Università
Ca'Foscari
Venezia
Facoltà
di Economia

Corso di Laurea in
Statistica e Informatica
per la Gestione dell'Impresa

Prova finale di Laurea

Gli *score* determinanti per la classifica del campionato italiano di pallavolo

Relatore
prof. Carlo Gaetan

Laureando
Paolo Favaretto
Matricola 810999

Anno Accademico
2008-2009

dedico questo lavoro a tutti coloro che,
attraverso piccoli ma importanti gesti,
mi hanno aiutato a raggiungere questo
importante traguardo.

INDICE

ELENCO DELLE TABELLE

ELENCO DELLE FIGURE

INTRODUZIONE E SOMMARIO

> Quando le regole della matematica si riferiscono alla realtà non sono certe e quando sono certe non si riferiscono alla realtà.
>
> *Albert Einstein*

La pallavolo ha subito importanti modifiche da alcuni anni a questa parte perché era uno sport poco televisivo visto che non aveva una durata fissa e una partita poteva durare da una a tre ore. Per questo motivo dall'anno 2000 la FIVB (Fédération Internazionale de Volleyball) ha deciso di eliminare definitivamente il cambio palla e di introdurre il *Rally Point System*. Questo nuovo sistema di gioco ha modificato diverse regole introducendo significative novità nel gioco che hanno determinato grandi modifiche nella tattica e nell'importanza dei diversi fondamentali che si sono adattati al nuovo sistema di gioco e quindi evoluti. Basta prendere in considerazione soltanto l'introduzione del libero[2, p. 102] e il fatto che se la palla tocca il nastro anche in battuta si continua a giocare per capire che adesso un giocatore di pallavolo deve essere meno completo nei fondamentali e più specializzato in quelli che svolge maggiormente per il proprio ruolo. Questa rivoluzione ha portato inoltre ad un significativo incremento della statura media dei giocatori, soprattutto dei centrali e degli opposti, che devono avere meno tecnica e braccia più lunghe e potenti perché alle loro carenze in difesa sopperisce il libero ed hanno un rendimento migliore a muro e in attacco grazie ai centimetri in più che li rendono meno dipendenti dalla loro capacità di salto e dal loro stato di forma.

In queste pagine si analizzano statisticamente i vari fondamentali della pallavolo ma-

schile nel nuovo sistema di gioco. I diversi fondamentali sono stati tutti scoutizzati, cioè valutati uno per uno secondo parametri prestabiliti, per poter valutare le differenze che ci sono tra le diverse squadre che compongono il campionato italiano e cercare di stabilire statisticamente quali sono i fondamentali che determinano la classifica finale della regular season.

Inizialmente, dopo un'introduzione sul tipo di dati esaminati e sulle componenti principali, si cercheranno le variabili con la migliore correlazione con la posizione in classifica, prima dimostrando che non è possibile farlo esaminando gli anni singolarmente, poi guardando alle medie dei vari anni e verranno selezionate le variabili più significative. Nel terzo capitolo, con un modello di regressione logistica a probabilità proporzionale, si cercheranno di stimare i coefficienti del modello esaminato precedentemente e di un nuovo modello che permetta di descrivere la posizione in classifica nel miglior modo possibile. Nelle conclusioni verranno illustrati i risultati ottenuti cercando di chiarire le relazioni che si sono potute vedere attraverso l'analisi statistica.

Capitolo 1

Il ruolo delle statistiche nella pallavolo

> Le sole statistiche di cui ci possiamo fidare sono quelle che noi abbiamo falsificato.
>
> *Winston Churchill*

1.1 La pallavolo e le statistiche

La pallavolo è stata uno dei primi sport ad avvalersi dei computer e di software per la rilevazione delle statistiche per analizzare sia gli schemi degli avversari che i vari fondamentali di gioco, cioè le prestazioni dei singoli giocatori e della squadra nel suo insieme. Il primo ad introdurre questo sistema in Italia fu Julio Velasco alla fine degli anni 80 che permise prima a Modena di vincere quattro scudetti consecutivi e poi alla nazionale italiana (sempre allenata da Velasco) di vincere il primo di tre mondiali consecutivi[3]. Negli anni, grazie al costante miglioramento dei programmi che hanno permesso di immagazzinare un numero sempre maggiore di informazioni utili, l'importanza delle statistiche di gioco è cresciuta tanto da poter affermare che una squadra che non si avvale di questo strumento, almeno per preparare un incontro, parte molto svantaggiata. L'importanza era già cresciuta quando si è passati dal vecchio sistema di gioco al nuovo perché in questo ogni pallone vale un punto ed è quindi più decisivo. Anche squadre di serie B (sia B1

che B2), che non sono professioniste, oggi si avvalgono delle statistiche ed è impensabile che giochino una partita senza aver studiato gli avversari. Sempre a questo livello alcune squadre hanno cominciato ad avere scout-man propri, cioè persone che raccolgono i dati sulla partita durante il suo svolgimento, cosa assolutamente normale e indispensabile a livelli più alti.

Nel caso della nazionale italiana di pallavolo e di alcuni club di altissimo livello, si arriva perfino a rilevare i dati durante l'allenamento nella fase di gioco grazie ad un software creato proprio per gli allenamenti. Questo serve per valutare, attraverso i dati che fanno da conferma empirica, la condizione degli atleti e i loro progressi nel tempo.

Anche i giocatori in campo sono molto condizionati dalle statistiche soprattutto nei fondamentali del muro e del palleggio. Ogni giocatore di buon livello deve infatti, prima di affrontare un incontro, aver presente certe situazioni che sono state rilevate e deve compiere durante la partita scelte coerenti con le informazioni delle quali è in possesso.

La pallavolo quindi, oltre ad essere facilmente scoutizzabile facilmente rispetto ad altri sport, ha sviluppato un sistema di gioco che si basa molto sullo studio degli avversari e sulle statistiche come misura della performance e della condizione dei giocatori.

1.2 Fonti e dati

I dati usati per questa tesi sono disponibili *on-line* sul sito ufficiale della lega pallavolo serie A (`http://www.legavolley.it/Statistiche.asp`) e precisamente sono state raccolte tutte le statistiche delle squadre maschili, dall'anno 2001 fino al 2009, riguardanti le 26 partite di *regular season* che ogni *club* ha avuto nel singolo anno. Questi dati sono stati inseriti nel database da scout man della lega pallavolo serie A quindi, nonostante siano fatti da persone diverse, hanno avuto tutti più o meno metri di giudizio simili durante la stagione e imparziali perché non sono sottomessi alle logiche particolari di un singolo club. La federazione cerca infatti di rendere gli scout più omogenei possibili durante la formazione del personale e fornendo diversi input ai vari scout man durante la stagione qualora si riscontrino possibili problemi. Da questa analisi sono stati omessi i

dati relativi al campionato 1999/2000, anche se disponibili, perchè il campionato in quell'anno aveva ancora 12 squadre e perciò non è confrontabile con gli anni successivi dove sono presenti 14 formazioni.

I dati disponibili sul sito internet sono i seguenti (tra parentesi il nome assegnato nella tabella):

Statistiche sui punti:

- *set* giocati: il numero di *set* che una squadra ha disputato durante la *regular season* di quell'anno (`set.giocati`).
- punti totali: il numero di punti realizzati dal club durante la *regular season* (`punti.tot`).
- punti vincenti: il numero di cambi-palla realizzati, cioŕ il numero di volte che una squadra ha fatto punto su un'azione iniziata con la battuta degli avversari (`punti.vin`).
- punti *break point*: numero di punti realizzati durante il proprio turno di battuta (`punti.BP`).

Statistiche sulla battuta:

- battute totali: il numero totale di battute fatte durante la *regular season* (`battuta.tot`).
- battuta *ace*: il numero di punti diretti fatti da una squadra battendo. Si considera ace una palla che tocca direttamente il campo avversario oppure viene ricevuta ma non è possibile eseguire il secondo dei tre tocchi a disposizione da parte di un compagno (`Ace`).
- battute sbagliate: sono il numero di battute sbagliate da parte di quel club. Viene considerato errore quando la battuta non oltrepassa il nastro superiore della rete o quando, pur avendo oltrepassato il nastro, finisce fuori dal campo di gioco (`batt.err`).

- *ace* per *set*: il numero di *ace* divisi per il numero di *set* giocati durante la stagione (`ace.set`).

- efficenza in battuta: è l'efficienza in battuta e si calcola sottraendo al numero di *ace* il numero di errori e poi dividendo il risultato per il numero totale di battute (`batt.Effic.`).

Statistiche sulla ricezione:

- ricezioni totali: il numero totale di ricezioni fatte durante la *regular season* (`rice.tot`).

- errori in ricezione: il numero di punti subiti da una squadra ricevendo. Vengono considerati tali gli *ace* subiti e i palloni ricevuti che finiscono nel campo avversario (`rice.err`).

- ricezioni negative: sono il numero di ricezioni che non permettono di far attaccare tutti gli elementi della squadra che potrebbero invece farlo con una ricezione buona (`rice.neg`).

- ricezioni perfette: il numero di ricezioni perfette fatte da una squadra. Una ricezione viene considerata perfetta quando risponde a determinati criteri che sono la traiettoria che la palla prende grazie al gesto di ricezione e la posizione del campo dove la palla cadrebbe ipoteticamente. È il parametro di valutazione più soggettivo (`rice.prf`).

- ricezioni perfette in percentuale: è il numero di ricezioni perfette diviso per il numero totale di ricezioni (`rice.prf.`).

- efficienza in ricezione: l'efficienza nella ricezione viene calcolata sottraendo al numero di ricezioni perfette (ricezione prf) il numero di ricezioni errate (ricezione err) e dividendo il risultato per il numero totale di ricezioni (`rice.eff`).

Statistiche sull'attacco:

- attacchi totali: il numero totale di attacchi fatti durante la *regular season* (`att.tot`).

- errori in attacco: il numero di attacchi che non superano il nastro superiore della rete oppure finiscono fuori campo (`att.err`).

- attacchi murati: sono il numero di attacchi che vengono murati dagli avversari (`att.murati`).

- attacchi perfetti: il numero di attacchi che permettono alla squadra di fare punto perché o cadono dentro al campo avversario, o vengono difesi ma gli avversari non riesco a mantenere il pallone in gioco oppure il pallone tocca il muro avversario e cade fuori campo (`att.prf`).

- attacchi perfetti in percentuale: è il numero di attacchi perfetti diviso per il numero totale di attacchi (`att.prf.`).

- efficienza in attacco: l'efficienza nell'attacco viene calcolata sottraendo al numero di attacchi perfetti, il numero di attacchi sbagliati e dividendo il risultato per il numero totale di attacchi (`att.eff`).

Statistiche sul muro:

- muro invasioni: il numero di volte che un qualsiasi giocatore della squadra tocca la rete mentre la palla è in gioco e viene sanzionato dall'arbitro per questo motivo (`muro.inv`).

- muri perfetti: il numero di muri che permettono a una squadra di ottenere un punto perché la palla cade nel campo avversario o gli avversari non riesco a mantenere il pallone in gioco (`muro.prf`).

- Muri perfetti per *set*: numero di muri punto divisi per il numero di set che una squadra ha giocato durante la stagione regolare (`muro.set`).

Nella Tabella 1.1 viene riportata una parte del dataset, precisamente le prime otto variabili e le prime quarantadue colonne, come esempio per vedere come sono stati aggregati i dati raccolti.

	X	set.giocati	punti.tot	punti.vin	punti.BP	battuta.tot	Ace	batt.err
1	1	103	1702	1096	606	2339	116	412
2	2	97	1663	1037	626	2292	153	290
3	3	94	1614	1039	575	2208	121	356
4	4	98	1650	1091	559	2248	142	443
5	5	99	1679	1078	601	2285	149	417
6	6	99	1633	1092	541	2266	110	392
7	7	96	1586	1045	541	2155	98	325
8	8	107	1768	1151	617	2406	132	384
9	9	109	1713	1135	578	2341	149	411
10	10	98	1529	1015	514	2138	131	398
11	11	101	1494	1003	491	2156	112	339
12	12	104	1609	1098	511	2259	109	356
13	13	97	1509	1077	432	2107	91	336
14	14	98	1556	1053	503	2160	153	429
15	1	100	1672	1055	617	2330	142	320
16	2	95	1582	995	587	2210	138	408
17	3	101	1629	1007	622	2298	152	380
18	4	107	1683	1138	545	2372	101	394
19	5	103	1647	1095	552	2308	104	427
20	6	105	1686	1111	575	2363	139	372
21	7	105	1701	1130	571	2341	141	393
22	8	111	1786	1184	602	2500	152	427
23	9	106	1641	1092	549	2272	137	403
24	10	101	1524	1020	504	2196	98	368
25	11	106	1658	1138	520	2342	130	461
26	12	104	1593	1075	518	2290	115	383
27	13	102	1540	1056	484	2207	127	414
28	14	94	1429	1003	426	2014	114	448

Tabella 1.1: Esempio del dataset con le prime 28 righe e 8 variabili

1.3 Perché analizzare i dati?

In questi ultimi anni, per misurare le prestazioni degli atleti, la prima cosa cui si guarda è generalmente l'efficienza nei vari fondamentali di gioco e in base a questa si determina la prestazione di un giocatore. L'efficienza in senso lato è la differenza tra punti ottenuti e qualunque errore fatto sul totale delle volte che quel fondamentale è stato eseguito. Il dato può essere letto in modo relativo o assoluto: relativo se confrontato con gli standard del giocatore (per esempio un atleta che normalmente attacca con il 50% in attacco e in una partita fa segnare un 60% in attacco ha fatto una buona prestazione perché superiore alla sua media), assoluto se confrontato con un modello di riferimento definito di solito dalle varie medie (per esempio normalmente in serie A si attacca con il 60% di efficienza quindi l'atleta di prima avrebbe fatto una partita assolutamente normale). L'efficienza è usata anche per definire i parametri della squadra tuttavia non spiega la classifica finale. Infatti, se osserviamo i grafici dell'efficienza nei vari fondamentali scopriamo che la prima in classifica solitamente non è mai la più efficiente in nessun fondamentale. Se per esempio consideriamo il 2009, l'ultima stagione regolare, vediamo che la squadra arrivata seconda presenta tutte efficienze superiori o uguali alla prima come illustrato nei grafici 1.1, 1.2, 1.3 e 1.4. Nemmeno una particolare combinazione che dia dei pesi diversi alle varie efficienze può essere allora presa in considerazione.

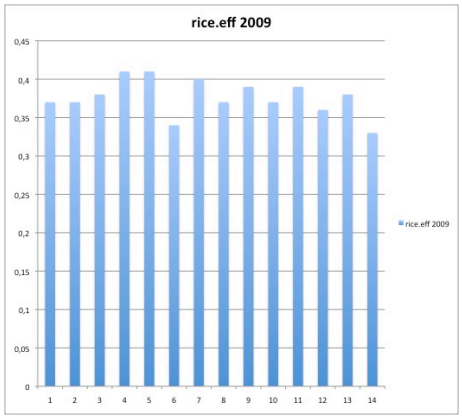

Figura 1.1: Grafico dell'efficienza in ricezione

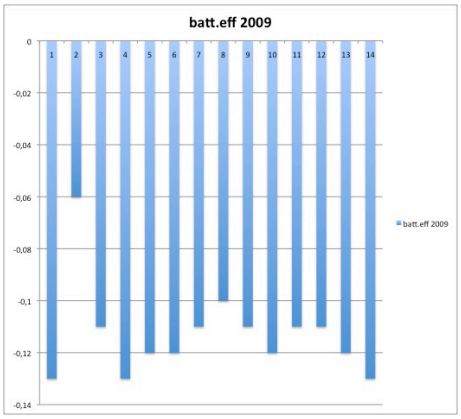

Figura 1.2: Grafico dell'efficienza in battuta

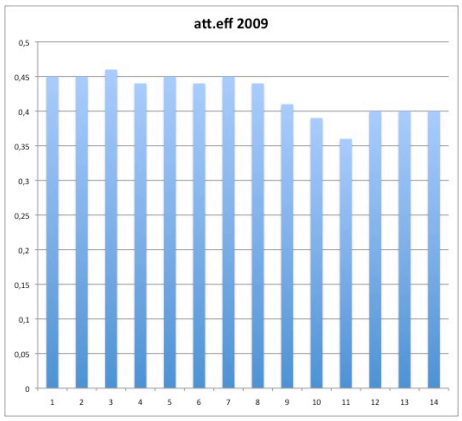

Figura 1.3: Grafico dell'efficienza in attacco

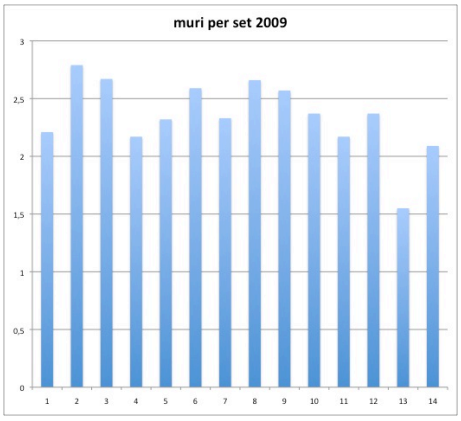

Figura 1.4: Grafico dei muri fatti sul numero di set giocati

Nel grafico 1.2 si deve ricordare che i valori, essendo negativi, sono maggiori quanto più vicini allo zero quindi l'efficienza in battuta della seconda classificata è molto superiore a quella della prima classificata. Non potendo essere l'efficienza la chiave di lettura giusta, si è cercato di vedere se la cosa più importante fosse la fase dove una squadra batte cioè la fase di BP (*break point*) figura 1.5. Anche in questo caso si nota che la seconda squadra in classifica supera la prima avendo ottenuto un maggior numero di punti sulla propria battuta.

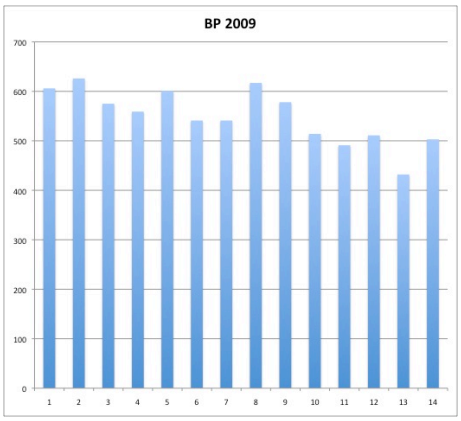

Figura 1.5: Grafico della variabile punti.BP

Dando uno sguardo più generale a tutti gli anni rilevati si delinea chiaramente il fatto che non ci sia un fondamentale che prevalga sugli altri. La tabella 1.6 riporta la posizione in classifica della squadra che ha il miglior fondamentale per ogni anno. Ad esempio

nella colonna 2009 la squadra con la migliore efficienza in attacco è arrivata terza. Nel grafico sono evidenziate tutti i fondamentali in cui le squadre prime classificate sono state le migliori.

Si può chiaramente distinguere che nell'anno 2001 ha vinto la squadra che ha avuto la miglior battuta e un buon muro che probabilmente le ha permesso di essere anche la migliore nel fare punti su battuta (`punti.BP`), mentre negli anni 2002, 2003, 2005 ha vinto la squadra che aveva l'attacco più forte. Nel 2006 si è classificata al primo posto la squadra che ha avuto la miglior ricezione e il miglior attacco quindi possiamo supporre con il miglior cambio palla visto. Nel 2008 ha vinto di nuovo la squadra con la miglior battuta e muro mentre nel 2009 e 2007 la squadra che è arrivata prima nella stagione regolare non è stata eccelsa, rispetto alle altre, in nessun fondamentale (la variabile att.prf infatti indica che ha fatto tanti punti ma non mi dice se perché ha giocato tanti *set*, se i *set* sono finiti con punteggi alti o se ha sbagliato anche tanti attacchi quindi non è significativa). L'anno 2004 merita una considerazione a parte perché ha vinto la squadra con la miglior fase punto su battuta ma non si può capirne il motivo. Forse è dipeso dalla difesa ma non essendo stata rilevata non è desumibile dai dati .

Alla luce di quanto detto, non è dunque così semplice trovare quali siano i fondamentali che permettono di delineare la posizione in classifica finale guardando solo alle statistiche normalmente usate per definire la bontà di una prestazione di squadra. Bisogna quindi studiare in maniera più approfondita le relazioni esistenti tra le variabili.

	nome var	2009	2008	2007	2006	2005	2004	2003	2002	2001
1	X									
2	set.giocati	9	8	7	10	9	13	6	4	5
3	punti.tot	8	8	7	4	3	11	6	4	5
4	punti.vin	8	8	9	4	9	13	6	4	5
5	punti.BP	2	3	7	4	6	1	4	7	1
6	battuta.tot	8	8	7	4	8	13	6	4	5
7	Ace	2	3	5	3	6	11	7	10	1
8	batt.err	4	11	5	4	6	6	6	11	1
9	ace.set	2	3	5	3	6	11	7	10	1
10	batt.eff	2	1	13	11	4	11	4	13	4
11	rice.tot	9	11	7	12	9	13	12	9	9
12	rice.err	14	9	11	14	13	9	10	14	11
13	rice.neg	12	8	9	13	9	11	10	7	9
14	rice.prf	9	11	7	12	7	6	6	8	5
15	rice,prf.	7	10	6	1	7	10	6	2	5
16	rice.eff	4	10	6	1	7	6	6	2	5
17	att.tot	13	8	7	10	10	13	6	8	9
18	att.err	11	12	7	12	9	14	10	11	9
19	att.murati	13	13	13	12	14	11	7	12	11
20	att.prf	1	8	7	4	9	13	6	4	5
21	att.prf.	3	2	3	1	1	3	1	1	3
22	att.eff	3	2	3	1	1	3	1	1	3
23	muri.inv	4	12	10	7	6	7	3	6	1
24	muri.prf	8	1	4	2	3	4	10	7	7
25	muri.set	2	1	4	2	3	4	1	7	7

Figura 1.6: Tabella con il rank delle squadre che hanno avuto i fondamentali migliori

Capitolo 2

UNA ANALISI DELLE COMPONENTI PRINCIPALI PER DETERMINARE I FONDAMENTALI PIÙ IMPORTANTI

> Le statistiche sono come i bikini. Ciò che rivelano è suggestivo, ma ciò che nascondono è più importante.
>
> *Aaron Levenstein*

2.1 Cosa sono le componenti principali

L'idea che sta alla base delle componenti principali è quella di sintetizzare le variabili scelte per analizzare dei dati e poterle rappresentare graficamente, in un piano cartesiano, per capire come e quanta variabilità nei dati spiegano senza però che questo provochi una consistente perdita di informazioni. Infatti, quando abbiamo p variabili, dovremmo rappresentare i dati in un grafico p-dimensionale ma quando p è maggiore di tre, oltre che essere molto difficile sia da rappresentare computazionalmente che da interpretare, è impossibile perché l'uomo riesce a percepire solo tre dimensioni dello spazio. L'analisi delle componenti principali è particolarmente utile quando un certo aspetto non è direttamente quantificabile, ma si dispone di più indicatori del medesimo [1, p. 217] perché sintetizza le p variabili del nostro modello, tra loro correlate, con delle nuove p variabili che posseggono queste proprietà:

1. sono incorrelate tra loro

2. hanno varianza decrescente

Queste variabili sono appunto dette componenti principali. La prima componente principale è la combinazione lineare delle p variabili di partenza con varianza massima mentre la seconda componente principale è la combinazione lineare delle p variabili di partenza con varianza massima dopo la prima e ortogonale rispetto alla prima, e così via per tutte le altre.

Per ottenere la prima componente principale Y_1 dobbiamo risolvere questa equazione:

$$\underline{y}_1 = \mathbf{X}\underline{a}_1 \tag{2.1}$$

con:

- $\mathbf{X}$ matrice dei dati con n righe e p colonne
- $\underline{a}_1 = [a_{11}, \ldots, a_{1p}]'$.

$\underline{y}_1$ è per definizione la combinazione lineare di massima varianza e dato che la varianza totale di una trasformazione lineare di $\mathbf{X}$ è esprimibile in funzione della matrice di covarianza $\mathbf{S}$:

$$Var(\mathbf{X}\underline{a}_1) = \underline{a}_1'\mathbf{S}\underline{a}_1. \tag{2.2}$$

Il vettore dei coefficienti $\underline{a}_1$ dovrà essere tale da massimizzare l'espressione precedente [1, p. 218] tenendo conto del vincolo di normalizzazione $\underline{a}_1'\underline{a}_1 = 1$, quindi:

$$\frac{\partial[\underline{a}_1'\mathbf{S}\underline{a}_1 - \lambda(\underline{a}_1'\underline{a}_1 - 1)]}{\partial \underline{a}_1} = 2\mathbf{S}\underline{a}_1 - 2\lambda\underline{a}_1 = 2(\mathbf{S} - \lambda\mathbf{I})\underline{a}_1 = 0 \tag{2.3}$$

dove $\mathbf{I}$ è la matrice identità di dimensione p per p. Il sistema lineare di p equazioni e p incognite ammette soluzioni diverse da zero solo nel caso che il determinante di $\mathbf{S} - \lambda\mathbf{I}$ sia diverso da zero. Da qui si ricava il polinomio caratteristico di ordine p e con p soluzioni,

cioè gli autovalori, non negativi per definizione. Per massimizzare la varianza bisogna prendere il più grande degli autovalori perché:

$$Var(\mathbf{X}a_1) = \underline{a}_1'\mathbf{S}\underline{a}_1 = \underline{a}_1'\lambda_1\underline{a}_1 = \lambda_1. \tag{2.4}$$

La seconda componente principale si calcola procedendo allo stesso modo ma ricordando che la nuova componente principale deve essere ortogonale alla precedente e risulta essere λ_2, cioè il massimo autovalore della matrice covarianza $\mathbf{S}$ dopo λ_1. In generale, si definisce v-esima componente principale di p variabili la combinazione lineare:

$$\underline{y}_v = \mathbf{X}\underline{a}_v \qquad \text{per} \qquad v = 1, ..., k \leq p \tag{2.5}$$

dove a_v è l'autovettore associato al v-esimo autovalore λ_v, in ordine decrescente, della matrice varianza. Tuttavia bisogna tenere in considerazione che per confrontare le diverse variabili è opportuno che esse siano espresse dalla medesima unità di misura perché l'analisi delle componenti principali guarda alla varianza totale e semplicemente cambiare l'unità di misura comporterebbe un cambiamento della varianza totale. Per questo i dati vengono solitamente standardizzati e questo fa si che, dati molto diversi, siano confrontabili perchè aventi la stessa scala di misura.

L'interesse operativo dell'analisi delle componenti principali si manifesta nel caso in cui poche componenti (le prime k) sono in grado di spiegare una percentuale elevata della varianza totale perché in questo caso possiamo sostituire le variabili di partenza con le componenti principali ottenendo una perdita d'informazione limitata nella speranza di migliorare l'interpretabilità dei grafici.

2.2 Un'analisi esplorativa dei dati

Vista l'incapacità di spiegare adeguatamente la posizione in classifica guardando semplicemente i dati si è deciso di affrontare un'analisi più approfondita. Il modo più semplice e intuitivo per riassumere la dipendenza è la correlazione tra le variabili e il *rank*. Subito

però si pongono una serie di problemi. Sono sempre le stesse variabili ad essere incorrelate con la posizione in classifica? E se cambiano nel tempo si può capire quali sono quelle realmente influenti nel determinare la posizione e quelle che magari lo sono solo casualmente? Per questo si è scelto di eliminare le variabili che, anche se avessero avuto una buona correlazione, non avrebbero fornito informazioni utili. Queste variabili sono: il numero di *set* giocati e il numero di punti fatti (perché si può arrivare primi vincendo tutti i *set* con uno scarto di 2 punti o con scarto di tanti punti, e si può perdere poche partite ma 3 a 0 o vincerne tante 3 a 2), il numero totale di battute, il numero totale di ricezioni e il numero totale di attacchi. Disegnando i grafici di dispersione delle singole variabili con il *rank* per ogni anno e calcolandone la correlazione, si nota che alcune variabili sono significative in certi anni ma in altri no. Avendo bisogno di un criterio più generale si è deciso di raggruppare tutte le variabili in base alla posizione in classifica finale delle squadre, in modo da avere una tabella con tutte le variabili delle prime classificate, una tabella con le variabili delle seconde classificate e così via fino alla quattordicesima. Questo procedimento è stato applicato agli anni dal 2001 al 2007 perché i successivi anni, 2008 e 2009, verranno usati successivamente. Poi sono state fatte le medie di tutte le variabili per ogni nuova tabella in modo tale da avere un anno ideale che contenga i dati di tutte le squadre arrivate prime, seconde, terze e così via fino all' ultima. Infine tutti questi valori medi sono stati riuniti in una tabella (2.1) che rappresenta quindi un anno ideale contenente i valori medi che ogni squadra, in base alla posizione in classifica, ha avuto nei diversi fondamentali.

	set giocati	punti totali	punti vincenti	punti *break point*	battute totali	Ace
1	97.29	1669.57	1067.57	602.00	2291.71	127.29
2	98.43	1653.71	1077.00	576.71	2290.86	115.43
3	99.57	1687.14	1100.71	586.43	2305.71	129.29
4	101.57	1691.71	1100.57	591.14	2334.71	107.43
5	102.57	1677.71	1101.29	576.43	2340.00	124.14
6	102.71	1705.71	1126.29	579.43	2351.00	134.86
7	103.57	1686.14	1119.57	566.57	2331.14	122.57
8	102.57	1671.57	1127.29	544.29	2317.43	111.57
9	103.86	1683.57	1146.86	536.71	2304.57	110.57
10	103.86	1656.43	1120.14	536.29	2317.43	112.14
11	102.00	1601.14	1088.14	513.00	2252.86	114.43
12	102.29	1576.71	1087.14	489.57	2227.14	114.29
13	101.71	1566.71	1089.57	477.14	2223.00	91.71
14	95.71	1412.57	1007.71	404.86	2022.71	83.14

Tabella 2.1: estratto della tabella con le medie di ogni fondamentale rispetto alla posizione in classifica

Per ogni fondamentale è stato poi rappresentato il diagramma di dispersione dove in ascissa troviamo la posizione in classifica mentre in ordinata abbiamo i valori medi del fondamentale rappresentato, inoltre viene riportata la correlazione.

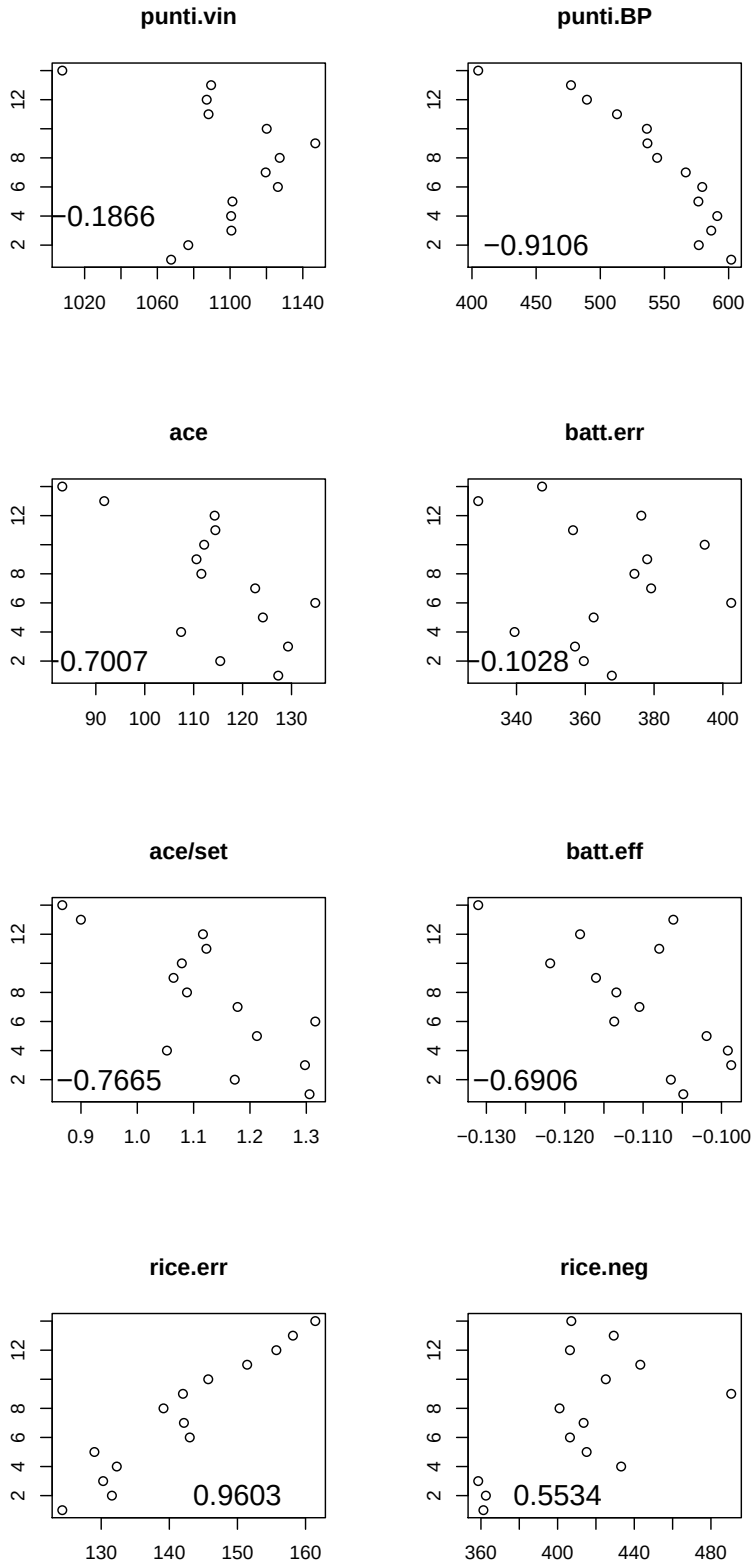

Figura 2.1: Grafico di dispersione con le correlazioni delle prime otto variabili dell'anno delle medie

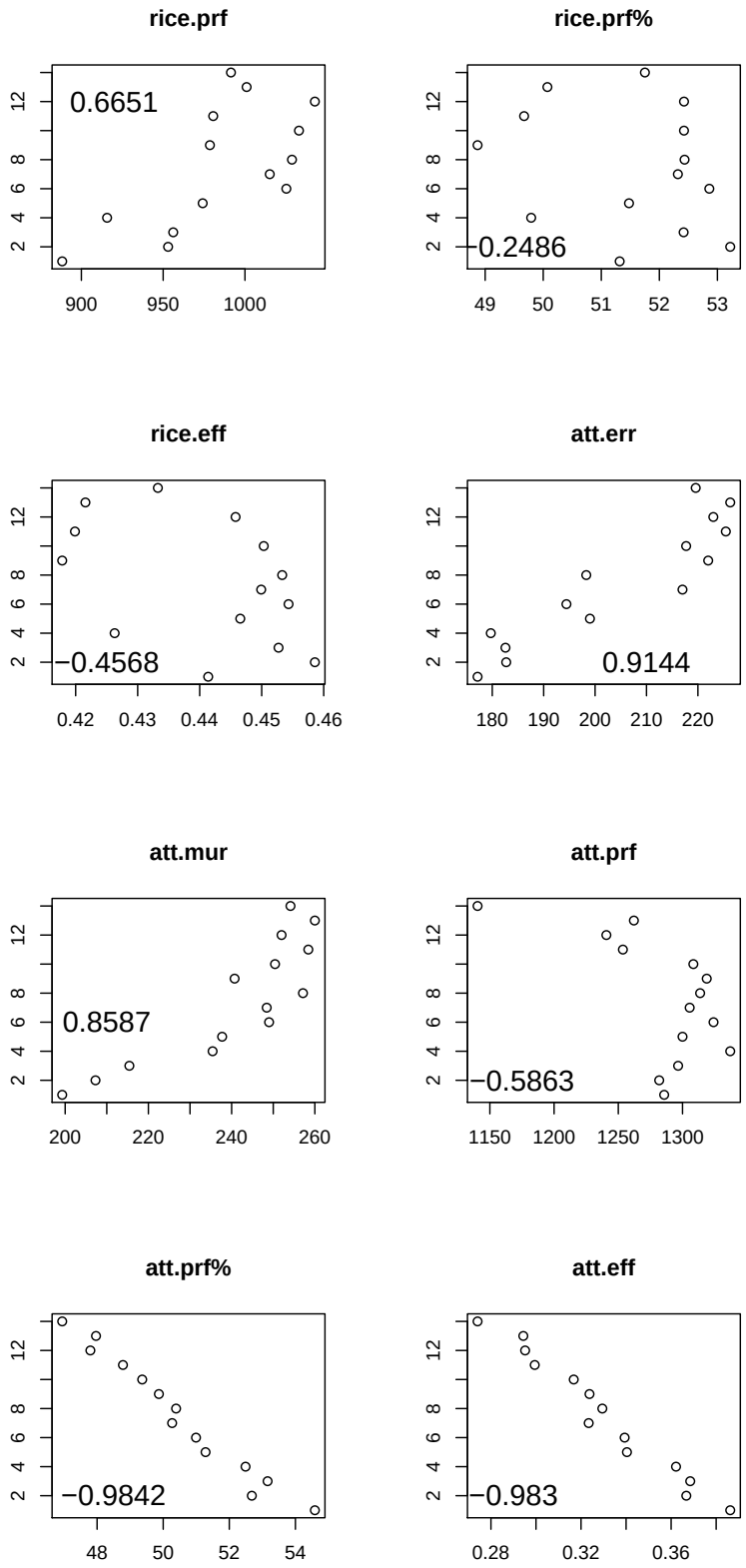

Figura 2.2: Grafico di dispersione con le correlazioni dalla nona alla sedicesima variabile dell'anno delle medie

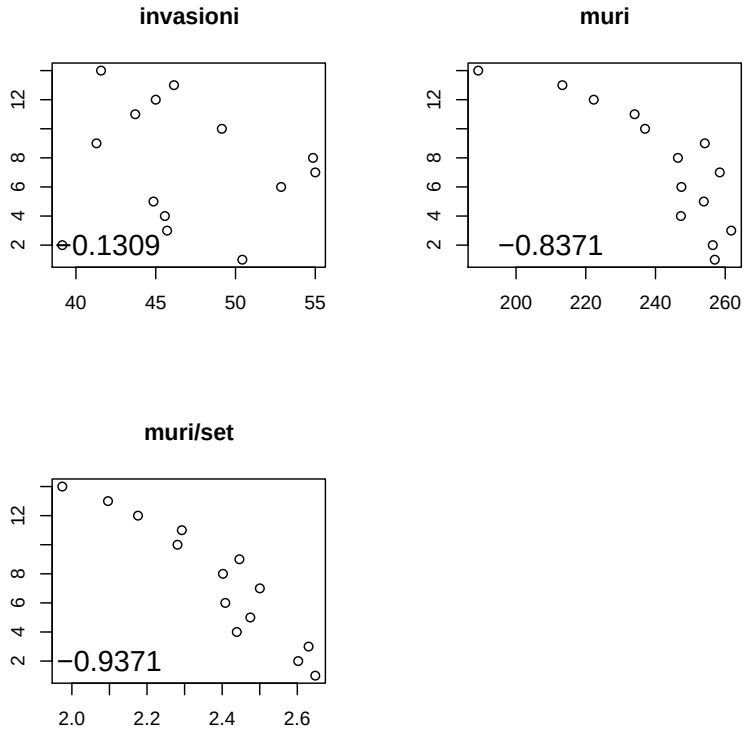

Figura 2.3: Grafico di dispersione con le correlazioni delle ultime tre variabili dell'anno delle medie

Le variabili che ci si aspetterebbe con una migliore correlazione sono quelle considerate tradizionalmente più importanti tra le presenti. La pallavolo è considerata uno sport di attacco perché questo fondamentale è quello che determina i punti più chiaramente (se non si fa punto con l'attacco, o si sbaglia e si concede il punto all'avversario, oppure si concede la possibilità all'avversario di attaccare e fare punto e questo, a differenza di altri sport, è molto grave perché tutte le azioni si concludono sempre con un punto per una squadra). È considerata fondamentale anche la fase del gioco dove si batte perché per vincere i *set*, e quindi la partita, deve esserci un divario di due punti tra le squadre. Infine è importante avere una solida ricezione per avere una buona fase di quello che una volta era il cambio palla. Dando uno sguardo alle efficienze in questi fondamentali si nota che questi dati hanno delle buone correlazioni: efficienza in attacco (-0.983), efficienza in battuta (-0.6906), efficienza in ricezione (-0.4568). È opportuno porre attenzione al

segno della correlazione perché, per avere la variabile risposta nell'asse delle ascisse, si ottiene una correlazione negativa, ma questo significa che al crescere della posizione in classifica, il rendimento migliora.

Proseguendo nell'analisi, bisogna notare come, eccetto l'efficienza in attacco, negli altri fondamentali si trovino migliori correlazioni per altre variabili. Nella battuta infatti, la variabile con una correlazione maggiore è il numero di ace per set con lo -0.7665, nella ricezione la correlazione migliore è degli errori in ricezione con ben il 0.9603. Inoltre hanno una buona dipendenza lineare anche i punti *break point* e i muri per set con rispettivamente il -0.9106 e il -0.9371.

2.3 L'analisi delle componenti principali sui dati

In questa parte del capitolo si cercherà di verificare se le variabili che hanno un'alta correlazione nei primi sette anni rilevati siano capaci di descrivere realmente la posizione in classifica finale al termine della stagione regolare. Per raggiungere lo scopo, per prima cosa è stata fatta una selezione delle variabili (necessaria dal punto di vista computazionale perché non è possibile applicare questo metodo avendo più variabili che posizioni in classifica) scegliendo quelle con una maggiore correlazione con il *rank*: i punti *break point*, il numero di ace, gli ace per set, l'efficienza in battuta, il numero di errori in ricezione, il numero di ricezioni perfette, gli errori in attacco, l'efficienza in attacco, il numero di muri e il numero di muri per set . Poi è stata controllata la matrice delle correlazioni tra le variabili (tabella 2.2) dalla quale si evince che tutte le variabili considerate, eccetto il numero di ricezioni perfette, sono molto incorrelate tra loro e questa è una condizione necessaria, come detto nella prima sezione di questo capitolo, per poter applicare il metodo delle componenti principali.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
break point	1.00	0.83	0.81	0.72	-0.91	-0.46	-0.75	0.91	0.94	0.93
ace	0.83	1.00	0.98	0.48	-0.67	-0.13	-0.51	0.66	0.82	0.79
ace per set	0.81	0.98	1.00	0.49	-0.72	-0.25	-0.60	0.73	0.79	0.81
effic. battuta	0.72	0.48	0.49	1.00	-0.69	-0.59	-0.60	0.70	0.65	0.67
errori ricez	-0.91	-0.67	-0.72	-0.69	1.00	0.65	0.85	-0.95	-0.87	-0.93
ricez perfette	-0.46	-0.13	-0.25	-0.59	0.65	1.00	0.67	-0.69	-0.35	-0.54
errori attacco	-0.75	-0.51	-0.60	-0.60	0.85	0.67	1.00	-0.92	-0.60	-0.75
effic. attacco	0.91	0.66	0.73	0.70	-0.95	-0.69	-0.92	1.00	0.82	0.92
muri punto	0.94	0.82	0.79	0.65	-0.87	-0.35	-0.60	0.82	1.00	0.96
muri per set	0.93	0.79	0.81	0.67	-0.93	-0.54	-0.75	0.92	0.96	1.00

Tabella 2.2: estratto della tabella con le correlazioni tra le variabili

Poi è stata fatta l'analisi delle componenti principali e come unità statistiche sono state inizialmente usate le posizioni in classifica ricavate dall'anno medio. Il risultato è riassunto dal grafico *biplot* (figura 2.5) che però va interpretato alla luce del grafico degli autovalori (figura 2.4) e della tabella 2.3.

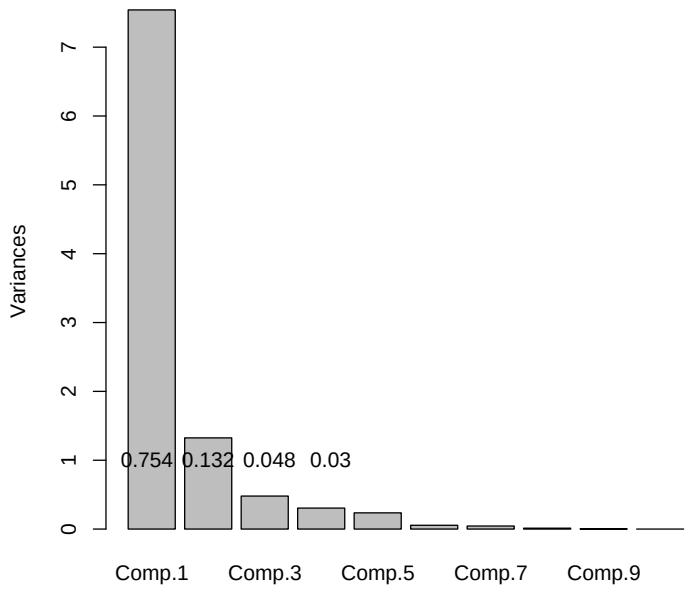

Figura 2.4: Grafico dei *loadings* dell'analisi sulla media degli anni

Come si può notare dal grafico 2.4 la maggior parte della variabilità è spiegata dalla prima componente principale, ben il 75.41%, e dalla seconda 13.25%. Le altre componenti principali sono sostanzialmente trascurabili perché sommate spiegano meno del 12% (solitamente due componenti sono ritenute sufficienti se queste su dieci variabili spiegano il 60% della variabilità [1, p. 238]) e, al contrario delle prime due, non hanno associati autovalori maggiori di uno. I risultati erano attesi perché c'è una forte dipendenza lineare nelle variabili scelte. Quando si andrà a leggere il grafico *biplot* si dovrà allora tenere conto soprattutto della posizione delle squadre rispetto all'asse orizzontale che è quella della prima componente principale. La tabella 2.3 invece, contiene i coefficenti di correlazione tra le prime due componenti principali e ognuna delle variabili. Da questa tabella possiamo dedurre l'intensità e, guardando il segno, il tipo di relazione che c'è tra una singola componente principale e una singola variabile. Per esempio possiamo notare che i punti *break point* hanno la maggiore correlazione con la prima componente principale e ad una peggiore posizione in classifica corrisponde un peggiore rendimento (figura 2.5). In generale la prima componente principale ha correlazione più alta con i fondamentali che determinano direttamente il punto, mentre la seconda è maggiormente legata alla ricezione perfetta che è l'unico dei fondamentali considerati che non serve a fare punto. Sarebbe stato interessante vedere se anche la difesa avesse avuto una buona correlazione con la seconda componente principale ma non era disponibile nel *dataset*.

	Prima componente principale	Seconda componente principale
break point	-0.97	0.12
ace	-0.81	0.55
ace per set	-0.84	0.44
efficienza in battuta	-0.75	-0.26
errori in ricezione	0.96	0.14
ricezioni perfette	0.59	0.73
errori in attacco	0.84	0.32
efficienza in attacco	-0.96	-0.19
muri punto	-0.92	0.24
muri per set	-0.97	0.06

Tabella 2.3: correlazione tra dati e prime due componenti principali

Il grafico *biplot* (figura 2.5) evidenzia chiaramente come le squadre siano disposte in maniera quasi orizzontale e si delineano chiaramente quattro sottogruppi formati praticamente dai quattro quadranti. Nel terzo quadrante in basso a sinistra troviamo le prime quattro classificate e si nota come siano influenzate dall'efficienza in attacco. Questo significa che le prime classificate hanno dei valori al di sopra della media in questo fondamentale rispetto alle altre squadre. Nel secondo quadrante ci sono le squadre dalla quinta alla settima posizione e notiamo che presentano valori al di sopra della media nel numero di *ace* e nel rapporto numero di *ace* e *set*. Nel primo quadrante ci sono le squadre dalla ottava alla dodicesima posizione. Da notare in particolare la posizione dell'ottava e nona classificata che, essendo vicino all'origine degli assi, indica che queste squadre hanno valori tutti vicini alla media praticamente in tutti i fondamentali. Le altre squadre in questo quadrante invece, hanno valori molto superiori alla media negli errori in ricezione e negli errori in attacco. La posizione delle ultime due classificate nel quarto quadrante invece, non ha una spiegazione chiara. Non sono influenzate particolarmente da nessun fondamentale perché probabilmente presentano sempre valori inferiori alla media.

Un'altra importante osservazione è che tra il segmento che rappresenta l'efficienza in attacco e quello che rappresenta gli errori in attacco c'è un angolo leggermente inferiore a 180 gradi che indica che la correlazione tra le due variabili è vicina a -1 e quindi sono fortemente incorrelate in senso negativo. Stupisce però che ci sia un angolo ancora più vicino ai 180 gradi tra l'efficienza in attacco e gli errori in ricezione. Questo fatto è molto rilevante perché si sta considerando solo il fatto di aver preso un punto dalla battuta dell'avversario e non si sta prendendo in esame se le ricezioni sono valutate "come buone" per consentire un buon sviluppo dell'azione. Si potrebbe pensare che prendere un *ace* dall'avversario, quindi aver sbagliato la ricezione, abbia un forte impatto sulla squadra perché induce poi all'errore in attacco (infatti le due variabili hanno una forte correlazione avendo un angolo molto piccolo tra loro) e penalizza ancora di più l'efficienza in attacco. Un ulteriore conferma di questo, deriva dal fatto che anche tra le variabili efficienza in battuta e errori in attacco c'è un angolo di 180 gradi a indicare che al crescere

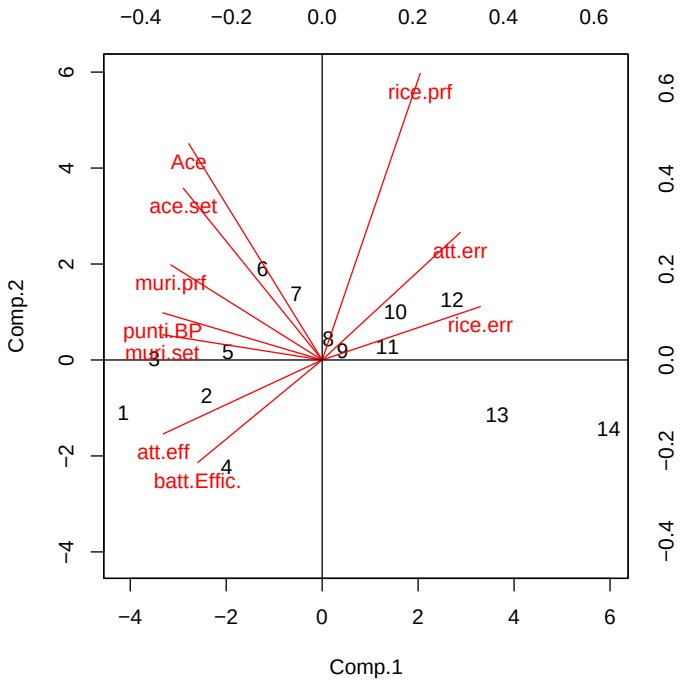

Figura 2.5: Grafico *biplot* dell'analisi sulla media degli anni

di una variabile l'altra diminuisce. A parere di chi scrive questa correlazione potrebbe evidenziare il fatto che chi ha grandi attaccanti in squadra, che fanno viaggiare la palla anche a 120 km/h, ha un occhio più allenato a seguire la traiettoria del pallone e quindi ha una maggiore facilità a ricevere rispetto a squadre che non hanno grandi attaccanti, quindi nemmeno grandi battitori, e non permettono ai ricevitori di allenarsi su questo aspetto. Bisogna specificare che nella precedente affermazione per grandi attaccanti si intende solamente i giocatori che riescono, sia in attacco che conseguentemente nella battuta in salto, a imprimere una elevata forza al pallone. Sarà interessante vedere se ci saranno cambiamenti tra alcuni anni, vista la sempre maggiore presenza di macchine chiamate "sparapalloni" che simulano la battuta in salto e fanno viaggiare il pallone alla velocità desiderata eliminando così il divario.

Alla luce dei risultati ottenuti con le componenti principali, si è cercato di capire se esse potessero essere usate come predittori. Per questo non sono stati inclusi gli anni 2008 e 2009 nella stima. Gli autovalori e autovettori stimati sono stati utilizzati per disegnare

dei nuovi grafici *biplot* dove però, come unità statistica, è stata presa la posizione in classifica nell'anno 2008 prima e nell'anno 2009 poi.

	Prima componente principale	Seconda componente principale
punti break point	-0.91	0.20
ace	-0.36	0.21
ace per set	-0.33	0.08
efficienza in battuta	-0.71	0.28
errori in ricezione	0.58	-0.01
ricezioni perfette	0.25	0.54
errori in attacco	0.31	0.50
efficienza in attacco	-0.70	-0.14
muri punto	-0.73	0.27
muri per set	-0.69	0.14

Tabella 2.4: correlazione tra dati del 2008 e prime due componenti principali

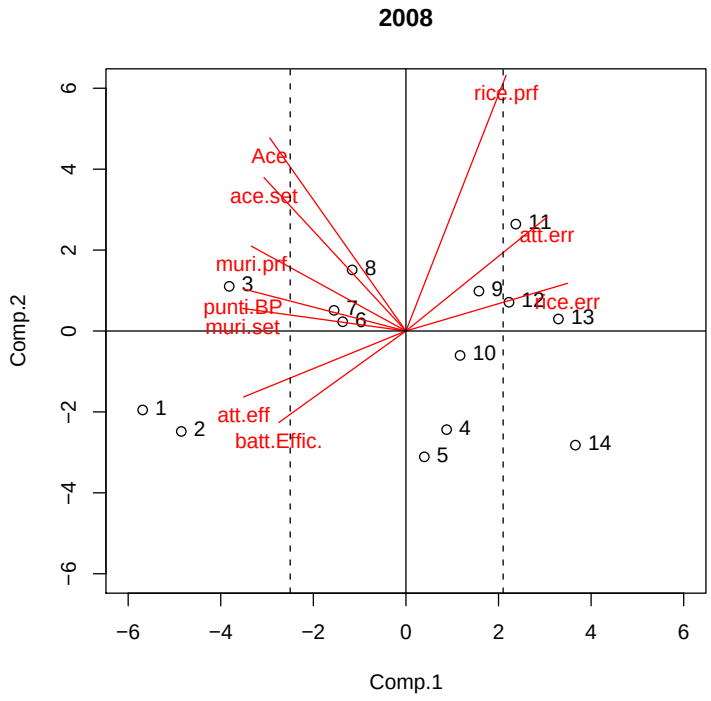

Figura 2.6: Grafico *biplot* del *rank* 2008 con le variabili stimate nell'anno medio

Nel primo grafico *biplot* relativo al 2008 (figura 2.6) si notano ancora chiaramente dei gruppi distinti di squadre con posizioni vicine nella classifica finale: un primo grup-

po alla sinistra del grafico comprende le prime tre classificate, al centro le squadre di medio-alta classifica e infine le ultime classificate si trovano sulla destra del grafico. Le variabili sono ovviamente quelle di prima e hanno la medesima interpretazione perché sono stati usati gli stessi autovalori del caso precedente in modo che si possa fare un confronto. Dal grafico emerge che le prime due classificate presentano valori superiori alla media nell'efficienza in attacco e in battuta, mentre le ultime classificate hanno fatto registrare un maggior numero di errori in attacco e ricezione; tuttavia notiamo che il modello non riesce a cogliere perfettamente le sfumature del singolo anno. Questo si deduce dalla presenza nel quarto quadrante del grafico di diverse squadre, anche con posizione in classifica lontana, che non presentano valori particolarmente lontani dalla media per nessun fondamentale particolare e non si può capire cosa le distingua dalle altre. Inoltre, se confrontiamo la tabella di correlazione tra i dati del 2008 e le componenti principali (tabella 2.4) con la stessa tabella riferita però all'anno ideale (tabella 2.3), notiamo come la correlazione peggiori sensibilmente per molte variabili e riamanga sostanzialmente immutata solo per i punti *break point*.

	Prima componente principale	Seconda componente principale
punti break point	-0.75	-0.03
ace	-0.08	-0.37
ace per set	-0.11	-0.42
efficienza in battuta	-0.20	0.12
errori in ricezione	0.67	0.38
ricezioni perfette	0.26	0.11
errori in attacco	0.48	-0.01
efficienza in attacco	-0.74	-0.07
muri punto	-0.42	0.43
muri per set	-0.51	0.43

Tabella 2.5: correlazione tra dati del 2009 e prime due componenti principali

Nel grafico *biplot* relativo al 2009 (figura 2.7) si delinea chiaramente il fatto che le variabili non descrivono più chiaramente i dati. Mantenendo la stessa suddivisione fatta in precedenza nel gruppo di sinistra sono contenute la seconda, terza, quinta e ottava

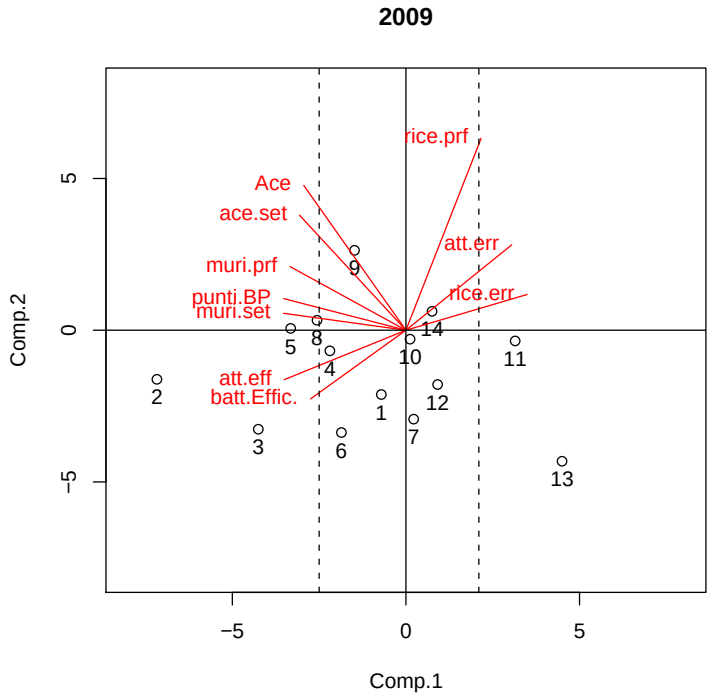

Figura 2.7: Grafico *biplot* del *rank* 2009 con le variabili stimate nell'anno medio

classificata, sulla destra del grafico ci sono solo l'undicesima e la tredicesima classificata e tutte le altre sono nel gruppo centrale. Inoltre per ben sei squadre (prima, sesta, settima, undicesima, dodicesima e tredicesima) non si riesce a capire da quali variabili siano influenzate e quindi non si riesce a spiegare la posizione in classifica. In generale, le variabili errori in attacco ed errori in ricezione non influenzano più così chiaramente le ultime classificate e tutte queste affermazioni trovano conferma dalla tabella 2.5, dove notiamo correlazioni molto inferiori ai casi precedenti.

In conclusione possiamo affermare che questa analisi sia valida in generale, cioè quando si considera l'insieme dei dati mediati, ma sia poco efficiente nei singoli anni se usata per fare previsioni. In particolare si nota un sensibile peggioramento con il passare del tempo, ovvero quando si considera l'anno più lontano dagli anni usati per trovare le medie.

Capitolo 3

IL MODELLO DI REGRESSIONE LOGISTICA A PROBABILITÀ PROPORZIONALE

> Se torturi i numeri abbastanza a lungo, confesseranno qualsiasi cosa.
>
> *Gregg Easterbrook*

3.1 Il modello di regressione logistica

Per vedere quali sono le variabili che effettivamente servono a individuare la posizione in classifica a fine campionato e quanto sono più importanti, ci si serve di un modello chiamato regressione logistica a probabilità proporzionale. Questo modello viene usato quando si ha di fronte una variabile ordinale, cioè una variabile i cui valori possono essere messi in ordine non decrescente o in ordine non crescente ma dei quali non è possibile conoscere la differenza esatta tra le variabili. Nel caso trattato in questo elaborato si ha appunto, come variabile risposta, la posizione in classifica a fine campionato che si può pensare come una variabile continua Z difficile da misurare ma al suo interno divisa da delle soglie (*cutpoints*), cioè dei punti che dividono nettamente una categoria dall'altra. In generale, si può partire da un modello log-lineare nel quale tutte le variabili sono trattate

allo stesso modo quindi, prendendo in considerazione una variabile Y con J categorie tale che:

$$Y = j \qquad \text{se} \qquad C_{j-1} \leq Z \leq C_j \tag{3.1}$$

avrà associata ad ogni categoria una probabilità π_j tale che $\sum_{j=1}^{J} \pi_j = 1$ e ci saranno $C_1, \ldots, C_{J-1}$ *cutpoints*. Nella figura 3.1 si riporta, a titolo esemplificativo, un disegno di una variabile continua con quattro categorie.

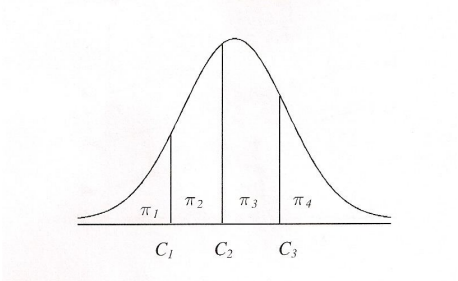

Figura 3.1: Distribuzione di una variabile continua con i *cutpoints* che definiscono l'ordine della variabile risposta.

La probabilità cumulata per la j-esima categoria sarà definita da questo rapporto:

$$\frac{P(Z \leq C_j)}{P(Z > C_j)} = \frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_j} \tag{3.2}$$

mentre il modello *logit* cumulativo sarà:

$$\log \frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_j} = \mathbf{X}_j^T \boldsymbol{\beta}_j \tag{3.3}$$

Quando il predittore lineare della 3.3 può essere scritto in questo modo:

$$\mathbf{X}_j^T \boldsymbol{\beta}_j = \beta_{0j} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} \tag{3.4}$$

cioè in modo tale da avere una intercetta β_{0j} che dipende da tutte le j categorie e le altre variabili esplicative che non dipendono da j, si parla di modello a probabilità proporzionale. Questa costruzione si basa sull'assunzione che gli effetti delle covariate

$x_1, x_2, \ldots, x_{p-1}$ sono gli stessi per tutte le categorie nella scala logaritmica [4], cioè ogni variabile ha lo stesso peso nel modello.

Una prima importante proprietà di questo modello è il fatto che se alcune categorie sono amalgamate o vengono eliminate, non cambia la stima dei coefficienti $\beta_1, \beta_2, \ldots, \beta_{p-1}$ nella 3.4 ma cambia solo la stima del termine β_{0j} (proprietà di collassamento) in quanto è l'unico dipendente dal numero di categorie j . Una seconda proprietà è che cambia solo il segno dei parametri se l'ordine delle categorie è invertito. Infatti:

$$\log \frac{\pi_j + \cdots + \pi_2 + \pi_1}{\pi_{j+1} + \cdots + \pi_j} = \beta_{0j} - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_{p-1} x_{p-1} \tag{3.5}$$

Molto importanti sono alcune statistiche che definiscono la bontà del modello:

`Test sulla devianza residua:`

$$RSS = \sum_{i=1}^{N} r_i^2 \,; \tag{3.6}$$

la somma dei residui al quadrato (*residual sum of squares*) è una misura di accostamento. I residui sono definiti come:

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}; \tag{3.7}$$

dove o_i sono le osservazioni e e_i sono le frequenze previste. Questa statistica si distribuisce asintoticamente come una χ^2 con $n - p$ gradi di libertà. Usando questa proprietà si può calcolare il *p-value* e capire se c'è differenza tra il modello saturo (cioè il modello che include tutte le variabili) e il modello corrente.

`AIC` (*Akaike's information criterion*).

Questo criterio in generale è definito come

$$AIC = -2\,log\,(\mathbf{L}) + 2p\,; \tag{3.8}$$

dove p è il numero di parametri nel modello statistico e $\mathbf{L}$ è il massimo della funzione di verosomiglianza per il modello considerato. Questo criterio permette di penalizzare i modelli che hanno un numero maggiore di parametri se questi non offrono una spiegazione sufficiente alla variabilità.

3.2 Identificazione del modello

In questa parte dell'elaborato verranno stimati i coefficienti del modello di regressione logistica a probabilità proporzionale descritto nella sezione precedente, cercando di identificare le variabili più significative. Con il programma `R`, attraverso la funzione `polr` della *library* `MASS`, si è arrivati a una prima stima dei coefficienti delle variabili punti *break point*, numero di *ace*, numero di *ace* per *set*, efficienza in battuta, errori in ricezione, ricezioni perfette, errori in attacco, percentuale di attacchi punto, efficienza in attacco, numero di muri fatti, muri fatti su numero di *set*:

	Value	Std. Error
punti break point	-0.06199	0.005945669
ace	0.05863	0.009916884
ace per set	-3.51175	0.004991709
effic. battuta	1.90291	0.001577336
errori ricez	0.01676	0.008686142
ricez perfette	-0.00017	0.002006494
errori attacco	0.01108	0.008007114
effic. attacco	-29.91082	0.002137389
muri punto	0.1232	0.008090255
muri per set	-12.39872	0.009551008

Tabella 3.1: Tabella dei coefficienti stimati

Com'è possibile capire però quali siano le variabili più importanti? Un modo molto semplice è ricorrere al test t. Questo test serve per verificare che i coefficienti stimati siano significativamente diversi da zero quindi importanti per descrivere le varie relazio-

ni. Questo valore viene calcolato dividendo ogni valore del coefficiente stimato per il relativo *standard error* ed elevando al quadrato il risultato. Questo valore può essere così confrontato con una χ^2_1 perché i coefficienti divisi per il loro *standard error* si distribuiscono come una normale di media nulla e varianza unitaria. Elevando al quadrato, la distribuzione che assumono questi valori diventa appunto una χ^2 con un grado di libertà. Tutti i coefficienti risultano significativamente diversi da zero perciò vanno presi in considerazioni tutti.

Un altro fatto importante da tenere presente è che, maggiore è il numero di parametri nel modello, migliore sarà il modo in cui questo modello interpreta la variabilità dei dati. Se per assurdo potessimo includere tutte le variabili del mondo avremmo un modello che spiega perfettamente tutta la realtà. Tuttavia questo fatto non è sempre utile perché un maggior numero di variabili implica uno sforzo computazionale maggiore, maggiori costi per reperire i dati e complica anche l'interpretazione. Per questo si è verificata la bontà del modello attraverso il criterio AIC (*Akaike's Information Criterion*).

Un modo per utilizzare questo criterio è, partendo dal modello creato, sottrarre una variabile a caso ottenendo due modelli, uno sottoinsieme dell'altro, che differiscono per una sola variabile. La differenza tra questi due modelli si distribuisce, anche in questo caso, come una χ^2_1 e quindi si può calcolare un nuovo livello di significatività per la variabile che questa volta però indica se c'è differenza oppure no tra i due modelli. Reinserendo la variabile tolta, sottraendo un'altra variabile e ripetendo questa procedura per tutte le variabili è possibile vedere qual è il coefficiente meno significativo e alla fine eliminare solo la variabile associata a questo coefficiente. A questo punto, viene calcolato l'AIC del nuovo modello e confrontato con quello del modello iniziale. Se è inferiore, il modello è migliorato. Si può poi ripetere tutto il procedimento ottenendo un nuovo modello con un'altra variabile in meno finchè l'AIC continua a diminuire. È importante eliminare una alla volta le variabili perché, ogni nuovo modello con una variabile in meno, ha dei coefficienti diversi perché devono essere stimati di nuovo e quindi possono cambiare sensibilmente la loro importanza. Nella tabella 3.2 si riportano i risultati della selezione.

	AIC	variabili incluse nel modello
1	473.64	punti.BP + Ace + ace.set + batt.Effic. + rice.err + rice.prf + att.err + att.eff + muri.prf + muri.set
2	471.66	punti.BP + Ace + ace.set + batt.Effic. + rice.err + att.err + att.eff + muri.prf + muri.set
3	469.71	punti.BP + Ace + batt.Effic. + rice.err + att.err + att.eff + muri.prf + muri.set
4	467.78	punti.BP + Ace + rice.err + att.err + att.eff + muri.prf + muri.set
5	467.15	punti.BP + Ace + rice.err + att.eff + muri.prf + muri.set

Tabella 3.2: Tabella dell'AIC relativo ai modelli ottenuti eliminando alcune variabili dal modello basato sulle correlazioni

3.3 Analisi su un modello generale

Nella sezione 3.2 si è visto come il metodo di regressione a probabilità proporzionale riesca a stimare dei coefficienti di significatività che non dipendono dalle correlazioni e le variabili considerate nel capitolo precedente si sono potute ridurre. Si è allora deciso di partire da un modello più generale che non comprende solo variabili mediamente correlate con la posizione in classifica per vedere quali risultati sarebbero emersi. Definito questo nuovo modello, che verrà chiamato modello generale per distinguerlo dal precedente, con le variabili punti vincenti, punti *break point*, numero di *ace*, numero di errori in battuta, numero di *ace* per *set*, efficienza in battuta, errori in ricezione, numero di ricezioni perfette, efficienza in ricezione, numero di errori in attacco, percentuale di attacchi perfetti, efficienza in attacco, numero di muri punto, numero di muri sul numero di *set*, si è proceduto stimando i coefficienti di questo secondo modello:

A questo punto si è ripetuto il procedimento applicato al modello precedente e si è calcolata la statistica test χ^2 per decidere quale variabile eliminare dal modello ad ogni passaggio con i seguenti risultati:

Confrontando la tabella 3.2 con la tabella 3.4 si nota che in quest'ultima si è ottenuto un modello migliore perché ha un AIC minore rispetto al precedente. Si può pensare che nel primo caso il problema fosse aver considerato le variabili sbagliate. Infatti, co-

	Value	Std. Error
punti vincenti	-0.0213	0.0051
punti break point	-0.0297	0.0092
ace	-0.14231	0.0135
errori battuta	-0.0972	0.0051
ace per set	25.2909	0.0044
effic. battuta	-213.9221	0.0006
errori ricez	0.5909	0.0101
ricez perfette	-11.2340	0.0445
effic ricezione	1120.2759	0.0005
errori attacco	-0.0031	0.0088
% attacchi punto	0.2709	0.0973
effic attacco	-15.7233	0.0014
muri punto	0.1038	0.0088
muri per set	-11.2382	0.0083

Tabella 3.3: Tabella dei coefficienti stimati per il modello generale

	AIC	variabili incluse nel modello
1	483.96	punti.vin+punti.BP+Ace+batt.err+ace.set+batt.Effic.+rice.err+ rice.prf.+rice.eff+att.err+att.prf.+att.eff+muri.prf+muri.set
2	460.74	punti.BP+Ace+batt.err+ace.set+batt.Effic.+rice.err+rice.prf. rice.eff+att.err+att.prf.+att.eff+muri.prf+muri.set
3	458.45	punti.BP+Ace+batt.err+ace.set+batt.Effic.+rice.err+rice.prf.+ rice.eff+att.err+att.prf.+muri.prf+muri.set
4	455.47	punti.BP+Ace+batt.err+batt.Effic.+rice.err+rice.prf.+rice.eff+ att.err+att.prf.+muri.prf+muri.set
5	454.81	punti.BP+Ace+batt.err+batt.Effic.+rice.err+rice.prf.+rice.eff+ att.err+muri.prf+muri.set
6	454.51	punti.BP+Ace+batt.err+batt.Effic.+rice.err+rice.prf.+rice.eff+ muri.prf+muri.set

Tabella 3.4: Tabella dell'AIC relativo ai modelli ottenuti eliminando alcune variabili dal modello generale

me sottolineato in precedenza, non si stanno più cercando le correlazioni ma il modo più semplice per descrivere le posizioni in classifica ottenute dalle squadre con le variabili a disposizione secondo i criteri definiti dal modello di regressione a probabilità proporzionale. Si può quindi affermare che, secondo questo modello, le migliori variabili da utilizzare per descrivere la posizione in classifica finale della *regular season* sono i punti

break point, il numero di *ace*, gli errori e l'efficienza in battuta, il numero di errori e l'efficienza in ricezione, il numero di muri totali e per set anche se stupisce un pò l'assenza di variabili legate all'attacco.

CONCLUSIONI

> Le statistiche dicono che uno su quattro soffre di qualche malattia mentale. Pensa ai tuoi tre migliori amici. Se stanno bene, vuol dire che sei tu.
>
> *Rita Mae Brown*

In questo lavoro si è inizialmente dimostrato che non è molto semplice e intuitivo capire cosa permetta alle diverse squadre di occupare una determinata posizione in classifica nel campionato italiano di pallavolo maschile. Si è visto infatti che guardando la tabella che indica dove le prime classificate sono state le migliori (tabella 1.6) non si notano fondamentali che rimangono costanti nel tempo. Inoltre, se si legge la tabella in senso orizzontale, in tutte le variabili cè una grande variabilità nella posizione in classifica che occupa la squadra che ha ottenuto i risultati migliori in quel fondamentale, eccetto per l'efficienza in attacco dove la miglior squadra è sempre tra le prime tre classificate. Tuttavia, nemmeno questo fatto può essere considerato decisivo perché non può essere dimostrato che anche per le posizioni inferiori della classifica valga la regola che un miglior attacco corrisponda a una migliore posizione. Per questo motivo sono state considerate le correlazioni e, capito che non è possibile fare un'analisi considerando i singoli anni, si è provveduto a stimare le medie dei vari anni rispetto alla posizione in classifica. Calcolate le correlazioni di queste nuove variabili medie, si sono notate diverse dipendenze lineari e per capirne esattamente le relazioni, è stata fatta un'analisi delle componenti principali considerando le variabili incorrelate più esplicative.

Da questa analisi è emersa innanzitutto la conferma che l'efficienza in attacco è fon-

damentale per determinare la posizione in classifica poichè nel grafico 2.5 si vede chiaramente che le prime classificate sono tutte vicine alla linea della variabile efficienza in attacco e questo indica che hanno valori in questo fondamentale superiori alla media. Un'altra importante osservazione riguarda la variabile errori in ricezione. Questa variabile è la più sorprendente in questa analisi per due motivi: il primo è che risulta più importante dell'efficienza in ricezione e, da un punto di vista pratico, questo indica che fa più la differenza il numero di palloni che si riesce a ricevere in modo giocabile rispetto alla qualità della ricezione. Se in via puramente teorica questa è una cosa ovvia, si riscontra che in pratica le squadre che retrocedono non riescono a lavorare efficacemente su questo aspetto al contrario delle migliori squadre. Il secondo importante aspetto della variabile errori in ricezione è che è fortemente correlata, in modo negativo, con l'efficienza in attacco e quindi, in qualche modo, prendere tanti *ace* significa anche attaccare male. La natura di questa relazione è opinabile tuttavia, a parere di chi scrive, avere in squadra grandi attaccanti e grandi battitori, soprattutto battitori in salto, allena i ricevitori a velocità del pallone superiori e questo provoca un sensibile miglioramento nella capacità di gestire palloni veloci. Questa analisi è stata poi usata per modellare i risultati ottenuti negli anni successivi con risultati incerti. Se infatti si può affermare che per l'anno 2008 si è ottenuto un risultato soddisfacente, la qualità dei risultati dell'anno 2009 non è buona anche se nemmeno insoddisfacente. Ciò che non funziona principalmente in questo tentativo di previsione sono soprattutto le variabili che permettono di distinguere le squadre di metà classifica. Ciò potrebbe dipendere dal fatto che la pallavolo ha sicuramente subito una evoluzione che probabilmente non poteva essere colta dal modello. Una possibile spiegazione consiste nell'evidenziare che gli anni considerati nell'analisi sono i primi con un nuovo sistema di gioco. Questo comporta che è necessario maggior tempo affinchè le squadre cambino metodologia di allenamento e tattica e affinassero questi aspetti. Maggiori osservazioni sono necessarie quindi per determinare gli aspetti del gioco che si sono evoluti e non sono stati pienamente colti da questa analisi.

Nel terzo capitolo di questo elaborato, si è adottato un approccio diverso, secondo il modello di regressione logistica a probabilità proporzionale che è sembrato il migliore per descrivere le posizioni in classifica a fine campionato. Partendo dal modello creato nel secondo capitolo, attraverso il test χ^2 e confrontando gli AIC, sono state eliminate le variabili meno importanti per ottenere un modello più semplice ma ugualmente esplicativo. Infine si è deciso ripetere questo approccio su un modello più generale che comprendesse molte più variabili perchè, questo secondo metodo di analisi sui dati, non si basa più sulle correlazioni. Da questa seconda analisi è emerso un insieme di variabili non così diverso dal precedente ma più complesso che descrive in modo migliore la variabilità dei dati. Alla conferma delle variabili punti *break point*, numero di *ace*, numero di muri punto e numero di muri sul numero di *set*, presenti già nel primo modello, si aggiunge una forte presenza delle variabili della battuta e della ricezione (numero di errori in battuta, efficienza in battuta, errori in ricezione, numero di ricezioni perfette, efficienza in ricezione) che completano e migliorano la descrizione della variabilità dei dati.

BIBLIOGRAFIA

[1] SERGIO ZANI, ANDREA CERIOLI, *Analisi dei dati e data mining per le decisioni aziendali*, Giuffrè editore, Milano, 2007

[2] MARCO PAOLINI, *Il nuovo sistema pallavolo*, Calzetti Mariucci editore, 2001

[3] GIAN LUCA PASINI, *Il computer un uomo in più*, articolo della Gazzetta dello Sport, 14/02/2004

[4] ANNETTE J. DOBSON, *An introduction to generalized linear models, second edition*, Chapman & Hall, Boca Raton, 2002