

UNIVERSITA' DEGLI STUDI DI PADOVA
FACOLTA' DI SCIENZE STATISTICHE
CORSO DI LAUREA IN STATISTICA E TECNOLOGIA INFORMATICA



TESI DI LAUREA

**CONFRONTO TRA MODELLI STATISTICI NON
PARAMETRICI :UNA APPLICAZIONE AI GOAL DEI
CAMPIONATI DI CALCIO**

Relatore: Prof . Stuart Coles

Laureando: Kenne Pagui Euloge Clovis

Anno Accademico 2006-2007

A mio Padre **Joseph Pagui** e a mia Madre **Bernadette Maffo**

INDICE

INTRODUZIONE E OBIETTIVI.....	9
CAPITOLO 1:CAMPIONATO FRANCESE	13
1.1 CAMPIONATO FRANCESE SERIE A.....	13
1.1.1 Analisi esplorativa.....	13
1.1.1.1 Primo approccio dell'analisi esplorativa	13
1.1.1.2 Secondo approccio dell'analisi esplorativa	14
1.1.2 Adattamento dei modelli per i nostri dati.....	17
1.1.2.1 Modello per i gol delle squadre in casa.....	17
1.1.2.2 Modello per i gol delle squadre fuori casa.....	18
1.1.2.3 Modello per il totale di gol fatti per partita.....	20
1.1.2.4 Modello per la differenza dei gol.....	21
1.2 CAMPIONATO FRANCESE SERIE B	22
1.2.1 Analisi esplorativa	22
1.2.1.1 Primo approccio dell'analisi esplorativa	22
1.2.1.2 secondo approccio dell'analisi esplorativa.....	24
1.2.2 Adattamento dei modelli per i nostri dati.....	26
1.2.2.1 Modello per i gol delle squadre in casa.....	26
1.2.2.2 Modello per i gol delle squadre fuori casa.....	27
1.2.2.3 Modello per il totale dei gol per partita.....	29
1.2.2.4 Modello per la differenza dei gol per partita	31
1.3 CAMPIONATO FRANCESE SERIE C	32
1.3.1 Analisi esplorativa.....	32
1.3.1.1 Primo approccio dell'analisi esplorativa	32
1.3.1.2 secondo approccio dell'analisi esplorativa.....	34
1.3.2 adattamento dei modelli ai nostri dati.....	35
1.3.2.1 Modello per i gol delle squadre in casa.....	35
1.3.2.2 Modello per i gol delle squadre fuori casa.....	37
1.3.2.3 Modello per il totale dei gol per partita.....	38
1.3.2.4 Modello per la differenza dei gol per partita	41
1.3.3 Differenza tra le varie serie.	42
1.3.3.1 prima analisi preliminare.	42
1.3.3.2 seconda analisi preliminare.	43

1.3.4 Modelli di analisi delle differenze tra le serie indipendentemente dal mese e dell'anno.....	43
1.3.4.1 Caso specifico dei gol delle squadre in casa per tutte le Serie messe insieme	43
1.3.4.2 Caso specifico dei gol delle squadre fuori casa.....	45
per tutte le Serie messe insieme.....	45
1.3.4.3 Caso specifico del totale dei gol per partita.....	46
1.3.5 Modelli di analisi delle differenze tra le serie dato il mese e l'anno.	47
1.3.5.1 Caso specifico dei gol delle squadre in casa.....	47
per tutte le Serie messe insieme.....	47
1.3.5.2 Caso specifico dei gol delle squadre fuori casa.....	49
per tutte le Serie messe insieme.....	49
1.3.5.3 Caso specifico del totale dei gol per partita.....	51
per tutte le Serie messe insieme.....	51
 CAPITOLO 2:CAMPIONATO ITALIANO	54
2.1 CAMPIONATO ITALIANO SERIE A	54
2.1.1 Analisi esplorativa.....	54
2.1.1.1 Primo approccio dell'analisi esplorativa	54
2.1.1.2 Secondo approccio dell'analisi esplorativa	56
2.1.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.....	58
2.2 CAMPIONATO ITALIANO SERIE B.....	59
2.2.1 Analisi esplorativa.....	59
2.2.1.1 Primo approccio dell'analisi esplorativa	59
2.2.1.1 Secondo approccio dell'analisi esplorativa	61
2.2.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.....	63
2.3 CAMPIONATO ITALIANO SERIE C.....	65
2.3.1 Analisi esplorativa.....	65
2.3.1.1 Primo approccio dell'analisi esplorativa	65
2.3.1.1 Secondo approccio dell'analisi esplorativa	67
2.3.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.....	69
2.3.3 Differenza tra le varie Serie indipendentemente dal mese e dell'anno.....	71
2.3.4 Modelli di analisi delle differenze tra le varie Serie dato il mese e l'anno.	73
 CAPITOLO 3:CAMPIONATO SPAGNOLO	76
3.1 CAMPIONATO SPAGNOLO SERIE A	76
3.1.1 Analisi esplorativa.....	76

3.1.1.1 <i>Primo approccio dell'analisi esplorativa</i>	76
3.1.1.1 <i>Secondo approccio dell'analisi esplorativa</i>	78
3.1.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.....	80
3.2 CAMPIONATO SPAGNOLO SERIE B.....	81
3.2.1 Analisi esplorativa.....	81
3.2.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.....	83
3.3 CAMPIONATO SPAGNOLO SERIE C.....	84
3.3.1 Analisi esplorativa.....	84
3.3.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.....	86
3.3.3 Differenze tra le varie Serie indipendentemente dal mese e dell'anno.....	87
3.3.4 Modelli di analisi delle differenze tra le varie Serie dato il mese e l'anno	90
CONCLUSIONI	93

INTRODUZIONE E OBIETTIVI

La nascita dello sport ha avuto numerosi sviluppi dai giochi e tornei medioevali ,ai ludi romani ed ai giochi greci,sino ad arrivare ai giochi popolari ed alle prime forme di ginnastica che si sono via via evolute giungendo alla nascita di quella che è l'attuale educazione fisica, strettamente legata con la pedagogia e,quindi, con l'educazione del fanciullo. Le scommesse sono nate quando è nato l'evento sportivo e l'evoluzione formale del mondo degli scommettitori,il passaggio da un rapporto orale ad un sistema strutturato, ha spinto il bisogno di qualcuno che controlli e generi i sistemi di pagamento intermedio: i bookmaker.Essendo arrivati ad un sistema di scommesse molto strutturato,con un bacino di utenti vasto ,è nata la necessità, soprattutto per i bookmaker, di stilare delle graduatorie probabilistiche per aiutare il popolo degli scommettitori a capire quali eventi sono meno, cioè più probabili, da quelli più remuneranti, quindi con meno possibilità. In questo campo entra la statistica, che con adeguati modelli facilita il lavoro dei bookmaker e dona speranze ai giocatori d'azzardo. In Italia, per esempio, per quanto riguarda il calcio il sistema più famoso è la Schedina del Totocalcio, sulla quale si decidono 14 risultati, 9 delle partite di serie A e 5 di serie minori. Inoltre, esistono le agenzie di scommesse che permettono di puntare su un singolo evento o partita.

Il lavoro svolto di seguito analizza la tendenza a segnare in una singola partita di calcio . Ma innanzitutto sottolineiamo il fatto che per uno scommettitore risulta importante sapere gli effetti specifici prima di fare una scommessa, cioè le squadre che giocano, quale delle due squadre gioca a casa, quanti gol vengono segnati in media sia dalle singole squadre che al termine della partita, risulta anche importante sapere la potenza delle squadre , cioè sapere le squadre che segnano di più o meno, questo ultimo effetto rimarrà fuori scopo del lavoro.

L'obiettivo del lavoro è anche di sfruttare i dati, cercando di identificare valori interessanti e, in particolare, cercare di cogliere il trend dei gol segnati al variare sia del mese che dell'anno. Qui sorge anche la necessità ai scommettitori la conoscenza degli effetti generali, cioè se il trend non potrebbe cambiare col tempo per i motivi vari tali:

- Alcune squadre cambiano il loro modo tattico di giocare
- L'incompetenze di certi allenatori
- Il cambio delle regole di gioco o la perdita dei talenti di alcuni giocatori
- Altro

Detto ciò ci rimane un aspetto molto importante ai confronti degli scommettitori da tenere presente che loro possono scommettere sia sui gol segnati da una squadra in casa o meno, o la differenza dei gol per partita, o la somma dei gol per partita. Per questo motivo, concentreremo le nostre analisi su queste quattro variabili (gol in casa, gol fuori casa, somma gol, differenza gol).

In questa seconda parte del lavoro mi sono soffermato sulla descrizione di alcuni metodi utilizzati. Dato che le nostre variabili di interesse sono variabili discrete del tipo conteggio, hanno una distribuzione di poisson e andrebbe bene, per la stima dei modelli statistici per i dati, il metodo parametrico della classe dei Modelli lineari generalizzati (GLM) con famiglia poisson o il metodo non parametrico della classe dei Modelli additivi generalizzati (GAM) con famiglia poisson. Uno degli obiettivi principali che è di cogliere la struttura delle variabili di interesse al variare del mese o dell'anno, ci ha spinto di scegliere il modello non parametrico di poisson per adattare i dati perché ci permette di introdurre una qualche forma di 'struttura', ovvero un modello sulla forma della funzione di regressione, e anche perché cerchiamo di mantenere il massimo di flessibilità. scegliendo il metodo parametrico di poisson noi non riusciamo a cogliere perfettamente il trend e bisogna operare delle trasformazioni sulle variabili che ci porta a delle perdite di informazione. Per quanto riguarda la parte delle analisi esplorativa, visto che desideriamo un qualcosa che ci indica una qualche struttura abbiamo scelto di usare una regressione non parametrica, perché non ci sono vincoli sulla distribuzione delle variabili, assume solo alcune condizioni di regolarità nell'andamento della funzione stimata. Abbiamo aggiunto anche degli

intervalli di confidenza dei valori predetti punto a punto lungo le curve stimate, assumendo che i valori predetti sono normali.

CAPITOLO 1: CAMPIONATO FRANCESE

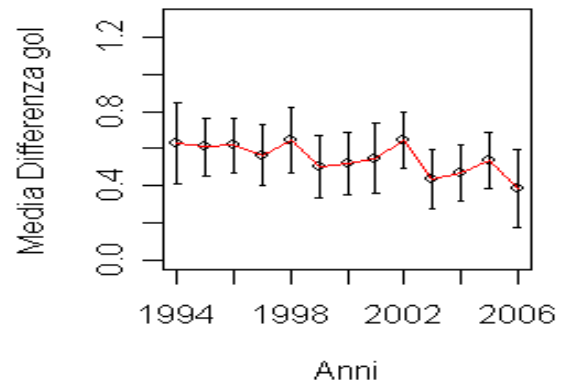
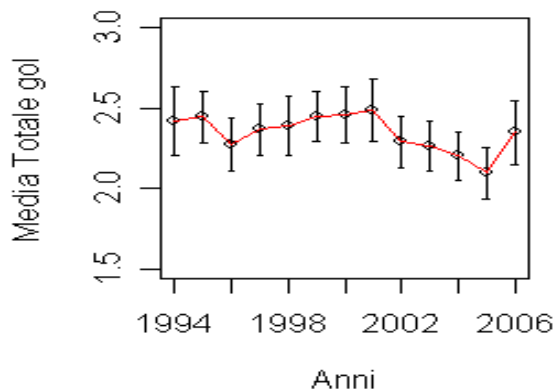
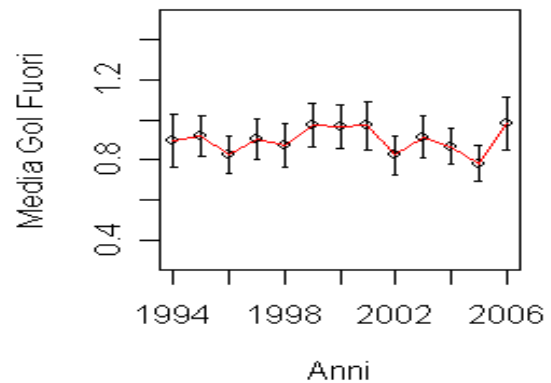
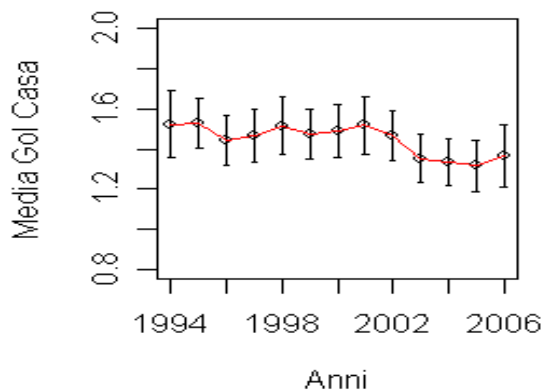
1.1 CAMPIONATO FRANCESE SERIE A

1.1.1 Analisi esplorativa

1.1.1.1 Primo approccio dell'analisi esplorativa

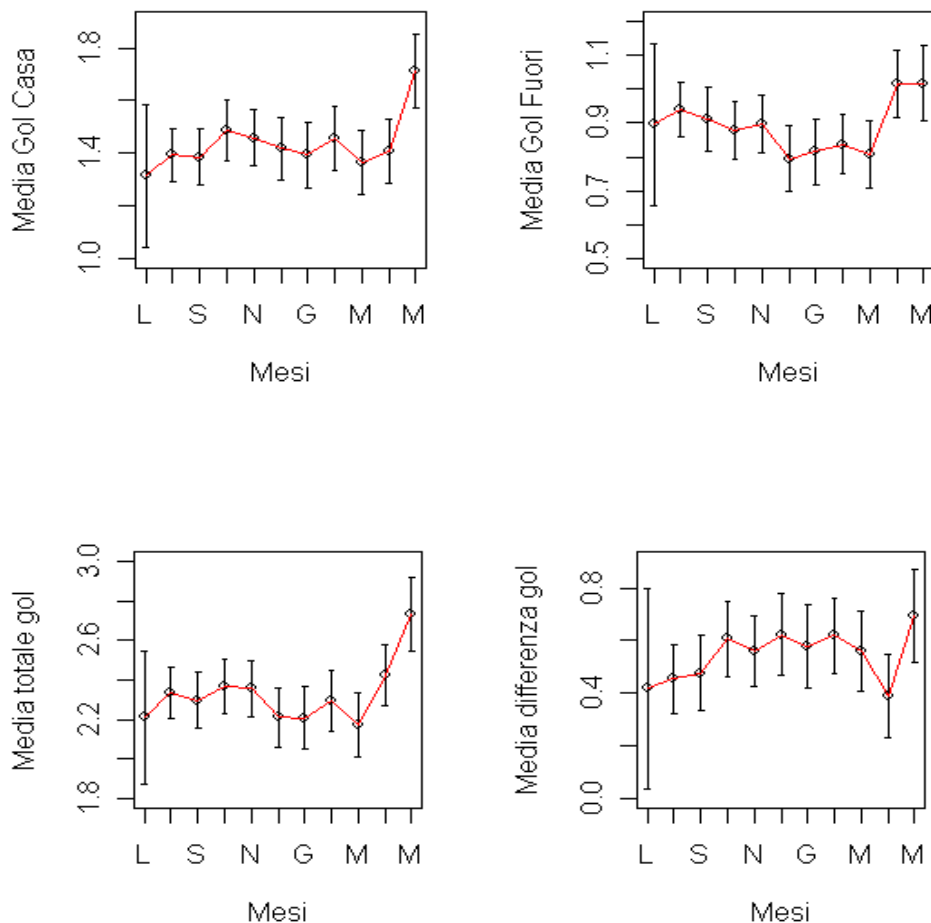
I grafici sotto esposti rappresentano gli intervalli di confidenza della media dei gol al variare dei mesi oppure al variare degli anni. Come detto al punto precedente siamo interessati a cogliere la struttura dei gol per le squadre in casa ,fuori casa , il totale dei gol e anche la differenza.

➔ **struttura dei gol delle squadre in casa al variare dell'anno.**



Considerando la media dei gol per le squadre in casa, fuori casa , il totale dei gol, e la differenza di gol non sembra esserci una particolare struttura nei dati. L'andamento è tendenzialmente stazionario.

→ **struttura dei gol delle squadre in casa al variare del mese.**



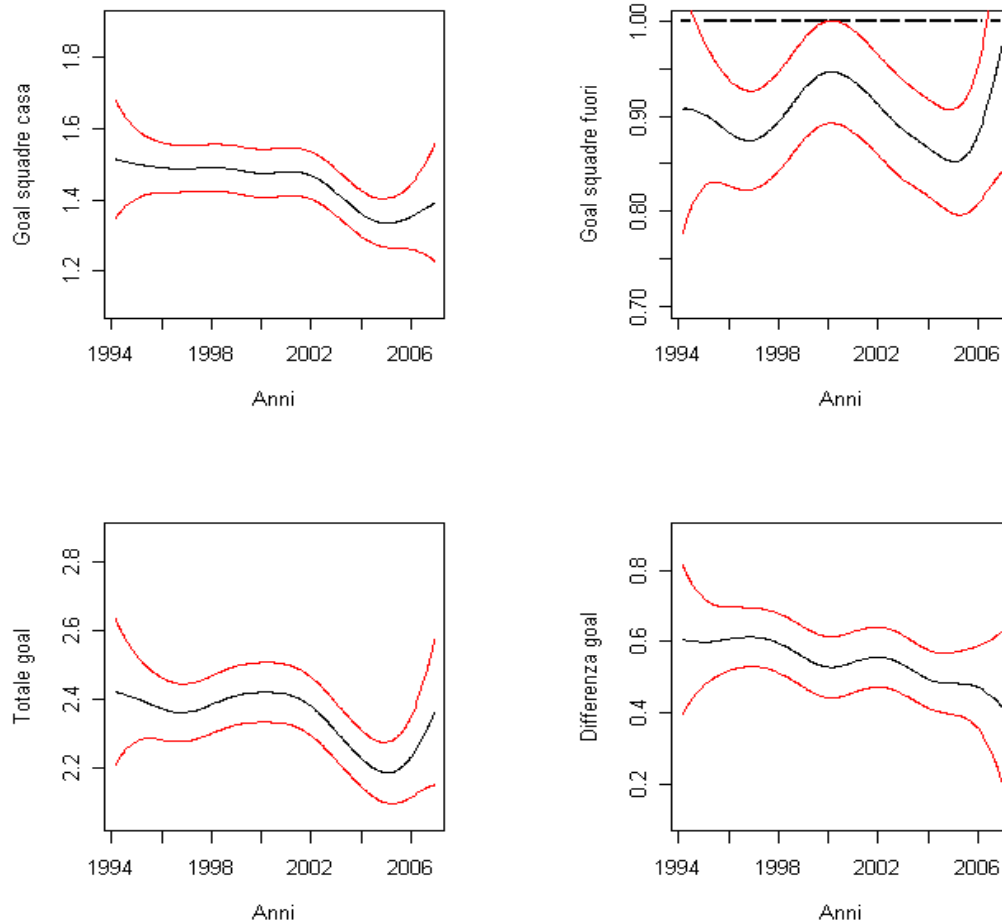
Nel campionato francese di primo livello, tranne la media della differenza dei gol, tendenzialmente abbiamo lo stesso andamento che sembra essere debolmente crescente nei primi tre grafici e poi, al variare dei mesi, sembrerebbe esserci una struttura particolare della media dei gol. In particolare nei primi tre grafici l'andamento sembra essere stazionario fino al mese di marzo e poi cresce fino alla fine della stagione(Maggio). Sembra, perciò, che si segna in media di più negli ultimi due mesi della stagione.

1.1.1.2 Secondo approccio dell'analisi esplorativa

Cerchiamo di lisciare la risposta per poter cogliere bene la struttura dei gol.

→ **Struttura dei gol al variare degli anni applicando regressione non parametrica**

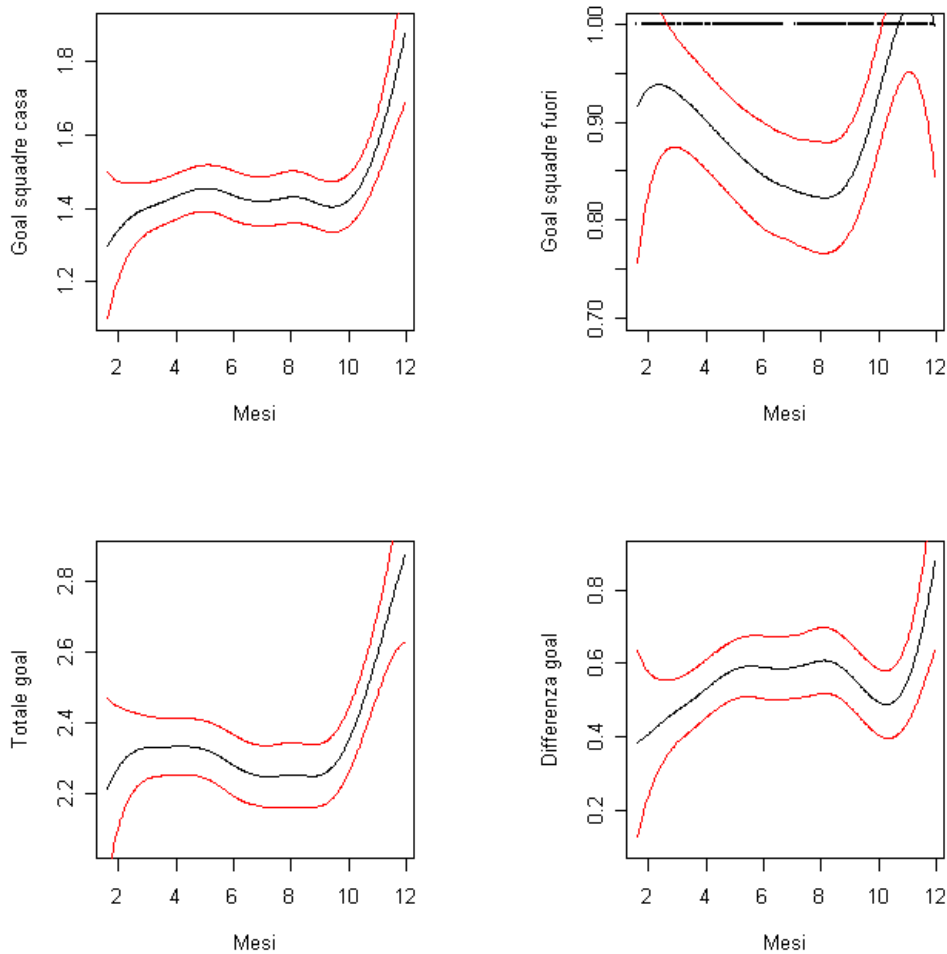
Legenda: In rosso abbiamo l'intervallo di confidenza.



In questo caso, se consideriamo i gol delle squadre fuori casa e il totale dei gol per partita tendenzialmente sembrerebbe esserci una qualche struttura, invece se consideriamo i gol per le squadre di casa sembra non esserci una qualche struttura.

→ **Struttura dei gol al variare dei mesi**

Legenda: 1='luglio'; 2='agosto'; ...; 11='maggio'



Se consideriamo i gol per le squadre in casa sembra che dall'inizio della stagione fino a marzo circa si segna in media lo stesso numero di gol, successivamente inizia a crescere fino alla fine della stagione. Osserviamo lo stesso andamento per il totale dei gol per partita. Considerando i gol per le squadre in trasferta, l'andamento decresce fino a febbraio e poi cresce fino alla fine della stagione. Quindi rispetto ai mesi sembrerebbe esserci una struttura particolare dei gol.

Osservazione: finora abbiamo notato che nella maggioranza dei grafici, al variare dei mesi sembra che si segni di più alla fine della stagione. Ciò può essere dovuto al fatto che alla fine della stagione ci sono delle squadre che competono per vincere il campionato, altre invece competono per non retrocedere.

1.1.2 Adattamento dei modelli per i nostri dati

1.1.2.1 Modello per i gol delle squadre in casa

Call: `gam(formula = golC ~ s(Mesi) + s(Anni), family = poisson, data = dat)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9214	-0.5642	-0.3248	0.4696	3.9163

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4990.213 on 4229 degrees of freedom
 Residual Deviance: 4954.680 on 4221 degrees of freedom
 AIC: 12869.02

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df Npar	Chisq	P(Chi)
(Intercept)	1				
s(Mesi)	1	3		14.1045	0.0028
s(Anni)	1	3		4.4864	0.2135

Da questo output si vede subito che la variabile Anni non risulta significativa al livello 5%, pero' risulta significativa la variabile Mesi: questo è cio' che potevamo aspettarci dall'analisi esplorativa. Vediamo la bontà del modello basato sulla devianza residua.

1-pchisq(4954.680,4221)
 1.842970e-14

Il valore del livello di significatività osservato ci suggerisce che il nostro modello corrente non spiega bene i dati rispetto al modello saturo. Vediamo cosa succede togliendo la variabile non significativa Anni.

Call: `gam(formula = golC ~ s(Mesi), family = poisson, data = dat)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8739	-0.5569	-0.3577	0.4661	3.7954

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4990.213 on 4229 degrees of freedom
 Residual Deviance: 4970.139 on 4225 degrees of freedom
 AIC: 12876.48

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			15.6693	0.0013

Qui si vede che la variabile Mesi rimane sempre significativa. Vediamo quanto è buono questo modello .

1-pchisq(4970.139,4225)
7.882583e-15

Il valore del p-value ci suggerisce che il modello saturo risulta preferibile rispetto al modello corrente. concludiamo dicendo che nella Serie A del campionato francese arriviamo ad un modello in cui la variabile Mesi risulta significativa; tuttavia, secondo un test basato sulla bontà del modello ,il modello richiederebbe altri parametri ,quindi diciamo che il modello non spiega bene i dati e di conseguenza le previsioni non sono molto affidabili. Il fatto che il nostro modello finale non spiega bene i dati, potrebbe essere dovuto al fatto che i gol segnati non hanno proprio una distribuzione di poisson e in parte c'è una extra-variabilità di conseguenza della variabilità delle squadre.

1.1.2.2 Modello per i gol delle squadre fuori casa

Call: gam(formula = golF ~ s(Mesi) + s(Anni), family = poisson, data = dat)
Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.50157	-1.31541	0.06536	0.20837	3.53421

(Dispersion Parameter for poisson family taken to be 1)

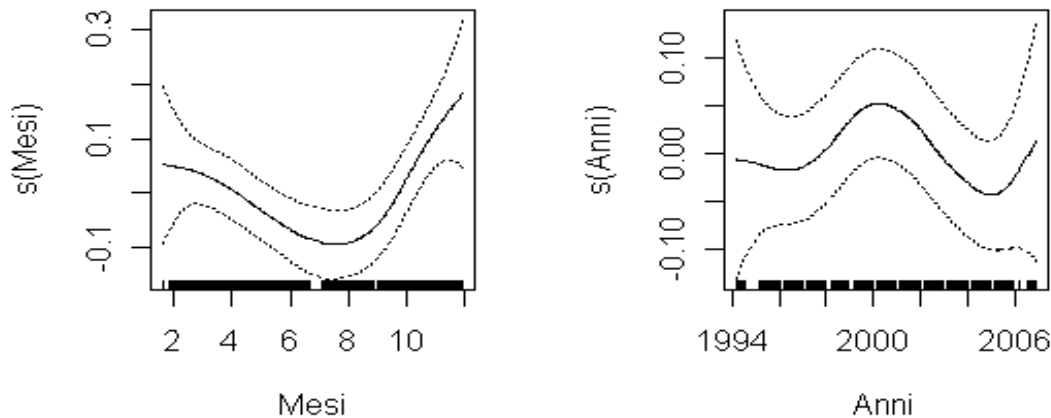
Null Deviance: 4968.799 on 4229 degrees of freedom
Residual Deviance: 4940.717 on 4221 degrees of freedom
AIC: 10592.22

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			20.8939	0.0001
s(Anni)	1	3			6.8301	0.0775

→ Grafico delle curve che sono state stimate



Questa è la struttura dei gol fuori casa ,al variare dei mesi e degli anni. Essa è simile quella che abbiamo visto al punto precedente nell'analisi esplorativa. Se guardiamo l'output di sopra , vediamo che, se fissiamo un livello approssimato del 5% per la non significatività delle variabili,risulta significativa solo la variabile mesi. Vediamo quanto è buono questo modello.

```
1-pchisq(4940.717,4221)
5.240253e-14
```

Il valore del livello di significatività ci suggerisce che il modello stimato non spiega bene i dati quanto il modello saturo che attribuisce un parametro ad ogni osservazione. Proviamo a togliere la variabile Anni che non risulta significativa.

```
Call: gam(formula = golF ~ s(Mesi), family = poisson, data = dat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.46336	-1.31360	0.06484	0.19480	3.58023

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4968.799 on 4229 degrees of freedom
Residual Deviance: 4947.629 on 4225 degrees of freedom
AIC: 10591.14

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			20.7309	0.0001

Togliendo la variabile Anni ,la variabile Mesi, come negli altri, rimane lo stesso significativa. È opportuno vedere comunque la sua bontà.

```
1-pchisq(4947.629,4225)
[1] 4.318768e-14
```

Anche togliendo la variabile Anni il modello non spiega bene la variabilità dei dati. Quindi anche con i gol delle squadre fuori casa arriviamo ad un modello che richiede l'aggiunta di ulteriori variabili; con le variabili a disposizione, infatti, non ce la facciamo a risolvere il problema e di conseguenza avremo delle previsioni dubbiose. La non adeguatezza del modello finale stimato potrebbe lo stesso essere dovuta ai motivi specificati al punto precedente.

1.1.2.3 Modello per il totale di gol fatti per partita

```
Call: gam(formula = (golC + golF) ~ s(Mesi) + s(Anni), family = poisson,
data = dat)
```

Deviance Residuals:

```
Min    1Q  Median    3Q   Max
-2.3852 -0.9606 -0.1922  0.5124  3.3157
```

(Dispersion Parameter for poisson family taken to be 1)

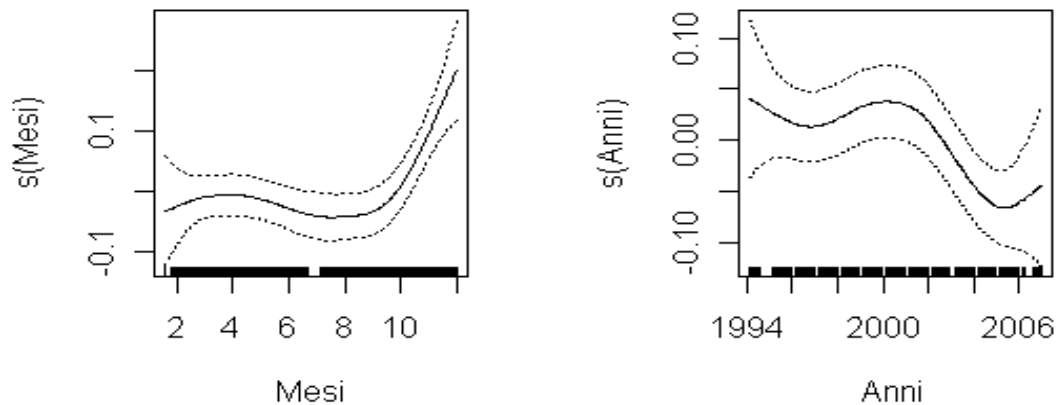
```
Null Deviance: 5034.798 on 4229 degrees of freedom
Residual Deviance: 4987.125 on 4221 degrees of freedom
AIC: 15328.08
```

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			24.6043	1.868e-05
s(Anni)	1	3			9.5476	0.0228

grafico delle curve stimate



Su questi due grafici sopra riportati si vede la struttura del totale dei gol rispettivamente al variare del mese e dell'anno. Abbiamo lo stesso andamento rispetto a quanto visto sui grafici nell'analisi esplorativa. Sull'output sopra riportato, in questo caso, si vede che al livello 5% per la non significatività delle variabili sia la variabile Mesi che Anni risulta significativa. Se, però, fissiamo il livello all' 1% solo la variabile Mesi risulta significativa. Vediamo quanto va bene il modello stimato.

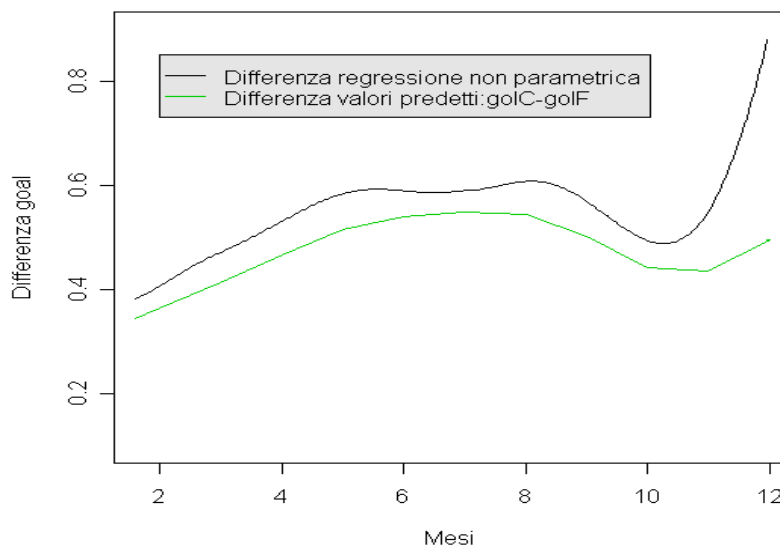
1-pchisq(4987.125,4221)
1.44329e-15

Questo valore del livello di significatività ci suggerisce che il nostro modello non spiega bene la variabilità dei dati quanto il modello saturo nel quale si associa un parametro ad ogni osservazione; di conseguenza il modello stimato richiederebbe variabili in più. In conclusione, con le variabili che abbiamo a disposizione noi non siamo in grado di migliorare il modello; quindi un tale modello non è adeguato per la spiegazione dei dati e per le previsioni. Anche qui il problema della non adeguatezza, può essere dovuto ai motivi specificati nei casi precedenti.

1.1.2.4 Modello per la differenza dei gol

Formulare un modello per la differenza di gol per partita risulta difficile, perché nelle analisi esplorative abbiamo usato la regressione non parametrica per cogliere la struttura della differenza di gol sia al variare del mese che dell'anno.

Questo è un metodo che non ha vincolo sulla distribuzione della risposta e, volendo adattare un modello con la *generalized additive models*, sorge il problema della specificazione della famiglia della variabile di interesse. Nel nostro caso avremmo valori negativi che non vanno bene per un modello con famiglia poisson. Come primo approccio proviamo a fare un grafico della differenza dei valori predetti dei gol per squadre in casa e quelli delle squadre fuori e lo confrontiamo con il grafico ottenuto con la regressione non parametrica.



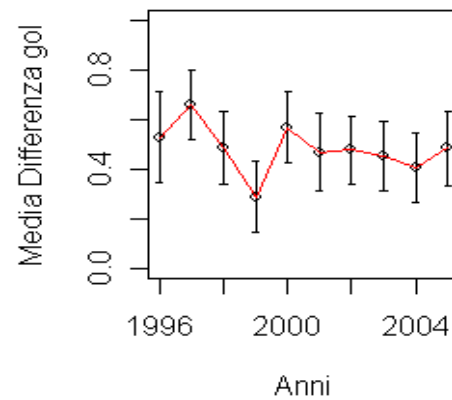
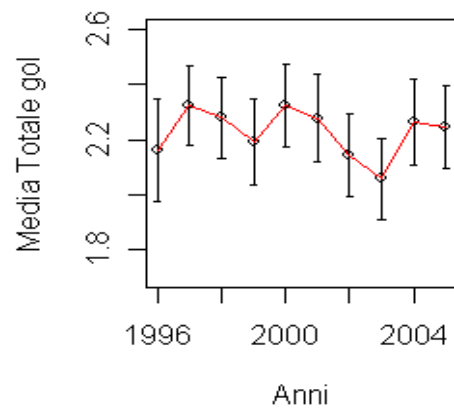
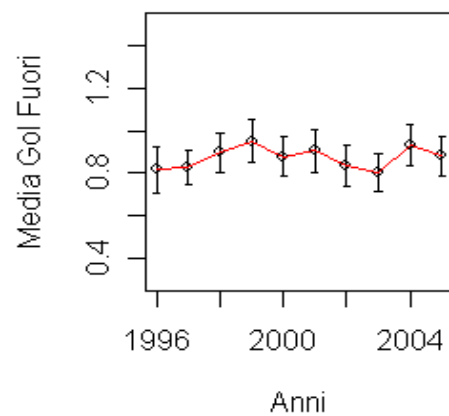
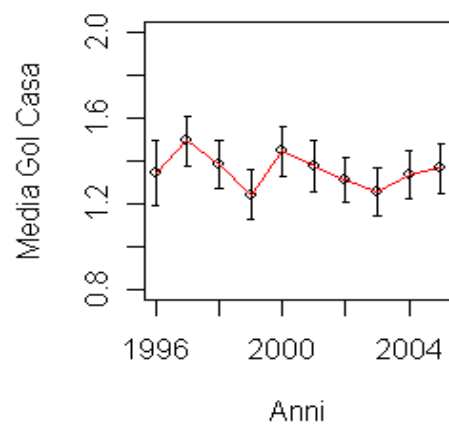
Si vede che le due curve hanno più o meno la stessa struttura; se avessimo adattato un modello per la differenza, ci saremmo aspettati che la variabile Mesi risultasse significativa. Si intuisce subito da questo grafico che nella Serie A francese le squadre in casa segnano in media di più rispetto a quelle fuori casa.

1.2 CAMPIONATO FRANCESE SERIE B

1.2.1 Analisi esplorativa

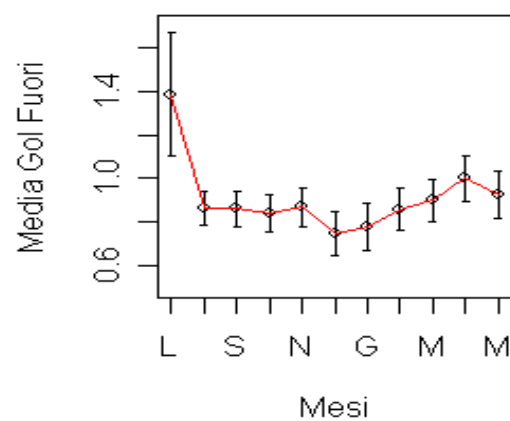
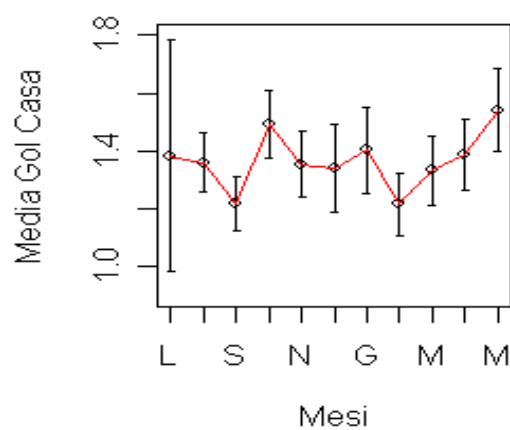
1.2.1.1 Primo approccio dell'analisi esplorativa

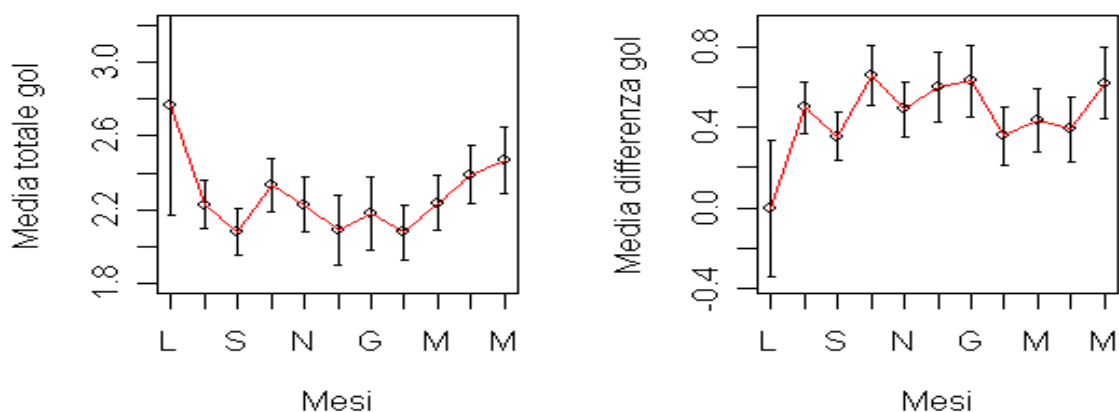
I grafici sotto esposti rappresentano gli intervalli di confidenza della media dei gol al variare dei mesi oppure al variare degli anni. Iniziamo con gli anni.



Guardando questi quattro grafici sembra non esserci una struttura particolare nei dati al variare degli anni.

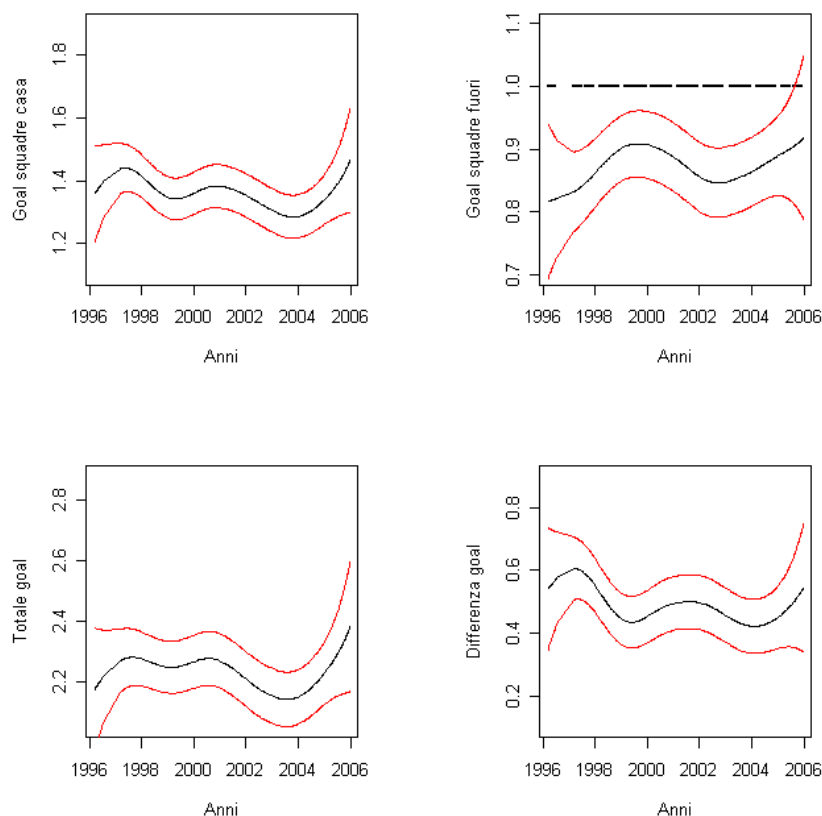
Struttura dei gol al variare dei mesi.





Come il campionato della Serie A ,anche la Serie B inizia in luglio e finisce in maggio. Se consideriamo la media dei gol per la squadre in casa sembra che non ci sia una struttura interessante, però se consideriamo quella delle squadre fuori casa e la media del totale di gol sembra che ci sia una struttura che decresce lentamente e poi cresce un po'; inoltre sembra che si segni di più all'inizio della stagione e di meno alla fine.

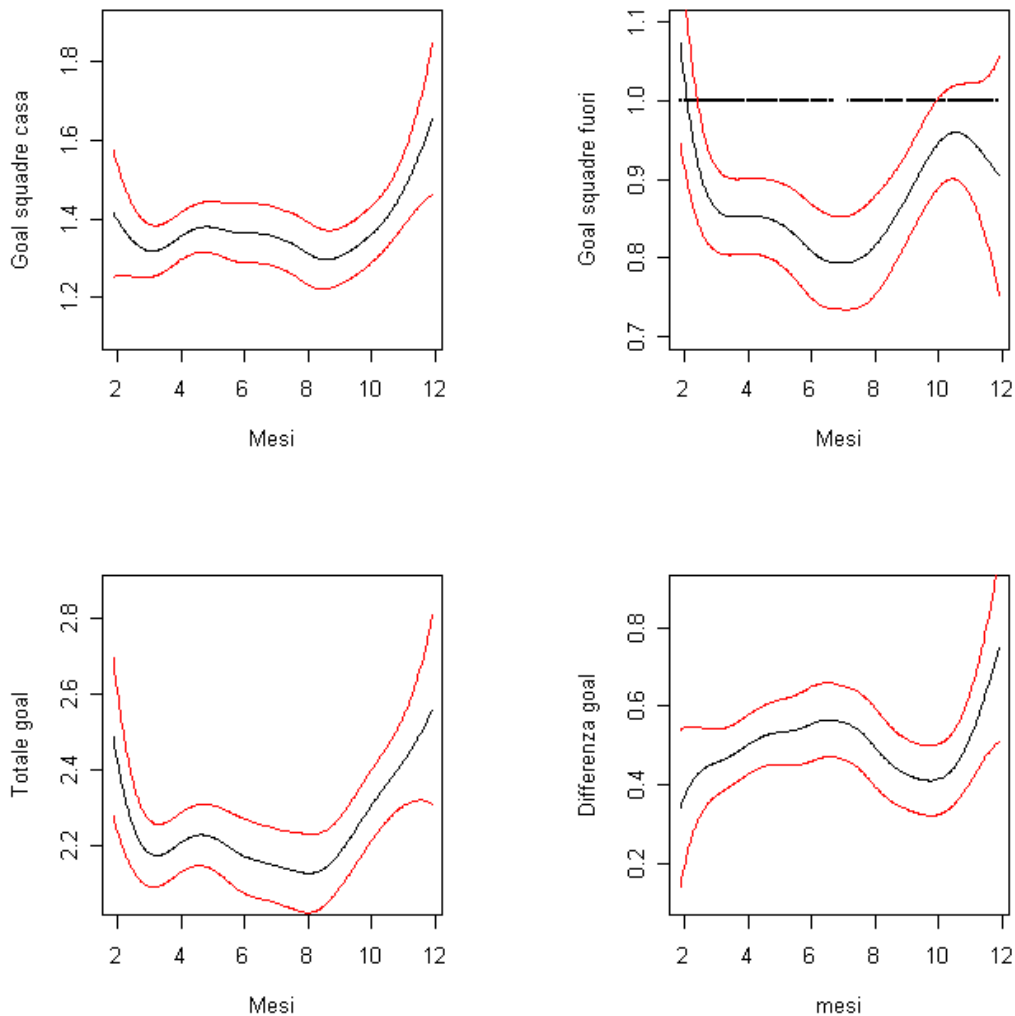
1.2.1.2 secondo approccio dell'analisi esplorativa



In questi grafici, se guardiamo l'andamento della curva e l'intervallo di confidenza, c'è un andamento che tende a crescere e decresce e, di conseguenza, sembra non esserci una struttura particolare al crescere degli anni.

Struttura dei gol al variare dei mesi.

Legenda Luglio='1',settembre='2',...,maggio='11'



Se consideriamo i gol delle squadre in casa e il totale di gol per partita sembra esserci una struttura interessante nei dati: dall'inizio della stagione l'andamento decresce fino a novembre, poi sembra essere costante fino a febbraio e infine cresce fino alla fine della stagione. Per quanto riguarda i gol delle squadre fuori casa non risulta facile cogliere la struttura perché all'inizio

scende lentamente fino a febbraio e poi cresce fino ad aprile e decresce nuovamente fino alla fine della stagione.

1.2.2 Adattamento dei modelli per i nostri dati

1.2.2.1 Modello per i gol delle squadre in casa

Call: `gam(formula = golC ~ s(Mesi) + s(Anni), family = poisson, data = dat1)`

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.8032	-1.5834	-0.2908	0.5406	3.4126

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4375.062 on 3761 degrees of freedom

Residual Deviance: 4354.109 on 3753 degrees of freedom

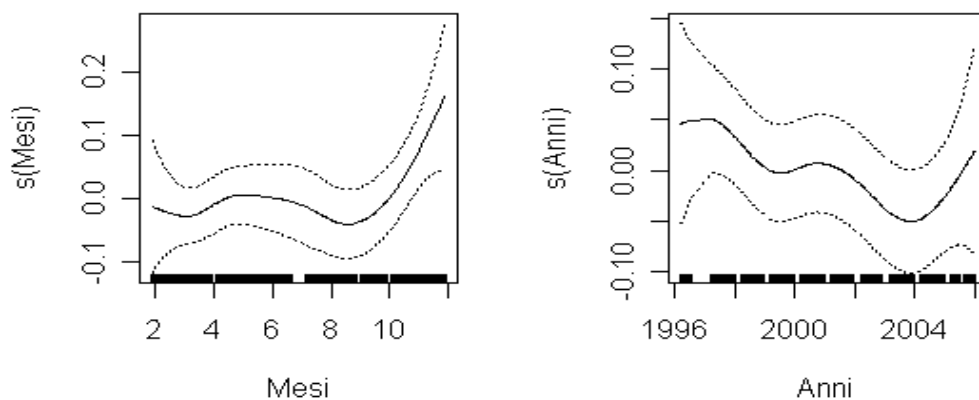
AIC: 11162.08

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3	9.9993	0.0186		
s(Anni)	1	3	5.4195	0.1435		

Questi grafici sotto esposti sono le curve che sono state stimate.



Più o meno la struttura sembra quella che abbiamo visto in precedenza nell'analisi esplorativa.

Dall'output di sopra risulta che al livello 1% nessuna delle due variabili è significativa, però al livello 5% la variabile Mesi risulta significativa come ci aspettavamo dall'analisi esplorativa. Vediamo quanto va bene questo modello.

1-pchisq(4354.109,3753)
1.966560e-11

Il valore del livello di significatività osservato ci suggerisce che il modello saturo risulta preferibile al modello corrente. Proviamo a togliere la variabile non significativa Anni.

Call: gam(formula = golC ~ s(Mesi), family = poisson, data = dat1)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7853	-1.6105	-0.2994	0.5361	3.4687

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4375.062 on 3761 degrees of freedom
Residual Deviance: 4362.283 on 3757 degrees of freedom
AIC: 11162.25

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			11.0447	0.0115

Come al punto precedente al livello 1% la variabile mese non risulta significativa, mentre al livello 5% è significativa.

1-pchisq(4362.283,3757)
1.495459e-11

Il valore del p-value ci suggerisce che il modello stimato non spiega altrettanto bene la variabilità dei dati quanto il modello saturo. Quindi, se ci fermiamo a questo modello, possiamo dire che non è adatto a spiegare i nostri dati ma ci vorrebbero probabilmente altre variabili. Anche qui il problema della non adeguatezza del modello finale ottenuto può essere dovuto agli stessi motivi visti nei casi precedenti.

1.2.2.2 Modello per i gol delle squadre fuori casa

Call: gam(formula = golF ~ s(Mesi) + s(Anni), family = poisson, data = dat1)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4278	-1.2976	0.0681	0.2137	3.1550

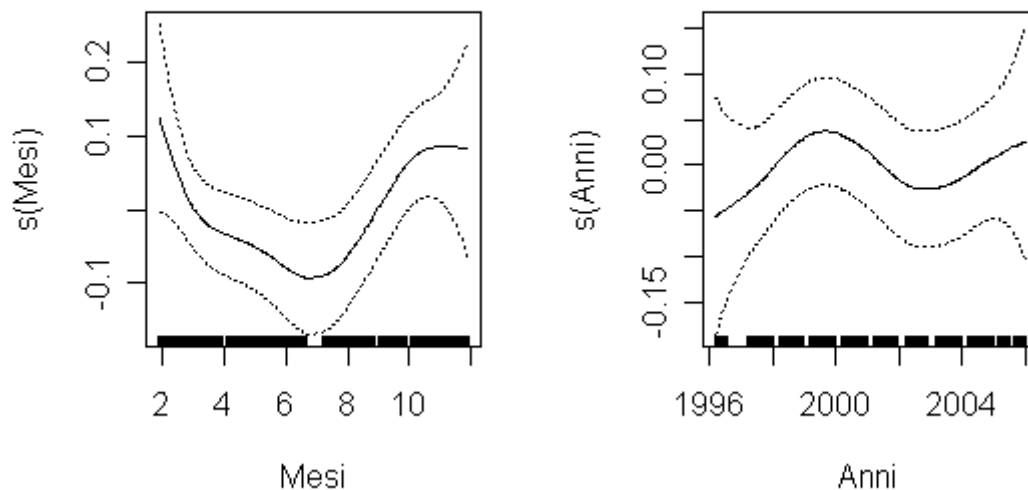
(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4217.881 on 3761 degrees of freedom
 Residual Deviance: 4195.814 on 3753 degrees of freedom
 AIC: 9213.173

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df Npar	Chisq	P(Chi)
(Intercept)	1				
s(Mesi)	1	3	16.4669	0.0009	
s(Anni)	1	3	3.3568	0.3398	



Questi grafici ci fanno vedere l'andamento dei gol delle squadre fuori casa rispettivamente al variare dei mesi e degli anni. Abbiamo la stessa informazione che abbiamo ricavato nell'analisi esplorativa. Sull'output di sopra vediamo che al livello 5% risulta significativa solo la variabile Mesi più o meno ciò che ci aspettavamo dall'analisi esplorativa. Proviamo a togliere la variabile non significativa Anni.

Call: gam(formula = golF ~ s(Mesi), family = poisson, data = dat1)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.40309	-1.29655	0.06844	0.20582	3.15436

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4217.881 on 3761 degrees of freedom
 Residual Deviance: 4199.14 on 3757 degrees of freedom
 AIC: 9208.499

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			16.9549	0.0007

Rimane sempre fortemente significativa la variabile Mesi. Proviamo a vedere la bontà di questo modello basato sulla devianza residua.

1-pchisq(4199.14,3757)
 4.308054e-07

Il valore del livello di significatività ci suggerisce che il nostro modello stimato non spiega bene la variabilità dei dati quanto il modello saturo(il modello che associa un parametro ad ogni osservazione). In altre parole il nostro modello richiederebbe variabili in più; il problema della non adeguatezza può essere dovuto ai motivi specificati ai punti precedenti.

1.2.2.3 Modello per il totale dei gol per partita

Call: gam(formula = (golC + golF) ~ s(Mesi) + s(Anni), family = poisson,
 data = dat1)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2769	-0.9052	-0.1379	0.5231	3.9271

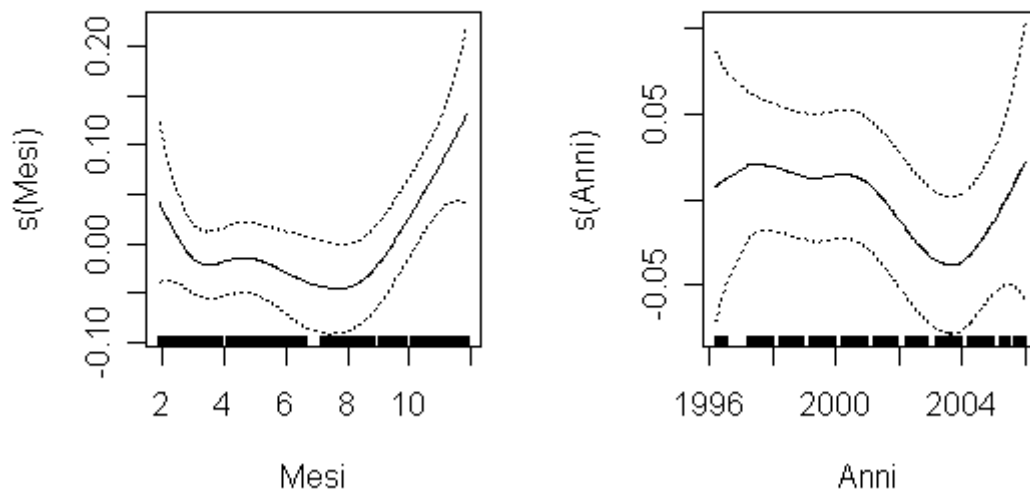
(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4518.319 on 3761 degrees of freedom
 Residual Deviance: 4492.653 on 3753 degrees of freedom
 AIC: 13447.32

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			15.8375	0.0012
s(Anni)	1	3			4.1175	0.2490



Dal grafico vediamo la stima della struttura dei gol al variare rispettivamente dei mesi e degli anni. Come visto anche nell'analisi esplorativa la struttura al variare degli anni non sembra interessante. Vediamo sull'output di sopra risulta significativa solo la variabile Mesi, come ci aspettavamo. Proviamo a togliere la variabile non significativa Anni.

Call: `gam(formula = (golC + golF) ~ s(Mesi), family = poisson, data = dat1)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2598	-0.9024	-0.1340	0.5172	3.8771

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 4518.319 on 3761 degrees of freedom
Residual Deviance: 4498.061 on 3757 degrees of freedom
AIC: 13444.73

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

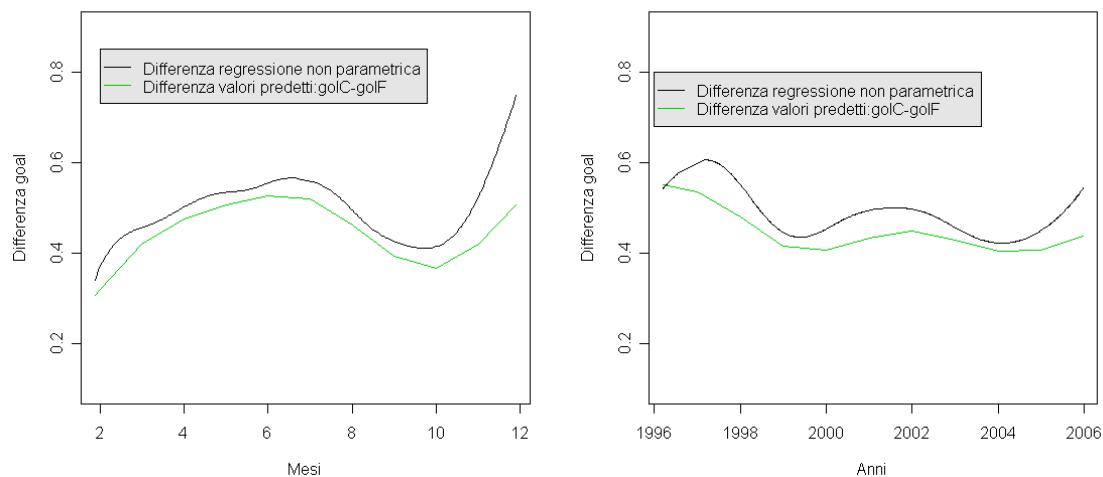
	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			16.7442	0.0008

La variabile Mesi rimane fortemente significativa. Vediamo quanto va bene il modello senza la variabile Anni.

1-pchisq(4498.061,3757)
3.330669e-16

Questo valore del livello di significatività ci suggerisce che il modello stimato non spiega bene la variabilità dei dati quanto il modello saturo ma richiede variabili in più. Anche in questo caso sorge lo stesso problema sottolineato negli altri casi precedenti.

1.2.2.4 Modello per la differenza dei gol per partita

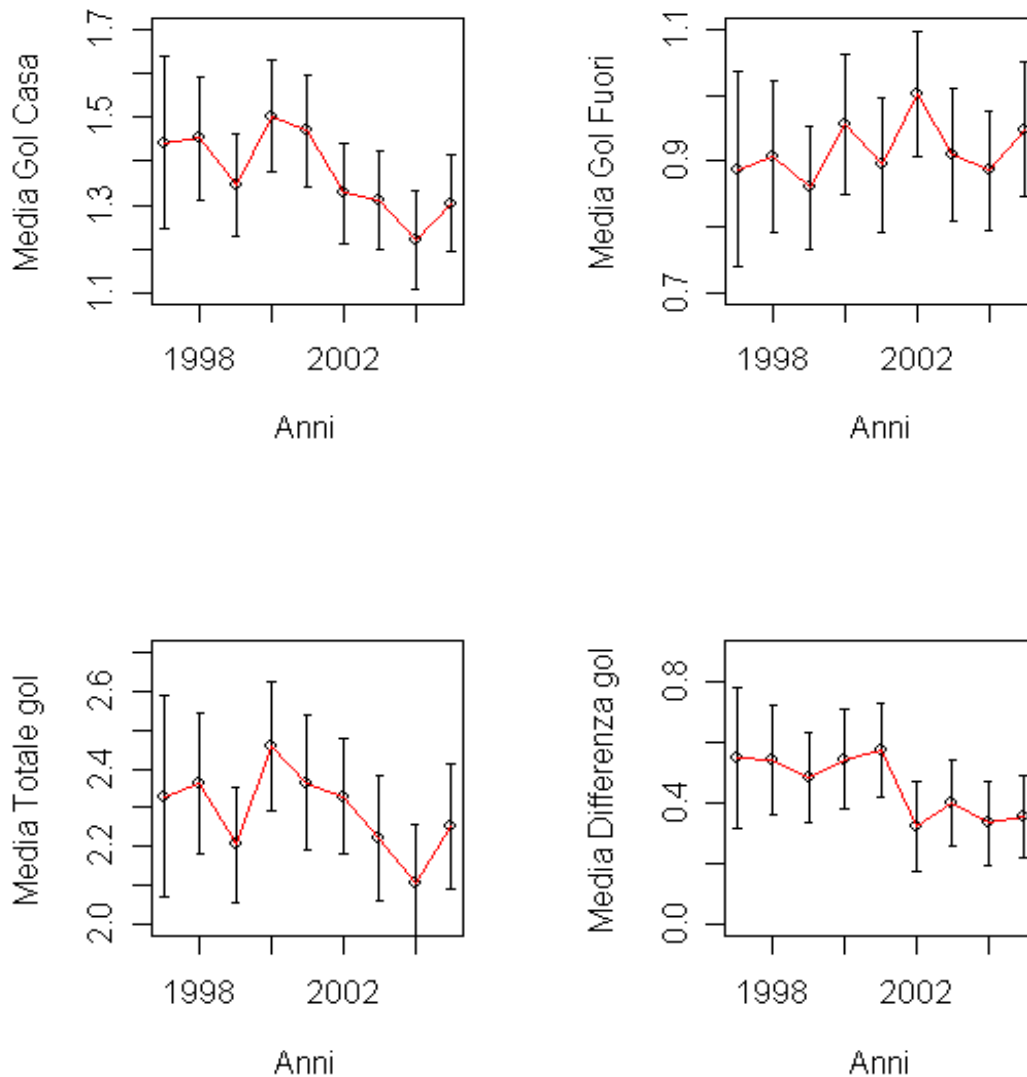


Secondo quanto spiegato su questo grafico, risulta che, se avessimo costruito un modello per la differenza di gol ci saremmo aspettati che la variabile Mesi fosse significativa, invece se guardiamo il grafico a destra sembra che al variare degli anni l'andamento è più o meno costante. Ci saremmo aspettati che l'anno non fosse una variabile significativa.

1.3 CAMPIONATO FRANCESE SERIE C

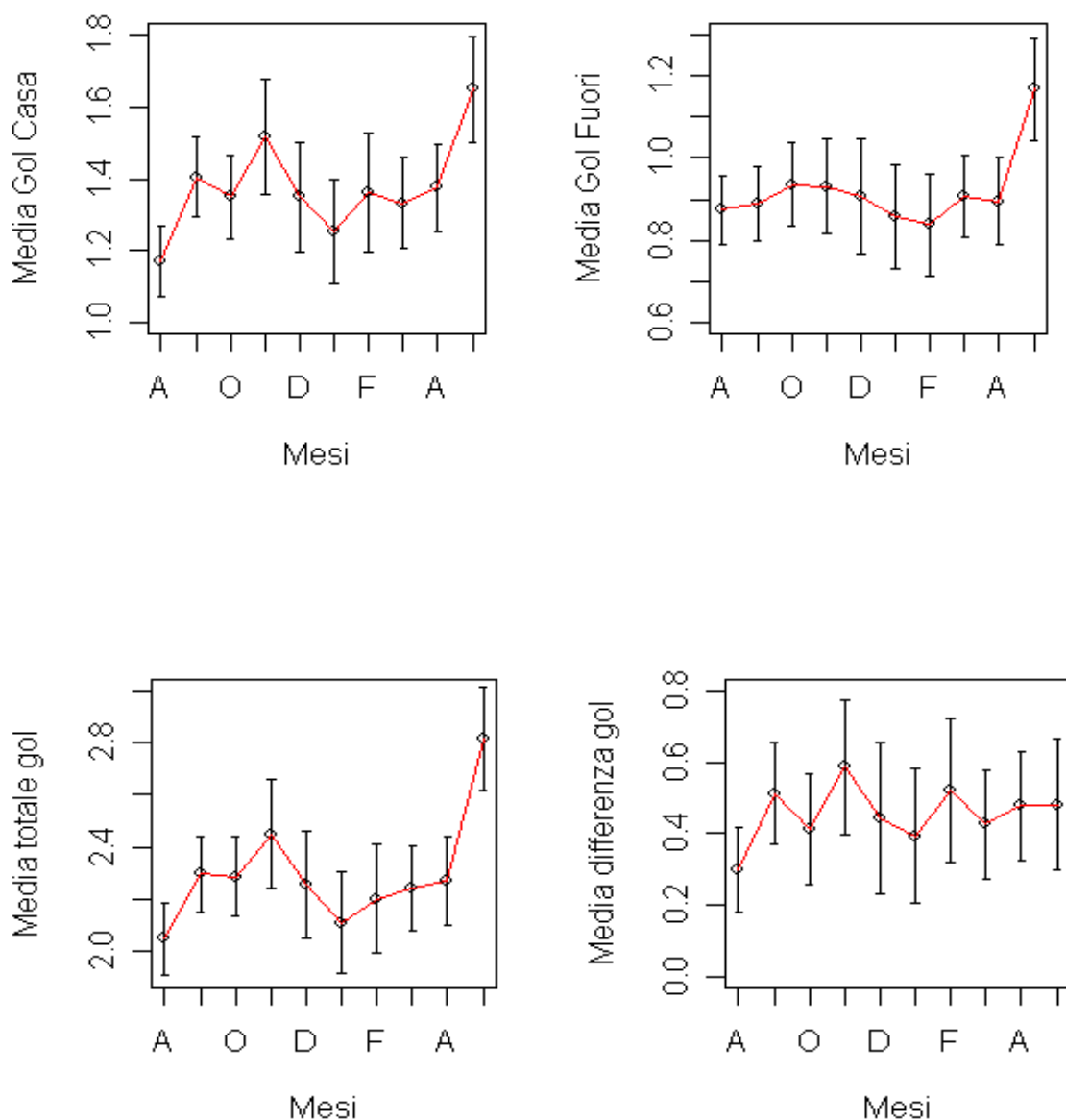
1.3.1 Analisi esplorativa

1.3.1.1 Primo approccio dell'analisi esplorativa



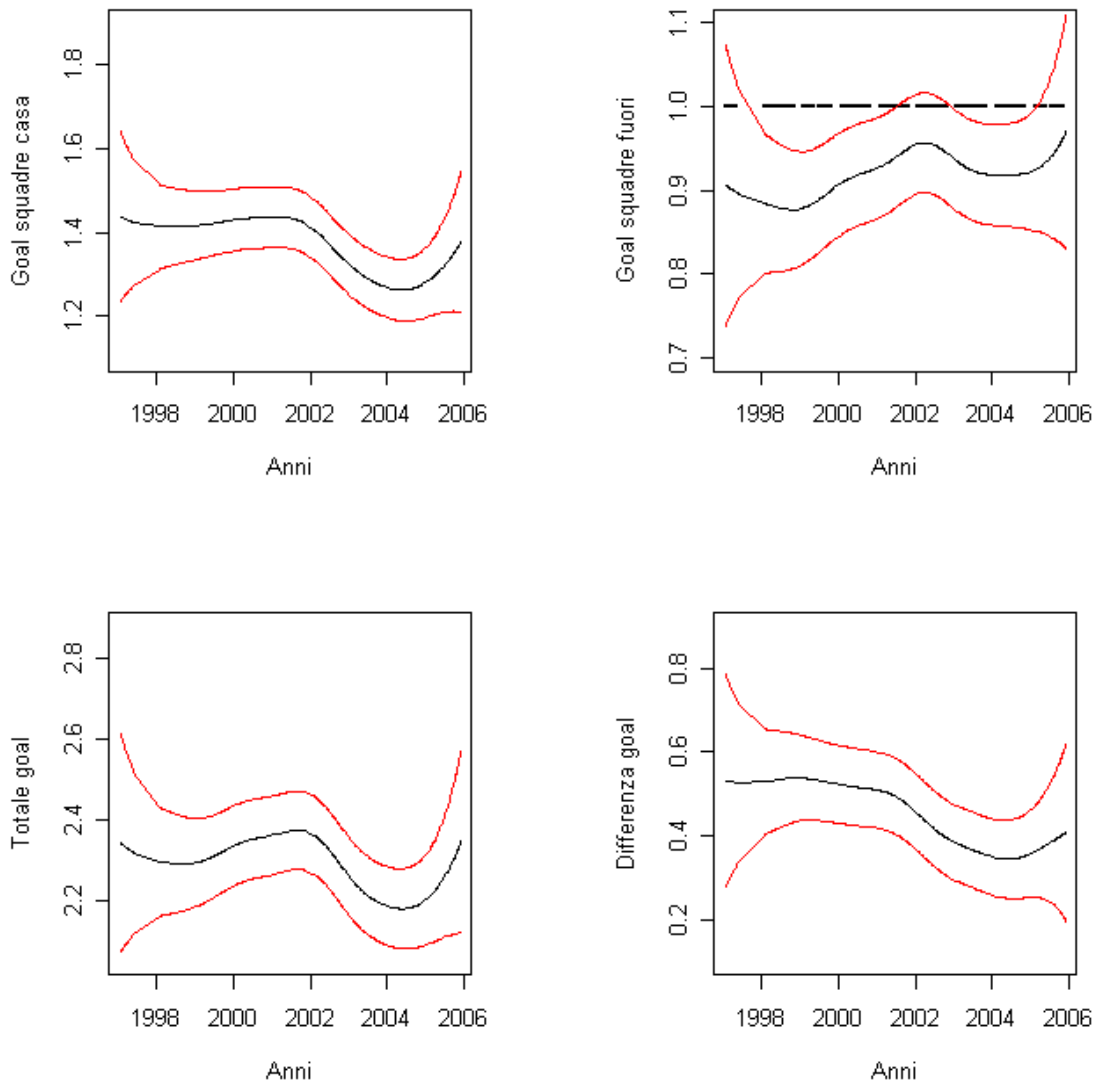
Su questi grafici se guardiamo la media dei gol per le squadre in casa e la media del totale dei gol tendenzialmente sembra esserci una struttura anche nel grafico della differenza al variare degli anni.

→ **Struttura dei gol al variare dei mesi**



Tranne il caso della differenza dei gol ,sembra esserci una struttura interessante nei dati sia per la media dei gol in casa, per la media dei gol fuori casa sia per la media del totale gol. Di preciso all’inizio si segna di meno poi cresce lentamente fino a novembre, infine scende fino a gennaio e ricresce fino a marzo, la fine della stagione.

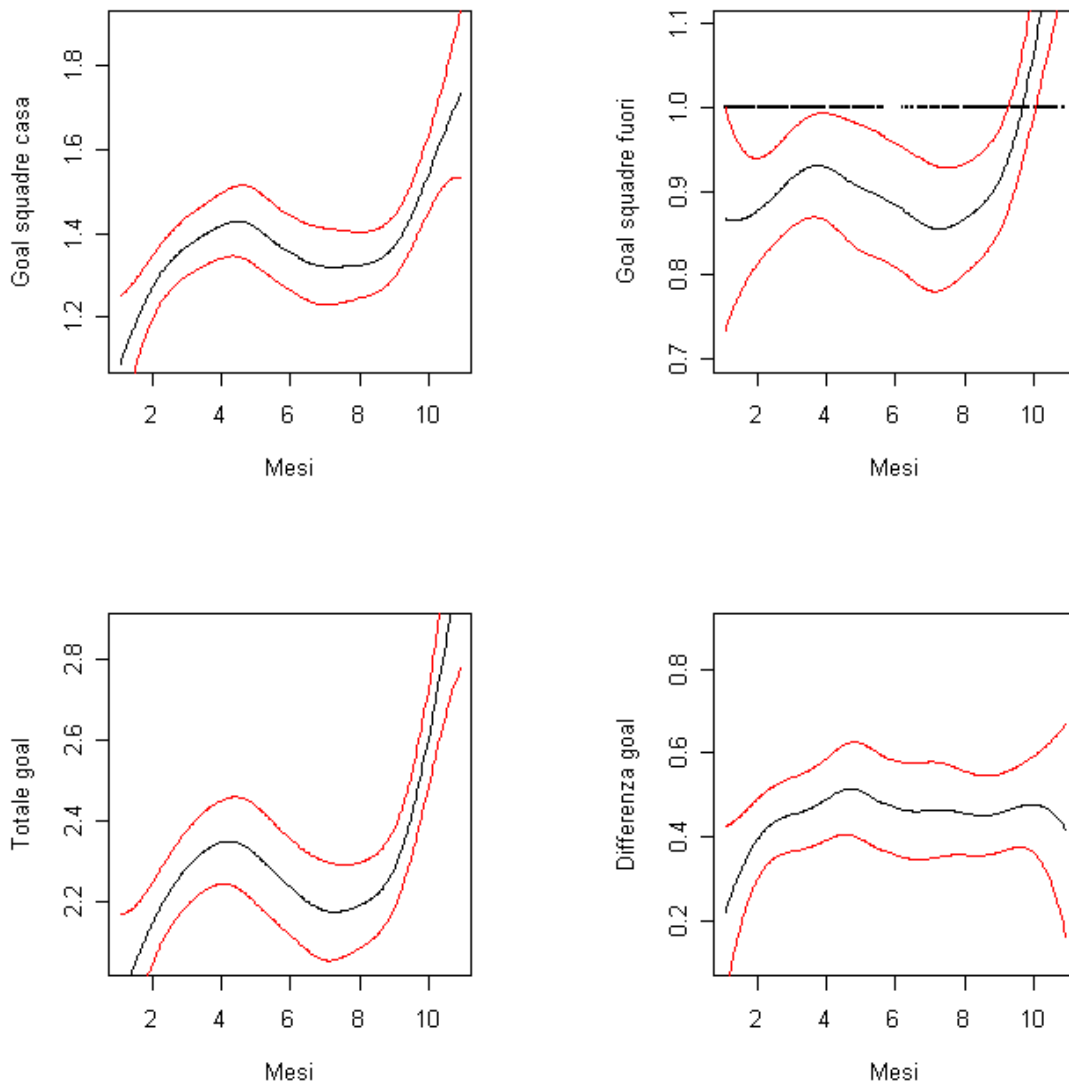
1.3.1.2 secondo approccio dell'analisi esplorativa



Se consideriamo sia il caso dei gol delle squadre in casa, sia il totale dei gol per partita, sia la differenza di gol per partita, tendenzialmente potrebbe esserci una struttura nei dati. Abbiamo che dal 1998 al 2002 l'andamento risulta più o meno costante poi decresce e ricresce lentamente fino al 2006. Questa struttura è quasi simile a quella vista al punto precedente con gli intervalli di confidenza della media dei gol.

→ **Struttura al variare dei mesi**

Legenda: 1='Agosto', 2='Settembre', ..., 10='Maggio'



Tranne il caso della differenza dei gol, se consideriamo sia il caso dei gol per la squadre in casa, sia i gol per le squadre fuori casa, sia il totale dei gol per partita, abbiamo una struttura particolare e simile nei tre casi, quasi uguali a quelli visti con gli intervalli di confidenza della media al punto precedente.

Si vede che dall'inizio fino a novembre l'andamento cresce poi scende fino a febbraio e infine ricresce fino a maggio, la fine della stagione. Si nota anche che all'inizio della stagione si segna di meno e di più alla fine.

1.3.2 adattamento dei modelli ai nostri dati

1.3.2.1 Modello per i gol delle squadre in casa

Call: `gam(formula = golC ~ s(Mesi) + s(Anni), family = poisson, data = dat2)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8871	-1.5037	-0.2638	0.5403	3.5459

(Dispersion Parameter for poisson family taken to be 1)

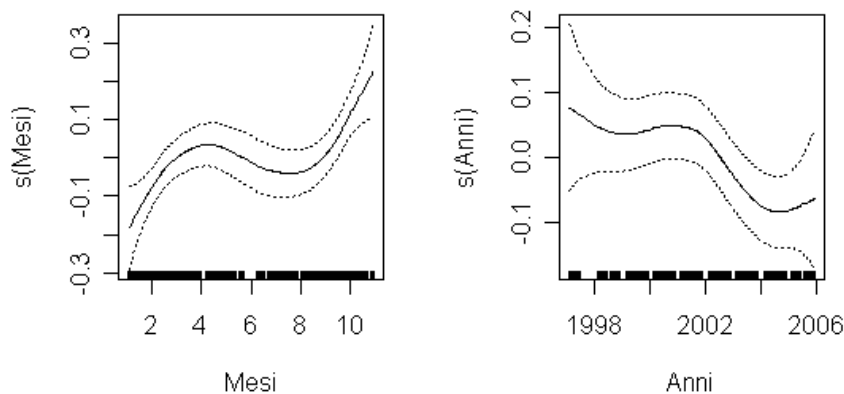
Null Deviance: 3657.449 on 3104 degrees of freedom
 Residual Deviance: 3612.402 on 3096 degrees of freedom
 AIC: 9238.497

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df Npar	Chisq	P(Chi)
(Intercept)	1				
s(Mesi)	1	3		19.8738	0.0002
s(Anni)	1	3		4.8447	0.1835

grafico delle curve che sono state stimate



E' quasi la stessa struttura di quella vista nell' analisi preliminare.

Sull'output di sopra, il valore del livello di significatività osservato ci suggerisce che al livello 5% la variabile Anni non risulta significativa. E' significativa sola la variabile Mesi. Togliamo la variabile Anni che non risulta significativa

Call: gam(formula = golC ~ s(Mesi), family = poisson, data = dat2)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8423	-1.5319	-0.2880	0.5349	3.4729

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 3657.449 on 3104 degrees of freedom

Residual Deviance: 3626.837 on 3100 degrees of freedom
AIC: 9244.932

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3	21.0021	0.0001		

Qui si nota che togliendo la variabile Anni il mese rimane lo stesso significativo al livello 5%. E' opportuno comunque vedere la sua bontà.

1-pchisq(3626.837,3100)
1.050300e-10

Il valore del livello di significatività ci suggerisce fortemente che il modello stimato non spiega bene la variabilità dei dati quanto il modello saturo. Il problema può essere dovuto ai motivi specificati nei casi precedenti. Diciamo che se ci fermiamo a questo modello, con i dati a disposizione per l'analisi non riusciremmo a migliorare la sua capacità predittiva.

1.3.2.2 Modello per i gol delle squadre fuori casa

Call: gam(formula = golF ~ s(Mesi) + s(Anni), family = poisson, data = dat2)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5685	-1.3273	0.0689	0.1698	3.6331

(Dispersion Parameter for poisson family taken to be 1)

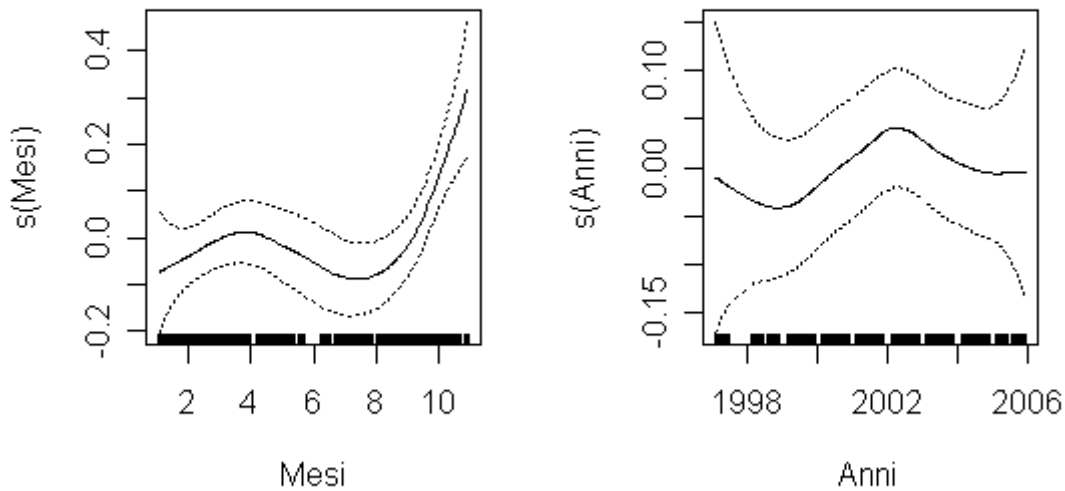
Null Deviance: 3626.789 on 3104 degrees of freedom
Residual Deviance: 3597.632 on 3096 degrees of freedom
AIC: 7853.589

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3	21.3809	0.0001		
s(Anni)	1	3	2.7044	0.4395		

Grafico delle curve che sono state stimate



Questi grafici sono simili a quelli visti nell' analisi esplorativa. Per quanto riguarda la significatività delle variabili, l'unica variabile significativa al 5% risulta essere il mese. Vediamo comunque la sua bontà.

1-pchisq(3597.632,3096)
6.468045e-10

Il valore del livello di significatività osservato ci suggerisce fortemente il rifiuto dell'ipotesi che il modello stimato spieghi la variabilità quanto il modello saturo. togliendo la variabile non significativa, come nei casi precedenti la variabile mese rimane significativa al livello 5%. E' opportuno vedere quanto è buono il modello a cui siamo arrivati.

1-pchisq(3600,3100)
7.429998e-10

Questo valore ci suggerisce fortemente che il modello stimato non spiega bene i dati rispetto al modello saturo. Ciò vuol dire che con i dati a disposizione abbiamo stimato un modello le cui le previsioni non sono affidabili. Il problema della non adeguatezza ,sarebbe dovuto ai motivi specificati nei altri casi precedenti.

1.3.2.3 Modello per il totale dei gol per partita

Call: `gam(formula = golC + golF ~ s(Mesi) + s(Anni), family = poisson, data = dat2)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4488	-0.9160	-0.1422	0.5315	3.4541

(Dispersion Parameter for poisson family taken to be 1)

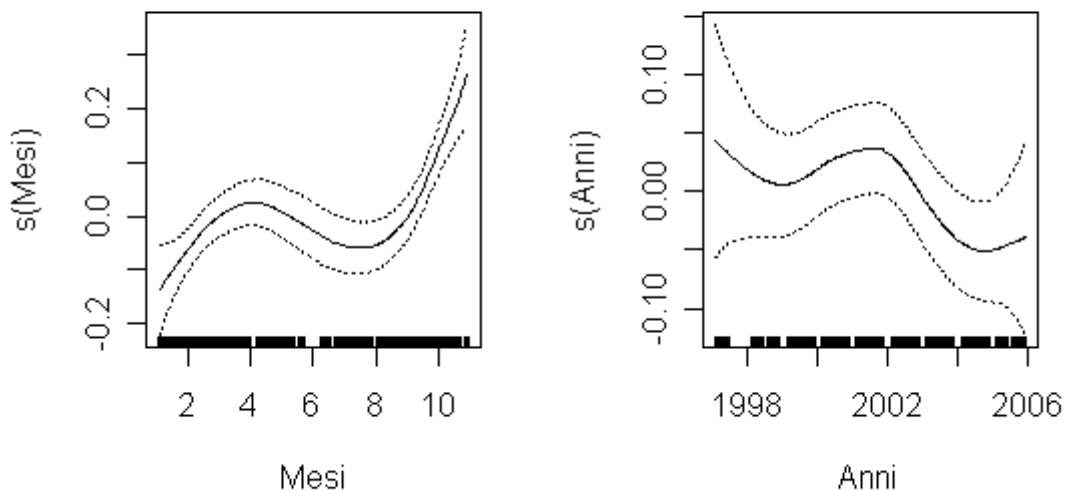
Null Deviance: 3854.512 on 3104 degrees of freedom
 Residual Deviance: 3791.731 on 3096 degrees of freedom
 AIC: 11261.20

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df Npar	Chisq	P(Chi)
(Intercept)	1				
s(Mesi)	1	3	37.273	4.028e-08	
s(Anni)	1	3	6.202	0.102	

grafico delle curve stimate



Da questi due grafici si vede una struttura interessante del totale di gol al variare dei mesi, mentre il grafico fatto rispetto al variare degli anni lo è molto di meno. Per quanto riguarda la significatività delle variabili, al livello 5% risulta fortemente significativa la variabile mese. Questo risultato è esattamente ciò che vediamo sul grafico ed anche ciò che ci aspettavamo dall'analisi esplorativa. Vediamo quanto risulta buono il modello stimato.

```
1-pchisq(3791.731,3096)
1.110223e-16
```

Il valore del livello di significatività osservato ci suggerisce che il modello stimato richiede altri parametri e di conseguenza altre variabili. E' opportuno togliere la variabile non significativa e vedere cosa succede al nuovo modello.

```
Call: gam(formula = golC + golF ~ s(Mesi), family = poisson, data = dat2)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4183 -0.9138 -0.1398  0.5277  3.3781
```

(Dispersion Parameter for poisson family taken to be 1)

```
Null Deviance: 3854.512 on 3104 degrees of freedom
Residual Deviance: 3802.024 on 3100 degrees of freedom
AIC: 11263.50
```

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

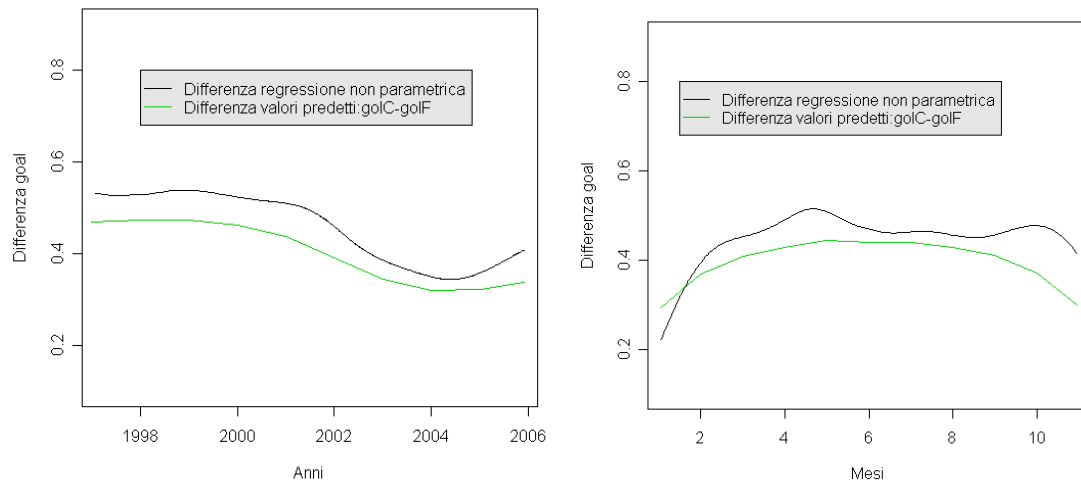
	Df	Npar	Df Npar	Chisq	P(Chi)
(Intercept)	1				
s(Mesi)	1	3		37.419	3.752e-08

Come nei casi precedenti il mese rimane sempre fortemente significativo. Risulta opportuno vedere quanto spiega i dati.

```
1-pchisq(3802.024,3100)
0
```

Il valore del livello di significatività osservato ci suggerisce che il modello stimato non spiega bene la variabilità dei dati quanto il modello saturo. Siamo di nuovo di fronte allo stesso problema incontrato nei precedenti casi. Diciamo solo che con le variabili a disposizione per l'analisi, non siamo in grado di migliorare la bontà del modello a cui siamo arrivati; nella speranza di ottenere le previsioni fidabili.

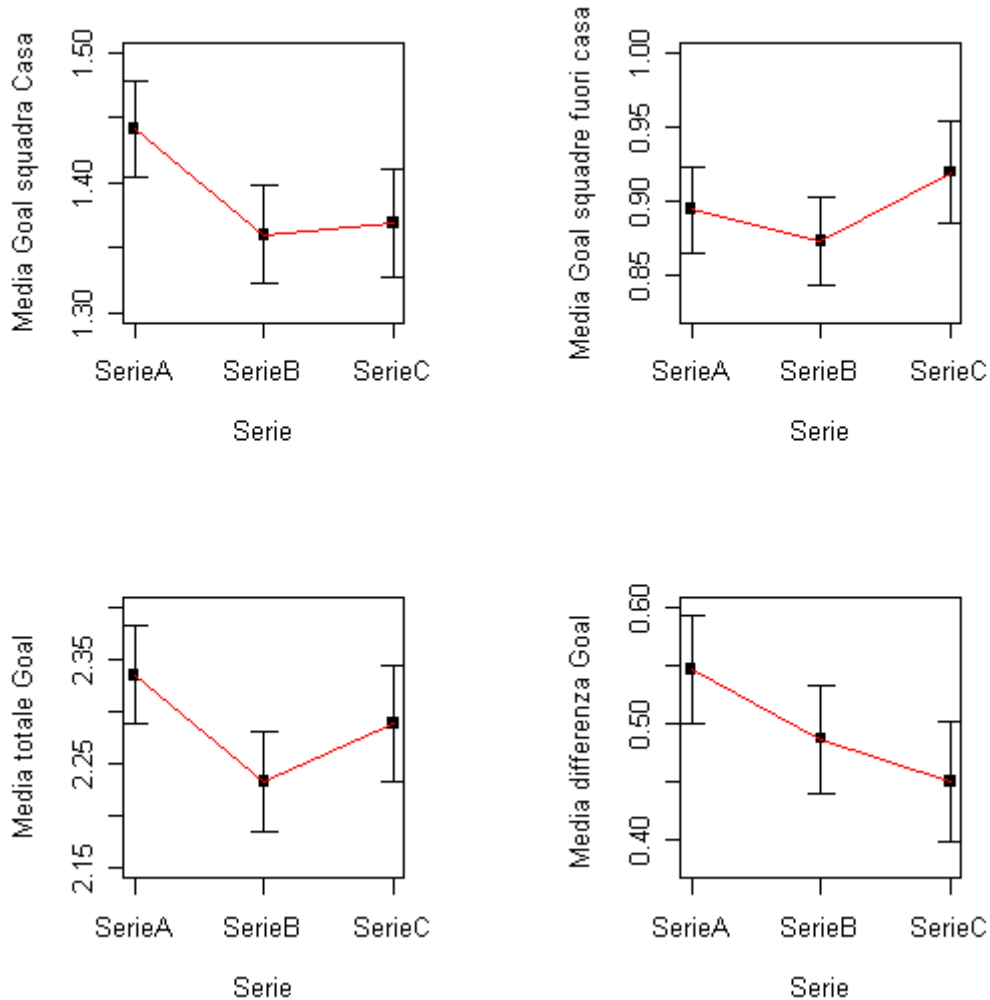
1.3.2.4 Modello per la differenza dei gol per partita



Su questi due grafici la curva della stima della differenza dei gol è simile a quella ottenuta con la regressione non parametrica, mentre se avessimo adattato un modello per la differenza dei gol le due variabili non sarebbero state significative.

1.3.3 Differenza tra le varie serie.

1.3.3.1 *prima analisi preliminare.*



Questi quattro grafici ci permettono di confrontare le varie Serie, in particolare se consideriamo il grafico in alto a sinistra, cioè il grafico della media dei gol delle squadre in casa al variare della categoria, salta fuori subito che in Serie A in media, si segnerebbe di più rispetto alla Serie B e C in cui si segna in media lo stesso numero di gol per partita, di conseguenza diciamo che il numero di gol segnato dalle squadre in casa è influenzato dalla Serie.

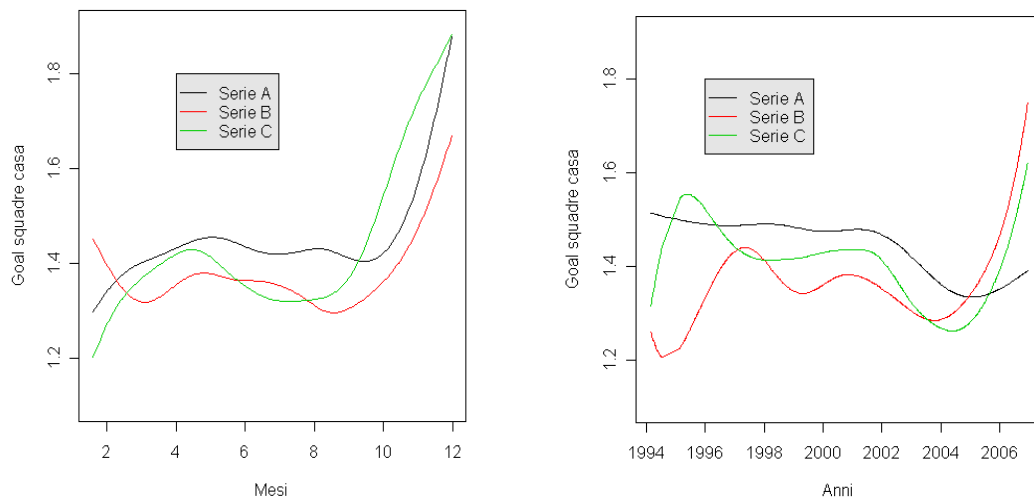
Il secondo grafico in alto a destra ci fa vedere come i tre intervalli di confidenza si sovrappongono, sembrerebbe che in media le squadre fuori segnano lo stesso numero di gol per partita, quindi in questo caso ci aspettiamo dopo aver adattato

un modello che i gol fatti dalle squadre fuori casa non siano influenzati dalla categoria.

Il terzo grafico in basso a sinistra si suggerisce che in media complessivamente si segnerebbe di più in Serie A rispetto alla Serie B, sembrerebbe anche che la categoria non ha un forte effetto sul totale dei gol per partita.

Il quarto grafico in basso a destra ci suggerisce che in media la differenza dei gol per partita è maggiore in Serie A rispetto alla Serie C, anche qui ci aspettiamo che la categoria non abbia un forte effetto sulla differenza dei gol per partita.

1.3.3.2 seconda analisi preliminare.



In questo caso stiamo facendo un'analisi esplorativa marginale. Si vede sul grafico in alto a sinistra che al crescere dei mesi la struttura è quasi la stessa tranne nel caso della curva rossa che rappresenta la serie B. Essa sembra leggermente diversa all'inizio della stagione. Per quanto riguarda gli anni, al crescere di questi la struttura è diversa.

1.3.4 Modelli di analisi delle differenze tra le serie indipendentemente dal mese e dell'anno.

1.3.4.1 Caso specifico dei gol delle squadre in casa per tutte le Serie messe insieme

Call: `gam(formula = golC ~ Serie, family = poisson, data =`

franciabis)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6979	-1.6492	-0.3240	0.5042	3.7821

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 13034.01 on 11096 degrees of freedom

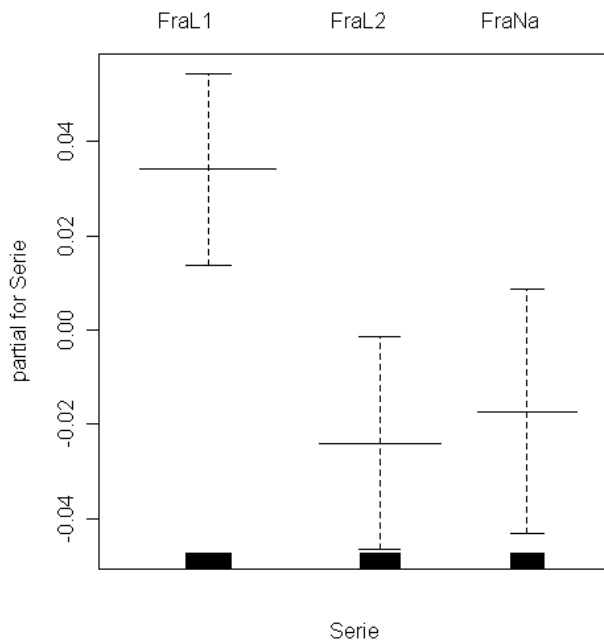
Residual Deviance: 13022.72 on 11094 degrees of freedom

AIC: 33323.13

Number of Local Scoring Iterations: 5

DF for Terms

	Df
(Intercept)	1
Serie	2



Entrambi gli intervalli di confidenza non sono sovrapposti, ci aspettiamo che ci siano differenze tra le varie Serie. Si può notare che nel campionato francese sembrerebbe che in media si segni di più in Serie A rispetto alle altre due Serie. Effettuiamo un test per vedere se ci sono differenze tra le varie Serie.

Analysis of Deviance Table

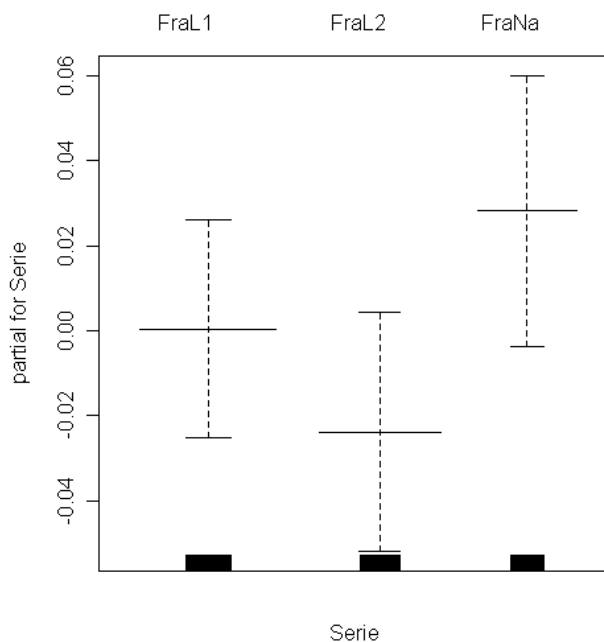
Model 1: golC ~ 1

Model 2: golC ~ Serie

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	11096	13034.0			
2	11094	13022.7	2	11.3	0.003545

Dal valore del livello di significatività risulta che la variabile Serie è significativa, cioè che, al variare della Serie, i gol segnati dalle squadre in casa variano.

1.3.4.2 Caso specifico dei gol delle squadre fuori casa per tutte le Serie messe insieme



Su questo grafico sembrerebbe esserci un effetto della Serie sui gol fatti dalle squadre fuori casa, cioè che, al cambio della Serie, i gol segnati dalle squadre fuori casa variano. Si nota anche che la Serie C risulta essere quella in cui in media si segna di più. E' opportuno fare un test per vedere se ci sono differenze tra le varie Serie o meno.

Analysis of Deviance Table

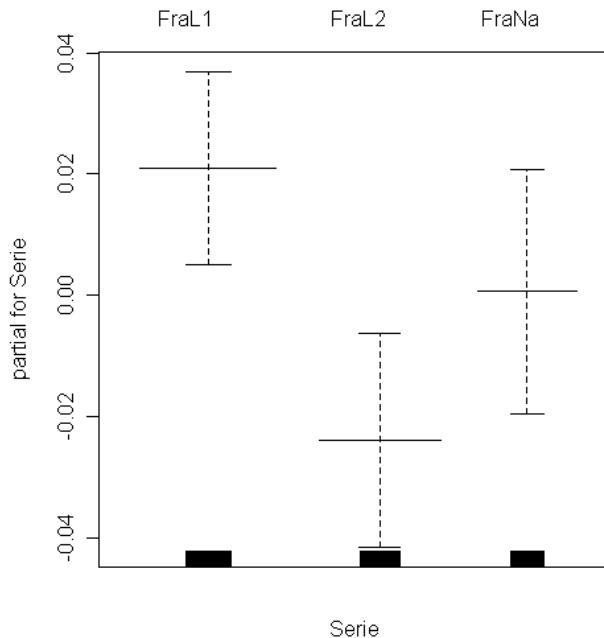
Model 1: $\text{golF} \sim 1$

Model 2: $\text{golF} \sim \text{Serie}$

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	11096	12817.6			
2	11094	12813.5	2	4.1	0.1

Il valore del livello di significatività ci suggerisce che la variabile Serie risulta significativa ma non tanto forte. Quindi se consideriamo i gol delle squadre fuori casa risulta che ci sono differenze al variare della Serie.

1.3.4.3 Caso specifico del totale dei gol per partita



Da questo grafico abbiamo che il cambiamento della serie influenza il totale di gol durante una partita. Anche qui è opportuno effettuare un test per evidenziare se ci sono differenze tra le varie Serie .

Analysis of Deviance Table

Model 1: (golC + golF) ~ 1

Model 2: (golC + golF) ~ Serie

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	11096	13416.8			
2	11094	13407.6	2	9.2	0.009963

Il valore del livello di significatività ci suggerisce che è significativa la variabile Serie, cioè che indipendentemente dal mese e dall'anno ci sono differenze tra le varie Serie.

1.3.5 Modelli di analisi delle differenze tra le serie dato il mese e l'anno.

1.3.5.1 Caso specifico dei gol delle squadre in casa per tutte le Serie messe insieme

Call: `gam(formula = golC ~ s(Mesi) + s(Anni) + FraL1, family = poisson, data = francia)`
 Deviance Residuals:
 Min 1Q Median 3Q Max
 -1.9097 -1.5274 -0.2989 0.5171 3.8470

(Dispersion Parameter for poisson family taken to be 1)

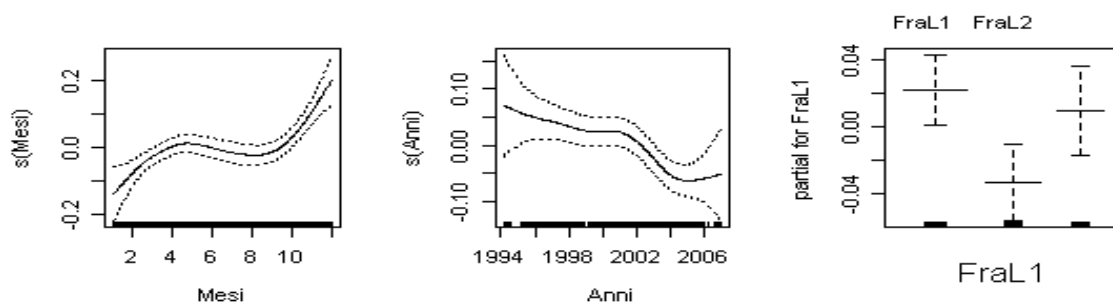
Null Deviance: 13034.01 on 11096 degrees of freedom
 Residual Deviance: 12946.83 on 11086 degrees of freedom
 AIC: 33263.23

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			29.8466	1.487e-06
s(Anni)	1	3			7.4905	0.0578
FraL1	2					

Queste sono le curve che vengono stimate dopo l'adattamento del modello.



Sui grafici sopra raffigurati, vediamo gli effetti marginali di ciascuna variabile, in particolare sul primo a sinistra si vede una forte struttura interessante dei gol delle squadre in casa al variare dei mesi, nel secondo si vede una qualche struttura rispetto agli anni. Infine, l'ultimo grafico a destra ci fa vedere che al

cambio della Serie cambiano anche i gol fatti dalle squadre in casa. Per quanto riguarda la significatività delle variabili, già dall'output di sopra si vede che al livello 5% la variabile Mesi è fortemente significativa e la variabile Anni è borderline cioè al limite. Proviamo a costruire un nuovo modello che ci aiuterà a sapere se risulta significativa o meno la variabile qualitativa che ci distingue le tre Serie.

Call: gam(formula = golC ~ s(Mesi) + s(Anni), family = poisson, data = francia)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8932	-1.5300	-0.2953	0.5140	3.8773

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 13034.01 on 11096 degrees of freedom

Residual Deviance: 12955.37 on 11088 degrees of freedom

AIC: 33267.78

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3	28.0378	3.567e-06		
s(Anni)	1	3	7.7866	0.0506		

→ confronto dei due modelli

Analysis of Deviance Table

Model 1: golC ~ s(Mesi) + s(Anni)

Model 2: golC ~ s(Mesi) + s(Anni) + FraL1

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	11088	12955.4			
2	11086	12946.8	2	8.5	0.01393

Il valore del livello di significatività ci suggerisce che al livello 1% ,dato il mese e l'anno, non ci sono differenze tra le Serie, invece al livello 5% ci suggerisce che è significativa la variabile qualitativa cioè che, dato il mese e l'anno, ci sono differenze tra la serie A, la Serie B e la Serie C del campionato francese. Vediamo quanto questo modello ci spiega i dati.

1-pchisq(12946.83,11086)

0

Il valore del p-value ci suggerisce che il modello non spiega altrettanto bene i dati quanto il modello saturo. Anche qui il problema della non adeguatezza, potrebbe essere dovuto agli stessi motivi specificati nei precedenti casi. Quindi con i dati che abbiamo a disposizione il modello stimato non risulta soddisfacente per le previsioni.

1.3.5.2 Caso specifico dei gol delle squadre fuori casa per tutte le Serie messe insieme

Call: gam(formula = golF ~ s(Mesi) + s(Anni) + Serie, family = poisson,
data = franciabis)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.47593	-1.31748	0.07415	0.19206	3.58508

(Dispersion Parameter for poisson family taken to be 1)

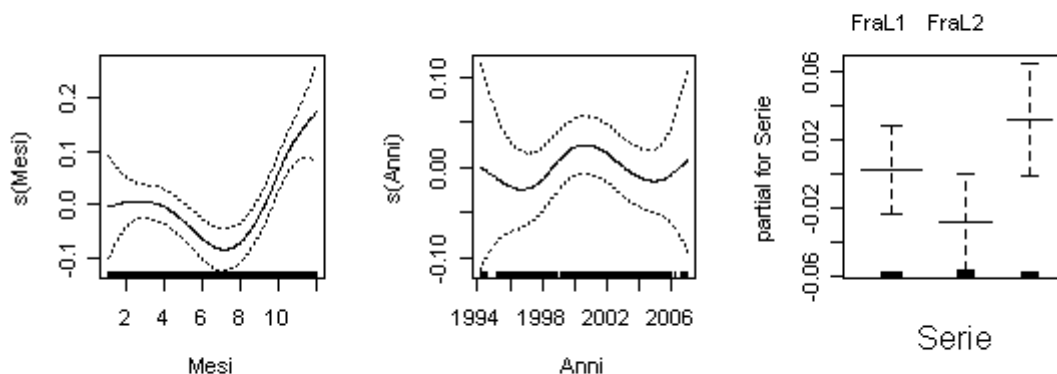
Null Deviance: 12817.58 on 11096 degrees of freedom
Residual Deviance: 12763.49 on 11086 degrees of freedom
AIC: 27656.31

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Df	Npar	Chisq	P(Chi)
(Intercept)	1					
s(Mesi)	1	3			39.051	1.694e-08
s(Anni)	1	3			5.014	0.171
Serie						

2



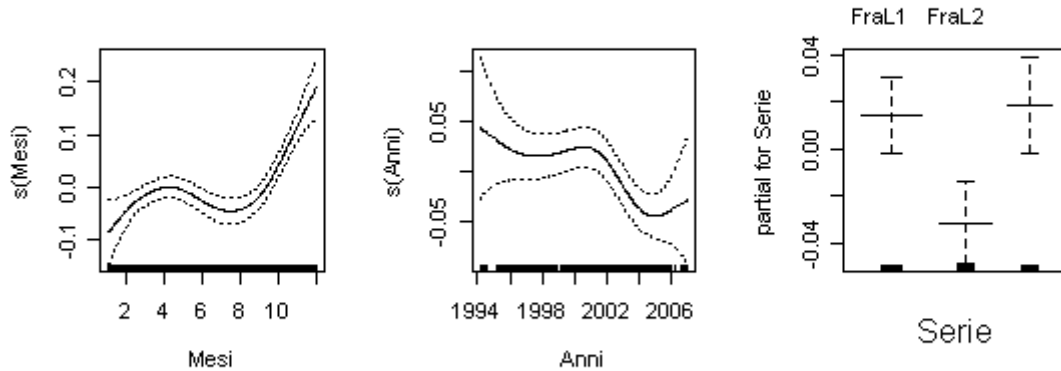
Dopo l'adattamento del modello si ottengono questi grafici che ci danno gli effetti marginali di ogni variabile. Unendo i dati si vede una struttura interessante dei gol per squadre fuori casa al variare dei mesi e abbiamo che dall'inizio della stagione fino a febbraio la curva decresce e poi comincia a crescere fino alla fine della stagione. La struttura rispetto agli anni non ci dice tanto. L'ultimo grafico sembra suggerirci che ci possiamo aspettare che ci siano differenze tra le varie serie e notiamo inoltre che le differenze non sono forti rispetto al grafico unico di prima in cui consideriamo l'effetto marginale della variabile qualitativa Serie che ci distingue le tre Serie. Tutto ciò si conferma guardando l'output di sopra che ci fa vedere che al livello approssimato 5% per la non significatività delle variabile risulta significativa solo il mese (proprio ciò che si vede su grafico). Risulta anche qui, opportuno effettuare un test per vedere se la variabile Serie è significativa o meno.

Analysis of Deviance Table**Model 1: $\text{golF} \sim \text{s(Mesi)} + \text{s(Anni)}$** **Model 2: $\text{golF} \sim \text{s(Mesi)} + \text{s(Anni)} + \text{Serie}$**

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	11088	12768.8			
2	11086	12763.5	2	5.3	0.1

Il valore del livello di significatività ci suggerisce che la variabile Serie (che ci distingue le tre Serie) risulta significativa. Ciò vuol dire che per quanto riguarda i gol delle squadre fuori casa, dato il mese o l'anno, ci sono differenze al cambio della Serie.

1.3.5.3 Caso specifico del totale dei gol per partita per tutte le Serie messe insieme



Questi sono i grafici che si ottengono dopo aver adattato il modello. Si vede che mettendo insieme i dati delle tre Serie abbiamo una struttura particolare del totale dei gol al crescere dei mesi: all'inizio sembrerebbe che nelle tre serie si segna di meno, poi cresce fino ottobre, poi decresce leggermente fino a febbraio e ricomincia a crescere fino alla fine delle stagioni. Per quanto riguarda l'anno è costante fino al 2002, decresce fino a 2005 e sembra ricominciare a crescere nel 2006. Guardando l'ultimo grafico in alto a destra ci aspettiamo che la Serie abbia un effetto sul totale di gol segnato per partita.

Call: `gam(formula = (golC + golF) ~ s(Mesi) + s(Anni) + Serie, family = poisson, data = franchiabis)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.3667	-0.9264	-0.1553	0.5136	3.9755

(Dispersion Parameter for poisson family taken to be 1)

Null Deviance: 13416.85 on 11096 degrees of freedom
Residual Deviance: 13310.63 on 11086 degrees of freedom
AIC: 40043.72

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

	Df	Npar	Chisq	P(Chi)
(Intercept)	1			
s(Mesi)	1	3	51.738	3.407e-11
s(Anni)	1	3	10.145	0.017

Serie 2

Si vede in questo output che al livello 5%, le due variabili risultano significative. Ciò è proprio quello che si vede sul grafico precedente. E' opportuno effettuare un test per vedere se dato il mese o l'anno, la variabile Serie risulta significativa.

Analysis of Deviance Table

Model 1: (golC + golF) ~ s(Mesi) + s(Anni)

Model 2: (golC + golF) ~ s(Mesi) + s(Anni) + Serie

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	11088	13323.3			
2	11086	13310.6	2	12.6	0.001815

Il valore del livello di significatività ci dice che la variabile Serie è significativa, cioè che, dato il mese o l'anno, il totale di gol per partita è influenzato dalla Serie.

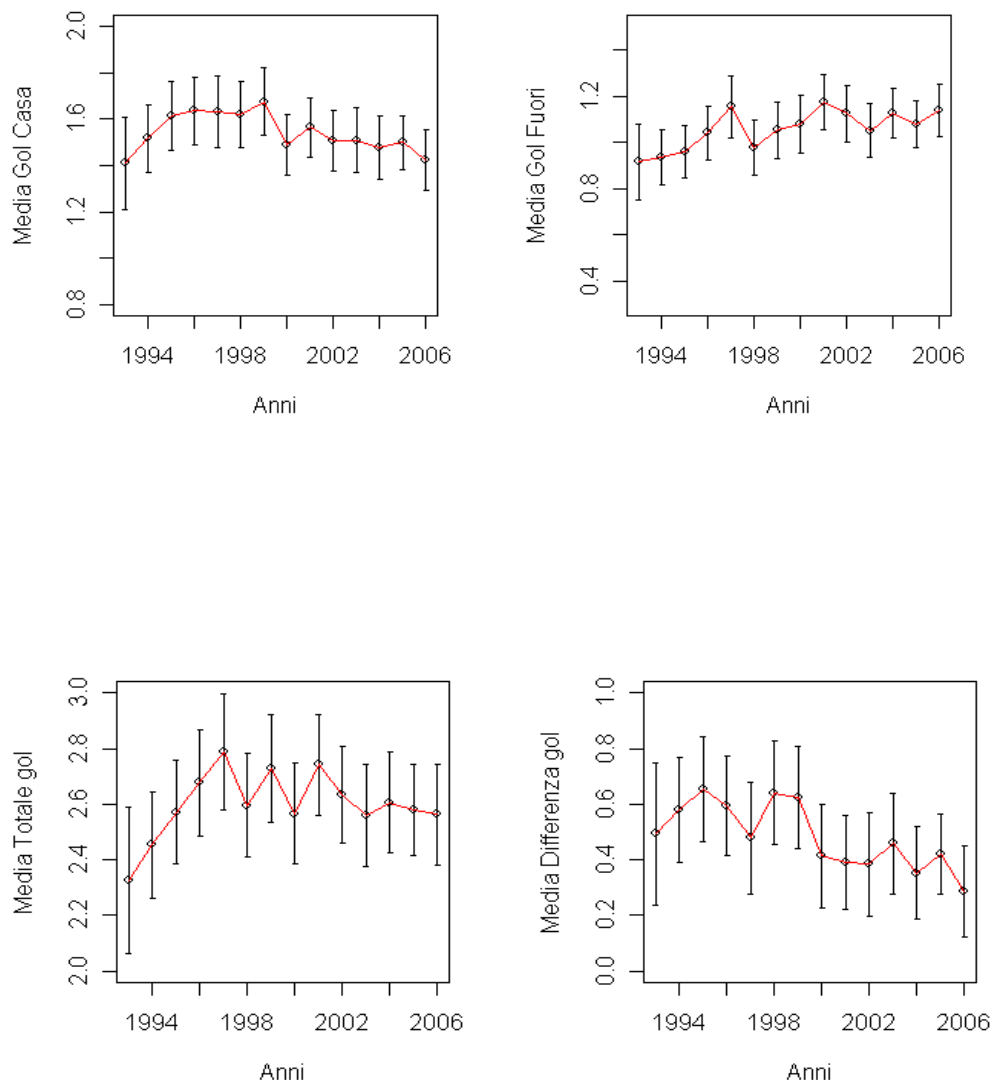
CAPITOLO 2: CAMPIONATO ITALIANO

2.1 CAMPIONATO ITALIANO SERIE A

2.1.1 Analisi esplorativa

2.1.1.1 Primo approccio dell'analisi esplorativa

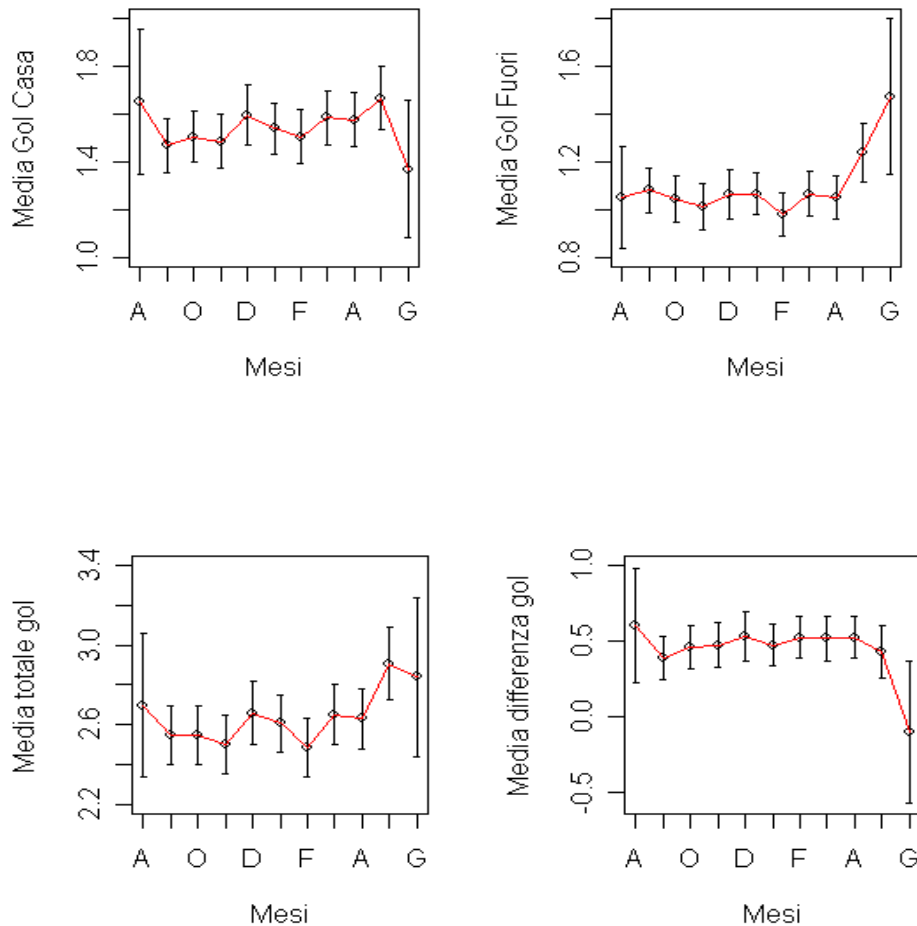
→ struttura della media dei gol al variare degli anni.



Guardando questi quattro grafici ,non si riesce a cogliere una struttura interessante della media dei gol al variare degli anni, ma se guardiamo bene il caso della media del totale di gol, il trend sembra crescere dal 1994 fino al 1997 e

poi prende un andamento fino all'ultimo anno; per gli altri tre casi il trend delle medie al variare degli anni non assume un andamento particolare.

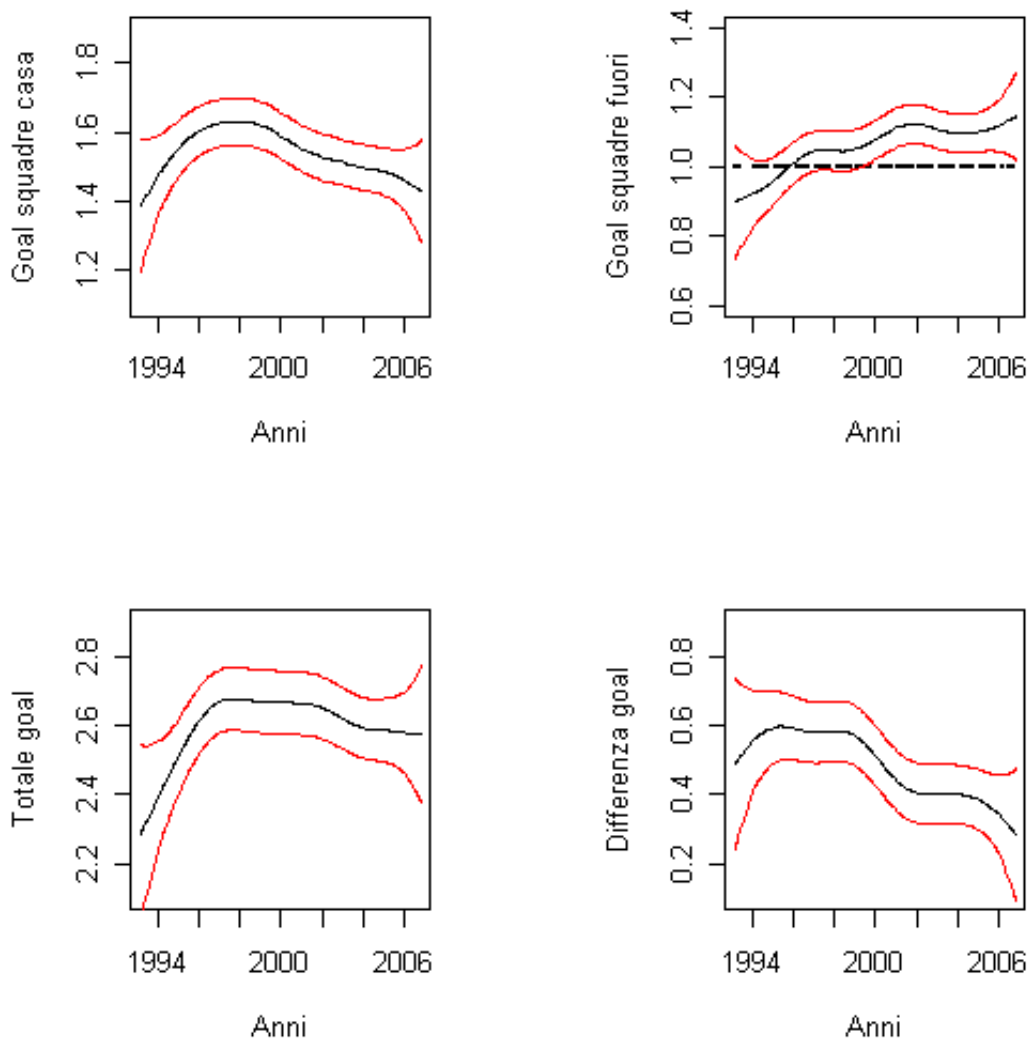
➔ **struttura della media dei gol al variare dei mesi.**



Anche se considerando il grafico della media dei gol al variare dei mesi sembra non esserci un andamento particolarmente interessante, si nota che la variabilità è molto grande all'inizio della stagione e alla fine. Però, considerando il grafico delle squadre fuori casa e del totale per partita vediamo una struttura interessante: dall'inizio della stagione fino a marzo l'andamento sembra costare, poi da aprile fino alla fine della stagione cresce; infine per quanto riguarda il grafico della media della differenza dei gol intuimmo subito il fatto che le squadre in casa segnano in media di più rispetto a quelle fuori casa, tranne che nell'ultimo mese della stagione.

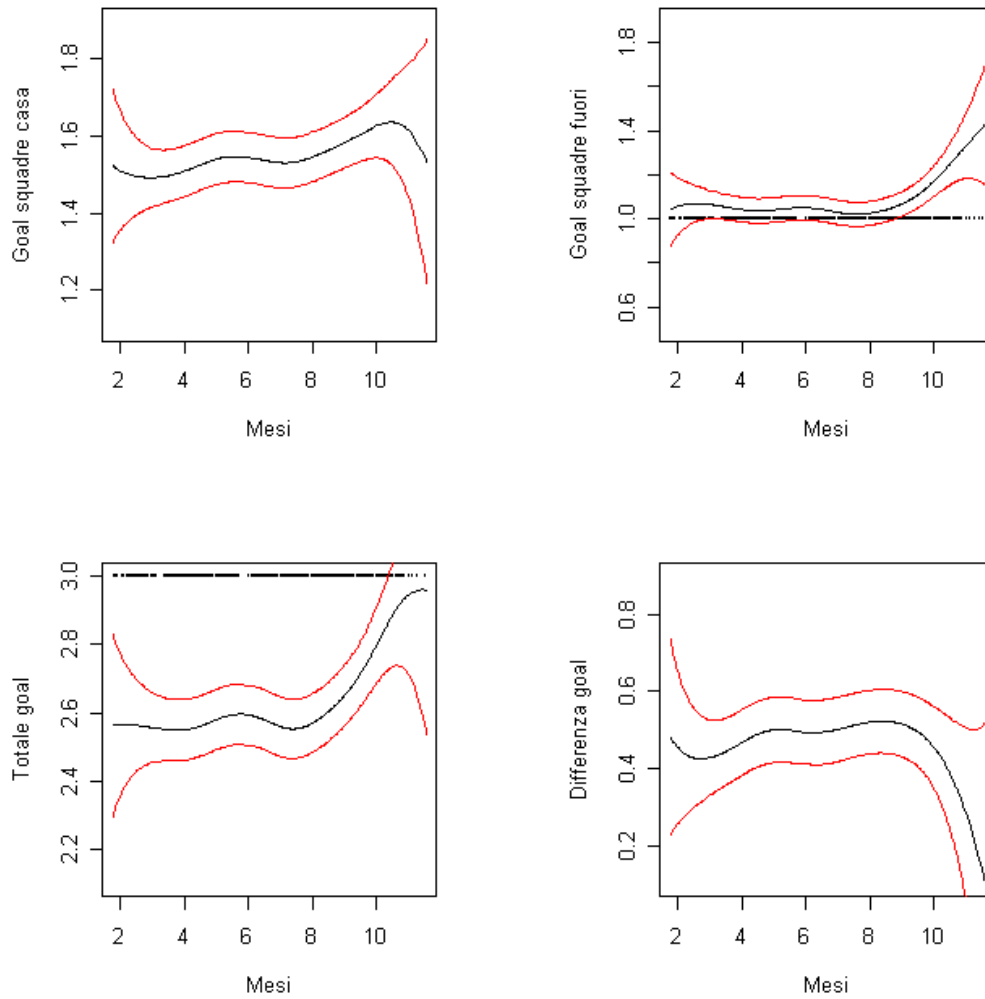
2.1.1.2 Secondo approccio dell'analisi esplorativa

→ **Struttura dei gol al variare degli anni.**



Se consideriamo i gol delle squadre in casa ed il totale dei gol per partita, sembra esserci un andamento particolare che cresce dal 1994 fino al 1998 e poi diventa più o meno costante. Per il caso dei gol delle squadre fuori casa, vediamo un andamento che cresce lentamente; infine per la differenza dei gol, l'andamento sembra essere costante nei primi anni e poi inizia a scendere lentamente fino al 2006.

→ **Struttura dei gol al variare dei mesi.**



Con i gol delle squadre in casa abbiamo un andamento che cresce lentamente fino alla fine di marzo e poi comincia a scendere fino alla fine della stagione. Considerando i gol delle squadre fuori casa e il totale dei gol per partita sembra esserci una struttura interessante che risulta costante dall'inizio della stagione fino a febbraio e poi cresce fino alla fine della stagione. Per il caso della differenza dei gol, la struttura sembra essere costante fino a marzo e poi decresce fino alla fine della stagione.

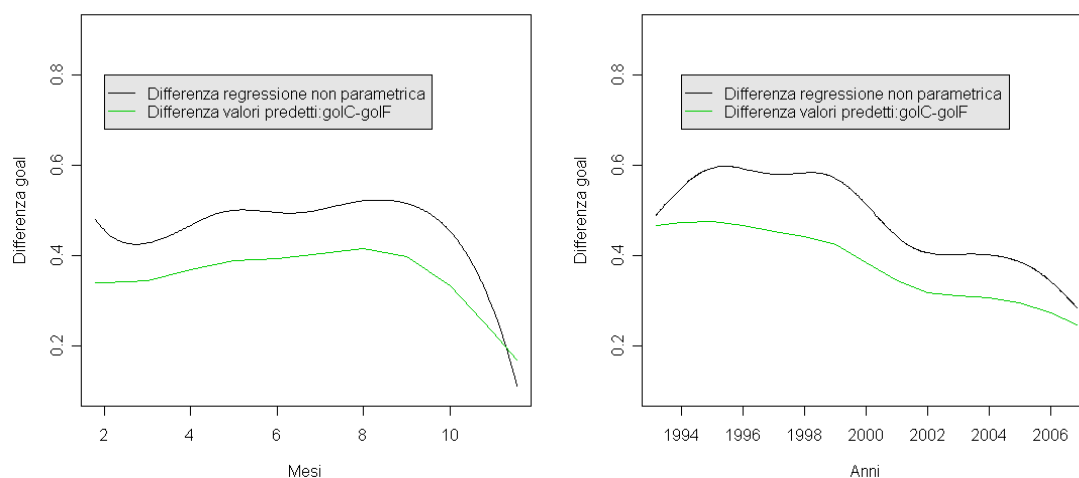
2.1.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati

→ Dopo aver adattato un modello per i gol in casa ci risulta, alla fine, che la variabile Anni è significativa. Ciò conferma la struttura vista nell'analisi esplorativa. Anche in questo caso dopo aver fatto un test di bontà del modello è risultato che il modello con la sola variabile Anni richiederebbe parametri in più; in altre parole, il modello non va tanto bene, cioè non spiega bene la variabilità dei dati.

→ Per quanto riguarda i gol delle squadre fuori casa, dopo aver adattato un modello, contrariamente al caso precedente, risulta che solo il mese è significativo. Però facendo un test basato sulla devianza residua, ci risulta che il modello non va tanto bene e di conseguenza non spiega bene la variabilità dei dati.

→ Considerando un modello per il totale dei gol per partita risulta che le due variabili Anni e Mesi sono significativi al 5%. Inoltre il test di bontà del modello ci ha suggerito che il nostro modello ha bisogno di ulteriori variabili, cioè che il modello stimato non spiega correttamente la variabilità dei dati e, quindi, non è un modello adatto per fare previsioni.

→ Per quanto riguarda il modello per la differenza ecco i grafici ottenuti con la procedura descritta precedentemente.



La prima cosa che evidenziamo in questi grafici è che, facendo la differenza dei valori predetti, la struttura sembra diversa da quella ottenuta con la

regressione non parametrica. Comunque sembra che, se avessimo adattato un modello per la differenza, ci saremmo aspettati che solo la variabile mese fosse significativa.

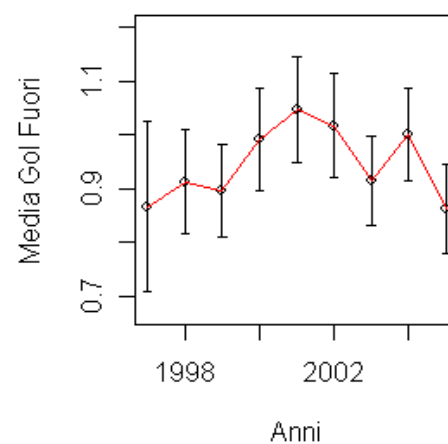
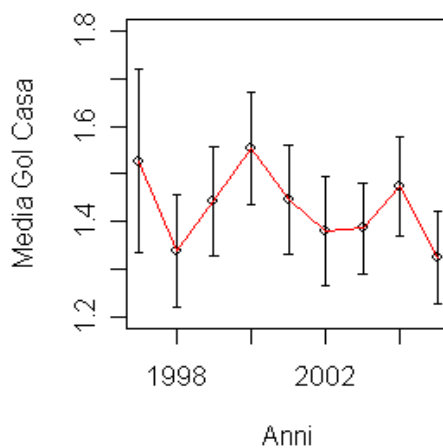
Osservazione: Anche in questo campionato, in tutti i modelli stimati per i dati della Serie A, succede che i modelli stimati non sono adeguati per la spiegazione dei dati. Il problema, potrebbe essere dovuto al fatto che i gol segnati non hanno proprio una distribuzione di poisson e in parte c'è una extra-variabilità di conseguenza della variabilità delle squadre.

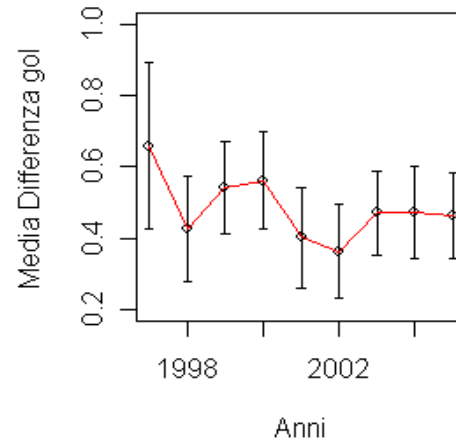
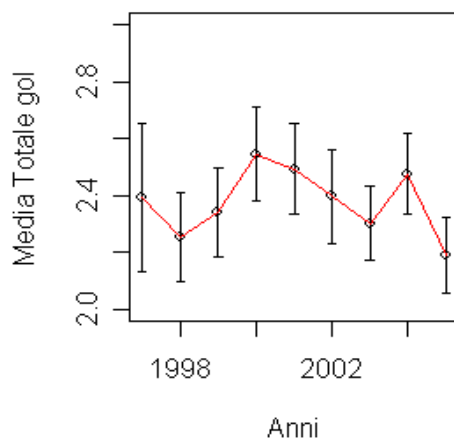
2.2 CAMPIONATO ITALIANO SERIE B

2.2.1 Analisi esplorativa

2.2.1.1 Primo approccio dell'analisi esplorativa

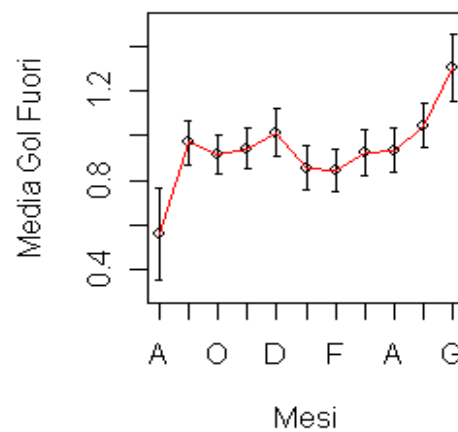
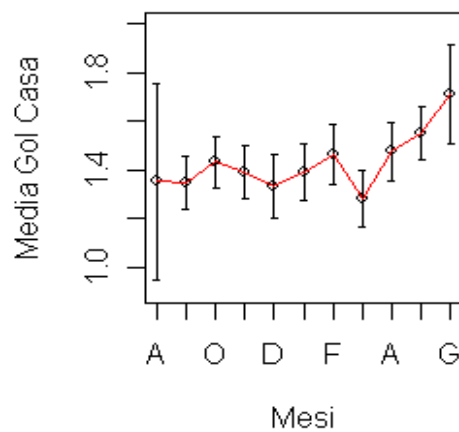
→ **Struttura della media dei gol al variare degli anni.**

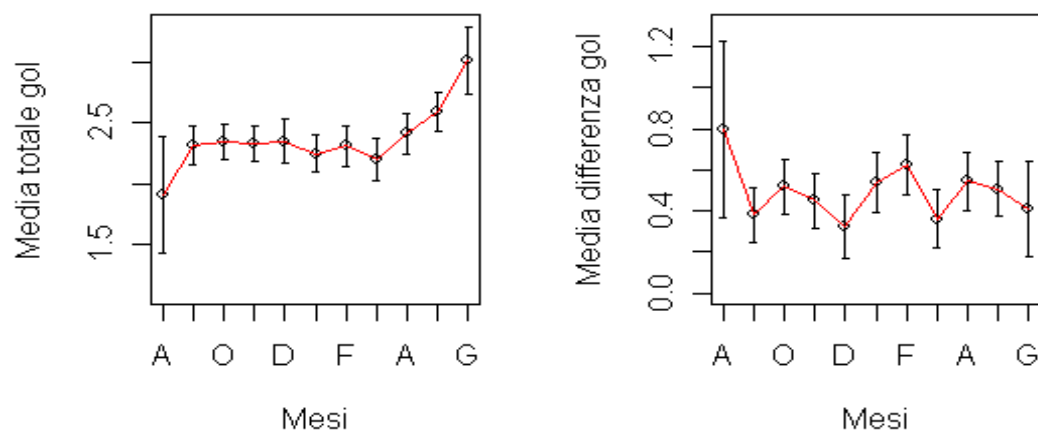




Considerando la media dei gol per le squadre in casa, fuori casa, per il totale dei gol e per la differenza, sembra difficile cogliere una struttura particolare al crescere degli anni.

→ **Struttura della media dei gol al variare dei mesi.**



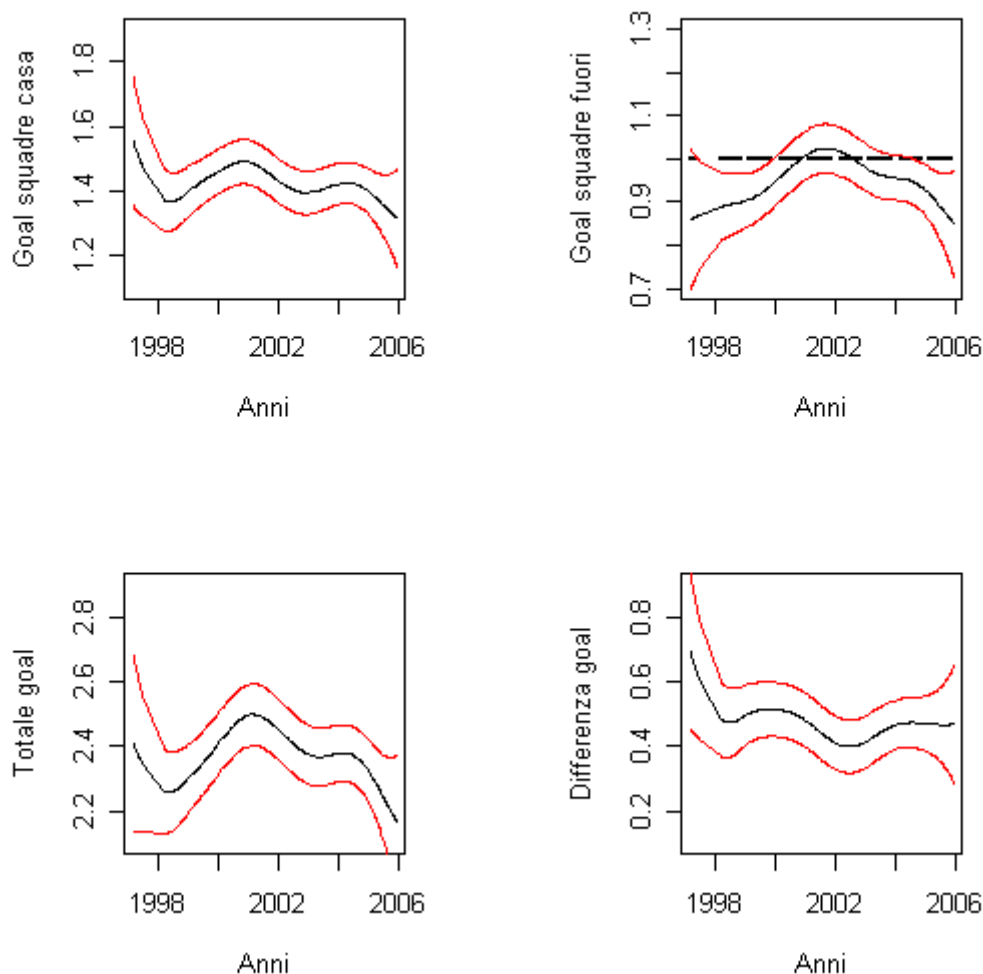


Considerando la media dei gol per le squadre in casa al variare dei mesi, sembra che da agosto fino a marzo l'andamento è più o meno costante; ciò vuol dire che in media si segna lo stesso numero di gol per partita, poi comincia a crescere fino alla fine della stagione(giugno).

Per quanto riguarda i gol per le squadre fuori casa e il totale dei gol per partita il grafico ci fa vedere una struttura molto interessante della media dei gol: si vede che all'inizio si segna poco cioè meno di un gol in media, poi aumenta leggermente nel successivo mese e segue un andamento costante fino marzo e infine comincia a crescere per gli ultimi mesi. Nel caso della media delle differenza dei gol per partita non sembra esserci una struttura particolare nei dati.

2.2.1.1 Secondo approccio dell'analisi esplorativa

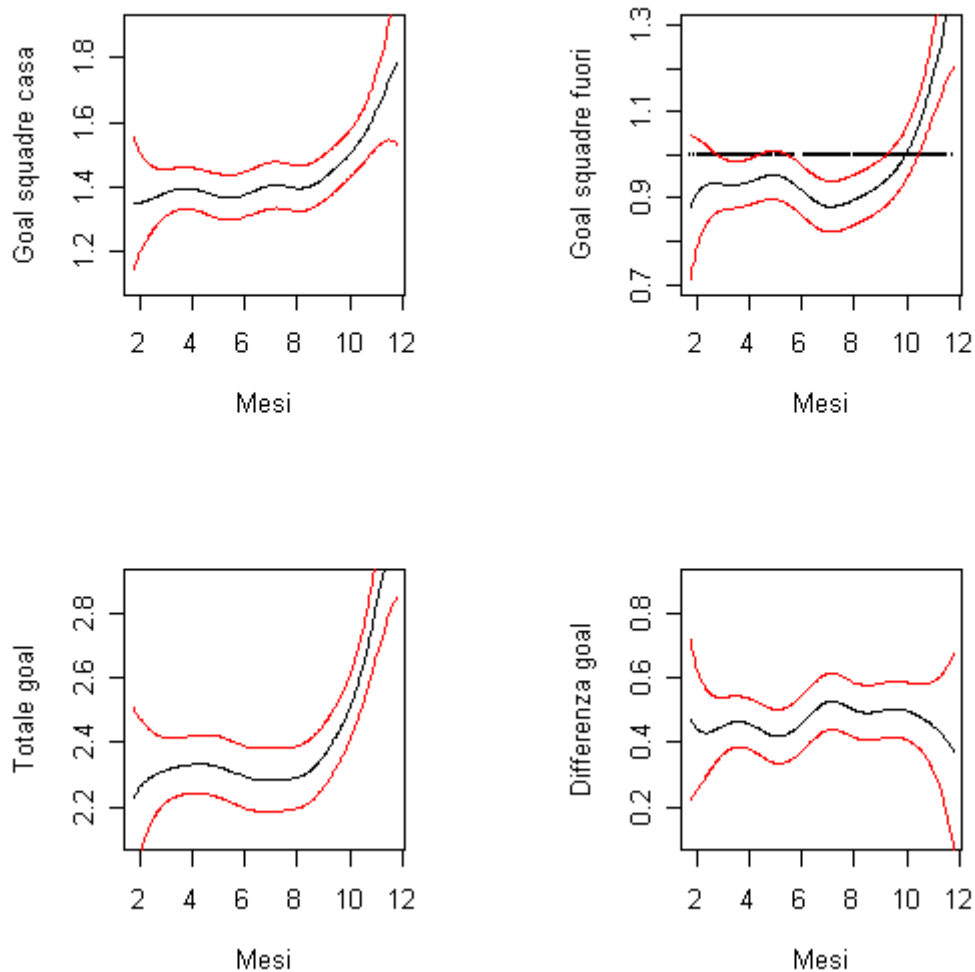
→ **Struttura dei gol al variare degli anni.**



In questi quattro grafici evidenziamo facilmente la struttura dei gol rispetto al grafico delle medie. Anche in questi quattro grafici vediamo delle strutture interessanti. In particolare per il caso dei gol per le squadre fuori casa e per il totale dei gol per partita, si vede la curva che cresce da 1998 fino al 2001 e poi comincia a decrescere fino al 2006. Invece per il caso dei gol per le squadre in casa e della differenza dei gol per partita, non sembra esserci un andamento particolare ma al variare degli anni ci sono delle oscillazioni.

→ **Struttura dei gol al variare dei Mesi.**

Legenda: 1='Agosto', 2='Settembre', ..., 11='giugno'



Tranne il caso della differenza dei gol per partita che assumere un andamento costante fino alla fine della stagione, negli altri tre grafici si vede che tendenzialmente si ha la stessa struttura che sembra essere più o meno costante dall'inizio della stagione fino a marzo e poi comincia a crescere fino alla fine della stagione (giugno). Anche nella Serie B italiana notiamo che si segna di più negli ultimi tre mesi.

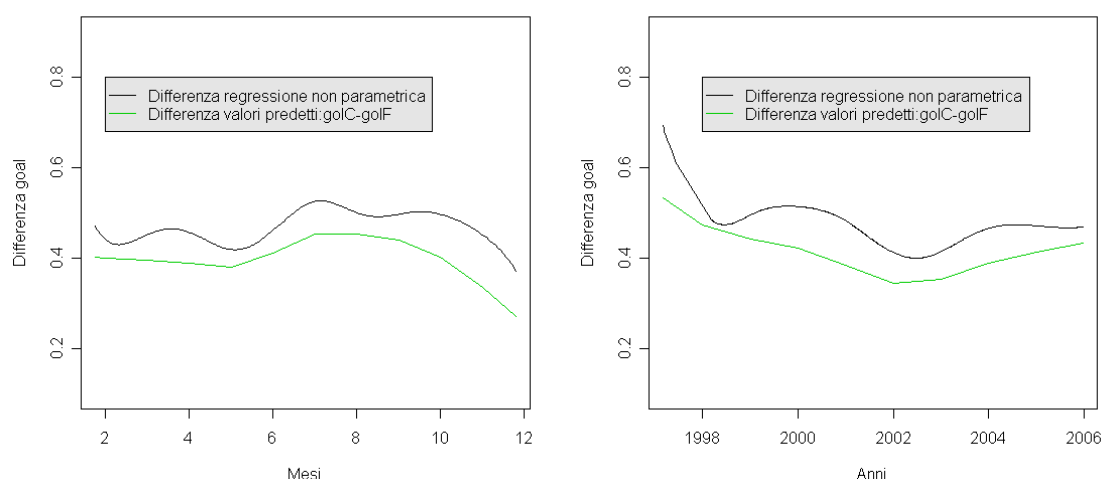
2.2.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.

→ Per quanto riguarda i gol per le squadre in casa della Serie B Italiana, dopo aver adattato un modello statistico per spiegare i dati, ci risulta che al livello 5% le due variabili Anni e Mesi non sono significative. Quindi, i gol per le squadre in casa non sono dipendenti da nessuna delle due variabili; togliendo, però, le due variabili, il modello con la sola intercetta non spiega bene i dati e richiede variabili in più.

→ Per il caso dei gol per le squadre fuori casa, dopo aver adattato un modello statistico per la spiegazione dei dati, ci risulta che le due variabili Anni e Mesi sono significative. Però, nel passo successivo come negli altri casi visti, il test per la bontà ci suggerisce che il modello stimato non è in grado di spiegare bene la variabilità dei dati, cioè ci servono parametri in più.

→ Per quanto riguarda il totale dei gol per partita, dopo l'adattamento del modello ci risulta anche qui che le due variabili Anni e Mesi sono significative, però, anche in questo caso, il test per la bontà del modello ci suggerisce che le nostre due variabili non bastano per analizzare la variabilità dei dati, cioè ci vorrebbero ulteriori variabili per migliorare la capacità predittiva del nostro modello.

→ Per il caso della differenza dei gol per partita, quelli che seguono sono i grafici ottenuti con l'approccio spiegato nei punti precedenti.



Evidenziamo subito in questi grafici che, facendo la differenza dei valori predetti dei gol delle squadre in casa con i gol delle squadre fuori casa, la struttura sembra diversa da quella ottenuta con la regressione non parametrica sia

al variare dei mesi che degli anni. Sembrerebbe che, se avessimo adattato un modello per la differenza dei gol, ci saremmo aspettati che le due variabili non fossero significative.

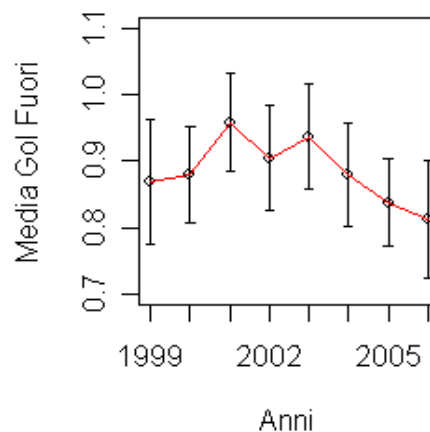
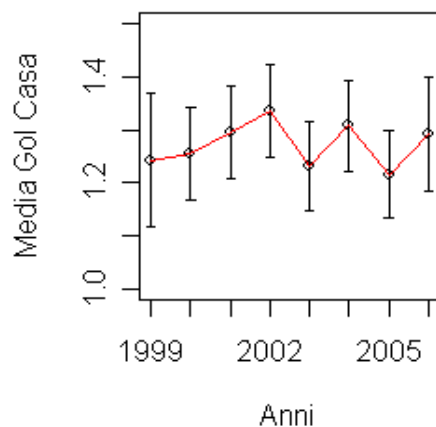
Osservazione: Anche in qui, in tutti i modelli stimati per i dati della Serie B, succede che i modelli stimati non sono adeguati per la spiegazione dei dati. Il problema, potrebbe essere dovuto al fatto che i gol segnati non hanno proprio una distribuzione di poisson e in parte c'è una extra-variabilità di conseguenza della variabilità delle squadre.

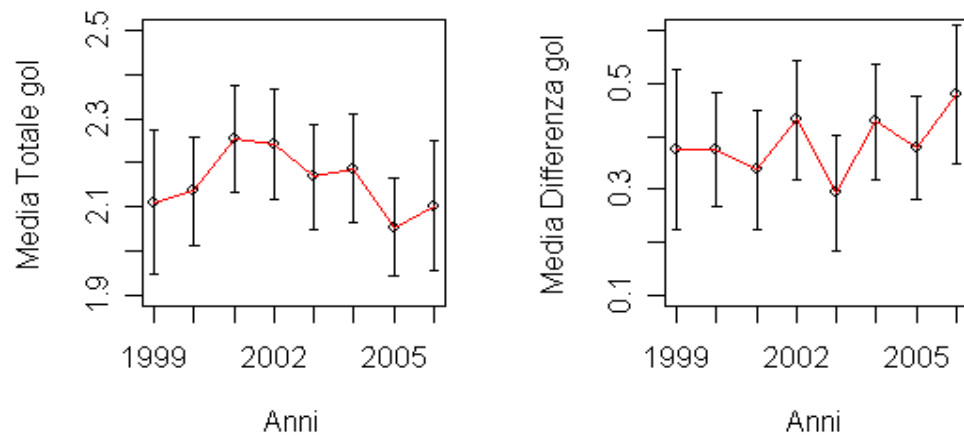
2.3 CAMPIONATO ITALIANO SERIE C

2.3.1 Analisi esplorativa

2.3.1.1 Primo approccio dell'analisi esplorativa

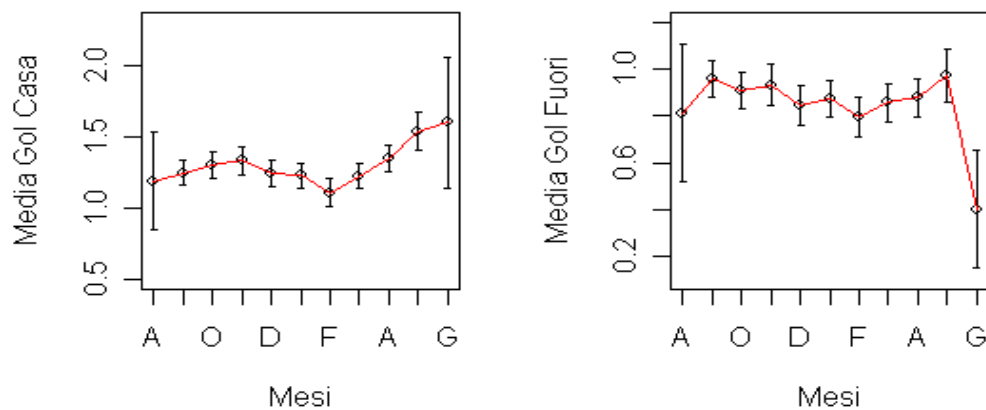
→ struttura della media dei gol al variare dell'anno.

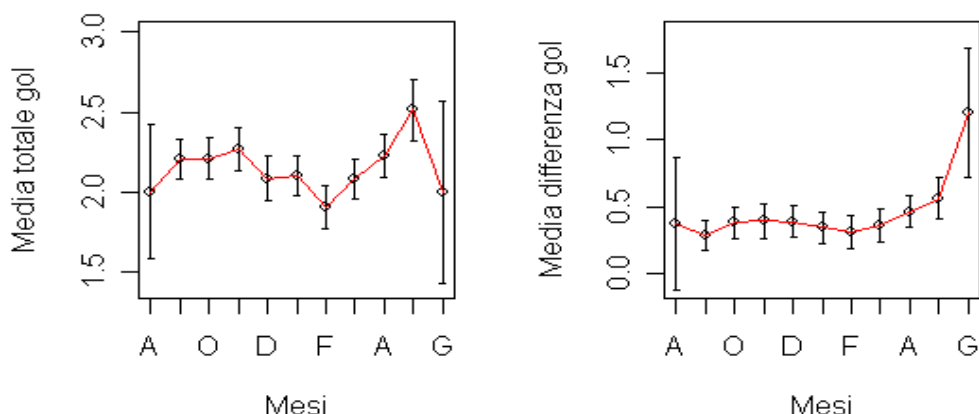




In questi quattro grafici al variare degli anni, non riusciamo ad evidenziare una struttura particolare della media dei gol in casa, fuori casa, totale e differenza. Quindi ci aspettiamo, dopo aver adattato un modello per i dati in tutti i quattro casi, che la variabile Anni non sia significativa.

➔ **struttura della media dei gol al variare del mese.**

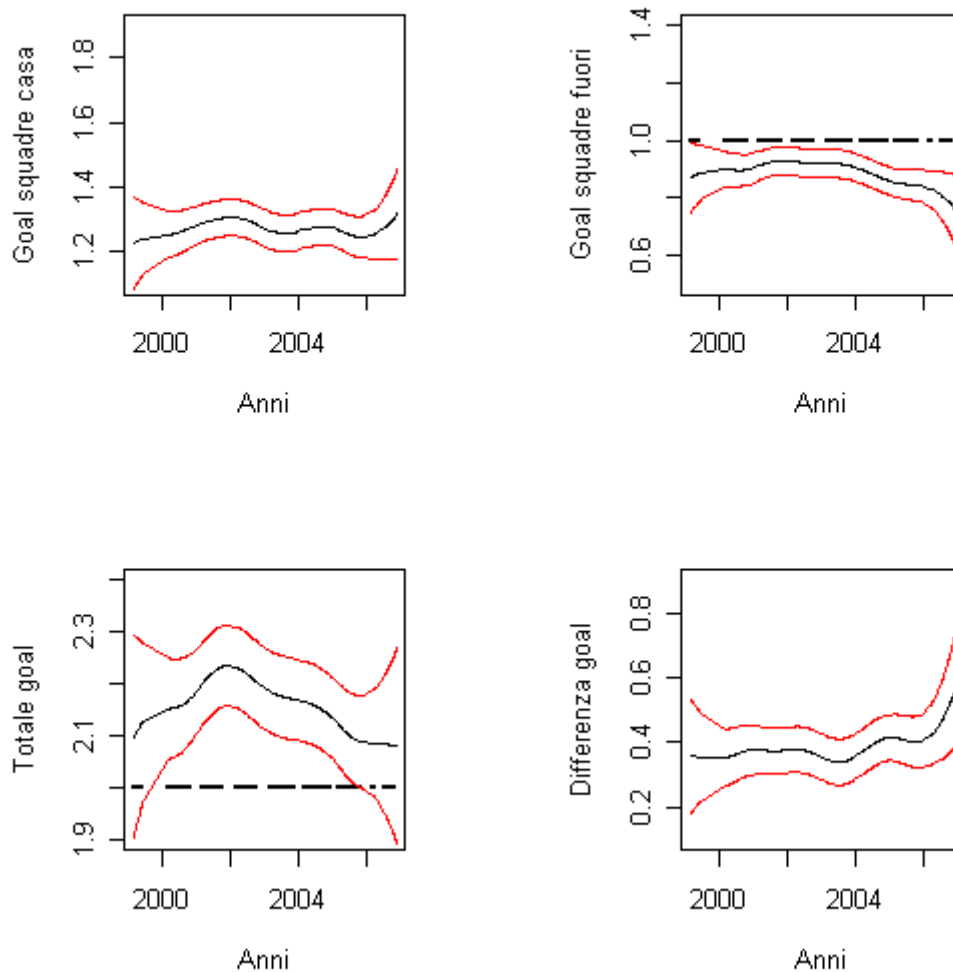




Guardando questi quattro grafici vediamo subito delle strutture particolarmente interessanti della media dei gol al variare dei mesi. Se consideriamo il caso della media dei gol delle squadre in casa, sembra crescere da agosto a novembre, poi decresce fino a febbraio e poi ricomincia a crescere fino alla fine della stagione. Per quanto riguarda la media dei gol delle squadre fuori casa, al variare dei mesi sembra essere costante fino a maggio seguito da un calo verso giugno. Per la media del totale, al variare dei mesi abbiamo quasi lo stesso andamento della media dei gol delle squadre in casa. La differenza tra i due casi è che decresce da maggio fino a giugno. Infine, al variare dei mesi, la media della differenza assume un andamento costante dall'inizio della stagione fino a febbraio e poi comincia a crescere fino alla fine della stagione. Quindi ci aspettiamo, dopo l'adattamento di un modello ai nostri dati, che la variabile mese sia significativa.

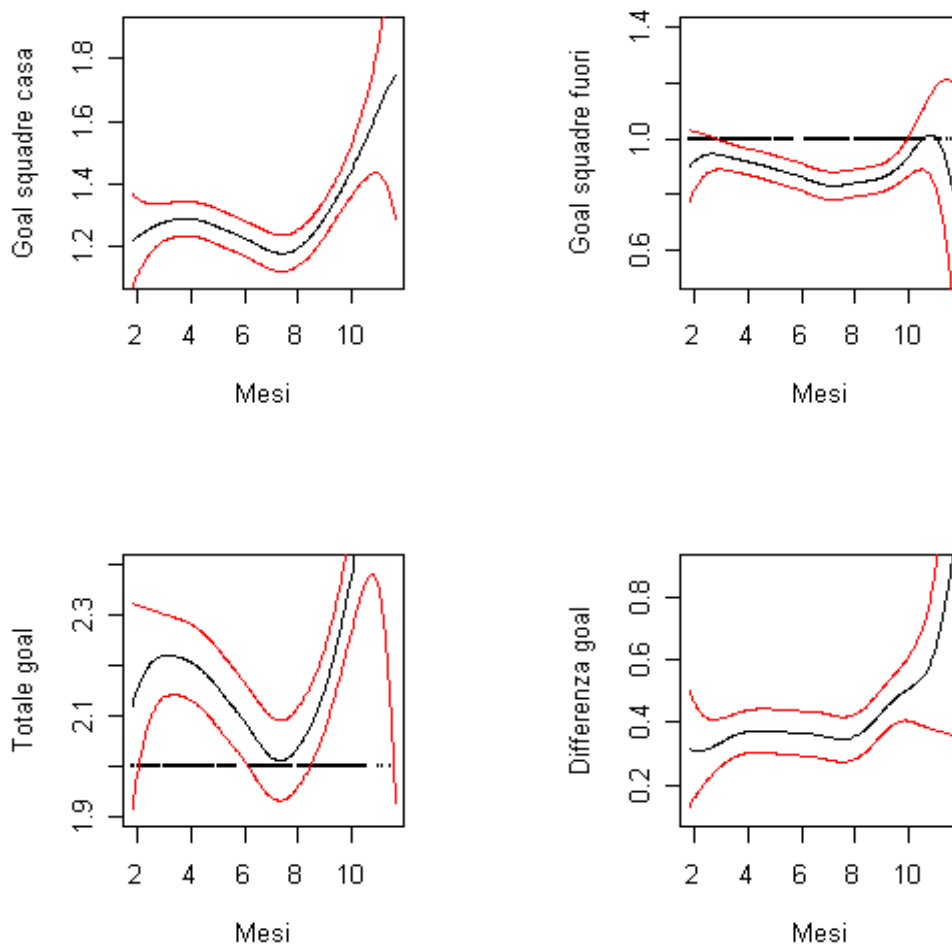
2.3.1.1 Secondo approccio dell'analisi esplorativa

→ struttura dei gol al variare dell'anno.



Guardando questi quattro grafici, risulta che l'andamento del totale dei gol al crescere degli anni sembra essere l'unico caso in cui la struttura è particolare, in quanto prima cresce e poi decresce. Negli altri tre casi non si riesce a evidenziare un andamento interessante.

→ **Struttura dei gol al variare del mese.**



Notiamo in questi quattro grafici che l'andamento è simile a quello visto considerando la media dei gol in funzione dei mesi. Comunque, si vede che in tutti i casi c'è una struttura. Quindi ci aspettiamo, dopo aver adattato un modello per i nostri dati, che la variabile Mesi abbia una relazione con i gol delle squadre in casa, fuori casa, il totale e la differenza dei gol.

2.3.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.

→ Per quanto riguarda i gol delle squadre in casa, dopo l'adattamento del modello, ci risulta che la variabile mese è fortemente significativa al livello 5% e la variabile Anni non lo è, cioè il risultato che ci aspettavamo dall'analisi esplorativa. Tuttavia, togliendo la variabile non significativa Anni, il mese ci risulta lo stesso fortemente significativo. Anche in questo caso succede che il test

di bontà del modello ci suggerisce che il modello finale a cui siamo arrivati, cioè il modello con sola la variabile mese, non spiega bene la variabilità dei dati quanto il modello saturo (modello che associa un parametro ad ogni osservazione). In altre parole il nostro modello richiederebbe ulteriori variabili per poter migliorare la capacità predittiva.

→ Considerando il caso dei gol delle squadre fuori casa, dopo l'analisi avvenuta, dall'adattamento del modello risulta che al livello approssimato 5%, la variabile Anni non è significativa, ma è significativa solamente la variabile mese. Anche questo è un risultato che ci aspettavamo, ma togliendo la variabile Anni, il mese rimane sempre significativo al livello 5%, però il test di bontà del modello ci informa che il nostro modello, che contiene sola la variabile mese, non spiega bene la variabilità dei dati. Quindi la capacità predittiva non è affidabile.

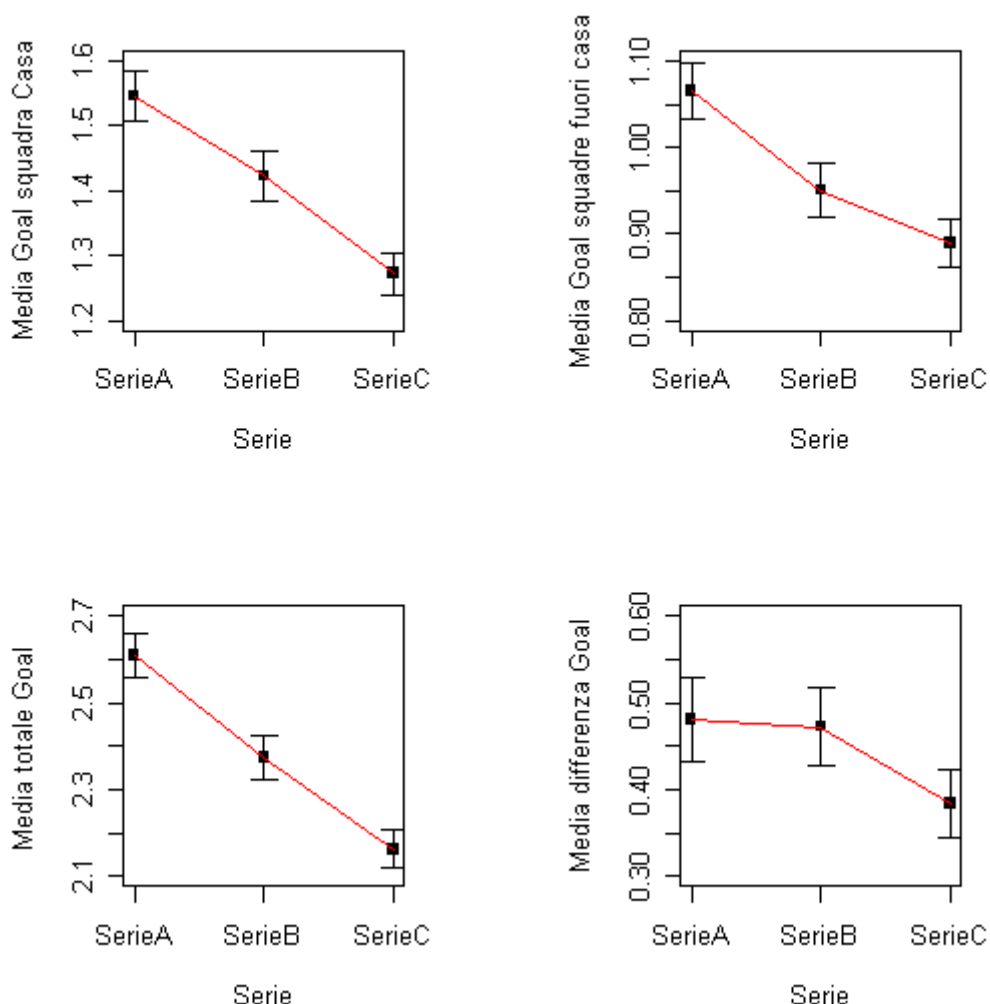
→ Per il caso del totale dei gol, dopo l'adattamento del modello, al livello 5% risulta significativa solo la variabile mese, la variabile Anni, invece, non è significativa. Anche togliendo la variabile Anni, la variabile Mesi rimane sempre significativa. Il test di bontà del modello che contiene sola la variabile Mesi, ci suggerisce che non è un modello che spiega bene la variabilità dei dati quanto il modello saturo, cioè il modello richiederebbe l'aggiunta di ulteriori variabili; con le variabili a disposizione infatti, non ce la facciamo a risolvere il problema e di conseguenza avremo delle previsioni dubbiose.

→ Per quanto riguarda la differenza dei gol per partita, applicando le procedure descritte nei punti sopra, succede che, se dovessimo adattare un modello specifico, ci aspetteremmo che sia significativa solamente la variabile Mesi.

Osservazione: Anche qui, in tutti i modelli stimati per i dati della Serie C, succede che i modelli stimati non sono adeguati per la spiegazione dei dati. Il problema, potrebbe essere dovuto al fatto che i gol segnati non hanno proprio una distribuzione di poisson e in parte c'è una extra-variabilità di conseguenza della variabilità delle squadre.

2.3.3 Differenza tra le varie Serie indipendentemente dal mese e dell'anno

Analisi esplorativa



Questi quattro grafici ci permettono di confrontare le varie serie, in particolare se consideriamo il grafico in alto a sinistra, cioè il grafico della media dei gol delle squadre in casa al variare della categoria, risulta subito che in Serie A in media, si segna di più, segue la Serie B e infine la Serie C in cui si segna in media di meno rispetto alle altre due Serie. Di conseguenza diciamo che il numero di gol segnato dalle squadre in casa è influenzato dalla Serie.

Il secondo grafico in alto a destra ci suggerisce che in media le squadre fuori casa segnano di più in Serie A, poi in Serie B e infine nella Serie C. Si vede anche in

questo caso che gli intervalli non si sovrappongono, quindi ci aspettiamo che i gol segnati dalle squadre fuori casa siano influenzati dalla Serie.

Il terzo grafico in basso a sinistra si suggerisce che in media complessivamente si segnerebbe di più in Serie A, poi segue nella serie B infine nella Serie C. Notiamo anche qui che gli intervalli non si sovrappongono e di conseguenza il totale dei gol per partita è molto influenzato dal tipo di Serie.

Il quarto grafico in basso a destra ci suggerisce che in media la differenza dei gol per partita è simile sia in Serie A che in Serie B ma è maggiore rispetto alla Serie C. Notiamo che due intervalli di confidenza si sovrappongono, quindi ci aspettiamo che il tipo di Serie non abbia effetto sulla differenza dei gol per partita. Effettuiamo un test per vedere se ci sono differenze tra le varie Serie indipendentemente del mese e dell'anno.

- **Caso in cui la variabile di interesse sono i gol delle squadre in casa.**

Analysis of Deviance Table

Model 1: golC ~ 1

Model 2: golC ~ categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	12172	13711			
2	12170	13596	2	115	1.061e-25

Il valore livello di significatività ci suggerisce, come ci aspettavamo, che la variabile categoria è fortemente significativa, cioè che i gol segnati dalle squadre in casa sono fortemente influenzati dal tipo di Serie.

- **Caso in cui la variabile di interesse sono i gol delle squadre fuori casa.**

Analysis of Deviance Table

Model 1: golF ~ 1

Model 2: golF ~ categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	12172	13959.8			
2	12170	13889.5	2	70.3	5.423e-16

Il valore del livello di significatività ci suggerisce che la variabile categoria risulta fortemente significativa al livello 5% e di conseguenza i gol segnati dalle squadre fuori casa sono fortemente influenzati dalla categoria.

- **Caso in cui la variabile di interesse è il totale dei gol per partita.**

Analysis of Deviance Table

Model 1: (golC + golF) ~ 1

Model 2: (golC + golF) ~ categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	12172	14939			
2	12170	14756	2	183	1.786e-40

Come negli altri due casi la variabile categoria risulta fortemente significativa al livello 5%, quindi al termine di una partita il totale dei gol risulta fortemente influenzato dalla categoria.

- **Caso in cui la variabile di interesse é differenza dei gol per partita.**

Per il caso della differenza, risulta difficile applicare un modello gam ai dati, perché avremmo una differenza negativa. Se ci fidiamo alle analisi esplorative, risulta che la differenza dei gol per partita non è influenzata dalla Serie.

2.3.4 Modelli di analisi delle differenze tra le varie Serie dato il mese e l'anno.

- **Caso in cui la variabile di interesse sono i gol delle squadre in casa.**

Analysis of Deviance Table

Model 1: golC ~ s(Mesi)

Model 2: golC ~ s(Mesi) + categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	12168	13671.5			
2	12166	13562.1	2	109.3	1.808e-24

Su questo output si vede che non c'è la variabile Anni: l'abbiamo tolta perché non era significativa. Il valore del livello di significatività ci suggerisce che il model 2 risulta preferibile al model 1 e di conseguenza la variabile categoria risulta fortemente significativa, quindi, dato il mese, i gol segnati dalle squadre in casa sono fortemente influenzati dalla categoria. Il test di bontà del modello, però, ci ha suggerito che il modello stimato non spiega bene la variabilità dei dati, quindi avremo previsioni dubbiose.

- **Caso in cui la variabile di interesse sono i gol delle squadre fuori casa.**

Analysis of Deviance Table

Model 1: golF ~ s(Mesi) + s(Anni)

Model 2: golF ~ s(Mesi) + s(Anni) + categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	12164	13905.5			
2	12162	13816.8	2	88.7	5.43e-20

Il valore del livello significatività ci suggerisce che è preferibile il Model 2; di conseguenza la variabile categoria è fortemente significativa, quindi, dato il mese o l'anno, i gol segnati dalle squadre fuori casa sono influenzati dalla categoria; sorge, tuttavia, il problema riscontrato in tutti gli altri casi che riguarda la bontà del modello: il test ci ha suggerito che il modello non spiega bene la variabilità dei dati e quindi la capacità predittiva è dubbiosa.

- **Caso in cui la variabile di interesse è il totale di gol per partita.**

Analysis of Deviance Table

Model 1: (golC + golF) ~ s(Mesi) + s(Anni)

Model 2: (golC + golF) ~ s(Mesi) + s(Anni) + categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	12164	14824.4			
2	12162	14664.5	2	159.9	1.928e-35

Il valore del test ci suggerisce che è preferibile il model 2, cioè la variabile categoria risulta significativa e di conseguenza, dato il mese o l'anno, il totale dei gol per partita è influenzato dalla categoria. Essendo rimasti con un modello in cui tutte le variabili sono significative, il test di bontà di modello ci ha suggerito comunque che il modello stimato non spiega bene la variabilità dei dati quanto il modello saturo, cioè che il modello richiederebbe l'aggiunta di ulteriori variabili e quindi la capacità predittiva è dubbiosa.

Osservazione: Per tutti i modelli stimati al punto 2.3.3 e 2.3.4, constatiamo che non sono ben adatti per spiegare i dati; quindi il problema può essere dovuto ai motivi specificati negli altri casi analizzati precedentemente.

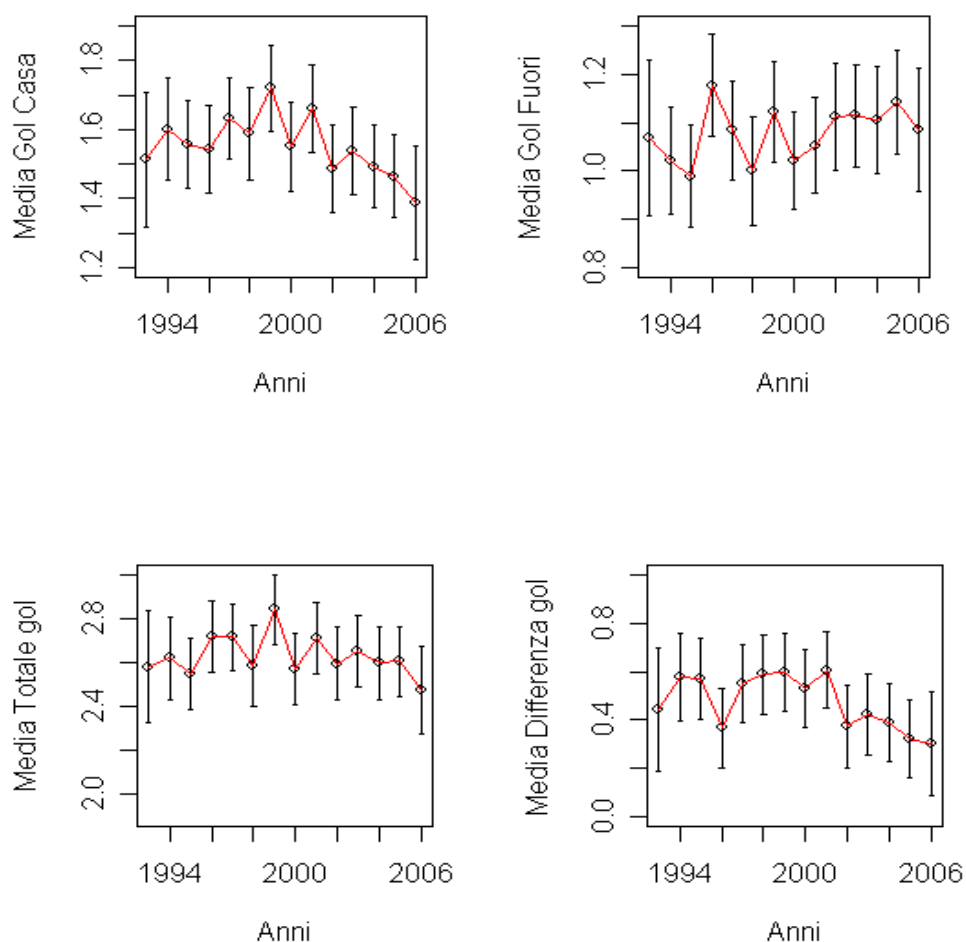
CAPITOLO 3: CAMPIONATO SPAGNOLO

3.1 CAMPIONATO SPAGNOLO SERIE A

3.1.1 Analisi esplorativa

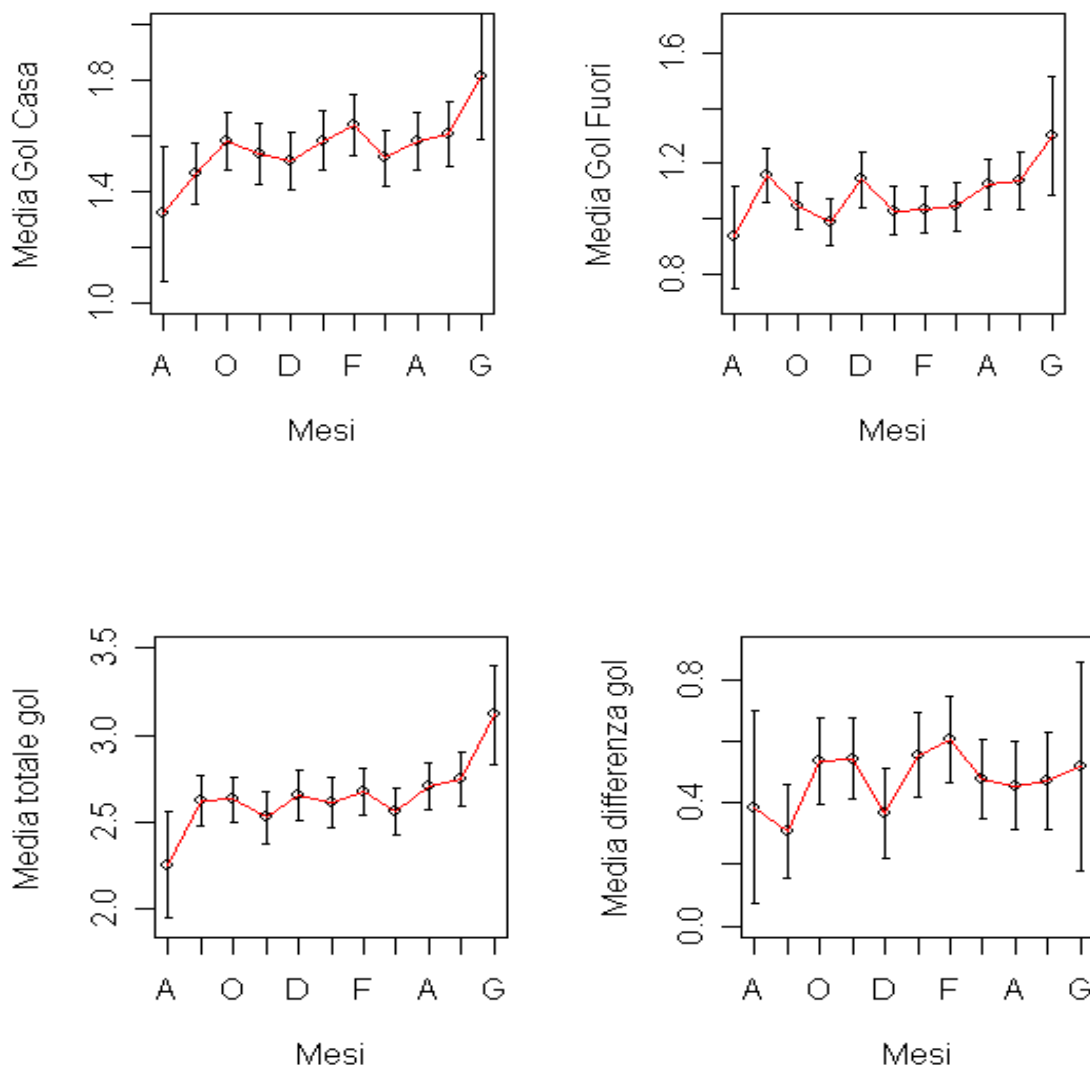
3.1.1.1 Primo approccio dell'analisi esplorativa

→ **struttura della media dei gol al variare degli anni.**



Considerando il grafico della media dei gol delle squadre in casa, fuori casa, del totale di gol e della differenza, al variare degli anni si coglie un andamento particolare nei dati. L'andamento sembra essere stazionario sia in media che in varianza.

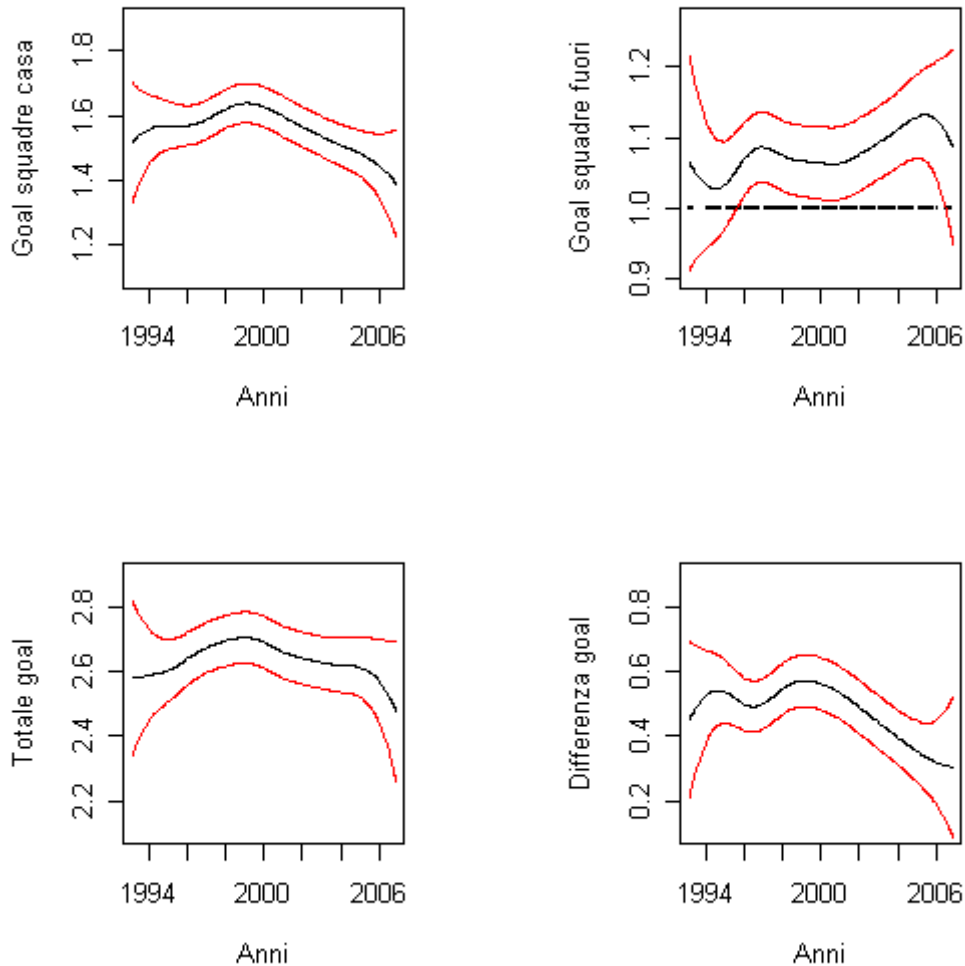
→ **struttura della media dei gol al variare del mese.**



Tranne il grafico in basso a destra che rappresenta la media della differenza dei gol al variare dei mesi, nel quale non c'è un andamento che ci dà una grande informazione, gli altri tre grafici ci fanno vedere un andamento abbastanza interessante e quasi simile; cioè all'inizio della stagione (ad agosto) in media si segna di meno rispetto agli altri mesi, poi il numero di gol cresce fino a ottobre e sembra essere costante fino a marzo e ricomincia a crescere fino alla fine della stagione (a giugno). Anche in questo caso sembra che si segni in media di più negli ultimi due o tre mesi.

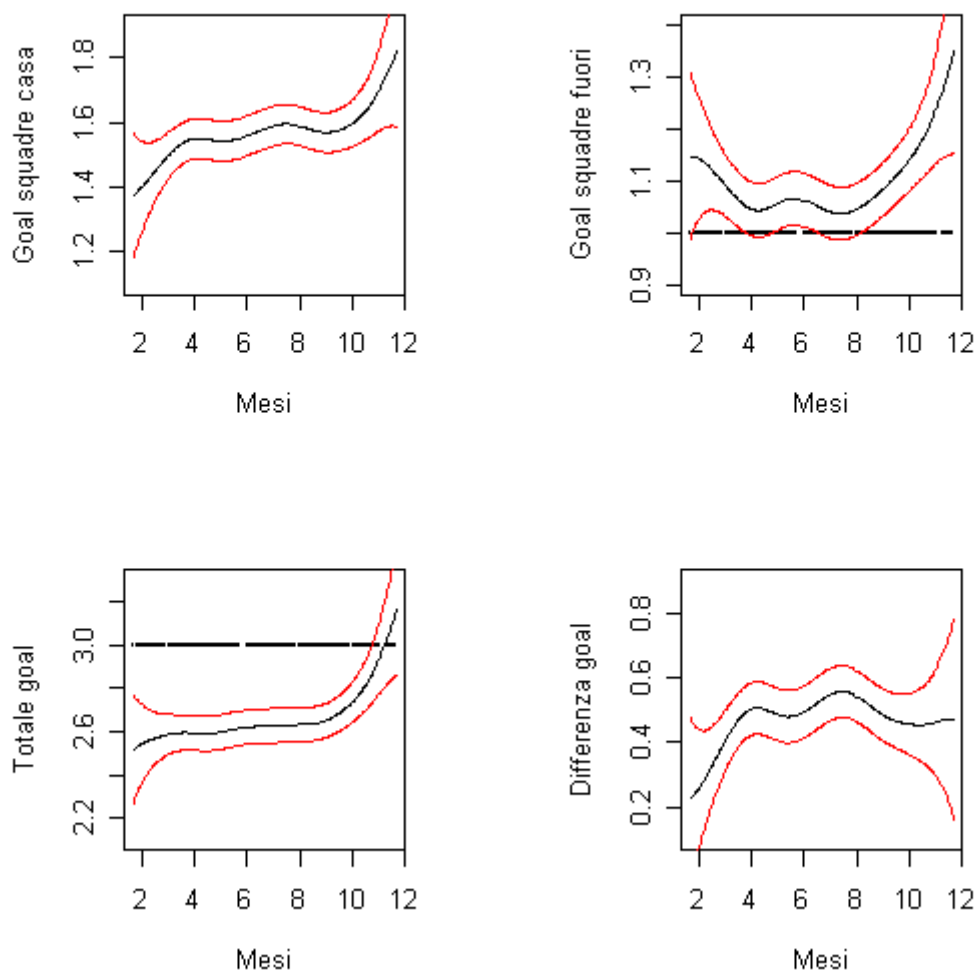
3.1.1.1 Secondo approccio dell'analisi esplorativa

→ **Struttura dei gol al variare degli anni.**



Con questo approccio si riesce a vedere l'andamento dei gol nei vari casi. Se consideriamo il caso dei gol delle squadre in casa vediamo un andamento interessante che cresce e poi decresce al crescere degli anni. Tale non è il caso dei gol delle squadre fuori casa: sembra non esserci, infatti, una struttura; per il totale dei gol per partita più o meno l'andamento sembra essere costante; per il caso della differenza l'andamento sembra essere costante dal 1994 fino al 2000 e poi inizia a decrescere.

→ **Struttura dei gol al variare dei mesi.**



Il grafico in alto a sinistra ci fa vedere l'andamento dei gol segnati dalle squadre in casa in funzione dei mesi. Notiamo che la curva cresce dall'inizio della stagione (ad agosto) fino a ottobre, poi sembra essere costante fino ad aprile e poi ricomincia a crescere fino alla fine della stagione, cioè a giugno.

Il grafico in alto a destra ci rappresenta i gol segnati dalle squadre fuori casa in funzione dei mesi: si vede che l'andamento decresce dall'inizio della stagione fino a novembre, poi sembra diventare costante fino ad aprile e ricomincia a crescere fino alla fine della stagione. Per quanto riguarda il grafico in basso a sinistra si vede che l'andamento del totale dei gol al crescere del mese assume un andamento costante dall'inizio della stagione fino al mese di maggio e poi comincia a crescere fino alla fine della stagione.

L'ultimo grafico ci fa vedere l'andamento della differenza al variare dei mesi: all'inizio della stagione cresce leggermente fino a novembre, poi sembra

essere costante fino a marzo e infine comincia a decrescere fino al termine della stagione.

3.1.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati

→ Caso in cui la variabile di interesse sono i gol segnati dalle squadre in casa

Dopo aver adattato un modello per i gol delle squadre in casa, abbiamo visto che la variabile Anni era significativa mentre la variabile Mesi non lo era. Pur togliendo la variabile non significativa Mesi, l'anno rimane significativo, però il test di bontà del modello ci suggerisce che il nostro modello finale, cioè modello con la sola variabile Anni non spiega bene la variabilità dei dati quanto il modello saturo(un modello che associa un parametro ad ogni osservazione); in altre parole il modello richiederebbe l'aggiunta di ulteriori variabili. Il fatto che il modello finale non sia adatto ai dati può essere ai motivi specificati nei casi simili visti precedentemente.

→ Caso in cui la variabile di interesse sono i gol segnati dalle squadre fuori casa

Dopo tutte le analisi per l'adattamento del modello per i gol delle squadre fuori casa e dopo l'eliminazione delle variabili non significative, siamo rimasti con un modello in cui la sola variabile Mesi risulta significativa, anche se il test di bontà del modello basato sulla devianza residua ci suggerisce che il nostro modello finale non spiega tanto bene la variabilità dei dati quanto il modello saturo; in altre parole, come detto sopra, richiederebbe l'aggiunta di ulteriori variabili e, di conseguenza, la capacità predittiva è dubbia. Anche qui il problema della non adeguatezza può essere dovuto ai motivi specificati nei casi simili visti ai punti precedenti.

→ Caso in cui la variabile di interesse è il totale dei gol segnati al termine di una partita.

Dopo tutte le analisi per l'adattamento del modello per il totale dei gol per partita e poi successivamente eliminando le variabili non significative, arriviamo

ad un modello con la sola intercetta, il test di bontà del modello ci suggerisce che il modello stimato non è adatto per spiegare la variabilità dei dati quanto il modello saturo. Di conseguenza avremo previsioni dubbiose.

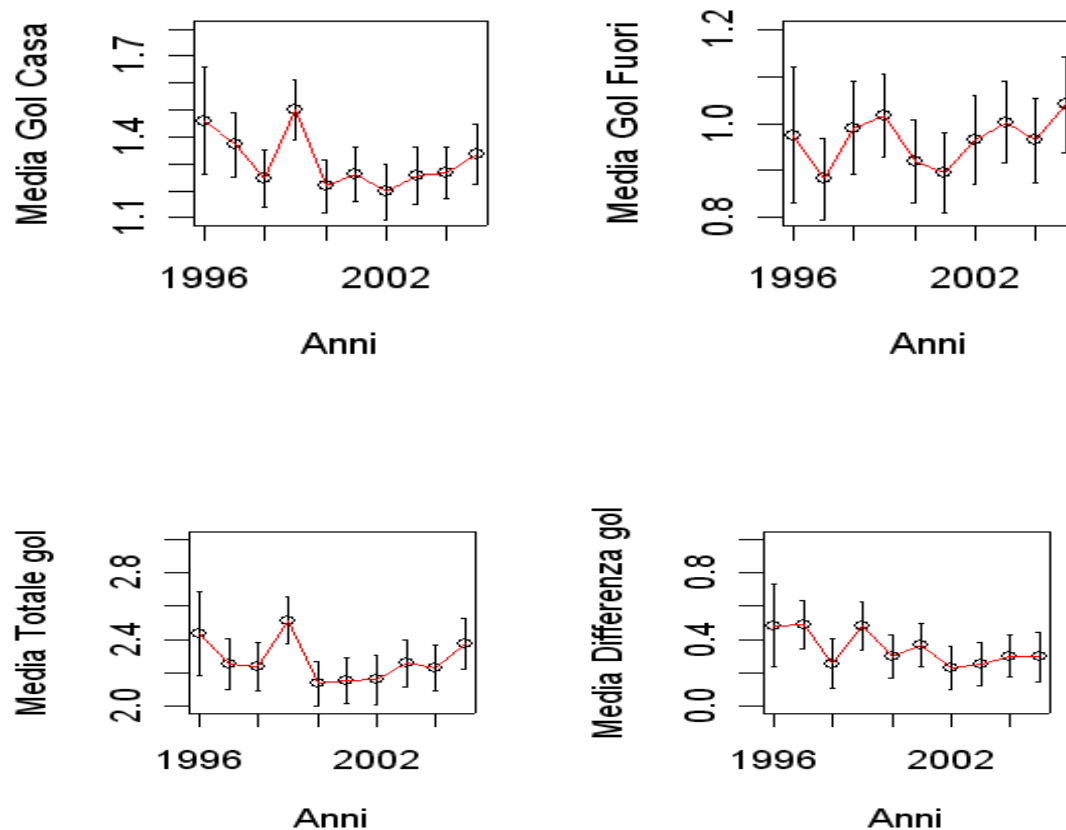
→ **Caso in cui la variabile di interesse è la differenza dei gol segnati al termine di una partita.**

Per quanto riguarda la differenza dei gol per partita, applicando le procedure descritte nei capitoli precedenti, succede che, se dovessimo adattare un modello, ci aspetteremmo che solo la variabile Mesi sia significativa.

3.2 CAMPIONATO SPAGNOLO SERIE B

3.2.1 Analisi esplorativa

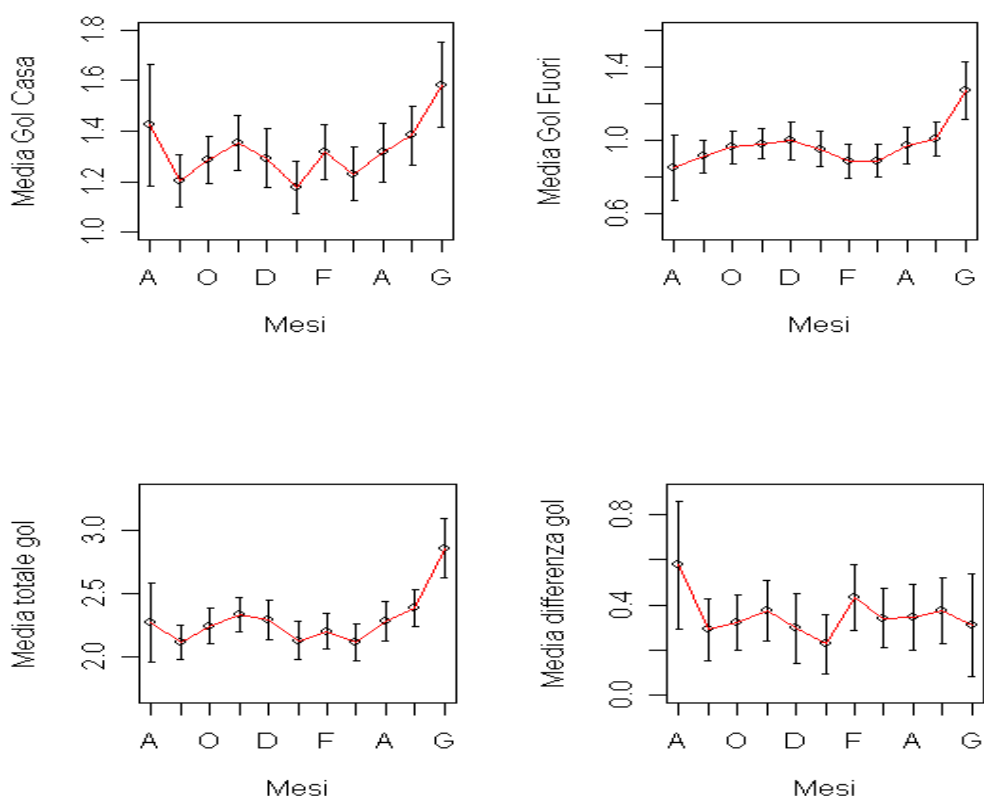
→ **Struttura della media dei gol al variare degli anni.**



Il grafico in alto a sinistra rappresenta la media dei gol segnati in casa in funzione degli anni. Vediamo un andamento abbastanza interessante che decresce da 1996 al 1998, poi fa un grande salto nel 1999 e infine sembra essere costante fino al 2005.

Il grafico in alto a destra ci fa vedere la media dei gol segnati dalle squadre fuori casa in funzione degli anni. Essa assume una struttura poco interessante, cioè si ha un andamento che cresce e decresce al crescere degli anni. Il grafico in basso a sinistra rappresenta la media del totale dei gol in funzione degli anni. Essa assume la stessa struttura dei gol delle squadre in casa. L'ultimo grafico ci presenta la media della differenza dei gol in funzione degli anni. Essa assume un andamento che tendenzialmente sembra essere costante.

➔ **Struttura della media dei gol al variare dei mesi.**



Tranne il caso della media della differenza dei gol in cui l'andamento non è molto interessante, se consideriamo il grafico della media dei gol delle squadre in casa, fuori casa ed il totale dei gol in funzione dei mesi, vediamo una struttura nei dati e inoltre si può notare che si segna in media di più negli ultimi tre mesi.

3.2.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati

→ Caso in cui la variabile di interesse sono i gol segnati dalle squadre in casa

Dopo tutte le analisi sull'andamento di modelli per i nostri dati, siamo rimasti con un modello in cui tutte e due variabili sono significative, ma la significatività non è tanto forte; successivamente il test di bontà del modello ci ha suggerito che il nostro modello non spiega bene la variabilità dei dati quanto il modello saturo, cioè il modello finale richiederebbe l'aggiunta di ulteriori variabili per poter migliorare la sua capacità predittiva.

→ Caso in cui la variabile di interesse sono i gol segnati dalle squadre fuori casa

Contrariamente al caso precedente, siamo rimasti con un modello in cui solamente la variabile Mesi risulta significativa (proprio ciò che ci aspettavamo dall'analisi esplorativa). Però anche in questo caso, il test di bontà del modello ci ha suggerito che il nostro modello finale non è adatto ai nostri dati, anzi non spiega bene la variabilità dei dati e di conseguenza richiederebbe l'aggiunta di ulteriori variabili per migliorare la capacità predittiva del modello.

→ Caso in cui la variabile di interesse è il totale dei gol segnati al termine di una partita

In questo caso dopo aver adattato un modello ai nostri dati risulta che al livello 5% le due variabili sono significative. Precisamente all' 1% la variabile anno non è significativa. Come in tutti gli altri casi, il test di bontà del modello ci ha suggerito che il modello finale al quale siamo arrivati non è adatto ai nostri dati quanto sarebbe stato il modello saturo. Quindi, richiederebbe l'aggiunta di ulteriori variabili nella speranza di migliorare la capacità predittiva del modello stimato.

→ Caso in cui la variabile di interesse è il totale dei gol segnati al termine di una partita

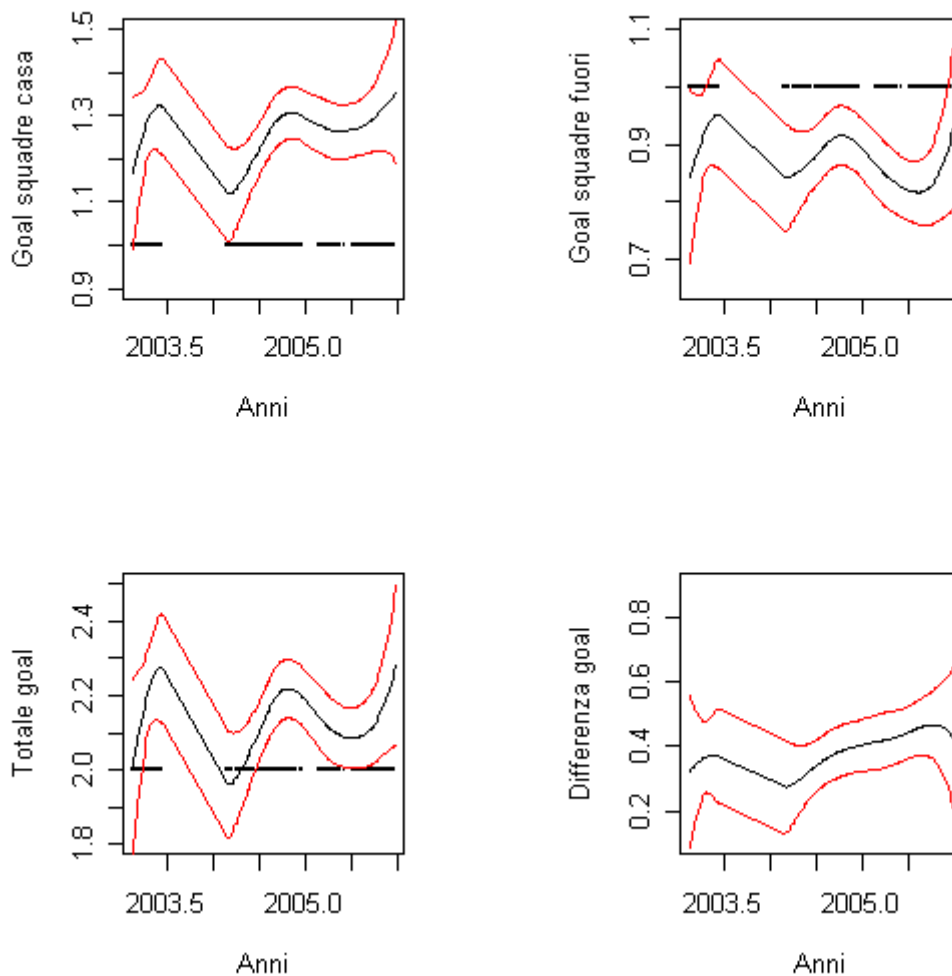
Per quanto riguarda la differenza dei gol per partita, applicando le procedure descritte nei capitoli precedenti, succede che, se dovessimo adattare un modello appropriato per i nostri dati, ci aspetteremmo che la sola variabile Anni sia significativa.

Osservazione: Anche qui, in tutti i modelli stimati per i dati della Serie B, succede che i modelli stimati non sono adeguati per la spiegazione dei dati. Il problema, potrebbe essere dovuto al fatto che i gol segnati non hanno proprio una distribuzione di poisson e in parte c'è una extra-variabilità di conseguenza della variabilità delle squadre.

3.3 CAMPIONATO SPAGNOLO SERIE C

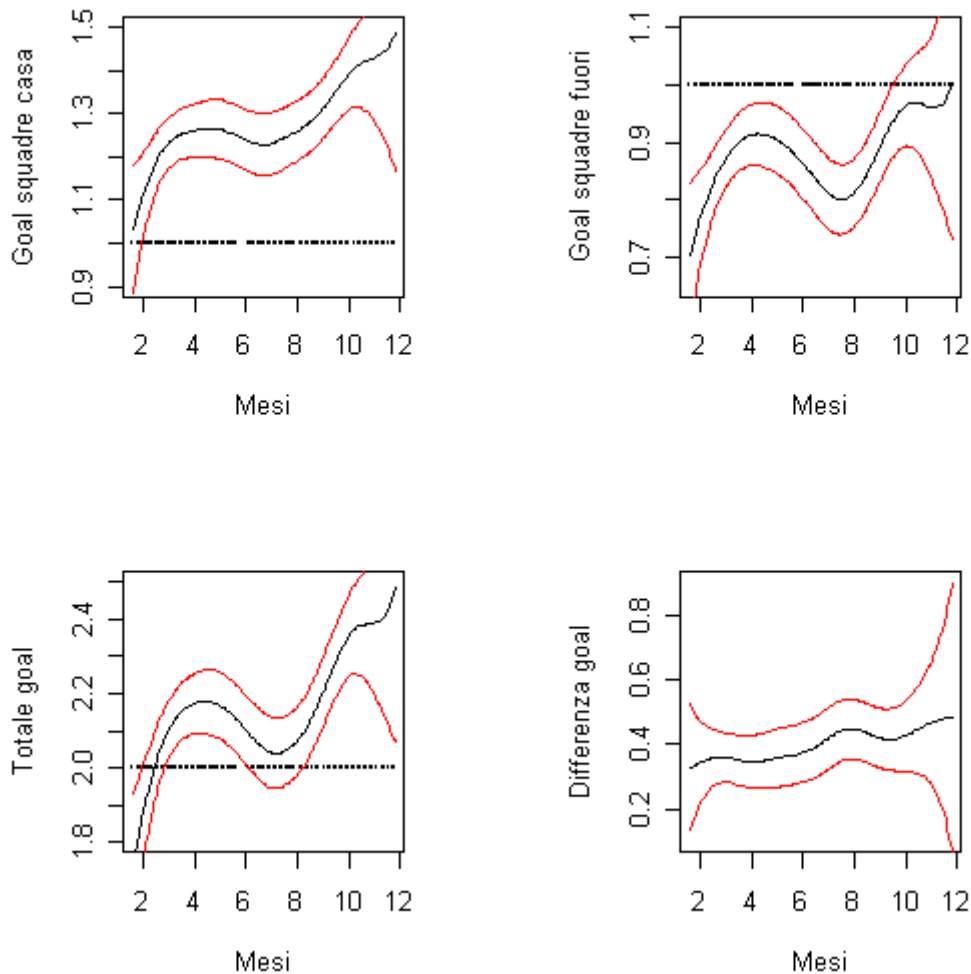
3.3.1 Analisi esplorativa

→ **Struttura della media dei gol al variare degli anni.**



La prima cosa da notare in questi grafici è che abbiamo i dati raccolti dal 2003 fino a 2005. Considerando i gol per le squadre in casa, fuori casa, per il totale dei gol e per la differenza, sembra difficile cogliere una struttura particolare al crescere degli anni, perché al crescere degli anni abbiamo delle oscillazioni. Nel caso della differenza dei gol, tuttavia, sembrerebbe esserci un andamento interessante in quanto prima decresce e poi cresce.

→ **Struttura della media dei gol al variare dei mesi.**



In questo caso, tranne la differenza dei gol per partita, in cui l'andamento sembra costante al crescere dei mesi, negli altri tre casi si evidenzia un struttura interessante simile in quanto all'inizio della stagione (ad agosto) cresce fino a novembre, poi decresce fino a febbraio e poi comincia a crescere fino alla fine della stagione (a giugno).

3.3.2 Riassunto dei risultati ottenuti dopo l'adattamento dei modelli per i dati.

→ Dopo aver adattato un modello per i gol segnati dalle squadre in casa, siamo arrivati ad un modello finale in cui al livello 5% le variabili Mesi e Anni sono significative, però è una significatività non tanto forte, cioè borderline. Successivamente, come in altri casi, il test di bontà di modello ci ha suggerito

che il modello a cui siamo arrivati non risulta adatto ai nostri dati in quanto non spiega bene la variabilità dei dati quanto il modello saturo; di conseguenza richiederebbe l'aggiunta di ulteriori variabili per poter migliorare la capacità predittiva, altrimenti avremmo delle previsioni dubbiose.

→ Per quanto riguarda i gol delle squadre fuori casa, siamo arrivati ad un modello finale in cui solo la variabile Mesi è rimasta significativa, poi nel successivo passo il test di bontà ci ha suggerito che il nostro modello non spiega bene la variabilità dei dati quanto quello saturo; sembrerebbe, infatti, richiedere l'aggiunta di ulteriori variabili per migliorare la sua capacità predittiva.

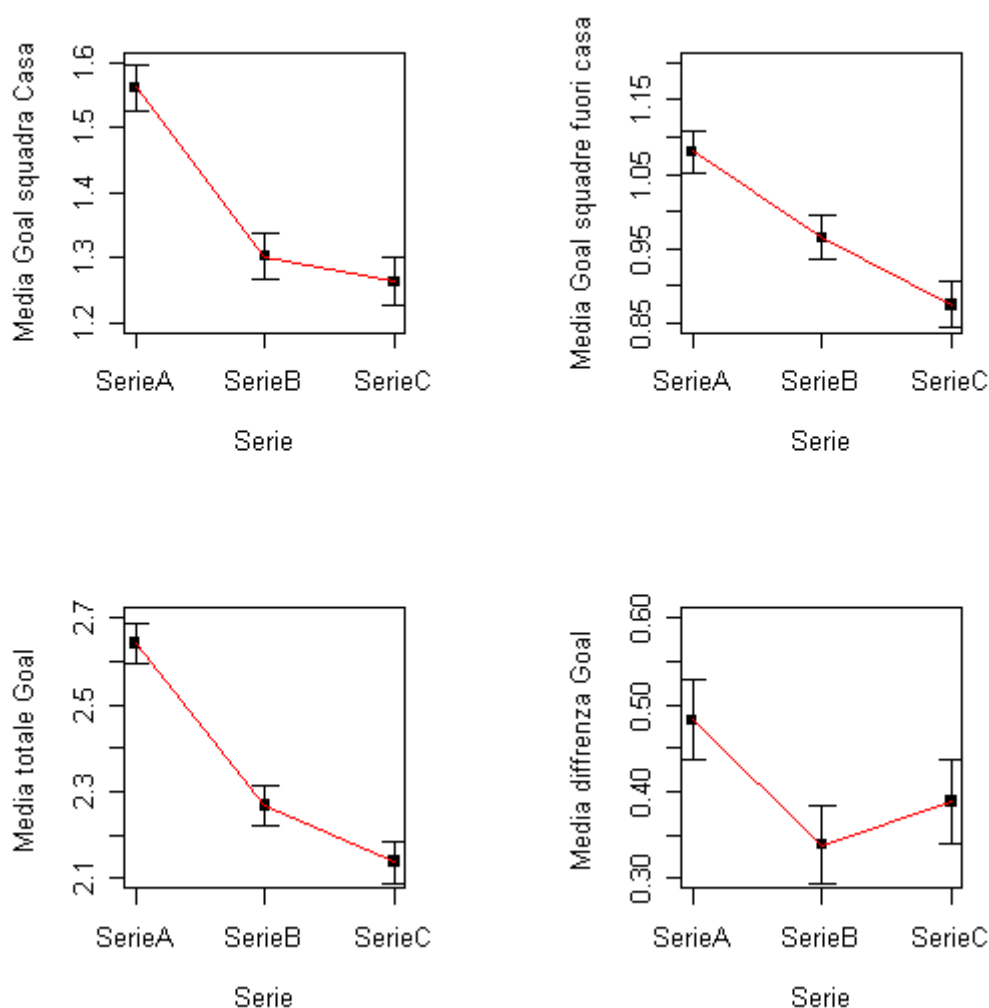
→ Per quanto riguarda il caso del totale dei gol per partita, siamo arrivati ad un modello in cui entrambe le variabili Mesi e Anni sono significative al 5%. Anche qui sorge lo stesso problema visto nei casi precedenti.

→ Per quanto riguarda la differenza dei gol per partita, applicando le procedure descritte nei capitoli precedenti, succede che, se dovessimo adattare un modello specifico per i nostri dati, ci aspetteremmo che una delle due variabili fosse significativa.

Osservazione: Anche qui, in tutti i modelli stimati per i dati della Serie C, succede che i modelli stimati non sono adeguati per la spiegazione dei dati. Il problema, potrebbe essere dovuto al fatto che i gol segnati non hanno proprio una distribuzione di poisson e in parte c'è una extra-variabilità di conseguenza della variabilità delle squadre.

3.3.3 Differenze tra le varie Serie indipendentemente dal mese e dell'anno

Analisi esplorativa



Questi quattro grafici ci permettono di confrontare le varie serie; in particolare se consideriamo il grafico in alto a sinistra, cioè il grafico della media dei gol delle squadre in casa al variare della categoria, risulta subito che in Serie A in media, si segna di maggiormente e più di un gol; seguono la Serie B e la Serie C in cui in media sembra che si segni lo stesso numero di gol. Di conseguenza diciamo che il numero di gol segnato dalle squadre in casa è influenzato dalla Serie.

Il secondo grafico in alto a destra ci suggerisce che in media le squadre fuori casa segnano maggiormente e più di un gol in Serie A, segue la Serie B che segna meno di 1 gol e infine la Serie C. Si vede anche in questo caso che gli intervalli non si sovrappongono, quindi ci aspettiamo che i gol segnati dalle squadre fuori casa siano fortemente influenzati dalla Serie.

Il terzo grafico in basso a sinistra si suggerisce che in media complessivamente si segna di più e più di 2 gol in Serie A, poi segue la serie B in cui si segnano 2 gol per partita e infine la Serie C in cui si segnano circa 2 gol per partita. Notiamo anche qui che gli intervalli non si sovrappongono e di conseguenza il totale dei gol per partita è fortemente influenzato dalla Serie.

Il quarto grafico in basso a destra ci suggerisce che in media la differenza dei gol per partita nei tre casi è meno di 1 gol, però è maggiore in Serie A, più o meno uguali in Serie B e in serie C, perché i due intervalli si sovrappongono. Quindi ci aspettiamo che la Serie abbia un effetto non tanto forte sulla differenza dei gol per partita. E' opportuno effettuare un test per vedere se ci sono differenze tra le varie Serie indipendentemente del mese o dell'anno.

- **Caso in cui la variabile di interesse sono i gol delle squadre in casa.**

Analysis of Deviance Table

Model 1: golC ~ 1

Model 2: golC ~ categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	13070	15882.3			
2	13068	15708.9	2	173.4	2.263e-38

Il valore del livello di significatività ci suggerisce, come ci aspettavamo, che la variabile categoria è fortemente significativa, cioè che i gol segnati dalle squadre in casa sono fortemente influenzati dalla Serie.

- **Caso in cui la variabile di interesse sono i gol delle squadre fuori casa.**

Analysis of Deviance Table

Model 1: golF ~ 1

Model 2: golF ~ categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	13070	15552.9			
2	13068	15457.5	2	95.4	1.925e-21

Anche qui, Il valore del livello di significatività ci suggerisce, come ci aspettavamo, che la variabile categoria è fortemente significativa, cioè che i gol segnati dalle squadre fuori casa sono fortemente influenzati dalla Serie.

- **Caso in cui la variabile di interesse è il totale dei gol per partita.**

Analysis of Deviance Table

Model 1: (golC + golF) ~ 1

Model 2: (golC + golF) ~ categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	13070	16216.0			
2	13068	15954.6	2	261.3	1.792e-57

Anche in questo caso, il valore del livello di significatività ci suggerisce, come ci aspettavamo, che la variabile categoria è fortemente significativa, cioè che il totale dei gol segnati per partita è fortemente influenzato dalla Serie.

- **Caso in cui la variabile di interesse è differenza dei gol per partita.**

Per il caso della differenza, risulta difficile applicare un modello gam ai dati, perché avremmo una differenza negativa. Se ci fidiamo delle analisi esplorative, risulta che la differenza dei gol per partita è influenzata dalla Serie.

3.3.4 Modelli di analisi delle differenze tra le varie Serie dato il mese e l'anno

- **Caso in cui la variabile di interesse sono i gol delle squadre in casa.**

Analysis of Deviance Table

Model 1: golC ~ s(Mesi)

Model 2: golC ~ s(Mesi) + categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	13066	15840.4			
2	13064	15674.8	2	165.6	1.088e-36

Su questo output si vede che non c'è la variabile Anni: l'abbiamo tolta perché non era significativa, inoltre il valore del livello di significatività ci suggerisce che il model 2 risulta preferibile al model 1 e di conseguenza la variabile categoria risulta fortemente significativa. Quindi, dato il mese, i gol segnati dalle squadre in casa sono fortemente influenzati dalla categoria. Considerando il modello finale, cioè il modello che contiene solo le variabili Mesi e categoria, il test di bontà del modello ci ha suggerito che il modello stimato non spiega bene la variabilità dei dati quanto il modello saturo e quindi richiederebbe l'aggiunta di ulteriori variabili nella speranza di migliorare la sua capacità predittiva.

- **Caso in cui la variabile di interesse sono i gol delle squadre fuori casa.**

Analysis of Deviance Table

Model 1: golF ~ s(Mesi)

Model 2: golF ~ s(Mesi) + categoria

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	13066.0000	15500.7			
2	13064.0000	15409.1	2.0000	91.6	1.268e-20

Anche in questo caso abbiamo tolto la variabile Anni perché non era significativa e con l'ultimo output stiamo cercando di vedere se la variabile categoria è significativa. Il valore del livello significatività ci suggerisce che è preferibile il Model 2; di conseguenza la variabile categoria è fortemente significativa; quindi, dato il mese, i gol segnati dalle squadre fuori casa sono influenzati dalla categoria; sorge, tuttavia, il problema riscontrato in tutti gli altri casi che riguarda la bontà del modello: il test ci ha suggerito che il modello finale cioè il modello contenente la variabile Mesi e categoria non spiega bene la variabilità dei dati e, quindi, la capacità predittiva è dubbiosa.

- **Caso in cui la variabile di interesse è il totale di gol per partita.**

Analysis of Deviance Table

Model 1: (golC + golF) ~ s(Mesi)

Model 2: (golC + golF) ~ s(Mesi) + categoria

Resid. Df Resid. Dev Df Deviance P(>|Chi|)

1 13066 16131.8

2 13064 15881.5 2 250.4 4.271e-55

Il valore del test ci suggerisce che è preferibile il model 2, cioè che la variabile categoria risulta fortemente significativa e di conseguenza, dato il mese, il totale dei gol per partita è influenzato dalla categoria. Essendo rimasti con un modello in cui solo il mese e la categoria sono significativi, il test di bontà di modello ci ha suggerito comunque che il modello stimato non spiega bene la variabilità dei dati quanto il modello saturo, cioè che il modello richiederebbe l'aggiunta di ulteriori variabili e quindi mette in dubbio la sua capacità predittiva.

Osservazione: Anche qui, in tutti i modelli stimati per i dati, succede che i modelli stimati non sono adeguati per la spiegazione dei dati. Il problema, potrebbe essere dovuto al fatto che i gol segnati non hanno proprio una distribuzione di poisson e in parte c'è una extra-variabilità di conseguenza della variabilità delle squadre.

CONCLUSIONI

Nella prima parte di questa tesi abbiamo passato in rassegna i diversi cambiamenti avvenuti nel settore di calcio ,con particolare riferimento al campionato di calcio francese ,italiano e spagnolo.Possiamo dire che nei tre campionati (francese,italiano,spagnolo) ci sono differenze nel segnare tra le varie Serie (Serie A,Serie B,Serie C). Un'altra cosa interessante che abbiamo notato nella maggioranza dei casi,nei tre campionati è che al crescere dei mesi sembrerebbe che all'inizio della stagione si segna di meno e di più negli ultimi tre o quattro mesi a seconda del campionato ,può essere dovuto al fatto che verso la fine del campionato ci sono squadre che competono per vincere il campionato e altre che competono per non essere retrocesse.Considerando i gol segnati dalle squadre in casa ,fuori casa,il totale e la differenza dei gol per partita ,indipendentemente dal campionato abbiamo notato che in pochi casi esisteva una struttura interessante al crescere degli anni ,invece rispetto ai mesi abbiamo evidenziato degli andamenti particolarmente interessante in cui sembrerebbe che dall'inizio della stagione cresceva o decresceva a seconda dei casi(gol squadre in casa,fuori casa,totale e differenza gol) ,poi sembrava essere costante e ricominciava a crescere fino alla fine della stagione. Dopo aver adattato dei modelli statistici per i nostri dati abbiamo scoperto che in 90% dei casi la variabile anno non risultava significativa,e in tutti i modelli finali in cui siamo arrivati ,il test di bontà del modello ci ha suggerito che il modello stimato non era in grado di spiegare correttamente la variabilità dei dati ,quindi richiedevano l'aggiunta di ulteriori variabili. Con le variabili messe a disposizione per la nostra analisi risulta che noi non siamo in grado di migliorare i modelli stimati per poter aver una capacità predittiva adeguata. Concludendo possiamo dire che il lavoro svolto finora può essere considerato come un primo passo verso la direzione di ricercare modelli più sofisticati che tengano in considerazione

eventuali squadre che hanno giocato le partite; ovvero analizzare tutti i modelli utilizzati includendo questa volta, gli effetti delle squadre.

BIBLIOGRAFIA

- Mark G. Dixon and Stuart G. Coles, "Modelling Inefficiencies for Betting on Association Football", S (1995)
- Dominico Piccolo *"statistica"* Bologna, Mulino, 1998
- A. Azzalini. *"Inferenza staticitica:una presentazione basato sul concetto di verosimiglianza"* Milano, Springer, 2004
- A. Azzalini, B. Scarpa *"Analisi dei dati e Dati mining"* Milano, Springer , 2004
- [http://it.wikipedia.org/wiki/Calcio_\(sport\)](http://it.wikipedia.org/wiki/Calcio_(sport))
- <http://www.calcio.sisal.it>
- <http://it.eurosport.yahoo.com/calcio/francia/ligue-1/index.html>
- <http://www.calcionew.it/squadrespagna.php>

