PRO GRADU -TUTKIELMA

Anne-Mari Asmala

Pienhiukkasaineiston analysointi yleistetyillä additiivisilla malleilla

TAMPEREEN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Tilastotiede

Joulukuu 2009

Tampereen yliopisto

Matematiikan ja tilastotieteen laitos

ASMALA, ANNE-MARI: Pienhiukkasaineiston analysointi yleistetyillä additiivisilla malleilla

Pro gradu -tutkielma, 38 s., 3 liites.

Tilastotiede

Joulukuu 2009

Tiivistelmä

Ilman epäpuhtaudet ja pienhiukkaset ovat maailmanlaajuinen ympäristöterveysongelma. Suomessa ilmansaasteiden pitoisuudet ovat melko alhaisia, mutta herkkyys niille vaihtelee yksilöittäin ja jo hyvin pienillä pitoisuuksilla näyttäisi olevan vaikutusta terveyteen.

Tutkielman tavoitteena on esitellä yleistettyjä additiivisia malleja sekä soveltaa niitä pienhiukkasaineistoon. Aineisto on Tampereen kaupungin ympäristönsuojeluyksikön keräämää pienhiukkasaineisto, joka ulottuu vuoden 2006 alusta vuoden 2008 heinäkuulle. Yleistetyillä additiivisilla malleilla mallinnetaan pienhiukkasten vaikutusta tamperelaisten terveyteen. Vasteena käytetään tehtyjä hengityselin- ja sydändiagnooseja sekä kuolemien lukumäärää. Vasteen jakaumaoletuksena on Poissonin jakauma ja linkkifunktiona vasteen ja selittävien termien välillä käytetään log-linkkiä. Osaan muuttujista käytetään tasoittavaa funktiota. Pienhiukkasten sekä muiden hiukkasmuuttujien vaikutusta diagnoosien määrään tutkitaan erilaisilla summaviiveillä, jotka ottavat huomioon mahdollisen useamman päivän kestäneen ilmansaasteiden pitoisuuden nousun vaikutuksen terveyskäynteihin. Mallia kehitetään edelleen tutkimalla muuttujien välisiä interaktioita eli yhdysvaikutuksia ja niiden vaikutusta vasteeseen.

Pienhiukkasten ja hengityselinoireiden sekä sydänoireiden välillä havaitaan selvä yhteys. Hiukkaspitoisuuden ilmassa kasvaessa tehtyjen diagnoosien määrä kasvaa. Vaikutukset näkyvät yleensä yhdestä kolmeen päivän summaviiveellä. Lukumääräpitoisuusmuuttujat antavat selitysasteen ja merkitsevyyksien mielessä parempia tuloksia kuin eri tavalla mitatut massapitoisuusmuuttujat, varsinkin hengityselinoireiden tapauksessa. Lukumäärä- ja massapitoisuusmuuttujien tai typen oksidien välillä ei havaita interaktiota. Eri hiukkasmuuttujien samanaikaisten pitoisuuksien välillä ei ole yhdysvaikutusta, joka olisi enemmän kuin muuttujien erillisten vaikutusten summa. Sen sijaan taustamuuttujien, varsinkin ajoneuvojen lukumäärän, ajan, lämpötilan sekä lämpötila eron välillä on interaktiota.

Asiasanat Poissonin regressio, splini

Sisältö

1 Johdanto

1.1 Johdatus pienhiukkasiin

Ilman epäpuhtaudet ja pienhiukkaset ovat maailmanlaajuinen ympäristöterveysongelma. Länsimaissa yhdyskuntailman pienhiukkasia pidetäänkin ihmisen terveydelle haitallisimpina ympäristötekijöinä (Salonen & Pennanen 2006). Euroopan unionin maat vaikuttavat lainsäädännöllään ilmanlaatuun yrittäen parantaa sitä ja pienentää pienhiukkaspitoisuuksia. Euroopan komissiossa muun muassa on tavoitteena vähentää ilman saastumisesta johtuvien kuolemien määrää 40 prosenttia vuoden 2000 tasosta vuoteen 2020 mennessä. (Euroopan komissio 2005; Pekkanen 2004).

Euroopassa on tutkittu viiden suurkaupungin ulkoilman partikkeleiden määriä ja vertailtu niitä. Tutkimuksessa todettiin, että Helsingin päivittäiset hiukkasten maksimiarvot ovat jopa pienempiä kuin etelä-Euroopan kaupunkien päivittäiset minimipitoisuudet. (Aalto et al. 2005). Vaikka Suomessa ilmansaasteiden pitoisuudet ovat melko alhaisia ja siten terveyshaitat useimmille pieniä, vaihtelee herkkyys ilmansaasteille kuitenkin yksilöittäin (Niemi et al. 2008). Ilmanlaadulle ja hiukkasten pitoisuuksille on asetettu erilaisia raja-, tavoite- ja ohjearvoja, joista raja-arvot ovat tiukimpia ja perustuvat EU-direktiiveihin. Esimerkiksi hiukkasten vuorokauden raja-arvo on 50 $\mu g/m^3$, joka saa ylittyä 35 kertaa vuodessa. (Elsilä 2006; Niemi et al. 2008.) Raja-arvoja pienemmät hiukkaspitoisuudet eivät kuitenkaan takaa hyvää ilmanlaatua, koska terveydelle täysin turvallista pienhiukkasten pitoisuutta ei ole pystytty määrittämään. Jo hyvin pienillä pitoisuuksilla näyttäisi olevan vaikutusta terveyteen. Myöskään pienhiukkasten terveysvaikutuksia selittävää keskeistä ominaisuutta ei ole täysin pystytty osoittamaan. (Pekkanen 2004.)

Tampereen alueella ilmanlaadun tarkkailusta vastaa Tampereen kaupungin ympäristönsuojeluyksikkö. Tärkeimmät kaupunki-ilman laatua heikentävät epäpuhtaudet ovat hiukkaset, typen oksidit, otsoni, hiilimonoksidi, rikkidioksidi sekä orgaaniset yhdisteet. Niillä on varsinkin runsaina pitoisuuksina negatiivisia vaikutuksia terveyteen, luontoon sekä viihtyvyyteen. (Niemi et al. 2008.) Hiukkaspäästöt ovat vähentyneet Tampereella noin 80 prosenttia 1990-luvun puolivälistä, mikä johtuu lähinnä teollisuuspäästöjen vähentymisestä. Suurimpien hiukkasten pitoisuudet sen sijaan eivät juuri ole muuttuneet 1990-luvusta. Ne ovat suurelta osin tienpinnasta autojen irroittamaa ainesta, joka ei ole vähentynyt liikennemäärien kasvaessa. Liikenteen ja teollisuuden lisäksi kaupungissa ilman epäpuhtauksien päästölähteitä ovat mm. energiantuotanto, pien-

poltto sekä kaukokulkeuma. (Niemi et al. 2008.)

Hiukkasilla on aiemmissa tutkimuksissa todettu olevan tiettyjä terveysvaikutuksia. Suurimmat hiukkaset aiheuttavat lähinnä hengityselinten tukkoisuutta sekä ympäristön likaantumista, kun taas pienet hiukkaset kulkeutuvat elimistössä pisimmälle ja ovat siten suurempia hiukkasia vaarallisempia. (Niemi et al. 2008; Salonen & Pennanen 2006.) Pienhiukkasten aiheuttamat terveysvaikutukset vaikuttavat paitsi yksilön terveyteen ja elämänlaatuun niin myös kansanterveyteen ja -talouteen erityisesti terveydenhuollon palvelujen kysynnän sekä työpoissaolojen muodossa. (Salonen & Pennanen 2006.)

1.2 Tutkielman tavoite

Tutkielman tavoitteena on esitellä yleistettyjä additiivisia malleja sekä soveltaa niitä pienhiukkasaineistoon. Aineistosta on tarkoitus tutkia tamperelaisten altistumista ulkoilman pienhiukkasille ja muille epäpuhtauksille erityisenä mielenkiinnonkohteena pienhiukkasten aiheuttamat terveysvaikutukset. Analysointiin käytetään yleistettyjä additiivisia malleja, jotka ovat joustavia kuvaamaan erilaisia monimutkaisia riippuvuuksia aineistossa. Tavanomaiset parametriset mallit eivät kykene kuvaamaan tämänkaltaisia riippuvuuksia yhtä joustavasti kuin additiiviset mallit. Yleistetyillä additiivisilla malleilla on tarkoitus mallintaa pienhiukkasten vaikutusta terveyteen. Vasteena ovat tehdyt hengityselin- ja sydändiagnoosit sekä kuolemien lukumäärät. Analyysimallin lähtökohtana käytetään jo tätä työtä edeltäneessä tutkimuksessa kehitettyä mallia, jonka soveltamista ja kehittämistä jatketaan.

Aiemmissa tutkimuksissa on myös havaittu, että ilmansaastepitoisuudet vaikuttavat diagnoosien määrään erilaisilla viiveillä. Ne otetaan malleissa huomioon summamuuttujina, joissa muuttujan viivästetyt arvot on laskettu yhteen. Mallia kehitetään edelleen tutkimalla muuttujien välisiä interaktioita eli yhdysvaikutuksia ja niiden vaikutusta. Käytännössä interaktio huomioidaan lisäämällä malliin interaktiotermi. Tutkimus liittyy vuonna 2008 alkaneeseen Tampereen kaupungin ympäristönsuojeluyksikön, Tampereen teknillisen yliopiston sekä Tampereen ammattikorkeakoulun yhteiseen ilman pienhiukkasten terveysvaikutuksia koskevaan tutkimushankkeeseen.

1.3 Tutkielman rakenne

Johdannon lisäksi tutkielma koostuu neljästä luvusta sekä lähdeluettelosta ja liitteistä. Ensin esitellään aineistoa ja menetelmiä, jonka jälkeen ne yhdistetään soveltamalla menetelmiä aineistoon ja lopuksi kokoamalla asiat yhteen yhteenvedossa. Aineistoa esitellään tarkemmin toisessa luvussa, jossa kerrotaan sen keräämisestä ja erilaisissa rooleissa olevista aineiston muuttujista. Muuttujista esitellään erikseen pienhiukkas- ja muita ilmanlaatumuuttujia, taustamuuttujia sekä vastemuuttujia. Kolmannessa luvussa käsitellään menetelmiä, päähuomion ollessa yleistetyissä additiivisissa malleissa. Muita malleja esitellään tuke-

na ja johdantona niille. Lisäksi käsitellään läheisesti yleistettyihin additiivisiin malleihin liittyviä Poissonin regressiota ja tasoitusmenetelmiä. Neljännessä luvussa yhdistetään kahden aiemman luvun asiat eli sovelletaan yleistettyjä additiivisia malleja aineistoon ja tutkitaan erilaisia malleja. Perusmallia yritään parantaa lisäämällä siihen erilaisia viivesummia ja interaktiotermejä. Tässä luvussa kerrotaan myös käytetystä tilasto-ohjelmasta. Viimeisessä viidennessä luvussa kootaan tutkielmasta yhteenveto. Lopussa on lähdeluettelon lisäksi kolme liitettä hiukkasmuuttujien merkitsevyyksistä, mallien selitysasteista sekä interaktiotermien merkitsevyyksistä ja vaikutuksesta selitysasteeseen.

2 Aineisto

Tutkimuksen aineistona käytetään Tampereen kaupungin ympäristönsuojeluyksikön keräämää pienhiukkasaineistoa. Aineisto ulottuu helmikuun alusta 2006 heinäkuun loppuun 2008. Havaintoyksikköjä eli vuorokausia tutkimuksessa käytetyssä vuorokausiaineistossa on 943. Aineisto koostuu useista muuttujista, joista käytetään tämän tutkimuksen kannalta tärkeimpiä. Seuraavassa esitellään tarkemmin aineiston keräämistä sekä käytettäviä muuttujia ryhmittäin.

2.1 Aineiston kerääminen

Aineiston pienhiukkastiedot on kerätty Tampereen keskustasta Pirkankadulta. Mittaamiseen käytettiin *Electrical Low Pressure Impactor* eli $ELPI$-hiukkasanalysaattoria. Laitteen toiminta perustuu hiukkasten saamaan sähköiseen varaukseen, joka jaottelee erikokoiset hiukkaset eri keräystasoille. $ELPI$:n mittaustuloksista pystytään laskemaan mm. hiukkasten lukumäärä- ja massapitoisuudet sekä aktiivisen pinta-alan. Samassa mittauspisteessä mitataan myös muiden ilmanlaatuun vaikuttavien seikkojen, hengitettävien hiukkasten ja typen oksidien, pitoisuuksia. Alkuperäinen aineisto, joka sisältää hiukkas- ja taustamuuttujat, on tuntiaineisto, josta on laskettu vuorokausittaiset mediaanit. Vuorokausittainen mediaani kertoo, mikä on ollut vuorokauden tuntien keskimmäisin arvo. Tarkempaa tietoa pienhiukkasaineiston keräämisestä ja $ELPI$-hiukkasanalysaattorista on saatavilla lähteissä Hilli-Lukkarinen (2009) ja Kariniemi (2006).

Säätiedot on saatu Ilmatieteen laitokselta ja ne on mitattu Näsinneulan sääasemalla. Liikennetiedot on saatu Tampereen kaupungin liikennesuunnitteluosastolta, joka on mitannut ajoneuvojen lukumäärät Pirkankadulla ajoradalla olevalla sensorilla. Terveysaineisto on koottu Tampereen kaupungin hyvinvointipalveluiden kehittämisyksiköstä saatujen tietojen mukaan ja se koostuu tutkimusaineiston päivinä Tampereella terveysasemilla ja päivystyksessä diagnosoiduista hengityselinoireista sekä sydänsairauksista. Tiedoista puuttuvat ainakin koulu- ja opiskelijaterveydenhuolto, yksityiset terveysasemat (työterveys) sekä keskussairaalan tiedot. Diagnoositietoja on saatu sekä ICPC- että ICD-10 -luokitusten mukaan, joista toisiaan vastaavat diagnoosit on yhdistetty. Tartuntatautirekisterin viikoittaiset lukumäärätiedot on haettu Kansanterveyslaitoksen, nykyisen Terveyden ja hyvinvoinnin laitoksen, tilastotietokannasta ja luku on yleistetty koskemaan koko viikkoa. Tampereen asukkaita koskevat kuolleisuustiedot on saatu Tilastokeskukselta. (Hilli-Lukkarinen 2009.)

2.2 Ilman pienhiukkaset ja hengitettävät hiukkaset

Ilmassa on jatkuvasti eri määriä, eri kokoisia ja eri ikäisiä hiukkasia, jotka ovat peräisin eri lähteistä. Myös hiukkasten kemiallinen koostumus vaihtelee paljon. Hiukkaset voidaan jakaa ryhmiin kokonsa perusteella. Halkaisijaltaan alle 10 μm kokoisia hiukkasia sanotaan hengitettäviksi hiukkasiksi, alle 2.5 μm pienhiukkasiksi ja alle 0.1 μm kokoisia hiukkasia ultrapieniksi hiukkasiksi. (Salonen & Pennanen 2006.) Hiukkasten pitoisuutta ilmassa voidaan mitata lukumäärän mukaan lukumääräpitoisuutena (esimerkiksi kpl/cm^3) tai massan mukaan massapitoisuutena (esimerkiksi mg/m^3).

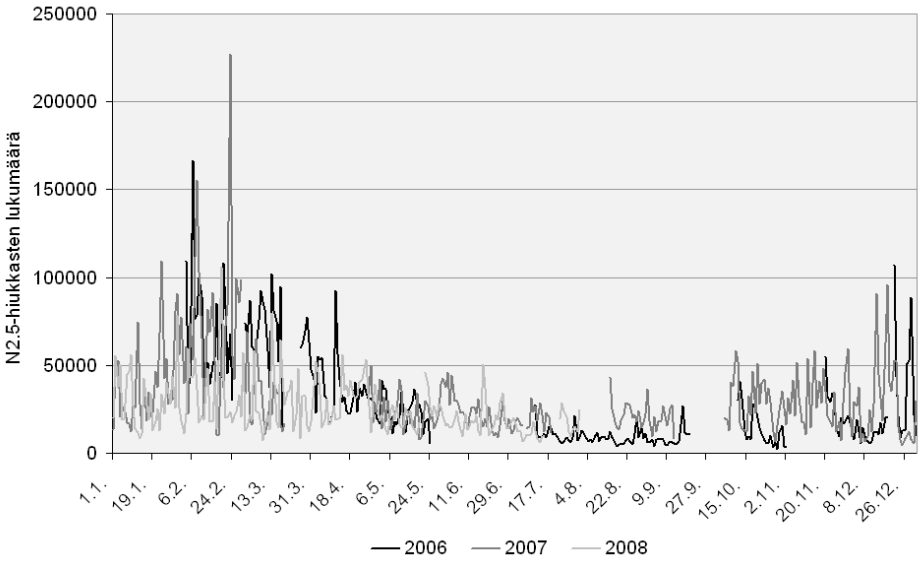
Hiukkasten lukumääräpitoisuuksia merkitään N:llä ja massapitoisuuksia PM:llä. Massapitoisuus PM_{10} eli hengitettävät hiukkaset koostuvat sekä pienhiukkasista ($PM_{2.5}$) että karkeammista hiukkasista ($PM_{10-2.5}$). Aineiston pienhiukkasmuuttujista käytetään neljää erilaista luokitusta. Kokoluokka $PM_{2.5}$ sisältää kaikkien pienhiukkasten eli halkaisijaltaan alle 2.5 μm olevien hiukkasten massapitoisuuden (mg/m^3). Myös $N_{2.5}$ sisältää kaikki vastaavan kokoiset hiukkaset, mutta nyt kyseessä on lukumääräpitoisuus (kpl/cm^3). $N_{0.1}$ on lukumääräpitoisuus, johon sisältyy kaikki ultrapienet hiukkaset. Kaikkein pienimpien hiukkasten ryhmä $N_{0.01}$ on vielä omana muuttujanaan. Näiden hiukkasten halkaisija on ainoastaan 0.01 μm eli 10 nm. Kaikilla pienhiukkasilla oli puuttuvia arvoja melko paljon, 225 vuorokautta. Puuttuvat tiedot johtuvat $ELPI$-hiukkasanalysaattorin mittausongelmista.

Suurin osa kaupunki-ilmassa olevista hengitettävistä hiukkasista on peräisin liikenteen ja tuulen nostattamasta katupölystä (Niemi et al. 2008). Nämä päästöt ovat arvioiden mukaan nykyään monta kertaa suurempia kuin ajoneuvojen pakokaasupäästöt. Karkeita hengitettäviä hiukkasia syntyy myös muun muassa teollisuuslaitoksissa, satamissa ja louhintatöissä materiaalien käsittelyn ja maansiirron seurauksena. (Salonen & Pennanen 2006.)

Ultrapieniä hiukkasia syntyy aina palamisessa. Kaupunkialueella niitä on runsaasti erityisesti pakokaasupäästöissä, varsinkin dieselmoottorisen auton tapauksessa. Puun poltossa pienhiukkaspitoisuuden pienentämisessä oleellista olisi puhdas palaminen. (Pekkanen 2004; Salonen & Pennanen 2006.) Osa pienhiukkaspitoisuudesta aiheutuu kaukokulkeumasta, joka tulee Suomeen varsinkin Venäjältä ja Baltian maista sekä Keski-Euroopasta, ja on usein peräisin maasto- tai metsäpaloista. (Pekkanen 2004)

Kuvassa 2.1 on $N_{2.5}$-muuttujan jakauma. Eri vuodet on piirretty päällekkäin. Nähdään, että suurimmat $N_{2.5}$-hiukkaspitoisuudet ovat kevät-talvella ja pienimmät kesällä ja alku-syksyllä. Vuoden 2008 leuto talvi näkyy matalina hiukkaspitoisuuksina verrattuna muihin vuosiin. Kuvasta erottuu myös ajanjaksot, joilta ei ole saatavilla tietoja.

Hiukkasten lukumääräpitoisuuksia ja massapitoisuuksia voidaan vertailla Pearsonin korrelaatiokertoimen avulla. Korrelaatiokerroin mittaa vain muuttujien välistä lineaarisen riippuvuuden voimakkuutta, eikä se takaa syy-seuraussuhdetta. Kerroin vaihtelee miinus yhdestä yhteen ja arvo nolla kertoo, ettei lineaarista riippuvuutta ole. (Heikkilä 2005.) Pienhiukkasten massapitoisuuden



Kuvio 2.1. $N_{2.5}$-hiukkasten lukumäärät vuorokausimediaaneina

ja lukumääräpitoisuuden välinen korrelaatio on melko heikko (0.27–0.29). Samansuuntaiseen tulokseen on päädytty myös tutkimuksessa (Ruuskanen et al. 2001), jossa ultrapienten- ja pienhiukkasten lukumääräpitoisuuksia verrattiin $PM_{2.5}$:een. Lukumäärä- ja massapitoisuuksia on tarpeen tarkastella erikseen.

2.3 Muut ilmanlaatumuuttujat

Muita tässä tutkimuksessa tutkittuja ilmanlaatumuuttujia ovat typen oksidit. Typen oksidit ovat ulkoilmassa suurelta osin peräisin liikenteen päästöistä. Niiden kohdalla raja-arvot ylittyvät kaupungeissa lähinnä tyyninä pakkaspäivinä. (Niemi et al. 2008.) Typen oksideista on tutkittu typpimonoksidia (NO) sekä typpidioksidia (NO_2). Niiden pitoisuudet on annettu massapitoisuuksina ja yksikkönä on $\mu g/m^3$.

Taulukossa 2.1 tarkastellaan tarkemmin pienhiukkasia sekä muita hiukkasia ja niiden tunnuslukuja kahden merkitsevän numeron tarkkuudella. Min tarkoittaa kyseisen muuttujan pienintä arvoa koko aineistossa ja Max vastaavasti suurinta. Md kertoo muuttujan mediaanin ja Ka keskiarvon. Keskihajontaa merkitään lyhenteellä Sd. Lukumääräpitoisuuksien maksimimäärät ovat noin satakertaisia verrattuna pienimpiin pitoisuuksiin. Hiukkasmuuttujien vuorokauden tuntien mediaaneista lasketut keskiarvot ovat säännöllisesti suurempia kuin vastaavat mediaanit. Yksittäiset suuret arvot vetävät keskiarvoja ylöspäin ja saavat jakaumat vinoiksi oikealle.

Pienhiukkasaineistosta on laskettu myös uusi muuttuja SA, joka kertoo hiukkasten aktiivisen pinta-alan, eli alueen, joka voi olla kosketuksissa esimerkiksi keuhkokudokseen. SA perustuu hiukkasen pinta-alaan ja sen sähkövirtaan.

Taulukko 2.1. Hiukkaset

Hiukkasmuuttuja	Min	Md	Ka	Max	Sd	Yksikkö
$PM_{2.5}$	0.0013	0.015	0.018	0.11	0.015	mg/m^3
$N_{2.5}$	2300	22000	29000	230000	23000	kpl/cm^3
$N_{0.1}$	2100	20000	27000	220000	23000	kpl/cm^3
$N_{0.01}$	1800	19000	25000	210000	21000	kpl/cm^3
NO	1.0	7.0	9.5	62	7.4	$\mu g/m^3$
NO_2	1.5	20	21	96	12	$\mu g/m^3$
PM_{10}	3.0	12	16	180	14	$\mu g/m^3$

Teollisuudesta ja energiantuotannosta tulevia rikkidioksidipäästöjä ei Tampereella enää vuoden 2003 jälkeen ole mitattu. Rikkidioksidipäästöt ovat niin pieniä, ettei niillä ole todettu olevan merkittäviä terveysvaikutuksia. (Elsilä 2006.)

2.4 Mallintamisessa käytettävät taustamuuttujat

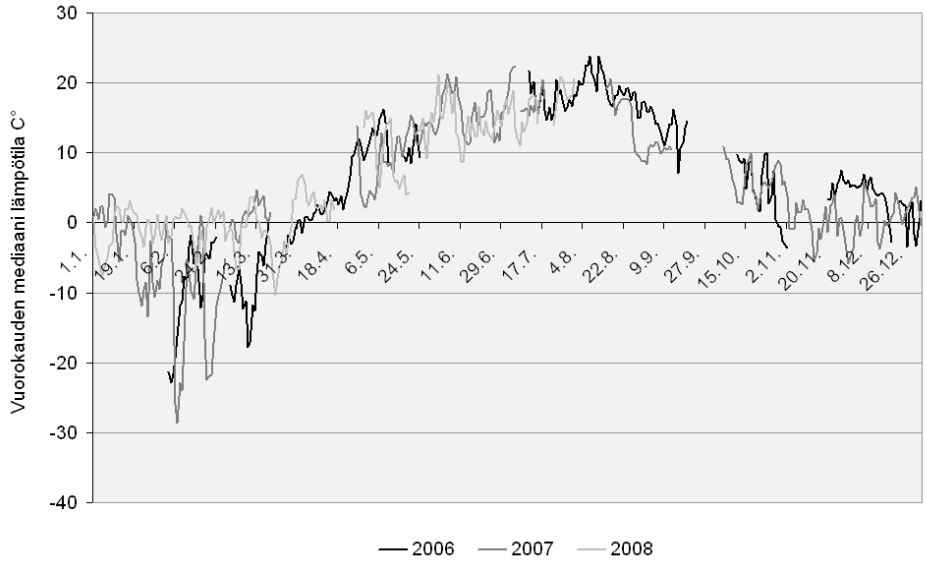
Tutkimuksessa käytetään taustamuuttujina lähinnä säätietoja: lämpötilaa, lämpötilaeroa ja ilmankosteutta. Muita taustamuuttujia ovat ajoneuvojen lukumäärä, aika sekä influenssaluokat. Kaikki arvot, aikaa ja influenssaluokkia lukuunottamatta, ovat vuorokausittaisia tunneista laskettuja mediaaneja.

Ilman lämpötila *temp*5 on mitattu 5 metrin korkeudelta maan pinnasta ja sen mittayksikkö on celsiusaste. Kun *temp*5 on vähennetty 135 metrin korkeudelta mitatusta lämpötilasta saadaan lämpötila ero *tempero*, joka kertoo mahdollisesta inversiotilanteesta. Inversiotilanteessa lämpötila on ylempänä lämpimämpää kuin lähempänä maanpintaa, toisin kuin tavallisesti. Inversiossa ilma ei sekoitu ja epäpuhtaudet jäävät lähelle maanpintaa. Ilmankosteus-muuttuja $RH5$ kertoo ilman suhteellisesta kosteudesta 5 metrin korkeudella ja sen yksikkönä on prosentti. Lämpötila ja ilmankosteus tiedot puuttuvat 243 vuorokaudelta. Ajoneuvojen lukumäärä $ajolkm$ on laskettu tuntikohtaisista lukumääristä muodostamalla vuorokauden mediaani. Tiedot puuttuvat 233 vuorokaudelta. *Aika*-muuttujan yksikkö on vuorokausi ja se käsittää kaikki aineiston 943 vuorokautta.

Kuvassa 2.2 on esitetty lämpötila-muuttujan jakauma. Eri vuosien vuorokausittaiset lämpötilojen mediaanit eroavat eniten toisistaan talvien kohdalla. Erityisesti erottuu leuto talvi 2008, jolloin kylmin vuorokausi koettiin vasta maaliskuun lopulla. Lämpimin vuorokausi on tutkimusajanjaksolla ollut elokuun alussa vuonna 2006.

Taulukossa 2.2 on kerrottu taustamuuttujien perustietoja kahden merkitsevän numeron tarkkuudella. Vuorokauden tuntien lämpötilojen mediaani on suurimmillaan ollut 24 astetta ja pienimmillään 29 pakkasastetta. $Temp5$-

muuttujan kohdalla keskiarvo on mediaania suurempi, mutta muiden taustamuuttujien tapauksessa tilanne on toisinpäin.



Kuvio 2.2. Lämpötila *temp*5 vuorokausimediaaneina

Taulukko 2.2. Taustamuuttujat

Taustamuuttuja	Min	Md	Ka	Max	Sd	Yksikkö
$Temp5$	-29	4.4	5.6	24	9.4	°C
$Tempero$	-12	0.90	0.68	2.1	1.1	°C
$RH5$	26	86	80	100	7.1	%
$Ajolkm$	140	610	550	810	140	kpl/h

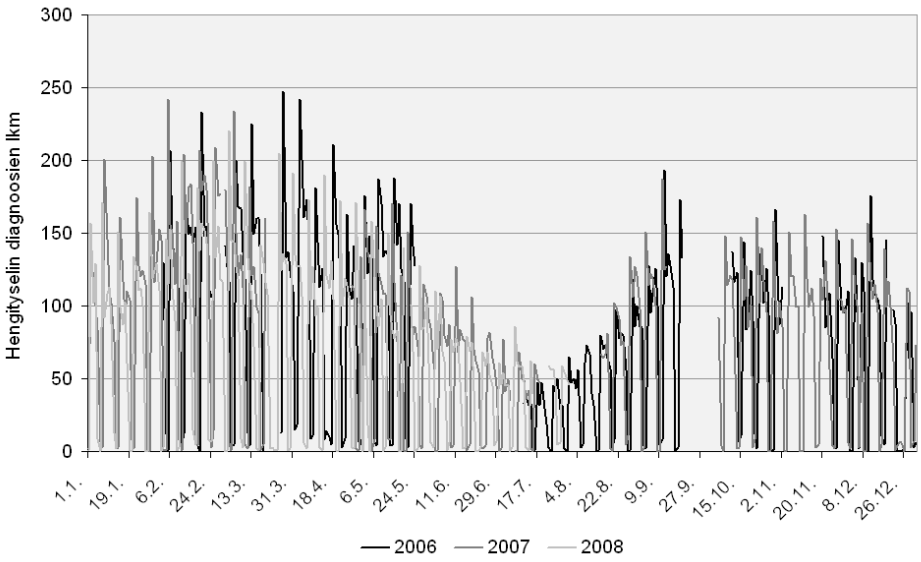
Terveysaineistoon kuuluvaa influenssaluokat-muuttujaa $influok$ käytetään taustamuuttujana. Se kertoo hengitystievirusten määrän luokiteltuna kolmeen luokkaan. Influenssaluokka 0 tarkoittaa alle 35 influenssatapausta päivässä, 1 tarkoittaa 35–180 tapausta ja 2 yli 180 tapausta päivässä. Alkuperäinen luku on ollut viikkokohtainen, joka on yleistetty koskemaan kaikkia viikon päiviä. (Hilli-Lukkarinen 2009.) Influenssaluokkien mediaani on nolla eli alle 35 tapausta päivässä. Päiviä, jolloin influenssatapauksia oli 35–180, oli vajaa 200 eli 27 prosenttia päivistä, joilta on influenssatapaustiedot ja yli 180 tapauksen päiviä noin 160 eli vajaa 23 prosenttia. Malliin tulleet taustamuuttujat valittiin ilman tilastollisia menetelmiä aiempien tutkimusten sekä asiantuntijoiden ehdotusten ja mielenkiinnon mukaan.

Tarkastellaan hieman muuttujan välisiä riippuvuuksia Pearsonin korrelaatiokertoimen avulla. Taustamuuttujista ainoastaan ajoneuvojen lukumäärän korrelaatio terveysmuuttujien kanssa ylittää 0.5:n. Lämpötilamuuttujat kor-

reloivat muiden, varsinkin influenssamäärien, pienhiukkasten lukumääräpitoisuuksien sekä ilmankosteuden kanssa, mutta korrelaatiokerroin on negatiivinen.

2.5 Vastemuuttujina käytettävät terveysmuuttujat

Tutkimuksessa vasteena käytettävät terveysaineiston muuttujat ovat kuolemien lukumäärä, hengityselinoireet sekä sydänoireet. Kuolemien lukumäärä *kuolkm* kertoo, kuinka monta tamperelaista on kuollut vuorokaudessa. Hengityselinoireet eli *keuh*-alkuiset muuttujat sisältävät diagnosoidut hengityselinoireet tautiluokitusten mukaan. Diagnoosit on luokiteltu neljään ikäluokkaan sekä laskettu yhteen *keuhkai*-muuttujaksi. Ikäryhmiksi on valittu alle 15-vuotiaat, 15–64-vuotiaat, 65–74-vuotiaat sekä vanhemmat eli yli 74-vuotiaat. Vastaavasti sydänoireet on luokiteltu *syd*-alkuisiin muuttujiin, joista *sydkai* sisältää kaikki tehdyt sydändiagnoosit päivittäin. Alle 15-vuotiaiden sydändiagnooseja on hyvin vähän, joten ne on yhdistetty seuraavaan ikäluokkaan. Yhdistetty muuttuja on siis kaikki alle 65-vuotiaat eli *syd*65.



Kuvio 2.3. Kaikkien hengityselindiagnoosien (*keuhkai*) lukumäärät

Kuvassa 2.3 on esitetty hengityselinoireiden jakauma kaiken ikäisillä. Kuvasta nähdään, että kesällä on selvä notkahdus diagnoosien määrässä ja suurimmat määrät ovat helmi-huhtikuussa, jolloin diagnooseja saatetaan tehdä yli 200 vuorokaudessa. Jakauman suuri vaihtelevuus johtuu viikonlopuista, jolloin diagnooseja tehdään paljon vähemmän kuin viikolla tai ei lainkaan. Eri vuosissa ei ole niin paljon eroavaisuutta kuin hiukkasmuuttujan ja lämpötilamuuttujan kohdalla oli. Kuitenkin vuoden 2008 talvella tehtiin jonkin verran vähemmän hengityselindiagnooseja kuin kahtena aiempana vuotena.

Taulukossa 2.3 on esitetty kuolemien lukumäärä -muuttujan sekä keuhko- ja sydändiagnoosi -muuttujien tietoja korkeintaan kahden merkitsevän nume-

ron tarkkuudella. Kaikilla terveysmuuttujilla vuorokauden pienin arvo on nolla. Kaikkiaan keuhkodiagnooseja on tehty keskimäärin noin 80 vuorokaudessa ja sydändiagnoodeja noin 30 maksimien ollessa yli kolme kertaa suurempia. Tamperelaisten keskimääräinen kuolleisuus vuorokaudessa on neljä tai viisi maksimin ollessa noin kolminkertainen määrä. Kaikilta terveysaineiston muuttujilta on poistettu tiedot 225 päivältä, koska kyseisiltä päiviltä ei ole saatu pienhiukkastietoja.

Taulukko 2.3. Terveystiedot

Vastemuuttuja	Min	Md	Ka	Max	Sd	Yksikkö
Kuolkm	0	4.5	4.8	14	2.3	*kpl*
*Keuh*15	0	14	14	54	12	*kpl*
*Keuh*1565	0	56	53	170	42	*kpl*
*Keuh*6574	0	7	7.0	32	6.3	*kpl*
*Keuh*74	0	5	5.3	20	4.7	*kpl*
Keuhkai	0	87	79	250	62	*kpl*
*Syd*65	0	9	9.1	32	7.8	*kpl*
*Syd*6574	0	9	9.4	40	8.4	*kpl*
*Syd*74	0	13	12	42	10	*kpl*
Sydkai	0	36	30	110	25	*kpl*

3 Menetelmät

Aineiston mallintamiseen käytetään *yleistettyjä additiivisia malleja (generalized additive models, GAM*), jotka perustuvat epäparametriseen regressioon ja tasoitusmenetelmiin. Nämä mallit saadaan yleistettyjen lineaaristen mallien yleistyksenä siten, että lineaarinen prediktori korvataan prediktorilla, joka muodostuu tasoitettujen termien summasta. (Wood 2006.) Yleistetyt additiiviset mallit tarjoavat joustavan keinon käsitellä epälineaaristen kovariaattien vaikutuksia ja mallintaa monimutkaisiakin riippuvuuksia. (Hastie & Tibshirani 1990; Wood 2006.) GAM-mallit ovat tulleet suosituiksi menetelmiksi varsinkin ympäristötieteissä ja -epidemiologiassa, ekologiassa, kansanterveystieteissä, poliittisissa tieteissä sekä ekonometriassa (French & Wand 2004).

Tässä luvussa esitellään ensin lineaarisia ja yleistettyjä lineaarisia malleja sekä additiivisia malleja. Niiden teoria antaa pohjan yleistetyille additiivisille malleille, joita käsitellään alaluvussa 3.4. Lisäksi tarkastellaan GAM-mallintamisessa erikoistapauksena saatavaa Poissonin regressiota ja interaktion eli yhdysvaikutuksen käsitettä yleistettyjen additiivisten mallien tapauksessa. Tärkeä osa additiivisia malleja ovat erilaiset tasoitusmenetelmät, joista käsitellään viimeisessä alaluvussa tarkemmin regressiospliniä, tasoittavaa spliniä sekä lokaaleja polynomeja, joista erityisesti Kernel-tasoitusta.

3.1 Lineaariset mallit

Lineaarisissa malleissa oletetaan, että satunnaismuuttujien odotusarvot $E(Y_i) = \mu_i$ riippuvat selittävistä muuttujista x_i lineaarisesti. Muuttujien välisen lineaarisen yhteyden kuvaamiseen sopii hyvin pienimmän neliösumman regressiosuora. Suora muodostetaan minimoimalla havaintopisteiden suorasta laskettujen etäisyyksien neliöiden summa. (Heikkilä 2005.) Yksinkertainen lineaarinen malli on muotoa

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

missä $i = 1,...,n$, y_i on selitettävä eli vastemuuttuja, β_0 on vakiotermi, x_i on selittävä muuttuja ja β_1 on x_i:n kerroin. Odotusarvo on $\mu_i = \beta_0 + \beta_1 x_i$. Jäännöstermi ϵ_i eli residuaali kuvaa mallin satunnaisvaihtelua, eli sitä osaa y:stä, jota mallin antama arvio ei pysty ennustamaan. Toisin sanoen $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$. Jäännöstermit ovat riippumattomia satunnaismuuttujia, joiden odotusarvo on nolla, $E(\epsilon_i) = 0$, ja varianssi σ^2 on vakio. (Wood 2006.) Lisäksi klassisessa lineaarisessa mallissa oletetaan, että jäännöstermit ovat riippumattomia ja noudattavat normaalijakaumaa $\epsilon_i \sim N(0, \sigma^2)$. Koska jäännöstermit noudattavat

normaalijakaumaa, myös y_i:t noudattavat normaalijakaumaa $N(\beta_0 + \beta_1 x_i, \sigma^2)$. (Wood 2006.)

Mallien hyvyyttä voidaan arvioida esimerkiksi selitysasteen R^2 perusteella. Se ilmaisee, kuinka suuri osa vasteen y vaihtelusta voidaan selittää selittävien muuttujien avulla. Regressioanalyysin tapauksessa selitysaste saadaan kaavasta

$$R^2 = SSR/SST = \sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2,$$

missä SSR on regressioneliösumma ja SST kokonaisneliösumma. Verrataan mallin avulla selitettyä vaihtelua kokonaisvaihteluun. SSR muodostuu sovitteiden eli estimoitujen arvojen $\hat{y}_i$ ja odotusarvon $\bar{y}$ erotuksien neliöiden summasta ja SST oikeiden arvojen y_i ja odotusarvon erotuksien neliöiden summasta. Jos SSR on yhtäsuuri kuin SST, on selitysaste yksi. Yhden selittäjän ja vakion tapauksessa R^2 saadaan korottamalla Pearsonin korrelaatiokerroin r toiseen potenssiin. (Leppälä 2006)

Lineaarinen malli on luonnollinen ja usein riittävä approksimaatio monimutkaisemmallekin riippuvuudelle, mutta se ei sovellu kaikkiin tilanteisiin (Heikkinen 2005). Malli voi olla selvästi epälineaarinen, eikä välttämättä muunnoksellakaan lineroitavissa. Tällöin tarvitaan muita mallintamismenetelmiä. (Nummi 2008.)

3.2 Yleistetyt lineaariset mallit

Yleistetyt lineaariset mallit (generalized linear models, GLM) sallivat vasteen jakauman olevan muutakin kuin normaalinen. (Wood 2006.) Ne ovat klassisten lineaaristen mallien perheen laajennus. Yleistetyillä lineaarisilla malleilla ei suoraan mallinneta odotusarvoa, vaan sen funktiota. Yleistetyssä sekä klassisessa lineaarisessa mallissa voi selittäjänä x_i olla jatkuva muuttuja, luokittelumuuttuja tai molempia. Jatkuvien muuttujien tapauksessa kyseessä on regressioanalyysi, luokittelumuuttujien tapauksessa varianssianalyysi ja molempien yhdistelmässä kyseessä on kovarianssianalyysi. Lisäksi malli voi sisältää vakiotermin, muuttujien muunnoksia sekä muuttujien yhdistelmiä eli yhdysvaikutuksia. (Heikkinen 2005.)

Yleistetyissä lineaarisissa malleissa on kolme komponenttia: satunnaiskomponentti, systemaattinen komponentti sekä linkkifunktio. Satunnaiskomponentti määrittää vasteen y sekä sen jakauman ja systemaattinen komponentti määrittää vastaavasti selittävän muuttujan x_i. Linkkifunktio g määrittää sen funktion rakenteen, jonka kautta vasteen y odotusarvo μ riippuu lineaarisesti selittävästä muuttujasta x_i. Toisin sanoen linkkifunktio yhdistää systemaattisen komponentin satunnaiskomponentin arvoihin.

Yksinkertainen yleistetty lineaarinen malli on vektorimuodossa muotoa

$$g(\mu) = \mathbf{x}'\boldsymbol{\beta},$$

missä

$$\mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

$\mu = E(y_i)$ ja g on linkkifunktio. (Isotalo 2009.) Yleistettyjen lineaaristen mallien tapauksessa yleensä oletetaan, että y_i noudattaa jotain eksponentiaaliseen perheeseen kuuluvaa jakaumaa. Käytännön mallintamisessa käytettyjä jakaumia ovat Poissonin, binomi-, gamma- ja normaalijakaumat.

Yleistettyjen lineaaristen mallien oletukset ovat pitkälti samoja kuin lineaaristen mallienkin, lisänä kuitenkin linkkifunktion ja jakauman määrittäminen. (Wood 2006.) Linkkifunktio g valitaan vasteen jakaumaoletuksen mukaan ja sen oletetaan olevan monotoninen sekä derivoituva. Yksinkertaisin linkkifunktio on identiteettilinkki. Identiteettilinkki on muotoa $g(\mu) = \mu$, jolloin y_i:n odotusarvon μ_i oletetaan olevan lineaarisesti riippuvainen selittävien muuttujien arvoista (Isotalo 2009; Heikkinen 2005). Jos linkkifunktioksi valitaan identiteettilinkki ja jakaumaksi normaalijakauma, saadaan erikoistapauksena tavallinen lineaarinen malli (Wood 2006).

Muut linkkifunktiot mahdollistavat odotusarvon epälineaarisen riippuvuuden selittävistä muuttujista. Muita käytettyjä linkkifunktioita ovat logistinen linkkifunktio eli log-linkki sekä esimerkiksi binomi-jakauman tapauksessa käytettävät logit-linkki sekä probit-funktio. (Heikkinen 2005; Isotalo 2009; Hastie & Tibshirani 1990.) Log-linkki, $g(\mu) = log(\mu)$, sopii tilanteisiin, joissa odotusarvo μ ei voi olla negatiivinen. Sitä käytetään muun muassa Poissonin jakauman tapauksessa. Logit-linkkiä, $g(\mu) = log(\mu/(1-\mu))$, käytetään, kun odotusarvo on välillä $0 \leq \mu \leq 1$. Tällöin yleistettyä lineaarista mallia kutsutaan logistiseksi regressiomalliksi. (Isotalo 2009.)

Yleistetyillä lineaarisilla malleilla tehtävä estimointi ja päättely perustuu usein suurimman uskottavuuden menetelmään. Lineaarisen mallin yleistämisestä seuraa kuitenkin tiettyjä asioita. Mallin sovittaminen joudutaan tekemään iteratiivisesti ja lisäksi jakaumatulokset ovat aproksimaatioita. (Wood 2006.)

3.3 Additiiviset mallit

Monet epäparametriset menetelmät eivät toimi kunnolla, jos mallissa on suuri määrä riippumattomia selittäjiä. Toinen ongelma muun muassa tasoittaviin spliniestimaatteihin perustuvissa epäparametrisissa regressiomalleissa on tulkittavuus. Näihin ongelmiin esitti Stone (1985) ratkaisuna additiiviset mallit.

Mallia sanotaan additiiviseksi, kun mallin termien yksittäiset vaikutukset lisätään toisiinsa ja saavutetaan niiden yhteinen vaikutus. Jos kaksi prediktoria eivät käyttäydy additiivisesti suhteessa vasteeseen, on niiden välillä interaktiota. (Ruppert, Wand & Carroll 2003.) Additiivinen malli on yleistys tavallisista lineaarisista malleista ja on niitä joustavampi (Hastie & Tibshirani 1990).

Epäparametrinen additiivinen malli on muotoa

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \epsilon_i,$$

missä y_i on mallin vaste, joka riippuu additiivisesti tasoittavan funktion f_j kautta kovariaateista x_j. Riippumattomien residuaalien ϵ_i odotusarvo on nolla ja varianssi σ^2. (Faraway 2006; Hastie & Tibshirani 1990.)

3.4 Yleistetyt additiiviset mallit

Yleistetty additiivinen malli (generalized additive model, GAM) saadaan yleistetystä lineaarisesta mallista korvaamalla lineaarinen prediktori tasoittavien funktioiden summalla. Malli perustuu sakotettuihin tasoittaviin regressiofunktioihin. Sakkotermistä enemmän kohdassa 3.7.2. GAM on laajennus additiivisesta mallista samoin kuin yleistetty lineaarinen malli on laajennus lineaarisesta mallista. (Wood 2006.)

Yleistetyissä additiivisissa malleissa jotkut lineaaristen prediktorien jatkuvia selittäjiä vastaavista komponenteista voidaan korvata x:n epäparametrisilla funktioilla. Niitä voidaan approksimoida esimerkiksi lokaaleilla polynomeilla tai tasoittavilla splineillä. Epäparametrisille termeille voidaan määrittää approksimatiiviset vapausasteet, jotka riippuvat tasoituksen voimakkuudesta. (Heikkinen 2005.) Yleistetyissä additiivisissa malleissa oletetaan, että vastemuuttujien odotusarvo riippuu additiivisesti ennustavasta muuttujasta linkkifunktion kautta.

Yleistetyissä additiivisissa malleissa lineaarinen prediktori on muotoa

$$\beta_0 + \sum_{j=1}^{p} f_j(x_j),$$

missä tasoittavat funktiot f_j estimoidaan aineistosta. (Faraway 2006.) Esimerkki GAM-mallista voisi olla

$$g(\mu_i) = \mathbf{x}_i'\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + ..., \tag{3.1}$$

missä μ_i on y_i:n odotusarvo $E(y_i)$ ja y_i on vastemuuttuja, joka noudattaa jotain eksponentiaaliseenperheeseen kuuluvaa jakaumaa. Funktio g on linkkifunktio, $\mathbf{x}_i'$ on rivi mallimatriisista, θ on vastaava parametrivektori ja f_i:t ovat kovariaattien, x_k, tasoittavia funktioita. Vastemuuttujan jakaumaksi voidaan määrittää mikä tahansa eksponentiaaliseen perheeseen kuuluva jakauma, esimerkiksi normaali- binomi-, gamma- tai Poissonin jakauma. Linkkifunktio g, jonka välityksellä prediktorimuuttujat ovat yhteydessä vasteeseen, valitaan kuten yleistettyjen lineaaristen mallienkin tapauksessa. Normaali-, gamma- ja Poissonin jakaumien tapauksessa vaihtoehtoina ovat muun muassa log-linkki ja identiteettilinkki. Binomi-jakauman tapauksessa logit-linkki. (Hastie & Tibshirani 1990.)

Malli (3.1) sallii melko joustavan määrittelyn vasteen riippuvuudesta kovariaateista. Tasoittavat funktiot on kuitenkin määriteltävä ja on valittava niiden tasoitusaste. Yleistettyjen additiivisten mallien esittämiseen käytetään *regressiosplinejä (regression splines)*, jotka on estimoitu regressiometodeilla. (Wood 2006.) Tasoitusmenetelmistä enemmän alaluvussa 3.7. Yleistettyjen additiivisten mallien joustavuus aiheuttaa helposti myös ylisovittamista. Aineistoon käytetään liian monimutkaista mallia. Näin saatava sovite harvoin toistuu, kun mallinnetaan aineistoa uudelleen. (Hastie & Tibshirani 1990.)

Yleistettyjä additiivisia ja (yleistettyjä) lineaarisia malleja voidaan verrata toisiinsa esimerkiksi käytettävyyden avulla. Lineaariset mallit ovat yleensä helposti ymmärrettäviä ja niiden tulosten yhteenveto ja vertailu onnistuu hyvin. Yleistetyt additiiviset mallit ovat vaikeampia tulkita, varsinkin jos niihin sisältyy prediktorien monimutkaisia epälineaarisia vaikutuksia. (Hastie & Tibshirani 1990.)

3.5 Poissonin regressio

GAM-mallintamisessa saadaan erikoistapauksena Poissonin regressio, jossa vastemuuttujana on yleensä lukumäärä. Vasteen arvot ovat kokonaislukuja, joiden voidaan olettaa olevan peräisin Poissonin prosessista, joka on Poissonin jakauman tuottama stokastinen prosessi. Tyypillinen esimerkki on tietyssä ajassa ja populaatiossa tiettyyn sairauteen sairastuvien lukumäärä.

Vaste y_i on tiettynä ajanjaksona ja tietyltä alueelta havaittujen kiinnostavien tapausten määrä. Jos tapaukset voidaan olettaa toisistaan riippumattomiksi ja niiden havaintoyksikön sisäinen intensiteetti vakioksi, niin y_i noudattaa Poissonin jakaumaa. (Heikkinen 2005.)

Poissonin jakauman todennäköisyysfunktio on muotoa

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!},$$

missä $y = 0, 1, 2,...$, ja odotusarvo $E(y) = \lambda$. Poissonin jakauman ominaisuuksiin kuuluu, että varianssi ja odotusarvo ovat yhtäsuuria

$$var(y_i) = E(y_i).$$

Tämä ei kuitenkaan aina käytännössä ole voimassa, jolloin jakauman todellinen vaihtelu eroaa Poissonin vaihtelusta. Tämä voi johtua havaintojen ryvästymisestä, havaitsemattomien selittäjien vaihtelusta tai siitä, että ajanjakso ja/tai alue vaihtelevat. Usein mallinnuksessa onkin tarpeen olettaa, että

$$var(y_i) = \sigma E(y_i).$$

Toisin sanoen varianssin ja odotusarvon välillä on suhde, joka riippuu hajontaparametrista σ. Useimmiten varianssi on odotusarvoa suurempi, joten hajontaparametri on suurempi kuin yksi. Kyseessä on *ylihajonta (overdispersion)*. Täl-

löin jakauman todellinen vaihtelu on Poissonin vaihtelua suurempi. Ylihajonnasta voi seurata esimerkiksi, että luottamusvälit muodostuvat liian kapeiksi. (Ruppert et al. 2003.)

Linkkifunktiona käytetään Poissonin jakauman tapauksessa log-linkkiä, $f(z) = log(z)$. Toisin sanoen funktio, joka yhdistää odotusarvon lineralisoituun prediktoriin, on log-funktio. (Hastie & Tibshirani 1990.)

Poissonin regressiomallien tapauksessa mallinnuksen hyvyyttä voidaan arvioida residuaalien perusteella. Residuaalien ϵ määrittely havainnolle i on

$$\epsilon_i = \frac{(y_i - \hat{y}_i)}{\sqrt{\hat{y}_i}},$$

jossa $\hat{y}_i$ on Poissonin regressiomalliin perustuva ennuste havainnolle i. Suuret positiiviset residuaalit tarkoittavat, että havaittu ennuste on paljon suurempi kuin mitä oli ennustettu mallin perusteella. Vastaavasti pienet negatiiviset residuaalit viittaavat siihen, että havaittu vaste on pienempi kuin mallin mukainen ennuste. (Hastie & Tibshirani 1990.)

3.6 Interaktio

Kahden muuttujan välillä on interaktiota, kun niiden yhteisvaikutus eroaa samojen muuttujien erillisten vaikutusten summasta. Ja toisaalta, mikäli muuttujan vaikutus on riippumaton toisen muuttujan vaikutuksesta, sanotaan niiden olevan additiivisia. Poikkeavuutta additiivisuudesta sanotaan interaktioksi (Ruppert et al. 2003). Esimerkiksi korkea verenpaine ja diabetes nostavat kumpikin kolesterolitasoa yksinäänkin, mutta jos molemmat sairaudet ovat yhtäaikaa, nousee kolesterolitaso enemmän kuin vain erillisten tekijöiden summan verran. (Wood 2006.) Interaktiotilanteessa yleistettyyn additiiviseen malliin on lisättävä interaktiotermi, jotta muuttujien yhdysvaikutusta pystytään mallintamaan.

Yleistetty additiivinen malli interaktioiden kanssa on muotoa

$$g(\mu_i) = \beta_0 + \sum f_i(x_i) + \sum f_{ij}(x_i, x_j),$$

missä $f_{ij}(x_i, x_j)$ on interaktiotermi (Roca-Pardiñas & Cadarso-Suárez 2005). Interaktiotermiä voidaan arvioida sen tilastollisen merkitsevyyden ja selitysasteen muutoksen avulla. Varsinainen tulkinta tapahtuu yleensä graafisesti kuvion avulla.

Interaktiotermin merkitsevyyttä voidaan testata eri testeillä. Testattava hypoteesi on tällöin $H_0 : f_{ij} = 0$. Jos H_0 on tosi, x_i:n ja x_j:n välillä ei ole interaktiota millään pareilla (i, j). Mahdollinen testi on muun muassa *uskottavuusosamäärätesti (likelihood ratio test)*. (Roca-Pardiñas & Cadarso-Suárez 2005.)

3.7 Tasoitusmenetelmät

Additiivisissa ja yleistetyissä additiivisissa malleissa sekä muissa epäparametrisissa regressiomalleissa voidaan käyttää *tasoittavaa funktiota (smooth function)* f. Tasoittava funktio voidaan estimoida esimerkiksi valitsemalla lineaarinen kanta ja määrittämällä funktioavaruus. (Wood 2006.)

Tasoittava funktio voidaan palauttaa lineaariseksi oikealla kantafunktiolla. Kantafunktiot saadaan yleensä jollakin sopivalla muunnoksella alkuperäisten selittäjien arvoista. (Nummi 2008.) Jos $b_j(\mathrm{x})$ on kantafunktio, niin tasoittava funktio on muotoa

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j. \tag{3.2}$$

Esimerkiksi yksinkertaisessa mallissa

$$y_i = f(x_i) + \epsilon_i, \tag{3.3}$$

funktio f on tasoittava funktio, y_i on vastemuuttuja, x_i on kovariaatti ja ϵ_i residuaali. Kun tasoittava funktio (3.2) sijoitetaan funktioon (3.3) saadaan periaatteessa tavallinen lineaarinen malli. (Wood 2006.)

Splinimenetelmät tarjoavat epäparametrisen lähestymistavan jakauman ennustamiseen sekä poikittais- että pitkittäisaineistojen tapauksessa. Epäparametrisia regressiomenetelmiä ovat muun muassa splinimenetelmät eli regressiosplinit ja tasoittavat splinit, Kernel-tasoitus sekä muut paikalliset eli *lokaalit polynomisovitteet (local polynomial fitting)* (Kääriä 2007). Seuraavaksi tarkastellaan tarkemmin regressiospliniä ja tasoittavaa spliniä sekä lokaaleja polynomeja, erityisesti Kernel-tasoitusta.

3.7.1 Regressiosplini

Regressiosplini on tasoitusmenetelmä, jossa solmukohtien määrän ja sijainnin määrittäminen on keskeistä. Solmukohdat ovat pisteitä $\tau_0, \tau_1, ...\tau_K, \tau_{K+1}$, jotka on määritelty tietyllä välillä $[a, b]$ siten, että $a = \tau_0 < \tau_1 < ... < \tau_{K+1} = b$. (Kääriä 2007.) Regressiosplini koostuu polynomifunktioiden paloista eli välin $[a, b]$ osaväleistä, joiden määrä määräytyy solmukohtien mukaan. Solmukohtien määrä ja sijainti on määritettävä oikein, jotta mitään yksityiskohtia ei häviä tai toisaalta paikallinen vaihtelu ei kasva liian suureksi. (Smith & Kohn 1996.) Solmukohtien valintaan voidaan käyttää esimerkiksi mallinvalintamenetelmiä. Toinen mahdollisuus on sijoitaa solmukohdat tasavälein havaintovälille tai valittuihin otoskvantiileihin. (Nummi 2008.)

Yleisimmin käytetty regressiosplini on kuutiollinen regressiosplini. Sillä on kaksi ehtoa: jokaisella osavälillään funktio on kolmannen asteen polynomi ja funktio on kaksi kertaa jatkuvasti derivoituva. (Nummi 2008.) Muita regressiosplinejä ovat muun muassa sakotettu regressiosplini sekä ”thin plate” regressiosplini. Viimeksi mainittu perustuu ”thin plate” splineihin, jotka ovat käteviä varsinkin usean prediktorin tasoitusfunktioden estimointiin. (Wood 2006)

3.7.2 Tasoittava splini

Solmukohtien määrän ja sijainnin valinnalta vältytään, kun käytetään tasoittavaa spliniä. Siinä solmukohtina toimivat kaikki mittauspisteet ja tasoituksen määrää ja voimakkuutta säädellään sakkotermin avulla. Ainoa määriteltävä parametri on tasoitusparametri $\lambda > 0$. (Kääriä 2007.)

Tasoitusparametrilla voidaan kontrolloida splinikäyrän tasaisuutta. Jos λ on pieni, kulkee splinikäyrä tarkasti aineiston havaintopisteiden kautta, mutta tällöin ylisovitetaan mallia ja mukaan tulee liikaa kohinaa. Suuri λ taas saa käyrän "oikaisemaan" enemmän. Toisin sanoen, kun λ lähestyy ääretöntä, lähestyy splinikäyrä regressiosuoraa. Sopiva tasoitusparametri olisikin hyvä löytää jostain ääripäiden väliltä. Hyvä tasoitusparametri valitaan siten, että estimoitu tasoittava splinifunktio $\hat{f}$ olisi mahdollisimman lähellä todellista jakaumaa f. (Wood 2006.)

Tasoitusparametrin valinta voidaan tehdä, varsinkin käytännön sovelluksissa, kuvien avulla tai erilaisia valintamenetelmiä käyttäen. Erilaisia valintamenetelmiä on paljon, muun muassa uskottavuusfunktioon perustuva menetelmä, ristiinvalidointi (cross-validation) sekä informaatiokriteereihin perustuvat menetelmät. (Nummi 2008; Kääriä 2007.) Eri menetelmiä esitellään tilastollisessa kirjallisuudessa (mm. Hastie & Tibshirani 1990) ja aiheesta on aktiivista tutkimusta (Jackman 2004).

3.7.3 Lokaalit polynomit

Lokaalisissa polynomimalleissa on tarkoituksena paikallisesti approksimoida funktiota g sopivan asteisella polynomilla. Lähtökohtana käytetään Taylorin sarjakehitelmää. Pisteen x_0 ympäristössä Taylorin sarjakehitelmä on muotoa

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + ...$$

Olkoon t_0 kohta, jossa halutaan approksimoida funktiota g ja oletetaan, että funktion $g(t)$ derivaatat ovat jatkuvia siinä pisteessä asteeseen $p + 1 \geq 0$ saakka. Nyt funktion $g(t)$ lokaali approksimaatio p-asteisella polynomilla on

$$g(t) \approx g(t_0) + g'(t_0)(t - t_0) + ... + \frac{g^{(p)}(t_0)}{p!}(t - t_0)^p,$$

missä $g^{(p)}(t_0)$ on funktion g p-asteinen derivaatta pisteessä t_0.

Lokaaleista polynomimalleista voidaan johtaa muun muassa niin sanottu Nadaray-Watson-estimaattori. Se on muotoa

$$\hat{g} = \sum_j \frac{y_j K(\frac{t_j - t_i}{b})}{K(\frac{t_j - t_i}{b})},$$

missä b on "ikkunan", leveys ja K on Kernel-funktio, jolla painotetaan havaintoja. (Nummi 2008.) Seuraavassa kappaleessa tarkastellaan tarkemmin lokaalisiin polynomimalleihin kuuluvaa Kernel-tasoitusta.

Kernel-tasoitus Kernel-tasoitus perustuu Kernel-funktioon $K(t)$, joka on yleensä symmetrinen tiheysfunktio. Sitä käytetään muun muassa polynomitasoittajissa havaintojen painottamiseen. Tavallisimmat Kernel-funktiot ovat Uniform-Kernel sekä niin sanottu Gaussian-Kernel, joista jälkimmäinen tuottaa tasaisemman sovitteen. Muita ovat Epanechnikov-, Biweight- ja Triweight-Kernelit.

Kernel-funktiot ovat erikoistapauksia symmetrisestä $Beta$-perheestä ja ovat muotoa

$$K(t) = \frac{1}{Beta(1/2, 1+\gamma)}(1-t^2)^\gamma, \tag{3.4}$$

missä $\gamma = 0, 1, ...$ ja $Beta(a, b)$ on $Beta$-funktio, joka on muotoa

$$Beta(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt.$$

Funktion (3.4) parametrin γ arvot 0, 1, 2 ja 3 vastaavat Uniform-, Epanechnikov-, Biweight- ja Triweight-Kerneleitä. Kun γ lähestyy ääretöntä, saadaan Gaussian-Kernel. Esimerkiksi Uniform-Kernelin tapauksessa sijoittamalla $\gamma = 0$ funktioon (3.4) saadaan vastaukseksi

$$K(t) = \frac{1}{Beta(1/2, 1)}(1-t^2)^0 = \frac{1}{2}.$$

Vastaavasti Gaussian-Kernelin tapauksessa, jolloin $\gamma \to \infty$, saadaan vastaukseksi $K(t) = e^{(-t^2/2)}/\sqrt{2\pi}$. Ja edelleen muut Kernelit saadaan muodostettua sijoittamalla niitä vastaava γ funktioon (3.4). (Nummi 2008.)

4 Menetelmien soveltaminen aineistoon

4.1 GAM-mallin sovittaminen aineistoon

Pienhiukkasaineiston ja vasteena tarkasteltavan terveysaineiston välillä ja sisällä on monimutkaisia riippuvuuksia, joita on vaikea mallintaa tavallisilla parametrisilla malleilla. Tarkoitukseen sopivampia ovat yleistetyt additiiviset mallit. Vasteen ja prediktorien välillä olevaa syy–yhteyttä voidaan kutsua stokastiseksi kausaliteetiksi. Se tarkoittaa, että seuraus eli tehdyt diagnoosit voivat esiintyä ilman tutkittavia syitä ja toisaalta syyt voivat esiintyä ilman seurausta. (Auvinen 2008.)

Jakaumaoletuksena on Poissonin jakauma, koska tarkasteltavat vastemuuttujat ovat lukumääriä. Vasteen ja selittävien termien välillä käytetään loglinkkiä eli tutkitaan vaikutuksia logaritmoituun vasteeseen. Mallissa käytetään osaan muuttujista tasoittavaa funktiota.

Perusmalli on muodostettu aiempien tutkimusten mallien (mm. Halonen et al. 2008) pohjalta ja on muotoa

$$log(E(y)) = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + f_5(x_5) + \beta_1 x_6 + \mu_i,$$

missä y noudattaa Poissonin jakaumaa ja f_i on tasoittava funktio. Ensimmäinen muuttuja mallissa, x_1, on hiukkasmuuttuja, jota vaihdellaan muiden termien pysyessä samoina. Taustamuuttujat x_2 - x_5 ovat lämpötila, lämpötila ero, ajoneuvojen lukumäärä sekä aika päivinä. Ilman tasoitusta malliin tulevat taustamuuttujat x_6, joka on ilmankosteus sekä μ_i, joka on kategorinen influenssamuuttuja. Ilmankosteuden vaikutus vasteeseen on lineaarinen, joten tasoitusta ei tarvita.

Viiveet Aiempien tutkimusten (Halonen et al. 2008) ja tutkittavan vasteen luonteen perusteella voidaan pitää perusteltuna tutkia eri pituisten viiveiden vaikutusta vasteeseen. Hoitoon ei päästä tai hakeuduta välttämättä samana päivänä, kun altistus on tapahtunut. Malli pysyy muuten samana, mutta testattavasta hiukkasmuuttujasta x_1 muodostetaan uusia muuttujia summaamalla yhteen muuttujan viivästettyjä arvoja. Käytetään 0, 1, 3, 5 ja 7 päivän summaviiveitä, joista esimerkiksi viiveen 3 summamuuttuja muodostetaan laskemalla yhteen tarkasteltavan päivän lisäksi kolmen edellisen päivän arvot. Viivesumma ottaa siis huomioon mahdollisen useamman päivän kestäneen ilmansaasteiden pitoisuuden nousun vaikutuksen terveyskäynteihin.

4.2 Interaktiotermin lisääminen malliin

Tutkitaan, onko mallissa muuttujien välillä interaktioita eli yhdysvaikutuksia. Interaktiotermin lisäyksen jälkeen alkuperäinen yleistetty additiivinen malli on muotoa

$$log(E(y)) = f(x_1) + f(x_2) + f(x_3) + f(x_4) + f(x_5) + \beta_1 x_6 + \mu_i + f(x_i, x_j),$$

missä $f(x_i, x_j)$ on lisätty interaktiotermi ja x_i ja x_j voivat olla muuttujia $x_1 - x_6$ tai muita aineiston muuttujia. Tutkitaan, parantaako interaktiotermin lisääminen mallia. Interaktiota voidaan tutkia esimerkiksi visuaalisesti kuvien avulla tai tarkkailemalla selitysasteen muutoksia ja malliin lisätyn interaktiotermin ja muiden selittäjien merkitsevyyksiä.

4.3 Tulokset eri vasteilla

4.3.1 Vasteena kuolemien lukumäärä

Kun mallissa on vasteena kaikkien kuolemien lukumäärät, jäävät selitysasteet odotetusti melko alhaisiksi (Liite 2). Paras selitysaste saadaan viiveellä nolla. Hiukkasmuuttujien saamat p-arvot eivät ole yhdessäkään mallissa merkitseviä viiden prosentin riskitasolla. Taustamuuttujista ainoastaan aika tulee viiden prosentin riskitasolla merkitseväksi ja lähinnä vain silloin, kun hiukkasmuuttujana ovat pienhiukkaset ja viiveenä 0, 1 tai 7.

Paras kuolemien lukumäärää selittävä malli, selitysasteen mielessä, saadaan, kun hiukkasmuuttujana on $PM_{2.5}$ summaviiveellä kolme. Selitysaste on tuolloin 3.35 prosenttia. Malliin lisätyistä interaktiotermeistä yksikään ei ole tilastollisesti merkitsevä. Mallien selitysasteista interaktiotermin lisäämisen jälkeen sekä interaktiotermien merkitsevyyksistä lisää liitteessä 3.

4.3.2 Vasteena hengityselinoireet

Hengityselin- eli keuhko-oireita on tutkittu eri ikäryhmissä sekä kaikkia ryhmiä yhteensä. Eri ikäryhmistä työikäisille eli 15–65-vuotiaille saadaan suurin selitysaste, noin 74 prosenttia. Heikoin selitysaste saadaan vanhimmassa, yli 74-vuotiaiden ryhmässä, jossa R^2 jää noin 56 prosenttiin. Kaikkia ikäryhmiä yhdessä tarkasteltaessa selitysasteeksi saadaan noin 77 prosenttia.

Taustamuuttujista ainoastaan ajoneuvojen lukumäärä on jokaisessa mallissa merkitsevä p-arvon ollessa alle 0.001. Se myös muodostaa yksin suurimman osan mallin selitysasteesta. Ajoneuvojen lukumäärän ollessa yksin mallissa prediktorina ja vasteen ollessa keuhkodiagnoosit, nousee selitysaste jo noin 60 prosenttiin. Yli 65-vuotiailla ajoneuvojen lukumäärä selittää hieman vähemmän, noin 50 prosenttia. Suurimmassa osassa malleista myös aika ja lämpötila ovat selvästi merkitseviä. Lämpötila ero ja ilman kosteus taas eivät ole yhdessäkään mallissa merkitseviä, kun rajaksi otetaan prosentin riskitaso. Vanhimpien, yli

74-vuotiaiden ryhmän, ollessa vasteena ainoa selvä selittäjä on ajoneuvojen lukumäärä. Muutamassa tapauksessa myös aika on merkitsevä taustamuuttuja.

Kun tarkastellaan malliin lisättyjä interaktiotermejä kuvaajien avulla näyttää, että selvimpiä interaktioita on lähes kaikkien vasteiden tapauksessa muuttujien aika ja ajoneuvojen lukumäärä välillä. Myös muuttujien aika ja lämpötila, aika ja hiukkasmuuttuja sekä ajoneuvojen lukumäärä ja hiukkasmuuttuja välillä on useiden vasteiden ja hiukkasmuuttujien tapauksessa interaktiota. Kun taas arviodaan interaktiotermin vaikutusta selitysasteen muutoksen perusteella, nostaa lämpötilan ja ajoneuvojen lukumäärän interaktio monessa mallissa eniten selitysastetta. Myös ajan interaktiot lämpötilan, lämpötila eron, ajoneuvojen lukumäärä ja hiukkasmuuttujan kanssa nostavat monessa mallissa selvästi selitysastetta. Kyseiset interaktiotermit ovat yleensä myös erittäin merkitseviä selittäjiä malleissa, varsinkin keuhkodiagnoosien tapauksessa. Muista merkitsevien interaktioiden termeistä poiketen, lämpötila ero ei yksinään ole malleissa merkitsevä. Sen ja samalla inversiotilanteen vaikutus vasteeseen saadaan esiin vasta interaktion muodossa.

Lähes kaikki merkitsevät interaktiot muodostuvat taustamuuttujien välillä tai taustamuuttujan ja mallin hiukkasmuuttujan välillä. Malliin alunperin kuulumattomia hiukkasmuuttujia sisältävät interaktiot eivät yleensä ole tilastollisesti selvästi merkitseviä. Kun lisätään malliin lukumääräpitoisuusmuuttujan interaktio massapitoisuusmuuttujan kanssa, muuttuu selitysaste korkeintaan 0.4 prosenttiyksikköä eikä interaktiotermi ole tilastollisesti merkitsevä. Ainoastaan aktiivinen pinta-ala SA muodostaa muutamassa mallissa toisen hiukkasmuuttujan kanssa merkitsevän interaktion. Tällöinkin selitysasteen muutos on korkeintaan vain noin 0.5 prosenttiyksikköä. Tarkastellaan tuloksia tarkemmin eri hiukkasmuuttujaryhmissä.

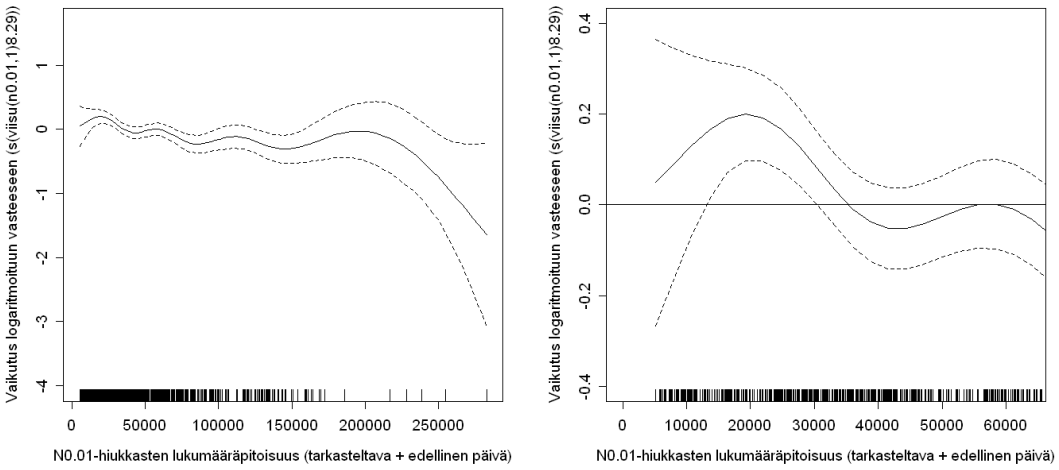
Lukumääräpitoisuusmuuttujat Lukumääräpitoisuusmuuttujilla selitysaste on suurin summaviiveellä yksi. Vanhimmissa ikäryhmissä myös nollaviiveisten mallien selitysasteet ovat suuria. Kun verrataan tarkemmin eri mallien selitysasteita, paras kaikkia hengityselindiagnooseja selittävä malli saadaan, kun hiukkasmuuttujana on $N_{2.5}$ yhden päivän summaviiveellä (Liite 2). Sama hiukkasmuuttuja antaa parhaan selitysasteen myös 65–74-vuotiaiden ryhmässä. Alle 15-vuotiaiden sekä yli 74-vuotiaiden tapauksessa suurin selitysaste saadaan, kun x_1 on $N_{0.1}$ myöskin yhden päivän summaviiveellä. Mallin termeistä hiukkasmuuttuja on merkitsevä yhden prosentin riskitasolla lähinnä silloin, kun x_1 on lukumääräpitoisuusmuuttuja. Esimerkiksi, kun vasteena on kaikki keuhkooireet *keuhkai* ovat merkitsevät hiukkasmuuttujat yhden ja kolmen summaviiveillä $N_{2.5}$, $N_{0.1}$ ja $N_{0.01}$

Kuvassa 4.1 on esitetty hiukkasmuuttujan $N_{0.01}$ vaikutus logaritmoituun vasteeseen, keuhkodiagnoosit 65–74-vuotiailla (*keuh*6574), summaviiveen ollessa yksi. Vaikutus on tilastollisesti merkitsevä, p-arvo on 0.00267. Kuvassa näkyy yhtenäisenä viivana funktion estimaatti sekä kahden keskivirheen verran suuntaansa oleva katkoviivoin rajattu alue, joka muodostaa 95 prosentin

luottamusvälin. (Wood 2001; Hastie & Tibshirani 1990.)

X-akselin piikit kuvaavat havaintojen frekvenssejä, tässä tapauksessa viivemuuttujan arvoja. Yksi piikki kuvaa yhtä tapausta. Summaviiveellä yksi yksittäinen tapaus sisältää tarkasteltavan sekä edellisen päivän eli kahden päivän $N_{0.01}$-hiukkasten lukumäärien summan. Esimerkiksi vasemmanpuoleisessa kuvassa oikeanpuoleisin, suurimman x-akselin arvon saanut piikki, kuvaa päivien 22.–23.2.2007 hiukkaspitoisuuksien summaa. Suurin osa hiukkasmuuttujan viivästetyistä arvoista on pieniä lukumääräpitoisuuksia ja tapausten määrä laskee oikealle mentäessä. Suurimmat arvot ovat siis vain yksittäisiä tapauksia, jolloin myös luottamusväli kasvaa leveäksi.

Y-akseli kuvaa hiukkasmuuttujan summaviiveen vaikutusta logaritmoituun vasteeseen, kun taustamuuttujat ovat mukana mallissa. Asteikko on logaritminen. Y-akselin otsikosta näkee, mihin muuttujan tasoitusfunktiota on käytetty sekä kuinka suuret ovat termin efektiiviset vapausasteet. (Wood 2001.) Jos efektiiviset vapausasteet ovat lähellä yhtä, on prediktorin vaikutus vasteeseen lineaarinen. Tässä tapauksessa vapausasteet ovat melko suuret, 8.29.



Kuvio 4.1. Hengityselin diagnoosit ikäluokassa 65-74-vuotiaat

Oikeanpuoleisessa kuvassa on sama kuvaaja kuin vasemmalla, mutta nyt yksittäiset suuret arvot on rajattu kuvasta pois ja y-akselin jakaumaa on kavennettu. Kuvassa näkyy kuitenkin edelleen 75 prosenttia hiukkasmuuttujan arvoista. Kuvaan on lisätty suora y-akselin kohtaan nolla, mikä helpottaa merkitsevyyden arviointia. Kun nolla ei sisälly 95 prosentin luottamusvälille, on hiukkasten lukumääräpitoisuuden vaikutus vasteeseen merkitsevä.

Kuvista nähdään, että vaikutus vasteeseen on epälineaarinen. Hiukkasten lukumääräpitoisuuden vaikutuksen merkitsevyys vaihtelee. Alueella, jolla vaikutus vasteeseen on tilastollisesti merkitsevä, on vaikutus selvästi positiivinen eli diagnoosilukumääriä nostava. Muilla alueilla vaikutus voidaan tulkita neutraaliksi. Jos tutkitaan saman hiukkasmuuttujan vaikutusta samaan vasteeseen

ilman, että mallissa on mukana muita muuttujia, on kuvaaja selvästi positiivinen hiukkasmuuttujien pitoisuuden kasvaessa. Tulkinnassa tulee siis ottaa huomioon myös muiden mallissa olevien muuttujien selittämä osuus.

Tutkitaan tarkemmin selitysasteella mitattaessa parhaita malleja ja yritetään vielä interaktiotermien lisäämisellä parantaa niiden selitysastetta. Kaikkien hengityselindiagnoosien tapauksessa suurin selitysaste saadaan lisäämällä malliin lämpötilan ja ajoneuvojen lukumäärän interaktio muiden mallin termien pysyessä entisellään (Liite 3). Tällöin selitysaste kasvaa 78.2 prosentista 81.7 prosenttiin eli 3.5 prosenttiyksikköä. Ajan ja ajoneuvojen lukumäärän tai ajan ja lämpötilan interaktioiden lisääminen nostaa alkuperäisen mallin selitysastetta noin kaksi prosenttiyksikköä. Kaikissa näissä tapauksissa lisätyn interaktiotermin p-arvo on erittäin merkitsevä, p-arvo < 0.001.

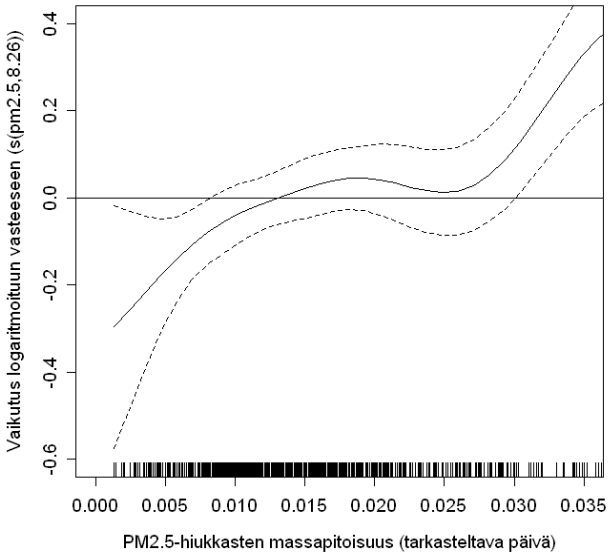
Kun tarkastellaan tarkemmin eri ikäryhmiä, alle 15-vuotiaiden tapauksessa suurin lisäys selitysasteeseen saadaan myöskin lisäämällä interaktiotermiksi lämpötila ja ajoneuvojen lukumäärä. Selitysaste kasvaa alkuperäisestä 70.9 prosentista lähes neljä prosenttiyksikköä. Seuraavassa ikäluokassa, 15–65-vuotiaat, selitysastetta kasvattaa lämpötilan ja ajoneuvojen lukumäärän interaktiotakin enemmän ajan ja lämpötilan interaktio, joka nostaa selitysastetta vajaa kolme prosenttiyksikköä. Vanhimmissa ikäluokissa selitysasteen nousu on hieman vähäisempää. Eniten R^2:sta kasvattavat edellisten interaktioiden ohella ajan ja ajoneuvojen lukumäärän yhteisvaikutus.

Massapitoisuusmuuttujat Massapitoisuusmuuttujilla selitysaste on suurin lähinnä summaviiveillä kolme tai viisi. Vanhimmissa ikäryhmissä myös nollaviiveisten mallien selitysasteet ovat suuria. Muuttujien $PM_{2.5}$ ja PM_{10} vaikutus ei kuitenkaan ole malleissa tilastollisesti kovin merkitsevä.

Tarkastellaan kuvaa, jossa vaikutus vasteeseen on tilastollisesti merkitsevä. Kuvassa 4.2 on $PM_{2.5}$-hiukkasten nollaviiveisten massapitoisuuksien vaikutus logaritmoituun vasteeseen. Vasteena on keuhkodiagnoosit 65–74-vuotiailla. Kuvasta on rajattu pois yksittäiset suuret arvot. Nähdään, että vaikutus vasteeseen kasvaa $PM_{2.5}$-pitoisuuden noustessa. Jos pitoisuus on pienempi kuin 0.013, on vaikutus vasteeseen negatiivinen eli diagnooseja on vähemmän. Jos taas $PM_{2.5} \geq 0.013$, on vaikutus positiivinen, mutta suurelta osin tilastollisesti ei-merkitsevä.

Muut ilmanlaatumuuttujat Muilla hiukkasmuuttujilla selitysaste on suurin lähinnä summaviiveillä kolme tai viisi. Typen oksidit ovat malleissa merkitseviä varsinkin yhden tai kolmen summaviiveillä prosentin riskitasolla. Esimerkiksi, kun vasteena on kaikki keuhko-oireet on NO_2 merkitsevä hiukkasmuuttuja summaviiveellä kolme. Eri ikäluokissa tulokset ovat samansuuntaisia. Tosin nuorimpien, alle 15-vuotiaiden ryhmässä, hiukkasmuuttujista ainoastaan NO_2 summaviivellä yksi on yhden prosentin riskitasolla merkitsevä.

Suurimman selitysasteen malli eri ikäryhmissä saadaan hiukkasmuuttujalla NO_2 summaviiveellä kolme vasteen ollessa 15–64-vuotiaiden hengityselin-



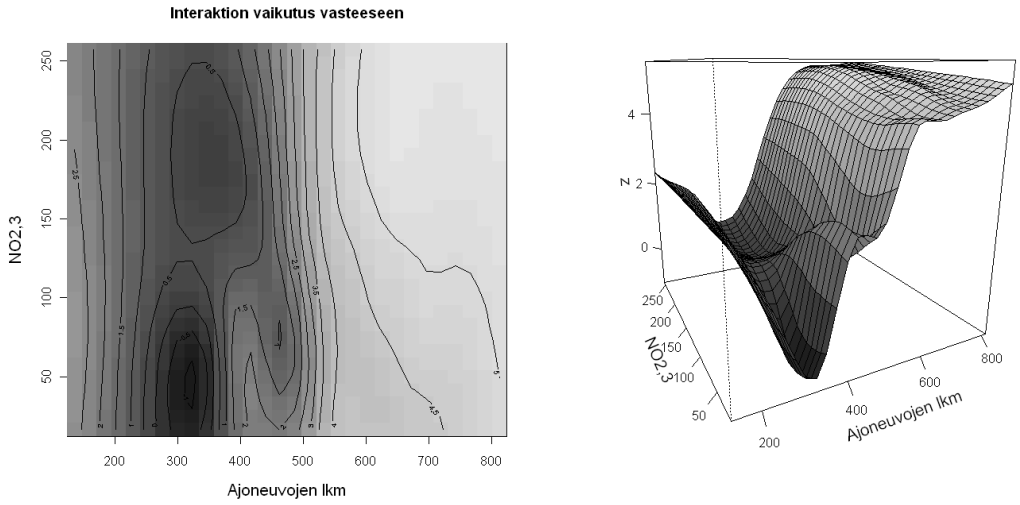
Kuvio 4.2. Hengityselin diagnoosit ikäluokassa 65-74-vuotiaat

diagnoosit. Tätä mallia on tutkittu edelleen lisäämällä eri interaktiotermejä. Kuvassa 4.3 on ajoneuvojen lukumäärän ja typpidioksidin interaktio, kun typpidioksidin viivesumma on kolme ja vasteena mallissa keuhkodiagnoosit 15–64-vuotiailla. Vasemmanpuoleisessa kuvassa interaktio on kuvattu kaksiulotteisesti ja oikeanpuoleisessa kolmiulotteisesti. Z-akseli kuvaa interaktion vaikutusta vasteeseen. Kuvista nähdään, että ajoneuvojen lukumäärän ja hiukkasmuuttujan välillä on interaktiota; vasemmanpuoleisessa kuvassa viivat eivät ole pystysuoria. Pienillä typpidioksidin arvoilla ajoneuvojen lukumäärä käyttäytyy erilailla kuin suuremmilla.

4.3.3 Vasteena sydänoireet

Vasteen ollessa sydändiagnoosit hiukkasmuuttujista ovat merkitseviä yhden prosentin riskitasolla osittain samat hiukkasmuuttujat ja samoilla viiveillä kuin hengityselinoireidenkin kohdalla. Selitysasteet jäävät hieman hengityselinoireiden selitysasteista, paitsi yli 74-vuotiaiden ryhmässä. Vanhimpien ryhmän R^2 on nyt ikäryhmien korkein, noin 69 prosenttia, kun alle 64-vuotiaiden ryhmässä selitysaste jää 64 prosenttiin ja 65–74-vuotiailla noin 66 prosenttiin. Kun tarkastellaan kaikkia ikäryhmiä yhdessä, saadaan selitysasteeksi noin 74 prosenttia.

Sydänoireidenkin tapauksessa taustamuuttujista ainoastaan ajoneuvojen lukumäärä on joka mallissa merkitsevä alle 0.1 prosentin riskillä. Sen ansiota on myös suurin osa selitysasteesta. Jos ajoneuvojen lukumäärä on ainoa prediktori mallissa vasteen ollessa sydändiagnoosit, saadaan selitysasteeksi noin 60 prosenttia kaikissa ikäluokissa. Toinen hyvin merkitsevä taustamuuttuja lähes joka



Kuvio 4.3. Ajoneuvojen lukumäärän ja NO_2-hiukkasten interaktio

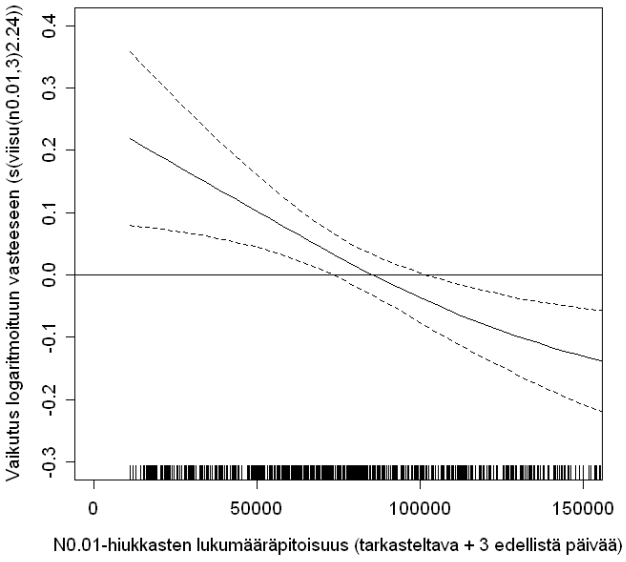
mallissa on aika. Lämpötila ei aivan yhtä selvästi kuulu malliin kuin keuhko-oireiden tapauksessa, mutta monessa mallissa p-arvo on kuitenkin pienempi kuin 0.01. Jälleen lämpötila ero ja ilman kosteus eivät ole merkitseviä taustamuuttujia. Myös influenssa-muuttuja tulee harvemmin malliin mukaan kuin keuhko-oireiden tapauksessa.

Interaktiot käyttäytyvät hyvin samoin tavoin kuin hengityselinoireiden tapauksessa. Selvimpiä interaktioita on varsinkin muuttujien aika ja ajoneuvojen lukumäärä välillä. Nyt selitysasteen nousu on kuitenkin vähäisempää, parhaimmillaan kahden ja kolmen prosenttiyksikön välillä. Sydändiagnoosienkin tapauksessa lähes kaikki merkitsevät interaktiot muodostuvat taustamuuttujien välillä tai taustamuuttujan ja mallin hiukkasmuuttujan välillä. Mikään interaktiotermeistä ei nouse selvästi parhaaksi eri malleissa. Kun kyseessä on kaikki sydändiagnoosit, aika ja lämpötila ovat eniten selitysastetta nostava interaktiopari; R^2 kasvaa 2.5 prosenttiyksikköä. Alle 65-vuotiaiden kohdalla selvästi eniten selitysastetta nostaa lämpötilan ja ajoneuvojen lukumäärän interaktio, 2.6 prosenttiyksikköä. Vanhempien ikäluokkien tapauksessa selitysaste nousee eniten, kun interaktion toinen termi on aika. Ajan ja lämpötilan interaktio on selvästi merkitsevä nuoremmissa ikäryhmissä, mutta yli 74-vuotiaiden kohdalla se on merkitsevä enää viiden prosentin riskitasolla. Kun verrataan interaktiotermien merkitsevyyksiä ja kuvaajia keuhkodiagnoosien tapauksen interaktioihin, havaitaan, että ne ovat hyvin saman suuntaisia. Samat interaktiotermit ovat yleensä yhtä merkitseviä molemmilla vasteryhmillä ja myös interaktiotermien kuvaajat vastaavat toisiaan.

Lukumääräpitoisuusmuuttujat Lukumääräpitoisuusmuuttujien kohdalla suurimmat selitysasteet saadaan lähinnä summaviiveilla yksi. Lukumääräpitoisuusmuuttujat ovat malleissa myös tilastollisesti merkitseviä varsinkin van-

himmassa 74-vuotiaiden ikäluokassa ja summaviiveillä yksi tai kolme. Paras kaikkia sydänoireita selittävä malli, selitysasteita verrattaessa, saadaan, kun hiukkasmuuttujana on $N_{0.01}$ summaviiveellä yksi. Erot selitysasteessa eri hiukkasmuuttujien välillä eivät kuitenkaan ole suuria.

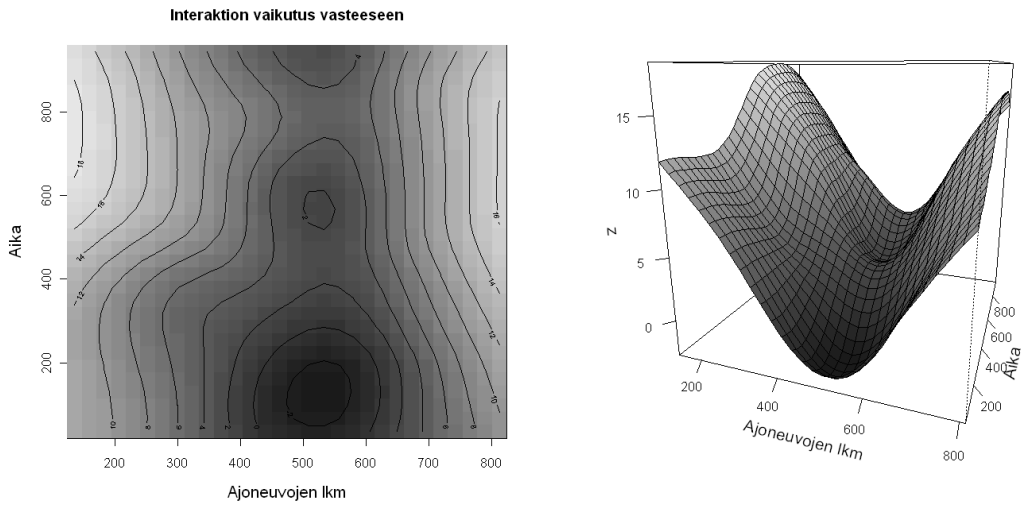
Kuvassa 4.4 on $N_{0.01}$-hiukkasten vaikutus logaritmoituun vasteeseen, kun vasteena on kaikki sydändiagnoosit ja hiukkasmuuttujan summaviiveenä kolme. Yhteys on tilastollisesti merkitsevä. Kuvassa x-akselilla on $N_{0.01}$-hiukkasmuuttujan lukumääräpitoisuudet yhden piikin kuvatessa tarkasteltavan sekä kolmen edellisen päivän tuntien lukumääräpitoisuutta. Y-akselilla on hiukkasmuuttujan vaikutus logaritmoituun vasteeseen $log(E(y))$, missä y on sydänoireiden lukumäärä kaikissa ikäluokissa. Y-akselin kohtaan $y = 0$ on piirretty merkitsevyyden hahmottamista helpottava suora. Kuvasta on rajattu ulkopuolelle yksittäiset suuret arvot, joiden vaikutus ei ole merkitsevä. Nähdään, että alueella, jolla vaikutus vasteeseen on tilastollisesti merkitsevä, on vaikutus aluksi positiivinen. Suurin osa arvoista sijoittuu tälle tai tilastollisesti eimerkitsevälle alueelle. Suurimpien pitoisuuksien alueella vaikutus vasteeseen on negatiivinen, mutta havaintoja on melko vähän.



Kuvio 4.4. Sydändiagnoosit kaikissa ikäluokissa

Tutkitaan tarkemmin suurimman selitysasteen saaneita malleja ja yritetään vielä interaktiotermien lisäämisellä parantaa niiden selitysastetta (Liite 3). Kuvassa 4.5 on ajoneuvojen lukumäärän ja ajan interaktio, kun mallissa vasteena on kaikki sydändiagnoosit ja hiukkasmuuttujana $N_{0.01}$ viivesummalla yksi. Vasemmanpuoleisessa kuvassa interaktio on kuvattu kaksiulotteisesti ja oikeanpuoleisessa sama interaktio on kolmiulotteisesti. Z-akseli kuvaa interaktion vaikutusta vasteeseen. Interaktiotermi on tilastollisesti merkitsevä. Kuvista nähdään, että muuttujien välillä on interaktiota. Ajoneuvojen lukumäärä

käyttäytyy erilailla eri aikoina.

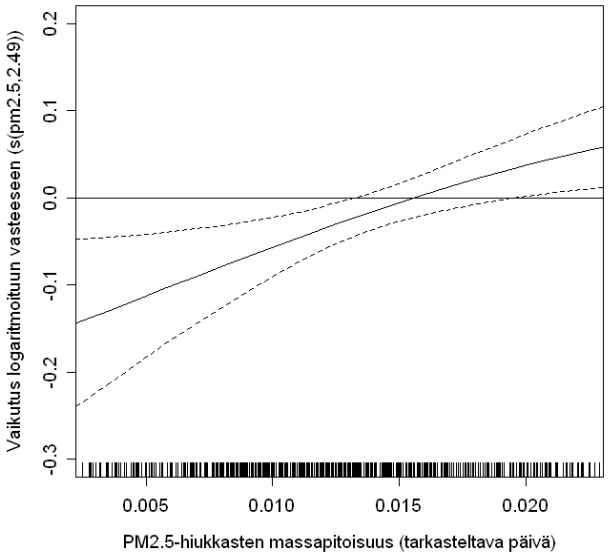


Kuvio 4.5. Ajoneuvojen lukumäärän ja ajan interaktio

Massapitoisuusmuuttujat Sydänoireiden tapauksessa muuttujien $PM_{2.5}$ ja PM_{10} selitysasteet ovat usein suurimmillaan nollaviiveellä. Kun tarkastellaan sydändiagnooseja eri ikäryhmissä, tuottaa $PM_{2.5}$ nollaviiveellä kaikista hiukkasmuuttujista suurimmat selitysasteet alle 65-vuotiaiden tapauksessa. Erot selitysasteessa eri hiukkasmuuttujien välillä eivät kuitenkaan ole suuria. Massapitoisuusmuuttujat ovat tilastollisesti merkitseviä prosentin riskitasolla eri ikäryhmistä vain alle 65-vuotiaiden kohdalla.

Kuvassa 4.6 on $PM_{2.5}$-hiukkasten vaikutus logaritmoituun vasteeseen, kun vasteena on sydändiagnoosit alle 65-vuotiailla ja hiukkasmuuttuja on mallissa viiveellä nolla. Yhteys on tilastollisesti merkitsevä. Kuvassa x-akselilla on $PM_{2.5}$-hiukkasmuuttujan massapitoisuudet ja y-akselilla on hiukkasmuuttujan vaikutus logaritmoituun vasteeseen, missä y on sydänoireiden massapitoisuus alle 65-vuotiailla. Kuvasta on rajattu ulkopuolelle yksittäiset suuret ja pienet arvot. Nähdään, että alueella, jolla vaikutus vasteeseen on tilastollisesti merkitsevä, on vaikutus aluksi negatiivinen, mutta kasvava. Suurin osa arvoista sijoittuu tälle tai tilastollisesti ei-merkitsevälle alueelle. Suurimpien pitoisuuksien alueella vaikutus vasteeseen on positiivinen.

Muut ilmanlaatumuuttujat Sydänoireiden tapauksessa muuttujan SA selitysasteet ovat usein suurimmillaan viiveellä nolla ja typen oksidien summaviiveellä yksi tai kolme. Typen oksidit ovat tilastollisesti merkitsevimpiä summaviiveellä kolme, kun taas SA ei ole merkitsevä yhdessäkään mallissa.



Kuvio 4.6. Sydändiagnoosit alle 65-vuotiailla

4.4 R-ohjelmisto ja gam-funktio

Analyysit suoritetaan tilasto-ohjelmisto R:llä versiolla 2.7.1 käyttämällä kirjastoa *mgcv* ja sen versiota 1.4-0. Kirjasto *mgcv* tarjoaa työkaluja yleistettyjen additiivisten mallien sekä muiden yleistettyjen harjaregressioiden (ridge regressio) käyttämiseen. Yleistettyjen additiivisten mallien esittämiseen käytetään *gam*-funktiota, jossa voidaan määritellä mallin ja aineiston lisäksi esimerkiksi haluttu funktioperhe, linkkifunktio sekä tasoitusfunktion solmukohtien määrän.

Gam-funktio on muotoa

```
gam(formula, family=gaussian(), data=list(), weights=NULL,
subset=NULL, na.action, offset=NULL, control=gam.control(),
method=gam.method(), scale=0, knots=NULL, sp=NULL,
min.sp=NULL, H=NULL, gamma=1, fit=TRUE, paraPen=NULL,
G=NULL, in.out, ...),
```

joka on tämän tutkielman tapauksessa muotoa

```
gam(formula, family=poisson, data=pht, na.action=na.omit,
scale=-1)
```

Jakaumaperheeksi valitaan siis Poissonin jakauma ja määritellään aineisto. Skaalaksi valitaan -1, mikä tarkoittaa, että tasausparametri valitaan yleistetyn ristiinvalidoinnin (generalized cross validation, *GCV*) avulla. *Na.action*-käskyllä määritellään puuttuvien arvojen käsittelyä. Muut parametrit ovat oletusten mukaisia. (Wood 2009)

Gam-funktion *formula*-kohta sisältää varsinaisen mallin, joka on esimerkiksi muotoa

```
y <- s(x0)+s(x1)+x2+factor(x3)+s(x4,x5),
```

missä s on tasoittava funktio. Tasoittavana kantana käytetään oletuksen mukaisesti regressiosplinejä, tarkemmin "thin plate regression splines" ($TPRS$). Muita mahdollisia tasoittavia kantoja ovat muun muassa kuutiollinen regressiosplini (cubic regression spline, CRS) ja P-splinit (P-splines)

Gam-mallia voidaan havainnollistaa piirtämällä esimerkiksi *plot*-käskyn avulla, jolloin saadaan jokaisesta tasoitetusta muuttujasta esimerkiksi kuvan 4.1 vasemmanpuolen kaltainen kuvaaja. Tarkempaa tietoa mallista saadaan esimerkiksi *summary*- tai *anova*-käskyillä. Tulostus kertoo mallin tasoitetuille termeille muun muassa estimoidut efektiiviset vapausasteet (effective degrees of freedom, *edf*). Niiden avulla voidaan arvoida sovitetun mallin joustavuutta. (Wood 2006.) Alle yhden suuruiset efektiiviset vapausasteet eivät ole enää luotettavia, eikä alle 0.5 vapausasteita edes raportoida (Wood 2009). Lisätietoa kirjastosta *mgcv* ja *gam*-funktiosta on saatavilla esimerkiksi lähteistä Wood (2001, 2006 ja 2009).

5 Yhteenveto

Tutkielmassa mallinnetaan pienhiukkasaineistoa yleistetyillä additiivisilla malleilla. Mallintamiseen käytettyä aineistoa esitellään tutkielman alussa. Aineistosta tutkitaan tamperelaisten altistumista ulkoilman pienhiukkasille ja muille epäpuhtauksille erityisenä mielenkiinnonkohteena pienhiukkasten aiheuttamat terveysvaikutukset. Yleistetyillä additiivisilla malleilla luodaan malli, jolla voidaan mallintaa pienhiukkasten vaikutusta terveyteen. Vasteena ovat tehdyt hengityselin- ja sydändiagnoosit sekä kuolemat. Jo aiemmin kehitettyä analyysimallia sovelletaan ja kehitetään eteenpäin. Uutena asiana malliin tulee muuttujien väliset interaktiot, joiden vaikutusta tutkitaan. Tulosten tulkinnassa täytyy huomioida melko suuri puuttuvien arvojen osuus. Myös diagnoositietojen puuttuminen tietyiltä yksiköiltä saattaa aiheuttaa harhaa eli systemaattista virhettä tuloksiin.

Yleistettyjä additiivisia malleja muodostetaan aineistosta kymmenelle eri vasteelle. Vasteen ollessa kuolemien lukumäärät, jäävät selitysasteet pieniksi, eikä eri hiukkasmuuttujien ja viiveiden välillä ole paljon eroa vaikutuksessa selitysasteeseen. Hiukkasmuuttujat eivät ole yhdessäkään mallissa merkitseviä viiden prosentin riskitasolla ja taustamuuttujistakin ainoastaan aika tulee osassa malleissa merkitseväksi. Myöskään interaktitermien lisäys ei juurikaan vaikuta selitysasteisiin. Interaktiotermit eivät ole malleissa merkitseviä selittäjiä. Näillä malleilla ja aineistolla ei siis onnistuta mallintamaan hyvin tamperelaisten kuolemia. Tilanne voisi muuttua, jos kuolemat olisi luokiteltu esimerkiksi eri kuolinsyihin ja ikäluokkiin.

Hengityselinoireiden ollessa vasteena, sopii valittu malli hyvin aineistoon. Mallien selitysasteet ovat suurinpia työikäisillä, 74 prosenttia, ja pienimpiä vanhuksilla, 56 prosenttia. Selitysasteet ja hiukkasmuuttujien merkitsevyydet ovat parhaimmillaan summaviiveillä yksi ja kolme. Kaikissa ikäluokissa ajoneuvojen lukumäärä on tärkein selittäjä ja kasvattaa selvästi eniten selitysastetta. Pienhiukkasmuuttujien kohdalla on havaittavissa selvä ero lukumäärä- ja massapitoisuusmuuttujien välillä. Lukumääräpitoisuudella mitatut hiukkasmuuttujat ovat malleissa useammin merkitseviä kuin massapitoisuusmuuttujat. Myös niiden kuvaajat käyttäytyvät eritavoin näillä taustamuuttujilla. Typpidioksidi on myös monessa mallissa merkitsevä prediktori.

Mallintamista jatketaan interaktioiden tutkimisella. Malleihin lisätyt interaktiotermit kasvattavat selitysastetta keuhkodiagnoosien tapauksessa parhaimmillaan noin 3.5 prosenttiyksikköä. Tilastollisesti merkitseviä interaktioita muodostuu mallissa jo mukana olevien taustamuuttujien ja hiukkasmuuttujan

kesken, tyypillisimpänä ajan ja ajoneuvojen lukumäärän interaktiot toistensa ja lämpötilan sekä lämpötila eron kanssa.

Kun vasteena ovat sydänoireet, jäävät selitysasteet hieman hengityselinoiremallien selitysasteista. Nyt vanhimmassa ikäryhmässä selityasteet ovat suurimpia, 69 prosenttia. Parhaat selitysasteet ja merkitsevyydet saavutetaan 0 – 3 päivän summaviiveillä. Taustamuuttujista jälleen ajoneuvojen lukumäärä on kaikissa malleissa merkitsevä ja selittää ylivoimaisesti eniten koko mallissa. Lukumäärä- ja massapitoisuuksisilla pienhiukkasilla on tälläkin vasteella eroa. Nyt kuitenkin massapitoisuusmuuttuja on useammin mallissa tilastollisesti merkitsevä, kuin keuhkodiagnoosien tapauksessa.

Sydänoireiden mallintamista jatketaan interaktioiden tutkimisella. Malleihin lisätyt interaktiotermit kasvattavat selitysastetta sydändiagnoosien tapauksessa hieman vähemmän kuin hengityselindiagnoosien tapauksessa. Eri vasteilla tilastollisesti merkitsevät interaktiotermit ja niiden kuvaajat ovat hyvin samanlaisia kuin hengityselindiagnoosien tapauksessa.

Muissa tutkimuksissa on saatu samansuuntaisia tuloksia. Pääkaupunkiseudun hiukkaspitoisuuksia tarkastelleessa tutkimuksessa Halonen et al. (2008) havaitsivat, että hiukkaspitoisuuden nousu lisäsi hengityselindiagnooseja. Eri ikäisten oireisiin hiukkaset vaikuttivat eri tavoin ja erilaisella viiveellä. Ruuskanen et al. (2001) totesivat omassa tutkimuksessaan, että lukumäärä- ja massapitoisuusmuuttujat käyttäytyvät eritavoin ja niitä on syytä vertailla erikseen. Myös tämä tutkimus vahvistaa asiaa. Lukumäärä- ja massapitoisuusmuuttujat eivät korreloi hyvin keskenään ja saavat aikaan erilaisia tuloksia.

Johtopäätöksenä voidaan sanoa, että pienhiukkasten sekä typen oksidien ja hengityselinoireiden sekä sydänoireiden välillä on selvä yhteys. Hiukkaspitoisuuden ilmassa kasvaessa tehtyjen diagnoosien määrä kasvaa. Vaikutukset näkyvät yleensä yhden – kolmen päivän summaviiveellä. Lukumääräpitoisuusmuuttujat antavat selitysasteen ja merkitsevyyksien mielessä parempia tuloksia kuin massapitoisuusmuuttujat, varsinkin hengityselinoireiden tapauksessa. Erikokoisten ja eri tavoin mitattujen hiukkasten tai typen oksidien välillä ei kuitenkaan ole interaktiota. Eri hiukkasmuuttujien samanaikaiset korkeat pitoisuudet eivät aiheuta yhdysvaikutusta, joka olisi enemmän kuin muuttujien erillisten vaikutusten summa. Sen sijaan taustamuuttujien, varsinkin ajoneuvojen lukumäärän, ajan, lämpötilan sekä lämpötila eron välillä on interaktiota. Kuolemien lukumäärää ei onnistuta näillä malleilla mallintamaan hyvin.

Lopuksi haluan kiittää ohjaajaani dosentti Tapio Nummea kannustavasta ja kärsivällisestä ohjauksesta. Kiitos kaikille pienhiukkastutkimuksessa mukana oleville kiinnostavasta aineistosta sekä rakentavista kommenteista. Atelle suuri kiitos sekä teknisestä että henkisestä tuesta.

Lähdeluettelo

Aalto, P., Hämeri, K., Paatero, P., Kulmala, M., Bellander, T., Berglind, N., Bouso, L., Castano-Vinyals, G., Sunyer, J., Cattani, G., Marconi, A., Cyrys, J., Von Klot, S., Peters, A., Zetzsche, K., Lanki, T., Pekkanen, J., Nyberg, F., Sjovall, B. & Forastiere, F. (2005), "Aerosol particle number concentration measurements in five cities using TSI3022 condensation particle counter over a threeyear period during health effects of air pollution on susceptible subpopulations", *Journal of the Air & Waste Management Association* 55, 1064-1076.

Auvinen, A. (2008), "Väestön terveys II", *Luentomoniste*.

Elsilä, A. (2006), "Ilmanlaatu Tampereella", *Ympäristövalvonta 4/2006*, Tampereen kaupunki.

Euroopan unioni (2009), "Air pollution 2005", *Summaries of EU legislation: Environment* `http://europa.eu/legislation_summaries/environment/air_pollution/index_en.htm(6.7.2009)`.

Faraway, J. J. (2006), "Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models", *Text in statistical science*.

French, J. L. & Wand, M. (2004), "Generalized additive models for cancer mapping with incomplete covariates", *Biostatistics*, 5, 177–191.

Halonen, J., Lanki, T., Yli-Tuomi T., Kulmala, M., Tiittanen, P. & Pekkanen, J. (2008), "Urban air pollution and asthma and COPD hospital emergency room visits", *Thorax*.

Hastie, T. J. & Tibshirani, R. J. (1990), "Generalized additive models", *Monographs on statistics and applied probability*.

Heikkilä, T. (2005), "Tilastollinen tutkimus", *Edita*.

Heikkinen, J. (2005), "Yleistetyt lineaariset mallit", *Luentomoniste* `http://www.rni.helsinki.fi/~jmh/glm05/glm05.pdf`.

Hilli-Lukkarinen, M. (2009), "Pienhiukkasten lukumääräpitoisuus ja terveysvaikutukset (aktiivinen pinta-ala) vuosina 2005-2007 Tampereella, Tutkimuksen väliraportti", *Ympäristönsuojelun julkaisuja 3/2009*, Tampereen kaupunki.

Isotalo, J. (2009), "Yleistetyt lineaariset mallit I", *Luentomoniste*.

Jackman, S. (2004), "Generalized additive models", *Stanford university*.

Kariniemi, H. (2006), "Kaupunkiaerosolin hiukkaspitoisuudet Tampereella keväällä 2006", *Diplomityö TTY*.

Kettunen, J., Lanki, T., Tiittanen, P., Aalto, P., Koskentalo, T., Kulmala, M., Salomaa, V. & Pekkanen, J. (2007), "Associations of fine and ultrafine particulate air pollution with stroke mortality in an area of low air pollution levels", *Stroke*, 38, 918–922.

Kääriä, S. (2007), "Odotusarvon ja kovarianssirakenteen estimointi splinien avulla", *Pro gradu -tutkielma TAY*.

Leppälä, R. (2006), "Tilastollisten menetelmien perusteet II", *Luentomoniste*.

Niemi, J.,Väkevä, O., Kousa, A., Weckström, M., Julkunen, A., Myllynen, M. & Koskentalo, T. (2008), "Ilmanlaatu pääkaupunkiseudulla vuonna 2007", *YTV:n julkaisuja 8/2008*.

Nummi, T. (2008), "Ei-parametrinen regressio", *Luentomoniste*.

Pekkanen, J. (2004), "Kaupunki-ilman pienhiukkasten terveysvaikutukset", *Duodecim*, 120, 1645–1652.

Roca-Pardiñas, J. & Cadarso-Suárez, C. (2005), "Testing for interactions in generalized additive models: Application to SO_2 pollution data", *Statistics and Computing*, 15, 289–299.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003), "Semiparametric regression", *Cambridge series in statistical and probabilistic mathematics*.

Ruuskanen, J.; Tuch, Th.; Ten Brink, H.; Peters, A.; Khlystov, A.; Mirme, A.; Kos, G. P. A.; Brunekreef, B.; Wichmann, H. E.; Buzorius, G.; Vallius, M.; Kreyling, W. G. & Pekkanen, J. (2001), "Concentrations of ultrafine, fine and $PM_{2.5}$ particles in three European cities", *Atmospheric environment*, 35, 3729–3738.

Salonen, R. & Pennanen, A. (2006), "Pienhiukkasten vaikutus terveyteen: Tuloksia ja päätelmiä teknologiaohjelmasta FINE Pienhiukkaset - Teknologia, ympäristö ja terveys", *Tekes*.

Smith, M. & Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection".

Stone, C. (1985), "Additive regression and other nonparametric models", *Annals of statistics*, 13, 689–705.

Venables, W. N. & Ripley, B. D. (2002), "Modern applied statistics with S", *Statistics and computing*.

Wood, S. N. (2001), "mgcv: GAMs and generalized ridge regression for R", *R Help*, 1/2, 20–25 `http://cran.r-project.org/doc/Rnews/Rnews_2001-2.pdf`.

——— (2006), "Generalized Additive Models: An Introduction with R", *Text in Statistical Science*.

——— (2009), "Generalized additive models with integrated smoothness estimation", *R help for package mgcv*.

Liite 1: Hiukkasmuuttujien p-arvot

Hiukkasmuuttujien tilastollinen merkitsevyys malleissa eri vasteilla ja eri viiveillä ilman interaktiotermiä. Alle 0.01:n p-arvot on lihavoitu.

	Viive	$PM_{2,5}$	$N_{2,5}$	$N_{0,1}$	$N_{0,01}$	SA	NO	NO_2	PM_{10}
HE <15	0	0.44243	0.98972	0.98410	0.96994	0.99295	0.92023	0.79332	0.40356
	1	0.5228	*0.02371* *	*0.01890* *	*0.02193* *	0.3019	*0.0157* *	**0.00842** **	0.95790
	3	0.6420	*0.0495* *	0.0537	0.0557	0.166	*0.0423* *	0.0658	0.915
	5	0.8460	0.7612	0.6828	0.6667	0.745	0.5587	0.915	0.2339
	7	0.0502	0.5112	0.5070	0.4783	0.8109	0.9872	0.98669	0.2538
HE 15-64	0	0.0674	0.9377	0.886	0.893	0.18648	0.8794	0.688	0.7641
	1	0.2649	**0.003425** **	**0.00204** **	**0.00196** **	0.65581	0.1123	*0.016456* *	0.3432
	3	0.37298	**0.002819** **	**0.001655** **	**0.001761** **	0.173569	**0.008639** **	**0.000931** ***	0.38285
	5	0.83324	0.44601	0.27045	0.46405	0.489157	0.125114	**0.003960** **	*0.0375* *
	7	0.06265	0.49526	0.53114	0.51358	0.123250	0.605816	0.11738	0.14828
HE 65-74	0	**0.000413** ***	0.23855	0.37010	0.22400	0.3196	0.49686	0.12545	*0.02629* *
	1	0.2427	**0.001762** **	**0.001586** **	**0.002673** **	0.68037	0.05116	0.07047	0.6490
	3	0.9285	*0.029233* *	*0.023641* *	*0.022759* *	0.51135	0.12349	*0.03946* *	0.16521
	5	0.43109	0.99534	0.98835	0.98658	0.8395	0.84689	0.8516	0.34286
	7	0.1301	0.96893	0.98296	0.97675	0.7331	0.94329	0.40235	*0.011304* *
HE >74	0	*0.0381* *	0.9406	0.9804	0.9780	0.1841	0.2142	0.1941	0.0840
	1	0.0653	*0.01704* *	*0.01162* *	*0.01290* *	0.5062	0.6812	0.3212	0.4627
	3	0.3526	**0.00509** **	**0.0041** **	**0.00522** **	0.1616	0.1730	*0.01786* *	0.7491
	5	0.4505	0.3213	0.3013	0.3218	0.5551	0.9690	0.10802	0.7139
	7	0.8248	0.7946	0.7881	0.2058	0.9577	0.9781	0.3245	0.7592
HE kaikki	0	0.0748	0.940	0.886	0.888	0.240	0.872	0.693752	0.693
	1	0.208	**0.005556** **	**0.001435** **	**0.001429** **	0.785	0.0753	*0.012727* *	0.745
	3	0.5357	**0.00204** **	**0.00126** **	**0.00134** **	0.151374	*0.01053* *	**0.00320** **	0.422
	5	0.70008	0.632144	0.601944	0.633328	0.714859	0.265914	*0.035722* *	0.061576
	7	*0.04546* *	0.446607	0.479655	0.448003	0.31803	0.78060	0.31954	0.055
S <65	0	**0.00586** **	0.6748	0.6037	0.6012	0.12981	0.9522	0.7928	0.4247
	1	**0.00383** **	*0.0338* *	*0.0263* *	*0.0276* *	0.6636	0.2393	0.2441	0.4067
	3	0.0932	0.0698	0.052	0.053	0.928	*0.0361* *	*0.0471* *	0.2677
	5	0.0841	0.9832	0.9491	0.9486	0.2152	0.869	0.432274	0.0912
	7	**0.00662** **	0.999	0.981	0.406719	0.1682	0.599	0.78113	**0.000424** ***
S 65-74	0	*0.026438* *	0.535610	0.52930	0.561153	0.16648	0.94515	0.145547	0.183424
	1	0.104135	0.07784	0.057852	0.061185	0.98440	0.08001	0.0846	0.50937
	3	0.59150	*0.0164* *	*0.01041* *	*0.01288* *	0.6371	**0.00845** **	*0.0117* *	0.5391
	5	0.9831	0.7937	0.7311	0.8000	0.9993	0.1593	0.1019	0.190703
	7	*0.011910* *	0.52683	0.44650	0.28683	0.9650	0.1510	0.2160	*0.0497* *
S >74	0	0.0593	0.3169	0.2933	0.2830	0.265	0.591	0.979872	0.651
	1	*0.031611* *	**0.00923** **	**0.00594** **	**0.00537** **	0.9851	0.511	0.186	0.4971
	3	0.282718	**0.00225** **	**0.00148** **	**0.00139** **	0.466	*0.045* *	**0.00239** **	0.1510
	5	0.198262	0.176	0.152	0.152	0.995181	0.533731	0.0646	0.131283
	7	0.260488	0.369976	0.36234	0.341980	0.987018	0.657	0.102647	0.157177
S kaikki	0	*0.01101* *	0.40822	0.35725	0.34706	0.13090	0.82945	0.87390	0.46549
	1	**0.00854** **	*0.01449* *	**0.00861** **	**0.00893** **	0.95731	0.14178	0.2549	0.54604
	3	0.21994	**0.00691** **	**0.00445** **	**0.00455** **	0.6614	**0.00902** **	*0.014* *	0.3145
	5	0.3019	0.6004	0.551	0.5598	0.7749	0.3672	0.148	0.1219
	7	0.0640	0.6516	0.7625	0.7708	0.6691	0.355	0.336	*0.0425* *

Liite 2: Mallien selitysasteet

Mallien selitysasteet eri vasteilla, hiukkasmuuttujilla ja viiveillä.

	Viive	$PM_{2,5}$	$N_{2,5}$	$N_{0,1}$	$N_{0,01}$	SA	NO	NO_2	PM_{10}
HE <15	0	0.698	0.697	0.697	0.697	0.697	0.695	0.696	0.699
	1	0.7	0.708	0.709	0.708	0.699	0.708	0.702	0.697
	3	0.702	0.701	0.701	0.701	0.703	0.696	0.697	0.695
	5	0.697	0.697	0.697	0.697	0.696	0.697	0.699	0.706
	7	0.699	0.692	0.692	0.692	0.69	0.69	0.69	0.695
HE 15-64	0	0.737	0.738	0.738	0.738	0.738	0.736	0.736	0.737
	1	0.736	0.747	0.748	0.748	0.738	0.743	0.745	0.74
	3	0.742	0.746	0.746	0.747	0.745	0.745	0.752	0.741
	5	0.735	0.738	0.738	0.738	0.738	0.737	0.749	0.741
	7	0.733	0.734	0.734	0.734	0.737	0.73	0.736	0.727
HE 65-74	0	0.645	0.636	0.634	0.636	0.628	0.631	0.635	0.64
	1	0.631	0.651	0.651	0.65	0.633	0.641	0.642	0.635
	3	0.629	0.633	0.633	0.633	0.631	0.631	0.631	0.637
	5	0.636	0.634	0.634	0.634	0.634	0.641	0.641	0.641
	7	0.634	0.634	0.634	0.634	0.633	0.638	0.642	0.646
HE >74	0	0.569	0.563	0.562	0.562	0.572	0.565	0.564	0.568
	1	0.569	0.571	0.572	0.571	0.567	0.563	0.563	0.568
	3	0.565	0.568	0.569	0.569	0.565	0.557	0.56	0.562
	5	0.558	0.555	0.555	0.555	0.559	0.554	0.556	0.56
	7	0.553	0.553	0.553	0.544	0.554	0.549	0.548	0.552
HE kaikki	0	0.769	0.768	0.768	0.768	0.772	0.767	0.767	0.768
	1	0.768	0.782	0.78	0.78	0.768	0.775	0.775	0.769
	3	0.772	0.777	0.777	0.778	0.776	0.775	0.78	0.768
	5	0.767	0.768	0.768	0.768	0.77	0.77	0.778	0.77
	7	0.766	0.766	0.766	0.766	0.768	0.765	0.769	0.758
S <65	0	0.649	0.644	0.646	0.646	0.646	0.643	0.644	0.644
	1	0.645	0.644	0.645	0.645	0.641	0.641	0.64	0.64
	3	0.641	0.643	0.643	0.643	0.641	0.646	0.646	0.64
	5	0.626	0.629	0.629	0.629	0.63	0.624	0.63	0.628
	7	0.626	0.625	0.625	0.63	0.629	0.617	0.616	0.634
S 65-74	0	0.661	0.659	0.66	0.659	0.658	0.655	0.661	0.662
	1	0.655	0.663	0.664	0.664	0.655	0.659	0.666	0.658
	3	0.654	0.657	0.657	0.657	0.654	0.656	0.66	0.657
	5	0.644	0.645	0.645	0.644	0.644	0.646	0.651	0.654
	7	0.647	0.642	0.642	0.642	0.639	0.64	0.641	0.645
S >74	0	0.692	0.694	0.693	0.693	0.691	0.688	0.688	0.69
	1	0.692	0.696	0.696	0.696	0.689	0.691	0.692	0.689
	3	0.691	0.696	0.696	0.696	0.692	0.698	0.695	0.696
	5	0.679	0.68	0.68	0.68	0.679	0.68	0.681	0.686
	7	0.672	0.673	0.673	0.673	0.672	0.671	0.673	0.674
S kaikki	0	0.751	0.75	0.75	0.75	0.749	0.746	0.746	0.748
	1	0.747	0.752	0.752	0.752	0.745	0.749	0.747	0.745
	3	0.745	0.748	0.748	0.748	0.745	0.75	0.749	0.745
	5	0.734	0.735	0.735	0.735	0.735	0.733	0.736	0.737
	7	0.734	0.735	0.733	0.733	0.732	0.728	0.729	0.732
Kuolkai	0	0.0161	0.0221	0.0226	0.0225	0.0165	0.0138	0.0136	0.0132
	1	0.0136	0.0173	0.0174	0.0172	0.0142	0.00951	0.0136	0.0146
	3	0.0335	0.0136	0.0136	0.0135	0.0213	0.0109	0.00561	0.019
	5	0.0159	0.013	0.0129	0.0128	0.0199	0.000187	0.000188	0.0131
	7	0.0129	0.0169	0.0168	0.0166	0.0198	0.00449	0.00205	0.019

Liite 3: Interaktiotermit

Eri interaktiotermien lisääminen parhaan selitysasteen tuottaviin malleihin. Raportoitu uusi selitysaste sekä interaktiotermin merkitsevyys. Mallien suurin selitysaste on lihavoitu, samoin p-arvot<0.01.

			Lisätty interaktiotermi						
Alkuperäinen malli			aika,ajolkm	aika,temp5	aika, tempero	temp5, ajolkm	ajolkm, tempero	aika,x	ajolkm,x
y=keuhkai	R^2	0.782	0.796	0.807	0.809	**0.817**	0.785	0.78	0.78
x=viisu(n2.5,1)	p-arvo		**0.000513** ***	**5.05e-08** ***	**2.3e-08** ***	**< 2e-16** ***	**< 2e-16** ***	0.404741	0.0404798
y=keuh15	R^2	0.709	0.727	0.737	0.739	**0.747**	0.714	0.709	0.709
x=viisu(n0.1,1)	p-arvo		**0.000761** ***	**7.63e-07** ***	**6e-07** ***	**3.06e-13** ***	**< 2e-16** ***	0.96263	0.95777
y=keuh1565	R^2	0.752	0.759	**0.781**	**0.781**	0.773	0.756	0.768	0.754
x=viisu(no2,3)	p-arvo		0.999620	**2.48e-09** ***	**9.77e-09** ***	**< 2e-16** ***	**< 2e-16** ***	**1.84e-05** ***	**< 2e-16** ***
y=keuh6574	R^2	0.651	**0.678**	0.669	0.671	**0.678**	0.665	0.651	0.651
x=viisu(n2.5,1)	p-arvo		**0.000416** ***	0.360436	0.252668	**8.74e-08** ***	**< 2e-16** ***	0.930029	0.930127
y=keuh74	R^2	0.572	**0.59**	0.583	0.582	0.588	0.58	0.572	0.572
x=viisu(n0.1,1)	p-arvo		**0.00674** **	0.01757 *	0.01798 *	**< 2e-16** ***	**0.00644** **	0.90673	0.91297
y=sydkai	R^2	0.752	0.774	**0.777**	0.775	0.773	0.76	0.752	0.752
x=viisu(n0.01,1)	p-arvo		**6.9e-07** ***	**0.000865** ***	**3.8e-07** ***	**1.27e-14** ***	**0.000382** ***	0.94127	0.94833
y=syd65	R^2	0.649	0.665	0.664	0.665	**0.675**	0.664	0.665	0.664
x=pm2.5	p-arvo		**0.00395** **	**0.000299** ***	**0.000313** ***	**6.24e-11** ***	**< 2e-16** ***	**0.000253** ***	**< 2e-16** ***
y=syd6574	R^2	0.666	0.678	0.681	0.681	0.679	0.677	**0.682**	0.667
x=viisu(no2,1)	p-arvo		**0.000714** ***	**0.000257** ***	**1.07e-05** ***	**5.85e-07** ***	**< 2e-16** ***	0.0121 *	**< 2e-16** ***
y=syd74	R^2	0.698	**0.724**	0.715	0.715	0.711	0.699	0.712	0.701
x=viisu(no,3)	p-arvo		**1.68e-09** ***	0.0170 *	0.0384 *	**7.81e-08** ***	0.0400 *	**3.17e-05** ***	**< 2e-16** ***
y=kuolkm	R^2	0.0335	0.0335	0.0335	0.0335	**0.0342**	0.0337	0.0335	0.0337
x=viisu(pm2.5,3)	p-arvo		0.9616	0.9310	0.9301	0.5807	0.5701	0.9445	0.5704