POLITECNICO DI TORINO

Corso di Laurea in Ingegneria informatica

Tesi di Laurea Magistrale

Caratterizzazione e visualizzazione della conoscenza estratta da attestati di prestazione energetica rilasciati dalla Regione Piemonte

Relatori:
Prof.ssa Tania Cerquitelli
Prof.ssa Elena Maria Baralis

Candidato:
Maria Giovanna Cassese

Anno accademico 2017/2018

Ringraziamenti

Innanzitutto vorrei ringraziare la Prof.ssa Tania Cerquitelli, per l'aiuto fornitomi, per la disponibilità e per la cortesia dimostratemi durante tutto il periodo di stesura.

Ringrazio la Prof.ssa Elena Maria Baralis per aver favorito la realizzazione di questo lavoro.

La presente tesi è stata svolta in collaborazione con il Dipartimento Energia del Politecnico di Torino. Ringrazio pertanto l'Ing. Alfonso Capozzoli per avermi fornito nozioni fondamentali per la stesura della tesi e per gli utili consigli.

Ringrazio la Dott.ssa Evelina Di Corso e il Dott. Stefano Proto per la pazienza, per la disponibilità e per il tempo che mi hanno dedicato in questo periodo.

Un ringraziamento speciale va ai miei genitori per la grande capacità di ascoltarmi e supportarmi in ogni momento ed a mio fratello per l'insostituibile affetto e sostegno.

Inoltre, vorrei ringraziare Leo per non avermi fatto sentire mai sola, per aver creduto in me e per essere riuscito sempre a comprendermi.

Infine, vorrei ringraziare tutti i miei amici, i vecchi, i nuovi e quelli che sono stati come una famiglia per me a Torino.

Indice

Introduzione 1

1 La certificazione energetica 3
1.1 La certificazione energetica nella regione Piemonte 4
1.2 Descrizione degli attributi delle certificazioni energetiche utilizzate . . 5

2 L'estrazione della conoscenza 12
2.1 Selezione, *preprocessing* e trasformazione 13
2.1.1 Metodologie per individuare gli *outlier* 14
2.2 *Data mining* 16
2.2.1 Algoritmi di *clustering* 17
2.2.2 Regole di associazione 21
2.3 Interpretazione e validazione della conoscenza estratta 25

3 Architettura sviluppata 26
3.1 Raccolta e integrazione dei dati 27
3.1.1 *Pulizia dei dati* 28
3.2 Preparazione dei dati 33
3.2.1 Selezione dei dati 33
3.2.2 Algoritmi basati sulla densità: DBSCAN 33
3.2.3 Normalizzazione dei dati 35
3.3 Analisi dei *cluster* 36
3.3.1 Algoritmo K-means 36
3.3.2 Regole di associazione 43
3.4 Validazione e visualizzazione della conoscenza 45
3.4.1 Realizzazione delle mappe 47

4 Risultati sperimentali 49
4.1 Raccolta e integrazione dei dati 49
4.1.1 Pulizia dei dati: algoritmo di correzione degli attributi per la geolocalizzazione 49

4.1.2 Pulizia dei dati: *scaling* 52
4.1.3 Pulizia dei dati: eliminazione degli outlier 52
4.2 Analisi dei *cluster* 55
4.2.1 Algoritmo K-means 57
4.2.2 Regole di associazione 70
4.3 Validazione e visualizzazione della conoscenza 71
4.3.1 Visualizzazione dei dati 71

5 Conclusioni e sviluppi futuri 73

Riferimenti bibliografici 75

Elenco delle figure

1.1 Scala classificazioni degli edifici tramite confronto con $EP_{gl,nren}$. ©Supplemento Ordinario n.39 alla Gazzetta Ufficiale Serie generale - n. 162 (15 luglio 2015) 11

2.1 Il processo di *Knowledge Discovery in Databases*. ©Fayyad, Piatetsky-Shapiro, Smyth 13
2.2 Descrizione di un esempio di boxplot. 16
2.3 Eempio di raggruppamenti in cluster. © Tan,Steinbach,Kumar 17
2.4 Clustering gerarchico rappresentato attraverso un dendrogramma. © Tan,Steinbach,Kumar 18
2.5 Esempio di clustering partizionale.© Tan,Steinbach,Kumar 19
2.6 Esempio di clustering. ©Dulli, Furini, Peron 20
2.7 Disuguaglianza triangolare. ©Dulli, Furini, Peron 21
2.8 Esempio di dataset transazionale 23
2.9 Descrizione passi dell' algoritmo Apriori.©Rakesh Agrawal, Ramakrishnan Srikant 25

3.1 Architettura 26
3.2 Esempio conversione della stringa "ac" in "ab" 29
3.3 Descrizione di pattern core, border o noise nel DBSCAN.© Tan,Steinbach,Kumar 34
3.4 Esempio fallimentare di clustering con DBSCAN.© Tan,Steinbach,Kumar 35
3.5 Esempio1. Importanza della scelta dei centroidi.© Tan,Steinbach,Kumar 37
3.6 Esempio2. Importanza della scelta dei centroidi.© Tan,Steinbach,Kumar 37
3.7 Due diversi clustering ottenuti con il K-means.© Tan,Steinbach,Kumar 38
3.8 Cluster di dimensioni diverse.© Tan,Steinbach,Kumar 41
3.9 Cluster di densità diverse. © Tan,Steinbach,Kumar 41
3.10 Cluster di forme diverse. © Tan,Steinbach,Kumar 41
3.11 Possibile soluzione per cluster di dimensioni diverse.© Tan,Steinbach,Kumar 42
3.12 Possibile soluzione per cluster di densità diverse. © Tan,Steinbach,Kumar 42
3.13 Possibile soluzione per cluster di forme diverse. © Tan,Steinbach,Kumar 42
3.14 Esempio di *elbow graph* in cui il valore ideale è K = 3 43
3.15 Esempio di mappa coropletica. ©Olivo, Hamidi, Ramamurthy 46

3.16 Esempio di mappa *scatter*. 46
3.17 Esempio di mappa coropletica con marker-cluster. 48
3.18 Esempio di mappa coropletica con marker-cluster con un livello di zoom maggiore. 48
3.19 Esempio di mappa coropletica con marker-cluster con il massimo livello di zoom. 48

4.1 Distribuzione del numero di certificati per CAP più significativi nella città di Torino prima della pulizia. 50
4.2 Distribuzione del numero di certificati per CAP nella città di Torino dopo la pulizia. 50
4.3 Esempio di indirizzi non risolti con Levenshtein, ma risolti con Geocoding. 51
4.4 Percentuale di valori ritenuti *outliers* per ogni metodo utilizzato. . . . 53
4.5 Range di validità degli attributi estratti con *gESD, MAD e Boxplot*. . 53
4.6 Distribuzione dei valori delle trasmittanze nel dataset. 54
4.7 Distribuzione dei valori dei rendimenti diversi da 0 nel dataset. 55
4.8 Distribuzione dei valori del rendimento di generazione che hanno causato l'eliminazione del relativo certificato. 56
4.9 Elbow graph per valori di K da 2 a 30. 57
4.10 Grafico dei centroidi degli attributi utilizzati per il clustering con K=4. 58
4.11 Grafico dei centroidi degli attributi utilizzati per il clustering con K=7. 59
4.12 Grafico dei centroidi degli attributi utilizzati per il clustering con K=9. 59
4.13 Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=4. . . 60
4.14 Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=7. . . 61
4.15 Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=9. . . 61
4.16 Distribuzione dei certificati per i 4 cluster. 62
4.17 Grafico dei centroidi degli attributi utilizzati per il clustering con K=4 (ottenuto per riaggregazione). 63
4.18 Distribuzione dei valori del rendimento medio globale stagionale invernale per i 4 cluster. 63
4.19 Distribuzione dei valori del fattore forma per i 4 cluster. 64
4.20 Distribuzione dei valori della trasmittanza media delle superfici opache per i 9 cluster. 64
4.21 Distribuzione dei valori della trasmittanza media delle superfici trasparenti per i 9 cluster. 65
4.22 Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=4 dopo la riaggregazione. 66
4.23 Distribuzione degli anni di costruzione per i 4 cluster dopo la riaggregazione. 67

4.24 Distribuzione delle classe energetiche raggruppate in base alle performance per i 4 cluster dopo l'aggregazione. 68
4.25 Distribuzione degli edifici ristrutturati per i 4 cluster dopo la riaggregazione. 68
4.26 Distribuzione degli anni di costruzione dei soli edifici ristrutturati per i 4 cluster dopo la riaggregazione. 69
4.27 Distribuzione dei valori dell' EP_H dei soli edifici ristrutturati per i 4 cluster dopo la riaggregazione. 69
4.28 Regole estratte dal cluster 2. 70
4.29 Regole estratte dal cluster 3. 71
4.30 Esempio di dettaglio della circoscrizione 8 per l'attributo ETAH. . . . 72

Introduzione

La *certificazione energetica* degli edifici è un attestato che fornisce delle informazioni circa le caratteristiche di isolamento termico e della tipologia degli impianti presenti, quindi del consumo energetico, attraverso un sistema di classificazione che consente di dedurre la qualità energetica di un immobile.
In Europa, circa il 40% dei consumi energetici finali globali è destinato agli edifici residenziali; pertanto, l'introduzione di tali certificazioni nasce dalla necessità di promuovere l'efficienza energetica, come soluzione ai grandi problemi che l'Europa deve affrontare in termini di importazioni di energia, scarsità di risorse energetiche e necessità di limitare i cambiamenti climatici.
A tal proposito, l'Unione europea, attraverso l'emanazione di direttive e norme, ha fornito agli Stati membri delle linee guida per la redazione di questi attestati di prestazione energetica, allo scopo di raggiungere gli obiettivi energetici prefissati. Tuttavia, è stata lasciata libertà a livello nazionale e regionale nella scelta della metodologia per il calcolo della prestazione energetica degli edifici.
In Italia, attualmente, il certificato energetico prende il nome di *APE (Attestazione di Prestazione Energetica)*, presenta un formato standard ed è in grado di fornire informazioni sul fabbisogno energetico, sull'efficienza e sulle prestazioni di un edificio.
Con l'introduzione dell' *APE*, è inoltre previsto un sistema informativo comune per tutto il territorio nazionale (*SIAPE*) che gestisce e archivia tali certificazioni.
Ciò fa sì che la disponibilità di dati sulle caratteristiche fisiche ed energetiche degli edifici e degli impianti termici sono aumentate in maniera consistente.

L'analisi effettuata in questa tesi ha lo scopo di caratterizzare l'efficienza energetica degli edifici residenziali attraverso tecniche di *machine learning*. L'estrazione delle caratteristiche termo-fisiche che accomunano gli edifici è stata effettuata attraverso tecniche di *data mining* non supervisionate, come gli algoritmi di *clustering*; in seguito, si è proceduto con la visualizzazione dei dati raccolti e della conoscenza estratta tramite le tecniche precedentemente citate.
L'obiettivo della presente tesi è stato quello di progettare e sviluppare un *framework* in grado di supportare l'analista durante l'analisi di grandi volumi di dati relativi a

diversi edifici localizzati nella regione Piemonte. Il *framework*, sviluppato in *Python*, utilizza una tecnica a due livelli: *(i)* attraverso tecniche non supervisionate partiziona il *dataset* in gruppi omogenei di edifici con caratteristiche termo-fisiche simili, *(ii)* esplora diverse tecniche di visualizzazione (e.g., tabelle, grafici, mappe geolocalizzate) per analizzare a diversi livelli di dettaglio la conoscenza estratta. Tale conoscenza è utile per supportare l'analista durante il *decison making* ma consente anche a un non esperto di dominio di comprendere problematiche e dedurre possibili soluzioni in termini di efficienza energetica edilizia.

La tesi è strutturata in 5 capitoli:
Il **Capitolo 1** descrive l'introduzione della certificazione energetica in ambito nazionale, focalizzandosi sulle normative della Regione Piemonte, a seguito dell'introduzione dell' *APE*, e descrivendo gli attributi fondamentali di tali attestati.
Nel **Capitolo 2** vengono descritti i concetti teorici legati all'estrazione della conoscenza, ponendo enfasi sulle fasi di *preprocessing* e *data mining*.
Nel **Capitolo 3** è descritta l'architettura sviluppata che ha permesso la caratterizzazione degli edifici della regione Piemonte, in particolare della città di Torino, tramite tecniche di *datamining* e la visualizzazione di tale conoscenza attraverso grafici e mappe.
Nel **Capitolo 4** vengono presentati i risultati sperimentali ottenuti dall'applicazione del *framework* sugli open-data disponibili.
Nel **Capitolo 5** vengono mostrate le conclusioni e gli sviluppi futuri.

Capitolo 1

La certificazione energetica

La certificazione energetica è diventata un attestato di rilevante importanza come risposta alla necessità europea di far fronte ai problemi economici ed ambientali odierni.
L'attenzione verso la tutela dell'ambiente era emersa già con il *protocollo di Kyoto* che aveva come obiettivo la riduzione delle emissioni di gas serra per limitare il surriscaldamento globale. In linea con tale trattato, l'Europa ha voluto riconoscere l'importanza dell'efficienza energetica come strumento per limitare le importazioni energetiche ed i cambiamenti climatici; in particolare, è emerso che il 40% dell'energia globale consumata è relativo all'edilizia, settore in forte espansione e sul quale, per tanto, risulta necessario un intervento consistente [4].
Ci sono state differenti indicazioni europee. La più recente è la direttiva *2010/31/UE*, pubblicata il 19 Maggio 2010 sulla *Gazzetta ufficiale dell'Unione europea*, che fornisce agli stati aderenti delle linee guida da seguire per il potenziamento delle prestazioni energetiche degli edifici; tuttavia, la metodologia di calcolo potrebbe essere differenziata a livello nazionale e regionale per tenere conto di fattori peculiari (ad esempio, climatici) di ogni zona.
In Italia, la normativa in merito alle certificazioni energetiche degli edifici è entrata in vigore nel 2005 con il decreto legislativo 192/2005, pur subendo innumerevoli modifiche negli anni. Tra queste ricordiamo il decreto 63/2013, con cui viene introdotto l'APE - Attestato di Prestazione Energetica in sostituzione dell' ACE - Attestato di Certificazione Energetica, ed i tre decreti ministeriali del 26 Giugno 2015, con cui viene completato il quadro normativo relativamente all'efficienza energetica degli edifici. In particolare, attraverso i tre suddetti decreti vengono dettagliati rispettivamente i requisiti minimi degli edifici e le metodologie di calcolo delle prestazioni energetiche, le linee guida per il nuovo APE e le modalità di compilazione della

relazione tecnica[1].
Ad ogni modo, spetta alle Regioni deliberare in merito a tali certificazioni, pur rispettando quanto stabilito nelle direttive nazionali.
In sintesi, l' APE è in grado di esprimere la prestazione energetica globale di un edificio, in termini di energia primaria totale e di energia primaria non rinnovabile, tenendo conto di tutti i servizi energetici (climatizzazione invernale e estiva, acqua calda sanitaria, ventilazione, illuminazione...). Quando parliamo di energia primaria intendiamo l'energia che non ha subito alcun tipo di trasformazione e che è dunque presente in natura; inoltre, è possibile distinguere tale energia in rinnovabile o non rinnovabile a seconda se la fonte sia rispettivamente rinnovabile (e.g.,energia eolica, solare...) oppure esauribile (e.g.,combustibili). La classe energetica viene definita sulla base di un confronto tra l'indice di prestazione energetica globale non rinnovabile $EP_{gl,nren}$ dell'immobile da certificare e quello relativo all'edificio di riferimento, attribuendo un indicatore alfanumerico in cui la lettera **A** rappresenta la classe con il migliore indice di prestazione mentre la lettera **G** esprime la classe con i maggiori consumi energetici[1].
Questo attestato, valido per 10 anni, viene prodotto da un soggetto accreditato e caricato in un sistema informativo nazionale noto come *SIAPE*.

1.1 La certificazione energetica nella regione Piemonte

Come detto precedentemente, la normativa che regola le attestazioni di prestazione energetica viene emanata dalla regione; perciò, poiché le certificazioni energetiche presenti nel nostro *dataset* riguardano la regione Piemonte, il seguente lavoro farà riferimento a regolamenti entrati in vigore nella suddetta area geografica a partire dal 2009, anno in cui viene approvata la delibera con le disposizioni in merito al rendimento energetico nell'edilizia[2].
Tale certificazione dovrà essere rilasciata obbligatoriamente per le nuove costruzioni, per le ristrutturazioni a fine lavori, per le compravendite o per le locazioni quando viene stipulato l'atto.
Anche la Regione Piemonte, nell' ottica di raggiungere gli obiettivi energetici europei

[1]Informazioni reperite da http://biblus.acca.it/

[2]Piemonte, Deliberazione della Giunta Regionale 21 settembre 2015, n. 14-2119

"20-20-20"[3] entro il 2020, si è impegnata ad incentivare un incremento dell'efficienza energetica, una riduzione dei consumi, un maggiore uso di fonti rinnovabili ed a sensibilizzare i cittadini sul tema dell'energia e del conseguente inquinamento; lo scopo ultimo è quello di ridurre di circa il 20% i consumi energetici e le emissioni di CO_2 e di aumentare del 20% l'uso delle fonti rinnovabili per produrre energia.
Nel 2015 anche il Piemonte si è uniformato alle direttive europee sulle prestazioni energetiche; per esempio il sistema informativo SICEE (Sistema Informativo Certificazione Energetica) è stato sostituito con il SIPEE (Sistema Informativo per la Prestazione Energetica degli Edifici), è stato stabilito che i soggetti accreditati a rilasciare tali certificazioni devono essere iscritti a un elenco regionale e sono state definite delle procedure di controllo della qualità delle attestazioni. Riguardo a quest'ultimo aggiornamento, è stato previsto che l' ARPA (Agenzia Regionale per la Protezione Ambientale) effettui controlli a campione, soprattutto per le classi energetiche migliori, allo scopo di verificare la validità dei certificati e il rispetto dei vincoli imposti[2].

1.2 Descrizione degli attributi delle certificazioni energetiche utilizzate

In questo paragrafo vengono descritti gli attributi significativi presenti nelle certificazioni APE e richiesti nella compilazione della procedura telematica attraverso il SIPEE.
I dati presenti sono di varia natura, dai dati catastali a quelli tecnici generali, dagli indici di fabbisogno alle informazioni sulle fonti rinnovabili utilizzate; ovviamente nelle analisi successive verranno utilizzate solo una parte di queste informazioni.
In particolare, ci siamo concentrati su edifici di tipo residenziale con carattere continuativo della città di Torino: l' attributo che ci consente di applicare questo filtro è la *'Destinazione d'uso'* con valore *E1 (1)*, come evidente nella tabella 1.1. Inoltre, per identificare univocamente un'unità immobiliare sono stati usati tre attributi di carattere catastale:

- **Foglio** identificativo numerico del foglio catastale con cui è possibile individuare l'area comunale in cui si trova l'edificio.
- **Particella** identificativo solitamente numerico della particella catastale con cui si rappresenta una porzione di terreno o un fabbricato.

[3]Il piano europeo "20-20-20" consiste nel ridurre entro il 2020 il 20% delle emissioni di gas serra, nell' aumento di circa il 20% dell'energia rinnovabile prodotta e nel ridurre i consumi di circa il 20%, come definito nella *Direttiva 2009/29/CE, pubblicata il 5 Giugno 2009 sulla Gazzetta ufficiale dell'Unione europea.*

Destinazione d'uso	Descrizione
E.1	Edifici adibiti a residenza e assimilabili
E.1 (1)	abitazioni adibite a residenza con carattere continuativo, quali abitazioni civili e rurali, collegi,conventi, case di pena, caserme
E.1 (2)	abitazioni adibite a residenza con occupazione saltuaria, quali case per vacanze, fine settimana e simili
E.1 (3)	edifici adibiti ad albergo, pensione ed attività similari
E.2	Edifici adibiti a uffici e assimilabili pubblici o privati, indipendenti o contigui a costruzioni adibite anche ad attività industriali o artigianali, purché siano da tali costruzioni scorporabili agli effetti dell'isolamento termico
E.3	Edifici adibiti a ospedali, cliniche o case di cura e assimilabili ivi compresi quelli adibiti a ricovero o cura di minori o anziani nonché le strutture protette per l'assistenza ed il recupero dei tossicodipendenti e di altri soggetti affidati a servizi sociali pubblici
E.4	Edifici adibiti ad attività ricreative o di culto e assimilabili
E.4 (1)	quali cinema e teatri, sale di riunioni per congressi
E.4 (2)	quali mostre, musei e biblioteche, luoghi di culto
E.4 (3)	quali bar, ristoranti, sale da ballo
E.5	Edifici adibiti ad attività commerciali e assimilabili quali negozi, magazzini di vendita all'ingrosso o al minuto, supermercati, esposizioni
E.6	Edifici adibiti ad attività sportive
E.6 (1)	piscine, saune e assimilabili
E.6 (2)	palestre e assimilabili
E.6 (3)	servizi di supporto alle attività sportive
E.7	Edifici adibiti ad attività scolastiche a tutti i livelli e assimilabili
E.8	Edifici adibiti ad attività industriali ed artigianali e assimilabili

Tabella 1.1: Elenco delle destinazioni d'uso come da *D.P.R. 26 agosto 1993, n. 412, pubblicato nella Gazzetta Ufficiale il 14 ottobre 1993, n. 242, S.O.* .

- **Subalterno** consente di identificare l'unità immobiliare presente su una particella.

Altri attributi di carattere più tecnico utilizzati nelle analisi sono:

- Fattore forma [m^{-1}]
- Trasmittanza media delle superfici opache [W/m^2K]

- Trasmittanza media delle superfici trasparenti [W/m^{2}K]
- Rendimento di generazione
- Rendimento di distribuzione
- Rendimento di regolazione
- Rendimento di emissione
- Indice di prestazione energetica globale [kWh/m^{2}anno]
- Indice di prestazione energetica invernale [kWh/m^{2}anno]
- Indice di prestazione energetica invernale dell'involucro [kWh/m^{2}anno]
- Area solare equivalente [m^2]
- Trasmittanza termica periodica [W/m^{2}K]
- Superficie utile [m^2]
- Classe energetica

Di seguito per ogni attributo è stata riportata una descrizione più approfondita.

- **Fattore forma**[m^{-1}]: misura il rapporto tra S, la superficie disperdente[4] dell'edificio considerato, e V, il relativo volume lordo riscaldato. L'importanza di questo attributo dipende dal fatto che il suo valore influenza fortemente le dispersioni termiche poiché tiene conto della superficie attraversata nello scambio termico tra l'interno e l'esterno dell'edificio. Il fattore forma non dipende da quanto l'edifico è isolato, ma solo dalla sua geometria. Questo rapporto viene anche definito *di compattezza* perché il fattore forma si riduce all'aumentare della compattezza dell'edificio; infatti, più un edificio è compatto, più è efficiente dal punto di vista energetico. Ovviamente, l'efficienza di cui parliamo è relativa alla climatizzazione invernale poiché nasce dalla riduzione della dispersione termica. Inoltre, anche la forma geometrica degli edifici influenza le perdite energetiche e, pertanto, vengono predilette le forme semplici, come quelle cubiche, piuttosto che quelle con molte sporgenze[5].

[4]Seconodo quanto definito nell' art.2 del D.M.26/06/2015 , la superficie disperdente è *"la superficie che delimita il volume climatizzato V rispetto all' esterno, al terreno, ad ambienti a diversa temperatura o ambienti non dotati di impianto di climatizzazione"*

[5]Informazioni reperite da https://www.e-genius.at/

Prima di descrivere gli attributi trasmittanza opaca e trasparente, occorre chiarire il concetto di trasmittanza termica. La trasmittanza termica assume un ruolo fondamentale nel calcolo della classe energetica di un edificio perché consente di quantificare il fabbisogno energetico dello stesso. Questa variabile è detta anche *trasmissione termica* perché rappresenta il passaggio di calore da un ambiente caldo a uno freddo, qualora la relativa superficie sia sottoposta a un gradiente termico[6]. Se una casa ha elementi dell'involucro edilizio con un alta trasmittanza termica, maggiore sarà il passaggio di calore e più la casa sarà calda in estate e fredda in inverno. La tramittanza termica si esprime attraverso il coefficiente U ed è espressa in W/m^2K, cioè i Watt di energia che si disperderebbero attraverso un metro quadro di superficie se ci fosse una differenza di temperatura di un grado Kelvin. Il calcolo della trasmittanza termica viene fatto sia per elementi opachi che trasparenti. Per i primi si tiene conto del materiale utilizzato e delle modalità di costruzione, per i secondi della tipologia di vetro degli infissi o della tipologia di telaio[7].

- **Trasmittanza opaca**[W/m^{2}K]: rappresenta la trasmittanza termica media ponderata relativa agli elementi opachi confinanti con l'ambiente esterno

- **Trasmittanza trasparente**[W/m^{2}K]: rappresenta la trasmittanza termica media ponderata relativa agli elementi trasparenti confinanti con l'ambiente esterno

Poiché i sistemi reali di riscaldamento presentano alcune perdite di calore, è necessario che l'energia primaria fornita al corpo scaldante sia maggiore di quella che esso emana e che quest'ultima sia maggiore di quella richiesta dall'ambiente. Per tale ragione, nel considerare le prestazioni energetiche di un edificio occorre considerare il contributo dei rendimenti dei quattro sottosistemi dell'impianto, ovvero il sistema di produzione di energia termica, quello di distribuzione del calore, quello di trasmissione dello stesso e infine quello di regolazione. I rendimenti considerati sono quelli medi stagionali e sono gli stessi che contribuiscono al calcolo del rendimento medio stagionale per la definizione della classe energetica.
Descriviamo ora nel dettaglio i rendimenti:

- **Rendimento di generazione**: esprime il rapporto tra il calore che viene prodotto dal generatore e l'energia fornita come energia elettrica o combustibile; questo valore tiene cioè conto del fatto che il passaggio di calore dal generatore al fluido termovettore è caratterizzato da perdite[2].

[6] www.ideegreen.it

[7] http://www.la-certificazione-energetica.net

- **Rendimento di distribuzione**: esprime il rapporto tra l'energia termica fornita dal corpo scaldante e dalle tubazioni che sono all'interno dell'involucro riscaldato (cioè quella parte di calore disperso ma recuperato) e l'energia termica immessa nella rete di distribuzione; questo valore tiene cioè conto delle dispersioni verso l'esterno dei tubi in cui scorre il fluido termovettore[2].

- **Rendimento di regolazione**: esprime il rapporto tra il calore che viene richiesto con regolazione teorica per riscaldare gli ambienti e quello con un sistema di regolazione reale[2]. Il primo tipo è quello che è in grado di ridurre l'emissione calorica appena rileva un apporto di calore diverso da quello prodotto dall'impianto di riscaldamento; il secondo, invece, varia l'emissione dopo che c'è stato un aumento della temperatura dell'ambiente[2]. Questo valore tiene conto di quanto è costante la temperatura dell'ambiente considerato.

- **Rendimento di emissione**: esprime il rapporto tra il calore che viene richiesto con un sistema di emissione teorico per riscaldare gli ambienti e quello con un sistema reale[2]. Il primo è in grado di fornire all'ambiente una temperatura uniforme in tutti i suoi punti, mentre nel secondo sistema ciò non avviene principalmente per i moti convettivi dell'aria e per la presenza di gradiente termico nel locale[2].

Al fine di stimare le prestazioni energetiche dell'edificio considerato, la certificazione APE fornisce differenti indicatori così da dare sia una valutazione generale dell'immobile sia una descrizione più dettagliata dei singoli aspetti. Ogni indicatore è in grado di esprimere la qualità energetica dell'aspetto considerato ed il suo valore è determinato da un confronto tra l'indice calcolato sulla base dei dati reali dell'edificio in questione e l'indice relativo al corrispondente edificio di riferimento. Quando parliamo di edificio di riferimento intendiamo un edificio che presenta le stesse caratteristiche in termini di geometria, destinazione d'uso e ubicazione dell'immobile da valutare, ma caratteristiche termiche ed energetiche degli impianti standard secondo quanto definito nell' art.2 del D.M.26/06/2015 requisiti minimi.
Di seguito verranno descritti questi indici ed alcuni degli attributi necessari per il calcolo degli stessi.

- $\mathbf{EP_{gl,nren}}$[kWh/m^{2}anno]: esprime l'indice di prestazione energetica globale non rinnovabile, ovvero quanta energia è necessaria affinché l'immobile abbia una condizione climatica confortevole[8]. Questo indice risulta fondamentale ai fini

[8] 15-7-2015, *Supplemento ordinario n. 39 alla GAZZETTA UFFICIALE, Serei generale - n. 162, Allegato 1, Articolo 3, "Linee guida nazionali per l'attestazione della prestazione energetica degli edifici"*

della classificazione poiché tiene conto del fabbisogno di energia per tutti i servizi energetici forniti, infatti deriva dalla somma dei seguenti attributi:

  - $EP_{H,nren}$: fabbisogno di energia primaria per la climatizzazione invernale
  - $EP_{C,nren}$: fabbisogno di energia primaria per la climatizzazione estiva
  - $EP_{W,nren}$: fabbisogno di energia primaria per l'acqua calda sanitaria
  - $EP_{V,nren}$: fabbisogno di energia primaria per la ventilazione
  - $EP_{L,nren}$: fabbisogno di energia primaria per l'illuminazione artificiale (se non residenziale)

- **Superficie utile**: è l'unione delle superfici climatizzate dell'edificio considerato, cioè quella riscaldata e quella raffrescata, ed è fondamentale per il calcolo di qualunque indice di prestazione[15].

- $\mathbf{EP_{H,nd}}$[kWh/m^{2}anno]: esprime l'indice di prestazione energetica invernale dell'involucro ed è calcolato come il rapporto tra il fabbisogno annuo di energia termica dell'edifico ($Q_{H,nd}$) e la superficie utile[8].

- $\mathbf{EP_H}$[kWh/m^{2}anno]: esprime l'indice di prestazione energetica in merito alla climatizzazione invernale ed è calcolato come il rapporto tra il valore di $EP_{H,nd}$ e il rendimento medio stagionale riferito all'impianto di riscaldamento[8].

- $\mathbf{A_{sol,est}/A_{sup,utile}}$: indica l'area solare equivalente estiva (cioè riferita alle componenti vetrate) per unità di superficie utile ed è necessaria per determinare l'indice di prestazione energetica estiva dell'involucro[15].

- $\mathbf{Y_{IE}}$: indica la trasmittanza termica periodica e viene utilizzata per determinare l'indice di prestazione energetica estiva dell'involucro[15].

- **Classe energetica**: è un indicatore alfanumerico che identifica il livello di prestazione energetica degli immobili. La lettera **A** indica un basso consumo energetico mentre la lettera **G** rappresenta un edificio con prestazioni più scarse e quindi maggiori consumi energetici. La classe **A** è inoltre suddivisa in quattro fasce, distinte dalla presenza di un numero che va da 4 a 1. La classe **A4** identificherà, tra gli edifici già ad alte prestazioni, i più performanti mentre la classe **A1** i peggiori. L'attribuzione della classe energetica per edificio è ottenuta sulla base del valore dell'indice di prestazione energetica non rinnovabile $EP_{gl,nren}$ confrontato con il corrispondente valore dell'edificio di riferimento, come in figura 1.1.

Altri attributi sono stati considerati rilevanti nell'analisi pur non contribuendo alla determinazione della classe energetica, ma con una valenza più descrittiva dell'immobile, tra cui l' *anno di costruzione*, il tipo di edificio e di costruzione.

	Classe A4	$\leq 0{,}40\ EP_{gl,nren,rif,standard\ (2019/21)}$
$0{,}40\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe A3**	$\leq 0{,}60\ EP_{gl,nren,rif,standard\ (2019/21)}$
$0{,}60\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe A2**	$\leq 0{,}80\ EP_{gl,nren,rif,standard\ (2019/21)}$
$0{,}80\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe A1**	$\leq 1{,}00\ EP_{gl,nren,rif,standard\ (2019/21)}$
$1{,}00\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe B**	$\leq 1{,}20\ EP_{gl,nren,rif,standard\ (2019/21)}$
$1{,}20\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe C**	$\leq 1{,}50\ EP_{gl,nren,rif,standard\ (2019/21)}$
$1{,}50\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe D**	$\leq 2{,}00\ EP_{gl,nren,rif,standard\ (2019/21)}$
$2{,}00\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe E**	$\leq 2{,}60\ EP_{gl,nren,rif,standard\ (2019/21)}$
$2{,}60\ EP_{gl,nren,rif,standard\ (2019/21)} <$	**Classe F**	$\leq 3{,}50\ EP_{gl,nren,rif,standard\ (2019/21)}$
	Classe G	$> 3{,}50\ EP_{gl,nren,rif,standard\ (2019/21)}$

Figura 1.1: Scala classificazioni degli edifici tramite confronto con $EP_{gl,nren}$. ©Supplemento Ordinario n.39 alla Gazzetta Ufficiale Serie generale - n. 162 (15 luglio 2015)

Capitolo 2

L'estrazione della conoscenza

Il *Data Mining* nasce dalla necessità di saper analizzare e ottenere conoscenze utili dalle grandi quantità di dati grezzi che si possono ottenere da innumerevoli contesti operativi presenti nella nostra società. Con il termine *'Data Mining'* si intende il processo di estrazione di conoscenza da *database* molto grandi, sui quali è possibile applicare algoritmi capaci di estrapolare associazioni ignote tra i dati e di renderle esplicite. Queste tecniche che permettono di analizzare grandi moli di dati e di estrarne informazioni significative risultano di fondamentale importanza nell' ambito del *decison making*. Il termine Data Mining viene spesso utilizzato in maniera impropria come sinonimo di *'Knowledge discovery in databases (KDD)'*; tuttavia tale uso risulta impreciso poiché il Data Mining è una particolare fase dell' intero processo di estrazione della conoscenza, noto come *'knowledge discovery'*. Il KDD risulta essere un processo iterativo ed interattivo, costituito da varie fasi con molte decisioni prese dall'utente fondamentali[10]. La figura 2.1 mostra uno schema di tale processo, includendo tutte le sue fasi (e.g., preparazione dei dati, selezione dei dati, pulizia dei dati), fondamentali per evitare di formulare un modello privo di significato.

La prima fase di *selezione* consiste nell'individuare i dati rilevanti rispetto all'obiettivo dell'analisi. Seguono la fase di *preprocessing*, in cui si cerca di eliminare informazioni inutili, di limitare l' effetto dei dati rumorosi, di identificare o rimuovere gli outliers e di gestire dati mancanti, e la fase di *transformation*, in cui vengono fatte delle correzioni sul dataset per evitare inconsistenze. Queste ultime due fanno uso di tecniche e strumenti che fanno parte del processo di ETL (*Extraction*, *Transformation* e *Loading*). Successivamente, si procede con la fase di *Data mining* che consiste nella scelta dell'algoritmo di *data mining* più opportuno e nell'applicazione dello stesso, allo scopo di ricercare i pattern di interesse in una particolare forma di rappresentazione[10]. L'ultima fase è basata sull'interpretazione e validazione dei dati ottenuti, così da consolidare la conoscenza estratta.

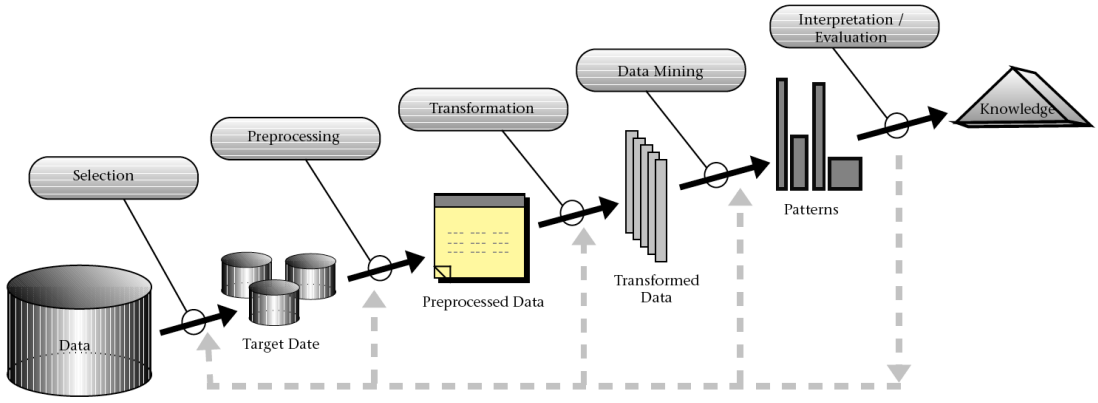

Figura 2.1: Il processo di *Knowledge Discovery in Databases*. ©Fayyad, Piatetsky-Shapiro, Smyth

2.1 Selezione, *preprocessing* e trasformazione

La fase di preparazione dei dati risulta essere fortemente *time consuming*, richiedendo circa il 70% del tempo necessario per il processo *'Knowledge discovery in databasese'*. Questo perché i dati nella realtà spesso sono incompleti a causa della mancanza di attributi ritenuti fondamentali per le analisi o perché tali attributi sono presenti come grandezze aggregate oppure sono rumorosi a causa della presenza di outliers o di errori. Infatti, nella fase successiva di *Data mining*, risulta necessario aver garantito la qualità dei dati per ottenere risultati attendibili, come definito nella cosiddetta *legge GIGO (garbage in, garbage out)*. A tal scopo, innanzitutto, occorre una comprensione del dominio applicativo e una definizione dell'obiettivo del processo di estrazione della conoscenza. Successivamente, si procede alla selezione dei dati, focalizzandosi sulle variabili fondamentali per l'analisi. Su questo sottoinsieme di dati scelti, si effettua allora la fase di *preprocessing* e trasformazione. In questa fase, risulta fondamentale la pulizia dei dati, o *data cleaning*, che tenta di risolvere tutte le incongruenze e gli errori sui dati. Tra le attività principali, risulta fondamentale la valorizzazione opportuna dei dati mancanti, o *missing values*; tale mancanza può dipendere da problemi nel funzionamento dei sistemi di raccolta dati oppure da incomprensioni da parte di chi si occupa della raccolta dati. Gli approcci proposti per la gestione dei dati mancanti sono:

- eliminare le righe che contengono dati mancanti
- ignorare i valori mancanti, se non sono un numero considerevole

- prevedere il dato mancante (attraverso una stima, sulla base dei dati noti...)

La seconda attività consiste nell'attenuare i dati rumorosi, dovuti a errori di immissione o di formato o alla presenza di *outliers*. Per quanto riguarda gli errori di immissione, che nella maggior parte dei casi riguardano l'inserimento dei nomi o degli indirizzi, ci sono numerosi *tools* in grado di estrarre, trasformare e validare i suddetti dati. Tuttavia, questi risultano essere a pagamento, per tanto una soluzione alternativa richiede un'ispezione manuale e l'applicazione di funzioni definite dall'utente, tali da poter essere riutilizzate facilmente per diverse attività di trasformazione. Invece, per l'*outlier detection*, sono disponibili in letteratura differenti soluzioni a seconda se si tratta di outlier univariati, cioè con valori anomali su una sola variabile, o multi variati, se si stanno considerando due o più variabili. Per l'analisi univariata, gli *outliers* possono essere identificati attraverso *(i)* tecniche che esaminino i principali indici statistici delle distribuzioni semplici (e.g., i valori minimi e massimi, le medie, le mediane e le relative deviazioni standard) oppure metodi grafici (e.g., istogrammi, boxplot). Per l'analisi multivariata si utilizzano tecniche che si basano su distanze o su metodi grafici (e.g., DBSCAN).

2.1.1 Metodologie per individuare gli *outlier*

In statistica, un outlier è un valore estremo di una distribuzione che si caratterizza per assumere dei valori troppo alti o troppo bassi rispetto al resto della distribuzione e che identifica quindi un caso isolato.
La definizione di outlier non è unica ed è influenzata dalla conoscenza personale del fenomeno. Quindi, lo stesso dato può essere classificato in maniera diversa da metodi e persone diverse.
In letteratura esistono diversi metodi per l'identificazione degli outlier all'interno di un dataset. I più diffusi sono:

- Un metodo parametrico, la **generalized Extreme Studentized Deviate (gESD)** [13], che fa uso della media e della deviazione del campione. Questo metodo individua gli outliers in un data-set univariato che segue approssimativamente una distribuzione normale (possibilmente ciò deve essere verificato con un plot). Un limite è che deve essere specificato il limite superiore del numero sospetto di outlier e questo valore deve essere corretto per evitare di distorcere le conclusioni.
  Dato il limite superiore r, il test gESD esegue essenzialmente r test separati: un test per un outlier, un test per 2 outlier e così via fino a r outlier.
  Il gESD è definito per le ipotesi:
  H_0 : non ci sono outlier nel dataset
  H_a : ci sono fino a r outlier nel dataset
  Il test calcola la statistica $R_i = \frac{max_i|x_i - \bar{x}|}{s}$ con $\bar{x}$ e s che indicano rispettivamente

la media e la deviazione standard.
Successivamente si rimuove l'osservazione che massimizza $|x_i - \bar{x}|$ e poi si ricalcola la statistica con *n-1* osservazioni. Si ripete questo processo finché r osservazioni sono state rimosse e così avremo R_1, R_2, ..., R_r.
I valori critici del test sono determinati specificando α e trovando poi β e $\lambda(\beta)$ tale che

$$\Pr[R_i > \lambda_i(\beta) \mid H_0] = \beta,\ i = 1,...,r$$
$$\Pr \bigcup_{i=1}^{r} [R_i > \lambda_i(\beta) | H_0] = \alpha$$

Se tutti $R_i \leqslant \lambda_i(\beta)$, allora non sono presenti outlier, altrimenti il numero di outlier è determinato dal più grande valore di i per cui $R_i > \lambda_i(\beta)$.

- La **Median Absolute Deviation (MAD)** [8],metodo non parametrico, utilizza la mediana e la deviazione mediana assoluta.
  Si ricorda che in statistica la deviazione media assoluta misura la dispersione statistica di un campione.
  Per un insieme X1, X2, ..., Xn, il valore di MAD è definito come la mediana del valore assoluto delle deviazioni dei dati dalla mediana, ovvero:

$$\text{MAD} = \text{median} \ (\ |\ \text{Xi} - \text{median}\ (\ \text{X}\ )\ )$$

  Attraverso il confronto del valore del MAD con una soglia stabilita (consigliato il valore 3.5), è possibile definire se il dato è da considerare un outlier o meno, a seconda se sia rispettivamente maggiore o minore di tale valore.

- Un metodo grafico, il **Boxplot**[7], descrive la distribuzione di un campione attraverso indici di dispersione e di posizione.
  Questo grafico è rappresentato come un rettangolo che può essere orientato in orizzontale o in verticale e che presenta due segmenti alle estremità terminati dal minimo e dal massimo dei valori.
  In particolare, la distanza tra il primo e il terzo quartile contiene il 50% dei valori e viene detta *distanza interquartilica*. Inoltre, è possibile dedurre la forma della distribuzione analizzando la distanza tra ciascun quartile e la mediana.
  I valori che sono esterni al limite inferiore o superiore rappresentano gli outlier e vengono individuati nel boxplot individualmente per poterne analizzare meglio la presenza e la posizione.

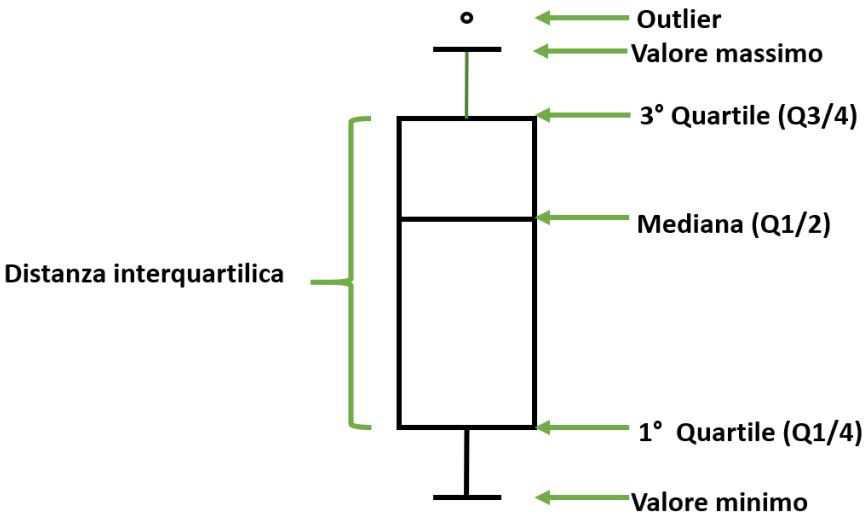

Figura 2.2: Descrizione di un esempio di boxplot.

2.2 *Data mining*

Una volta ottenuto un *dataset* pulito, si procede con la fase di *Data mining* in cui vengono selezionati gli algoritmi di *data mining* necessari per l'estrazione di *pattern* significativi.
Tali tecniche possono avere fondamentalmente due obiettivi, la cui importanza varia a seconda del contesto:

- la **predizione**: si basa sull'uso di variabili o attributi noti con lo scopo di riuscire a predire altre variabili di interesse.
- la **descrizione**: si basa sul ricercare e analizzare i pattern significativi per descrivere i dati.

Nel caso del KDD, l'aspetto descrittivo risulta essere più importante. Inoltre, queste tecniche si dividono in due categorie in base al grado di intervento dell'utente:

- Tecniche **supervisionate**: prevedono la conoscenza delle *label* dei dati che, date in input ad un algoritmo di classificazione, consentono di costruire un modello in grado di classificare correttamente futuri campioni.
- Tecniche **non supervisionate**: non prevedono nessun tipo di conoscenza pregressa dei dati ma cercano di estrapolare dagli stessi le correlazioni più interessanti.

In questo lavoro, vengono usate principalmente tecniche non supervisionate a carattere descrittivo.

2.2.1 Algoritmi di *clustering*

Tra gli algoritmi non supervisionati, di fondamentale importanza sono gli algoritmi di *clustering* che hanno lo scopo di individuare sottogruppi di dati, tali che i membri di uno stesso gruppo mostrano caratteristiche simili tra loro e dissimili dai dati appartenenti agli altri gruppi.
L' algoritmo si basa sull'ipotesi che si possa stabilire una misura di similarità, che nella maggior parte dei casi è vista come distanza dei dati in uno spazio multidimensionale. Lo scopo generale è quello di individuare dei cluster omogenei che minimizzino la distanza intracluster e massimizzino la distanza intercluster (Fig 2.3).
Le tecniche di clustering si distinguono anche in base alla possibilità di ammettere

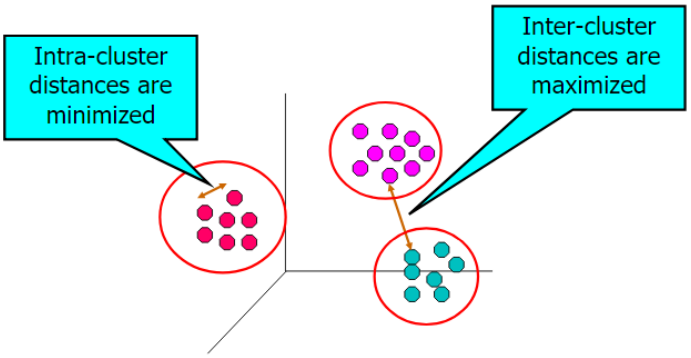

Figura 2.3: Eempio di raggruppamenti in cluster. © Tan,Steinbach,Kumar

che un elemento appartenga nello stesso momento a più cluster (*clustering non-esclusivo*) o meno (*clustering esclusivo*).
Inoltre, gli algoritmi di clustering possono seguire due approcci:

- **Bottom-up**: questo approccio prevede che all'inizio ogni elemento costituisce un cluster separato, successivamente l'algoritmo aggrega i singoli cluster in base alla loro distanza. Cluster più vicini, quindi simili, vengono accorpati per creare un unico cluster, fino a quando determinate condizioni non vengono verificate (e.g., numero di cluster ottenuti, limite delle distanza minima,...).

- **Top-down**: questo approccio, invece, inizialmente considera tutti gli oggetti come appartenenti ad un unico cluster, in seguito separa il cluster in cluster più piccoli. Questo procedura di separazione viene eseguita finché determinate

condizioni non vengono soddisfatte, ad esempio fino a quando non si è ottenuto il numero di cluster prefissato.

Le tecniche di clustering si dividono inoltre sulla modalità di divisione dello spazio. Gli **algoritmi gerarchici** producono una serie di cluster annidati che vengono rappresentati attraverso un dendrogramma, una rappresentazione ad albero come in figura 2.4. Anche per questo tipo di algoritmo si possono ancora suddividere due

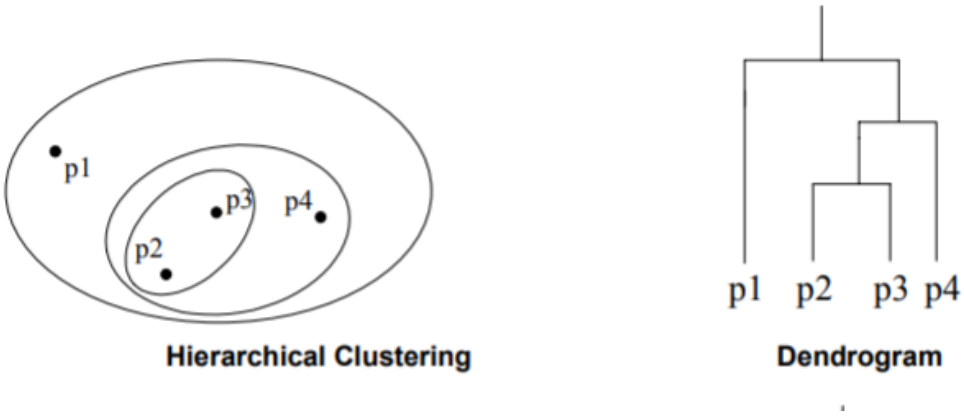

Figura 2.4: Clustering gerarchico rappresentato attraverso un dendrogramma. © Tan,Steinbach,Kumar

approcci: quello agglomerativo e quello divisivo. Il primo segue il principio della filosofia bottom-up, per cui si assume che inizialmente ogni cluster contiene un unico punto e via via i cluster più vicini vengono fusi finché tutti i punti sono accorpati in un unico cluster; il secondo, usando la filosofia top-down, considera tutti i dati appartenenti ad un unico cluster e ad ogni iterazione sceglie quale cluster dividere in due. Entrambi gli approcci necessitano di una misura di similarità che permetta di scegliere quale coppia di cluster fondere o dividere a seconda dell'approccio scelto.
Uno dei vantaggi di questo tipo di clustering è che non è necessario avere un numero prefissato di cluster, ma qualora si decida di ottenere un certo numero di sottogruppi, sarà sufficiente tagliare il dendrogramma ad un opportuno livello; tuttavia, non è adatto per dataset molto grandi e presenta una rigidità dettata dal fatto che, una volta che viene effettuata una fusione o una separazione tra due oggetti, non si può più tornare indietro.
Gli **algoritmi partizionali** generano una singola partizione dei dati, costituita da un inseme di cluster disgiunti la cui unione restituisce il dataset originale, come evidente nella figura 2.5.
Questo algoritmo richiede una conoscenza apriori del numero K di cluster in cui partizionare il dataset. L'insieme dei dati viene diviso in K cluster e per ogni cluster

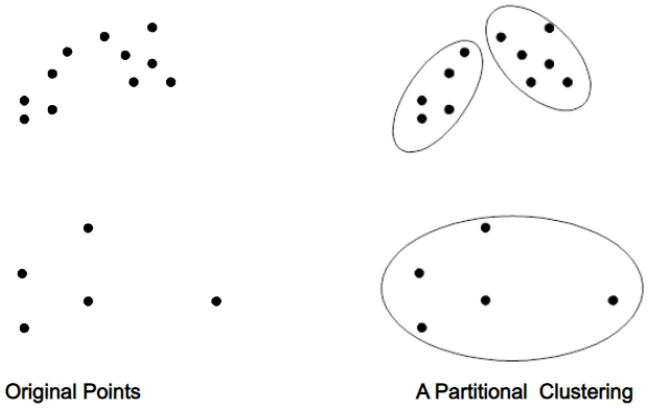

Figura 2.5: Esempio di clustering partizionale.© Tan,Steinbach,Kumar

vengono calcolati i centroidi (definiti nel paragrafo 3.3.1). Successivamente, vengono riassegnati tutti gli elementi in base al centroide più vicino finché non viene raggiunto il valore minimo dell'errore totale, calcolato come somma degli errori quadratici di ogni singolo cluster.
Tuttavia, il rischio è quello di rimanere nell'ottimo locale, perciò sarebbe opportuno enumerare tutte le possibili partizioni oppure utilizzare tecniche euristiche che considerano ogni cluster come un unico punto. Ad ogni modo, queste tecniche euristiche risultano efficienti se ci si interfaccia con dataset dalla forma sferica e non troppo grandi.

Il concetto di distanza nel clustering.
Come detto precedentemente, la misura di similarità tra un elemento e un altro è fondamentale perché consente di definire l'appartenenza a un cluster o meno; inoltre, possiamo affermare che due punti o due cluster sono considerati simili se sono abbastanza vicini. Il concetto di somiglianza è legato a quello di distanza e, per tale ragione, si usa una funzione di distanza, che può essere diversa per tipi di dato diversi, per misurare quanto due oggetti sono simili. L'idea alla base è quella di immaginare ogni dato come un punto in uno spazio multidimensionale, o per semplicità in uno spazio bidimensionale. Osservando la figura 2.6, è immediato notare la presenza di due cluster ognuno dei quali include elementi la cui distanza è decisamente inferiore rispetto a quella tra elementi appartenenti a cluster diversi.
Nel lavoro svolto in questa tesi, è stata utilizzata come distanza per gli algoritmi di clustering la distanza euclidea.
Di seguito, sono riportati brevemente i concetti di distanza e similarità.
Dati tre punti qualsiasi x,y,z, appartenenti a S, rappresentazione simbolica di uno

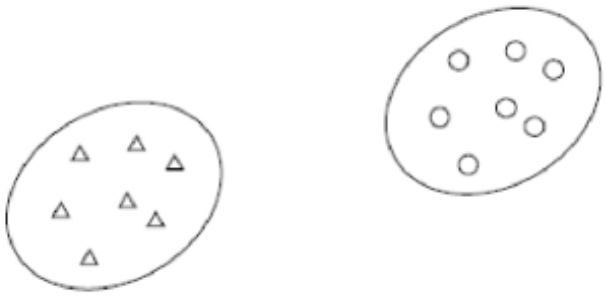

Figura 2.6: Esempio di clustering. ©Dulli, Furini, Peron

spazio, si definisce *distanza (o metrica)* tra due punti una funzione *d(x,y)* che gode delle seguenti proprietà:

1. $d(x, y) \geq 0, \forall x, y \in S$
2. $d(x, y) = 0 \iff x = y$
3. $d(x, y) = d(y, x), \forall x, y \in S$
4. $d(x, y) \leq d(x, z) + d(z, y)\ , \forall x, y, z \in S$

Le condizioni elencate richiedono che la funzione distanza non sia nulla (1), goda della proprietà riflessiva (2), sia una funzione simmetrica (3) e soddisfi la cosiddetta *disuguaglianza triangolare* (4), ovvero la distanza fra due punti deve essere minore o uguale alla somma delle distanze tra i due punti precedentemente considerati e un terzo punto distinto, come mostrato in figura 2.7.

Dati due punti qualsiasi *x,y*, appartenenti a *S*, rappresentazione simbolica di uno spazio geometrico, si definisce *misura di similarità* tra due punti una funzione *s(x,y)* che gode delle seguenti proprietà:

1. $s(x, y) = 1 \iff x = y$
2. $s(x, y) = s(y, x), \forall x, y \in S$

Le condizioni elencate richiedono che, se due punti sono uguali, la similarità sia massima e che la funzione sia simmetrica. Infine, dati due vettori X e Y di lunghezza l:

$$X = (x_1, x_2, ..., x_\mathrm{i}, ..., x_l)$$
$$Y = (y_1, y_2, ..., y_\mathrm{i}, ..., y_l)$$

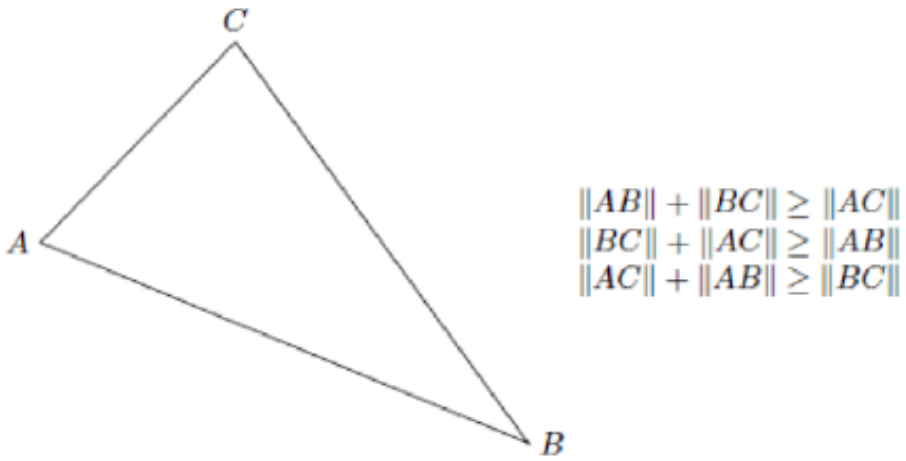

Figura 2.7: Disuguaglianza triangolare. ©Dulli, Furini, Peron

possiamo definire la *distanza euclidea* come:

$$D(X,Y) = \sqrt{\sum_{i=1}^{l}(x_\mathrm{i} - y_\mathrm{i})^2}$$

2.2.2 Regole di associazione

Nel processo di estrazione della conoscenza, ricoprono un ruolo di fondamentale importanza le cosiddette *regole di associazione* poiché permettono di estrarre da un ampio *database* transazionale correlazioni a diverso livello di dettaglio tra i dati, utili per estrarre conoscenza. La grande efficacia di questo processo dipende dal fatto che l'analisi effettuata non prevede una conoscenza apriori delle relazioni cercate tra i dati. Da ciò ne consegue che il numero di regole estratte può essere considerevole. Il concetto di regole associative nasce grazie ad un articolo pubblicato nel 1993 da *Rakesh Agrawal*[9], in cui si descrive la *basket market analysis*, cioè un processo che analizza le abitudini di acquisto dei clienti per estrarre le relazioni tra i prodotti comprati e usarle per supportare le decisione delle azienda in merito ai prodotti da vendere, alla posizione degli stessi sugli scaffali o anche per capire su quali prodotti fare le offerte promozionali. Un esempio comunemente utilizzato di regola potrebbe essere quella che ha estratto una relazione tra la vendita della birra e quella dei pannolini dalle transazioni delle casse di un supermercato e che potrebbe essere utilizzata per capire qual è la posizione migliore in cui collocare i suddetti prodotti in un supermercato. Ovviamente, le correlazioni possono essere estratte tra diversi tipi di dati, non solo prodotti commerciali; basti pensare a quanto può essere utile e interessante cogliere tra le schede mediche dei pazienti delle ipotetiche relazioni

tra effetti di terapie diverse. Le uniche caratteristiche che i *dataset* devono avere per estrarre correttamente le regole è che siano abbastanza grandi e transazionali, cioè costituiti da una collezione di transazioni ognuna delle quali contiene almeno un item e può essere rappresentata in differenti formati(e.g., documenti testuali, dati strutturati).

Definizione formale.

Data una regola di associazione nella forma:

$$A \Rightarrow B$$

si intende che esiste una coesistenza di A e B, e non una causalità tra gli elementi, come si potrebbe dedurre intuitivamente.
Definiamo una collezione di n oggetti, detti *item*, $I = i_1, i_2, ..., i_n$ nota come *itemset*. Un sottoinsieme dell' *itemset* costituito da k elementi verrà invece definito *k-itemset*. L'insieme delle transazioni che costituiscono il *dataset* di riferimento è definito come $D = t_1, t_2, ..., t_m$ ed ogni transazione è un sottoinsieme di attributi presenti in I.
Con riferimento alla regola descritta precedentemente, vengono chiamati rispettivamente gli *itemset* A e B, *antecedente* e *conseguente*. Per una corretta interpretazione delle regole, occorre definire delle metriche che esprimano la qualità della regola estratta. Il *supporto* è il rapporto tra il numero di transazioni che contengono sia l'*itemset* A sia l'*itemset* B e la cardinalità dell'intero *database* transazionale D. La *confidenza* esprime invece il numero di volte in cui compare B nelle transazioni che contengono l'*itemset* A.
Queste misure possono essere interpretate equivalentemente da un punto di vista probabilistico. Per cui definiremo il supporto della regola estratta $A \Rightarrow B$

$$sup(A \Rightarrow B) = P(A \cap B)$$

Possiamo perciò interpretare il supporto come la probabilità che gli *itemset* A e B siano presenti nella stessa transazione. Inoltre, la confidenza della suddetta regola potrà essere espressa statisticamente nel seguente modo:

$$conf(A \Rightarrow B) = \frac{sup(A,B)}{sup(A)} = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

Questa interpretazione esprime la confidenza della regola $A \Rightarrow B$ come la probabilità condizionata di trovare in una transazione B avendo già trovato A.
Allo scopo di estrarre le regole di associazione più significative, occorre che esse rispettino alcuni vincoli, in particolar modo è necessario che il supporto sia maggiore di una soglia minima definita sup_{min} e che la confidenza superi il valore minimo consentito $conf_{min}$. Il risultato sarebbe definito completo se entrambi i vincolo fossero

soddisfatti da tutte le regole estratte.

Esempio di calcolo del supporto e della confidenza.
Dato l' itemset $I = \{I1, I2, I3, I4, I5, I6\}$ e il dataset transazionale in figura 2.8, fissiamo i limiti sup_{min} e $conf_{min}$ uguali al 50%.

ID Transazione	Itemset
1	{A1,A2,A3}
2	{A1,A4}
3	{A1,A3}
4	{A2,A5,A6}

Figura 2.8: Esempio di dataset transazionale

Risultati ottenuti:

(I) $sup(A1) = 75\%$

(II) $sup(A1, A2) = 25\%$

(III) $sup(A1 \Rightarrow A3) = 50\%$

(IV) $conf(A1 \Rightarrow A3) = 66\%$

Dal risultato *(I)* otteniamo che l' itemset $\{A1\}$ è frequente poiché ha il supporto maggiore della soglia sup_{min} mentre dal risultato *(II)* si evince che l' itemset $\{A1, A2\}$ non è frequente. Dalla confidenza e dal supporto calcolati nei punti *(III)* e *(IV)* deduciamo che la regola $A1 \Rightarrow A3$ è forte perché soddisfa le soglie minime imposte sup_{min} e $conf_{min}$.

Un approccio per estrarre le regole associative potrebbe essere quello di estrarre tutte le regole, calcolare di ognuna di esse supporto e confidenza e poi eliminare le regole che non rispettano le soglie minime. Questo metodo risulta essere computazionalmente dispendioso, anche perché possono esistere due regole che hanno lo stesso supporto ma diversa confidenza. A questo punto, sarebbe più opportuno calcolare separatamente supporto e confidenza. A tal proposito, l'approccio più comunemente utilizzato si basa su due fasi:

- Estrazione degli item più frequenti: attraverso tecniche differenti vengono estratti gli item con il supporto maggiore del limite scelto e, pur essendo lo step più complesso da un punto di vista computazionale, è possibile ridurre tale complessità agendo sulla soglia del supporto.

- Estrazione delle regole associative: vengono estratte le regole con un livello di confidenza alto, ovvero superiore al limite scelto, e sono generate usando tutte le possibili combinazioni binarie degli itemset più frequenti.

Il primo approccio descritto, ovvero quello che considera ogni itemset come un possibile frequente calcolando per ognuno il supporto e la confidenza, risulta essere computazionalmente oneroso perché ha una complessità pari a $O(M2^{n}w)$, dove w rappresenta la lunghezza della transazione. Per tale ragione sono stati proposti tre tipi di soluzioni:

- **Diminuire il numero di candidati** usando tecniche di pruning per diminuire lo spazio di ricerca

- **Diminuire il numero di transazioni** se si ritiene che il numero degli itemset è troppo grande

- **Diminuire il numero dei confronti** ottimizzare la ricerca memorizzando le transazioni e i candidati in strutture dati più efficienti.

I metodi comunemente più usati per l'estrazione degli itemset più frequenti sono l'algoritmo **Apriori** e l' **FP-growth**.
L'algoritmo **Apriori** è basato sul principio per cui "se un itemset è frequente, allora saranno frequenti anche tutti suoi sottoinsiemi"; ciò dipende dalla proprietà anti-monotona che caratterizza il supporto e secondo la quale il supporto di un itemset non potrà mai essere più grande di quello dei suoi sottoinsiemi[1]. Definiamo C_k la collezione di candidati di dimensione k e L_k la collezione degli itemset frequenti; il primo passo dell'algoritmo consiste nel generate i candidati in C_{k+1} attraverso la *join* di L_k con sè stesso e selezionando solo quelli che hanno lunghezza pari a *k+1*. Tra questi vengono scelti i candidati che hanno un supporto maggiore del sup_{min}. A questo punto entra in gioco il principio di base dell'algoritmo per cui vengono eliminati i *k-itemset* che contengono almeno un sottoinsieme già ritenuto non frequente. In questo modo si cerca di ridurre il numero di candidati da esplorare pur dovendo comunque esplorare il dataset ad ogni iterazione e, inoltre, si potrebbero generare allo stesso modo un numero consistente di candidati. I passi dell'algoritmo sono descritti in pseudocodice nella figura 2.9.
L'algoritmo **FP-growth** si differenzia da quello Apriori poiché non ha la fase di generazione dei candidati e, per tanto, presenta delle perfomance migliori. L'estrazione dei pattern frequenti invece viene effettuata creando un *FP-tree* che permette di comprimere il dataset in una struttura più compatta e attraverso una visita ricorsiva dell'albero, così con solo due scansioni del database è possibile contare i supporti degli item e costruire l' FP-tree.

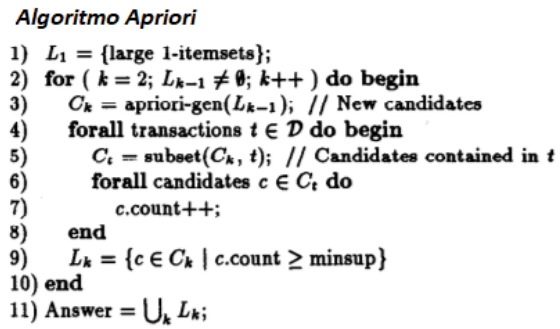

Figura 2.9: Descrizione passi dell' algoritmo Apriori.©Rakesh Agrawal, Ramakrishnan Srikant

2.3 Interpretazione e validazione della conoscenza estratta

I modelli ottenuti dalle analisi precedenti richiedono un processo di studio e di verifica. In particolare, risulta fondamentale il supporto dell'esperto di dominio per confermare l'interpretazione di tali modelli o per evidenziare possibili inconsistenze. La presenza di inconsistenze nei risultati ottenuti implica un processo di retroazione agli step precedenti per effettuare migliorie o ulteriori iterazioni al fine di potenziare l'efficacia dei modelli estratti.
Inoltre, occorre un consolidamento della conoscenza estratta che, oltre a essere validata dall'esperto di dominio, sia oggettivamente verificata attraverso metodi statistici o tramite test sui dati reali.
Infine, alcuni strumenti molto utili ai fini di una corretta interpretazione e validazione del modello, ma anche per una maggiore fruibilità della conoscenza estratta, sono i *tool* di visualizzazione.

Capitolo 3

Architettura sviluppata

L'architettura sviluppata per l'estrazione e la visualizzazione dei pattern relativi alle certificazioni energetiche della regione Piemonte è rappresentata schematicamente nella figura 3.1. Questo sistema, che prende il nome di TUCANA(**TU**rin **C**ertificates **ANA**lysis), è costituito da quattro blocchi fondamentali : *la raccolta e l' integrazione dei dati*, *la preparazione dei dati*, *l'analisi dei cluster* e *la validazione e la visualizzazione della conoscenza*.

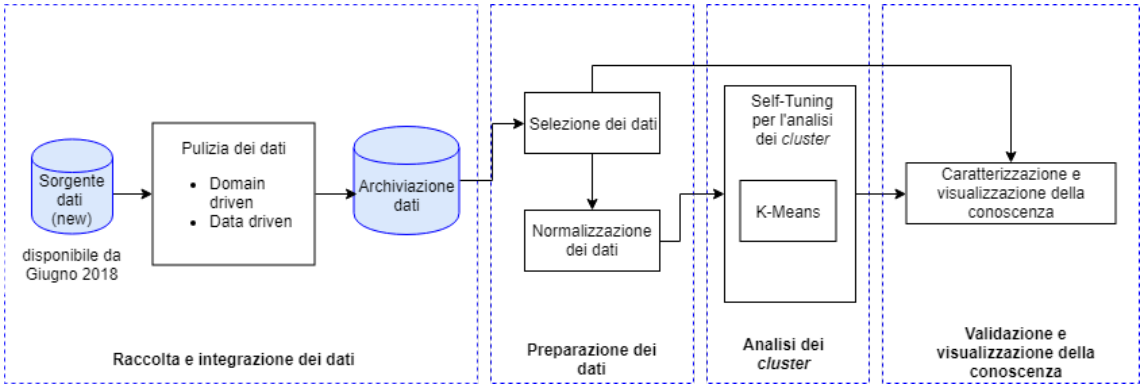

Figura 3.1: Architettura

Ognuno di questi blocchi rappresenta uno step fondamentale per la definizione dell'obiettivo finale.
La *raccolta e integrazione dei dati* estrae dal Catasto energetico relativo agli edifici della regione Piemonte i dati delle certificazioni energetiche emesse ed effettua una pulizia dei dati, preparatoria per l' uso degli stessi negli step successivi. La fase di *preparazione dei dati* si basa su di un processo di selezione degli attributi fondamentali e di normalizzazione degli stessi. L' *analisi dei cluster* consiste nell'applicazione dell' algoritmo K-means per individuare gruppi di unità abitative accomunate da

caratteristiche termo-fisiche simili. La *validazione e visualizzazione della conoscenza* consente di rappresentare efficacemente sia i dati raccolti dalle sorgenti identificate sia la conoscenza prodotta attraverso l' esplorazione.

3.1 Raccolta e integrazione dei dati

Per il lavoro mostrato in questa tesi, è stato utilizzato il *database* contenente una parte del Catasto energetico degli edifici delle Regione Piemonte; in particolare, ci riferiamo a quella porzione di *dataset* che contiene le certificazioni emesse dal 2016 fino al primo semestre del 2018. Il *database* utilizzato contiene i certificati raggruppati in diversi file Excel in base alle caratteristiche descritte. Data la grande quantità di attributi disponibili nei differenti file, non sono stati utilizzati tutti, ma è stata fatta una selezione degli stessi in collaborazione con esperti del Dipartimento di Energia del Politecnico di Torino, i quali hanno individuato gli attributi dei certificati ritenuti fondamentali, ovvero quelli che influiscono maggiormente sulle prestazioni energetiche degli edifici.

Di seguito sono riportati gli attributi considerati.

- Fattore forma [m^{-1}]
- Trasmittanza trasparente [W/m^2K]
- Trasmittanza opaca [W/m^2K]
- Rendimento di generazione
- Rendimento di emissione
- Rendimento di regolazione
- Rendimento di distribuzione
- Indice di prestazione energetica globale [kWh/m^2anno]
- Indice di prestazione energetica invernale [kWh/m^2anno]
- Indice di prestazione energetica invernale dell'involucro [kWh/m^2anno]
- Superficie utile [m^2]
- Superficie disperdente [m^2]

3.1.1 *Pulizia dei dati*

La fase di estrazione della conoscenza necessita di dati consistenti e di qualità. Per tale motivo, la pulizia dei dati garantisce l'eliminazione degli errori e delle incongruenze. Ciò è reso possibile grazie a tre attività fondamentali: la *correzione di errori di immissione in campi testuali*, lo *scaling dei dati* e l' *eliminazione degli outlier*.

- **Correzione di errori di immissione in campi testuali.**

  Poiché per la visualizzazione dei dati su mappe è necessario avere una localizzazione precisa degli edifici associati ai certificati disponibili, occorre che gli attributi che contribuiscono alla geolocalizzazione degli edifici siano corretti e presentino un formato standard. Purtroppo questi campi presentano alcune criticità. In particolar modo, l'attributo relativo agli indirizzi presenta errori di battitura, caratteri codificati erroneamente e formati non standard; inoltre, è spesso presente il CAP generico della città di riferimento e le coordinate non sempre sono correttamente associate al corrispondente edificio.
  L' architettura sviluppata risolve i problemi sopracitati utilizzando due tecniche che verranno applicate successivamente alla conversione dei caratteri in formato ASCII.
  La prima tecnica fa uso di un servizio, basato sulle API geocoding, in grado di convertire gli indirizzi stradali in coordinate geografiche (latitudine e longitudine), correggendo automaticamente gli indirizzi che presentano errori nel formato o errori di battitura e fornendo l'indirizzo completo, comprensivo di CAP.
  Tuttavia, questo servizio presenta un forte limite in quanto risulta avere una versione *free* che permette di fare un numero limitato di richieste al giorno.
  Per tale ragione, è stato necessario sviluppare un secondo metodo di sanitizzazione degli indirizzi. Questa seconda tecnica si basa sul confronto tra gli indirizzi da verificare e gli indirizzi presenti nel viario di Torino, ottenuto dal «Geoportale Comune di Torino[1]».
  Per calcolare l'indice di similarità tra gli indirizzi, si utilizza la *distanza di Levenshtein*.
  La *distanza di Levenshtein* consente di determinare quanto due stringhe siano simili, contando il numero minimo di modifiche elementari (cancellazione, sostituzione o inserimento di un carattere) che permettono di trasformare la stringa A nella stringa B.
  L'indice di similarità, che consente di esprimere quanto le due stringhe siano simili, viene calcolato nel seguente modo:

[1] http://geoportale.comune.torino.it/web/

$$Indice\_di\_similarità = \frac{(Len_{\text{SUM}} - DistanzadiLevenshtein)}{Len_{\text{SUM}}}$$

dove :
NS = numero di sostituzioni necessarie per la trasformazione da A a B
NI = numero di inserimenti necessari per la trasformazione da A a B
NC = numero di cancellazioni necessarie per la trasformazione da A a B
$DistanzadiLevenshtein = 2NS + NI + NC$
$Len_{\text{SUM}} = len(A) + len(B)$

Questo indice sarà uguale a 1 se le due stringhe sono identiche.

Esempio di calcolo dell'indice di similarità usando la distanza di Levenshtein.
A = "ac"
B = "ab"
Distanza di Levenshtein = 1
Indice di similarità = $\frac{(4-2*1)}{4} = 0.5$

		a	c
	0	1	2
a	1	0	1
b	2	1	2

Figura 3.2: Esempio conversione della stringa "ac" in "ab"

Allo scopo di ottenere una geolocalizzazione attendibile degli edifici, verranno utilizzate entrambe le tecniche. In particolare, come primo step, si utilizzerà la tecnica basata sul confronto; se da questa analisi non si riuscirà a risolvere l'indirizzo, ovvero se il massimo indice di similarità ottenuto sarà minore di 0.9, allora si utilizzeranno le API geocoding.
Di seguito è riportato in pseudocodice l'algoritmo di sanitizzazione degli attributi geografici (Algoritmo 1).

Gli input richiesti sono il nome del file che contiene il dataset (*datasetName*) e il nome del file che contiene tutti gli indirizzi della città considerata (*fileViarioName*).
Innanzitutto, viene eseguita la funzione *conversioneCaratteriASCII(datasetName)*

Algorithm 1 Algoritmo di sanitizzazione degli attributi geografici

Input: *datasetName, fileViarioName*
Output: *datasetNameCorrect*

conversioneCaratteriASCII($datasetName$)
$DizViario \leftarrow$ creoDizionarioDaViario($fileViarioName$)
$DizViarioCombinazioni \leftarrow$ creoDizionarioCombinazioniDaViario()
for $Certificato$ in $dataset$ **do**
  $DELETED = False$
  $Indirizozo, NumCiv \leftarrow$ estraiIndirizzoNumCivDaCertificato($Certificato$)
  if $Indirizzo$!= "" **then**
    convertiFormatoCorretto($Indirizzo, NumCiv$)
    if $NUMCIV$ è assente **then**
      eliminaCertificato($Certificato$)
      $DELETED \leftarrow True$
    else
      if $Indirizzo$ è già stato analizzato **then**
        $IndirizzoTrovato, LevIndice \leftarrow$ prendiIndirizzoEIndiceDaVieProcessate($Indirizzo$)
      else
        $IndirizzoTrovato, LevIndice \leftarrow$ cercaIndirizzoDaViario($Indirizzo, DizViario$)
      end if
    end if
  else
    $eliminaCertificato(certificato)$
    $DELETED \leftarrow True$
  end if
  if $DELETED == False$ **then**
    if $LevIndice < 0.875$ **then**
      $RESULT, IndirizzoTrovato \leftarrow$ cercaSimileIndirizzoDaGeocoding($Indirizzo, NumCiv$)
      if $RESULT == ERROR$ **then**
        $eliminaCertificato(Certificato)$
        $CertificatoAccettato \leftarrow False$
      else
        $CertificatoAccettato \leftarrow True$
      end if
    else
      $CertificatoAccettato \leftarrow True$
    end if
  end if
  if $CertificatoAccettato == True$ **then**
    $Cap, Lat, Long \leftarrow$ cercoCAPeCoordinate($IndirizzoTrovato, NumCiv$)
    $datasetNameCorrect \leftarrow$ aggiornaCertificatoCorretto($Certificato, IndirizzoTrovato$, $NumCiv, Cap, Lat, Long$)
  end if
end for

che converte in ASCII tutti i caratteri codificati con un formato differente. Successivamente, viene eseguita la funzione *creoDizionarioDaViario(fileViarioName)* che, a partire dal viario della città di Torino, crea una struttura dati opportuna. In particolare, è stato creato un oggetto che, data una via, permetterà di ottenere tutti i CAP ai quali essa è associata e altre informazioni utili alla localizzazione.
Le strutture dati utilizzate sono note come dizionario e, attraverso una chiave, consentono di accedere alla corrispondente cella del dizionario. La struttura dati più esterna creata avrà come chiave la via e come cella un oggetto della classe *ClasseVia*. Ogni istanza della *ClasseVia* è costituita da un ulteriore dizionario che ha come chiave uno dei CAP associati a quella via e come *item* un oggetto della *ClasseCAP*. Ognuno di questi ultimi oggetti ha come chiave il numero civico estratto dal viario e come *item* le relative coordinate.
In seguito, viene eseguita la funzione *creoDizionarioCombinazioniDaViario()* che crea un altro dizionario il quale contiene come chiave la via estratta dal viario e come *item* una lista contenente tutti i modi in cui quella via può essere scritta. Ciò è stato possibile creando tutte le combinazioni delle sottostringhe che costituiscono il nome completo della via.
Una volta che sono state generate le strutture dati necessarie per l'elaborazione dei dati, si procede con l'analisi di ogni singolo certificato. Per ogni certificato, viene estratto l'indirizzo e il numero civico attraverso la funzione *estraiIndirizzoNumCivDaCertificato*; qualora l'indirizzo sia un campo vuoto, quel certificato viene eliminato, altrimenti si cerca di ottenere la via corrispondente all'indirizzo estratto dal dataset. Per ottimizzare la ricerca, vengono memorizzate le informazioni ottenute dall'elaborazione di un indirizzo. Per tanto, prima di chiamare la funzione *cercaIndirizzoDaViario*, si verifica se l'indirizzo corrente è già stato analizzato e, in caso positivo, viene estratto l'indirizzo già ottenuto e il relativo indice di similarità attraverso la funzione *prendiIndirizzoEIndiceDaVieProcessate*. Nel caso in cui l'indirizzo da elaborare non è mai stato processato, viene eseguita la funziona *cercaIndirizzoDaViario* che ottiene dal dizionario delle combinazioni delle vie l'indirizzo più simile e il relativo indice di similarità.
Ottenuta la via più simile, si verifica se il corrispondente indice di similarità ricavato è inferiore a 0.875, valore stabilito a seguito di diverse prove sperimentali. Qualora il valore fosse inferiore a tale soglia, viene effettuata una richiesta al servizio che fa uso delle API Geocoding per ottenere un risultato più affidabile. Se tale servizio dovesse restituire un errore, poiché incapace di individuare un indirizzo simile, allora il certificato viene eliminato.
Infine, ottenuta la via corrispondente all'indirizzo di riferimento, viene ricavato il corrispondente CAP e le relative coordinate geografiche. Se la richiesta è stata fatta con le API Geocoding, questi valori sono restituiti dal servizio

stesso; se invece la via è stata ottenuta dal confronto con le vie del dizionario, tali informazioni vengono ricavate consultando il *DizViario*.
Tuttavia un limite nell'uso del viario è l'assenza di tutti i numeri civici e delle rispettive coordinate. Per tanto, se il numero civico estratto dal dataset non fosse presente nel dizionario, vengono attribuite le informazioni del numero civico più prossimo.
Attraverso questa metodologia, il 97% degli indirizzi è stato risolto con la tecnica del confronto, il 2,8% con l'uso delle API Geocoding e lo 0,2% degli indirizzi non è stato risolto ed è stato eliminato dal dataset.

- ***Scaling* dei dati.**
  Durante una prima fase esplorativa dei dati, sono stati analizzati i grafici delle distribuzioni degli attributi ritenuti maggiormente significativi per comprendere l'andamento dei valori e la presenza di alcune anomalie.
  Da questo studio è emerso che i dati relativi ai rendimenti, il cui range di validità include valori da 0 a 1, assumono anche valori maggiori dell'unità; tali valori potrebbero esprimere il rendimento in forma percentuale.
  Per tale ragione, è stato necessario scalare di un fattore 100 i valori dei rendimenti maggiori del limite superiore del relativo range di validità.

- **Eliminazione degli *outlier*.**
  Dalle analisi esplorative iniziali del *dataset* è emersa la presenza di numerosi *outlier*. Ciò ha reso necessaria l'applicazione di tecniche di individuazione di questi valori anomali, per evitare che le statistiche e le deduzioni successive potessero essere fuorvianti.
  Nella nostra analisi sono state utilizzate tre tecniche di riconoscimento degli *outlier* univariate: la *Generalized Extreme Studentized Deviate (gESD)*[13], la *Median Absolute Deviation (MAD)*[8] e il *Boxplot*[7].
  Queste tecniche sono state applicate su alcuni attributi ritenuti fondamentali e hanno fornito dei limiti indicativi circa il range di validità di ogni singolo attributo. Tali intervalli sono stati forniti come supporto all' esperto di dominio, che ha estratto i range di validità di questi attributi scelti, secondo i limiti fisici e tenendo conto della distribuzione degli stessi nel *dataset* a disposizione.
  Inoltre, è stato possibile osservare dall' eliminazione degli *outlier* per questi attributi, quale dei suddetti metodi fosse il migliore in termini di minor numero di falsi positivi individuati.

3.2 Preparazione dei dati

3.2.1 Selezione dei dati

Allo scopo di estrarre i pattern significativi, occorre raggruppare i dati in *cluster* così da associare edifici con caratteristiche termo-fisiche tra loro simili. Per applicare questi algoritmi, è necessario eliminare le differenze numeriche tra le variabili. Questo perché se una *feature* è caratterizzata da valori molto più grandi, avrà un peso maggiore rispetto alle altre e, pertanto, risulterà dominante. A questo scopo, analizzando l'andamento delle variabili utilizzate nel nostro lavoro, è stato necessario normalizzarle, cosicché i valori 'riscalati' ricadessero in un range definito. Tuttavia, affinché la normalizzazione rispecchi l'effettivo andamento delle variabili, è necessario che non sia influenzata dalla presenza di *outlier*. L'approccio per la pulizia degli *outlier* descritto precedentemente è stato *domain driven* ed usava tecniche univariate; pertanto è stato applicato solo ai sette attributi ritenuti fondamentali. Per gli altri attributi che vengono utilizzati per il clustering o per altre prove sperimentali, è preferibile usare un algoritmo che sia in grado di individuare dati rumorosi valutando contemporaneamente più variabili. A questo scopo si presta molto bene un algoritmo che lavora sulla densità dei dati, come il DBSCAN[18].

3.2.2 Algoritmi basati sulla densità: DBSCAN

Tra gli algoritmi di clustering, occorre citare sicuramente gli algoritmi basati sulla densità. Questi algoritmi sono in grado di determinare i cluster non basandosi sulla distanza tra i pattern, bensì sulla loro densità; questa peculiarità fa sì che questi algoritmi riescano intrinsecamente a fornire un'ulteriore funzionalità, ovvero quella di riconoscere i dati rumorosi poiché appartenenti a zone a più scarsa densità. Tra questi algoritmi di clustering, il più noto è il *DBSCAN*.
Il DBSCAN è in grado di definire i cluster basandosi su un principio fondamentale noto come *density-reachability*, secondo il quale due punti sono ritenuti parte di un cluster se la loro distanza è inferiore ad una soglia stabilita ϵ e se ogni pattern è circondato da un numero di pattern superiore ad un certo valore $MinPts$. Con il termine distanza non si fa riferimento necessariamente alla distanza euclidea, ma si può scegliere quella più opportuna in base all'applicazione, anche perché i pattern coinvolti non necessariamente devono appartenere ad uno spazio vettoriale. Per descrivere formalmente il concetto di *density-reachability*, occorre definire in primo luogo cosa si intende per *Eps-neighborhood*.
Dato un pattern p e un database D costituito da m pattern, definiamo *Eps-neighborhood* di p:

$$Neps(p) = \{q \in D | distanza(p, q) \leq \epsilon\} \text{ [18]}$$

A questo punto è possibile determinare l'appartenenza di un pattern p ad un cluster verificando se il relativo *Eps-neighborhood* $Neps(p)$ possiede un numero di pattern maggiore di $MinPts$. In seguito, definiamo un pattern p *directly densitiy-reachable* da un pattern r se, definiti ϵ e $MinPts$, sono verificate le seguenti condizioni:

- $p \in Neps(r)$, p appartiene al vicinato di r [18]
- $|Neps(r)| \geq MinPts$, r è un *core point* [18]

I punti che soddisfano la seconda condizione sono noti come *core point*; tuttavia, ci possono essere pattern che non sono circondati da abbastanza punti nel raggio di ϵ ma sono molto vicini ad un *core point* e che, per tanto, vengono definiti *border point*. Il DBSCAN individua inoltre dei pattern che in base alla distanza utilizzata non sono identificabili né come *core point* né come *border point* (Fig 3.3); infatti, questi punti non sono stati attribuiti a nessun cluster e vengono perciò considerati dati rumorosi, definendoli *outlier o noise point*.
"Un pattern p è definito allora *density-reachable* da un punto r se c'è una catena di punti $p_1, ..., p_n$, con $p_1 = r, p_n = p$ tale che $p_(i + 1)$ sia *directly densitiy-reachable* da p_i"[18]. La relazione di *densitiy-reachability* è simmetrica infatti dato un pattern p che sia *core point*, una volta che viene trovato il suo cluster di appartenenza raggruppando tutti i pattern che sono *densitiy-reachable* da p, il cluster risulta indipendente dal *core point* individuato. Questo algoritmo presenta il vantaggio di riuscire a trova-

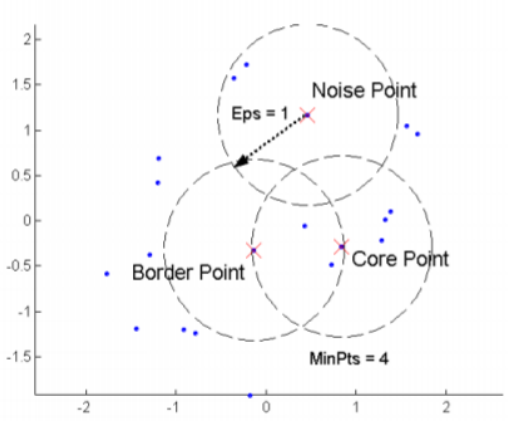

Figura 3.3: Descrizione di pattern core, border o noise nel DBSCAN.© Tan,Steinbach,Kumar

re cluster di forme molto diverse senza conoscerne il numero a priori e di necessitare solo di due parametri (ϵ e numero minimo di punti per cluster), oltre a riconoscere la presenza di outliers. Tuttavia, il DBSCAN risulta fallimentare nel caso in cui si abbiano cluster con densità molto diverse e a causa della cosiddetta *high dimensional*

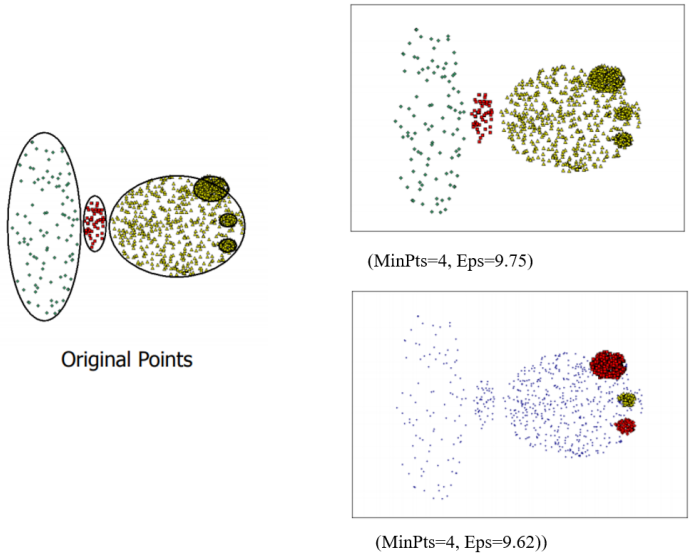

Figura 3.4: Esempio fallimentare di clustering con DBSCAN.© Tan,Steinbach,Kumar

data (Fig 3.4). Questi svantaggi sono fortemente connessi alla scelta dei parametri richiesti dall'algoritmo; nel primo caso, perché non è possibile scegliere la coppia di valori di ϵ e di $MinPts$ in maniera opportuna per ogni cluster e, nel secondo caso, perché spesso la scelta della distanza utilizzata non risulta appropriata. A tal proposito è stata proposta un'euristica per la determinazione dei suddetti parametri basata sull'idea che i punti appartenenti ad un cluster abbiano i loro *k nearest neighbors* più o meno alla stessa distanza e che i punti rumorosi abbiano i relativi *k nearest neighbors* a una distanza maggiore. Così, rappresentando su un grafico le distanze ordinate dei punti da i loro *k nearest neighbors*, è possibile avere un' idea della distribuzione di densità. Tuttavia la mancata certezza circa la validità di tale euristica e l'alternativa basata su tentativi per cercare la configurazione corretta di tali parametri, hanno incrementato i dubbi sul metodo descritto. In merito alla complessità, nel caso peggiore risulta pari a $O(n^2)$, ma potrebbe diventare $O(nlogn)$ se si usassero strutture dati con indici spaziali.

3.2.3 Normalizzazione dei dati

La normalizzazione effettuata in questa architettura ha trasformato i dati di interesse in valori compresi tra 0 e 1. Per fare ciò, esistono fondamentalmente due tecniche:

- **min-max**[16]: utilizzando questo metodo, i valori vengono normalizzati cosicché ricadono in un intervallo definito $[\min_{new}; \max_{new}]$.

Dato un attributo V, l'elemento v di V viene trasformato in v' secondo la seguente formula:

$$v' = \frac{v - min_{\mathrm{V}}}{max_{\mathrm{V}} - min_{\mathrm{V}}}(max_{\mathrm{new}} - min_{\mathrm{new}}) + min_{\mathrm{new}}$$

Questo metodo risulta essere molto sensibile agli outlier, oltre a richiedere chiaramente la conoscenza del valore minimo e massimo del nuovo intervallo.

- **Z-score**[16]: questa tecnica, detta anche *standardizzazione*, fa sì che la distribuzione di una variabile abbia media pari a 0 e deviazione standard uguale a 1.
  Dato un attributo V, di cui è nota la media μ e la deviazione standard σ, l'elemento v di V viene trasformato in v' secondo la seguente formula:

$$v' = \frac{v - \mu}{\sigma}$$

  Questo metodo risulta essere meno sensibile agli outlier e non richiede la conoscenza del valore minimo e massimo del nuovo intervallo.

3.3 Analisi dei *cluster*

Per poter individuare gli edifici che possiedono caratteristiche termo-fisiche simili sono state utilizzate tecniche di clustering, le quali per definizione aggregano dati con caratteristiche comuni. L'applicazione di tali algoritmi richiede che i dati siano normalizzati, come specificato nella sezione precedente, e poi elaborati. Nel nostro lavoro è stato utilizzato come algoritmo di clustering il K-means, descritto di seguito.

3.3.1 Algoritmo K-means

L'algoritmo k-means è stato sviluppato nel '67 da MacQueen. Questo algoritmo fa parte della famiglia degli algoritmi partizionali e non supervisionati e ha come scopo principale quello di minimizzare la distanza tra i punti di uno stesso cluster, cioè la distanza intra-cluster.
In questa tecnica, i cluster vengono considerati *center-based*, ovvero il centro di un cluster è il *centroide*, che rappresenta la media di tutti i punti del cluster, o il *medioide*, che rappresenta il punto più rappresentativo del cluster[17].
Nel K-means ogni cluster viene associato ad un centroide e il numero dei cluster K deve essere noto a priori.

Lo sviluppo dell'algoritmo segue un processo iterativo per cui si scelgono K punti casuali che rappresentano i centroidi iniziali dei cluster; a questo punto si assegnano i punti restanti al cluster del centroide più vicino, utilizzando il criterio di similarità stabilito, e si ricalcolano i centroidi dei cluster definiti come media dei punti del cluster. Questo procedimento viene rieseguito finché i centroidi non si spostano più, cioè finché tutte le istanze non sono assegnate allo stesso cluster per due o più iterazioni successive.
Poiché l'insieme iniziale dei centroidi è scelto in maniera *random*, i cluster che vengono generati variano ad ogni esecuzione, producendo risultati differenti ogni volta(vedi fig. 3.5 e fig. 3.6).

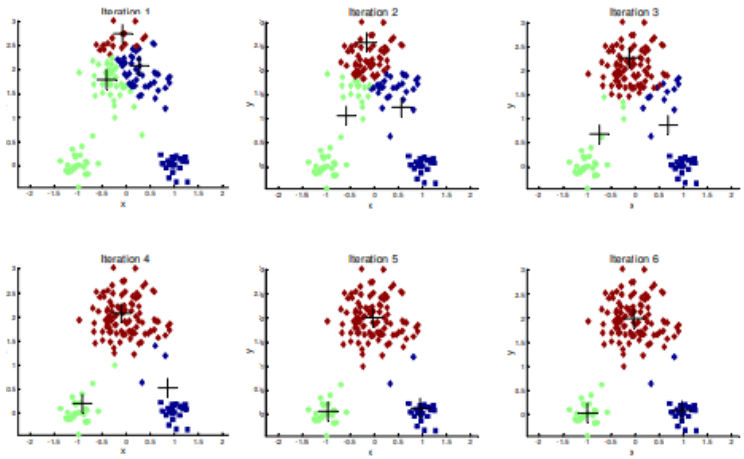

Figura 3.5: Esempio1. Importanza della scelta dei centroidi.© Tan,Steinbach,Kumar

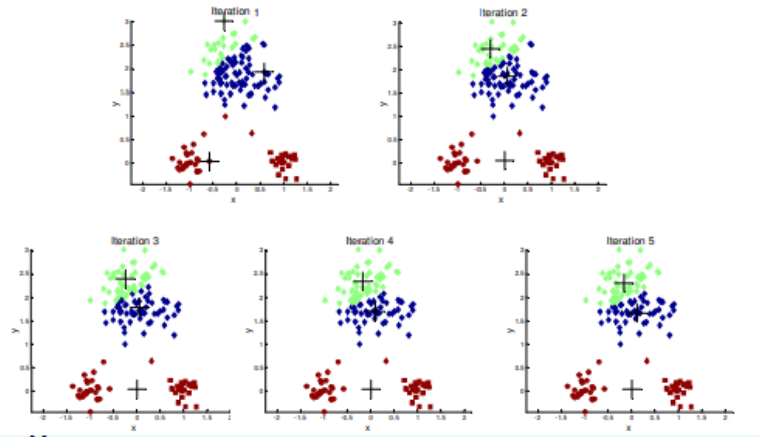

Figura 3.6: Esempio2. Importanza della scelta dei centroidi.© Tan,Steinbach,Kumar

Il K-means tendenzialmente converge già nelle prime iterazioni, infatti spesso come

condizione di fine si impone che non vengano più ricalcolati i centroidi se i punti riassegnati sono relativamente pochi.
Tuttavia non è detto che la convergenza sia ottimale poiché ciò dipende dalla scelta iniziale dei cluster e spesso la soluzione trovata risulta sub-ottima, come è possibile notare nella figura 3.7.

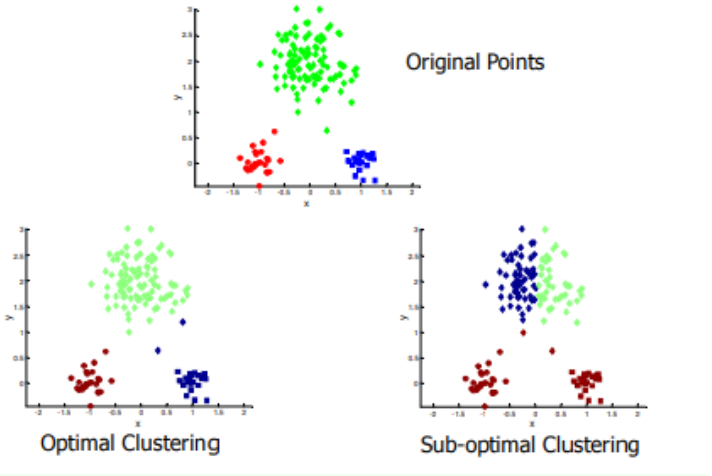

Figura 3.7: Due diversi clustering ottenuti con il K-means.© Tan,Steinbach,Kumar

La complessità dell'algoritmo è $O(n * K * I * d)$, dove n è il numero di punti, K è il numero di cluster, I è il numero di iterazioni e d è il numero di attributi[17].
Poiché il numero di iterazioni e di cluster è di solito molto minore del numero di oggetti da considerare, uno dei punti di forza di questo algoritmo è sicuramente il fatto di essere relativamente efficiente. Certamente, il calcolo diventa computazionalmente molto più oneroso con l'aumentare del numero di attributi da valutare.

Descrizione formale del K-means.

Dato un problema di clustering, i dati di partenza necessari sono:

- L'insieme di partenza dei dati $P \subseteq R^{\mathrm{n}}$

  $$P = p^{1}, ..., p^{\mathrm{m}}$$

  costituito da m vettori.

- Una funzione $D : R^{\mathrm{n}} \times R^{\mathrm{n}} \rightarrow R$ che rappresenti una misura di similarità tra due vettori.

Questa misura dipende dalla natura del problema e , in particolare, dai dati trattati. Tuttavia, risulta essere una scelta molto comune l'uso del quadrato della norma euclidea. Per tanto, se venisse usata questa metrica, dati due vettori $x,y \in R^{n}$, la funzione che calcola la similarità dei due vettori sarà la seguente:

$$D(x, y) = \| z - y \|^{2}$$

Il K-means è un problema di clustering e, in quanto tale, ha l'obiettivo di partizionare l'insieme P in k sottoinsiemi $P^{i} \subseteq P$, con i=1,...,k, tali da contenere vettori più simili possibili secondo la funzione di similarità *D(x,y)* precedentemente definita. Questi sottoinsiemi generati P^{i} devono godere delle seguenti proprietà:

1. $\bigcup_{i=1}^{k} P^{i} = P$
2. $P^{i} \cap P^{j} = \emptyset, \forall i \neq j$
3. $\emptyset \subset P^{i} \subset P$

Per formalizzare matematicamente il problema, occorre associare ad ogni sottoinsieme P^{i}, con $i = 1, ..., k$ un vettore $x^{i} \in R^{n}$ che rappresenterà il centroide del cluster i. A questo punto, è possibile definire i sottoinsiemi P^{i} associando ogni punto p^{j} al centro di cluster x^{i} più vicino, secondo il valore determinato dalla funzione $D(p^{j}, x^{i})$. A questo punto, il problema diventa scegliere quali sono i cluster i cui centroidi x^{i} sono tali per cui le distanze dei punti p^{j} dai centri più vicini sia la minore possibile. La funzione obiettivo che sintetizza tale problema è la seguente:

$$min_{\,x^{1},...,x^{k}}(\sum_{j=1}^{m} (min_{i=1,...,k} D(p^{j}, x^{i})))$$

La condizione di convergenza può essere:

- Il valore della funzione obiettivo non varia per due iterazioni consecutive
- I centroidi non variano.

Valutazione della bontà dei cluster prodotti.

Una volta prodotti i cluster occorre verificarne la bontà. Una delle misure che viene maggiormente utilizzata per questo scopo è lo scarto quadratico medio, noto anche come *SSE*.
Per ciascun punto p_i, l'errore corrisponde alla distanza dal centroide x_i del cluster C_i. Il valore di *SSE* sarà la somma dei quadrati degli errori, ovvero:

$$SSE = \sum_{i=i}^{K} \sum_{p \in C_\mathrm{i}} d(p_\mathrm{i}, x_\mathrm{i})^2$$

Poiché l'obiettivo è di minimizzare l'errore, un modo per ridurre il valore di *SSE* è quello di aumentare il numero dei cluster K, considerando che l'SSE ha un andamento decrescente; tuttavia un buon cluster con un valore di K piccolo può avere il valore di SSE minore rispetto a quello di un cluster cattivo ma che ha un valore K maggiore[17].
La scelta dei centroidi iniziali è quindi uno dei punti chiavi per ottenere una buona clusterizzazione dei dati e, per tanto, sono state proposte diverse soluzioni per ovviare a tale problema:

- Molteplici esecuzioni dell'algoritmo K-means utilizzando diversi centroidi di partenza.
- Effettuare un campionamento e usare tecniche di clustering gerarchico per trovare i K centroidi iniziali.
- Selezione di più centroidi iniziali e estrazione, tra questi, dei K da utilizzare, selezionando quelli che sono più separati.
- Fare uso di tecniche di *post-processing* con lo scopo di eliminare quei cluster che sono stati individuati in maniera sbagliata.

Un altro problema che si può presentare è quello che l' algoritmo generi dei cluster vuoti. Questo cluster potrebbe essere generato se al relativo centroide non venisse mai attribuito alcun elemento, e potrebbe determinare un valore di *SSE* molto elevato, poiché sarebbe come se un cluster non fosse stato utilizzato.
Le soluzioni che sono state individuate sono le seguenti:

- Scegliere un centroide alternativo, per esempio quello che ha più impatto sul valore di *SSE*
- Fare lo *split* di un cluster in due cluster che includono i punti più vicini, per esempio scegliendo l'elemento del cluster con il maggiore *SSE*.

Ciò che può inoltre influire sulla bontà dei cluster è la presenza di *outlier*, poiché i punti molto lontani hanno un impatto pesante sul valore del'*SSE* , essendo quest'ultimo un quadrato di distanza. Questa è un'ulteriore motivazione per cui è stata approfondita precedentemente la fase di *pre-processing*. Le potenziali limitazioni del K-means riguardano anche caratteristiche intrinseche dei cluster. Ad esempio, la scelta del valore K sulla base del valore di *SSE*, determina una scelta dei centroidi in maniera tale che tutti i cluster abbiano più o meno le stesse dimensioni; questo

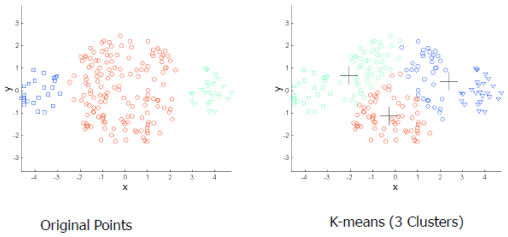

Figura 3.8: Cluster di dimensioni diverse.© Tan,Steinbach,Kumar

comportamento risulta essere un problema nel caso i cluster abbiano dimensioni diverse, come è possibile vedere in figura 3.11.
Inoltre, se un cluster è più denso, allora la distanza tra i punti di questo cluster è minore; pertanto, le zone con una densità inferiore necessiteranno di un numero maggiore di mediani per minimizzare il valore di *SSE*. La situazione appena descritta è rappresentata graficamente in figura 3.9.
Infine, anche la forma dei cluster influisce nel risultato finale, poichè l'*SSE* non tiene conto della forma degli oggetti. Se i cluster presentano delle forme non globulari, si possono verificare clusterizzazioni come in figura 3.10.

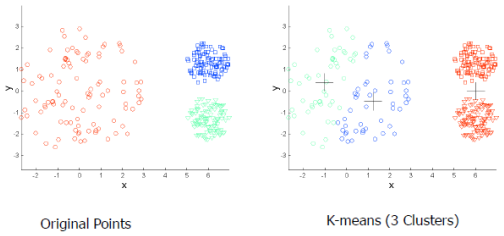

Figura 3.9: Cluster di densità diverse.
© Tan,Steinbach,Kumar

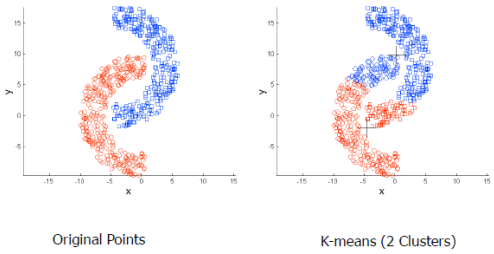

Figura 3.10: Cluster di forme diverse.
© Tan,Steinbach,Kumar

Una possibile soluzione è quella di aumentare il valore di K per trovare più cluster e poi definire una tecnica che riassembli i cluster scomposti.

Definizione del parametro *K*

Dalla descrizione finora fatta è evidente che determinare un corretto numero di cluster in cui partizionare il dataset risulta fondamentale per l'esito positivo dell'algoritmo.
A tale scopo si fa uso del cosiddetto *Elbow method*[17] che individua quale valore di

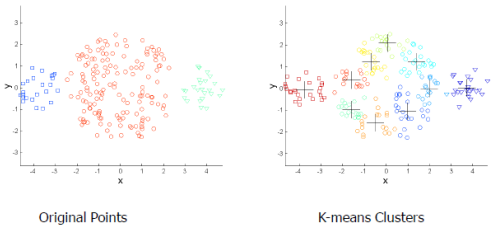

Figura 3.11: Possibile soluzione per cluster di dimensioni diverse.© Tan,Steinbach,Kumar

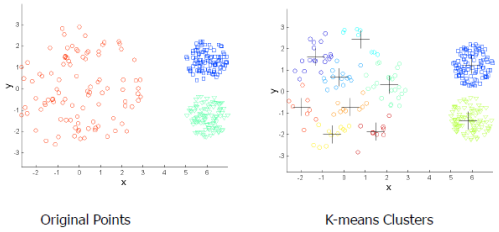

Figura 3.12: Possibile soluzione per cluster di densità diverse.
© Tan,Steinbach,Kumar

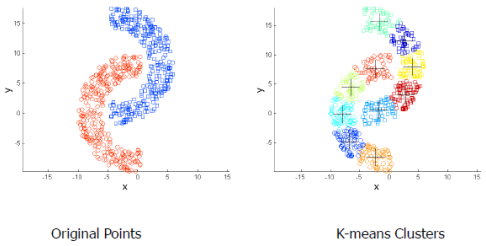

Figura 3.13: Possibile soluzione per cluster di forme diverse.
© Tan,Steinbach,Kumar

K risulta essere più indicato.
Per fare ciò, viene mostrato graficamente l'andamento della metrica usata per dedurre la bontà dei cluster, nel nostro caso *SSE*, al variare del numero di cluster K. L'obiettivo è quello di individuare qual è il punto in cui l'aumento di K causerà una diminuzione molto piccola del valore di *SSE*, mentre la diminuzione di K aumenterà bruscamente tale valore.
Quando la curva presenta un angolo 'a gomito', significa che, dopo il primo momento in cui la distanza dei punti dai corrispondenti centroidi è molto alta, diminuisce progressivamente fino a raggiungere una condizione in cui i miglioramenti ottenibili sono trascurabili e, quindi, si è ottenuto il valore ideale di K.

Decomposizione ai valori singolari

La *decomposizione ai valori singolari* è una tecnica dell'algebra lineare che effettua una fattorizzazione di una matrice attraverso l'utilizzo di autovettori e autovalori. Questa tecnica, detta anche *SVD*, afferma che data un matrice A reale o complessa di dimensioni $m \times n$, allora esiste una matrice $U \in O(m)$ e una matrice $V \in O(n)$, tali che:

$$U^{\mathrm{T}}AV = \sum \text{ , ovvero } A = U \sum V^{\mathrm{T}},$$

con la matrice diagonale $\sum$ che ha gli elementi

$$o_{ij} = \begin{cases} 0 & se \quad i \neq j, \\ o_i & se \quad i = j \end{cases}$$

con $o_1 \geq o_2 \geq ... \geq o_r \geq o_{r+1} = ... = o_p = 0$, $p = min(m, n)$.
E' una tecnica molto usata per la riduzione di dimensionalità; in particolare, quando si ha un dataset con un numero molto grande di colonne si può ridurre il numero di questi attributi pur conservando le caratteristiche più significative del problema.

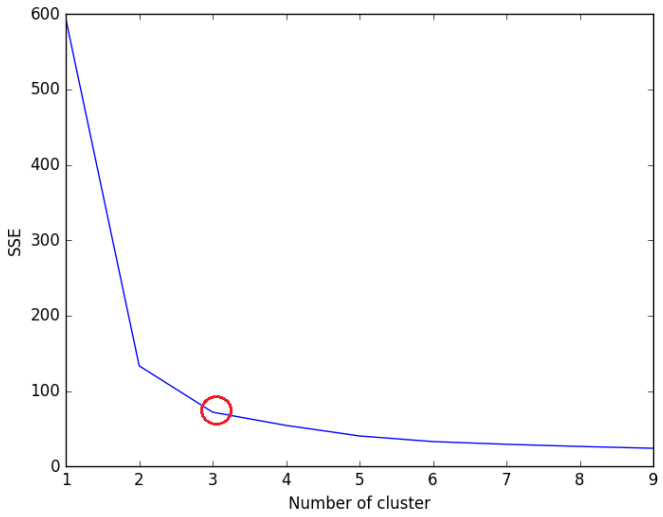

Figura 3.14: Esempio di *elbow graph* in cui il valore ideale è K = 3

3.3.2 Regole di associazione

Come descritto nella sezione 2.2.2, successivamente alla realizzazione dei cluster che raggruppano gli edifici in base alle proprietà termo-fisiche, risulta molto interessante estrarre ulteriore conoscenza da ogni cluster attraverso le regole di associazione. Riprendiamo alcuni concetti descritti precedentemente. Una regola di associazione viene definita nella seguente forma:

$$A \Rightarrow B$$

ed esprime la coesistenza tra gli *itemset* A e B, definiti rispettivamente *antecedente* e *conseguente*. Allo scopo di valutare nel miglior modo possibile le regole estratte,

esistono delle misure di interesse che, tenendo conto di ciascun itemeset coinvolto e del dataset completo, permettono di eliminare regole inutili.
Il *supporto* è definito come la probabilità che l'antecedente e il conseguente siano nella stessa transazione, ovvero esprime la frequenza della regola.

$$sup(A \Rightarrow B) = P(A \cap B)$$

La confidenza è definita come la percentuale di transazioni che contengono B tra quelle che contengono A, ovvero esprime una probabilità condizionata.

$$conf(A \Rightarrow B) = \frac{sup(A,B)}{sup(A)} = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

Queste due metriche presentano forti limiti ed è per questo motivo che per valutare la bontà di una regola associativa sono state introdotte altre metriche.
Il *lift* nasce come risposta ai limiti della confidenza. Infatti, nel calcolo della confidenza non viene considerato il supporto del conseguente e quindi se ci fossero gruppi di item che non sono stocasticamente indipendenti la valutazione della regola sarebbe scorretta. Definiamo due eventi A e B stocasticamente indipendenti se

$$P(A|B) \quad = \quad P(B|A)$$

cioè se la probabilità che si verifichi l'evento A non modifica quella che si verifichi B e viceversa. Possiamo allora definire il *lift* nel seguente modo:

$$lift(A \Rightarrow B) = \frac{conf(A,B)}{sup(B)} = \frac{P(B|A)}{P(B)} = \frac{P(B,A)}{P(A)} \frac{1}{P(B)}$$

Il lift indica perciò in che modo l'occorrenza di un evento è in grado di far aumentare l'occorrenza dell'altro. Se il valore del lift sarà uguale a 1, gli eventi A e B saranno indipendenti, mentre se il lift sarà maggiore di 1 vorrà dire che gli eventi hanno una correlazione positiva e cioè che la probabilità che si verifichi B è maggiore se è noto che si è verificato A. Una correlazione negativa si verificherà se il lift assumerà un valore inferiore a 1 e vorrà dire che quella regola non è molto interessante.
La *convinction* è una metrica di valutazione delle regole che, come il *lift*, cerca di superare alcuni limiti del *supporto* e della *confidenza*, ma a differenza dello stesso è più sensibile alla direzione della regola[3]. Definiamo la *conviction* nel seguente modo:

$$conviction(A \Rightarrow B) = \frac{supp(A)supp(\neg B}{sup(A,\neg B)} = \frac{P(A)P(\neg B)}{P(A,\neg B)}$$

Questa metrica cerca di dare quindi un peso alla direzione dell'implicazione della regola. Se il valore della conviction sarà compreso tra 0 e 1, gli eventi A e B avranno una dipendenza negativa, mentre se tale valore sarà maggiore di 1, allora la dipendenza avrà un significato positivo. Infine, se il valore di conviction sarà uguale a 1, i due eventi risulteranno indipendenti.

3.4 Validazione e visualizzazione della conoscenza

Allo scopo di dare una corretta interpretazione della conoscenza estratta finora è stato necessario mostrare i risultati ottenuti attraverso tecniche di visualizzazione.
Nello specifico, poiché il dominio utilizzato tratta certificazioni energetiche di edifici geolocalizzati, è stato possibile introdurre l'utilizzo di mappe che dessero informazioni sull'efficienza energetica degli edifici, localizzando opportunamente le aree di interesse.
Per tanto, oltre ai più comuni metodi di visualizzazione della conoscenza, come grafici a torta o distribuzioni dei valori degli attributi di interesse, sono state realizzate mappe a diverso livello di aggregazione. Nel dettaglio, è stato possibile realizzare delle mappe che esprimessero le informazioni energetiche della città considerata, ma anche di specifiche aree metropolitane. Per esempio, è possibile distinguere i consumi energetici per circoscrizione, per isolato o per unità abitativa.

Mappe coropletiche

Le *mappe coropletiche*, create nel 1826 da Charles Dupin[5], sono mappe tematiche in cui le aree che costituiscono la zona delimitata sono colorate in funzione della misurazione della variabile visualizzata sulla mappa attraverso l'uso di scale cromatiche. Questo genere di mappa è in grado di visualizzare in maniera molto semplice come una misura vari in un'area geografica o all'interno di una regione (Fig 3.15).
Nelle nostre analisi il numero di certificazioni disponibili non è il medesimo per ogni circoscrizione o isolato e, poiché in queste mappe ogni area è colorata con un unico colore rappresentativo di essa, occorre trovare una metrica che possa esprimere l'informazione richiesta tenendo conto del numero di certificati da cui è stata estratta.
E' evidente che queste mappe sono indicate qualora le regioni delimitate sono importanti per la discussione; nel nostro caso, sono state usate per quelle mappe che vogliono fornire informazioni circa una particolare circoscrizione o uno specifico isolato.

Mappe *scatter*

Le mappe *scatter* servono per visualizzare la distribuzione nello spazio di un certo attributo attraverso una collezione di simboli grafici, come è possibile vedere nella figura 3.16. Nelle mappe *scatter*, per ogni certificato geolocalizzato, è possibile visualizzare il relativo attributo di interesse attraverso un simbolo (noto come *marker*) posizionato esattamente nel punto in cui è ubicato l'edifico corrispondente.
Dato il consistente numero di certificati, questa mappa non viene usata tanto per avere una visione d'insieme dell'informazione, quanto per analizzare nel dettaglio gli

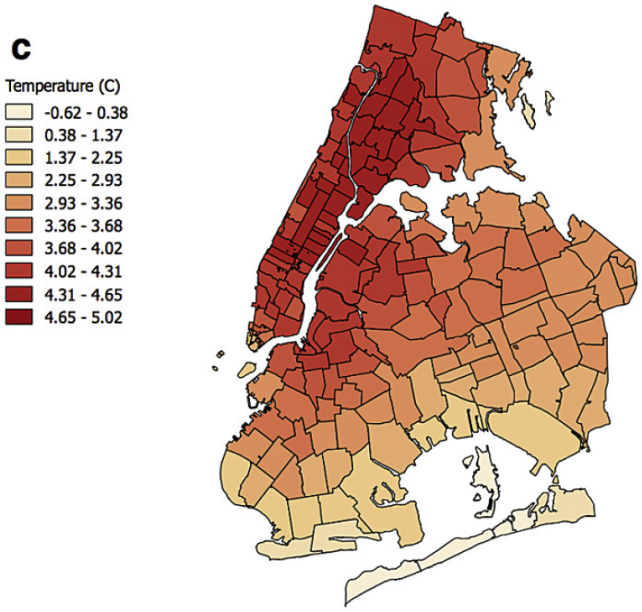

Figura 3.15: Esempio di mappa coropletica. ©Olivo, Hamidi, Ramamurthy

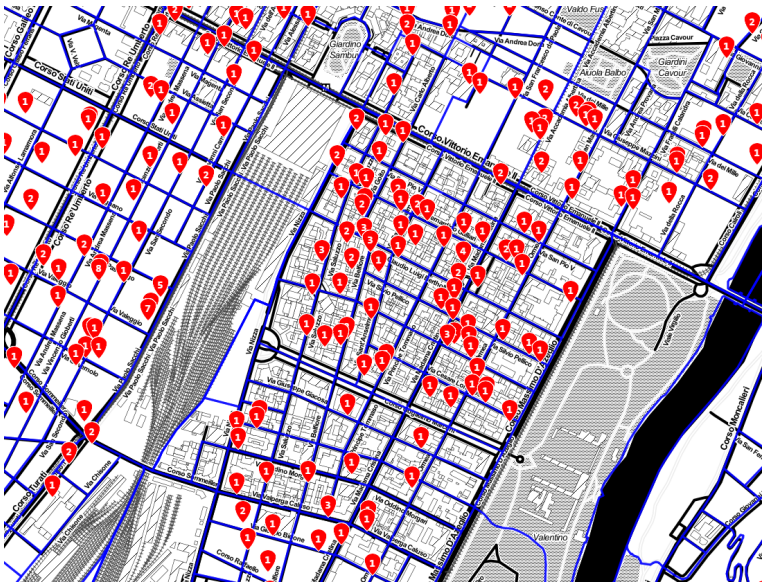

Figura 3.16: Esempio di mappa *scatter*.

attributi dei certificati relativi ad una certa circoscrizione o ad un certo isolato.

3.4.1 Realizzazione delle mappe

Nel presente lavoro, per rappresentare le caratteristiche energetiche degli edifici analizzati sulle mappe, si è utilizzata la libreria *folium* di Python che permette la visualizzazione di mappe generando dei file html. In particolare, sono state realizzate delle mappe coropletiche che usano come metrica per colorare le zone la media dei relativi valori. Con questo metodo tuttavia non viene completamente risolto il forte limite per cui le mappe coropletiche sono fortemente dipendenti dal numero di certificati presenti in quell'area; per tale ragione, ogni mappa coropletica è stata completata con l'introduzione dei *marker cluster*, presenti nella libreria *folium*. Questi marker, rispetto a quelli presenti nelle mappe scatter, sono in grado di mostrare informazioni aggregate. La suddetta libreria permette quindi di realizzare delle *dashboard*, mostrando delle mappe interattive che consentono di analizzare le aree geografiche a differenti livelli di dettaglio.
Per la realizzazione delle mappe coropletiche viene utilizzata la funzione di folium *Map.choropleth()*, che richiede un file *geojson* per definire i limiti delle aree da considerare (e.g. circoscrizioni, isolati,...). Inoltre, una volta specificato come parametro quale attributo si vuole mostrare, è possibile stabilire i valori limite per la scala di colori da utilizzare. La mappa coropletica che viene generata di *default* dalla libreria python, non utilizza la media dei valori per colorare ogni area; per questo motivo è stato necessario modificare la funzione *colorscalefun()* affinché il colore fosse assegnato sulla base del valore medio degli attributi presenti, invece che considerare solo l'ultimo valore letto per quell'area, come prevede il comportamento di default.
Per sovrapporre i *marker cluster* alla mappa coropletica, si è utilizzato un *plugin* della libreria folium, noto come *marker cluster*. Questi ultimi oggetti sono dei marker dinamici basati principalmente sulla cardinalità dei valori presenti nella zona di riferimento e che si aggregano o si separano a seconda del livello di zoom dell'utente, modificando il colore in base alla numerosità dei campioni nell'area considerata. La possibilità di modificare il comportamento custom dei marker è reso possibile grazie alla funzione *iconcreatefunction()* che viene passata nel costruttore dell'oggetto e che rende possibile personalizzare la dimensione e il colore del cluster modificando il codice html che permette la visualizzazione dei marker. Nel nostro caso tale funzione è stata modificata cosicché cambi dimensione in base al numero di valori present nell'area di riferimento del marker e cambi colore in base al valore medio dell'attributo.
Infine, sovrapponendo la mappa coropletica e i marker-cluster si può avere lo stesso contenuto informativo ad un livello più alto, ma si può comprendere meglio la distribuzione dei valori facendo lo zoom sull'area di interesse; dalla lettura della mappa allora sarà possibile comprendere il valore medio di ogni macro area vedendo il colore della mappa coropletica e capire a diversi livelli di dettaglio la cardinalità degli edifici considerati e i valori assunti. Nelle figure 3.17, 3.18 e 3.19 sono mostrati degli

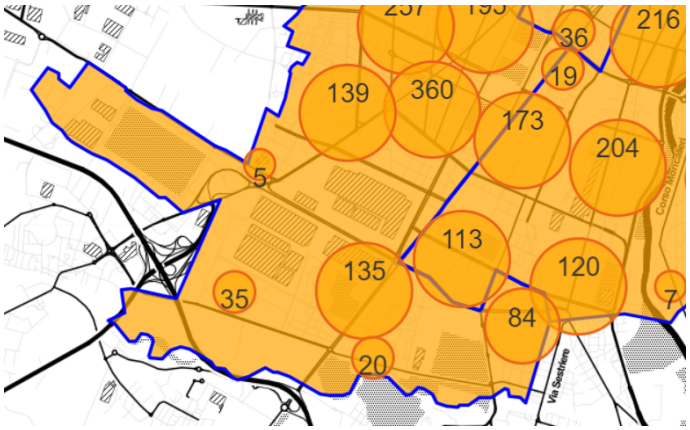

Figura 3.17: Esempio di mappa coropletica con marker-cluster.

esempi di mappe a differenti livelli di dettaglio.

Figura 3.18: Esempio di mappa coropletica con marker-cluster con un livello di zoom maggiore.

Figura 3.19: Esempio di mappa coropletica con marker-cluster con il massimo livello di zoom.

Capitolo 4

Risultati sperimentali

4.1 Raccolta e integrazione dei dati

Durante la prima fase del lavoro, è stata effettuata un'analisi esplorativa del dataset per cogliere gli aspetti più critici dei dati. Innanzitutto, si è proceduto con una fase di *preprocessing* eliminando gli attributi che presentavano lo stesso valore per tutti i record o che avevano circa il 70% di *missing values*. Da tale operazione sono stati eliminati circa 46 attributi, ottenendo un dataset con 70 attributi utilizzabili.
In seguito, è stata condotta un'analisi più approfondita con lo scopo di estrarre un sottoinsieme di record significativi sui quali sviluppare un modello di estrazione della conoscenza, estendibile successivamente a sottoinsiemi differenti. Analizzando le distribuzioni è emerso che circa il 90% dei certificati presenta come destinazione d'uso il valore **E.1(1)**, ovvero edifici residenziali a carattere continuativo, e come *oggettoAPE* la stringa 'Unità immobiliare'. Alla luce di tanto, per costituire il dataset di riferimento, sono stati estratti circa 30000 certificati riferiti ad unità immobiliari di carattere residenziale nella città di Torino.

4.1.1 Pulizia dei dati: algoritmo di correzione degli attributi per la geolocalizzazione

Nella presente tesi, i dati utilizzati e la conoscenza estratta saranno visualizzati su mappe per poter avere una comprensione degli stessi più intuitiva. Ciò richiede che gli attributi dei certificati necessari per la geolocalizzazione siano corretti e con un formato standard. Dalla figura 4.1 è possibile notare che circa un terzo del dataset, ovvero più di 9000 certificati, ha il CAP generico, e non più in vigore, della città di Torino 10100 e nella figura 4.2 sono mostrati i CAP più frequenti a seguito del processo di pulizia. Inoltre, nel tentativo di recuperare i CAP corretti, sono emerse

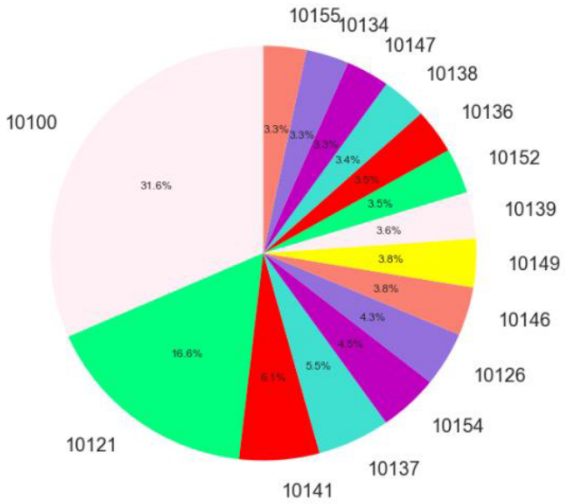

Figura 4.1: Distribuzione del numero di certificati per CAP più significativi nella città di Torino prima della pulizia.

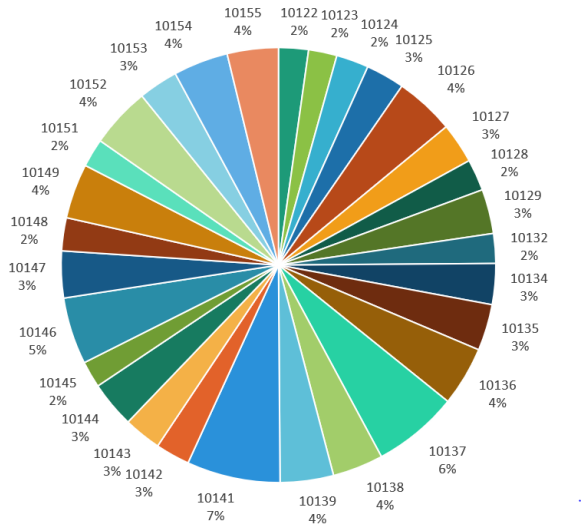

Figura 4.2: Distribuzione del numero di certificati per CAP nella città di Torino dopo la pulizia.

innumerevoli criticità nell'attributo relativo agli indirizzi, poiché sono presenti caratteri codificati erroneamente, errori di battitura e formati non standard. Per tale ragione, si è proceduto nello sviluppo di una tecnica che risolvesse tutte le criticità relative agli attributi geografici dei certificati. Come detto nel Capitolo 3, le tecniche utilizzate sono due: una che fa uso di un servizio web con un numero limitato di

richieste e una che si basa sul confronto degli indirizzi con il viario di Torino.
Nell'algoritmo sviluppato, in primo luogo viene usata la tecnica di confronto che fa uso della distanza di Levenshtein per verificare la similarità tra le stringhe che rappresentano gli indirizzi. Per effettuare questo confronto, il viario è stato memorizzato in una struttura dati che tiene conto di tutti i possibili modi in cui è possibile combinare le parole che costituiscono la via. Attraverso questa tecnica sono stati risolti circa il 97% degli indirizzi elaborati. Gli indirizzi che non sono stati risolti sono quelli per cui non è stato trovato un indice di similarità maggiore di 0.875. Dunque, questi indirizzi sono stati elaborati attraverso le API *geocoding* e solo il 7% è stato eliminato perché ritenuto invalido. In particolare, la metà dei certificati eliminati aveva un indirizzo inesistente nella città di Torino e l'altra metà aveva il corrispondente campo vuoto.
In figura 4.3 sono mostrati alcuni esempi di indirizzi esclusi dal procedimento che usa la distanza di Levenshtein e risolti da *geocoding*, ed è evidente che i limiti del primo metodo dipendono dalla lunghezza delle stringhe confrontate. Inoltre, è possibile notare che non è stato possibile manipolare il campo indirizzo per eliminare stringhe che non descrivono la via, tuttavia ciò non rappresenta un problema per la risoluzione con geocoding che invece è in grado di gestire tali anomalie. Attraverso

ORIGINAL	RESULTLEVENSHTEIN	VALUE	RESULTGEOCODING
VIA EANDI	VIA CANDIA	0.823	VIA ABATE VASSALLI EANDI
VIA RAFFAELLO MORGHEN	VIA RAFFAELLO LAMBRUSCHINI	0.774	VIA RAFFAELLO MORGHEN
VIA ASTI - EDIFICIO F	VIA CASTELDELFINO	0.687	VIA ASTI
VIA CAROSSIO	VIA CARISIO	0.857	STR DEL CAROSSIO
STRADA PROVINCIALE DI LANZO	STRADA S. VINCENZO	0.736	VIA LANZO
TORINO - VIA ASTI	VIA TRINO	0.761	VIA ASTI
VIA GAMBA	VIA GAMBASCA	0.842	CORSO ENRICO GAMBA
ROVIGOVIA ROVIGO	VIA ROVIGO	0.750	VIA ROVIGO

Figura 4.3: Esempio di indirizzi non risolti con Levenshtein, ma risolti con Geocoding.

questo algoritmo perciò sono stati corretti gli indirizzi di circa il 99.8% dei certificati e sono stati eliminati quelli ritenuti invalidi, cioè quelli che avevano errori di immissione durante la fase di compilazione del certificato non risolvibili.

4.1.2 Pulizia dei dati: *scaling*

La fase esplorativa iniziale ha permesso di individuare un' ulteriore aspetto da valutare nei dati. In particolare, è stato possibile notare che circa il 30% dei certificati aveva almeno un valore di rendimento con un ordine di grandezza maggiore del previsto, esprimendo quindi un valore percentuale. Pertanto è stato necessario scalare tali valori di un fattore 100. In questo modo, se si ha un valore di rendimento maggiore di 1 verrà scalato ma conserverà la sua appartenenza o meno al range di validità, anche se, in alcuni casi, non perché è maggiore del limite superiore ma perché è più piccolo del limite inferiore. Ad esempio, se un rendimento ha valore 1,65 e il relativo range di validità è [0,70; 1,00], allora tale valore dovrebbe essere considerato invalido perché maggiore del limite superiore; allo stesso modo anche il valore scalato, che è pari a 0,0165, viene considerato fuori dai range di validità, ma perché è inferiore a 0,70, limite inferiore del relativo intervallo di ammissibilità.

4.1.3 Pulizia dei dati: eliminazione degli outlier

Prima di procedere con la pulizia degli outlier, si è notato che nel dataset c'erano più certificati relativi allo stesso edificio; ciò è possibile perché, ad esempio, è necessario richiedere una nuova certificazione se l'edificio ha subito una ristrutturazione. E' stato necessario pertanto selezionare solo un APE per ogni immobile grazie a tre attributi che descrivono univocamente un' unità abitativa, ovvero gli attributi *particella, foglio e subalterno*, e selezionando tra i certificati estratti quello più recente. Questo processo ha permesso di selezionare circa l' 85% dei certificati del *database*. Grazie al supporto degli esperti di dominio, è stato possibile estrarre un set di attributi significativi per la determinazione delle prestazioni energetiche degli edifici. Quindi si è ritenuto opportuno procedere inizialmente con l'eliminazione degli *outliers* per i suddetti attributi seguendo un approccio principalmente *domain driven* allo scopo di avere un *dataset* di certificati dai quali estrarre una conoscenza attendibile; in seguito, per gli attributi non ritenuti fondamentali dagli esperti di dominio, ma ugualmente utilizzati nelle fasi successive, è stato seguito un approccio *data driven*, con il quale è stato possibile eliminare gli *outliers* applicando l'algoritmo DBSCAN.
Per la prima fase di pulizia, in supporto agli esperti di dominio, sono state applicate tre tecniche univariate di riconoscimento degli outlier e, per ognuna di esse, sono stati analizzati i range di validità estratti e la percentuale di valori ritenuti *outliers*. In figura 4.4 è possibile notare la percentuale di outliers che ognuno dei tre metodi proposti (*gESD, MAD e Boxplot*) è in grado di individuare per ogni attributo; da questa tabella emerge che la tecnica nota come MAD riesce ad individuare un numero più corposo di outlier, mentre il gESD è il metodo che riconosce un numero più esiguo di valori anomali. Analizzando più approfonditamente i range di validità

METODO	FATTORE FORMA	TRASMITTANZA OPACA	TRASMITTANZA TRASPARENTE	RENDIMENTO DISTRIBUZIONE	RENDIMENTO EMISSIONE	RENDIMENTO GENERAZIONE	RENDIMENTO REGOLAZIONE
GESD	0,29 %	0,14 %	0,01 %	1,62 %	2,43 %	1,63 %	1,75 %
MAD	2,69 %	1,23 %	0,02 %	42,45 %	41,91 %	14,79 %	41,73 %
BOXPLOT	2,06 %	3,50 %	0,05 %	1,62 %	2,43 %	14,92 %	1,75 %

Figura 4.4: Percentuale di valori ritenuti *outliers* per ogni metodo utilizzato.

estratti da ognuna delle tecniche sopracitate, come mostrato in figura 4.5, è stato possibile notare che i metodi di analisi individuano comportamenti differenti nella valutazione dei rendimenti. In particolare, è emerso che solamente il MAD è stato in grado di riconoscere correttamente gli outlier per i rendimenti; questo perché, sebbene i rendimenti ammettano un valore compreso tra 0 e 1 per definizione, in relazione al contesto trattato, non possono assumere valori troppo piccoli perché descriverebbero un impianto non funzionante che non produce energia. Per quanto riguarda invece il fattore forma e le trasmittanze, tutti e tre i metodi forniscono dei range di validità accettabili che permettono all'esperto di dominio di definire degli intervalli di ammissibilità coerenti con i limiti fisici e tali da preservare quasi la totalità dei certificati.

	GESD				MAD				BOXPLOT			
	Range di esclusione		Range di validità		Range di esclusione		Range di validità		Range di esclusione		Range di validità	
	MIN	*MAX*	*MIN*	*MAX*	*MIN*	*MAX*	*MIN*	*MAX*	*MIN*	*MAX*	*MIN*	*MAX*
FATTORE FORMA	1,50	42,96	0,00	1,47	0,97	42,96	0,00	0,97	1,03	42,96	0,00	1,03
TRASMITTANZA OPACA	3,05	5,65	0,00	3,03	0,00	5,65	0,02	2,35	0,00	5,65	0,29	2,08
TRASMITTANZA TRASPARENTE	25,01	25000,00	0,00	9,17	9,17	25000,00	0,00	8,98	7,97	25000,00	0,00	7,58
RENDIMENTO DISTRIBUZIONE	81,00	100,00	0,00	1,21	0,00	100,00	0,56	1,21	81,00	100,00	0,00	1,21
RENDIMENTO EMISSIONE	9,60	131,83	0,00	1,10	0,00	131,83	0,63	1,10	9,60	131,83	0,00	1,10
RENDIMENTO GENERAZIONE	11,42	299,89	0,00	8,78	1,78	299,89	0,00	1,76	1,64	299,89	0,00	1,62
RENDIMENTO REGOLAZIONE	3,43	100,00	0,00	2,16	0,00	100,00	0,57	1,12	3,43	100,00	0,00	2,16

Figura 4.5: Range di validità degli attributi estratti con *gESD, MAD e Boxplot.*

Una volta che sono stati definiti i vincoli fisici che i sette attributi fondamentali devono rispettare, si è proceduto applicando i suddetti filtri al dataset. L' approccio che è stato seguito è stato quello di applicare prima i filtri sul fattore forma e sulle trasmittanze, analizzando i risultati ottenuti, e poi applicare i filtri anche sui rendimenti e ripetere le analisi, soprattutto sui certificati eliminati, per comprendere la

natura dell'esclusione dei record. A seguito dell'applicazione del primo filtro sul fattore forma e sulle trasmittanze, è stato eliminato circa il 7,6% dei certificati a causa dell'invalidità di almeno un attributo dei tre precedentemente citati. Analizzando i certificati prima dell'applicazione del filtro, è emerso che circa il 99.6% del database soddisfaceva i limiti delle transimttanze (Fig 4.6); per tanto, si può dedurre che l'eliminazione dei certificati è stata dettata dal fattore forma. In seguito, sul dataset

Figura 4.6: Distribuzione dei valori delle trasmittanze nel dataset.

validato rispetto ai primi vincoli, sono stati applicati i filtri sui rendimenti e circa il 60% del dataset è stato eliminato. Dall'analisi dei certificati eliminati è emerso che circa il 60% dei valori del rendimento di regolazione, di distribuzione e di emissione è uguale a 0. Si è supposto che tale anomalia potesse dipendere da un arrotondamento di valori molto piccoli nel dataset originale o da errori di trasmissione dei sofware per la redazione degli APE. Il restante 40% invece è stato analizzato più approfonditamente per capire quale attributo ha determinato l'esclusione dei certificati. Nella figura 4.7, per questi certificati, sono mostrate le distribuzioni dei valori dei quattro rendimenti dopo lo scaling. In particolare viene mostrata sia la frequenza dei certificati per ogni intervallo di rendimento considerato sia la frequenza cumulata per

comprendere meglio in quale range di valori sono concentrati i rendimenti del maggior numero di certificati. Dalla suddetta figura è possibile notare che i rendimenti di distribuzione, di emissione e di regolazione hanno valori concentrati in un intervallo che è molto vicino a quello definito dall'esperto di dominio, mentre il rendimento di generazione presenta un andamento molto più distribuito, presentando circa il 50% dei valori tra 0,2 e 0,7, intervallo ritenuto invalido dagli esperti. E' stato possibile dunque affermare che il 40% dei certificati è stato eliminato a causa del rendimento di generazione. Questa anomalia potrebbe essere dettata da una incomprensione nel-

Figura 4.7: Distribuzione dei valori dei rendimenti diversi da 0 nel dataset.

la fase di compilazione del certificato, poiché il rendimento di generazione potrebbe esprimere anche il valore del COP, se relativo alle pompe di calore come impianto; in quel caso, valori dell' ordine di grandezza della decina sarebbero ritenuti ammissibili. Ciò trova conferma nei risultati di un'ulteriore analisi condotta solo sui valori del rendimento di generazione che hanno causato l'eliminazione dell' APE e da cui è emerso che circa il 90% di tali valori ha un valore compreso tra 5 e 10 (cioè 0.05 e 0.1 dopo lo scaling), come mostrato in figura 4.8. Tuttavia, dalle informazioni a nostra disposizione non è stato possibile confermare tale assunzione, per tanto i relativi certificati sono stati ritenuti invalidi.

4.2 Analisi dei *cluster*

La fase di *clustering* è la fase più importante del processo di estrazione della conoscenza poiché consente di individuare gruppi di certificati che condividono le stesse

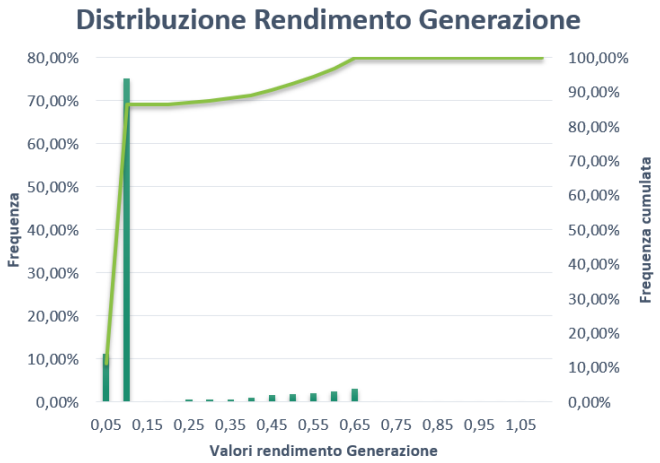

Figura 4.8: Distribuzione dei valori del rendimento di generazione che hanno causato l'eliminazione del relativo certificato.

caratteristiche, da ognuno dei quali è possibile estrarre conoscenza utile. Ovviamente, i raggruppamenti non possono essere realizzati utilizzando tutti gli attributi disponibili e, per tale ragione, sono stati selezionati dall'esperto di dominio un sottoinsieme di attributi che maggiormente influenzano le prestazioni energetiche di un edificio da un punto di vista fisico-tecnico. Gli attributi utilizzati sono:

- il fattore forma,
- il rendimento medio globale dell'impianto ottenuto dal prodotto del rendimento di emissione, distribuzione, regolazione e generazione,
- la trasmittanza media delle superfici opache,
- la trasmittanza media delle superfici trasparenti,
- la superficie riscaldata.

Su queste cinque variabili è stato applicato un processo di normalizzazione allo scopo di eliminare differenze numeriche tra variabili. In particolare, è stato ritenuto necessario applicare questa tecnica poiché le variabili in questione presentano ordini di grandezza differenti tra di loro; in particolare, le trasmittanze presentano valori di un ordine di grandezza superiore rispetto ai rendimenti e gli attributi che esprimono le superfici degli edifici assumono valori anche dell'ordine delle centinaia. Tra i vari metodi presenti in letteratura, è stato utilizzato il metodo *min-max*. Tale metodo risulta essere sensibile alla presenza di outlier ma, nel nostro caso, tale limite non sussiste poiché le variabili utilizzate sono state già processato con tecniche di *outlier*

detection. In dettaglio, il fattore forma, le trasmittanze ed i rendimenti sono stati filtrati, applicando i limiti definiti attraverso un approccio *domain driven*, mentre i restanti campi sono stati puliti attraverso l'uso del DBSCAN.

4.2.1 Algoritmo K-means

Una volta selezionati gli attributi più significativi ed elaborati per poterli utilizzare, è stato applicato l'algoritmo Kmeans per la determinazione dei cluster. Questo algoritmo richiede, come parametro per poter convergere, il numero di cluster K in cui raggruppare le certificazioni. Questa scelta risulta fondamentale per poter estrarre una buona conoscenza e, per questo motivo, è stato utilizzato lo scarto quadratico medio, noto come SSE, quale metro di valutazione della bontà di un cluster. In particolare, sono stati calcolati i valori del SSE per valori di K compresi tra 2 e 30 e sono stati rappresentati graficamente mostrando il cosiddetto *Elbow graph*. Dal punto di vista teorico il numero ottimale K di cluster è quello per cui, utilizzando K+1 si avrebbe un decremento ridotto di SSE e utilizzandone K-1 si avrebbe un aumento rilevante di SSE. Questo valore dovrebbe corrispondere nel grafico al punto in cui la curva mostra un gomito. Dalla figura 4.9 è evidente che nel nostro *elbow graph* non sia presente un angolo a gomito evidente, ma piuttosto è possibile individuare una zona in cui scegliere il K che va dal valore 4 a 12. Questo può succedere perché non è detto che ci sia un esatto valore in cui si manifesta il comportamento descritto. In questi casi, viene evidenziata un'area di interesse e si procede con prove sperimentali che consentono di scegliere il valore adeguato. Nel range individuato tra 4

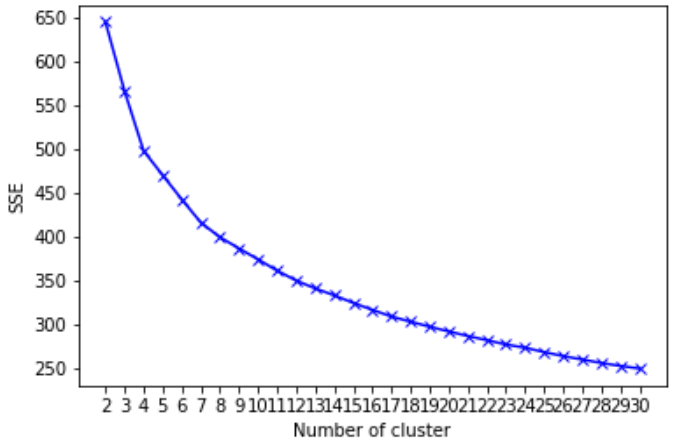

Figura 4.9: Elbow graph per valori di K da 2 a 30.

e 12, sono stati selezionati i valori di $K = 4$, $K = 7$ e $K = 9$ per effettuare le prove

sperimentali, poiché sembravano i valori più indicati per riuscire a comprendere i macro comportamenti da estrarre. Nelle figure 4.10, 4.11 e 4.12 vengono mostrati gli andamenti dei centroidi per gli attributi utilizzati per il clustering; è possibile notare che, per tutti gli esperimenti, gli attributi con la maggiore variabilità tra i cluster sono il rendimento medio globale (ETAH) e la trasmittanza trasparente. Tuttavia, dai risultati ottenuti con i valori di K pari a 7 e 9, emerge che ci sono cluster in cui, oltre ai due attributi citati precedentemente, anche il fattore forma e, nel caso con K uguale a 9, la superficie riscaldata influiscono nella determinazione del cluster di appartenenza. Come metro di valutazione per stabilire la bontà di un cluster nell'i-

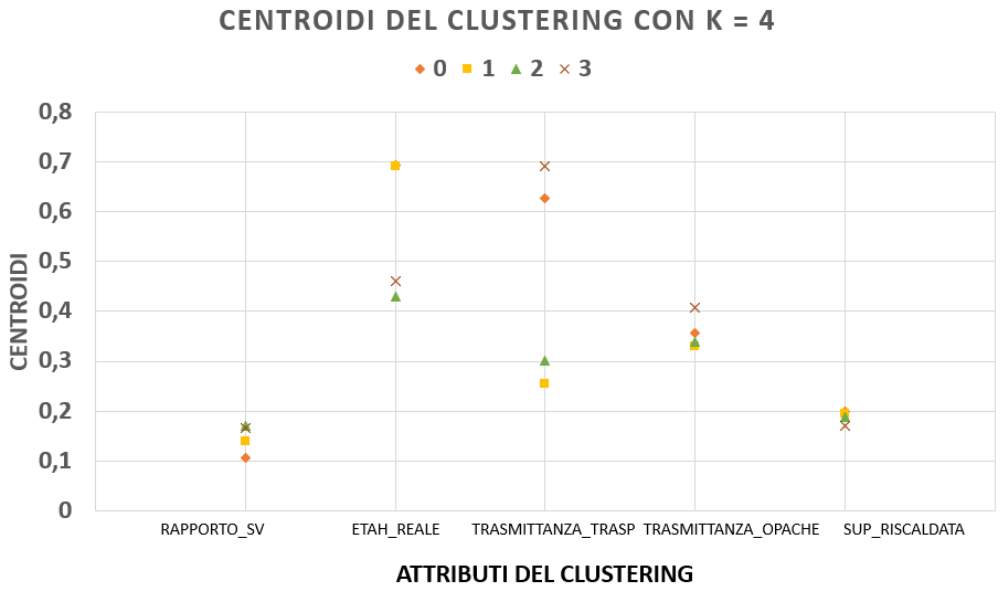

Figura 4.10: Grafico dei centroidi degli attributi utilizzati per il clustering con K=4.

solare certificati con determinate caratteristiche da quelli con altre, si è utilizzato il valore dell' EP_H, cioè l'indice di prestazione energetica per la climatizzazione invernale; si è scelto tale valore perché dà un'indicazione in merito all'efficienza energetica di un edificio ed, in particolare, per quanto riguarda il riscaldamento invernale che è il servizio più dispendioso in termini di consumi energetici. Per una migliore comprensione della relazione tra il valore assunto dalla variabile EP_H e la qualità della performance energetica in merito al relativo edificio, sono stati definiti dall'esperto di dominio dei limiti che ci consentono di definire la qualità energetica dell'edificio; in particolare:

- $0 \leq EP_h \leq 50 \Rightarrow$ *high*, edificio performante

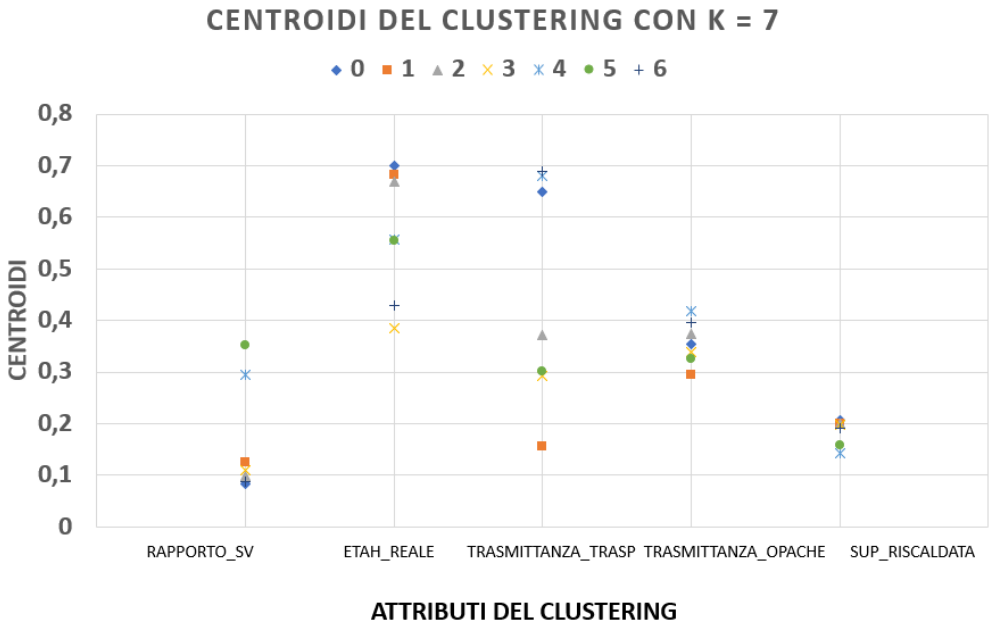

Figura 4.11: Grafico dei centroidi degli attributi utilizzati per il clustering con K=7.

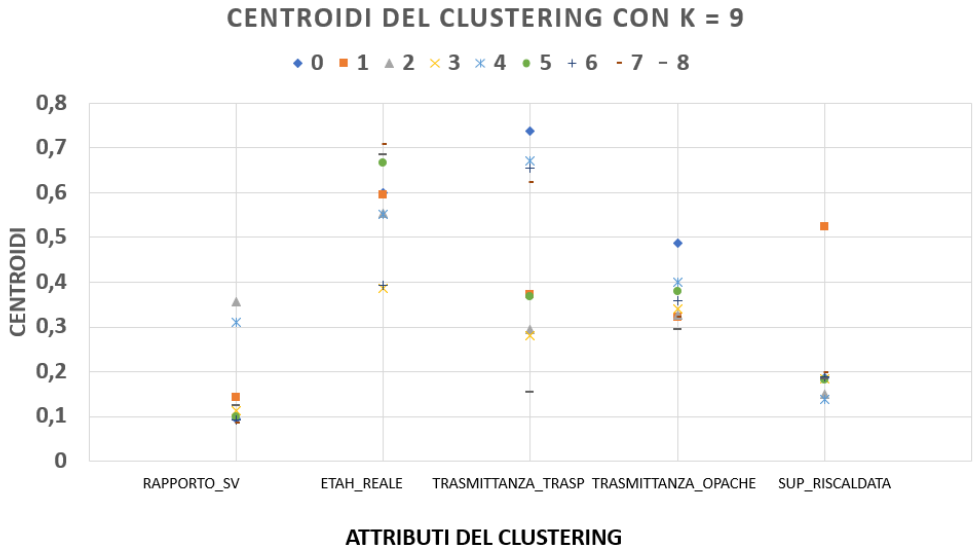

Figura 4.12: Grafico dei centroidi degli attributi utilizzati per il clustering con K=9.

- $50 < EP_h \leq 100 \Rightarrow$ *medium*, edificio poco performante
- $EP_h > 100 \Rightarrow$ *low*, edificio non performante

L'obiettivo è quello di individuare dei cluster che riescano a isolare bene gli edifici con i valori di EP_H in un certo range. Dalle figure 4.13, 4.14 e 4.15 è possibile notare che, sebbene si riescano ad ottenere alcuni cluster con uno dei tre intervalli di EP_H predominante, sono presenti molti cluster con l'intervallo *medium* presente per almeno il 40% dei record. A partire da queste considerazioni, si è ritenuto opportuno

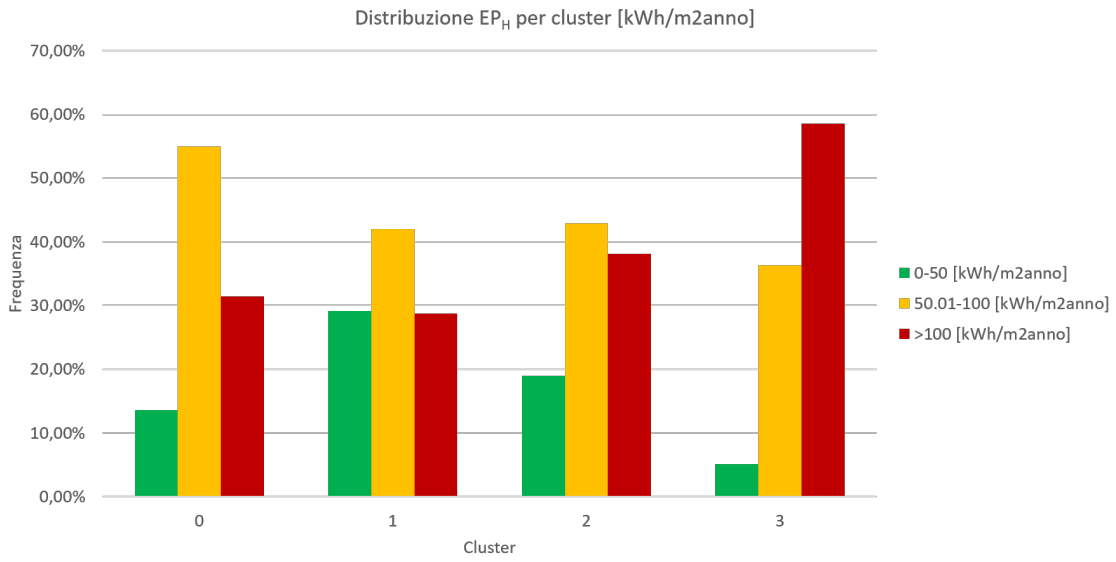

Figura 4.13: Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=4.

utilizzare un approccio leggermente differente. Si è pensato di applicare l'algoritmo *Kmeans* per generare un numero abbastanza grande di cluster, così da individuare più gruppi con caratteristiche singolari e, in seguito, riaggregarli utilizzando una combinazione di metriche. Per stabilire quali cluster riaggregare viene calcolata la distanza euclidea tra i centroidi di tutti i cluster generati e vengono selezionati i due cluster più vicini. Come indice per determinare se la riaggregazione dei due cluster scelti con il metodo precedente è la scelta migliore, è stato utilizzato l'indice di *Silhouette*[14], poiché consente di capire quanto un campione sia stato assegnato al cluster corretto. La *Sihouette* viene calcolata per ogni campione, confrontandone la distanza dal centroide del cluster a cui appartiene e la distanza dal centroide del cluster più vicino. Tuttavia, è necessario avere un'idea globale circa la corretta appartenenza dei campioni ad un certo cluster e, pertanto, è stata calcolata per ogni cluster la media dei valori di *Silhouette* dei campioni appartenenti a quel custer. In questo modo, una volta che sono stati scelti due cluster attraverso la minima

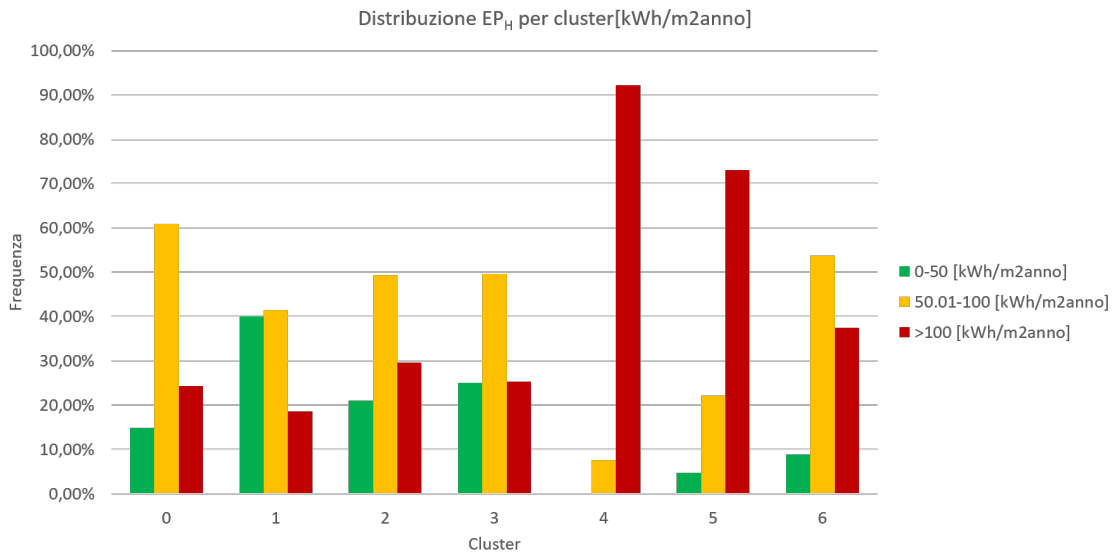

Figura 4.14: Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=7.

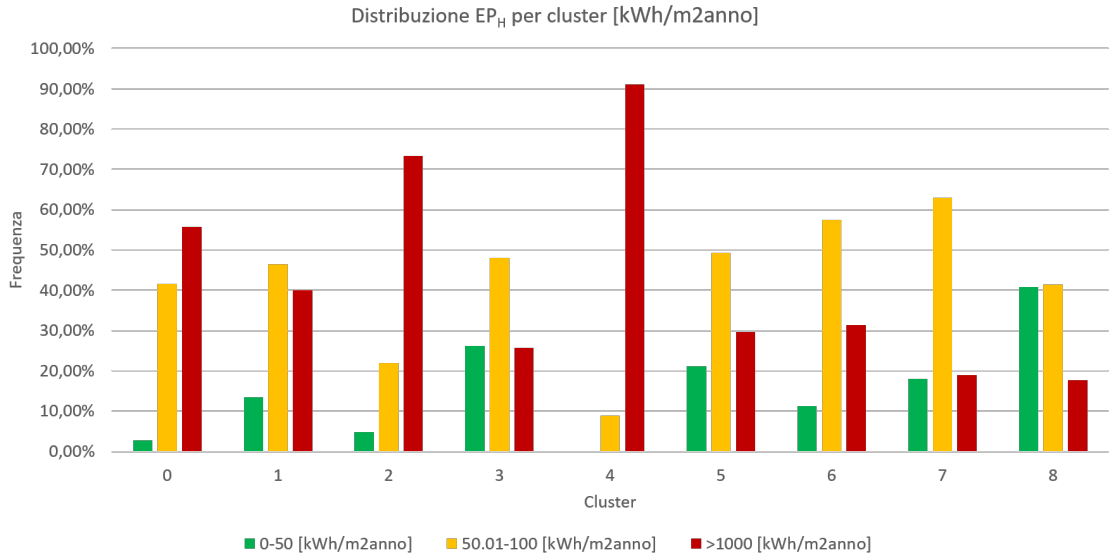

Figura 4.15: Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=9.

distanza euclidea, è possibile confermare tale scelta, valutando il valore medio di *Silhouette* per quei cluster, e reiterare, scegliendo altri due cluster, qualora la scelta non fosse approvata.
Applicando questo metodo al nostro dataset, si è deciso di partire dal clustering con K uguale a 12 e si è proceduto con la riaggregazione fino ad ottenere 4 cluster.

Analizziamo i 4 cluster ottenuti da tale processo.
In figura 4.16, viene mostrata la cardinalità di ogni cluster ed è possibile notare che, rispetto al clustering ottenuto dal Kmeans con K uguale a 4, i cluster hanno diversa cardinalità; i cluster più piccoli sono quelli che sono riusciti maggiormente ad isolare gli edifici molto performanti e poco performanti, mentre i cluster con maggiore cardinalità sono quelli che hanno una composizione molto equilibrata in termini di efficienza energetica. Dalla figura 4.17 è possibile notare che la variabile

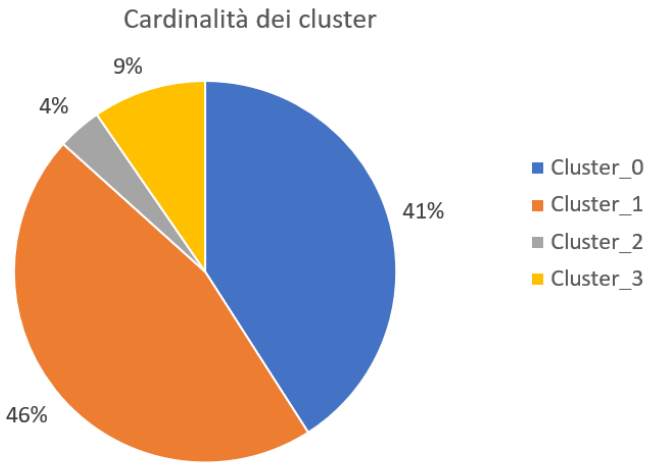

Figura 4.16: Distribuzione dei certificati per i 4 cluster.

meno significativa per il clustering continua ad essere la trasmittanza opaca, mentre acquisiscono importanza il fattore forma e la superficie riscaldata. Di seguito verranno riportati i *boxplot* di ogni attributo utilizzato per ogni cluster in modo da poter fare un confronto immediato tra i vari gruppi sulla stessa variabile. Dalla figura 4.18 è possibile notare che i valori dei rendimenti si distribuiscono in maniera molto simile per i primi tre cluster, mentre l'ultimo cluster isola gli edifici con i rendimenti migliori, come è possibile vedere dal relativo boxplot che presenta poca varianza ed un valore della mediana maggiore. Tra gli attributi utilizzati, è evidente che la trasmittanza trasparente sia quella più significativa; lo si può notare dalla figura 4.21, in cui la variabile ha un andamento abbastanza differente tra i vari cluster. In particolare il cluster 3 presenta i valori di trasmittanza migliori mentre il cluster 1 i peggiori. La trasmittanza opaca è la variabile meno significativa per la suddivisione dei cluster; infatti, dall'immagine 4.20 è evidente che per tutti i cluster i relativi valori hanno la stessa variabilità. Per quanto riguarda il fattore forma (Fig 4.19), ci si aspettava un andamento poco significativo, visto che edifici molto diversi in grandezza possono avere lo stesso rapporto tra superficie disperdente e volume lordo

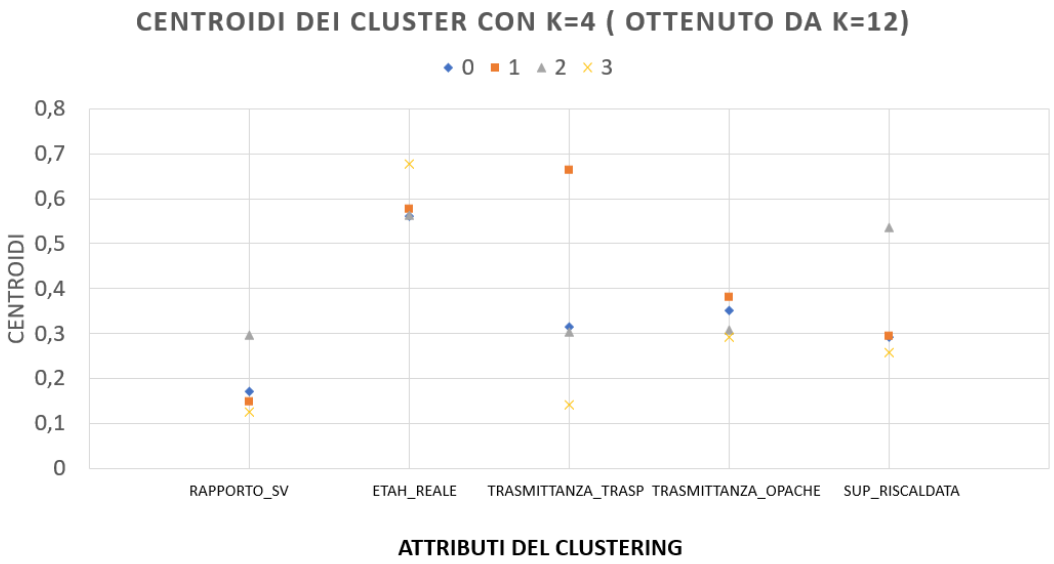

Figura 4.17: Grafico dei centroidi degli attributi utilizzati per il clustering con K=4 (ottenuto per riaggregazione).

riscaldato; tuttavia, è possibile notare come il cluster 2 sia riuscito a distinguere gli edifici meno compatti e quindi meno efficienti.

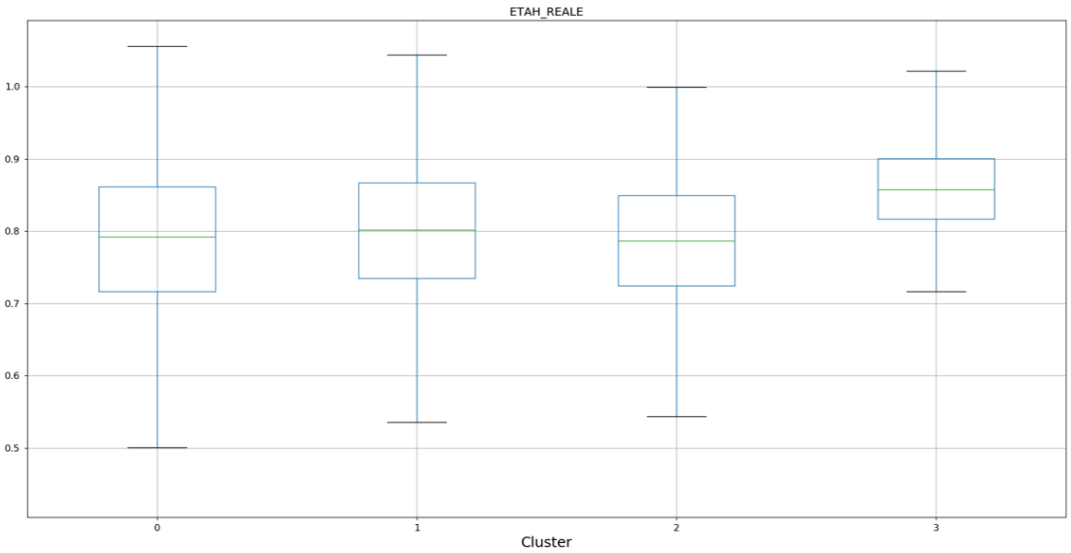

Figura 4.18: Distribuzione dei valori del rendimento medio globale stagionale invernale per i 4 cluster.

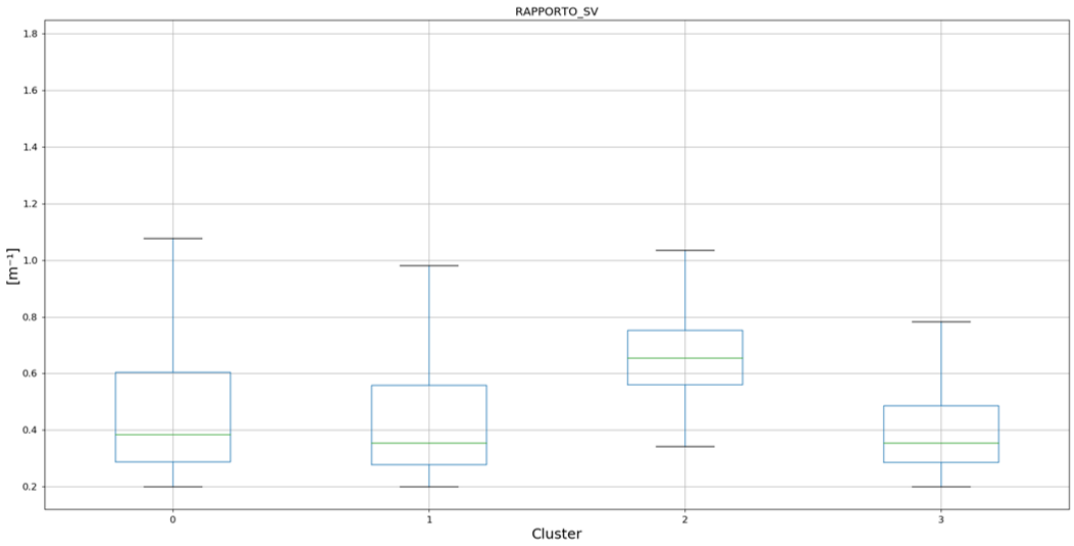

Figura 4.19: Distribuzione dei valori del fattore forma per i 4 cluster.

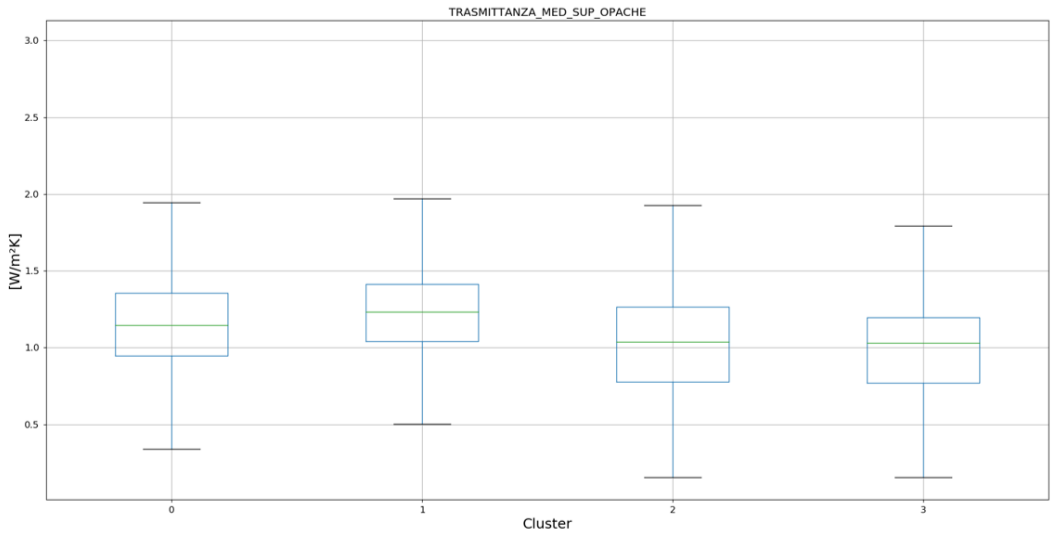

Figura 4.20: Distribuzione dei valori della trasmittanza media delle superfici opache per i 9 cluster.

Per ogni cluster sono stati analizzate alcune informazioni dei certificati appartenenti per comprenderne al meglio la caratterizzazione. Innanzitutto, si è voluto analizzare l'andamento dei valori dell'EP_H per comprendere se il limite precedentemente descritto, in merito all'isolamento delle performance energetiche, è stato superato. A tale scopo, sono stati leggermente modificati i limiti dei range dell'EP_H per definire

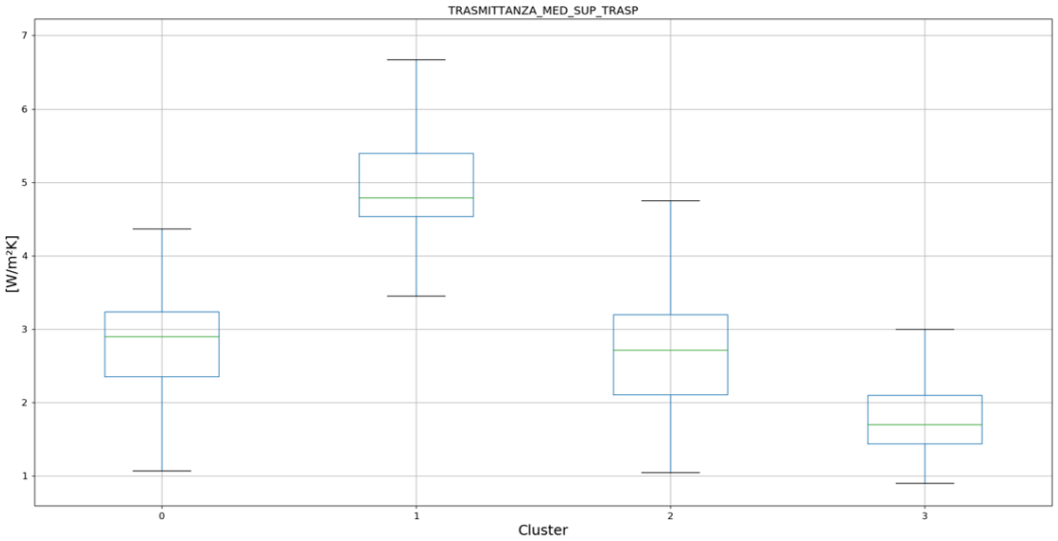

Figura 4.21: Distribuzione dei valori della trasmittanza media delle superfici trasparenti per i 9 cluster.

un certificato come *high*, *medium* e *low*, seguendo un approccio maggiormente *data driven*. I nuovi limiti sono:

- $0 \leq EP_H \leq 57 \Rightarrow$ *high*, edificio performante
- $57 < EP_H \leq 82 \Rightarrow$ *medium*, edificio mediamente performante
- $EP_H > 82 \Rightarrow$ *low*, edificio non performante

Dalla figura 4.22 possibile notare nei cluster con una maggiore predominanza del range *high* (cluster 3) e del range *low* (cluster 2), una forte riduzione della percentuale di valori *medium*. Questo sta ad indicare una maggiore capacità del metodo di riaggregazione nell' isolare i cluster con significative caratteristiche energetiche. In seguito, si è voluto comprendere per ogni cluster le epoche di costruzione degli edifici; per fare ciò, sono stati selezionati degli intervalli temporali che si differenziano o per l'entrata in vigore di particolari leggi che regolamentano i consumi energetici degli edifici o per epoche storiche. I primi due intervalli separano gli edifici costruiti rispettivamente prima e durante i due conflitti mondiali. In seguito, sono stati selezionati come limiti per gli intervalli temporali i seguenti anni[1]:

[1]Questi dati sono stati ottenuti da una guida rilasciata da ENEA, in cui viene descritta la legislazione nella regione Lombardia in merito alle certificazioni energetiche

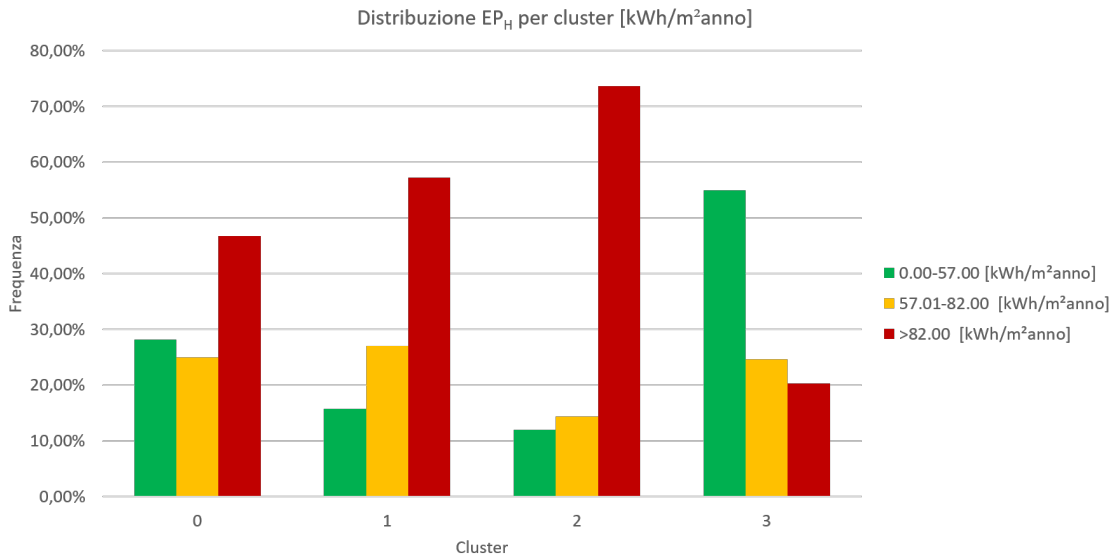

Figura 4.22: Distribuzione dei valori dell' EP_H per i clsuter ottenuti con K=4 dopo la riaggregazione.

- il 1973, anno in cui è stata emanata una legge che imponeva dei limiti in merito alla dispersione termica verso l'esterno,
- il 1991, anno della legge per l'attuazione del piano energetico nazionale in merito all'uso di fonti rinnovabili e al risparmio energetico,
- il 2005, anno in cui sono state attuate le direttive europee sul rendimento energetico nell'edilizia

Dalla figura 4.23 è possibile notare che in ogni cluster circa il 60% degli edifici è stato costruito prima del 1973 e ciò è giustificato dal fatto che circa l'80% del dataset iniziale presenta edifici costruiti prima dello stesso anno. Per tanto, da questa analisi possiamo rilevare il perché l'attributo relativo alla trasmittanza trasparente non è significativo nel clustering. Infatti, se gli anni di costruzione degli edifici seguono la stessa distribuzione per ogni cluster, vorrà dire che le modalità di costruzione degli edifici ed i materiali saranno per lo più simili per edifici costruiti nello stesso periodo. Inoltre, si è pensato potesse essere interessante analizzare come si sono distribuite le etichette delle classi energetiche tra i cluster. Anche in questo caso, gli edifici più performanti, ovvero quelli di classe A1, A2, A3 e A4, sono solo il 3% del dataset; così, per tenere in considerazione anche questa porzione, è stato ritenuto opportuno accorpare le classi energetiche nel seguente modo:

- A4, A3, A2, A1, B, C: edifici molto performanti

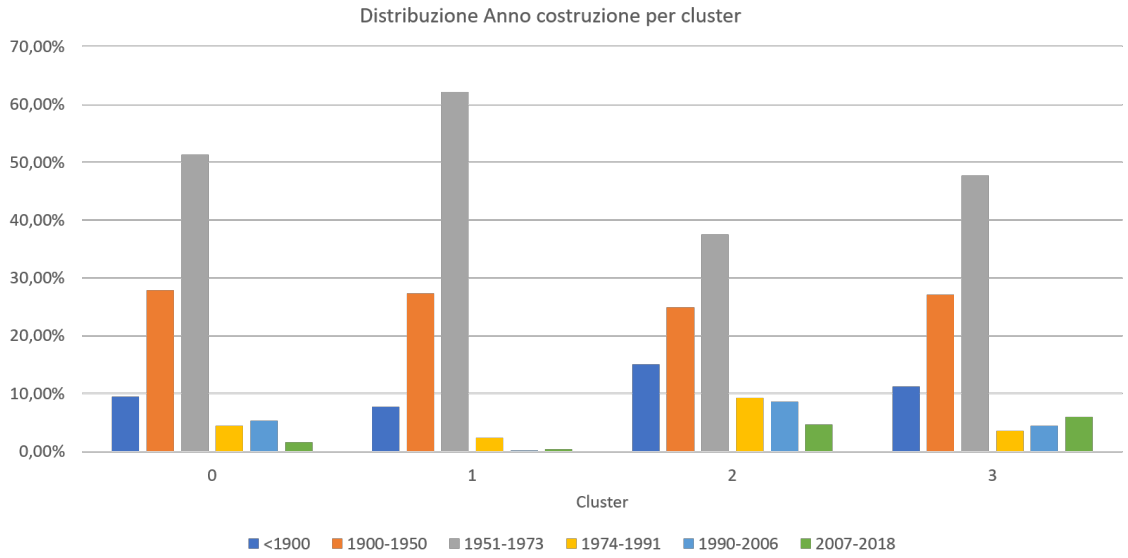

Figura 4.23: Distribuzione degli anni di costruzione per i 4 cluster dopo la riaggregazione.

- D, E : edifici con performance nella media
- F, G : edifici poco performanti

Dalla figura 4.24 è possibile notare che, anche in questo caso, poiché il 60% degli edifici nel dataset iniziale ha etichetta F o G, anche nei cluster queste risultano essere la maggioranza, tranne per il cluster 3, che è il cluster in cui sono presenti circa il 25% degli edifici ritenuti più performanti ed in cui la maggioranza degli edifici ha etichetta D o E. Inoltre, da questa figura risalta il cluster 1 come quello con quasi l' 80% di edifici poco performanti. Tuttavia, dai risultati ottenuti si è ritenuto opportuno non considerare la classe energetica nella caratterizzazione del cluster, poiché fuorviante per quanto riguarda l'efficienza energetica; ciò nasce dal fatto che la classe energetica è determinata dal confronto dell' EP_{GL} reale con il relativo valore di riferimento e dalla possibilità di poter assumere la classe F nel caso di autocertificazione. Dalle analisi finora condotte, considerando complessivamente le riflessioni fatte, si è ritenuto possibile individuare il cluster 3 come rappresentativo degli edifici più performanti. Tuttavia, occorre considerare un'ulteriore aspetto a supporto della suddetta tesi. Infatti, il cluster 3 presenta una corposa porzione di edifici più vecchi ma non mantiene la stessa proporzione per quanto riguarda i valori di EP_H considerati poco performanti, come invece succede per gli altri cluster. Per tale ragione, si è pensato di approfondire le analisi, andando a considerare anche l'attributo relativo alla ristrutturazione. Questo campo non è obbligatorio nella redazione dell' APE e, pertanto, può essere lasciato vuoto. Dalla figura 4.25 è infatti possibile no-

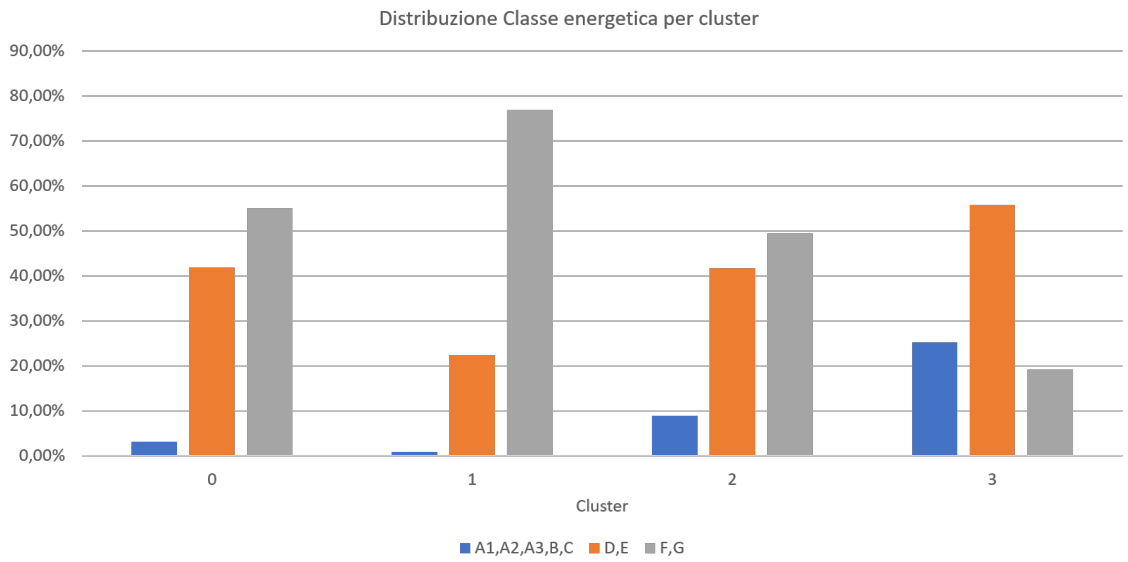

Figura 4.24: Distribuzione delle classe energetiche raggruppate in base alle performance per i 4 cluster dopo l'aggregazione.

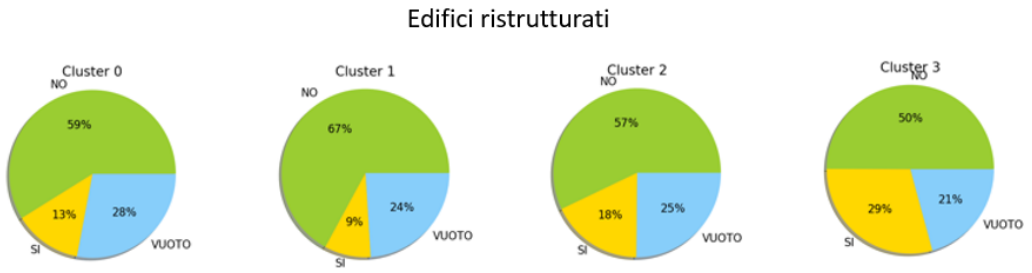

Figura 4.25: Distribuzione degli edifici ristrutturati per i 4 cluster dopo la riaggregazione.

tare che, per ogni cluster, circa il 20-30% dei certificati presenta il suddetto campo vuoto. Analizzando come sono distribuiti gli edifici ristrutturati per cluster, emerge che il cluster 3 è l'unico ad avere una porzione consistente di edifici ristrutturati, circa il 30%. Questo spiega il perché il cluster 3 risulta il cluster con gli edifici più performanti; infatti, se consideriamo che una ristrutturazione importante non viene fatta di solito su edifici di recente costruzione, vuol dire che una buona parte degli edifici ristrutturati era vecchia ed ha migliorato le sue prestazioni energetiche. Ciò può essere spiegato attraverso la figura 4.26, in cui si evince che l'80% degli edifici ristrutturati è stato costruito prima del 1973, ed attraverso la figura 4.27, da cui

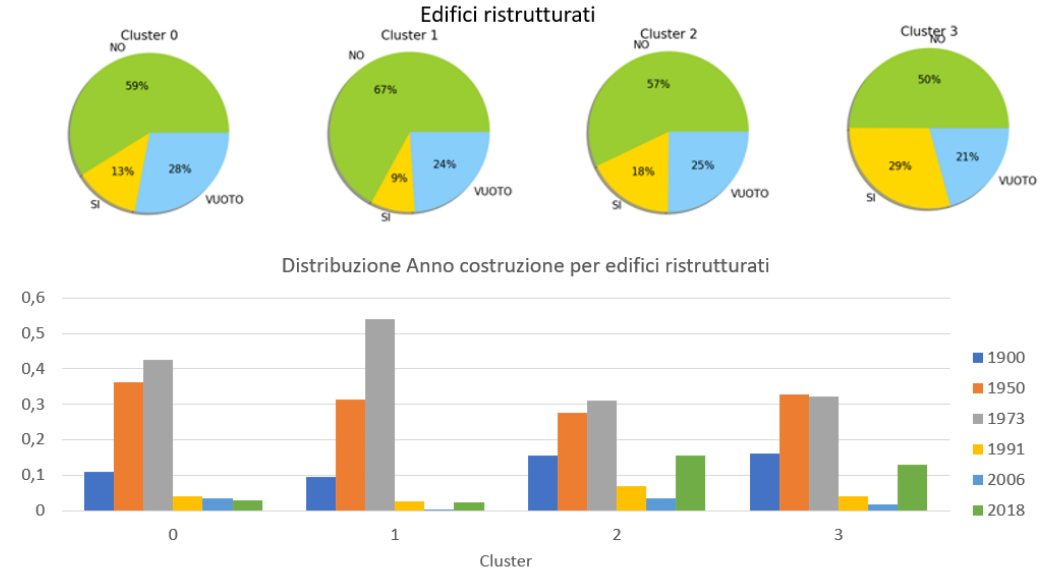

Figura 4.26: Distribuzione degli anni di costruzione dei soli edifici ristrutturati per i 4 cluster dopo la riaggregazione.

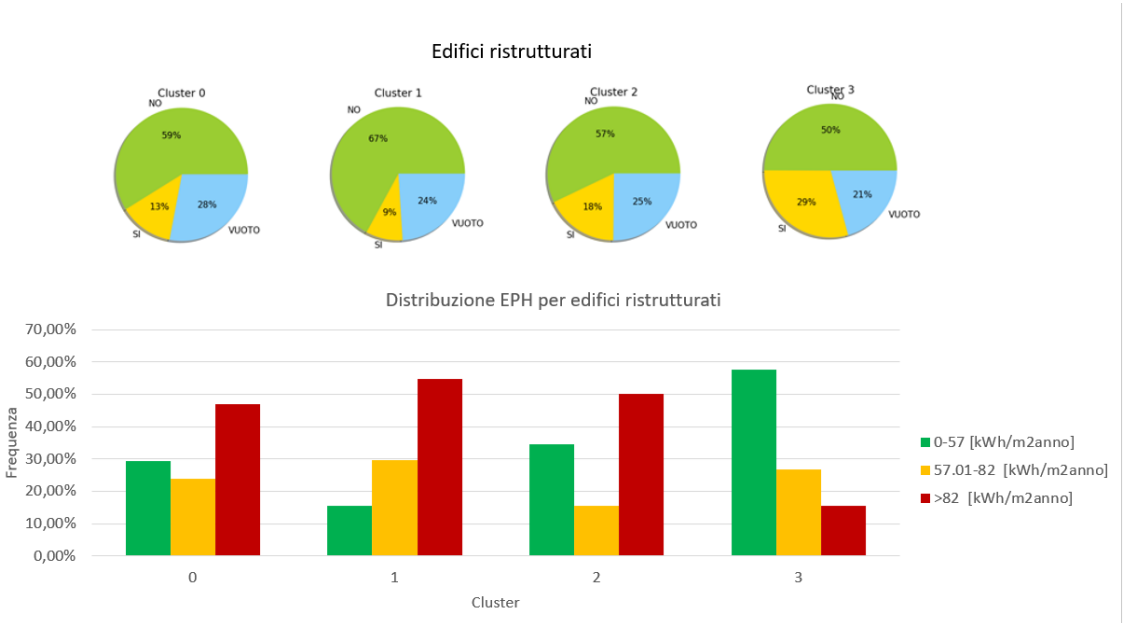

Figura 4.27: Distribuzione dei valori dell' EP_H dei soli edifici ristrutturati per i 4 cluster dopo la riaggregazione.

si rileva che il 60% degli edifici ha assunto buon valore di EP_H. Dalle stesse figure, si nota inoltre che il cluster 1 è quello che racchiude il minor numero di edifici ristrutturati e, tra questi, circa il 60% continua ad avere un valore di EP_H molto alto; probabilmente perché la ristrutturazione non ha intaccato l'edificio dal punto di vista energetico.

4.2.2 Regole di associazione

Dopo che è stato applicato il processo di *clustering*, si è proceduto con l'estrazione delle regole di associazione da ogni cluster. A tale scopo si è utilizzato l'algoritmo FP-growth in quanto computazionalmente migliore dell'Apriori, al quale sono stati fissati il valore di supporto minimo uguale a 0,2 e il valore di confidenza minimo uguale a 0,5. Tale processo è stato applicato agli attributi che sono risultati più significativi nella fase di clustering (trasmittanze e rendimento medio globale dell' impianto), congiuntamente con l'indice di prestazione energetica per la climatizzazione invernale (EP_H) e l'anno di costruzione. In questo modo, è stato possibile caratterizzare ulteriormente i cluster sia da un punto di vista termo-fisico che energetico, oltre che in base al periodo di costruzione. Inoltre, per discretizzare gli attributi scelti, sono stati utilizzati i valori del primo quartile e della mediana delle distribuzioni di ciascun attributo, tranne per la variabile EP_H, per la quale sono stati utilizzati i range definiti dall'esperto di dominio nella fase di valutazione dei clustering (paragrafo 4.2.1). Data la numerosità delle regole generate, è stato necessario introdurre due metriche per valutare la bontà di una regola: il *lift* e la *conviction*. Dunque, nel processo di selezione delle regole, sono state scelte quelle relazioni il cui valore del *lift* fosse maggiore di 1,05 e la *conviction* maggiore di 3,05. Nella figura 4.28 e 4.29

CLUSTER 2						
	Antecedent	**Consequent**	**Confidence**	**RHS_support**	**Lift**	**Conviction**
	PERFORMANTE RECENTE COSTRUZIONE	RENDIMENTO DELL'IMPIANTO BUONO TRASMITTANZE BUONE	0,91	0,62	1,46	4,16
	PERFORMANTE RENDIMENTO DELL'IMPIANTO BUONO RECENTE COSTRUZIONE	TRASMITTANZE BUONE	0,91	0,67	1,36	3,63
	RENDIMENTO DELL'IMPIANTO BUONO TRASMITTANZE BUONE RECENTE COSTRUZIONE	PERFORMANTE	0,83	0,28	2,93	4,29
	MEDIAMENTE PERFORMANTE RENDIMENTO DELL'IMPIANTO MEDIO COSTRUZIONE NON RECENTE	TRASMITTANZE BUONE	0,92	0,67	1,38	4,29

Figura 4.28: Regole estratte dal cluster 2.

sono mostrati alcuni esempi di regole estratte rispettivamente dal cluster 2 e dal

CLUSTER 3						
	Antecedent	Consequent	Confidence	RHS_support	Lift	Conviction
	PERFORMANTE TRASMITTANZE BUONE RECENTE COSTRUZIONE	RENDIMENTO DELL'IMPIANTO MEDIO	0,92	0,66	1,4	4,12
	MEDIAMENTE PERFORMANTE RENDIMENTO DELL'IMPIANTO SCARSO RECENTE COSTRUZIONE	TRASMITTANZE MEDIE	0,88	0,47	1,86	4,23

Figura 4.29: Regole estratte dal cluster 3.

cluster 3. In particolare, è possibile notare che tra i certificati presenti nel cluster 2, come ci si potrebbe aspettare, risultano essere altamente performanti (valore di EP_H inferiore a 50 kWh/m^{2}K) gli edifici di recente costruzione (anni 2006-20018) che hanno degli ottimi valori di trasmittanza trasparente (minore di 3,00 W/m^{2}K) e di rendimento medio globale (maggiore di 0,90). Mentre analizzando le regole estratte dal cluster 3, risulta che un edificio mediamente performante (EP_H minore di 100 kWh/m^{2}K), con un cattivo rendimento medio globale dell'impianto (minore di 0,75) ma costruito recentemente (anni 2006-20018), presenta dei valori di trasmittanza nella media (compresi tra 3,00 W/m^{2}K e 5,00 W/m^{2}K).

4.3 Validazione e visualizzazione della conoscenza

Allo scopo di visualizzare i dati ed i risultati ottenuti nella maniera più intuitiva possibile, sono state realizzate delle mappe interattive che hanno permesso una visualizzazione della conoscenza a diversi livelli di dettaglio.

4.3.1 Visualizzazione dei dati

Nella fase iniziale, è stata fatta un'analisi esplorativa del dataset per comprendere la distribuzione dei dati ed i relativi aspetti critici. A supporto di tale analisi, è stata realizzata una mappa interattiva che consente di ottenere le informazioni utili sulla distribuzione di un determinato attributo in una specifica area della città di Torino. Nella figura 4.30 è mostrato un esempio di risultato ottenuto dalla navigazione della suddetta mappa per la circoscrizione 8 della città di Torino. In particolare, nella porzione della mappa relativa a tale circoscrizione, è possibile vedere ogni isolato colorato in base al valore medio del valore di ETAH assunto dai certificati in quella zona, secondo la scala cromatica stabilita. La presenza di isolati di colore grigio indica l'assenza di certificati in quell'isolato o la presenza di un numero troppo piccolo per poter essere preso in considerazione nelle analisi. I *marker cluster* presenti

assumono una forma circolare di dimensione proporzionale al numero di record presenti nella zona a cui si riferiscono e sono colorati secondo lo stesso principio sopra citato. Nella finestra mostrata accanto alla circoscrizione sono presenti le distribuzioni della variabile considerata in quella specifica circoscrizione, utilizzando gli stessi colori delle mappe per coerenza e per suggerire un' associazione immediata tra la rappresentazione su mappa e quella più analitica.

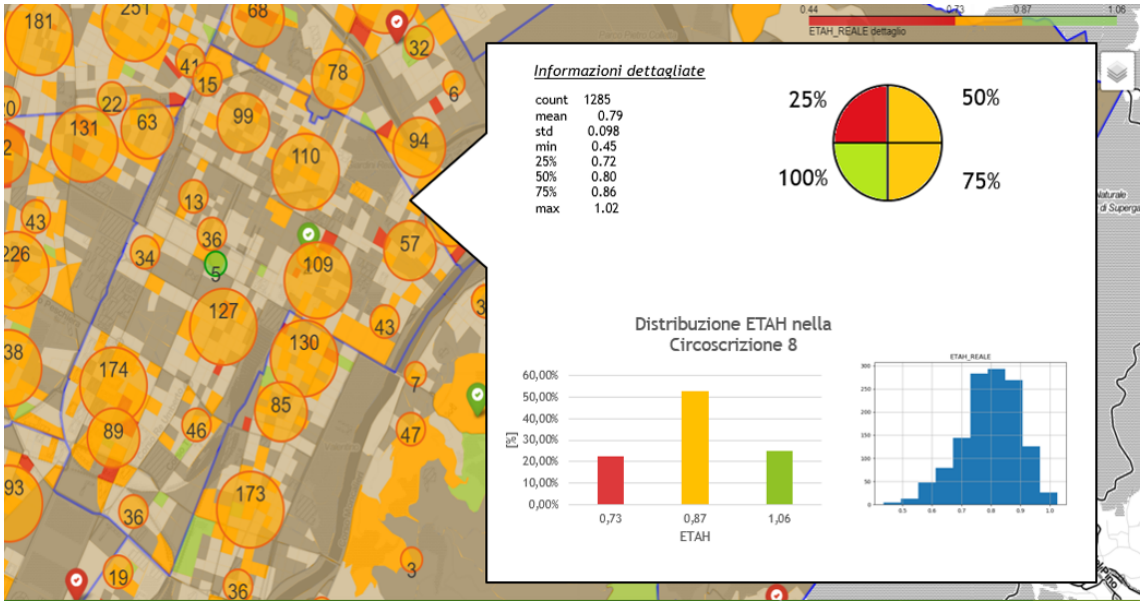

Figura 4.30: Esempio di dettaglio della circoscrizione 8 per l'attributo ETAH.

Capitolo 5

Conclusioni e sviluppi futuri

L'obiettivo della presente tesi è stato quello di progettare e sviluppare un architettura che possa supportare l'analista durante l'analisi di grandi volumi di dati relativi ad edifici geolocalizzati. A tal proposito è stato sviluppato TUCANA, un tool automatico in grado di caratterizzare l'efficienza energetica degli edifici residenziali, presenti nella città di Torino, attraverso tecniche di *data mining*. Questo tool consente anche una visualizzazione dei dati su mappe geolocalizzate così da supportare l'analista nella prima fase esplorativa e nella visualizzazione della conoscenza estratta. L'architettura sviluppata è stata applicata ad una porzione del Catasto energetico degli edifici della Regione Piemonte; in particolare, le analisi sono state effettuate sui certificati rilasciati dal 2016 al primo semestre del 2018. Nello specifico, attraverso il *framework* TUCANA, è possibile individuare i certificati invalidi, cioè quelli che presentano errori soprattutto nei campi fondamentali, determinando l'inattendibilità dell'intero APE. Inoltre, l'architettura sviluppata garantisce un supporto all'analista nel *decision making*, attraverso la rappresentazione dei dati con mappe geolocalizzate, consentendogli di analizzare velocemente insiemi di dati più interessanti, rilevandone gli aspetti più significativi. Inoltre, attraverso la caratterizzazione degli edifici, è stato possibile individuare le relazioni tra le caratteristiche termo-fisiche degli edifici e le relative performance energetiche e visualizzare tale conoscenza sulle mappe geolocalizzate, così da permettere anche a un non esperto di dominio di comprendere le problematiche in termini di efficienza energetica edilizia e di poter individuare aree precise in cui proporre soluzioni adatte alla tipologia di edifici presenti.

La metodologia proposta è stata applicata solo ad edifici residenzali, ma potrebbe essere estesa ad edifici con altre tipologie di destinazioni d'uso. Inoltre, sarebbe interessante integrare i dati del catasto con i dati degli impianti, così da valutare altre componenti determinanti per valutare l'efficienza energetica. Valutare altri algoritmi

di clustering, come quelli gerarchici, potrebbe permettere un'ulteriore caratterizzazione degli edifici. Infine, sarebbe possibile estendere la tecnica di visualizzazione dei dati sulle mappe per consentire una selezione dinamica della zona di interesse, rendendola maggiormente navigabile.

Bibliografia

[1] Rakesh Agrawal e Ramakrishnan Srikant. «Fast Algorithms for Mining Association Rules ». In: *Proc. of the 20th VLDB Conference,* (1994), pp. 478–499.

[2] Assotermica. *Impianti termici: concetti innovativi della normativa vigente*. A cura di Anima. 1998.

[3] Motwani R. Ullman J.D. Tsur S. Brin S. «Dynamic itemset counting and implication rules for market basket data». In: *Proceedings of the 1997 ACM SIGMOD international conference on Management of data* 26 (1997), pp. 255–264.

[4] Stefano Cascio. *Guida alla certificazione energetica degli edifici*. A cura di Grafill. 2016.

[5] Michael Friendly. «Milestones in the history of thematic cartography, statistical graphics, and data visualization». In: (2008).

[6] Capozzoli Serale Piscitelli Grassi. «Data mining for energy analysis of a large data set of flats». In: *ICE Publishing* (2017), pp. 1–16.

[7] Tukey J.W. Hoaglin D. Mosteller F. «Understanding robust and exploratory data analysis». In: (1983).

[8] Boris Iglewicz e David Hoaglin. *Volume 16: How to Detect and Handle Outliers*. A cura di The ASQC Basic References in Quality Control: Statistical Techniques. Edward F. Mykytka, Ph.D., Editor, 1993.

[9] Rakesh Agrawal Tomasz Imielinski e Arun Swami. «Mining Association Rules between Sets of Items in Large Databases». In: (1993).

[10] Usama Fayyad Gregory Piatetsky-Shapiro e Padhraic Smyth. «From Data Mining to Knowledge Discovery in Databases». In: *AI Magazine* 17 (1996).

[11] Erhard Rahm e Hong Hai Do. «Data Cleaning: Problems and Current Approaches». In: *Data Engineering* 23 (2000).

[12] Y. Olivo A. Hamidi P. Ramamurthy. «Spatiotemporal variability in building energy use in New York City». In: *Energy* 141 (2017), pp. 1393–1401.

[13] B. Rosner. «Percentage Points for a Generalized ESD Many-Outlier Procedure.» In: *Technometrics* 25 (1983), pp. 165–172. DOI: 10.2307/1268549.

[14] P. J. Rousseeuw. «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. » In: *Journal of Computational and Applied Mathematics* (1987), pp. 53–65.

[15] ANGELA SANCHINI. «Manuale operativo redazione APE». In: *Insiel* (2017).

[16] Luai Al Shalabi Zyad Shaaban e Basel Kasasbeh. «Data Mining: A Preprocessing Engine». In: *Journal of Computer Science* (2006), pp. 735–739.

[17] Kumar Tan Steinbach. *Introduction to Data Mining*. A cura di Pearson. 2006.

[18] Martin Ester Hans-Peter Kriegel Jiirg Sander Xiaowei X. «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise». In: *KDD-96 Proceedings* (1996).