Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica e Tecnologie Informatiche

Relazione Finale
Mercato Ittico di Chioggia: Analisi della Banca Dati sul Pescato Giornaliero

Relatore: Prof. Livio Finos
Dipartimento di Psicologia dello sviluppo e della socializzazione

Correlatore: Prof. Carlotta Mazzoldi
Dipartimento di Biologia

Laureando: Alberto Peres
Matricola: 1010663

Anno Accademico 2014/2015

Indice

Introduzione

La Banca Dati della pesca a Chioggia, nasce intorno al 2011 con il desiderio di rendere disponibile le informazioni storiche sui prodotti ittici del Mar Adriatico Settentrionale a diversi tipi di utenti: non solo biologi marini o operatori della pesca, ma anche a semplici cittadini, studenti di vario tipo (grado di istruzione, ordine di studio).

Analizzare questi dati può fornire molte differenti informazioni per comprendere i cambiamenti della biodiversità e come l'uomo stia influenzando tali cambiamenti. Comprendere le variazioni temporali delle specie può aiutare nello sviluppo di strategie mirate per la salvagurdia di specie a rischio, dato il sovrasfruttamento delle risorse. Inoltre queste informazioni possono risultare utili anche agli adetti ai lavori, i pescatori per primi, i quali possono così confrontare la loro esperienza e conoscenza con dei dati quantitativi e su lunga scala temporale. Infine i dati della Banca Dati possono risultare molto utili anche ai giovani, agli studenti che vogliono informarsi sulle risorse presenti nel territorio marino circostante, per capire quanto e come viene sfruttato il mare o semplicemente per curiosità.

La Banca Dati, infatti, contiene tutte le statistiche sul pescato, raccolte al mercato ittico di Chioggia, dal 1945 ad oggi, è liberamente accessibile all'indirizzo `http://chioggia.biologia.unipd.it/banche-dati/` da dove, in pochi passaggi, è possibile conoscere quanto è stato pescato di ogni categoria di pesce anno per anno o mese per mese. I dati sono liberamente scaricabili in un foglio di calcolo, inoltre la Banca Dati contiene anche tutte le informazioni fornite dalla capitaneria di porto su numero, tonnellaggio e caratteristiche dei pescherecci locali.

La Banca Dati riporta:

- Le statistiche ufficiali dei prodotti ittici locali sbarcati al Mercato Ittico di Chioggia. Le informazioni presenti vanno dal 1945 ad oggi su base mensile.

- Le statistiche della flotta peschereccia di Chioggia dal 1991 ad oggi, su base mensile. I dati grezzi sono stati ottenuti dal registro delle Marinerie dell'Unione Europea.

- L'elenco delle segnalazioni di dati e informazioni raccolte direttamente da chi opera in mare e si dedica alle attività di pesca, dai pescatori professionisti, a quelli sportivi, a chi si immerge in acqua per osservare gli organismi marini .

Questa tesi si propone di presentare il lavoro svolto durante il periodo di stage presso la Stazione Idrobiologica "Umberto D'Ancona" di Chioggia, seguito dalla responsabile della stazione la dott.ssa Carlotta Mazzoldi, co-creatrice della banca dati, del Dipartimento di Biologia di Padova. L'idea di fondo è quella di creare un procedimento di lavoro, una sorta di "manuale", per gli adetti ai lavori nel campo biologico ma che non sono specializzati in elaborazioni statistiche. Questo manuale permetterà loro di poter analizzare i dati più comodamente avendo già a disposizione tutti i codici per l'eleborazione.

Nel Capitolo 1 verrà presentato il lavoro di input e pulizia dei dati, volto alla creazione del dataset finale sul quale si andrà a lavorare per lo studio e l'elaborazione delle funzioni riutilizzabili nei prossimi anni, tutto, a partire dai dati dei foglio di calcolo della Banca Dati. Inoltre verranno presentati i problemi riscontrati in fase di pulizia e le ricodifiche attuate.

Il Capitolo 2 si occuperà delle serie storiche, cioè dell'andamento dei dati nel corso degli anni per studiare la dipendenza tra osservazioni successive, se ci possono essere fattori di stagionalità che influiscono sui dati e per prevedere l'andamento futuro che si potrà avere.

Nel Capitolo 3 sono studiate le correlazioni che intercorrono tra le varie specie pescate nell'Adriatico Settentrinale (118 quelle dal 1997) nel passare

degli anni e nei cambi stagionali, specificatamente si vedrà come le quantità di pescato siano cambiate nel corso dei 17 anni di studio.

Per questi studi e per la creazione del dataset finale sono state utilizzate delle funzioni elaborate col software statistico *R*, tali funzioni sono riportate in Appendice.

Capitolo 1

Input Dati e Pulizia

In questo capitolo verrà presentato il lavoro svolto per l'input dei dati, la pulizia e la loro automazione. Il lavoro può essere diviso in due blocchi, una parte preliminare, dove principalmente si è creato il dataset finale e la parte di elaborazione dei dati.

1.1 Creazione del dataset finale

Per la creazione del dataset finale, dati i dataset dei vari anni, si sono dovute compiere alcune correzioni. Lo sviluppo per arrivare al dataset finale si divide in due parti, *DATA CLEANING* dove si è svolto il lavoro di pulizia dei dati ed il *DATA MANAGEMENT* dove si è sistemato definitivamente il dataset in modo da arrivare ad un risultato chiaro per la lettura finale, togliendo quegli elementi che non servivano e rendendolo il più leggibile possibile. I codici *R* utilizzati per questi processi sono riportati in appendice.

1.1.1 Data Cleaning

Questo processo in informatica fa riferimento alla capacità di assicurare correttezza ed affidabilità dei dati in uso, così da evitare errori e garantire accuratezza sia dal punto di vista sintattico che semantico e avere la possibilità di rilevare dati non desiderati, inesatti o scoretti. Dopo aver analizzato visivamente i dataset annuali si è proceduto a pulirli (cleaning) per uniformarli

tra loro. Come si vedrà successivamente nella sezione *Errori dei Dati e Ricodifiche* i file annuali excel non sono stati creati tutti nello stesso formato, ma presentano impaginazioni molto differenti tra gi anni con l'inserimento di campi aggiuntivi non utili al nostro studio e la ridenominazione di altri. Inoltre anche le denominazioni delle specie sono mutate, quindi, si è dovuto trovare il modo per uniformare tutti gli anni. Il primo passo dopo l'apertura del file è stata quella di selezionare le colonne utili all'analisi, viste le numerose colonne che non sarebbero servite, cioè, la data del giorno di pesca, la specie, il codice della specie, il peso, il prezzo di vendita, la zona di pesca e numero di pescatori. Successivamente si è proceduto alla modifica di alcuni nomi di specie denominati in modo errato (si veda sezione 1.1.3 per approfondire), così da rendere omogeneo l'intero dataset.

Risultato finale del data cleaning per l'anno 2010:

```
       data mese anno cod.spe specie peso.kg prezzo valore tpr
7  20100105    1 2010    1060  ALICE       7   0.71      5  MA
8  20100105    1 2010    1060  ALICE       7   0.71      5  MA
9  20100105    1 2010    1060  ALICE       7   0.71      5  MA
10 20100105    1 2010    1060  ALICE       7   0.86      6  MA
11 20100105    1 2010    1060  ALICE       7   0.86      6  MA
12 20100105    1 2010    1060  ALICE       7   2.14     15  MA
                            produttore
7  SNC DI PERINI GIMMY E PERINI DANILO
8  SNC DI PERINI GIMMY E PERINI DANILO
9  SNC DI PERINI GIMMY E PERINI DANILO
10 SNC DI PERINI GIMMY E PERINI DANILO
11 SNC DI PERINI GIMMY E PERINI DANILO
12 SNC DI PERINI GIMMY E PERINI DANILO
```

1.1.2 Data Management

In questo processo, in italiano gestione dei dati, dopo aver ripulito il dataset nella fase precedente, si passa allo sviluppo per arrivare al dataset finale vero e proprio, sul quale svolgeremo le analisi. Dato il dataset ripulito si vuole arrivare ad avere, alla fine, lo stesso dataset ma con l'aggiunta del conteggio di pescherecci che hanno pescato quella specie in quel giorno, inolte si vuole avere la divisione del peso nei vari luoghi di pesca (mare, laguna, valle e acqua dolce). Queste informazioni aggiuntive non serviranno alla nostra analisi ma possono risutare molto utile per altri studi su questi dati, e non sono facilmente estrapolabili dai dati iniziali.

Di seguito un un'esempio di come risulta il dataset alla fine dei due processi, utilizziamo sempre il file del 2010:

```
      data mese anno cod.spe  specie peso.kg prezzo valore peso.MA
1 20100302    3 2010    1020 AGUGLIA    1.50    4.0   6.00       0
2 20100320    3 2010    1020 AGUGLIA    3.00    3.0   9.00       0
3 20100323    3 2010    1020 AGUGLIA    3.75    2.5   9.37       0
4 20100330    3 2010    1020 AGUGLIA    9.25    4.5  41.62       0
5 20100403    4 2010    1020 AGUGLIA    5.00    2.0  10.00       0
6 20100407    4 2010    1020 AGUGLIA    3.75    4.0  15.00       0
  peso.LA peso.AD peso.VA nprod
1    1.50       0       0     1
2    3.00       0       0     1
3    3.75       0       0     1
4    9.25       0       0     1
5    5.00       0       0     1
6    3.75       0       0     1
```

Conclusi i due processi per i tredici anni di studio si procede a creare un dataset unico, il dataset finale, sul quale andremo poi a lavorare ed a compiere il nostro studio.

Il codice *R* completo dei processi precedentemente spiegati si può trovare in appendice al punto 1.

1.1.3 Errori dei Dati e Ricodifiche

Durante il processo di pulizia dei dati (*DATA CLEANING*) si sono presentati alcuni problemi, il principale dei quali è stato il cambio nella redazione dei dataset annuali da parte dei responsabili del mercato ittico.

Per i primi dataset (1997-2006) la struttura del file era sempre la stessa, venivano indicate data di pesca, codice specie, nome specie, peso, prezzo, valore, zona di pesca e alcuni dati sulla vendita, che nell'elaborazione non sarebbero serviti e quindi sono stati eliminati. Nel proseguo degli anni si sono cominciate ad adottare diverse modalità di redazione dei dataset, più di una dopo il 2010, aggiungendo molte più variabili rispetto a quante ce ne fossero prima, ad esempio codici intermedi di vendita oppure non più solo chi vendeva il pescato ma anche chi lo comprava e per conto di chi. Tutti dati non utili all'elaborazione e che quindi non presi in considerazione nello studio successivo.

Inoltre dall'anno 2011, oltre all'aggiunta di molte variabili, è cominciato un processo di ridefinizione di quella già esistenti, ad esempio, prima per indicare il giorno di pesca veniva usata la variabile "data" successivamente si è passati ad un meno comprensibile "DocDtHH" oppure per indicare il prezzo al kg veniva usata la variabile "prezzo" divenuta poi "DocRgPre". Questo cambio di definizioni ha reso la lettura dei file più complessa e meno immediata, soprattutto ai non adetti ai lavori, rispetto a quella dei primi anni che risultava molto più semplice alla lettura e all'interpretazione.

Un'altro problema collegato al precedente riscontrato durante la pulizia dei dati è stato quello del cambio di denominazione del luogo di pesca, passato da MA,LA,AD e VA a Mare, Laguna, Acqua Dolce e Valle. Questo cambiamento, a differenza del precedente, non ha creato problemi di comprensione, anzi, ma ha creato complicanze dal punto di vista della creazione dello *script*, che a causa di questa mutazione non risulta omogeneo alla lettura e chiaro all'utilizzo.

Infine per quanto concerne i problemi nei vari dataset, in alcuni casi (dal

2010 al 2013) i dataset presentavano anche prodotti non locali, ma importati sia dall'Italia che dall'estero. Il più delle volte erano segnalati ma in alcuni casi, che vedremo successivamente, questa informazione era omessa.

Per quanto rigurda gli errori, per la quasi totalità dei casi, sono dovuti ad un'errata trascrizione dei nomi delle specie. Il punto principale è che i dati del pescato giornaliero vengono dati dai pescatori ai responsabili del mercato, i pescatori nel porto di Chioggia sono per la maggior parte del luogo e come lingua "ufficiale" adottano il dialetto, da qui la maggior parte degli errori riscontrati nei dataset.

Ripartiamo alcuni esempi di modifiche effettuate, per la lista completa si rimanda alla Tabella 2.1:

- "ARINGA" questa specie è diffusa lungo le coste dell'Atlantico settentrionale, quindi è impensabile possa trovarsi in Adriatico. L'errore è dovuto al fatto che i vecchi pescatori chiamano le "ALICE" *renghe* da qui aringhe.

- "GRANCHIO DA MOLECA" si definisce Moleca, termine dialettale, il granchio verde in fase di muta, quando è senza carapace e si presentano teneri e molli, quindi per differenziarli dai granchi comuni verranno rinominati "GRANCHIO(Moleca)". Uguale discorso per la femmina, "GRANCHIO(Mazaneta)".

- "BISO" è una specie che vive nell'Atlantico nella fascia tropicale, viene spesso confuso per le caratterstiche esteriori con il "TOMBARELLO" o "TONNO TOMBARELLO", questo tipo di errore non è dovuto alla lingua ma bensì al riconoscimento della specie da parte del pescatore, da cui l'errata trascrizione.

In alcuni casi si è preferito utilizzare e mantenere il termine dialettale per uniformare il dataset finale, per quanto rigurda le specie non presenti nell'Adriatico che accidentalmente sono rientrate nei dataset, non sono state prese in considerazione e sono state eliminate dallo studio.

Nome Specie da Modificare	Nome Finale	Motivazioni
ARINGA	ALICE	Errata traduzione
ALICE (cxkg)	ALICE	Differenziazione inutile
BISO O TOMBARELLO	TONNO TOMBARELLO	Errato riconoscimento pesce
CANNOLICCHIO O CAPPALUNGA	CANNOLICCHIO	Nome comune dialettale
CAPPELLANO O BUSBANA	BUSBANA	Nome comune dialettale
CEFALO O BOSEGA	CEFALO BOSEGA	Uniformare dataset
CEFALO O CALAMITA	CEFALO CALAMITA	Uniformare dataset
CEFALO O LOTREGANA	CEFALO LOTREGANA	Uniformare dataset
CEFALO O VERZELATA	CEFALO VERZELATA	Uniformare dataset
CEFALO O VOLPINA	CEFALO VOLPINA	Uniformare dataset
GALLINELLA O CAPPONE	GALLINELLA	Uniformare dataset
GHIOZZETTO	GHIOZZETTO MINUTO	Uniformare dataset
GHIOZZETTO Gò	GHIOZZO	Nome comune dialettale
GRANCHIO DA MOLECA	GRANCHIO(Moleca)	Uniformare dataset
GRANCHIO MAZANETA	GRANCHIO(Mazanta)	Uniformare dataset
GRANSEOLA O GRANCEOLA	GRANCEOLA	Uniformare dataset
LATTERINO O ACQUADELLA	LATTERINO	Nome comune dialettale
LUCCIOPERCA O SANDRA	LUCCIOPERCA	Nome comune dialettale
MERLANO O MOLO	MOLO	Nome comune dialettale
NASELLO O MERLUZZO	MERLUZZO	Uniformare dataset
PANNOCCHIA O CANNOCCHIA	PANNOCCHIA	Nome comune dialettale
SEPPIA (T/PICCOLA)	SEPPIA (tag/P)	Uniformare dataset
SPIGOLA O BRANZINO	BRANZINO	Errato riconoscimento pesce
SURO O SUGARELLO	SURO	Nome comune dialettale
ZANCHETTA O SUACIA	ZANCHETTA	Nome comune dialettale

Tabella 1.1: Modifiche eseguite ad alcune specie

Capitolo 2

Analisi delle Serie Storiche

2.1 Serie Storica

Per *serie statistica* si intende un'insieme di variabili casuali ordinate secondo un criterio qualitativo. Quando il criterio ordinatore dei dati è il tempo, inteso come pregressione cronologica, si ha una *serie storica* (o temporale). Possiamo pertanto definire serie storica *una successione di dati numerici nella quale ogni dato è associato ad un particolare istante od intervallo del tempo* (Vianelli, 1983).

In una serie storica è lecito presumere che vi sia dipendenza tra osservazioni successive e che essa sia legata alla posizione dell'osservazione della sequenza. Le serie storiche vengono studiate sia per interpretare un fenomeno, individuando componenti di trend, di ciclicità, di stagionalità e/o di accidentalità, sia per prevedere il suo andamento futuro.

L'analisi statistica di una serie storica si propone di chiarire il meccanismo casuale che l'ha generata, o per dare una breve descrizione delle caratteristiche della serie, oppure per prevedere l'evoluzione del fenomeno osservato, di cui è nota la storia passata.

L'analisi della serie storica può avere diversi obbiettivi:

- descrivere sinteticamente l'andamento nel tempo di un fenomeno; il

grafico di una serie, in particolare, mette facilmente in evidenza sia eventuali irregolarità, sia valori anomali;

- spiegare il fenomeno, individuare il suo meccanismo generatore ed eventuali relazioni con altri fenomeni;
- filtrare la serie; con ciò si intende la composizione della serie stessa nelle sue componenti non osservabili;
- prevedere l'andamento futuro del fenomeno.

Nel nostro studio l'utilizzo di serie storiche è fondamentale in quanto avendo dati a cadenza annuale possiamo vedere l'andamento nel tempo della pesca, in particolare vedremo come le quantità di pescato cambiano durante l'anno e nei vari anni, se ci sono fattori che influenzano l'andamento della pesca e come quasta sia cambiata nel corso dei 17 anni in esame.

Vedremo ora un'esempio di ciò, presa in considerazione una delle specie tra le più pescate nell'Adriatico, la SARDINA, mediante un preliminare studio, con l'apporto di un grafico di serie[1], si noteranno alcune particolarità.

[1] codice visibile in appendice al punto 2.

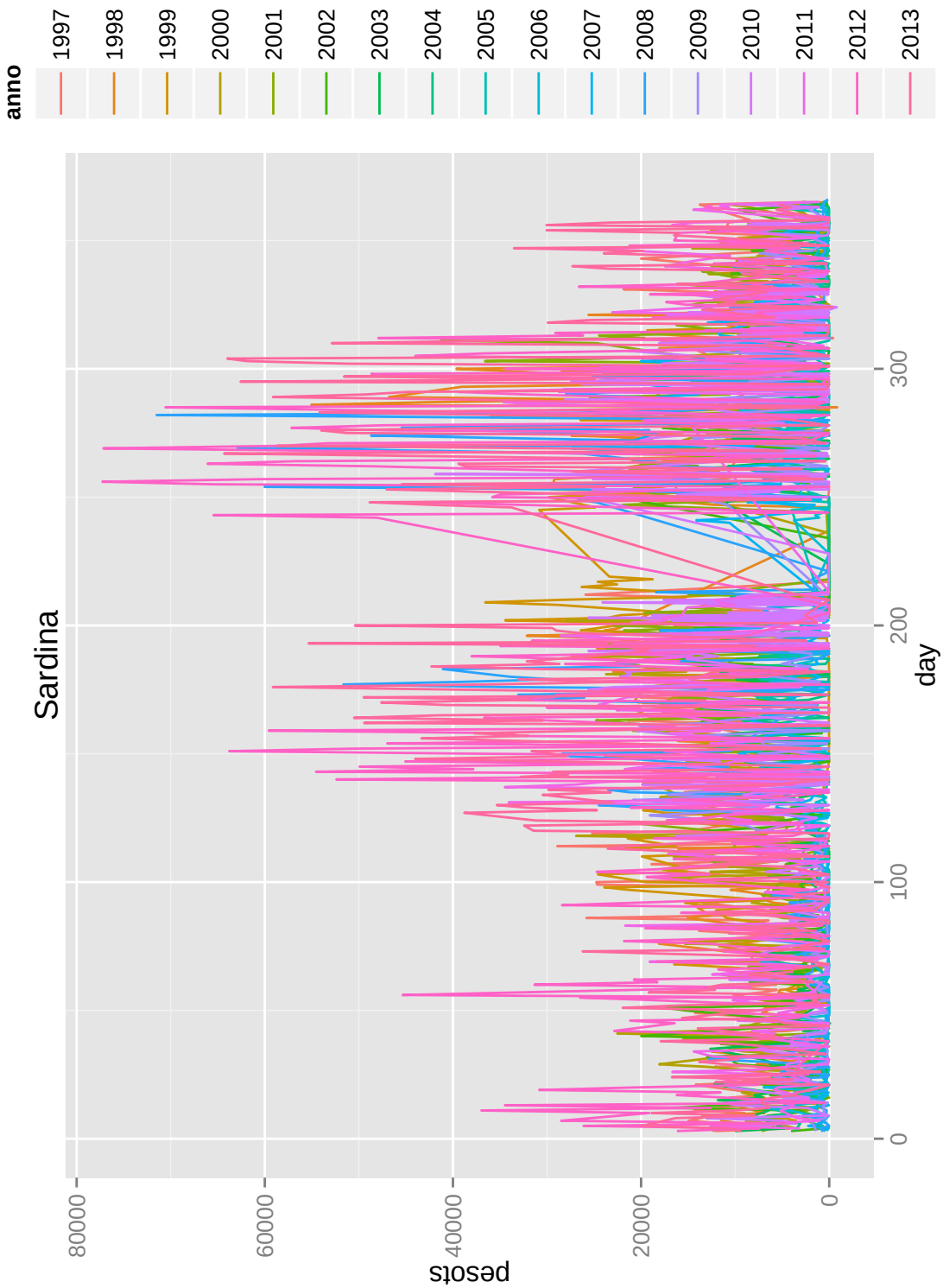


Figura 2.1: Grafico della Serie Storica dell'Alice nel corso degli anni di studio

Il grafico in Figura 2.1, rappresenta la serie storica relativa alla specie *ALICE* per gli anni dal 1997 al 2013, dove viene messo in relazione il peso pescato per ogni giorno di pesca effettuato. Il grafico non è di facile interpretazione a causa della numerosità campionaria molto elevata, ma si possono notare due particolari molto significativi:

- Una "pausa": intorno al 200° giorno possiamo notare un "buco" che dura circa una ventina di giorni, questo fenomeno, che si può riscontrare per qualsiasi specie di pesce per ogni anno di studio, è dovuto al fermo pesca[2], periodo, circa il mese di Agosto, in cui nessun peschereccio, tranne poche eccezzioni, può uscire a pescare.

- Valori negativi: questi valori, molto pochi rispetto al totale, sono relativi alle note di addebito e accredito sulle vendite[3]. Quando una pesata è registrata in modo errato durante la settimana, il lunedì successivo il peso errato viene segnato in negativo, così da bilanciare il totale. Infatti, andando a verificare, ogni valore negativo del grafico corrisponde ad un lunedì. Questi valori possono procurare dei problemi all'analisi, quindi si è definita una strategia di procedimento differente per far si che questi dati non influiscano sull'analisi.

Per ovviare a quest'ultimo problema si è valutata una modifica in termini di valori di riferimento temporali. In pratica si è passati da rilevazioni giornaliere a rilevazioni settimanali, questo cambiamento ha fatto sì che si potessero sommare i valori del peso per specie settimanalmente così facendo i valori negativi presenti (cioè le rettifiche sul venduto) si sono sommati con i valori errati precedentemente trascritti, annullandosi tra loro.

Si è successivamente creata una variabile "*week*" per il conteggio delle settimane di pesca, la quale ha permesso di poter sommare i pesi delle specie nell'intera settimana di pesca, questa modifica ci garantisce di non avere più valori negativi. Inoltre se si fossero presentati giorni i cui non si pescava,

[2]Divieto di pesca, applicato in periodi e zone determinate, per favorire il ripopolamento e la riproduzione delle specie ittiche.

[3]Giro di partite addebitate a commercianti per errore che a loro volta vengono girate ai commercianti che veramente hanno acquistato le merci

anche più di uni consecutivo, la nostra serie ne sarebbe stata influenzata ma lavorando per settimane questo problema viene superato, ad eccezzione del periodo di fermo pesca.

Si è giunti ad un dataset di questo formato[4]:

```
  week anno cod.spe   specie peso.kg   prezzo valore
1   13 2008    1092 ABRAMIDE   43.0 0.500000  21.50
2    2 1997    1020  AGUGLIA   95.5 2.780000 215.41
3    6 1997    1020  AGUGLIA   14.0 0.555000   7.77
4    8 1997    1020  AGUGLIA   14.5 1.805000  29.66
5    9 1997    1020  AGUGLIA    4.0 2.580000  10.32
6   10 1997    1020  AGUGLIA   83.5 2.189206 187.54
  peso.MA peso.LA peso.AD peso.VA nprod
1    43.0     0.0       0       0     1
2    95.5     0.0       0       0     1
3    14.0     0.0       0       0     1
4     9.5     5.0       0       0     1
5     0.0     4.0       0       0     1
6     0.0    83.5       0       0     2
```

A questo punto il nostro dataset è definitivamente pronto per l'analisi, non sono più presenti valori negativi e si può procedere nuovamente con l'analisi grafica vista in precedenza, il codice utilizzato è lo stesso del punto 2 in appendice. Il nuovo grafico, in Figura 2.2, risulta molto più leggibile del precedente, qui la serie storica della $SARDINA$ è sempre dal 1997 al 2013 ma ha frequenza setimanale. Si può notare anche in questo caso il "buco" presente nel mese di Agosto, corrispondente circa alle settimane 33-35.

Questo grafico può essere fatto per tutte le specie di pesce da analizzare, basta modificare la specie nei codici in appendice al punto 2. Vediamo di seguito, in Figura 2.3, invece la serie storica completa, sempre per la specie SARDINA, per tutti i 17 anni:

[4]i codici sono visibili in appendici al punto 3

Figura 2.2: Grafico per la specie Sardina, espresso in peso pescato settimanalmente.

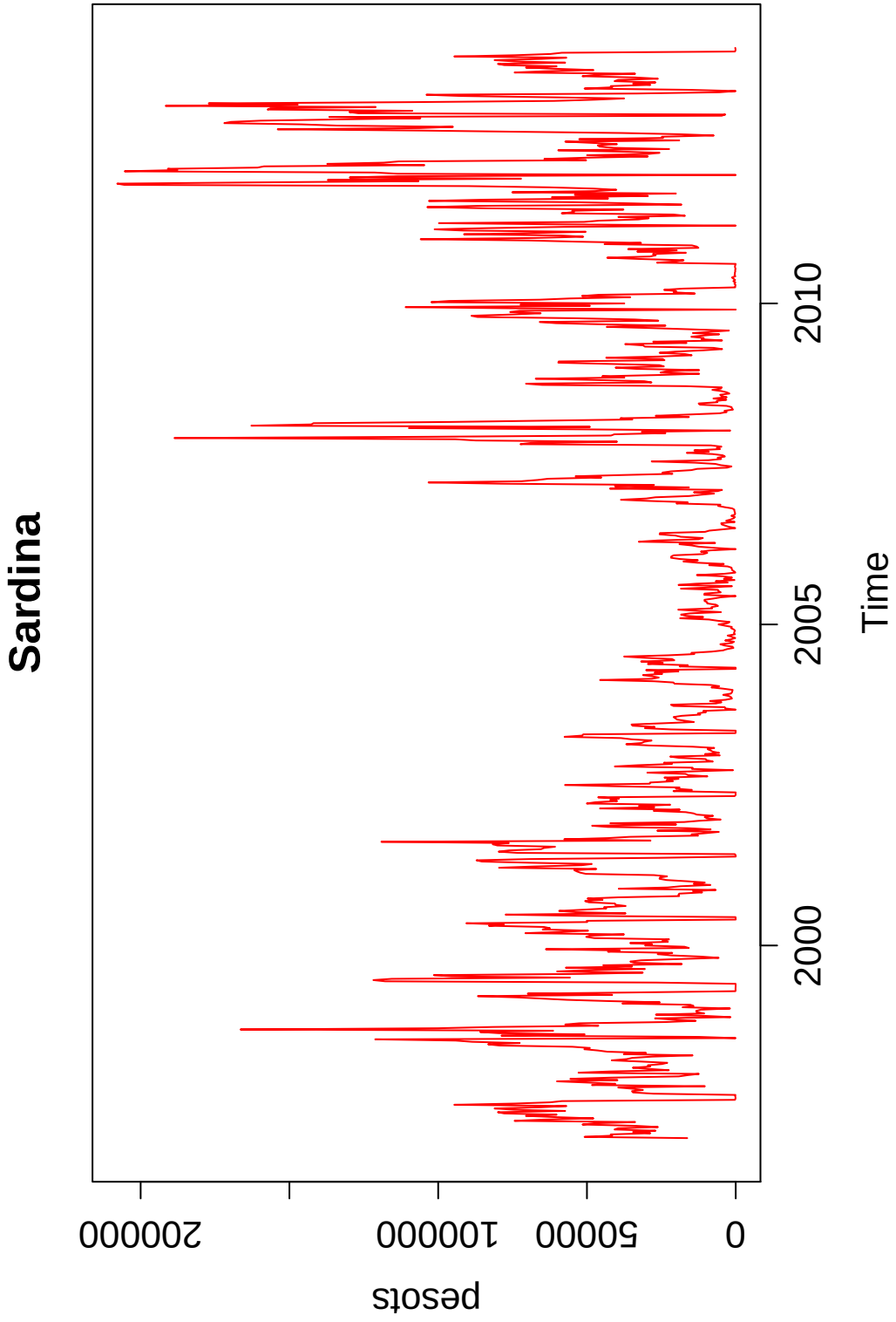


Figura 2.3: Serie storica della Sardina per tutti gli anni in esame.

In questa figura la serie è completa, sono presenti tutti gli anni in successione. Da questo grafico si può notare come l'andamento del peso del pescato della SARDINA ogni anno abbia due fasi, una creswcita fino a metà anno, per poi diminuire drasticamente fino ad Agosto (fermo pesca) ed una successiva crescita che culmine tra Novembre e Dicembre. I codici sono visibili in appendice al punto 3.

2.2 Il Modello ARIMA, Processo AutoRegressivo a Media Mobile

In statistica per modello ARIMA (da *AutoRegressive Integrated Moving Avarange*) si intende un modelo usato per indagare serie storiche, sia per comprendere al meglio i dati sia per prevedere i punti futuri nella serie. Questo modello viene applicato nei casi in cui i dati mostrano evidenza di non-stazionarietà[5].

Sia $\{\varepsilon_t\}$ un processo *white noise* di media 0 e varianza σ_ε^2. Si indichi con X_t la d-esima differenza di Y_t, $X_t = (1-B)^d Y_t$. Si dice che $\{Y_t\}$ è un processo autoregessivo integrato a media mobile di ordine (p,d,q), e lo si indica con ARIMA(p,d,q), se $\{X_t\}$ è un processo ARMA(p,q). In sintesi, pertanto, valgono le seguenti relazioni:

$$X_t = (1-B)^d Y_t,$$

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}.$$

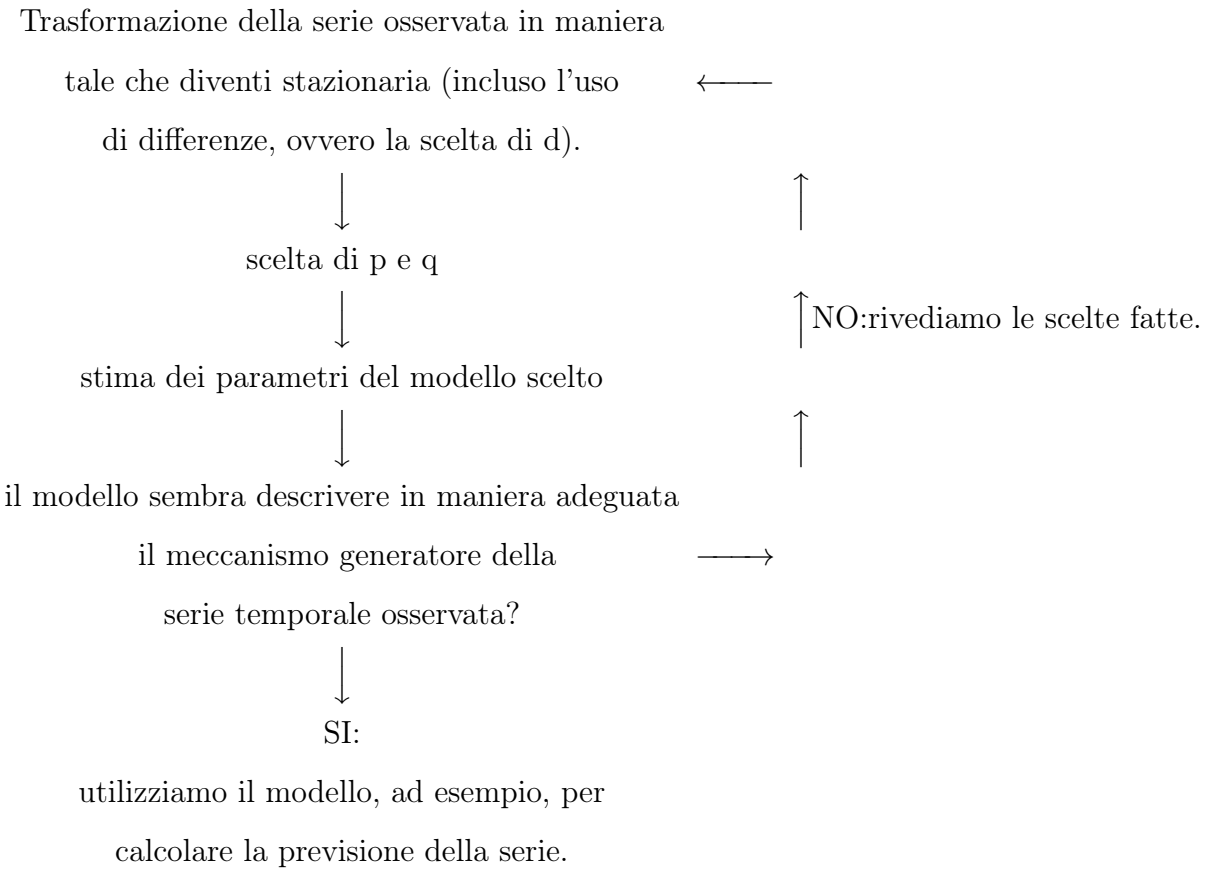
dove indichiamo con B l'operatore ritardo che trasforma la serie Y_t nella serie ritardata di periodo $Y_{t-1} = BY_t$. Utilizzando l'opertore differenza ed i polinomi, rispettivamente, autoregressivo ed a media mobile, la relazione precedente può anche essere scritta in termini di Y_t:

$$\phi(B)(1-B)^d Y_t = \phi_0 + \theta(B)\varepsilon_{t-j}.$$

[5]Un processo si dice non-stazionario quando la sua distribuzione di probabilità congiunta non cambia nel tempo, cioè quando parametri quali media e varianza, non cambiano nel tempo e non seguono una tendenza, non sono costanti nel tempo.

In generale il modello viene indicato come ARIMA(p,d,q), dove AR, autoregressione ha ordine p, I, integrazione, ha ordine d e MA, media mobile, ha ordine q. Questo modello costituisce una parte importante dell'approccio Box-Jenkins (1976) alle serie temporali, la procedura che seguiremo è di tipo iterativo e serve per: l'identificazione, la stima dei parametri e la verifica di un modello ARIMA. In alcuni casi queste fasi possono essere ripetute più volte come riportato nello schema successivo. Lo scopo è costruire un modello che si adatti alla serie storica.

Schema per l'identificazione di un modello ARIMA.



Procediamo con lo studio delle specie, prenderemo in esame sei specie, quelle più pescate in adriatico: *"ALICE"*, *"LATTERINO"*, *"PANNOCCHIA"*, *"SARDINA"*, *"SEPPIA"* e *"SOGLIOLA"*.

Primo passo è quello di specificare l'ordine del modello, ovvero identificare i parametri p,d,q. I principali strumenti da usare sono la funzione di autocorrelazione (ACF) e la funzione di autocorrelazione parziale (PACF). Utilizziamo i dati del peso, precedentemente trasformati in serie storica, e creiamo i grafici di ACF e PACF. Dato che, come detto nell'introduzione, questa tesi ha lo scopo di creare un manuale per facilitare l'analisi di questi tipi di dataset, dopo un controllo grafico, per implementare lo studio utilizzeremo il comando *auto.arima*, il quale ci restituisce il migliore modello ARIMA secondo i vincoli di ordine previsti. Questo comando aiuta molto l'automazzione dello studio, ma non può sostituirsi in toto all'analisi analitica, è comunque un'ottimo sostegno per chiunque voglia fare un'analisi delle serie storiche ma è alle prime armi. Per far ciò utilizzeremo i comandi R che si trovano in appendice al punto 4.

Vediamo di seguito i grafici delle autocorrelazioni e delle autocorrelazioni parziali per le sei specie in esame e l'uso del comando *auto.arima*. Andiamo ad analizzare i risultati.

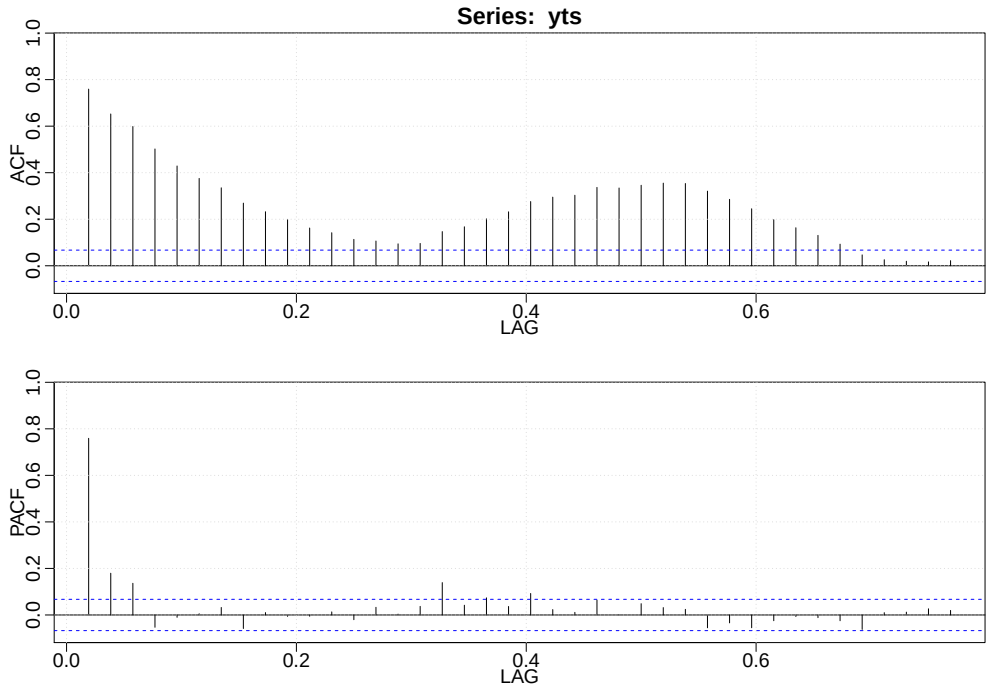


Figura 2.4: Autocorrelazione e autocorrelazione parziale specie ALICE.

Dal grafico in Figura 2.4 si nota come le autocorrelazioni tendano ad avere un'andamento decrescente e notiamo una stagionalità nei dati, come ci aspetteremo data la natura dei nostri dati. Le autocorrelazioni tendono ad annullarsi lentamente, quindi c'è la possibilità che i dati per questa specie non siano stazionari. Vediamo ora il risultato dell'*auto.arima*:

```
Series: yts
ARIMA(3,1,2)(2,0,0)[52]

Coefficients:
          ar1     ar2     ar3     ma1      ma2    sar1   sar2
      -0.4594  0.6129  0.1492  0.0254  -0.9347  0.0454  0.066
s.e.   0.0387  0.0413  0.0362  0.0179   0.0165  0.0324  0.029

sigma^2 estimated as 501311351:  log likelihood=-10097.38
AIC=20190.49   AICc=20190.66   BIC=20228.76
```

Dall'output del comando *auto.arima* notiamo un particolare nuovo, il modello è nella forma ARIMA(p,d,q)x(P,D,Q). Un modello di questo tipo viene chiamato ARIMA stagionale (SARIMA, Seasonal ARIMA), sono dei modelli misti di componenti a media mobile e di componenti autoregressive, che tengono conto dell'eventuale non stazionarietà e stagionalità di una serie. Tali modelli cercano di spiegare l'andamento di una serie storica basandosi sulla storia passata, descrivendo il fenomeno attraverso l'adattamento sia della parte stagionale sia della parte non stagionale, perchè ciò riusulti possibile è necessario che la serie studiata sia caratterizzata da una forte correlazione seriale ai ritardi stagionali (a distanza settimanale nella nostra analisi). La nuova equazione per questo modello è:

$$\phi(B)\Phi(B^S)(1-B)^d(1-B^S)^D Y_t = \phi_0 + \theta(B)\Theta(B^S)\varepsilon_t$$

dove, S è il periodo stagionale, nel nostro caso lavorando in settimane S=52. L'idea che sta alla base dei modelli SARIMA è che il processo deve poter descrivere due tipi di relazioni all'interno della serie osservata: la correlazione tra valori consecutivi, che è quella modellata dai normal modelli ARIMA, e la correlazione tra osservazioni che distano tra loro di un multiplo del periodo.

Analizzando il risultato di *auto.arima* e confrontandolo con i grafici ACF e PACF, possimo vedere come i valori dei parametri p,d,q siano coerenti, anche per quanto riguarda i valori degli "operatori stagionali", P,D,Q, suggeriti non sono necessari cambiamenti. Possiamo quindi utilizzare questo modello per calcolare delle previsioni future, usando il comando *forecast*, contentuto nella *library(forecast)*, sul modello eseguiamo il grafico di Figura 2.5.

Da questo grafico possiamo vedere la previsione a cinque anni del peso per l'ALICE, la linea blu è la linea di previsione, mentre le due aree colorate mostrano l'intervallo di predizione all'80%, l'area più scura, ed al 95% quella più chiara.

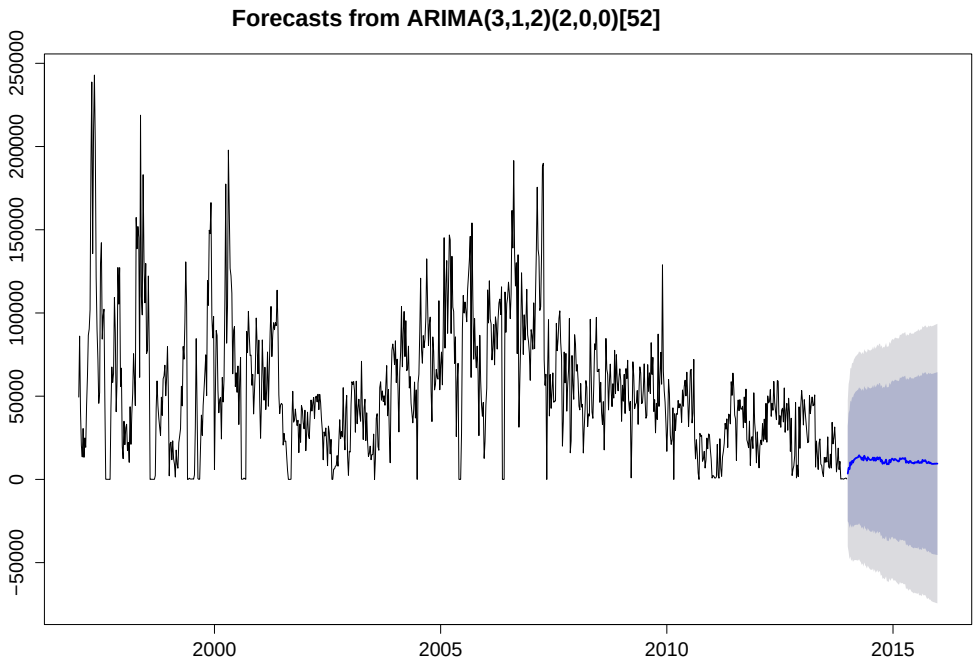


Figura 2.5: Previsione a cinque anni ALICE.

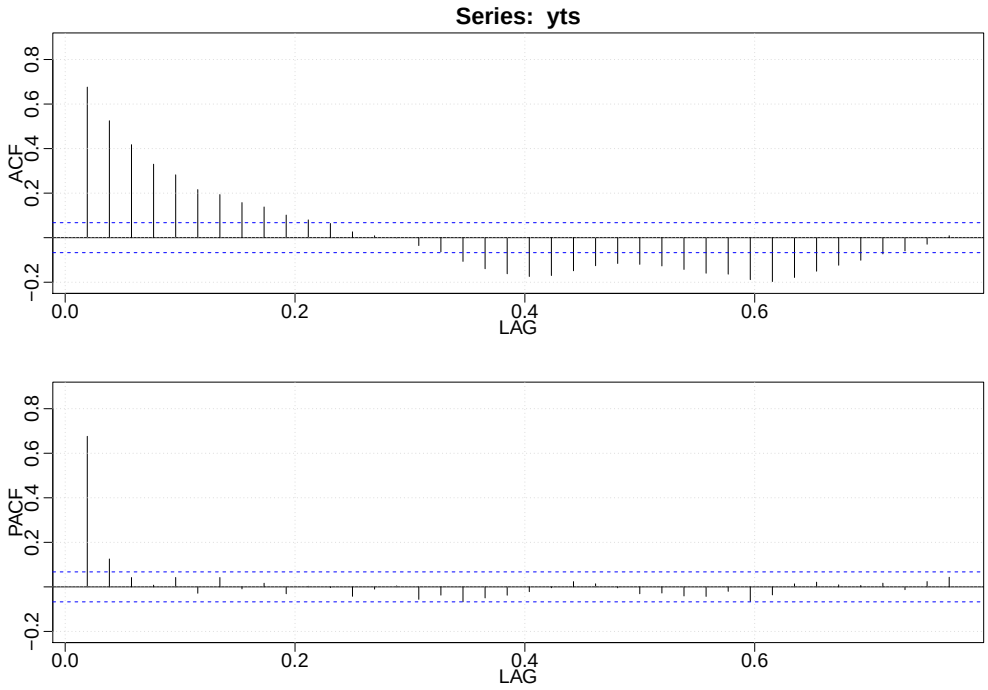


Figura 2.6: Autocorrelazione e autocorrelazione parziale specie LATTERINO.

I grafici in Figura 2.6 rappresentano l'autocorrelazione e l'autocorrelazione parziale per la specie LATTERINO. A differenza dei precedenti in questi grafici le autocorrelazioni decrescono velocemente a zero, per poi leggeremente tornare sopra lo zero a fine dati, possiamo affermare che i dati sono stazionari. Anche in questo caso c'è un leggero andamento oscillatorio. I dati presentano una stagionalità dovuta al raggruppamento settimanale.

Eseguendo l'*auto.arima* risulta:

```
Series: yts
ARIMA(2,0,1)(2,0,0)[52] with non-zero mean

Coefficients:
         ar1      ar2      ma1    sar1    sar2  intercept
      1.1645  -0.2605  -0.6655  0.2231  0.0795  3094.3741
s.e.  0.1153   0.0805   0.1071  0.0308  0.0301   309.3537

sigma^2 estimated as 3355108:  log likelihood=-7895.83
AIC=15806.27   AICc=15806.4   BIC=15839.76
```

Dal confronto tra i grafici ACF, PACF ed il risultato dell'*auto.arima* ci suggeriscono una probabile modifica dei valori dei parametri, da una prima analisi potremmo affermare che il modello più adatto sembrerebbe essere un ARIMA(1,0,0)(1,0,0). Proviamo quindi a ricalcolarci il modello con i nuovi parametri e confrontiamolo con il precedente.

```
Series: yts
ARIMA(2,0,0)(1,0,0)[52] with non-zero mean

Coefficients:
         ar1     ar2    sar1  intercept
      0.5534  0.1243  0.2076  3393.3847
s.e.  0.0340  0.0335  0.0347   259.2249

sigma^2 estimated as 4014556:  log likelihood=-7976.56
AIC=15963.12   AICc=15963.19   BIC=15987.04
```

Questo nuovo modello nel quale è stato tolto un termine a media mobile ed un termine regressivo del primo ordine dalla stagionalità, risulta adeguato, quindi, possiamo usarlo per calcolare le predizioni dei dati con il comando *forecast* come si vede dal grafico in Figura 2.7.

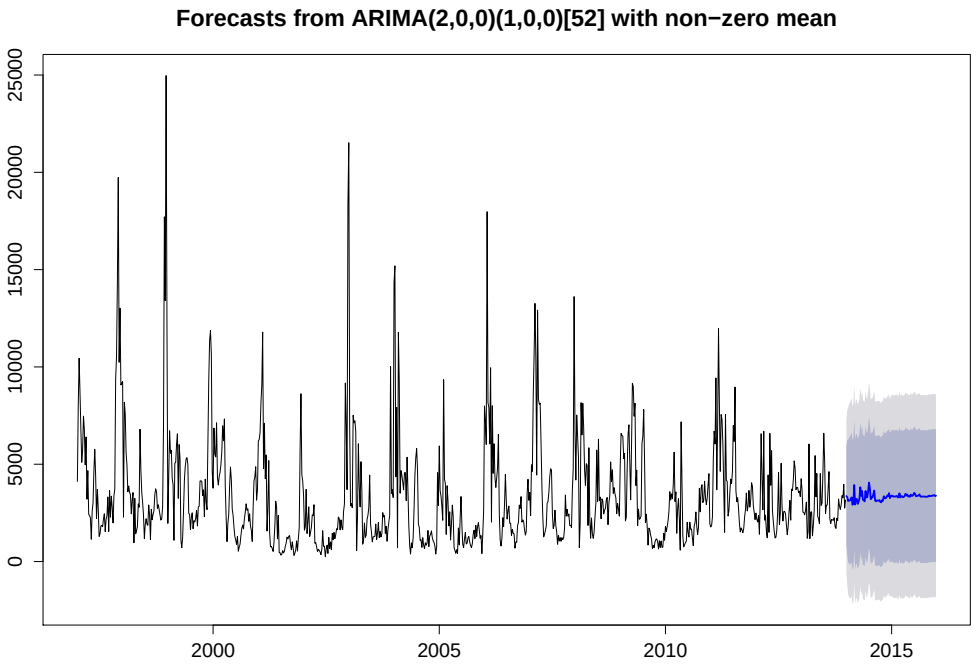


Figura 2.7: Previsione a cinque anni LATTERINO.

In Figura 2.7, la previsione per il LATTERINO nei successivi cinque anni, in blu la linea di previsione, mentre le aree colorote mostrano l'intervallo di predizione all'80% ed al 95% rispettivamente l'area più scura e la più chiara.

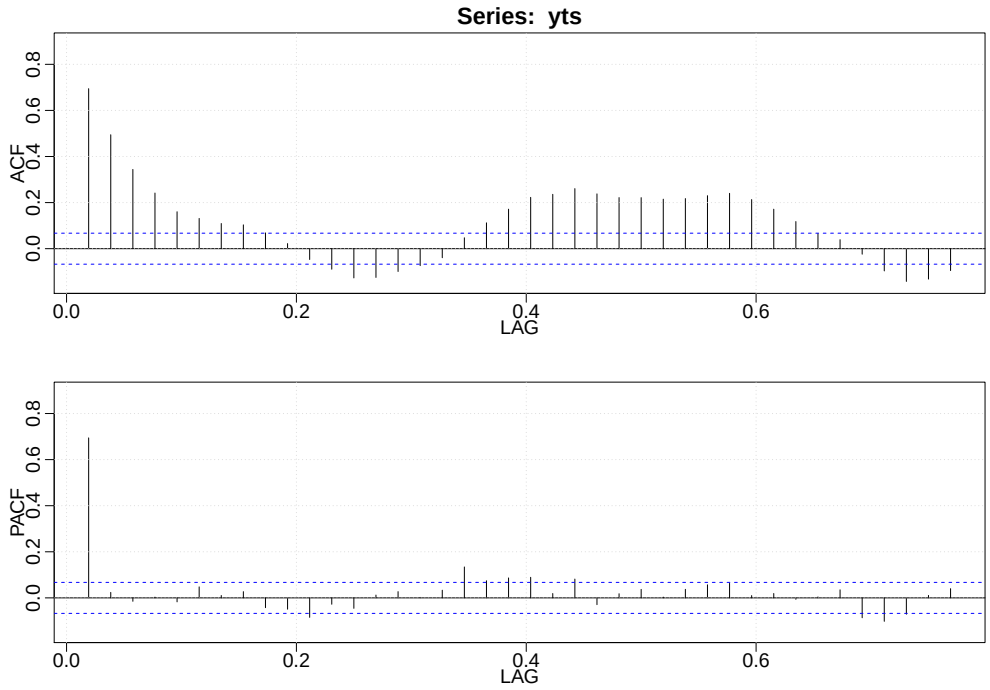


Figura 2.8: Autocorrelazione e autocorrelazione parziale specie PANNOCCHIA.

Ora vedremo lo studio sulle serie storiche per le altre quattro specie prese in esame nell'analisi, il lavoro svolto è lo stesso delle precedenti due, i comandi *R* si trovano in appendice al punto 4. Procediamo quindi come fatto finora, visualizzando i grafici ACF e PACF, dopodichè li confronteremo con il risultato dell'*auto.arima* e, se necessario, si modificheranno i parametri del modello, infine verranno calcolate le previsioni a cinque anni per il modello scelto.

Di seguito l'output dell'*auto.arima* per la specie PANNOCCHIA, che confronteremo con il grafcico in Figura 2.8.

auto.arima per la PANNOCCHIA:

```
Series: yts
ARIMA(2,1,2)(0,0,2)[52] with drift

Coefficients:
         ar1     ar2      ma1      ma2    sma1    sma2   drift
      0.3952  0.1791  -0.7777  -0.2223  0.1577  0.0699  2.6147
s.e.     NaN     NaN      NaN      NaN  0.0349  0.0315  0.7642

sigma^2 estimated as 2978983:  log likelihood=-7837.64
AIC=15691.29   AICc=15691.45   BIC=15729.55
```

Il modello stimato dall'*auto.arima* non è adeguato, andiamo a ristimare i parametri, e "passiamo" ad un modello ARIMA(2,0,0)(0,0,1) da cui:

```
Series: yts
ARIMA(2,0,0)(0,0,1)[52] with non-zero mean

Coefficients:
         ar1     ar2    sma1  intercept
      0.6393  0.0447  0.1573  3114.0255
s.e.  0.0345  0.0339  0.0325   212.2595

sigma^2 estimated as 3031651:  log likelihood=-7851.97
AIC=15713.95   AICc=15714.02   BIC=15737.87
```

Questo modello stimato è adeguato, possiamo usare i modello per fare le previsioni. Si veda in Figura 2.9 le previsioni a cinque anni.

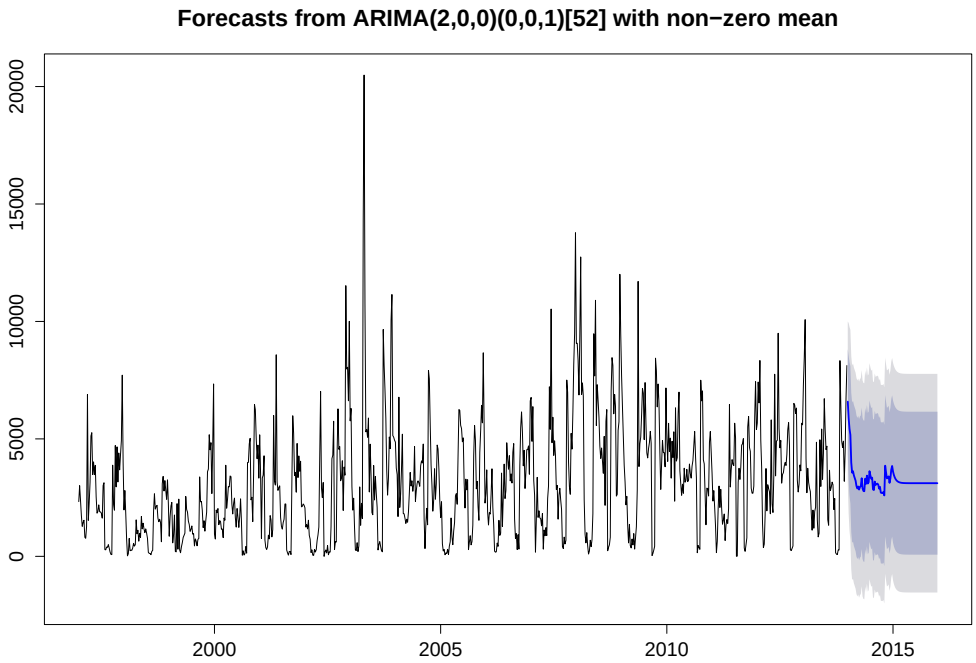


Figura 2.9: Previsione a cinque anni PANNOCCHIA.

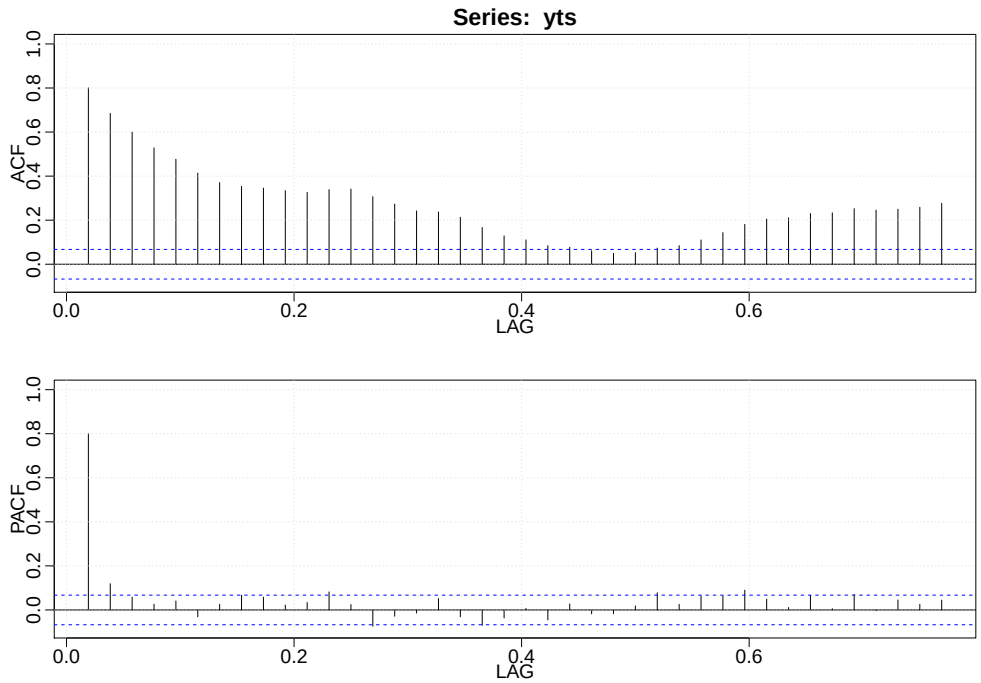


Figura 2.10: Autocorrelazione e autocorrelazione parziale specie SARDINA.

La specie SARDINA, come in precedenza l'ALICE, ha un andamento delle autocorrelazioni decrescente ma molto lento, si veda Figura 2.10, ma che non si azzera mai, questo potrebbe essere sintomo di non-stazionarietà. Analizziamo l'*auto.arima*:

```
Series: yts
ARIMA(2,1,1)(0,0,1)[52]

Coefficients:
         ar1     ar2      ma1   sma1
      0.6685  0.0862  -0.9816  0.067
s.e.  0.0346  0.0344   0.0077  0.034

sigma^2 estimated as 470731226:  log likelihood=-10070.38
AIC=20150.76   AICc=20150.83   BIC=20174.68
```

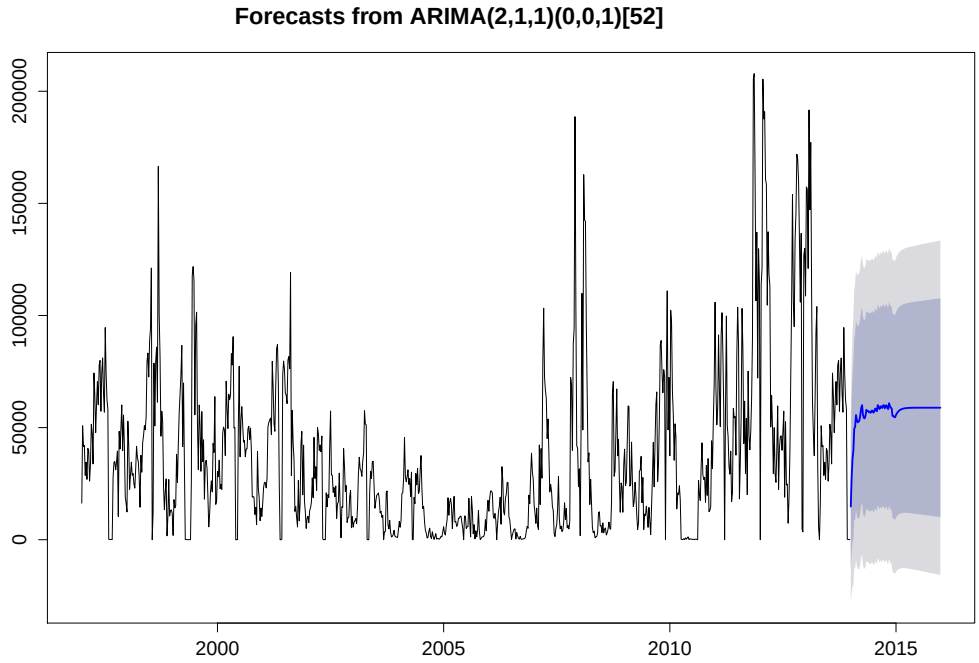


Figura 2.11: Previsione a cinque anni SARDINA.

Dai grafici ACF e PACF e dall'*auto.arima*, il modello sembra adeguato ai nostri dati, quindi possiamo proseguire con la predizione senza dover andare a modificare i parametri. Il grafico in Figura 2.11 mostra la predizione a cinque anni e gli intervalli di predizione all'80% e al 95%.

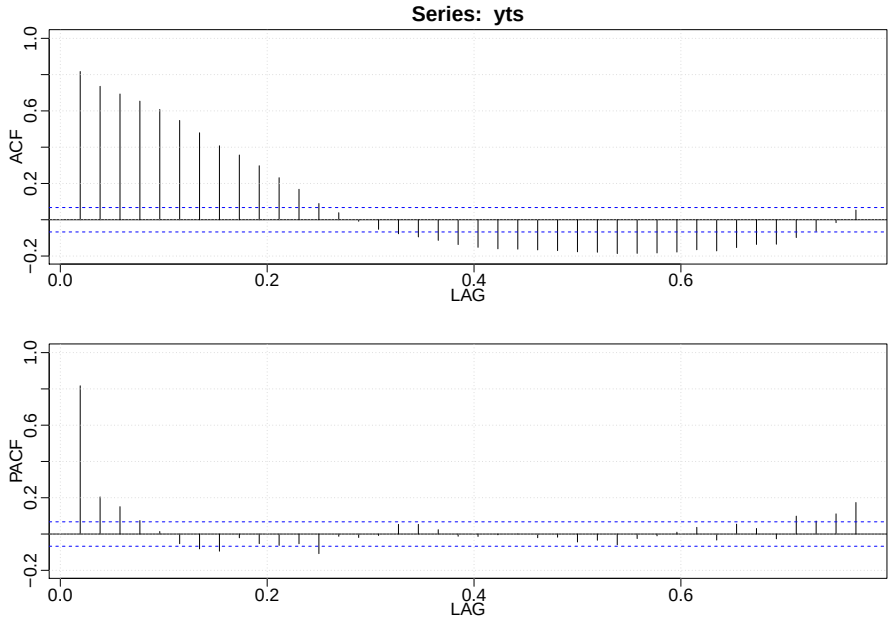


Figura 2.12: Autocorrelazione e autocorrelazione parziale specie SEPPIA.

In Figura 2.12 i grafici ACF e PACF per la SEPPIA ci indicano, come ci aspettiamo, una stagionlità dei dati. L'andamento decrescente che si azzera velocemente indica una stazionarità nei dati. Passiamo quindi all'analisi del modello usando *auto.arima* che ci fornisce il seguente output:

```
Series: yts
ARIMA(4,0,3)(2,0,2)[52] with non-zero mean
Coefficients:
         ar1     ar2      ar3     ar4      ma1      ma2
      1.2398  0.4566  -1.0836  0.3471  -0.7095  -0.7752
s.e.     NaN  0.0000      NaN     NaN      NaN      NaN
         ma3    sar1    sar2     sma1    sma2    intercept
      0.7473  0.1511  0.2728  -0.0194   -0.1787  14256.917
s.e.   0.0030     NaN     NaN   0.0055     NaN    1920.502

sigma^2 estimated as 43629486:  log likelihood=-9031.09
AIC=18088.18   AICc=18088.6   BIC=18150.38
```

Il modello stimato da *auto.arima* non risulta adeguato ai nostri dati, questo è probabilmente dovuto alla numerosità campionaria molto elevata, ma molto influenzata dalla stagionalità della specie. Ristimiamo il nostro modello abbassando i valori del parametro di autoregressione e la media mobile, e togliendo un termine nella media mobile dell'operatore stagionale. Il nuovo modello sarà dunque un ARIMA(2,0,1)(1,0,1) da cui:

```
Series: yts
ARIMA(2,0,1)(1,0,1)[52] with non-zero mean

Coefficients:
         ar1      ar2      ma1    sar1     sma1  intercept
      1.1559  -0.2031  -0.6500  0.8964  -0.7490  14880.115
s.e.  0.0511   0.0463   0.0418     NaN   0.0174   3066.081

sigma^2 estimated as 43645525:  log likelihood=-9036.19
AIC=18086.38   AICc=18086.5   BIC=18119.87
```

Nel grafico di Figura 2.13, le previsioni a cinque anni, con il comando *forecast*, per la specie SEPPIA.

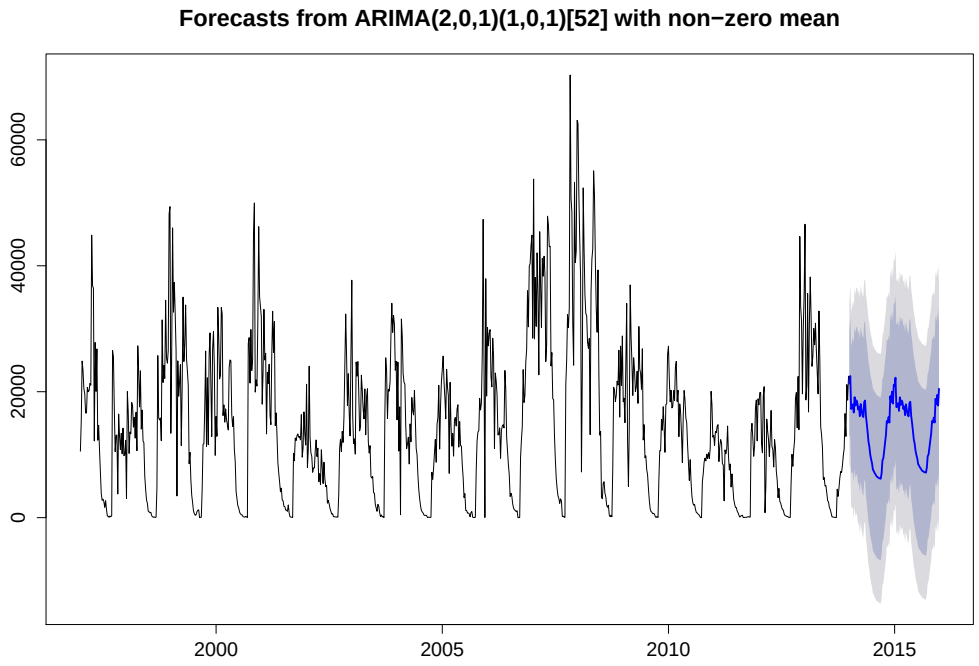


Figura 2.13: Previsione a cinque anni SEPPIA.

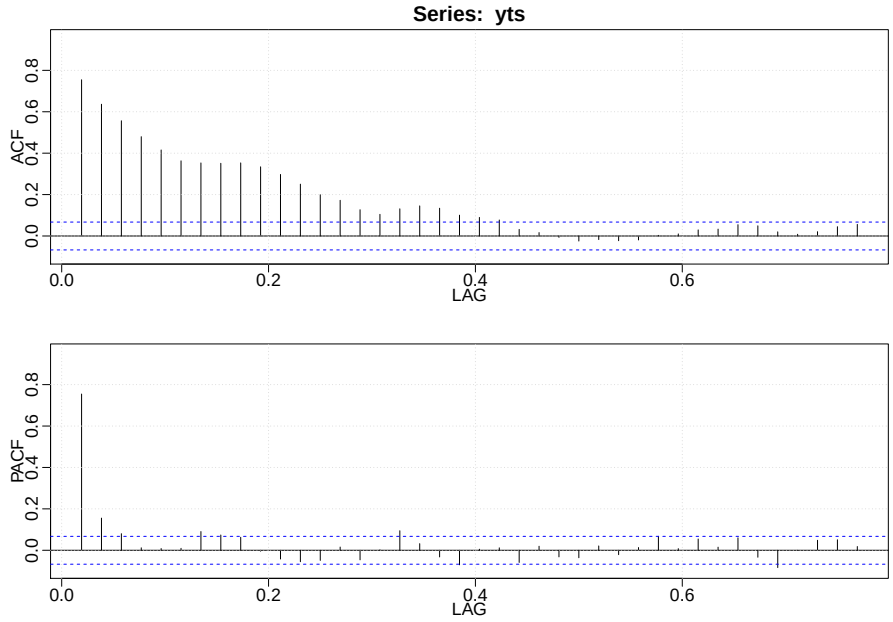


Figura 2.14: Autocorrelazione e autocorrelazione parziale specie SOGLIOLA.

Dalla Figura 2.14, i grafico ACF e PACF mostrano l'andamento delle autocorrelazioni per la specie SOGLIOLA decresce a zero per poi risalire, mantenendo una certa oscillazione, da questo possiamo dedurre una stazionarietà nei dati che ci porta a proseguire l'analisi andando a confrontare i grafici con i risulati dell'*auto.arima*.

```
Series: yts
ARIMA(1,0,1)(0,0,2)[52] with non-zero mean

Coefficients:
         ar1      ma1    sma1    sma2  intercept
      0.8620  -0.3369  0.2344  0.1304  6107.7813
s.e.  0.0242   0.0485  0.0358  0.0336   451.1815

sigma^2 estimated as 4414358:  log likelihood=-8019.54
AIC=16051.08   AICc=16051.17   BIC=16079.78
```

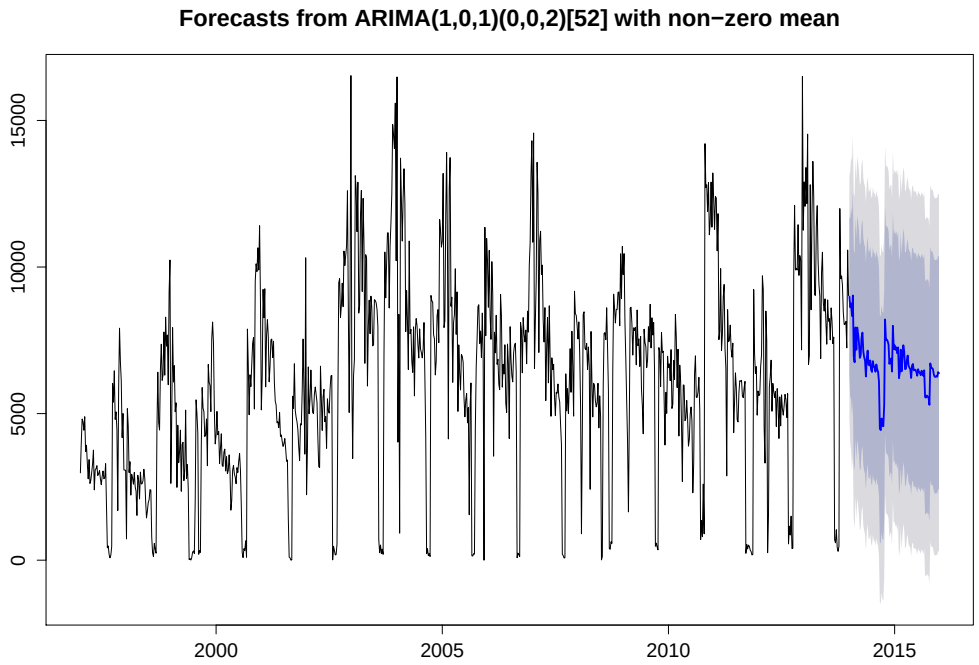


Figura 2.15: Previsione a cinque anni SOGLIOLA.

I risultati dell'*auto.arima* mostrano dei parametri adeguati al modello, quindi non occorrono modifche e si può procedere all'uso statistico per il calcolo delle previsioni. Il grafico in Figura 2.15 mostra le previsioni a cinque anni per la specie SOGLIOLA, dove la linea blu è la predizione e le due aree invece, sono le predizioni all'80% l'area più scura ed al 95% quella più chiara.

Tutte le funzioni finora utilizzate si trovano in appendice al punto 4.

Capitolo 3

Analisi delle Correlazioni Canoniche

L'obbiettivo dell'analisi delle correlazioni canoniche (ACC), è quello di identificare la relazione lineare esistente tra due insiemi di variabili quantitative. Nell'ambito ecologico è spesso necessario prendere in considerazione insiemi di variabili qualitativamente eterogenei.

Ad esempio, è frequente disporre di una lista di specie e di misure di parametri fisico-chimici relative ad un insieme di osservazioni distribuite nello spazio e/o nel tempo, nel nostro studio infatti disponiamo di una lista di specie con relativo peso pescato settimanalmente distribuito negli anni. Un insieme di dati organizzato in questo modo non può essere analizzato esaustivamente mediante le consuete tecniche di ordinamento, le quali non consentono di isolare i due sottoinsiemi di variabili e di valutarne il grado di correlazione globale.

Lo scopo è trovare una combinazione lineare della variabili del primo gruppo e una combinazione lineare delle variabili del secondo gruppo che siano le più correlata possibile.

L'analisi delle correlazioni canoniche ha come fine prorpio l'esame di tali correlazioni.

Data una matrice di dati tipo:

$$Z_{nx(p+q)} = \begin{pmatrix} y_{11} & \dots & y_{1p} & x_{11} & \dots & x_{1q} \\ y_{21} & \dots & y_{2p} & x_{21} & \dots & x_{2q} \\ \vdots & & & \vdots & & \\ y_{n1} & \dots & y_{np} & x_{n1} & \dots & x_{nq} \end{pmatrix}$$

dove ci sono x_p variabili esplicative, predittrici, e y_q variabili di risposta, predette, e S_{11}, S_{22} e S_{12} vengono stimate sulla base dei dati osservati. Dove:

$$S_X = \frac{1}{n} X'HX$$
$$S_Y = \frac{1}{n} Y'HY$$
$$S_{XY} = \frac{1}{n} X'HY$$

Questa matrice è detta *MATRICE CAMPIONARIA DEI DATI.*

Si definiscano due generiche combinazioni lineari di **x** e **y**:

$$\eta = \mathbf{a'x}, \qquad \varphi = \mathbf{b'y}$$

L'obbiettivo è cercare i vettori **a** e **b** per i quali sia massima la correlazione tra η e φ. Esistono però inifiniti vettori $\mathbf{a^*}=k_1\mathbf{a}$ e $\mathbf{b^*}=k_2\mathbf{b}$ per i quali le combinazioni $\eta* = \mathbf{a'^*x}$ e $\varphi* = \mathbf{b'^*y}$ hanno lo stesso coefficiente di correlazione di η e φ. D'altra parte il coefficiente di correlazione è invariante per cambiamenti di scala. Per rendere possibile la soluzione di tale problema si può, senza perdita di generalità, cercare le variabili canoniche **a'x** e **b'y** tra quelle di varianza fissata, in particolare unitaria.

Si procede dunque a massimizzare il **primo coefficiente di correlazione canonica** e ricavare la prima copoia di variabili canoniche :

$$\rho_{\eta,\varphi}(\mathbf{a,b}) = \frac{\mathbf{a'}S_1 2\mathbf{b}}{[\mathbf{a'}S_1 1\mathbf{ab'}S_2 2\mathbf{b}]^{1/2}}$$

Il problema di massimo vincolato da risolvere è:

$$\underset{a,\,b}{max} \qquad \mathbf{a'}S_{12}\mathbf{b}$$

$$\mathbf{a'}S_{11}\mathbf{a} = 1$$
$$\mathbf{b'}S_{22}\mathbf{b} = 1$$

La soluzione al problema del massimo vincolato, è detta *prima correlazione canonica*, è il primo valore singolare d_1 della decomposizione in valori singolari (SVD), della matrice $G_{mxn} = S_{11}^{-1/2} S_{12} S_{22}^{-1/2} = UDV'$ e i primi vettori delle matrici U e V definiscono le combinazioni lineari che massimizzano questa correlazione. Le successive correlazioni canoniche sono identificate dai successivi valori (e vettori) singolari.

L'ACC ci permette di costruire $r{=}r(S_{12})$ nuove coppie di variabili canoniche, combinazioni lineari di quelle originali, **massimamente correlate fra loro, incorrelate entro ogni gruppo** e con tutte le altre variabili nell'altro gruppo eccetto quella con cui sono accoppiate.

Vediamo di seguito l'analisi delle correlazioni canoniche per le sei specie in esame. Il primo passo è quello di creare una matrice con i dati che serviranno all'analisi, metteremo in colonna le settimane di pesca, il peso pescato settimanalmente, diviso per le sei specie in esame, e aggiungeremo un variabile, la stagione[1], che ci permetterà di studiare le correlazioni tra peso pescato per specie e stagione atmosferica. Di seguito la nuova matrice su cui lavoreremo, i codici per realizzarla si trovano in appendice al punto 5.

stag	anno	sett	Alice	Latterino	Pannocchia	Sardine	Seppia	Sogliola
1	1997	1	49462.0	4118.0	2329.5	16282.0	10548.5	2977.5
1	1997	2	86093.0	7988.8	3010.0	50813.0	15312.8	3978.2
1	1997	3	41251.0	10442.0	2388.7	41496.0	24865.3	4816.5
1	1997	4	20419.0	8717.5	1741.0	42021.0	23417.3	4734.9
1	1997	5	13517.0	6829.4	1275.6	28679.0	20504.0	4433.5
1	1997	6	30555.0	5096.3	1446.1	34510.0	19709.3	4893.9
1	1997	7	13293.0	5725.3	1537.2	26936.0	16624.5	3704.5
1	1997	8	24906.0	7472.8	810.9	40733.0	16618.0	3924.1
1	1997	9	19299.0	6786.0	769.3	39536.0	20687.5	3299.3
1	1997	10	44611.0	4968.5	1272.1	26139.8	19996.3	2773.3
2	1997	11	61981.5	6401.0	6884.8	33827.5	20167.9	3431.0
2	1997	12	86625.0	3203.3	1532.6	51504.3	21321.8	2616.3
2	1997	13	90384.0	4669.3	2286.3	47369.0	21030.4	2721.7
2	1997	14	102004.0	2430.3	3157.4	33796.0	44866.3	2993.3
2	1997	15	189014.0	2358.8	5123.6	74333.0	36827.3	3271.6
2	1997	16	238766.5	2242.0	5265.0	64456.0	36427.5	3755.5
2	1997	17	135667.0	1145.8	3480.0	47796.0	12177.0	2397.6

Nella prima colonna compaiono le stagioni, da 1 a 4, da inverno ad autunno, nella seconda gli anni di riferimento, nella terza ci sono le settimane e nelle restanti sei i pesi per le sei specie. Creata la matrice dei dati, andremo a creare le due sottomatrici X e Y mediante le quali si andranno a studiare

[1]Si è cercato il più possibile di seguire la stagionalità reale, facendo concludere l'inverno alla terza settimana di Marzo, per primavera, estate ed autunno le succesive trentanove settimane dell'anno e per l'inverno le ultime due settimane di Dicembre.

le correlazioni canoniche. Nella prima matrice X saranno presenti i dati sul peso pescato, vale a dire le ultime sei colonne della matrice, nella matrice Y invece ci saranno i dati temporali, le prime due colonne, le stagioni e gli anni di riferimento. Prima di creare le due matrici X e Y, dovremmo utilizzare alcune funzioni, visibili in appendice, che ci permettano di trasformare le nostre variabili categoriali (le stagioni) in variabili dummy, le quali miglioreranno l'adattamento della regressione. Successivamente verranno create X e Y con le quali si potrà eseguire l'analisi delle correlazioni con i seguenti risultati.

Il grafico in Figura 3.1 mostra le correlazione canoniche delle matrici X (in rosso) e Y (in nero), sullo sfondo in grigio chiaro, invece, sono riportati i punteggi delle nuove combinazioni canoniche della matrice X, cioè gli anni di pescato e la settimana di riferimento nel tempo (esempio 1997.1 è la prima settimana del 1997).

Diamo una prima interpretazione dei dati, guardando il grafico possiamo possiamo vedere come le specie LATTERINO e SEPPIA siano correlate tra loro e con la stagione *inverno*, ed incorrelate con l'*estate*, la quale si trova all'opposto, cioè possiamo dire che queste due specie si pescano maggiormente in inverno rispetto che in estate. Un'altra affermazione che possiamo fare è la notevole incorrelazione tra ALICE, e SOGLIOLA e PANNOCCHIA, il pescare tanto o poco di una non implica pescare tanto o poco delle altre due, e viceversa.

Sempre per l'ALICE si può anche affermare che la quantità di pescato non è correlata al procedere dell'*anno*, cosa che invece influenza SARDINE e PANNOCCHIE, questo vuol dire che, mediamente, il pescato di SARDINE e PANNOCCHIE e regolare nell'anno, mentre quello delle ALICE subisce la stagionalità, infatti è maggiore nella *primavera*, come si vede dal grafico. Un'ultima osservazione possiamo farla a proposito della specie SEPPIA la quale è incorrelata con l'*estate*, possiamo infatti notare che i valori temporali attorno alla specie, si noti 2008.1, 2007.52, 2007.44 e 2003.1, siano tutti all'interno dell'*inverno* che, per appunto, è opposta all'*estate* sia graficamente che nella realtà.

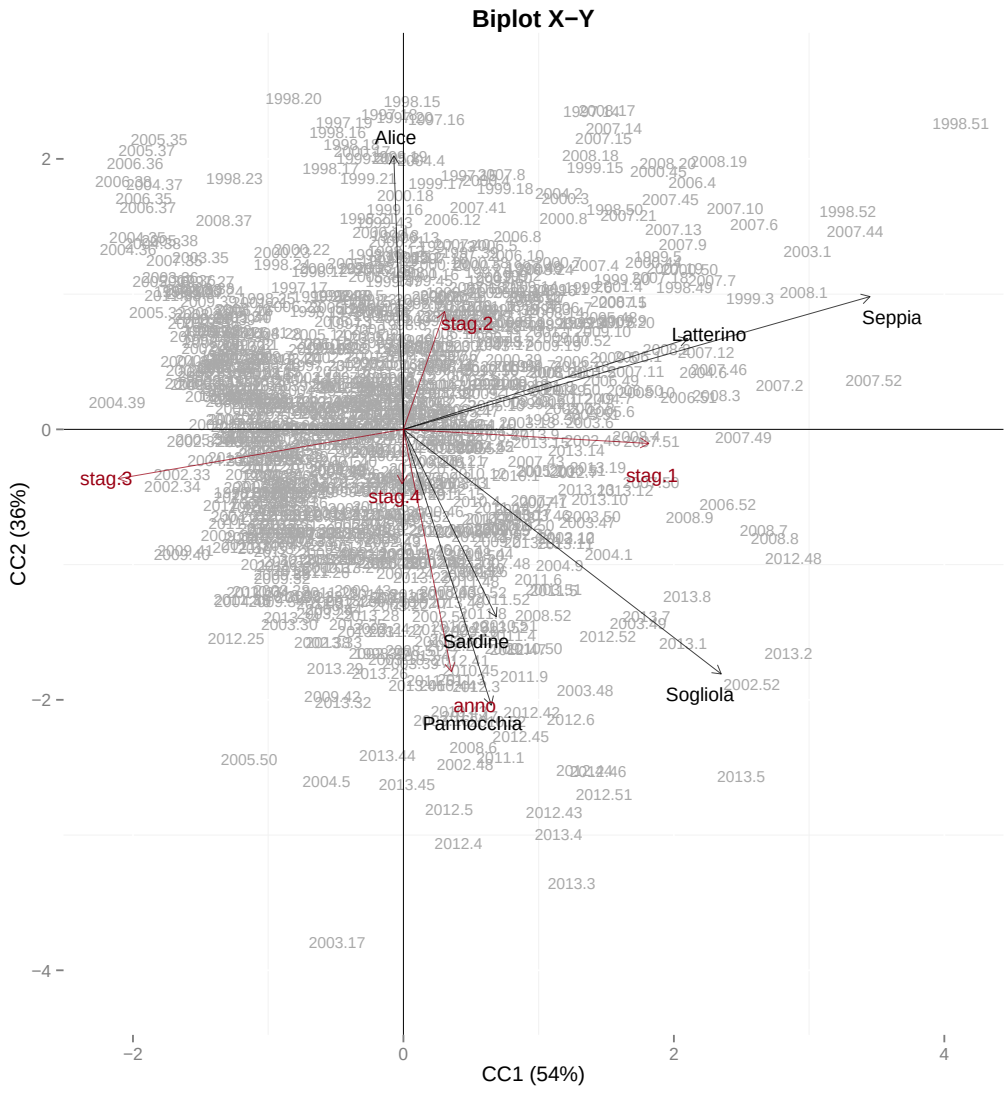


Figura 3.1: Prima e seconda correlazione canonica.

In modo similare al precedente grafico possiamo vedere la seconda e la terza correlazione canonica in Figura 3.2, dove si nota la forte correlazione tra SEPPIE, ALICE e la *primavera*, e la correlazione, come visto precedentemente, tra *anno*, SPIGOLE, PANNOCCHIE e SOGLIOLA. Risulta inoltre evidente l'incorrelazione tra LATTERINO e *autunno*.

Possiamo inoltre analizzare i vari trend presenti, come si evince dalla Figura 3.1 e dalla Figura 3.2 il trend annuo sembra legato alla seconda componente, le osservazioni sono all'incirca in ordine crescente. Invece le variabili stagionali sono più legate alla prima componente canonica. Per quanto rigurda le specie, come accennato in precedenza, dal grafico in Figura 3.2 si vede come la specie LATTERINO sia più legata al trend stagionale rispetto che a quello annuo. Viceversa la specie ALICE è legata più ad un trend annuale che stagionale.

Questi ed altre analisi si possono fare mediante lo studio delle correlazioni canoniche. Si possono cambiare le variabili, avendo a disposizione le temperature si potrebbe effettuare uno studio più approfondito sul pescato correlato al meteo, oppure studiare se esiste una correlazione tra zone specifiche di pesca e le varie specie, come ad esempio, uno studio preda-predatore.

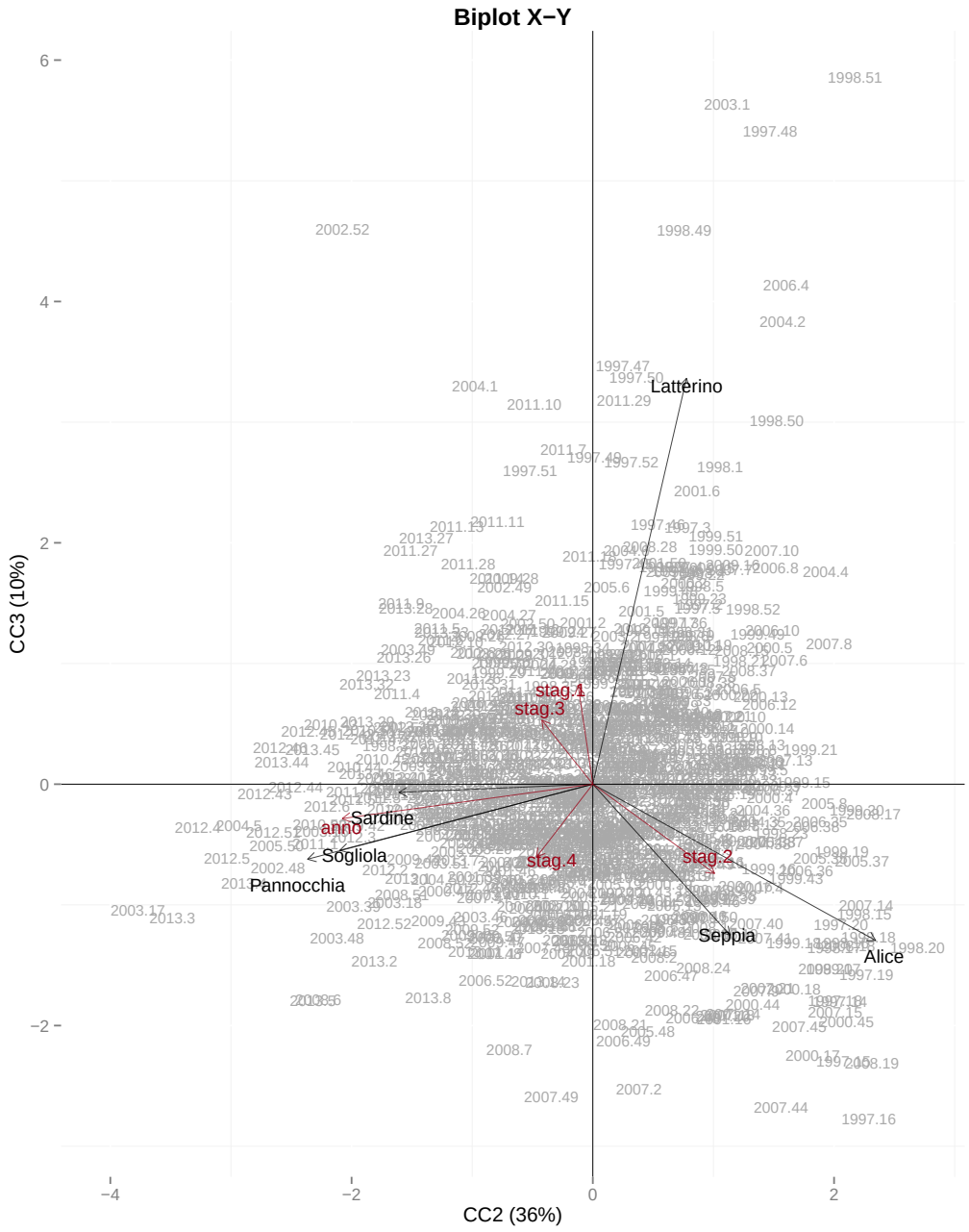


Figura 3.2: Seconda e terza correlazione canonica.

Conclusioni

In questa tesi si è cercato di creare un manuale per lo studio della banca dati del pescato di Chioggia, l'obbiettivo era quello di permettere, anche le persone meno avvezze all'uso di un software statistico come R, di poter senza troppa fatica compiere un'analisi con questi dati.

Il primo lavoro effettuato è stato quello di pulizia dei dati, parte fondamentale e molto delicata del procedimento, in quanto, essendoci molti dati su cui lavorare e numerose variabili, si è cercato di togliere quelle meno utili e che sarebbero state solo di peso per dataset e cercando di mantenere quelle che realmente servivano per un'analisi. Questo passaggio, causa le differenti impostazioni dei dataset annuali risulta quindi molto importante, poichè se la pulizia e la gestione dei dati non vengono effettuate con criterio questo comprometterà l'intera analisi.

L'analisi delle serie storiche, tramite l'analisi grafica ed analitica, ha pemesso uno studio dell'andamento del pescato negli anni, si è potuto verificare le variazioni subite nel tempo e calcolare una predizione futura, a cinque anni, del suo andamento. Questa analisi risoluta molto importante, sia ai fini dell'interpretazione di un fenomeno, cercando di chiarire il meccanismo che l'ha generata, sia per prevedernene il suo andamento futuro, di cui è nota la storia passata.

Nell'ultima analisi si sono studiate le correlazioni canoniche, cioè lo studio delle correlazioni che possono intercorrere tra le variabili in esame. Questo studio può essere di portata molto elevata, in base alle variabili a disposizione, basti pensare a tutte le correlazioni che possono esserci nel mare, dalla temperatura alla marea, dal vento alla pressione. Nel nostro studio abbiamo voluto studiare le correlazioni tra le varie specie, aggiungendo alla

nostra analisi la variabile stagionale e guardando come questa si comportava al passare del tempo.

Il manuale quindi è pronto per essere utilizzato e perfezionato con l'aggiunta, fin da subito, dei nuovi dataset annuali così da creare un dataset finale aggiornato annualmente e analizzabile con le funzioni utilizzate.

Appendice A

Codici R

1. **Creazione del dataset finale**

   -Lettura dei file:

```
h=read.csv("ANNO 2010.csv")
head(h)
```

   -Data Cleaning, per anni da 1997 a 2006:

```
h=subset(h,select=-c(data.vendita,pr,cd.prod,mp))
head(h)

h$specie=as.character(h$specie)
h$specie[h$specie=="ARINGA"]="ALICE"
h$specie=factor(h$specie)

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)
```

per l'anno 2007:

```
h=subset(h,select=c(data.f.a.,mese,anno,cod.sp,specie,quantità.kg,
prezzo.kg,valore,pr,venditore))

h$specie=as.character(h$specie)
h$specie[h$specie=="ALICE (c3kg)"]="ALICE"
h$specie[h$specie=="ALICE (c4kg)"]="ALICE"
h$specie[h$specie=="ALICE (c6kg)"]="ALICE"
h$specie[h$specie=="ARINGA"]="ALICE"
h$specie=factor(h$specie)

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)
```

per l'anno 2008:

```
h=subset(h,select=c(data.f.a.,mese,anno,cod.sp,specie,quantità.kg,
prezzo.kg,valore,pr,venditore))

h$specie=as.character(h$specie)
h$specie[h$specie=="ALICE (c4kg)"]="ALICE"
h$specie[h$specie=="ALICE (c6kg)"]="ALICE"
h$specie[h$specie=="ARINGA"]="ALICE"
h$specie[h$specie=="SARDINA (c4kg)"]="SARDINA"
h$specie=factor(h$specie)

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)
```

per l'anno 2009:

```
h=subset(h,select=c(data.f.a.,mese,anno,cod.sp,specie,quantità.kg,
prezzo.kg,valore,pr,venditore))

h$specie=as.character(h$specie)
h$specie[h$specie=="ALICE (c4kg)"]="ALICE"
h$specie[h$specie=="ALICE (c6kg)"]="ALICE"
h$specie[h$specie=="ARINGA"]="ALICE"
h$specie[h$specie=="SARDINA (c4kg)"]="SARDINA"
h$specie[h$specie=="SARDINA (c6kg)"]="SARDINA"
h$specie[h$specie=="SEPPIOLAA"]="SEPPIOLA"
h$specie=factor(h$specie)

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)
```

per l'anno 2010:

```
h=subset(h, tipo.prodotto=="Pescato locale")
h=subset(h,select=c(data.f.a.,mese,anno,cod.sp,specie,quantità.kg,
prezzo.kg,valore,pr,venditore))
head(h)

h$specie=as.character(h$specie)
h$specie[h$specie=="ALICE (c3kg)"]="ALICE"
h$specie[h$specie=="ALICE (c4kg)"]="ALICE"
h$specie[h$specie=="ALICE (c6kg)"]="ALICE"
h$specie[h$specie=="ARINGA"]="ALICE"
h$specie[h$specie=="BISO O TOMBARELLO"]="TONNO TOMBARELLO"
h$specie[h$specie=="SARDINA (c4Kg)"]="SARDINA"
h$specie[h$specie=="SARDINA (c6Kg)"]="SARDINA"
h$specie[h$specie=="SEPPIOLA (Zotolo)"]="ZOTOLO"
```

```
h$specie=factor(h$specie)

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)
```

per l'anno 2011:

```
h=subset(h, DocRgP1Desc=="01-Locale")

h=subset(h, select=c(DocDtHH,DocMese,DocAnno,ArtIDStat,
ArtDescrComStat2,Qta,DocRgPre,DocRgImpN,
DocRgP2Desc,VendPescVal))

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)

h$specie=as.character(h$specie)
h$specie[h$specie=="ALICE O ACCIUGA"]="ALICE"
h$specie[h$specie=="AQUILA DI MARE"]="AQUILA DI MARE/Razza"
h$specie[h$specie=="BISO O TOMBARELLO"]="TONNO TOMBARELLO"
h$specie[h$specie=="CANNOLICCHIO O CAPPALUNGA"]="CANNOLICCHIO"
h$specie[h$specie=="CAPPASANTA O CONCHIGLIA DI S. GIACOMO"]="CAPPASANTA
h$specie[h$specie=="CAPPELLANO O BUSBANA"]="BUSBANA"
h$specie[h$specie=="CEFALO DORATO O LOTREGANO"]="CEFALO LOTREGANO"
h$specie[h$specie=="CEFALO O BOSEGA"]="CEFALO BOSEGA"
h$specie[h$specie=="CEFALO O CALAMITA O BOTOLO"]="CEFALO CALAMITA"
h$specie[h$specie=="CEFALO O VERZELATA"]="CEFALO VERZELATA"
h$specie[h$specie=="CEFALO O VOLPINA"]="CEFALO VOLPINA"
h$specie[h$specie=="COREGONE O LAVARELLO"]="LAVARELLO"
h$specie[h$specie=="CORIFENA O LAMPUGA"]="LAMPUGA"
h$specie[h$specie=="GALLINELLA O CAPPONE"]="GALLINELLA"
h$specie[h$specie=="GHIOZZETTO"]="GHIOZZETTO MINUTO"
```

```
h$specie[h$specie=="GHIOZZO GÒ"]="GHIOZZO"
h$specie[h$specie=="GRANCHIO DA MOLECA"]="GRANCHIO (Moleca)"
h$specie[h$specie=="GRANCHIO DA MOLECA (MUTATO/MOLECA)"]
="GRANCHIO (Moleca)"
h$specie[h$specie=="GRANCHIO DA MOLECA (FEMM/MAZANETA)"]
="GRANCHIO (Mazaneta)"
h$specie[h$specie=="GRANSEOLA O GRANCEOLA"]="GRANSEOLA"
h$specie[h$specie=="LANZARDO O SGOMBRO OCCHIONE"]="LANZARDO"
h$specie[h$specie=="LATTERINO O ACQUADELLA"]="LATTERINO"
h$specie[h$specie=="LUCCIOPERCA O SANDRA"]="LUCCIOPERCA"
h$specie[h$specie=="MERLANO O MOLO"]="MOLO"
h$specie[h$specie=="MISTO"]="MISTO PESCE"
h$specie[h$specie=="MOSCARDINO BIANCO (T/EXTRA PICCOLA)"]
="MOSCARDINO BIANCO (tag/XP)"
h$specie[h$specie=="NASELLO O MERLUZZO"]="MERLUZZO"
h$specie[h$specie=="OSTRICA O OSTRICA PIATTA"]="OSTRICA"
h$specie[h$specie=="PANNOCCHIA O CANOCCHIA"]="PANNOCCHIA"
h$specie[h$specie=="PAPALINA O SPRATTO"]="PAPALINA"
h$specie[h$specie=="PESCE S. PIETRO"]="PESCE S.PIETRO"
h$specie[h$specie=="POLPESSA O POLPO MACCHIATO"]="POLPESSA"
h$specie[h$specie=="ROMBO LISCIO O SOASO"]="SOASO"
h$specie[h$specie=="ROSPO O RANA PESCATRICE"]="RANA PESCATRICE"
h$specie[h$specie=="ROSPO O RANA PESCATRICE  (CODA)"]
="RANA PESCATRICE (Coda)"
h$specie[h$specie=="SEPPIA (T/PICCOLA)"]="SEPPIA (tag/P)"
h$specie[h$specie=="SEPPIETTA (SEPPI)"]="SEPPIETTA"
h$specie[h$specie=="SEPPIOLA (ZOTOLO)"]="ZOTOLO"
h$specie[h$specie=="SPIGOLA O BRANZINO"]="BRANZINO"
h$specie[h$specie=="SURO O SUGARELLO"]="SURO"
h$specie[h$specie=="TONNO O TONNO ROSSO"]="TONNO"
h$specie[h$specie=="ZANCHETTA O SUACIA"]="ZANCHETTA"
h$specie=factor(h$specie)
```

per l'anno 2012:

```
h=subset(h, DocRgP1Desc=="01-Locale")

h=subset(h, select=c(DocDtHH,DocMese,DocAnno,ArtIDStat,
ArtDescrComStat2,Qta,DocRgPre,DocRgImpN,
DocRgP2Desc,VendPescVal))

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)

h$specie=as.character(h$specie)
h$specie[h$specie=="ALICE O ACCIUGA "]="ALICE"
h$specie[h$specie=="ALOSA O CHEPPIA"]="CHEPPIA"
h$specie[h$specie=="AQUILA DI MARE"]="AQUILA DI MARE/Razza"
h$specie[h$specie=="BISO O TOMBARELLO"]="TONNO TOMBARELLO"
h$specie[h$specie=="CANNOLICCHIO O CAPPALUNGA"]="CANNOLICCHIO"
h$specie[h$specie=="CAPPASANTA O CONCHIGLIA DI S. GIACOMO"]
="CAPPASANTA C.G."
h$specie[h$specie=="CAPPELLANO O BUSBANA"]="BUSBANA"
h$specie[h$specie=="CEFALO CALAMITA O BOTOLO"]="CEFALO CALAMITA"
h$specie[h$specie=="CEFALO DORATO O LOTREGANO"]="CEFALO LOTREGANO"
h$specie[h$specie=="CEFALO O BOSEGA"]="CEFALO BOSEGA"
h$specie[h$specie=="CEFALO O VOLPINA"]="CEFALO VOLPINA"
h$specie[h$specie=="GALLINELLA O CAPPONE"]="GALLINELLA"
h$specie[h$specie=="GHIOZZETTO QUADRIMACULATO (SCAGIOTO)"]
="GHIOZZETTO MINUTO"
h$specie[h$specie=="GHIOZZO GÒ"]="GHIOZZO"
h$specie[h$specie=="GRANCHIO DA MOLECA  "]="GRANCHIO (Moleca)"
h$specie[h$specie=="GRANCHIO DA MOLECA (FEMM/MAZANETA) "]
="GRANCHIO (Mazaneta)"
h$specie[h$specie=="GRANCHIO DA MOLECA (MUTATO/MOLECA)"]
```

```
="GRANCHIO (Moleca)"
h$specie[h$specie=="GRANSEOLA O GRANCEOLA"]="GRANSEOLA"
h$specie[h$specie=="LANZARDO O SGOMBRO OCCHIONE"]="LANZARDO"
h$specie[h$specie=="LATTERINO O ACQUADELLA"]="LATTERINO"
h$specie[h$specie=="LUCCIOPERCA O SANDRA"]="LUCCIOPERCA"
h$specie[h$specie=="MERLANO O MOLO"]="MOLO"
h$specie[h$specie=="MISTO"]="MISTO PESCE"
h$specie[h$specie=="MOSCARDINO BIANCO (T/EXTRA PICCOLA)"]
="MOSCARDINO BIANCO (tag/XP)"
h$specie[h$specie=="NASELLO O MERLUZZO"]="MERLUZZO"
h$specie[h$specie=="OSTRICA CONCAVA"]="OSTRICA"
h$specie[h$specie=="OSTRICA PIATTA"]="OSTRICA"
h$specie[h$specie=="PANNOCCHIA O CANOCCHIA"]="PANNOCCHIA"
h$specie[h$specie=="PAPALINA O SPRATTO"]="PAPALINA"
h$specie[h$specie=="PESCE PRETE O LUCERNA"]="PESCE PRETE"
h$specie[h$specie=="PESCE S. PIETRO"]="PESCE S.PIETRO"
h$specie[h$specie=="POLPESSA O POLPO MACCHIATO"]="POLPESSA"
h$specie[h$specie=="ROMBO LISCIO O SOASO"]="SOASO"
h$specie[h$specie=="ROSPO O RANA PESCATRICE"]="RANA PESCATRICE"
h$specie[h$specie=="SARDINA (BIANCHETTO - NOVELLAME)"]="SARDINA"
h$specie[h$specie=="SEPPIA (T/PICCOLA)"]="SEPPIA (tag/P)"
h$specie[h$specie=="SEPPIETTA (SEPO)"]="SEPPIETTA"
h$specie[h$specie=="SEPPIOLA (ZOTOLO)"]="ZOTOLO"
h$specie[h$specie=="SMERIGLIO O MAKO"]="SMERIGLIO"
h$specie[h$specie=="SPIGOLA O BRANZINO"]="BRANZINO"
h$specie[h$specie=="SURO O SUGARELLO"]="SURO"
h$specie[h$specie=="TONNO ROSSO"]="TONNO"
h$specie[h$specie=="VONGOLA O LUPINO"]="VONGOLA"
h$specie[h$specie=="ZANCHETTA O SUACIA"]="ZANCHETTA"
h$specie=factor(h$specie)

head(h)
```

per l'anno 2013:

```
h=subset(h, DocRgP1Desc=="01-Locale")

h=subset(h, select=c(DocDtHH,DocMese,DocAnno,ArtIDStat,
ArtDescrComStat2,Qta,DocRgPre,DocRgImpN,
DocRgP2Desc,VendPescVal))

dimnames(h)[[2]]=list("data","mese","anno","cod.spe","specie",
"peso.kg","prezzo","valore","tpr","produttore")
head(h)

h$specie=as.character(h$specie)
h$specie[h$specie=="ALICE O ACCIUGA "]="ALICE"
h$specie[h$specie=="ALOSA O CHEPPIA"]="CHEPPIA"
h$specie[h$specie=="AQUILA DI MARE"]="AQUILA DI MARE/Razza"
h$specie[h$specie=="BISO O TOMBARELLO"]="TONNO TOMBARELLO"
h$specie[h$specie=="CANNOLICCHIO O CAPPALUNGA"]="CANNOLICCHIO"
h$specie[h$specie=="CAPPASANTA O CONCHIGLIA DI S. G."]
="CAPPASANTA C.G."
h$specie[h$specie=="CAPPELLANO O BUSBANA"]="BUSBANA"
h$specie[h$specie=="CARASSIO DORATO O PESCE ROSSO"]="CARASSIO DORATO"
h$specie[h$specie=="CEFALO CALAMITA O BOTOLO"]="CEFALO CALAMITA"
h$specie[h$specie=="CEFALO DORATO O LOTREGANO"]="CEFALO LOTREGANO"
h$specie[h$specie=="CEFALO O BOSEGA"]="CEFALO BOSEGA"
h$specie[h$specie=="CEFALO O VOLPINA"]="CEFALO VOLPINA"
h$specie[h$specie=="GALLINELLA O CAPPONE"]="GALLINELLA"
h$specie[h$specie=="GHIOZZETTO QUADRIMACULATO (SCAGIOTO)"]
="GHIOZZETTO MINUTO"
h$specie[h$specie=="GHIOZZO GÒ"]="GHIOZZO"
h$specie[h$specie=="GRANCHIO DA MOLECA  "]="GRANCHIO (Moleca)"
h$specie[h$specie=="GRANCHIO DA MOLECA (MAZANETA) "]
="GRANCHIO (Mazaneta)"
```

```
h$specie[h$specie=="GRANCHIO DA MOLECA (MUTATO/MOLECA)"]
="GRANCHIO (Moleca)"
h$specie[h$specie=="GRANSEOLA O GRANCEOLA"]="GRANSEOLA"
h$specie[h$specie=="LANZARDO O SGOMBRO OCCHIONE"]="LANZARDO"
h$specie[h$specie=="LATTERINO O ACQUADELLA"]="LATTERINO"
h$specie[h$specie=="LUCCIO DI MARE O BARRACUDA"]="BARRACUDA"
h$specie[h$specie=="LUCCIOPERCA O SANDRA"]="LUCCIOPERCA"
h$specie[h$specie=="MERLANO O MOLO"]="MOLO"
h$specie[h$specie=="MISTO"]="MISTO PESCE"
h$specie[h$specie=="NASELLO O MERLUZZO"]="MERLUZZO"
h$specie[h$specie=="OSTRICA PIATTA"]="OSTRICA"
h$specie[h$specie=="PANNOCCHIA O CANOCCHIA"]="PANNOCCHIA"
h$specie[h$specie=="PAPALINA O SPRATTO"]="PAPALINA"
h$specie[h$specie=="PESCE PRETE O LUCERNA"]="PESCE PRETE"
h$specie[h$specie=="PESCE S. PIETRO"]="PESCE S.PIETRO"
h$specie[h$specie=="POLPESSA O POLPO MACCHIATO"]="POLPESSA"
h$specie[h$specie=="ROMBO LISCIO O SOASO"]="SOASO"
h$specie[h$specie=="ROSPO O RANA PESCATRICE"]="RANA PESCATRICE"
h$specie[h$specie=="SARDINA (BIANCHETTO - NOVELLAME)"]="SARDINA"
h$specie[h$specie=="SEPPIA (T/PICCOLA)"]="SEPPIA (tag/P)"
h$specie[h$specie=="SEPPIETTA (SEPO)"]="SEPPIETTA"
h$specie[h$specie=="SEPPIOLA (ZOTOLO)"]="ZOTOLO"
h$specie[h$specie=="SPIGOLA O BRANZINO"]="BRANZINO"
h$specie[h$specie=="SPINAROLO SAGRÌ"]="SPINAROLO"
h$specie[h$specie=="SURO O SUGARELLO"]="SURO"
h$specie[h$specie=="TONNO ROSSO"]="TONNO"
h$specie[h$specie=="ZANCHETTA O SUACIA"]="ZANCHETTA"
h$specie=factor(h$specie)

head(h)
```

-Data Management, per gli anni dal 1997 al 2010:

```
#Peso pescato MA
h$peso.MA = 0
h$peso.MA[h$tpr=="MA"] = h$peso.kg[h$tpr=="MA"]

#Peso pescato LA
h$peso.LA = 0
h$peso.LA[h$tpr=="LA"] = h$peso.kg[h$tpr=="LA"]

#Peso pescato AD
h$peso.AD = 0
h$peso.AD[h$tpr=="AD"] = h$peso.kg[h$tpr=="AD"]

#Peso pescato VA
h$peso.VA = 0
h$peso.VA[h$tpr=="VA"] = h$peso.kg[h$tpr=="VA"]

conteggio <- function(x,...) length(unique(x))

h.out = cbind(
  aggregate(peso.kg ~ data + mese + anno + cod.spe + specie,
  data=h, sum),
  prezzo = aggregate(prezzo ~ data + mese + anno + cod.spe
  + specie, data=h, mean)$prezzo,
  aggregate(cbind(valore,peso.MA,peso.LA,peso.AD,peso.VA)
  ~ data + mese + anno + cod.spe + specie ,data=h, sum),
  nprod = aggregate(produttore ~ data + mese + anno + cod.spe
 + specie, data=h, conteggio)$produttore
)

h.out=h.out[,c(-8,-9,-10,-11,-12)]
head(h.out)
```

per gli anni dal 2011 al 2013:

```
#Peso pescato mare
h$peso.MA = 0
h$peso.MA[h$tpr=="Mare"] = h$peso.kg[h$tpr=="Mare"]

#Peso pescato LA
h$peso.LA = 0
h$peso.LA[h$tpr=="Laguna"] = h$peso.kg[h$tpr=="Laguna"]

#Peso pescato AD
h$peso.AD = 0
h$peso.AD[h$tpr=="Acqua Dolce"] = h$peso.kg[h$tpr=="Acqua Dolce"]

#Peso pescato VA
h$peso.VA = 0
h$peso.VA[h$tpr=="Valle"] = h$peso.kg[h$tpr=="Valle"]

conteggio <- function(x,...) length(unique(x))

h.out = cbind(
  aggregate(peso.kg ~ data + mese + anno + cod.spe + specie,
  data=h, sum),
  prezzo = aggregate(prezzo ~ data + mese + anno + cod.spe
  + specie, data=h, mean)$prezzo,
  aggregate(cbind(valore,peso.MA,peso.LA,peso.AD,peso.VA)
  ~ data + mese + anno + cod.spe + specie ,data=h, sum),
  nprod = aggregate(produttore ~ data + mese + anno + cod.spe
  + specie, data=h, conteggio)$produttore
)

h.out=h.out[,c(-8,-9,-10,-11,-12)]
head(h.out)
```

2. **Grafico serie storica**

-Apire file finale:

```
h=read.csv("Finale.csv")
head(h)
```

-Creare il grafico:

```
library(ggplot2)

h$date=as.Date(as.character(h$data),"%Y%m%d")
h$day=as.numeric(as.character(strftime(h$date, format = "%j")))
h$anno=as.factor(h$anno)


##scegliere la specie di pesce interessata

dati=subset(h, specie=="SARDINA")
h$pesots=ts(h$peso.kg)

bp=ggplot(data=dati, aes(x=day,
      y=pesots, group=anno, colour=anno)) + geom_line()
bp + ggtitle("Sardina")

#oppure si usa la media smussata:
ggplot(data=dati), aes(x=day, y=pesots,
       group=anno, colour=anno)) + geom_smooth(se=FALSE)
bp + ggtitle("Sardina")
```

3. **Nuova suddivisione in settimane**

-Creazione variabile week:

```
h$date=as.Date(as.character(h$data),"%Y%m%d")
h$anno=as.factor(h$anno)
h$week=as.numeric(as.character(strftime(h$date,format="%W")))
head(h)
```

-Sistemazione dataset:

```
conteggio <- function(x,...) length(unique(x))

h=cbind(
aggregate(pesots~week+anno+cod.spe+specie,data=h,sum),
prezzo=aggregate(prezzo~week+anno+cod.spe+specie,
   data=h,mean)$prezzo,
aggregate(cbind(valore,peso.MA,peso.LA,peso.AD,peso.VA)~
   week+anno+cod.spe+specie,data=h,sum),
nprod=aggregate(nprod~week+anno+cod.spe+specie,
   data=h,conteggio)$nprod
)

h=h[,c(-7,-8,-9,-10)]

head(h)
```

Scegliere la specie sulla quale si vuole lavorare, dopodiche si ripetarà il codice per fare il grafico del punto 2.

-Serie storica completa per i 17 anni di studio

```
plot(dati$pesots,type="l",col="red",main="Sardina")
a=c(53,106,159,212,265,318,371,424,477,530,583,636,689,742,795,848,901)
abline(v=a,col="black",lty=2)
```

4. **Serie Storiche**

-Grafici ACF e PACF

```
dati=subset(h.out, specie=="ALICE")
yts=ts(dati$peso.kg,start=c(1997,1), end=c(2014,0),frequency=52)
acf2(yts)

dati=subset(h.out, specie=="LATTERINO")
yts=ts(dati$peso.kg,start=c(1997,1), end=c(2014,0),frequency=52)
acf2(yts)

dati=subset(h.out, specie=="PANNOCCHIA")
yts=ts(dati$peso.kg,start=c(1997,1), end=c(2014,0),frequency=52)
acf2(yts)

dati=subset(h.out, specie=="SARDINA")
yts=ts(dati$peso.kg,start=c(1997,1), end=c(2014,0),frequency=52)
acf2(yts)

dati=subset(h.out, specie=="SEPPIA")
ytsS=ts(dati$peso.kg,start=c(1997,1), end=c(2014,0),frequency=52)
acf2(yts)

dati=subset(h.out, specie=="SOGLIOLA")
yts=ts(dati$peso.kg,start=c(1997,1), end=c(2014,0),frequency=52)
acf2(yts)
```

-auto.arima e modello ARIMA

```
auto.arima(yts)
```

```
m1=arima(yts, c(p,d,q), c(P,D,Q))
m1
```

inserire al posto di p,d,q e P,D,Q i valori per l'ARIMA.

-Grafico predizioni

```
pred=forecast(m1)
plot(pred)
```

5. **Correlazioni Canoniche**

   -Creare la matrice

```
dati=subset(h.out, specie=="ALICE")
Alice=ts(dati$peso.kg,start=c(1997,1), end=c(2014,0),frequency=52)

dati=subset(h.out, specie=="LATTERINO")
Latterino=ts(dati$peso.kg,start=c(1997,1),end=c(2014,0),frequency=52)

dati=subset(h.out, specie=="PANNOCCHIA")
Pannocchia=ts(dati$peso.kg,start=c(1997,1),end=c(2014,0),frequency=52)

dati=subset(h.out, specie=="SARDINA")
Sardine=ts(dati$peso.kg,start=c(1997,1),end=c(2014,0),frequency=52)

dati=subset(h.out, specie=="SEPPIA")
Seppia=ts(dati$peso.kg,start=c(1997,1),end=c(2014,0),frequency=52)

dati=subset(h.out, specie=="SOGLIOLA")
Sogliola=ts(dati$peso.kg,start=c(1997,1),end=c(2014,0),frequency=52)

sett=rep(c(1:52),17)
anno=rep(c(1997:2013), each=52)

datiLarge= cbind(anno,sett, Alice, Latterino, Pannocchia,
Sardine, Seppia, Sogliola)

datiLarge=cbind(stag=ceiling(datiLarge[,"sett"]/13),datiLarge)
colnames(datiLarge)=gsub("datiLarge.","",colnames(datiLarge))

stag=model.matrix(~0+.,data=data.frame(factor(datiLarge[,"stag"])))
colnames(stag)=c("inv","prim","est","aut")
```

-Funzione per trasformare una variabile categoriale in una matrice di variabili dummy

```
dummyfy <- function(vettore){
  res=outer(vettore,unique(vettore),"==")*1
  colnames(res)=unique(vettore)
  res
}
```

-Genera i nomi delle osservazioni

```
get.row.names <- function(datiLarge){
  paste(datiLarge[,"anno"],datiLarge[,"sett"],sep=".")
}
```

-Calcolo correlazioni canoniche e grafico

```
Y=cbind("stag"=dummyfy(datiLarge[,1]),anno=datiLarge[,2])
X=datiLarge[,-(1:3)]
row.names(X)<-row.names(Y) <- get.row.names(datiLarge)

library(CCA)

cc1=cc(X,Y)

plt.cc(cc1,var.label = TRUE,ind.names=rownames(datiLarge))
```

-Funzione PCbipolt per i grafici

```
PCbiplot <- function(PC, x=1, y=2, title="Biplot",
                 obs.names=NULL, obs.shape=NULL, obs.color=NULL,
obs.size=3, obs.label.size=obs.size,
obs.col.palette=NULL,
var.names=NULL, var.color=NULL,
filename=NULL,
loadingsTextJitter= position_jitter(w = 0.2, h = 0.2),
scoresJitter=  position_jitter(w = .2, h = .2),
addPercEV=TRUE,
obs.color.title=NULL,obs.shape.title=NULL,
var.suppl=NULL,var.suppl.color=NULL,var.text.size=3,
var.arrow.size=.2,...) {
#modificata da
#stackoverflow.com/questions/6578355/plotting-pca-biplot-with-ggplot2

##PC is a list with 1) x. the score matrix:
# nxp matrix with colnames (es "PC1" "PC2" etc.)
##2) rotation the matrix of loadings original variables as rows and
#PCs as columns.(required row/col-names)

  redUnipd="#9b0014"
  greyUnipd="#444F51"
  n=nrow(PC$x)

require(ggplot2,quietly =TRUE)
require(grid,quietly =TRUE)
if(is.null(var.names)) var.names=rownames(PC$rotation)
if(is.null(var.color)) var.color=redUnipd
if(is.null(var.suppl.color)) var.suppl.color=greyUnipd

if(is.numeric(x)) {idx=x; x=colnames(PC$x)[x]}
```

```
else
idx=which(colnames(PC$x)==x)
if(is.numeric(y)) {idy=y; y=colnames(PC$x)[y]}
else
idy=which(colnames(PC$x)==y)

if(is.null(obs.names)) obs.names=FALSE
if(is.logical(obs.names) &&
(obs.names==TRUE)) obs.names=if(is.null(row.names(PC$x)))
    as.character(1:n)
else
row.names(PC$x)

  if(is.null(obs.color)) obs.color=1
  if(obs.color[1]=="each.obs") {
    obs.color=1:n
    if(is.null(obs.col.palette))
      obs.col.palette=heat.colors(n)
  }
  if(is.null(obs.col.palette))
    obs.col.palette=heat.colors(length(unique(obs.color)))

if(is.null(obs.shape)) obs.shape=1

  if(is.null(obs.shape.title)) obs.shape.title="obsshape"
if(is.null(obs.color.title)) obs.color.title="obscolor"

# PC being a prcomp object
data <- data.frame(obsnames=obs.names, obscolor=
obs.color, obsshape=obs.shape, PC$x)

names(data)[names(data)==x]="x"
names(data)[names(data)==y]="y"
```

```
data$obsshape=as.character(data$obsshape)
#  obsorder<-data$obsorder<-as.numeric(factor(data$obscolor))
data$obscolor=as.character(data$obscolor)
data$obsnames=as.character(data$obsnames)

#                                  order=obsorder
#per evitare legende non volute: (NON FUNZIONA)
#rivedere
#stackoverflow.com/questions/11714951/remove-extra-legends-in-ggplot2
if(length(unique(data$obscolor))==1){
  if(length(unique(data$obsshape))==1){
#     plotout=ggplot(data, aes(x=x, y=y, label=obsnames),
#                         colour=obscolor,shape=obsshape)
    plotout=ggplot(data,aes(x=x,y=y,label=obsnames,shape=obsshape,
                                colour=obscolor))

  } else {
    plotout=ggplot(data,aes(x=x,y=y,label=obsnames,shape=obsshape),
                    colour=obscolor)
  }
} else{
  if(length(unique(data$obsshape))==1){
    plotout=ggplot(data,aes(x=x,y=y,label=obsnames,colour=obscolor),
                    shape=obsshape)
  } else {
    plotout=ggplot(data,aes(x=x,y=y,label=obsnames,shape=obsshape,
                    colour=obscolor))
  }

}

  plotout=plotout + geom_point(size=obs.size,fill="gray",
     position = scoresJitter)
```

```
  plotout=plotout + scale_shape(obs.shape.title)
if(!is.null(obs.names) && obs.names!=FALSE)
plotout=plotout + geom_text(vjust=0, position = scoresJitter,
                          size=obs.label.size)
if(!is.null(obs.col.palette))
    plotout <- plotout + scale_colour_manual(obs.color.title,
                         values=obs.col.palette) else
      plotout <- plotout + scale_colour_manual(obs.color.title,
                         values="darkgrey")

     xLabel=ifelse(addPercEV,paste(x," (",round(PC$sdev[idx]^2/
                  sum(PC$sdev^2)*100,0),"%)",sep=""),x)
     yLabel=ifelse(addPercEV,paste(y," (",round(PC$sdev[idy]^2/
                  sum(PC$sdev^2)*100,0),"%)",sep=""),y)
plotout <- plotout +  theme(panel.background =
      element_rect(fill='gray100'),
       legend.background = element_rect(fill='gray100')) +
        scale_x_continuous(xLabel)+scale_y_continuous(yLabel)
plotout=plotout + geom_hline(aes(0), size=.2)
     + geom_vline(aes(0), size=.2)
plotout=plotout + guides(fill=guide_legend(title=NULL))
  plotout <- plotout + ggtitle(title) +
  theme(plot.title = element_text(lineheight=.8, face="bold"))
# Using a manual scale instead of hue
# plotout=plotout+ scale_fill_manual(values=
       c("#999999", "#E69F00", "#56B4E9"),
            # name="Experimental\nCondition",
            # breaks=c("ctrl", "trt1", "trt2"),
            # labels=c("Control","Treatment 1","Treatment 2"))

datapc <- data.frame(obsnames=var.names, PC$rotation)
```

```
names(datapc)[names(datapc)==x]="x"
names(datapc)[names(datapc)==y]="y"

    # mult <- min(
#(max(data[,y])-min(data[,y])/(max(datapc[,y])-min(datapc[,y]))),
#(max(data[,x])-min(data[,x])/(max(datapc[,x])-min(datapc[,x])))
#)
   mult <- min(
     (max(data$y)-min(data$y)/(max(datapc$y)-min(datapc$y))),
     (max(data$x)-min(data$x)/(max(datapc$x)-min(datapc$x)))
      )

    datapc <- transform(datapc,
            v1 = .7 * mult * x,
            v2 = .7 * mult * y,
            arrowshape= "1")


plotout=plotout + coord_equal() +
  geom_segment(data=datapc, aes(x=0, y=0, xend=v1, yend=v2,
   shape=arrowshape), arrow=arrow(length=unit(0.2,"cm")),
     alpha=0.75, color=var.color, size=var.arrow.size)

plotout=plotout+geom_text(data=datapc,
   position=loadingsTextJitter,
  aes(x=v1, y=v2, label=obsnames,shape="1"),
    size = var.text.size, vjust=1, color=var.color)


  if(!is.null(var.suppl)){
    datapc.suppl=data.frame(x=cov(var.suppl,data$x),
                      y=cov(var.suppl,data$y),
                      obsnames=colnames(var.suppl))
```

```
    mult <- min(
      (max(data$y) - min(data$y)/(max(datapc$y)-min(datapc$y))),
      (max(data$x) - min(data$x)/(max(datapc$x)-min(datapc$x)))
    )
 datapc.suppl$v1 <-  .7 * mult * datapc.suppl$x
 datapc.suppl$v2 <-  .7 * mult * datapc.suppl$y
 datapc.suppl$arrowshape <- "1"
 plotout=plotout+geom_segment(data=datapc.suppl,
   aes(x=0, y=0, xend=v1, yend=v2,shape=arrowshape),
     arrow=arrow(length=unit(0.2,"cm")), alpha=0.75,
       color=var.suppl.color, size=.5)

plotout=plotout + geom_text(data=datapc.suppl,
  position = loadingsTextJitter ,
    aes(x=v1, y=v2, label=obsnames,shape="1"),
      size = var.text.size, vjust=1, color=var.suppl.color)
  }

#mi pare che non funzioni
unici_gruppi=apply(cbind(data$obscolor,data$obsshape),
   1,paste,collapse="")
if(length(unique(unici_gruppi))==1) {
  plotout=plotout+theme(legend.position="none")
}


  if(!is.null(filename)) {plotout
                          ggsave(file=filename)
                          } else plotout

}
```

-Funzione CCbiplot per i grafici

```
CCbiplot<-  function(PC, x=1, y=2, title="Biplot",
                obs.names=NULL, obs.shape=NULL, obs.color=NULL,
                obs.size=3, obs.label.size=obs.size,
                obs.col.palette=NULL,
                var.names.x=NULL, var.names.y=NULL,
                var.color.x=NULL, var.color.y=NULL, ...) {
  nvar.x=nrow(PC$scores$corr.X.xscores)
  nvar.y=nrow(PC$scores$corr.Y.xscores)
  PC=list(x=PC$scores$xscores,
        rotation=rbind(PC$scores$corr.X.xscores,
          PC$scores$corr.Y.xscores),
        sdev=PC$cor)
  colnames(PC$rotation)=paste("CC",1:ncol(PC$rotation),sep="")
  colnames(PC$x)=paste("CC",1:ncol(PC$x),sep="")

  var.names=c(var.names.x, var.names.y)

  if(is.null(var.color.x))
    var.color.x="#444F51"
  if(is.null(var.color.y))
    var.color.y="#9b0014"

  var.color=c(rep(var.color.x,length.out=nvar.x),
              rep(var.color.y,length.out=nvar.y))
  PCbiplot(PC, x=x, y=y,  title=title,
     obs.names=obs.names, obs.shape=obs.shape, obs.color=obs.color,
     obs.size=obs.size, obs.label.size=obs.label.size,
     obs.col.palette=obs.col.palette,
     var.names=var.names, var.color=var.color,
     ...)
}
```

-biplot delle correlazioni X-Y

```
my.biplot <-function(cc1)
  CCbiplot(mod, x=1, y=2, title="Correlazioni X-Y",
                      obs.names=row.names(Y), obs.shape=NULL,
          obs.size=0,obs.label.size=3,obs.col.palette="darkgrey",
          var.color.x="black", var.color.y=NULL,var.text.size=4)
my.biplot(cc1)
```

Bibliografia

[1] Mazzoldi C., Sambo A. and Riginella E. (2014). The Clodia database: a long series of fishery data from the Adriatic Sea. *Background & Summary*

[2] Vianelli S. (1983). L'analisi delle serie temporali nello sviluppo storico e metodologico della statistica, in Analisi moderna delle serie storiche, a cura di Piccolo, D., Franco Angeli, Milano.

[3] Di Fonzo T., Lisi F. (2009), Serie storiche economiche, Carocci, Roma.

[4] Box G.E.P., Jenkins G.M. (1976), Time Series Analysis, Forecasting an Control, Holden-Day, San Francisco.

[5] Masarotto G. (2005), Analisi delle Serie Temporali - Lucidi delle lezioni.

[6] Piccolo D. (1990), Introduzione all'analisi delle serie storiche, Carocci, Roma.

[7] Finos L. (2013), Analisi dei dati multidimensionali - Appunti delle lezioni.

[8] Mardia K.V., Kent J.T. and Bibby J.M. (1980), Multivariate Analysis (Probability and Mathematical Statistics), Academic Press Edition.

[9] Crivellari F. (2006), Analisi statistica dei dati con R, Apogeo.

[10] Scardi M. (2009), Tecniche di analisi dei dati in ecologia - Slide del corso.

[11] Gabriel K.R. (1971), The biplot graphical display of matrices with application to principal component analysis. Biometrika 58, 453–467.

Ringraziamenti

Desidero ringraziare in primis il Prof. Livio Finos e la Prof.ssa Carlotta Mazzoldi, non solo per essere relatore e correlatore della mia tesi, ma anche per avermi permesso di vivere quella grande esperienza che è stato il mio stage a Chioggia.

Un ringraziamento va alla mia famiglia, a mamma, papà e mia sorella Elena, sia per il sostegno economico, i primi due, non da poco, ma anche per quello morale, visto che hanno sempre creduto in me, forse anche più di quanto ci abbia mai creduto io. Un ringraziamento a tutti i nonni e gli zii che mi stanno sempre accanto e fanno il tifo per me.

Da ultimi ma non ultimi, volevo ringraziare tutti gli amici, sia quelli storici, che sono la mia seconda famiglia, e che nel bene nel male ci saranno sempre, sia tutti quelli che ho conosciuto in questi ultimi anni e che spero rimarranno per molto tempo. Un particolare grazie agli amici della facoltà, con i quali ho condiviso questi tre (anche cinque) anni di studio matto e disperatissimo.

Anni fa non avrei mai detto possibile arrivare a questo traguardo, adesso invece è tutto vero.

... e quindi uscimmo a riveder le stelle.