

Hyvinvointialueiden stokastinen hierarkkinen klusterointi Kansallisen terveysindeksin pohjalta

Iina Manninen

Perustieteiden korkeakoulu

Kandidaatintyö
Espoo 6.3.2026

Vastuuopettaja

Prof. Ahti Salo

Työn ohjaaja

DI Leevi Olander



**Aalto-yliopisto
Perustieteiden
korkeakoulu**

Copyright © 2026 Iina Manninen

The document can be stored and made available to the public on the open internet pages of Aalto University.
All other rights are reserved.



Tekijä Iina Manninen

Työn nimi Hyvinvointialueiden stokastinen hierarkkinen klusterointi Kansallisen terveysindeksin pohjalta

Koulutusohjelma Teknistieteellinen kandidaattiohjelma

Pääaine Matematiikka ja systeemitieteet

Pääaineen koodi SCI3029

Vastuopettaja Prof. Ahti Salo

Työn ohjaaja DI Leevi Olander

Päivämäärä 6.3.2026

Sivumäärä 33+6

Kieli Suomi

Tiivistelmä

Sairastavuus ja työkyvyttömyys kuormittavat sosiaali- ja terveydenhuoltoa sekä heikentävät yksilön mahdollisuuksia osallistua yhteiskunnassa. Suomalaisen sosiaali- ja terveydenhuoltojärjestelmän tehtäviin kuuluu tukea väestön hyvinvointia ja toimintakykyä sekä kaventaa terveyseroja. Vuonna 2023 toimeenpannun sosiaali-, terveys- ja pelastuspalveluiden uudistuksen päämääränä on edistää näiden tavoitteiden toteutumista väestön ikääntyessä ja keskittyessä kaupunkeihin. Uudistuksessa sosiaali- ja terveydenhuollon palvelut siirtyivät kunnilta hyvinvointialueille. Väestön terveyden tilan kehitystä voidaan seurata hyvinvointialuekohtaisesti Terveyden ja hyvinvoinnin laitoksen sekä Kansaneläkelaitoksen koostamasta Kansallisesta terveysindeksistä.

Tässä työssä tutkitaan alueellisten terveyserojen tilaa sosiaali- ja terveydenhuollon uudistuksen toimeenpanon aikaan. Kokoavan hierarkkisen klusterianalyysin avulla tarkastellaan hyvinvointialueiden klusterointia pohjautuen Kansallisen terveysindeksin sairastavuuden ja työkyvyttömyyden osaindekseihin vuosilta 2021–2023. Euklidiseen etäisyyteen ja keskiarvolinkkifunktioon perustuvan hierarkkisen klusteroinnin ohella esitetään euklidisten etäisyyksien käänteislukuihin perustuviin todennäköisyyksiin pohjautuva stokastinen klusterointialgoritmi, jonka simulointi antaa tietoa deterministisellä menetelmällä tuotetun klusterointiratkaisun herkkyydestä.

Klusteroinnin tuloksena erottuu vähäisen sairastavuuden ja työkyvyttömyyden klusteri, joka kattaa Uudenmaan hyvinvointialueet ja Pohjanmaan hyvinvointialueen sekä korkean sairastavuuden ja työkyvyttömyyden klusteri Itä- ja Pohjois-Suomessa. Lounais-Suomen, Hämeen ja Keski-Suomen alueet sijoittuvat stokastisen klusteroinnin tulosten perustella piirteiltään Etelä-Suomen ja Itä- ja Pohjois-Suomen klusterien välimaastoon. Havainnot siitä, että Etelä-Suomessa on vähemmän työkyvyttömyyttä ja harvaan asutuilla alueilla on enemmän sairastavuutta ja työkyvyttömyyttä ovat linjassa aikaisemman alueellisia terveyseroja tarkastelleen tutkimuksen kanssa.

Avainsanat Kansallinen terveysindeksi, hyvinvointialueet, sote-uudistus, terveyserot, sosiaali- ja terveydenhuoltojärjestelmä, sairastavuus, työkyky, hierarkkinen klusterointi, Monte Carlo -simulointi

Author Iina Manninen

Title Stochastic Hierarchical Clustering of Finnish Wellbeing Services Counties
Based on the National Health Index

Degree programme Bachelor's Programme in Science and Technology

Major Mathematics and Systems Sciences

Code of major SCI3029

Teacher in charge Prof. Ahti Salo

Advisor M.Sc. (Tech.) Leevi Olander

Date 6.3.2026

Number of pages 33+6

Language Finnish

Abstract

Morbidity and work disability place a burden on social and health care services and reduce individuals' opportunities to participate in society. The responsibilities of the Finnish social and health care system include supporting the wellbeing of the population and reducing health differences. The aim of the health and social services reform implemented in 2023 is to promote the achievement of these objectives as the population ages and concentrates in urban areas. In the reform, the responsibility for organizing social and health care services was transferred from municipalities to wellbeing services counties. The development of the population health can be studied by region using the National Health Index compiled by the Finnish Institute for Health and Welfare and the Social Insurance Institution of Finland.

This thesis examines the state of regional health differences at the time of the implementation of the social and health services reform. Agglomerative hierarchical clustering is used to study the clustering of the wellbeing services counties based on the morbidity and work disability sub-indices of the National Health Index in 2021–2023. In addition to hierarchical clustering based on the Euclidean distance and the average linkage function, a stochastic clustering algorithm based on probabilities derived from the reciprocals of the Euclidean distances is presented. Simulation of the stochastic clustering algorithm provides information on the sensitivity of the deterministic clustering solution.

As a result, a cluster of low morbidity and work disability covering the wellbeing services counties in Uusimaa and the county of Ostrobothnia stands out, as well as a cluster of high morbidity and work disability in Eastern and Northern Finland. Based on the results of stochastic clustering, the regions of Southwest Finland, Häme, and Central Finland resemble both the cluster of Southern Finland and the cluster of Eastern and Northern Finland. The findings of lower disability in Southern Finland and higher morbidity and work disability rates in rural areas are consistent with previous research.

Keywords National Health Index, wellbeing services counties, social and health services reform, health differences, social and health care system, morbidity, work ability, hierarchical clustering, Monte Carlo simulation

Sisällys

Tiivistelmä	3
Tiivistelmä (englanniksi)	4
Sisällys	5
1 Johdanto	6
2 Työn taustaa	7
2.1 Sairastavuus ja työkyky	7
2.2 Sote-uudistus ja uudet hyvinvointialueet	8
2.3 Kansallinen terveysindeksi	9
2.4 Klusterointi	11
2.4.1 Kokoava hierarkkinen klusterointi	11
2.5 Monte Carlo -simulointi	13
3 Tutkimusaineisto ja -menetelmät	13
3.1 Klusteroinnin tavoitteet	13
3.2 Klusteroitava aineisto	14
3.3 Aineiston valmistelu	14
3.4 Deterministinen kokoava hierarkkinen klusterointi	15
3.5 Stokastinen kokoava hierarkkinen klusterointi	16
4 Tulokset	17
4.1 Klusterointi sairastavuuden osaindeksien perusteella	19
4.2 Klusterointi työkyvyttömyyden osaindeksien perusteella	23
4.3 Sairastavuus ja työkyvyttömyys hyvinvointialueilla	26
5 Pohdinta	28
A Stokastisen klusteroinnin visualisoinnit	34
B Stokastisen klusteroinnin Python-koodi	38
C Kirjaus tekoälyn hyödyntämisestä	39

1 Johdanto

Sairastavuus ja työkyvyttömyys aiheuttavat inhimillisiä ja taloudellisia menetyksiä yksilöille, yhteisöille ja yhteiskunnalle. Sairaudesta aiheutuvat poissaolot ja työkyvyttömyys heikentävät yksilön ja perheiden toimeentuloa sekä sosiaalisia suhteita (Bryngelson, 2009), mikä voi puolestaan entisestään alentaa toimintakykyä (Elovainio & Komulainen, 2023). Työnantajille sairauspoissaolot ja työkyvyttömyyseläkkeet aiheuttavat sekä suoria kustannuksia (ETK, 2021) että tuottavuuden menetyksiä. Yhteiskunnallisia haasteita ovat muun muassa työpanoksen ja verotulojen menetys, maksettavat sairauspäivärahat, työttömyysetuudet ja työkyvyttömyyseläkkeet (Kela, 2026) sekä kasvavat sosiaali- ja terveydenhuollon menot. Lisäksi sairastavuuden ja työkyvyttömyyden myötä yhteisöt ja yhteiskunta voivat menettää inhimillistä luovuutta.

Väestön ikääntyessä sosiaali- ja terveydenhuollon menojen ennakoita kasvavan (Tynkkynen ym., 2025). Samalla työssäkäyvien veronmaksajien suhteellinen osuus on aikaisempaa pienempi. Väestörakenteessa on myös merkittäviä maantieteellisiä eroja (Kestilä, 2025). Ikärakenteen muutoksen vaikutuksiin varautumiseksi sosiaali-, terveys- ja pelastuspalveluiden järjestämisvastuu siirrettiin laajassa hallinnollisessa uudistuksessa kunnilta hyvinvointialueille vuoden 2023 alussa (STM, 2025). Väestömäärältään riittävän suurten alueiden ajatellaan pystyvän tuottamaan sosiaali- ja terveyspalvelut laadukkaasti ja yhdenvertaisesti. Palveluiden saatavuuden yhdenmukaistamisen tavoitteena on myös alueellisten terveys- ja hyvinvointierojen loiventaminen (Tynkkynen ym., 2025). Yhtenä indikaattorina väestön terveyden kehityksestä sekä sote-uudistuksen tavoitteiden toteutumisesta on Terveyden ja hyvinvoinnin laitoksen sekä Kansaneläkelaitoksen yhteisesti tuottama Kansallinen terveysindeksi (THL, 2025a), joka sisältää tietoa sairastavuudesta ja työkyvyttömyydestä aluekohtaisesti.

Tässä työssä tarkastellaan hyvinvointialueiden välisiä eroja ja yhtäläisyyksiä soveltamalla kokoavaa hierarkkista klusterointialgoritmia hyvinvointialueiden tietoihin Kansallisen terveysindeksin sairastavuuden sekä työkyvyttömyyden osaindekseistä vuosilta 2021–2023. Lisäksi vertaillaan deterministisen ja stokastisen kokoavan hierarkkisen klusteroinnin tuloksia. Stokastisessa klusteroinnissa alueet jaetaan klustereihin euklidisten klusterietäisyyksien käänteislukuihin pohjautuviin todennäköisyyksiin perustuen ja klusterointia toistetaan Monte Carlo -simuloinnin avulla. Tämän jälkeen deterministisen klusterointiratkaisun herkkyyttä tarkastellaan stokastisella klusteroinnilla. Klusterointi sairastavuuden ja työkyvyttömyyden perusteella tarjoaa rajallisen näkökulman alueiden välisten erojen tarkasteluun, eikä se ota huomioon esimerkiksi maantieteellisiä piirteitä tai alueiden sisäistä vaihtelua. Tulokset kuitenkin luotaavat kokonaiskuvaa terveyserojen tilasta sosiaali- ja terveydenhuollon uudistuksen (sote-uudistuksen) toimeenpanon aikaan ja voivat tarjota mielekkäitä lähtökohtia tarkastella klusterien mahdollisen muodostumisen taustalla olevia tekijöitä.

Työn luvussa 2 taustoitetaan sairastavuuden ja työkyvyn näkökulmia, sote-uudistusta, Kansallista terveysindeksiä sekä klusterointimenetelmää. Aineiston esikäsittelyprosessi sekä työssä sovellettavat klusterointimenetelmät esitellään tarkemmin luvussa 3. Klusteroinnin tulokset kuvataan luvussa 4. Pohdinta ja nostot jatkotutki-

muksen näkökulmista sisältyvät lukuun 5.

2 Työn taustaa

2.1 Sairastavuus ja työkyky

Sairastavuus kuvaa sairauksien yleisyyttä väestössä ja se voidaan ilmaista ilmaantuvuutena eli uusien tapausten määränä tai esiintyvyytenä, jolloin mukaan lasketaan kaikki henkilöt, jotka ovat saaneet hoitoa tai etuuksia tai joilla on diagnoosiin perustuva lääkekorvausoikeus tietyllä ajanjaksolla ([Lindell ym., 2025](#)). Mikäli yksi tai useampi diagnosoitu sairaus estää työnteon, on mahdollista hakea sairauspäivärahaa tai työkyvyttömyyseläkettä ([Kela, 2025c](#)). Työkyvyttömyyseläke voi olla määräaikainen kuntoutustuki tai jatkuva työkyvyttömyyseläke ([Kela, 2026](#)). Osa-kuntoutustuki ja -työkyvyttömyyseläke mahdollistavat työskentelyn osa-aikaisesti ([Kela, 2026](#)). Yleensä työkyvyttömyyseläkettä edeltää vuoden kestänyt sairauspäiväraajakso ([Kela, 2025c](#)). Työkyvyttömyyseläkkeen sijaan voidaan myös tarjota ammatillista kuntoutusta, joka tukee työhön paluuta ja työllistymistä esimerkiksi koulutus- ja työkokeilujen avulla ([Kela, 2025b](#)).

Työkyky rakentuu ihmisen voimavarojen ja kuormitustekijöiden tasapainosta ja on sidoksissa työtehtäviin ([TTL, 2025](#)). Sairauspäivärahaa, kuntoutusta tai työkyvyttömyyseläkettä haettaessa arvio työkyvystä perustuu ensisijaisesti lääketieteellisiin diagnooseihin. Laaja-alaisemmin työkyvyn voidaan käsittää rakentuvan terveyden ohella muun muassa osaamisesta, mielekkäistä työtehtävistä, lähiyhteisön ja -ympäristön sekä yhteiskunnallisten tilanteiden vaikutuksista ([TTL, 2025](#)). Yksilöiden työkyky heijastelee myös yhteisön tilaa.

Työkyvyn edistämisessä merkittävä rooli on toistaiseksi sote-uudistuksen ulkopuolelle jäänellä työterveyshuollolla, jonka vahvempi kytkeminen muuhun sote-järjestelmään voisi vähentää hoidon pirstaloitumista ([Tynkkynen ym., 2025](#)). Kaksi osittain päällekkäistä järjestelmää ja erityisesti työterveyshuollon lakisääteisten palveluiden ulkopuolelle jäävät vapaavalintaiset palvelut herättävät myös keskustelua yhdenvertaisuudesta ([Koponen & Tynkkynen, 2023](#)). Koska työikäinen työssäkäyvä väestö on työnantajien kustantaman työterveyshuollon piirissä, hyvinvointialueiden asiakaskunta on etenkin peruspalveluissa verrattain pienituloista ([Saari, 2024](#)). Samalla palvelutarve on yleisesti suurempi niillä, jotka käyttävät vain julkisen sektorin palveluita ([Koponen & Tynkkynen, 2023](#)). Nämä kaksi erillistä järjestelmää kilpailevat samasta työvoimasta. Lääkärikato julkiselta puolelta voi viivästyttää vakavien sairauksien hoitoa ([Koponen & Tynkkynen, 2023](#)). [Koponen & Tynkkynen \(2023\)](#) painottavat, että jatkossa olisi hyvä kehittää tapoja, joilla työterveyshuollon osaamista voitaisiin hyödyntää laajemmin koko työikäisen väestön toimintakyvyn edistämisessä.

Suomessa työkyvyttömyyseläkkeelle siirtyminen on 2000-luvun alusta vähentynyt. Tuki- ja liikuntaelinten sairaudet sekä mielenterveyden häiriöt vastaavat kumpikin noin yhdestä kolmasosasta työkyvyttömyyseläkkeiden taustalla vaikuttavista diagnooseista ([ETK, 2025](#)). Ikä- ja ammattiryhmien välillä on kuitenkin huomattavia eroja sekä työkyvyttömyyseläkkeelle siirtymisen yleisyydessä että taustalla vaikuttavissa

sairauksissa. [Laaksonen ym. \(2025\)](#) ovat nostaneet esille maksettujen työkyvyttömyyseläkkeiden alueelliset erot, joiden on todettu selittyvän ensisijaisesti yksilötason muuttujilla, mutta myös alueellisten tekijöiden, kuten alueen työllisyystilanteen, on havaittu olevan yhteydessä työkyvyttömyysriskiin. Myös [Virtanen ym. \(2022\)](#), jotka ovat tutkineet työkyvyttömyyseläkkeiden ja sosioekonomisen huono-osaisuuden alueellista jakautumista Ruotsin kuntatasolla, ovat havainneet alueellisia eroja. Heidän tulostensa mukaan työkyvyttömyyseläkkeitä ja useilla mittareilla todettua huono-osaisuutta esiintyy samoilla maantieteellisillä alueilla. Lisäksi työkyvyttömyyttä on vähiten pääkaupungin lähistöllä.

2.2 Sote-uudistus ja uudet hyvinvointialueet

Vuoden 2023 alussa sosiaali- ja terveydenhuollon sekä pelastustoimen palveluiden uudistuksen myötä Suomessa aloitti toimintansa 21 hyvinvointialuetta ([STM, 2025](#)). Uudistuksessa palvelut siirtyivät kunnilta hyvinvointialueille, joilla on kunnista ja valtiosta erillinen hallinto. Valtio kuitenkin vastaa hyvinvointialueiden rahoituksesta ([Tynkkynen ym., 2025](#)). Hyvinvointialueiden ohella Helsingin kaupunki ja Ahvenanmaa vastaavat lakisääteisistä sosiaali- ja terveystalvveluista omilla alueillaan. Helsingin kaupungin ja Uudenmaan hyvinvointialueiden erikoissairaanhoido on HUS-yhtymän vastuulla. Kukin hyvinvointialue kuuluu yhteen viidestä yhteistyöalueesta, jotka muodostuvat yliopistollisten sairaaloiden ympärille. Laajan rakenteellisen uudistuksen tavoitteena on ollut muun muassa varmistaa yhdenvertainen palveluiden saatavuus koko maassa sekä hidastaa kustannusten kasvua samalla, kun väestön ikärakenne muuttuu ([Jalonen, 2020](#)).

Sosiaali- ja terveydenhuollon uudistus ei ole vain yhden toimialan rakenteellinen muutos, vaan sote-järjestelmä on vahvasti kytkeytynyt muihin yhteiskunnan toimintoihin. Lisäksi järjestelmän sisällä on runsaasti toimijoita, sillä hyvinvointialueiden ohella palvelukentällä toimii myös yksityisiä yrityksiä ja järjestöjä ([STM, 2026](#)). Palveluiden tehokas järjestäminen vaatii sujuvaa yhteistyötä näiden eri toimijoiden välillä. Myös ajassa ilmenevät muutokset, kuten ikääntyminen, digitalisaatio, ilmastomuutos sekä talouden ja turvallisuuden toimintaympäristön muutokset tuovat uusia haasteita ja mahdollisuuksia sosiaali- ja terveydenhuollon sektorilla ([Jalonen, 2020](#)). Ikääntyminen todennäköisesti kasvattaa palveluiden tarvetta, mutta työikäisen väestön vähentyessä sote-järjestelmän haasteena ovat sekä osaaapula että kapeneva veropohja, josta palvelut rahoitetaan. Työvoimapulaan voidaan saada helpotusta maahanmuutosta. Myös digitaaliset ratkaisut voivat lieventää kestävyysvajetta vähentämällä sote-henkilöstön työkuormaa sekä mahdollistamalla palveluista saatavan tiedon hyödyntämisen johtamisessa, ennakoivassa hyvinvoinnin tukemisessa ja palveluiden kehittämisessä ([STM, 2023](#)). Toisaalta on hyvä kiinnittää huomiota muun muassa digitaalisten palveluiden saavutettavuuteen, inhimillisten kohtaamisten merkitykseen terveyden edistämisessä sekä tietojen hyödyntämisen eettisiin näkökulmiin.

Pitkään hahmoteltua uudistusta ja hyvinvointialueita on kritisoitu erityisesti kustannusvaikutustensa osalta. [Rimpelä & Laihonon \(2025\)](#) ovat tarkastelleet sote-uudistuksesta esitettyjä tilannekuvia viranomaisjulkaisuissa sekä viestintäaineistoissa vuodesta 2020 vuoteen 2024. Analyysin keskeisenä tuloksena on se, että joukkoviestin-

tävälineiden tilannekuvat keskittyvät uudistuksen ja hyvinvointialueiden ongelmiin, valtionohjaus vahvistuu, yhtenäinen tilannetietoisuus puuttuu ja kustannusten kasvu yllättää. Vuonna 2024 julkaistun ulkoisen arvionnin ([Owal Group Oy, 2024](#)) mukaan sisäisesti sekä toisiinsa nähden väestörakenteeltaan ja -tiheydeltään erilaisten alueiden tuen tarpeisiin olisi pitänyt kiinnittää uudistuksen valmisteluvaiheessa enemmän huomiota. Alueiden eriytymisen välttäminen sekä rahoitusmallin toimivuus ovat kehittämiskohteita myös hyvinvointialueuudistuksen väliarvioinnissa ([Haveri ym., 2025](#)), jossa alueiden eriytymiskehityksen taustalla tunnistetaan vaihtelevat väestömäärät ja ikärakenteet, sairastavuuserot, kaupungistumiskehitys sekä aikaisemmat aluekohtaiset uudistukset. [Haveri ym. \(2025\)](#) esittävät uudistamissuosituksiksi hyvinvointialueiden määrän arviointia ja alueiden mahdollista yhdistämistä sekä alueiden verotusoikeutta. Verotusoikeuden taustalla on ajatus, että kun itsehallinnollisten alueiden rahoitus pohjautuu pääosin sairastavuuden perusteella jaettavaan valtion rahoitukseen, alueilla ei välttämättä ole riittäviä kannustimia palveluiden tehokkaaseen toteuttamiseen ja ennaltaehkäisevään terveyden edistämiseen ([Kortelainen ym., 2021](#)). Uuden veron käyttöönotto voisi kuitenkin vaatia kokonaisverotuksen tarkastelua ja alueiden erilaisten taloudellisten tilanteiden huomioimista.

Uudistumista voi tarkastella päämäärän sijaan myös sote-järjestelmän ominaisuutena, jolloin uudistuksen voi nähdä vasta aluillaan olevana jatkuvaa hienosäätöä vaativana prosessina ([Jalonen, 2020](#)). Päätöksenteon tasolla sote-järjestelmän kehittämisessä on kuitenkin yhdisteltävä erilaisia ideologisia näkemyksiä keinoista tarjota sosiaali- ja terveydenhuollon palvelut tehokkaasti ja oikeudenmukaisesti. Sote-uudistuksen valmistelua sekä hyvinvointialueiden ensimmäisiä toimintavuosia ovat leimanneet myös toimintaympäristön muutokset, pandemia sekä turvallisuusympäristön ja taloustilanteen epävarmuus. Ulkoiset muutosvoimat voivatkin viitoittaa tietä ajatteluun, jossa sote-järjestelmästä ei pyritä luomaan kerralla valmista, vaan uudistamista pilkotaan pienempiin osiin ([Jalonen, 2020](#)). Hallinnollisten muutosten ohella tulevaisuusorientoitunutta kehittämistä on sote-järjestelmää koettelevien tekijöiden juurisyihin pureutuminen esimerkiksi ennaltaehkäisevällä terveyden ja toimintakyvyn edistämisellä ([Haveri ym., 2025](#)). Pidemmän aikavälin jaettujen tavoitteiden kirkastaminen ja ideologisten kompromissien haasteiden tunnistaminen sekä pienet uudistukset ja niiden vaikutusten seuraaminen koko sote-järjestelmän pitkän aikavälin näkökulmasta voivat vahvistaa järjestelmän ennakoivaa kehittämistä ongelmiin reagoimisen sijaan. Uudistuksen suunnan selkeyttäminen vahvistaa myös kansalaisten luottamusta sosiaali- ja terveydenhuoltojärjestelmään ([Tynkkynen ym., 2025](#)).

2.3 Kansallinen terveysindeksi

Sote-uudistuksen yhtenä keskeisenä tavoitteena on kaventaa väestön terveyseroja ([STM, 2025](#); [Terveydenhuoltolaki 1326/2010, § 2, 2010](#); [Laki sosiaali- ja terveydenhuollon järjestämisestä 612/2021, § 1, 2021](#)). Tavoitteen toteutumista voidaan seurata tarkastelemalla Kansallisen terveysindeksin sairastavuuden ja työkyvyttömyyden muutoksia sekä alueellisia eroja hyvinvointialue- ja kunta- ja yhteistyöaluekohtaisesti ([THL, 2025a](#)). Indeksillä hyödyntää usean rekisterinpitäjän, Eläketurvakeskuksen,

Kansaneläkelaitoksen, Terveyden ja hyvinvoinnin laitoksen, Tilastokeskuksen sekä Syöpärekisterin tietoja. Indeksiluvut kuvaavat sairastavuuden ja työkyvyttömyyden alueellista yleisyyttä verrattuna yleisyyteen koko maan väestössä. Satunnaisvaihtelun vaikutuksen vähentämiseksi indeksiluvut esitetään kolmevuosittaisina keskiarvoina. Indeksillä voidaan arvioida väestön terveydentilaa, sosiaali- ja terveystalouden sekä sosiaaliturvaetuuksien tarvetta sekä sosiaaliturvan vaikuttavuutta. Lisäksi se mahdollistaa resurssien ennakoivaa kohdentamista.

Indekseistä on mahdollista tarkastella ikävakioituja ja ikävakioimattomia aluekohtaisia lukuja. Useat tilastossa huomioitavat sairaudet ovat yleisempiä iäkkäillä, mutta ikärakenteiden erojen vaikutus voidaan häivyttää ikävakioinnilla. Ikävakioidut luvut ovat mielekkäitä, kun vertaillaan alueita, joiden ikäjakaumat poikkeavat toisistaan tai tarkastellaan ikärakenteen muutoksesta riippumatonta kehitystä. Ikävakioimattomat luvut puolestaan soveltuvat tilanteisiin, joissa halutaan havainnoida alueen sairastavuuden ja työkyvyttömyyden todellista vaikutusta palvelutarpeeseen.

Kansallisen terveysindeksin sairastavuusindeksiin kuuluvat sairastavuuden kokonaisindeksin ohella kymmenen sairausryhmäkohtaista osaindeksiä. Sairastavuuden kokonaisindeksin laskennassa hyödynnetään näitä osaindeksejä, joilla on erilaiset painotukset sairausryhmien kuolleisuus-, työkyvyttömyys- elämänlaatu-, sekä sosiaali- ja terveystaloudellisuuden kustannusvaikutusten mukaan. Huomioitavat sairausryhmät ovat syöpäsairaudet, sepelvaltimotauti, aivoverisuonitaudit, tuki- ja liikuntaelinten sairaudet, mielenterveyden häiriöt, tapaturmat, muistisairaudet, keuhkosairaudet, diabetes ja alkoholisairaudet. Osaindeksit kuvaavat vaihtelevasti sairausryhmän yleisyyttä joko ilmaantuvuuden tai esiintyvyyden perusteella. Kansallisen terveysindeksin tilastoraportin 2021–2023 ([Lindell ym., 2025](#)) mukaan ikävakioitu sairastavuusindeksi on korkein Pohjois-Savossa, Pohjois-Pohjanmaalla, Pohjois-Karjalassa sekä Kymenlaaksossa. Sairausryhmäkohtaisesti vakavat mielenterveyden häiriöt, tuki- ja liikuntaelinten sairaudet sekä alkoholisairastavuus ovat näillä alueilla yleisempiä verrattuna koko maan tasoon. Matalimman sairastavuuden alueet ovat Pohjanmaa, Uudenmaan hyvinvointialueet sekä Helsinki. Koko maan tasolla diabetes sekä krooniset keuhkosairaudet yleistyvät. Ikävakioimattomassa indeksissä näkyy myös muistisairauksien yleistymistä.

Työkyvyttömyysindeksi puolestaan sisältää työkyvyttömyysindeksin ja sen osaindeksit, jotka kuvaavat tietynä ajanjaksona maksettua työkyvyttömyyseläkkeitä, yli kolmen kuukauden mittaisia sairauspäiväraajaksoja sekä ammatillista kuntoutusta. Myös näillä osaindekseillä on erilaiset painotukset yleisindeksissä. [Lindell ym. \(2025\)](#) kuvaavat raportissaan työkyvyttömyysindeksin olevan korkein Kainuussa ja Pohjois-Savossa. Ammatillisessa kuntoutuksessa on raportin mukaan maksettua työkyvyttömyyseläkkeitä ja sairauspäivärahoja enemmän aluekohtaista vaihtelua, mikä kertoo todennäköisesti hoidon saatavuuden alueellisista eroista.

Sairastavuuden ja työkyvyttömyyden alueellisten erojen taustalla voi ikärakenteiden ohella olla koulutukseen, työllisyyteen, taloudelliseen tilanteeseen, elintapoihin sekä hoitoon pääsyyn ja ennaltaehkäiseviin toimiin liittyviä tekijöitä. Toisaalta palveluiden saatavuus voi myös näyttäytyä korkeampana sairastavuutena, kun sairauksia havaitaan ja hoidetaan tehokkaasti. Hyvinvointialueiden rahoitusmallin perustuminen diagnoosi- ja sairastavuuskirjauksiin voi myös näkyä luvuissa ([Kortelainen ym.,](#)

2021; Haveri ym., 2025). Ahvenanmaan osalta luvuissa voi olla puutteita johtuen Ruotsissa saaduista hoidoista (Lindell ym., 2025).

2.4 Klusterointi

Luokittelu on ihmiselle luontainen tapa jäsentää maailmaa ja luoda järjestystä suureen tietomäärään. Vaikka luokittelu voi toisinaan olla keinotekoisia, sen avulla voidaan pyrkiä ymmärtämään tarkasteltavan kohteen luonnetta ja siihen vaikuttavia taustatekijöitä sekä ennakoimaan kohteen tulevaisuutta. Lisäksi erilaiset luokittelut ovat keskeisiä useimmilla tieteenaloilla. Usein luokittelun täsmällistä oikeellisuutta kiinnostavampaa on sen onnistuneisuus aineiston luonteen valottamisessa. (Everitt ym., 2011)

Klusterointi on numeerinen menetelmä, jonka avulla aineisto voidaan luokitella valittujen ominaisuuksien perusteella klustereihin eli luokkiin siten, että saman klusterin havainnoilla on mahdollisimman samankaltaiset piirteet, ja eri klusterit erottuvat ominaisuuksiltaan mahdollisimman selkeästi (Hair ym., 2019; Everitt ym., 2011). Kun tarkasteltava aineisto sisältää $n \in \mathbb{N}$ kappaletta heterogeenisiksi oletettuja havaintoja $x_1, \dots, x_n$, joista kullakin on $m \in \mathbb{N}$ arvoa, klusteroinnin tavoite on luokitella havainnot $g \in \mathbb{N}$ kappaleeseen keskenään mahdollisimman homogeenisia luokkia. Luokkien lukumäärää g ei välttämättä oleteta etukäteen, mutta se on yleensä pienempi kuin havaintojen määrä n (Mardia ym., 1979).

Koneoppimisen näkökulmasta klusterointi kuuluu ohjaamattoman oppimisen menetelmiin, sillä aineistosta pyritään havaitsemaan mahdollisia luokkia, jotka eivät ole valmiiksi tunnettuja (Everitt ym., 2011). Klusterointi voidaan näin ollen hahmottaa eksploratiivisena menetelmänä (Hair ym., 2019), joka mahdollistaa uusien rakenteiden havaitsemisen aineistosta. Yleisimpien klusterointialgoritmien lopputuloksena jokaisella aineiston havainnolla on vain yksi klusterijäsenyys ja jokainen havainto kuuluu johonkin klusteriin (Everitt ym., 2011). Jos klusterit eivät kuitenkaan erotu toisistaan kovin selkeästi, useampia klusterijäsenyyksiä salliva ratkaisu voi olla sopeva. Klusterointiratkaisua tulkittaessa voidaan päätyä myös johtopäätökseen, että yksikään aineiston klusterointi ei ole mielekäs (Everitt ym., 2011), vaan aineistosta havaitaan luokkia vain algoritmin toimintaperiaatteen vuoksi.

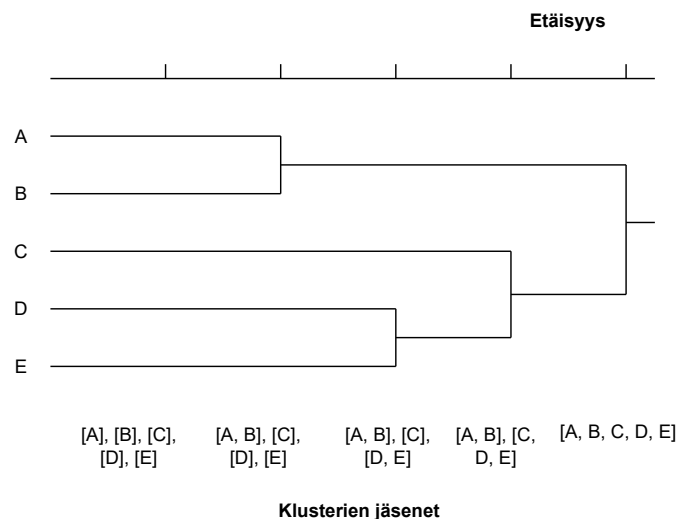
Klusterointiin liittyy sarja päätöksiä, joissa tieto klusteroitavan aineiston ominaisuuksista on keskeinen. Päätökset koskevat klusteroinnin tavoitteiden hahmottelua, muuttujien valintaa ja muuttujien arvojen mahdollista standardisointia, samankaltaisuutta mittaavan suureen valintaa, oletusten tarkastelua, klusterointialgoritmin valintaa sekä lopputulosten tulkintaa ja arviointia. (Hair ym., 2019)

2.4.1 Kokoava hierarkkinen klusterointi

Hierarkkisen klusterointimenetelmän avulla on mahdollista tarkastella klusterien muodostumista ilman, että niiden lukumäärä on tiedossa etukäteen (Everitt ym., 2011), sekä havainnoida klusterien muodostumista klusteroinnin eri vaiheissa. Hierarkkisen klusteroinnin perusalgoritmit ovat joko kokoavia tai jakavia. Kokoavassa menetelmässä jokainen klusteroitava havainto on aluksi omassa luokassaan ja kluste-

roinnin edetessä kaksi samankaltaisinta luokkaa yhdistetään yhdeksi luokaksi niin kauan, kunnes kaikki havainnot kuuluvat samaan luokkaan (Everitt ym., 2011; Hair ym., 2019). Jakava klusterointi toimii käänteisesti niin, että kaikki havainnot kuuluvat alussa samaan luokkaan ja klusteroinnin edetessä luokat jaetaan pienempiin osiin (Hair ym., 2019). Kokoava hierarkkinen klusterointi on jakavaa klusterointia yleisemmin hyödynnetty ja yksinkertaisempi toteuttaa (Everitt ym., 2011).

Klusteroinnin tuloksena syntynyt sisäkkäisten klusterien rakenne voidaan esittää visuaalisesti kuvan 1 havainnollistamana dendrogrammina (Hair ym., 2019). Myöhemmin dendrogrammi voidaan katkaista kohdasta, jossa klusterien määrä vaikuttaa kontekstiedon valossa sopivalta (Everitt ym., 2011). Hierarkkisissa menetelmissä klusterijäsenyys on luonteeltaan pysyvä, jolloin kokoavan klusteroinnin tapauksessa kahta toisiinsa liitettyä luokkaa ei voida jälkikäteen erottaa (Everitt ym., 2011).



Kuva 1: Havainnollistus hierarkkisen klusteroinnin tuloksista muodostetusta dendrogrammista mukaillen lähdettä Everitt ym. (2011).

Klusteroitavien havaintojen samankaltaisuuden tai etäisyyden määrittämiseksi valitaan sopiva samankaltaisuusmitta. Kun klusteroitavan aineiston muuttujat ovat reaaliarvoisia, hyödynnetään yleisesti joko etäisyyteen tai korrelaatioon pohjautuvaa mitta (Everitt ym., 2011). Yleisimmin käytetty etäisyysmitta on L^2 -normi eli euklidinen etäisyys (Everitt ym., 2011).

Ensimmäisen yhdistämisen jälkeen klustereissa voi olla useampia havaintoja, jolloin etäisyysmitan rinnalle tarvitaan klusterien yhdistämiskriteerinä toimiva linkkifunktio (Hair ym., 2019). Yleisimmät linkkifunktiot perustuvat klusterien yhdistämiseen muun muassa lähimpien tai kauimpien havaintojen tai klusterien havainnoista muodostettujen yhdistelmien perusteella (Hair ym., 2019). Keskiarvoetäisyyteen perustuva linkkifunktio kuvaa kaikkien klusterin havaintojen keskiarvoista etäisyyttä kaikkiin toisen klusterin havaintoihin (Hair ym., 2019). Kun klusterin kaikki havainnot huomioidaan, poikkeavien havaintojen vaikutus on vähäisempi (Hair ym., 2019).

Tyypillisesti yhdistäminen tehdään linkkifunktion pienimmän arvon perusteella (Everitt ym., 2011).

Hierarkkisen klusteroinnin haasteena on kuitenkin algoritmin ahneus eli se, että klusterit yhdistetään kussakin yksittäisessä tilanteessa optimaalisella tavalla kaikkien päätösten kokonaiskuvaa huomioimatta (Everitt ym., 2011). Myöskään aikaisempia päätöksiä ei voida muuttaa. Lisäksi yleisimpiä linkkifunktioita hyödynnettäessä päädytään aina tiettyyn valitun linkkifunktion ohjaamaan ratkaisuun. Kun mukana ei ole satunnaisuutta, deterministinen hierarkkinen klusterointi antaa tuloksena selkeän luokittelun silloinkin, kun aineisto ei ole yksiselitteisesti jaettavissa klustereihin. Muun muassa Heller & Ghahramani (2005) ovat lähestyneet näitä haasteita soveltamalla bayesiläistä päättelyä hierarkkiseen klusterointiin perustaen klusterien yhdistämisen suurimpaan posterioritodennäköisyyteen, jolla kahden klusterin havainnot ovat peräisin samasta jakaumasta. Lisäksi heidän algoritminsa antaa posterioritodennäköisyyden arvoon perustuvan tavan päättää klusterien määrästä.

2.5 Monte Carlo -simulointi

Monte Carlo -simuloinnin yleistema on ottaa satunnaisotoksia tarkasteltavaa kohdetta mallintavista todennäköisyysjakaumista ja laskea otosten perusteella halutut tilastolliset suureet (Harrison, 2010). Simulointien toistaminen parantaa tuloksen luotettavuutta. Kokoava hierarkkinen klusterointi voidaan toteuttaa satunnaisuuselementin sisältävänä stokastisena klusterointina, jossa simuloidaan klusterien muodostumista hierarkian eri tasoilla. Tämän jälkeen voidaan tarkastella, kuinka usein tietyt havainnot esiintyvät samassa klusterissa.

3 Tutkimusaineisto ja -menetelmät

3.1 Klusteroinnin tavoitteet

Tässä työssä klusteroinnilla haetaan lisänäkemyksiä sairastavuuden ja työkyvyttömyyden eroista ja yhtäläisyyksistä Suomen hyvinvointialueiden välillä sosiaali- ja terveydenhuollon uudistuksen toimeenpanon aikaan vuosina 2021–2023. Vaikka aineistona hyödynnettävän Kansallisen terveysindeksin pohjalta on tunnistettu alueellisia eroja yksittäisten osaindeksien sekä kokonaisindeksien tasolla, eri osaindeksien tietoja laajemmin huomioivaa luokittelua ei ole etukäteen määritetty. Näin ollen hierarkkinen klusterointi soveltuu mahdollisten uusien rakenteiden havainnointiin. Deterministisen kokoavan klusteroinnin tavoitteena on löytää yksiselitteinen alueiden klusterointi jokaisella hierarkkisen klusteroinnin tasolla. Stokastisen kokoavan hierarkkisen klusteroinnin simuloinnilla saadaan tietoa deterministisen klusteroinnin tuloksen herkkyydestä. Siinä etäisyyksien käänteislukuihin perustuvia todennäköisyyksiä ei määritetä analyttisesti, sillä tämä olisi työlästä.

3.2 Klusteroitava aineisto

Klusterointi tehdään erikseen kahdelle eri aineistolle, joista toinen koostuu Kansallisen terveysindeksin sairastavuuden ikävakioiduista osaindekseistä ja toinen terveysindeksin työkyvyttömyyden ikävakioiduista osaindekseistä. Sairastavuuden osaindeksit kuvaavat diagnoosiryhmäkohtaista tietoa ja työkyvyttömyyden osaindeksit puolestaan erilaisten eläkkeiden ja etuuksien tietoja, joten luonteeltaan erilaiset tiedot pidetään tässä tarkastelussa erillään. Tuloksia tulkitessa on siten myös mahdollisuus tarkastella eroja diagnoosiryhmien ja työkyvyttömyyden klusterointien välillä. Sairastavuuden osaindeksien tiedot ovat saatavissa Terveiden ja hyvinvoinnin laitoksen tilasto- ja indikaattoripankki Sotkanetistä ([THL, 2025b](#)) ja työkyvyttömyyden osaindeksien tiedot Kansaneläkelaitoksen Tietotarjottimelta ([Kela, 2025a](#)).

Klusteroinnin kohteena olevat hyvinvointialueet $i \in \{1, \dots, n\}$ ($n = 23$) ovat:

1 Ahvenanmaa	7 Kainuu	13 Lappi	19 Pohjois-Savo
2 Etelä-Karjala	8 Kanta-Häme	14 Länsi-Uusimaa	20 Päijät-Häme
3 Etelä-Pohjanmaa	9 Keski-Pohjanmaa	15 Pirkanmaa	21 Satakunta
4 Etelä-Savo	10 Keski-Suomi	16 Pohjanmaa	22 Varsinais-Suomi
5 Helsinki	11 Keski-Uusimaa	17 Pohjois-Karjala	23 Vantaa ja Kerava
6 Itä-Uusimaa	12 Kymenlaakso	18 Pohjois-Pohjanmaa	

Klusterointimuuttujat sairastavuusindeksistä $k_a \in \{1, \dots, m_a\}$ ($m_a = 10$) ovat:

1 Aivoverisuonitauti-indeksi	6 Muistisairausindeksi
2 Alkoholisairastavuusindeksi	7 Sepelvaltimotauti-indeksi
3 Diabetesindeksi	8 Syöpäindeksi
4 Keuhkosairausindeksi	9 Tapaturmaindeksi
5 Mielenterveysindeksi	10 Tuki- ja liikuntaelinsairausindeksi

Klusterointimuuttujat työkyvyttömyysindeksistä $k_b \in \{1, \dots, m_b\}$ ($m_b = 3$) ovat puolestaan:

1 Työkyvyttömyyseläkeindeksi	3 Ammatillisen kuntoutuksen indeksi
2 Sairauspäiväraha indeksi	

Sairastavuuden osaindeksien arvoja kuvaava aineisto esitetään $X_a = [x_{ik_a}]_{i=1, \dots, n, k_a=1, \dots, m_a}$, ja työkyvyttömyyden osaindeksien arvot sisältävä aineisto $X_b = [x_{ik_b}]_{i=1, \dots, n, k_b=1, \dots, m_b}$.

3.3 Aineiston valmistelu

Kansallisen terveysindeksin aluekohtaisia lukuja voidaan tulkita suhteessa koko maan tasoon, jota kuvataan luvulla 100 ([Lindell ym., 2025](#)). Klusterointia varten terveysindeksin luvut standardisoidaan, sillä indeksilukujen vaihtelu koko maan tasosta on osassa indekseistä toisia suurempi. Ilman standardisointia suuremman vaihtelun indeksit hallitsisivat analyysiä, sillä ne vaikuttaisivat eniten klusteroinnissa käytettävän euklidiseen klusterietäisyyteen. Standardisoituja indeksilukuja kuvaa

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}, \quad (1)$$

missä $\bar{x}_k$ on osaindeksin k keskiarvo ja s_k kyseisen osaindeksin keskihajonta. Indeksien standardisoitujen arvojen keskiarvo on nolla ja keskihajonta yksi. Keskiarvo ja keskihajonta lasketaan kaavoilla

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}, \quad (2)$$

$$s_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}, \quad (3)$$

Tässä työssä kaikilla sairastavuuden osaindekseillä sekä työkyvyttömyyden osaindekseillä on standardisoinnin jälkeen omissa klusteroinneissaan sama painoarvo. Lisäksi osaindekseistä muodostetut sairastavuuden ja työkyvyttömyyden kokonaisindeksit jätetään tarkastelun ulkopuolelle. Kokonaisindeksien laskennassa eri sairausryhmiä sekä työkyvyttömyysetuuksia ja -eläkkeitä painotetaan eri tavoin. Esimerkiksi osa sairauksista on luonteeltaan kroonisia ja toiset puolestaan akuutimpia, jolloin vaikutukset hoidon ja elämänlaadun kannalta voivat olla erilaisia. Tässä työssä indeksien painokertoimet jätetään kuitenkin huomiotta. Standardisoinnin jälkeen pienemmän alueellisen vaihtelun indeksien merkitys voi hieman ylikorostua.

Klusterointi suoritetaan standardisoiduille indeksitiedoille $X_{\text{standardisoitu}} = [z_{ik}]_{i=1,\dots,n, k=1,\dots,m}$, missä m kuvaa muuttujien määrää. Sairastavuusindeksin tietojen klusteroinnissa $m = m_a$ ja työkyvyttömyysindeksin tietojen klusteroinnissa $m = m_b$. Jatkossa merkitään hyvinvointialueen standardisoituja indeksitietoja $x_i \in X$, missä $x_i \in \mathbb{R}^{1 \times m}$ sisältää alueen i indeksiluvut ja $X \in \mathbb{R}^{n \times m}$ sisältää kaikkien alueiden indeksitiedot. Tässä työssä hyödynnetään kaikkia osaindeksien muuttujia. Analyysivaiheessa aineistosta voidaan kuitenkin tarvittaessa rajata muista hyvinvointialueista selkeästi poikkeavat havainnot, jotta ne eivät vaikuta tuloksiin liikaa.

3.4 Deterministinen kokoava hierarkkinen klusterointi

Kokoavan hierarkkisen klusteroinnin tuloksena syntyvät klusterit $c \in C$ ovat joukkoja hyvinvointialueista $c \subseteq X$. Kullakin hierarkkisen klusteroinnin tasolla kaksi aikaisemman tason luokkaa yhdistetään toisiinsa ja yhdistämistä jatketaan niin kauan, että jokainen alue on samassa klusterissa. Klusteroinnin ensimmäisellä tasolla $l = 0$ kaikki alueet ovat omissa klustereissaan. Tasolla l muodostuvaa klusterointia kuvataan joukkona klustereita $y_l = \{c_1, \dots, c_{n-l}\}$, missä $c_1, \dots, c_{n-l}$ kuvaavat tason l klustereita. Kokonaisuudessaan kokoavan hierarkkisen klusteroinnin tulos voidaan tällöin ilmaista $y = (y_1, \dots, y_{n-1}) \in Y$, missä Y on joukko kaikki hierarkian tasot sisältäviä klusterointiratkaisuja. Alussa siis $y_0 = \{\{x_i\} | x_i \in X\}$. Klusteroinnin etäisyysmittana hyödynnetään euklidistä etäisyyttä

$$d(x_i, x_j) = \left[\sum_{k=1}^m (x_{ik} - x_{jk})^2 \right]^{1/2}, \quad (4)$$

missä x_{ik} ja x_{jk} ovat muuttujan k arvot hyvinvointialueille i ja j . Yhdistämisperusteena käytetään pienintä euklidistä etäisyyttä. Klusterien yhdistämisessä sovellettava keskiarvolinkkifunktio on

$$d(c_A, c_B) = \frac{1}{n_{c_A} \cdot n_{c_B}} \sum_{x_i \in c_A} \sum_{x_j \in c_B} d(x_i, x_j), \quad (5)$$

missä c_A ja c_B ovat erilliset klusterit, n_{c_A} alkioden määrä klusterissa c_A ja n_{c_B} alkioden lukumäärä klusterissa c_B . Deterministisen kokoavan hierarkkisen klusteroinnin pseudokoodi kuvataan algoritmissa 1. Klusteroinnin tulokset esitetään dendrogrammina tai yhden hierarkian tason klusterirakenteena valitsemalla aikaisempi pysäytyskohta, jossa klusterien määrä vaikuttaa aineistoon nähden sopivalta.

Algoritmi 1 Deterministinen kokoava hierarkkinen klusterointi

- 1: Jokainen alkio x_i on omassa klusterissaan, klusterien määrä on n .
 - 2: Lasketaan euklidiset etäisyydet $d(x_i, x_j)$ kaikkien alkioerien välillä.
 - 3: Lyhin etäisyys $d(x_i, x_j)$ määrittää samaan klusteriin yhdistettävät alkiot.
 - 4: **Toistetaan**
 - 5: Lasketaan ryhmien väliset etäisyydet linkkifunktiota $d(c_A, c_B)$ hyödyntäen.
 - 6: Yhdistetään klusterit pienimmän linkkifunktion $d(c_A, c_B)$ arvon perusteella.
 - 7: **kunnes** kaikki alkiot kuuluvat samaan klusteriin.
-

3.5 Stokastinen kokoava hierarkkinen klusterointi

Stokastisessa kokoavassa klusteroinnissa yhdistämistä ei toteuteta suoraan lyhimmän etäisyyden tai pienimmän linkkifunktion arvon perusteella, vaan keskiarvolinkkifunktion antamien etäisyyksien käänteislukuista muodostetaan jokaisella hierarkian tasolla todennäköisyysjakauma, josta arvotaan yhdistettävät klusterit. Tällöin ei voida tulkita pelkästään yksittäistä klusterointiratkaisua $y = (y_1, \dots, y_{n-1}) \in Y$, vaan klusterointi tehdään useita kertoja. Monte Carlo -simuloinnin hengessä voidaan ajatella, että toistamalla satunnaistettua klusterointia ja havainnoimmalla klusterien muodostumista kullakin tasolla saadaan tietoa mahdollisesta aineiston klusterirakenteesta. Tässä työssä todennäköisyys klusterien c_A ja c_B yhdistämiseen tasolla l lasketaan

$$\mathbb{P}[\text{Yhdistetään } c_A \text{ ja } c_B \text{ tasolla } l] = p_{(c_A, c_B)}^{(l)} = \frac{1/d(c_A, c_B)}{\sum_{\{a < b \mid c_a \in y_l, c_b \in y_l\}} 1/d(c_a, c_b)}. \quad (6)$$

Todennäköisyyksien (6) perusteella arvotaan, mitkä klusterit yhdistetään. Stokastisen klusteroinnin tulokset voidaan esittää kullakin tasolla kuvaamalla, kuinka usein tietyt

alkiot esiintyvät samassa klusterissa. Tässä hyödynnetään listaa hierarkkisen klusteroinnin tasokohtaisista kumulatiivisista klusterijäsenyysmatriiseista. Tasokohtaisen kumulatiivisen klusterijäsenyysmatriisin elementtiä voidaan kuvata kaavalla

$$M_{ij}^{(l)} = \sum_{t=1}^N \mathbf{1}\{x_i \text{ ja } x_j \text{ samassa klusterissa tasolla } l \text{ toistolla } t\}, \quad (7)$$

missä $N \in \mathbb{N}$ on simulointikertojen määrä ja $\mathbf{1}\{\cdot\}$ on indikaattorifunktio, joka saa arvon 1, kun x_i ja x_j ovat samassa klusterissa ja muuten arvon 0. Tällöin tason l todennäköisyysmatriisin elementti voidaan laskea $P_{ij}^{(l)} = M_{ij}^{(l)}/N$. Pseudokoodi stokastiseen klusterointiin esitetään algoritmissa 2. Algoritmissa ei ensisijaisesti tarkastella klusterijäsenyysmatriisia tasolla $l = 0$, sillä tällä tasolla jokainen alkio on omassa klusterissaan. Ensimmäisellä kumulatiivisen klusterijäsenyysmatriisien listan elementillä kuvataan näin ollen todennäköisyyksiä tasolla, jolla jotkin kaksi alkioita on yhdistetty.

Algoritmi 2 Stokastinen kokoava hierarkkinen klusterointi

- 1: Alustetaan lista, jossa on valmiina kumulatiiviset klusterijäsenyysmatriisit jokaiselle hierarkian tasolle. Kaikki matriisien arvot ovat alussa nollia.
 - 2: Lasketaan euklidiset etäisyydet $d(x_i, x_j)$ kaikkien alkioerien välillä.
 - 3: **Toistetaan** vaiheet 4.–11. $N \in \mathbb{N}$ kertaa:
 - 4: Jokainen alkio x_i on omassa klusterissaan, klusterien määrä on n .
 - 5: **Toistetaan**
 - 6: Lasketaan ryhmien väliset etäisyydet linkkifunktiota $d(c_A, c_B)$ hyödyntäen.
 - 7: Lasketaan yhdistämisen todennäköisyys $p_{(c_A, c_B)}^{(l)}$ kaikkien klusteriparien välillä.
 - 8: Arvotaan yhdistettävät klusterit todennäköisyyksien perusteella.
 - 9: Luodaan jäsenyysmatriisi kuvaamaan alkioiden esiintymisestä samassa klusterissa.
 - 10: Summataan tason jäsenyysmatriisin tiedot tason kumulatiiviseen jäsenyysmatriisiin.
 - 11: **kunnes** kaikki alkiot kuuluvat samaan klusteriin.
 - 12: Lasketaan todennäköisyydet alkioiden esiintymiselle samassa klusterissa jakamalla kumulatiivisen klusterijäsenyysmatriisin luvut kaikilla tasoilla toistojen määrällä.
-

4 Tulokset

Hierarkkinen klusterointi sekä klusteroinnin tulosten visualisointi toteutetaan Python-kielellä. Aineiston valmistelussa sekä analyysissä käytetään Pandas- ja NumPy-kirjastoja (McKinney, W., ym., 2010; Harris, C. R., ym., 2020). Deterministisessä klusteroinnissa hyödynnetään Scikit-learn- ja SciPy-kirjastojen (Pedregosa, F., ym., 2011; Virtanen, P., ym., 2020) valmiita hierarkkisen klusteroinnin funktioita. Stokastista klusterointia varten rakennetaan algoritmin 2 pohjalta erillinen funktio, joka löytyy liitteestä B. Visualisoinneissa käytetään Matplotlib- sekä Seaborn-kirjastoja (Hunter, J. D., 2007; Waskom, M., ym., 2017) ja kartta-aineiston käsittelyssä GeoPandas-kirjastoa (Jordahl, 2020). Tässä työssä ei analysoida laajasti yksittäisten osaindeksien

alueellisia tuloksia, sillä alueellisia indeksilukuja on avattu kattavasti Kansallisen terveysindeksin tilastoraportissa (Lindell ym., 2025). Klusteroinnin tuloksia tulkittaessa nojataan tilastoraportin päätelmiin.

Stokastista klusterointia simuloidaan standardisoiduilla aineistoilla $N = 10000$ kertaa. Tällä toistomäärällä saadaan luotettavahkoja tuloksia koodin suoritusajan pyssyessä kohtuullisena. Tuloksena saadaan lista todennäköisyysmatriiseja hierarkkisen klusteroinnin eri tasoilla. Matriisien elementit kuvaavat kahden hyvinvointialueen todennäköisyyttä esiintyä samassa klusterissa tietyllä hierarkian tasolla. Todennäköisyyksiä kuvataan lämpökarttana. Todennäköisyyksien tulkinnan selkeyttämiseksi simulaatio toteutetaan aluksi (22×1) -nollamatriisille, joka havainnollistaa tilannetta, jossa 22:n hyvinvointialueen (Ahvenanmaan tiedot poistetaan ennen stokastista klusterointia, osioissa 4.1 ja 4.2) standardisoiduissa indeksitiedoissa ei olisi lainkaan vaihtelua. Sovellettava algoritmi löytää aineistosta jokaisella toistokerralla jonkin rakenteen, vaikka alueiden välillä ei olisi lainkaan eroa. Tasojen todennäköisyyksien vertailu ei myöskään ole yksiselitteistä, sillä klusteroinnin edetessä samassa klusterissa esiintymisen todennäköisyydet kasvavat, kun aikaisemmilla tasoilla samaan klusteriin päätyneet alueet ovat hierarkkisessa klusteroinnissa samassa klusterissa myös ylemmillä tasoilla. Nollamatriisin stokastisen klusteroinnin avulla saatavien referenssitodennäköisyyksien minimi, mediaanit ja toiseksi suurimmat arvot eri tasoilla esitetään taulukossa 1. Maksimien sijaan esitetään toiseksi suurimmat arvot, sillä tarkastelussa ei haluta huomioida jokaisella tasolla varmaa tapahtumaa alueiden esiintymisestä samassa klusterissa itsensä kanssa. Taulukossa ja seuraavien osioiden kuvissa ei esitetä tasoa $l = 0$, jolla kaikki alueet ovat omissa klustereissaan, eikä tasoa $l = 21$, jolla kaikki 22 aluetta kuuluvat samaan klusteriin. Sairastavuuden ja työkyvyttömyyden osaindeksilukujen pohjalta toteutettavassa stokastisessa klusteroinnissa jokaisen tason todennäköisyysasteikon arvoja voidaan verrata näihin nollamatriisin avulla saatuihin referenssiarvoihin.

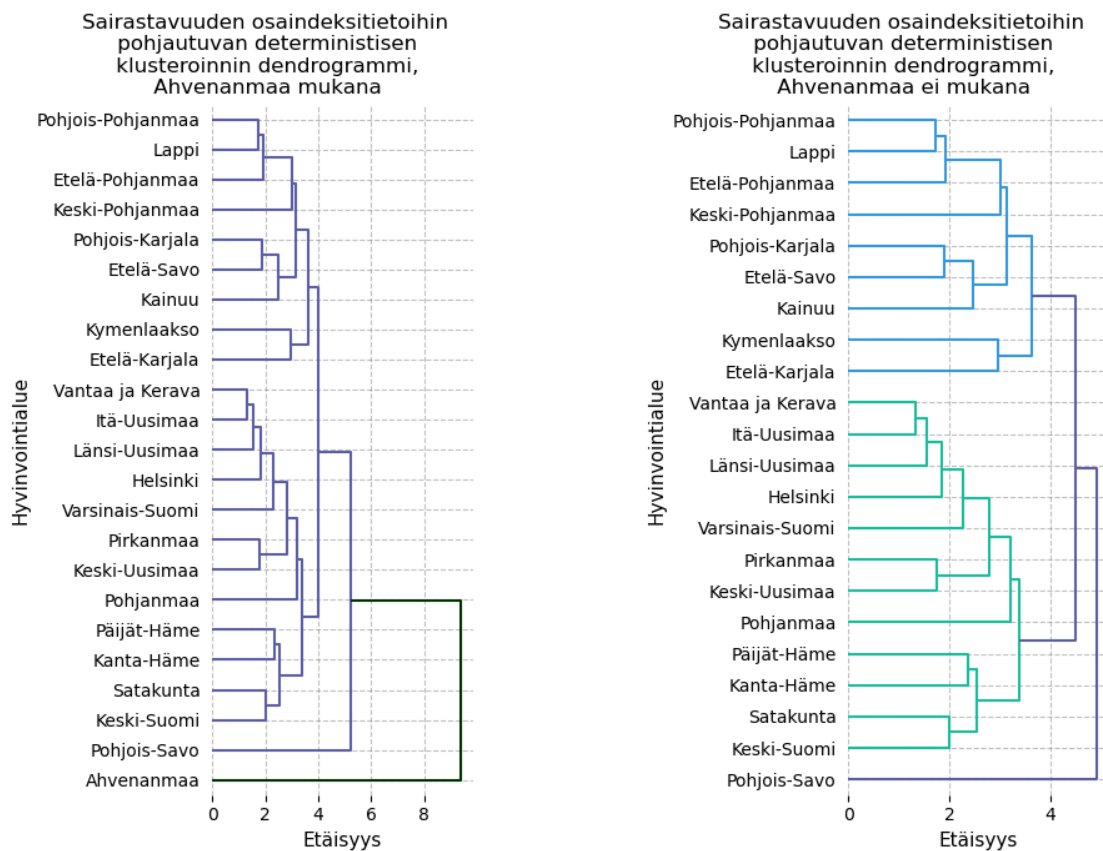
Taso l	Klusterien määrä $n - l$	Minimi	Mediaani	Toiseksi suurin arvo
1	21	0.003	0.004	0.007
2	20	0.007	0.009	0.012
3	19	0.011	0.014	0.018
4	18	0.016	0.020	0.024
5	17	0.022	0.026	0.032
6	16	0.029	0.034	0.040
7	15	0.036	0.042	0.048
8	14	0.046	0.051	0.057
9	13	0.055	0.061	0.068
10	12	0.067	0.073	0.081
11	11	0.080	0.088	0.095
12	10	0.096	0.104	0.112
13	9	0.114	0.124	0.132
14	8	0.137	0.149	0.156
15	7	0.167	0.179	0.187
16	6	0.207	0.218	0.226
17	5	0.259	0.270	0.279
18	4	0.329	0.342	0.351
19	3	0.442	0.452	0.464
20	2	0.621	0.633	0.645

Taulukko 1: Stokastisen klusteroinnin tasojen todennäköisyysjakaumia nollamatriisille.

Jos stokastisen klusteroinnin antama kahden alueen todennäköisyys esiintyä samassa klusterissa tietyllä tasolla on suurempi kuin taulukon 1 todennäköisyyksien toiseksi suurin arvo samalla hierarkian tasolla, niin alueet esiintyvät yleisemmin samassa klusterissa kuin silloin, kun minkään alueiden välillä ei ole eroja.

4.1 Klusterointi sairastavuuden osaindeksien perusteella

Deterministisen klusteroinnin tulokset sairastavuuden osaindeksin standardisoidulle aineistolle visualisoidaan kuvissa 2a ja 2b dendrogrammeina. Kuvassa 2a visualisoidaan klusteroinnin tulos koko aineistolla. Ahvenanmaan indeksiluvut ovat syöpäindeksiä lukuun ottamatta merkittävästi alhaisempia kuin muiden alueiden indeksiluvut. Ahvenanmaan sairastavuuden osaindeksien tiedot poistetaan aineistosta, eikä niitä käsitellä analyysin seuraavissa vaiheissa, sillä tiedoissa tiedetään olevaan puutteita (Lindell ym., 2025). Terveysindeksin tilastoraportin mukaan on kuitenkin todennäköistä, että sairastavuus on Ahvenanmaalla tilastointipuutteista huolimatta vähäisempää Manner-Suomeen verrattuna (Lindell ym., 2025). Kuvassa 2b voidaan havaita suurempi etäisyyden kasvu ennen toiseksi viimeistä yhdistämistä, joten dendrogrammin sopivana leikkauskohtana voidaan alustavasti ajatella kolmen klusterin tasoa.

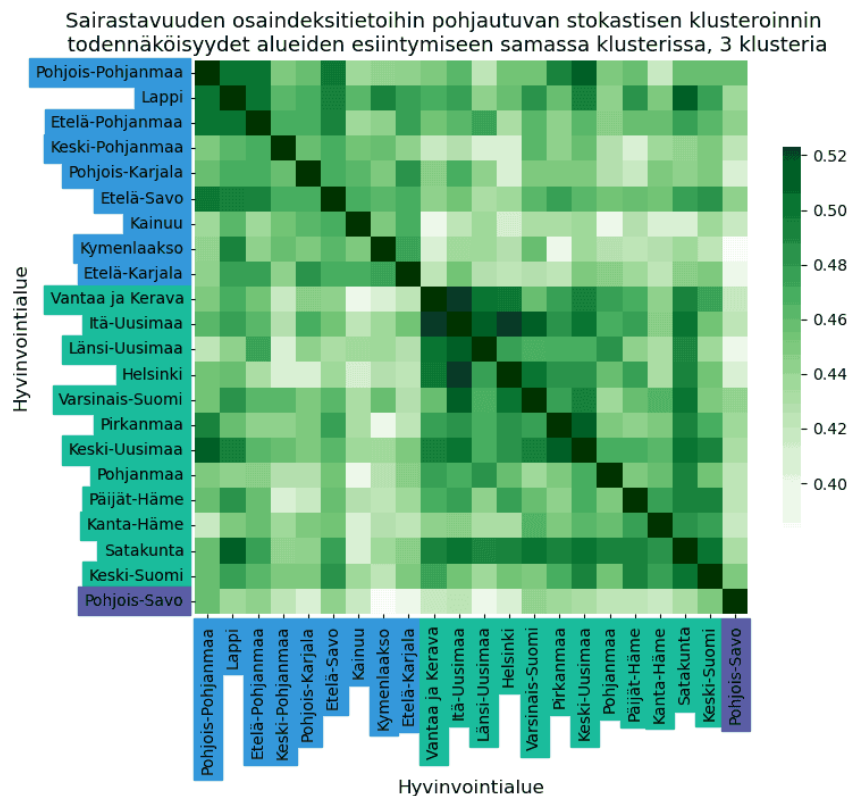


(a) Ahvenanmaa mukana aineistossa.

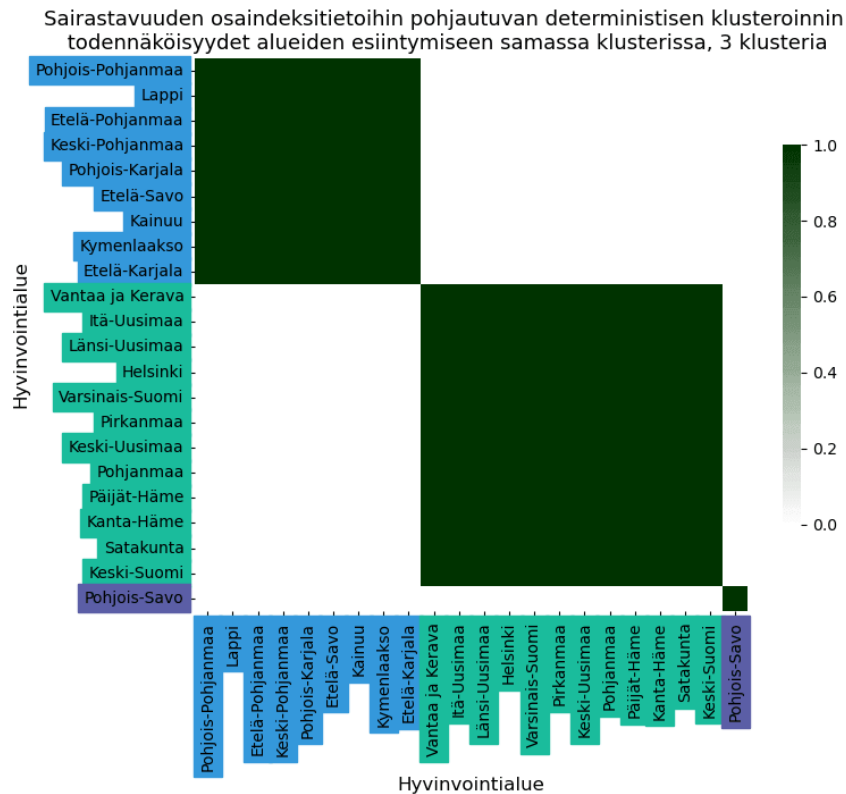
(b) Ahvenanmaa poistettu aineistosta.

Kuva 2: Sairastavuuden deterministisen hierarkkisen klusteroinnin dendrogrammit.

Sairastavuuden osaindeksilukuihin perustuvan stokastisen klusteroinnin tasokohtaiset todennäköisyydet esitetään liitteessä A kuvissa A1 ja A2. Kuvassa A1 todennäköisyysasteikko on tasokohtainen, jolloin alueiden parittaisten todennäköisyyksien erot ovat selkeästi tarkasteltavissa kullakin tasolla. Lämpökartan diagonaalilla olevat elementit vastaavat aina todennäköisyyttä 1. Kuvassa A2 todennäköisyysasteikon ollessa jokaisella tasolla $[0, 1]$ tasokohtaisen rakenteen havaitseminen on vaikeaa. Tasokohtaisissa lämpökartoissa on havaittavissa rakenne, joka tukee deterministisen klusteroinnin tulosten pohjalta muodostettua hypoteesia kolmesta erillisestä klusterista. Kolmen klusterin taso (taso 19) on esitetty suurempana kuvassa 3. Kuvan asteikon pienin arvo vastaa tason todennäköisyyksien pienintä arvoa ja suurin tason todennäköisyyksien suurinta arvoa. Asteikko on selvästi leveämpi kuin kolmen klusterin tasolla taulukossa 1. Parittaisten todennäköisyyksien välillä on satunnaisuudesta poikkeavaa vaihtelua. Stokastisen klusteroinnin lämpökartan rakenteiden hahmottamiseksi kuvassa 4 esitetään deterministisen klusteroinnin todennäköisyyksien lämpökartta kolmen klusterin tasolla. Deterministisessä klusteroinnissa ei ole mukana satunnaisuutta, joten alueet joko esiintyvät tai eivät esiinny samassa klusterissa. Tällöin kaikki parittaiset todennäköisyydet ovat arvoltaan joko 0 tai 1.



Kuva 3: Sairastavuuden stokastisen klusteroinnin todennäköisyydet kolmen klusterin tasolla.



Kuva 4: Sairastavuuden deterministisen klusteroinnin todennäköisyydet kolmen klusterin tasolla.

Stokastisen klusteroinnin lämpökartoista voi tarkastella valittujen klusterien sisäisiä rakenteita sekä vaihtoehtoisia ryhmittelyjä. Tarkastelemalla kuvan 2b dendrogrammia yhdessä stokastisen klusteroinnin lämpökarttojen kuvien A1, A2 sekä 3 kanssa voidaan havaita, että merkittävintä samankaltaisuutta on Uudellamaalla sijaitsevien hyvinvointialueiden välillä sekä Itä- ja Pohjois-Suomen hyvinvointialueiden välillä. Uudenmaan hyvinvointialueet taas poikkeavat piirteiltään Itä- ja Pohjois-Suomen hyvinvointialueista.

Taulukko 2 kuvaa klusterien jäsenet kolmen klusterin tasolla sairastavuuden osaindeksien deterministisen klusteroinnin tuloksena. Taulukko 3 sisältää näiden kolmen klusterin keskiarvot, mediaanit, minimi ja maksimit sairastavuuden alkuperäisistä standardisoimattomista osaindeksiluvuista. Taulukon 3 värit kuvaavat rivikohtaisesti tunnusluvun suhdetta saman rivin lukuihin. Värit on valittu siten, että ne noudattavat ensisijaisesti samankaltaisia värivalintoja kuin muissa kuvaajissa on käytetty sille klusterille, joka saa taulukossa eniten arvoja, jotka ovat yhdellä värillä kuvattavissa olevassa suhteessa muihin saman rivin arvoihin. Esimerkiksi klusterin 1 arvot ovat usein rivin pienimpiä, joten rivien pienimmät arvot kuvataan klusterin 1 värillä, joka on turkoosi. Taulukosta voidaan havaita, että Länsi- ja Etelä-Suomen hyvinvointialueita sisältävä klusteri 1 saa pienimpiä arvoja lähes kaikissa osaindekseissä.

Syöpäindeksin suurempi luku voi kertoa myös varhaisdiagnostiikan onnistumisesta (Lindell ym., 2025). Klusteri 2 sisältää Itä- ja Pohjois-Suomen hyvinvointialueita ja klusteri 3 Pohjois-Savon hyvinvointialueen. Molemmissa klustereissa tunnusluvut saavat taulukon 3 mukaan lähes kaikissa osaindekseissä klusterin 1 arvoja suurempia arvoja. Klusteri 3 saa useiden indeksien kohdalla kaikista suurimmat arvot. Länsi- ja Etelä-Suomessa on terveysindeksin lukujen mukaan Itä- ja Pohjois-Suomea vähemmän sairastavuutta lähes kaikissa diagnoosiryhmissä. Pohjois-Savossa sairastavuus on maan suurinta useissa diagnoosiryhmissä.

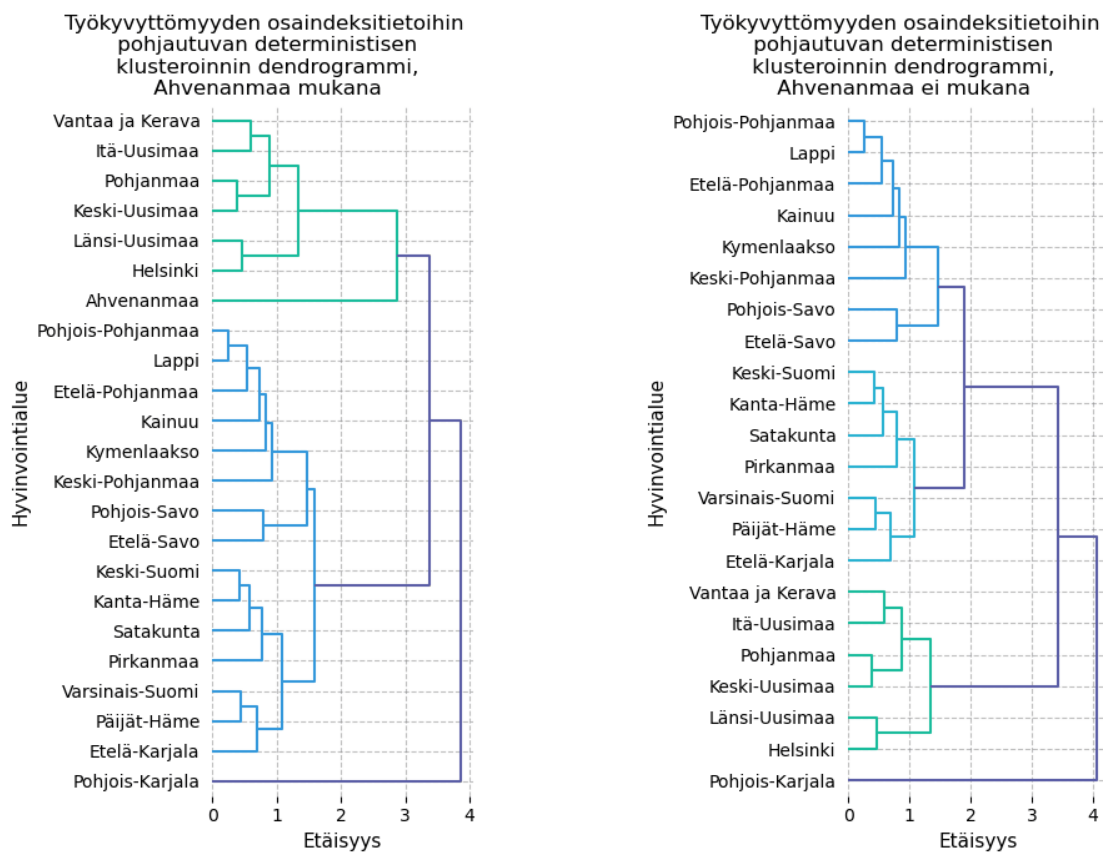
Klusteri 1	Osaindeksi	Klusteri 1	Klusteri 2	Klusteri 3
Helsinki	<i>Keskiarvo</i>			
Itä-Uusimaa	Aivoverisuonitauti-indeksi	101.1	104.2	95.2
Kanta-Häme	Alkoholisairastavuusindeksi	94.7	117.7	114.1
Keski-Uusimaa	Diabetesindeksi	97.3	106.7	112.1
Keski-Suomi	Keuhkosairausindeksi	98.3	106.7	110.0
Länsi-Uusimaa	Mielenterveysindeksi	91.3	105.5	144.3
Pirkanmaa	Muistisairausindeksi	95.0	108.8	110.2
Pohjanmaa	Sepelvaltimotauti-indeksi	94.4	116.6	103.4
Päijät-Häme	Syöpäindeksi	100.2	99.3	94.9
Satakunta	Tapaturmaindeksi	98.0	113.3	73.1
Vantaa ja Kerava	Tuki- ja liikuntaelinsairausindeksi	93.3	111.5	132.9
Varsinais-Suomi				
Klusteri 2	<i>Mediaani</i>			
Etelä-Karjala	Aivoverisuonitauti-indeksi	93.4	103.1	95.2
Etelä-Pohjanmaa	Alkoholisairastavuusindeksi	95.2	118.2	114.1
Etelä-Savo	Diabetesindeksi	98.4	105.1	112.1
Kainuu	Keuhkosairausindeksi	96.6	106.9	110.0
Keski-Pohjanmaa	Mielenterveysindeksi	91.2	104.2	144.3
Kymenlaakso	Muistisairausindeksi	96.9	110.3	110.2
Lappi	Sepelvaltimotauti-indeksi	93.1	116.3	103.4
Pohjois-Karjala	Syöpäindeksi	100.7	101.4	94.9
Pohjois-Pohjanmaa	Tapaturmaindeksi	98.2	111.3	73.1
	Tuki- ja liikuntaelinsairausindeksi	94.9	113.4	132.9
Klusteri 3	<i>Minimi</i>			
Pohjois-Savo	Aivoverisuonitauti-indeksi	89.6	93.8	95.2
	Alkoholisairastavuusindeksi	60.1	93.8	114.1
	Diabetesindeksi	83.6	98.7	112.1
	Keuhkosairausindeksi	84.6	90.6	110.0
	Mielenterveysindeksi	74.7	86.3	144.3
	Muistisairausindeksi	80.5	89.3	110.2
	Sepelvaltimotauti-indeksi	80.9	105.1	103.4
	Syöpäindeksi	93.3	88.2	94.9
	Tapaturmaindeksi	86.6	93.5	73.1
	Tuki- ja liikuntaelinsairausindeksi	72.5	94.5	132.9
	<i>Maksimi</i>			
	Aivoverisuonitauti-indeksi	123.6	114.4	95.2
	Alkoholisairastavuusindeksi	119.1	142.3	114.1
	Diabetesindeksi	104.6	116.1	112.1
	Keuhkosairausindeksi	111.9	119.6	110.0
	Mielenterveysindeksi	106.0	122.8	144.3
	Muistisairausindeksi	101.6	119.7	110.2
	Sepelvaltimotauti-indeksi	109.4	126.0	103.4
	Syöpäindeksi	105.3	104.7	94.9
	Tapaturmaindeksi	109.4	136.3	73.1
	Tuki- ja liikuntaelinsairausindeksi	110.4	124.1	132.9

Taulukko 2:
Hyvinvointialueet
klustereittain
sairastavuuden
osaindekseihin
perustuen.

Taulukko 3: Sairastavuuden osaindeksien tilastollisia tunnuslukuja klustereittain.

4.2 Klusterointi työkyvyttömyyden osaindeksien perusteella

Dendrogrammit työkyvyttömyyden deterministisen klusteroinnin tuloksista esitetään kuvissa 5a ja 5b. Työkyvyttömyyden osalta Ahvenanmaan luvut eivät poikkea yhtä merkittäväällä tavalla muiden hyvinvointialueiden luvuista, mutta sairastavuuden ja työkyvyttömyyden analyysien yhdenmukaisuuden vuoksi Ahvenanmaan tiedot poistetaan myös työkyvyttömyyden osaindeksilukuja kuvaavasta aineistosta dendrogrammin 5a piirtämisen jälkeen. Työkyvyttömyyseläkeindeksi sekä ammatillisen kuntoutuksen indeksi saavat Ahvenanmaalla selvästi koko maan tasoa pienemmät indeksiluvut, mutta sairauspäivärahoissa indeksiluku on jonkin verran koko maan tasoa suurempi. Kuvassa 5b luonnollisena kohtana dendrogrammin katkaisemiseen voisivat olla kolmen tai neljän klusterin tasot.



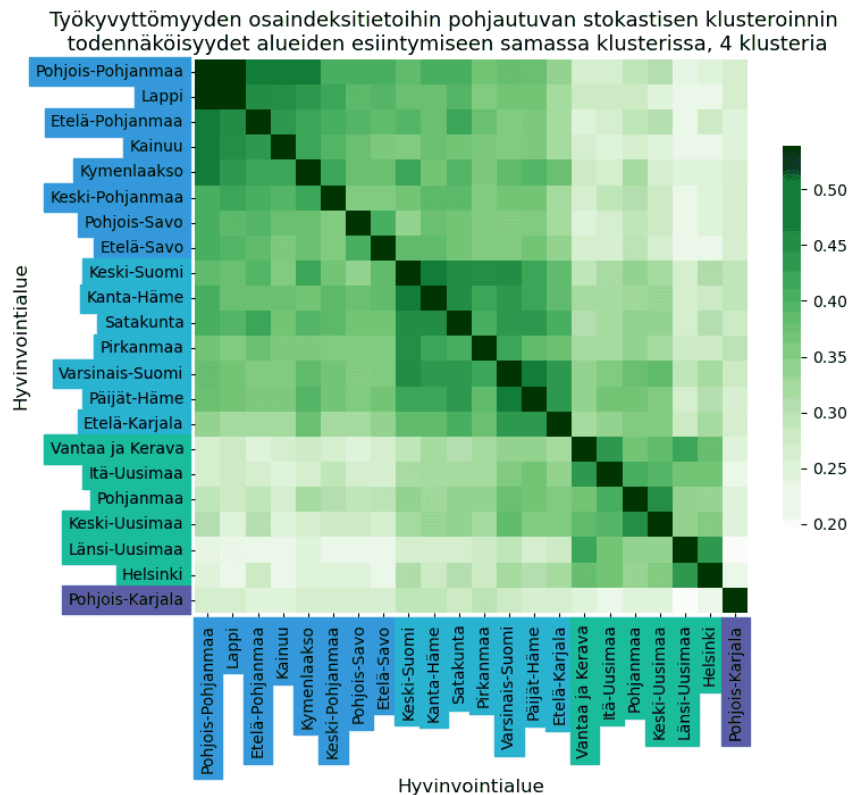
(a) Ahvenanmaa mukana aineistossa.

(b) Ahvenanmaa poistettu aineistosta.

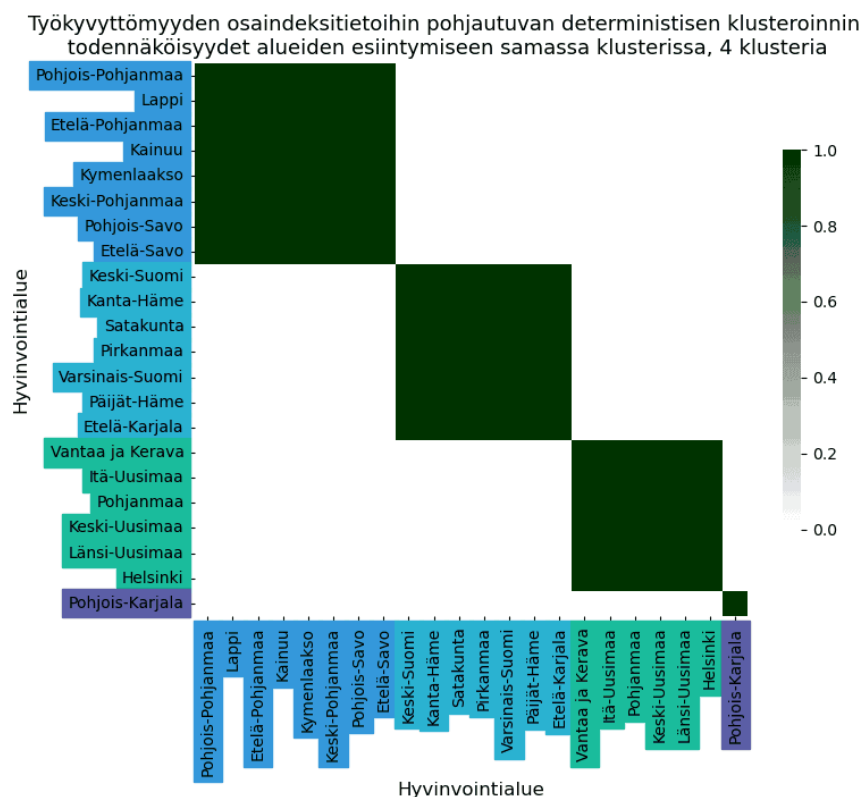
Kuva 5: Työkyvyttömyyden deterministisen hierarkkisen klusteroinnin dendrogrammit.

Stokastisen klusteroinnin tuloksena saadut todennäköisyydet esitetään tasokoh-
taisesti liitteen A kuvissa A3 ja A4. Kuvassa A3 jokaisella tasolla on oma todennä-
köisyysasteikkonsa ja kuvassa A4 kaikkien tasojen todennäköisyysasteikkona on $[0, 1]$.
Havaitaan, että useilla tasoilla hahmottuu neljän klusterin rakenne, jossa keskim-
mäinen klusteri osoittaa kuitenkin samankaltaisia piirteitä kahden muun klusterin
kanssa. Uudenmaan alueet ja Pohjanmaa esiintyvät selkeästi omana klusterinaan.

Myös Itä- ja Pohjois-Suomen alueet yhdistyvät omaan klusteriinsa. Lounais-Suomen, Hämeen ja Keski-Suomeen alueita sisältävä klusteri voisi kuulua samaan klusteriin sekä Uudenmaan alueiden että Itä- ja Pohjois-Suomen alueiden kanssa. Samankaltaisuus on vahvempi Itä- ja Pohjois-Suomen alueiden kanssa. Kahden klusterin yhdistämisen jälkeen suurin osa hyvinvointialueista olisi samassa klusterissa, joten jatkossa työkyvyttömyyden osaindeksilukujen perusteella toteutettu klusterointi esitetään neljän klusterin tasolla, jotta työkyvyttömyyden eroista saadaan tarkempaa näkemystä. Stokastisen klusteroinnin todennäköisyydet neljän klusterin tasolla (taso 18) esitetään kuvassa 6. Todennäköisyysasteikko on leveämpi kuin taulukon 1 referenssiarvoilla neljän klusterin tasolla. Todennäköisyyksien erot ovat hierarkian tason huomioiden selkeämpiä kuin sairastavuuden stokastisessa klusteroinnissa. Tätä voi osaltaan selittää se, että työkyvyttömyyden osaindeksejä on vähemmän kuin sairastavuuden osaindeksejä. Työkyvyttömyyden klusteroinnissa yksittäisen osaindeksin merkittävällä alueellisella vaihtelulla on suurempi painoarvo. Kuvassa 7 esitetään deterministisen klusteroinnin lämpökartta neljän klusterin tasolla.



Kuva 6: Työkyvyttömyyden stokastisen klusteroinnin todennäköisyydet kolmen klusterin tasolla.



Kuva 7: Työkyvyttömyyden deterministisen klusteroinnin todennäköisyydet kolmen klusterin tasolla.

Taulukko 4 kuvaa työkyvyttömyyden osaindeksien pohjalta toteutetun deterministisen klusteroinnin tuottamia klustereita neljän klusterin tasolla. Taulukossa 5 esitetään klusterien piirteitä kuvaavia tilastollisia tunnuslukuja työkyvyttömyyden standardisoimattomista osaindeksiluvuista. Värit on valittu pääosin samoilla periaatteilla kuin sairastavuuden klusterien tunnuslukuja esittävässä taulukossa 3. Muutoksena väreissä on vaalean sinisen ja vaalean violetin värin keskinäisen suhteen vaihtuminen niin, että vaalea sininen kuvaa taulukossa 5 rivien suurimpia indeksilukuja. Havaitaan, että kaikki Uudenmaan alueet sekä Pohjanmaan hyvinvointialueen sisältävällä klusterilla 1 on pienimmät tunnusluvut kaikissa työkyvyttömyyden osaindekseissä. Työkyvyttömyys on tässä klusterissa vähäisintä. Itä- ja Pohjois-Suomen alueita sisältävässä klusterissa 3 työkyvyttömyys on yleisintä. Lounais-Suomen, Hämeen ja Keski-Suomen alueita kuvaava klusteri 2 on piirteiltään välimuoto Uudenmaan klusterista ja Itä- ja Pohjois-Suomen klusterista, sillä työkyvyttömyys on yleisempää kuin klusterissa 1, mutta pääosin vähäisempää kuin klusterissa 3. Klusteri 4 sisältää Pohjois-Karjalan hyvinvointialueen, jossa erityisesti ammatillisen kuntoutuksen saajia on väestöön suhteutettuna paljon. Tämä voi kertoa myös alueen mahdollisuuksista tarjota kuntoutusta maksettavien työkyvyttömyyseläkkeiden sijaan.

Klusteri 1	Osaindeksi	Klusteri 1	Klusteri 2	Klusteri 3	Klusteri 4
Helsinki	<i>Keskiarvo</i>				
Itä-Uusimaa	Työkyvyttömyyseläkkeet	75.3	107.0	128.5	120.0
Keski-Uusimaa	Sairauspäivärahat	81.5	108.4	125.0	104.0
Länsi-Uusimaa	Ammatillinen kuntoutus	63.8	109.4	121.4	222.0
Pohjanmaa	<i>Mediaani</i>				
Vantaa ja Kerava	Työkyvyttömyyseläkkeet	74.00	107.00	129.0	120.0
Klusteri 2	Sairauspäivärahat	83.00	108.0	126.00	104.0
Etelä-Karjala	Ammatillinen kuntoutus	62.50	110.00	114.00	222.0
Kanta-Häme	<i>Minimi</i>				
Keski-Suomi	Työkyvyttömyyseläkkeet	67.00	100.00	113.00	120.0
Pirkanmaa	Sairauspäivärahat	63.00	97.00	112.0	104.0
Päijät-Häme	Ammatillinen kuntoutus	52.00	82.00	96.00	222.0
Satakunta	<i>Maksimi</i>				
Varsinais-Suomi	Työkyvyttömyyseläkkeet	84.00	116.00	139.0	120.0
Klusteri 3	Sairauspäivärahat	97.00	115.00	132.0	104.0
Etelä-Pohjanmaa	Ammatillinen kuntoutus	82.00	143.00	166.00	222.0
Etelä-Savo					
Kainuu					
Keski-Pohjanmaa					
Kymenlaakso					
Lappi					
Pohjois-Pohjanmaa					
Pohjois-Savo					
Klusteri 4					
Pohjois-Karjala					

Taulukko 5: Työkyvyttömyyden osaindeksien tilastollisia tunnuslukuja klustereittain.

Taulukko 4:
Hyvinvointialueet
klustereittain
työkyvyttömyyden
osaindekseihin
perustuen.

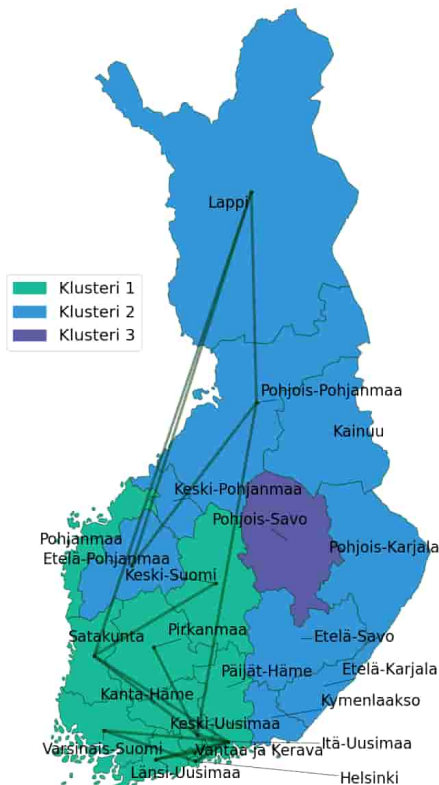
4.3 Sairastavuus ja työkyvyttömyys hyvinvointialueilla

Deterministisen ja stokastisen kokoavan hierarkkisen klusteroinnin tuloksena hahmotuu jako vähäisemmän sairastavuuden ja työkyvyttömyyden Länsi- ja Etelä-Suomeen sekä suuremman sairastavuuden ja työkyvyttömyyden Itä- ja Pohjois-Suomeen. Sairastavuus ja työkyvyttömyys ovat vähäisintä Uudellamaalla pääkaupungin lähellä. Tulos työkyvyttömyyden vähäisestä alueellisesta esiintyvyydestä pääkaupunkiseudulla on samanlainen kuin Ruotsin kuntatasolla tarkastellussa alueiden jaottelussa työkyvyttömyyden korkean ja matalan esiintyvyyden alueisiin (Virtanen ym., 2022).

Kuvissa 8a ja 8b esitetään väreillä deterministisen klusteroinnin tuloksena syntyvät klusterit ja alueiden välisillä linkeillä stokastisen klusteroinnin valitun todennäköisyyskynnyksen ylittävät yhteydet klusterimäärää vastaavalla hierarkian tasolla. Kahden hyvinvointialueen välinen linkki kuvaa, että kyseisellä stokastisen klusteroinnin tasolla todennäköisyys kyseisten hyvinvointialueiden esiintymiselle samassa klusterissa on toteutetun simulaation perusteella valittua todennäköisyyskynnystä suurempi. Kuvissa todennäköisyydet on valittu taulukon 1 nollamatriisiin referens-

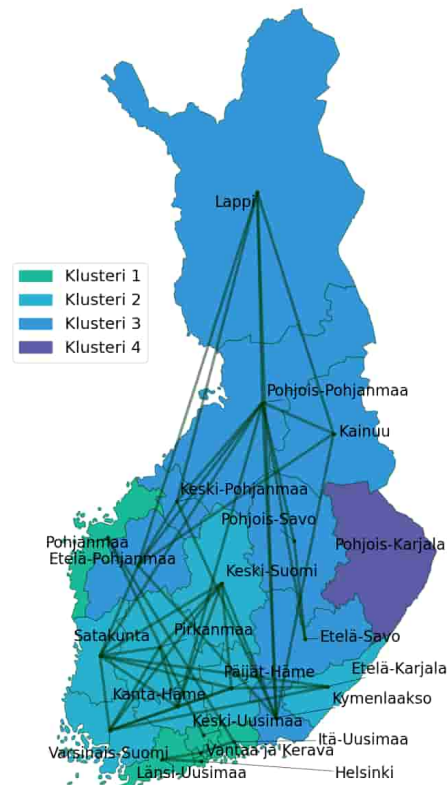
siarvoja suuremmiksi sairastavuuden osalta kolmen klusterin tasolla ja työkyvyttömyydelle neljän klusterin tasolla. Sairastavuuden linkkien todennäköisyytenä on kuvassa 0.5 ja työkyvyttömyyden linkeillä 0.4. Nämä ovat vain yhdet mahdolliset valinnat, kynnsarvoina voisi käyttää vaihtoehtoisiaakin todennäköisyyksiä.

Sairastavuuden osaindeksitietoihin pohjautuvan klusteroinnin tulokset kartalla, 3 klusteria, todennäköisyyskynnys 0.5



(a) Sairastavuuden klusterit.

Työkyvyttömyyden osaindeksitietoihin pohjautuvan klusteroinnin tulokset kartalla, 4 klusteria, todennäköisyyskynnys 0.4



(b) Työkyvyttömyyden klusterit.

Kuva 8: Klusterointi kartalla, linkit kuvaavat todennäköisyyskynnyksen ylittäviä yhteyksiä. Karttapohjan lähde: [Tilastokeskus \(2025\)](#).

Tässä työssä kaikki linkit esitetään kuvissa yhtä vahvoina, mutta niiden värillä tai leveydellä voitaisiin myös kuvata todennäköisyyden arvoa. Linkkien avulla voidaan arvioida determinististä klusterointia tarkastelemalla esimerkiksi, kuinka monet linkeistä pysyvät yhdellä värillä kuvatus klusterin sisällä. Sairastavuuden ja työkyvyttömyyden klusteroinnin tarkastelussa linkkien alueellisuuden hahmottaminen onnistuu verrattain helposti, kun klusterit ovat pääosin maantieteellisesti yhtenäisiä. Kuvista on havaittavissa, että stokastisen klusteroinnin linkit esiintyvät suurilta osin klusterien sisällä, mutta työkyvyttömyyden klusterien 2 ja 3 samankaltaisuudet näkyvät klusterien hyvinvointialueiden välisinä linkeinä. Työkyvyttömyydessä alueiden väliset erot ovat suurempia kuin sairastavuudessa, joten stokastisen klusteroinnin todennäköisyyksien erot ovat myös suurempia. Klustereiden peilaaminen

viiteen hyvinvointialueiden yhteistyötä ohjaavaan yliopistollisten sairaaloiden ympärille muodostuvaan yhteistyöalueeseen tuo esille sairastavuuden ja työkyvyttömyyden suurimman yleisyyden Oulun ja Kuopion yliopistollisten sairaaloiden vastuualueilla.

5 Pohdinta

Tarkastelu terveyserojen jakautumisesta Suomen hyvinvointialueilla Kansallisen terveysindeksin tietojen pohjalta vuosina 2021–2023 piirtää kuvan alhaisen sairastavuuden ja työkyvyttömyyden Länsi- ja Etelä-Suomen hyvinvointialueista sekä asteittain kohti Itä- ja Pohjois-Suomea kasvavasta alueellisesta sairastavuuden ja työkyvyttömyyden esiintyvyydestä. Työssä hyödynnetyn deterministisen klusterien keskiarvoetäisyyksiin perustuvan hierarkkisen klusterointimenetelmän avulla sairastavuusindeksien pohjalta päädytään kolmeen erilliseen klusteriin, joista Etelä- ja Länsi-Suomen klusterissa sairastavuus on vähäisintä ja Itä- ja Pohjois-Suomen kahdessa klusterissa yleisintä. Työkyvyttömyyden osalta hahmottuu Etelä-Suomen klusterin ja Itä- ja Pohjois-Suomen klusterin välissä Lounais-Suomen, Hämeen ja Keski-Suomen hyvinvointialueita sisältävä klusteri. Työkyvyttömyyden osaindeksiin perustuvan stokastisen keskiarvoetäisyyksien käänteisluvuista johdettuihin todennäköisyyksiin pohjautuvan klusterointimenetelmän tuloksena havaitaan, että keskimmäinen klusteri muistuttaa piirteiltään sekä vähäisen työkyvyttömyyden Etelä-Suomen klusteria että korkeamman työkyvyttömyyden esiintyvyyden Itä- ja Pohjois-Suomen klusteria. Tulokset vahvistavat käsitystä alueellisista terveyseroista Suomessa.

Työssä esitelty keskiarvoisista etäisyyksistä laskettuihin todennäköisyyksiin perustuva stokastinen klusterointimenetelmä antaa lisänäkemystä, joka helpottaa sekä klusterien määrän valintaa että vaihtoehtoisten klusterointiratkaisujen tarkastelua. Stokastisen klusterointi ei kuitenkaan anna yksiselitteistä klusterointiratkaisua tai klusterien todennäköisyyksiä, vaan parittaiset todennäköisyydet alueiden esiintymisestä samassa klusterissa tietyllä hierarkian tasolla. Tässä työssä tulosten tulkinna apuna hyödynnetään referenssitodennäköisyyksiä, joissa tasojen todennäköisyysjakautumaa verrataan täysin satunnaisesti muodostettujen parittaisten todennäköisyyksien jakaumaan. Mahdollisena jatkotarkasteluna voisi olla deterministisen klusteroinnin tietyllä tasolla muodostaman klusterin kokonaistodennäköisyyden tarkastelu. Yksittäisten simulaatioiden tasolla stokastisen klusteroinnin tausta-ajatuksena on ratkaisun löytyminen silloinkin, kun aineistossa ei ole havaittavissa klusterirakennetta. Jatkossa olisikin hyvä pohtia, tarjoaisiko suurempi toistomäärä vielä luotettavampia tuloksia. Suurempi toistomäärä voisi myös vaatia klusterointialgoritmin optimointia suoritusajan lyhentämiseksi. Lisäksi klusterien yhdistäminen perustuu stokastisessa algoritmissa keskiarvoetäisyyksiin, minkä vuoksi muun muassa klusterien poikkeavat havainnot voivat vaikuttaa klusterien muodostumiseen. Algoritmia voitaisiin kehittää hyödyntämällä vaihtoehtoisia linkkifunktiota.

Aineiston näkökulmasta hyvinvointialueiden tarkastelu on suhteellisen karkealla tasolla ja todellisiin terveyseroihin kiinni pääsemiseksi alueellisia eroja olisi hyvä tarkastella hienojakoisemmalla tasolla. Esimerkiksi kuntatason analyysi voisi nostaa

esille merkittävää hyvinvointialueiden sisäistä vaihtelua. Suuret maantieteelliset etäisyydet sekä sosiaali- ja terveydenhuollon toimipisteiden sijainnit voivat vaikuttaa merkittävästi alueellisiin eroihin. Lisäksi aikaisemman sosiaali- ja terveydenhuollon kuntapohjaisen järjestelmän vaikutukset voivat näkyä luvuissa vielä pitkään tulevaisuudessa. Hienojakoisempi tarkastelu vaatisi myös stokastisen klusteroinnin algoritmin optimaalisuuden pohdintaa aineiston kasvaessa. Viimekädessä terveyserojen tutkimisessa tarvitaan kuntatasoakin tarkempaa aineistoa, sillä esimerkiksi kaupungeissa sisäiset erot voivat olla suuria. Tässä työssä on hyödynnetty ikävakioituja indeksitietoja. Mikäli tavoitteena olisi tutkia palvelutarpeen todellisia eroja, ikävakioimattomien indeksilukujen hyödyntäminen voisi myös olla hyvä ratkaisu. Lisäksi Kansallisen terveysindeksin tietoihin voivat jossain määrin vaikuttaa alueiden sairastavuuden kirjauskäytännöt. Klusteroinnin tulosten pohjalta ei voida päätellä sote-uudistuksen vaikutuksista terveyserojen kehitykseen, sillä aineisto koskee uudistuksen toimeenpanon ajankohtaa. Ajassa ilmenevien muutosten seuraaminen toistamalla analyysi esimerkiksi kymmenen tai kahdenkymmenen vuoden kuluttua voisi olla kiinnostavaa. Samalla on pidettävä mielessä, että terveyserojen kehitykseen vaikuttavia hallinnollisesta uudistuksesta riippumattomia tekijöitä on runsaasti.

Tässä työssä sairastavuuden ja työkyvyttömyyden alueellisia eroja tarkastellaan pääasiassa sosiaali- ja terveystalouden uudistuksen yhteydessä, mutta terveyserojen vähentämiseen tähtäävät toimenpiteet vaativat hallinnollisten uudistusten ohella poikkileikkaavaa yhteistyötä monella eri sektorilla. Sote-järjestelmän toimintaa voidaan tukea järjestelmän ulkopuolelta esimerkiksi vahvistamalla lasten ja nuorten terveyttä liikuntaharrastusmahdollisuuksilla ja koulutuksen toimenpiteillä (OPH, 2025). Kansanterveydellisestä näkökulmasta järkevää olisi sosiaali- ja terveydenhuollon resurssien kohdentaminen monialaiseen ennaltaehkäisevään työhön sekä toimiin, joista on koko väestön kannalta suurin terveyshyöty (Mikkonen, 2015). Tutkimusta tarvitaan sekä erilaisten terveystaloudellisten toimintalinjausten vaikuttavuudesta että päätösten taustalla vaikuttavista toimijoista, vallankäyttäjistä, tavoitteista, arvoista sekä ongelmien määrittelyn ja ratkaisujen oletuksista (Mikkonen, 2015). Hyvinvoinnin ja toimintakyvyn edistämisessä sekä terveyserojen kaventamisessa voi olla merkityksellistä myös luoda taiteiden, tieteiden ja ihmisten arjen kohtaamispaikkoja (Dow ym., 2023; Vuolasto ym., 2025; Mäenpää ym., 2025), hienovaraista yhteisöllisyyttä.

Viitteet

- Bryngelson, A. Long-term sickness absence and social exclusion. *Scandinavian Journal of Public Health*, 37(8):839–845, 2009. Saatavissa: <https://doi.org/10.1177/1403494809346871>.
- Dow, R., Warran, K., Letrondo, P. & Fancourt, D. The arts in public health policy: progress and opportunities. *The Lancet Public Health*, 8(2):e155–e160, 2023. Saatavissa: [https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667\(22\)00313-9/fulltext](https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(22)00313-9/fulltext).
- Elovainio, M. & Komulainen, K. Yksinäisyys, sosiaalinen eristyneisyys ja terveys.

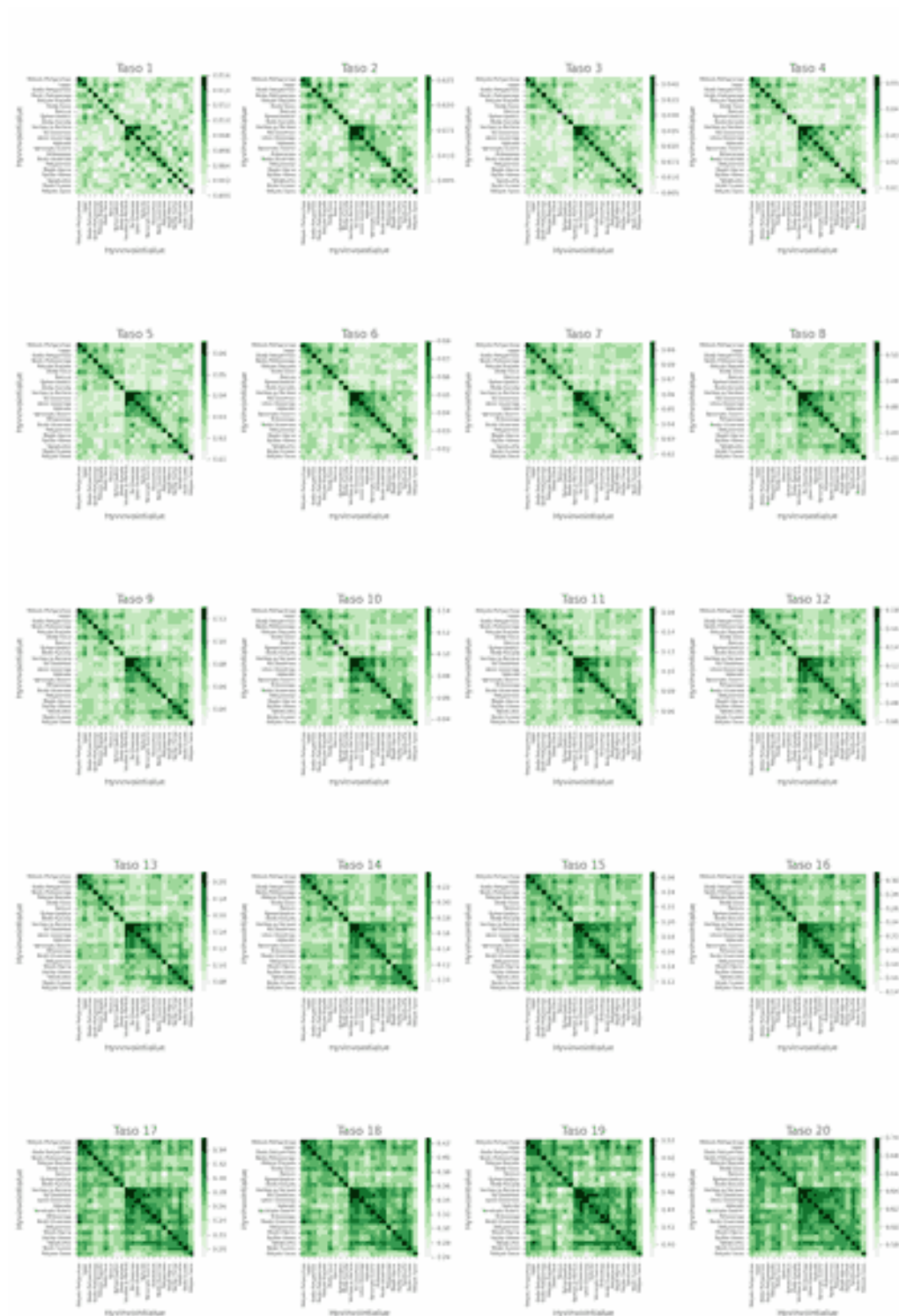
- Duodecim*, 139(7):571–576, 2023. Saatavissa: <https://www.duodecimlehti.fi/xmedia/duo/duo17628.pdf>.
- ETK, Eläketurvakeskus. Työeläkemaksut, 2021. Viitattu 15.1.2026. Saatavissa: <https://www.etk.fi/suomen-elakejarjestelma/rahoitus-ja-sijoitustointa/tyoelakemaksut/>.
- ETK, Eläketurvakeskus. Suomalaisia työkyvyttömyyseläkkeelle yhä vähemmän, 2025. Päivitetty 7.3.2025. Viitattu 15.12.2025. Saatavissa: <https://www.etk.fi/ajankohtaista/suomalaisia-tyokyvyttomyyselakkeelle-ya-vahemman/>.
- Everitt, B. S., Landau, S., Leese, M. & Stahl, D. *Cluster Analysis: 5th Edition*. John Wiley & Sons, Ltd, Chichester, 2011. 5. painos.
- Hair, J. F. Jr., Black, W. C., Babin, B. J. & Anderson, R. E. *Multivariate Data Analysis*. Cengage Learning EMEA, Hampshire, 2019. 8. painos.
- Harris, C. R., ym. Array programming with NumPy. *Nature*, 585:357–362, 2020. Saatavissa: <https://doi.org/10.1038/s41586-020-2649-2>.
- Harrison, R. L. Introduction to Monte Carlo simulation. Teoksessa *AIP Conference Proceedings*, volyymi 1204, sivut 17–21. American Institute of Physics, 2010. Saatavissa: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2924739/>.
- Haveri, A., Jalonen, H., Keskimäki, I., Saari, J., Tynkkynen, L.-K., Pekurinen, M., Kortelainen, M., Lyytikäinen, T., Ristikari, T. & Sorsa, V.-P. Hyvinvointialueuudistuksen väliarviointi, 2025. Valtioneuvoston julkaisuja 2025:109. Valtioneuvosto, Helsinki, 2025. Saatavissa: <https://urn.fi/URN:ISBN:978-952-383-022-6>.
- Heller, K. A. & Ghahramani, Z. Bayesian hierarchical clustering. Teoksessa De Raedt, L. & Wrobel, S. (toim.), *Proceedings of the 22nd International Conference on Machine Learning*, sivut 297–304. ACM Press, Bonn, 2005. Saatavissa: <https://dl.acm.org/doi/10.1145/1102351.1102389>.
- Hunter, J. D. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. Saatavissa: <https://doi.org/10.1109/MCSE.2007.55>.
- Jalonen, H. Sote-uudistus – kompleksisuusteoreettinen tulkinta. *Hallinnon tutkimus*, 39(4):302–309, 2020. Saatavissa: <https://journal.fi/hallinnontutkimus/article/view/102254>.
- Jordahl, K. geopandas/geopandas: v0.8.1, 2020. Saatavissa: <https://doi.org/10.5281/zenodo.3946761>.
- Kela, Kansaneläkelaitos. Työkyvyttömyysindeksi, 2025a. Päivitetty 17.9.2025. Viitattu 1.1.2026. Saatavissa: <https://tilastot.kela.fi>.

- Kela, Kansaneläkelaitos. Ammatillinen kuntoutus, 2025b. Päivitetty 3.11.2025. Viitattu 15.1.2026. Saatavissa: <https://www.kela.fi/ammattillinen-kuntoutus>.
- Kela, Kansaneläkelaitos. Sairastaminen, 2025c. Päivitetty 4.11.2025. Viitattu 14.12.2025. Saatavissa: <https://www.kela.fi/sairastaminen>.
- Kela, Kansaneläkelaitos. Kuntoutustuki ja työkyvyttömyyseläke, 2026. Päivitetty 15.1.2026. Viitattu 15.1.2026. Saatavissa: <https://www.kela.fi/kuntoutustuki-ja-tyokyvyttömyyselake#kuka-maksaa-kuntoutustukea-ja-tyokyvyttömyyselaketta>.
- Kestilä, L. Demografinen kehitys ja yhteiskunnan kestävyys. Teoksessa Kestilä, L. & Karvonen, S. (toim.), *Ratkaisuja kestävästä yhteiskunnan rakentamiseen: Väestön terveys- ja hyvinvointikatsaus 2025*, sivut 20–24. THL, Terveysten ja hyvinvoinnin laitos, Helsinki, 2025. Saatavissa: <https://urn.fi/URN:ISBN:978-952-408-444-4>.
- Koponen, P. & Tynkkynen, L.-K. Työterveyshuollon sairaanhoitopalvelut: Näkökulmia suomalaisen terveydenhuoltojärjestelmän oikeudenmukaisuuteen, 2023. Työpäperi 8/2023. THL, Terveysten ja hyvinvoinnin laitos, Helsinki, 2023. Saatavissa: <https://urn.fi/URN:ISBN:978-952-408-040-8><https://urn.fi/URN:ISBN:978-952-408-459-8>.
- Kortelainen, M., Kotakorpi, K. & Lyytikäinen, T. Hyvinvointialueiden rahoitusmallin kannustinvaikutukset. *Kansantaloudellinen aikakauskirja*, 117(2):203–210, 2021. Saatavissa: <https://www.doria.fi/handle/10024/182853>.
- Laaksonen, M., Perhoniemi, R. & Blomgren, J. Municipality-level differences in disability retirement in Finland: The contribution of local social characteristics. *Scandinavian Journal of Public Health*, 53(7):807–811, 2025. Saatavissa: <https://urn.fi/URN:NBN:fi-fe2024062457100>.
- Laki sosiaali- ja terveydenhuollon järjestämisestä 612/2021, § 1, 2021. Finlex. Viitattu 15.1.2026. Saatavissa: <https://www.finlex.fi/fi/lainsaadanto/2021/612>.
- Lindell, E., Koponen, P., Parikka, S., Sarttila, K., Luoto, R. & Sarnola, K. Kansallinen terveysindeksi 2021–2023: Sairastavuus ja työkyvyttömyys edelleen yleisintä Itä- ja Pohjois-Suomessa, 2025. Tilastoraportti 26/2025. THL, Terveysten ja hyvinvoinnin laitos, Helsinki, 2025. Saatavissa: <https://urn.fi/URN:NBN:fi-fe2025050637337>.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. *Multivariate Analysis*. Academic Press, London, 1979. 10. painos. 1995.
- McKinney, W., ym. Data structures for statistical computing in Python. Teoksessa *Proceedings of the 9th Python in Science Conference*, volyymi 445, sivut 51–56. Austin, TX, 2010. Saatavissa <https://doi.org/10.25080/Majora-92bf1922-00a>.

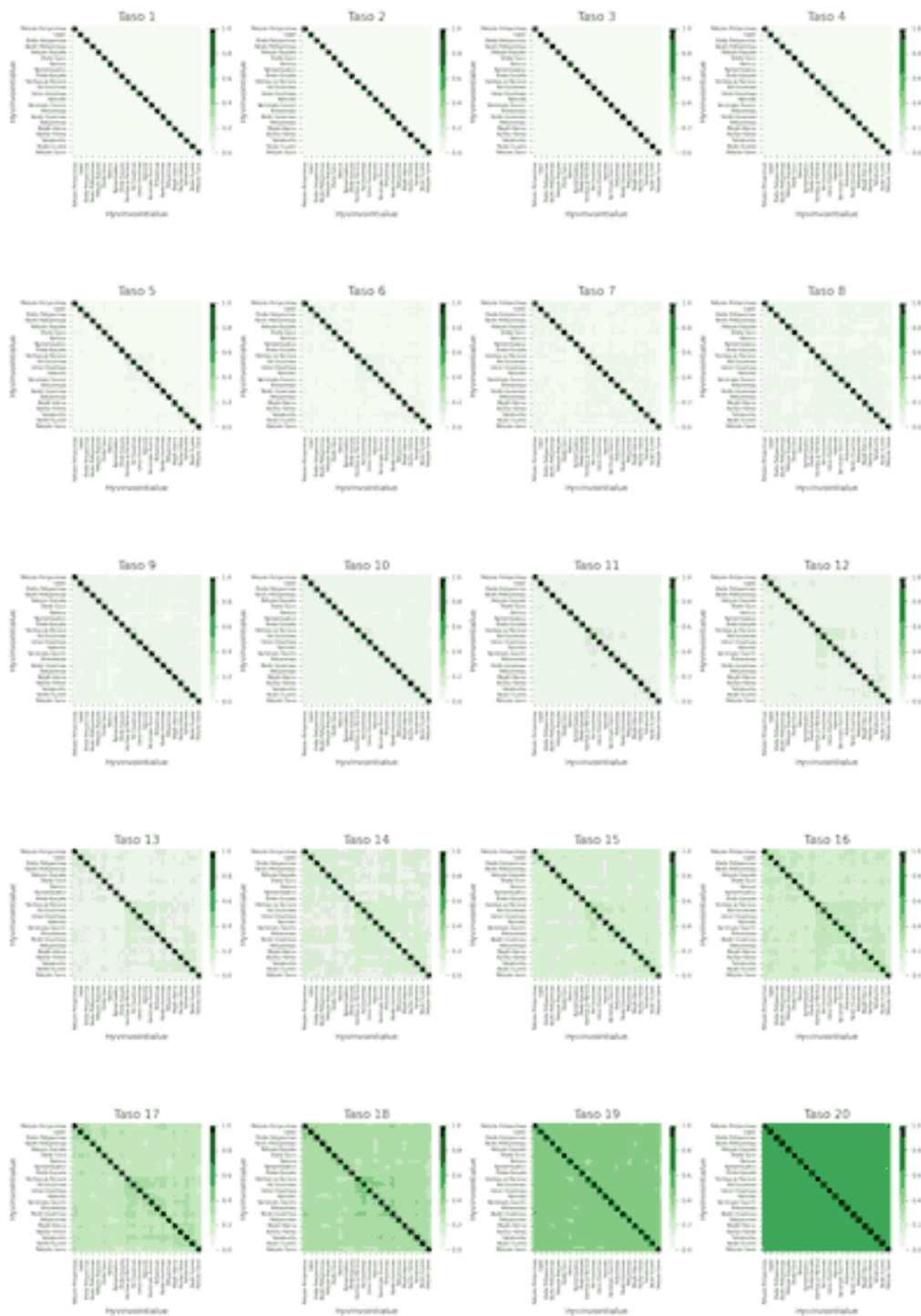
- Mikkonen, J. Terveyden tasa-arvon tulevaisuus, 2015. Briefing-paperi. Demos Helsinki, 2015. Saatavissa: <https://demoshelsinki.fi/publication/terveyden-tasa-arvon-tulevaisuus/>.
- Mäenpää, P., Faehnle, M., Grönlund, H., Pitkänen, K., Alanko, L., Sarkia, K. & Kuosmanen, S. Reaktionopeutta, paikallistietoa ja huolenpitoa: Paikallinen kansalaistoiminta vahvistaa yhteiskunnan pitkäjänteistä resilienssiä ja kriisivalmiutta, 2025. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2025:2. Valtioneuvoston kanslia, Helsinki, 2025. Saatavissa: <https://urn.fi/URN:ISBN:978-952-383-340-1>.
- OPH, Opetushallitus. Move! 2025: Koululaisten ja varusmiesten fyysisen toimintakyvyn myönteinen kehitys jatkuu – toimintakyvyssä alueellisia eroja, 2025. Päivitetty 11.12.2025. Viitattu 13.12.2025. Saatavissa: <https://www.oph.fi/fi/uutiset/2025/move-2025-koululaisten-ja-varusmiesten-fyysisen-toimintakyvyn-myonteinen-kehitys>.
- Owal Group Oy. Sosiaali- ja terveydenhuollon ja pelastustoimen uudistuksen ulkoinen arviointi: Loppuraportti, 2024. Sosiaali- ja terveysministeriön raportteja ja muistioita 2024:14. STM, Sosiaali- ja terveysministeriö, Helsinki, 2024. Saatavissa: <https://urn.fi/URN:ISBN:978-952-00-6876-9>.
- Pedregosa, F., ym. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. Saatavissa: <http://jmlr.org/paper/s/v12/pedregosa11a.html>.
- Rimpelä, M. & Laihonon, H. Ratkaisusta tuli ongelma: Tuliko hyvinvointialueista epäsiikioitä? *Yhteiskuntapolitiikka-lehti*, 90(1):58–66, 2025. Saatavissa: <https://urn.fi/URN:NBN:fi-fe2025021211740>.
- Saari, J. *Harvinainen hetki - Hyvinvointivaltio valinkauhassa*. Kunnallissalan kehittämissäätiö, Helsinki, 2024. Saatavissa: <https://kaks.fi/julkaisut/harvinainen-hetki-hyvinvointivaltio-valinkauhassa/>.
- STM, Sosiaali- ja terveysministeriö. Digitaalisuus sosiaali- ja terveydenhuollon kivijalaksi: Sosiaali- ja terveydenhuollon digitalisaation ja tiedonhallinnan strategia 2023–2035, 2023. Sosiaali- ja terveysministeriön julkaisuja 2023:32. STM, Sosiaali- ja terveysministeriö, Helsinki, 2023. Saatavissa: <https://urn.fi/URN:ISBN:978-952-00-9889-6>.
- STM, Sosiaali- ja terveysministeriö. Hyvinvointialueet vastaavat sosiaali- ja terveyspalvelujen ja pelastustoimen järjestämisestä, 2025. Päivitetty 28.11.2025. Viitattu 3.12.2025. Saatavissa: <https://stm.fi/hyvinvointialueet>.
- STM, Sosiaali- ja terveysministeriö. Sosiaali- ja terveydenhuollon järjestelmä ja vastuut, 2026. Päivitetty 15.1.2026. Viitattu 15.1.2026. Saatavissa: <https://stm.fi/sotepalvelut/jarjestelma-vastuut>.

- Terveydenhuoltolaki 1326/2010, § 2, 2010. Finlex. Viitattu 15.1.2026. Saatavissa: <https://www.finlex.fi/fi/lainsaadanto/2010/1326>.
- THL, Terveyden ja hyvinvoinnin laitos. Kansallinen terveysindeksi, 2025a. Päivitetty 8.10.2025. Viitattu 12.12.2025. Saatavissa: <https://thl.fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/kansallinen-terveysindeksi>.
- THL, Terveyden ja hyvinvoinnin laitos. Sotkanet – indikaattoripankki, 2025b. Päivitetty 5.5.2025. Viitattu 1.1.2026. Saatavissa: <https://sotkanet.fi/sotkanet/fi/index>.
- Tilastokeskus. Tilastokeskuksen hyvinvointialueet (WFS: SHAPE-ZIP), 2025. Avoin paikkatietoaineisto: kuntapohjaiset tilastointialueet – hyvinvointialueet, saatavilla WFS:llä. Valmis URL-osoite kartta-aineiston lataamiseen: <https://geo.stat.fi/geoserver/wfs?service=WFS&version=2.0.0&request=GetFeature&typeName=tilastointialueet:hyvinvointialue4500k&outputFormat=SHAPE-ZIP>.
- TTL, Työterveyslaitos. Työkyky, 2025. Viitattu 14.12.2025. Saatavissa: <https://www.ttl.fi/teemat/tyohyvinvointi-ja-tyokyky/tyokyky>.
- Tynkkynen, L.-K., Paatela, S., Aalto, A.-M., Keskimäki, I., Nykänen, E., Peltola, M., Sinervo, T., Tammi, T. & Viita-aho, M. Tilannekuvia hyvinvointialueilta: muutokset palvelujärjestelmässä sote-uudistuksen alkuvuosina, 2025. THL, Terveyden ja hyvinvoinnin laitos, Helsinki, 2025. Saatavissa: <https://urn.fi/URN:ISBN:978-952-408-459-8>.
- Virtanen, M., Heikkilä, K., Vahtera, J., Kivimäki, M., Halonen, J. I., Alexanderson, K., Rautiainen, S., Lallukka, T. & Mittendorfer-Rutz, E. Clustering of disability pension and socioeconomic disadvantage in Sweden: a geospatial analysis. *European Journal of Public Health*, 32(5):703–708, 2022. Saatavissa: <https://urn.fi/URN:NBN:fi-fe2022102462967>.
- Virtanen, P., ym. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. Saatavissa: <https://doi.org/10.1038/s41592-019-0686-2>.
- Vuolasto, J., Turpeinen, I., Hätönen, H., Norppa, L. & Seemer, S. Kulttuurihyvinvointitoiminnan kehittyminen ja ohjaus, 2025. OKM, Opetus- ja kulttuuriministeriön politiikka-analyysijä 2025:2. Opetus- ja kulttuuriministeriö, Helsinki, 2025. Saatavissa: <https://urn.fi/URN:ISBN:978-952-415-397-3>.
- Waskom, M., ym. mwaskom/seaborn: v0.8.1, 2017. Saatavissa: <https://doi.org/10.5281/zenodo.883859>.

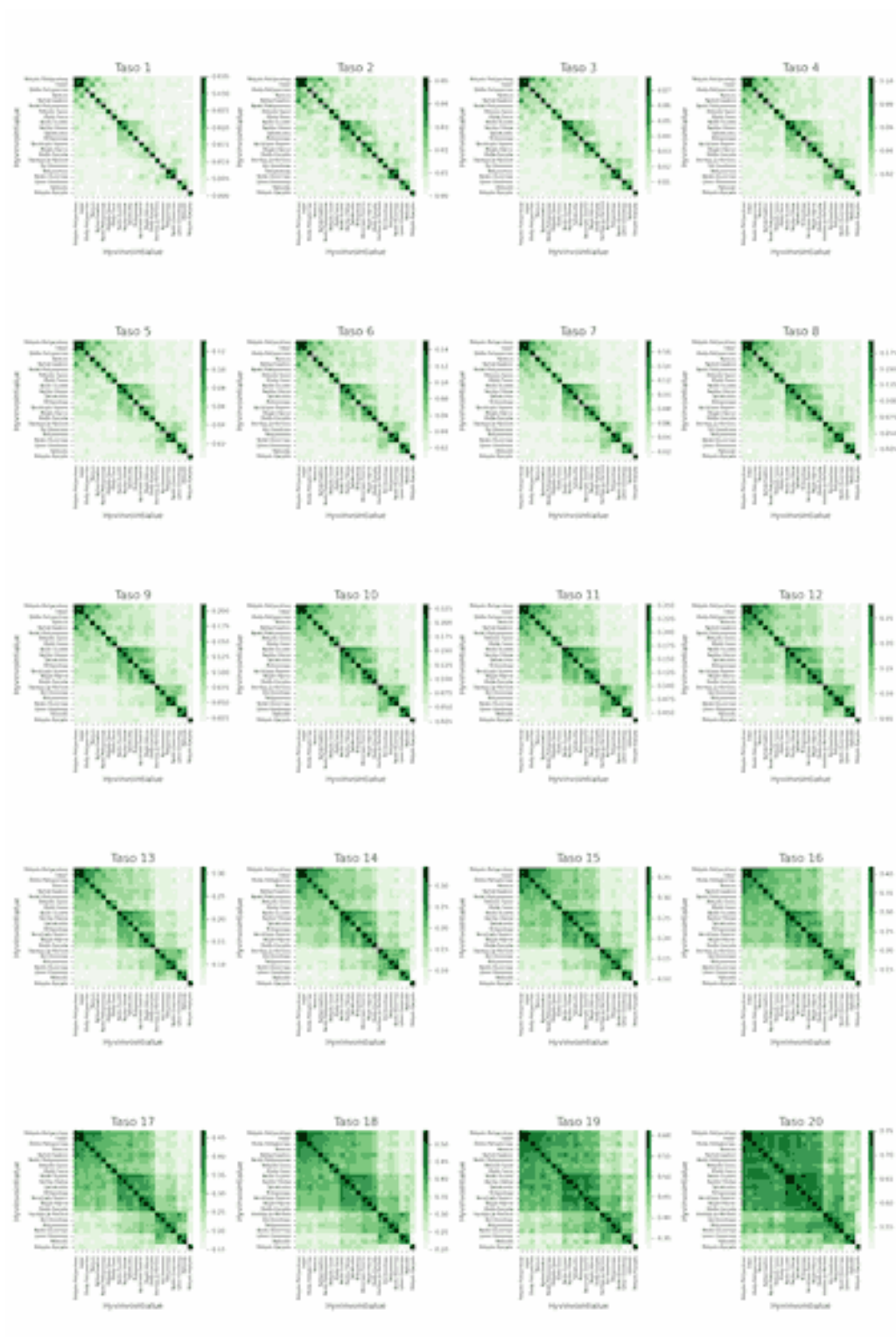
A Stokastisen klusteroinnin visualisoinnit



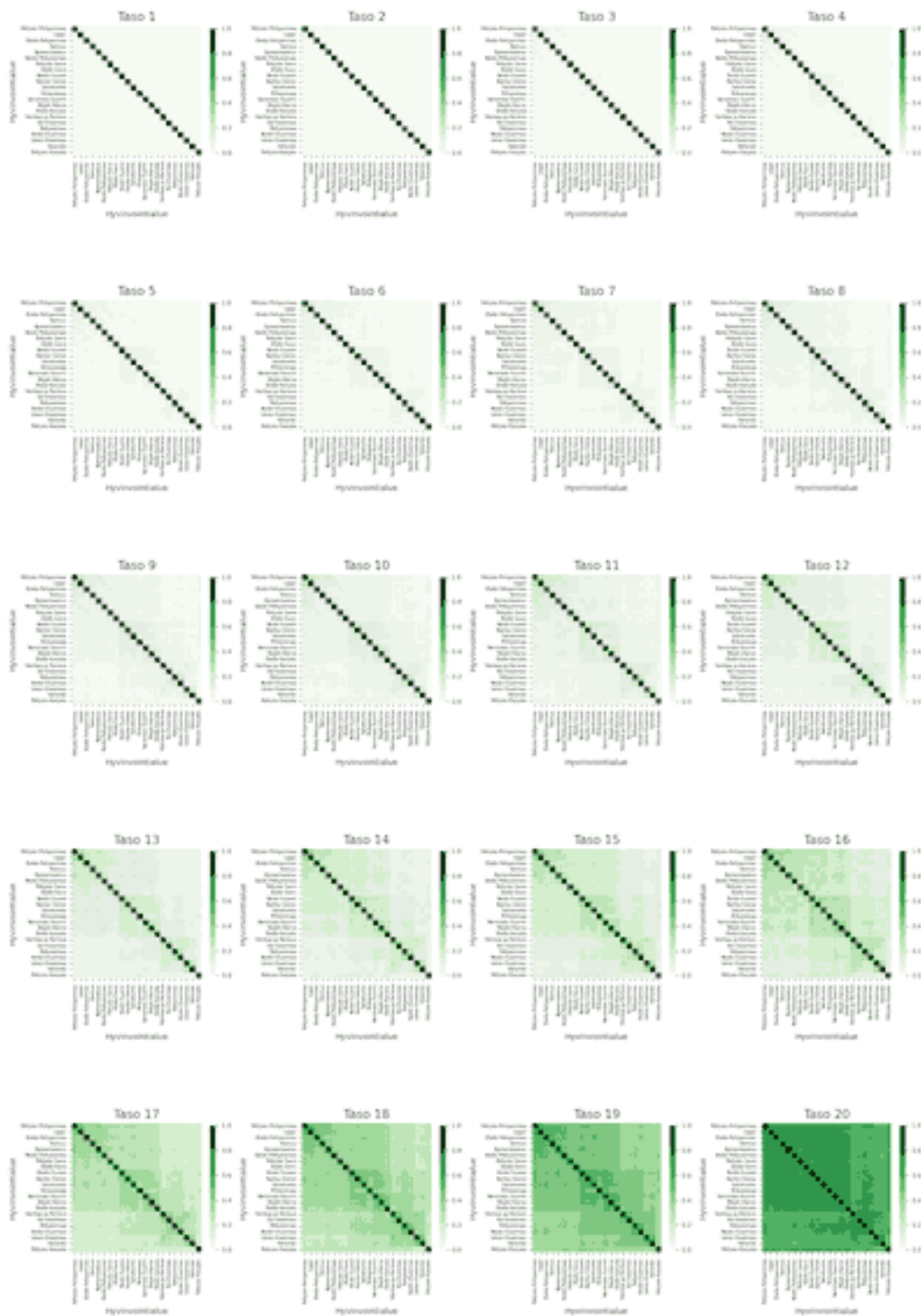
Kuva A1: Sairastavuuden stokastisen klusteroinnin tulokset vaihtuvilla asteikoilla.



Kuva A2: Sairastavuuden stokastisen klusteroinnin tulokset tasaisilla asteikoilla.



Kuva A3: Työkyvyttömyyden stokastisen klusteroinnin tulokset vaihtuvilla asteikoilla.



Kuva A4: Työkyvyttömyyden stokastisen klusteroinnin tulokset tasaisilla asteikoilla.

B Stokastisen klusteroinnin Python-koodi

```

1 def stokastinen_klusterointi(X, N=10000, seed=None):
2     if seed is not None:
3         np.random.seed(seed)
4
5     n = X.shape[0] # Alkioiden määrä
6     l = n - 1 # Hierarkian tasojen määrä
7     l_p = [np.zeros((n, n)) for _ in range(1)] # Todennäköisyydet eri tasoilla
8
9     etaisyysmatriisi = squareform(pdist(X, metric='euclidean')) # Alkioiden etaisyysmatriisi, alkuperäinen
10    ↪ etaisyysmatriisi
11
12    for _ in range(N):
13        klusterit = [[i] for i in range(n)] # Aloitetaan siitä, kun kaikki alkiot ovat omissa klustereissaan
14        l_indeksi = 0 # Aloitus aina tasolta 0
15
16        # Keskiarvolinkkifunktio
17        while len(klusterit) > 1: # Toistetaan, kunnes kaikki alkiot ovat samassa klusterissa
18            n_klusterit = len(klusterit) # Klusterien määrä
19            klusterietaisyydet = np.zeros((n_klusterit, n_klusterit)) # Alustetaan klusterietaisyyksien matriisi,
20            ↪ kaikki etaisyydet alussa nollia
21            for a in range(n_klusterit): # Lasketaan ryhmien väliset etaisyydet
22                for b in range(a+1, n_klusterit):
23                    alkioparit = np.array(np.meshgrid(klusterit[a], klusterit[b])).T.reshape(-1,2) # Kaikki
24                    ↪ alkioparit klusterien a ja b alkiosta
25                    klusterietaisyydet[a, b] = etaisyysmatriisi[alkioparit[:,0], alkioparit[:,1]].mean() # Keskiarvo
26                    ↪ alkioparien etaisyyksistä
27                    klusterietaisyydet[b, a] = klusterietaisyydet[a, b] # Täydennetään klusterietaisyyksien matriisi
28                    ↪ symmetria (ei sinänsä välttämätön)
29
30            # Etaisyydet todennäköisyyksiksi
31            with np.errstate(divide='ignore', invalid='ignore'): # Muutetaan etaisyydet
32                ↪ yhdistämistodennäköisyyksiksi, jos nollalla jakamista tms. ei virhettä
33            etaisyyksien_kaanteisluvut = 1 / (klusterietaisyydet + 1e-6) # Lasketaan klusterietaisyyksien
34            ↪ käänteisluvut ja varmistetaan, että ei jaeta nollalla
35            np.fill_diagonal(etaisyyksien_kaanteisluvut, 0) # Täydennetään diagonaaliin nollat, klusterin
36            ↪ etaisyydet itsestään on nolla
37            p = etaisyyksien_kaanteisluvut / etaisyyksien_kaanteisluvut.sum() # Lasketaan klusterien
38            ↪ yhdistämisen todennäköisyydet
39
40            # Yhdistettävän klusteriparin valinta
41            ylakolmio_indeksit = np.triu_indices(n_klusterit, k=1) # Poistetaan kaksoiskappaleet
42            p_flat = p[ylakolmio_indeksit] # Levitetään matriisi jonoksi, voidaan laskea kumulatiivinen summa
43            ↪ todennäköisyyksistä
44            p_kumulatiivinen = np.cumsum(p_flat) # Luodaan "todennäköisyyksien jana"
45            R = np.random.rand() * p_kumulatiivinen[-1] # R on oikeassa skaalassa
46            indeksi_flat = np.searchsorted(p_kumulatiivinen, R) # A:n ja B:n indeksit
47            A = ylakolmio_indeksit[0][indeksi_flat]
48            B = ylakolmio_indeksit[1][indeksi_flat]
49
50            # Uusi yhdistetty klusteri
51            uusi_klusteri = list(klusterit[A]) + list(klusterit[B]) # Yhdistetään klusterit
52            klusterit = [c for k, c in enumerate(klusterit) if k not in (A,B)] # Poistetaan klusterien listasta
53            ↪ klusterit indekseillä A ja B
54            klusterit.append(uusi_klusteri) # Lisätään uusi klusteri klusterien listaan
55
56            # Alustetaan jäsenyyssmatriisi tälle simulaatiolle
57            hierarkian_tason_jasenyydet = np.zeros((n, n))
58            for klusteri in klusterit: # Päivitetään tason tiedot
59                for i in klusteri:
60                    for j in klusteri:
61                        hierarkian_tason_jasenyydet[i, j] = 1 # Päivitetään, mitkä alkiot ovat samassa klusterissa
62
63            # Tallennetaan tason kumulatiiviseen matriisiin
64            l_p[l_indeksi] += hierarkian_tason_jasenyydet
65
66            l_indeksi += 1 # Seuraava taso
67
68    for indeksi in range(len(l_p)):
69        l_p[indeksi] = l_p[indeksi] / N # Jaetaan kumulatiivisten jäsenyyssmatriisien luvut toistokertojen määrällä,
70        ↪ jolloin saadaan todennäköisyydet alkioiden esiintymiselle samassa
71        ↪ klusterissa jokaisella tasolla
72
73    return l_p

```

C Kirjaus tekoälyn hyödyntämisestä

Tämän työn klusterianalyysin tuloksia kuvaavan luvun 4 taulukoiden muotoilussa ja visualisointien viimeistelyssä, sopivan karttapohja-aineiston etsimisessä sekä stokastisessa klusteroinnissa hyödynnettävän Python-funktion (Liite B) ohjelmoinnillisten ratkaisuvaihtoehtojen hahmottelussa on käytetty apuvälineenä kielimallia (OpenAI GPT-5). Kaikki analyysissä hyödynnettävät koodit, valmiit kuvaajat sekä taulukot on käyty läpi ja tarkistettu. Kielimallia ei ole käytetty tekstin työstämisessä.