Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Analisi di Mobilità Pedonale Mediante Dati di Telefonia Georeferenziati

Relatore:
Prof. Armando Bazzani

Correlatore:
Dott. Alessandro Fabbri

Presentata da:
Andrea Checcoli

Anno Accademico 2016/2017

Abstract

Al fine di organizzare al meglio le città del futuro occorrono nuovi strumenti in grado di analizzare e comprendere il comportamento delle persone nelle aree urbane.
In questo elaborato viene illustrata la costruzione di un modello teorico relativo alla mobilità pedonale nella città di Venezia a partire dall'analisi di dati di telefonia mobile, rilevati nella giornata del 26 Febbraio 2017.
Vengono in seguito mostrate le differenti fasi necessarie alla realizzazione del modello a partire dall'elaborazione preliminare dei data set disponibili e focalizzando poi l'attenzione sugli algoritmi di georeferenziazione disponibili in letteratura.
Una volta ultimata l'analisi dei dati ed illustrate le prime osservazioni, vengono in seguito enunciati i concetti teorici che stanno alla base del modello di mobilità ponendo l'accento sulla natura Markoviana del processo stocastico osservato.
Infine si volge lo sguardo al risultato ottenuto, mostrando le verifiche a cui viene sottoposto il modello realizzato e le criticità emerse nell'affrontare questo studio.

Indice

Introduzione

Lo studio di molti problemi importanti e alla frontiera tra diverse discipline, dalla Fisica alla Biologia, dalla Economia alla Sociologia, ha visto emergere negli ultimi decenni nuove metodologie di analisi. Una di queste è una fra le più recenti branche della Fisica, la Fisica dei Sistemi Complessi.

Un *Sistema Complesso* è un sistema dinamico dotato di numerosi gradi di libertà; le componenti che lo costituiscono sono diverse sia per natura che per dinamica. I meccanismi di interazione sono di tipo collettivo e non sono inotre riconducibili ad interazioni fondamentali.
In relazione a ciò le interazioni che possono sussistere fra diverse componenti assumono carattere di scambio di *informazione* e la dinamica macroscopica del sistema ha proprietà di poter influenzare la dinamica microscopica del sistema il quale è caratterizzato talvolta dalla presenza di meccanismi evolutivi.
La modellizzazione di un sistema complesso ha come fine unltimo quello di evidenziare le proprietà emergenti del sistema e si focalizza dunque sulle proprietà del "Network" di interazione e non sulle proprietà delle singole particelle che costituiscono il sistema.

Questi sistemi sono studiati tramite diversi strumenti teorici, come la Teoria dei Network e la Teoria dei Grafi, e attraverso l'analisi di moli di dati, ottenuti da sorgenti diverse e pertanto non omogenei fra loro in merito alla tipologia.

Un esempio significativo di Sistema Complesso è rappresentato dalle città.[7]
Attualemente il 54 % della popolazione mondiale vive all'interno di agglomerati urbani. Le proiezioni delle organizzazione internazionali transgovernative[1] mostrano un trend crescente, dando al 66 % la percentuale della popolazione nel mondo che abiterà all'interno di aree urbane entro il 2050. Inoltre, ulteriori proiezioni segnalano un significativo aumento della popolazione all'interno della stessa finestra temporale, pari a circa 2.5

[1] dati forniti dalle Nazioni Unite: www.un.org

miliardi di persone.
Appare dunque chiara l'esigenza di elaborare strumenti di analisi con cui studiare e comprendere il comportamento degli utenti delle aree urbane, per meglio organizzare le città e le loro infrastrutture. I dispositivi di indagine tradizionali (*e.g.* imbuti ditraffico, contapersone etc.) risultano oggi insufficienti nel descrivere il comportamento degli esseri umani all'interno di una città, in quanto non rapidi nelle rilevazioni e dispendiosi in termini economici; emergono, invece, nuove tipologie di dati più adatti a questo scopo e già ottenibili tramite le infrastrutture digitali oggi presenti.
In questo elaborato si intende sperimentare questa metodologia di analisi per studiare un modello di mobilità pedonale all'interno di un contesto urbano molto singolare, ovvero il centro della città di Venezia, durante il corso di una giornata nel periodo di Carnevale. Più precisamente, si ci è interessati alla dinamica di transizione pedonale tra le differenti parti, storicamente dette *sestieri*, del centro città. Il campione analizzato è quello degli utenti di telefonia mobile *Tim* in un certo lasso di tempo. I dati sullo spostamento dei pedoni sono stati forniti da *Tim*, in completo anonimato e nel pieno rispetto della privacy.
Il fine di questo elaborato finale è quello di dimostrare sulla base dei dati ottenuti la validità di un modello che rappresenti la mobilità pedonale all'interno dellla città di Venezia attraverso l'utilizzo di dati di telefonia georeferenziati.
La nostra analisi mostra che il modello approssima il fenomeno osservato entro un certo intervallo di confidenza. Vi sono infatti delle discrepanze legate presumibilmente all' esistenza di effetti che allo stato dell'arte non sono stati integrati all'interno del modello e sui quali si possono solo avanzare ipotesi.

La struttura della tesi è la seguente. Nel Capitolo 1 è descritta la fase di data set e parsing preliminare all'analisi dati. In questa i dati sono stati estratti e catalogati, modellizzando i sestieri di Venezia tramite dei poligoni.

Lo studio della mobilità si è basato, tra le altre cose, sulla scelta di un algoritmo efficace che permetta di capire se un certo pedone si trovi a un dato momento all'interno di un dato sestiere (algoritmo di tipo *points in polygons*). Questa parte è l'oggetto del Capitolo 2.

Nel Capitolo 3 è esposta l'analisi dei dati. In particolare, si spiega come la scelta del sestiere si sia preferita ad altre scelte di siti, come si sia affrontato il problema degli utenti *fantasma* (utenti che vengono rilevati, ma non geolocalizzati) e quale sia la dinamica della popolazione osservata durante il corso della giornata.

Il Capitolo 4 si compone di due sezioni: la prima comprende alcuni richiami di teoria sui sistemi stocastici discreti, le catene di Markov e le variabili random con cui si è costruito il modello teorico che descrive i dati analizzati in 3. Il capitolo prosegue fornendo nella seconda sezione una panoramica sul procedimento utilizzato al fine di costruire il modello, con riferimento ai teoremi e alle definizioni usate nella sezione precedente. Infine vengono esposti i risultati ottenuti, le verifiche realizzate per validarli ed eventuali ipotesi su come interpretare alcune criticità emerse nell'analisi dei dati e nella loro modellizzazione.

Infine nell'ultimo capitolo si conclude questa trattazione mettendo in luce gli elementi chiave che hanno portato alla costruzione del modello; a partire dalla struttura dei dati fino alla loro interpretazione probabilistica, passando attraverso le scelte degli algoritmi e degli oggetti realizzati ed utilizzati nell'analisi dei dati a disposizione.

Capitolo 1

Data set e parsing

In questo capitolo descriveremo la fase preliminare all'analisi dati necessaria alla formulazione del modello studiato in questo progetto di tesi.

Lo studio presentato in questo elaborato si concentra sulla creazione di un modello di mobilità pedonale nel centro cittadino di Venezia basato sull'analisi di dati di telefonia georeferenziati.

1.1 Data Set

Si tratta di uno studio *spin-off* di un progetto affrontato dal gruppo di ricerca di Fisica dei Sistemi Complessi in collaborazione con una fra le maggiori compagnie di telefonia mobile italiane (*Tim*) e con il Comune di Venezia; tale collaborazione ha permesso di usufruire di un bacino di informazioni altrimenti non reperibile.

I dati forniti dal gestore di telefonia mobile *"Tim"* sono degli elenchi in formato *.csv* che rappresentano le attività geolocalizzate ad un certo istante, di un certo utente.

Più precisamente, l'utente è identificato da un ID univoco (*GEID*). Ciascuna tipologia di attività svolta è invece caratterizzata da un ulteriore ID (*CallID*). Il gestore fornisce anche informazioni in merito alla nazionalità della SIM che utilizza le sue piattaforme per le attività telefoniche. Tale tipologia di dato aggregato prende nome di *"Call Detail Record"* (*CDR*).

Ai fini del modello che si vuole elaborare sono necessarie solo:

- le informazioni relative all'ID dell'utente,

- le informazioni relative alle coordinate geografiche che lo referenziano al territorio;
- le informazioni relative all'ora in cui il dato è stato catalogato.

Il campione di dati utilizzati è stato rilevato durante la giornata di domenica 26 Febbraio 2017, giornata di particolare interesse per quanto riguarda la possibilità di rilevare attività telefoniche in quanto al centro del periodo del Carnevale.

Il processo di estrazione dati da un file prende nome di *parsing* ed il programma che esegue tale compito prende nome di *parser*. Una volta estratti i dati aggregati attraverso il parser, questi sono stati catalogati all'interno di strutture, discriminandoli in base all'orario. In particolare, all'interno di questo progetto di tesi sono state utilizzate librerie messe a punto dal gruppo di ricerca di "Fisica dei Sistemi Complessi" del Dipartimento di Fisica dell'Università di Bologna[1].
Il campione di dati fornito da *Tim* risulta però essere affetto da alcune criticità che ne compromettono lo sfruttamento al fine di realizzare uno studio su intervalli temporali ristretti; ciò lo si può evincere dal fatto che: a fronte di più di 9709852 milioni di righe presenti in un file csv, ai fini della nostra indagine e dunque per le nostre esigenze di georeferenziazione del dato, ne vengono esserne utilizzate olamente 1849782 righe. Complessivamente i dati estratti dal fil utilizzato sono stati pesantemente penalizzati da alcune caratteristiche non ottimizzabili.

STATS		
Total lines	9709852	
Errors	149680	1.54%
Not Georef	5436680	56%
Out of bbox	2272640	23.4%
No GeID	1065	0.01%
Valid	1849782	19.05%
GeID georef	5832	
CallID georef	28502	

Tabella 1.1: In tabella, accanto al numero di linee presenti nel *.csv*, sono presentate le percentuali relative ai dati analizzati e quelli scartati per differenti ragioni.

[1] consultabili sulla pagina *github* del gruppo (https://github.com/physycom)

Si può notare che la maggior parte del campione preso in analisi è privo di georeferenziazione, vale a dire che nel file appaiono le informazioni relative agl ID (GeID, CallID) ma non le coordinate associate ad un'attività telefonica svolta dall'utente.
Si tratta di un segnale che sagnala solo la presenza di un attività telefonica ma senza riferirla nè a qualcuno ne al luogo in cui essa è stata effettuata.
Una ulteriore significativa frazione di dati, circa un quarto, pur essendo valida sul profilo dell'identificazione e della georeferenziazione non viene considerata nell'analisi dati in quanto costituita da dati che vengono mappati al di fuori di un area che ragionevolmente approssima la città e le acque limitrofe (*"Bounding Box"*).
L'uso di un *"Bounding Box"* in qualità di discriminante preliminare dei dati garantisce di poter lavorare con dati che saranno sicuramente circoscritti nella stessa area geografica.
Una porzione trascurabile di dati, invece risulta inutilizzabile in quanto sprovvista di dati e pertanto sarebbe impossibile poter discriminare in modo opportuno gli individui, condizione indispensabile al fine di effettuare misure consistenti e coerenti.
Risulta dunque che la frazione di dati che sopravvive a questa catalogazione preliminare si aggira intorno al 19% del campione fornito da *"Tim"*.
Il comune di Venezia afferma che durante la giornata del 26/2/2017 fossero presenti fra le 80,000 e le 100,000 persone e confrontando tali numeri con quanto riportato dal provider *Tim*, ovvero di gestire circa il 30% del mercato di telefonia mobile, possiamo stimare in prima battuta che in base al numero di GeID rilevati la percentuale di penetrazione del campione statistico dovrebbe aggirarsi intorno al $5 - 7\%$.

1.2 Creazione delle Mappe

Per creare un modello che rappresenti il comportamento del campione di dati analizzato, è necessario costruire una lista di siti, approssimati da poligoni che ne circoscrivono il perimetro.
La piattaforma online *Google Maps* prevede la possibilità di svolgere questa funzione manualmente tramite l'applicazione *myMaps*.
É stato dunque "ritagliato" ciascun sito di interesse all'interno della città di Venezia e le coordinate dei vertici di ciascun poligono rappresentante il sito d'interesse sono così state esportate su *file*. I siti di interesse sono stati scelti seguedo la divisione storica della città di Venezia in sei zone chiamate *"Sestieri"*. Le zone di interesse, o sestieri, considerate sono: Cannaregio, Castello, Dorsoduro, Santa Croce, San Marco, San Polo.

Il progetto di analisi dati presentato in questo elaborato, nato per studiare la dinamica di mobilità pedonale, era inizialmente rivolto al considerare una mappa (1.2) costituita da siti di superficie minore rispetto quelli utilizzati in seguito in questa trattazione. La scelta di considerare i Sestieri come zone di interesse, rispetto a siti più piccoli, è motivata nel Capitolo 3 (Sezione 3.1)

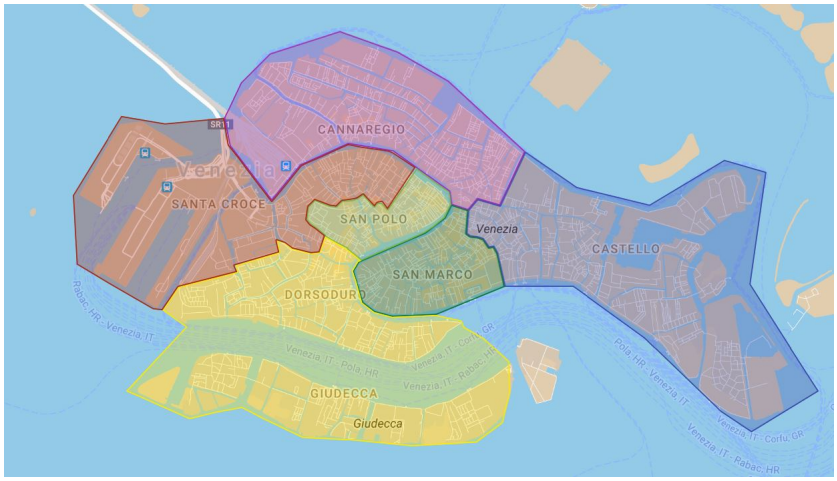

Figura 1.1: La mappa rappresenta Venezia, i poligoni che appaiono colorati circoscrivono i quartieri in cui è suddivisa la città (immagine da:*"Google My Maps"* ©)

Figura 1.2: l'immagine illustra una porzione della mappa su cui il progetto era stato concepito, ma che in seguito è stata abbandonata e sostituita con 1.1

Capitolo 2

Algoritmi Points in Polygon

Uno dei problemi affrontati durante lo studio del nostro modello è stato quello di determinare algoritmicamente se, ad un certo momento, un utente si trovi o meno all'interno di un sito di interesse.

Questo problema può essere studiato ricorrendo al seguente modello nel piano $\mathbb{R}^2$. La posizione di ogni utente è modellizzata da un *punto* nel piano. Più precisamente, ogni riga di file *.csv* contiene le informazioni necessarie all'identificazione e alla geolocalizzazione di un attività svolta da un utente. L'identificazione è tale da garantire un ID (GeID) univoco nel tempo per ciascun utente, senza però far trapelare alcuna informazione sensibile di quest'ultimo.
Ogni sito di interesse è invece rappresentato da un *poligono* o, più precisamente, da un insieme di punti del piano formanti i vertici del poligono semplice (i.e. poligono i cui lati si intersecano solo nei vertici) che meglio approssima il sito considerato.

Il problema di partenza è quindi ridotto al problema di geometria computazionale di determinare l'inclusione o meno di un punto all'interno di un poligono nel piano. Gli algoritmi disponibili per la realizzazione di questo *task* prendono nome di "Algoritmi Points in Polygon". I due algoritmi classici in questo contesto sono:

- il *Crossing Number algorithm* (E.Haines, 1994),(W.R.Franklin, 2000)[1]
- il *Winding Number algorithm* (J.O'Rourke, 1998)[2]
- il *Winding number Inclusion algorithm* (D.Sunday, 2001)[9]

2.1 L'algoritmo *Crossing Number*

Sia Σ un poligono in $\mathbb{R}^2$ e P un punto del piano di cui si vuole determinare la posizione rispetto a Σ. L'algoritmo *Crossing number* conta il numero di intersezioni tra una semiretta uscente da P e i lati del poligono Σ. Intuitivamente, il punto P è *interno* al poligono se il numero di punti di intersezione è dispari, mentre è *esterno* altrimenti.

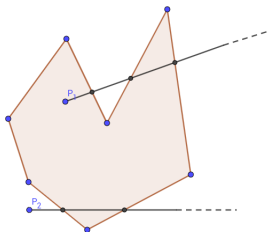

Figura 2.1: Un esempio del metodo *Crossing Number*

Questo risultato si può dimostrare utilizzando il teorema della curva di Jordan. Infatti il bordo di un poligono semplice del piano è una curva di Jordan, ovvero l'immagine di una mappa continua iniettiva dal cerchio $\mathcal{S}^1$ in $\mathbb{R}^2$. Il teorema di Jordan afferma che, data una curva di Jordan che non si intersechi con se stessa in $\mathbb{R}^2$, il suo complementare ha due componenti connesse, una limitata (l'*inside*) e una illimitata (l'*outside*) e la curva è il bordo di ciascuna componente. Quindi un punto esterno a Σ che viaggia sulla semiretta in direzione del punto P cambierà componente connessa ad ogni itersezione con un lato del poligono (transizione *in-out* o *out-in*) e, se il numero di intersezioni è pari, alla fine si troverà all'esterno di Σ.
L'algoritmo non funziona se il punto P è esattamente sul bordo del poligono.

Nell'implementare questo algoritmo, risultano opportuni alcuni accorgimenti. Ad esempio, bisogna garantire che nel conteggio delle intersezioni figurino solo quelle del tipo *in-out* o *out-in* descritto sopra (ovvero i cambiamenti di componente connessa) mentre le possibili intersezioni della semiretta con i vertici di Σ vanno trattate separatamente.

Inoltre, nel caso in cui si voglia considerare l'appartenenza di P o meno rispetto a un'unione di poligoni è opportuno che il conteggio risulti univoco nel caso in cui i siano poligoni adiacenti, per il quale il punto potrebbe giacere sul lato in comune ai due poligoni. In questo caso si dovrà considerare come *bordo del poligono* l'unione dei lati non adiacenti, in modo che l'intersezione fra la semiretta uscente da P ed un lato in comune a due poligoni non cambi la parità del conteggio delle intersezioni.

2.2 L'algoritmo *Winding Number*

Questo algoritmo si basa sul calcolo del numero di avvolgimenti (*winding number*) che compie il poligono attorno al punto.

Ricordiamo che una curva chiusa continua in $\mathbb{R}^2$ è l'immagine di un'applicazione continua $\mathcal{C} : [0,1] \to \mathbb{R}^2$ tale che $\mathcal{C}(0) = \mathcal{C}(1)$.

Sia $\mathcal{C}$ una curva continua chiusa in $\mathbb{R}^2$ e sia P un punto non appartenente alla curva. Il *winding number* $wn(P,\mathcal{C})$ *di* $\mathcal{C}$ *intorno a* P è il numero di avvolgimenti che la curva $\mathcal{C}$ compie intorno al punto P e può essere calcolato nel modo seguente.

Per ogni $u \in [0,1]$, definiamo il vettore $\mathbf{c}(P,u) = \mathcal{C}(u) - P$, congiungente P e $\mathcal{C}(u)$ ed il vettore unitario $\mathbf{w}(P,u) = \frac{\mathbf{c}(P,u)}{|\mathbf{c}(P,u)|}$. A questo è associata l'applicazione lineare continua

$$W(P) : \mathcal{C} \to S^1$$

dalla curva alla sfera unitaria $S^1 = \left\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\right\}$ tale che

$$W(P)\left(\mathcal{C}(u)\right) = \mathbf{w}(P,u).$$

Sia $W(P)(u) = (cos(\theta(u), sin\theta(u))$ la rappresentazione in coordinate polari di $W(P)(u)$, usando la solita convenzione per cui $\theta(u) > 0$ è un angolo in radianti contato in senso antiorario.

Il numero di avvolgimento $wn(P,\mathcal{C})$ di $\mathcal{C}$ intorno a P è dato allora dal seguente integrale:

$$wn(P,\mathcal{C}) = \frac{1}{2\pi}\int_{W(P)} d\theta = \frac{1}{2\pi}\int_{u=0}^{1} \theta(u)du.$$

Osserviamo che il *winding number* $wn(P,\mathcal{C})$ ha un'interpretazione topologica in termini di classe di omotopia. Infatti una curva intorno al punto P può essere deformata in modo continuo in un cerchio S^1 che gira intorno a P un certo numero di volte. Questo numero è dato dalla classe di omotopia della curva $\mathcal{C}$ nel gruppo fondamentale del cerchio, che è isomorfo al gruppo degli interi $\mathbb{Z}$, ed è uguale a $w(P,\mathcal{C})$.

Se invece di una curva chiusa qualsiasi consideriamo un poligono di vertici $V_1, \ldots, V_n$ con $V_n = V_1$, il problema si riduce al calcolo della somma degli angoli orientati sottesi dai lati del poligono rispetto al punto P. Se θ_i è l'angolo sotteso dal lato V_iV_{i+1} rispetto a P, si ha dunque:

$$wn(P,\mathcal{C}) = \frac{1}{2\pi}\sum_{i=1}^{n-1}\theta_i = \frac{1}{2\pi}\sum_{i=1}^{n-1}\arccos\left(\frac{(V_i - P)\cdot(V_{i+1} - P)}{|V_i - P||V_{i+1} - P|}\right) \tag{2.1}$$

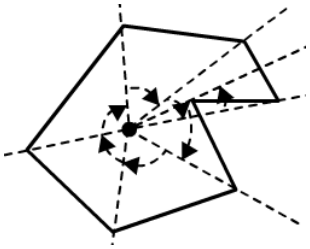

Figura 2.2: Un esempio del metodo *Winding Number* [1]

dove il simbolo $\cdot$ indica il prodotto scalare.

La formula (2.1) mostra immediatamente la dipendenza dell'algoritmo dal calcolo della funzione trigonometrica arccos e dalla funzione radice quadrata $\sqrt{\ }$, necessaria al calcolo delle norme dei vettori al denominatore. Questa dipendenza risulta gravare sul'efficienza computazionale del metodo. Ciò può essere in parte ottimizzato utilizzando l'algoritmo *Fast Inverse Square Root* per calcolare il prodotto fra i quadrati delle norme di $(V_i - P)$ e $(V_{i+1} - P)$, molto famoso nell'ambiente della computer-grafica[2]il quale produce l'inverso della radice quadrata di un numero.
Nonostante questo accorgimento le prestazioni del codice restano comunque affette da un sostanziale gap di efficienza sistematico.

2.3 L'Algoritmo *Winding Number Inclusion*

Esiste un ulteriore metodo per determinare l'inclusione di un punto all'interno di un poligono. Tale algoritmo si basa sul confronto della posizione del punto rispetto ciascun lato del poligono.
Osservando il poligono e scorrendo lungo i lati di questo in senso antiorario è necessario distinguere fra segmenti diretti verso l'alto e segmenti diretti verso il basso. Assumiamo di avere un punto P ed un poligono Σ che lo contiene. Qualora si scorresse lungo i vertici del poligono, si osserverebbe che il punto interno risulta trovarsi a sinistra dei segmenti diretti verso l'alto e a destra di quelli diretti verso il basso.
Consideriamo, ad esempio tre punti P_0,P_1,P_2 ed assumiamo di voler determinare la

[1]Immagine presa da *https://code.tutsplus.com/tutorials/euclidean-vectors-in-flash–active-8192*

[2]La prima applicazione commericale di questo algoritmo risale alla pubblicazione del video-game *'Quake Arena III'* nel 1999 (J.Carmack et al.), anche se è probabile che esistano implementazioni ottenute gia a partire dalla seconda metà degli anni '80.

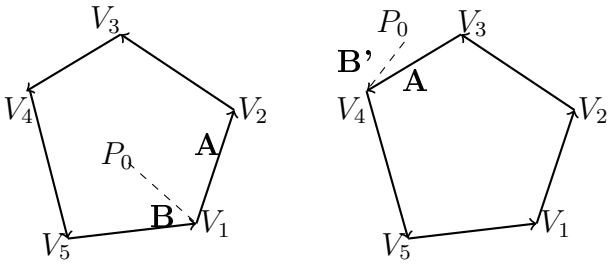

Figura 2.3: Esempio relativo all'applicazione dell'algoritmo (2.3): sia per un punto interno confrontanto con un segmento diretto verso l'alto ed un punto esterno confrontato con un segmento diretto verso il basso.

posizione di P_0 rispetto al segmento formato dagli altri due punti. Costruendo due vettori

$$\mathbf{A} = (A_x, A_y) = (\mathbf{P_2} - \mathbf{P_1}) \text{ e } \mathbf{B} = (B_x, B_y) = (\mathbf{P_0} - \mathbf{P_1})$$

si può notare che la posizione di P_0 rispetto al segmento P_1P_2 dipende dal segno di

$$(\mathbf{A_x B_y}) - (\mathbf{A_y B_x}).$$

La posizione di P_0 rispetto al lato $\mathbf{A}$ in generale, è quindi determinata :

- per un lato $\mathbf{A}$ diretto nel verso delle $y > 0$, dal prodotto vettoriale fra $\mathbf{A}$ e $\mathbf{B}$
- per un lato $\mathbf{A}$ diretto verso le $y < 0$, dal prodotto vettoriale fra $\mathbf{A}$ e $\mathbf{B'}= (\mathbf{P_0} - \mathbf{P_2})$.

Sia ora $P = (x, y)$ un punto e Σ un poligono con vertici $V_1, \ldots, V_n = V_1$ di coordinate $V_i = (x_i, y_i)$. L'algoritmo scorre tutti i vertici del poligono, considera solo quei vertici V_i tali che $y_i < y < y_{i+1}$ (*crossing verso l'alto*) oppure $y_{i+1} < y < y_i$ (*crossing verso il basso*) e confronta la posizione di P rispetto a questi vertici (con il prodotto vettore come spiegato sopra). Incrementando di 1 il valore di un contatore *wn* per ciascuna volta in cui un punto risulta essere "a sinistra" di un segmento e decrementandolo qualora invece si trovi "a destra" di un segmento, si può affermare se il punto risulta essere incluso nel poligono se alla fine risulta $wn \geqslant 1$.

2.4 Scelta e confronto tra gli algoritmi

L'algoritmo scelto in questo progetto è l'algoritmo *Winding Number Inclusion*. In questa scelta l'elemento determinante è stato il confronto della velocità computazionale dei

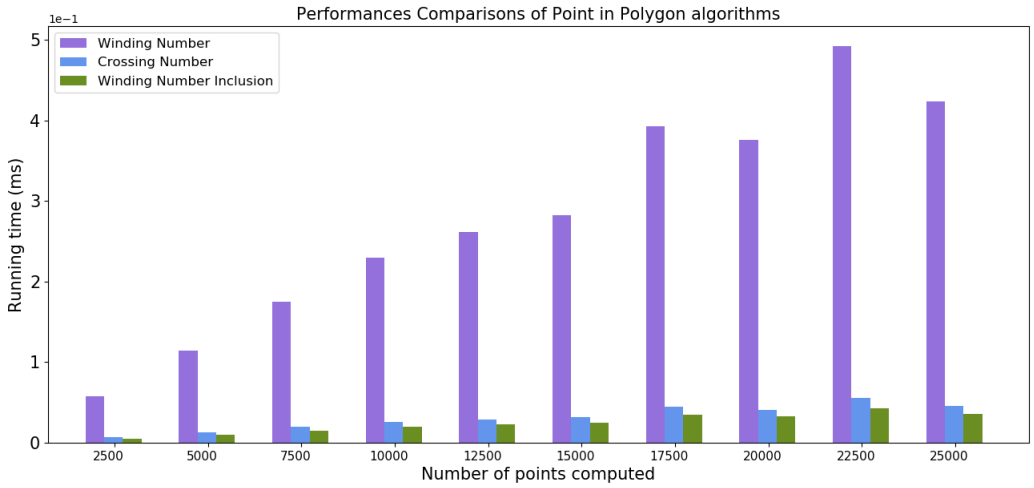

Figura 2.4: Il grafico mostra le prestazioni in termini di tempo di calcolo dei differenti algoritmi utilizzabili per l'inclusione

diversi algoritmi.
Per quanto riguarda la complessità dei poligoni analizzati i metodi *Winding Number Inclusion* e *Winding Number* risultano essere gli algoritmi migliori con cui operare, essendo, per costruzione, meno affetti dell'algoritmo *Crossing Number* dalle caratteristiche geometriche del poligono.
Per quanto riguarda invece le performance in termini di tempi di elaborazione, come accennato nei paragrafi precedenti, *Winding Number* è un algoritmo penalizzato dalla necessità di utilizzo di funzioni matematiche che rallentano i tempi di calcolo. Non è affetto da tale dipendenza, invece, l'algoritmo *Winding Number Inclusion*, che opera solamente attraverso comparazioni ed incrementi (senza ricorrere alle funzioni matematiche già citate in 2.2).

Il seguente grafico 2.4 è stato ottenuto attraverso una simulazione per un solo poligono e per un set di punti generati casualmente, sui quali sono stati applicati serialmente tutti e tre gli algoritmi: *Winding Number*, *Crossing Number* e *Winding Number Inclusion*. Quanto emerge dall'esperienza risulta essere quanto atteso, ovvero che sia l'algoritmo *Winding Number Inclusion* il migliore fra quelli a disposizione in termini di velocità computazionale, versatilità nei confronti dei poligoni scelti ed eleganza.

Capitolo 3

Analisi Dati

3.1 La scelta del *Sestiere*

La scelta di considerare il *Sestiere* come zona di interesse da considerare è stata operata in qualità di compromesso fra le necessità di tassellazione della superficie da analizzare e la garanzia di avere un certo margine di confidenza nei confronti del campione di dati considerato, cercando così di sfruttare a pieno tutta l'informazione disponibile.

Infatti, all'inizio del progetto, i dati raccolti erano stati analizzati cercando di georeferenziarli ad un set ristretto di poligoni non adiacenti, rappresentanti non i Sestieri, ma piccoli siti sparsi in giro per la città, come ad esempio piazzette o locali all'interno dei quali hanno avuto luogo eventi legati alle celebrazioni del Carnevale Veneziano.
Nonostante questo modello ambisse ad un' analisi molto più precisa rispetto a quella elaborata utilizzando la suddivisione della città in Sestieri, i risultati ottenuti utilizzando questo set di poligoni hanno prodotto misure troppo deboli per avanzare ipotesi su un modello di mobilità globale tra i diversi siti considerati.
Il numero totale di persone mappate in almeno uno dei siti che costituiscono la mappa considerata in principio, ovvero quella dei *siti isolati*, risulta essere molto inferiore a quello calcolato per la mappa dei *"Sestieri"*. In media per la prima delle due mappe, la porzione di dato che viene mappato è circa il 24.2% del totale dei GeID registrati nel singolo slot temporale.
Per quanto riguarda la mappa formata dai *Sestieri*, la percentuale di dati georeferenziati rispetto al totale degli utenti registrati fluttua attorno all' $89,9\%$. Graficamente questo fatto appare chiaro osservando come si distribuiscono le prestazioni, in termini di dati

utilizzabili, che si possono ottenere a seconda della mappa utilizzata.

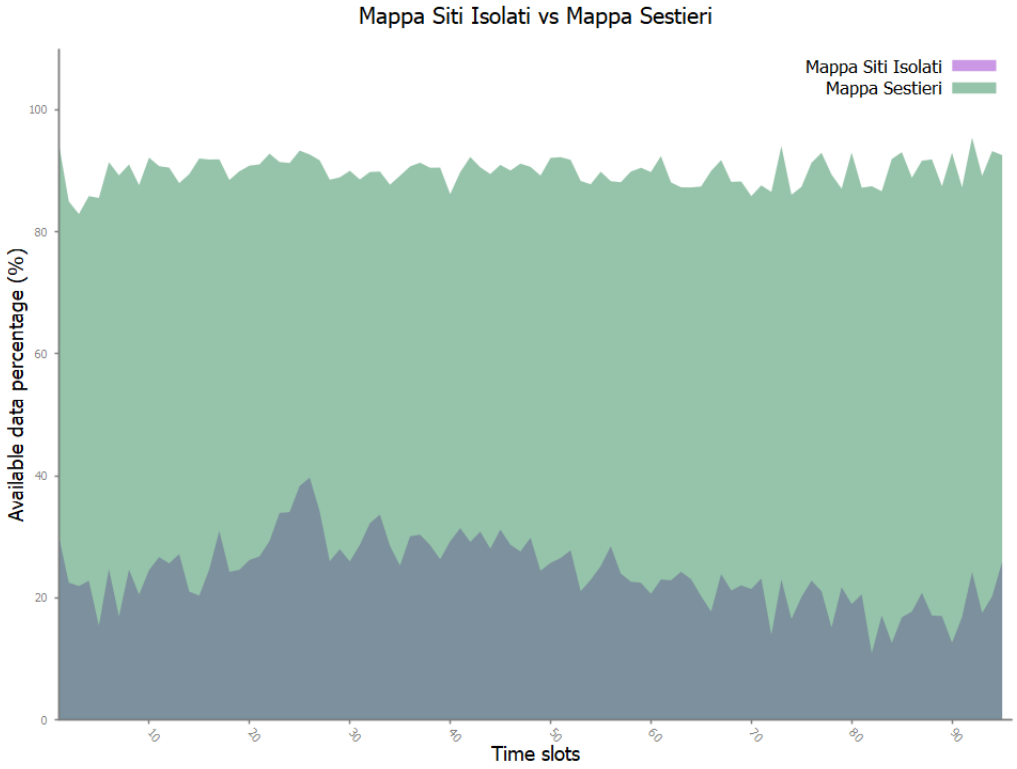

Figura 3.1: Il grafico illustra le percentuali di dati che vengono mappate per ciascun slot temporale durante il corso della giornata a seconda della mappa utilizzata: appare evidente il vantaggio nell'utilizzare la mappa dei Sestieri.

Per questo motivo si è preferita la suddivisione della città in sestieri.

3.2 Gli utenti *fanstama* e il sito *Nowhere*

Un altro problema considerato durante l'analisi dati è quello degli utenti *fantasma*, ovvero utenti che vengono registrati come attivi in una certa finestra temporale, ma che non vengono localizzati in nessun sestiere.

É ragionevole assumere che queste persone, probabilmente, si trovino nei paraggi delle zone considerate in quanto i dati aggregati nel file utilizzato riguardano solo persone all'interno di una area che racchiude il centro di Venezia, qualche isolotto limitrofo, le acque che che la circondano e la percorrono. Si possono avanzare svariate ipotesi sulle attività che stanno svolgendo tali utenti rilevati ma non localizzati, ad esempio che si

trovino su uno dei vaporetti di linea o che stiano raggiungendo la città attraverso il Ponte della Libertà.
Ad ogni modo, possiamo assumere che in media qualcuno fra questi individui riapparirà all'interno delle zone considerate.
Per giustificare la discontinuità che si manifesta fra il numero totale degli utenti attivi in una finestra temporale e la somma degli utenti localizzati nei siti che costituiscono il "Sestiere", é stato istituito un ulteriore sito artificiale ribattezzato *Nowhere*; tale oggetto assume significato, in quanto è un primo passo in avanti nel garantire la conservazione del numero di persone rilevate nell'area di Venezia.

3.3 Una giornata a Venezia: dinamica della popolazione

Nell'andare a misurare le presenze su ciascun sito, si è osservato, confrontando la somma delle popolazioni presenti su ogni sito (considerando anche il nodo "*Nowhere*") in un certo slot temporale, che il numero totale di attività telefoniche e dunque di individui presenti in quell'arco di tempo risulta essere altamente diverso lungo il corso della giornata.

Questo è un fatto sia per quanto riguarda l'alta variabilità in termini di frequenza con cui un individuo utilizza il telefono sia perchè l'effettivo numero di soggetti rilevati in uno slot temporale può risultare estremamente variabile rispetto a quello precedente.
Al di là di questo, si può ipotizzare che in media risultino sempre attivi un numero di utenti tale da permettere l'osservazione di un comportamento di gruppo come l'aumento di popolazione nel corso della giornata.
Questo fatto appare ragionevole, in quanto una buona percentuale di persone che arriva in città per il Carnevale viene appunto rilevata e localizzata man mano che la giornata scorre, facendo registrare un picco delle presenze intorno alle 12:20-12:30.
Il picco di attività registrate in uno dei nodi centrali, *"Castello"*, si registra intorno a tale orario. Ciò può essere dovuto, presumibilemente, al fatto che molti utenti fossero presenti nello stesso sito per assistere ad uno degli eventi previsti nel programma delle celebrazioni.

Allo stesso ci si aspetta che la popolazione complessiva in città decresca nel corso

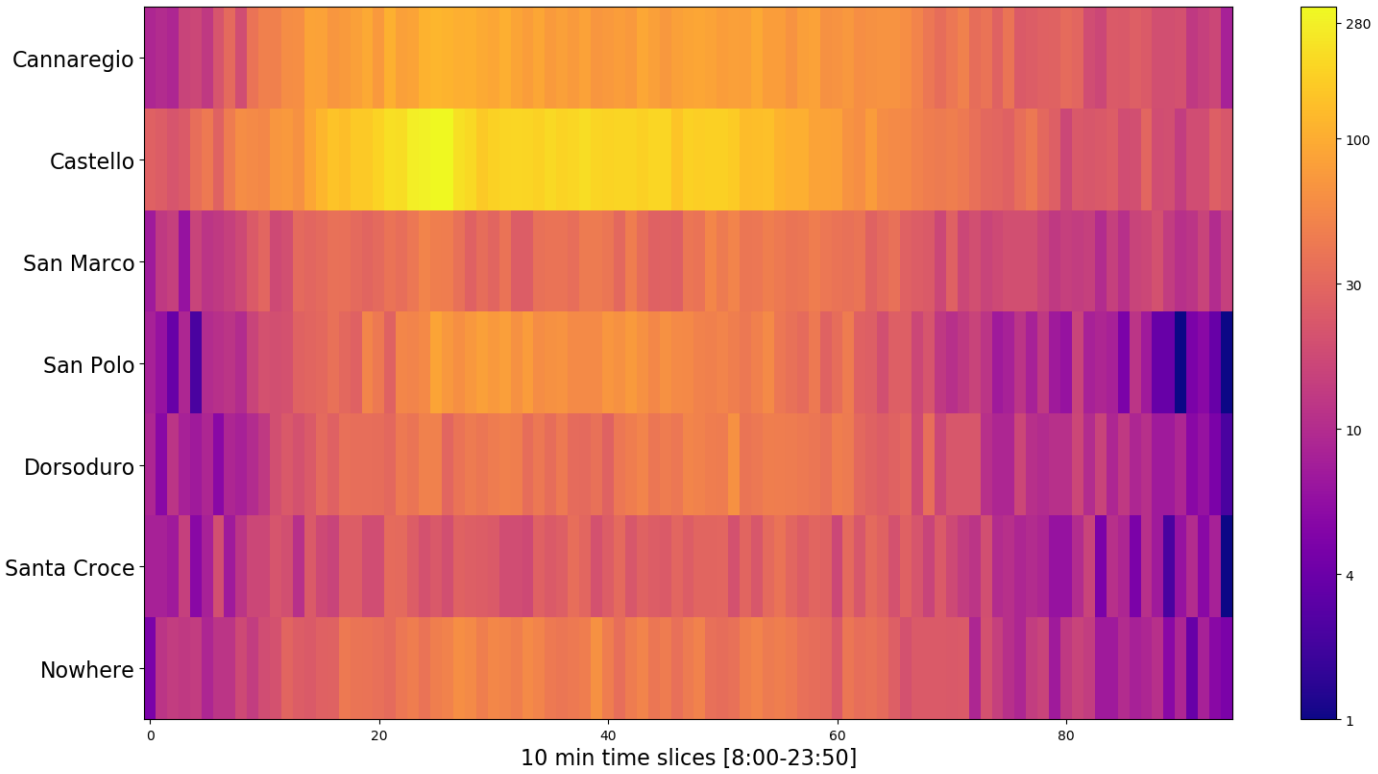

Figura 3.2: Il grafico rappresenta, in scala *log*, i profili registrati per ciascun Sestiere durante il corso della giornata. La colorbar laterale fornisce indicazioni in merito al numero di utenti presenti ad una certa ora in un certo sito

del pomeriggio fino alla serata e in effetti è quanto si registra. In pratica la città si "carica" e si "scarica" nel corso della giornata e nel formulare un modello che esprima la probabilità di transizione di un campione di popolazione da un sito ad un altro bisognerà necessariamente tener conto di questo fenomeno.

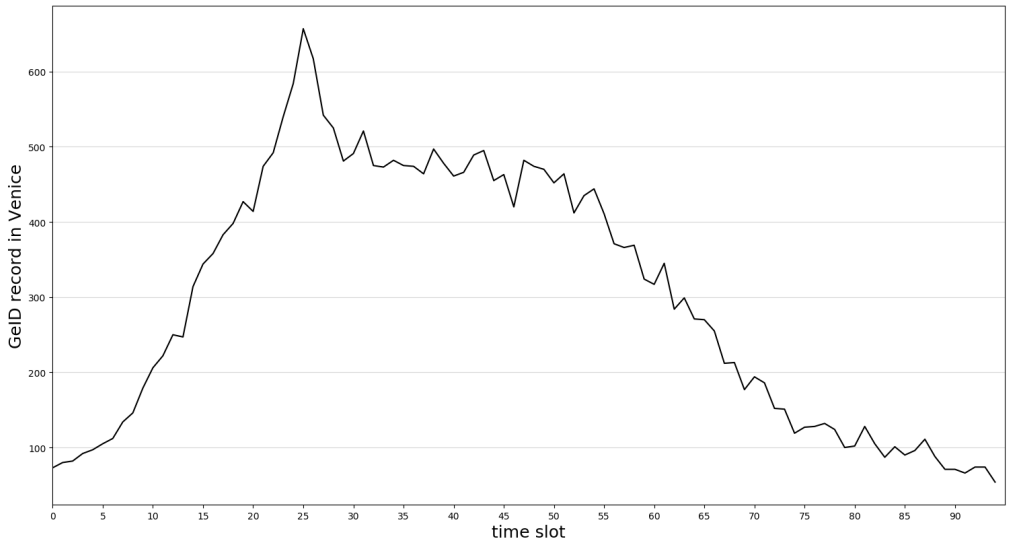

Figura 3.3: Il grafico rappresenta il numero di attività rilevate per tutta l'area inscritta nel *Bounding Box* durante il corso della giornata

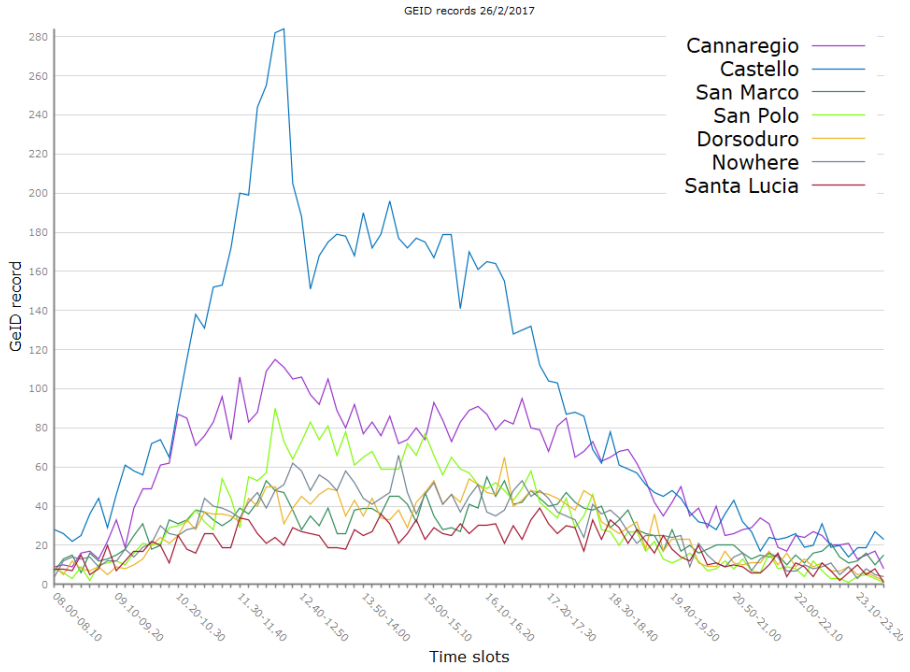

Figura 3.4: Il grafico rappresenta il numero di attività rilevate per tutta l'area inscritta nel *Bounding Box* durante il corso della giornata

Capitolo 4

Modelling

Il seguente capitolo illustra dapprima alcuni brevi concetti preliminari all'elaborazione dei dati analazzati: si intende fornire il lettore di qualche semplice argomento teorico con il quale è stato messo a punto il modello con cui andare ad interpretare i dati analizzati. La sezione successiva, infatti, sarà incentrata su quanto è stato fatto sulla base dei dati catalogati ed analizzati, come si è proceduti nella costruzione della matrice di transizione basandoci sugli elementi teorici esposti nella sezione.

4.1 Richiami Teorici

4.1.1 Notazione

Consideriamo uno spazio delle probabilità $(\mathbf{\Omega}, \mathbf{A}, \mathbf{P})$ dove $\mathbf{\Omega}$ è un insieme di campioni misurati, $\mathbf{A}$ è un insieme di eventi possibili (dotato di σ- algebra) mentre $\mathbf{P}$ è una misura di probabilità sull'insieme $\mathbf{\Omega}$.
Si assuma di avere una variabile random $X(t)$, a valori $\mathbb{R}$ e dipendente in modo continuo da un certo parametro t (in genere il tempo). Per descrivere un processo che si voglia definire *stocastico* occorre descrivere i possibili valori che la $X(t)$ può assumere ad ogni istante t (o $\forall$ valore del parametro t), i cambiamenti che può eventualmente subire ed il grado di correlazione con gli eventi che hanno preceduto tali cambiamenti.
Senza determinare tutto questo, non si può affermare di riconoscere un processo stocastico.

Definizione 1. *Si chiama* processo stocastico*[8] un processo per il quale le funzioni*

$$F(t_1, t_2, \ldots, t_n, x_1, x_2, \ldots, x_n) = P[X(t_1) < x_1, X(t_2) < x_2, \ldots, X(t_n) < x_n]$$

sono fissate $\forall n \in \mathbb{N}$ *e per ogni istante* $t_1, t_2, \ldots, t_n$.

Definizione 2. *Dato* $\omega \in \mathbf{\Omega}$ *, il set di valori* $X(\omega) = \{X_0(\omega), X_1(\omega), X_2(\omega), ...\}$ *si chiama* realizzazione *del processo stocastico* X *associato a* ω.

4.1.2 Catene di Markov

Definizione 3. *Un processo stocastico discreto nel tempo* $X = \left\{X_k\right\}_{k\in\mathbb{N}}$ *si dice* catena di Markov[1] omogenea *su uno spazio di stati numerabili* $\mathbf{S}$ *se vale la seguente proprietà (detta* di Markov*):*

$$\mathbf{P}[X_{k+1} = z \mid X_k = y, X_{k-1} = x_{k-1}, ..., X_0 = x_0] = \mathbf{P}[X_{k+1} = z \mid X_k = y] \tag{4.1}$$

$$\forall k \in \mathbb{N} \text{ e } \forall x_0, x_1, .., x_{k-1}, y, z \in \mathbf{S} \tag{4.2}$$

dove ricordiamo che se A *e* B *sono due eventi nello spazio di probabilità si ha*

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

Osserviamo, in particolare, che per una catena di Markov si ha

$$\mathbf{P}[X_{k+1} = z \mid X_k = y] = ... = \mathbf{P}[X_1 = z \mid X_0 = y] \tag{4.3}$$

Per una catena di Markov omogenea, la funzione $P(y, z) = \mathbf{P}[X_{k+1} = z \mid X_k = y]$ prende nome di *funzione di transizione* e i valori assunti da $P(y, z)$ sono le probabilità di transizione dallo stato y allo stato z.
Denotiamo $\mu_0(x) = \mathbf{P}[X_0 = x]$ la *distribuzione iniziale.*
Inoltre se esiste $x \in \mathbf{S}$ tale che $\mu_0(x) = 1$, x prende valore di stato iniziale.

Definizione 4. *Una matrice* $\mathbf{\Pi} = (\pi_{xy})_{x,y\in\mathbf{S}}$ *si definisce* stocastica *se i suoi elementi* $\pi_{xy} \geq 0$ *e se* $\sum_{y\in S} \pi_{xy} = 1$, $\forall x, y \in \mathbf{S}$

[1] dal nome del matematico russo A.A. Markov (Rjazan' 1856 - San Pietroburgo 1922)

Ogni catena di Markov definisce attraverso la sua funzione di transizione una matrice stocastica.
Più in generale un processo stocastico può essere definito sulla base delle ditsribuzioni congiunte:

$$\mathbf{P}[X_m = x_m, X_{m-1} = x_{m-1}, ..., X_0 = x_0] \tag{4.4}$$

con $m \in \mathbb{N}$ e x_m nello spazio degli stati possibili. Utilizzando la proprietà di Markov si ha

$$\mathbf{P}[X_m = x_m, X_{m-1} = x_{m-1}, \dots, X_0 = x_0] = \mathbf{P}[X_m = x_m \mid X_{m-1} = x_{m-1}, \dots, X_0 = x_0] \cdots$$
$$\cdots \mathbf{P}[X_2 = x_2 \mid X_1 = x_1 \mid X_0 = x_0] \cdot \mathbf{P}[X_1 = x_1 \mid X_0 = x_0] \cdot \mathbf{P}[X_0 = x_0] =$$
$$= \mathbf{P}[X_m = x_m, X_{m-1} = x_{m-1}] \cdots \mathbf{P}[X_2 = x_2 \mid X_1 = x_1] \cdots$$
$$\cdots \mathbf{P}[X_1 = x_1 \mid X_0 = x_0] \cdot \mathbf{P}[X_0 = x_0].$$

Quindi per calcolare le probabilità di transizione di un cammino semplice si deve iniziare dal considerare la probabilità dello stato iniziale ed si effettua il prodotto fra questa e le probabilità di transizione condizionali.

4.1.3 Costruzione di una catena di Markov

Teorema 1. *Sia $\{\xi_k\}_{k\in\mathbb{N}}$ un set di variabili aleatorie indipendenti ed egualmente distribuite, i cui valori appartengono ad un qualche spazio misurabile $\mathbf{Y}$. Sia inoltre X_0 una variabile aleatoria a valori in $\mathbf{S}$ e indipendente da $\{\xi_k\}_{k\in\mathbb{N}}$.*

*Si consideri una funzione $\mathbf{f} : \mathbf{S} \times \mathbf{Y} \to \mathbf{S}$, allora è possibile definire un **sistema stocastico dinamico** definito dall'equazione di ricorrenza: $X_{k+1} = f(x_k,\xi_k)$ che definisce una* catena di Markov omogenea *$X = \{x_k\}_{k\in\mathbb{N}}$ nello spazio degli stati $\mathbf{S}$[4].*

Sia ora $(\xi_k)_{k\in\mathbb{N}}$ una sequenza di variabili aleatorie, egualmente distribuite, indipendenti fra loro e indipendenti da X_0 con valori in $\mathbf{Y} = \{-1, +1\}$.
Vale

$$\mathbf{P}[\xi_k = 1] = q \qquad \text{mentre} \qquad \mathbf{P}[\xi_k = -1] = 1 - q$$

per $q \in (0, 1)$.
Allora la catena di Markov $(\xi_k)_{k\in\mathbb{N}}$ con $\mathbf{S} = \mathbb{Z}$ è definita da: $X_{k+1} = X_k + \xi_k$.
Qualora la catena, invece, fosse definita in termini di una certa matrice stocastica di transizione Π, allora si andranno a definire la famiglia $(X_k)_{k\in\mathbb{N}}$ assieme con i $(\xi_k)_{k\in\mathbb{N}}$,

indipendenti ed identicamente distribuiti con continuità sull'intervallo $[0, 1]$.
La relazione di ricorrenza ora viene ad essere descritta come: $X_{k+1} = f(x_k, \xi_k)$, dove però $\mathbf{f} : \mathbf{S} \times [0, 1] \to \mathbf{S}$ e $f(x, y) = z$ con $\sum_{y=1}^{z-1} \mathbf{P}(x, y) \leq y \leq \sum_{1}^{z} \mathbf{P}(x, y)$.

4.1.4 Variabili random e processi di Markov

Risulteranno importanti nella trattazione seguente una particolare categoria di variabili random implicate nei processi Markoviani, sia in quelli continui che in quelli discreti. Si consideri dunque una variabile random continua $\tau : \Omega \to \mathbb{R}^+$ che soddisfi la condizione $\mathbb{P}[\tau > s] = \exp(-\lambda s)\ \forall s \geq 0$ e con $\lambda \geq 0$; essa prende nome di *"random exponential variable"*.
La densità di probabilità di tale variabile è una certa funzione $f : \mathbb{R}^+ \to \mathbb{R}^+$ con: $f(s) = \lambda e^{-\lambda s}$ per $s \geq 0$ ed $f(s) = 0$ per $s < 0$.
Di particolare rilievo è la proprietà che assumono queste variabili considerate all'interno di un sistema che evolve nel tempo, vale a dire quella di omogeneità.

$$\mathbf{P}[\tau > t + s \mid \tau > t] = \mathbf{P}[\tau > s] \qquad \forall s, t \geq 0$$

Ai fini dell' analisi svolta nella sezione successiva, utilizzeremo una variante discreta delle exponential random variable, le cosiddette **"Poisson random variable"** definite come $\mathbf{N} : \mathbf{\Omega} \to \mathbb{N}$ e con distribuzione di probabilità

$$\mathbf{P}[N = k] = \frac{\lambda^k e^- k}{k!}$$

dove $k \in \mathbb{N}$.

4.2 Il Modello

Il modello che si intende costruire è una generalizzazione del *template* di *Master Equation*. Con il termine "Master Equations" si intende l'insieme delle relazioni che permettono di modellizzare l'evoluzione di un sistema sulla base di considerazioni probabilistiche, dove il sistema è generalmente descritto con un numero discreto di stati ed in cui la transizione fra stati è trattata appunto in modo probabilistico.
Una Master Equation ha generalmente la struttura di equazione differenziale:

$$\dot{\vec{P}} = A\vec{P}$$

In prima battuta la differenza che sussiste fra il nostro modello ed una qualunque master equation è caratterizzato dal fatto che la soluzione delle equazione rappresentante il modello non viene interpretata come distribuzione di probabilità cumulativa di un insieme di eventi ma, generalizzando, come distirbuzione di un certo numero di utenti telefonici all'interno della città ad un certo orario.
Tale quantità non può conservarsi in quanto il numero totale di utenti rilevati può variare fra slots temporali ed occorrono termini supplementari al fine di comprendere e poter rappresentare in modo opportuno il fenomeno.
Il maggior effetto sui quali occorre lavorare è il fatto che la popolazione entra ed esce dalla città rendendo il numero di persone presenti in un intervallo temporale, non omogeno nel corso della giornata.
Non è nemmeno trascurabile il fatto che il momento in cui un utente decide di utilizzare il telefono risulta non essere prevedibile.
A tal proposito viene giustificato l'uso delle *poisson random variable*(4.1.4).

Cominciamo dunque col delineare gli elementi che concorrono alla costruzione del modello.

Viene indicata con $n_i(t)$ la popolazione misurata nel sito i-esimo al tempo t quella presente invece al tempo $t + \Delta\ t$ viene indicata con $n_i(t + \Delta t)$.
L'ipotesi che viene messa a punto è che la popolazione che viene misurata in un qualunque altro sito j in un istante temporale successivo alla prima misura (cioè a $t + \Delta t$), sia in realtà l'evoluto della popolazione misurata al tempo t, attraverso la relazione:

$$n_i(t + \Delta t) = \Big[\Pi_{i,j} + c(t)\delta_{ij}\Big] n_j(t) + \xi_i \tag{4.5}$$

dove ξ_i è un ipotetico rumore da cui è affetto il segnale relativo al dato della popolazione nel sito i-esimo al tempo t, e che per costruzione si assume che:

$$\sum_i \xi_i(t) = 0 \qquad e \qquad \sigma^2_{\xi_i} = 1 \tag{4.6}$$

cioè che esso abbia media nulla e varianza unitaria σ^2_x.

Tale scelta fornisce al sistema la caratteristica di essere sprovvisto di memoria, in riferimento (4.1.4). Si assume dunque che la distribuzione seguita dalla variabile aleatoria *rumore*, cioè ξ_i, sia di tipo *poissoniano*: $\xi \propto \frac{e^{-1}}{k!}$ e che giustifichi sul piano teorico la fluttuazione sul dato misurato.

Per quanto riguarda la $\Pi_{i,j}$, alla luce della generalizzazione fatta a partire dal modello di master equation utilizzato come template della nostra indagine, essa andrà a rappresentare la matrice di scambio (def.4) di un campione di popolazione da un sito ad un altro. Una volta costruita è opportuno verificare che essa rispetti due caratteristiche:

$$\sum_i \Pi_{i,j} = 1 \qquad e \qquad \Pi_{ij} \geq 0 \tag{4.7}$$

in quanto Π_{ij} deve soddisfare le condizioni che la rendono una matrice stocastica.
La $c(t)$ è una costante che è stata introdotta per compensare l'aumento o la diminuzione della popolazione presente in un sito rispetto quello precedente ed è dunque da determinare per ciascun *bin* temporale.

Questo elemento si è reso necessario, in quanto l'attività telefonica è altamente aleatoria e non è detto che un individuo che utilizzi il telefono in un certo sito lo riutilizzi nuovamente o nello stesso sito o in uno differente. Inoltre l'aumento della popolazione totale registrata fra un intervallo temporale e il consecutivo risulta spesso avere effetti non trascurabili.
Si può dire che in un certo senso, $c(t)$ intervenga in qualità di moderatore del dato rilevato fra un certo istante t e quello consecutivo; $c(t)$ è normalizzato sulla popolazione presente nel primo dei due *bin* considerati. In altri termini rappresenta l'aumento o la diminuzione relativa del campione misurato in funzione del tempo.
Quello che si ottiene è un modello *predittivo* ad uno step; si ipotizza in tal modo che gli eventi rilevati costituiscano una *catena di Markov*.

Cominciamo con l'osservare che:

$$g_i(t + \Delta t) = [\Pi_{i,j} + c(t)\delta_{ij}]n_j(t) \tag{4.8}$$

$$\sum_i g_i(t + \Delta t) = \sum_i n_i(t) + c(t) \sum_i n_i(t) \tag{4.9}$$

$$= \sum_i n_i(t + \Delta t) \tag{4.10}$$

da ipotesi, inoltre:

$$N(t+\Delta t) = N(t) + c(t)N(t) \tag{4.11}$$

$$c(t) = \frac{N(t+\Delta t) - N(t)}{N(t)} \tag{4.12}$$

che è calcolato a parte, utilizzando i dati elaborati.
Definiamo dunque ϵ l'errore associato alla predizione che può essere ottenuta sulla base del modello:

$$\epsilon = \frac{1}{2}\sum_t \sum_i \left[n_i(t+\Delta t) - (\Pi_{ij} + c(t)\delta_i j) n_j(t) \right]^2 \tag{4.13}$$

Tale errore sulla predizione del sistema al tempo $(t+\Delta t)$ deve essere necessariamente minimizzato:

$$\frac{\partial \epsilon}{\partial \Pi_{ij}} = \sum_t n_j(t) \left[n_i(t+\Delta t) - \sum_k (\Pi_i k + c(t)\delta_i k) n_k(t) \right] = 0 \tag{4.14}$$

e tale scrittura rappresenta la condizione di minimo. Assumendo ragionevolmente valida l'ipotesi secondo cui ξ_i i campioni di popolazione rilevati in ciascun sito ad un certo istante, $n_i(t)$, siano indipendenti nel tempo si può affermare che:

$$\sum_t n_j(t) n_i(t+\Delta t) - \sum_t n_j(t)[\Pi_{ij} + c(t)\delta_i k] n_k(t) = \tag{4.15}$$

$$= \langle n_j(t) n_i(t+\Delta t)\rangle_t - \Pi_{ik}\langle n_k(t) n_j(t)\rangle_t - \langle c(t) n_j(t) n_i(t)\rangle_t \tag{4.16}$$

In forma matriciale:

$$C_{kj}(t) = \langle n_k(t) n_j(t)\rangle_t, \quad C_{ij}(t+\Delta t) = \langle n_i(t+\Delta t) n_j(t)\rangle_t, \quad T_{ij}(t) = \langle c(t) n_k(t) n_j(t)\rangle_t \tag{4.17}$$

É quindi possibile ottenere finalmente l'espressione relativa alla Π_{ij} risolvendo il sistema in (4.15)

$$\Pi_{ik} = C_{ij}(t;\Delta t) C^{-1}_{jk}(t) - T_{ij}(t) C^{-1}_{jk} \tag{4.18}$$

$$\langle n_k(t) n_i(t+\Delta t)\rangle_t = \Pi_{ij}\langle n_k(t) n_j(t)\rangle_t + \langle c(t) n_k(t) n_i(t)\rangle_t \tag{4.19}$$

4.3 Risultati

La soluzione ottenuta in (4.18) rappresenta la matrice che descrive le transizioni da uno stato ad un altro sulla base dei dati catalogati ed analizzati.
Ora occorre che vengano soddisfatte delle condizioni che rendano la soluzione trovata, consistente con il modello e dunque con le premesse istanziate in precedenza. Ciò che bisogna verificare inizialmente è che Π_{ij} rappresenti davvero la soluzione dell'equazione che descrive il modello.
Definiamo $\alpha_k = \sum_i \Pi_{ik}$ come vettore colonna i cui elementi sono le somme dei termini su ciascuna riga delle Π_{ik}.
Deve valere la condizione:

$$\sum_k \alpha_k C_{kj}(t) - \sum_t N(t) n_j(t) = 0 \tag{4.20}$$

$$\sum_t \Big(\alpha_k n_k(t) - N(t)\Big) n_j(t) = 0 \qquad \forall j. \tag{4.21}$$

Inoltre se vale () devono valere anche:

$$\sum_k \alpha_k n_k(t) - N(t) = 0 \qquad \forall t$$

e che α_k sia un vettore di dimensione pari al numero dei siti e composto da 1.
Tali proprietà sono soddisfatte dalla soluzione ottenuta.
Per verificare inoltre l'attendibilità del modello abbiamo calcolato la discrepanza fra il dataset misurato e quello ottenuto attraverso la matrice di scambio Π_{ij} ed il coefficiente $c(t)$ applicati dapprima al vettore delle popolazioni nei siti al tempo t_0 dal quale si è poi ricostruito un nuovo dataset applicando consequenzialmente la trasformazione al vettore popolazione precedentemente ottenuto.

Ciò significa:

$$\epsilon_i = \frac{\sqrt{\sum_t \left[n_\pi(t+\Delta t)_i - n_{mis}(t+\Delta t)_i \right]^2}}{\sqrt{\sum_t \left[\sum_i n_{mis}(t+\Delta t)_i \right]^2}}$$

Dove $n_\Pi(t+\Delta t)_i = \Pi_{ij} n(t)_j + c(t) n(t)_i$ ed $n_{mis}(t+\Delta t)$ rappresenta il vettore contente le popolazioni su ciascun sito misurate al tempo $(t+\Delta t)$.
Allo stesso modo abbiamo calcolato la discrepanza calcolata fra il dataset misurato e quello generato da una trasformazione identità a cui sono stati sommati i vettori moltiplicati per la costante $c(t)$, ovvero $n_\mathbb{I}(t+\Delta t)_i = \mathbb{I}_{ij} n(t)_j + c(t) n(t)_i$. Da cui:

$$\epsilon_i = \frac{\sqrt{\sum_t \left[n_\mathbb{I}(t+\Delta t)_i - n_{mis}(t+\Delta t)_i \right]^2}}{\sqrt{\sum_t \left[\sum_i n_{mis}(t+\Delta t)_i \right]^2}}$$

In praticà così sono stati confrontati due diversi approcci allo stesso problema, cioè quello della ricostruzione della dinamica di scambio fra i siti.

Quanto ci si aspetta è in effetti che il modello costruito sulla base di una matrice di scambio Π_{ij} ed un coefficiente $c(t)$ che modera l'aumento o la diminuzione della popolazione presente i città ad una certa ora risulti essere più efficace di un modello che invece afferma che il sistema evolva solo attraverso il contributo di una trasformazione Identità a cui vanno sommati i contributi dati dall'incremento o la diminuzione della popolazione presente in città ad un certo istante t. In figura (4.1) viene riportato l'errore calcolato ϵ_i in funzione dei "Sestieri".

Va detto a proposito della Π_{ij}, che essa presenta elementi negativi, in apparente contraddizione con le ipotesi. Tale contraddizione può essere ricondotta ad effetti presenti nel fenomeno reale che il modello teorico, allo stato dell'arte, non è in grado di catturare e per i quali, al momento, è possibile solo avanzare ipotesi.
L'equazione del bilancio dettagliato (4.5) permette di costruire una soluzione al sistema tale da fornire le probabilità di transizione, i valori Π_{ij}, della popolazione dal nodo i al nodo j.

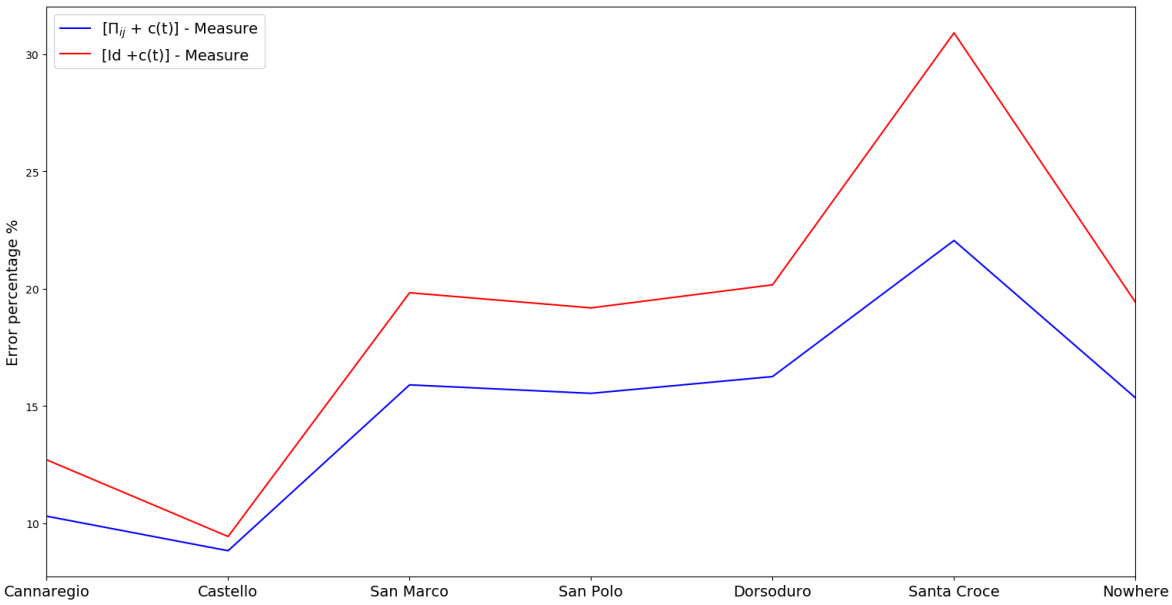

Figura 4.1: Il grafico rappresenta l'errore relativo associato ai risultati ottenuti attraverso due modelli differenti.

L'immagine in fig (4.2) rappresenta una binarizzazione della matrice Π_{ij}, costruita a scopo qualitativo per evidenziare quali siano i siti in cui il modello riscontra maggiori criticità.

Possiamo confrontare inoltre, questa immagine con un grafico rappresentante l'andamento nel tempo di ciascun sito nel corso della giornata, ed evidenziare dunque quelli per cui alcuni elementi della matrice Π_{ij} risultano negativi.
Qualitativamente, ad esempio, si può osservare che ad un picco associato al sestiere "Castello" osserviamo invece un minimo locale per quanto riguarda i sestieri "Santa Croce" e "Dorsoduro". L'ipotesi che si può avanzare osservando il grafico in corrispondenza del picco del segnale misurato per "Castello" è che i coefficienti della matrice di scambio associati a questa interazione fra nodi sia negativa in virtù del fatto che l'aumento massiccio di utenti rilevati in quel sito implichi l'afflusso di utenti dagli altri siti misurati. Talvolta però, come nel caso del sestiere "Castello" è ragionevole affermare che i siti limitrofi non riescano a compensare l'ingente aumento di utenti rilevati nel sestiere corrispondente ai coefficienti negativi.

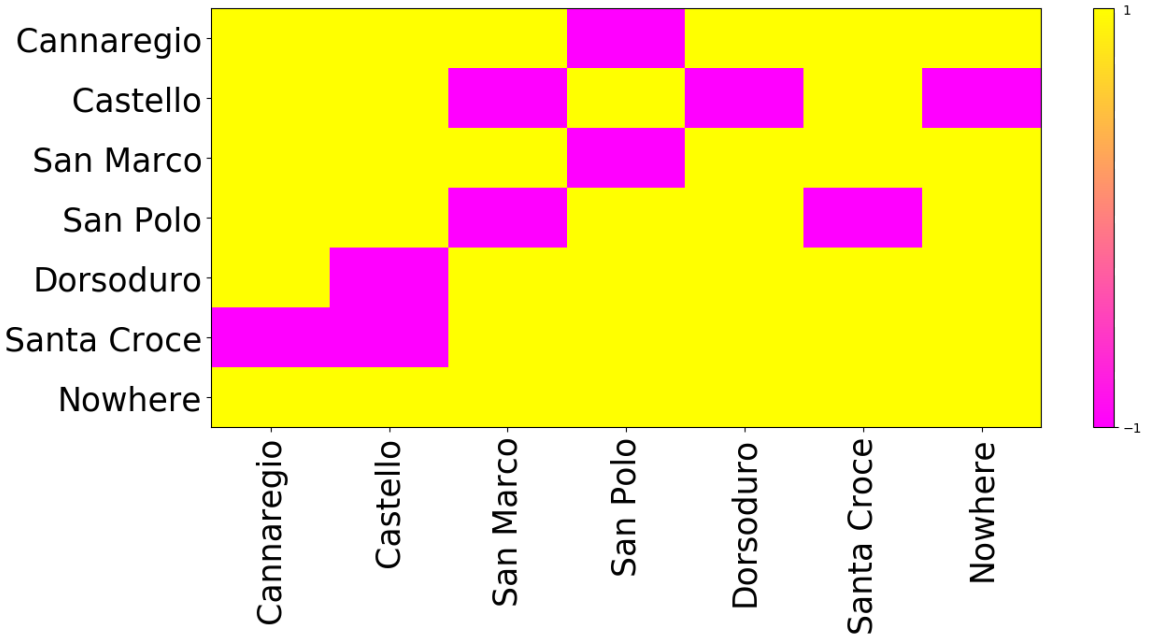

Figura 4.2: I rettangoli in giallo nella figura rappresentano le intersezioni fra siti per i quali i coefficienti della Π_{ij} sono negativi. Tale immagine è stata ottenuta tramite binarizzazione della matrice ed è un indicatore puramente qualitativo delle criticità manifestate e di dove esse si collocano"

Fra due slot temporali possono esistere però variazioni significative in termini del numero di utenti rilevati, ed il modello non è in grado di bilanciare tali variazioni. Tale situazione si manifesta tramite l'apparizione di coefficienti negativi laddove ci si aspettavano coefficienti solamente positivi.

Dal momento che gli elementi di Π_{ij}, rappresentano con quale probabilità di transizione da un nodo a quelli ad esso collegati, una possibile interpretazione è che il modello compensi eventuali discontinuità mostrando che il nodo che dovrebbe cedere popolazione a quelli limitrofi in realtà si ritrova ad assorbirne. Si veda fig(4.3).

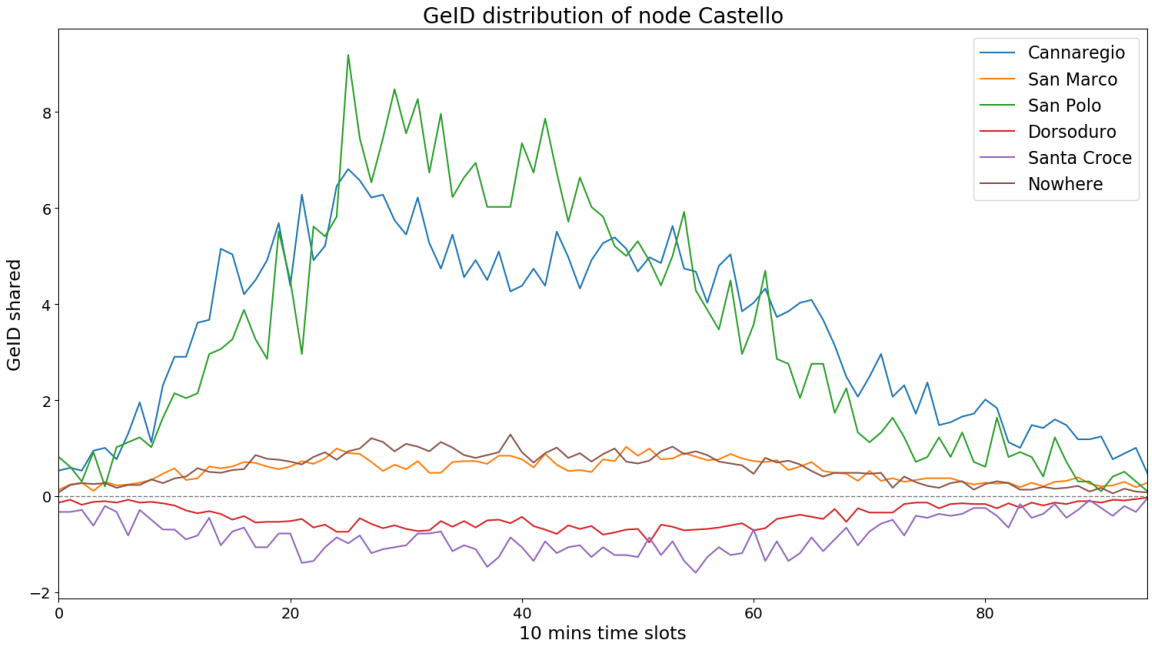

Figura 4.3: Il grafico rappresenta la distribuzione degl utenti in uscita dal nodo "Castello"; da notare sono le corrispondenze negative per i siti "Dorsoduro" e "Santa Croce".

Giunti a questo punto possiamo aspettarci che reiterando il processo di analisi dati e di modelling su data set costruiti a partire da giornate *medie*(probabilmente un qualsiasi giorno della settimana al di fuori del periodo di Carnevale o del "Redentore"[2]), allora il comportamento medio dovrebbe essere attendibilmente rappresentato dal modello.

[2]celebrazione estiva che richiama a Venezia un numero di persone confrontabile con quelle presenti al Carnevale

Conclusioni

Attualmente il progresso in termini di infrastrutture digitali ed interesse nelle loro potenziali applicazioni coinvolge differenti campi d'indagine. Per ciò che concerne questa tesi l'attenzione è stata focalizzata sullo studio delle città e della mobilità pedonale all'interno di esse, in relazione alle esigenze sempre maggiori di organizzazione e di controllo in termini di flussi di persone. É stata infatti messa in luce l'importanza che assumono i dati di telefonia per ciò che riguarda il loro utilizzo in qualità di indicatore della popolazione presente all'interno di Venezia e come tale dato possa essere utilizzato con finalità di modellizzazione del comportamento di essa.
In principio sono stati introdotti gli elementi che hanno costituito il data set con il quale è stato condotto lo studio e gli algoritmi di geometria computazionale con cui è stato possibile procedere alla georeferenziazione dei dati disponibili.
L'analisi dati condotta sulla base dei dati georeferenziati insieme alla formulazione di un modello probabilistico, hanno portato alla costruzione di una matrice di transizione i cui termini rappresentano le probabilità con cui si ridistribuiscono le persone all'interno della varie zone della città.
Il modello ottenuto rappresenta dunque un' approssimazione di quello che può essere il comportamento medio delle persone presenti a Venezia; si rilevano discrepanze all'interno dei risultati che indicano come il modello elaborato non sia in grado di cogliere quelle situazioni che si discostano in modo significativo dal comportamento medio. Si lascia, in questo modo, spazio a future analisi volte a risolvere questo problema o, quanto meno, ad includere nel modello quegli effetti che fino ad ora sono stati trascurati e che possono, una volta presi in considerazione, rendere un futuro modello in grado di descrivere quei contesti che si discostano dal comportamento medio.

Bibliografia

[1] Eric Haines, "Point in Polygon strategies", article in *"Graphic Gems IV"*, Paul S. Heckbert ed. (1994)

[2] Joseph O'Rourke, *"Computational Geometry in C"* II edition, Cambridge University Press, (1998)

[3] S. Gambs, M. Killijian et al, "Next Place Prediction Using Mobility Markov Chains" , article in*" Proceedings of the First Workshop on Measurement, Privacy, and Mobility"*, ACM press, (2012)

[4] W. Huisinga, E.Meerbach, *"Markov Chain for Everybody - An Introduction to the theory of discrete time Markov chains on countable state spaces"*, notes, Free University of Berlin online available pdf, `http://compphysiol.mi.fu-berlin.de/teaching/downloads/SS05_MP/MarkovProzessesScript_FuerDieKlausur.pdf`, (2005)

[5] A. Eberle, *"Markov Processes Skript"*, notes, University of Bonn online available pdf,`https://wt.iam.uni-bonn.de/fileadmin/WT/Inhalt/people/Andreas_Eberle/MarkovProcesses/MPSkript1415.pdf`, (2017)

[6] F.Alhasoun, *"Understanding and Modelling Human Movement in Cities using Phone Data"*, thesis, Massachussets Institute of Technology, (2016)

[7] M.Batty, KW. Axhausen, F.Giannotti, A.Pozdnoukhov, A.Bazzani, M.Wachowicz, G.Ouzounis, *"Smart cities of the future"*, article in "The European Physical Journal Special Topics",(2012)

[8] B.V.Gnedenko, *"Theory of probability"*,Mir Publishers, IV edition,(1978)

[9] D.Sunday, "Geometry Algorithms Home, web page, http://geomalgorithms.com/, (2001)