Dezagregacja danych przedziałowych

Michał Ramsza

M. Ramsza, Warsaw School of Economics
Department of Mathematics and Mathematical Economics

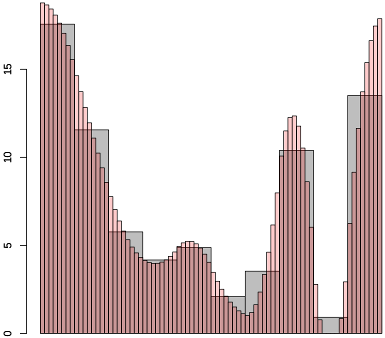
Why R? 2017

Plan prezentacji

- Problem
- Algorytm matematyczny
- Implementacja w R
- Przykłady

Problem (1)

- Mamy dane przedziałowe.
- Chcemy zdezagregaować dane na drobniejsze przedziały.



Tego typu zadanie jest często rozwiązywane w zakresie szeregów czasowych ale nie dla danych przedziałowych.

Problem (2)

- Dana jest funkcja $y : \mathbb{R} \to \mathbb{R}$.
- Dana jest siatka $\mathcal{A} = (a_0, a_1, \ldots, a_N)$, gdzie $N \in \mathbb{N}$ zawierająca $N + 1$ elementów.
- Dane jest N wartości y_n, $n = 1, \ldots, N$.
- Dana jest nadsiatka $\mathcal{A}' = (a'_0, a'_1, \ldots, a'_K)$, gdzie
  - $K \in \mathbb{N}$ i $K > N$
  - dla dowolnego $n = 0, \ldots, N$ zachodzi $a_n \in \mathcal{A}'$
  - $a'_0 = a_0$ i $a'_K = a_N$.

Zadanie polega na znalezieniu "optymalnych" wartości y'_k, $k = 1, \ldots, K$.

Algorytm matematyczny (1)

- Związek wartości y_n z funkcją y i siatką $\mathcal{A}$ jest zadany operatorem agregacji ϕ w następujący sposób

$$y_n = \phi(y, a_{n-1}, a_n).$$

- Typowy przykład takiego agregatora to suma

$$y_n = \int_{a_{n-1}}^{a_n} y(\tau) d\tau.$$

- Nowe, nieznane wartości y'_k są zadane tym samym operatorem agregacji

$$y'_k = \phi(y, a'_{k-1}, a'_k).$$

Algorytm matematyczny (2)

- Operator agregacji ϕ i zależności na siatkach definiują funkcję wiążącą ξ zadającą zależności

$$y_n = \xi\left(\{y'_k\}_{k \in K_n}\right),$$

gdzie K_n jest zbiorem indeksów postaci

$$K_n = \{k : a_{n-1} \leq a'_k \leq a_n\}.$$

- Przykładowo, dla agregatora będącego sumą zachodzi

$$y_n = \int_{a_{n-1}}^{a_n} y(\tau)d\tau = \sum_{j=1,\ldots,\ell_n} \int_{a'_{k_{j-1}}}^{a'_{k_j}} y(\tau)d\tau = \sum_{j=1,\ldots,\ell_n} y'_j,$$

gdzie $\{k_j\}_j = K_n$. Zatem ξ jest sumą.

Algorytm matematyczny (3)

- Dodatkowo do warunków spójności dodany jest selektor $\mathcal{S}$. Ściśle, selektor jest funkcją $\mathcal{S} : \mathbb{R}^K \to \mathbb{R}$ określoną na wartościach y'_k.
- Przykładowy selektor może być np. średnią zmianą

$$S(y'_0, \ldots, y'_K) = \sum_{k=1}^{k=K} (y'_k - y'_{k-1})^2.$$

Algorytm matematyczny (4)

Zadanie uzyskania nowych wartości y'_k sprowadza się do rozwiązania zadania optymalizacji postaci

$$\min_{y'_k} \mathcal{S}(y'_1, \ldots, y'_K)$$

z warunkami ograniczającym postaci

$$y_n = \xi\left(\{y'_k\}_{k \in K_n}\right), \quad n = 1, \ldots, N$$

oraz ewentualnymi dodatkowymi ograniczeniami (równości lub nierówności) dodanymi przez analityka.

Implementacja w R

- Model obiektowy S3
- Obiekt **`gridData`**.
- Typowe metody usługowe typu **`print`**, **`barplot`** itd.
- Podstawowa metoda to **`resample`** obliczającą wartości y_k'.
- Podstawowe narzędzie to pakiet **`alabama`** (Ravi Varadhan) implementujący algorytm Augmented Lagrangian Adaptive Barrier Minimization Algorithm

  `https://cran.r-project.org/web/packages/alabama/index.html`.

Great. Let's see the code ...