Evaluatie van de effectiviteit van diverse activerende arbeidsmarktmaatregelen bij VDAB

Contactpersoon:

Joost Bollens, Studiedienst VDAB

joost.bollens@vdab.be

INHOUD

Evaluatie van de effectiviteit van diverse activerende arbeidsmarktmaatregelen bij VDAB

VDAB, December 2022

SITUERING

VDAB (Vlaamse Dienst voor Arbeidsbemiddeling en Beroepsopleiding), de Vlaamse PES, werkt al sinds decennia samen met het ESF aan het versterken van het Vlaamse arbeidsmarktbeleid. Binnen de huidige ESF-programmeringsperiode is binnen het zogenaamde actor-dossier – dat betrekking heeft op maatregelen die door VDAB zelf worden uitgevoerd – de beroepsopleiding van werkzoekenden een belangrijke gecofinancierde maatregel. In het zogenaamde regie-dossier – dat betrekking heeft op maatregelen die VDAB samen met partners uitvoert – is dan weer de TIBB-maatregel ("Tender intensieve begeleiding en bemiddeling") de belangrijkste gecofinancierde maatregel. Binnen het actor-dossier, waarin naast de opleidingen in eigen beheer ook de (kleinere) IBO-maatregel zit, wordt over de periode 2016-2022 een ESF-inbreng voorzien van circa 112 miljoen €. Voor het regie-dossier is er, alvast voor de jaren 2016 tot en met 2019, onder prioriteit 8.1 (TIBB4) een ESF-inbreng van circa 12,1 miljoen €.

In de loop van 2019 werd door prof. Bart Cockx (Universiteit Gent) en prof. Michael Lechner (Universiteit van St. Gallen), in samenwerking met de studiedienst van VDAB, een onderzoek gevoerd waarin recente Machine Learning technieken werden gebruikt om de effectiviteit te schatten van diverse actief arbeidsmarktmaatregelen (ALMP's) die door VDAB worden beheerd. Vanaf het begin was het de bedoeling dat zowel de beroepsopleiding voor werkzoekenden als TIBB mee deel zouden uitmaken van de geëvalueerde maatregelen.

De resultaten van dit onderzoek werden in de loop van november 2022 geaccepteerd voor publicatie in Labour Economics, en zullen binnenkort aldaar kunnen worden geraadpleegd.

Cockx, B., Lechner, M., Bollens, J. (2022) Priority to Unemployed Immigrants? A Causal Machine Learning Evaluation of Training in Belgium, In Press, Labour Economics

Het onderstaand verslag is in belangrijke mate gebaseerd op deze paper, maar wijkt er op een aantal punten van af, de voornaamste verschilpunten zijn de volgende:

- In de loop van het onderzoek werd besloten om in de eindspecificatie de TIBB-maatregel niet mee te nemen. In de Engelse paper zal men dus geen resultaten voor TIBB vinden. In de voorliggende tekst worden wel resultaten voor TIBB gegeven, en wordt toegelicht waarom ze niet in de eindpublicatie zijn opgenomen;
- We voegden een sectie 2 toe waarin de onderzoeksvragen expliciet worden opgenomen;
- In sectie 7.4 is een aanvullend resultaat opgenomen;.
- Bovendien werd een sectie 9 toegevoegd waarin beleidsaanbevelingen voor VDAB worden opgesomd;
- De Engelse paper bevat een aantal appendices die in dit verslag niet zijn herhaald.

1 INLEIDING

De werkloosheid blijft een belangrijk economisch en sociaal probleem in Europa, ook al is de (algemene) werkloosheid in de Europese Unie (EU) gestaag gedaald van 10,9% in 2013 tot 6,8% in 2018 (Eurostat) (al is er natuurlijk recent weer sprake van een opstoot). Het probleem treft in het bijzonder sommige kwetsbare groepen, zoals jongeren, oudere werknemers en migranten. Beleidsmakers zijn daarom blijvend geïnteresseerd in een beter inzicht in welk arbeidsmarktbeleid werkt voor wie. Een dergelijk inzicht draagt bij tot de verbetering van het begeleidingsproces, het opzet en de allocatie van werkzoekenden naar het actief arbeidsmarktbeleid.

Het blootleggen van heterogeniteit in de effectiviteit van het beleid ("wat werkt voor wie?") is echter een uitdaging vanuit econometrisch standpunt, omdat hiervoor schatters nodig zijn die tegelijkertijd voldoende flexibel en voldoende nauwkeurig zijn bij het voorspellen van de causale effecten op een gedesaggregeerd niveau. Recente ontwikkelingen op het gebied van het Causal Machine Learning (CML) hebben dit probleem aangepakt en veelbelovende oplossingen aangedragen. In dit onderzoek gebruiken we een dergelijke CML-benadering om de heterogeniteit in de effectiviteit van diverse arbeidsmarktmaatregelen van VDAB te evalueren. Vervolgens gebruiken we deze schattingen om specifieke heterogeniteit aan het licht te brengen en om aan te tonen in welke mate de VDAB de effectiviteit van deze programma's kan vergroten door de toewijzing van werkloze werkzoekenden aan deze programma's te wijzigen.

Voor het maken van voorspellingen wordt reeds geruime tijd gebruik gemaakt van benaderingen van Machine Learning (bvb. Hastie, Tibshirani en Friedman, 2009). Meer recentelijk zijn deze methoden zodanig aangepast dat ze ook bruikbaar zijn binnen het domein van de causale gevolgtrekking (zie Athey 2019, en Athey en Imbens, 2019, voor overzichten). Deze literatuur laat zien hoe het counterfactual causaal probleem (bijvoorbeeld Imbens en Wooldridge 2009) kan worden omgezet in een combinatie van specifieke voorspellingsproblemen. Voor het voorliggend onderzoek zijn deze methoden bijzonder interessant omdat ze op een systematische manier mogelijk maken om de onderliggende heterogeniteit van de causale effecten bloot te leggen, iets waarvoor de traditionele econometrische methoden geen systematische oplossing bieden. Tegelijkertijd zullen we laten zien dat de Machine Learning-methoden zeer vergelijkbare en dus betrouwbare schattingen van de gemiddelde behandelingseffecten zullen opleveren als de traditionele methoden.

De identificatie van het causale effect berust op de veronderstelling van "unconfoundedness"[1]. Knaus, Lechner en Strittmatter (2018) evalueerden de prestaties van verschillende CML-methodieken voor binaire behandelingen, alle gebaseerd op unconfoundedness, met behulp van een Empirische Monte Carlo-benadering (zie bijvoorbeeld Huber et al. 2013; Lechner en Wunsch 2013). In tegenstelling tot een standaard Monte Carlo-analyse, wordt in een dergelijke aanpak het gegevensgenererende proces (DGP, "data generating process") zoveel mogelijk gebaseerd op reële gegevens en reduceert het de synthetische componenten in het DGP tot een minimum. De reële gegevens zijn ontleend aan de Zwitserse sociale zekerheidsgegevens die werden gebruikt voor de evaluatie van een programma dat werklozen helpt bij het zoeken naar werk (Knaus, Lechner en Strittmatter 2017). Knaus, Lechner en Strittmatter (2018) komen tot de conclusie dat de op het Causal Forest gebaseerde ML-methoden, met name het General Forest van Athey, Tibshirani en Wager (2018), tot de best presterende schatters behoren, tenminste als ze expliciet worden aangepast om rekening te houden met confounding. Vervolgens stelt Lechner (2018) de Modified Causal Forest (MCF) schatter voor. Deze bouwt voort op de door Wager en Athey (2018) en Athey et al. (2018) voorgestelde schatters. Een innovatie is dat Lechner (2018) de objectieffunctie die gebruikt wordt om de bomen van het Causal Forest te bouwen, in één enkele stap verbetert door het penaliseren van splitsingen die de selectie-bias niet verminderen. De tweede voorgestelde innovatie is het gebruik van weight-based inferentiemethoden, die op een rekenkundig goedkope en betrouwbare manier toelaten om de precisie van de geschatte behandelingseffecten op de verschillende aggregatieniveaus te schatten, van de geïndividualiseerde tot de (gegroepeerde) gemiddelde behandelingseffecten.

Op basis van een empirische Monte Carlo-analyse toont Lechner (2018) aan dat de MCF-schatter beter presteert dan eerder gesuggereerde schatters in niet-experimentele omgevingen. Aangezien de MCF zowel de heterogeniteit van het programma als de individuele heterogeniteit op verschillende niveaus effectief kan aanpakken, aantrekkelijke theoretische eigenschappen heeft en in de praktijk goed lijkt te presteren, is het onze voorkeurschatter.

[1] De "unconfoundedness"- veronderstelling stelt o.m. dat er informatie beschikbaar is over alle variabelen die zowel van invloed zijn op de selectiebeslissing als op de uitkomst, zie sectie 4.2 voor een meer nauwkeurige bepaling.

Voor zover wij weten is dit onderzoek één van de eerste die CML-methodieken toepast om treatment-heterogeniteit te analyseren bij de evaluatie van actief arbeidsmarktbeleid (ALMP). Het blijkt ook de eerste paper te zijn die een dergelijke analyse uitvoert in een context met meer dan één treatment. Knaus, Lechner en Strittmatter (2017) gebruiken op Lasso gebaseerde methoden om de effect-heterogeniteit van een (enkel) jobsearch-programma in Zwitserland te evalueren met behulp van administratieve gegevens uit 2003. Zij vinden substantiële effect-heterogeniteit, maar alleen gedurende de eerste 6 maanden na de start van het programma. Bertrand, Crépon, Marguerie en Premand (2017) passen de Causal Random Forest methode van Wager en Athey (2018) toe binnen een experimentele benadering (RCT) om de programma-heterogeniteit te evalueren van tijdelijke openbare werken in een minder ontwikkeld land. Hun analyse brengt belangrijke heterogeniteit aan het licht, maar ook hier weer vooral tijdens de programmadeelname.

In tegenstelling tot de eerder genoemde papers, wordt in deze paper gewerkt met verschillende programma's tegelijkertijd en wordt expliciet gebruik gemaakt van de geschatte effect-heterogeniteit om herverdelingsregels voor te stellen die de prestaties van het programma zouden kunnen verbeteren.

De rest van de paper is als volgt gestructureerd. In sectie 2 lijsten we de onderzoeksvragen op, en in sectie 3 beschrijven we de maatregelen en data die in de analyse worden gebruikt. Sectie 4 bespreekt de econometrische methoden. Sectie 5 presenteert de resultaten met een focus op de analyse van de effect-heterogeniteit. Sectie 6 simuleert verschillende alternatieve toewijzingsregels. In sectie 7 volgt een robuustheidsanalyse, inclusief de placebo-analyse, en in sectie 8 volgt de conclusie. In sectie 9 wordt tot slot stilgestaan bij de beleidsaanbevelingen voor VDAB.

2 ONDERZOEKSVRAGEN

Naast een aantal methodologische vragen (zie sectie 1 en verder), zijn er in dit onderzoek ook een aantal meer inhoudelijke onderzoeksvragen:

[1] VDAB heeft diverse ALMP's in zijn aanbod. Voor wat betreft de in dit onderzoek opgenomen ALMP's: is er een verschil tussen deze programma's onderling wat betreft hun effectiviteit op het vlak van het verhogen van de kans op werk, het verminderen van de kans op werkloosheid en het verminderen van het verblijf buiten-de-arbeidsmarkt? Zit er een bepaald tijdspatroon in het verloop van deze effecten, en is dat verschillend voor de verschillende ALMP's?

[2] Is er sprake van heterogeniteit in de effectiviteit van de programma's? Zo ja, kan dan worden geëxpliciteerd of deze heterogeniteit gerelateerd is aan bepaalde kenmerken van de werkzoekenden? Is dit gelijk of verschillend voor de diverse bestudeerde ALMP's? Is er, binnen ieder specifiek bestudeerd ALMP ook sprake van een grote variatie in de individuele effectiviteit?

[3] Is er een verschil tussen het gemiddeld populatie-effect (ATE) enerzijds, en de gemiddelde effectiviteit voor de groep van personen die daadwerkelijk deelnamen aan de diverse ALMP's (ATET) anderzijds?

[4] Is de allocatie van werkzoekenden naar de verschillende programma's optimaal? Zou men door het maken van andere keuzes bij de allocatie het resultaat op het vlak van effectiviteit kunnen verbeteren? Welke keuzes zou men hier kunnen overwegen?

3 DATA

3.1 BESTUDEERDE SUBPOPULATIES

De gegevens voor de analyse zijn afkomstig uit administratieve dossiers van alle personen die zich tussen januari 1991 en februari 2019 als werkloze werkzoekende bij de VDAB hebben ingeschreven. Deze dossiers bevatten rijke socio-demografische informatie, alsook de evolutie van de individuele arbeidsposities sinds 1991. Uit deze databank selecteren we 148942 personen die na een onvrijwillig ontslag tussen december 2014 en juni 2016 een werkloosheidsuitkering zijn gaan claimen. Individuen die na juni 2016 werkloos worden, werden niet geselecteerd omdat we de deelnemers gedurende een voldoende lange periode willen opvolgen. Omdat we in wat volgt ALMP's (active labour market policies) selecteren die tot 9 maanden na het begin van de

werkloosheidsperiode zijn gestart en omdat de observatieperiode in september 2019 afloopt, behouden we de arbeidsmarktresultaten tot 30 maanden na de start van het programma.

We sluiten schoolverlaters die aanspraak maken op een werkloosheidsuitkering en personen jonger dan 21 jaar uit. We sluiten ook werknemers met een handicap en degenen die ouder zijn dan 55 jaar aan het begin van de werkloosheidsperiode uit, omdat deze groepen niet volledig beschikbaar hoeven te zijn voor de arbeidsmarkt of kunnen profiteren van alternatief beleid. We hebben ook personen die niet in Vlaanderen wonen en personen die tijdens de analyseperiode zijn overleden, laten vallen. 73.582 personen worden behouden na het opleggen van deze selectiecriteria.

3.2 DE ARBEIDSMARKTPROGRAMMA'S

Initieel werden binnen het programma-aanbod van VDAB een vijftal grote programma's afgebakend die potentieel interessant waren voor het onderzoek: omwille van inhoudelijke redenen, maar ook omdat ze een voldoende groot aantal deelnemers hadden. In wat volgt, beschrijven we kort deze programma's. Vervolgens wordt toegelicht welke selectie er uiteindelijk voor dit onderzoek werd gemaakt.

Wanneer werkzoekenden geen realistisch jobdoel hebben, kan men ze doorverwijzen naar een Oriëntatie training. Binnen zo een programma zal men trachten om samen met de werkzoekende toch één of meer concrete jobdoelen af te bakenen, en zal men nagaan welke verdere stappen er kunnen worden ondernomen op weg naar dat doelwit.

Voor werkzoekenden die wel een realistisch jobdoel hebben, maar die tekort schieten op het vlak van de beroepsgerichte competentie die nodig is om dit doel te bereiken, kan de meer klassieke beroepsopleiding worden ingezet. VDAB beschikt over een uitgebreid netwerk van beroepsopleidingen die voorbereiden op beroepen in de diverse economische sectoren. Dit gaat van korte opleidingen van enkele dagen tot onderwijsopleidingen van verschillende studiejaren.

TIBB staat voor Tender Intensieve Begeleiding en Bemiddeling. Deze maatregel is toegankelijk voor werkzoekenden en voor jongeren in beroepsinschakelingstijd. Het programma is gericht op personen met een duidelijk jobdoel, die geen nood hebben aan beroepsgerichte competentieversterking, maar die omwille van diverse redenen wél nood hebben aan een intensieve bemiddeling. Dit kan dan bvb. over één van de volgende redenen gaan: gebrek aan juiste sollicitatievaardigheden, een gebrekkige kennis van de arbeidsmarkt, gebrekkige arbeidsattitudes, of andere belemmerende randvoorwaarden (mobiliteit, combinatie werk en vrije tijd, tijdsbeheer, eigen middelenbeheer, huishouden, persoonlijke administratie,…).

Nederlands voor anderstaligen is een programma gericht op werkzoekenden die willen geholpen worden bij het leren van de Nederlandse taal.

Bij een Individuele Opleiding in de Onderneming (IBO) zal een werkzoekende gedurende een periode van maximaal 6 maanden worden opgeleid op de werkvloer binnen een onderneming. De onderneming wordt geacht de werkzoekende op het einde van de opleiding een contract aan te bieden.

Naast deze vijf grote programma's, zijn er nog vele, kleinere programma's.

Tabel 3.1 geeft aan hoe de populatie uiteindelijk over de vier volgende subgroepen is verdeeld:

(1) Er zijn 56.324 personen die in de eerste negen maanden van hun werkloosheid aan geen enkele ALMP hebben deelgenomen, d.w.z. de nog niet behandelde groep (Sianesi, 2004);

(2) 3.640 personen die binnen de eerste 9 maanden aan een ALMP zijn begonnen, die in de analyse zijn aangehouden;

(3) 13.618 personen die in de eerste negen maanden aan een ALMP zijn begonnen en niet voor de uiteindelijke evaluatie zijn geselecteerd, omwille van één van de volgende redenen:

> (i) bijna alleen buitenlanders met een beperkte talenkennis nemen deel aan de Nederlandse taaltraining, zodat er te weinig vergelijkingsobservaties beschikbaar waren;
>
> ii) de IBO-maatregel (individuele beroepsopleiding in de onderneming) is niet vergelijkbaar met de andere ALMP, aangezien deelnemers onmiddellijk aan het werk gaan;
>
> iii) een aantal eerder kleine ALMP's konden niet worden samengevoegd in betekenisvolle groepen die groot genoeg zijn voor een empirische analyse, of behoorden tot een categorie die gemiddeld te lang

duurde (10 maanden of meer) voor een evaluatie van de effecten op de middellange termijn binnen de waarnemingsperiode van 30 maanden;.

(i) in de laatste modelspecificatie, waarin de deelnemers tot 30 maanden na de start worden opgevolgd, gaven placebo-tests (zie sectie 7.1) aanleiding tot bezorgdheid over een mogelijke aanwezigheid van een selectiebias in de resultaten van de TIBB-maatregel. Omwille van deze reden werden de TIBB-resultaten voor de uiteindelijke publicatie niet weerhouden. In een vroegere specificatie, waarin de deelnemers tot 23 maanden na de start werden opgevolgd (in plaats van tot 30 maanden), passeerden de TIBB-resultaten evenwel nog wel de placebo-testen. Aangezien TIBB een belangrijke maatregel is binnen het ESF-programma, zullen daarom in wat volgt toch een aantal van deze vroegere TIBB-resultaten worden gerapporteerd. Dit betekent dus dat naast resultaten voor de 3.640 personen waarvan sprake in (2), er ook een aantal resultaten zullen worden besproken voor 3695 deelnemers aan TIBB.

De deelnemers aan het programma worden ingedeeld volgens het eerste programma waaraan zij deelnemen. Binnen het beroepsopleidingsprogramma wordt een onderscheid gemaakt tussen korte (minder dan 6 maanden, gemiddeld 3,8 maanden) en lange (meer dan 6 en minder dan 10 maanden, gemiddeld 7,8 maanden) programma's (SVT en LVT). Zeer lange beroepsopleidingsprogramma's, met een duur van 10 maanden of meer, werden ondergebracht bij maatregel 7, “Andere ALMP's” en werden vervolgens uit de analyse geschrapt omdat daar geen voldoende lange follow-up mogelijk is.

Ten derde is de oriëntatietraining (OT) gericht op het helpen bepalen van een duidelijk beroepsdoel. Dit programma is relatief kort. Het duurt gemiddeld slechts ongeveer een maand.

Voor een aantal resultaten (zie tabel 5.2) zal ten vierde ook gerapporteerd worden over TIBB.

De dataset bevat 45 geordende en 9 categorische variabelen (met 3 tot 44 ongeordende categorieën). Alle tijdsvariabelen worden gemeten aan het begin van de werkloosheidsperiode. Rekening houdend met het feit dat gewoonlijk categorische variabelen worden omgezet in dummy-variabelen in een regressie-achtige methode, zou dit overeenkomen met 175 variabelen, en zelfs veel meer als men parametrische beperkingen wil vermijden door het opnemen van interacties en hogere orde-termen (kwadraat, kubiek enz.), dingen waarmee de MCF automatisch rekening zal houden.

Tabel 3.1: ALMP gevolgd door cohortes die werkloos werden tussen december 2014 en juni 2016

Type	Gemiddelde duur (maanden)	Aantal	Aandeel
Geen deelname aan ALMP gedurende 1e 9 maand (NOP)[1]	-	56,324	76.5%
ALMP gedurende de eerste 9 maanden, opgenomen in de analyse		3,640	4.9%
1. Korte (< 6 months) beroepsopleiding (SVT)	3.83	1,305	1.8%
2. Lange (< 10 months) beroepsopleiding (LVT)	7.18	1,220	1.7%
3. Orientation training (OT)	1.05	1,115	1.5%
ALMP gedurende de eerste 9 maanden, uitgesloten uit de analyse (maar wél een aantal resultaten voor TIBB)		13,618	18.5%
4. TIBB (TIBB)		3,695	5.0%
5. Nederlands als vreemde taal (NED)	2.56	991	1.3%
6. IBO (IBO)	-	2,045	2.8%
7. Andere ALMP, waaronder zeer lange opleiding	-	6,887	9.4%
Total		73,582	100.0%

Note: De geselecteerde personen zijn tussen 21 en 55 jaar en claimden een werkloosheidsuitkering na ontslag in de periode december 2014-juni 2016. Uitgesloten werd wie (i) niet in Vlaanderen woont; (ii) een arbeidshandicap had; (iii) overleden is gedurende de analyseperiode

[1] Deze personen kunnen wél in één van de maatregelen gestart zijn na de eerste 9 maanden.

Deze variabelen geven informatie over persoonlijke sociaal-demografische kenmerken, de arbeidsmarktgeschiedenis, inclusief ziekte, in de voorafgaande 2, 5 en 10 jaar, de ALMP-participatiegeschiedenis

tijdens eerdere werkloosheidsperiodes, informatie over de werkvoorkeuren van de werkzoekende en de bijbehorende beroepservaring, de kalendermaand waarin de werkloosheidsperiode startte (19 indicatoren) en de dag waarop het ALMP begon (of werd voorspeld te beginnen in geval van geen deelname) in de werkloosheidsperiode (maximaal 274 dagen).

Tabel 3.2 geeft een overzicht van de gemiddelde waarden van een aantal covariaten (panel A) en van een aantal uitkomstvariabelen (panel B): het gemiddelde per programma en de gestandaardiseerde verschillen (in %) voor elke programma (SVT, LVT, OT) ten opzichte van de NOP-groep die in de eerste negen maanden van de werkloosheidsperiode niet aan een ALMP heeft deelgenomen. De gestandaardiseerde verschillen zijn vaak groter dan 20%, een getal dat Rosenbaum en Rubin (1985) als 'groot' beschouwen. Dit geeft aan dat de conditioneringsvariabelen van deelnemers aan ALMP zeer onevenwichtig zijn ten opzichte van de NOP-groep en dat controle voor de selectievertekening cruciaal is in deze setting.

We constateren dat er veel meer mannen dan vrouwen deelnemen aan de beroepsopleiding. Deelnemers aan het beroepsonderwijs zijn iets vaardiger in het Nederlands, met name in de groep LVT. Deelnemers aan het een hebben een vergelijkbare werkeloosheidservaring als niet-deelnemers, terwijl de deelnemers aan een LVT de afgelopen 2 en 10 jaar duidelijk minder werkloos zijn geweest. Deelnemers aan het beroepsonderwijs zijn minder goed opgeleid dan niet-deelnemers, terwijl deelnemers aan het beroepsonderwijs gemiddeld hoger opgeleid zijn.

Tabel 3.2: Gemiddelde en gestandaardiseerde verschillen voor een aantal variabelen

Variabele	Geen ALMP deelname (NOP)	Korte beroeps-opleiding (SVT)		Lange beroeps-opleiding (LVT)		Orientatie-training (OT)	
A. Covariaten	Steekproefgemiddelde (gestandaardizeerd verschil**100 in verhouding tot NOP*)[1]						
Vrouw	0.49	0.31	*(36)*	0.40	*(16)*	0.46	*(3)*
Leeftijd in jaren	35	34	*(12)*	34	*(12)*	34	*(16)*
Kennis van Nederlands (0-3)[2]	2.4	2.5	*(8)*	2.7	*(40)*	2.6	*(27)*
Aantal maanden werkloos in de laatste 10 jaar	18	19	*(5)*	16	*(11)*	17	*(4)*
Aantal maanden werkloos in de laatste 2 jaar	3.9	3.8	*(2)*	3.0	*(18)*	3.2	*(14)*
Opleidingsniveau (1 to 13)	7.2	5.9	*(38)*	7.9	*(22)*	7.2	*(1)*
B. Uitkomsten							
# maanden gewerkt 10 maand na start ALMP[3]	4.0	3.9	*(2)*	2.8	*(33)*	2.4	*(45)*
# maanden gewerkt 20 maand na start ALMP[3]	9.8	11	*(16)*	9.4	*(5)*	7.8	*(27)*
# maanden gewerkt 30 maanden na start ALMP[3]	16	18	*(24)*	17	*(11)*	15	*(12)*
Number of observations	59964	1305		1220		1115	

Notes: [1] Het gestandaardizeerde verschil wordt als volgt gedefinieerd: $\left|\bar{x}^j - \bar{x}^{NOP}\right| / \sqrt{[Var(x^j) + Var(x^{NOP})]/2} * 100$, met $\bar{x}^j$ en $Var(x^j)$ het gemiddelde en de variantie van de variabele x^j voor $j \in \{SVT, LVT, DLT, IC\}$.

[2] Kennis Nederlands = 0 als geen; = 1 als beperkt; = 2 als goed; =3 als zeer goed.

[3] Voor niet-deelnemers (NOP) is dit de datum van de voorspelde start (zie verder).

Deelnemers aan een oriëntatietraining hebben veel meer kans op een goede kennis van het Nederlands. Bovendien zijn ze de afgelopen twee jaar gemiddeld veel minder werkloos geweest. Ze hebben een grotere kans om een gemiddeld niveau van scholing te hebben. Dit profiel lijkt dus overeen te komen met het profiel van een gemiddeld geschoolde die een routinefunctie heeft vervuld en zijn baan heeft verloren in de geleidelijke tendens naar meer polarisatie op de arbeidsmarkt (zie bijvoorbeeld Autor et al. 2003; Goos et al. 2009). Deze werknemers moeten zich meestal heroriënteren naar een ander beroep, omdat ze vaardigheden hebben die niet meer gevraagd worden.

Panel B van tabel 3.2 rapporteert de samenvattende statistieken voor drie belangrijke uitkomstvariabelen, namelijk het cumulatief aantal maanden dat een werknemer 10, 20 en 30 maanden na de start van het ALMP in dienst is. In het empirische deel worden de laatste (30 maanden) en enkele aanvullende uitkomstvariabelen in aanmerking genomen.

Uit panel B van tabel 3.2 kan worden afgeleid dat de uitkomsten per programma aanzienlijk verschillen. Gezien de grote variatie in de covariaten (panel A), zijn deze beschrijvende statistieken echter niet noodzakelijkerwijs informatief over causale gemiddelde programma-effecten, mogelijk speelt hier een selectievertekening. Hoe dan

wel conclusies kunnen worden getrokken over de effecten van deze programma's, wordt in de volgende sectie besproken.

4 METHODOLOGIE

4.1 HET KADER VAN DE CAUSALE MODELLERING EN DE RELEVANTE PARAMETERS

We gebruiken Rubins (1974) model van potentiële uitkomsten om een model met verschillende behandelingen onder *unconfoundedness*, of voorwaardelijke onafhankelijkheid, te beschrijven (Imbens, 2000, Lechner, 2001). Weze D de behandeling die in ons geval niet-deelname en deelname aan één van de drie programma's is. D kan dus vier verschillende gehele waarden aannemen, gaande van 0 tot 3[2]. De (potentiële) uitkomst die onder behandeling d wordt gerealiseerd, wordt aangeduid met Y^d. Voor elk individu observeren we alleen het specifieke potentiële resultaat dat verband houdt met de behandelingsstatus die dit individu heeft gekozen,

$$y_i = \sum_{d=0}^{3} \underline{1}(d_i = d) y_i^d$$

($\underline{1}(\cdot)$ geeft een indicatorfunctie aan, die één is als het argument waar is en nul anders). Er zijn twee groepen van variabelen waarop wordt geconditioneerd, $\tilde{X}$ en Z. $\tilde{X}$ bevat die covariaten die nodig zijn om te corrigeren voor selectievertekeningen (confounders), terwijl Z variabelen bevat die (groepen van) leden van de populatie definiëren waarvoor een gemiddelde schatting van het causale effect gewenst is. Voor de identificatie kunnen $\tilde{X}$ en Z discreet of continu, of beide, maar voor de schatting zullen we alleen discrete Z overwegen. Ze kunnen elkaar op enigerlei wijze overlappen. In overeenstemming met de literatuur m.b.t. machine leren noemen we ze voortaan 'features'. Stel de unie van de twee groepen variabelen gelijk aan X, $X = \left\{ \tilde{X}, Z \right\}, \dim(X) = p$.

In wat volgt, onderzoeken we de volgende gemiddelde causale effecten:

$$IATE(m,l;x,\Delta) = E(Y^m - Y^l \mid X = x, D \in \Delta) \ ,$$

$$GATE(m,l;z,\Delta) = E(Y^m - Y^l \mid Z = z, D \in \Delta) = \int IATE(m,l;x,\Delta) \, f_{X|Z=z,D\in\Delta}(x) dx \ ,$$

$$ATE(m,l;\Delta) = E(Y^m - Y^l \mid D \in \Delta) = \int IATE(m,l;x,\Delta) \, f_{X|D\in\Delta}(x) dx \ .$$

De geïndividualiseerde gemiddelde behandelingseffecten (IATE's, “Individualized average treatment effect”), $IATE(m,l;x,\Delta)$, meten het gemiddelde effect van de behandeling m in vergelijking met behandeling l voor eenheden met features x die behoren tot behandelingsgroepen Δ, waarbij Δ staat voor alle behandelingen van belang. De IATE's vertegenwoordigen de causale parameters op het laagste aggregatieniveau van de beschikbare kenmerken.

Aan het andere uiterste zijn er dan de gemiddelde behandelingseffecten (ATE's, “Average treatment effect”) die de populatiegemiddelden vertegenwoordigen. In het geval dat Δ betrekking heeft op de populatie met D=m, bekomt men het Gemiddelde behandelingseffect op de behandelden (ATET, “average treatment effect on the treated”) voor behandeling m. ATE en ATET zijn de klassieke parameters die in veel econometrische causale studies worden onderzocht. De parameters m.b.t. het gegroepeerde gemiddelde behandelingseffect (GATE, “Group Average Treatment Effect”) liggen tussen deze twee uitersten wat betreft hun aggregatieniveau. Voorafgaand aan de schatting wordt door de analist beslist welke covariaten beleidsrelevant zijn, deze vormen dan de verzameling Z, waarvoor GATE’s worden geschat. De IATE's en de GATE's zijn bijzondere gevallen van de zogenaamde Conditional Average Treatment Effects (CATE's).

[2] Of gaande van 0 tot 4 in de specificatie waarin ook TIBB wordt meegenomen, cf. tabel 5.2.

4.2 IDENTIFICATIE

De klassieke unconfoundedness-veronderstellingen bestaat uit de volgende delen (zie Imbens, 2000, Lechner 2001):

$$\{Y^0, Y^1, Y^2, Y^3\} \coprod D \mid X = x, \qquad \forall x \in \chi; \qquad (CIA)$$

$$0 < P(D = d \mid X = x) = p_d(x), \qquad \forall x \in \chi, \forall d \in \{0, ..., 3\}; \qquad (CS)$$

$$Y = \sum_{d=0}^{3} \underline{1}(D = j) Y^d; \qquad (SUTVA)$$

De Conditional Independence-veronderstelling (CIA) houdt in dat er geen andere features dan X zijn die de behandeling en de potentiële uitkomsten gezamenlijk beïnvloeden. De Common Support-veronderstelling (CS) bepaalt dat voor elke waarde in de support χ , er de mogelijkheid moet zijn om alle behandelingen te observeren. De aanname van de stable-unit-treatment (SUTVA) houdt in dat de concrete behandeling die een bepaalde persoon volgt, niet afhankelijk is van de behandelingen die door andere leden van de populatie worden gevolgd. Meestal is het voor een interessante interpretatie van de effecten nodig te veronderstellen dat X niet beïnvloed wordt door de behandeling (exogeneiteit). Als deze aanname geldt, zijn alle IATE's geïdentificeerd:

$$\begin{aligned} IATE(m, l; x, \Delta) &= E(Y^m - Y^l \mid X = x, D \in \Delta) \\ &= E(Y^m - Y^l \mid X = x) \\ &= E(Y^m \mid X = x, D = m) - E(Y^l \mid X = x, D = l) \\ &= E(Y \mid X = x, D = m) - E(Y \mid X = x, D = l) \\ &= IATE(m, l; x); \qquad \forall x \in \chi, \forall m \neq l \in \{0, ..., 3\}. \end{aligned}$$

Het is natuurlijk belangrijk om aan te tonen dat deze voorwaarden aannemelijk zijn in ons onderzoek. Laten we ze op hun beurt bekijken. Voorheen stelden we al dat de beschikbaarheid van een breed scala aan sociaal-demografische informatie en van rijke informatie over de arbeidsmarktgeschiedenis van individuen de plausibiliteit van de CIA vergroot. Dit zijn in essentie de features die in andere evaluatiestudies als de belangrijkste confounders zijn geïdentificeerd (bijvoorbeeld Heckman et al., 1998; Lechner en Wunsch, 2013). Dit zijn ook de features waarover de bemiddelaar tijdens het interview beschikt en die zij dus vooral moet gebruiken om haar beslissing te onderbouwen. Voordelen van onze studie ten opzichte van de literatuur over de evaluatie van het opleidingsprogramma zijn de beschikbaarheid van ziektegegevens en het werkloosheidspercentage in het arrondissement van de woonplaats. Een gebrek is dan weer de afwezigheid van informatie over het vroegere loon. Dit is echter misschien niet zo belangrijk, omdat er proxies voor het vroegere loon beschikbaar zijn, zoals opleiding, nationaliteit, de sector van de vorige jobs, de duur van de vorige tewerkstellingsperiode en het gewenste voorkeurberoep van de werkzoekende. Over het geheel genomen kunnen we concluderen we CIA aannemelijk is voor deze studie. De placebo-resultaten (zie sectie 7.1) hebben als doel om mogelijke overtredingen van de CIA te detecteren. Voor de maatregelen SVT, LVT en OT werden geen problemen gedetecteerd, voor TIBB evenmin in een versie van het onderzoek dat zich beperkte tot de periode tot 23 maanden na de start (maar wél in de definitieve versie, waar er werd opgevolgd tot 30 maanden na de start).

Aannemelijk is dat aan SUTVA wordt voldaan, aangezien alle onderzochte programma's vrij klein zijn in vergelijking met de beroepsbevolking. Common Support is een voorwaarde die in de gegevens kan worden gecontroleerd. We hebben geen common support-problemen met de onderzochte programma's ontdekt. Ten slotte wordt de exogeneiteit van confounders gewaarborgd door alle tijdsvariabelen aan het begin van de werkloosheidsperiode te meten. Op dat moment wist het individu niet of, en wanneer ze aan een opleiding zou beginnen.

4.3 SCHATTING

In dit artikel maken we gebruik van de recent verschenen literatuur m.b.t. causal machine learning (zie Athey 2019, en Athey en Imbens, 2019, voor overzichten). Het combineert de voorspellende kracht van de machine learning-literatuur (zie voor een overzicht bijvoorbeeld Hastie, Tibshirani en Friedman, 2009) met de micro-economische literatuur over het definiëren en identificeren van causale effecten (bijvoorbeeld Imbens en

Wooldridge, 2009). De laatste tijd is er in deze literatuur een aanzienlijke toename van voorgestelde methoden, en met name in de epidemiologie en de econometrie. Knaus, Lechner en Strittmatter (2018) vergelijken in een simulatieoefening veel van deze methoden systematisch met betrekking tot hun opzet en hun prestaties. Een van de conclusies van hun paper is dat methoden op basis van een random forest beter presteren dan alternatieve schattingsmethoden.

Het uitgangspunt van de causal forest-literatuur is de causale boom die werd geïntroduceerd in een paper van Athey en Imbens (2016). In een causale boom wordt de steekproef achtereenvolgens opgesplitst in kleinere en kleinere strata, waarbij de waarden van X steeds homogener worden, om de selectie-effecten te beperken en om de heterogeniteit van de effecten aan het licht te brengen. Als de splitsing eenmaal is beëindigd op basis van een of ander stopcriterium, wordt het behandelingseffect binnen elk stratum (een "blad" genoemd) berekend door het verschil te berekenen tussen de gemiddelde resultaten van de behandelden en de vergelijkingsgroep. In de literatuur over regressiebomen wordt echter erkend dat deze benadering nogal onstabiel kan zijn vanwege de sequentiële aard ervan (als de eerste splitsing anders is, zal de volledige boom waarschijnlijk tot verschillende eindlagen leiden). Een oplossing voor dit probleem is de zogenaamde random forest-schatter. Het belangrijkste idee is om enige willekeur in het proces van de boomopbouw te veroorzaken, veel bomen te bouwen en dan het gemiddelde te nemen van de voorspellingen van die vele bomen. De geïnduceerde willekeurigheid wordt gegenereerd door gebruik te maken van willekeurig gegenereerde substeekproeven (of bootstrap steekproeven) en door voor elke splitsingsbeslissing alleen een willekeurige selectie van de covariaten te maken. Wager en Athey (2018) gebruiken dit idee om causal forests voor te stellen, die gebaseerd zijn op een verzameling van causale bomen met kleine eindbladeren. Lechner (2018) ontwikkelt deze ideeën verder door de splitsingsregel voor de individuele bomen te verbeteren, door met name splitsingen die de selectiebias niet verminderen te bestraffen, en door methoden aan te reiken om heterogene effecten voor een beperkt aantal discrete beleidsvariabelen te schatten (Group Average Treatment Effects, GATE) en dit aan lage rekenkosten, naast de sterk uitgesplitste effecten waar de literatuur zich tot nu toe op richtte (Individualized Average Treatment Effects, IATE). Verder suggereert Lechner (2018) een manier om eenduidige conclusies te trekken voor alle aggregatieniveaus. Tot slot is de aanpak toepasbaar op een meervoudig, discreet behandelingskader. Aangezien veel van deze voordelen belangrijk zijn in de empirische analyse van dit onderzoek, wordt deze aanpak, die Modified Causal Forests (MCF) wordt genoemd, hieronder gebruikt. Voor alle verdere technische details van de schatter wordt verwezen naar Lechner (2018).

4.4 PRAKTISCHE UITVOERING

4.4.1 Resultaat- en controlevariabelen

We beschouwen drie soorten van uitkomstvariabelen: werk, werkloosheid en een restcategorie die we “buiten de arbeidsmarkt” noemen. Deze wordt gedefinieerd als niet werkend noch werkloos. Deze variabelen worden ofwel op een bepaalde afstand tot het begin van het programma gemeten, ofwel op een gecumuleerde manier als een som over een bepaalde periode.

De controlevariabelen zijn in de voorgaande paragrafen al besproken. Het is natuurlijk interessant om te begrijpen welke van deze features belangrijk zijn in de schatting. In de klassieke programma-evaluatie van gemiddelde populatie-effecten zou dergelijke informatie worden afgeleid uit de geschatte propensity score. Voor random forest schatters is de zogenaamde “variable importance measure” informatief wat betreft de relevantie van één variabele, gegeven alle andere. In ons geval bestonden de belangrijkste variabelen uit het geboorteland, de taalvaardigheid, de gesimuleerde startdatum, de arbeidsmarktgeschiedenis (werkgelegenheid en werkloosheid over verschillende horizonten in het verleden), de regio en de sector van de laatste tewerkstelling. Het is echter belangrijk om op te merken dat deze test variabelen zal kiezen die ofwel relevant zijn voor selectievertekening, ofwel heterogeniteit tot gevolg hebben, of beide. Deze twee aspecten kunnen niet worden gescheiden, aangezien ze allebei de grootte van de objectieffunctie van de MCF bepalen.

4.4.2 Verschillen in de start van de programma’s

Omdat individuen op elk moment in hun werkloosheidsperiode aan een ALMP kunnen worden toegewezen (hoewel ze meestal in het begin worden toegewezen), hebben we te maken met een dynamische toewijzing. In een dergelijke omgeving is de veronderstelling van “geen anticipatie” vereist naast de CIA en is de opbouw van een geschikte vergelijkingsgroep ingewikkeld, zoals Fredriksson en Johansson (2008) voor het eerst erkenden. Geen anticipatie betekent dat individuen hun gedrag niet veranderen naar aanleiding van een toekomstige toewijzing aan een ALMP. Aangezien in de analyseperiode de opleidingscapaciteit de neiging had de vraag te

overstijgen, is de tijd tussen de toewijzing en de daadwerkelijke start van het programma kort, zodat de vertekening die door het falen van deze veronderstelling wordt veroorzaakt, waarschijnlijk klein is.

Om een dynamische programma-toewijzing om te zetten in een statische, worden niet-deelnemers gedefinieerd als de populatie die niet heeft deelgenomen aan het programma binnen een bepaalde periode, in onze studie in de eerste 9 maanden. Fredriksson en Johansson (2008) leggen uit dat een dergelijke definitie de inschatting van de effecten naar beneden toe vertekent, omdat niet-deelnemers minder geneigd zijn om aan een programma deel te nemen, omdat ze misschien al een baan hebben gevonden. Om deze vertekening te voorkomen, stellen zij voor om de vergelijkingsgroep te definiëren als degenen die nog niet zijn behandeld. Op basis van deze inzichten zijn in de literatuur twee benaderingen ontwikkeld. Een eerste benadering is gericht op het identificeren van de effecten van degenen die nog niet zijn behandeld (bijvoorbeeld Sianesi 2004, 2008 en Biewen, Fitzenberger, Osikominu en Paul 2014). Nadeel van deze aanpak is dat het effect opnieuw wordt gedefinieerd en afhankelijk wordt gemaakt van het deel van de niet-deelnemers dat (kort na deze periode) deelneemt. Een andere benadering binnen de literatuur is dan ook gericht op het identificeren van het effect ten opzichte van het nooit ontvangen van de behandeling. Dit gebeurt in essentie door het rechts censureren van niet-deelnemers die vervolgens aan het programma deelnemen: informatie m.b.t. deze personen wordt gebruikt tot op het moment van deelname, vanaf dat punt wordt alle latere informatie gecensureerd. Fredriksson en Johanson (2008) gaan uit van een onafhankelijke rechtscensuur, terwijl Crépon, Ferracci, Jolivet en van den Berg (2009) en Vikström (2017) dit veralgemenen door selectieve rechtscensuur toe te staan. Van den Berg en Vikström (2019) houden rekening met de effecten op de lange termijn van de nabehandeling, zoals die van de beroepsprogramma's op de inkomsten.

Het identificeren van de effecten ten opzichte van het nooit ondergaan van de behandeling met CML-methoden valt buiten het bestek van deze studie. We volgen hier de eerste benadering van de literatuur. Daarbij volgen we de aanpak die is voorgesteld door Lechner, Miquel en Wunsch (2011), die de door Lechner voorgestelde aanpak (1999, 2002) aangepast hebben aan de kritiek van Fredriksson en Johansson (2008). In plaats van de log van de verstreken tijd tot de start aan het programma door de deelnemers te regresseren op een selectie van de beschikbare verklarende variabelen die belangrijk lijken voor de timing van het programma, gebruiken we een post-LASSO-schatter (d.w.z. OLS met de door een LASSO-schatting geselecteerde variabelen) om de relevante variabelen en de coëfficiënten van deze regressie te bepalen. Vervolgens gebruiken we de geschatte coëfficiënten samen met een trekking uit de restverdeling om de start van het 'pseudo'-programma voor niet-deelnemers te voorspellen. De onderliggende veronderstelling is dus dat de toewijzing van de startdata van het programma willekeurig is, afhankelijk van de variabelen die in de post-LASSO-procedure zijn opgenomen. We sluiten de niet-deelnemers uit voor wie deze gesimuleerde startdatum buiten het 9 maanden durende behandelingsvenster ligt en - om rekening te houden met de kritiek - degenen die niet langer werkloos zijn op de toegewezen startdatum.

5 ONDERZOEKSRESULTATEN

In dit hoofdstuk bespreken we de belangrijkste resultaten. We beginnen met het bekijken van de gemiddelde populatie-effecten voor verschillende beleidsrelevante uitkomsten en de evolutie daarvan in de tijd. Dit informeert ons over de algemene effectiviteit van de verschillende programma's en de dynamiek van de effecten. Vervolgens onderzoeken we of de gemiddelde populatie-effecten (ATE) verschillen van de effecten van de werklozen in een bepaald programma (ATET). Deze vergelijkingen zijn informatief om de effecten van de selectie van de werklozen tot op zekere hoogte te begrijpen. Als bemiddelaars programma's selecteren die het meest effectief zijn voor hun specifieke werkloze, dan zou ATET groter moeten zijn dan ATE.

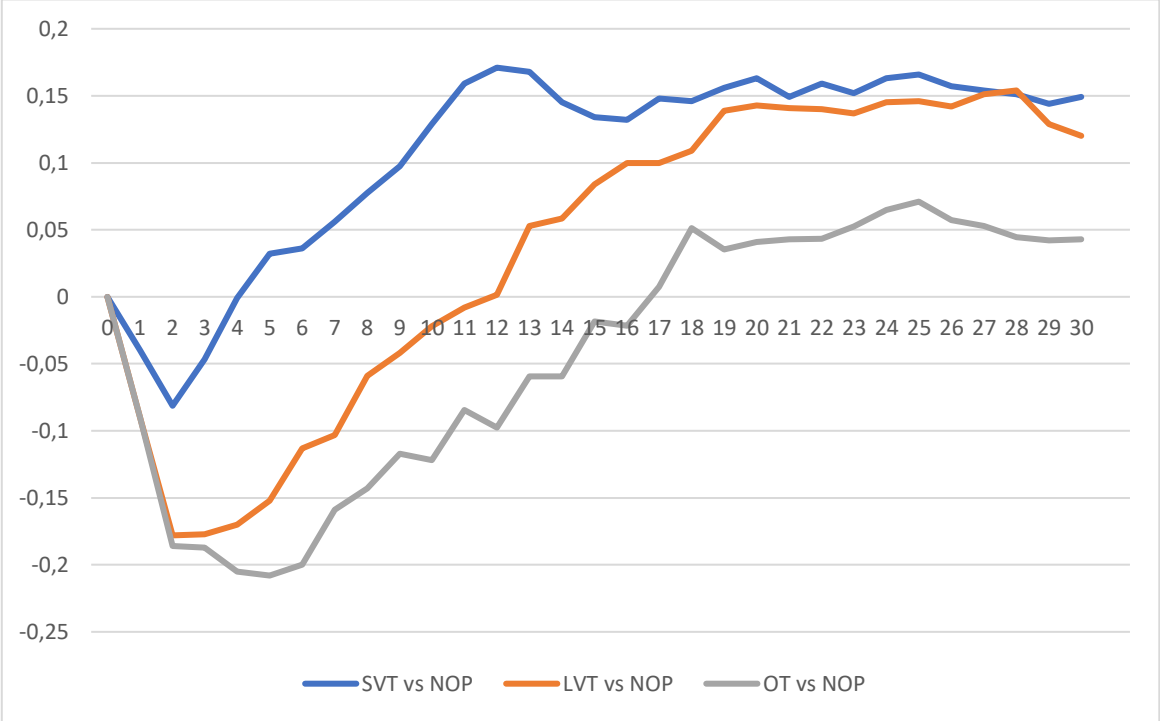
Voor het wellicht meest belangrijkste resultaat op korte en middellange termijn, namelijk de werkgelegenheid, onderzoeken we vervolgens de heterogeniteit met betrekking tot de programma's. Vervolgens wordt de heterogeniteit van een aantal middellange termijn effecten onderzocht met betrekking tot enkele variabelen die van belang worden geacht voor het beleid. Tot slot presenteren we in de laatste paragraaf een analyse van de IATE's, d.w.z. de effectschattingen op het laagst mogelijke niveau van granulariteit.

5.1 GEMIDDELDE POPULATIE-EFFECTEN

5.1.1 Dynamiek en heterogeniteit van de programma's

In deze paragraaf rapporteren we de gemiddelde populatie-effecten (ATE) van de verschillende programma's in vergelijking met geen ALMP-participatie (NOP) en met elkaar.

Figuur 5.1: De evolutie over de tijd van ATEs m.b.t. de kans op werk 2 tot 30 maand na de start van het programma



In figuur 5.1 wordt voor de verschillende programma's de evolutie over 30 maanden bestudeerd van het effect op de kans op werk, steeds in vergelijking met geen deelname aan een ALMP (NOP). Deelname aan kortlopende beroepsopleidingen (SVT) vermindert de kans op een baan alleen in de eerste vier maanden met maximaal 8 procentpunten (pp) ten opzichte van de counterfactual van geen deelname (NOP). Daarna is de werkgelegenheidswinst positief. Het lock-in-effect duurt ongeveer even lang als de gemiddelde duur van het programma zelf, wat erop wijst dat de deelname aan het programma na afloop van het programma gepaard gaat met een toename van de kans op het vinden van werk. Vervolgens blijft het ATE m.b.t. de kans op werk stijgen tot ongeveer 12 maanden na de start van het programma. Daarna stabiliseert het zich tot ongeveer 15 pp, wat een substantieel effect is, vooral als het stabiel blijft in de tijd, zoals uit het dynamische patroon zou kunnen worden afgeleid.

Deelname aan langdurige beroepsopleiding (LVT) leidt tot een kans op werk die in de eerste twee maanden veel sterker afneemt, waardoor deze tot 18 pp daalt ten opzichte van de counterfactual van niet deelname (NOP). Dit vloeit natuurlijk voort uit de langere duur van het programma (gemiddeld 7,2 maanden), maar de lock-in periode duurt nog langer, tot ongeveer een jaar na de start van het programma. Het uiteindelijke effect van deelname aan het programma is vergelijkbaar met dat van SVT, dat wil zeggen ongeveer 15 pp, maar het wordt pas na ongeveer 20 maanden bereikt. Dit betekent dat de langere tijdsinvestering in de toename van menselijk kapitaal niet wordt weerspiegeld door een hogere kans op werk. Het is mogelijk dat de hogere tijdsinvestering van LVT leidt tot hogere productiviteit en/of looneffecten, maar omdat er hierover geen gegevens beschikbaar zijn, kon dit niet worden getest. Gemiddeld genomen lijkt SVT dus te domineren, aangezien de opleidingen goedkoper zijn en de indirecte kosten (lock-in-periode) ook lager zijn.

De negatieve effecten tijdens de lock-in periode van de oriëntatietraining (OT) zijn nog meer uitgesproken omdat het effect in termen van de kans op werk zelfs afneemt tot min 21 pp, en het duurt 17 maanden voordat het positief wordt. Dit lange lock-in effect is vermoedelijk gerelateerd aan de 45% van de OT-deelnemers die binnen 3 maanden na afronding van de OT aan andere programma's deelnemen. OT (inclusief de vervolgprogramma's)

is echter ook minder effectief op de middellange termijn, omdat het effect zich stabiliseert rond 5 pp, 10 pp onder het niveau van VT. Concluderend kan worden gesteld dat SVT gemiddeld genomen de LVT domineert in termen van effectiviteit, dat op zijn beurt de OT domineert.

Om een beter algemeen beeld van de effecten te krijgen, onderzoeken we (i) drie samenvattende maatregelen van de werkgelegenheidseffecten (samengevat over de eerste en de laatste 9 maanden en over alle 30 maanden), en (ii) twee alternatieve uitkomstmaten (maanden in werkloosheid, maanden buiten-de-arbeidsmarkt).

Tabel 5.1: Effecten van de verschillende programma's op het cumulatief aantal maanden werk, werkloosheid en buiten-de-arbeidsmarkt (ATE), tot 30 maanden na de start van de programma's

	Geen deelname (NOP)	Korte beroepsopleiding (SVT)	Lange beroepsopleding (LVT)	Orientation training (OT)
Cumulatief # maanden *werk gedurende 9 maanden* na de start van het programma				
NOP	**3.5 (0.0)**			
SVT	0.1 (0.2)	**3.6 (0.2)**		
LVT	-1.1 (0.1) ***	-1.2 (0.2) ***	**2.4 (0.1)**	
OT	-1.6 (0.1) ***	-1.6 (0.2) ***	-0.4 (0.2) **	**2.0 (0.1)**
Cumulatief # maanden *werk tussen maand 22 en maand 30* na de start van het programma				
NOP	**5.7 (0.0)**			
SVT	1.4 (0.2) ***	**7.1 (0.2)**		
LVT	1.3 (0.2) ***	-0.1 (0.3)	**7.0 (0.2)**	
OT	0.4 (0.2) **	-0.9 (0.3) ***	-0.8 (0.3) ***	**6.2 (0.2)**
Cumulatief # maanden *werk gedurende 30 maanden* na de start van het programma				
NOP	**16.0 (0.1)**			
SVT	3.4 (0.5) ***	**19.4 (0.5)**		
LVT	1.0 (0.5) **	-2.4 (0.7) ***	**17.1 (0.5)**	
OT	-1.4 (0.5) ***	-4.8 (0.7) ***	-2.4 (0.7) ***	**14.7 (0.5)**
Cumulatief # maanden *werkloos gedurende 30 maanden* na de start van het programma				
NOP	**10.9 (0.1)**			
SVT	-1.9 (0.4) ***	**9.0 (0.3)**		
LVT	0.9 (0.4) **	2.8 (0.5) ***	**11.8 (0.4)**	
OT	2.7 (0.5) ***	4.5 (0.6) ***	1.8 (0.6) ***	**13.6 (0.5)**
Cumulatief # maanden *buiten-de-arbeidsmarkt gedurende 30 maanden* na de start van het programma				
NOP	**3.1 (0.1)**			
SVT	-1.6 (0.3) ***	**1.6 (0.3)**		
LVT	-1.8 (0.3) ***	-0.4 (0.3)	**1.1 (0.3)**	
OT	-1.4 (0.3) ***	0.2 (0.4)	0.7 (0.4)	**1.7 (0.2)**

Noot: Uitkomsten worden gemeten in maanden. Niveau van potentiële resultaten voor het specifieke programma op de hoofddiagonaal in vetgedrukt. Alle effecten zijn populatiegemiddelden (ATE). Standaardfouten staan tussen haakjes. *, **, *** geven de nauwkeurigheid van de schatting aan door aan te geven of de p-waarde van een tweezijdige significancetest respectievelijk lager is dan 10%, 5%, 1%.

Uit het eerste luik van tabel 5.1 blijkt dat SVT na 9 maanden tot hetzelfde aantal maanden werk leidt als NOP (3,5), terwijl LVT en OT te kampen hebben met aanzienlijke gemiddelde verliezen van respectievelijk 1,1 en 1,6 maanden. De cumulatieve effecten in de laatste 9 maanden van het waarnemingsvenster (maand 22 tot 30) zijn allemaal positief. Ze zijn het grootst en vergelijkbaar voor SVT en LVT (1,3-1,4) en merkelijk kleiner bij OT (0,4). Deze statistisch significante resultaten worden ook bevestigd wanneer de effecten van de verschillende programma's rechtstreeks met elkaar worden vergeleken. Het derde luik vat deze werkeffecten over alle 30 maanden samen. Terwijl SVT (3.4) en LVT (1.0) positieve effecten hebben, is het effect van OT negatief (-1.4), vanwege de grote lock-in component. De laatste twee luiken van tabel 5.1 maken melding van het gemiddelde totale effect van de verschillende programma's op de tijd die wordt besteed in werkloosheid en buiten-de-arbeidsmarkt gedurende 30 maanden na de start van het programma. Terwijl alle programma's de tijd buiten de arbeidsmarkt met ongeveer anderhalve maand verminderen, vermindert alleen SVT ook de tijd in werkloosheid

(met 1,9 maanden). LVT verlengt het verblijf in de werkloosheid met bijna 1 maand, terwijl OT het verblijf in werkloosheid met bijna 3 maanden verhoogt. Ook dit zijn in ieder geval gedeeltelijk de gevolgen van de verschillende lock-in effecten.

Op dit punt aangekomen, grijpen we terug naar oudere resultaten (met een meting tot 23 maanden na de start van de programma's), omdat daar nog resultaten te vinden zijn m.b.t. TIBB. Tabel 5.2 geeft de resultaten.

Tabel 5.2: Effecten van de verschillende programma's op het cumulatief aantal maanden werk, werkloosheid en buiten-de-arbeidsmarkt (ATE), tot 23 maanden na de start van de programma's

	Geen deelname (NOP)	Korte beroepsopleiding (SVT)	Lange beroepsopleiding (LVT)	Orientation training (OT)	TIBB (TIBB)
Cumulatief # maanden werk gedurende 8 maanden na de start					
NOP	***2.88(0.1)***				
SVT	-0.01 (0.1)	***2.87 (0.1)***			
LVT	-1.18 (0.1)	-1.16 (0.2)	***1.70 (0.1)***		
OT	-1.29 (0.1)	-1.27 (0.2)	-0.11 (0.1)	***1.60 (0.1)***	
TIBB	0.24 (0.1)	0.25 (0.1)	1.41 (0.1)	1.52 (0.1)	***3.12 (0.1)***
Cumulatief # maanden werk gedurende 23 maanden na de start					
NOP	***10.9 (0.1)***				
SVT	2.12 (0.4)	***13.0 (0.3)***			
LVT	0.08 (0.3)	-2.03 (0.5)	***11.0 (0.3)***		
OT	-1.21 (0.3)	-3.33 (0.5)	-1.30 (0.5)	9.71 (0.3)	
TIBB	1.08 (0.2)	-1.04 (0.4)	0.99 (0.4)	2.29 (0.4)	***12.0 (0.2)***
Cumulatief # maanden werkloos gedurende 23 maanden na de start					
NOP	***9.94 (0.1)***				
SVT	-0.77 (0.3)	***9.17 (0.3)***			
LVT	1.01 (0.4)	1.78 (0.4)	***10.95 (0.3)***		
OT	2.30 (0.4)	3.07 (0.4)	1.29 (0.5)	***12.2 (0.4)***	
TIBB	-0.19 (0.)	0.58 (0.3)	-1.20 (04)	-2.49 (0.4)	***9.75 (0.2)***
Cumulatief # aantal maanden buiten-de-arbeidsmarkt gedurende 23 maanden na de start					
NOP	***2.12 (0.1)***				
SVT	-1.23 (0.1)	***0.89 (0.1)***			
LVT	-1.02 (0.2)	0.21 (0.2)	***1.10 (0.2)***		
OT	-1.02 (0.2)	0.20 (0.2)	0 (0.2)	***1.09 (0.2)***	
TIBB	-0.80 (0.1)	0.43 (0.2)	0.22 (0.2)	0.23 (0.1)	***1.32 (0.1)***

Noot: Uitkomsten worden gemeten in maanden. Niveau van potentiële resultaten voor het specifieke programma op de hoofddiagonaal in vetgedrukt. Alle effecten zijn populatiegemiddelden (ATE). Standaard fouten staan tussen haakjes.

De (evoluties van de) resultaten voor TIBB zijn interessant: op korte termijn (eerste 8 maanden) heeft TIBB als enige maatregel een beduidend positief treatment effect. Na verloop van tijd (zie resultaten na 23 maanden) wordt TIBB echter ingehaald en voorbijgestoken door competentieversterking SVT, al blijft het effect van TIBB ook daar nog beduidend positief. Een zeer logisch resultaat: van een intensieve begeleiding en bemiddeling kan relatief snel een resultaat verwacht worden. Van opleiding daarentegen, weten we dat het meestal een wat langere aanlooptijd nodig heeft voor er een effect zichtbaar wordt, maar het kan dan ook gedurende een langere tijd aanhouden, het is meer duurzaam. Zo vinden Card, Kluve en Weber (2018) in hun metastudie dat over het algemeen programma's met meer accumulatie van menselijk kapitaal (d.w.z. opleiding) effectiever zijn op de langere termijn.

5.1.2 Heterogeniteit m.b.t. het programma

Terwijl we in tabel 5.1 (en 5.2) de effecten voor de volledige werklozenpopulatie onderzochten (ATE), analyseren we nu hoe de effecten verschillen voor de verschillende populaties die effectief aan de verschillende programma's deelnemen (ATET). Stel dat bemiddelaars een werkzoekende naar dat programma sturen waar

voor hem of voor haar de grootste effectiviteitswinst te verwachten is, dat kunnen we verwachten dat de effecten voor de eigen deelnemers (bijvoorbeeld de effecten van SVT voor de deelnemers aan SVT) groter zullen zijn dan het populatiegemiddelde (ATE) dat we tot dusver bekeken. Er blijkt echter dat dit in het algemeen niet het geval is. In tabel 5.3 tonen we formele statistische (Wald-) testen voor de gelijkheid van de effecten over de vier populaties. We zien dat er van de 30 testen slechts één duidelijke afwijzing is op conventionele significantieniveaus. En dat is dan bovendien omdat in dat geval het ATET lager is dan het ATE.

Tabel 5.3: Wald test voor gelijkheid van de effecten in ieder van de 4 specifieke subpopulaties

Uitkomst	SVT – NOP	LVT – NOP	OT – NOP	LVT – SVT	OT – SVT	OT – LVT
Cumulatief # maanden werk 0-9 maanden na start	3.6	6.1	16.5***	0.7	1.6	2.7
Cumulatief # maanden werk *21-30 maanden* na start	1.9	2.8	0.3	0.9	0.6	1.1
Cumulatief # maanden werk 0-30 maanden na start	3.2	3.2	2.5	0.6	0.4	0.1
Cumulatief # maanden werkloos 0-30 maanden na start	4.6	3.0	2.8	0.1	0.2	0.1
Cumulatief # maanden buiten de arbeidsmarkt 0-30 maanden na start	4.9	7.3*	6.7*	2.2	0.5	1.0

Noot: Bij een nulhypothese van gelijkheid, is de test statistiek $\chi^2(3)$ verdeeld. *, **, *** verwijzen naar respectievelijk een p-waarde beneden 10%, 5% en 1%.

Dit betekent dat ofwel de behandelingseffecten voor deze programma's vrij homogeen zijn, ofwel dat de toewijzing van medewerkers aan de verschillende programma's niet beter dan random is. Hieronder zullen we laten zien dat de effecten duidelijk heterogeen zijn, zodat we kunnen concluderen dat bemiddelaars de werklozen niet toewijzen aan de programma's waarvan ze het meest zouden profiteren. Verder zullen we bespreken welke voordelen VDAB zou kunnen behalen door te werken met een meer optimale toewijzing.

5.2 HETEROGENITEIT MET BETREKKING TOT BELEIDSRELEVANTE VARIABELEN

Er zijn diverse heterogeniteitsvariabelen waar beleidsmakers belang aan hechten. In dit hoofdstuk analyseren we dergelijke variabelen aan de hand van de hierboven geïntroduceerde GATE-parameter. We presenteren de resultaten voor de totale populatie, aangezien de programma-populatie-specifieke effecten niet veel lijken af te wijken van de populatiegemiddelden. We richten ons opnieuw op het belangrijkste resultaat op de middellange termijn: het cumulatief aantal maanden werk in de 30 maanden na de start van het programma. Uiteraard brengt het a priori specificeren van een lange lijst van beleidsrelevante variabelen en het rapporteren van significante resultaten het gevaar met zich mee op p-hacking. We gaan er hier van uit dat de Vlaamse arbeidsmarktautoriteiten de volgende variabelen als bijzonder belangrijk beschouwen: werkloosheidsverleden (laatste 2 en 10 jaar), werkloosheidsduur bij aanvang van het programma (onder of boven de mediaan), leeftijd (jonger dan 25 of ouder dan 50 jaar, onder of boven de mediaan), geslacht, beheersing van de Nederlandse taal, werkloosheidspercentage in het arrondissement van de woonplaats bij aanvang van de werkloosheidsperiode, geboorteland (6 groepen): België, zuidelijke EU-landen, oostelijke EU-landen, andere EU-landen, Turkije of Marokko, de rest van de wereld) en 13 onderwijsniveaus (vanaf het tweede jaar van de middelbare school of lager tot aan de masteropleiding). Deze keuzes zijn gebaseerd hypotheses die we hier toelichten:

De steekproef voor dit onderzoek werd getrokken uit personen die werkloos zijn geworden nadat ze hun baan hadden verloren. De recente werkloosheidsgeschiedenis helpt dus bij het identificeren van een subpopulatie die minder sterk verbonden is met de arbeidsmarkt in die zin dat zij personen kan identificeren die niet alleen werk hebben gehad in de afgelopen twee jaar, maar ook enige tijd in de werkloosheid hebben gezeten. Vanuit een beleidsperspectief kan het interessant zijn om programma's te identificeren die voor een dergelijke populatie effectief zijn. A priori zou men kunnen verwachten dat het aanbieden van beroepsopleidingen de competenties van deze groep kan versterken en een stabielere werksituatie kan opleveren. Het kan echter nuttig zijn om na te gaan of deze hypothese geldt.

De geschiedenis van de werkloosheid over de laatste 10 jaar, daarentegen, kan bijvoorbeeld helpen bij het identificeren van een groep van werknemers die een stabiele baan hadden, maar die hun baan abrupt kwijtraakten. Dit zou de groep kunnen zijn waarop OT zich richt en het is van belang om te weten of een dergelijke strategie werkt.

In België hebben jongeren en oudere werknemers moeite om een baan te vinden, zodat het interessant is om te weten welk beleid effectief is voor mensen jonger dan 25 of ouder dan 50 jaar.

Discriminatie op het vlak van zowel geslacht als migratieachtergrond (geboorteland en vaardigheid in het Nederlands zijn hiervoor proxies) is een gevoelig politiek thema. In België hebben personen met een migratieachtergrond veel meer moeite dan elders in de EU om werk te vinden (Piton en Rycx, 2020). Het is daarom zeer relevant om vast te stellen welk beleid het beste werkt om dergelijke discriminatie in te dammen en om migranten aan een vaste baan te helpen.

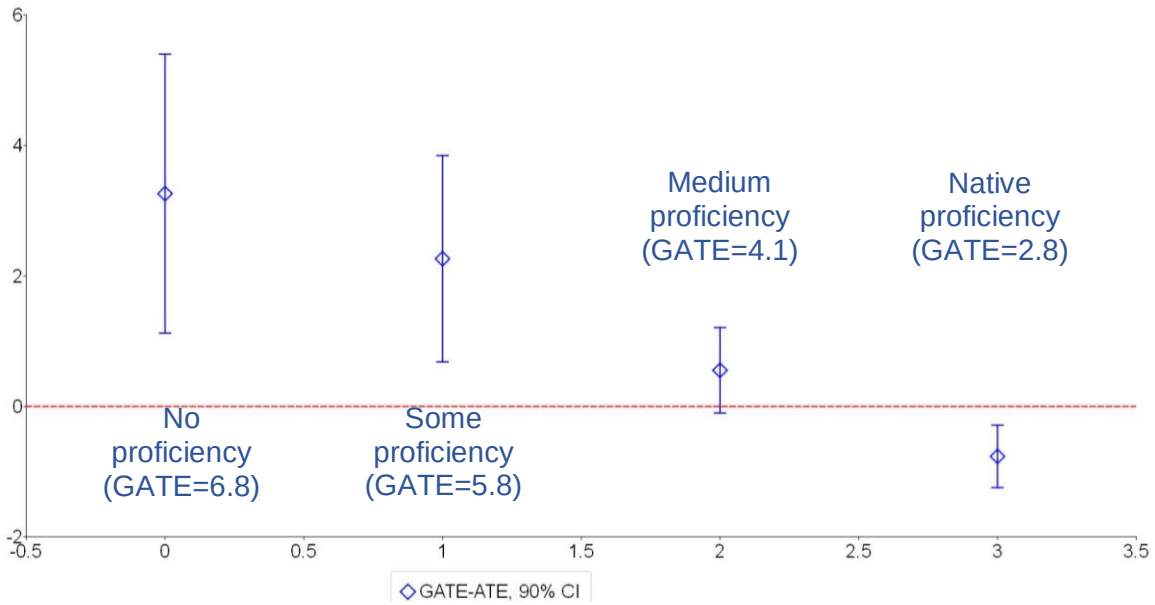
Op de Belgische arbeidsmarkt lopen laagopgeleide werknemers een bijzonder groot risico om werkloos te worden, zodat kennis over de relatieve effectiviteit van het beleid, gegeven het opleidingsniveau van de deelnemers, eveneens waardevol is.

Ondanks het feit dat Vlaanderen een kleine regio is, verschillen de werkloosheidscijfers aanzienlijk tussen de arrondissementen. Dit houdt verband met de beperkte geografische mobiliteit binnen de regio, die onder meer wordt veroorzaakt door een beleid dat het huisbezit sterk ondersteunt en dat verkeersopstoppingen stimuleert.

Ten slotte heeft de doeltreffendheid van de opleidingsprogramma's in functie van de duur van de werkloosheid te maken met de discussie over de vraag of preventieve dan wel curatieve interventies effectiever zijn.

De in dit onderzoek gebruikte machine learning benadering, is bij uitstek geschikt om deze hypothesen te toetsen.

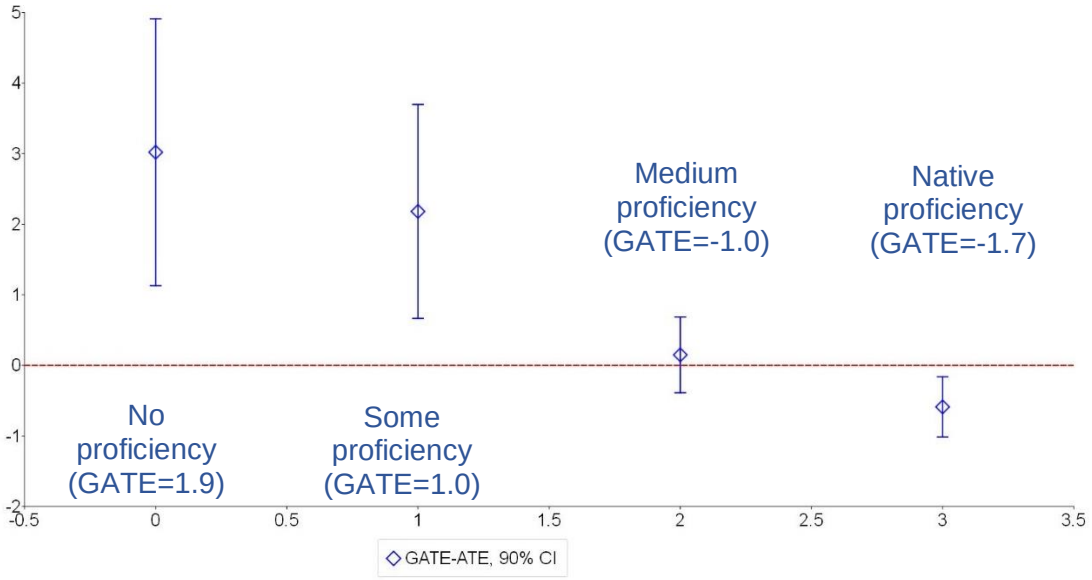
Figuur 5.2: Verschil tussen de GATEs en het ATE van SVT (in vergelijking met NOP) voor de vier niveaus van kennis Nederlands – Cumulatief # maanden werk in de 30 maanden sinds de start



Noot: De kennis van Nederlands staat op de horizontale as. De verticale as geeft het verschil aan tussen de respectievelijke GATE en het ATE. (GATE-ATE), telkens met een 90%-betrouwbaarheidsinterval.

Wanneer we de heterogeniteit van de ATE's nagaan voor de uitkomst "cumulatief aantal maanden werk gedurende 30 maanden na de start van het programma", vinden we statistisch significante verschillen (op 10% niveau) in drie dimensies (taalvaardigheid in het Nederlands, geboorteland en opleidingsniveau), maar niet voor alle programma's. Figuur 5.2 illustreert het verschil tussen de GATE's (minus ATE) van SVT in verhouding tot NOP voor het kennisniveau Nederlands. De horizontale lijn op nul geeft het niveau van het ATE aan. Er is duidelijk een afname van de GATE's naarmate niveau in de kennis van het Nederlands toeneemt. Bovendien hebben de lagere beheersingsniveaus significant hogere GATE's dan het ATE. Zo is het GATE van degenen die geen kennis van het Nederlands hebben 3,7 maanden hoger dan het ATE (p-waarde van 2%).

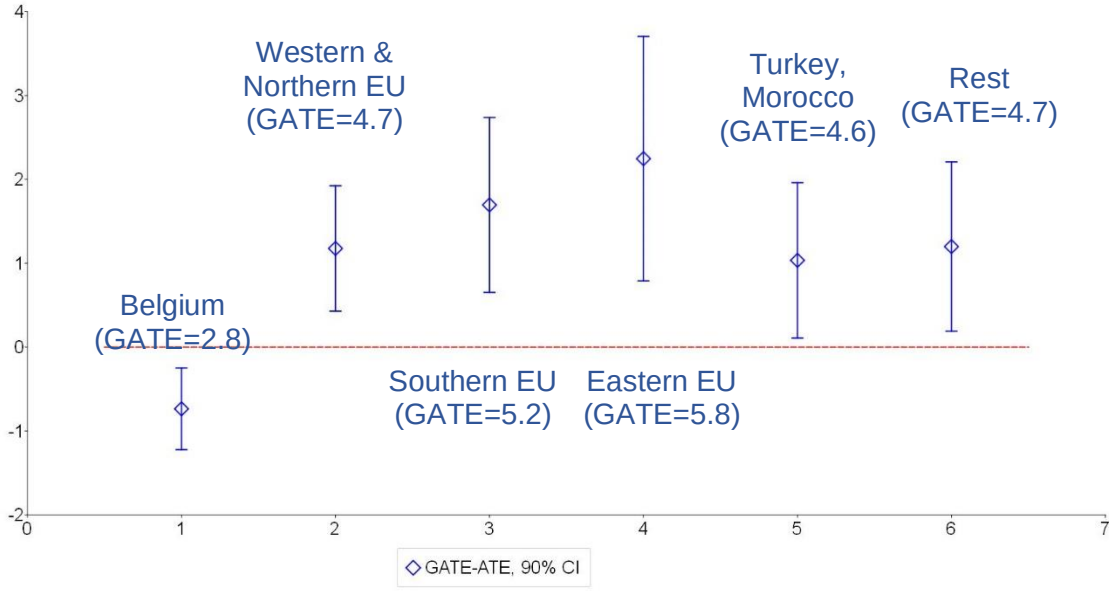
Figuur 5.3: Verschil tussen de *GATEs* en het *ATE* van OT (in vergelijking met *NOP*) voor de vier niveaus van kennis Nederlands – *Cumulatief # maanden werk in de 30 maanden sinds de start*



Noot: De kennis van Nederlands staat op de horizontale as. De verticale as geeft het verschil aan tussen de respectievelijke GATE en het ATE. (GATE-ATE), telkens met een 90%-betrouwbaarheidsinterval.

Figuur 5.3 geeft de GATE's van OT versus NOP weer. Interessant is dat de puntschattingen van de GATE's voor de twee laagste kennisniveaus levels positief zijn, terwijl de puntschatting van het ATE negatief was. Hoewel deze GATE's statistisch niet significant verschillen van nul, zijn ze statistisch wel significant verschillend van het ATE, op het 3%-niveau voor kennisniveau nul en op het 8%-niveau voor kennisniveau één. De puntschattingen van de GATE's van LVT versus NOP (geen figuur opgenomen in de tekst) vertonen eveneens een vergelijkbaar negatief verband met het kennisniveau Nederlands. Geen van deze verschillen is echter significant verschillend van het ATE.

Figuur 5.4: Verschil tussen de *GATEs* en het *ATE* van *SVT* (in vergelijking met *NOP*) volgens land van herkomst – *Cumulatief # maanden werk in de 30 maanden sinds de start*



Note: De regio van herkomst staat op de horizontale as. De verticale as geeft het verschil aan tussen de respectievelijke GATE en het ATE. (GATE-ATE), telkens met een 90%-betrouwbaarheidsinterval.

In figuur 5.4 wordt geïllustreerd hoe de GATE per geboorteland verschilt. De figuur suggereert dat de GATE's van SVT ten opzichte van NOP het hoogst zijn voor personen die in zuidelijke landen van de Europese Unie zijn geboren (6,8 maanden). Opvallend is dat de effecten voor personen die in Turkije en Marokko geboren zijn, d.w.z. voor wie de tewerkstellingsgraad lager is dan voor andere buitenlanders, aanzienlijk hoger blijven (5,1 maanden) dan voor Belgen (2,5 maanden). Ook al is de nauwkeurigheid lager en vinden we geen statistisch significante verschillen bij de andere ALMP's, het patroon van de overeenkomstige GATE's is gelijkaardig en wordt dus niet gerapporteerd. Als bijkomend bewijs hebben we de GATE's vergeleken voor het al dan niet in België geboren zijn. Voor deelnemers aan SVT die buiten België geboren zijn, is de GATE van SVT ten opzichte van NOP 5,8 maanden tegenover 2,1 maanden voor diegenen die in België geboren zijn (significant verschillend aan 5%). Samen met de vorige bevinding over kennis van het Nederlands wijst dit er sterk op dat SVT effectiever is voor migranten die recentelijk naar België zijn gemigreerd.

Tot nu toe hebben we gewerkt met GATE's voor het cumulatief aantal maanden werk gedurende 2,5 jaar na de start van het programma. De boven vastgestelde heterogeniteit in de effecten, vermengt twee bronnen van heterogeniteit: één tijdens de lock-in fase en een ander tijdens de nabehandelingsperiode. Zowel in de studie van Knaus, Lechner en Strittmatter (2017) als in de studie van Bertrand, Crépon, Marguerie en Premand (2017) wordt de heterogeniteit van de effecten vooral gevonden tijdens de lock-in fase en niet zozeer na de behandeling. In onze evaluatie bevestigen we dat heterogeniteit belangrijker is in deze beginfase, maar vinden we ook aanwijzingen voor heterogeniteit in het nabehandelingseffect.

Tijdens de lock-in fase wordt heterogeniteit voornamelijk veroorzaakt door het verschil in de snelheid waarmee verschillende soorten werklozen in een toestand zonder deelname (NOP) werk vinden. Voor werklozen met een lage inzetbaarheid zal lock-in typisch minder spelen, omdat deze personen ook zonder deelname aan het programma weinig kans hebben om aan het werk te gaan. Om een idee te krijgen over de heterogeniteit tijdens de lock-in fase, bekijken we de GATE voor het cumulatief aantal maanden werk in de 9 maanden na de start van het programma. Op die manier vinden we inderdaad aanwijzingen voor heterogeniteit in meer dimensies dan de boven vernoemde (die betrekking hadden op het benchmarkresultaat m.b.t. de periode van 30 maanden na de start van het programma). Zo zijn in de eerste 9 maanden alle programma's significant minder effectief voor jongeren onder de 25 jaar en effectiever voor oudere werknemers boven de 50 jaar, waarbij de effectiviteit over het algemeen toeneemt met de leeftijd, zijn de programma's effectiever voor degenen die in een stad wonen,

effectiever voor langdurig werklozen (met uitzondering van SVT), en minder effectief voor degenen met veel werkloosheidservaring in de afgelopen twee jaar.

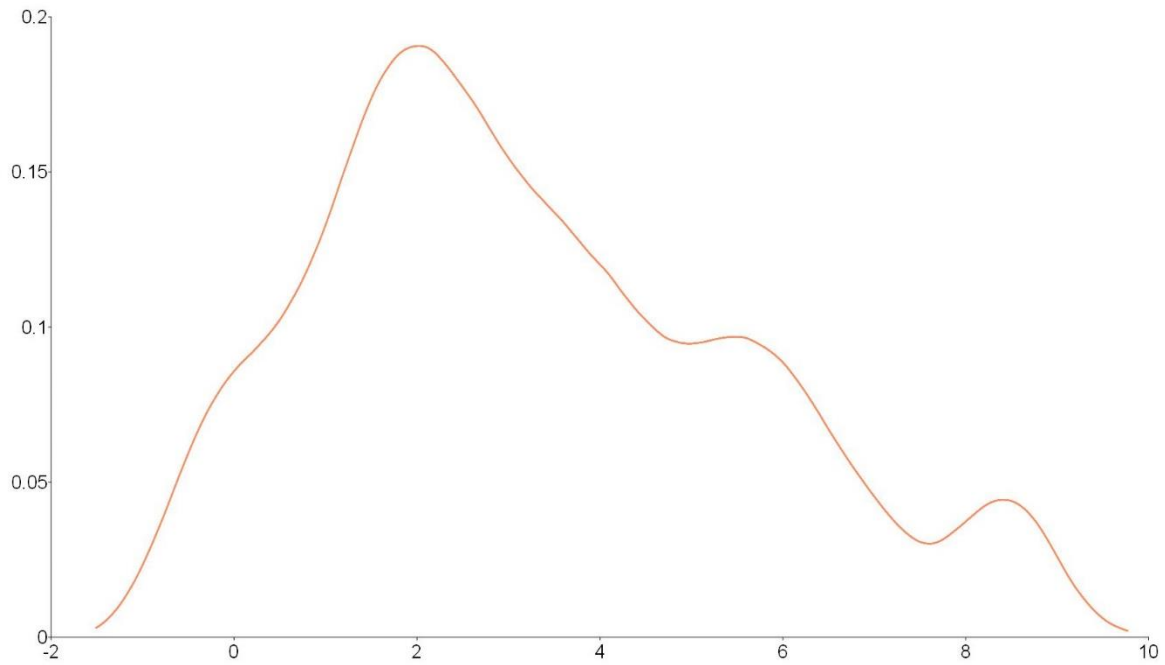
Om de heterogeniteit in de nabehandeling te evalueren, houden we rekening met het cumulatieve werkgelegenheidsresultaat tussen de maanden 22 en 30 na de start van het programma. Dit is het moment waarop de effecten van het programma op de lange termijn in evenwicht zijn (zie figuur 5.1). In dit geval blijft de heterogeniteit in de volgende dimensies bestaan: hogere effectiviteit voor mensen met een lager niveau van kennis Nederlands, voor mensen die in het buitenland zijn geboren, in het bijzonder voor mensen die in een zuidelijk of oostelijk EU-land zijn geboren. Voor het onderwijs vertonen de puntschattingen vergelijkbare verschillen als voor het benchmarkresultaat, maar ze zijn niet meer statistisch significant.

5.3 HETEROGENITEIT OP INDIVIDUEEL NIVEAU (IATE'S)

In deze sectie presenteren we de resultaten voor de geïndividualiseerde gemiddelde effecten (IATE's), die het laagste niveau van de beschikbare granulariteit vertegenwoordigen. We richten de discussie hier op de cumulatieve werkresultaten op de middellange termijn, relatief t.o.v. NOP (geen deelname aan het programma), wat waarschijnlijk het meest beleidsrelevant is. We beschrijven eerst de mate van heterogeniteit in de effecten van het programma. We presenteren vervolgens de resultaten van een k-means-clusteranalyse om een informele beschrijving van de subgroepen te krijgen, geclusterd naar de effectiviteit van de programmadeelname.

Figuur 5.6 toont de verdeling van de IATE's van SVT relatief t.o.v. NOP. 99% van de geschatte effecten zijn positief. Het gemiddelde van deze effecten is 3,4 maand (zoals weergegeven in tabel 5.1) en de standaardafwijking 2.2. Ongeveer 34% van de geschatte IATE's is significant verschillend van nul. Dit wijst op twee belangrijke punten: (i) Er is een aanzienlijke heterogeniteit in de IATE's, ten dele te wijten aan schattingsfouten; (ii) Het is veel moeilijker om een precieze schatting te krijgen (zonder bepaalde functionele veronderstelingen te maken) voor de IATE's dan voor de GATE's en ATE's die met een vrij grote nauwkeurigheid werden geschat. Dit is ook zichtbaar in figuur 5.7, waarin de gesorteerde effecten samen met een 90%-betrouwbaarheidsinterval op basis van de geschatte standaardfouten worden gegeven (zie ook Chernozhukov, Fernandez-Val en Luo 2018). Ook hier zien we een aanzienlijke variatie in de effecten, maar ook dat de onzekerheid van het ATE veel lager is dan bij de IATE's.

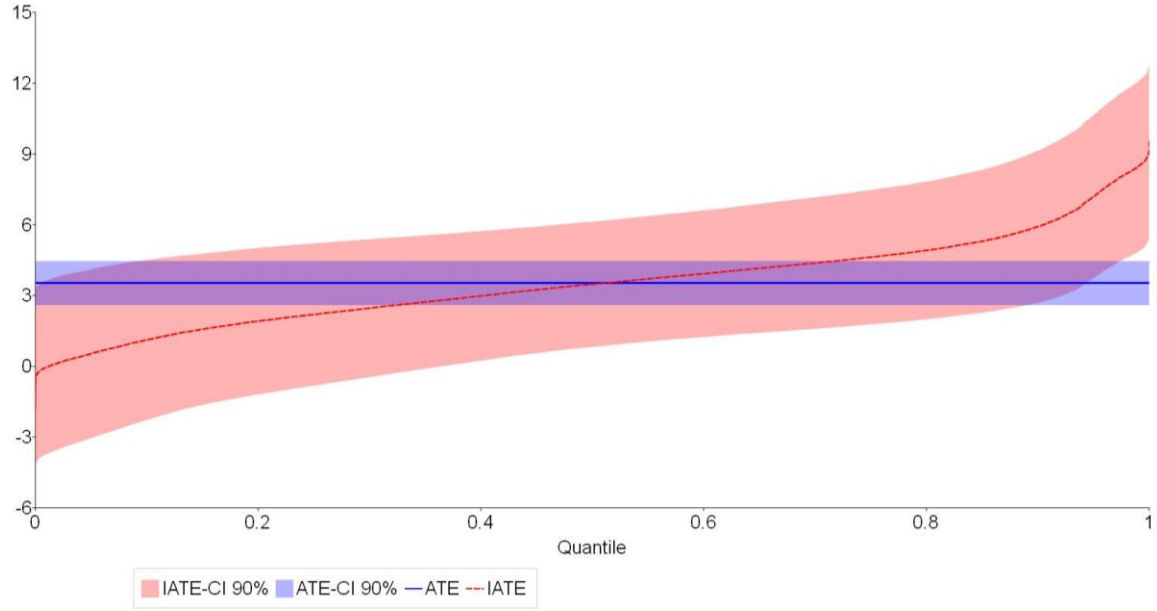
Figuur 5.6: Verdeling van de geschatte IATE's van SVT relatief t.o.v. NOP



Noot: Kernel smooth met Epanechnikov Kernel en Silverman (normality) bandwidth.

De respectievelijke cijfers en de gesorteerde effecten voor de andere programma's m.b.t. het verschil tussen het IATE en het ATE zijn kwalitatief vergelijkbaar met de degene die hier worden gepresenteerd.

Figuur 5.7: Globale heterogeniteit: gesorteerde effecten van SVT relatief t.o.v. NOP – Werk gedurende 30 maanden na de programmastart



Noot: De IATE's zijn gesorteerd volgens hun grootte. Het 90%-betrouwbaarheidsinterval van de IATE's is gebaseerd op de geschatte standaardfouten en de normale verdeling. Standaard fouten worden afgevlakt via een Nadaraya-Watson regressie (Epanechnikov kernel met Silverman bandwidth).

Uit de vorige discussie over de GATE's en de verdeling van de IATE's blijkt dat er sprake is van een aanzienlijke mate van heterogeniteit. In de vorige paragraaf hebben we besproken in hoeverre deze heterogeniteit verband houdt met een aantal beleidsrelevante variabelen. Om verdere patronen van heterogeniteit op dit fijne niveau te kunnen detecteren, is het aan te raden enige bijkomende structuur op te leggen. Daarom doen we een beroep op k-means clustering.

Aanvullend bij de eerdere heterogeniteitsanalyses beschrijven we in wat volgt de afhankelijkheid van de effecten op covariaten, en wel door door k-means++ clustering (Arthur en Vassilvitskii 2007). De clustering wordt geïmplementeerd door de IATE's van de 3 programma-effecten (relatief t.o.v. NOP), gezamenlijk te gebruiken om 8 clusters te vormen. We beperken ons hier tot de effecten op werk, over de 30 maanden na de start van het programma. De resultaten zijn opgenomen in tabel 5.4.

De clustering is in de effectiviteit van alle programma's bijna overal uniform monotoon en de kolommen in tabel 5.4 zijn dienovereenkomstig gerangschikt, van klein naar groot. Uit de analyse blijkt opnieuw de belangrijke heterogeniteit van de programma-effecten. De werkgelegenheidswinst varieert van 1,2 tot 8,0 maanden voor SVT, van -1,7 tot +4,9 maanden voor LVT en van -2,8 tot +1,9 voor OT. De programma's zijn duidelijk het meest effectief voor degenen met een beperkte kennis van het Nederlands, geboren in het buitenland en met zeer weinig recente werkloosheids- en werkervaring, gemiddeld respectievelijk één en vijf maanden in de afgelopen twee jaar. Dit profiel kan alleen maar passen voor personen die recent toetraden tot de arbeidsmarkt. Oostelijke en zuidelijke EU-landen zijn de meest representatieve landen van herkomst binnen deze meest effectieve cluster, maar het is opvallend dat personen uit Turkije en Marokko en uit de rest van de wereld - migranten met zeer slechte arbeidsmarktprestaties in Vlaanderen - het meest vertegenwoordigd zijn in de tweede meest effectieve cluster. Samen met het feit dat de meest effectieve clusters bestaan uit individuen met de minst recente en minder recente arbeids- en werkloosheidservaring (en dat de gemiddelde leeftijd recente schoolverlaters uitsluit), wordt het duidelijk dat de groep waarvoor de programma's het meest effectief zijn, voornamelijk bestaat uit recente migranten. Bovendien is ook het wonen in een stad in plaats van op het platteland geassocieerd met grotere programma-effecten.

De twee minst effectieve groepen zijn autochtonen met een uitstekende beheersing van het Nederlands en relatief veel recente (laatste 2 jaar) en minder recente (laatste 10 jaar) werkervaring. Hun eerste toetreding tot

de arbeidsmarkt was meestal als werkloze werkzoekende. Opvallend is ook dat er geen duidelijk verband bestaat tussen de effectiviteit van het programma enerzijds, en de leeftijd en het geslacht anderzijds.

Ten slotte kan men de inzetbaarheid van de werklozen ook beoordelen op basis van hun geschatte aantal maanden werk zonder het programma (NOP). De laatste rij in tabel 5.4 laat zien dat de effectiviteit van het programma in die 8 groepen monotoon afneemt met de inzetbaarheid, hetgeen in overeenstemming is met het beeld van heterogeniteit dat tot nu toe aan het licht is gekomen.

Tabel 5.4: Beschrijving van de clusters gebaseerd op k-means clustering

Cluster	**Minst voordelig**	**2**	**3**	**4**	**5**	**6**	**7**	**Meest voordelig**
Aandeel van de observaties in %	**20**	16	18	16	10	11	5	**5**
	Gemiddelde							
	IATE relatief t.o.v. geen deelname							
SVT-NOP	**1.2**	1.4	2.4	4.2	5.0	6.0	6.0	**8.0**
LVT-NOP	**-0.1**	-1.7	0.5	1.1	2.7	3.9	2.1	**4.9**
OT-NOP	**-2.6**	-2.8	-1.7	0.0	-2.8	-0.6	2.5	**1.9**
	Selectie van features							
Leeftijd	**32**	28	40	39	36	36	37	**35**
Vrouw (in %)	**74**	0	60	40	61	65	12	**48**
Woont in een stad (in %)	**29**	32	29	28	51	45	48	**47**
Kennis Nederlands (3: hoog, 0: geen)	**3.0**	3.0	2.8	1.9	2.4	2.3	1.1	**0.8**
Geboorteland: België (in %)	**100**	100	95	72	1	0	9	**0**
Geboorteland: Westen en Noorden van EU (in %)	**0**	0	0	1	12	38	12	**14**
Geboorteland: Zuiden van EU (in %)	**0**	0	0	1	0	4	5	**18**
Geboorteland: Oosten van EU (in %)	**0**	0	0	0	1	24	4	**58**
Geboorteland: Turkije & Marokko (in %)	**0**	0	0	6	0	16	42	**7**
Geboorteland: Rest van de wereld (in %)	**0**	0	5	18	86	18	28	**3**
BIT (werlkoos bij eerste arbeidsmarktintrede; in %)	**52**	72	23	18	27	16	5	**0**
# maanden werkloos in de laatste *10 jaren*	**14**	19	24	18	29	15	13	**1.6**
# maanden werk in de laatste *10 jaren*	**57**	49	56	60	52	37	23	**7**
# maanden werkloos in de laatste *2 jaren*	**4**	4	5	4	5	3	3	**1**
# maanden werk in de laatste *2 jaren*	**18**	17	16	17	17	14	10	**5**
# dagen tot programma-start sinds begin werkloosheid	**75**	84	109	96	97	106	98	**108**
Voorspelde uitkomst zonder deelname (NOP)	**19**	18	17	15	13	13	13	**13**

Noot: De uitkomst is de cumulatieve tewerkstelling over 30 maanden na de start van het programma. Alle IATE's voor alle vergelijkingen met niet-deelname worden gebruikt om de 8 clusters te vormen. Covariaten worden niet gebruikt om clusters te vormen. Hiervoor werd het K-means ++ algoritme gebruikt (Vassilvitskii, 2007).

6 BELEIDSSIMULATIES

Tot nu toe hebben we een aanzienlijke heterogeniteit in de effectiviteit van de programma's gevonden. Wordt deze heterogeniteit door bemiddelaars gebruikt om de werklozen toe te wijzen aan de programma's die het best voor hen werken, en zo niet, in hoeverre zou dan een andere allocatie de prestaties van de VDAB kunnen verbeteren? Om deze vragen te beantwoorden, worden diverse hypothetische programma-allocaties gesimuleerd, die we vergelijken met de waargenomen allocatie. Voor het genereren van de nieuwe allocaties, gebruiken we twee benaderingen. De eerste is een "Black-Box" benadering die gebaseerd is op het individueel toewijzen van de behandeling met het hoogste geschatte potentiële resultaat. Een dergelijke aanpak in praktijk brengen, bvb. door de aanbevolen allocatie via een AI-systeem ter beschikking te stellen van de bemiddelaars, kan echter op weerstand stuiten. Omdat van hen niet kan worden verwacht dat ze begrijpen hoe de regels tot stand zijn gekomen, zullen ze de aanbeveling niet altijd vertrouwen, en er zich dan ook niet aan houden.

Onlangs hebben Zhou, Athey en Wager (2019) deze leemte opgevuld door een methode voor te stellen die geldig is in de context van meerdere behandelingen en die het mogelijk maakt om beleidsregels af te leiden op basis van beslissingsbomen (“decision trees”). Wanneer deze methode wordt beperkt tot ondiepe bomen met een klein aantal knopen, zal de resulterende regel eenvoudig en gemakkelijk te begrijpen en te implementeren zijn. Misschien zullen bemiddelaars er zich dan wel aan houden, althans als ze het er niet te sterk mee oneens zijn.

Een potentieel nadeel is echter dat een ondiepe boom misschien niet het optimale beleid benadert, in tegenstelling tot een Black-Box AI recommendersysteem.

De twee eerste regels van tabel 6.1 bevatten de feitelijk waargenomen allocatie en een gerandomiseerde allocatie. De volgende 5 regels geven de Black-Box regels weer, terwijl het onderste deel van de tabel de allocaties op basis van beslissingsbomen met een diepte van drie en van vier niveaus weergeven. De eerste kolom van deze tabel bevat de beschrijving van de toewijzing, gevolgd door het aandeel van de bevolking dat aan de verschillende programma's is toegewezen (in %), en de gemiddelde effecten van het beleid op de belangrijkste resultaten, namelijk het gemiddelde aantal maanden werk (Emp), het gemiddelde aantal maanden werkloosheid (UE) en het gemiddelde aantal maanden buiten-de-arbeidsmarkt (OLF). Deze drie uitkomsten tellen op tot 30. We gaan er steeds van uit dat de beleidsregels gericht zijn op het maximaliseren, respectievelijk minimaliseren, met een gelijk gewicht, van het aantal maanden werk, respectievelijk van het aantal maanden werkloosheid. De laatste twee kolommen geven de relatieve verandering (in %) in werk en werkloosheid weer voor de subpopulatie van die individuen die bij een toepassing van de regel in een andere behandelingstoestand zouden terecht komen (de "switchers").

Uit een vergelijking van de waargenomen allocatie, die als referentiepunt dient voor de simulaties, met de toevallige allocatie ("random", waarbij de waarschijnlijkheid van allocatie gelijk is aan het waargenomen aandeel van de deelname aan de verschillende programma's) blijkt dat de gemiddelde effecten van de twee toewijzingen aan de programma's zeer vergelijkbaar zijn. De vergelijking van het resultaat van de toevallige allocatie met de waargenomen allocatie wijst erop dat het onwaarschijnlijk is dat bemiddelaars hun toewijzingsbeslissingen baseren op de in dit document beschreven heterogeniteit van de effecten. Dit lijkt in overeenstemming te zijn met het officiële beleid. VDAB gebruikte, althans in het verleden, allocatieregels zoals "70% van de deelnemers moet binnen 6 maanden na het einde van de opleiding aan het werk zijn", i.p.v. de allocatie te baseren op effectiviteitsdoelstellingen. Een PES zoals VDAB kan m.a.w. de resultaten van de opleidingen verbeteren door de allocatie te baseren op de verwachte programmaprestaties van personen zoals geschat door hun IATE.

Tabel 6.1: Globale effecten van een aantal gesimuleerde hypothetische alternatieve programma allocaties

Aandeel van de verschillende programma's in %				**Cumulatief # maanden binnen de 30 na de programmastart**			**Winst voor switchers in %**	
	SVT	LVT	OT	Emp	UE	OLF	Emp	UE
Geobserveerd	2.1	2.0	1.8	16.1	10.9	3.0	-	-
Random	2.0	1.9	1.9	16.1	10.9	3.0	0.9	-0.1
Black-Box – zonder beperking	97.3	0.2	0.2	19.4	8.9	1.6	21.9	-18.5
Black-Box – zonder beperking, alleen de significante	58.1	1.6	0.5	18.8	9.4	1.8	33.4	-23.0
Black-Box – beperkt, voorkeur voor grootste winst	2.1	2.0	1.8	16.4	10.8	2.8	13.6	-6.3
Black-Box – beperkt, sequentiële optimisatie	2.1	2.0	1.8	16.4	10.8	2.9	19.3	-8.0
Black-Box – beperkt, voorkeur voor werkloosheid voordien*)	2.1	2.0	1.8	16.2	10.9	2.9	6.6	-2.5
Simpel – Beleidsboom van 3 niveaus, beperkt	2.0	2.0	1.8	16.3	10.8	2.9	15.0	-4.2
Simpel – Beleidsboom van 4 niveaus, beperkt	2.1	2.0	1.8	16.4	10.8	2.8	14.6	-6.7

Noot: Emp: Werk; OFL: Buiten-de-arbeidsmarkt; UE: Werkloos. Allocaties minimaliseren het aantal maanden werkloosheid en maximaliseren het aantal maanden werk (met gelijk gewicht). *) Als de capaciteit van de programma's een beperking wordt, wordt de voorkeur gegeven aan degenen met het hoogste aantal maanden werkloosheid in de afgelopen tien jaar. De sequentiële optimalisering houdt in dat, te beginnen met de waargenomen toewijzing, de programmatoestanden van elk paar individuen paarsgewijs worden omgeschakeld als het resultaat over het geheel genomen beter is en de begrotingsrestrictie wordt gehandhaafd.

Vervolgens bekijken we verschillende Black-Box allocatieresultaten die afhankelijk zijn van de beschikbare programmacapaciteit en de mate van zekerheid over de effectiviteit van de programma's en één of andere prioriteitsregel. De derde regel (geen beperking) geeft de resultaten weer voor het geval dat er geen capaciteits- of budgetbeperkingen zouden zijn. In dit geval zou de PES meer dan 97% van de werklozen aan SVT toewijzen, ongeveer 0,2% aan LVT en aan OT, en minder dan 3% zou ongetraind blijven. Een dergelijke toewijzing zou gemiddeld het aantal maanden werk kunnen doen toenemen van 16,1 tot 19,4 (een toename met meer dan 20%), het aantal maanden in de werkloosheid kunnen doen afnemen van 10,9 tot 8,9 (een afname met ongeveer 18%) en buiten-de-arbeidsmarkt van 3,0 tot 1,6 maand (een afname met 53%). Dit zou uiteraard erg duur zijn,

omdat het aantal deelnemers aan SVT hierdoor enorm zou toenemen. Of de totale baten opwegen tegen deze kosten is onduidelijk, omdat we geen informatie hebben over o.m. de programmakosten, wat nodig is om een kosten-batenanalyse uit te voeren.

Bij de vorige allocatie is voorbijgegaan aan het feit dat sommige geschatte IATE positief (of omgekeerd) zijn, maar niet statistisch significant verschillend van nul. Daarom melden we in regel vier (geen beperking, alleen de significante) het resultaat van een simulatie waarbij we personen alleen aan programma's toewijzen als de overeenkomstige IATE's significant positief of negatief zijn op het niveau van 2,5% van een eenzijdige statistische test wat betreft het effect ervan op de tijd die in respectievelijk werk of werkloosheid wordt doorgebracht. Uit deze simulatie blijkt dat de IATE's voor een groot deel van de populatie (39%) niet significant verschillen van nul, en relatief klein zijn, aangezien het gemiddelde effect van het programma op de populatie in veel mindere mate afneemt. Desondanks wordt bijna 60% van de werklozen toegewezen aan SVT. De gemiddelde toename in termen van werk is nog steeds in totaal 2,7 maanden en komt overeen met een verbetering van 33% voor de waarnemingen die opnieuw zijn toegewezen.

In de volgende scenario's wordt uitgegaan van gevallen waarin de opleidingscapaciteit van de VDAB beperkt is tot de waargenomen capaciteit (en de kosten van het programma dus ongeveer gelijk blijven). Aangezien slechts 6% van de populatie aan opleiding deelneemt, zal een wijziging van allocatie geen drastisch gevolg hebben voor het totale resultaat. Voor degenen die daadwerkelijk door de re-allocatie worden getroffen, kunnen de voordelen echter zeer aanzienlijk zijn. In het eerste scenario wordt voorrang gegeven aan de personen die het hoogste rendement op de deelname aan het programma behalen (beperkt, voorkeur voor de grootste winst). De gemiddelde werkgelegenheidswinst voor de populatie is nog steeds 9,3 dagen (d.w.z. 0,31 maanden). Per ge-realloceerde werkzoekende komt dit overeen met een aanzienlijke werkgelegenheidswinst van ongeveer 14%. Het volgende scenario probeert dicht bij een beperkt maximum te komen door het vinden van paren van waarnemingen waarvoor het uitwisselen van het behandelingsstatuut werk, resp. werkloosheid verbetert, binnen de bestaande capaciteitsbeperking. Zulke uitwisselingen worden achtereenvolgens berekend tot er geen dergelijke paren meer kunnen worden gevonden. Dit leidt tot een aanzienlijke winst in termen van maanden werk van ongeveer 19% en een vermindering van de maanden werkloosheid van ongeveer 8% voor degenen die werden gerealloceerd, een verbetering van respectievelijk 42% en 21% ten opzichte van het vorige scenario. In de laatste Black-Box-aanpak gebruiken we een prioriteitsregel die niet is gebaseerd op effectiviteit. Dit kan worden gerechtvaardigd door bezorgdheid over billijkheid of positieve actie, of door politieke beperkingen. Concreet betekent dit dat we de programma's voor elk individu eerst rangschikken op basis van het "beste" geschatte counterfactual-resultaat. Als dit leidt tot minder deelnemers dan programmaslots, dan worden al deze personen aan dit programma toegewezen. Zo niet, dan worden de individuen volgens een andere prioriteitsregel toegewezen en niet aan hun volgende beste programma. Dit gaat door totdat alle programmaslots zijn gevuld. In tabel 6.1 wordt prioriteit gegeven aan personen met de meeste maanden werkloosheid in de afgelopen 10 jaar. Ten opzichte van de vorige regel vermindert dit de relatieve winst van de switchers tot ongeveer een derde. Dit is in overeenstemming met de verwachtingen, omdat we hebben vastgesteld dat de effectiviteit van het programma het grootst is voor mensen met de minste maanden werkloosheid (tabel 5.3).

Vervolgens nemen we het voorstel van Zhou, Athey en Wager (2019) over en gebruiken we ondiepe beslissingsbomen van diepte 3 (wat resulteert in 8 lagen met mogelijk verschillende behandelingstoewijzingen) en 4 (16 lagen) om een meer intuïtieve, interpreteerbare regel te verkrijgen. De onbeperkte toewijzingen worden verkregen door een lichte wijziging van algoritme 2 van Zhou, Athey en Wager (2019), die zou moeten leiden tot enkele betere toewijzingen ten koste van iets hogere rekenkosten.

De eenvoudige regel met drie niveaus die door deze procedure wordt bepaald (uitgaande van de bestaande, dus beperkte capaciteit), is volgt[3]:

“Alleen personen die ouder zijn dan 28,2 jaar worden naar een programma gealloceerd. Binnen deze groep worden enkel degenen die in een oostelijke EU-land zijn geboren, aan beroepsopleiding toegewezen: werklozen die de afgelopen twee jaar minder dan 10 maanden hebben gewerkt aan SVT en personen die in deze periode meer dan 9 maanden hebben gewerkt aan LVT. OT is voor werknemers die niet in een Oost-Europees land zijn geboren en geen kennis van het Nederlands hebben.”

[3] In de working paper waarnaar werd verwezen in de inleiding, vindt men in tabel 6.2 een grafische weergave van het resultaat van de regels met 3 en met 4 niveaus.

Wanneer er geen capaciteitsbeperkingen zijn (niet in tabel 6.1), is een ondiepe boom van diepte 3, bijna net zo goed als de Black-Box-regel.

Wanneer wel wordt vertrokken van de bestaande capaciteit, leiden de eenvoudige regels op basis van een boom van diepte 3 tot iets hogere verliezen voor gerealloceerde personen ten opzichte van de Black-Box aanpak met sequentiële optimalisatie, met name voor de werkloosheid-uitkomst: Bij deze groep vermindert het aantal maanden in werkloosheid met 4,2 in plaats van met 8 maanden (= -48%), terwijl het aantal maanden werk toeneemt met 15 in plaats van 19,3 maanden (= -22%).

Een meer complexe beslissingsboom met 4 niveaus doet het, in vergelijking met de meer eenvoudige regel met 3 niveaus, beter op het vlak van werkloosheid, maar niet op het vlak van werk.

7 SENSITIVITEITSANALYSE

7.1 PLACEBO-ANALYSE

Om de lezer (en onszelf) er verder van te overtuigen dat de matchende variabelen de CIA ondersteunen, doen we een placebo validatietest zoals voorgesteld door Imbens en Wooldridge (2009, pp. 48-50). Deze validatie bestaat uit het schatten met dezelfde methodologie van de ATE's van deelname aan een toekomstig opleidingsprogramma binnen een voorafgaande werkloosheidsperiode. Aangezien de (onverwachte) toekomstige deelname aan opleidingen geen invloed zou mogen hebben op de huidige resultaten, biedt het vinden van een effect dat dicht bij nul ligt enige steun voor de CIA.

Om deze placebotest uit te voeren, selecteren we uit de analysepopulatie de subpopulatie die ten minste één voorafgaande werkloosheidsperiode heeft meegemaakt - in het geval van meerdere perioden, behouden we de eerste waargenomen periode - die begint tussen september 2008 en februari 2014. Februari 2014 wordt gebruikt, omdat dit leidt tot een verschil van 9 maanden tussen het begin van de laatste werkloosheidsperiode die werd aangehouden voor de placebo-steekproef en de eerste onderzochte periode in de hoofdanalyse, d.w.z. december 2014. Dit verschil maakt het mogelijk om de placebo-behandelingseffecten te schatten gedurende 9 maanden sinds het begin van de vorige werkloosheidsperiode. Deze keuze van 9 maanden is gericht op het vinden van een evenwicht tussen het niet te veel reduceren van de omvang van de placebopopulatie - deze omvang neemt snel af met de omvang van het verschil in de tijd - en het hebben van een voldoende lange periode om de placebo-effecten te meten. Om contaminatie te voorkomen, hebben we alle personen die tijdens deze voorafgaande periode van werkloosheid aan een ALMP hebben deelgenomen, laten vallen. De uiteindelijke steekproef waarop deze placebo-analyse wordt uitgevoerd bestaat uit 17.943 niet-deelnemers en 360, 336 en 285 deelnemers aan respectievelijk SVT, LVT en OT.

Tabel 7.1 geeft de resultaten weer voor drie uitkomsten: cumulatief aantal maanden werk, werkloosheid én buiten-de-arbeidsmarkt 9 maanden na de start van de vorige werkloosheidsperiode. Uit de resultaten blijkt dat alle ATE's bijna nul zijn en ondanks de vrij kleine programmagroepen nauwkeurig worden geschat.

7.2 TUNING PARAMETERS

Om de stabiliteit van de MCF-schattingen met betrekking tot diverse afstemmingsparameters te onderzoeken, werden de volgende sensitiviteitsanalyse uitgevoerd: (i) het aantal bootstrap replicaties werd verhoogd van 1000 naar 2000 replicaties; (ii) de minimale bladgrootte werd gevarieerd van 5 tot 3 naar 7; (iii) het aandeel van de subsampling werd verlaagd van 67% tot 50%; (iv) er werd gevarieerd met het aantal variabelen dat wordt gebruikt voor het splitsen van een bepaald blad; (v) de schatting is uitgevoerd met en zonder voorafgaande deselectie van irrelevante kenmerken, en (vi) de penalty term in de MCF-objectiefunctie is verhoogd met een factor 10 ten opzichte van de basiswaarde die gelijk is aan de variantie van de respectievelijke uitkomstvariabelen. Geen van deze variaties heeft geleid tot substantiële veranderingen in de schattingsresultaten.

Tabel 7.1: Placebo-effecten voor verschillende toekomstige programma's, cumulatief aantal maanden werk, werkloosheid en buiten-de-arbeidsmarkt (ATE)

	Geen deelname (NOP)	Korte beroepsopleiding (SVT)	Lange beroepsopleiding (LVT)	Orientatie training (OT)
Cumulatief aantal maanden werk gedurende 9 maanden na de start van de vroegere werkloosheidsperiode				
NOP	**3.9 (0.1)**			
SVT	0.01 (0.3)	**3.9 (0.3)**		
LVT	0.5 (0.3)	0.5 (0.4)	**4.3 (0.3)**	
OT	0.001 (0.4)	-0.02 (0.4)	0.5 (0.4)	**3.9 (0.2)**
Cumulatief aantal maanden werkloosheid gedurende 9 maanden na de start van de vroegere werkloosheidsperiode				
NOP	**4.8 (0.04)**			
SVT	-0.1 (0.3)	**4.8 (0.3)**		
LVT	-0.4 (0.3)	-0.5 (0.4)	**4.4 (0.3)**	
OT	-0.002 (0.3)	-0.1 (0.4)	0.4 (0.4)	**4.8 (0.3)**
Cumulatief aantal maanden buiten de arbeidsmarkt gedurende 9 maanden na de start van de vroegere werkloosheidsperiode				
NOP	**0.4 (0.01)**			
SVT	-0.1 (0.1)	**0.3 (0.02)**		
LVT	-0.1 (0.1)	-0.005 (0.1)	**0.3 (0.1)**	
OT	-0.03 (0.1)	0.04 (0.2)	0.04 (0.2)	**0.3 (0.1)**

Noot: Uitkomsten zijn gemeten in maanden. Het niveau van potentiële resultaten voor het specifieke programma op de hoofddiagonaal is vetgedrukt. Alle effecten zijn populatiegemiddelden voor de respectievelijke deelnemers aan het placeboprogramma die in de kolom zijn vermeld. Standaardfouten staan tussen haakjes. *, **, *** geven de nauwkeurigheid van de schatting aan, waarbij de p-waarde van een tweezijdige significancetest respectievelijk lager is dan 10%, 5%, 1%.

7.3 VERDELING VAN DE GEWICHTEN

De geschatte causale effecten van Causal Forests kunnen worden gezien als het gewogen gemiddelde van de uitkomstvariabele. Deze gewichten kunnen worden onderzocht om de stabiliteit van de schatting te controleren. Als een beperkt aantal gewichten zeer groot is, wijst dit erop dat een zeer beperkt aantal observaties een zeer belangrijke rol spelen bij het schatten van de counterfactual. Zo beschouwden Huber, Lechner, Wunsch (2013) bijvoorbeeld gewichten met waarden groter dan 4% van de totale (absolute) som van de gewichten als een aandachtspunt. Het bleek dat voor de ATE's en de GATE's geen van de gewichten boven de 1%, respectievelijk 3% ligt. Uitzondering is de GATE voor het land van herkomst, waarvoor in de deelsteekproeven van de deelnemers aan de training ongeveer 0,4% van de waarnemingen gewichten tussen de 4% en 10% hebben. Voor de IATE's is de situatie extremer (zoals blijkt uit de grotere standaardfouten die hierboven zijn weergegeven), aangezien de gewichten van nature meer geconcentreerd zijn: in de trainings-substeekproeven ligt ongeveer 1-2% van de gewichten boven de 4% en in zeer zeldzame gevallen kunnen ze zelfs 25% bereiken. Verder onderzoek zal de implicaties van dergelijke grote gewichten moeten onderzoeken, en hoe ze kunnen worden aangepast om kleine steekproefproblemen te voorkomen. Hieruit volgt ook dat er veel grotere steekproeven nodig zijn voor een betrouwbare schatting van de IATE's dan de voor ATE of GATE.

7.4 VERGELIJKING MET PROPENSITY SCORE MATCHING

We hebben ook de resultaten van de MCF voor het ATE vergeleken met conventionele matchingschattingen met de propensityscore-matching-met-biascorrectie, een schatter gesuggereerd door Lechner, Miquel, en Wunsch (2011) die goede prestaties leverde in de uitgebreide Empirische Monte Carlo-studie van Huber, Lechner, en Wunsch (2013). Zowel de algemene puntschattingen als de standaardfouten leiden tot zeer vergelijkbare conclusies als de MCF. Een dergelijke vergelijking kan natuurlijk alleen worden gemaakt voor het ATE (of ATE's voor sommige grote subgroepen), aangezien dergelijke schattingen niet zijn aangepast voor de schatting van meer gedesaggregeerde effecten.

8 CONCLUSIES EN AANBEVELINGEN

In dit onderzoek hebben we gebruik gemaakt van recente ontwikkelingen op het gebied van causal machine learning om de gemiddelde en heterogene effecten van een aantal recent uitgevoerde ALMP's in Vlaanderen te onderzoeken, met behulp van administratieve individuele gegevens van de VDAB. De cijfer tussen rechte haken verwijzen naar de onderzoeksvragen uit sectie 2.

[C1. Verschillende effectiviteit voor de diverse ALMP's] We vonden dat alle programma's gemiddeld genomen positieve werkgelegenheidseffecten hebben op de middellange termijn, hoewel ze niet altijd voldoende groot zijn om na 2,5 jaar de negatieve effecten in de lock-in periode te compenseren. Het bleek dat gemiddeld genomen korte beroepsopleidingen effectiever zijn dan TIBB, langere beroepsopleidingen en oriënterende opleidingen. Wat betreft andere effecten, wordt vastgesteld dat alle programma's de tijd buiten de arbeidsmarkt verminderen, maar alleen de korte beroepsopleidingen (en in beperkte mate ook TIBB) reduceren ook de tijd in werkloosheid.

[C2. Effect-heterogeniteit] Uit de analyse van de heterogeniteit van de effecten bleek dat de programma's (ook na de lock-in periode) beter lijken te werken voor werklozen met een lage inzetbaarheid, met name recente migranten met een beperkte taalvaardigheid (deze effect-heterogeniteit is niet identiek voor alle programma's). Na het exploiteren van de grote variatie in de individuele effectiviteit voor een analyse van het toewijzingsbeleid van de VDAB bleek dat er sprake is van een aanzienlijke inefficiëntie. Een andere, op de te verwachten effecten gebaseerde, toewijzing van werklozen aan bestaande programmaslots kan leiden tot een aanzienlijke verbetering van de arbeidsmarktprestaties tegen geen of geringe extra kosten.

We kunnen deze bevindingen vergelijken met de metastudie van Card, Kluve en Weber (2018) die de effectiviteit van het actief arbeidsmarktbeleid (ALMP) analyseerden op basis van meer dan 200 papers. In lijn met hun algemene bevindingen detecteren we bijna geen effecten op de korte termijn als gevolg van lock-in. Ook stemt overeen dat opleidingsprogramma's na twee tot drie jaar effectief worden. Zij vinden over het algemeen dat programma's met meer accumulatie van menselijk kapitaal (d.w.z. opleiding) effectiever zijn op de langere termijn. Onze bevindingen nuanceren deze conclusies, aangezien ons bewijs aantoont dat SVT zelfs op de lange termijn net zo effectief is als LVT. Card et al. (2018) vinden dat heterogeniteit relevant is, maar ze rapporteren niet, zoals wij, hogere effectiviteit voor (recente) migranten of met betrekking tot het opleidingsniveau. Zij maken melding van hogere effecten voor vrouwen, langdurig werklozen en tijdens periodes van recessie. We vinden nooit een differentiële impact voor wie woont in een regio's met een hoge werkloosheid.

De werkloosheidsduur, naast andere factoren die negatief gerelateerd zijn met de inzetbaarheid, zijn van belang in de lock-in fase (9 maanden na de start van het programma): het negatieve effect van lock-in is voor minder inzetbare deelnemers minder sterk, omdat zij bij niet-deelname ook een veel kleinere kans op werk hebben.

[C3. ATE versus ATET] Er kon, op één uitzondering na, geen statistisch significant verschil worden gevonden tussen het gemiddeld populatie-effect (ATE) enerzijds, en de gemiddelde effectiviteit voor de groep van personen die daadwerkelijk deelnamen aan de diverse ALMP's (ATET) anderzijds.

[C4. Alternatieve allocatie] We hebben onze schattingen van de individuele programma-effecten gebruikt om de bestaande toewijzing door VDAB aan de beschouwde opleidingsprogramma's te evalueren. We stelden vast dat een wijziging van de toewijzingsregels de tijd die de toegewezen personen aan het werk zijn, met ongeveer 20% kan doen toenemen. Dit is een aanzienlijke winst en illustreert de maatschappelijke waarde van het gebruik van CML-methoden bij de toewijzing van werklozen aan het actieve arbeidsmarktbeleid.

[C5. Slotbeschouwing en toekomst] Vanuit methodologisch oogpunt lijkt de Modified Causal Forest (MCF)-aanpak (Lechner, 2018) goed geschikt voor een dergelijke analyse en bleek deze te leiden tot plausibele en informatieve resultaten tegen redelijke rekenkosten.

In de toekomst zouden veel openstaande kwesties kunnen worden aangepakt, zoals de uitbreiding van de databank met (i) programmakosten, zodat de afleiding van optimale beleidsregels kan worden gebaseerd op de netto maatschappelijke waarde van programma's, en met (ii) bijkomende controlevariabelen, zodat de effecten van andere programma's van het actieve arbeidsmarktbeleid van Vlaanderen die hier buiten beschouwing worden gelaten, ook geloofwaardig kunnen worden geëvalueerd. Bovendien zal het interessant zijn om te zien of er in andere landen met een vergelijkbaar beleid een vergelijkbare heterogeniteit optreedt.

Uit dit onderzoek volgen een aantal duidelijke aanbevelingen (hier is er niet langer sprake van een één-op-één link met de onderzoeksvragen, zodat een gewone oplopende nummering wordt gebruikt):

[1] Er wordt aanbevolen dat VDAB bij de inrichting van het actief arbeidsmarktbeleid meer dan nu het geval is rekening houdt met de effectiviteit van de acties. De “inrichting” van het actief arbeidsmarktbeleid heeft overigens betrekking op verschillende aspecten:

-de maatregelkeuze of maatregelmix (welke maatregelen wel, welke niet?);

-de te voorziene capaciteit voor iedere specifieke maatregel (waarbij meer capaciteit soms wel, maar zeker niet altijd optimaal is);

-de allocatie, meer in het bijzonder door die werkzoekenden toe te wijzen die het meeste baat hebben bij deelname aan een gegeven maatregel;

[2] Wat dit laatste aspect betreft, blijkt meer in het bijzonder dat een andere, op de te verwachten effecten gebaseerde, toewijzing van werkzoekenden aan bestaande programmaslots kan leiden tot een aanzienlijke verbetering van de arbeidsmarktprestaties tegen geen of geringe extra kosten. Er wordt dan ook aanbevolen om bij voorrang in te focussen op dit aspect.

Hierbij valt op te merken dat in het onderzoek slechts een beperkte selectie van (weliswaar kwantitatief belangrijke) maatregelen werd opgenomen, waardoor de analyse voor een stuk partieel blijft. Daarom wordt aanbevolen dat de oefening wordt uitgebreid tot alle relevante maatregelen (die een voldoende grote omvang hebben).

[3] De vraag hoe men bemiddelaars kan ondersteunen bij het toewijzen van werkzoekenden aan de maatregel waar deze het meeste baat bij hebben, kan op verschillende manieren worden beantwoord. Men kan kiezen voor een op AI gebaseerd recommender-systeem, dat aanbevelingen op maat produceert (maar misschien te veel als een black-box-oplossing zal worden gepercipieerd). Anderzijds kan men kiezen voor meer transparante beslissingregels op basis van een meer beperkt aantal kenmerken van de werkzoekende.

Dit onderzoek toont alvast aan dat beide benaderingen kunnen worden gegrondvest op een zelfde wetenschappelijke basis. De in dit onderzoek gebruikte MCF-methodiek, of een soortgelijke benadering, vormt hier zeker een interessant vertrekpunt.

[4] Als men van de voorgaande aanbevelingen werk wil maken, volstaat het niet om zich te beperken tot een eenmalige schatting van de effect-heterogeniteit van de diverse maatregelen. Er wordt aanbevolen om werk te maken van een systeem dat toelaat om met een zekere regelmaat een actualisatie te maken van de schattingen, zodat men onder meer rekening kan houden met wijzigingen in de arbeidsmarktconjuctuur én in het aanbod van maatregelen.

Een op machine learning-technieken gebaseerde benadering leent zich bij uitstek tot automatisering. Men maakt dan wellicht ook best werk van een geautomatiseerde aanlevering van de data-input die nodig is voor dergelijke modellen.

[5] Een beter zicht op de effectiviteit van het arbeidsmarktbeleid is dus zonder meer fundamenteel voor de inrichting van dat beleid. Maar eigenlijk moet men nog een stap verder zetten door met name ook rekening te houden met de kostprijs van het beleid. Daarom wordt aanbevolen dat ook op dit vlak wordt gewerkt aan het berekenen van correcte kostprijzen, zodanig dat er invulling kan worden gegeven aan de notie “kosten-effectiviteit”.

REFERENTIES

Arthur, D., Vassilvitskii, S. (2007): k-means++: the advantages of careful seeding, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1027–1035.

Athey, S. (2019): The Impact of Machine Learning on Economics, in Agrawal, A., J. Gans, and A. Goldfarb (eds.), The Economics of Artificial Intelligence: An Agenda. 507-547, Chicago: Chicago University Press.

Athey, S., and G. W. Imbens (2016): Recursive Partitioning for Heterogeneous Causal Effects, *Proceedings of the National Academy of Sciences of the United States of America*, 113(27), 7353–7360.

Athey, S., and G. W. Imbens (2019): Machine Learning Methods Economists Should Know About, *arXiv*.

Athey, S., and S. Wager (2019a): Estimating Treatment Effects with Causal Forests: An Application, forthcoming in *Observational Studies*.

Athey, S., and S. Wager (2019b): Efficient Policy Learning, *arXiv*.

Athey, S., J. Tibshirani, and S. Wager (2019): Generalized Random Forests, *The Annals of Statistics*, 47, 1148-1178.

Autor, D. H., F. Levy, and R. J. Murnane (2003): The Skill-Content of Recent Technological Change: An Empirical Investigation, *Quarterly Journal of Economics*, 118, 1279-1333.

Bertrand, M., B. Crépon, A. Marguerie and P. Premand (2017): Contemporaneous and Post-Program Impacts of a Public Works Program: Evidence from Côte d'Ivoire, *mimeo*, University of Chicago, Crest and World Bank.

Biewen, M., B. Fitzenberger, A. Osikominu, and M. Paul (2014): The effectiveness of public sponsored training revisited: the importance of data and methodological choices, *Journal of Labor Economics*. 32 (4), 837–897.

Card, D., J. Kluve and A. Weber (2018): What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations, *Journal of the European Economic Association*, 16(3). 894-934.

Chernozhukov, V., I. Fernandez-Val, and Y. Luo (2018): The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages, *Econometrica*, 86, 1911-1938.

Chou, Philip A. (1991): Optimal partitioning for Classification and Regression Trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13-4, 340-354.

Crépon, B., Ferracci, M., Jolivet, G. van den Berg, G.J. (2009): Active labor market policy effects in a dynamic setting, *Journal of the European Economic Association*, 7, 595–605.

Fredriksson P., and P. Johansson (2008): Dynamic treatment assignment: the consequences for evaluations using observational data, *Journal of Business and Economic Statistics*, 26(4), 435–445.

Friedman, J., Hastie, T. and Tibshirani, R. (2008): Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33(1), 1-22. Goos, M., A. Manning and A. Salomons (2009): The Polarization of the European Labor Market, *American Economic Review Papers and Proceedings* 99, 58-63.

Goos, M., A. Manning, and A. Salomons (2009): Job Polarization in Europe, *American Economic Review, P&P*, 99:2, 58-63.

Hastie, T., R. Tibshirani, and J. Friedman (2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer (10th printing with corrections, 2013).

Heckman, J. J., H. Ichimura, J. A. Smith, and P. E. Todd (1998): Characterizing selection bias using experimental data, *Econometrica* 66, 1017–1098.

Huber, M., M. Lechner, and C. Wunsch (2013): The performance of estimators based on the propensity score, Journal of Econometrics, 175 (1), 1-21.

Imbens, G. W. (2000): The Role of the Propensity Score in Estimating Dose-Response Functions, *Biometrika*, 87, 706-710.

Imbens, G. W., and J. M. Wooldridge (2009): Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, 47 (1), 5-86.

Knaus, M. C., M. Lechner, and A. Strittmatter (2017): Heterogeneous Employment Effects of Job Search Programs: A Machine Learning Approach, *arXiv: 1709.10279v2*.

Knaus, M. C., M. Lechner, and A. Strittmatter (2018): Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence, *arXiv: 1810.13237v1*.

Langenbucher, K. (2015): How demanding are eligibility criteria for unemployment benefits, quantitative indicators for OECD and EU countries, *OECD Social, Employment and Migration Working Papers*, No. 166, OECD Publishing, Paris, doi: 10.1787/5jrxtk1zw8f2-en.

Latham, G. P., M. B. Mawritz and E. A. Locke (2018): Goal Setting and Control Theory: Implications for Job Search, in: Klehe U.-C. and E. van Hooft (eds.), *The Oxford Handbook of Job Loss and Job Search*, Chapter 8.

Lechner, M. (1999): Earnings and Employment Effects of Continuous Off-the-job Training in East Germany after Unification, *Journal of Business Economics and Statistics*, 17, 74–90.

Lechner, M. (2001): Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in: M. Lechner and F. Pfeiffer (eds.), *Econometric Evaluation of Active Labour Market Policies*, 43-58, Heidelberg: Physica.

Lechner, M. (2002): Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programs by Matching Methods, *Journal of the Royal Statistical Society, Series A*, 165, 59–82.

Lechner, M. (2018): Modified Causal Forests for Estimating Heterogeneous Causal Effects. Version 2. arXiv: 1812.09487v2.

Lechner, M. and C. Wunsch (2013): Sensitivity of matching based program evaluations to the availability of control variables, *Labour Economics* 21, 111-121.

Lechner, M., R. Miquel, and C. Wunsch (2011): Long-run effects of public sector sponsored training in West-Germany, *Journal of the European Economic Association*, 9 (4), 742–784.

Piton, C., and F. Rycx (2020): The heterogeneous employment outcomes of first- and second-generation immigrants in Belgium, National Bank of Belgium discussion paper N°381, Brussels: National Bank of Belgium.

Rosenbaum, P.R., and D.B. Rubin (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70, 41-50.

Rosenbaum, P.R., and D.B. Rubin (1985): Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, *The American Statistician*, 39, 33-38.

Rubin, D. B. (1974): Estimating Causal Effects of Treatments in Randomized and Nonrandomized *Studies, Journal of Educational Psychology, 66, 688-701.*

Sianesi, B. (2004): An evaluation of the Swedish system of active labour market programs in the 1990s, *Review of Economics and Statistics* 86, 133–155.

Sianesi, B. (2008): Differential effects of active labour market programs for the unemployed, *Labour Economics*, 15 (3), 370–399.

Van den Berg, G., and J. Vikström (2019): Long-Run Effects of Dynamically Assigned Treatments: A New Methodology and an Evaluation of Training Effects on Earnings, IZA Discussion paper 12470, Bonn: IZA.

Van Hooft, E. A. J., and G. Noordzij (2009): The Effects of Goal Orientation on Job Search and Reemployment: A Field Experiment Among Unemployed Job Seekers, *Journal of Applied Psychology* 94(6), 1581-1590.

Vikström, J. (2017): Dynamic treatment assignment and evaluation of active labor market policies. *Labour Economics*, 49, 42–54.

Wager, S., and S. Athey (2018): Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, *Journal of the American Statistical Association*, 113:523, 1228-1242.

Zhou, Z., S. Athey, and S. Wager (2018): Offline Multi-Action Policy Learning: Generalization and Optimization, arXiv:1810.04778v2.