Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan koulutusohjelma

Pekka Alli

# OECD-maiden työttömyyden klusterirakenteista

Diplomityö
Espoo, 7. syyskuuta 2016

Valvojat: Apulaisprofessori Pauliina Ilmonen, Aalto-yliopisto

Ohjaaja: Apulaisprofessori Pauliina Ilmonen, Aalto-yliopisto

Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

ABSTRACT OF
MASTER'S THESIS

**Author:** Pekka Alli

**Title:**
On cluster structures in unemployment of OECD-nations

**Date:** September 7, 2016 **Pages:** 84

**Major:** Systems and Operations Research **Code:** F3008

**Supervisors:** Assistant professor Pauliina Ilmonen, Aalto University

**Advisor:** Assistant professor Pauliina Ilmonen, Aalto University

The purpose of this thesis work is to study unemployment related cluster structures of OECD member countries. Hierarchical clustering methods are applied in the analysis. Previous studies have found geographical, lingual, and cultural clusters of OECD countries, when clustering was based on variables related to unemployment and other measures of welfare.

In this thesis, instead using of simple unemployment rates, we in addition consider different aspects of unemployment, including youth not in education, employment or training. We then perform cluster analysis using these several unemployment related variables. In the analysis, we have data from 27 different OECD member countries from years 2010 and 2014.

In addition to applying cluster analysis, this thesis gives a review of current clustering methods.

Cluster structures were found with every hierarchical clustering method we applied. The found clusters can be characterized using common regional labels such as Nordic Countries and Southern Europe. Other significant result is that the data contains some outliers. Greece, Turkey, Mexico, Israel and Spain have high enough unemployment figures to stand out clearly from the rest of the countries. The results of this thesis are more descriptive than determining; nevertheless the repeated results of different approaches speak in behalf of the stability of the results.

**Keywords:** Clustering, Agglomerative Hierarchical Clustering, Unemployment, OECD member countries

**Language:** Finnish

Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan koulutusohjelma

DIPLOMITYÖN
TIIVISTELMÄ

**Tekijä:** Pekka Alli

**Työn nimi:**
OECD-maiden työttömyyden klusterirakenteista

| | | | |
|---|---|---|---|
| **Päiväys:** | 7. syyskuuta 2016 | **Sivumäärä:** | 84 |
| **Pääaine:** | Systeemi- ja operaatiotutkimus | **Koodi:** | F3008 |
| **Valvojat:** | Apulaisprofessori Pauliina Ilmonen, Aalto-yliopisto | | |
| **Ohjaaja:** | Apulaisprofessori Pauliina Ilmonen, Aalto-yliopisto | | |

Tämän työn tarkoituksena on tutkia OECD-maiden työttömyyden klusterirakenteita käyttäen hierarkkisia klusterointimenetelmiä. Työttömyydestä sekä muista kansantalouden hyvinvoinnin mittareita on aikaisemmin löydetty klusteroitumista maantieteellisten, kulttuurillisten sekä kielellisten tekijöiden perusteella.

Tässä työssä työttömyyden käsitettä laajennetaan muihin työvoiman hyödyntämisen mittareihin ja nuorten työelämästä syrjäytymiseen, sekä etsitään 27:stä OECD-maasta klusterirakenteita koskien näitä tekijöitä. Työn aineisto on vuosilta 2010 ja 2014.

Klusterianalyysin suorittamisen lisäksi työssä keskitytään eri klusterointimenetelmiin.

Työn tuloksena on, että jokaisella työssä käytetyllä hierarkkisella klusterointimenetelmällä aineistosta on tunnistettavissa klusterirakenteita. Näitä tunnistettuja klusterirakenteita voidaan luonnehtia maantieteellisten tekijöiden perusteella kuten Pohjoismaat ja Etelä-Eurooppa. Muuta merkittävää tuloksissa on, että otosmaat sisältävät muutamia poikkeavia havaintoja (Kreikka, Turkki, Meksiko, Israel, Espanja), joiden työttömyys- ja työelämästä syrjäytymisluvut ovat huomattavasti erillään muista maista. Työn tulokset ovat suuntaa-antavia enemmän kuin määrittäviä, mutta tulosten stabiilisuutta vahvistaa eri menetelmillä saadut samankaltaiset tulokset.

| | |
|---|---|
| **Asiasanat:** | Klusterointi, Hierarkkinen klusterointi, Työttömyys, OECD-maat |
| **Kieli:** | Suomi |

# Sisältö

# Luku 1

# Johdanto

## 1.1 Työttömyys OECD-maissa

Vuonna 2007 alkaneen maailmanlaajuisen talouskriisin seurauksena työttömyys on ollut korkealla lähes kaikkialla maailmassa. Työttömyys ja työelämästä syrjäytyminen on kohdistunut suurissa osin nuoriin ja onkin aivan ennätyslukemilla. Tämä tekee työttömyydestä mielenkiintoisen tutkimuskohteen.

Aikaisemmassa tutkimuksessa OECD-maista on löydettävissä valtioiden rajoja rikkovia alueita, joissa monet kansantalouden mittarit, kuten työttömyysaste, yhtenevät.

Työttömyysaste ei yksinään kuvaa maan ekonomista tilaa työvoiman kannalta, vaan se jättää huomioimatta monia työttömyyden osa-alueita. Tämä työ tarkastelee useita työvoiman hyötykäyttöön liittyviä mittareita.

Työttömyydelle selittäviä tekijöitä ja teoreettisia lähtökohtia on useita. Monet teoriat ovat todettu vanhentuneiksi, eikä yhtä selittävää lähtökohtaa löydy. Työttömyyden tunnistaminen on tärkeää niin kansantalouden teorian kuin hyvinvoinnin näkökulmasta.

## 1.2 Klusterointianalyysin tavoite ja aineisto

Tämän työn tavoitteena on tutkia OECD-maiden työttömyyden klusterirakenteita. Työssä tutkitaan useita työttymyyden, työvoiman hyödyntämisen sekä työelämästä syrjäytymisen mittareita ja etsitään niissä esiintyviä klusterirakenteita. OECD tulee sanoista The Organisation for Economic Co-

operation and Development. OECD-maat koostavat ryhmän, jonka tarkoituksena on tukea ja edistää taloudellista kehitystä.

Klusterointianalyysiä sovelletaan kahteen eri aineistoon: työttömyysmittareihin sekä nuorten työelämästä syrjäytymiseen. Näitä kahta aineistoa tarkastellaan erillään sekä vertaillaan, onko näiden lukujen klusteroitumisella yhtäläisyyksiä.

Klusterirakenteita etsitään vuosien 2010 ja 2014 aineistoista. Näiden vuosien välisiä eroja ja yhtäläisyyksiä tutkitaan ja pyritään selittämään tapahtuneita rakenteellisia muutoksia. Tässä työssä tehdään klusterirakenteiden pohjalta tulkintaa maiden yhtäläisyyksistä ja selitetään, mitä klusterointitekniikat tuottavat. Jos mahdollista, tunnistettuja rakenteita luokitellaan tunnetuilla maa ryhmiä kuvaavilla käsitteillä.

Työttömyyttä rajaudutaan tutkimaan vain itseään kuvaavilla muuttujilla. Työhön ei valita ulkoisia selittäviä tekijöitä, joilla tavanomaisesti yritettäisiin selittää selitettävää muuttujaa. Sen sijaan, työttömyydestä yritetään löytää sisäisiä rakenteita, joiden perusteella työttömyyttä OECD-maissa voidaan kuvailla ja asettaa jatkokysymyksiä aiheen tutkimista varten.

Klusterointimenetelmät ja niiden teoria on työn keskeisessä osassa ja menetelmissä rajaudutaan vain koviin klusterointimenetelmiin.

## 1.3 Klusterointimenetelmät

Tämän työn menetelminä ovat klusterointimenetelmät. Työn klusterianalyysiluvussa tehdään katsaus tämän hetkisiin klusterointimenetelmiin ja vertaillaan erilaisia lähestymistapoja klusteroinnin ongelmaan.

Erilaisten klusterointitapojen lisäksi työssä syvennytään klusteroinnissa keskeiseen asioihin: mittauspisteiden erilaisuuteen ja metriikkoihin.

Työssä tehdyn menetelmäkatsauksen perusteella tehdään valinta sovellettavasta klusterointialgoritmista. Sovelletavaa algoritmin implementaatiota etsitään R-ohjelmointikielen kirjastosta ja käytetään tätä valittuun aineistoon. Klusterointianalyysin pohjalle suoritetaan aineistoon yksi- ja useampiulotteinen tilastollinen analyysi pohjustamaan klusterointia.

## 1.4 Työn rakenne

Työn ensimmäinen luku on johdanto, jossa esitellään työn menetelmät, taustа, tavoitteet, rajaukset sekä rakenne.

Toisessa luvussa perehdytään työttömyyteen ja työvoiman hyötykäyttöön. Tässä luvussa perehdytään työttömyyden kansantaloudelliseen pohjaan ja tämän hetkiseen työttömyyden tilanteeseen ja lähihistoriaan OECD-maissa. Tässä luvussa myös tarkastellaan tutkittuja klusterirakenteita OECD-maissa, mitkä koskevat kansantaloudellisia mittareita ja työttömyyttä.

Kolmas luku kertoo klusterointimenetelmistä. Kappale alkaa katsauksella klusterointiin ylätasolla: mitä klusteroinnilla haetaan ja minkälaisia matemaattisia ominaisuuksia tarvitaan asioiden saavuttamiseen. Tämän jälkeen syvennytään tämän hetken menetelmiin ja lähestymistapojen eroavaisuuksiin sekä siihen miten eri menetelmät sopivat sovelluskohteisiin. Osiossa luodaan syvempi teoreettinen perusta menetelmään, joka valitaan työn empiiriseen osuuteen.

Neljäs luku esittelee työttömyyden aineistoa, joka on kerätty OECD-maista. Aineistossa on yhteensä kymmenen eri muuttujaa ja 27 eri maata. Aineisto on kerätty vuosilta 2010 ja 2014.

Viidennessä luvussa sovelletaan kolmannessa luvussa valittua menetelmään esiteltyyn aineistoon. Kappale sisältää lopputuloksena saatujen klusterien esittelyn, tulkinnan ja vertailun aiemmin esitettyyn kirjallisuuteen.

Kuudennessa luvussa tehdään kriittistä pohdintaa saatujen tulosten pohjalta ja esitetään jatkokysymyksiä kuinka tutkimusta voisi viedä eteenpäin ja mitä merkitystä saaduilla tuloksilla on.

Viimeinen, seitsemäs, luku sisältää yhteenvedon koko työstä, sen tuloksista ja haasteista.

# Luku 2

# Työttömyys ja työelämästä syrjäytyminen

## 2.1 Työttömyyden kehitys ja kategorisointi

### 2.1.1 Työttömyys OECD-maissa

Maailmanlaajuinen talouden notkahtaminen vuosien 2007-2009 aikana teki suuren vaikutuksen työttömyyslukuihin läpi maailman. Työttömyys oli ennätyskorkealla ja Euroopassa joidenkin maiden kohdalla kriisin seuraukset näyttävät enemmän pysyviltä vaikutuksilta kuin poikkeustilalta. Kriisin vaikutukset näkyvät eritoten eurooppalaisten nuorten ja matalasti kouluttautuneiden työttömyydessä. Vuonna 2009 Euroopassa oli jopa 9 miljoonaa 15-24 vuotiasta työtöntä, ja maissa joissa tilanne oli pahimmillaan, nuorten työttömyysaste kohosi yli 50 prosenttiin. [1]

Työttömyys OECD-maiden välillä on vaihdellut suuresti jo 60-luvulta lähtien ja vielä 2000-luvulla OECD-maissa on tunnistettu matalan ja korkean työttömyyden polarisaatiota [2]. Talouskriisin myötä nämä maiden väliset erot ovat korostuneet. Näin ollen heikossa taloudellisessa tilanteessa olevien valtioiden on ollut haastavaa saada työttömyystilannetta kuriin.

### 2.1.2 Työttömyys käsitteenä

Työttömyys ilmiönä kuvaa maan tai ekonomisen alueen kansantaloudellista tilaa työvoiman näkökulmasta. Työttömyyttä pidetään yhtenä tärkeimmistä kansantaloudellisista mittareista. Mittarina työttömyys ei kuitenkaan ole

aivan yksiselitteinen, vaan se pitää sisällään monia asioita, joita esimerkiksi yleisesti käytetty työttömyysaste ei tuo esiin. [3]

Työttömiksi lasketaan ne ihmiset jotka ovat yli 15 vuotiaita, joilla ei ole työpaikkaa, kuuluvat työvoimaan ja ovat etsineet töitä. Työttömiksi lasketaan myös ne jotka eivät saa työstään rahallista korvausta tai eivät työskentele itselleen. Työvoimaan kuuluvien joukko määritellään joukoksi, jotka ovat käytettävissä palkkatyöhön tai ovat itsensä työllistäviä mittausajankohtana. [4]

Myös muita työllistymisen mittareita käytetään kuin työttömyysasteetta. Matalasuhdanteen aikaan tarjolla olevien työtuntien määrä vähenee, jolloin osa-aikatyöllisten määrä kasvaa. Tämän nähdään myös yhtenä työvoiman hyödyntämisen mittarina. Osa-aikatyöllisyys voi olla myös toivottua. Esimerkiksi opiskelijat, jotka eivät ehdi työskennellä täysipäiväisesti, voivat olla omasta tahdostaan osa-aikatyössä. Tällöin ehdotetaankin, että omasta tahdosta riippumaton osa-aikatyöllisyys olisi nousussa kun talous kohtaa matalasuhdanteen. USA:ssa vuonna 2009 puolet osa-aikatyöllisistä ilmoitti olevansa osa-aikatyössä kansantaloudellisista syistä. Omasta tahdosta riippumaton osa-aikatyöllisyys voidaan nähdä työttömyyden mittarina, mutta tätä mittaria on haastavaa erottaa vapaaehtoisesta osa-aikatyöllisyydestä. [3]

Vapaaehtoisen osa-aikatyöllisyyden seurauksena nähdään myös työvoiman hyödyntämisen suuruusluokan väärin arvioiminen, sillä työllisyysaste nousee vaikka tehtyjen työtuntien määrä ei kasva samassa suhteessa kuin täysipäiväisten työntekijöiden tapauksessa. Tällöin työttömyysasteen kasvu ei tuo samanlaista talouskasvua kuin täysipäiväisten työntekijöiden tuoma kasvu ja näin ollen taloudelliset seuraukset ovat väärin arvioituja. [3]

On olemassa osajoukko ihmisistä jotka kuuluvat marginaalisesti työvoimaan, mutta eivät kuulu tilastollisten määritelmien mukaan. Nämä ovat ihmisiä, jotka ovat olleet työttöminä pitkän aikaa ja lopettaneet työn etsimisen, eikä heitä enää lasketa työvoimaan, vaikka he olisivatkin valmiita työn tekoon tilaisuuden tullen. Tätä osuutta ei lasketa työttömiin, sillä työttömiin kuuluu vain työtä hakevien joukko. Tätä osuutta kutsutaan piilotyöttömyydeksi. Näistä osuutta, jotka eivät hae töitä työhön liittyvistä syistä kutsutaan lannistuneiksi työntekijöiksi. Tämä johtuu yleisesti negatiivisista odotuksista liityen työhön ja työnhakuun. [3]

Näin ollen on tärkeää puhua muustakin kuin työttömyysasteesta, kun tarkastellaan ekonomista työvoiman hyödyntämisen astetta. Työttömyysaste onkin vain yksi käsite, joka ei kuvaa työvoiman tehokkuutta täydellisesti. [5]

### 2.1.3 Työttömyyden mallintaminen

Työttömyydestä tilastollisena ilmiönä ja sen mallista on useita kiisteltyjä näkökulmia. Menetelmät, kuten Phillipsin käyrä ja NAIRU ovat näyttäneet suuntaviivoja työttömyyden kehityksessä, mutta näitä menetelmiä pidetään kuitenkin nykyään epätarkkoina. [6]

Tilastollisena ilmiönä työttömyyttä voidaan tarkastella kahdelta kantilta: työttömyys on stationäärinen ilmiö tai työttömyydellä on hystereesi vaikutus. [7]

Hystereesivaikutus työttömyydessä tarkoittaa sitä, että poikkeaminen luonnollisesta asapainosta aiheuttaa vastareaktion tasapainoon palaamiseen, jolloin suuri hetkellinen työttömyys aiheuttaa pysyvän muutoksen työttömyyden tasapainoon. Tällöin työttömyyttä tarkastellessa satunnaismuuttujana tarkoittaa kyseinen tulkinta sitä, että satunnaismuuttujan arvoon vaikuttaa siihen mennessä kuljettu polku. [7]

Hystereesiteorian mukaiset seuraukset olisi Euroopassa synkät, sillä nykyinen korkea työttömyyden tilanne aiheuttaisi pysyvän nousun työttömyyden tasapainotasoon.

Stationäärinen työttömyys tarkoittaa sitä, että työttömyyden jakauma on kaikkialla sama, eikä jakauma ole ajasta tai paikasta riippuvainen. Tällöin korkeat työttömyysasteet ovat vain poikkeamia, jotka palaavat ajallaan takaisin perustasolleen. Tällöin työttömyyteen vaikuttaa vain hetkellisesti kansantaloudelliset ilmiöt ja makroekonomiset päätökset, mutta hetkelliset muutokset palaavat kuitenkin takaisin kohti tasapainotilaa. Tällöin tasapaino tilaan vaikuttaa vain edellisen jakson tasapainotila eikä hetkittäiset muutokset. [7]

Työttömyys voidaan myös jakaa rakenteelliseen ja sykliseen työttömyyteen. Rakenteellisen teorian pohjalla on oletus, että työttömyys saavuttaa sille ominaisen tason pitkän ajan kuluessa jos ulkopuolisia shokkivaikutuksia ei ole. Tällöin työttömyyden aste määräytyy muiden kyseistä taloutta koskevien mittareiden perusteella. Näitä tekijöitä ovat mm. inflaatio, verotuspolitiikka, sekä useat sisä- ja ulkopoliittiset tekijät. Myös kysynnän rakenteellinen muuttuminen kehityksen edetessä johtaa rakenteelliseen työvoiman kysynnän ja tarjonnan kohtaamattomuuteen. Monissa rakenteellisissa työttömyysteorioissa lyhyen ajan kausittaiset vaihtelut halutaan poistaa. Kausittaiset vaihtelut voidaan poistaa tilastollisilla menetelmillä, kuten Kalmansuotimella. NAIRU menetelmä on eräs tunnetuista rakenteellisista tilastollisista työttömyyden mallintamisen menetelmistä. Ominaista rakenteellisille työttömyysmalleille on, että tarkkaa työttömyyttä ei yritetä mallintaa, vaan luoda lä-

hestymistapa, jolla tarkastellaan työttömyyden rakenteita. [7],[3]

Eräs rakenteellisen työttömyyden malli ehdottaa, että inflaatio olisi työttömyyden rakenteellisena aiheuttajana, eli inflaation noustessa työttömyys paranee, jolloin ilman rahan arvon laskua ei voi olla täydellistä työllisyyttä. Nämä inflaatioon perustuvat teoriat saavat erilaista empiiristä tukea, jos verrataan lukuja esimerkiksi Amerikan ja Ranskan välillä. [8]

Syklinen työttömyyden teoria tarkoittaa työttömyyden luonnollista kausittaista vaihtelua. Kausittainen vaihtelu nähdään monesti heijastuvan työvoiman kysynnän kausittaisesta vaihtelusta, joka johtaa esimerkiksi työvoiman kysynnän ja tarjonnan kohtaamattomuuteen. Kausittaiselle työttömyydelle löytyy useita tekijöitä, joilla vaihtelua yritetään selittää. [9], [3]

### 2.1.4 Työttömyyden seuraus

Työttömyys jättää ison jäljen talouteen. Kun työtä ei ole tarjolla, suuri osa ihmisistä ei tuota mitään, vaan tarvitsevat tukea valtiolta. Tällä on vaikutus valtion kykyyn myötävaikuttaa maailman talouteen. Työttömyyden jälki ei rajoitu pelkästään talouteen, vaan nuorena koettu työttömyys aiheuttaa pitkän aikavälin vaikutukset supistaen eliniän ansaitsemisen odotusarvoa sekä nostaen tulevaisuuden työttömyyden todennäköisyyttä. Tällä nähdään olevan negatiivisia seurauksia elämänlaadun kanssa. Korkean työttömyyden vallitessa monet ovat loukussa alipalkatuissa työpaikoissa siinä pelossa, että he eivät saa koulutustaan vastaavaa työtä. Usko työmarkkinoihin katoaa ja tämä luo negatiivisen kuvan työelämästä. [1]

Työttömyys vaikuttaa eritavoin sen kestosta riippuen. Pidempikestoista työttömyyttä pidetään vakavampana, koska tällä nähdään olevan pysyvämpiä seurauksia kohteena olevan elämään. Työttömyyden seurauksena on monesti elämänlaadun lasku, sillä työttömyyden suorana seurauksena on rahavirtojen pysähtyminen, vaikka ihmisen kulutuksentarve ei poistu. Tällä nähdään yhteyksiä moniin ongelmiin, kuten mielialaongelmiin, rikollisuuteen sekä yleiseen terveyteen ja elämänlaatuun. [10]

## 2.2 Työelämästä ja koulutuksesta syrjäytytyminen

Kuten aiemmassa luvussa todettiin, viime vuosikymmenen lopun talouskriisin tuoma työttömyys on koetellut rajusti eurooppalaisia nuoria. Nuorten

työttömyys ja työelämästä syrjäytyminen on yksi tärkeä työttömyyden osa-alue, jolla on vakavia pitkän ajan seurauksia. Jotta tätä ilmiötä voidaan hoitaa, on se ensin tunnistettava. Tätä varten on kehitetty käsite NEET.

NEET tulee sanoista Not in Education, Employment or in Training. Täysin työelämästä ja kouluttautumisesta ulos jääneet nuoret. Tämä on tärkeä mitta, sillä nuorten työttömyysaste ei sisällä osuutta nuorista, jotka ovat kokonaan syrjäytynyt ulos työelämästä. NEET koskettaa ainoastaan nuoria ja yleisesti määritelmänä on ikäryhmä 15-24-vuotiaat. NEET on alumperin Iso-Britanniassa keksitty määritelmä, joka on myöhemmin tullut käyttöön muihin maihin kuten Japaniin ja Kiinaan. [11], [12]

NEET-tilastoa pidetään vähemmän harhaisena kuin yksistään nuorten työttömyystilastoa, sillä työttömyystilasto ei ota kantaa nuorten osallistumisesta työvoimaan. [13]

Nuorten työttömyyttä selitetään monesti sillä, että nuorten tulevaisuuden potentiaalia liioitellaan kuluihin nähden. Näin ollen on helpompaa antaa nuorten hakea turvaa perheestään ja selvitä tällä tavoin työttömyydestä. Nuoret ovat kokemattomia, joten työnantajalla on riski palkatessaan nuori kokematon työntekijä. Tämä johtaa kierteeseen, jossa nuoret jäävät ilman työkokemusta eikä palkkaussyyt parane. Muita syitä ilmiölle ovat samoja kuin yleiselle työttömyydelle: osaamisen ja tarpeen kohtaamattomuus sekä taloudelliset shokit. [11]

Kirjallisuudessa esitetään, että NEET-ilmiötä on ollut vaikea hahmottaa. Konsensusta, joka ohjaa nuoren siirtymisvaihessa aikuisuuteen väärälle polulle, ei ole ollut ja määritelmät ovat olleet puutteellisia, joten ilmiöön on ollut hankala puuttua [12].

Keskeisenä ongelmana nuorten työelämästä syrjäytymisessä on pitkän ajan vaikutukset, jotka kantautuvat kokonaistyöttömyyteen ja työhön osallistuvien asteeseen. Tästä johtuen NEET itsessään on mielenkiintoinen ja tärkeä tutkimuksen kohde, sillä nuoreen kohdistuvat negatiiviset vaikutukset ovat pitkäkestoisia ja täten niillä on vakavat tulevaisuuden vaikutukset. Nuorten työelämästä syrjäytymisellä on todettu olevan vaikutusta seitsemän vuoden päähän aikuisena työllisyyteen osallistumiseen [12].

## 2.3 OECD-maiden klusterirakenteista

Kirjallisuudessa usein esiintyviä maiden luokittelumalleja ja -ryhmiä käytetään jatkuvasti osana maan identiteettiä. Luokittelut, kuten Pohjoismaat ja

Baltia ovat vakiintuneita käsitteitä. Nämä käsitteet sisältävät implisiittisesti oletuksia maiden politiikasta ja taloudesta.

Tutkimus [14] koostaa aikaisempaa tutkimusta OECD-maiden klusterirakenteista useiden eri kansantalouden mittarien perusteella. Osassa tutkimuksia tarkesteltavina muuttujina ovat olleet muun muassa työttömyysaste ja bruttokansantuote. Näissä tutkimuksissa on todettu, että maantieteellinen klusteroituminen on enimmillään heikkoa. Kuitenkin sama tutkimus [14] ehdottaa, että eri vuosikymmenillä klusteroituminen olisi selkeämpää ja että OECD-maissa nähtävistä klustereista pystytytään tunnistamaan tekijöitä kuten yhteinen kieli tai yhteinen kulttuuri. Aiemmassa työssä tunnistettuja rakenteita ovat englanninkieliset alueet, Skandinavia, Manner-Eurooppa ja Etelä-Eurooppa. Nämä löydökset on saavutettu käyttämällä kokoavaa hierarkkista klusterointia. [14]

Klusteroitumista vastaan puhuu Euroopan yhtenäistyminen ja yhtenäistynyt politiikka. Liitot kuten Euroopan unioni ja yhtenäinen rahapolitiikka nähdään tekijöinä, jotka yhdistävät maiden kansantaloudellista tilaa. [14]

Työttömyyden klustereita tutkivat tutkimukset löytävät Euroopasta maiden rajoja rikkovia alueita, joissa työttömyystilanteissa on yhteneväisyyksiä. Yhtenä huomiona kuitenkin nähdään, että maiden sisällä alueelliset erot ovat suuria. Esimerkkinä Italian Campania, jossa työttömyysaste oli vuonna 1996 4,4 kertaa niin suuri kuin Italian Valle d'Aossa. Tämä ero on samaa luokkaa kuin pienimmän ja suurimman työttömyysasteen ero OECD-maiden välillä. Työttömyyden klusterien tunnistamisen seurauksena pystyttäisiin muuttamaan työttömyyden korjausten politiikkaa vaikuttamaan yli maiden rajojen.[15]

# Luku 3

# Klusterianalyysi

Tässä luvussa tehdään katsaus yleisimpiin klusterointimenetelmiin, klusteroinnin tulosten arvioimiseen sekä kyseisille menetelmille tärkeään käsitteeseen, pisteiden etäisyyteen ja samanlaisuuteen. Kappaleen lopussa esitetään valinta työn aineistoon sovellettavasta menetelmästä.

## 3.1 Klusterointi

Klusterointi on yläkategoria luokittelumenetelmille, joiden tarkoituksena on koota aineistosta osa-joukkoja, klustereita, siten, että osa-joukkojen sisällä olisi mahdollisimman paljon samanlaisuutta ja osa-joukkojen välillä mahdollisimman paljon erilaisuutta. Klusteroinnin tavoitteena on kuvata aineiston sisäisiä rakenteita ottamatta kantaa ulkopuolisiin selittäviin tekijöihin. [16]

Klusterointi luokitellaan ohjaamattomiin oppimisalgoritmeihin. Tämä tarkoittaa sitä, että lopputuloksena syntyviä klustereita ei etukäteen tunneta, toisin kuin ohjattavissa oppimisalgoritmeissa, kuten klassifioinnissa ja diskriminanttianalyysissä. Klusterointi menetelmien tarkoituksena on enemmän kuvailla kuin määritellä aineistoa. [16], [17]

Klusterointi on merkittävä menetelmä tekoälyn, tiedonlouhinnan, koneoppimisen sekä tilastotieteen tieteenaloilla. Käytännön sovelluksia klusteroinnille löytyy myös lukuisista sovellusaloista kuten biologiasta, markkinoinnista sekä kansantaloudesta. Näin ollen myös mielipiteet eri klusterointimenetelmistä jakautuvat sovelluskohteen mukaan, eikä kirjallisuudesta löydy yhtä oikeaa tapaa klusteroinnille. [18]

Klusterointimenetelmät voidaan luokitellaan sillä perusteella, mitä ominai-

suutta menetelmä mittaa, sekä kuinka kuvataan asioiden erilaisuutta. Jako voidaan tehdä myös mittausten lähestymistavan perusteella. Asioiden erilaisuudella tarkoitetaan esimerkiksi yksittäisten avaruuden pisteiden välistä erilaisuutta tai klusterien välistä erilaisuutta. Tässä kontekstissa on hyvä puhua erilaisuudesta, eikä etäisyydestä, sillä etäisyysmetriikoiden lisäksi klusteroinnissa käytetään samanlaisuusfunktioita kuvaamaan pisteiden erilaisuutta. Mitattavia asioita pisteiden välisten erilaisuuksien lisäksi on esimerkiksi pistejoukkojen tiheys. [19]

## 3.2 Aineiston esikäsittely

Kuten monille tilastollisenanalyysin menetelmille, myös klusteroinnille on yleistä aineiston esikäsittely. Yleensä aineisto halutaan standardoida. Tämä tarkoittaa aineistoon kohdistuvaa affiinimuunnosta, jossa muunnetun aineiston keskiarvo on nolla ja varianssi tai absoluuttinen keskipoikkeama on yksi. Absoluuttinen keskipoikkeama ei ole yhtä altis poikkeaville havainnoille kuin keskihajonta, jolloin hajonta ei ole aivan niin suurta kuin kyseinen mittari voisi antaa olettaa [20].

Standardoinnin tarkoituksena on samanarvoistaa kaikki muuttujien ominaisuudet, jotta ominaisuuksien väliset skaalaerot eivät hallitsisi. Tämä tosin on tarkoituksenmukaista vain, jos aineiston muuttujien väliset skaalaerot halutaan hävittää. Tässä työssä skaalaeroista halutaan eroon ja aineisto standardoidaan ennen klusterointimenetelmien soveltamista.

Tässä työssä aineiston standardoinnissa keskiarvo skaalataan nollaan sekä varianssi yhteen.

Standardointoi pisteelle $x_{ij}$ määritellään kaavalla

$$z_{ij} = \frac{x_{ij} - m_j}{\sigma_j}, \tag{3.1}$$

jossa $m_j$ on ominaisuuden j keskiarvo sekä $\sigma_j$ on ominaisuuden j otoskeskihajona. Muuttuja $z_{ij}$ on uusi kuvattu arvo muuttujan $x_i$ ominaisuudelle $j$. Otoskeskiarvo määritetään seuraavasti:

$$m_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{3.2}$$

ja otoskeskihajonta määritetään seuraavasti:

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - m_j)^2}. \tag{3.3}$$

[19]

## 3.3 Etäisyysfunktio

Klusteroinnissa tarkoituksena on yhdistää samanlaiset pisteet ja erottaa erilaiset pistejoukot. Tällöin suuren merkityksen lopputulokseen muodostaa pisteiden välinen erilaisuusmitta. Asioiden välinen erilaisuus on arkinen asia, mutta sen määrittäminen tarkasti on todella vaikeaa. Tälle ei löydy absoluuttista ratkaisua, mutta useita vakiintuneita lähestymistapoja. Tämä johtaakin siihen, että pisteiden väliselle erilaisuudelle ei ole oikeaa mittaa, vaan valinta metriikasta tai samanlaisuusfunktiosta on tehtävä kontekstista riippuen.

Metriikalle eli etäisyysfunktiolle on olemassa määritelmät, siitä mitä mittareiden on täytettävä, jotta asiayhteydestä voidaan tehdä metrisessä avaruudessa päteviä jatko päätelmiä.

Tässä työssä kahden vektorin $x_i$ ja $x_j$ välistä etäisyyttä merkitään $d_(x_i, x_j)$. Metriikka eli etäisyysfunktio $d$ joukolle $x_i \in X$ määritellään kuvauksena

$$d : X \times X \to [0, \infty]. \tag{3.4}$$

Avaruuden metriikalle on määritetty neljä ehtoa täytettäväksi:

1. Etäisyys on aina nolla tai positiivinen

$$d(x, y) \geq 0, \quad \forall x, y \in X. \tag{3.5}$$

2. Etäisyys on nolla vain ja ainoastaan pisteen itsensä kanssa

$$d(x, y) = 0 \Leftrightarrow x = y \quad \forall x, y \in X. \tag{3.6}$$

3. Etäisyysfunktio on symmetrinen

$$d(x, y) = d(y, x) \quad \forall x, y \in X. \tag{3.7}$$

4. Etäisyys täyttää kolmioepäyhtälön

$$d(x, z) \leq d(x, y) + d(y, z) \quad \forall x, y, z \in X.[21] \tag{3.8}$$

Etäisyysmetriikka voidaan määrittää erikseen numeerisille muuttujille, binäärimuuttujille, nominaalisille muuttujille sekä järjestyslukumuuttujille. [16]

Numeeristen arvojen, eli määritelmäjoukko on $X = \mathbb{R}^{\prime}$, yleistetty metriikka määritellään Minkowskin etäisyydellä:

$$d_g(x_i, x_j) = (\sum_{k=1}^{k=p} |x_{ik} - x_{jk}|^g)^{1/g}, \quad x_i, x_j \in \mathbb{R}^p. \tag{3.9}$$

Minkowskin etäisyys sisältää määritelmän Euklidiselle etäisyydelle ($L^2$) kun g=2 ja Manhattanin etäisyydelle kun g=1 ($L^1$).

Klusterointianalyysissä Minkowskin etäisyyttä käytettäessä aineisto on hyvä standardoida jos ei haluta, että etäisyys riippuu mitattavien ominaisuuksien suuruusluokasta. [16].

Minowskinin etäisyys voidaan myös määrittää painotettuna, jos muuttujan eri ominaisuuksille halutaan luoda eriarvoisuutta:

$$d_g(x_i, x_j) = (\sum_{k=1}^{k=p} w_k|x_{ik} - x_{jk}|^g)^{1/g}, \quad x_i, x_j \in \mathbb{R}^p, w_k \in [0, \infty). \tag{3.10}$$

Binäärimuuttuja on muuttuja joka saa arvon 0 tai 1. Binääriarvoisten vektorien etäisyysmetriikka esitetään seuraavasti:

$$d_(x_i, x_j) = \frac{r + s}{q + r + s + t}, \tag{3.11}$$

jossa q on ominaisuuksien määrä, jotka ovat 1 kummallekin vektorille, t on niiden ominaisuuksien lukumäärä jotka ovat 0 kummallekin vektorille ja s ja r ovat vastaavat luvut, mutta jotka ovat erisuuria vektorien välillä [16].

Nominaalisille muuttujille esitetään kaksi erilaista lähestymistapaa etäisyyden määrittämiseen:

$$d_(x_i, x_j) = \frac{p - m}{p}, \tag{3.12}$$

jossa p on dimensioiden lukumäärä vektorissa $x_i$ ja m on ominaisuuuksien lukumäärä, joiden arvot ovat samoja. [16]

Toinen lähestymistapa on luoda nominaalisista muuttujista binäärimuuttujia, jossa jokainen binäärimuuttuja edustaa tiettyyn kategoriaan kuulumista ja binäärimuuttujien etäisyys lasketaan kaavan 3.11 mukaan.

Järjestysluvuille esitetään seuraavanlaista lähestymistapaa. Vektori $x_i$ standardoidaan muutujaksi $z_i$, jossa n:s komponentti määritellään seuraavasti:

$$z_{in} = \frac{x_{in} - 1}{M_n - 1}, \tag{3.13}$$

jossa $M_n$ on järjestyslukujen oletettu yläraja. Kaavassa 3.13 on oletettu, että järjesyslukujen alaraja on yksi. Standardoinnin jälkeen etäisyyden voi laskea käyttäen muuttujaa $z_i$ sekä edellä esitettyä Minowskin etäisyyttä. [16].

Etäisyysmetriikka voidaan määrittää myös pisteiden välille, jossa vektorien eri ominaisuudet eivät välttämättä ole samantyyppisiä tai kaikkien pisteiden kaikki ominaisuudet eivät ole tiedossa.

$$d(x_i, x_j) = \frac{\sum_{k=1}^{p} \delta_{ij}^k d'(x_{ik}, x_{jk})}{\sum_{k=1}^{p} \delta_{ij}^k}, \quad x_i, x_j \in \mathbb{R}^p, \tag{3.14}$$

jossa $d'$ on etäisyysmetriikka vektorien välille joka määritellään binäärimuuttujille:

$$d'(x_{ik}, x_{jk}) = 0, \tag{3.15}$$

jos $x_{ik} = x_{jk}$, muulloin $d'(x_{ik}, x_{jk})$ on yksi binäärimuutujien tapauksessa.

Jatkuville muuttujille:

$$d'(x_{ik}, x_{jk}) = \frac{x_{ik} - x_{jk}}{max_h x_{hk} - min_h x_{hn}}, \tag{3.16}$$

jossa h käy pisteiden kaikki arvot ominaisuudelle k.

Järjestyslukuiselle muuttujille ensin muuttuja skaalataan yhtälön 3.13 avulla ja toimitaan samalla tavalla kuin jatkuvan muuttujan tapauksessa yhtälössä 3.16.

## 3.4 Samanlaisuusfunktiot

Toinen vastaava käsite, kuin pisteiden välinen etäisyys on samanlaisuusfunktio. Samanlaisuusfunktiot eivät välttämättä täytä metrisen avaruuden määritelmiä, jotka on esitetty kaavoissa 3.5, 3.6, 3.7 ja 3.8. Samanlaisuusfunktioille kuitenkin olennaista on, että ne ovat symmetrisiä ja funktioiden arvo kasvaa, kun pisteet ovat lähempänä toisiaan. [16]

Merkitään pisteiden välistä samanlaisuutta funktiolla $s(x_i, x_j)$.

Yksi esitetty tapa on mitata pisteiden samanlaisuutta on vektorien välisen kulman kosini. Tämä määritellään pistetulon avulla:

$$s(x_i, x_j) = \frac{x_i^T \cdot x_j}{||x_i|| \cdot ||x_j||}. \tag{3.17}$$

Tällöin vektorien suunnalla on enemmän väliä kuin vektorien pituudella.

Toinen vastaava samanlaisuuden mitta on Pearsonin korrelaatiomitta:

$$s(x_i, x_j) = \frac{(x_i - \bar{x}_i)^T \cdot (x_j - \bar{x}_j)}{||x_i - \bar{x}_i|| \cdot ||x_j - \bar{x}_j||}. \tag{3.18}$$

Pearsonin korrelaatio mittaa muuttujien välistä lineaarista riippuvuutta. Tämä arvo on välillä [-1, 1], jossa 1 tarkoittaa täydellistä lineaarista riippuvuutta, 0 ei korrelaatiota ja -1 täydellistä negatiivista lineaarista riippuvuutta. [16]

Binäärisille vektoreille toimii Jaccardin indeksi, joka mittaa joukkojen leikkauksen suhdetta joukkojen unioniin. Jaccardin mitta voidaan yleistää muotoon:

$$s(x_i, x_j) = \frac{x_i^T \cdot x_j}{||x_i||^2 + ||x_j||^2 - x_i^T \cdot x_j}. \tag{3.19}$$

[16]

## 3.5 Klusterointimenetelmät

Klusterointimenetelmät voidaan jakaa usealla tavalla eri kategorioihin. Eräs karkea jako on pehmeä ja kova klusterointi. Kovassa klusteroinnissa jokainen piste kuuluu tasan yhteen klusteriin, kun taas pehmeässä klusteroinnissa pisteen kuulumista klusteriin kuvataan yleensä luvulla joka on välillä [0, 1], jolloin piste voi kuulua useaan klusteriin. Pehmeät klusterointimenetelmät ovat uudempia, ja tälle alalle tehtyä tutkimusta on huomattavasti vähemmän kuin koviin menetelmiin [22]. Pehmeät klusterointimenetelmät tarjoavat erilaisen lähestymistavan klusterointiin, joka on hyvin tärkeä, sillä mustavalkoisuus asioiden kuulumisesta vain ja ainoastaan yhteen joukkoon rajoittaa aineistosta mallien tunnistamisen mahdollisuuksia.

Tässä työssä tarkastellaan vain kovia klusterointimenetelmiä. Kovat klusterointimenetelmät ovat hyvin tunnistettuja ja tutkittuja [22]. Kovat klusterointimenetelmät määritellään seuraavasti:

Olkoon $X$ pisteiden joukko joka halutaan klusteroida. Jokainen klusteri esitetään osajoukkona $C_k$. Osajoukoille $C_k$ pätee

$$S = \cup_k C_k, \tag{3.20}$$

ja

$$C_i \cap C_j = \emptyset \quad \forall i \neq j. \tag{3.21}$$

Klusterointimenetelmät voidaan jakaa myös seuraaviin kategorioihin: osittavat menetelmät, hierarkkiset menetelmät, tiheyteen perustuvat menetelmät, tilastolliseen malliin perustuvat menetelmät sekä ruudukkoperustaiset menetelmät [19].

### 3.5.1 Osittavat menetelmät

Osittavissa menetelmissä aineisto jaetaan $k$ kappaleeseen osajoukkoja, jotka edustavat omaa klusteriaan. Menetelmässä klusterien lukumäärä on $k \leq n$, jossa $n$ on aineiston pisteiden määrä. Osittamismenetelmillä on ominaista, että menetelmä kehittää aluksi $k$ kappaletta osajoukkoja, joita menetelmä muokkaa iteratiivisesti, siten että muodostuneiden klusterien sisäinen samanlaisuus on mahdollisimman suuri ja klusterien välinen samanlaisuus mahdollisimman pieni. [19] [17]

Monet osittavat menetelmät käyttävät heuristiikkoja saavuttaakseen lopputuloksen nopeasti, sillä yleinen aineiston osittaminen halutulla tavalla on laskennallisesti raskas prosessi. Tämä prosessi on NP-täydellinen tehtävä, eli ratkaisuaika kasvaa eksponentiaalisesti suhteessa aineiston kokoon. Tämä kombinatorinen haaste on helposti nähtävissä siitä, että joukko, jossa on $N$ kappaletta pisteitä, voidaan jakaa k-lukumäärään joukkoja $k^N$:llä eri tavalla. Tällöin eri vaihtoehtojen läpikäyminen ja parhaimman vaihtoehdon löytäminen on todella aikaavievää. Heuristisista menetelmistä K-means-menetelmä on tunnettu esimerkki.

K-means-menetelmässä valitaan klusterien lukumäärä $k$ ja klusterien alkusijaintia kuvaavat vektorit. Jokaisella iteraatiolla klusteriin $k_j$ kuuluu ne pisteet joille kyseinen klusterin keskipiste on lähin klusteri. Eli datapiste $x_i$ kuuluu klusteriin $arg\min_j d(x_i, k_j)$. Jokaisen klusterin sijainti päivitetään klusteriin kuuluvien pisteiden keskiarvoksi. Tätä iteraatiota toistetaan kunnes menetelmä konvergoi. K-means menetelmälle on lopputuloksen kannalta hyvin tärkeää, minkälainen metriikka on valittu pisteiden väliseksi etäisyydeksi. [23],[19]

K-means-menetelmien heikkoutena nähdään vahva riippuvuus alkuratkaisuun ja etukäteen klusterien lukumäärän valitseminen. Alkuratkaisusta riippuvuus monesti ratkaistaankin jalostamalla menetelmää esimerkiksi käyttämään useita satunnaisia alkuarvoja. Sama pätee myös klusterien määrän valitsemiseen. [17]

### 3.5.2 Hierarkkiset menetelmät

Hierarkisessa klusteroinnissa lopputuloksena on puu sisäkkäisistä klustereista. Tätä klusteri puuta kutsutaan dendrogrammiksi. Dendrogrammin solmut ovat aineiston osajoukkoja ja lehdet yksittäisiä datapisteitä. Solmukohdat kuvaavat aineiston yhdistämistä tai hajottamista algoritmin suunnasta riippuen. Solmukohtien korkeudet kertovat jakaantuvien osajoukkojen erilaisuuden, jolloin algoritmin kulun voi lukea dendrogrammista kulkemalla solmukohdat ylhäältä alaspäin. [24] [17] [25]

Hierarkkinen klusterointi jaetaan kahteen alalajiin: jakavat ja yhdistävät menetelmät. Yhdistävässä lähestymistavassa aluksi jokainen piste on oma klusterinsa, joista iteraatiokierrosten edetessä muodostetaan isompia klustereita kunnes kaikki pisteet ovat yhdessä klusterissa. [24]

Jakavassa lähestymistavassa aluksi kaikki pisteet kuuluvat yhteen klusteriin, jota aletaan purkamaan pienempiin klustereihin. [24]

Hierarkkisessa klusteroinnissa valitaan pisteiden välinen etäisyysmetriikka sekä klusterien välinen samanlaisuusfunktio. Kummassakin, jakavassa ja yhdistävässä, lähestymistavassa klusterien yhdistäminen ja purkaminen perustuu valittuun erilaisuusmetriikkaan.

Hierarkkista klusterointia pidetään klusteroinnin teoreettisenä pohjana. [26]

### 3.5.3 Yhdistävä hierarkkinen klusterointi

Hierarkkinen yhdistävä klusterointi koostaa klustereita pienemmistä klustereista.

Merkitään $s_1, s_2, s_3$..:lla jokaista klusterien yhdistämiskerran samanlaisuutta. Yhdistävän klusteroinnin olettamuksena on että, $s1 \geq s2 \geq s3 \geq s4$... Tällöin klusterointi algoritmia kutsutaan monotoniseksi. Algoritmi on epämonotoninen, jos on olemassa $s_i > s_{i+1}$.

Merkitään klustereiden joukkoa $\Omega$:lla

$$\Omega = \{\omega_1, \omega_2, ...\omega_K\}, \tag{3.22}$$

jossa $\omega_i$ kuvaa yhtä klusteria.

Yhdistelmien samanlaisuus määritellään klusteroinnille

$$s(\Omega) = \min_k s(\omega_k), k \in [1, K]. \tag{3.23}$$

Klusterointi $\Omega^{'}$ on optimaalinen jos kaikille klustroinneille $\Omega$, joissa on enintään K klusteria, klusteroinnin samanlaisuus on pienempi.

eli

$$|\Omega'| \leq |\Omega|, k \leq K, \tag{3.24}$$

kun $\Omega'$:ssa on k kappaletta klustereita ja $\Omega$:ssa K kappaletta.

Klusteroinnille voidaan määrittää erilaisia klusterin sisäisiä samanlaisuusmittoja:

$$s_{single-link}(\omega) = \min_{\omega' \subset \omega} \max_{x_i \in \omega'} \max_{x_j \in \omega - \omega'} d(x_i, x_j) \tag{3.25}$$

, jossa $\omega'$ ja $\omega - \omega'$ ovat klusterin $\omega$ osajoukkoja.

$$s_{complete-link}(\omega) = \min_{\omega' \subset \omega} \min_{x_i \in \omega'} \min_{x_j \in \omega - \omega'} d(x_i, x_j) \tag{3.26}$$

$$s_{gaac}(\omega) = 1/(n_A n_B) \sum_{x_i \in \omega_A} \sum_{x_j \in \omega_B} d(x_i, x_j) \tag{3.27}$$

Voidaan todistaa induktiolla, että $s_{single-link}$ tuottaa optimaalisen tuloksen yllä mainitun monotonisuuskriteerin näkökulmasta.

Induktion perusaskel K=N, jolloin jokainen piste on oma klusterinsa $s(\Omega)$=1, joka on optimaalinen määritelmän mukaan, sillä $s \in [0, 1]$.

Oletetaan, että $\Omega_K$ on optimaalinen klusterointitapa klusterien lukumäärällä K $|\Omega_K| \geq |\Omega_{K'}|$.

Klusterointi $\Omega_{K-1}$ saadaan aikaiseksi yhdistämällä kaksi klusteroinnin $\Omega_K$ samanlaisinta klusteria. Tehdään vastaoletus, että löytyy klusterointi $\Omega'_{K-1}$ joka saadaan eri polulla klusteroineja $\Omega'_K, \Omega'_{K-1}$ ja että $|\Omega'_{K-1}| \geq |\Omega_{K-1}|$. Todistus jakautuu kahteen osaan:

1. Pisteet, jotka ovat linkitetty klusteroinnissa $\Omega_{K-1'}$ ovat samassa klusterissa klusteroinnissa $\Omega_K$. Ne voivat olla linkitetty, vain jos niiden yhdistäminen, jonka samanlaisuus on pienempi kuin $s(\Omega'_{K-1})$, on tapahtunut ennen klusterointia $\Omega_K$. Tällöin $s(\Omega'_{K-1}) \geq s(\Omega_K) \geq s(\Omega'_K) \geq s(\Omega'_{K-1})$, joka on johtanut ristiriitaan.

2. Pisteet, jotka linkitettiin klusteroinnissa $\Omega_{K-1'}$ eivät ole samassa klusterissa klusteroinnissa $\Omega_K$. Kuitenkin $s(\Omega'_{K-1}) > s(\Omega_{K-1}$, jolloin single-link-samanlaisuusfunktion perusteella viimeisen askeleen pitäisi yhdistää nämä prosessoidessa $\Omega_K$:ta. Tämä johtaa ristiriitaan, joten $\Omega_{K-1}$ on optimaalinen.

Samanlaisuusfunktioille $s_{compelete-link}$ ja $s_{gaac}$ ei vastaavasti päde, jolloin klusterointi ei ole tässä mielessä optimaalinen. Käytännön kannalta tämä on monesti väärä optimaalisuuden mitta. [25]

Yhdistävän hierarkkisen klusterointialgoritmin voi kuvata algoritmina seuraavasti:

**Data:** $x_i \in X \subset \mathbb{R}^p$
Jokainen piste $x_i \in X$ on oma klusterinsa;
Laske etäisyysmetriikat jokaisen datapisteen välille $d(x_i, x_j)$ ;
**while** *Useampi kuin yksi klusteri jäljellä* **do**
  Laske jokainen joukon välinen erilaisuusmetriikka $d(A, B)$, jossa A ja B ovat eri klustereita. ;
  Yhdistä klusterit joiden välinen erilaisuusmetriikka on pienin. ;
**end**

Edellä esitetyt klusteroinnin samanlaisuusfunktiot voidaan tulkita yksinkertaisemmin algoritmin valitsemis vaiheeseen:

Minimietäisyys klusterien välillä

$$d_{min}(A, B) = \min \quad d(x_i, x_j) \quad x_i \in A, x_j \in B. \tag{3.28}$$

Tällöin klusterien yhdistämisvaiheessa yhdistetään klusterit, joiden lähimmäiset pisteet ovat lähimpinä toisiaan.

Maksimietäisyys klusterien välillä

$$d_{max}(A, B) = \max \quad d(x_i, x_j) \quad x_i \in A, x_j \in B. \tag{3.29}$$

Tälläin klusterien yhdistämisvaiheessa yhdistetään klusterit, joiden kauimmaiset pisteet ovat lähimpänä toisiaan.

Keskimääräinen etäisyys klusterien välillä

$$d_{ave}(A, B) = 1/(n_A n_B) \sum_{x_i \in A} \sum_{x_j \in B} d(x_i, x_j), \tag{3.30}$$

jossa $n_A$ ja $n_B$ ovat pisteiden lukumäärä joukoissa A ja B. Klusterit, joiden keskimääräiset pisteiden väliset etäisyydet ovat lähimpänä tosiaan, yhdistetään.

Eräs tapa yleistää klusterien yhdistämisvaiheen päätös kohdefunktion avulla on Wardin menetelmä. Tarkemmin sanottuna Wardin minimivarianssimenetelmä, jossa tarkoituksena on optimoida tehtävään soveltuvaa kohdefunktiota. Esimerkkinä tämänlaisesta kohdefunktiosta on minimoida klusterien sisäistä varianssia. Juuri tämä menetelmä kuvaa hyvin klusteroinnin tarkoitusta, sillä klusteroinnin toinen päätavoite on, että joukkojen sisäinen samanlaisuus olisi mahdollisimman suurta.

Yhdistävän hierarkkisen klusterointimenetelmän tuloksena saadaan dendrogrammi. Dendrogrammista voidaan lukea pisteiden kuuluminen klustereihin ja se kuinka pienemmät klusterit ovat muodostaneet suurempia klustereita. Dendrogrammi-puusta voidaan nähdä tulokset eri abstraktiotasoilla. Puusta nähdään muodostuvatko isot klusterit isoista pistejoukoista vai onko kyseessä esimerkiksi chaining effect. Chaining effect tarkoittaa tässä yhteydessä tilanetta, jossa klusterien yhdistämisvaiheessa isoon klusteriin yhdistyy toistuvasti peräkkäin yksipisteinen klusteri.

Kokoamismetriikan valitseminen vaikuttaa suuresti lopputulokseen. Erityisesti minimietäisyyskokoamistavalla on ominaista ns. chaining effect, eli klustereiden yhdistäminen yhden paikallisen etäisyyden perusteella. Aineiston jakaminen klustereihin tuottaa monesti klustereita, jossa on vain yksi datapiste. Tämä on huonoa käytännön kannalta vaikka menetelmä onkin optimaalinen monotonisuus-kriteerin suhteen [24]. Maksimietäisyysmetriikka kärsii herkkyydestä aineistossa olevien poikkeavia havaintoja kohtaan.

Hierarkkisessa klusteroinnissa on merkittävää koko algoritmin kulku ja näin ollen yksi päätös puun rakentamisessa vaikuttaa suuresti koko lopputulokseen eikä algoritmilla ole tapaa mennä taaksepäin ja korjata tilannetta. Tämä perustuu monotonisuusolettamisukseen, jota rikotaan monessa eri algoritmin toteutuksessa. Tämän seurauksena on kritisoitavaa, että puun haarautuessa tasapelitilanteessa, eli algoritmi on indifferentti päätöksestä, se tekee kuitenkin merkittävän päätöksen, jota se ei pysty korjaamaan. Kuitenkin tätä varten hierarkkiseen klusterointiin voidaan suorittaa keskiarvoistetuilla etäisyyksillä tai Wardin etäisyydellä, jotta vainta tehostuisi ja tasapelien määrä vähenisi. [27]

Hierarkkinen klusterointi ei tarvitse etukäteen klusterien lukumäärää, mutta tulosten tulkintaa varten tarvitaan dendrogrammista jokin lukumäärä klustereita. Eräs tapa on leikata dendrogrammista solmukohdat joista lähtevä viiva on tiettyä lukuarvoa pidempi, eli tällöin klusterien yhdistämisen samanlaisuus vastaa jotakin ennalta määritettyä tasoa.

Toinen vaihtoehto on minimoida kohdefunktiota

$$K = arg \min_{K}^{'} E(\Omega'_K) + \lambda K', \tag{3.31}$$

jossa E on jokin virhefunktio kuvaamaan klusterien jakautumista. Kuten esimerkiksi 3.39 ja $\lambda$ on sakkovakio klusterien lukumäärälle.

### 3.5.4 Tiheyteen perustuvat menetelmät

Klusterien muodostaminen voi perustua muuhunkin kuin pisteiden tai klusterien väliseen etäisyyteen. Yksi vaihtoehtoinen tapa on koota klustereita pisteiden tiheyden perusteella.

Tiheyteen perustuvissa klusterointimenetelmissä perusideana on, että klusterin kokoa voidaan kasvattaa vain jos sen naapurustossa on tarpeeksi paljon pisteitä.

Tiheyteen perustuvia klusterointimenetelmiä voidaan käyttää suodattamaan kohina ja poikkeavuudet pois aineistosta. [19]

Esimerkkinä tiheyteen perustuvasta menetelmästä toimii DBSCAN. Menetelmä voidaan kuvata iteratiivisena algoritmina. Ennen algoritmin kuvaamista on kuitenkin määritettävä seuraavat käsiteet:

*Eps*-naapurusto on joukko pisteitä, jotka ovat pisteestä $p$ enintään Eps-mitan päässä.

$$N_{Eps}(p) = q \in D | d(p, e) \leq Eps, \tag{3.32}$$

jossa D on pisteiden perusjuokko.

Algoritmille toinen määriteltävä vakio on $MinPts$. Luku on vähimmäinen pisteiden lukumäärä, jonka perusteella joukko voidaan määrittää klusteriksi.

Määritellään tiheyssaavutettavuus pisteelle $p$ pisteestä $q$: Jos 1) $p \in N_{Eps}(q)$ 2) $|N_{Eps}(q)| \geq MinPts$, tällöin piste $p$ on tiheyssaavuttava pisteen $q$.

Määritetään tiheys-yhdistyvyys pisteelle $p$ ja $q$: Jos on olemassa piste o, jolle pisteet $p$ ja $q$ ovat tiheyssaavutettavia ovat pisteet $p$ ja $q$ tiheysyhdistettyjä.

Klusteri määritellään D:n osajoukolle siten, että olkoon joukko C joukon D:n osajoukko. Oletetaan että, piste $p$ kuuluu klusteriin C ja $q$ on tiheyssaavuttava $p$:n. Tällöin piste $q$ kuuluu klusteriin C. Toinen ehto klusterille on, että jos $p$ ja $q$ kuuluvat klusteriin C on $p$:n oltava tiheys-yhdistetty pisteesen $q$.

Kohinaksi määritellään kaikki aineiston D pisteet, jotka ei kuulu mihinkään klusteriin.

Piste on sisäpiste, jos se on klusterin sisällä ja reunapiste, jos se on klusterin reunalla.

DBSCAN-algormitmi voidaan määritellä seuraavasti:

Aloitetaan mielivaltaisesta pisteestä $p$.

Jos $p$ on sisäpiste, voidaan muodostaa klusteri. Jos kyseessä on reunapiste, siirrytään seuraavaan aineiston pisteeseen, kunnes saadaan sisäpiste.

Algoritmi voi yhdistää kaksi klusteria jos ne täyttävät yllämääritetyt klusterin pisteiden ehdot. [28]

Tämän jälkeen kaksi osajoukkoa, joiden tiheys on vähintään harvimman klusterin verran, voidaan erottaa toisistaan, jos näiden joukkojen etäisyys on vähintään *Eps*.

### 3.5.5 Ristikkoperustaiset menetelmät

Ristikkoperustaiset menetelmät saavat nimensä tavastaan kuvata aineistoa. Aineisto kuvataan ruudukkoon, minkä jälkeen klusterointimenetelmä sovelletaan tähän uuteen aineistoon. Tämän tarkoituksena on tehdä isoista aineistoista laskennallisesti saavutettavia. Ruudukon avulla aineiston resoluutio laskee, mutta myös laskenta aika vähenee merkittävästi. Tämä menetelmä soveltuukin hyvin, kun tarkastellaan isoja tietovirtoja [19]. Ristikkoperustaiset menetelmät voidaan jakaa ristikointivaiheeseen ja tämän jälkeiseen klusterointiin [29].

Yleistä on, että ristikkoperustaiset menetelmissä aineistoon sovelletaan tasaisen resoluution ruudukkoa, jolloin ongelmaksi muodostuu resoluution lasku. Tällöin epäsäännöllisen aineiston erityispiirteet voivat jäädä huomioimatta. [29]

Ristikkoperustaisissa menetelmissä ruudukoinnin jälkeen aineistoon voidaan suorittaa mielivaltainen klusterointialgoritmi. Tämän algoritmin valinta on tehtävä tapauksesta riippuen. Monesti suosittu valinta onkin K-means-menetelmät, sillä ruudukkomenetelmiä yleisesti suositaan, kun tarvitaan nopeaa suoritusta ja K-means sopii tämänkaltaiseen tehtävään. [29]

Matalan resoluution ongelmaa varten on ehdotettu menetelmä Adaptive Mesh Refinement (AMR). AMR sisältää ruudukoinnin lisäksi täyden implementaation klusterointiin. AMR koostaa aineistosta useita eritiheyksisiä ruudukkoja, joiden tiheys perustuu paikalliseen tiheyteen. Nämä ruudukot tuottavat hierarkkisen puun ruudukkoja, jossa ruudukkojen resoluutio kasvaa edeten hierarkkiassa alaspäin. Ruudukkohierarkkian muodostamisen jälkeen algorit-

mi muodostaa jokaisen puun lehden klusterin keskipisteeksi ja lisää rekursiivisesti lehden isäntäsolmuja jäseniksi klusteriin, kunnes saavutetaan puun juuri. Eteneminen puussa menee pienimmän etäisyyden perusteella. [29]

Muita ristikkoperustaisia menetelmiä ovat STING ja CLIQUE. [19], [30]

### 3.5.6 Tilastolliseen malliin perustuvat menetelmät

Tilastolliseen malliin perustuvat klusterointimenetelmät esittävät tilastollisen mallin klustereille ja sovittavat aineiston vastaamaan mallia mahdollisimman hyvin[19]. Malli yleisesti ottaen maksimoi odotusarvoa tai samankaltaisuutta. Menetelmälle jää mallin parametrien estimoiminen ja sovittaminen siten, että tämä valittu arvo saavuttaa maksiminsa.

Tilastolliseen malliin perustuvat klusterointimenetelmät ovat kasvattaneet suosiotaan ja niiden vahvuutena nähdään se, että menetelmä huomioi aineiston muodon ja rakenteen, eikä pelkästään pisteiden välisiä etäisyyksiä. [31]

Tilastoon perustuva klusterointimalli voidaan karakterisoida seuraavasti:

Olkoon $x_i \in X$ datapisteiden joukko ja $z_i$ pisteen kuuluminen klusteriin 1 ... M.

Merkitään klusterien frekvenssiä

$$P(z_i = m) = \pi_m. \tag{3.33}$$

Oletetaan, että saman klusterin sisällä aineisto on generoitu samasta jakaumasta:

$$P(x_k|z_k = m) = f(x_k, \lambda_m), \tag{3.34}$$

jossa $\lambda_m$ on klusterille vapaa parametri.

Näin ollen pisteen kuuluminen klusteriin on todennäköisyys

$$P(x_k) = \Sigma_{m=1}^{M} P(x_k|z_k = m)P(z_k = m) \tag{3.35}$$

koko pistejoukolle, olettaen, että havainnot ovat riippumattomia.

$$P(x_1, x_2, ...X_n) = \Pi_{k=1}^{N} P(x_k) \tag{3.36}$$

Näin ollen mallin parametrit voidaat estimoida maksimoimalla mallin uskottavuutta.

Parametrien estimoinnin jälkeen todennäköisyys pisteen kuulumisesta klusteriin saadaan:

$$P(z|x) = P(x|z)\frac{P(z)}{P(x)}. \tag{3.37}$$

Edellistä varten on ehdotettu odotusarvoa maksimoivaa algoritmia, sillä mallin uskottavuutta on haastava määrittää.

EM-algoritmi haluaa maksimoida integraalin logaritmia

$$\log P(x|\theta) = \log \int p(x, z|\theta)dz. \tag{3.38}$$

Algoritmin jaetaan kahteen askeleeseen: E ja M.

E-askel määritellään seuraavasti:

Approksimoidaan puuttuvan tiedon jakaumaa. Merkitään sitä $u(z)$:lla. Ideaalitilanteessa tämä on $p(z|x, \theta)$.

Olkoon $Q(\theta) = \int logP(x, z|\theta)u(z)dz$. Tämä on uskottavuuden funktio keskiarvoitettuna $u(z)$:lla.

M-askel määritellään seuraavasti: maksimoi $\theta_n = arg\max Q(\theta)$. $\theta = \theta_n$

Toistetaan E ja M-askelia kunnes algoritmi konvergoi.

EM-menetelmä on vain yksi tapa estimoida tilastollisen mallin parametrit. Muita ehdotettuja menetelmiä ovat muun muassa. neuroniverkkoihin perustuvia algoritmeja sekä COWEB-oppimisalgoritmi [19].

### 3.5.7 Menetelmien aikakompleksisuus

Tänä päivänä datamäärät ovat kasvaneet käsittämättömiin lukemiin. Tietokannat käsittävät useissa tapauksissa terabiteittäin tavaraa, eikä ole enää harvinaista, että tiedonlouhinnan kohteena olevat aineistot sisältävät kymmeniä miljoonia datapisteitä. Vaikka laskentateho on kasvanut huomattavasti teknologian kehityksen mukana, on silti tärkeää kuinka nopea tiedonkäsittelyalgoritmi on suhteessa aineiston kokoon. Algormitmin suoritumisnopeudella on tärkeä vaikutus algoritmin valinnassa kun suoriutumisajalla on väliä. Käytännön sovelluksia löytyykin nykyään niin suuri määrä, että algoritmien suoriutumisnopeus nousee yhä tärkeämmäksi tekijäksi menetelmiä valittaessa. [32]

Hierarkkiset klusterointialgoritmit eivät ole kärkipäässä suoriutumisnopeuden näkökulmasta. Naivi yhdistävä hierarkkinen klusterointi algoritmi suoriutuu $O(n^3)$ ajassa. Merkintä $O(n^3)$ tarkoittaa, että algoritmi joutuu keskimäärin suorittamaan lukumäärän operaatioita, joka on verrannollinen datapisteiden lukumäärän kolmanteen potenssiin. Merkintä ei suoraan ota kantaa algoritmin implementaation absoluuttisen suoritusaikaan, vaan kertoo teoreettisen suoritusajan suhteen syötteen pituuteen. [24]

Yhdistävän hierarkkisen klusteroinnin voi optimoida suoriutumaan ajassa $(n^2 logn)$. Tietyt eritystapaukset, kuten minimi- ja maksimietäisyyskriteerifunktiot saavuttavat $O(n^2)$-aikakompleksisuuden jos menetelmässä hyödynnetään ennaltalaskettuja etäisyysmatriiseja. Huomioitavaa on, että mielivaltaiselle kriteerifunktiolla ei voida määrittää aikakompleksisuutta, mutta menetelmän kokoamisosa on itsessään $O(n^2)$ aikavaatimukseltaan, joten tätä paremmaksi ei teoriassa näytetä pääsevän. [24]

Jakavien hierarkkisten klusterointin algoritmien aikakompleksisuus on exhaustive searchilla $O(2^n)$, joka skaalautuu erittäin huonosti aineiston koon kasvaessa. Heuristisia algoritmeja löytyy, mutta jakavat algoritmit suorituvat aikaan nähden huonommin kuin yhdistävät.

Aikakompleksisuus $O(n^2)$ on algoritmisesti hidas. Tämä tarkoittaa sitä, että syötteen pituuden ollessa miljoonia tai kun sovelluskohde vaatii lähes realiaikaista ulosantia on tämän kaltainen algoritmi harvoin käytäntöön soveltuva.

Osittavat menetelmät ovat suosittuja, kun menetelmän aikavaativuus on merkittävä tekijä. Globaali optimaalinen aineiston partitiointi on NP-vaikea ongelma, ja ongelman ratkaisuaika kasvaakin eksponentiaalisesti. Kuitenkin K-means-menetelmille löytyy heuristiikkoja, jonka avulla klusterointi tapahtuu lineaarisessa ajassa $O(n)$. Tästä esimerkinä Lloydin algoritmi. Heuristiset K-means-menetelmät saattaavat helposti konvergoida lokaaliin optimiin globaalin optimin sijaan. Tämä on kuitenkin selkeä vaihtokauppa hyvän aikavaativuuden saavuttamiseen. K-means voidaan esimerkiksi lineaarisen sekalukuoptimoinnin keinoin määritellä siten, että tuloksena on globaali optimi, mutta tämän globaalin varman ratkaisun saavuttaminen on NP-vaikea ongelma. [19]

## 3.6 Klusteroinnin tuottamien tulosten arviointi

Klusterointianalyysi on yleensä kuvaileva, eikä optimaalista lopputulosta välttämättä ole olemassa, tai ainakin se on subjektiivista. On kuitenkin olemassa menetelmiä, joilla voidaan analysoida lopputuloksena saavutettujen klustereiden laatua.

Laadun tarkkailumenetelmät jakautuvat klusterien sisäisiin ja klusterien välisiin mittoihin.

Klusterien sisäiset mitat mittaavat kuinka hyvin klusteri kuvaa sen osajoukon sisältävää aineistoa. Mittaustapana voi olla aineiston homogeenisuuden tai hajonnan mittaaminen. Ominaista menetelmälle kuitenkin on, että sisäistä virhemittaa käytettäessä ei käytetä ulkoista aineistoa.

Klustereita voidaan mallintaa niiden sisältämien pisteiden keskiarvolla. Tällöin eräs virhemitta on keskiarvosta poikkeavuuden neliöiden summa, jota merkitään $SSE$:llä.

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_i - \mu_k|^2, \tag{3.39}$$

jossa $\mu_k$ kuvaa klusterin k vektorikeskipistettä. Tämän menetelmän selkeä haittapuoli on, että virhemitta pienenee klusterien määrän kasvaessa. Tilanteessa, jossa jokainen piste on oma klusterinsa, virhe on nolla. Tähän lähestymistapaan voidaan lisätä sakkoparametri $\lambda K'$ klusterien lukumäärälle, jolloin klusterien määrän kasvattaminen ei välttämättä paranna kohdefunktion arvoa. Tällöin tosin tullaan tilanteeseen jossa klusteroinnin hyvyyden arviointiin tarvitaan ulkopuolinen tekijä $\lambda$.

Esitetyn menetelmän voi yleistää vaihtamalla samanlaisuusmetriikan kuvaamaan klusteroinnin tavoitetta. Yleisessä muodossa virhetermi voidaan esittää:

$$SSE = \sum_{k=1}^{K} N_k S_k, \tag{3.40}$$

jossa $S_k$ on määritetty virhetermi. Tarkemmin:

$$S_k = \min s(x_i, x_j), \quad x_i, x_j \in C_k. \tag{3.41}$$

Klusteroinnille voidaan esittää todennäköisyys, kuinka merkittäviä klusterit ovat. Toisin sanoen määritetään todennäköisyys sille, että klusteria ei oikeasti ole, mutta on saatu kyseiset havainnot. Tämä on klassinen tilastollinen testaus estimoidulle parametrille. Tällä menetelmällä monesti huomataankin, että valitulla metriikalla on huomattavasti väliä, sillä monesti eri metriikat tuottavat eri tason tilastollisen merkitsevyyden klustereille. [33]

Eräs tapa mitata klusteroinnin stabiiliutta on lisätä aineistoon kohinaa ja suorittaa klusterointi uudestaan. Jos klusteroinnin tulos pysyy ennallaan, on syytä olettaa, että menetelmä on tehokas.

Klusteroinnille on vakiintunut eräitä mittoja mittaamaan klusterointialgoritmin suoriutumista, kuten Dunnin etäisyys. [34]

### 3.6.1 Dunnin etäisyys

Dunnin etäisyys on määritelty kuvaamaan klusterien sisäisen samanlaisuuden suhdetta ulkoiseen erilaisuuteen ja toimii näin klusteroinnin hyvyysmittarina.

Dunnin etäisyys määritellään seuraavasti

$$DI(c) = \min_{i \in c}\Big(\min_{j \in c, j i}\Big(\frac{\delta(A_i, A_j)}{\max_{k \in c}\{\Delta(A_k)\}}\Big)\Big), \tag{3.42}$$

jossa $\delta$-funktio kuvastaa klusterien i ja j pienintä etäisyyttä ja $\Delta$ klusterin sisäistä suurinta etäisyyttä. [34]

$$\delta(A_i, A_j) = \min\{d(x_i, x_j) | x_i \in A_i, x_j i \in A_j\} \tag{3.43}$$

$$\Delta(A_k) = \max\{d(x_i, x_j) | x_i, x_j \in A_k\}. \tag{3.44}$$

[34]

## 3.7 Sovellettavan menetelmän valinta

Tässä työssä klusterointia sovelletaan aineistoon, joka on suhteellisen pieni, jolloin menetelmän suoriutumisnopeudella ei ole väliä. Työssä korostuu klusteroinnin tulosten tulkitseminen sekä mahdollisimman pieni ulkoinen syöte algoritmille, jolloin algoritmin annetaan tehdä päätöksensä klustereista täysin itsenäisesti.

Näin ollen tässä työssä sovelletaan hierarkkisia klusterointimenetelmiä ja tarkennettuna yhdistäviä hierarkkisia menetelmiä. Yhdistävistä menetelmistä

käytetään kolmea eri klusterien kokoamistapaa ja metriikkana toimii Euklidinen etäisyys. Myös algoritmin selkeys ja eri vaiheiden tulkinnat ovat hierarkkisten menetelmien vahvuus kyseiseen tehtävään. Tämän työn kannalta menetelmän valintaa painottaa myös se, että klusterien lukumäärää ei määritetä ennalta, vaan klusterit ilmaantuvat semmoisenaan, jos ovat ilmaantuakseen.

# Luku 4

# Työttömyysaineisto

Tämän työn aineisto koostuu kahdesta erillisestä aineistosta. Ensimmäinen aineisto käsittelee yleisiä työttömyyttä kuvaavia mittareita. Aineisto sisältää maiden työttömyysasteen, harmonisoidun työttömyysasteen, työvoimaan osallistuvien asteen, osa-aikatyöllisten osuuden, pitkäaikaistyöttömien asteen sekä nuorten työttömien osuuden. Näiden muuttujien on tarkoitus kuvastaa valtion työvoiman hyödyntämisen tilannetta. Toinen aineisto kuvaa nuorten syrjäytymistä työelämästä. Aineisto sisältää NEET-asteen eri ikäryhmistä. Aineisto on jaettu neljään ryhmään: 15-19-vuotiaat miehet, 20-24-vuotiaat miehet, 15-19-vuotiaat miehet sekä 20-24-vuotiaat miehet.

Kummatkin aineistot koostuvat OECD-maiden työttömyyden tunnusluvuista, ja ovat kerätty vuosilta 2010 ja 2014. Vuosiluvut ovat valittu siten, että maksimoidaan maiden lukumäärä, joilta aineistoa löytyy valituista muuttujista.

Yksi OECD-järjestön tavotteista on tukea taloudellista kehitystä ja tutkimusta. Näin ollen järjestö tarjoaa kaikille saatavilla olevaa tilastollista aineistoa järjestömaiden taloudellisista luvuista. Tämän työn aineisto on peräisin OECD-järjestön aineistosta [35], [36], [37], [38], [39], [40], [41].

Aineistoon kuuluvat maat ovat esitetty taulukossa 4.1. Taulukossa on esitetty tässä työssä maista käytetyt lyhenteet.

| Lyhenne | Maan nimi |
|---|---|
| AUS | Australia |
| AUT | Itävalta |
| BEL | Belgia |
| CAN | Kanada |
| CHE | Sveitsi |
| CZE | Tsekki |
| DEU | Saksa |
| DNK | Tanska |
| ESP | Espanja |
| EST | Viro |
| FIN | Suomi |
| FRA | Ranska |
| GBR | Iso-Britannia |
| GRC | Kreikka |
| HUN | Unkari |
| IRL | Irlanti |
| ISR | Israel |
| ITA | Italia |
| MEX | Meksiko |
| NLD | Alankomaat |
| NZL | Uusi-Seelanti |
| POL | Puola |
| PRT | Portugali |
| SVK | Slovakia |
| SVN | Slovenia |
| SWE | Ruotsi |
| TUR | Turkki |

Taulukko 4.1: Tutkimuksen otosmaat ja niiden lyhenteet

## 4.1 Työttömyyttä kuvaava aineisto

### 4.1.1 Työttömyysaste

Tässä työssä työttömyysastetta mitataan erään viikon työttömien lukumäärän perusteella. Mittausaika on referenssiviikko, jolloin mittaukset ovat suoritettu. EU-maissa tätä mittausaikaa jatketaan neljä viikkoa muiden maiden jälkeen. Tässä työssä työttömyyden kriteereihin kuuluu, että työtön on etsinyt töitä viimeisen neljän viikon aikana.

Kuvassa 4.1 on esitetty maiden työttömyysasteet kuvaajana vuosina 2010 ja 2014, ja taulukossa 4.2 esitetään kyseisen aineiston tilastollisen analyysin tunnusluvut.

Vuosina 2010 ja 2014 otosmaiden työttömyysasteen vaihteluväli on suuri. Vuonna 2010 väli vaihtelee vajaan viiden prosentin ja vajaan 20 prosentin välillä. Vuoteen 2014 mennessä alaraja on pysynyt samana, mutta suurin arvo on kasvanut yli 25:een prosenttiin.

Työttömyysasteen keskiarvo on pysynyt lähes samana vuosien 2010 ja 2014 välillä, mutta hajonta on kasvanut merkittävästi.

Vuoden 2010 matalan työttömyyden maat ovat pääsääntöisesti säilyttäneet hyvän työttömyysasteen. Vuoden 2010 matalimmat työttömyysasteet ovat Alankomaissa, Sveitsissä, Itävallassa ja Meksikossa. Vastaavasti vuonna 2014 matalimmat luvut löytyvät Sveitsistä, Meksikosta, Saksasta ja Itävallasta.

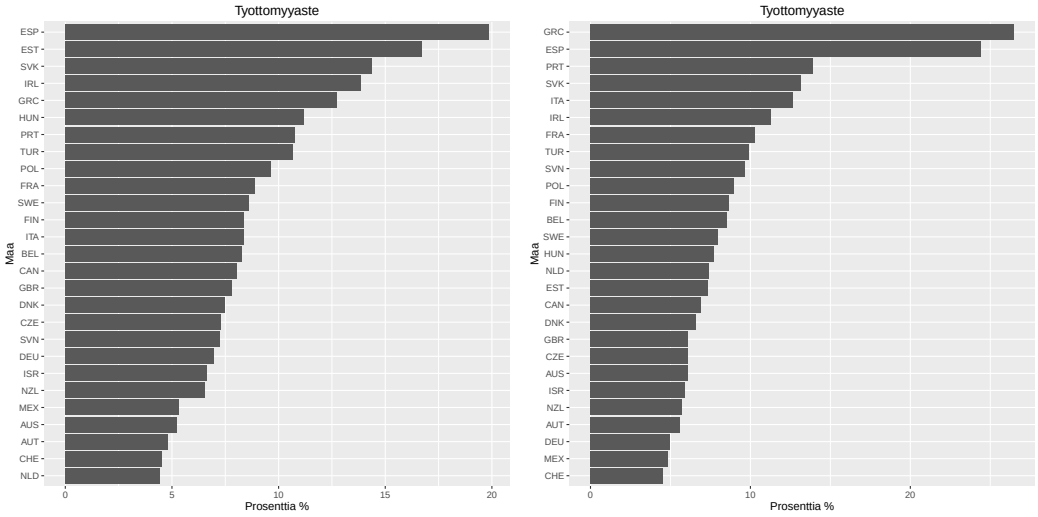
Korkeimmat työttömyysasteet vuonna 2010 ovat Espanjassa, Virossa, Slovakiassa sekä Irlannissa. Vastaavat korkeimmat luvut vuonna 2014 löytyvät Kreikasta, Espanjasta, Portugalista ja Slovakiasta.

Vuoden 2010 tilannetta selkeästi eniten parantanut maa on Viro, joka on puolittanut työttömyysasteensa vuosien 2010 ja 2014 välillä.

Eniten tilanne on huonontunut Kreikalla, jonka työttömyysaste on kasvanut jopa 10 prosenttiyksikköä tämän neljän vuoden aikana. Muita työttömyysasteen selkeitä kasvattajia ovat olleet Portugali ja Italia. Maat, joiden työttömyyslukemat ovat olleet korkealla 2010, ovat Viroa lukuun ottomatta säilyttäneet korkeat työttömyyslukemat.

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 19,86 | 26,49 |
| Minimi | 4,45 | 4.54 |
| Keskiarvo | 9,06 | 9,33 |
| Mediaani | 8,29 | 7,73 |
| Keskihajonta | 3,76 | 5,31 |

Taulukko 4.2: Työttömyysasteaineiston tilastolliset tunnusluvut




(a) Työttömyysaste vuonna 2010 (b) Työttömyysaste vuonna 2014

Kuva 4.1: Työttömyysaste otosmaissa vuosina 2010 ja 2014 [35]

### 4.1.2 Harmonisoitu työttömyysaste

Harmonisoitu työttömyysaste mittaa työttömien osuutta työvoimasta. Harmonisoitu työttömyysaste määritellään työttömien prosentuaaliseksi osuudeksi työvoimasta, josta on poistettu kausittaisvaihtelut. Kausittaisen vaihtelun poistaminen tarkoittaa kausittaisten jaksojen "piikkivaikutusten"poistamisen aikasarjasta, jotta pidemmän aikavälin trendi näkyisi paremmin.

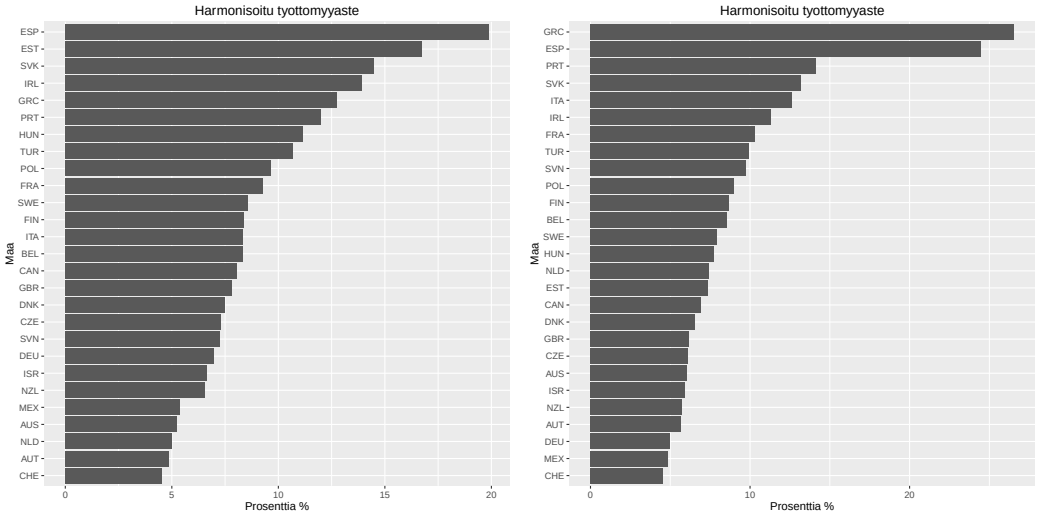
Harmonisoidun työttömyysasteen aineisto on esitetty kuvaajina kuvassa 4.2 ja aineiston tilastolliset tunnusluvut taulukoissa taulukossa 4.3.

Harmonisoidusta työttömyysasteesta nähdään, että kausittaisvaihtelun poistamisella on merkitystä vain Ranskassa, Portugalissa sekä Alankomaissa. Tällöinkin kyse on vain noin prosenttiyksikön suuruisesta erosta. Muiden maiden

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 19,88 | 26.55 |
| Minimi | 4,54 | 4.54 |
| Keskiarvo | 9,15 | 9,35 |
| Mediaani | 8,31 | 7,74 |
| Keskihajonta | 3,78 | 5,33 |

Taulukko 4.3: Harmonisoitutyöttömyysasteaineiston tilastolliset tunnusluvut

tapauksessa ero on alle 0,1 prosenttiyksikköä ja harmonisoitu työttömyysaste vastaa työttömyysastetta.




(a) Harmonisoitu työttömyysaste vuonna 2010

(b) Harmonisoitu työttömyysaste vuonna 2014

Kuva 4.2: Harmonisoitu työttömyysaste otosmaissa vuosina 2010 ja 2014 [36]

### 4.1.3 Osa-aikatyöllisten osuus

Osa-aikatyölliseksi lasketaan ne työssä kävijät, jotka tekevät pääsääntöisesti enintään 30 tuntia töitä viikossa. Tämä esitetään prosenttiosuutena työssäkäyvien joukosta. Osa-aikatyöllisyyttä on vapaaehtoista sekä olosuhteiden pakosta johtuvaa. Tämä tarkoittaa sitä, että on olemassa joukko osa-aikatyöllisiä, jotka ovat omasta tahdostaan vain osan aikaa viikosta töissä. Tästä esimerkkinä monet opiskelijat tai urheilijat. Toisaalta osa-aikatyöllisyyttä voidaan mitata myös työttömyytenä, jos töissäkäyvä haluaisi olla täyden työviikon

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 37,15 | 38,45 |
| Minimi | 3,66 | 4,50 |
| Keskiarvo | 15,91 | 16,33 |
| Mediaani | 15,24 | 15,99 |
| Keskihajonta | 7,9 | 7,92 |

Taulukko 4.4: Osa-aikatyöllisyysaineiston tilastolliset tunnusluvut

töissä, mutta tätä mahdollisuutta ei tarjota. Tämä aineisto ei erottele perusteita, joiden takia ihmiset ovat vain osan aikaa viikosta töissä.

Osa-aikatyöllisten osuuden aineiston on esitetty kuvaajassa 4.3 ja aineiston tilastolliset tunnusluvut taulukossa 4.4.

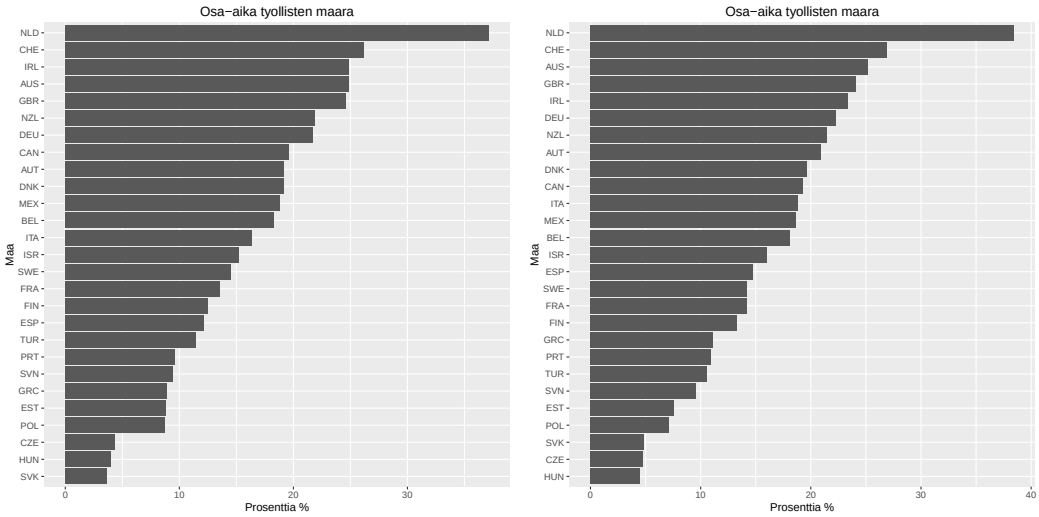
Osa-aika työllisten osuus vaihtelee maittain suuresti. Vuonna 2010 vaihteluväli on 3,65-37,1 prosenttiyksikköä ja vuonna 2014 tämä väli on 4,5-38,45 prosenttiyksikköiä.

Keskihajonta on pysynyt lähes samana (7,9) vuosina 2010 ja 2014.

Keskiarvo ja mediaani ovat nousseet muutamalla prosentilla vuoteen 2014 mennessä.

Alankomaat poikkeavat suuresti muista maista suurimmalla osa-aikatyöllisten osuudellaan. Ero seuraavaan on kumpanakin vuotena kymmenen prosenttiyksikön luokkaa.

Suuria nousijoita tai laskijoita ei ole vuosien 2010 ja 2014 välillä. Tsekki, Unkari ja Slovakia pitävät häntäpäätä osa-aikatyöllisten osuudessa kumpanakin tarkasteluvuotena, ja Alankomaita seuraa Sveitsi, Irlanti, Australia sekä Iso-Britannia kärkipäässä.




(a) Osa-aikatyöllisten osuus työvoimasta vuonna 2010

(b) Osa-aikatyöllisten osuus työvoimasta vuonna 2014

Kuva 4.3: Osa-aikatyöllisten osuus työvoimasta otosmaissa vuosina 2010 ja 2014 [37]

### 4.1.4 Työvoimaan osallistuvien aste

Työttömyyden indikaattoreihin kuuluu työvoimaan osallistuvien aste. Työvoimaan osallistuvien aste on se prosenttiosuus työikäisistä, jotka osallistuvat palkkatyöhön tai ovat itsensä työllistäviä mittausajanjaksossa tai ovat tilapäisesti poissa töistä mittausajanjakson aikana. Työikäisiksi määritellään kaikki 15-64 vuotiaat.

Työvoimaan osallistumisasteen aineisto on esitetty kuvaajassa 4.4 ja aineiston tilastolliset tunnusluvut taulukossa 4.5.

Työvoimaan osallistuvien aste sisältää vähemmän hajontaa maiden välillä kuin muut muuttujat. Suurin työvoimaan osallistuvien aste löytyy Ruotsista, mutta vain pienellä erolla seuraavana tuleviin Uuteen-Seelantiin, Sveitsiin, Kanadaan, Suomeen sekä Viroon. Samat maat ovat kärjessä vuosina 2010 ja 2014, vain Suomen, Viron ja Kanadan järjestys vaihtelee.

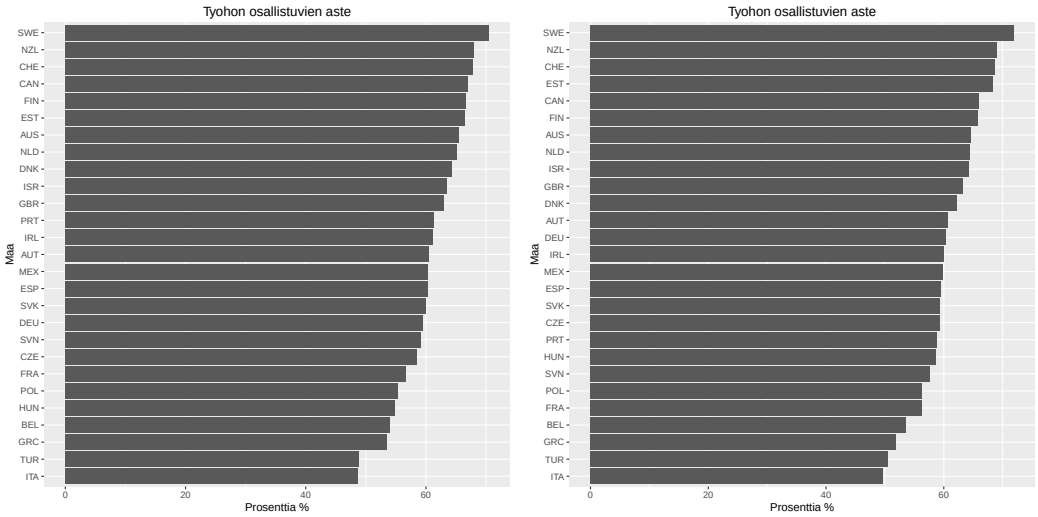
Pienimmät työvoimaan osallistuvien asteet löytyvät Italiasta, Turkista, Kreikasta ja Belgiasta. Nämä neljä maata ovat viimeisenä kumpanakin tarkasteluvuotena ja samassa järjestyksessä.

Suurta vaihtelua ei missään tarkastelumaassa tapahdu vuosien 2010 ja 2014 välillä. Myös tunnusluvut (keskihajonta, mediaani, keskiarvo, maksimi ja mi-

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 70,48 | 71,94 |
| Minimi | 48,74 | 49,60 |
| Keskiarvo | 60,73 | 60,77 |
| Mediaani | 60,40 | 59,99 |
| Keskihajonta | 5,69 | 5,69 |

Taulukko 4.5: Työvoimaan osallistumis aste-aineiston tilastolliset tunnusluvut

nimi) pysyvät paikoillaan vuosien välillä.




(a) Työvoimaan kuuluvien osuus työikäisistä vuonna 2010

(b) Työvoimaan kuuluvien osuus työeikäisistä vuonna 2014

Kuva 4.4: Työvoimaan kuuluvien osuus työikäisistä otosmaissa vuosina 2010 ja 2014 [38]

### 4.1.5 Pitkäaikaistyöttömien osuus työttömistä

Pitkäaikaistyöttömiksi lasketaan ne työttömät, jotka ovat olleet mittausviikolla vähintään 12 kuukautta tai pidempään työttömiä. Tämä muuttuja kuvaa sitä, kuinka suuri osa työttömistä on pitkäaikaistyöttömiä.

Pitkäaikaistyöttömyien osuus työttömistä aineisto on esitetty kuvaajassa 4.5 ja aineiston tilastolliset tunnusluvut taulukossa 4.6.

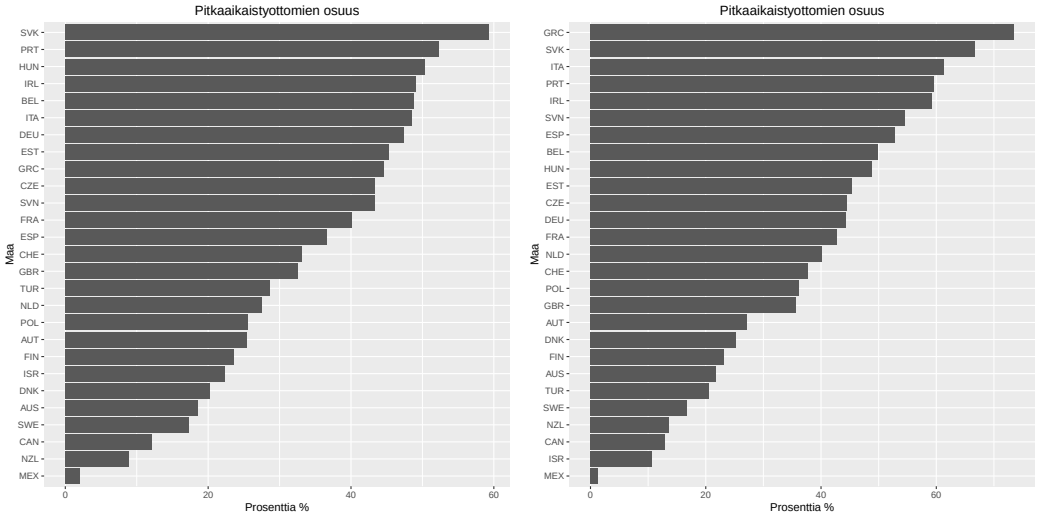
Pitkäaikaistyöttömien osuus on keskimäärin kasvussa vuosien 2010 ja 2014

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 59,31 | 73,50 |
| Minimi | 2,03 | 1,23 |
| Keskiarvo | 33,58 | 38,02 |
| Mediaani | 33,14 | 40,15 |
| Keskihajonta | 14,97 | 18,93 |

Taulukko 4.6: Pitkäaikaistyöttömien osuus työttömistä-aineiston tilastolliset tunnusluvut

välillä. Hajonta on kumpanakin vuonna todella suurta.

Vuoden 2010 minimi on vain 2,0 prosenttia, joka saavutetaan Meksikossa, kun taas vastaava suurin arvo Slovakiassa on 59,3 prosenttia. Vastaavat maat ja luvut vuonna 2014 ovat: minimi 1,2 prosenttia Meksikossa ja maksimi Kreikassa 73,5 prosenttia. Myös keskiarvo on noussut 33,6 prosentista 38,0 prosenttiin ja hajonta on noussut selkeästi vuosien välillä.




(a) Pitkäaikaistyöttömien osuus työttömistä vuonna 2010

(b) Pitkäaikaistyöttömien osuus työttömistä vuonna 2014

Kuva 4.5: Pitkäaikaistyöttömien osuus työttömistä otosmaissa vuosina 2010 ja 2014 [39]

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 41,48 | 53,20 |
| Minimi | 7,82 | 7,76 |
| Keskiarvo | 20,39 | 21,16 |
| Mediaani | 20,30 | 17,92 |
| Keskihajonta | 8,77 | 12,24 |

Taulukko 4.7: Nuorten työttömien osuus vastaavasta ikäryhmästä -aineiston tilastolliset tunnusluvut

### 4.1.6 Nuorten työttömien osuus

Nuorten työttömien osuus mittaa nuorten eli 15-24-vuotiaiden työttömien osuutta vastaavan ikäluokan työvoimasta.

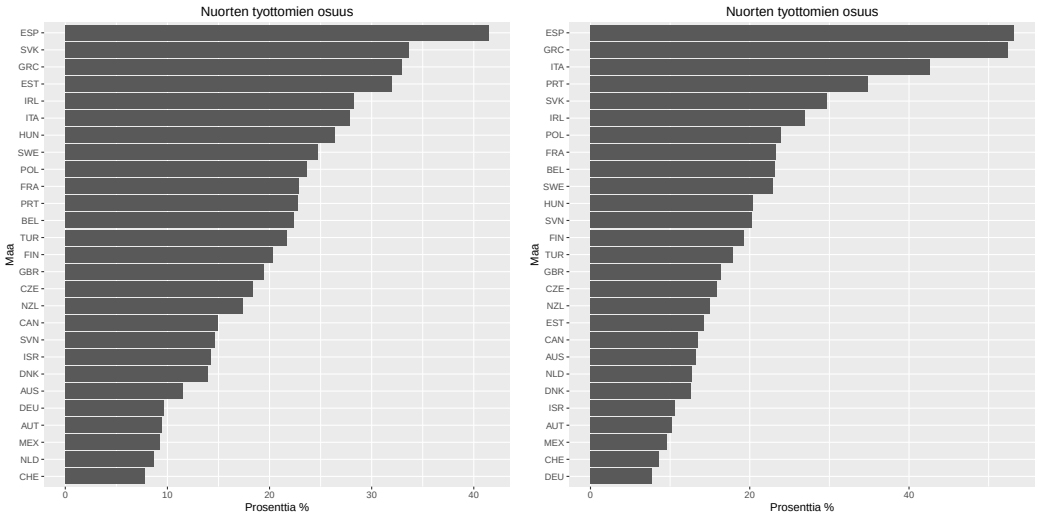
Nuorten työttömien osuuden aineisto on esitetty kuvaajassa 4.6 ja aineiston tilastolliset tunnusluvut taulukossa 4.7.

Myös nuorten työttömyys vaihtelee suuresti maiden välillä. Vuoden 2010 suurin nuorten työttömyys menee Espanjalle, jossa 41,5 prosenttia nuorista on työttömiä. Vuoteen 2014 mennessä Espanja on säilyttänyt paikkansa suurimpana nuorisotyöttömyyden maana ja kohottanut vastaavan luvun 53,2 prosenttiin.

Kreikka on myös tehnyt korkean nousun vuoteen 2014 ja onkin lähes samoissa lukemissa Espanjan kanssa. Vuoden 2010 suurimpiin nuorten työttömyyden maihin kuulunut Slovakia on tehnyt suurimman laskun vuoteen 2014 mennessä.

Pienimpiä nuoriso työttömyyden arvoja löytyy vuonna 2010 Sveitsistä, Alankomaista, Meksikosta, Itävallasta sekä Saksasta. Vastaavasti vuonna 2014 pienimmät nuorisotyöttömyyden luvut löytyvät Saksasta, Sveitsistä, Meksikosta Itävallasta sekä Israelista. Pienin arvo vuonna 2010 on 7,8 prosenttia ja vuonna 2014 7,7 prosenttia.

Vuosien 2010 ja 2014 välillä keskiarvo on kasvanut, mutta mediaani on laskenut. Tästä nähdään, että pylväskuvaajien massa on vetäytynyt selkeästi, mutta kärkimaiden nuorisotyöttömyys on pahentunut selkeästi. Matalan työttömyyden maissa muutosta ei ole suuresti tapahtunut.




(a) Nuorten työttömien osuus työvoimasta vuonna 2010

(b) Nuorten työttömien osuus työvoimasta vuonna 2014

Kuva 4.6: Nuorten työttömien osuus työvoimasta otosmaissa vuosina 2010 ja 2014 [40]

## 4.2 Nuorten syrjäytyminen työelämästä (NEET)

Tässä työssä tarkastellaan työelämästä syrjäytymisessä erikseen ryhmiä 15-19-vuotiaat miehet, 15-19-vuotiaat naiset, 20-24-vuotiaat miehet sekä 20-24-vuotiaat naiset.

Työelämästä syrjäytyminen esitetään prosenttiosuutena ikäryhmästä.

### 4.2.1 NEET-aste 15-19-vuotiaat miehet

15-19-vuotiaiden miesten työelämästä syrjäytymisaste on esitetty kuvaajassa 4.7 ja aineiston tilastolliset tunnusluvut taulukossa 4.8.

Nuorten 15-19-vuotiaiden miesten työelämästä syrjäytymisaste on vuonna 2010 enimmillään 24 prosenttia ikäryhmästä ja 2014 10 prosenttia ikäryhmästä. Suurimmat luvut vuonna 2010 löytyvät Israelista, Turkista sekä Espanjasta, joista Israel on selkeässä yli viiden prosenttiyksikön johdossa.

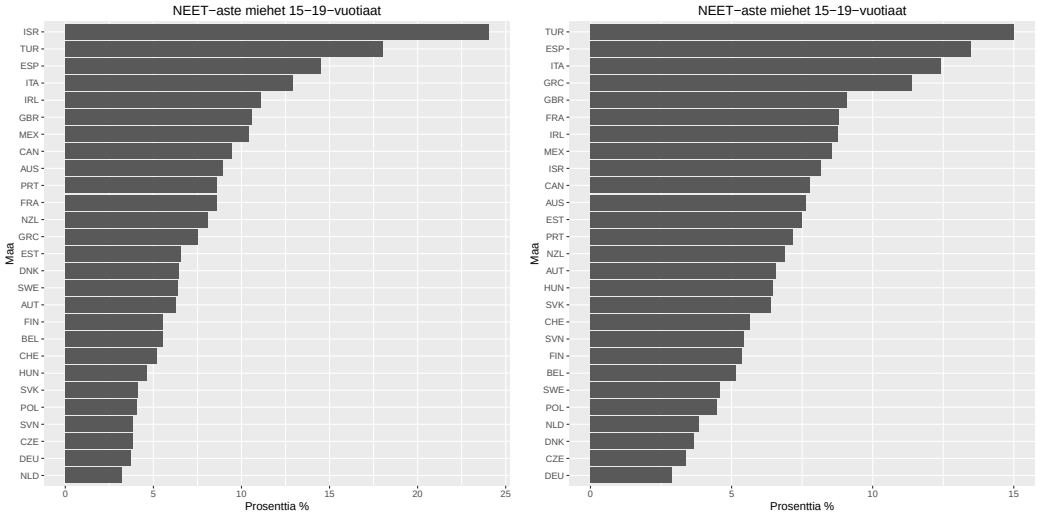
Vuonna 2014 Israelin tilanne on parantunut huomattavasti ja suurimmat luvut löytyvät Turkista, Espanjasta Italiasta sekä Kreikasta. Pienimmät luvut löytyvät vuonna 2010 Alankomaista, Saksasta, Tsekistä sekä Sloveniasta.

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 24,05 | 15,01 |
| Minimi | 3,21 | 2,9 |
| Keskiarvo | 8,22 | 7,28 |
| Mediaani | 6,54 | 6,91 |
| Keskihajonta | 4,79 | 3,05 |

Taulukko 4.8: 15-19 vuotiaiden miesten NEET-aineiston tilastolliset tunnusluvut

Vuonna 2014 vastaavat maat ovat Saksa, Tsekki, Tanska ja Alankomaat.

Vuosien 2010 ja 2014 välillä 15-19-vuotiaiden miesten työelämästä syrjäytyminen on selkeässä laskussa. Keskiarvo on tippunut 8,2:sta 7,3:meen, sekä maksimi arvo 24,0 prosentista 15,0 prosenttiin. Myös hajonta maiden välillä on pienentynyt vuoteen 2014 mennessä.




(a) 15-19-vuotiaiden miesten NEET-aste vuonna 2010

(b) 15-19-vuotiaiden miesten NEET-aste vuonna 2014

Kuva 4.7: 15-19-vuotiaiden miesten NEET-aste otosmaissa vuosina 2010 ja 2014 [41]

### 4.2.2 NEET-aste 20-24-vuotiaat miehet

20-24-vuotiaiden miesten työelämästä syrjäytymisaste on esitetty kuvaajassa 4.8 ja aineiston tilastolliset tunnusluvut taulukossa 4.9.

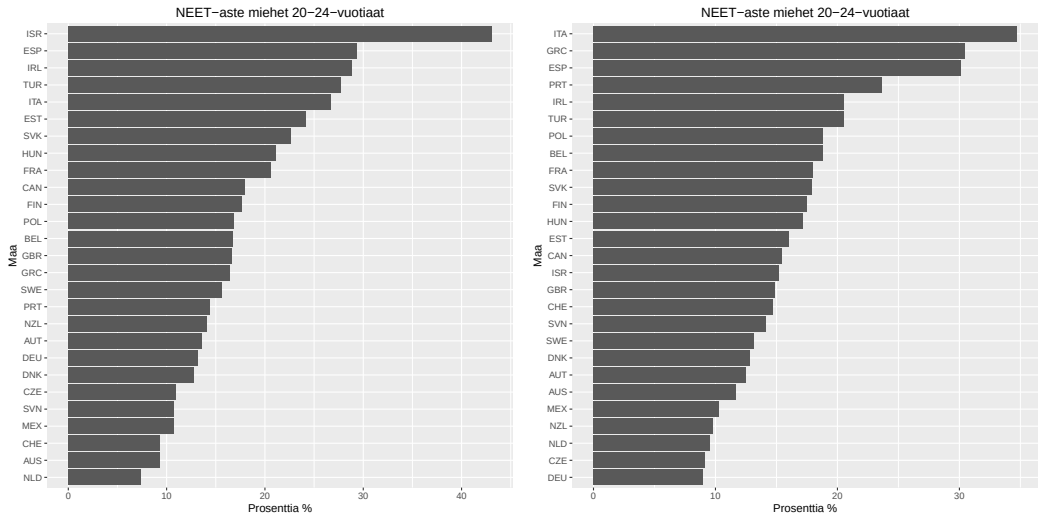
| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 43,09 | 34,73 |
| Minimi | 7,39 | 8,99 |
| Keskiarvo | 18,09 | 16,92 |
| Mediaani | 16,68 | 15,49 |
| Keskihajonta | 7,97 | 6,59 |

Taulukko 4.9: 20-24 vuotiaiden miesten NEET-aineiston tilastolliset tunnusluvut

20-24-vuotiaiden miesten työelämästä syrjäytyminen on suurempaa kuin vastaava luku 15-19-vuotiaiden miesten keskuudessa.

Vuonna 2010 suurin arvo on 43,1 prosenttia, joka löytyy Israelista ja pienin arvo on 7,3, joka löytyy Alankomaista. Vuonna 2014 suurin arvo löytyy Italiasta (34,73), Kreikan ja Espanjan seuraamana. Pienin arvo vuonna 2014 löytyy Saksasta.

Vuoden 2010 ja 2014 välillä myös 20-24-vuotiaiden miesten työelämästä syrjäytyminen on vähentynyt. Keskiarvo on laskenut yli yhden prosenttiyksikön, ja maksimi on laskenut 10 prosenttiyksikköä.




(a) 20-24-vuotiaiden miesten NEET-aste vuonna 2010

(b) 20-24-vuotiaiden miesten NEET-aste vuonna 2014

Kuva 4.8: 20-24-vuotiaiden miesten NEET-aste otosmaissa vuosina 2010 ja 2014 [41]

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 33,99 | 27,33 |
| Minimi | 2,48 | 2,95 |
| Keskiarvo | 8,45 | 7,49 |
| Mediaani | 6,21 | 6,44 |
| Keskihajonta | 7,4 | 5,52 |

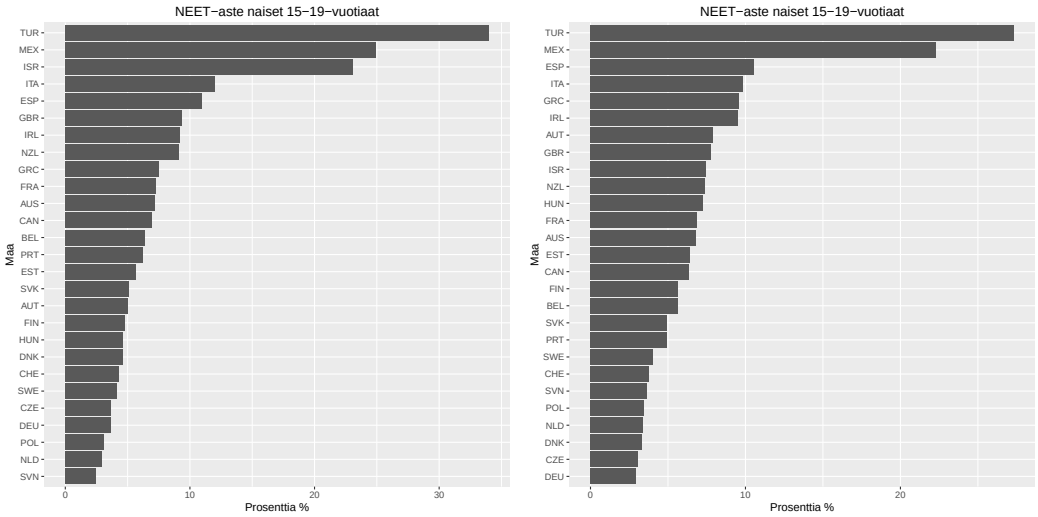
Taulukko 4.10: 15-19 vuotiaiden naisten NEET-aineiston tilastolliset tunnusluvut

### 4.2.3 NEET-aste 15-19-vuotiaat naiset

15-19-vuotiaiden naisten työelämästä syrjäytymisaste on esitetty kuvaajassa 4.9 ja aineiston tilastolliset tunnusluvut taulukossa 4.10.

15-19-vuotiaiden naisten työelämästäsyrjäytymisluvut ovat suurempia kuin vastaavan ikäluokan miesten.

Vuonna 2010 suurin prosenttiosuus löytyy Turkista, jossa 34 prosenttia 15-19-vuotiaista naisista ei kuulu opetuksen, harjoittelun tai työelämän pariin. Kärkikolmikosta löytyy Turkin lisäksi Meksiko ja Israel. Näiden kolmen maiden luvut ovat huomattavasti muita maita korkeammat. Vaikka suurin arvo onkin 34 prosenttia, keskiarvo on vain 8,5 prosenttia ja mediaani 6,2. Vastaavasti vuonna 2014 maksimi arvo on edelleen Turkilla 27 prosenttia, mutta Israelin luku on laskenut huomattavasti. Tuloksena on että vain Turkki ja Meksiko ovat poikkeavan korkeita maita muihin verrattuna. Vuoden 2014 keskiarvo on prosenttiyksikön vähemmän kuin vuonna 2010, mutta mediaani vastaavasti noussut kaksi prosenttiyksikön kymmenystä.




(a) 15-19-vuotiaiden naisten NEET-aste vuonna 2010

(b) 15-19-vuotiaiden naisten NEET-aste vuonna 2014

Kuva 4.9: 15-19-vuotiaiden naisten NEET-aste otosmaissa vuosina 2010 ja 2014 [41]

### 4.2.4 NEET-aste 20-24-vuotiaat naiset

20-24-vuotiaiden naisten työelämästä syrjäytymisaste on esitetty kuvaajassa 4.10 ja aineiston tilastolliset tunnusluvut taulukossa 4.11.

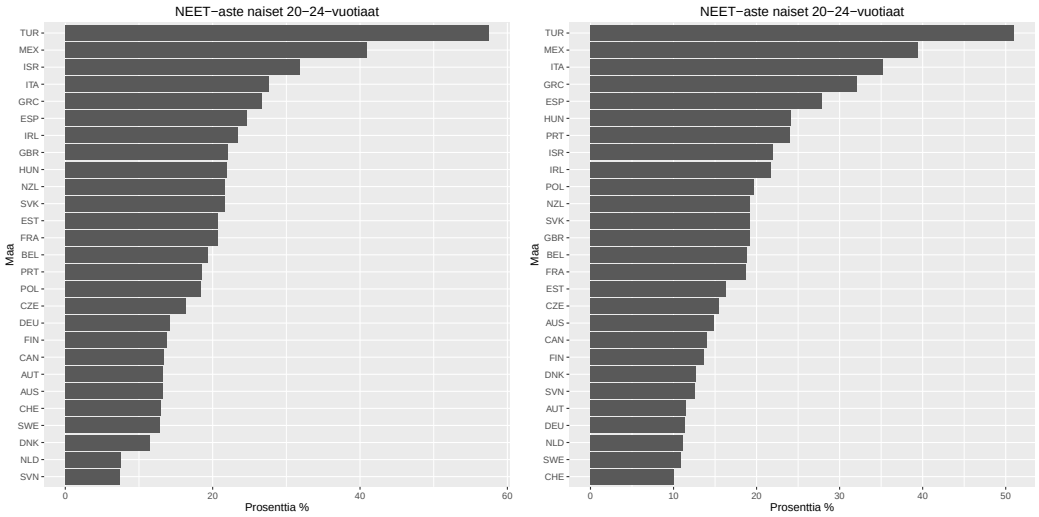
20-24-vuotiaiden naisten työelämästä syrjäytyminen on keskimäärin pahempaa kuin vastaava 15-19-vuotiaiden naisten.

Vuonna 2010 pahiten työelämästä syrjäytyneet naiset löytyvät Turkista, jossa tämä koskee jopa 57,5 prosenttia ikäluokan naisista. Turkin lisäksi pahimmat työelämästä syrjäytymisen luvut löytyvät Meksikosta ja Isralista, joiden ero muihin maihin on huomattava. Vuonna 2014 vain Turkin ja Meksikon vastaavat luvut ovat enää muista suuresti poikkeavia. Suurin arvo on 51 prosenttia ja tämä löytyy Turkista.

Vuoden 2010 ja 2014 välillä ei keskiarvossa suurta muutosta näy. Kumpanakin vuonna se on noin 20 prosenttia. Vuonna 2010 suurin arvo on viisi prosenttiyksikköä suurempi kuin 2014, mutta myös pienin arvo on pienempi kuin vuonna 2014.

| Tunnusluku | vuonna 2010 | vuonna 2014 |
|---|---|---|
| Maksimi | 57,49 | 51,01 |
| Minimi | 7,38 | 10,07 |
| Keskiarvo | 20,49 | 20,26 |
| Mediaani | 19,28 | 18,87 |
| Keskihajonta | 10,46 | 9,75 |

Taulukko 4.11: 20-24 vuotiaiden naisten NEET-aineiston tilastolliset tunnusluvut




(a) 20-24-vuotiaiden naisten NEET-aste vuonna 2010

(b) 20-24-vuotiaiden naisten NEET-aste vuonna 2014

Kuva 4.10: 20-24-vuotiaiden naisten NEET-aste otosmaissa vuosina 2010 ja 2014 [41]

## 4.3 Yhteenveto yksiulotteisista aineistoista

Aineistosta nähdään, että maiden välillä on suuria eroja. Espanja ja Kreikka eroavat muista maista lähes poikkeuksetta. Matalan työttömyyden maat monesti ovatkin myös matalan nuorisosyrjäytymisen maita, mutta toisinpäin tämä ei näytä aina pitävän.

Vuosien 2010 ja 2014 maiden väliset erot näyttävät vain kasvavan. Hajonta on kasvanut lähes kaikissa ulottuvuuksissa, ja "huonoimman"ja "parhaimman"maan välinen ero vain kasvaa. OECD-maiden työllisyystilanne on suu-

rimmassa osissa maita parantunut, mutta kriisin pahimmissa kohteissa tilanne on vain pahentunut. Varsinkin nuorison osalta tilanne on vaikea ja pitkäaikaistyöttömyys on nousussa.

Yksittäiset maat ovat onnistuneet korjaamaan työttömyystilannettaan merkittävästi näiden neljän vuoden aikana. Tästä hyvänä esimerkkinä Viro, joka on nostanut työllisyyttään merkittävästi ja Israel joka on tehnyt merkittävän muutoksen nuorison syrjäytymisen hoitamisessa.

## 4.4 Kaksiulotteinen aineiston analyysi

### 4.4.1 Työttömyyttä kuvaava aineisto

Kuvissa 4.11 ja 4.12 on esitetty työttömyyttä kuvaavat muuttujat pareittain pistekaaviossa. Pistekaavien tarkoituksena on nähdä poikkeavat havainnot, muuttujien väliset riippuvuudet sekä muuttujien klusteroituminen kahdessa ulottuvuudessa.

Pistekaavioista nähdään, että aineistot kerääntyvät huomattavasti pienemmälle alueelle vuonna 2014 kuin 2010. Työttömyysastemuuttujan kanssa muut muuttujat muodostavat vuonna 2014 selkeitä erillisiä alueita. Muutkin muuttujat kerääntyvät alueille tavalla joka ehdottaa, että aineistosta on löydettävissä klustereita.

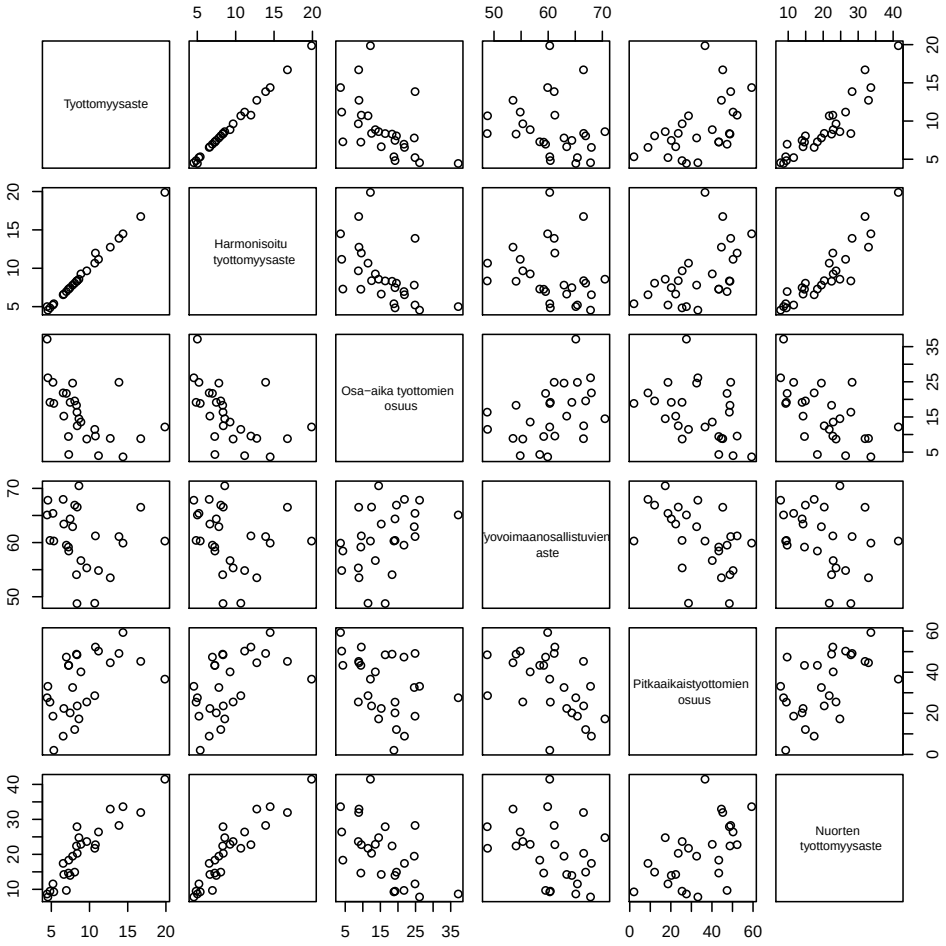
Pistekaavioista nähdään myös, että aineisto sisältää muutamia poikkeavia havaintoja jokaisessa ulottuvuudessa.

Muuttujien työttömyysaste ja harmonisoitu työttömyysaste välinen kuvaaja on muodoltaan suora, jonka kulmakerroin on yksi, mikä tarkoittaa että työttömyyden kausittaisvaihtelut ovat todella pienet. Tämä ilmiö tapahtuu kumpanakin tarkasteluvuotena. Myös nuorten työttömyyden ja työttömyysasteen välinen riippuvuus on lähes lineaarinen. Vuoden 2014 aineistossa riippuvuus ei ole aivan yhtä selvää kuin 2010, mutta lineaarinen riippuvuus on silti havaittavissa.

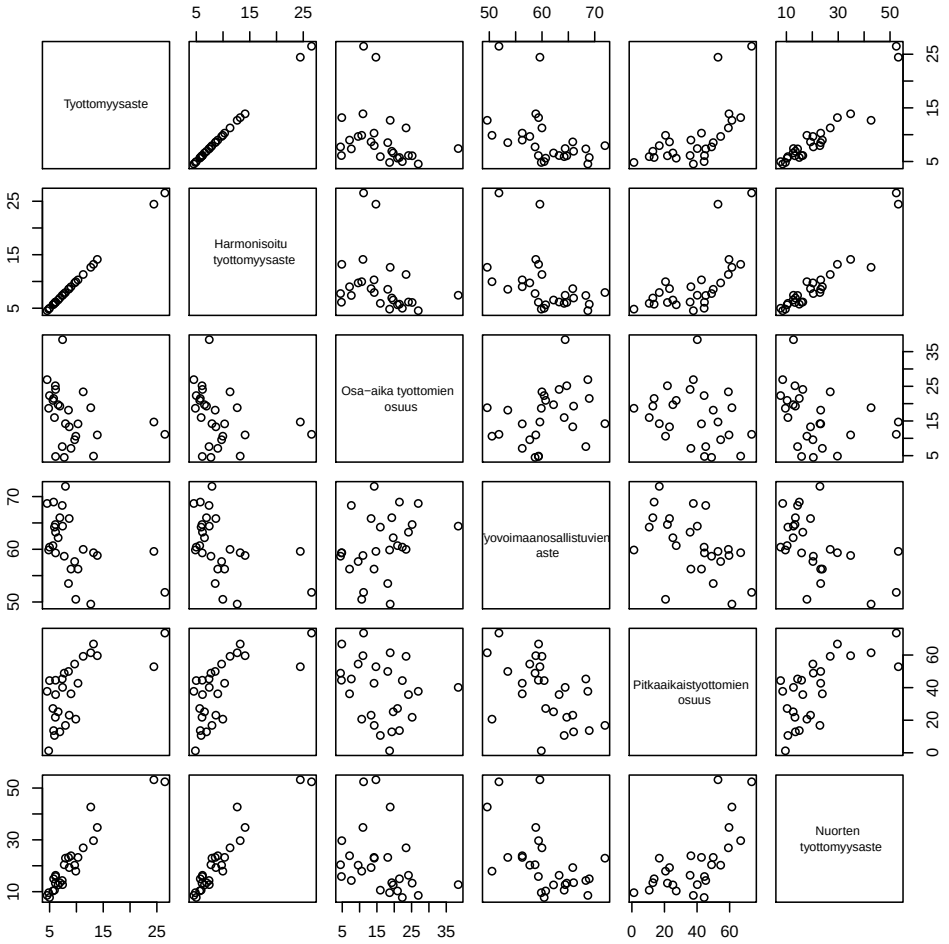
Havaintoa lineaarisesta riippuvuudesta vahvistaa kuvan 4.13 korrelaatiomatriisi, jonka mukaan työttömyysaste ja harmonisoitu työttömyysaste ovat täysin lineaarisesti riippuvia toisistaan, ja nuorten työttömyys on lähes lineaarisesti riippuvainen työttömyysasteesta. Huomionarvoista on se, että eri työttömyysmittareiden välinen korrelaatio ei ole aina positiivinen vaan välillä vahvasti negatiivinen. Nuorten työttömyyden ja osa-aikatyöllisten osuuden välinen korrelaatio on -0.52, joka on vahva negatiivinen korrelaatio. Tämä viit-

taa siihen, että varsinkin nuoret tekevät osa-aikatöitä. Osa-aikatyöllisyyden vaikutus näkyy myös työttömyysasteen kanssa negatiivisena korrelaationa luonnollisesti.

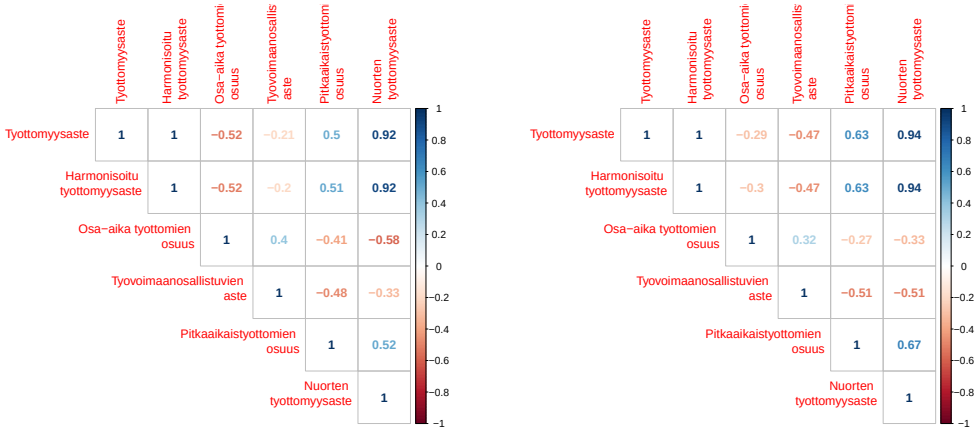
Työvoimaan osallistuvien aste laskee työttömyysastetta huomattavasti vähemmän vuonna 2010 kuin 2014. Tässä nähdään vaikutus sille, kuinka työttömyysaste on riippuvainen siitä, kuinka suuri osa ihmisistä lasketaan työvoimaan. Negatiivinen korrelaatio kuitenkin antaisi vihjettä siitä, että töitä löytyy kun työvoimaa kasvattaa.




Kuva 4.11: Vuoden 2010 työttömyyttä kuvaavat muuttujat pareittain pistekaavioissa




Kuva 4.12: Vuoden 2014 työttömyyttä kuvaavat muuttujat pareittain pistekaavioissa




(a) Korrelaatiot vuonna 2010 (b) Korrelaatiot vuonna 2014

Kuva 4.13: Työttömyyttä kuvaavien muuttujien väliset korralaatiot

### 4.4.2 Nuorten syrjäytymistä työelämästä kuvaava aineisto

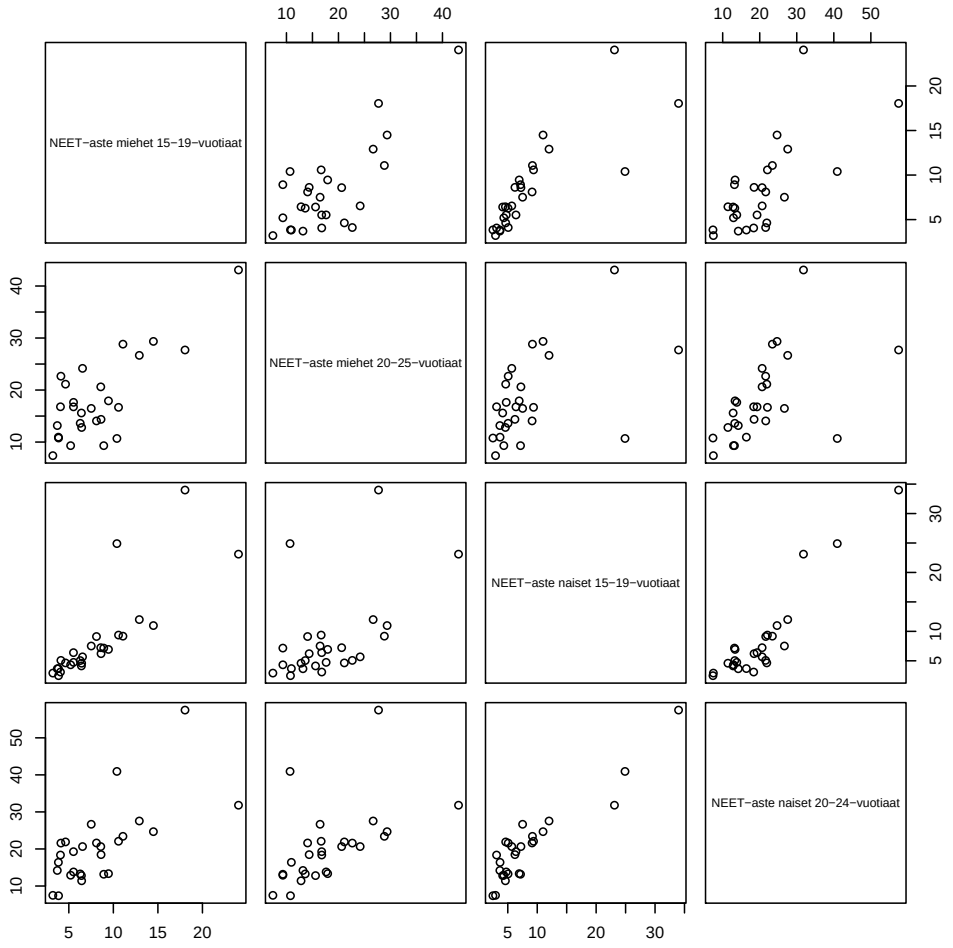
Kuvissa 4.14 ja 4.15 on esitetty nuorten syrjäytymistä kuvaavat muuttujat pareittain pistekaavioissa. Kaavioista nähdään, että pareittain aineisto kasaantuu tiiviisti pienille alueille. Poikkeavia havaintoja nähdään jokaisessa ulottuvuudessa. Vuoden 2010 aineistossa Meksiko, Israel sekä Turkki näkyvät irrallisina pisteinä muusta aineistosta ja vuonna 2014 aineistossa nähdään kaksi erillistä yksittäisten pisteiden muodostamaa pientä joukkoa. Toinen näistä joukoista on Kreikka, Espanjan ja Italia. Toinen joukko on Meksikon ja Turkin muodostama ryhmä.

Lineaarista riippuvuutta näkyy muuttujien 15-19-vuotiaiden miesten NEET-asteen ja 15-19-vuotiaiden naisten välillä. Muiden muuttujien välillä lineaarinen riippuvuus ei ole yhtä selkeää.

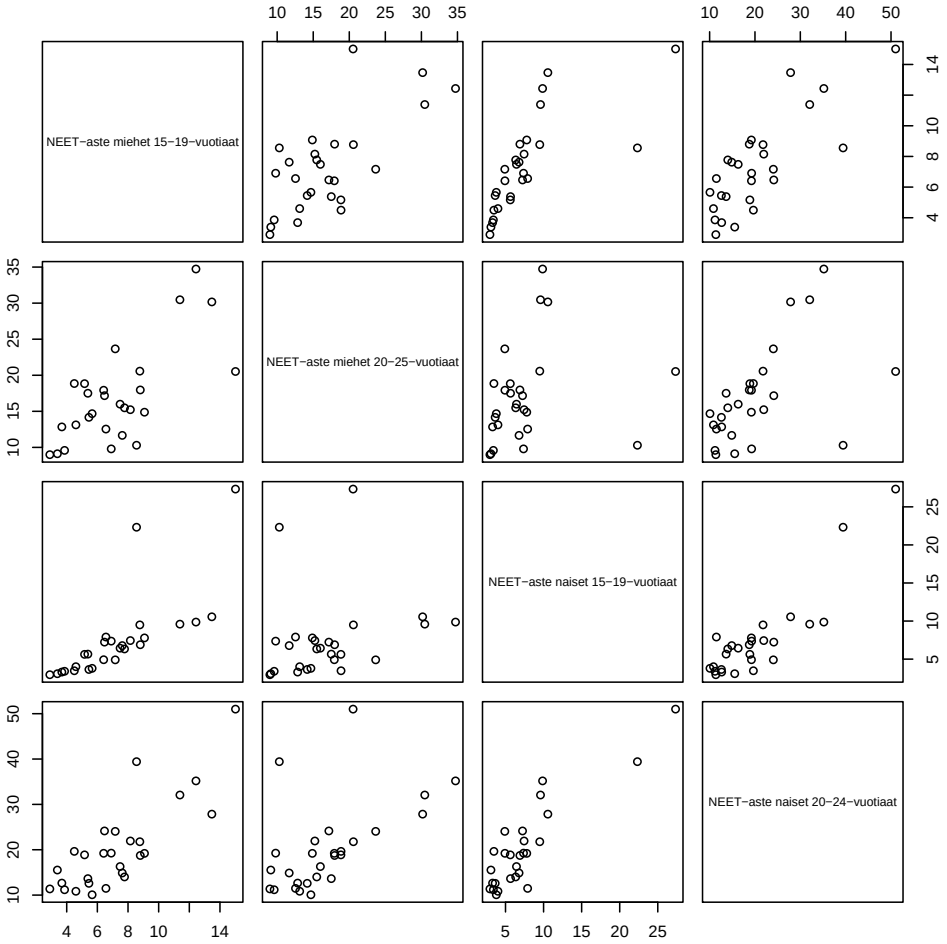
Kuvan 4.16 korrelaatiomatriisi vahvistaa havaintoja lineaarisesta riippuvuudesta. Kaikkien muuttujien välillä jonkin asteista linearista riippuvuutta. Vahvin riippuvuus on 15-19-vuotiaiden miesten ja 15-19-vuotiaiden naisten välillä ja heikoin 15-19-vuotiaiden naisten ja 20-24-vuotiaiden miesten välillä. Merkittävää kuitenkin on, että korrelaatio on aina positiivinen.

Korrelaatio pysyy vuosien välissä samassa suuruusluokassa kaikkien paitsi miesten 20-24- ja naisten 15-19-vuotiaiden ikäryhmien välissä. Näiden välinen

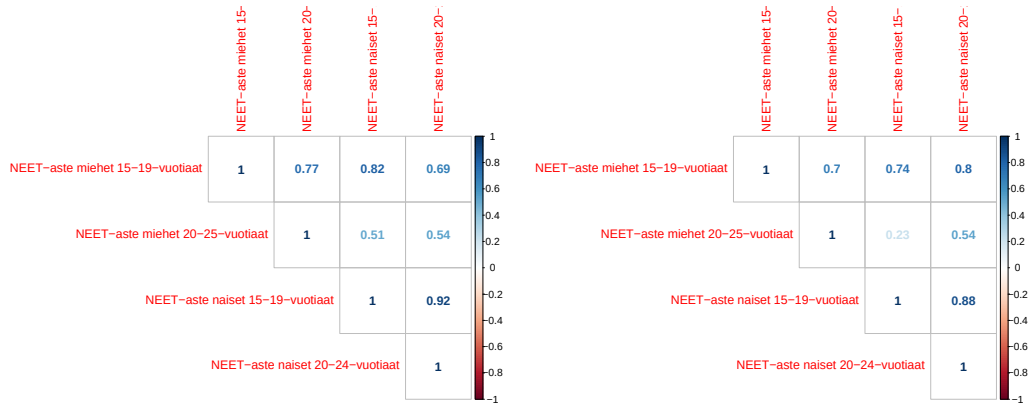
korrelaatio pienenee huomattavasti vuosien 2010 ja 2014 välillä.




Kuva 4.14: Vuoden 2010 nuorten työelämästä syrjäytymistä kuvaavat muuttujat pareittain pistekaavioissa




Kuva 4.15: Vuoden 2014 nuorten työelämästä syrjäytymistä kuvaavat muuttujat pareittain pistekaavioissa




(a) Korrelaatiot vuonna 2010 (b) Korrelaatiot vuonna 2014

Kuva 4.16: Työelämästä syrjäytymistä kuvaavien muuttujien väliset korralaatiot

# Luku 5

# Työttömyysaineiston klusterointi

Työssä on suoritettu klusterointianalyysi työttömyyden ja työelämästä syrjäytymisen aineistoon. Tämän tarkoituksena on löytää aineistosta klustereita eli maajoukkoja, joiden keskinäiset piirteet olisivat mahdollisimman samanlaisia ja klusterien välillä mahdollisimman erilaisia. Tässä kappaleessa esitetään suoritetun klusterointianalyysin tulokset.

Menetelmänä käytetään yhdistävää hierarkkista klusterointia. Menetelmä on kuvattu kolmannessa luvussa. Ennen klusterointia luvussa neljä esitetyt aineistot on standardoitu menetelmällä, joka esitetään luvussa kolme. Klusteroinnin etäisyysmetriikkana käytetään euklidista etäisyyttä. Klusterien kokoamismenetelmänä käytetään maksimietäisyyden minimointia, keskimääräisen etäisyyden minimointia sekä Wardin menetelmää. Myös minimietäisyyden minimointia on käytetty vertailun vuoksi, mutta chaining effectin takia nämä tulokset on esitetty vain liitteessa A.

Tämä kappale rakentuu siten, että eri aineistojen klusteroinnit esitetään erikseen, vielä siten, että eri vuosien aineistojen tulokset ovat erillään. Yksittäisten vuosien tulosten jälkeen tuloksista koostetaan yhteenveto aineistojen sisällä sekä eri aineistojen klustereita vertaillen.

Klusterien muodostaminen on tehty R-ohjelmointikielen hclust-klusterointimenetelmää käyttäen.

Tulosten esittämisen lisäksi tässä kappaleessa arvioidaan menetelmien tuottamia tuloksia kolmannen kappaleen arviointimenetelmien mukaisesti.

Tämän luvun kuvissa 5.1, 5.2, 5.3 ja 5.2 on esitetty klusteroinnin tuottamat dendrogrammit. Kuvia tulkitaan siten, että alhaalla olevat puun lehdet esittävät maita ja solmu kohdasta alaspäin olevat lehdet kuuluvat solmukohdan määrittämään klusteriin. Solmukohtien etäisyyttä mittaava viiva kertoo,

kuinka erilaisia eri klusterit ovat.

## 5.1 Työttömyysaineiston klusterointi

### 5.1.1 Vuoden 2010 aineiston klusterointi

Kuvassa 5.1 on esitetty klusteroitu työttömyysaineisto. Kuvassa esitetään tulokset eri kokoamismetriikoita käyttäen. Jokaista käytettyä klusterien kokoamistapaa käyttäen aineistosta voidaan tunnistaa klusteroitumista. Dunnin indeksin perusteella jako kolmeen klusteriin selittäisi aineistoa parhaiten. Maksimietäisyyden minimoinnissa Dunnin indeksiä tulkiten tulos on indifferentti neljän ja kolmen klusterin välillä. Tunnistetut kolme klusteria voidaan lokeroida maiden sijainnin, kulttuurin, kielen sekä taloudellisen tilanteen perusteella seuraavasti:

1. Poikkeavat havainnot, jotka kärsivät talouskriisistä pahiten. 2. Etelä-Eurooppa ja Itä-Eurooppa. 3. Manner-Eurooppa, Pohjoismaat sekä englanninkieliset maat.

Tämä luokittelu on suuntaa-antava ja jokainen ryhmä sisältää pari poikkeavaa havaintoa.

Kokoamisperusteen ollessa keskimääräinen etäisyys klusterien sisältö olisi seuraava:

Talouskriisitä pahiten kärsineet: Irlanti, Espanja ja Viro.

Etelä-Eurooppa ja Itä-Eurooppa: Slovakia, Portugali, Kreikka, Unkari, Tsekki, Slovenia, Puola, Turkki, Italia, Belgia ja Ranska.

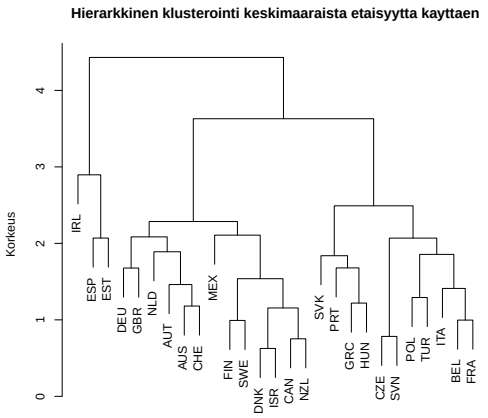
Manner-Eurooppa, Pohjoismaat sekä englannin-kieliset maat: Saksa Iso-Britannia, Alankomaat, Australia, Itävalta, Sveitsi, Meksiko, Suomi, Ruotsi, Tanska, Israel, Canada ja Uusi-Seelanti.

Kokoamistavan ollessa maksimietäisyyden minimointi kolmen klusterin sisältöinä olisivat:

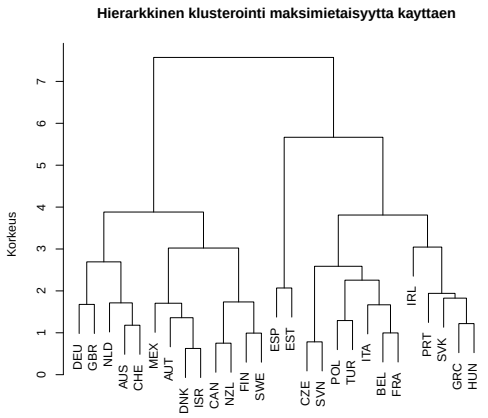
Talouskriisitä pahiten kärsineet: Espanja ja Viro.

Etelä-Eurooppa ja Itä-Eurooppa: Tsekki, Slovenia, Puola, Turkki, Italia, Belgia, Ranksa, Irlanti, Portugali, Slovakia, Kreikka ja Unkari.

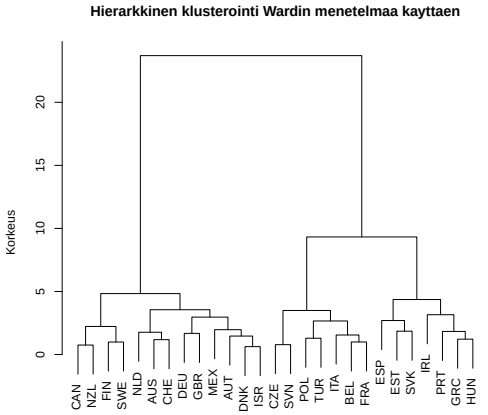
Manner-Eurooppa, Pohjoismaat sekä englanninkieliset maat: Saksa, Iso-Britannia, Alankomaat, Australia, Sveitsi, Meksiko, Itävalta, Tanska, Israel, Kanada, Uusi-Seelanti, Suomi ja Ruotsi.




(a) Dendrogrammi käyttäen keskimääräistä etäisyyttä yhdistämiskriteerinä




(b) Dendrogrammi käyttäen maksimi etäisyyttä yhdistämiskriteerinä




(c) Dendrogrammi käyttäen Wardin menetelmää

Kuva 5.1: Työttömyyttä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2010 aineistolla eri yhdistämismenetelmillä

Wardin menetelmällä klusterien sisältönä olisivat:

Talouskriisitä pahiten kärsineet: Espanja, Viro, Kreikka, Unkari, Irlanti, Portugali ja Slovakia.

Manner-Eurooppa, Pohjoismaat sekä englanninkieliset maat: Saksa, Iso-Britannia,

Alankomaat, Australia, Sveitsi, Meksiko, Itävalta, Tanska, Israel, Kanada, Uusi-Seelanti, Suomi ja Ruotsi.

Etelä-Eurooppa ja Itä-Eurooppa: Tsekki, Slovenia, Puola, Turkki, Italia, Belgia, Ranska.

Menetelmät jakavat aineiston hyvin samanlaisiin joukkoihin. Vain pienimmän talouskriisitä pahiten kärsineiden-klusterin sisältö vaihtelee kokoamistapaa vaihtamalla. Wardin menetelmää käyttäen tämä kyseinen klusteri laajenisi. Kaikilla menetelmillä isoin klusteri, joka sisältää muun muassa Suomen ja Ruotsin, pysyy samana.

Jos klusterien lukumäärää kasvattaa, alkaa menetelmien erot näkyä. Maksimietäisyyden ja keskimääräisen etäisyyden menetelmät erottaisivat Irlannin omaksi klusterikseen, toisin kuin Wardin menetelmä, joka lähtee puolittamaan isointa klusteria.

Aivan kaikki maat eivät näihin luokitteluun mahdu, kuten Israel ja Meksiko, mutta muuten tämä luokittelu kuvaa hyvin klustereiden sisältöä.

### 5.1.2 Vuoden 2014 aineiston klusterointi

Kuvassa 5.2 on vuoden 2014 työttömyysaineiston klusteroinnin tuottamat dendrogrammit. Eri menetelmien tuottamat tulokset eivät ole aivan yhtä yhtenevät kuin vuoden 2010 tapauksessa.

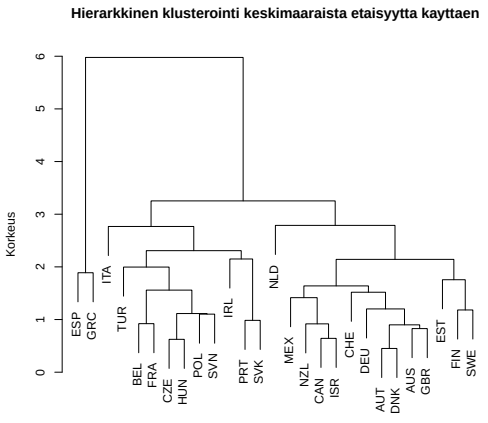
Käyttäen keskimääräistä etäisyyttä kokoamisperusteena klusterit jakautuvat parhaiten kahteen tai kuuteen klusteriin. Kummassakin tapauksessa Espanja ja Kreikka muodostavat yhdessä oman klusterinsa. Kuuden klusterin tapaus on jakautunut samalla tavoin kuin vuoden 2010 tapauksessa kyseisellä menetelmällä, mutta Italia ja Alankomaat ovat muodostaneet yksinäiset klusterit.

Käyttäen maksimietäisyyttä kokoamisperusteena klusterit jakautuvat Dunnin indeksin näkökulmasta parhaiten kahteen klusteriin. Tämä jako olisi samanlainen kuin keskimääräistä etäisyyttä käyttäen, eli Kreikka ja Espanja muodostaisi oman joukkonsa ja loput maat olisivat yhdessä klusterissa. Toiseksi paras vaihtoehto olisi kahdeksan klusteria.

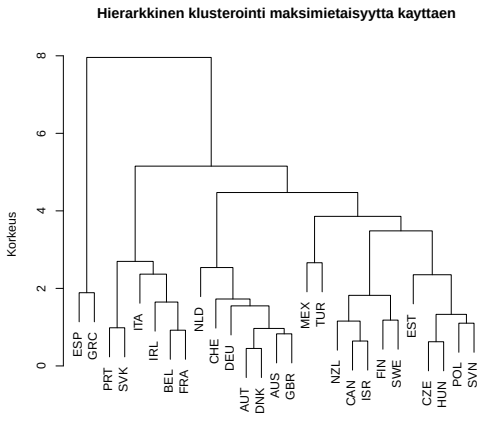
Wardin menetelmällä jako olisi paras kun klusterien määrä olisi kolme. Tällöin jako olisi hyvin samanlainen kuin vuoden 2010 tapauksessa. Tuloksista pystytään tekemään samankaltainen jako kuin vuoden 2010 tapauksessa ja klusterit luokittelaan samoiksi kuin 2010:

Talouskriisitä pahiten kärsineet: Espanja ja Kreikka.

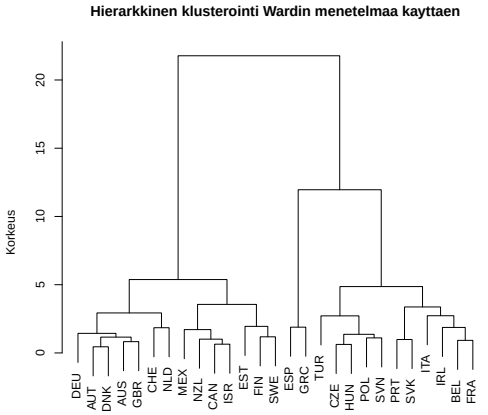
Itä- ja Etelä-Eurooppa: Slovakia, Portugali,Viro, Irlanti, Unkari, Tsekki, Slo-




(a) Dendrogrammi käyttäen klusterien välistä keskimääräistä etäisyyttä yhdistämiskriteerinä




(b) Dendrogrammi käyttäen klusterien välistä maksimi etäisyyttä yhdistämiskriteerinä




(c) Dendrogrammi käyttäen Wardin menetelmää

Kuva 5.2: Työttömyyttä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2014 aineistolla eri yhdistämismenetelmillä

venia, Puola, Turkki, Italia, Belgia ja Ranska.

Pohjoismaat, Manner-Eurooppa, englannin-kieliset valtiot: Saksa Iso-Britannia, Alankomaat, Australia, Itävalta, Sveitsi, Meksiko, Suomi, Ruotsi, Tanska, Israel, Canada ja Uusi-Seelanti.

Vuoden 2014 aineistolla tuloksissa on eroa mitä kokoamistapaa käyttää klusteroinnissa. Maksimietäisyyttä käyttäessä toinen klustereista jää pieneksi, mutta sisältää samat maat kuin keskimääräistä etäisyyttä käyttäessä toisessa klusterissa. Kreikka ja Espanja ovat tavasta riippumatta omana klusterinaan.

Menetelmien eron huomaa siinä, kuinka Kreikka ja Espanja eroavat omaksi joukokseen Wardin menetelmässä vasta kolmannen klusterin kohdalla, kun taas muissa menetelmissä ero tapahtuu jo toisen klusterin kohdalla.

### 5.1.3 Vuosien 2010 ja 2014 klustereiden vertailu

Kumpanakin vuotena OECD-maissa näkyy samanlaiset pääpiirteet työttömyyden klusteroitumisessa. Muutama valtio eroaa todella korkealla työttömyydellään, Etelä- ja Itä-Euroopassa vallitsee korkea työttömyys. Pohjoismaat, Manner-Eurooppa sekä englanninkieliset valtiot pitävät matalan työttömyyden. Liikehdintää klusterien välillä ei tapahdu paljoa vuosien välillä.

Espanjan ja Viron (sekä Irlannin) muodostama korkean työttömyyden klusteri hajoaa ja Viron (ja Irlannin) tilalle siirtyy Kreikka. Menetelmästä riippuen korkean työttömyyden klusteriin, jossa Espanja ja Viro sijaitsevat, kuuluu eri lukumäärä maita.

Kreikka ja Espanja ovat ottaneet todella paljon etäisyyttä muihin maihin. Tämä oli jo nähtävissä aineiston tarkastelussa erillään, sillä maiden väliset erot ovat korostuneet vuoteen 2014 mennessä. Näyttää myös siltä, että muiden maiden työttömyyserot ovat keskenään pienentyneet, sillä Dunnin indeksiin vedoten Kreikan ja Espanjan jälkeen maita on vaikea jakaa enää sisäisesti eri joukkoihin.

Muita klusterien välisiä selkeitä siirtymisiä on tapahtunut Viron ja Irlannin tapauksessa. Viro on tehnyt selkeitä muutoksia. Talouskriisin iski Viroon todella pahasti ja ajoi Viron OECD-maiden matalimmalle työttömyysklusteriin, josta Viron talous nousi vuonna 2011 euroon liittymisen siivittämänä. Nykyinen arvio Viron taloudellisesta tilanteesta onkin lupaava ja Viro nähdään EU:n yhtenä tehokkaimpana taloutena. [42]

Irlantiin talouskriisi iski pahasti. Tästä seurasi monta vuotta kestänyt nouseva työttömyys. Irlannin tilanne alkoi kohentua vuonna 2012 ja vuonna 2014 Irlanti ei enää luokitukaan tämän tutkimuksen klusterianalyysin mukaan OECD-maiden pahimpaan ryhmään.

## 5.2 Nuorten työelämästä syrjäytymisaste-aineiston klusterointi

### 5.2.1 Vuoden 2010 aineiston klusterointi

Kuvassa 5.3 on vuoden 2010 nuorten työelämästä syrjäytymistä kuvaava aineisto klusteroituna.

Israel, Meksiko ja Turkki eroavat muista maista merkittävästi ja vielä siten, että näiden väliset erot ovat suurempia kuin muun massan sisällä olevien klustereiden väliset etäisyydet. Tämä näkyy eri menetelmien tuottamissa klustereissa.

Käyttäen keskimääräistä etäisyyttä kokoamisperusteena klusterit jakautuvat parhaiten kolmeen tai neljään klusteriin. Tällöin tuloksena olisi joukot

1: Israel

2: Meksiko ja Turkki

3: Loput maat

Jakaessa neljään klusteriin:

1: Israel

2: Meksiko

3: Turkki

4: Loput maat

Viiden klusterin tapauksessa Espanja, Irlanti sekä Italia irrottautu omaksi klusterikseen.

Näyttäisi siltä, että Isrealissa miesten työelämästä syrjäytyminen on niin korkealla, että tämä tekijä erottaa maan täysin muista maista. Vastaavasti Meksikossa ja Turkissa naisten työelämästä syrjäytyminen on omissa lukemissaan vuonna 2010.

Näin ollen muiden maiden väliset erot eivät korostu kovin helpolla. Dunnin indeksin perusteella tehty jako näyttäisi siltä, että kolmeen tai neljään klusteriin jako selittäisi aineistoa parhaiten. Jakamalla aineiston useampaan klusteriin antaisi mahdollisuutta tulkita maiden välisiä eroja, mutta Dunnin indeksi, joka mittaa juokkojen sisäisen samanlaisuuden suhdetta joukkojen ulkoiseen erilaisuuteen, tippuu välittömästi. Tällöin Irlannin, Espanjan ja Italian muodostaman klusteri jäisi huomioimatta.

Myös maksimietäisyyden yhdistämistapaa käyttämällä Israel, Meksiko ja Turkki eroavat muusta massasta. Erona keskimääräiseen etäisyyteen on se, että Isreal, Meksiko, Turkki-klusterin sisäiset erot eivät ole niin suuret, että tämä jättäisi muut maat yhtä pahasti yhdeksi massaksi kuin keskimääräisen etäisyyden tapauksessa.

Käyttäen maksimietäisyyttä klusterien yhdistämisperusteena klusterit jakautuvat seuraavasti: Jakaessa kolmeen klusteriin:

1: Israel

2: Meksiko ja Turkki

3: Loput maat.

Kolmannessa klusterissa on matalampi NEET-aste kuin muissa kahdessa.

Jakaessa neljään klusteriin:

1: Israel

2: Meksiko ja Turkki

2: Manner-Eurooppa, Pohjoismaat ja puolet englanninkielisitä maista: Alankomaat, Slovenia, Australia, Kanada, Suomi, Ruotsi, Belgia, Puola, Tsekki, Saksa, Sveitsi, Itävalta ja Tanska

4: Etelä- ja Itä-Eurooppa sekä puolet englanninkielisistä maista: Irlanti, Espanja, Italia, Viro, Unkari, Slovakia, Uusi-Seelanti, Portugali, Kreikka, Ranska ja Iso-Britannia

Neljään klusteriin jako ei Dunnin indeksin näkökulmasta ole kovin hyvä jako, vaan onkin klusterien määrillä 2 - 8 huonoin ratkaisu. Tässä luokittelussa englanninkieliset maat jakautuisivat kahtia Manner- ja Etelä-Eurooppa klusterien välille.

Wardin menetelmällä paras jako Dunnin indeksin näkökulmasta on hieman erilainen. Klusterien määräksi tulisi seitsemän. Espanja, Irlanti ja Italia erottautuvat omaksi klusterikseen ja Meksikon, Turkin ja Israelin joukko onkin lähes samalla tavalla kuin muissa menetelmissä. Tässä jaossa englanninkieliset maat erottuisivat omana joukkonaan ja Pohjoismaat olisivat samassa kategoriassa Itä-Euroopan kanssa.

Kahdeksaan klusteriin jako toisi huomattavasti maiden välisiä eroja esiin sisältäen samat pääpiirteet kuin tämän aineiston muut esitetyt klusteroinnit: Meksiko, Israel ja Turkki ovat omissa kokonaisuuksissan ja Espanja, Irlanti ja Italia muodostavat oman ryhmänsä. Tämän jälkeen maat jakautuvat vielä neljään ryhmään. Alankomaan ja Slovenian muodostama klusteri edustaa matalimman tason nuoren syrjäytymistä jokaisessa kategoriassa.

Merkittävä ero muihin menetelmiin Wardin menetelmässä on, että tämä erottaa Meksikon, Turkin ja Israelin myöhäisessä vaiheessa omiksi klustereikseen. Tässä nähdään Wardin menetelmän eroavaisuus kokonaisvarianssin minimoinnissa, jolloin yksittäiset tapaukset harvoin hallitsevat jakoa. Kuitenkin menetelmästä riippumatta Israel, Meksiko ja Turkki jäävät omaksi kokonaisuudekseen loppujen maiden ulkopuolelle. Klusterien määrä Dunnin indeksin näkökulmasta vaihtelee reilusti riippuen käytetystä menetelmästä.

### 5.2.2 Vuoden 2014 aineiston klusterointi

Kuvassa 5.4 on vuoden 2014 nuorison työelämästä syrjäytymistä kuvaava aineisto klusteroituna.

Vuoden 2014 aineistossa pääpiirteittäin aineisto jakautuu siten, että Meksiko ja Turkki näyttävät olevan omassa kokonaisuudessaan ja Espanja, Kreikka ja Italia omassaan sekä muut maat erillään näistä kahdesta joukosta.

Käyttäen maksimietäisyysmetriikkaa Dunnin etäisyyteen perustuen jako pitäisi tehdä kolmeen tai neljään klusteriin. Kahden klusterin tapauksessa jako ei tuota aivan yhtä hyvää tulosta, mutta merkittävää siinä on, että jaossa Meksiko ja Turkki muodostaisivat oman klusterin ja jättäisivät kaikki muut omaan joukkoonsa.

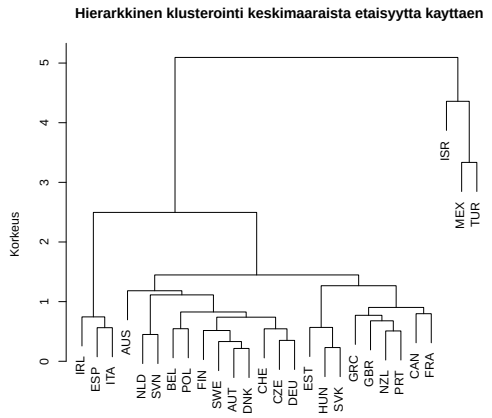
Kolmen klusterin tapauksessa Meksiko ja Turkki ovat oma joukkonsa sekä Italia, Kreikka ja Espanja omansa ja muut maat jäisivät erilleen. Neljään klusteriin jaossa tapahtuisi vain yksi muutos: Meksikon ja Turkin klusteri jakautuisi kahtia. Näin ollen muiden maiden erot eivät meinaa korostua tässä klusterianalyysissä.

Käyttäen keskimääräistä etäisyyttä kokoamismetriikkana aineisto jakaantuu selkeämmin samaan kolmeen klusteriin kuin yllä esitetty maksietäisyysmetriikkaa käyttäen. Tällöin Dunnin indeksi maksimoituu kolmessa klusterissa selkeästi verrattuna muihin.

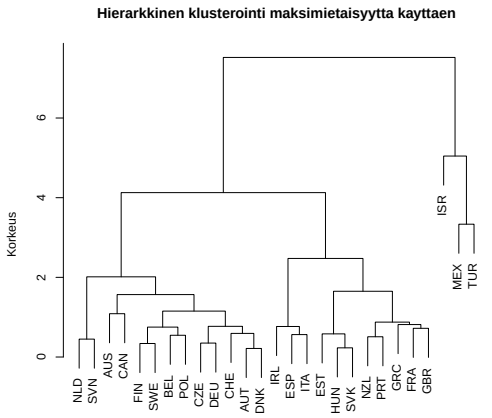
Wardin etäisyyttä käyttäen jako olisi paras, jos klustereita olisi vain kaksi. Tällöin jako olisi:

1: Etelä-Eurooppa (Espanja, Kreikka, Italia, Meksiko ja Turkki) 2: Loput maat

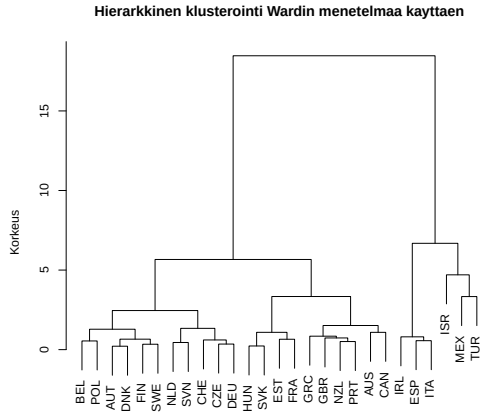
Dunnin indeksi alkaa kohota seuraavan kerran vasta seisemän klusterin kohdalla. Tällöin Espanja, Kreikka ja Italia ovat omassa klusterissaan sekä Meksiko ja Turkki yksin omissaan sekä englanninkieliset maat näyttäisi pysyvän omana kokonaisuutenaan pienin poikkeuksin.




(a) Dendrogrammi käyttäen klusterien välistä keskimääräistä etäisyyttä yhdistämiskriteerinä

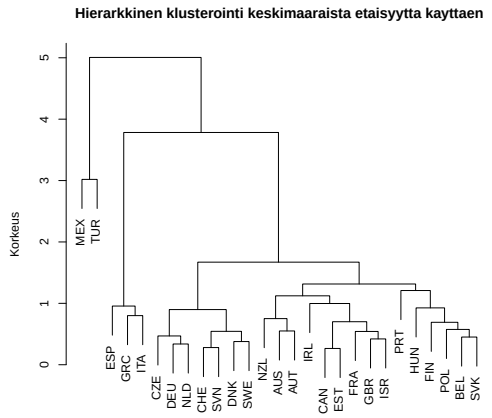



(b) Dendrogrammi käyttäen klusterien välistä maksimi etäisyyttä yhdistämiskriteerinä

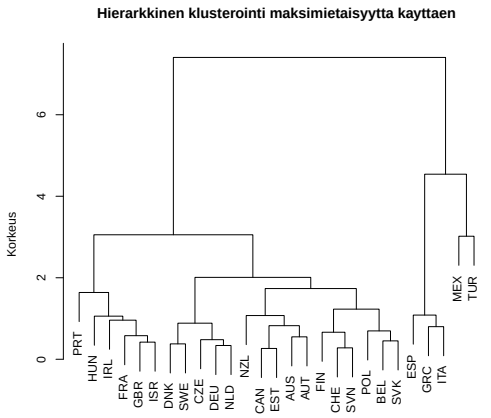



(c) Dendrogrammi käyttäen Wardin menetelmää

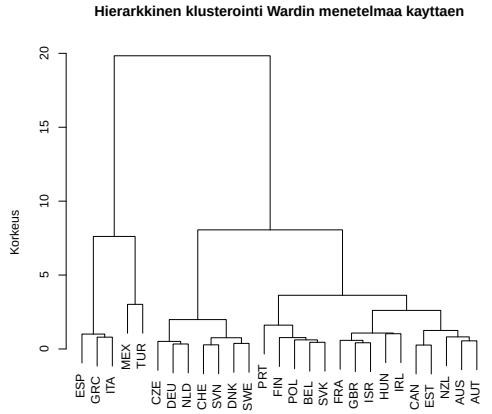
Kuva 5.3: Nuorten työelämästä syrjäytymistä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2010 aineistolla eri yhdistämismenetelmillä




(a) Dendrogrammi käyttäen klusterien välistä keskimääräistä etäisyyttä yhdistämiskriteerinä




(b) Dendrogrammi käyttäen klusterien välistä maksimi etäisyyttä yhdistämiskriteerinä




(c) Dendrogrammi käyttäen Wardin menetelmää

Kuva 5.4: Nuorten työelämästä syrjäytymistä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2014 aineistolla eri yhdistämismenetelmillä

### 5.2.3 Vuosien 2010 ja 2014 klustereiden vertailu

Meksiko ja Turkki ovat säilyttäneet yhtäläisyydet vuosien välillä, mutta Israel on vaihtanut klusteria maiden joukkoon, jossa nuorten työelämästä syrjäy-

tyminen on vähäisempää. Israel on vuoden 2011 aikana tehnyt huomattavia lakimuutoksia, joiden tarkoituksena on ollut parantaa nuorten työllisyystilannetta [43]. Ennen tätä nuorten palkkaamista Israelissa on ollut säännöstöjä, jotka ovat vaikeuttaneet nuorten palkkaamista. Israel muutenkin on oma tapauksensa kun puhutaan nuorista miehistä ja naisista. Missään muussa OECD-maassa ei nuoria velvoita asevelvollisuus yhtä voimakkaasti kuin Israelissa.

Huomattavaa on, että Espanjan, Italian ja Kreikan nuorison työtilanne muuttuu selkeästi yhtenäisemmäksi vuoteen 2014 menessä. Klusterien määrää ja kokoamismetriikkaa vaihtamalla nähdään Espanjan ja Italian samanlaiset rakenteet, mutta niiden eriytyminen muista maista on huomattavasti selkeämpää vuonna 2014. Hieman menetelmästä riipppuen myös englanninkielisistä maista (Kanada, Uusi-Seelanti, Australia, Iso-Britannia ja Irlanti) voidaan tunnistaa yhteisiä piirteitä nuorison työelämästä syrjäytymisessä.

## 5.3 Tulkinta

Aineisto sisältää selkeitä poikkeavia havaintoja: Meksiko, Turkki, Israel, Kreikka ja Espanja. Osa näistä tapauksista selittyy talouskriisillä, osa kulttuurisilla eroilla.

Espanjan ja Kreikan tilanne on vain pahentunut vuoden 2010 jälkeen ja ero muihin talouksiin kasvanut. Italia ja Portugali ovat osassa luokitteluissa samassa kategoriassa, mutta Espanja ja Kreikka ovat työttömyydessä kuitenkin oma joukkonsa.

Israel nähdään nuorten työelämästä syrjäytymisessä omana kategorianaan varsinkin nuorten naisten kohdalla. Tämän tulkitsemiseen vaatii enemmän taustatietoja, sillä Israelin asevelvollisuus tekee kohteesta uniikin muihin nähden. Israelissa on todella hyvä koulutus taso, ja vuonna 2011 tehdyn työlaki uudistuksen mukana maa ei enää poikkea nuorten naisten työelämästä syrjäytymisessä. Israelin kokonaistyöttömyys on lähes kärkitasolla vertailumaista, joten poikkeus nuorten naisten työelämästä syrjätymisessä nähdään selkeästi erillään muista.

Turkin ja Meksikon nuorten miesten työelämästä syrjätyminen on niin suurta, että vain Kreikka on vuonna 2014 kyennyt samoihin lukemiin. Näiden maiden ero kuitenkin on siinä, että Meksikossa ja Turkissa on hyvä työllisyysaste toisin kuin Kreikassa, jossa kaikki työttömyysmittarit näyttävät pahalta. Meksikossa matalaa työttömyysastetta selitelläänkin paljon kulttuurisella erolla, jossa monet työt ovat epävirallisia, joten luku ei välttämättä ole

verrattavissa muihin maihin.

Muita ääritapauksia klusterianalyysissä huomataan työttömyystilanteiden parantumiset: Viro ja Irlanti, jotka kokivat kovan iskun talouteen vuosina 2007-2009, mutta ovat saaneet vuoteen 2014 mennessä taloutensa hyvälle mallille.

Klusterointimenetelmät eivät suuresti eroa tuloksissa, mutta enemmänkin ero näkyy klusteroinnin eri vaiheissa. Wardin menetelmässä yksittäisen erilliset pisteet eivät muodosta omia klustereitaan kovin helpolla toisin kuin muissa menetelmissä. Dunnin indeksin perusteella valituissa lopputuloksissa on kuitenkin hyvin samanlaiset tulokset.

Työttömyyttä ja nuorten työelämästä syrjäytymistä kuvaavat klusterit sisältävät yhtäläisyyksiä ja eroavaisuuksia. Yhtäläistä on vuoden 2014 Kreikka-Espanja-klusteri, joka näkyy molemmissa aineistoissa omana joukkonaan. Myös Italian liittyminen Espanjan ja Kreikan joukkoon näkyy molemmissa aineistoissa. Nuorison työelämästä syrjäytymisen klusterit alkavat muistuttamaan työttömyyden klustereita, kun Meksikon, Turkin ja Israelin klusterit sivuutetaan. Tämä samanlaisuus tietenkin riippuu valituista klusterointimenetelmistä, mutta esimerkiksi vuoden 2010 NEET-klusteri maksimietäisyydellä jaettuna viiteen klusteriin sisältää lähes samat ryhmät kuin saman vuoden työttömyysklusterit pois lukien ääritapaukset (Espanja, Viro, Turkki, Meksiko ja Israel).

Aineistojen välisenä eroavaisuutena on, että nuorten syrjäytymisessä erot ovat todella suuria. Yksittäiset maat muodostavat omia klustereitaan, sillä maidet erot ovat niin radikaalit. Työttömyydessä erot kärjistyvät vasta vuonna 2014 ja niidenkin osalta vain kahdessa maassa ja nämä ovat lähellä toisiaan.

Kummassakin aineistossa yksittäiset maat tekevät hurjia parannuksia luvuissaan. Viron ja Israelin poliittiset muutokset näkyvät vuosien 2010 ja 2014 välillä, mutta vain toisessa aineistossa.

Huomattava ero löytyy Meksikon ja Turkin muodostamasta nuorten syrjäytymisklusterista, jota ei erota laisinkaan tarkastellessa muuta työttömyys aineistoa.

Nuorten syrjäytymisaineisto ei myöskään jaa selkeästi muuta aineistoa kahtia yleisen työttömyyden tavoin, ellei klusterien lukumäärää kasvateta suuremmaksi kuin työttömyyden klusteroinnissa. Tällöin samantyylinen maantieteellinen tai kulttuurillinen jako olisi oikeutettua, kuten työttömyysaineiston klusteroinnissa.

Kreikan nouseminen työttömyyden kärkeen näkyy myös nuorison työelämästä syrjätymisessä, mutta se ei silti ole aivan samalla tasoilla kuin muissa

tämän kategorian kärjissä.

Työttömyyden klusteroinnissa eri menetelmät tuottavat tuloksia, jotka ovat hyvin lähellä toisiaan, mikä viittaisi siihen, että tuloksista voidaan tehdä päätelmiä eteenpäin. Työttömyyden klusterointi ehdottaa tuloksekseen, että OECD-maat voitaisiin jakaa kolmeen joukoon, ja nämä kolme joukkoa edustavat todella korkeaa työttömyyttä, korkeaa työttömyyttä ja matalaa työttömyytä. Korkean työttömyyden maita yhdistää maantieteellinen sijainti Etelä-Euroopa sekä Itä-Eurooppa. Matalan työttömyyden maat kuuluvat joko Pohjoismaihin tai Manner-Euroopaan, englanninkielisiin maihin (Kanada, Australia, Uusi-Seelanti, Iso-Britannia, Irlanti). Tätä tulosta tukee aikaisempi kirjallisuus, jossa Euroopasta todetaan työttömyyden olevan maantieteellisesti klusteroitunutta [15]. Myös aikaisempi tutkimus OECD-maiden kansantaloudellisten mittaireiden, jotka sisältävät myös työttömyysasteen, tunnistavat hyvin lähellä samat klusterit kuin tässä työttä [14].

Työttömyyden klusterissa näkyvä matalantyöttömyyden klusteri nähdään myös työelämästä syrjäytymisen klusteroinnissa, joten tämä voidaan yleistää näkyväksi koko aineistossa. Muita yhteisiä piirteitä tälle joukolle maantieteellinen sijainnin lisäksi ovat sosiaalidemokraattinen politiikka Pohjoismaiden tapauksessa tai liberaali politiikka kuten Kanadan ja Uuden-Seelannin tapauksessa. Yhteistä näille maille on myös matalat tuloerot sekä järjestäytynyt työvoima muiden kuin Meksikon ja Turkin tapauksessa. Näiden kahden maan työttömyysmittarit ovat muutenkin kiisteltyjä tässä kontekstissa.

Samoin kuin työttömyyden klusteroinnissa, nuorison työelämästä syrjäytymisessä eri menetelmien väliset tulokset ovat hyvin samanlaiset, mikä vahvistavat sitä, että tuloksista voidaan tehdä jatkotulkintoja. Nuorten työelämästä syrjätymisessä nähdään hieman erimuotoisia klustereita kuin työttömyydessä. Korkean ja matalan syrjäytymisen klusterin lisäksi nähdään sukupuolien välinen ero. Nuorten miesten työelämästä syrjäytyminen on joissain maissa aivan eri suuruinen kuin naisten ja vielä siten, että eri maat ovat kärjessä eri sukupuolen tilastoissa.

Nuorten työelämästä syrjätymisen klusteroinnin perusteella ei voida tehdä tulkintaa, että jos maassa on paljon nuoria, jotka ovat syrjätyneitä työelämästä, olisi taloudessa myös korkea työttömyys. Israel, Meksiko ja Turkki näyttävät, että tämä ei ole aivan näin suoraviivaista. Tässä voitaisiin pohtia, indikoiko asia kuitenkin huonosta talouden tilasta, mutta työttömyysluvut eivät osaa kertoa sitä. Toisinpäin tilanne näyttäisi kuitenkin toimivan. Korkean työttömyyden maat ilmestyvät korkean työelämästä syrjäytymisen klustereihin. Tässä korkean NEET-asteen klusteri alkaa näyttää melkein samalta kuin korkean työttömyyden ja todella korkean työttömyyden klusterin

yhdistelmä.

# Luku 6

# Pohdinta

Analyysin antamat tulokset herättävät paljon pohdintaa. Menetelmät ja tulkinnat sisältävät monta osaa, joihin ei ole yhtä oikeaa vastausta. Ihan perustavanlaatuiset kysymykset menetelmistä: miten tämän menetelmän valitseminen vaikuttaa tulokseen.

Klusterointiin ei ole oikeaa tapaa ja tulosten tulkitseminen pitää tehdä tapauskohtaisesti. Vaikka työssä käytetään klusterointimenetelmien vertailuun tunnuslukua, olisi tämä sama päättely voitu tehdä hyvin perusteluin aivan toisella tapaa. Esimerkiksi olisi voitu muodostaa klusterit jollakin ennalta määritetyllä merkitsevyystasolla. Tämän työn tuloksia kuitenkin vahvistaa eri menetelmillä saadut samankaltaiset tulokset

Valitun klusterointimenetelmän heikkoutena nähdään tulosten tulkinnassa se, että menetelmä ei kerro missä suhteessa muuttujat myötävaikuttavat tulokseen. Aineistossa nähdään suuria eroja jo yksiulotteisessa analyysissä, jonka perusteella voi epäillä yksittäisten muuttujien hallitsevan klusterien muodostamisessa.

Muuttujien kaksiulotteisessa analyysissä nähdään, että työttömyysasteen ja harmonisoidun työttömyysasteen välillä ei käytännössä ole eroa. Tämän seurauksena on, että menetelmä luo painoarvoa työttömyysasteelle enemmän kuin muille muuttujille. Tämän vaikutusta voisi pohtia poistamalla harmonisoidun työttömyysaste muuttujan.

Näin ollen päästään takaisin pohtimaan mittaammeko asioiden erilaisuutta sille ominaisella tavalla. Perustelemattomaksi jää, onko euklidinen etäisyys usean muuttujan aineistossa oikea valinta. Tälle kysymykselle ei liene oikeaa tai väärää vastausta. Tämän vuoksi on kuitenkin hyvä esitellä vaihoehtosia tapoja, joita on pohdittu työn kolmannessa luvussa.

Aineisto on yleisesti katsottuna huomattavan pieni oppimisalgoritmille. Näin ollen yhden maan pois jättäminen tai uuden lisääminen olisi voinut vaikuttaa huomattavasti tuloksiin. Vaikutus olisi todella merkittävä jos aineistosta poistettaisiin maa, joka on muodostanut yksinään klusterin tuloksissa.

Aineistot käsittelevät vain yhtä vuotta kerrallaan ja näin ollen shokkivaikutukset saattavat tuoda suuria vaikutuksia. Varsinkin talouskriisin seurauksena epästabiilit vuodet voivat luoda tilanteen, joka muuttuu huomattavan nopeasti. Näin ollen suositeltavaa voisi olla tarkastella esimerkiksi vuosien 2010-2014 keskiarvoja ja tarkastella olisiko tulokset samankaltaisia.

Käytetyt menetelmät ovat kuitenkin enemmän kuvaavia kuin määrittäviä, joten tämä enemmänkin avaa kysymyksiä jatkotutkimukselle. Työttömyydessä, kuten muissakin ilmiöissä, on tärkeää ensin tunnistaa ilmiö ennen kuin asiaan voi ennustaa tai suunnitella tekevänsä siihen muutosta. Saatuja klustereita voisi lähteä selittämään esimerkiksi inflaation tai bruttokansantuotteen avulla tai valitsemalla esimerkiksi tuonti-vienti-suhteen, joka voisi selittää Etelä-Euroopan maiden yhtenäisen tilanteen.

# Luku 7

# Yhteenveto

Tässä työssä tutkitaan OECD-maiden työttömyyden klusterirakenteita. Tarkasteltavana olevat työttömyysmuuttujat ovat työttömyysaste, harmonisoitu työttömyysaste, työvoimaan osallistuvien aste, pitkäaikaistyöttömien osuus työttömistä ja nuorisotyöttömyyden aste. Nuorten työelämästä syrjäytymistä mittaava aineisto on jaettu neljään muuttujaan: 15-19-vuotiaat miehet, 20-24-vuotiaat miehet, 15-19-vuotiaat naiset ja 20-24-vuotiaat naiset.

Työttömyyden klusterirakenteita tutkittiin kokoavan hierarkkisen klusteroinnin keinoin. Tässä työssä esiteltiin useita vaihtoehtoisia klusterointimenetelmiä, mutta kokoava hierarkkinen menetelmä sopii tämän työn tehtävään. Valintaan vaikutti muun muassa algoritmin riippumattomuus parametreistä.

Klusterirakenteita löydettiin työttömyys aineistosta sekä työelämästä syrjäytymisen aineistosta. Löydettyjä rakenteita pystytään luokittelemaan käyttäen maantieteellisiä ja kulttuurillisia tekijöitä. Tunnistettuja alueita ovat muun muassa Pohjoismaat, Manner-Eurooppa sekä englanninkieliset valtiot, Etelä- ja Itä-Eurooppa. Tarkastelussa löytyi myös pieni joukko maita, jotka poikkesivat muusta aineistosta suuresti korkeilla työttömyyden luvuilla. Nuorten työelämästä syrjätymisen klusteroituminen ei ollut aivan yhtä selkeää ja tässä näkyi selkeästi poikkeavat yksilöt: Turkki, Meksiko ja Israel.

Tuloksena tulleita klusterirakenteita pystytytään selittämään poliittisilla yhteneväisyyksillä kuten Pohjoismaiden sosiaalidemokratia, englanninkielisten maiden liberaali politiikka ja näille maille yhteistä olevat matalat tuloerot ja järjestäytynyt työvoima.

Nuorten syrjäytyminen työelämästä ei kulje aivan samaa rataa kuin työttömyys. Monet maat, kuten Espanja ja Portugali, joissa on korkea nuorten työelämästä syrjäytymisaste, kokevat myös kovaa työttömyyttä, mutta tämä

ei ole sääntö. Israel, Meksiko ja Turkki ovat todella korkean nuorisosyrjäytymisen maita, mutta eivät kärsi kovasta työttömyydestä. Tätä yritetään selittää maiden eri lähestymistavoilla töiden tekoon sekä esimerkiksi naisen ja nuorten asemalla kotona.

Menetelmän ja lopputuloksen on tarkoitus olla aineistoa kuvaavia, eikä regressiivisiä aineistoa ennustavia. Menetelmän soveltuvuutta voisi kuitenkin mitata tilastollisilla testeillä ja tutkimalla kuinka herkkiä lopputulokset ovat pieneille muutoksille.

Tuloksena saatuja klustereita ei kannata pitää absoluuttisina, mutta suuntaaantavina. Klusteorintianalyysi toimii hyvänä lähestymistapana työttömyyden analysointiin ja tämän perusteella voidaan tehdä parempia tutkimuspäätöksiä siitä mitä aineisto voisi sisältää ja minkälaisia riippuvuuksia voisi löytyä.

Tässä työssä tuloksena saadaan, että OECD-maat sisältävät maantieteellisia ja kulttuurillisesti yhteisiä alueita, jotka rikkovat maiden rajoja, joissa työttömyys ja työelämästä syrjäytyminen on samankaltaista. Tämä on tärkeä asia tunnistettavaksi, sillä se mahdollistaa työttömyyttä korjaavat toimenpiteet, jotka myös koskettavat näitä kokonaisuuksia yksittäisten valtioiden sijaan.

# Kirjallisuutta


[1] Jacqueline O'Reilly, Werner Eichhorst, András Gábos, Kari Hadjivassiliou, David Lain, Janine Leschke, Seamus McGuinness, Lucia Mỳtna Kureková, Tiziana Nazio, Renate Ortlieb, et al. Five characteristics of youth unemployment in europe. *Sage Open*, 5(1):2158244015574962, 2015.

[2] Richard Layard and Stephen Nickell. 9 unemployment in the oecd. *Labour Market and Economic Performance: Europe, Japan and the USA*, page 253, 2016.

[3] Mehmet Odekon. *Booms and Busts: An Encyclopedia of Economic History from the First Stock Market Crash of 1792 to the Current Global Economic Crisis: An Encyclopedia of Economic History from the First Stock Market Crash of 1792 to the Current Global Economic Crisis*. 2015.

[4] International Labour Organization (ILO). Resolution concerning statistics of the economically active population, employment, unemployment and underemployment, 1982.

[5] Mehmet Odekon. *Booms and Busts: An Encyclopedia of Economic History from the First Stock Market Crash of 1792 to the Current Global Economic Crisis*. Routledge, 2015.

[6] Pete Richardson, Laurence Boone, Claude Giorno, Mara Meacci, David Rae, and David Turner. The concept, policy use and measurement of structural unemployment. 2000.

[7] Naceur Khraief, Muhammad Qaiser Shahbaz, Almas Heshmati, and Muhammad Azam Azam. Are unemployment rates in oecd countries stationary? evidence from univariate and panel unit root tests. 2015.

[8] Philipp Heimberger, Jakob Kapeller, and Bernhard Schütz. What's 'structural'about unemployment in europe: On the determinants of the european commission's nairu estimates. 2016.

[9] Katharine G Abraham and Lawrence F Katz. Cyclical unemployment: sectoral shifts or aggregate disturbances?, 1984.

[10] Stephen Machin and Alan Manning. The causes and consequences of longterm unemployment in europe. *Handbook of labor economics*, 3:3085–3139, 1999.

[11] David NF Bell and David G. Blanchflower. Youth unemployment in greece: measuring the challenge. *IZA Journal of European Labor Studies*, 4(1):1–25, 2015.

[12] Karen Robson and Marie Curie Excellence Team. Becoming neet in europe: A comparison of predictors and later-life outcomes. In *Global Network on Inequality Mini-Conference*, volume 22, 2008.

[13] Angana Banerji, Ms Hannah Huidan Lin, and Mr Sergejs Saksonovs. *Youth unemployment in advanced Europe: Okun's law and beyond*. Number 15. International Monetary Fund, 2015.

[14] Francis G Castles and Herbert Obinger. Worlds, families, regimes: Country clusters in european and oecd area public policy. *West European Politics*, 31(1-2):321–344, 2008.

[15] Henry G Overman and Diego Puga. Unemployment clusters across europe's regions and countries. *Economic policy*, 17(34):115–148, 2002.

[16] Lior Rokach. A survey of clustering algorithms. In *Data mining and knowledge discovery handbook*, pages 269–298. Springer, 2009.

[17] Z. Nazari, D. Kang, M. R. Asharif, Y. Sung, and S. Ogawa. A new hierarchical clustering algorithm. In *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 148–152, Nov 2015.

[18] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)19th International Conference in Pattern Recognition (ICPR).

[19] *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[20] Daniel G. Goldstein and Nassim Nicholas Taleb. We don't quite know what we are talking about when we talk about volatility. 2007.

[21] Encyclopedia of Mathematics. Metric.

[22] M.-S. Yang. A survey of fuzzy clustering. *Mathematical and Computer Modelling*, 18(11):1 – 16, 1993.

[23] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584, 2001.

[24] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 515–524, New York, NY, USA, 2002. ACM.

[25] Prabhakar Raghavan Christopher D. Manning and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[26] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[27] Wilmer Leal, Eugenio J. Llanos, Guillermo Restrepo, Carlos F. Suárez, and Manuel Elkin Patarroyo. How frequently do clusters occur in hierarchical clustering analysis? a graph theoretical approach to studying ties in proximity. *Journal of Cheminformatics*, 8(1):1–16, 2016.

[28] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[29] Wei-keng Liao, Ying Liu, and Alok Choudhary. A grid-based clustering algorithm using adaptive mesh refinement. In *7th Workshop on Mining Scientific and Engineering Datasets of SIAM International Conference on Data Mining*, pages 61–69, 2004.

[30] Nancy P Lin, Chung-I Chang, Hao-En Chueh, Hung-Jen Chen, and Wei-Hua Hao. A deflected grid-based algorithm for clustering analysis. *WSEAS Transactions on Computers*, 7(4):125–132, 2008.

[31] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1):9–29, 2001.

[32] Athman Bouguettaya, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, 42(5):2785 – 2797, 2015.

[33] Mark Heckmann and Richard C Bell. A new development to aid interpretation of hierarchical cluster analysis of repertory grid data. *Journal of Constructivist Psychology*, pages 1–14, 2016.

[34] C.-H. Chou, M.-C. Su, and E. Lai. A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications*, 7(2):205–220, 2004.

[35] OECD (2016). Unemployment rate (indicator), 2016. Accessed on 09 April 2016.

[36] OECD (2016). Harmonised unemployment rate (hur) (indicator), 2016. Accessed on 10 April 2016.

[37] OECD (2016). Part-time employment rate (indicator), 2016. Accessed on 10 April 2016.

[38] OECD (2016). Labour force participation rate (indicator), 2016. Accessed on 10 April 2016.

[39] OECD (2016). Long-term unemployment rate (indicator), 2016. Accessed on 10 April 2016.

[40] OECD (2016). Youth unemployment rate (indicator), 2016. Accessed on 10 April 2016.

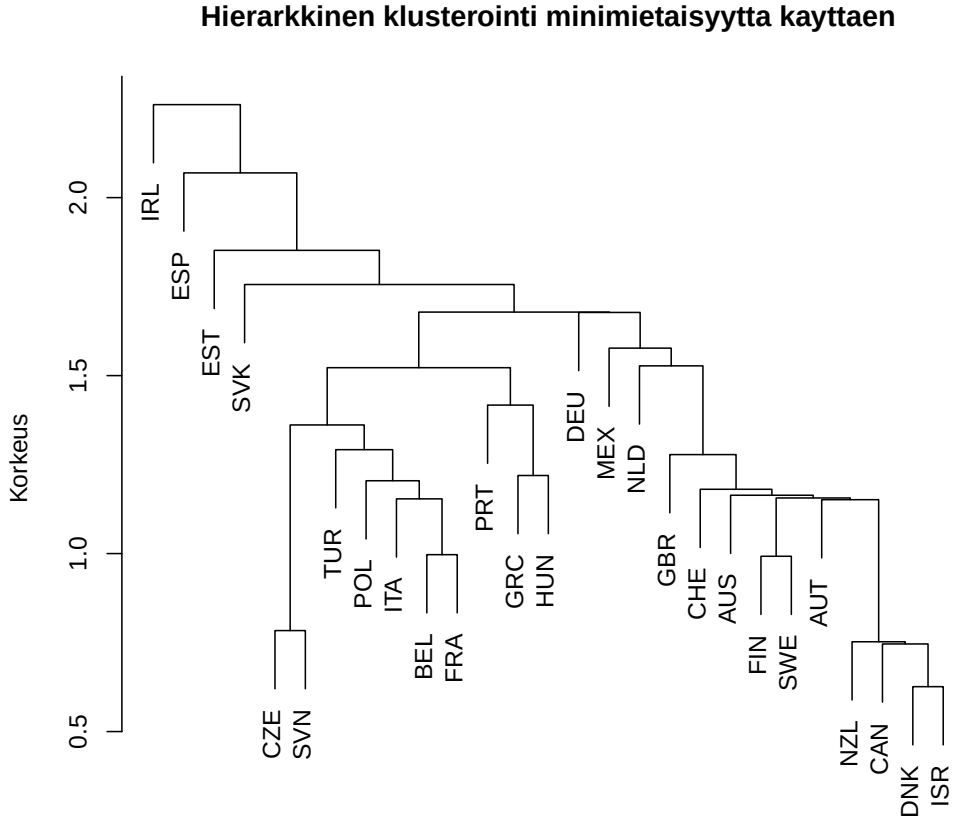
[41] OECD (2016). Youth not in employment, education or training (neet) (indicator), 2016. Accessed on 10 April 2016.

[42] Marge Unt. *Boom and bust effects on youth unemployment in Estonia.* Friedrich-Ebert-Stiftung, Internat. Dialogue, 2012.

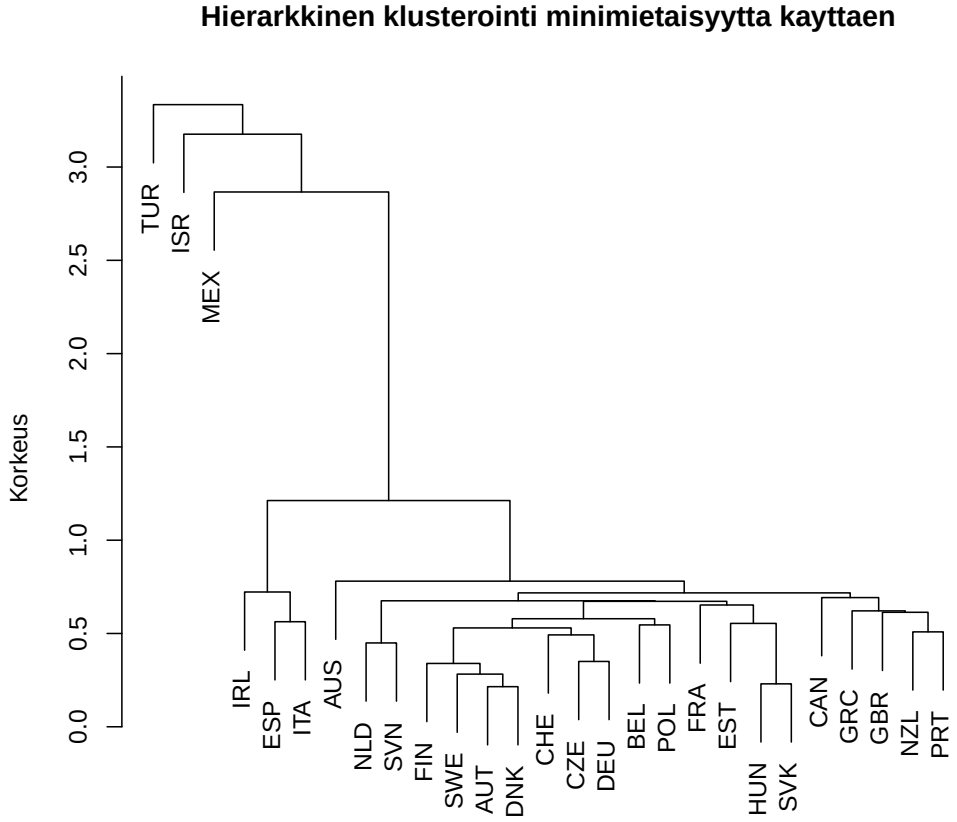
[43] European Training Foundation. Labour market and employment policy in israel, 2015.

Liite A

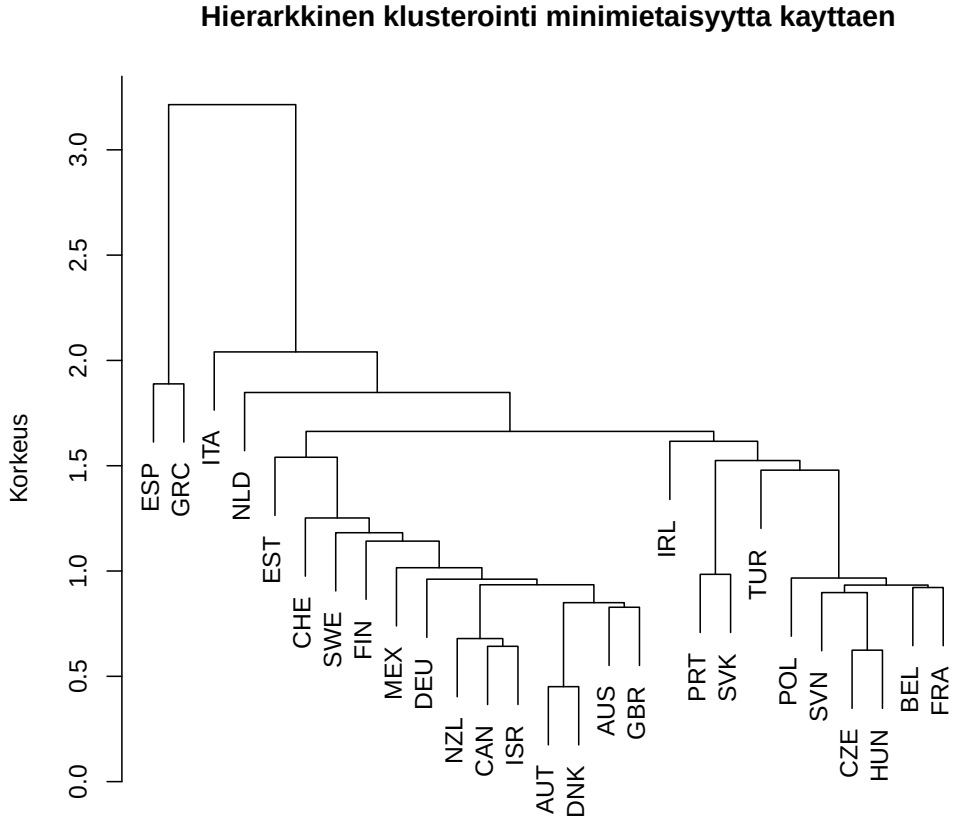
# Ensimmäinen liite: Dendrogrammit minimietäisyyteen perustuvalla klusterien kokoamismenetelmällä



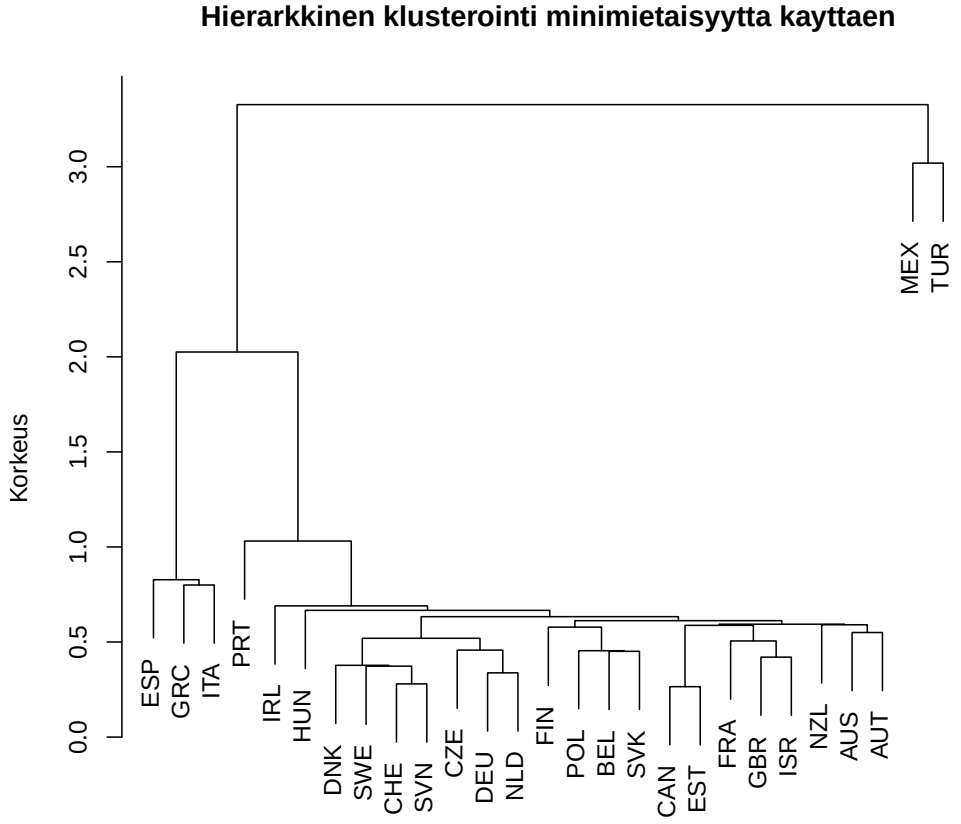

Kuva A.1: Työttömyyttä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2010 aineistolla käyttäen minimietäisyyteen perustuvaa klusterien yhdistämismenetelmää




Kuva A.2: Nuorten työelämästä syrjäytymistä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2010 aineistolla käyttäen minimietäisyyteen perustuvaa klusterien yhdistämismenetelmää




Kuva A.3: Työttömyyttä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2014 aineistolla käyttäen minimietäisyyteen perustuvaa klusterien yhdistämismenetelmää




Kuva A.4: Nuorten työelämästä syrjäytymistä kuvaavan aineiston klusteroinnin tuottamat dendrogrammit vuoden 2014 aineistolla käyttäen minimietäisyyteen perustuvaa klusterien yhdistämismenetelmää