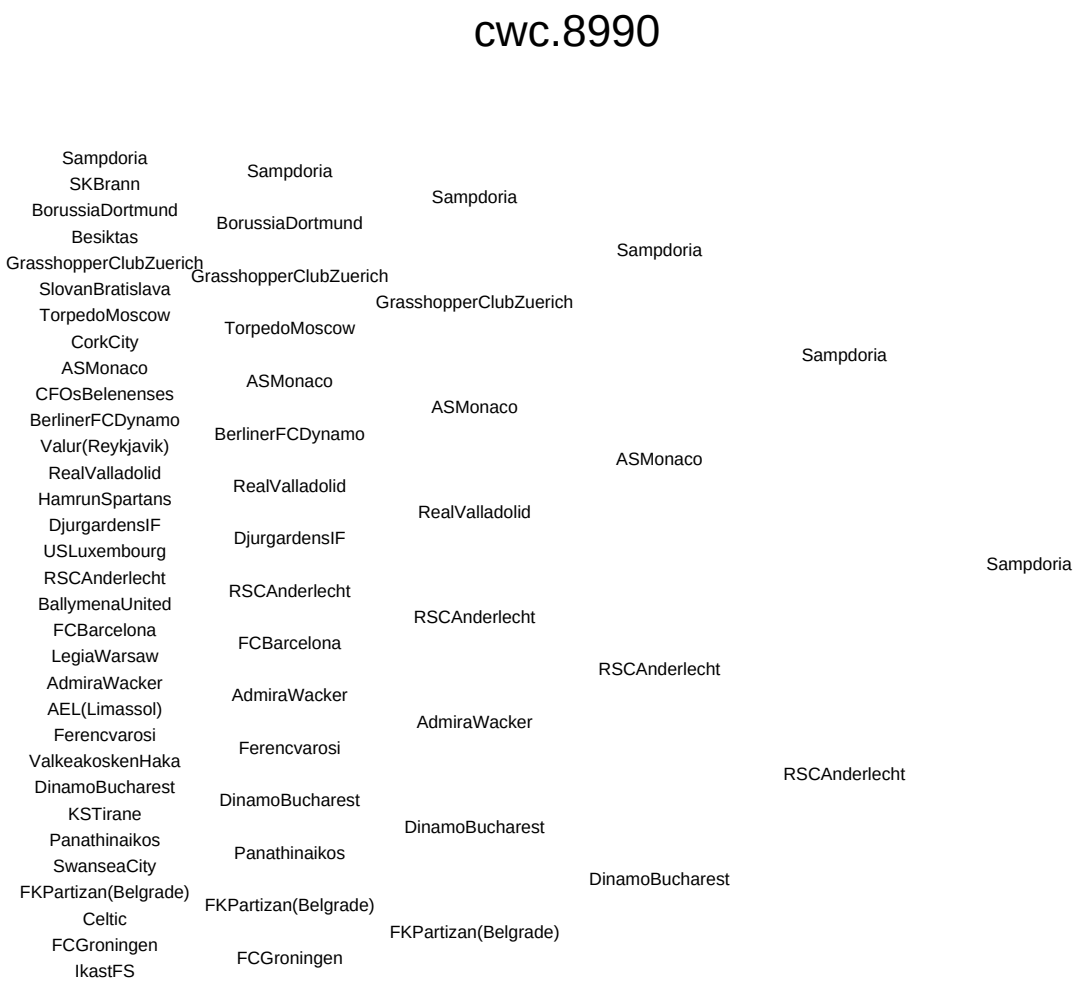


Figure 14 *The Cup Winners Cup 1989/90.*

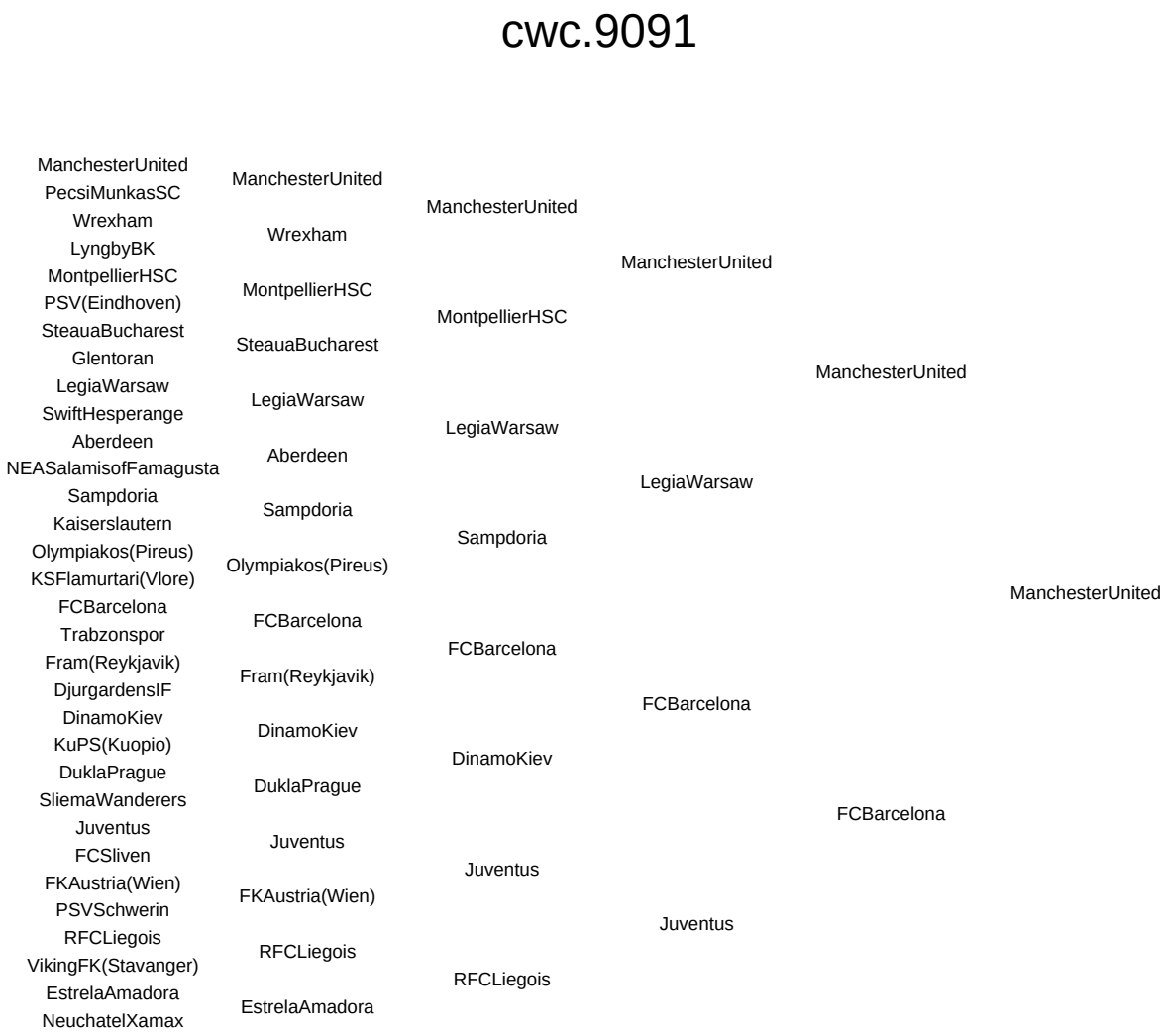


Figure 15 *The Cup Winners Cup 1990/91.*

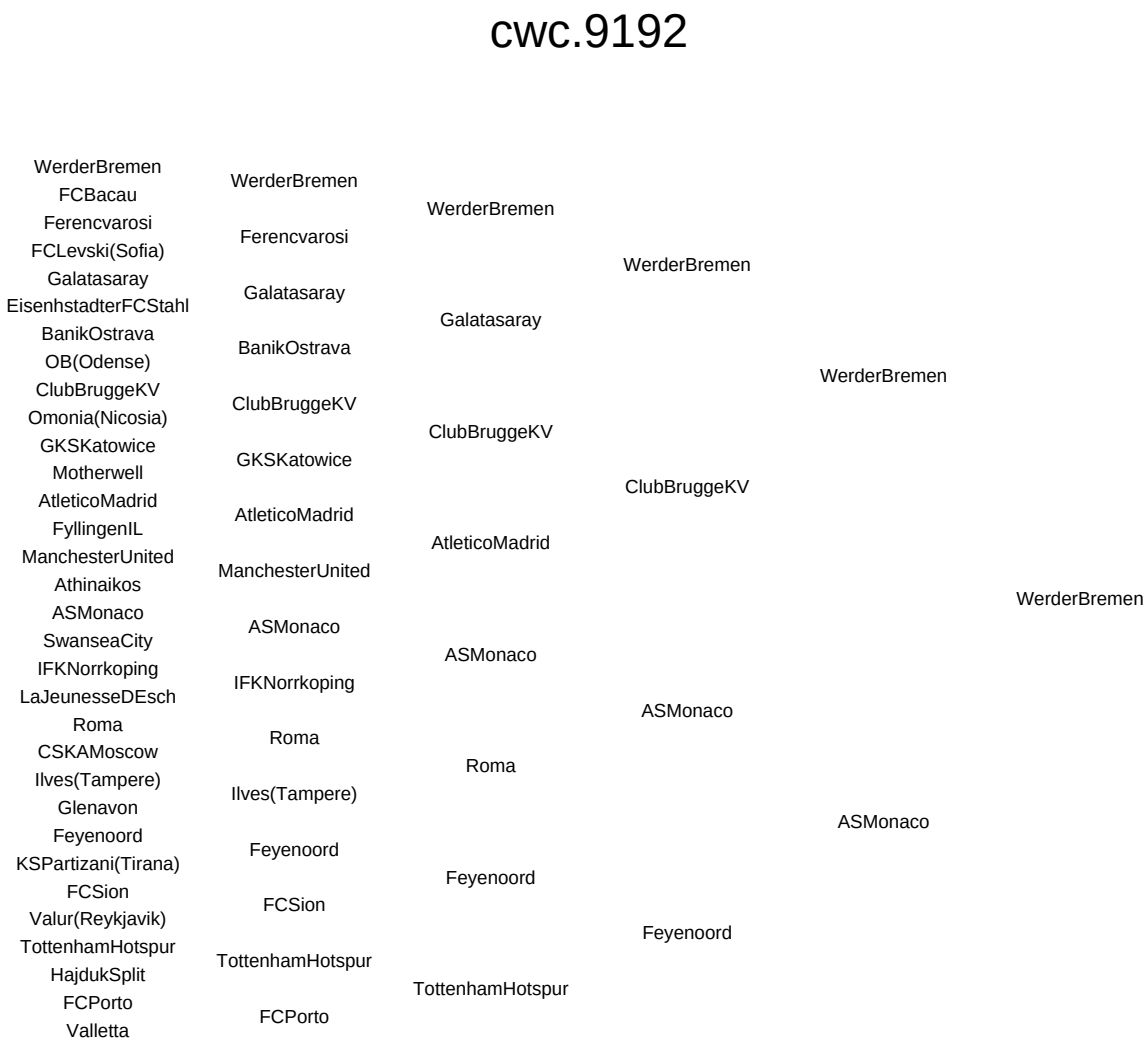


Figure 16 *The Cup Winners Cup 1991/92.*

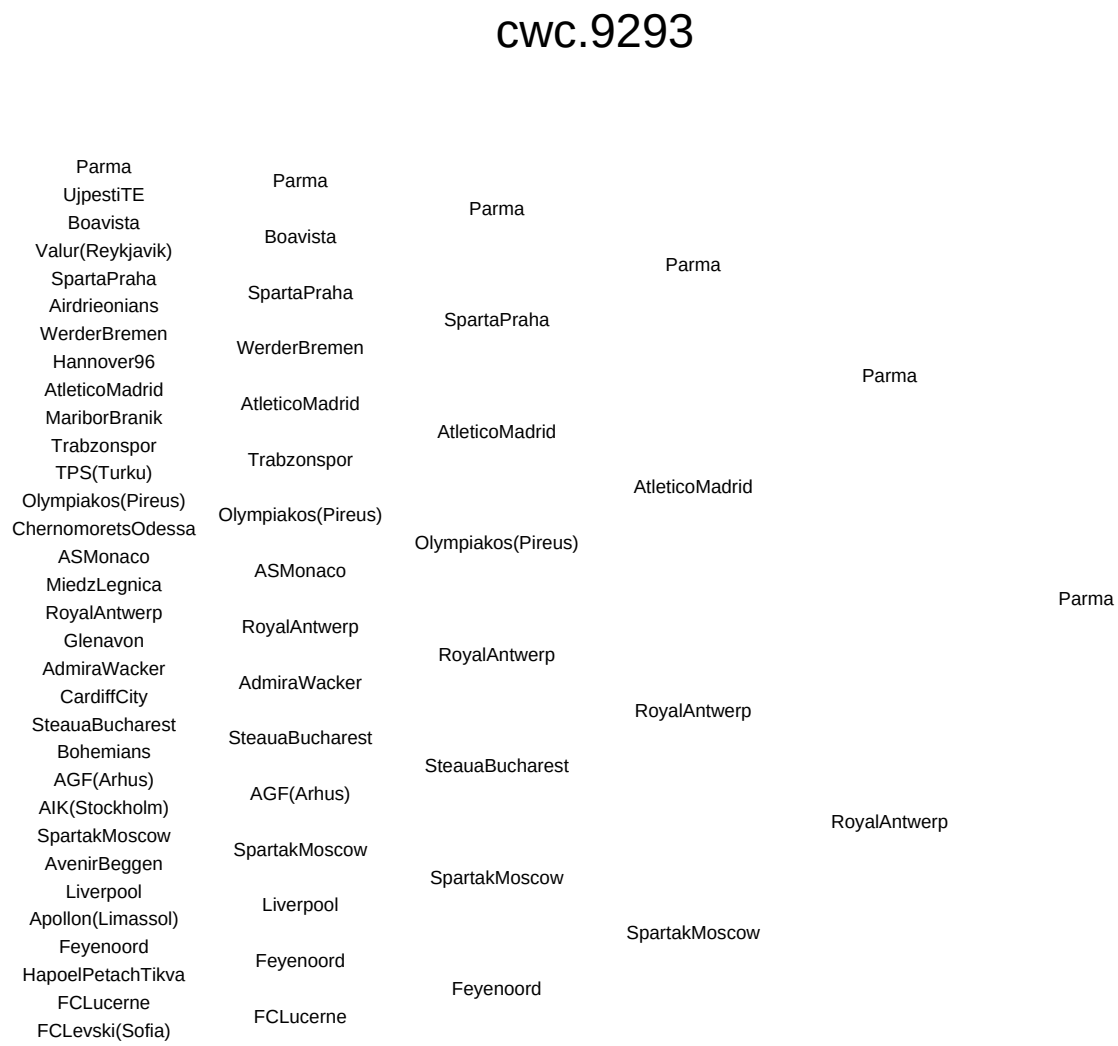


Figure 17 *The Cup Winners Cup 1992/93.*

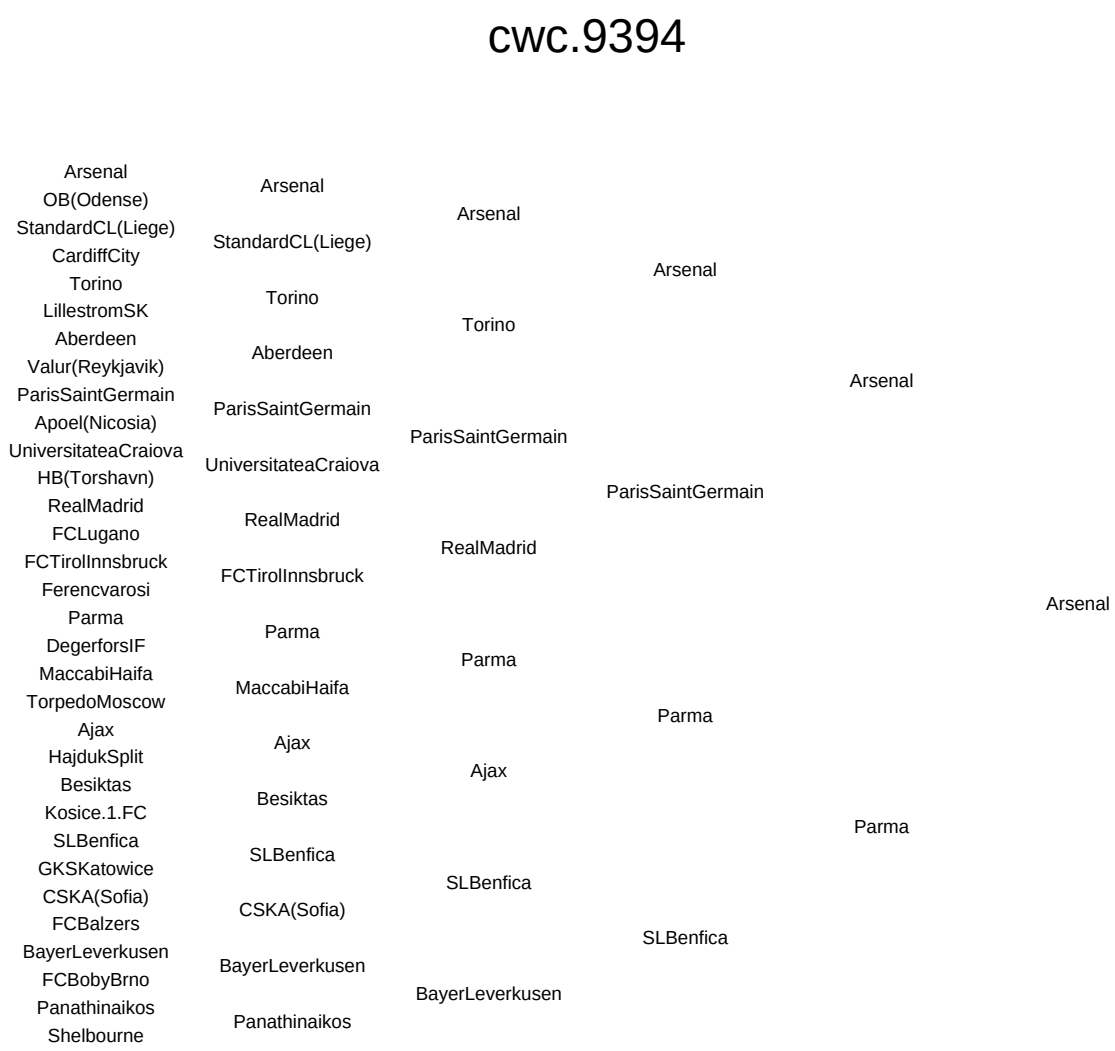


Figure 18 *The Cup Winners Cup 1993/94.*

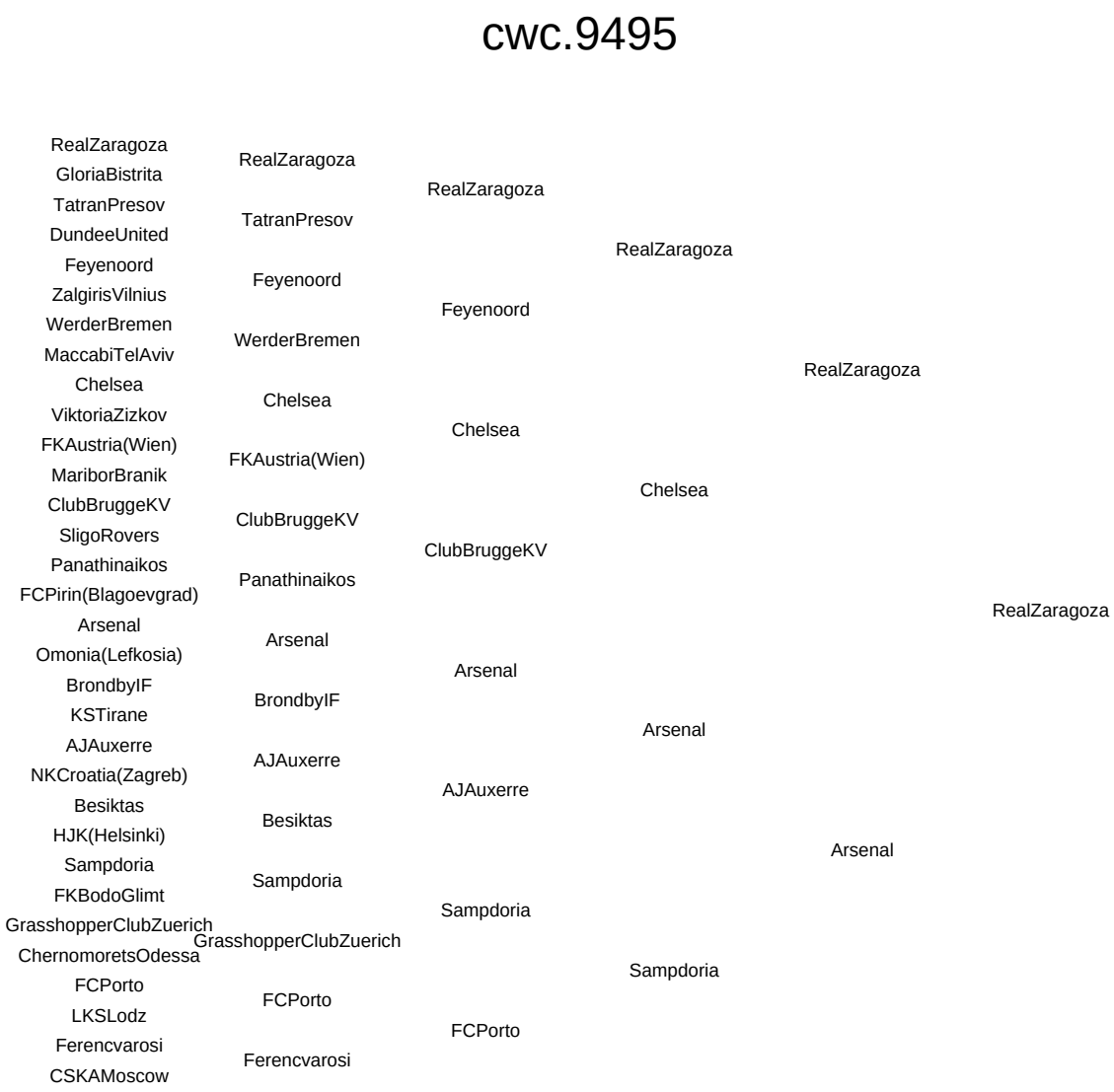


Figure 19 *The Cup Winners Cup 1994/95.*

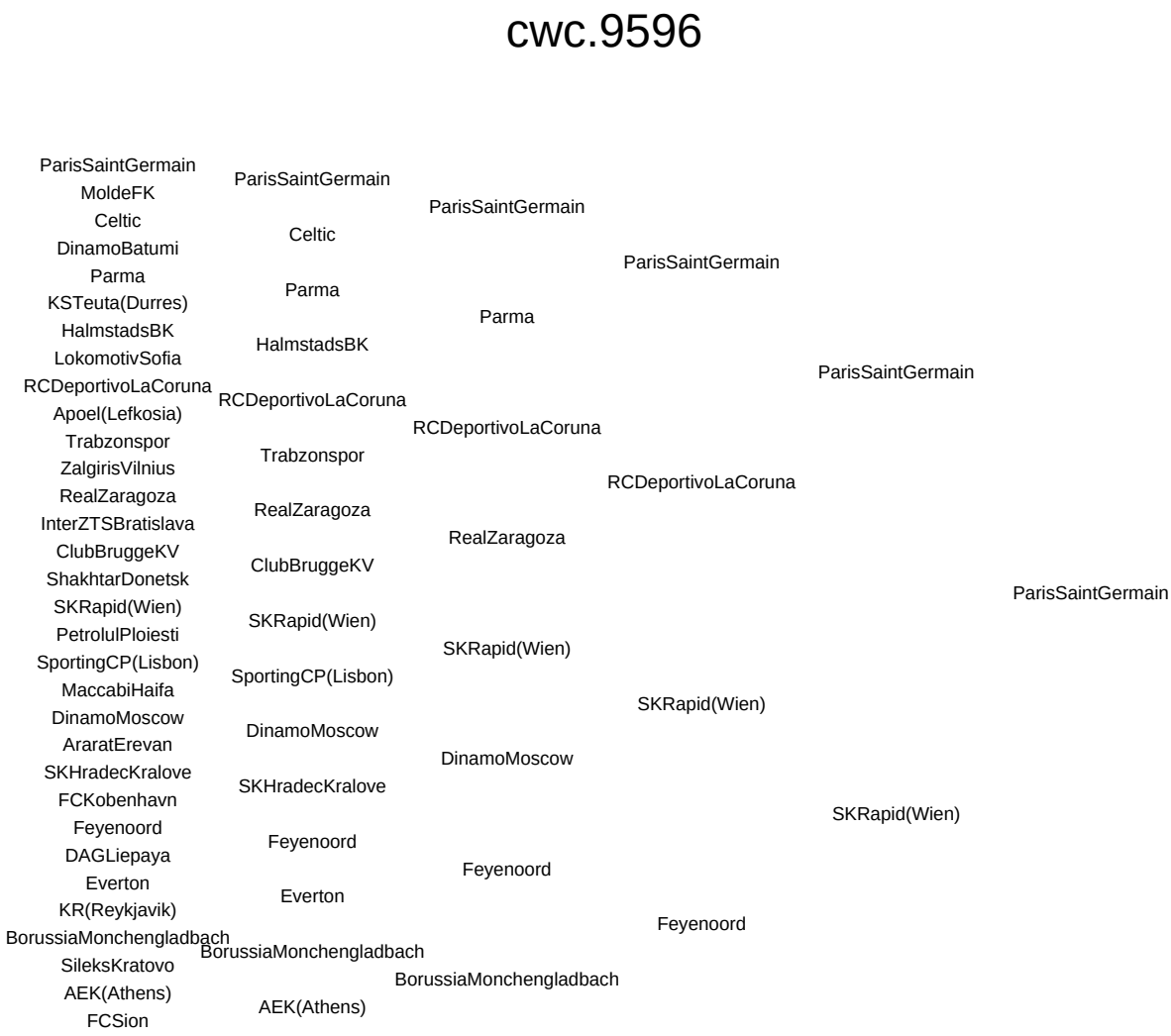


Figure 20 *The Cup Winners Cup 1995/96.*

A.2.2 UEFA Cup

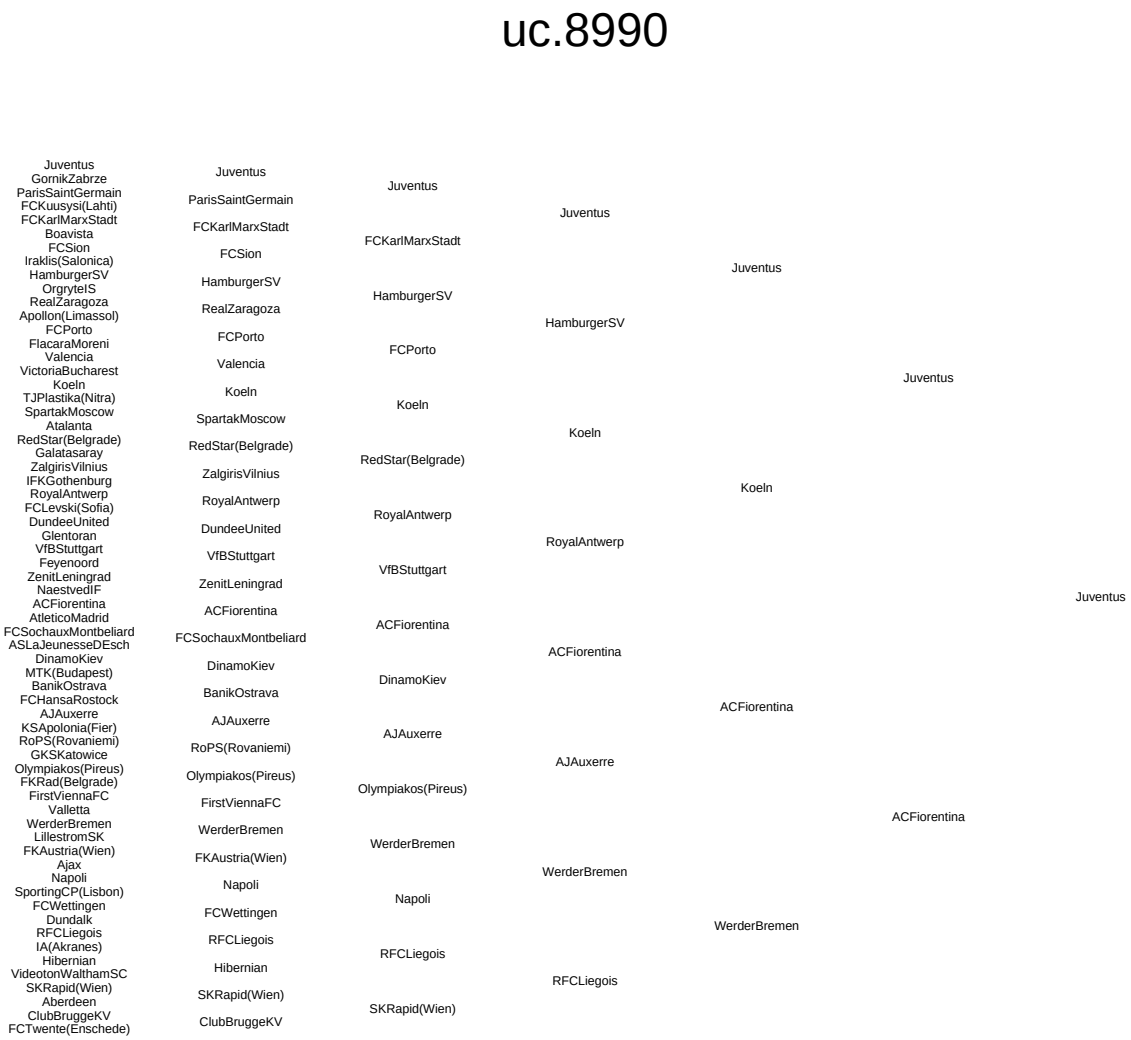


Figure 21 *The UEFA Cup 1989/90.*

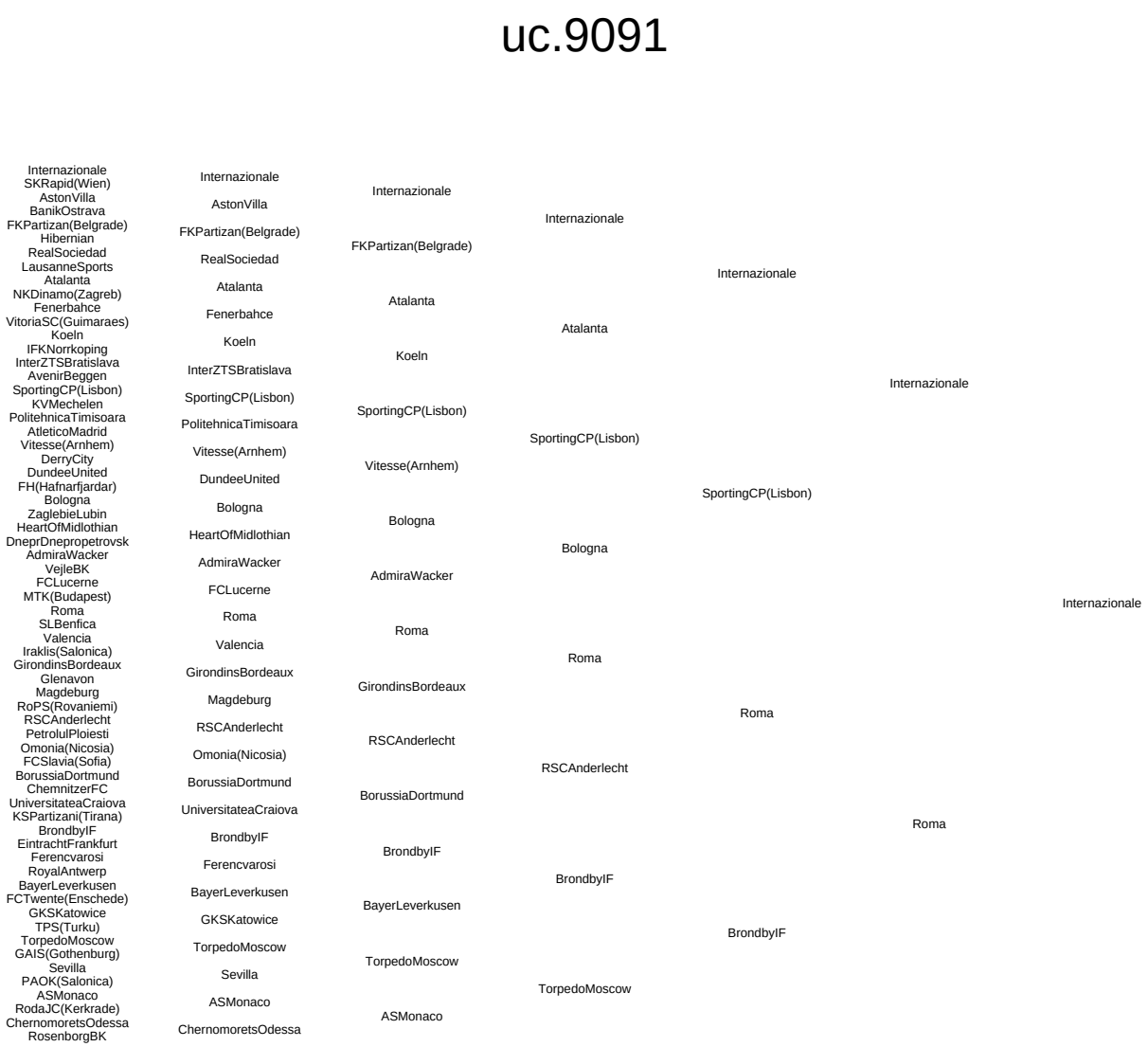


Figure 22 *The UEFA Cup 1990/91.*

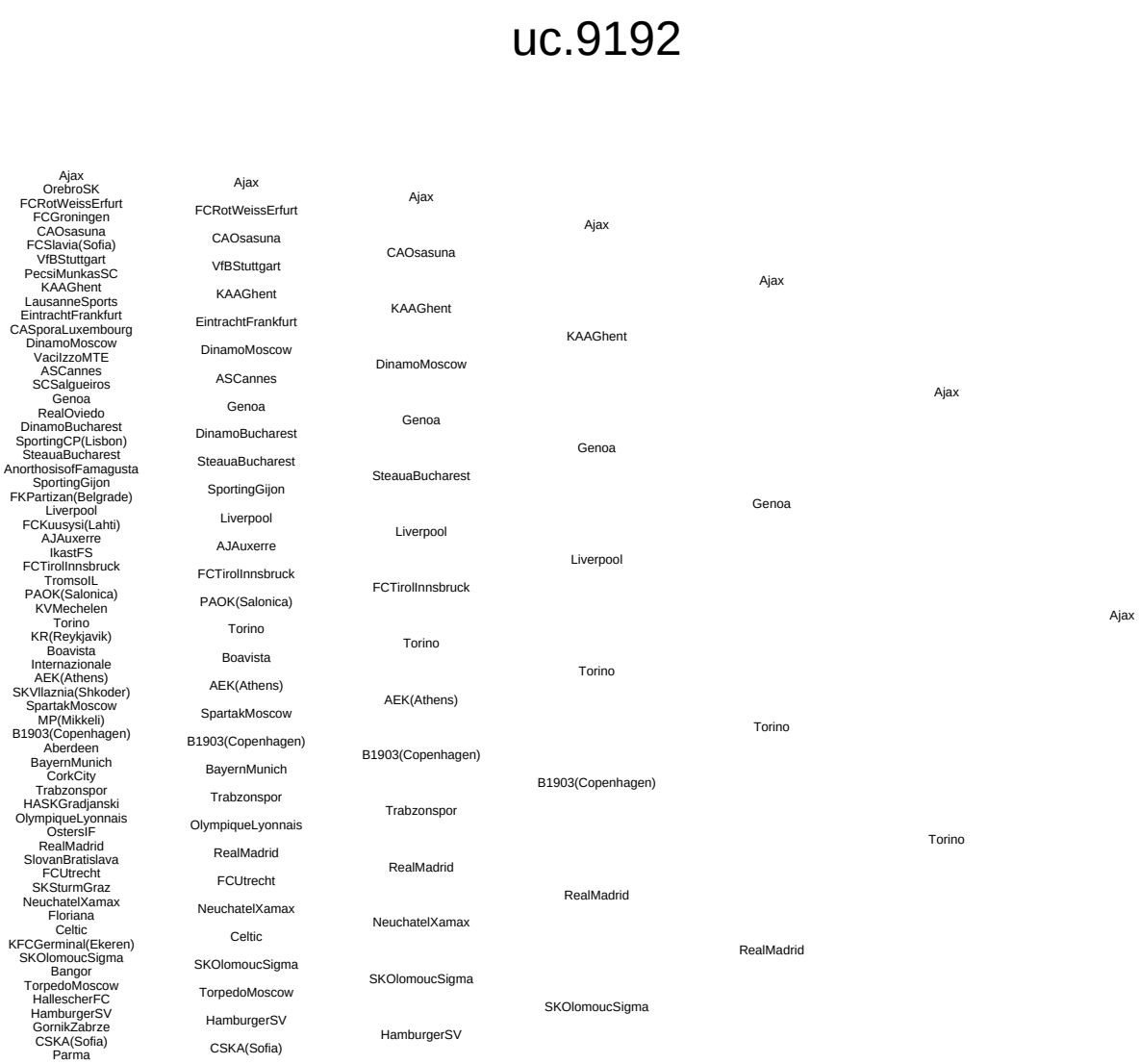


Figure 23 *The UEFA Cup 1991/92.*

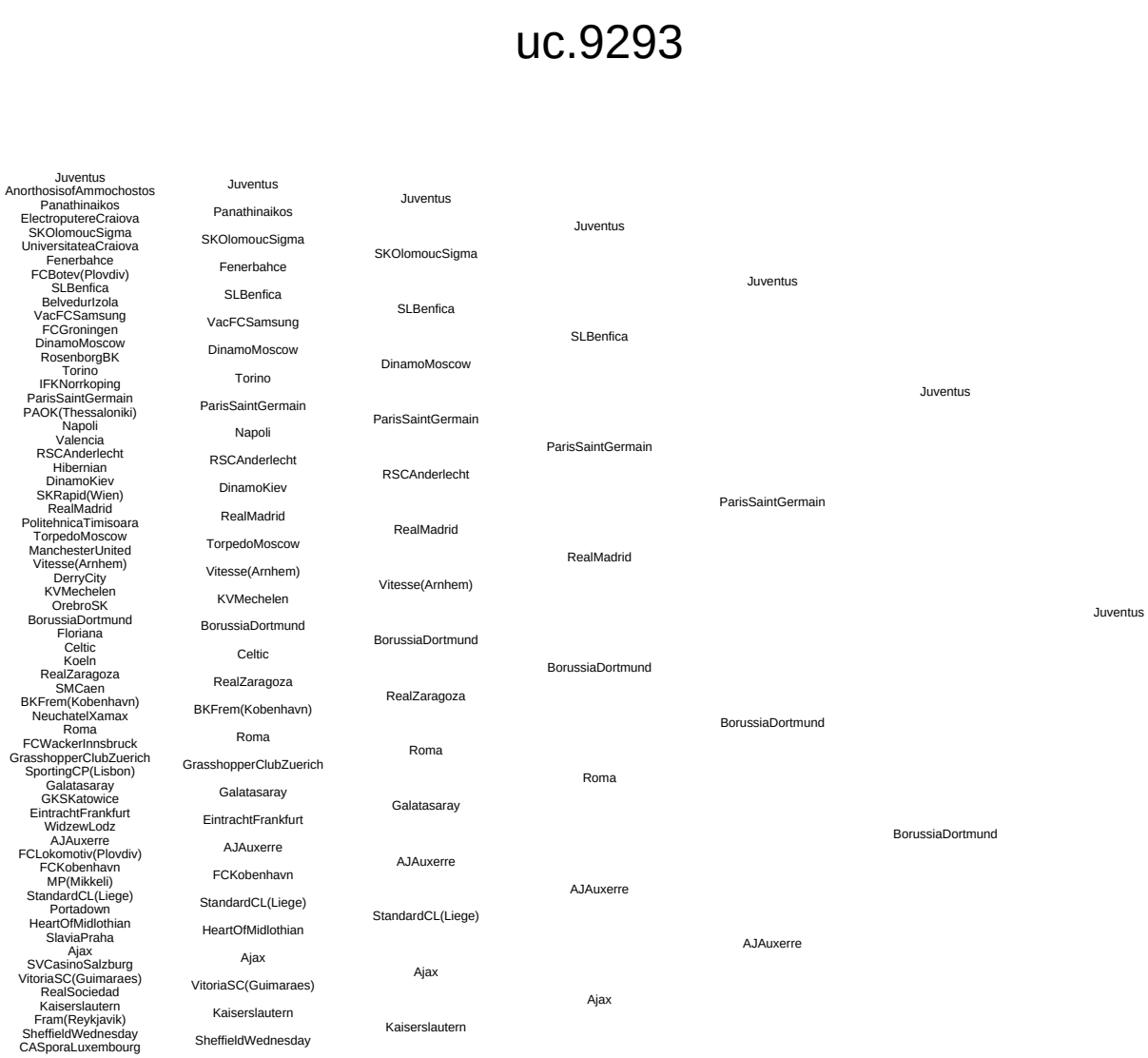


Figure 24 *The UEFA Cup 1992/93.*

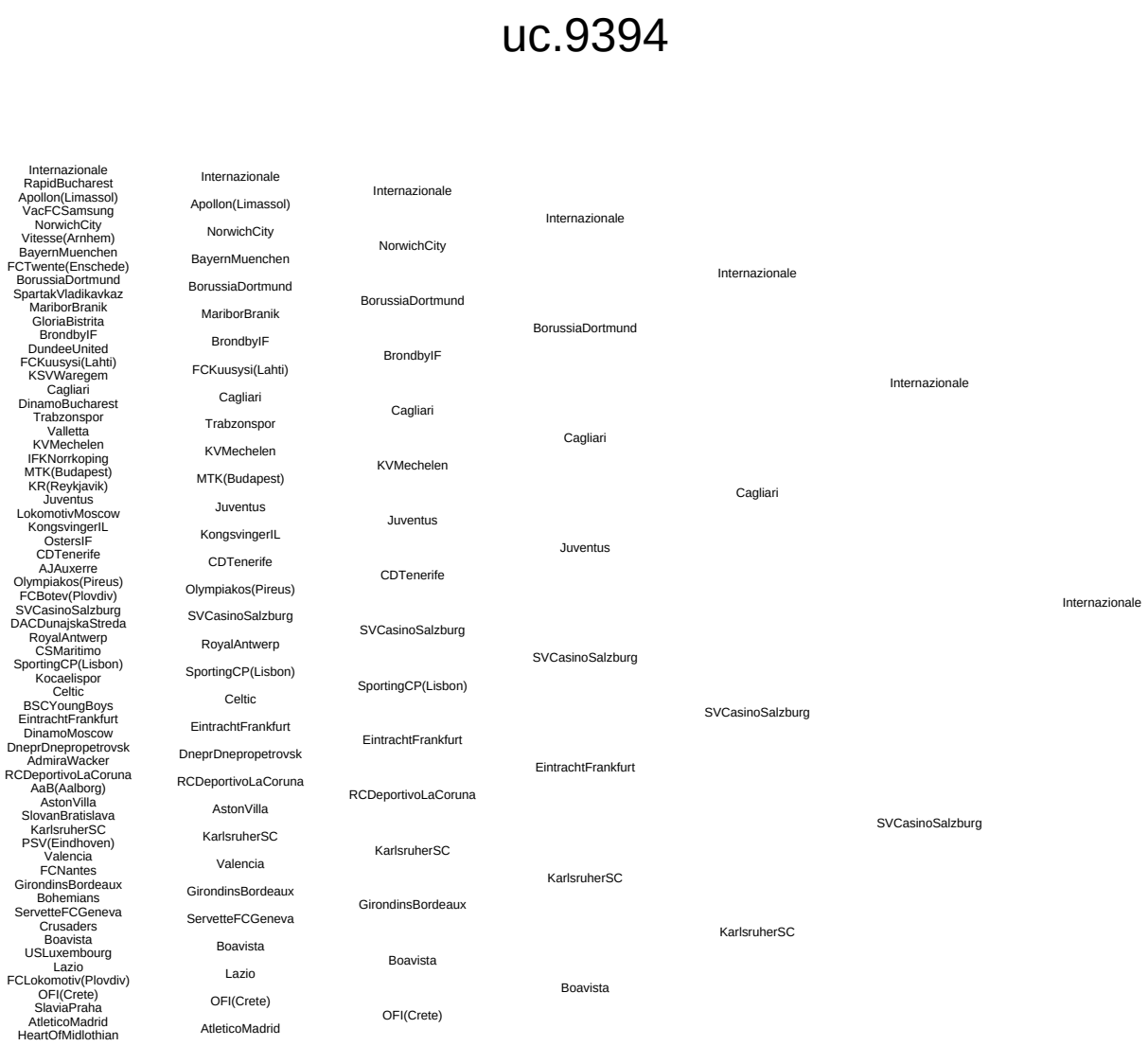


Figure 25 *The UEFA Cup 1993/94.*

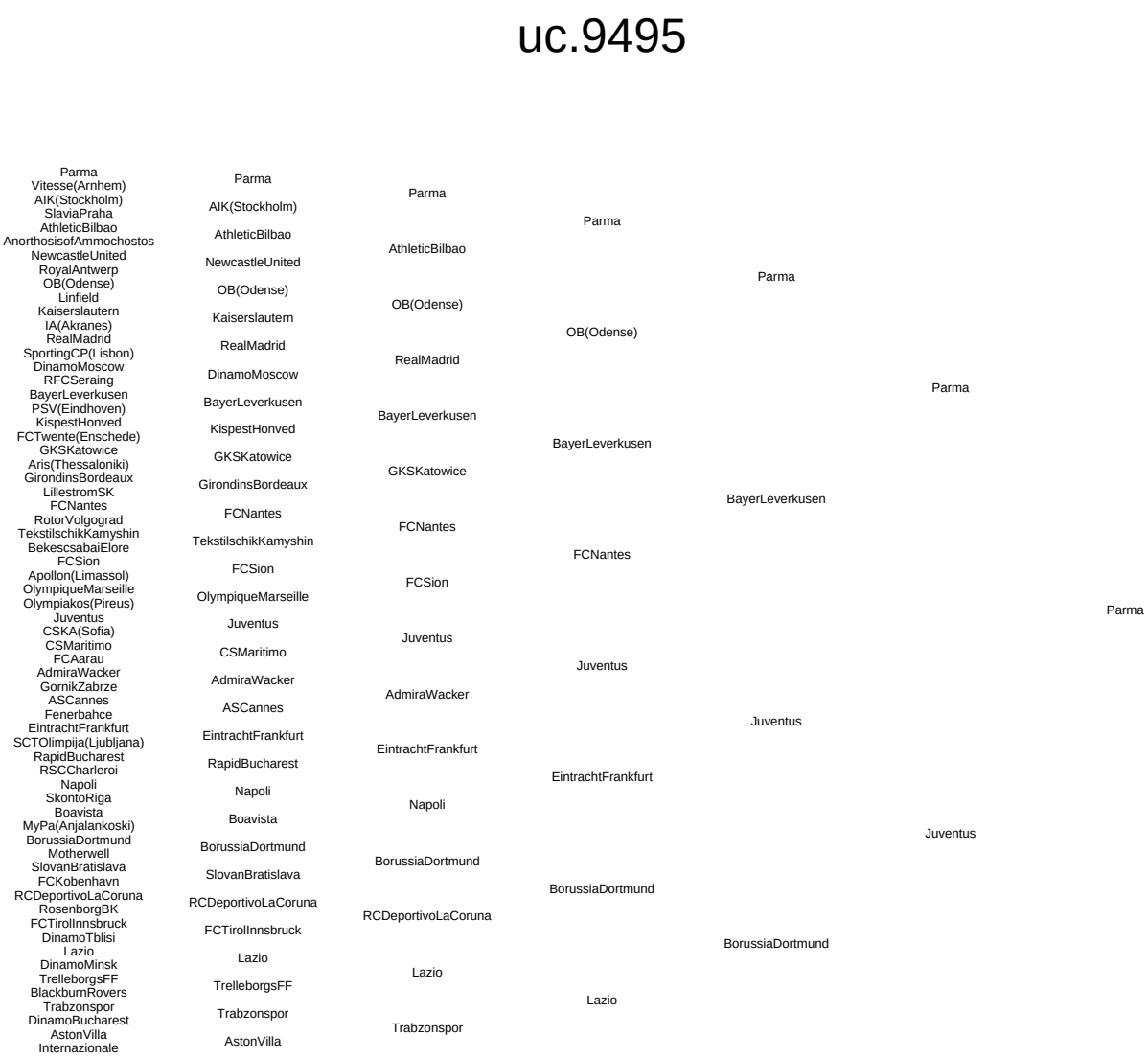


Figure 26 *The UEFA Cup 1994/95.*

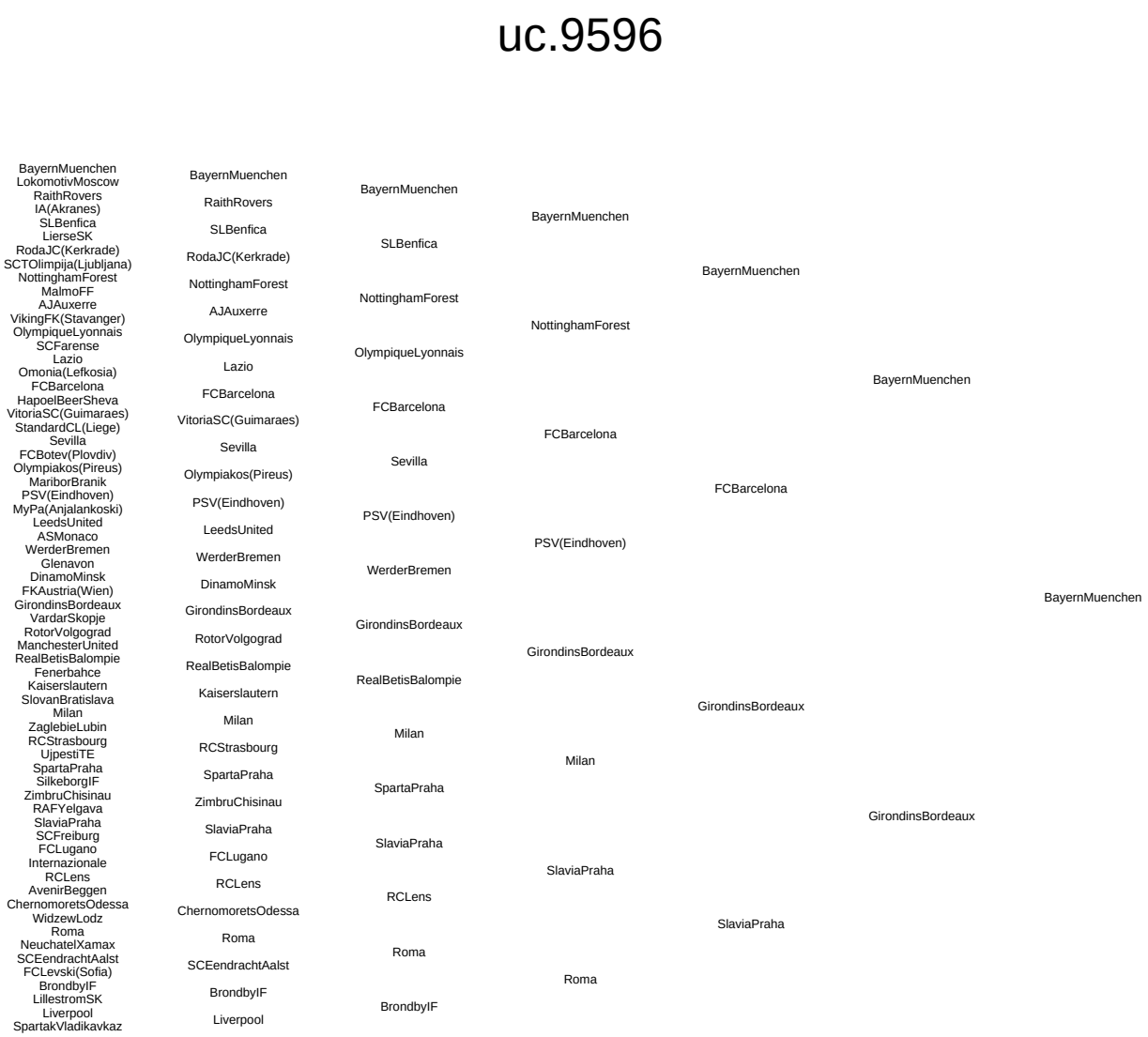


Figure 27 *The UEFA Cup 1995/96.*

A.3 Evaluation of the Coefficients

The UEFA Cup 1995/96 has been considered.

A.3.1 Using Probable Opponents

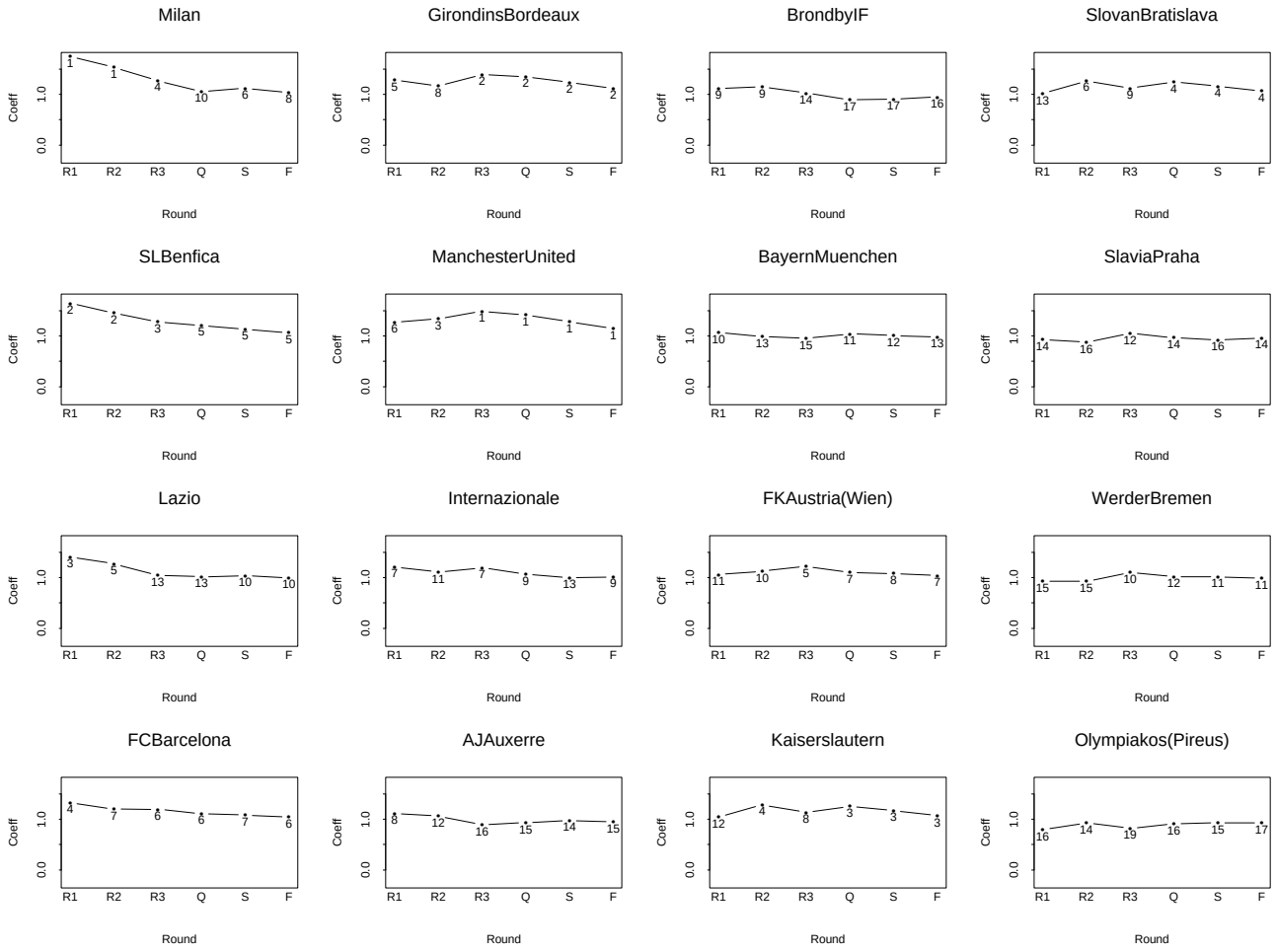


Figure 28 *Using probable opponents (UEFA Cup 1995/96).*

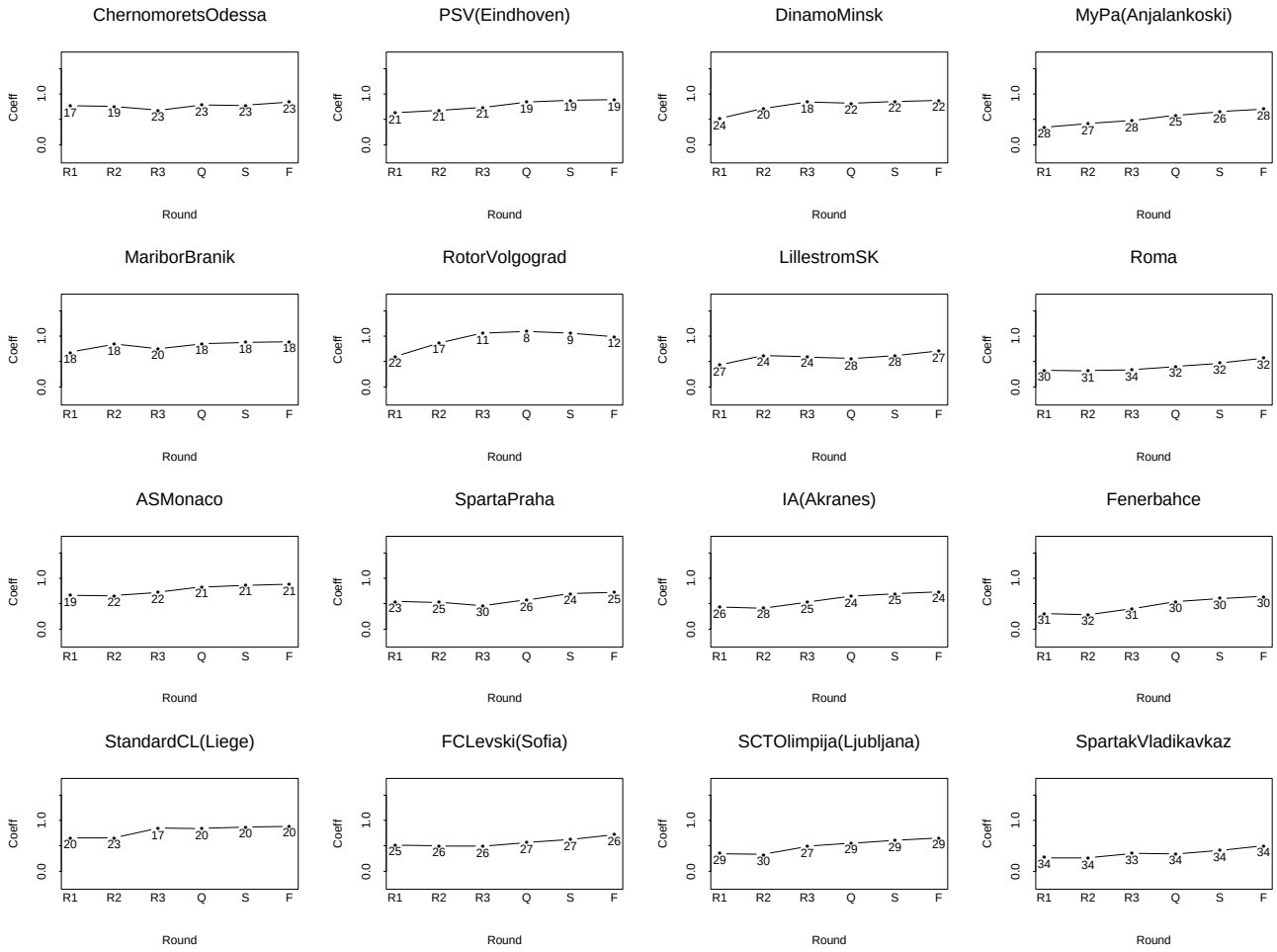


Figure 29 *Using probable opponents (UEFA Cup 1995/96).*

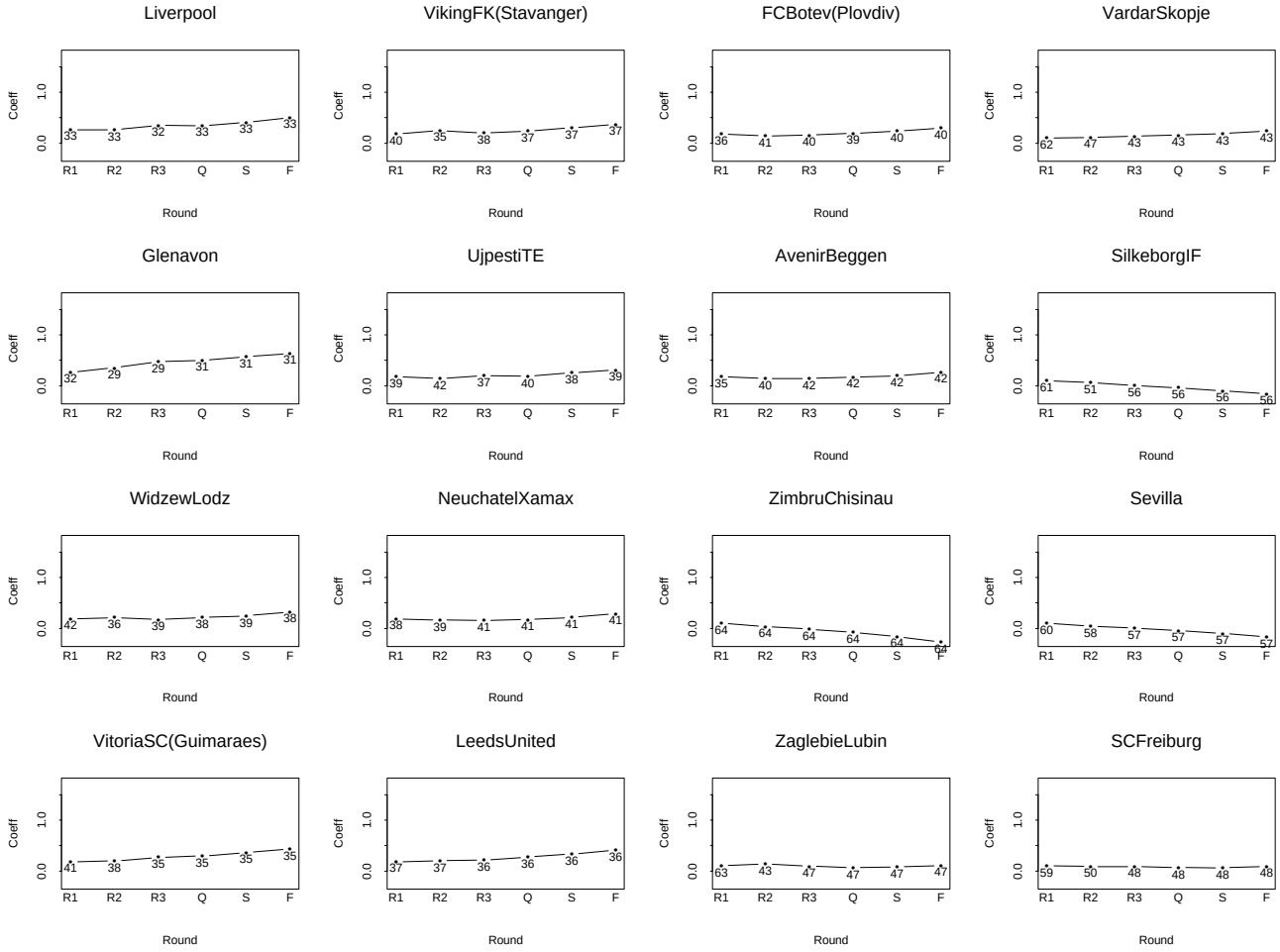


Figure 30 *Using probable opponents (UEFA Cup 1995/96).*

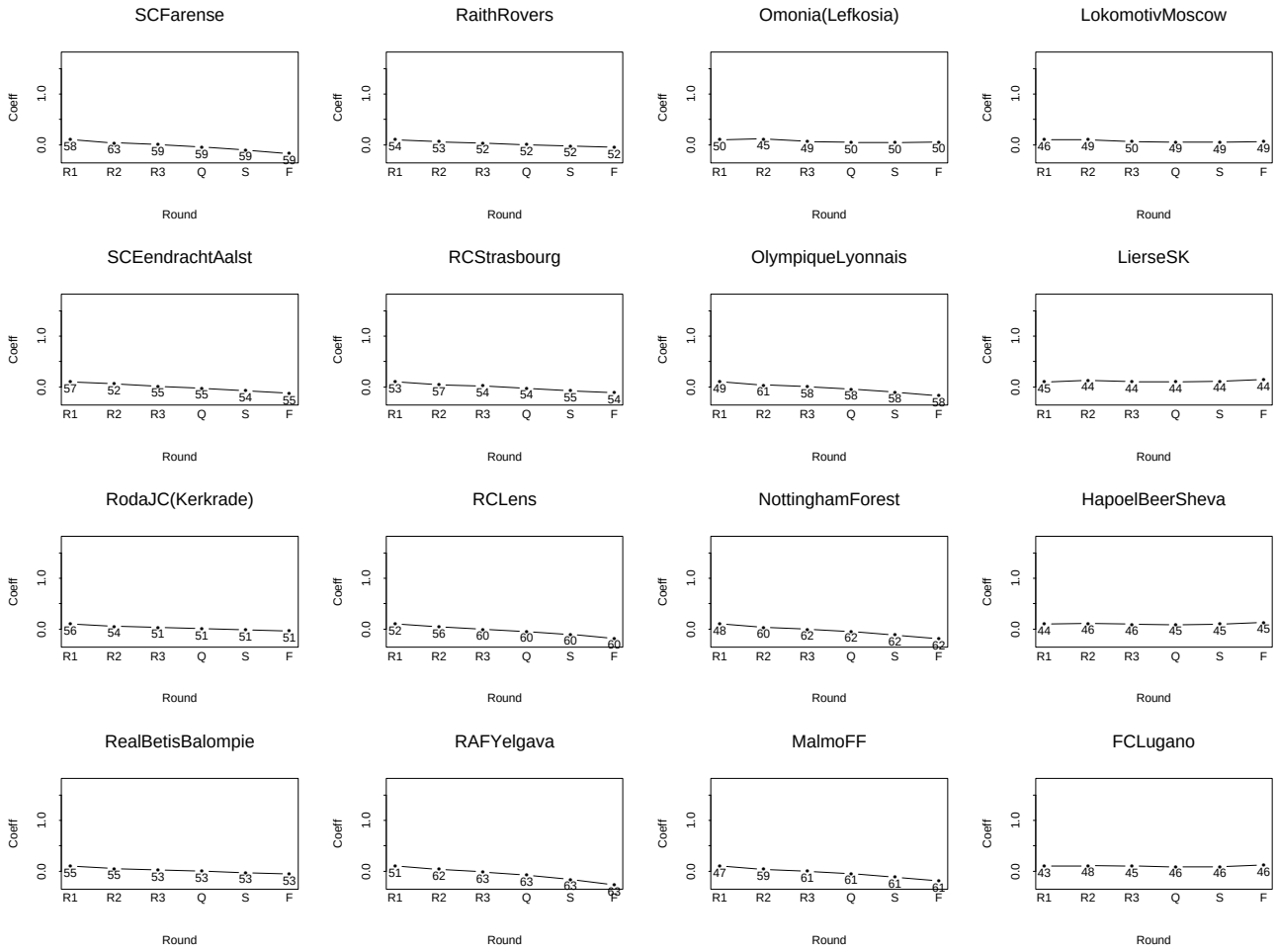


Figure 31 *Using probable opponents (UEFA Cup 1995/96).*

A.3.2 Using All Potential Opponents

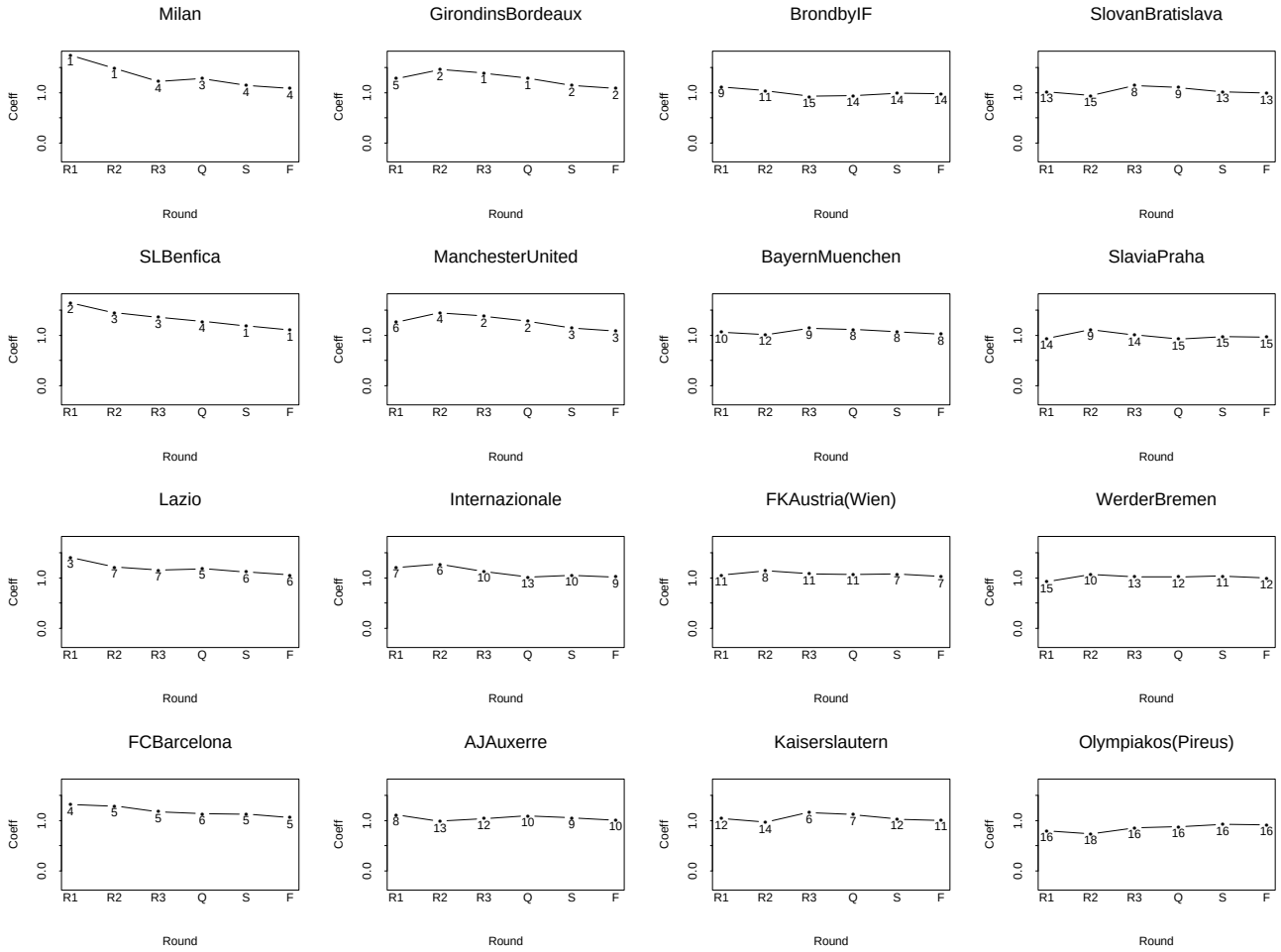


Figure 32 *Using all potential opponents (UEFA Cup 1995/96).*

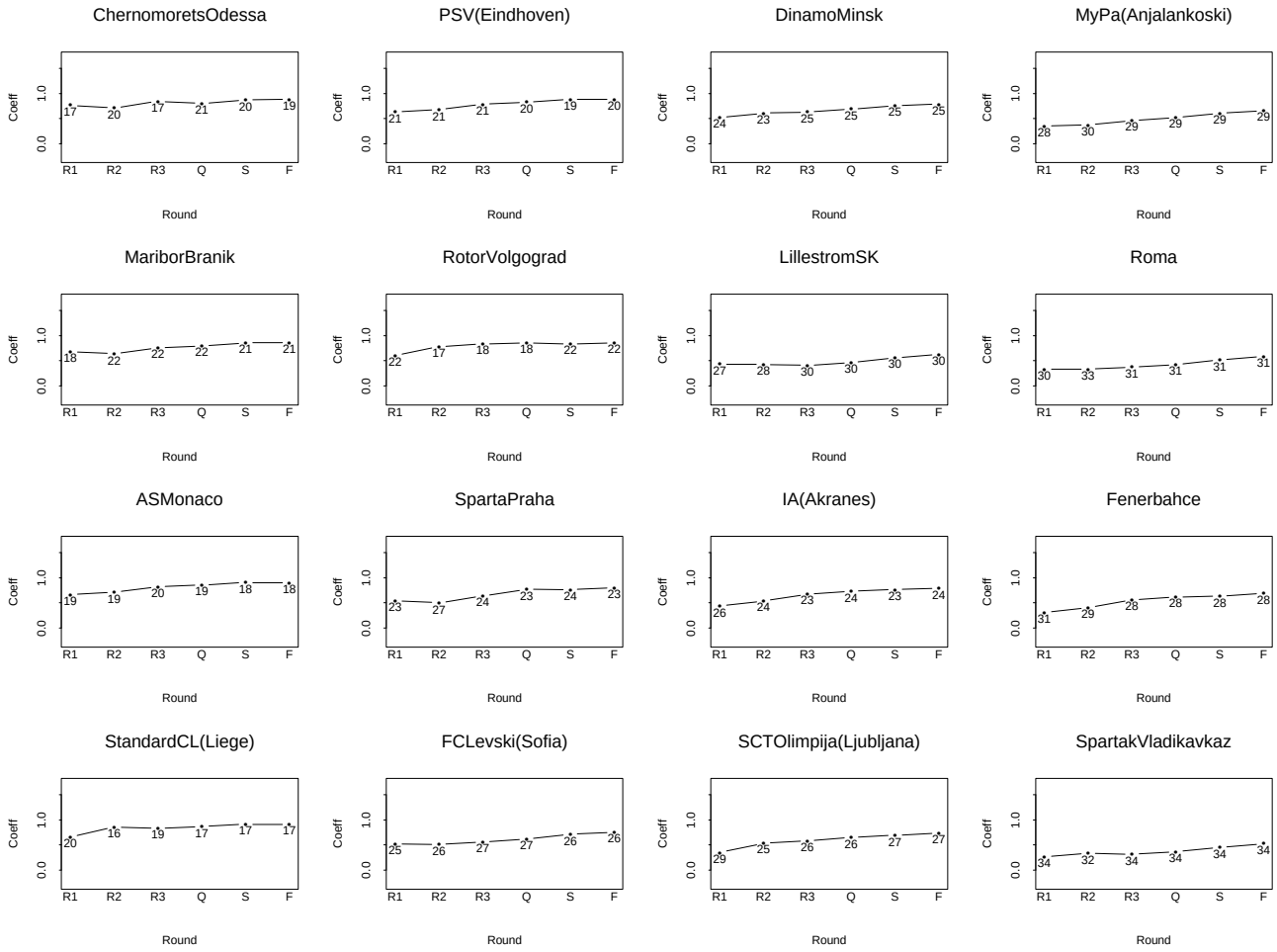


Figure 33 *Using all potential opponents (UEFA Cup 1995/96).*

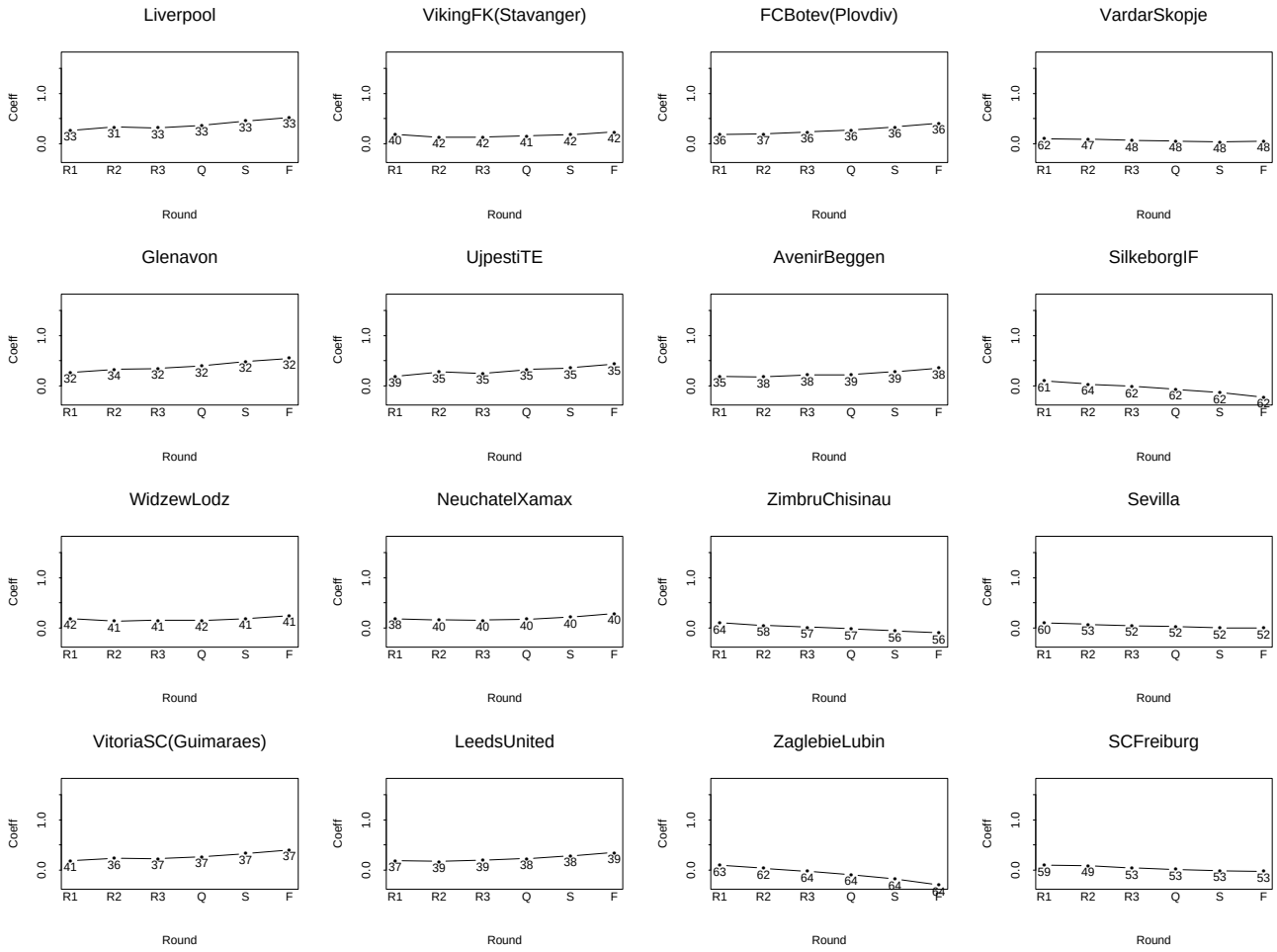


Figure 34 *Using all potential opponents (UEFA Cup 1995/96).*

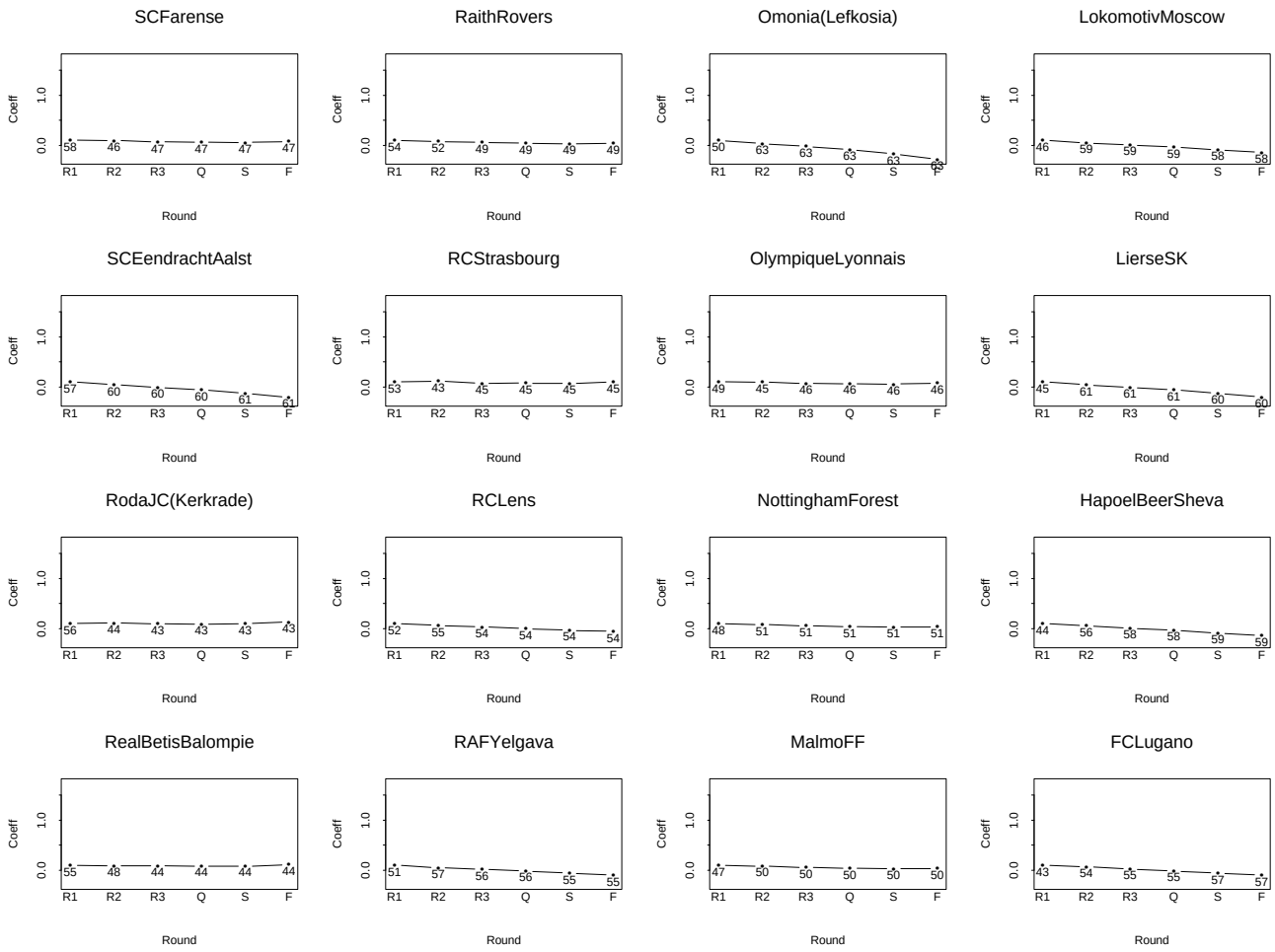


Figure 35 *Using all potential opponents (UEFA Cup 1995/96).*

B References

Bradley R.A. and Terry M.E. (1952).
Rank Analysis of Incomplete Block Designs - I. The Method of Paired Comparisons.
Biometrika, **39**, 324-345.

Protzen Martin and RSSSF (Rec.Sport.Soccer Statistics Foundation) (1996).
Qualification and Seeding for the European Cups (FAQ).
martin@inferenzsysteme.informatik.th-darmstadt.de.

Carlin B.P. (1996).
Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information.
The American Statistician, **50**, 39-43.

Schwertman N. C., McCready T.A. and Howard L. (1991).
Probability Models for the NCAA Regional Basketball Tournaments.
The American Statistician, **45**, 35-38.

Schwertman N. C., Schenk L. and Holbrook B.C. (1996).
More Probability Models for the NCAA Regional Basketball Tournaments.
The American Statistician, **50**, 34-38.