Kausaalimalli Osuuskauppa Hämeenmaan Facebook-julkaisujen sitoutuneisuusasteelle ja tavoittavuudelle

Tilastotieteen pro gradu -tutkielma

Lauri Valkonen

Matematiikan ja tilastotieteen laitos
Jyväskylän yliopisto

toukokuu 2020

JYVÄSKYLÄN YLIOPISTO, Matematiikan ja tilastotieteen laitos
Valkonen Lauri
Kausaalimalli Osuuskauppa Hämeenmaan Facebook-julkaisujen sitoutuneisuusasteelle ja tavoittavuudelle
pro gradu -tutkielma, tilastotiede
38 sivua + liitteet (5 sivua)
toukokuu 2020

Tiivistelmä

Sosiaalisesta mediasta on tullut sen kasvun myötä merkittävä osa yritysten digitaalista markkinointia. Hoitaakseen asiakassuhteitaan kannattavasti, on yritysten mietittävä tavoitteitaan ja markkinointistrategiaansa myös sosiaalisen median osalta. Digitaalisina palveluina sosiaalisen median kanavat mahdollistavat monipuolisen tiedon keruun käyttöön liittyen, ja siten datapohjaisen päätöksenteon hyödyntämisen asiantuntemuksen tueksi.
Tässä pro gradu -tutkielmassa muodostetaan toimintasuosituksia Osuuskauppa Hämeenmaan Facebook-julkaisuille kausaalimallin avulla, kun tarkasteltavina suorituskykymittareina ovat julkaisun sitoutuneisuusaste, sekä julkaisun tavoittavuus. Toimintasuositukset sitoutuneisuusasteen osalta käsittävät julkaisun sisältöön liittyviä valintoja. Tavoittavuuden osalta sisällöllisten tekijöiden lisäksi annetaan suosituksia julkaisuajankohtaan liittyvään toteutukseen vuorokaudenajan tasolla. Aineistona käytetään Hämeenmaan Facebook-julkaisuja noin kahden vuoden ajalta, jonka pohjalta muodostetaan edellä mainitut suorituskykymittarit, sekä muut mallinnuksessa käytettävät Facebook-julkaisuihin liittyvät muuttujat. Mallinnuksessa käytettävien muuttujien välisten syy-seuraussuhteiden kuvaamiseen esitetään kausaaligraafit molempien suorituskykymittareiden osalta. Kausaaligraafien pohjalta identifioituvat kausaalivaikutukset estimoidaan yleistettyjä additiivisia malleja käyttäen. Tehtyjen sisällöllisten ja ajallisten valintojen vaikutusta julkaisun suorituskykymittareihin tarkastellaan keskimääräisten kausaalivaikutusten avulla.
Tutkielmasta saatujen tulosten perusteella julkaisujen sitoutumisasteen osalta parhaiten toimiva julkaisu on sisällöltään viihdyttävä. Myös vakuuttavan sisällön julkaisu sitouttaa opastavaa ja inspiroivaa sisältöä paremmin. Julkaisujen tavoittavuuden kohdalla sisällöltään toimivin julkaisu on viihdyttävä, jonka jälkeen tulevat vakuuttavan, opastavan ja inspiroivan sisällön julkaisut. Vuorokaudenajan osalta havaitaan epälineaarinen suhde julkaisun tavoittavuudessa siten, että ennen kello yhdeksää aamulla toteutetut julkaisut tavoittavat kokonaisuudessaan parhaiten ja kello 12-15 julkaistut vähiten.

Avainsanoja: kausaalimalli, kausaalivaikutusten estimointi, keskimääräinen kausaalivaikutus, yleistetty additiivinen malli, sosiaalinen media, Facebook, julkaisun sitoutuneisuusaste, julkaisun tavoittavuus, digitaalinen markkinointi, ohjaileva analytiikka

Sisältö

1. Johdanto

Yritystoiminnassa markkinointi on oleellinen osa liiketoimintaa. Markkinoinnissa voidaan hyödyntää sosiaalista mediaa, joka mahdollistaa laajan yleisön tavoittamisen ja viestinnän asiakkaiden kanssa. Sosiaalisen median käyttöön liittyy keskeisesti sisällön julkaiseminen, johon käyttäjät voivat reagoida eri tavoin. Julkaisujen ominaistietojen lisäksi sosiaalisen median palvelut keräävät tietoja esimerkiksi julkaisujen näkyvyyteen ja käyttäjien julkaisuihin kohdistamiin toimiin liittyen. Näitä tietoja analysoimalla yritykset voivat kehittää sosiaalisen median julkaisujaan, parantaen markkinointiviestintäänsä, ja saaden mahdollisesti liiketoiminnallista hyötyä. Tässä tutkielmassa käsitellään Osuuskauppa Hämeenmaan sosiaalisen median julkaisuja Facebookin osalta pyrkimyksenä muodostaa toimintasuosituksia julkaisuille kausaalimallin avulla.

S-ryhmä on yritysverkosto, joka koostuu osuuskaupoista ja Suomen Osuuskauppojen keskuskunnasta ja sen tytäryhtiöistä. Alueosuuskauppoja on yhteensä 19, joista Hämeenmaa on yksi. Yritysmuotona osuuskaupoissa on osuuskunta, jonka päämääränä on tarjota hyötyjä asiakasomistajilleen. (S-ryhmä, 2019.) Kanta- ja Päijät-Hämeen alueella toimiva Hämeenmaa on yli 165 000 asiakkaan omistama osuuskunta, joka harjoittaa liiketoimintaa useilla eri toimialoilla, ollen myynnissä mitattuna alueensa suurin yritys. Keskeisenä asiana Hämeenmaan, kuten muidenkin S-ryhmän osuuskauppojen toiminnassa on myös bonus-järjestelmä, jolla palkitaan asiakkaita ostoista. (Osuuskauppa Hämeenmaa, 2020.) Laajasta ja monipuolisesta liiketoiminnasta, sekä suuresta omistajapohjasta johtuen sosiaalisen median hyödyntäminen markkinoinnissa on erittäin tarpeellista.

Hämeenmaalla on käytössään useita sosiaalisen median kanavia, joista yhtenä tärkeimmistä on Facebook. Sosiaalisen median käytön ensisijaiset tavoitteet liittyvät muun muassa asiakkuuksien hallintaan, kuten sitouttamiseen ja tavoittamiseen. Näiden tavoitteiden saavuttamisella pyritään osaltaan edesauttamaan myynnin lisäämistä. Tavoitteet huomioiden, sosiaalisen median kanavien suorituskykyä voidaan tarkastella useilla erilaisilla KPI-mittareilla (*Key Performance Indicators*), jotka määräytyvät osittain kanavakohtaisesti. (Valtari & Inkinen, 2018.) Tässä tutkielmassa keskitytään KPI-mittareiden osalta julkaisujen sitoutuneisuusasteeseen ja tavoittavuuteen. Julkaisuihin liittyvät ominaisuudet käsittävät muun muassa sisällöllisiä ja aikasidonnaisia muuttujia, joiden vaikutusta edellä mainittuihin KPI-mittareihin on mielekästä selvittää. Esimerkiksi lähes 800 yrityksen yli 100 000 Facebook-viestiä käsittävässä tutkimuksessaan Lee, Hosanagar ja Nair (2018) havaitsivat assosiaatioita erilaisten sisällöllisten

tekijöiden ja käyttäjien viestiin sitoutumisen välillä.

Kiinnostuksena on selvittää, millainen kausaalivaikutus Hämeenmaan Facebook-julkaisujen sisällöllisiin ja ajallisiin tekijöihin kohdistetuilla valinnoilla on julkaisujen sitoutuneisuusasteeseen ja tavoittavuuteen. Tarkasteltavana on julkaisun sisältötyypin keskimääräinen kausaalivaikutus sitoutuneisuusasteeseen, sekä sisältötyypin ja vuorokaudenajan keskimääräinen kausaalivaikutus julkaisun tavoittavuuteen. Vastaavanlaista kvantitatiivista tutkimusta aiheesta ei Hämeenmalla ole aiemmin tehty.

Tutkielman aluksi esitellään analyysissä käytettyä aineistoa, joka koostuu Hämeenmaan Facebook-julkaisuista, noin kahden vuoden aikaväliltä. Aineiston avulla määritellään vastemuuttujina toimivat Facebook-julkaisujen sitoutuneisuusaste ja tavoittavuus, sekä johdetaan julkaisuihin liittyviä muita muuttujia. Menetelmät-osiossa esitellään tutkielmassa käytettävien kausaalipäättelyn ja yleistettyjen additiivisten mallien perusteita, sekä mallin arviointiin liittyviä tekniikoita. Kappaleessa 4 käydään läpi varsinaisen analyysin vaiheet muodostamalla aluksi vastemuuttujiin liittyvät kausaaligraafit ja tutkimalla kiinnostavien kausaalivaikutusten identifioituvuutta. Identifioituvat kausaalivaikutukset estimoidaan yleistettyjen additiivisten mallien avulla, jonka jälkeen lasketaan keskimääräisten kausaalivaikutusten estimaatit. Estimointien tarkkuutta arvioidaan keskimääräisten kausaalivaikutusten luottamusvälitarkasteluilla, sekä tutkitaan kausaalivaikutusten estimaattien jakautumista kvantiilivälien avulla. Lopuksi tarkastellaan saatujen tulosten ja niistä tehtyjen päätelmien hyödyntämistä Hämeenmaan Facebook-markkinoinnissa. Lisäksi pohditaan käytettyyn mallinnukseen liittyviä huomioita ja kehittämiskohteita, sekä esitellään mahdollisia jatkotutkimusten aiheita.

2. Aineisto

Tutkielman aineisto koostuu Hämeenmaan Facebook-julkaisuista aikaväliltä 21.8.2017 – 30.8.2019, jolta on kertynyt yhteensä 791 julkaisua. Julkaisut elokuusta 2017 heinäkuuhun 2019 tietoineen on kerätty aineistoon elokuussa 2019. Julkaisuista kerättävät tiedot päivittyvät sitä mukaan kun käyttäjät reagoivat julkaisuihin (esimerkiksi tykkäämällä). Tämän huomioon ottamiseksi aineiston osa elokuulta 2019 on kerätty vasta lokakuussa 2019. Seuraavissa alakappaleissa esitellään tutkielmassa käytettäviä vastemuuttujia ja kovariaatteja.

2.1. Vastemuuttujat: sitoutuneisuusaste ja tavoittavuus

Aineistossa on lukumäärätieto jokaisen julkaisun nähneistä ja niihin sitoutuneista käyttäjistä. Näiden avulla määritellään mielenkiinnon kohteena olevat vastemuuttujat tämän tutkielman osalta:

Julkaisun tavoittavuus on niiden käyttäjien lukumäärä, jotka ovat nähneet kyseiseen julkaisuun liittyvää sisältöä näytöllään. Tällä tarkoitetaan julkaisun kokonaisuudessaan tavoittamia yksilöllisiä käyttäjiä julkaisuhetkestä alkaen. Tätä muuttujaa käytetään vasteena julkaisun tavoittavuuden mallinnuksessa sellaisenaan.

Julkaisun sitoutuneisuusaste saadaan julkaisuun sitoutuneiden käyttäjien suhteena julkaisun tavoittavuuteen. Julkaisuun sitoutuneet käyttäjät kuvaa niiden käyttäjien lukumäärää, jotka ovat reagoineet julkaisuun esimerkiksi tykkäämällä, jakamalla, kommentoimalla tai klikkaamalla jotain elementtiä julkaisussa.

Taulukko 1: Vastemuuttujiin liittyviä tunnuslukuja.

Vastemuuttujien tunnuslukuja ja jakaumia					
	Keskiarvo	**Mediaani**	**Keskihajonta**	**Minimi**	**Maksimi**
Tavoittavuus (käyttäjää)	11070	4880	37507	391	748699
Sitoutuneisuusaste (%)	3.98	2.88	3.26	0.44	27.45

Taulukkoon 1 on koottu molempiin vastemuuttujiin liittyviä tunnuslukuja. Sekä julkaisun tavoittavuuden, että sitoutuneisuusasteen osalta havaitaan suurta vaihtelua niiden arvoissa keskihajontaa ja vaihteluväliä tarkasteltaessa, ja siten niitä on mielekästä lähteä mallintamaan. Molempien vastemuuttujien osalta keskiarvo on suurempi kuin mediaani ja jakaumat siten oikealle vinoja, jolloin pienempiä arvoja on havaittu enemmän. Lisäksi huomattavaa on, että molemmat vastemuuttujat ovat saaneet myös poikkeuksellisen suuria arvoja (maksimit).

2.2. Kovariaatit

Esitellään seuraavaksi julkaisun sitoutuneisuusasteen ja tavoittavuuden mallinnukseen käytettäviä kovariaatteja, jotka käsittävät sekä kategorisia, että jatkuvia muuttujia. Kovariaattien muodostamisessa on hyödynnetty aineiston julkaisuihin liittyviä tietoja sellaisenaan, sekä johtamalla niiden avulla uusia muuttujia. Taulukossa 2 on esitelty kovariaatteihin liittyviä jakaumia ja tunnuslukuja.

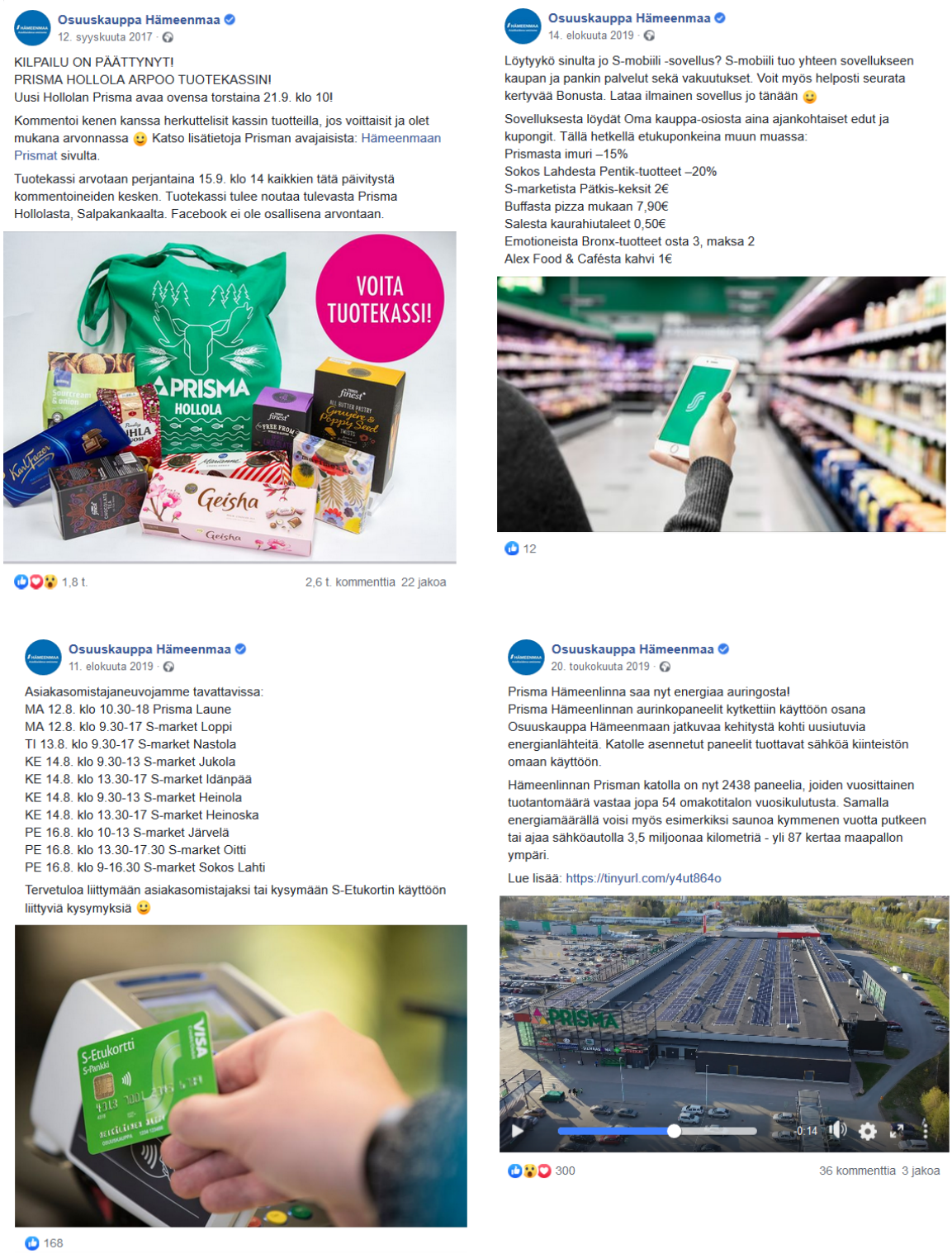
Taulukko 2: Kovariaatteihin liittyviä jakaumia ja tunnuslukuja.

	Frekvenssi	**%-osuus**		**Frekvenssi**	**%-osuus**
Julkaisun sisältötyyppi			**Julkaisun viikonpäivä**		
Inspiroiva	298	38 %	*maanantai*	137	17 %
Opastava	196	25 %	*tiistai*	160	20 %
Vakuuttava	183	23 %	*keskiviikko*	131	17 %
Viihdyttävä	114	14 %	*torstai*	117	15 %
Julkaisun vuorokaudenaika			*perjantai*	127	16 %
–8:59	136	17 %	*lauantai*	41	5 %
9:00–11:59	241	30 %	*sunnuntai*	78	10 %
12:00–14:59	242	31 %	**Julkaisun mediatyyppi**		
15:00–	172	22 %	*Kuva*	675	85 %
Julkaisun vuodenaika			*Video & jaettu video*	53	7 %
Kevät (kk = 3,4,5)	199	25 %	*Muu*	63	8 %
Kesä (kk = 6,7,8)	174	22 %	**Ostettu näkyvyyttä**		
Syksy (kk = 9,10,11)	247	31 %	*Ei*	588	74 %
Talvi (kk = 12,1,2)	171	22 %	*Kyllä*	203	26 %

	Keskiarvo	**Mediaani**	**Keskihajonta**	**Minimi**	**Maksimi**
Julkaisun pituus (merkkiä)	450	320	449	0	3230
Julkaisujen välinen aikaero (tuntia)	22	19	24	0	155
CLS-luku	0.012	0.007	0.015	0.000	0.120
Kumulatiivinen aika (tuntia)	**Vaihteluväli**: [0, 17361]				

Hämeenmaan julkaisuissa toistuu tiettyjä asiasisältöjä ja näiden sisällöllisten tekijöiden huomiointiin käytetään tässä tutkielmassa neliluokkaista muuttujaa. Määrittelyn pohjana on hyödynnetty Valtarin ja Inkisen (2018) raportin nelikenttäjakoa Hämeenmaan sosiaalisen median julkaisujen sisällöistä (Liite 1). Sen perusteella sisällöt on jaettavissa neljään eri kategoriaan: inspiroiviin, opastaviin, vakuuttaviin ja viihdyttäviin. Nelikentän pystyakseli kuvaa julkaisun sisällön luokittelua tunnepitoiseksi tai faktapohjaiseksi. Vaaka-akselilla arvioidaan julkaisun sisältöä tietoisuutta lisääväksi tai toisaalta sitotuttavaksi ja mahdollisesti ostoja lisääväksi. Julkaisuissa ei ole valmiiksi tietona mihin sisältöluokkaan se kuuluu, vaan luokittelu on tehty tutkielman tekijän toimesta lukemalla ja katsomalla julkaisut läpi.

Viihdyttävä julkaisu on tunnepitoisempi ja tietoisuuteen tähtäävä. Tällaisia ovat esimerkiksi arvonnat, henkilökunnan arkeen liittyvät julkaisut ja osaltaan myös videot. Näitä julkaisuja on aineistossa vähiten (14 %). Inspiroiva julkaisu on myös tunteisiin vetoava, mutta tavoitteeltaan sitouttavampi ja ostoihin tähtäävä. Aineistossa on eniten tämän sisältötyypin julkaisuja (38 %), joista esimerkkeinä ovat tapahtumat, kampanjat ja tarjoukset. Vakuuttava julkaisu puolestaan on enemmän faktaperusteinen, mutta myös sitouttamiseen ja mahdollisesti ostoihin pyrkivä. Erilaiset tiedotteet, vastuullisuusasiat ja osuustoiminnan esittely liittyvät tähän sisältötyyppiin. Viimeisen kategorian opastavat julkaisut ovat faktapohjaisia ja tietoisuutta lisääviä, joista esimerkkeinä asiakasomistaja-infot, aukioloajat ja toimipaikkojen esittely. Myös rekrytoinnit on laskettu tähän kategoriaan. Vakuuttavien ja opastavien julkaisujen osuudet aineistossa ovat lähes samat (23 % ja 25 %). Esimerkki kunkin sisältötyypin julkaisusta on esitelty kuvassa 1.



Kuva 1: Tyypillisiä julkaisujen sisältöjä (ylhäältä oikealle): Viihdyttävä julkaisu (Osuuskauppa Hämeenmaa, 2017), Inspiroiva julkaisu (Osuuskauppa Hämeenmaa, 2019b), Opastava julkaisu (Osuuskauppa Hämeenmaa, 2019a) ja Vakuuttava julkaisu (Osuuskauppa Hämeenmaa, 2019c).

Aineistossa on tieto täsmällisestä ajasta, milloin julkaisut on luotu. Kategoriset ajalliset muuttujat perusmuodossaan tuottavat kuitenkin paljon luokkatasoja (esim. kuukausi, kellonaika), jolloin voi olla järkevää yhdistellä luokkia. Julkaisun vuorokaudenaikamuuttuja on muodostettu luokittelemalla julkaisu kellonajan suhteen neljään luokkaan. Klo 9-12 ja 12-15 ovat eniten julkaisuja sisältäviä vuorokaudenaikoja, käsittäen yhteensä 61 % julkaisuista. 17 % julkaisuista on toteutettu ennen klo 9:ää ja loput 22 % klo 15 jälkeen. Julkaisun ajankohdalle kuukausitasolla on toteutettu jako neljään luokkaan vuodenaikojen suhteen. Vuodenaikamuuttujan luokista syksy (syys- , loka-, marraskuu) sisältää eniten julkaisuja (31 %), jonka jälkeen toiseksi eniten (25 %) on julkaistu keväällä (maalis-, huhti-, toukokuu). Kesä (kesä-, heinä-, elokuu) ja talvi (joulu-, tammi-, helmikuu) kattavat molemmat saman osuuden aineistosta (22 %). Viikonpäivistä julkaiseminen on painottunut arkipäiviin siten, että eniten on julkaistu tiistaisin (20 %) ja muiden arkipäivien osalta jakauma on tasainen. Sen sijaan viikonlopun (lauantai-sunnuntai) osuus julkaisuista on vain 15%.

Facebook luokittelee julkaisuja eri mediatyyppeihin, joita ovat kuva, video, jaettu video, status ja linkki. Aineiston julkaisuista suurin osa (85 %) on luokiteltu kuviksi. Videot ja jaetut videot on yhdistetty yhdeksi muuttujan tasoksi johtuen jaettujen videoiden vähäisestä määrästä. Loput julkaisujen mediatyypit on yhdistetty kategoriaan 'muut', joka sisältää statukset, linkit ja luokittelemattomat julkaisut, joita on seitsemän kappaletta. Julkaisuista on lisäksi mitattu niiden käyttäjien lukumäärä, jotka ovat nähneet julkaisun maksetun jakelun, esimerkiksi mainoksen kautta. Tästä on muodostettu kaksiluokkainen muuttuja kuvaamaan onko julkaisulle ostettu näkyvyyttä. Aineiston osalta 26 %:lle julkaisuista on ostettu näkyvyyttä ja loput 74 % ovat niin sanottuja orgaanisia julkaisuja.

Edellä kuvattujen kategoristen muuttujien lisäksi analyyseihin sisältyy jatkuvia kovariaatteja. Julkaisusta on laskettu siihen liittyvän viestin pituus summaamalla viestissä esiintyvien merkkien määrä. Yhtenä jatkuvana aikasidonnaisena kovariaattina on julkaisujen välinen aikaero, joka on edeltävästä julkaisusta kulunut aika tunnin tarkkuudella mitattuna. Vastemuuttujiin läheisesti liittyvä kovariaatti CLS-luku (*Comment, Like, Share*) käsittää yksilöllisten käyttäjien julkaisuun suorittamien tykkäysten, jakojen ja kommentointien summan suhteessa julkaisun nähneisiin käyttäjiin. Aineiston avulla ei voida yksilöidä käyttäjiä tehtyjen tykkäysten, jakojen ja kommentointien osalta. Huomioitavaa onkin, että yksittäinen käyttäjä on voinut suorittaa useampaa kuin yhtä edellä mainituista toiminnoista, jolloin CLS-luvun arvot eivät teoriassa rajoitu välille [0,1]. Käytännössä tykkäyksiä, kommentointeja ja jakoja suhteessa julkaisun näh-

neisiin on kuitenkin selvästi vähemmän, ja muuttujan arvot painottuvat taulukon 2 perusteella hyvin pieniin arvoihin. Kuten vastemuuttujien tapauksessa, myös julkaisun pituuden, aikaeron ja CLS-luvun osalta jakaumat ovat oikealle vinoja keskiarvon ollessa mediaania suurempi. Keskihajontaa ja vaihteluväliä tarkastelemalla havaitaan myös selvää vaihtelua näiden muuttujien arvojen osalta.

Koska tavoitteena on rakentaa kausaalimalli, voidaan olettaa havaintoaineiston kahden vuoden aikajänteen olevan oleellinen tekijä muuttujien välisiä syy-seuraussuhteita arvioitaessa. Tämän huomioimiseksi jokaiselle julkaisulle lasketaan kumulatiivisen aikamuuttujan arvo havaintoaineiston käsittämältä aikaväliltä. Muuttujan arvot julkaisuille mitataan tunnin tarkkuudella, lähtien ensimmäisen julkaisun ajanhetkestä nolla.

3. Menetelmät

Seuraavissa alakappaleissa esitellään kausaalimallien estimointiin liittyviä vaiheita ja niissä käytettävien menetelmien teoriaa. Ensin tarkastellaan kausaalipäättelyä yleisesti, sekä kausaalisuhteiden esittämistä graafeilla. Graafien avulla voidaan muodostaa käsityksiä kovariaattien välisistä keskinäisistä suhteista, sekä niiden yhteyksistä vastemuuttujiin. Kausaalivaikutusten estimointiin liittyy identifioituvuustarkastelut, joita sivutaan lyhyesti. Varsinaista estimointia varten esitellään tutkielmassa käytettävien yleistettyjen additiivisten mallien teoriaa, sekä vastemuuttujille valittavia jakaumia, joita ovat tämän tutkielman osalta binomijakauma ja negatiivinen binomijakauma. Kausaalimallien estimointimenetelmien jälkeen kuvaillaan mallin valinnassa käytettäviä kriteerejä, kuten ristiinvalidointia ja jäännöstarkastelua. Lopuksi esitellään menetelmä keskimääräisten kausaalivaikutusten luottamusvälien laskemiseen prosenttipiste bootstrap-menetelmällä. Analyysissä käytettävien menetelmien toteutukseen R-ympäristössä (R Core Team, 2020) hyödynnetään kausaaligraafien osalta Csardin ja Nepuszin (2006) igraph-pakettia. Identifioituvuuden selvittämiseen käytetään causaleffect-pakettia (Tikka & Karvanen, 2017), ja kausaalivaikutusten estimointiin yleistettyjen additiivisten mallien avulla käytetään mgcv-paketin gam-funktiota (Wood, 2017). Tulosten visuaalisessa esittämisessä hyödynnetään ggplot2-pakettia (Wickham, 2016).

3.1. Kausaalimallit

Kausaalimallinnukseen liittyy kysymys asioiden välisistä syy-seuraussuhteista ja niiden vaikutuksista. Kiinnostuksen kohteena voi olla selvittää miten tekijä X vaikuttaa tekijään Y ja tämän vaikutuksen estimointi. Näiden vaikutusten ymmärtämisellä voidaan ohjata toimintaa haluttuun suuntaan. Tämä tutkimusasetelma liittyy vahvasti kokeelliseen tutkimukseen, mutta sen toteuttamiseen voidaan käyttää joissakin tilanteissa myös havainnoivaa tutkimusta. (Pearl, Glymour & Jewell, 2016.)

Pelkän aineiston pohjalta tehdyssä päättelyssä voidaan päätyä tilanteeseen, jossa tulkinta asioiden välisistä vaikutuksista voi olla jopa epäloogisia. Esimerkiksi tutkittaessa tekijän X vaikutusta tekijään Y, voi tulkinta olla erilainen verrattuna tilanteeseen, jossa otetaan huomioon kolmas tekijä Z. (Pearl ym., 2016.) Tämänkaltainen tilanne tunnetaan myös Simpsonin paradoksina (Simpson, 1951). Oikeanlaisiin pää-

telmiin kausaalivaikutuksista tarvitaankin perinteisen tilastotieteen ja ilmiön tuoman datan lisäksi juuri ymmärrystä asioiden välisistä suhteista (Pearl ym., 2016).

Käsitys tutkittavan ilmiön kausaalimekanismista voidaan ottaa huomioon suunatulla asyklisellä graafilla (*Directed acyclic graph*, DAG), jossa graafin solmut kuvaavat muuttujia ja särmät niiden välisiä yhteyksiä. Suunnatussa graafissa kaikkien solmujen välisillä särmillä on suunta (nuoli) osoittamaan syy-seuraussuhdetta. Graafi on lisäksi asyklinen, jos siinä ei ole solmujonoa, joita pitkin särmien nuolten suuntaisesti päästäisiin lähtösolmusta takaisin lähtösolmuun. (Pearl, 2009.)

Kausaalimalliin M liittyvä graafi G sisältää muuttujia X_p $(p = 1, .., m)$, joiden välillä vallitsee erilaisia funktionaalisia suhteita. Mahdollisia muuttujaan X_p suoraan vaikuttavia muuttujia graafissa G kutsutaan muuutujan X_p vanhemmiksi, joita merkitään $\boldsymbol{Pa}(X_p)$. Kausaalimalli M voidaan siten esittää joukon

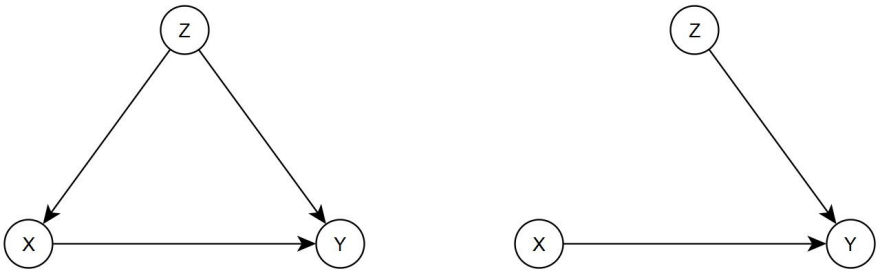
$$\{X_1, ..., X_m\} = \{f_1(\boldsymbol{Pa}(X_1), \boldsymbol{U}_1), ..., f_m(\boldsymbol{Pa}(X_m), \boldsymbol{U}_m)\} ,$$

sekä taustamuuttujien yhteisjakauman $P(\boldsymbol{U})$ yhdistelmänä, missä muuttuja X_p riippuu vanhemmistaan $\boldsymbol{Pa}(X_p)$, sekä havaitsemattomista taustamuuttujistaan $\boldsymbol{U}_p$. (Pearl, 2009.)

Kiinnostuksen kohteen ollessa muuttujan X_p kausaalivaikutus muuttujaan Y, voidaan kausaalivaikutus määritellä ehdollisen jakauman $P(Y|do(X_p))$ avulla. Interventiossa (merkitään $do(.)$) muuttujan X_p arvoa halutaan kontrolloida tutkijan toimesta asettamalla $X_p = x_p$, mikä johtaa muuttujan X_p funktionaalisen suhteen $f_p(\boldsymbol{Pa}(X_p), \boldsymbol{U}_p)$ häviämiseen kausaalimallista M. Muut mallin muuttujat eivät tällöin vaikuta muuttujaan X_p, mistä seuraa myös graafin G muuttujaan X_p tulevien nuolten poistaminen. Lisäksi interventiosta johtuen muut mallin funktionaaliset suhteet saavat muuttujan X_p kohdalla arvon x_p. (Pearl, 2009.) Esimerkki yksinkertaisesta DAG:sta ja muuttujaan kohdistetun intervention vaikutuksesta on esitetty kuvassa 2.

Oleellinen kysymys kausaalimallinnuksessa liittyy kausaalivaikutuksen $P(Y|do(X_p))$ identifioituvuuteen, jolloin kausaalijakauma voidaan esittää havaintoaineiston jakaumien avulla (Pearl, 2009). Identifioituvuuden selvittämiseen on olemassa erilaisia kausaaligraafiin liittyviä päättely- ja laskusääntöjä, jotka tunnetaan myös nimellä do-calculus (Pearl, 1995). Nämä laskusäännöt perustuvat graafin muuttujien ehdollisten riippumattomuuksien tutkimiseen redusoiduissa kausaaligraafeissa, pyrkimyksenä päästä havaintojakaumien esitysmuotoon (Pearl, 2009). Monimutkaisemmissa graafeissa hyödyllinen keino identifioituvuuden selvittämiseen on ID-algoritmi (Shpitser

& Pearl, 2006), jossa hyödynnetään identifioituvuuden tarkastelemista graafin osissa rekursiivisesti. ID-algoritmin avulla saadaan identifioituvuuden tapauksessa kausaalivaikutuksen esitys havaintojakaumien muodossa (Shpitser & Pearl, 2006).



Kuva 2: Esimerkki suunnatuista asyklisistä graafeista (DAG). Vasemmassa graafissa on kolme solmua (muuttujaa) $\{X, Y, Z\}$ ja kolme särmää osoittamaan solmujen väliset yhteydet. Oikeassa graafissa on tilanne, jossa muuttujaan X on kohdistettu interventio $do(X = x)$.

Usein kausaalivaikutusten identifioituessa päädytään kausaalijakauman lausekkeessa takaovikorjauksena tunnettuun havaintojakaumien esitysmuotoon

$$P(Y|do(X)) = \sum_{\boldsymbol{Z}} P(Y|X, \boldsymbol{Z})P(\boldsymbol{Z}) \, ,$$

missä muuttujajoukko $\boldsymbol{Z}$ voi edustaa yhtä tai useampaa muuttujaa graafissa G (Pearl, 2009). Kun halutaan selvittää muuttujan X keskimääräistä kausaalivaikutusta muuttujaan Y, tarkastellaan jakauman $P(Y|do(X))$ odotusarvoa $E(Y|do(X))$. Odotusarvon määritelmän mukaan

$$E(Y|do(X = x)) = \sum_{y} y \, P(Y = y|do(X = x)) \, .$$

Nyt takaovikorjauksen tilanteessa päästään odostusarvon kohdalla esitysmuotoon

$$
\begin{aligned}
E(Y|do(X=x)) &= \sum_y y \sum_{\boldsymbol{z}} P(Y=y|X=x,\boldsymbol{Z}=\boldsymbol{z})P(\boldsymbol{Z}=\boldsymbol{z}) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y|X=x,\boldsymbol{Z}=\boldsymbol{z}_i)\,,
\end{aligned}
$$

missä viimeinen yhtäsuuruus seuraa suurten lukujen lakiin perustuvasta approksimaatiosta. (Shalizi, 2019.) Merkitään edellä määritettyä keskimääräistä kausaalivaikutusta

$$
\theta_{Y;X=x} := E(Y|do(X=x)) = \frac{1}{n}\sum_{i=1}^{n} E(Y|X=x,\boldsymbol{Z}=\boldsymbol{z}_i)\,.
$$

Keskimääräisen kausaalivaikutuksen estimaattori on siten

$$
\widehat{\theta}_{Y;X=x} := \widehat{E}(Y|do(X=x)) = \frac{1}{n}\sum_{i=1}^{n} \widehat{E}(Y|X=x,\boldsymbol{Z}=\boldsymbol{z}_i)\,,
$$

jonka sisältämän odotusarvon $E(Y|X,\boldsymbol{Z})$ estimaattorin $\widehat{E}(Y|X,\boldsymbol{Z})$ muodostamiseen voidaan käyttää seuraavassa kappaleessa esiteltävää yleistettyä additiivista mallia.

3.2. Yleistetyt additiiviset mallit

Tarkastellaan seuraavaksi keskimääräisen kausaalivaikutuksen sisältämän ehdollisen odotusarvon $E(Y|X, \boldsymbol{Z})$ mallintamista. Tavallinen tapa odotusarvon mallintamiseen on käyttää yleistettyjä lineaarisia malleja. On kuitenkin tilanteita, jolloin jatkuvien kovariaattien ja vasteen välinen yhteys ei ole lineaarinen. Odotusarvon mallintamiseen voidaan tällöin käyttää yleistettyjä additiivisia malleja, jotka saadaan yleistetyn lineaarisen mallin laajennuksena ottamalla nämä epälineaarisuudet huomioon.

Satunnaismuuttujan saadessa arvoja {0,1} todennäköisyyksillä $\{(1-\pi), \pi\}$, sen sanotaan noudattavan Bernoullin jakaumaa parametrilla π. Merkitään arvojoukkoa siten, että 0="epäonnistuminen" ja 1="onnistuminen". m riippumattoman Bernoulli-jakautuneen satunnaismuuttujan summa noudattaa binomijakaumaa, jonka pistetodennäköisyysfunktio on muotoa

$$p(y_i; m_i, \pi_i) = \binom{m_i}{y_i} {\pi_i}^{y_i} (1-\pi_i)^{m_i - y_i} ,$$

missä m_i on havainnon i Bernoulli-jakautuneiden satunnaismuuttujien lukumäärä, y_i onnistumisten lukumäärä ja π_i onnistumisen todennäköisyys. (McCullagh & Nelder, 1989.)

Positiivista lukumäärämuuttujaa mallinnettaessa tavallinen valinta on käyttää Poisson-jakaumaa, joka olettaa odotusarvon ja varianssin yhtäsuuruuden. Usein kuitenkaan tämä oletus ei toteudu ja varianssi voi olla odotusarvoa selvästi suurempi. (McCullagh & Nelder, 1989.) Edellä kuvatussa ylihajontatilanteessa satunnaismuuttujan mallintamiseen on mahdollista käyttää negatiivista binomijakaumaa, joka voidaan johtaa ja parametrisoida usealla eri tavalla. Negatiivisen binomijakautuneen satunnaismuuttujan voidaan oletta olevan niin sanottu sekoitus Poisson- ja gamma-jakaumasta, jolloin jakauman pistetodennäköisyysfunktio on

$$p(y_i; \mu_i, \alpha) = \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} ,$$

missä μ_i on odotusarvoparametri, ja α ylihajonnan huomioiva parametri. (Hilbe, 2011.)

Yleistetyssä lineaarisessa mallissa kiinnostuksena on mallintaa vastemuuttujan Y odotusarvoa $\boldsymbol{E(Y)} = (E(Y_1), ..., E(Y_n))$, kun vastemuuttujasta on havaittu otos riip-

pumattomia havaintoja $\mathbf{y} = (y_1, ..., y_n)$. Lisäksi on havaittu p kpl muita satunnaismuuttujia $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$, joita voidaan käyttää mallinnuksessa kovariaatteina. Yleistetyssä lineaarisessa mallissa käytetään linkkifunktiota $g(.)$ liittämään vasteen odotusarvo $E(Y_i)$ lineaariseen ennustimeen η_i, joka sisältää kovariaatit $\mathbf{X}$ sekä estimoitavat malliparametrit $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$. Binomijakauman kohdalla mallin satunnaisosan ja systemaattisen osan liittäväksi linkkifunktioksi tavallinen valinta on logit-linkki, jolloin $g(E(Y_i)) = log(\pi_i/(1 - \pi_i))$. (McCullagh & Nelder, 1989.) Negatiivisen binomijakauman tapauksessa linkkifunktioksi käy logaritminen linkki $g(E(Y_i)) = log(\mu_i)$ (Hilbe, 2011).

Seuraavaksi esitettävä yleistettyjen additiivisten mallien teoria pohjautuu teokseen *Generalized Additive Models: An Introduction with R* (Wood, 2017). Yleistetty additiivinen malli voidaan muodostaa yleistetyn lineaarisen mallin laajennuksena. Merkitään mallin parametriseen komponenttiin liityviä kovariaatteja indekseillä $j \in J$ ja kovariaatteja joihin on sovitettu splinifunktio indekseillä $k \in K$, missä $J \cap K = \varnothing$. Nyt malliyhtälö voidaan kirjoittaa muodossa

$$g(E(Y_i)) = \beta_0 + \sum_j \beta_j x_{ji} + \sum_k s_k(x_{ki}) \, .$$

Malliyhtälössä $s_k(.)$ on kovariaattiin x_k sovitettu silotusfunktio, jonka estimaatiksi eräs valinta on regressiosplini (*thin plate regression spline*). Mallin systemaattinen ennustinosa koostuu siten vakiotermin β_0, parametristen komponenttien $\beta_j x_j$ ja splinifunktioiden $s_k(x_k)$ summasta. Uskottavuuspäättelyyn perustuvassa estimoinnissa käytetään sakotettua logaritmista uskottavuutta, joka huomioi käytettävän silotuksen. Maksimoitava sakotettu logaritminen uskottavuusfunktio on muotoa

$$l_{sakotettu}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi} \sum_k \lambda_k \boldsymbol{\beta}^T \boldsymbol{S}_k \boldsymbol{\beta} \, ,$$

missä $l(\boldsymbol{\beta})$ on mallin logaritminen uskottavuus ja ϕ malliin liittyvä skaalaparametri. Silotusparametri λ_k ja sakkomatriisi $\boldsymbol{S}_k$ määrittävät yhdessä sakon muuttujaan x_k käytettävästä silotuksesta suhteessa mallin sopivuuteen. Tämän avulla pyritään ottamaan huomioon kovariaattien ja vasteen epälineaarisuudet, välttämällä liian monimutkaisen silotusfunktion käyttöä. $\boldsymbol{\beta}$-parametrien estimointi toteutetaan sakotetulla iteratiivisella uudelleenpainotetulla pienimmän neliösumman (PIRLS) menetelmällä, annetulla parametrilla λ. Lisäksi estimoidaan vastemuuttujan jakaumasta riippuen skaalaparametri

ϕ, sekä siloitusparametrit λ_k. Lähdekirjallisuudessa (Wood, 2017) on esitetty yksityiskohtaisemmin mallin parametrien estimointitapoja ja toteutuksia.

3.3. Mallin valinta ja diagnostiikka

Odotusarvon mallin $\widehat{E}(Y|X, \boldsymbol{Z})$ valinta ja diagnostiikka käsittää useita tarkasteluvaiheita. Ehdokkaiden joukosta sopivimman mallin valitsemisen lisäksi tarkastellaan mallin sopivuutta tehtyjen valintojen ja oletusten osalta. Lisäksi voidaan selvittää mallin avulla toteutetun keskimääräisen kausaalivaikutuksen estimoinnin tarkkuutta tulosten luottettavuuden arviointiin.

Perinteisten mallinvalintakriteerien (mm. AIC) lisäksi mallien vertailua on mahdollista toteuttaa aineiston uusiokäyttöön pohjautuvilla menetelmillä. Aineiston koon ollessa suhteellisen pieni, voidaan käyttää K-ristiinvalidointia mallin sopivuuden arviointiin. (Hastie, Tibshirani & Friedman, 2009.) Olkoon aineisto $\mathcal{D}$ jaettu opetusaineistoon $\mathcal{O}_k$ ja testiaineistoon $\mathcal{T}_k$ (k=1,...,K). Opetusaineistoon sovitetulla mallilla lasketaan ennusteet testiaineistolle ja verrataan niitä testiaineiston todellisiin havaintoihin. Vaihtoehtoisia malleja vertailtaessa, pienemmän keskimääräisen virheen

$$CV(h) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i:y_i \in \mathcal{T}_k} \Big(h_k(\boldsymbol{x}_i) - y_i \Big)^2$$

tuottavaa mallia pidetään sopivampana. Kaavassa K on ositusten lukumäärä ja n_k on k:nnen osituksen havaintojen lukumäärä. $\mathcal{T}_k$ on k:nnen osituksen testiaineisto ja $h_k(.)$ on k:nnen osituksen opetusaineistoon sovitetun mallin antamat ennusteet testiaineistolle. (Bengio & Grandvalet, 2004.) Valittaessa K=10, aineisto jaetaan satunnaisesti kymmeneen noin yhtä suureen osaan, missä jokainen k:nnes osa on kerran testiaineistona, ja loppuosa aineistosta opetusaineistona (Hastie ym., 2009). Ristiinvalidointeja on mahdollista toteuttaa myös useaan kertaan, jolloin voidaan tarkastella keskiarvoa ristiinvalidointitoistojen tuottamista testivirheistä $CV(h)$ (Kim, 2009).

Mallin estimoinnin jälkeen on hyvä tutkia sen sopivuutta tehtyjen mallioletusten ja valintojen osalta. Tätä voidaan toteuttaa erilaisilla jäännöstarkasteluilla, kuten Pearsonin jäännösten avulla. Pearsonin jäännökset määritellään

$$\epsilon_i = \frac{y_i - \widehat{E}(Y_i)}{\sqrt{V(\widehat{E}(Y_i))}},$$

missä mallin jäännöksiä $y_i - \widehat{E}(Y_i)$ skaalataan mallin varianssifunktion neliöjuurella. Mallin sopiessa Pearsonin jäännösten ja mallin sovitteiden välisen hajontakuvion tulisi näyttää satunnaisesti jakautuneelta nollan ympäristössä. (Wood, 2017.)

Keskimääräiselle kausaalivaikutukselle θ voidaan laskea luottamusväli kuvaamaan havaintoaineistosta johtuvaa epävarmuutta estimoinnissa. Seuraavaksi esiteltävä prosenttipiste bootstrap-menetelmä on eräs tapa luottamusvälin laskemiseen. Esitettävä teoria pohjautuu teokseen *An Introduction to the Bootstrap* (Efron & Tibshirani, 1993). Olkoon aineistossa $\mathcal{D}$ n kpl havaintoja, jolloin yksi boostrap-otos $\mathcal{D}^*$ on n-havainnon kokoinen otos aineistosta. Havaintojen poiminta bootstrap-otokseen tehdään satunnaisesti ja takaisin palauttaen, ja siten yksittäisen havainnon otokseen päätymistodennäköisyys on $1/n$. Näitä bootstrap-otoksia voidaan luoda B kpl, joka valitaan riittävän suureksi (esimerkiksi B=1000). Kiinnostavalle parametrille θ lasketaan estimaatti $\widehat{\theta}_b^*$ erikseen jokaisesta bootstrap-otoksesta $\mathcal{D}_b^*$, jonka jälkeen estimaatit järjestetään suuruusjärjestykseen. Parametrin θ luottamusväli merkitsevyystasolla α konstruoidaan laskemalla näistä järjestetyistä estimaateista prosenttipisteet $\widehat{\theta}^{*(\alpha/2)}$ ja $\widehat{\theta}^{*(1-\alpha/2)}$, jolloin luottamusväliksi saadaan

$$(\widehat{\theta}^{*(\alpha/2)}, \widehat{\theta}^{*(1-\alpha/2)}).$$

4. Kausaalimalli sitoutuneisuusasteelle ja tavoittavuudelle

Ajatellaan Facebook-julkaisujen kehityskaarta sitoutuneisuusasteen ja tavoittavuuden osalta seuraavasti: Sisällöntuottaja julkaisee tiettynä ajanhetkenä sisällöltään tietyntyypisen julkaisun. Tämän julkaisun näkee tietty määrä käyttäjiä, joista osa reagoi siihen. Käyttäjien julkaisun tykkäämisten, jakamisten ja kommentointien voidaan olettaa osaltaan edesauttavan julkaisun leviämistä. Tämä taas generoi uusia tavoitettuja käyttäjiä, joista osa reagoi julkaisuun ja niin edelleen. Tältä ajattelupohjalta voidaan muodostaa asetelmat mielenkiinnon kohteena olevia tutkimuskysymyksiä pohjustamaan: Tavoitteena on sitouttaa käyttäjiä julkaisuun tuottamalla merkityksellistä sisältöä. Toisaalta Facebookin jatkuvassa syötevirrassa oikeaan aikaan lähetetyllä julkaisulla (vuorokaudenaika) ja käyttäjille merkityksellisellä asiasisällöllä (sisältötyyppi) voi olla mahdollista tavoittaa enemmän käyttäjiä.

Esitellään seuraavaksi kappaleen 3 menetelmien soveltamista keskimääräisten kausaalivaikutusten estimointiin sitoutuneisuusasteelle ja tavoittavuudelle. Kiinnostuksena on estimoida keskimääräinen kausaalivaikutus julkaisun sitoutuneisuusasteelle, kun julkaisun sisältötyyppiin kohdistetaan tietoista valintaa. Tämän lisäksi estimoidaan keskimääräinen kausaalivaikutus julkaisun tavoittavuudelle, kun valintaa toteutetaan julkaisun sisältötyypin lisäksi myös julkaisun vuorokaudenajan osalta.

4.1. Kausaaligraafi sitoutuneisuusasteelle

Muodostetaan ensin julkaisun sitoutuneisuusasteen kausaaligraafi (graafi 1), joka on esitetty kuvassa 3. Sitoutuneisuusaste voidaan ajatella todennäköisyysmielessä julkaisuun sitoutuneiden suhteena julkaisun nähneisiin käyttäjiin, missä julkaisun nähneiden lukumäärä oletetaan kiinteäksi. Graafi 1 koostuu seuraavista tutkijan tekemistä oletuksista aineiston muuttujien välisille suhteille.

Kumulatiivisella ajalla tarkoitetaan yleistä ajan kulumista. Ajan kuluessa käyttäjien toiminta sosiaalisessa mediassa voi muuttua, mikä vaikuttaa suoraan sitoutuneisuusasteeseen. Ajan kuluminen tuo muutoksia myös yrityksen sosiaalisen median strategian käytännön toteutukseen, kun toimintaa kehitetään analyysien perusteella säännöllisesti. Kumulatiivisen ajan oletetaankin vaikuttavan suoraan kaikkiin graafin

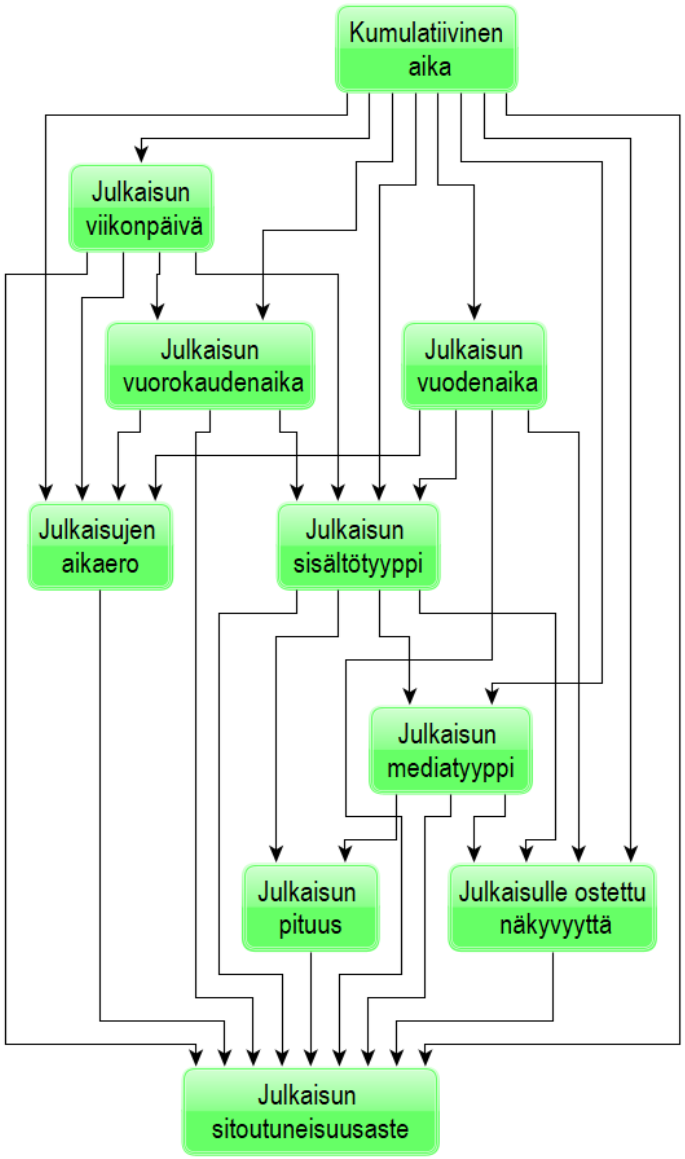
muuttujiin, paitsi julkaisun pituuteen. Julkaisun viikonpäivä vaikuttaa julkaisujen sitoutuneisuusasteeseen, aikaeroon, vuorokaudenaikaan ja sisältöön. Tarkasteltaessa viikonpäiviä jaoteltuna esimerkiksi arkipäivien ja viikonlopun suhteen, voidaan eroavaisuuksia näiden välillä olettaa sekä julkaisutahdissa, että julkaisun vuorokaudenajoissa. Samoin julkaisujen sisältötyypeillä voi olla erilaisia painotusajankohtia viikonpäivien välillä. Julkaisun viikonpäivän vaikutusta suoraan sitoutuneisuuteen voidaan myös pitää mahdollisena.

Julkaisun vuodenaika vaikuttaa puolestaan ostettuun näkyvyyteen, julkaisujen aikaeroon, sisältötyyppiin ja sitoutuneisuusasteeseen. Tiettyihin vuodenaikoihin osuvat sesongit ja juhlapyhät voivat vaikuttaa viestintään myös pidemmällä aikavälillä: Julkaisutiheys voi kasvaa, julkaisulle voidaan haluta ostaa lisää näkyvyyttä ja sisällöntuotannossa voi painottua tietty sisältöteema.

Julkaisun vuorokaudenajan osalta tilanne on samankaltainen kuin edellä. Vuorokaudenaika vaikuttaa siihen, kuinka kauan edellisestä julkaisusta on kulunut aikaa. Oletettavasti vuorokauden eri aikoina voivat painottua erilaiset julkaisujen sisältötyypit, mutta vuorokauden ajalla ei oleteta kuitenkaan olevan vaikutusta suoraan julkaisulle ostettuun näkyvyyteen. Sekä julkaisun vuorokaudenajan, että vuodenajan kohdalla vaikutusta suoraan sitoutuneisuuteen voidaan pitää yhtä lailla mahdollisena, kuin julkaisun viikonpäivän kohdalla.

Julkaisun aikaero vaikuttaa ainoastaan suoraan sitoutuneisuusasteeseen. Oletuksena tämä liittyy käsitykseen olemassa olevasta optimaalisesta julkaisutiheydestä sitoutumisen toteutumiselle. Julkaisun sisältötyypillä on vaikutus ostettuun näkyvyyteen, julkaisun pituuteen, mediatyyppiin ja sitoutuneisuusasteeseen. Tietyn sisällön julkaisun leviämistä voidaan haluta lisätä ostamalla sille näkyvyyttä. Julkaisujen viestejä muotoillaan myös eri tavoilla: esimerkiksi osuustoimintaa esittelevässä julkaisussa voidaan hyödyntää viestin kerrontaa lyhyen tarinan muodossa, kun taas tarjouksia sisältävässä julkaisussa listatyyppinen esitystapa on yleistä. Asian sisältö vaikuttaa kirjoitustapaan, jolla on vaikutusta suoraan viestin pituuteen. Lisäksi tietyn tyyppistä sisältöä julkaistaan eri mediatyypin muodossa. Tähän antaa viitteitä myös esimerkiksi videoiden sijoittuminen Valtarin ja Inkisen (2018) toteuttamassa nelikenttäjaossa (Liite 1). Toimiva sisältö sitouttaa, jolloin oletettavasti sisältötyyppi vaikuttaa myös suoraan sitoutuneisuusasteeseen.

Julkaisun mediatyypillä on puolestaan vaikutusta julkaisun sitoutuneisuusasteeseen, pituuteen ja ostettuun näkyvyyteen. Esimerkiksi videot ja linkit voivat toimia parempina klikkausten tuottajina verrattuna tavalliseen kuvaliitteiseen julkaisuun. Myös



Kuva 3: Kausaaligraafi julkaisujen sitoutuneisuusasteelle (Graafi 1).

julkaisujen viestien pituuksissa voi olla eroavaisuuksia eri mediatyyppien välillä. Tietynlaisen mediatyypin julkaisulle voidaan haluta ostaa lisää näkyvyyttä, esimerkiksi videoiden levittämisen tehostamiseen. Sekä ostetun näkyvyyden, että julkaisun pituuden osalta vaikutus on ainoastaan suoraan sitoutuneisuusasteeseen. Esimerkiksi julkaisun viestin loppuosa saatetaan piilottaa linkin taakse viestin ollessa pitkä, mikä toimii potentiaalisena klikkauksien lähteenä. Ostettu näkyvyys puolestaan lisää julkaisun näkyvyyttä, jolloin on mahdollista tavoittaa uusia sitoutuvia käyttäjiä.

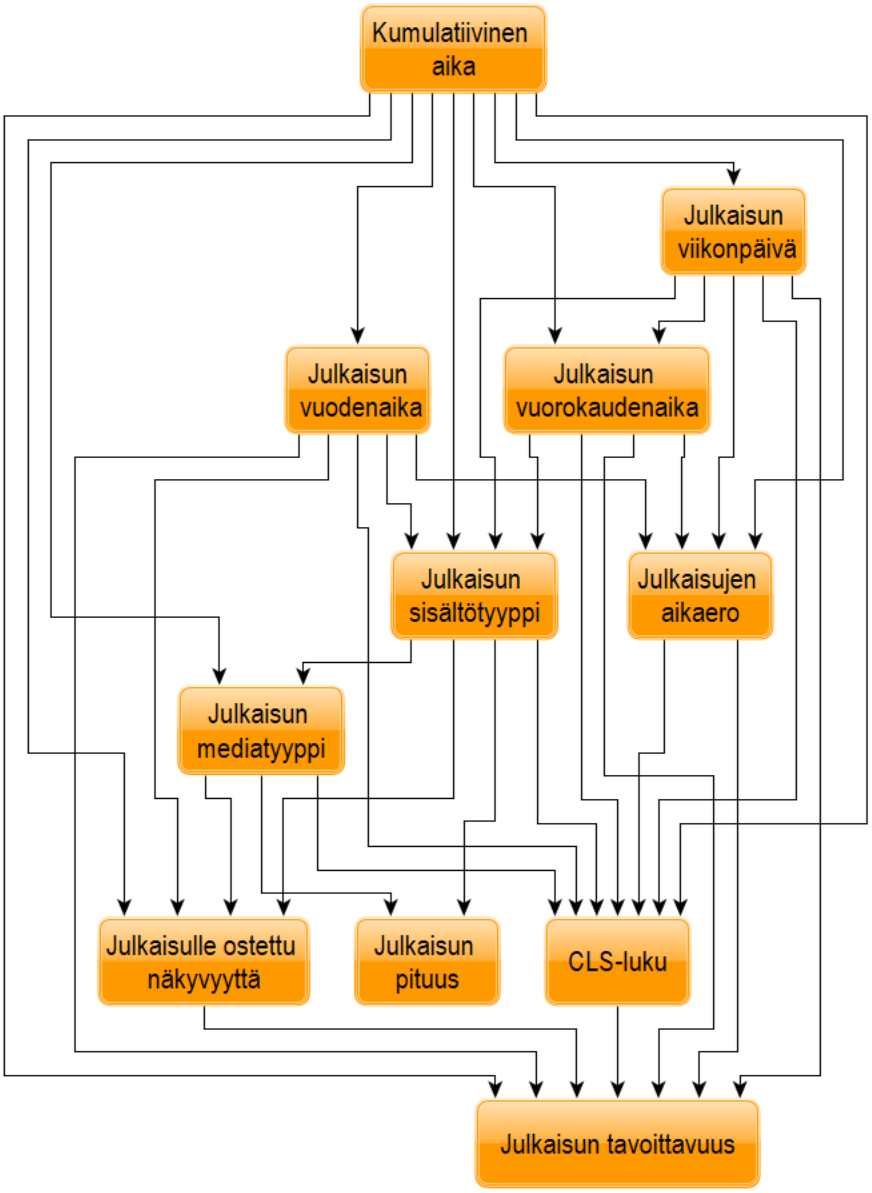
4.2. Kausaaligraafi tavoittavuudelle

Tarkastellaan seuraavaksi kausaaligraafia julkaisujen tavoittavuudelle (graafi 2), joka on esitetty kuvassa 4. Graafi 2 on perusrakenteeltaan hyvin samankaltainen kuin graafi 1, muutamia poikkeuksia lukuunottamatta. Kuten kappaleen 4 alussa kuvattiin, on oletettavaa, että julkaisuun suoritetut tykkäykset, kommentoinnit ja jaot voivat mahdollistaa uusien käyttäjien tavoittamisen julkaisulle. Tämän huomiointiin on graafiin 2 lisätty muuttuja CLS-luku, jonka oletetaan vaikuttavan suoraan julkaisun tavoittavuuteen. Suurin osa muuttujien välisistä suhteista perusteltiin graafin 1 kohdalla, joten täydennetään oletuksia nyt graafin 2 osalta.

Kumulatiivisella ajalla on suora vaikutus julkaisun tavoittavuuteen. Esimerkiksi ajan kuluessa tapahtuvat Facebookin käyttäjämäärien muutokset voivat heijastaa myös Hämeenmaan julkaisujen tavoittavuuteen. Lisäksi voidaan olettaa ajassa tapahtuvilla muutoksilla olevan vaikutusta käyttäytymiseen sosiaalisessa mediassa, mikä voi näkyä tykkäyksien, jakojen ja kommentointien määrässä.

Julkaisun vuodenaika, viikonpäivä ja vuorokaudenaika vaikuttavat suoraan sekä julkaisun tavoittavuuteen, että julkaisun CLS-lukuun. Julkaisu voi tavoittaa kokonaisuudessaan erilailla riippuen julkaisemisen ajankohdasta vuodenajan, viikonpäivän ja vuorokaudenajan tasoilla. Esimerkiksi tietty ajankohta voi olla erityisen otollinen julkaisun leviämiseen. Vastaavasti joku toinen ajankohta voi vaikuttaa siihen, kuinka aktiivisesti käyttäjät suorittavat julkaisuun kohdistettuja tykkäyksiä, jakoja ja kommentointeja.

Julkaisujen mediatyypin, sisältötyypin ja aikaeron osalta oletetaan myös suora vaikutus CLS-lukuun. Sisällöltään ja medialtaan erilaisten julkaisujen oletetaan eroavan toisistaan tykkäysten, kommentointien ja jakojen yhteismäärässä. Julkaisun aikaeron ajatellaan puolestaan vaikuttavan CLS-luvun lisäksi tavoittavuuteen. Tämän perusteena on aiemmin graafin 1 kohdalla esitetty käsitys sopivista julkaisuväleistä. Julkaisulle ostettu näkyvyys vaikuttaa suoraan tavoittavuuteen, mutta nyt graafissa 2 julkaisun pituuden ei oleteta vaikuttavan mihinkään toiseen muuttujaan.



Kuva 4: Kausaaligraafi julkaisujen tavoittavuudelle (Graafi 2).

4.3. Kausaalivaikutusten identifiointi ja estimointi

Tässä kappaleessa tutkitaan kiinnostavien kausaalivaikutusten identifioituvuutta edellä esitettyjen graafien pohjalta. Lisäksi estimoidaan keskimääräiset kausaalivaikutukset kiinnostuksen kohteena oleville identifioituville kausaalivaikutuksille.

Käytetään seuraavia merkintöjä muuttujien osalta:

$$
\begin{aligned}
Y_1 &= \textit{Julkaisun sitoutuneisuusaste} \\
Y_2 &= \textit{Julkaisun tavoittavuus} \\
X_1 &= \textit{Julkaisun sisältötyyppi} \\
X_2 &= \textit{Julkaisun vuorokaudenaika} \\
Z_1 &= \textit{Kumulatiivinen aika} \\
Z_2 &= \textit{Julkaisun vuodenaika} \\
Z_3 &= \textit{Julkaisun viikonpäivä} \\
Z_4 &= \textit{Julkaisujen välinen aikaero} \\
Z_5 &= \textit{Julkaisun mediatyyppi} \\
Z_6 &= \textit{Julkaisulle ostettu näkyvyyttä} \\
Z_7 &= \textit{Julkaisun pituus} \\
Z_8 &= \textit{Julkaisun CLS-luku.}
\end{aligned}
$$

Merkitään lisäksi $\boldsymbol{Z} = \{Z_1, ..., Z_8\}$. Kiinnostuksena olevat kausaalivaikutukset voidaan nyt esittää kausaalijakaumina

$$
\begin{gathered}
P(Y_1 \,|\, do(X_1)) \ , \\
P(Y_2 \,|\, do(X_1), do(X_2)) \ .
\end{gathered}
$$

Molempien kausaalijakaumien kohdalla päädytään ID-algoritmin avulla takaovikorjauksen tilanteisiin. Identifioituvien kausaalivaikutusten jakaumat saadaan siten esitettyä havaintojakaumien avulla, jolloin

$$
P(Y_1 \,|\, do(X_1)) = \sum_{X_2, \boldsymbol{Z} \setminus \{Z_8\}} P(Y_1 \,|\, X_1, X_2, \boldsymbol{Z} \setminus \{Z_8\}) P(X_2, \boldsymbol{Z} \setminus \{Z_8\}) \ ,
$$

$$
P(Y_2 \,|\, do(X_1), do(X_2)) = \sum_{\boldsymbol{Z} \setminus \{Z_7\}} P(Y_2 \,|\, X_1, X_2, \boldsymbol{Z} \setminus \{Z_7\}) P(\boldsymbol{Z} \setminus \{Z_7\}) \ .
$$

Edellä identifioituvien kausaalivaikutusten kohdalla keskimääräisten kausaalivaikutusten estimointiin voidaan käyttää kappaleessa 3.1 esiteltyä suurten lukujen lakia hyödyntävää kaavaa. Tällöin keskimääräiset kausaalivaikutukset ovat muotoa

$$\begin{aligned}\theta_{Y_1;X_1=x_1} &= E(Y_1 \mid do(X_1 = x_1)) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(Y_1|X_1 = x_1, X_2 = x_{2i}, \boldsymbol{Z} \setminus \{Z_8\} = \boldsymbol{z_i}) \,,\end{aligned}$$

$$\begin{aligned}\theta_{Y_2;X_1=x_1,X_2=x_2} &= E(Y_2 \mid do(X_1 = x_1), do(X_2 = x_2)) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(Y_1|X_1 = x_1, X_2 = x_2, \boldsymbol{Z} \setminus \{Z_7\} = \boldsymbol{z_i}) \,.\end{aligned}$$

Seuraavaksi on valittava ja estimoitava edeltäville odotusarvoille $E(Y_1|X_1, X_2, \boldsymbol{Z} \setminus \{Z_8\})$ ja $E(Y_2|X_1, X_2, \boldsymbol{Z} \setminus \{Z_7\})$ tilastolliset mallit. Sitoutuneisuusasteen Y_{1i} voidaan olettaa noudattavan binomijakaumaa parametreilla m_i ja π_i, missä m_i on i:nnen julkaisun tavoittama käyttäjämäärä, ja π_i i:nnenteen julkaisuun sitoutumisen todennäköisyys. Julkaisun pituuden ja sitoutuneisuusasteen välinen suhde on epälineaarinen, jolloin sovitetaan pituuden kovariaattiin regressiosplini (k=4). Kantafunktioiden dimension k valinnassa hyödynnetään ristiinvalidointia kappaleessa 3.3 esitetyllä tavalla, laskemalla kymmenen kertaa suoritetuista K-10 ristiinvalidoinneista keskiarvo. Ristiinvalidoinnin lisäksi valinnassa hyödynnetään graafista tarkastelua, sekä AIC-kriteeriä.

Estimoitu yleistetty additiivinen malli odotusarvolle $E(Y_1|X_1, X_2, \boldsymbol{Z} \setminus \{Z_8\})$ on

$$\begin{aligned}&logit(\widehat{E}(Y_{1i}|X_1 = x_{1i}, X_2 = x_{2i}, \boldsymbol{Z} \setminus \{Z_8\} = \boldsymbol{z_i})) \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \widehat{\beta}_3 z_{1i} + ... + \widehat{\beta}_8 z_{6i} + \widehat{s}(z_{7i}) \quad ,\end{aligned}$$

missä $\widehat{s}(z_7)$ on julkaisun pituuteen Z_7 sovitettu regressiosplinifunktio. Liitteessä 2 on esitetty edelläolevan sitoutuneisuusasteen mallin estimoidut regressiokertoimet, niiden keskivirheet, testisuureet ja p-arvot, sekä mallin ristiinvalidointiin liittyvän suureen arvo ja AIC-arvo. Pearsonin residuaalien ja mallin antamien sovitteiden välinen jäännöskuvio (Liite 4) ei paljasta huomattavaa systemaattisuutta jäännöksissä ja mallia voidaan tämän tarkastelun osalta pitää sopivana.

Mallinnettaessa odotusarvoa $E(Y_2|X_1, X_2, \boldsymbol{Z} \setminus \{Z_7\})$, tavoittavuuden Y_{2i} voidaan olettaa noudattavan negatiivista binomijakaumaa parametreilla μ_i ja α, missä μ_i on julkaisun i tavoittavuuden odotusarvo ja α mallin ylihajontaparametri. Julkaisun tavoittavuuden ja kumulatiivisen ajan osalta havaittu epälineaarinen yhteys huomioidaan sovittamalla regressiosplini (k=5) kumulatiiviseen aikaan. Myös tämän sovituksen valintaa kantafunktioden dimension k osalta toteutetaan tarkastelemalla kymmenen kertaa toteutettujen K-10 ristiinvalidointien keskiarvoa. Lisäksi hyödynnetään graafisia tarkasteluja, sekä AIC-kriteeriä.

Estimoitu yleistetty additiivinen malli odotusarvolle $E(Y_2|X_1, X_2, \boldsymbol{Z} \setminus \{Z_7\})$ on

$$\begin{aligned} &log(\widehat{E}(Y_{2i}|X_1 = x_{1i}, X_2 = x_{2i}, \boldsymbol{Z} \setminus \{Z_7\} = \boldsymbol{z_i})) \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \widehat{\beta}_3 z_{2i} + ... + \widehat{\beta}_7 z_{6i} + \widehat{\beta}_8 z_{8i} + \widehat{s}(z_{1i}). \end{aligned}$$

Mallissa $\widehat{s}(z_1)$ on kumulatiiviseen aikaan Z_1 sovitettu regressiosplinifunktio. Liitteeseen 3 on koottu tämän tavoittavuusmallin estimoidut regressiokertoimet, niiden keskivirheet, testisuureet, sekä p-arvot. Liitteessä 3 on esitetty lisäksi mallin ristiinvalidointiin liittyvän suureen arvo, AIC-arvo, sekä estimoitu ylihajontaan liittyvä parametri. Tarkasteltaessa Pearsonin jäännösten ja mallin antamien sovitteiden hajontakuviota (Liite 5), havaitaan jäännösten jakautuvan suhteellisen tasaisesti nollan ympäristöön tyypillisimmillä arvoilla. Sen sijaan suuremmilla sovitteiden arvoilla havaitaan painotusta negatiivissa jäännöksissä, sekä lievää heteroskedastisuutta.

Mallien avulla voidaan nyt estimoida kausaalivaikutukset kiinnittämällä kontrolloinnin kohteena olevien muuttujien arvot. Tällöin sitoutuneisuusasteen mallissa asetetaan julkaisun sisältötyypin arvoksi $X_1 = x_1$ (havainnon x_{1i} sijaan). Sisältötyypin muuttuja on neliluokkainen, ja siten kausaalivaikutusten estimaatit lasketaan vuorotellen eri sisältötyypin arvoilla yli havaintojoukon. Keskimääräisten kausaaalivaikutusten estimaatit asetetulla X_1:n arvolla saadaan nyt laskemalla mallin antamien kausaalivaikutusten estimaattien keskiarvo

$$\begin{aligned} \widehat{\theta}_{Y_1;X_1=x_1} &= \widehat{E}(Y_1 \,|\, do(X_1 = x_1)) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{E}(Y_{1i}|X_1 = x_1, X_2 = x_{2i}, \boldsymbol{Z} \setminus \{Z_8\} = \boldsymbol{z_i}) \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} logit^{-1} \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_{2i} + \widehat{\beta}_3 z_{1i} + ... + \widehat{\beta}_8 z_{6i} + \widehat{s}(z_{7i}) \right). \end{aligned}$$

Julkaisun tavoittavuuden mallissa kontrolloinnin kohteena olevia muuttujia on kaksi: julkaisun sisältötyyppi ja vuorokaudenaika. Kausaalivaikutukset estimoidaan nyt erikseen kaikkien valintojen yhdistelmillä, laskemalla mallin antamat estimaatit yli havintojoukon. Vastaavasti asetetuilla muuttujien X_1 ja X_2:n arvoilla keskimääräisten kausaaalivaikutusten estimaatit saadaan mallin antamien kausaalivaikutusten estimaattien keskiarvona, jolloin

$$\begin{aligned}
\widehat{\theta}_{Y_2;X_1=x_1,X_2=x_2} &= \widehat{E}(Y_2 \mid do(X_1 = x_1), do(X_2 = x_2)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{E}(Y_{2i}|X_1 = x_1, X_2 = x_2, \boldsymbol{Z} \setminus \{Z_7\} = \boldsymbol{z_i})\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 z_{2i} + ... + \widehat{\beta}_7 z_{6i} + \widehat{\beta}_8 z_{8i} + \widehat{s}(z_{1i})\right).
\end{aligned}$$

Kausaalivaikutusten estimaateista voidaan lisäksi laskea 95 %:n kvantiilivälit havainnollistamaan, mille välille 95 % estimaateista sijoittuu. Keskimääräisen kausaalivaikutuksen estimoinnin tarkkuutta voidaan tutkia luottamusvälitarkasteluilla, käyttämällä kappaleessa 3.3 esitettyä prosenttipiste bootstrap-menetelmää. Tällöin saadaan arvio havaintoaineistosta johtuvalle keskimääräisen kausaalivaikutuksen θ estimoinnin epävarmuudelle. Merkitsevyystasoksi valitaan $\alpha = 0.05$, jolloin lasketaan 95 %:n luottamusvälit.

4.4. Tulokset ja johtopäätökset

Tarkastellaan seuraavaksi keskimääräisten kausaalivaikutusten estimaatteja, 95 %:n luottamusvälejä, sekä estimoitujen kausaalivaikutusten 95 %:n kvantiilivälejä. Sitoutuneisuusasteen osalta tulokset on koottu taulukkoon 3.

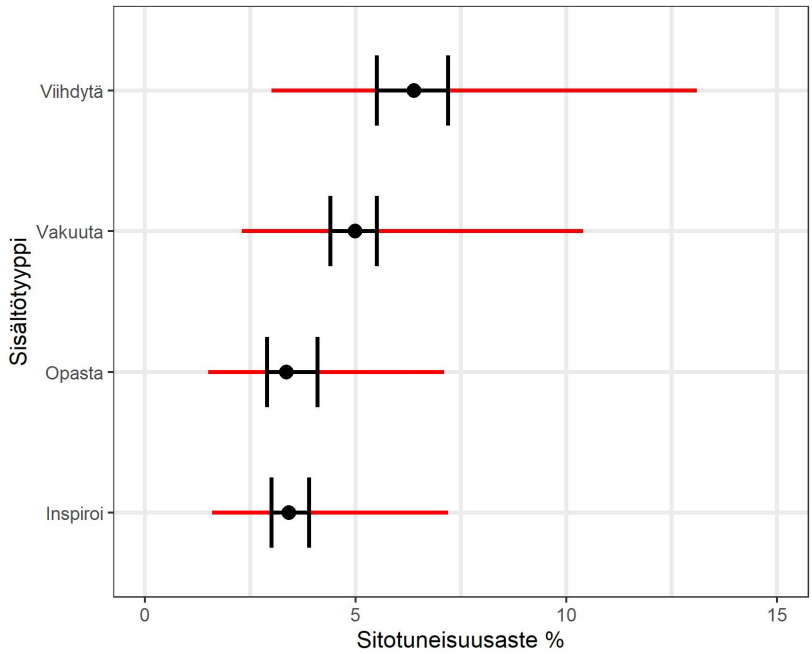
Taulukko 3: Eri sisältötyyppien keskimääräisten kausaalivaikutusten estimaatit sitoutuneisuusasteelle, keskimääräisten kausaalivaikutusten 95 %:n luottamusvälit ja estimoitujen kausaalivaikutusten 95 %:n kvantiilivälit.

Sisältötyyppi $X_1 = x_1$	Keskimääräisen kausaalivaikutuksen estimaatti $\widehat{\theta}_{Y_1;X_1=x_1}$	$\theta_{Y_1;X_1=x_1}$:n 95 %:n luottamusväli	$\widehat{E}(Y_1\|X_1 = x_1, X_2, \boldsymbol{Z} \setminus \{Z_8\})$:n 95 %:n kvantiiliväli
Inspiroi	3.4 %	(3.0 %, 3.9 %)	(1.6 %, 7.2 %)
Opasta	3.4 %	(2.9 %, 4.1 %)	(1.5 %, 7.1 %)
Vakuuta	5.0 %	(4.4 %, 5.5 %)	(2.3 %, 10.4 %)
Viihdytä	6.4 %	(5.5 %, 7.2 %)	(3.0 %, 13.1 %)

Taulukosta 3 havaitaan, että viihdyttävä ja vakuuttava julkaisu erottuvat muista sisältötyypeistä keskimääräisen kausaalivaikutuksen estimaateiltaan sitoutuneisuusasteen osalta. Viihdyttävä julkaisu toimii sitouttavimpana julkaisuna 6.4 %:n sitoutuneisuusasteellaan (luottamusväli 5.5 % – 7.2 %) ja vakuttava julkaisu toiseksi sitouttavimpana julkaisuna 5.0 %:n sitoutuneisuusasteellaan (luottamusväli 4.4 % – 5.5 %). Estimoidut kausaalivaikutukset jakautuvat 95 % kvantiilivälin mukaan viihdyttävän julkaisun osalta välille 3.0 % – 13.1 % ja vakuuttavan julkaisun osalta välille 2.3 % – 10.4 %. Inspiroivan ja opastavan sisältötyypin osalta sitoutuneisuusaste on molempien osalta 3.4 % (luottamusvälit 3.0 % – 3.9 % ja 2.9 % – 4.1 %). Kausaalivaikutusten 95 %:n kvantiilivälit ovat näiden sisältötyyppien osalta myös hyvin lähellä toisiaan siten, että inspiroivan julkaisun osalta se on 1.6 % – 7.2 % ja opastavan julkaisun osalta 1.5 % – 7.1 %.

Kuvasta 5 nähdään tarkemmin julkaisun sitoutuneisuusasteen ja sisältötyyppien väliset suhteet keskimääräisten kausaalivaikutusten estimaattien, 95 %:n luottamusvä-

lien ja kausaalivaikutusten estimaattien 95 %:n kvantiilivälien osalta. Luottamusväli on viihdyttävän julkaisun osalta levein, ja siten sitoutuneisuusasteen keskimääräisen kausaalivaikutuksen estimointiin liittyy tämän sisältötyypin osalta eniten havaintoaineistosta johtuvaa epävarmuutta. Sen sijaan inspiroivan julkaisun osalta luottamusväli on kapein ja estimointi näin ollen tarkin. Taulukosta 1 nähtiin, että inspiroivia julkaisuja on havaittu selvästi muita sisältötyyppejä enemmän. Viihdyttäviä julkaisuja on puolestaan huomattavasti vähemmän suhteessa muihin. Nämä huomiot tukevat edellä tehtyjä päätelmiä sisältötyyppien keskimääräisten kausaalivaikutusten estimoinnin tarkkuuden osalta. Kvantiilivälien osalta voidaan havaita kausaalivaikutusten estimaattien käsittävän laajan vaihteluvälin. Viihdyttävä julkaisu poikkeaa selvästi muista sisältötyypeistä kvantiilivälin ylärajan suhteen, mutta myös vakuuttavan julkaisun ero ylärajassa inspiroivaan ja opastavaan julkaisuun on huomattava. Tämä antaa viitettä viihdyttävän ja vakuuttavan julkaisun kyvystä sitouttaa myös poikkeuksellisen paljon käyttäjiä julkaisuun.



Kuva 5: Keskimääräisten kausaalivaikutusten estimaatit (•) sitoutuneisuusasteelle, 95 %:n luottamusvälit (mustat pystyviivat), ja kausaalivaikutusten 95 %:n kvantiilivälit (punaiset poikittaiset viivat).

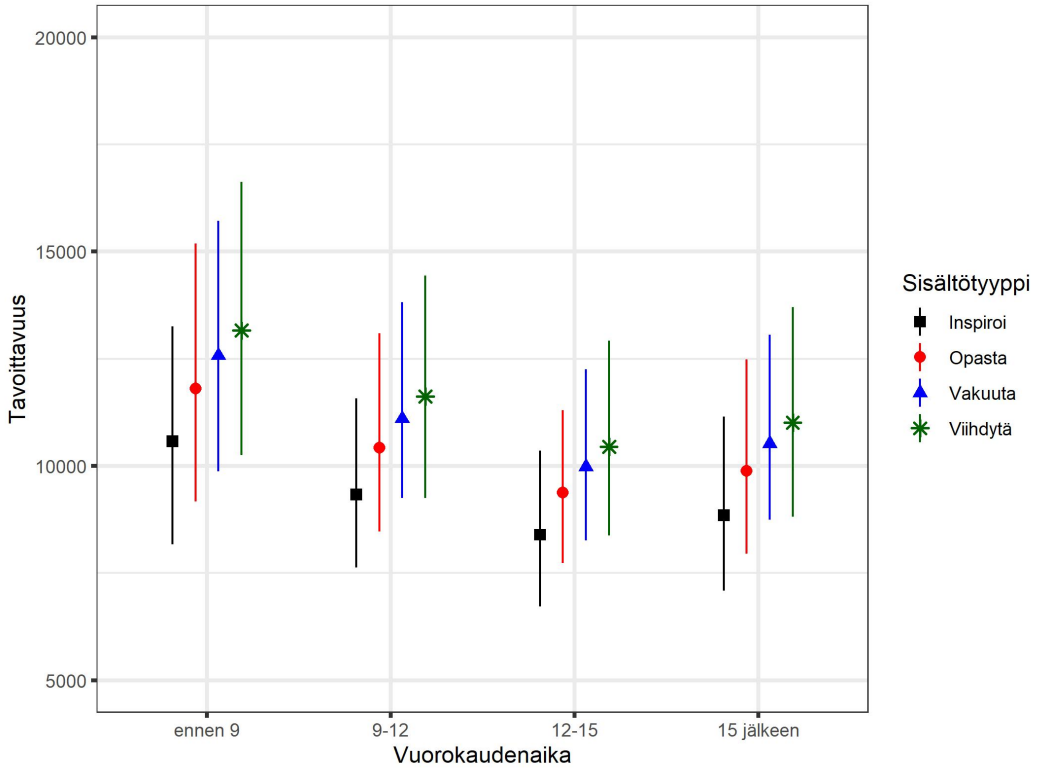
Tarkastellaan seuraavaksi julkaisun tavoittavuutta. Taulukossa 4 on esitelty keskimääräisten kausaalivaikutusten estimaatit, 95 %:n luottamusvälit, sekä estimoitujen kausaalivaikutusten 95 %:n kvantiilivälit tämän vastemuuttujan osalta.

Taulukko 4: Eri sisältötyyppien ja vuorokaudenaikojen keskimääräisten kausaalivaikutusten estimaatit tavoittavuudelle, keskimääräisten kausaalivaikutusten 95 %:n luottamusvälit ja estimoitujen kausaalivaikutusten 95 %:n kvantiilivälit.

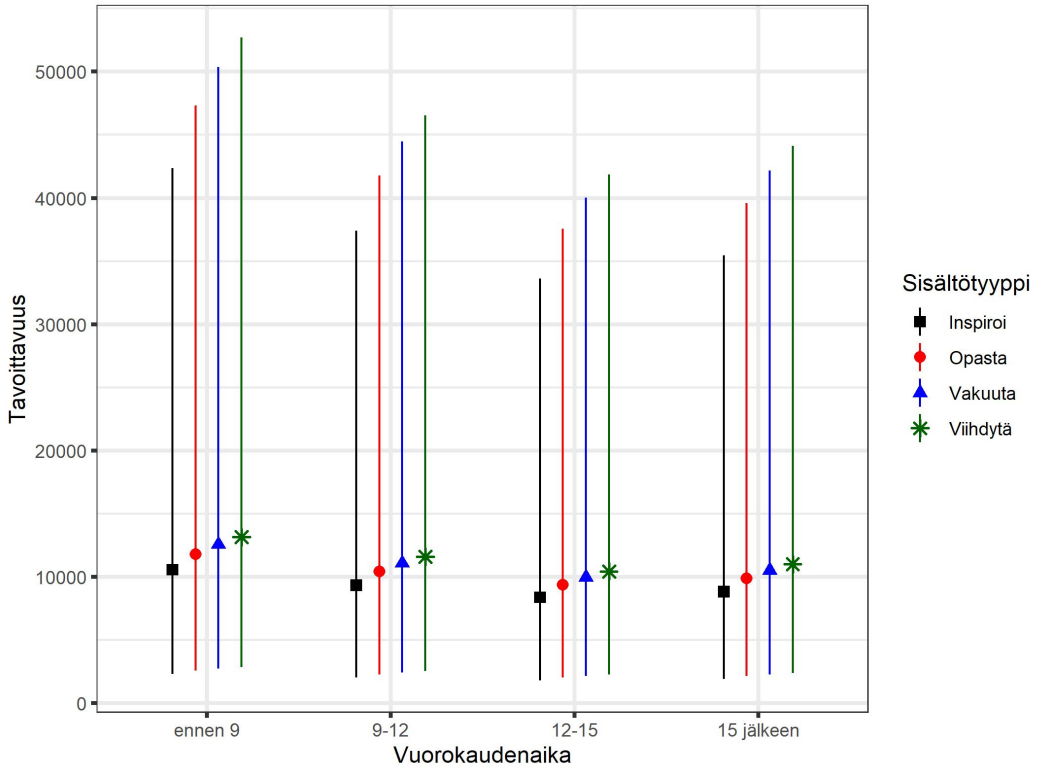
Vuorokaudenaika $X_2 = x_2$	Sisältötyyppi $X_1 = x_1$	Keskimääräisen kausaalivaikutuksen estimaatti $\widehat{\theta}_{Y_2;X_1=x_1,X_2=x_2}$	$\theta_{Y_2;X_1=x_1,X_2=x_2}$:n 95 %:n luottamusväli	$\widehat{E}(Y_2\|X_1 = x_1, X_2 = x_2, \boldsymbol{Z} \setminus \{Z_7\})$:n 95 %:n kvantiiliväli
– 8:59	Inspiroi	10570	(8173, 13256)	(2307, 42354)
– 8:59	Opasta	11804	(9166, 15182)	(2576, 47299)
– 8:59	Vakuuta	12573	(9871, 15718)	(2744, 50378)
– 8:59	Viihdytä	13158	(10254, 16623)	(2872, 52722)
9:00 – 11:59	Inspiroi	9333	(7630, 11572)	(2037, 37398)
9:00 – 11:59	Opasta	10423	(8463, 13095)	(2275, 41764)
9:00 – 11:59	Vakuuta	11101	(9250, 13820)	(2423, 44482)
9:00 – 11:59	Viihdytä	11618	(9254, 14442)	(2536, 46553)
12:00 – 14:59	Inspiroi	8393	(6717, 10354)	(1832, 33631)
12:00 – 14:59	Opasta	9373	(7731, 11293)	(2046, 37557)
12:00 – 14:59	Vakuuta	9983	(8261, 12253)	(2179, 40002)
12:00 – 14:59	Viihdytä	10448	(8382, 12920)	(2280, 41864)
15:00 –	Inspiroi	8847	(7084, 11153)	(1931, 35450)
15:00 –	Opasta	9880	(7945, 12485)	(2157, 39589)
15:00 –	Vakuuta	10523	(8746, 13054)	(2297, 42166)
15:00 –	Viihdytä	11013	(8810, 13703)	(2404, 44128)

Tarkasteltaessa julkaisun tavoittavuuden keskimääräisten kausaalivaikutusten estimaatteja, huomataan selviä eroja vuorokaudenaikojen välillä. Korkeimmat tavoittavuudet saavutetaan ennen klo 9:ää toteutetuilla julkaisuilla. Tähän verrattuna klo 9-12 julkaistujen osalta tavoittavuus on yli tuhat käyttäjää vähemmän kaikkien sisältötyyppien osalta. Pienimmät tavoittavuudet ovat klo 12-15 toteutetuilla julkaisuilla, tavoittavuuden ollessa noin tuhat käyttäjää vähemmän klo 9-12 julkaisuihin verrattuna. Klo 15 jälkeisillä julkaisuilla havaitaan tavoittavuudessa noin viidensadan käyttäjän suuruinen lisäys klo 12-15 toteutettuihin julkaisuihin verrattuna. Kun tarkastellaan keskimääräisen kausaalivaikutuksen estimaatteja sisältötyyppien osalta, havaitaan parhaiten tavoittavaksi julkaisuksi viihdyttävä julkaisu ja toiseksi parhaiten tavoittavaksi vakuuttava julkaisu. Tämän jälkeen tulevat opastava ja inspiroiva julkaisu.

Sisältötyypin ja vuorokaudenajan osalta parhaiten tavoittava julkaisu on ennen klo 9 toteutettu viihdyttävä julkaisu, keskimääräisen kausaalivaikutuksen estimaatin ollessa 13158 tavoitettua käyttäjää. Vähiten tavoittava julkaisu on puolestaan klo 12-15 julkaistu inspiroiva julkaisu, jonka keskimääräisen kausaalivaikutuksen estimaatti on 8393 käyttäjää. Näiden kahden julkaisun yhdistelmän erotus on peräti 4765 tavoitettua käyttäjää. Kuvassa 6 on esitetty keskimääräisten kausaalivaikutusten estimaatit tavoittavuuden osalta, sekä keskimääräisten kausaalivaikutusten 95 %:n luottamusvälit. Kuvasta nähdään havaittu epälineaarinen suhde estimaateissa vuorokaudenaikojen välillä. Keskimääräisten kausaalivaikutusten luottamusvälitarkastelujen osalta leveimmät luottamusvälit ovat julkaisuilla ennen klo 9:ää. Luottamusvälien leveydet muiden vuorokaudenaikojen suhteen ovat lähellä toisiaan. Kuvaa 6 ja taulukkoa 4 tarkastelemalla voidaan havaita sisältötyyppien osalta leveimmät luottamusvälit viihdyttävälle julkaisulle. Kapeimmat luottamusvälit ovat pääsääntöisesti inspiroivalla julkaisulla, lukuunottamatta klo 12-15 osalta. Kuvassa 7 on keskimääräisten kausaalivaikutusten estimaatit tavoittavuudelle esitettynä estimoitujen kausaalivaikutusten 95 %:n kvantiiliväleillä. Kuten sitoutuneisuusasteen osalta, myös tavoittavuuden kohdalla voidaan havaita kausaalivaikutusten estimaattien kvantiilivälien käsittävän laajan vaihteluvälin. Kuvasta huomataan kvantiilivälien alarajojen pysyvän suhteellisen lähellä toisiaan sekä sisältötyyppien, että vuorokaudenajan suhteen. Sen sijaan kvantiilivälien ylärajat eroavat selvemmin toisistaan. Näiden avulla havaitaan, että ennen klo 9:ää julkaistuilla on leveimmät kvantiilivälit ja klo 12-15 julkaistuilla kapeimmat. Myös sisältötyyppien osalta havaitaan eroja kvantiiliväleissä siten, että viihdyttävällä julkaisulla se on levein ja inspiroivalla kapein. Nämä havainnot tukevat jo aiemmin tehtyjä päätelmiä toimivimmista julkaisuista tavoittavuuden osalta vuorokaudenajan ja sisältötyypin suhteen.



Kuva 6: Sisältötyyppien ja vuorokaudenaikojen keskimääräisten kausaalivaikutusten estimaatit ja keskimääräisten kausaalivaikutusten 95 %:n luottamusvälit tavoittavuudelle.



Kuva 7: Sisältötyyppien ja vuorokaudenaikojen keskimääräisten kausaalivaikutusten estimaatit ja kausaalivaikutusten estimaattien 95 %:n kvantiilivälit tavoittavuudelle.

4.5. Kausaalimallin hyödyntäminen Facebook-markkinoinnissa

Edellä estimoituja keskimääräisiä kausaalivaikutuksia (taulukot 3 ja 4) voidaan hyödyntää Hämeenmaan Facebook-markkinoinnin suunnittelussa julkaisujen toteutuksen osalta. Havainnollistetaan tätä olettamalla seuraava hypoteettinen esimerkki:

Markkinointiyksikön suunnitelmissa on toteuttaa tulevaan kampanjaan liittyvää viestintää, jossa halutaan käyttää Facebookia. Yksikkö pohtii, pitäisikö julkaisu toteuttaa esimerkiksi aamulla iltapäivän sijaan, jotta julkaisu lähtisi leviämään tehokkaasti ja tavoittaisi siten mahdollisimman monta käyttäjää? Data-analyytikko esittää tekemänsä kausaalimallin pohjalta ehdotuksen julkaisuajankohdasta aamuksi iltapäivän sijaan. Kyseessä on inspiroiva julkaisu (kampanja) ja toteutusajankohdan vaihtoehdot ovat ennen klo 9 ja klo 12-15: Keskimääräisten kausaalivaikutusten estimaattien erotus julkaisun tavoittavuudessa on tällöin

$$
\begin{aligned}
&E(Y_2 \,|\, do(X_1 = Inspiroi), do(X_2 = ennen\, klo\, 9)) - \\
&E(Y_2 \,|\, do(X_1 = Inspiroi), do(X_2 = klo\, 12-15)) \\
&= 10570 - 8393 = 2177.
\end{aligned}
$$

Ero on siten 2177 tavoitettua käyttäjää. Edelleen voidaan laskea lisäys sitoutuneissa käyttäjissä käyttämällä sitoutuneisuusasteen kausaalimallia. Estimoitu keskimääräinen kausaalivaikutus inspiroivan julkaisun sitoutuneisuusasteelle on $E(Y_1 \,|\, do(X_1 = Inspiroi)) = 3.4\%$. Tällöin ennen klo 9:ää toteutettu julkaisu saa $2177 \cdot 0.034 \approx 74$ sitoutunutta käyttäjää enemmän klo 12-15 toteutettuun julkaisuun verrattuna.

Kampanjaan liittyvien Facebook-markkinoinnin kautta toteutuneiden ostojen estimoiminen ei luonnollisestikaan ole edeltävien mallien avulla mahdollista. Julkaisun nähnyttä käyttäjää ajatellaan kuitenkin markkinointiviestin nähneenä potentiaalisena asiakkaana. Julkaisuun sitoutunutta käyttäjää ei voida välttämättä suoraan pitää todennäköisempänä ostajana. Sen sijaan sitoutuneen käyttäjän voidaan olettaa olevan markkinointiviestin asiasisällöstä mahdollisesti kiinnostuneempi ja sitä kautta vahvemman asiakassuhteen omaava.

5. Pohdinta

Tämän tutkielman tavoitteena oli muodostaa toimintasuosituksia Osuuskauppa Hämeenmaan Facebook-julkaisuille kausaalimallin avulla, kun tarkasteltavina KPI-mittareina ovat julkaisun sitoutuneisuusaste ja tavoittavuus. Tutkielmasta saatuja tuloksia voidaan käyttää ohjaamaan julkaisujen toteutusta tavoittavuuden ja sitoutuneisuusasteen näkökulmasta toimivampiin julkaisuihin, ja siten antamaan tukea Hämeenmaan Facebook-markkinoinnin toteutukseen asiantuntemuksen rinnalle. Havaintoaineiston avulla tehty kausaalimallinnus mahdollistaa Facebookissa toteutettavan markkinointiviestinnän kehitystoiminnan suunnittelun ilman tarvetta erilliselle kokeelliselle tutkimukselle, jolloin säästetään resursseja. Kokeellisen tutkimuksen tekeminen Facebook-julkaisujen suorituskyvyn arvioinnissa ei välttämättä ole edes mahdollista, johtuen käytännön Facebook-viestinnän ja sitä kautta julkaisujen sidonnaisuudesta senhetkisiin liiketoiminnallisiin tavoitteisiin.

Käytetty kausaalimallinnus voidaan nähdä relevantimpana lähestymistapana tutkimusongelmaan verrattuna ennustemalliin. Kun Facebook-julkaisemiseen liittyviä toimintatapoja halutaan lähteä kehittämään tai kokeilemaan muutosta, ajatellaan, että halutaan tietoisesti muuttaa jotakin toiminnassa. Tähän ajatustapaan sopii kausaalipäättely, jossa otetaan samalla huomioon myös asioiden välisiä syy-seuraussuhteita.

Toteutetusta analyysistä nousee esille huomioita liittyen aineistoon, mallinnukseen ja tutkijan subjektiivisten näkemysten osuuteen. Kuten kappaleessa 3.1 mainittiin, sisältyy kausaalipäättelyyn tutkijan näkemystä asioiden välisistä suhteista. Tämä näkemys otetaan huomioon kausaaligraafeja muodostettaessa, jolla on oleellinen vaikutus lopputuloksiin. Tutkielman muuttujien välisten syy-seuraussuhteiden määrityksessä oltaisiinkin voitu päätyä erilaisiin graafeihin toisten tutkijoiden toimesta. Huomoitavaa on myös se, että graafissa ei ole otettu huomioon mahdollisia havaitsemattomia taustamuuttujia.

Tutkielmassa käytetty aineisto on suhteellisen pieni 791:llä havainnollaan. Suuremmalla käytettävissä olevalla aineistolla keskimääräisten kausaalivaikutusten estimaattien osalta voitaisiin päästä saatuja kapeampiin luottamusväleihin ja siten tarkempiin estimaatteihin. Luonnollisesti tarkkuuden parantamiseksi tulee pohdittavaksi myös vaihtoehtoisen bootstrap-toteutuksen mahdollisuus luottamusvälien muodostamisessa.

Aineiston pienestä koosta johtuen myös kategoristen muuttujien luokkia päädyttiin yhdistelemään, jossa mahdollisia luokitteluvaihtoehtoja on useita. Luokkien yh-

distelyä toteutettiin tutkijan harkinnan mukaan, pyrkien huomiomaan myös luokkien kokojen tasapainoisuus. Näin meneteltiin julkaisun vuorokaudenaikojen, kuukausien ja mediatyypin kohdalla.

Julkaisun sisältötyyppi kontrolloitavan muuttujan roolissa on olennainen osa tutkimusta ja myös tähän muuttujaan liittyy tutkijan subjektiivista näkemystä. Luokittelussa on huomioitavaa, että julkaisun oletetaan kuuluvan sisältötyypin osalta vain ja ainoastaan yhteen luokkaan. Tämä voi olla ongelmallista niiden julkaisujen luokittelussa, joiden sisällöllinen luokka ei välttämättä ole yksiselitteinen. Lisäksi toisen luokittelijan toimesta oltaisiin voitu päätyä erilaisiin luokitteluihin, esimerkiksi erilaisten rajatapausten osalta.

Julkaisun sisältö ei luonnollisestikaan rajoitu neliluokkaiseen muuttujaan. Tutkielmassa käytetyn sisältötyyppi-muuttujan sijasta sisällöllisten tekijöiden huomiointiin voitaisiin pyrkiä käyttämään paljon informatiivisempia muuttujia. Aineisto ei kuitenkaan itsessään sisältänyt tietoa sisällöllisistä tekijöistä, jolloin päädyttiin yhteen yksinkertaisempaan muuttujaan. Varteenotettava keino sisällöllisten muuttujien muodostamisessa olisi erilaisten koneoppimis- ja tiedonlouhintamenetelmien käyttö, koska julkaisu käsittää runsaasti sekä visuaalista, että tekstillistä informaatiota. Julkaisujen viesteihin toteutettiin tekstianalytiikkaa, etsimällä usein esiintyviä sanoja ja sanapareja. Merkittäviä hyötyjä ei näillä menetelmillä kuitenkaan tämän tutkielman kannalta saatu.

Julkaisun tavoittavuuden mallin valinnassa päädytiin negatiiviseen binomijakaumaan Poisson-jakauman sijaan ylihajonnan huomioimiseksi. Oletettaessa Poisson-jakauma tavoittavuudelle, voidaan ylihajonta havaita suoraan taulukosta 1 tarkastelemalla tavoittavuuden varianssin suurutta keskiarvoon verratuna. Myös sitoutuneisuusasteeseen käytetyn binomimallin osalta havaittiin ylihajontaa. Tutkittaessa mallin jäännösdevianssia, havaittiin sen olevan selvästi suurempi mallin vapausasteisiin verrattuna, ilmentäen ylihajonnan mallissa. Vaihtoehtoisena mallina binomijakaumalle voitaisiin käyttää esimerkiksi beta-binomiaalista jakaumaa, joka on eräs vaihtoehto ylihajontatilanteessa (McCullagh & Nelder, 1989). Tätä mallinnusta ei kuitenkaan toteutettu johtuen estimoinnissa käytetystä R-paketista. Hilbe (2011) esittää erääksi ylihajonnan syyksi oleellisten kovariaattien puuttumisen mallista. Tähän ongelmaan liittyy vastemuuttujien luonne. Julkaisuun sitoutumisen osalta voidaan ajatella olevan Facebookin käyttäjiin liittyviä inhimillisiä piirteitä, joita ei pystytä ottamaan huomioon aineiston keruussa, eikä edes välttämättä havaitsemattomina taustamuuttujina kausaalisuhteita määritettäessä. Facebookilla on palvelun tarjoavana yrityksenä

mahdollisuus vaikuttaa alustan toiminta-algoritmeihin, jonka voi olettaa vaikuttavan julkaisujen näkyvyyteen ja sitä kautta myös julkaisun tavoittavuuteen. Oletettavasti malleista puutuu siten kovariaatteja, jotka voisivat läsnäolollaan poistaa ylihajontaa.

Myös interaktiotermien sisällyttämistä malleihin tutkittiin. Mallin regressiokertoimien lukumäärän suuresta lisääntymisestä ja interaktioiden tulkinnallisista vaikeuksista johtuen pitäydyttiin yksinkertaisemmissa malleissa, vaikka ristiinvalidoinnin perusteella osa interaktioista vaikutti parantavan lievästi mallin sopivuutta. Jäännösten avulla mallien sopivuutta tarkasteltaessa havaittiin tavoittavuusmallissa suurempien arvojen osalta heikommat sovitteet. Vastemuuttujien luonnostaan saamat suuret havaintoarvot aiheuttavat osaltaan epäsopivuutta malliin, mikä näkyy myös jäännöksissä. Näitä suuria havaintoarvoja ei voida kuitenkaan perustellusti pitää poikkeavina havaintoina ja siten mallinnuksesta poistettavina.

Usein halutaan tarkastella muuttujia niiden keskimääräisen käyttäytymisen osalta, johon tässä tutkielmassa on käytetty keskimääräistä kausaalivaikutusta. Estimoitujen kausaalivaikutusten osalta huomataan kuitenkin jakauman vinous tarkasteltaessa keskimääräisten kausaalivaikutusten ja kvantiilivälien suhdetta (Kuva 5 ja 7). Kausaalivaikutusten keskimääräisen käyttäytymisen kuvaamiseen voitaisiin käyttää myös robustimpaa vaihtoehtoa, kuten mediaanikausaalivaikutusta (Rubin, 1974).

Julkaisun tavoittavuuden mallinnuksessa otettiin huomioon tykkäykset, kommentoinnit ja jaot CLS-muuttujan muodossa. Nämä sitoutumisen muodot voitaisiin ottaa tarkasteluun myös erikseen omina muuttujinaan, suhteuttamalla kyseinen sitoutumisen muoto julkaisun tavoittavuuteen. Tavoittavuuden mallinnuksessa olisi tällöin mahdollista ottaa huomioon näiden kolmen eri sitoutumisen muodot erikseen ja tutkia saatujen regressiokertoimien estimaattien eroavaisuuksia. Niin ikään julkaisuun sitoutuneisuutta voitaisiin mallintaa erikseen julkaisun tykkäysten, jakojen ja kommentointien osalta. Kiinnostavaa olisi myös selvittää julkaisun ostetulle näkyvyydelle kohdistetun valinnan vaikutusta julkaisun tavoittavuuteen. Tämä edellyttäisi kuitenkin ostetun näkyvyyden muuttujan havaitsemista jatkuvana, jotta vaikutuksen selvittäminen olisi mielekästä.

Lähteet

Bengio, Y. & Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research*, 5(Sep):1089–1105.

Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695. http://igraph.org

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, New York, 2nd edition.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.158.8831&rep=rep1&type=pdf

Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press, Cambridge; New York, 2nd edition.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11): 3735–3745.

Lee, D., Hosanagar, K. & Nair, H. S. (2018). Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Management Science*, 64(11):5105–5131. https://doi.org/10.1287/mnsc.2017.2902

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd edition.

Osuuskauppa Hämeenmaa (2017). Facebook-julkaisu 12.9.2017.
https://www.facebook.com/hameenmaa/posts/1244490762322160. Viitattu 18.3.2020.

Osuuskauppa Hämeenmaa (2019a) Facebook-julkaisu 11.8.2019.
https://www.facebook.com/hameenmaa/posts/2079420838829144. Viitattu 18.3.2020.

Osuuskauppa Hämeenmaa (2019b) Facebook-julkaisu 14.8.2019.
https://www.facebook.com/hameenmaa/posts/2089104444527450. Viitattu 18.3.2020.

Osuuskauppa Hämeenmaa (2019c) Facebook-julkaisu 20.5.2019.
https://www.facebook.com/hameenmaa/posts/1944359575668605. Viitattu 18.3.2020.

Osuuskauppa Hämeenmaa (2020). Perustietoa Hämeenmaasta.
https://www.s-kanava.fi/web/hameenmaa/perustietoa-hameenmaasta. Viitattu 28.1.2020.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition.

Pearl, J., Glymour, M. & Jewell, N. P. (2016) *Causal Inference in Statistics : A Primer*. Wiley, West Sussex, England.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701.

Shalizi, C. R. (2019). Advanced Data Analysis from an Elementary Point of View.
`http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf`
Viitattu 12.5.2020.

Shpitser, I. & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 1219–1226.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

S-ryhmä (2019). Tietoa S-ryhmästä. https://s-ryhma.fi/tietoa-meista/tietoa-s-ryhmasta Luettu 28.1.2020.

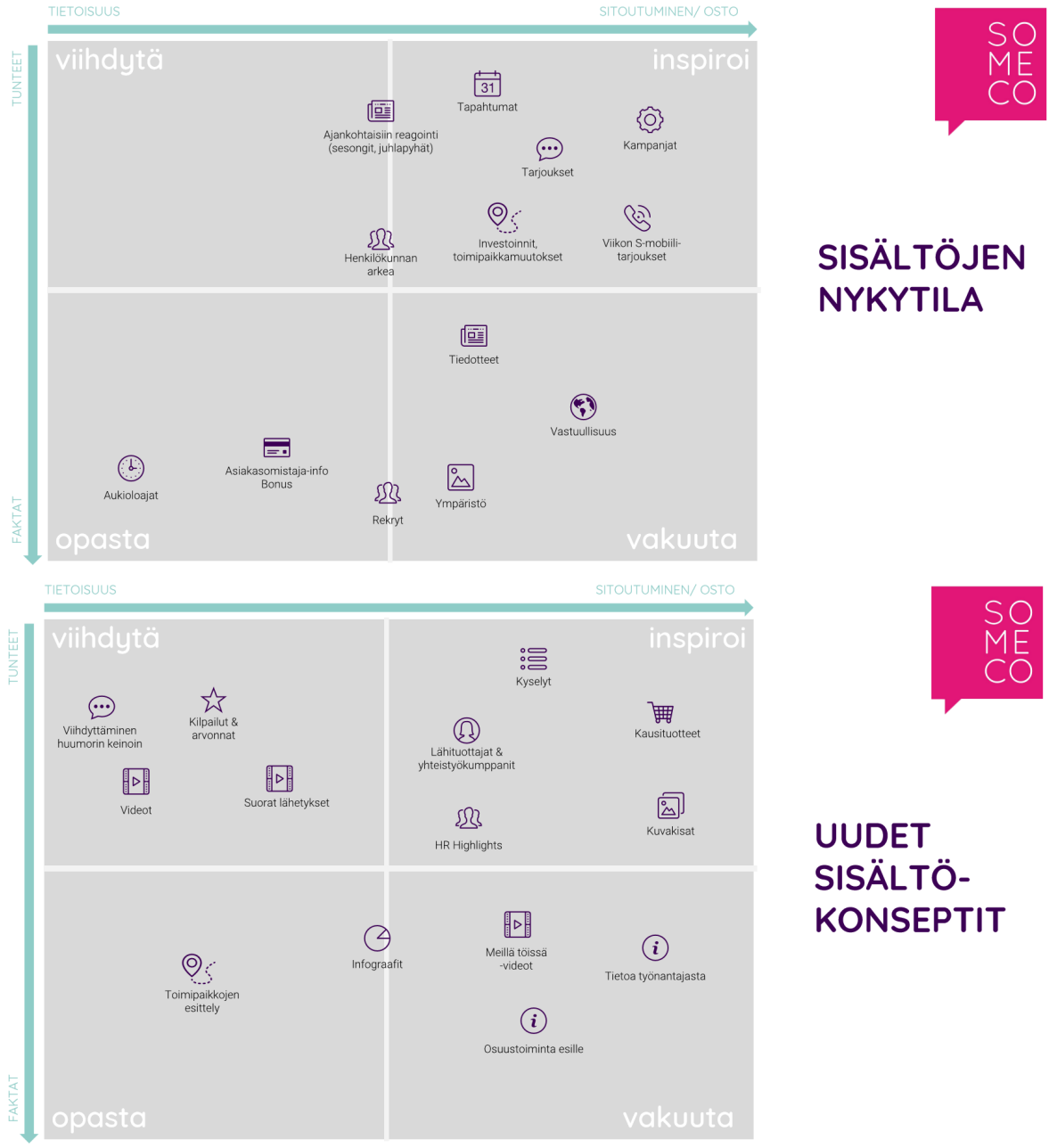
Tikka, S. & Karvanen, J. (2017). Identifying Causal Effects with the R Package causaleffect. *Journal of Statistical Software*, 76(12):1–30.

Valtari, M. & Inkinen, W. (2018). Osuuskauppa Hämeenmaan sosiaalisen median strategiset suuntaviivat. (Julkaisematon raportti).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, 2nd edition.

Liitteet



Liite 1: Sisältöjen nelikenttäjaottelu (Valtari & Inkinen, 2018).

Julkaisun sitoutuneisuusasteen malli
Splini: pituus (k=4)
CV-suure = 0.00075
AIC = 90315.77

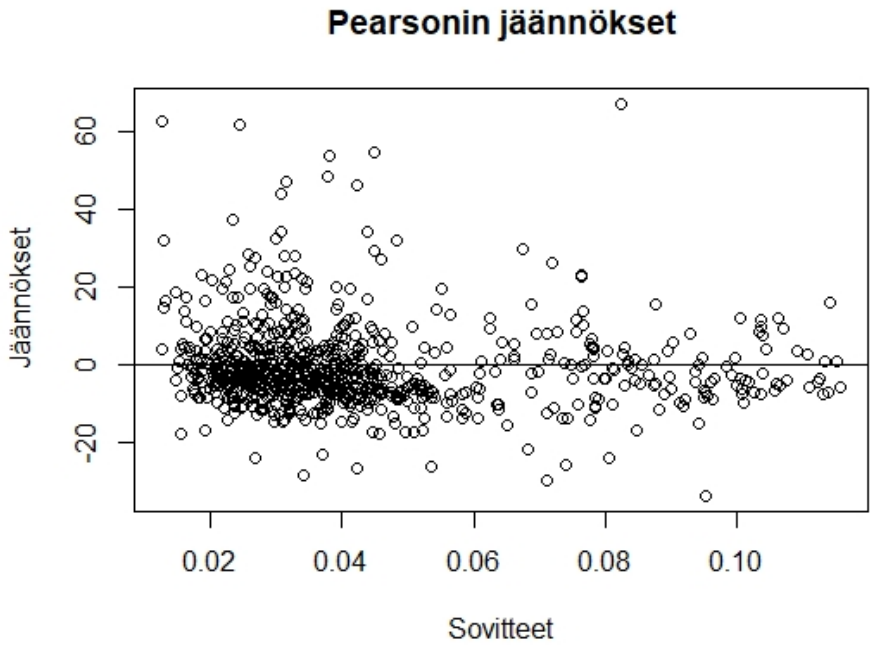
kovariaatti	$\widehat{\beta}$	$se(\widehat{\beta})$	z-arvo	p-arvo
(vakio)	-3.34978	0.0133207	-251.473	< 0.001
Vuodenaika				
kesä	-0.09704	0.0056144	-17.284	< 0.001
syksy	0.11889	0.0059985	19.820	< 0.001
talvi	0.07333	0.0050757	14.448	< 0.001
Viikonpäivä				
tiistai	-0.06972	0.0053833	-12.951	< 0.001
keskiviikko	0.06065	0.0049539	12.243	< 0.001
torstai	0.07472	0.0052277	14.294	< 0.001
perjantai	-0.01675	0.0071495	-2.343	0.019
lauantai	-0.11817	0.0098792	-11.962	< 0.001
sunnuntai	-0.19927	0.0088585	-22.494	< 0.001
Vuorokaudenaika				
klo 9-12	-0.10538	0.0042601	-24.736	< 0.001
klo 12-15	-0.07768	0.0052345	-14.841	< 0.001
klo 15 jälkeen	0.03099	0.0059157	5.238	< 0.001
Sisältötyyppi				
opastava	-0.01471	0.0067044	-2.194	0.028
vakuuttava	0.40099	0.0071863	55.799	< 0.001
viihdyttävä	0.66267	0.0057498	115.251	< 0.001
Mediatyyppi				
kuva	0.14810	0.0109012	13.586	< 0.001
videot	-0.34630	0.0132096	-26.215	< 0.001
Ostettu näkyvyyttä				
kyllä	0.09213	0.0045577	20.214	< 0.001
kumulatiivinen aika	-0.00002	0.0000005	-44.561	< 0.001
aikaero	-0.00036	0.0000910	-3.913	< 0.001

Liite 2: Julkaisun sitoutuneisuusasteen malli, sekä siihen liittyvä splini-termi, CV-suureen arvo ja AIC-arvo. Kovariaattitaulukossa estimoidut regressiokertoimet $\widehat{\beta}$, niiden keskivirheet $se(\widehat{\beta})$, sekä regressiokertoimien testisuureet z ja p-arvot.

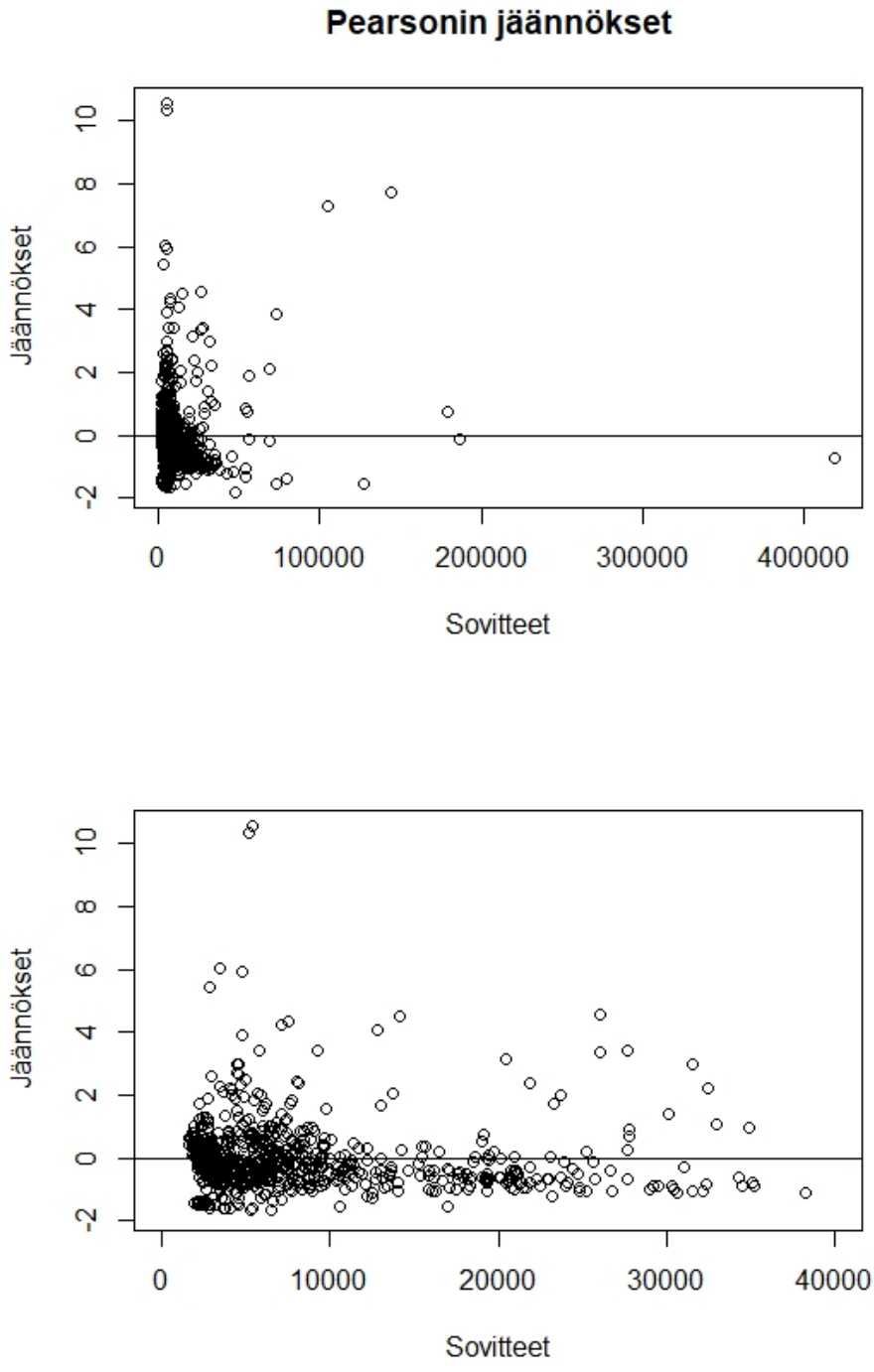
Julkaisun tavoittavuuden malli
ylihajontaparametri $\alpha = 0.29$
Splini: kumulatiivinen aika (k=5)
CV-suure = 2460423127
AIC = 14988.76

kovariaatti	$\widehat{\beta}$	$se(\widehat{\beta})$	z-arvo	p-arvo
(vakio)	8.4243	0.1031	81.725	< 0.001
Vuodenaika				
kesä	-0.0718	0.0786	-0.913	0.361
syksy	-0.2315	0.0748	-3.094	0.002
talvi	-0.1006	0.0640	-1.571	0.116
Viikonpäivä				
tiistai	-0.0565	0.0650	-0.868	0.385
keskiviikko	-0.1379	0.0674	-2.047	0.041
torstai	-0.1667	0.0693	-2.405	0.016
perjantai	-0.1371	0.0680	-2.017	0.044
lauantai	-0.1925	0.1004	-1.918	0.055
sunnuntai	-0.1398	0.0819	-1.707	0.088
Vuorokaudenaika				
klo 9-12	-0.1245	0.0598	-2.080	0.038
klo 12-15	-0.2306	0.0611	-3.775	< 0.001
klo 15 jälkeen	-0.1779	0.0666	-2.673	0.008
Sisältötyyppi				
opastava	0.1104	0.0554	1.992	0.046
vakuuttava	0.1735	0.0578	3.000	0.003
viihdyttävä	0.2190	0.0689	3.179	0.002
Mediatyyppi				
kuva	0.0392	0.0750	0.523	0.601
videot	0.4445	0.1049	4.238	< 0.001
Ostettu näkyvyyttä				
kyllä	0.6674	0.0534	12.494	< 0.001
aikaero	-0.0007	0.0009	-0.804	0.421
CLS-luku	31.4171	1.5382	20.424	< 0.001

Liite 3: Julkaisun tavoittavuuden malli, sekä siihen liittyvä splini-termi, CV-suureen arvo ja AIC-arvo. Kovariaattitaulukossa estimoidut regressiokertoimet $\widehat{\beta}$, niiden keskivirheet $se(\widehat{\beta})$, sekä regressiokertoimien testisuureet z ja p-arvot.



Liite 4: Sitoutuneisuusasteen mallin Pearsonin jäännösten ja mallin ennusteiden välinen hajontakuvio



Liite 5: Tavoittavuusmallin Pearsonin jäännösten ja mallin ennusteiden välinen hajontakuvio; alempana vastaava kuvio skaalattuna x-akselin suhteen.