Analisi aggregata per ore

Michele Carignani, Alessandro Lenzi

3 marzo 2014

1 Generazione dei grafi orari

Per prima cosa i dati sono stati aggregati per ora. I dati originali del dataset Telecommunications - MI to MI sono nel formato:

```
timestamp \t SourceId \t DestId \t Stregth
```

e sono stati suddivisi in 24 file (uno per ogni ora) e aggregati, per cui ogni file contiene (al massimo[1]) un record per ogni nodo nel formato:

```
SourceId \t DestId:Strength [\t DestId:Strength]
```

A questo punto i pesi sugli archi (sopra chiamati `Strength`) sono stati riscalati rispetto alla somma dei valori della stella uscente di un nodo, ottenendo la probabilità di transire dal nodo i al nodo j, ovvero:

$$sumStrength_i = \sum_{j \in FS(i)} Strength_{ij}$$

$$probability_{ij} = \frac{Strength_{ij}}{sumStrength_i}$$

2 Ricerca delle componenti fortemente connesse

Per ricercare le componenti fortemente connesse (in seguito CFC) è stato utilizzato l'algoritmo Tarjan su un sotto insieme degli archi "tagliati" secondo il peso percentuale.

2.1 Tagli

Un taglio degli archi su un valore x significa utilizzare per la visita solo gli archi con peso (ovvero valore di probabilità) maggiore o uguale a x. L'algoritmo esegue sampling sugli archi (10^6 sample) e calcola la distribuzione delle probabilità nella fascia oraria in esame. Su questa distribuzione calcola i percentili. L'algoritmo è stato eseguito con diversi tagli ai percentili 99, 95, 90 e 80.

2.2 Strategie di visita

Le strategie di visita utilizzate sono 5 e impiegano i dati del dataset "Telecommunications - SMS, Call, Internet - MI": a partire dai record del formato

```
SquareID \t Timestamp \t .. ChiamateInUscita ..
```

per ogni ora sono stati generati file con record

```
SquareID \t AggregatedCalls
```

[1] poichè certi nodi possono non avere chiamate in uscita in una certa fascia oraria.

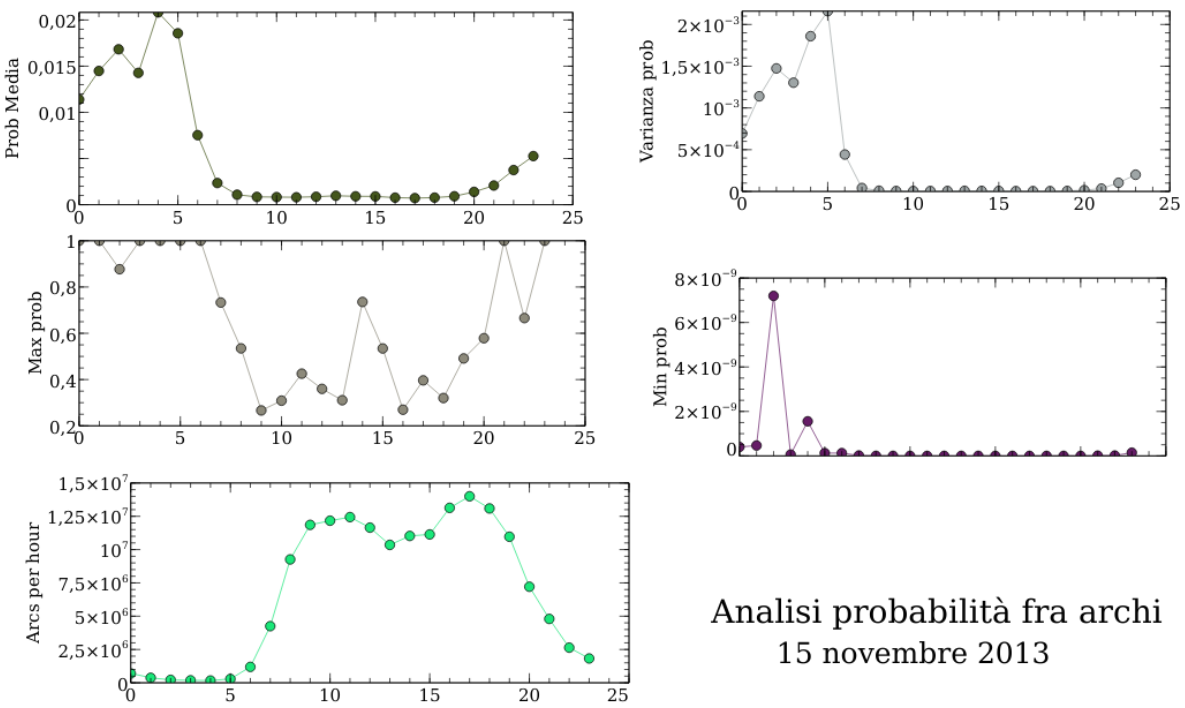

Figura 1: Statistiche sui pesi degli archi, 15 novembre. Sulle ascisse le fascie orarie, sulle ordinate le probabilità.

che permettono di capire quale sia il valore assoluto proporzionale a tutte le chiamate in uscita dallo square ID in una certa fascia oraria. In questo modo è possibile iniziare la visita del grafo non da nodi ordinati lessicograficamente ma in ordine (crescente o decrescente) di traffico in uscita. Le strategie inoltre si differenziano per il modo di ordinare la stella uscente da un nodo. Sono state provate diverse strategie:

- SCC1: visita il grafo in ordine crescente di traffico uscente, selezionando prima gli archi con probabilità maggiore;
- SCC2: visita i nodi per traffico decrescente e con archi selezionati per probabilità crescente;
- SCC3: visita i nodi per traffico decrescente e gli archi per probabilità crescente;
- SCC4: visita dei nodi per traffico crescente e archi per probabilità decrescente;
- Stable: Esegue la ricerca delle componenti fortemente connesse selezionando i nodi in ordine crescente di traffico telefonico **giornaliero** uscente e gli archi in ordine di probabilità decrescente.

3 Risultati

3.1 Statistiche

In fig. 1 sono mostrate le statistiche sui pesi degli archi come probabilità sulle diverse fasce orarie della giornata del 15 novembre.

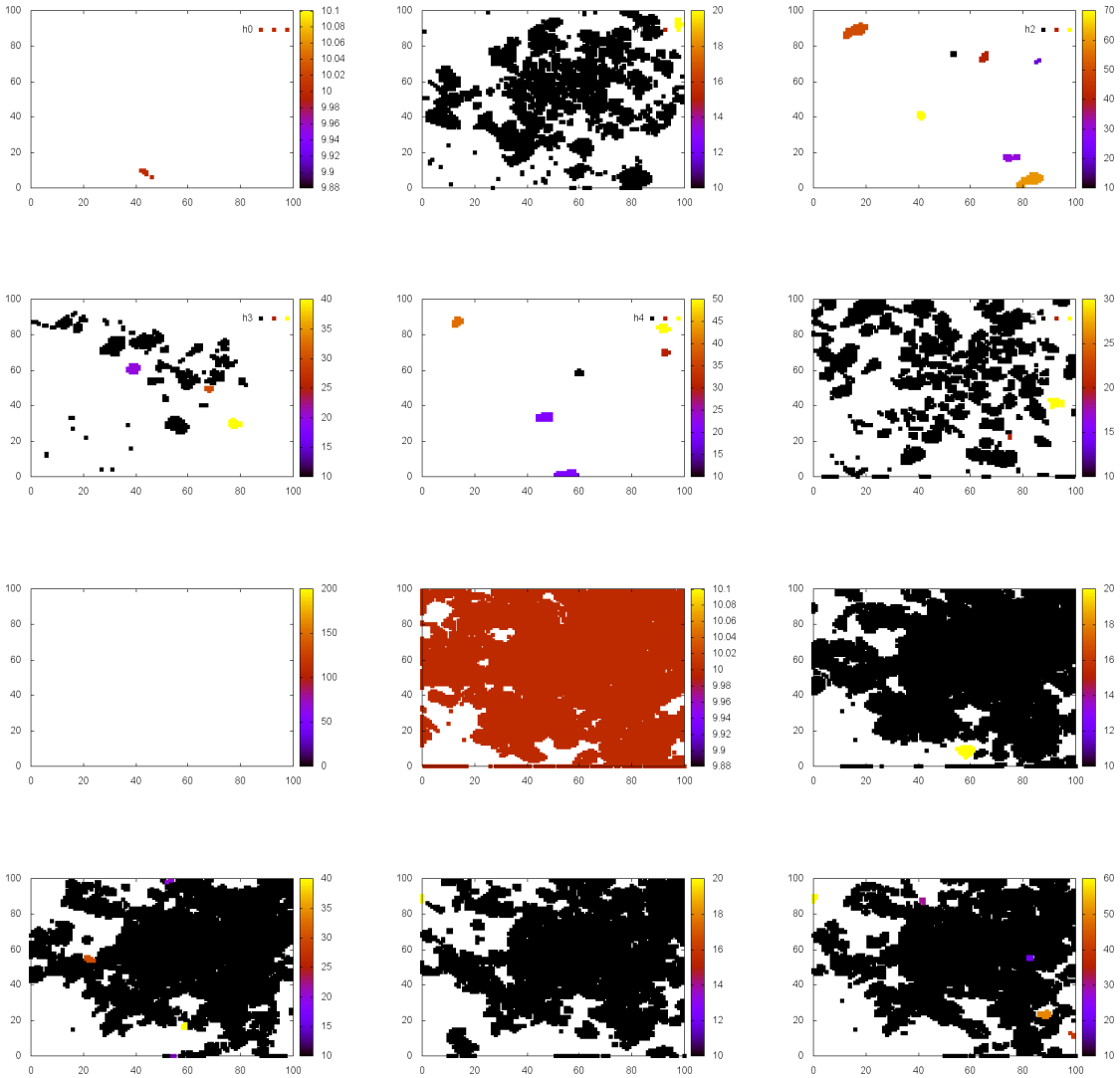

Figura 2: SCC1, Taglio 0.005, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

3.2 Componenti Fortemente Connesse

Le CFC sono state disegnate su mappe (utilizzando `gnuplot`) assegnando una funzione z alle celle di coordinate (x, y) così definita:

$$z(x,y) = \begin{cases} 0, & \text{se } (x,y) \text{ non appartiene ad alcuna CFC,} \\ 10k, & \text{se } (x,y) \text{ appartiene alla k-esima CFC trovata.} \end{cases}$$

Perciò i seguenti grafici disegnagno le CFC nella loro posizione sulla Milano Grid.

3.2.1 SCC1, taglio 0.005

Come esempio per tutti i casi SCC riportiamo il caso SCC 1 con taglio degli archi con valore minore di 0.005. In questo caso, le componenti sono per la maggior parte delle ore estremamente ampie, come si può vedere in seguito:

Si noti in 4 come la dimensione delle componenti può variare anche notevolmente a seconda del traffico presente in un'ora. Questo potrebbe essere dovuto al fatto che il taglio è estremamente basso, e

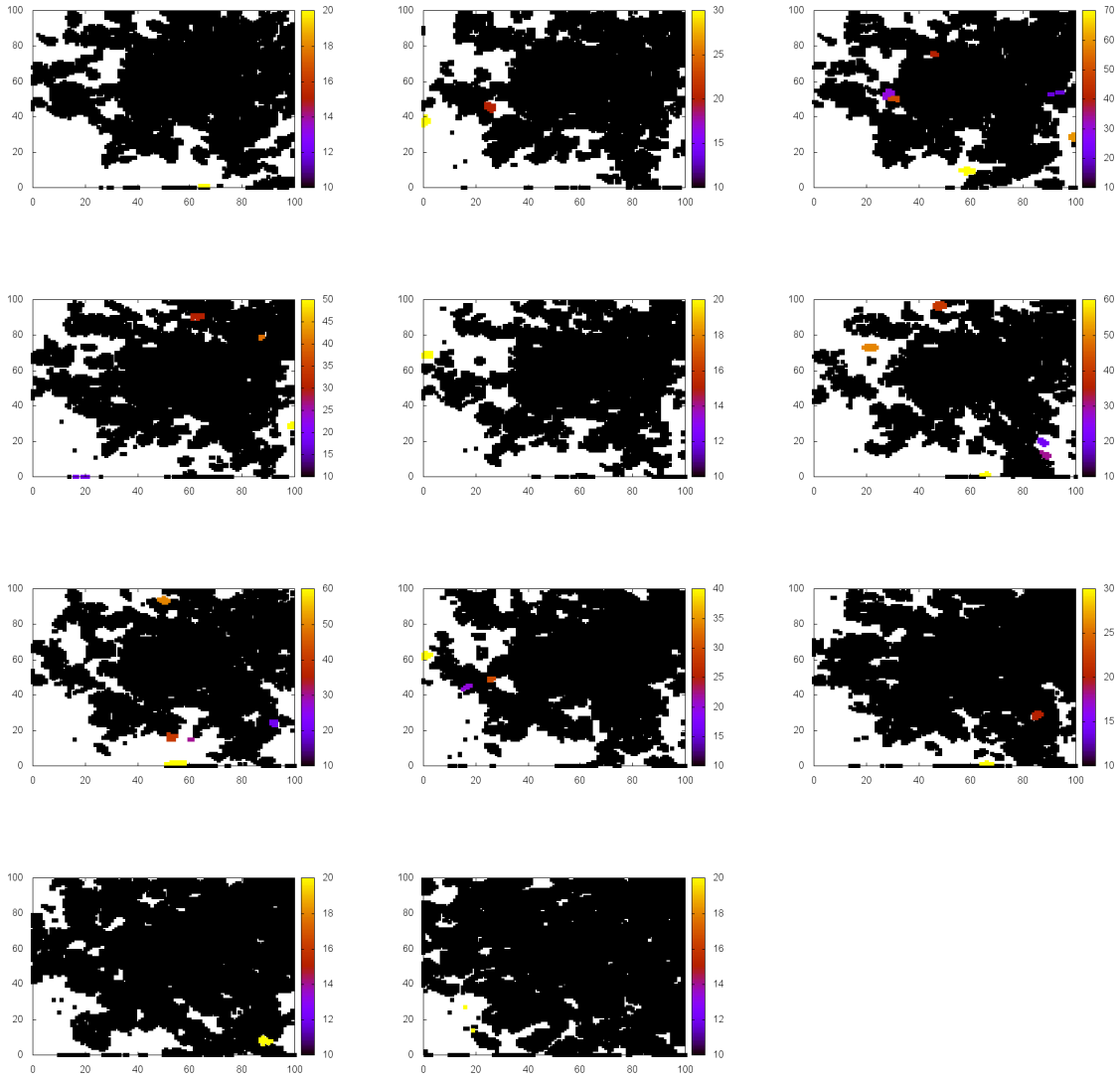

Figura 3: SCC1, Taglio 0.005, h 12-22; da vedersi da sinistra verso destra e dell'alto verso il basso.

```
0:      6,
1:      2972,11,
2:      4,2,13,10,38,40,9,
3:      462,19,7,16,
4:      7,41,7,14,14,
5:      2563,2,18,
6:
7:      8010,
8:      6605,29,
9:      5677,5,8,5,
10:     5638,4,
11:     5492,4,5,3,13,6,
12:     5729,4,
13:     5692,14,11,
14:     5818,5,13,5,8,6,12,
15:     5784,5,12,3,5,
16:     5362,11,
17:     5118,9,7,15,14,5,
18:     5623,8,2,11,11,14,
19:     5951,5,5,10,
20:     6676,11,6,
21:     7449,12,
22:     7971,2,
```

Figura 4: Nel listato, per ogni ora (alla sinistra), una lista delle dimensioni delle CFC trovate

pertanto agisce solamente su una parte minima degli archi. Nelle ore in cui il traffico è minore, anche le componenti hanno dimensioni notevolmente minori: probabilmente questo è dovuto al fatto che il numero di archi uscenti (fuori dall'orario lavorativo) è minore e con probabilità maggiore.

3.2.2 Stable, taglio al 99-esimo percentile

3.2.3 Stable, taglio al 95-esimo percentile

3.2.4 Stable, taglio al 90-esimo percentile

4 Alcune conclusioni

1. Occorre effettuare tagli con valore molto alto perchè il grafo è inerentemente fortemente connesso.
2. Le statistiche sulle probabilità degli archi denotano l'effettiva esistenza di square che si chiamano più degli altri negli orari non lavorativi.
3. La strategia di visita incide molto nella formazione del risultato e probabilmente il problema di non considerare mai due volte lo stesso nodo può eliminare delle CFC interessanti.
4. Il numero di CFC è più basso di quello attesa, tuttavia le CFC sono ben distribuite spazialmente sulla grid; si noti anche come al variare delle fasce orarie variano le zone in cui vengono scoperte le CFC.
5. Sopratutto nelle ore di minor traffico si individuano principalmente CFC che potrebbero essere fatte risalire alla presenza di paesi della periferia di Milano (la grid copre un'area molto più grande della sola zona urbana di Milano).
6. Le CFC trovate non mostrano persistenza (salvo rare eccezioni) in fasce orarie contigue ma tendono ad apparire e scomparire. Si noti come questo non può essere dovuto al valore del taglio, perchè i percentili vengono calcolati per ogni fascia oraria.

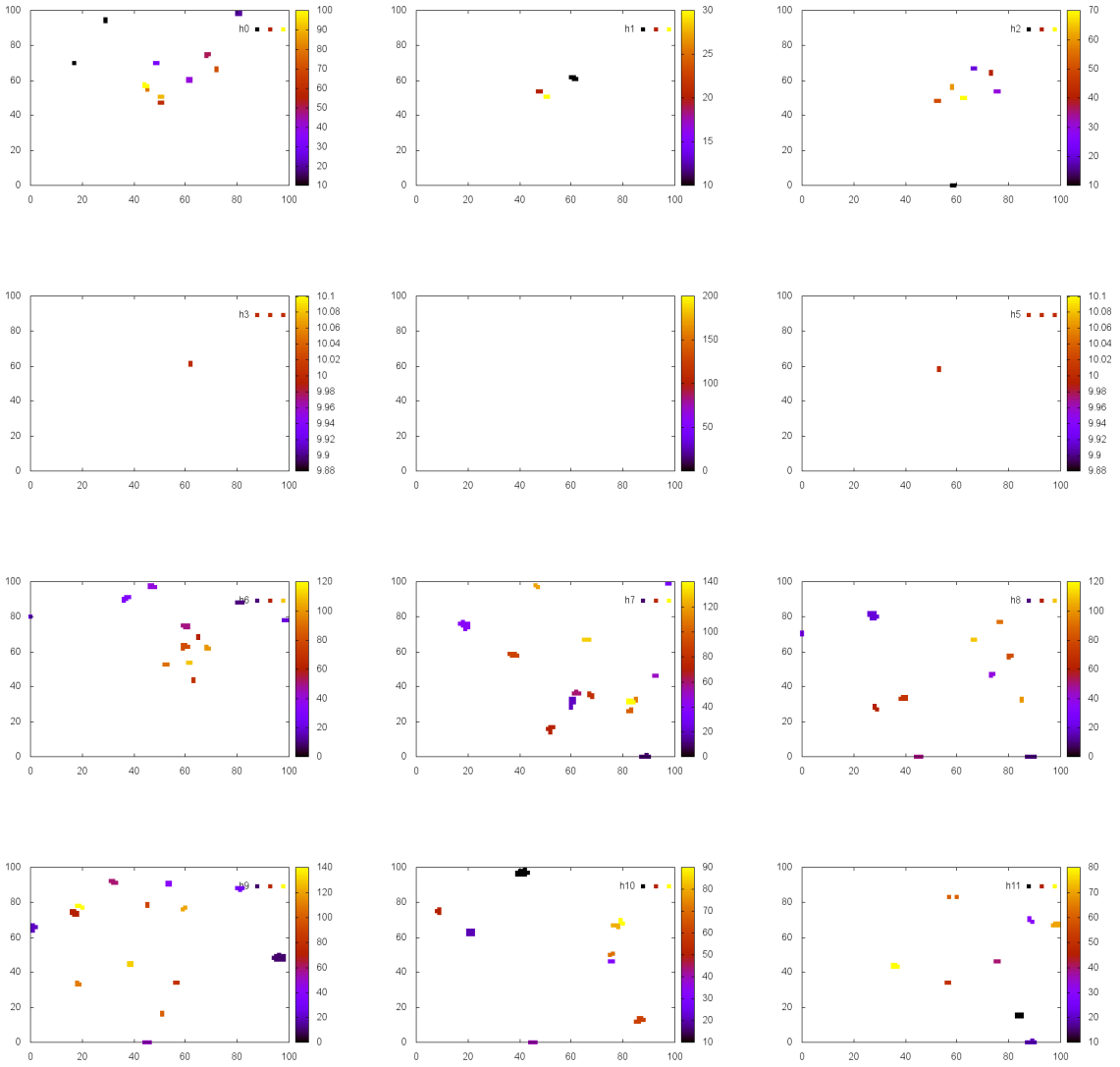

Figura 5: Stable, Taglio 99-esimo percentile, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

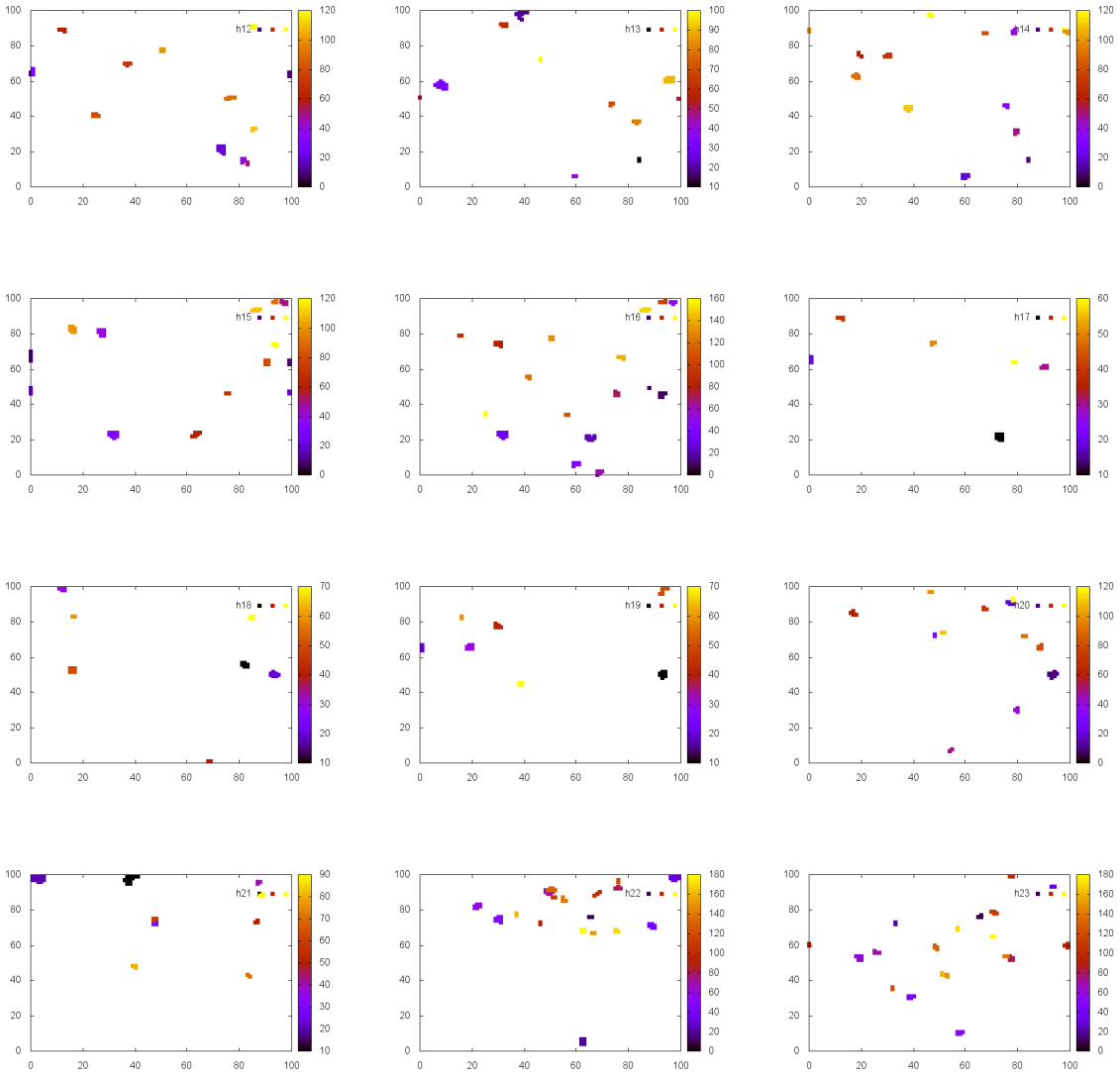

Figura 6: Stable, Taglio 99-esimo percentile, h 12-23. Da vedere da sinistra verso destra e dall'alto verso il basso

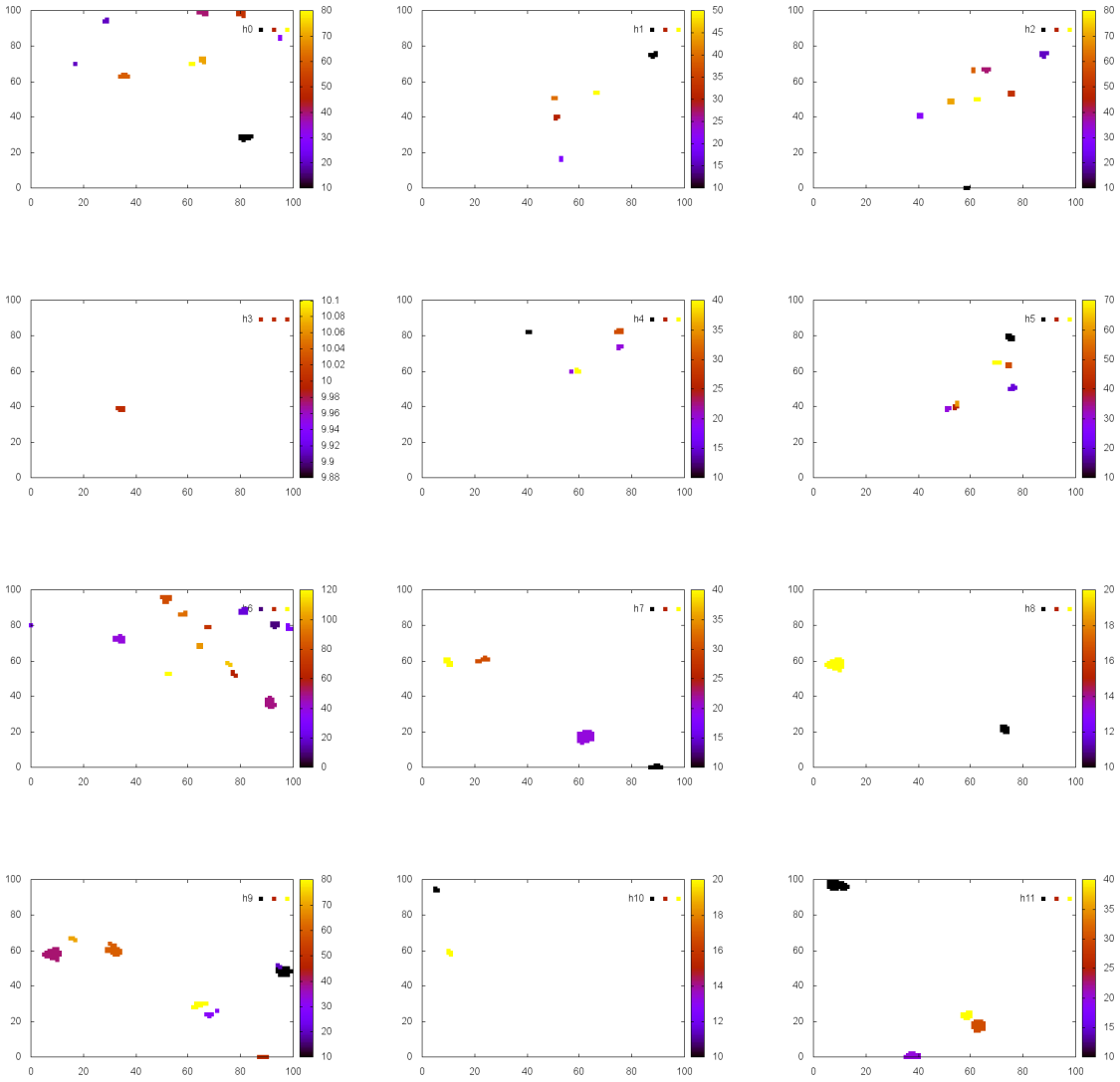

Figura 7: Stable, Taglio 95-esimo percentile, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

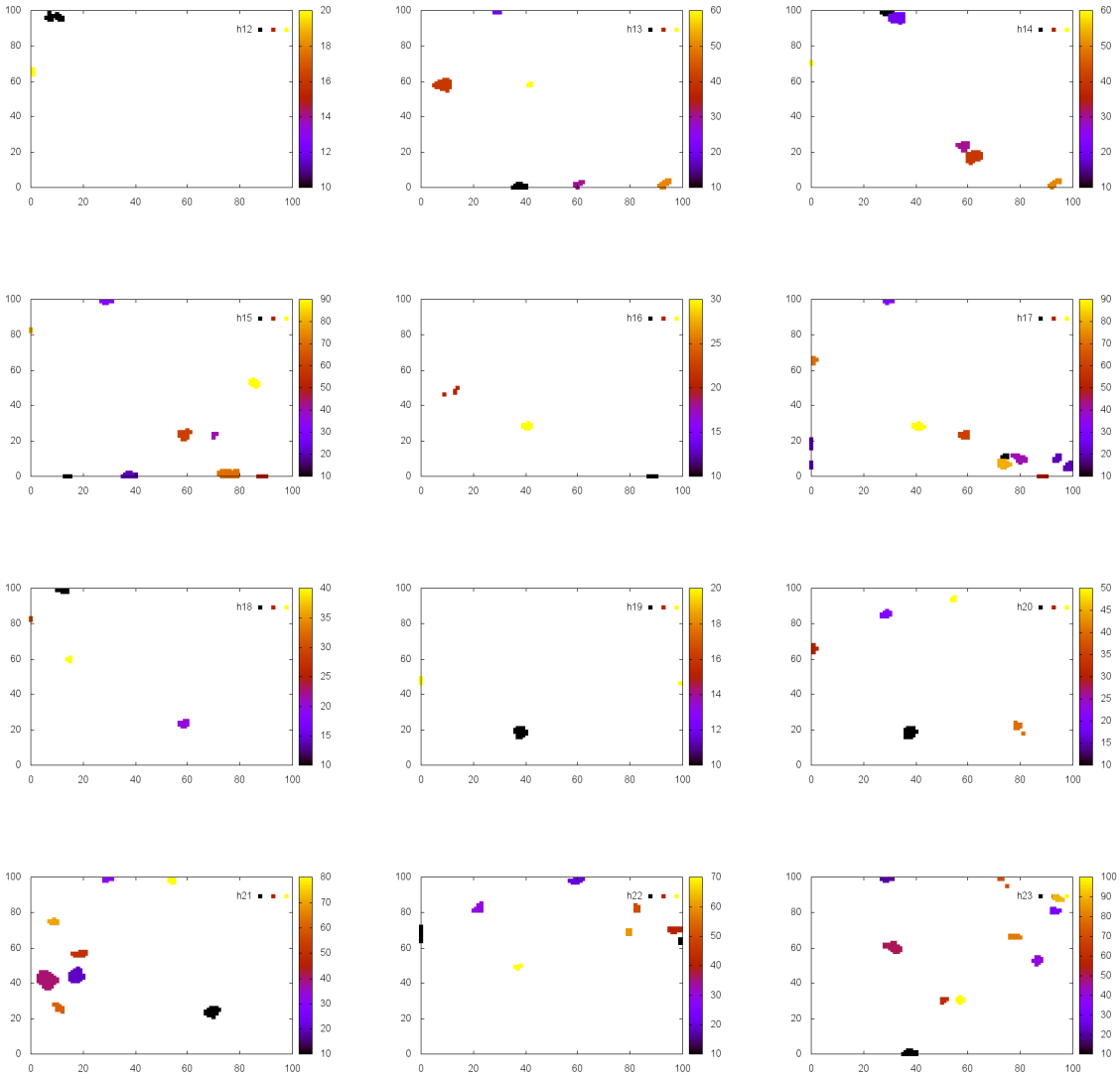

Figura 8: Stable, Taglio 95-esimo percentile, h 12-23. Da vedere da sinistra verso destra e dall'alto verso il basso

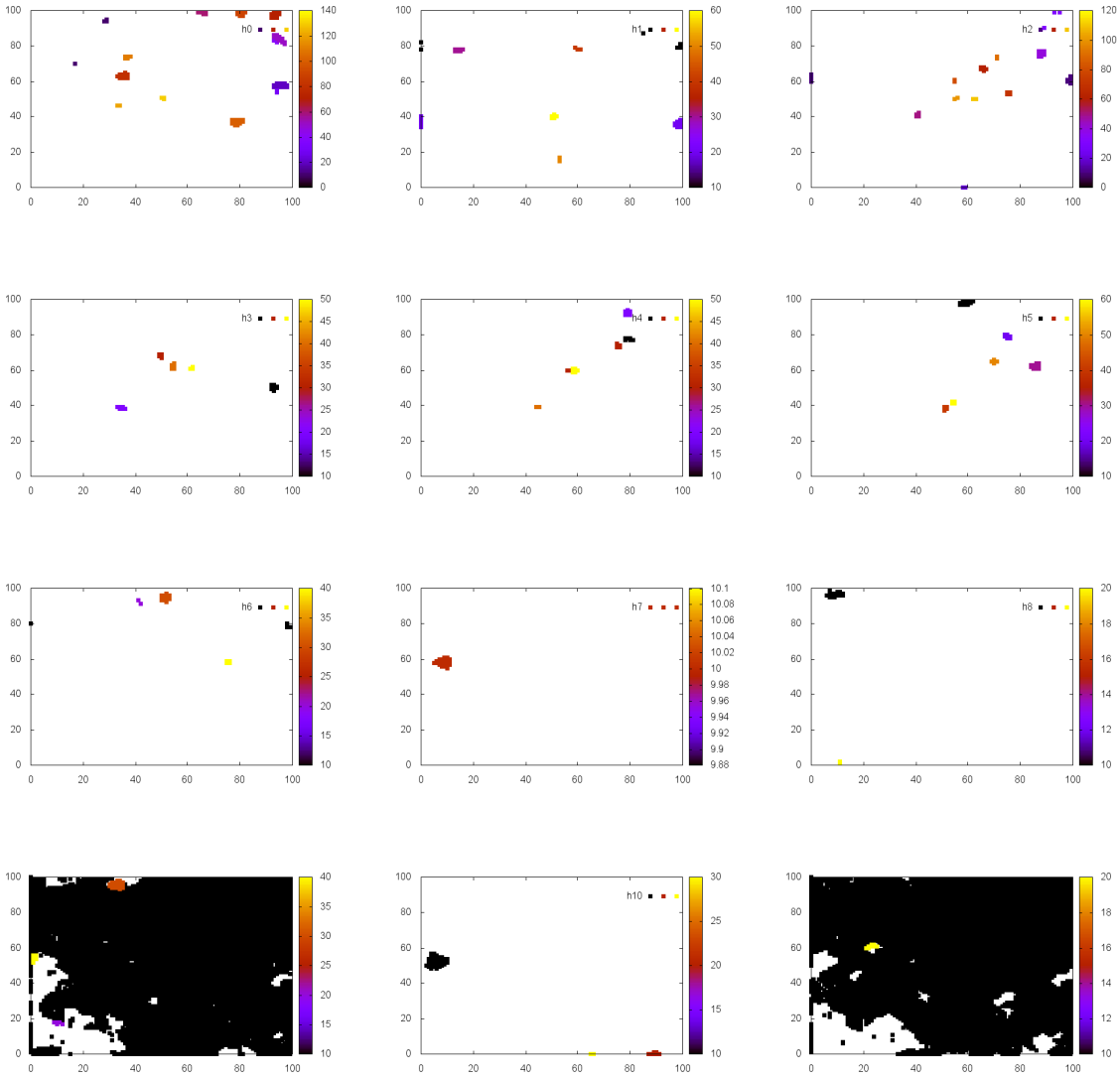

Figura 9: Stable, Taglio 90-esimo percentile, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

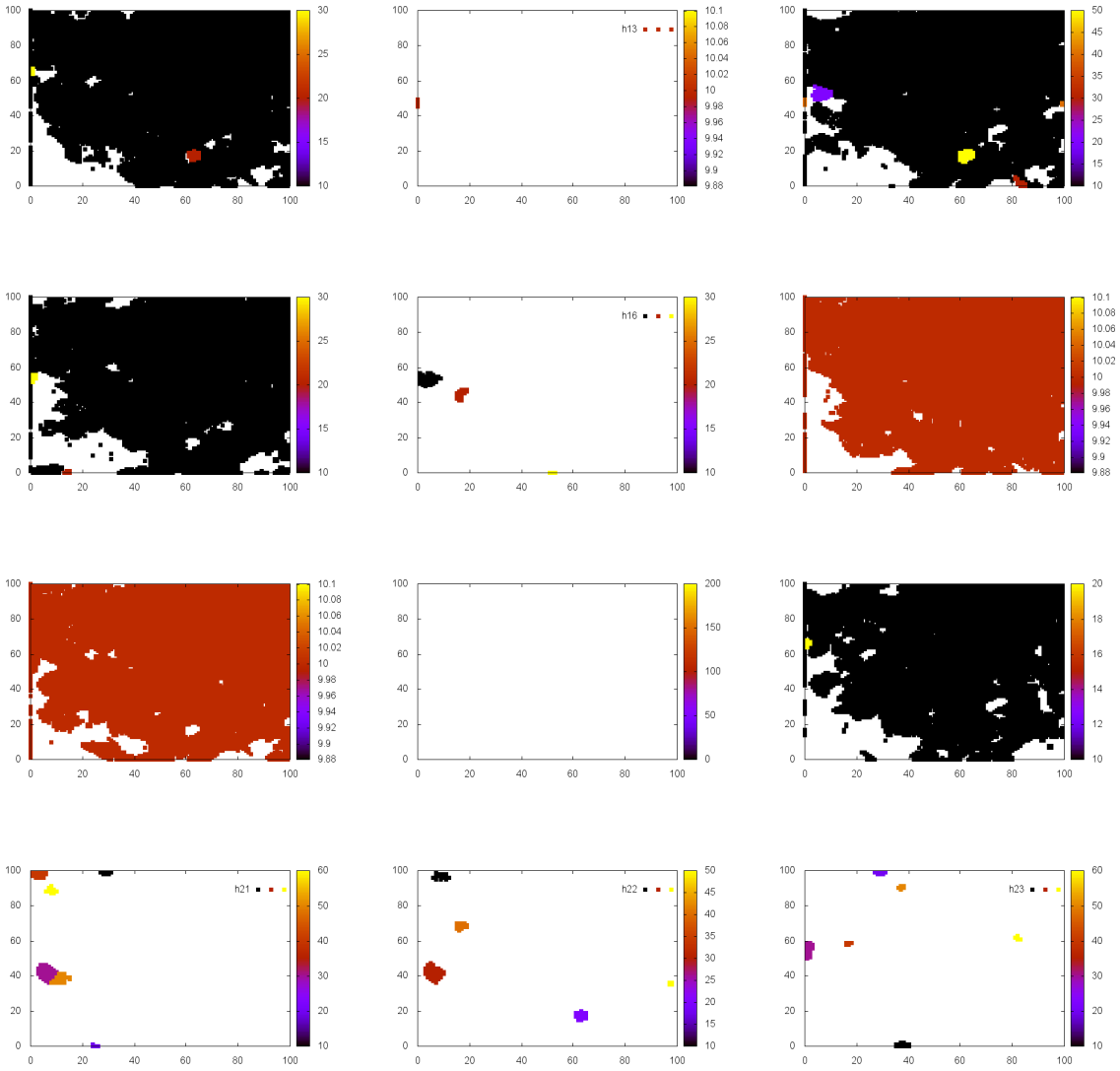

Figura 10: Stable, Taglio 90-esimo percentile, h 12-23. Da vedere da sinistra verso destra e dall'alto verso il basso

5 Prosecuzione del lavoro

1. Implementazione di tutta l'analisi su Hadoop.
2. Cercare le componenti debolmente connesse.
3. Strategie di clustering.
4. Ampliare il periodo analizzato per includere tutti i giorni di novembre e dicembre (da decidere come aggregare i risultati giornalieri)
5. affinare il periodo di aggregazione? (e.g. passare a fasce di 30 min)