Analisi aggregata per ore

Michele Carignani, Alessandro Lenzi

5 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 effettuare la ricerca delle componenti fortemente connesse (in seguito CFC) è stato utilizzato l'algoritmo Tarjan, il cui pseudocodice (tratto da Wikipedia, Tarjan's Strongly Connected Components algorithm) è mostrato in 1. L'approccio iniziale è stato quello di utilizzare una semplice visita ordinata sull'identificatore dell'arco (i.e. `for i = 0 to 10000 do strongconnect(i)`). Tale visita è stata effettuata con diversi tagli. In tutti i vari esperimenti, è stato possibile notare che l'ordine di visita dei nodi ha una notevole influenza sulle CFC trovate. Ad esempio, effettuando un taglio *semplice* (per la spiegazione riguardo tagli si veda 2.1) al valore 0.005, si trova in pressoché tutte le ore una componente fortemente connessa molto grande, con nodi con valori nelle prime decine e che arrivano fino a 10000, per poi trovarne alcune altre estremamente piccole. É evidente che in questo caso il taglio è estremamente basso, e perciò non necessariamente significativo: quello che invece viene messo in evidenza è che l'ordine di visita del grafo ha una notevole influenza sulle CFC individuate, come si può capire dal primo `for` nello pseudocodice in 1.

Il secondo passagio è stato dunque quello di ottenere una **strategia di visita** che potesse essere consistente con il traffico effettivo nella giornata. Lo strumento al quale abbiamo pensato è stato quello di utilizzare un secondo dataset, ovvero 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.

```
input: graph G = (V, E)
output: set of strongly connected components (sets of vertices)
index := 0
S := empty
for each v in V do
  if (v.index is undefined) then
    strongconnect(v)
  end if
end for

function strongconnect(v)
  v.index := index
  v.lowlink := index
  index := index + 1
  S.push(v)
  for each (v, w) in E do
    if (w.index is undefined) then
      strongconnect(w)
      v.lowlink  := min(v.lowlink, w.lowlink)
    else if (w is in S) then
      v.lowlink  := min(v.lowlink, w.index)
    end if
  end for
  if (v.lowlink = v.index) then
    start a new strongly connected component
    repeat
      w := S.pop()
      add w to current strongly connected component
    until (w = v)
    output the current strongly connected component
  end if
end function
```

Figura 1: Pseudocodice dell'algoritmo di Tarjan per la ricerca delle CFC

Inoltre, è stato generato un ulteriore file in cui, per il giorno analizzato, viene mostrato il traffico totale uscente da ogni griglia, sempre nel formato di cui sopra.

I file generati sono stati dunque impiegati al fine della definizione di un ordinamento nella visita. Il primo approccio è stato quello di compiere la visita con criteri diversi per ogni ora, per una cui analisi dettagliata si rimanda alla rispettiva sottosezione nel listato sottostante:

1. SCC1, visibile in **??** è una visita del grafo con nodi ordinati per **traffico orario crescente** e con archi della stella uscente selezionati per **probabilità crescente**
2. SCC2 visita i nodi per **traffico orario decrescente** ordinando gli archi per **probabilità crescente**
3. SCC3 visita i nodi per **traffico orario decrescente** e gli archi per **probabilità decrescente**
4. SCC4 visita i nodi per **traffico orario crescente** e gli archi per **probabilità decrescente**

Si noti come tale approccio in realtà si sia mostrato poco soddisfacente in tutti i quattro casi in analisi. Difatti, al fine della verifica di persistenza delle CFC individuate, una potenziale variazione di ordinamento nella visita avrebbe potuto comportare una distruzione delle CFC precedentemente individuate.

Il passo successivo è stato quindi quello di definire un ordinamento giornaliero sui nodi, al fine di effettuare delle visite possibilmente coerenti sui vari grafi orari generati. L'esperimento effettuato in questo caso può essere visto in **??**

2.1 Tagli

I tagli sono stati effettuati secondo due approcci distinti:

1. **tagli semplici**, in cui semplicemente tutti gli archi al di sotto del valore di probabilità specificato sono stati eliminati
2. **tagli percentili**[2], in cui i pesi degli archi nel grafo sono analizzati prima di effettuare la visita eseguendo un campionamento casuale con l'algoritmo di `Reservoir Sampling` (si veda Wikipedia: Reservoir Sampling) per un numero prefissato di al più 10^6 valori, sui quali è stata stimata la soglia del percentile desiderato per poi eliminare tutti i valori al di sotto della stessa.

Ci sono due diverse *ratio* dietro i due tipi di taglio utilizzati. Nel primo caso, l'idea è stata quella di eliminare tutti gli archi la cui probabilità fosse estremamente bassa: l'ipotesi è che se da una certa area la probabilità di chiamare l'altra è estremamente bassa, è poco probabile che siano correlate. Il problema di questo approccio è quello inerente i diversi *valori degli archi* all'interno della giornata. Se in ore di poco traffico, infatti, possiamo contare su pochi archi con probabilità estremamente alte, all'interno delle ore in cui il traffico è maggiore ritroviamo valori molto più concentrati intorno alla media, dando pertanto luogo a grafi molto più connessi anche nel caso di taglio come si può vedere in 3.2.1. Il taglio con i percentili è stato appunto immaginato per risolvere questo problema, introducendone però un secondo: quello della *casualità* del campione, che potenzialmente potrebbe influenzare i risultati ottenuti. Ciononostante si suppone che il campione di 10^6 valori (pari all'1% al caso pessimo, ma sovente quando il traffico è minore sufficiente a coprire la totalità degli archi) sia abbastanza elevato da attenuare tale influenza.

3 Risultati

3.1 Statistiche

Al fine di comprendere quanto fosse effettivamente possibile trovare le CFC, sono state effettuate delle analisi sui pesi degli archi. In questo frangente, il fatto che i pesi degli archi rappresentino in realtà delle probabilità può essere trascurato e si può considerare tale analisi come effettuatasi su una variabile aleatoria (in seguito anche v.a.) $Prob$, rappresentante i pesi degli archi in un certo grafo orario. Visto il dominio applicativo la v.a. è discreta e con valori nell'intervallo $[0, 1] \subset \mathbb{Q}$.[3] Si noti che $Prob$ può essere ritenuta a pieno titolo una v.a., poiché:

[2] sempre che questa sia la dicitura corretta

[3] Sappiamo che l'idea di calcolare la probabilità di probabilità può sembrare strana; in ogni caso possiamo pensare che, dal nostro punto di vista, l'esperimento in questione sia il fatto che un arco abbia una certo valore.

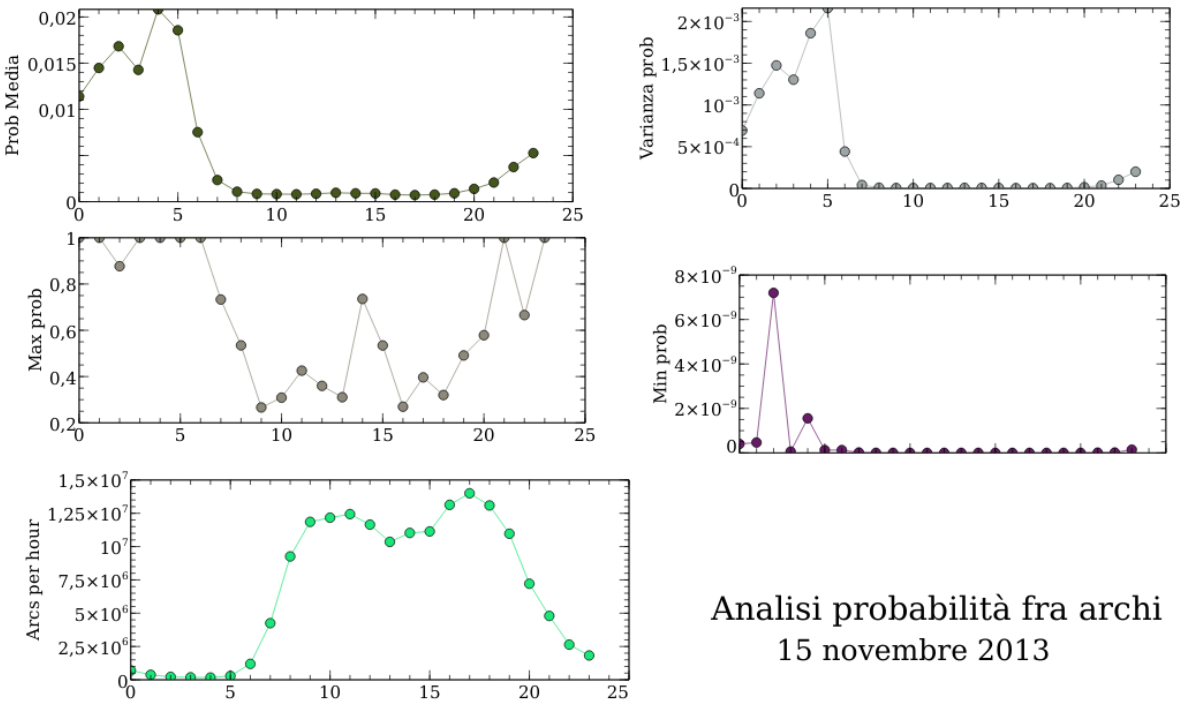

Figura 2: Statistiche sui pesi degli archi, 15 novembre. Sulle ascisse le fascie orarie. Da sinistra a destra e dal basso verso l'alto abbiamo (a)...(e)

1. $\forall x \in [0,1], \mathbb{P}(Prob = x) = \mathbb{P}(Prob = x^+)$
2. $\sum_{x\in[0,1]} \mathbb{P}(Prob = x) = 1$
3. $\forall x_1 < x_2 \in [0,1] \Rightarrow \mathbb{P}(Prob \leq x_1) \leq \mathbb{P}(Prob \leq x_2)$

In fig. 2 sono mostrate alcune statistiche sulla variabile aleatoria di cui sopra. Nella figura *(a)* è stato evidenziata la media empirica della v.a. per ogni ora del 15 novembre. Si noti come durante la notte i valori medi siano sensibilmente più alti, a riprova del fatto che il numero di archi nella stella uscente di ogni nodo è minore in tali grafi orari. Tale ipotesi viene suffragata dal grafico mostrato in *e*, in cui viene mostrato il numero di archi totali per ogni grafo orario.

In *(b)* abbiamo invece scelto di evidenziare la varianza della v.a. *Prob*. Si noti che la varianza è estremamente basse negli orari lavorativi, mostrando come in questo caso i molti archi uscenti mostrati in *(e)* tendano a stabilizzarsi sul valore medio.

In *(c), (d)* mostriamo invece i massimi valori assunti dagli archi su ogni grafo orario. Di particolare interesse è *(c)*, da cui si nota come anche nelle ore di traffico molto elevato esistano dei picchi notevoli nei valori degli archi. I minimi rimangono sempre invece nell'ordine di 10^{-9}, mostrando comunque un incremento considerevole, seppur in scala, nelle ore notturne.

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:

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

Il valore della funzione z non ha alcun significato: è stato semplicemente utilizzato al fine di assegnare differenti colori a differenti CFC; la costante 10 è un *numero magico* che consente una differenza cromatica

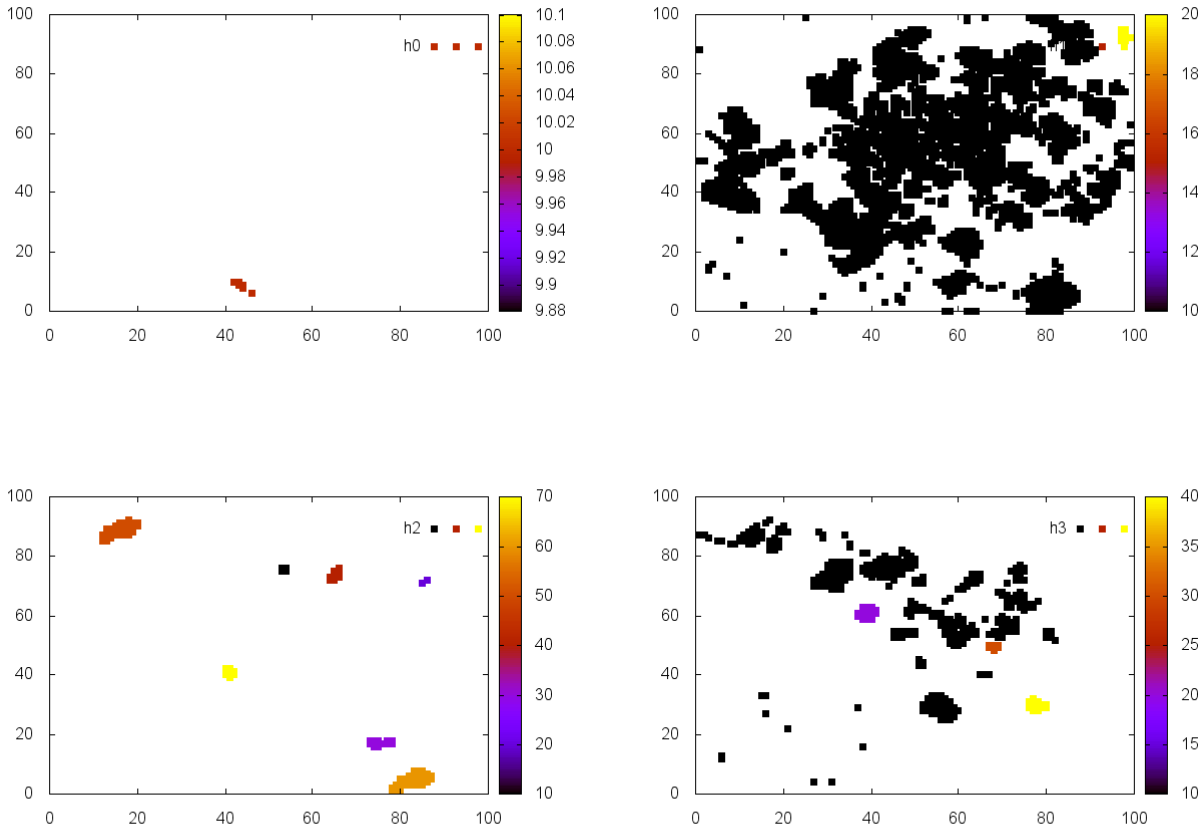

Figura 3: comportamento nelle ore notturne (0-3) con SCC1 e taglio a 0.005

sufficiente a poter individuare differenti CFC in maniera pressoché immediata. Si noti come i grafici disegnino unicamente le celle appartenenti a delle CFC individuate e con cardinalità maggiore di uno e lo facciano nella posizione corretta secondo la griglia di separazione definita nei dataset.

3.2.1 SCC1

Come esempio per tutti i casi in cui l'ordine di visita dei nodi è indotto dal traffico orario, riportiamo il caso SCC1 con taglio degli archi con probabilità minore o uguale a 0.005. In questo caso, le componenti sono per la maggior parte delle ore estremamente ampie, come si può vedere in seguito in 3 e **??**. In particolare in 3 il grafo è probabilmente composto per la maggiorparte da archi di probabilità piuttosto bassa, per cui si ottengono molto semplicemente delle CFC piuttosto distinte. Invece in **??** si vede come il taglio si riveli insufficente e le componenti individuate coprano buona parte dell'area presa in considerazione. La notte il comportamento è simile a quello mostrato in figura. Alle 6 del mattino si nota invece una completa assenza di componenti, mentre dalle 8 alle 23 il comportamento è pressoché il medesimo di quello mostrato in **??**. Si noti in 5 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 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. Questo può essere visto anche da 2*(a), (c), (e)*. Il comportamento in questo caso non è quindi stato considerato soddisfacente, visto l'eccessivo sbilanciamento nelle dimensioni delle CFC e la

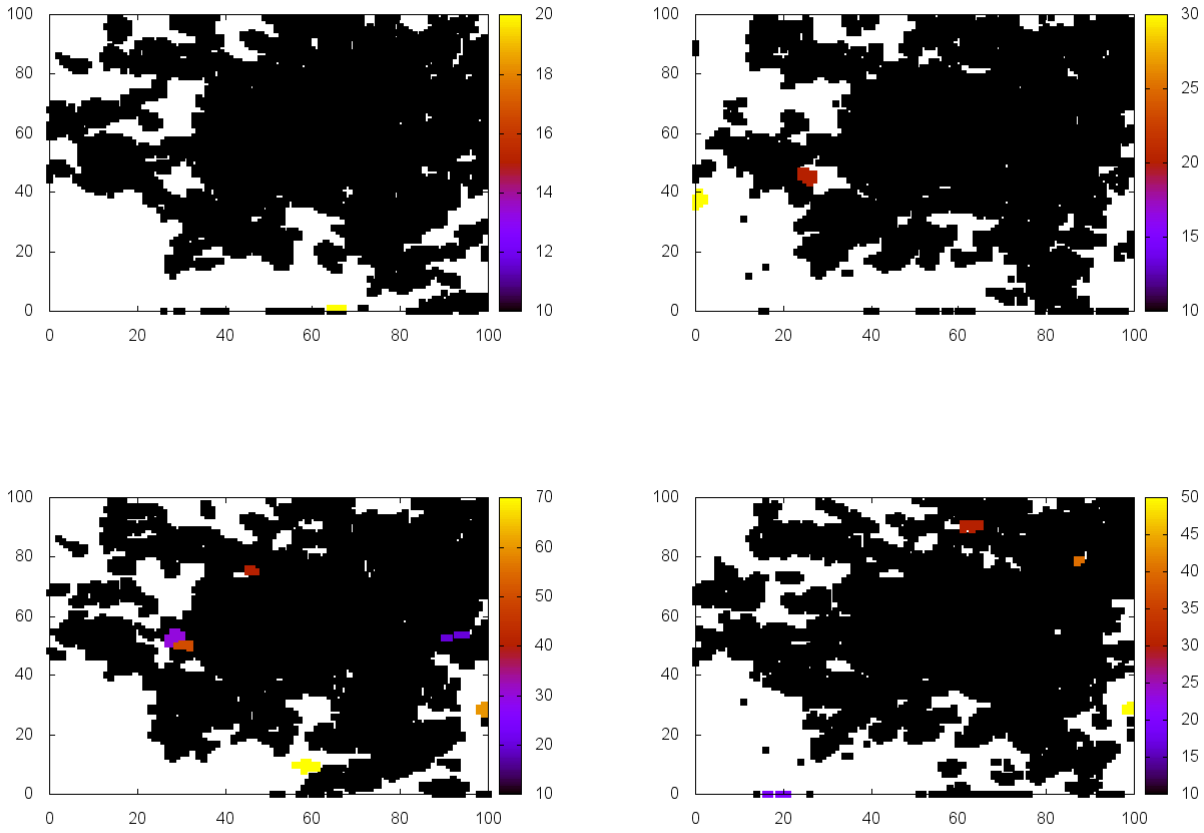

Figura 4: comportamento nelle ore diurne (12-15) con SCC1 e taglio a 0.005

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

difficoltà di individuarne una logica o dei confini a livello geografico. Con tagli maggiori i risultati non si rivelano differenti. In si può vedere il comportamento notturno e in quello diurno:

Le componenti individuate nel taglio maggiore non sempre sono sottoinsiemi di quelle individuate in precedenza: anzi, ciò avviene in generale di rado e quando la CFC individuata ha come archi a se interna principalmente quelli con probabilità più elevata. Come si può vedere nell'algoritmo 1, infatti, viene svolta una visita di profondità. I nodi inoltre possono essere visitati una sola volta e una sola volta sono considerati come possibili appartenenti a una componente fortemente connessa.

Dato il grafo senza tagli $G = (V, E)$ e sia $P : V \times V \to [0, 1]$ la funzione che assegna un valore di probabilità agli archi, definiamo $P_{cut}(u, v) = \begin{cases} P(u, v) & \text{se } P(u, v) > cut \\ 0 & \text{altrimenti} \end{cases}$

Sia ora $G_{cut} = (V, E')$ il grafo tagliato, in cui $E' = \{(u, v) \in E : P_{cut}(u, v) > 0\}$. Definiamo ora la CFC ottenuta partendo da un nodo $v \in V$ come $C_{cut}(v) = W \subseteq V.W$ è una CFC in G_{cut}. Infine sia $T_{cut_{dfs}}(v) = T \subseteq V.T$ è l'insieme degli archi nell'albero generato dalla visita dfs in G_{cut} originata in v e $\alpha_{cut}(v, e) = \{v_0, v_1, v_2, ..., v_{n-1}, v_n = e : \forall i = 0...n - 1.(v_i, vi + 1) \in E'\} \subseteq T_{cut_{dfs}}$ la sequenza di archi orientati che conduce dal nodo v al nodo e.

Sia $a, b \in G.a < b$, dove il $<$ significa che a viene visitato prima di b. Supponiamo inoltre che esista $C_{c_2}(b)$ con cardinalità maggiore di 1; sia infine $c_1 < c_2$, è possibile trovare un esempio che dimostri che:

$$\exists x \in V.\alpha_{c_1}(a, x) \cap C_{c_2}(b) \neq \varnothing \Rightarrow C_{c_2}(b) \neq C_{c_1}(b) \wedge |C_{c_2}(b)| > |C_{c_1}(b)|$$

In poche parole, è possibile che un taglio su un valore più basso possa reintrodurre nella ricerca delle CFC un cammino con almeno un arco il cui valore di P, contenuto tra c_1 e c_2, porti alla rimozione di nodi o addirittura alla disturzione della CFC individuata con un taglio su un valore maggiore.

Nel nostro contesto, anche minime variazioni nel taglio possono reintrodurre un gran numero di cammini e questo spiega perché non sempre si possono ritrovare su tagli con valori più bassi tutte le CFC individuate con tagli su valori più alti, che sarebbero intuivitamente più selettivi.

Una possibile soluzione a questo problema potrebbe essere individuare una variazione a 1 per cui una precedente visita di un nodo non lo escluda successivamente dalla ricerca di ulteriori CFC, **pur**

mantenendo le CFC disgiunte. In questo frangente presumibilmente la relazione di inclusione varrebbe.

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.

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)

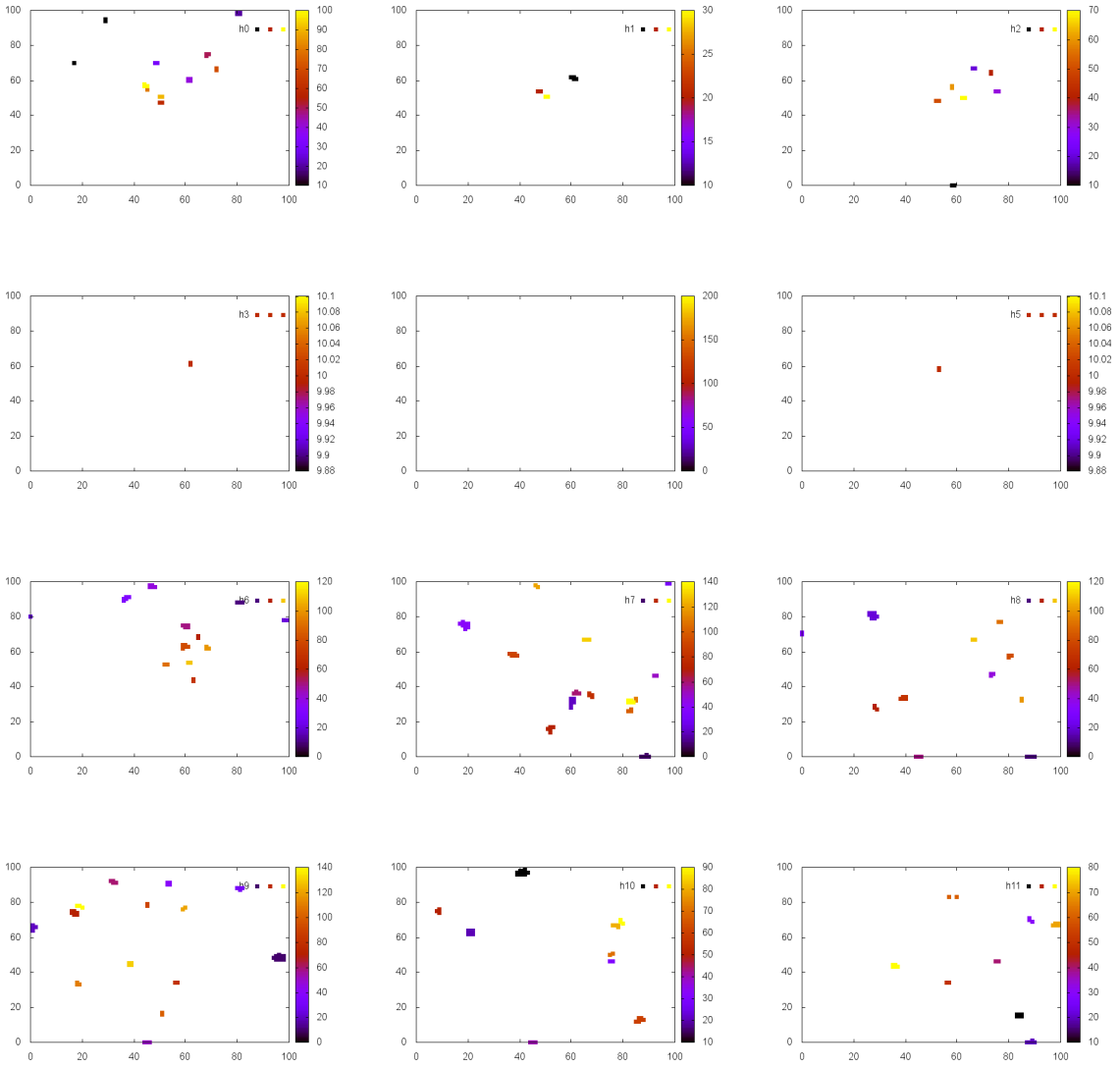

Figura 6: Stable, Taglio 99-esimo percentile, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

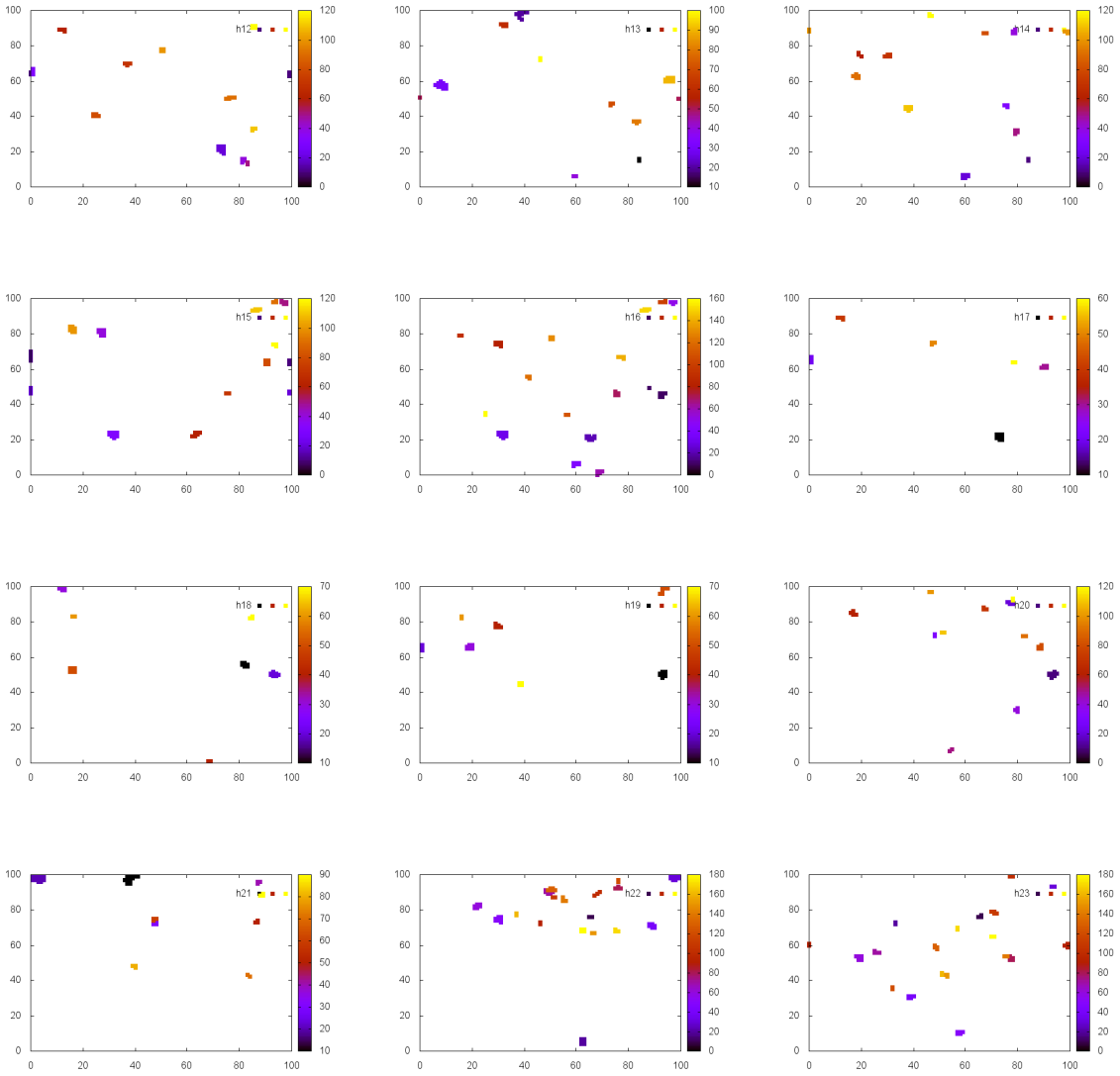

Figura 7: Stable, Taglio 99-esimo percentile, h 12-23. Da vedere da sinistra verso destra e dall'alto verso il basso

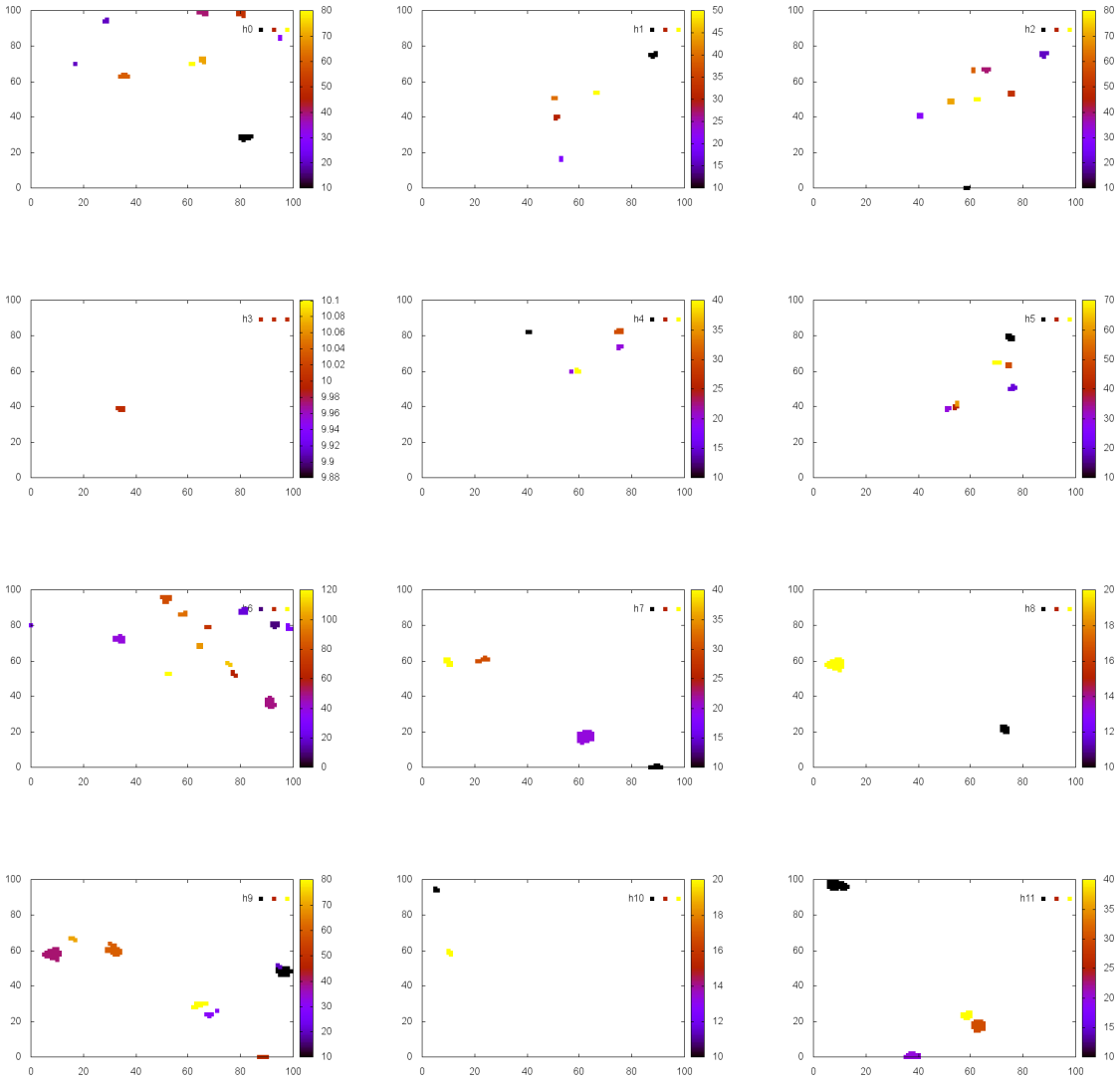

Figura 8: Stable, Taglio 95-esimo percentile, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

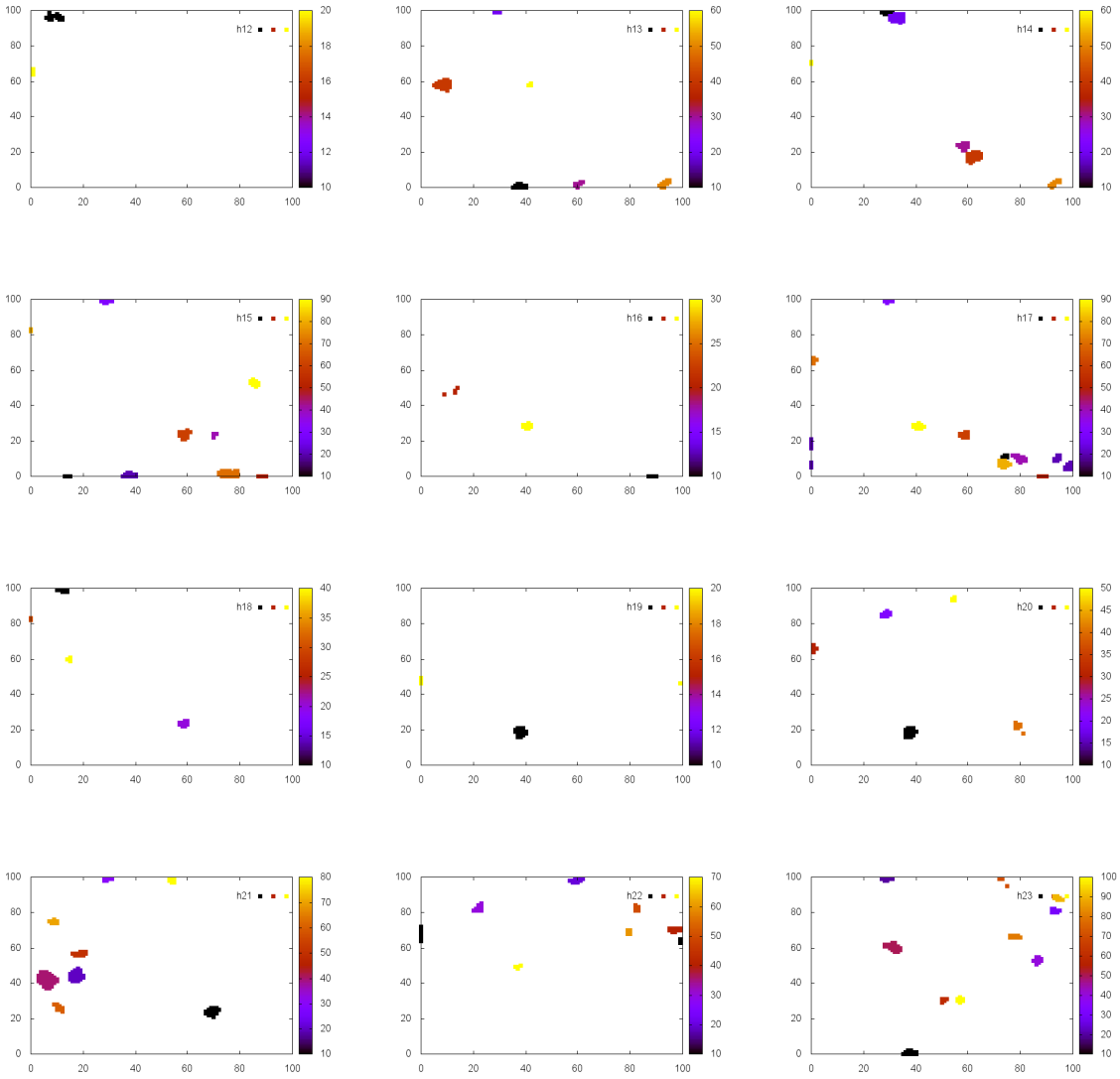

Figura 9: Stable, Taglio 95-esimo percentile, h 12-23. Da vedere da sinistra verso destra e dall'alto verso il basso

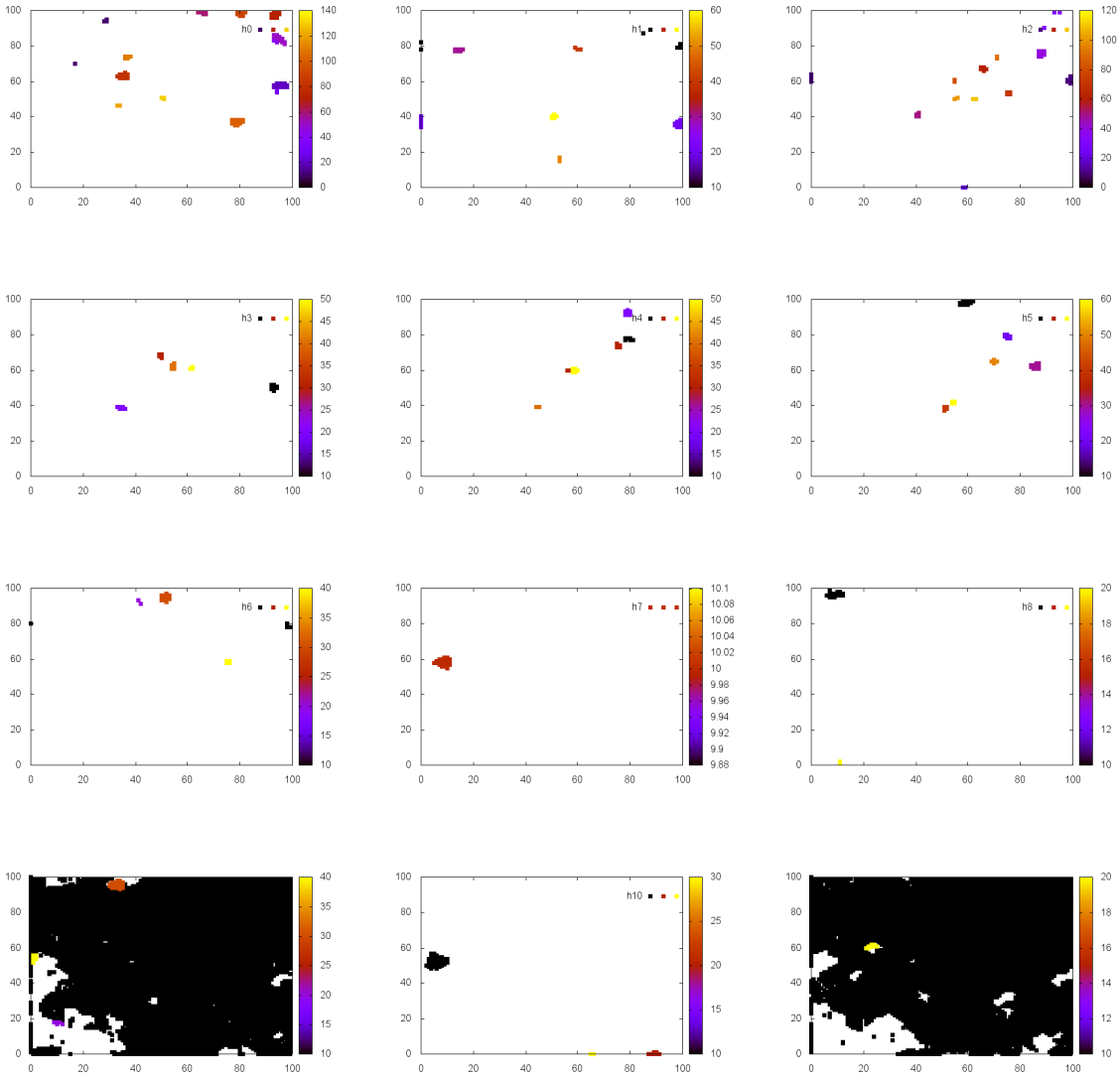

Figura 10: Stable, Taglio 90-esimo percentile, h 0-11. Da vedere da sinistra verso destra e dall'alto verso il basso

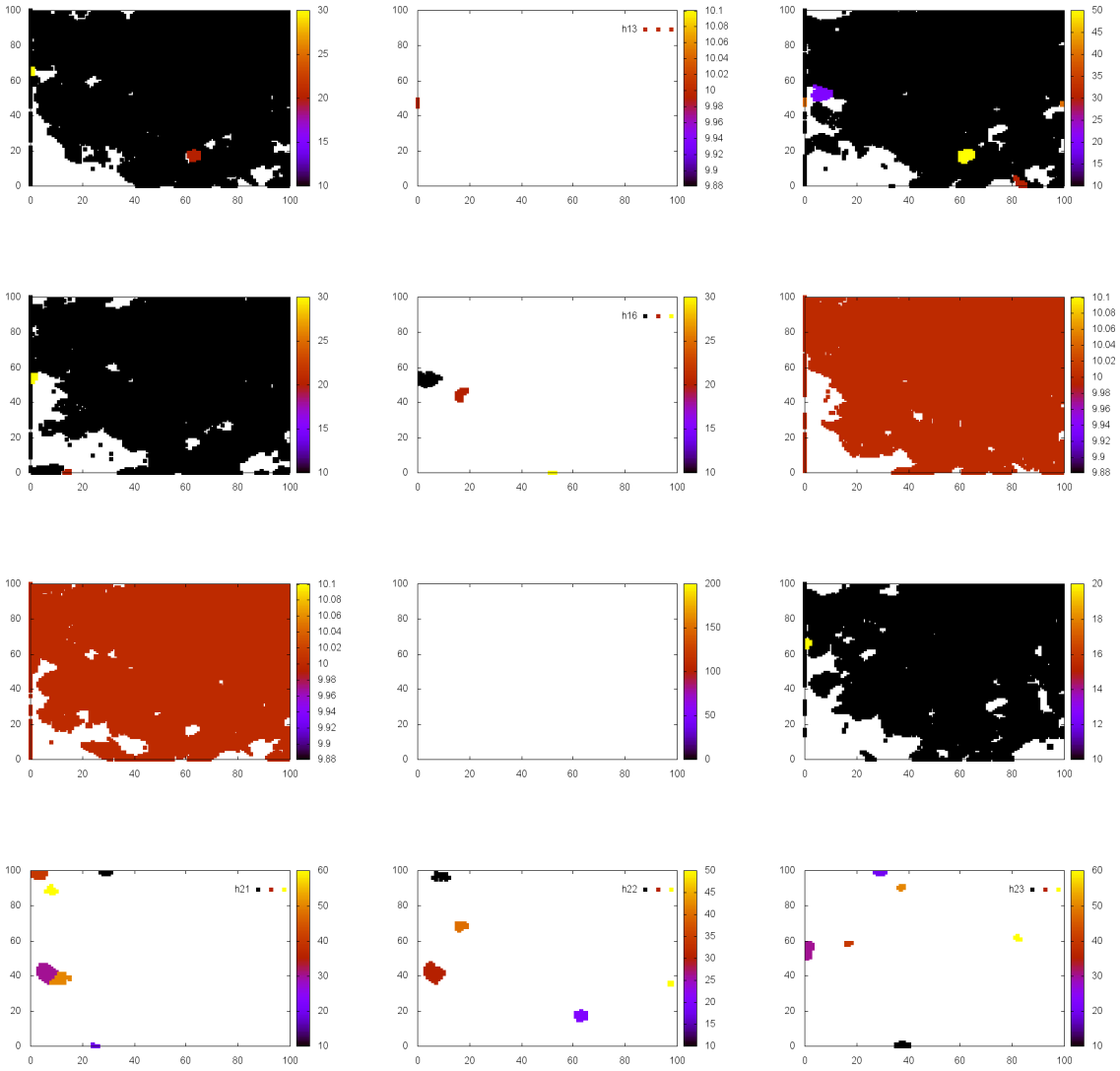

Figura 11: Stable, Taglio 90-esimo percentile, h 12-23. Da vedere da sinistra verso destra e dall'alto verso il basso