

Sports Data Mining

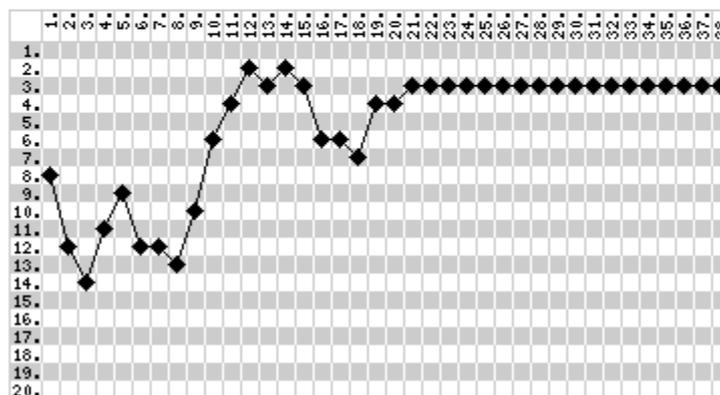
Tools and Systems for Sports Data Analysis

Inhalt

1. Überblick
2. Tools
 - a. Spezielle
 - b. Allgemeine
3. Anwendungsbeispiel

PLATZ	VEREIN	SPIELE	S	U	N	TORE	DIFF	PKT
2. (2.)	 RB Leipzig	38	24	7	7	65:34	+31	79
3. (3.)	 SV Darmstadt 98	38	21	9	8	58:29	+29	72
4. (4.)	 SV Wehen Wiesbaden	38	15	11	12	43:44	-1	56

SAISONVERLAUF DER 3. LIGA 2013/2014



Wieso spezielle Tools?

- Video Analysen
- Benutzbarkeit
- Vorschläge für profitable Visualisierungen von Statistiken

Anwendung

- Spieler/Teambewertung
 - hier auch Semi-Professionelle, daher sind die Tools oft relativ einfach gehalten
- Statistiken um Entscheidungen zu unterstützen
- Wettbüros
- Fans/Fantasy Sports
- Statistiken verkaufen an
 - Zeitungen
 - Spieleentwickler
 - Teams
 - Sportmoderatoren

Sports Data Mining Tools

- Häufig Video Auswertungen
- Grafische Aufbereitung von Statistiken

Advanced Scout (IBM)

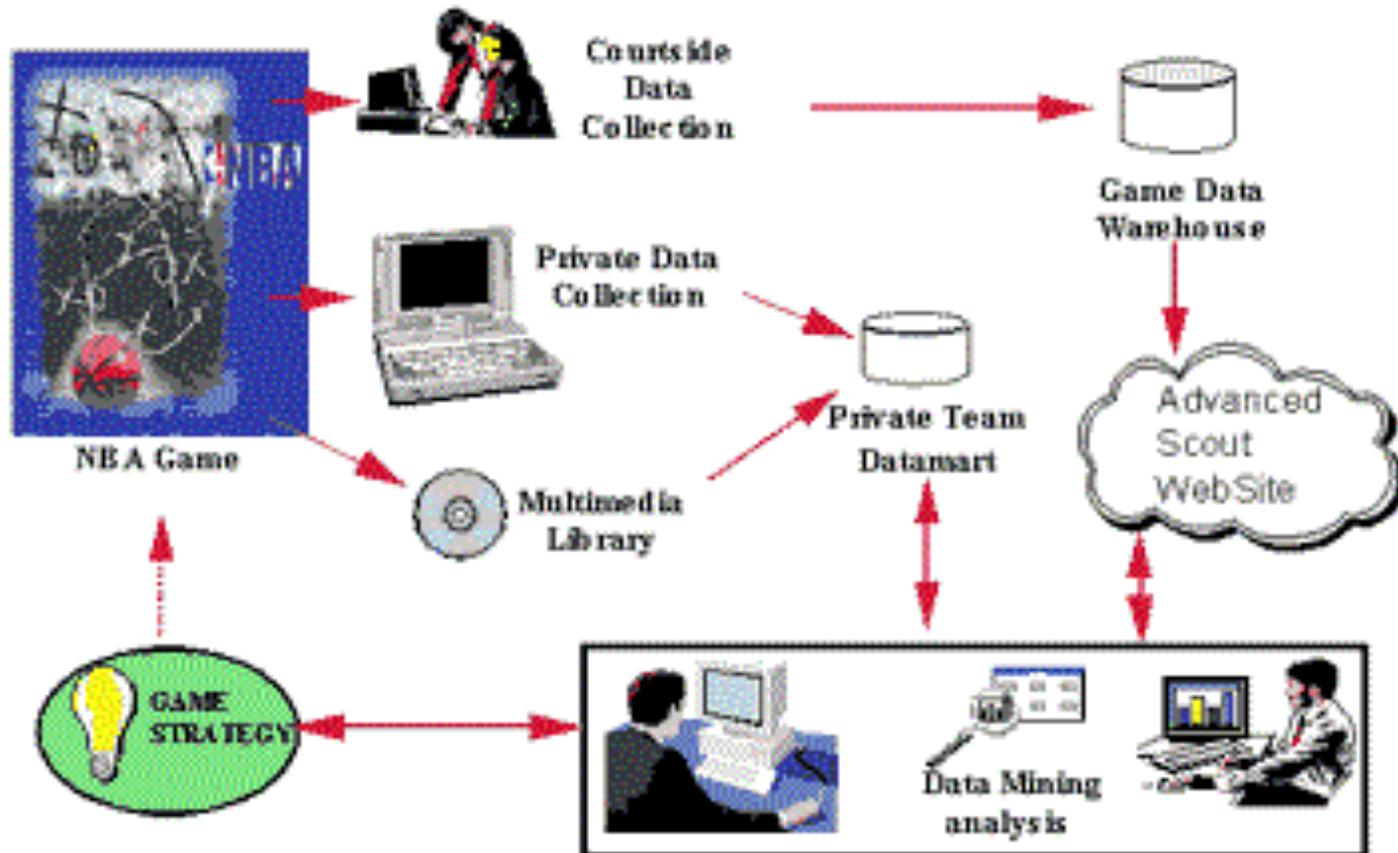
- Mitte der 90er
- Gegnereinschätzung anhand Historischer Daten
- Multimedia Part: finden wichtiger Spielsituationen
- Textuelle Beschreibung von Anomalien

[...]a customized software application that IBM created for NBA coaches. IBM Advanced Scout automatically mines massive amounts of player statistical data, condenses it into useful pieces of coaching information, and gives coaches insight into how their teams and opponents are performing. Coaches on 22 NBA teams use IBM Advanced Scout.

NEW YORK- 04 Nov 1999:

Zur Zeit scheint es 30 NBA Teams zu geben.

IBM Advanced Scout for NBA Coaches Information Flow



from IBM Webcast 4/26/2000

Synergy

- Live durchsuchbarer Index aus Statistiken und Video Material



Scouting Tools

- Aufzeichnen von Spielerleistungen

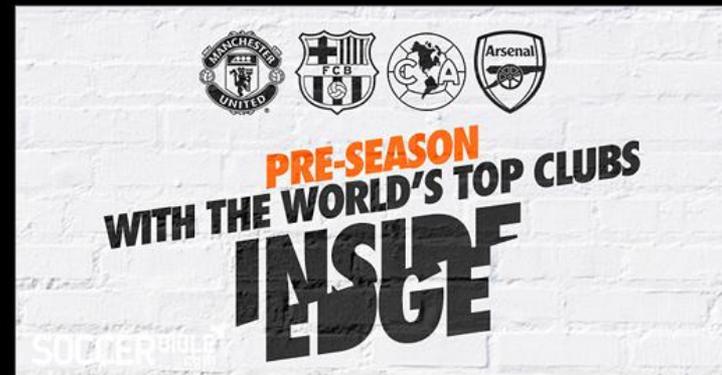
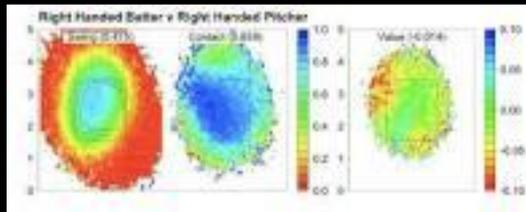
Digital Scout



Inside Edge



- “Inside Edge data, analysis tools and reports provide unique and enlightening facts to enhance any broadcast, print, video game or fantasy application.” [<http://inside-edge.com/>]



Sports Fraud Detection

- Finden von Unregelmäßigkeiten
- Hauptsächlich von Wettbüros eingesetzt

Allgemeine Tools

Weka

Weka 3.5.5 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None

Current relation: Relation: iris, Instances: 150, Attributes: 5

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> sepalength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Selected attribute: Name: sepalength, Type: Numeric, Missing: 0 (0%), Distinct: 35, Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

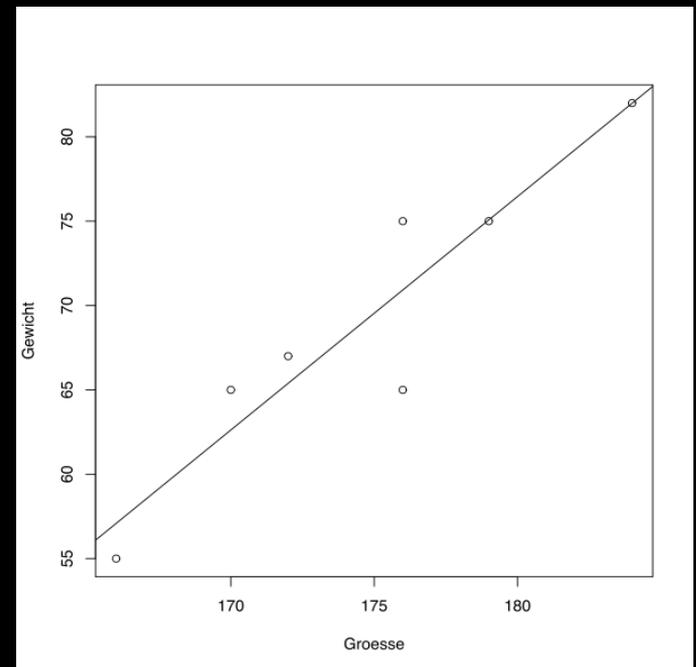
Class: class (Nom) Visualize

Region	Sub-Region	Count
Blue	-	16
Red	-	30
Red	Sub-Red 1	28
Red	Sub-Red 2	25
Cyan	-	34
Cyan	Sub-Cyan 1	10
Cyan	Sub-Cyan 2	7

R

- Programmiersprache für statistisches Rechnen und statistische Grafiken

```
# Punktdiagramm der Daten:  
plot(Gewicht~Groesse)  
# Regressionsgerade:  
abline(reg)
```



Wieso spezielle Tools? (2)

- Spezielle Visualisierungen bieten sich an
- Datenbasen werden häufig mitgeliefert
- Häufig nicht von Statistikern genutzt
- Mobilität

Fragen? Anmerkungen?

Anwendung

Datensatz 1

- 2004/05 - Alle Events der Portugisischen Liga
 - Event: Zeitpunkt, Spiel, Art des Events, Spieler
 - Auswechseln: Event 1: Spieler raus,
Event 2 Spieler rein
 - 18 Teams, 305 Spiele, 711 Tore und 1771 Karten

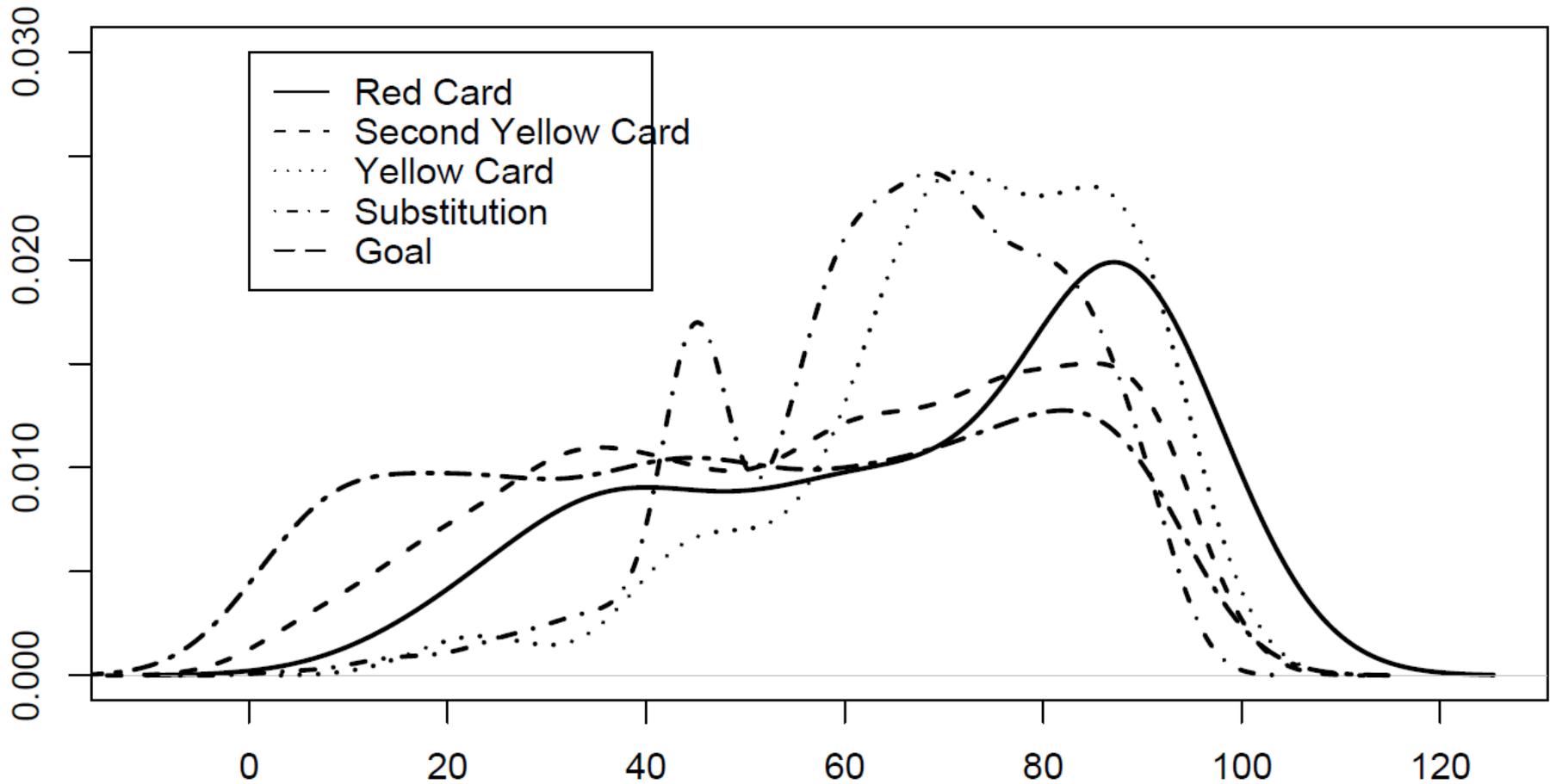
Table 1. Features available in the Portuguese Championship dataset.

Group	Related Features
Event	Related to each event: type, minute and half within the match.
Match	Related to the match being played: date, start time, score, TV channel transmitting, referee, number of spectators and total overtime granted.
Teams	Related to each team involved in the match: name, coach and current position, number of points, victories, defeats and draws in the championship.
Location	Related to the place where the match takes place: stadium and city.
Player	Related to the player involved the event: name, age, playing position, nationality, birth country, weight and height.

Table 2. Event types in the Portuguese Championship dataset.

Event Type	Description
Starter	Represents a starter player included in the initial lineup. For each match there are 22 events of this type, 11 for each team, occurring in the minute 0 of the match.
Substitute	Represents a substitute player for the match. For each match there are 14 events of this type, occurring in the minute 0 of the match.
In	Represents the exiting of a player during a substitution.
Out	Represents the entering of a player during a substitution.
Yellow	Represents the showing of an yellow card to a player.
Second Yellow	Represents the showing of the second yellow card to a player.
Red	Represents the showing of a direct red card to a player.
Goal	Represents the scoring of a standard goal.
Penalty	Represents the scoring of a penalty.
AutoGoal	Represents the scoring of an auto goal.

Dichte der Events



Datensatz 2

- Meisterschaften und Spiele verschiedener Europäischer Länder
 - Portugal - 15.382 Spiele seit 1934
 - England - 43.730 Spiele seit 1888
 - Spanien - 19.846 Spiele seit 1930
 - Italien - 17.680 Spiele seit 1946
 - Frankreich - 22.702 Spiele seit 1933
 - Deutschland - 13.406 Spiele seit 1963
 - Insgesamt 132.749 Spiele

Table 3. Features available in the European Championships dataset.

Feature

Visited and Visiting team's name.

For each match, the number of goals scored, the number of goals suffered and the winner.

Country's name, year and decade of the match.

For each team, the number of goals scored and suffered for each specific championship (total, in and out).

For each team, the number of points, victories, draws and defeats for each specific championship (total, in and out).

Portugal

Portugal - 3 dominierende Teams:
FC Porto, Benfica, Sporting

Analyse ergab:

- FC Porto meiste Schwankungen
- Benfica geringste Schwankungen
- Alle Teams haben hin und wieder außergewöhnlich schlechte Saisons
- Nur 2 Meisterschaften nicht von den 3 Teams gewonnen

Suche nach Assoziationen (Apriori)

Gefundene Regeln:

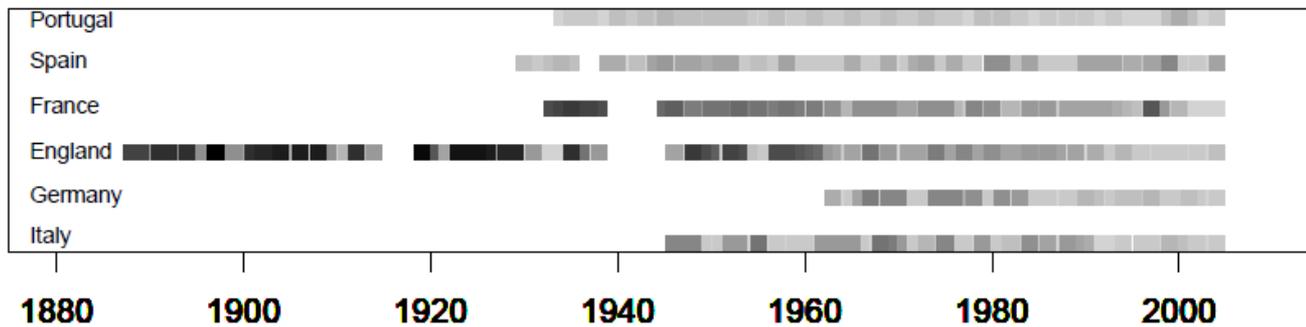
- Spiele die zwischen 15:30 und 16:30 beginnen sind Sonntags. (84% confidence)
- Spiele die nicht übertragen werden, sind Sonntags. (80% confidence)

Klassifikation (J48)

Klassifikation nach Sieg

- Datensatz 1: Klassifikation sehr einfach, da alle Events des Spiels bekannt (insbesondere die Tore)
- Datensatz 2: Nach Ländern aufgeteilt, Name beider Teams und Jahr (70% Trainingsset)
 - Portugal: (59.81%)
Wenn das Gastteam FC Porto Benefica oder Sporting ist -> defeat
 - true -> victory
 - Andere Länder: true -> victory

Visualisierung



1. Plätze in den jeweiligen Ländern, jeder Verein hat eine andere Graustufe.



Beobachtungen:

- Portugal hat eine geringe Varianz
- England hat die höchste Varianz und eine klare grenze in den 50ern
- 1. und 2. Weltkrieg sind leicht zu sehen

Spielergebnisse im Jahresmittel

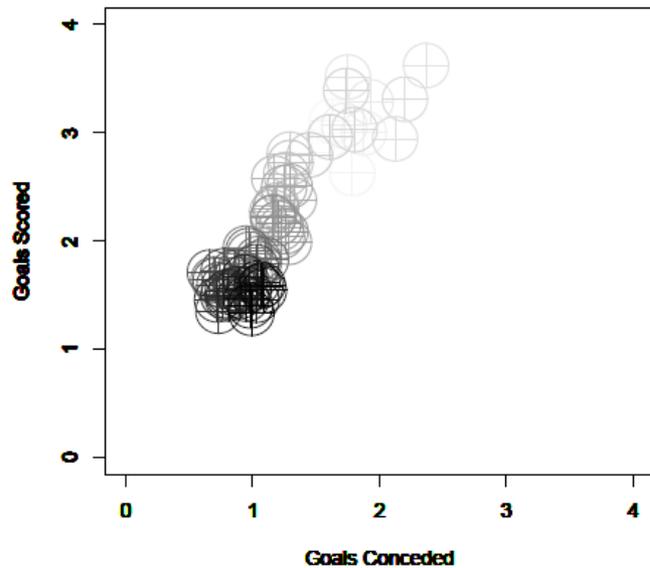


Fig. 6. Centroids for match results in Portugal (1934-2004).

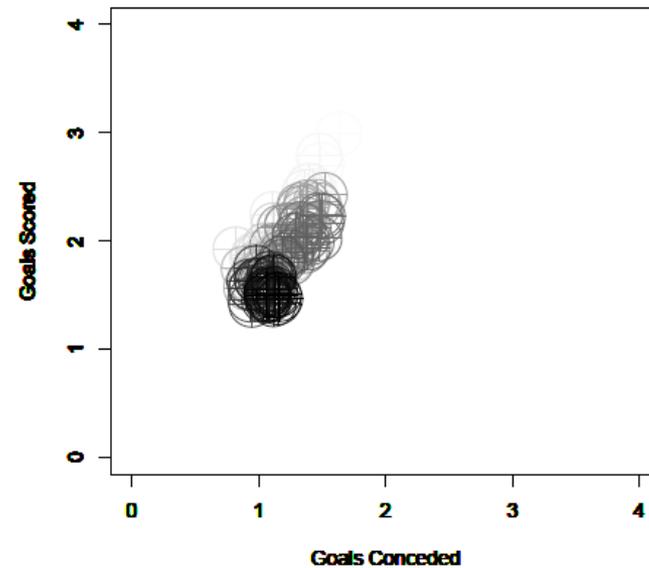
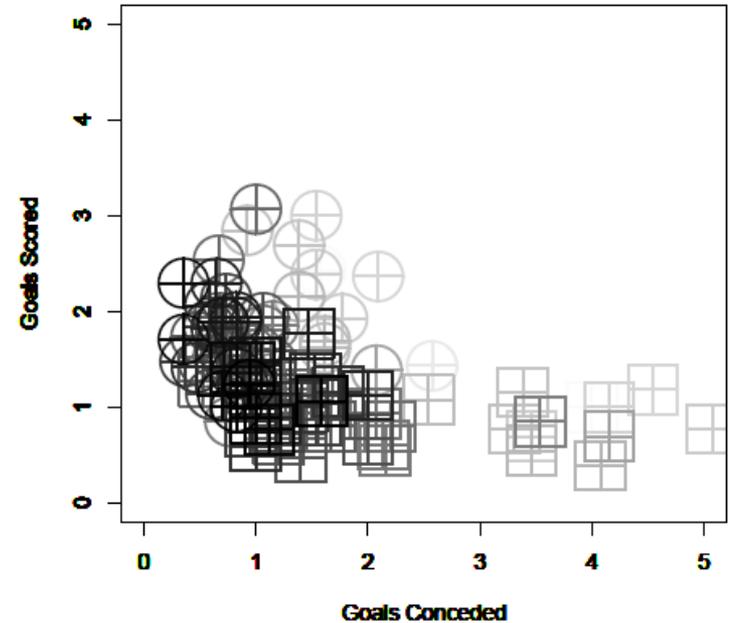
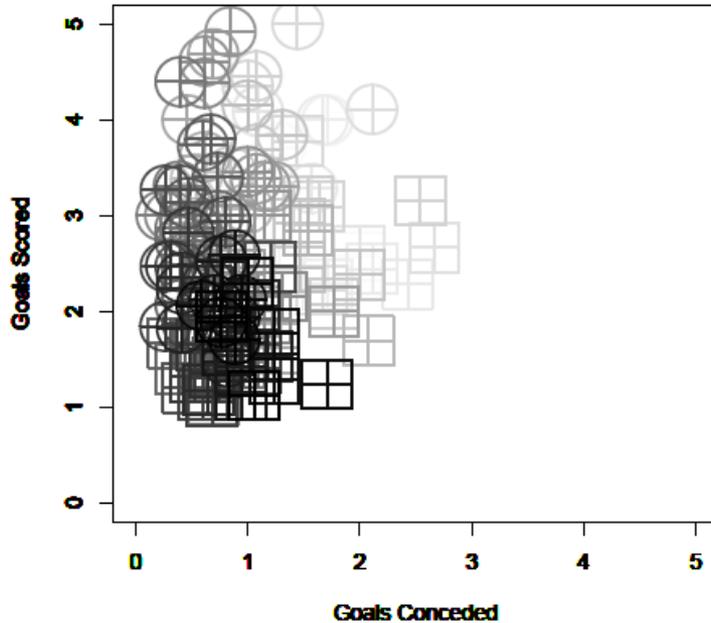


Fig. 7. Centroids for match results in England (1988-2004).

Torverhältnisse



Grauwert: Jahr (heller=älter)

Form: Ort (Rechteckig = Auswärtsspiel)

Spielergesultate

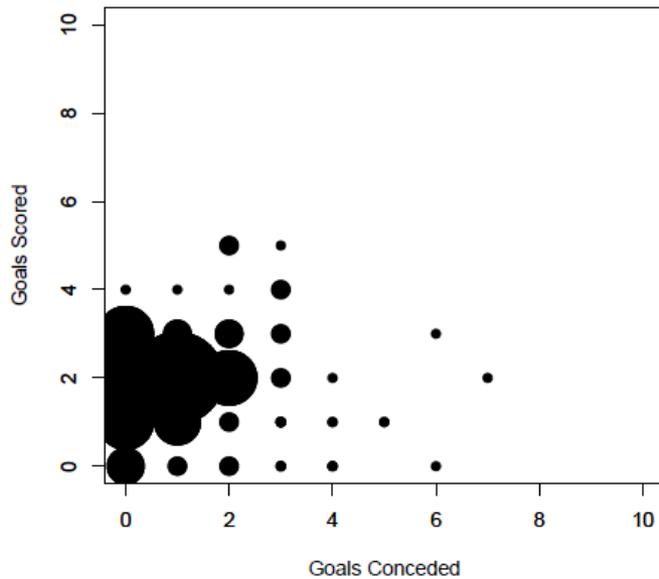


Fig. 10. FC Porto versus Benfica in the Portuguese Championship (1934-2004).

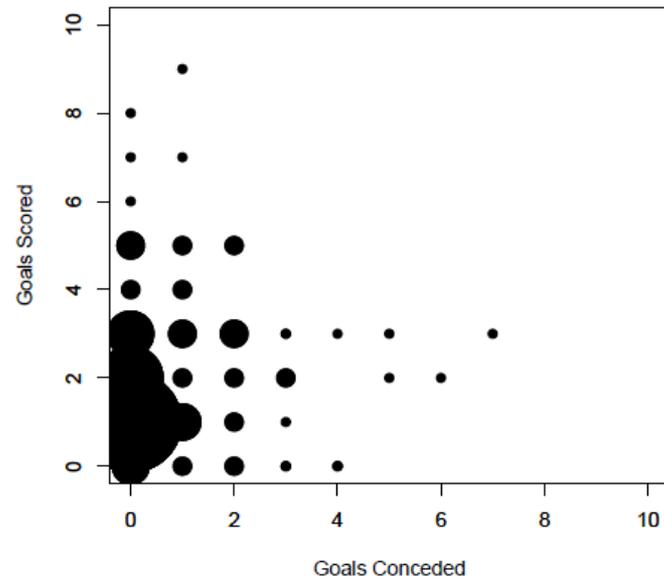


Fig. 11. FC Porto versus Belenenses in the Portuguese Championship (1934-2004).

Größere Punkte -> häufiger

Auswertung

- Mehr als 66% Zeitaufwand für Vorbereitung der Daten

Zusammenfassung

- Anwendungen und Nutzer sind vielfältig
 - Teams, Zeitungen, Fans,...
 - Fraud Detection, Scouting,...
- Visualisierungen sind häufig
Kernkomponente
- Häufig kann man simpel Tendenzen sichtbar machen

Sources

1. Sérgio Nunes, Marco Sousa - Applying Data Mining Techniques to Football Data from European Championships
2. Schuhmaker, Solieman, Chen - 2010, Springer - Sports Data Mining