Statistische beveiliging van output

Eric Schulte Nordholt (e.schultenordholt@cbs.nl)

1 november 2018, CBS Den Haag

Centraal Bureau voor de Statistiek

Inhoud

- Inleiding
- Noodzaak van statistische beveiliging
- Geschiedenis
- Outputcontrole
- Soorten output
- Criteria
- Meer informatie, Conclusies en Discussie

Inleiding (1)

Wat zijn vertrouwelijke data?
Data die niet ‘zomaar’ kunnen worden gepubliceerd

- Vanwege wetgeving
- Als ze gevoelige gegevens bevatten
- Als de respondent/registerhouder de aan het CBS toevertrouwde gegevens als gevoelig beschouwt

Wat is statistische beveiliging?

- Fysieke beveiliging (toegang tot gebouw en netwerk)
- Wettelijke beveiliging (eed zweren en veiligheidsverklaring tekenen)
- Beveiligen van de statistische output!

Inleiding (2)

Vijf kernstappen:

1. **Waarom is statistische beveiliging nodig?**
2. **Wat zijn de kenmerkende karakteristieken en soorten gebruik van de data?**
3. **Tegen welke onthullingsrisico's moet worden beschermd?**
4. **Beveiligingsmethoden**
5. **Uitvoering**

Noodzaak van statistische beveiliging (1)

- Is statistische beveiliging echt noodzakelijk in dit specifieke geval?

- Wat voor soort informatie kan worden afgeleid?
  - Gevoelige informatie?
  - Algemeen bekende informatie?
    - Vrij beschikbaar?

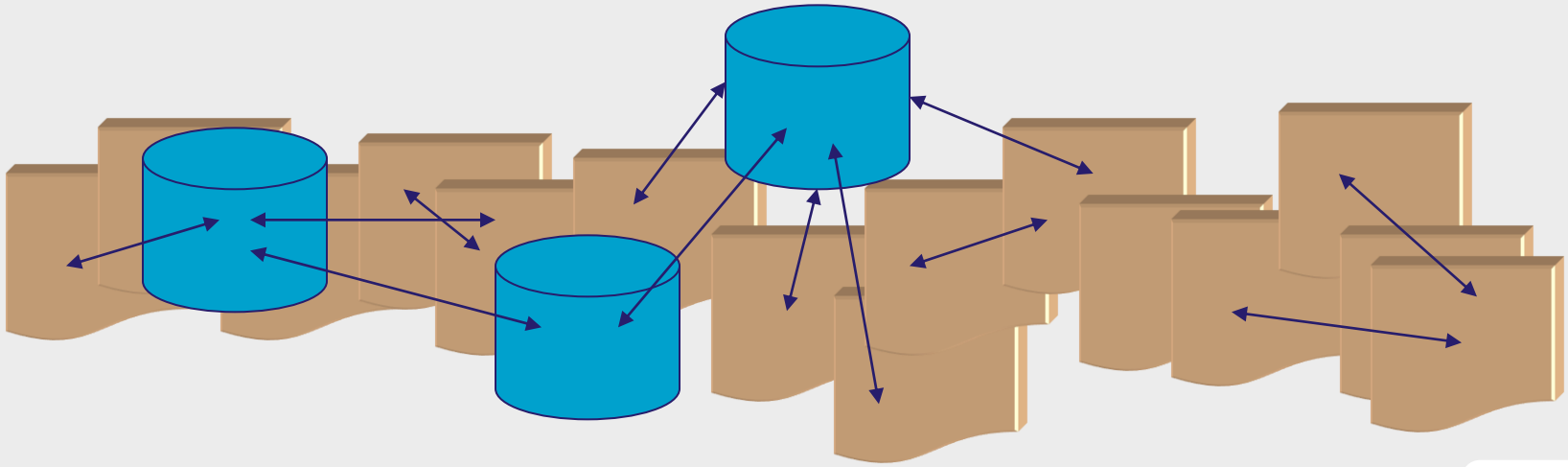
- Groepsonthulling of statistiek?

Noodzaak van statistische beveiliging (2)

- Wetten
  - Internationaal (EU)
  - Nationaal (in Nederland)
- Respecteren respondent
  - Huidige respons
  - Toekomstige respons
- Respecteren eigenaren van registraties

Noodzaak van statistische beveiliging (3)

- Linken
  - Registraties
  - Surveys
- Datamining technieken
- Gemakkelijker zoeken



Noodzaak van statistische beveiliging (4)

Veranderingen van datakarakteristieken:

- Beschikbaarheid
  - Administratieve registraties
  - Krachtige computers
- Mate van detail
- Actualiteit

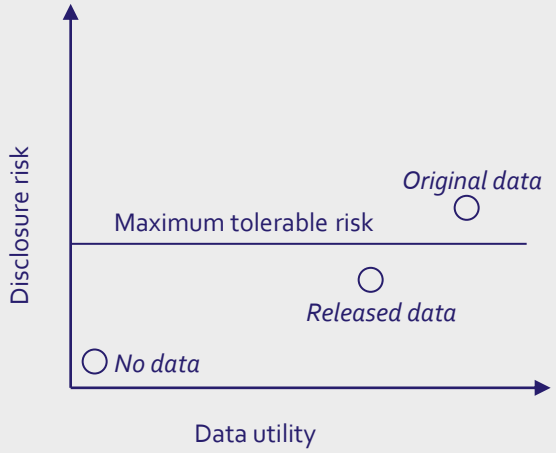
Noodzaak van statistische beveiliging (5)

Pas beveiligingstechnieken zodanig toe dat

- ⓘ De resulterende data veilig zijn
- ⓘ Het informatieverlies minimaal is

Problemen

- Definitie veilige data
- Definitie informatieverlies



Geschiedenis (1)

Traditionele output van een statistisch bureau:

- Tabellen
- Grafieken

Echter:

- Groeiende behoefte aan informatie
- Groeiende behoefte aan microdata
- IT-mogelijkheden
- Mogelijkheden voor analyse
- Steeds meer samenwerkingsverbanden

Geschiedenis (2)

Eerste stap:

- Grotere tabellen
- Beveiligingsrisico's voor kwantitatieve tabellen (τ-ARGUS)
- Online publicatie van tabellen / geaggregeerde data
    - Eerst beveiligen, dan publiceren (voorbeelden: StatLine, Census Hub)
    - Eerst tabellen aanvragen, dan beveiligen (voorbeelden: Factfinder van het U.S. Census Bureau, Table Builder van het ABS)

Geschiedenis (3)

Tweede stap:

- PUF (Public Use Files) (μ-ARGUS)
- MUC (Microdata files Under Contract) (μ-ARGUS)

Derde stap:

- Remote access

Het CBS moet de privacy van individuele bedrijven en personen beschermen (AP)

Outputcontrole (1)

Hoe moeten we output checken?

Output kan van alles zijn …

Bewerkte data, gecombineerde data, …

Set regels die alle output checkt? ***Mission impossible?***

Handen ineengeslagen:
https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf

Outputcontrole (2)

Algemene aanpak

- Output kan worden verdeeld in verschillende categorieën (b.v. tabellen, regressie-analyses, enz.)
- Elke categorie kan worden geclassificeerd als ‘veilig’ of ‘onveilig’
- Output checking is context specifiek
- Het doel van het classificeren is om criteria te geven of de output kan worden vrijgegeven

Outputcontrole (3)

Controle van output is optimaal als

- Maximaal gebruik van de data kan worden gemaakt (minimaliseren informatieverlies)
  met
- Minimaal onthullingsrisico (binnen toegestane kaders)

Outputcontrole (4)

Twee soorten fouten bij controle van output

- Beveiligingsfout: vrijgeven van onveilige output
- Inefficiencyfout: niet vrijgeven van veilige output

Doel: beide soorten fouten voorkomen!

Outputcontrole (5)

Hoe proberen we dat te benaderen?

- Veel moeite van output checkers om aangeboden output die niet aan criteria voldoet nader te bekijken
- Hulp van onderzoekers die output goed documenteren en omvangrijke output vooraf aankondigen

Key word: samenwerking!

Outputcontrole (6)

Praktijk:

- Criteria als startpunt
- Focus op output die niet aan criteria voldoet
- Met elkaar bekijken of die output toch kan worden vrijgegeven

Soorten output (1)

Algemeen criterium veilig / onveilig

Veilig:
kan worden vrijgegeven zonder of met minimale veranderingen
(slechts beperkt aantal uitzonderingen) (✓)

Onveilig:
kan niet worden vrijgegeven, tenzij wordt aangetoond dat de output niet onthullend is ( )

Soorten output (2)

Beschrijvende statistiek

- Frequentietabellen 
- Kwantitatieve tabellen 
- Minimum, maximum, percentielen 
- Modus ✓
- Gemiddelden, indices, ratios, indicatoren 
- Concentratiequotiënten ✓
- Hogere momenten (b.v. varianties) ✓
- Grafieken 

Soorten output (3)

Verklarende statistiek

- Lineaire regressiecoëfficiënten ✓
- Niet-lineaire regressiecoëfficiënten ✓
- Schattingsresiduen 
- Samenvattende maten (b.v. R^2) ✓
- Toetsen (b.v. χ^2) ✓
- Correlatiecoëfficiënten ✓
- Factoranalyse ✓
- Correspondentieanalyse ✓

Criteria (1)

Drempelwaarde:

– Minimaal 10 eenheden (ongewogen) per cel, datapunt, enz.

Model:

– Minimaal 10 vrijheidsgraden, model gebaseerd op minimaal 10 eenheden

(aantal vrijheidsgraden = # eenheden - # parameters - # andere restricties van het model)

Criteria (2)

Groepsonthulling:

– Elke cel bevat minder dan 90% van de eenheden in zijn rij/kolom

Dominantie:

– Grootste bijdrager aan een cel heeft minder dan 50% van het celtotaal

Meer informatie

CBS-wet:

- Autonome overheidsorganisatie
- Vrije toegang tot andere overheidsdata
- Statistische beveiliging is een wettelijke plicht
- Toegang verlenen tot microdata voor wetenschappelijk onderzoek

CBS-methoden voor statistische beveiliging:

- Methodenreeks (https://www.cbs.nl/nl-nl/onze-diensten/methoden/statistische-methoden/output/output/statistische-beveiliging)

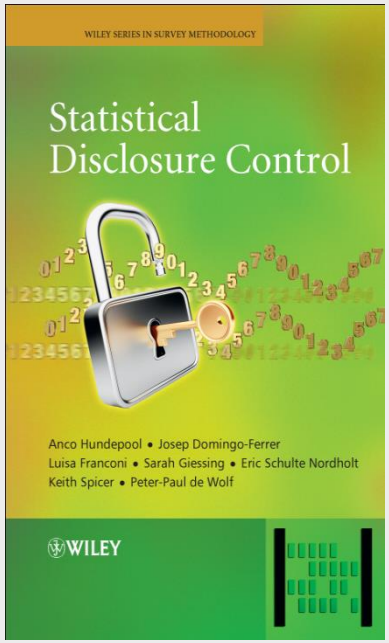
Boek:

WILEY

Statistical Disclosure Control

Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, Peter-Paul de Wolf

Order your copy online at www.wiley.com



ISBN: 978-1-1199-7815-2
Hardcover, 302 pages
September 2012, US $99.95

Statistical Disclosure Control:

- Presents a combination of both theoretical and practical solutions
- Introduces all the key concepts and definitions involved with statistical disclosure control
- Provides a high level overview of how to approach problems associated with confidentiality
- Provides a broad-ranging review of the methods available to control disclosure
- Explains the subtleties of group disclosure control
- Features examples throughout the book along with case studies demonstrating how particular methods are used

Conclusies

Ik heb geprobeerd een kort overzicht te geven over:

- Noodzaak van statistische beveiliging
- Geschiedenis van de statistische beveiliging
- Outputcontrole, soorten output en criteria

Bedenk echter:

– In een beperkte tijd kan niet alles aan de orde komen
– Er zijn nog vele andere beveiligingsonderwerpen, b.v. tabelbeveiliging, beveiliging van microdata, methoden, en software

Discussie

- Zijn er vragen of opmerkingen?
- Is verdere communicatie gewenst (b.v. externe presentatie)?
- Is er behoefte aan opleidingen?