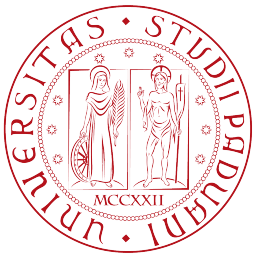
Università degli Studi di Padova
DIPARTIMENTO DI SCIENZE STATISTICHE
Corso di Laurea Magistrale in
Scienze Statistiche



CLASSIFICAZIONE DELLE SERIE STORICHE: ALCUNE ANALISI

Relatore:
Prof Lisi Francesco
Dipartimento di Scienze Statistiche

Laureanda:
Crepaldi Marica
Matricola: 1082716

Anno Accademico
2015/2016

Indice

Introduzione

In questo lavoro si vuole effettuare Cluster Analysis nell'ambito delle serie storiche. Con Cluster Analysis, detta anche analisi dei grappoli, si intende il processo che suddivide un insieme generico di osservazioni in gruppi di osservazioni simili. Tali metodi sono stati sviluppati fin dalla fine del XIX secolo e si valuta che gli algoritmi elaborati fino ad oggi siano circa un migliaio. L'uso della Cluster Analysis pone problemi relativi alla scelta di una metrica che sia in grado di esprimere sinteticamente la distanza tra gli elementi del campione che si vogliono raggruppare. Nello specifico contesto dei dati temporali, il concetto di somiglianza è particolarmente complesso date le caratteristiche dinamiche della serie. Le misure di somiglianze generalmente considerate nella analisi di cluster per dati indipendenti non funzionano in maniera adeguata con dati che hanno una dipendenza temporale perché ignorano le relazioni di interdipendenza tra i valori. È opportuno quindi usare delle misure di distanza che tengano conto del comportamento della serie nel tempo.

In questo lavoro, oltre a presentare delle misure di distanza per serie storiche già presenti in letteratura, verrà posta particolare attenzione a due misure. La prima misura, proposta in un recente articolo di Lafuente-Rego e Vilar (2015) e la seconda è una misura mai proposta prima.

Nella seconda parte di questo lavoro verrà presentata una procedura, mai proposta prima, in grado di discriminare serie storiche provenineti da processi molto differenti tra loro. Questa procedura è stata proposta per due motivi. Il primo è che nell'ambito delle serie storiche una divisione ragionevole può essere effettuata in base al processo generatore dei dati e, in letteratura, non esistono delle misure di dissimilarità che siano in grado di dividere le serie storiche provenienti da processi con caratteristiche molto differenti. Il secondo motivo che ha portato all'introduzione di questa procedura è che, oltre che avere una divisione dei dati in base al processo generatore, è utile avere anche un'etichettatura che permetta il riconoscimento del tipo di processo sottostante.

Nell'ultima parte del lavoro verrà effettuata Cluster Analysis su dati rela-

tivi ai prezzi di energia del mercato elettrico inglese. In particolare verranno applicate le due misure di dissimilarità trattate in maniera particolare in questo lavoro.

Questo lavoro è diviso in 4 capitoli.

Nel primo capitolo viene introdotto il concetto di Cluster Analysis nella sua generalità: viene definito il concetto di matrice delle distanze, vengono presentati i tipi di legame maggiormente usati, e vengono presentati alcuni indici per la valutazione dei raggruppamenti ottenuti. Nel secondo capitolo vengono presentate alcune misure di dissimilarità tipiche delle serie storiche. Viene posta particolare attenzione a due misure di dissimilarità particolari. Per valutare la performance di queste due nuove misure di dissimilarità viene anche condotto uno studio di simulazione. Nel terzo capitolo viene spiegata la nuova procedura di discriminazione proposta in questo lavoro. In particolare, prima vengono spiegati gli obiettivi e i passi che si intendono seguire, poi viene effettuato uno studio di simulazione. Nel quarto capitolo si effettua Cluster Analysis su dei dati reali relativi ai prezzi di energia del mercato elettrico inglese.

Capitolo 1

La Cluster Analysis

In questo Capitolo si parlerà di tecniche di Cluster Analysis. Non si farà particolare attenzione al caso delle serie storiche perché i concetti introdotti di seguito sono di carattere generale. Alcuni riferimenti utili per le tecniche spiegate in questo Capitolo sono Zani e Cerioli (2007), Fabbris (1997) e Liao (2005).

1.1 Introduzione

La Cluster Analysis gioca un ruolo centrale in molti campi tra i quali l'economia, la finanza, la medicina, l'ecologia, studi ambientali, e molti altri. Essa è un importante strumento per l'esplorazione della struttura presente nei dati. Può essere definita come un insieme di metodi e procedure finalizzati al raggruppamento degli oggetti in categorie o classi, in base a delle caratteristiche di somiglianza. Lo scopo della Cluster Analysis è quindi quello di raggruppare le unità sperimentali in gruppi secondo criteri di similarità, cioè determinare un certo numero di classi in modo tale che le osservazioni siano il più possibile omogenee all'interno delle classi ed il più possibile disomogenee tra le diverse classi. Tuttavia, vale la pena sottolineare che la scelta di una adeguata misura di dissimilarità deve tener conto principalmente dello scopo specifico del compito di clustering.

La Cluster Analysis ha, tra gli altri, i seguenti obiettivi:

- *Ridurre i dati* in forma (anche grafica) in modo tale da rendere facile la lettura delle informazioni rilevate e rendere parsimoniosa la presentazione dei risultati;

- *Generare ipotesi di ricerca* per effettuare un'analisi di raggruppamento. Per effettuare un'analisi di clustering non è necessario avere in mente alcun modello interpretativo;

- *Ricerca tipologica* per individuare gruppi di unità statistiche con caratteristiche distintive che facciano risaltare la fisionomia del sistema osservato;

- *Ricerca di classi omogenee*, dentro le quali si può supporre che i membri siano mutuamente surrogabili;

- *Costruzione di sistemi di classificazione automatica* attraverso la definizione di un "classificatore accurato" che consenta di classificare nuove unità;

Le tecniche di Cluster Analysis si possono suddividere in due grandi categorie: *analisi gerarchica* e *analisi non gerarchica*. Nel primo tipo di analisi ogni classe fa parte di una classe più ampia, la quale è contenuta a sua volta in una classe di ampiezza superiore, e così in progressione fino alla classe che contiene l'intero insieme di dati. Nel secondo tipo di analisi invece i gruppi sono non gerarchizzabili, e per questo motivo si deve decidere a priori il numero dei gruppi.

Le tecniche di analisi gerarchica si possono ulteriormente distinguere in:

- *agglomerative*, se prevedono una successione di fusioni delle n unità, a partire dalla situazione di base nella quale ognuna costituisce un gruppo a sé stante e fino allo stadio $(n-1)$ nel quale si forma un gruppo che le comprende tutte;

- *divisive* o *scissorie*, quando l'insieme delle n unità, in $(n-1)$ passi si ripartisce in gruppi che sono, ad ogni passo dell'analisi, sottoinsiemi di un gruppo formato allo stadio di analisi precedente, e che termina con la situazione in cui ogni gruppo è composto da un'unità.

Le tecniche di analisi non gerarchica possono essere divise in:

- *partizioni*, ossia classi mutuamente esclusive e tali che, per un numero di gruppi prefissato, è possibile classificare un'entità in una e una sola classe;

- *classi sovrapposte*, per le quali si ammette la possibilità che un'entità possa appartenere contemporaneamente a più di una classe.

In questo lavoro ci si concentrerà sulle analisi gerarchiche agglomerative. Un'eccellente rassegna delle tecniche di Cluster Analysis per serie storiche è riportata da Liao (2005).

1.2 Distanze e dissimilarità

Nelle analisi di clustering di tipo gerarchico un ruolo fondamentale è giocato dalle misure di vicinananza tra gli individui o, equivalentemente, dalla loro dissomiglianza. Dato un insieme di serie storiche $X^i = \{X_1^i, \cdots, X_T^i\}$ per $i = 1, \cdots, N$, la funzione dissomiglianza $d(X^i, X^j)$, nella sua definizione più generale, è una funzione non negativa definita su ogni coppia di serie tale che, un elevato livello di similarità fra due serie è caratterizzato da un piccolo valore della loro dissomilgianza.

Nella pratica le misure di distanza devono rispettare alcune di queste condizioni:

- *Identità* $d(X^i, X^i) = 0$;
- *Non negatività* $d(X^i, X^j) \geq 0$;
- *Simmetria* $d(X^i, X^j) = d(X^j, X^i)$;
- *Disuguaglianza triangolare* $d(X^i, X^j) + d(X^j, X^k) \geq d(X^i, X^k)$
- $d(X^i, X^j) \leq max\left\{d(X^i, X^k), d(X^k, X^j)\right\}$

Una misura di dissomiglianza che gode delle proprietà di non negatività, identità e simmetria prende il nome di *indice di dissimilarità*. Si parla invece di *distanza* o di *metrica* nel caso di funzioni $d(X^i, X^j)$ che rispettano, oltre alle prime tre proprietà, anche la disuguaglianza triangolare. Se inoltre soddisfa anche la quinta proprietà si parla di *ultrametrica*.

La scelta della funzione distanza da utilizzare dipende dal problema che si vuole risolvere. Vengono esaminate, nel Capitolo 2, alcune misure di dissimilarità già presenti in letteratura assieme a due nuove misure di dissimilarità. La prima si basa sulla funzione dei quantili di autocovarianza, proposta nel recente lavoro di Lafuente-Rego e Vilar (2015). La seconda misura proposta nasce dall'idea che le serie storiche spesso possono essere espresse come la somma di componenti deterministiche, quali Componente di lungo periodo trend e/o componenti cicliche, e da una componente residuale d'errore. Tale misura di dissimilarità vuole cercare di distinguere le serie storiche in base alla stima non parametrica delle componenti deterministiche.

1.3 Metodi di raggruppamento

Un'ampia classe di metodi gerarchici si fonda sull'impiego iniziale di un matrice $\mathbf{D}$ di distanze, calcolata per le n unità statistiche. In tal caso la procedura seguita per individuare la partizione successiva si articola nelle seguenti fasi:

1. Si individuano le unità con la minore distanza nella matrice $\mathbf{D}$, ossia quelle più simili tra loro, e si riuniscono a formare il primo gruppo: si ottiene così una partizione con $(n-1)$ gruppi di cui $(n-2)$ costruiti dalle singole unità e l'altro formato da due unità.

2. Si ricalcola, adottando un certo metodo, la distanza del gruppo ottenuto dagli altri gruppi (eventualmente costruiti da una sola unità), ricavando una nuova matrice, con dimensione diminuita di uno.

3. Si individua nella nuova matrice delle distanze, la coppia di unità, o gruppi, con la minore distanza, riunendole in un unico gruppo.

4. si ripetono le fasi 2) e 3) sino a quando tutte le unità sono riunite in un solo gruppo.

Le differenze tra i vari metodi gerarchici consistono nel criterio utilizzato per calcolare la distanza tra i gruppi di unità appena formati con i gruppi formati in precedenza, eventualmente formati da una sola unità.

Si considerino due gruppi C_1 e C_2 e siano n_1 e n_2 le rispettive numerosità. Sono possibili diverse definizioni di distanza tra due gruppi, in particolare si considerano: il metodo del *legame singolo*, il metodo del *legame completo*, il metodo del *legame medio* e il metodo di *Ward*. Tali legami richiedono esclusivamente la conoscenza della matrice di distanze.

1.3.1 Metodo del legame singolo

Il metodo del legame singolo, chiamato anche *single linkage* o del vicino più vicino (*nearest neighbour*) definisce la distanza tra due gruppi come il minimo delle $n_1 n_2$ distanze tra ciascuna delle unità di un gruppo in ciascuna delle unità dell'altro gruppo:

$$d(C_1, C_2) = min(d_{rs}), \text{ per } r \in C_1, s \in C_2.$$

Il metodo del legame singolo presenta il cosiddetto effetto catena, cioè può riunire in un unico gruppo elementi anche di molto distanti in $\mathbb{R}^p$ quando tra essi esiste una successione di punti intermedi. Questo effetto è posto in evidenza in Figura 1.1(a), in cui si vedono chiaramente tre nuvole di punti nel piano, ma la presenza di alcune unità (gli anelli della catena) conduce, con questo metodo, alla riunione in un unico gruppo delle due nuvole a sinistra ed in alto: tale gruppo non presenta tuttavia coesione interna. L'effetto catena, se da un lato rappresenta uno svantaggio, dall'altro ha il pregio di consentire l'individuazione di gruppi anche con forme molto diverse da quelle sferiche. Si veda a questo proposito la Figura 1.1(b): vi sono due gruppi nel piano,

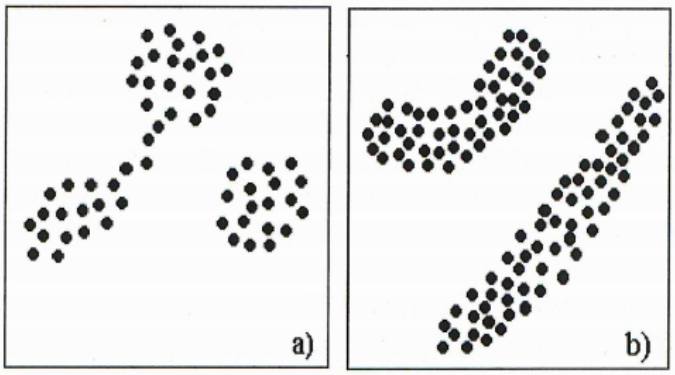


Figura 1.1: Rappresentazione grafica dell'effetto a catena (a) e di gruppi a forma non circolare (b).

nettamente separati tra loro, l'uno a forma di fagiolo e l'altro a forma di sigaro, che il metodo del legame singolo riesce ad individuare grazie all'effetto catena.

1.3.2 Metodo del legame completo

Il metodo del legame completo, detto anche *complete linkage* o del vicino più lontano (*further neighbourd*) definisce la distanza tra due gruppi come il massimo delle $n_1 n_2$ distanze tra ciascuna della unità d'un gruppo e ciascuna delle unità dell'altro gruppo:

$$d(C_1, C_2) = max(d_{rs}), \text{ per } r \in C_1, s \in C_2.$$

Adottando questo criterio, tutte le distanze tra le unità del primo gruppo e quelle del secondo gruppo sono minori o uguali alla distanza tra i due gruppi così definita. Con tale metodo si individuano gruppi compatti al loro interno, ma di forma approssimativamente circolare (in $\mathbb{R}^2$). Nell'esempio di Figura 1.1(a), il metodo del legame completo trova correttamente i tre gruppi assegnando punti intermedi ai gruppi più prossimi. Nell'esempio della Figura 1.1(b) però il metodo del legame completo non è in grado di individuare quei due gruppi, data proprio la loro non sfericità.

1.3.3 Metodo del legame medio

Il metodo del legame medio o *average linkage* definisce la distanza tra due gruppi come la media aritmetica delle $n_1 n_2$ distanze tra ciascuna delle unità

di un gruppo e ciascuna delle unità dell'altro:

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_r \sum_s d_{rs} \text{ per } r \in C_1, s \in C_2.$$

Il metodo del legame medio costituisce un compromesso ragionevole tra una discreta coesione interna e separazione esterna.

1.3.4 Metodo di Ward

Il metodo di Ward definisce la distanza tra due gruppi tramite la minimizzazione della varianza delle variabili entro ciascun gruppo. Il metodo è quello della minimizzazione di una funzione obiettivo che vuole realizzare la massima coesione interna a ciascun gruppo e la massima separazione esterna tra gruppi diversi. La tecnica è iterativa e ad ogni passo vengono fusi i gruppi che presentano la minima variazione della varianza entro i gruppi. La devianza totale viene scomposta in devianza nei gruppi e devianza fra i gruppi, e, ad ogni passo della procedura gerarchica, si aggregano tra loro i gruppi che comportano il minore incremento della devianza nei gruppi e il maggiore incremento della devianza tra gruppi.

Tutte le tecniche gerarchiche esaminate finora dette tecniche *gerarchico-agglomerative*, possono essere viste come varianti di un unica tecnica generale (Lance e Williams, 1967) che può essere espressa in forma compatta e ricorsiva nei termini seguenti:

1. si parte da una situazione con n cluster con un oggetto ciascuno;

2. si uniscono i due gruppi i e j che minimizzano la misura di dissimilarità d_{ij};

3. si ripete il passo (2) finché tutti gli oggetti non formano un solo gruppo.

La misura di dissimilarità fra gruppi può essere calcolata ricorsivamente. All'inizio, nel passo (1) le dissimilarità degli n gruppi coincidono con le dissimilarità tra gli n oggetti. Nei passi successivi la misura della dissimilarità fra il gruppo k-esimo e il gruppo ottenuto dalla fusione dei gruppi i-esimo e j-esimo si calcola sulla base della seguente espressione:

$$d_{k,ij} = \alpha(i)d_{ki} + \alpha(j)d_{kj} + \beta d_{ij} + \Gamma|d_{ki} - d_{kj}|,$$

nella quale i parametri $\alpha(i)$, $\alpha(j)$, β, Γ si possono determinare dipendentemente dalla tecnica adottata, come illustrato in Tabella 1.1.

	$\alpha(i)$	$\alpha(j)$	β	Γ
L. singolo.	1/2	1/2	0	-1/2
L. completo	1/2	1/2	0	1/2
L. medio	$\frac{n_i}{(n_i+n_j)}$	$\frac{n_j}{(n_i+n_j)}$	0	0
L. di Ward	$\frac{n_i+n_k}{n_k+n_i+n_j}$	$\frac{n_j+n_k}{n_k+n_i+n_j}$	$-\frac{n_k}{n_k+n_i+n_j}$	0

Tabella 1.1: Parametri per il calcolo delle misure di dissimilarità fra gruppi.

1.4 Indici di validità esterna

L' approccio esterno alla validazione presuppone la conoscenza del vero numero di gruppi e della loro composizione. Se si hanno a disposizione le etichette delle classi, si può eseguire il clustering per comparare risultati provenienti dall'applicazione di diversi algoritmi, con l'obiettivo di individuare l'algoritmo ottimale per uno specifico insieme di dati. Gli indici esterni forniscono una misura dell'accuratezza dei risultati, in termini di quante osservazioni sono correttamente classificate secondo le etichette fornite a priori e quante, invece, risultano appartenenti ad una classe cui non dovrebbero essere associate. Di seguito vengono presentati l'indice Γ di Hubert, l'indice di Gavrilorv, l'indice di Rand e l'indice di Rand corretto, e si indicheranno con $U = \{u_1, ..., u_R\}$ e $V = \{v_1, ..., v_C\}$ la partizione vera e la partizione trovata con la procedura di clustering, rispettivamente.

1.4.1 Indice Γ di Hubert

L'indice Γ di Hubert (Hubert e Schultz, 1976) misura la correlazione tra le partizioni U e V. Esso è definito come:

$$\begin{aligned}\Gamma &= Corr(Y_U, Y_V) \\ &= \frac{1}{\binom{n}{2}\sigma_{Y_U}\sigma_{Y_V}} \sum_{i<j} (Y_U(i,j) - \mu_{Y_U})(Y_V(i,j) - \mu_{Y_V})\end{aligned}$$

dove $Y_U(i,j)$ e $Y_V(i,j)$ sono due variabili casuali binarie che valgono 1 se X^i e X^j sono classificati nello stesso modo nella partizione U e V rispettivamente, e μ_{Y_a} e $\sigma^2_{Y_a}$ sono definite come segue:

$$\mu_{Y_a} = \frac{1}{N} \sum_{i<j} Y_a(i,j),$$
$$\sigma^2_{Y_a} = \frac{1}{N} \sum_{i<j} (Y_a(i,j)^2),$$

per a=U,V. Tale indice varia da -1 a 1.

1.4.2 Indice di Gavrilov

Un altro modo per calcolare un indice di validità esterna è quello di calcolare una misura di accordo tra la partizione vera U e la partizione trovata dalla procedura di clustering che si sta valutando V. L'indice di Gavrilov o indice di similarità (Gavrilov *e altri*, 2000) è calcolato come:

$$Sim(U,V) = \frac{1}{k} \sum_{i=1}^{K} \max_{i \leq j \leq K} Sim(u_i, v_j),$$

dove

$$Sim(u_i, v_j) = \frac{\mid u_i \cap v_j \mid}{\mid u_i \mid + \mid v_j \mid},$$

con $|\cdot|$ che indica la cardinalità degli elementi nel cluster.

Tale indice varia da 0 a 1 e vale 1 quando le due partizioni coincidono.

1.4.3 Indice di Rand

Le informazioni relative alle due partizioni U e V possono essere riassunte in una tabella di contingenza, come la Tabella 1.2, dove n_{ij} denota il numero di osservazioni che fanno parte dei cluster u_i e v_j, con $i = 1, \cdots R$, $j = 1, \cdots, C$. Siano inoltre $n_{i\cdot} = \sum_{j=1}^{C} n_{ij}$ e $n_{\cdot j} = \sum_{i=1}^{R} n_{ij}$ le somme per riga e per colonna rispettivamente, ossia il numero di osservazioni nei gruppi u_i e v_j, e sia $Z = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}^2$.

L'indice di Rand (Rand, 1971) può essere calcolato come:

$$Rand = 1 + \frac{Z - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i\cdot}^2 + \sum_{j=1}^{C} n_{\cdot j}^2\right)}{\binom{n}{2}}. \tag{1.1}$$

Un altro modo per esprimere l'indice di Rand è quello di considerare le seguenti quattro possibilità per una data coppia di punti, x_i e x_j:

(a) x_i e x_j sono nello stesso cluster in entrambe le partizioni;

	v_1	v_2	$\cdots$	v_C	
u_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1\cdot}$
u_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot C}$	n

Tabella 1.2: Tabella di contingenza per le due partizioni.

(b) x_i e x_j sono nello stesso cluster nella partizione trovata ma sono in cluster differenti nella partizione vera;

(c) x_i e x_j sono in cluster differenti nella partizione trovata ma sono nello stesso cluster nella partizione vera;

(d) x_i e x_j sono in cluster diversi sia nella patizione vera sia nella partizione trovata.

Siano a, b, c e d il numero di coppie di punti che appartengono alle situazioni (a), (b), (c) e (d) rispettivamente. L'indice di Rand misura la proporzione di coppie di punti che concordano con l'appartenenza allo stesso cluster (a) o di diversi cluster (d) in entrambe le partizioni. Esso è quindi definito come:

$$Rand = \frac{a+b}{a+b+c+d}.$$

Gli indici di fatto risultano numericamente uguali in quanto:

$$a = \sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}$$

$$b = \sum_{j=1}^{C}\binom{n_{\cdot j}}{2} - \sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}$$

$$c = \sum_{i=1}^{R}\binom{n_{i\cdot}}{2} - \sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}$$

$$d = \binom{n}{2} - a - b - c$$

ma la prima versione dell'indice non necessita di $n(n-1)/2$ confronti a coppie ma solo della conoscenza del cluster di apparteneza.

L'indice di Rand varia da 0 a 1, e vale 1 quando le due partizioni coincidono.

1.4.4 Indice di Rand corretto

Hubert e Arabie (Hubert e Arabie, 1985) hanno proposto una correzione dell'indice di Rand, in modo che assuma valore atteso 0 quando le due partizioni sono determinate casualmente, e che assuma il valore 1 quando c'è perfetta corrispondenza tra esse. L'indice di Rand corretto può essere calcolato nei modi seguenti:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]},$$

oppure come

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_j \binom{n_{\cdot j}}{2}\sum_i \binom{n_{i\cdot}}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_j \binom{n_{\cdot j}}{2} + \sum_i \binom{n_{i\cdot}}{2}] - [\sum_j \binom{n_{\cdot j}}{2}\sum_i \binom{n_{i\cdot}}{2}]/\binom{n}{2}}$$

dove $n_{ij}, n_{\cdot j}, n_{i\cdot}$, sono i relativi valori corrispondenti alla Tabella di contingenza 1.2 e a,b,c e d sono il numero di coppie di punti che appartengono alle situazioni (a), (b), (c) e (d) decritte in 1.4.3. Tale indice varia da $-\infty$ a un massimo di 1, che viene raggiunto in caso di perfetta corrispondenza tra le due partizioni. Di fatto tale indice difficilmente risulta negativo.

1.5 Indici di validità interna

Una volta ottenuta la divisione in cluster è utile capire quanti cluster tenere in considerazione per poter deteminare la soluzione ottimale di clustering. Per far ciò vengono proposti degli indici di validità interna che tengono conto unicamente dei dati a disposizione e della misura di dissimilarità in questione. Questo genere di indicatori hanno l'obiettivo di trovare la soluzione di clustering più compatta e meglio separata.

1.5.1 Indice di Dunn generalizzato

L'indice di Dunn (Dunn, 1974) è così definito:

$$Dunn(n_c) = \min_{i=1,\cdots,n_c}\left\{\min_{j=i+1,\ldots n_c}\left\{\frac{\min_{X\in C_i, X'\in C_j} d(X,X')}{\max_{k=1,\cdots,n_c}\{\max_{X,X'\in C_k} d(X,X')\}}\right\}\right\}.$$

L'obiettivo è di massimizzare la distanza inter-cluster e di minimizzare la distanza intra-cluster. Risulta quindi il rapporto tra la minima distanza tra cluster diversi e la massima distanza tra elementi dello stesso cluster. L'indice varia da 0 a ∞. Valori alti dell'indice indicano che la minima distanza tra

i cluster, riportata al numeratore, è molto elevata e c'è quindi una buona separazione esterna, e che la distanza tra gli elementi dello stesso cluster, riportata al denominatore, è molto bassa e quindi c'è una buona coesione interna. Valori alti indicano quindi l'esistenza di gruppi compatti e ben separati.

1.5.2 Coefficiente Silhouette

Il calcolo del coefficiente Silhouette non è immediato. Si considerino le serie storiche $X^i \in C_k$ e n_k la numerosità di ogni cluster, con $i = 1, ..., N$ e $k = 1, ..., K$. Una definizione della coesione interna può essere data dalla varianza entro-cluster $a(X^i)$ data dalla distanza media della serie X^i dalle altre serie del cluster a cui essa appartiene. Una definizione della separatezza esterna $b(X^i)$ può essere data dalla minima distanza media di X^i dagli altri cluster a cui X^i non appartiene. Nel dettaglio si ha

$$a(X^i) = \frac{1}{n_k - 1} \sum_{\substack{j=1 \\ j \neq i}}^{n_k} d(X^i, X^j)$$

$$b(X^i) = \min_{k' \neq k} \left\{ \frac{1}{n_{k'}} \sum_{j \in C_{k'}} d(X^i, X^j) \right\}.$$

Per ogni serie X^i, una volta calcolati $a(X^i)$ e $b(X^i)$ si calcola la larghezza Silhouette di ogni serie definita come:

$$s(X^i) = \frac{b(X^i) - a(X^i)}{max\left\{a(X^i), b(X^i)\right\}}.$$

La media delle larghezze Silhouette di ogni cluster è chiamata *Silhouette media*, e l'indice Silhouette globale è definito come la media delle larghezze Silhuoette di ogni cluster (Rousseeuw, 1987), ossia come

$$S = \frac{1}{K} \sum_{k=1}^{K} s_k,$$

con $s_k = \frac{1}{n_k} \sum_{i \in C_k} s(X^i)$.

Tale indice varia tra -1 e 1. È auspicabile che il valore assunto dall'indice sia quanto più possibile vicino a 1. In questo caso si ha che $a(X^i)$ è molto basso, ossia la coesione interna è molto elevata, mentre $b(X^i)$ è alto, ossia i cluster sono ben separati.

1.5.3 Indice di Caliński-Harabasz

L'indice di Caliński-Harabasz (Caliński e Harabasz, 1974) è chiamato anche criterio della varianza. L'obiettivo di tale indice è di minimizzare la varianza intra-cluster e di massimizzare la varianza entro-cluster così da raggiungere una situazione di cluster ben separati tra loro e ben coesi al loro interno. Esso è definito come:

$$CH_k = \frac{SS_B/(K-1)}{SS_W/(N-K)},$$

dove SS_B è la varianza intra-cluster complessiva, SS_W è la varianza entro-cluster, K è il numero di cluster e N è il numero totale delle osservazioni. Si ha quindi che

$$SS_B = \sum_{k=1}^{K} n_k \|\bar{X}^k - \bar{X}\|^2$$

e

$$SS_W = \sum_{k=1}^{K} \sum_{i \in C_k} \|X^k - \bar{X}^k\|^2$$

dove n_k è la numerosità del singolo cluster k, $\bar{X}^k$ è il baricentro del cluster k, $\bar{X}$ è il baricentro dei dati ed X_i^k sono le singole osservazioni dentro al cluster k.

Cluster ben definiti hanno una grande variazione tra cluster (SS_B) e una piccola varianza all'interno del cluster (SS_W). Maggiore è il rapporto CH_k, migliore è la partizione dati. Il numero ottimale di cluster è la soluzione con il più alto valore dell'indice Caliński-Harabasz.

Capitolo 2

Le misure di dissimilarità per serie storiche

In questo capitolo verranno presentate le misure di dissimilarità principali usate nell'ambito delle serie storiche. Un'ottima rassegna si trova nell'articolo di Montero e Vilar (2014)

Il metodo di raggruppamento usato in questo lavoro è il metodo del legame completo, e la misura di validità esterna per il confronto tra partizione vera e la partizione trovata è l'*indice di Rand corretto (ARI)*, perchè il più facile da interpretare. Il software utilizzato per le analisi è R (www.R-project.org).

2.1 Introduzione

Fare Cluster Analysis nell'ambito delle serie storiche è del tutto differente da fare Cluster Analysis quando si ha a che fare con dati indipendenti. Nel caso di dati indipendenti la procedura di clustering si basa su alcune variabili di interesse. Si può fare Cluster Analysis semplicemente calcolando la distanza euclidea tra esse. Nell'ambito delle serie storiche, invece, non è sufficiente calcolare la distanza euclidea tra i diversi valori delle serie storiche al variare del tempo. Questo perché, in questo modo, si stanno erroneamente trattando le osservazioni temporale come delle variabili indipendenti. Per poter fare Cluster Analysis tra serie storiche occorre introdurre delle misure appositamente studiate per questo ambito che tengano conto della dipendenza temporale intrinseca nei dati di serie storiche. Di seguito verranno proposte delle misure di dissimilarità che si basano s caratteristiche comuni alle serie quali, ad esempio, l'autocorrelazione o la diversità tra i parametri di modelli specificamente studiati per serie storiche,

Le misure presenti in letteratura spesso vengono divise in tre grandi gruppi: *approcci model-free*, ossia metodi che si basano solo su specifiche caratteristiche delle serie , *approcci model-based*, ossia metodi che si basano su assunzioni parametriche sulle serie storiche e *approcci prediction-based*, ossia metodi che si basano su previsioni di serie storiche. Nell'ultima parte del Capitolo vengono presentate nel dettaglio due misure di dissimilarità. La prima proposta in un recente articolo di Lafuente-Rego e Vilar (2015) e la secondo proposta per la prima volta in questo lavoro.

2.2 Approcci model-free

In questa sezione verranno descritte tre misure che si basano su approcci model-free. Tali misure non necessitano di assunzioni su eventuali modelli per serie storiche.

2.2.1 Distanza basata sulla correlazione

Il problema di stabilire relazioni di somiglianza fra serie storiche non è nuovo in statistica. Infatti, il concetto classico di correlazione è stato introdotto per caratterizzare processi che hanno un andamento simile. Dato un insieme di serie storiche, la matrice di correlazione o di varianza-covarianza misura, infatti, il grado di similarità fra le serie storiche. Un primo e semplice criterio di dissimilarità è quello di considerare il coefficiente di correlazione di Pearson tra la serie $X = \{X_1, \cdots, X_T\}$ e la serie $X' = \{X'_1, \cdots, X'_T\}$, dato da:

$$COR(X, X') = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(X'_t - \bar{X}')}{\sqrt{\sum_{t=1}^{T}(X_t - \bar{X})^2}\sqrt{\sum_{t=1}^{T}(X'_t - \bar{X}')^2}},$$

dove $\bar{X}$ e $\bar{X}'$ sono le medie delle realizzazioni delle serie X e X'. Si noti tuttavia che i coefficienti di correlazione non costituiscono una distanza perché possono assumere valori compresi da -1 a +1. E' stato mostrato (Golay *e altri*, 1998) come si possa costruire una distanza fra serie in modo molto semplice a partire dai consueti coefficienti di correlazione fra serie, definendo la distanza come:

$$d_{COR}(X, X') = \sqrt{2(1 - COR(X, X'))}.$$

Tale distanza ora gode delle proprietà tipiche di una misura di dissimilarità.

2.2.2 Distanza basata sull'autocorrelazione

Si considerano ora delle misure basate sulle funzioni di autocorrelazione stimate. Siano $\hat{\rho}_X = (\hat{\rho}_{1,X}, ..., \hat{\rho}_{L,X})'$ e $\hat{\rho}_{X'} = (\hat{\rho}_{1,X'}, ..., \hat{\rho}_{L,X'})'$ i vettori di autocorrelazione stimati di X e X' rispettivamente, per alcuni L tale che $\hat{\rho}_{i,X} \approx 0$ e $\hat{\rho}_{i,X'} \approx 0$ per $i > L$. Galeano e Peña (2001) definiscono una distanza tra X e X' come segue

$$d_{\Omega ACF}(X, X') = \sqrt{(\hat{\rho}_X - \hat{\rho}_{X'})' \, \Omega \, (\hat{\rho}_X - \hat{\rho}_{X'})}$$

dove Ω è una matrice di pesi. Scelte comuni di Ω sono:

- considerare pesi uniformi prendendo $\Omega = I$. In tal caso $d_{\Omega ACF}$ diventa la distanza euclidea tra le stime di autocorrelazione. È data da

$$d_{ACF}(X, X') = \sqrt{\sum_{i=1}^{L} (\hat{\rho}_{i,X} - \hat{\rho}_{i,X'})^2}. \tag{2.1}$$

- considerare pesi che decadono geometricamente con il ritardo di autocorrelazione, quindi $d_{\Omega ACF}$ assume la seguente forma:

$$d_{GACF}(X, X') = \sqrt{\sum_{i=1}^{L} \rho(1-\rho)^i \, (\hat{\rho}_X - \hat{\rho}_{X'})^2}.$$

Distanze analoghe possono essere costruite considerando le funzioni di autocorrelazione paziale invece che le funzioni di autocorrelazione globale. D'ora in avanti verrà considerata la distanza d_{ACF} che considera pesi uniformi.

2.3 Approcci model-based

In questa sezione verranno descritte due misure che si basano su un approccio model-based. Tali misure presuppongono la presenza di un modello statistico per la descrizione delle osservazioni.

2.3.1 Distanza di Piccolo

Piccolo (1990) definisce una misura di dissimilarità nella classe dei processi ARMA invertibili come la distanza euclidea tra i coefficienti della rappresentazione $AR(\infty)$ e delle corrispondenti misure $ARMA$. Piccolo sostiene

che i coefficienti autoregressivi trasmettono tutte le informazioni utili sulla struttura stocastica di questo tipo di processi. Nella pratica, si tronca la rappresentazione $AR(\infty)$ nei modelli di ordini k_1 e k_2 che approssimano i processi di generazione di X e X' rispettivamente. Questo approccio consente di superare il problema di ottenere un modello ARMA ad hoc per ciascuna serie sottoposta al clustering. Se $\hat{\Pi}_X = (\hat{\pi}_{1,X}, ..., \hat{\pi}_{k_1,X})'$ e $\hat{\Pi}_{X'} = (\hat{\pi}_{1,X'}, ..., \hat{\pi}_{k_2,X'})'$ denotano i vettori delle stime dei parametri $AR(k_1)$ e $AR(k_2)$ per X e X', rispettivamente, allora la distanza di Piccolo prende la forma:

$$d_{PIC}(X, X') = \sqrt{\sum_{i=1}^{k} \left(\hat{\pi}'_{i,X} - \hat{\pi}'_{i,X'}\right)^2}, \tag{2.2}$$

dove $k = \max\{k_1, k_2\}$, $\hat{\pi}'_{i,X} = \hat{\pi}_{i,X}$ se $j \leq k_1$ e $\hat{\pi}_{i,X} = 0$ altrimenti, in analogia $\hat{\pi}'_{i,X'} = \hat{\pi}_{i,X'}$ se $j \leq k_2$ e $\hat{\pi}_{i,X'} = 0$ altrimenti. La misura d_{PIC} esiste per un qualsiasi processo $ARMA$ invertibile, e soddisfa le proprietà di una misura di dissimilarità (non negatività, simmetria e triangolarità).

2.3.2 Distanza di Otranto

Quando si ha a che fare con serie storiche finanziarie una misura rilevante è data del rischio, che può essere misurato tramite la volatilità, e quindi tramite la varianza condizionale. Otranto (2008) propone un'estensione della misura di dissimilarità proposta da Piccolo che considera la distanza euclidea dei parametri di un processo $GARCH$.

Si consideri per la serie storica X, un modello autoregressivo a eteroschedasticità condizionata (o modello $GARCH$) è una funzione dei valori assunti dal processo negli istanti precedenti. Secondo una tipica formulazione, un processo $GARCH(p, q)$ è dato da:

$$X_t = \varepsilon_t = \sigma_t z_t, \qquad z_t \sim N(0, 1)$$

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2,$$

con $\omega > 0$, $0 \leq \alpha_i < 1$, $0 \leq \beta_j < 1$, $(i = 1, \cdots, p,\ j = 1, \cdots q)$, e $\left(\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j\right) < 1$.

Com'è noto, si può rappresentare la serie storica degli ε_t^2 come un processo $ARMA(p*, q)$, con $p* = max\{p, q\}$, ossia

$$\varepsilon_t^2 = \omega + \sum_{i=1}^{p*} \alpha_i \varepsilon_{t-i}^2 - \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \eta_t,$$

con $\eta_t = \varepsilon_t^2 - \sigma_t^2$, $\alpha_i = 0$ per $i > q$, e $\beta_i = 0$ per $i > p$. Si può facilmente vedere che η_t ha le proprietà di un white noise, almeno in senso debole. Infatti $\mathbb{E}[\eta_t] = 0$, e il processo η_t risulta non autocorrelato.

Una volta calcolato il modello $ARMA(p, q)$ per la serie degli ε_t è possibile ottenere la forma $AR(\infty)$ nel modo seguente:

$$\varepsilon_t^2 = \frac{\omega}{1 - \sum_{j=1}^{q} \beta_j} + \sum_{k=1}^{\infty} \pi_k^* \varepsilon_{t-k}^2 + \eta_t.$$

Dalla rappresentazione $AR(\infty)$ è immediato calcolare la distanza di Otranto come la distanza euclidea tra i coefficienti AR in analogia con la distanza di Piccolo in 2.2. Si ottiene dunque

$$d_{OTR}(X, X') = \sqrt{\sum_{i=1}^{k} \left(\hat{\pi}_{i,X}^{*'} - \hat{\pi}_{i,X'}^{*'} \right)^2}. \tag{2.3}$$

2.4 Approcci prediction-based

La dissomiglianza che si vuole descrivere in questa sezione si basa sulla previsione di serie storiche. Due serie storiche saranno considerate simili se la loro previsione, per un dato istante temporale, è vicina. Ovviamente, una procedura di clustering di questo tipo può produrre risultati differenti da quelli generati con metodi gerarchici model-free. Ad esempio, due serie storiche provenienti dalla stesso processo di generazione sono in grado di produrre differenti previsioni ad orizzonti pre-specificati, quindi queste serie storiche potrebbero non essere raggruppate assieme secondo questo nuovo criterio di dissimilarità. Ci sono molte situazioni pratiche in cui il vero interesse di raggruppamento si basa direttamente sulle previsioni e non sul modello sottostante. Alonso *e altri* (2006) hanno proposto una misura di dissimilarità basata sul confronto delle densità di previsione per ogni serie in un futuro istante temporale. Nella pratica, le densità di previsioni sono ignote e devono essere approssimate dai dati. Nella misura di dissimilarità proposta tali densità vengono approssimate con una procedura bootstrap combinata con tecniche di stima kernel. Questa procedura richiede di assumere che le serie storiche ammettano una rappresentazione $AR(1)$ perché il bootstrap si basa su un ricampionamento da approssimazioni autoregressive. Vilar *e altri* (2010) hanno esteso questa metodologia per coprire il caso di modelli non parametrici. In questo nuovo scenario, le densità di previsione sono approssimate considerando una procedura di bootstrap che imita i processi di generazione, senza assumere alcun modello parametrico per la vera struttura autoregressiva della serie.

In particolare siano X e X' realizzazioni di processi stazionari che ammettono una rappresentazione autoregressiva generale nella forma $S_t = \varphi(S_{t-1}) + \varepsilon_t$ con ε_t una sequenza IID e $\varphi(\cdot)$ una funzione regolare e non limitata a qualsiasi modello parametrico. Dato un particolare istante futuro $T + h$, Vilar introduce la seguente distanza tra X e X':

$$d_{PRED,h}(X, X') = \int \left| \hat{f}_{X_{T+h}}(u) - \hat{f}_{X'_{T+h}}(u) \right| du,$$

dove $\hat{f}_{X_{T+h}}(u)$ e $\hat{f}_{X'_{T+h}}(u)$ indicano le stime di densità di previsione all'orizzonte $T + h$ per X e X' rispettivamente. Le vere densità di previsione sono sostituite da stimatori kernel basati sulle previsoni bootstrap. Sebbene possano essere utilizzate differenti procedure di ricampionamento per ottenere le previsioni bootstrap , si considera una procedura basata sulla generazione di un processo

$$S^*_T = \hat{\varphi}_g(S^*_{t-1}) + \varepsilon^*_t,$$

dove $\hat{\varphi}_g$ è una stimatore non parametrico di φ e ε^*_t è un ricampionamento IID dei residui non parametrici. Tale metodo di ricampionamento utilizza un approccio simile a quello del ricampionamento residuo basato modelli autoregressivi lineari, ma si avvale di essere libero da assunzioni di linearità, e quindi può essere applicato ad una classe più ampia di modelli.

2.5 Due nuove misure di dissimilarità

In questa sezione vengono descritte due nuove misure di dissimilarità usate in questo lavoro.

2.5.1 Distanza basata sui quantili di autocovarianza

Questa misura di dissimilarità può essere inserita nelle misure di dissimilarità model-free dato che le informazioni contenute nei dati vengono riassunte con un ridotto numero di caratteristiche che descrivono la struttura temporale delle serie.

Molti autori hanno considerato misure basate sul confronto delle stime dell'autocorrelazione semplice o parziale. L'autocorrelazione mostra buone proprietà di discriminazione tra diversi tipi di processo, ma spesso presenta debolezze in presenza di outlier e code pesanti e non è in grado di rilevare la coda della dipendenza. Da notare che code pesanti e non esistenza di momenti alti sono caratteristiche frequenti nella distribuzione di molte serie finanziarie, (serie dei log-rendimenti, indici azionari, i prezzi delle azioni, tassi

di cambio, ecc). Per superare queste limitazioni, viene proposta (Lafuente-Rego e Vilar, 2015) una misura di dissimilarità che confronta le funzioni dei quantili di autocovarianza. Siano $X_1, \cdots, X_T$ le osservazioni di un processo stazionario e con $q_\tau = F^{-1}(\tau)$, $\tau \in [0,1]$, la funzione quantile tale per cui $\mathbb{E}[I(X_t \leq \tau)] = \mathbb{P}(X_t \leq q_\tau) = \tau$. Per un fissato $l \in \mathbb{Z}$ e un'arbitraria coppia di quantili di livello $(\tau, \tau') \in [0,1]^2$, si considera la covarianza delle funzioni indicatrici $I(X_t \leq q_\tau)$ e $I(X_{t+l} \leq q_{\tau'})$ data da

$$\begin{aligned} \gamma_l(\tau, \tau') &= cov(I(X_t \leq q_\tau), I(X_{t+l} \leq q_{\tau'})) \\ &= \mathbb{P}(X_t \leq q_\tau, X_{t+l} \leq q_{\tau'}) - \tau\tau', \end{aligned} \tag{2.4}$$

La funzione $\gamma_l(\tau, \tau')$, con $(\tau, \tau') \in [0,1]^2$, è chiamata *funzione quantile di autocovarianza a ritardo l*.

A differenza delle usuali covarianze e della autocorrelazioni, γ_l è definita in termini di indicatori di covarianza e non richiede condizioni sui momenti. Una dissimilarità basata sulla funzione dei quantili di autocovarianza può avere il vantaggio di discriminare tra serie generate da processi con code pesanti nella distribuzione marginale o che seguono modelli a eteroschedasticità condizionale. Da notare che $\gamma_l(\tau, \tau')$ riportata in (2.4), include i quantili q_τ e $q_{\tau'}$ della distribuzione marginale di F, che in pratica non è nota. Si sostituiscono quindi i quantili teorici con i corrispondenti quantili empirici $\hat{q}_\tau$ e $\hat{q}_{\tau'}$ basati sulle osservazioni $X_1, \cdots, X_T$, e si ottiene una stima di $\gamma_l(\tau, \tau')$ data da

$$\hat{\gamma}_l(\tau, \tau') = \frac{1}{T-l} \sum_{t=1}^{T-l} I(X_t \leq \hat{q}_\tau) I(X_{t+l} \leq \hat{q}_{\tau'}) - \tau\tau', \tag{2.5}$$

dove il quantile empirico $\hat{q}_\alpha$ per $0 \leq \alpha \leq 1$, può essere visto formalmente come soluzione di un problema di minimizzazione data da

$$\hat{q}_\alpha = \arg\min_{q \in \mathbb{R}} \sum_{t=1}^{T} \rho_\alpha \left(X_t - q\right),$$

con $\rho_\alpha(x) = x(\alpha - I(x < 0))$.

Per un dato set di L differenti ritardi temporali, $l_1 < l_2 < \cdots < l_L$ e r quantili di livello $0 < \tau_1 < \cdots < \tau_r < 1$, ogni serie storica $X^{(u)}$, $u = 1, 2$, è caratterizzata per mezzo del vettore $\Gamma(u)$ costruito come segue

$$\Gamma(u) = \left(\Gamma_{l_1}^{(u)}, \cdots, \Gamma_{l_L}^{(u)}\right),$$

dove ogni $\Gamma_{l_i}^{(u)}$ per $i = 1, \cdots, L$ è un vettore di lunghezza r^2 formato riorganizzando le righe degli elementi della matrice

$$\left(\hat{\gamma}_{l_i}^{(u)}(\tau_j, \tau_{j'});\ j, j' = 1, \cdots, r\right),$$

con $\hat{\gamma}_l$ data in (2.5). In questo modo la dissimilarità tra X_t^1 e X_t^2 è definita come la distanza euclidea tra $\Gamma^{(1)}$ e $\Gamma^{(2)}$. Siano $\hat{\gamma}_l^1$ e $\hat{\gamma}_l^2$ le funzioni di autocovarianza di X^1 e X^2 rispettivamente si ha

$$\begin{aligned} d_{QAF}(X, X') &= \|\Gamma^{(1)} - \Gamma^{(2)}\|^2 \\ &= \sum_{i=1}^{L} \sum_{j=1}^{r} \sum_{j'=1}^{r} (\hat{\gamma}_{l_i}^1(\tau_j, \tau_{j'}) - \hat{\gamma}_{l_i}^2(\tau_j, \tau_{j'}))^2. \end{aligned} \tag{2.6}$$

2.5.2 Distanza basata sul parametro di lisciamento

La misura di dissimilarità proposta in questo lavoro, si basa sull'idea che nell'ambito delle serie storiche spesso si possono stimare i modelli con un approccio classico, ossia considerando le serie storiche come somma di una parte deterministica data, ad esempio, da componente di trend e/o da componenti stagionali e da una parte erratica residuale. Quando ci si trova in questi contesti è utile stimare le componenti deterministiche o usando una procedura parametrica o usando una procedura non parametrica. Per quanto riguarda la misura di dissimilarità qui proposta si utilizzano tecniche non parametriche di stima e la misura si baserà sulla distanza euclidea tra i parametri di lisciamento scelti. Prima di introdurre la misura di dissimilarità è utile richiamare alcuni concetti di statistica non parametrica, in particolare sui modelli additivi non parametrici.

Quando si vuole superare l'ipotesi di linearità, molto frequente in statistica, una delle soluzioni che si possono attuare è quella considerare un modello non parametrico, con la seguente forma

$$\begin{aligned} y &= f(x) + \varepsilon \\ &= f(x_1, \cdots, x_p) + \varepsilon. \end{aligned} \tag{2.7}$$

Per poter ovviare al problema della maledizione della dimensionalità occorre introdurre una qualche forma di struttura per $f(x)$. Una possibile opzione è quella di considerare una rappresentazione additiva del tipo

$$f(x_1, \cdots, x_p) = \alpha + \sum_{j=1}^{p} f_j(x_j) \tag{2.8}$$

dove le funzioni $f_1, \cdots, f_p$ sono funzioni in una variabile con andamento sufficientemente regolare e α una costante. Si osservi che, per evitare quello che sostanzialmente è un problema di identificabilità del modello, occorre che le varie f_j siano centrate sullo 0, ossia

$$\sum_{i=1}^{n} f_j(x_{ij}) = 0, \qquad j = 1, \cdots, p.$$

La formulazione (2.8) per $f(x)$ costituisce un *modello additivo* che è più restrittivo del modello di regressione non parametrica generale dato in (2.7), ma meno restrittivo del modello di regressione lineare, che assume che tutte le funzioni di regressione parziale siano funzioni lineari. Ogni singola funzione $f_j(x_j)$ può essere stimata con metodi di regressione non parametrica univariati, come ad esempio la regressione locale o le splines. Non è cruciale quale metodo di stima non parametrica si utilizza, e addirittura si potrebbero scegliere metodi diversi per le diverse stime di f_j. Usualmente però si usa un unico metodo per la stima delle f_j .

Ai fini di trovare una misura di dissimilarità, è utile ricordare che le tecniche di regressione non parametrica più note necessitano della selezione di un parametro h detto *parametro di lisciamento* o *smoothers*. É quindi necessario riscrivere ogni singola la funzione $f_j(x_j)$ data in (2.8) come una funzione, oltre che della variabile x_j, anche del parametro di lisciamento h_j, ottenendo così

$$\begin{aligned} f(x,h) &= f(x_1,\cdots,x_p;\, h_1,\cdots h_p) \\ &= \alpha + \sum_{j=1}^{p} f_j(x_j;h_j). \end{aligned}$$

Nell'ambito delle serie storiche questo tipo di approccio può essere utilizzato per stimare le componenti deterministiche come il trend o eventuali componenti stagionali. Un modello generale può quindi essere il seguente:

$$X_t = f_1(T_t,h_1) + f_2(Y_t,h_2) + f_3(W_t,h_3) + C_t + \varepsilon_t,$$

dove

- X_t è la serie storica che si vuole stimare;
- T_t è la componente di lungo periodo;
- Y_t è la componente periodica annuale;
- W_t è la componente periodica settimanale;
- C_t rappresenta gli effetti di calendario;
- ε_t rappresenta l'errore del modello.

In questo modo, una volta stimati gli h ottimali per ogni funzione f_j, è immediato calcolare la dissimilarità basata sul parametro di lisciamento come

la distanza euclidea tra gli h stimati, ossia

$$d_{SM}(X, X') = \sqrt{\sum_{i=1}^{K} \left(\hat{h}_{i,X} - \hat{h}_{i,X'}\right)^2}, \tag{2.9}$$

con K numero di h presenti nel modello di riferimento, in questo caso $K = 3$.

2.6 Uno studio sulle misure d_{QAF} e d_{SM}

È stato condotto uno studio di simulazione per capire meglio come si comportano le due nuove misure di dissimilarità. Tale studio ha lo scopo di valutare le performance di d_{QAF} e d_{SM} in differenti contesti.

2.6.1 Studio delle performance di d_{QAF}

Allo scopo di valutare la misura di dissimilarità basata sui quantili di autocovarianza si effettuano delle simulazioni che prendono in considerazione differenti tipi di processi. Questi vengono divisi in tre categorie, e, all'interno di ogni categoria, vengono simulati diversi tipi di processo. Nel dettaglio si ha:

- **Scenario 1**: Modelli della classe $ARMA$
  - AR(1) $X_t = 0.9X_{t-1} + \varepsilon_t$
  - AR(2) $X_t = 0.3X_{t-1} - 0.1X_{t-2} + \varepsilon_t$
  - MA(1) $X_t = -0.7\varepsilon_{t-1} + \varepsilon_t$
  - MA(2) $X_t = 0.4\varepsilon_{t-1} - 0.2\varepsilon_{t-2} + \varepsilon_t$
  - ARMA(1,1) $X_t = 0.8X_{t-1} + 0.2\varepsilon_{t-1} + \varepsilon_t$
- **Scenario 2**: Modelli non lineari
  - TAR $X_t = 0.5X_{t-1}I(X_{t-1} \leq 0) - 2X_{t-1}I(X_{t-1} > 0) + \varepsilon_t$
  - EXPAR $X_t = [0.3 - 10\exp(-X_{t-1}^2)]X_{t-1} + \varepsilon_t$
  - MA $X_t = -0.4\varepsilon_{t-1} + \varepsilon_t$
  - NLMA $X_t = -0.5\varepsilon_{t-1} + 0.8\varepsilon_{t-1}^2 + \varepsilon_t$
- **Scenario 3**: Modelli a eteroschedasticità conidizonale. Si considera $X_t = \sigma_t\varepsilon_t$.
  - ARCH $\sigma_t^2 = 0.1 + 0.8X_{t-1}^2$
  - GARCH $\sigma_t^2 = 0.1 + 0.1X_{t-1}^2 + 0.8\sigma_{t-1}^2$
  - GJR-GARCH $\sigma_t^2 = 0.1 + [0.25 + 0.3I(X_{t-1} < 0)]X_{t-1}^2 + 0.5\sigma_{t-1}^2$
  - EGARCH $ln(\sigma_t^2) = 0.1 + 0.3\varepsilon_{t-1} + 0.3|\varepsilon_{t-1}| + 0.4\ln(\sigma_{t-1}^2)$

In tutti i casi si ha ε_t una sequenza di variabili casuali Gaussiane indipendenti a media zero e varianza unitaria. Per ogni tipologia di processo sono state generate $n = 5$ serie storiche di lunghezza $T = 365$. La scelta dei vari modelli è stata fatta con l'obiettivo di dare una classificazione generale, così da coprire molti campi di applicazione.

Per l'implementazione pratica della misura d_{QAF} è richiesto di fissare un numero di ritardi L, il numero di quantili e il loro relativo livello τ_i, per $i = 1, \cdots, r$. Di seguito viene effettuato uno studio di simulazione usando $L = 1, 2, 3$ e le sequenze di quantili seguenti

- $\tau_1 = (0.1, 0.9)$,
- $\tau_2 = (0.1, 0.5, 0.9)$,
- $\tau_3 = (0.1, 0.3, 0.5, 0.7, 0.9)$,
- $\tau_4 = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$.

Sono state effettuate $B = 100$ simulazioni all'interno di ogni Scenario e sono stati applicati gli indici di dissimilarità d_{QAF} con le caratteristiche appena riportate. Gli indici ARI medi con i relativi standard error sono riportati in Tabella 2.1

(a) Scenario 1

	τ_1	τ_2	τ_3	τ_4
L=1	0.50698	0.53455	0.53552	0.53989
	(0.0451)	(0.0687)	(0.0651)	(0.0664)
L=2	0.54739	0.64862	0.68509	0.68006
	(0.0629)	(0.1109)	(0.1216)	(0.1068)
L=3	0.56421	0.64459	0.65665	0.66980
	(0.0615)	(0.1053)	(0.1106)	(0.1273)

(b) Scenario 2

	τ_1	τ_2	τ_3	τ_4
L=1	0.94621	0.98908	0.99332	0.99576
	(0.1037)	(0.0411)	(0.0341)	(0.0242)
L=2	0.95793	0.95010	0.98409	0.97629
	(0.0870)	(0.1132)	(0.0605)	(0.0749)
L=3	0.94243	0.92904	0.96523	0.97944
	(0.0994)	(0.1304)	(0.0948)	(0.0746)

(c) Scenario 3

	τ_1	τ_2	τ_3	τ_4
L=1	0.46019	0.33226	0.30785	0.30785
	(0.1536)	(0.1249)	(0.1267)	(0.1253)
L=2	0.45043	0.31104	0.30353	0.29258
	(0.1480)	(0.1197)	(0.1235)	(0.1162)
L=3	0.43694	0.29712	0.27886	0.25706
	(0.1650)	(0.1120)	(0.1138)	(0.1157)

Tabella 2.1: Media e standard error (in parentesi) degli indici ARI in B=100 simulazioni per ogni scenario cambiando il numero di quantili e il ritardo considerato.

Grazie agli indici riportati in Tabella 2.1, si può concludere che non esiste una combinazione ottimale di L e di τ. Per quanto riguarda la scelta di L si può vedere che in due Scenari su tre si ha l'accordo maggiore in $L = 1$, mentre si ha l'accordo maggiore in $L = 2$ nello Scenario 1. Questo perché nel primo Scenario ci sono serie che considerano anche ritardi superiori al primo. Per quanto riguarda la scelta del numero di quantili r si può vedere che per

quanto riguarda lo Scenario 1 e lo Scenario 2 si ha l'accordo maggiore in τ_4, mentre nello Scenario 3 si ha l'accordo maggiore in τ_1. Questo porta alla conclusione che il numero di quantili da scegliere varia con la complessità delle serie con cui si ha a che fare, in particolare più aumenta la complessità più diminuiscono i quantili da prendere in considerazione. Bisogna considerare che a livello computazionale il calcolo di d_{QAF} con $r = 9$ quantili è molto oneroso e quindi è conveniente usare τ_3.

Una volta capito che ritardo usare e quanti quantili tenere in considerazione è utile chiedersi quali quantili riescono a spiegare meglio l'informazione contenuta nei dati. Non sempre infatti è conveniente usare quantili che considerino tutta la distribuzione. In molto ambiti, soprattutto in quello finanziario, molta dell'informazione sulla distribuzione è contenuta nelle code e non al centro della distribuzione. A questo proposito vengono ora considerati i seguenti quantili di livello:

- $\tau_{TUTTO} = (0.1, 0.3, 0.5, 0.7, 0.9)$,
- $\tau_{SX} = (0.05, 0.1, 0.15, 0.20, 0.25)$,
- $\tau_{DX} = (0.75, 0.80, 0.85, 0.90, 0.95)$,
- $\tau_{CODE} = (0.05, 0.1, 0.9, 0.95)$.

con $L = 1$. Vengono quindi usati $r = 4, 5$ quantili per costruire quattro sequenze di quantili, pensate per tenere conto di tutta la distribuzione, dell'informazione contenuta solo nella code sinistra, dell'informazione contenuta solo nella coda destra e dell'informazione contenuta nelle due code considerate congiuntamente. Anche in questo caso è stato condotto uno studio di simulazione con *B=100* replicazioni, e i risultati sono riportati in Tabella 2.2.

Dalla Tabella 2.2 si può vedere che, come nel caso della valutazione del numero di quantili da considerare, non c'è una scelta preferibile alle altre: non si può dire quali quantili sia meglio considerare, dipende molto dal tipo di serie storica in esame. Si nota però che considerare solo i quantili di una delle due code della distribuzione considerata singolarmente non è mai una scelta che porta ad avere accordi alti, è quindi preferibile considerare tutta al distribuzione o l'informazione contenuta in entrambe le code.

Di seguito quindi verranno usati $L = 1$, e i quantili di livello sono $\tau_{TUTTO} = (0.1, 0.3, 0.5, 0.7, 0.9)$ e $\tau_{CODE} = (0.05, 0.1, 0.9, 0.95)$.

2.6.2 Studio delle performance di d_{SM}

In questa sezione si vuole vedere come funziona la nuova misura di dissimilarità d_{SM} proposta in questo lavoro. Per poter testare la misura basata sul

(a) Scenario 1

	Media	St.Error
$d_{QAF\,TUTTO}$	0.53552	0.0650
$d_{QAF\,SX}$	0.52260	0.0619
$d_{QAF\,DX}$	0.55074	0.0794
$d_{QAF\,CODE}$	0.51182	0.0482

(b) Scenario 2

	Media	St.Error
$d_{QAF\,TUTTO}$	0.99332	0.0341
$d_{QAF\,SX}$	0.59811	0.0799
$d_{QAF\,DX}$	0.81322	0.1786
$d_{QAF\,CODE}$	0.96608	0.0850

(c) Scenario 3

	Media	St.Error
$d_{QAF\,TUTTO}$	0.33085	0.1267
$d_{QAF\,SX}$	0.26657	0.1112
$d_{QAF\,DX}$	0.19632	0.1203
$d_{QAF\,CODE}$	0.50112	0.1430

Tabella 2.2: Media e standard error dell'indice ARI in B=100 simulazioni per ogni scenario cambiando il livello dei quantili considerati.

parametro di lisciamento si considera un nuovo scenario che prende in considerazione serie storiche che al loro interno hanno una componente annuale e una componente settimanale. In particolare si considera:

- **Scenario 4**: Modelli con componenti deterministiche
  - $X_t = TREND_1 + STAG_1 + \eta_t$
  - $X_t = TREND_2 + STAG_2 + \eta_t$
  - $X_t = TREND_3 + STAG_3 + \eta_t$
  - $X_t = TREND_4 + STAG_4 + \eta_t$

dove η_t è una serie storica della classe dei modello $ARMA$, in particolare $\eta_t \sim ARMA(1,1)$ del tipo $\eta_t = 0.2\eta_{t-1} + 0.4\varepsilon_{t-1} + \varepsilon_t$ con $\varepsilon_t \sim N(0,1)\,\forall t$, e le componenti $TREND_i$, $STAG_i$ per $i = 1, \cdots, 4$ sono state stimate dalle serie dei prezzi di energia del mercato elettrico inglese. Anche in questo caso sono state simulate $n = 5$ serie storiche da ogni processo con lunghezza pari a $T = 365$.

Per calcolo di d_{SM} si regrediscono le variabili $T = \mathsf{Tempo}$ e $W = \mathsf{Settimana}$ sulla serie storica di interesse attraverso un modello additivo non parametrico del tipo

$$X_t = \alpha + f_1(T_t, h_1) + f_2(W_t, h_2) + \varepsilon_t,$$

dove entrambe le funzioni f_1 ed f_2 sono della funzioni *splines*. Per la scelta di h, in fase simulativa, si procederà in due passi: al primo passo si stimano

modelli del tipo

$$X_t = \alpha + f_1(T_t, h_1) + \varepsilon_t$$

e si sceglierà h_1 in maniera tale da rendere minimo

$$EQ_1 = \sum_{t=1}^{T} (T_t^* - \hat{T}_t)^2$$

dove il vettore $T^* = \{T_1^*, \cdots, T_T^*\}$ rappresenta il vettore contenente i veri valori del trend da cui è stata generata Y_t; al secondo passo, invece, si stima il modello basato sui residui del modello precedente $\hat{\varepsilon}_t$, ossia

$$\hat{\varepsilon}_t = f_2(W_t, h_2) + \eta_t$$

e, anche in questo caso, si sceglierà h_2 in modo tale da minimizzare

$$EQ_2 = \sum_{t=1}^{T} (W_t^* - \hat{W}_t)^2$$

dove $W^* = \{W_1^*, \cdots, W_T^*\}$ rappresenta il vettore dei veri valori della componente stagionale da cui è stata generata X. L'accordo calcolato con l'indice ARI ottenuto con la misura basata sui parametri di lisciamento su $B = 100$ simulazioni è pari a 0.92652 con uno standard error di 0.1229. Tale misura di dissimilarità risulta quindi avere buone proprietà di discriminazione di serie storiche che hanno al loro interno delle componenti deterministiche.

2.7 Conclusioni

In questo capitolo sono state introdotte alcune misure di dissimilarità specifiche per serie storiche. Nelle prima parte sono state presentate delle misure di dissimilarità già presenti in letteratura, mentre nella seconda parte sono state presentate due nuove misure di dissimilarità. La prima misura è stata introdotta da Lafuente-Rego e Vilar in un recente articolo del 2015, mentre la seconda misura non è mai stata proposta prima ed è stata creata con l'idea che le serie storiche spesso contengono al loro interno delle componenti deterministiche. Tali componenti possono essere modellizzate tramite approcci parametrici o tramite approcci non parametrici. Se si decide di modellizzarle tramite approcci non parametrici allora si avrà a che fare con dei parametri di lisciamento. La misura proposta si basa proprio sulla distanza euclidea tra i parametri di lisciamento.

Nell'ultima parte del Capitolo sono state studiate meglio queste due nuove misure di dissimilarità. Per farlo si è preso in considerazione uno studio

di simulazione che considera quattro Scenari tipici delle serie storiche. Lo Scenario 1 considera serie lineari, lo Scenario 2 considera serie non lineari e lo Scenario 3 considera serie con eteroschedasticità condizionata e lo Scenario 4 che considera serie storiche che hanno all'interno delle componenti deterministiche. In primo luogo si è valutato quali parametri è meglio usare per la misura di dissimilarità che si basa sui quantili di autocovarianza (il numero di ritardi, quanti e quali quantili usare) arrivando alla conclusione che non esiste una combinazione ottimale del numero di quantili e del numero di ritardi da considerare. La scelta dipende molto dal contesto applicativo. In secondo luogo sono state valutate le performance della nuova misura di dissimilarità basata sui parametri di lisciamento. Tale misura è in grado di discriminare bene i modelli che hanno componenti deterministiche diverse.

Nel prossimo Capitolo si vuole proporre una procedura di clustering supervisionata che ha lo scopo di dividere i dati in categorie, in maniera tale da poter applicare all'interno di ogni categoria, le misure di dissimilarità specifiche per serie storiche. Quest'idea nasce dal fatto che non esiste una msura di dissimilarità in grado di dividere serie storiche provenienti da processi molto differenti, ed è quindi necessario catalogare prima i dati in macrocategorie così da poter procedere con l'applicazione delle misure di dissimilarità.

Capitolo 3

Procedura di discriminazione

In questo Capitolo si propone una procedura in grado di discriminare serie storiche provenienti da processi molto diversi. Tale procedura non è mai stata proposta prima.

3.1 Introduzione e obiettivi

La procedura di discriminazione proposta in questo Capitolo nasce dai seguenti due presupposti:

1. nell'ambito delle serie storiche, a differenza dell'ambito che considera dati indipendenti, è possibile dare delle etichette alle serie considerate in base al processo di generazione sottostante;

2. non esiste una misura di dissimilarità per serie storiche che sia in grado di dividere serie provenienti da processi molto differenti. Ogni misura di dissimilarità ipotizza che le serie storiche da dividere abbiano caratteristiche simili, come, ad esempio, la provenienza da processi di tipo $ARMA$ o di tipo $GARCH$.

L'obiettivo di questo Capitolo è di proporre una tecnica che sia in grado di ottenere una partizione quanto più generale possibile, dividendo le serie storiche in base a caratteristiche generali. Si è pensato, quindi, di proporre una procedura strutturata in più passi che permetta la suddivisione e l'etichettatura di serie storiche provenienti da processi molto differenti. Tutti i passi della procedura sono fatti con lo scopo di osservare caratteristiche sempre meno generali. In particolare, gli obiettivi che si vogliono raggiungere sono:

- divisione delle serie storiche in stazionarie, trend stazionarie e con radici unitarie;

- individuazione, nel gruppo delle serie stazionarie di serie di tipo White Noise, serie di tipo ARMA, serie con correlazione non lineare nei livelli e serie di tipo GARCH.

Per raggiungere gli obiettivi prefissati in ogni passo non si utilizzano misure di dissimilarità: tali misure, oltre a considerare caratteristiche troppo specifiche delle serie storiche, non sono in grado di etichettare le partizioni ottenute. Inoltre, le tecniche che considerano una matrice delle distanze dividono sempre i dati in $K > 1$ gruppi, anche se i dati non necessitano di una divisione. Nel caso in cui le serie storiche siano generate dallo stesso processo con le misure di dissimilarità si avrà comunque una divisione in gruppi mentre, con la procedura proposta che non utilizza misure di dissimilarità, non si avrà alcuna divisione.

Una volta terminata la procedura si possono effettuare due scelte in base agli obiettivi delle analisi. Se si vuole ottenere un partizione ancora più fine si può procedere con l'applicazione delle misure di dissimilarità specifiche per ogni gruppo trovato. Per ogni gruppo identificato dalla procedura esistono delle misure di dissimilarità che garantiscono una divisione in gruppi ottimale. Se invece si ha a che fare con serie storiche molto differenti tra loro la sola procedura può essere sufficiente per ottenere una buona discriminazione.

3.2 Descrizione dei passi della procedura

Per poter raggiungere gli obiettivi spiegati precedentemente si procede in 4 passi e ad ogni passo si utilizzano tecniche diverse. In Figura 3.1 vengono riportati schematicamente i passi che si è scelto di seguire. Come si può notare dalla Figura 3.1 per raggiungere il primo obiettivo della procedura, ossia quello di dividere le serie storiche in stazionarie, con componenti deterministiche e a radici unitarie, si è dovuto procedere in due passi. Questo perché in statistica non esiste un test per la stazionarietà in senso stretto ma esistono una serie di test che verificano la presenza o meno di radici unitarie. Le tecniche usate in ogni passo sono le seguenti:

Primo passo. Al primo passo della procedura si applica un test per le radici unitarie a tutte le serie in esame. In particolare, si applica il test di Dickey-Fuller per verificare l'ipotesi $H_0 : \rho = 1$ contro l'alternativa $H_1 : \rho < 1$ considerando la seguente regressione:

$$X_t = \beta_0 + \rho X_{t-1} + \varepsilon_t.$$

Sotto H_0, il processo considerato per X è un random walk senza drift mentre, sotto H_1, si ha un processo stazionario con eventuale media

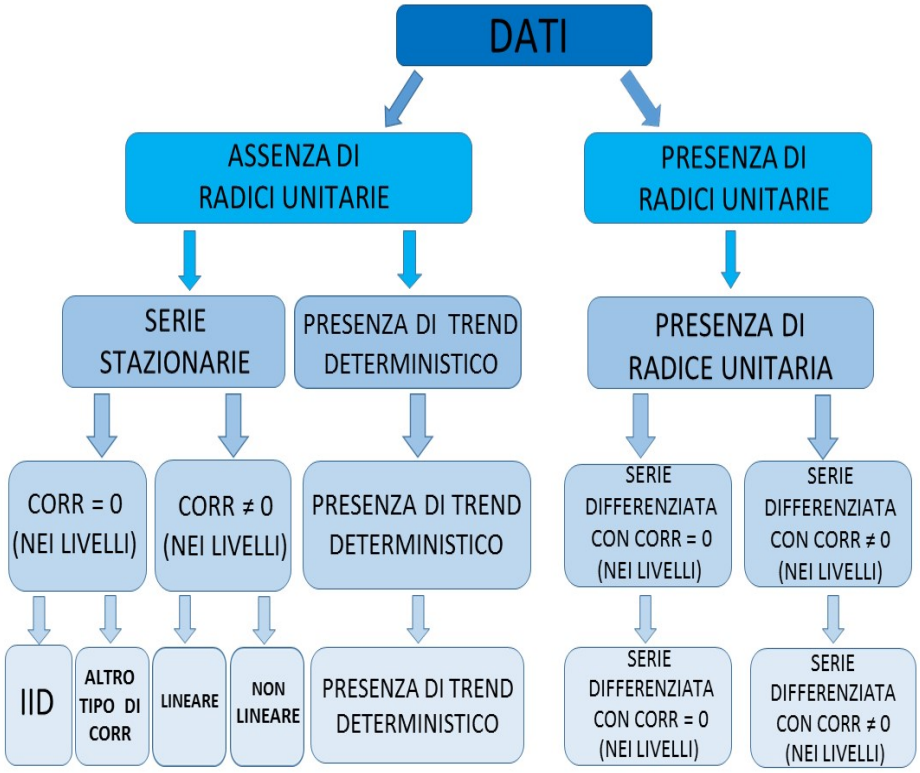


Figura 3.1: Schematizzazione dei passi utilizzati nella procedura di discriminazione.

diversa da zero. La statistica test calcolata con il metodo dei minimi quadrati è

$$t_c = \frac{\hat{\rho} - 1}{se(\hat{\rho})}, \tag{3.1}$$

che, sotto H_0, risulta avere una distribuzione asintotica non standard, tabulata dagli stessi Dickey e Fuller (1979). Si considerano, quindi, serie storiche con radici unitarie tutte le serie storiche che forniscono un p-value relativo al test t_c riportato in (3.1) maggiore o uguale del livello di significatività $\alpha = 0.05$.

Secondo passo. Al secondo passo ci si concentra esclusivamente sul gruppo delle serie senza radici unitarie. L'interesse è di capire se in tale gruppo ci sono delle serie che hanno al loro interno una componente di trend. Per poter verificare questo si utilizza un modello additivo non

parametrico del tipo:

$$X_t = \beta_0 + f_1(T_t) + f_2(X_{t-1}) + \varepsilon_t$$

dove X_t rappresenta la serie in esame e T_t rappresenta la componente di trend ed ε_t la componente erratica del modello. Una volta stimato il modello usando splines per la stima di $f_1(\cdot)$ e di $f_2(\cdot)$ si effettua un test ANOVA per effetti parametrici per verificare la significatività della funzione $f_1(\cdot)$. Le serie storiche considerate in questo passo della procedura vengono considerate con trend deterministico se il p-value della statistica F del test ANOVA accetta l'ipotesi di significatività della componente $f_1(T_t)$ ad un livello $\alpha = 0.05$. Si è deciso di utilizzare un modello non parametrico per cercare di rimanere in un contesto il più generale possibile.

Terzo passo. Al terzo passo della procedura ci si concentra solo sul gruppo delle serie identificate come stazionarie e sulle differenze prime delle serie appartenenti al gruppo delle radici unitarie. Il gruppo identificato come serie con componenti deterministiche non verrà più considerato: tale informazione è già sufficiente per applicare la misura di dissimilarità d_{SM} proposta nel Capitolo 2.

In questo passo si valuta la correlazione presente nei livelli. Sia ρ_i la funzione di autocorrelazione relativa al ritardo i calcolata su X. Si considerano serie storiche a correlazione nulla le serie storiche che soddisfano queste due condizioni sull'autocorrelazione:

$$|\rho_1| \leq \frac{z_{1-\alpha/2}}{\sqrt{T}}, \qquad |\rho_s| \leq \frac{z_{1-\alpha/2}}{\sqrt{T}}$$

dove s rappresenta il ritardo stagionale e $z_{1-\alpha/2}$ rappresenta il quantile di livello $(1-\alpha/2)$ della distribuzione Normale Standard e T rappresenta la lunghezza della serie storica X. Il livello di significatività considerato è $\alpha = 0.05$ e la scelta di s dipende dalle serie storiche che si stanno considerando. Le serie storiche che non soddisfano almeno una delle due condizioni su ρ_i vengono considerate a correlazione non nulla sui livelli.

Quarto Passo Nell'ultimo passo della procedura ci si concentra solo sul gruppo delle serie stazionarie a correlazione nulla e non nulla sui livelli. Non vengono più le serie storiche con radici unitarie per non andare troppo nel dettaglio: si può applicare la procedura alle serie storiche differenziate se nei dati in esame ci sono tante serie con radici unitarie.

Nel gruppo delle serie a correlazione non nulla si vuole vedere se la correlazione presente nella media è di tipo lineare o non lineare. Per far ciò si stima un modello autoregressivo non parametrico del tipo:

$$X_t = \beta_0 + f_1(X_{t-1}) + \cdots + f_k(X_{t-k}) + \varepsilon_t. \tag{3.2}$$

Una volta stimato il modello in (3.2) si effettua un test ANOVA per valutare la presenza degli effetti non parametrici. Si considerano serie storiche con componenti non lineari nella media le serie che hanno tutti i test F relativi alle funzioni $f_j(\cdot)$ che accettano l'ipotesi di significatività della componente non parametrica per $j = 1, \cdots, k$ a livello $\alpha = 0.05$. In questo lavoro si fissa il numero di ritardi pari a 2 per semplicità.

Nel gruppo delle serie storiche a correlazione nulla nei livelli si vuole vedere se l'eventuale dipendenza è da imputarsi alla varianza o se si stanno trattando serie incorrelate, tipo White Noise. La tecnica utilizzata è la stessa usata al passo precedente per dividere le serie storiche tra serie con correlazione nulla e serie con correlazione non nulla, con la differenza che la correlazione non viene calcolata sui livelli della serie storica ma sui quadrati.

3.3 Studio di simulazione

Per testare le performance della procedura di discriminazione proposta si effettua una studio di simulazione. Sono state generate serie storiche con caratteristiche molto differenti in maniera tale da avere delle serie rappresentative per ogni gruppo finale della partizione al quale arriva la procedura proposta. Gli scenari qui considerati sono diversi dagli scenari pensati per il Capitolo 2. Questo perché ora l'obiettivo è più generale e richiede l'applicazione su un set di serie storiche più grande e più complesso. Le serie storiche simulate sono state divise in due grandi gruppi. Ogni gruppo è stato diviso in altri due sotto gruppi per rendere più chiaro l'obiettivo che si vuole raggiungere. In particolare si ha:

- **Scenario 1: Serie stazionarie**
  - ∗ **Scenario 1.1:** Serie a correlazione non nulla sui livelli
    - AR $X_t = 0.4X_{t-1} + \varepsilon_t$
    - SAR $X_t = -0.3X_{t-1} + 0.5X_{t-7} + \varepsilon_t$
    - ARMA $X_t = 0.5X_{t-1} + 0.4\varepsilon_{t-1} + \varepsilon_t$
    - LIN-QUAD $X_t = -0.5X_{t-1} + 0.1X_{t-1}^2 + \varepsilon_t$
    - LOG $X_t = -0.7X_{t-1} + 0.3\log(X_{t-1}^2) + \varepsilon_t$
  - ∗ **Scenario 1.2:** Serie a correlazione nulla sui livelli
    - WHITE NOISE $X_t \sim N(0,1)$
    - GARCH $X_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = 0.3X_{t-1} + 0.5\sigma_{t-1}^2$
    - EGARCH $X_t = \sigma_t\varepsilon_t,$
      $$ln(\sigma_t^2) = 0.1 + 0.3X_{t-1} + 0.3|X_{t-1}| + 0.4\ln(\sigma_{t-1}^2)$$
- **Scenario 2: Serie non stazionarie**
  - ∗ **Scenario 2.1:** Serie con componenti deterministiche
    - $X_t = TREND_1 + STAG_1 + \eta_t,\ \eta_t = 0.2\eta_{t-1} + 0.4\varepsilon_{t-1} + \varepsilon_t$
    - $X_t = TREND_2 + STAG_2 + \eta_t,\ \eta_t = 0.2\eta_{t-1} + 0.4\varepsilon_{t-1} + \varepsilon_t$
    - $X_t = TREND_3 + STAG_3 + \eta_t,\ \eta_t = 0.2\eta_{t-1} + 0.4\varepsilon_{t-1} + \varepsilon_t$
  - ∗ **Scenario 2.2:** Serie con radice unitaria
    - ARI $(1 - 0.5L)(1 - L)X_t = \varepsilon_t$
    - ARIMA $(1 - 0.3L)(1 - L)X_t = (1 - 0.5L)\varepsilon_t$
    - I-GARCH $(1 - L)X_t = \varepsilon_t,\ \sigma_t^2 = 0.1\varepsilon_{t-1} + 0.6\sigma_{t-1}^2$

dove $\{\varepsilon_t\}_{t=1,\cdots,T}$ è una sequenza di variabili indipendenti e identicamente distribuite provenienti da una distribuzione Normale Standard e L indica l'operatore ritardo. Le serie denominate con $LIN-QUAD$ e LOG sono state appositamente create per avere una correlazione con X_{t-1} e una componente non lineare per X_{t-1}. Le componenti $TREND_i$ e $STAG_i$ per $i = 1, 2, 3$ sono state stimate nei dati relativi ai prezzi di energia elettrica del mercato elettrico inglese. Per ogni serie storica sono state stimate $n = 5$ serie storiche di lunghezza $T = 365$. In totale si hanno quindi $N = 70$ serie storiche. La procedura viene applicata in $B = 100$ simulazioni e gli indici ARI medi con i relativi standard error trovati vengono riportati in Tabella 3.1.

Come si può notare dalla Tabella 3.1 gli accordi, inevitabilmente, si abbassano ad ogni passo. Questo è un punto a svafore della procedura di discriminazione qui proposta: ad ogni passo il livello di significatività fissato

	Media	St. Error
Primo Passo	0.94475	0.0516
Secondo Passo	0.84175	0.0727
Terzo Passo	0.67921	0.0732
Quarto Passo	0.30741	0.0235

Tabella 3.1: Media e standard error dell'indice ARI in B=100 simulazioni applicando la procedura di discriminazione.

è di $\alpha = 0.05$ ma tale livello non può essere controllato ad ogni passo. Infatti, più si procede con la procedura più si commettono errori che non si riescono a controllare. Il punto dove le procedura sbaglia di più è l'individuazione delle serie che hanno solo componenti lineari e dalle serie che hanno anche componenti non lineari. Applicando questo passo della procedura solo alla parte del dataset che contiene le serie storiche dello Scenario 1.1 si ha un indice di accordo ARI medio in $B = 100$ simulazioni pari a 0.59893 ($\pm$ 0.2085). Se si va a guardare nel dettaglio si nota che gli errori più frequenti sono nelle 5 serie storiche provenienti dal processo ARMA. Tali serie storiche vengono spesso classificate come non lineari.

3.4 Conclusioni

In questo Capitolo è stata proposta una nuova procedura in grado di discriminare serie storiche provenienti da processi con caratteristiche molto diverse tra loro. Questa procedura è stata proposta per diversi motivi. Innanzitutto perché in letteratura non esistono misure di dissimilarità in grado di dividere serie storiche provenienti da processi molto differenti tra loro. Ogni misura di dissimilarità si basa su caratteristiche molto precise e non esiste una misura che guarda la serie storica nella sua totalità. Inoltre, nell'ambito di serie storiche, a differenza dell'ambito che considera dati indipendenti, c'è la possibilità di avere un'etichettatura dei gruppi trovati che si basa sul processo di generazione sottostante, che tale procedura è in grado di fornire.

La procedura proposta si compone di 4 passi ognuno dei quali con obiettivi differenti. L'obiettivo finale è quello di trovare un partizione che permetta l'applicazione delle misure di dissimilarità tipiche per le serie storiche in gruppi che siano omogenei al loro interno. Nei primi due passi l'obiettivo è quello di dividere le serie storiche in stazionarie, con componenti deterministiche e con radici unitarie. Per arrivare a questa divisione sono necessari due passi perché in statistica non esiste un test in grado di dare la divisione richiesta. È quindi necessario prima dividere le serie storiche che hanno radice unitaria

e, in un secondo momento, dividere le serie stazionarie dalle serie che contengono componenti deterministiche. Al terzo passo l'obiettivo è quello di dividere le serie storiche stazionarie e le differenze prime delle serie storiche con radici unitarie in base alla presenza o meno di correlazione nei livelli. Nell'ultimo passo della procedura si vuole capire se l'eventuale correlazione trovata nei livelli è di tipo lineare o non lineare e, nel gruppo delle serie senza correlazione nei livelli, si vuole capire se c'è una correlazione nei quadrati o se si ha a che fare con serie di tipo White Noise. Quest'ultimo passo delle procedura viene applicato, per semplicità, solo al gruppo delle serie storiche stazionarie e non al gruppo delle serie storiche che contengono radice unitaria. Tale passo però può essere applicato anche alle serie storiche con radice unitarie opportunamente differenziate.

La procedura è stata testata grazie ad uno studio di simulazione. Sono state generate $N = 70$ serie storiche proveniente da processi con caratteristiche diverse. L'indice ARI calcolato nell'ultimo passo della procedura in $B = 100$ simulazioni è circa 0.31 ($\pm$ 0.02). Tale accordo non risulta molto alto: ad ogni passo si commettono degli errori, e tali errori non si riescono a controllare soprattutto se si procede troppo con la procedura. L'indice ARI ottenuto al terzo passo è di circa 0.68 ($\pm$ 0.07) che risulta un accordo accettabile.

Nel prossimo Capitolo verranno applicate le due misure di dissimilarità proposte in questo lavoro su dei dati reali. Non verrà applicata la procedura proposta in questo Capitolo perché i dati considerati riguardano i prezzi di energia del mercato elettrico inglese ed è noto che tali serie sono caratterizzate da componenti deterministiche.

Capitolo 4

Applicazione al caso reale: prezzi dell'energia del mercato elettrico inglese

In questo Capitolo si vuole vedere come funzionano alcune misure di dissimilarità su dati reali. In particolare si fa riferimento alle 48 serie semi-orarie dei prezzi di energia elettrica del mercato elettrico inglese. Il periodo considerato va dal 01 Luglio 2009 al 30 Giugno 2014.

4.1 Scelta delle misure di dissimilarità

L'analisi delle serie storiche può essere fatta secondo due approcci. Il primo approccio, quello classico, assume che il processo abbia una parte deterministica, che consente la scomposizione del processo in componenti tendenziali, cicliche e/o stagionali, e che la differenza tra i dati teorici del modello deterministico ed i dati osservati sia attribuibile ad una componente casuale residuale. Il secondo approccio, quello moderno, assume che il processo sia stato generato da un processo stocastico descrivibile mediante un modello probabilistico di tipo parametrico.

Le serie storiche dei prezzi del mercato elettrico inglese sono caratterizzate da componenti deterministiche. Per questo motivo si decide di modellizzarle tramite l'approccio classico. In particolare, si considera un modello additivo non parametrico del tipo

$$X_t = f_1(T_t; h_1) + f_2(Y_t; h_2) + \varepsilon_t, \qquad t = 1, \cdots, T \tag{4.1}$$

dove X_t rappresenta una delle 48 serie semi orarie dei prezzi di energia, T_t rappresenta la componente di lungo periodo, Y_t rappresenta la componente annuale e ε_t rappresenta la componente erratica.

Per poter fare della Cluster Analysis con questo tipo di serie storiche si può considerare la misura basata sul parametro di lisciamento proposta in questo lavoro. Tale misura, infatti, è appositamente studiata per questo tipo di serie storiche.

Di seguito, i risultati ottenuti dalla Cluster Analysis fatta usando la nuova misura di dissimilarità verranno confrontati con i risultati ottenuti usando la misura basata sui quantili di autocovarianza. Tale misure si adatta bene al caso in esame perché considera le informazioni contenute nella distribuzione delle serie storiche. Non verranno usate le altre misure di dissimilarità presenti in questo lavoro perché non si adattano bene, per costruzione, a questo tipo di serie storiche.

4.1.1 Dissimilarità d_{SM}

Per quanto riguarda la scelta dei parametri di lisciamento del modello non parametrico in (4.1) si decide di usare il criterio informativo di Akaike corretto in maniera da penalizzare la complessità del modello (Hurvich *e altri*, 1998).

Le partizioni ottenute con questa misura di dissimilarità al variare del numero di gruppi considerato sono riportate in Figura 4.1 e in Tabella 4.1 vengono riportate nel dettaglio le serie storiche che rientrano nella partizione considerata ottimale dell'indice di Dunn.

C_1	C_2	C_3
$n_1 = 26$	$n_2 = 16$	$n_3 = 6$
19 - 19.30 - 20 - 20.30		
21 - 21.30 - 22 - 22.30	8 - 8.30 - 9 - 9.30	
23 - 23.30 - 24 - 00.30	10 - 10.30 - 11 - 11.30	16 - 16.30 - 17 - 17.30
1 - 1.30 - 2 - 2.30 - 3	12 - 12.30 - 13 - 13.30	18 - 1.30
3.30 - 4 - 4.30 - 5	14 - 14.30 - 15 - 15.30	
5.30 - 6 - 6.30 - 7 - 7.30		

Tabella 4.1: Dettaglio delle serie storiche relative alle serie semi orarie dei prezzi dell'energia del mercato elettrico inglese che rientrano in ogni cluster considerato della misura d_{SM}.

Dalle partizioni riportate in Figura 4.1 si vede che la giornata viene divisa in fasce orarie contigue. Tale caratteristica è ragionevole e facilita di molto

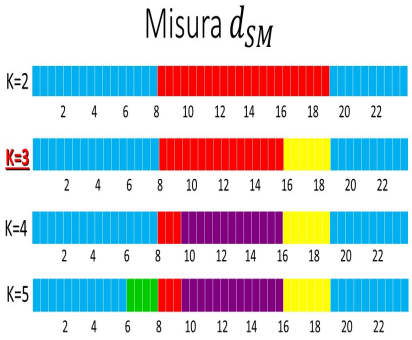


Figura 4.1: Partizioni ottenute per le 48 serie semi orarie dei prezzi dell'energia nel mercato elettrico inglese usando la misura d_{SM}. Ogni riga rappresenta le varie ore della giornata e ogni colore rappresenta i diversi gruppi ottenuti. In rosso è riportato il numero di gruppi ottimale secondo l'indice di Dunn.

l'interpretabilità dei risultati. Nel caso $K = 2$ la giornata viene divisa in maniera tale da separare le ora centrali della giornata dalla ore che riguardano la sera, la notte e le prime ora del mattino. L'indice di Dunn suggerisce di considerare una partizione più fine che considera tre gruppi. Tale partizione risulta avere gruppi meglio separati tra loro e meglio coesi al loro interno. In particolare, si divide la giornata nel modo seguente:

- fascia notturna (19.00-8.00);
- fascia giornaliera (8.00-16.00);
- fascia serale (16.00-19.00).

In Figura 4.2 vengono riportate tre serie storiche rappresentative di ogni fascia considerata da tale misura di dissimilarità.

Dalla Figura 4.2 si vede che le serie storiche all'interno di ogni fascia hanno un andamento differente: si nota un andamento oscillatorio nelle fasce del mattino senza forti picchi, si nota la presenza di alcuni picchi nelle ore centrali della giornata e la presenza di picchi molto elevati nelle ore serali.

Se si vuole considerare un numero di gruppi maggiore di 3, la misura di dissimilarità basata sul parametro di lisciamento suggerisce di dividere ulteriormente la fascia relativa alle ore centrali della giornata. In un secondo momento, tale misura suggerisce di dividere le prime ore del mattino dalle ore che riguardano la notte.

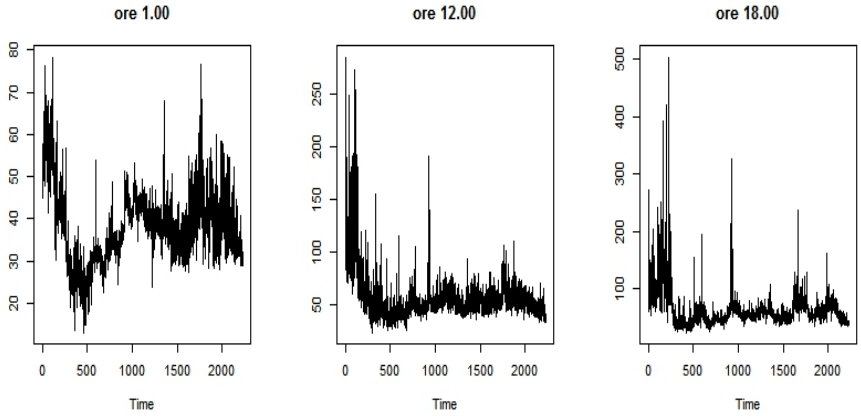


Figura 4.2: Serie storiche rappresentative dei tre gruppi ottenuti con la misura di dissimilarità d_{SM}.

4.1.2 Dissimilarità d_{QAF}

In questa sezione si vuole fare della Cluster Analysis considerando la misura basata sui quantili di autocovarianza. Tale misura considera le informazioni contenute nella distribuzione senza fare nessun tipo di ipotesi su possibili modelli da applicare ai dati. Per il calcolo della misura d_{QAF} si considerano i seguenti livelli di probabilità:

- $\tau_{TUTTO} = (0.1, 0.3, 0.5, 0.7, 0.9)$,
- $\tau_{CODE} = (0.05, 0.1, 0.9, 0.95)$,
- $\tau_{DX} = (0.75, 0.80, 0.85, 0.90, 0.95)$.

Si utilizzano quindi i quantili che considerano l'informazione contenuta in tutta la distribuzione dei prezzi di energia, l'informazione contenuta in entrambe le code e l'informazione contenuta solo nella coda destra. Si è deciso di includere anche i quantili della coda destra perché le serie storiche considerate sono relative a dei prezzi e le informazioni più rilevanti sono contenute proprio nella parte destra della distribuzione. Il parametro L è posto pari a 1.

Le partizioni ottenute con i vari quantili e con diverso numero di gruppi sono riportate in Figura 4.3,e in Tabella 4.2 vengono riportate nel dettaglio le serie che rientrano nelle partizioni ottenute dalle tre misure basate sui quantili di autocovarianza che risultano migliori in base all'indice di Dunn.

Dalle partizioni relative alle misure di dissimilarità basate sui quantili di autocovarianza riportate in Figura 4.3 si vede che i gruppi ottenuti cambiano

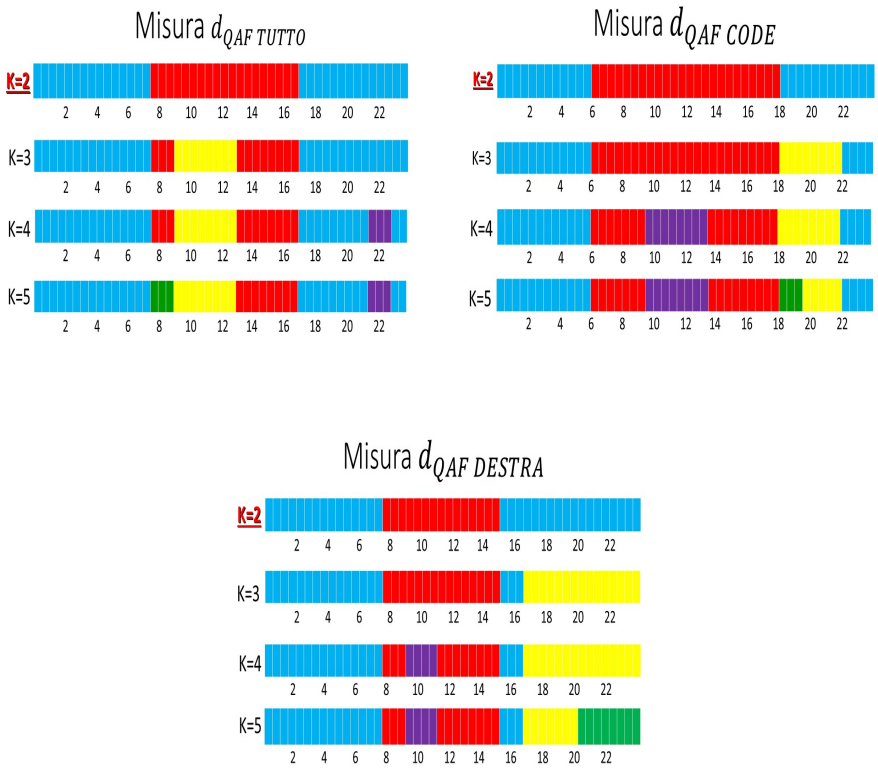


Figura 4.3: Partizioni ottenute nei dati relativi alle 48 serie semi orarie dei prezzi dell'energia nel mercato elettrico inglese con la misura d_{QAF}. Ogni riga rappresenta le varie ore della giornata e ogni colore rappresenta i diversi gruppi ottenuti. In rosso è riportato il numero di gruppi ottimale secondo l'indice di Dunn.

in base ai quantili utilizzati. L'indice di Dunn, calcolato per ogni misura, suggerisce la divisione della giornata in soli due gruppi. Tale divisione, secondo l'indice scelto, risulta quella con gruppi meglio separati tra loro e con maggior coesione interna. Tutte e tre le partizioni che considerano $K = 2$ concordano nel dividere la giornata in:

- fascia che comprende le ore centrali della giornata;
- fascia che comprede le ore serali, notturne e mattiniere.

La differenza delle tre partizioni sta nelle ore che vengono considerate per ogni fascia oraria. In particolare, la misura che considera tutta l'informazione della distribuzione considera come ore centrali le ore della fascia 07.00-17.00,

(a) Misura $d_{QAFTUTTO}$

C_1	C_2
$n_1 = 29$	$n_2 = 19$
17 - 17.30 - 18 - 18.30 - 19 - 19.30	7.30 - 8 - 8.30 - 9 - 9.30 - 10
20 - 20.30 - 21 - 21.30 - 22 - 22.30	10.30 - 11 - 11.30 - 12 - 12.30
23 - 23.30 - 24 - 00.30- 1 - 1.30	13 - 13.30 - 14 - 14.30 - 15
2 - 2.30 - 3 - 3.30 - 4 -4.30 - 5	15.30 - 16 - 16.30
5.30 - 6 - 6.30 - 7	

(b) Misura $d_{QAFCODE}$

C_1	C_2
$n_1 = 24$	$n_2 = 24$
18 - 18.30 - 19 - 19.30 - 20 - 20.30	6 - 6.30 - 7 - 7.30 - 8 - 8.30 - 9
21 - 21.30 - 22 - 22.30 - 23 - 23.30	9.30 - 10 - 10.30 - 11 - 11.30
24 - 00.30 - 1 - 1.30 - 2 - 2.30 - 3	12 - 12.30 - 13 - 13.30 - 14
3.30 - 4 - 4.30 - 5 - 5.30	14.30 - 15 - 15.30 - 16
	16.30 - 17 - 17.30

(c) Misura d_{QAFDX}

C_1	C_2
$n_1 = 34$	$n_2 = 14$
15 - 15.30 - 16 - 16.30 - 17 - 17.30	8 - 8.30 - 9 - 9.30 - 10 - 10.30
18 - 18.30 - 19 - 19.30 - 20 - 20.30	11 - 11.30 - 12 - 12.30 - 13
21 - 21.30 - 22 - 22.30 - 23 - 23.30	13.30 - 14 - 14.30
24 - 00.30 - 1 - 1.30 - 2 - 2.30 - 3	
3.30 - 4 - 4.30 - 5 - 5.30 - 6	
6.30 - 7 - 7.30	

Tabella 4.2: Dettaglio delle serie storiche relative alle serie semi orarie dei prezzi dell'energia del mercato elettrico inglese che rientrano in ogni cluster considerato della misura d_{QAF} con vari quantili di livello.

la misura che utilizza l'informazione delle code considera come fascia delle ore centrali la fascia 06.00-18.00, mentre la misura che utilizza solo i quantili della coda destra considera come ore centrali della giornata una fascia oraria più ristretta che va dalle ore 7.30 alle ore 15.00.

All'aumentare del numero di gruppi considerati, ogni misura di dissimilarità divide in maniera diversa le fasce orarie. In particolare, la misura che considera i quantili di tutta la distribuzione divide la fascia centrale in sottoclassi scomponendo la prima parte delle ore dalla mattina dalle ore del tardo pomeriggio, mentre le misure che considerano i quantili delle code del-

la distribuzione dividono prima la fascia serale e, in un secondo momento, la fascia centrale della giornata. La divisione con $K = 3$ gruppi risulta, in base all'indice di Dunn, peggiore rispetto alla divisione con $K = 2$ gruppi in termini di separatezza tra i gruppi e di coesione nei gruppi.

4.2 Conclusioni

In questo Capitolo sono state applicate le nuove misure di dissimilarità proposte in questo lavoro su dati relativi ai prezzi dell'energia del mercato elettrico inglese. Sono state considerate solo le misure basate sul parametro di lisciamento e sulla correlazione tra i quantili di autocovarianza perchè meglio si adattano a questo tipo di serie. Infatti, è noto che le serie storiche relative ai prezzi di energia elettrica sono caratterizzate da componenti deterministiche e quindi, considerare misure che si basano ad esempio sulla distanza tra i parametri di un modello $ARMA$, non è il modo migliore per vedere differenze significative in questo tipo di applicazione. La misura basata sui quantili di autocovarianza è una dissimilarità che considera le informazioni contenute nella distribuzione, e quindi può adattarsi bene al caso in esame. In particolare, per il calcolo di d_{QAF} si utilizzano dei quantili che considerano l'informazione relativa a tutta la distribuzione, alle code e alla sola coda destra della distribuzione. In tre casi su quattro si sceglie la partizione che considera $K = 2$ gruppi, mentre nel restante caso si considera la partizione con $K = 3$ gruppi. Tutte le misure di dissimilarità concordano sul fatto che le 48 serie semi-orarie considerate possono essere divise nelle seguenti fasce:

- **Fascia notturna** (indicativamente dalle ore 19.00 alle ore 06.00). È una fascia caratterizzata dallo stesso andamento delle componenti deterministiche e dallo stesso tipo di correlazione tra i quantili di autocovarianza.

- **Fascia centrale della giornata**. È una fascia caratterizzata da presenza di picchi elevati che porta ad avere distribuzioni con la coda destra molto pesante.

Questa divisione è plausibile: è ragionevole che i picchi più elevati siano nella parte centrale della giornata, ossia negli orari di lavoro, mentre nella parte relativa alla notte e alle prime ore del mattino avvengono delle movimentazioni meno significative e non si verificano picchi molto elevati di prezzo. Per quanto riguarda la misura basata sul parametro di lisciamento occorre considerare 3 gruppi. Per ottenere una migliore separatezza tra i gruppi e una migliore coesione all'interno dei gruppi è necessario dividere ulteriormente

la parte centrale della giornata in due gruppi: il primo gruppo riguarda la mattina e le prime ore del pomeriggio e il secondo gruppo riguarda le ore del tardo pomeriggio. In base alle misure che si basano sui quantili si ritiene sufficientemente separata e coesa la partizione con $K = 2$ gruppi.

Conclusioni

In questo lavoro si è parlato di Cluster Analysis nell'ambito delle serie storiche. Per prima cosa si è spiegato cos'è e quali obiettivi ha la Cluster Analysis. In un secondo momento ci si è preoccupati di studiare delle misure di dissimilarità appositamente create per le serie storiche. Tali misure di dissimilarità sono costruite in modo tale da considerare la struttura di dipendenza intrinseca nelle serie storiche. Oltre a considerare delle misure già proposte in letteratura ci si è concentrati su due misure particolari. La prima, proposta da Lafuente-Rego e Vilar (2015), si basa sulla funzione dei quantili di autocovarianza. Questo tipo di dissimilarità funziona particolarmente bene quando si ha a che fare con processi a code pesanti nella distribuzione marginale, con modelli non lineari o con processi che seguono modelli a eteroschedasticità condizionale. La seconda misura approfondita in questo lavoro è una misura mai proposta prima e appositamente creata per le serie storiche che hanno al loro interno delle componenti deterministiche che non possono essere ignorate. Tale misura si basa sulla distanze euclidea tra i parametri di lisciamento di un modello lineare non parametrico per le componenti deterministiche.

Dopo aver introdotto e studiato alcune misure di dissimilarità tipiche per serie storiche ci si è chiesti se è possibile ottenere una divisione basata sui processi generatori delle serie storiche. In letteratura non esistono misure di dissimilarità in grado di dividere serie storiche con caratteristiche tanto diverse, inoltre, con le misure di dissimilarità non si è in grado di etichettare i gruppi trovati. In questo lavoro è stata proposta una procedura in grado di dividere ed etichettare le serie storiche in base a caratteristiche generali. Tale procedura si compone di 4 passi ognuno dei quali utilizza tecniche diverse per la divisione dei dati. Tra le tecniche considerate non si utilizzano mai le misure di dissimilarità: tali misure hanno, in questo contesto, due svantaggi. Il primo è che per definizione dividono sempre i dati in $K > 1$ gruppi anche se i dati non necessitano di divisione, e il secondo è che non sono in grado di etichettare le divisioni trovate. Per verificare il funzionamento della procedura di discriminazione proposta si è fatto uno studio di simulazione che comprende serie storiche provenienti da processi molto differenti. In $B = 100$

simulazione l'indice di accordo ARI calcolato nei 4 passi della procedura risulta pari a 0.9475 ($\pm$ 0.05) al primo passo, 0.8418 ($\pm$ 0.073) al secondo passo, 0.6792 ($\pm$ 0.073) al terzo passo e 0.3074 ($\pm$ 0.024) al quarto e ultimo passo. L'accordo all'ultimo passo non risulta molto elevato principalmente a causa degli errori commessi ai passi precedenti. Una volta applicata la procedura di discriminazione si può applicare, all'interno di ogni gruppo trovato, una particolare misura di dissimilarità che meglio si adatta ai gruppi etichettati della procedura. Se i dati provengono da contesti differenti e quindi sono molto diversi tra loro, la sola procedura può essere sufficiente per una buona partizione dei dati.

Nell'ultima parte di questo lavoro sono state applicate le due misure di dissimilarità ai dati relativi ai prezzi di energia elettrica del mercato elettrico inglese. Dato che le serie storiche provenivano tutte dallo stesso contesto ed è noto che tali serie hanno al loro interno delle componenti deterministiche che non possono essere ignorate, si è deciso di applicare direttamente, senza l'utilizzo della procedura, la misura che si basa sul parametro di lisciamento e la misura che si basa sui quantili di autocovarianza, considerando prima i quantili di tutta la distribuzione, poi i quantili delle code della distribuzione e poi i quantili della code destra dato che si stanno trattando serie di prezzi. I gruppi trovati con le due misure sono diversi, ma tutte e quattro le misure tengono separata la fascia delle ore del mattino dalle altre fasce della giornata. Per quanto riguarda la misura basata sul parametro di lisciamento si ha una divisione in tre gruppi: il primo gruppo coinvolge le ore della sera, della notte e del primo mattino, in particolare considera le ore della fascia 19.00-8.00, il secondo gruppo coinvolge la fascia del giorno, in particolare considera le ore della fascia 8.00-16.00, e l'ultimo gruppo coinvolge la fascia delle ore serali, in particolare coinvolge le ore della fascia 16.00-19.00. Per quanto riguarda le misure che si basano sui quantili di autocovarianza è risultato che la divisione più compatta e meglio coesa è la divisione che coinvolge soli due gruppi: il primo gruppo coinvolge la parte di giornata che riguarda le ore serali, della notte e del primo mattino, mentre il secondo gruppo riguarda le ore della giornata.

Bibliografia

Alonso A. M.; Berrendero J. R.; Hernández A.; Justel A. (2006). Time series clustering based on forecast densities. *Computational Statistics & Data Analysis*, **51**(2), 762–776.

Caliński T.; Harabasz J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, **3**(1), 1–27.

Dickey D. A.; Fuller W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, **74**(366a), 427–431.

Dunn J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, **4**(1), 95–104.

Fabbris L. (1997). *Statistica multivariata: analisi esplorativa dei dati*. McGraw-Hill Libri Italia.

Galeano P.; Peña D. (2001). Multivariate analysis in vector time series.

Gavrilov M.; Anguelov D.; Indyk P.; Motwani R. (2000). Mining the stock market (extended abstract): which measure is best? In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 487–496. ACM.

Golay X.; Kollias S.; Stoll G.; Meier D.; Valavanis A.; Boesiger P. (1998). A new correlation-based fuzzy logic clustering algorithm for fmri. *Magnetic Resonance in Medicine*, **40**(2), 249–260.

Hubert L.; Arabie P. (1985). Comparing partitions. *Journal of classification*, **2**(1), 193–218.

Hubert L.; Schultz J. (1976). Quadratic assignment as a general data analysis strategy. *British journal of mathematical and statistical psychology*, **29**(2), 190–241.

Hurvich C. M.; Simonoff J. S.; Tsai C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**(2), 271–293.

Lafuente-Rego B.; Vilar J. A. (2015). Clustering of time series using quantile autocovariances. *Advances in Data Analysis and Classification*, pp. 1–25.

Lance G. N.; Williams W. T. (1967). A general theory of classificatory sorting strategies ii. clustering systems. *The computer journal*, **10**(3), 271–277.

Liao T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, **38**(11), 1857–1874.

Montero P.; Vilar J. A. (2014). Tsclust: An r package for time series clustering. *Journal of*.

Otranto E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis*, **52**(10), 4685–4698.

Piccolo D. (1990). A distance measure for classifying arima models. *Journal of Time Series Analysis*, **11**(2), 153–164.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66**(336), 846–850.

Rousseeuw P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65.

Vilar J. A.; Alonso A. M.; Vilar J. M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics & Data Analysis*, **54**(11), 2850–2865.

Zani S.; Cerioli A. (2007). *Analisi dei dati e data mining per le decisioni aziendali*. Giuffré Editore.