VERO SKATT | HARMAAN TALOUDEN SELVITYSYKSIKKÖ

Miten verovelka ja harmaa talous liittyvät toisiinsa?

Selvitys perustuu verovelka- ja -tarkastustietoihin vuosilta 2017–2020

Selvitys 3/2022

Julkaisun nimi:

Miten verovelka ja harmaa talous liittyvät toisiinsa? – Selvitys perustuu verovelka- ja -tarkastustietoihin vuosilta 2017–2020

Tekijät:
Mikko Kotiranta ja Lasse Winter

Diaarinumero: VH/Harmaan talouden selvitysyksikkö/2021/057

Julkaisija: Verohallinto, Harmaan talouden selvitysyksikkö

Julkaistu: 27.9.2022

Julkaisutapa: Sähköinen (PDF)

Julkisuus: Julkinen

Lisätietoja medialle:
Johtaja Janne Marttinen, puh. 029 512 6066
Apulaisjohtaja Marko Niemelä, puh. 029 512 6070
Muut yhteydenotot htsy@vero.fi

Julkaisualustat harmaa-talous-rikollisuus.fi

Velvoitteidenhoitoselvityksen voi pyytää ilmiöselvityksessä kuvatun ryhmän perusteella. (Laki Harmaan talouden selvitysyksiköstä (1207/2010) 5:3 §)

Harmaan talouden selvitysyksikön selvityksissä esitetyt näkemykset ja tulkintakannanotot ovat yksikön omia, eivätkä ne sido Verohallintoa tai muita viranomaisia.

Tiivistelmä

Selvityksessä on tarkasteltu, miten verovelka ja harmaa talous liittyvät toisiinsa. Harmaan talouden osuutta veroveloista on ainakin osa verotarkastusten ja arvioverotuksen perusteella määrätyistä veroista (tahallinen väärinilmoittaminen). Muu harmaaseen talouteen kuuluva verovelka liittyy tarkoitukselliseen maksamattomuuteen.

Verovelan merkitys harmaan talouden ennustamisessa

Saatujen tulosten perusteella verovelkaisuus ei näytä johtavan merkittävästi kohonneeseen riskiin harjoittaa harmaata taloutta. Sama johtopäätös on tehtävissä yrityksen verovelkarekisteriin kuulumisen osalta.

Harmaan talouden yritysten verovelkojen euromäärä

Koneoppimismallia ja verotarkastusten tuloksia on hyödynnetty sen selvittämiseksi, onko osalla toimivista osakeyhtiöitä havaittavissa merkkejä harmaan talouden toiminnasta ja mikä on kyseisten yritysten osuus veroveloista. Yritykset on jaettu mallin ennustaman harmaan talouden todennäköisyyden perusteella kolmeen luokkaan: 1) valkoiset 2) harmaat 3) rajatapaukset.

Toiminnassa olleiden osakeyhtiöiden verovelan euromäärä on pysynyt suhteellisen samana, noin 200 miljoonassa eurossa, sekä harmaiden että rajatapauksiksi tulkittujen yritysten joukoissa kaikkina tarkasteluvuosina 2017–2020. Mallin mukaan suurin osa verovelasta on kuulunut odotetusti valkoisille yrityksille. Vuosien 2017–2019 verovelan määrän putoaminen 740 miljoonasta eurosta 300 miljoonaan euroon ja vuoden 2020 hetkellinen uudelleennousu lähemmäs miljardiin euroon tapahtuivat lähes täysin valkoisten yritysten joukossa.

Yritysten taloudelliset ongelmat ovat verovelkojen suurin syntymissyy - ei harmaa talous

Verovelan syntymissyitä on lähestytty substanssilähtöisen mallin avulla. Tällöin asiantuntija määrittelee säännöstön, joka luokittelee yritykset oletetun verovelan syntymissyyn mukaisiin luokkiin.

Selkeästi suurin määrä eli 13 500 yritystä (40 prosenttia kaikista verovelkaisista) kuuluu talousongelmallisten yritysten luokkaan. Juuri taloudellisten vaikeuksiensa vuoksi kyseiset yritykset eivät selviydy veroveloistaan. Toiseksi suurin sääntöjä käyttämällä syntyvä luokka on noin 5 000 myöhästelijäyritystä (15 prosenttia). Arvioverotettujen osuus on lähes yhtä suuri kuin myöhästelijöiden. Verotussäännöksiä noudattamatta jääneiksi yrityksiksi on luokiteltu hieman yli 750 yritystä (hieman yli kaksi prosenttia). Verovelkaisten yritysten syyluokittelu on tehty ainoastaan vuoden 2019 tiedoilla.

Verotussäännösten noudattamattomuus ja arvioverotus nostavat yksittäisen yrityksen harmaan talouden riskiä ja verovelan määrää

Kaikista suurin keskimääräinen verovelka, 90 000 euroa, on verotussäännöksiä noudattamattomilla yrityksillä. Näillä yrityksillä on myös korkeaksi arvioitu harmaan talouden riski (55 prosenttia). Arvioverotettujen yritysten keskimääräinen verovelka on toiseksi korkein eli liki 40 000 euroa.

Avainsanat: Ilmiöt; Maksukyvyttömyys; Verovaje

Selvityksessä käytetyt keskeiset määrittelyt

VEROVELKA

Verovelka

Verojen maksuerä, jota ei ole maksettu eräpäivään mennessä.

Verovelkainen yritys

Verovelkaisella yrityksellä on kolme päivää vanhempaa verovelkaa kalenterivuoden viimeisenä päivänä tai yritys on ollut merkittynä verovelkarekisteriin kalenterivuoden aikana yhdenkin päivän ajan.

Verovelan perussyy

Verovelan teoreettisia perussyitä ovat: 1) Osaamattomuus, tietämättömyys tai välinpitämättömyys 2) Tahallinen väärinilmoittaminen 3) Maksukyvyttömyys 4) Tarkoituksellinen maksamattomuus.

Verovelkojen syntymiseen vaikuttavat syyt

Selvityksessä muodostettujen asiantuntijasääntöjen mukaan verovelkojen syntymiseen vaikuttavia syitä ovat: 1) Verotussäännösten noudattamattomuus (valvontatoimenpide tai maksuunpanoon johtanut tarkastuskäynti) 2) Myöhästely (myös veroilmoituksista yli 20 prosenttia annettu myöhässä) 3) Arvioverotus (vähintään kaksi arvioverotusta) 4) Poikkeuksellisen suuren maksuerän maksuunpano (yli kahdeksankertainen yrityksen kalenterivuoden maksuerien keskiarvoon verrattuna) 5) Taloudelliset ongelmat (tilikauden tulos negatiivinen, maksuvalmius huono tai omavaraisuusaste negatiivinen) 6) Muut syyt.

YRITYSTOIMINTA

Toimiva yritys

Rekistereissä toimiva yritys ei ole lopettanut toimintaansa ja on tarkasteluhetkellä ainakin yhdessä seuraavista rekistereistä: ennakkoperintärekisteri, alv-rekisteri (liiketoiminta, kiinteistön käyttöoikeuden luovutus, alkutuotanto) tai työnantajarekisteri.

Harmaa yritys

Verotarkastettu yritys on määritelty harmaan talouden yritykseksi, jos verotarkastuksen yksi tai useampi löydös viittaa harmaan talouden tekoon. Vaihtoehtoisesti verotarkastuksessa maksettavaksi määrättyjen verojen euromäärä on ylittänyt 10 000 euroa, ja löydös liittyy mahdollisesti harmaaseen talouteen.

Selvityksessä käytetyn ennustemallin mukaan harmaiden yritysten joukossa ovat yritykset, joiden todennäköisyys harmaan talouden toimintaan on mallin perusteella yli 80 % tai jotka ovat verotarkastusten tulosten perusteella harmaan talouden yrityksiä.

On huomattava, että tässä selvityksessä käytetty harmaan talouden yrityksen määrittely poikkeaa Verohallinnon vastaavasta. Verohallinnon harmaan talouden kohteiksi luokitellaan rikosilmoitusharkintaan saatetut yritykset.

Valkoinen yritys

Verotarkastettu yritys on määritelty valkoiseksi yritykseksi, jos yksikään verotarkastuksen löydös ei viittaa harmaan talouden tekoon. Vaihtoehtoisesti tarkastuksen aikana todennettu löydös on määritelty liittyvän mahdollisesti harmaaseen talouteen, mutta maksettavaksi määrättyjen verojen euromäärä ei ole ylittänyt 10 000 euroa.

Selvityksessä ennustemallin mukaan valkoisten yritysten alajoukkoon kuuluvat kaikki yritykset, joiden mallin mukainen todennäköisyys olla harmaa on alle 50 prosenttia tai jotka näyttäisivät verotarkastuksen tulosten perusteella noudattavan verotussäännöksiä.

Rajatapausyritys

Yritykset, joita ei ole luokiteltu valkoisiksi tai harmaiksi menevät rajatapausten luokkaan. Rajatapausyrityksen harmaan talouden todennäköisyys on 50–80 prosenttia ja sitä ei ole verotarkastettu.

Sisällys

1 Harmaa talous ja verovelat

Harmaa talous on toimintaa, jossa laiminlyödään lakisääteisiä velvoitteita verojen suorittamisen välttämiseksi. (Laki Harmaan talouden selvitysyksiköstä, 2010) Velvoitteiden laiminlyönti voi koskea rekisteröimistä, ilmoittamista ja maksamista. Harmaata taloutta on myös se, että kolmannen osapuolen, kuten tilitoimiston verovelvollisen puolesta antama ilmoitus sisältää virheitä ja verovelvollinen on tietoinen asiasta.

Harmaata taloutta on ainakin osa verotarkastusten ja arvioverotuksen perusteella määrättyjen verojen veroveloista (yrityksen tahallinen väärin ilmoittaminen, joka johtaa uuteen verojen maksuerään, jota ei kyetä hoitamaan). Osaamattomuudesta, tietämättömyydestä tai huolimattomuudesta johtuvat ilmoituslaiminlyönnit eivät ole sen sijaan harmaata taloutta. Esimerkiksi tietämättömyyteen vetoaminen ei kuitenkaan poista automaattisesti harmaan talouden toiminnan mahdollisuutta, koska veroilmoituksen antamiseen liittyvä selvitysvelvollisuus on voitu laiminlyödä tietoisesti.

Muu harmaaseen talouteen kuuluva erääntynyt verovelka liittyy tarkoituksellisen maksamattomuuteen. Harmaan talouden toimintaa ei ole toiminta, jossa maksuvelvollisuus olisi laiminlyöty taloudellisista vaikeuksista johtuen. Sen sijaan maksukyvytön voi lisävelkaantumista välttääkseen laiminlyödä ilmoitusvelvoitteitaan verojen maksamisen välttämiseksi. Sellaisessa tilanteessa maksukyvyttömänkin organisaation voitaisiin katsoa harjoittavan harmaata taloutta. (Viranomaistyöryhmän loppuraportti, 2014)

Tietyissä tilanteissa verovelkojen vuosittaisella kohdentamisella on suurtakin merkitystä. Osa asiakkaista pystyy kuitenkin maksamaan edellisen verovuoden erääntyneet verovelat vasta seuraavana vuonna. Tällöin seuraavan verovuoden verokertymä yliarvioidaan, jos siihen sisällytetään edellisenä vuoden erääntyneiden verovelkojen maksuja. Tämän perusteella taas voidaan tehdä vääriä johtopäätöksiä erääntyneiden verovelkojen pienentämiseksi tehtyjen toimenpiteiden vaikuttavuudesta. (Viranomaistyöryhmän loppuraportti, 2014)

Aikaisemmissa tutkimuksissa on selvitetty, mikä osa verovajeesta liittyy harmaaseen talouteen. Tässä selvityksessä on pyritty arvioimaan sen sijaan sitä, kuinka suuri osuus veroveloista liittyy harmaan talouden toimintaan. Ensimmäiseksi on selvitetty, miten yrityksen verovelka ennustaa harmaata taloutta. Toiseksi tarkastusaineistoa ja koneoppimismallia on hyödynnetty sen selvittämisessä, millä osalla toimivista yrityksiä on havaittavissa merkkejä harmaan talouden toiminnasta, ja mikä on kyseisten yritysten osuus veroveloista. Kolmanneksi aihetta on lähestytty tarkastelemalla verovelan aiheuttaneita erilaisia syitä. Jokaiselle määritellylle verovelan syyluokalle voi laskea siihen kuuluvien yritysten lukumäärän kaikkien yritysten joukosta, yksittäisen toimijan harmaan talouden todennäköisyyden ja keskimääräisen verovelan määrän.

Pääosa Verohallinnon harmaan talouden kohteista on verotarkastettuja yrityksiä, jotka on saatettu rikosilmoitusharkintaan. Tässä selvityksessä tarkastettu yritys on määritelty harmaan talouden toimijaksi, jos verotarkastuksen yksi tai useampi löydös viittaa harmaan

talouden tekoon. Vaihtoehtoisesti verotarkastuksessa maksettavaksi määrättyjen verojen euromäärä on ylittänyt 10 000 euroa, ja löydös liittyy mahdollisesti harmaaseen talouteen.

Erääntynyt verovelka on tärkeä tietoerä monelle viranomaiselle heidän toiminnassaan ja sen olemassaolo voi estää esimerkiksi luvan myöntämisen yritykselle. Verovelan rakenteen tarkastelu voi auttaa tunnistamaan harmaan talouden riskiryhmiä. Tällä voi olla merkitystä veronkantoon, perintä- ja turvaamistoimien valintaan ja niiden suorittamisen nopeuteen. Tietoa voidaan hyödyntää myös viranomaisten yhteisten toimenpiteiden tehokkaassa kohdentamisessa. (Verovelkaselvitys , 2012)

Selvitys jäsentyy johdannon ja yhteenvedon lisäksi neljään päälukuun. Aluksi luvussa kaksi on kuvattu selvityksen tutkimusongelma ja alatavoitteet. Luvussa kolme on esitetty tilastotietoja veroveloista ja -tarkastuksista. Luvussa neljä on tarkasteltu, mikä on verovelan vaikutus harmaan talouden toiminnan todennäköisyyteen. Tämän jälkeen luvussa viisi on selvitetty, missä suhteessa verovelka jakaantuu harmaan talouden yrityksille ja muille yrityksille. Luvussa kuusi on tarkasteltu verovelkaa erilaisin syntymissyihin perustuen. Lopuksi yhteenvedossa on vastattu tutkimusongelmaan (miten verovelka ja harmaa talous liittyvät toisiinsa) tehdyn selvitystyön perusteella.

2 Tutkimusongelma ja -aineisto

Selvityksen tarkoituksena on selvittää, miten verovelka ja harmaa talous liittyvät toisiinsa. Harmaan talouden osuutta veroveloista on ainakin osa verotarkastusten ja arvioverotuksen perusteella maksettavaksi määrätyistä veroista. Muu harmaaseen talouteen liittyvä verovelka liittyy tahalliseen maksamattomuuteen.

Alatavoitteina on selvittää:

- Miten yrityksen verovelka vaikuttaa harmaan talouden toiminnan todennäköisyyteen? (luku 4)
- Missä määrin verovelat jakaantuvat harmaan talouden toimintaa harjoittaville yrityksille, valkoisille ja muille yrityksille? (luku 5)
- Voidaanko yrityksiä ryhmitellä erilaisten verovelan syntymissyiden perusteella? (luku 6)

Verovelalla tarkoitetaan tässä selvityksessä yritykselle maksettavaksi määrättyä veroa, jonka maksaminen on laiminlyöty. Verovelan laiminlyöminen on määritelty kahdella eri tavalla, joita on tarkasteltu erikseen. Verovelan vaihtoehtoiset määritelmät ovat: 1) yrityksellä on verovelkaa[1] kalenterivuoden viimeisenä päivänä 2) yritys on merkitty verovelkarekisteriin[2] kalenterivuoden aikana tilapäisesti tai pysyvästi. Vaikka joillakin yrityksillä olisi haettuna ja myönnettynä maksunlykkäyksiä vuosien 2017–2020 veroveloille, on ne otettu mukaan tarkasteluun.

Perinnän tilastoissa maksamatta oleva vero tilastoidaan sille vuodelle, jona sen eräpäivä on ollut. Verotarkastuksen perusteella asiakkaalle on voitu määrätä maksettavaksi veroja, jotka olisi pitänyt maksaa aikaisempana vuonna.

Tässä selvityksessä verotarkastusaineisto muodostaa pohjan *harmaan talouden yrityksen*[3] määritelmälle. Verotarkastuksen kohteeksi joutunut yritys määritellään harmaan talouden yritykseksi, jos verotarkastuksen yksi tai useampi löydös viittaa harmaan talouden tekoon. Vaihtoehtoisesti tarkastuksen aikana todennettu löydös on määritelty mahdollisesti harmaaseen talouteen liittyväksi ja teon maksuunpanojen euromäärä on ylittänyt 10 000 euroa. Tällä tavalla toimimalla vuosina 2017–2020 verotarkastetuista reilusta 6 200 yrityksestä luokitellaan lähemmäs 40 prosenttia harmaan talouden toimijoiksi.

[1] Erääntyneen verovelan tulee olla vähintään kolmen päivän ikäinen.
[2] Verovelkarekisteristä näkyy, onko yrityksellä verovelkaa vähintään 10 000 euroa tai oma-aloitteisten verojen ilmoituslaiminlyöntejä viimeisten 6 kuukauden ajalta. Verovelkarekisterissä ei näytetä verovelan euromäärää, vaan onko verovelkaa vähintään 10 000 euroa vai ei. Verovelkamerkintä päivittyy pois viivytyksettä sen jälkeen, kun tieto maksusta tai ilmoituksesta on saapunut Verohallinnolle. Jos yrityksellä on voimassa Verohallinnon kanssa tehty maksujärjestely, verovelkarekisteritietoa ei julkaista verovelkarekisterissä. (Vero, 2021)
[3] Yleisesti harmaan talouden toiminnalla tarkoitetaan sellaista organisatorista toimintaa, josta aiheutuvia lakisääteisiä velvoitteita laiminlyödään verojen ja maksujen suorittamisen välttämiseksi tai perusteettoman palautuksen saamiseksi. Toiminta on tapahtunut viranomaisilta salassa tai siitä on annettu väärää tai puutteellista tietoa. (Laki harmaan talouden selvitysyksiköstä, 2010)

On huomattava, että tässä selvityksessä käytetty harmaan talouden yrityksen määrittely poikkeaa Verohallinnossa käytetystä. Verohallinnossa harmaan talouden yrityksiksi on määritelty vain ne yritykset, jotka on saatettu rikosilmoitusharkintaan.

Yksityisten elinkeinonharjoittajien ja henkilöyhtiöiden verovelkaa ei ole tarkasteltu tässä selvityksessä, koska kyseisistä yritysmuodoista ei ollut saatavilla riittävän kattavasti verotarkastustietoja. Lisäksi rajaamalla tarkastelu vain yksityisiin osakeyhtiöihin, ovat käsiteltävät yritys- ja verotustiedot suoraan keskenään suoraan vertailukelpoisia. Selvityksessä ei ole mukana ollenkaan ulkomaisia yrityksiä (verovelkarekisterissä on tieto vain sellaisista ulkomaisista yhteisöistä, jotka on rekisteröity sivuliikkeenä kaupparekisteriin) eikä lepääväksi tai lopetetuksi ilmoitettuja osakeyhtiöitä (merkittävä osa niiden veroilmoitustiedoista puuttuu).

Selvityksessä käytetyistä tilastollisista menetelmistä on kerrottu lyhyesti kunkin pääluvun yhteydessä. Lisäksi näitä menetelmiä on kuvattu yksityiskohtaisesti kolmessa liitteessä: 1) Verovelan selitysvoiman menetelmät 2) Yrityksen harmaata taloutta ennustava malli, ja 3) Yrityksen syntysyyt -aineistossa.

3 Tilastoja veroveloista ja tehdyistä tarkastuksista

Tässä luvussa on aluksi kuvattu sitä, mikä on osakeyhtiöiden verovelkojen osuus organisaatioiden ja yksityishenkilöiden kokonaisveroveloista. Samassa yhteydessä on esitetty tilastolukuja osakeyhtiöiden verovelan määrästä. Toisessa alaluvussa on tarkasteltu selvityksessä käytettyä osakeyhtiöiden verovelka- ja tarkastusaineistoa. Lopuksi on esitetty tilastoja verotarkastuksen lopputuloksen vaikutuksesta yrityksen verovelkaisuuteen.

3.1 Kokonaisverovelka ja osakeyhtiöiden verovelka

Verohallinto on kerännyt veroja vajaa 70 miljardia euroa vuosittain vuosina 2017–2020. Samana ajanjaksona yritysten ja yksityishenkilöiden verovelan kokonaismäärä on noussut reilusta kolmesta miljardista eurosta hetkellisesti viiteen miljardiin euroon. Osakeyhtiöiden verovelan määrä (600–1 300 miljoonaa euroa vuosina 2017–2020) ja sen osuus kokonaisveroveloista (vaihteluväli 15–31 prosenttia vuosina 2017–2020) on ollut laskussa lukuun ottamatta koronapandemian alkamisvuotta 2020.

Taulukko 1. Kokonaisverovelka ja osakeyhtiöiden verovelka kalenterivuoden viimeisenä päivänä.

Verovuosi	Organisaatioiden ja yksityishenkilöiden kokonaisverovelka (miljoonaa euroa)	Osakeyhtiöiden kokonaisverovelka (miljoonaa euroa)	Osakeyhtiöiden kokonaisverovelan osuus
2017	3 202	996	31,1 %
2018	3 464	963	27,8 %
2019	3 953	611	15,5 %
2020	4 923	1 276	25,9 %

Osakeyhtiön keskimääräisen verovelan määrä on pieni, suuruudeltaan vain muutama tuhat euroa[4]. Alla olevasta taulukosta selviää, että mediaaniverovelka on ollut kuitenkin vuodesta toiseen kasvussa ja on jo lähes 5 000 euroa vuonna 2020. Iso osa osakeyhtiöiden verovelasta syntyy yksittäisten yritysten miljoonaluokan veloista. Esimerkiksi erään yrityksen yli 300 miljoonan euron velat vastaavat noin kolmasosaa koko osakeyhtiöiden verovelan määrästä vuosina 2017–2018. Pelkästään tämän yhden yrityksen velkojen hoitaminen on pienentänyt osakeyhtiöiden verovelan määrän vajaasta miljardista vuosina 2017–2018 reiluun 600 miljoonaan vuonna 2019. Yhdellä prosentilla yritysten lukumäärästä laskettuna on ollut yli 250 000 euron verovelat vuosina 2017–2019. Koronavuonna 2020 velkaisimpaan prosenttiin pääsy vaati jo yli 500 000 euron velat.

[4] Suurimmalla osalla velallisista verovelan määrä on pääomaltaan alle 5 000 euroa. Verovelkaisista 5 prosentilla on verovelkaa 100 000 euroa tai enemmän. Selvityksestä ilmenee, että noin 72 prosenttia elinkeinotoiminnassa syntyneestä perimiskelpoisesta verovelasta on toimimattomilla yrityksillä. Rakentamisen toimialalla osuus on vieläkin suurempi, 80 prosenttia toimialan yritysten verovelasta on toimimattomilla yrityksillä. Suurin osa tästä verovelasta on osakeyhtiöillä, joiden toiminta on päättynyt. Yritystoiminnan päättymisen jälkeen kertymä verovelalle on hyvin epätodennäköistä. Yksityisellä elinkeinonharjoittajalla ja yhtymän vastuunalaisella yhtiömiehellä on henkilökohtainen vastuu elinkeinotoiminnassa syntyneistä veroveloista. Pieni osa verovelasta voi vielä kertyä toiminnan loppumisen jälkeen yrityksen omaisuudesta ulosmittauksen kautta tai konkurssipesän jakoosana. (Harmaan talouden selvitysyksikkö, 2012)

Taulukko 2. Osakeyhtiöiden verovelan maksimi, 99 % persentiili sekä mediaani.

Verovuosi	Maksimi	99 % persentiili	Mediaani
2017	343 923 494	262 024	2 563
2018	355 667 910	246 822	2 848
2019	9 846 845	247 479	3 038
2020	58 129 145	502 061	4 737

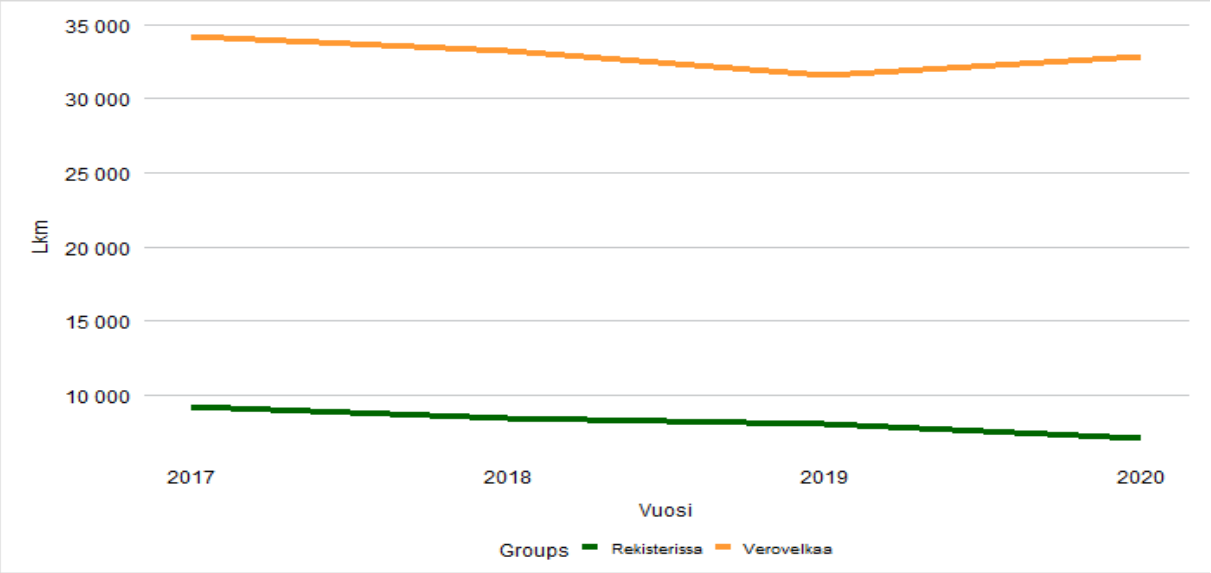
Persentiili eli sadannes- tai prosenttipiste kuuluu ns. fraktiileihin eli jakauman osuuspisteisiin. Se ilmoittaa muuttujan arvon, jonka alapuolelle jakaumassa jää tapauksista 99 % (99. persentiili).

3.2 Selvityksessä käytetty verovelka- ja tarkastusaineisto

Selvityksen perusjoukkona ovat toiminnassa olevat osakeyhtiöt Verohallinnon rekisterien mukaan. Toiminnassa olevia yrityksiä on vuosittain noin 200 000 ja lukumäärä on ollut hieman kasvussa vuosina 2017–2020. Selvityksessä toimivaksi yritykseksi katsotaan yritys, joka on tarkasteluaikana ainakin yhdessä seuraavista rekistereistä: ennakkoperintärekisteri, alv-rekisteri (liiketoiminta, kiinteistön käyttöoikeuden luovutus, alkutuotanto) tai työnantajarekisteri.

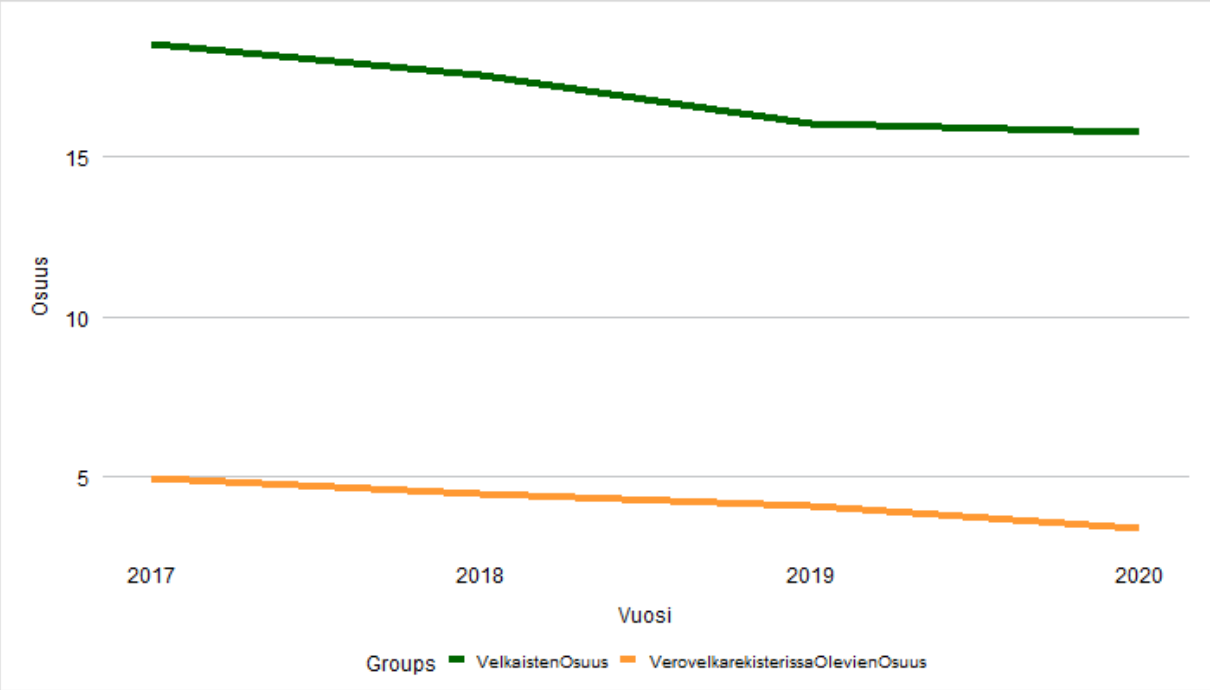
Verovelkaisia osakeyhtiöitä on sen sijaan ollut keskimäärin 35 000 kappaletta vuoden viimeisenä päivänä. Yritysten verovelan määrä on ollut silloin suurempi kuin 0 euroa. Verovelkarekisterissä on ollut vuosittain 7 000–9 000 yritystä. Molempien tarkasteltavien asioiden trendi on ollut laskeva vuosina 2017–2020. Ainoastaan vuonna 2020 verovelkaisten yritysten määrä nousi koronapandemian seurauksena verrattuna edelliseen vuoteen.

Kuvio 1. Verovelkaisten yritysten määrä vuoden viimeisenä päivänä sekä verovelkarekisterissä vuoden aikana olleiden yritysten lukumäärät.



Verovelkaisten yritysten suhteellinen osuus kaikista yrityksistä laski 18 prosentista 16 prosenttiin vuosina 2017–2020. Samana ajankohtana verovelkarekisterissä olleiden yritysten suhteellinen osuus kaikista yrityksistä väheni neljästä prosentista kahteen prosenttiin.

Kuvio 2. Verovelkaisten yritysten sekä verovelkarekisterissä vuoden aikana olleiden yritysten osuus kaikista aktiivisista osakeyhtiöistä.



Verotarkastukset

Selvityksen toinen keskeinen aineisto perustuu Verohallinnon tekemiin riskiperusteisiin ja satunnaisotannalla toteutettuihin verotarkastuksiin, jotka kohdistuvat vuosina 2017–2020 päättyneisiin tilikausiin. Verotarkastettuja yrityksiä on ajanjaksolla yhteensä 6 200 ja aineisto koostuu noin 11 000 havainnosta. Yksi verotarkastus kohdistuu usein useampaan vuoteen, jolloin jokainen yrityksen verotarkastettu vuosi esiintyy omana havaintonaan. Esimerkiksi jos yritys on tarkastettu vuosilta 2017–2018, niin yritys esiintyy aineistossa vuosien 2017 ja 2018 tarkastustulosten osalta.

Verotarkastusaineistojen käyttöön harmaan talouden määrän arvioinnissa on alan kirjallisuudessa suhtauduttu jossain määrin varauksellisesti. OECD:n julkaiseman harmaan talouden mittaamista koskevan käsikirjan mukaan ongelmana tarkastustulosten yleistämisessä on erityisesti se, ettei tarkastuskohteita valita sattumanvaraisesti vaan erilaisten riskitekijöiden tai häiriökäyttäytymisen perusteella. Tarkastukset eivät kata kaikkia toimialoja ja yritystyyppejä tasaisesti eikä tarkastuksilla välttämättä havaita kaikkea tosiasiassa esiintyvää harmaata taloutta. (OECD, 2002)

Verotarkastetut harmaan talouden yritykset ovat pääsääntöisesti liikevaihdoltaan alle 10 miljoonaa euroa. Näistäkin merkittävä osa on yrityksiä, joiden liikevaihto on alle 2 miljoonaa euroa. Tämän lisäksi on suuri määrä yrityksiä, jotka harjoittavat harmaan talouden toimintaa, mutta joista Verohallinnolla ei ole liikevaihtotietoa. Tällaisia ovat esimerkiksi ulkomaiset yritykset sekä lepääviksi tai lakkautetuiksi ilmoitetut yritykset. (Harmaa talous & talousrikollisuus, 2021)

Tässä selvityksessä pyritään ensisijaisesti verotarkastusaineistojen avulla päättelemään harmaan talouden toiminnan riski kullekin tarkastamattomalle yritykselle, jonka avulla päästään arvioon tarkastamattomien, verovelkaisten yritysten mahdollisesta kytköksestä harmaan talouden toimintaan. Tämän lisäksi tarkastettujen yritysten alajoukon sisällä pyritään päättelemään verovelkaisuuden selitysvoimaa suhteessa harmaan talouden toimintaan indikoivaan tarkastustulokseen. Näidenkin arvioiden laatuun vaikuttaa luonnollisesti verotarkastusaineiston valikointiharha. Valikointiharhan aiheuttama ongelma ei ole kuitenkaan niin iso kuin tutkimuksissa, jotka pyrkivät arvioimaan harmaan talouden kokoa tarkastustulosten perusteella.

3.3 Verotarkastuksen lopputuloksen vaikutuksesta verovelkaisuuteen

Verotarkastetut yritykset ovat muita yrityksiä useammin verovelkaisia. Lisäksi ne ovat saaneet verovelkarekisterimerkinnän suuremmalla todennäköisyydellä kuin muut yritykset. Koska tarkastusaineisto kattaa useamman vuoden ja yksi yritys voi esiintyä siinä useamman kerran, esitettyjä lukuja ei pidä tulkita prosenttina tarkastetuista yrityksistä vaan prosenttina kaikista yrityksiin kohdistuneista tarkastusvuosista ajalla 2017–2020.

Taulukko 3. Verotarkastettujen yritysten verovelkaisuus vuosina 2017–2020. Yritysten luokittelu tarkastuksen lopputuloksen mukaan.

Verotarkastuksen lopputulos	Verovelkaisten osuus	Verovelkarekisterissä olevien osuus	Verovelan mediaani	Rekisterin saldon mediaani
Harmaa	37,5 %	16,2 %	8 887 €	29 278 €
Valkoinen	23 %	8,8 %	9 103 €	25 352 €

Verotarkastettu yritys on määritelty harmaan talouden yritykseksi, jos verotarkastuksen yksi tai useampi löydös viittaa harmaan talouden tekoon. Vaihtoehtoisesti verotarkastuksessa maksettavaksi määrättyjen verojen euromäärä on ylittänyt 10 000 euroa, ja löydös liittyy mahdollisesti harmaaseen talouteen.

Vaihtoehtoisesti verotarkastettu yritys on määritelty valkoiseksi yritykseksi, jos yksikään verotarkastuksen löydös ei viittaa harmaan talouden tekoon. Vaihtoehtoisesti tarkastuksen aikana todennettu löydös on määritelty liittyvän mahdollisesti harmaaseen talouteen, mutta maksettavaksi määrättyjen verojen euromäärä ei ole ylittänyt 10 000 euroa.

Verovelkaisten yritysten osuus kaikista yrityksistä on ollut 15–20 prosenttia[5] kalenterivuoden viimeisinä päivänä vuosina 2017–2020. Tähän nähden verotarkastusten perusteella harmaan talouden yrityksiksi todettujen yritysten lähes 40 prosentin verovelkaisuus on huomattavan korkea. Verotarkastuksissa verotussäännöksiä noudattavaksi todetuista yrityksistä vajaa neljännes oli verovelkaisia, mikä sekin on verotarkastamattomia yrityksiä korkeampi osuus.

Verotarkastuksissa harmaan talouden yrityksiksi todetuista on ollut 16 prosenttia verovelkarekisterissä. Sen sijaan valkoisella yrityksellä on ollut puolet pienempi riski eli kahdeksan

[5] Tähän lukuun eivät sisälly ne yritykset, jotka ovat maksaneet verovelkansa pois kalenterivuoden aikana.

prosentin todennäköisyys verovelkarekisteriin joutumiselle. Sekä harmaat että valkoiset verotarkastetut yritykset ovat olleet verovelkarekisterissä merkittävästi todennäköisemmin kuin tarkastamattomat yritykset.

Vuoden lopussa oleva verovelan mediaani on verotarkastusten lopputulosten perusteella valkoisiksi yrityksiksi luokitelluilla noin 9 000 euroa. Harmaan talouden yritysten verovelan mediaani oli muutaman sataa euroa pienempi. Verotarkastamattomien yritysten verovelan mediaani on vain noin 3 000 euroa eli vain kolmasosa tarkastettujen yritysten vastaavasta.

Verovelkarekisterin saldon mediaani on verotarkastetuilla harmailla yrityksillä 29 000 euroa. Verotarkastettujen valkoisten yritysten verovelan saldo jäi 4 000 euroa pienemmäksi. Kaikkien verovelkarekisteriin joutuneiden yritysten verovelkasaldo oli sen sijaan hieman tätä pienempi (noin 22 000 euroa).

4 Verovelan yhteys harmaan talouden todennäköisyyteen

Tässä luvussa tarkastellaan, miten verovelka selittää harmaan talouden verotarkastuksiksi luokiteltujen tarkastusten lopputulosta. Tarkastusaineistoa on käytetty verovelan ja verovelkarekisterissä olon harmaan talouden selitysvoimaisuuden arvioimiseen. Tällä tavalla on voitu tehdä yleistyksiä verovelan ja harmaan talouden yhteydestä, vaikka asiaa voidaan tutkia pelkästään riskiperusteisesti tarkastuskohteeksi valikoituneiden yritysten joukossa.

Yrityksen verovelkaisuuden ja harmaan talouden välille ei ole pyritty tekemään kausaalipäättelyä, jolloin ensimmäinen tapahtuma olisi syy ja toinen sen seuraus. Muuttujien välinen vaikutussuhde voi hyvinkin kulkea molempiin suuntiin. Yritys voi käyttää verovelkojen maksamatta jättämistä yhtenä harmaan talouden toiminnan työkaluna. Yhtä lailla taloudelliset ongelmat ja siitä johtuvat verovelat voivat johtaa yrityksen myös harmaan toiminnan poluille, kun verotussäännöksiä noudattamalla yritys ei enää selviydy velvoitteiden hoitamisestaan. Kausaalipäättelyn sijaan on tutkittu verovelan sekä harmaan talouden toiminnan yhteyttä eli kuinka usein havaitsemme molempia samanaikaisesti.

Verovelan merkitystä on selvitetty kahden tilastollisen menetelmän avulla. Ensimmäinen on bayesilaisittain toteutettu logistinen regressiomalli ja jälkimmäinen on Shapley-arvot. Kummassakin mallissa on huomioitu joukko muitakin harmaata taloutta ennustavia muuttujia. Verovelan vakioimisella muiden muuttujien suhteen on saatu poistettua estimaatista harhaa. Harhan olemassaolo johtuu siitä, että jokin kolmas muuttuja voi ennustaa voimakkaasti sekä harmaan talouden toimintaa että verovelan esiintymistä. Esimerkiksi alhainen liikevaihto voi olla tällainen muuttuja. Näin jäljelle jää verovelan oma itsenäinen ennustevoima suhteessa harmaaseen talouteen.

Logistisen regressiomallin avulla voidaan tehdä päätelmiä verovelkaisuuden selitysvoimasta. Logistinen regressioanalyysi[6] on käyttökelpoinen silloin, kun selitettävän muuttujan arvot rajoittuvat vain kahteen vaihtoehtoon (esimerkiksi kyllä/ei). Logistinen regressioanalyysi ei pyri ennustamaan määriä, vaan todennäköisyyksiä. Kyse on siis siitä, millä todennäköisyydellä tarkasteltavana oleva asia tapahtuu tai ei tapahdu.

Shapley-arvojen avulla voidaan avata kunkin valitun muuttujan vaikutus lopulliseen ennusteeseen riippumatta mallintamiseen käytetystä algoritmista. Tämä mahdollistaa monimutkaisempien ja ennustevoimaisempien algoritmien käyttämisen, jotka sallivat myös monimutkaiset, epälineaariset assosiaatiot muuttujien välillä. Tässä yhteydessä käytetty XGBoost-päätöspuumalli on suunniteltu hyvin tehokkaaksi koneoppimisen menetelmäksi, joka ei tee oletuksia aineiston eikä assosiaation muodoista.

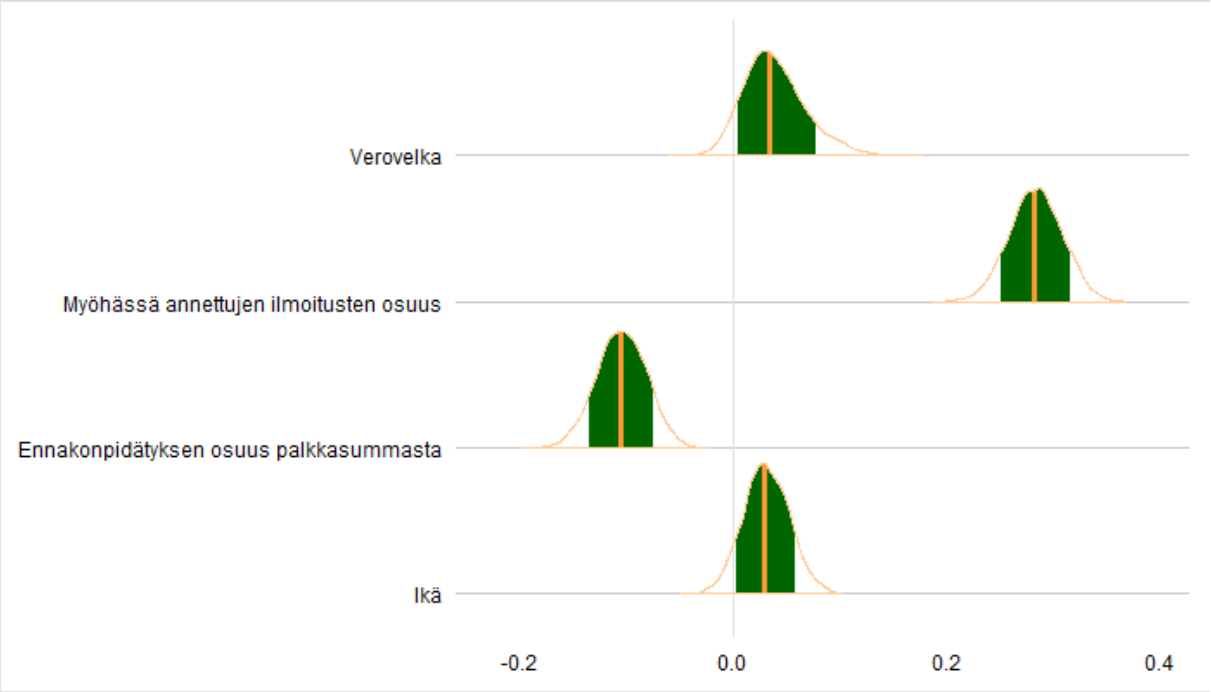
[6] Bayesilainen regressio on valittu perinteisen frekventistisen version sijaan yksinkertaisesti siksi, että sen lopputulos on yksinkertaisempi tulkita sekä visualisoida. Käytännön vaikutusta tehtäviin päätelmiin valinnalla ei ole.

4.1 Selitysvoimaisuus logistista regressioanalyysia käyttäen

Selvityksessä on rakennettu kaksi logistista regressiomallia, joissa on verovelan tai verovelkarekisterisaldon lisäksi kymmenen muuta harmaata taloutta ennustavaa muuttujaa. Muiden muuttujien ottamisella mukaan malliin on saatu verovelan euromäärän tai verovelkarekisterin saldon vaikutus vakioitua[7]. Lisäksi malleista on tehty versiot, joissa verovelkaisuus ja verovelkarekisterissä olo annettiin mallille kategorisena "kyllä/ei" -muuttujana. Kaikki mallien muuttujat ovat skaalattu samalle asteikoille, jolloin eri muuttujien selitysvoimaisuutta voidaan suoraan verrata.

Regressiokertoimen arvo 0 tarkoittaa, ettei selittävän ja selitettävän muuttujan välillä ole tilastollista yhteyttä. Mitä kauempana regressiokertoimen arvo on nollasta, joko positiiviseen tai negatiiviseen suuntaan, sitä voimakkaampi assosiaatio on. Alla olevasta kuviosta 4 nähdään verovelan ja kolmen muun muuttujan regressiokertoimien piste-estimaatti kuvattuna oranssilla palkilla sekä 80 prosentin uskottavuusväli (väli, jossa kertoimen oikea arvo sijaitsee 80 prosentin todennäköisyydellä).

Kuvio 3. Logistisen regression regressiokertoimet verovelan ja kolmen muun muuttujan osalta.

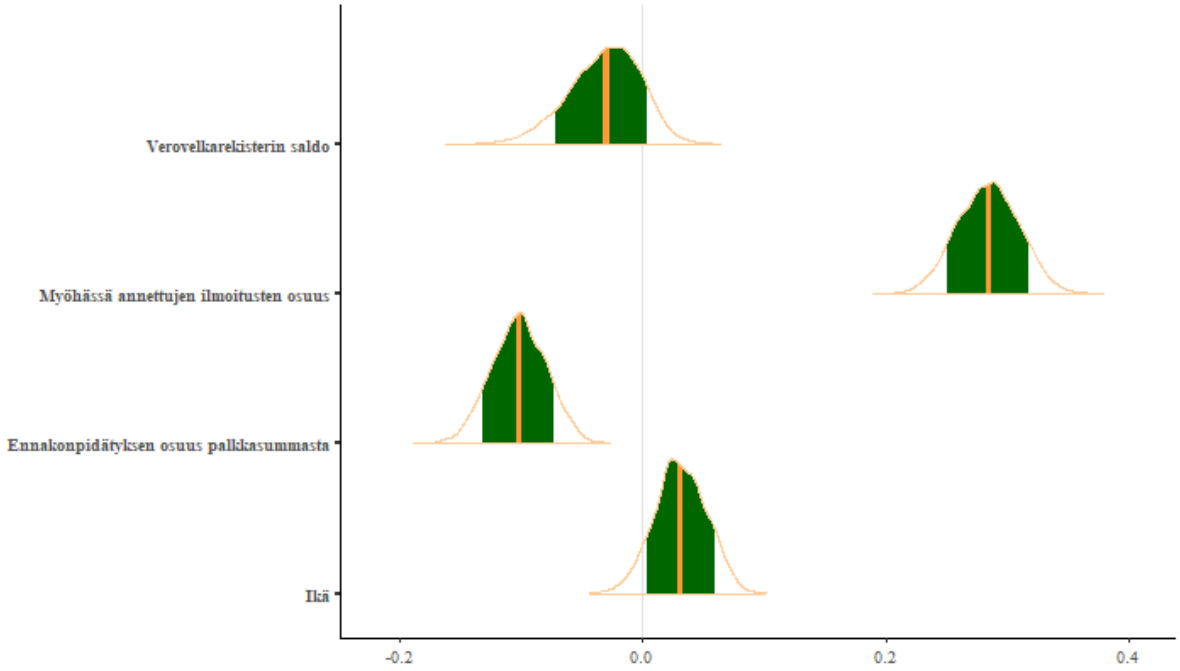


Kun verovelan olemassaolo ovat vakioitu, sen regressiokertoimen estimaatti osuu suhteellisen lähelle nollaa ja uskottavuusväli kattaa arvoja sen molemmin puolin. Piste-estimaatti verovelan regressiokertoimelle on 0,035 ja kerroin on noin 93 prosentin todennäköisyydellä positiivinen. Mallilla ei siis ole täyttä varmuutta, että olemassa oleva assosiaatio olisi

[7] Esimerkiksi alhainen liikevaihto voi korreloida sekä erääntyneen verovelan että harmaan talouden kanssa. Se voi indikoida joko ilmoittamatta jääneitä tuloja tai aitoja talousvaikeuksia, jotka molemmat voivat johtaa erääntyneen verovelan syntymiseen. Jos liikevaihto jätetään mallista pois, voisi verovelka napata osan liikevaihdon selitysvoimasta itselleen. Tällöin verovelka näyttäytyisi paremmin harmaata taloutta selittävänä muuttujana kuin mitä se todellisuudessa on.

positiivinen. Se tarkoittaisi, että lisääntyvä verovelka johtaisi lisääntyvään harmaan talouden toiminnan riskiin. Verovelkarekisterin saldon vaikutus näyttäisi sen sijaan jopa alentavan harmaan talouden toiminnan riskiä muiden muuttujien vakioinnin jälkeen.

Kuvio 4. Logistisen regression regressiokertoimet verovelkarekisterisaldon ja kolmen muun muuttujan osalta.



Sekä verovelan määrä että verovelkarekisterin saldo vertautuvat harmaan talouden ennustusteholtaan yrityksen ikään, joka sekin on heikko indikaattori. Jos verovelkaisuutta olisi käsitelty kategorisena muuttujana[8], muuttujan selitysvoima olisi entistä heikompi. Selkeästi selitysvoimaisempia harmaan talouden muuttujia ovat esimerkiksi myöhässä annettujen ilmoitusten suuri osuus ja ennakonpidätyksen pieni osuus ilmoitetusta palkkasummasta.

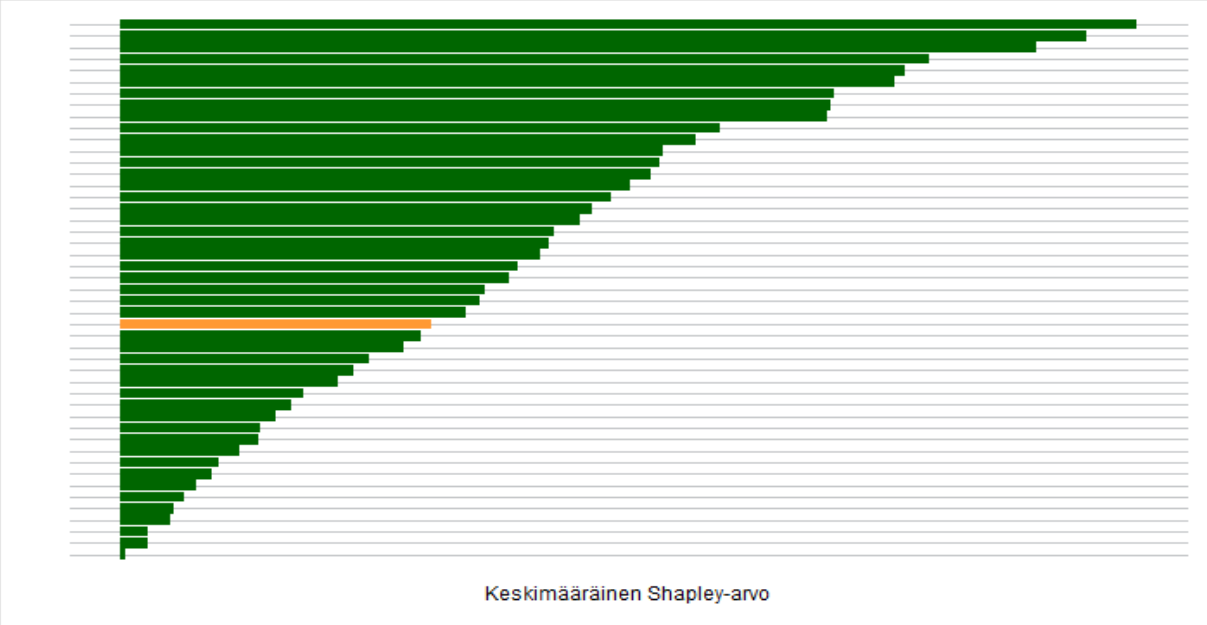
Verovelka on siis logistisessa regressiomallissa heikohko muuttuja. Muiden muuttujien huomioimisen jälkeen sen harmaan talouden selitysvoimaisuus jää suhteellisen alhaiseksi, mutta ei kuitenkaan olemattomaksi. Verovelkarekisteriin kuuluminen vuoden aikana on verovelkaakin heikompi muuttuja. On mielenkiintoista huomata, että vakioinnin jälkeen verovelkarekisteriin kuuluminen vähentäisi harmaan talouden toiminnan todennäköisyyttä. Koska vaikutus on niin heikko, on luultavaa, että tämä on pienestä aineistosta johtuva sattumanvarainen löydös. Tätä tukee myös hyvin leveä uskottavuusväli, joka sijoittuu nollan molemmin puolin.

[8] Kategorisella muuttujalla tarkoitetaan muuttujaa, joka voi saada vain rajatun määrän eri arvoja. Esimerkiksi yrityksen toimiala voi saada arvoja vain virallisen toimialaluokituksen mukaisesti. Verovelka itsessään on numeerinen muuttuja, jossa yrityksen verovelka voi olla kaikkea yhden ja miljoonien eurojen välillä. Verovelka voidaan muuttaa kategoriseksi muuttujaksi esimerkiksi niin, että yritykset, joilla on verovelkaa yli nolla euroa, saavat arvon "kyllä" ja muut saavan arvon "ei".

4.2 Selitysvoimaisuus Shapley-arvoilla pääteltynä

Shapley-arvoilla voidaan arvioida muuttujan selitysvoimaa suhteessa muihin muuttujiin sekä tutkia tarkemmin, miten yksittäinen muuttuja vaikuttaa lopulliseen ennusteeseen. Muuttujien merkityksellisyyden arvioinnissa on käytetty keskimääräistä Shapley-arvoa, jonka avulla muuttujat on voitu asettaa tärkeysjärjestykseen sen mukaan, kuinka paljon ne osallistuvat lopulliseen XGBoost-mallin ennusteeseen. Tämä järjestys on nähtävissä alla olevassa kuviossa 5 niin, että muuttujien nimet on häivytetty. Verovelka on näytetty oranssilla.

Kuvio 5. Muuttujien merkityksellisyys Shapley-arvoilla mitattuna.



Kuviosta voi nähdä, että Shapley-arvojen näkökulmasta verovelka näyttäytyy varsin keskimääräisesti harmaata taloutta ennustavana muuttujana. Verovelka on kaukana ennusteisiin eniten vaikuttavista muuttujista. Yrityksen verovelkaisuus ei ole kuitenkaan merkityksetön ja mallista löytyy useampi sitäkin heikompi muuttuja.

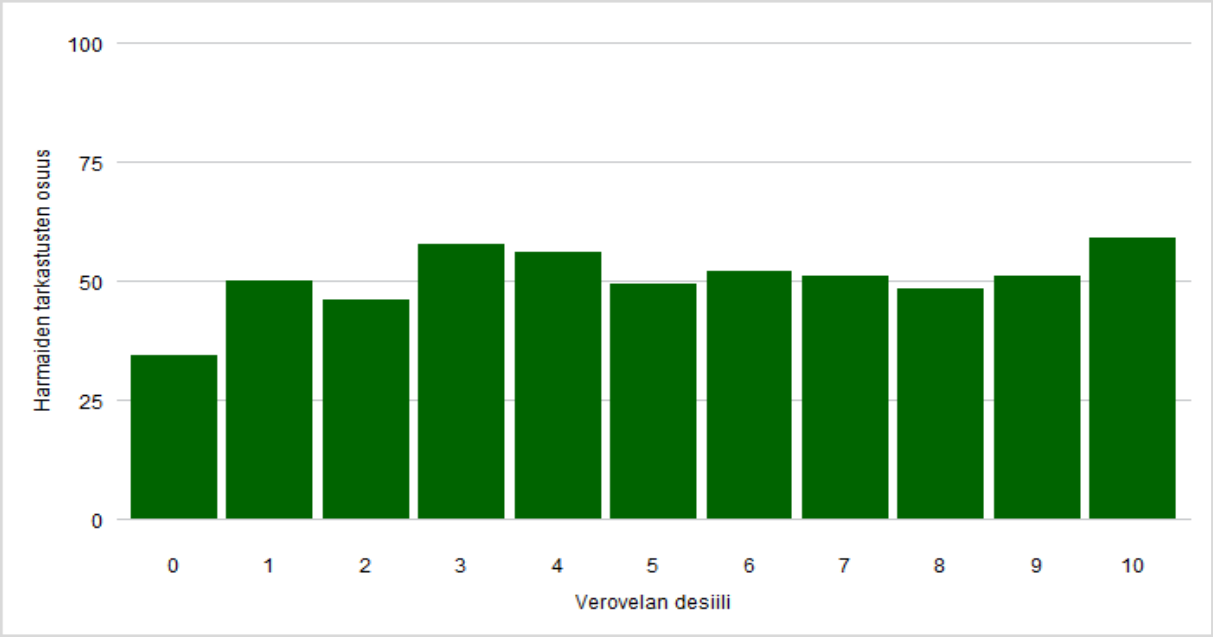
Johtopäätös on se, että yrityksen verovelkaisuus ei näytä johtavan merkittävästi kohonneeseen harmaan talouden riskiin. Yrityksen verovelkaisuus antaa kuitenkin jotain indikaatiota yrityksen riskisyydestä. Sillä on oma rajattu merkitys arvioitaessa yrityksen harmaan talouden toiminnan todennäköisyyttä. Sama johtopäätös on tehtävissä verovelkarekisteriin kuulumisen osalta.

Verovelan epälineaarinen suhde harmaaseen talouteen

XGBoost-koneoppimismallin ja Shapley-arvojen avulla voi huomata, että verovelan assosiaatio harmaiden verotarkastusten kanssa ei ole lineaarinen vaan muistuttaa enemmän paraabelia. Aluksi verovelan kasvaessa harmaan tarkastuksen todennäköisyys kasvaa, sen jälkeen laskee keskimääräisillä verovelkasaldoilla ja aivan isoimmilla verovelan määrillä taas kasvaa.

Alla olevassa kuviossa näytetään verotarkastukset luokiteltuna yrityksen verovelan desiilin mukaan. Verovelattomat ovat desiilissä 0. Suurimmat harmaiden yritysten osuudet ovat desiileissä 3,4 ja 10. Huomionarvoista kuitenkin on, että harmaan talouden yrityksiksi luokiteltujen osuus on selvästi korkeampi ihan pienimmissäkin verovelan desiileissä kuin verovelattomien joukossa.

Kuvio 6. Verovelka desiileittäin ja harmaiden tarkastusten osuus kaikista.



Aikaisempia tuloksia tulkittaessa on muistettava käytetyn aineiston aiheuttamat rajaukset. Totuus yrityksen harmaan talouden toiminnasta tunnetaan vain hyvin pienen tarkastettujen yritysten osajoukon osalta. Lisäksi tarkastuskohteet on valikoitu pääsääntöisesti riskiperusteisesti tai valvonnan havaintojen perusteella, jolloin aineisto on vinoutunut ja se kuvaa huonosti koko yritysjoukkoa.

Itse verovelankin suhteen tarkastettujen yritysten aineisto on vinoutunut suhteessa koko yrityskantaan, sillä tarkastettujen yritysten joukossa on suhteellisesti paljon enemmän verovelkaisia yrityksiä kuin koko yrityspopulaatiossa. Tällöin verovelan assosiaatio tulee luultavasti aliarvioiduksi. Aineiston vinoumista seuraa, että yllä nähdyn perusteella ei voida tehdä kovin pitkälle meneviä johtopäätöksiä muuttujien välisistä assosiaatioista harmaan talouden toiminnan kanssa.

Verovelan yhteys harmaaseen talouteen on todennäköisesti ennustettua suurempi. Tämä johtuu siitä, että jo valmiiksi verovelkaiset yritykset joutuvat verotarkastuksen kohteeksi useammin kuin velattomat yritykset. Lisäksi harmaan talouden ennustevirhettä ylöspäin lisää epäilemättä se, että tarkastukseen joutuneilla harmaan talouden piiriin kuulumattomilla yrityksillä oli myöskin suhteellisesti useammin verovelkaa kuin keskimääräisellä yrityksellä.

Tässä luvussa on haettu vastausta kysymykseen, miten yrityksen verovelkaisuus liittyy yrityksen harmaan talouden riskiin. Johtopäätöksenä voidaan todeta, että verovelkaisuus näyttää liittyvän harmaan talouden toimintaan. Verovelka ja harmaan talouden toiminta kulkevat

käsi kädessä, mutta molempiin vaikuttaa lukuisa joukko muita muuttujia, jolloin verovelan itsenäinen ennustevoima jää maltilliseksi. Siitä, onko yrityksen verovelkaisuus harmaan talouden toiminnan syy vaiko seuraus, ei voida tehdyn tarkastelun perusteella sanoa mitään.

5 Harmaan talouden yritysten suhteellinen verovelkaisuus ja niiden osuus kokonaisveroveloista

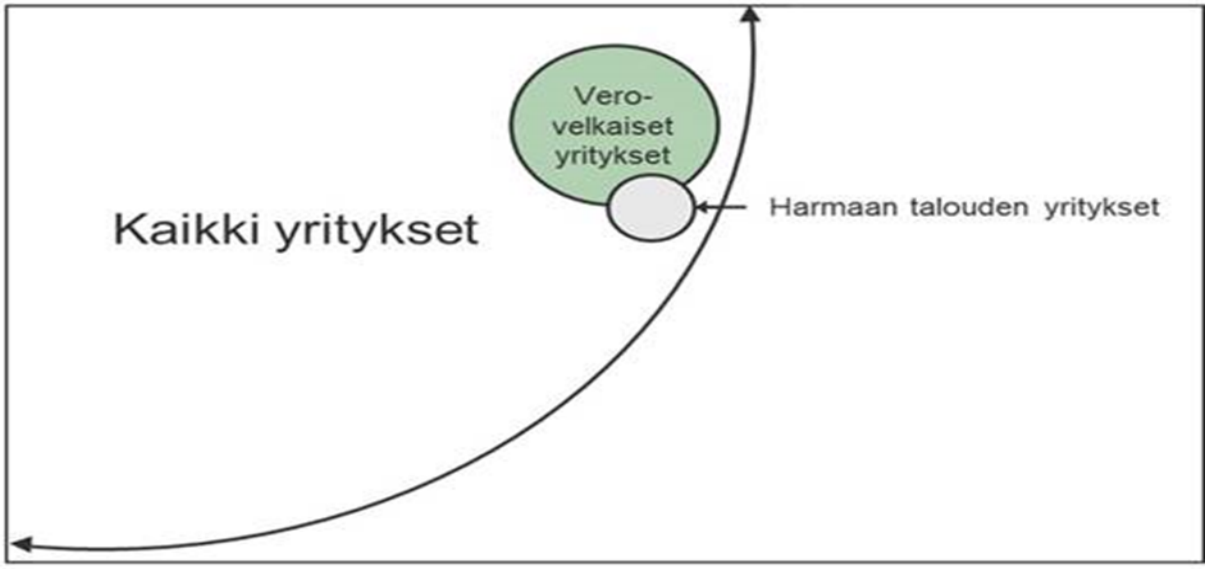
Tämän osion tavoitteena on hyödyntää koneoppimismallia sen selvittämisessä, millä osalla kaikista toimivista yrityksiä on havaittavissa merkkejä harmaan talouden toiminnasta ja mikä on kyseisten yritysten osuus kaikista veroveloista. Verovelkaa koskeva tarkastelu on tehty eri yritysluokille sekä verovelan olemassaoloon että verovelkarekisteriin kuulumiseen perustuen.

Suurin osa verovelkaisista yrityksistä on verotarkastamattomia, jolloin niiden mahdollista osallisuutta harmaan talouden toimintaan ei ole todennettu. Hyödyntämällä verotarkastusten lopputuloksia ja havaintoja voidaan päätellä jotain myös tarkastamattomista yrityksistä ja niiden harmaan talouden toiminnan riskistä. Kun koneoppimismalli on opetettu tunnistamaan harmaan talouden yritys verotarkastettujen yritysten joukossa, voidaan tätä mallia hyödyntää tietyin varauksin myös tarkastamattomien yritysten joukkoon harmaan talouden riskisyyden päättelemiseksi.

Verotarkastuksia on valmistunut suurempi määrä tarkastelussa olleille varhaisemmille vuosille. Harmaan talouden tarkastukset ovat kestoltaan keskimäärin pidempiä kuin tarkastukset, joissa ei ole havaittu harmaata taloutta. Mitä lähemmäksi nykyhetkeä sen sijaan tullaan, sitä vähemmän on suhteellisesti harmaan talouden tarkastuksia. Syynä tälle muutokselle on se, että harmaan talouden yritysten tarkastukset eivät ole monesti vielä ehtineet valmistumaan.

Verotarkastetut yritykset eivät ole kuvaava otos kaikkien yritysten ja verovelkaisten yritysten joukosta. Suurin osa näistä yrityksistä on valikoitunut tarkastuksen kohteeksi riskiperusteisesti ominaisuuksiensa vuoksi. Tämä johtaa siihen, että harmaan talouden toiminta näyttäytyy suhteellisesti paljon yleisempänä kuin se oikeasti on kaikkien yritysten joukossa. Verotarkastettujen yritysten joukosta tehdyt yleistykset kaikkien yritysten joukkoon ovat siten vinoutuneita ja mallin toiminta tarkastettujen yritysten ulkopuolella epäluotettavampaa.

Kuvio 7. Yhteiskunnassa toimivien harmaan talouden yritysten määrän estimointi.

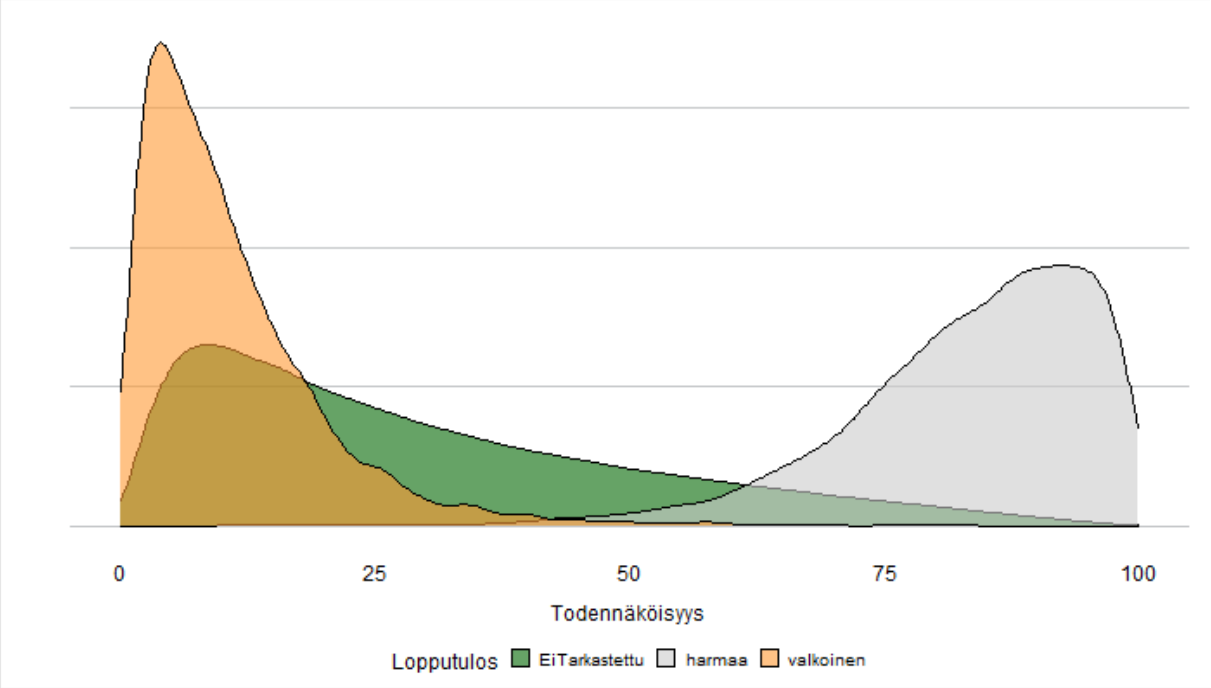


Malli antaa kullekin yritykselle oman käsityksensä mukaisen todennäköisyyden sille, että kyseinen yritys on harmaan talouden toimija kaikkien yritysten joukossa. Lopputuloksena ei ole suoraviivainen jaottelu kahteen luokkaan vaan todennäköisyysjakauma, jota voidaan hyödyntää monin tavoin. On muistettava, että kunkin yrityksen saama todennäköisyys ei vastaa todellisuutta, vaan on vain mallin käyttämien tietojen mukainen ennuste.

5.1 Kaikkien toimivien yritysten luokittelu

Toiminnassa olevat yritykset on jaettu mallin ennustaman harmaan talouden todennäköisyyden perusteella kolmeen luokkaan: 1) valkoiset 2) harmaat 3) rajatapaukset. Valkoisten yritysten alajoukkoon kuuluvat kaikki yritykset, joiden mallin mukainen todennäköisyys olla harmaa on alle 50 prosenttia tai jotka on verotarkastuksessa todettu verotussäännöksiä noudattaviksi. Harmaiden yritysten joukossa ovat vastaavasti ne yritykset, joiden todennäköisyys harmaan talouden toimintaan on mallin perusteella yli 80 % tai jotka on verotarkastuksissa todettu harmaan talouden yrityksiksi. Loput tapaukset menevät rajatapausten luokkaan. Kolmiosainen luokittelu antaa mahdollisuuden luokitella osan kaikista toiminnassa olevista yrityksistä melko todennäköisesti harmaiden ja valkoisten yritysten osajoukoiksi.

Kuvio 8. Verovelkaisten yritysten alaluokkien muodostaminen harmaan talouden toiminnan todennäköisyysjakaumiin perustuen.



Yllä oleva kuvio näyttää mallin antaman todennäköisyysjakauman kullekin tarkastustuloksen mukaan luokitellulle yritysjoukolle. Koska mallille on annettu opetusvaiheessa verotarkastuksissa harmaaksi ja valkoisiksi todettujen yritysten tiedot, osaa se päätellä niille totuutta kuvaavan arvion riskisyydestä.

Tarkastamattomien yritysten osalta malli kokee, että valtaosalla yrityksistä on vain pieni, alle 25 prosentin todennäköisyys olla harmaan talouden toimija. Jakauman oikea häntä kuitenkin pysyy verrattain paksuna pitkään, eli malli arvioi melko suuren yritysjoukon hyvin riskiseksi. Jos selvityksessä olisi ollut käytettävissä laaja satunnaistarkastusaineisto, olisi jakauma luultavasti vieläkin enemmän painottunut alhaisiin todennäköisyysarvioihin.

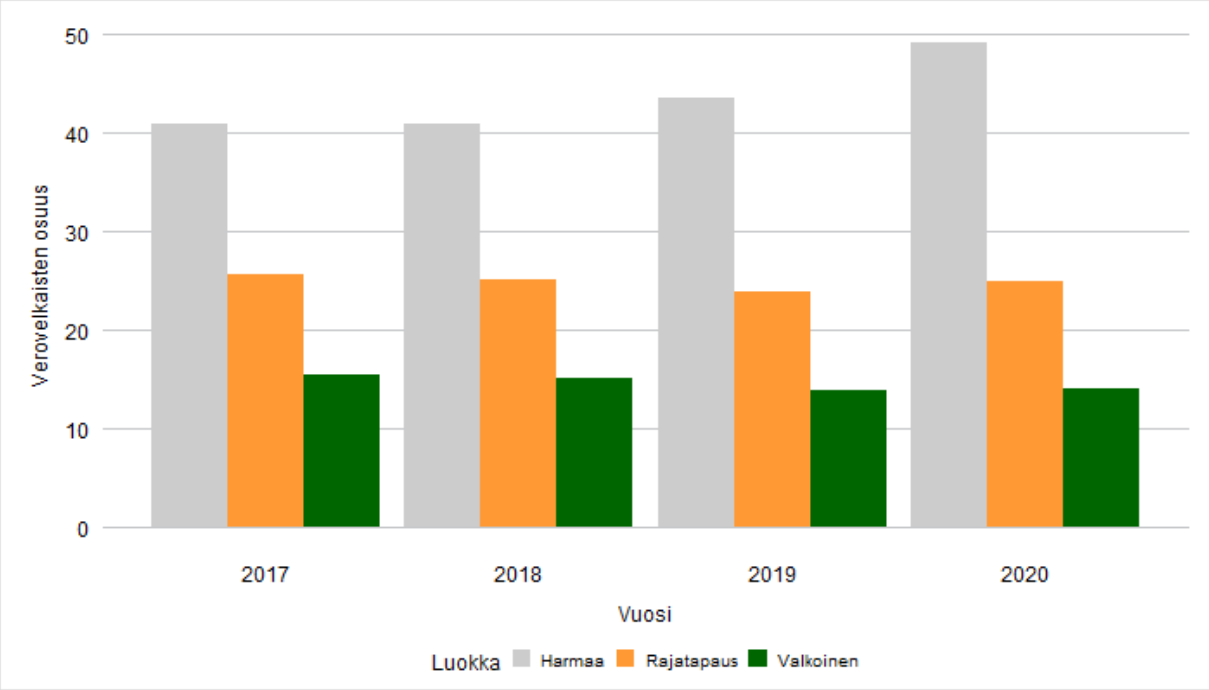
Ennustemallin algoritmina on käytetty tässä yhteydessä aikaisemmin hyödynnettyä XGBoostia. Malli toimii riittävän hyvin tarkastettujen yritysten joukossa ja sen ennustevoimaisuutta kuvaava mittari AUC[9] (area under curve) on kelvollinen 0,73 tarkastettujen yritysten testijoukkoon ennustaessa.[10]

[9] AUC eli area under curve on luokittelumallin hyvyyttä mittaava luku, joka on riippumaton valitusta luokittelun raja-arvosta. Se saa arvoja 0,5 sekä 1 väliltä, jossa ensimmäinen vastaa kolikonheittoa ja toinen täydellistä mallia.
[10] Koneoppimismallia rakentaessa on tyypillistä jakaa aineisto kahteen osaan niin, että malli opetetaan näyttämällä niin sanottuun harjoitusaineistoon kuuluvat havainnot. Tämän jälkeen mallin toimivuutta mitataan ennustamalla sovitetun mallin avulla irralliseen testijoukkoon, johon kuuluu havaintoja, joita malli ei ole koskaan aikaisemmin nähnyt.

5.2 Verovelan jakautuminen muodostetuissa yritysluokissa

Koneoppimismallin harmaiksi luokitelluista yrityksistä on ollut keskimäärin verovelkaisia noin 40 prosenttia vuoden lopussa. Ainoastaan poikkeusvuonna 2020 osuus on lähemmäs 50 prosenttia. Valkoisten yritysten joukossa verovelkaisten osuus on reilusti alle puolet harmaiden yritysten vastaavasta (valkoisista noin 15 prosenttia verovelkaisia). Rajatapausten joukossa verovelkaisuutta on ollut joka neljännellä yrityksellä vuosittain.

Kuvio 9. Verovelkaisten yritysten suhteellinen osuus omassa alaluokassa (harmaat yritykset, rajatapaukset ja valkoiset yritykset).

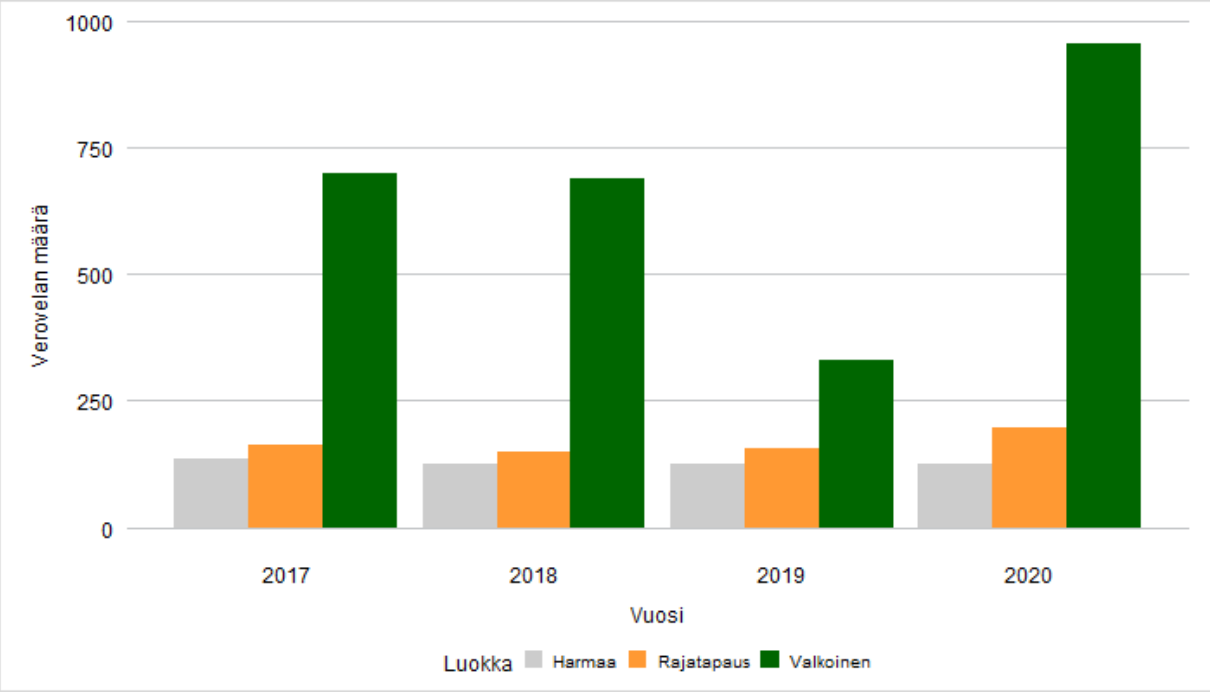


Mallin mukaan suurin osa verovelasta on kuulunut odotetusti valkoisille yrityksille vuosina 2017–2020. Vuosien 2017–2018 verovelan määrän putoaminen 740 miljoonasta eurosta 300 miljoonaan euroon vuonna 2019 ja vuoden 2020 hetkellinen uudelleennousu lähemmäs miljardiin euroon tapahtuivat lähes täysin valkoisten yritysten joukossa.

Kuten verovelka- ja tarkastusaineiston tarkastelun kohdalla aikaisemmin on huomattu, niin suuri osa kokonaisveroveloista voi johtua yksittäisten yritysten veloista. Äärimmäisenä esimerkkinä on yhden yrityksen yli 300 miljoonan euron verovelka vuosina 2017–2018, jonka yritys maksoi pois vuonna 2019. Tämä näkyy selkeästi valkoisten yritysten verovelan kokonaismäärän romahtamisena vuonna 2019. Valkoisten yritysten kokonaisverovelan kasvu 700 miljoonalla eurolla seuraavana vuonna johtuu puolestaan koronapandemian alkamisesta.

Sen sijaan verovelan euromäärä on pysynyt suhteellisen samana noin 200 miljoonassa eurossa sekä harmaiden että rajatapauksiksi tulkittujen yritysten joukoissa kaikkina tarkasteluvuosina. Edes koronapandemian alkamisvuotena 2020 ei näiden yritysten osalta ole havaittavissa merkittävää verovelan määrän kasvua.

Kuvio 10. Verovelan jakaantuminen euromääräisesti harmaille yrityksille, rajatapauksille ja valkoisille yrityksille.

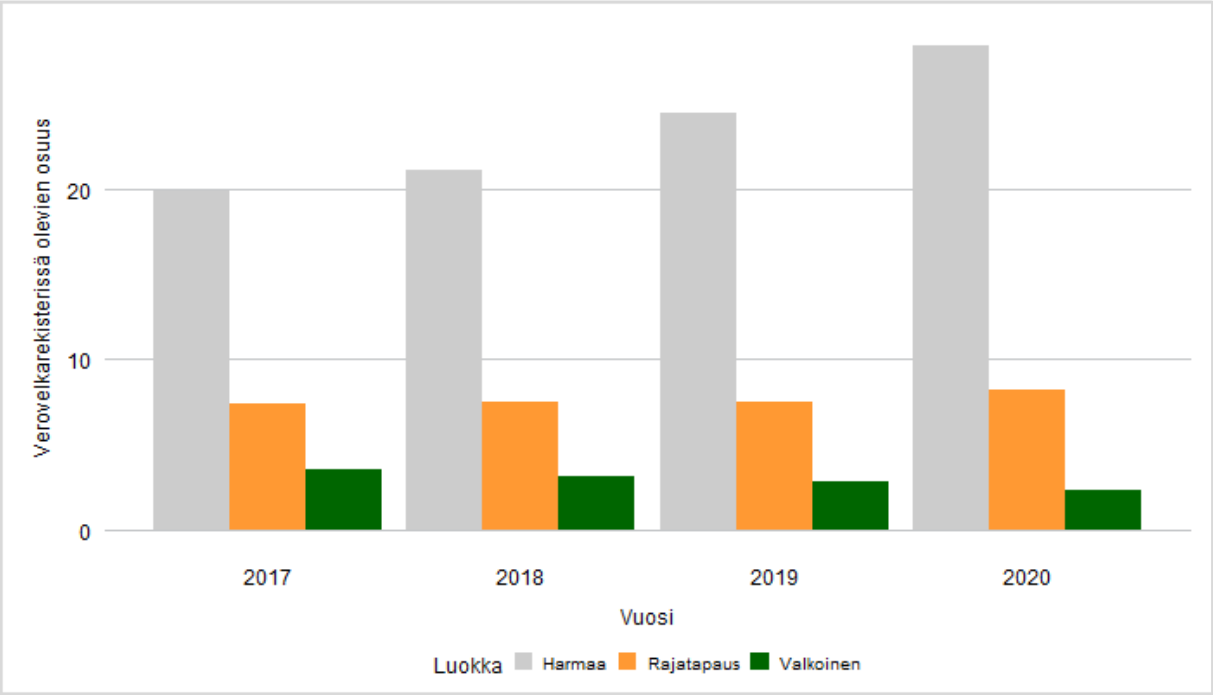


5.3 Verovelkarekisteriin kuuluminen eri yritysluokissa

Verovelkarekisteriin on kuulunut 20–30 prosenttia mallin harmaiksi yrityksiksi luokittelemista yrityksistä vuosina 2017–2020. Rajatapausten suhteellinen osuus verovelkarekisteriin kuulumisessa säilyy hieman yli viiden prosentin tasolla vuodesta toiseen. Valkoisten yritysten osuus on ollut sen sijaan 2–3 prosentissa.

Tämän perusteella verovelkarekisteriin kuuluminen on mallin mukaan harmaan talouden yritykselle puolet harvinaisempaa kuin aikaisemmin tarkasteltu verovelan olemassaolo vuoden viimeisenä päivänä. Valkoisista yrityksistä ja rajatapausyrityksistä joutuu verovelkarekisteriin vain hyvin pieni osa verrattuna verovelan olemassaolon todennäköisyyksiin.

Kuvio 11. Verovelkarekisteriin kuuluneiden yritysten suhteellinen osuus eri alaluokissa (harmaat yritykset, rajatapaukset ja valkoiset yritykset.



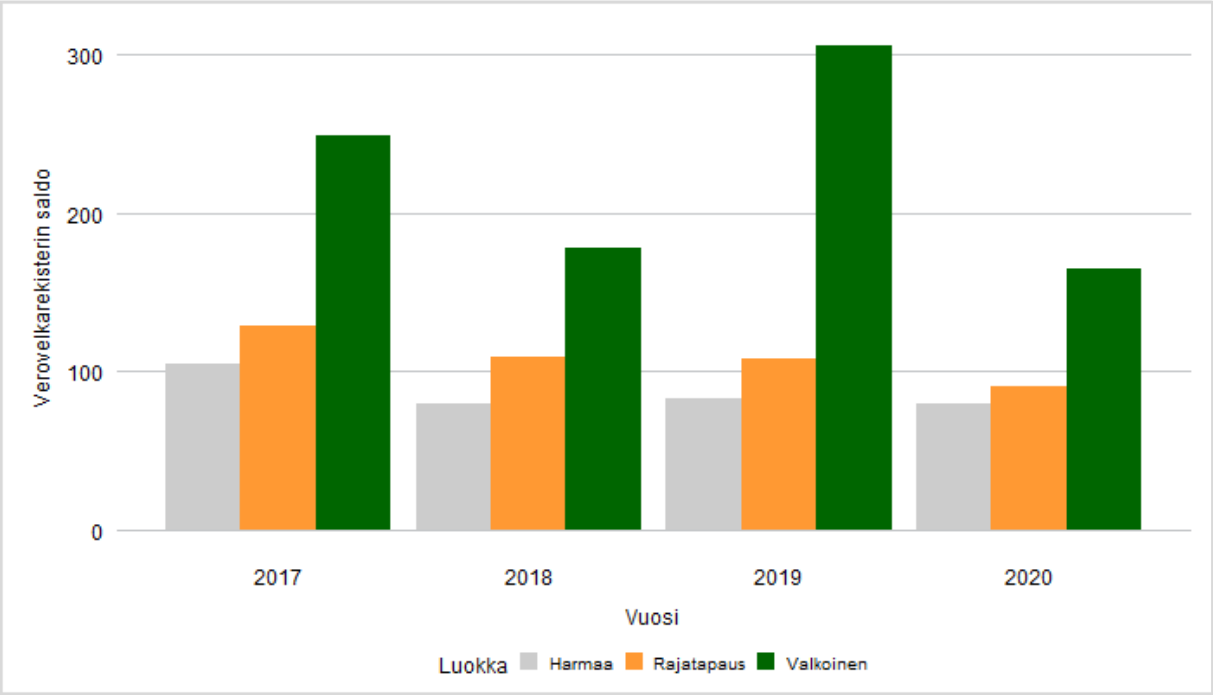
Verovelkarekisteriin merkittyjen yritysten verovelkojen yhteissaldo vaihtelee vuosittain merkittävästi. Vuosittainen vaihtelu johtuu siitä, että yksittäisen yrityksen verovelkarekisterin saldo voi heilauttaa huomattavasti verovelkarekisterissä olevien yritysten vuosittaista yhteissaldoa. Pienin verovelkarekisteriin merkityn verovelan yhteissaldo on noin 340 miljoonaa euroa (vuosi 2020) ja suurin 500 miljoonaa euroa (vuosi 2019). Vuonna 2019 oli suurin yhden yksittäisen yrityksen verovelkasaldo 138 miljoonaa euroa eli yli neljännes kaikkien yritysten verovelkarekisterin yhteissaldosta.

Vuosittainen vaihtelu verovelkarekisteriin merkittyjen yritysten verovelan yhteissaldossa tapahtuu lähes kokonaan valkoisiksi luokiteltujen yritysten kohdalla. Sekä harmaiksi luokiteltujen yritysten että rajatapausten vuosittaiset verovelkarekisterin yhteissaldot ovat pysyneet vakioina.

Suhteellisesti suurempi osa verovelkarekisteriin merkitystä yhteissaldosta kuuluu harmaan talouden toimijoille kuin aikaisemmin tarkastellusta verovelan yhteismäärästä. [11]Verovelkarekisteriin merkitystä verovelasta vuosittain noin 100 miljoonaa euroa kuuluu harmaille yrityksille, mikä vastaa noin viidennestä yhteismäärästä.

[11] Selvityksen alaluvussa 4.1 on selvitetty verovelan ja verovelkarekisterisaldon selitysvoimaa harmaan talouden ennustamisessa.

Kuvio 12. Verovelkarekisterin kokonaissaldon mukainen jakautuminen alaluokkiin (harmaat yritykset, rajatapaukset ja valkoiset yritykset).

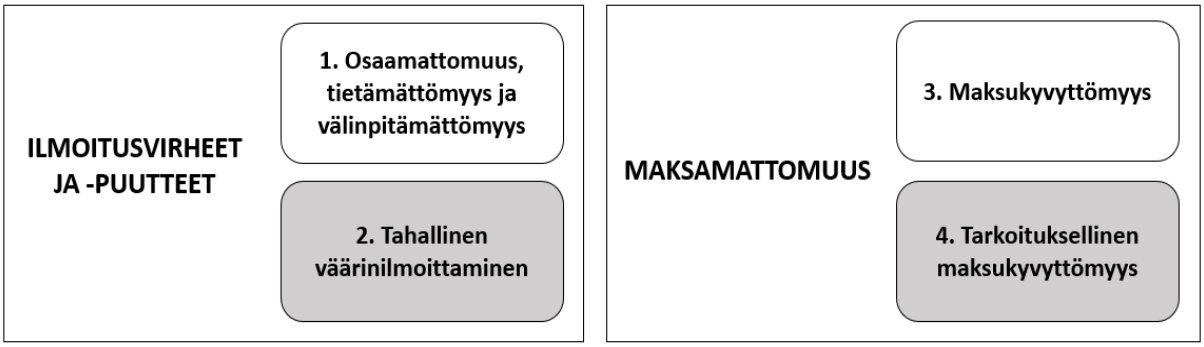


6 Yritysten verovelan syyt ja harmaa talous

Verovelkaa voidaan lähestyä tutkimalla sen teoreettista syytä. Erilaisista verovelan syistä on helppo laatia hyvinkin yksityiskohtaisia listoja asettamatta niitä keskenään tärkeysjärjestykseen tai ajalliseen järjestykseen.

On oletettavaa, että harmaan talouden toimijoiden erääntyneen verovelan syyt ovat pääsääntöisesti erilaisia kuin valkoisten yritysten. Harmaan talouden toimintaa harjoittavien yritysten verovelka johtuu oletettavasti useammin tahallisesta väärinilmoittamisesta tai maksuhaluttomuudesta (muu syy kuin heikko taloudellinen tilanne). Valkoisten yritysten verovelkojen taustalla on useimmissa tapauksissa taloudellisesta tilanteesta johtuva tilapäinen tai pysyvä maksukyvyttömyys. Valkoisten yritysten maksukyvyttömyys voi johtua normaalin liiketoiminnan kausiluonteisuudesta tai yksittäisestä suuresta kaupasta, jolloin yrityksen rahat ovat tiukalla. Toisaalta valkoisten yritysten verovelkoihin johtaneiden virheiden taustalla voi olla niiden vastuuhenkilöiden osaamattomuutta, tietämättömyyttä ja välinpitämättömyyttä, mutta ei suoranaista tahallisuutta. Heidän tarkoituksellinen tietämättömyys (olisi pitänyt ottaa selvää) tai välinpitämättömyys (olisi pitänyt tietää) on sen sijaan harmaan talouden puolella.

Kuvio 13. Verovelan syyt ja siihen liittyvä harmaa talous.



Koska aitoja verovelan syntysyitä ei pystytä havaitsemaan, täytyy teoreettinen malli operationalisoida käytettävissä olevien muuttujien avulla. Todelliset verovelan syyhyn vaikuttaneet tekijät ovat havaitsemattomia, niin kutsuttuja latentteja muuttujia. Yksittäisen yrityksen kohdalla ei voida kuin arvailla, johtuuko verovelka lopulta osaamattomuudesta, välinpitämättömyydestä, tahallisuudesta vaiko aidosta maksukyvyttömyydestä.

Lopuksi yritysten verovelan syntymissyitä on lähestytty substanssilähtöisen mallin avulla. Tällöin asiantuntija määrittelee säännöstön, joka luokittelee yritykset oletetun verovelan syntymissyyn mukaisiin luokkiin.

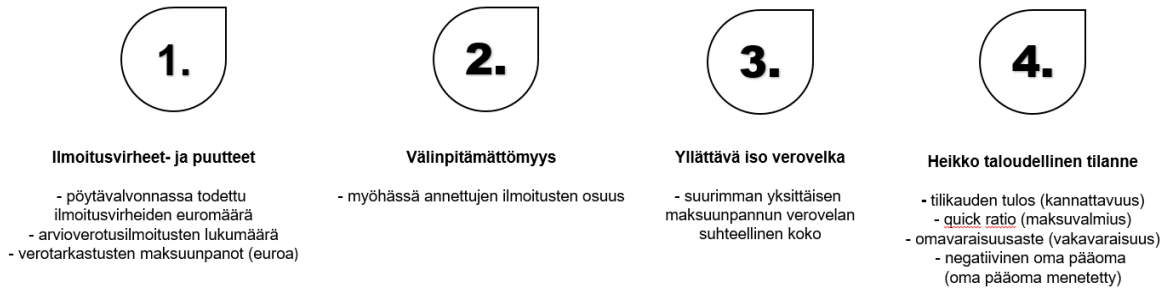
6.1 Verovelan teoreettiset syyt ja niitä indikoivat muuttujat

Yritysten verotukseen liittyvän velvoitteidenhoidon tason ja taloudellisen tilanteen selvittämiseksi on kuitenkin olemassa valmiiksi muuttujia. Osaa niistä voidaan käyttää sellaisenaan verovelkojen syiden ymmärtämiseksi ja selvittämiseksi.

Erilaisiin verovelan syihin vaikuttavat muuttujat on operationalisoitu verotus- ja taloustietoja käyttäen. Jokaisen syyn kohdalla on käytetty eri muuttujia ja raja-arvoja. Esimerkiksi jos yrityksen erääntynyt verovelka aiheutuu yksinkertaisesti siitä syystä, että maksu muistetaan maksaa vasta eräpäivän jälkeen, on asiakkaan huolimattomuutta kuvaavana muuttujana käytetty myöhässä annettujen ilmoitusten osuutta. Operationalisointi ei pidä paikkaansa, jos säännöksiä noudattamaton yritys antaa aina veroilmoituksensa tahallisesti myöhässä. Tällöin voidaan tehdä virheellinen tulkinta, että välinpitämättömyys on aiheuttanut verovelan. Syy on tällöin ennemminkin tahallinen toiminta.

Verovelan syihin vaikuttavien muuttujien on ajateltu kuvastavan seuraavia dimensioita: 1) Ilmoitusvirheet ja -puutteet 2) Välinpitämättömyys 3) Yllättävä verovelka 4) Heikko taloudellinen tilanne. Tarkastelussa on käytetty yhteensä kahdeksaa muuttujaa.

Kuvio 14. Erääntyneen verovelan syihin vaikuttavien muuttujien operationalisointi dimensioiden kautta.



Ilmoitusvirheiden ja -puutteiden dimension eri muuttujat kuvaavat osaamattomuutta, tietämättömyyttä sekä tahallista väärinilmoittamista verovelan syynä. Välinpitämättömyyttä on kuvattu yksittäisellä muuttujalla, joka on myöhässä annettujen ilmoitusten osuus. Kahden viimeisen dimension (yllättävä iso verovelka ja heikko taloudellinen tilanne) muuttujat kuvastavat ensisijaisesti yrityksen maksukyvyttömyyttä verovelan syynä. Samoja muuttujia eri raja-arvoilla voidaan käyttää käänteisesti kuvaamaan yrityksen maksuhaluttomuutta.

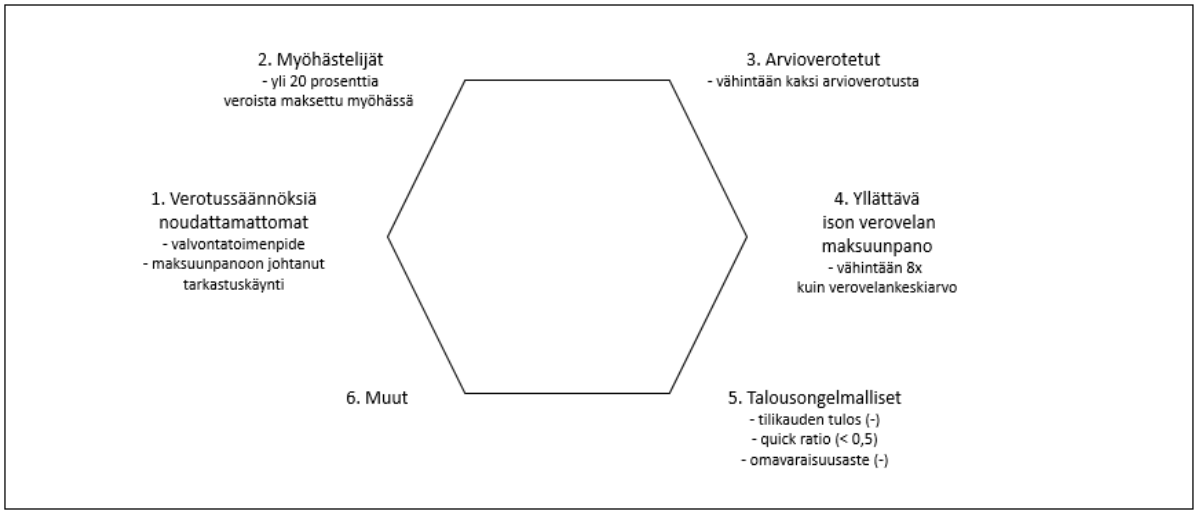
6.2 Verovelan syntymiseen vaikuttavat syyt aineistossa

Verovelkaiset yritykset on jaettu siten kuuteen eri luokkaan sääntöpohjaisesti. Kuhunkin luokkaan kuuluvien yritysten erääntyneen verovelan on tulkittu aiheutuneen luokan määrittelevän muuttujan tai useamman muuttujan mukaisella tavalla.

Asialähtöisen määrittelyn avulla on voitu tehdä juuri sellaisia luokkia kuin on haluttu. Huonona puolena on se, että sääntöjä on lähes mahdoton muodostaa siten, että ne huomioisivat useamman kuin muutaman muuttujan kerrallaan. Lisäksi yksi yritys voi täyttää yhtäaikaisesti useamman säännön kriteerit, jolloin luokkien välille on määriteltävä mielivaltainen prioriteetti. Yritys tuleekin luokitelluksi siihen luokkaan, minkä ehdot se ensiksi toteuttaa. Jokaiselle muodostetulle luokalle voi laskea erilaisia asioita, kuten siihen kuuluvien yritysten lukumäärän, harmaan talouden todennäköisyyden ja keskimääräisen verovelan määrän.

Tehdyssä luokittelussa oli mukana ainoastaan yritykset, joilla on ollut verovelkaa vuoden 2019 lopussa. Kyseisen vuoden on ajateltu edustavan tyypillistä vuotta, eikä suuria eroavaisuuksia luokkien osuuksien välillä kuvitella olevan eri vuosina. Tulosten esittäminen on selkeämpää, kun sitä ei tarvitse tehdä kullekin vuodelle erikseen.

Kuvio 15. Sääntöperusteisesti luodut verovelkaisten yritysten syyluokat prioriteettijärjestyksessä.



Selkeästi suurin määrä eli 13 500 verovelkaista yritystä (40 prosenttia) kuuluu talousongelmallisten yritysten luokkaan. Nämä yritykset ovat noudattaneet toiminnassaan verotussäännöksiä, mutta ajautuneet taloudellisiin ongelmiin. Juuri taloudellisten vaikeuksiensa vuoksi kyseiset yritykset eivät selviydy verojenmaksuvelvoitteistaan. Toiseksi suurin sääntöjä käyttämällä syntyvä luokka on noin 5 000 myöhästelijäyritystä (15 prosenttia). Arvioverotettujen osuus on lähes yhtä suuri kuin myöhästelijöiden. Verotussäännöksiä noudattamatta jääneiksi yrityksiksi on luokiteltu lähemmäs reilu 750 yritystä (hieman yli kaksi prosenttia).

Taulukko 4. Yritysten lukumäärät, harmaan talouden todennäköisyys ja keskimääräisen verovelan euromäärät sääntöperusteisesti määritellyissä luokissa vuonna 2019.

Luokka	Lukumäärä	Harmaan talouden keskimääräinen todennäköisyys	Verovelan määrä keskimäärin
1. Verotussäännöksiä noudattamattomat	762 (2,3 %)	0,5	90 550
2. Myöhästelijät	5 060 (15,1 %)	0,44	14 911
3. Arvioverotetut	4 415 (13,2 %)	0,55	39 488

4. Yllättävän ja ison verovelan maksuunpano	2 071 (6,2 %)	0,32	21 198
5. Talousongelmalliset	13 435 (40,0 %)	0,3	15 066
6. Muut	7 824 (23,3 %)	0,31	10 648

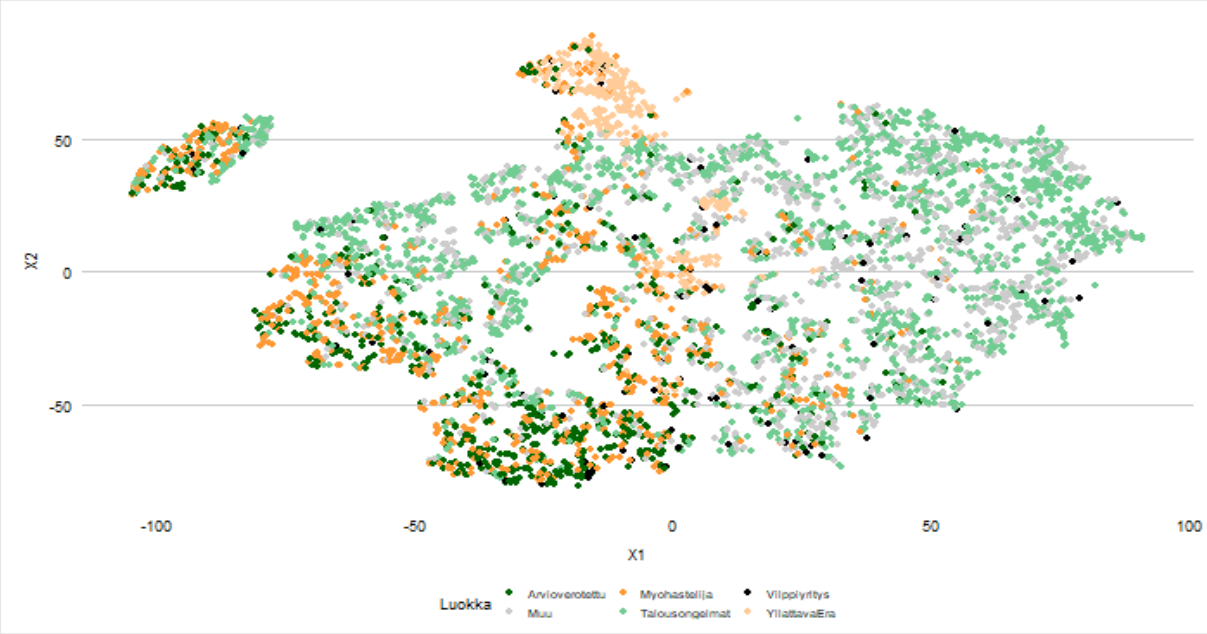
Kunkin luokan osalta voidaan päätellä sen harmaan talouden riskisyys hyödyntämällä selvityksen osiossa neljä esiteltyä harmaan talouden ennustemallia. Tällä tavalla voidaan löytää yhteyksiä verovelan syytyypin sekä harmaan talouden toimijoiksi luokiteltujen välillä.

Kaikista suurin keskimääräinen verovelka 90 000 euroa löytyy verotussäännöksiä noudattamattomien yritysten luokasta. Näillä on myös melkoisen korkeaksi päätelty harmaan talouden riski eli keskimäärin noin 50 prosenttia. Suurin riski harmaaseen talouteen luokittelumallin mukaan löytyy kuitenkin arvioverotetuilta yrityksiltä, joiden keskimääräinen todennäköisyys olla harmaan talouden toimija on 55 %. Arvioverotettujen yritysten keskimääräinen verovelka on toiseksi korkein eli liki 40 000 euroa. Kolmanneksi eniten verovelkaa on yllättävän verovelvoitteen kohdanneilla yrityksillä määrältään 20 000 euroa ja harmaan talouden toiminnan keskimääräinen todennäköisyys on noin 30 prosenttia.

Substanssilähtöisen mallin tulosten visualisointi

Visualisointia varten aineiston dimensionaalisuus on pienennetty vain kahteen hyödyntäen t-SNE-algoritmia. T-SNE-algoritmi pyrkii pienentämään aineiston dimensionaalisuutta niin, että se synnyttää mielekkäitä alaryhmiä ja samankaltaiset yritykset löytävät toisensa uudessa avaruudessa lähekkäin. T-SNE:n huono puoli on se, että sen tuottamat uudet dimensiot eivät ole enää tulkittavissa suhteessa alkuperäisten muuttujien arvoihin. T-SNE on lisäksi laskennallisesti raskas, joten visualisointi on tehty vain 5 000 yrityksen otokselle.

Kuvio 16. Verovelan syntysyyn luokat visualisoituna kahdessa dimensiossa.



Yllä olevasta sirontakuviosta on nähtävissä, että loogisilla säännöillä luodut luokat asettuvat suhteellisen säännönmukaisesti uuteen kaksiulotteiseen avaruuteen. Esimerkiksi yllättävän

erän luokka asettuu suurilta osin X1-akselin keskivaiheille ja erityisesti X2-akselin yläosaan. Myöhästelijöiden sekä arvioverotettujen valtaosa löytyy X1-akselin negatiivisten arvojen puolelta. Tämä verrattain järjestelmällinen asettuminen uuteen kaksiulotteiseen avaruuteen antaa uskoa siihen, että syntyneet luokat ovat järkeviä ja erottelevat yrityksiä mielekkäällä tavalla toisistaan verovelan luonteen osalta.

7 Yhteenveto

Selvityksen tarkoituksena on tarkastella, miten verovelka ja harmaa talous liittyvät toisiinsa. Harmaan talouden osuutta on ainakin osa verotarkastusten ja arvioverotuksen perusteella maksuunpannuista veroista. Muu harmaan talouden toimintaan kuuluva verovelka liittyy tahalliseen maksamattomuuteen.

On oletettavaa, että harmaan talouden toimijoiden verovelan syyt ovat keskimäärin erilaisia kuin verotussäännöksiä noudattaneiden yritysten. Harmaan talouden toimintaa harjoittavien yritysten verovelka johtuu oletettavasti useammin tahallisesta väärinilmoittamisesta tai tarkoituksellisesta maksamattomuudesta. Verotussäännöksiä noudattavien yritysten verovelkojen taustalla on sen sijaan useimmiten taloudellisesta tilanteesta johtuva tilapäinen tai pysyvä maksukyvyttömyys. On myös mahdollista, että yllättäviin verovelkoihin johtaneiden virheiden taustalla on niiden vastuuhenkilöiden osaamattomuutta, tietämättömyyttä ja välinpitämättömyyttä.

1. Miten verovelka vaikuttaa harmaan talouden toiminnan todennäköisyyteen?

Selvityksen tulosten perusteella verovelkaisuus ei näytä johtavan merkittävästi kohonneeseen riskiin harjoittaa harmaata taloutta. Sama johtopäätös on tehtävissä yrityksen verovelkarekisterisaldon osalta. Arvion harhaa lisää riskiperusteisesta tarkastuskohteiden valinnasta johtuva vinoutunut aineisto.

Verojen maksamisen laiminlyönnit johtavat rekistereistä poistamiseen, mikä vaikeuttaa toiminnan jatkamista. Tämän johdosta yritykset, joilla on tarkoitus harjoittaa elinkeinotoimintaa myös tulevaisuudessa, pyrkivät maksamaan verovelkansa pois yhdessä tai useammassa erässä.

2. Missä määrin verovelat jakaantuvat harmaan talouden toimintaa harjoittaville yrityksille, valkoisille ja muille yrityksille?

Toiminnassa olevat osakeyhtiöt on jaettu mallin ennustaman harmaan talouden toiminnan todennäköisyyden perusteella kolmeen luokkaan: 1) valkoiset 2) harmaat 3) rajatapaukset. Valkoisten yritysten alajoukkoon kuuluvat kaikki yritykset, joiden mallin mukainen todennäköisyys olla harmaa on alle 50 prosenttia tai jotka on verotarkastuksessa todettu verotussäännöksiä noudattaviksi. Harmaiden yritysten joukossa ovat vastaavasti ne yritykset, joiden todennäköisyys harmaan talouden toimintaan on yli 80 % tai niiden on havaittu verotarkastuksessa harjoittaneen harmaan talouden toimintaa. Loput tapaukset menevät rajatapausten luokkaan. Kolmiosainen luokittelu antaa mahdollisuuden luokitella osan kaikista toiminnassa olevista yrityksistä melko varmasti harmaiden ja valkoisten yritysten osajoukoiksi.

Mallin mukaan suurin osa verovelasta on kuulunut odotetusti valkoisille yrityksille vuosina 2017–2020. Vuosien 2017–2019 verovelan määrän putoaminen 740 miljoonasta eurosta 300 miljoonaan euroon ja vuoden 2020 hetkellinen uudelleennousu lähemmäs miljardiin euroon tapahtuivat lähes täysin valkoisten yritysten joukossa. Sen sijaan verovelan euromäärä

on pysynyt suhteellisen samana noin 200 miljoonassa eurossa sekä harmaiden että rajatapauksiksi tulkittujen yritysten joukoissa kaikkina tarkasteluvuosina.

3. Voidaanko verovelkaisia yrityksiä ryhmitellä erilaisten syntymissyiden perusteella?

Selvityksessä on lähestytty verovelan syitä substanssilähtöisen mallin avulla. Tällöin asiantuntija määrittelee säännöstön, joka luokittelee yrityksen oletetun verovelan syntymiseen vaikuttavan syyn mukaiseen luokkaan. Tehty luokittelu on tehty ainoastaan vuoden 2019 tietoja käyttäen.

Selkeästi suurin määrä eli 13 500 verovelkaista yritystä (40 prosenttia) kuuluu talousongelmallisten yritysten luokkaan. Nämä yritykset ovat noudattaneet toiminnassaan verotussäännöksiä, mutta ajautuneet taloudellisiin ongelmiin. Juuri taloudellisten vaikeuksiensa vuoksi kyseiset yritykset eivät selviydy veroveloistaan. Toiseksi suurin sääntöjä käyttämällä syntyvä luokka on noin 5 000 myöhästelijäyritystä (15 prosenttia). Arvioverotettujen osuus on lähes yhtä suuri kuin myöhästelijöiden. Verotussäännöksiä noudattamatta jääneiksi yrityksiksi on luokiteltu lähemmäs reilu 750 yritystä (hieman yli kaksi prosenttia).

Kaikista suurin keskimääräinen verovelka 90 000 euroa löytyy verotussäännöksiä noudattamattomilta yrityksiltä. Näillä on myös melkoisen korkeaksi päätelty harmaan talouden riski. Suurin riski harmaaseen talouteen luokittelumallin mukaan löytyy kuitenkin arvioverotetuilta yrityksiltä, joiden keskimääräinen todennäköisyys harmaan talouden toimintaan on 55 %. Arvioverotettujen yritysten keskimääräinen verovelka on toiseksi korkein eli liki 40 000 euroa.

8 Luettelo kuvioista ja taulukoista

9 Lähteet

Eduskunnan tarkastusvaliokunnan julkaisu. (2010). *Suomen kansainvälistyvä harmaa talous.* https://www.eduskunta.fi/FI/naineduskuntatoimii/julkaisut/Documents/trvj_1+2010.pdf.

Harmaa talouden selvitysyksikkö. (2018). *Verovelkarekisterin vaikuttavuus.* https://www.vero.fi/globalassets/harmaatalous/selvitykset/valmistuneet-selvitykset/selvitykset-2018/2018_029-verovelkarekisterin-vaikuttavuus.pdf.

Harmaa talous & talousrikollisuus. (2021). *Veropetosten torjunnassa hyviä tuloksia vuonna 2020.* https://www.vero.fi/harmaa-talous-rikollisuus/torjunta/torjuntatilastot/verotus/.

Harmaan talouden selvitysyksikkö. (2012). *Verovelat.*

Laki harmaan talouden selvitysyksiköstä. (2010). https://www.finlex.fi/fi/laki/ajantasa/2010/20101207.

Laki Harmaan talouden selvitysyksiköstä. (2010).

OECD. (2002). *Measuring the Non-Observed Economy - A Handbook.*

Vero. (2021). *Verovelkarekisteri.* https://www.vero.fi/henkiloasiakkaat/maksaminen/maksuvaikeudet/verovelkarekisteri/.

Verovelkaselvitys. (2012). Harmaan talouden selvitysyksikkö.

Viranomaistyöryhmän loppuraportti. (2014). *Verovajeen arviointimenetelmien kehittäminen.* https://www.vero.fi/contentassets/7c4eea34ab7f4b2398d7c583f37bb8e6/loppuraportti_verovajeen_arviointimenetelmien_kehittaminen.pdf.

10 Liitteet

Liite 1 Verovelan selvitysvoiman menetelmät

Liite 2 Yrityksen harmaan ennustava malli

Liite 3 Yrityksen verovelan syntysyyn menetelmät

Liite 1. Verovelan selitysvoiman menetelmät

Selvityksen luvussa kolme tutkitaan, kuinka paljon yrityksen verovelkaisuus kertoo yrityksen harmaan talouden riskistä. Selvityksessä ei lähdetty tekemään kausaalipäättelyä muuttujien välisestä suhteesta, vaan tyydyttiin tilastollisiin assosiaatioihin. Jos muuttujien välillä ei ole tilastollista assosiaatiota, silloin ei kausaliteettisuhdettakaan voi esiintyä.

Aineistona käytetään Verohallinnon vuosina 2017–2020 tekemiä verotarkastuksia. Verotarkastuksista suurin osa on tehty riskiperusteisesti, joten verotarkastukseen joutuneiden yritysten joukko edustaa lähinnä riskiyritysten alijoukkoa, ei koko yrityspopulaatiota. Näin aineistosta tehtävät päätelmät eivät automaattisesti yleisty koko yritysjoukkoon. Koska tarkastetut yritykset ovat ylipäätään keskimääräistä useammin verovelkaisia, on todennäköistä, että verovelan ja harmaan talouden toiminnan välinen aito assosiaatio on nähtyä voimakkaampi. Selvityksessä ei kuitenkaan koettu tarpeelliseksi erikseen huomioida riskiperusteisen valikoinnin aiheuttamaa harhaa erityisin tilastollisin menetelmin.

Luvussa käytetään ensiksi logistista regressiota bayesilaisittain. Logistinen regressio on regressioanalyysin erikoistyyppi, jossa selitettävä muuttuja on binäärinen, ts. se voi saada vain kaksi eri arvoa. Logistinen regressio ei siten pyri ennustamaan *määriä* vaan *todennäköisyyksiä*; kuinka todennäköisesti verotarkastuksessa löytyy yrityksen X kohdalla harmaan talouden toimintaa, ei kuinka suuri euromäärä jää valtiolta tämän seurauksena saamatta. Vastemuuttuja tulkitaan erityisesti vedonlyönnistä tuttuna ns. *riskivetosuhteena* (eng. **odds**) eli todennäköisyyksien suhteena. Jos esimerkiksi nopanheitossa on vaihtoehdot "kuusi" ja "ei-kuusi" niin kuutosen saamisen todennäköisyys on 1/6 ja riskivetosuhde sen puolesta on (1/6)/(5/6) = 0.2. Riskivetosuhde tulkitaan niin, että kuinka paljon todennäköisemmin tapahtuma A tapahtuu kuin ei tapahdu.

Logistisen regression kaava a + bx on täysin vastaava kuin lineaarisessa regressiossa. Regressiokertoimen b tulkinta ei kuitenkaan ole yhtä suoraviivaista kuin lineaarisessa versiossa. Selitettävän ja selittävän muuttujan suhde oletetaan seuraavan logistista käyrää ja kertoimen arvo 0 tarkoittaa olematonta suhdetta, positiiviset arvot taas kasvavasti voimakkaampaa assosiaatiota tapahtuman A tapahtumisen vs ei-tapahtumisen puolesta ja negatiiviset päinvastoin.

Bayesilainen tilastotiede on toinen koulukunta perinteisen ns. frekventistisen tilastotieteen rinnalla. Käytännössä kahden koulukunnan välillä suurimmat erot ovat todennäköisyyden käsitteessä sekä siinä, että bayesilaisessa tilastotieteessä voidaan hyödyntää mallin ulkopuolista informaatiota priorijakaumien avulla. Käytännön vaikutusta lopputuloksiin ei valinnalla tässä käyttötapauksessa ollut. Bayesilainen logistinen regressio valittiin, koska bayesilainen todennäköisyyden käsite on arkijärkisempi sekä tulosten tulkitseminen sekä visualisointi posteriorijakauman sekä uskottavuusvälien avulla on helpompaa.

Logistisen regression etuna on sen täydellinen avoimuus ja regressiokertoimen tulkittavuus sekä vakiointi muiden muuttujien suhteen. Regressiokertoimien avulla on yksinkertaista todeta yksittäisen muuttujan assosiaatio vastemuuttujaan siten, että muiden muuttujien

mahdollinen vaikutussuhde on jo huomioitu. Sen haittapuolena on, että logistinen regressio olettaa lineaarisen assosiaation muuttujien välille. Tätä voidaan kiertää lisäämällä polynomisia termejä regressioon. Tätä ei kuitenkaan tässä selvityksessä tehdä vaan hyödynnetään epäparametrista ja epälineaariset vaikutussuhteet sallivaa XGBoost-algoritmia, jonka tulokset avataan Shapley-arvojen avulla.

Shapley-arvot ovat nk. selkoälyn (Explainable AI eli XAI) menetelmä, jolla voidaan malliagnostisesti avata algoritmin tekemät ennustukset. Shapley-arvot ovat alun perin kehitetty peliteoriassa jakamaan yhteistyöpelin pelaajille palkinto heidän panoksensa mukaisesti. Koneoppimisen kentällä tätä sovelletaan niin, että kullekin muuttujalle annetaan palkinto muuttujan lopulliselle ennusteelle antaman panoksen mukaan. Käytännössä tämä tapahtuu pitämällä yhden muuttujan arvoa vakiona ja kokeilemalla muiden muuttujien erilaisia arvokombinaatioita, jolloin saadaan selville halutun muuttujan vaikutus ennustuksen lopputulokseen.

XGBoost taas on päätöspuihin perustuva koneoppimisen menetelmä, joka sallii epälineaariset assosiaatiot muuttujien välille. Se on hyvin tehokas malli, joka useissa ennustustehtävissä tuottaa huipputason ennustuksia ilman sen kummempia säätöjä. Sen ideana on rakentaa sekvenssi pieniä päätöspuu-malleja, joista kukin malli korjaa edellisen virheitä. Päätöspuu-pohjaisena mallina se mahdollistaa epälineaariset vaikutussuhteet muuttujien välille, minkä vuoksi se toimii erinomaisena pohjamallina, jonka perusteella muuttujien vaikutussuhteita voidaan arvioida hyödyntäen Shapley-arvoja.

Liite 2. Yrityksen harmauden ennustava malli

Selvityksen luvussa neljä jaetaan verovelan kokonaissaldo joko harmaiksi, valkoisiksi tai rajatapauksiksi luokitelluille yrityksille. Seuraavassa kerrotaan tarkemmin, miten yritysten luokittelu näihin luokkiin suoritetaan.

Verotarkastusaineisto sisältää tiedon tarkastuksessa tehdyistä havainnoista sekä yritykselle määrätyistä sanktioista. Harmaan talouden selvitysyksikössä on tehty mahdollisten eri tarkastushavaintojen pohjalta luokittelu, joka määrittää verotarkastuksen lopputuloksen joko "harmaaksi" tai "valkoiseksi". Harmaan talouden selvitysyksikön luokittelu on laajempi kuin Verohallinnon määritelmä harmaalle tarkastukselle. Näin verotarkastukset luokittelemalla saadaan binäärinen 0/1-vastemuuttuja, jota vasten voidaan rakentaa luokitteleva malli. Tehty verotarkastusten luokittelu on karkea yksinkertaistus monimutkaisesta reaalimaailman ilmiöstä. Näin myös koneoppimismalli saa vain yksinkertaistetun kuvan todellisuudesta nähdäkseen.

Luokiteltu verotarkastusaineisto muodostaa harjoitusaineiston oppivalle koneoppimismenetelmälle, joka on edellisestäkin luvusta tuttu XGBoost. Verotarkastetuille yrityksille liitetään joukko verotarkastukseen liittymättömiä yrityksen toimintaa ja riskisyyttä kuvaavia muuttujia, kuten myöhässä annettujen veroilmoitusten osuus kaikista yrityksen vuoden aikana antamista veroilmoituksista sekä liikevaihto. Malli oppii tilastollisen suhteen verotarkastuksen lopputuloksen sekä näiden muuttujien välille ja tämän opitun mallin avulla voidaan tehdä

päättelyä yrityksen harmaan talouden toiminnan riskisyydestä tarkastamattomien yritysten joukossa, joilla emme tätä tietoa ole muuten havainneet. Esimerkiksi jos malli havaitsee, että tarkastettujen yrityksen joukossa "harmailla" yrityksillä on keskimäärin alhaisempi liikevaihto kuin "valkoisilla", tarkastamattomat alhaisen liikevaihdon yritykset saavat korkeamman riskiarvion kuin muuten vastaavat korkeamman liikevaihdon yritykset.

Tarkastetut yritykset eivät ole koko yrityspopulaatiota hyvin edustava satunnaisotos vaan pääsääntöisesti riskiperusteisesti valikoitunut joukko. Tarkastettujen yritysten aineistoon rakennettu malli näkee siis hyvin vinoutuneen näkymän koko harmaan talouden kuvasta, jolloin sen tekemät ennusteet ovat myös harhaisia. Käytännössä tämä johtaa siihen, että saadut todennäköisyysarviot ovat korkeampia kuin oikeasti olisi odotettavissa. Valikoitumisharhaa voitaisiin vähentää niin kutsutuilla propensiteettipisteillä, joilla mallinnetaan kunkin yrityksen todennäköisyyttä tulla verotarkastetuksi tai hyödyntäen vanhoja satunnaistarkastuksia joko skaalauskertoimen tai bayesilaisen lähestymistavan avulla. Tässä selvityksessä emme koe näitä toimenpiteitä kuitenkaan tarpeelliseksi, sillä tarkoituksena ei ole selvittää harmaan talouden kokoluokkaa Suomessa, vaan hyödyntää mallia ainoastaan yritysten luokitteluun. Tietäen mallin harhaisuuden, otamme sen huomioon asettamalla korkean raja-arvon harmaaksi päätymiselle ja lisäämällä "epävarmojen" luokan harmaiden ja valkoisten yritysten väliin.

Liite 3. Yrityksen verovelan syntysyyn menetelmät

Verovelan syntytapaa tutkittiin substanssilähtöisesti hyödyntäen asiantuntijasääntöjä. Näiden lisäksi hyödynnettiin visualisoinnin apuna t-SNE-algoritmia.

T-SNE-algoritmi on menetelmä, jota käytetään aineiston dimensionalisuuden pienentämiseen erityisesti monimuuttuja-aineiston visualisointitarkoituksiin. Tämä tarkoittaa sitä, että aiemmin esimerkiksi 20 eri muuttujaan sisältyvä informaatio puristetaan kahteen tai kolmeen uuteen muuttujaan. Näin t-SNE on sukulainen esimerkiksi pääkomponenttianalyysille, mutta siitä poiketen uusilla muuttujilla ei ole mitään tulkinnallista arvoa.

Pääkomponenttianalyysin sijaan t-SNE ei pyri optimoimaan säilytettävän informaation määrää vaan se pyrkii projisoimaan samankaltaiset havainnot lähelle toisiaan uudessa muuttuja-avaruudessa. Tämä johtaa siihen, että havainnot tyypillisesti ryhmittyvät uudessa avaruudessa toisistaan eroaviin ryppäisiin, joten t-SNE:tä voidaan pitää tietyllä tavalla myös visuaalisena klusterointialgoritmina. Toisaalta tämän suhteen tulee olla varovainen, sillä t-SNE voi asettaa havainnot klustereita muistuttaviin ryppäisiin silloinkin, kun varsinaisesta datasta ei merkityksellisiä klustereita löydy. T-SNE on laskennallisesti raskas menetelmä ja tässä selvityksessä jouduttiin ottamaan vuoden 2019 lopussa verovelkaisten yritysten joukosta satunnaisotos, johon algoritmi ajettiin.