Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea in Fisica

Strutture dinamiche nei dati di telefonia mobile in contesto urbano

Relatore:
Chiar.mo Prof.
ARMANDO BAZZANI

Presentata da:
GIOVANNI CANTALINI

Sessione II
Anno Accademico 2013/2014

Un particolare ringraziamento al Prof. Giorgini

Abstract

Lo scopo del presente lavoro é di stabilire se l'attivitá di una rete telefonica obbedisce a leggi statistiche e verificare se sono presenti delle fluttuazioni rispetto a queste ultime. Nella prima parte vengono illustrati alcuni modelli di *Human Dynamics*. Nella seconda parte viene descritta l'analisi dei dati soffermandosi su alcuni punti chiave che determinano il significato del risultato finale.

Indice

Introduzione

Il comportamento umano regola fenomeni sociali, tecnologici ed economici e quindi la sua comprensione attira un forte interesse. La costruzione di modelli che riescano a descrivere la *Human Dynamics* é oggi legata alla disponibilitá di una grande mole di dati relativi alle attivitá umane.
Una tipologia di dati che viene comunemente analizzata é messa a disposizione dalla telefonia mobile. Uno dei pregi dei dati di telefonia mobile é che sono ottenuti da uno strumento molto diffuso tra la popolazione e quindi forniscono una descrizione molto particolareggiata del fenomeno in analisi.
Le compagnie telefoniche raccolgono un grande numero di dati non solo sulla chiamata in sé ma anche sulla posizione spaziale e temporale dei loro utenti e ció permette diversi tipi di analisi. É possibile ad esempio studiare lo schema di attivitá di un certo individuo o di un certo gruppo, oppure studiare la mobilitá all'interno della cittá.
I dati in nostro possesso riguardano l'attivitá telefonica (intesa come numero di chiamate e messaggi in entrata e in uscita) della cittá di Marsiglia in un periodo di sei giorni. L'obiettivo di questo studio é di stabilire se l'attivitá telefonica della rete obbedisce a leggi statistiche e verificare se sono presenti delle fluttuazioni rispetto a queste ultime. Nell'analisi verrá anche controllata la compatibilitá con il modello della *Human Dynamics* proposto da Barabasi[1].
Nel primo capitolo vengono illustrati due modelli che si contrappongono nell'interpretazione della *Human Dynamics*. Il primo prevede attivitá umane descritte da processi di Poisson, ammettendo quindi la Markovianitá del sistema. Il secondo prevede leggi a potenza che rispecchiano l'adozione di protocolli di selezione e quindi introduce una memoria. Vengono inoltre descritti i risultati dell'analisi di Jiang [5], da cui emerge che, anche se il campione totale segue una legge a potenza, il singolo individuo potrebbe seguire una distribuzione diversa. Nel secondo capitolo vengono discussi i diversi *steps* che sono stati seguiti nell'analisi. In particolare abbiamo dovuto adottare delle strategie per ridurre il rumore presente nel sistema e poi creare dei clusters per rinforzare il segnale. Infine abbiamo ricavato la distribuzione delle fluttuazioni ed eseguito un confronto con il modello.

Capitolo 1

Modelli per la *Human Dynamics*

1.1 *Poisson Statistics*

Il modello che sembra fornire la descrizione migliore di molte attivitá umane é quello dei processi di Poisson. Questi processi stocastici vengono introdotti per tener conto di tutte le variabili ignorate.

Consideriamo un processo di Poisson di eventi puntuali e sia $N(t, t+\delta t)$ il numero di eventi avvenuti in $(t, t+\delta t]$. Valgono allora le seguenti proprietá per $\delta t \to 0$:

$$prob(N(t, t+\delta t) = 0) = 1 - p \cdot \delta t + o(\delta t) \tag{1.1}$$

$$prob(N(t, t+\delta t) = 1) = p \cdot \delta t + o(\delta t) \tag{1.2}$$

dove p é una costante positiva; inoltre $N(t, t+\delta t)$ risulta completamente indipendente dagli eventi in $(0, t]$.

I processi di Poisson appartengono quindi alla categoria dei processi di Markov continui nel tempo.

Sia $t_0 + Z$ il momento in cui avviene il primo evento a partire da un certo t_0 fissato. Indicata la probabilitá che questo evento accada dopo un tempo Z con $P(x) = P(Z > x)$ allora:

$$P(x + \delta x) = P(x) \cdot prob(N(t_0 + x, t_0 + x + \delta x) = 0) \tag{1.3}$$

$$P(x + \delta x) = P(x) \cdot (1 - p \cdot \delta x + o(\delta t)) \tag{1.4}$$

da cui:

$$P'(x) = -p \cdot P(x) \tag{1.5}$$

Osservando che P(0)=1 si ottiene:

$$P(x) = e^{-p \cdot x} \tag{1.6}$$

La probabilitá dell' *interevents time*, ossia dell'intervallo temporale tra due eventi successivi, segue una legge di decadimento esponenziale.

Nel nostro modello $N(t, t+\delta t)$ va interpretato come il numero di volte che una persona svolge una determinata attivitá (di cui viene trascurata la durata) nell'intervallo δt. Le proprietá del processo di Poisson equivalgono quindi a considerare che l'uomo svolga le sue attivitá ad un tasso costante (vedi 1.2) e in modo indipendente dagli eventi precedenti (ipotesi Markoviana). La distribuzione di probabilitá risultante ci dice che lunghi *interevents times* sono poco probabili.

1.2 *Non-Poisson Statistics*

Studi piú recenti propongono una *human dynamics* governata da leggi a potenza:

$$P(x) \sim x^{-\alpha} \tag{1.7}$$

Le distribuzioni di questo genere tendono a zero piú lentamente dell'esponenziale, per $x \to \infty$, e la probabilitá di lunghi *interevents times* non é trascurabile come nel caso della distribuzione di Poisson. In questo contesto Barabasi [1] parla di *bursts*, ossia momenti in cui un lungo *interevents time* viene seguito da molti altri piú brevi.

Nell'articolo di Barabasi vengono anche ipotizzati diversi modelli per giustificare alcuni valori dell'esponente α della legge a potenza. Essi assumono che il processo sia regolato da un protocollo di selezione in cui vengono aggiunti nuovi compiti ad intervalli di tempo che seguono una distribuzione di Poisson. Questi compiti hanno una prioritá fornita da una distribuzione di probabilitá generica e hanno la stessa durata. Le assunzioni di questo modello fanno quindi perdere la proprietá di Markov dal momento che il protocollo costituisce la memoria del processo. A seconda del protocollo seguito, vengono distinti diversi comportamenti:

1. first-in-first-out (FIFO) per cui i compiti vengono eseguiti nell'ordine in cui sono aggiunti alla lista;

2. i compiti vengono eseguiti in ordine casuale;

3. i compiti vengono eseguiti secondo un'ordine di prioritá (questo protocollo puó essere raffinato considerando che ci sia anche una probabilitá non nulla di eseguire un compito in modo casuale);

Ai primi due protocolli corrisponde una distribuzione di Poisson. L'ultimo protocollo é quello in cui Barabasi trova la giustificazione di una legge a potenza. In base al rapporto tra la velocitá di esecuzione dei compiti e la velocitá con cui nuovi compiti si aggiungono alla lista, vengono stabiliti diversi regimi a cui corrispondono diversi valori per α:

1. sottocritico: vengono aggiunti nuovi compiti ad un tasso minore rispetto a quello con cui sono eseguiti; di conseguenza la lista sará spesso vuota; a questo regime corrisponde una distribuzione di Poisson;

2. critico: vengono aggiunti nuovi compiti allo stesso tasso con cui sono eseguiti; la lunghezza della lista segue quindi un random walk, che porta ad una distribuzione con $\alpha = 1.5$;

3. supercritico: vengono aggiunti nuovi compiti ad un tasso maggiore rispetto a quello con cui sono eseguiti; porta allo stesso esponente del caso precedente: $\alpha = 1.5$;

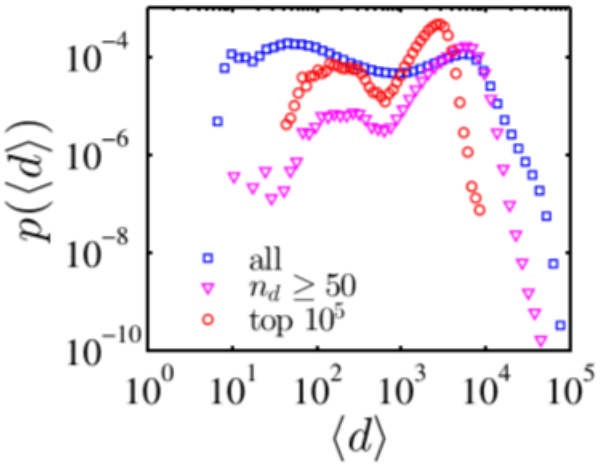

Figura 1.1: Probabilitá dell'*interevents time* medio (l'*interevents time* viene indicato con d, duration); la curva in blu indica la distribuzione ottenuta considerando il campione intero; quella viola é riferita agli utenti con attivitá superiore a 50; quella rossa é riferita gruppo costituito dai 10^5 utenti piú attivi; si distinguono due picchi che corrispondo a gruppi con durata media di chiamata molto differente

Fin'ora si é considerata una lunghezza variabile della lista. In caso contrario si possono effettuare simulazioni che mostrano che la distribuzione é una legge a potenza con $\alpha = 1.0$. L'assunzione di una lunghezza fissa sembra descrivere meglio il comportamento umano dato che la memoria a breve termine di una persona puó contenere un numero limitato di attivitá .

Quest'ultimo é proprio l'esponente trovato da Barabasi nel caso di:

1. Email (*interevent time* = tempo tra due email successive inviate dallo stesso utente)
2. *Web browsing* (*interevent time* = tempo tra due *page download* consecutivi, ossia due *clicks*, dello stesso visitatore)
3. Prestiti in biblioteca (*interevent time* = tempo tra due prestiti successivi dello stesso cliente)

per i singoli individui. Le piccole deviazioni dal valore $\alpha = 1.0$ vengono spiegate dall'intervallo temporale finito e dalla variabilitá dell'attivita' degli utenti.

Pur avendo elaborati questi modelli, Barabasi non esclude la possibilitá di altri modelli che giustifichino altri esponenti.

Nell'articolo di Jiang [5] sono analizzate le chiamate telefoniche e viene utilizzata una legge a potenza solo per descrivere il comportamento collettivo, mentre il comportamento dei singoli sembra meglio approssimato da una distribuzione di Weibull

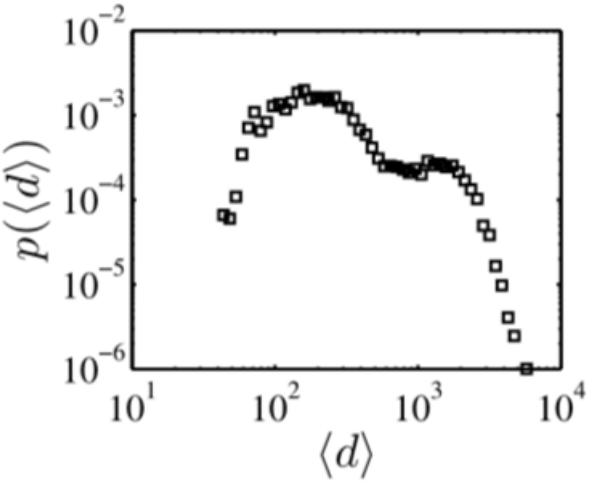

Figura 1.2: Probabilitá dell'*interevents time* medio per un campione che segue la distribuzione di Weibull

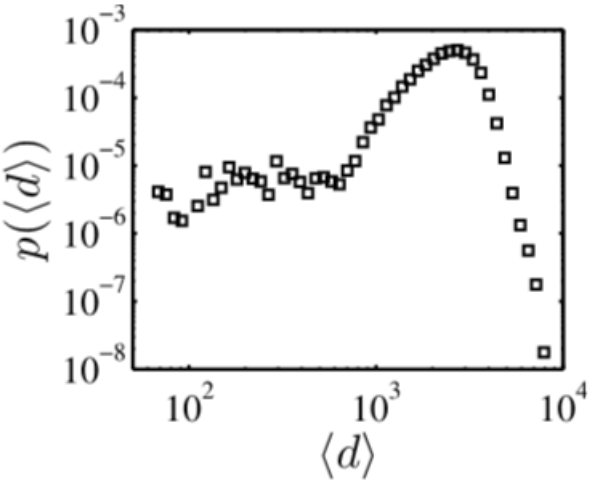

Figura 1.3: Probabilitá dell'*interevents time* medio per un campione che segue una legge a potenza

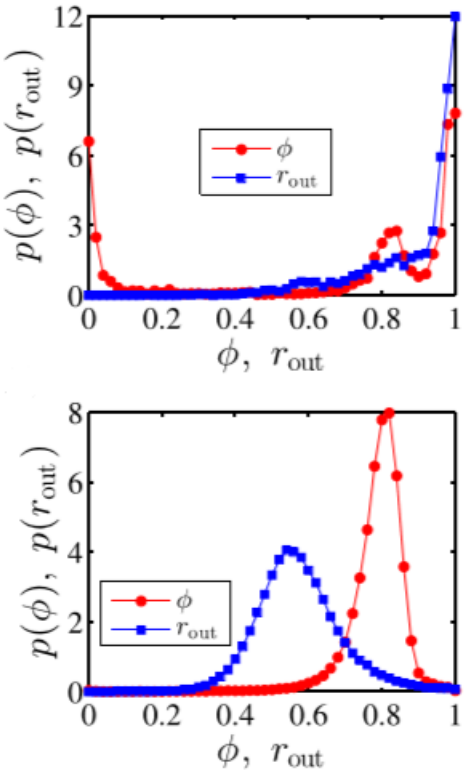

Figura 1.4: ϕ rappresenta la suddivisione delle chiamate totale tra i diversi numeri chiamati; un alto valore di ϕ indica chiamate in un'ampia cerchia di persone, mentre un valore nullo indica che tutte le chiamate sono indirizzate ad un unico numero; r_{out} rappresenta la percentuale di chiamate in uscita. Confronto fra un campione che segue una legge a potenza (in alto) e uno che segue una distribuzione di Weibull (in basso)

($p(d) = \alpha\beta d^{\beta-1} \cdot e^{-\alpha \cdot d^{\beta}}$). Jiang trova che nel gruppo selezionato dei 10^5 utenti piú attivi il 3.46% segue una legge a potenza, mentre il 73.34% segue una distribuzione di Weibull (la rimanente percentuale non ha un comportamento ben definito). La distribuzione del campione totale é invece una legge a potenza con $\alpha = 0.942$. Dall'analisi emerge anche che all'interno della popolazione sono presenti due gruppi con *interevents times* medi molti differenti (un lungo *interevents time* corrisponde ad una persona che utilizza poco il cellulare). Il picco di *interevents times* medi piú brevi viene ritrovato nel gruppo che segue una legge a potenza, mentre il picco di *interevents times* medi piú lunghi viene riconosciuto nel gruppo che segue la distribuzione di Weibull. Analizzando le chiamate fatte all'interno dei diversi gruppi emerge inoltre che il gruppo della legge a potenza ha un comportamento piú estremo. Esso é infatti caratterizzato da un'alta probabilitá di avere solo chiamate in uscita, le quali risultano spesso distribuite uniformemente tra i numeri chiamati oppure nettamente indirizzate verso un'unico numero. Jiang ipotizza quindi che le caratteristiche del gruppo della legge a potenza descrivano vendite telefoniche (quando le chiamate sono distribuite uniformemente) e *robot-based users* (quando le chiamate sono concentrate su un numero).

Capitolo 2

Analisi dati

2.1 *Mobile phone data*

Lo studio della dinamica delle chiamate telefoniche é guidato dalle applicazioni che puó avere nella gestione e ottimizzazione del servizio da parte della compagnia telefonica, ma anche dallo studio delle dinamiche della popolazione.

Dai dati telefonici (intesi come numero di cellulari collegati ad una determinata antenna ad un certo momento) si possono ricavare informazioni riguardo [4]:

1. la dinamica della densitá della popolazione: dal numero di cellulari collegati si puó infatti stimare il numero di persone presenti nella regione assegnata ad una specifica antenna; l'analisi puó essere effettuata sia su lunghi periodi (come nel caso della densitá di abitanti) sia su brevi (migrazioni interne ad una cittá durante la giornata).

2. il turismo: i cellulari stranieri possono essere individuati grazie al fenomeno del *roaming* e quindi i loro dati possono essere isolati e studiati separatamente.

3. l'impatto di fenomeni speciali: questi possono corrispondere ad un comportamento anomalo di una rete o di una singola antenna; viene riconosciuto per esempio l'avvenimento di un concerto o di un disastro naturale.

4. la classificazione delle attivitá dominanti di una regione (si possono definire regioni residenziali, commerciali, lavorative): viene analizzata la corrispondenza del profilo di attivitá dell'antenna con profili che corrispondono alla prevalenza di una attivitá specifica.

I dati di telefonia mobile sono particolarmente studiati dal momento che permettono di risolvere in parte il problema della scarsitá di dati, dato che la quasi totalitá della popolazione in certe regioni del mondo accede a questo servizio (anche se rimane l'ostacolo della *privacy*). Inoltre questa tipologia di dati si pu avvalere di un'adeguata precisione spaziale e temporale.

2.2 Descrizione dell'analisi

I dati di cui siamo in possesso ci riportano in forma anonima l'attivitá delle antenne telefoniche dell'area di Marsiglia nei giorni 19,20,21 ottobre e 29,30,31 marzo (entrambe le serie si riferiscono alla successione sabato, domenica, lunedí). L'assenza di un codice identificativo del cellulare ci impedisce di avere una corrispondenza tra l'attivitá dell'antenna e una determinata persona, e quindi non possiamo cercare informazioni sulla mobilitá.

Le informazioni di cui disponiamo riguardano il numero di utenti collegati ad una determinata antenna telefonica e la posizione spaziale (coordinate della torre GSM) e temporale (istante a cui é stata osservata l'attivitá) di questo dato. Piú precisamente con la parola *attivitá* faremo riferimento d'ora in poi al numero di chiamate e messaggi, inviati e ricevuti, sommato su tutte le antenne localizzate sulla stessa torre GSM (vedi appendice).

La nostra analisi si focalizza sulla ricerca di legge statistiche che governano il sistema, ma potremo controllare anche la presenza di anomalie. La nostra scelta é motivata dalla possibile presenza di comportamenti anomali connessi alle elezioni municipali del 30 marzo. La ricerca di anomalie puó avere sia applicazioni pratiche (per esempio in una compagnia telefonica) sia puramente scientifiche, facendo emergere nuovi dettagli nella descrizione dell'attivitá telefonica.

Durante l'elaborazione dei dati abbiamo dovuto creare un algoritmo di *Clustering* per ricercare elementi che potessero essere considerati simili ai fini della nostra analisi e quindi raggruppati. Dato che i nostri risultati sono fortemente dipendenti dalla decisione di questo algoritmo, quest'ultimo ha acquistato una rilevanza notevole. Il secondo elemento che caratterizza l'analisi é la definizione di fluttuazione (che chiameremo anche segnale), dal momento che indica l'interpretazione dei dati finali ottenuti.

Il metodo di analisi a cui ci siamo attenuti é riassunto nei seguenti punti:

1. Organizzazione dei dati
   (a) Identificazione dell'attivitá associata ad ogni antenna (vedi appendice)
   (b) Individuazione ed eliminazione delle antenne con bassa attivitá
   (c) Definizione dei profili di attivitá
2. Individuazione del segnale
   (a) Definizione del segnale
   (b) *Clustering*
   (c) Calcolo del segnale
3. Studio del segnale

(a) Distribuzione dei valori del segnale
(b) Descrizione e confronto del *fitting* dei dati
(c) Indicazione delle anomalie

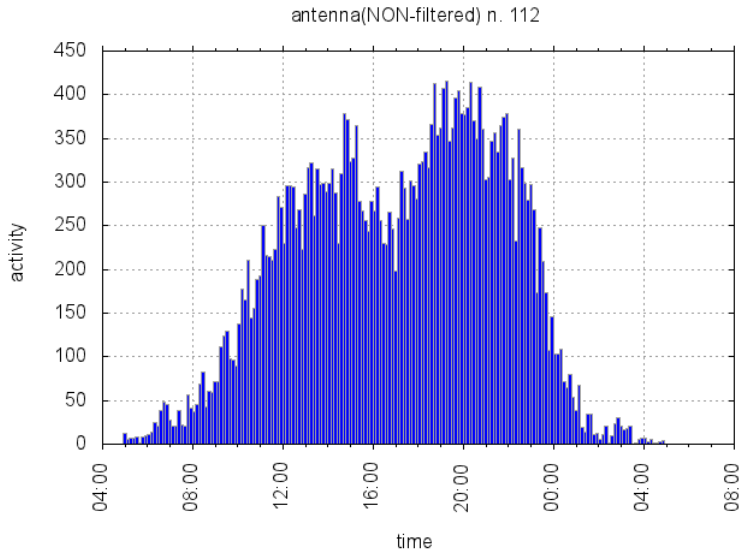

Figura 2.1: Esempio di Profilo

2.3 Individuazione ed eliminazione delle antenne con bassa attivitá

Selezioniamo le antenne piú attive di ogni giorno sulla base dei grafici di distribuzione dell'attivitá giornaliera tra le diverse antenne (vedi figura 2.3). La selezione é necessaria dal momento che queste antenne sono maggiormente influenzate dal rumore e quindi hanno un profilo di attivitá meno definito che le rende difficili da raggruppare (*Clustering*). Dal momento che sono proprio le antenne meno attive quelle che vogliamo unire per migliorare il segnale, questo problema non puó essere trascurato. Abbiamo mostrato subito un esempio di una delle strategie che saremo obbligati ad attuare per risolvere il nostro principale ostacolo, ossia separare il rumore dal segnale.

Nel grafico abbiamo distribuito le antenne in base alla loro attivitá totale.

Ogni antenna appartenente ad un bin ($x - \frac{\delta x}{2} < total\ activity < x + \frac{\delta x}{2}$) fornisce a quel bin un contributo sulle ordinate pari alla sua attivitá totale e quindi l'area del grafico corrisponde all'attivitá totale della rete.

Per decidere la soglia di attivitá sotto cui eliminare le antenne, valutiamo i profili di attivitá che si ottengono eseguendo il taglio in punti differenti. Ponendo la soglia pari a zero, e quindi senza effettuare il taglio, otteniamo che le antenne con attivitá minore hanno profili che non possono essere confrontati con le altre (figura 2.2).

Aumentiamo quindi la soglia fino al primo minimo che incontriamo nel grafico (vedi figura 2.3 e successive). In questo modo otteniamo un buon rapporto tra attivitá totale persa e numero di antenne poco attive eliminate (siamo interessati a mantenere se

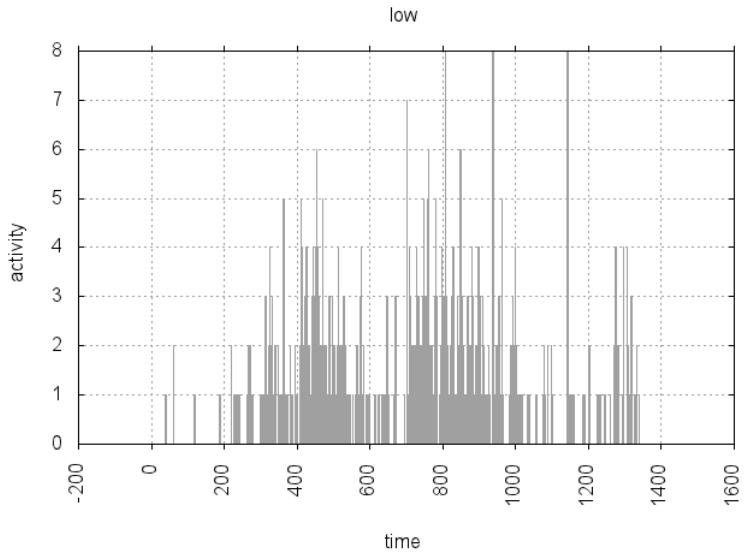

Figura 2.2: Esempio di profilo eliminato (attivitá totale pari a 754); in ascissa il tempo é indicato in minuti

possibile una bassa percentuale di attivitá persa dato che essa rappresenta una delle caratteristiche con cui si valuta un metodo di elaborazione dati). L'esistenza di un minimo locale viene ipotizzata considerando che la rete sia stata costruita in modo tale da limitare la presenza di antenne con bassa attivitá. Nei casi in cui non sia presente un minimo nella prima regione dell'asse x, il taglio viene effettuato eliminando una percentuale di attivitá comparabile con quella minore eliminata negli altri giorni.

Nel nostro caso l'attivita' persa e':

sabato 19 ottobre: attivitá totale 6883063, considerata 6015883.
domenica 20 ottobre: attivitá totale 5966103, considerata 5380616.
lunedí 21 ottobre: attivitá totale 7852054, considerata 7255054.

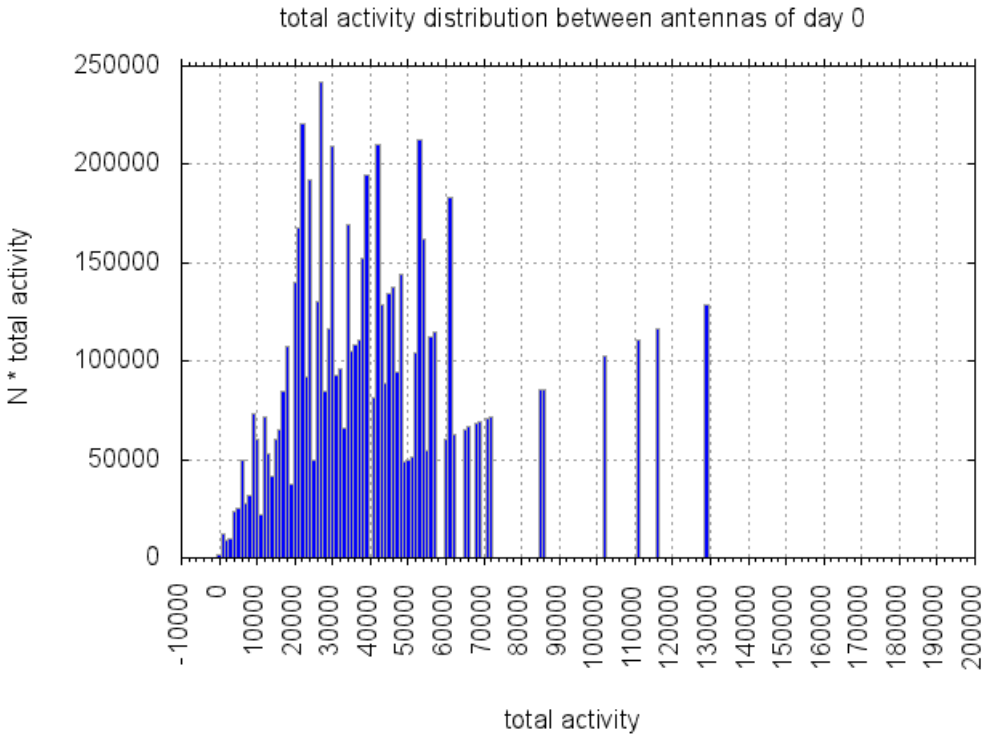

Figura 2.3: Sabato 19 ottobre, la soglia di attivitá é pari a 19000

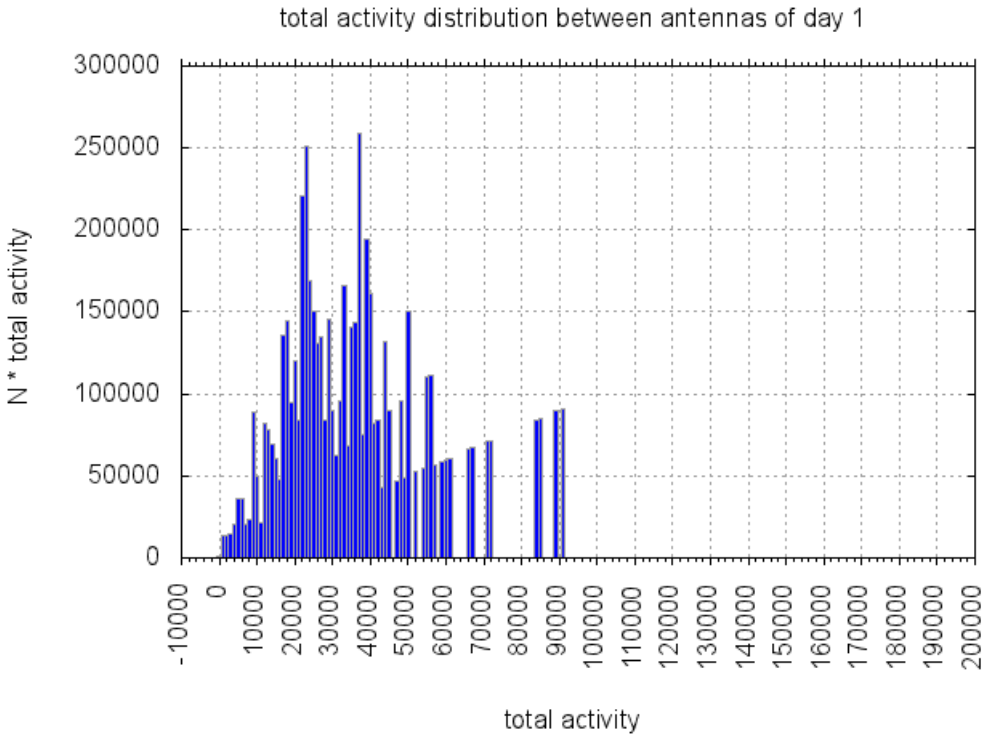

Figura 2.4: Domenica 20 ottobre, la soglia di attivitá é pari a 15000

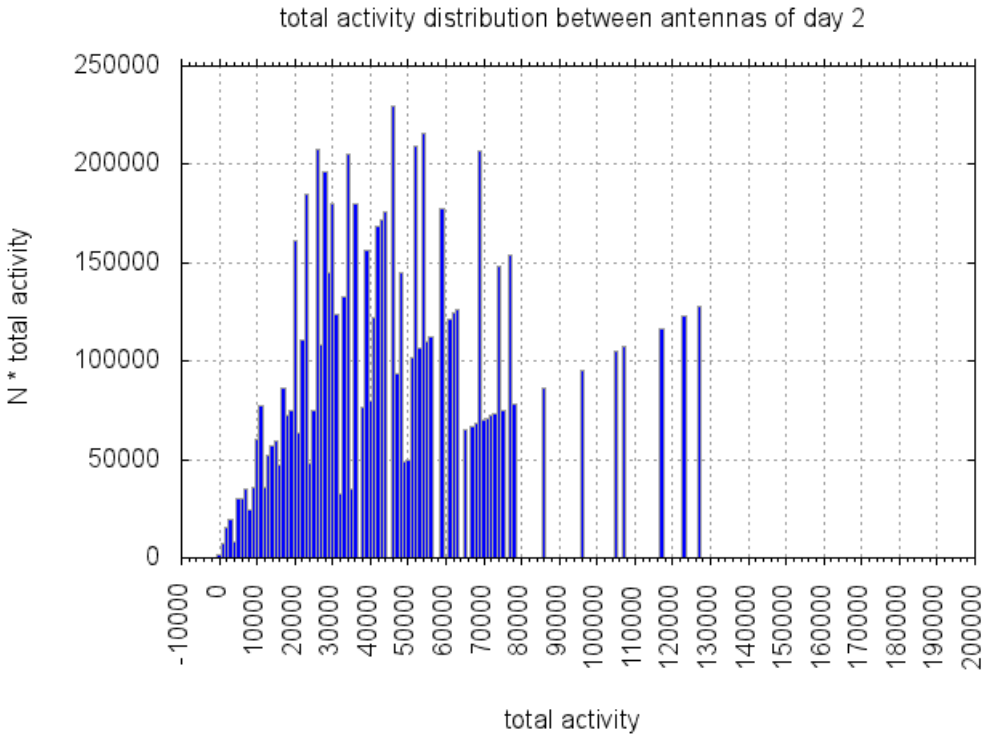

Figura 2.5: Lunedí 21 ottobre, la soglia di attivitá é pari a 16000

2.4 Definizione dei profili di attivitá

Ipotizziamo che l'intervallo di tempo piú breve su cui i dati ci permettono di descrivere l'attivita' delle antenne non sia inferiore al minuto e quindi definiamo l'unitá di misura della scala temporale dell'analisi pari a 1 min.

Se nel proseguire l'analisi avremo bisogno di una migliore risoluzione temporale abbiamo la possibilitá di raggiungere la risoluzione di 1 s.

Gli eventi relativi ad ogni antenna vengono raccolti in istogrammi di attivitá giornaliera. Il numero di bins di questo istogramma viene posto pari a:

$$n_{bins} = min(\sqrt{total\ activity}, 288) \tag{2.1}$$

in modo da imporre una lunghezza minima dei bins pari a 5 min. Viene usato questo criterio per permettere che in ogni bin, indipendentemente dall'attivitá totale dell'antenna, cada un numero sufficiente di eventi tale da attenuare il rumore presente nel segnale. In questo modo possiamo inoltre rappresentare i profili delle antenne meno attive in modo meno dettagliato (vedi figura: passaggio da bin=1 min a bin=10min; confronto con antenna attiva). Evidentemente questa perdita di risoluzione influirá sul confronto delle antenne nella fase della clusterizzazione.

L'inizio della giornata viene posto alle 5:00 am sulla base dei grafici dell'attivitá totale della rete (vedi Figura 2.6), da cui possiamo notare un taglio netto tra due profili appartenenti a giorni consecutivi proprio in corrispondenza di questo orario. Cosí facendo evitiamo di spezzare possibili eventi appartenenti alla vita notturna.

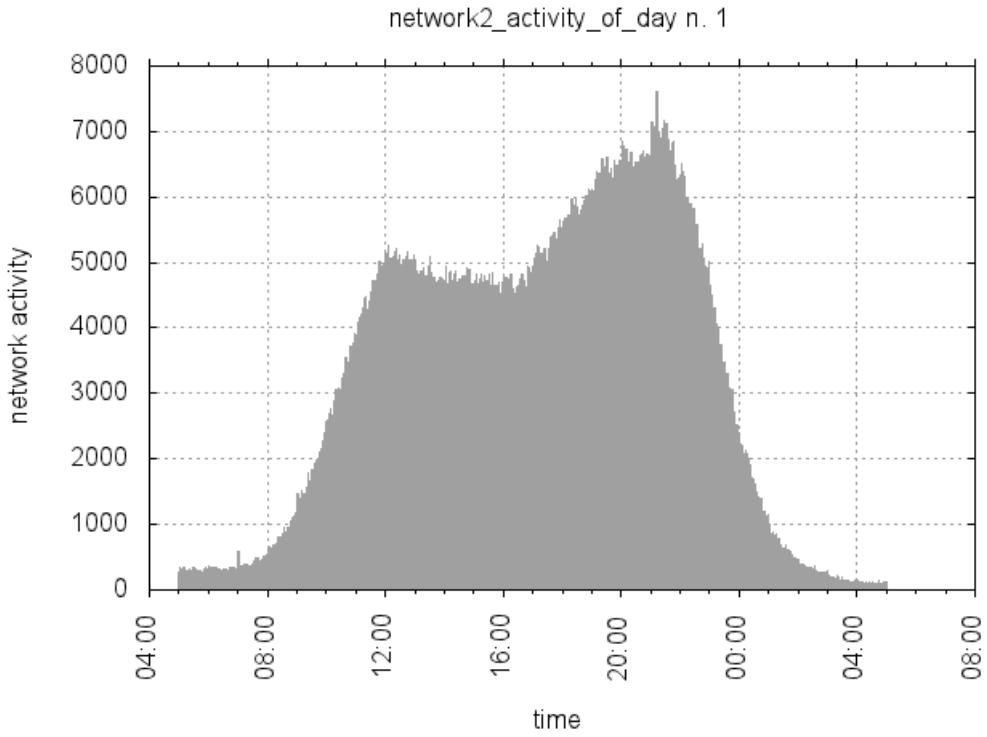

Figura 2.6: profilo di attivitá della rete domenica 20 ottobre

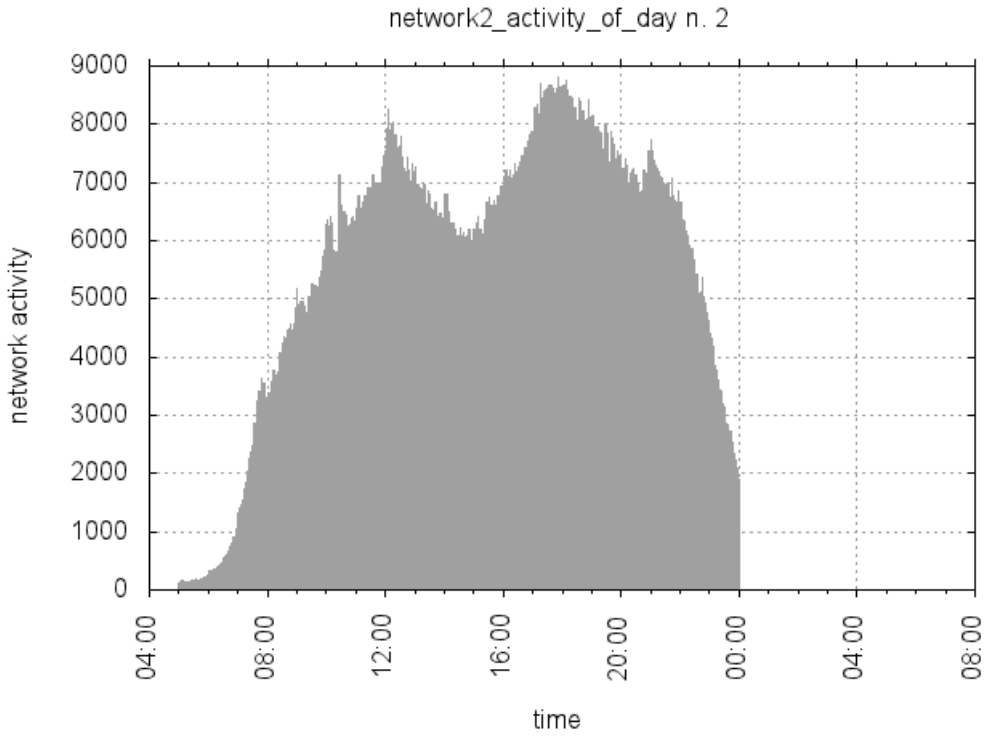

Figura 2.7: profilo di attivitá della rete lunedí 21 ottobre

2.5 Definizione del segnale

La nostra analisi é volta alla ricerca e allo studio delle fluttuazioni che possono essere presenti nell'attivitá di telefonia mobile. Le fluttuazioni vengono definite nel momento in cui si definisce il valore atteso, quest'ultimo dipende dal tipo di informazione che vogliamo estrarre. A livello teorico, per associare un evento straordinario ad una fluttuazione, dovremmo utilizzare un valore atteso che tenga conto di tutti gli altri eventi che stanno accadendo, tralasciando solo quello di cui vogliamo stabilire l'influenza. Questo modo di procedere comporterebbe senz'altro una conoscenza dettagliata dell'impatto di un determinato evento sull'attivitá telefonica, ma é altrettanto sicuro che comporterebbe un dispendio di energia superiore ad ogni altro metodo, a paritá di risultati ottenuti.

In questa analisi, invece, sacrifichiamo la conoscenza dettagliata dei meccanismi alla base dell'attivitá telefonica e costruiamo il nostro valore atteso direttamente dal valore osservato.

2.5.1 Media temporale

Tuttavia possiamo continuare a seguire il ragionamento precedente definendo il nostro valore atteso per il profilo di una determinata antenna come la media temporale del profilo della stessa antenna in giorni differenti. Assumiamo poi che le fluttuazioni si compensino nel calcolo della media a causa della loro natura stocastica.

Cosí facendo il nostro valore atteso dovrebbe proprio tener conto di tutti gli altri eventi che stanno accadendo, tralasciando solo quello di cui vogliamo stabilire l'influenza. Arriveremmo cosi allo stesso risultato dell'analisi piú faticosa perdendo peró la conoscenza dei meccanismi che hanno originato sia il profilo medio che la (eventuale) fluttuazione trovata. Questi ultimi andrebbero ritrovati con altri tipi di analisi.

Il numero di giorni che possiamo analizzare é per sia limitato temporalmente sia fortemente eterogeneo. Disponiamo infatti di soli sei giorni incentrati nei finesettimana. Gran parte dei nostri dati si riferisce quindi a momenti in cui le persone non svolgono attivitá lavorativa e si attengono a schemi meno regolari. Dai profili di attivitá emerge infatti che la maggior parte delle antenne ha caratteristiche differenti nei diversi giorni (vedi figura 2.8).

Una differenza che viene individuata in certi casi é che i profili sono rapportati all'attivitá totale giornaliera (domenica<sabato<lunedí normalmente). Per eliminare questa discrepanza possiamo utilizzare l'attivitá percentuale dell'antenna rispetto alla attivitá della rete. Un'altra differenza é costituita dalla posizione della curva del profilo sull'asse temporale.

Osserviamo che il profilo della domenica tende ad essere traslato verso tempi maggiori rispetto al sabato e ancor di piú rispetto alla domenica.

I differenti profili tendono peró a convergere verso sera e quindi viene resa impossibile una traslazione del profilo per eliminare questa seconda discrepanza. Infine, anche i profili

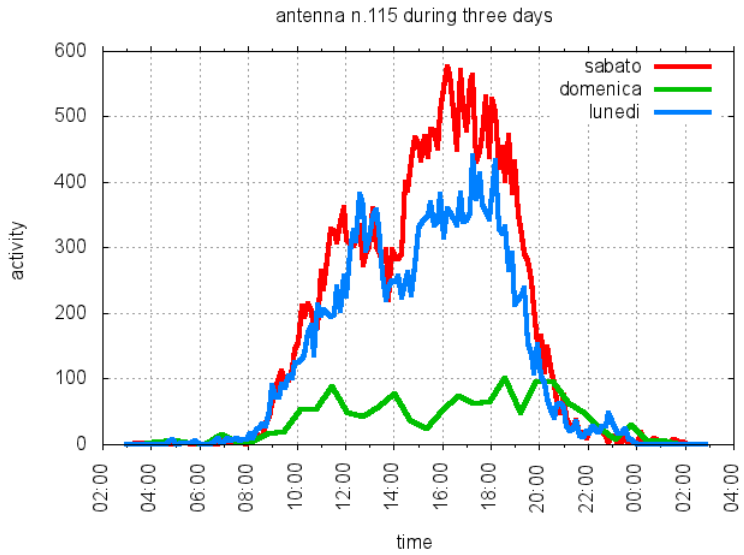

Figura 2.8: Diverso fattore di scala

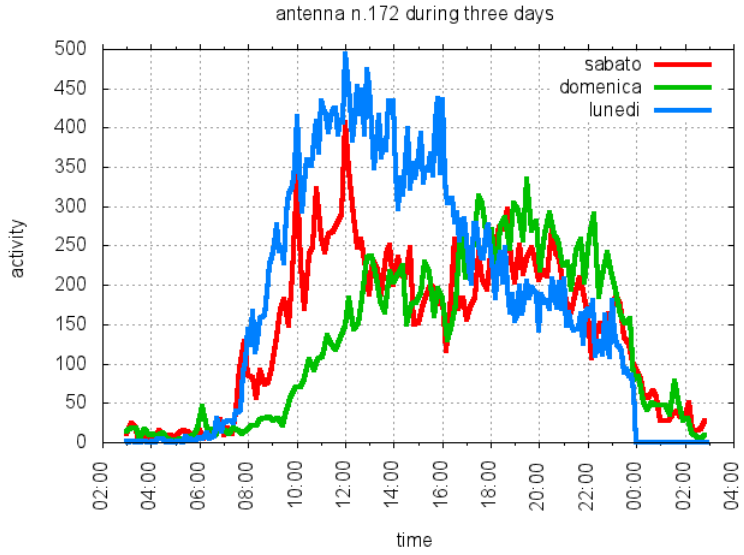

Figura 2.9: Profili incompatibili

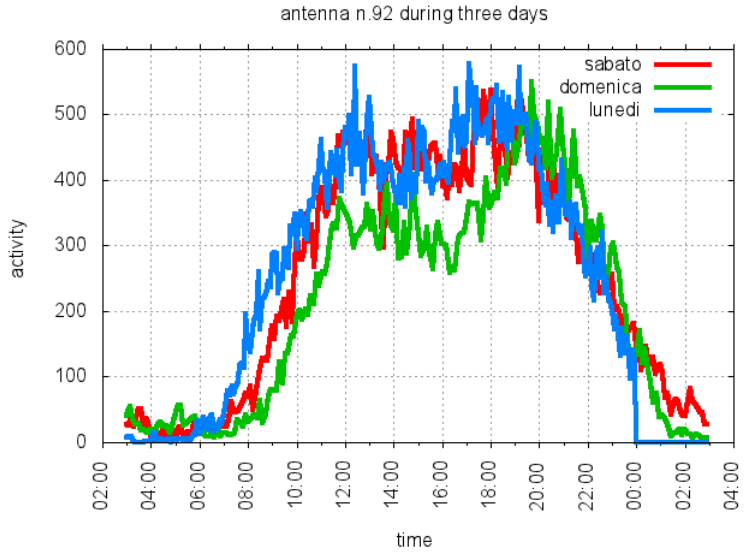

Figura 2.10: Traslazione del profilo

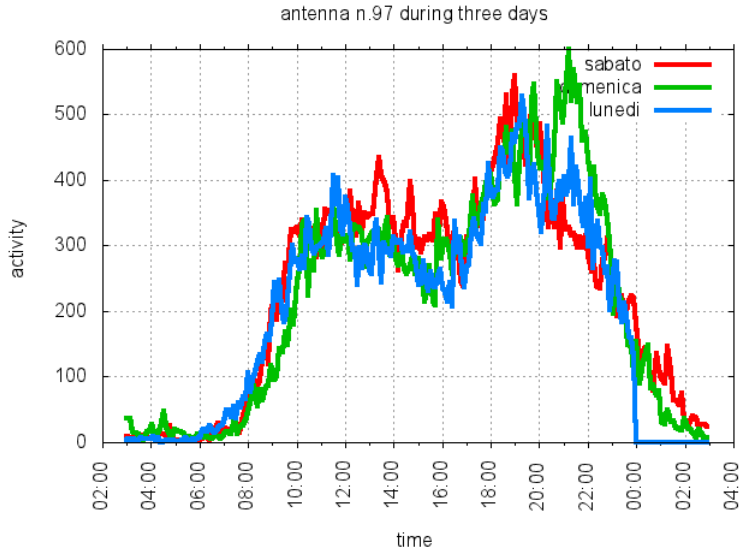

Figura 2.11: Profili con incongruenze non trascurabili

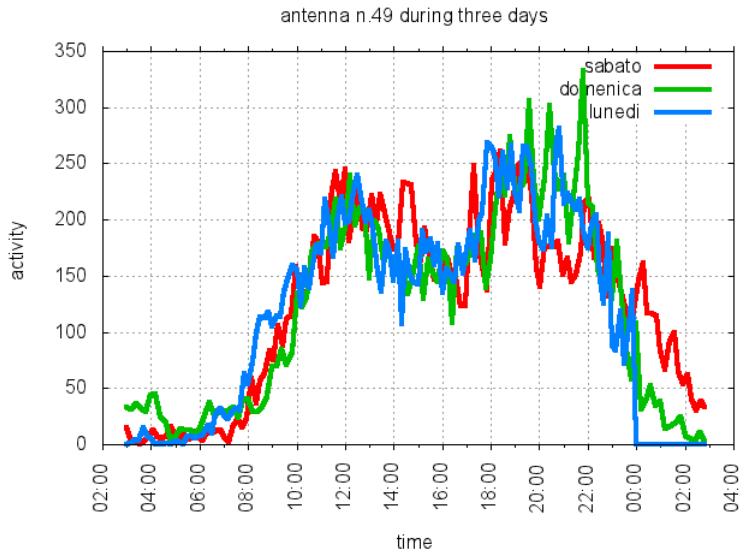

Figura 2.12: Profili con incongruenze non trascurabili

che si differenziano meno, mostrano delle inevitabili differenze di forma. In particolare notiamo un secondo picco di attivitá (oppure un picco di estensione temporale maggiore) la domenica sera verso le 21.

Da quanto detto, considerato che ogni correzione andrebbe adattata nello specifico ad ogni singola antenna (fattore di scala, traslazione temporale, individuazione di forme incompatibili), é chiaro che anche nel caso in cui otteniamo dei profili comparabili, é difficile che siano in numero sufficiente a determinare una media in cui le fluttuazioni si siano compensate.

2.5.2 Media spaziale

Nell'attesa di ulteriori dati decidiamo quindi di cambiare strategia ed eseguire una media spaziale. Ipotizziamo che gruppi di antenne vicine possono essere considerati diverse realizzazioni dello stesso processo stocastico.

Considerata la prima antenna del cluster, tutte le antenne in un certo raggio fissato R sono considerate vicine. Nel caso di zone centrali $R = 2\,km$, nelle regioni periferiche $R = 4\,km$. A causa dell'eterogeneitá che possiamo trovare anche tra i profili di antenne vicine (che corrisponde a gruppi eterogenei di persone) controlliamo che i profili non si discostino in maniera eccessiva. La strategia adottata consiste nel calcolare la norma L2 tra i diversi profili e confrontarla con la discrepanza attesa pari ad N deviazioni standard. Al variare del parametro N non viene trovato un equilibrio tra numero di cluster e somiglianza effettiva dei profili.

Maggiore é infatti il numero di elementi del cluster, maggiore sará la probabilitá che nel valore medio non siano presenti fluttuazioni. Profili troppo differenti portano d'altra parte ad un valore medio che non rappresenta piú la singola antenna e quindi crea un segnale che comprende anche porzioni del profilo originale (sostanzialmente non previste).

Dobbiamo concludere anche questa seconda strategia non consente di trovare un valore atteso per il profilo di attivitá delle antenne a causa delle persistenti disomogeneitá tra i profili.

2.5.3 Applicazione di un filtro

Il metodo che riteniamo piú efficiente per individuare un'attivitá straordinaria, che si possa definire tale anche a livello sociale, consiste nel calcolare un valore atteso che elimini dal valore osservato le variazioni troppo rapide di attivitá. Le nostre fluttuazioni vengono cosí a identificare un aumento o un calo improvviso di chiamate e SMS in entrata e in uscita. Riteniamo infatti che questa sia una semplice e chiara indicazione di quello che comunemente parlando si intende per segnale.

L'applicazione del filtro consiste, per ogni bin, nel calcolare l'attivitá media in un intervallo temporale di un'ora centrato nel bin e imporre questo valore al bin. I bin con cui operiamo sono in questo caso i bin originari di lunghezza pari ad un minuto. Questo filtro é stato fatto partire dalle 5:30 e spostato verso i bin temporalmente successivi, fino alle 4:30 del giorno successivo. É stato applicato un filtro rettangolare (con pesi tutti uguali) perché permette di eliminare in modo netto eventuali segnali (che non devono essere presenti nel valore atteso) rispetto ad un filtro triangolare (con pesi decrescenti in modo simmetrico a partire dal bin in considerazione fino agli estremi dell'intervallo).

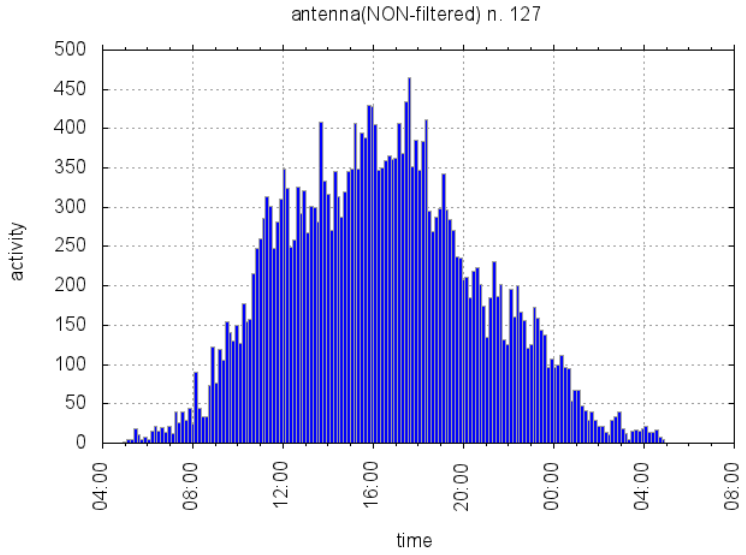

Figura 2.13: Esempio di profilo senza l'applicazione del filtro

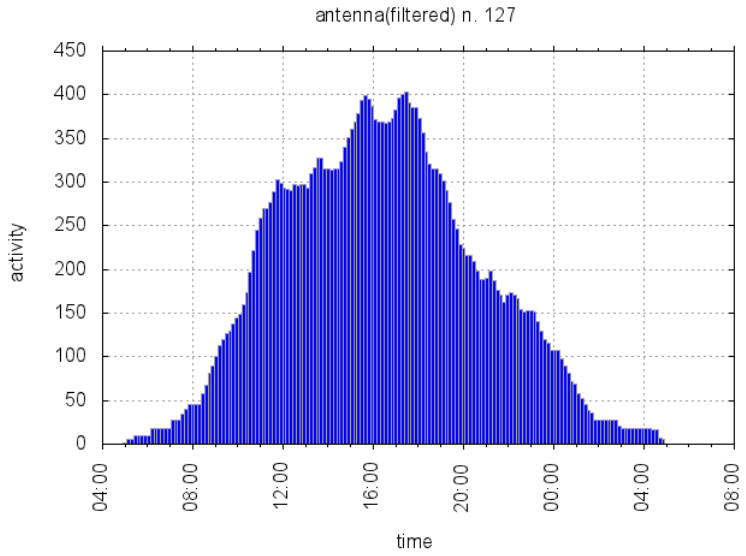

Figura 2.14: Esempio di profilo con l'applicazione del filtro (da riferire alla figura precedente)

2.6 *Clustering*

2.6.1 Definizione di Cluster

Anche se l'applicazione di un filtro ci permette di individuare un profilo atteso osserviamo che il segnale rimane in certi casi abbastanza debole. Riprendendo il discorso della media spaziale, decidiamo di cercare antenne simili e vicine e di unire i loro segnali. L'assunzione che viene fatta in questo caso non é peró che le antenne siano diverse realizzazioni dello stesso processo, ma é che raccolgano dati relativi ad un unico profilo, frammentandolo.

Questo fenomeno é particolarmente rilevante nel caso in cui la densitá delle antenne sia alta: suddividere lo stesso segnale tra piú antenne vicine ha il solo effetto di indebolire l'applicabilitá degli strumenti statistici.

Per chiarezza poniamo l'accento sul fatto che nel ragionamento del paragrafo precedente abbiamo considerato un unico segnale suddiviso. Bisogna infatti considerare anche l'eventualitá per cui antenne vicine delimitino effettivamente realtá sociali differenti a cui corrispondono segnali qualitativamente differenti, come giá osservato nella strategia precedente.

Inoltre, sempre sull'esempio dei tentativi fatti, imponiamo un raggio R di grandezza massima del cluster. A differenza dei casi precedenti, il clustering ha il solo obiettivo di migliorare la sensibilitá dei nostri risultati e puó considerarsi errato solo nel caso in cui non riesca a rinforzare un segnale esistente. Questo avviene nel caso in cui siano presenti all'interno del cluster due antenne con segnali di segno opposto nello stesso momento. Tale fenomeno non dovrebbe accadere nel caso in cui le antenne condividano uno stesso segnale. Occorrerebbe in ogni caso un'analisi differente per controllare che ció non sia avvenuto.

2.6.2 Algoritmo di *Clustering*

Nella creazione dei cluster teniamo conto dei seguenti fenomeni:

1. i profili di attivitá di antenne corrispondenti allo stesso segnale differiscono in proporzione all'attivitá totale di ciascuna antenna, occorre quindi normalizzare il profilo di ogni antenna in modo da riconoscere lo stesso profilo su scale diverse

2. ogni elemento del cluster deve essere simile ad ogni altro all'interno dello stesso cluster. Questo ci assicura che le piccole differenze tollerate in un confronto non si propaghino fino ad arrivare a profili incompatibili con alcune antenne del cluster.

3. a partire da una antenna di base viene fissato un raggio di ricerca di antenne simili che definisce la grandezza massima possibile del cluster. Questo criterio rappresenta la nostra fiducia nel fatto che probabilmente un segnale simile sará condiviso da antenne vicine geograficamente. Non abbiamo la certezza che questo procedimento

ci permetta di individuare tutte le antenne con profilo simile, ma ci non é importante finché i cluster ottenuti hanno un'attivitá totale significativa. Nel caso in cui analisi piú dettagliate rendano evidente che due cluster che abbiamo considerato distinti vadano in realtá uniti, non si avrebbero conseguenze sui nostri risultati. Uno dei motivi per cui non ci interessiamo di questo fenomeno é che un'analisi con un tale obiettivo risulterebbe complessa e pericolosa (portando alla considerazione di problematiche pi sottili, come ad esempio il punto 5, nella definizione di cluster) e ci obbligherebbe ad analizzare ogni cluster singolarmente.

4. Le grandezze massime dei cluster sono, come prima, $R = 2\,km$ nelle zone centrali, e $R = 4\,km$ nelle regioni periferiche. Queste distanze sono state stimate sulla base della densitá delle antenne in modo da assicurarci che vengano sempre confrontate un numero ragionevole di antenne. Viene adottato tale criterio perché si considera che un eventuale segnale comune venga suddiviso tra antenne adiacenti indipendentemente dalla distanza a cui si trovano. Questi ragionamenti sono inoltre compatibili con la mobilitá umana dato che $2\,km$ é una distanza ragionevole per lo spostamento della maggior parte delle persone in una zona densamente popolata e $4\,km$ lo é per la periferia (ovviamente ci possiamo permettere di non scendere in un'analisi pi sofisticata solo perché stiamo costruendo questo modello per cittá come Marsiglia, le quali possono essere schematizzate in modo semplice).

   L'aumento del raggio di ricerca riesce, in parte, a mantenere costante il numero totale di confronti; esso non é in ogni caso da interpretare come un peggioramento della qualitá del cluster dato che i confronti sono soggetti agli stessi parametri di tolleranza. L'uso di due diversi raggi serve unicamente a tener conto della scala del problema.

5. non viene tenuto conto della distribuzione spaziale delle antenne simili effettivamente unite all'interno di un cluster; in altre parole, nel caso peggiore, possono essere tollerati cluster costituiti da due antenne poste a una distanza pari alla metá della grandezza del cluster (distanza massima), le quali siano inoltre isolate in mezzo a molte altre antenne classificate come diverse.

Vengono inoltre adottate le seguenti tecniche per ottimizzare i risultati:

1. la scelta dell'antenna di base viene effettuata in ordine decrescente di attivitá totale, in modo iniziare l'algoritmo sulla base di un profilo con poco rumore;

2. tra tutte le antenne classificate simili che possono essere aggiunte al cluster, vengono aggiunte prima le piú simili; in questo modo il cluster rimane il pi omogeneo possibile al suo interno e i profili che hanno una distanza L2 vicina alla soglia sono aggiunti per ultimi e solo se sono simili a tutti quelli giá presenti nel cluster

2.6.3 Confronto dei profili

Decidiamo di seguire due metodi diversi per arrivare ad una descrizione pi completa del nostro gruppo di antenne. Allo stesso modo possiamo sfruttare metodi differenti per controllare che (alcune tra) le nostre ipotesi non influenzino in modo imprevisto i risultati.

1. il profilo dell'antenna meno attiva viene riportato (quasi) inalterato sull'istogramma dell'antenna piú attiva. Dal momento che l'antenna piú attiva ha un numero maggiore di bin, ci ritroviamo a dover decidere come dividere l'attivitá di un bin dell'antenna meno attiva sui numerosi bin dell'antenna piú attiva. Stabiliamo che il bin di lunghezza maggiore imponga il suo valore a tutti i bin contenuti al suo interno.

2. il profilo dell'antenna giá assegnata al cluster rimane invariato, mentre quello dell'altra antenna viene ricalcolato su un istogramma con un numero di bin dettato dall'antenna del cluster. L'obiettivo di questo tipo di confronto é stabilire in modo preciso un profilo per il cluster. Le antenne che si aggiungono ad esso infatti devono essere risultate compatibili con il profilo originario (ossia calcolato con il criterio della radice) di tutte le antenne del cluster.

Da qui in poi i due procedimenti seguono le stesse procedure:

1. ad ogni bin viene assegnato un errore pari a $\sqrt{(N \cdot \sigma_1)^2 + (N \cdot \sigma_2)^2}$.

   Nel nostro modello l'attivitá segue una distribuzione di Poisson e quindi l'errore sul singolo bin di ogni antenna viene stimato come un multiplo N della radice dell'attivitá presente nel bin.

2. viene confrontata la somma dei quadrati delle differenze tra i due istogrammi con la somma dei quadrati degli errori. Detta A_{ik} l'attivitá dell'antenna i nel bin numero k, se

$$\sum_{k=0}^{N_{bins}} \sqrt{(N \cdot \sigma_{1k})^2 + (N \cdot \sigma_{2k})^2} > \sum_{k=0}^{N_{bins}} (A_{ik} - A_{j_k}) \tag{2.2}$$

   le antenne vengono considerate compatibili. Nel caso sia verificata la compatibilitá con tutti gli elementi del cluster, l'antenna viene aggiunta al cluster; in caso contrario l'antenna viene esclusa dal cluster in modo definitivo (dato che non ha la possibilitá di aggiungersi al cluster neanche in seguito ad un confronto futuro con un altro elemento del cluster).

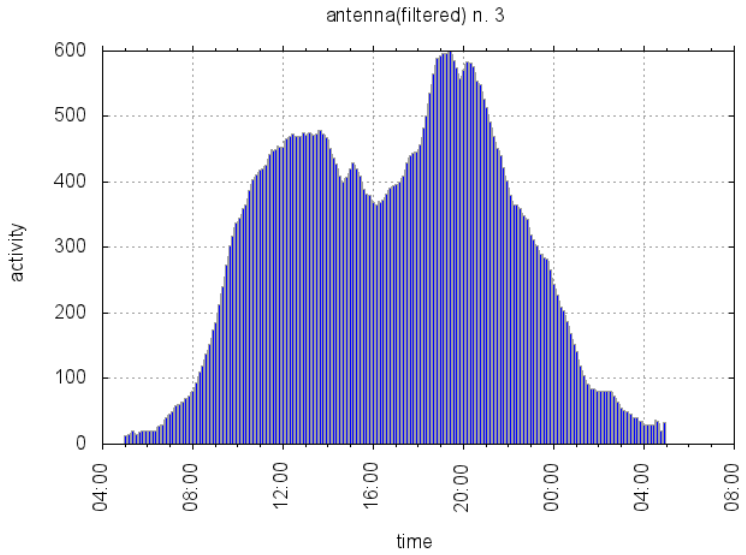

Figura 2.15: Elemento n.0 del cluster n.4

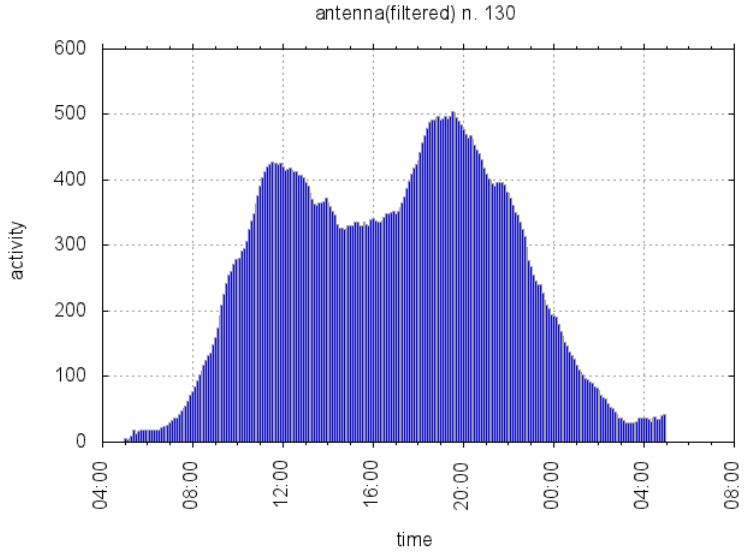

Figura 2.16: Elemento n.1 del cluster n.4

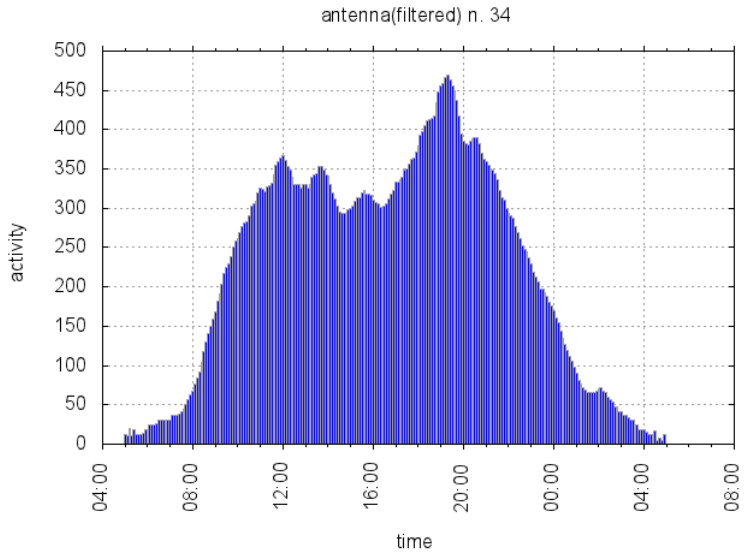

Figura 2.17: Elemento n.2 del cluster n.4

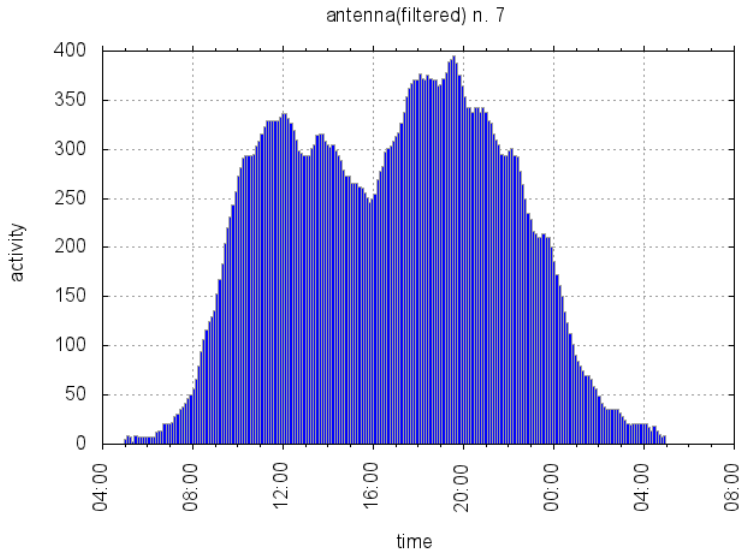

Figura 2.18: Elemento n.3 del cluster n.4

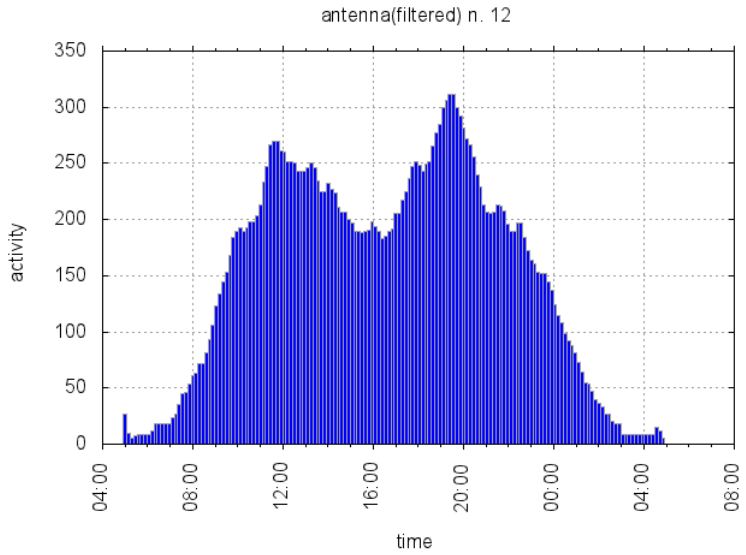

Figura 2.19: Elemento n.4 del cluster n.4

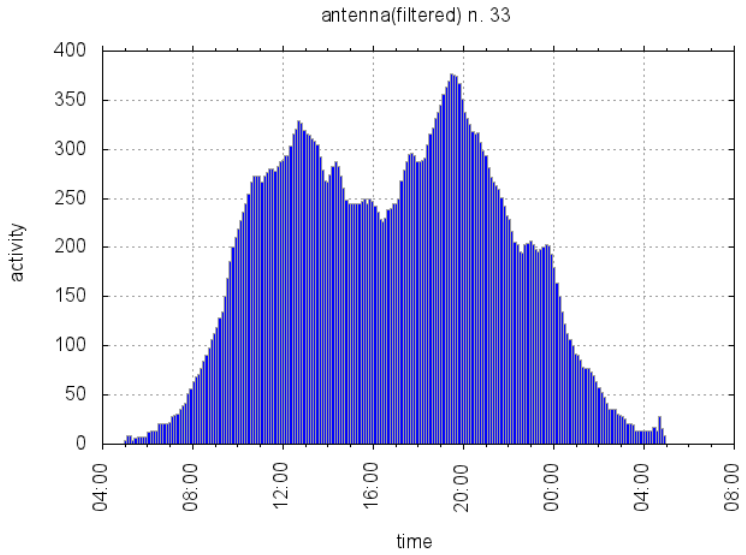

Figura 2.20: Elemento n.5 del cluster n.4

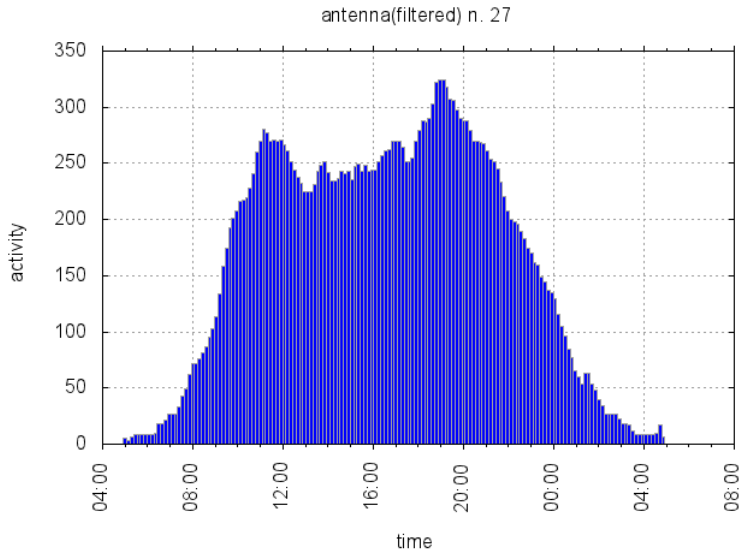

Figura 2.21: Elemento n.6 del cluster n.4

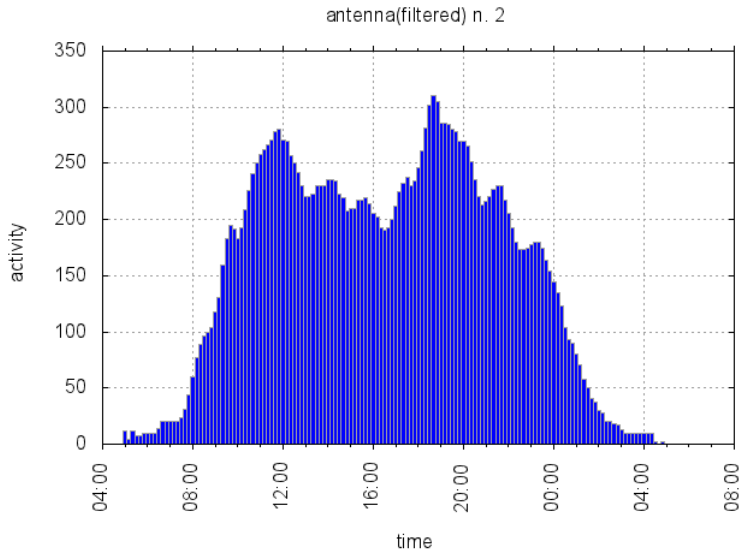

Figura 2.22: Elemento n.7 del cluster n.4

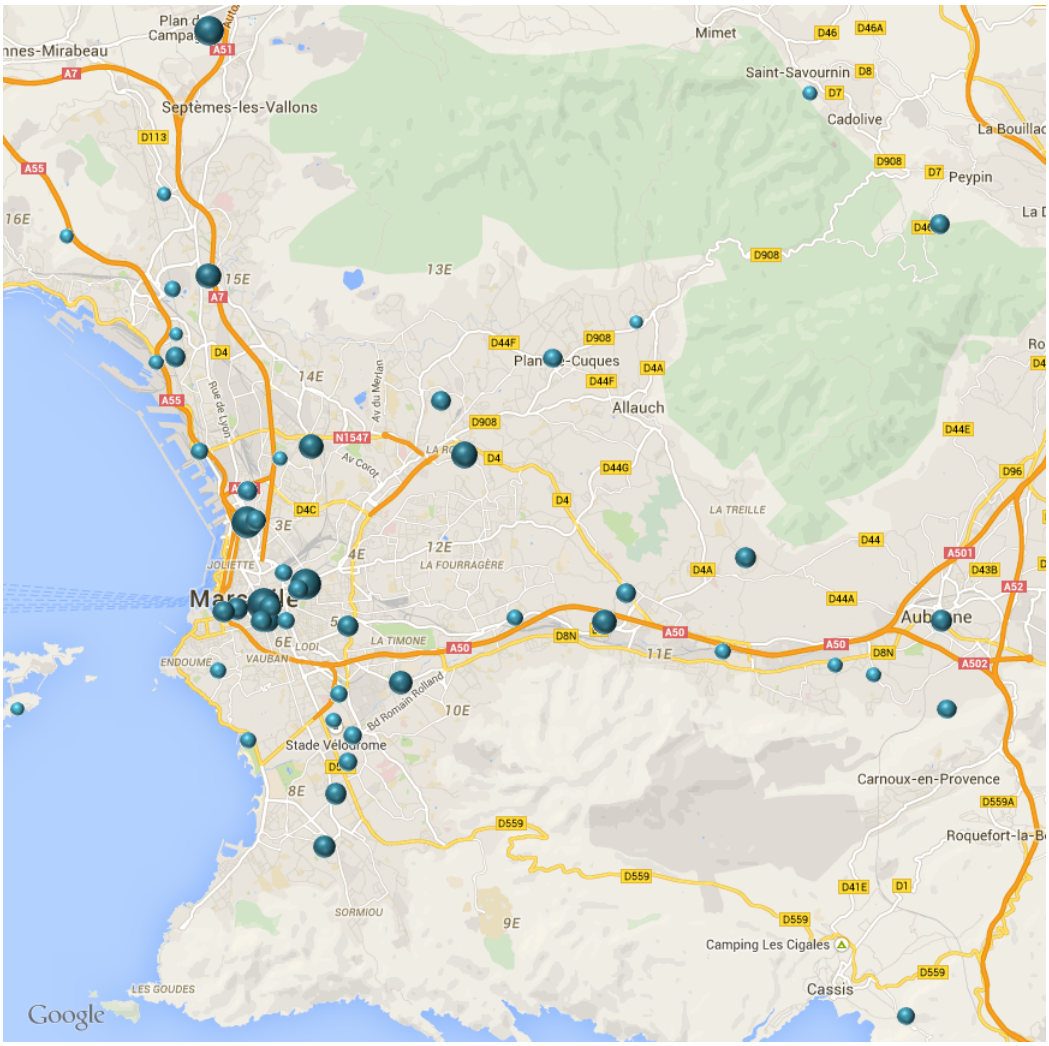

Figura 2.23: posizione dei cluster di sabato 19 ottobre; la grandezza delle sfere é proporzionale all'attivitá del cluster

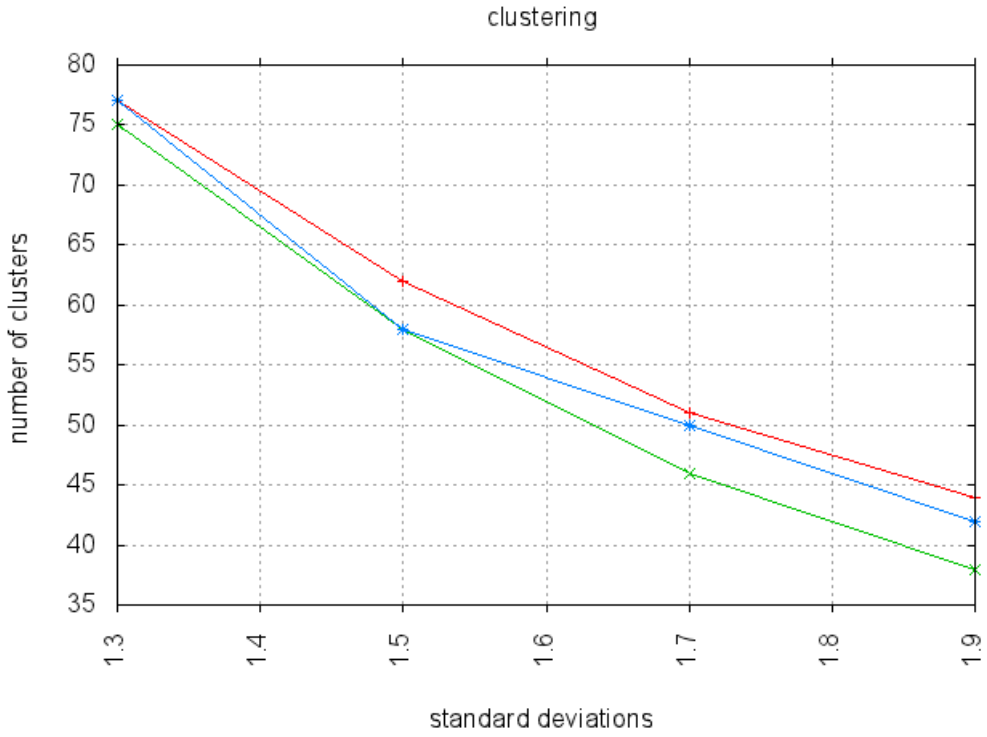

Figura 2.24: Numero di cluster risultati dall'analisi (metodo 2, dati di ottobre) in funzione della tolleranza accettata; il valore scelto é 1.7

Il parametro N viene scelto in modo da ottenere un buon compromesso fra tolleranza all'interno del cluster e numero di elementi del cluster (vedi figura 2.24).

Definiamo (x,y,z) per indicare che abbiamo ottenuto x clusters il primo giorno (sabato), y il secondo e z il terzo. Riportiamo quindi i valori scelti per N:

- $N_{ottobre}^{metodo_1} = 2.3$, da cui (65,72,81)
- $N_{ottobre}^{metodo_2} = 1.7$, da cui (51,46,50)
- $N_{marzo}^{metodo_1} = 3.3$, da cui (66,62,76)
- $N_{marzo}^{metodo_2} = 1.9$, da cui (52,42,53)

Il secondo metodo richiede l'utilizzo di un numero maggiore di deviazioni standard perché l'algoritmo prevede che spesso le antenne con bassa attivitá siano trasformate in antenne molto piú attive. Questo metodo viene abbandonato dal momento che la tolleranza che richiede lo porta a creare clusters poco omogenei al loro interno.

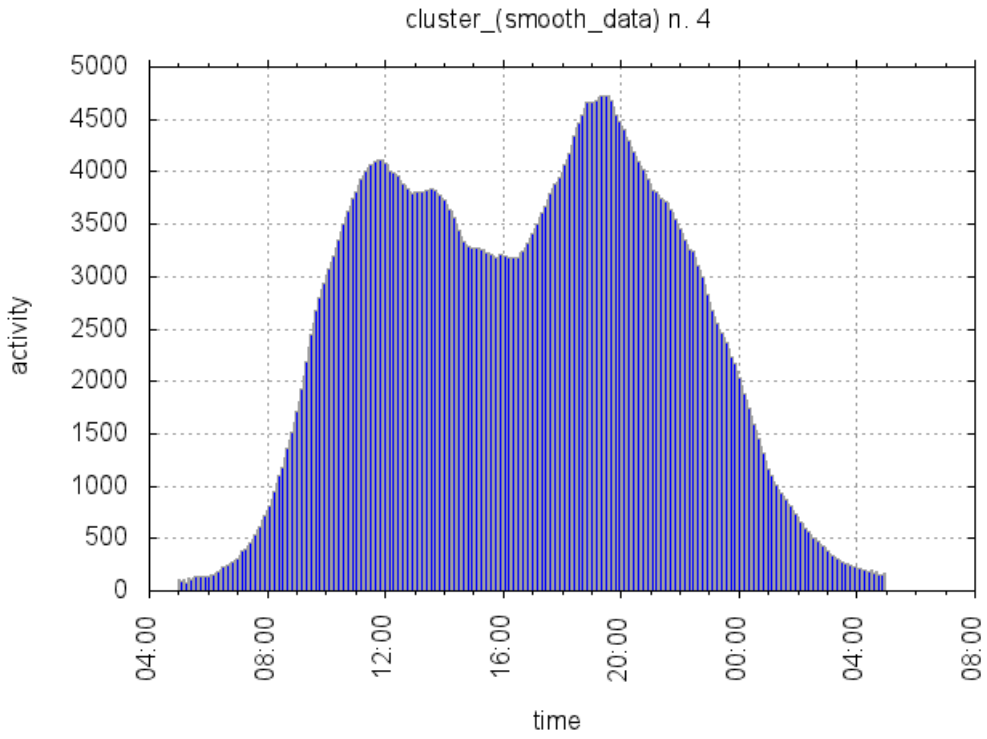

Figura 2.25: Profilo medio del cluster n.4

2.7 Calcolo del segnale

Dopo aver definito i clusters viene raccolta l'attivitá totale del cluster in un unico istogramma, e questo viene ripetuto due volte: la prima per i profili originari e la seconda per i profili filtrati.

A questo punto viene calcolata, per ogni singola antenna, la differenza tra il profilo originale e quello filtrato e si ottengono grafici come quella di figura 2.26.

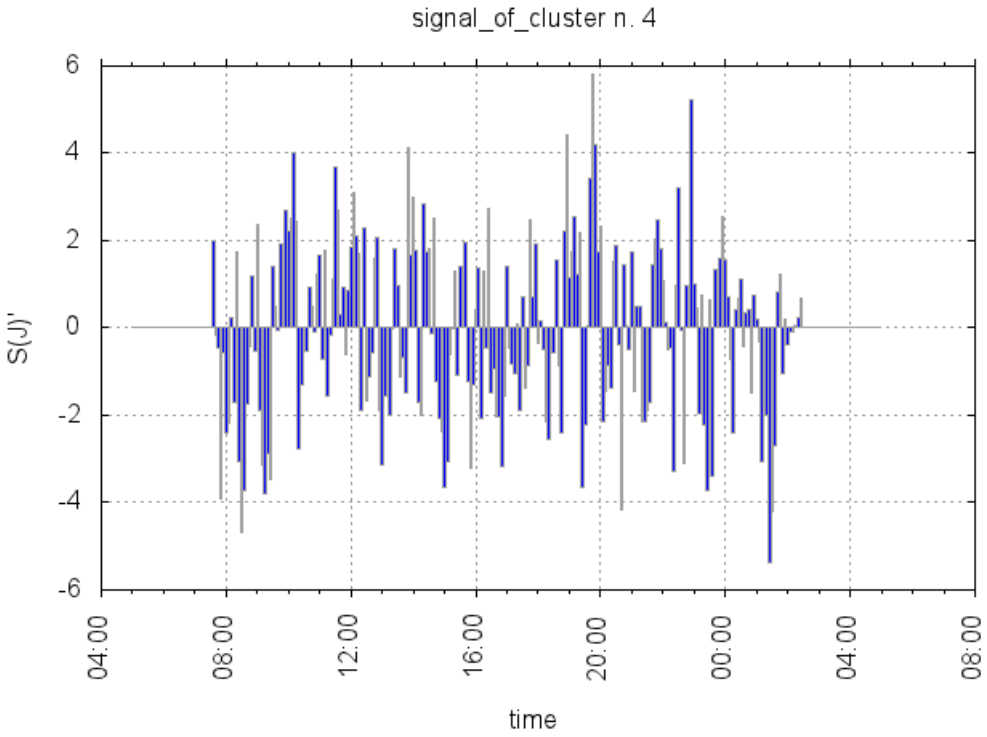

Figura 2.26: Segnale associato al cluster n.4

2.8 Distribuzione dei valori del segnale

I segnali ottenuti da ogni singolo bin del cluster vengono rapportati all'errore associato al bin, secondo la formula:

$$S'_{Jk} = \frac{S_{Jk}}{\sqrt{A_{Jk}}} \tag{2.3}$$

in cui S_{Jk} é il segnale del cluster J nel bin k e A_{Jk} l'attivitá del cluster J (non filtrata) nel bin k; detta B_{ik} l'attivitá dell'antenna filtrata e $S_{ik} = A_{ik} - B_{ik}$ il segnale dell'antenna i nel bin k, S_{Jk} é definito da:

$$S_{Jk} = \sum_{i=0}^{n_{elements}} S_{ik} \tag{2.4}$$

Gli S'_{Jk} vengono raccolti in un istogramma e disposti per grandezza crescente.

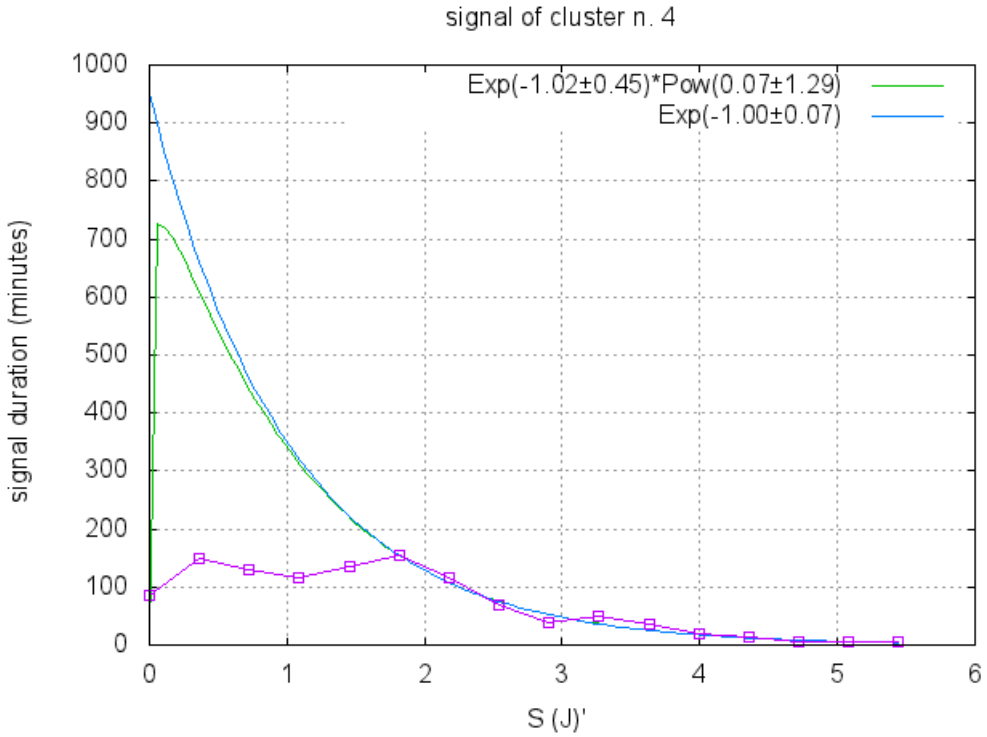

Figura 2.27: Fit per il cluster n.4

2.9 Descrizione e confronto del *fitting* dei dati

La distribuzione osservata sembra avere un'andamento esponenziale. Il nostro modello prevede una distribuzione Poissoniana degli *interevents time* e la stessa distribuzione sará seguita dalle fluttuazioni sulle antenne telefoniche. Se consideriamo la parte centrale dei profili, possiamo ritenere che l'attivitá attesa sia circa costante per un determinato intervallo δt. A questo punto possiamo considerare di avere un'unico processo nell'arco δt per cui prevediamo una distribuzione di Poisson. Eseguiamo un fit esponenziale dei minimi quadrati della forma:

$$f(x) = A \cdot e^{-b \cdot x} \tag{2.5}$$

Il fit viene eseguito solo nel range $x > 1.5$ (che corrisponde alle fluttuazioni superiori ad 1.5 deviazioni standard) al fine di individuare al meglio il comportamento delle code. Vedi ad esempio figura 2.38 (e successive).

Gli esponenti b risultano distribuiti attorno ad un valore di $b = 1$ in tutti e sei i giorni. Riportiamo media e deviazione standard sui sei giorni:

$$b_1 = (1.0 \pm 0.4) \tag{2.6}$$

$$b_2 = (0.9 \pm 0.2) \tag{2.7}$$

$$b_3 = (0.9 \pm 0.2) \tag{2.8}$$

$$b_4 = (1.0 \pm 0.4) \tag{2.9}$$

$$b_5 = (1.0 \pm 0.3) \tag{2.10}$$

$$b_6 = (0.9 \pm 0.3) \tag{2.11}$$

Il coefficiente b dell'esponenziale puó essere interpretato come una temperatura dal momento che regola la velocitá di decadimento dei valori delle fluttuazioni. Data la compatibilitá di tutti i coefficienti, la rete puó essere definita all'equilibrio termico. Per controllare la correttezza del fit calcoliamo il χ^2 ridotto (a cui d'ora in poi ci riferiremo solo con χ^2); questo parametro dovrebbe assumere un valore vicino ad uno se abbiamo fatto una corretta stima degli errori (vedi figura 2.31). Decidiamo di eseguire un fit per la distribuzione prevista da Barabasi:

$$f(x) = A \cdot x^c \cdot e^{-b \cdot x} \tag{2.12}$$

Troviamo un valore medio di $c > 0$ non previsto dai modelli Barabasi. Il valore del χ^2 non viene influenzato significativamente dall'introduzione della termine a potenza. Concludiamo dunque che dai nostri dati si puó affermare solo una compatibilitá con la distribuzione prevista da Barabasi.

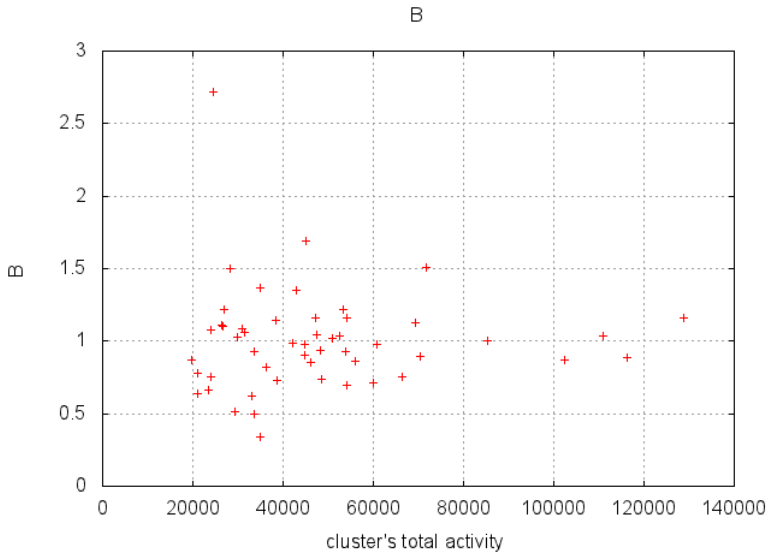

Figura 2.28: coefficiente b di sabato 19 ottobre

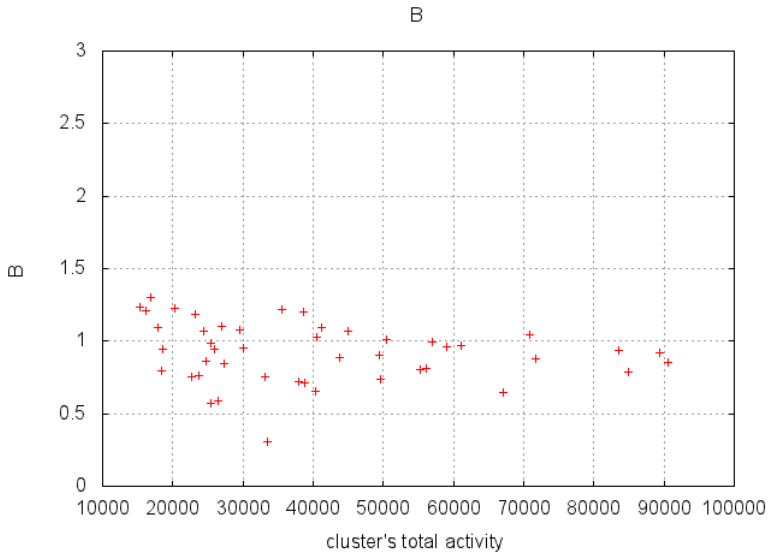

Figura 2.29: coefficiente b di domenica 20 ottobre

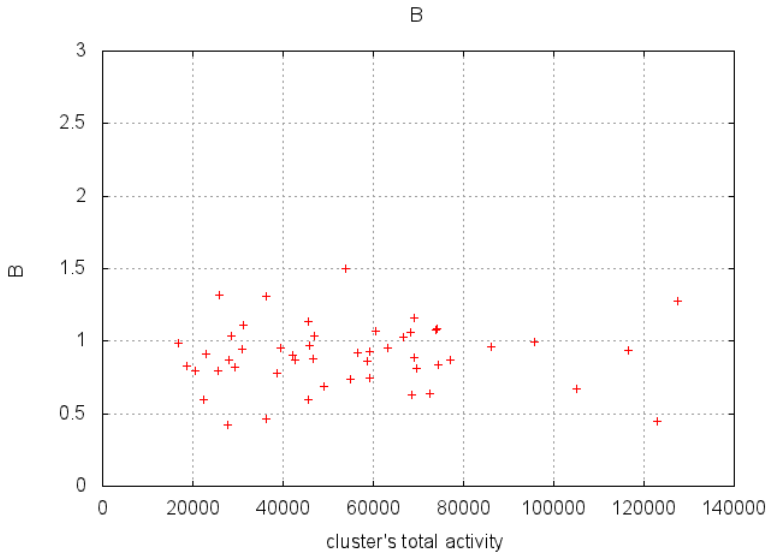

Figura 2.30: coefficiente b di lunedí 21 ottobre

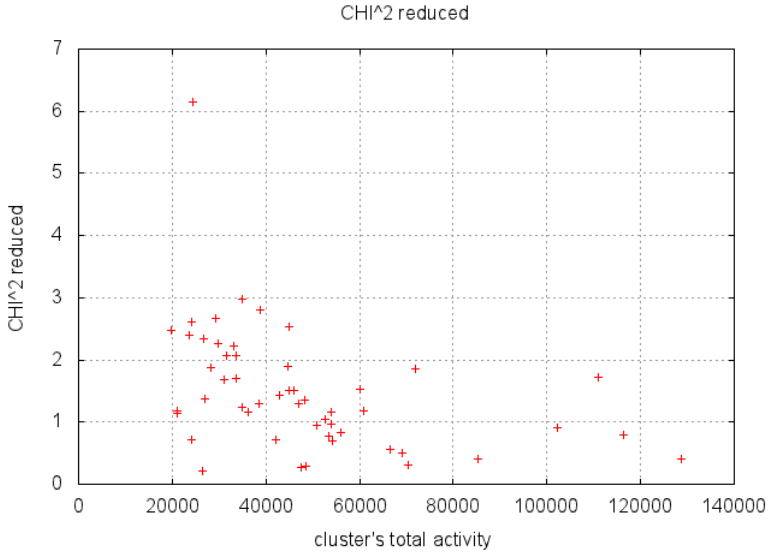

Figura 2.31: χ^2 ridotto di sabato 19 ottobre per la distribuzione di Poisson

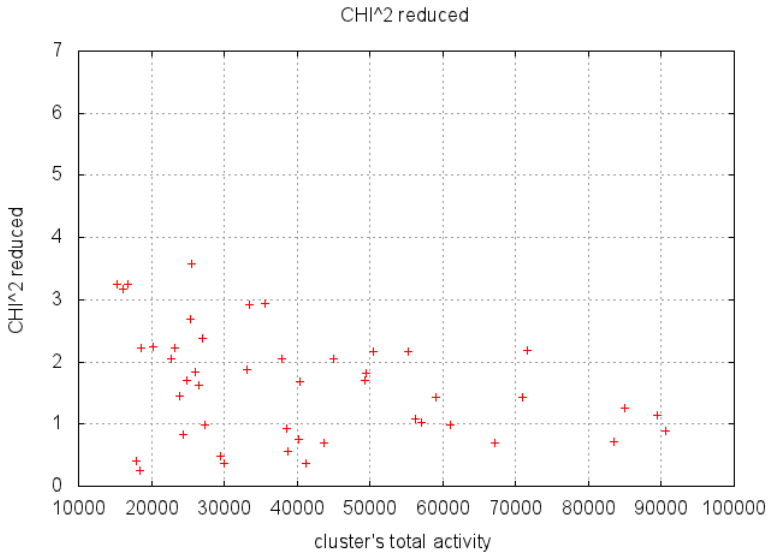

Figura 2.32: χ^2 ridotto di domenica 20 ottobre per la distribuzione di Poisson

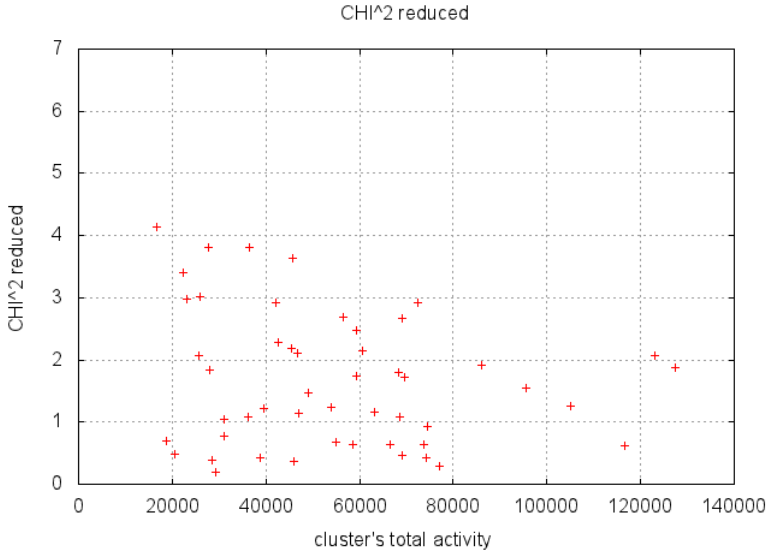

Figura 2.33: χ^2 ridotto di lunedí 21 ottobre per la distribuzione di Poisson

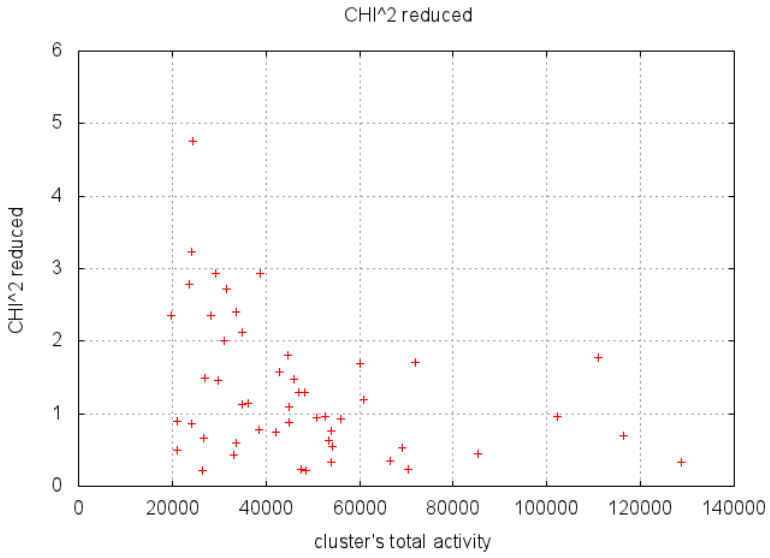

Figura 2.34: χ^2 ridotto di sabato 19 ottobre per la distribuzione di Barabasi

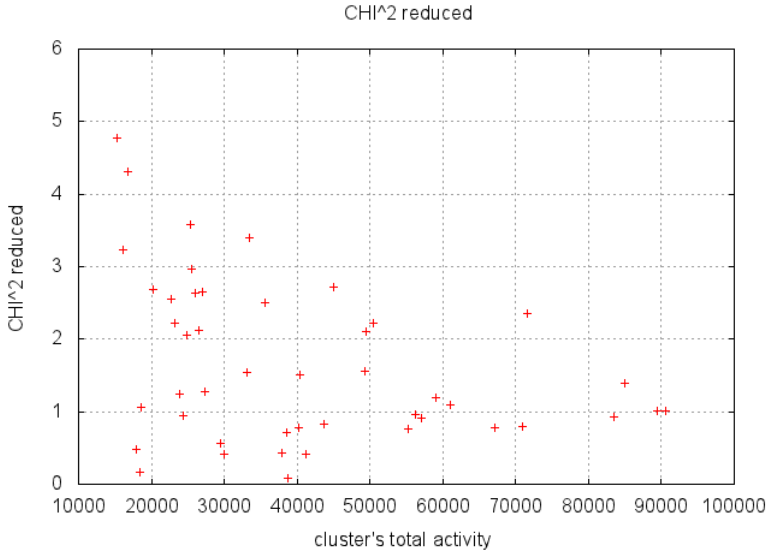

Figura 2.35: χ^2 ridotto di domenica 20 ottobre per la distribuzione di Barabasi

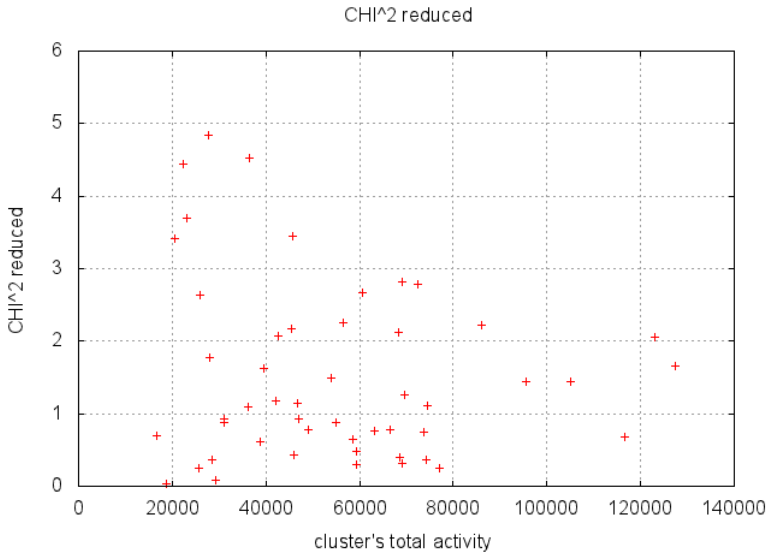

Figura 2.36: χ^2 ridotto di lunedí 21 ottobre per la distribuzione di Barabasi

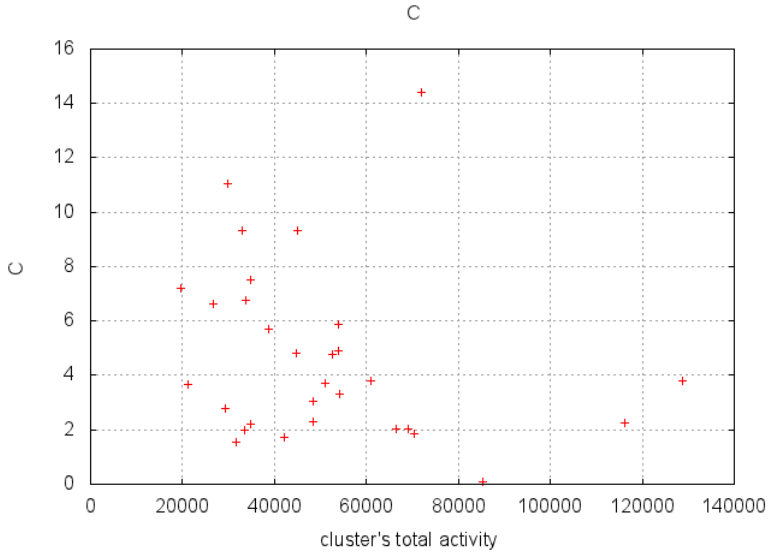

Figura 2.37: esponente c di sabato 19 ottobre della potenza prevista da Barabasi

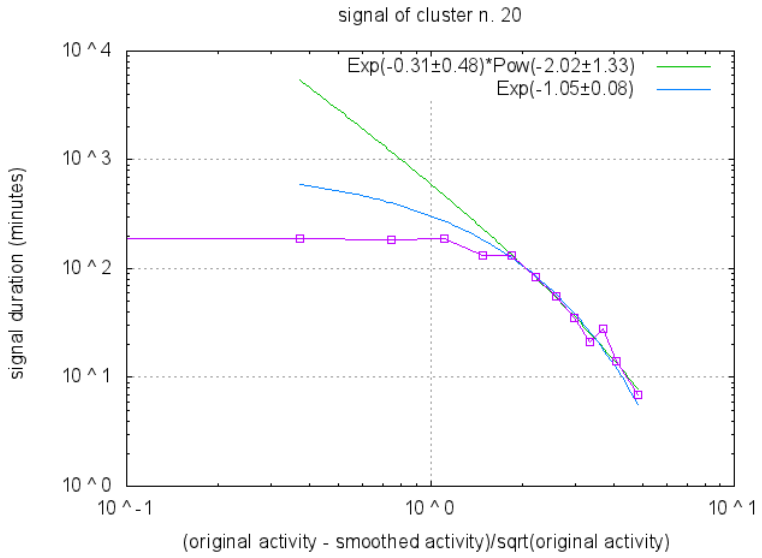

Figura 2.38: Profilo compatibile con entrambe le distribuzioni

2.10 Indicazione delle anomalie

Riassumiamo le diverse situazioni che abbiamo ritrovato nei fit.

1. $\chi^2 < 1$ per entrambe le distribuzioni
2. $\chi^2 < 1$ solo per la legge a potenza con esponente negativo
3. $\chi^2 < 1$ solo per la legge a potenza con esponente positivo
4. $\chi^2 >> 1$ per entrambe le distribuzioni ipotizzate

Quando $\chi^2 >> 1$ per la distribuzione esponenziale, il profilo risulta nettamente distinto dalla media e questo potrebbe indicare la presenza di un'anomalia.

Notiamo che sono tutte antenne appartenenti a cluster singoli: probabilmente il forte segnale che contenevano ha impedito che si raggruppassero con altre antenne.

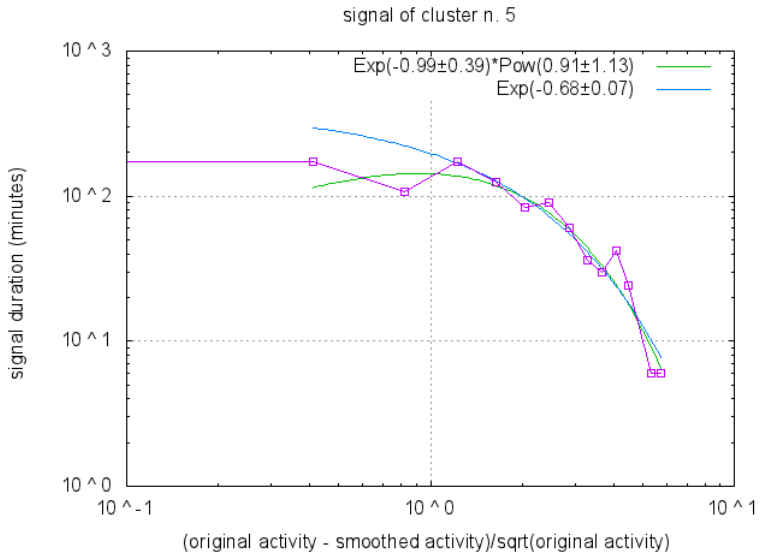

Figura 2.39: Profilo compatibile con entrambe le distribuzioni

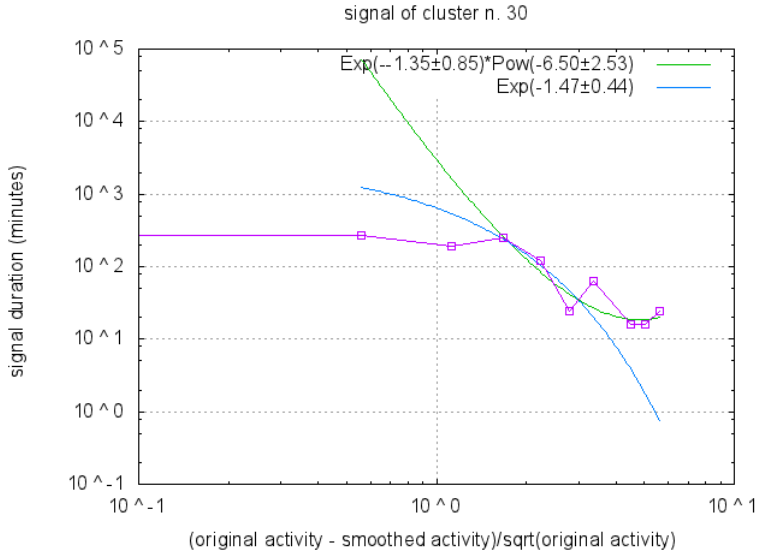

Figura 2.40: Profilo compatibile solo con la legge a potenza con esponente negativo

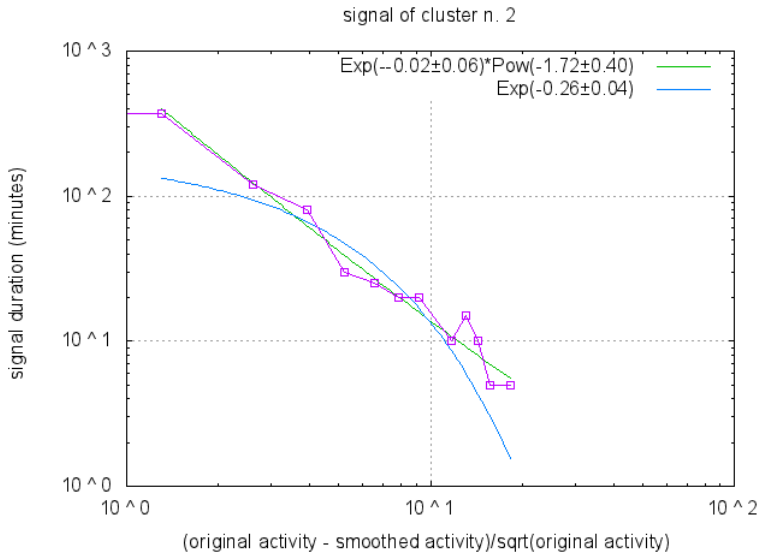

Figura 2.41: Profilo compatibile solo con la legge a potenza con esponente negativo

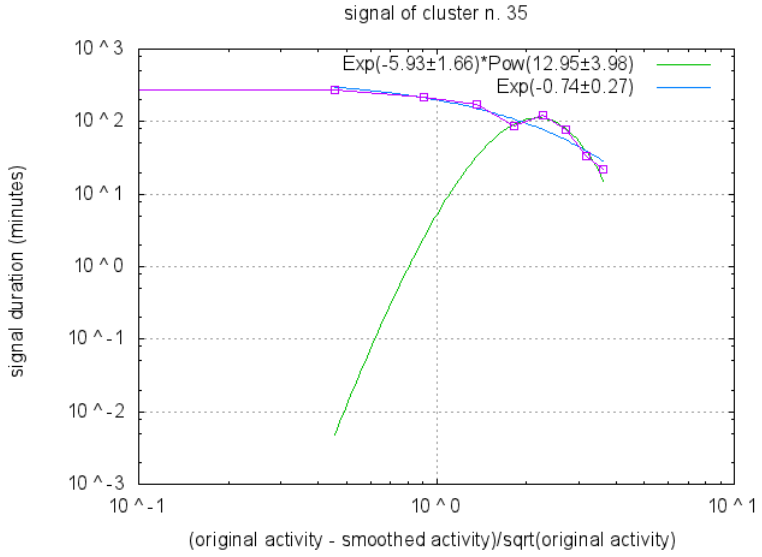

Figura 2.42: Profilo compatibile solo con la legge a potenza con esponente positivo

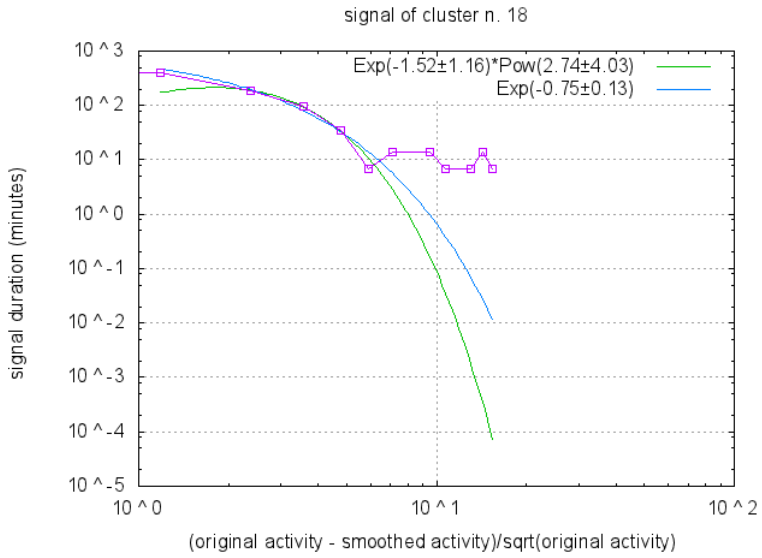

Figura 2.43: Profilo incompatibile con entrambe le distribuzioni

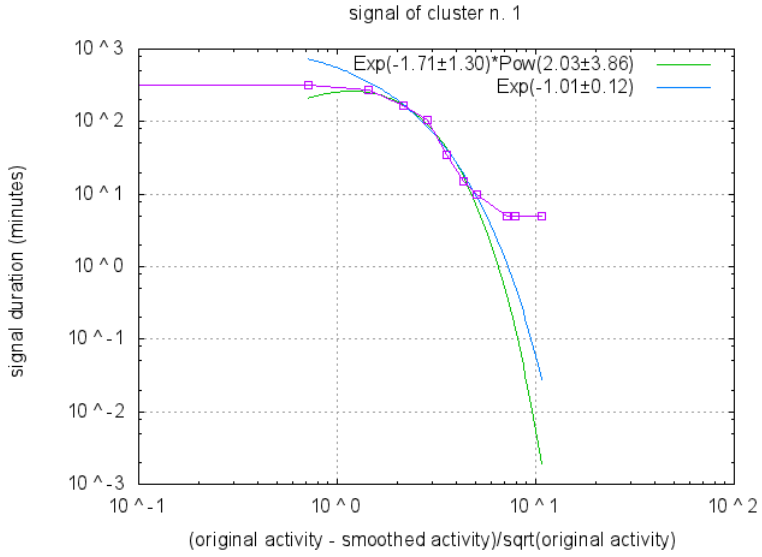

Figura 2.44: Profilo incompatibile con entrambe le distribuzioni

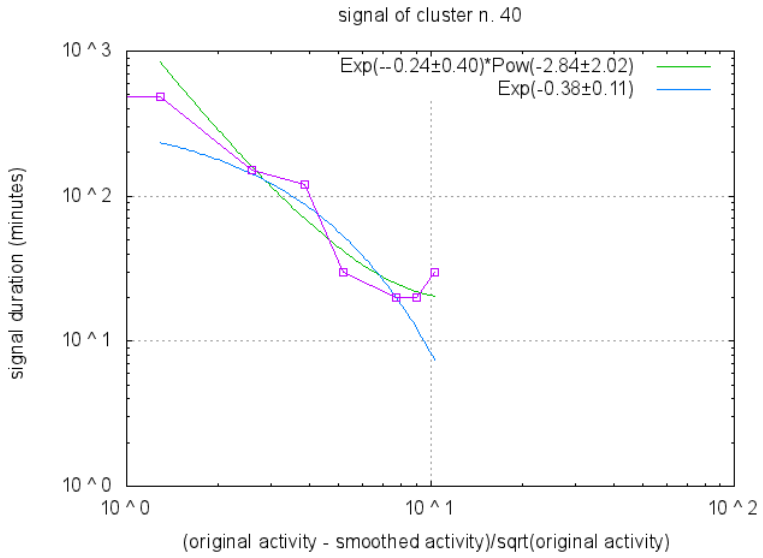

Figura 2.45: Profilo incompatibile con entrambe le distribuzioni

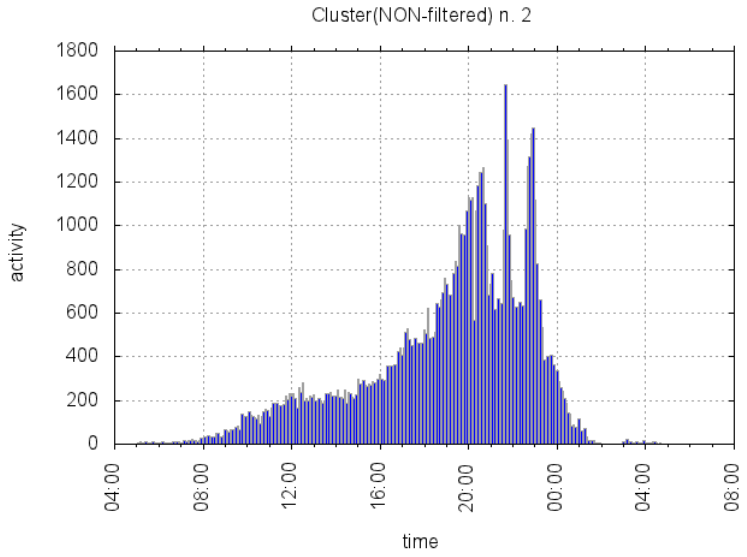

Figura 2.46: profilo d'attivitá di Figura 2.41 osservato sabato 29 marzo 2014

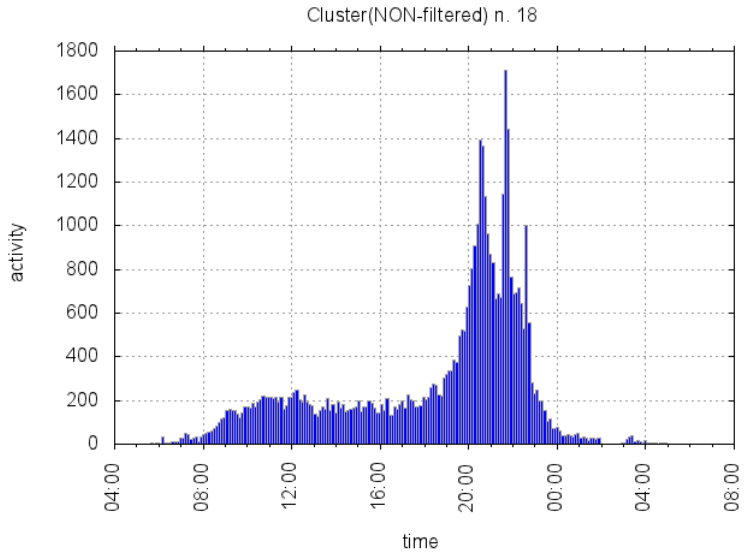

Figura 2.47: profilo d'attivitá di Figura 2.43 osservato sabato 29 marzo 2014

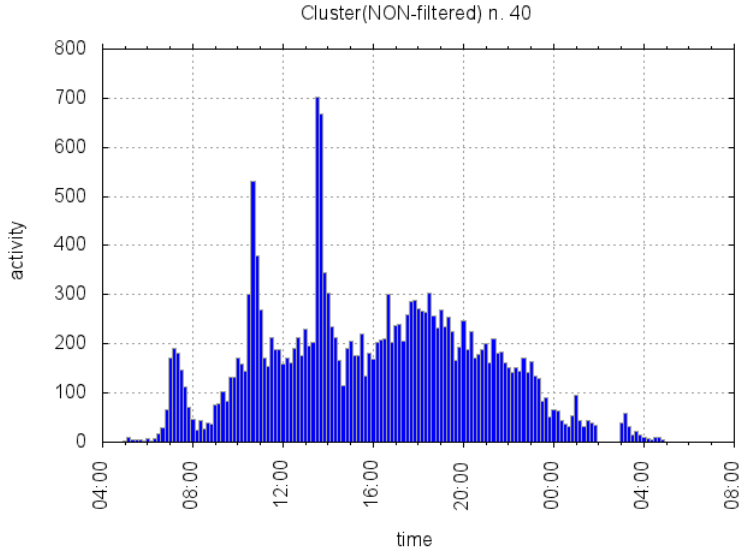

Figura 2.48: profilo d'attivitá di Figura 2.45 osservato sabato 29 marzo 2014

2.11 Conclusione

Nel corso di questo lavoro sono state incontrate numerose difficoltá riguardanti il *Clustering* e abbiamo dovuto tentare diversi approcci prima di raggiungere l'algoritmo finale. Riteniamo che questi fallimenti siano imputabili alla mancanza dei dati necessari per calcolare un profilo atteso che non contenesse fluttuazioni. I profili ottenuti con una media temporale o spaziale non sono infatti mai riusciti a rappresentare correttamente l'identitá dell'antenna. Il profilo medio ottenuto con il filtro, tuttavia, non puó discostarsi eccessivamente dal profilo di partenza e quindi non puó essere una soluzione definitiva. Siamo riusciti in ogni caso a trovare dei possibili segnali anomali. I grafici mostrati in figura 2.43 e 2.47 sembrano infatti individuare un'anomalia, la quale sarebbe anche vicina temporalmente alle elezioni municipali. Non possiamo peró trarre conclusioni definitive a causa di un numero di dati non sufficiente. Per questo stesso motivo non possiamo assumere una posizione definita per quanto riguarda l'applicabilitá del modello Poissoniano o di una legge a potenza:l'errore da cui sono affetti i risultati ci porta ad accettare entrambe le possibilitá. Per quanto riguarda la distribuzione dei coefficienti dell'esponenziale Poissoniana possiamo dire di avere ottenuto solo delle piccole fluttuazioni rispetto ad un valore medio e quindi ci sembra corretto assumere che il sistema sia pressoché all'equilibrio. I risultati ottenuti suggeriscono di approfondire l'analisi, preferibilmente con un numero di dati maggiore. Una volta individuate le anomalie, potremo anche interrogarci su una loro correlazione.

Appendice

Dati spaziali

I dati forniti presentano la posizione delle antenne nella seguente maniera:
in cui:

- *lac* (*Location Area Code*) rappresenta una certa regione della cittá
- *ci* (*Cell Id*) indica le diverse antenne in una determinata *lac*
- *nidt* é il codice del sito (torre GSM) su cui si trova l'antenna; ogni torre puó contenere piú antenne
- *techno* é la generazione tecnologica a cui appartiene l'antenna (questo dato non verrá preso in considerazione)
- x,y: coordinate Lambert 2 Etendu della posizione dell'antenna (sistema usato principalmente in Francia)
- xlon, ylat: coordinate WGS84 della posizione dell'antenna
- *insee* é il codice dell'arrondissement in cui si trova l'antenna (questo dato non verrá preso in considerazione)

ci	lac	nidt	techno	x	y	xlon	ylat	insee
1	5383	00013774J1	3G	837789	1825378	526.738	433.907	13088
1	5388	00000244J1	3G	838654	1821745	527.635	433.578	13088
2	5383	00013774J1	3G	837789	1825378	526.738	433.907	13088
2	5388	00000244J1	3G	838654	1821745	527.635	433.578	13088
3	5383	00013774J1	3G	837789	1825378	526.738	433.907	13088
3	5388	00000244J1	3G	838654	1821745	527.635	433.578	13088
4	5388	00000244J1	3G	838654	1821745	527.635	433.578	13088

Figura 2.49: dati di posizione

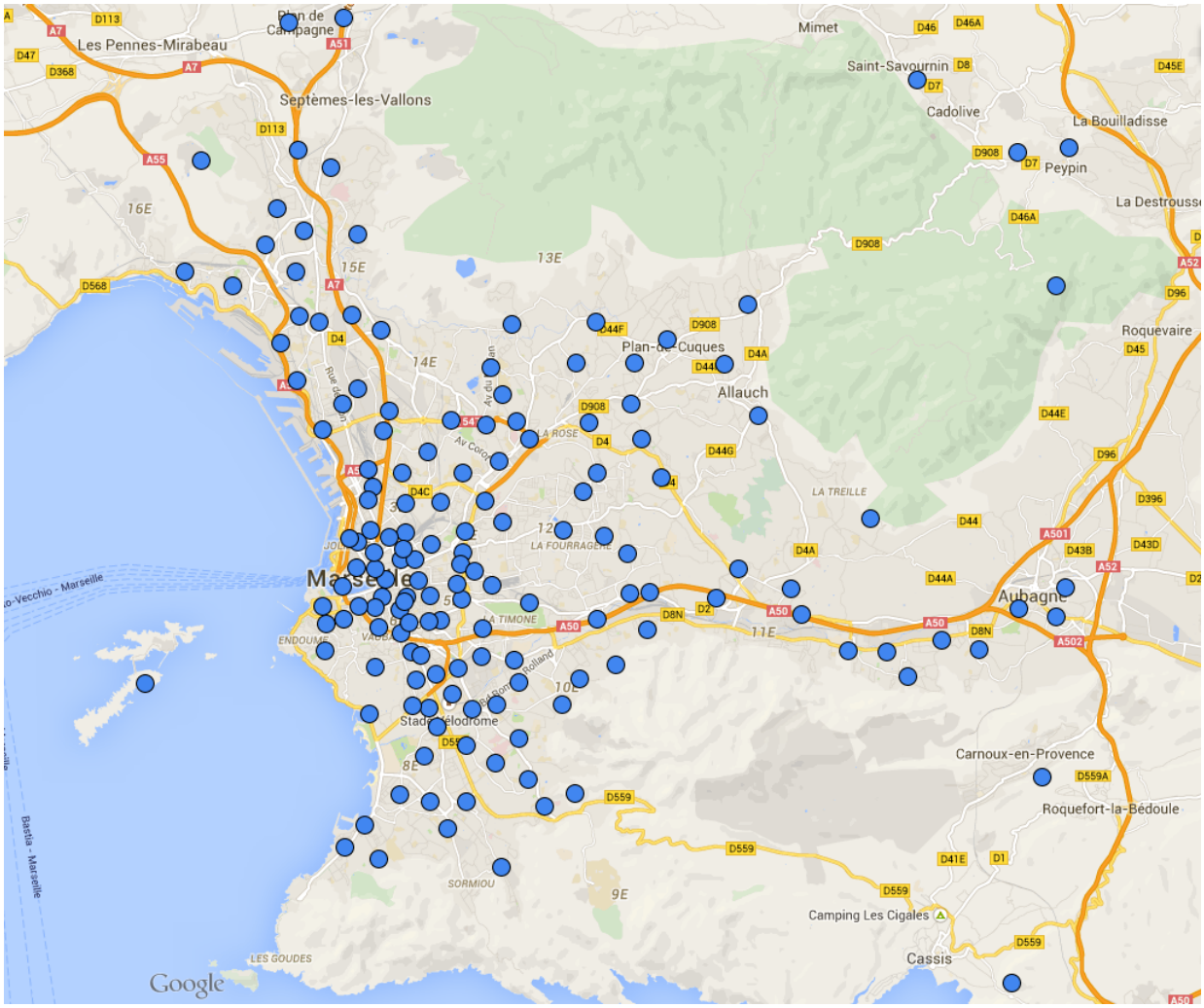

Figura 2.50: posizione dei cluster di sabato 19 ottobre

In questi dati viene quindi fatta una distinzione tra le antenne vere e proprie, identificate da *lac* e *ci*, e la torre GSM su cui ci possono essere una decina di antenne, identificata dal *nidt*. Dal momento che tutte le antenne sulla medesima torre hanno la stessa posizione, questa distinzione non ci interessa nell'analisi, ma saremo obbligati a utilizzare il codice *lac* e *ci* per leggere i dati temporali. Le coordinate della posizione delle torri ci vengono fornite sia secondo il sistema Lambert 2 Etendu che il sistema WGS84; scegliamo questo secondo sistema dato che é lo stesso sistema usato dai GPS e quindi rende pi immediata la georeferenziazione delle antenne. Le antenne totali sono 303.

A questo punto possiamo ricostruire la mappa delle antenne.

dt		evtype	lac	ci
21/10/2013	00.00.01	1	20482	3397
21/10/2013	00.00.01	4	5388	1263
21/10/2013	00.00.01	1	5388	38299
21/10/2013	00.00.01	1	5388	269
21/10/2013	00.00.01	3	5391	34975
21/10/2013	00.00.01	3	5388	299
21/10/2013	00.00.01	4	20481	1431
21/10/2013	00.00.01	3	5388	7145

Figura 2.51: dati di tempo

Dati temporali

I dati relativi alle attivitá delle antenne sono presentati nella seguente maniera:
in cui:

- *dt* é la data a cui é avvenuto l'evento
- *evtype* é la tipologia dell'evento
    1. = SERVICE EVENT: un cellulare vuole fare una chiamata o inviare un SMS/MMS
    2. =DETACH EVENT: un cellulare viene spento
    3. = LOCATION UPDATE EVENT: un cellulare si é spostato in una differente *lac*
    4. = PAGING RESPONSE EVENT: un cellulare ha accettato una chiamata o ricevuto un SMS
    5. = HANDOVER EVENT: un cellulare ha cambiato *ci* durante una chiamata
    6. = UNKNOWN EVENT: evento non identificato
- *lac* e *ci* identificano l'antenna che ha presentato tale evento

Possiamo quindi associare gli eventi con una determinata *lac* e *ci* all'antenna localizzata precedentemente dagli stessi codici. Una volta completata questa operazione, possiamo trascurare questi codici e riferirci unicamente alla posizione in coordinate WGS84. Nella nostra analisi siamo interessati solo all'attivitá intesa come l'atto di inviare o ricevere chiamate/SMS/MMS e quindi selezioniamo solo gli eventi di tipo 1 e 4.

Bibliografia

1. Modeling bursts and heavy tails in human dynamics A. Vzquez, J.G.Oliveira, Z. Dezs, 2 K.I. Goh, I. Kondor, A.L. Barabsi PHYSICAL REVIEW E 73, 036127 (2006)

2. From mobile phone data to the spatial structure of cities T. Louail, M. Lenormand, O. G. Cantu, M. Picornell, R. Herranz, E. F. Martinez, J. J. Ramasco, M.Barthelemy Sci. Rep. PY (2014)

3. Uncovering individual and collective human dynamics from mobile phone records J. Candia, M.C. Gonzalez, P. Wang, T. Schoenhar, G. Madey, A.L. Barabasi J. Phys. A: Math. Theor. 41 (2008) 224015 (11pp)

4. Time patterns, geospatial clustering and mobility statistics based on mobile phone network data E. de Jonge, M.van Pelt, M. Roos

5. Calling patterns in human communication dynamics Z.Q. Jiang, W.J. Xie, M.X.Li, B. Podobnik, W.X.Zhou, H.E. Stanley 10.1073/pnas.1220433110 (2013)

6. Does Urban Mobility Have a Daily Routine? Learning from Aggregate Data Mobile Networks A. Sevtsuk, C. Ratti Journal of Urban Technology, 17:1, 41-60 (2010)

7. Impact of Non-Poissonian Activity Patterns on Spreading Processes A. Vazquez, B. Racz, A. Lukacs, A.L. Barabasi PRL 98, 158702 (2007)

8. The Theory of Stochastic Processes D.R. Cox, H.D. Miller