

# UNIVERSITÀ DEGLI STUDI DI PADOVA

Facoltà di Ingegneria

Corso di laurea magistrale in Ingegneria Informatica


# Approccio e strumenti per individuare errori di classificazione e di incompletezza nei DB Topografici


RELATORE
Prof. Massimo Rumor

LAUREANDO
Lorenzo Valerio

CORRELATORE
Ing. Sandro Savino



Anno Accademico 2011/2012

*A papà*

# SOMMARIO

All'interno di un Sistema Informativo Territoriale i dati sono una risorsa cruciale, in quanto sono costosi da acquisire, archiviare e manipolare. Questo ci fa capire l'importanza che le informazioni fornite dai dati geografici siano il più possibili affidabili; si deve cioè verificare la qualità dei dati stessi in tutte le fasi, dalla rilevazione fino alla pubblicazione.

Una parte importante della ricerca sulla qualità dei dati riguarda la descrizione delle incertezze nei dati spaziali, incertezze dovute dalla vaghezza della documentazione, dall'ambiguità sulle definizioni degli oggetti di un dataset e dagli errori, definiti come la differenza tra la realtà e la sua rappresentazione.

Con l'avvento dei GIS e l'evoluzione degli strumenti di acquisizione, analisi e correzione dei dati, la quantità di errori presenti nei dataset finali si è ridotta notevolmente. Questi strumenti permettono di rilevare e correggere soprattutto errori sulle geometrie, come ad esempio *self-intersection* e punti doppi.

Esistono però degli errori che non sono rilevabili con strumenti formali: sono situazioni particolari in cui le caratteristiche del dato cartografico si scostano da quelle volute dalla logica comune. Queste circostanze sono collegate al valore semantico degli oggetti. In questo lavoro di tesi si definirà anomalia queste situazioni in cui l'informazione cartografica potrebbe essere, secondo logica comune, errata ma per le quali non esiste garanzia che esse siano effettivamente un errore.

Lo scopo di questa tesi è quello di formalizzare un modello concettuale e sviluppare degli strumenti per la rilevazione automatica di queste anomalie; l'output degli algoritmi fornisce un indicazione su possibili errori, da verificare con ulteriori analisi.

Nel primo capitolo viene introdotto il tema della cartografia e del processo cartografico, il progetto CARGEN e gli strumenti di lavoro. Nel secondo capitolo viene definita la qualità dei dati con particolare riferimento agli errori e alle anomalie presenti in un DB e con un esempio di caso reale di controllo della qualità. Nel terzo capitolo vengono descritti gli approcci e i metodi implementati per la ricerca di eventuali errori di classificazione e incompletezza dei dati. Nel quarto capitolo infine vengono analizzati i risultati dei test su un campione di dati.

# INDICE

# ELENCO DELLE FIGURE

# 1

# LA CARTOGRAFIA E IL PROGETTO CARGEN

## 1.1 ORIGINI DELLA CARTOGRAFIA

L'Associazione Cartografica Internazionale nel 1966 definiva la cartografia come "il complesso degli studi e delle operazioni scientifiche, artistiche e tecniche che si svolgono a partire dai risultati delle osservazioni dirette o dalla utilizzazione di una documentazione al fine di elaborare ed allestire carte, piante ed altri modi d'espressione atti a risvegliare l'immagine esatta della realtà". Da quella data, la richiesta di dati sulla topografia e su altri temi della superficie terrestre ha avuto una enorme accelerazione.

Fin dall'antichità c'era la necessità di rappresentare in piano la superficie terrestre; è del 2400-2200 a.c. la prima rappresentazione della regione dell'Eufrate in Mesopotamia, mentre nel 1200 a.c gli egiziani produssero piani catastali per la delimitazione delle proprietà terriere.

Furono poi i greci, a partire dal IV secolo a.c., a concepire una cartografia scientifica, con i primi esempi di rappresentazione concepiti da Dicearco di Messina e da Eratostene: il primo compose una carta in cui compare un parallelo passante per i luoghi ritenuti alla stessa latitudine, mentre il secondo, due secoli dopo, introdusse una prima grezza struttura di reticolato geografico composta da diverse linee di riferimento a distanze variabili passanti per luoghi noti e linee perpendicolari alle precedenti.

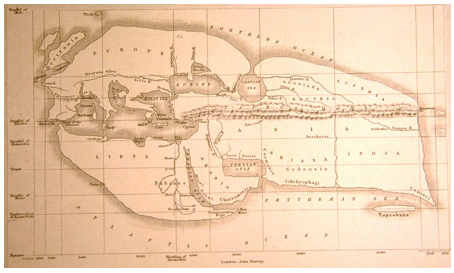

Figura 1: Mappa di Eratostene

Le necessità di espansione commerciale produssero in epoca romana un miglioramento delle conoscenze geografiche; risale infatti al III o IV secolo d.c. la rappresentazione dell'impero romano in una striscia lunga 6,75 m e larga 34 cm, rappresentazione nota come *Tabula Peutigeriana* dal nome dell'austriaco Corrado Peutiger che per primo la studiò. In quest'epoca compaiono anche i primi atlanti, grazie a Marino di Tiro e Claudio Tolomeo, delle terre fino ad allora conosciute (dalle Canarie fino alla Cina).

Nel medioevo non ci furono miglioramenti nella conoscenza cartografica, anzi, si ebbe una notevole regressione nel ritorno all'idea di una terra piana, sostenuta dalla Chiesa e dalle Sacre Scritture. Il mondo veniva rappresentato tramite le *Mappae Mundi*, con Gerusalemme al centro, l'Asia in alto, l'Europa in basso a sinistra e l'Africa in basso a destra.

Tra il '400 e il '500, con gli studi umanistici dei classici provenienti dalla Grecia, con l'invenzione della stampa e con la successiva scoperta dell'America da parte di Cristoforo Colombo, si determina un'improvvisa accelerazione nello sviluppo della cartografia, con un proliferare di cartografia a stampa e manoscritta e con i primi rilievi topografici regolari. Mentre la cartografia generale continua la sua evoluzione con la grande carta dell'Europa (1544) e del planisfero (1569) del Mercatore, nei diversi paesi europei ed anche negli stati italiani si avvia, sulla spinta delle autorità pubbliche, la redazione di una cartografia ufficiale, attraverso rilevamenti sistematici.

Nel '700 e nell'800 in seguito alle maggiori conoscenze della forma e della dimensione terrestre e all'invenzione di nuovi strumenti di rilevazione, nasce la geodesia e quindi la cartografia geodetica, cioè basata sulla conoscenza esatta della posizione di alcuni punti, e l'introduzione di segni convenzionali e nuove proiezioni.

Nel '900 inizia la costruzione di carte topografiche attraverso fotografie (fotogrammetria), fino all'adozione della stereofotogrammetria, possibile attraverso lo stereoautografo in grado di effettuare la restituzione cartografica di fotografie. Negli ultimi decenni si è imposta la proiezione U.T.M. e si è passati ad una nuova generazione di rilevamento con la produzione delle carte a partire da foto satellitari, scattate ad altezze che superano i 900 km. Si possono ricordare i satelliti della serie Landsat in grado di studiare le forme e le risorse terrestri per la produzione di una cartografia tematica.

## 1.2 LE CARTE GEOGRAFICHE: DEFINIZIONE E CONCETTI PRINCIPALI

Se dovessimo definire così su due piedi il concetto di carta geografica, potremmo dire che è una rappresentazione attraverso simboli grafici ben definiti che rispecchia il mondo reale. L'Associazione Internazionale di Cartografia definisce la *carta geografica* 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". Questa definizione evidenzia quelli che sono i concetti principali.

*Proiezioni e deformazioni*

La proiezione non è altro che il sistema matematico-geometrico per mezzo del quale la superficie sferica della Terra viene trasferita sulla superficie piana di una carta. Poiché una proiezione deforma la geometria del globo e poiché non esiste una proiezione migliore delle altre, è compito del cartografo scegliere la giusta proiezione (ad esempio per le carte topografiche si usano le proiezioni isogone che mantengono le forme degli oggetti geografici, oppure proiezioni equiareali, che mantengono il rapporto delle aree corrispondenti, per rappresentare fenomeni economici e vaste aree geografiche);

*Scala*

La scala è definibile come il rapporto tra la lunghezza misurata sulla carta e la corrispondente lunghezza reale sul terreno; ad esempio una mappa in scala 1:5.000 indica che 1 mm nella mappa corrisponde a 5 m nel mondo reale. Le carte vengono suddivise in carte geografiche a piccola o grande scala. Le carte geografiche a larga scala sono quelle che cercano di riportare con un alto grado di dettaglio i singoli oggetti in uno spazio ristretto, mentre quelle a bassa scala rappresentano una vasta regione che può essere uno Stato, un continente o il globo, con un basso livello di dettaglio ma col vincolo che le relazioni tra gli elementi siano rispettate. Quindi più piccola è la scala minore è la quantità di informazioni che figurano nella carta, maggiore è la scala maggiore la quantità di informazioni. In funzione della scala le carte possono essere classificate nel seguente modo:

- **piante** e **mappe**, in scala maggiore di 1:10.000, usate per mappe catastali e per carte tecniche dei centri abitati;
- **carte topografiche**, in scala compresa tra 1:10.000 e 1:200.000, che rappresentano con molta accuratezza zone limitate della superficie terrestre, come le carte dell'Istituto Geografico Militare;
- **carte corografiche**, in scala compresa tra 1:200.000 e 1:1.000.000, che rappresentano zone abbastanza estese della superficie terrestre come regioni o intere nazioni;
- **carte generali**, in scala minore di 1:1.000.000, con basso livello di dettaglio e che rappresentano continenti o l'intero globo.

La scala quindi regola una serie di parametri per la creazione di una carta geografica, tra cui il grado di risoluzione, l'errore massimo di posizionamento e il livello di dettaglio. Il *grado di risoluzione* è la dimensione lineare del più piccolo particolare rappresentabile ed è dato dal minimo spessore del tratto grafico con cui la carta viene

disegnata (per convenzione viene assunto pari a 0,2 mm); l'*errore massimo di posizionamento* di un punto rappresenta il diametro del cerchio al cui interno il punto è contenuto sicuramente e corrisponde all'incertezza con cui è rappresentata la posizione di un generico punto sulla carta (per convenzione viene assunto pari a 0,5 mm); il *livello di dettaglio* di una mappa influenza la precisione con cui i dati sono riportati su di essa, sulla quantità di informazione che è possibile inserire per non avere una mappa con una eccessiva densità di informazione rispetto all'area rappresentata, ma anche sul costo di produzione e di manutenzione della mappa stessa (questo giustifica proprio la necessità di presentare dati geografici già disponibili ad un rapporto di riduzione differente).

## 1.3 I GIS

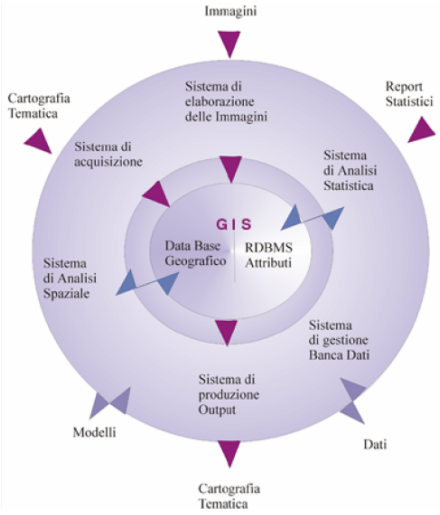


Figura 2: Organizzazione dei GIS

Dalla seconda metà del ventesimo secolo la richiesta di dati sulla topografia e su altri temi della superficie terrestre ha avuto una enorme accelerazione. Questa accelerazione è stata anche favorita dallo sviluppo delle fotografie aeree e dalle immagini riprese da satellite; queste due fonti hanno permesso di poter cartografare ampie aree della superficie terrestre con maggior dettaglio e precisione rispetto al passato.

Queste evoluzioni, insieme all'utilizzo delle tecnologie informatiche, sono state di grande aiuto per tutti gli studiosi che hanno a che fare con lo studio delle carte. Inizialmente infatti l'informazione spaziale era semplicemente un disegno su carta, in cui tutte le informazioni erano codificate da simboli che venivano associati agli attributi tramite legende ed, in alcuni casi, una memoria scritta contenente ulteriori informazioni veniva associata alla carta. Tutto ciò portava a ridurre le moli di dati iniziali (con conseguente perdita di precisione), al disegno a mano libera della mappa (che richiedeva estrema precisione) e alla divisione delle regioni di studio in più fogli (probabili aree di interesse a cavallo di due fogli). La carta che ne risultava era pertanto un documento puramente statico e qualitativo. Questo poteva andare bene in epoche in cui si supponevano valide le carte anche per periodi pari a vent'anni, ma non più ora in quanto è frequente il bisogno di poter aggiornare continuamente i dati spaziali relativi ai mutamenti dei fenomeni naturali.

A partire dagli anni '60 fu necessario perciò sviluppare nuove tecnologie a supporto della cartografia, ma fu negli anni '90 che avvenne il vero e proprio boom grazie all'estrema diffusione di PC sempre più potenti e di software a prezzi sempre meno elevati; si passò perciò da una cartografia in prevalenza cartacea ad una cartografia su base informatica denominata *cartografia numerica*. Nascono perciò una serie di strumenti di supporto per la gestione ottimale della cartografia digitale che prendono il nome di *Sistemi Informativi Geografici* o GIS.

Esistono varie definizioni possibili di Sistema Informativo Territoriale (GIS), ma quella che più sembra in linea con la definizione di Sistema Informativo in generale, definito come un insieme organizzato di procedure, risorse umane e risorse materiali utilizzato per la raccolta, l'archiviazione, l'elaborazione e la comunicazione di informazioni necessarie ad una organizzazione per gestire le proprie attività operative e di governo, è la seguente:

> Un sistema informativo geografico è un sistema composto da banche dati, hardware, software ed organizzazione che gestisce, elabora ed integra informazione su una base spaziale o geografica.

La tecnologia dei GIS è "l'insieme degli strumenti usati per acquisire, gestire e rendere disponibile l'informazione territoriale". Gli strumenti, nella realtà informatica, sono l'hardware e il software, il quale a sua volta si divide in software di base, d'ambiente e applicativo. Facciamo quindi un'analisi veloce di questi strumenti:

- Hardware: costituito da computer, stampanti, plotter, monitor ad alta risoluzione;
- Software di base: è formato dal sistema operativo e dalle utility di sistema;
- Software d'ambiente: è formato da linguaggi e da gestori di dati. In questo ambito si collocano i prodotti della tecnologia GIS, specializzati per l'analisi, la gestione e la rappresentazione dell'informazione territoriale, e i prodotti della tecnologia CAD per la gestione dell'informazione spaziale non territoriale;
- Software applicativo: è un prodotto specifico per supportare delle funzioni predefinite, realizzato sulla base di uno o più pacchetti GIS e altre componenti del software d'ambiente.

Il GIS memorizza le informazioni geografiche attraverso strati separati rappresentati sullo schermo geometricamente da punti, linee o aree. Ad ogni elemento geografico corrisponde un attributo o elemento descrittivo che indica cosa rappresenta l'elemento spaziale e la sua esatta posizione geografica espressa in coordinate. Questo sistema è risultato fondamentale per la risoluzione di molti problemi del mondo

reale. L'informazione geografica contiene un riferimento spaziale esplicito (come latitudine e longitudine) o implicito (come un indirizzo, un codice postale, una denominazione stradale). Grazie all'utilizzo di un processo automatico chiamato *geocoding* è possibile ottenere riferimenti geografici espliciti da riferimenti impliciti, consentendo di localizzare oggetti ed eventi sulla superficie terrestre.

L'obiettivo principale dei GIS è quello di svolgere i seguenti compiti:

- *Inserimento*: prima di essere usati, i dati devono essere convertiti in un idoneo formato digitale. Questo processo viene chiamato *digitalizzazione*;
- *Manipolazione*: i dati richiesti da uno specifico progetto GIS possono necessitare una trasformazione per essere compatibili con il sistema. Un esempio può essere l'unione di due carte a scala differente: prima di essere unite è richiesto di portare le due carte alla medesima scala, cioè ad uno stesso livello di dettaglio o accuratezza;
- *Gestione*: è svolto dai software di gestione di dati come i DBMS;
- *Ricerca e analisi*: il GIS fornisce semplici funzionalità di ricerca e sofisticati strumenti di analisi per fornire informazioni tempestivamente ad analisti. Gli strumenti più importanti di analisi nei moderni GIS sono l'analisi di prossimità (buffering) e l'overlay;
- *Visualizzazione*: il risultato finale è rappresentato nel modo ottimale come mappa o grafico

## 1.4 IL PROCESSO CARTOGRAFICO

Il processo cartografico è definito come un insieme di operazioni e procedimenti che sono necessari per la creazione di una carta geografica. Nonostante il progresso delle scienze e delle tecnologie abbia portato ad uno sviluppo delle tecniche e degli strumenti utili al cartografo, questo procedimento è rimasto pressoché invariato nel tempo.

Il processo cartografico può essere schematizzato nelle seguenti operazioni:

- analisi e definizione delle caratteristiche della mappa;
- raccolta dei dati;
- produzione della mappa;
- collaudo.

Nella fase di *analisi e definizione* si decidono quali sono le caratteristiche principali del prodotto. Inizialmente, in base ai contenuti della

mappa, cioè le informazioni che si vogliono collocare nella mappa stessa, si determinano i livelli di astrazione e di rappresentazione che saranno utilizzati per trattare i dati in ingresso. Vengono successivamente decise le caratteristiche tecniche della carta, tra cui la scala, la proiezione, la superficie di riferimento, il tipo di rappresentazione (conforme, equivalente, equidistante) e le caratteristiche semantiche, relative cioè ai contenuti: in base alle decisioni semantiche, la carta assumerà una natura tematica, cioè focalizzata solo su alcuni aspetti, oppure una natura olistica, mirata a dare una definizione d'insieme. Una volta determinata l'astrazione della realtà, si produce anche, per le tradizionali mappe cartacee, la legenda della carta, e per le mappe digitali la definizione di un GeoDB.

Terminata la fase di analisi e definizione, si passa alla *raccolta dei dati*. I dati possono essere raccolti in due modi, a seconda che la mappa sia il frutto di una rilevazione o di una derivazione. Nel primo caso, gli elementi del territorio possono venire rilevati o durante un sopralluogo con l'aiuto di sistemi di posizionamento satellitare (GPS), o con tecniche di rilevazioni fotogrammetriche che ricavano la posizione dei punti mediante l'utilizzo di immagini fotografiche stereoscopiche del terreno, oppure con una scansione e successiva digitalizzazione delle mappe analogiche. Nel secondo caso, se la mappa viene derivata, vengono utilizzati dati provenienti da una cartografia preesistente, la quale dovrà avere un livello di dettaglio maggiore (e quindi un'alta scala) rispetto alla carta che si deve realizzare (questo è necessario in quanto solamente se si è in possesso di tutti i dettagli è possibile astrarre e rappresentare correttamente la realtà). In entrambi i casi, gli operatori dovranno assegnare ad ogni elemento individuato un codice diverso a seconda della natura dell'elemento (strada, ferrovia, edificio, fiume etc.), in accordo col modello già stabilito nella fase di analisi e definizione.

Una volta che il cartografo entra in possesso di tutte le informazioni (dati e specifiche) necessarie, avvia il processo di generalizzazione, fondamentale per la *produzione della mappa*. Durante questa fase, il cartografo deve prendere una serie di decisioni atte a soddisfare le specifiche definite, ma anche a garantire i requisiti classici di ogni mappa, quali la leggibilità e l'usabilità. In pratica egli dovrà creare un'astrazione della realtà geografica che faciliti la comprensione e la comunicazione dell'informazione.

L'ultima fase del processo cartografico consiste nel *collaudo della mappa*, durante il quale si verifica la correttezza e la consistenza della carta. Tra le attività più importanti di collaudo vi è il controllo finale sul terreno mediante operazioni di misura e verifica della rappresentazione cartografica: vengono confrontati sul campo i dati riportati nella mappa con rilevazioni effettuate tramite strumenti ad alta precisione, come ad esempio il GPS differenziale.

## 1.5 IL PROGETTO CARGEN

Il progetto CARGEN, CARtographic GENeralization, è un progetto di ricerca, partito nel 2006, che vede coinvolti il Dipartimento di Ingegneria Informatica dell'Università di Padova, la Regione Veneto, e l'Istituto Geografico Militare. L'obiettivo principale era quello di elaborare nella sua interezza (progettazione, sviluppo e test) un processo automatizzato di generalizzazione cartografica, che partisse dai dati in scala 1:5.000 del modello GeoDBR, sviluppato dalla Regione Veneto, per restituire una base di dati coerente con il modello DB25 dell'IGM.

Avendo ottenuto dei risultati soddisfacenti, il lavoro di ricerca fu poi esteso nel 2009 alla generalizzazione di mappe con scala 1:50.000. Sebbene i dati di partenza si riferiscono appunto al più recente modello GeoDBR, in questo progetto si é fatto uso anche della CTRN, fornita sempre dalla Regione Veneto. In particolare, la relazione che intercorre tra i modelli ha come punto di partenza la CTRN, che viene utilizzata per popolare il GeoDBR, dal quale si derivano successivamente i dati per popolare il DB25. Seppur già da tempo, in Europa, la produzione automatizzata di carte a differenti scale attraverso la generalizzazione è prassi ormai consolidata, in Italia la situazione risulta essere diametralmente opposta. Ecco perchè questo progetto, essendo tra i primi in Italia, cerca anche di essere uno stimolo per questo tipo di produzione. Sin dall'inizio si é deciso di intraprendere la strada del raggiungimento di risultati concreti: ad innovativi approcci teorici fini a se stessi si é preferita la realizzazione di un prodotto reale, ossia un software che effettivamente realizzasse la generalizzazione cartografica. Nei paragrafi successivi viene fatta una panoramica sulla situazione cartografica in Italia, sui modelli dei dati utilizzati nel progetto, l'architettura e gli strumenti usati.

## 1.6 LA SITUAZIONE CARTOGRAFICA IN ITALIA

La competenza cartografica istituzionale italiana è cura dell'Istituto Geografico Militare Italiano (I.G.M.I.), che sin dal 1875 si è occupato della redazione della cartografia ufficiale italiana. La prima produzione cartografica dell'istituto è stata la *Nuova Carta Topografica d'Italia*, realizzata alla scala 1:100.000. Il territorio Italiano ricade nei fusi 32, 33 e in parte del 34 e nelle fasce S e T del Sistema UTM. L'IGM ha prodotto diverse carte che coprono il territorio a livello nazionale e solo una parte di queste vengono mantenute aggiornate: per la scala 1:25.000 esistono le serie 25V, 25 e 25DB, per la scala 1:50.000 le serie 50 e 50/L e per la scala 1:100.000 le serie 100V e 100L. Tra le varie serie sussiste un preciso rapporto matematico: ogni foglio di una mappa in scala 1:100.000 è divisa in quattro *settori*, rappresentati in altrettante mappe della serie 1:50.000; queste mappe vengono dette *quadranti*, e ogni quadrante è a sua volta diviso in quattro parti, le *tavolette*, che

sono in scala 1:25.000. Delle serie cartografiche alla scala 1:25.000, solo la più antica, la 25V (vecchio taglio), è stata completata. I dati di questa serie sono aggiornati mediamente al 1960, ad eccezione di alcuni elementi, che sono stati prodotti durante una campagna di aggiornamento parziale del 1984. La serie 25 copre, dunque, solo il 36% del territorio nazionale; la sua produzione è stata interrotta per fare spazio alla nuova serie 25DB. Questa serie introduce la cartografia digitale, e, su un totale di 2298 sezioni, ne sono state realizzate 68, ottenute tramite stereorestituzione numerica o come derivazione dalla cartografia tecnica regionale.

Oltre all'IGM, che lavora su cartografie con scala inferiore, da 1:25.000 in giù, vi sono altri enti che si occupano di cartografia e ai quali vengono affidate le realizzazioni di carte a media e grande scala (1:10.000 e superiore). Inizialmente, in base alla legge n°68/1960, "Norme sulla cartografia ufficiale dello Stato e sulla disciplina della produzione e dei rilevamenti terrestri e idrografici", era il Catasto a dover adempiere a questo compito; successivamente con il DPR n°616/77 tale funzione è stata trasferita alle Regioni che hanno potuto così gestire in modo autonomo la creazione delle carte regionali, definite tecniche in quanto create specificamente per i tecnici delle amministrazioni. Le carte prodotte dalle Regioni sono carte ricche di particolari e vengono aggiornate frequentemente, operazione facilitata dalla ristretta porzione del territorio nazionale che viene rappresentata. La *Carta Tecnica Regionale* (CTR), e la sua versione digitale (*Carta Tecnica Regionale Numerica* - CTRN), costituisce la base di riferimento per la redazione degli strumenti urbanistici comunali, per i Piani di Coordinamento Provinciali, per i Piani d'Area e per i vari piani di settore della pianificazione e della programmazione regionale.

In Veneto vengono prodotte due serie di CTR: una in scala 1:10.000, composta da *sezioni*, e una in scala 1:5.000, composta da *elementi*. Il taglio e l'inquadramento le rendono sovrapponibili alle carte IGM di nuova produzione. In particolare, ogni foglio IGM in scala 1:50.000 è diviso in 4 quadranti della serie 25, ognuno dei quali ripartito in 4 sezioni in scala 1:10.000 della CTR, divisi a loro volta in 4 elementi in scala 1:5.000.

La mancanza di standard nella redazione delle CTRN e la sempre crescente esigenza di una collaborazione tra le regioni italiane, hanno portato nel 2010 alla produzione del documento "Catalogo dei Dati Territoriali - Specifiche di contenuto per i DB Geotopografici" che definisce un modello nazionale per le mappe in scala 1:1.000, 1:2.000, 1:5.000 e 1:10.000 e fissa i requisiti minimi che ogni cartografia regionale dovrebbe soddisfare. Una volta adottato dalle regioni, il modello di dati costituirà la base per una facile condivisione dei dati geografici tra le diverse regioni d'Italia.

## 1.7 MODELLI DEI DATI

In questa sezione vengono approfonditi i tre modelli dati che sono stati usati nel progetto, ovvero i modelli CTRN e GeoDBR per i dati geografici in scala 1:5.000 e il modello DB25 dell'IGM per i dati geografici in scala 1:25.000

### *CTRN*

La Carta Tecnica Regionale Numerica è una cartografia generale e metrica, in formato vettoriale, prodotta dalla Regione Veneto. La carta, la cui produzione trova la sua principale fonte di dati nel rilievo fotogrammetrico, gode di campagne d'aggiornamento piuttosto frequenti ed offre quindi un dato geografico piuttosto recente e di buona qualità.Le scale di rappresentazione adottate sono la scala 1:5.000 per la quasi totalità del territorio regionale e la scala 1:10.000 per le zone montane scarsamente urbanizzate.

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.

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.

### *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, le cui finalità principali consistono nel "Deliberare... una strategia unitaria e definire norme di comportamento comuni fra le regioni e le province autonome nella materia delle informazioni aventi rilevanza territoriale e con particolare riferimento alla programmazione della produzione cartografica...". Aderendo alle specifiche IntesaGis, il GeoDBR si presta come perfetto tramite tra i dati della CTRN e quelli del DB25 (e perciò nel futuro sarà destinato a sostituire la CTRN). Le specifiche di IntesaGis sono infatti mirate proprio alla definizione di un modello di dati che permetta una facile derivazione del DB25.

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 raggruppati 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.

*DB25*

Il DB25 è il modello dati creato dall'IGM per la compilazione della serie topografica 25DB. È un modello di rappresentazione del mondo reale per mezzo di *feature* suddivise in oggetti identificati dall'attributo LAB (*Label*). L'attributo LAB contiene un codice che identifica in maniera univoca l'oggetto cartografico ed è presente nello standard DIGEST. Il carattere identificativo dell'attributo LAB fa si che gli siano legati altri attributi indispensabili per descrivere le proprietà di un oggetto. La prima lettera del codice identifica a quale tipo di primitiva geometrica fa parte la *feature*:

- **A** per tipologia areale;
- **L** per tipologia lineare;
- **P** per tipologia puntuale;
- **T** per feature testuali.

Ad esempio, la codifica FACC del DIGEST identifica le strade con il codice AP030; nel DB25 la *feature* è lineare e quindi le strade saranno identificate dal codice LAP030.

Ad ogni *feature* è associato un insieme di attributi, attributi che possono essere variabili o fissi (che dipendono cioè dal codice LAB e che quindi non compaiono nel database ma che sono resi disponibili in fase di cessione dati).

Le *feature* che popolano il DB25 si riferiscono a 291 particolari topografici, suscettibili di restituzione, di eventuale ricognizione e/o

di acquisizione da Banche Dati di Enti Pubblici e Privati, e a 48 tipologie di testi, per la maggior parte legati a particolari topografici

A differenza del GeoDBR, l'organizzazione gerarchica delle *feature* del DB25 è minima: ogni oggetto topografico è di norma descritto direttamente da un codice LAB univoco. Questa scelta si spiega nella finalità tipografica di questo modello dati: ad ogni codice LAB è infatti associata anche una vestizione grafica, che viene usata nella stampa delle carte e riportata nella legenda delle cartografie in serie 25DB. Conseguenza di questa scarsa organizzazione gerarchica è la presenza, nel DB25, di più *feature* che afferiscono allo stesso oggetto reale; la scelta di quale tra queste *feature* utilizzare per l'oggetto si basa generalmente sui limiti di acquisizione: un esempio è l'oggetto edificio industriale, che, nonostante sia descritto dal medesimo codice FACC AC000 *Processing Plant/Treatment Plant*, viene rappresentato dalla *feature* Opificio Generico, con codice LAB C406 se la sua superficie è inferiore a 1500 metri quadri, oppure con la *feature* Stabilimento Industriale, LAB C405A, se la sua area risulta superiore a tale dimensione.

## 1.8 L'AMBIENTE DI LAVORO

Il modello adottato all'interno del progetto CARGEN per la gestione dei dati è stato quello client-server. Il server 'e costituito da una macchina con installato un DBMS Oracle Spatial 10g, la cui funzione è quella di memorizzare e mantenere i dati spaziali, accessibili tramite query. Nel lato client, invece, sono stati installati i software Geomedia Professional 6 e Dynamo/Dynagen, entrambi di proprietà della Intergraph. Geomedia viene utilizzato principalmente come strumento d'accesso ai dati spaziali; Dynamo/Dynagen sono invece gli strumenti usati durante il processo di generalizzazione cartografica. Questa architettura, tuttavia, è stata in parte abbandonata recentemente: infatti, sia allo scopo di semplificare la sviluppo di nuovi algoritmi, sia allo scopo di migliorare le prestazioni temporali, è stato cambiato metodo d'accesso ai dati ed il software per visualizzarli. I dati vengono caricati in RAM e gestiti proprio come se fossero delle tabelle, grazie ad una libreria sviluppata all'interno del Progetto. Inoltre, una libreria potentissima sviluppata in Java, la JTS, fornisce una vasta serie di operatori spaziali, evitando così di ricorrere al DBMS di Oracle per effettuare le interrogazioni spaziali.

Descriviamo ora brevemente gli strumenti più importanti usati durante l'attività di tesi.

### *Java Topology Suite*

La *Java Topology Suite* (JTS) è una libreria open source che fornisce una modellazione ad oggetti per le geometrie lineari in uno spazio

euclideo.

In questa libreria sono definite tre geometrie fondamentali, *Point*, *LineString* e *Polygon*, che rappresentano rispettivamente la geometria puntuale, lineare e areale. La JTS mette a disposizione numerose funzioni geometriche, tra le quali possiamo citare:

- gli operatori topologici, che realizzano le funzioni di intersezione, differenza, unione;
- la funzione per la creazione del buffer intorno alla geometria;
- la funzione per la costruzione dell'inviluppo convesso;
- alcune funzioni per la semplificazione delle geometrie;
- funzioni per determinare orientamento di un segmento e angolo interno tra due lati consecutivi;
- la funzione per la costruzione del Minimum Bounding Box.

Oltre a queste funzioni, la JTS fornisce l'implementazione di indici spaziali, come il quadtree, che offrono un modo veloce per la risoluzione di query spaziali.

La versione utilizzata per questa tesi è la versione 1.12 datata Giugno 2011

*OpenJUMP*

*OpenJUMP* è un Desktop GIS open source che permette di visualizzare, modificare ed interrogare dati spaziali. OpenJUMP è scritto in Java, si basa sulla JTS ed è in grado di gestire file raster, vettoriali e database (PostGis, Oracle); una caratteristica degna di nota è la sua architettura modulare, che permette di estendere di molto le funzionalità di base, potendo integrare, per esempio, il proprio codice mediante la realizzazione di un plugin.

In OpenJUMP, la creazione di un plugin, relativo al proprio codice, offre al programmatore il grosso vantaggio di poter visionare tramite l'interfaccia grafica gli effetti della propria applicazione. Il plugin diventa così uno strumento essenziale nello sviluppo di nuovi algoritmi che manipolano geometrie: avere una risposta grafica e istantanea è un grande aiuto per semplificare e velocizzare la fase di testing e debug.

La modalità con cui OpenJUMP gestisce le *feature* si basa sull'utilizzo dei *layer*, o livelli, che svolgono il ruolo di contenitori di *feature*. Ogni layer 'e in grado di contenere le *feature* relative ad uno specifico schema dati, chiamato *FeatureSchema*; quest'ultimo specifica il nome e la tipologia degli attributi che costituiscono la *feature*, simile a quanto accade nelle tabelle dei database. Un *layer* rappresenta, quindi, una vera e propria tabella, il cui schema dati è specificato dal *FeatureSchema*.

Il *layer* è interrogabile per mezzo di query, che possono essere sia spaziali che non spaziali. Per migliorare le query spaziali, è possibile associare al *layer* uno degli indici spaziali forniti dalla JTS.

La versione utilizzata per questa tesi è la versione 1.4.2 datata Settembre 2011.

### *Eclipse*

*Eclipse* è un ambiente di sviluppo integrato, multi-linguaggio e multipiattaforma. Ideato nel 2001 da un consorzio no-profit di grandi società quali Borland,IBM, Red Hat, SUSE, Ericsson, HP, Fujitsu, Intel, MontaVista Software, QNX, SAP e Serena Software e chiamata *Eclipse Foundation*, viene supportato da una comunità strutturata sullo stile dell'open source. La licenza dunque di Eclipse è la *Eclipse Public License*, permette cioè di creare prodotti derivati ridistribuibili gratuitamente. Importante caratteristica deriva dal fatto che, essendo scritto in Java, Eclipse è multipiattaforma. È disponibile infatti per le piattaforme Linux, HP-UX, AIX, Mac OSX e Windows. Dal 2006, la Eclipse Foundation ha prefissato un'uscita annuale del suo software.

La versione utilizzata per questa tesi è la versione Indigo datata Febbraio 2012.

# 2

# LA QUALITÀ DEI DATI

All'interno di un GIS i dati sono una risorsa cruciale. La prima cosa di cui bisogna tener presente è che i dati di tipo geografico sono costosi da acquisire, archiviare e manipolare, anche perché solitamente occorrono grandi volumi di dati per risolvere problemi geografici di tipo sostanziale. Una stima indica che i costi di acquisizione dei dati per i GIS sono più del doppio del costo dell'hardware e del software; un'altra stima indica che tali costi influenzano il 70% del costo totale di implementazione di un intero GIS. Questo ci fa capire che, quando si acquisiscono i dati per un GIS, è importante che le informazioni fornite dai dati geografici siano affidabili il più possibile: si deve cioè verificare qual è la *qualità dei dati*.

La qualità dei dati spaziali è sempre stato un problema fondamentale e di interesse nella ricerca; prima degli anni '90, la qualità era associata all'accuratezza posizionale nel passaggio da rilevazione fotogrammetrica a cartina. Successivamente il concetto di qualità si estese anche ai processi di produzione dei dati spaziali, che vanno dalla acquisizione fino alla elaborazione, in quanto tutte le sorgenti e i metodi di produzione presentano errori e il tipo, la gravità e le implicazioni di questi errori determinano la qualità del dato spaziale.

Negli ultimi 30 anni, due rivoluzioni hanno incrementato l'interesse nel campo della qualità dei dati spaziali:

1. il passaggio da mappe cartacee a mappe digitali e lo sviluppo dei GIS. Le mappe digitali restituiscono quella sensazione di grande accuratezza perché possono calcolare le distanze in maniera molto precisa e perché possono visualizzare i dati ad un alto livello di dettaglio.

2. Lo sviluppo di internet e di altre forme di comunicazione che hanno facilitato lo scambio di dati tra individui ed organizzazioni.

Un GIS è un sistema fondamentale per pianificare una gestione del territorio e degli interventi. Proprio per questi aspetti è indispensabile che ci sia un controllo della qualità in tutte le fasi di produzione del dato. L'analisi di qualità deve essere pesata in base all'uso che se ne deve poi fare di tali informazioni, in quanto richiede notevoli risorse (bisogna conoscere l'intero processo di acquisizione, gli strumenti di misura, la fonte).

Nei paragrafi successivi daremo una definizione della qualità secondo le norme ISO e vedremo quali sono le possibili cause della non qualità dei dati spaziali, approfondendo la tipologia degli errori che si possono incontrare ed introducendo il concetto di anomalia, il cui rilevamento è lo scopo di questa tesi. Infine viene presentato un esempio reale di controllo degli errori nei DB Topografici.

## 2.1 DEFINIZIONE

All'inizio del 20° secolo, per la prima volta la parola qualità è stata usata nella produzione di beni. Taylor fu il primo ad approfondire tale concetto nel campo ingegneristico: descrisse un insieme di principi di gestione del lavoro per migliorare la produzione. Questi principi formarono la base del *Taylorismo* e furono per esempio adottati da Ford nei suoi impianti automobilistici. Nel 2005 la qualità è stata definita nella norma ISO 9000 come il "grado in cui un insieme di caratteristiche intrinseche soddisfano i requisiti".

Lo standard ISO 19113 del 2002 stabilisce i principi per la descrizione della qualità dei dati geografici e definisce i componenti per documentare le informazioni relative alla qualità. La norma si applica sia ai produttori di dati che forniscono informazioni di qualità per descrivere e verificare in quale misura i dati corrispondono alla realtà come definito nelle specifiche di prodotto, sia agli utenti per stabilire se la qualità di determinati dati geografici è adeguata per l'applicazione richiesta. La norma dovrebbe essere seguita dagli organismi responsabili dell'acquisizione e vendita in modo da rendere possibile l'adeguamento alle specifiche di prodotto ed essere utilizzata per la definizione di schemi applicativi e per definire requisiti di qualità. Oltre ad essere applicata ai dati geografici codificati numericamente, i principi della norma possono essere estesi per identificare, raccogliere e documentare informazioni relative alla qualità di insiemi di dati o raggruppamenti più piccoli di insiemi di dati. Benché la norma si applichi ai dati geografici rappresentati numericamente, i principi definiti possono essere estesi a molteplici forme di dati geografici quali mappe, carte e documenti testuali.

Gli standard affermano che la qualità deve essere espressa sia a livello di dataset che a livello di singolo oggetto e che è importante, per mantenere una congruenza logica nell'operazione di controllo della qualità, confrontare i dati in possesso con riferimento alla fonte di acquisizione; perciò se il dato è stato ricavato da una cartografia, il riferimento sarà la cartografia stessa e non il mondo reale.

L'ISO 19113 definisce inoltre gli *elementi di qualità* dei dati:

- *Completezza*: grado di conformità del dato digitale rispetto alla fonte in termini di omissioni o eccedenze di oggetti. Si esprime come rapporto percentuale tra il numero di oggetti mancanti o

sovrabbondanti rispetto al numero di oggetti totali presenti sul terreno;

- *Consistenza Logica*: grado di conformità del dato digitale rispetto alla fonte; riguarda gli aspetti topologici, la struttura del file e la validità dei valori tematici. I controlli di consistenza vengono eseguiti con procedure automatiche (tipicamente i controlli riguardano chiusura dei poligoni, connessione dei grafi, intervalli di validità o valori non consistenti di un attributo);

- *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 INCERTEZZE NEI DATI SPAZIALI

Una parte importante della ricerca sulla qualità dei dati spaziali riguarda la descrizione degli errori e delle incertezze nei dati spaziali. In [6] sono state distinte tre forme di incertezza che si presentano durante il processo di derivazione dei dati spaziali dal mondo reale.

Una prima forma di incertezza è data dalla *vaghezza* che dipende fondamentalmente da una documentazione scadente della classe degli oggetti o del singolo oggetto. La seconda forma di incertezza è data dall'*ambiguità* che deriva dal disaccordo sulla definizione degli oggetti in un dataset. Tale disaccordo può nascere perché la definizione non è precisa e per differenze di opinione nell'assegnare un oggetto ad una classe.

Mentre le prime due forme di incertezza derivano da una documentazione scadente, la terza forma di incertezza, l'*errore*, nasce quando le classi degli oggetti sono ben definite. Un errore è definito come la differenza che esiste tra la realtà e la rappresentazione della realtà.

Oltre a queste tre categorie, in un precedente lavoro di tesi [22] realizzato all'interno del progetto CARGEN e argomento di approfondimento in questa tesi, è stato individuato il concetto di *anomalia*, ossia situazioni in cui all'interno di database formalmente corretti l'informazione potrebbe essere, secondo logica comune, errata.

Per capire bene la differenza tra errore ed anomalia, diamo una ulteriore spiegazione dei due concetti.

### 2.2.1 *Errori*

Un errore è definito come la differenza che esiste tra la realtà e la rappresentazione della realtà. È stata proposta in [6] una classificazione di questi errori e delle sorgenti d'errore:

1. errori relativi all'acquisizione dei dati;
2. errori relativi all'elaborazione dei dati;
3. errori relativi all'uso dei dati.

Queste tre categorie di sorgenti d'errore creano *errori primari* (errori posizionali ed errori sugli attributi) ed *errori secondari* (errori di consistenza logica e completezza). È stato individuato che i dati spaziali possono cadere in errore durante le seguenti fasi:

- *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 della informazione, uso scorretto dei dati.

Generalizzando, ogni fase di produzione del dato può portare ad errori di:

- *Misurazione*: errata misurazione di una proprietà di un oggetto. Sono gli errori più facili da individuare perché negli anni sono stati sviluppati molte procedure avanzate di analisi statistica;
- *Assegnazione*: l'oggetto è stato assegnato ad una classe sbagliata a causa di errori di misurazione fatti dai tecnici, o nel campo o in laboratorio, o dai topografi;
- *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.

### 2.2.2 *Anomalie*

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. 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, oppure di inserire un vertice nell'intersezione tra due o più linee. Altri strumenti, come ad esempio il DBTopoCheck sviluppato all'interno del progetto, permettono di verificare la correttezza dei vincoli topologici tra feature definiti nelle specifiche. Possiamo allora affermare che un dataset che risulta formalmente corretto sia esente da errori? La risposta è no. Per capire il perché di questa risposta, consideriamo alcuni esempi significativi con l'aiuto delle figure:

- In Figura 3 in verde sono rappresentati degli oggetti che indicano la presenza di un'area boschiva. Due sono le cose che saltano subito all'occhio: la prima cosa è l'eccessiva regolarità nella forma dell'area rappresentata in centro; la seconda cosa è la dimensione dell'area in alto a destra che sembra troppo piccola. Niente ci impedisce di dire, oltre a questioni di risoluzione della mappa, che tale oggetto non sia effettivamente un bosco, ma d'altro canto il valore semantico dell'oggetto ci porterà a sospettare un qualche errore di classificazione o di rappresentazione;

Figura 3: Possibile errore di classificazione di un bosco

- In Figura 4 è rappresentato un tratto di un grafo stradale. In questo caso i due punti rossi rappresentano una interruzione tra due vie di una città. Anche in questo caso, vista la vicinanza tra i due punti, ci sorge il dubbio che le due strade nella realtà siano in qualche modo collegate;

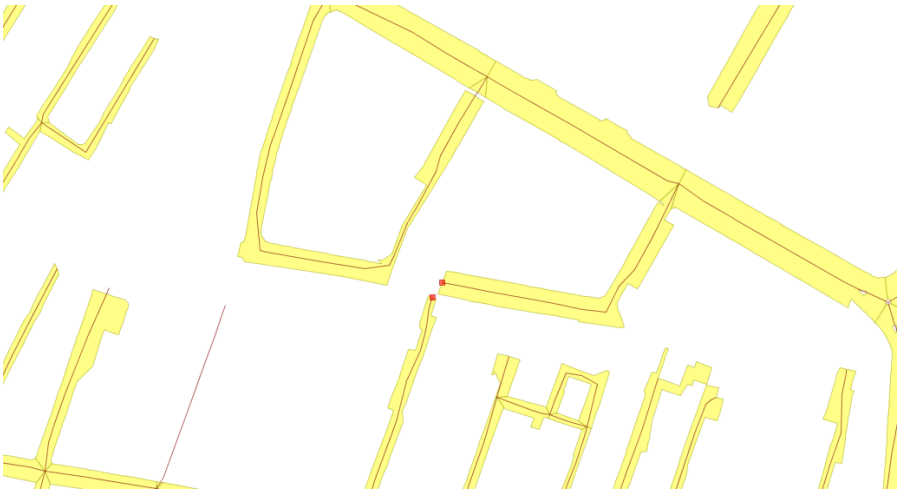

Figura 4: I punti in rosso segnalano una possibile interruzione nel grafo stradale

- Infine, consideriamo il caso visibile in Figura 5, dove in rosso sono identificati gli oggetti classificati come baracche. In questo caso è facile notare come nella parte destra della mappa non ci sia alcuna baracca, a differenza della parte sinistra. Questa situazione si può creare perché le due aree sono state create da produttori differenti che potrebbero aver usato delle specifiche differenti (e per cui da una parte le baracche sono state inserite, dall'altra no), oppure per un effettivo errore di classificazione delle entità.

Gli esempi qui sopra mostrano come questi errori, che d'ora in poi chiameremo anomalie, riguardano la corretta classificazione di un oggetto (come avviene in Figura 3) e l'incompletezza dei dati (Figure 4 e 5). Ciò che però le differenzia dagli errori tradizionali è il fatto che non possiamo classificarli come tali finché non viene fatto un confronto effettivo con la realtà d'interesse (la Figura 4 rappresenta in maniera chiara questo problema).

Partendo da queste considerazioni, possiamo finalmente definire più formalmente un'*anomalia* come una situazione in cui l'informazione rappresentata in un database cartografico potrebbe essere, secondo logica comune, errata e che fornisce una indicazione su un probabile errore non rilevabile automaticamente con i normali controlli formali ma con indagini sul campo.

Principalmente le cause della presenza di un'anomalia sono due: la prima è che le specifiche falliscono nel fornire una formalizzazione

Figura 5: In rosso sono evidenziate le baracche presenti in una mappa. È chiaro come solo la metà sinistra della mappa contiene tutte le baracche

completa dei requisiti necessari a garantire la reale correttezza dell'informazione, mentre la seconda è che ci sono degli errori che non sono rilevabili verificando i vincoli imposti.

Le anomalie si possono trovare solamente con la logica e l'osservazione. Alcuni esempi individuati in [22] possono essere zone di mappa prive di edifici, interruzioni nella strada, laghi naturali di forma quadrata, edifici di forma molto spigolosa, curve brusche nei binari ferroviari, strade a una corsia intervallate da brevi tratti autostradali, rifugi non in montagna, fiumi in salita, ponti senza strade vicine, ospedali senza edifici vicini, boschi di piccole dimensioni. Per capire i tipi di anomalie che possiamo incontrare, nel prossimo paragrafo viene proposta una possibile tassonomia.

## 2.3 TASSONOMIA DELLE INCERTEZZE NEI DATABASE

In questa sezione vedremo una possibile tassonomia dei fattori che causano errori in un database spaziale. La tassonomia proposta riguarda principalmente la violazione delle specifiche e le anomalie. Oltre a queste due categorie, verranno anche analizzati gli errori che si possono verificare in un database spaziale dopo il processo di generalizzazione cartografica, errori che si vengono a creare in quanto gli algoritmi comportano manipolazioni e spostamenti delle geometrie che possono causare errori di classificazione e di violazione topologica.

### 2.3.1 *Violazione specifiche*

Sono errori che derivano dal fatto che alcune geometrie, in base alle specifiche, non dovrebbero essere presenti nel database finale.

Sono soprattutto difetti che derivano dal mancato rispetto dei criteri di selezione per la costruzione del database.

Individuiamo due tipi di violazioni:

1. **Semantiche**: avviene quando non c'è corrispondenza tra la realtà e l'attributo qualitativo associato all'oggetto.
2. **Geometriche**:
   - *Topologiche*: non sono rispettate le proprietà spaziali tra gli oggetti definiti nelle specifiche (relazioni tra feature), come ad esempio distanze minime tra le feature.
   - *Di misura*: violazione dei limiti di acquisizione sulle feature. Riguardano l'acquisizione o no di un elemento lineare in base alla sua lunghezza o di un elemento areale in base alla sua superficie.

### 2.3.2 *Anomalie*

La classificazione tassonomica è stata realizzata in [22] usando un approccio bottom-up con l'individuazione, secondo logica ed osservazione, di una serie di particolari casi base e la creazione di una classificazione generale che può portare successivamente al riconoscimento di nuovi casi base.

Le categorie di anomalie individuate sono:

1. **Forma**: riguarda il contorno e le dimensioni di un oggetto. All'interno di questa categoria possiamo identificare i seguenti casi:
   - forma di un elemento naturale troppo regolare e, viceversa, forma di un elemento artificiale troppo irregolare;
   - angoli troppo acuti tra due lati consecutivi di un elemento artificiale (ad esempio tra due pareti di un edificio);
   - cappi che si formano nelle strutture lineari o sui bordi di strutture areali. Possono essere molto piccoli ma in alcuni casi portano al fallimento degli strumenti di analisi topologici.
   - improvvisa divergenza tra un vertice e i due vicini a esso in una linea o in un bordo di un'area.

   La forma inoltre è fortemente correlata con le dimensioni, dimensioni legate a limiti di:
   - *Risoluzione*: ad esempio nella scala 1:25.000, assumendo che la distanza minima che un occhio umano è in grado di percepire è pari a $0,1$ mm e che rappresentano $2,5$ m nel mondo reale, non sono possibili pareti di edifici inferiori a $2,5$ m nelle strade e laghi con aree inferiori a $6,25$ m$^2$ ;

- *Semantico*: ad esempio è accettabile una casa di 1 m² (baracca) ma non un bosco di 2 m².

2. **Distribuzione**: situazioni in cui la distribuzione degli oggetti di una classe in una mappa, in assenza di ulteriori informazioni, risulta anomala rispetto a quanto ci si può aspettare;

3. **Posizione**: posizione errata nello spazio di un elemento:
   - *Posizionamento logico*: incongruenza tra il contesto in cui si trova l'oggetto e la classificazione dell'oggetto stesso. Ad esempio in questa categoria si possono includere strutture portuali in montagna, ponti senza strade, ospedali senza edifici vicini, casa in mezzo ad un lago senza la presenza di un'isola a cui può appartenere;

4. **Grafi**: riguarda anomalie che alterano le proprietà di un grafo. Questa categoria è stata creata per raggruppare anomalie di diverso tipo ma tutte relative a strutture a grafo. Questo è stato fatto per sottolineare che molte di queste anomalie sono frutto delle particolari caratteristiche che un grafo deve rispettare come ad esempio la connettività:
   - *Interruzioni*: vuoto di piccole dimensioni tra due elementi lineari che sembrano congiunti e che portano al fallimento degli algoritmi di navigazione. Ad esempio strada o binario che s'interrompe per poi ripartire dopo una distanza superiore alla risoluzione della scala ma inferiore ad un certo valore di soglia fissato;
   - *Tratto isolato o troppo corto*: tratto di strada non connesso al resto della rete stradale;
   - *Mancanza vertice comune incrocio*: caso in cui dei tratti del grafo si sovrappongono senza la presenza di un punto in comune. Capita soprattutto nel grafo stradale: esempio classico è l'incrocio a T dove o le due strade si vanno a sovrapporre perché una delle due è rappresentata più lunga del vero, o manca effettivamente il punto che rappresenta l'incrocio, oppure una delle strade potrebbe essere effettivamente una sopra l'altra e manca l'informazione;
   - *Incoerenza di tratti adiacenti*: causato da brevi cambiamenti nel tipo di dato che generano soprattutto errori di vestizione. Ad esempio una strada che da locale diventa improvvisamente autostrada.

### 2.3.3 *Errori di generalizzazione*

Gli errori della generalizzazione violano norme/consigli/regole del processo di generalizzazione. Il risultato del processo restituisce sem-

pre un dataset, per cui in questo caso troviamo anomalie e violazioni delle specifiche che si possono trovare anche nel dataset originale.

Gli algoritmi però modificano ed eliminano in buona parte la geometria e i pattern presenti nei dati originali in quanto lo scopo è quello di semplificare la struttura.

Oltre a questi errori, possiamo aggiungere ulteriori due tipi di difetti/errori:

1. **Di Traduzione**: traduzione errata dei dati sorgente nel modello target, poiché non c'è un match perfetto tra i due modelli. Può anche derivare da errori di geometria, come ad esempio se la larghezza di un fiume è valutata incorrettamente, il fiume può essere generalizzato in una classe sbagliata.

2. **Cartografici**: alcuni difetti per cui risulta che i dati generalizzati non diano una buona rappresentazione. Sono difetti che non derivano da errori sulle specifiche o sulle geometrie, ma che influiscono in generale nell'abilità di una mappa di essere una buona fonte di informazione per l'utente, come ad esempio la leggibilità. Sono difetti che possono ad esempio emergere nella vestizione, la quale esalta le reali dimensioni dei particolari geografici rappresentati e per cui è necessaria la variazione della posizione planimetrica e/o alla modifica della forma degli oggetti in modo da evitare sovrapposizioni grafiche tra i vari segni.

## 2.4 UN CASO REALE: IL DB TOPOGRAFICO

Abbiamo già parlato in precedenza del GeoDBR, il DBT implementato dalla Regione Veneto. In questo paragrafo approfondiamo i controlli di qualità previsti da IntesaGIS [9] e uno strumento di verifica dei dati realizzato, il GeoUML Validator [21].

### 2.4.1 *Le specifiche di contenuto*

Una specifica di contenuto di norma descrive il contenuto informativo di un database per mezzo di diversi tipi di definizioni che hanno scopi diversi. Tali definizioni sono:

1. definizione degli *elementi informativi*, cioè degli elementi che devono essere rappresentati nel database, indipendentemente dalla tecnologia di memorizzazione dei dati. Alcuni di questi elementi sono la classe, gli attributi, la componente spaziale, la chiave primaria e lo strato topologico;

2. definizione dei *vincoli d'integrità*, cioè delle proprietà intrinseche (proprietà verificabili sugli elementi informativi stessi senza os-

servazioni nel mondo reale) che gli elementi informativi devono soddisfare. Possono essere suddivisi in:

a) vincoli topologici, che definiscono le relazioni topologiche che devono sussistere tra le classi, come ad esempio adiacenza, sovrapposizione, contenimento;

b) vincoli di composizione, che definiscono vincoli di appartenenza (ad esempio "la superficie dell'area stradale è composta di oggetti delle classi area di circolazione veicolare, pedonale, ciclabile ed eventuali manufatti dell'infrastruttura di trasporto") e vincoli di partizionamento (ad esempio "il territorio di una regione è partizionato nel territorio delle provincie in cui è scomposta").

3. *elementi descrittivi*, ossia informazioni utilizzabili per interpretare il contenuto del database in termini di realtà rappresentata e, viceversa, informazioni su come tradurre un oggetto della realtà in un dato spaziale.

Le specifiche di contenuto sono il più possibile astratte, nel senso che non si vanno a definire gli aspetti tecnologici. Quindi per poter poi effettivamente creare un dataset bisognerà specificare in che modo verrà implementato, definendone la struttura (schema fisico) e le corrispondenze tra elementi dello schema fisico e delle specifiche di contenuto (mapping fisico).

Affinché il database sia popolato correttamente, i dati spaziali devono essere conformi alle specifiche; in particolare la conformità di un dataset ad una specifica di contenuto è composta da due aspetti:

1. la *conformità reale* che riguarda la corrispondenza tra il contenuto informativo del dataset e il territorio di riferimento al quale il dataset si riferisce; tale corrispondenza è valutata in base alla definizione degli elementi informativi e dei relativi elementi descrittivi definiti nelle specifiche;

2. la *conformità intrinseca* che riguarda la consistenza dell'informazione contenuta nel dataset; diremo che un dataset è intrinsecamente conforme a una specifica di contenuto se e solo se tutti i dati contenuti nel dataset corrispondono a elementi informativi e soddisfano tutti i vincoli di integrità della specifica

Gli strumenti di validazione automatica, come ad esempio GeoUML, prendono in considerazione solamente la conformità intrinseca di un dataset in quanto esistono delle metodologie implementate in linguaggio macchina che permettono di valutare i vincoli topologici e di composizione.

### 2.4.2 *Controlli di qualità*

La determinazione della qualità del DBT viene fatta mediante l'analisi di alcune grandezze detti *elementi di qualità*.

Nella fase di progettazione del DBT vengono definiti dei limiti sui valori di ogni singolo elemento, ed è quindi importante che la qualità finale dei dati del DBT, o un suo campione significativo, rispetti tali valori.

Gli elementi di qualità sono stati estratti dai documenti ISO 19113 – Quality Principles, ISO 19114 – Quality Evaluation Procedures, ISO 19138 – Quality Measures, e che abbiamo già trattato nel primo paragrafo. Per ogni elemento di qualità viene indicato un insieme di controlli indicativi, flessibili in base ai bisogni dell'utente finale, che si possono effettuare sui dati nelle varie fasi di produzione del dato.

1. *Consistenza logica*:
    - di *formato*, cioè la struttura fisica dei dati deve essere conforme alle specifiche riguardo al formato del DB, delle tabelle che lo compongono e degli attributi. Nello specifico il *data type*, i codici degli oggetti e degli attributi e la componente spaziale di ogni oggetto deve corrispondere a quelle indicate in tabella, mentre gli attributi obbligatori devono essere tutti popolati;
    - di *dominio*, cioè i valori degli attributi devono essere coerenti con quanto definito nelle specifiche. Nello specifico il valore degli attributi numerici devono essere riportati nell'unità di misura richiesta e deve rientrare nel range definito mentre le stringhe alfanumeriche devono rispettare le regole delle specifiche;
    - di *geometria*, cioè non ci devono essere difetti nella geometria. Nello specifico non ci devono essere duplicazione di parti di un oggetto e la presenza di vertici doppi in una geometria, i poligoni devono essere chiusi, i vertici quotati e tutta la geometria deve essere contenuta all'interno dell'area prefissata (foglio, sezione, ...);
    - di *topologia*, cioè devono essere rispettate le relazioni topologiche tra gli oggetti. Nello specifico tutte le aree devono essere classificate, la rappresentazione lineare di un oggetto deve essere completamente interna alla sua rappresentazione areale, i poligoni devono avere nodi in corrispondenza della intersezione con altre geometrie lineari, areali o puntuali, il grafo delle reti deve essere connesso.
2. *Accuratezza posizionale*: è riferita alla posizione assoluta del singolo punto e vertice. Nello specifico vengono valutate separatamente sia la componente altimetrica che quella planimetrica. Per

la componente altimetrica viene indicato il valore di scostamento relativo in altezza ($\Delta H$), mentre per la componente planimetrica viene indicato o le componenti di scostamento est-nord($\Delta E$, $\Delta N$) oppure più semplicemente la distanza. La valutazione di questa correttezza viene fatta tramite un confronto tra i dati ottenuti per restituzione fotogrammetrica e i dati rilevati da misurazioni dirette sul terreno.

3. *Completezza*: la valutazione consiste nel verificare che per ogni classe l'errore percentuale in termini di completezza sia tale da rispettare l'attendibilità ammessa, quantità che esprime in percentuale la presenza attesa della classe nel dataset. Tali valori di attendibilità sono espressi in una tabella.

4. *Accuratezza tematica*:
   - di *classificazione*, cioè l'oggetto e i suoi attributi devono essere correttamente codificati in base alla specifica e completati al meglio delle informazioni previste dai relativi attributi. Nello specifico in fase di editing si possono modificare i valori degli attributi che non rispettano le specifiche di classe.
   - di *toponomastica*, cioè le classi di toponimi e attributi devono corrispondere a quelli reali. Nello specifico, ad esempio, nella rappresentazione grafica, si esamina la corrispondenza tra oggetto topografico e font/dimensione/colore assegnati al relativo toponimo.

Gli elementi di qualità indicano in termini quantitativi il livello di qualità di un dataset, per cui per ogni elemento viene definito un valore percentuale di tolleranza minimo da rispettare (sull'intero dataset o su un campione rappresentativo) che può dipendere oppure no dalla scala e/o dalla classe.

### 2.4.3 *GeoUML Methodology*

Quando parliamo di valutazione della qualità di un dato bisogna tenere conto sia della conformità reale sia della conformità intrinseca delle specifiche. Nei database topografici i vincoli riguardano principalmente la conformità intrinseca e sono gli unici che possono essere valutati automaticamente. Per la gestione delle specifiche di contenuto del DBT e di verifica di conformità intrinseca dei dati a tali specifiche sono stati realizzati alcuni strumenti inseriti nella **GeoUML Methodology**, realizzati dallo SpatialDBGroup del Politecnico di Milano, con un progetto cofinanziato dalle Regioni attraverso il CISIS (Centro Interregionale per i Sistemi informatici, geografici e statistici). Due di questi strumenti implementati sono il GeoUML Catalogue e il GeoUML Validator.

Il **GeoUML Catalogue** permette di visualizzare e modificare le specifiche di contenuto arricchite con testi, immagini e diagrammi descrittivi. Per ogni classe sono poi elencati i vincoli topologici o di composizione che devono essere rispettati. Permette inoltre di tradurre una specifica nelle strutture fisiche disponibili.

Il **GeoUML Validator** è invece lo strumento atto a verificare il controllo di conformità intrinseca di un generico dataset relativamente ad una specifica di contenuto gestita dal Catalogue. Il Validator esegue l'importazione di una specifica di contenuto generato con Catalogue, la configurazione dei controlli da eseguire, la validazione dei dati e la produzione della diagnostica.

# 3

# APPROCCI E STRUMENTI PER LA RICERCA DELLE ANOMALIE

Scopo principale di questo lavoro di tesi è quello di trovare delle metodologie automatiche per rilevare le anomalie presenti in un dataset. In questo capitolo vengono perciò spiegati gli approcci e gli algoritmi realizzati per la valutazione di alcune tipologie di anomalie elencate precedentemente, suddivise in base alla tassonomia proposta.

## 3.1 DETTAGLI IMPLEMENTATIVI DEL SOFTWARE

La struttura del software realizzato segue nella concezione la struttura del Topological Tester realizzato all'interno del progetto; ho scelto questa strada visto lo scopo comune dei due software (entrambi eseguono delle analisi di controllo topologico sui dati di un DBMS) e per semplificarne l'utilizzo ad un utente pratico del Topological Tester stesso, il quale si troverà di fronte ad una organizzazione concettuale degli strumenti familiare. Entrando più nel dettaglio, il software realizzato ha lo scopo di controllare le anomalie sui dati memorizzati in un DBMS in base ad una serie di regole. Le regole sono memorizzate dall'utente in un file .txt secondo una sintassi predefinita. Il file di controllo è un semplice file ASCII che contiene le istruzioni su come il software si deve comportare. Ogni file di controllo contiene:

- informazioni sulla connessione al DBMS dove sono presenti le tabelle da analizzare e dove saranno memorizzate quelle dei risultati. Il formato è il seguente:

  ```
  Conn: type_of_connection username password host
  port database_name
  ```

  dove in particolare `type_of_connection` indica il tipo di DBMS a cui ci si collega (Oracle o PostGIS), `username` e `password` i dati di autenticazione al DBMS, `host` l'indirizzo IP di collocazione del DBMS, `port` la porta sul server per accedere al servizio, `database_name` il nome del database dove sono contenute le tabelle;

- Il nome della tabella di default dei risultati;

- una o più regole, dove ogni regola deve almeno specificare il tipo di anomalia da controllare, ossia l'operatore da chiamare, e una lista delle tabelle coinvolte nel check. In base poi al tipo di operatore, si potranno inserire ulteriori parametri, come ad esempio vincoli su distanze o lunghezze, filtri su un attributo di una tabella per la selezione di determinati elementi oppure opzioni supplementari.

Il software si occupa poi di caricare le varie regole e di controllarle una ad una, memorizzando poi i dati in un DBMS che l'utente potrà poi visualizzare in OpenJUMP o in un qualsiasi altro software GIS, in modo da controllare visivamente i risultati. In questa versione, sia per una migliore visione e suddivisione dei risultati in fase di test, sia per non avere primitive geometriche differenti in una stessa tabella, ho deciso di creare una tabella per ogni tipo di check (ogni tabella ha però lo stesso schema, simile a quello proposto nel Topological Tester), aggiungendo al nome della tabella di default un suffisso che indica l'operatore. In futuro, se ritenuto necessario, i risultati possono venire inseriti in un unica tabella. Il software realizzato ha tre componenti principali:

1. il primo componente consiste nel file `Start.java` che contiene il codice per selezionare il file di controllo, per leggerlo e per lanciare l'operatore di controllo della anomalia;

2. un secondo componente consiste nei vari operatori organizzati in classi Java; ogni operatore è implementato estendendo la classe astratta `AbstractCheck.java` e deve implementare un metodo `compute()` che rappresenta il suo corpo principale e un metodo `getName()` che restituisce il nome dell'operatore;

3. il terzo componente è il file che contiene le regole di controllo.

La struttura del software è modulare e permette in futuro l'inserimento di nuovi operatori; i nomi dei nuovi operatori verranno inseriti in una HashMap creata nel momento in cui viene lanciato `Start.java`. La HashMap contiene il nome simbolico e l'implementazione corretta dell'operatore. È stato inoltre realizzato un plugin per OpenJUMP che fornisce una visione immediata dei risultati caricando i file shape necessari e che permette di non dover passare ogni volta attraverso un DBMS.

Gli algoritmi sono stati sviluppati usando le solide librerie JTS (*Java Topology Suite*), quelle di OpenJUMP e altre librerie sviluppate all'interno del progetto CARGEN, soprattutto per le interazioni con il database e per la creazione di indici spaziali. All'interno del progetto è stato scelto come ambiente di sviluppo Eclipse, strumento che ci permette anche di sviluppare e testare facilmente dei plugin per OpenJUMP, utili per visualizzare rapidamente i risultati di applicazione degli algoritmi.

Nei paragrafi successivi vengono descritti gli operatori realizzati in questo lavoro di tesi.

## 3.2 FORMA

### 3.2.1 *Contorni*

Analizzando il contorno di un edificio possiamo imbatterci in un paio di situazioni anomale: una prima situazione è il caso di pareti troppo corte per essere visibili ad una certa scala mentre una seconda situazione è la presenza di spigolosità eccessiva tra due pareti. Queste situazioni si possono creare sia in fase di digitalizzazione della mappa che, soprattutto, durante la generalizzazione degli edifici, che va a modificare in maniera profonda la forma delle geometria. Per capire quanto la forma di un edificio possa cambiare prima e dopo la generalizzazione, diamo un'occhiata al processo, riferendoci agli algoritmi di manipolazione sviluppati nel progetto CARGEN, nello specifico fusione, aggregazione, semplificazione e squaring. Nella fusione, due edifici che sono adiacenti vengono fusi insieme; gli edifici che non possono essere fusi insieme ma che sono entro una certa distanza, vengono uniti tramite il MBR[1] dell'inviluppo convesso della intersezione tra i buffer di raggio R (raggio pari alla distanza massima tra due edifici data delle specifiche) dei due edifici candidati (solo se l'intersezione è abbastanza grande). Dopo l'unione, la geometria viene semplificata riducendo il numero di vertici, con algoritmo di Douglas-Peucker, ed eliminando le piccole pareti, con l'algoritmo di Sester. Infine, l'operazione di squaring mira a dare una forma squadrata ad alcuni edifici con forma più o meno rettangolare per migliorarne l'aspetto estetico.

L'algoritmo sviluppato si preoccupa perciò di valutare queste due situazioni:

- *Lunghezza minima dei lati*: intende valutare il rispetto del valore minimo di lunghezza di un lato nella scala di rappresentazione (ad esempio, nella scala 1:25.000, il lato minimo è di $2,5$ m, che equivale a $0,1$ mm). Tali lati non sono visibili nella scala di rappresentazione. L'algoritmo si occupa di dividere una linea o i bordi di un poligono in un insieme di segmenti costruiti partendo dai vertici e di analizzarne poi le loro lunghezze. Questo metodo banale non prende però in considerazione il fatto che molte geometrie presentano dei vertici supplementari che solitamente vengono inseriti per rispettare vincoli topologici tra oggetti differenti. Questi vertici, nella maggior parte dei casi, non creano una variazione sensibile di orientamento tra i due segmenti con questi vertici in comune. L'idea perciò è

1 Minimum Bounding Rectangle

quella di, dopo aver determinato l'insieme dei segmenti, unire in un'unica linea tutti quei segmenti consecutivi che hanno lo stesso orientamento entro una certa tolleranza. Per capire bene il funzionamento, consideriamo una linea definita da $n$ vertici e che quindi avrà $n\text{-}1$ segmenti. A partire dal primo segmento, si va a valutare se l'angolo interno tra questo segmento e il successivo è, entro una certa tolleranza, assimilabile ad un angolo piatto. Se ciò accade, allora si crea una linea unendo i due segmenti; al secondo passo, il calcolo dell'angolo interno avviene considerando *l'anchor line* della linea (ovvero il segmento che si crea unendo i due estremi) e il segmento successivo. Se il test è ancora superato, la linea e il segmento vengono uniti e si procede così fino all'ultimo segmento. Se il test invece non è superato, allora la linea viene inserita in un array e si ricomincia come nel caso base. Quando tutte le linee sono state trovate ed inserite nell'array, si procede al calcolo della loro lunghezza e si memorizzano quelle che stanno sotto il valore di soglia. Un caso particolare si ha nel caso dei poligoni. Per costruzione, i loro bordi sono sempre chiusi. L'algoritmo calcola le linee procedendo sempre lungo una direzione fino ad arrivare di nuovo al vertice di partenza. Perciò può capitare che l'unione tra la prima e l'ultima linea trovata sia anch'essa assimilabile ad una linea retta. Perciò, prima del calcolo delle lunghezze, le due linee sono sottoposte allo stesso test sull'angolo interno ed eventualmente unite.

- *Spigolosità*: il metodo implementato è molto semplice: dato un valore d'angolo $\phi$ in input, viene calcolato l'angolo interno tra due pareti consecutive e se il valore è compreso tra $\pm\phi$ il vertice d'unione delle due pareti viene segnalato come anomalia.

### 3.2.2 *Divergenza punto*

Le divergenze (*Kink*) si verificano quando in un elemento lineare o nel bordo di un'area si crea un'improvviso scostamento tra un vertice e i due vicini ad esso. L'algoritmo si occupa di individuare gli angoli dei vertici, come avviene per l'algoritmo precedente, dato un valore di soglia. Se il test nel vertice $i$ non è superato, allora $i$ è candidato ad essere segnalato come eventuale anomalia. A questo punto, per vedere se effettivamente siamo di fronte ad una divergenza e per non finire nel caso precedente di valutazione della spigolosità, si prendono in considerazione i vertici $i\text{+}1$ e $i\text{-}1$ e si controlla che i segmenti composti dai vertici ($i\text{-}2$,$i\text{-}1$) e ($i\text{+}1$,$i\text{+}2$) abbiano lo stesso orientamento entro una certa tolleranza e che i vertici $i\text{-}1$ e $i\text{+}1$ siano entro una certa distanza ragionevolmente bassa: se questo accade allora siamo di fronte ad una possibile divergenza. Il calcolo dell'orientamento di un segmento viene fatto utilizzando il metodo `angle()` della classe

`LineSegment` di JTS, che fornisce un valore compreso tra $-\pi$ e $+\pi$. Un caso particolare si ha per le linee, dove se il secondo vertice della linea non supera il test sul valore di soglia dell'angolo, la divergenza viene immediatamente segnalata in quanto non è possibile creare il segmento (*i-2*,*i-1*) in quanto il vertice *i-2* non esiste. La stessa cosa vale per il penultimo vertice in quanto non si può creare il segmento (*i+1*,*i+2*).

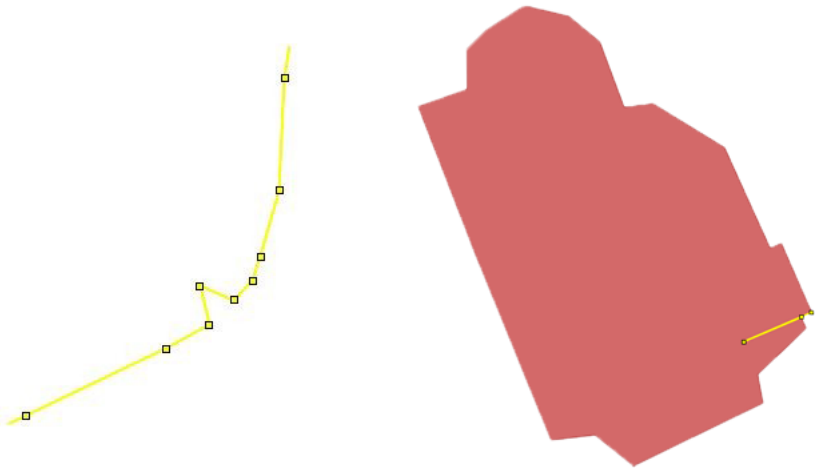

Figura 6: Divergenza in una linea a sinistra e del bordo di un edificio a destra

### 3.2.3 *Forme regolari ed irregolari*

Questo tipo di anomalia si verifica quando:

- forme di elementi naturali sono troppo regolari: un esempio può essere dato da un lago artificiale (Figura 7)
- forme di elementi artificiali sono troppo irregolari: un esempio può essere dato da una pista da go-kart che può essere rilevata come anomalia data la sua tortuosità (Figura 8).

D'ora in poi con il termine edificio per i poligoni e con ferrovia per le linee intendiamo elemento regolare, mentre con lago e corsi d'acqua intendiamo elemento irregolare.

L'idea alla base di questo algoritmo è quello di sviluppare una serie di criteri per permettere di capire la forma di una determinata geometria. Un modo per descriverla è quello di generare un vettore i cui valori vengono calcolati da misure specifiche sulla forma. Lo scopo è quello di ottenere, a partire dal vettore, una classificazione delle geometrie in modo tale che forme diverse possano essere distinte tra loro e forme simili abbiano un vettore simile. Una scelta adeguata di queste misure ci darà la possibilità di valutare le anomalie.

Definiamo innanzitutto una misura come una procedura per compiere misurazioni - cioè assegnare un valore numerico ad una osserva-

Figura 7: Forma naturale troppo regolare

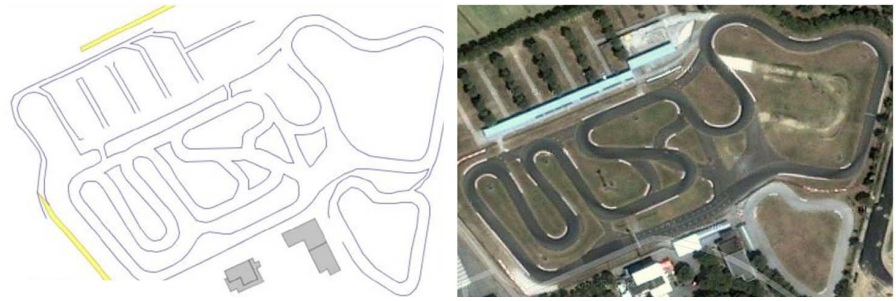

Figura 8: Forma artificiale troppo irregolare

zione che riflette l'importanza di una caratteristica - in modo tale da valutare le caratteristiche di una entità geografica. Una misura, per essere utile, deve soddisfare alcuni criteri:

- idealmente deve descrivere la proprietà in maniera più precisa possibile,
- deve essere invariante alle trasformazioni geometriche come traslazione, rotazione e cambiamento di scala,
- deve produrre risultati diversi in caso di configurazioni differenti e risultati simili per configurazioni simili,
- deve essere facile da calcolare, interpretare ed usare, con un numero limitato di parametri in input.

Una classificazione delle misure propone la loro suddivisione in base alla principale caratteristica che intende rappresentare. Abbiamo perciò misure di dimensione, di forma, di distanza, di topologia, di densità e distribuzione, sui pattern, semantiche. Queste vengono usate soprattutto per studiare delle buone strategie di generalizzazione (ad esempio nel progetto AGENT [1]) in quanto, essendo veloci da calcolare rispetto ad una trasformazione geometrica, minimizzano l'uso di meccanismi di *backtracking* in caso di conflitti.

Per l'implementazione di questo algoritmo sono state scelte ed introdotte misure sia per oggetti areali che per oggetti lineari. Per quanto riguarda i poligoni abbiamo:

- *Squareness*: misura che indica se un oggetto assomiglia ad un quadrato. Calcolato come $\frac{16Area}{Perimetro^2}$. Va da 0 per i segmenti a $4/\pi$ 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. Varia tra 0 e 1, più è alto e più una figura è spigolosa. Per non inficiare la corretta misurazione, come per l'algoritmo di valutazione della spigolosità tra due lati, non vengono presi in considerazioni i vertici che i cui angoli interni sono simili ad un angolo piatto.

- *Perpendicolarità*[2]: percentuale del numero di angoli retti rispetto al numero di vertici (vertici contanti come in precedenza). Anch'essa varia tra 0 e 1.

Per gli elementi lineari invece sono state implementate le seguenti misure:

- *Inflection points ratio*: lo scopo di questa misura è quello di determinare i "punti di flessione" di una linea, ovvero i punti in cui cambia la curvatura della linea. Per evitare che il calcolo dei punti di flessione sia influenzato dai disturbi sulla linea, come prima cosa l'algoritmo esegue una sua semplificazione utilizzando l'algoritmo di Douglas-Peucker. Nella semplificazione è importante la scelta del parametro di distanza, in quanto più alto è il suo valore e minore saranno i punti di flessione. Successivamente, per ogni vertice della linea eccetto gli estremi, viene trovata la direzione della curvatura. Per trovare la direzione, si immagini di costruire un parallelogramma che ha come base di partenza i due segmenti $S_1$ e $S_2$ della linea che hanno il vertice $v$ in comune, in modo così da trovare il vertice opposto a $v$ e quindi una delle due diagonali del parallelogramma. La direzione dipende da dove si trova la diagonale rispetto a un segmento tra $S_1$ e $S_2$; per cui diremo che la curvatura è verso sinistra se, fissato un segmento $S \in \{S_1, S_2\}$, la diagonale è situata nel semipiano sinistro formato da $S$ e, viceversa, diremo che è a destra se la diagonale è situata nel semipiano destro di $S$. Di seguito è riportato il codice realizzato in Java di questa prima parte:

```
1 int[] tpi;
2 Geometry g;
```

2 La perpendicolarità è in realtà un concetto geometrico che indica la presenza di un angolo retto tra due entità geometriche come ad esempio due rette in un piano, una retta ed un piano o due piani incidenti nello spazio. In questa tesi viene usata tale parola per una questione di comodità

```
3  if(g.getCoordinates().length>3){
4    Coordinate coord[] = g.getCoordinates();
5    boolean[] tab_condition = new boolean[coord.length-1];
6    tab_condition[0] = tab_condition[tab_condition.length-1] = true;
7
8    for(int i=1; i<coord.length-2;i++){
9        LineSegment ls1 = new LineSegment(coord[i], coord[i-1]);
10       LineSegment ls2 = new LineSegment(coord[i], coord[i+1]);
11
12       double x1 = coord[i-1].x;
13       double y1 = coord[i-1].y;
14       double x2 = coord[i].x;
15       double y2 = coord[i].y;
16       double x3 = coord[i+1].x;
17       double y3 = coord[i+1].y;
18
19       //trovo punto per la diagonale del parallelogramma
20       double x4 = x1-x2+x3;
21       double y4 = y1-y2+y3;
22
23       Coordinate pt4th = new Coordinate(x4,y4);
24
25       LineSegment vprod = new LineSegment(coord[i],pt4th);
26       //0: destra, 1: sinistra
27       tab_condition[i] = vprod.orientationIndex(ls1) == 1;
28   }
```

Le direzioni della curvatura sono memorizzate in un vettore booleano (sinistra=*true*, destra=*false*).

Per ogni sequenza di vertici che hanno la stessa direzione, viene calcolato il numero di vertici nella sequenza e memorizzato in un vettore. Per eliminare ulteriori punti di micro-flessione, se l'elemento i-esimo del vettore ha valore '1', ovvero un gruppo con un unico elemento, questo comporta una fusione tra l'elemento $i$-esimo e gli elementi $(i-1)$-esimo e $(i+1)$-esimo.

Ogni punto corrispondente alla fine di un gruppo è considerato come il punto che precede il punto di flessione. Il vettore dei punti di flessione è dato dalla somma progressiva dei valori del vettore precedente e dal primo e ultimo punto della linea; da questo vettore vengono trovate le coordinate nella linea. È da tenere presente che il vettore è calcolato nella linea semplificata, ma è comunque valido nella linea originale. Viene infine restituito il rapporto tra il numero di punti di flessione e la lunghezza della linea;

- *Anchor line ratio*: definito come il rapporto tra la lunghezza di una linea e la lunghezza della sua anchor line. L'anchor line è costruita collegando gli estremi della linea presa in considerazione. Più il valore di questo rapporto è vicino ad 1 e più la linea è rettilinea.

- *Concurrence ratio*: è definito come il numero di volte in cui la linea attraversa l'anchor line. In pratica viene calcolato come il numero di punti di una linea che interseca la sua anchor line. Il valore viene poi normalizzato rispetto alla lunghezza della linea.

Un approccio generale nel risolvere problemi di classificazione è quello di fornire inizialmente un training set dove la classe d'appartenenza di ogni oggetto è conosciuta. Il training set è usato per costruire un modello di classificazione il quale poi viene successivamente applicato al test set. La valutazione della classificazione è basata sul conteggio del numero di record di test classificati correttamente e incorrettamente dal modello. Un modo per risolvere un problema di classificazione è quello di porre una serie di domande astute sui record di test. Ad ogni risposta, viene posta una ulteriore domanda fintanto che riceviamo una risposta sulla possibile classe del record. Uno dei metodi che si possono usare è la tecnica dell'albero delle decisioni.

L'*albero delle decisioni* è strutturato in modo tale che le foglie rappresentino le classificazioni e le ramificazioni l'insieme delle proprietà che hanno portato a quella classificazione. Ad ogni nodo interno viene associata una condizione di split sulla base del quale avviene la ripartizione dei dati. I nodi sono etichettati con il nome degli attributi mentre i rami sono etichettati con i possibili valori dell'attributo su cui si fa lo split. Per determinare il miglior modo di dividere i rami dell'albero, tra gli indici di split più usati abbiamo la variazione d'entropia, che si basa sul concetto di entropia definito in teoria dell'informazione[3], e il *Gini Index*, definito come

$$Gini(t) = 1 - \sum_{j=1}^{m} p(i|t)^2$$

dove $p(i|t)$ rappresenta la frazione dei record che appartengono alla classe $i$ nel nodo $t$. Questi indici rappresentano il grado di impurità dei nodi. Per determinare qual è il miglior attributo da usare come condizione di test, viene confrontato il grado di impurità del nodo genitore, prima di subire split, con il grado di impurità del nodo figlio, dopo aver subito lo split. Maggiore è questa differenza e migliore è la condizione di test. Il guadagno $\Delta$ è un criterio per determinare la bontà di uno split ed è definita come:

$$\Delta = I(genitore) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

dove $I(\cdot)$ è il grado di impurità in un dato nodo, $N$ è il numero totale di record nel nodo genitore, $k$ è il numero di valori dell'attributo e $N(v_j)$ è il numero di record associati al nodo figlio $v_j$. Un albero di decisione usa perciò una condizione di test che massimizza il guadagno. Se gli attributi sono continui, si suddivide in $n$ parti il range dei valori in cui una misura può stare e si valuta la media pesata del grado di impurità dei record che stanno sopra e sotto tale valore. Si sceglierà

3 In teoria dell'informazione, l'entropia è data da $\sum_{i=1}^{n} P(x_i) \log_2 \frac{1}{P(x_i)}$

come candidato il valore che minimizza il grado di impurità. Dopo aver valutato tutti gli attributi, si andrà a scegliere come attributo di split quello che ha il grado di impurità minore tra tutti i candidati e si procede poi in maniera ricorsiva per creare i sottoalberi del nodo preso in considerazione. L'albero ad un certo punto subisce *pruning*, nel momento in cui o tutti gli attributi sono stati valutati oppure verificando se il miglioramento nell'accuratezza giustifica la presenza aggiuntiva di ulteriori nodi.

Per risolvere il nostro problema di classificazione per i poligoni, è stato scelto come test set un insieme di 100 edifici e 100 laghi rappresentati in scala 1:5.000 presi rispettivamente dal DBT della regione Liguria e dal CTRN della regione Veneto. Seguendo i passi per la costruzione dell'albero, l'attributo che minimizza il grado di impurità, ossia il *Gini Index*, è risultato essere l'indice di spigolosità, scegliendo come valore di soglia dell'angolo un ampiezza pari a $\frac{3}{4}\pi$. In particolare, lo split viene fatto per un valore pari a 0.6 come possiamo vedere in Figura 9.

Per quanto riguarda la classificazione delle linee, è stato scelto come test set un insieme di 100 corsi d'acqua e 100 binari ferroviari rispettivamente dal DBT della regione Liguria e della regione Veneto. Applicando l'algoritmo per la costruzione dell'albero di decisione, si è visto che il migliore split iniziale si ha con l'indice *inflection point ratio*. Scegliendo questo indice otteniamo un *Gini index* minore rispetto agli altri indici quando lo split viene fatto per un valore pari a 0.0116. In Figura 10 l'albero di decisione ottenuto.

## 3.3 DISTRIBUZIONE

Analizzando gli elementi di una mappa digitale, possono accadere alcune situazioni anomale nella loro distribuzione nello spazio. Prendiamo ad esempio in considerazione gli elementi classificati come 'baracca' in una determinata zona dello spazio e supponiamo di trovare che una parte della mappa contiene molte baracche mentre l'altra metà no, situazione ebn visibile in Figura 5. Ovviamente questo è possibile nella realtà, ma è comunque una situazione sospetta. Perciò, in assenza di ulteriori informazioni sul contesto questa è un'anomalia che deve essere segnalata.

Un approccio generale è quella di contare il numero di oggetti della stessa tipologia presenti in una porzione dello spazio. Per fare questo, ci viene in aiuto un algoritmo realizzato in precedenza in [14] per il calcolo della densità di elementi in una mappa. L'idea è quella di valutare l'affollamento di un'area in base al numero di vertici dell'elemento, sia esso un punto, una linea o un'area, presente all'interno della stessa. Per fare questo viene creata una griglia che divide la zona d'interesse in celle quadrate di lato arbitrario, dove ogni cella tiene conto del numero di vertici presenti all'interno del

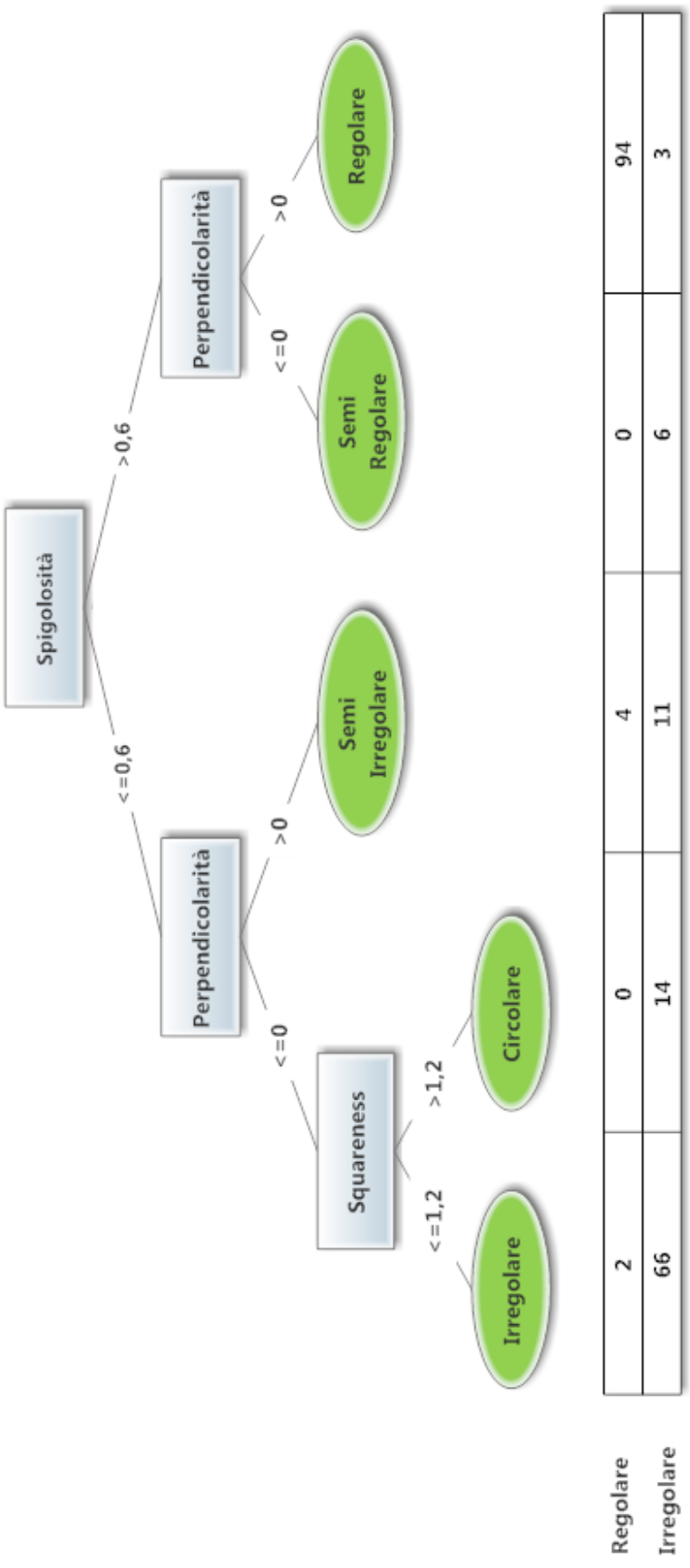


Figura 9: Albero di decisione per elementi areali

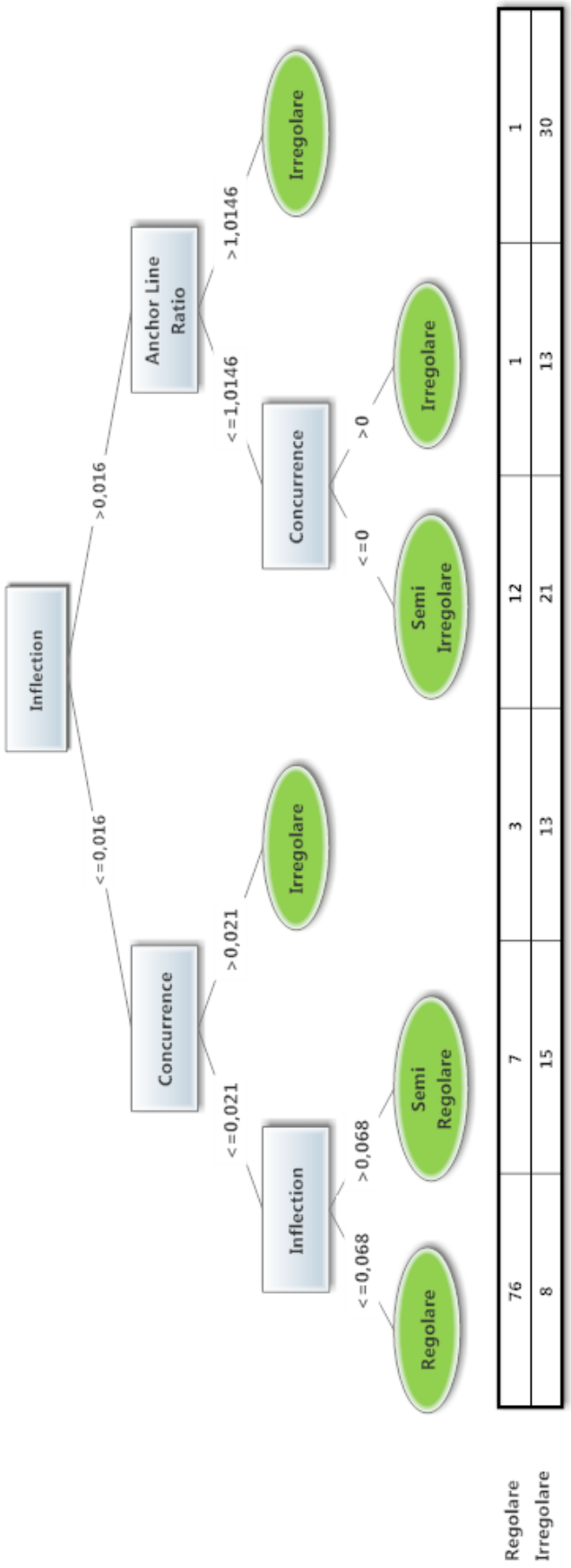


Figura 10: Albero di decisione per elementi lineari

proprio perimetro. In questo modo si ha un parametro qualitativo che rispecchia la densità della mappa. Per rendere omogenea la valutazione, l'algoritmo aggiunge vertici a linee ed aree perché queste possono contribuire in maniera significativa all'affollamento della mappa.

Per il nostro scopo, l'algoritmo può andar bene nel caso di elementi puntuali ed elementi lineari, ma la situazione è differente quando l'oggetto è di tipo areale in quanto l'algoritmo aggiunge vertici solo sui bordi. Per ovviare a questo problema, si possono pensare due soluzioni:

1. una prima soluzione è quella di trovare l'area della intersezione tra una cella e gli elementi areali che cadono all'interno della cella stessa. La singola cella tiene poi conto della quantità d'area occupata.

2. una seconda soluzione, più efficiente in termini di complessità temporale in quanto non deve essere calcolata l'intersezione, è invece quella di creare una "sottogriglia" di punti posti ad una ugual distanza tra loro e poi contare, per ogni cella, quanti di questi punti intersecano gli elementi dentro la cella. La cella infine tiene conto del numero totale di questi punti.

Ad ogni cella della matrice di densità viene poi assegnato un colore, colore che sarà più saturo per le celle con all'interno un maggior numero di vertici. In questo modo una rapida occhiata ci dà l'opportunità di vedere all'istante quali sono le zone più affollate della mappa e capire, in base ai cambiamenti di densità, la distribuzione dell'oggetto considerato nella mappa, oppure valutare la congruenza rispetto alle informazioni sull'uso del suolo (ad esempio una alta densità di edifici in una zona classificata come agricola è sintomo di anomalia).

Con qualche modifica, l'algoritmo può essere usato anche per altri scopi. Uno di questi è quello ad esempio di valutare la correttezza dei punti quotati, ossia che non siano presenti punti che hanno un valore di quota che si scosta in maniera significativa dai punti vicini.

Bisogna però procedere ad una serie di valutazioni prima di poter utilizzare correttamente l'algoritmo. Ad esempio, con riferimento alla Figura 11, se la griglia ha una maglia stretta, un'analisi della distribuzione degli edifici in una zona urbana come quella in figura restituisce, in assenza di ulteriori informazioni, un falso positivo in corrispondenza della piazza a destra in quanto le celle che la contengono avranno sicuramente un valore di densità minore rispetto alle altre celle. Oppure la valutazione sui punti quotati avrà un significato diverso se l'analisi viene fatta in una zona di pianura, dove è abbastanza facile catturare un punto anomalo, o in una zona di alta montagna, dove sono frequenti strapiombi o campanili.

Quindi l'algoritmo, in tutte le sue possibili varianti, deve comunque essere usato con cautela, valutando attentamente in partenza alcuni

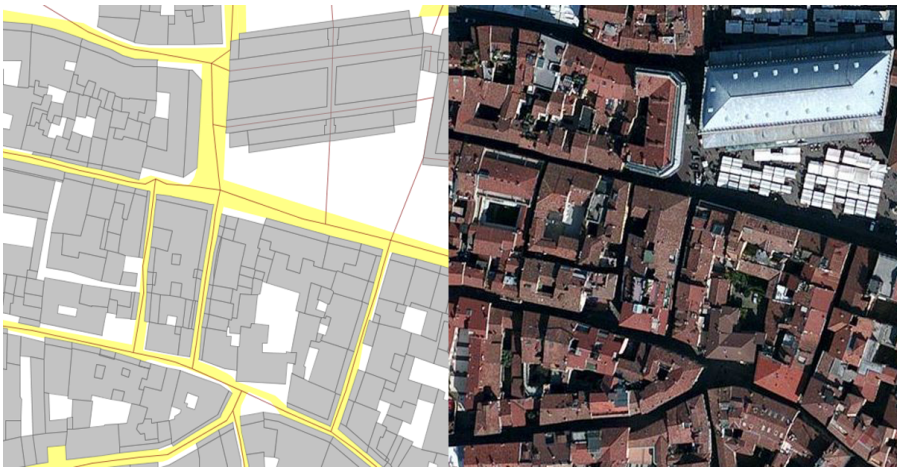

Figura 11: Se la griglia ha una maglia stretta, la presenza della piazza causa una anomalia di distribuzione

punti chiave, tra cui la dimensione e il contesto della zona da esaminare e/o le dimensioni delle singole celle. In generale un modello di confronto ci permette effettivamente di capire se siamo di fronte ad una situazione anomala oppure no.

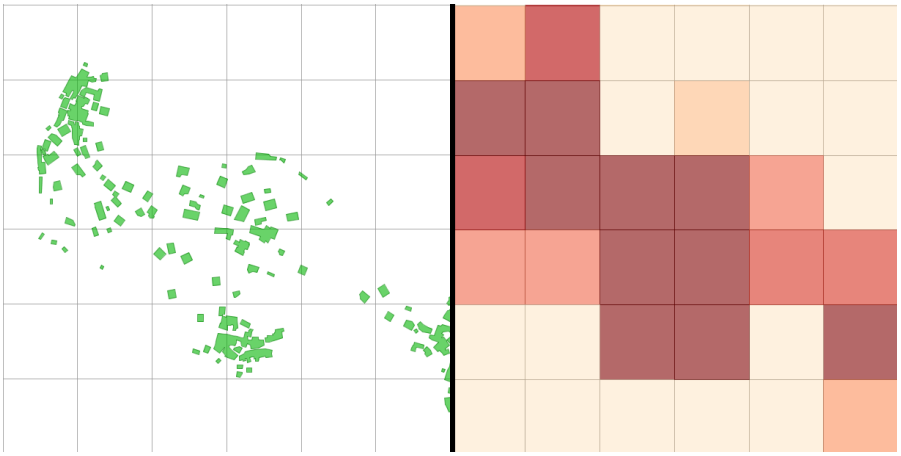

Figura 12: Un esempio di mappa e sua matrice di densità

## 3.4 POSIZIONE

### 3.4.1 *Posizionamento logico*

Nelle specifiche di realizzazione di un database topografico, vengono definiti dei vincoli topologici che devono essere assolutamente rispettati. Ad esempio nel GeoUML Catalogue [21] viene indicato che "la galleria con uso ciclabile deve contenere una corrispondente sede di area di circolazione ciclabile" oppure che "il territorio della specifica provincia è partizionato nel territorio dei comuni che la compongono,

tra loro disgiunti; viceversa ogni territorio comunale deve appartenere al territorio della provincia di cui è parte".

Ma le specifiche non definiscono vincoli di posizionamento logico di un oggetto, dove per anomalia di *posizionamento logico* intendiamo una incongruenza tra il contesto in cui si trova un oggetto e la classificazione dell'oggetto stesso. Un tipico esempio di questa anomalia può essere la presenza di un rifugio in una zona non di alta montagna o di un elemento portuale distante dal mare o da un corso d'acqua.

A partire da un insieme di esempi, sono state individuate alcune categorie di anomalie di posizionamento logico possibile e, per ognuna di queste, sono stati realizzati dei metodi all'interno dell'algoritmo:

- *Overlay*: situazione in cui un'entità geografica non può trovarsi in una particolare zona di una mappa. Un esempio può essere la presenza di una casa in mezzo ad un lago senza che esista una isola a cui appartiene. Il metodo implementato restituisce anomalia nel momento in cui è verificata una relazione spaziale definita da una maschera d'errore tra le due entità prese in considerazione;

- *Vicinanza spaziale*: situazione in cui un'entità geografica non si trova vicino ad un'altra. Ad esempio è possibile trovare una chiesa senza campanile, ma la situazione inversa è molto meno comune. Il metodo implementato non fa altro che valutare se l'entità è entro una certa distanza dall'altro oggetto di riferimento;

- *Touch*: situazione in cui un'entità geografica tocca almeno un'altra entità. Un esempio sono quelle situazioni in cui un ponte non è collegato ad una strada o ad una ferrovia oppure un passo alpino non è collegato ad una strada o ad un sentiero. Analizzando le situazioni tipiche di questa categoria, il metodo è stato pensato soprattutto per elementi lineari. Pensando al tipico esempio del ponte, il procedimento si occupa di verificare che entrambe le estremità siano collegate con due oggetti della stessa classe o tipologia, segnalando anomalia se ciò non accade;

- *Altitudine*: situazione in cui un elemento geografico si trova ad una quota anomala. Tipico esempio è la presenza di un rifugio in una zona non di montagna. Il metodo valuta una media dei valori $z$ delle coordinate spaziali dell'oggetto, il quale viene segnalato se non rispetta un valore di soglia definito con cura in input.

## 3.5 GRAFI

Questa tipologia di anomalie riguarda quelle situazioni che vanno a compromettere la correttezza di un grafo. In questa categoria troviamo anomalie che riguardano la connettività nei grafi e la coerenza tra i

tratti adiacenti. Per quanto riguarda questa categoria, l'implementazione di algoritmi si è focalizzata principalmente sulla ricerca dei tratti isolati e sulle possibili interruzioni.

### 3.5.1 *Interruzioni*

Per comprendere questo tipo di anomalia sui grafi, prendiamo in considerazione la viabilità stradale. Analizzando i tratti che ne compongono il grafo, possiamo imbatterci spesso in situazioni in cui un tratto è *dangling* (ovvero ad una sua estremità non è collegato alcun altro tratto) ma solo perché c'è una zona di vuoto tra due elementi lineari. Nella realtà due strade considerate troppo vicine potrebbero essere davvero così oppure invece una delle due linee effettivamente potrebbe arrivare "corta". Questo esempio è esplicativo per capire la differenza tra anomalia ed errore.

Definiamo una interruzione come la presenza di un vuoto di piccole dimensioni tra due elementi lineari che sembrano connessi. La presenza di interruzioni in un grafo è una situazione spiacevole in quanto porta al fallimento degli strumenti di ricerca sui grafi, come ad esempio gli algoritmi di navigazione.

Figura 13: Esempio d'interruzione in una strada

Per la ricerca delle possibili interruzioni in un grafo, l'idea dell'algoritmo è quella di restituire una coppia di punti che potrebbero nella realtà essere uniti. A partire dall'insieme degli archi della *feature* lineare, vengono presi in considerazione i punti iniziale e finale di una linea, e si verifica se tali punti toccano almeno un punto di un'altra linea vicina. Per evitare di prendere in considerazione un numero

eccessivo di punti, in fase di pre-processing viene fatta una fusione delle linee senza preoccuparci della classe della linea stessa, utilizzando i metodi della classe `LineMerger` di JTS. Se un punto non tocca nient'altro, allora bisogna vedere se effettivamente c'è un altro punto di un altro arco del grafo che è entro una certa distanza (ad esempio potrebbe capitare che due strade siano divise da un tratto ferroviario senza la presenza di un qualsiasi tipo di attraversamento). Se un tale punto viene trovato, non possiamo ancora concludere d'essere di fronte ad una anomalia di interruzione. Infatti potrebbe capitare che le due strade coinvolte si trovino ad altezze differenti e per cui una loro unione sarebbe impossibile. Perciò in ultima analisi si possono confrontare le coordinate $z$ dei due punti e, se la differenza d'altezza è sotto un certo limite fissato, allora possiamo segnalare la coppia di punti come anomalia.

### 3.5.2 *Tratti isolati*

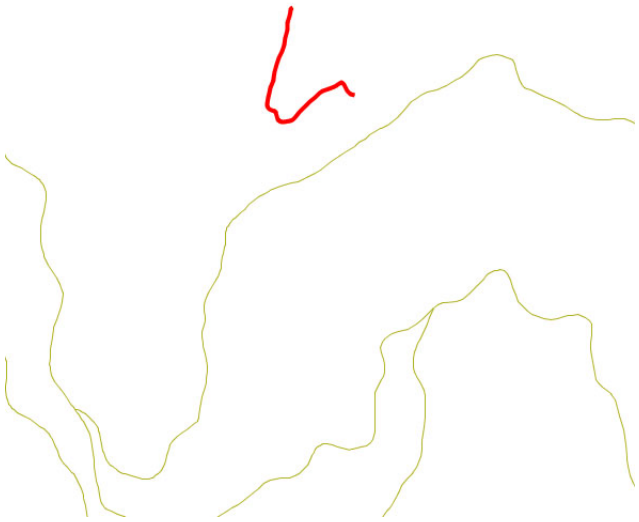

Figura 14: In rosso è segnalato il tratto di strada isolato dal resto del grafo

Un'altra delle situazioni che influenzano la connettività nei grafi riguarda la presenza di tratti di viabilità stradale o ferroviaria che sono isolati dal resto del grafo, situazione visibile in Figura 14. L'idea è quella di cercare l'insieme dei grafi presenti nell'area della mappa memorizzandone la cardinalità (ossia il numero di archi che lo compongono). In questo modo si può ragionevolmente pensare che il grafo a cardinalità più alta sia quello principale. Per non essere influenzato da errori di interruzione, viene chiamato il metodo `vertexSnapper` che si occupa di collegare un vertice *dangling* al più vicino vertice di un'altra linea entro una certa distanza uguale al valore di risoluzione della mappa.

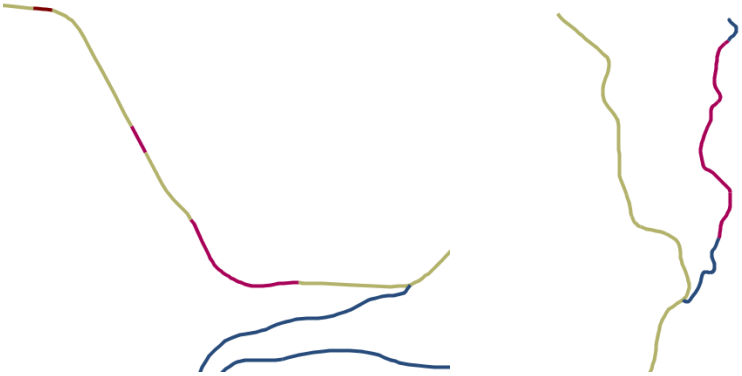

Figura 15: Incoerenza di classificazione di tratti stradali

Una volta identificato il grafo principale, i restanti grafi sono candidati ad essere segnalati. Si prendono perciò in considerazione i punti estremi del grafo e si valuta la distanza minima di tali punti dal grafo. Se però il nodo di un grafo candidato è nel bordo della mappa (*bounding box* o limite amministrativo), allora è probabile che sia collegato ad altri tratti lineari facenti parte di un'altra area, e per cui il grafo non viene segnalato. Altrimenti, se nessuno dei punti del grafo candidato è dentro il limite di distanza fissato, allora viene segnalata l'anomalia.

### 3.5.3 *Uniformità dei tratti del grafo*

Un altro caso di anomalie in cui si può imbattere è la presenza di valori semantici o geometrici in disaccordo tra tratti adiacenti del grafo. Un primo caso è dato dal cambio continuo di classe di un tratto del grafo: ad esempio la classificazione di una strada che da locale diventa per un tratto autostrada e che poi ritorna ad essere nuovamente locale. Queste situazioni tipicamente indicano un errore di classificazione. L'incoerenza di classificazione in una zona di grafo stradale è visibile in Figura 15. Un altro caso di anomalia sul grafo può essere dato dalla non monotonicità: se prendiamo ad esempio il grafo dei fiumi, ci si aspetta che un fiume non abbia dei tratti in salita ma che proceda sempre in discesa. Il caso in cui tratti consecutivi del grafo presentino valori di pendenza contrastanti rappresenta di fatto una anomalia e può indicare un'errata attribuzione delle quote sui vertici del tratto del grafo.

# 4

# ANALISI DEI RISULTATI

In questo capitolo vengono presentati i risultati più significativi dell'applicazione degli algoritmi realizzati sui dataset specificati nel paragrafo successivo. Per una migliore comprensione, i risultati sono corredati da una serie di figure.

## 4.1 CARTOGRAFIA USATA PER I TEST

La cartografia usata per effettuare i test di ricerca delle anomalie è il DBT in scala 1:5.000 del Comune di Padova aggiornato a Luglio 2011. La cartografia della città di Padova presenta numerosi elementi che la rendono appetibile per i test; infatti, come tutte le città, è una zona ricca di edifici, strade, binari ed altri elementi più o meno naturali come parchi, zone boschive, fiumi e, nelle zone di periferia, fossi. L'area del Comune di Padova consiste in un territorio di circa $92,85$ $km^2$, contenente ad esempio 85894 unità volumetriche, 10454 segmenti stradali, 338 segmenti ferroviari, 2441 fossi, 12 specchi d'acqua, 79 boschi. Nei dati in possesso però non ci sono informazioni di interesse per quanto riguarda la tipologia di edifici, in quanto siamo a disposizione solamente della classe UN_VOL (Unità volumetrica) e non della classe EDIFC che può fornire l'informazione. Questa situazione risulta essere un problema soprattutto nella ricerca di anomalie di posizionamento logico degli edifici ma risulta comunque utile nelle valutazioni sulla forma degli oggetti.

Per ovviare al problema, alcuni test vengono compiuti sul DBT sempre in scala 1:5000 di una porzione del Comune di Borgomaro (in particolare le frazioni Conio, Poggiato e Ville San Pietro), in provincia di Imperia (Liguria), una zona collinare che si sviluppa su una superficie di circa 23 $km^2$ e che contiene 646 edifici.

Inoltre, per valutare il posizionamento logico, prendiamo in esami i dati generalizzati in scala 1:25.000 nell'ottobre 2011 dal progetto CARGEN dell'intera area comunale che include anche le restanti frazioni. Quest'area presenta 908 edifici.

## 4.2 FORMA

### 4.2.1 *Contorni*

L'algoritmo si occupa di valutare il numero di lati che sono sotto un certo limite di lunghezza e il numero di vertici il cui angolo è sotto un certo valore di soglia. I test sono stati effettuati sulle classi degli elementi stradali e ferroviari alla ricerca di improvvisi cambi di direzione del loro tracciato, e sulla classe unità volumetrica alla ricerca di pareti troppo corte per essere rappresentate e di spigoli troppo acuti tra due pareti consecutive. Sono stati scelti come soglia per la lunghezza un valore pari a 0.5m, e come limite d'angolo un valore pari a 20°. In Figura 16 si può notare un tipico caso di divergenza della rete stradale; in questo caso è presente in prossimità di una curva dove l'area stradale aumenta la sua larghezza. Per quanto invece riguarda

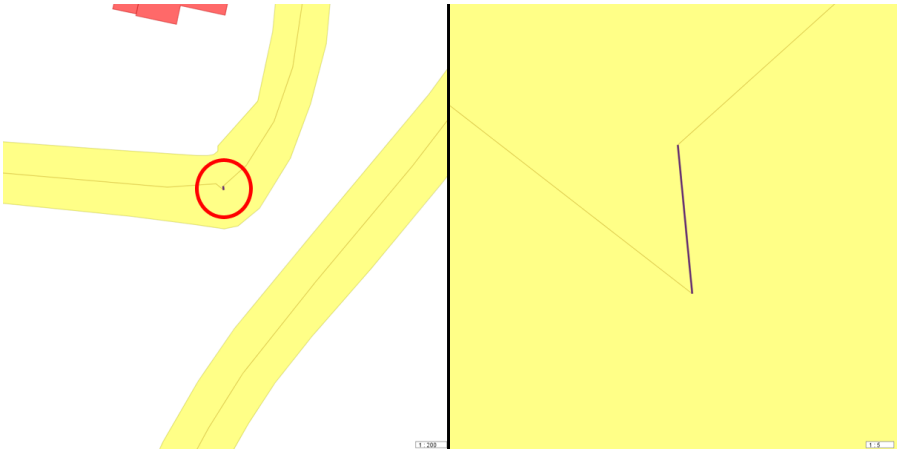

Figura 16: Divergenza della rete stradale

le unità volumetriche, la ricerca ha portato all'individuazione di una serie di casi differenti per cui si verifica tale anomalia:

- la non perfetta acquisizione dei contorni delle unità che hanno lati in comune; l'esempio in Figura 17 mostra come un elemento sembra "entrare" dentro l'altro;
- un livello di dettaglio maggiore di quanto la scala possa rappresentare; la Figura 18 mostra come i lati segnalati concorrono a formare uno spigolo dolce tra due lati, situazione che non è visibile nella scala 1:5.000 del DBT;
- la presenza di pareti effettivamente troppo corte; un esempio è mostrato in Figura 19 dove la parete segnalata può essere tolta in modo poi da formare un unico lato più lungo (può essere eliminato ad esempio utilizzando l'algoritmo di Sester);
- gli spigoli di una unità rappresentati da lati di piccola dimensione invece che da semplici vertici (Figura 20);

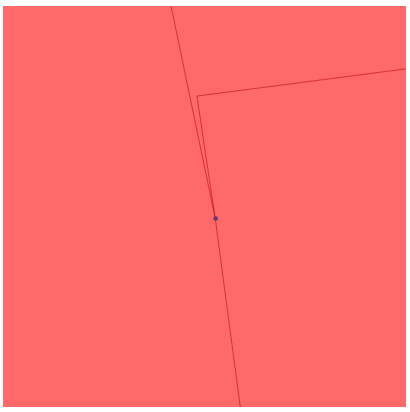

Figura 17: Divergenza nel bordo comune di due entità

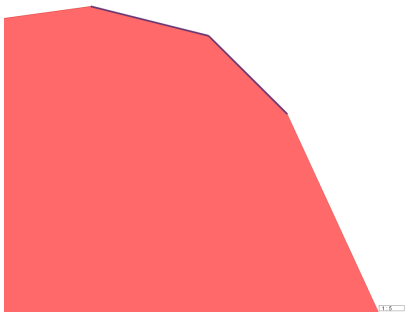

Figura 18: Eccessivo dettaglio nella rappresentazione dello spigolo nella scala 1:5.000

- la non perfetta acquisizione dei lati che causa distorsioni del contorno.

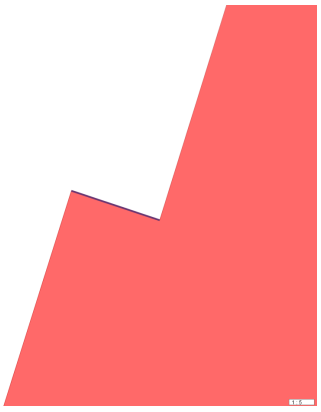

Figura 19: Parete di piccole dimensioni

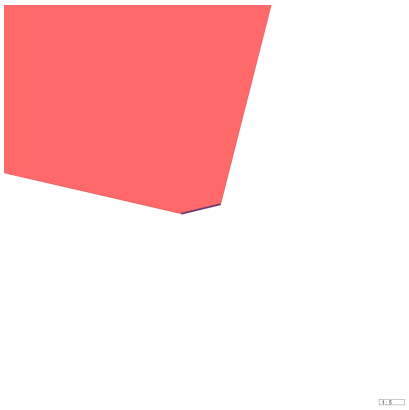

Figura 20: Spigolo rappresentato da un lato invece che da un singolo vertice

### 4.2.2 *Divergenza punto*

L'algoritmo ricerca possibili divergenze nei contorni di un poligono o lungo il percorso di una linea. È stato scelto come parametro per l'angolo un valore pari a 20° e il test è stato effettuato, per gli elementi lineari, per le classi elementi stradali, elementi ferroviari e scolina e, per i poligoni, per la classe unità volumetrica.

In Figura 21 sono raffigurate un paio di divergenze trovate nella classe degli elementi stradali. Si è notato che, per come è realizzato l'algoritmo, il risultato è fortemente influenzato dalla presenza di punti doppi (o addirittura tripli) che provocano errori nel calcolo del valore dell'angolo di un vertice, ciò che accade soprattutto nella classe

delle scoline.

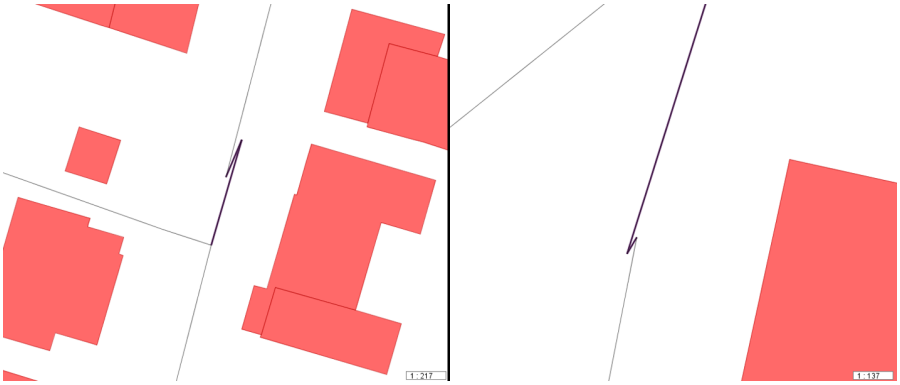

Figura 21: Divergenze nel grafo stradale

### 4.2.3 *Forme regolari ed irregolari*

L'algoritmo cerca di dividere gli oggetti di una classe in base alla regolarità e alla irregolarità del suo contorno, in modo tale da individuare possibili errori di classificazione dell'oggetto stesso. Per non associare una forma solamente a due categorie, l'algoritmo restituisce come anomalie anche quelle situazioni intermedie in cui le misure non definiscono un oggetto esattamente come regolare o irregolare, ma i cui valori delle misure si avvicinano ad uno dei due casi; d'ora in poi parleremo perciò di elementi semi-regolari e semi-irregolari.

Per i poligoni è stato fatto un test di regolarità sulla classe unità volumetrica e un test di irregolarità sulle classi specchio d'acqua e bosco. Nelle figure 22 e 23 in blu sono segnate le unità che sono state

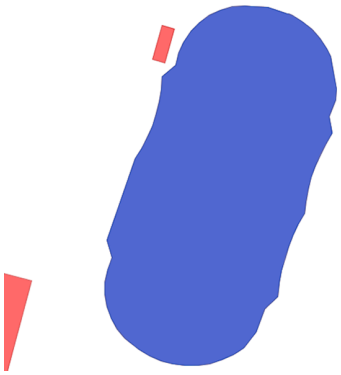

Figura 22: Irregolarità di un edificio

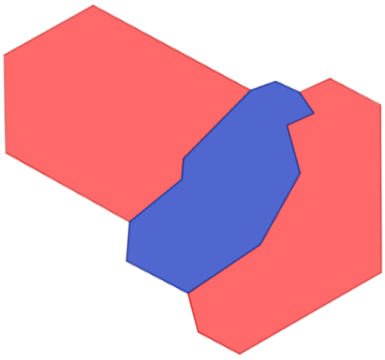

Figura 23: Irregolarità di un edificio

identificate come irregolari; in effetti le due forme possono portare a qualche dubbio sia sulla loro corretta classificazione che eventualmente ad errori di digitalizzazione del contorno.

Invece, nelle figure 24 e 25 possiamo notare due casi in cui l'irregolarità è palesemente non vera; nel primo caso la valutazione è

influenzata dal valore di spigolosità appena inferiore al valore di split, mentre per il secondo caso sia il valore di perpendicolarità che di spigolosità è pari a 0, dovuto proprio al fatto che tutti gli spigoli sono arrotondati. I risultati ottenuti utilizzando le misure di spigolosità, perpendicolarità e squareness sono comunque soddisfacenti in quanto il 97% delle unità viene giustamente classificato come regolare mentre solamente lo 0,22% è classificato come irregolare, principalmente a causa dei due casi descritti in precedenza. Per quanto riguarda la

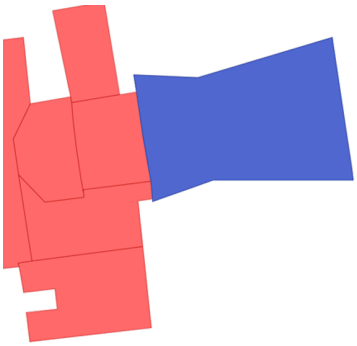

Figura 24: Entità segnalata come irregolare a causa del valore di spigolosità

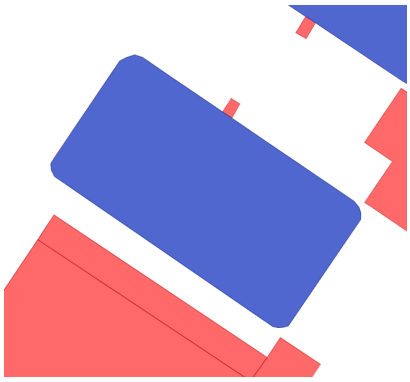

Figura 25: Entità segnalata come irregolare a causa degli spigoli troppo arrotondati

classe Bosco la situazione è differente. Quando si prende in considerazione un oggetto di questo tipo, ci si aspetta di trovare una forma abbastanza irregolare. Si trova che, su un totale di 79 elementi, ben 12 elementi sono classificati come regolari: alcuni di questi sono boschi artificiali mentre altri sono inseriti all'interno di proprietà private e/o pubbliche e sono perciò delimitati da confini ben precisi; sono comunque delle situazioni che ci possiamo aspettare di trovare in zone densamente popolate. Tenendo però presente che lo scopo dell'algoritmo è quello di cercare errori nella classificazione degli oggetti, sono stati trovati che alcuni di questi boschi addirittura sembrano non esistere (Figura 26) e che circa 30 elementi non rispettano il solo vincolo di estensione minima (2000 $m^2$) definito nelle specifiche per la costruzione del DBT. Per quanto riguarda gli specchi d'acqua, l'analisi di irregolarità è stata eseguita per il valore '02' (stagno) dell'attributo SP_ACQ_TY; in questo caso solamente due elementi sono stati classificati come semi-irregolari: come infatti si nota nella Figura 27 le loro forme sono influenzate da alcuni tratti più "squadrati" e ciò ci porta a pensare a stagni artificiali (ed effettivamente lo sono).

Per quanto riguarda infine gli elementi lineari, l'analisi è stata eseguita sui binari ferroviari presenti all'interno della mappa. In questo caso le misure indicano che solo 10 elementi sono irregolari. Ma, come si può vedere dalla Figura 28, l'elemento non è poi così irregolare come ci si può aspettare; questo è dovuto al fatto che c'è

un confine molto sottile tra la regolarità e l'irregolarità di una linea, a causa soprattutto degli elementi scelti nel test set. Se infatti la scelta di prendere i binari ferroviari come rappresentati degli elementi regolari è soddisfacente per la loro tipica linearità, non altrettanto felice è la scelta dei corsi d'acqua (sebbene siano presi in zone collinari) per gli elementi irregolari, in quanto molti tratti, anche se brevi, sono lineari e perciò causano disturbi nella classificazione. Risulta però difficile trovare elementi lineari più tortuosi dei corsi d'acqua.

## 4.3 POSIZIONE

### 4.3.1 *Posizionamento logico*

Per quanto riguarda il posizionamento logico degli oggetti nella mappa, sono stati effettuati alcuni test significativi su varie classi presenti nei dataset disponibili; vediamo in dettaglio i vari casi considerati e i risultati ottenuti:

- *Giunzione ferroviaria su elemento ferroviario*: dalla definizione in [8], la giunzione ferroviaria è un "punto di inizio/fine o di confluenza/diramazione di elementi ferroviari o di intersezione con altri grafi topologici della viabilità". Si è deciso perciò di fare un test di intersezione tra questi elementi e i binari presenti nella classe "elemento ferroviario". Si è ottenuto che 8 elementi su 358 si scostano dal grafo; tali punti assumono un valore pari a '05' nel campo TIPO, valore che però nelle specifiche del DBT non è presente. Visivamente gli scostamenti si verificano nel cambio d'attributo degli elementi ferroviari (in Figura 29 l'anomalia indicata dal punto blu si ha proprio nel passaggio da binario a raso a binario su ponte). Oltre perciò a questo errore di scostamento, si può evidenziare anche un errore di classificazione, in quanto, secondo le specifiche [8], l'attributo TIPO permette di classificare questi punti con il valore '11' (cambio attributo elemento ferroviario).

- *Ponti/viadotti/cavalcavia su corsi d'acqua/strade/binari*: il test viene fatto per valutare la corretta classificazione degli elementi stradali. In questo caso l'idea è quella di valutare la correttezza della classificazione dei tratti di grafo della viabilità stradale classificati come "ponti/viadotti/cavalcavia" (valore '02' dell'attributo SEDE). Perciò, come nel caso precedente, viene fatto un test d'intersezione tra questo elemento e alcuni dei possibili elementi sottostanti, ovvero binari, corsi d'acqua ed altre strade. A differenza del caso precedente, i risultati non hanno fornito alcuna anomalia.

- *Campanili in prossimità di chiese*: in questo caso si è deciso di fare un test sulla bontà dei dati generalizzati della Liguria. Nella

mentalità comune, l'idea è quella che in vicinanza di un campanile ci sia sempre una chiesa (il viceversa non è sempre vero). Secondo le specifiche dell'IGM [13], la feature "Campanile" è presente nella classe PAL220 con codice pari a P422A, mentre la feature "Chiesa cristiana" ha codice C431 ed è presente nella classe AAL015. Prendendo un valore di distanza massima tra il campanile ed una chiesa pari a 30m, l'analisi ha restituito il risultato di Figura 30 , dove il punto blu rappresenta il campanile "isolato". L'errore nasce dal fatto che, durante la fase di generalizzazione, tutti gli edifici adiacenti alla chiesa (segnalato in verde oliva l'oggetto originale in scala 1.5.000) sono stati uniti in un unica entità, perdendo così di fatto l'informazione sulla presenza del luogo di culto.

- *Ponti collegati alle strade*: questo è il tipico caso di verifica di situazioni in cui un'entità geografica tocca almeno un'altra entità. Nel caso esaminato, il procedimento si occupa di verificare che entrambe le estremità dell'oggetto "ponti/viadotti/cavalcavia" siano collegate con il resto del grafo stradale. Il risultato di questo test è visibile in Figura 31, dove è possibile vedere chiaramente la differenza tra la realtà e la sua rappresentazione; infatti nella mappa digitale a sinistra entrambe le estremità del viadotto non sono collegate al grafo stradale, mentre nel mondo reale a destra il viadotto è perfettamente collegato. In questo caso perciò siamo di fronte ad una situazione di incompletezza dei dati per la mancanza appunto di tratti stradali.

## 4.4 GRAFI

### 4.4.1 *Interruzioni*

L'algoritmo si occupa di ricercare eventuali interruzioni in un grafo. La ricerca viene fatta per la rete stradale e per quella ferroviaria, dove l'interruzione massima possibile è stata fissata a 10m. Durante la ricerca sono stati individuati i seguenti casi:

1. zona di vuoto di dimensioni minore dell'accuratezza planimetrica tra due punti di due segmenti che nella realtà appartengono allo stesso arco. In questo caso è evidente l'errore che può essere corretto senza ulteriori controlli (vedi Figura 32);

2. zona di vuoto tra due linee che potrebbero essere collegate. Come si può vedere in Figura 33, le parti terminali delle due strade sono sotto il valore di soglia. Senza un controllo diretto sul campo o senza ulteriori informazioni nella mappa (come ad esempio una recinzione che delimita la proprietà dell'edificio) non è possibile decidere se siamo di fronte ad un errore di collegamento.

Lo stesso accade per la rete ferroviaria (Figura 34), dove manca il collegamento tra le due parti del grafo per la presenza di un tratto stradale che provoca l'interruzione; in questo caso però siamo di fronte ad un errore, sia per le tipiche caratteristiche della rete ferroviaria, sia perché quel tratto stradale è in realtà un cavalcavia;

3. situazione di *undershoot*, cioè una situazione in un cui una linea si trova ad una distanza inferiore ad una specifica tolleranza da un'altra linea. Il caso di Figura 35 mostra come la linea arriva "corta" e non va a collegarsi all'incrocio, al contrario di quanto accade per le aree stradali corrispondenti segnate in giallo;

4. binari che terminano alla stessa "altezza" (Figura 36) la cui distanza tra i punti finali sono entrambi sotto il valore di soglia. Questa è una situazione tipica nell'ambito ferroviario, soprattutto all'interno di stazioni e scali merce e che quindi non deve essere segnalato come errore.

In conclusione, per capire se effettivamente siamo di fronte ad una situazione d'errore oppure no, è evidente che è importante conoscere sia la copertura della mappa che la tipologia del grafo su cui si fanno le valutazioni, come si evince ad esempio dal caso (2).

### 4.4.2 *Tratti isolati*

L'algoritmo si occupa di ricercare tratti non connessi di un grafo e quindi isolati. Si specifica in input un parametro di distanza abbastanza ragionevole che, in questo caso è stato scelto essere pari a 5m. Come per l'algoritmo precedente, sono stati presi in considerazione i grafi stradale e ferroviario, mentre come bordo mappa è stato scelto il bordo comunale. Nelle figure successive in blu sono identificati i tratti isolati. La ricerca ha evidenziato il caso tipico: nella Figura 37 il cerchio rosso mostra che l'anomalia è causata dalla mancanza di uno scambio ferroviario di collegamento al resto dei binari, mentre in Figura 38 la strada è scollegata dal resto del grafo su tutti e due gli estremi.

Un altro tipo di errore frequente è la presenza di tratti isolati in prossimità dei bordi della mappa: un esempio è visibile in Figura 39, dove la via è segnalata come anomalia poiché il tratto di strada principale a cui la via fa riferimento si trova appena fuori dal confine del comune identificato dalla linea rossa e la via non è stata prolungata fino al confine, come invece succede in altri casi (vedi Figura 40).

Figura 26: A sinistra in verde è rappresentato l'oggetto bosco, che nel mondo reale sembra non esserci

Figura 27: Laghi segnalati come semi-irregolari

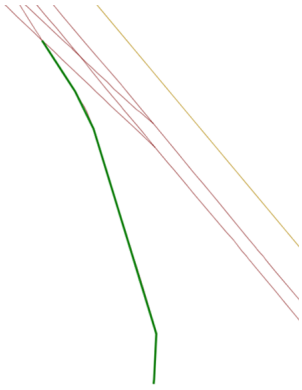

Figura 28: Binario segnalato come irregolare

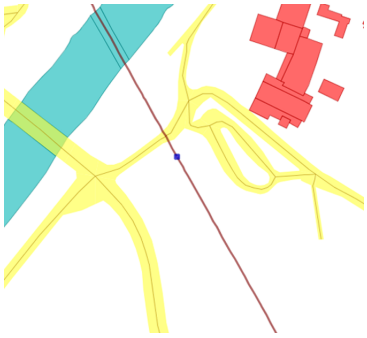

Figura 29: In blu è segnalata la giunzione ferroviaria che si distacca dal grafo ferroviario

Figura 30: In blu è segnalato il campanile non in prossimità di una chiesa

Figura 31: A sinistra in blu scuro è segnalato il ponte non collegato al resto del grafo stradale, situazione che non si presenta nel mondo reale com'è possibile vedere a sinistra

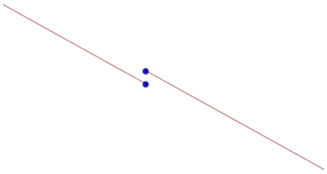

Figura 32: Zona di vuota di dimensioni minori dell'accuratezza planimetrica

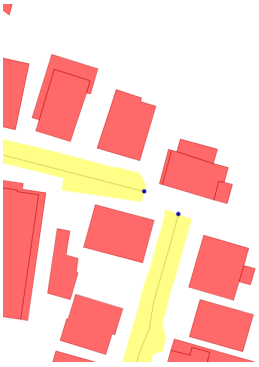

Figura 33: Interruzione possibile tra due strade

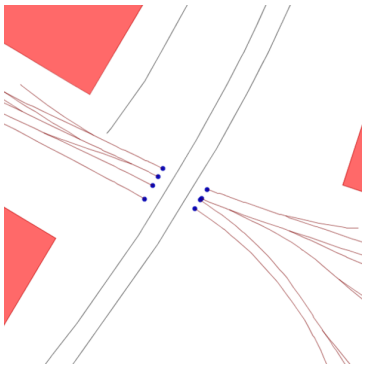

Figura 34: Interruzione dovuta alla mancanza di tratti ferroviari in corrispondenza di un sottopasso

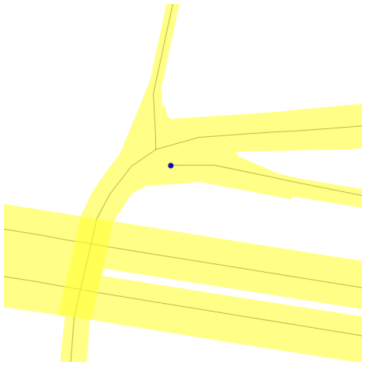

Figura 35: *Undershoot* nei pressi di un incrocio

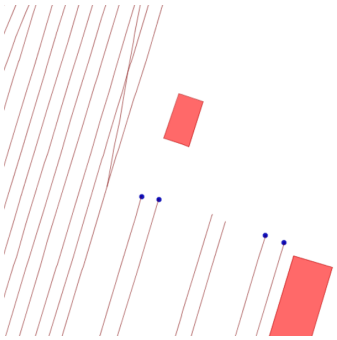

Figura 36: Due binari che terminano alla stessa altezza la cui distanza tra i punti terminali è sotto la soglia

Figura 37: Tratto di binari isolati

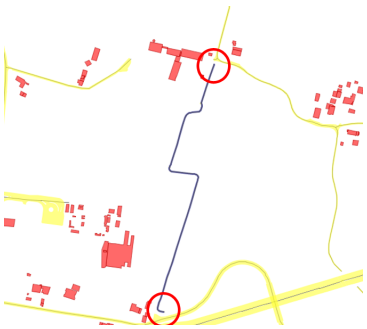

Figura 38: Tratto di strade isolato

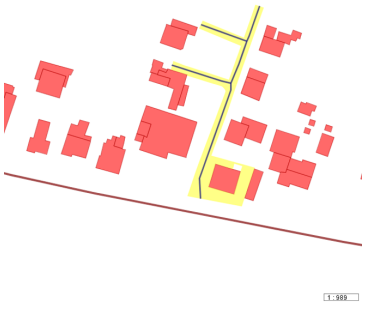

Figura 39: Tratti isolati a causa della mancanza della strada principale nel confine

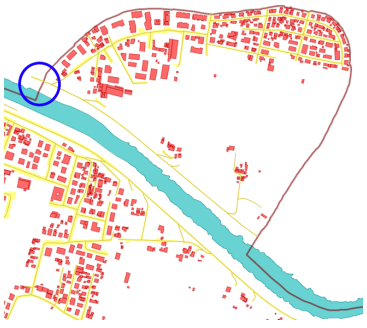

Figura 40: Tratto adiacente al confine non isolato

# 5

# CONCLUSIONI E POSSIBILI SVILUPPI

Questo lavoro di tesi si poneva l'obiettivo della ricerca di metodi e strumenti per individuare automaticamente errori non rilevabili con i normali controlli formali di correttezza, quali errori di classificazione o incompletezza dei dati. Gli algoritmi che sono stati sviluppati – che hanno seguito la tassonomia proposta in un precedente lavoro di tesi – insieme agli strumenti formali di controllo della topologia e della geometria, possono rappresentare un solido ed originale punto di partenza per valutare la correttezza dei dati, sia prima sia dopo la fase di generalizzazione.

Mentre alcuni di questi algoritmi, con un accurato *tuning* dei parametri e un'ottima conoscenza delle specifiche di un DB topografico, permettono di trovare immediatamente errori all'interno di una database spaziale, è importante notare come altri (o gli stessi) strumenti non sono effettivamente in grado di trovare un errore senza ulteriori informazioni aggiuntive. Un caso significativo a supporto di questa considerazione è la ricerca delle interruzioni in un grafo. Quando l'interruzione è inferiore alla risoluzione possiamo affermare quasi certamente di essere di fronte ad un errore; quando invece è maggiore abbiamo bisogno di informazioni maggiori sulla copertura del territorio, come ad esempio la presenza di un elemento divisorio, di un edificio, ecc.... Un altro caso è la distribuzione di oggetti dello stesso tipo in una mappa, in cui il risultato è influenzato dalla copertura e dalla gestione del territorio.

L'approccio scelto è quello che ogni algoritmo che gestisce un singolo caso restituisce una tabella delle anomalie riscontrate. Un miglioramento a questo approccio potrebbe essere quello di unire i risultati di differenti valutazioni su uno stesso oggetto per avere un risultato più affidabile: come visto precedentemente, un oggetto classificato come bosco può essere un'anomalia sia se la forma è troppo regolare, sia se la sua area è sotto un certo valore; un unione ragionata dei due risultati potrebbe scremare ulteriormente il numero di situazioni anomale.

Infine, grazie all'efficienza delle librerie sviluppate in JTS e all'interno del progetto CARGEN, gli algoritmi creati in generale riescono ad operare su una gran quantità di dati con buone prestazioni. Alcuni

numeri[1]: considerando la classe "Unità Volumetrica" siamo andati alla ricerca del numero di lati e angoli sotto un certo valore di soglia e analizzate le forme degli oggetti che la popolano. La prima esecuzione ha richiesto circa 2,5s mentre la seconda 1,5s, su un totale di 732.422 segmenti suddivisi tra gli 85.894 oggetti.

Resta il fatto che, vista la vastità dell'argomento, delle situazioni nel mondo reale e delle problematiche, esistono possibilità di una estensione della tassonomia proposta e delle implementazioni e una ottimizzazione dell'approccio e dei singoli algoritmi sviluppati per raccogliere più casistiche possibili.

1 Test eseguiti su notebook dotato di processore Pentium Dual Core a 2,10 GHz, 4 GB di RAM, Sistema Operativo Windows 7 64bit SP1, Eclipse Indigo, JTS 1.12, Java 1.7, OpenJUMP 1.4.2

# BIBLIOGRAFIA


[1] AGENT (2000), "D C1 - Selection of Basic Measures"

[2] AGENT (2000), "D C2 - Specification of Internal Measures"

[3] AGENT (2000), "D D1 - Specification of Basic Algorithms"

[4] Buttenfield B. (1991), "A rule for describing line feature geometry"

[5] De Gennaro M., Rumor M., Savino S. (2009), "Le procedure per la derivazione del DB25 dal DBT della Regione del Veneto: risultati del progetto CARGEN", In: *Bollettino della Associazione Italiana di Cartografia, 135*

[6] Devillers R., Jeansoulin R. (2006), "Fundamentals of Spatial Data Quality", ISTE Ltd, Londra

[7] Frank R., Ester M. (2006), "A Quantitative Similarity Measure for Maps"

[8] IntesaGIS (2006), "Il catalogo degli oggetti: revisione delle specifiche di contenuto 1n1007_1 e 1n1007_2"

[9] IntesaGIS (2007), "Specifiche per la realizzazione dei database topografici di interesse generale: Linee guida per l'implementazione"

[10] Istituto Geografico Militare (2007), "Criteri di generalizzazione per la derivazione del DB25 e relativa cartografia alla scala 1:25.000 da CTRN e/o db cartografici"

[11] Istituto Geografico Militare (2009), "Relazioni topologiche del DB25: vincoli e controlli geometrici"

[12] Istituto Geografico Militare (2006), "La selezione degli oggetto topografici"

[13] Istituto Geografico Militare (2006), "Struttura del DB25: Feature, attributi e domini"

[14] Lazzaro S. (2011), "Un approccio al posizionamento di etichette su cartografia generata da database cartografici"

[15] Mackaness W.A, Ruas A. (2007), "Evaluation of the Map Generalization Process", In: *Generalisation of geographic information: Cartographic modelling and applications*, William A. Mackaness, Anne Ruas, L. Tiina Sarjakoski (Editors), Elsevier Science; 1st edition (June 6, 2007), 89-111

[16] Mara S., Mara H., Aktug B. (2010), "Topological error correction of GIS vector data", In: *International Journal of Physical Sciences vol.5*, 467-483

[17] Regione Veneto (2008), "DB Topografico", Segreteria Regionale all'Ambiente e Territorio, Unità di Progetto per il SIT e la cartografia. Online: `http://www.regione.veneto.it/NR/rdonlyres/50B13921-AD61-41A0-98CD-9429940E7B38/0/DBTopografico26.pdf`

[18] Peter B. (2001), "Measures for the Generalization of Polygonal Maps with Categorical Data"

[19] Savino S. (2011), "A solution to the problem of the generalization of the Italian geographical databases from large to medium scale: approach definition , process design and operators"

[20] Sole A., Scuccimarra V. (2002), "La Terra vista dai GIS"

[21] SpatialDBgroup (2011), "GeoUML Methodology e Tools: Organizzazione Complessiva"

[22] Stocco A. (2011), "Controllo per la qualità dei database topografici a grande scala"

[23] Stoter J., Xiang Z. (2008), "The Evaluation of Spatial Distribution Density in Map Generalization"

[24] Stoter J. et al (2009), "Methodology for Evaluating Automated Map Generalization in Commercial Software"

[25] Tan P.N., Steinbach M., Kumar V. (2006), "Introduction to Data Mining." Addison Wesley

[26] Ubeda T., Egenhofer M.J. (1997), "Topological Error Correcting in GIS", In: *Advances in Spatial Databases - Fifth International Symphosium on Large Spatial Database*, M. Scholl, A. Voisard (Editors), Vol. 1262, Springer-Verlag, 283-297

[27] Valerio L. (2008), "Applicazioni Open Source per Database Spaziali"