POLITECNICO DI TORINO

Collegio di Ingegneria Informatica, del Cinema e Meccatronica

Corso di Laurea Magistrale
in Ingegneria Informatica (Computer Engineering)

Tesi di Laurea Magistrale

Esplorazione di indici di qualità per configurare gli algoritmi di clustering

Relatore
Prof.ssa Tania Cerquitelli

Candidato
Andrea Fanara

Ottobre 2017

Indice

1. Introduzione

In questa tesi sarà affrontato il tema di una tra le principali tecniche di ***Data Mining***, l'analisi di *clustering*, mediante due tra i più comuni algoritmi: ***K-means*** **e** ***Dbscan***.

L'obiettivo di quest'analisi è la suddivisione degli oggetti in gruppi con un certo grado di omogeneità; ciò lo si ottiene formando delle collezioni di *oggetti/dati* che presentano caratteristiche simili rispetto a ciascun oggetto nello stesso cluster e, contemporaneamente, dissimili agli oggetti di ogni altro cluster.

Il clustering, definito come *unsupervised classification*, ha lo scopo di segmentare i dati ma senza assegnare etichette di classe; non si hanno, infatti, classi predefinite ma ogni cluster può essere inteso come una classe contenente oggetti simili e le applicazioni tipiche possono essere:

- strumento stand-alone per cercare di capire come i dati sono distribuiti;
- passo di preprocessing per algoritmi.

Le applicazioni tipiche dell'utilizzo di tale tecnica sono:

- image processing & pattern recognition;
- analisi di dati spaziali;
- scienze economiche (*market research*)
- WWW (*raggruppamento di documenti simili e clustering di weblog per scoprire gruppi di pattern di accesso simili ad un sito web*).

Di seguito alcuni esempi di utilizzo:

- **marketing:** vengono scoperti gruppi distinti e poi viene usata questa conoscenza per sviluppare programmi di *targeted marketing*;
- **land use:** identifica aree terrestri simili rispetto all'osservazione della terra (*satellite*);
- **assicurazioni:** identifica gruppi di assicurati con caratteristiche comuni;
- **city-planning:** identifica gruppi di case sulla base di tipo, valore e localizzazione geografica;
- **studi di terremoti:** clustering di epicentri;
- **astronomia:** analisi delle immagini del cielo e delle esplosioni dei raggi gamma.

2. Stato dell'arte

2.1 Data mining e clustering

Il *data mining* è l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di un sapere o di una conoscenza a partire da grandi quantità di dati *(attraverso metodi automatici o semi-automatici)* e l'utilizzo scientifico, industriale o operativo di questo sapere. Mentre la statistica permette di elaborare informazioni generali riguardo ad una popolazione *(es. percentuali di disoccupazione, nascite, ecc…)*, il *data mining* è utilizzato per cercare correlazioni tra più variabili relativamente ai singoli individui. Ad esempio: conoscendo il comportamento di un cliente di una compagnia telefonica si cerca di effettuare una previsione di quanto spenderà nell'immediato futuro. In sostanza, il *data mining* rappresenta *"l'analisi da un punto di vista matematico eseguita su un database di grandi dimensioni".*

Oggi il *data mining* ha una duplice valenza:

- estrazione, con tecniche analitiche all'avanguardia, di informazione implicita, nascosta, da dati già strutturati, per renderla disponibile e direttamente utilizzabile;
- esplorazione ed analisi, eseguita in modo automatico o semiautomatico, su grandi quantità di dati al fine di scoprire *pattern* (*schemi*) significativi.

In entrambi i casi i concetti di informazione e di significato sono legati strettamente al dominio applicativo in cui si esegue il *data mining*; in altre parole, un dato può essere interessante o trascurabile a seconda del tipo di applicazione in cui si vuole operare. Questo tipo di attività è cruciale sia in molti ambiti della *ricerca scientifica* sia in altri settori *(per esempio in quello delle ricerche di mercato)*. Nel mondo professionale è utilizzata per risolvere problematiche diverse tra loro, che vanno dalla gestione delle relazioni con i clienti, all'individuazione di comportamenti fraudolenti, sino all'ottimizzazione di *siti web*.

Ecco alcuni esempi per comprendere meglio il concetto espresso sino a questo momento:

- cercare un numero di telefono nell'elenco (*non è data mining*);
- fare una ricerca in internet su "*vacanze alle Maldive*" (*non è data mining*);
- fare una ricerca nel web su una parola chiave e classificare i documenti trovati secondo un criterio semantico (*per esempio "corriere": nome di giornale, professione, ecc.) (è data mining);*
- scoprire chi sono i clienti che hanno maggiore propensione di acquisto su certi prodotti o campagne pubblicitarie *(è data mining)*;

2.1.1 Applicazioni del data mining nella ricerca scientifica

I fattori principali che hanno contribuito allo sviluppo del data mining sono:

- le grandi accumulazioni di dati in formato elettronico;
- il *data storage* poco costoso;
- i nuovi metodi e tecniche di analisi *(apprendimento automatico, riconoscimento di pattern).*

Le tecniche di *data mining* sono fondate su specifici ***algoritmi***. I *pattern* identificati possono essere, a loro volta, il punto di partenza per ipotizzare, e quindi verificare, nuove relazioni di

tipo causale fra fenomeni; in generale, possono servire in senso statistico per formulare previsioni su nuovi insiemi di dati.

Un concetto correlato al *data mining* è quello di apprendimento automatico (*machine learning*); infatti, l'identificazione di pattern può paragonarsi all'apprendimento, da parte del sistema di data mining, di una relazione causale precedentemente ignota, cosa che trova applicazione in ambiti come quello degli algoritmi euristici e dell'intelligenza artificiale. Tuttavia, occorre notare che il processo di *data mining* è sottoposto al rischio rivelare relazioni causali che poi si rivelano inesistenti.

Tra le tecniche maggiormente utilizzate in questo ambito vi sono:

- clustering;
- reti neurali;
- alberi di decisione;
- analisi delle associazioni (*individuazione dei prodotti acquistati congiuntamente).*

2.1.2 Clustering

Un *cluster* è una collezione di oggetti ***simili*** tra loro che sono contemporaneamente ***dissimili*** rispetto agli oggetti degli altri *cluster* ***(Figura 2.1.2.1)****;* la qualità del risultato di tale tecnica dipende dalla misura di similarità usata o dallo specifico algoritmo usato.

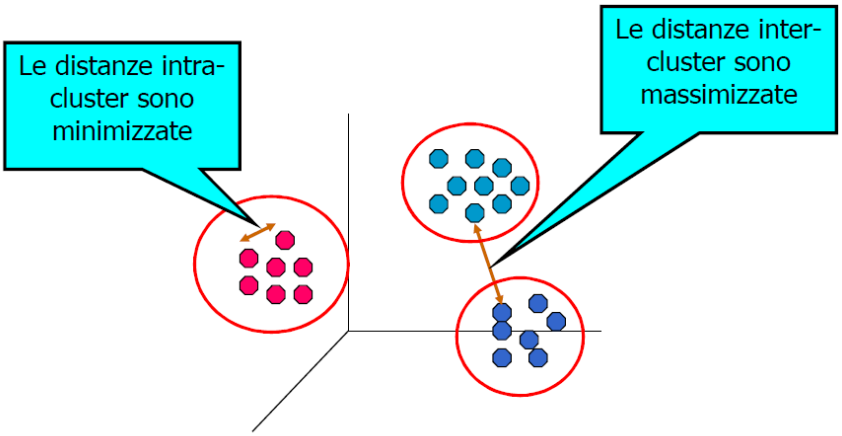


Figura 2.1.2.1: *Ecco un esempio di Clustering Analysis [1].*

Un clustering è un insieme di cluster; di essi può esserne fatta una distinzione importante:

- ***Clustering partizionante:*** una divisione degli oggetti in sottoinsiemi *(cluster)* non sovrapposti. Ogni oggetto appartiene esattamente a un cluster;

Figura 2.1.2.2: *Esempio di clustering partizionante [1].*

- ***Clustering gerarchico:*** un insieme di cluster annidati organizzati come un albero gerarchico.

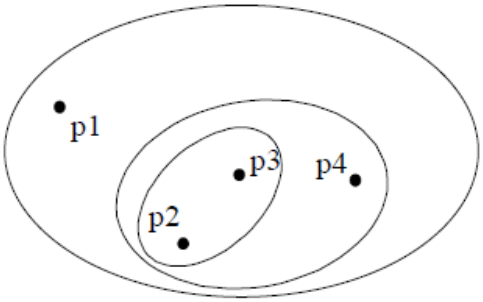


Figura 2.1.2.3: *Esempio di clustering gerarchico tradizionale[1].*

Oltre alla differenziazione appena citata esistono altre distinzioni tra insiemi di cluster:

- ***esclusivo/non esclusivo***
  - ✓ in un clustering non esclusivo i punti possono appartenere a più cluster;
  - ✓ utile per rappresentare punti di confine o più tipi di classi.
- ***fuzzy/non fuzzy***
  - ✓ in un fuzzy clustering un punto appartiene a tutti i cluster con un peso tra 0 e 1;
  - ✓ la somma dei pesi per ciascun punto deve essere 1;
  - ✓ i clustering probabilistici hanno caratteristiche similari.
- ***parziale/completo***
  - ✓ in un clustering parziale alcuni punti potrebbero non appartenere a nessuno dei cluster.
- ***eterogeneo/omogeneo***
  - ✓ in un clustering eterogeneo i cluster possono avere dimensioni, forme e densità molto diverse.

Detto ciò è possibile spiegare in dettaglio le due tecniche di *clustering* usate all'interno della tesi: *K-means* e *Dbscan*.

2.2 K-means

L'algoritmo K-means è un algoritmo di *clustering* partizionale che permette di suddividere un insieme di oggetti in "K" gruppi sulla base dei loro attributi. È una variante dell'algoritmo di aspettativa-massimizzazione (EM) il cui obiettivo è determinare i "K" gruppi di dati generati da distribuzioni gaussiane. Si assume che gli attributi degli oggetti possano essere rappresentati come vettori e che quindi formino uno spazio vettoriale.

L'obiettivo che l'algoritmo propone è di minimizzare la varianza totale intra-cluster. Ogni cluster è identificato mediante un centroide o punto medio. L'algoritmo segue una procedura iterativa. Inizialmente crea "K" partizioni e assegna ad ogni partizione i punti d'ingresso, o casualmente o usando alcune informazioni euristiche. Quindi calcola il centroide di ogni gruppo. Costruisce, quindi, una nuova partizione associando ogni punto d'ingresso al cluster il cui centroide è più vicino ad esso. Di conseguenza sono ricalcolati i centroidi per i nuovi cluster e così via, finché l'algoritmo non converge.

2.2.1 Pregi e difetti

L'algoritmo ha acquistato notorietà giacché converge molto velocemente. Infatti, si è osservato che generalmente il numero di iterazioni è minore del numero di punti. Comunque, l'algoritmo può essere molto lento nel caso peggiore: D. Arthur e S. Vassilvitskii hanno mostrato che esistono certi insiemi di punti per i quali l'algoritmo impiega un tempo superpolinomiale, $2^{\Omega(\sqrt{n})}$ a convergere. Più recentemente A. Vattani ha migliorato questo risultato mostrando che l'algoritmo può impiegare tempo esponenziale ($2^{\Omega(n)}$) a convergere anche per certi insiemi di punti sul piano. D'altra parte D. Arthur, B. Manthey e H. Roeglin, hanno mostrato che la *smoothed complexity* dell'algoritmo è polinomiale; la qual cosa è a supporto del fatto che l'algoritmo è veloce in pratica. In termini di qualità delle soluzioni l'algoritmo non garantisce il raggiungimento dell'ottimo globale. La qualità della soluzione finale dipende largamente dal set di cluster iniziale e può, in pratica, ottenere una soluzione ben peggiore dell'ottimo globale. Dato che l'algoritmo è di solito estremamente veloce, è possibile applicarlo più volte e, fra le soluzioni prodotte, scegliere quella più soddisfacente. Un altro svantaggio dell'algoritmo è che esso richiede di scegliere il numero di cluster (k) da ottenere. Se i dati non sono naturalmente partizionati, si ottengono risultati strani. Inoltre l'algoritmo funziona bene solo quando sono individuabili cluster sferici nei dati.

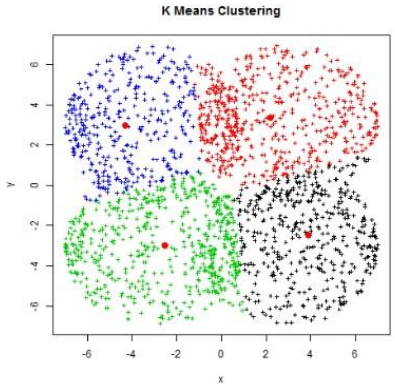


Figura 2.2.1.1: *Esempio di utilizzo dell'algoritmo K-means. In rosso è raffigurato il centroide di ogni cluster[4].*

2.2.2 Valutazione della bontà dei cluster e convergenza

La misura più comunemente utilizzata è lo scarto quadratico medio (*SSE – Sum of Squared Error*); per ogni punto l'errore è la distanza dal centroide del cluster a cui è assegnato:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

*"**x**"* è un punto appartenente al cluster C_i e m_i è il rappresentante del cluster C_i; è possibile dimostrare che il centroide che minimizza SSE quando si utilizza come misura di prossimità di distanza euclidea è la media dei punti del cluster.

$$m_i = \sum_{x \in C_i} x$$

Ovviamente il valore di SSE si riduce incrementando il numero dei cluster K. Un buon clustering con K ridotto può avere un valore di SSE più basso di un cattivo clustering con K più elevato.

C'è soltanto un numero finito di modi di partizionare *n* record in *k* gruppi. Quindi c'è soltanto un numero finito di possibili configurazioni in cui tutti i centri sono centroidi dei punti che possiedono. Se la configurazione cambia in un'iterazione, deve avere migliorato la distorsione. Quindi, ogni qualvolta la configurazione cambia, deve portare in uno stato mai visitato prima:

- **il riassegnamento dei record ai centroidi è fatto sulla base delle distanze minori;**
- **il calcolo dei nuovi centroidi minimizza il valore di SSE per il cluster.**

L'algoritmo deve fermarsi nel momento in cui termina la disponibilità di visitare ulteriori configurazioni. Non è detto, tuttavia, che la configurazione finale sia quella che in assoluto presenta il minimo valore di SSE (***Figura 2.1.2.1)***:

- **spostare un centroide della soluzione sul lato destro comporta sempre un aumento di SSE, mentre la configurazione sul lato sinistro presenta un SSE minore.**

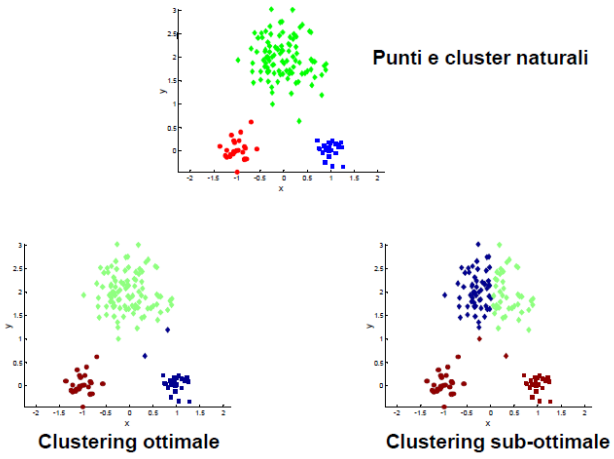


Figura 2.2.2.1: *ecco la rappresentazione dei cluster dopo le iterazioni[1].*

2.3 Dbscan

Il DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) è un metodo di *clustering* proposto nel 1996 da Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu e rientra nell'ambito delle tecniche di *clustering*.

Grazie ad esso è possibile connettere differenti regioni di punti usando come indicatore la densità; aree con densità simili vengono connesse, permettendo cosi di individuare ed isolare gli *outlier* presenti nell'insieme dei dati.

2.3.1 L'algoritmo

Per capire il funzionamento di questo algoritmo basato sulla densità, è necessario introdurre una serie di concetti che saranno poi utili per la sua descrizione:

- ***Direttamente raggiungibile in densità***: Siano **q** e **p** due punti, "**q"** è detto direttamente raggiungibile da "**p"** se essi non sono distanti più di una certa quantità "**ε"; ;** tale quantità è definita generalmente dall'utente che sta utilizzando l'algoritmo ed è utile a indicare l'area entro cui cercare punti vicini. Secondo il numero di dimensioni utilizzate la quantità "**ε**" può quindi indicare una distanza, una superficie o un volume. Assegnando ad "**ε**" un valore ritenuto grande l'algoritmo utilizzato sarà impostato in un modo più permissivo, in altre parole andrà a ricercare tutti i punti adiacenti a quello di partenza prendendo in considerazione un ritorno più ampio;
- ***Raggiungibile in densità*****:** Siano **q** e **p** due punti, **q** si dice raggiungibile in densità **p** se esiste una sequenza $p_1 \ldots p_n$ di punti con $p_1 = p$ e $p_n = q$ dove ognuno di essi e direttamente raggiungibile dal suo predecessore;
- ***Densamente connesso*****:** Due punti **p** e **q** sono connessi in densità se esiste un punto **o** tale che sia ***o-p*** che ***o-q*** siano raggiungibili in densita.

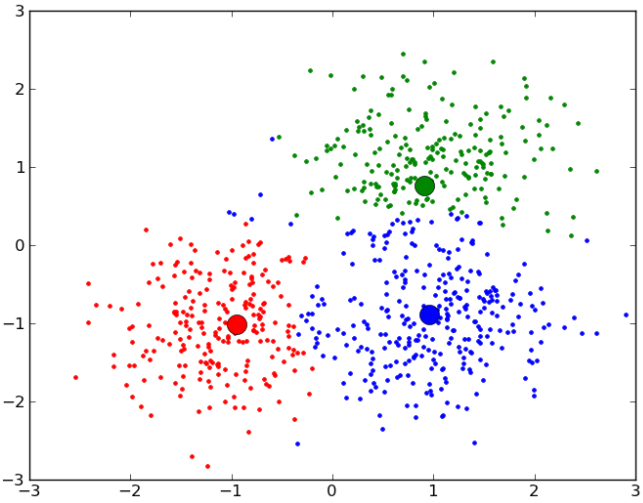


Figura 2.3.1.1: *Applicazione del DBSCAN su un insieme di dati[2].*

Nella ***figura 2.3.1.1*** è possibile osservare come in un insieme di punti sono identificati dall'algoritmo tre aree a densità differente e creati tre differenti *cluster*.

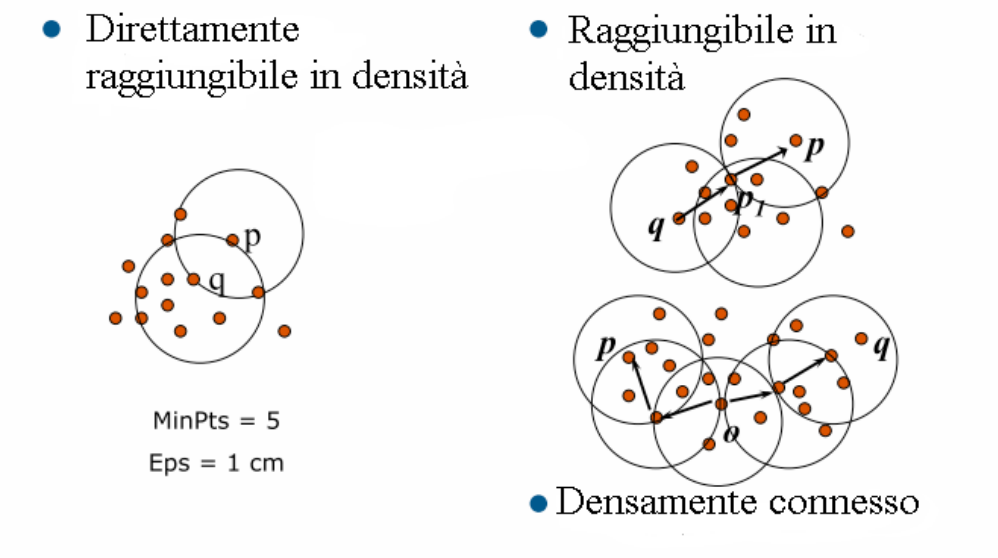


Figura 2.3.1.2: *Rappresentazione dei casi direttamente raggiungibile in densità, densamente raggiungibile e densamente connesso per coppie di punti [2].*

Partendo dai concetti esposti in precedenza, si può introdurre ora la definizione di *cluster* per l'algoritmo DBSCAN. Un *cluster* può essere definito come un insieme in cui tutti i punti al suo interno sono mutualmente connessi in densità. Inoltre, se un punto e connesso in densità ad un altro punto del *cluster* allora anch'esso ne è parte.

Per spiegare l'algoritmo sono inoltre necessari i tre concetti seguenti:

- **Core point:** Punto all'interno di un *cluster:* da esso è possibile raggiungere in densità un numero di punti maggiore del numero minimo definito per creare un *cluster* (***minpts***);
- **Border point:** Punto posto sul bordo di un *cluster*: esso è raggiungibile in densità da un *core point* ma, partendo da un *border point*, non è piu possibile espandere ulteriormente il *cluster*;
- **Noise point:** Rappresenta un *outlier*: l'algoritmo non è in grado di collocarlo in alcun *cluster* con i parametri ***ε*** e ***minpts*** forniti.

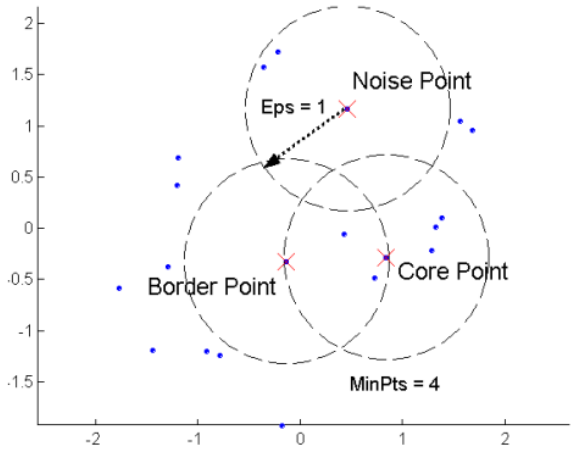


Figura 2.3.1.3: *Rappresentazione di esempi di noise point, border point e core point[2].*

Ecco il funzionamento spiegato nel dettaglio:

- è selezionato arbitrariamente un punto ***p***;
- se ***p*** è un *punto core*, sono individuati tutti i punti **raggiungibili in densità** da ***p*** rispetto a *Eps* e *MinPts*, ed è formato un cluster;
  1. sono aggiunti al cluster ***p*** i punti **direttamente raggiungibili in densità;**
  2. è controllato in modo ricorsivo se tali punti siano a loro volta core, ecc…;
  3. sono uniti via via i punti **raggiungibili in densità.**
- Se da *p* è possibile raggiungere un punto border *p'*:
  1. nessun punto è *raggiungibile in densità* da *p'*;
  2. si passa a esaminare un altro punto del database non ancora considerato.

Si continua fino alla completa visita di tutti i punti.

2.3.2 Complessità

DBSCAN visita ogni punto del database anche più volte nel caso di punti candidati a cluster differenti. DBSCAN esegue esattamente una invocazione per ogni punto e se è utilizzata una struttura indicizzata che esegue un'interrogazione del vicinato in ***O(log n)*** si ottiene un tempo globale di esecuzione pari a ***O(n * log n)*** . Senza l'uso di strutture indicizzate il tempo di esecuzione è pari a $O(n^2)$. Spesso la matrice delle distanze di dimensione $(n^2-n)/2$ è creata per evitare, appunto, il ricalcolo delle distanze riducendo il tempo di elaborazione a spese della memoria utilizzata pari a $O(n^2)$.

2.3.3 Vantaggi e svantaggi

DBSCAN presenta i seguenti vantaggi:

- non richiede di conoscere il numero di cluster a priori, al contrario dell'algoritmo *K-means*;
- può trovare cluster di forme arbitrarie. Può anche trovare un cluster completamente circondato da un cluster differente a cui non è connesso (dato il parametro MinPts, il cosiddetto effetto single-link, cluster differenti connessi da una sottile linea di punti è ridotto);
- possiede la nozione di rumore;
- richiede soltanto due parametri ed è per lo più insensibile all'ordine dei punti nel database: solo i punti posti sull'arco fra 2 cluster differenti possono cambiare la loro appartenenza se l'ordine dei punti è cambiato mentre l'assegnazione ai cluster è unico solo su isomorfismi.

Svantaggi:

- La qualità del *clustering* generato da DBSCAN dipende dalla sua misura della distanza che è riconducibile alla funzione ***getVicini(P, epsilon)***. La più comune misura usata è la distanza euclidea. In particolare per il *high-dimensional data*, questa misura della distanza diventa quasi inutile tanto da esser denominata *"Maledizione della dimensionalità";* nei fatti diventa difficile trovare un valore appropriato per epsilon. Tuttavia questo effetto è presente anche in altri algoritmi basati sulla distanza euclidea.
- DBSCAN non è in grado di classificare insiemi di dati con grandi differenze nelle densità giacché la combinazione minPts-epsilon non può poi essere scelta in modo appropriato per tutti i cluster.

3. Configurazione automatica dei parametri di input degli algoritmi di clustering

3.1 K-means (in dettaglio)

Dopo la scelta di 5 dataset *(di cui 3 di medio/alte dimensioni e 2 con una bassa quantità di dati)* e dei dati che lo compongono *(devo, infatti, necessariamente escludere le colonne che rappresentano attributi categorici e che porterebbero ad un inquinamento dei risultati)*, si è reso necessario applicare su essi la **normalizzazione**. Con il termine normalizzazione si intende l'azione necessaria affinché importanti variabili di processo con piccole ampiezze non siano considerate meno rilevanti di variabili con ampiezza maggiore; in questo caso saranno considerati tre differenti tipi di normalizzazione: MIN-MAX, Z-SCORE e 0-1 *(il cui funzionamento sarà spiegato in seguito).*

Per estrapolare i dati che sono stati normalizzati è necessario l'utilizzo dei due algoritmi su citati: *K-means* e *DbScan*.

All'interno del *K-means* utilizzato nell'applicazione creata, ricopre un ruolo fondamentale l'estrapolazione del valore di *silhouette* legato agli **"N"** *clusters* presi in considerazione. Ecco i cinque punti che sono possibili selezionare:

- silhouette media per "N" *clusters (con "N" che va da 2 a 20);*
- silhouette per il cluster selezionato: per ogni punto del cluster considerato, viene presentato il rispettivo andamento della *silhouette*;
- silhouette media per "N" *clusters*: il grafico si presenta con "N" rettangoli *(a seconda del numero di cluster coinvolti)* raffiguranti la *silhouette* media; in questo modo all'utente è rivelato in modo evidente lo scostamento dal valore 0;
- andamento della *silhouette* per "N" *clusters*: selezionato il numero di cluster da visualizzare vengono presentati a video "N" grafici raffiguranti l'andamento della *silhouette* per ogni punto di ogni cluster coinvolto *(es. se N=2 saranno presenti 2 clusters con un certo numero di punti che appartengono al primo ed i restanti al secondo; l'applicazione si occuperà di mostrare i due grafici presentando per ognuno di essi, sulle ascisse, i relativi punti che ne fanno parte e sulle ordinate il rispettivo andamento della silhouette);*
- percentuale dei punti oltre al valore 0 di *silhouette*: per ogni "N" *clusters* (2-20) viene verificata la percentuale dei punti che superano la soglia impostata. Nell'analisi dei dati raccolti ci si baserà di quest'ultimo punto per scegliere quale numero di *clusters* selezionare per ottenere le tre migliori percentuali ed utilizzare questa informazione nei punti spiegati in precedenza (dal 2° al 4°).

Per capire meglio come venga calcolata all'interno dell'applicazione la distanza di un generico punto **p** da **q** si può vedere la formula di seguito:

$$\boldsymbol{d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}}$$

3.2 Dbscan (in dettaglio)

Allo stesso modo di come si è operato con il K-means si scelgono le percentuali di tuple *(4 differenti)* appartenenti al dataset in questione e si implementa il **k-dist** creando una matrice con le dimensioni del numero di tuple considerate per ogni percentuale:

	1	2	N
1			
2			
N			

Figura 3.2.1: *Tabella delle distanze; la prima riga/colonna è formata dai punti appartenenti alla percentuale di dati considerati. All'interno della tabella sono inserite, in corrispondenza di colonna i-esima e riga i-esima, la distanza tra i due punti.*

Nella ***figura 3.2.1*** è rappresentata la matrice delle distanze; ogni riga è ordinata in modo decrescente ed è scelta la colonna *(che corrisponderà al K)*. In seguito è mostrato un grafico con i punti della colonna *"**K-esima**"* ordinata in modo crescente; mostrato il grafico con la porzione di dati scelta, il punto in cui l'andamento farà una curva decisa (circa 90°) potrà offrire sull'asse delle ordinate il valore "**ε**" legato al **minPoint(K)** scelto.

3.3 Normalizzazione dei dati raccolti

- Le variabili d'ingresso e di uscita, considerate in un processo di acquisizione dati, in genere sono grandezze fisiche di diverso tipo.
- Esse presentano valori numerici spesso assai diversi che non possono essere forniti direttamente a un sistema poiché si rischierebbe di penalizzare quelle grandezze che, a causa del loro range di variazione, assumono dei valori numerici più piccoli delle altre.
- La fase di normalizzazione, altrimenti detta di *"scaling",* riveste un'importanza rilevante in fase di *pre-processing* dei dati.
- E' necessaria affinché importanti variabili di processo con piccole ampiezze non siano considerate meno rilevanti di variabili con ampiezza maggiore.

Noi prenderemo in considerazione tre diversi tipi di normalizzazione: **MIN-MAX**, **Z-SCORE** e **0-1**.

3.3.1 Min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Dove:

v = valore da normalizzare;

v' = valore normalizzato;

min_A = valore minimo della variabile A;

max_A = valore massimo della variabile A;

new_min_A = valore minimo del nuovo range che si vuole definire;

new_max_A = valore massimo del nuovo range che si vuole definire.

3.3.2 Z-score normalization

$$v' = \frac{v - mean_A}{\sigma_A}$$

Dove:

v = valore da normalizzare;
v' = valore normalizzato;
$mean_A$ = valore medio della variabile A;

σ_A = deviazione standard della variabile A.

3.3.3 0/1 normalization

$$v' = \frac{v - min_A}{max_A - min_A}$$

Dove:

v = valore da normalizzare;
v' = valore normalizzato;

min_A = valore minimo della variabile A;

max_A = valore massimo della variabile A;

Questa trasformazione produce valori che variano tra 0 e 1.

3.4 Silhouette Coefficient

In dettaglio, il *Silhouette Coefficient* combina le idee della coesione e della separazione *(per singoli punti, cluster singoli o risultati del clustering)*.

Per un punto "**i**":

- sia C_i il cluster di *i;*
- **calcola**: a_i = distanza media di *i* dagli altri punti di *Ci;*
- **calcola**: b_i = *min per ogni cluster C,* $C \neq C_i$ *(distanza media di i dai punti del cluster C)*
- Silhouette Coefficient s_i per il punto *i*:

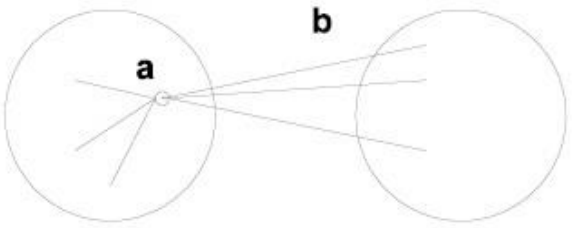


Figura 3.4.1: *Combinazione dei fattori di coesione e separazione per il calcolo della silhouette [1].*

$$s_i = \frac{(b_i - a_i)}{max(a_i, b_i)}$$

- tra -1 e 1;
- caso -1 non desiderabile, perché accadrebbe se $a_i > b_i$;
- ciò che si vorrebbe ottenere è un valore positivo *(ovvero* $a_i < b_i$*),* con a_i molto piccolo (≅ 0) *[in questo caso* s_i *tende a 1]*;
- il coefficiente per un singolo cluster;
- media dei coefficienti di tutti i punti del cluster;
- il coefficiente per un *clustering* completo;
- media dei coefficienti di tutti i punti.

4. I dataset

4.1 3D Road Network (North Jutland, Denmark) dataset

[Rete stradale 3D con le informazioni di elevazione ad altra precisione (+/- 20cm) usato in eco-routing e per gli algoritmi di routing per la stima carburante/Co2]

Questo dataset proviene dal dipartimento di "*Computer Science*" dell'Aarhus University in Danimarca.

Il database è stato composto sommando le informazioni riguardanti l'elevazione di una rete stradale 2D nello Jutland settentrionale in Danimarca (che copre una regione di 185 x 135 km^2). I valori di elevazione sono stati estratti grazie ad una scansione laser e questa rete stradale 3D è stata utilizzata per l'analisi comparativa di diversi algoritmi di carburante e di stima Co2; questo dataset può essere usato da tutte quelle applicazioni che richiedono di conoscere informazioni accurate riguardo l'elevazione di una rete stradale per eseguire un "*routing*" più preciso per l'eco routing, percorsi per ciclisti ecc. Per la comunità che si occupa di *data mining* e *machine learning* questo set di dati può essere usato come conferma ulteriore dei dati ottenuti grazie alle immagini satellitari.

Ogni riga di siffatto dataset è formata da quattro valori che rappresentano le seguenti informazioni:

1. **OSM_ID**: ID OpenStreetMap associato a ogni tratto di strada;
2. **LONGITUDINE**: Web Mercaptor (formato Google) che indica la longitudine;
3. **LATITUDINE**: Web Mercaptor (formato Google) che indica la latitudine;
4. **ALTITUDINE**: altezza in metri.

Nota: Ogni OSM_ID rappresenta l'ID assegnato da OpenStreetMaps per i segmenti stradali e ogni ID inserito con lo stesso ordine con cui appare il segmento lungo la strada. Ciò significa che per ottenere una "poli-linea 3D" può essere disegnata tramite l'unione degli OSM_ID.

4.2 User Knowledge Modeling dataset

[E' il vero e proprio set di dati sullo stato di conoscenza degli studenti sul tema delle macchine elettriche a corrente continua]

Per comprendere come sono stati raccolti i dati presenti nel dataset è importante comprendere cosa sia uno *"user model"*, ovvero la raccolta e classificazione dei dati personali associati ad un utente specifico. I dati che sono inclusi nel modello sono dipendenti dallo scopo dell'applicazione *(possono includere informazioni personali quali nomi degli utenti e le età, i loro interessi, le loro competenze e conoscenze, i loro obiettivi e piani, le loro preferenze e le loro antipatie, o dati sul loro comportamento e le loro interazioni con il sistema*).

Per questo dataset gli utenti sono stati classificati dagli autori utilizzando un classificatore di conoscenza intuitiva *(una tecnica ML ibrida composta di k-NN e metodi di esplorazione meta-euristici).*

Ogni riga di tale dataset è formata da valori che rappresentano le seguenti informazioni:

1. **STG:** il grado del tempo di studio sull'argomento (*valore in input*);
2. **SCG:** grado del numero di ripetizioni dell'utente per il materiale in questione (*valore in input*);
3. **STR:** è il grado di tempo di studio dell'utente per gli argomenti legati all'argomento in questione (*valore in input*);
4. **LPR:** le prestazioni in esame dell'utente per gli argomenti legati al tema in questione (*valore in input*);
5. **PEG:** le prestazioni in esame dell'utente per l'argomento in questione (*valore in input*);
6. **UNS:** Il livello di conoscenza dell'utente (*valore obiettivo*)

4.3 Wholesale customers dataset

[Il set di dati si riferisce ai clienti di un distributore all'ingrosso. Esso comprende la spesa annuale in unità monetarie (m.u.) su diverse categorie di prodotto]

Ogni riga di tale dataset è formata da valori che rappresentano le seguenti informazioni:

1. **FRESH**: spesa annuale (m.u.) sui prodotti freschi;
2. **LATTE**: spesa annuale (m.u.) sui prodotti lattiero-caseari;
3. **GROCERY**: spesa annuale (m.u.) sui prodotti alimentari;
4. **FROZEN**: spesa annuale (m.u.) sui prodotti surgelati;
5. **DETERGENTS_PAPER**: spesa annuale (m.u.) sui detergenti e prodotti di carta;
6. **DELICATESSEN**: spesa annuale (m.u.) sui prodotti e specialità gastronomiche;
7. **CHANNEL**: Horeca (Hotel/Restaurant/Cafè) o canale di vendita al dettaglio;
8. **REGIONE**: Lisbona, Porto o altra regione.

4.4 Turkiye Student Evaluation dataset

[Questo set di dati contiene la totalità dei 5820 punteggi di valutazione forniti da studenti provenienti da Gazi University di Ankara (Turchia).C'è un totale di ventotto domande specifiche del corso e ulteriori cinque attributi].

Ogni riga di questo dataset è formata da valori che rappresentano le seguenti informazioni:

1. **INSTR**: Identificatore di istruttore, valori scelti tra 1,2,3;
2. **CLASS**: Codice del corso (descrittore), valori compresi tra 1 e 13;
3. **REPEAT**: Numero di volte in cui lo studente partecipa al corso, valori compresi tra 0,1,2 ecc….
4. **ATTENDANCE**: Codice del livello di partecipazione, valori compresi tra 0,1,2,3,4;
5. **DIFFICULTY**: Grado di difficoltà del percorso *(come percezione da parte dello studente)*, valori compresi tra 1,2,3,4,5;
6. **Q1**: Nel semestre il contenuto, il metodo di insegnamento e di valutazione, sono stati forniti all'inizio del corso;
7. **Q2**: Gli obiettivi erano chiaramente indicati all'inizio del periodo;
8. **Q3**: Il corso era proporzionato al numero dei crediti assegnati;
9. **Q4**: Il corso è stato insegnato secondo il programma mostrato il primo giorno di lezione;

10. **Q5**: Le discussioni in classe, i compiti a casa, le applicazioni e gli studi, sono stati soddisfacenti;
11. **Q6**: Il libro di testo e le risorse concernenti i corsi sono stati sufficienti ed aggiornate;
12. **Q7**: Il corso ha permesso il lavoro sul campo, le applicazioni, laboratori, discussioni e altri studi;
13. **Q8**: I quiz, le assegnazioni, i progetti e gli esami hanno contribuito ad aiutare l'apprendimento;
14. **Q9**: Ho molto gradito la classe ed ero impaziente di partecipare attivamente durante le lezioni;
15. **Q10**: Le mie aspettative iniziali sul corso sono state raggiunte al termine del periodo o l'anno;
16. **Q11**: Il corso è stato rilevante e utile per la mia crescita professionale;
17. **Q12**: Il corso mi ha aiutato a guardare la vita e il mondo con una nuova prospettiva;
18. **Q13**: La conoscenza del docente era adeguata e aggiornata;
19. **Q14**: L'istruttore è arrivato preparato per le classi;
20. **Q15**: L'istruttore ha insegnato coerentemente al piano di lezione annunciato;
21. **Q16**: L'istruttore era impegnato al corso ed era comprensibile;
22. **Q17**: L'istruttore è arrivato in orario per le classi;
23. **Q18**: L'istruttore ha un modo di parlare liscio e facile da seguire;
24. **Q19**: L'istruttore ha fatto un uso efficace delle ore di lezione;
25. **Q20**: L'istruttore ha dato spiegazioni e non perdeva occasioni per essere utile agli studenti;
26. **Q21**: L'istruttore ha dimostrato un approccio positivo per gli studenti;
27. **Q22**: L'istruttore era aperto e rispettoso dei punti di vista degli studenti sul corso;
28. **Q23**: L'istruttore ha incoraggiato alla partecipazione al corso;
29. **Q24**: L'istruttore ha consegnato rilevanti compiti a casa, progetti, ed ha aiutato/guidato gli studenti;
30. **Q25**: L'istruttore ha risposto alle domande relative il corso dentro e fuori dallo stesso;
31. **Q26**: Il sistema di valutazione dell'istruttore (domande finali, i progetti, le assegnazioni, ecc.) era adatto agli obiettivi del corso;
32. **Q27**: L'istruttore ha fornito soluzioni per gli esami e ne ha discusso con gli studenti;
33. **Q28**: L'istruttore ha trattato tutti gli studenti in modo corretto.

Per i valori che vanno da Q1-Q28 sono stati dati valori di gradimento compresi tra 1 e 5.

4.5 US Census Data (1990) Data Set

[Il set di dati US Census Data 1990 contiene un campione (1%) del Public Use Microdata Samples (PUMS) ottenuto dal campione completo del censimento del 1990]

Il set di dati USCensus1990 è stato ottenuto dal sito Census Bureau *(U.S. Department of Commerce)* utilizzando il sistema di estrazione dei dati.

Ci sono 68 attributi categorici ed il set di dati USCensus1990 è stato ottenuto grazie alla seguente sequenza di operazioni:

1. **Randomizzazione**: l'ordine dei casi nei dati originali è stato permutato in modo casuale;
2. **Selezione di attributi**: i sessantotto inclusi nel set di dati sono riportati di seguito. Nel set di dati *USCensus1990* è stata aggiunta una sola lettera di prefisso al nome originale. Aggiungiamo la lettera "i" per indicare che sono usati valori degli attributi originali e "d" per indicare che i valori degli attributi originali (*per ciascun caso*) siano stati mappati a nuovi valori *(la mappatura precisa è descritta di seguito)*;

Ogni riga di questo dataset è formata da valori che rappresentano le seguenti informazioni:

1. **ID**;
2. **dAge**: età del soggetto;
3. **dAncstry1**: identità di origine (*etnia*);
4. **dAncstry2**: identità di origine (*etnia*);
5. **iAvail**: disponibile per lavorare *(è possibile indicare se si ha già un lavoro, se non si ha l'età per lavorare, ecc…);*
6. **iCitizen**: cittadinanza *(è possibile scegliere tra diverse zone);*
7. **iClass**: classe di lavoratore *(es. Lavoratore per il governo, impiegato/a per associazione no profit ecc.);*
8. **dDepart**: tempo di partenza per il lavoro *(ore e minuti)*;
9. **iDisabl1**: lavoro con limitazioni (*prima versione*);
10. **iDisabl2**: lavoro con limitazioni (*seconda versione*);
11. **iEnglish**: capacità di parlare *inglese (diversi gradi, es. Molto bene, bene,non bene ecc...)*;
12. **iFeb55**: servito dal Febbraio 1955 al Luglio 1964;
13. **iFertil**: numero di bambini nati *(es. Nessuno, un bambino, due bambini, ecc.);*
14. **dHispanic**: origini ispaniche *(es. No origine ispanica, origine messicana, origine cubana, ecc….);*
15. **dHour89**: ore di lavoro nell'ultima settimana dell'anno 1989;
16. **dHours**: ore di lavoro nell'ultima settimana;
17. **iImmigr**: anno di immigrazione *(ci sono diverse fasce selezionabili, es. 1987-1990, 1985-1986, ecc……)*;
18. **dIncome1**: salario o stipendio nel 1989;
19. **dIncome2**: reddito non agricolo (*lavoro autonomo*) nel 1989;
20. **dIncome3**: reddito agricolo *(lavoro autonomo)* nel 1989;
21. **dIncome4**: entrate di quarta tipologia (*es. Affitti*);
22. **dIncome5**: entrate di quinta tipologia;
23. **dIncome6**: entrate di sesta tipologia;
24. **dIncome7**: entrate di settima tipologia;
25. **dIncome8**: tutte le restanti entrate;
26. **dIndustry**: indica a quale delle 235 categorie appartiene l'individuo;
27. **iKorean**: l'individuo ha preso servizio durante il conflitto con la Corea nel 1950;
28. **iLang1**: l'individuo parla altri linguaggi in famiglia oltre l'inglese;
29. **iLooking**: l'individuo sta cercando lavoro;
30. **iMarital**: stato matrimoniale *(es. non sposato/a, sposato/a, divorziato/a, separato, ecc….);*
31. **iMay75880**: ha prestato servizio dal maggio 1975 all'agosto 1980;
32. **iMeans**: mezzi di trasporto per andare al lavoro (*es. bicicletta, moto, ecc…..);*

33. **iMilitary**: l'individuo sta prestando *(o ha prestato)* il servizio militare;
34. **iMobility**: l'individuo vive nella stessa casa dal 1° aprile;
35. **iMobilism**: l'individuo ha limitazioni nella mobilità;
36. **dOccup**: occupazione dell'individuo *(es. amministratori, manager, agenti, ingegneri, ecc….)*;
37. **iOthserv**: ha prestato servizio in qualsiasi altro momento;
38. **iPerscare**: limitazione "*personal care*";
39. **dPOB**: luogo di nascita *(tra quelli disponibili in lista);*
40. **dPoverty**: l'individuo è in stato di povertà;
41. **dPwgt1**: peso dell'individuo;
42. **iRagechild**: presenza ed età di un proprio figlio *(range di età indicato tra le possibilità)*;
43. **dRearning**: totale guadagni;
44. **iRelat1**: relazione di primo grado *(es. marito/moglie, figlio/figlia, padre/madre, ecc…..)*;
45. **iRelat2**: relazione di secondo grado *(es. nipote, nonno, zio, cugino, ecc…..)*;
46. **iRemplpar**: stato di occupazione dei genitori *(es. entrambi i genitori lavorano 35 o più ore, solo la madre lavora più di 35 ore, ecc….);*
47. **iRiders**: occupanti del veicolo guidato dal soggetto *(es. nel veicolo c'è solo il soggetto, ci sono due persone, ci sono tre persone, ecc….);*
48. **iRlabor**: stato di occupazione *(es. impiegato civile, membro delle Forze Armate, ecc….)*;
49. **iRownchild**: l'individuo ha un proprio figlio;
50. **dRpincome**: il totale delle persone aggiunte al nucleo familiare;
51. **iRPOB**: zona di nascita del soggetto *(es. Nord-Est, Sud, Ovest ecc…..);*
52. **iRrelchid**: l'individuo è in relazione con bambini di altri nuclei familiari;
53. **iRspouse**: lo stato dell'individuo *(es. sposato/a - con marito/moglie assente, vedovo/a, divorziato/a, separato/a, ecc…..)*;
54. **iRvetserv**: veterano di guerra *(es. del Vietnam, della Corea, ecc…..);*
55. **iSchool**: iscrizione scolastica;
56. **iSept80**: in servizio dal settembre 1980 o nel periodo successivo;
57. **iSex**: sesso, uomo/donna;
58. **iSubfam1**: relazione dell'individuo con la famiglia *(es. marito/moglie, figlio/a, ecc.....)*;
59. **iSubfam2**: numero di membri della famiglia;
60. **iTmpabsnt**: periodo di assenza dal lavoro *(es. in vacanza, in malattia, ecc...);*
61. **dTravtime**: tempo di viaggio per recarsi al lavoro;
62. **iVietnam**: servizio in Vietnam dall'Agosto del 1964;
63. **dWeek89**: settimane di lavoro nel 1989;
64. **iWork89**: indica se l'individuo ha lavorato nel 1989;
65. **iWorklwk**: indica se l'individuo ha lavorato nell'ultima settimana;
66. **iWWII**: l'individuo ha prestato servizio nella seconda guerra mondiale dal Luglio del 1940;
67. **iYearsch**: grado di istruzione raggiunto dall'individuo *(es. nessuna scuola terminata, asilo, scuola elementare, ecc…..);*
68. **iYearwrk**: anno dell'ultimo lavoro *(es. 1990, 1989, 1985-1987, ecc….);*
69. **dYrsserv**: anni di servizio militare;

4.6 Tamilnadu Electricity Board Hourly Readings Data Set

[Il set di dati rappresenta il consumo di energia elettrica attorno al Thanjavur].

Il set di dati Tamilnadu Electricity Board Hourly Readings raccoglie dati riguardanti il consumo elettrico nella zona residenziale *(industrie, esercizi commerciali, ecc….)* nelle vicinanze del Thanajvur *(Tamil Nadu)*.

Ogni *tupla* di questo *dataset* è formata da valori che rappresentano le seguenti informazioni:

1. **ForkVA**;
2. **ForkW**;
3. **Type** *(i valori possibili sono: banca, industria automobilistica, industria del cemento, fattoria del primo tipo, fattoria del secondo tipo, clinica privata, industria tessile, industria del pollame, villa, appartamento, industria alimentare, industria chimica, industria tessile, industria di fertilizzanti, ostello, ospedale, supermercato, teatro, università).*

5. Analisi dei dati raccolti (K-means)

Dopo la lettura dei dati è possibile fare alcune considerazioni a riguardo. Come anticipato, si sono presi in esame 3 *dataset* di medio/alte dimensioni tra quelli a disposizione *(US Census dataset, Turkiye Student Evaluation dataset ed Electricity dataset)* decidendo di considerare dalle 3000 *tuple* in avanti e 2 *dataset* di ridotte dimensioni *(User Knowledge Modeling dataset e Wholesale customers dataset).*

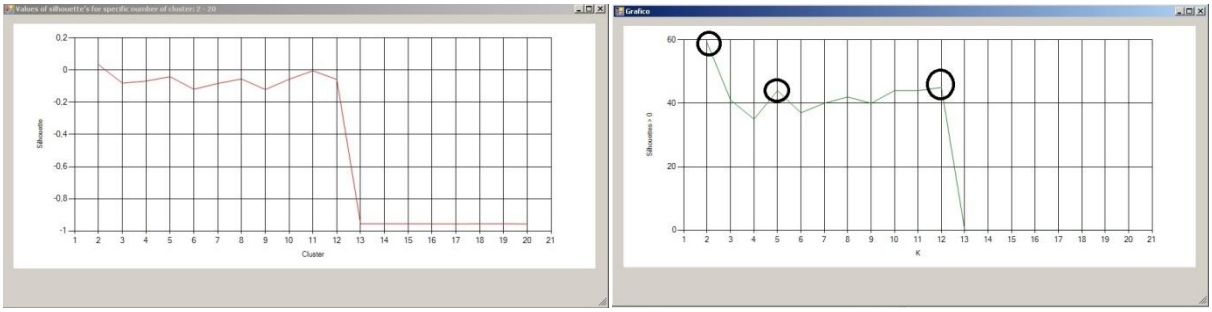
5.1 US Census Data (1990) Data Set

Questo *dataset* contiene circa 70000 *tuple* (70477); di queste sono state considerate il 10%, 20%, 30% e 40%.

10% (7047 tuple)

Di seguito sono presentati due grafici; il primo descrive la *silhouette* media per ogni *clusters* (*2-20*) e il secondo serve ad indicare la percentuale del numero di punti superanti la soglia "0" della *silhouette* per lo stesso numero di *clusters*.

Come riferito in precedenza le attenzioni saranno rivolte al secondo grafico scegliendo tra le tre migliori percentuali; in questo caso per le elaborazioni successive utilizzeremo 2, 5 e 12 *clusters*.



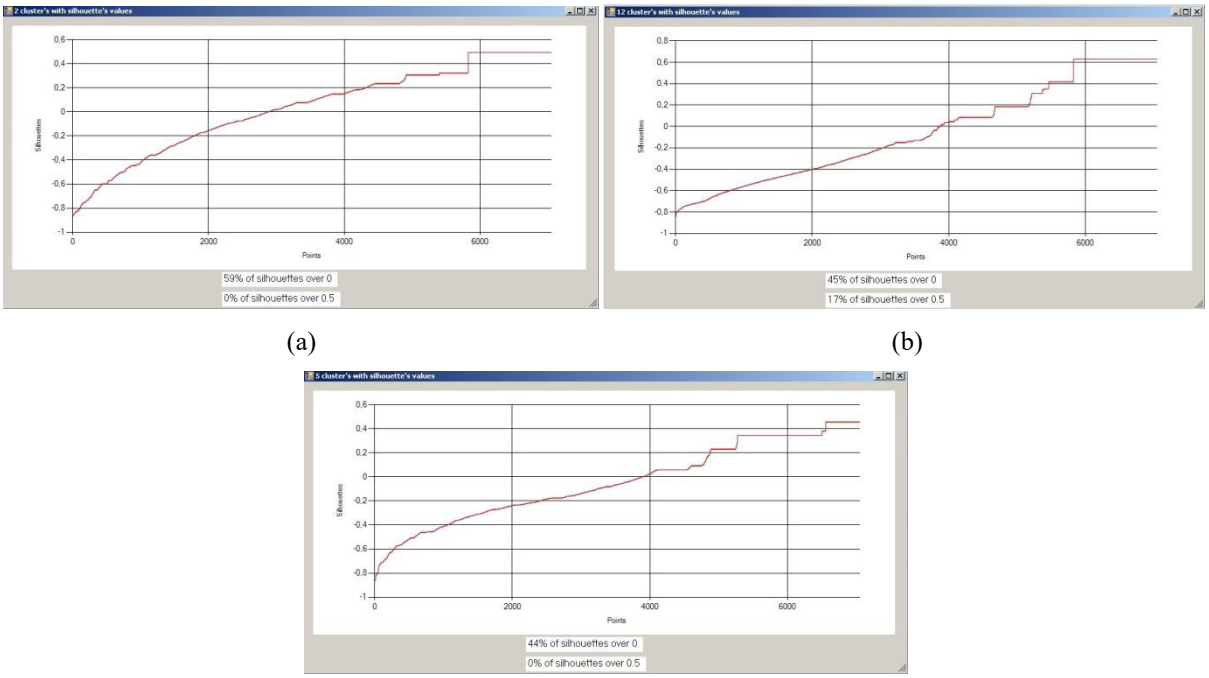
(a) (b)

Figura 5.1.1: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

Esaminando i grafici prodotti ***(Figura 5.1.2)*** si nota l'andamento della *silhouette* per i punti considerati. Sotto il grafico sono indicate le percentuali riguardanti il numero di punti superiori ai valori 0 e 0,5 della soglia di *silhouette*; nel caso in specie:

1. **2 clusters** *(1° miglior valore di silhouette)*: 59% dei punti superiori a 0 e nessun punto superiore allo 0,5;
2. **12 clusters** *(2° miglior valore di silhouette):* 45% dei punti superiori a 0 e 17% dei punti superiori allo 0,5;

3. **5 clusters** *(3° miglior valore di silhouette):* 44% dei punti superiori a 0 e nessun punto superiore allo 0,5;

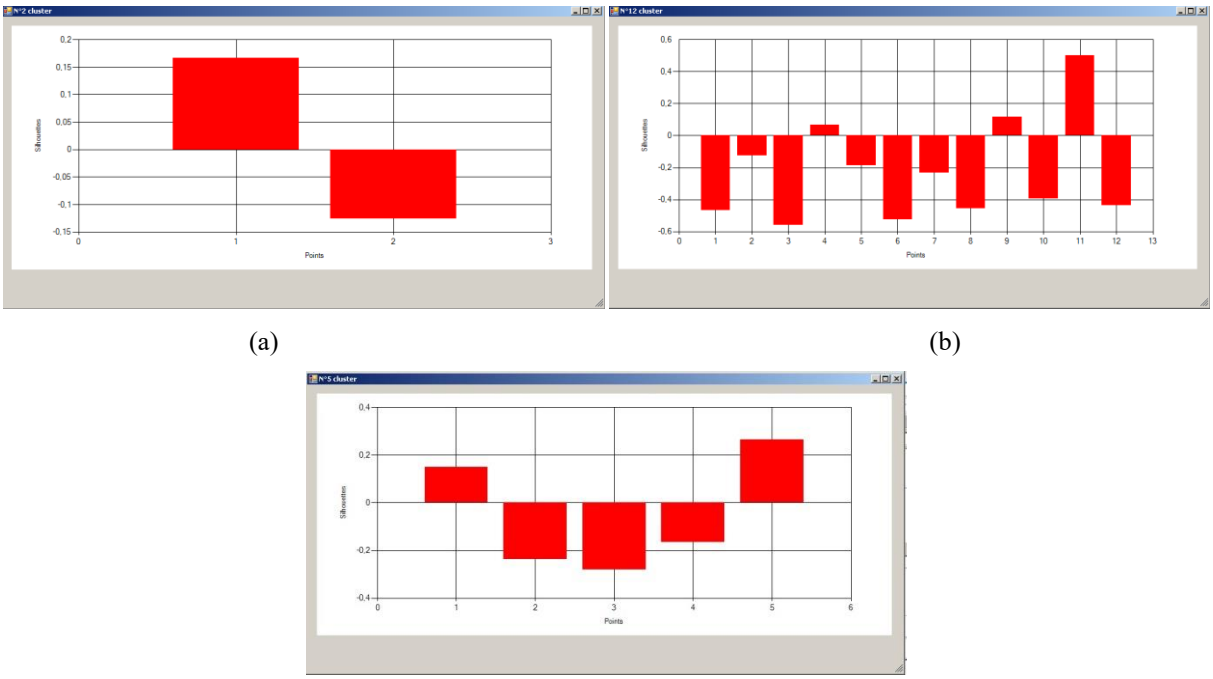


(a) (b)

(c)

Figura 5.1.2: *Si prendono in considerazione gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valor medio (c).*

Dei *clusters* in esame è stata verificata la *silhouette* media ottenendo i seguenti andamenti ***(Figura 5.1.3)***:

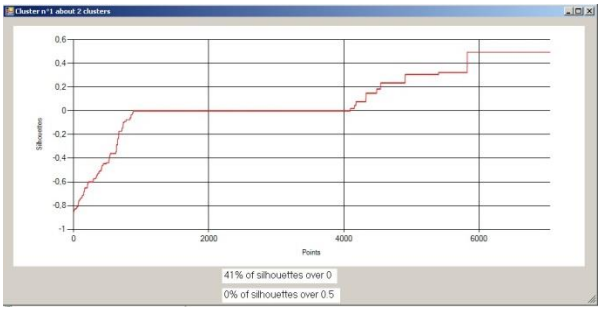


(a) (b)

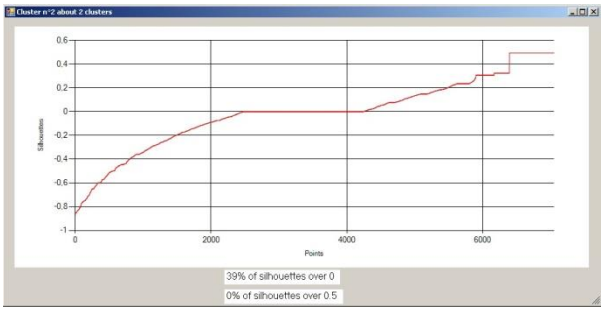
(c)

Figura 5.1.3: *Come in precedenza, si considerano gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valor medio (c) e per ogni clusters che ne faccia parte viene mostrato lo scostamento dalla soglia "0".*

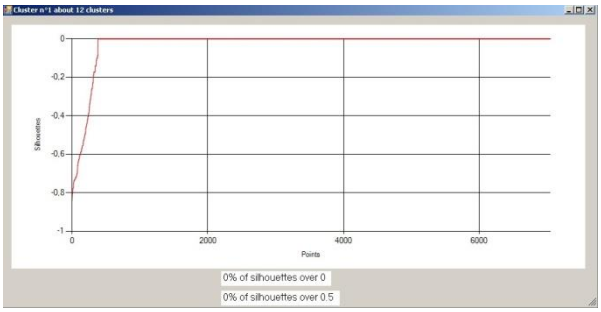
Infine, viene rilevato l'andamento della *silhouette* per i due casi migliori *(2 e 12 clusters)* ottenendo i trend mostrati ***(Figura 5.1.4)***. Considerando l'uso di 2 *clusters (e mostrandone l'andamento della silhouette per ogni valore di "N clusters" che ne faccia parte)* si ha, rispettivamente, il 41% e 39% dei valori oltre la soglia "0". Allo stesso modo si considera l'uso di 12 *clusters* e si nota come dal 4° grafico in avanti si ha, rispettivamente, il 3%, 5%, 4%, 4%, 4%, 20%, 19%, 36% e 35%, dei punti superiori al valore "0" di *silhouette*. Gli ultimi due grafici presentano inoltre il 17% e 16% dei punti oltre la soglia 0,5 di *silhouette*.



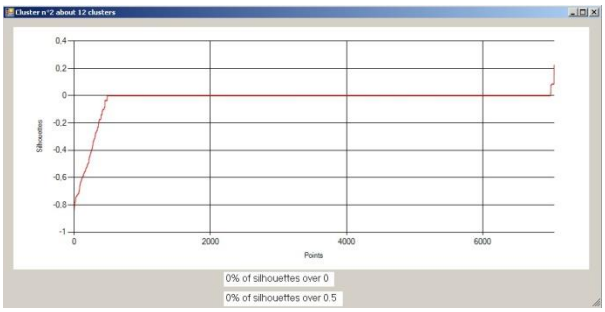
(a)



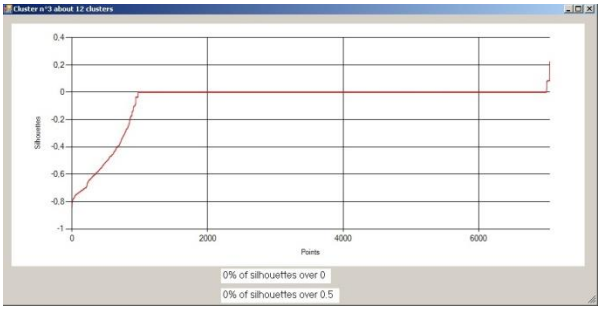
(b)



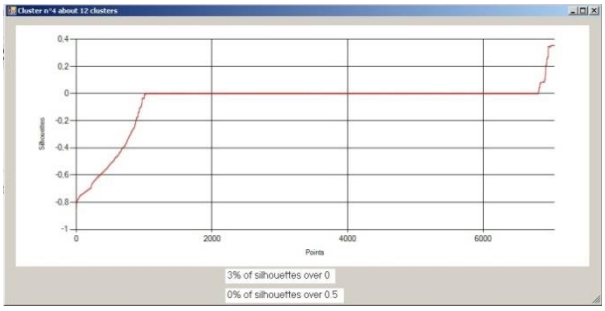
(c)



(d)



(e)



(f)

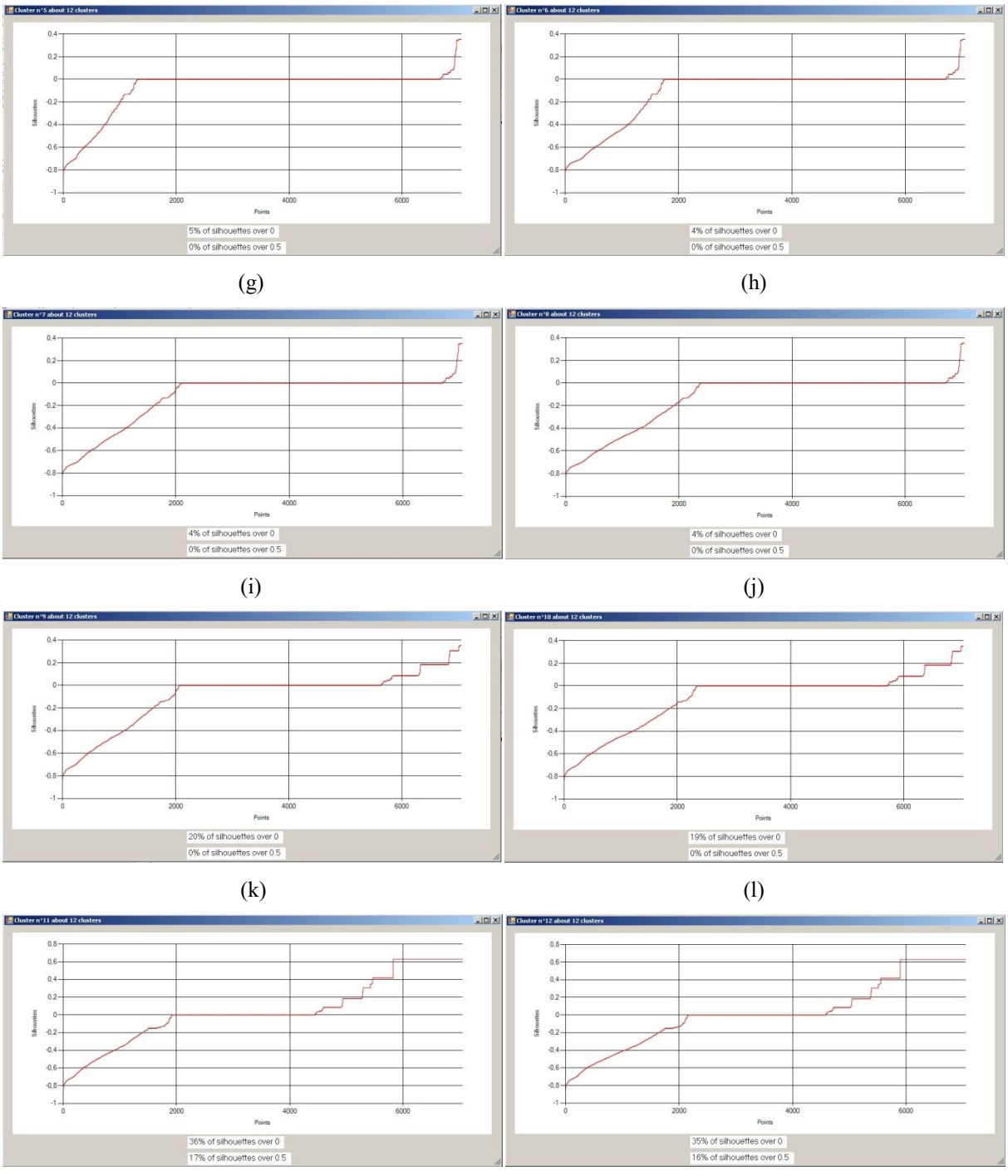
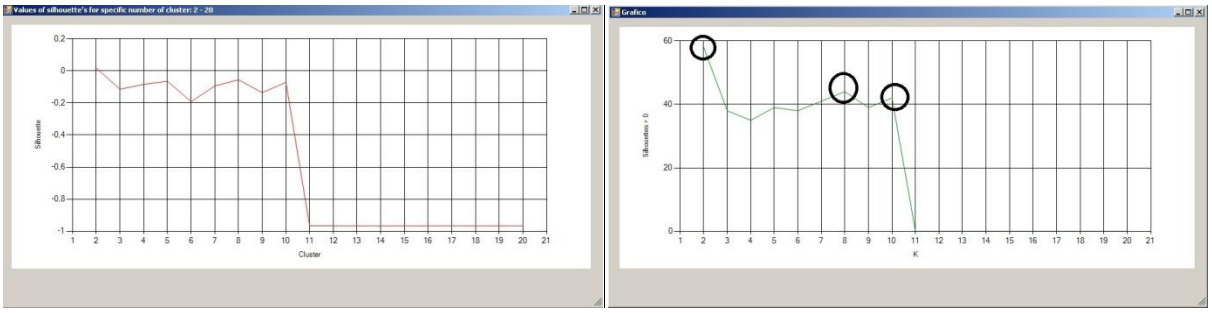


Figura 5.1.4: *Considerando gli "N" clusters con miglior valor medio di silhouette e secondo miglior valor medio di silhouette, si assiste all'andamento della silhouette per gli "N" clusters considerati. Il miglior valor medio di silhouette (con 2 clusters) è rappresentato dagli andamenti raffigurati nelle configurazioni (a) e (b). Il secondo miglior valor medio (12 clusters) dalle restanti figure (c), (d), (f), (g), (h), (i), (j), (k), (l), (m) e (n).*

20% (14094 tuple)

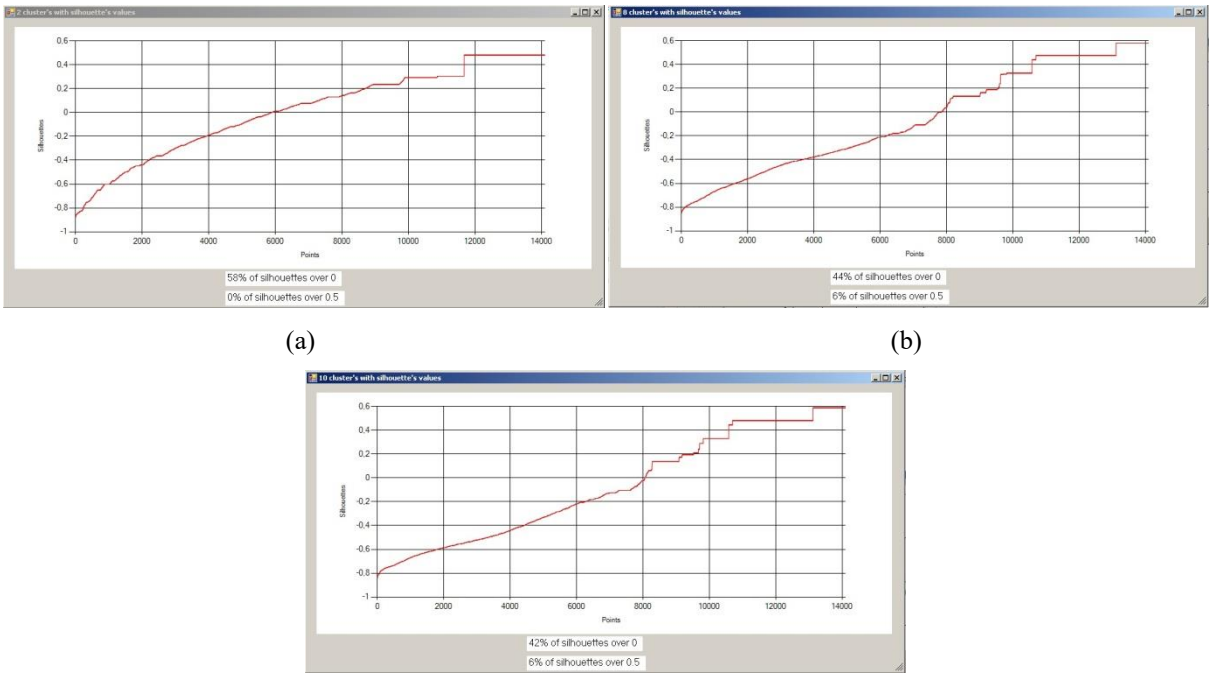
Anche in questo caso si considerano i valori di *silhouette* per i *clusters* 2-20 ottenendo il grafico ***(Figura 5.1.5)***. Osservando il secondo, che raffigura le percentuali dei punti oltre la soglia "0", si percepisce come i migliori valori di *silhouette* si ottengono con 2, 8 ed 11 *clusters*.



(a) (b)

Figura 5.1.5: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

Andamento della *silhouette* per "N" cluster ***(Figura 5.1.6)***: con 58%, 44% e 42% dei valori superiori allo "0" si ha una ulteriore conferma della bontà dell'uso di 2 *clusters* per ottenere i migliori valori di *silhouette*; oltre a ciò, in questo caso, per 8 ed 11 *clusters*, il 6% dei punti riescono a superarne la soglia dello 0,5.



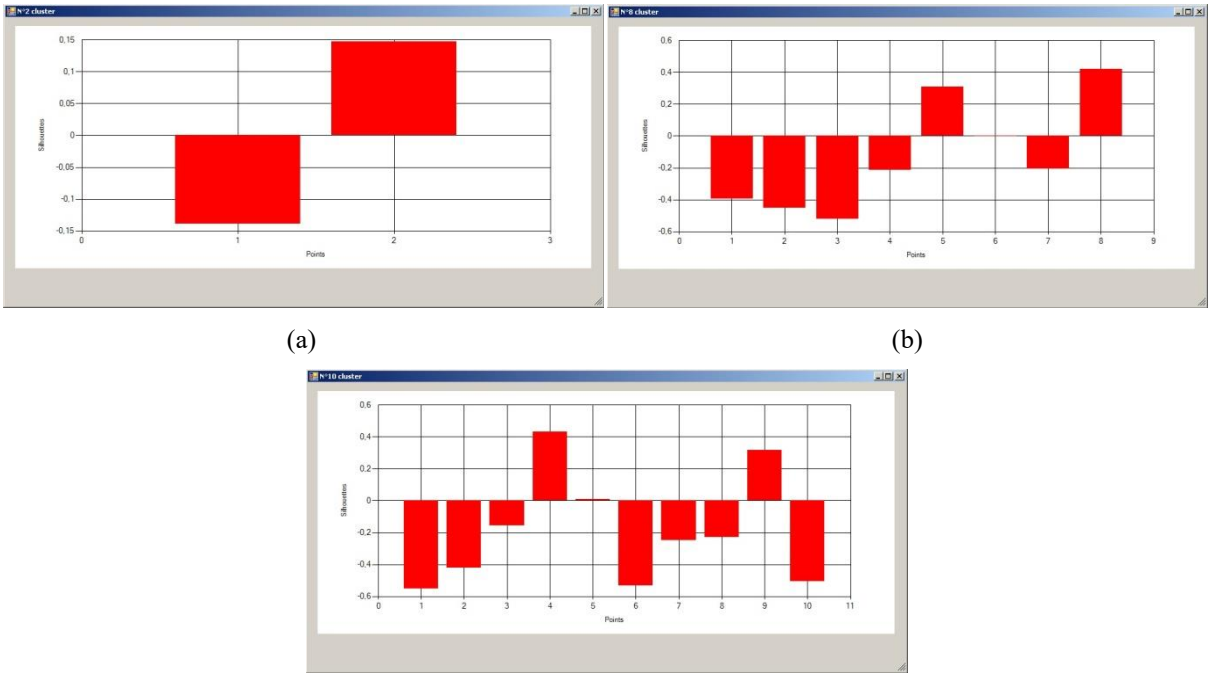
(a) (b)

(c)

Figura 5.1.6: *Si prendono in esame gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di*

silhouette (b) e quelli che offrono il terzo miglior valor medio (c).

Dei cluster in esame, si è deciso di verificare la *silhouette* media per gli "N" *clusters* considerati, ottenendo i seguenti andamenti ***(Figura 5.1.7)***:

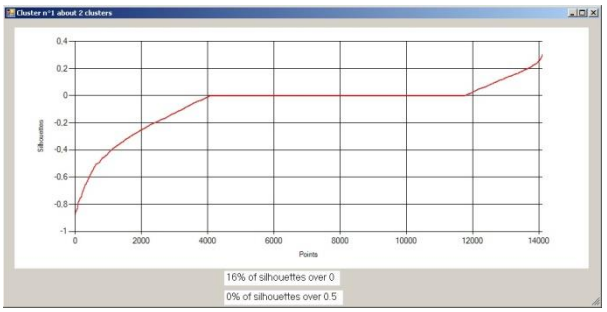


(a) (b)

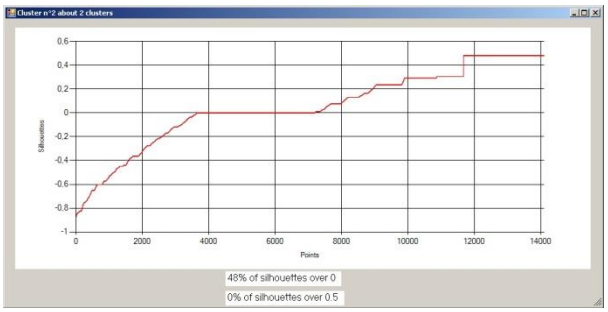
(c)

Figura 5.1.7: *Si considerano gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valor medio (c); per ogni clusters che ne fanno parte ne viene mostrato lo scostamento dalla soglia "0".*

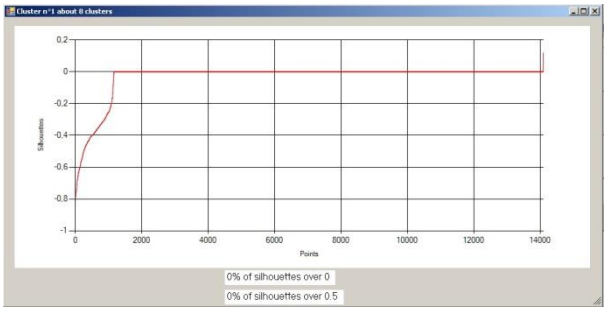
Concludendo, sarà mostrato l'andamento della *silhouette* con 2 e 8 *clusters* ***(Figura 5.1.8)***. Dai grafici ottenuti si registra che con 2 *clusters* i valori di *silhouette* superiore alla soglia "0" (grafico 1° e 2°), per i *clusters* considerati, risultano essere del 16% e 48%; invero, con 8 *clusters* i valori maggiori alla soglia sono (dal 4° cluster in poi) dell'1%, 29%, 31%, 29% e 36%.


Cluster n°1 about 2 clusters
Silhouettes
Points
16% of silhouettes over 0
0% of silhouettes over 0.5

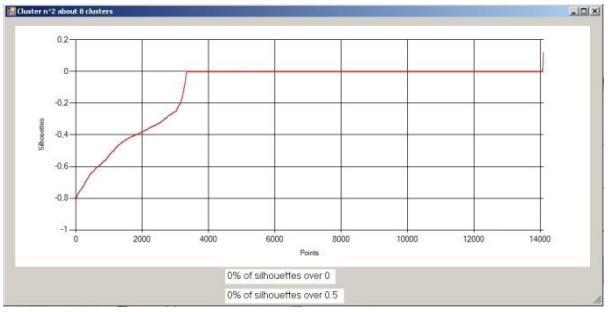
(a)


Cluster n°2 about 2 clusters
Silhouettes
Points
48% of silhouettes over 0
0% of silhouettes over 0.5

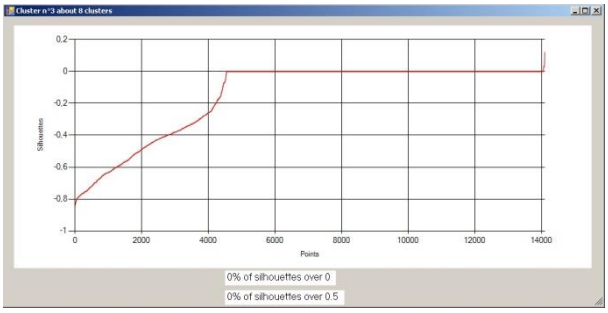
(b)


Cluster n°1 about 8 clusters
Silhouettes
Points
0% of silhouettes over 0
0% of silhouettes over 0.5

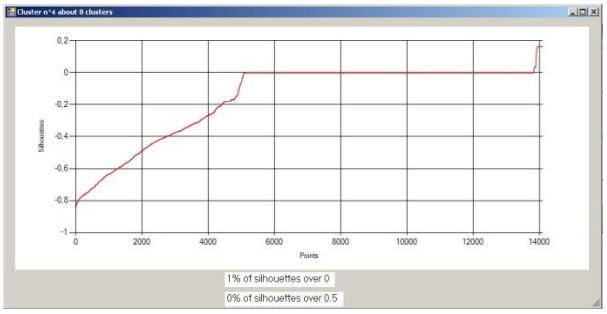
(c)


Cluster n°2 about 8 clusters
Silhouettes
Points
0% of silhouettes over 0
0% of silhouettes over 0.5

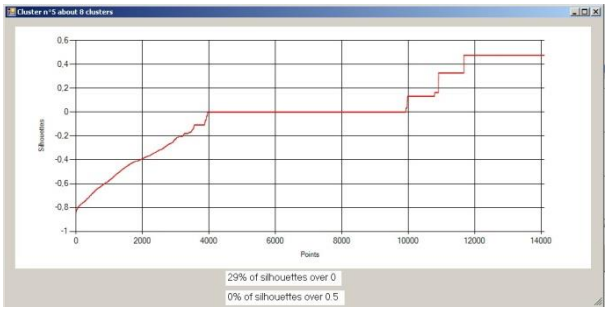
(d)


Cluster n°3 about 8 clusters
Silhouettes
Points
0% of silhouettes over 0
0% of silhouettes over 0.5

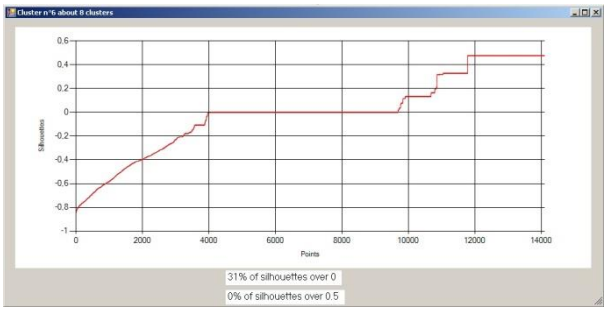
(e)


Cluster n°4 about 8 clusters
Silhouettes
Points
1% of silhouettes over 0
0% of silhouettes over 0.5

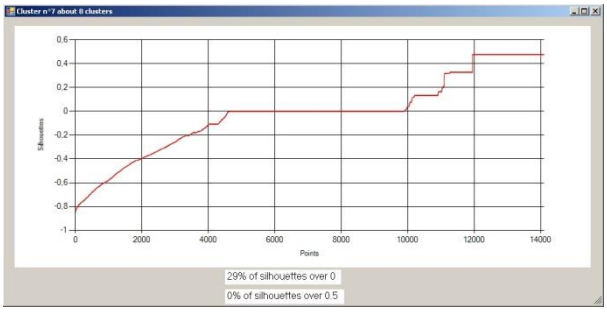
(f)


Cluster n°5 about 8 clusters
Silhouettes
Points
29% of silhouettes over 0
0% of silhouettes over 0.5

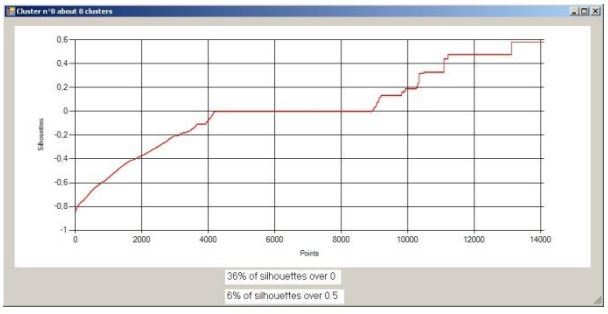
(g)


Cluster n°6 about 8 clusters
Silhouettes
Points
31% of silhouettes over 0
0% of silhouettes over 0.5

(h)


Cluster n°7 about 8 clusters
Silhouettes
Points
29% of silhouettes over 0
0% of silhouettes over 0.5

(i)


Cluster n°8 about 8 clusters
Silhouettes
Points
36% of silhouettes over 0
6% of silhouettes over 0.5

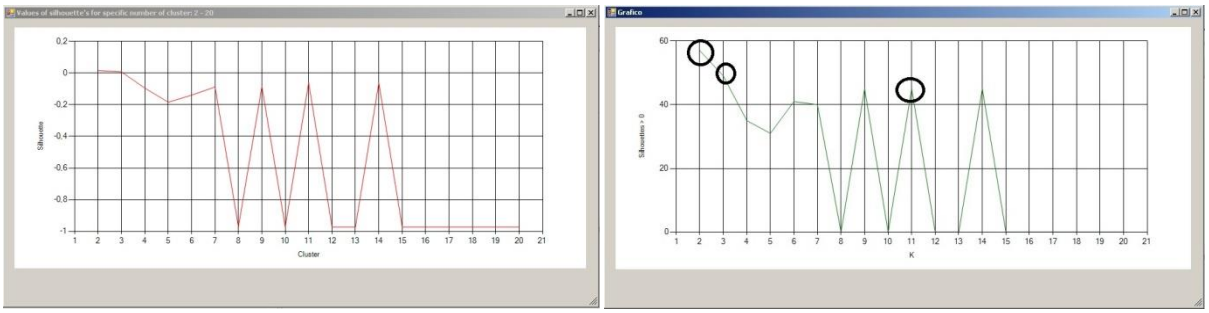
(l)

Figura 5.1.8: *Considerando gli "N" clusters con miglior valor medio di silhouette e secondo miglior valor medio di silhouette, mostro l'andamento della silhouette per gli "N" clusters considerati.*

Il miglior valor medio di silhouette (con 2 clusters) è rappresentato dagli andamenti illustrati nelle figure (a) e (b). Il secondo miglior valor medio (8 clusters) dalle restanti figure (c), (d), (e), (f), (g), (h), (i), (l).

30% (21143 tuple)

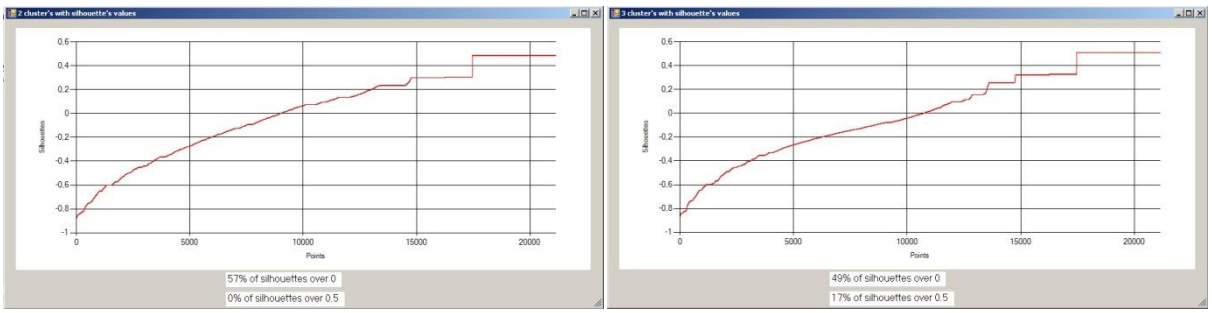
In questo caso si considera la *silhouette* per i *clusters* 2-20 e si ottiene il grafico ***(Figura 5.1.9)***. I migliori valori di *silhouette* si raggiungono con 2 e 3 *clusters*, mentre il terzo miglior valore si ottiene con l'uso di 11 *clusters*.



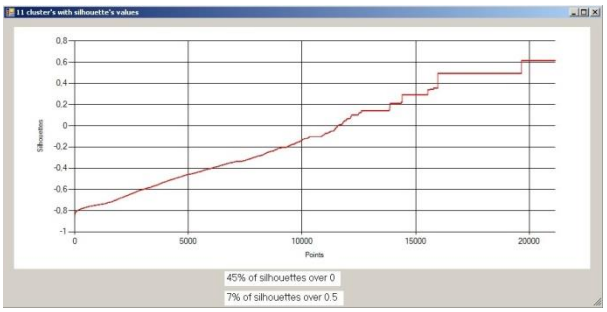
(a) (b)

Figura 5.1.9: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b)*

Andamento della *silhouette* per "N" cluster ***(Figura 5.1.10)***. Con 57%, 49% e 45% dei valori superiori alla soglia si ottiene un'altra conferma della bontà dell'uso di 2 cluster per avere i valori di *silhouette* superiori. Per 3 e 11 *clusters* si hanno, inoltre, rispettivamente il 17% e 7% di punti oltre la soglia dello 0,5.



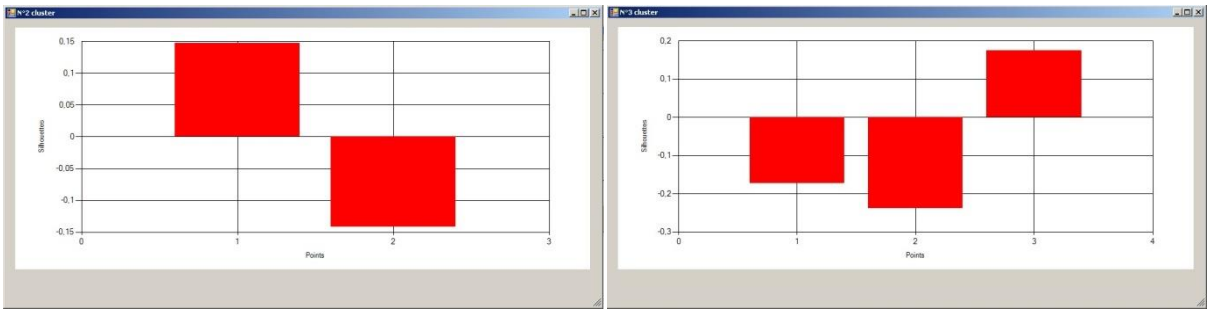
(a) (b)



(c)

Figura 5.1.10: *Si considerano gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e quelli che offrono il terzo miglior valor medio (c).*

Dei *clusters* in esame si è deciso di verificare la *silhouette* media per gli "N" *clusters* considerati e sono stati ottenuti i seguenti andamenti ***(Figura 5.1.11)***:



(a) (b)

(c)

Figura 5.1.11: *Si considerano gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e quelli che offrono il terzo miglior valor medio (c) e per ogni clusters ci cui fanno parte viene mostrata la media.*

Dai grafici ottenuti ***(Figura 5.1.12)*** si nota che con 2 *clusters* i valori di *silhouette* superiore la soglia "0" (1° e 2° grafico) sono 41% e 38%; invero, con l'uso di 3 *clusters* i valori maggiori a "0" risultano essere del 6%, 7% e 44%.

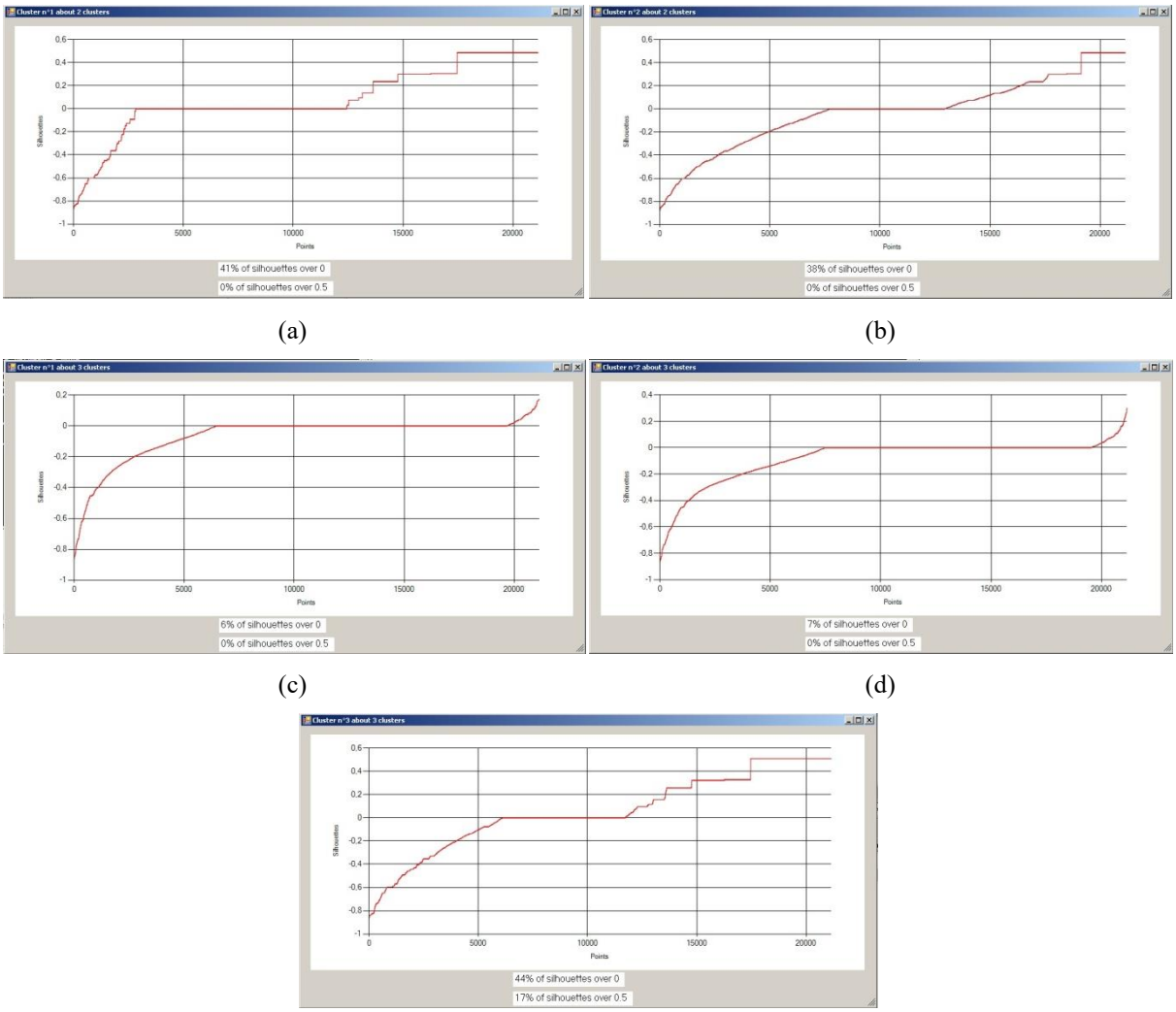


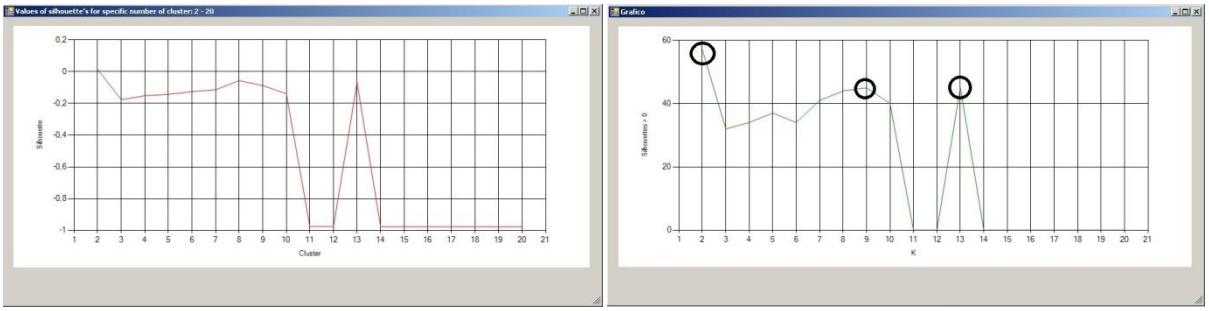
Figura 5.1.12: *Considerando gli "N" clusters con miglior valor medio di silhouette e secondo miglior valor medio di silhouette, è mostrato l'andamento della silhouette per gli N clusters studiati.*

Il miglior valor medio di silhouette (con 2 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a) e (b). Il secondo miglior valor medio (3 clusters) dalle restanti figure (c), (d), (e).

40% (28190 tuple)

Anche in questo caso è stata valutata la *silhouette* per i *clusters* 2-20 ottenendo il grafico ***(Figura 5.1.13)***. I migliori valori di *silhouette* si ottengono con 2 e 9 *clusters*, mentre il terzo miglior valore medio si raggiunge con l'uso di 13 *clusters*. Si noti che dei 2 *clusters* con

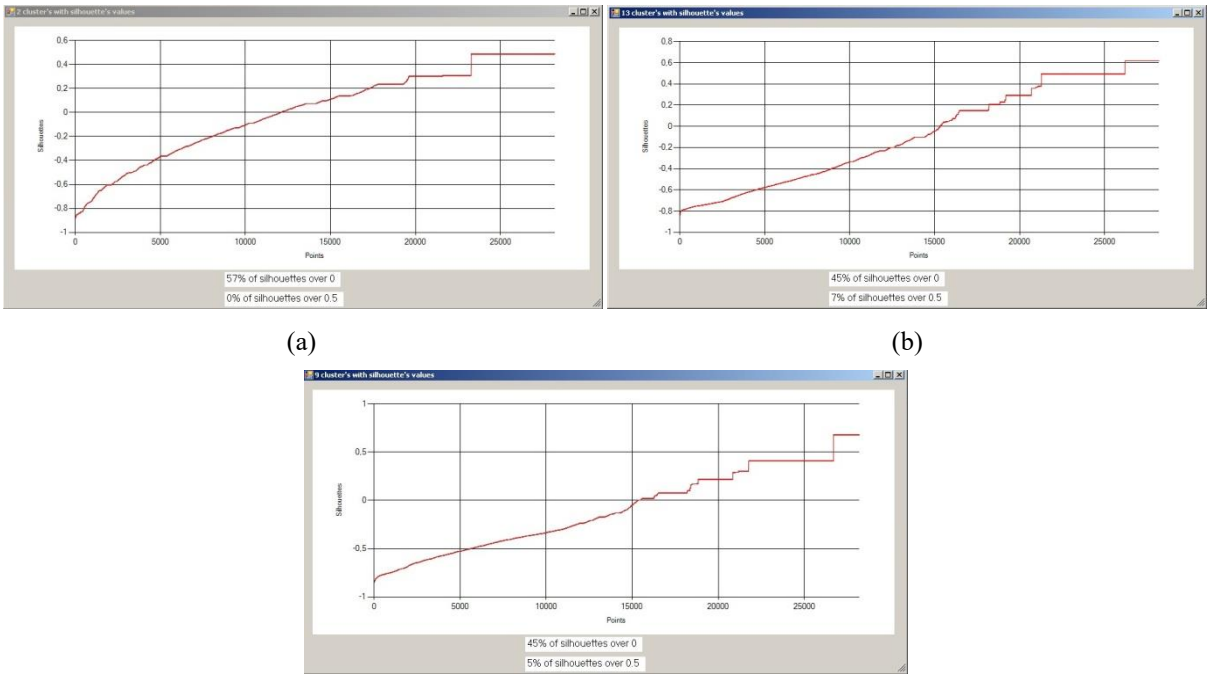
miglior *silhouette* media, solo il primo ha un valore di *silhouette* di poco superiore alla soglia "0".



(a) (b)

Figura 5.1.13: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

Andamento della *silhouette* per "N" cluster (***Figura 5.1.14)***. Con 57%, 45% e 45% dei valori superiori allo "0" si ha un'ulteriore conferma della bontà dell'uso di 2 cluster per avere i valori di *silhouette* migliori. Per 9 e 13 *clusters* si hanno, invero, rispettivamente il 7% e 5% di punti superiori allo 0,5 di *silhouette*.

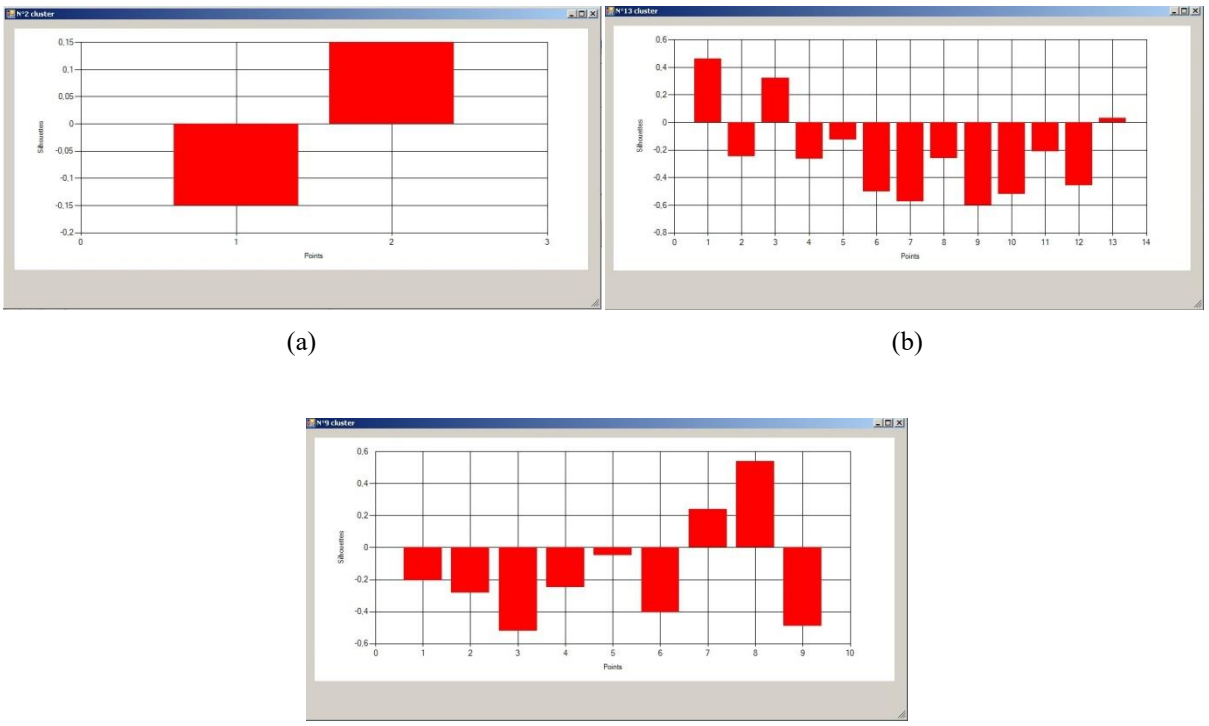


(a) (b)

(c)

Figura 5.1.14: *Sono mostrati gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e quelli che offrono il terzo miglior valor medio (c).*

Dei cluster in esame è stato deciso di verificare la *silhouette* media per gli "N" *clusters* considerati, ottenendo i seguenti andamenti ***(Figura 5.1.15)***:

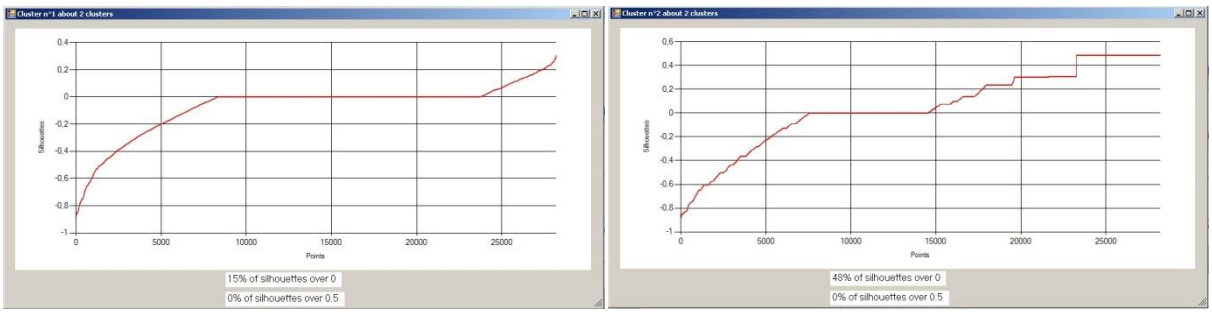


(a) (b)

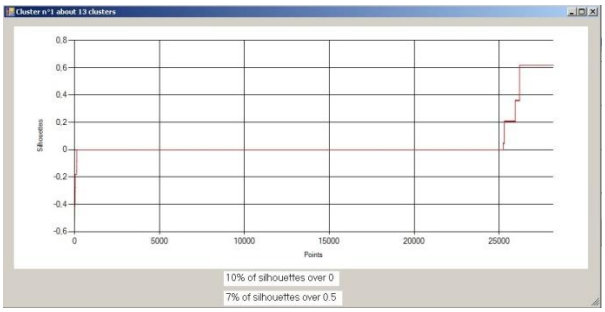
(c)

Figura 5.1.15: *Si considerano gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valor medio (c) e per ogni clusters che ne fanno parte ne viene mostrata la media.*

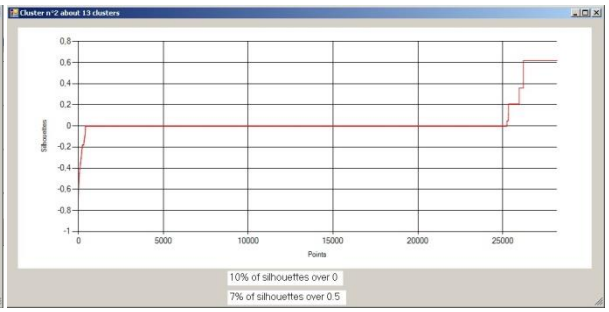
I risultati ottenuti ***(Figura 5.1.16)*** dimostrano che con 2 *clusters* i valori di *silhouette* superiore alla soglia "0" per i *clusters* considerati (1° e 2°) risultano 15% e 48%; con l'uso di 13 *clusters* i valori percentuali sono rispettivamente del 10%, 10%, 35%, 35%, 35%, 33%, 30%, 30%, 28%, 26%, 25%, 23% e 25%.



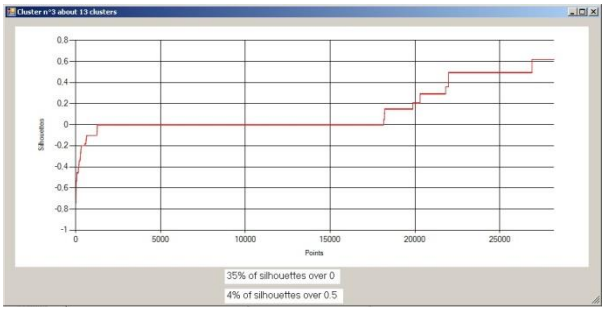
(a) (b)


Cluster n°1 about 13 clusters
Silhouettes
Points
10% of silhouettes over 0
7% of silhouettes over 0.5

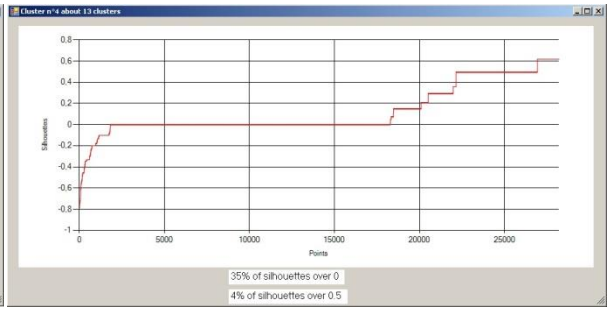
(c)


Cluster n°2 about 13 clusters
Silhouettes
Points
10% of silhouettes over 0
7% of silhouettes over 0.5

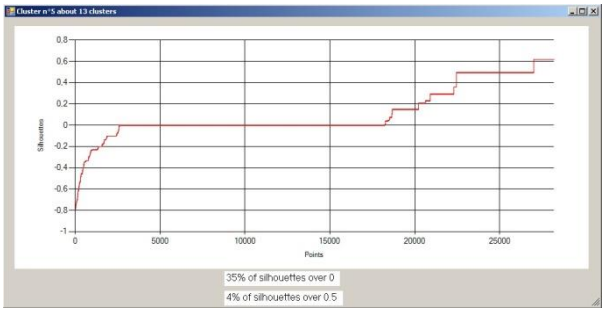
(d)


Cluster n°3 about 13 clusters
Silhouettes
Points
35% of silhouettes over 0
4% of silhouettes over 0.5

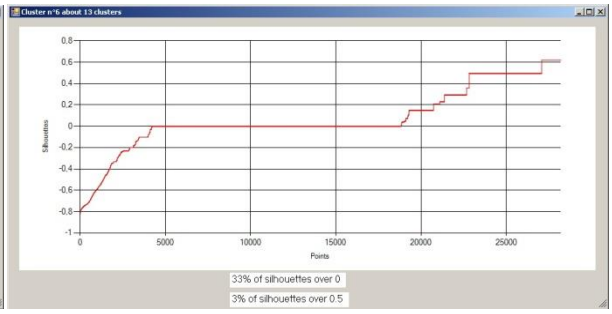
(e)


Cluster n°4 about 13 clusters
Silhouettes
Points
35% of silhouettes over 0
4% of silhouettes over 0.5

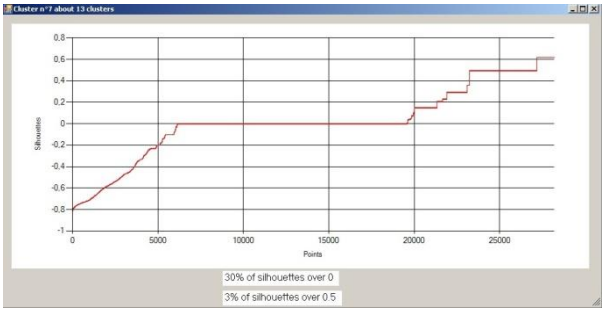
(f)


Cluster n°5 about 13 clusters
Silhouettes
Points
35% of silhouettes over 0
4% of silhouettes over 0.5

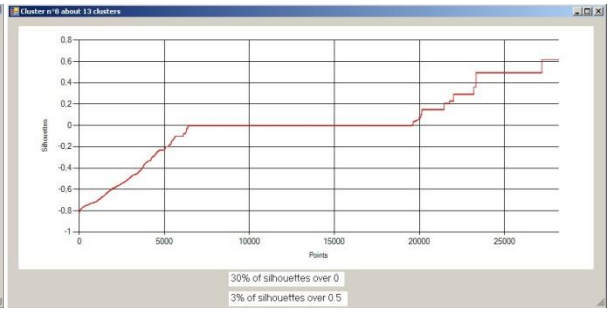
(g)


Cluster n°6 about 13 clusters
Silhouettes
Points
33% of silhouettes over 0
3% of silhouettes over 0.5

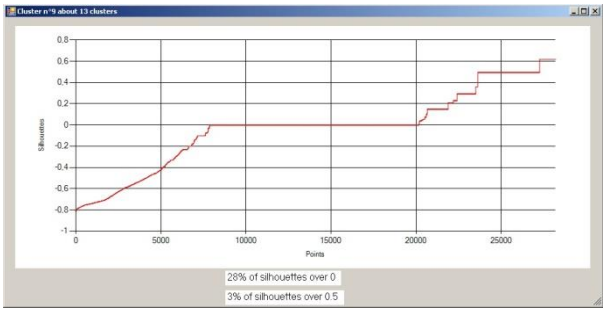
(h)


Cluster n°7 about 13 clusters
Silhouettes
Points
30% of silhouettes over 0
3% of silhouettes over 0.5

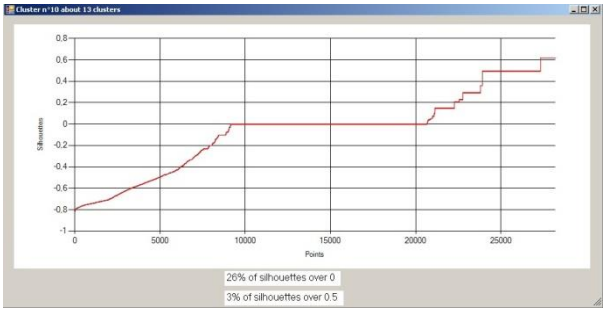
(i)


Cluster n°8 about 13 clusters
Silhouettes
Points
30% of silhouettes over 0
3% of silhouettes over 0.5

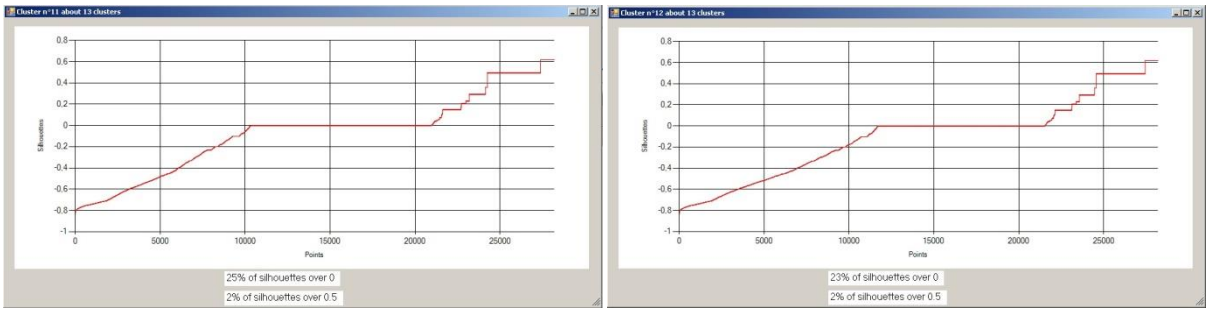
(j)


Cluster n°9 about 13 clusters
Silhouettes
Points
28% of silhouettes over 0
3% of silhouettes over 0.5

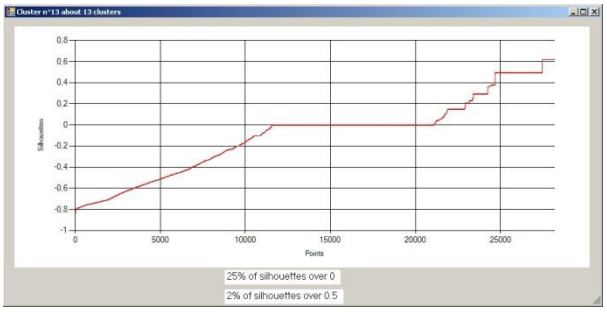
(k)


Cluster n°10 about 13 clusters
Silhouettes
Points
26% of silhouettes over 0
3% of silhouettes over 0.5

(l)



(m) (n)



(o)

Figura 5.1.16: *Valutando gli "N" clusters con miglior valore medio di silhouette e secondo migliore valore medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

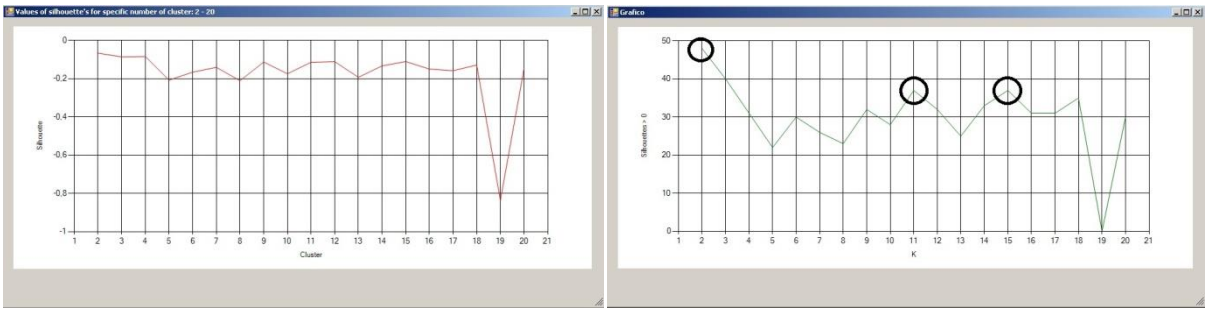
Il migliore valore medio di silhouette (con 2 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a) e (b). Il secondo migliore valore medio (13 clusters) dalle restanti figure (c), (d), (e), (f), (g), (h), (i), (j), (k), (l), (m), (n), (o).

5.2 Turkiye Student Evaluation Data Set

Questo *dataset* contiene circa 5800 righe *(5820)* e si è deciso di considerare il 70%, 80%, 90% e 100%.

70% (4074 tuple)

Si è deciso di considerare l'andamento della *silhouette* media osservando i *clusters* che vanno da 2-20 ***(Figura 5.2.1)***. In tal caso si nota che il picco della *silhouette* è in corrispondenza di 2 *clusters*. A fronte di ciò si è deciso di verificare l'andamento della *silhouette* media per i 2 *clusters* che offrono la *silhouette* migliore *(2 e 11 clusters)* e quello con il terzo miglior andamento *(in questo caso rappresentato dall'utilizzo di 15 clusters).*

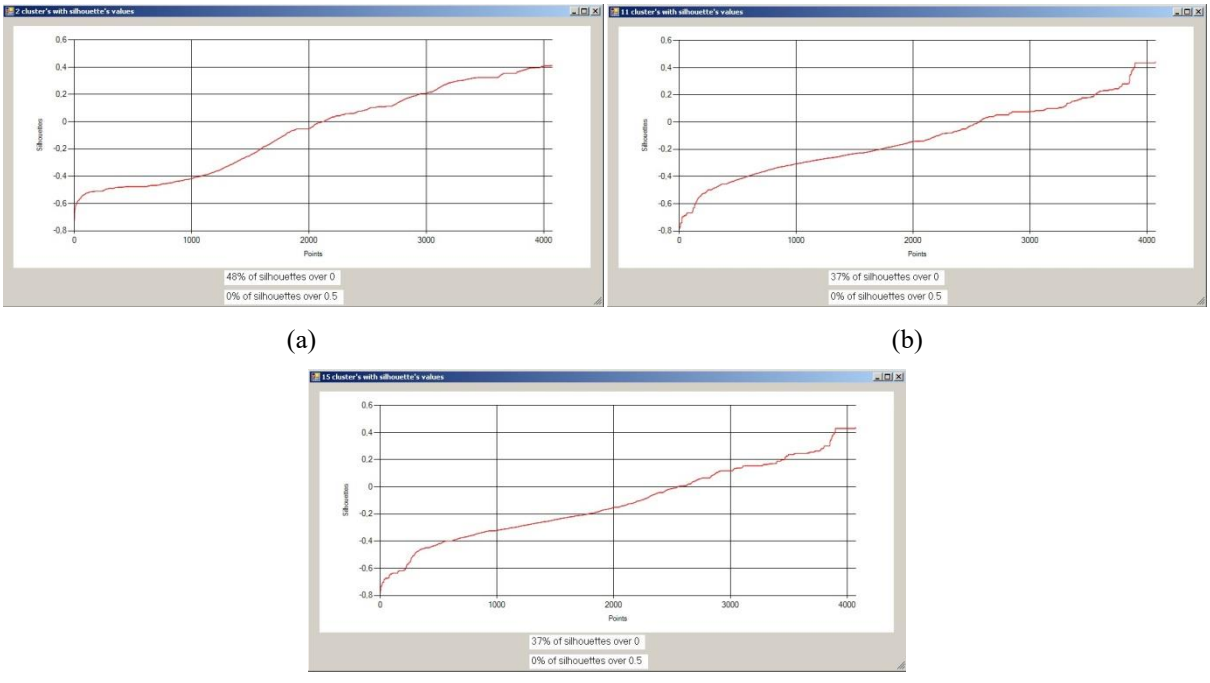


(a) (b)

Figura 5.2.1: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli N clusters con silhouette superiori alla soglia 0 (b).*

Esaminando i grafici prodotti dall'applicazione ***(Figura 5.2.2)*** si osserva l'andamento della *silhouette* per i punti considerati; sotto, invero, è possibile esaminare le percentuali attinenti il numero di punti superiori al valore 0 e 0,5 di *silhouette*. Nel caso specifico si mostra:

1. **2 cluster** (*1° miglior valore di silhouette*): 48% dei punti superiore la soglia 0 e nessun punto superiore allo 0,5;
2. **11 cluster** (*2° miglior valore di silhouette*): 37% dei punti superiore a 0 e nessun punto superiore allo 0,5;
3. **15 cluster** (*peggior valore di silhouette*): 37% dei punti superiore a 0 e nessun punto superiore allo 0,5.

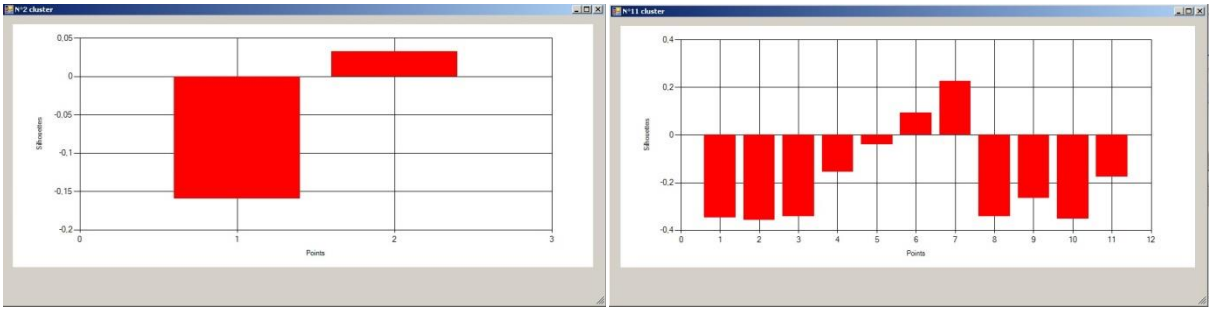


(a) (b)

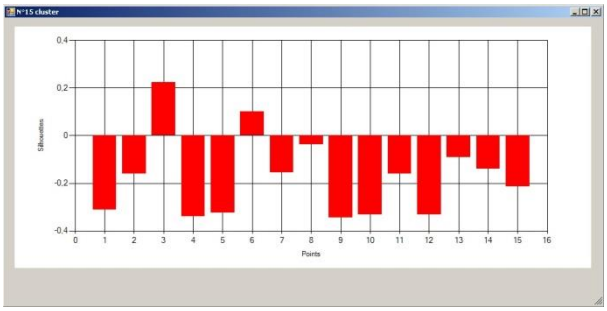
(c)

Figura 5.2.2: *Si considerano gli "N" clusters che offrono il migliore valore medio di silhouette (a), gli "N" clusters che offrono il secondo migliore valore medio di silhouette (b) e gli "N" clusters che offrono il terzo migliore valore medio (c).*

Dei *clusters* in esame si è deciso di verificare la *silhouette* media per gli "N" *clusters* considerati ottenendo i seguenti andamenti ***(Figura 5.2.3)***:



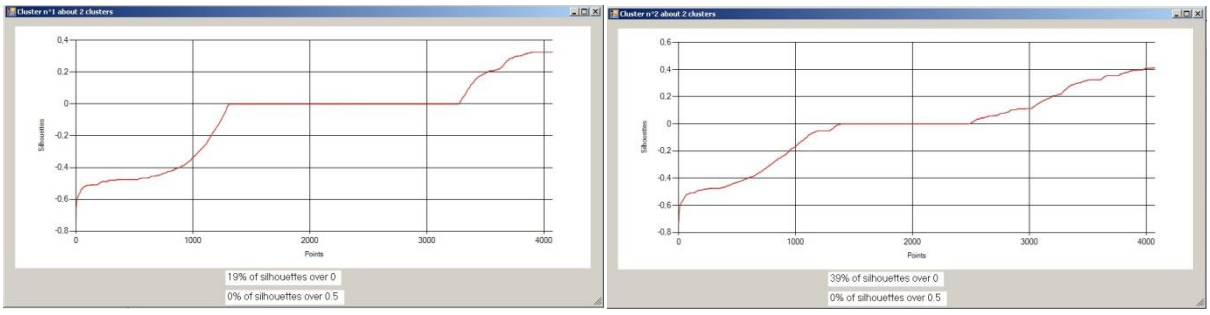
(a) (b)



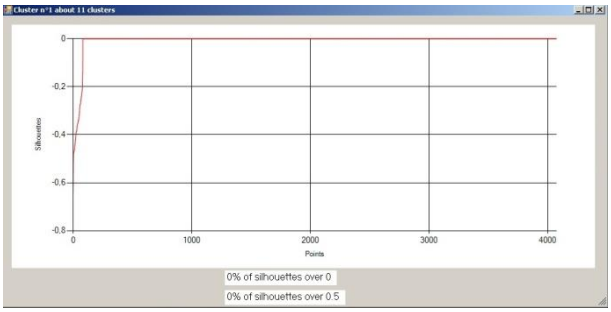
(c)

Figura 5.2.3: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c); per ogni clusters che ne fanno parte viene mostrata la media.*

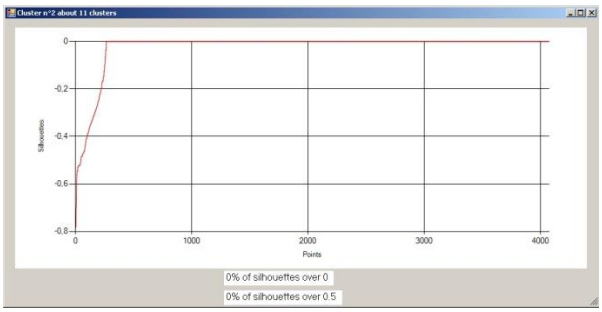
Dai grafici ottenuti ***(Figura 5.2.4)*** si può costatare come con 2 *clusters* i valori di *silhouette* superiore a "0" (1° e 2°) siano 19% e 39%; con l'uso di 11 *clusters* i valori maggiori di "0" sono (*dal 4° grafico in poi*) del 3%, 14%, 22%, 29%, 27%, 25%, 22% e 21%.



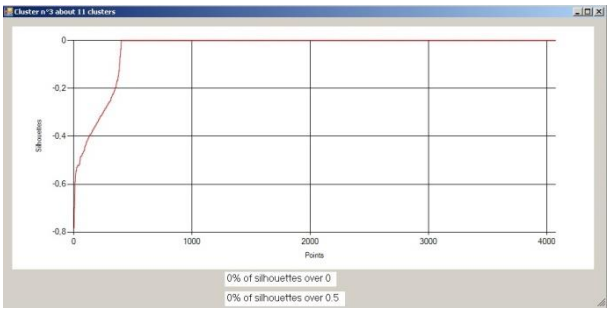
(a) (b)


Cluster n°1 about 11 clusters
Silhouettes
Points
0% of silhouettes over 0
0% of silhouettes over 0.5

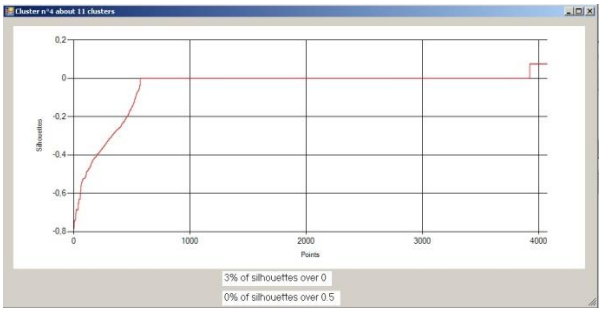
(c)


Cluster n°2 about 11 clusters
Silhouettes
Points
0% of silhouettes over 0
0% of silhouettes over 0.5

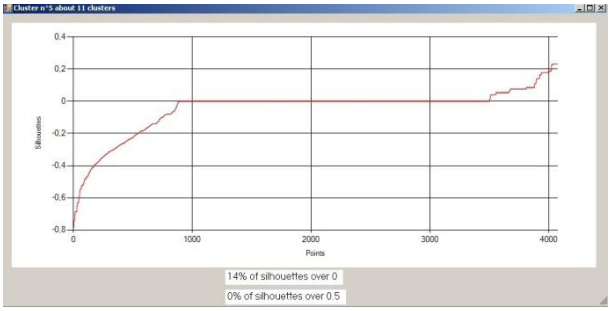
(d)


Cluster n°3 about 11 clusters
Silhouettes
Points
0% of silhouettes over 0
0% of silhouettes over 0.5

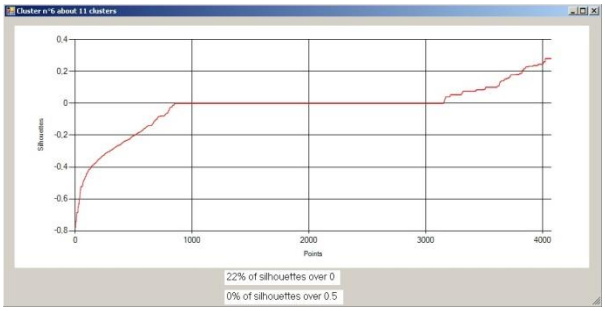
(e)


Cluster n°4 about 11 clusters
Silhouettes
Points
3% of silhouettes over 0
0% of silhouettes over 0.5

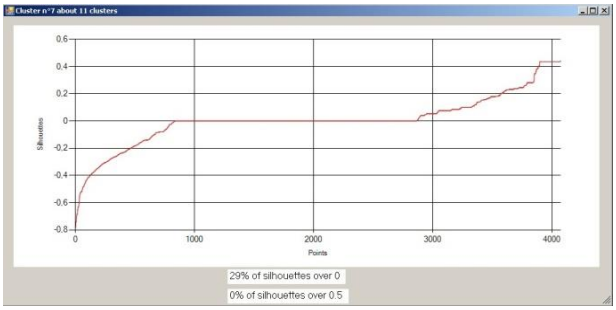
(f)


Cluster n°5 about 11 clusters
Silhouettes
Points
14% of silhouettes over 0
0% of silhouettes over 0.5

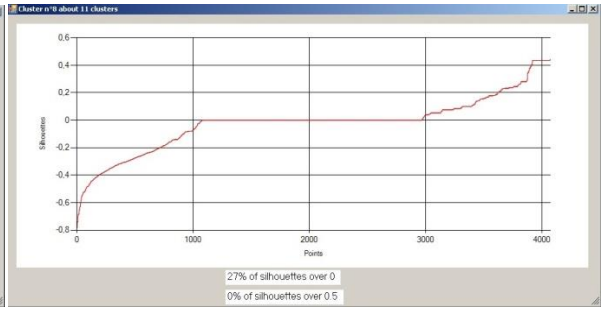
(g)


Cluster n°6 about 11 clusters
Silhouettes
Points
22% of silhouettes over 0
0% of silhouettes over 0.5

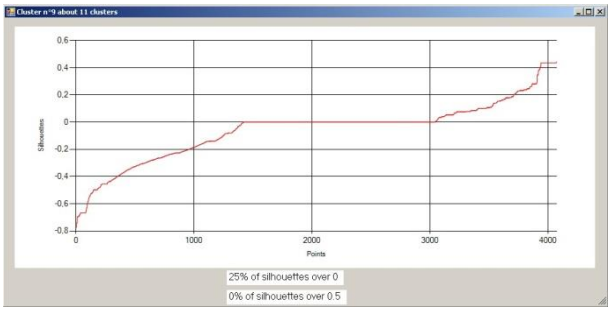
(h)


Cluster n°7 about 11 clusters
Silhouettes
Points
29% of silhouettes over 0
0% of silhouettes over 0.5

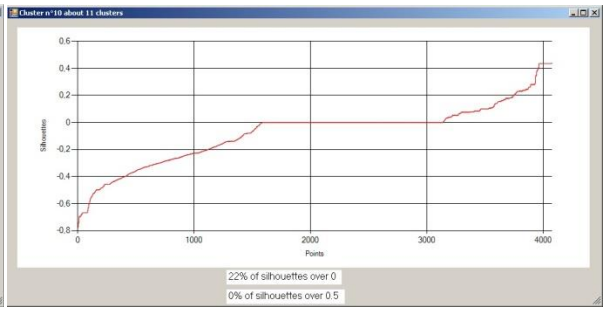
(i)


Cluster n°8 about 11 clusters
Silhouettes
Points
27% of silhouettes over 0
0% of silhouettes over 0.5

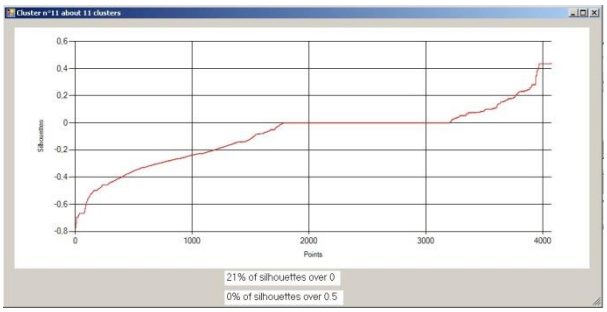
(j)


Cluster n°9 about 11 clusters
Silhouettes
Points
25% of silhouettes over 0
0% of silhouettes over 0.5

(k)


Cluster n°10 about 11 clusters
Silhouettes
Points
22% of silhouettes over 0
0% of silhouettes over 0.5

(l)



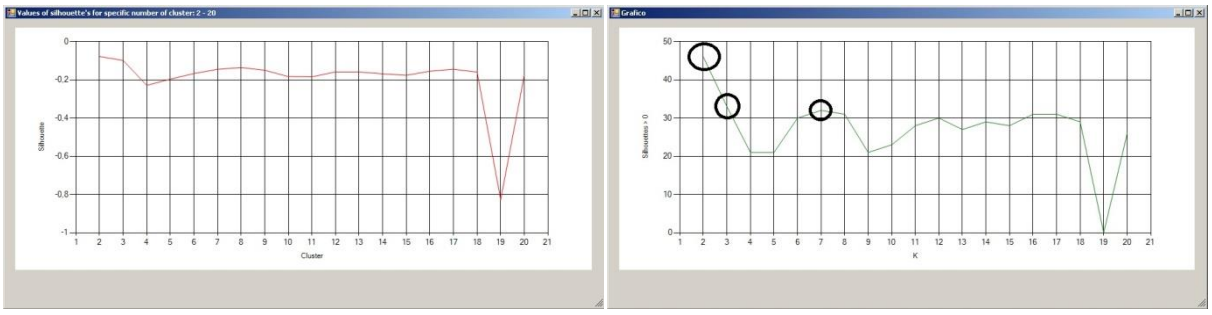
(m)

Figura 5.2.4: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

Il miglior valore medio di silhouette (con 2 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a) e (b). Il secondo miglior valore medio (3 clusters) dalle restanti figure (c), (d), (e), (f), (g), (h), (i), (j), (k), (l), (m).

80% (4656 tuple)

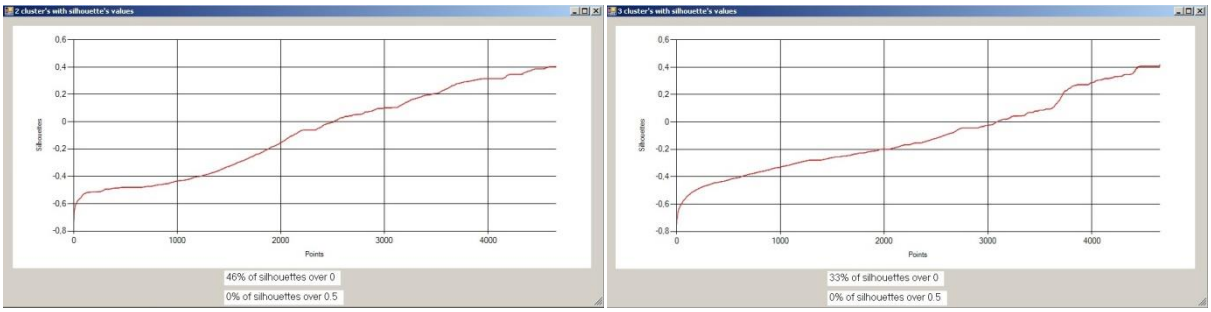
Sarà valutato l'andamento della *silhouette* media considerando i *clusters* che vanno da 2-20 ***(Figura 5.2.5)***. I migliori valori di *silhouette* si raggiungono con 2 e 3 *clusters*, mentre il terzo miglior valore si ottiene con l'uso di 7 *clusters*.



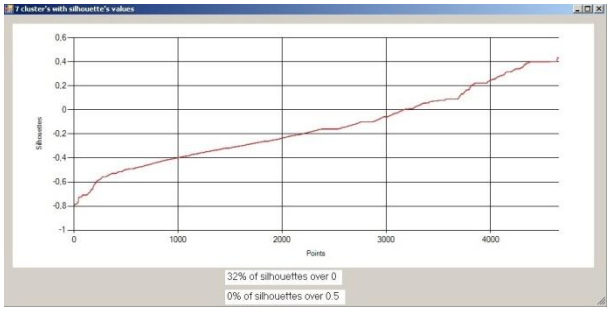
(a) (b)

Figura 5.2.5: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli N clusters con silhouette superiori alla soglia 0 (b).*

L'andamento della *silhouette* per "N" cluster è mostrato in ***Figura 5.2.6***. Con 46%, 33% e 32% dei valori superiori alla soglia 0 si ha un'ulteriore conferma della bontà sull'uso di 2 *clusters* per ottenere i valori di *silhouette* migliori.



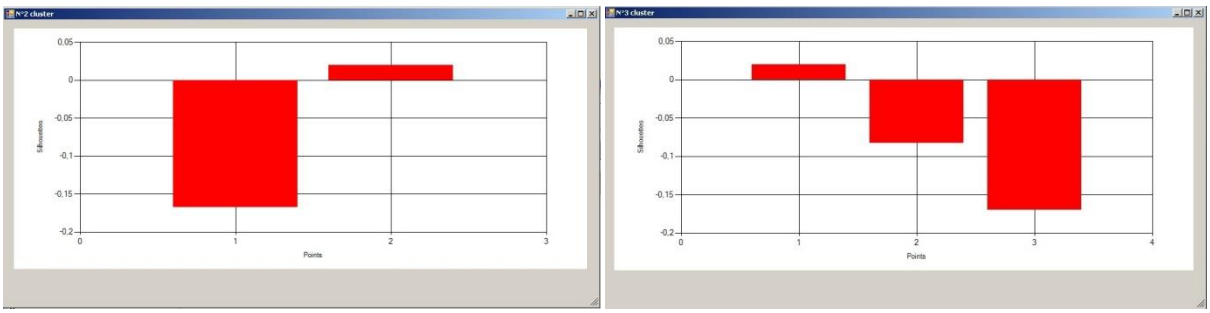
(a) (b)



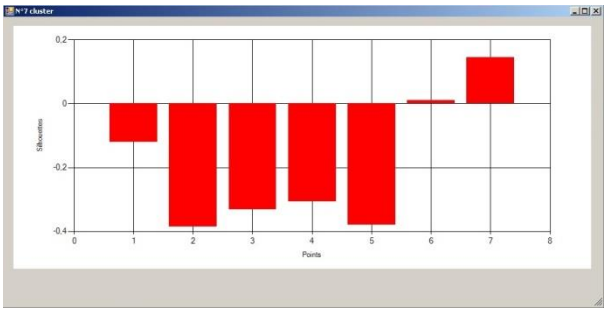
(c)

Figura 5.2.6: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c).*

Dai cluster in esame si è scelto di verificare la *silhouette* media per gli "N" *clusters* considerati ottenendo i seguenti andamenti ***(Figura 5.2.7)***:



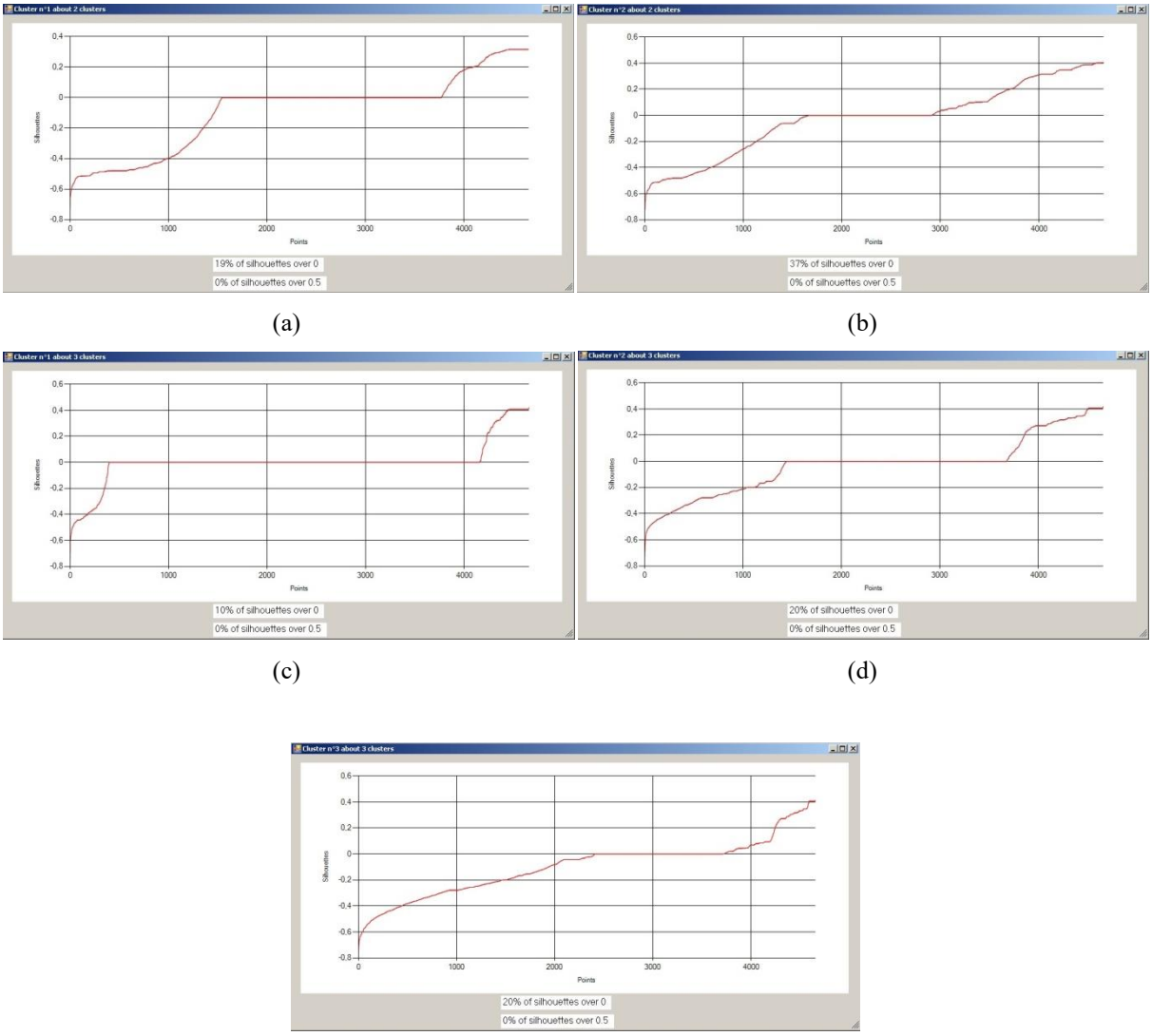
(a) (b)



(c)

Figura 5.2.7: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c). Per ogni clusters che ne fanno parte ne viene mostrata la media.*

Dai risultati ottenuti si nota come, con 2 *clusters*, i valori di *silhouette* superiore la soglia "0", per i *clusters* considerati (1° e 2°), siano 19% e 37%; con l'uso di 3 *clusters* i valori maggiori di 0 risultano del 10%, 20%, 20%.



(a) (b)

(c) (d)

(e)

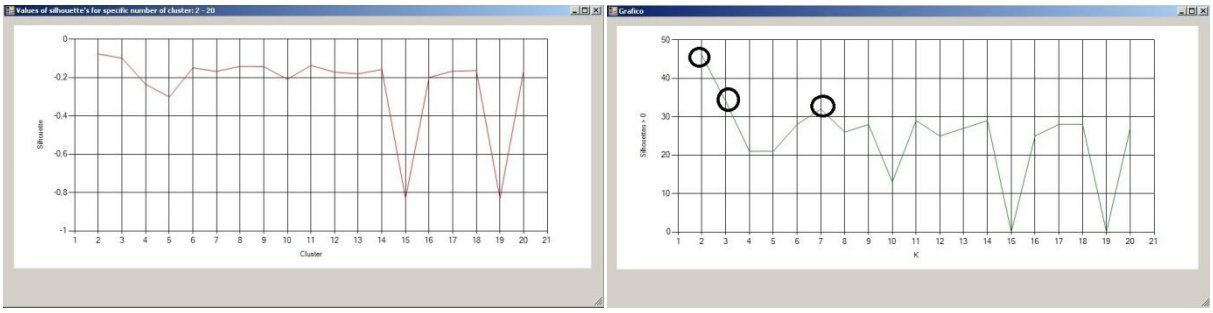
Figura 5.2.8: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valor medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

Il miglior valor medio di silhouette (con 2 clusters) è rappresentato dagli andamenti rappresentati nelle figure

(a) e (b). Il secondo miglior valore medio (3 clusters) dalle restanti figure (c), (d) e (e).

90% (5238 tuple)

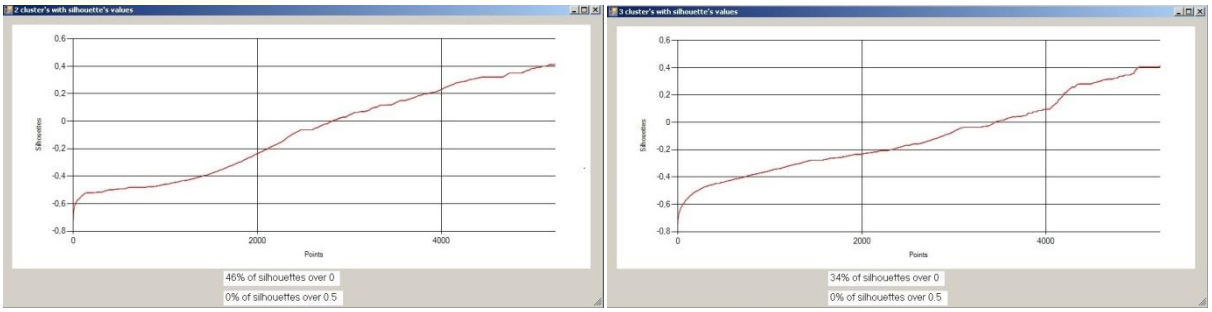
In questo caso si considera la *silhouette* per i *clusters* da 2-20 ***(Figura 5.2.9)***. I migliori valori di *silhouette* si ottengono con 2 e 3 *clusters* mentre il terzo miglior valore si ottiene con 7 *clusters*.



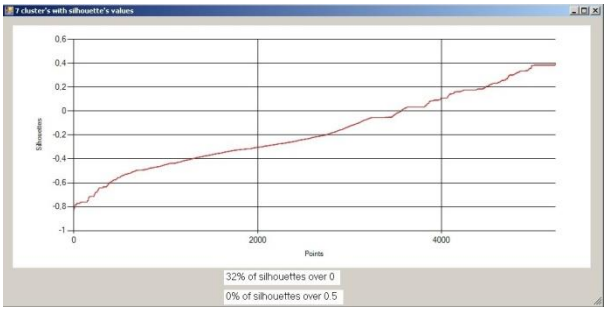
(a) (b)

Figura 5.2.9: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

L'andamento della *silhouette* per "N" cluster è mostrato nella ***Figura 5.2.10 .*** Con 46%, 34% e 32% dei valori superiori allo "0" si ha un'altra conferma della bontà sull'uso di 2 *clusters* per avere i valori di *silhouette* migliori.



(a) (b)

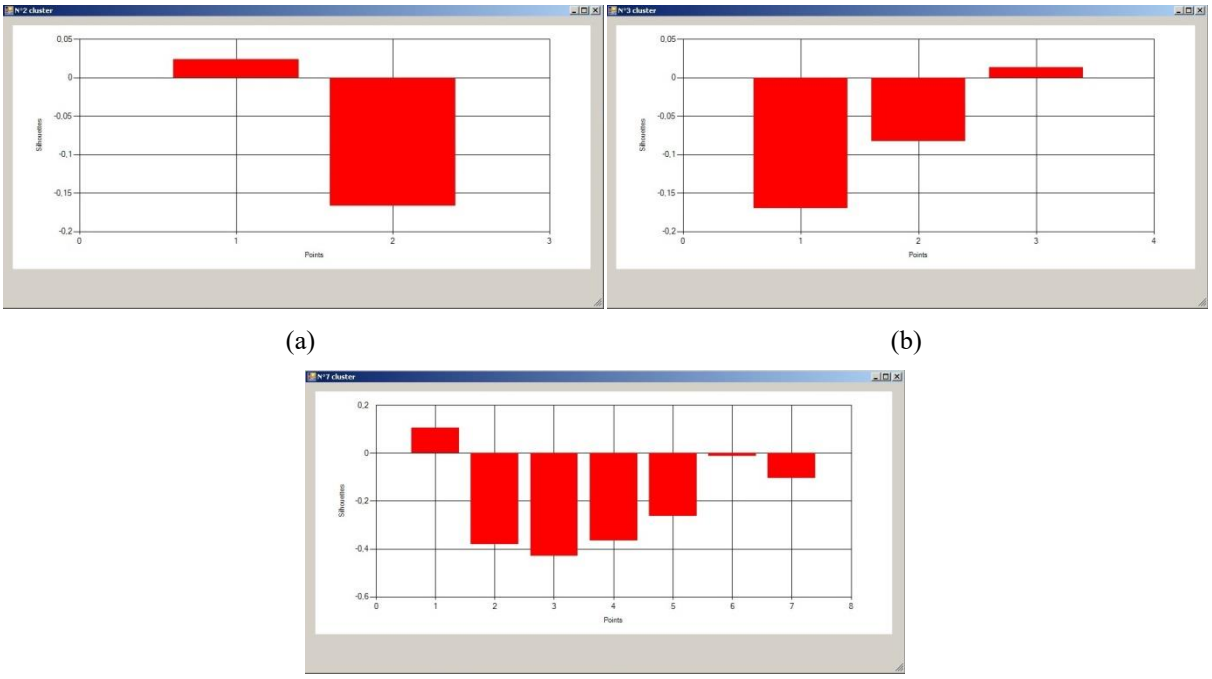


(c)

Figura 5.2.10: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N"*

clusters con il secondo miglior valore medio di silhouette (b) e gli "N" clusters con il terzo miglior valore medio (c).

Dai cluster in esame si è scelto di verificare la *silhouette* media per gli "N" *clusters* considerati ottenendo i seguenti andamenti ***(Figura 5.2.11)***:

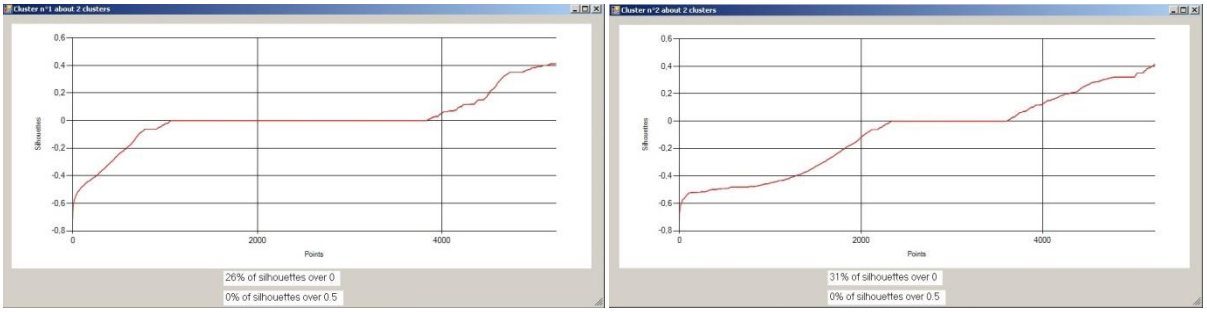


(a) (b)

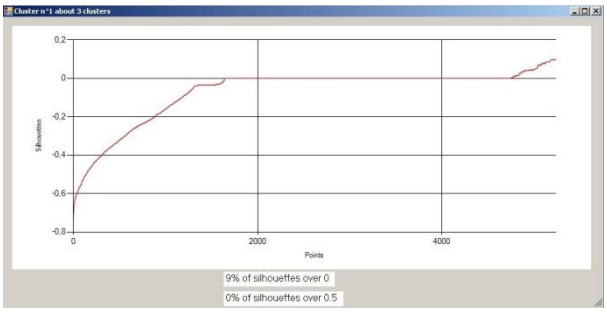
(c)

Figura 5.2.11: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters con il secondo miglior valore medio di silhouette (b) e gli "N" clusters con il terzo miglior valore medio (c). Per ogni clusters ne viene mostrata la media.*

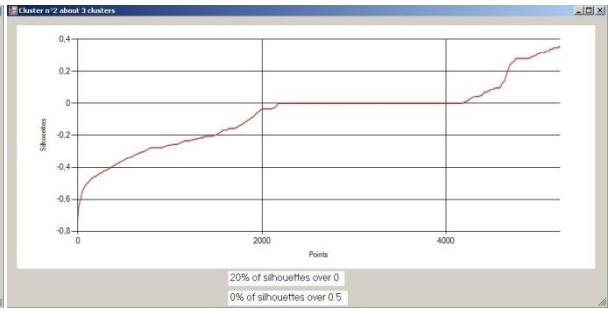
Dai risultati raggiunti ***(Figura 5.2.12)*** si può costatare come con 2 *clusters* i valori di *silhouette* superiore la soglia "0" (1° e 2°) risultano 26% e 31%; con l'uso di 3 *clusters* i valori maggiori di 0 risultano del 9%, 20%, 27%.



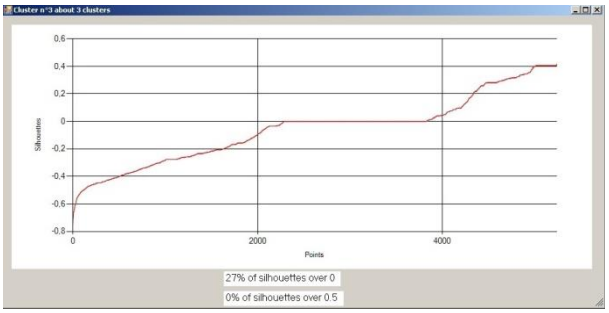
(a) (b)



(c)



(d)



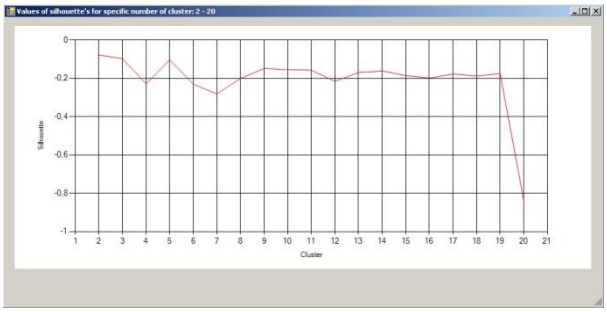
(e)

Figura 5.2.12: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

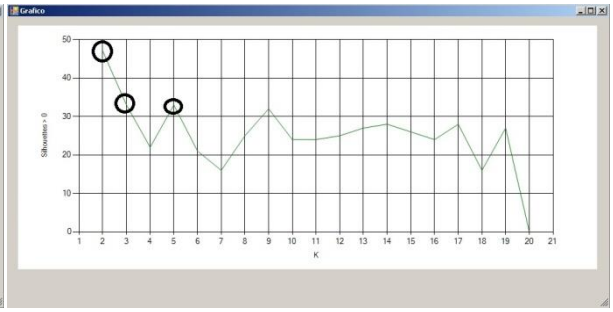
Il miglior valor medio di silhouette (con 2 clusters) è mostrato dagli andamenti rappresentati nelle figure (a) e (b). Il secondo miglior valore medio (3 clusters) dalle restanti figure (c), (d) ed (e).

100% (5820 tuple)

E' stato considerato l'andamento della *silhouette* media utilizzando i *clusters* da 2-20 ***(Figura 5.2.13)***. I migliori valori di *silhouette* si ottengono con 2 e 3 *clusters*, mentre il terzo miglior valore lo ottiene con 5 *clusters*.



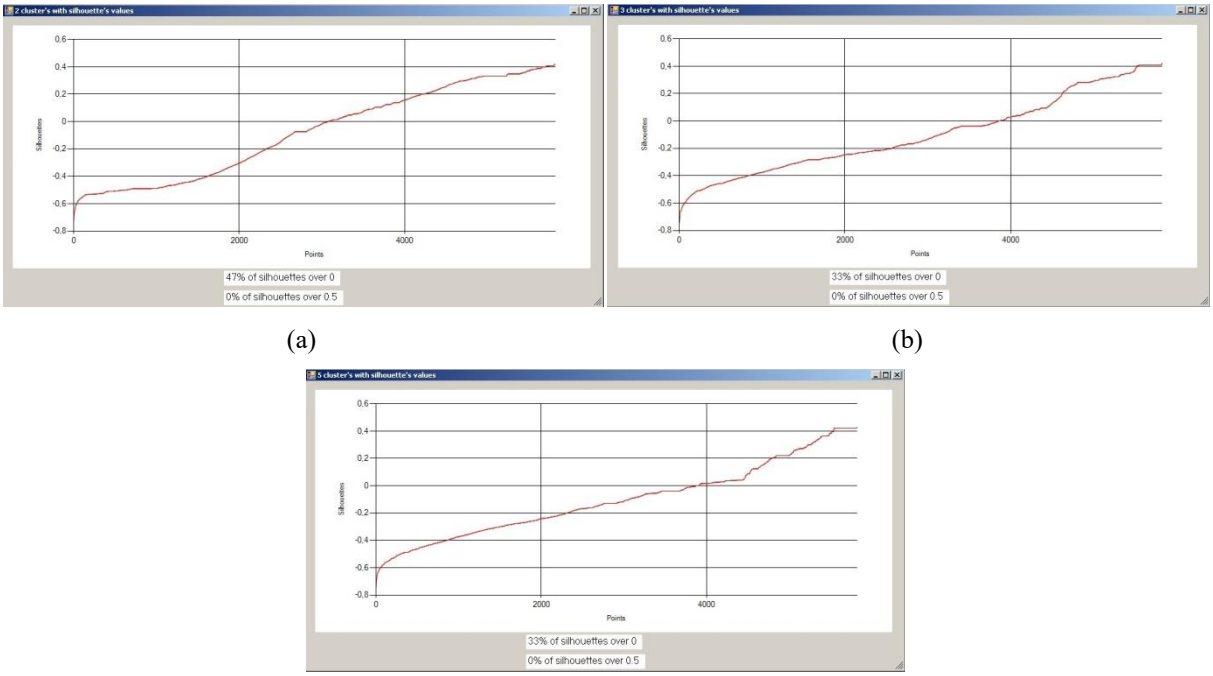
(a)



(b)

Figura 5.2.13: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

L'andamento della *silhouette* per "N" cluster è mostrato in ***Figura 5.2.14***. Con 47% e 33%, rispettivamente per 2 e 3 *clusters*, si ha la conferma che per avere migliori valori di *silhouette* si debba considerare l'uso di 2 *clusters*.

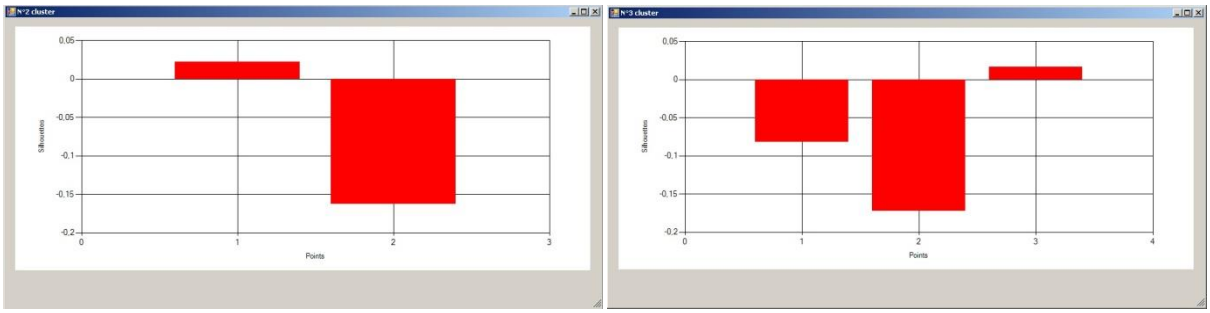


(a) (b)

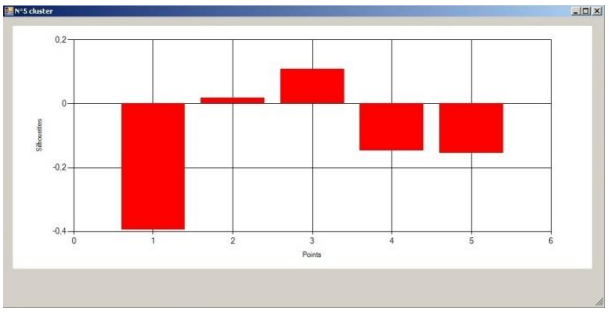
(c)

Figura 5.2.14: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c).*

Dai cluster in esame si è deciso di verificare la *silhouette* media per gli "N" *clusters* considerati ottenendo i seguenti andamenti ***(Figura 5.2.15)***:



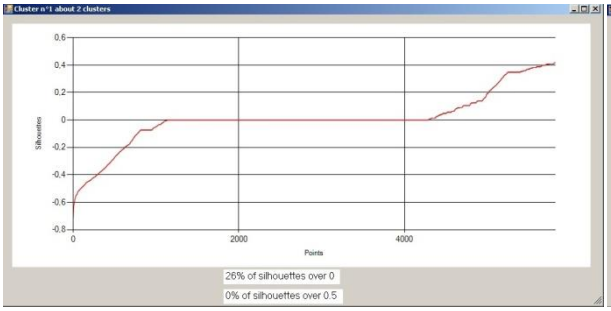
(a) (b)



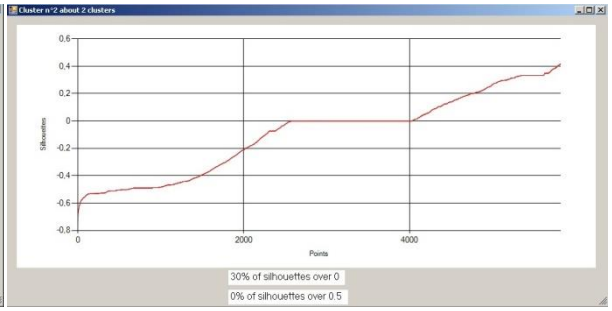
(c)

Figura 5.2.15: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c). Per ogni clusters che ne fanno parte ne viene mostrata la media.*

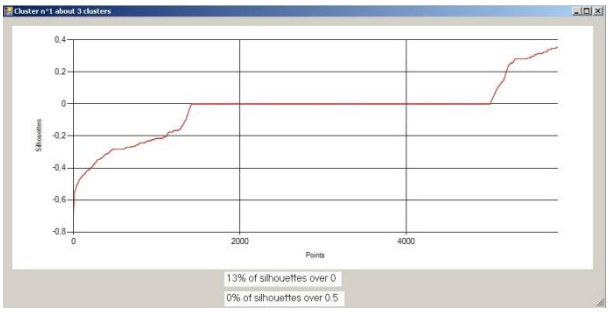
I risultati ottenuti ***(Figura 5.2.16)*** dimostrano che con 2 *clusters* i valori di *silhouette* superiore la soglia "0" (1° e 2°) risultano 26% e 30%; con l'uso di 3 *clusters* i valori maggiori di 0 risultano essere del 13%, 16%, 23%.



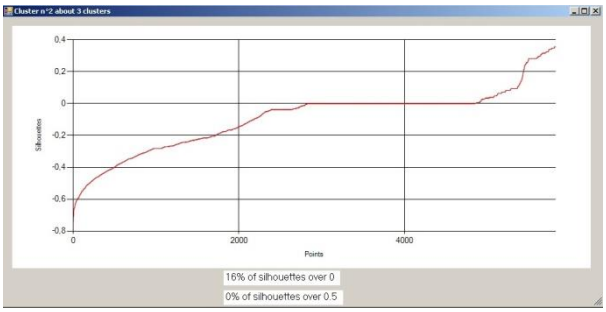
(a)



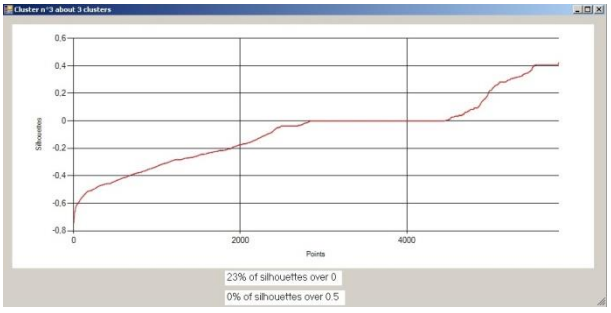
(b)



(c)



(d)



(e)

Figura 5.2.16: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

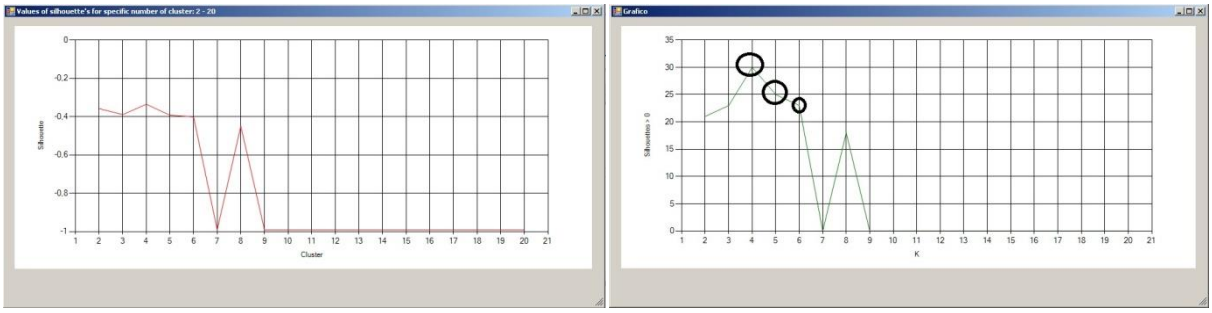
Il miglior valore medio di silhouette (con 2 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a) e (b). Il secondo miglior valore medio (3 clusters) dalle restanti figure (c), (d) ed (e).

5.3 Electricity Data Set

Tale *dataset* contiene circa 46000 righe *(45781)* e si è deciso di considerare il 30%, 50%, 70% e 80%.

30% (13761 tuple)

Sarà esaminato l'andamento della *silhouette* media considerando i *clusters* da 2-20 ***(Figura 5.3.1)***. In questo caso si nota che il picco della *silhouette* è in corrispondenza di 4 *clusters*. A fronte di ciò si è deciso di verificare l'andamento della percentuale di *silhouette* superiore alla soglia "0" per i 2 cluster che offrono i risultati migliori *(4 clusters e 5 clusters)* e quello con il terzo miglior andamento *(in questo caso rappresentato dall'utilizzo di 6 clusters)*.



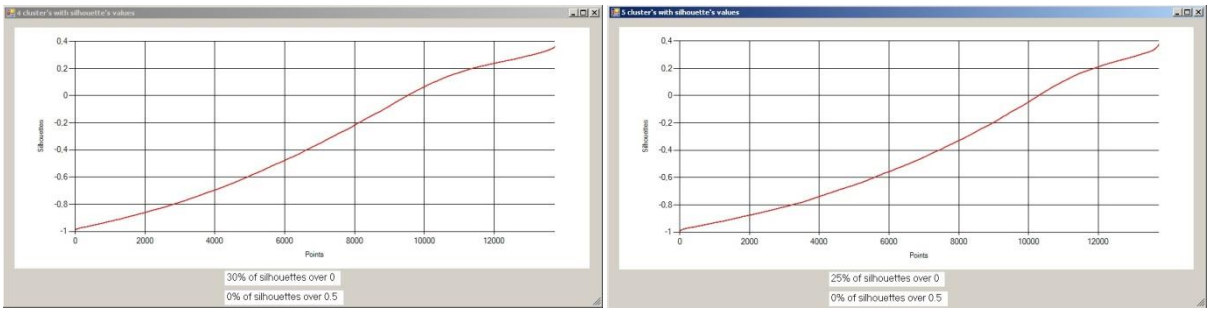
(a) (b)

Figura 5.3.1: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

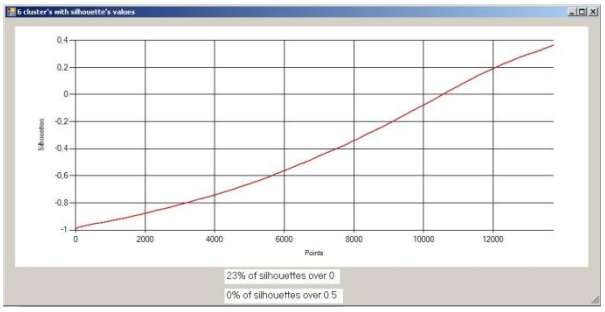
Osservando i grafici ***(Figura 5.3.2)*** prodotti dall'applicazione si comprende quale sia l'andamento della *silhouette* per i punti considerati.

Al di sotto del grafico sono indicate le percentuali relative il numero di punti superiori ai valori 0 e 0,5 di *silhouette*. Nel caso specifico:

1. **4 cluster** *(1° miglior valore di silhouette):* 30% dei punti superiore a 0 e nessun punto superiore allo 0,5;
2. **5 cluster** *(2° miglior valore di silhouette):* 25% dei punti superiore a 0 e nessun punto superiore allo 0,5;
3. **6 cluster** *(peggior valore di silhouette):* 23% dei punti superiore a 0 e nessun punto superiore allo 0,5;



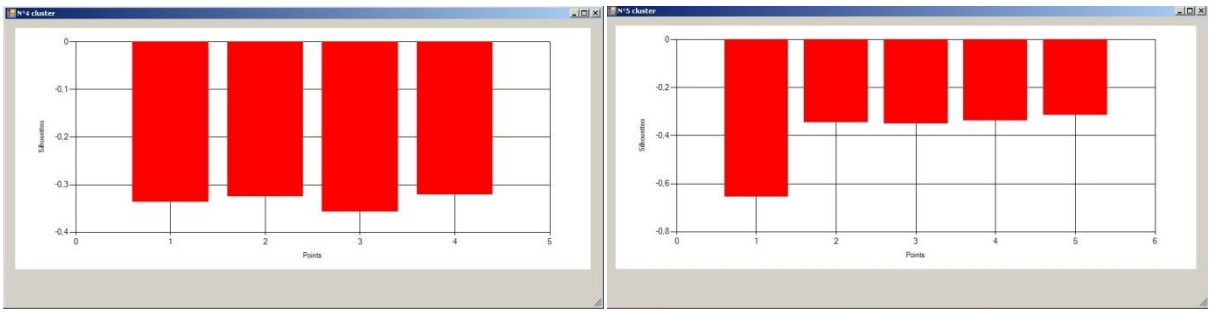
(a) (b)



(c)

Figura 5.3.2: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c).*

Dai *clusters* in esame è stata analizzata la *silhouette* media per gli "N" *clusters* considerati ottenendo i seguenti andamenti ***(Figura 5.3.3)***:



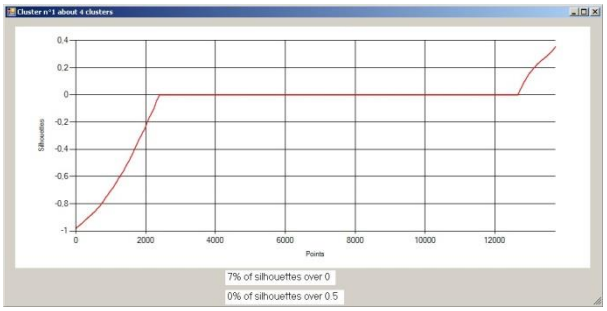
(a) (b)

(c)

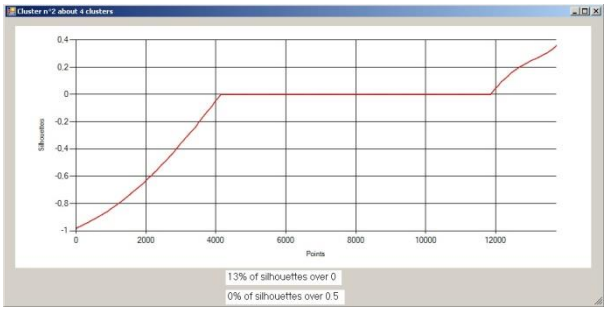
Figura 5.3.3: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c). Per ogni clusters che ne fanno parte viene mostrata la media.*

Considerando l'uso di 4 *clusters* ***(Figura 5.3.4)*** *(e mostrandone l'andamento della silhouette per ogni "N" clusters che ne fa parte)* si percepisce come i grafici mostrino rispettivamente il 7%, 13%, 17% e 20% dei punti superiori alla soglia "0" di *silhouette*.

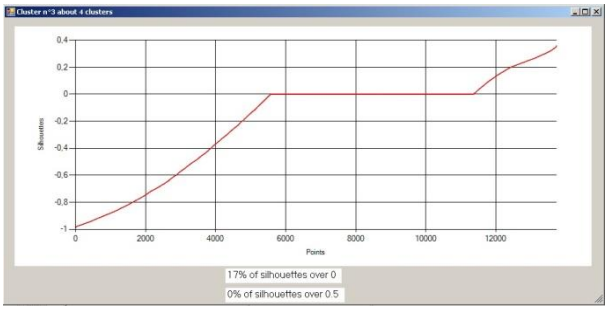
E' stato valutato, poi, l'uso di 5 *clusters* e si è potuto constatare che tranne che per il 1° grafico, ove la percentuale di valori che superano la soglia "0" di *silhouette* è pari a "0", il 2°, 3°, 4°, 5° grafico hanno un numero di valori di *silhouette* superiori alla soglia "0" ovvero il 5%, 9%, 14%, 18%.



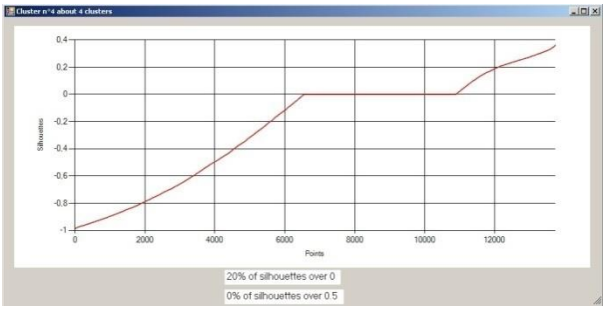
(a)



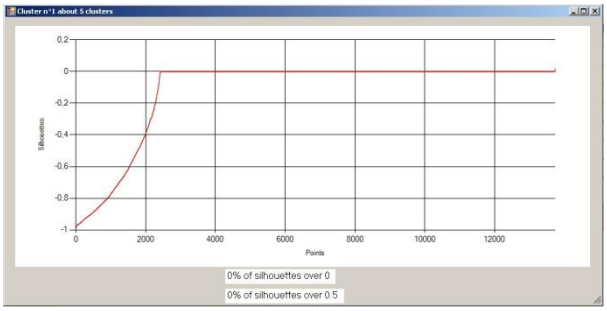
(b)



(c)

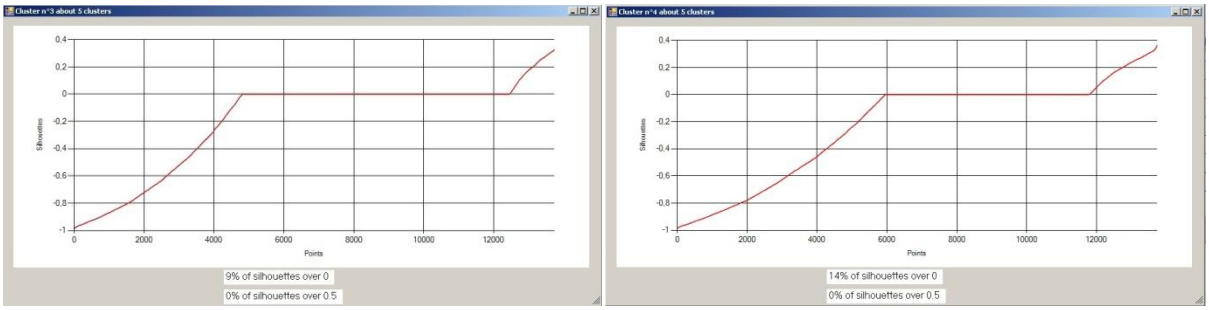


(d)

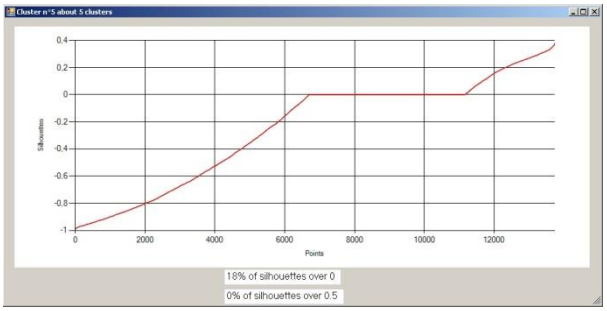


(e)

(f)



(g) (h)



(i)

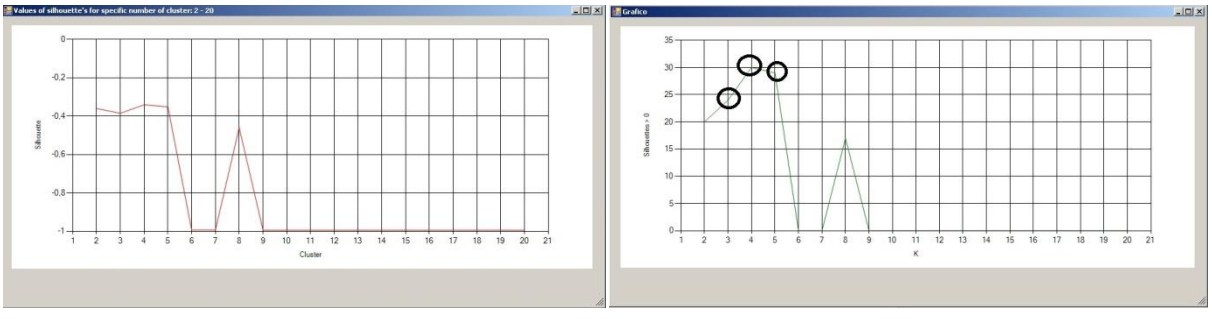
Figura 5.3.4: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

Il miglior valore medio di silhouette (con 4 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a), (b), (c), (d). Il secondo miglior valore medio (5 clusters) dalle restanti figure (e) ,(f), (g), (h), (i).

50% (22890 tuple)

Si consideri la *silhouette* per i *clusters* da 2-20 ***(Figura 5.3.5)***.

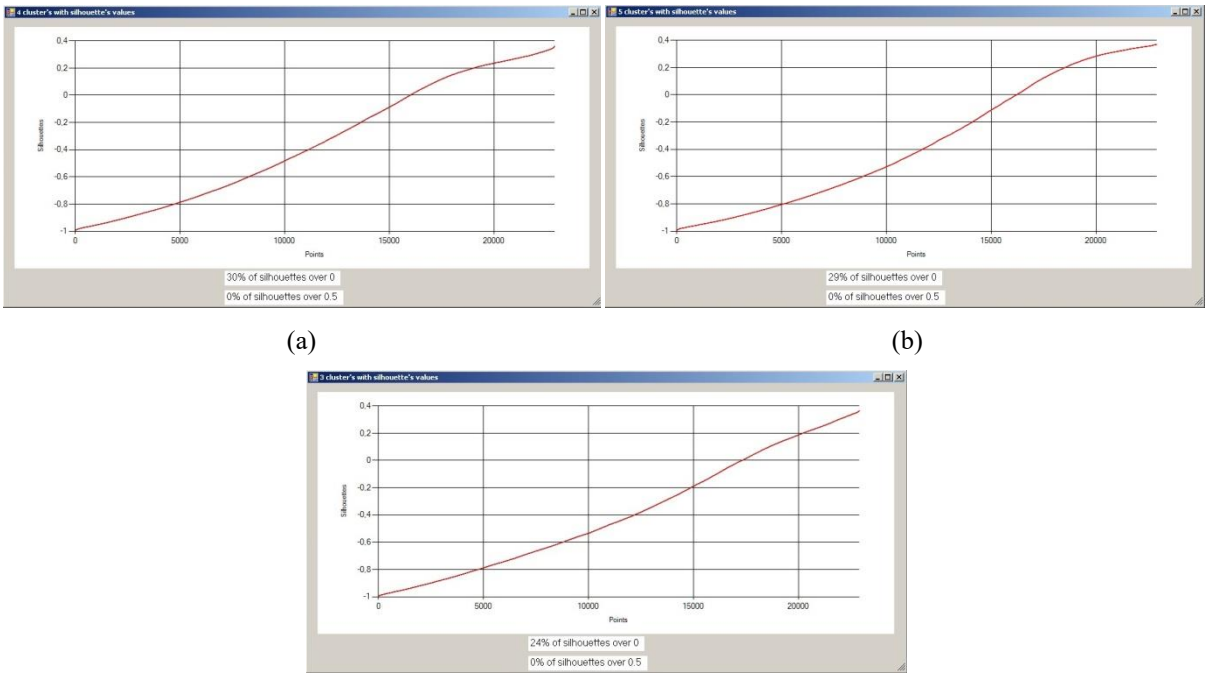
I migliori valori si ottengono con 4 e 5 *clusters* mentre il terzo miglior valore medio si raggiunge con l'uso di 3 *clusters*.



(a) (b)

Figura 5.3.5: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

L'andamento della *silhouette* per "N" cluster è mostrato in ***Figura 5.3.6***. Si ottengono il 30%, 29% e 24% dei valori di *silhouette* superio allo "0". Anche in questo caso non si ha alcun valore di *silhouette* superiore lo 0,5.

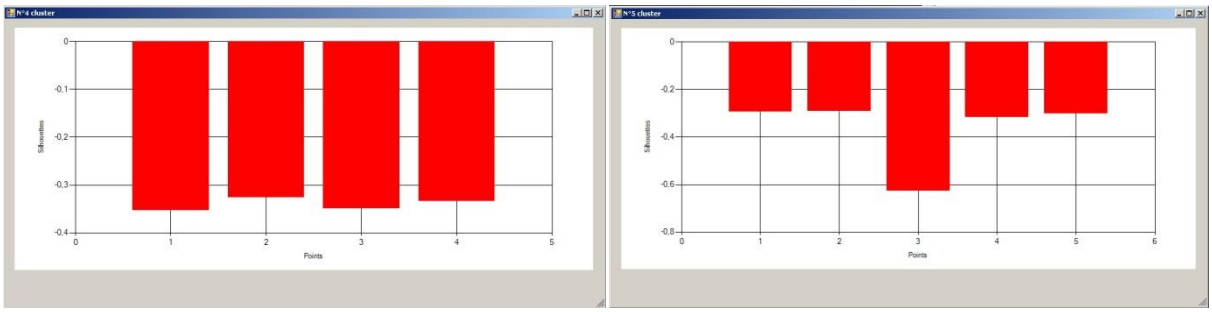


(a) (b)

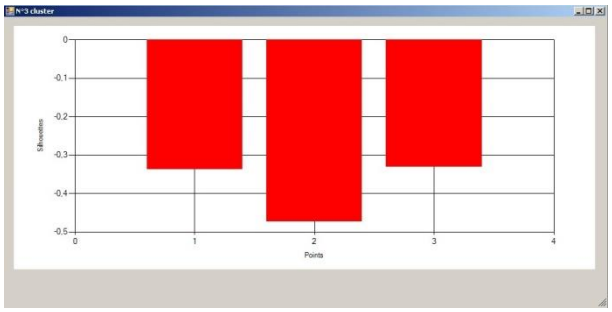
(c)

Figura 5.3.6: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valor medio (c).*

Dai cluster in esame si è deciso di verificare la *silhouette* media per gli "N" *clusters* considerati ottenendo i seguenti andamenti ***(Figura 5.3.7)***:



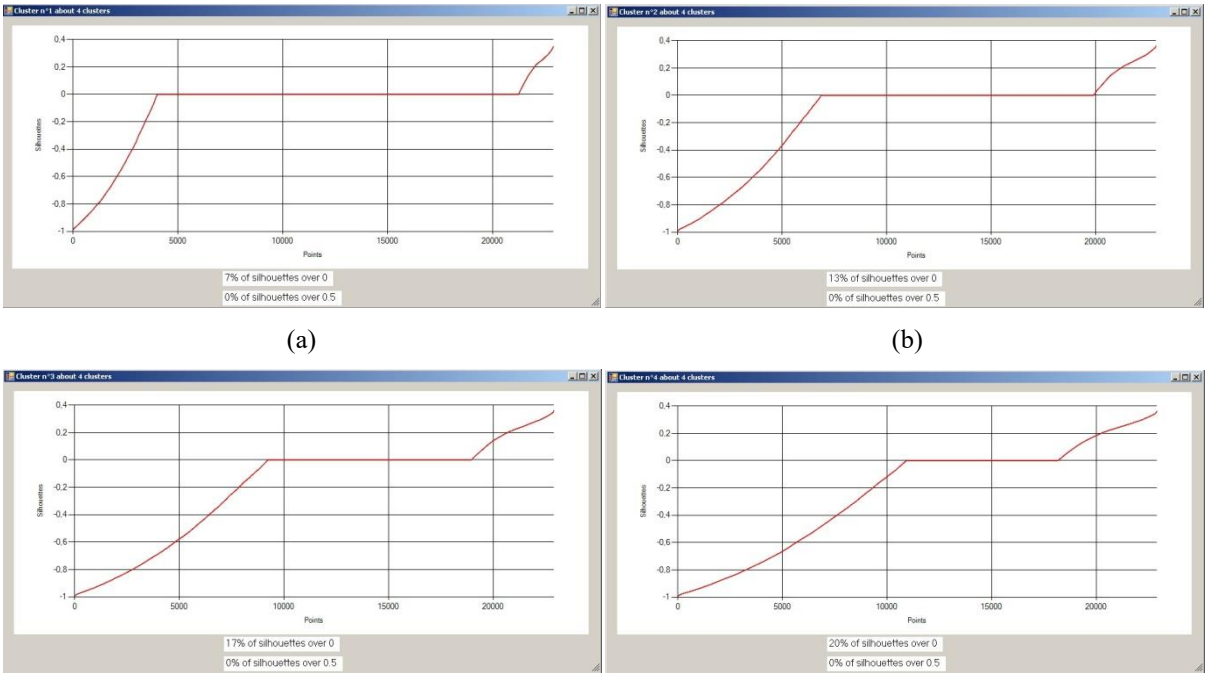
(a) (b)



(c)

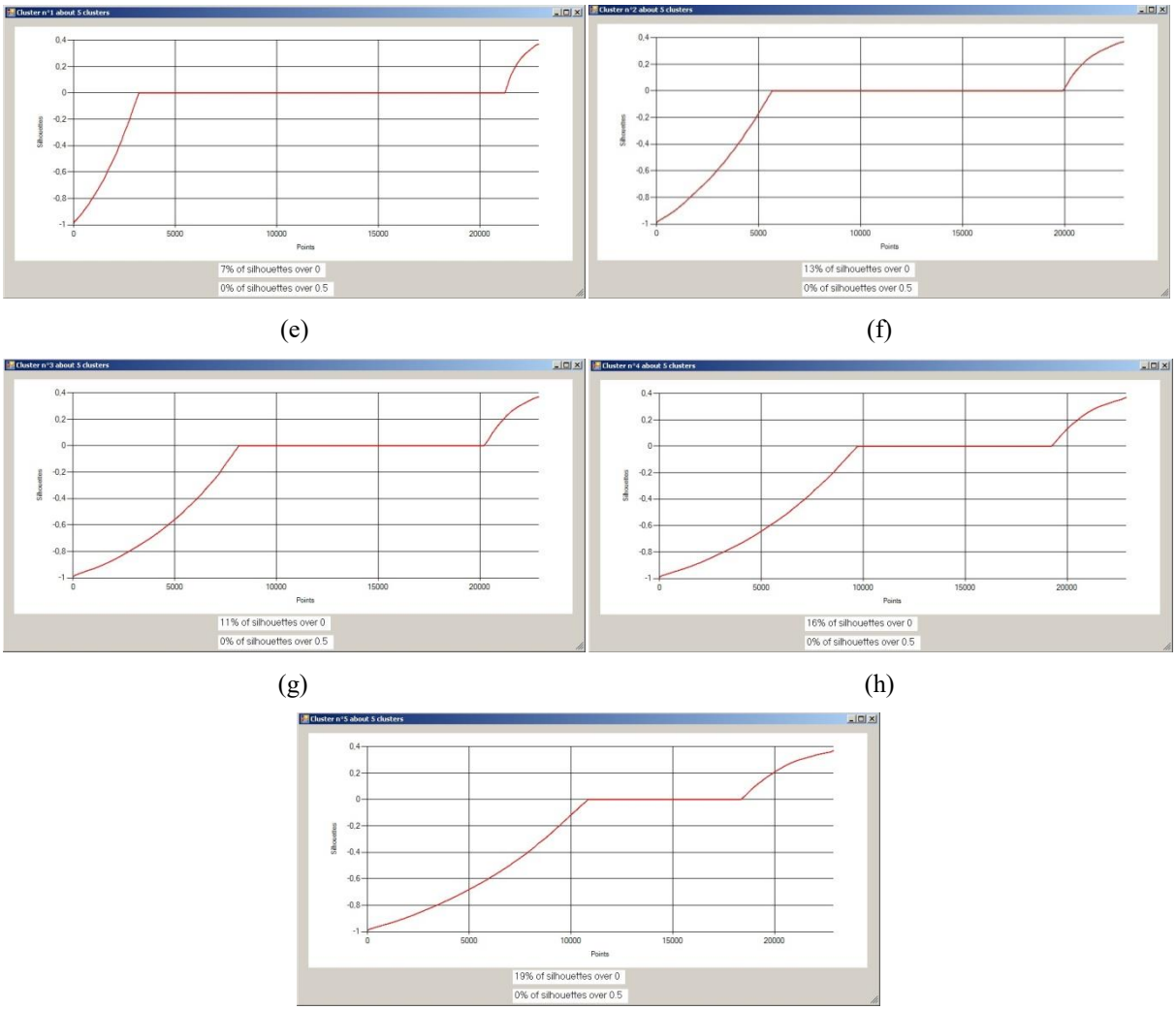
Figura 5.3.7: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c). Per ogni clusters ne viene mostrata la media.*

I risultati ottenuti ***(Figura 5.3.8)*** dimostrano che con 4 *clusters* i valori di *silhouette* superiore la soglia "0" risultano il 7%, 13%, 17% e 20%; con l'uso di 5 *clusters* i valori maggiori di 0 risultano essere del 7%, 13%, 11%, 16% e 19%.



(a) (b)

(c) (d)



(e) (f)

(g) (h)

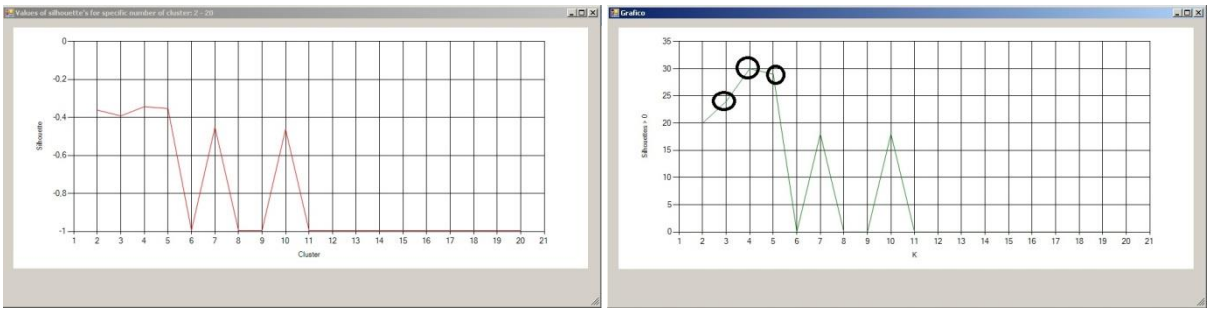
(i)

Figura 5.3.8: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, viene mostrato l'andamento della silhouette per gli "N" clusters considerati.*

Il miglior valore medio di silhouette (con 4 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a), (b), (c) e (d). Il secondo miglior valor medio (5 clusters) dalle restanti figure (e), (f), (g), (h), (i).

70% (32046 tuple)

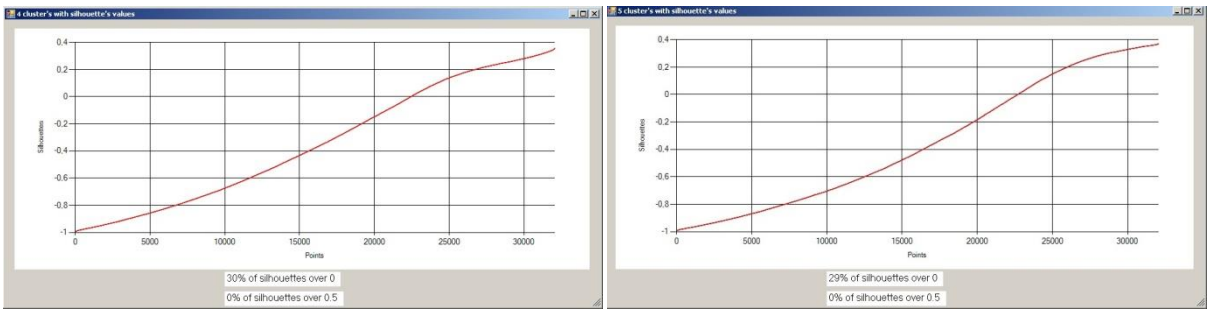
In questo caso si considerano le *silhouette* per i *clusters* da 2 a 20 ***(Figura 5.3.9)***. I migliori valori di *silhouette* si raggiungono con 4 e 5 *clusters* mentre il terzo miglior valore medio si ottiene con l'uso di 3 *clusters*.



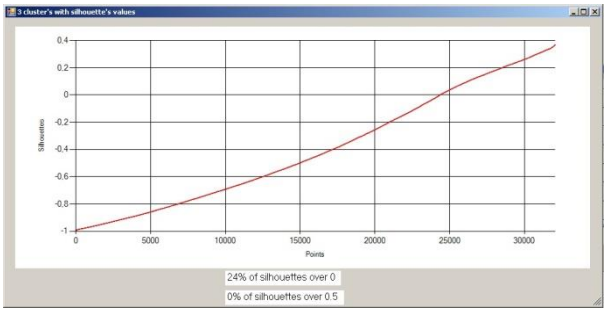
(a) (b)

Figura 5.3.9: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

L'andamento della *silhouette* per "N" cluster è mostrato nella ***Figura 5.3.10***. Con 30%, 29%, 24%, rispettivamente per 4, 5 e 3 *clusters*, si ha la conferma che per avere migliori valori di *silhouette* si debba considerare la percentuale ottenuta con il primo grafico.



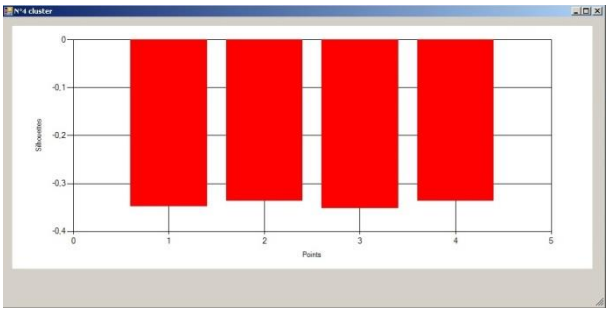
(a) (b)



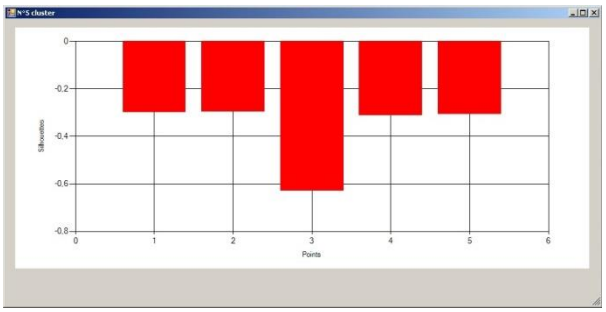
(c)

Figura 5.3.10: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c).*

Dai *clusters* in esame si è deciso di verificare la *silhouette* media per gli "N" *clusters* considerati raggiungendo i seguenti andamenti ***(Figura 5.3.11)***:



(a)

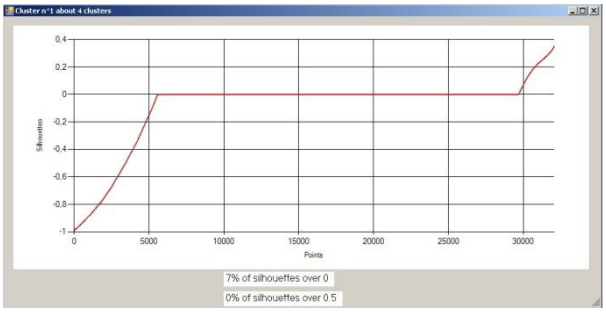


(b)

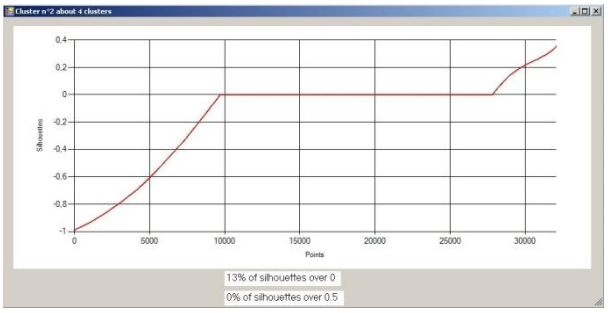
(c)

Figura 5.3.11: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c). Per ogni clusters viene mostrata la media.*

Dai grafici ottenuti ***(Figura 5.3.12)*** si nota come con 4 *clusters* i valori di *silhouette* superiore a "0" risultano rispettivamente del 7%, 13%, 17%, 20%; con l'uso di 5 *clusters* si ottengono rispettivamente il 7%, 13% , 11%, 16%, 19% di valori che superano la soglia "0".



(a)



(b)

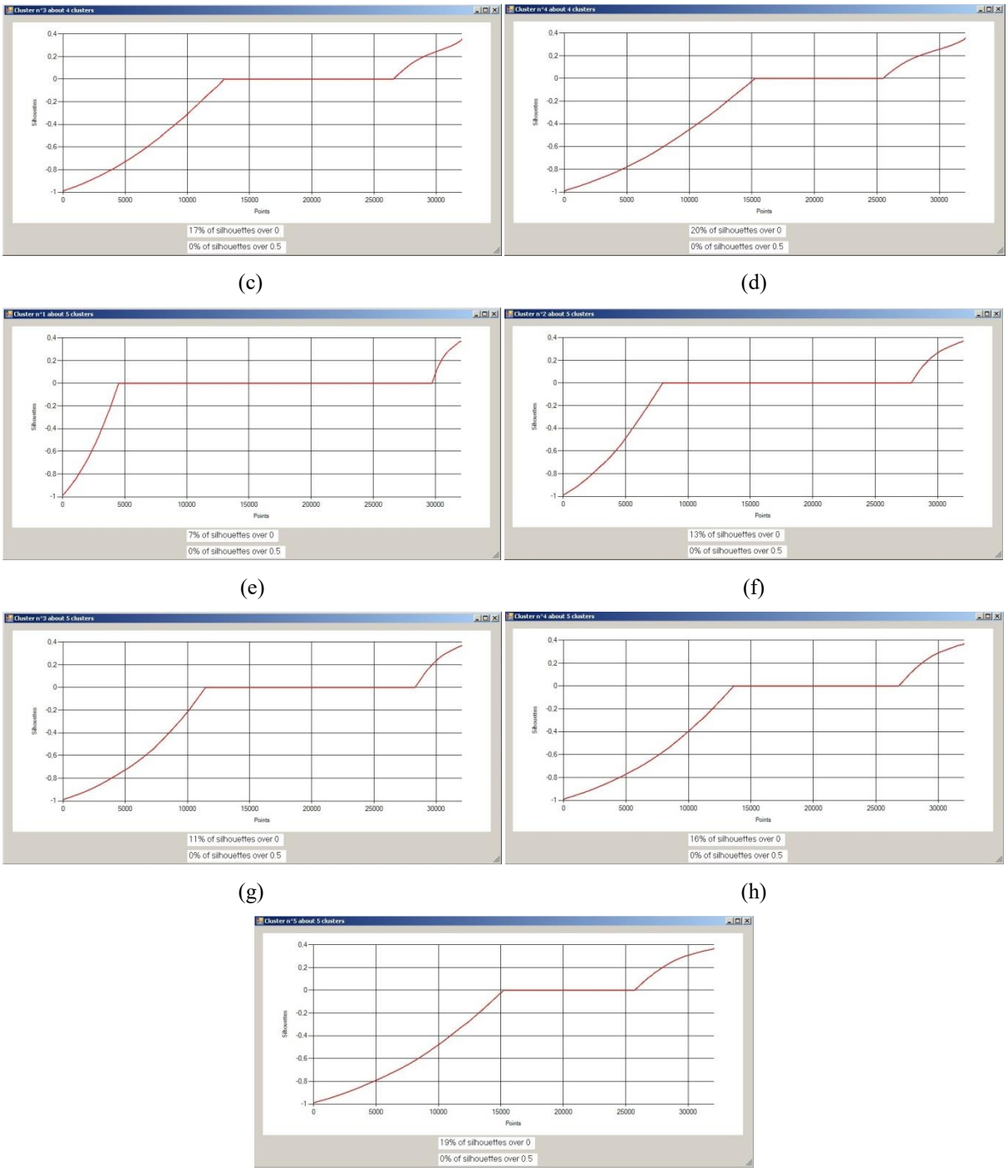
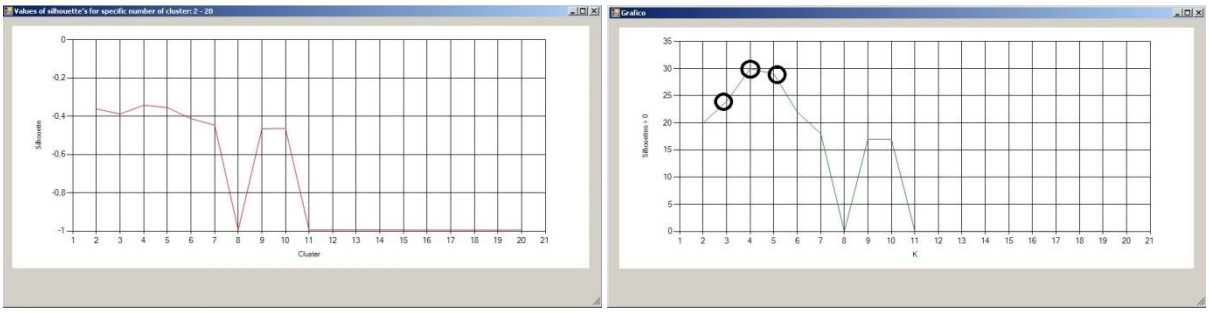


Figura 5.3.12: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, viene mostrato l'andamento della silhouette per gli "N" clusters considerati.*

Il miglior valore medio di silhouette (con 4 clusters) è rappresentato dagli andamenti descritti nelle figure (a), (b), (c), (d). Il secondo miglior valore medio (5 clusters) dalle restanti figure (e), (f), (g), (h), (i).

80% (36624 tuple)

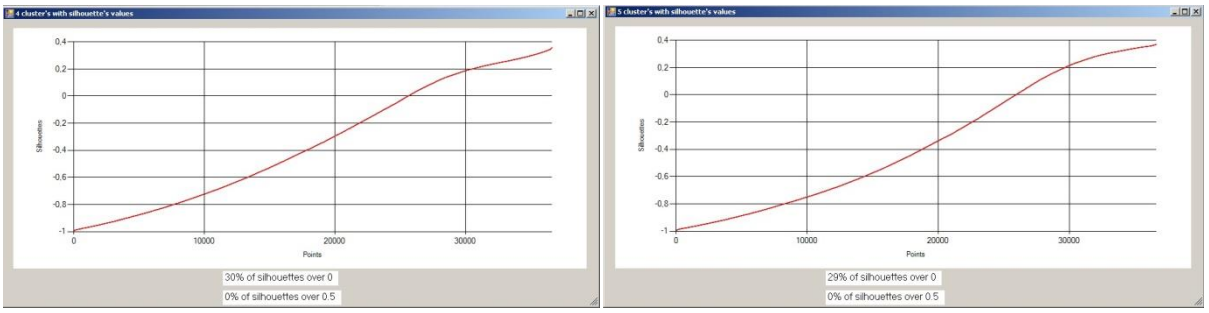
In questo caso è considerata la *silhouette* per i *clusters* da 2 a 20 ***(Figura 5.3.13)***. I migliori valori di *silhouette* si raggiungono con4 e 5 *clusters* mentre il terzo miglior valore medio si ottiene con l'uso di 3 *clusters*.



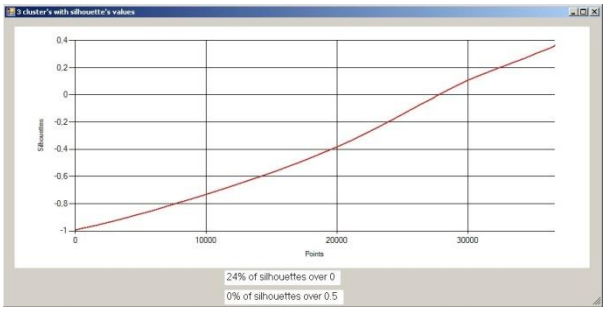
(a) (b)

Figura 5.3.13: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

L'andamento della *silhouette* per "N" *clusters* è mostrato nella ***Figura 5.3.14***. Per 4, 5 e 3 *clusters* si raggiungono rispettivamente il 30%, 29% e 24% dei valori superiori alla soglia "0".



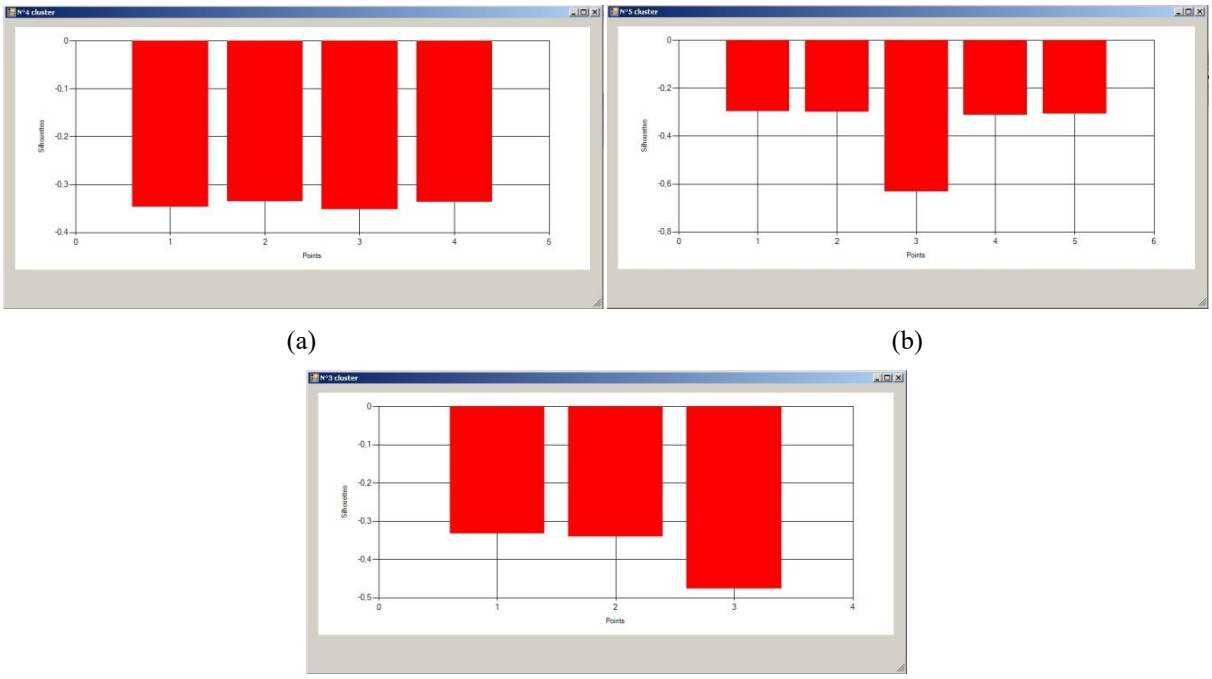
(a) (b)



(c)

Figura 5.3.14: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c).*

Dai *clusters* in esame si è scelto di verificare la *silhouette* media per gli "N" *clusters* considerati conseguendo i seguenti andamenti ***(Figura 5.3.15)***:

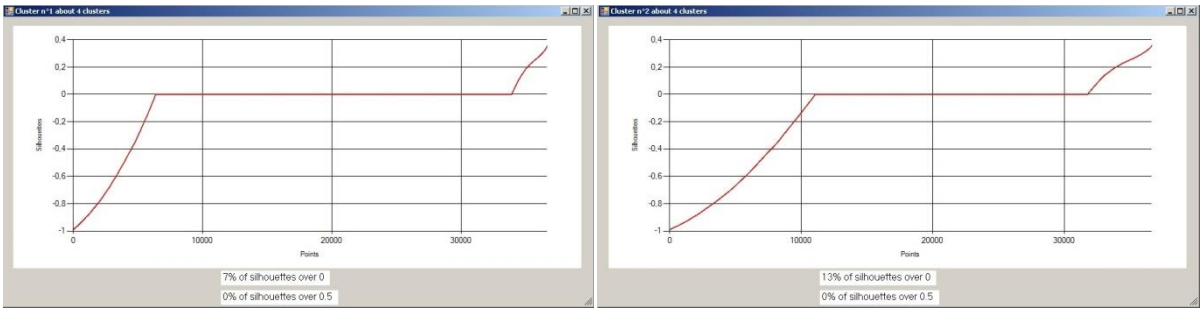


(a) (b)

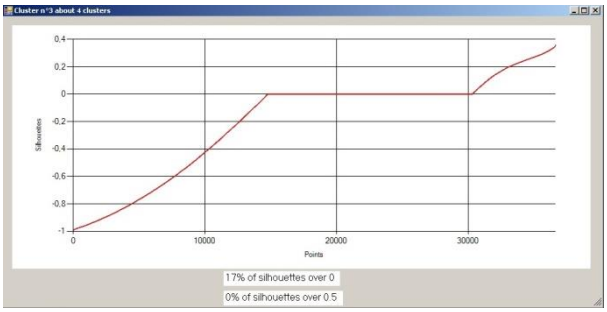
(c)

Figura 5.3.15: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c). Per ogni clusters viene mostrata la media.*

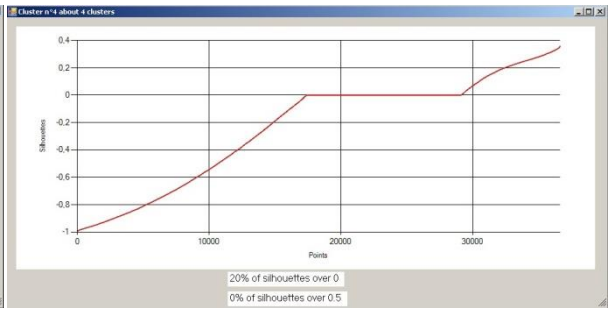
Dai grafici ottenuti ***(Figura 5.3.16)*** si nota che con 4 *clusters* i valori di *silhouette* superiore a "0" sono rispettivamente del 7%, 13%, 17%, 20. Con 5 *clusters,* invero, si ottiene il 7%, 13% , 11%, 16%, 19% dei valori che superano la soglia "0".



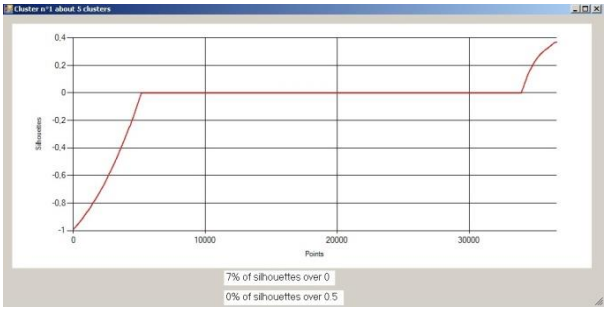
(a) (b)



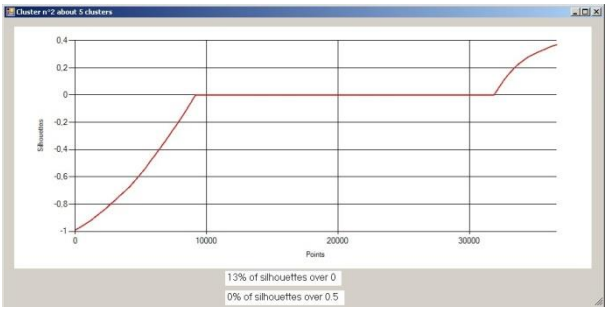
(c)



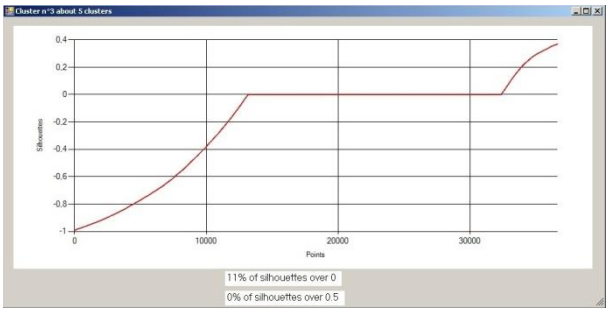
(d)



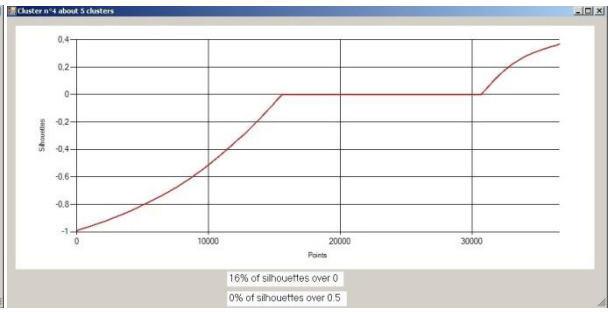
(e)



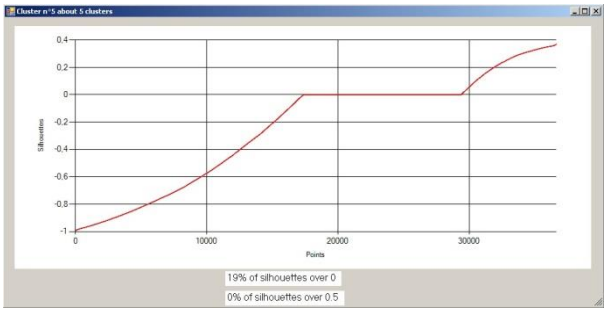
(f)



(f)



(g)



(i)

Figura 5.3.16: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

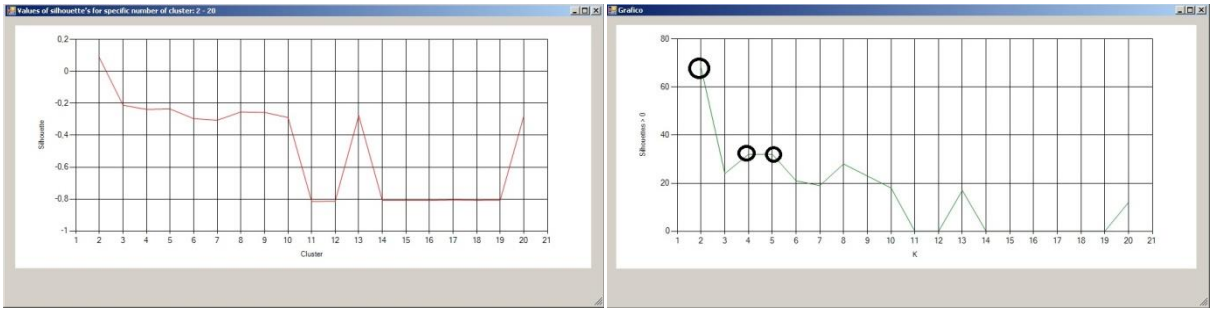
Il miglior valore medio di silhouette (con 4 clusters) è rappresentato dagli andamenti descritti nelle figure (a), (b), (c) e (d). Il secondo miglior valore medio (5 clusters) dalle restanti figure (e), (f), (g), (h) ed (i).

5.4 Wholesale customers dataset

Questo dataset contiene circa 400 righe *(440)* e si è scelto di considerarne il 100%.

100% (440 tuple)

In questo caso è considerata la *silhouette* per i *clusters* da 2 a 20 ***(Figura 5.4.1)***. I migliori valori di *silhouette* si raggiungono con 2 e 4 *clusters* mentre il terzo miglior valore medio si ottiene con l'uso di 5 *clusters*.



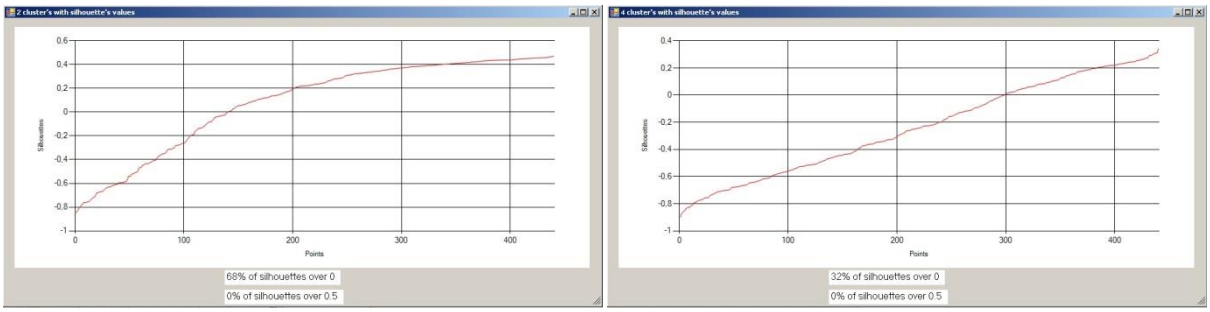
(a) (b)

Figura 5.4.1: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

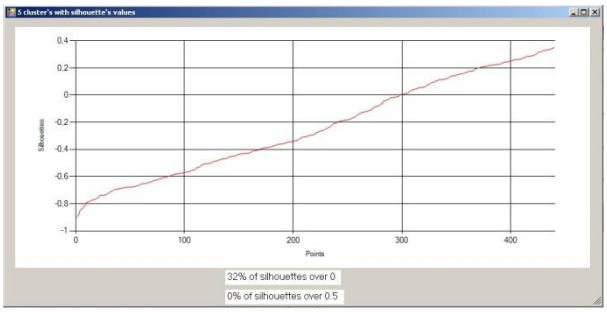
Osservando i grafici ***(Figura 5.4.2)*** prodotti dall'applicazione si comprende quale sia l'andamento della *silhouette* per i punti considerati.

Al di sotto del grafico sono indicate le percentuali relative il numero di punti che sono superiori ai valori 0 e 0,5 di *silhouette*. Nel caso in specie:

1. **2 cluster** *(1° miglior valore di silhouette):* 68% dei punti superiore a 0 e nessun punto superiore allo 0,5;
2. **4 cluster** *(2° miglior valore di silhouette):* 32% dei punti superiore a 0 e nessun punto superiore allo 0,5;
3. **5 cluster** *(peggior valore di silhouette):* 32% dei punti superiore a 0 e nessun punto superiore allo 0,5;



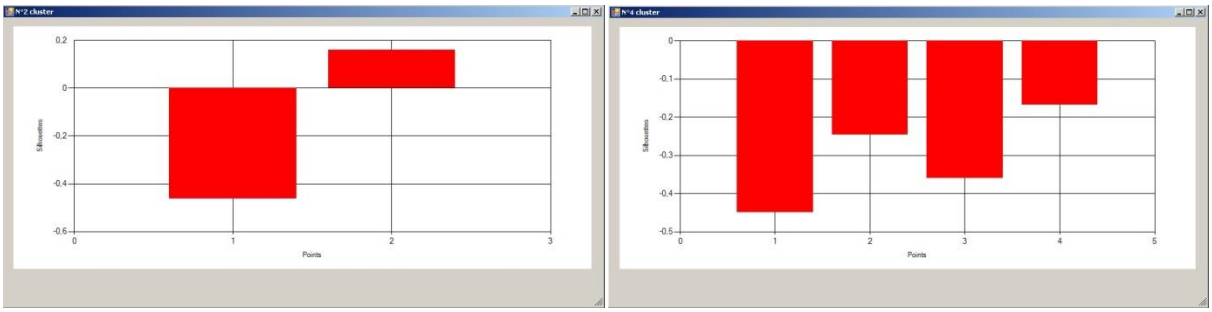
(a) (b)



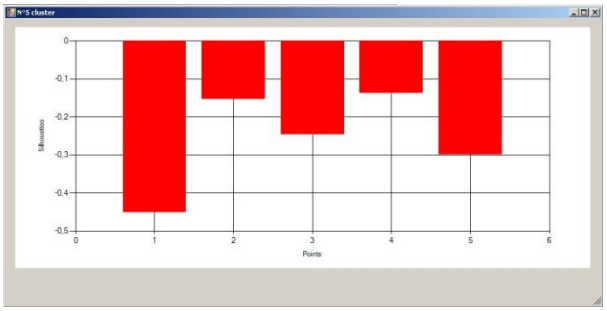
(c)

Figura 5.4.2: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c).*

Dai *clusters* in esame si è deciso di verificare la *silhouette* media per gli "N" *clusters* considerati conseguendo i seguenti andamenti ***(Figura 5.4.3)***:



(a) (b)



(c)

Figura 5.4.3: *Si considerano gli "N" clusters che offrono il miglior valor medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valor medio di silhouette (b) e gli "N" clusters che offro"N"o il terzo miglior valor medio (c) e per ogni clusters che ne fanno parte ne viene mostrata la media.*

Considerando l'uso di 2 *clusters* ***(Figura 5.4.4)*** e mostrandone l'andamento della *silhouette* per ogni "N" *cluster* che ne fa parte ci si rende conto come i grafici mostrino rispettivamente l'1% e 66% dei punti superiori alla soglia "0" di *silhouette*.

E’ stato valutato in seguito l’uso di 4 *clusters* e si è potuto appurare che i grafici hanno valore di *silhouette* maggiore alla soglia “0” ovvero il 2%, 2%, 4%, 29%.

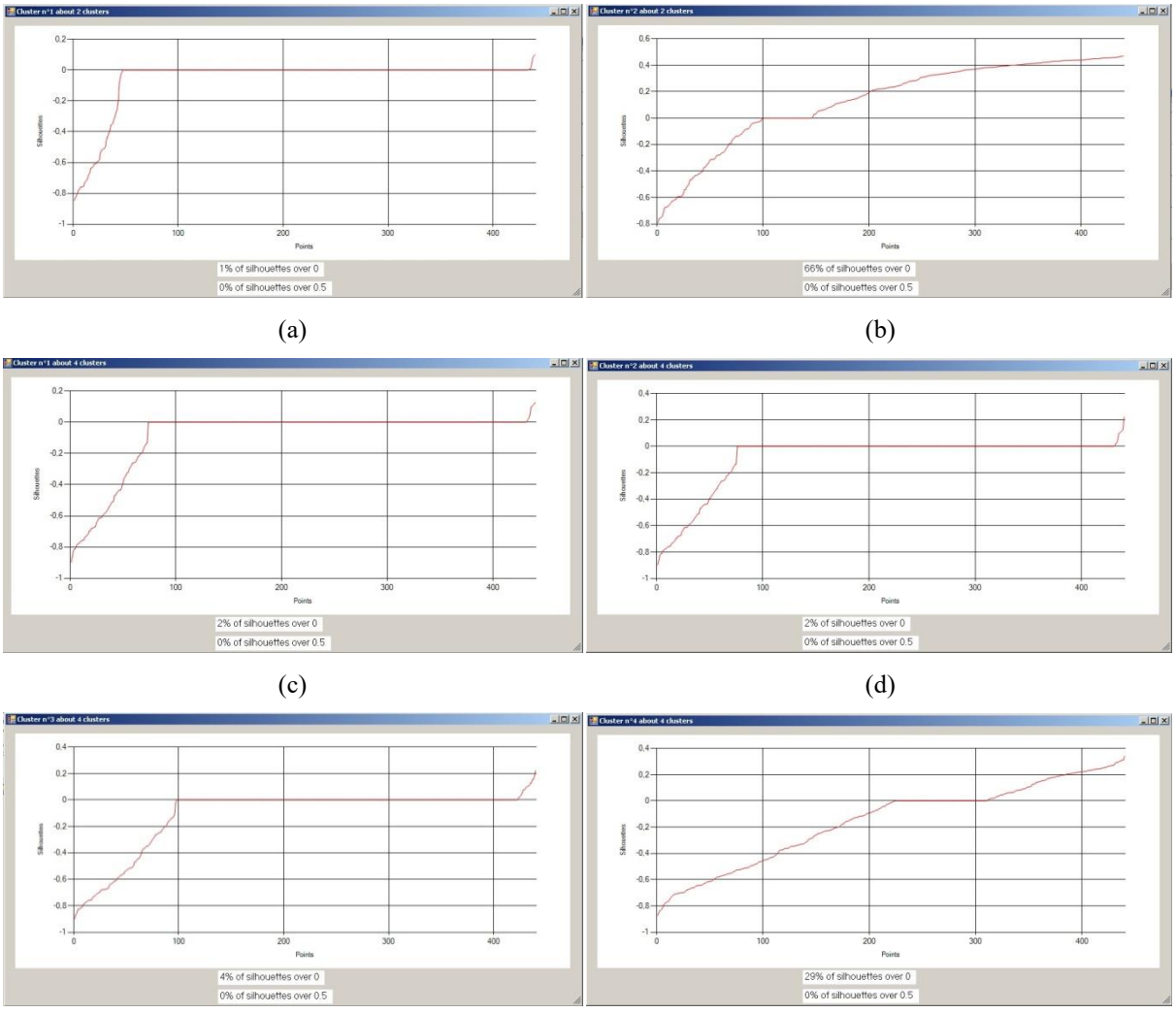


Figura 5.4.4: *Considerando gli “N” clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, si mostra l’andamento della silhouette per gli “N” clusters considerati.*

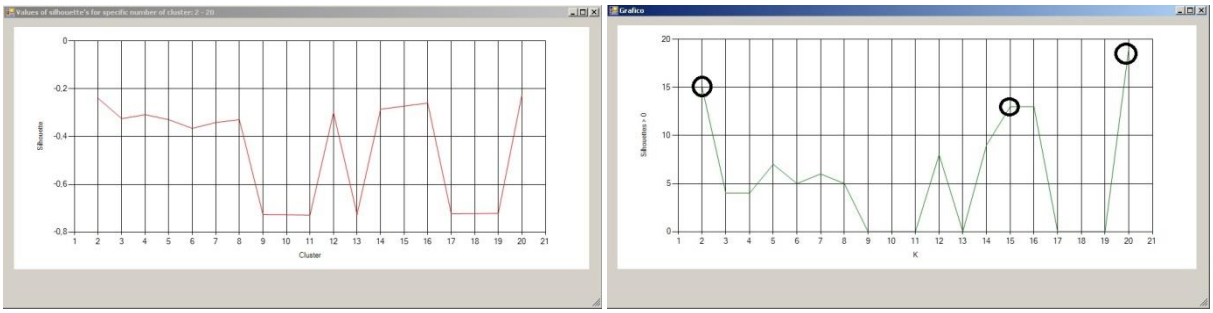
Il miglior valore medio di silhouette (con 2 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a) e (b). Il secondo miglior valore medio (4 clusters) dalle figure (c), (d), (e), (f).

5.5 Knowledge Modeling dataset

Tale *dataset* contiene più di 200 *tuple (258)* e si è scelto di considerarne il 100%.

100% (258 tuple)

E' stato studiato l'andamento della *silhouette* media esaminando i *clusters* da 2-20 ***(Figura 5.5.1)***. In questo caso si nota come il picco della *silhouette* sia in corrispondenza di 20 *clusters*. A fronte di ciò, si è deciso di verificare l'andamento della *silhouette* media per i 2 *cluster* che offrono la *silhouette* migliore (20 e 2 *clusters*) e quello con il terzo andamento migliore rappresentato con l'utilizzo di 15 *clusters*.



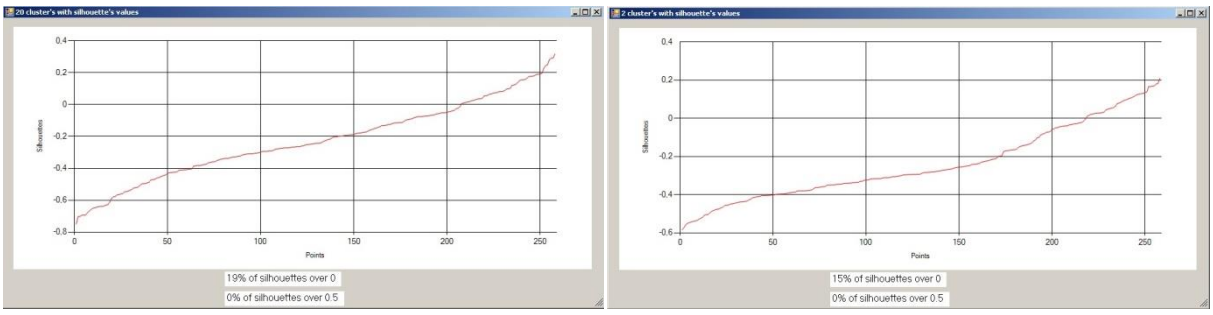
(a) (b)

Figura 5.5.1: *Andamento della silhouette per i clusters da 2 a 20 (a) e grafico delle percentuali degli "N" clusters con silhouette superiori alla soglia 0 (b).*

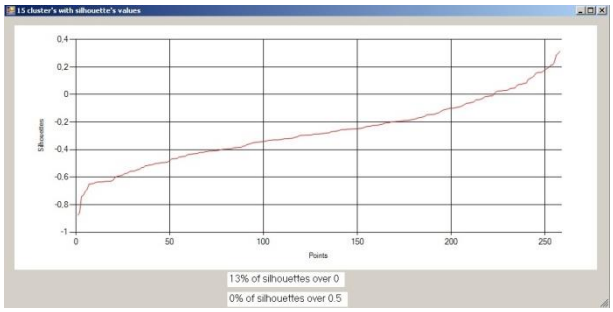
Osservando i grafici ***(Figura 5.5.2)*** prodotti dall'applicazione si comprende quale sia l'andamento della *silhouette* per i punti considerati.

Al di sotto del grafico sono indicate le percentuali riguardanti il numero di punti che sono superiori ai valori 0 e 0,5 di *silhouette*. Nel caso in specie:

1. **20 cluster** *(1° miglior valore di silhouette):* 19% dei punti superiore a 0 e nessun punto superiore allo 0,5;
2. **2 cluster** *(2° miglior valore di silhouette):* 15% dei punti superiore a 0 e nessun punto superiore allo 0,5;
3. **15 cluster** *(peggior valore di silhouette):* 13% dei punti superiore a 0 e nessun punto superiore allo 0,5;



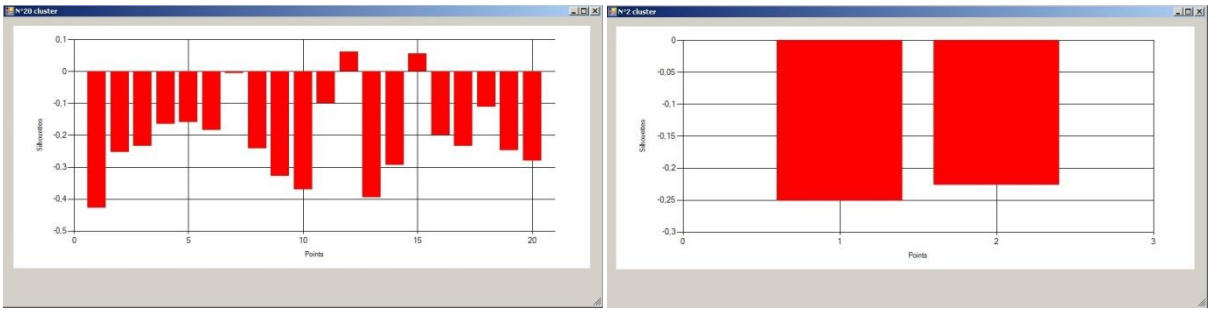
(a) (b)



(c)

Figura 5.5.2: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c).*

Dai cluster in esame è stata analizzata la *silhouette* media per gli "N" clusters considerati ottenendo i seguenti andamenti ***(Figura 5.5.3)***:



(a) (b)

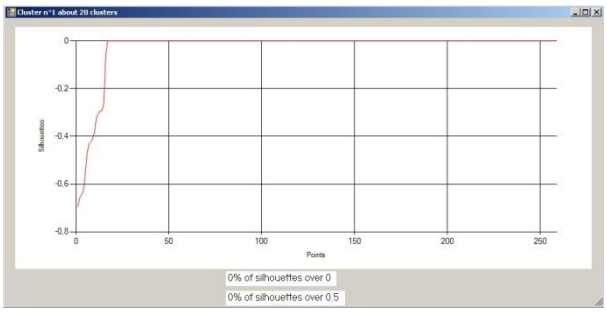
(c)

Figura 5.5.3: *Si considerano gli "N" clusters che offrono il miglior valore medio di silhouette (a), gli "N" clusters che offrono il secondo miglior valore medio di silhouette (b) e gli "N" clusters che offrono il terzo miglior valore medio (c). Per ogni clusters viene mostrata la media.*

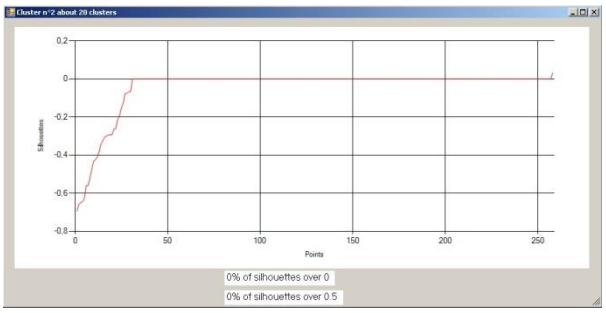
Considerando l'uso di 20 *clusters* ***(Figura 5.5.4)*** e mostrando l'andamento della *silhouette* per ogni "N" *cluster* che ne fa parte si percepisce come i grafici mostrino rispettivamente lo 0%,

0%, 2%, 3%, 5%, 6%, 8%, 9%, 9%, 9%, 10%, 10%, 8%, 7%, 10%, 11%, 12%, 13%, 13% e 10% dei punti superiori alla soglia “0” di *silhouette*.

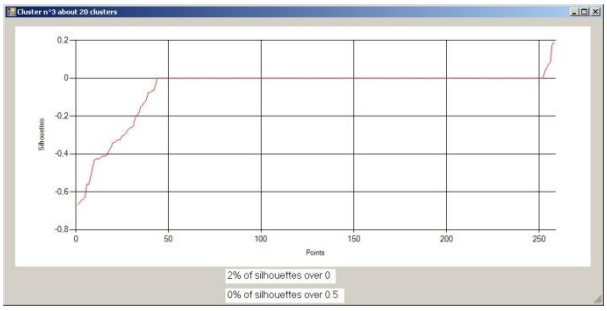
In seguito è stato valutato l’uso di 2 *clusters* costatando che, tranne che per il 1° grafico ove la percentuale dei valori che superano lo “0” di *silhouette* è pari a “0”, gli altri grafici hanno un valore di *silhouette* maggiore allo “0” ovvero l’8% e 13%.



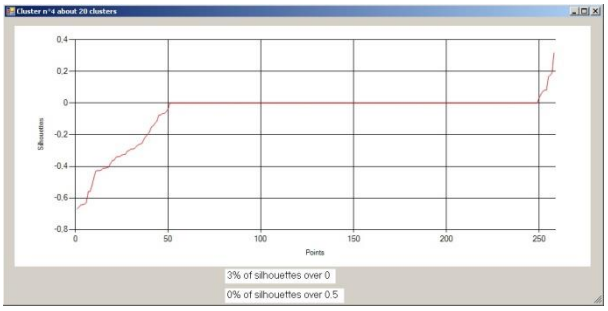
(a)



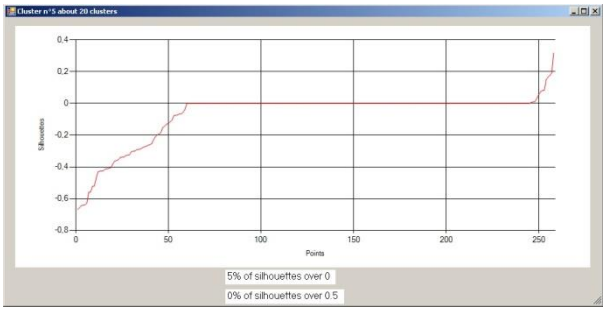
(b)



(c)

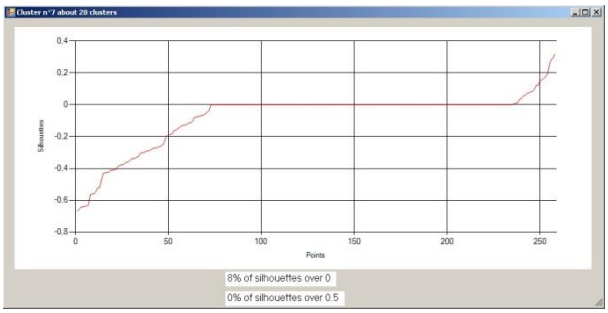


(d)

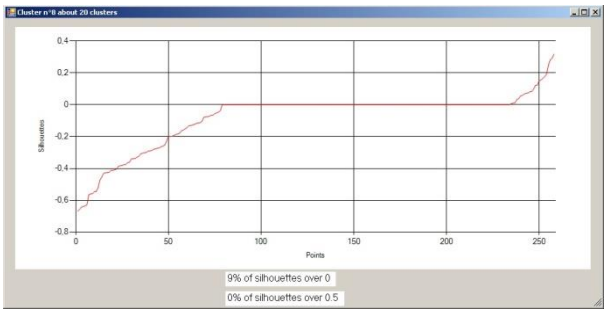


(e)

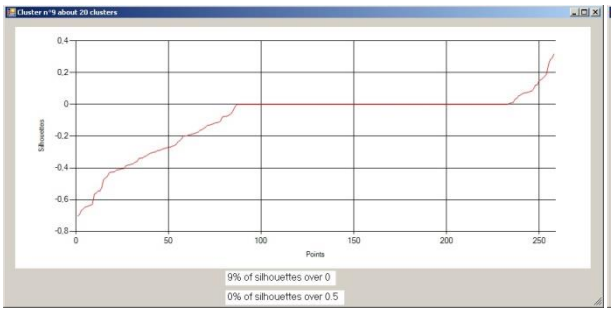
(f)



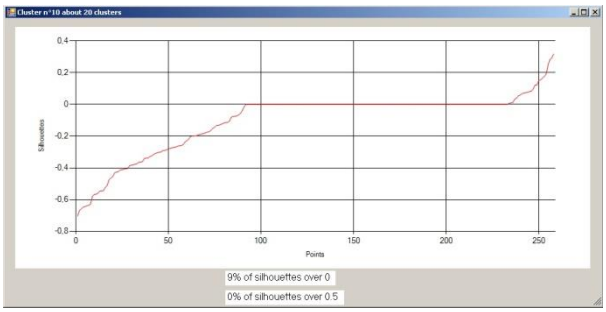
(g)



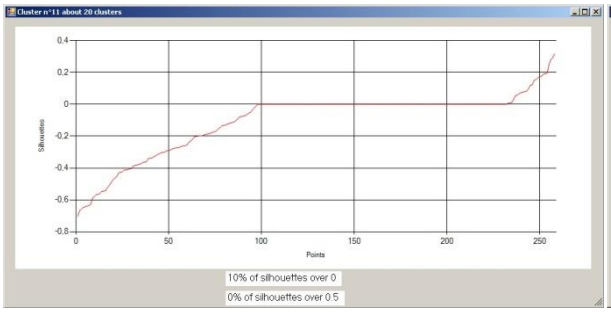
(h)


Cluster n°9 about 20 clusters
Silhouettes
Points
9% of silhouettes over 0
0% of silhouettes over 0.5

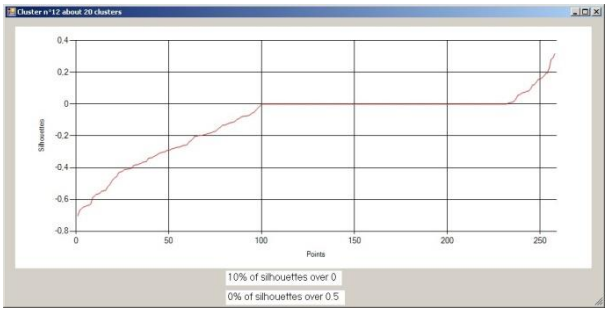
(i)


Cluster n°10 about 20 clusters
Silhouettes
Points
9% of silhouettes over 0
0% of silhouettes over 0.5

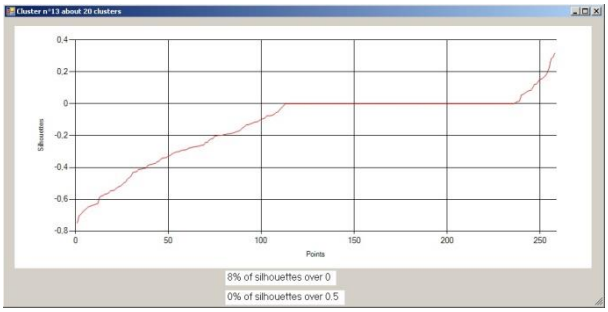
(j)


Cluster n°11 about 20 clusters
Silhouettes
Points
10% of silhouettes over 0
0% of silhouettes over 0.5

(k)


Cluster n°12 about 20 clusters
Silhouettes
Points
10% of silhouettes over 0
0% of silhouettes over 0.5

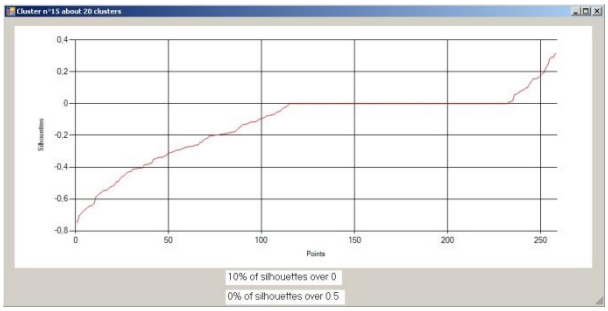
(l)


Cluster n°13 about 20 clusters
Silhouettes
Points
8% of silhouettes over 0
0% of silhouettes over 0.5

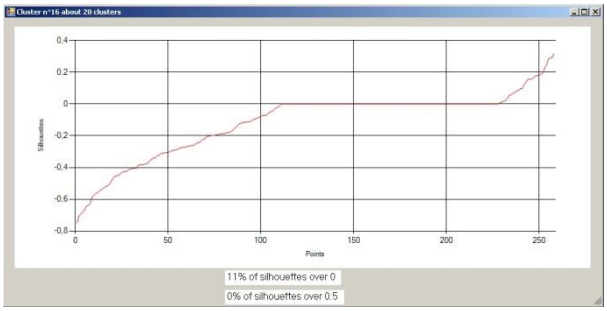
(m)


Cluster n°14 about 20 clusters
Silhouettes
Points
7% of silhouettes over 0
0% of silhouettes over 0.5

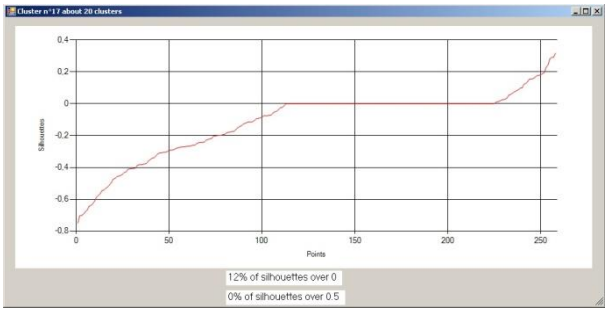
(n)


Cluster n°15 about 20 clusters
Silhouettes
Points
10% of silhouettes over 0
0% of silhouettes over 0.5

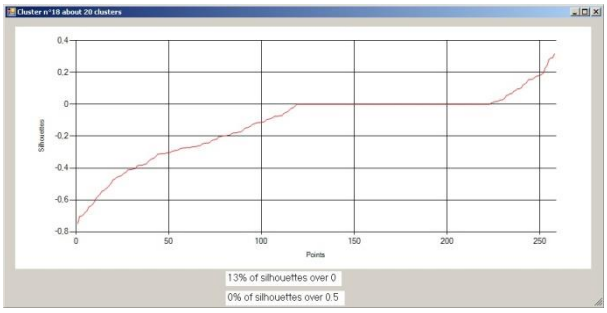
(o)


Cluster n°16 about 20 clusters
Silhouettes
Points
11% of silhouettes over 0
0% of silhouettes over 0.5

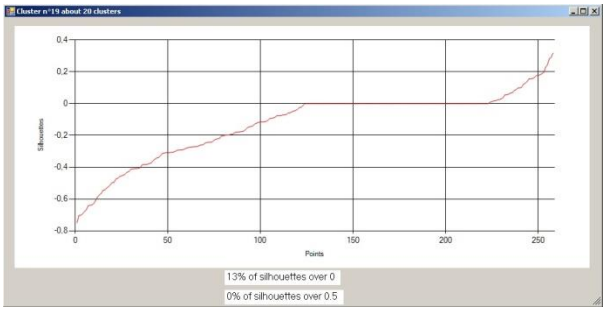
(p)


Cluster n°17 about 20 clusters
Silhouettes
Points
12% of silhouettes over 0
0% of silhouettes over 0.5

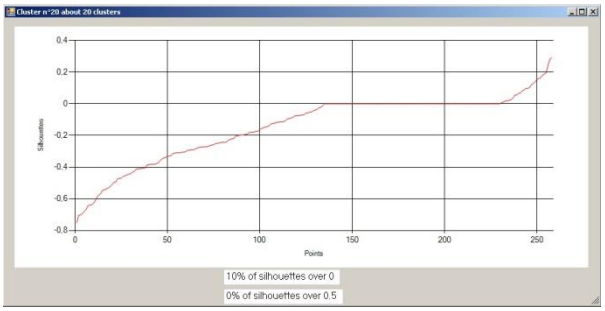
(q)


Cluster n°18 about 20 clusters
Silhouettes
Points
13% of silhouettes over 0
0% of silhouettes over 0.5

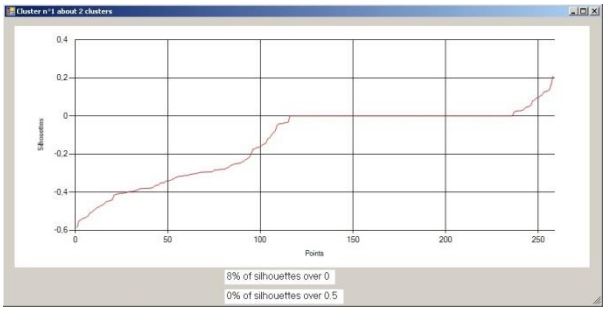
(r)



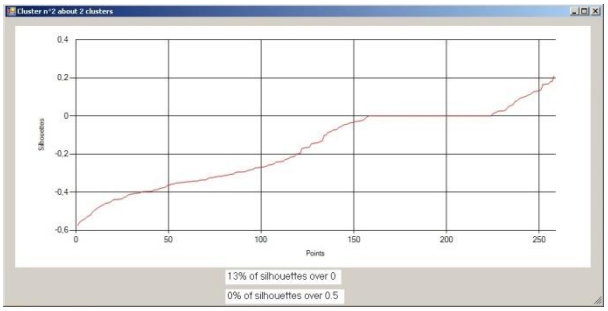
(s)



(t)



(u)



(v)

Figura 5.5.4: *Considerando gli "N" clusters con miglior valore medio di silhouette e secondo miglior valore medio di silhouette, si mostra l'andamento della silhouette per gli "N" clusters considerati.*

Il miglior valor medio di silhouette (con 20 clusters) è rappresentato dagli andamenti rappresentati nelle figure (a), (b), (c) ,(d), (e), (f), (g), (h), (i), (j), (k), (l), (m), (n), (o), (p), (q), (r), (s), (t). Il secondo miglior valore medio (2 clusters) dalle restanti figure (u) e (v).

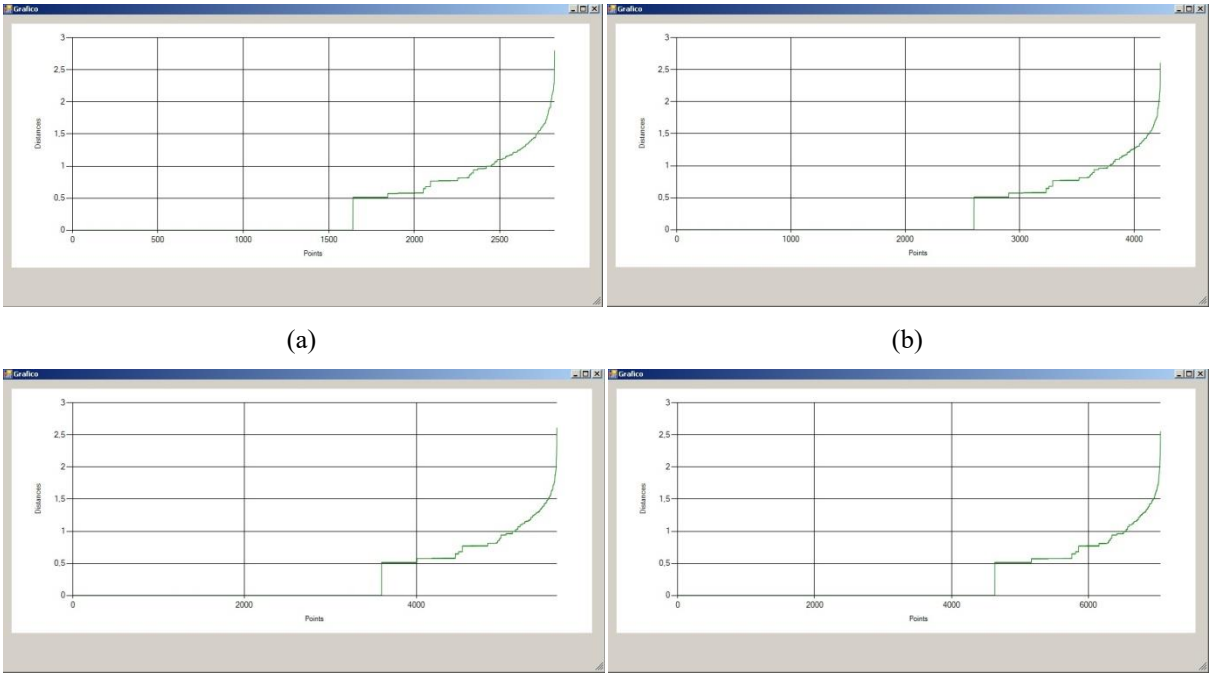
6. Analisi dei dati raccolti (Dbscan)

Eseguendo il *K-Dist* si determinano i valori di *Epsilon* e *MinPts*. Mostrando i grafici con i punti ordinati per distanza crescente, quello che ci si aspetta è che i *"k^{th} nearest neighbors"* stiano più o meno alla stessa distanza *(a meno che non ci sia un'alta variabilità della densità).*

I punti *Noise* tuttavia avranno il loro *k^{th} nearest neighbor* ad una distanza più alta.

6.1 US Census dataset

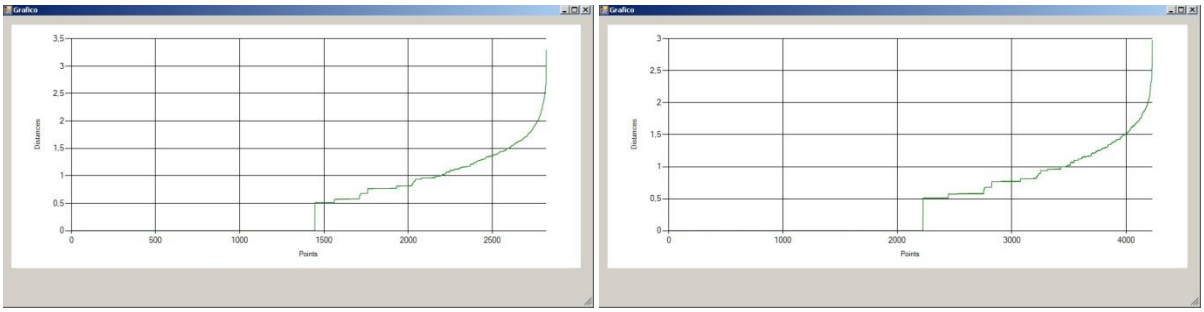
Tale *dataset* contiene circa 70000 *tuple* (70477). Di queste si è deciso di considerarne il 4%, 6%, 8% e 10%. Sotto sono mostrati, in percentuale, i grafici ottenuti:



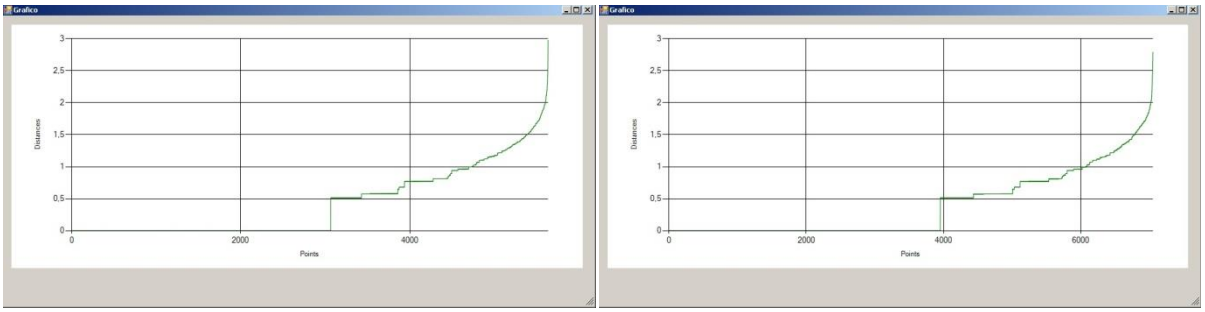
(a) (b)

(c) (d)

Figura 6.1.1: *I grafici (a), (b), (c) e (d) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 2.*

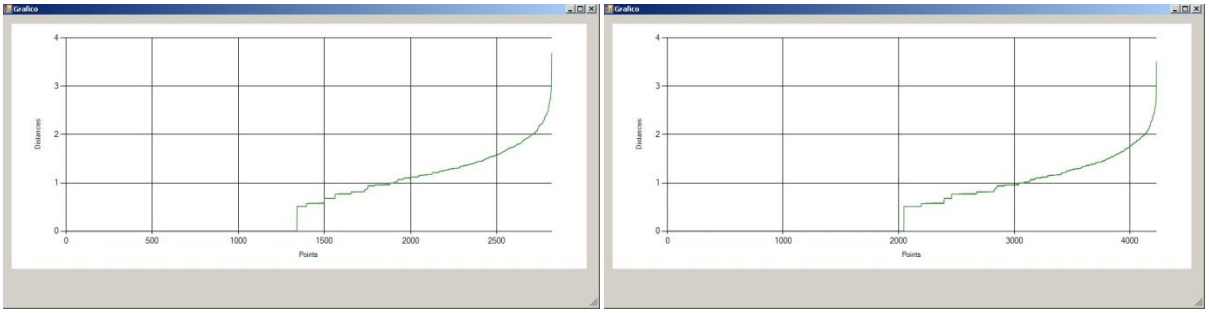


(e) (f)

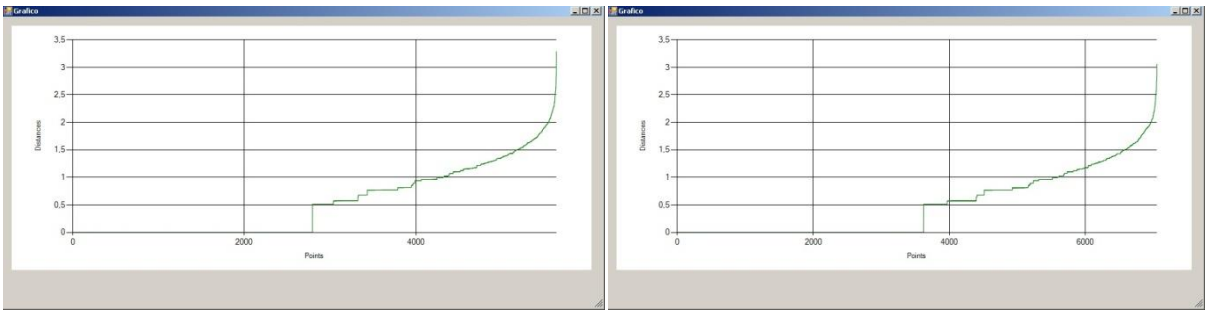


(g) (h)

Figura 6.1.2: *I grafici (e), (f), (g) e (h) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 5.*



(i) (j)

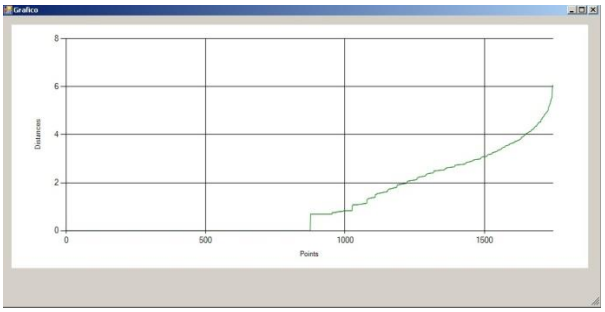


(k) (l)

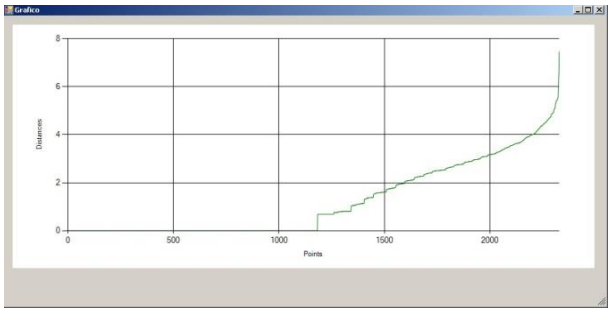
Figura 6.1.3: *I grafici (i), (j), (k) e (l) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 10.*

6.2 Turkiye Student Evaluation - Dataset

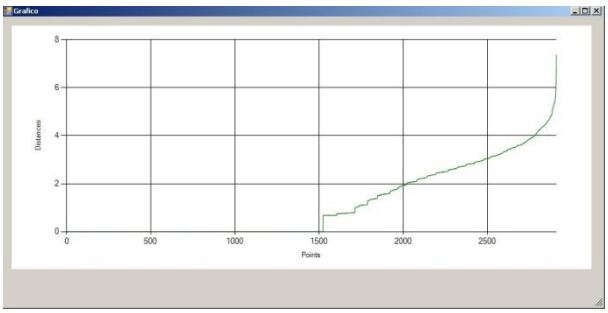
Tale *dataset* contiene circa 5800 *tuple (5820).* Si è deciso di considerare il 30%, 40%, 50% e 60%:



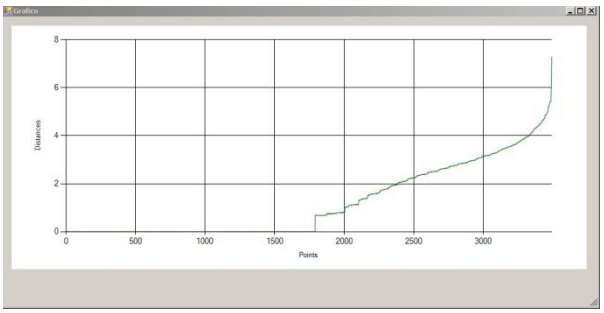
(a)



(b)

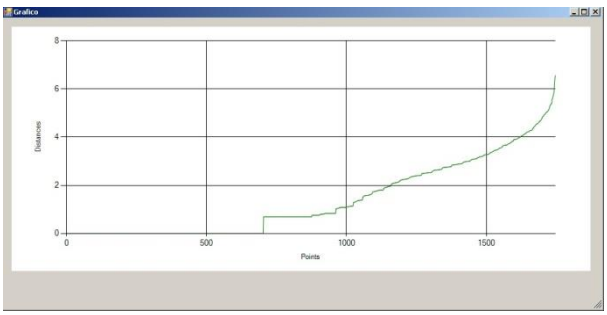


(c)

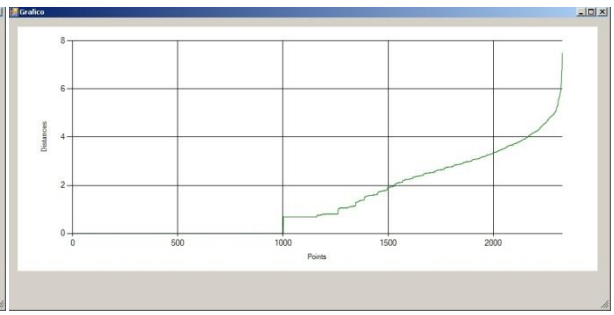


(d)

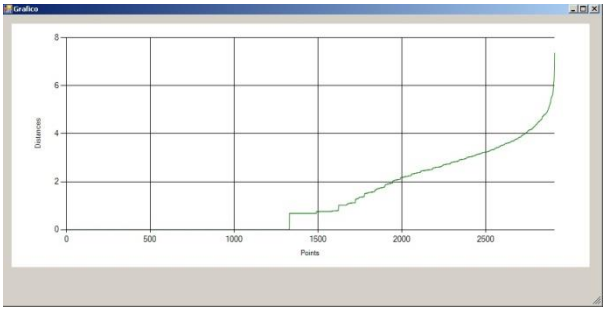
Figura 6.2.1: *I grafici (a), (b), (c) e (d) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 2.*



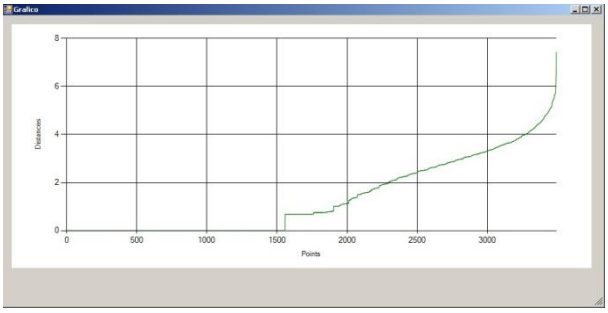
(e)



(f)

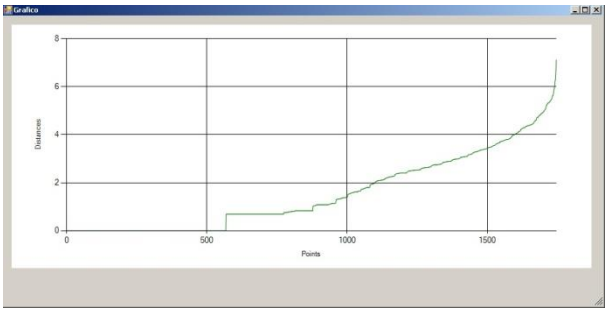


(g)

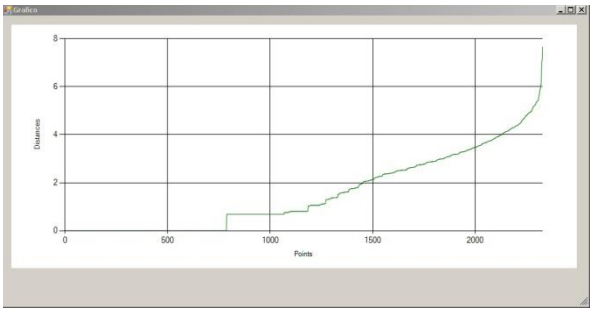


(h)

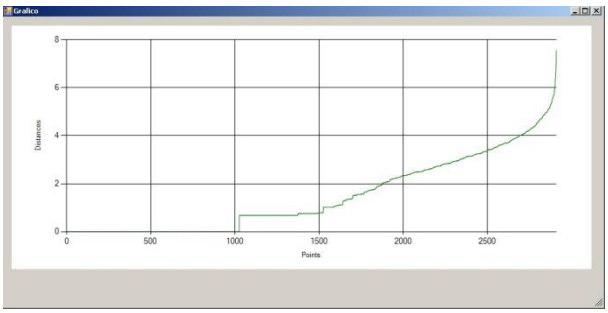
Figura 6.2.2: *I grafici (e), (f), (g) e (h) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 5.*



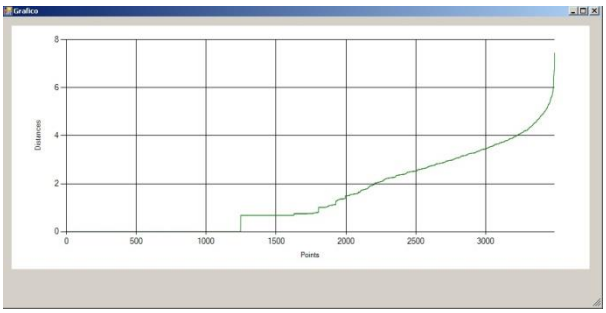
(i)



(j)



(k)

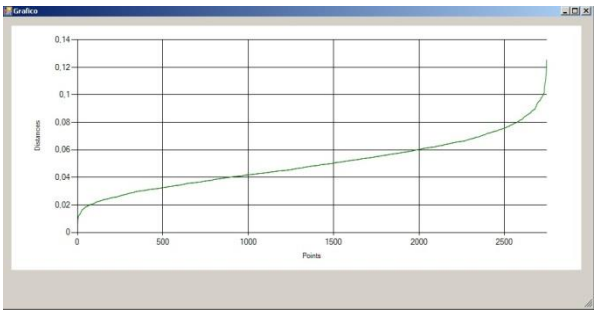


(l)

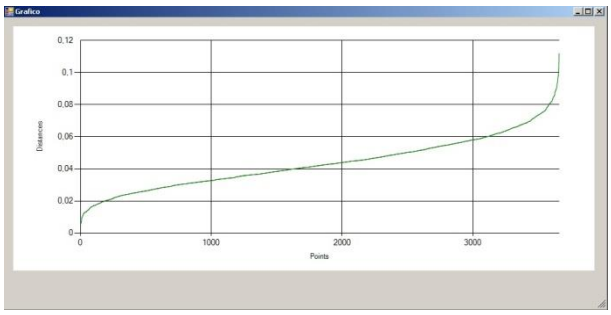
Figura 6.2.3: *I grafici (i), (j), (k) e (l) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 10.*

6.3 Electricity - Dataset

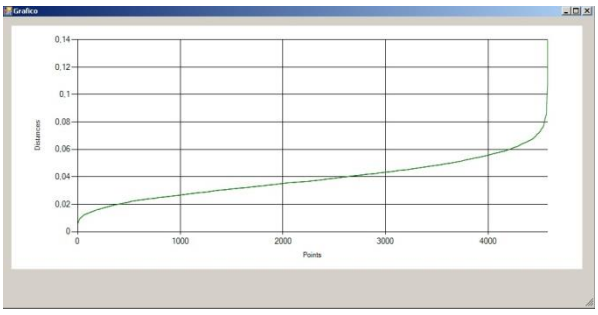
Tale *dataset* contiene circa 46000 *tuple (45781).* Si è deciso di considerare il 6%, 8%, 10% e 12%:

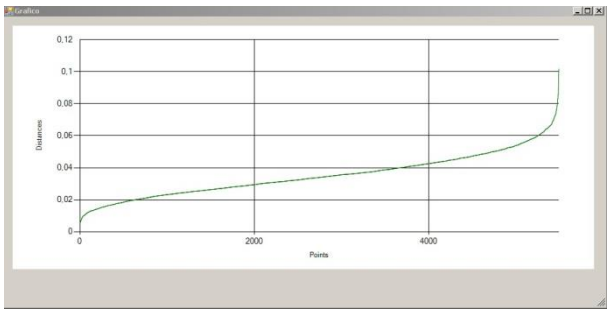


(a)



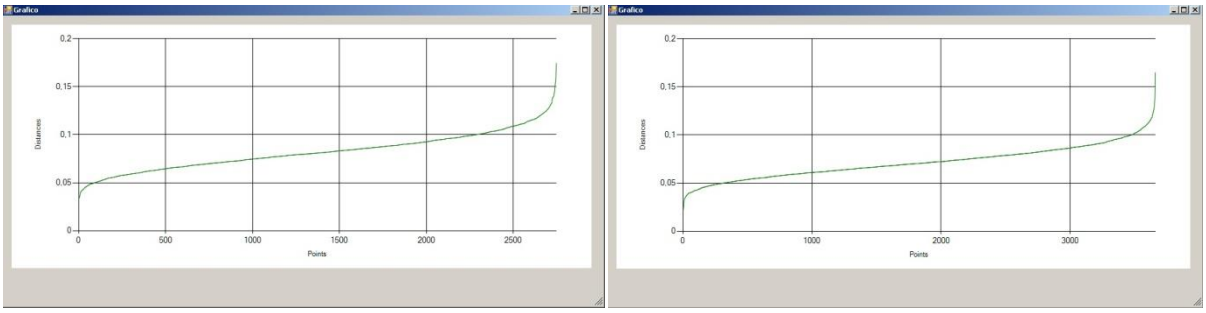
(b)



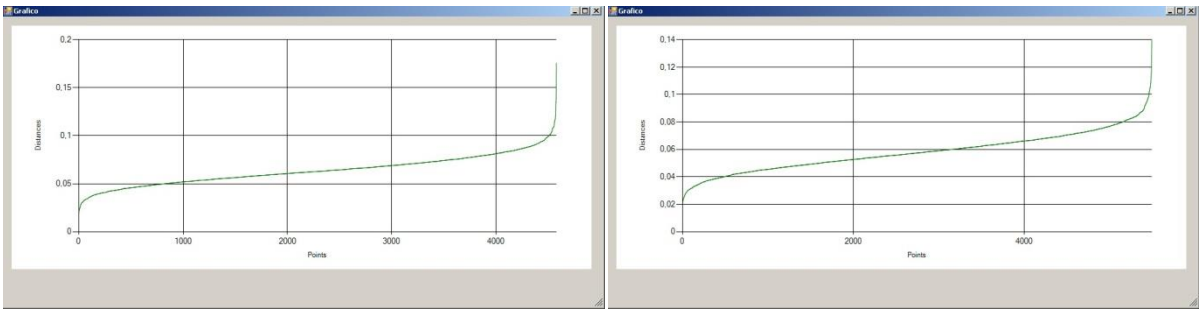


(c) (d)

Figura 6.3.1: *I grafici (a), (b), (c) e (d) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 2.*

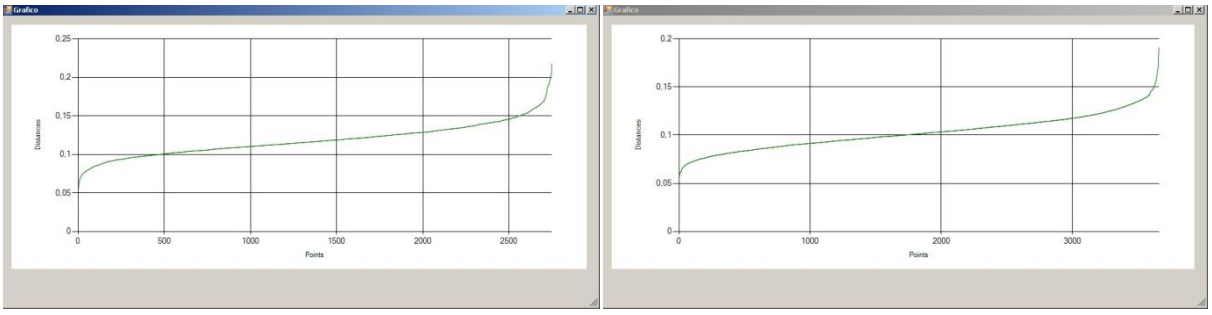


(e) (f)

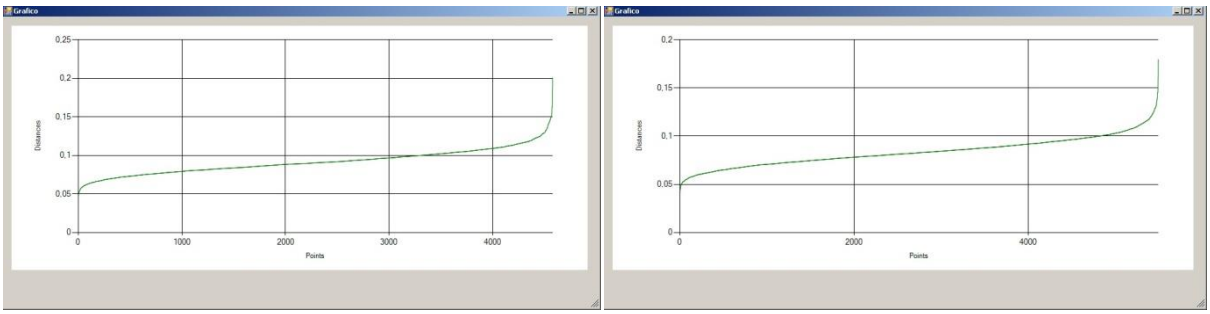


(g) (h)

Figura 6.3.2: *I grafici (e), (f), (g) e (h) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 5.*



(i) (j)

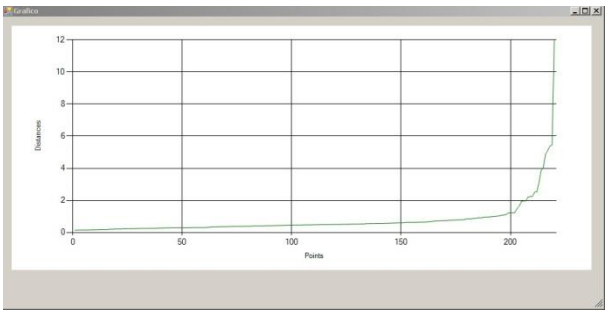


(k) (l)

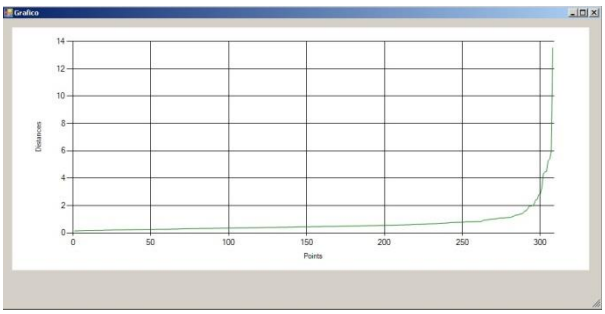
Figura 6.3.3: *I grafici (i), (j), (k) e (l) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 10.*

6.4 Wholesale customers - Dataset

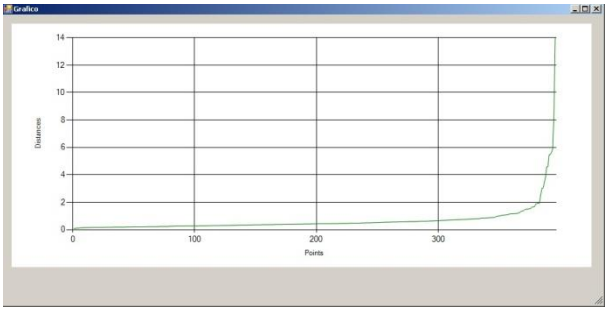
Tale *dataset* contiene più di 400 *tuple (440)*. Si è deciso di considerare il 50%, 70%, 90% e 100%:



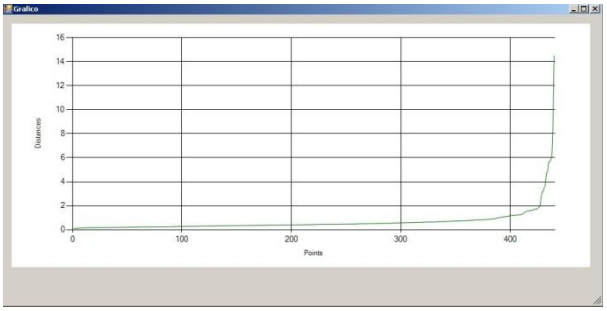
(a)



(b)

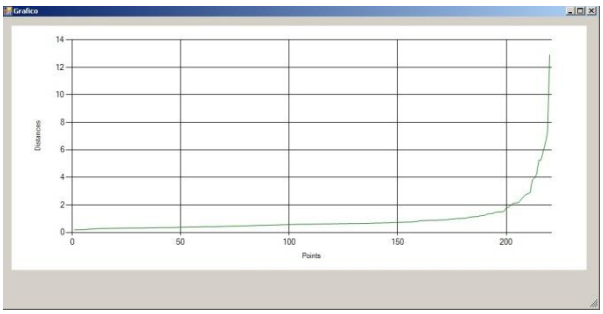


(c)

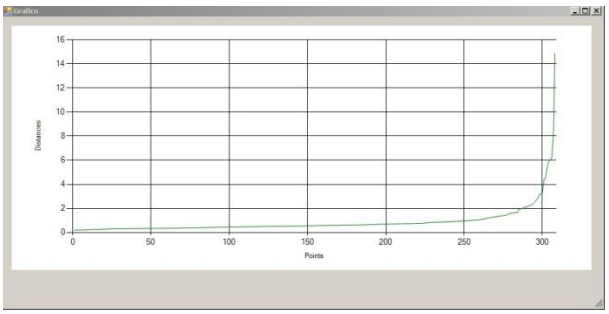


(d)

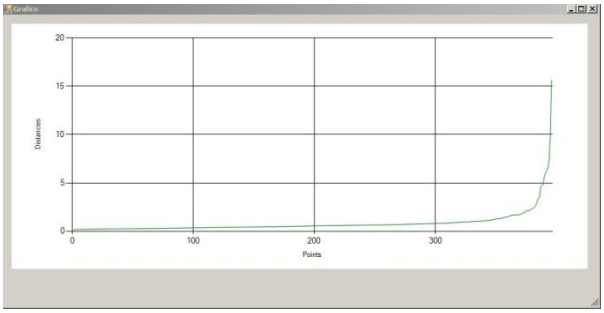
Figura 6.4.1: *I grafici (a), (b), (c) e (d) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 2.*



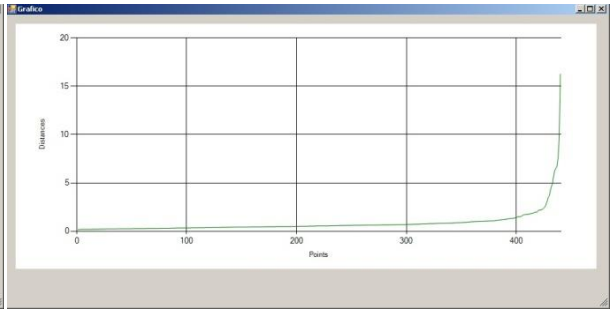
(e)



(f)

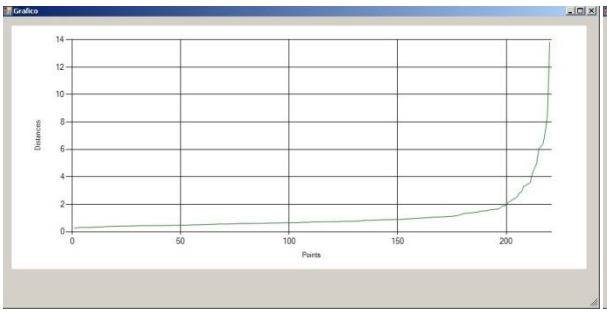


(g)

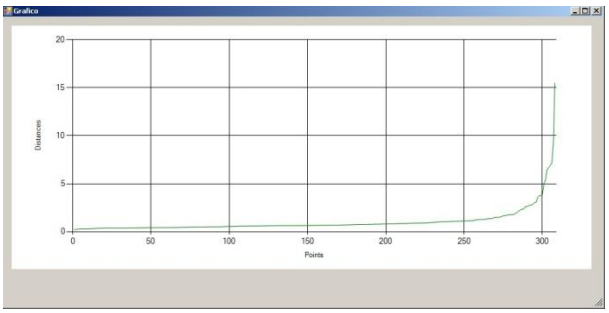


(h)

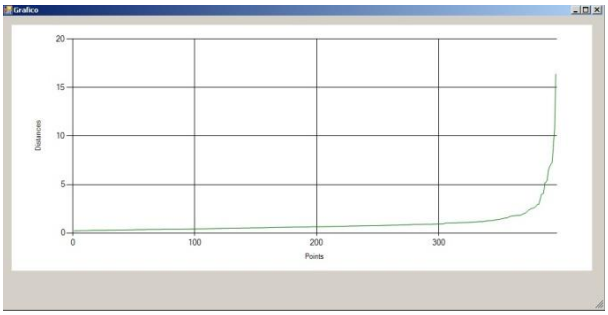
Figura 6.4.2: *I grafici (e), (f), (g) e (h) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 5.*



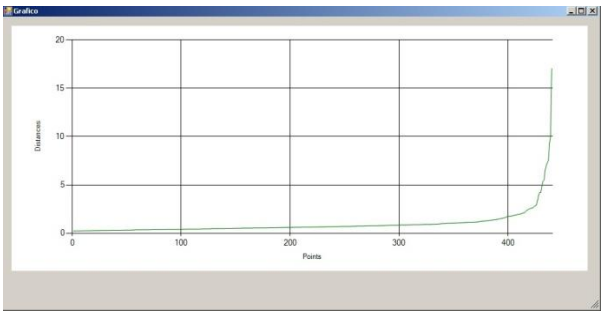
(i)



(j)



(k)

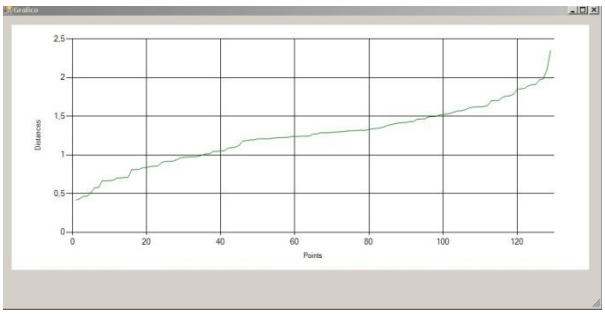


(l)

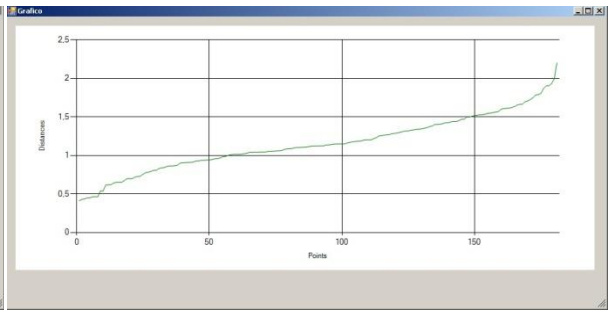
Figura 6.4.3: *I grafici (i), (j), (k) e (l) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 10.*

6.5 User Knowledge Modeling - Dataset

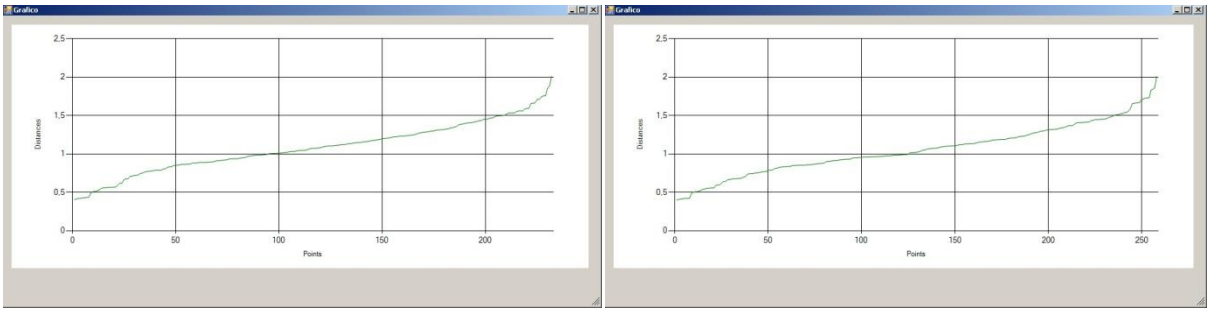
Tale dataset contiene più di 200 *tuple (258)*. Si è deciso di considerare il 50%, 70%, 90% e 100%:



(a)

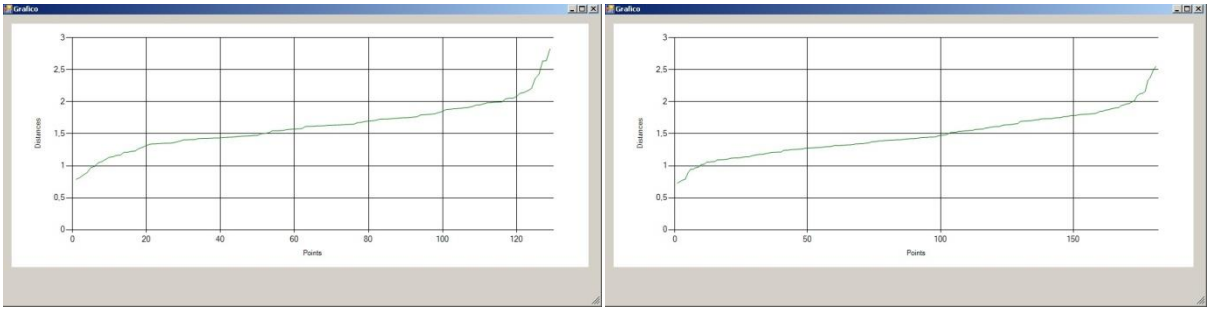


(b)

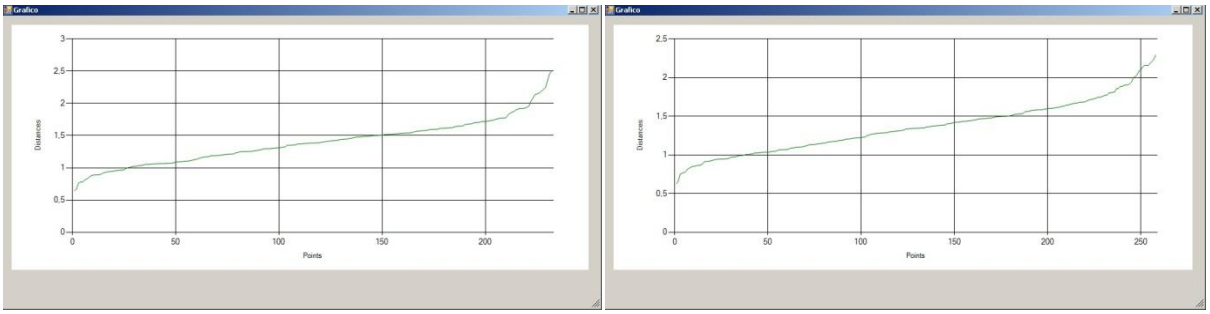


(c) (d)

Figura 6.5.1: *I grafici (a), (b), (c) e (d) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 2.*

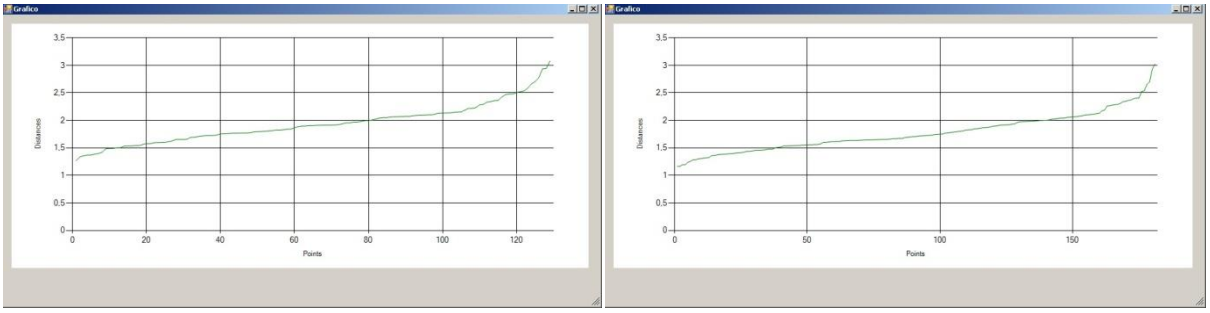


(e) (f)

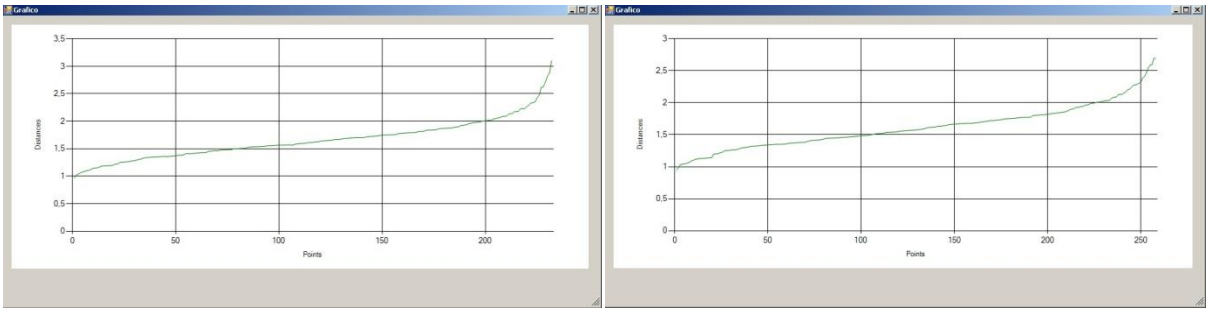


(g) (h)

Figura 6.5.2: *I grafici (e), (f), (g) e (h) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 5.*



(i) (j)



(k) (l)

Figura 6.5.3: *I grafici (i), (j), (k) e (l) mostrano, per ogni punto esaminato (in base alla percentuale di punti), l'andamento della distanza considerando la colonna k = 10.*

6.6 Risultati

Dopo aver ottenuto i grafici (*vds. sopra*) e scelto la percentuale di punti da mostrare *(in questo caso si è stabilito di esibire per ogni grafico il 100% dei punti)*, è stato ricavato il valore "***K***" ed il corrispondente valore *Epsilon* (***ε***) ***[Tabelle 6.6.1, 6.6.2, 6.6.3, 6.6.4, 6.6.5, 6.6.6, 6.6.7, 6.6.8, 6.6.9, 6.6.10, 6.6.11, 6.6.12, 6.6.13, 6.6.14, 6.6.15].***

Così facendo sono stati raggiunti i seguenti risultati:

Tabella 6.6.1: *US Census Data (1990) Data Set con k=2 (70477 tuple)*

% of data	Minpoint (K)	ε
4%	2	2700
6%	2	4200
8%	2	5500
10%	2	7100

Tabella 6.6.2: *US Census Data (1990) Data Set con k=5 (70477 tuple)*

% of data	Minpoint (K)	ε
4%	5	2700
6%	5	4200
8%	5	5600
10%	5	7100

Tabella 6.6.3: *US Census Data (1990) Data Set con k=10 (70477 tuple)*

% of data	Minpoint (K)	ε
4%	10	2700
6%	10	4200
8%	10	5600
10%	10	7000

Tabella 6.6.4: *Turkiye Student Evaluation Data Set con k=2 (5820 tuple)*

% of data	Minpoint (K)	ε
30%	2	1700
40%	2	2800
50%	2	2900
60%	2	3400

Tabella 6.6.5: *Turkiye Student Evaluation Data Set con k=5 (5820 tuple)*

% of data	Minpoint (K)	ε
30%	5	1700
40%	5	2800
50%	5	2900
60%	5	3400

Tabella 6.6.6: *Turkiye Student Evaluation Data Set con k=10 (5820 tuple)*

% of data	Minpoint (K)	ε
30%	10	1700
40%	10	2800
50%	10	2900
60%	10	3400

Tabella 6.6.7: *Electricity Data Set con k=2 (45781 tuple)*

% of data	Minpoint (K)	ε
6%	2	2300
8%	2	3500
10%	2	4500
12%	2	5600

Tabella 6.6.8: *Electricity Data Set con k=5 (45781 tuple)*

% of data	Minpoint (K)	ε
6%	5	2200
8%	5	3400
10%	5	4400
12%	5	5700

Tabella 6.6.9: *Electricity Data Set con k=10 (45781 tuple)*

% of data	Minpoint (K)	ε
6%	10	2200
8%	10	3500
10%	10	4400
12%	10	5700

Tabella 6.6.10*: Wholesale customers Data Set con k=2 (440 tuple)*

% of data	Minpoint (K)	ε
50%	2	220
70%	2	280
90%	2	380
100%	2	430

Tabella 6.6.11*: Wholesale customers Data Set con k=5 (440 tuple)*

% of data	Minpoint (K)	ε
50%	5	220
70%	5	300
90%	5	380
100%	5	430

Tabella 6.6.12*: Wholesale customers Data Set con k=10 (440 tuple)*

% of data	Minpoint (K)	ε
50%	10	200
70%	10	300
90%	10	380
100%	10	430

***Tabella 6.6.13**: User Knowledge Modeling Data Set con k=2 (258 tuple)*

% of data	Minpoint (K)	ε
50%	2	130
70%	2	180
90%	2	220
100%	2	260

***Tabella 6.6.14**: User Knowledge Modeling Data Set con k=5 (258 tuple)*

% of data	Minpoint (K)	ε
50%	5	125
70%	5	170
90%	5	220
100%	5	240

***Tabella 6.6.15**: User Knowledge Modeling Data Set con k=10 (258 tuple)*

% of data	Minpoint (K)	ε
50%	10	120
70%	10	170
90%	10	210
100%	10	230

Con la variazione dei "*minpoint*" scelti *(il quale valore corrisponde a K)* la corrispondente $\boldsymbol{\varepsilon}$ non rivela una correlazione direttamente proporzionale; infatti, il valore $\boldsymbol{\varepsilon}$ di volta in volta può leggermente variare a seconda del cambiamento sull'uso del K scelto.

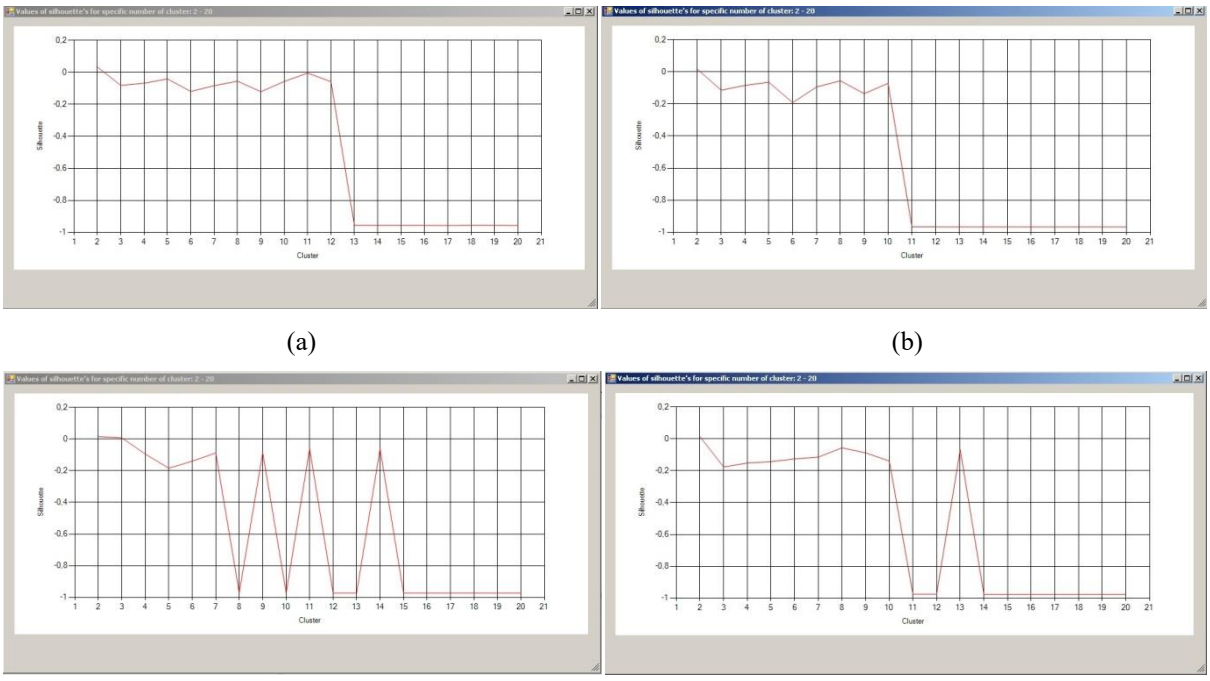
Tuttavia, all'aumentare del K scelto, è possibile notare come di volta in volta i grafici in questione dimostrino un cambiamento sempre più netto nei k^{th} *nearest neighbors.*

7. Discussioni

7.1 Silhouette media

In seguito all'estrapolazione dei dati, ecco alcune considerazioni visionando i grafici ***(Figura 7.1.x)*** che riguardano la *silhouette* media. Dai grafici ottenuti si nota come non ci sia una correlazione tra il numero di *tuple* considerate per ogni *dataset* e i picchi di *silhouette* da attribuire ad uno specifico cluster. Infatti, per ogni grafico i picchi di *silhouette* per ogni cluster possono variare parecchio *(anche per grafici appartenenti ad uno stesso dataset, ma con un maggior numero di tuple).*

Nei grafici proposti i valori di *silhouette* sono prevalentemente al di sotto della soglia; l'unica eccezione riguarda l'uso di 2 *clusters*, il cui uso consente di ottenere una *silhouette* media di poco oltre lo "0".



(a) (b)

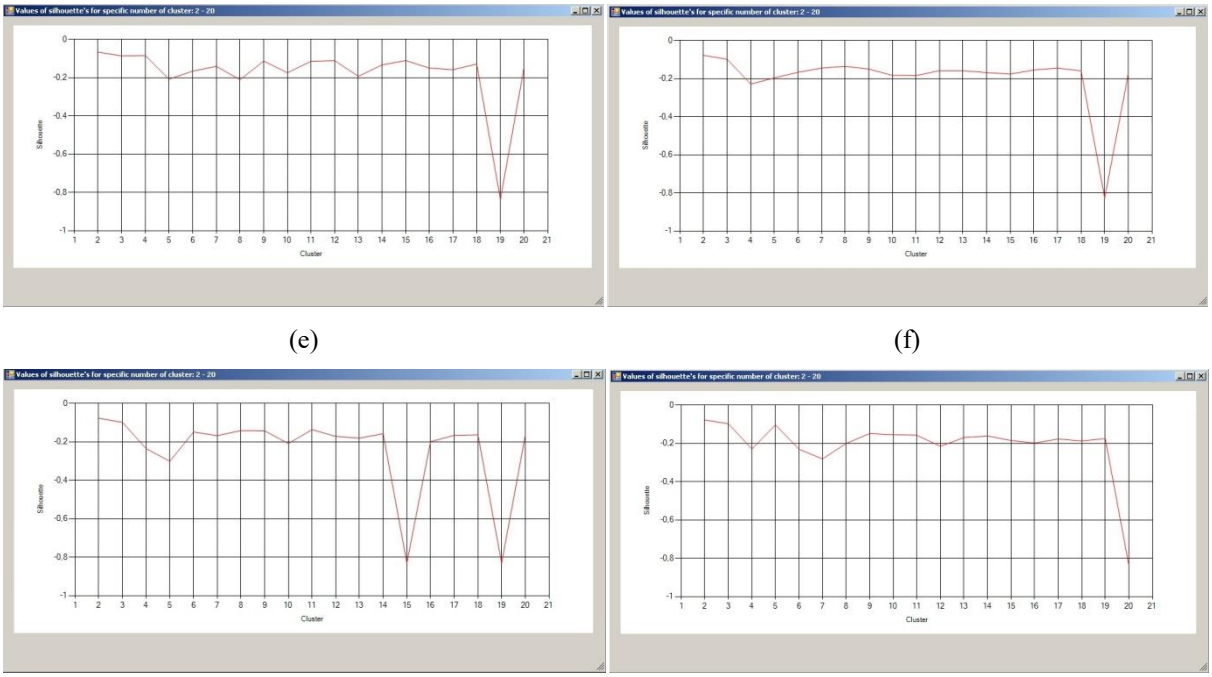
(c) (d)

Figura 7.1.1: *Sono qui raccolti i grafici degli andamenti della silhouette per i clusters 2-20:*

US Census Dataset: *10% [7047 tuple] (a);*
20% [14094 tuple] (b);
30% [21143 tuple] (c);
40% [28190 tuple] (d).

Nell'esempio sottostante non si ottengono *clusters* con valori medi di silhouette superiori la soglia. Anche in questo caso con 2 *clusters* si ottengono i migliori risultati, ma il notevole peggioramento della silhouette lo si raggiunge oltre l'uso di 13 *clusters*. I grafici, pur essendo in ordine crescente di tuple (70%, 80%, 90% e 100%) variano abbastanza; ad esempio, nel

terzo grafico si ha una considerevole diminuzione di silhouette media in corrispondenza dei 15 *clusters* (*cosa che nei restanti grafici non si verifica*).



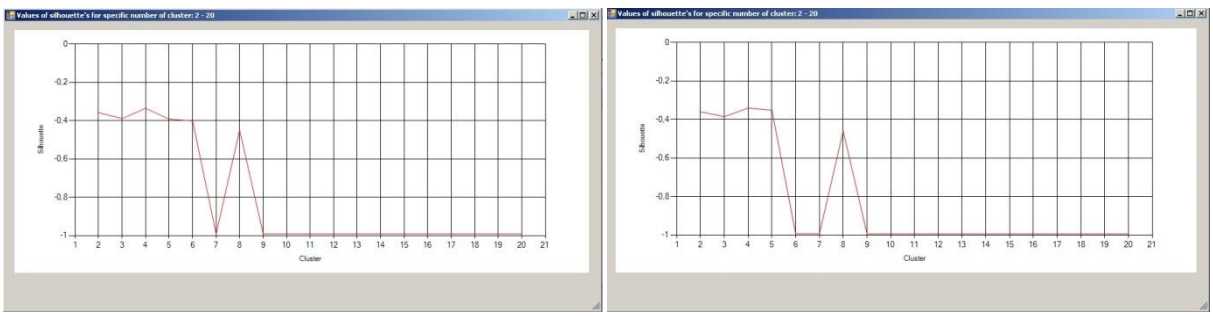
(e) (f)

(g) (h)

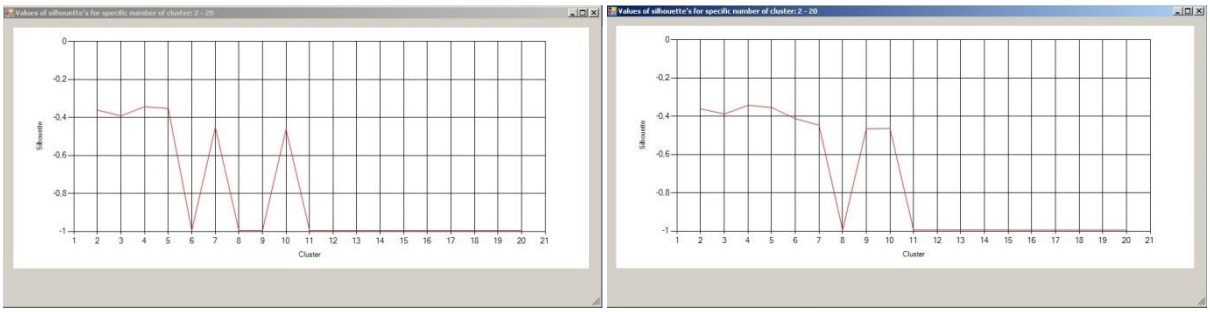
Figura 7.1.2: *Sono qui raccolti i grafici degli andamenti della silhouette per i clusters 2-20:*

Turkiye Student Evaluation - Dataset: *70% [4074 tuple] (e);*
80% [4656 tuple] (f);
90% [5238 tuple] (g);
100% [5820 tuple] (h).

Il terzo dataset consente di ottenere dei grafici con scarsi valori di silhouettes medi. In questo caso specifico, in corrispondenza dei 4 *clusters*, per tutti i grafici si ottengono i migliori valori di *silhouette (di poco superiore -0,4);* inoltre, oltrepassando la soglia dei 5 *clusters* si ottiene un peggioramento dei valori, cosa che ci permette di ottenere addirittura *clusters* con *silhouette* pari a -1 ovvero, il peggior valore di *silhouette* che si possa ottenere.



(i) (j)

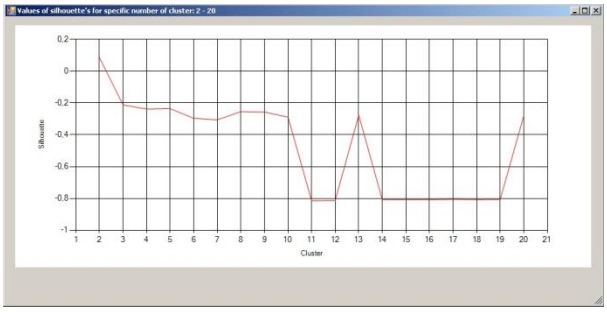


(k) (l)

Figura 7.1.3: *Sono qui raccolti i grafici degli andamenti della silhouette per i clusters 2-20:*

Electricity – Dataset: *30% [13761 tuple] (i);*
50% [22890 tuple] (j);
70% [32046 tuple] (k);
80% [36624 tuple] (l).

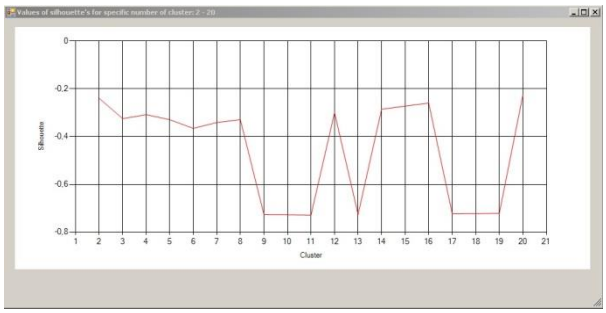
Per i due restanti dataset *(il quarto ed il quinto)* non è possibile fare considerazioni attraverso paragoni di differenti grafici giacchè entrambi sono provvisti di un solo grafico come possibile confronto *(è stato considerato esclusivamente il 100% delle tuple, in quanto i due dataset sono di ridotte dimensioni).*



(m)

Figura 7.1.4: *Sono qui raccolti i grafici degli andamenti della silhouette per i clusters 2-20:*

Whosale customers - Dataset: *100% [440 tuple] (m).*



(n)

Figura 7.1.5: *Sono qui raccolti i grafici degli andamenti della silhouette per i clusters 2-20:*

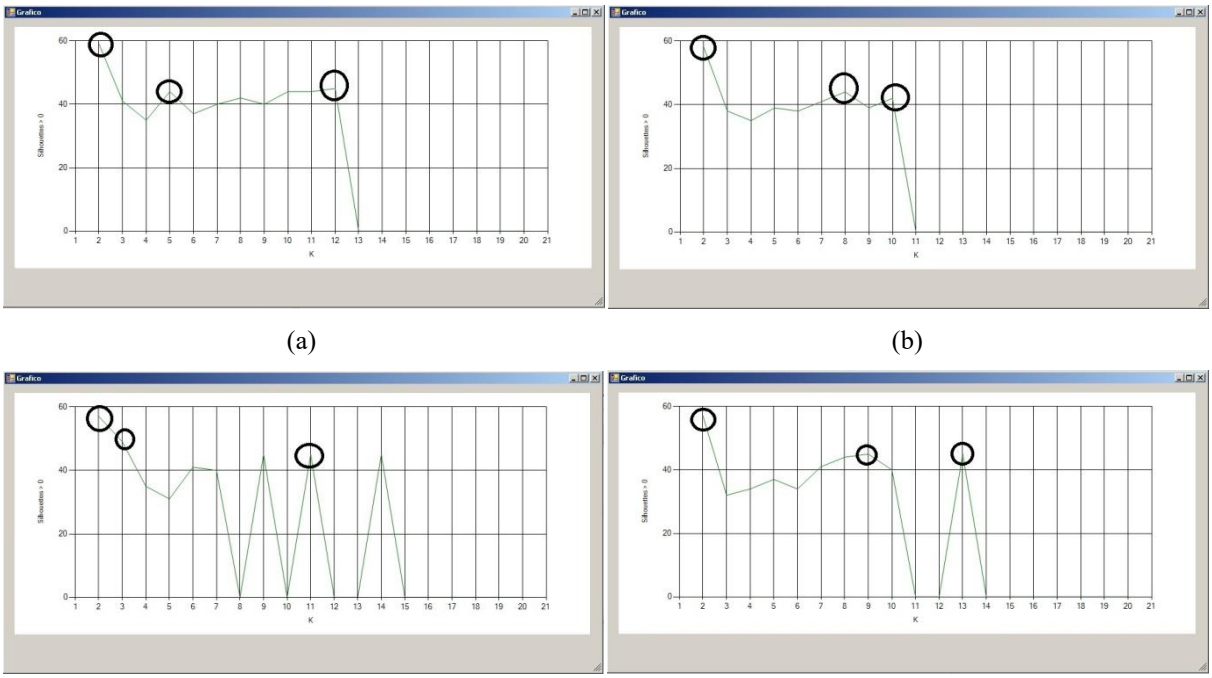
User Knowledge Modeling – Dataset: *100% [258 tuple] (n).*

In conclusione, si può affermare come solo il primo dataset sia riuscito a soddisfare le condizioni di *silhouette media* oltre la soglia "0"; infatti, tra i *clusters* a disposizione il primo corrisponde a quello con il maggior numero di dati sotto elaborazione *(con il 40% dei dati sotto elaborazione si arriva alle 28190 tuple con l'impiego di 9 attributi differenti per ogni singola tupla).*

7.2 Percentuale di silhouette oltre la soglia "0"

Sotto è possibile osservare i grafici che mostrano *(in percentuale),* per ogni cluster, il numero di punti superanti la soglia "0" di *silhouette*. Dai grafici ottenuti ci si rende conto come spesso il picco dei valori superiori alla soglia si ottenga con l'uso di un ridotto numero di cluster *(spesso 2, ma di solito inferiore ai 10);* inoltre, esaminando gli ultimi due grafici, formati da dataset con un ridotto numero di tuple, si nota come i cluster con i migliori valori di *silhouette* superiori la soglia risultino essere quelli di maggiore dimensione *(es. cluster 20 per l'ultimo grafico).*

Nei grafici sottostanti con 2 *clusters* si ottiene la più alta percentuale di punti oltre la soglia "0". Per quanto riguarda le seconde e terze percentuali più alte i *clusters* da considerare variano molto.



(a) (b)

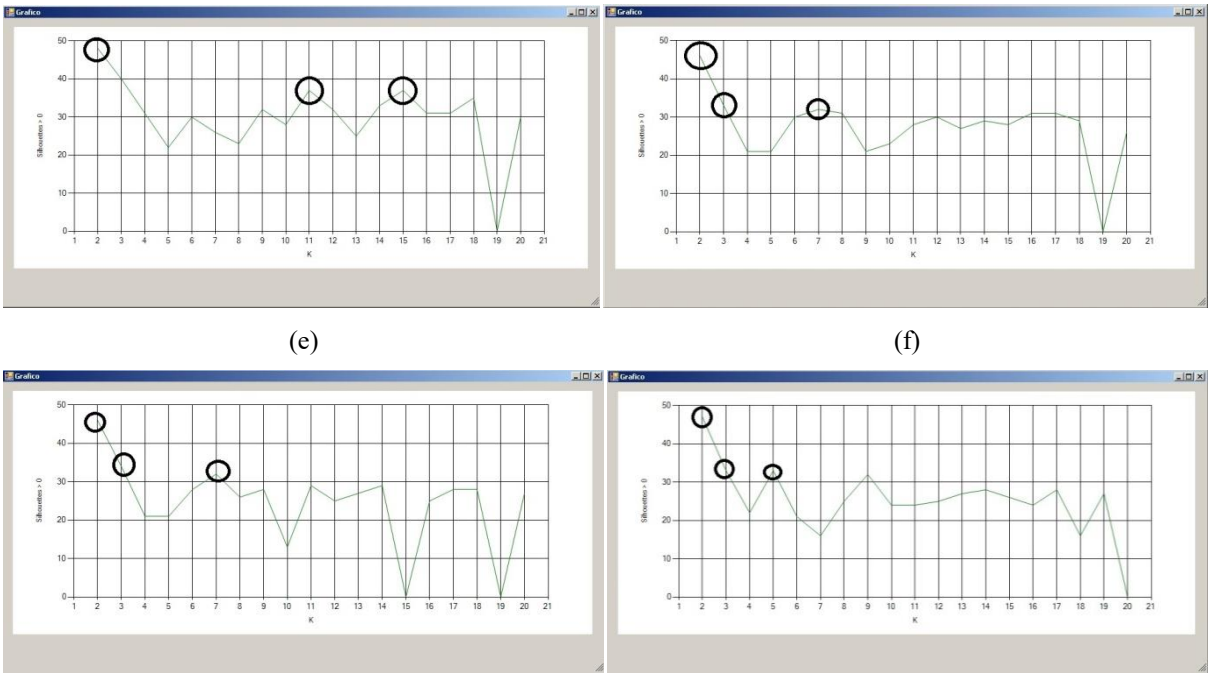
(c) (d)

Figura 7.2.1: *Sono qui raccolti i grafici con le percentuali degli andamenti della silhouette per i clusters 2-20:*

US Census Dataset: *10% [7047 tuple] (a);*
20% [14094 tuple] (b);
30% [21143 tuple] (c);

40% [28190 tuple] (d).

Anche nel caso sottostante i più alti valori percentuali si ottengono con l'uso di 2 *clusters*, mentre in tre casi su quattro il secondo miglior valore lo si ha con l'uso di 3 *clusters*.



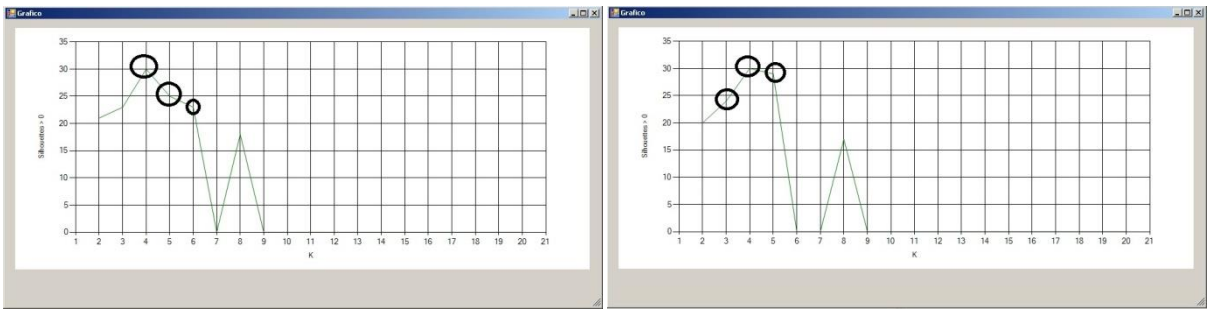
(e) (f)

(g) (h)

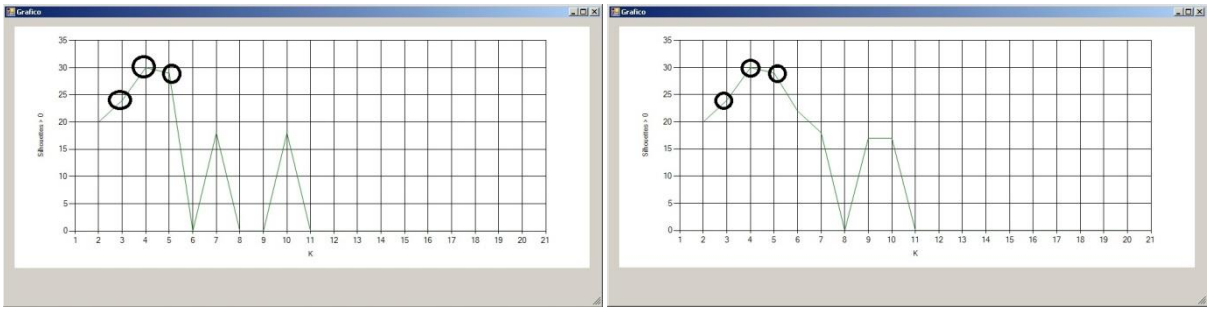
Figura 7.2.2: *Sono qui raccolti i grafici con le percentuali degli andamenti della silhouette per i clusters 2-20:*

Turkiye Student Evaluation - Dataset: *70% [4074 tuple] (e);*
80% [4656 tuple] (f);
90% [5238 tuple] (g);
100% [5820 tuple] (h).

Nei grafici sottostanti le più alte percentuali di *silhouette* oltre la soglia si raggiungono con 4 *clusters*; in questo specifico dataset è evidente come i tre risultati più soddisfacenti si ottengono con l'uso di 3, 4 e 5 *clusters (senza considerare l'ordine preciso di come si presentano rispetto alla percentuale dei valori oltre la soglia).*



(i) (j)

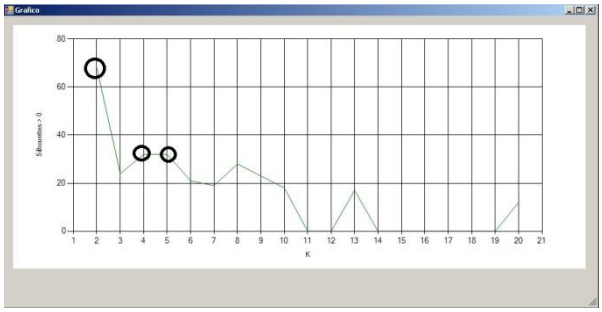


(k) (l)

Figura 7.2.3: *Sono qui raccolti i grafici con le percentuali degli andamenti della silhouette per i clusters 2-20:*

Electricity – Dataset: *30% [13761 tuple] (i);*
50% [22890 tuple] (j);
70% [32046 tuple] (k);
80% [36624 tuple] (l).

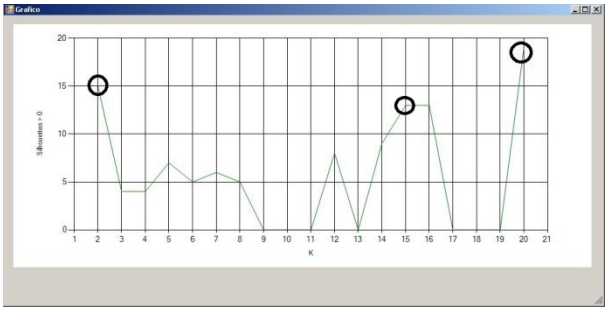
Per i due restanti dataset *(il quarto ed il quinto)* non è possibile fare considerazioni attraverso paragoni di differenti grafici giacchè entrambi sono provvisti di un solo grafico come possibile confronto *(è stato considerato esclusivamente il 100% delle tuple, in quanto i due dataset sono di ridotte dimensioni).*



(m)

Figura 7.2.4: *Sono qui raccolti i grafici con le percentuali degli andamenti della silhouette per i clusters 2-20:*

Whosale customers – Dataset: *100% [440 tuple] (m).*



(n)

Figura 7.2.5: *Sono qui raccolti i grafici con le percentuali degli andamenti della silhouette per i clusters 2-20:*

User Knowledge Modeling – Dataset: *100% [258 tuple] (n).*

Tranne che per il terzo, i restanti dataset coinvolti nelle elaborazioni hanno mostrato un alto valore percentuale di *silhouette* oltre la soglia "0" con l'uso di 2 *clusters*.

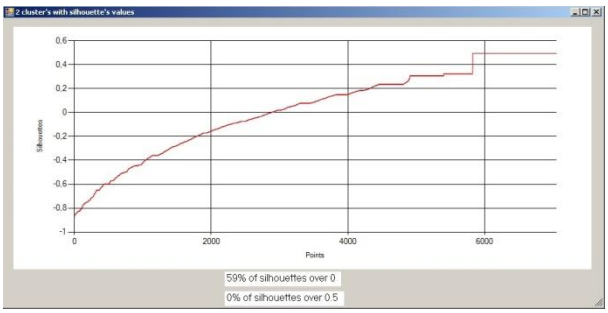
7.3 Silhouette media per "N" clusters

Prendendo spunto dai dati elaborati si può costatare per i 5 dataset *(considerati ad ogni percentuale),* quale sia il numero di *clusters* che consenta di ottenere il miglior valore di *silhouette* media ***(Figura 7.3.1)***.

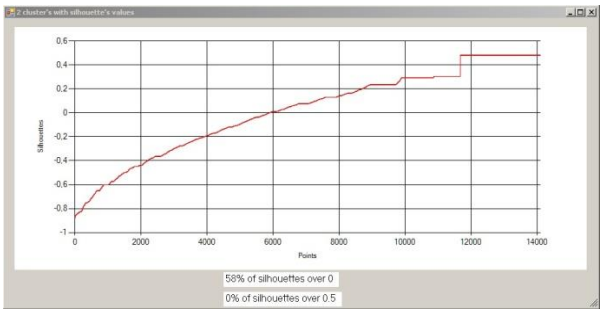
Nonostante sia stato deciso di considerare per ogni dataset una differente percentuale di *tuple*, dai risultati ottenuti non si riscontra alcun legame tra la quantità di dati elaborati e l'andamento della *silhouette*. Infatti, a dimostrazione di ciò, si può notare come per il primo dataset, considerato all'aumentare delle *tuple,* si ottenga una diminuzione del numero di valori di *silhouette* superanti la soglia *(passando da 59% a 57%)*, cosa che non si ha riscontro con l'utilizzo del dataset successivo *(l'aumento delle istanze provoca una diminuzione da 48% al 46% di valori superiori alla soglia, per salire poi al 47%).*

A fronte di questi risultati si può sostenere come non sia la quantità dei dati a permettere il conseguimento di una buona percentuale di valori di *silhouette* superiori la soglia "0", ma la "*qualità*" degli stessi.

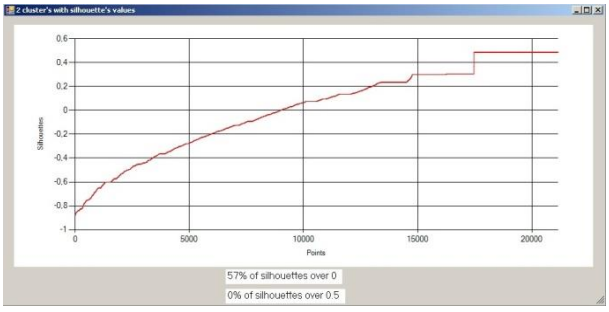
Nei grafici sottostanti *(disposti in ordine crescente di tuple considerate)* si ottengono andamenti decrescenti per quanto riguarda la percentuale di punti oltre la soglia "0" di *silhouette*.



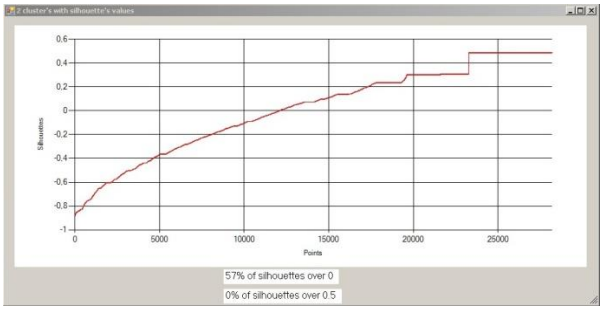
(a)



(b)



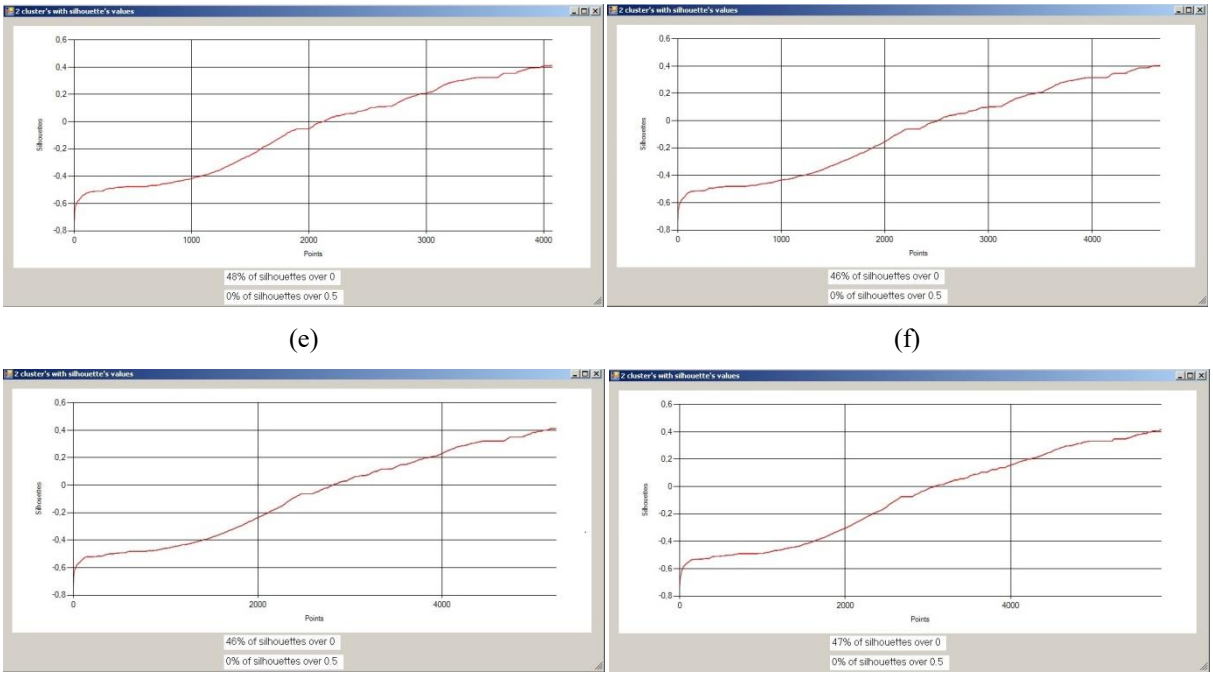
(c)



(d)

Figura 7.3.1: *Grafici degli andamenti della silhouette (disposti in andamento crescente per percentuale di dati elaborati) per gli "N" clusters con miglior valor medio di silhouette. Il primo dataset dimostra di avere il 59% (a), 58% (b), 57% (c) e 57% (d) di valori superiori alla soglia "0" di silhouette;*

Nell'esempio sottostante, diversamente dal caso mostrato in precedenza, la percentuale dei valori di *silhouette* oltre la soglia zero dimostra un andamento altalenante (48%, 46%, 46% e 47%).

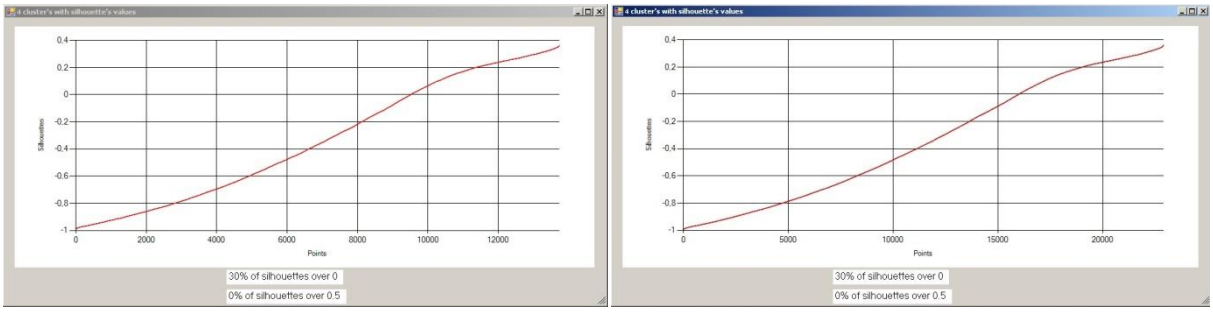


(e) (f)

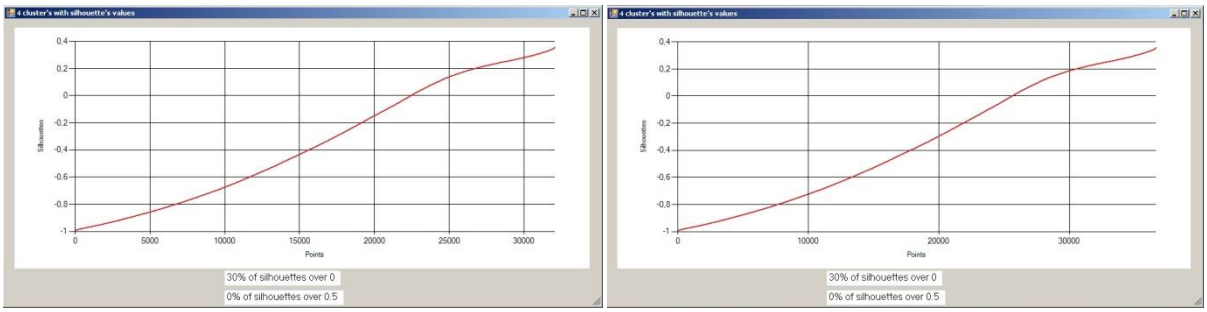
(g) (h)

Figura 7.3.2: *Grafici degli andamenti della silhouette (disposti in andamento crescente per percentuale di dati elaborati) per gli "N" clusters con miglior valor medio di silhouette. Il secondo dataset contiene il 48% (e), 46% (f), 46% (g) e 47% (h) dei valori superiori la soglia.*

In questo caso, nonostante l'aumento del numero di tuple considerate, i quattro grafici consentono di raggiungere la stessa percentuale di punti oltre la soglia "0".



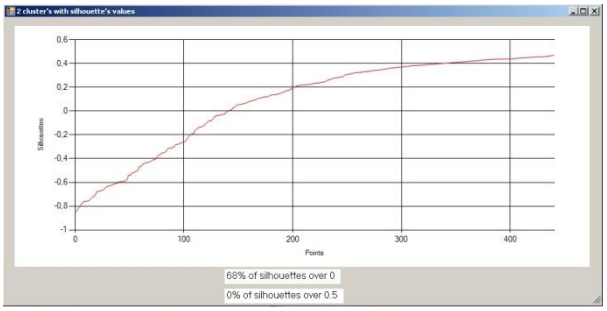
(i) (j)



(k) (l)

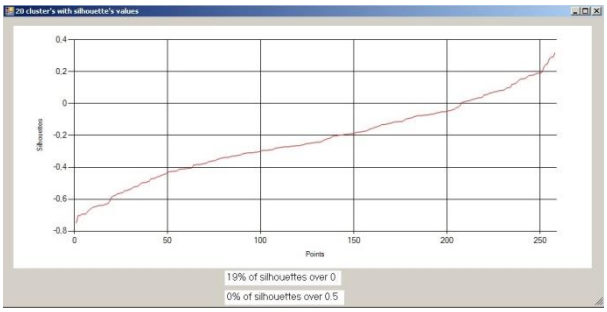
Figura 7.3.3: *Grafici degli andamenti della silhouette (disposti in andamento crescente per percentuale di dati elaborati) per gli "N" clusters con miglior valor medio di silhouette. Il terzo dataset ne contiene il 30% (i), 30% (j), 30% (k) e 30% (l).*

Per i due restanti dataset *(il quarto ed il quinto)* non è possibile fare considerazioni attraverso paragoni di differenti grafici giacchè entrambi sono provvisti di un solo grafico come possibile confronto *(è stato considerato esclusivamente il 100% delle tuple, in quanto i due dataset sono di ridotte dimensioni).*



(m)

Figura 7.3.4: *Grafici degli andamenti della silhouette (disposti in andamento crescente per percentuale di dati elaborati) per gli "N" clusters con miglior valor medio di silhouette. Il quarto dataset contiene il 68% (m).*



(n)

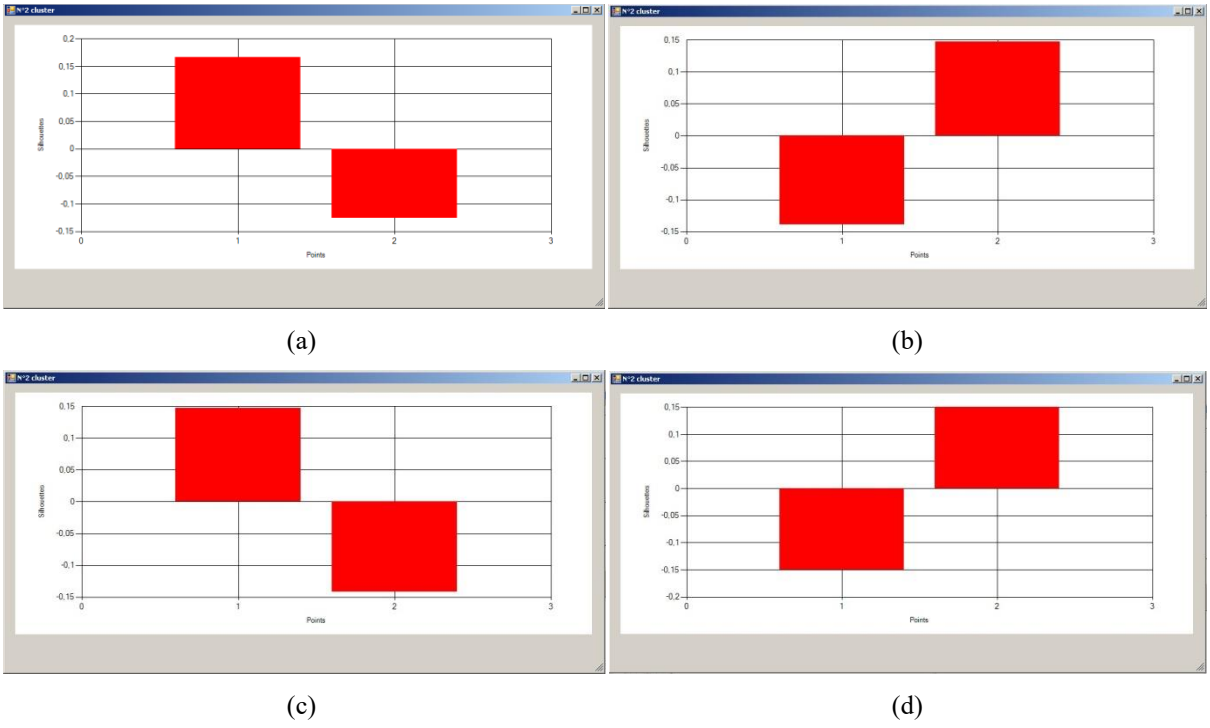
Figura 7.3.5: *Grafici degli andamenti della silhouette (disposti in andamento crescente per percentuale di dati elaborati) per gli "N" clusters con miglior valor medio di silhouette. Il quinto il 19% (n).*

7.4 Silhouette media per clusters

Saranno analizzati, ora, i grafici ottenuti ***(Figura 7.4.1)*** attraverso la lettura dei dati relativi la *silhouette* media conseguita simultaneamente con l'uso di "N" *clusters*; appare evidente che con l'uso di un numero di cluster superiore a due il valore di *silhouette* media per gli "N" cluster coinvolti tenda ad abbassarsi in modo sensibile.

Dai grafici si riesce facilmente a notare che quando il numero di cluster coinvolti è superiore a due il valore medio di *silhouette*, relativo ad ogni singolo cluster, tenda a scendere sotto la soglia "0".

I quattro grafici sottostanti presentano l'uso di 2 *clusters* uno con valore medio di *silhouette* pari o superiore allo 0,15 ed il secondo sotto la soglia.



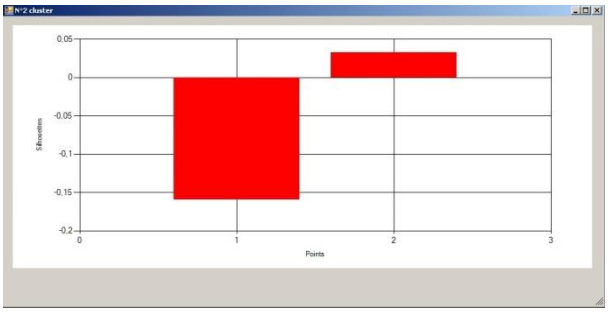
(a) (b)

(c) (d)

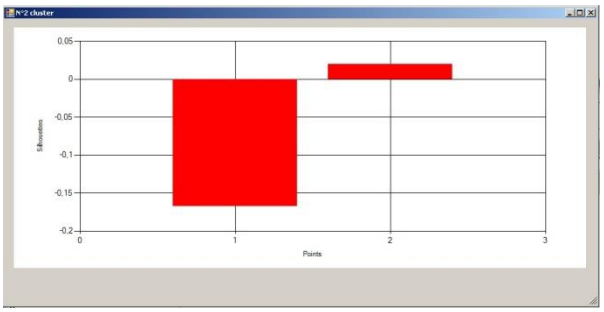
Figura 7.4.1: *Grafici riguardanti la silhouette media; per ogni dataset sono mostrati (per ogni percentuale di dati elaborata) quattro grafici che con "N" clusters hanno dato la miglior silhouette media:*

US Census Dataset: *10% [7047 tuple] (a);*
20% [14094 tuple] (b);
30% [21143 tuple] (c);
40% [28190 tuple] (d).

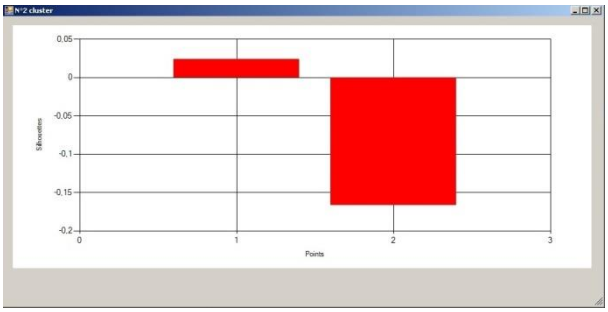
I quattro grafici sottostanti presentano l'uso di 2 *clusters*, uno con valore medio di *silhouette* di poco superiore alla soglia ed il secondo al di sotto.



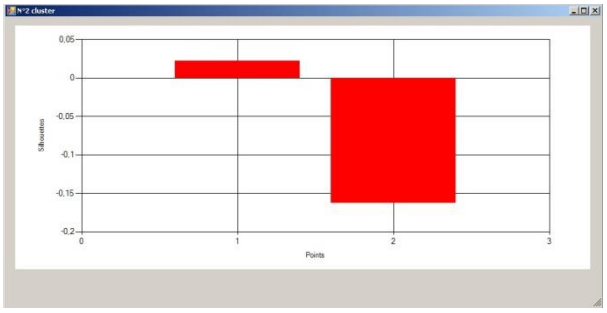
(e)



(f)



(g)



(h)

Figura 7.4.2: *Grafici riguardanti la silhouette media; per ogni dataset sono mostrati (per ogni percentuale di dati elaborata) quattro grafici che con "N" clusters hanno fornito la miglior silhouette media:*

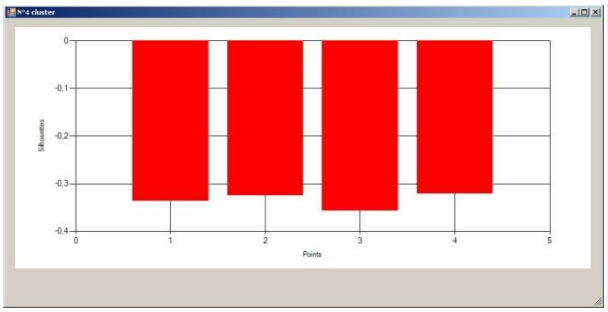
Turkiye Student Evaluation - Dataset: *70% [4074 tuple] (e);*

80% [4656 tuple] (f);

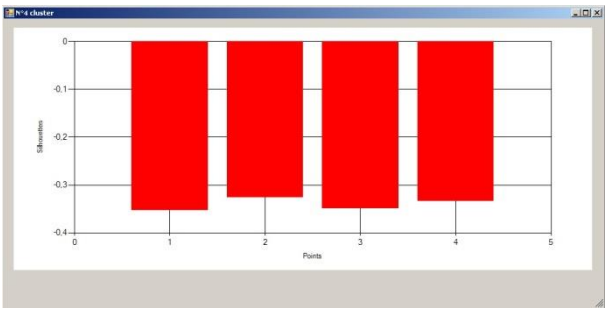
90% [5238 tuple] (g);

100% [5820 tuple] (h).

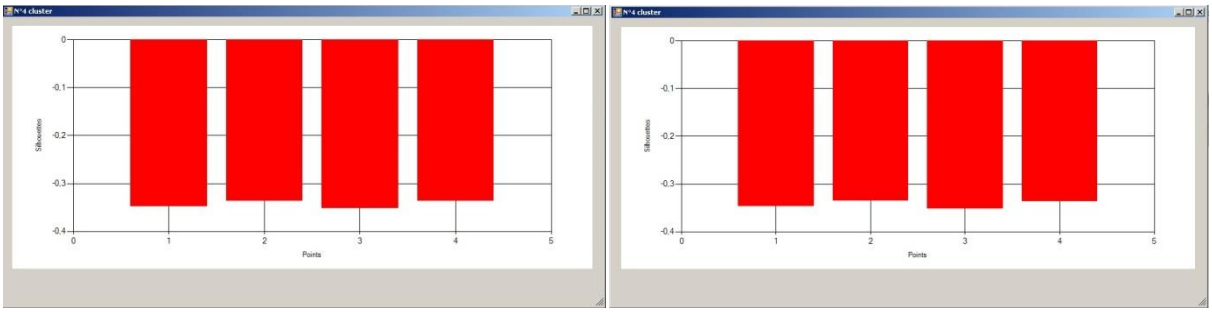
I quattro grafici sottostanti presentano l'uso di 4 *clusters* e nessuno presenta un valor medio di *silhouette* oltre la soglia.



(i)



(j)

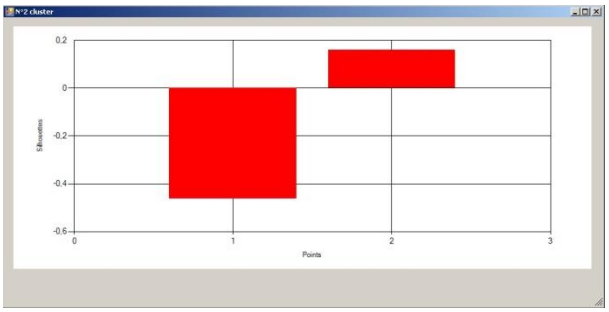


(k) (l)

Figura 7.4.3: *Grafici riguardanti la silhouette media; per ogni dataset sono mostrati (per ogni percentuale di dati elaborata) quattro grafici che con "N" clusters hanno fornito la miglior silhouette media:*

Electricity – Dataset: *30% [13761 tuple] (i);*
50% [22890 tuple] (j);
70% [32046 tuple] (k);
80% [36624 tuple] (l).

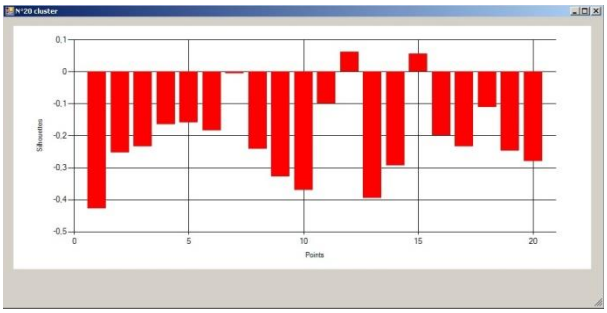
Per i due restanti dataset *(il quarto ed il quinto)* non è possibile fare considerazioni attraverso paragoni di differenti grafici giacchè entrambi sono provvisti di un solo grafico come possibile confronto *(è stato considerato esclusivamente il 100% delle tuple, in quanto i due dataset sono di ridotte dimensioni).*



(m)

Figura 7.4.4: *Grafici riguardanti la silhouette media; per ogni dataset sono mostrati (per ogni percentuale di dati elaborata) quattro grafici che con "N" clusters hanno fornito la miglior silhouette media:*

Whosale customers – Dataset: *100% [440 tuple] (m).*



(n)

Figura 7.4.5: *Grafici riguardanti la silhouette media; per ogni dataset sono mostrati (per ogni percentuale di dati elaborata) quattro grafici che con "N" clusters hanno fornito la miglior silhouette media:*

User Knowledge Modeling – Dataset: *100% [258 tuple] (n).*

7.5 Andamento della silhouette per "N" clusters

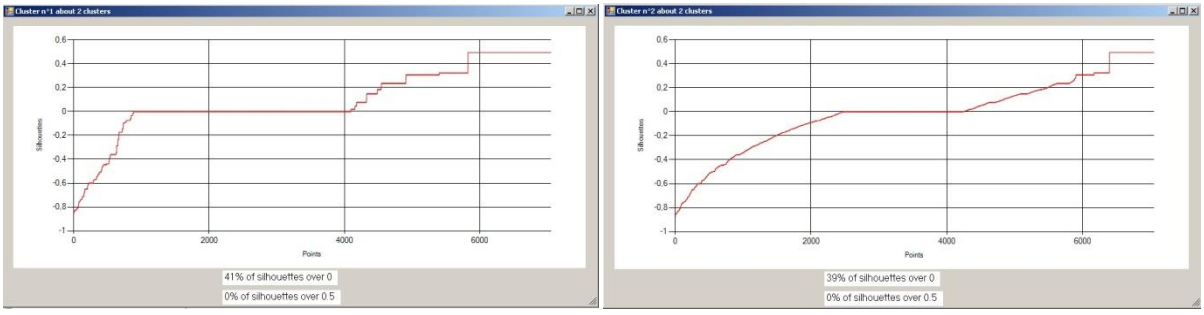
Oltre quanto è stato mostrato sinora, si è tenuto conto di raccogliere i dati riguardanti l'andamento della *silhouette* per ***"N"*** *clusters (considerando i due con andamento medio di silhouette migliore)*. In questo modo si può compiere un'esatta valutazione su quale tra gli ***"N"*** *clusters* abbia il miglior valore di *silhouette (valore percentuale ottenuto da quelli superiori alla soglia 0)*. Si prenderanno in considerazione solo gli "N" *clusters* che hanno consentito di raggiungere il miglior valor medio di *silhouette*.

Dai risultati ottenuti ***(Figura 7.5.x)*** si nota che, quando i *clusters* coinvolti sono di un numero ridotto *(solitamente 2),* l'andamento della *silhouette* abbia il maggior numero di punti superiori alla soglia "0". Esaminando il dataset ***Us Census Data*** *(2 cluster)* si nota come i grafici abbiano come percentuale di punti oltre la soglia del 41%-39%, 16%-48%, 41%-38%, 15%-48%; anche per il secondo dataset ***Turkiye Student Evaluation Data*** vengono considerati 2 cluster per ogni percentuale di tuple ottenendo il 19%-39%, 19%-37%, 26%-31%, 26%-30%. Per il terzo dataset ***Electricity Data*** *(4 cluster)* si hanno invece percentuali differenti; per le 4 percentuali si hanno i medesimi valori ovvero il 7%, 13%,17%,20%. Infine, per gli ultimi due dataset ***Whosale Customers Data*** e ***User Knowledge Modeling Data*** si hanno rispettivamente le seguenti percentuali: con l'impiego di 2 cluster l'1% e 66%, mentre con l'uso di 20 cluster lo 0%, 0%, 2%, 3%, 5%, 6%, 8%, 9%, 9%, 9%, 10%, 10%, 8%, 7%, 10%, 11%, 12%, 13%, 13%, 10%.

E' molto probabile che la maggiore/minore quantità di punti superiori allo "0" sia dovuta ad una ripartizione dei punti superiori alla soglia che, essendo posta in più cluster, è suddivisa abbassando, di conseguenza, la percentuale per ogni singolo cluster.

7.5.1 US Census Data - Dataset

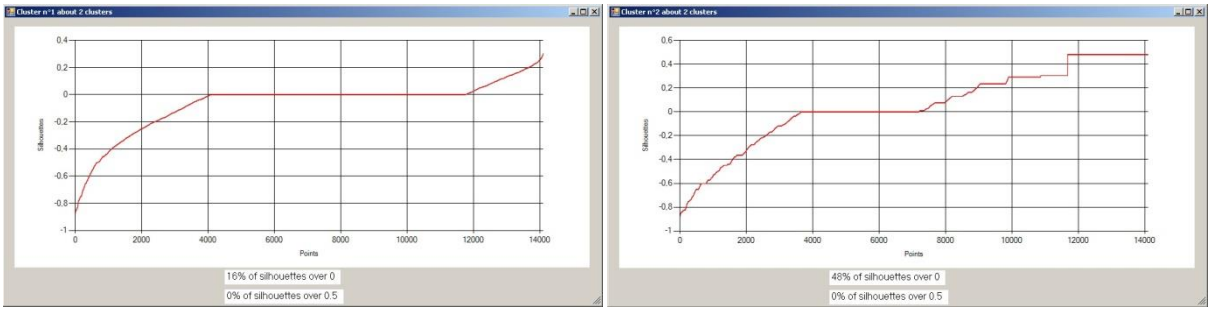
10% (7047 number of instance) – 2 clusters best silhouette's value



(a) (b)

Figura 7.5.1.1: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0" sia del 41% (a) e 39% (b).*

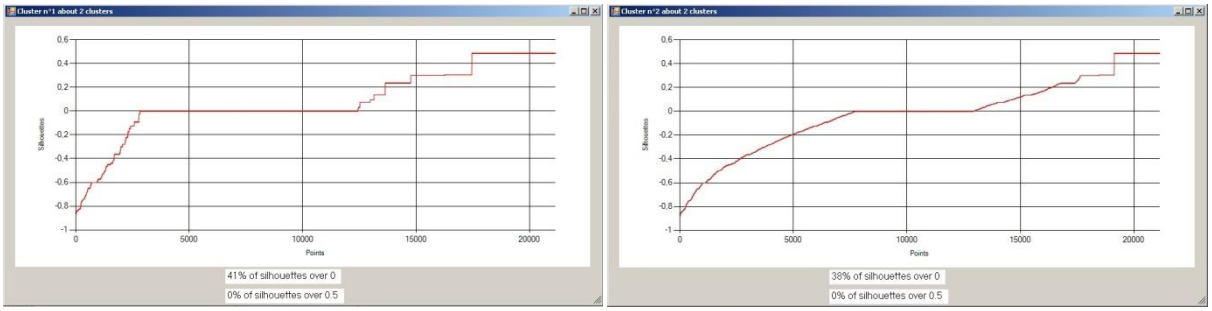
20% (14094 number of instance) – 2 clusters best silhouette's value



(a) (b)

Figura 7.5.1.2: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0" sia del 16% (a) e 48% (b).*

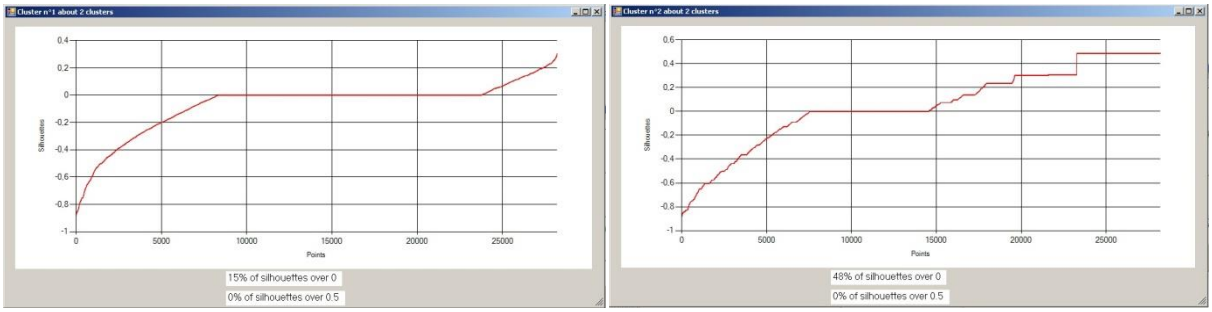
30% (21143 number of instance) – 2 clusters best silhouette's value



(a) (b)

Figura 7.5.1.3: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0" sia del 41% (a) e 38% (b).*

40% (28190 number of instance) – 2 clusters best silhouette's value

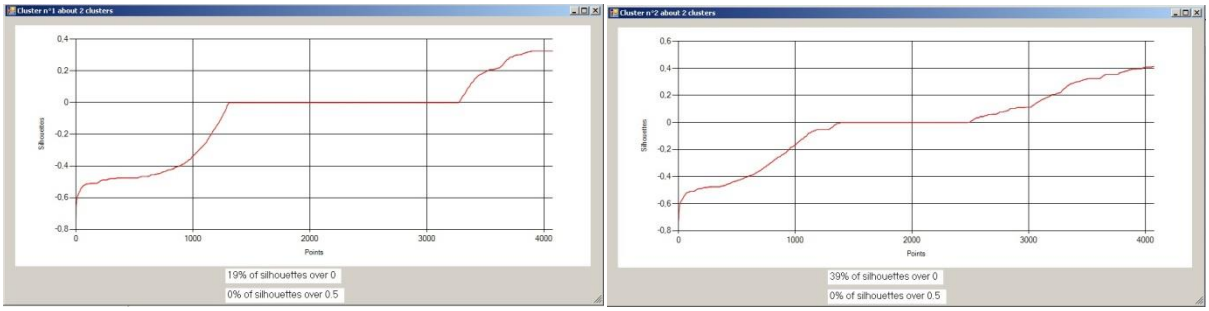


(a) (b)

Figura 7.5.1.4: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine d visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0"sia del 15% (a) e 48% (b).*

7.5.2 Turkiye Student Evaluation - Dataset

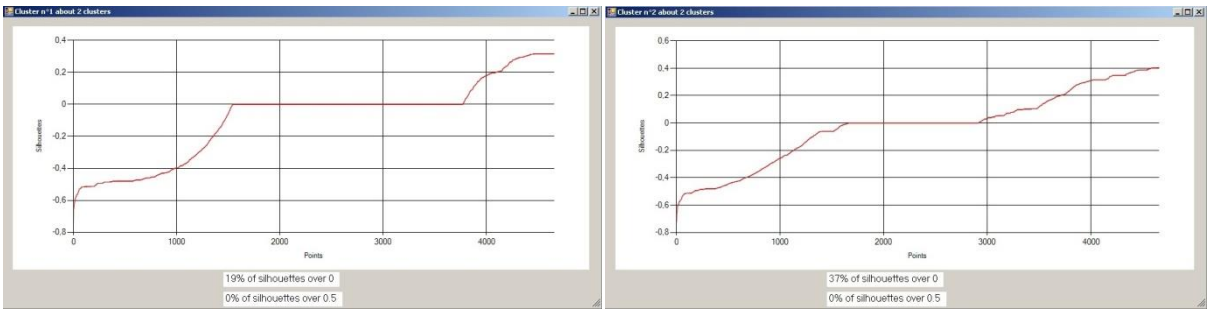
70% (4074 number of instance) – 2 clusters best silhouette's value



(a) (b)

Figura 7.5.2.1: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0" sia del 19% (a) e 39% (b).*

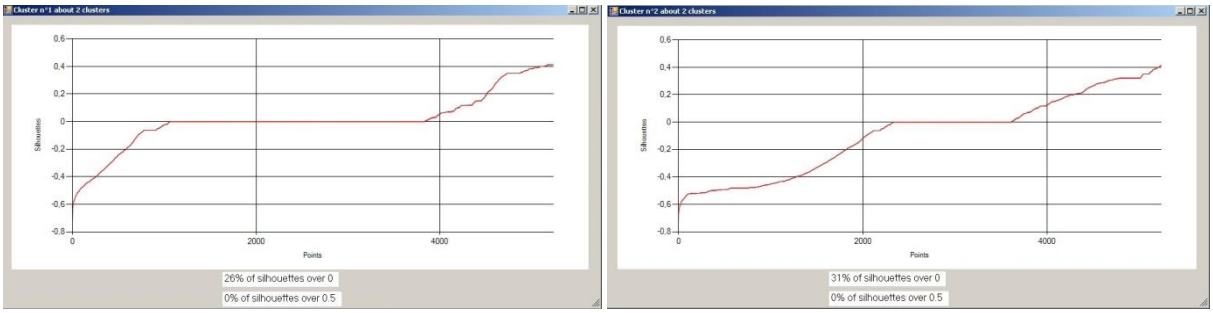
80% (4656 number of instance) – 2 clusters best silhouette's value



(a) (b)

Figura 7.5.2.2: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0"sia del 19% (a) e 37% (b).*

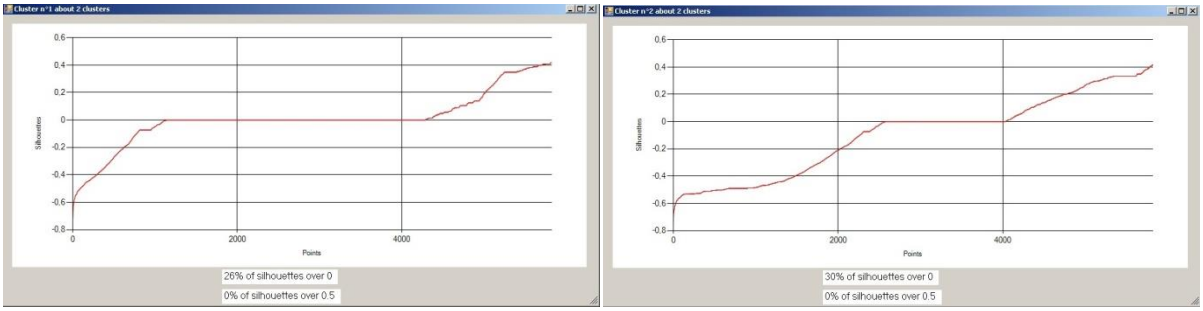
90% (5238 number of instance) – 2 clusters best silhouette's value



(a) (b)

Figura 7.5.2.3: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0" sia del 26% (a) e 31% (b).*

100% (5820 number of instance) – 2 clusters best silhouette's value

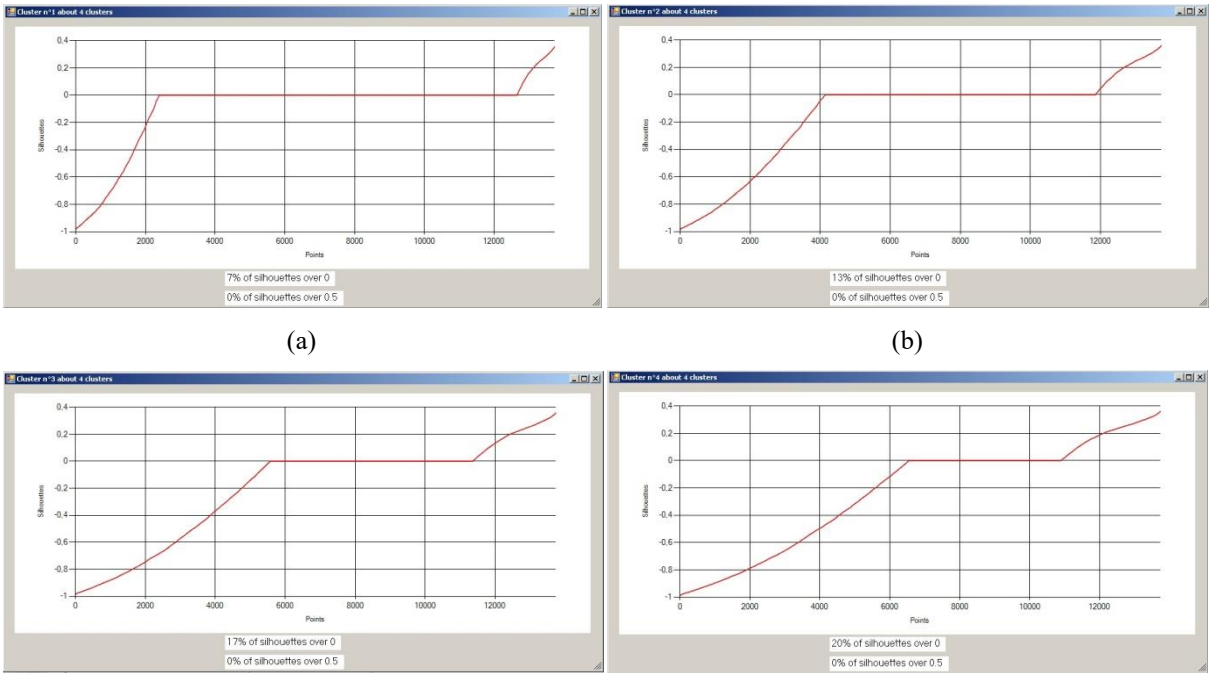


(a) (b)

Figura 7.5.2.4: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0"sia del 26% (a) e 30% (b).*

7.5.3 Electricity - Dataset

30% (13761 number of instance) – 4 clusters best silhouette's value

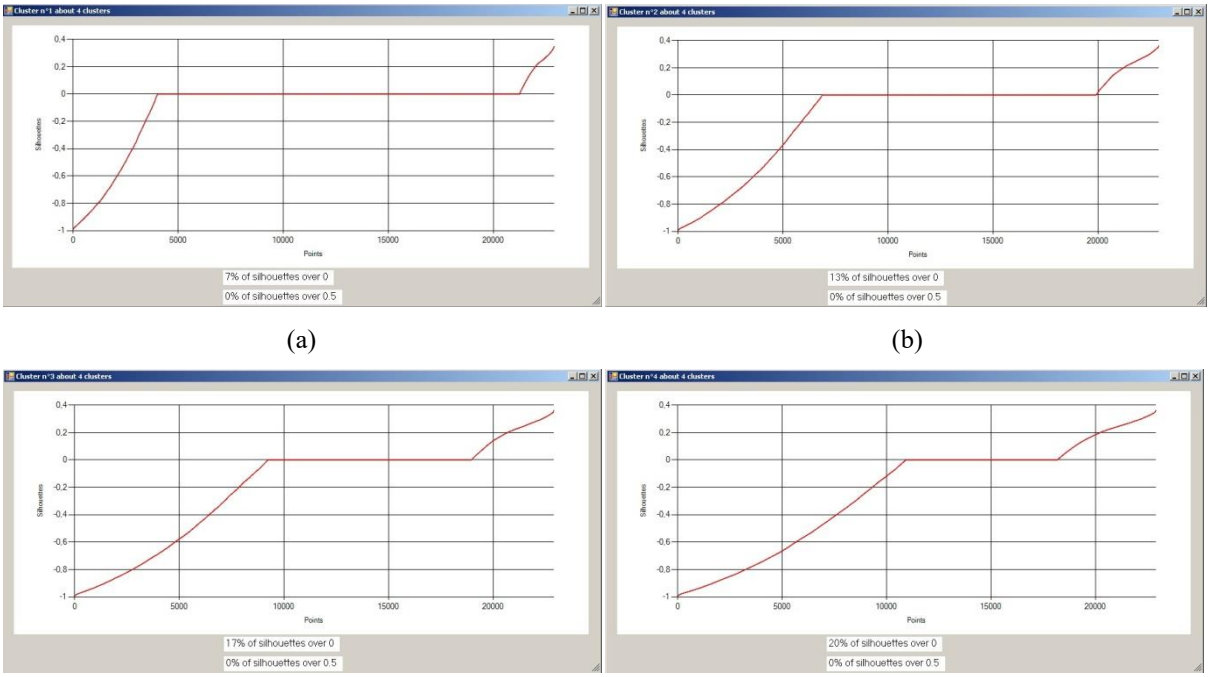


(a) (b)

(c) (d)

Figura 7.5.3.1: *I quattro grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0"sia del 7% (a), 13% (b), 17% (c) e 20%(d).*

50% (22890 number of instance) – 4 clusters best silhouette's value

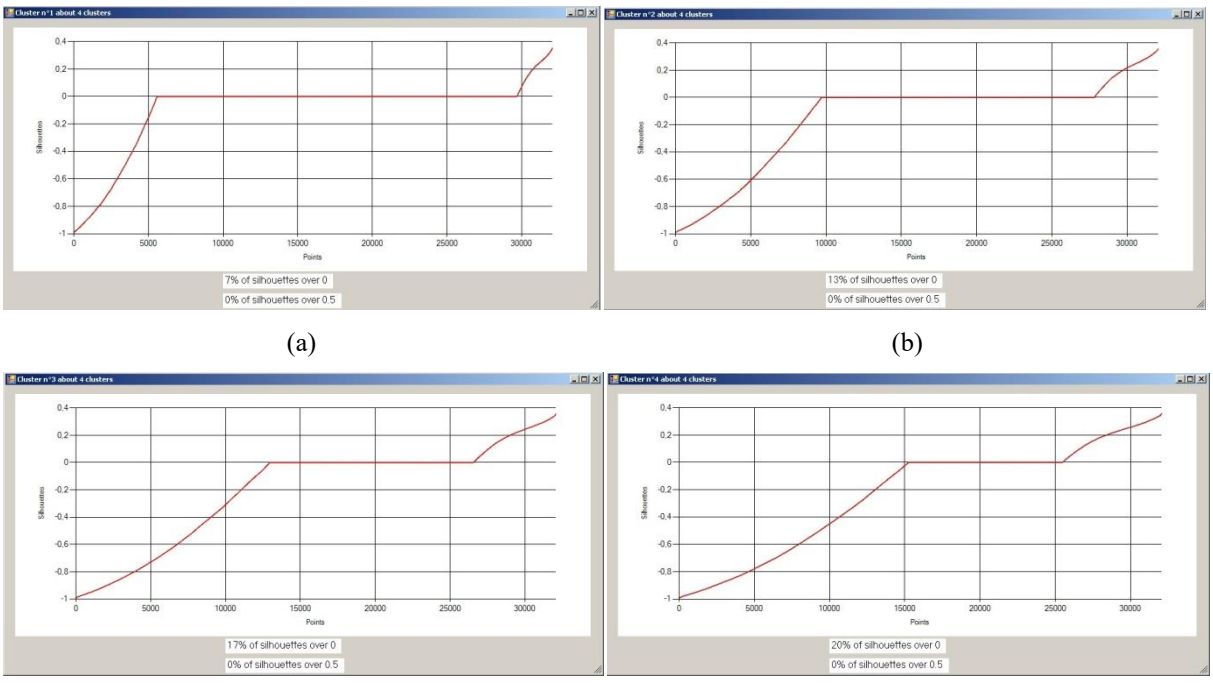


(a) (b)

(c) (d)

Figura 7.5.3.2: *I quattro grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0" sia del 7% (a), 13% (b), 17% (c) e 20%(d).*

70% (32046 number of instance) – 4 clusters best silhouette's value

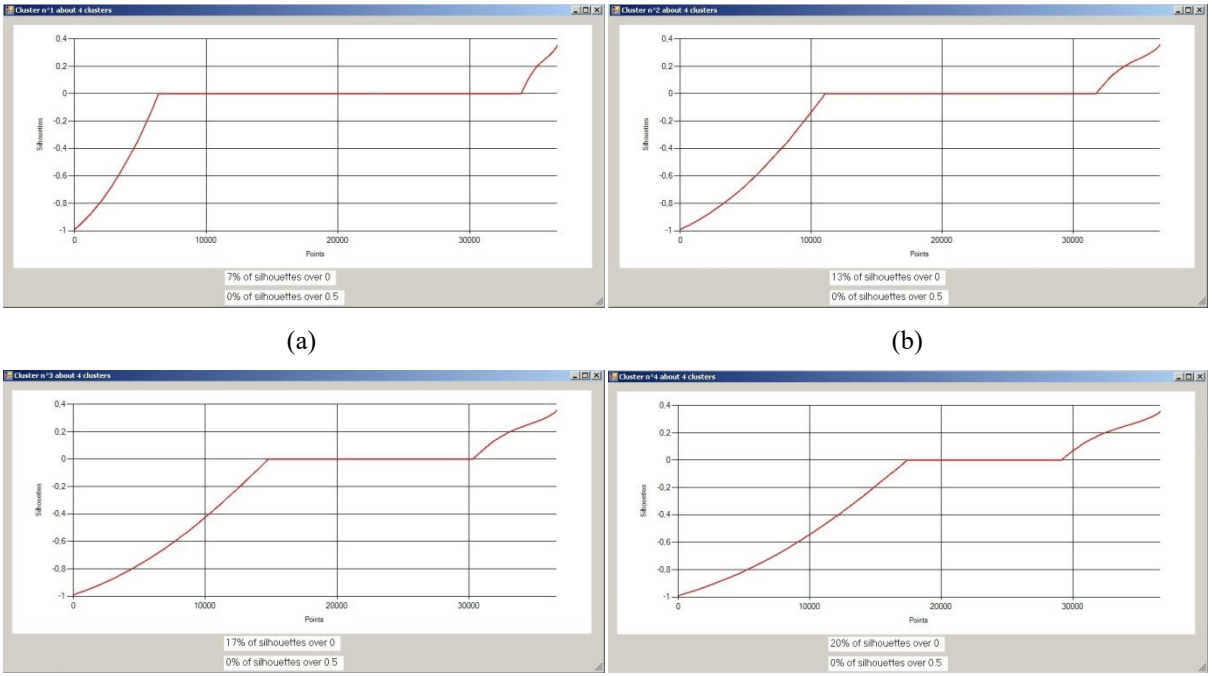


(a) (b)

(c) (d)

Figura 7.5.3.3: *I quattro grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0"sia del 7% (a), 13% (b), 17% (c) e 20%(d).*

80% (36624 number of instance) – 4 clusters best silhouette's value



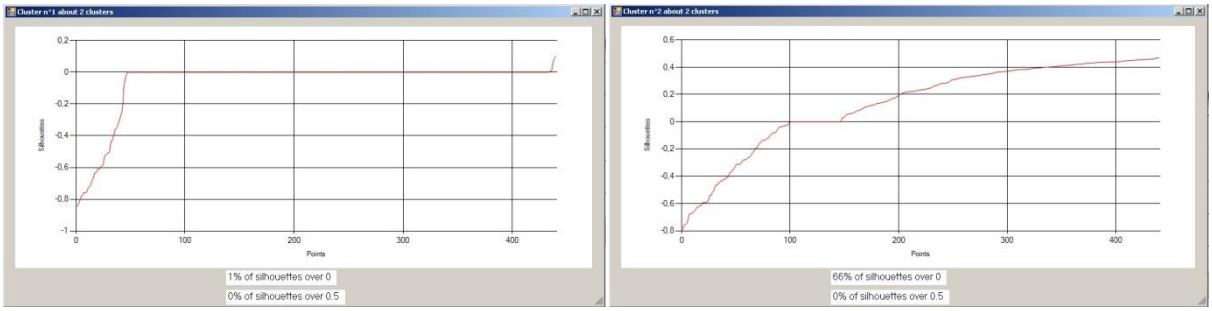
(a) (b)

(c) (d)

Figura 7.5.3.4: *I quattro grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0"sia del 7% (a), 13% (b), 17% (c) e 20%(d).*

7.5.4 Whosale customers - Dataset

100% (440 number of instance) – 2 clusters best silhouette's value

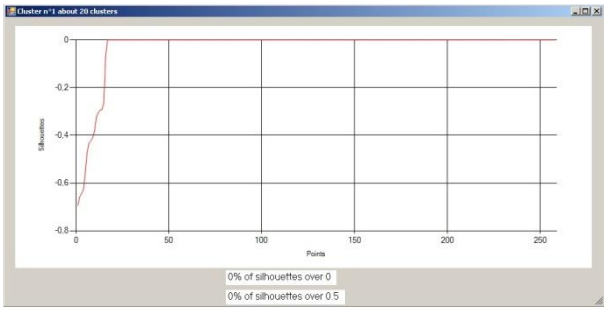


(a) (b)

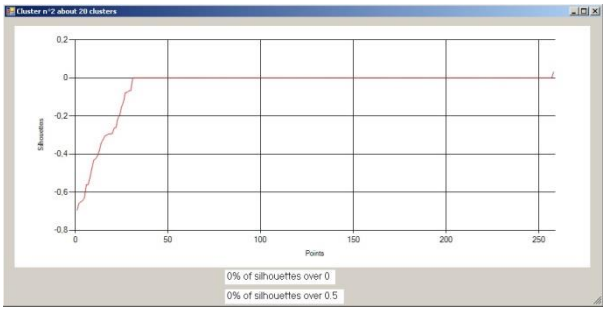
Figura 7.5.4.1: *I due grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0"sia del 1% (a) e 66% (b).*

7.5.5 User Knowledge Modeling - Dataset

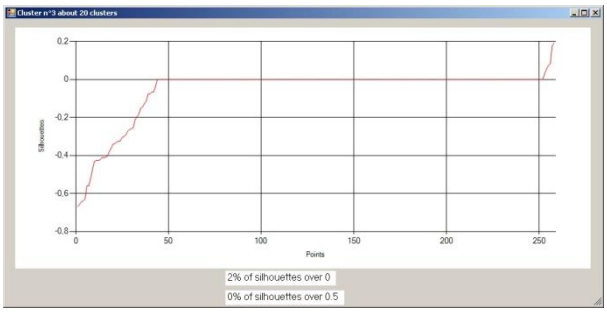
100% (258 number of instance) – 2 clusters best silhouette's value



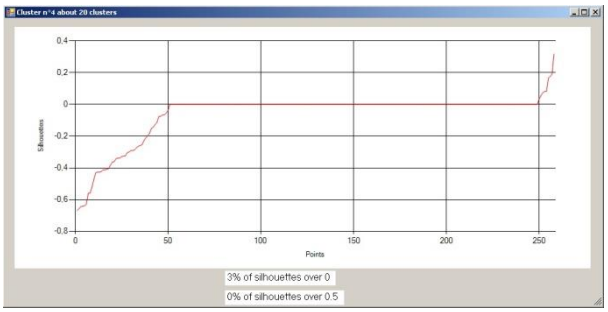
(a)



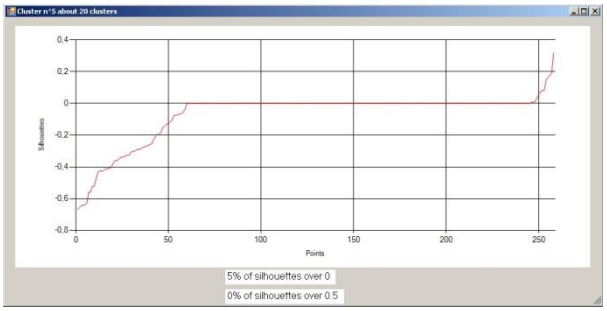
(b)



(c)

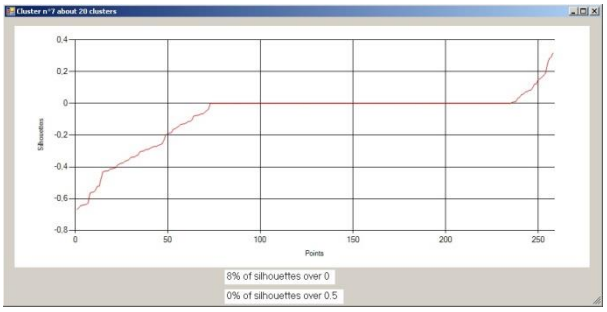


(d)

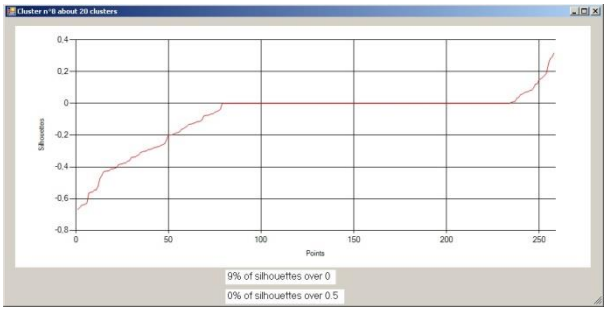


(e)

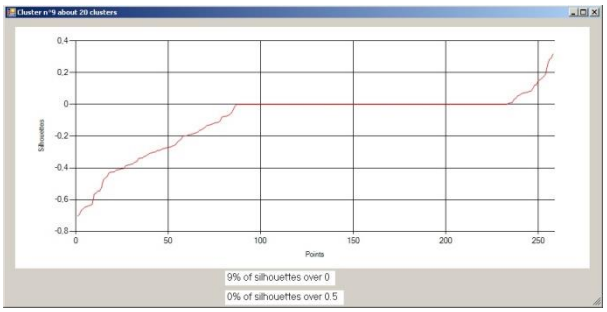
(f)



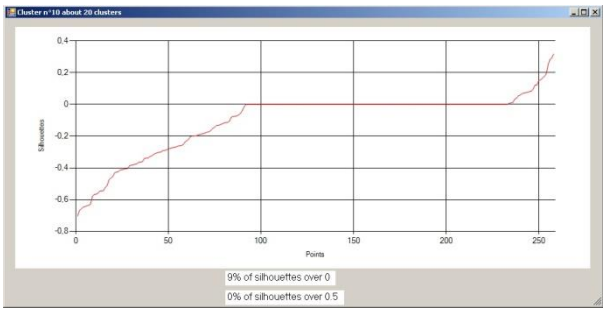
(g)



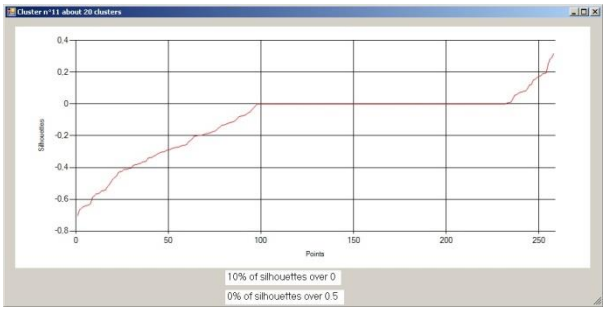
(h)


Cluster n°9 about 20 clusters
Silhouettes
Points
9% of silhouettes over 0
0% of silhouettes over 0.5

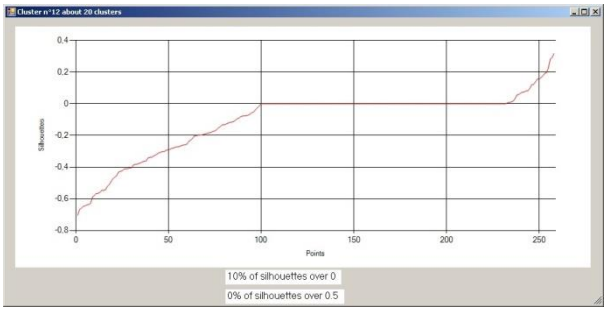
(i)


Cluster n°10 about 20 clusters
Silhouettes
Points
9% of silhouettes over 0
0% of silhouettes over 0.5

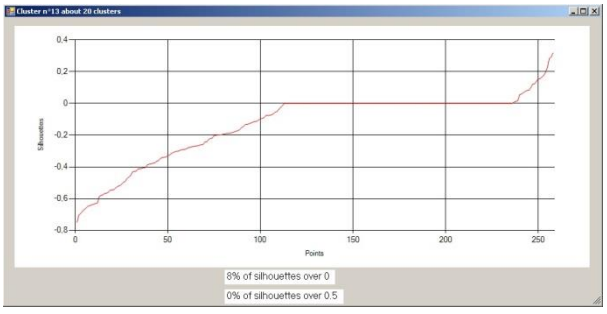
(j)


Cluster n°11 about 20 clusters
Silhouettes
Points
10% of silhouettes over 0
0% of silhouettes over 0.5

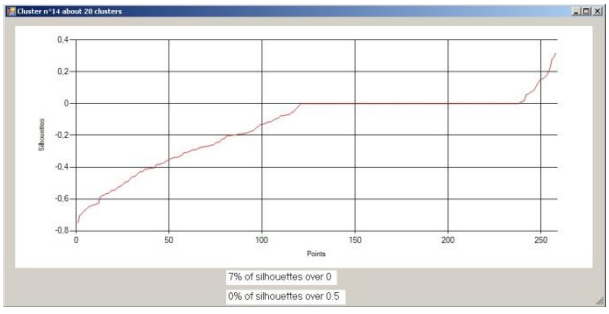
(k)


Cluster n°12 about 20 clusters
Silhouettes
Points
10% of silhouettes over 0
0% of silhouettes over 0.5

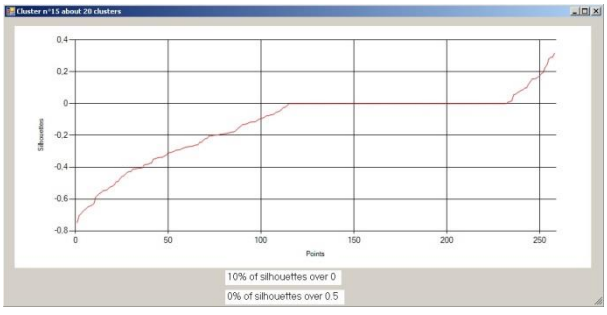
(l)


Cluster n°13 about 20 clusters
Silhouettes
Points
8% of silhouettes over 0
0% of silhouettes over 0.5

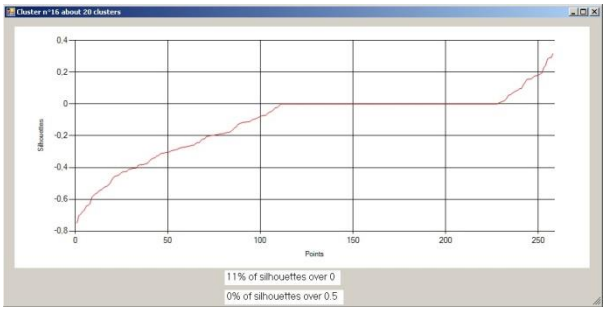
(m)


Cluster n°14 about 20 clusters
Silhouettes
Points
7% of silhouettes over 0
0% of silhouettes over 0.5

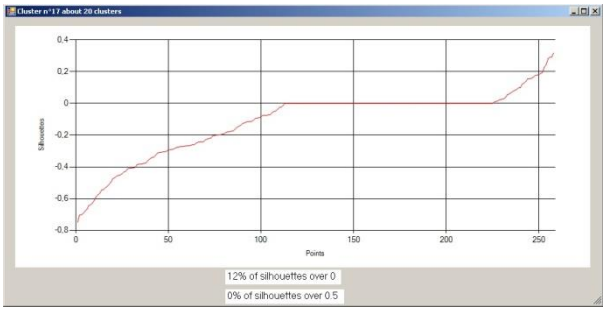
(n)


Cluster n°15 about 20 clusters
Silhouettes
Points
10% of silhouettes over 0
0% of silhouettes over 0.5

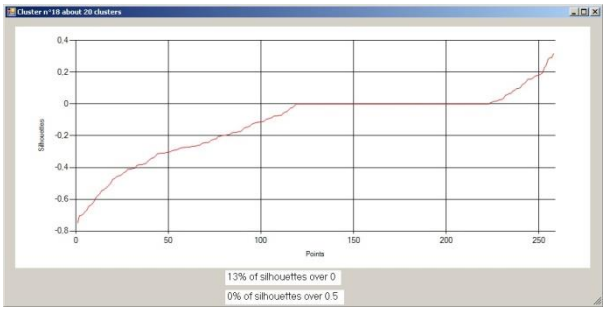
(o)


Cluster n°16 about 20 clusters
Silhouettes
Points
11% of silhouettes over 0
0% of silhouettes over 0.5

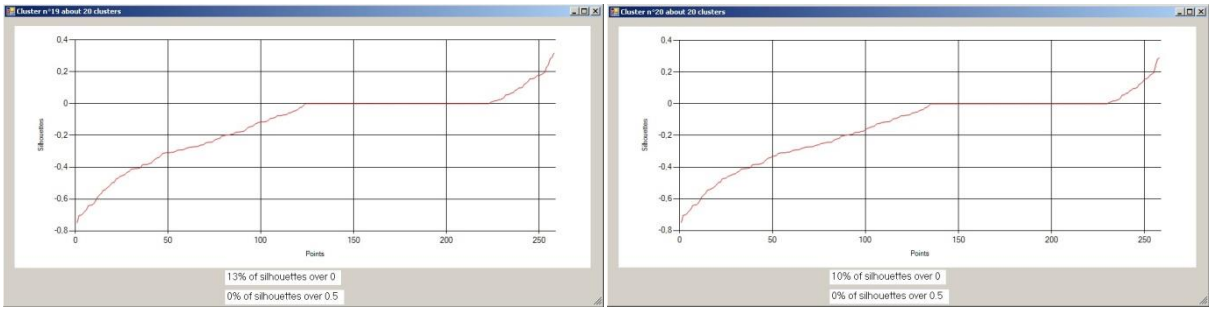
(p)


Cluster n°17 about 20 clusters
Silhouettes
Points
12% of silhouettes over 0
0% of silhouettes over 0.5

(q)


Cluster n°18 about 20 clusters
Silhouettes
Points
13% of silhouettes over 0
0% of silhouettes over 0.5

(r)



(s) (t)

Figura 7.5.5.1: *I venti grafici mostrano l'andamento della silhouette per ognuno dei clusters in esame. In ordine di visualizzazione dei grafici si nota come l'andamento di silhouette superiore alla soglia "0" sia del 13% e 10%.*

8. Conclusioni

In questa tesi è stato esaminato in dettaglio l'utilizzo delle due tecniche di clustering *K-means* e *Dbscan* allo scopo di verificare, per la prima tecnica, quale fosse il numero di cluster K *(per ogni dataset coinvolto)* che potesse offrire i migliori valori di *silhouette*; per la seconda, dopo aver ottenuto la matrice delle distanze *(implementando il k-dist)* ed aver scelto il valore K l'obiettivo era quello di raggiungere il miglior *epsilon* ottenibile.

Sono stati scelti cinque dataset, di cui tre di grandi dimensioni e due con dimensioni *medio - piccola*. Per ogni dataset, usando il ***K-means,*** è stata scelta la percentuale di *tuple* da sottoporre ad analisi e, dai dati "grezzi", si è ottenuta la *silhouette media* per "N" *clusters (con "N" da 2 a 20)*; la *silhouette* per il *cluster* selezionato *(con andamento di silhouette per ogni punto del cluster)*, la *silhouette* media per "N" *clusters*, l'andamento della *silhouette* per "N" *clusters (con N grafici rappresentanti l'andamento della silhouette per ogni punto del cluster coinvolto)* e, per ogni *cluster* 2-20, la percentuale di punti oltre la soglia "0" di *silhouette*.

Implementando il ***DbScan***, per gli stessi dataset, sono state scelte le quattro percentuali di tuple da analizzare usando il *k-dist*. Per creare ciò, è stato necessario realizzare una matrice delle distanze *N x N* ponendo come riga/colonna i punti della porzione di dataset scelta. Successivamente si è ordinata per righe tale matrice e si è scelta la colonna *K (minPoint)*. Tale colonna, allorché ordinata in modo crescente, è rappresentata in un grafico X/Y con i punti sull'asse delle ascisse e la distanza *(contenuta nelle celle della matrice)* su quella delle ordinate.

Dal grafico ottenuto occorre considerare la zona in cui avviene la curvatura di esso allo scopo di comprendere a quale **epsilon** *(leggibile sull'asse delle ascisse)* corrisponda il **minPoint(K)** prescelto.

Dai risultati conseguiti con l'uso dei due algoritmi appena citati, si può affermare che gli effetti raggiunti siano senza dubbio promettenti; nel prossimo futuro la ricerca potrà essere estesa all'uso dei dati testuali.
Infatti, l'obiettivo è quello di estendere le conoscenze sino ad ora apprese impiegandole all'uso di una forma particolare di data mining: il **text mining**. Il *data mining* permette la scoperta di nuovi punti di vista e correlazioni trovando dei *pattern* in dati che non sarebbero correlabili con le tradizionali query e le tecniche di reporting.
Tali tecniche permettono di confrontare dati provenienti da fonti eterogenee di diverso tipo ed estrarre informazioni che non sarebbero visibili all'utente, organizzare documenti e informazioni per soggetto e argomento. Il *text mining* è l'applicazione delle metodologie del *data mining* a dati poco, o per nulla, strutturati; esso, infatti, opera in un mondo meno organizzato ove i documenti hanno raramente una strutturazione e, laddove esiste, è inerente al formato del documento e non al contenuto.

Il *text mining* permette l'estrazione di metadati dai documenti e il conseguente inserimento in un DB sul quale si potranno compiere analisi di *data mining*.

Attualmente tale tecnica è impiegata nei seguenti settori di mercato:

- information technology e internet;
- aziende di telecomunicazione;
- editoria;
- pubblica amministrazione;

- aziende finanziarie ed assicuratrici;
- aziende farmaceutiche.

Ancorché in apparenza si possano riscontrare scarsi legami tra i settori sopra citati, è opportuno precisare che il ***text mining*** non si occupa solo di documenti ma di dati testuali in genere: quindi, che si tratti di brevetti, di e-mail dei propri clienti, di sondaggi, di articoli di giornale, di informazioni pubblicate su un sito web, di pratiche amministrative e/o legali, di curriculum vitae, il text mining viene in aiuto per estrarre e organizzare l'informazione.

9. Riferimenti bibliografici

[1] http://bias.csr.unibo.it/golfarelli//DataMining/MaterialeDidattico/DMISI-Clustering.pdf (26/07/2017)

[2] http://www.dsi.unive.it/~dm/Slides/2_Cluster.pdf (26/07/2017)

[3] http://www.i-dome.com/articolo/9586-Il-Text-Mining-che- cosa-è-(parte-I).html (27/07/2017)

[4] http://www.learnbymarketing.com/methods/k-means-clustering/ (20/04/2017)

[5] DbScan from https://it.wikipedia.org/wiki/Dbscan and K-means algorithm from https://it.wikipedia.org/wiki/K-means (20/04/2017)

Building Accurate 3D Spatial Networks to Enable Next Generation Intelligent Transportation Systems (Accepted and to be published in June). Proceedings of International Conference on Mobile Data Management (IEEE MDM), June 3-6 2013, Milan, Italy. Authors: Chenjuan Guo, Yu Ma, Bin Yang, Christian S. Jensen, Manohar Kaul: EcoMark: evaluating models of vehicular environmental impact. SIGSPATIAL/GIS 2012: 269-278

H. T. Kahraman, Sagiroglu, S., Colak, I., Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, Knowledge Based Systems, vol. 37, (pp. 283-295, 2013). 1)H. T. Kahraman, Sagiroglu, S., Colak, I., Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, Knowledge Based Systems, vol. 37, pp. 283-295, 2013.
2) Kahraman, H. T. (2009). Designing and Application of Web-Based Adaptive Intelligent Education System. Gazi University Ph. D. Thesis, Turkey, 1-156.

Gunduz, G. & Fokoue, E. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Zhiyuan Chen and Johannes Gehrke and Flip Korn. Query Optimization In Compressed Database Systems. SIGMOD Conference. 2001. David R. Musicant. DATA MINING VIA MATHEMATICAL PROGRAMMING AND MACHINE LEARNING. Doctor of Philosophy (Computer Sciences) UNIVERSITY.Chris Giannella and Bassem Sayrafi. An Information Theoretic Histogram for Single Dimensional Selectivity Estimation. Department of Computer Science, Indiana University Bloomington. David R. Musicant and Alexander Feinberg. Active Set Support Vector Regression. Meek, Thiesson, and Heckerman (2001), "The Learning Curve Method Applied to Clustering", to appear in The Journal of Machine Learning Research.

Efficient Electricity Utilization By IHBMO.

Cardoso, Margarida G.M.S. (2013). Logical discriminant models " Chapter 8 in Quantitative Modeling in Marketing and Management Edited by Luiz Moutinho and Kun-Huang Huarng. World Scientific. p. 223-253. ISBN 978-9814407717 Jean-Patrick Baudry, Margarida Cardoso, Gilles Celeux, Maria Josè.

Amorim, Ana Sousa Ferreira (2012). Enhancing the selection of a model-based clustering with external qualitative variables. RESEARCH October 2012, Project-Team SELECT.,

Projet select, Universitàˆ© Paris-Sud 11

The data set is originated from a larger database referred on: Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon

10. Ringraziamenti

Desidero ricordare tutti coloro che mi hanno aiutato nella stesura con suggerimenti, critiche e osservazioni: a loro va la mia gratitudine.

Ringrazio anzitutto la professoressa Tania Cerquitelli per il supporto e l'aiuto nella realizzazione di questa tesi. Un ringraziamento particolare va ai colleghi e agli amici che mi hanno incoraggiato o che hanno speso parte del proprio tempo a leggere e discutere con me le bozze del lavoro.

Vorrei infine esprimere gratitudine alle persone a me più care: i miei amici, i miei genitori Chiara e Giuseppe, mio fratello Luca.

Questo traguardo desidero dedicarlo al ricordo dei miei nonni, di Piera e di mio zio Aldo.