GDA of *engsoccerdata2* (from **engsoccerdata**)

Background The **engsoccerdata** package contains results of professional soccer in a number of European countries over their entire histories. The package is not yet on CRAN, but is available on GitHub.

Aims How has soccer changed over the years in England? Are there more draws now than before? Has home advantage always been the same? Do teams score fewer goals than in the old days? Can we display the histories of teams over the years? How do teams perform over a season?

Source James P. Curley (2015). engsoccerdata: English Soccer Data 1871-2015. R package version 0.1.4

Structure 190096 observations on 12 variables (6 factors, 1 date variable, 5 discrete variables)

Data preparation

The dataset is a list of the results of each game played in the top professional soccer leagues in England. For many purposes, especially for looking at performances by team, it would be better to have two lines for each game, one for the home team and one for the away team. The following code does this and reformats some of the data.

```
library(dplyr)
library(engsoccerdata)
data(engsoccerdata2)
esd <- tbl_df(engsoccerdata2 %>% filter(Season!=1939))
esd <- esd %>% mutate(Date=as.Date(Date, format="%Y-%m-%d"), tier=factor(tier))
allhome <- esd %>% mutate(team = home,
                          opp = visitor,
                          GF=as.numeric(as.character(hgoal)),
                          GA=as.numeric(as.character(vgoal)),
                          venue="home")
allaway <- esd %>% mutate(team = visitor,
                             opp = home,
                             GF=as.numeric(as.character(vgoal)),
                             GA=as.numeric(as.character(hgoal)),
                             venue="away")
allboth <- rbind(allhome,allaway) %>% mutate(GD = GF-GA,
                             result=(GD>0)*1 + (GD==0)*0.5) %>%
  select(Date, Season, division, tier, team, opp,
         GF, GA, GD, result, venue)
```

The league started out with one division and quickly added another. After the First World War a third division started and this was doubled and split into North and South divisions one year later. In 1958 these divisions were turned into new third and fourth divisions. In 1992 the old first division became the Premier League and the other divisions also changed their names (and again later). For simplicity's sake this report, like the dataset, refers to the different levels as the four tiers.

Much information on the intricate details of football history can be found on the Wikipedia sites for individual league seasons. These have been very helpful for explaining the basics and understanding various initally surprising features.

James Curley has published three interesting articles on the 538 website using this dataset (http://fivethirtyeight.com/contributors/james-curley/).

Draws

The proportion of draws per game over time for the top two tiers of the English league is shown with a gam smoother in Figure 1.

```
library(mgcv)
Draws <- esd %>% group_by(Season, tier) %>%
  summarise(ngames=n(), draws=sum(result=="D"), drx=draws/ngames)
ggplot(Draws %>% filter(tier %in% c(1,2)), aes(Season, drx)) + geom_point(size=0.25) +
  ylab("Proportion of draws per game") + geom_vline(xintercept=1981, colour="red") +
  ylim(0, 0.4) + stat_smooth(method = "gam", formula = y ~ s(x))
```

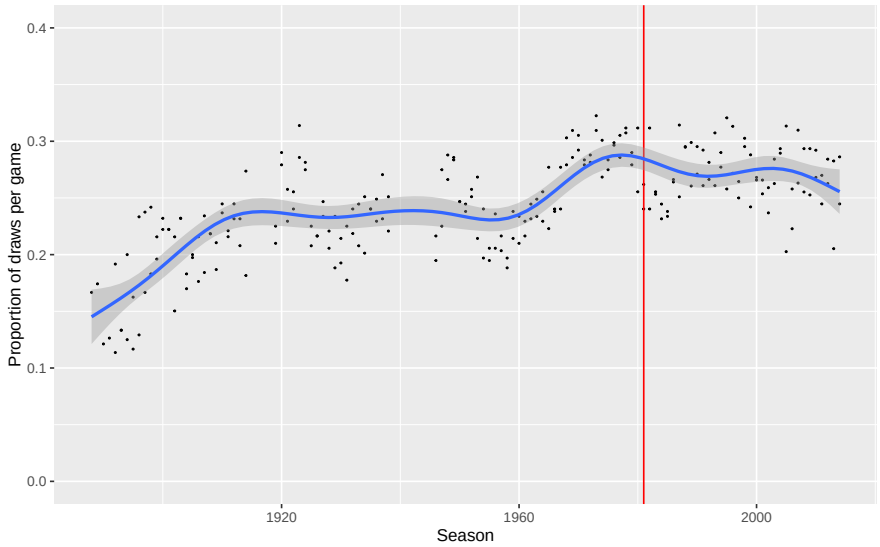

Fig 1: The proportion of draws in the top two tiers climbed initially until the First World War. Afterwards it remained relatively constant till the 1960s, when it rose again. After the introduction of three points for a win in 1981-2 (the red line) it dropped a little again. There was quite a lot of variability by season.

The rates of draws in other countries were sometimes quite different as Figure 2 shows.

```
data(italycalcio)
DrawsI <- italycalcio %>% mutate(result=(hgoal>vgoal)*1+(hgoal==vgoal)*0.5) %>%
group_by(Season) %>% summarise(ngames=n(), draws=sum(result==0.5), drx=draws/ngames)
data(spainliga)
DrawsS <- spainliga %>% mutate(result=(hgoal>vgoal)*1+(hgoal==vgoal)*0.5) %>%
group_by(Season) %>% summarise(ngames=n(), draws=sum(result==0.5), drx=draws/ngames)

ggplot(Draws %>% filter(tier==1), aes(Season, drx))  + labs(y=NULL) + ylim(0, 0.4) +
  stat_smooth(method = "gam", formula = y ~ s(x), colour="red", se=FALSE) +
  geom_vline(xintercept=1981, colour="red", linetype="dashed") +
  geom_smooth(data=DrawsI, aes(Season, drx), method = "gam", formula = y ~ s(x),
  colour="blue", se=FALSE) +
  geom_vline(xintercept=1994, colour="blue", linetype="dashed") +
  geom_smooth(data=DrawsS, aes(Season, drx), method = "gam", formula = y ~ s(x),
  colour="gold2", se=FALSE) +
  geom_vline(xintercept=1995, colour="gold2", linetype="dashed")
```

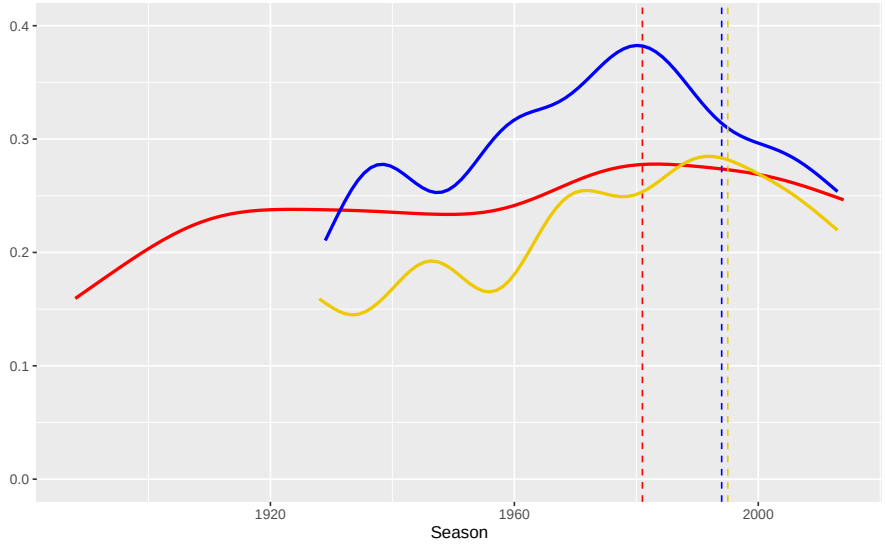

Fig 2: The proportions of draws over time for the top tiers in Italy (blue), England (red), and Spain (gold). The corresponding dotted lines show when the 3 points for a win rule was introduced in the different countries. Italy had a higher rate of draws, particularly in the 1970s and 1980s. The rate in Spain was lower, rose above the English rate, then fell again.

Home advantage

Home advantage is a common feature of team sports. One way to measure it is to consider the proportion of a team's points won at home. As the number of points awarded for a win was increased in 1981 from 2 to 3, the proportions before and after 1981 would not be comparable. Figure 3 assumes that a win was always worth 3 points.

```
library(tidyr)
HA3a <- allboth %>% filter(venue=="home") %>% group_by(Season, tier) %>%
  summarise(ng=n(), nh=sum((result==1)*3 + (result==0.5)*1),
            na=sum((result==0)*3 + (result==0.5)*1)) %>% mutate(HomeAway=nh/(nh+na))
HA3a1 <- complete(data.frame(HA3a), Season=full_seq(Season, 1), tier)
ggplot(HA3a1, aes(Season, HomeAway, group=tier, colour=tier)) + geom_line() +
  ylab("Home proportion of points") +
  theme(legend.position="bottom") + guides(col=guide_legend(nrow=1, byrow=TRUE))
```

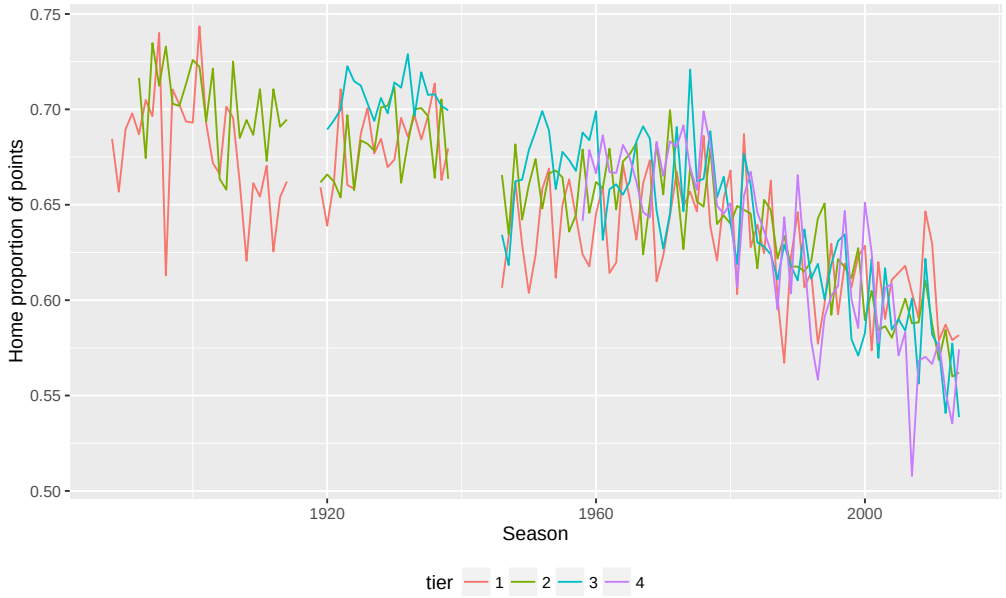

Fig 3: The proportion of points won at home over time in the various tiers. The most striking feature is the relative steady decline since 1980 for the lower three tiers. The sharp drop in levels when the leagues restarted after the Second World War is a little surprising, why would that have happened? There is quite a lot of variability over the years and some relatively extreme values, for example the fourth tier in 2007 when there were almost as many away wins as home wins.

You can also look at the home away records for individual teams and Figure 3 shows the results for the four tiers separately (the gaps are the First and Second World Wars).

```
HA3x <- allboth %>%  group_by(Season, tier, team, venue) %>%
  summarise(ng=n(), np=sum((result==1)*3 + (result==0.5)*1))
HA3y <- spread(HA3x %>% select(-ng), venue, np)
HA3y <- HA3y %>% mutate(HomeAway=home/(home+away))
ggplot(HA3y, aes(Season, HomeAway)) + geom_point() + facet_wrap(~tier, nrow=2) +
  ylim(0,1) + ylab("Home proportion of points by team")
```

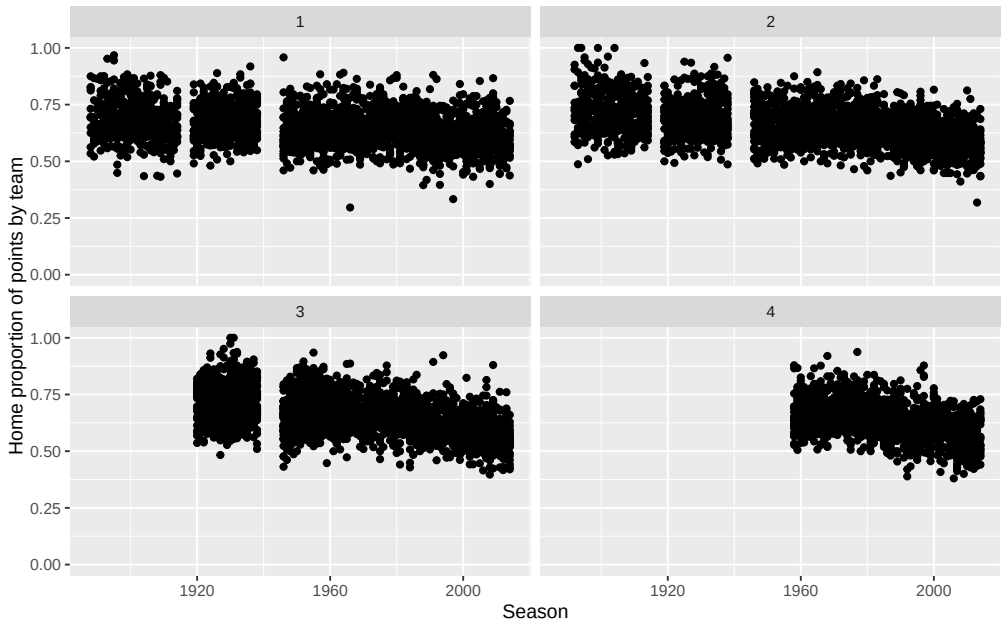

Fig 4: The individual home away points ratio for every team for each season. The overall trend is the same as in Figure 3 (but note the difference in vertical scale). There are some curious individual outliers.

The four outliers in the early days of the second division who got no points away from home (and the two from the third division later on) were:

```
data.frame(HA3y %>% filter(HomeAway > 0.999))
```

```
##   Season tier               team away home HomeAway
## 1   1893    2 Northwich Victoria    0   12        1
## 2   1894    2    Crewe Alexandra    0   13        1
## 3   1899    2       Loughborough    0    9        1
## 4   1904    2   Doncaster Rovers    0   11        1
## 5   1930    3             Nelson    0   25        1
## 6   1931    3      Wigan Borough    0   10        1
```

They all finished last except for Wigan Borough in 1931, who folded. In the first season after the Second World War Leeds United drew one away game and lost the other 20, although they managed six wins and five draws at home. (And yes, they finished last as well.)

The teams with the best relative away record were:

```
data.frame(HA3y %>% filter(HomeAway < 0.35))
```

```
##   Season tier            team away home  HomeAway
## 1   1966    1       Blackpool   19    8 0.2962963
## 2   1997    1  Crystal Palace   22   11 0.3333333
## 3   2013    2 Birmingham City   30   14 0.3181818
```

The first two were relegated, while Birmingham survived on goal difference.

Goals

The number of goals scored per game in the first few years of the league was a lot higher than now, but it declined quickly. In 1925 the offside law was changed and that led to a short-term dramatic increase. After the Second World War levels rose again to peak around 1960. There was then a steady decline till 1970 and since then goal scoring has remained fairly constant, close to 2.5 goals a game. This is all shown in Figure 5.

```
SGoals <- allboth %>% group_by(Season, tier) %>% summarise(gg=sum(GF+GA), ngames=n()) %>%
  mutate(GPG=gg/ngames)
SG <- complete(data.frame(SGoals), Season=full_seq(Season, 1), tier)
ggplot(SGoals, aes(Season, GPG)) + ylim(0,5) + ylab("Goals per game") +
  geom_line(aes(group=factor(tier), colour=factor(tier))) +
  theme(legend.position="bottom", legend.title=element_blank()) +
  geom_vline(xintercept=1925, linetype="dashed", colour = "red")
```

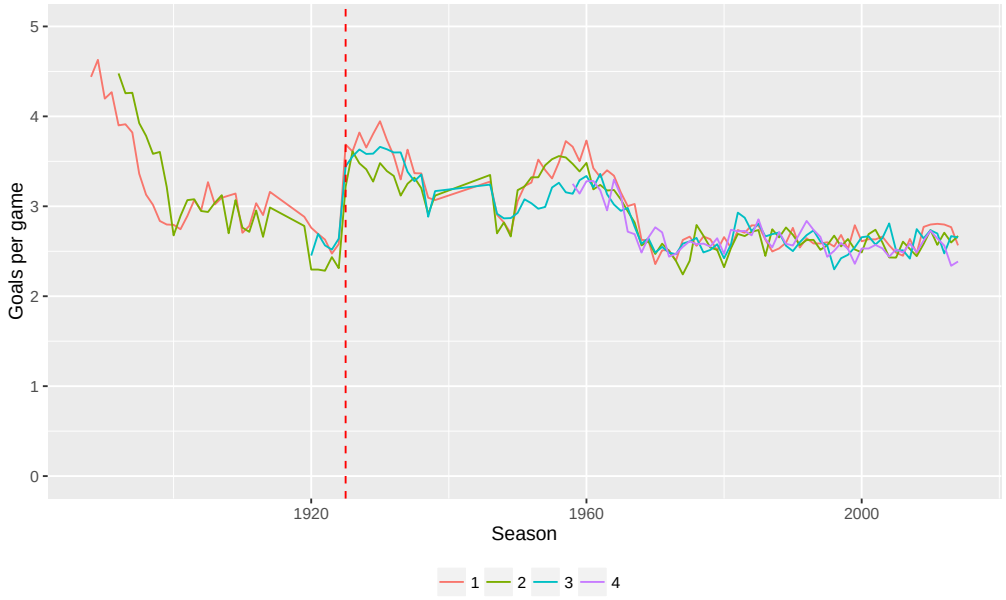

Fig. 5: The numbers of goals per game for the four tiers of the English league. The offside law was changed in 1925. The level has been fairly constant across all four tiers for the past 45 years.

Team rankings over time

Everyone has their favorite team and generally remembers the times the team were strongest. There is also a lot of interest in local rivalries: which team has done better over the years. The following code constructs league tables for every division for every season and gives teams an overall ranking for each season. The top teams come from the first tier, the next ones from the second tier, and so on. For the period when there were two third tiers, the Third Divisions North and South, the tables for the two have been interwoven and the rankings calculated accordingly.

```
newX <- allboth %>%
  group_by(Season, team) %>%
  mutate(gameno = dense_rank(Date)) %>%
  arrange(Season, team, gameno) %>%
  mutate(CumGF = cumsum(GF), CumGA = cumsum(GA), CumGD = cumsum(GD))  %>%
  select(Season, tier, division, team, gameno, result, CumGF, CumGA, CumGD)
newY <- newX %>% group_by(Season, team) %>%
  mutate(pts = (result==1)*(Season < 1981)*2 + (result==1)*(Season > 1980)*3 +
  (result==0.5)*1, Cumpts = cumsum(pts), GoalAve=CumGF/CumGA)
#Sort and rank by Season, division, points and goals (average/ difference)
Z1 <- newY %>% ungroup() %>% filter(Season < 1976) %>%
  arrange(Season, division, gameno, desc(Cumpts), desc(GoalAve))
Z2 <- newY %>% ungroup() %>% filter(Season > 1975) %>%
  arrange(Season, division, gameno, desc(Cumpts), desc(CumGD))
Zall <- rbind(Z1, Z2)

Zall <- Zall %>% mutate(drank=ave(team, Season, division, gameno, FUN=seq_along)) %>%
  mutate(drank=as.numeric(drank))

#Rank all tiers together at end of seasons
aZ <- Zall %>% group_by(Season, division) %>% filter(gameno==max(gameno))
aZ <- aZ %>% ungroup() %>% arrange(Season, tier, desc(Cumpts), desc(CumGD)) %>%
  mutate(lrank=ave(team, Season, FUN=seq_along)) %>% mutate(lrank=as.numeric(lrank))
#Add a square root function to emphasise higher ranks more
aZ <- aZ %>% mutate(srank=lrank^0.5)

#Add NAs for years teams were not present
aZM <- complete(aZ, Season, team)
aZM[is.na(aZM)] <- NA

#Numbers of teams by division by Season
nTeams <- Zall %>% filter(gameno==1) %>% group_by(Season, tier) %>% summarise(nt=n()) %>%
  mutate(cnt=cumsum(nt), snt=cnt^0.5)
nD <- split(nTeams, nTeams$tier)
```

Until 1976 teams equal on points were ranked on the ratio of the number of goals they scored at home to the number they scored away. This was called, slightly misleadingly, goal average. From 1976 on, goal difference was used. The calculations reflect these rules (and the introduction of three points for a win in 1981).

Goal average was actually only introduced in 1894. For the first few seasons of the league's history there was no official way of ranking teams with equal points. These tables use goal average for those years. You would think it might have been an issue as in the very first season with 12 clubs, there were three pairs on equal points. What are the chances of that? Two teams finished at the foot of the table on equal points, Notts County and Stoke City. Notts County had the better goal average and Stoke City the better goal difference. It did not really matter, as both—and the two teams above them—had to apply for re-election.

Individual team histories

To plot a small group of teams first specify a subset containing just their data. The rankings of all teams are plotted in light grey in the background and the selected teams are plotted in colour in the foreground. (Specifying colours to match team colours and keep the teams distinct is easier for some groups of teams than for others.) The lowest level of each tier is marked with a black line.

```
aS <- aZM %>% filter(team %in% c("Portsmouth", "Southampton", "AFC Bournemouth"))
ggplot(aZM, aes(Season, lrank)) + geom_line(aes(group=team), alpha=0.05) +
  geom_line(data=aS, aes(Season, lrank, group=team, colour=team), size=1.5) +
  scale_colour_manual(values=c("violet", "royalblue", "red2")) +
  theme(legend.position="bottom", legend.title=element_blank(),
        axis.ticks = element_blank(), axis.text.y=element_blank()) +
  labs(x=NULL, y=NULL) + scale_y_reverse() + scale_x_continuous(expand=c(0.02, 0)) +
  geom_step(data=nD[[1]], aes(Season, cnt)) + geom_step(data=nD[[2]], aes(Season, cnt)) +
  geom_step(data=nD[[3]], aes(Season, cnt)) + geom_step(data=nD[[4]], aes(Season, cnt))
```

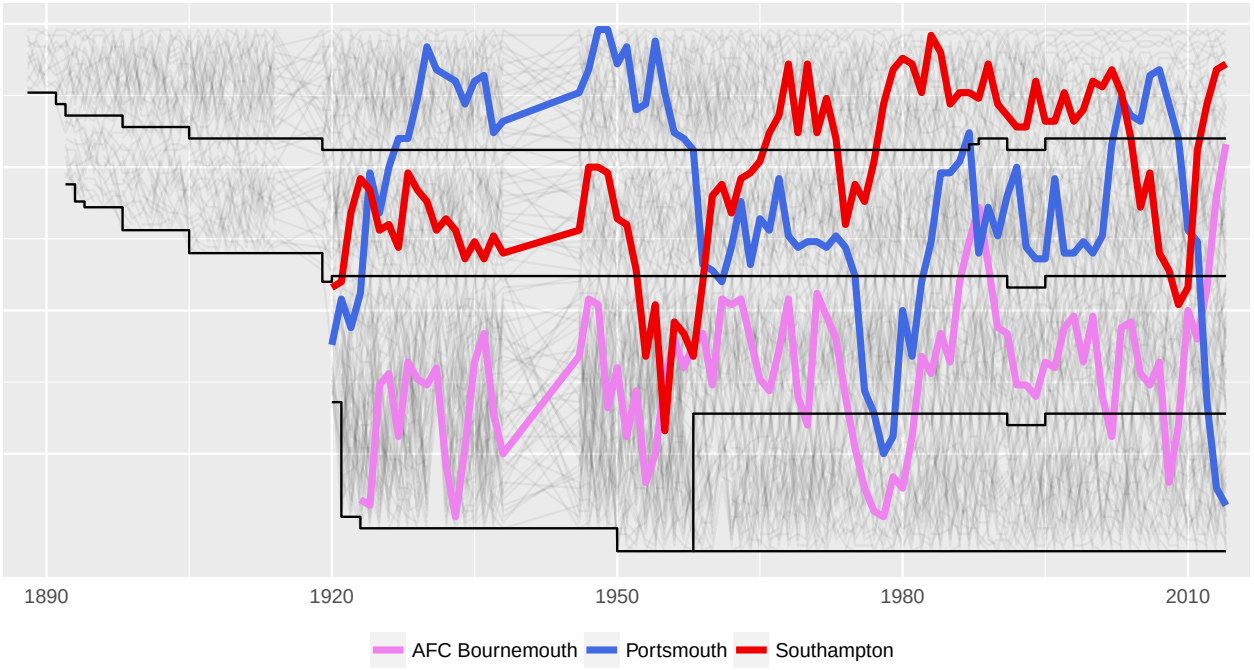

Fig. 6: A comparative plot of the three South Coast teams, Bournemouth, Portsmouth, and Southampton. Although Portsmouth was the most successful of the three for a long time (including two league titles), Southampton was mostly best from the 1960s on. They briefly dropped to the third tier in 2009 from which they bounced straight back, just as Portsmouth were going up and down in the opposite direction. Bournemouth have pretty well always been below the other two until the last couple of years.

Linear ranks mean that the difference between first and second at the very top gets the same weight as any difference between consecutive rankings. One possibility is to stretch the scale using a square root function and Figure 7 shows the same data as Figure 6 on this alternative scale.

```
ggplot(aZM, aes(Season, srank)) + geom_line(aes(group=team), alpha=0.05) +
  geom_line(data=aS, aes(Season, srank, group=team, colour=team), size=1.5) +
  scale_colour_manual(values=c("violet", "royalblue", "red2")) +
  theme(legend.position="bottom", legend.title=element_blank(),
        axis.ticks = element_blank(), axis.text.y=element_blank()) +
  labs(x=NULL, y=NULL) + scale_y_reverse() + scale_x_continuous(expand=c(0.02, 0)) +
  geom_step(data=nD[[1]], aes(Season, snt)) + geom_step(data=nD[[2]], aes(Season, snt)) +
  geom_step(data=nD[[3]], aes(Season, snt)) + geom_step(data=nD[[4]], aes(Season, snt))
```

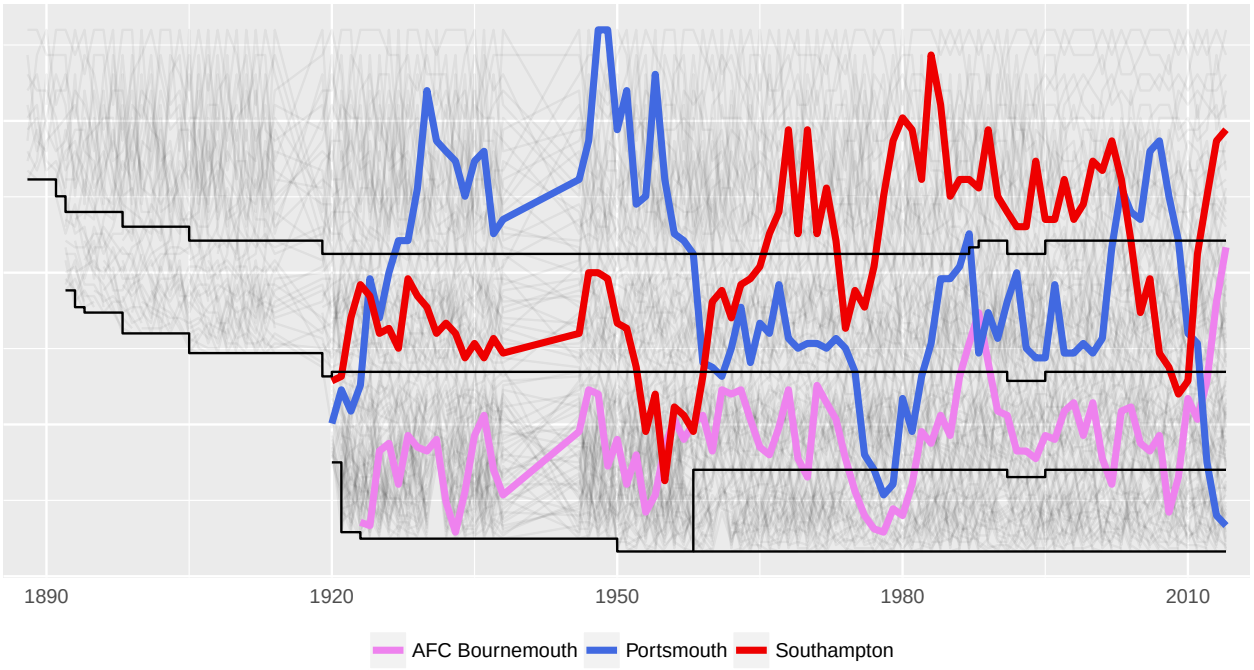

Fig. 7: The same display as in Figure 5, but using the square roots of the ranks to emphasise differences at the top. It is now clearer that Southampton did not ever win the league (they finished second in 1983).

Plots of historical trajectories of team performance can be found on various Wikipedia sites for individual teams.

Wormcharts for individual seasons and divisions

Wormcharts can show how a season develops. The numbers of points won by each team are plotted over the length of the season. For reasons of simplicity and consistency, the unit of time used is the number of games played and not the actual date. To make the display more readable it is not the actual numbers of points that are plotted, but the differences to the current mean across all teams. There are too many teams to colour them all individually without overcrowding the display, so the choice was made to highlight the top four teams and the bottom four teams. The division variable is now used instead of the tier variable to distinguish between the Third Divisions North ("3a") and South ("3b").

```
Seax <- 1991; divx <- 1; nsel <- 4
SeasonX <- Zall %>% filter(Season==Seax, division==divx)
SeasonX <- SeasonX %>% ungroup() %>% group_by(gameno) %>% mutate(meanP=mean(Cumpts))
nteams <- with(SeasonX, nlevels(factor(team)))
ngames <- 2*(nteams-1)
pal1 <- palette_pander(2*nsel)
ZY <- SeasonX %>% filter(gameno==ngames)
SelTeam <- ZY %>% filter(drank < (nsel+1) | drank > (max(SeasonX$gameno)/2-(nsel-1))) %>%
  select(team, drank)
Sy <- SeasonX %>% filter(team %in% SelTeam$team)
Sy <- within(Sy, Team <- reorder(team, -drank, last))
lS <- ZY %>% summarise(ul=ceiling(max(Cumpts-meanP)), ll=floor(min(Cumpts-meanP)))
ZY <- ZY %>% mutate(xx=gameno+2, yy=lS$ul-(drank-1)*(lS$ul-lS$ll)/(nteams-1))
ZYt <- ZY %>% filter(team %in% SelTeam$team)
ggplot(SeasonX, aes(gameno, (Cumpts-meanP))) + geom_line(aes(group=team), alpha=0.1) +
  theme(legend.position="none") + labs(y=NULL, x="Number of games") + xlim(0, ngames+12) +
  geom_line(data=Sy, aes(group=Team, colour=Team), size=1.5) +
  scale_colour_manual(values=pal1) + geom_text(data=ZY, aes(xx+7, yy, label=team)) +
  geom_segment(data=ZY, aes(x=gameno+1, xend=gameno+5, y=Cumpts-meanP, yend=yy,
  group=team), linetype=3) + geom_text(data=ZYt, aes(xx+7, yy, label=team, colour=team))
```

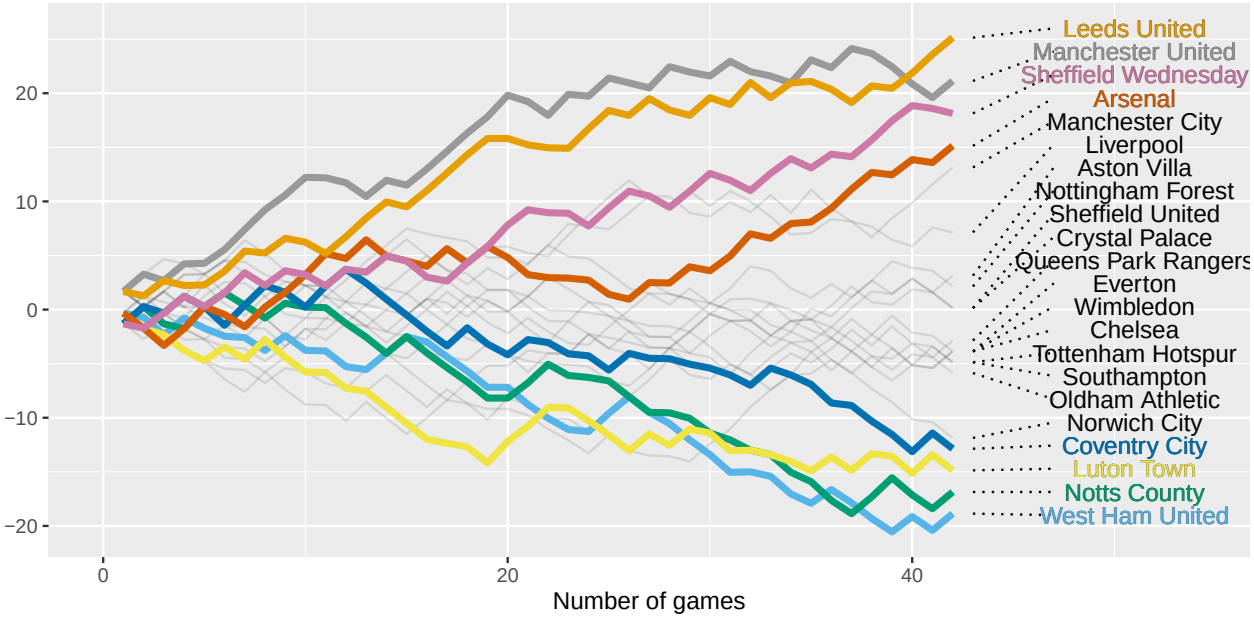

Fig. 8: A wormchart for Season 1991 of the English First Division, the final season before the Premier League started. Manchester United led for almost the whole of the season, but were caught right at the end and overtaken by Leeds United. Notts County and West Ham were both relegated, although neither was in a relegation position in the first half of the season.