# 4 Метод динамічного програмування. Практична частина

Проведемо оцінку роботи алгоритму на штучних даних. Опишемо концепцію.

Спочатку змоделюємо таку вибірку об'єму $n$, де на кожному сегменті математичне сподівання дорівнює деякій константі. Покладемо $X = (X_1, X_2, \dots, X_n)^\top$ в якості такої. Далі, для $X$ знаходимо оптимальну сегментацію $\hat{t}_X$ за вказаним методом. У цьому випадку ми її будемо називати істинною сегментацією вибірки $X$. Для кожного $\sigma \in S = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$ генеруємо по $N$ варіантів $X$, модифікованої за допомогою білого шуму. Тобто, покажемо наступне:

$$\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_N),\ \tilde{X}_j = X + e_j = (X_1 + e_j^{(1)}, X_2 + e_j^{(2)}, \dots, X_n + e_j^{(n)})^\top\ (j = \overline{1, N}),$$

де $e_j = (e_j^{(1)}, e_j^{(2)}, \dots, e_j^{(n-1)}, e_j^{(n)})^\top$, $e_j^{(k)} \sim N(0, \sigma^2)$, $k = \overline{1, n}$ - вектор похибок, компоненти якого нормально розподілені.

Задіємо алгоритм розбиття для отриманих екземплярів.

Показники відхилень отриманного розбиття від істиного будемо рахувати двома способами. Тут $K \in \overline{1, K_{max}}$.

**Перший спосіб** базується на підрахунку пропорцій кожного сегмента вибірки.

Нехай $X$ - вибірка, $\hat{t}_X = \{\hat{t}_1, \hat{t}_2, \dots, \hat{t}_K\}$ - її розбиття. Тоді визначимо частку кожного сегмента вибірки $\hat{t}_X$:

$$p_1 = \frac{t_1}{\#X},\ \Delta_i = t_i - t_{i-1},\ p_i = \frac{\Delta_i}{\#X},\ i \in \overline{2, K}$$

де $\#A$ - кількість елементів множини $A$, $t_0 := 0$. Зауважимо, що

$$\forall i \in \overline{1, K}\colon 0 < p_i < 1;\ \sum_{i=1}^{K} p_i = 1.$$

Маємо набір часток сегментів за сегментацією $\hat{t}_X$: $P = \{p_1, p_2, \dots, p_K\}$

Беремо $X_j \in \tilde{X}$ та набір пропорцій $P_j = \{q_1^{(j)}, q_2^{(j)}, \dots, q_K^{(j)}\}$ за сегментацією $\hat{t}_{X_j}$. Вводимо вектор, компоненти якого сильно нагадують дисперсію, де $q_j^{(i)}$ - частка $i$-го сегмента у $j$-му екземплярі, $p_j$ - частка $i$-го сегмента за істинним розбиттям.

$$\vec{V}^{(K)} = \frac{1}{K} \begin{pmatrix} \sum_{j=1}^{K} (q_j^{(1)} - p_j)^2 \\ \sum_{j=1}^{K} (q_j^{(2)} - p_j)^2 \\ \vdots \\ \sum_{j=1}^{K} (q_j^{(N)} - p_j)^2 \end{pmatrix} = \begin{pmatrix} v_1^{(K)} \\ v_2^{(K)} \\ \vdots \\ v_N^{(K)} \end{pmatrix}$$

Кінцевим результатом такої оцінки буде середнє арифметичне коренів квадратних компонент $\vec{V}^{(K)}$.

$$\tilde{p}^{(K)} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{v_i^{(K)}}$$

**Другий спосіб** полягає в застосуванні спеціальної метрики для обчислення відхилень. Побудуємо її наступним чином:

Для всіх $i, j \in \mathcal{S}$ визначимо функцію приналежності пари елементів одному числовому сегмента з розбиття $t$.

$$\delta_t(i,j) = \sum_{s=0}^{K-1} \mathbb{1}_{\{t_s \leqslant i \leqslant t_{s+1}\}}(i) \mathbb{1}_{\{t_s \leqslant j \leqslant t_{s+1}\}}(j),$$

де $\mathbb{1}_A(x)$ - індикатор множини $A$.

$$\rho^{(K)}(t,s) = \sum_{i=1}^{T} \sum_{j=1}^{T} |\delta_t(i,j) - \delta_s(i,j)| \tag{5}$$

**Твердження.** Функція (5) задає відстань між обраними сегментаціями.

□ 1. $\forall x, y$: $\rho(x,y) \geqslant 0$, а $\rho(x,y) = 0 \Leftrightarrow x = y$.
2. $\rho(x,y) = \rho(y,x)$ - тривіально.
3. Беремо $x \leqslant z \leqslant y$. Тоді:
$\rho(x,y) = \sum_{i=1}^{T} \sum_{j=1}^{T} |\delta_x(i,j) - \delta_y(i,j)| = \sum_{i=1}^{T} \sum_{j=1}^{T} |\delta_x(i,j) - \delta_y(i,j) + \delta_z(i,j) - \delta_z(i,j)| \leqslant$
$\leqslant \sum_{i=1}^{T} \sum_{j=1}^{T} |\delta_x(i,j) - \delta_z(i,j)| + \sum_{i=1}^{T} \sum_{j=1}^{T} |\delta_y(i,j) - \delta_z(i,j)| = \rho(x,z) + \rho(z,y)$. ■

**Зауваження.** Функцію (5) називають робастною метрикою Біфермана[1]. Своїх властивостей вона не втрачає, якщо $t, s$ вважати не за сегментації вибірки, а за числові вектори в $\mathbb{R}^d$.

Необхідно модифікувати (5) принаймні по причині великих затрат по часу на перебір усіх можливих пар $(i,j)$. Крім того, функція повертає достатньо громіздкі значення, тому її необхідно нормувати.

Біферман пропонує скоротити час на обчислення операцій за допомогою наступної модифікації з використанням нової змінної $k$ - половина від середньої довжини сегментів у $t$ та $s$.

$$\rho_k^{(K)}(t,s) = \sum_{i=1}^{T-k-1} |\delta_t(i, i+k+1) - \delta_s(i, i+k+1)|, \tag{6}$$

За допомогою множення (5) на $(T)^{-2}$ ( або (6) на $(T-k-1)^{-1}$ ) отримуємо нормованість метрики.

$$\hat{\rho}^{(K)}(t,s) = \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} |\delta_t(i,j) - \delta_s(i,j)| \tag{7}$$

$$\hat{\rho}_k^{(K)}(t,s) = \frac{1}{T-k-1} \sum_{i=1}^{T-k-1} |\delta_t(i, i+k+1) - \delta_s(i, i+k+1)| \tag{8}$$

[1]У своїй роботі, Біферман ввів задану метрику в якості оцінки сегментацій текстових документів. Однак така функція згодиться і для нашої задачі. (https://arxiv.org/abs/cmp-lg/9706016)

**Зауваження.** Нормована метрика (5) носить більш загальну назву - індекс Ренда. Вона широко застосовується у кластерному аналізі в якості міри подібності розбиттів.

Для $N$ отриманих сегментацій будемо знаходити показники метрики відносно істинної сегментації. Кінцевим буде середнє по всім значенням.

$$\bar{\rho}^{(k)} = \frac{1}{N}\sum_{i=1}^{N}\rho^{(k)}(\hat{t}_{\tilde{X}_i}, \hat{t}_X)$$

Вказані обчислення проводимо для кожної сегментації $j$ порядку ($j \in \overline{1, K}$). Кінцевими показниками відхилень $t_{X_k}$ від $t_X$, $k \in \overline{1, N}$ будуть:

$$\tilde{p} = \frac{1}{K_{max}}\sum_{k=1}^{K_{max}}\tilde{p}^{(k)}, \;\; \bar{\rho} = \frac{1}{K_{max}}\sum_{k=1}^{K_{max}}\bar{\rho}^{(k)}.$$

Змодельовано три вибірки обсягу $n = 100$. Після цього, для кожної вибірки та кожного $\sigma \in S = \{0, 0.05, \ldots, 0.25\}$, створено по $N = 1000$ модифікованих екземплярів. Далі все було виконано в порядку показаних раніше інструкцій. У кінці будуть показані графіки зміни показників відхилень в залежності від зміни параметрів розподілу.

У даному експерименті за вказаним вище способом було змодельовано три тестові вибірки обсягу $n = 100$:

- Першу вибірку ($A$) можна розбити на максимальну кількість частин рівній 3. Перші 35 спостережень матимуть середнє $\mathbb{M}X_j = 0.25$, ($1 \leqslant j \leqslant 25$); наступні 50: $\mathbb{M}X_j = 0.65$, ($26 \leqslant j \leqslant 75$); останні 25 - $\mathbb{M}X_j = 0.45$ ($76 \leqslant j \leqslant 100$). (Числові характеристики будуть наведені лише для $A$, спостереження з довільної вибірки набувають значень з $(0, 1)$);
- Для другої вибірки ($B$), на відміну від першої, характерна наступна властивість: математичні сподівання спостережень на різних сегментах утворюють монотонну послідовність. Максимальна кількість доречних розбиттів аналогічна першому випадку;
- Остання вибірка ($C$) розбивається на 7 частин. Має східчасту форму - математичні сподівання для кожного сегменту з першої половини вибірки утворюють зростаючу послідовність, а з другої - спадну.

На рис. (1) маємо зображення для $A, B, C$ та модифікації кожної з них внаслідок додавання вектора похибок з гаусовим розподілом.

Графічна інтерпретація показників якості сегментації, тобто значень індексу Ренда та оцінки невідповідності пропорцій, можна побачити на рисунках:

- (2) - графіки двох оцінок на одному рисунку для кожної вибірки;
- (3), (4), (5) - для $A$, $B$ і $C$ відповідно.

З графіків видно, що показник похибки та індекс Ренда зростає у разі збільшення середньоквадратичного відхилення в розподілі. Ліпше це бачити для модифікацій над $C$.

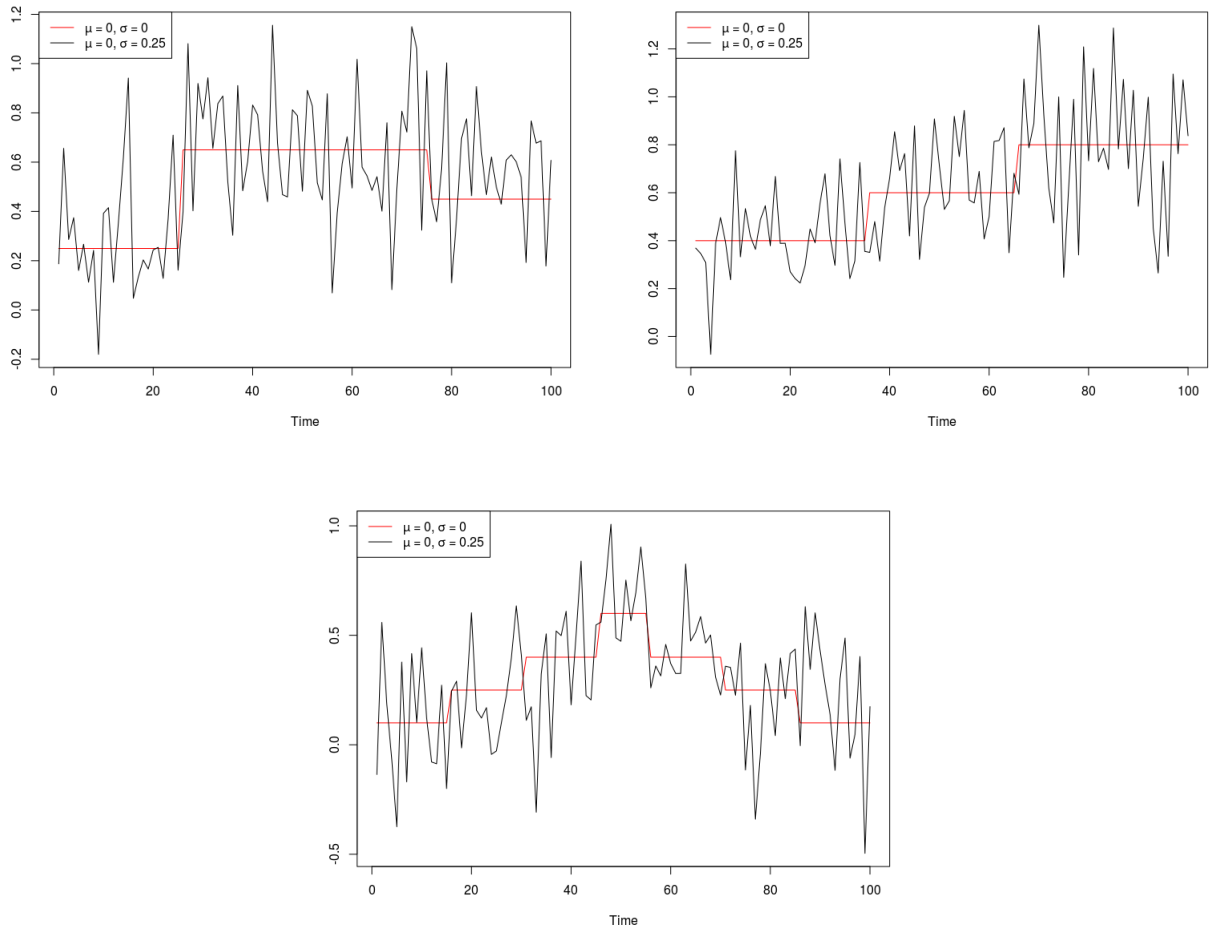


Рис. 1: Вручну створені числові набори даних $A, B, C$ з $n$ компонент (червоний колір) та результат взаємодії на елементи з кожної гаусівськими величинами (чорний колір)

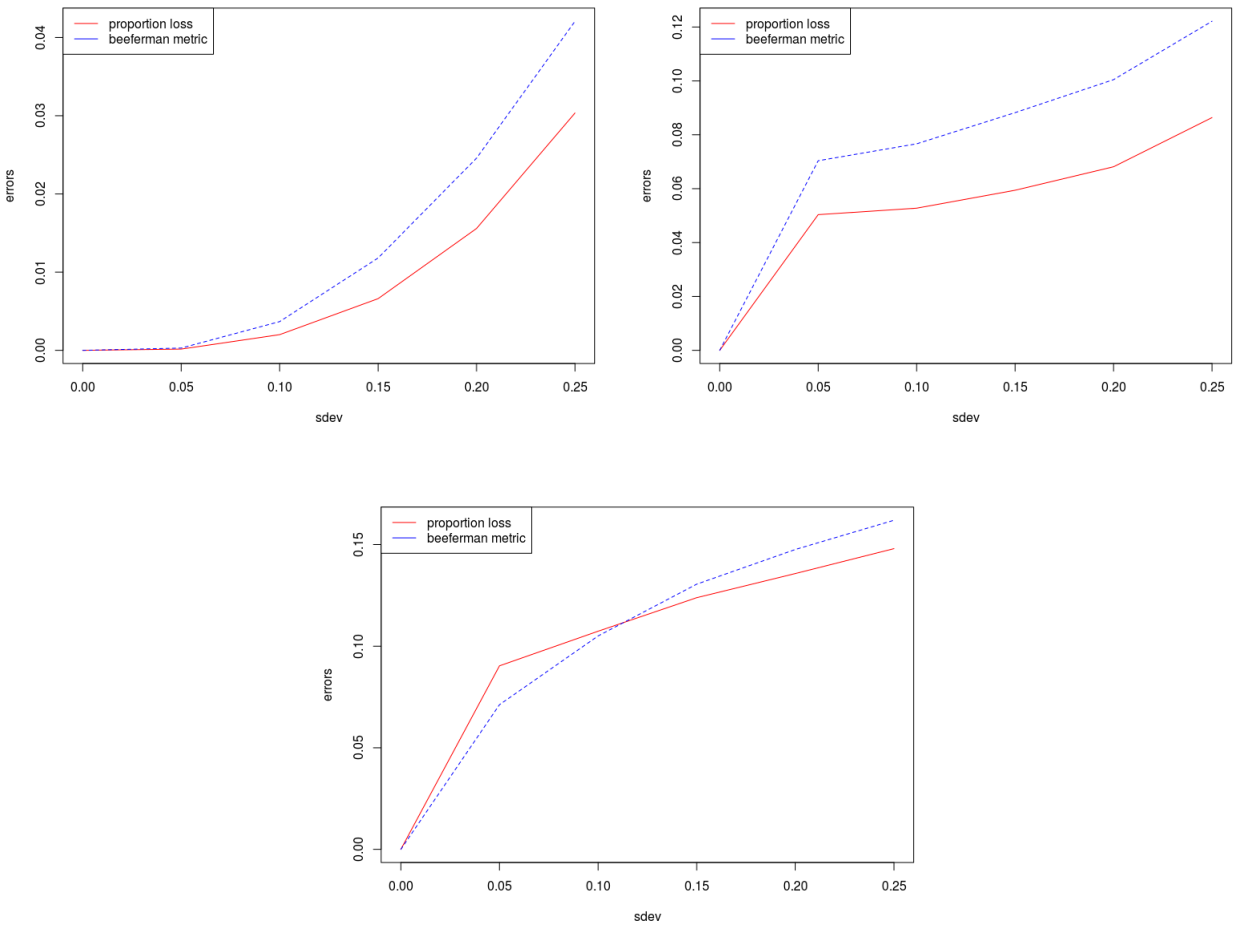


Рис. 2: Графіки зміни значень метрики (синій колір) та оцінки невідповідності пропорцій (червоний колір) в залежності від зміни параметра $\sigma$ в $N(0, \sigma^2)$ для модифікованих вибірок $A, B, C$ відповідно

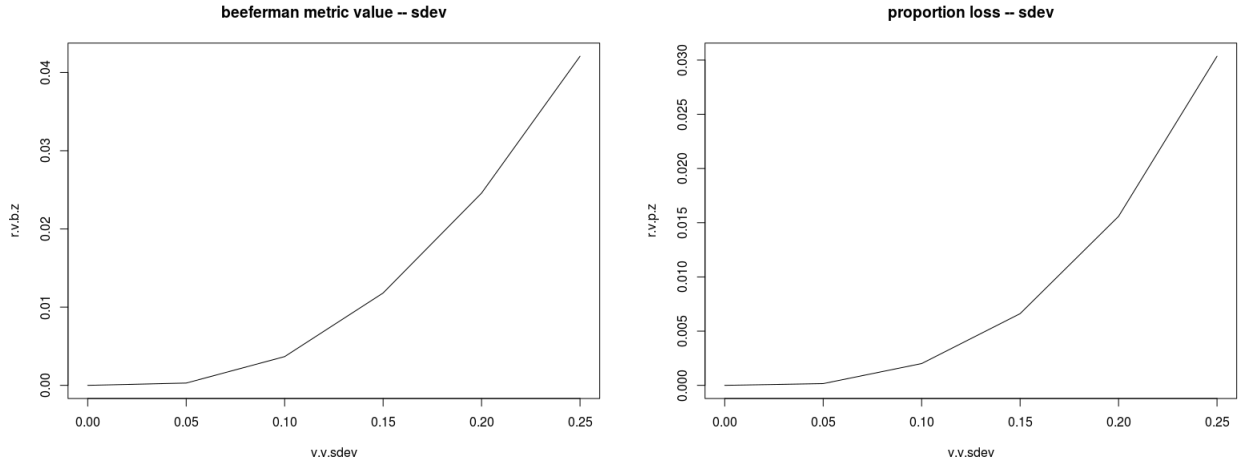


Рис. 3: Графіки зміни значень метрики (правий) та оцінки невідповідності пропорцій (лівий) для сегментацій модифікованих вибірок $A$

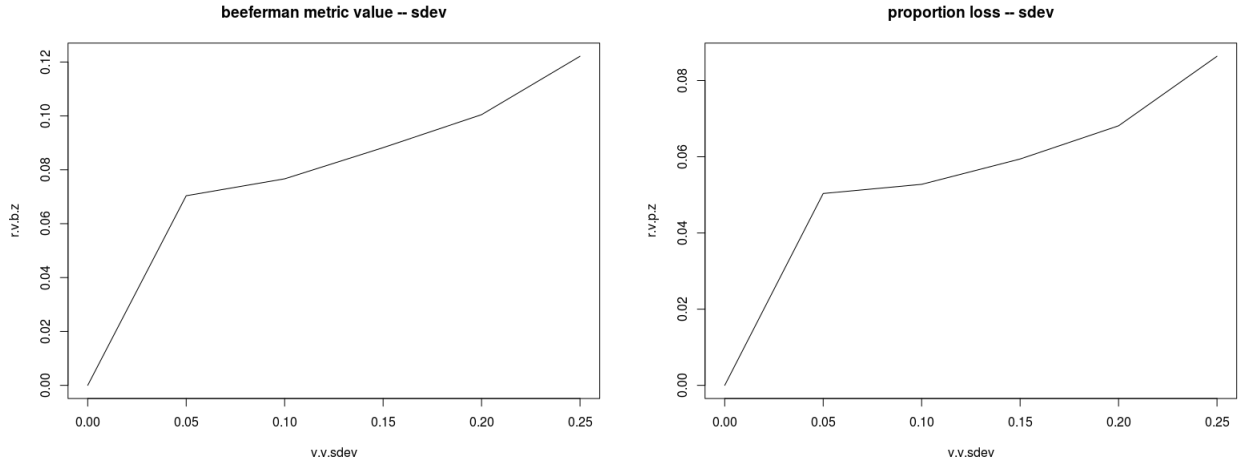


Рис. 4: Графіки зміни значень метрики (правий) та оцінки невідповідності пропорцій (лівий) для сегментацій модифікованих вибірок $B$

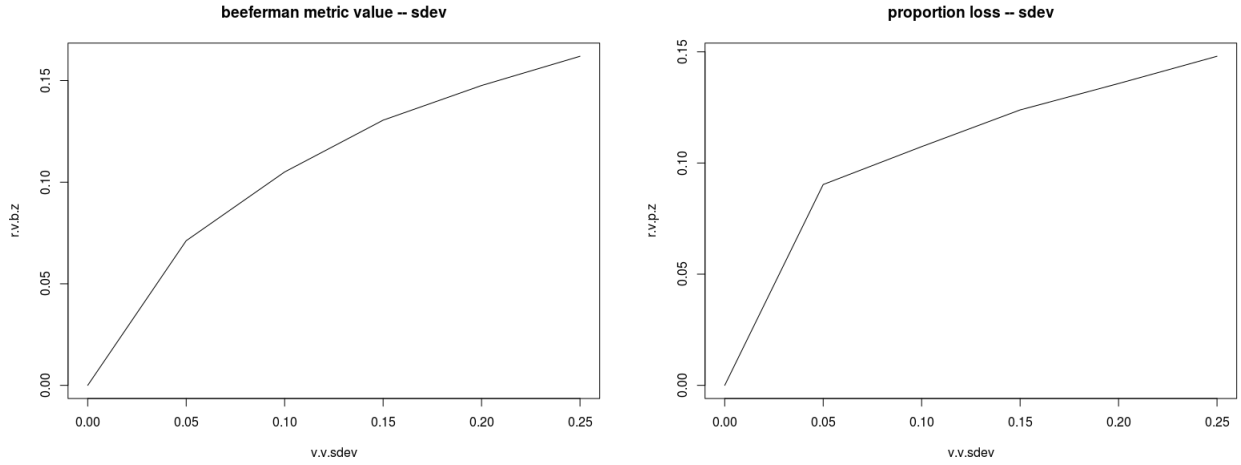


Рис. 5: Графіки зміни значень метрики (правий) та оцінки невідповідності пропорцій (лівий) для сегментацій модифікованих вибірок $C$

# 5 Модифікація алгоритму. Принцип блокової сегментації.

Наступним завданням було порівняння результатів описаного алгоритму при фіксованому значенні дисперсії похибок, збільшуючи обсяг початкової вибірки. Основна реалізація виконується досить швидко, опрацьовуючи вибірки невеликих розмірів та для малих кількостей розбиттів.

У такому випадку використання принципу блокової сегментації розбиттів є доречним. Він спрощує кількість операцій на обчислення значень функції похибок. Якщо для $d_{s,t}$ часові границі перебиралися з множини $\mathcal{S}$, то в цьому разі значення беруться з множини меншої потужності: $\mathcal{S}_N = \{1, 1+N, 1+2N, \ldots, 1+(M-1)N, T\}$, де $N$ визначає розмір блока, а $M$ - кількість елементів зазначеної множини. Тобто показник відхилень визначається аналогічно, але для інших сегментів:

$$d_{s,t}^{(N)} = \sum_{m=s}^{t} (X_m - \hat{X}_m)^2,$$

де $N$ - розмір блока, $s, t \in \mathcal{S}_N$, інші позначення відомі. Припущення щодо різниці між двома останніми показниками часу $1+(M-1)N, T$ була рівною $N$ можна знехтувати, але бажано, щоб це виконувалося в загальному. Для цього $N$ обирається таким чином, щоб число $(\#X - 1)$ ділилося націло на нього.

Інші математичні викладки, наведені для базового алгоритму фрагментації, залишаються незмінними. Звичайно, якість отриманого розбиття буде гіршою, але таким чином проводиться менша кількість операцій. Крім того, для скінченних часових рядів великих розмірів, отримана похибка такої сегментації не є значущою.

# 6 Порівняння результатів методу фрагментації часового ряду в залежності від його обсягу.

Подивимося на поведінку алгоритму сегментації при збільшенні обсягу вибірки при деякому фіксованому значенні дисперсії похибок $e_j$.

Для кожного $m \in S \subset \mathbb{N}$ змоделюємо вибірку $X^{(m)}$ об'єму $m$ за аналогічним принципом, як робили раніше, при цьому зберігаючи початкові пропорції середніх значень на кожному сегменті. Далі, знаходимо оптимальну сегментацію $\hat{t}_{X^{(m)}}$ та генеруємо по $N$ варіантів $\tilde{X^{(m)}} = X^{(m)} + e_j^{(m)}$, де $e_j^{(m)}$ - гаусовий вектор розмірності $m$ з незалежних в сукупності та однаково розподілених координат, дисперсія яких рівна $\sigma^2$. Для отриманих сегментацій обчислюємо індекс Ренда та робимо висновки.

Зафіксували $S = \{100, 250, 500, 1000, 2000\}$ та $N = 1000$. Моделюємо часовий ряд $X$ з такими ж відношеннями середніх, як в ряді $B$ з розділу 4. Проводимо вищезазначені викладки для кожного $\sigma \in \{0.1, 0.15, 0.25\}$. Знаходження сегментацій для великої кількості спостережень було прискорено з використанням блокової сегментації (розмір блока рівний п'яти).

Нижче наведені отримані значення індексу для різних значень $\#X$ та $\sigma$.

| | 100 | 250 | 500 | 1000 | 200 |
|---|---|---|---|---|---|
| $\sigma = 0.1$ | 0.1280756 | 0.1175231 | 0.1146085 | 0.1123248 | 0.1111846 |
| $\sigma = 0.15$ | 0.1471239 | 0.132588 | 0.1259694 | 0.1227693 | 0.1099334 |
| $\sigma = 0.25$ | 0.2007336 | 0.1504958 | 0.1420582 | 0.131018 | 0.1174093 |

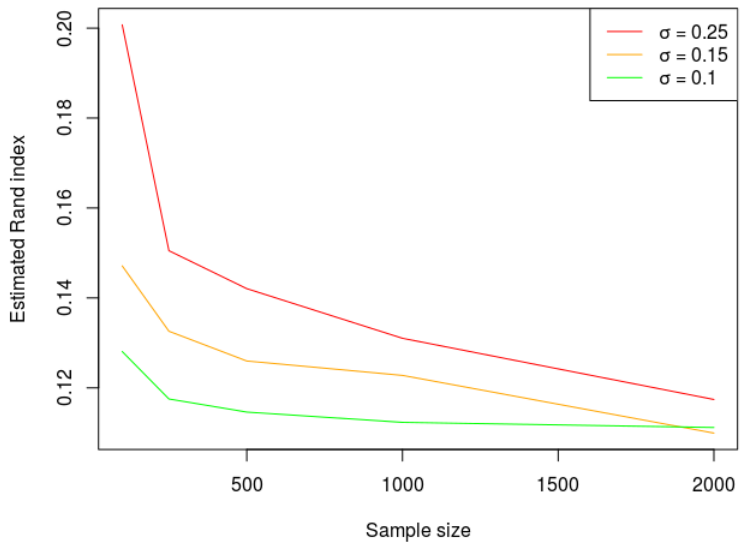


Рис. 6: Графік зміни значень індексу Ренда в залежності від обсягу початкового часового ряду та стандартного відхилення похибок.

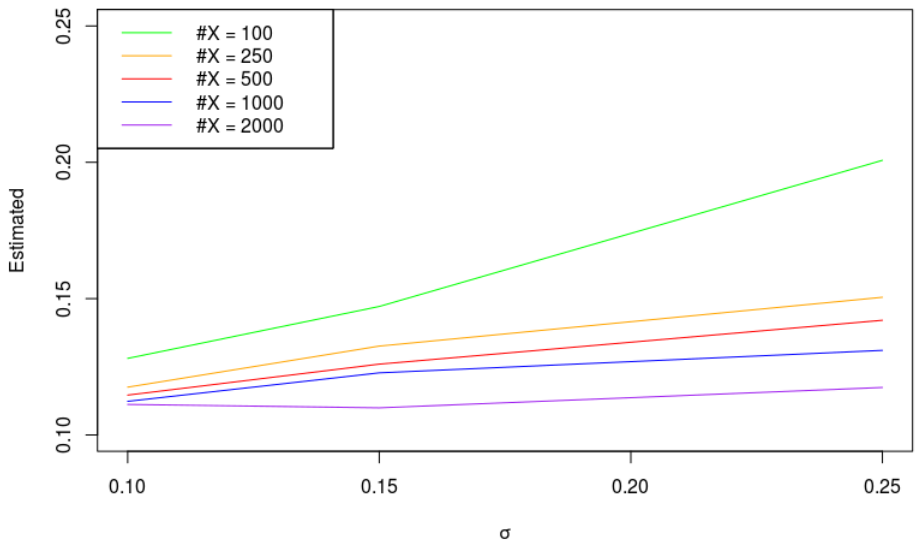


Рис. 7: Графік зміни значень індексу Ренда в залежності від збільшення дисперсії похибок. Різними кольорами побудовані для різної кількості спостережень.

З рисунків (6), (7) можна побачити, що з більшою потужністю часового ряду можна отримати кращі сегментації. Точну форму залежності між $\#X$ та $\sigma$ проблемно описати, але є деяка схожість зі степеневими функціями.

# 7 Вибір оптимального розбиття алгоритму. Інформаційний Байєсівський критерій.

Останнім кроком до розв'язання поставленої задачі на розбиття часового ряду було обрання оптимальної кількості розбиттів для початкової вибірки. Для цього використаємо байєсівський інформаційний критерій. У загальному випадку вводиться статистика:

$$BIC(X,\theta) = k\ln(n) - 2\ln\left(\max_{\theta\in\Theta} L(X,\theta)\right), \tag{9}$$

де $\theta$ визначається як вектор параметрів розподілу з деякої параметричної множини $\Theta$, кількість невідомих параметрів - $k \leqslant \dim\theta$, $n$ - обсяг кратної вибірки $X$ і $L(X,\theta)$ - емпірична функція вірогодності.

У рамках поставленої задачі ситуація інакша, тому необхідно переформулювати[2] (9):

$$\tilde{J}(K) = T\ln\left(\frac{\hat{J}(K)}{T-1}\right) + 2K\ln(T), \tag{10}$$

де $\hat{J}(K)$ - мінімальне значення функції витрат (2) за $K$ розбиттями, $K = \overline{1, K_{max}}$.

Статистика 10 є допоміжним засобом для визначення найбільш оптимальної сегментації серед $K_{max}$ можливих. З меншими значеннями можна вважати, що сегментація більше відповідає початковому часовому ряду.

Ідея наступна. Нехай $X$ - часовий ряд, а $\hat{t}$ - матриця оптимальних розбиттів $X$ від 1 до $K_{max}$. Для кожного $K = \overline{1, K_{max}}$ обчислимо $\tilde{J}(K)$ і обираємо $K_* = \arg\min_{1\leqslant K\leqslant K_{max}} \tilde{J}(K)$. Тоді $\hat{t}_* = \hat{t}_{K_*}$ є оптимальною сегментацією $X$ за байєсівським критерієм.

Беремо в якості прикладу вибірку $A$ з $K_{max} = 3$ з четвертого розділу. Отримали такі значення функції витрат та статистики в залежності від сегментацій $\hat{t}_K$, $K = \overline{1, K_{max}}$:

| $K$ | 1 | 2 | 3 |
|---|---|---|---|
| $\hat{J}(K)$ | 2.75 | 0.7763 | 0.1953 |
| $\tilde{J}(K)$ | -349.1416 | -466.4109 | -595.1904 |

Звідси маємо, що $K_* = \arg\min_{1\leqslant K\leqslant K_{max}} \tilde{J}(K) = 3$ та $\hat{t}_{K_*} = \hat{t}_3 = (25, 75, 100)^\top$.

[2]Насправді форма статистики була переписана на основі критерію для оцінки якості регресійних моделей.

# 9 Графічна інтерпретація результатів DTW.

У цьому розділі об'єми часових рядів $X = (X_j)_{j=1}^N$ та $Y = (Y_i)_{i=1}^N$ однакові: $N = 100$. Значення спостережень змодельовані за формулами:

$$X_j = \cos(I_j) + \frac{\varepsilon}{10},\ Y_i = \sin(I_i),\ 1 \leqslant i, j \leqslant N$$
$$\varepsilon \sim U[0,1],\ I = \{I_1, \ldots, I_N\} = \{0, 0.0634, \ldots, 2\pi\}$$

1. **Теплова матриця відстаней для усіх спостережень з двох часових рядів. Крива оптимального шляху.**

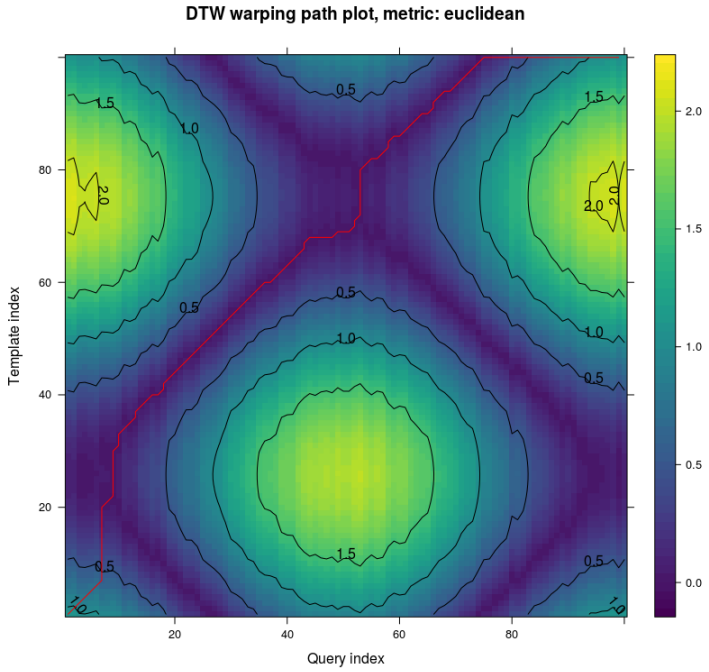


Рис. 8: Прокладення отпимального шляху алгоритму на тепловій матриці відстаней.

Спочатку малюємо теплову матрицю відстаней для усіх можливих пар $(X_i, Y_j)$, $X_i \in X$, $Y_j \in Y$. Далі, маючи оптимальних шлях алгоритму, прокладемо його на матриці таким чином: значення по осі абсцис визначають ліві елементи пар, по осі ординат, відповідно, праві елементи.

2. **Графіки двох рядів та вирівнювання кожного з них.**

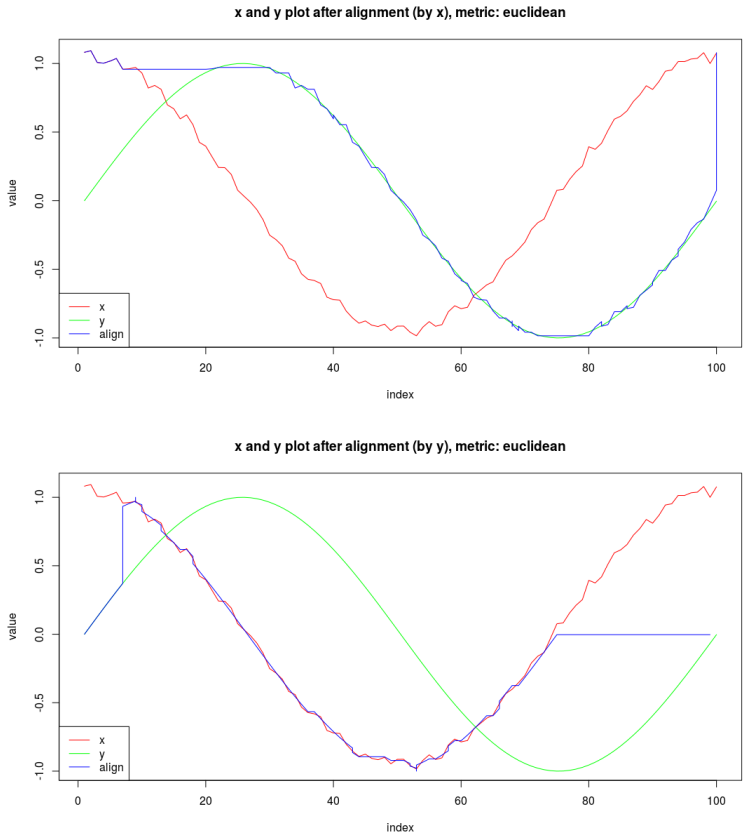


Рис. 9: Графіки з вирівнюванням відносно спостережень першої та другої вибірки відповідно.

Будуємо графіки часових рядів. Через $p_{left}$, $p_{right}$ позначено набори лівих та правиї елементів кожної пари з обраного шляху. Вирівнювання визначено так:

$$\hat{X} = X_p = X_{p_{left}},\ \hat{X} = Y_p = Y_{p_{right}}$$

3. **Графічне зображення спостережень з обох рядів та відповідностей між ними.**

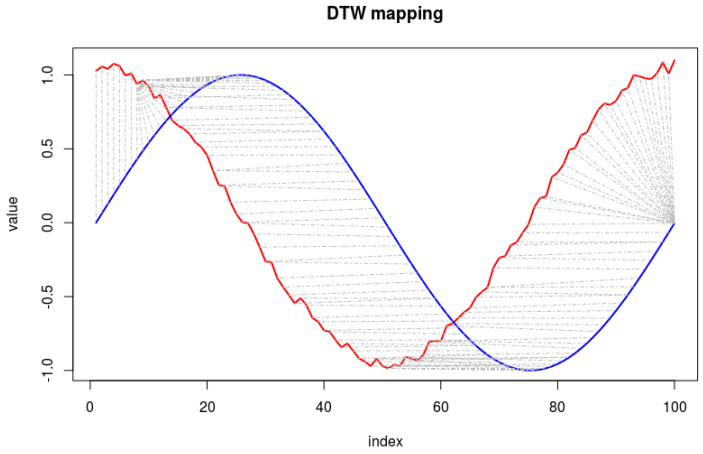


Рис. 10: Графік відповідностей між двома рядами.

Будуємо графіки часових рядів. Маючи оптимальний шлях алгоритму, ставимо у відповідність елементи першого ряду до елементів другого, показуючи це на рисунку пунктирними лініями.

# 10 Застосування базового алгоритму DTW. Приклад.

У світлі останніх подій, будемо порівнювати часові ряди, які містять спостереження про щоденну кількість захворювань на $COVID$-19 для країн Європи за період з початку березня та кінця квітня цього року. Дані наведено для наступних країн: Франція, Німеччина, Іспанія, Велика Британія. У вибірці для третьої країни маємо єдине спостереження з від'ємним значенням - невідомо, чи це значення було помилково введено, або цьому слугували інші фактори. За допомогою техніки трансформації часової шкали, зробимо короткі висновки щодо можливої ситуації в загальному. Для зручності в подальшому, дані пронормовано.

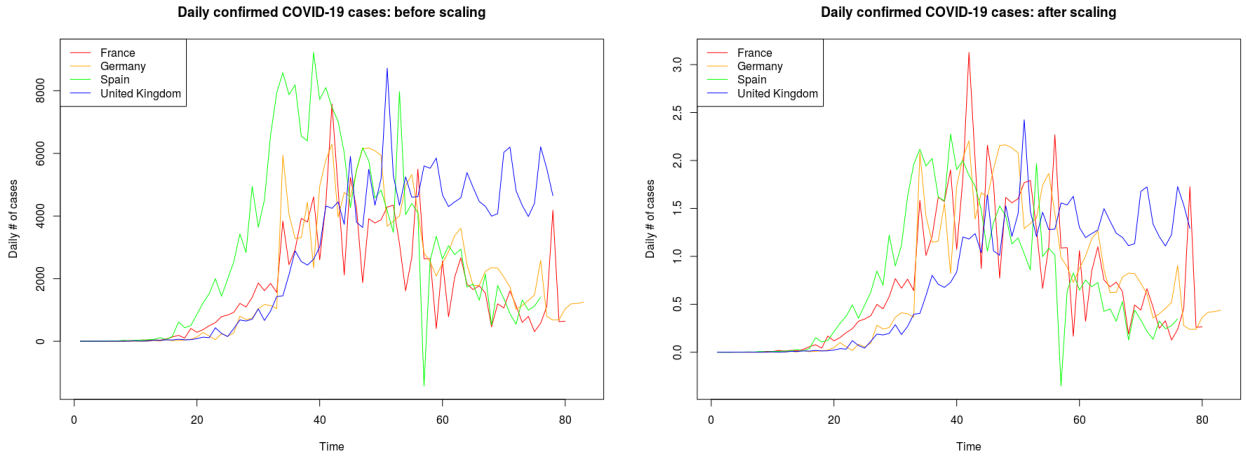


Рис. 11: Спостереження наведених часових рядів до нормування (зліва) та після (справа).

Нижче наведено мінімальні значення функції витрат для кожної пари часових рядів.

| Pair | Min.Cost |
|---|---|
| Germany - France | 0.1501 |
| Spain - France | 0.1216 |
| Spain - Germany | 0.1081 |
| United Kingdom - France | 0.2167 |
| United Kingdom - Germany | 0.2184 |
| United Kingdom - Spain | 0.2215 |

Відомо, що карантин у Франції та Іспанії було запроваджено всередині березня, трохи пізніше - в Німеччині. З (10) видно, що для цих країн форми ламаних кривих. Для Британії вона є більш розтягнутою. З отриманих результатів DTW, маємо значущу відмінність між спостереженнями про захворювання у Великій Британії: високі значення функції витрат, оптимальний шлях, у всіх трьох випадках, зсунутий від побічної діагоналі локальної матриці витрат, розглядаючи інші пари.

Якщо б у Британії карантин було введено раніше, то, ймовірно, розподіл спостережень за вказаний проміжок часу відповідного часового ряду мав більшу схожість з розподілами спостережень для перших трьох країн. Тому і щоденна кількість нових випадків на захворювання могла б стати меншою ніж тепер.

Далі показано графіки відповідностей, отриманих попарно за DTW, для рядів.

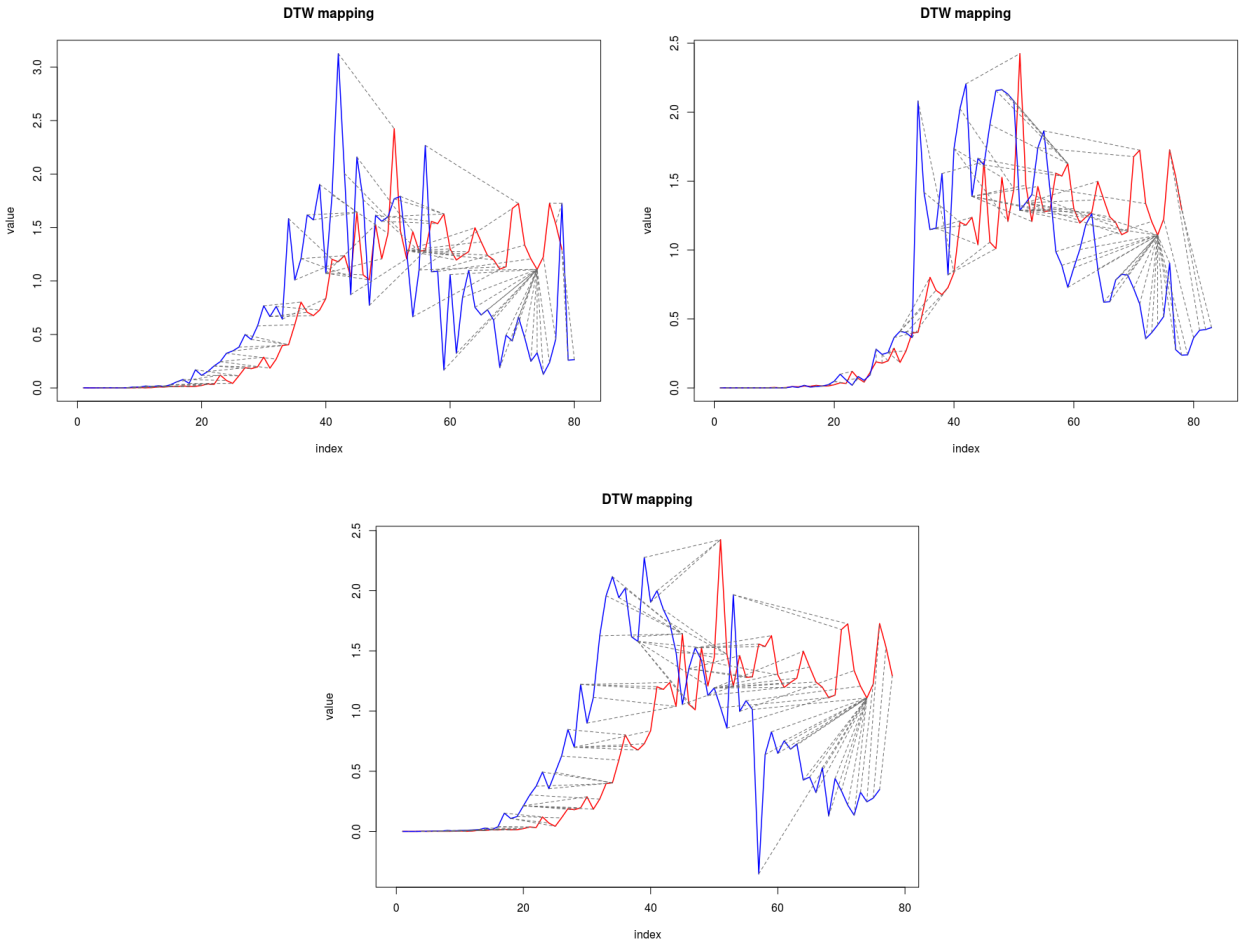


Рис. 12: Графіки відповідності спостережень для пар: ”Британія-Франція” (зліва), ”Британія-Німеччина” (справа), ”Британія-Іспанія” (внизу).

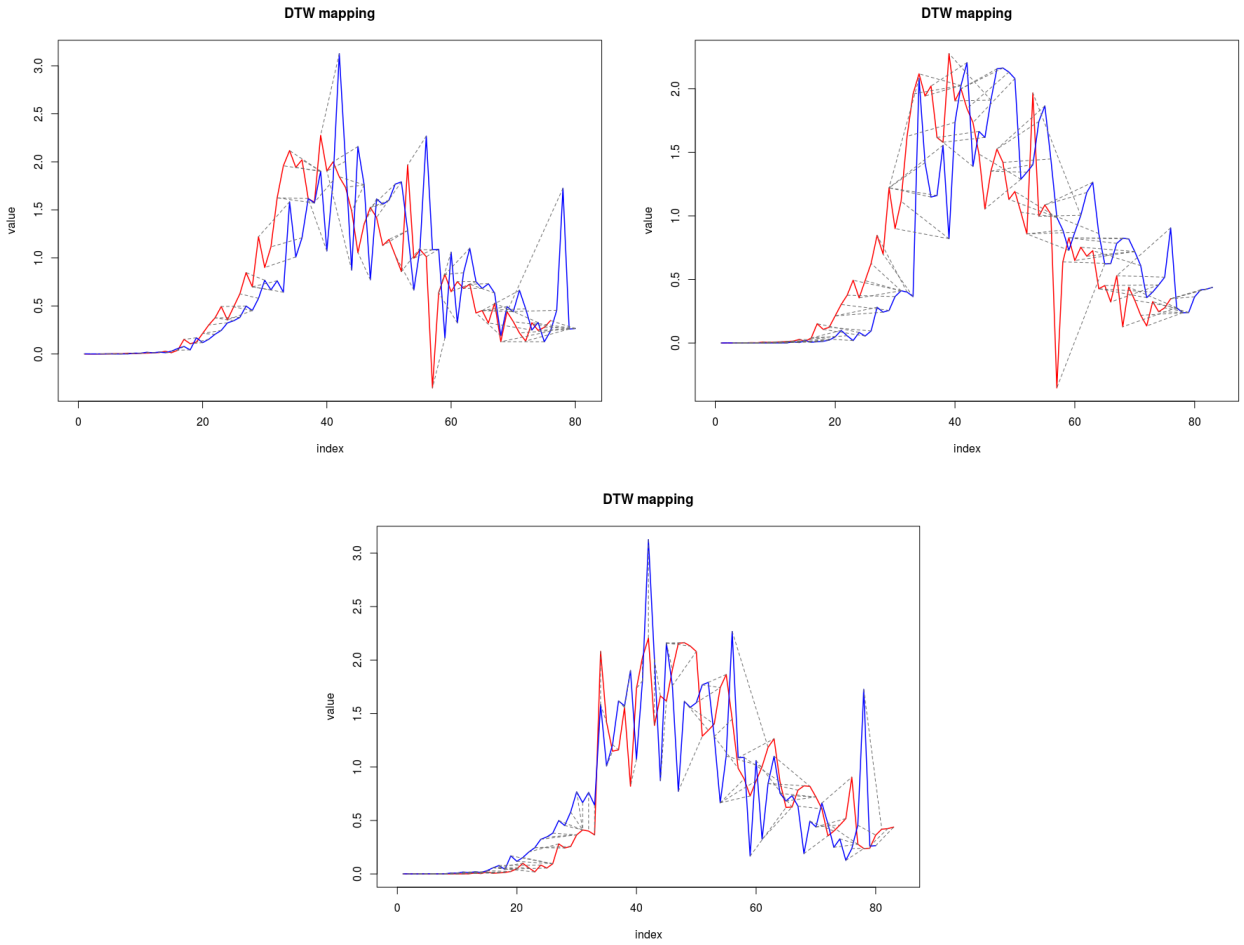


Рис. 13: Графіки відповідності спостережень для пар: "Іспанія-Франція" (зліва), "Іспанія-Німеччина" (справа), "Німеччина-Франція" (внизу).

Нарешті, покажемо теплові матриці відстаней для кожної пари.

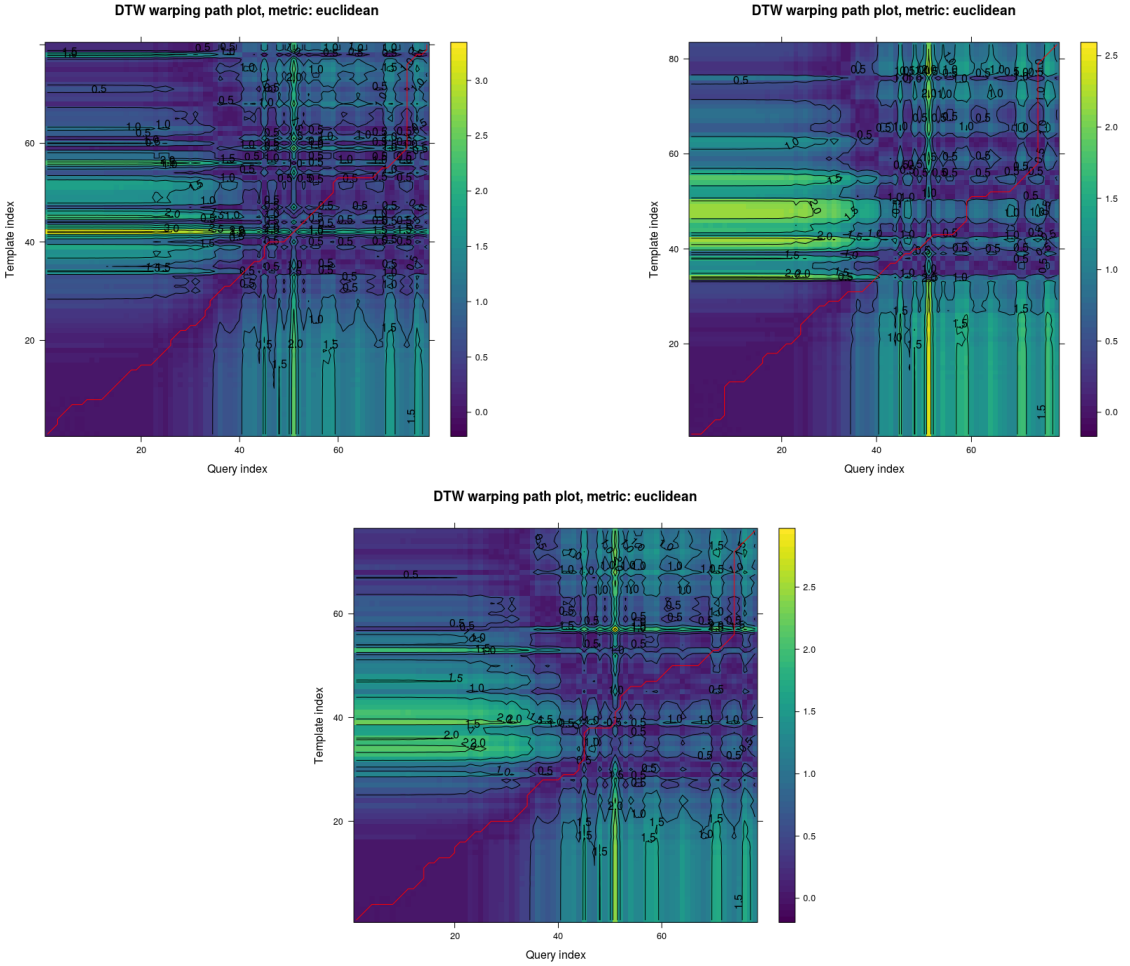


Рис. 14: Теплові матриці для пар: ”Британія-Франція” (зліва), ”Британія-Німеччина” (справа), ”Британія-Іспанія” (внизу).

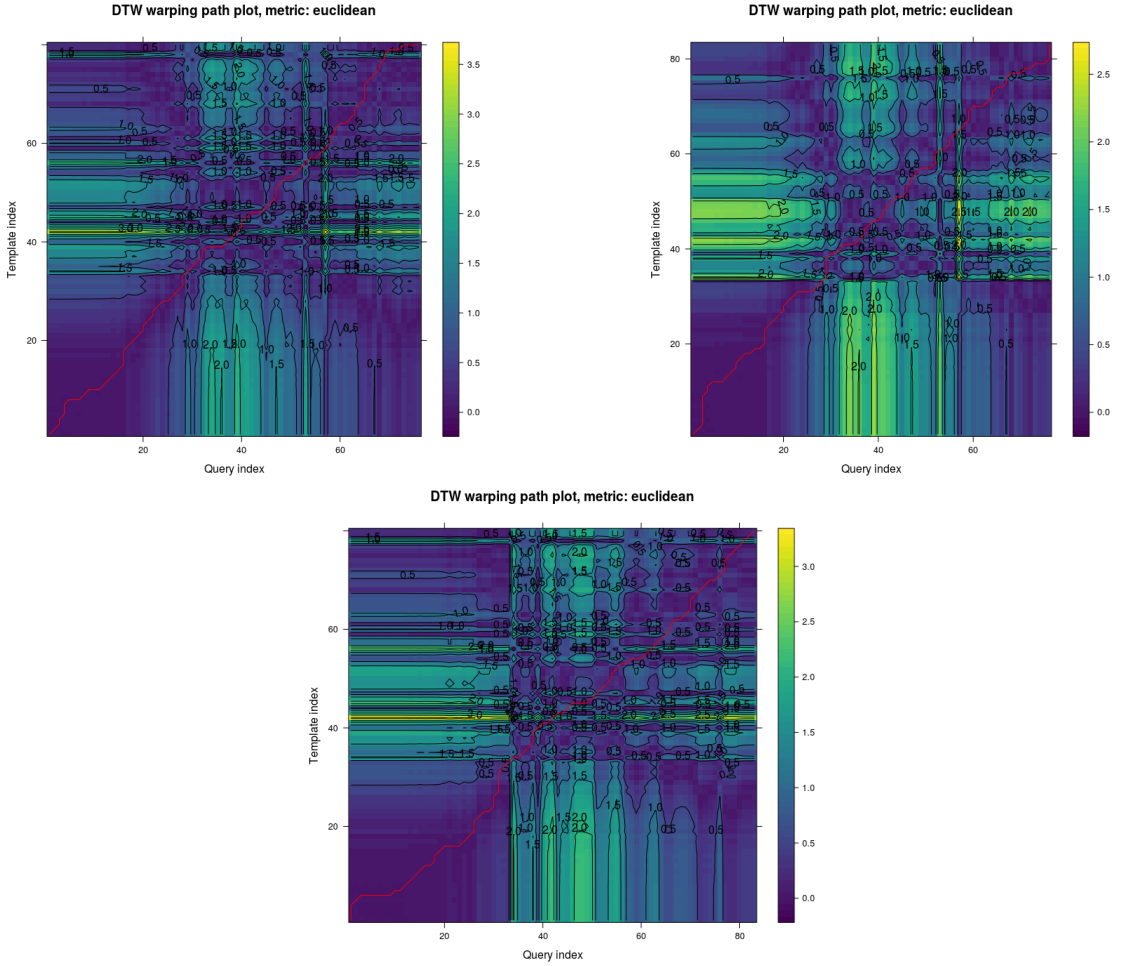


Рис. 15: Теплові матриці для пар: "Іспанія-Франція" (зліва), "Іспанія-Німеччина" (справа), "Німеччина-Франція" (внизу).