# Analisi della dipendenza tra eventi distribuiti su uno spazio monodimensionale tramite test di Hopkins e studio delle distanze tra eventi consecutivi

Tommaso Di Luciano

## I. INTRODUZIONE

L'obbiettivo di questo report è quello di definire e testare un procedura attraverso la quale analizzare l'interazione tra eventi distribuiti su uno spazio monodimensionale e caratterizzarli come eventi indipendenti o dipendenti. Un gran numero di fenomeni può essere infatti modellizzato come eventi distribuiti lungo uno spazio rappresentabile da un segmento (basti pensare ad un qualsiasi evento localizzato nel tempo) e spesso sapere se la realizzazione di un evento in un punto dello spazio influenza le successive realizzazioni può rivelarsi un informazione importante.
Un fenomeno caratterizzato da eventi indipendenti è ben modellizzato da un **processo omogeneo di Poisson (HPPP homogeneous poisson point process)** che in linea con l'ipotesi di indipendenza presenta una distribuzione degli eventi uniforme nello spazio. Eventi dipendenti invece vengono modellizzati attraverso due fenomeni distinti: **clustering** nel caso in cui tendano ad aggregarsi o **eventi equidistanziati** nel caso in cui tendano a respingersi all'interno dello spazio.
Viene inoltre effettuata un analisi di dati reali attraverso suddetta procedura.

## II. ANALISI

Per analizzare una distribuzione di eventi su spazio monodimensionale ho scelto di effettuare sui samples di dati il **test di Hopkins**, per poi andare a studiare il comportamento della distribuzione in maniera più approfondita verificando che rispettasse le proprietà dei modelli testati.

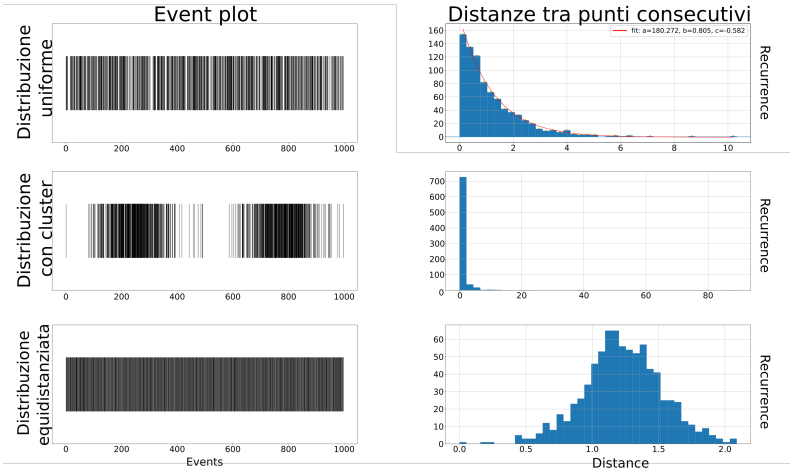


FIG. 1 Nel caso di HPPP (riga 1) la distribuzione delle distanze ha decrescita esponenziale
Nel caso di Clustering (riga 2) la distribuzione delle distanze è piccata verso l'origine
Nel caso di eventi equidistanziati (riga 3)la distribuzione delle distanze è molto stretta e piccata attorno al valore del periodo che idealmente intercorre tra gli eventi

### A. Hopkins Test

Il test di Hopkins restituisce un parametro H che idealmente assume:

- valore 1 in corrispondenza di fenomeni di clustering
- valore 0.5 in corrispondenza di distribuzioni uniformi
- valore 0 in corrispondenza di distribuzioni di eventi equidistanziati nello spazio (caso limite di eventi tra loro repulsivi)

**Definizione:** Sia X un set di n punti. Consideriamo un random sample di $m << n$ punti con membri $x_i$.
Generiamo un set Y di m punti distribuiti random uniformemente.
Definiamo due misure di distanza:

- $u_i$ la distanza di $y_i \in Y$ dal suo vicino più prossimo in X
- $w_i$ la distanza di m numeri di $x_i \in X$ scelti randomicamente dal loro vicino piu' prossimo in X

se i dati sono d-dimensionali, il test di Hopkins e' definito come: $H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d}$

### B. Distribuzioni uniformi corrispondono ad un valore H≃0.5

Si osserva dalle simulazioni effettuate che **il parametro H tende a distribuirsi attorno al valore 0.5** per distribuzioni uniformi.

Nel caso in cui il test di Hopkins fornisca un valore compatibile a quello di una distribuzione uniforme si procederà ad analizzare nel dettaglio la **distribuzione delle distanze tra punti consecutivi**: questa per un HPPP **decade esponenzialmente**, attraverso un fit vengono quindi ricavati i parametri relativi al decadimento esponenziale.
Inoltre perchè la distribuzione sia uniforme l'andamento del numero di punti al crescere della distanza considerata si deve comportare come una retta avente come coefficiente la probabilità che accada un evento in quel punto dello spazio, che in un HPPP è considerata costante.

### C. Fenomeni caratterizzati da Clustering corrispondono ad un valore H≃1

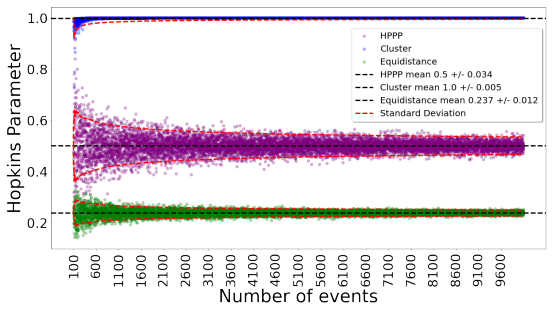


FIG. 2 Coefficiente di Hopkins calcolato da simulazioni di Processi di Poisson(viola), Clustering(blu) e eventi equidistanziati(verde). In rosso la deviazione standard che decresce circa come $1/\sqrt{N}$

Si osserva dalle simulazioni effettuate che **il parametro H tende a distribuirsi attorno ad un valore di poco inferiore ad 1**. Questo è probabilmente legato al fatto che non potendo produrre valori superiori ad 1 la distribuzione presenta una coda sola e la media risulta spostata verso il basso.
Nel caso in cui il test di Hopkins resituisca un valore compatibile con un fenomeno caratterizato da clustering si procede ad osservare come si dispongono le distanze tra i punti ordinate secondo la disposizione delle coppie da cui provengono: osservando la distanza tra un evento e il suo successivo, si registra la presenza di **zone caratterizzate da distanze molto ridotte intervallate da picchi di distanze significative (in relazione alle precedenti) che stanno ad indicare la zona di separazione tra cluster distinti** (se questi non si sovrappongono).
A questo punto **si applica l'algoritmo DBSCAN** per vedere quanti cluster emergono e come questi sono distribuiti nello spazio, regolandone i parametri (dispersione del cluster e numero minimo di eventi) in base ai picchi osservati sulle distanze.
Nel caso in cui il fenomeno sia ben modellizzato da cluster **la distribuzione delle distanze mostrerà un picco verso l'origine ed una rapida decrescita all'aumentare di queste**.
Analizzando il numero degli eventi registrati al crescere della distanza considerata si osserverà un grafico simile a discontinuità di tipo salto in corrispondenza di ogni cluster.

### D. Fenomeni caratterizzati da eventi equidistanziati corrispondono ad un valore H≃0.2

Si osserva dalle simulazioni effettuate che **il parametro H tende a distribuirsi attorno ad un valore di circa 0.2**. Questo è probabilmente legato al fatto che il coefficiente H non può produrre valori inferiori allo 0 dunque la media risulta spostata verso l'alto. Inoltre molto probabilmente le simulazioni sono state effettuate con una quantità eccessiva di rumore aggiunto che ha portato a questo innalzamento.
Nel caso limite di eventi equidistanziati **la distribuzione delle distanze tra punti consecutivi tende ad una distribuzione piccata nel periodo che intercorre tra due eventi e nulla altrimenti** (le distanze tra questi tendono cioè ad un valore costante).

### E. Valori di H non riconducibili a nessuno dei tre modelli

Nel caso in cui il test di Hopkins non produca alcun risultato riconducibile ai tre tipi di distribuzione utilizzati per modellizzare il fenomeno viene effettuato il test di Kolmogorov-Smirnov.
Se quest'ultimo produce un valore sufficiente a rigettare l'ipotesi di distribuzione uniforme viene analizzato il processo attraverso il modello che più si avvicina al valore di H ottenuto inizialmente.

## III. DA SIMULAZIONI CONTROLLATE EMERGE CHE IL PARAMETRO H CONVERGE AI VALORI ATTESI COME $1/\sqrt{(N)}$ E PER FENOMENI DI CLUSTERING RISULTA PIÙ EFFICACE IN PRESENZA DI DUE O PIÙ CLUSTER

Effettuando 10000 simulazioni per ognuno dei modelli ipotizzati per il comportamento di eventi su spazio monodimensionale, risulta che i valori del parametro di Hopkins tendono a distribuirsi per la maggior parte attorno ai 3 particolari valori medi sovracitati con una **dispersione** (distanza dai valori corrispondenti) **inversamente proporzionale al numero di eventi**. In particolare questa ha un comportamento **simile a** $1/\sqrt{N}$ come si può vedere in figura 2.
Inoltre nel caso di eventi distribuiti in cluster **il test di Hopkins restituisce risultati più efficaci in presenza di due o più cluster**, mentre in presenza di un unico cluster i valori delle simulazioni risultano molto dispersi e attorno ad un valore medio di circa 0.75. Questo è probabilmente dovuto al fatto che in presenza di un solo cluster il test è più propenso ad interpretare gli eventi come un HPPP con probabilità molto alta considerando l'intervallo più piccolo su cui si distribuisce quest'ultimo. Per questo motivo il test è stato applicato a simulazioni con 2 o più cluster.

### A. Metodo simulazioni

- *Processo omogeneo di poisson*:
  Nel simulare un processo omogeneo di Poisson vengono generati un numero N di eventi appartenente ad una distribuzione di Poisson con valore atteso $\lambda|W|$ dove $\lambda(x)$ è la probabilità che un evento accada alla posizione x, questa è costante in processi omogenei di Poisson, e $|W|$ indica la larghezza della finestra su cui si desidera ottenere la distribuzione.
  Per le simulazioni la probabilità che accada un evento è stata fissata a 0.8 per ogni punto dello spazio.
- *Cluster*:
  Per simulare processi di clustering sono stati prodotti un numero N di eventi con lo stesso metodo utilizzato per il caso HPPP (per poter confrontare distribuzioni con un numero di eventi paragonabile). Ogni simulazione è caratterizzata da un numero variabile di cluster distinti, in maniera tale da osservare il comportamento del test in presenza di più di un cluster. I singoli cluster sono stati prodotti attraverso distribuzioni normali con una sigma di molto inferiore alla distanza tra due cluster consecutivi. Sono stati testati più valori di sigma selezionati tramite questo criterio per osservare eventuali variazioni senza registrare particolari anomalie.
- *Eventi repulsivi*:
  Per simulare eventi di natura repulsiva è stato prodotto un numero N di eventi con lo stesso metodo utilizzato per il caso HPPP. Nella simulazione è stata fissata la frequenza con cui generare gli eventi all'interno dell'intervallo ed in seguito questi sono stati spostati di valori generati attraverso distribuzioni normali per simulare la presenza di rumore bianco all'interno del set di misure.

Per le simulazioni effettuate per testare l'andamento del parametro H sono state utilizzate sia simulazioni che generavano numeri di eventi partendo dalle dimensioni dell'intervallo come previsto da un processo omogeneo di Poisson, sia simulazioni che generavano le distribuzioni a partire dal numero di eventi desiderato.

#### 1. Possibili limiti nell'esecuzione delle simulazioni

Il metodo con cui sono state effettuate le simulazioni è fortemente influenzato da una serie di limiti:

- L'ipotesi di eventi repulsivi tende ad eventi equidistanziati solo come caso limite.
- I cluster prodotti da gaussiane sono solo una sottofamiglia dei possibili cluster producibili.

## IV. LA PAROLA "IL" SEGUE LOCALMENTE UNA DISTRIBUZIONE UNIFORME

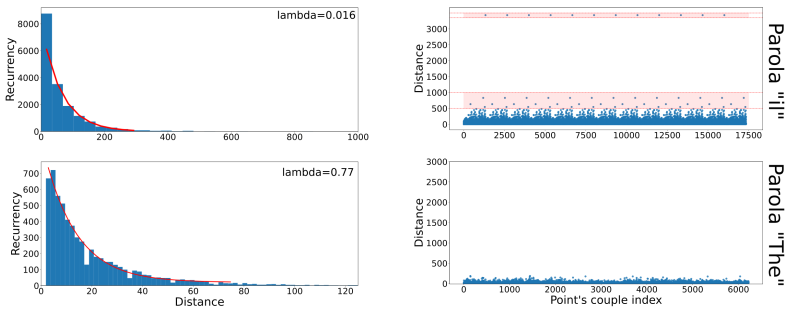


FIG. 3 Distribuzione delle distanze e posizione in ordine di apparizione. In particolare si può notare nello scatterplot di riga 1 come "Il" si distribuisca in zone nettamente separate dalle distanze evidenziate

Il dataset analizzato consiste nella distribuzione nel testo "Frammenti letterari e filosofici" di Leonardo Da Vinci della parola 'Il'.
Utilizzando la procedura sovracitata **il valore del test di Hopkins restituito si assesta attorno a 0.7** quindi non immediatamente riconducibile ad uno dei tre modelli analizzati. Si osserva che la **distribuzione delle distanze tra punti consecutivi ha un comportamento esponenziale decrescente**, ciò fa supporre che i dati si distribuiscano uniformemente. Osservando in ordine di apparizione le distanze tra punti consecutivi possiamo notare zone di **punti molto ravvicinati intervallate da picchi di distanza**, ma escludendo questi ultimi **il comportamento è di tipo uniforme** ed un analisi su ognuno di questi intervalli produce un valore di H corrispondente a distribuzioni uniformi.

La parola 'Il' si distribuisce quindi in maniera **uniforme su diversi sottointervalli tra loro separati**. Per altre parole testate si è visto che queste tendono ad equispaziarsi (probabilmente perchè si tende ad utilizzare dei sinonimi evitando ripetizioni nel testo) mentre l'articolo 'Il' viene utillizzato uniformemente nelle frasi. Le zone di separazione tra diverse distribuzioni uniformi sono probabilmente dovute al fatto che essendo 'il' un articolo maschile non viene utilizzato in frasi in cui il soggetto è sempre femminile o neutro, nè in soggetti maschili che prevedono articolo indeterminativo. Per testare questa ipotesi eliminando la differenza di genere è stata analizzata la parola 'The' in un testo inglese, in particolare il libro "A Connecticut Yankee in King Arthur's Court" di Mark Twain. Come si può notare in figura 3 la parola 'The' non presenta picchi marcati nella distanza ma si distribuisce su distanze molto più basse. Effettuando un fit esponenziale sulla distribuzione delle distanze (ipotizzando quindi che si tratti di un processo di poisson) si ottiene un **coefficiente $\lambda$ di 0.16 per "Il" e di 0.77 per "The"**, questo sta a rappresentare in un processo di Poisson la probabilità che un evento si realizzi.