

Representativiteit van respons in de Korte Termijn Statistiek van bedrijven

Guido de Nooij, Barry Schouten, Ger
Snijkers, Pieter Vlag

Centraal Bureau voor de Statistiek



Korte Termijn Statistiek (STS)

- **Verplicht vanuit Eurostat**
- **Bedrijven zijn verplicht deel te nemen**
- **Maandstatistiek over productie in alle branches**
- **Steekproeven uit Algemeen Bedrijven Register (ABR)**
- **Opgave elektronisch of schriftelijk**
- **Publicatie volgens 1:1 norm; na ongeveer 30 dagen**

- **KS 2007**
- **Bekend vooraf: aantal werknemers, branche (SBI), BTW 2006, mode van waarneming, datum enquête binnen**
- **Bekend achteraf: BTW 2007**
- **Probleem: fiscale eenheden en ABR eenheden verschillend**



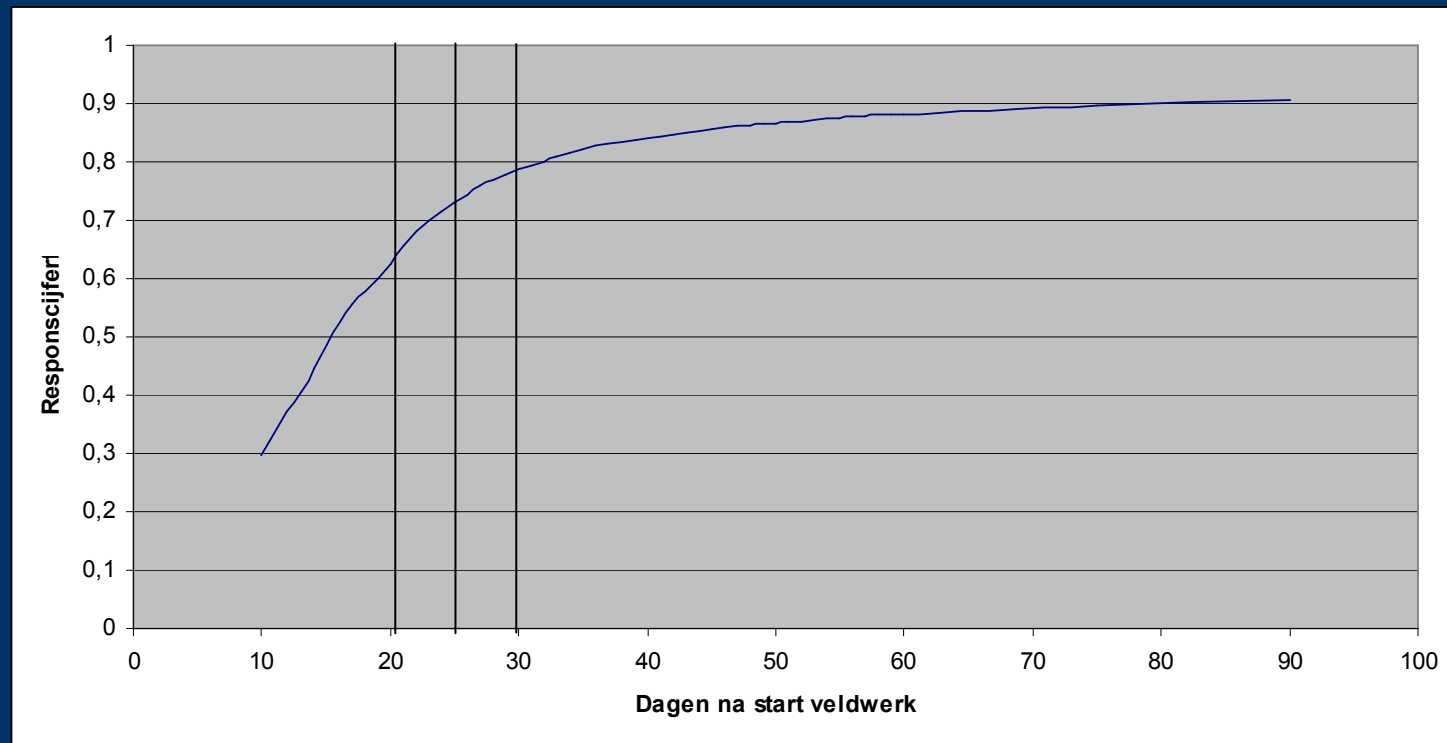
Onderzoeksvragen

- **Is de KS representatief na ongeveer 30 dagen?**
- **Welke bedrijven verdienen meer/minder aandacht?**
- **Maakt de mode van waarneming uit?**
- **Maakt de maand van waarneming uit?**

Kunnen we representativiteit meten als functie van tijd?



Respons in KS



Representativiteit

Wanneer is respons representatief?

- Vanuit oogpunt gebruiker van de survey
- Vanuit oogpunt veldwerk management

Definitie (sterk): De respons op een onderzoek is representatief t.o.v. de steekproef als individuele responskansen gelijk zijn en als de respons van een eenheid onafhankelijk is van de respons van andere eenheid.

Definitie (zwak): De respons op een onderzoek is representatief voor een variabele X als de gemiddelde responskansen over categorieën van X constant is.



Het meten van representativiteit

Via spreiding in geschatte responskansen

$$\hat{R}^t(\hat{\rho}^t) = 1 - 2\sqrt{\frac{1}{N-1} \sum_{i=1}^N \frac{s_i}{\pi_i} (\hat{\rho}_i^t - \hat{\rho}^t)^2}$$

Interpretatie via maximale vertekening

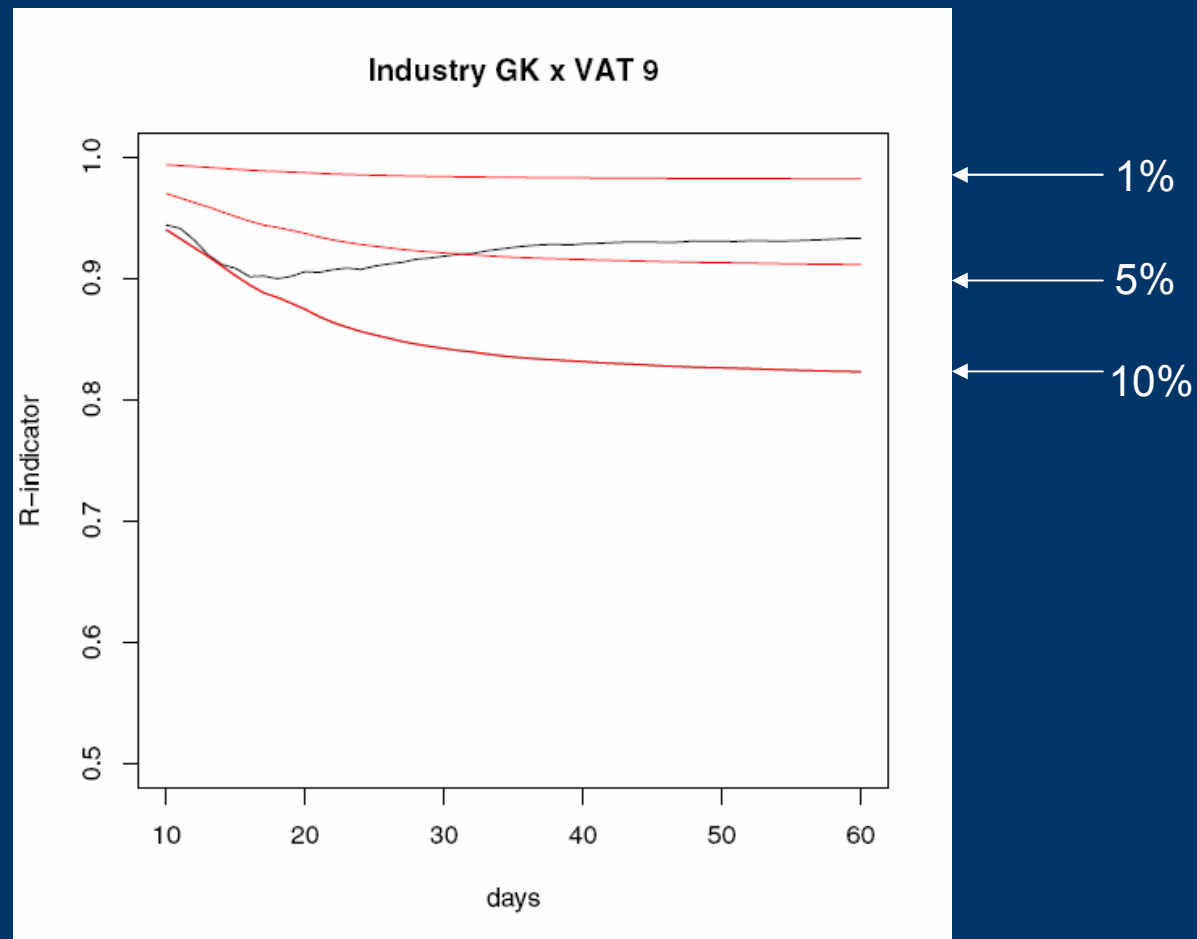
$$\left| B(\hat{y}_{HT}^t) \right| \leq \frac{S(\rho^t)S(y)}{\bar{\rho}^t} = \frac{(1 - R(\rho^t))S(y)}{2\bar{\rho}^t} \qquad \frac{\left| B(\hat{y}_{HT}^t) \right|}{S(y)} \leq \frac{S(\rho^t)}{\bar{\rho}^t} = \frac{1 - R(\rho^t)}{2\bar{\rho}^t}$$

Respons-Representativiteit curves

$$\gamma \geq \frac{1 - R(\hat{\rho}^t)}{2\hat{\rho}^t} \qquad R(\hat{\rho}^t) \geq 1 - \gamma 2\hat{\rho}^t$$



Respons-Representativiteit curves



Het schatten van responskansen

Representativiteit

- Vanuit veldwerk: vast model en variabelen waarop je kunt sturen
- Vanuit onderzoeker: beste model en variabelen die gerelateerd zijn aan doelvariabelen

Voor KS

- Kandidaten: aantal werknemers, branche, BTW 2006
- Vast model: aantal werknemers x BTW in klassen



Validatie met KS en BTW 2007

Wanneer zijn indicatoren bruikbaar?

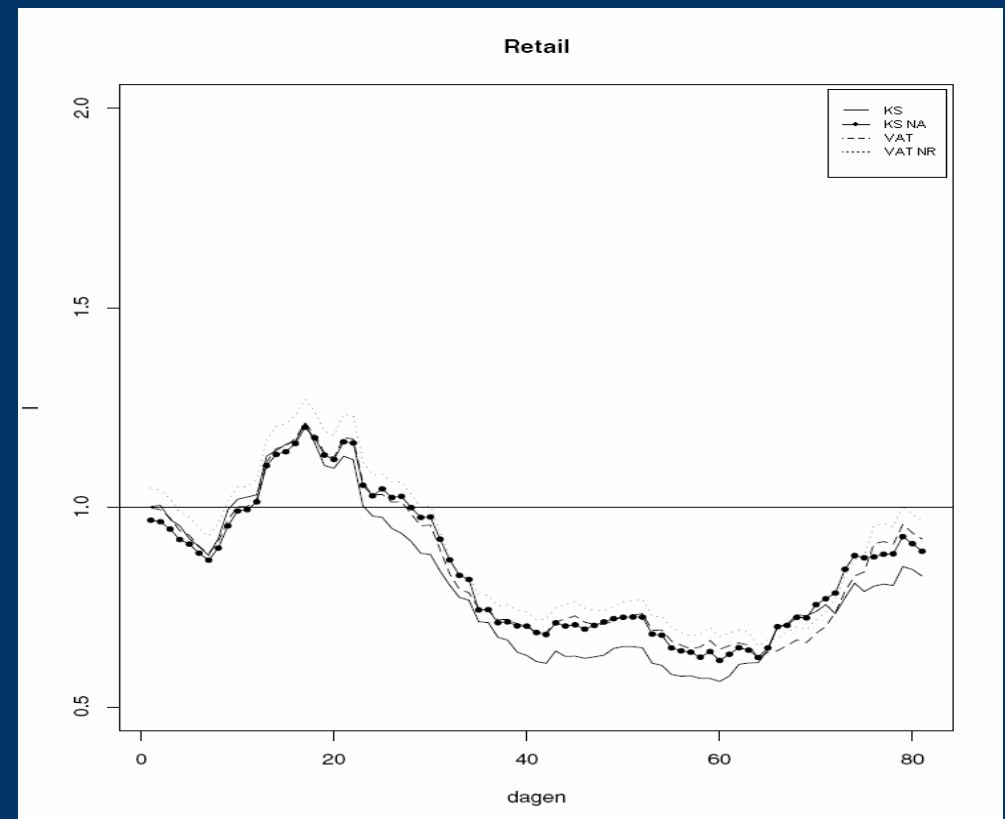
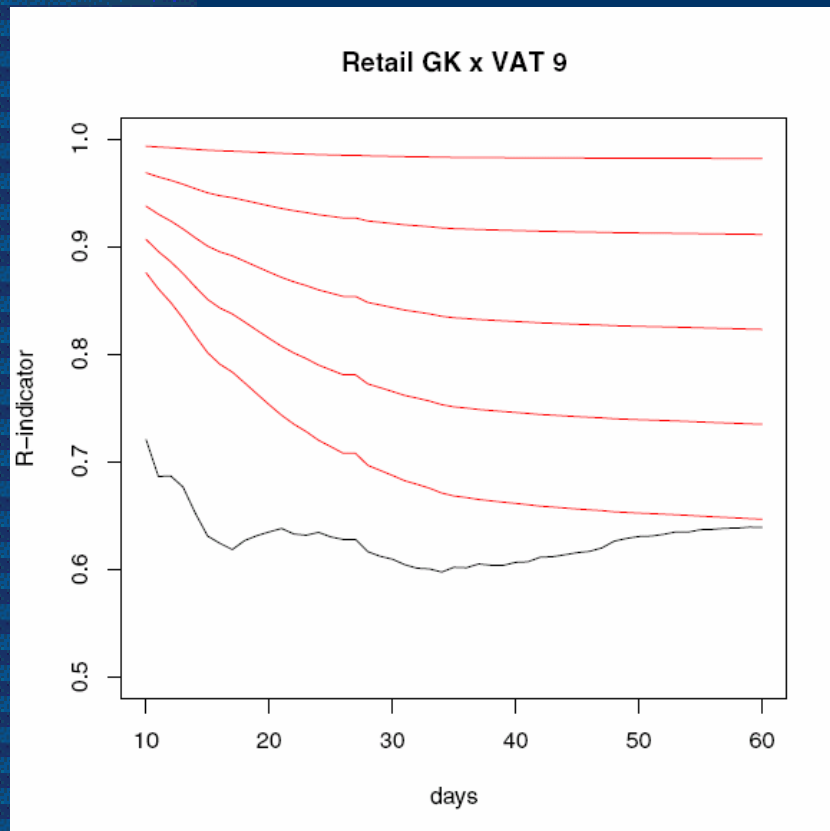
- Als patronen R-indicator en vertekening overeenkomen
- Als vertekening binnen worst-case-scenario vertekening blijft
- Als betrouwbaarheidsintervallen relatief klein zijn

Hoe te valideren?

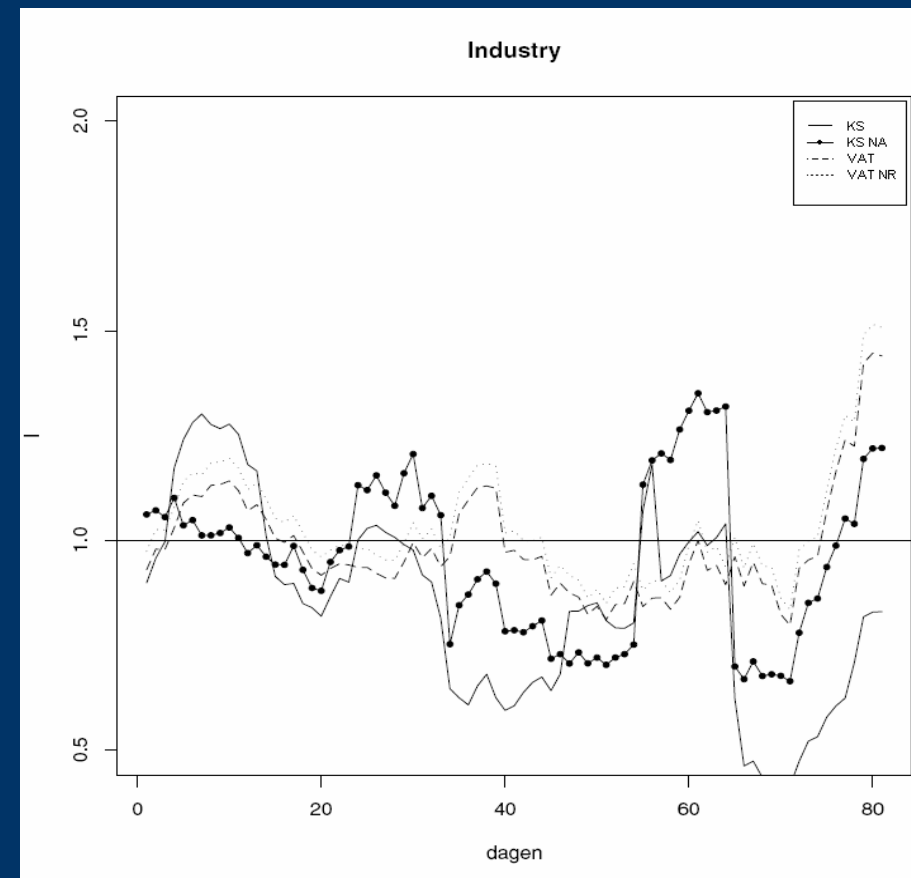
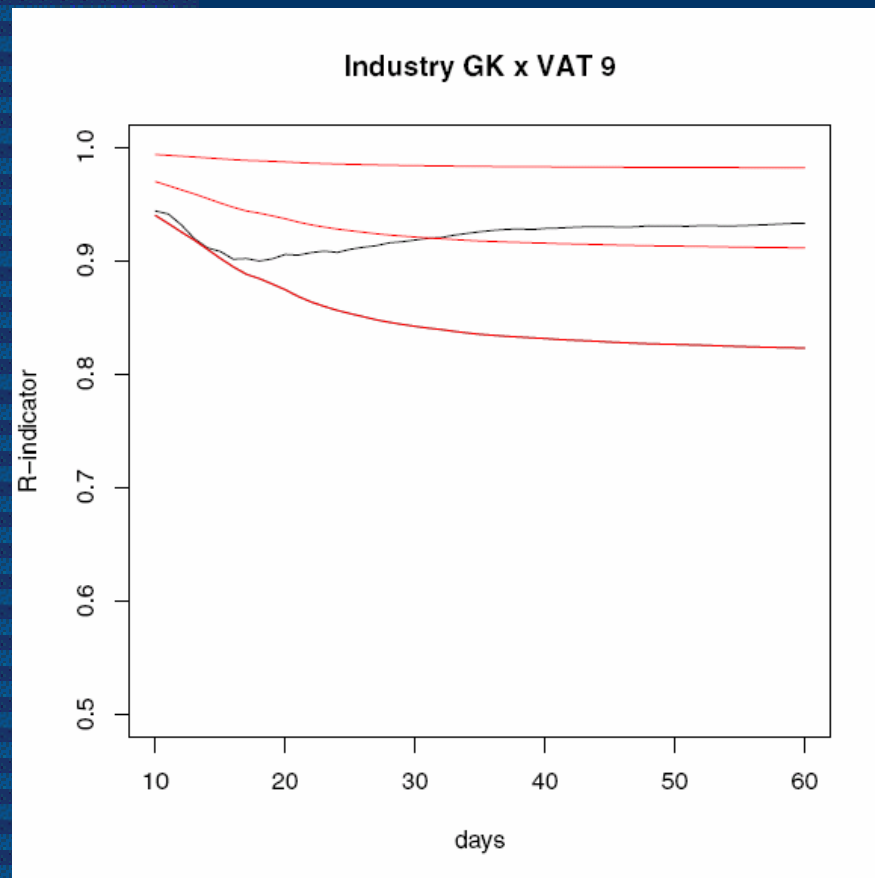
- **Cumulatieve gemiddelde (gewogen) omzet op t / gemiddelde gewogen omzet op $t=90$**
- **Gemiddelde (gewogen) omzet op t / gemiddelde gewogen omzet op $t=90$**
- **Voor zowel KS als BTW mogelijk**
- **BTW ook voor non-respons, maar mislukte koppelingen**



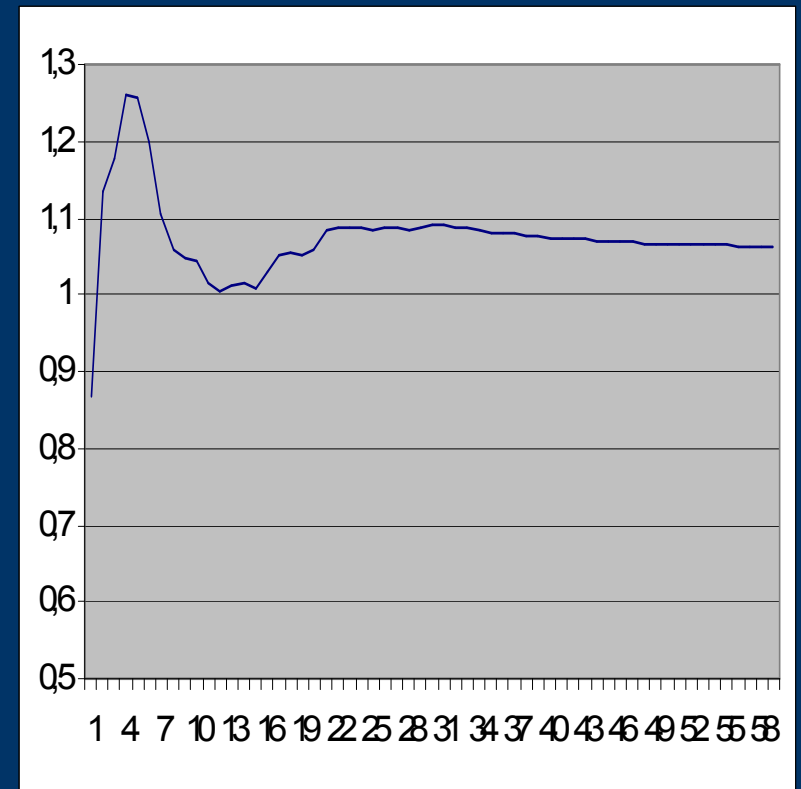
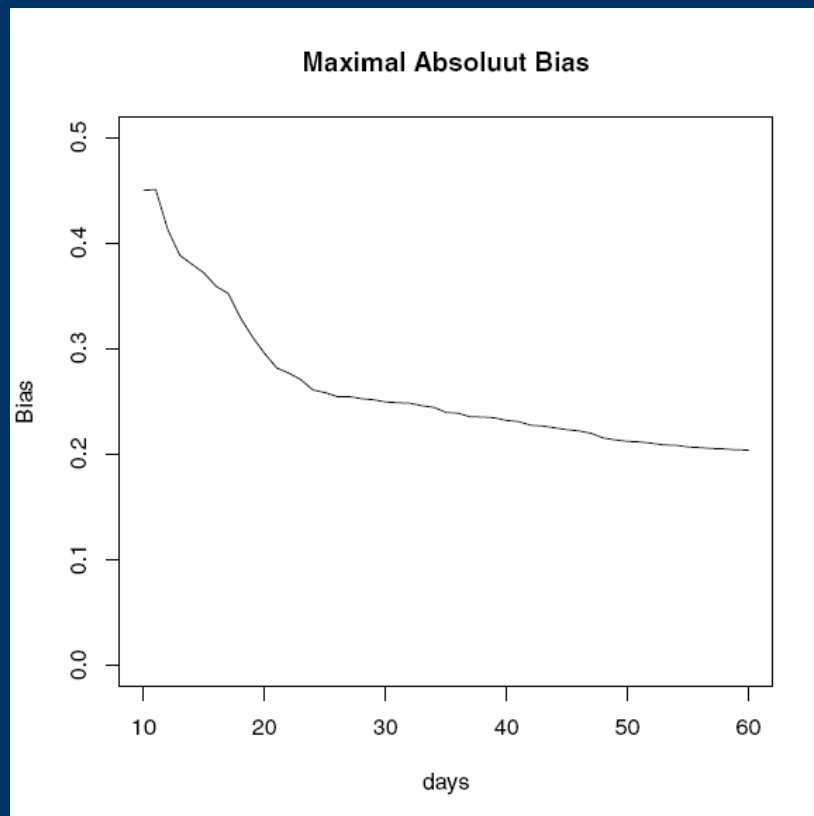
R-indicator en gemiddelde omzet: detailhandel



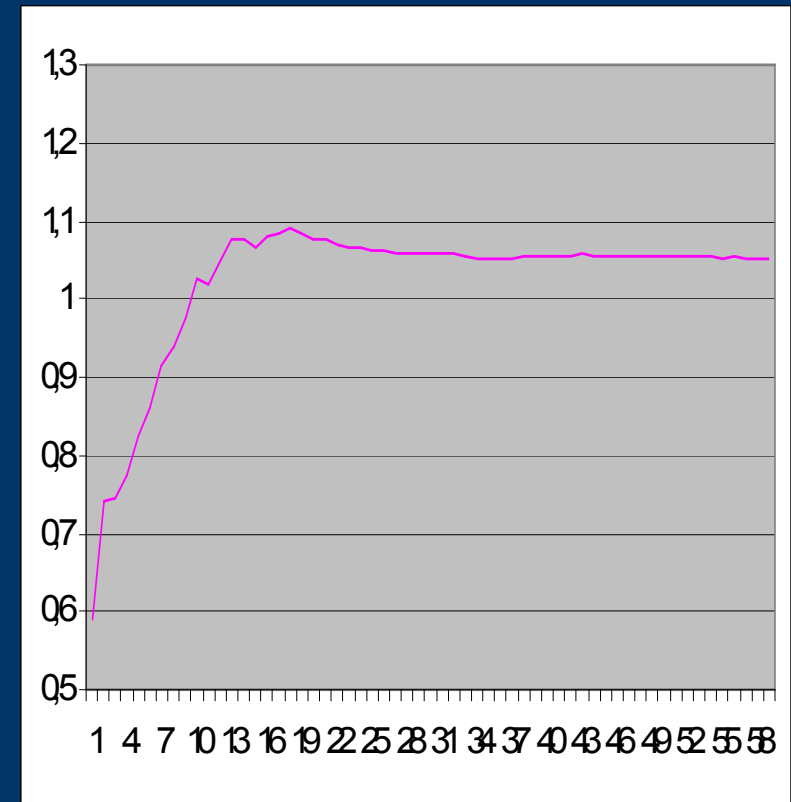
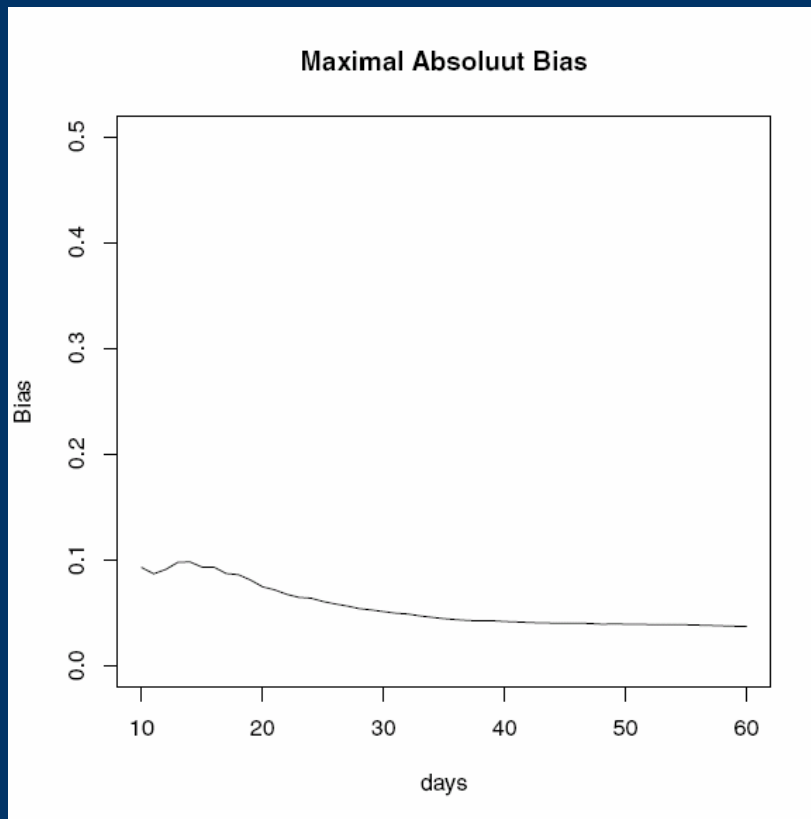
R-indicator en gemiddelde omzet: industrie



Vertekening KS via BTW 2007: detailhandel



Vertekening KS via BTW 2007: industrie



Conclusies

- Maximale vertekening groter dan vertekening via BTW
- Patronen R-indicator en vertekening KS komen overeen
- Betrouwbaarheidsintervallen klein: tussen 3,5% en 5% breed

- Grote verschillen tussen branches in representativiteit
- In algemeen: kleine bedrijven late of geen respons
- Mode: industrie nauwelijks verschil, detailhandel verschillen vrij groot
- November maand geeft lage score op R-indicator



Algemene discussie

- Benaderingsstrategieën in een mixed-mode setting: duur van veldwerk, aanschrijfbrieven, interviewers
- Gedifferentieerde benaderingsstrategieën en responsive designs: efficiënt toewijzen van capaciteit gegeven beschikbare achtergrondkenmerken
- Modellen voor paradata/procesdata: Dergelijke data zijn functie van bereikbaarheid/responsgeneigdheid. Hoe leiden we deze latente variabelen af?
- Correctie voor nonrespons die zo nauw mogelijk aansluit op veldwerk (dus paradata, mode)
- Combinatie met meetfouten: Leidt meer respons tot minder valide antwoorden?

