# Mathematical Programming for Change Point Detection

Lisa Vecchi · Università di Bologna

*Mentor: Prof. Vittorio Maniezzo · Università di Bologna*

# Struttura della presentazione

01

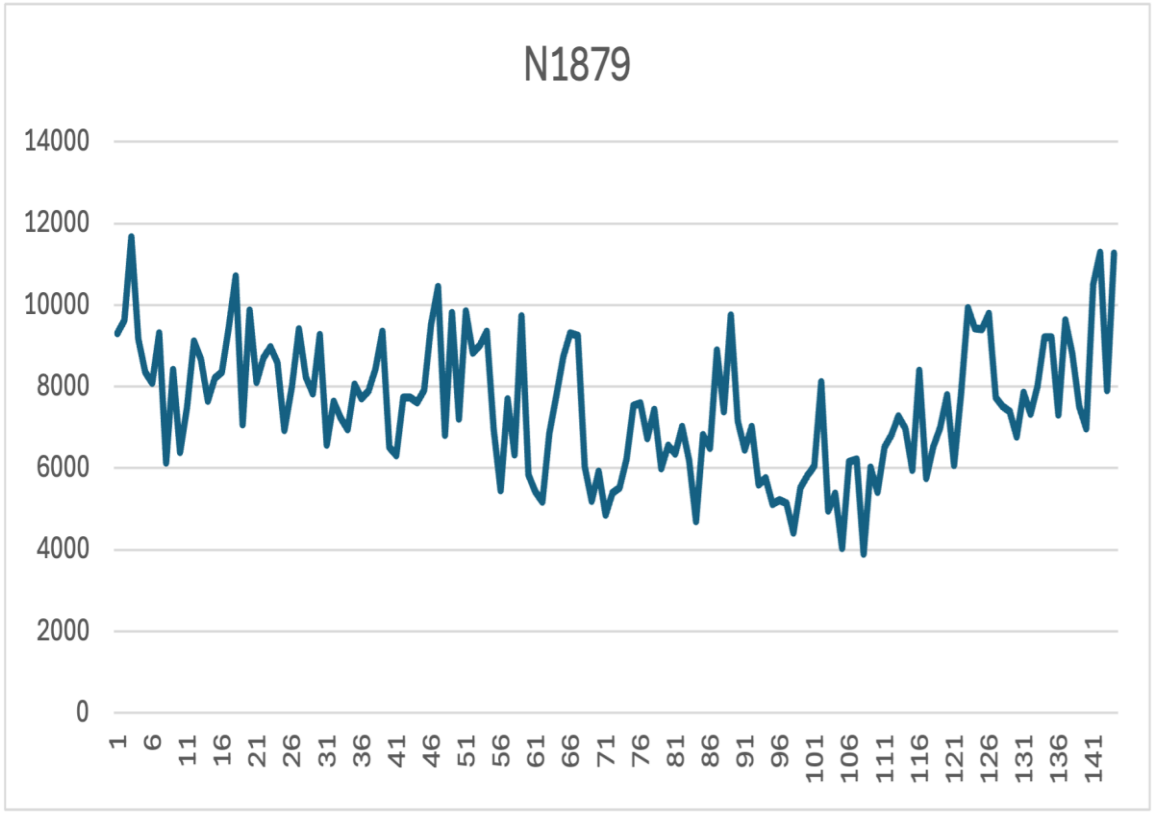
# Serie Storiche

# Serie Storiche

Una serie storica è una sequenza di osservazioni ordinate nel tempo:

$$y_1, y_2, \ldots, y_n \quad \text{con} \quad y_t \in \mathbb{R},\ t = 1, \ldots, n$$

- **Analisi finanziaria**
- **Epidemiologia**
- **Monitoraggio industriale**
- **Macroeconomia**



# Obiettivi

### Descrizione

trend, stagionalità

### Previsione

Stimare i valori futuri sulla base di quelli passati

### Segmentazione

Dividere la serie in parti omogenee

### Rilevare anomalie

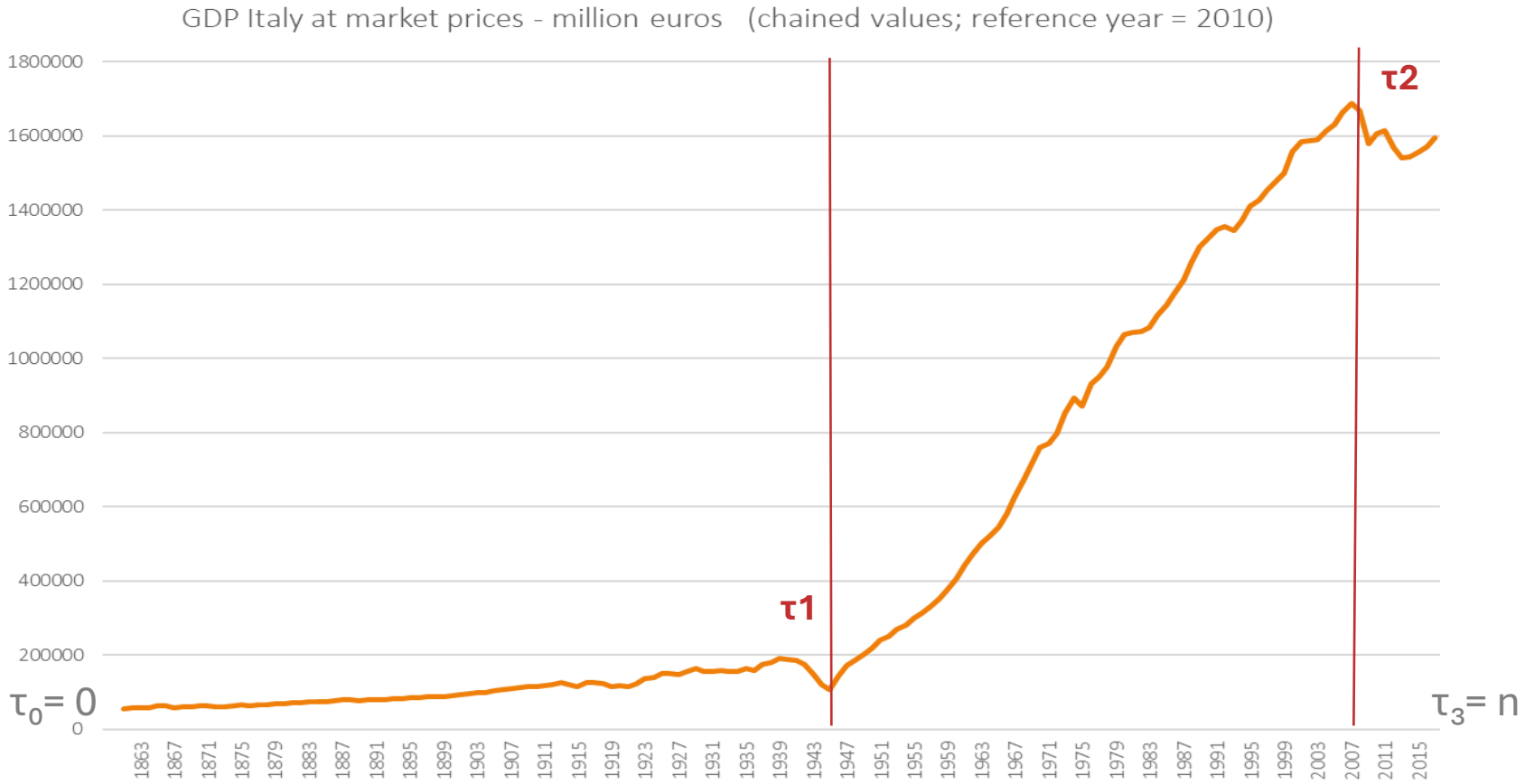
Identificare osservazioni inattese rispetto al modello

02

# Il Problema: Change Point Detection

# Change Point Detection

**Definizione**
Dato $y_{1:n}$ , il CPD consiste nell'identificare k indici $0 = \tau_0 < \tau_1 < \ldots < \tau_k < \tau_{k+1} = n$ tali che le sotto-sequenze $y(\tau_{j-1}+1):\tau_j$ siano statisticamente omogenee internamente e statisticamente distinte tra loro.



# Regressione Piecewise Linear

Ogni segmento è modellato tramite regressione lineare locale:

$$y_t = \alpha_j + \beta_j t + \varepsilon_t$$
$$\varepsilon_t \sim N(0, \sigma^2), \quad t \in (\tau_{j-1}, \tau_j]$$

Obiettivo di ottimizzazione (struttura additiva):

$$\min \ \Sigma_j C(\tau_{j-1}, \tau_j)$$

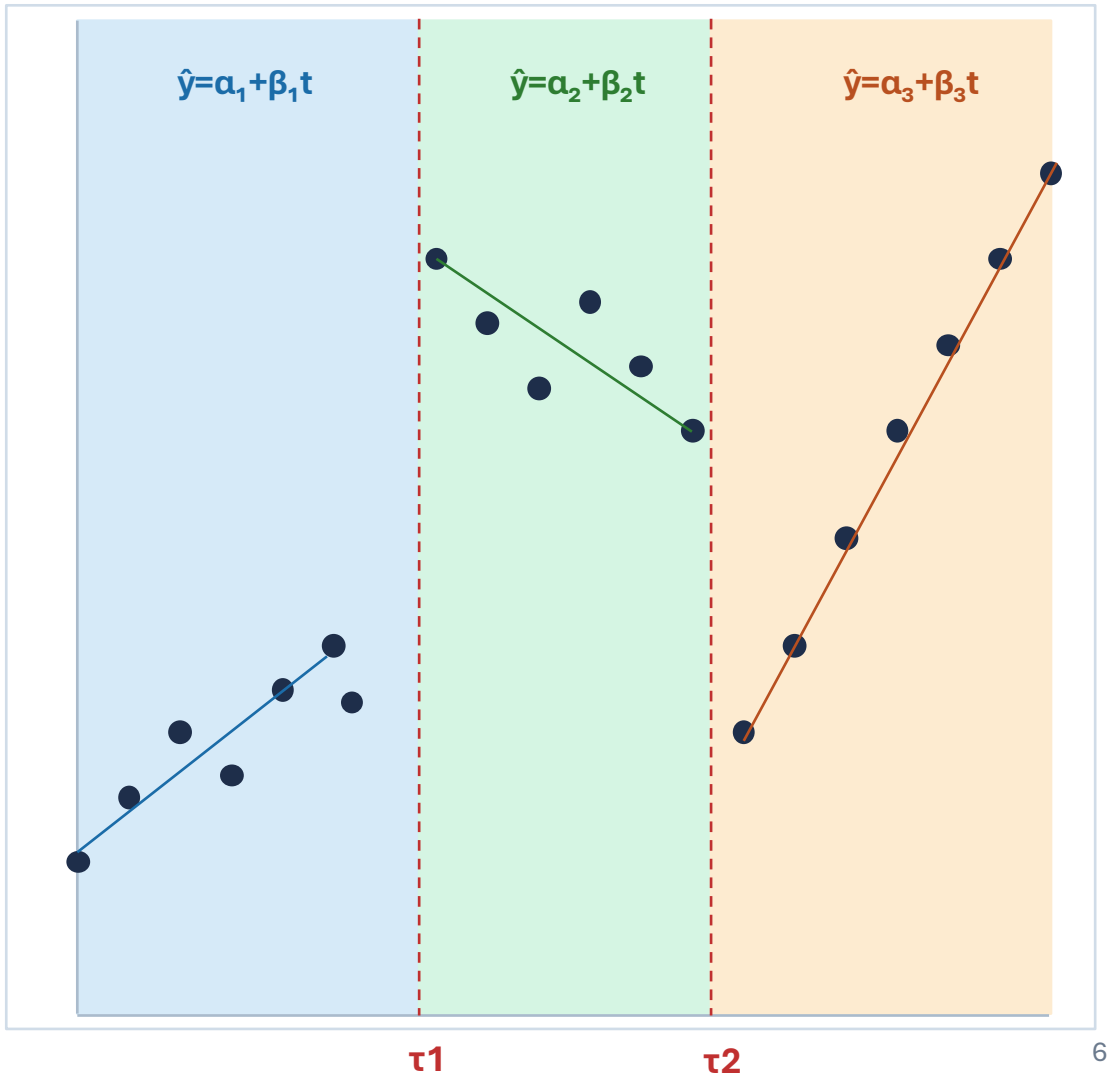
**Funzioni di costo:**

**AIC**

$$= 2k - 2\log(L)$$

**QRMSE**

$$= \sqrt{SSE/\sqrt{n}}$$



# 03

# Stato dell’arte

# Metodi esatti

## Idea chiave della Programmazione Dinamica:

La segmentazione ottima di una serie di n punti con k changepoint contiene al suo interno la segmentazione ottima di ogni sua sotto-serie.

## Due varianti:

### Formulazione VINCOLATA - Dynp

Il numero di changepoint k è fisso.

$$\min_{\tau_1,..,\tau_k} \sum_{j=1}^{k+1} C(\tau_{j-1}, \tau_j)$$

### Formulazione PENALIZZATA - PELT

Il numero di changepoint è determinato automaticamente:

$$\min_{k,\tau_1,...,\tau_k} \sum_{j=1}^{k+1} C(\tau_{j-1}, \tau_j) + \beta k$$

# Dynp

*Algoritmo Segment Neighbourhood (Auger & Lawrence, 1989).*

$$F(t, k) = \min_{s<t} \{ F(s, k-1) + C(s, t) \}$$

*F(t,0) = C(0,t) — condizione iniziale*

**F(t,k)**

*costo minimo fino a t con k CP*

**F(s,k−1)**

*sotto-problema già risolto*

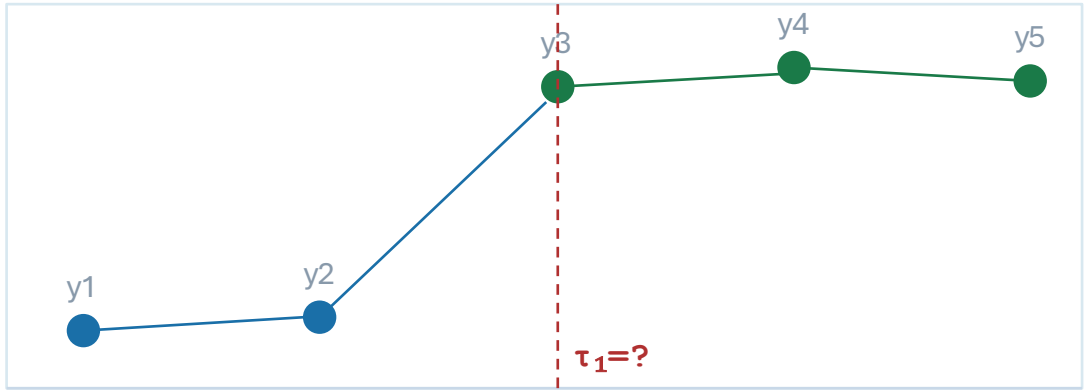
**C(s,t)**

*costo segmento*

Complessità: $O(kn^2)$ · Ottimalità: globale per k fisso

# Esempio: n=5, k=1

y = [ 2,  2.2,  5.8,  6.1,  5.9 ] — cambio strutturale attorno a t=2




**Riga k=0: F(t,0) = C(0,t)**

| | |
|---|---|
| F(1,0) | = 1.2 |
| F(2,0) | = 1.4 |
| F(3,0) | = 4.2 |
| F(4,0) | = 6.1 |
| F(5,0) | = 6.8 |

**Riga k=1: calcolo F(5,1)**

**F(5,1) = min di:**

| | | |
|---|---|---|
| **s=1** | F(1,0)+C(1,5) | = 1.2 + 3.12 = 4.32 |
| **s=2** | F(2,0)+C(2,5) | **= 1.4 + 0.22 = 1.62 ← min ✓** |
| **s=3** | F(3,0)+C(3,5) | = 4.2 + 0.80 = 5.00 |
| **s=4** | F(4,0)+C(4,5) | = 6.1 + 0.00 = 6.10 |

$\tau_1 = 2$

→ $y_{1:2}$ = {2, 2.2} e $y_{3:5}$ = {5.8, 6.1, 5.9}

# PELT — Pruned Exact Linear Time

*Killick, Fearnhead & Eckley (2012).*

$$G(t) = \min_{s<t} \{ G(s) + C(s,t) + \beta \} \quad \text{con } G(0) = -\beta$$

*β = costo fisso per ogni changepoint aggiunto.*

**Regola di potatura**

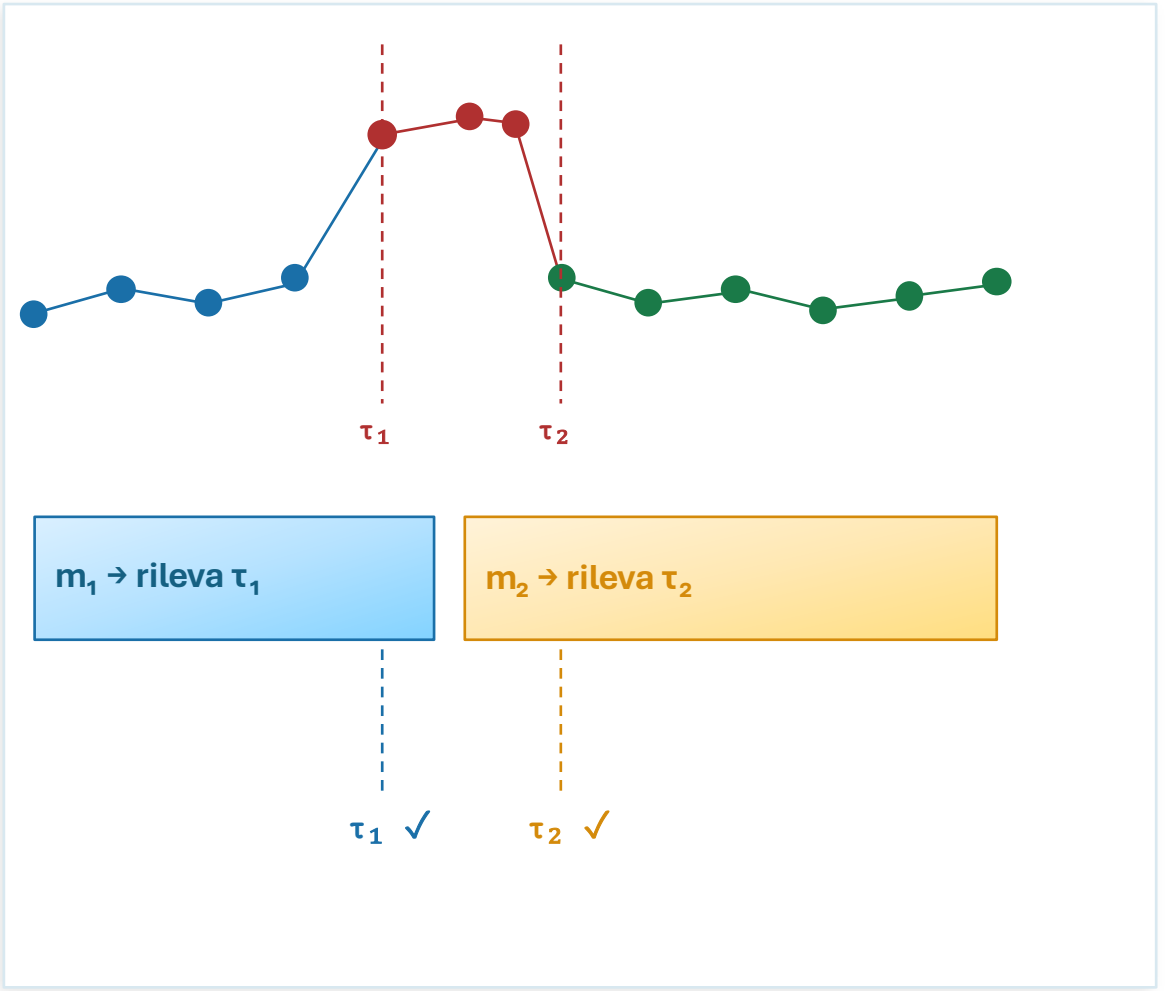
Se $G(s) + C(s,t) + \beta \geq G(t)$ allora s si scarta definitivamente

Complessità: O(n) · Ottimalità: globale

# Wild Binary Segmentation (WBS)

*Fryzlewicz (2014).*

| | | |
|---|---|---|
| 1 | **Campiona M intervalli casuali** | $[s_m, e_m] \subseteq \{1,\dots,n\}$ casuali, m=1,…,M=400 |
| 2 | **Calcola il contrasto su ogni intervallo** | su ogni intervallo: trova b* che massimizza la statistica di constrasto |
| 3 | **Seleziona il massimo globale** | se contrasto > soglia → τ* è il prossimo changepoint |
| 4 | **Ricorsione** | ripete su [1,τ*] e [τ*+1,n] con gli intervalli contenuti nel sotto-problema |




Complessità: O(Mn log n) · Euristica

# 04

# Il nostro modello: SC-CPD

# Limiti dei metodi classici

**Super-additività obbligatoria per Pelt**

Dividere un segmento non può aumentare il costo totale.

**Nessun vincolo globale**

**Calibrazione manuale dei parametri**

**Soluzione: formulazione MILP (Set Covering)**

# SC-CPD: Formulazione Set Covering

**Idea fondamentale: ogni possibile segmento della serie diventa una variabile di decisione.**

1. **Genera tutti i segmenti candidati**

   Ogni segmento (s, t) con lunghezza ≥ min_size diventa una variabile $x_j \in \{0,1\}$

2. **Assegna un costo a ciascuno**

   $c_j$ = costo del segmento j

3. **Risolve il problema di set covering**

   Sceglie il sottoinsieme di segmenti che copre tutta la serie al costo minimo

# Il modello TSSP: Time Series Set Partitioning

## Modello TSSP

$$\min \quad \sum_j c_j x_j$$

$$\text{s.t.} \quad \sum_j a_{ij} x_j = 1 \quad \forall i = 1,\dots,n$$

$$\sum_j x_j \leq \text{maxruns}$$

$$x_j \in \{0,1\}$$

*$a_{ij}$ = 1 se il segmento j copre il punto i*

- **Qualsiasi funzione C**
- **Vincoli globali**
- **Ottimalità garantita**
- **Flessibile**

### TSSP — Set Partitioning

- Vincoli di uguaglianza.
- Ogni punto coperto esattamente una volta.
- Soluzione ottima.
- Più lenta su serie lunghe.

### TSSC — Set Covering (rilassato)

- Vincoli di disuguaglianza (≥ 1).
- Richiede postprocessing.
- GAP medio: < 1% dall'ottimo.
- Più veloce

# 05

# Vincoli locali e globali

# Vincoli locali

**Lunghezza minima e massima** $\ell_{min} \leq \text{lunghezza} \leq \ell_{max}$

Evita segmenti troppo corti o troppo lunghi.

**Vincolo sulla pendenza** $|\beta| \leq \beta_{max}$

Limita la ripidità del trend nel segmento.

**Posizioni ammissibili per i changepoint** $\tau \in A$

I breakpoint possono essere solo in certe posizioni.

# Vincoli globali

## Numero massimo di segmenti

$$\Sigma x_i \leq n\_max$$

Esempio: voglio al massimo 5 segmenti.

## Budget sul costo totale

$$\Sigma c_i x_i \leq B$$

Esempio: la segmentazione non può costare più di B.

## Variazione totale limitata

$$v\_max - v\_min \leq \delta\_max$$

Esempio: la serie non può "saltare" più di δ_max tra un segmento e l'altro.

## Pattern direzionali

**Aumento → plateau → discesa**

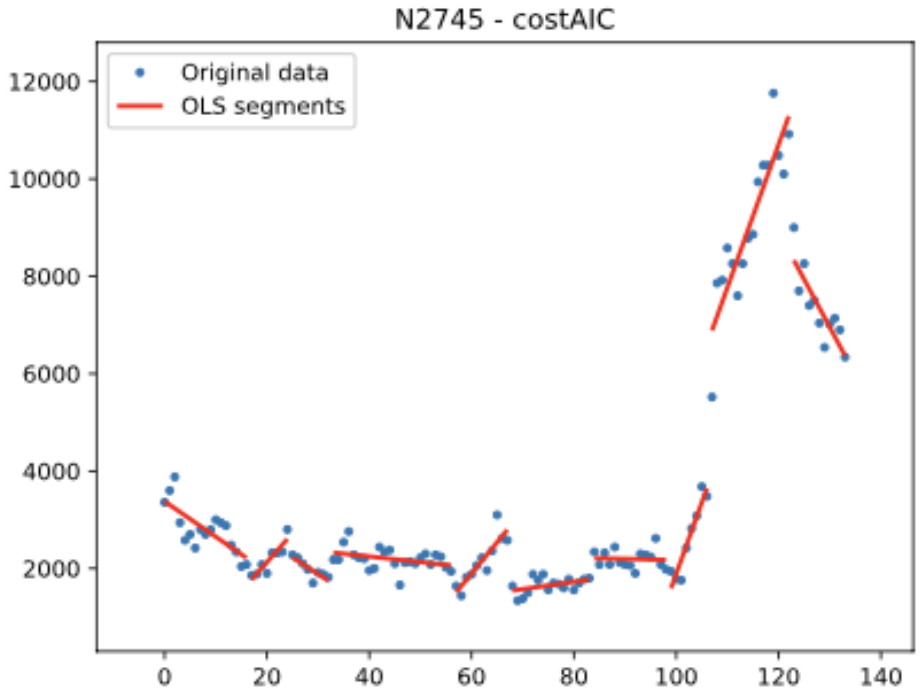
06

# Risultati sperimentali

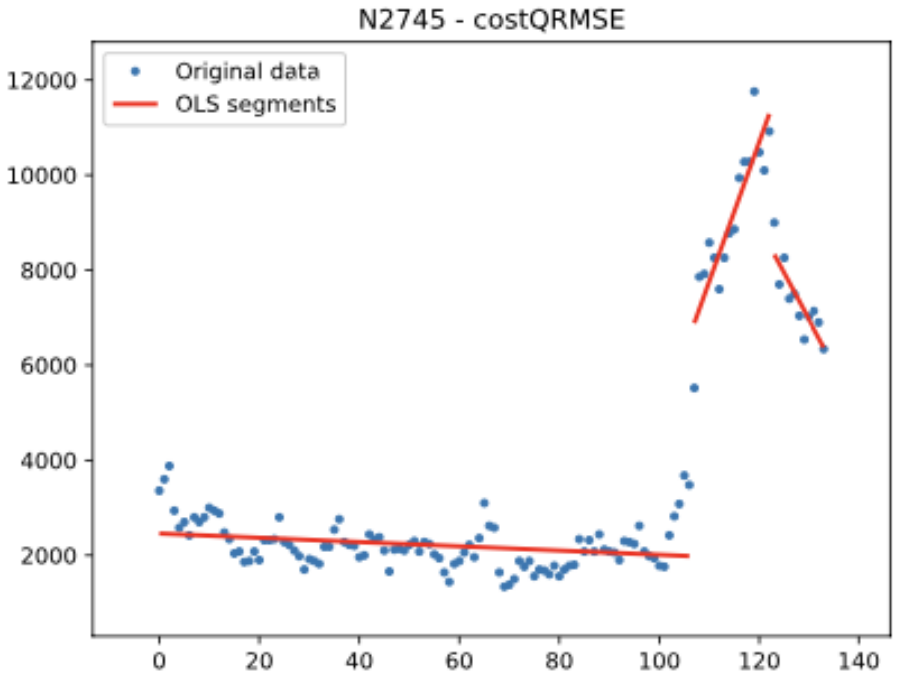
# Il benchmark: le M competizioni

| | |
|---|---|
| **M3**<br>**(2000)**<br>**3.003 serie** | **Frequenze:** annuali, trimestrali, mensili<br>**Domini:** Finanza, Industria, Economia, Demografia<br>**Test set:** 5 serie per categoria<br>Brevi (60–144 punti). |
| **M4**<br>**(2018)**<br>**100.000 serie** | **Frequenze:** annuale, trimestrale, mensile, settimanale, giornaliera, oraria<br>**Domini:** Macro/Micro-Economia, Finanza, Industria, Demografia<br>**Test set:** 5 serie per categoria<br>Serie lunghe (fino a 700+ punti). |
| **M6**<br>**(2022)**<br>**100 asset finanziari** | **Frequenze:** Serie finanziarie ad alta frequenza<br>**Domini:** Health Care, Energy, Real Estate, IT, Consumer Discretionary, Equities, Commodities, Volatility<br>**Test set:** 10 serie per categoria |

# Funzioni di costo




(a) AIC cost segmentation




(b) QRMSE cost segmentation

**AIC** $= 2k - 2\log(L)$

**QRMSE** $= \sqrt{SSE/\sqrt{n}}$

# Risultati: Metodi esatti

*Parametri: min_size=8 · PELT: β=40 · Dynp: n_bkps=9*

**Tabella 1 — PELT**

| Bench | Cat. | AIC med.n | AIC max.n | QRMSE med.n | QRMSE max.n | AIC med.t | AIC max.t | QRMSE med.t | QRMSE max.t | TSSP med.t | TSSP max.t |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **M3** | Demographic | 3 | 7 | 3 | 4 | 0 | 0 | 0 | 0 | 0.41 | 0.59 |
| **M3** | Finance | 3 | 9 | 2 | 7 | 0 | 0 | 0 | 0 | 0.40 | 0.56 |
| **M4** | Industry | 5 | 10 | 6 | 7 | 0 | 0.1 | 0 | 0 | 0.58 | 71.77 |
| **M4** | Micro | 3 | 9 | 4 | 6 | 0 | 0.1 | 0 | 0 | 0.30 | 69.87 |
| **M6** | Finance | 6.5 | 8 | 1 | 1 | 0 | 0 | 0 | 0 | 2.72 | 3.10 |

**Tabella 2 — DYNP**

| Bench | Cat. | AIC med.n | AIC max.n | QRMSE med.n | QRMSE max.n | AIC med.t | AIC max.t | QRMSE med.t | QRMSE max.t | TSSP med.t | TSSP max.t |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **M3** | Demographic | 10 | 10 | 10 | 10 | 0 | 0 | 0 | 0.1 | 0.41 | 0.59 |
| **M3** | Finance | 10 | 10 | 10 | 10 | 0 | 0.1 | 0 | 0 | 0.40 | 0.56 |
| **M4** | Industry | 10 | 10 | 10 | 10 | 0 | 2.5 | 0 | 2.7 | 0.58 | 71.77 |
| **M4** | Micro | 10 | 10 | 10 | 10 | 0 | 2.5 | 0 | 2.7 | 0.30 | 69.87 |
| **M6** | Finance | 10 | 10 | 10 | 10 | 0.2 | 0.2 | 0.2 | 0.2 | 2.72 | 3.10 |

# Risultati: WBS

*Parametri: min_size=8 · WBS: M=400, thr_AIC=20, thr_QRMSE=10*

**Tabella 3 — WBS**

| Bench | Cat. | AIC med.n | AIC max.n | QRMSE med.n | QRMSE max.n | AIC med.t | AIC max.t | QRMSE med.t | QRMSE max.t | TSSP med.gap | TSSP max.gap |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **M3** | Demographic | 10 | 11 | 5 | 8 | 0.3 | 0.3 | 0.2 | 0.3 | 1.30 | 9.26 |
| **M3** | Finance | 9 | 11 | 6 | 12 | 0.2 | 0.2 | 0.2 | 0.3 | 1.28 | 8.62 |
| **M4** | Industry | 10 | 27 | 8 | 11 | 0.3 | 2.7 | 0.3 | 2.0 | 3.31 | 26.15 |
| **M4** | Micro | 8 | 22 | 7 | 9 | 0.2 | 2.6 | 0.2 | 2.1 | 2.29 | 19.43 |
| **M6** | Finance | 17 | 21 | 1 | 3 | 0.7 | 0.8 | 0.2 | 0.4 | 0.00 | 33.05 |

# Risultati: TSSP

*Vincoli: min_size=8, max 10 segmenti.*

**Tabella 4 — TSSP**

| Bench | Cat. | avg.n | max.n | AIC avg.nseg | AIC max.nseg | AIC avg.t | AIC max.t | QRMSE avg.nseg | QRMSE max.nseg | QRMSE avg.t | QRMSE max.t |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **M3** | Demographic | 122.6 | 138 | 9.4 | 10 | 0.2 | 1 | 3.8 | 7 | 0.6 | 1 |
| **M3** | Finance | 124.8 | 144 | 9.6 | 10 | 0.1 | 0.1 | 3.6 | 9 | 0.1 | 0.1 |
| **M4** | Industry | 258.4 | 737 | 8.8 | 10 | 62.4 | 310 | 4.6 | 7 | 60 | 298 |
| **M4** | Micro | 232.8 | 737 | 10 | 10 | 60.4 | 302 | 2.8 | 7 | 58.4 | 292 |
| **M6** | Finance | 257.8 | 265 | 10 | 10 | 5.9 | 7 | 4.2 | 7 | 7.0 | 25 |

**~40%**

**più veloce**

*TSSC rispetto a TSSP*

**< 1%**

**gap medio**

*TSSC vs ottimo TSSP*

# Conclusioni

### Framework MILP

SC-CPD supera i limiti strutturali della DP

### TSSC: veloce e quasi-ottimo

~40% più veloce di TSSP con GAP medio <1%.

### Flessibilità del modello

Non limitato alla regressione lineare.

# Grazie per l'attenzione.

## Domande?

Lisa Vecchi · Università di Bologna

*Mentor: Prof. Vittorio Maniezzo · Università di Bologna*