

# Data Integration Methods

R. Urso

1 Giugno 2023

## 1 Introduzione

## 2 Data Fusion

## 3 Record Linkage

## 4 Statistical Matching

## 5 File Grafting

## 6 Riferimenti

# Progetto DORA

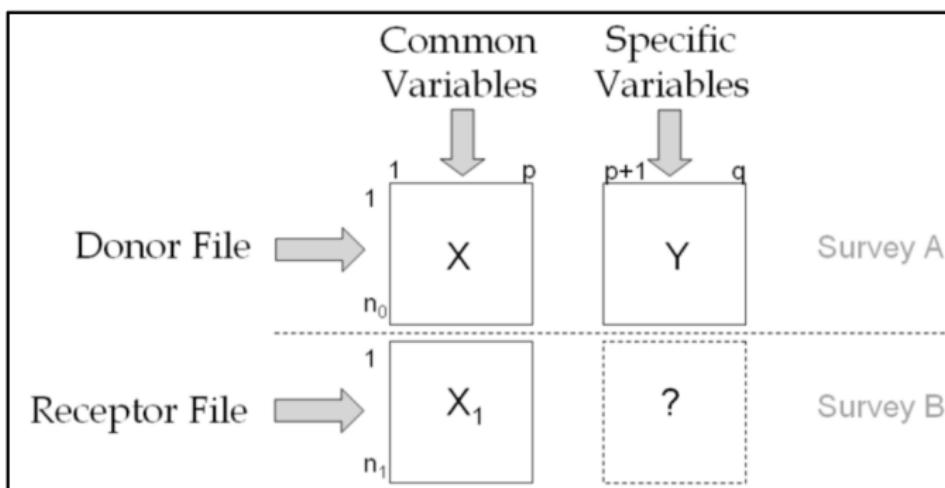
- Il progetto DORA<sup>1</sup> mira a combattere e prevenire la violenza contro i bambini (VAC), che rappresenta uno dei principali obiettivi dell'UE.
- I dati raccolti dall'UE non forniscono sufficienti informazioni, motivo per il quale l'analisi statistica non risulta immediata.
- Lo scopo è quindi quello di mettere a punto soluzioni statistiche che siano in grado di combinare, integrare e produrre dati "*utilizzabili*".
- Sono state adottati due differenti approcci per l'imputazione dei dati mancanti:
  - ① data fusion
  - ② approccio *model-based*.

---

<sup>1</sup>Acronimo di Data Integration for acknowledging risks and protecting children from violence

# Data Fusion

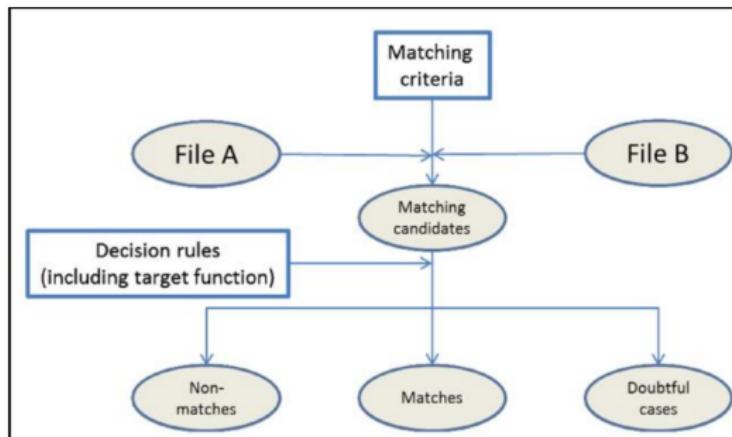
- Le tecniche di data fusion, sviluppate negli anni 60, nascono dagli studi di mercato, nei casi in cui il questionario presenta troppe domande.
- Il problema può essere formalizzato in due file di dati provenienti da fonti diverse: **ricettore-donatore**



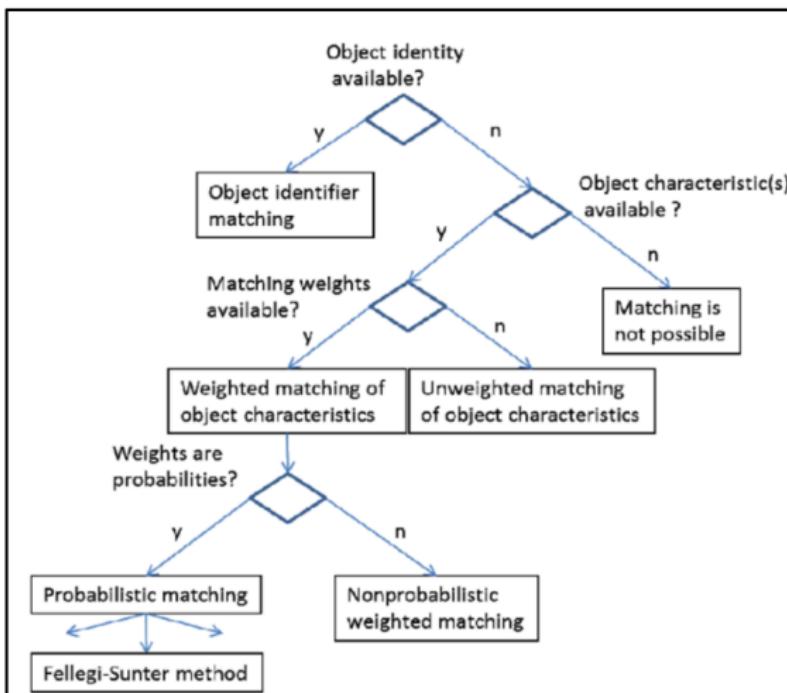
- Il file donatore contiene  $p + q$  variabili misurate su  $n_0$  unità statistiche, mentre il file ricettore include  $p$  variabili su  $n_1$  osservazioni, solitamente con  $n_0 > n_1$ .
- Le variabili  $X$  sono definite **comuni** (solitamente rappresentate da variabili socio-demografiche), mentre le  $Y$  **specifiche**. Obiettivo è trasferire le variabili specifiche dal donatore al recettore.
- Esistono diverse classificazioni riguardo le tecniche di data fusion: matching con certezza/incertezza, metodi impliciti/esplícitos e approccio parametrico/non parametrico.

# Record Linkage

- Se le variabili possono essere abbinate con certezza non è necessaria alcuna misurazione statistica, per cui abbiamo una corrispondenza esatta. Tale approccio (non parametrico) è denominato **Record Linkage**.
- Le componenti chiave nel Record Linkage sono il criterio di matching, la regola di decisione e le procedure di blocco.



- Per quanto concerne i metodi, invece, ritroviamo il matching identificatore oggetto/caratteristica, il matching ponderato/non ponderato e il matching probabilistico



# Statistical Matching

- Rispetto al Record Linkage, il **matching statistico** è utilizzato quando le variabili comuni differiscono, per cui sono necessarie misure di distanza per determinare corrispondenza più vicina.
- Quindi, se la corrispondenza è incerta, ritroviamo metodi esplicativi ed impliciti:
  - ① I metodi esplicativi riguardano la costruzione di un modello di regressione  $Y = f(X)$ , ma anche PLS o NN.
  - ② I metodi impliciti trovano il match più vicino tra ricevitore e donatore mediante il principio del k-nn (*metodo hot-deck*)
$$r(x_i) \simeq \text{ave}[r(x), x \in L(x_i)]$$
- Se è presente assunzione sulla distribuzione di probabilità congiunta della variabili, ci troviamo in un'ottica parametrica. In caso contrario, il processo è non parametrico.

## Approccio parametrico

- L'approccio parametrico riguarda un processo di stima della covarianza delle variabili specifiche (definito obiettivo **micro**) o dei coefficienti di regressioni (definito obiettivo **macro**) sul dataset "*fuso*".
- Per le variabili di tipo continuo vengono utilizzate le stime di massima verosimiglianza mediante algoritmo E-M; per variabili di tipo categoriale il modello log-lineare.
- Inconvenienti:
  - ① Non viene considerata la correlazione tra le variabili.
  - ② Possono essere prodotti risultati poco coerenti.

# Metodi esplicativi

Nel caso di approccio parametrico possiamo:

- ➊ Considerare la distribuzione multivariata  $f(X, Y|\theta)$ , con  $X$  e  $Y$  indipendenti, ovvero consideriamo dati mancanti di tipo MAR e MCAR.

Tale distribuzione può essere scomposta in:

$$f(X, Y|\theta) = f(Y|X, \theta_Y|\theta_X)f(X, \theta_X)$$

- ➋ Modellare direttamente tra le variabili  $X$  ed  $Y$  nel file donatore, per cui:

$$f(X, Y) = r(X) + \epsilon$$

Rileviamo due step:

- ① La concatenazione dei file
- ② Imputazione delle variabili mancanti con regressione.

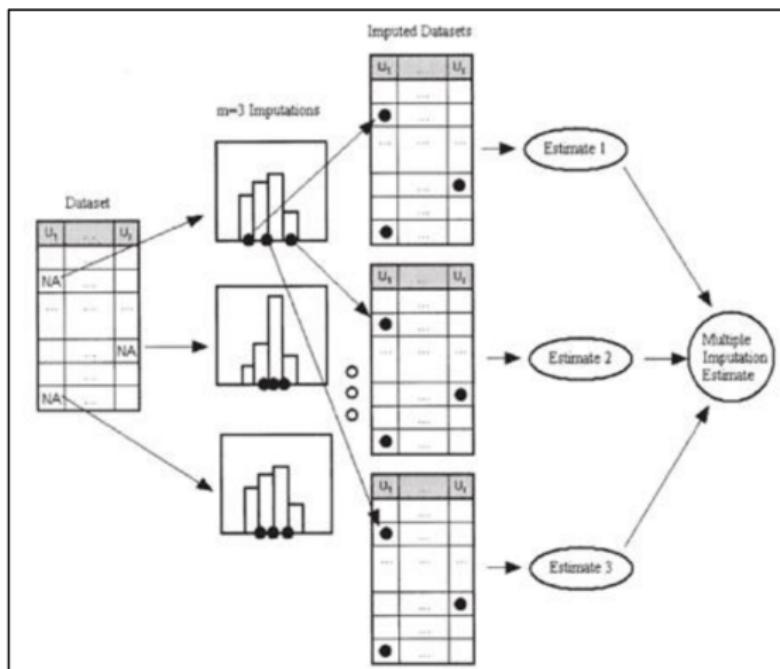
Unit no.	Common var Z								
	$Z_1$	$\dots$	$Z_k$	$X_1$	$\dots$	$X_q$	$Y_1$	$\dots$	$Y_p$
1				missing					
$\dots$									
$n_A$									
$n_A+1$							missing		
$\dots$									
$n$									

**File A:**  
 $U = (Z, X, Y)$   
 $U_{mis} = (X)$   
 $U_{obs} = (Z, Y)$

**File B:**  
 $U_{mis} = (Y)$   
 $U_{obs} = (Z, X)$

**k common variables Z**  
**q specific variables X**  
**p specific variables Y**

Rubin ha proposto un metodo di imputazione multiplo, dove a  $n$  imputazioni sono associati  $n$  dataset completi:



## Metodi impliciti

In ottica non parametrica, come visto in precedenza, registriamo il **k-nn**, in cui viene definita una misura di distanza tra le variabili. L'obiettivo è minimizzare questa distanza, ovvero cercare le coppie di  $(i, j)$  record di donatori e ricettori:

$$\sum_{j=1}^{n_B} \sum_{i=1}^{n_A} d_{ij} w_{ij} \quad w_{ij} \geq 0, i = 1, \dots, n_A, j = 1, \dots, n_B$$

dove:

- $i$  è il set di dati del donatore A
- $j$  è il set di dati del recettore B
- $d_{ij}$  è la distanza tra le variabili
- $w_{ij}$  è il peso della funzione distanza

Rodger nel 1984 ha presentato due tecniche basate sul  $k\text{-nn}$ :

- Matching non vincolato, in cui non ci sono restrizioni su quante volte il record del donatore debba essere utilizzato

File A				
Unit $i$	Weight $w_i^A$	$Z_1^A$	$Z_2^A$	$X$
A1	3	1	42	$x_1^A$
A2	3	1	35	$x_2^A$
A3	3	0	63	$x_3^A$
A4	3	1	55	$x_4^A$
A5	3	0	28	$x_5^A$
A6	3	0	53	$x_6^A$
A7	3	0	22	$x_7^A$
A8	3	1	25	$x_8^A$

File B								
Unit $j$	Weight $w_j^B$	$Z_1^B$	$Z_2^B$	$Y$	$Z_1^A$	$Z_2^A$	$Z_2^B$	$X$
B1	4	0	33	$y_1^B$				
B2	4	1	52	$y_2^B$				
B3	4	1	28	$y_3^B$				
B4	4	0	59	$y_4^B$				
B5	4	1	41	$y_5^B$				
B6	4	0	45	$y_6^B$				

Statistically matched file, recipient file A								
Matched unit $ij$	Weight $w_{ij}$	$Z_1^A$	$Z_2^A$	$Z_2^B$	Distance $d_{ij}$	$X$	$Y$	$Z_1^A$
A1, B5	3	1	42	41	1	$x_1^A$	$y_5^B$	
A2, B5	3	1	35	41	6	$x_2^A$	$y_5^B$	
A3, B4	3	0	63	59	4	$x_3^A$	$y_4^B$	
A4, B2	3	1	55	52	3	$x_4^A$	$y_2^B$	
A5, B1	3	0	28	33	5	$x_5^A$	$y_1^B$	
A6, B4	3	0	53	59	6	$x_6^A$	$y_4^B$	
A7, B1	3	0	22	33	11	$x_7^A$	$y_1^B$	
A8, B3	3	1	25	28	3	$x_8^A$	$y_3^B$	

dove  $Z_1^A$  sono definite variabile critiche, utili per definire la cd. **classe di matching**.

Il match, però, non conserva la distribuzione marginale dei dati originali e genera una distorsione. Utilizziamo per questo un altro matching

- Matching vincolato, che richiede la conoscenza dei pesi originari, per cui:

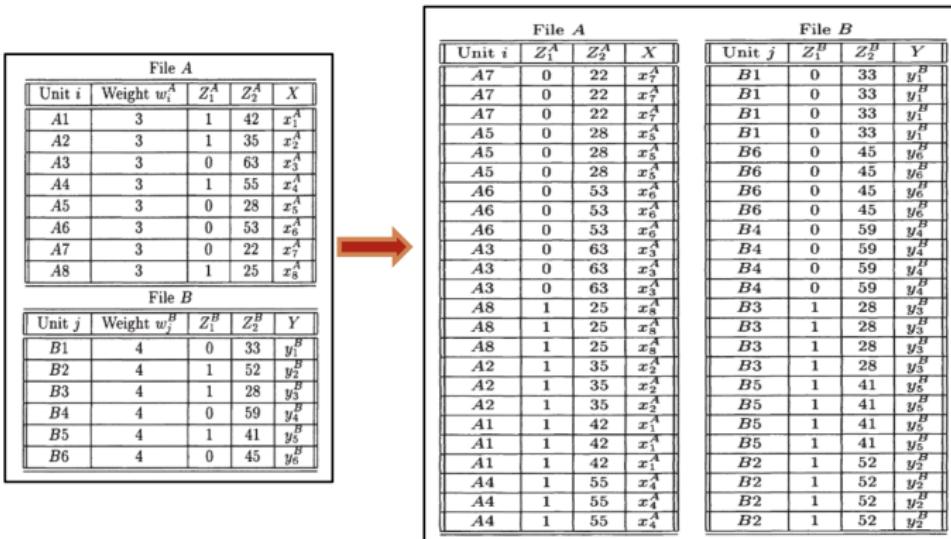
- $\sum_{j=1}^{n_B} w_{ij} = w_i \quad i = 1, \dots, n_A$
- $\sum_{i=1}^{n_A} w_{ij} = w_j \quad j = 1, \dots, n_B$

Step:

- ① "Esplosione" dei file originari: se un record in A ha peso  $w_{Ai}$ , esso viene ripetuto  $w_{Ai}$  volte
- ② "Abbinamento" dei file

Problemi:

- ① Non viene garantita la corrispondenza più vicina
- ② Costo computazionale elevato



Sia il matching vincolato che non vincolato, è soggetto all'ipotesi di indipendenza condizionata (CIA):

$$f(x, y, z) = f(y|x)f(z|x)f(x)$$

In alcuni casi, applicare una metrica di distanza potrebbe non essere semplice. È utilizzata la fusione in uno spazio fattoriale o massimizzando la coerenza interna.

- La fusione in uno spazio fattoriale concerne:
  - ① Effettuare un MCA su tutte le unità utilizzando metrica euclidea
  - ② Determinare la vicinanza tra donatore e recettore
  - ③ Scelta dei potenziali donatori
- La fusione massimizzando la coerenza interna riguarda, invece, la minimizzazione di una funzione di perdita:

$$\sigma(X, Y) = \frac{1}{m} \sum_{j=1}^m (X - G_j Y_j)' (X - G_j Y_j)$$

dove  $X = (G_1 | G_2 | \dots | G_m)$  è una matrice disgiuntiva e  $Y_j$  è la matrice delle coordinate delle categorie.

# File Grafting

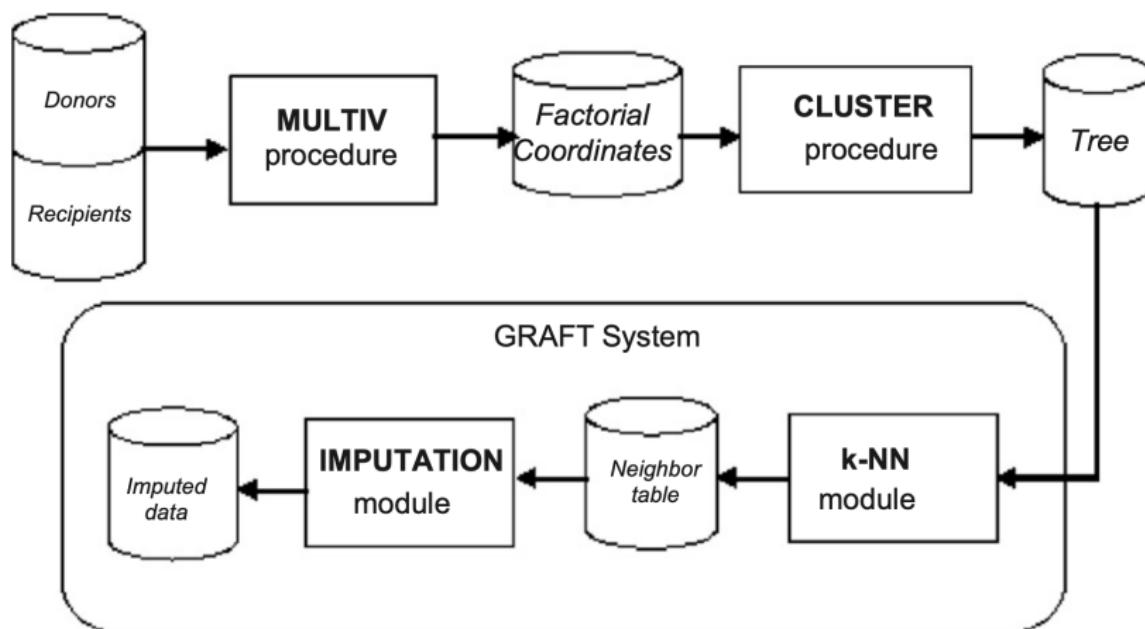
L'architettura GRAFT è un approccio che segue due sequenze di procedure:

- Procedura MULTIV, in cui donatore e recettore sono collocati in uno spazio comune.
- Procedura CLUSTER, in cui viene eseguito un algoritmo di clustering di tipo gerarchico con criterio di Ward.

A queste due procedure, segue il sistema GRAFT composto da 2 moduli:

- Modulo k-nn, in cui viene prodotto una tabella di vicini tra donatori e riceventi
- Modulo di imputazione, in cui vengono imputate le variabili nel file destinatario.

# Architettura GRAFT



L'algoritmo utilizzato è quello proposto da *Fukunaga e Narendra* nel 1975. Tale algoritmo segue due regole di pruning:

- Regola sui nodi dell'albero, ovvero un nodo  $t$  non può contenere un "vicino più vicino" se:

$$d(x_{1j}, knn) + r_t < d(x_{1j}, m_t)$$

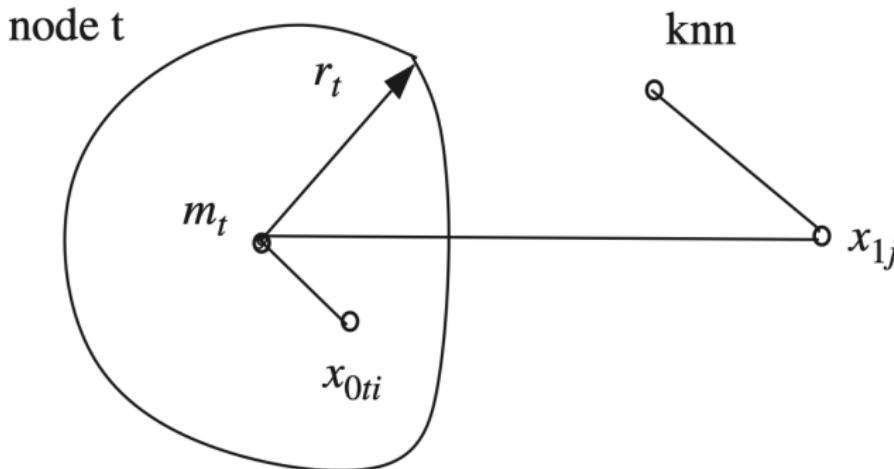
dove  $x_{1j}$  sono le coordinate del destinatario ed  $r_t$  è il suo raggio.

- Regola sugli individui dei nodi selezionati, per cui, per ogni nodo  $t$ ,  $x_{0ti}$  non può essere "il vicino più vicino" se:

$$d(x_{1j}, knn) + d(m_t, x_{0ti}) < d(x_{1j}, m_t)$$

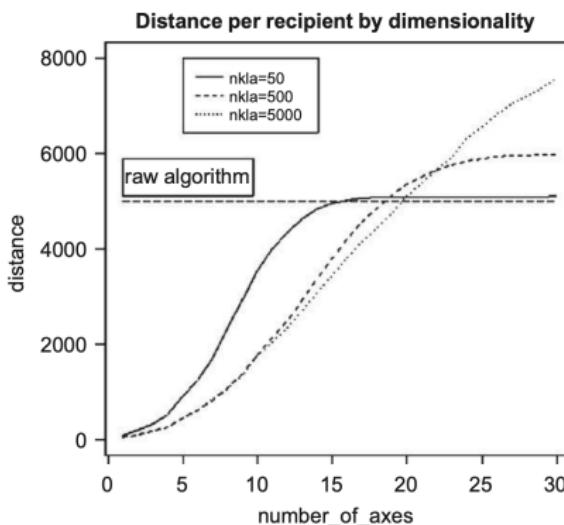
Quindi i parametri dell'algoritmo sono  $n_0, n_1$ , dimensionalità  $n_{axe}$ , numero di nodi terminali  $n_{kla}$  e numero di vicini  $n_{knn}$ .

Lo schema che viene seguito è così riassunto:



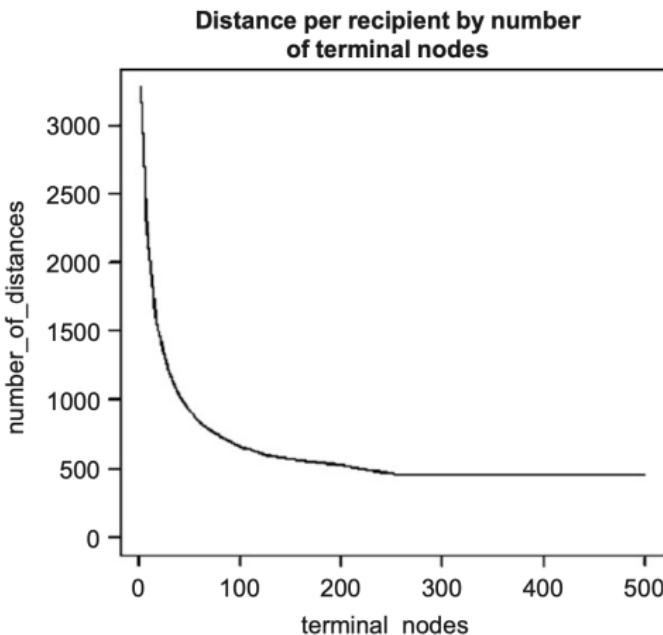
## Costo Computazionale

Il costo del modulo K-NN dipende molto dalla struttura dei dati.  
Se consideriamo  $n_{axe}$  notiamo che:



Il costo computazionale aumenta in maniera esponenziale all'aumentare della dimensionalità.

Invece, se consideriamo altro parametro quale il numero di nodi terminali  $n_{kla}$  emerge che:



In questo caso, il costo computazionale diminuisce all'aumentare dei nodi terminali.

# Riferimenti

- Gao, D., Srikuenthiran, S., Habib, K. N., Miller, E. J. DATA FUSION: TECHNIQUES AND APPLICATIONS.
- Aluja-Banet, T., Daunis-i-Estadella, J., Pellicer, D. (2007). GRAFT, a complete system for data fusion. Computational statistics Data analysis, 52(2), 635-649
- Saporta, G. (2002). Data fusion and data grafting. Computational statistics data analysis, 38(4), 465-473