Een gegeneraliseerde aanpak voor automatische foutlocalisatie

Sander Scholtus

(s.scholtus@cbs.nl)

Centraal Bureau voor de Statistiek

Automatische controle en correctie

- Doel: geautomatiseerd verbeteren fouten in microdata
- Twee stappen:
  - detecteren foutieve waarden (*foutlocalisatie*)
  - imputeren van nieuwe waarden
- Voordelen automatische controle en correctie:
  - efficiënt
  - snel
  - reproduceerbaar
- In evaluaties: systematische verschillen tussen handmatig en automatisch gecontroleerde data
  - aanname: handmatig gecontroleerde data zijn correct

Automatische foutlocalisatie

- Stel controleregels op
  - idee: fouten <--> geschonden controleregels
- Pas de data aan zodat ze voldoen aan de regels

Omzet ≥ 0

ALS (Leeftijd < 12) DAN (Inkomen = 0)

Omzet – Kosten = Resultaat

0 ≤ Leeftijd ≤ 120

Automatische foutlocalisatie

– Principe van Fellegi en Holt (1976):

> Bepaal de kleinste verzameling variabelen die kan worden geïmputeerd zodat voldaan wordt aan alle controleregels.

– Voorbeeld:

Omzet – Kosten = Resultaat
Kosten ≥ 60% × Omzet
Omzet ≥ 0
Kosten ≥ 0

	Omzet	Kosten	Resultaat
oorspronkelijke data	200	125	755
gecorrigeerde data	200	125	75

Automatische foutlocalisatie

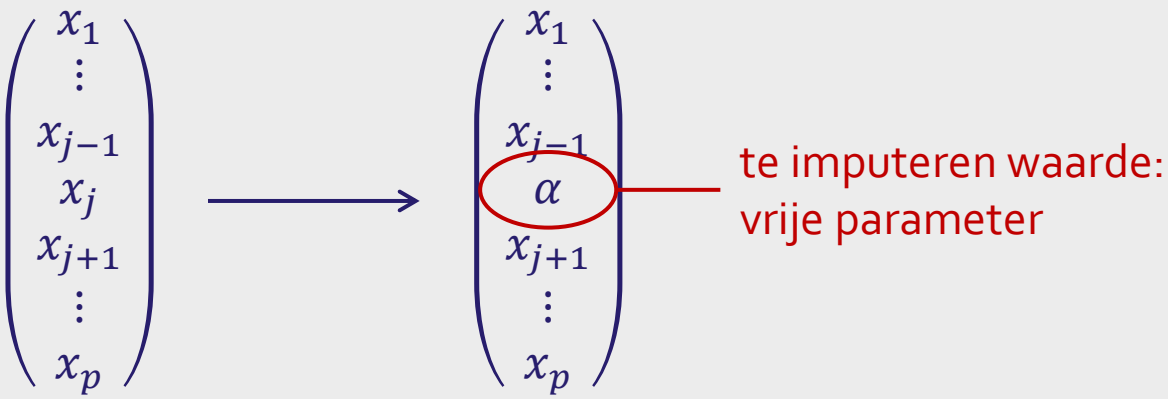
- Aanname bij Fellegi-Holt: elke variabele onafhankelijk wel/niet fout
- In de praktijk zijn er fouten waarvoor dit niet geldt
  - Voorbeeld 1: baten-lasten-verwisselingen

	Omzet	Kosten	Resultaat
oorspronkelijke data	70	130	60
data na automatische correctie	190	130	60
data na handmatige correctie	130	70	60

  - Voorbeeld 2: overhevelingen tussen variabelen
    - bijv. *omzet uit groothandel / omzet uit detailhandel*

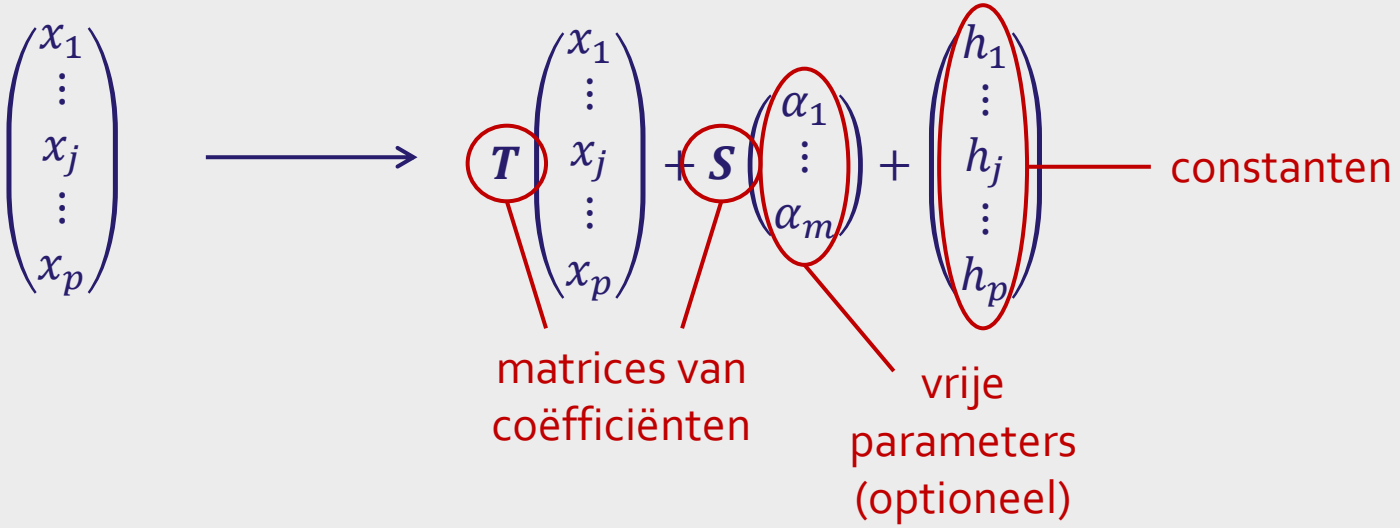
Edit-acties

– Principe van Fellegi en Holt: één soort edit-actie



Edit-acties

– Idee: algemenere lineaire edit-actie



Edit-acties

– Voorbeelden:

- Tekenwisseling

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \rightarrow \begin{pmatrix} -x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

- Verwisseling van twee aangrenzende waarden

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \rightarrow \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

- Overheveling tussen twee variabelen

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \rightarrow \begin{pmatrix} x_1 + \alpha \\ x_2 \\ x_3 - \alpha \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \cdot (\alpha) + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

- De “Fellegi-Holt”-actie

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \rightarrow \begin{pmatrix} x_1 \\ \alpha \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \cdot (\alpha) + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Edit-acties

- Specificeer toegelaten edit-acties (+ gewichten)
- Paden tussen records:

$$\boldsymbol{x} = \boldsymbol{x}_0 \xrightarrow{\text{actie 1}} \boldsymbol{x}_1 \xrightarrow{\text{actie 2}} \cdots \xrightarrow{\text{actie } t} \boldsymbol{x}_t = \boldsymbol{y}$$

- Padlengte: som van de gewichten bij de acties

- Gegeneraliseerd principe van Fellegi en Holt:

  Bepaal het kortste pad van toegelaten gaafmaakacties dat leidt tot een record dat voldoet aan alle controleregels.

- Oorspronkelijk principe van Fellegi en Holt als speciaal geval: alleen de “Fellegi-Holt”-acties zijn toegelaten

Simulatiestudie

- Vijf variabelen: OmzetHoofd, OmzetOverig, OmzetTot, KostenTot, Resultaat
- Controleregels:

  OmzetHoofd + OmzetOverig = OmzetTot
  OmzetTot – KostenTot = Resultaat
  OmzetHoofd ≥ OmzetOverig
  –10% × OmzetTot ≤ Resultaat ≤ 50% × OmzetTot
  OmzetHoofd, OmzetOverig, OmzetTot, KostenTot ≥ 0

- Toegelaten edit-acties:
  - de vijf “Fellegi-Holt”-acties
  - verwisseling van OmzetTot en KostenTot (Verw)
  - overheveling van OmzetOverig naar OmzetHoofd (niet andersom) (Ovh)
  - tekenwisseling in KostenTot (Tek1)
  - tekenwisseling in Resultaat (Tek2)
- Gesimuleerde data:
  - Foutloze data: afgeknotte normale verdeling
  - Ruwe data: fouten toegevoegd die horen bij toegelaten edit-acties
  - 1025 records met 1, 2 of 3 fouten

Simultatiestudie

– Resultaten m.b.t. kiezen juiste edit-acties

aanpak	α	β	δ	ρ^C
alleen Fellegi-Holt-acties	74%	12%	23%	80%
alle edit-acties	14%	3%	5%	24%
…behalve Verw	29%	5%	9%	35%
…behalve Ovh	34%	5%	10%	37%
…behalve Tek1	28%	6%	9%	39%
…behalve Tek2	35%	7%	10%	47%

Simultatiestudie

– Resultaten m.b.t. aanwijzen foute variabelen

aanpak	α	β	δ	ρ^c
alleen Fellegi-Holt-acties	19%	10%	13%	32%
alle edit-acties	10%	5%	7%	17%
…behalve Verw	15%	9%	11%	29%
…behalve Ovh	10%	5%	7%	18%
…behalve Tek1	10%	5%	7%	17%
…behalve Tek2	11%	6%	7%	18%

Algoritmes

- Scholtus (2016): algoritme voor het oplossen van het gegeneraliseerde foutlocalisatieprobleem
  - Gebaseerd op generalisatie van een techniek (Fourier-Motzkin-eliminatie) die eerder is gebruikt voor foutlocalisatie volgens het oorspronkelijke Fellegi-Holt-principe
  - Toegepast in simulatiestudie
  - Niet toepasbaar op (veel) grotere problemen, i.v.m. rekentijd

Algoritmes

- Nieuw onderzoek (met Jacco Daalmans):
  - Foutlocalisatie als Mixed Integer Linear Programming-probleem
  - Bekende aanpak voor oorspronkelijk Fellegi-Holt-principe (bijv. geïmplementeerd in R-pakket **`editrules`**)
  - Gegeneraliseerd probleem ook in deze vorm geschreven
  - In principe toepasbaar op grote problemen

  - Toepassing op echte data van de zgn. Productiestatistieken
    - ruim 100 variabelen
    - ongeveer 200 controleregels

Conclusie

- Nieuwe formulering foutlocalisatieprobleem
  - Houd rekening met fouten die meerdere variabelen tegelijk raken
  - Meer flexibiliteit bij automatische foutlocalisatie
- Resultaten op gesimuleerde data positief
- Voor toepassing in de praktijk nodig:
  - Meer inzicht in werkwijze handmatige foutlocalisatie
    - Wat zijn relevante edit-acties?
    - Wat zijn relevante controleregels?
    - Wat zijn ‘optimale’ gewichten?

– Referenties:

- I.P. Fellegi en D. Holt (1976), A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* **71**, 17–35.
- S. Scholtus (2016), A generalized Fellegi-Holt paradigm for automatic error localization. *Survey Methodology* **42**, 1–18.