



Centraal Bureau voor de Statistiek

Divisie Keten Economische Statistieken

Sector Ontwikkeling en ondersteuning

Postbus 5.1.2.e

5.1.2.e

Rapportage gebruik VpB data

NU!

5.1.2.e

Versie 1.0

Kennisgeving:

De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.

Projectnummer:

KOO-

BPA-nummer:

-KOO

Datum:

16 januari 2012

RAPPORTAGE GEBRUIK VPB DATA

Versie	Datum	Beschrijving	Auteur
0.1	15-12-2011	Eerste versie	5.1.2.e 5.1.2.e
0.2	9-1-2012	Commentaar ALMA verwerkt	5.1.2.e
1.0	12-1-2012	Commentaar GMUK, PVRT, GBTN, ALMA verwerkt	5.1.2.e

Actieve distributie

Versie	Distributie
0.1	5.1.2.e 5.1.2.e
0.2	5.1.2.e 5.1.2.e
1.0	

Afkortingen

Versie	Distributie
PS	productiestatistieken
VpB	vennootschapsbelasting
beid	bedrijfsidentificatie

Samenvatting	3
1. Inleiding	4
1.1 Achtergrond	4
1.2 Beschikbare data	4
1.3 Doel	7
2. Haalbaarheidsonderzoek inzet VpB in NOPS	8
2.1 Vragen 1 en 2: Wanneer is welk percentage van de aangiften binnen naar grootteklassegroep en kerncel	8
2.2 Vraag 3: Wanneer is de verbruiksstructuur stabiel?	11
2.3 Vraag 4+5: Vergelijkbare tabellen voor alle grootteklassen zoals in het PS+ onderzoek	15
2.4 Vraag 6+7: Wat is de verdeling VPB en PS gegevens binnen kerncellen en grootteklassen	21
2.5 Vraag 8: Wat is de volgtijdelijke vergelijkbaarheid van de VpB gegevens. 22	
2.5.1 Volgtijdelijke vergelijkbaarheid op microniveau	22
2.5.2 Volgtijdelijke vergelijkbaarheid op geaggregeerd niveau	23
2.6 Vraag 9+10: Levensduur van beids in de VPB gegevens en in het ABR. 25	
3. Conclusie	25
4. Discussie	27
5. Referenties	28
Bijlage 1: berekening van grootteklasse-indeling op basis van de omzet	28

Samenvatting

Aan de hand van een aantal onderzoeksvragen is onderzocht of het mogelijk is om de PS te vervangen door de VpB om tot lastendrukreductie te komen. In lijn met eerdere conclusies (bijvoorbeeld Lammertsma et al (2008)), is de algemene conclusie dat met de huidige kwaliteit van de VpB bestanden het niet mogelijk is om de VpB te gebruiken voor grootschalige vervanging van de PS op korte termijn. Een aantal belangrijke oorzaken zijn:

- De gewenste datum van september $t+1$ voor het begin van de analyse van de VpB data is onmogelijk omdat er onvoldoende respons is, maart $t+2$ is reëler, dan is er ongeveer 70 % respons.*
- De verbruiksstructuur in de VpB heeft nog een flinke tijd (minimaal een jaar) nodig na maart $t+2$ voordat deze stabiel is. Als het niet mogelijk is om betere gevulde en gecontroleerde bestanden van de belastingdienst te krijgen moet door middel van een filter voor foutieve opgave en een imputatie methode de convergentie van de verbruiksstructuur verbeterd worden. Bij veel variabelen speelt de toevalscomponent een grote rol of deze gevuld is in de opgave van een bedrijf.*
- De methodiek van het PS+ onderzoek geeft de mogelijkheid om voor een beperkt aantal kerncellen maal grootteklasse de PS te vervangen door VpB. Maar blijkt ook zeer gevoelig te zijn voor de gebruikte versie van het VpB-bestand of grootteklasse-indeling (op basis van omzet of werkzame personen). In de PS+ methodiek wordt naar records gezocht die een vergelijkbare opgave hebben in de PS en in de VpB.*
- De volgtijdelijkheid op kerncel maal grootteklasse niveau is bepaald door de VpB ontwikkeling te leggen naast die van de PS. Slechts een zeer beperkt deel van de cellen heeft een ontwikkeling die in de beurt komt van de PS.*

Vervolgonderzoek naar de VpB heeft alleen zin als er veel tijd voor wordt uitgetrokken om de genoemde problemen op te lossen of de belastingdienst op korte termijn betere bestanden gaat leveren.

Trefwoorden: VpB, secundaire bronnen, PS, PS+

1. Inleiding

In het kader van het project NU! (Nieuwe Uitvraagstrategie) wordt gekeken naar manieren om tot een vereenvoudiging van de uitvraag door enerzijds een betere aansluiting op de bedrijfsadministraties en anderzijds vermindering van het aantal uitgevraagde variabelen. Hiervoor worden meerdere mogelijkheden onderzocht. Een van de mogelijkheden is het gebruik van data van de belastingdienst voor de vennootschaps- en inkomstenbelasting (VpB).

1.1 Achtergrond

Voor VpB is aan het begin van het HEcS-programma al veel uitgezocht, zie Lammertsma et al (2008) en de daarin bekeken onderzoeken. Conclusie van dat traject was dat VpB-data niet ingezet kunnen worden voor gebruik in de PS, daar weliswaar een deel van de VpB-omschrijvingen overeenkomsten vertonen met die van vergelijkbare variabelen in de PS, maar de inhoud van de data tussen de bedrijfseenheden (beids) onderling grote verschillen vertonen o.a. als gevolg van andere invulling en salderingen. Verder is geconcludeerd dat het VpB-bestand te weinig informatie bevat om de kostenstructuur te ramen, dat de kostenaandelen instabiel zijn en de tijdigheid een probleem is; slechts 35-40% van de aangiften is na 13 á 14 maanden binnen. In het huidige haalbaarheidsonderzoek worden de conclusies van dit onderzoek nogmaals tegen het licht gehouden. Bekeken wordt of de kennis van nu en de huidige context wellicht tot een andere conclusie leiden. Hierbij wordt ook beoordeeld welke van de VpB-variabelen mogelijk wél inzetbaar zijn voor de PS. Indien het onderzoek alsnog tot een positieve conclusie leidt, wordt een vervolgonderzoek opgestart waarin o.a. het eenhedenprobleem, d.w.z. hoe de VpB-stroom te koppelen aan de PS, wordt opgepakt (van Mourik 2011).

1.2 Beschikbare data

De volgende data zijn voor het onderzoek beschikbaar:

- DRT data 2008, 2009 en 2010 met een beid en omzet. De kwaliteit van het 2008 bestand is twijfelachtig, 2009 is beter maar ook nog niet goed genoeg om in productie gebruikt te worden. DRT 2010 wordt (voor een aantal branches) in de productie van statistieken gebruikt;
- Gaafgemaakte PS data van 2006 t/m 2009. 2006 t/m 2008 zijn onder SBI93 in Impect 1 verwerkt. 2008 en 2009 zijn geclassificeerd onder SBI2008 in NOPS. Voor 2010 zijn er al wel gaafgemaakte PS-resultaten, echter het aantal records is zeer beperkt;
- Niet gaafgemaakte PS data van 2009 en 2010 onder SBI2008. PS2010 zijn nog niet volledig;

- De ongecontroleerde VPB data¹ van 2006 t/m 2010 met datum dat een beid volledig gerespondeerd heeft onder SBI93 (alle jaren) en SBI2008 (2009 en 2010). Vanuit Baseline worden nog steeds mutaties in de bestanden doorgevoerd.

Zowel in de PS als de VPB zijn verschillende variabelen opgenomen. De onderzochte variabelen en de relaties tussen daartussen wordt gegeven in Tabel 1.

Omschrijving	Samenstelling PS	Samenstelling VPB ²
Netto omzet	OMZETPS210000 ³	Netto_omzet_c_w
Inkoopwaarde	INKWRDE100000	T_ktn_gr_uwe_c_w
Arbeidskosten	LOONSOM110002 ⁴ + LOONSOM121100+ LOONSOM121200+ LOONSOM122000	Loon_salaris_c_w + soc_ltn_c_w +pens_ltn_c_w - ont_uitk_subs_c_w
Afschrijving	AFSCHR110000	T_afs_c_w
Verkoopkosten	BEDRLST346000	Verk_ktn_c_w
Bedrijfsresultaat bruto	RESULTS120000 + AFSCHR100000	(Netto_omzet_c_w + ov_opbr_c_w + geact_prod_eb_c_w + ont_uitk_subs_c_w + w_voorr_ondh_werk_c_w) - (t_ktn_gr_uwe_c_w + loon_salaris_c_w+ soc_ltn_c_w + pens_ltn_c_w + ov_pers_ktn_c_w + auto_tr_ktn_c_w + huisv_ktn_c_w + ondh_ktn_ov_mva_c_w + verk_ktn_c_w+andere_ktn_c_w)
Bedrijfsresultaat netto	RESULTS120000	(Netto_omzet_c_w + ov_opbr_c_w + geact_prod_eb_c_w+ont_uitk_subs_c_w + w_voorr_ondh_werk_c_w) - (t_ktn_gr_uwe_c_w+loon_salaris_c_w+ soc_ltn_c_w+pens_ltn_c_w+ ov_pers_ktn_c_w+auto_tr_ktn_c_w+

¹ Deze data zijn nog niet door de belastingdienst gecontroleerd op juistheid. Het gecontroleerde bestand is te laat beschikbaar om een alternatief te kunnen zijn voor de PS.

² In de tabel worden de commerciële VPB-variabelen genoemd, voor de fiscale variabelen moet “_c” vervangen worden door “_f”. Vanaf 2010 zijn er alleen nog maar fiscale data beschikbaar.

³ OMZETPS210000 is de vervanger van VERKOOP210000 in Ouwehand (2011b)

⁴ LOONSOM110002 is de vervanger van LOONSOM110000 in Ouwehand (2011b)

		huisv_ktn_c_w+ondh_ktn_ov_mva_c_w + verk_ktn_c_w+andere_ktn_c_w+ t_afs_c_w)
Bedrijfslasten	BEDRLST310000	(t_ktn_gr_uwe_c_w+loon_salaris_c_w+ soc_ltn_c_w+pens_ltn_c_w+ ov_pers_ktn_c_w+auto_tr_ktn_c_w+ huisv_ktn_c_w+ondh_ktn_ov_mva_c_w + verk_ktn_c_w+andere_ktn_c_w+ t_afs_c_w)
Overige bedrijfs- opbrengsten	OPBRENG100000	Ov_opbr_c_w
Huisvestings- kosten	BEDRLST343000	Huisv_ktn_c_w
Overige personeelskoste n	PERSLST100000 - LOONSOM110002- LOONSOM121100- LOONSOM121200- LOONSOM122000	Ov_pers_ktn_c_w

Tabel 1: Vergelijking tussen PS en VPB variabelen

In de analyses worden bij voorkeur de commerciële data gebruikt, waarbij als de commerciële omzet ≤ 0 of leeg is, voor die betreffende beid de fiscale data gebruikt worden voor alle variabelen. In van de Burgt (2010) is uitgezocht dat commerciële waarden beter aansluiten bij de primaire waarneming dan fiscale waardes. Veel eenheden geven echter geen commerciële waarden op, daarom vullen we deze aan met fiscale waarden. Dit levert een structuur op die voor 70% van de kerncellen beter aansluit bij de primair waargenomen structuur dan de structuur op basis van alleen fiscale data. Kanttekening daarbij is dat vanaf 2010 alleen nog maar fiscale data beschikbaar zijn. Dit is een nadeel van de huidige opzet van het onderzoek waardoor een positieve bias ontstaat, maar door ook de commerciële waarden te gebruiken is het onderzoek beter vergelijkbaar met eerdere onderzoeken, zoals van der Stegen (2011).

Daarnaast worden de grootteklassen gegroepeerd in drie groepen, namelijk

1. grootteklasse 1 t/m 3
2. grootteklasse 4 en 5
3. grootteklasse 6 t/m 9

Bij de VPB is een complicatie dat waarnemingen niet volledig kunnen zijn. Daarom worden alleen de waarnemingen van beids gebruikt waarvoor geldt dat zowel het dekkingspercentage als het vullingpercentage van de beid 100 % is. Voor een

uitgebreidere beschrijving van deze algemene BASELINE richtlijnen zie Ouwehand (2011a) of Rademaker (2005).

1.3 Doel

Het doel van dit rapport is een haalbaarheidsonderzoek naar het gebruik van VPB-data ter (gedeeltelijke) vervanging van de PS-data voor de variabelen uit Tabel 1. Voor dit onderzoek zijn verschillende uitgangspunten mogelijk. Ten eerste kan de VpB data bekeken worden aan de hand van de methode voor vervanging van de PS, bijvoorbeeld de PS+ methodiek (van der Stegen 2011), een gewone micro-aanpak of een meso-aanpak waarbij de VpB wordt gecombineerd met een andere bron. Elke methode heeft andere eisen aan de data. Een tweede invalshoek kan de kwaliteit van de data zijn. Met behulp van zogenaamde rustpunten kunnen verschillende stadia worden onderscheiden met verschillende kwaliteitseisen. Voor deze invalshoeken zijn de volgende vragen van belang, welke onderzocht worden in dit rapport:

1. Wanneer is welk percentage van de VPB-aangiften binnen;
2. Wanneer is welk percentage van de VPB-aangiften binnen waarbij gelet wordt op grootteklassen en kerncellen;
3. Wanneer is de verbruiksstructuur afgeleid uit de VPB-aangiften stabiel en of dit op tijd is gegeven de doorlooptijd van het PS-proces;
4. Voor welke kerncellen maal grootteklassengroepen kan op basis van de PS+ methodiek⁵ de PS vervangen worden door de VpB, zie bijvoorbeeld van der Stegen (2011). Hierbij moet gekeken worden de invloed van beslisriteria, de invloed van de grootteklasse-indeling op werkzame personen of omzet en het gaafmaakproces van de PS;
5. Hetzelfde onderzoek als in punt 4 maar dan met de VpB-data die op september T+1 binnen is, in plaats van het meest definitieve bestand voor de grootteklasse-indeling in werkzame personen en gaafgemaakte PS-data;
6. Wat is de spreiding van de gegevens binnen kerncellen maal grootteklasse binnen de VPB;
7. Wat is de spreiding van de gegevens binnen kerncellen maal grootteklasse binnen de PS;
8. Wat is de volgtijdelijke vergelijkbaarheid van de VPB op zowel kerncel* grootteklasse niveau als microniveau;
9. Wat is de levensduur van de berichtgevers in de VPB;
10. Wat is de levensduur van de berichtgevers in het ABR.

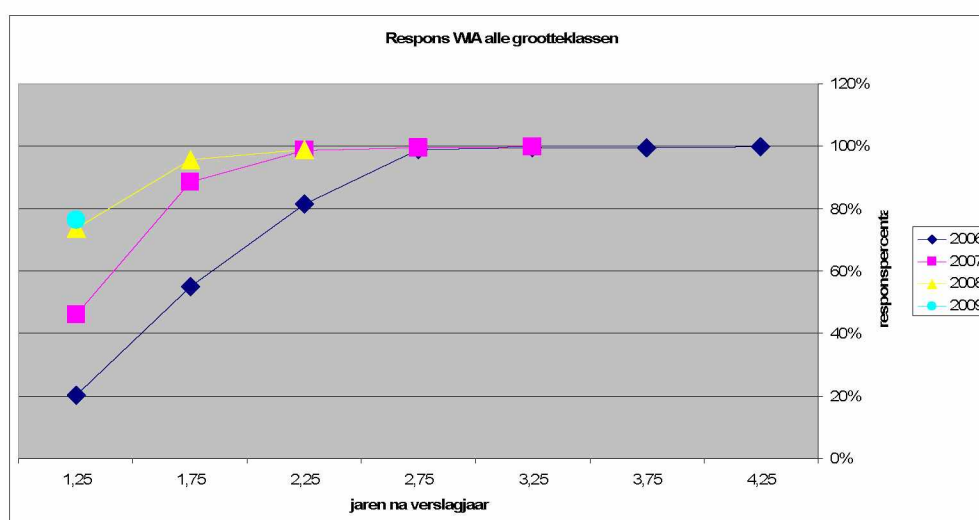
⁵ De methodiek van het PS+ onderzoek wordt gekenmerkt door per kerncel en grootteklasse beids te zoeken die binnen een bepaalde kleine marge dezelfde opgave doen voor de PS als voor de VpB.

Deze vragen worden in paragraaf 2 besproken. In een volgende versie van het rapport zullen de uitkomsten niet anders zijn maar wordt het rapport herschreven naar de verschillende invalshoeken. Gezien de planning was hiervoor in deze fase geen tijd.

2. Haalbaarheidsonderzoek inzet VpB in NOPS

2.1 Vragen 1 en 2: Wanneer is welk percentage van de aangiften binnen naar grootteklassegroep en kernel.

Het is moeilijk om het responspercentage te bepalen omdat in de bestanden nog steeds mutaties verwerkt worden en daarom nog niet definitief zijn. Als benadering wordt daarom de meest recente versie van het bestand als definitief bestand genomen in de analyse.



Figuur 1: Respons percentage VPB voor alle grootteklassen.

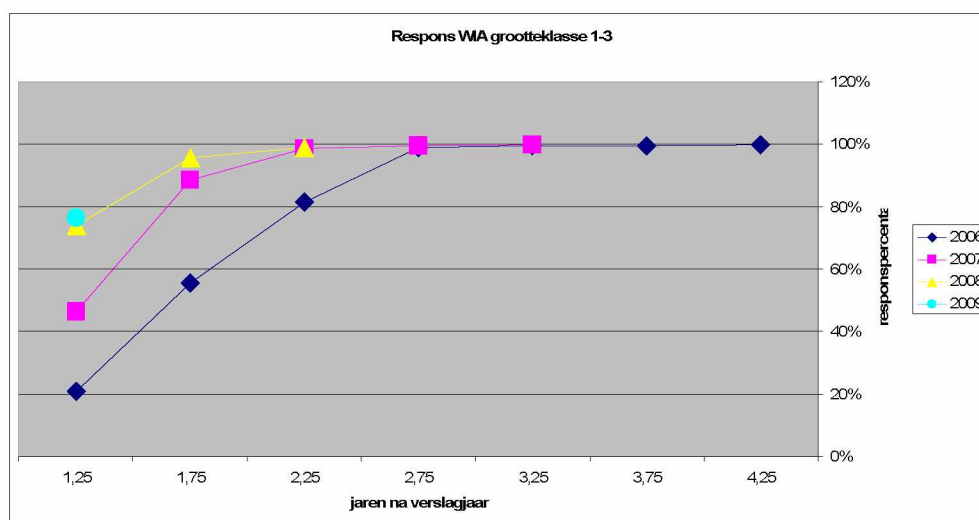
Figuur 1 geeft het responspercentage als functie van de tijd na afloop van het verslagjaar. In geen enkel bestand was er respons na 9 maanden⁶ na het verslagjaar. Het responspercentage is gebaseerd op het aantal beids dat gerespondeerd heeft. Van 2006 en 2007 kan eenvoudig worden aangenomen dat deze bestanden bij benadering volledig zijn omdat het laatste punt niet meer verandert. De mutaties zijn dusdanig minimaal dat ze grafisch niet zichtbaar meer zijn. 2008 lijkt dit punt ook te benaderen. 2009 is nog niet volledig. 2010 heeft al wel respons maar dat bestand is nog lang niet definitief waardoor ook niet te bepalen is wat het responspercentage is, omdat niet bekend is hoe groot de respons gaat worden.

⁶ In het huidige proces begint de analyse van de PS 9 maanden na het verslagjaar, maar in de praktijk wordt pas na 14 maanden met de analyse van secundaire bronnen begonnen bij KAD.

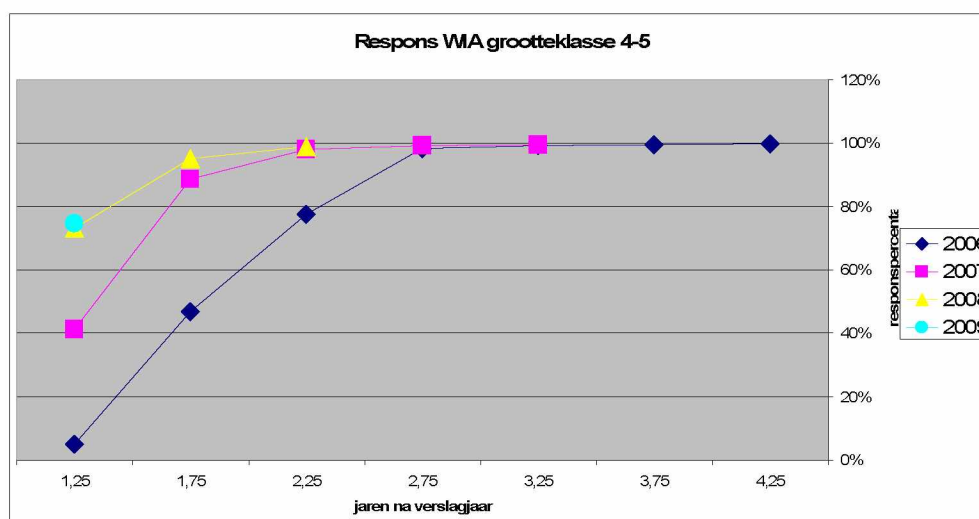
Een vergelijking tussen 2006, 2007, 2008 en 2009 levert dat er steeds sneller gerespondeerd wordt t/m 2008. Een vergelijking tussen 2008 en 2009 is moeilijk want 2009 is nog niet definitief en de verschillen zijn te klein.

Voor de jaren 2006-2009 was niet te achterhalen of er een versnelling van de interne processen geweest is waardoor de VpB sneller beschikbaar komt. Voor het 2010-bestand was dit wel mogelijk en bleek dat andere prioriteiten bij Baseline voor een vertraging van enkele weken hebben gezorgd bij de werking van de eerste respons.

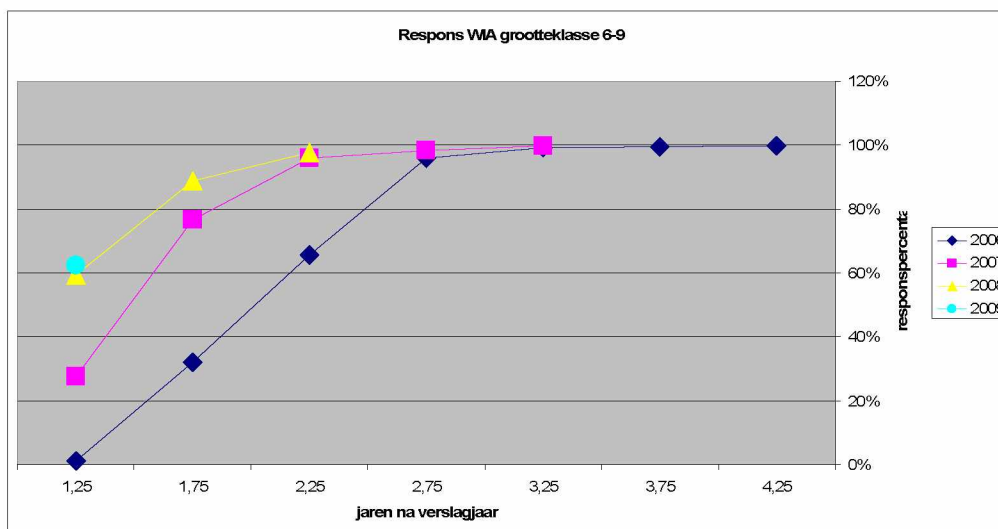
Figuur 2, Figuur 3 en Figuur 4 geven de responspercentages voor respectievelijk grootteklasse 1-3, 4-5 en 6-9. Het globale verloop is hetzelfde als voor alle grootteklassen, alleen naar mate de grootteklasse groter wordt, komt de respons later.



Figuur 2: Responspercentage VPB voor grootteklasse 1-3



Figuur 3: Responspercentage VPB voor grootteklasse 4-5



Figuur 4: Responspercentage VPB voor grootteklasse 6-9

In Tabel 2 staat het responspercentage naar SBI 2-digit onder SBI93 voor VPB van 2008. Dit jaar is gekozen om de SBI verdeling van het responspercentage te bekijken omdat dit het meest recente jaar is dat bijna volledig binnen is. Uit de verdeling blijkt dat deze vrij uniform is over de bedrijfstakken. Rond de 70 % na 1,25 jaar, ongeveer 95 % na 1,75 jaar en na 2,25 jaar 99%.

Deze resultaten kunnen vergeleken worden met die in Lammertsma (2008) worden gegeven. Op basis van deze vergelijking lijkt de respons van de VPB versneld te zijn.

SBI93-2d	1,25 jaar	1,75 jaar	2,25 jaar
	73%	96%	99%
01	74%	95%	99%
11	61%	80%	94%
14	78%	96%	99%
15	72%	95%	98%
16	82%	100%	100%
17	76%	96%	99%
18	80%	97%	99%
19	79%	96%	98%
20	76%	96%	99%
21	73%	94%	98%
22	74%	95%	99%
23	67%	96%	100%
24	69%	93%	99%
25	71%	95%	99%
26	74%	96%	99%
27	76%	92%	99%
28	76%	96%	99%
29	75%	96%	99%
30	71%	91%	99%
31	71%	96%	99%
32	70%	95%	99%
33	72%	95%	98%
34	73%	95%	99%
35	73%	96%	99%
36	75%	96%	99%
37	66%	91%	98%
40	70%	96%	99%
41	67%	100%	100%
45	76%	96%	99%
50	71%	95%	99%
51	70%	94%	99%
52	73%	95%	99%
55	68%	94%	98%
60	75%	95%	99%
61	74%	96%	99%
62	73%	95%	99%
63	71%	94%	98%
64	71%	94%	99%
70	70%	94%	99%
71	71%	95%	99%
72	76%	96%	99%
73	72%	95%	99%
74	75%	96%	99%
85	73%	96%	99%
90	73%	95%	99%
92	73%	96%	99%
93	81%	97%	99%

Tabel 2: Responspercentage naar SBI-2digit in 2008

2.2 Vraag 3: Wanneer is de verbruiksstructuur stabiel?

Er wordt gekeken naar de gemiddelde waarde per kerncel en grootteklasse van de volgende variabelen uit Tabel 1:

- Omzet
- Inkoopwaarde
- Arbeidskosten
- Afschrijvingen
- Verkoopkosten
- Bruto bedrijfsresultaten
- Netto bedrijfsresultaten
- Bedrijfslasten
- Overige opbrengsten
- Huisvestingskosten
- Overige personeelskosten

Het bestand van 2008 wordt hiervoor gebruikt omdat dat dit een redelijk recent en definitief bestand is en hiervoor ook DRT data beschikbaar zijn. Om geen last te hebben van ophoogfactoren wordt gekeken naar de gemiddelde waarde per kerncel per grootteklasse (grootteklasse in 3 groepen). Hierbij zijn 4 verschillende setups doorgerekend:

- Gewoon gemiddelde berekenen van de respons;
- Gewoon gemiddelde berekenen van de respons over de beids waarvoor DRT2008 respons heeft (ongeveer de helft) en met een filter hierop gebaseerd goedgekeurde beids, volgens de methode beschreven in paragraaf 2.3 op pagina 15 onder punt 1. Er is niet gekeken waarom er geen DRT2008 respons is. Dit kan selectiviteit veroorzaken in de populatie;
- De berekening van het gemiddelde verbeteren door het imputeren van ontbrekende waarden. Een combinatie van de respons 2008 en voor de ontbrekende beids, de respons halen uit 2007. Hierbij wordt aangenomen dat je weet welke beids gaan responderen zodat je weet welke beids geïmputeerd moeten worden. Andere complexere en mogelijk nauwkeurigere imputatie-methoden worden bijvoorbeeld beschreven in van der Stegen (2005);
- Combinatie van de 2 voorgaande bewerkingen, eerst filteren met DRT2008 data en daarna de nauwkeurigheid vergroten met de imputatie techniek uit het vorige punt;

In Tabel 3 worden de resultaten voor de omzet gegeven voor de 4 bovenstaande methoden uitgesplitst naar grootteklassengroepen. Daarbij gekeken wordt naar de convergentie van de gemiddelde omzet per kerncel, grootteklassegroep naar de definitieve waarde. Voor de convergentie wordt het gemiddelde absolute procentuele verschil berekend tussen de definitieve waarde en na 1,25, 1,75 en 2,25 jaar. Door het gebruik van DRT-data wordt slechts 57 % van de VPB data gebruikt,

met name omdat het DRT-bestand veel minder beids bevat dan het VPB bestand. Hierdoor ontbreken een paar kerncellen en de overige kerncellen convergeren gemiddeld naar een iets andere waarde⁷. Zowel het gebruik van DRT-data als een imputatietechniek verbeteren de convergentie sterk. Een combinatie van beide methoden levert het beste resultaat. Het verschil is het grootst 1,25 jaar na het verslagjaar. In dit geval heeft meer dan de helft van de kerncellen een fout $\leq 10\%$. Het aantal echt foute waarden is beperkt (groter dan 100%) tot 13. Vergelijkingen met VpB bestanden van 2008 van een latere versie datum laten een minder groot verschil zien. Mogelijk dat verbeteringen in de gehanteerde imputatie-methode dit kunnen verbeteren. Bij de gewone methode leveren een aantal kerncellen na 1,25 jaar een zeer slecht resultaat, tot meer dan een factor 600 fout. Dit betekent dat een goed filter voor de selectie van waarden onontbeerlijk is. Daarnaast convergeren grootteklasse 4 en 5 beter dan de kleine bedrijven en de grote bedrijven.

methode	grootteklasse	2,25 jaar	1,75 jaar	1,25 jaar
gewoon	alles	6%	32%	163%
	1-3	7%	51%	72%
	4,5	3%	10%	36%
	6-9	9%	35%	378%
DRT	alles	3%	11%	45%
	1-3	3%	5%	43%
	4,5	1%	7%	40%
	6-9	5%	21%	52%
imputeren	alles	3%	19%	26%
	1-3	1%	44%	52%
	4,5	1%	3%	7%
	6-9	6%	9%	19%
DRT+imputeren	alles	1%	3%	8%
	1-3	2%	3%	6%
	4,5	0%	2%	7%
	6-9	1%	4%	12%

Tabel 3: Gemiddelde fout ten opzichte van definitieve waarde voor de omzet

⁷ Hoe groot dit verschil is moet verder worden uitgezocht.

		2,25 jaar	1,75 jaar	1,25 jaar
inkoopwaarde	gewoon	9%	53%	102%
	DRT	24%	83%	191%
	imputeren	10%	90%	128%
	DRT+imputeren	8%	13%	33%
arbeidskosten	gewoon	4%	13%	28%
	DRT	7%	22%	66%
	imputeren	6%	15%	36%
	DRT+imputeren	3%	10%	26%
afschrijvingen	gewoon	6%	27%	199%
	DRT	14%	62%	404%
	imputeren	9%	26%	72%
	DRT+imputeren	5%	17%	40%
verkoopkosten	gewoon	10%	78%	201%
	DRT	12%	113%	275%
	imputeren	36%	127%	180%
	DRT+imputeren	6%	15%	43%
bruto bedrijfsresultaat	gewoon	16%	40%	77%
	DRT	20%	68%	476%
	imputeren	14%	159%	176%
	DRT+imputeren	8%	19%	80%
netto bedrijfsresultaat	gewoon	15%	49%	75%
	DRT	20%	81%	203%
	imputeren	20%	192%	230%
	DRT+imputeren	12%	33%	97%
bedrijfslasten	gewoon	26%	59%	115%
	DRT	11%	35%	159%
	imputeren	10%	62%	88%
	DRT+imputeren	8%	14%	33%
overige opbrengsten	gewoon	125%	313%	3254%
	DRT	340%	5801%	8790%
	imputeren	334%	383%	519%
	DRT+imputeren	296%	638%	606%
huisvestings- kosten	gewoon	7%	20%	578%
	DRT	6%	26%	83%
	imputeren	19%	36%	3319%
	DRT+imputeren	2%	10%	37%
overige personeels- kosten	gewoon	6%	111%	63%
	DRT	9%	83%	211%
	imputeren	13%	127%	174%
	DRT+imputeren	4%	116%	161%

Tabel 4: Gemiddelde fout ten opzichte van definitieve waarde voor overige variabelen.

In Tabel 4 staan de resultaten voor de overige variabelen. Deze zijn slechter dan die voor de omzet in Tabel 3. Er is ook geen methode die in alle gevallen de beste resultaten levert. Hiervoor zijn een aantal verklaringen:

- Het DRT filter kijkt alleen naar de omzet, dus filtert ook alleen beids met foutieve omzetten. Het heeft geen invloed op andere variabelen;
- In sommige gevallen is de waarde van een variabele in het vorige jaar geen goede voorspeller, bijvoorbeeld als deze eenmaal negatief is. De imputatiemethode kan hier nog niet mee omgaan;
- Sommige variabelen zijn het ene jaar wel en het andere jaar niet ingevuld bij een beid. De imputatiemethode kan hier nog niet mee omgaan. Het percentage beids dat wel een waarde heeft voor 2007 en niet voor 2008 of andersom wordt voor de verschillende variabelen berekend in Tabel 14. In deze tabel blijkt dit percentage voor alle variabelen behalve de omzet groot te zijn (groter dan 15 %).

Verbeteringen in de imputatiemethode en in de filtering van foutieve waarden kunnen mogelijk tot verbeterde resultaten leiden. Op basis van dit onderzoek is nog niet te zeggen in welke mate dit kan, met name het feit dat veel variabelen het ene jaar wel een waarde hebben en het andere jaar niet is een complicerende factor. Maar met de huidige resultaten is het niet mogelijk om van een stabiele structuur te spreken op het moment dat met de analyse begonnen moet worden, namelijk na ongeveer een 1,25 jaar. Alleen de omzet is na 1,75 jaar redelijk uitgeconvergeerd (mutatie kleiner dan 4 %), de overige variabelen halen dit percentage in een enkel geval na 2,25 jaar. Andere overige variabelen convergeren nog slechter.

Gezien het vrij hoge respons percentage na 1,25 jaar uit voorgaande paragraaf zijn de slechte resultaten voor de convergentie opmerkelijk. Een mogelijke verklaring is dat de laatste respons zeer afwijkende waarden heeft die een sterke invloed hebben. Nader onderzoek is hier gewenst.

De beste oplossing om de convergentie te verbeteren is nog dat de belastingdienst completere records levert, maar het is de vraag of dit mogelijk is.

2.3 Vraag 4+5: Vergelijkbare tabellen voor alle grootteklassen zoals in het PS+ onderzoek

In het PS+ onderzoek wordt gekeken naar de PS en de VPB data van 2009 op basis van SBI2008. Het onderzoek is uitgebreid beschreven in van der Stegen (2011). De belangrijkste stappen in dit onderzoek zijn:

1. Verwijderen van foutieve VPB data aan de hand van een vergelijking op basis van de omzet met DRT data. Omdat je de structuur uit de VPB haalt en dus alles uitdrukt als fractie van de omzet wil je dat de VPB de goede eenheid beschrijft. Als maat daarvoor gebruiken we de omzet. De omzet in VPB moet aansluiten bij de echte omzet. Als benadering voor de echte omzet gebruiken we de BTW-omzet. Door DRT is uitgezocht voor welke beids je omzet kunt schatten met BTW. Er is voor gekozen om alleen te filteren op de omzet omdat daarvoor gegevens met een jaarlijkse frequentie

beschikbaar zijn⁸. Daarom wordt door het filter niet voor de andere variabelen dan de omzet gefilterd. Om een foutieve waarde te detecteren worden per beid twee bronnen voor dezelfde variabele met elkaar vergeleken voor alle beids die in beide bronnen respons hebben. Analoog aan Ouwehand (2011a) wordt dit gedaan aan de hand van het delta criterium:

$$\text{delta} = 2 * \text{bron1} * \text{bron2} / (\text{bron1}^2 + \text{bron2}^2)$$

Een uitbijter is gedefinieerd als een beid waarvoor geldt dat $\text{delta} \leq 0,8$ (deze waarde wordt in DRT gebruikt, zie UC0407 van NOPS) is. Voor $\text{delta} \leq 0,8$ is $\text{bron1} \geq 2 * \text{bron2}$ of $\text{bron1} \leq 0,5 * \text{bron2}$. Deze grenzen zijn bewust ruim gekozen omdat alleen de records gefilterd moeten worden waarvoor een andere manier/eenheid van opgeven van de omzet bij de belastingdienst wordt gehanteerd. Dit filter is niet waterdicht. Als een waarde voor zowel VPB als DRT op dezelfde manier fout is, laat het filter hem door. De kans hierop is erg klein.

2. Vervolgens wordt op de records die goedgekeurd zijn door het foutieve waarde-filter een analyse uitgevoerd aan de hand van een aantal criteria die overgenomen zijn uit eerder onderzoek. Ook is een beoordeling van de score op de criteria gebruikt die aansluit bij de hiervoor vroeger gehanteerde normeringen. De analyses worden per variabele uitgevoerd. Per kerncel per grootteklassegroep wordt een lineaire regressie $Y = aX + b$ uitgevoerd door de puntenwolk van beids met voor zowel $x = \text{bron1}$ als $y = \text{bron2}$ een waarneming. Met de kleinste kwadraten formule worden de a en de b in deze formule bepaald. Er is gekozen om een constante b in de formule te plaatsen, maar alleen te toetsen op de waarde van a , zoals in Ouwehand (2011a) en de Jong (2004). Systematische afwijkingen tussen de bronnen komen in een waarde van a afwijkend van 1 tot uiting, waarbij een waarde van b ongelijk aan 0 als meetruis niet in de selectiecriteria wordt meegenomen. Er zijn twee beslisbomen aan de hand van vier selectiecriteria waarvan er twee gebaseerd op een regressie analyse. De criteria worden per grootteklassegroep per kerncel berekend:

- a. de regressiecoëfficiënt a in de lineaire regressie⁹
- b. de R^2 van de lineaire regressie
- c. verhouding $\text{verh} = \text{ABS}((\Sigma \text{bron1} - \Sigma \text{bron2}) / \Sigma \text{bron2})$ waar $\Sigma \text{bron1} = \Sigma \text{PS}$ en $\Sigma \text{bron2} = \Sigma \text{VPB}$ de som over de beids zijn voor de

⁸ Een filter op de andere variabelen uit tabel 1 is gewenst maar alleen voor de omzet is een referentiewaarde voor bijna alle beids aanwezig. Voor andere variabelen zou een filter bijvoorbeeld gebaseerd kunnen worden op een plausibele ontwikkeling in de tijd ten opzichte van vorig jaar.

⁹ Er wordt niet gekeken naar regressieparameter b vanwege tijdsdruk en omdat a de belangrijkste is. Daarnaast wordt als b een grote waarde heeft niet voldaan aan criterium 3.

verschillende bronnen voor dezelfde variabele met respons van een beid in zowel bron1(=PS) en als bron2(=VPB).

- d. het percentage foutieve waarden (dit percentage wordt bepaald op het moment dat de foutieve waarden worden verwijderd, dus tijdens het foutieve waarden filter).

Voor elk van deze criteria zijn classificatiegrenzen bepaald in een eerder PS+ onderzoek zie bijvoorbeeld Jong (2004):

Criterium	Goed(g)	Medium(m)	Slecht(s)
A	$0,9 < a < 1,1$	$0,8 < a \leq 0,9$ of $1,1 \leq a < 1,2$	$a \leq 0,8$ of $a \geq 1,2$
B	$1 \geq R2 > 0,8$	$0,8 \geq R2 > 0,7$	$0,7 \geq R2$
C	$\text{verh} < 0,1$	$0,1 \leq \text{verh} < 0,2$	$\text{verh} \geq 0,2$
D	$\% \text{uitb} < 10\%$	$10\% \leq \% \text{uitb} < 15\%$	$\% \text{uitb} \geq 15\%$

Tabel 5: Grenzen van de criteria

De criteria geven aan dat er niet te veel afgeweken mag worden van de relatie bron1 = bron2 per beid. De afwijking wordt gemeten aan de hand van de selectiecriteria. Met deze selectiecriteria worden beslisbomen samengesteld om een totaal oordeel per kerncel per grootteklasse te krijgen. Er zijn 2 beslisbomen, namelijk beslisboom 1 en beslisboom 2, zie Tabel 6 en Tabel 7. In eerste instantie wordt beslisboom 1 gebruikt voor alle variabelen uit Tabel 1. Als de omzet geen “g” scoort in beslisboom 1, dan wordt de beslisboom 2 voor deze variabele gebruikt om alsnog eventueel een “g” te kunnen geven aan die kerncel maal grootteklassegroep.

Criterium 1	Criterium 2	Criterium 3	Totaal
g	g	g	g
m	g	g	g
g	m	g	g
g	g	m	g
Overige combinaties			s

Tabel 6: Beslisboom 1

Criterium 1	Criterium 2	Criterium 3	Criterium 4	Totaal
m	m	g	g	g

m	g	m	g	g
g	m	m	g	g
Overige combinaties				s

Tabel 7: Beslisboom 2

Een groot nadeel van deze procedure is dat het percentage foutieve waarden geen grote rol speelt in de selectie criteria. Dit betekent als er in een kerncel 2 goede beids zijn waarvoor beslisboom 1 geldt en er tig worden afgekeurd op basis van het BTW-filter, deze cel toch een “g” kan krijgen. Als verbetering op de PS+ methodiek wordt daarom beslisboom 3 voorgesteld die beslisbomen 1 en 2 vervangt. Deze beslisboom is strenger dan de methodiek in het PS+onderzoek het mogelijk is dat er minder kerncellen worden aangewezen voor vervangen van PS-waarneming door VpB-data.

Criterium 1	Criterium 2	Criterium 3	Criterium 4	Totaal
g	g	g	g	g
m	g	g	g	g
g	m	g	g	g
g	g	m	g	g
g	g	g	m	g
Overige combinaties				s

Tabel 8: Beslisboom 3

- Als alle analyses uitgevoerd zijn worden de resultaten gewogen om tot een resultaat per kerncel maal grootteklassengroep te komen. In Tabel 1 wordt de koppeling gemaakt tussen de PS en de VPB voor een aantal variabelen. Maar niet alle vergelijkingen hebben hetzelfde gewicht, zie Tabel 9. Omdat VPB dezelfde eenheid moet beschrijven als de PS en we de hele structuur uitdrukken in de omzet heeft deze een hoog gewicht gekregen. Een andere reden is dat veel variabelen uit de PS-vragenlijst in de VPB het beste linken aan de omzet. Het aantal variabelen vanuit de PS-vragenlijst dat linkt aan een variabele in VPB, zoals in NOPS vastgelegd in de tabel tbl_PS_koppeling_EDW was ook de leidraad voor de verdeling van de overige gewichten. Het bedrijfsresultaat zou dan een gewicht 0,25 moeten krijgen. Deze variabele heeft een gewicht 0, omdat deze post de resultante is van omzet, overige bedrijfsopbrengsten en bedrijfslasten.

Vergelijking	gewicht
--------------	---------

PS-VPB-omzet	5
PS-VPB-inkoopwaarde	1
PS-VPB-arbeidskosten	0,5
PS-VPB-afschrijving	0,25
PS-VPB-verkoopkosten	0
PS-VPB-netto-bedrijfsresultaat	0
PS-VPB-bruto-bedrijfsresultaat	0
PS_VPB-overige opbrengsten	0,5
PS-VPB-bedrijfslasten	1
PS-VPB-bedrijfslasten-huisvesting	0,25
PS-VPB-bedrijfslasten-overige personeelskosten	0

Tabel 9: Gewichten voor vergelijking

Er zijn 7 sets met berekeningen uitgevoerd, zie Tabel 10. In alle gevallen is de delta voor het foutieve waarden filter in gesteld op 0,8, de waarde die ook in DRT wordt gebruikt. Als beslisboom 1 en 2 zijn gebruikt, dan is de methodiek identiek aan het PS+onderzoek. Beslisboom 3 is de verbeterde versie hiervan. Daarnaast zijn 2 afslagen van de VPB respons gebruikt, namelijk na 15 maanden op 31 maart en de maximaal beschikbare dataset ten tijde van het onderzoek, na 23 maanden. Omdat er geen VPB data beschikbaar zijn op september t+1 (hetgeen gevraagd werd in de opdracht), is deze verplaatst naar maart t+2 als de werkelijke analyse van de data begint. Ook de invloed van het gaafmaken van de PS is bekeken. Ten slotte wordt de invloed onderzocht van een grootteklasse-indeling op basis van de omzet in plaats van werkzame personen.

methode	delta	beslisboom	responsdatum	PS-data	grootteklasse
a	0,8	1 en 2	30-11-2011	gaaf	Werkz. personen
b	0,8	1 en 2	31-3-2011	gaaf	Werkz. personen
c	0,8	3	30-11-2011	gaaf	Werkz. personen
d	0,8	3	31-3-2011	gaaf	Werkz. personen
e	0,8	1 en 2	30-11-2011	niet-gaaf	Werkz. personen
f	0,8	1 en 2	30-11-2011	gaaf	omzet
g	0,8	1 en 2	30-11-2011	niet-gaaf	omzet

Tabel 10: Berekeningen volgens PS+ methodiek

De resultaten zijn bijgevoegd in de excel sheet.

	grootteklasse-groep		
methode	1	2	3
a	85	104	75
b	82	106	53
c	38	30	9
d	33	42	16
e	20	43	68
f	87	40	83
g	13	29	69

Tabel 11: Aantal kerncellen per grootteklasse-groep dat geselecteerd is voor vervanging PS door VpB

Op het blad samenvatting in de Excel sheet staat aangegeven welke kerncellen maal grootteklasse-groep volgens welke methode een score van 6,25 of hoger heeft, een samenvatting hiervan wordt in Tabel 11 gegeven. Opvallend is

- Een goede score voor responsdatum 31 maart T+2 betekent niet dat ook 30 november T+2 goed scoort, dit is wel vaak het geval. De aantallen kerncellen komen redelijk overeen, maar een flink aantal kerncellen heeft op 31 maart T+2 wel een goede score en op 30 november T+2 niet en andersom. Voor methode a en b is in Tabel 12 weergegeven. De aantallen kerncellen die slechts door methode a of door methode b worden geselecteerd zijn niet te verwaarlozen.

	grootteklasse-groep		
methode	1	2	3
a	22	22	38
b	19	24	16
ab	63	82	37

Tabel 12: Invloed van versie VpB bestand op selectie van kerncellen

- Daarnaast is het duidelijk dat beslisboom 3 strenger is dan beslisboom 1 en 2, waardoor veel minder kerncellen maal grootteklasse-groepen worden geselecteerd. Als beslisboom 3 goed scoort, dan kun je ervan uitgaan dat volgens beslisboom 1 en 2 de score ook goed is.
- Een goede score voor de grootteklasse-indeling op basis van werkzame personen betekent niet dat ook een indeling op basis van de omzet ook goed scoort, dit is wel vaak het geval. De aantallen kerncellen komen redelijk overeen, maar een flink aantal kerncellen heeft op basis van werkzame personen wel een goede score en op basis van omzet niet en andersom. Met name voor grootteklasse-groep 2 zijn de verschillen groot, de indeling op basis van werkzame personen scoort daarin veel beter.

Samenvattend is er groep kerncellen maal grootteklasse die in alle gevallen goed scoort, zowel naar datum van het VpB-bestand, als grootteklasse-indeling, als naar beslisboom. Echter een flink deel van de kerncellen maal grootteklasse dat op een bepaalde manier geselecteerd wordt, wordt voor een andere beslisboom, grootteklasseindeling of datum van het VpB bestand niet geselecteerd.

2.4 Vraag 6+7: Wat is de verdeling VPB en PS gegevens binnen kerncellen en grootteklassen.

De oorspronkelijke vraag is om de spreiding op microniveau in de PS kerncellen maal grootteklasse te bekijken en dit ook voor de VPB data te doen. In plaats daarvan is gekeken naar de spreiding van de PS microdata ten opzichte de VPB microdata. De reden hiervoor is dat de spreiding van de PS of de VPB niet zo veel zegt, alleen de verhouding is interessant. Hiervoor wordt per kerncel per grootteklasse in 3 groepen de relatieve standaardfout=(standaardfout/ gemiddelde waarde) berekend voor 5 datasets:

1. De gaafgemaakte PS data van 2009;
2. De VPB data van 2009, waarbij alleen Beids gebruikt zijn die in de PS voorkomen;
3. De VPB data van 2009 gefilterd met DRT2009, waarbij alleen Beids gebruikt zijn die in de PS voorkomen;
4. Alle beschikbare VPB data van 2009;
5. Alle beschikbare VPB data van 2009 gefilterd met DRT;

De 4 VPB datasets worden vergeleken met de PS dataset in Tabel 13. In vergelijking 1 wordt de relatieve standaardfout van de tweede dataset gedeeld door de eerste, in vergelijking 2, de derde dataset door de eerste enz. Per vergelijking worden het percentage kerncellen maal grootteklassegroepen berekend dat valt in de volgende categorieën:

- 0 betekent ongeveer gelijke relatieve standaardfout (+/- 5 %) voor VPB en PS;
- >0 dat is de relatieve standaardfout van de VPB data groter dan van de PS data en
- <0 andersom.

Algemene conclusies zijn dat filteren met DRT data de spreiding verkleint en ook de selectie op basis van microdata ook de spreiding verkleint. De spreiding op de VpB gegevens die door PS+ worden geselecteerd is kleiner dan de PS spreiding. Aan de andere kant is de spreiding van de VpB over alle records juist weer groter.

	vergelijking 1 0 >0 <0				vergelijking 2 0 >0 <0				vergelijking 3 0 >0 <0				vergelijking 4 0 >0 <0		
omzet	21%	30%	49%		16%	20%	64%		9%	58%	33%		10%	48%	41%
inkoop- waarde	18%	32%	50%		15%	21%	63%		5%	74%	21%		5%	67%	28%
arbeids- kosten	23%	30%	47%		19%	23%	58%		7%	39%	54%		7%	32%	60%
afschrij- vingen	12%	20%	68%		12%	15%	74%		5%	67%	28%		7%	60%	33%
verkoop- kosten	14%	36%	50%		12%	30%	59%		3%	79%	18%		4%	74%	22%
netto bedrijfs- resultaat	10%	54%	36%		6%	27%	67%		3%	85%	12%		3%	80%	16%
bruto bedrijfs- resultaat	8%	28%	64%		8%	20%	72%		6%	75%	19%		7%	67%	26%
bedrijfs- lasten	10%	72%	18%		16%	21%	64%		8%	58%	34%		9%	50%	42%
overige opbreng- sten	9%	28%	62%		10%	20%	70%		3%	92%	5%		2%	91%	7%
huisvestin- gskosten	14%	28%	58%		10%	23%	67%		8%	58%	34%		9%	53%	39%
overige personeel kosten	11%	41%	48%		10%	34%	56%		4%	76%	20%		3%	74%	23%

Tabel 13: Spreiding van de microdata binnen kerncel en grootteklasse

2.5 Vraag 8: Wat is de volgtijdelijke vergelijkbaarheid van de VpB gegevens.

De volgtijdelijke vergelijkbaarheid van de VpB wordt op zowel micro- als op geaggregeerd niveau bekeken.

2.5.1 Volgtijdelijke vergelijkbaarheid op microniveau

Om de volgtijdelijke vergelijkbaarheid op microniveau te bekijken zijn de VpB-bestanden van 2006, 2007 en 2008 op beide niveau gekoppeld. Vervolgens zijn de jaar-op-jaar mutaties van de verschillende variabelen uit Tabel 1 per beide berekend. Dan worden er 3 indicatoren onderscheiden¹⁰;

1. Ten eerste wordt het % beids berekend dat geen informatie heeft in beide jaren voor die variabele maar wel voor andere variabelen;
2. Ten tweede wordt het percentage van de beids gegeven waarvoor geldt $1/1,2 < \text{waarde_jaar_1} / \text{waarde_jaar_2} < 1,2$. Dit zijn de mutaties waar niets bijzonders aan de hand is;
3. Als laatste wordt het percentage van de beids gegeven waarvoor geldt $1/2 < \text{waarde_jaar_1} / \text{waarde_jaar_2} < 2$. Als de mutatie groter is, dan is de kans erg groot dat gaat om een foutieve aangifte.

De groep bedrijven die wel voldoen aan het criterium onder 3 en niet aan het criterium 2 zijn de zogenaamde bekijk-bedrijven.

¹⁰ De grenzen zijn in overleg met 5.1.2.e vastgesteld.

Deze 3 indicatoren worden uitgerekend voor 3 vergelijkingen, namelijk 2006 met 2007, 2006 met 2008 en 2007 met 2008. De resultaten staan in Tabel 14. De resultaten van de verschillende vergelijkingen verschillen niet veel. Hierdoor is kan dit resultaat eenvoudig worden gegeneraliseerd worden naar andere jaren.

	% beids bij 2006 tov 2007				% beids bij 2006 tov 2008				% beids bij 2007 tov 2008		
	missing	1,2	2		missing	1,2	2		missing	1,2	2
Omzet	4,8	55,5	83,4		3,7	46,0	79,5		4	55,5	83,9
Inkoopwaarde	30,3	30,1	52,5		28,7	24,3	49,1		30	29,3	52,0
Arbeidskosten	54,7	28,2	38,9		55,5	21,9	35,3		55,7	27,6	38,2
Afschrijvingen	15,7	34,7	61,8		14,4	25,0	53,7		16,8	40,6	64,4
Verkoopkosten	38,3	18,0	39,3		36,8	15,1	36,2		37,9	17,7	39,5
Bruto bedrijfsresultaat	19,1	28,2	54,1		21,5	22,9	48,5		19,6	28,2	54,0
Netto bedrijfsresultaat	19,1	24,9	50,4		21,6	20,1	44,7		19,7	25,6	50,9
Bedrijfslasten	15	49,2	75,5		17,2	39,1	69,0		15,4	48,4	74,8
Overige opbrengsten	77,8	6,5	10,9		77,9	4,7	8,9		78,9	6,2	10,4
Huisvestingskosten	39,6	33,1	48,8		38,4	28,8	45,9		40,9	32,2	47,5
Overige personeelslasten	65,9	10,9	21,1		65,7	8,2	18,1		66,3	-8,9	21,0

Tabel 14: Het % beids bij vergelijking van 2 jaarbestanden dat ontbreekt, en een mutatiefactor van kleiner dan 1,2 of 2 heeft per variabele.

Opvallend in Tabel 14 is dat sommige variabelen bij de meeste beids te berekenen zijn en maar een paar % ontbrekend hebben, zoals de omzet, maar andere weer niet zoals de overige opbrengsten. Van bijvoorbeeld de arbeidskosten kun je verwachten dat deze bij meeste bedrijven gemaakt worden, echter bij slechts minder dan 50 % van de bedrijven zijn deze in 2 jaar aanwezig.

2.5.2 Volgtijdelijke vergelijkbaarheid op geaggregeerd niveau

Om de volgtijdelijke vergelijkbaarheid te bepalen op geaggregeerd niveau worden de resultaten vergeleken met de ontwikkeling van de PS in de jaren 2006, 2007 en 2008. Nadeel van deze keuze is dat er geen DRT-foutieve waarden filter beschikbaar is, deze data is gecodeerd volgens SBI93 en bewerkt in Impect 1 en niet in NOPS. De reden om toch voor deze data te kiezen is de lengte van de tijdreeks, namelijk 3 jaar in plaats van 2 jaar voor NOPS-data.

Het PS-bestand bestaat uit beids, met kerncel, grootteklasse en een ophoogfactor voor de verschillende variabelen uit Tabel 1. Met deze ophoogfactoren kunnen totalen per kerncel, grootteklasse en verder geaggregeerd berekend worden. Van de VpB bestanden is niet bekend of ze volledig zijn (zeer waarschijnlijk niet) omdat er geen gezien de tijd geen koppeling gemaakt is met het volledige ABR-bestand. Hierdoor is het niet mogelijk om door sommatie een totaal te berekenen bijvoorbeeld per kerncel en grootteklasse. In plaats daarvan wordt de gemiddelde waarde per kerncel en grootteklasse¹¹ berekend. Door de respons van de beids in het PS-bestand te vervangen door de gemiddelde waarde volgens de VpB per kerncel en grootteklasse kan met de PS-ophoogfactoren toch een totaal per kerncel per grootteklasse berekend worden voor de VpB. De gemiddelde waarde per kerncel en grootteklassegroep (dus geaggregeerd over de grootteklassen) en variabele wordt op 3 manieren berekend worden:

¹¹ Dit is gedetailleerder dan het niveau waarop de resultaten worden gepresenteerd om vertekening te voorkomen van een foutieve verhouding over de grootteklassen in de VpB.

1. als gewoon gemiddelde over alle respons per kerncel en grootteklassegroep, in de berekeningen hieronder hebben deze het label “alles”;
2. als gemiddelde van de panel-respons van de bedrijven die in twee jaren in het bestand zijn opgenomen per kerncel en grootteklassegroep in de berekeningen hieronder hebben deze het label “panel”;
3. als gemiddelde van de panel-respons van de bedrijven die in twee jaren in het bestand zijn opgenomen per kerncel en grootteklassegroep, waarbij de mutatie van de omzet tussen die 2 jaren binnen een factor 2 groter of kleiner is¹². In de berekeningen hieronder hebben deze het label “selectie”.

Omdat er niveau verschillen tussen de opgehoogde PS en de WIA respons zijn, wordt er gekeken naar de mutatie tussen de jaren 2006 ten opzichte van 2007, 2006 ten opzichte van 2008 en 2007 ten opzichte van 2008. De uitgebreide resultaten per kerncel en grootteklassegroep zijn opgenomen in de ingevoegde file.



Alle kerncellen maal grootteklassegroepen zijn op 9 manieren bekeken, namelijk 3 verschillende combinaties van jaren en 3 verschillende selecties van VpB data. Per variabele uit Tabel 1 wordt de absolute afwijking berekend tussen de Impect 1 ontwikkeling en de VpB ontwikkeling in %. Dit percentage wordt vermenigvuldigd met de gewichten uit Tabel 9 om tot een score te komen voor die cel. De score van een cel wordt op de volgende manieren ingedeeld:

- <50: score groen, dit komt neer op een gemiddelde fout per variabele van iets minder dan 6%.
- >50 en <100: score oranje, dit komt neer op een gemiddelde fout per variabele van iets minder dan 12%.
- >100: score rood, de grotere gemiddelde fouten.

De gebruikte grenzen zijn zodanig gezet dat onderscheid gemaakt kon worden in cellen die eventueel bruikbaar kunnen zijn en cellen die dat zeker niet zijn.

¹² Deze analyse is alleen op de omzet uitgevoerd. In een vervolg onderzoek kan dit worden uitgebreid naar een selectie van beids die verschilt per variabele.

	geen oplossing	groen	oranje	rood
06_07 alles	37	43	92	503
06_07 panel	42	46	122	465
06_07 selectie	45	44	114	472
06_08 alles	57	27	73	518
06_08 panel	69	24	85	497
06_08 selectie	76	28	87	484
07_08 alles	59	39	98	493
07_08 panel	67	39	136	447
07_08 selectie	68	44	132	445

Tabel 15: Samenvatting van de scores van de kerncellen maal grootteklasse

In Tabel 15 staat een samenvatting van de resultaten. Slechts een klein deel van de cellen scoort “groen”. Dit lijkt iets hoger te zijn naar mate het VpB bestand meer bewerking heeft ondergaan, het selectie bestand op basis van de omzet levert de beste score.

2.6 Vraag 9+10: Levensduur van beids in de VPB gegevens en in het ABR.

Om de levensduur van de beids in de VPB te bepalen worden de bestanden van 2006 en 2009 met elkaar vergeleken. In Tabel 16 staat het aantal beids per grootteklasse in 2006 in de kolommen tegen de rijen het aantal beids per grootteklasse in 2009. Missing in 2009 betekent dat een beids verdwenen is sinds 2006 en missing in 2006 betekent dat een beid erbij gekomen is sinds 2006. Bijna 70 % van de beids uit 2006, bestaat nog steeds in 2009. Dit verschilt niet sterk van de situatie in het ABR.

	missing in 2006	0	1	2	3	4	5	6	7	8	9
missing in 2009											
0	0	33783	106827	62720	11881	6073	3530	895	387	191	90
1	46112	12063	3207	937	174	71	17	0	0	0	0
2	198326	9631	231762	33689	1260	246	31	1	0	0	0
3	71861	2437	14055	103546	7125	595	51	2	1	0	0
4	12363	623	1338	12986	17689	2640	131	7	3	2	0
5	7299	262	331	1691	5506	8623	1101	18	0	0	0
6	4141	100	66	270	477	2618	4543	203	9	1	0
7	1229	12	9	21	24	97	815	921	71	4	1
8	580	0	2	1	3	13	54	247	342	23	1
9	282	2	2	1	2	1	6	20	73	159	11
9	112	0	0	0	0	0	1	1	6	17	59

Tabel 16: Levensduur van beids uit 2006 en 2009

3. Conclusie

Grootschalige toepassing van de VpB in plaats van de PS is op de lange termijn moeilijk en op korte onmogelijk. Deze conclusie ligt in de lijn van eerdere conclusies over inzetbaarheid van VpB data, zoals bijvoorbeeld Lammertsma (2008). Er spelen een aantal problemen, deze kunnen verdeeld worden in 2 groepen, technische problemen met het bestand en inhoudelijke problemen:

De technische problemen zijn:

- De respons van de VpB is te langzaam om in september t+1 te beginnen met verwerking van de data. Maar t+2 is reëler. Maar dan blijft nog steeds het

probleem dat de resultaten nog niet volledig geconvergeerd zijn. Om een voorbeeld te noemen, de convergentie van de omzet is versneld met een DRT-filter en een imputatie methode. Maar er is nog veel onderzoek nodig om te bepalen wat voor de andere variabelen de beste methode is.

- Respons voor een variabele is het ene jaar wil nog niet zeggen dat er respons is voor die variabele in een ander jaar, zoals blijkt uit Tabel 14, terwijl aangenomen mag worden dat als er in beide jaren omzet is, er ook in beide jaren bijvoorbeeld arbeidskosten zijn. De vulling van de geleverde bestanden is dus niet compleet. Het heeft de voorkeur dat de belastingdienst completere records opstuurt waardoor de convergentie verbetert, een arbeidsintensief alternatief is het ontwikkelen van een imputatiemethode om de lege gaten in het bestand vol te schieten.

Inhoudelijk staan de volgende zaken toepassing van de VpB in de weg:

- Uit dit onderzoek blijkt dat er wel mogelijkheden zijn om voor bepaalde kerncellen volgens de PS+ methodiek en volgens een methodiek gebaseerd op het gemiddelde van een kerncel maal grootteklasse (paragrafen 2.3 en 2.5.2) maar dat dit zeker niet toepasbaar is voor alle kerncellen maal grootteklassen.
- Daarnaast roept dit onderzoek nog veel vragen op, voornamelijk als gevolg van de vele ontwikkelingen op dit gebied in de afgelopen jaren waardoor veel meer onderzoek gewenst en nodig is voordat de VpB ingezet kan worden. Impact 1 is vervangen door NOPS. DRT kan als filter voor foutieve waarden worden ingezet.
- Als de VpB toegepast zou gaan worden, is een goed filter nodig om records met foutieve waarden te verwijderen. Zoals in paragraaf 2.5.1 aangetoond kan de respons in de VpB records sterk fluctueren hetgeen de resultaten op macroniveau sterk beïnvloedt.
- Er zijn verschillende setups van de PS+ methodiek onderzocht, waarin zaken gevarieerd zijn zoals gaaf vs niet gaafgemaakte data, beslisriteria, datum voor download van het bestand en grootteklasse indeling. Hieruit blijkt dat de selectie kerncellen maal grootteklassen die gebruikt kunnen

- worden voor vervanging door VpB erg gevoelig is de gebruikte setup.
- Uit de volgtijdelijke vergelijkbaarheid met de PS van 2006 t/m 2008 blijkt slechts enkele kerncellen maal grootteklasse geselecteerd kunnen worden voor vervanging van de PS door VpB De levensduur van de beids in de VpB verschilt nauwelijks van die in het ABR. Dit levert dus geen probleem op.

Afsluitend kan gezegd worden dat met de huidige kennis het niet mogelijk is om de VpB op grote schaal in te zetten in plaats van de PS, in eerste instantie door de technische problemen met de bestanden en in de tweede plaats door inhoudelijke

problemen. Als de technische problemen niet opgelost kunnen worden, heeft vervolgonderzoek verder weinig zin om snel tot een grootschalige vervanging van PS door VpB data te komen.

4. Discussie

Naar aanleiding van dit rapport blijven er veel open vragen over. Meer onderzoek is nodig maar niet mogelijk als gevolg van de beschikbare tijd waarin het onderhavige onderzoek uitgevoerd moest worden en de opleverdatum van dit rapport. Als er tijd is voor uitgebreid onderzoek naar de toepassing van VpB-data, dan moeten de volgende vragen opgelost worden:

- De vulling van andere variabelen dan de omzet is slecht in records. (het aantal beids met respons is nog niet eens het grootste probleem). Kan met een imputatie methode de vulling verbeterd worden en daarmee de convergentie versneld worden?
- De convergentie van andere variabelen dan de omzet met een convergentie versnellingsmethode gebaseerd op de omzet is slecht. Als de versnellingsmethode meer zou worden toegespitst op die variabelen, kan dan winst behaald worden?
- Er is alleen een filter voor de omzet, de convergentie van de andere variabele wordt verstoord doordat deze niet gefilterd worden voor foutieve waarden. Hier is ook een filter voor nodig. In het overzicht dat Lammertsma (2008) presenteert van onderzoeken naar het gebruik van VPB data wordt regelmatig genoemd dat er 0 of negatieve waarden zitten bij bedragen die positief moeten zijn. In dit onderzoek is hier niet naar gekeken. Eenvoudige methoden om deze waarden of zelfs records te verwijderen zijn niet getest.
- De volgtijdelijke vergelijkbaarheid is nu gebaseerd op 2006-2008 data. In deze periode is geen DRT filter beschikbaar en is de indeling gebaseerd op SBI93. In de loop van 2012 komt PS2010 beschikbaar. Met deze data is het mogelijk om een vergelijking PS-VPB te maken over 2008-2010 met gebruik van DRT filter voor de VPB data en een indeling gebaseerd op SBI2008. Dit is een betere vergelijking omdat foutieve VPB data gefilterd wordt en de indeling recenter is.
- Daarnaast is het interessant om te onderzoeken of er beids zijn met een structureel andere omzet in de VpB dan in DRT door te kijken in de bestanden van 2008, 2009 en 2010 of als een beid een jaar een foutieve waarde is, dat in de andere jaren dat ook geldt.

5.1.2.e

- b. bij het een na kleinste bedrijf wordt de cumulatieve omzet berekend door de cumulatieve omzet van het kleinste bedrijf te nemen plus de wortel van de omzet van het bedrijf zelf.
 - c. Deze procedure wordt herhaald tot en met het bedrijf met de grootste omzet. Voor dit bedrijf is de cumulatieve omzet de som van wortels uit de omzetten van alle bedrijven.
4. De grootteklasse-indeling wordt bepaald op basis van de cumulatieve omzet.
- a. Grootteklasse 0 zijn alle bedrijven met omzet 0 of een negatieve omzet.
 - b. Grootteklasse 1 zijn alle bedrijven met een cumulatieve omzet kleiner dan $1/9$ van de maximale cumulatieve omzet
 - c. Grootteklasse 2 zijn alle bedrijven met een cumulatieve omzet groter dan $1/9$ en kleiner dan $2/9$ van de maximale cumulatieve omzet
 - d. Enz.

		grootteklasse omzet										Total
		0	1	2	3	4	5	6	7	8	9	
grootte- klasse werkz. Personen	1	157	10303	2402	170	64	23	8	5	4	1	13137
	2	195	3723	5878	1055	242	111	49	35	8	5	11301
	3	59	144	2138	1653	598	261	90	59	24	12	5038
	4	44	69	741	2057	1357	675	332	135	65	26	5501
	5	57	21	150	710	1408	1271	844	487	235	89	5272
	6	18	5	15	79	205	422	595	533	369	153	2394
	7	11	2	2	9	39	75	159	234	274	197	1002
	8	4	0	2	0	5	13	24	44	109	191	392
	9	4	0	0	0	0	0	2	9	13	91	119
Total		549	14267	11328	5733	3918	2851	2103	1541	1101	765	44156

Tabel 17: verschil tussen grootteklasse op basis van werkzame personen en omzet

In Tabel 17 staat het aantal bedrijven uit de PS2009 volgens de grootteklasse met werkzame personen en omzet. De diagonaal van de tabel bevat de meeste bedrijven, echter opvallend is dat vrijwel de gehele tabel ongelijk aan 0, zelfs in de meest extreme gevallen.