Università degli studi di Padova

Tesi Magistrale

Analisi e sperimentazione di approcci statistici per l'individuazione di errori non formali nei database topografici

Autore:
Davide Ereno

Relatore:
Prof. Massimo Rumor

Laboratorio G.I.R.S.
Dipartimento di ingegneria dell'informazione

1 luglio 2015

"I have not failed. I've just found 10,000 ways that won't work."

Thomas A. Edison

UNIVERSITA‘ DEGLI STUDI DI PADOVA

Abstract

Ingegneria Informatica
Dipartimento di ingegneria dell’informazione

Analisi e sperimentazione di approcci statistici per l’individuazione di errori non formali nei database topografici

by Davide Ereno

L’individuazione degli errori non formali nei database topografici richiede la definizione di regole che non sono facilmente individuabili e che comunque dipendono dal modello dati utilizzato. Sono quindi necessarie delle tecniche che possano semplificare questo compito. In questa tesi verranno presentati alcuni approcci, prevalentemente statistici, per l’individuazione di tali errori. Inoltre verranno presentati i risultati di questi procedimenti applicati su database reali.

Indice

Capitolo 1

Introduzione

1.1 Cartografia

Sin dai tempi più antichi vi fu la necessità di rappresentare il mondo che ci sta intorno. Le popolazioni babilonesi furono le prime a creare mappe tramite la misurazione del territorio imprimendole su tavolette di creta. Furono però i greci a dare le basi della cartografia come scienza, con Dicearco da Messina che creò la prima rappresentazione in cui venivano riportati i meridiani e le località ritenute essere alla stessa latitudine. Successivamente con l'aumento delle esplorazioni e delle tecnologie per la rilevazione dei dati è stato possibile creare mappe sempre più dettagliate.

1.1.1 Definizione e Principi delle carte geografiche

La carta geografica è stata definita dall'associazione internazionale di cartografia come:

"la rappresentazione in piano dei fenomeni e delle condizioni di fatto della Terra resa in proiezione orizzontale, rimpicciolita, generalizzata e dichiarata nei suoi segni"

Da questa definizione si può capire che i principali concetti riguardanti le carte geografiche sono:

Proiezione e deformazione

Dato che si vuole proiettare una parte di una superficie curva su un piano, dovranno inevitabilmente essere fatte delle deformazioni geografiche. Le proiezioni sono appunto i processi matematico-geometrico per mezzo dei quali la superficie sferica della terra viene trasformata in una superficie piana.

Scala

La scala è definibile come il rapporto tra la lunghezza misurata sulla carta, e la stessa lunghezza misurata nella realtà. Pensando alla scala come al valore di una frazione, che diminuisce all'aumentare del denominatore della frazione, si può dire che una mappa in

1:5000 ha una scala più grande di una mappa in 1:25000. Il valore della scala è quindi inversamente proporzionale al fattore di riduzione adottato: all'aumentare del secondo diminuisce il primo.

Tramite la scala può essere fatta la distinzione tra carte a larga scala, in cui vengono rappresentati con grande dettaglio gli elementi di una zona ristretta, e carte a bassa scala, in cui sono rappresentate grandi aree geografiche con un livello di dettaglio minore. In particolare possono essere distinte queste categorie:

- **piante e mappe**: di scala superiore a 1:10.000, vengono usate per la rappresentazione tecnica di ambienti ed edifici, o per mappe catastali.
- **carte topografiche**: con una scala tra 1:10.000 e 1:200.000, vengono usate per rappresentare con molta accuratezza zone geografiche limitate, un esempio sono le carte dell'istituto Geografico Militare.
- **carte corografiche**: con una scala tra 1:200.000 e 1: 1.000.000, vengono usate per rappresentare zone abbastanza estese come regioni o stati.
- **carte generali**: con scale superiori a 1:1.000.000, sono quelle che vengono usate per la rappresentazione dei continenti o dell'intero globo.

La scala quindi influisce anche in una serie di altri parametri propri di una mappa, come per esempio il *grado di risolutezza* cioè la lunghezza lineare del più piccolo elemento rappresentabile nella mappa. *L'errore massimo di posizionamento* cioè la massima incertezza della posizione di un punto nella cartina rispetto alla realtà. *Ed ancora il livello di dettaglio* che rappresenta la quantità di informazione presente nella cartina.

1.2 Il processo cartografico

Con processo cartografico si definisce tutto l'insieme di procedimenti e operazioni necessarie alla creazione di una carta geografica. Nonostante nel corso dei secoli il progresso delle scienze e delle tecnologie abbia portato ad uno sviluppo delle tecniche e degli strumenti in possesso dei cartografi, il processo di produzione cartografica è rimasto sostanzialmente invariato nel tempo, e può essere schematizzato nei seguenti passi:

- **Analisi e definizione delle caratteristiche**: Il primo passo del processo riguarda la progettazione della mappa: durante la fase di definizione ed analisi, si decidono le caratteristiche che il prodotto finito dovrà possedere. I parametri tra cui scegliere sono molteplici, quali ad esempio la scala del prodotto finito, la proiezione da utilizzare e non da ultimo la sua destinazione: se si tratta di una mappa tradizionale a supporto cartaceo, oppure una carta numerica, destinata ad una fruizione digitale. Durante la fase di definizione ed analisi viene deciso cosa rappresentare sulla mappa e come rappresentarlo: le scelte effettuate in questa fase influenzeranno sia le caratteristiche tecniche della carta, sia quelle semantiche, relative cioè ai suoi contenuti. Per quanto riguarda le caratteristiche tecniche di una carta geografica, possiamo citare la superficie di riferimento e di proiezione adottata, il tipo di rappresentazione (conforme, equivalente, equidistante) e la tipologia della sua generazione (geometrica, semi-geometrica, analitica) e non da ultimo la

scala. Dal lato delle caratteristiche semantiche, in base al prodotto finale che si vuole ottenere è possibile condurre l'attività di astrazione e semplificazione dei dati in ingresso in modo tale da focalizzare l'attenzione solo su alcuni particolari aspetti della realtà da rappresentare, oppure cercare di avere una rappresentazione il più completa possibile. Nel primo caso parleremo di cartografia meristica, o tematica, contenente un elevato livello di informazione su uno o più specifici argomenti - detti tematismi - , nel secondo di cartografia olistica, o analitica. E' importante sottolineare come le decisioni prese in questa fase corrispondono a definire una prima astrazione della realtà, soprattutto a livello semantico, la scelta di come distinguere e classificare gli elementi che si trovano sul territorio coincide col creare un modello di realtà. Nella produzione delle tradizionali mappe cartacee, questo processo di modellazione porta alla definizione della legenda della carta, che stabilisce quali elementi saranno presenti e come saranno rappresentati, mentre nella produzione di mappe digitali, questa fase porta alla definizione di un GeoDB.

- **Raccolta dati**: La fase successiva alla definizione e analisi, è la raccolta dei dati. In questo passo si possono presentare due scenari differenti: il dato di partenza può essere la realtà, e quindi la raccolta dei dati sarà effettuata tramite campagne di acquisizione eseguite sul territorio, oppure come dati di partenza verranno usati quelli di una cartografia pre-esistente. Nel primo caso si parla di carte rilevate, nel secondo di carte derivate. E' importante sottolineare che questa seconda alternativa è percorribile solo nel caso sia già esistente una cartografia con un livello di dettaglio maggiore (e quindi una scala maggiore) di quella da produrre. L'acquisizione dei dati per la creazione di una cartografia rilevata è un'attività lunga e composta da varie fasi. In passato l'attività di rilevazione veniva compiuta esclusivamente con ricognizioni a terra, utilizzando strumenti quali la tavoletta pretoriana e il teodolite, misurando angoli e distanze degli oggetti partendo da punti di posizione nota, ad esempio i vertici delle reti geodetiche. Al giorno d'oggi l'acquisizione dei dati è velocizzata dal ricorso a nuove tecnologie: è possibile effettuare il rilievo della posizione tramite GPS, e la restituzione del territorio si può eseguire tramite fotogrammetria. La fotogrammetria è una tecnica di rilevazione della posizione di punti mediante l'utilizzo di immagini fotografiche stereoscopiche del terreno. Si tratta per lo più di immagini fotografiche riprese da aereo, in sequenze chiamate strisciate, o strip, utilizzate a coppie, e tali che ciascun fotogramma si sovrapponga per circa il 60% con quelli adiacenti e ciascuna strisciata si sovrapponga per il 15% con quelle adiacenti (sidelap). Appositi strumenti, detti stereorestitutori, permettono di rilevare le coordinate degli oggetti presenti nelle strisciate e ricostruirne la posizione in latitudine, longitudine ed elevazione, posti come noti la quota dell'aereo al momento dello scatto e la posizione di alcuni punti di riferimento. L'utilizzo della fotogrammetria integra le operazioni di rilievo topografico, consentendo un rilievo di dettaglio del territorio a costi notevolmente ridotti rispetto al rilievo diretto (per ampie zone). Durante le fasi di acquisizione e restituzione, gli operatori assegnano ad ogni oggetto rilevato un codice che ne identifica la natura (casa, ferrovia, strada, ..) basandosi sul modello stabilito durante la precedente fase di analisi e definizione delle specifiche della mappa. Queste annotazioni verranno poi utilizzate nella successiva fase di costruzione della mappa.

- **Produzione della mappa**: Una volta raccolti i dati, si passa alla fase di costruzione della mappa: qui il cartografo, sfruttando gli strumenti a disposizione, e la sua conoscenza ed esperienza, produce una carta che deve soddisfare non

solo tutte le specifiche decise nella fase di definizione ed analisi, ma anche requisiti irrinunciabili per una mappa quali la leggibilità e l'usabilità. Per realizzare questo, il cartografo deve estrarre dai dati di partenza una loro rappresentazione astratta ma al contempo efficace e rappresentativa: il suo lavoro è cioè mirato alla creazione di un'astrazione della realtà geografica che faciliti la comprensione e la comunicazione dell'informazione. In questi termini, è possibile definire questo processo come un processo di generalizzazione ("Generalisation aims to provide an abstraction of geographic reality to enhance comprehension and communication of information", Agent Esprit 2001). Parlando di generalizzazione, in ambito cartografico, ci si può scontrare con una certa ambiguità nell'uso di questo termine: nel creare una rappresentazione cartografica della realtà sono infatti coinvolte due attività di astrazione che riguardano aspetti diversi del dato geografico: da una parte il contenuto semantico, dall'altra l'informazione spaziale (posizione, forma). La definizione di generalizzazione precedente accorpa entrambe queste due attività, conosciute anche come generalizzazione del modello e generalizzazione cartografica, ma durante la fase di costruzione della mappa viene eseguito il solo processo di generalizzazione cartografica, quando il cartografo sceglie e posiziona uno ad uno gli oggetti sulla carta finale. Il processo di generalizzazione del modello invece viene posto in essere già durante la fase di analisi e definizione delle specifiche della mappa. Il capitolo successivo di questa tesi tratterà dettagliatamente l'argomento della generalizzazione. La costruzione di una carta geografica include anche una fase di raffinamento estetico che si può considerare non completata durante il solo processo di generalizzazione cartografica: questa fase include attività quali il posizionamento della toponomastica, la creazione di ombreggiature o sfumature per la delineazione dell'orografia, il posizionamento di riferimenti come la griglia di inquadramento geografico e in generale la vestizione dei particolari.

- **Collaudo della mappa**: L'ultima fase del processo, quella di collaudo, prevede che la carta venga sottoposta ad una serie di test per verificarne la correttezza e la consistenza. Durante questa fase si può assistere ad un raffinamento estetico del prodotto, oltre che ad una verifica della validità della rappresentazione creata. Tra le attività svolte durante questa fase di collaudo si può citare il controllo della ripresa aerea, della determinazione della rete di raffinamento e dei punti d'appoggio, della restituzione, del disegno e della ricognizione e il controllo finale sul terreno mediante operazioni di misura e verifica della rappresentazione cartografica. Quest'ultima è una delle attività più importanti della fase di collaudo e viene eseguita confrontando sul campo i dati riportati nella mappa con dati rilevati tramite strumenti ad alta precisione, come il GPS differenziale.

1.3 Passaggio ai dati digitali

Nella seconda metà del ventesimo secolo, con l'utilizzo delle fotografie aeree e satellitari e l'avvento della digitalizzazione dei dati è stato possibile velocizzare la produzione e l'aggiornamento delle mappe. Soprattutto negli anni 90 con l'espansione dei personal computer si è cominciato a vedere la nascita dei primi sistemi per la gestione dei dati territoriali chiamati GIS (Geographic Information Systems). Con questo termine si intende l'insieme degli strumenti usati per acquisire, gestire e rendere disponibile l'informazione territoriale. Si possono suddividere gli obbiettivi dei GIS nei seguenti punti:

- **Inserimento**: i dati per essere utilizzati da questi sistemi devono essere memorizzati in convenienti formati. Il processo per la conversione dei dati *"analogici"* (rilevazioni, ortofoto) a digitali viene chiamato *digitalizzazione*.
- **Manipolazione**: rientrano in questa categoria tutte quelle operazioni per il cambiamento di scala delle mappe, quindi generalizzazioni oppure unione di mappe a differente scala.
- **Gestione**: è svolto dai software di gestione di dati come i DBMS
- **Ricerca e analisi**: qualsiasi operazione per ottenere informazioni dai dati geografici, alcuni esempi potrebbero essere indici di distribuzione, gli strumenti più utilizzati per ottenere queste informazioni sono il *buffering* e *overlay*.
- **Visualizzazione**: La rappresentazione visuale dei dati che può essere sia sotto forma di mappa oppure di grafici, per aiutare la comprensione di questi.

Vale la pena soffermarsi nella fase di inserimento dei dati in quanto fonte principale di errori nei dati geografici. I dati geografici vengono salvati in particolari strutture che sono di fatto dei database con particolari attributi per la gestione delle geometrie. Infatti oltre alle informazioni spaziali-geometriche vengono memorizzate anche informazioni più astratte, come ad esempio l'uso di un edificio o la tipologia di una strada.

1.4 Errori nei dati

Come si può immaginare guardando il processo di creazione delle mappe dato che vi è una diretta interazione dell'uomo con i dati, vi è una probabilità praticamente pari a 1 che vengano commessi degli errori. Proprio sul concetto di errore si incentra questa tesi, in particolare nell'individuare una determinata tipologia di questi che può essere definita come: *errori non formali*. Con questo termine si vogliono indicare quegli errori che per poter essere scoperti necessitano di un confronto con la realtà. Nel capitolo successivo verrà data una definizione più precisa di questo concetto, prendendo in considerazione anche alcune tecniche per la loro rilevazione attualmente disponibili. Con questa tesi si è cercato di sviluppare dei sistemi per l'individuazione di questa tipologia di errori. Per la precisione le tecniche che saranno descritte mirano ad individuare un'altra categoria di dato, *le anomalie*. Ricadono in questa categoria quei dati che si discostano dalla normalità. Dato che le anomalie hanno una forte correlazione con gli errori non formali, possono essere sfruttate per l'individuazione di questi ultimi.

Capitolo 2

Qualità dei dati

2.1 Definizione

La qualità è un concetto generale e di grande importanza tanto da essere regolamentato da uno standard l'ISO 9000 che definisce la qualità come

Il grado in cui un insieme di caratteristiche intrinseche soddisfano i requisiti.

La qualità perché possa essere considerata come una grandezza, deve essere definita una misura. Ciò può essere fatto definendo quanto un prodotto è lontano dall'ideale. Per i dati geografici è stato sviluppato per questo scopo lo standard ISO 19113, nel quale vengono elencati quali sono gli elementi di qualità dei dati:

- **Completezza**: rappresenta quanto il dato rispecchia la realtà in termini di elementi mancanti o eccedenti. Si esprime come rapporto percentuale tra il numero di oggetti mancanti o sovrabbondanti rispetto al numero di oggetti totali presenti sul terreno.
- **Consistenza Logica**: dice quanto un dato rappresenta la realtà sotto l'aspetto topologico, della struttura del file e della validità dei valori tematici. I controlli per eliminare difetti di questo tipo possono essere automatizzati, un esempio sono i controlli per la chiusura dei poligoni.
- **Accuratezza posizionale**: accuratezza della posizione geografica di un oggetto rispetto alla fonte. Si valuta lo scostamento delle coordinate dalla reale posizione sul terreno rispetto alla tolleranza indicata. La verifica viene fatta su un campione di punti di controllo, utilizzando strumenti di misura che garantiscono una precisione maggiore. Si differenzia dall'accuratezza relativa che riguarda il posizionamento relativo tra gli oggetti
- **Accuratezza temporale**: accuratezza degli attributi e delle relazioni temporali degli oggetti. Per ogni attributo si può indicare la data dell'ultimo aggiornamento o modifica.
- **Accuratezza tematica**: correttezza di classificazione di un oggetto e degli attributi descrittivi.

2.2 Gli errori

Una componente molto importante nella descrizione della qualità dei dati è sicuramente la presenza di errori. Nel campo considerato in questa tesi, con errore si intende la differenza che vi è tra la rappresentazione della realtà e la realtà stessa. In [1] è stata data una suddivisione delle varie fasi da cui può provenire l'errore:

- **Raccolta dati**: imprecisione degli strumenti, incorrettezza nelle procedure di memorizzazione, errori nell'analisi dei dati rilevati da satelliti.
- **Inserimento dati**: errori di digitalizzazione, tortuosità dei bordi naturali, altre forme di inserimento.
- **Memorizzazione dati**: precisione numerica, precisione spaziale.
- **Manipolazione dei dati**: errori di adiacenza nei contorni, errori semantici, poligoni impuri e propagazione degli errori con le operazioni di overlay.
- **Dati in uscita**: dispositivi di output imprecisi.
- **Uso dei dati**: incomprensione dell' informazione, uso scorretto dei dati.

Inoltre sempre in [1] viene anche fornita una classificazione degli errori riportata di seguito:

- **Misurazione**: errata misurazione di una proprietà di un oggetto. Sono gli errori più facili da individuare perché negli anni sono state sviluppate molte procedure avanzate di analisi statistica;
- **Assegnazione**: l'oggetto viene classificato in modo sbagliato a causa di errori di misurazione sul campo o in laboratorio;
- **Generalizzazione di classe**: a seguito di osservazioni sul campo e per ragioni di semplicità, l'oggetto è raggruppato insieme ad altri oggetti che però hanno qualche proprietà differente;
- **Generalizzazione spaziale**: generalizzazione della rappresentazione cartografica di un oggetto prima di essere digitalizzato;
- **Inserimento**: i dati non sono codificati correttamente durante l'inserimento nel database;
- **Temporale**: l'oggetto cambia di tipologia nel tempo trascorso tra l'inserimento nel database e l'effettivo uso dei dati;
- **Elaborazione**: durante la trasformazione dei dati nascono errori dovuti agli algoritmi e agli arrotondamenti.

2.2.1 Individuazione degli errori

La presenza di errori nei dati spaziali rappresenta un indice di scarsa qualità dei dati, si deve quindi cercare di eliminarli o quantomeno ridurli il più possibile. Con l'avvento dei GIS e l'evoluzione di strumenti di acquisizione, analisi e correzione dei dati, la quantità di errori presenti nei dataset finali vengono ridotti notevolmente. Questo perché gli strumenti moderni permettono ad esempio di eliminare self-intersection in un poligono e punti doppi in una linea o in un contorno di un poligono. Uno di questi strumenti è ad esempio il DBTopoCheck sviluppato all'interno del progetto CARGEN [2], il quale permette di verificare la correttezza dei vincoli topologici tra feature definiti nelle specifiche.

Questo tipo di errori possono essere completamente eliminati con controlli automatici, chiamati controlli formali. Vi sono però altri errori, che tali controlli non possono evidenziare. Per questo motivo si può fare la seguente distinzione:

- **Errori Formali**: In questa tipologia ricadono gli errori che sono stati descritti precedentemente, cioè quelli per cui possono essere scritti dei controlli formali per la loro individuazione. I vincoli topologici e di dominio sono due esempi di questi errori(come la sovrapposizione di due edifici come mostrato nell'esempio in figure 2.1).
- **Errori non Formali**: il dato affetto da questo errore risulta essere comunque "possibile" nella rappresentazione della realtà cui appartiene, diventa quindi impossibile applicare delle regole certe che permettano di individuarli. In questa tipologia di errori ricadono gli errori di classificazione.

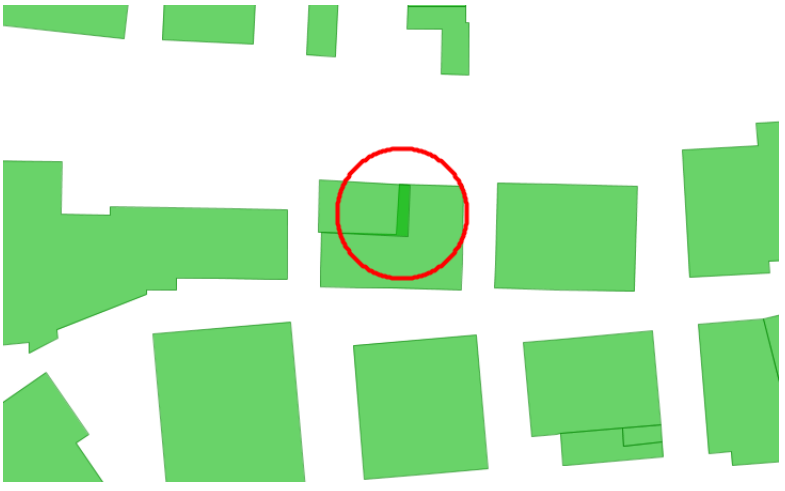

FIGURA 2.1: Esempio di errore topologico in cui due edifici si sormontano

2.2.2 Gli errori non formali

Come detto in precedenza questo tipo di errore è difficilmente individuabile, se non impossibile in alcuni casi. Questo perché il dato affetto da questo errore è comunque un dato possibile. Ora visto che questi errori per essere scoperti richiedono il confornto puntuale con la realtà, evidentemente non praticabile, un approccio possibile è quello

di ridurre lo spazio di ricerca, consentendo quindi controllare un sottoinsieme di dati in cui la concentrazione di questi errori è più alta. Questa strategia è proprio quella intrapresa in questa tesi. I dati che vengono individuati per costituire tale insieme vengono denominati *anomalie*.

2.2.3 Le anomalie

Con il termine anomalie si intende un oggetto che si discosta notevolmente dalla media degli altri oggetti, viene anche definito *outlayer*. Possiamo anche dire che le anomalie sono quei dati che nel normale senso comune possono sembrare errati, ciò non significa che lo siano realmente. Possiamo vedere due esempi, il primo in figura 2.2 che mostra un anomalia in cui i due elementi colorati in verde sono classificati come bosco. Si intuisce la stranezza del dato, dal fatto che si tratterebbe di un bosco dalle dimensioni e forme di un edificio, per di più in una zona residenziale. In questo caso si è rivelato essere effettivamente un errore. Il secondo esempio in figura 2.3 mostra due figure abbastanza regolari classificate come lago. Questa regolarità di un oggetto con forme normalmente irregolari fà sembrare il dato anomalo. In questo caso però la classificazione risulta corretta, in quanto si tratta di un lago artificiale.

FIGURA 2.2: Le aree verdi sono classificate come Bosco, è chiaro che c'è un errore di attribuzione della classe.

Con questi esempi si è potuto vedere che all'interno dell'*insieme* dei dati anomali si possono trovare sia dati affetti da errori non formali, sia dati in realtà corretti. Ma come detto precedentemente, questo andrebbe bene perché lo scopo di individuare queste anomalie è quello di individuare un sottoinsieme dei dati dove la concentrazione degli errori non formali è più alta.

Quindi gli errori che potenzialmente si andranno a correggere sono quelli contenuti nell'intersezione dei due insiemi. L'obbiettivo delle tecniche proposte è anche quello di restituire un insieme di anomalie tale da massimizzare il numero degli errori formali contenuti all'interno di essi, minimizzando allo stesso tempo il numero delle anaomalie stesse.

Per capire questo concetto si possono osservare le immagini sottostanti, in cui in 2.4 viene mostrata una generica distribuzione dei dati. In 2.5 nella stessa distribuzione vengono mostrati i dati ritenuti anomali tramite le palline nere. In 2.6 vengono evidenziati anche gli errori non formali tramite la doppia cerchiatura come si può vedere, i due insiemi non

FIGURA 2.3: Le zone in azzurro sono classificate come lago, ma la forma regolare potrebbe trarre in inganno facendo pensare ad un errore. In realta si tratta di un lago artificiale come si vede dalle immagini aeree. In questo caso non vi è alcun errore, il problema deriva dallo scarso livello di dettaglio del modello dati che raggruppa tutti i laghi sotto un unico attributo.

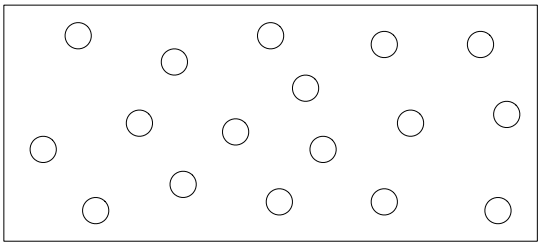

FIGURA 2.4: Generica distribuzione di dati

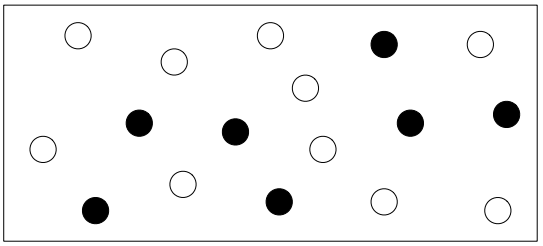

FIGURA 2.5: Evidenziazione delle anomalie

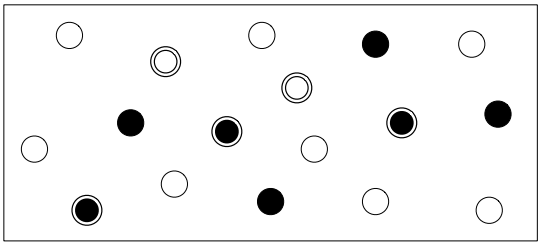

FIGURA 2.6: Evidenziazione degli errori non formali

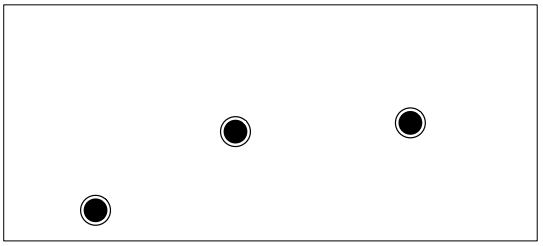

FIGURA 2.7: Errori non formali potenzialmente corretti

coincidono. Nella figura 2.7 infine sono mostrati gli errori che si riusciranno a correggere sfruttando le anomalie. Si vuole quindi individuare un insieme di anomalie tale che la sua intersezione con l'insieme degli errori sia la più ampia possibile, contemporaneamente minimizzando il numero di anomalie che non costituiscono errori formali.

2.2.4 Tassonomia delle anomalie

Ora daremo la tassonomia che è stata definita da Alan Stocco in [3] per poi successivamente ampliarla nella sottosezione successiva.

Anomalie di forma

In questa categoria ricadono tutte quelle anomalie che sono sulla forma e le dimensioni degli elementi.

- **Dimensioni coerenti con il valore semantico**: il valore semantico di un elemento, potrebbe far presupporre dei limiti di alcune dimensioni dello stesso. Un esempio potrebbe essere l'altezza dei campanili, elementi classificati come tali con un altezza di qualche metro risulterebbero anomali.
- **Forme artificiali troppo regolari e forme naturali troppo irregolari**: è l'equivalente del caso precedente solamente considerando le forme anziché le dimensioni, un esempio e quello mostrato in figura 2.3

Anomalie di distribuzione

Questa categoria non riguarda un elemento preso singolarmente, ma la distribuzione di questi su un area. Un esempio può essere la presenza di aree senza edifici in zone dove la densità di questi è molto alta, un altro esempio può essere quello dell'assenza di una chiesa entro i confini di un paese (chiaramente questa è una "regola" che deve essere contestualizzata, può andare bene nei paesi italiani ma non in altre realtà estere).

Anomalie di posizione

In questa categoria ricadono le anomalie che riguardano la posizione dell'elemento, oppure deformazioni delle geometrie:

- **Posizionamento logico**: anche in questo caso la semantica dell'elemento comanda in quali zone sia naturale trovare un oggetto ed in quali invece no. Esempi di questi errori possono essere fari in città, strutture portuali in zone di montagna oppure edifici al centro di un lago.
- **Divergenza punto**: questa categoria prende in considerazione il posizionamento dei vertici dei bordi delle geometrie. Scostamenti troppo repentini o che comunque non seguono il naturale andamento della linea sono esempi di questa categoria.

Anomalie sui grafi

In questa tipologia vengono raggruppate tutte quelle anomalie che riguardano i dati organizzati a grafo o che fanno parte di un grafo, come ad esempio le reti stradali. Possono essere suddivise in:

- **Interruzione**: in questa categoria ricadono quegli errori o anomalie in un tratto del grafo che è interrotto per poi riprendere ad una breve distanza. Con breve distanza si intende una distanza maggiore della precisione posizionale del dato ma inferiore di una certa soglia.

- **Tratto isolato**: in questa categoria ricadono i lati del grafo che sono sconnessi e con una lunghezza tale da rispettare la risoluzione, ma che comunque risultano anomali secondo logica. Un esempio può essere visto in figura 2.8.

FIGURA 2.8: Esempio di anomalia su grafo per tratto isolato.

- **Assenza vertice comune**: due lati del grafo si intersecano, senza però avere un vertice in comune, alcuni esempi possono essere:
  - un incrocio a T nel quale la strada che si interseca è stata riportata più lunga.
  - potrebbe mancare effettivamente il punto che identifica l'incrocio
  - le strade effettivamente si sovrappongono (ad esempio un cavalcavia) ma non è stata riportata l'informazione.
- **Incoerenza di tratti adiacenti**: ricadono in questa categoria gli errori di classificazione di tratti del grafo, ad esempio una strada di tipo provinciale diventa autostrada, per poi ritornare provinciale poco più avanti, tutto lungo lo stesso tratto.

2.2.5 Strumenti per l'individuazione delle anomalie

Per alcune delle anomalie presentate qui sopra in [4] sono stati proposti alcuni procedimenti per individuarle. Verranno ora brevemente illustrate.

- **Individuazione delle irregolarità nei contorni degli edifici**: Per individuare le irregolarità presenti nei bordi degli edifici, quindi che ricadono nelle anomalie di forma, è stato sviluppato un algoritmo che controlla che la lunghezza dei segmenti non sia troppo corta, e che gli angoli interni o esterni dei vertici siano sopra una certa soglia.
- **Individuazione delle forme regolari/irregolari**: L'idea alla base di questo algoritmo è l'individuazione di alcune feature geometriche che possano essere calcolate sui poligoni o sugli elementi lineari e che vengono utilizzate per discriminare la regolarità o irregolarità dell'elemento. Sono state individuate una decina di queste feature con le quali, grazie a tecniche di machine learning, si sono ricavati due alberi decisionali. Uno per gli elementi areali e uno per quelli lineari. Fornendo le opportune feature geometriche di un elemento a questi alberi si riesce a predire a quale categoria appartiene.

- **Individuazione delle anomalie di distribuzione**: Per individuare delle distribuzioni anomale di oggetti all'interno di una determinata area è stato sviluppato questo semplice algoritmo che prende come input una tipologia di oggetto da analizzare e un'area su cui eseguire la ricerca. L'area viene suddivisa in una griglia regolare e viene calcolata la media e la deviazione standard del numero di elementi appartenenti alla classe considerata all'interno di ogni cella. Vengono poi segnate le celle con il loro valore di scostamento, un esempio può essere visto in figura.

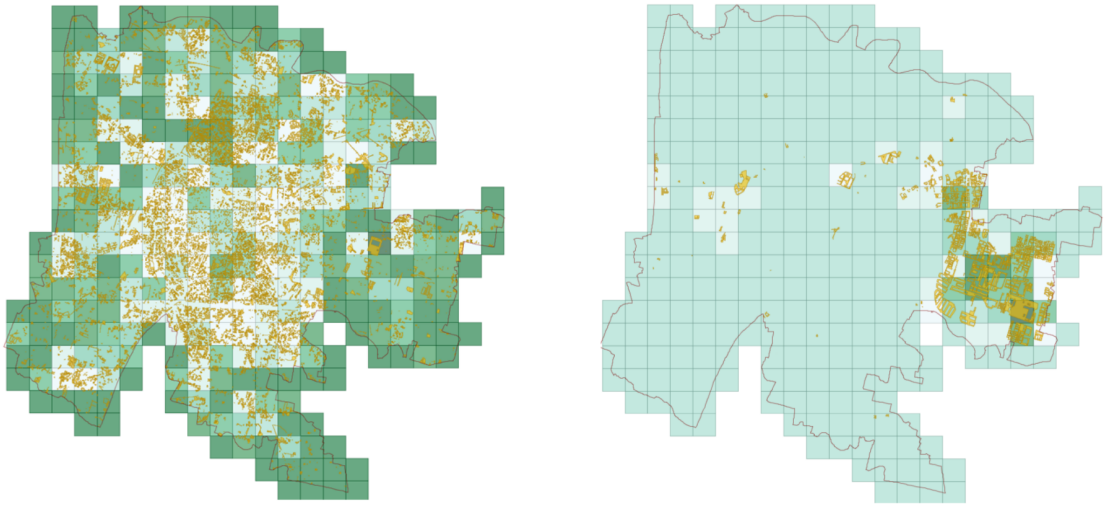

FIGURA 2.9: Risultato dell'algoritmo per l'individuazione delle anomalie di distribuzione sviluppato da Valerio Lorenzo

- **Individuazione delle divergenze dei punti**: Le divergenze si verificano quando in un elemento lineare o nel bordo di un'area si crea una improvvisa divergenza tra un vertice e i due vicini ad esso. L'algoritmo si occupa di individuare gli angoli dei vertici, come avviene per l'algoritmo di forma, dato un valore di soglia.

- **Individuazione delle anomalie di posizionamento logico**: Per questa tipologia di anomalie non è stato individuato un vero e proprio algoritmo bensì una serie di regole basate sull'esperienza. Ad esempio alcune di queste regole sono la presenza di edifici sopra dei laghi oppure chiese che non hanno un campanile vicino.

- **Individuare le interruzioni anomale nei grafi**: Questo algoritmo cerca di collegare il grafo "principale" con tutte le componenti connesse che sono abbastanza vicine al grafo principale, le restanti componenti sconnesse sono candidate anomalie.

2.2.6 Ampliamento della tassonomia

Durante lo svolgimento di questa tesi, in particolare per valutare la correttezza di classificazione degli edifici, ci si è resi conto che si poteva ampliare la tassonomia con delle anomalie che non rientravano in nessuna delle categorie elencate sopra.

Queste sono state chiamate anomalie di cambiamento e anomalie di compatibilità semantica. Qui di seguito verranno brevemente spiegate, saranno poi riprese nei capitoli

successivi mostrando come possono essere individuate e quali sono stati i risultati che si sono ottenuti sui dati reali.

Anomalie di cambiamento

Questa tipologia di anomalie può essere individuata se si ha a disposizione un' altra mappa con un contenuto paragonabile, non necessariamente con lo stesso modello dati, che possa essere in qualche modo confrontata con la mappa che si sta testando, per vedere se i vari elementi corrispondono. Naturalmente maggiore è la qualità di questa mappa migliore sarà il risultato. Può essere utilizzata anche una mappa dello stesso tipo ma più vecchia per vedere se vi sono delle anomalie nelle cose che sono cambiate, per esempio in zone del centro storico dove la demolizione di edifici è rarissima, può essere segnalata un' anomalia se nella mappa più vecchia è presente un edificio mentre nella nuova no.

Anomalie compatibilità semantica

Questa tipologia di anomalia può essere ritrovata solamente in quegli elementi che possiedono due o più attributi semantici, se questi due attributi possiedono dei valori che non sono compatibili tra loro, o che comunque appaiono strani può essere segnalata una anomalia. Un esempio concreto che è stato trovato durante gli esperimenti di questa tesi è l'associazione degli edifici di tipologia campanile e destinazione d'uso pinacoteca.

Capitolo 3

Metodi statistici per l'individuazione delle anomalie

3.1 Introduzione

Come abbiamo visto dal capitolo precedente, fino ad ora l'approccio che si è utilizzato per individuare le anomalie è stato quello di ideare ed implementare delle regole basate sull'esperienza di persone del settore. Quello che si è cercato di fare in questa tesi invece è stato ideare e testare tecniche per l'individuazione di tali regole partendo dai dati e cercando di minimizzare la conoscenza a priori di questi.

Partendo da una definizione alternativa di anomalie data in [5]:

"Un anomalia è un oggetto che differisce in modo significativo dal resto degli altri oggetti, come se fosse stato generato da un differente meccanismo."

Per poter individuare queste anomalie si deve capire quali sono i pattern o come si comportano i dati "normalmente". Alcuni strumenti statistici permettono di eseguire delle analisi e generare dei modelli che possono essere utilizzati per questo scopo. Infatti è proprio quello che è stato fatto in questa tesi. Questo tipo di tecniche vengono anche chiamate anomaly detection o outlayer detection, e sono utilizzate in altri campi come ad esempi le frodi bancarie.

In questa tesi si sono utilizzate differenti tecniche per la costruzioni dei modelli statistici, e gli aspetti che sono stati ricavati sono anch'essi molteplici. Nei paragrafi che seguono saranno presentati tutti i modelli che sono stati pensati, da un punto di vista generale e indipendente dalla strutturazione dei dati (modello dati). Nel capitolo successivo invece saranno presentati i risultati dell'applicazione di questi approcci su dati reali. Quello che si è cercato di fare è stato di sviluppare dei metodi che, analizzando i dati, riescano in qualche modo ad estrapolare delle "regole" per l'individuazione delle anomalie. Un altro importante obbiettivo che si è voluto raggiungere è stato quello dell'indipendenza dal modello dati. Si è cercato di estrarre delle informazioni direttamente dai dati cercando di ridurre al minimo le conoscenze sui dati stessi. Questo perché si vuole che queste tecniche possano essere applicate in qualsiasi modello dati.

La maggior parte degli approcci utilizzati e descritti di seguito utilizzano strumenti statistici ma non solo, infatti sono state utilizzate anche delle tecniche di machine learning.

3.2 Caratterizzazione tramite elementi presenti nel vicinato

In questa sezione verrà spiegato come si è pensato di utilizzare gli elementi nei dintorni di un elemento per caratterizzare la classe dell'elemento stesso. Per fare ciò si presenteranno due tecniche, la prima prenderà in considerazione il numero degli elementi vicini divisi per classi. Per ognuna delle classi si guarda se questa risulta essere un fattore discriminante sufficiente. Il secondo approccio invece cerca di utilizzare tutte le informazioni del numero di elementi vicini, cercando di estrarre un modello con tecniche di machine learning.

3.2.1 Notazione e Formalismo

Per cominciare diamo la notazione per ciò che andremo a trattare in questa sezione. I dati spaziali che analizzeremo sono composti da un insieme di Feature, ognuna delle quali può essere definita come:

$$f_j^i = \{g_j, a_1^i, \cdots, a_m^i\}$$

l'insieme di una geometria g_j più un numero m di attributi a_t^i ognuno dei quali avente un dominio di diverso tipo. In questa sezione considereremo solamente quelli con un dominio a valori enumerati, quindi:

$$a_t^i \in [v_1^{a_t^i}, \ldots, v_l^{a_t^i}]$$

Dove $v_p^{a_t^i}$ sono i valori che possono essere assunti dall'atributo a_t^i

Le feature sono suddivise in classi, che possono essere formalizzate come un insieme di queste:

$$\mathcal{C}_i = \{f_1^i, \cdots, f_n^i\}$$

Feature cha appartengono alla stessa classe condividono gli stessi attributi.

Un altro importante concetto che varrà utilizzato è quello del vettore del *vicinato*

$$\mathcal{N}_{f_j^i}^{dist} = \{n_{v_1}^{a_{t_1}}, \cdots, n_{v_{l_1}}^{a_{t_1}}, \cdots, n_{v_1}^{a_{t_b}}, \cdots, n_{v_{l_b}}^{a_{t_b}}\}$$

in cui dato un intorno di raggio *dist* centrato in f_i^j il generico elemento n_v^a corrisponde al numero di feature non più lontane di *dist* da f_j^i, che hanno nell'attributo a il valore v. Per capire meglio questo concetto si può guardare la figura3.1. Come si vede il vettore, in questo caso associato all'elemento a forma di rombo al centro, conta quanti elementi con determinate classificazioni sono contenute all'interno del raggio di ricerca.

Un errore di classificazione può essere sia errata assegnazione di una feature ad una classe $\mathcal{C}_i$ sia un errato valore dell'attributo a_t^i.

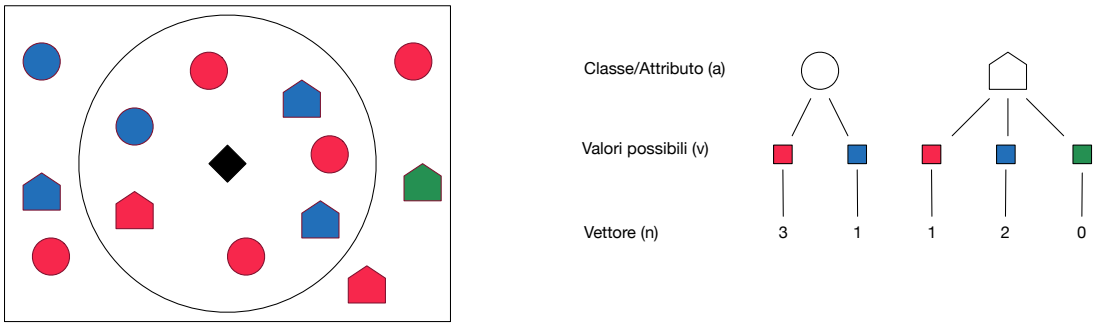

FIGURA 3.1: Rappresentazione stilizzata di come viene popolato il vettore del vicinato per per l'elemento *rombo*. Come si vede il vettore, in questo caso associato all'elemento a forma di rombo al centro, conta quanti elementi con determinate classificazioni sono contenute all'interno del raggio di ricerca.

3.2.2 Caratterizzazione Statistica

Utilizzando i singoli valori del vettore del vicinato $\mathcal{N}_{f_j^i}^{dist}$, si può provare a caratterizzare la classe o i valori della feature f_j^i ad esso associata. Ciò equivale a valutare se esistono delle tipologie di feature che possono influenzare le sue vicine.

Per una miglior comprensione di questo concetto, verrà ora fatto un esempio emblematico che prende in considerazione due categorie di edifici fortemente correlate: Il campanile e la chiesa. La presenza di un campanile nelle immediate vicinanze di un edificio implica che, con alta probabilità, questo sia una chiesa. Naturalmente vale anche il viceversa. Inoltre si può trovare un'anomalia anche nel caso in cui un elemento non sia presente nel vettore del vicinato, utilizzando sempre l'esempio sopra se si vede una chiesa che non ha un campanile nelle vicinanze.

Per valutare se vi siano altri casi simili si devono controllare le distribuzioni degli elementi nei dintorni dei tipi di feature che si vogliono caratterizzare. Possiamo organizzare le distribuzioni secondo una matrice in cui le righe sono associate ai tipi di feature da caratterizzare. Mentre le colonne sono associate ai tipi di feature utilizzati per la caratterizzazione. La generica cella ij quindi conterrà la distribuzione del numero delle feature di tipo j nei dintorni delle feature di tipo i. Possiamo rappresentare le distribuzioni in vario modo, supponiamo per esempio di utilizzare l'intervallo formato da media $\mu \pm$ deviazione standard σ. calcolati in questo modo:

$$\mu = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \mu)^2}{n-1}}$$

Utilizzando questa matrice possono essere fatti due tipi di analisi per l'individuazione delle anomalie: valutando la singola cella, oppure valutando le colonne. Valutando la generica cella ij si può controllare se la distribuzione associata ha una bassa deviazione standard. In questo caso si può dire che se una feature f_i di tipo i ha un numero di feature f_j di tipo j che si discosta dalla media di più della deviazione standard si può

ritenere un anomalia. Valutando invece le colonne si controllano tutti gli intervalli della colonna. Se ce n'è uno, ad esempio quello associato con la cella ij che non si interseca con i restanti allora possiamo dire che se una feature f_x, di tipo incognito x, ha nei suoi dintorni un numero di feature f_j di tipo j che sta nell'intervallo ij allora si può dire che $x = i$. Se cosi non fosse si avrà un anomalia.

Un altro utile strumento grafico che potrebbe essere utilizzato in questo caso per la valutazione delle distribuzioni è sicuramente il grafico a box. Questo particolare grafico che può essere visto in figura 3.2 mostra cinque valori statistici che fanno capire la distribuzione dei dati:

- La linea che si trova all'interno del rettangolo è la mediana dei valori.
- I limiti del rettangolo, indicati con Q1 e Q3 in figura 3.2 indicano le mediane dei valori contenuti negli intervalli formati dal valore minimo e la mediana,da la mediana e il valore massimo.
- I due baffi che escono dal rettangolo danno una misura della dispersione dei dati, e sono posizionati ad una distanza di $1.5 \times IQR$ dai lati in cui escono. IQR si riferisce alla larghezza del rettangolo.

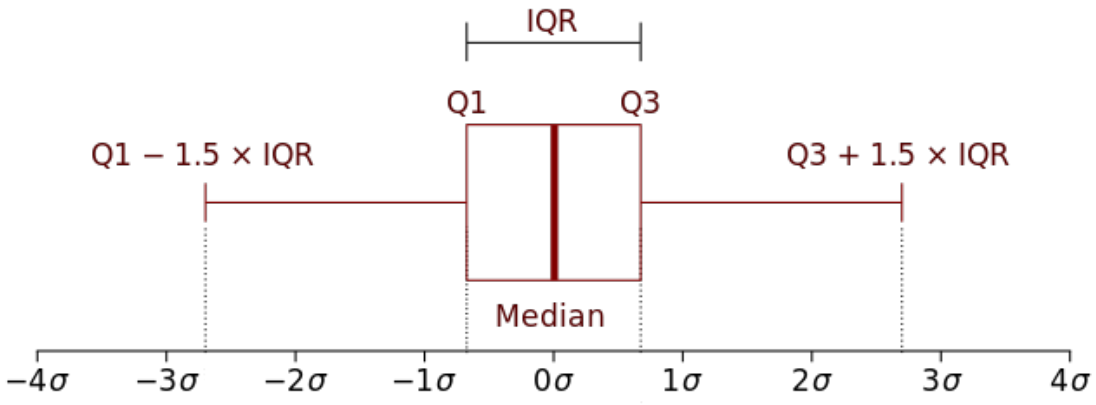

FIGURA 3.2: Esempio di grafico a box

Questo grafico può essere utilizzato in un modo simile a quello spiegato precedentemente con la media e la deviazione standard. Plottando assieme tutti i box, se ce ne sono alcuni che sono disposti in intervalli di valori dove non sono posizionati altri box, allora il valore che rappresenta può essere utilizzato per la caratterizzazione.

3.2.3 Utilizzo di tecniche di machine learning : reti neurali

L'approccio considerato in precedenza tiene in considerazione solamente i valori delle calle del vettore prese singolarmente, questo equivale a considerare che questi valori siano indipendenti. Potrebbero esserci invece delle relazioni molto più complicate che permettono anch'esse la discriminazione di classi e attributi. Per scoprire queste relazioni complicate un buon approccio è quello di utilizzare tecniche di apprendimento automatico. Questi sistemi possono essere "allenati" fornendo un certo numero di esempi, e quindi costruire un modello che riesca a individuare in modo autonomo dei pattern che potrebbero essere presenti nei dati, che possono essere utilizzati per i nostri scopi.

Nel caso qui trattato, in cui dato un vettore del vicinato si vuole che venga restituito la classe o il valore dell'attributo della feature ad esso associato, potrebbe essere una buona scelta l'utilizzo delle reti neurali.

Le reti neurali artificiali sono dei particolari modelli di apprendimento ispirate al collegamento dei neuroni nel cervello. Servono per stimare una funzione con un arbitrario numero di input che può essere anche molto grande. Questa capacità delle reti neurali è stato uno dei motivi, per cui si è scelto di utilizzare questa tipologia di apprendimento automatico. La rete neurale è composta da una serie di strati, un primo strato contenente i nodi di input, uno strato finale composto da i nodi di output più una serie di strati interni chiamati livelli nascosti. Ognuno di questi livelli contiene un certo numero di nodi chiamati anche neuroni. Ognuno di questi nodi è collegato con tutti i nodi del precedente livello. Ad ogni collegamento è associato un peso che inizialmente ha un valore di default. Quando l'informazione arriva dai vari collegamenti viene pesata con i relativi pesi e sommata. Il risultato di questa somma passa per una funzione di attivazione che genererà un segnale che il nodo propagherà ai livelli successivi. Questo processo si ripeterà contemporaneamente in tutti i nodi di un livello nascosto, e si propagherà fino ad arrivare al livello di output. L'algoritmo che esegue l'allenamento della rete neurale inserirà l'input e controllerà che l'output sia corretto, se cosi non fosse modifica i pesi di tutti i collegamenti tra neuroni per correggere l'errore.

Lq rete neurale richiede un processo di tunning di vari parametri, che sono il layout della rete piu la scelta della funzione di attivazione.

3.3 Utilizzo di descrittori geometrici per la discriminazione delle classi

Dato che i dati geografici possiedono una rappresentazione geometrica è utile scoprire se esiste una correlazione tra la loro forma e la loro classe di appartenenza, per fare ciò devono essere trovati dei descrittori geometrici che permettano di caratterizzare al meglio la geometria.

Qui di seguito verranno elencati i descrittori utilizzati:

- **Area, perimetro, altezza, numero di lati**
- **Verticalità**: il rapporto tra la l'altezza e l'area dell'edificio, questa misura da un idea di quanto un edificio si sviluppa verso l'alto.
- **Squareness**: misura che indica se un oggetto assomiglia ad un quadrato. Calcolato come $\frac{16Area}{Perimetro^2}$ và da 0 per i segmenti a 4/p per i cerchi (1 per quadrati). Bassi valori sono influenzati da concavità ed elongazione;
- **Spigolosità**: percentuale del numero di angoli interni sotto una certa soglia rispetto al numero di vertici.
- **Perpendicolarità**: percentuale del numero di angoli retti rispetto al numero di vertici (vertici contati come in precedenza). Varia tra 0 e 1.
- **Compattezza**: indice che misura quanto una figura è compatta, calcolata come $\frac{Perimetro^2}{Area}$.

- **Eccentricità**: rapporto tra l'asse minore e l'asse maggiore dove l'asse maggiore coincide con la linea su cui è stato valutato il diametro del contorno, l'asse minore è una linea perpendicolare all'asse maggiore e di lunghezza tale che il rettangolo passante per i quattro estremi dei due assi contiene completamente il contorno.
- **Convessità**: misura di convessità ed è calcolata come il rapporto tra area della figure e l'area del convex hull.
- **Simmetria**: un indice che misura quanto è simmetrica una figura, l'algoritmo per ottenere questo indice è stato preso da [6]. L'algoritmo prende come input una figura geometrica e un numero intero pari al numero di assi di simmetria con cui si vuole fare il test. I punti della figura originale vengono ridistribuiti in modo da formare la figura più "vicina" con un numero di simmetrie rotazionali pari a quello specificato in input. Viene infine calcolata la somma delle distanze tra la posizione dei vertici della figura originale e i corrispondenti nella figura simmetrica.

Tutti questi indici posso essere raccolti ed analizzati per cercare di valutare se è possibile discriminare alcune classificazioni degli elementi. Chiaramente questa analisi ha senso soprattutto per gli attributi che hanno una qualche relazione logica con la geometria, per esempio nei test eseguiti in questa tesi si è provato a discriminare la tipologia edilizia degli edifici, che dà una descrizione, seppur grossolana, di come l'edificio è fatto.

Queste valutazioni possono essere fatte utilizzando dei modelli statistici ottenuti dai dati stessi. Per poter generare questi modelli possono essere utilizzate due strade, la prima consiste nell'utilizzare modelli standard come la distribuzione Gaussiana e quindi ricavare, tramite delle formule note, i parametri necessari alla realizzazione del modello. Oppure utilizzando delle tecniche per la determinazione della funzione di distribuzione, direttamente dai dati. Queste tecniche vengono chiamate *Kernel Density Estimation* (KDE) [7] e servono per la determinazione, in maniera non parametrica, della funzione di densità di probabilità di una variabile casuale da un numero finito di esempi di tale variabile. Per stimare la densità di distribuzione viene calcolata la seguente funzione:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

in cui K è la funzione di kernel, che è una finzione a media zero che nella formula viene posizionata ad ogni valore dell'esperimento e sommata al resto. Nella figura 3.3 viene illustrato in modo più chiaro questo concetto.

Questo tipo di stima è ben rappresentabile con gli istogrammi degli esperimenti, infatti vine anche considerato lo smussamento di quest'ultimo per ottenere la densità di probabilità.

Una volta ottenute tutte le funzioni di distribuzione di un particolare indice suddivisi per le possibili categorie, per vedere se è possibile fare una discriminazione, si deve guardare se qualcuna di queste non intersechi o comunque intersechi in minima parte le altre. In questo caso se si trovasse un elemento il cui valore dell'indice sia all'interno dei valori della funzione di densità di probabilità allora si potrebbe dire che quell'elemento dovrebbe appartenere alla categoria associata alla funzione di distribuzione, se non è cosi può considerarsi un anomalia.

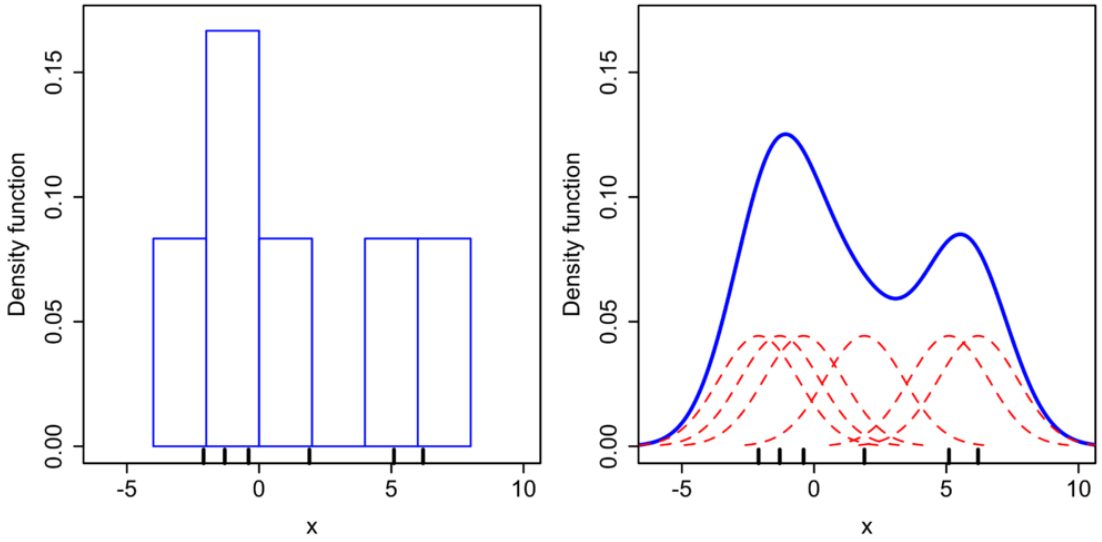

FIGURA 3.3: In questa figura viene mostrata come la funzione di distribuzione ottenuta tramite kde sia una sorta di levigatura dell'istogramma, e come i kernel (le campane rosse) centrati nei vari campioni, sommati diano la curva blu

3.4 Compatibilità semantica

Questo procedimento può essere applicato solamente per quelle feature che possiedono più di un attributo che fornisca una classificazione. L'idea alla base di questo approccio è che eventuali associazioni tra coppie di questi attributi che non siano molto ricorrenti e che quindi, presumibilmente si trovino collegati a causa di un errore. Per fare ciò si deve creare una tabella di contingenza che conta il numero delle ricorrenze di tutte le possibili combinazioni dei valori degli attributi, e vengono considerate come anomalie tutte quelle con cardinalità molto bassa. La tabella delle contingenze è una tabella con doppia etichettatura (cioè sia sulle righe che sulle colonne) che mette in relazione due o più variabili. In questo caso le variabili sono gli attributi di classificazione delle feature considerate.

3.5 Difference Detection

Un altro approccio per l'individuazione delle anomalie consiste nell'utilizzare un ulteriore mappa che abbia un contenuto informativo paragonabile alla mappa da tastare ed utilizzarla come confronto per trovare le differenze. Le due mappe devono preferibilmente aver avuto un processo produttivo differente, almeno per le informazioni che si vogliono testare, in questo modo si può affermare con alta probabilità che gli errori che si trovano in una delle due mappe non sono presenti nell'altra. Confrontando le due mappe possono essere trovate delle non corrispondenze che sono interpretate come anomalie. Chiaramente solo dopo aver controllato i dati reali si potrà stabilire in quale delle due mappe è presente l'errore. Il confronto delle mappe non è comunque un procedimento così immediato infatti richiede alcuni accorgimenti a seconda delle caratteristiche delle due mappe che verranno analizzati qui di seguito:

- **Modello dati differente**: nel caso in cui il modello dati sia differente, cosa molto probabile, deve essere eseguita una mappatura tra le classi equivalenti. Potrebbe capitare che una mappa abbia una classe che contenga degli elementi che nell'altra sono classificati in più classi. Questo potrebbe essere causato da un maggiore livello di dettaglio della seconda mappa.

- **Sistemi di riferimento, risoluzioni differenti**: se le cartografie hanno un differente sistema di riferimento, oppure hanno una diversa risoluzione, si dovranno fare delle ulteriori elaborazioni per capire quali sono gli edifici corrispondenti nelle due mappe. Esistono dei tool specifici per cambiare sistema di riferimento e proiezione, un esempio è pro4j. Può capitare che le mappe non coincidano perfettamente questo potrebbe essere dato da una serie di fattori, ad esempio l'utilizzo di strumenti di rilevazione differenti e con differenti precisioni. Oppure le due mappe potrebbero avere un livello di dettaglio differente, un edificio potrebbe essere suddiviso in più unità volumetriche da una parte mentre, come un unico edificio dall'altra. O ancora si utilizzano mappe a scala differente. Il risultato in entrambi i casi è che gli elementi nelle due mappe non si sovrappongono in modo corretto in figura3.4 si può notare molto bene questo fenomeno. Problemi di questo tipo vengono definiti

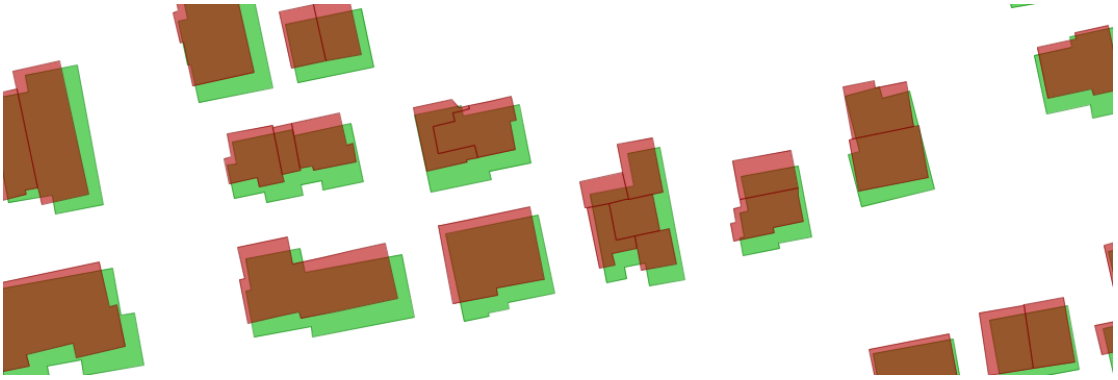

FIGURA 3.4: Problemi di sovrapposizione tra due mappe con differente dettaglio e differente sistema di riferimento

  problemi di conflaction, per risolverli possono essere utilizzate differenti strategie a seconda dei casi. Per esempio le feature con geometria areale di una delle mappe possono essere trasformate nei rispettivi centroidi, cosi da ricadere in al più una delle geometrie dell'altra mappa. Un altro procedimento più sofisticato potrebbe essere valutare la percentuale di area occupata dalle feature che si sovrappongono parzialmente e valutare con delle soglie se le due feature rappresentano lo stesso elemento.

Difficilmente si troveranno delle mappe contemporanee della stessa zona provenienti da fonti differenti. Il motivo principale è sicuramente l'enorme costo di produzione che queste hanno. Possono essere utilizzate comunque altre tipologie di dati che comunque possiedono un minimo di localizzazione geografica. Un esempio può essere un l'elenco di negozi con le rispettive coordinate. L'utilizzo di mappe non contemporanee è invece un caso più plausibile. Questa particolare forma di difference detection viene chiamata change detection.

3.5.1 Change detection

Nel caso in cui le mappe non siano contemporanee, cioè abbiano molti (>5) anni di differenza nella data di creazione, si parla di change detection. In questo caso oltre alla differenza di modello che potrebbe esserci tra le due mappe vi è la possibilità che realmente gli elementi siano cambiati nel tempo. Inoltre alcuni elementi potrebbero addirittura mancare, perché sono stati eliminati, oppure perché sono stati costruiti. Questo fatto può essere utilizzato per la determinazione di cambiamenti anomali di questo tipo, ad esempio demolizioni di edifici dei centri storici, fatto estremamente raro che può benissimo essere considerato un anomalia.

Capitolo 4

Implementazioni e risultati

In questo capitolo verranno illustrate le implementazioni dei metodi descritti nel capitolo precedente. Verrà fatta un'analisi sui dati che sono stati utilizzati e verranno mostrati i risultati che si sono ottenuti, utilizzando i vari modelli come indice di normalità per valutare la presenza di anomalie nei dati.

4.1 Dati utilizzati

I dati che sono stati analizzati sono di alcune delle aree del veneto, possono essere liberamente scaricati dal sito del geoportale della regione veneto http://idt.regione.veneto.it/. Per la precisione sono stati utilizzati i dati del DBTopografico, più le mappe della CTRN nel caso della change detection.

4.1.1 DB Topografico : GeoDBR

Il GeoDBR è un modello dati sviluppato dalla Regione Veneto nell'ambito di un progetto per l'aggiornamento del proprio sistema informativo territoriale, realizzato secondo le specifiche definite all'interno del progetto IntesaGis. L'organizzazione degli oggetti all'interno del DB topografico è realizzata mediante l'uso di classi, a sua volta raggruppati in temi (come ad esempio strade, edificato). I temi a loro volta sono raggruppata in strati. Ad esempio nello strato "Viabilità, mobilità e trasporti" è presente il tema "Strade" e la classe "Area di circolazione veicolare". Tale gerarchia non compare in modo esplicito nel DB topografico, la cui struttura contiene solamente classi. Ad ogni oggetto geografico vengono associati gli opportuni attributi, la componente spaziale (realizzata tramite primitive geometriche, cioè punto, linea ed area, in base alla loro dimensione e forma) e i vincoli topologici. È richiesta la copertura totale del territorio in forma topologica e, tranne qualche eccezione, non ci devono essere né sovrapposizioni né buchi nell'informazione. Le proprietà degli oggetti geografici sono esplicitate tramite delle codifiche numeriche; in particolare nel modello ad ogni classe è attribuito un codice di 6 cifre, agli attributi è assegnato il codice di classe più ulteriori due cifre finali, mentre ai valori degli attributi è assegnato il codice dell'attributo più ulteriori due cifre: anche i codici presentano quindi una struttura estremamente gerarchica.

4.1.2 CTRN

Questi sono i dati cartografici precedenti al Data Base Topografico, sono carte per la maggior parte con scale 1:5000 e per alcune zone 1:10000. Gli oggetti e le informazioni territoriali contenute nella Carta Tecnica Regionale, acquisiti in forma vettoriale, sono organizzati in livelli e codici: i livelli costituiscono una primaria classe di aggregazione degli oggetti, che a loro volta sono suddivisi nei codici, relativi alle caratteristiche particolari di ciascun oggetto. In totale sono presenti 16 livelli principali, 12 livelli di servizio e 6 livelli funzionali per la gestione informatica dei grafi (assi e nodi di viabilità, idrografia e ferrovia); ciò consente la codifica di 480 oggetti ed informazioni. Le cartine sono suddivise in fogli i quali, non sono altro che una suddivisione a griglia del territorio, questa suddivisione riguarda anche gli elementi della mappa nel senso che, quelli che si trovano a cavallo di due fogli sono stati tagliati.

I dati della CTRN, però, non si prestano bene all'analisi spaziale e ad un diretto utilizzo, in quanto sono realizzati prevalentemente tramite tecniche CAD, e perciò non offrono alcuna forma di controllo di coerenza topologica. Questo fatto si ripercuote nella necessità di attuare una lunga fase di controllo e pulizia dei dati.

4.1.3 Zonizzazione del territorio: Corine Land Cover

Come già accennato nei capitoli precedenti i dati geografici hanno sicuramente una correlazione con la zona del territorio in cui si trovano. Quindi fare un'analisi in base alle diverse zone ha sicuramente più senso che non trattare un modello generale per la totalità dei dati. Purtroppo nel GeoDBR non è prevista alcuna informazione che possa essere utilizzata per caratterizzare il territorio. Si è quindi pensato di utilizzare le informazioni del Progetto Corine Land Cover (CLC). Questo progetto è nato a livello europeo specificatamente per il rilevamento e il monitoraggio delle caratteristiche di copertura e uso del territorio, con particolare attenzione alle esigenze di tutela ambientale. Secondo le specifiche il territorio viene suddiviso in una gerarchia di classi che partono da cinque classi generali le quali si sviluppano ad albero fino a tre livelli di profondità nell'ultima specifica del 2006. Nella tabella 4.1 sono riportati tutti i livelli previsti.

In base alla suddivisione delle aree Corine sono stati suddivisi anche i dati, e di conseguenza anche le analisi che sono state fatte. Quindi per ognuno di questi insiemi è stato ricavato un differente modello statistico. Il vantaggio è stato quello di ottenere dei modelli più precisi in quanto le concentrazioni dei vari elementi risultano essere più omogenee e quindi con una minore varianza. Il tipo di suddivisione fatto da Corine risulta essere ottimo per i test fatti in questa tesi, che prendono in considerazione prevalentemente gli edifici. La suddivisione è stata fatta utilizzando il seconodo livello di Corine perché fornisce un giusto compromesso tra la dimensione della zona e il dettaglio relativo all'uso.

4.1.4 Problemi relativi ai dati da modellare

La costruzione dei modelli utilizzati richiede l'elaborazione di un numero abbastanza elevato di dati e che siano anche rappresentativi per ciò che si vuole analizzare. Per alcuni dei dati geografici questi prerequisiti potrebbero non essere soddisfatti, un esempio concreto può essere visto nelle categorizzazioni degli edifici. Indipendentemente da come

TABELLA 4.1: Classi della Corine Land Cover, Suddivise per i tre livelli gerarchici

Level 1	Level 2	Level 3
1. Artificial Surface	11 Urban fabric	111 Continuous urban fabric
		112 Discontinuous urban fabric
	12 Industrial, commercial and transport units	121 Industrial or commercial units
		122 Road and rail networks and associated land
		123 Port areas
		124 Airports
	13 Mine, dump and construction sites	131 Mineral extraction sites
		132 Dump sites
		133 Construction sites
	14 Artificial, non-agricultural vegetated areas	141 Green urban areas
		142 Sport and leisure facilities
2 Agricultural areas	21 Arable land	211 Non-irrigated arable land
		212 Permanently irrigated land
		213 Rice fields
	22 Permanent crops	221 Vineyards
		222 Fruit trees and berry plantations
		223 Olive groves
	23 Pastures	231 Pastures
	24 Heterogeneous agricultural areas	241 Annual crops associated with permanent crops
		242 Complex cultivation patterns
		243 Land principally occupied by agriculture, with significant areas of natural vegetation
		244 Agro-forestry areas
3 Forest and semi natural areas	31 Forests	311 Broad-leaved forest
		312 Coniferous forest
		313 Mixed forest
	32 Scrub and/or herbaceous vegetation associations	321 Natural grasslands
		322 Moors and heathland
		323 Sclerophyllous vegetation
		324 Transitional woodland-shrub
	33 Open spaces with little or no vegetation	331 Beaches, dunes, sands
		332 Bare rocks
		333 Sparsely vegetated areas
		334 Burnt areas
		335 Glaciers and perpetual snow
4 Wetlands	41 Inland wetlands	411 Inland marshe
		412 Peat bogs
	42 Maritime wetlands	421 Salt marshes
		422 Salines
		423 Intertidal flats
5 Water bodies	51 Inland waters	511 Water courses
		512 Water bodies
	52 Marine waters	521 Coastal lagoons
		522 Estuaries
		523 Sea and ocean

sono state rappresentate le categorie nel modello dati, indubbiamente vi saranno alcune di queste che conterranno un numero molto alto di elementi (e.g. edifici residenziali) ed altre che ne conterranno un numero molto basso, in figura 4.1 sono mostrati un numero di edifici per ogni categoria d'uso nella zona a nord della provincia di Padova. Il grafico è in scala logaritmica, quindi si può vedere che ci sono quasi 4 ordini di grandezza tra la classe che contiene più elementi e quella che ne contiene di meno. Questo fatto deve essere tenuto in considerazione soprattutto nelle tecniche di machine learnig.

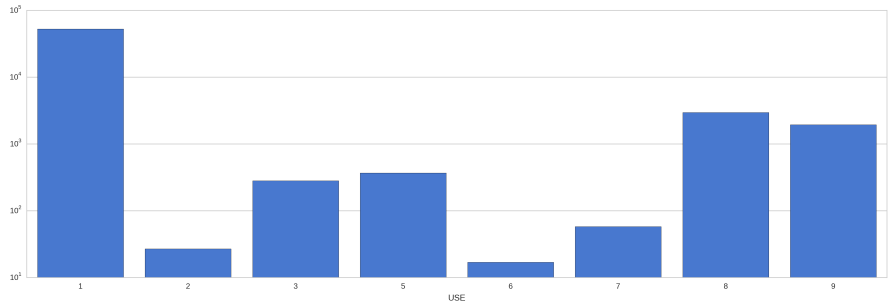

FIGURA 4.1: Distribuzione del numero di edifici divisi per categoria d'uso su scala logaritmica. La zona raffigurata è la zona nord della provincia di Padova.
Leggenda : (1) Residenziale, (2) Amministrativo, (3) Servizio pubblico, (5) Luogo di culto, (6) Servizi di trasporto, (7) Commerciale, (8) Industriale, (9) Agricolturale

La presenza dei tipi di errore che si sta cercando di correggere, comporta l'impossibilità di stabilire con certezza se i dati che stiamo analizzando sono totalmente corretti. Perché se così fosse esisterebbe già un modo per valutare la correttezza non formale dei dati.

Questo potrebbe essere un problema per la costruzione dei modelli. La soluzione consisterebbe nella costruzione a mano di un dataset con dati completamente testati, ma questa possibilità è di fatto impossibile in quanto il numero di dati necessario è molto elevato, ed inoltre dato che questi modelli sono altamente contestualizzati dovrebbe essere realizzato un dataset per ogni zona che si ritenga avere una distribuzione differente nei dati. L'unica possibile alternativa, è l'utilizzo dei dati con gli errori, questo può essere fatto solo nell'ipotesi che il numero degli errori sia molto basso. La sicurezza che questa affermazione sia vera deriva da diversi fattori, il primo è lo stesso che garantisce l'esatto opposto, e cioè che il processo di costruzione delle mappe è fortemente manuale. Infatti la continua supervisione di un essere umano durante il processo produttivo garantisce che la qualità del risultato sia abbastanza alta, sicuramente non perfetta. Inoltre i dati su cui questi controlli dovranno essere eseguiti sono passati per una fase di collaudo in cui vengono fatti dei controlli a campione che riescono a garantire degli errori inferiori a qualche punto percentuale.

4.2 Caratterizzazione del vicinato

Per testare le tecniche di caratterizzazione del vicinato sono state utilizzate le classi riguardanti i principali edifici del GeoDBR, qui di seguito vengono riportate le classi utilizzate e le definizioni date nelle specifiche intesa:

- **EDIFC**: Si intende un corpo costruito che non presenta soluzione di continuità, ha un'unica tipologia edilizia, può avere più categorie d'uso ha un dato stato di conservazione e può eventualmente essere sotterraneo. Questa classe contiene due attributi di classificazione:
  - **Tipologia Edilizia (EDIFC_TY)**: specifica le caratteristiche strutturali di un edificio, definisce la tipologia di edificio desumibile dalla ripresa aerofotogrammetrica e quindi dall'osservazione della pianta dell'edificio.
  - **Destinazione d'uso (EDIFC_USO)**: specifica le varie destinazioni d'uso di un edificio.
- **EDI_MIN**: nelle specifice intesa questa classe è descritta contenente tutti quegli edifici minori che partecipano alla definizione del territorio antropizzato in quanto costruzioni che integrano e supportano l'edificato e le attività dell'uomo, caratterizzati dalla permanenza non continuativa delle persone. Per esempio baracche chioschi garage ecc.
- **MN_EDI**: Sono definiti in questa classe i manufatti di varia natura accessori allo sviluppo di attività o servizi industriali, all'interno di aree specifiche o opportunamente recintati.

Quello che si è provato a caratterizzare sono state queste classi associate ad uno dei loro attributi, per quanto riguarda EDIFC è stato utilizzato EDIFC_USO, mentre per le altre due classi è stato utilizzato l'unico attributo che serve ad indicare il loro tipo cioè rispettivamente EDI_MIN_TY e MN_EDI_TY.

4.2.1 Composizione del vettore del vicinato

Come detto nel capitolo precedente, il vettore del vicinato è composto da tutti i possibili valori degli attributi che caratterizzano le classi scelte. In sostanza ogni casella è associata ad una particolare tipologia di elemento e il valore al suo interno conta quanti di questi elementi sono presenti entro una certa distanza dalla feature cui è associato il vettore. La composizione del vettore usato in queste prove prende in considerazioni diverse classi, più precisamente oltre alle tre esposte nel paragrafo precedente sono state utilizzate anche:

- **AB_CDA**: aree bagnate da corsi d'acqua.
- **AC_VEI**: aree di circolazione veicolare.
- **AR_VRD**: aree adibite a scopo ornamentale o inserite in aree ricreative.
- **CL_AGR**: colture agricole.

Tutte le classi utilizzate possiedono un solo attributo di classificazione a parte EDIFIC che come già detto ne ha due.

Per la generazione del vettore del vicinato è stato seguito per ognuna delle feature il seguente procedimento:

- **Bufferizzazione**: Si è eseguito un buffer attorno alla feature che si sta considerando, l'ampiezza del buffer corrisponde alla distanza associata al vettore che si vuole ottenere. In figura 4.2 il buffer è mostrato in verde e appartine all'edificio posto nel suo centro.
- **Join Spazziale**: Utilizzando il buffer del passo precedente viene eseguito un join spaziale tra questo e tutti gli elementi della mappa che si vogliono utilizzare per caratterizzare le feature.
- **Enumerazione**: vengono contati quanti sono gli elementi appartenenti alle differenti tipologie, e viene generato il vettore.

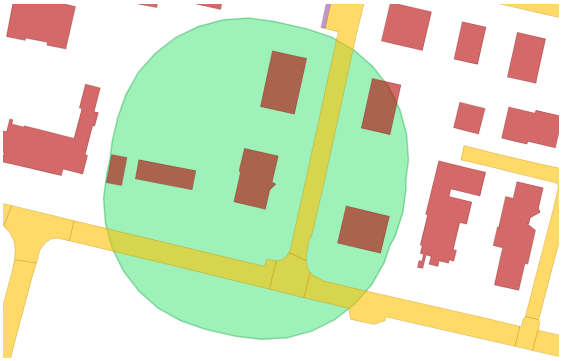

Figura 4.2: Processo di generazione del vettore del vicinato

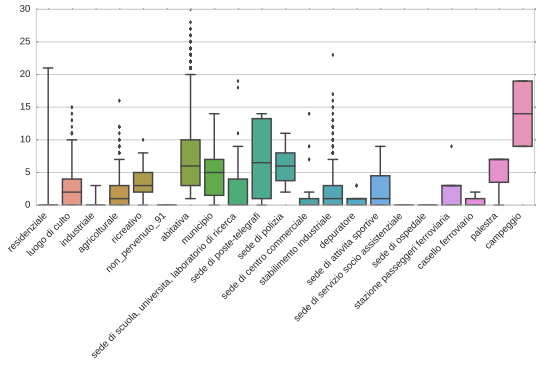

FIGURA 4.3: Grafico a box rappresentante la distribuzione del numero degli edifici abitativi nel raggio di 50 metri dagli edifici indicati in ascissa.

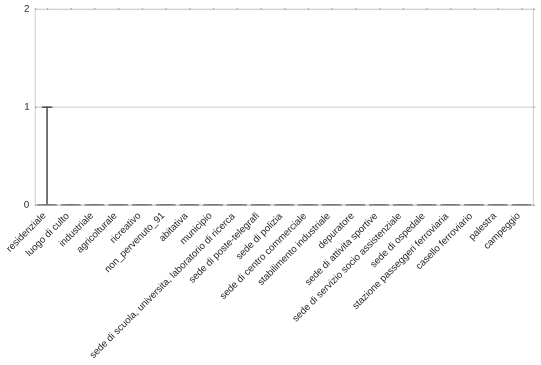

FIGURA 4.4: Caso di grafico a box che potrebbe generare una regola, ma che in realtà è frutto di un errore di classificazione.

4.2.2 Caratterizzazione statistica dei singoli elementi

Un primo controllo che si è provato ad eseguire è stato quello di valutare se vi siano delle tipologie di elementi che, se presenti nel vettore del vicinato, riescano a caratterizzare le possibili classificazioni dell'oggetto cui il vettore appartiene. L'esempio emblematico (frutto dell'esperienza e non supporatato da nessun dato) che ha fatto pensare a questo approccio è quello della chiesa e del campanile, cioè il fatto che se un edificio nei suoi dintorni prossimi ha un altro edificio classificato come campanile, allora molto probabilmente il primo sarà una chiesa. Per poter valutare se una particolare tipologia di elemento può essere utilizzata allo stesso modo, si deve valutare la distribuzione che questa ha nelle vicinanze di tutte le classi degli elementi di cui si vuole fare la caratterizzazione.

Un utile strumento che si può utilizzare è sicuramente il grafico a box che può essere visto nella figura 4.3. In questo particolare grafico si sono presi in considerazione gli edifici ad uso abitativo, e si è valutata la loro distribuzione nei dintorni (cinquanta metri) dei vari edifici specificati in ascissa. Come si può notare da questa immagine non ci sono box che assumono dei valori che si discostano in maniera significativa da quelli degli altri, di conseguenza la presenza di edifici ad uso abitativo non può essere utilizzata per la discriminazione.

Risultati

Utilizzando questa tecnica non si sono trovate regole che possono essere utili allo scopo qui presentato. In realtà una regola è stata trovata, nello specifico quella mostrata in figura 4.4. Dove è riportata la distribuzione delle *"ciminiere - torri industriali"*.

Come si può vedere il diagramma dice che questa tipologia di costruzione risulta essere vicino solamente ad edifici residenziali, questo perché i box di tutte le altre classi risultano essere concentrati in zero. Inoltre si capisce che vi sono solo pochi casi di edifici residenziali che hanno nelle vicinanze una ciminiera. Questa regola, che risulta essere anomala anche nel senso comune, è dovuta ed un effettivo errore di classificazione degli edifici attorno alla ciminiera come si può vedere dalla figura 4.5.

FIGURA 4.5: In questa figura viene mostrato un edificio classificato interamente come edificio residenziale. Questo errore è stato trovato tramite la generazione di una regola anomala che diceve che un edificio vicino ad una ciminiera probabilmente è ad uso residenziale.

4.2.3 Utilizzo delle reti neurali

Per poter trovare delle relazioni più complicate rispetto all'approccio precedente si è provato ad utilizzare una tecnica di apprendimento automatico, cioè le reti neurali. Quello che si è voluto fare con questa tecnica è stato passare l'intero vettore del vicinato ad una rete neurale e valutare se questa restituiva la stessa classe cui è associata al vettore del vicinato.Per la realizzazione della rete neurale è stato utilizzato un framework visuale scritto in java chiamato H_2O. Questo framework permette di creare delle reti neurali modificando tutti i parametri propri di questo genere di modello, come ad esempio il layout della rete e la funzione di attivazione dei nodi. La configurazione che è stata utilizzata per questa tesi non si discosta molto da quella proposta di base dal framework, visto che alcuni tentativi di modifica non hanno portato a grossi miglioramenti del risultato finale.

La dimensione del vettore, contando tutte le possibili tipologie di edificio utilizzate per la caratterizzazione, è di 186 elementi. Questi sono stati utilizzati direttamente come input della rete neurale,quindi il primo livello conta lo stesso numero di nodi, sono stati poi predisposti due layer interni di 200 nodi ciascuno ed infine nel layer di output un numero di nodi pari al numero di tipologie di edifici che si è voluto caratterizzare. Il totale di tutte le tipologie sono 127, in realtà però tra i dati a disposizione ne risultano utilizzate solamente 35 ed è quindi questo il numero di nodi che sono stati implementati.

Il framework naturalmente permette di eseguire anche la fase di trining che è stata eseguita utilizzando il 75% dei dati a disposizione. I dati sono stati ottenuti da una piccola parte dell'area totale disponibile, in quanto alcune mappe risultavano danneggiate e parzialmente accessibili. In ogni caso il numero delle feature a disposizione, all'incirca 100.000 ,è risultato essere più che sufficiente. Il restante 25% dei dati è stato invece utilizzato per eseguire i test della rete e valutarne le prestazioni su dati nuovi. Il modello ottenuto è stato poi utilizzato per classificare nuovamente la totalità dei dati, e quelli che risultavano classificati in modo diverso sono stati interpretati come anomalia.

Risultati

I test che si sono effettuati con i quattro tipi di vettori, differenti per la distanza considerata per la generazione del buffer, hanno ottenuto gli errori complessivi mostrati nella tabella 4.2. Come si può vedere si ottiene un minore errore utilizzando la distanza più bassa, comunque l'errore totale, rimane molto alto. Se si dovessero considerare tutte le feature erroneamente classificate come anomalie, si avrebbe sicuramente una riduzione dello spazio di ricerca degli errori non formali, ma ancora troppo grande per giustificare un controllo manuale su questi. Se però si analizza l'errore compiuto sulle singole classi si vede che alcune di queste hanno un errore minore. In tabella 4.3 sono elencate alcune classi/tipologie con un errore basso, si è quindi pensato di controllare solamente le anomalie di queste classi che complessivamente contano qualche centinaio di entità.

Tabella 4.2: Errori ottenuti dalla rete neurale nel classificare le feature tramite il vettore del vicinato

Distanza Utilizzata (m)	**Errore Training**	**Errore Test**
10	0.2613	0.2804
25	0.3783	0.3777
50	0.3563	0.3734
100	0.4068	0.4156

Tabella 4.3: Errori ottenuti dalla rete neurale con il vettore del vicinato da 10 metri su particolari classi

Classe/tipologia	**Errore Training**	**Errore Test**
EDIFC/abitativa	0.0084	0.0120
EDIFC/stabilimento industriale	0.0870	0.0967
EDIFC/agricolturale	0.0942	0.1635

Anche se nella maggior parte dei casi si sono ottenuti dei falsi positivi (cioè anomalie che non sono errori) qualche caso di possibile errore si è ottenuto come quello mostrato in figura4.6 in cui l'edificio è classificato nel GeoDBR come edificio industriale, mentre la rete neurale lo classifica come edificio ad uso agricolturale. Come si può vedere dalla foto dell'edificio sembrerebbe che la classificazione della rete neurale sia corretta. Un secondo esempio è mostrato in figura4.7 in cui la torre all'interro di un probabile monastero viene classificata come luogo ricreativo, mentre la rete neurale lo classifica come luogo di culto, in questo caso la presenza dell'errore non è cosi scontato in quanto all'interno delle categorie luogo ricreativo vi sono le sottoclassi museo, biblioteca e pinacoteca, tutti possibili usi che potrebbe avere l'edificio in esame. Questo comunque è un caso interessante in quanto solleva un altro problema e cioè l'utilizzo di super-classi di categorie anziché l'utilizzo di quelle più specifiche. Infatti come detto nella descrizione nei dati utilizzati le categorie possiedono una gerarchia che molto spesso viene utilizzata in modo improprio, seppur non sbagliato, dai produttori di mappe, in quanto, come è successo in questo caso utilizzano le classi di più alto livello anziché tutte le specializzazioni a disposizione. Inoltre dato che le mappe sono create da differenti entità, ognuna di queste utilizza un livello di dettaglio più o meno elevato, generando nella mappa zone più o meno dettagliate. Questo fatto rende anche per la rete neurale un più difficoltoso riconoscimento dato che tutte le varie categorie vengono trattate come se fossero differenti.

FIGURA 4.6: Edificio classificato come industriale, ma valutato come agriculturale dalla rete neurale

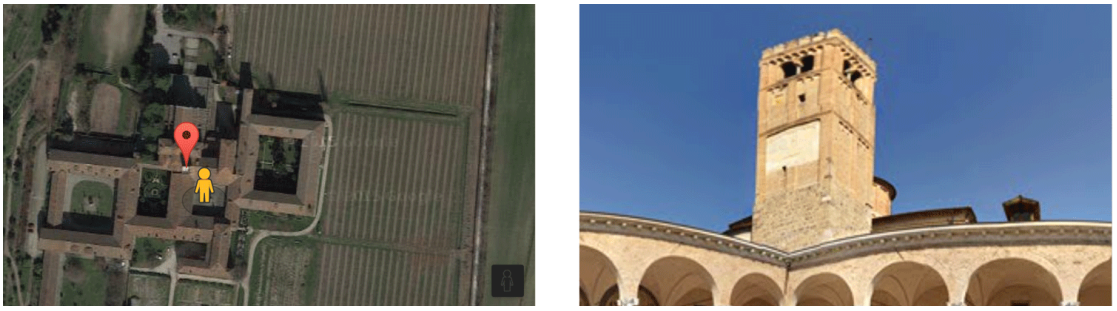

FIGURA 4.7: Edificio classificato ad uso ricreativo, ma valutato come luogo di culto dalla rete neurale

4.3 Caratterizzazione geometriche delle feature

Per eseguire la caratterizzazione geometrica degli edifici, sono stati raccolti gli indici che sono stati esposti nel capitolo precedente. Per poter ottenere l'altezza è stato necessario eseguire l'intersezione con la classe del GeoDBR UN_VOL il quale contiene tutti i dati delle unità volumetriche compresa l'altezza. Molte volte le linee che rappresentano una geometria hanno un livello di dettaglio molto alto, questo potrebbe essere un problema in quanto comporterebbe un aumento del numero di vertici che comunque aggiungerebbero poca se non nessuna informazione alla geometria, andando però ad inficiare in alcuni coefficienti che saranno descritti di seguito. Potrebbe essere conveniente quindi utilizzare un algoritmo di semplificazione delle geometrie, come ad esempio l'algoritmo di Douglas–Peucker [8] che serve appunto per ridurre il numero di punti in una poligonale, un esempio del risultato può essere visto in 4.8.

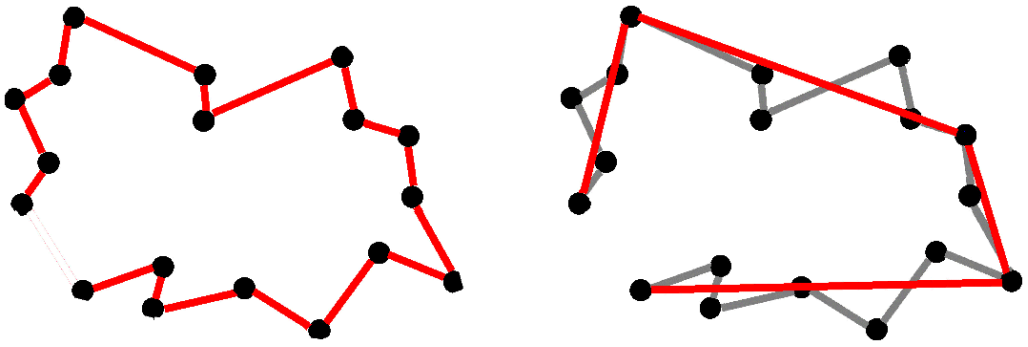

FIGURA 4.8: Esempio di Douglas–Peucker applicato ad una poligonale

Successivamente si sono calcolati tutti i descrittori geometrici ed è stata fatta un' analisi di questi. Si è utilizzato il metodo delle KME utilizzando come funzione di kernel una gaussiana.

Risultati

I risultati di questa analisi sono mostrati in figura 4.9 in cui sono stati utilizzatati dei grafici a violino. Questo tipo di grafico è molto simile al grafico a box, solamente che viene mostrato l'andamento della funzione di distribuzione trovata dal KME. Come si può intuire dove la forma è più larga, significa che vi è una maggiore concentrazione di dati.

Cercando di interpretare i dati si può vedere che purtroppo questi indici non permettono di fare nessuna discriminazione netta. solamente l'altezza e la verticalità presentano per alcune tipologie di edificio una correlazione abbastanza forte tra la loro classe e questi indici. Queste tipologie sono campanile e torre. Eseguendo dei controlli su questi due indici nelle altre classi, tutti i casi riscontrati si sono rivelati dei falsi positivi. Questo è dovuto al fatto che alcuni edifici sono rappresentati con un unica feature, e quindi con un unica geometria ed un unica classificazione. Mentre le unità volumetriche hanno un livello di dettaglio maggiore in quanto devono rappresentare tutti i volumi di cui l'edificio è composto. In particolare si sono ottenuti molti casi in cui la classificazione era chiesa basilica, eseguendo una analisi sul modello dati del GeoDBR si è scoperto che la rappresentazione delle chiese che hanno il campanile attaccato viene raggruppata in un' unica feature , mentre le altezze sono riferite alle unità volumetriche, quindi si sono ottenute delle chiese alte come campanili che hanno sporcato i dati.

4.4 Compatibilità semantica

Questa tipologia di anomalie può essere trovata solamente nelle feature con più di un attributo di classificazione. Nel GeoDBR le feature che hanno questa proprietà sono quelle della classe EDIFC,le quali possiedono un attributo che descrive la destinazione d'uso, ed un altro per la tipologia edilizia. Il primo indica a quale scopo viene utilizzato quell'edificio, mentre il secondo riguarda le caratteristiche fisiche dello specifico.

L'implementazione di questo procedimento conta le occorrenze delle coppie uso-tipo che compaiono su una feature. Questo viene fatto per tutte le feature della tabella EDFIC. Le coppie che appaiono meno di frequente sono considerate come anomalie. L'algoritmo genera una tabella di contingenza dove in figura 4.10 ne è mostrato un esempio.

4.4.1 risultati

Controllando tutte le feature a disposizione si sono trovate alcune anomalie che si sono dimostrate essere anche errori non formali e sono mostrati nelle immagini 4.11 e 4.12. Nella prima i due edifici mostrati sono stati classificati come tipologia edilizia campanile e chiaramente non è così. Nella seconda il capannone ha un uso impianto di produzione energia il cui codice è 0802 anche in questo caso si tratta di un errore non formale che è facilmente dovuto ad un errata battitura, in quanto il codice 0801, stabilimento industriale, potrebbe essere la corretta classificazione.

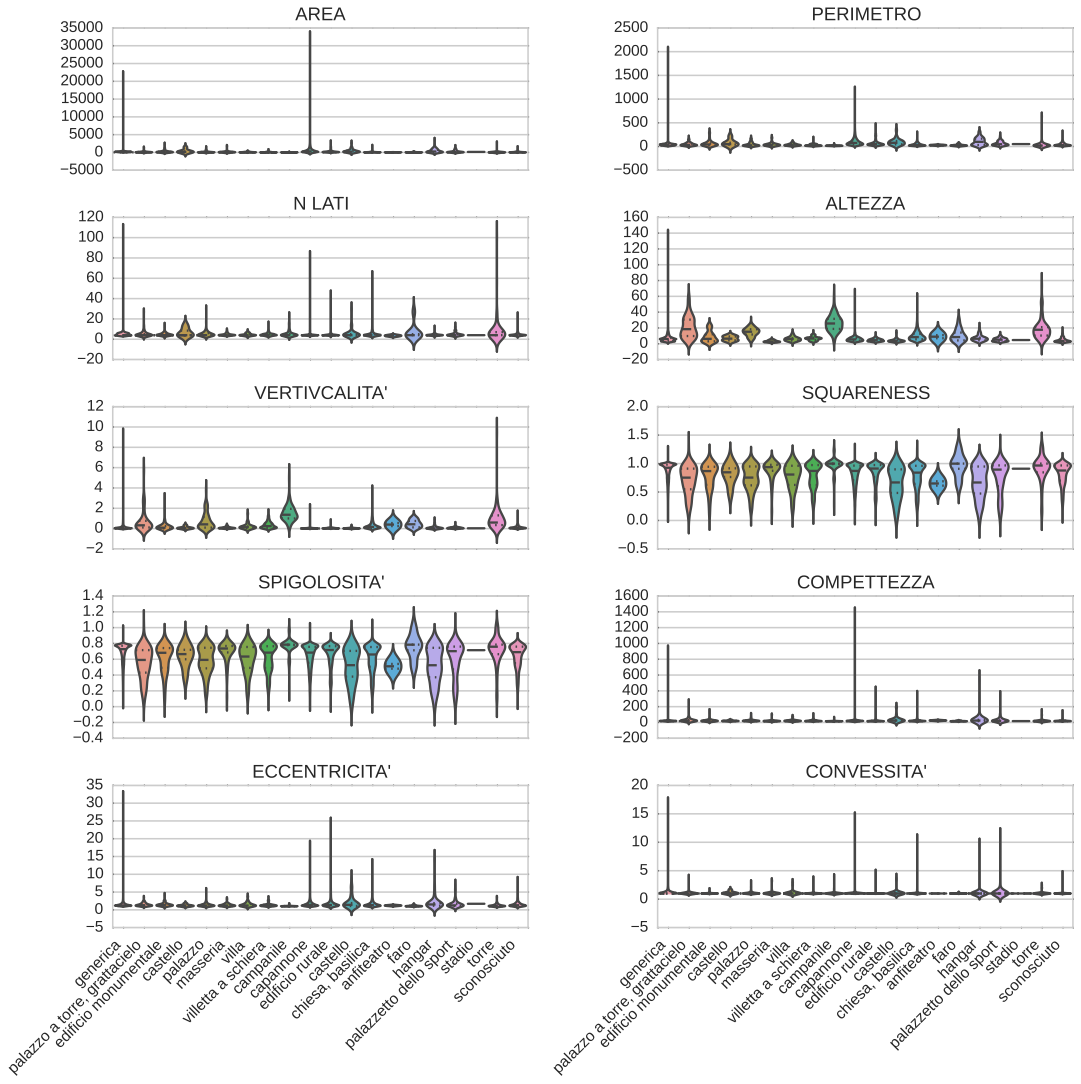

FIGURA 4.9: In questa figura vengono mostrate le distribuzioni dei vari indici geometrici per alcune delle tipologie edilizie del GeoDBR

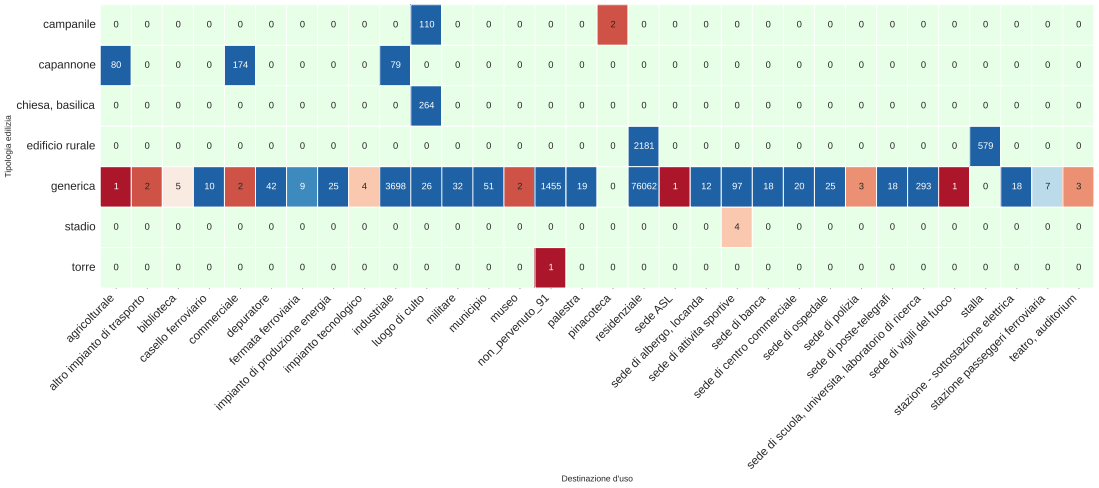

FIGURA 4.10: Tabella delle contingenze per le coppie di classi appartenenti alle feature della classe EDIFC delle aree della provincia a sud di Padova

FIGURA 4.11: Anomalia si trova controllando la coerenza semantica della classe EDIFC, l'edificio mostrato ha come destinazione d'uso pinacoteca mentre come tipologia edilizia campanile il quale è chiaramente sbagliato.

FIGURA 4.12: Anomalia si trova controllando la coerenza semantica della classe EDIFC, l'edificio mostrato ha come destinazione d'uso centrale elettrica mentre come tipologia edilizia capannone, in questo caso è l'uso ad essere sbagliato.

4.5 Change Detection

Per trovare le anomalie di cambiamento è stato implementato un algoritmo per confrontare le tre classi degli edifici (EDIFC, EDI_MIN, MN_EDI) del GeoDBR con le carte della CTRN. I periodi di realizzazione delle due carte non coincidono, in quanto i fogli della CTRN hanno una data di produzione che va dal 1997 al 2005, mentre il GeoDBR risale al 2009. Quindi d'ora in poi, in questa sezione, chiameremo feature nuove quelle appartenenti al GeoDBR e feature vecchie quelle appartenenti alla CTRN.

L'algoritmo esegue tre controlli, uno per confrontare quali sono le associazioni tra le diverse categorizzazioni date dai due modelli agli elementi che si sovrappongono. Uno per valutare quali elementi sono presenti nella mappa più vecchia ma non in quella nuova, e l'ultimo che al contrario controlla quali sono gli elementi che sono nella mappa nuova ma non in quella vecchia. Di seguito verranno spiegati come sono stati implementati i vari step.

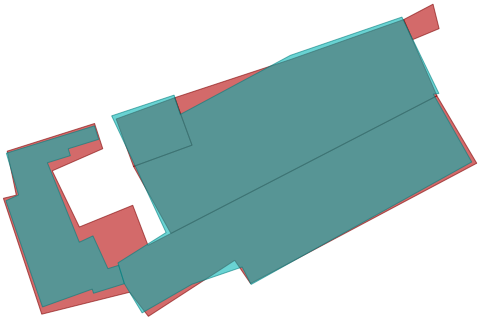

FIGURA 4.13: Le due featre rappresentano lo stesso edificio

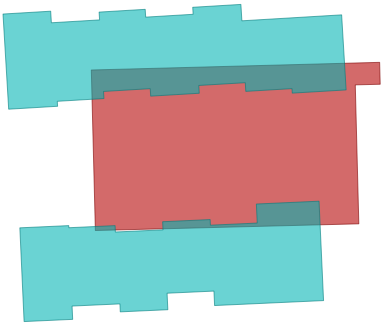

FIGURA 4.14: Le feature non rappresentano gli estessi edifici

- **Edifici che si sovrappongono**: Per questo controllo si deve fare un join spaziale tra le mappa "nuova" e quella vecchia. Il Join Spaziale è una particolare operazione geometrica che, data una feature trova tutte le feature che si sovrappongono ad essa. In questo caso la sovrapposizione può essere anche parziale. Sostanzialmente viene fatto il join spaziale utilizzando le feature dei layer GeoDBR come feature di ricerca sulla mappa della CTRN. Il risultato è che per ogni feature del GeoDBR si ottiene una lista di feature della CTRN le quali si intersecano con essa. Ora per capire se queste intersezioni sono dovute al fatto che realmente le features vecchie e quelle nuove rappresentano lo stesso elemento viene calcolata l'area di intersezione, se questa supera il 60% dell'area della feature nuova allora questo sarà considerato vero. Come dimostrazione di ciò che si è appena detto si possono vedere le immagini 4.13 e4.14 in cui le feature azzurre sono quelle nuove, mentre quelle rosse sono le vecchie. Un altro problema che è stato necessario affrontare è stato quello di capire a quale classe vecchia si dovesse associare quella nuova. Questo problema si pone solo nel caso in cui nella lista delle feature vecchie ve ne siano di classi diverse, per risolvere questo problema si è scelta la classe associata alla maggior parte dell'area intersecata.

- **Edifici presenti nella cartina vecchia e non in quella nuova**: Per trovare questi edifici si è utilizzato lo stesso risultato del join spaziale del passo precedente. Controllando tutte le feature vecchie che sono presenti nei risultati dello spatial join si possono ottenere tutti gli edifici che in qualche modo hanno un corrispondente edificio nella mappa nuova. Facendo il complementare di questo insieme, nell'insieme delle feature vecchie si trovano le feature che si stanno cercando. Qui in realtà si escludono quelle feature che erroneamente si intersecano, ma si è preferito prendere quelle che sicuramente mancano per minimizzare i controlli da fare.

- **Edifici presenti nella cartina nuova e non in quella vecchia**: Il procedimento è identico al punto precedente solo che con le feature nuove.

4.5.1 Risultati

Per quanto riguarda l'associazione delle classi dei fabbricati della CTRN con quelle degli edifici del GoeDBR sono riportate nella tabella delle contingenze in figura 4.15, in cui le varie sfumature di rosso rappresentano le possibili anomalie,in quanto sono le relazioni che si verificano più di rado, crescendo verso le sfumature blu che rappresentano invece le

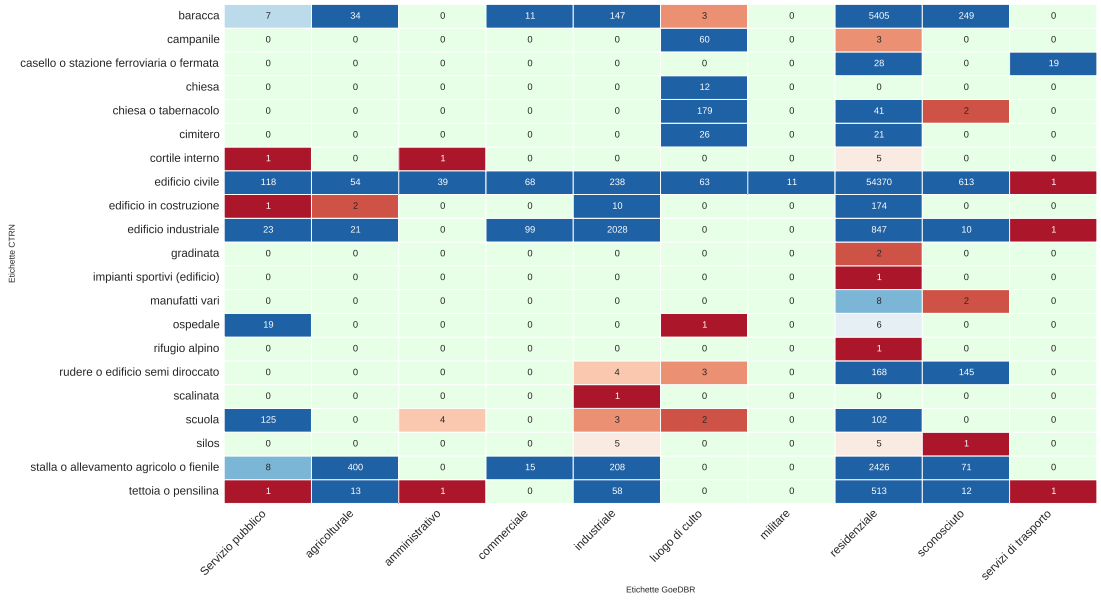

FIGURA 4.15: In questa figura è mostrata la matrice delle contingenze che mostra le relazioni tra le categorizzazioni della classe fabbricati della CTRN e quelle della classe EDIFC del GeoDBR

relazioni più frequenti. Controllando le feature che corrispondono alle relazioni anomale si sono trovate alcune anomalie che sono mostrate qui sotto.

FIGURA 4.16: Anomalia trovata controllando le relazioni con bassa cardinalità, in questo caso l'edificio èra classificato nella CTRN come edificio civile, mentre nel GeoDBR come edificio adibito a servizi di trasporto.

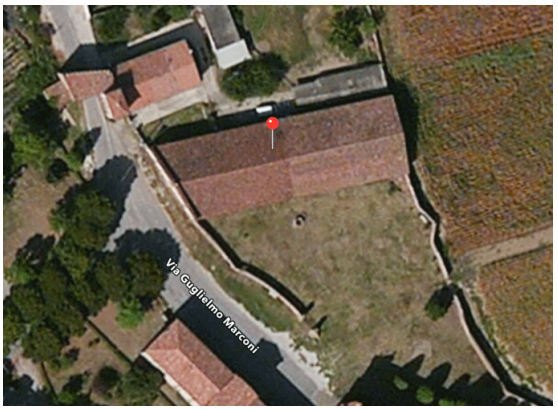

FIGURA 4.17: Anomalia trovata controllando le relazioni con bassa cardinalità, in questo caso l'edificio era classificato nella CTRN come edificio industriale, mentre nel GeoDBR come edificio adibito a servizi di trasporto.

Capitolo 5

Conclusioni

Questo lavoro di tesi si poneva l'obbiettivo di ideare e testare delle tecniche per l'individuazione automatica degli errori non rilevabili con i normali controlli formali. Su questi argomenti si erano già fatti degli studi che hanno contribuito alla definizione di una tassonomia e hanno inoltre portato allo sviluppo di alcuni metodi per l'individuazione di questi errori. Queste tecniche però si sono basate su esperienze di persone esperte nel settore, utilizzando delle regole mirate ad individuare un particolare errore. Quello che si è fatto invece in questa tesi è stato di sviluppare tecniche che riuscissero a generare delle regole attraverso alcune analisi sui dati stessi. Sono stati utilizzati degli strumenti statistici per creare dei modelli che sono stati usati per individuare dati anomali e quindi potenzialmente errati. Le tecniche sono state studiate in maniera completamente indipendente dal modello dati, risultando applicabili in qualsiasi database geografico. Sono stati fatti dei test su dati reali nello specifico, utilizzando il modello dati del GeoDBR. Gli errori che si sono trovati non sono stati molti e sicuramente sono una piccola percentuale degli errori realmente presenti. Un fattore che ha influito sui risultati è sicuramente la scarsa qualità dei dati utilizzati, Per primo l'utilizzo improprio della gerarchia dei valori degli attributi, cioè l'utilizzo di classi di alto livello anziché quelle più specializzate. La quantità di informazione quindi è sicuramente minore e i modelli ne risentono. I risultati sono comunque sufficienti per far sperare che attraverso una raffinazione di queste tecniche si possano ottenere delle prestazioni migliori.

Questo lavoro potrebbe essere ampliato provando degli altri strumenti statistici, oppure cercando di applicare su altre classi i modelli studiati in questa tesi.

Bibliografia

[1] Nicholas Chrisman. *Fundamentals of Spatial Data Quality*. ISTE, 2010. ISBN 9780470612156. doi: 10.1002/9780470612156.ch1.

[2] Zanon M. Lissandron I. Savino S., Rumor M. Data enrichment for road generalization through analysis of morphology in the cargen project, 2010.

[3] Alan Stocco. Controlli per la qualità dei database topografici a grande scala, 2011.

[4] Valerio Lorenzo. Approccio e strumenti per individuare errori di classificazione e di incompletezza nei db topografici, 2012.

[5] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011. ISBN 0123814790, 9780123814791.

[6] David Avnir Hagit Zabrodsky, Shmuel Peleg. Continuous symmetry measures. *J. Am. Chem. Soc.*, 114, September 1992.

[7] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3):832–837, 09 1956. doi: 10.1214/aoms/1177728190. URL http://dx.doi.org/10.1214/aoms/1177728190.

[8] DAVID H DOUGLAS and THOMAS K PEUCKER. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122, 1973. doi: 10.3138/FM57-6770-U75U-7727. URL http://dx.doi.org/10.3138/FM57-6770-U75U-7727.

Ringrazziamenti

Ringrazio i miei genitori per avvermi permesso di inziare e concludere questo mio percorso di studi, nonchè per avermi sempre sostenuto e supportato.

Ringrazio il Prof. Massimo Rumor e Sandro Savino per avermi aiutato durante tutto lo svolgimento di questa tesi.

Ringrazio i miei amici per avermi sostenuto e per tutti i bei momenti passati assieme.

Ringrazio i miei compagni di corso (che fanno parte anche della categoria precedente) per tutti i momenti condivisi assieme e per aver permesso di trasformare una passione in qualcosa che continuerà anche dopo la conclusione di questo percorso.