Euroopan taloudellisesta yhtenäisyydestä

Jani Strandberg

Perustieteiden korkeakoulu

Kandidaatintyö
Espoo 16.4.2016

Vastuuopettaja ja ohjaaja:

Prof. Pauliina Ilmonen

Aalto-yliopisto
Perustieteiden
korkeakoulu

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

KANDIDAATINTYÖN
TIIVISTELMÄ

Tekijä: Jani Strandberg

Työn nimi: Euroopan taloudellisesta yhtenäisyydestä

Päivämäärä: 16.4.2016 Kieli: Suomi Sivumäärä: 4+39

Koulutusohjelma: Matematiikka ja systeemianalyysi

Vastuuopettaja ja ohjaaja: Prof. Pauliina Ilmonen

Tässä kandidaatintyössä on tutkittu Euroopan maiden taloudellista yhtenäisyyttä tarkastelemalla yhteensä 24 Euroopan maan osalta kuuden eri OECD:n tietokannasta saatavien talousindikaattorin arvoja. Näiden talousindikaattorien perusteella on klusterianalyysin keinoin pyritty havaitsemaan Euroopan maiden ryhmittymistä toisistaan poikkeaviin ryhmiin eli klustereihin.

Ensisijainen analyysi on suoritettu vuoden 2014 aineistolla, jonka lisäksi pyrkimyksenä on ollut selvittää sekä pohtia syitä mahdollisiin klustereissa ajan yli tapahtuneisiin muutoksiin toistamalla klusterointi samoilla valtioilla ja talousindikaattoreilla erikseen myös varhaisemmalla, vuonna 2006 kerätyllä aineistolla.

Tuloksena on havaittu, että talousindikaattorien perusteella Euroopan valtiot voidaan luokitella ainakin maantieteellisesti jokseenkin eheisiin ryhmiin. Lisäksi on havainnoitu, etteivät klusterit ole ajan suhteen täysin vakioita, vaan vuosien 2006 ja 2014 välilä on tapahtunut yksittäisten valtioiden uudelleenryhmittymistä.

Avainsanat: Bruttokansantuote (BKT), Velkaantumisaste, Hintatasoindeksi, Nettovienti, Klusterianalyysi, Agglomeratiivinen Hierarkkinen Klusterointi, Etäisyysmitta, Wardin Algoritmi

Sisällysluettelo

Symbolit ja lyhenteet

Symbolit

$\mathcal{I}_T$	Pistepilven kokonaisvaihtelu
$\mathcal{I}_B$	Pistepilven ryhmien välinen vaihtelu
$\mathcal{I}_W$	Pistepilven ryhmien sisäinen vaihtelu
Δ_{ij}	Matriisin Δ rivin i, sarakkeen j elementti

Operaattorit

$d(A, B)$	Klusterien A ja B välinen etäisyys (tietyllä etäisyysmitalla)
$Q^{'}$	Matriisin Q transpoosi
$\sum_{i\|x_i \in A_j}$	summa indeksin i yli ehdolla $x_i \in A_j$

Lyhenteet

GDPcap	BKT per capita
GDPh	BKT työtuntia kohden
D/E (ratio)	Yrityksen velkaantumisaste
PLI	Hintatasoindeksi
HUR	Harmonisoitu työttömyysprosentti
Net trade	Nettovienti

1 Johdanto

Nykypäivän ajankohtaisten kriisien keskellä Euroopan maiden johtajat pyrkivät parhaansa mukaan ylläpitämään euroalueen sekä laajemmankin Euroopan alueen yhtenäisenä säilyttääkseen useiden vuosikymmenien aikana luodut kumppanuussuhteet ehjinä. Syy näihin ponnisteluihin liittynee vahvasti Euroopan historiaan, jossa se on kerta toisensa jälkeen jakautunut vastakkaisiin leireihin, jonka seurauksena naapurivaltiot ovat joutuneet keskenään tuhoisiin konflikteihin.

Erityisesti kahden peräkkäisen maailmansodan jälkeen heräsi poliitikkojen ja kansankin keskuudessa yhteisymmärrys siitä, että samankaltaisten konfliktien välttämiseksi olisi Euroopan sisällä tapahduttava jonkinlaista yhtenäistymistä. Aiemmin riitoja aiheuttaneen, yhteisen ideologian etsimisen, sijaan useat poliitikot näkivät instrumentin konfliktien välttämiseen Euroopan taloudellisessa yhtenäistymisessä. [1]

Tämän yhteinäistymisen voidaan katsoa alkaneen Euroopan hiili- ja teräsyhteisön perustamisesta, jolla pyrkimyksenä oli asettaa Euroopan hiili- ja teräsvarannot yksittäisten valtioiden ulottumattomille siten, että yksittäiset valtiot eivät pystyisi saamaan mielivaltaisesti sotaan tarvittavia raaka-aineita. Euroopan taloudellista yhtenäisyyttä on edelleen kasvattanut tämän yhteistyön pohjalta syntynyt Euroopan Unioni, ja sitä kautta Euroopan markkinoiden vapautuminen sekä hieman myöhemmin myös yhteinen valuutta, euro. [2]

Euroopan taloudellinen yhtenäistyminen on siis ollut Euroopassa useiden tahojen tavoitteena jo vuosikymmenten ajan. Miltä tilanne näyttää nykypäivän Euroopassa?

Tämän kandidaatintyön tavoitteena on tutkia Euroopan maiden taloudellista yhtenäisyyttä selvittämällä, onko Euroopan maissa todellisuudessa havaittavissa jonkinlaista ryhmittymistä toisistaan poikkeaviin havaintoryhmiin, eli klustereihin. Tutkimukseen on otettu mukaan yhteensä 24 Euroopan maata, joista jokaisesta on kerätty OECD:n julkisesta tietokannasta tietoa yhteensä kuudesta erilaisesta, taloudelliseen toimintaan liittyvästä indikaattorista.

Tutkimus toteutetaan klusterianalyysin keinoin, jossa pyrkimyksenä on siis matemaattisesti ryhmitellä moniulotteiset havaintopisteet, tässä tapauksessa Euroopan valtiot, erilaisiin ryhmiin. Klusterointiin liittyy vahvasti "samankaltaisuuden" käsite, eli se, miten klusterien "samankaltaisuutta" mitataan. Tähän ei ole yhtä oikeaa vastausta, vaan huomioon tulee ottaa analyysin asiayhteys.

Kun klusterien väliseen etäisyyteen sopiva mitta on määritelty, on luotava tai käytettävä valittua etäisyysmittaa soveltava algoritmi, joka ryhmittelee havainnot ryhmiin tämän etäisyysmitan perusteella. Tässä työssä keskitytään hierarkkisiin klusterointialgoritmeihin, joissa ryhmien lukumäärää ei kiinnitetä etukäteen.

Hierarkkisessa klusteroinnissa lisähaasteena on "oikean" klusterien lukumäärän arviointi tulosten perusteella, eikä tähänkään ole saatavissa selkeästi perusteltua, asiayhteydestä riippumatonta ratkaisua. Valitun "katkaisutason" perusteella voidaan klusteroinnin tulokset raportoida, sekä tulkita syntyvien klusterien taustalla vaikuttavia tekijöitä.

Klusterianalyysi suoritetaan ensisijaisesti uusimmalla saatavissa olevalla aineistolla, joka on kerätty vuodelta 2014. Tämän lisäksi työssä arvioidaan myös klustereissa mahdollisia ajassa tapahtuneita muutoksia. Tämä arviointi suoritetaan toistamalla

vastaava klusterianalyysi samoilla mailla ja talousindikaattoreilla käyttämällä varhaisempaa, vuodelta 2006 saatua aineistoa. Tuloksien perusteella voidaan arvioida, millaisia muutoksia aineiston perusteella on havaittavissa Euroopan maiden taloudellisessa ryhmittymisessä, ja miten stabiililta Euroopan talouden ryhmittyminen näiden tulosten valossa näyttää.

Kaikki työssä suoritettavat laskut toteutetaan R-ohjelmointikielellä.

2 Tutkimusaineisto ja -menetelmät

2.1 Klusterianalyysin perusidea

Klusterianalyysissa on pohjimmiltaan kyse erilaisten ryhmien, niin kutsuttujen klusterien, rakentaminen moniulotteisista havainnoista koostuvasta aineistosta. Tavoitteena on luokitella moniulotteiset havainnot eri ryhmiin jonkinlaisen kriteerin perusteella siten, että ryhmän sisällä olevat havainnot ovat jollakin määrittelytavalla "lähempänä" toisiaan kuin muihin ryhmiin kuuluvat havainnot.

Klusterianalyysi ei ole mikään spesifi algoritmi, vaan yleisemmällä tasolla määritelty ryhmittelyongelma, johon voi löytää ratkaisun monin eri tavoin. Näissä tavoissa voi olla suuriakin eroja siinä, miten klusteri määritellään, ja miten niitä tulisi etsiä. Yleinen periaate on pyrkiä klusteroinnissa siihen, että ryhmien väliset erot ovat mahdollisimman suuria, mutta samalla ryhmien sisäiset erot mahdollisimman pieniä. Kenties johtuen sen yleisestä luonteesta, on klusterianalyysilla paljon sovelluskohteita eri aloilla, kuten sosiologiassa, lääketieteessä ja tietotekniikassa.

Mikä tahansa klusterianalyysi perustuu kahden toisiinsa liittyvän tekijän määrittelyyn [3]:

1. Havaintojen välisen etäisyyden tai ”erilaisuuden” mitta. Tämän mitan määrittely on olennainen klusteroinnin kannalta, sillä sen perusteella havainnot luokitellaan ryhmiin. Etäisyys kertoo myös ryhmän sisäisestä vaihtelusta: mitä lähempänä ryhmän havainnot ovat toisiaan, sitä homogeenisempi kyseinen ryhmä on.

2. Klusteroinnissa käytettävä algoritmi. Tämän tarkoituksena on rakentaa edellä mainitun etäisyysmitan perusteella havainnoista ryhmät siten, että ryhmät ovat sisäisesti mahdollisimman homogeenisia, mutta ryhmät keskenään mahdollisimman erilaisia.

Näiden asioiden määrittely on välttämätöntä klusterianalyysin kannalta, ja niissä on otettava konteksti ja käytettävän aineiston luonne huomioon. Seuraavissa kappaleissa käydään näihin liittyviä seikkoja tarkemmin läpi, ja samalla myös pyritään perustelemaan tässä työssä tehdyt valinnat näiden suhteen.

2.1.1 Klusterien etäisyyksien mittaaminen

Havaintojen välisien etäisyyksien laskemiseen mittaamiseen liittyy kaksi ongelmaa:

- Miten mitataan kahden *havaintopisteen* välistä etäisyyttä?
- Miten mitataan kahden *klusterin* välistä etäisyyttä, kun klustereissa on enemmän kuin yksi havaintopiste?

Kahden havaintopisteen välisen etäisyysmitan tulisi olla sopiva käytettävän aineiston tyyppiin nähden. Tässä tutkimuksessa käytettävä aineisto on luonteeltaan

numeerista, minkä takia luonteva etäisyysmitta tälle moniulotteiselle, kvantitatiiviselle aineistolle on yleistetty euklidinen etäisyys: p-ulotteisille havainnoille $x, y \in \mathbb{R}^p$, tämä etäisyys määritellään seuraavasti:

$$d^2(x, y) = (x - y)'Q(x - y) \tag{1}$$

,jossa $Q \in \mathbb{R}^{p \times p}$ on etäisyyksissä käytettävä metriikka. Tämä on siis matriisi, jolla kutakin havainnon p:stä dimensiosta voidaan painottaa etäisyyden laskemisessa halutulla tavalla. Yksinkertaisin ja paljon käytetty metriikka on normaali euklidinen etäisyys: $Q = I$, eli Q on identiteettimatriisi. Tässä tapauksessa jokaisella x:n komponentilla on painoarvo, joka riippuu kyseisen komponentin yksiköstä tai varianssista.

Varianssista riippuva painoarvo ei välttämättä ole mielekäs tutkimusasetelmissa, jossa moniulotteisten havaintojen komponenttien skaalat eroavat toisistaan huomattavasti. Tämä johtuu siitä, että tällaisissa tapauksissa skaalaltaan ja varianssiltaan suuremmat komponentit usein dominoivat havaintojen välisten etäisyyksien saamia arvoja, jolloin ne saavat siis suuremman painoarvon skaalaltaan ja varianssiltaan pienempiin komponentteihin verrattuna.[4]

Kun halutaan, että jokaisella komponentilla on yhtä suuri painoarvo sen mittayksiköstä ja varianssista riippumatta, niin havaintojen komponentit voidaan skaalata siten, että jokaisen muuttujan varianssi vakioidaan jakamalla komponentti sitä vastaavalla varianssilla. Tällöin käytetään niin kutsuttua pääkomponenttimetriikkaa, jossa

$$Q = \text{diag}(1/s_1^2, ..., 1/s_p^2) \tag{2}$$

Tällöin Q on siis diagonaalimatriisi, jonka alkiot ovat havaintojen komponenttien $1...p$ varianssit. Tämän metriikan käytöllä on sama vaikutus kuin muuttujien standardointiin liittyvällä skaalauksella, jossa havaintoarvojen komponentit jaetaan vastaavan komponentin keskihajonnalla, joka vakioi kaikkien havaintojen komponenttien varianssin ykkösen suuruiseksi.

Muita mahdollisia kvantitatiiviseen aineistoon käytettäviä etäisyysmittoja ovat esimerkiksi Manhattan-metriikka $||x-y||_1 = \sum_{i=1}^{p} |x_i - y_i|$, eli havaintojen koordinaattien erotusten itseisarvojen summa, sekä maksimietäisyys $||x - y||_\infty = \max_i |x_i - y_i|$, eli suurin havaintojen välisten komponenttien etäisyys. Näitä etäisyysmittoja ei kuitenkaan käytetä tässä työssä johtuen siitä, että käytettävään aineistoon soveltuu paremmmin edellä mainittu pääkomponenttimetriikka.

Tämän työn aineiston tapauksessa havaintoarvot ovat kvantitatiivisia, ja komponenttien skaalat ovat toisistaan selvästi poikkeavia (vrt. esimerkiksi BKT per capita, työttömyysprosentti), joten havaintojen komponenttien varianssin vakiointi on tässä tapauksessa perusteltua. Siitä syystä etäisyysmittana työssä käytetään edellä mainittua pääkomponenttimetriikkaa, tai vaihtoehtoisesti ajateltuna normaalia euklidista etäisyyttä skaalatulle aineistolle.

Kvalitatiivisten muuttujien tapauksessa käytetään usein hieman erilaisia etäisyysmittoja, joista kenties tyypillisin on khiin neliö -etäisyys, jonka avulla esimerkiksi kategorisesta aineistosta syntyvän kontingenssitaulun riviprofiilien väliset etäisyydet lasketaan. Toisin kuin euklidinen etäisyys, joka antaisi riviprofiilien muodostaman

taulun kaikille sarakkeille yhtä suuren painoarvon, khiin neliö- etäisyys painottaa etäisyyksiä vastaavan sarakkeen frekvenssillä, millä on varianssia standardoiva vaikutus:

$$Q = \text{diag}(1/f_{.1}, ..., 1/f_{.K}) \tag{3}$$

,jossa $f_{.k}$ on kontingenssitaulun sarakkeen k reunafrekvenssi. Tässä työssä käytettävä aineisto on kuitenkin luonteeltaan kvantitatiivista, joten kvalitatiiviset etäisyysmitat eivät ole asiayhteyteen nähden relevantteja.

Metriikan päättämisen jälkeen tärkeää on määritellä se tapa, jolla eri klusterien välistä etäisyyttä mitataan, ja lisäksi formuloida tapa laskea klusterien väliset etäisyydet uudelleen algoritmissa tapahtuvan yhdistämisen jälkeen. Kenties yksinkertaisin tapa tähän on määritellä kahden klusterin etäisyys asettamalla se yhtä suureksi kuin pienin eri klusterien sisältämien havaintojen etäisyys. Tällöin klusterien A ja B etäisyys määritellään

$$d(A, B) = \min_{a_i \in A, b_j \in B} d(a_i, b_j) \tag{4}$$

,jossa d on valittu kahden pisteen etäisyys $\mathbb{R}^p$:ssä. Yhdistämisen jälkeiset etäisyydet seuraavat helposti: esimerkiksi kun $C = A \cup B$, niin etäisyydet tästä yhdistetystä klusterista muihin klustereihin saadaan suoraan määritelmän mukaan

$$d(C, D) = \min\{d(A, D), d(B, D)\} \tag{5}$$

Tämä ”lähimmän naapurin” menetelmä on suosiossa sen yksinkertaisten laskutoimitusten vuoksi. Ongelmana tällä etäisyysmitalla on kuitenkin se, että klusterien ”läheisyyteen” riittää pelkästään yksi pari (a_i, b_j), mutta kaikki muut pisteparit voivat olla hyvin kaukana toisistaan. Kyseisellä etäisyysmitalla onkin taipumus tuottaa pitkiä ja ohuita klustereita, jossa samassa klusterissa olevat lähinaapurit ovat lähellä toisiaan, mutta saman klusterin vastakkaiset päät voivat olla huomattavasti kauempana toisistaan kuin joistakin muiden klusterien elementeistä. Tätä ilmiötä kutsutaan ketjuuntumiseksi (eng.*chaining*) [5]

Toinen, jokseenkin vastakkainen vaihtoehto on mitata klusterien välistä etäisyyttä minimietäisyyden sijaan maksimietäisyydellä, jolloin klusterien A ja B etäisyys on määritelmän mukaan:

$$d(A, B) = \max_{a_i \in A, b_j \in B} d(a_i, b_j) \tag{6}$$

Yhdistämisen $C = A \cup B$ jälkeen etäisyydet muihin klustereihin saadaan jälleen suoraan määritelmästä:

$$d(C, D) = \max\{d(A, D), d(B, D)\} \tag{7}$$

Tämä "kauimman naapurin" menetelmä ei kärsi edellä mainitusta ketjuuntumisesta, ja sillä on taipumus tuottaa usein kompakteja ja läpimitaltaan saman kokoisia klustereita. Se ei myöskään ota huomioon klusterien sisäistä rakennetta. [5]

Kolmas, jokseenkin sofistikoituneempi tapa etäisyysmitan vaihtoehto on ottaa huomioon kahden eri klusterin kaikkien mahdollisten parien välinen etäisyys. Tämä voidaan laskea ottamalla keskiarvo kaikkien mahdollisten klusterien A ja B havaintojen välisistä etäisyyksistä:

$$d(A,B) = \frac{1}{n_A n_B} \sum_{a_i \in A} \sum_{b_j \in B} d(a_i, b_j) \tag{8}$$

Tällöin yhdistämisen $C = A \cup B$ jälkeiset etäisyydet voidaan todistaa saatavan seuraavassa muodossa:

$$d(C,D) = \frac{n_A d(A,D) + n_B d(B,D)}{n_A + n_B} \tag{9}$$

Tällä etäisyysmitalla on taipumusta liittää yhteen klustereita, joiden varianssi on pieni. Sitä voidaan pitää eräänlaisena minimi-ja maksimietäisyyksien välimuotona, joka kykenee myös paremmin ottamaan huomioon klusterien rakenteen. [5]

Edellä mainituissa etäisyysmitoissa ei eksplisiittisesti huomioida klusterien homogeenisuutta. Haluaisimme kuitenkin työssä käyttää sellaista etäisyysmittaa, jossa huomioidaan sekä klusterien sisäinen että niiden välinen vaihtelu. Eräs tällainen algoritmi, jossa ryhmittelykriteeri perustuu klusterien muodostamien pistepilvien väliseen ja sisäiseen hajontaan, on nimeltään Wardin algoritmi. Tämä esitellään perusteluineen tarkemmin kappaleessa 2.1.3.

On tärkeää huomioida, että edellä mainittujen lisäksi on toki myös paljon muita tapoja klusterien etäisyyksien mittaamiseen, joskin edellä mainitut ovat eniten käytettyjä ja tunnnetuimpia metodeja. "Oikean" klusterien välisen etäisyyden mittaamistavan valitsemiseen ei kuitenkaan ole selkeää, kaikissa tilanteissa parhaiten toimivaa tapaa.

2.1.2 Algoritmin valinta

Klusterianalyysissa on lähtökohtaisesti tärkeää tietää myös se, onko haettavien klusterien lukumäärä etukäteen tiedossa. Mikäli ennakkotiedon perusteella tiedetään tarkkaan ryhmien "oikea" määrä, nin klusterien lukumäärä on järkevää kiinnittää analyysissa, ja tällä on vaikutus myös klusterointialgoritmin toimintaan. Tällaisille tapauksille onkin olemassa monia erilaisia klusterointialgoritmeja, joista tunnetuin lienee liikkuvien keskipisteiden (eng.*K-means*) – algoritmi. [6] Tämän algoritmin tavoitteena on siis ryhmitellä havaintopisteet k eri ryhmään, jossa parametri k on käyttäjän valitsema.

Tämän työn tapauksessa klusterien lukumäärän kiinnittäminen etukäteen ei kuitenkaan ole mielekästä, sillä emme halua asettaa mitään ennakko-oletuksia, joilla voisi olla vaikutusta lopputulokseen. Täten näihin algoritmeihin, joissa klusterien lukumäärä on kiinnitettävä parametri, ei perehdytä tässä sen tarkemmin.

Sen sijaan työssä on mielekästä käyttää sellaista algoritmia, jossa klusterien lukumäärää ei kiinnitetä etukäteen. Tällaisia asetelmaa kutsutaan hierarkkiseksi klusteroinniksi, sillä ajatuksena on nimen mukaisesti rakentaa klustereista eräänlainen hierarkia. Tällaiset algoritmit voidaan jakaa kahteen toiminnaltaan vastakkaissuuntaiseen luokkaan:[5]

1. Agglomeratiiviset algoritmit, joissa aloitetaan hienoimmasta mahdollisesta ryhmäjaosta, eli jokainen havainto muodostaa oman ryhmänsä, ja lopetetaan karkeimpaan mahdolliseen, eli jokainen havainto kuuluu samaan klusteriin.

2. Divisiiviset algoritmit: aloitetaan karkeimmasta mahdollisesta ryhmäjaosta ja lopetetaan hienoimpaan mahdolliseen ryhmäjakoon.

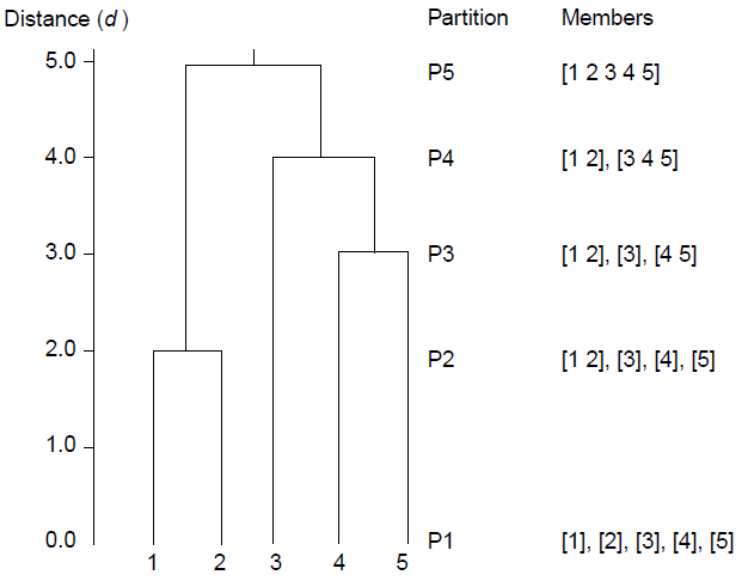
Tässä yhteydessä keskitytään ensimmäiseen luokkaan, eli agglomeratiivisiin algoritmeihin, sillä nämä ovat hierarkkisista metodeista laajimmin käytettyjä. Kuten aiemmin mainittu, näissä algoritmeissa ideana on aloittaa hienoimmasta ryhmäjaosta ja päätyä karkeimpaan ryhmäjakoon. Jos käytössä on yhteensä n havaintoa, niin algoritmi koostuu kokonaisuudessaan $n-1$ askeleesta: siirrytään alun n luokasta askel kerrallaan $n-1, n-2, \ldots$, ja lopulta yhteen luokkaan, joka sisältää kaikki havaintopisteet. Jokaisella askeleella ryhmitellään yhteen kaksi määrätyltä etäisyysmitaltaan lähimpänä toisiaan olevaa ryhmää. Tällöin siis askeleella, jossa jäljellä on k ryhmää, on laskettava $\binom{k}{2}$ etäisyyttä näiden ryhmien välillä, jonka jälkeen kaksi lähintä ryhmää muodostavat uuden ryhmän. Algoritmi tarvitsee edellä mainitun klusterien välisen etäisyysmitan, joka on havaintojen ryhmittelyn kannalta avainasemassa algoritmin toiminnassa.

Koska algoritmi on konstruoitu siten, että jokaisella askeleella ryhmitellään yhteen kaksi lähimpänä toisiaan olevaa ryhmää, niin nämä algoritmin minimietäisyydet kasvavat algoritmin edetessä. Tällöin klustereista syntyvä hierarkia voidaan indeksoida näiden pienimpien etäisyyksien perusteella, jossa nämä arvot ovat kasvavassa järjestyksessä. Tämä indeksi kertoo siis algoritmin kullakin askeleella olevan ”aggregaatiotason”. Matala taso viittaa siihen, että yhdistetyt ryhmät ovat lähellä toisiaan, ja korkea taso viittaa taas ryhmien olevan kauempana toisistaan. [3]

Algoritmissa järjestyksessä tapahtuvat ryhmittelyt voidaan esittää graafisesti niin kutsutun dendrogrammin eli puukaavion avulla. Tässä kaaviossa esitetään havaintoyksilöt ja kuhunkin aggregaatiotasoon liittyvät havaintoyksilöiden klusterit. Toisella akselilla esitetään siis havaintoyksilöiden nimet, ja toisella akselilla aggregaatiotaso. Koska korkea aggregaatiotaso viittaa toisistaan poikkeavien ryhmien yhdistämiseen, niin ryhmittelyistä syntyvä ”puu” on usein mielekästä katkaista jollakin halutulla tasolla. Koska klusteroinnissa on tavoitteena löytää aineistosta erilaisia ryhmiä, on selkeästi järkevää katkaista jollekin mielekkäälle tasolle ennen kaikkien havaintojen yhdistymistä samaan klusteriin. Tällä valitulla tasolla katkaistun puun oksat antavat suoraan tähän tasoon liittyvät havaintojen klusterit. Esimerkki dendrogrammista on esitetty kuvassa 1: [5]

Tässä esimerkissä ensimmäinen askel ryhmittelisi havainnot 1 ja 2 omaan klusteriinsa, sillä niiden etäisyys on pienin mahdollinen, noin 2.0. Tämän jälkeen ryhmitellään havainnot 4 ja 5 yhteen, jonka jälkeen havainto 3 ryhmitellään vielä klusterin [4 5] kanssa yhteen, ja lopulta jäljelle jäävät klusterit yhdistämällä kaikki havaintopisteet ovat samassa klusterissa.

Mitään yleispäteviä metodeita ”oikean” katkaisun tason valitsemiseen ei kuitenkaan ole olemassa. On täysin asiayhteydestä riippuvaista, mikä aggregaatiotaso on miellettävissä "pieneksi" ja mikä "suureksi". Katkaisun tasoa voidaan toki pyrkiä validoimaan tilastollisen analyysin keinoin, esimerkiksi tarkastelemalla ryhmien tilastollisia tunnuslukuja.



Kuva 1: Esimerkki dendrogrammista

2.1.3 Wardin algoritmi

Wardin algoritmin perusajatuksena on se, että ryhmittelyn laatua voidaan arvioida jakamalla aineiston kokonaisvaihtelu (dispersio) ryhmien väliseen ja ryhmien sisäiseen vaihteluun:

$$\mathcal{I}_T = \mathcal{I}_B + \mathcal{I}_W \tag{10}$$

Olkoon käytettävissä oleva aineisto n pisteen muodostama p-ulotteinen pilvi, $x_i \in \mathbb{R}^p, i = 1 \dots n$, jossa jokaiselle pisteelle asetetaan painoarvo $p_i > 0$ siten että $\sum_{i=1}^{n} p_i = 1$. Tämä painoarvo on useimmiten $p_i = \frac{1}{n}$, kuten myös tämän työn tapauksessa, mutta kyseinen merkintätapa on yleistettävissä myös tilanteisiin, joissa painoarvot poikkeavat toisistaan. Pistepilven keskipiste g on tällöin:

$$g = \sum_{i=1}^{n} p_i x_i \tag{11}$$

Aineiston kokonaisvaihtelu $\mathcal{I}_T$ on määritelmän mukaan jokaisen havaintopisteen painoarvolla painotettu keskimääräinen etäisyys havaintojen keskipisteeseen:

$$\mathcal{I}_T = \sum_{i=1}^{n} p_i d^2(x_i, g) = \sum_{i=1}^{n} p_i (x_i - g)' Q (x_i - g) \tag{12}$$

Oletetaan sitten, että havainnot on jaettu k eri klusteriin esimerkiksi jollakin klusterialgoritmin askeleella. Klusterin j painoarvo saadaan laskemalla siihen kuuluvien havaintopisteiden painoarvot yhteen:

$$P_j = \sum_{i|x_i \in A_j} p_i, \ j = 1, \ldots, k \tag{13}$$

Tällöin klusterin j keskipiste on siis

$$g_j = \frac{1}{P_j} \sum_{i|x_i \in A_j} p_i x_i \tag{14}$$

Ryhmien välistä vaihtelua voidaan mitata tällöin laskemalla kunkin klusterin keskipisteiden etäisyys kaikkien havaintojen yhteiseen keskipisteeseen, ottaen kunkin ryhmän painot P_j huomioon:

$$\mathcal{I}_B = \sum_{j=1}^{k} P_j d^2(g_j, g) = \sum_{j=1}^{k} P_j (g_j - g)' Q (g_j - g) \tag{15}$$

Ryhmien sisäinen vaihtelu taas lasketaan painotettuna keskiarvona yksittäisten klusterien sisäisistä vaihteluista. Soveltamalla yhtälöä 12 yksittäiseen klusteriin A_j, yksittäisen klusterin sisäinen vaihtelu saadaan seuraavalla tavalla:

$$\mathcal{I}_{A_j} = \frac{1}{P_j} \sum_{i|x_i \in A_j} p_i d^2(x_i, g_j) = \frac{1}{P_j} \sum_{i|x_i \in A_j} p_i (x_i - g_j)' Q (x_i - g_j), \tag{16}$$

jolloin ryhmien sisäinen vaihtelu on painotettu summa yksittäisten ryhmien sisäistestä vaihtelusta:

$$\mathcal{I}_W = \sum_{j=1}^{k} P_j \mathcal{I}_{A_j} \tag{17}$$

Mitä suurempi on ryhmien välinen vaihtelu $\mathcal{I}_B$ osuus kokonaisvaihtelusta, $\mathcal{I}_T$, sitä parempana ryhmittelyä voidaan pitää. Tämä on kuitenkin suurimmillaan agglomeratiivisen algoritmin alussa, jossa jokainen havaintopiste muodostaa oman klusterinsa, sillä tällöin pätee

$$\mathcal{I}_B = \mathcal{I}_T, \ \mathcal{I}_W = 0 \tag{18}$$

Algoritmin jokaisella askeleella siirrytään $k+1$ klusterista k klusteriin ryhmittelemällä kaksi toisiaan lähimpänä olevaa klusteria yhteen. Tällöin väistämättä ryhmien sisäinen vaihtelu kasvaa ja ryhmien välinen vaihtelu pienenee. Algoritmin loppuvaiheessa ollaan taas tilanteessa, jossa ryhmien välinen vaihtelu on nolla, ja ryhmien sisäinen vaihtelu on yhtä suuri kuin kokonaisvaihtelu, koska kaikki havainnot ovat yhdessä ryhmässä:

$$\mathcal{I}_B = 0, \ \mathcal{I}_W = \mathcal{I}_T \tag{19}$$

Wardin algoritmissa ideana onkin valita "pisin mahdollinen polku" alkutilanteesta $\mathcal{I}_B = \mathcal{I}_T$ lopputilanteeseen $\mathcal{I}_B = 0$. Tämä tapahtuu määrittelemällä klusterien läheisyyskriteeriksi se ryhmien välisen vaihtelun määrä, joka menetetään ryhmittelemällä kyseiset kaksi klusteria yhteen. Jokaisella askeleella siis ryhmitellään yhteen ne

kaksi klusteria, joiden yhdistämisen seurauksena ryhmien välinen vaihtelu pienenee mahdollisimman vähän. [3]

Klusterien A ja B yhdistämisestä aiheutuva ryhmävaihtelun menetys saadaan laskemalla:

$$d(A,B) = \frac{P_A P_B}{P_A + P_B} d^2(g_A, g_B) \tag{20}$$

Tämä on se etäisyysmitta, jota käytetään Wardin algoritmissa ryhmittelykriteerinä kappaleessa 2.1.1 kuvattujen etäisyysmittojen sijaan. Voidaan myös näyttää, että yhdistetyn klusterin $C = A \cup B$ etäisyys muihin klustereihin on tällä kriteerillä mitattuna:

$$d(C,D) = \frac{(P_A + P_D)d(A,D) + (P_B + P_D)d(B,D) - P_D d(A,B)}{P_A + P_B + P_D} \tag{21}$$

Wardin algoritmi voidaan siis kokonaisuudessaan jakaa seuraaviin askeliin:

1. Aloitetaan klusterointi hienoimmasta jaosta, jossa jokainen havainto muodostaa oman klusterinsa.

2. Käyttämällä kaavaa 20, lasketaan kaikkien mahdollisten klusteriparien väliset etäisyydet. Näistä muodostuu symmetrinen $n \times n$ etäisyysmatriisi Δ, jonka elementit ovat
$$\Delta_{ij} = \frac{p_i p_j}{p_i + p_j} d^2(i,j) \tag{22}$$

3. Etsitään pienin arvo matriisista Δ, ja muodostetaan siihen liittyvistä klustereista uusi klusteri, jonka paino on siihen liittyvien klusterien painojen summa.

4. Lasketaan uuden klusterin ja muiden klusterien väliset etäisyydet käyttämällä kaavaa 21. Nämä muodostavat puolestaan $(n-1) \times (n-1)$ etäisyysmatriisin Δ.

5. Toistetaan kohtia 3-4 niin kauan kunnes kaikki havainnot on ryhmitelty samaan klusteriin.

Kohdassa 3 laskettavat minimietäisyydet antavat kullakin algoritmin askeleella k aggregaatiotason $\delta^{(k)}$, jotka summautuvat koko algoritmin $n-1$ askeleessa kokonaisvaihtelun suuruiseksi. Lopussa ryhmien välinen vaihtelu on nolla, jolloin menetetty ryhmien välinen vaihtelu on yhtä suuri kuin kokonaisvaihtelu:

$$\sum_{k=1}^{n-1} \delta^{(k)} = \mathcal{I}_T \tag{23}$$

Tässä työssä käytetään tätä agglomeratiivista Wardin algoritmia valtioiden ryhmittelyyn. Etuna tässä on se, ettei se kärsi yksittäisiin etäisyyksiin perustuvien metodien ongelmista, ja se perustuu puhtaiden etäisyyksien sijasta jokseenkin sofistikoituneempaan tapaan mitata klusterien eroja niiden välisen vaihtelun avulla.

Lisäksi tulosten tarkasteluissa todettiin, että eri etäisyysmittojen antamat tulokset eivät muutenkaan poikenneet toisistaan huomattavasti, minkä takia keskittyminen yksittäiseen metodiin on enemmän perusteltua. Liitteessä A on esitetty muita etäisyysmittoja (minimi, maksimi, keskiarvo) käytettäessä tuloksena syntyvät dendrogrammit. Näistä on todettavissa, että algoritmien toiminnassa ei tämän aineiston tapauksessa ole kovin huomattavia eroja.

Toisin kuin muut edellä mainitut etäisysmitat, Wardin kriteerin käyttäminen etäisyysmittana vaatii sen, että aineisto voidaan esittää euklidisessa avaruudessa. Tämä ei ole käyttämämme aineiston tapauksessa ongelma, sillä kaikki aineisto on kvantitatiivista, eikä sisällä kategorisia muuttujia.

2.2 Aineisto

Työssä käytettävä tutkimusaineisto on haettu julkisesti saatavilla olevasta OECD:n tietokannasta (OECD.org). Tutkimuksessa on otettu aineiston saatavuus huomioon ottaen seuraavat 24 Euroopan maata. Tässä yhteydessä esitetään myös maihin liittyvät lyhenteet, jotka esiintyvät klusteroinnin tuloksena syntyvissä dendrogrammeissa:

Taulukko 1: Työssä käytettävät Euroopan valtiot

Valtio	Lyhenne
Alankomaat	NLD
Belgia	BEL
Espanja	ESP
Irlanti	IRL
Iso-Britannia	GBR
Italia	ITA
Itävalta	AUT
Kreikka	GRC
Luxemburg	LUX
Norja	NOR
Portugali	PRT
Puola	POL
Ranksa	FRA
Ruotsi	SWE
Saksa	DEU
Slovakia	SVK
Slovenia	SVN
Suomi	FIN
Sveitsi	CHE
Tanska	DNK
Tsekki	CZE
Turkki	TUR
Unkari	HUN
Viro	EST

Nämä valtiot siis muodostavat aineiston moniulotteiset havaintopisteet, kun jokaisen valtion taloudelliset indikaattorit kerätään yhteen. Havaintopisteiden kokonaismäärä on siis tässä tapauksessa $n = 24$. Havaintopisteiden valinnassa on päällimmäisenä tavoitteena ollut pyrkimys ottaa mahdollisimman monta valtiota mukaan siten, että puuttuvia tai selkeästi vanhentuneita tietoja ei ole.

Euroopan maista tutkimuksen ulkopuolelle on jätetty siis Bulgaria, Kroatia, Kypros, Latvia, Liettua, Malta ja Romania. Tämä johtuu siitä, että kyseisistä maista oli puuttuvaa tai selkeästi vanhentunutta tietoa joko osassa tai kaikissa tutkimukseen mukaan valituissa muuttujissa. Tutkimuksen eheyden vuoksi puuttuvien tietojen maita ei otettu analyysiin mukaan, sillä tämä vaatisi erillisiä toimenpiteitä, eikä puuttuvasta aineistosta voine tehdä yhtä luotettavia tulkintoja.

2.3 Käytettävät muuttujat

Tutkimuksessa käytetään seuraavia muuttujia:

- BKT per capita, 2014
- Hintatasoindeksi, 2014
- BKT työtuntia kohden, 2014
- Finanssiyritysten velkaantumisaste, 2014
- Nettovienti (% BKT), 2014
- Harmonisoitu työttömyysprosentti, 2014

Yhteensä jokaisesta valtiosta on tutkimukseen otettu siis 6 erilaista taloudellista indikaattoria. Tässä tapauksessa siis käytetty havaintoaineisto on dimensioltaan 6-ulotteinen: $p = 6$.

Kokonaisuudessaan työhön on pyritty ottamaan sellaisia muuttujia, jotka yhdessä kuvaavat valtioiden taloudellista toimintaa useasta eri näkökulmasta: Taloudellinen hyvinvointi, hintataso, työn tuottavuus, yritysten pääomarakenne, kansainvälinen kaupankäynti sekä työttömyys ovat paljolti yhteydessä toisiinsa, mutta kukin kertoo omaa tarinaansa maan taloudellisesta toiminnasta.

Euroopan sisältä löytyy huomattavan erikokoisia talouksia, minkä takia muuttujien valinnassa on olennaista huomioida myös niiden mittayksiköt. Työssä on tarkoituksellisesti vältetty sellaisia mittareita, joiden arvo riippuu olennaisesti maan talouden koosta. Tällainen on esimerkiksi puhdas bruttokansantuote, jolla suuremman tuotannon valtioilla on huomattavasti suurempi arvo kuin pienemmän tuotannon mailla. Mikäli tällaisia mittareita otettaisiin analyysiin mukaan, niin nämä talouksien kokoeroista johtuvat erot mahdollisesti dominoisivat muita eroja klusteroinnissa, jolloin ryhmittely tapahtuisi enimmäkseen talouden koon perusteella. Tästä syystä analyysiin otettavat muuttujat on pyritty valitsemaan siten, että valtioiden vertailukelpoisuus säilyy.

Seuraavaksi esitellään ja taustoitetaan kutakin tutkimukseen valittua muuttujaa.

2.3.1 Bruttokansantuote (BKT) per capita

Bruttokansantuote eli BKT on kaikkien yksittäisen valtion rajojen sisällä tuotettujen lopputuotteiden -ja palveluiden markkina-arvo tietyllä aikavälillä, yleensä vuodessa tai neljännesvuodessa. [7].

BKT siis mittaa valtion kokonaistuotannon arvoa, mutta samalla se mittaa myös kokonaistuloa. Tästä syystä sitä usein käytetään taloudellisen suorituskyvyn ja elintason mittana. Bruttokansantuote on kuitenkin aggregaattisuure, eikä ota huomioon valtion kokoa. Tästä syystä BKT ilmaistaan usein myös muodossa BKT per capita, joka saadaan yksinkertaisesti jakamalla valtion BKT sen väkiluvulla. Tällöin eri kokoiset maat tulevat siis näiltä arvoiltaan paremmin vertailukelpoisiksi. OECD:n ylläpitämässä tilastossa indikaattorin arvo on annettu dollareissa per capita. [8]

Bruttokansantuotetta käytetään sen ongelmista huolimatta elintason mittarina, koska sen sisältö on laajasti ymmärretty ja muiden kvantitatiivisten hyvinvoinnin mittareiden rakentaminen on hankalaa. [9] Myös suurin osa muista työssä käytettävistä muuttujista on jollakin tavalla liitoksissa bruttokansantuotteeseen sen yleismaailmallisen luonteen vuoksi.

2.3.2 Hintatasoindeksi (PLI)

Hintatasoindeksillä tarkoitetaan tässä yhteydessä OECD- tilastoissa olevaa price level index (PLI)- tilastoa, joka on määritelmän mukaan ostovoimapariteetin suhde nimelliseen valuuttakurssiin. Ostovoimapariteetti taas on valuuttojen välinen arvosuhde, eli se valuuttakurssi, jolla laskettuna kahden maan hyödykekorin hinta on täysin sama yhteiseksi valuutaksi muutettuna.[10] Indeksin arvo saadaan tästä jakamalla arvosuhde nimellisellä valuuttakurssilla, joka puolestaan on valuuttamarkkinoilla määräytyvä valuuttojen vaihtosuhde.

Näillä hintatasoindekseillä voidaan siis mitata hintatasossa eri maiden välillä olevia eroja indikoimalla sen yhteisen valuutan määrän, joka tarvitaan samansuuruisen tuotekorin ostamiseen eri maissa. Hintatasoindeksin tarkoituksena ei kuitenkaan ole asettaa valtioita hintatason suhteen tiukkaan järjestykseen, vaan se kertoo pikemminkin valtion hintatason suuruusluokasta suhteessa vertailutasoon. OECD:n tilastoimassa aineistossa indeksin vertailutaso on OECD-maiden keskiarvo, joka saa indeksissä arvon 100.[11] Jos valtion PLI on suurempi kuin 100, niin kyseisen valtion hintataso on suhteessa kalliimpi vertailutasoon (OECD) nähden. Samaten valtioiden, joiden PLI on alle 100:n, hintataso on vertailutasoon nähden matalampi.

2.3.3 BKT työtuntia kohden

BKT työtuntia kohden on yksi tapa mitata työn tuottavuutta. OECD:n määritelmän mukaan työn tuottavuus on yleisemmin määriteltynä talouden ulostulon (eng. *output*) eli tuotannon määrän suhde saman talouden sisääntulojen (eng. *input*) määrään. [12]

Talouden tuotantoa mitataan tyypillisesti jo edellä mainitulla bruttokansantuotteella, mutta talouden sisääntuloja voidaan mitata usealla eri tavalla. OECD:n tilastoissa lähtökohdaksi on otettu työvoiman panos työtunneissa mitattuna, joka

määritellään kaikkien tuotantoon osallistuvien henkilöiden yhteenlaskettuna tuntimääränä, ja koko indikaattorin yksikkönä toimii yhdysvaltain dollari. [13] Muita mahdollisia tapoja työvoiman panoksen mittaamiseen olisivat esimerkiksi töissä käyvien ihmisten lukumäärä tai tarjottavien töiden lukumäärä. [14]

BKT työtuntia kohden kertoo, miten tehokkaasti maan työvoimaa käytetään yhdessä muiden tuotantotekijöiden kanssa valtion sisäisissä tuotantoprosesseissa. Mittari ei sen sijaan kykene arvoimaan työn tuottavuuden yksilötason elementtejä, kuten työntekijöiden omaa kapasiteettia tai heidän työskentelytehokkuuttaan. Mittarin arvo myös riippuu paljolti myös muiden tuotantotekijöiden, kuten pääoman, teknologian ja organisaatioiden olemassaolosta.

2.3.4 Finanssiritysten velkaantumisaste (D/E -ratio)

Yritykset voivat rahoittaa toimintaansa joko velalla tai pääomalla. Yrityksen velkaantumisasteella (eng. *debt-equity ratio*) tarkoitetaan yleisellä tasolla yrityksen osakeomistajien saatavien, eli oman pääoman, suhde yrityksen ulkopuolisten tahojen saatavien, eli velan, määrään.[15] Yrityksen velka ja oma pääoma yhdessä muodostavat yrityksen varallisuuden, ja velkaantumisaste kertoo, miten tämä varallisuus on jakautunut velkaan ja omaan pääomaan.

Velkaantumisasteen mittaamiselle ei ole yhtä oikeaa mittaamistapaa. OECD:n tilastoissa tämä mittari lasketaan jakamalla yritysten kokonaisvelka kaikkien osakkeiden ja muiden pääomasitoumusten kokonaisarvolla.[16] Tässä mittarissa velka on määritelty tyypilliseen tapaan yrityksen lainojen, vakuutusten, eläkkeiden ja ostovelkojen summana. Nimittäjä eli oma pääoma on määritelty osakkeiden ja muun pääoman summaksi, mutta lähdesivuston mukaan tätä parempi olisi niin kutsuttu "omat varat", joka kattaisi edellä mainitun lisäksi myös yritysten nettovarallisuuden. Ei-rahallisesta omaisuudesta ei kuitenkaan ole riittävästi aineistoa saatavilla, jonka takia tätä ei ole otettu nimittäjään mukaan.

Kun velkaantumisaste on 1, on yrityksillä keskimäärin yhtä paljon velkoja kuin omaa pääomaa. Tätä suurempi velkaantumisaste yleisesti tarkoittaa sitä, että yritykset rahoittavat toimintaansa suuremmalta osin velalla kuin omalla pääomalla. Vastaavasti tätä pienempi velkaantumisaste viittaa siihen, että yritykset on rahoitettu suuremmalta osin omalla pääomalla velan sijaan.

Velkaantumisaste on suoraan yhteydessä yritysten rahoitusriskiin. Ottaessaan velkaa yritykset sitoutuvat maksamaan lainan korkoineen takaisin. Yritysten voiton kasvaessa kaikki hyödyt menevät pääoman omistajille lainantajien saadessa vain kiinteää korkomaksua. Tässä mielessä korkea velkaantumisaste siis voi parantaa oman pääoman tuottoa, mutta asialla on myös kääntöpuoli: voittojen pienentyessä lainat korkoineen tulee edelleen maksaa pois, jolloin oman pääoman omistajat puolestaan kärsivät eniten. Pahimmassa tapauksessa yritys ajautuu konkurssiin, jolloin jäljellä olevasta omaisuudesta maksetaan ensin velat pois, jolloin osakkeenomistajille ei välttämättä jää mitään jäljelle. [15]

Velanoton hyöty yrityksissä perustuu paljolti siihen, että veloista maksettavat korot ovat verotuksessa vähennyskelpoinen erä. Tällöin suurempi velan osuus rahoituksesta tarkoittaa potentiaalisesti suurempia verovähennyksiä. Korkea velkaisuusas-

te toisaalta tarkoittaa myös suurempia korkomaksuja, jolloin yrityksen konkurssin todennäköisyys kasvaa. Sekä velanantajat että osakkeenomistajat ottavat tähän liittyvät riskit ja kustannukset huomioon arvottaessaan yritystä. Lisäksi verovähennyksistä on hyötyä vain jos yritys tekee positiivista liikevoittoa, mikä on vähemmän todennäköistä yritysten maksaessa suuria korkomaksuja. [17]

Työssä käytetään vuoden 2014 aineistoa kaikkien muiden valtioiden paitsi Sveitsin osalta. Kyseisen tiedon puutteesta johtuen Sveitsille kyseinen arvo on otettu edeltävältä vuodelta 2013.

2.3.5 Nettovienti (% BKT)

Euroopan valtiot ovat niin kutsuttuja avoimia talouksia, jotka ovat vapaasti jatkuvassa interaktiossa keskenään: ne ostavat toisiltaan ja myyvät toisilleen erilaisia palveluja ja tuotteita markkinoiden välityksellä. Nämä interaktiot muodostavat olennaisen osan kansantalouksien toiminnasta. Valtion viennillä tarkoitetaan kotimaassa tuotettujen palveluita ja tuotteita, jotka myydään ulkomaille, ja tuonnilla tarkoitetaan ulkomailla tuotettuja palveluita ja tuotteita, jotka myydään kotimaahan. Valtion nettoviennillä tarkoitetaan jonakin rajattuna ajanjaksona, yleensä yhden vuoden aikana, tapahtuvan viennin ja tuonnin rahallisen arvon erotusta kyseisen valtion rahayksiköissä mitattuna. Se kertoo siis valtion viennin ja tuonnin välisestä suhteesta.[18] Nettovienti on vahvasti yhteydessä myös BKT:een, sillä se on yksi puhtaan BKT:n useista komponenteista.

Positiivinen nettovienti kertoo valtion vaihtotaseen ylijäämästä, jolloin valtion ulkomaille vietyjen tuotteiden ja palveluiden kokonaisarvo on ulkomailta maahan tuotujen tuotteiden ja palveluiden kokonaisarvoa suurempi. Tätä pidetään yleisesti hyvän taloustilanteen indikaattorina valtiolle, koska tällöin valtion sisään virtaa enemmän rahaa kansantalouden kiertoon kuin sitä virtaa ulos muihin kansantalouksiin. Vastaavasti negatiivinen nettovienti kertoo valtion kokonaistuonnin arvon olevan kokonaisvientiä suurempi, jolloin valtion vaihtotase on alijäämäinen ja vaihtokauppaan liittyvät nettorahavirrat suuntautuvat ulkomaille. Jos taas vienti ja tuonti ovat yhtä suuret, niin vaihtokauppa on tasapainossa ja ulkomaille suuntautuva nettorahavirta on nolla. [18]

Samoin kuin yksittäisten yritysten velan tapauksessa, ei koko kansantalouden velkaantuminen vaihtotaseen alijäämäisyyden seurauksena ole yksinkertaisesti tulkittavissa pelkästään ongelmaksi, sillä esimerkiksi ulkomailta hankittavat palvelut voivat parantaa kotimaisen tuotannon tehokkuutta ja olla sitä kautta taloudellisesti kannattavaa velkojen kertymisestä huolimatta.

OECD:n tilastoissa saatavilla on valtioiden vuotuisen viennin ja tuonnin prosentuaaliset osuudet kyseisen valtion BKT:sta, sekä nettoviennin suuruus Yhdysvaltain dollareissa. [19] Maiden vertailukelpoisuuden ylläpitämiseksi on mielekästä käyttää prosenttiosuuksia valtioiden bruttokansantuotteesta, koska tällöin valtion koolla ei enää ole merkittävää vaikutusta. Lisäksi prosenttiosuudesta mielenkiintoisen tekee se, että se kertoo osittain myös siitä, kuinka olennaista kansainvälinen kauppa on kyseiselle kansantaloudelle. Nettoviennin prosenttiosuutta ei ole OECD:n tilastoista suoraan saatavilla, mutta tämä on helposti laskettavissa vähentämällä viennin

prosenttiosuudesta tuonnin prosenttiosuus:

$$V_{netto} = V_{\%} - T_{\%} \tag{24}$$

, jossa $V_{\%}$ ja $T_{\%}$ ovat saatavilla olevat viennin ja tuonnin prosenttiosuudet. Vastaava tulos saataisiin myös jakamalla valtion nettoviennin suuruus dollareina sen BKT:lla.

2.3.6 Harmonisoitu työttömyysprosentti (HUR)

Valtion elintasosta kertoo paljon suorien taloudellisten indikaattorien lisäksi myös se työttömyystaso, joka valtiossa tyypillisesti vallitsee. Työttömyyttä mitataan jatkuvasti erilaisten toimijoiden puolesta kyselytutkimusten avulla. Näiden perusteella ihmiset voidaan luokitella kolmeen ryhmään:

- Työllistetyt: Ne henkilöt, jotka työskentelevät palkattuna työntekijänä, yrittäjänä tai palkattomana työntekijänä perheenjäsenen yrityksessä. Tähän ryhmään lasketaan sekä täysi- että osa-aikaiset työntekijät, sekä syystä tai toisesta väliaikaisesti töistään poissaolevat henkilöt.

- Työttömät: Ne henkilöt, jotka eivät ole töissä, mutta ovat halukkaita ja kykeneviä työskentelemään. Tällä tarkoitetaan siis aktiivisesti töitä etsiviä ihmisiä, jotka eivät tällä hetkellä osallistu valtion tuotteiden ja palveluiden tuotantoon.

- Ei työvoimassa: Ne henkilöt, jotka eivät sovi edellä mainittuihin ryhmiin, kuten täysiaikaiset opiskelijat, kotiäidit -ja isät sekä eläkeläiset.

Tällöin työvoiman kokonaismäärä on yhtä suuri kuin työllistettyjen ja työttömien summa. Työttömyysprosentti on tällöin työttömien lukumäärän suhde koko työvoimaan: [18]

$$\text{Työttömyysprosentti} = \frac{\text{Työttömien lkm}}{\text{Työvoima}} \times 100\% \tag{25}$$

Nykyajan suurissa ja monimutkaisissa talouksissa tuhansine yrityksineen ja miljoonine työntekijöineen on käytännössä mahdotonta päästä työttömyydestä kokonaan eroon. Taloustieteessä on useita perusteluja tälle, mutta yleisen teorian mukaan työttömyys vaihtelee niin sanotun luonnollisen työttömyystason ympärillä. Tämä "luonnollinen työttömyys" on pitkällä aikavälillä havaittava työttömyystaso, jonka ympärillä tapahtuvaa vaihtelua kutsutaan sykliseksi työttömyydeksi. Nämä sykliset vaihtelut johtuvat esimerkiksi taloussuhdanteista sekä muusta kausittain tapahtuvasta työn tarjonnan vaihtelusta. Pitkän aikavälin työttömyydelle on useita perusteluja, mutta yleisimpiä ovat kitkatyöttömyys sekä rakenteellinen työttömyys. Kitkatyöttömyydellä tarkoitetaan sitä työttömyyttä, joka johtuu siitä, että työntekijöillä kuluu aikaa heille sopivien töiden etsimiseen. Tästä johtuvat työttömyyskaudet ovat usein lyhyitä. Rakenteellinen työttömyys puolestaan johtuu siitä, että joillakin työmarkkinoilla on liian vähän töitä tarjolla kaikille niitä haluaville. Tällä selitetään usein pitempiä työttömyyskausia. [18]

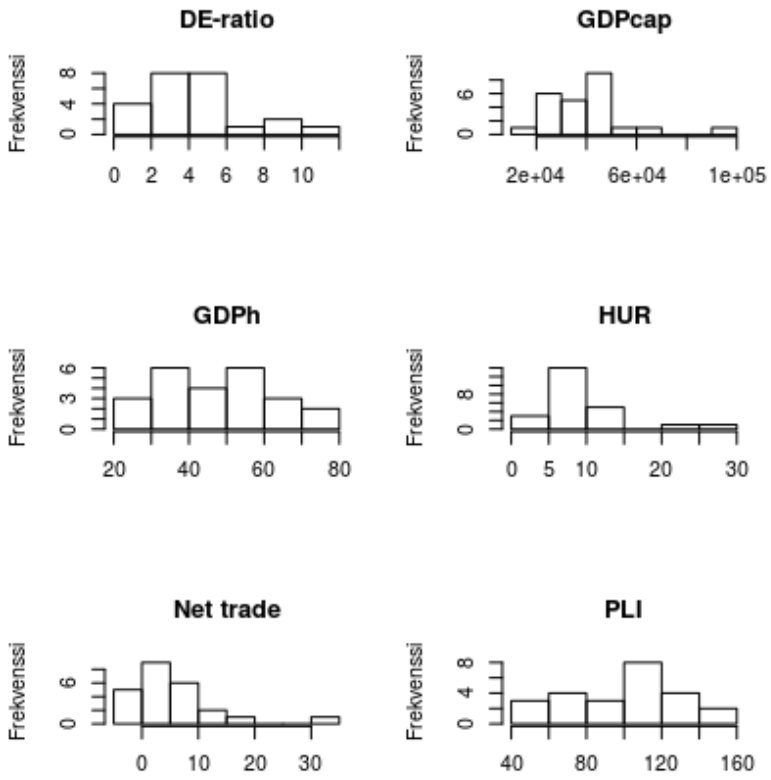
OECD:n tilastoimassa harmonisoidussa työttömyysprosentissa (HUR) työttömiksi määritellään "ne henkilöt, jotka ovat ilman töitä, ovat valmiita töihin, ja ovat tehneet

toimia töiden löytämiseksi." Tämän määritelmän mukaan tehdyt arviot johtavat kansainvälisesti paremmin vertailukelpoisiin lukuihin kuin kansallisiin määritelmiin perustuvat estimaatit. Kyseinen mittari on annettu prosenteissa kunkin valtion kokonaistyövoimasta, ja siitä on poistettu edellä mainittu kausittaisvaihtelu. [20]

3 Yksi-ja kaksiulotteinen analyysi

Ennen varsinaista klusterointia tarkastelemme, miltä kerätty aineisto näyttää yksiulotteisena sekä parettain vertailtuna. Aineiston muuttujien havainnointi selkeyttää sitä, miten työhön valitut talouisindikaattorit ovat jakautuneet Euroopan tasolla, sekä millaisia riippuvuussuhteita näissä on aineiston perusteella havaittavissa.

Aineiston yksiulotteiset histogrammit esitetään kuvassa 2, ja taulukossa 2 puolestaan esitetään joitakin tyypillisiä tilastollisia tunnuslukuja kustakin talousindikaattorista. Tarkastellaan näitä hieman tarkemmin.



Kuva 2: Yksiulotteiset histogrammit

Velkaantumisaste (D/E): Mukaan otetuissa maissa finanssialan yritysten velkaantumisaste on keskimäärin noin 4.28, ja mediaani on hieman tätä pienempi (3.92). Histogrammin perusteella suurimmmalla osalla valtioista velkaantumisaste on välillä 2-6, ja keskihajonta puolestaan on noin 2.62. On havaittavissa selkeää vaihtelua valtioiden velkaantumisasteissa: useammalla valtiolla se on myös välillä 0-2, ja toisaalta on myös toisessa ääripäässä havaittavia, suuren velkaantumisasteen valtioita. Pienimmällä havainnolla velka on jopa alle yksi-, suurimmalla taas yli yksitoistakertainen pääoman määrään verrattuna.

BKT per capita (GDPcap): Valtaosassa havainnoista kyseinen muuttuja sijoittuu

Taulukko 2: Muuttujien tilastolliset tunnusluvut

Tunnus/Muuttuja	D/E	GDPcap	GDPh	HUR	Net trade	PLI
Keskiarvo	4.28	40941.28	48.39	9.69	5.34	99.69
Mediaani	3.92	39783.16	48.67	8.23	3.65	107.00
Keskihajonta	2.62	16694.34	15.29	5.59	7.77	28.96
Maksimi	11.32	98110.11	79.28	26.55	32.39	151.00
Minimi	0.67	19610.30	27.61	3.53	-4.45	54.20

välille 20000-50000 (dollaria), ja keskiarvo onkin noin 41000. Keskihajonta tälle muuttujalle on noin 16700, eli vaihtelua on maiden välillä tässäkin muuttujassa huomattavan paljon. Minimihavainto (19610.30) ei aivan yllä alimpaan enemmän havaintoja sisältävästä 20000-30000 intervallista, mutta on kuitenkin hyvin lähellä sitä. Sen sijaan valtaosan yläpuolella havaitaan useampia yksittäisiä poikkeavia maita, joissa BKT per capita on selvästi muita korkeampi. Maksimihavainto on hyvin kaukana muista , ja yli kaksinkertainen valtioiden kokonaiskeskiarvoon nähden.

BKT työtuntia kohden (GDPh): Työn tuottavuudessa valtiot saavuttavat keskimäärin noin 48.39 dollarin kokonaistuotannon työtuntia kohden. Tuottavuudessa ei aineiston perusteella ole havaittavissa kovin selkeästi poikkeavia havaintoja, sillä jokaisessa 10 yksikön suuruisessa intervallissa on useampia havaintoja. Suurella osalla valtioista tuottavuus on välillä 30-60 dollaria/tunti, mikä käy myös hyvin yhteen keskihajonnan (15.29) kanssa. Useampia tämän välin ulkopuolellekin osuvia havaintoja on: parhaimpaan tuottavuuteen yltävällä valtiolla on tuottavuus tasolla 79.28 dollaria/tunti, kun taas heikoimman tuottavuuden valtiolla se on vain 27.61 dollaria/tunti.

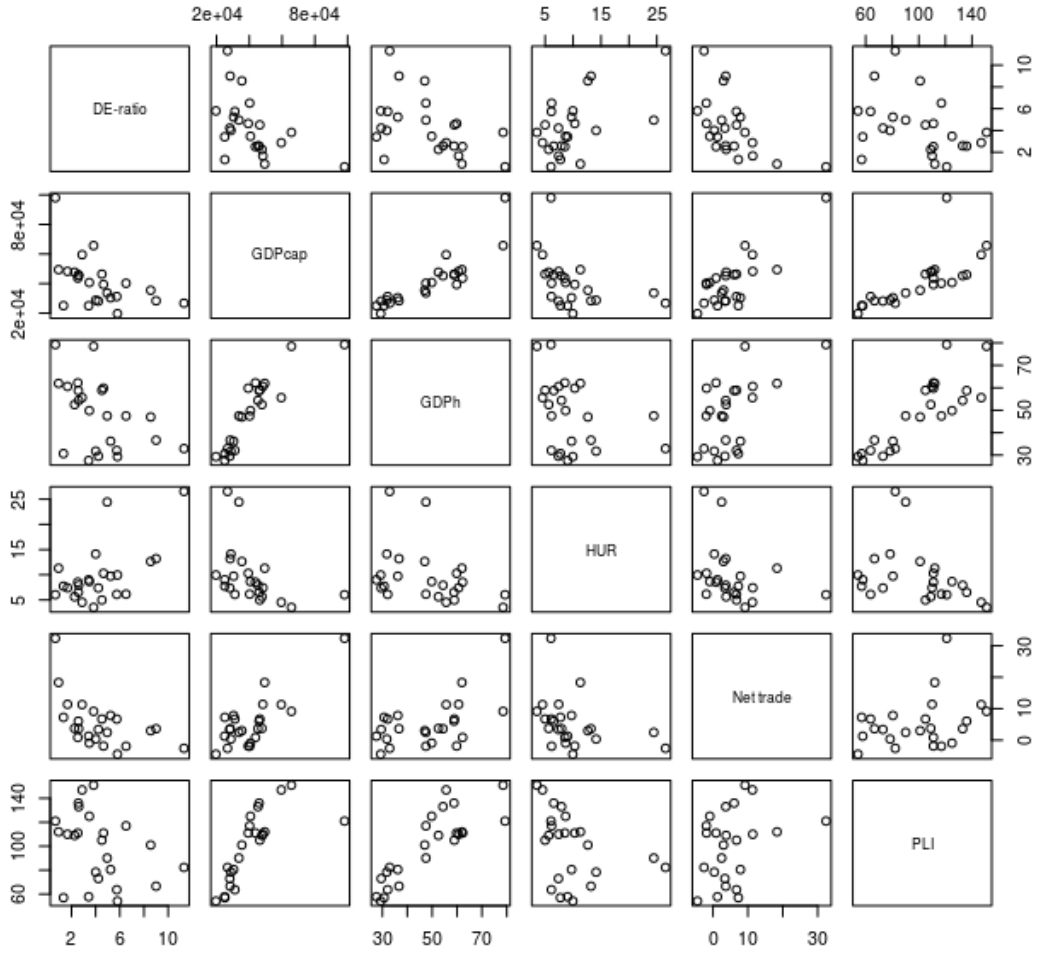
Harmonisoitu työttömyysprosentti (HUR): Keskiarvoisesti tarkastelluilla mailla työttömyysprosentti on noin 9.69. Mediaani (8.23) on keskiarvoa hieman pienempi, mikä viittaa havaintojen olevan määrällisesti enemmän painottuneita pienempiin työttömyyslukuihin. Tämä on havaittavissa myös vastaavasta histogrammista: Selkeällä enemmistöllä työttömyysprosentti on välillä 5-10%, joskin tämän läheisimpiin intervalleihin 0-5% sekä 10-15% kuuluu myös useampia valtioita, jotka ovat paljolti työttömyysprosentin keskihajonnan (5.59) rajoissa. Työttömyysluvuissa on havaittavissa myös muutamia poikkeavia, valtaosaa selvästi suurempia valtioita, joissa työttömyysprosentti on yli 20%.

Nettovienti (Net Trade): Viennissä valtaosa tarkastelluista maista on melko lähellä nollaa, jossa siis valtioiden vienti ja tuonti ovat arvoiltaan yhtäsuuret. Keskimäärin valtiot ovat vaihtosuhteeltaan noin 5.34 % kokonaistuotannon positiivisella puolella, joskin mediaani (3.65) on taas keskiarvoa pienempi, mikä viittaa valtioiden enemmistön sijoittuvan nettovienniltään lähemmäksi nollaa kuin keskiarvo antaisi ymmärtää. Tämä havaitaan myös histogrammista ,jonka perusteella selkeä enemmistö valtioista sijoittuu nettovienniltään intervalleille -5-0%, 0-5% sekä 5-10% BKT:sta. Aineiston perusteella Euroopassa on myös joitakin poikkeuksellisen vientivetoisia valtioita, joiden nettovaihto on yli 10%, ja maksimihavainnolla tämä on jopa 32.39%.

Hintatasoindeksi (PLI): Maiden keskiarvo hintatasoindeksillä on hyvin lähellä

100:aa, mikä ei ole yllättävää, sillä tarkastelun maat muodostavat valtaosan OECD-maista, joskin osa on jätetty tarkastelun ulkopuolelle. Mediaani (107.00) on hieman tätä korkeampi, eli määrällinen enemmistö tarkastelun valtioista sijoittuu kuitenkin standardiarvon 100 yläpuolelle. Hintatasoindeksissä on havaittavissa Euroopan maiden välillä merkittäviä eroja: Maksimihavainnolla hintataso on jopa 151.00, eli puolitoistakertainen OECD-maiden keskiarvoon verrattuna, kun taas minimihavainnolla (54.20) se on vain noin puolet OECD-maiden keskiarvosta. Vaihtelusta kertoo myös suurehko keskihajonta (28.96) ja se, että kaikilla histogrammin 20 yksikön pituisilla intervalleilla on useampia havaintoja.

Tarkastellaan seuraavaksi muuttujien välisiä lineaarisia riippuvuussuhteita käyttämällä kaksiulotteisia pistekaavioita sekä laskemalla muuttujien välisiä korrelaatiokertoimia. Kaksiulotteiset pistekaaviot muuttujaparien välillä esitetään kuvassa 3, ja muuttujien väliset korrelaatiot taulukossa 3.



Kuva 3: Kaksiulotteiset pistekaaviot muuttujaparien välillä

Taulukko 3: Muuttujien väliset korrelaatiot

Muuttuja	DE-ratio	GDPcap	GDPh	HUR	Net trade	PLI
DE-ratio	1.00	-0.50	-0.47	0.57	-0.54	-0.36
GDPcap	-0.50	1.00	0.88	-0.42	0.79	0.73
GDPh	-0.47	0.88	1.00	-0.36	0.57	0.85
HUR	0.57	-0.42	-0.36	1.00	-0.33	-0.36
Net trade	-0.54	0.79	0.57	-0.33	1.00	0.32
PLI	-0.36	0.73	0.85	-0.36	0.32	1.00

Kuvasta 3 voidaan tehdä seuraavat havainnot muuttujien välisiin riippuvuuksiin liittyen. Tässä yhteydessä käytetään muuttujien nimilyhenteitä selkeyden vuoksi.

D/E näyttäisi olevan jokseenkin negatiivisesti korreloitunut GDPcap:n, GDPh:n sekä Nettoviennin kanssa, eli aineiston perusteella valtion finanssiyritysten suuri velkaantumisaste on jokseenkin yhteydessä heikompaan kokonaistuotantoon, työn tuottavuuteen sekä nettovientiin. Samanaikaisesti D/E näyttäisi olevan positiivisesti korreloitunut HUR:n kanssa, eli suuremmat velkaantumisasteet voivat viitata suurempaan työttömyysprosenttiin. On tosin huomioitava, että kaikki tähän liittyvät korrelaatiot ovat itseisarvoltaan noin 0.5, eli kyse ei kuitenkaan ole selkeästä lineaarisesta riippuvuudesta.

Sen sijaan melko selkeä positiivinen lineaarinen riippuvuus on havaittavissa GDPcap:n ja GDPh:n välillä, joiden korrelaatiokerroin on noin 0.88. Tämä tulos ei ole erityisen yllättävä, sillä kummatkin suureet liittyvät pohjimmiltaan samaan kokonaistuotantoon. On selvää, että valtion BKT per capita:n suuruus, joka kertoo kokonaistuotannon arvon jakautumisesta asukasta kohti, on kiinteästi yhteydessä siihen, miten saman kokonaistuotannon arvo jakautuu käytettyä työtuntia kohden. GDPcap näyttäisi olevan myös melko vahvassa positiivisessa lineaarisessa riippuvussuhteessa Nettoviennin (0.79) ja PLI:n (0.73) kanssa. Nämäkään riippuvuudet eivät yllätä: niissä maissa, joissa elintaso (BKT per capita:lla mitattuna) on korkealla, vallitsee usein myös korkea hintataso, ja toisin päin. Nettovienti taas on yksi kokonaistuotannon osa, jolloin suurempi nettovienti kasvattaa suoraan kokonaistuotantoa, mikä näkyy kasvuna myös muissa, BKT-johdannaisissa indikaattoreissa. Toisaalta tässä muuttujana käytetään nettoviennin osuutta kokonaistuotannosta, eikä suoraa nettovientiä, mutta sama positiivinen yhteys on silti havaittavissa.

Työn tuottavuus (GDPh) ei kuitenkaan ole aivan yhtä selkeässä riippuvuudessa nettoviennin kanssa (0.57), mutta huomattavan voimakkaassa positiivisessa riippuvuudessa hintatason kanssa (0.85). Aineiston perusteella se, miten paljon työ valtiossa keskimäärin tuottaa, on melko voimakkaasti yhteydessä siihen, miten paljon hyödykkeistä ja palveluista samassa valtiossa joutuu maksamaan.

Jäljelle jääneistä riippuvuuksista harmonisoitu työttömyysprosentti (HUR) ei edellä mainittua riippuvutta lukuun ottamatta ole selkeästi lineaarisessa riippuvuussuhteessa muiden tarkasteltujen muuttujien kanssa: se on vain hieman negatiivisesti korreloitunut muiden muuttujien (GDPcap,GDPh,Net trade,PLI) kanssa. Aineiston perusteella siis työttömysprosentista ei voida juurikaan päätellä valtion kokonaistuotantoon tai hintatasoon liittyviä arvoja. Myöskään nettoviennin ja hintatasoindeksin

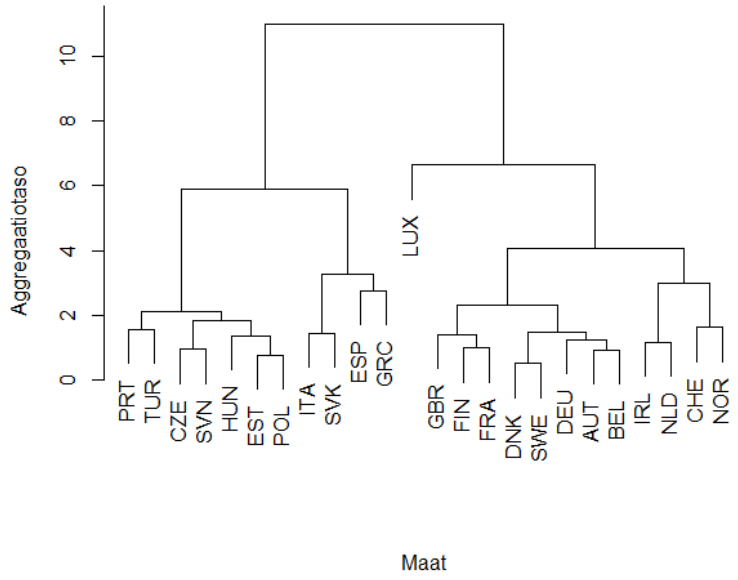
(PLI) välillä ei ole havaittavissa selkeää yhteyttä.

Seuraavaksi suoritetaan varsinainen klusterointi kappaleessa 2.1.3 kuvattua Wardin algoritmia käyttäen. Tämän jälkeen tarkastellaan klusteroinnin tuloksia sekä pohditaan klusteroinnin tuloksiin vaikuttavia syitä.

4 Klusteroinnin tulokset

4.1 Klusterointi vuoden 2014 aineistolla

Euroopan maiden klusteroinnin tulokset Wardin algoritmilla ovat selkeästi esitettävissä dendrogrammin avulla. Tässä dendrogrammissa näkyy siis koko algoritmin kulku alkupisteestä, jossa kaikki valtiot ovat itsessään klustereita, aina loppupisteen yksittäiseen klusteriin saakka. Tuloksena saatava dendrogrammi esitetään kuvassa 4:



Kuva 4: Wardin algoritmin tuottama dendrogrammi

Kuten aiemmin mainittu, ei ole olemassa yhtä oikeaa katkaisutasoa, joka erottelisi aineistosta "oikeat" ryhmät, vaan katkaisutasossa on otettava asiayhteyteen liittyvät seikat, esimerkiksi syntyvien klusterien järkevyys, huomioon. Tarkastellaan siis hieman eri katkaisutasolla syntyviä klustereita:

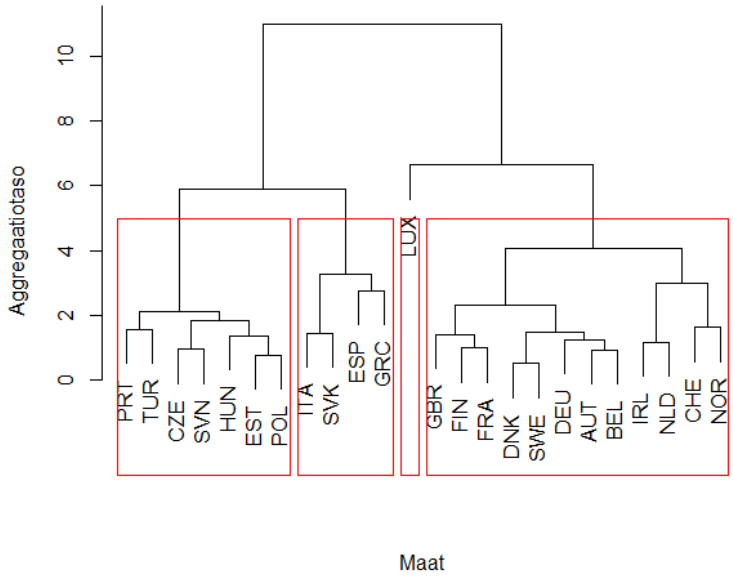
Kun katkaisutaso asetetaan siten, että jäljelle jää kolme klusteria, niin aineiston perusteella Euroopan valtioista muodostuu kaksi suurempaa, toisistaan poikkeavaa ryhmää: Portugali, Turkki, Tsekki, Slovenia, Unkari, Viro, Puola, Italia, Slovakia, Espanja sekä Kreikka muodostavat toisen näistä suurista ryhmistä. Tämä ryhmä koostuu selvästi Etelä- ja Itä-Euroopan maista. Sen sijaan toinen ryhmä koostuu

Keski-, Länsi- sekä Pohjois-Euroopan maista: Iso-Britannia, Suomi, Ranska, Tanska, Ruotsi, Saksa, Itävalta, Belgia, Irlanti, Alankomaat, Sveitsi sekä Norja.

Dendrogrammin perusteella on siis havaittavissa Euroopan maissa jonkinasteista kahtiajakautumista: toisella puolella on perinteisesti vauraina pidettyjä valtioita, ja toisella taas vähemmän vauraina pidettyjä valtioita. Tällainen jako on ainakin käytettyjen taloudellisten indikaattorien valossa myös perusteltua.

Dendrogrammista kenties silmiinpistävin havainto on se, että Luxemburg on selvästi muista maista poikkeava, joskin lähempänä muita Keski-, Länsi- ja Pohjois-Euroopan maita kuin toisessa isossa ryhmässä olevia Etelä- ja Itä-Euroopan maita. Valtio on talousindikaattoriensa perusteella jopa niin poikeava, että sen yhdistäminen toiseen suuremmita klustereista on toisiksi viimeinen koko klusterointihierarkiassa tapahtuvista klusterien yhdistämisistä.

Asetettaessa katkaisutaso siten, että lopulliseksi klusterien lukumääräksi tulee neljä, saadaan Euroopan maista seuraavat kuvassa 5 esitettävät ryhmittymät:



Kuva 5: Klusterien lkm = 4

Tässä tapauksessa Luxemburg muodostaa siis edelleen oman ryhmänsä, ja Länsi- ja Pohjois-Euroopan valtioiden muodostama klusteri on myös sama kuin kolmen klusterin tapauksessa. Sen sijaan erona aiempaan on, että Etelä-ja Itä-Euroopan erottuvat nyt aiempaa selkeämmin.

Mielenkiintoinen kysymys onkin, miten ryhmät eroavat toisistaan niiden taloudellisten indikaattorien perusteella, eli mitä ovat klusterien ryhmäkohtaiset erot. Tätä voidaan tutkia tarkastelemalla erikseen kunkin ryhmän ryhmäkeskiarvoja, jotka esitetään taulukossa 4.

Taulukko 4: Klusterien ryhmäkeskiarvot, k=4

Muuttuja	Ryhmä 1	Ryhmä 2	Ryhmä 3	Ryhmä 4
DE-ratio	3.20	4.25	8.46	0.67
GDPcap	47688.18	26869.75	31033.54	98110.11
GDPh	58.39	31.04	41.03	79.28
HUR	7.13	9.15	19.21	6.05
Net trade	5.56	3.21	1.65	32.39
PLI	122.25	66.37	85.00	121.00

Ryhmä 4: Pelkkä Luxemburg. Muuttujien arvoista huomataan välittömästi Luxemburgin poikkeavuus: Maan BKT per capita (GDPcap), työn tuottavuus (GDPh) sekä nettovienti (Net trade) ovat kaikki *huomattavasti* suurempia kuin muilla ryhmillä. Luxemburg eroaa muista valtioista erityisesti voimakkaan vientivetoisella taloudellaan, jossa nettoviennin osuus vuonna 2014 oli yli 30% valtion BKT:sta. Luxemburg on siis eräs kuvan 2 GDPcap- sekä Net trade- histogrammeissa havaittu, muista tarkastelluista valtioista selkeästi poikkeava havainto.

Luxemburgissa toimii vahva pankki- ja finanssisektori, josta kertoo esimerkiksi se, että pinta-alaltaan melko pienessä valtiossa toimi vuonna 2010 yhteensä 149 eri pankkia. Finanssisektorin vahvuuteen liittyvät Luxemburgin keskeinen sijainti Euroopassa, poliittinen vakaus, valtion rakentamat tehokkaat tietoliikenneyhteydet sekä pankkitoiminnan vankat salaisuusperiaatteet. Luxemburg myös tarjoaa ulkomaalaisille sijoittajille otollisen ympäristön esimerkiksi tarjoamalla verovapauksia maassa toimiville yrityksille. [21] Havaintoarvojen perusteella pääomasijoitusten vahvuus näkyy myös velkaantumisasteessa: Luxemburgissa toimivat finanssialan yritykset ovat muista ryhmistä poiketen jopa enemmän pääomalla kuin velkarahalla rahoitettuja yritysten velkaantumisasteen ollessa alle yksi. Hintatasoltaan Luxemburg on hyvin lähellä ryhmän 1 keskiarvoa. Molemmissa ryhmissä hintataso on selvästi korkeampi kuin keskimäärin OECD-maissa. Työttömyysluvuissa Luxemburg on myös lähellä saman ryhmän keskiarvoa. Muissa muuttujissa Luxemburg kuitenkin poikkeaa selvästi ryhmästä 1, ja vielä selkeämmin muista ryhmistä.

Luxemburgin suuri BKT per capita, työn tuottavuus, pieni velkaantumisaste sekä vahva vientivoittoisuus tekevät siitä selkeästi muista Euroopan maista poikkeavan havainnon taloudellisilla indikaattoreilla mitattuna. Perinteisesti vauraana veroparatiisina pidetty Luxemburg näyttäisi myös käytetyn aineiston valossa ansaitsevan oman luokkansa Euroopan maiden keskuudessa.

Ryhmä 1: Iso-Britannia, Suomi, Ranska, Tanska, Ruotsi, Saksa, Itävalta, Belgia, Irlanti, Alankomaat, Sveitsi sekä Norja.

Lukuun ottamatta Luxemburgin muodostamaa ryhmää, on kyseisessä ryhmässä selkeästi suurin BKT per capita, työn tuottavuus, nettovienti sekä hintataso. Samanaikaisesti ryhmän velkaantumisaste sekä työttömyysprosentti ovat jokseenkin alhaisempia kuin muilla ryhmillä, joskin nämä muuttujat eivät profiloi ryhmää aivan yhtä voimakkaasti. Ryhmäkeskiarvojen poikkeavuuden perusteella ryhmä eroaa muista ryhmistä merkittävästi, minkä takia ryhmittelyä voidaan pitää perusteltu-

na. Kyseinen ryhmä näyttäisi koostuvan valtioista, joissa on keskiarvoisesti muihin Euroopan maihin nähden BKT:lla mitattuna selvästi korkeampi elintaso, työn tuottavuus ja hintataso. Ryhmän sisällä lähimpänä toisiaan ovat pohjoismaat Tanska ja Ruotsi. Länsi-Euroopan maat Saksa, Itävalta ja Belgia ryhmittyvät myös varhaisessa vaiheessa yhteen. Tulokset kertovat siitä, että taloudellisella ja maantieteellisellä ryhmittymisellä vaikuttaisi olevan ainakin jonkilainen yhteys. Tähänkin löytyy ryhmästä toki poikkeuksia: hieman yllättävä tulos on, että Suomi on käytettyjen talousindikaattorien perusteella lähempänä Ranskaa ja Iso-Britanniaa kuin muita Pohjoismaita. Samoin esimerkiksi maantieteellisesti erillään toisistaan olevat Norja ja Sveitsi ovat taloudellisten indikaattorien perusteella hyvin lähellä toisiaan: molempia maita profiloi hyvin korkea bruttokansantuote, hintataso sekä vientivoittoisuus. Kokonaisuudessaan ryhmä koostuu kaikista analyysissa mukana olleista Länsi- ja Pohjois-Euroopan maista. Nämä maat näyttäisivätkin taloudellisten indikaattorien perusteella olevan selvästi lähempänä toisiaan kuin muita analyysissa huomioituja maita, jotka sijoittuvat myös maantieteellisesti eri alueisiin. Ryhmä muodostaa kokonaisuudessaan maantieteellisesti hyvin yhtenäisen alueen.

Ryhmä 2: Tsekki, Viro, Unkari, Puola, Slovenia, Turkki.

Tässä ryhmässä on kaikista matalin elintaso BKT:lla mitattuna, samoin myös työn tuottavuus. Samanaikaisesti myös hintataso on selkeästi matalin, keskimäärin vain noin 66% OECD- maiden keskiarvosta. Velkaantumisasteeltaan se on ryhmien keskitasoa: Keskiarvolla 4.25 ryhmän valtioiden velkaantumisaste on keskiarvoisesti suurempi kuin ryhmän 1, mutta toisaalta selvästi pienempi kuin ryhmän 3. Myös nettoviennissä ryhmä on muihin nähden keskitasoa: Vienti prosenteissa BKT:sta on muutaman prosenttiyksikön ryhmän 1 arvoa pienempi, mutta toisaalta lähes saman verran ryhmän 3 arvoa suurempi. Sama ilmiö toistuu työttömyysprosentissa, joka on hieman ryhmän 1 keskiarvoa suurempi, mutta toisaalta selkeästi ryhmän 3 keskiarvoa pienempi. Voimakkaimmin ryhmää profiloivat kuitenkin matala BKT per capita, työn tuottavuus ja hintataso.

Myös tässä ryhmässä on havaittavissa maantieteellistä yhteneväisyyttä: Ryhmään kuuluvat Slovakiaa lukuun ottamatta kaikki analyysissa mukana olevat Keski-ja Itä-Euroopan maat. Maantietellisesti lähellä toisaan olevat Tsekki, Puola, Slovenia sekä Unkari yhdistyvät klusteroinnissa melko aikaisessa vaiheessa, mikä viittaa niiden talouksien olevan jokseenkin samankaltaisia. Toisaalta ryhmässä on näiden lisäksi etelämmässä sijaitsevat Välimeren maat Portugali ja Turkki, jotka ovat myös ryhmän sisäisesti lähimpänä toisiaan. Maantieteellisesti ryhmä ei siis ole yhtä selkeästi keskittynyt kuin ryhmä 1, eli maiden maantieteellinen sijainti ei mene aivan yksi yhteen maiden "taloudellisen sijainnin" kanssa.

Ryhmä 3: Italia, Slovakia, Espanja, Kreikka

Kyseisellä ryhmä on elintason, työn tuottavuuden sekä hintatason osalta ryhmien keskikastia: Hintatasoltaan ryhmän valtiot ovat keskimäärin 85%:a OECD-maiden keskiarvosta, BKT per capita vuodessa on noin 4000 dollaria suurempi kuin ryhmässä 2, mutta noin 16000 pienempi kuin ryhmässä 1. Työn tuottavuus taas on noin 10 dollaria tunnnissa suurempi kuin ryhmässä 2, mutta toisaalta noin 17 dollaria

tunnissa pienempi kuin ryhmässä 1. Nettovienniltään ryhmä 3 on keskimäärin kaikista ryhmistä pienin, noin 1.7 prosenttia BKT:sta. Kaikista voimakkaimmin ryhmää kuitenkin profiloivat loput muuttujat: Yritysten velkaantumisaste sekä työttömyysprosentti. Ryhmän maissa työttömyysprosentti on keskimäärin lähes 20%, mikä on huomattavasti enemmän kuin missään muussa ryhmässä. Samoin yritysten velkaantumisaste näissä maissa on jopa noin 8.5, tarkoittaen sitä että kyseisten maiden taloudelliset instituutiot ovat keskimäärin melko hurjissa veloissa. Maantieteellisesti ryhmän valtiot keskittyvät Etelä-Eurooppaan: kolme neljästä klusterin sisältämästä valtiosta on Välimeren yhteydessä olevia maita. Ryhmässä poikkeavana havaintona on kuitenkin Slovakia, joka sopisi maantieteellisen sijaintinsa perusteella paremmin edelliseen ryhmään. Taloudellisten indikaattorien perusteella Slovakia on kuitenkin naapurimaidensa sijasta lähempänä Etelä-Euroopan maita.

Ryhmä 3 ja ryhmä 2 ovat klusteroinnin perusteella myös lähempänä toisiaan kuin ryhmä 1:tä. Tämä vaikuttaa järkevältä, sillä elintasoon ja työn tuottavuuteen liittyvät luvut eivät poikkea toisistaan merkittävästi ryhmien välillä. Huomattavat erot tulevat näiden sijaan valtioiden työttömyysluvuissa sekä yritysten velkaantumisasteissa, jotka ovat molemmat ryhmässä 3 keskimäärin selvästi korkeammalla tasolla kuin ryhmässä 2.

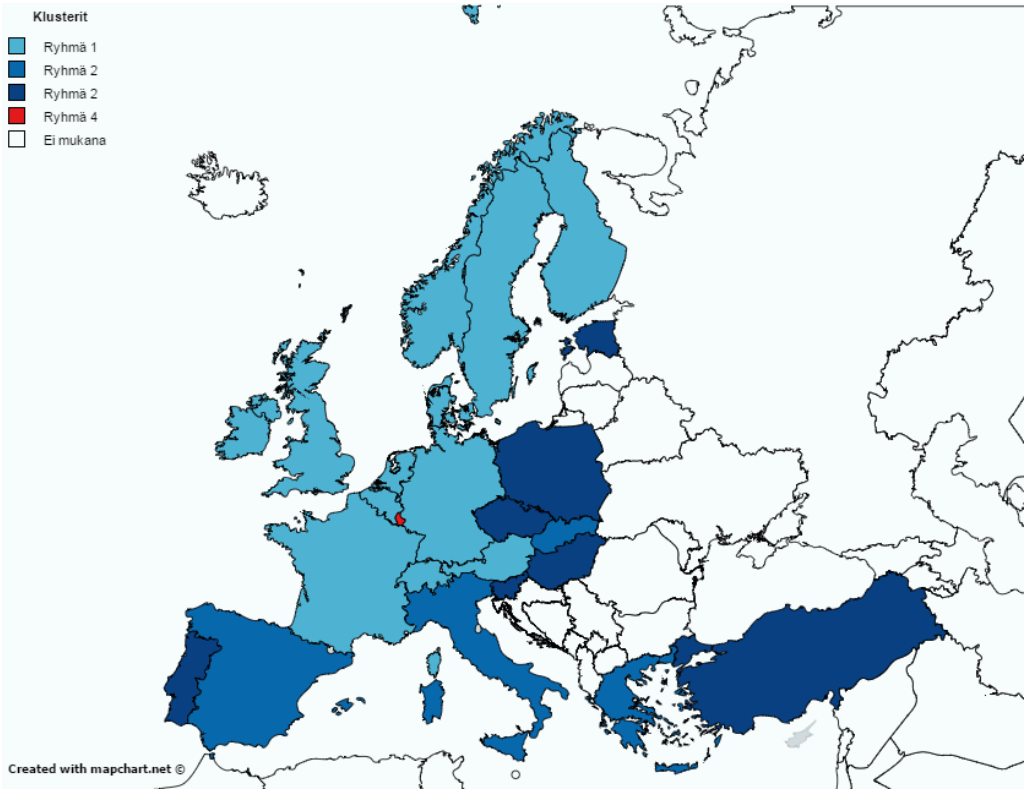
Kuvassa 6 esitetään klusteroinnin tulokset Euroopan kartalla. Tämä havainnollistaa klusterien sijoittumista Euroopan eri osiin: Kuten mainittu, ryhmä 1 muodostaa jokseenkin selkeän maantieteellisen alueen, joka kattaa Länsi- ja Pohjois- Euroopan. Ryhmä 4 eli Luxemburg on taloudellisilta indikaattoreilta niin poikkeava, että se muodostaa kokonaan oman ryhmänsä. Toisaalta, kuten kuvan 4 dendrogrammista havaitaan, on Luxemburg lähempänä ryhmän 1 valtioita, joten maantietellinen sijainti pysyisi yhtenäisenä myös tämä nämä klusterit yhdistettäessä. Myös ryhmät 2 ja 3 ovat muutamia poikkeusvaltioita (Portugali, Slovakia) lukuun ottamatta maantieteellisesti melko yhtenäisiä.

4.2 Klusterointi vuoden 2006 aineistolla

Suoritetaan seuraavaksi sama maiden klusterointi käyttämällä täysin samoja muuttujia, mutta vuoden 2006 aineistoa. Tällöin voidaan tarkastella, onko Euroopan valtioiden ryhmittymisessä tapahtunut muutoksia viimeisten vuosien aikana. Kaikki tarvittavat tiedot on saatavilla samasta OECD:n tietokannasta, muutamaa poikkeusta lukuun ottamatta:

- Sveitsin harmonisoitu työttömyysprosentti. Vuodelta 2006 ei ole saatavissa OECD:n tietokannasta aineistoa HUR:n osalta. World Bank Group- nimisen talousinstituution tietokannasta [22] kuitenkin löytyy Sveitsin työttömyysprosentille estimaatit, jotka vaikuttaisivat myös olevan hyvin yhteneviä OECD:n vuodesta 2010 eteenpäin löytyviin lukuihin. Puuttuvan tiedon välttämiseksi Sveitsin osalta käytetään analyysissa tätä estimaattia, jonka uskotaan olevan riittävällä tarkkuudella oikeansuuruinen.

- Turkin finanssiyritysten velkaantumisaste (D/E). Varhaisin tieto OECD:n tie-



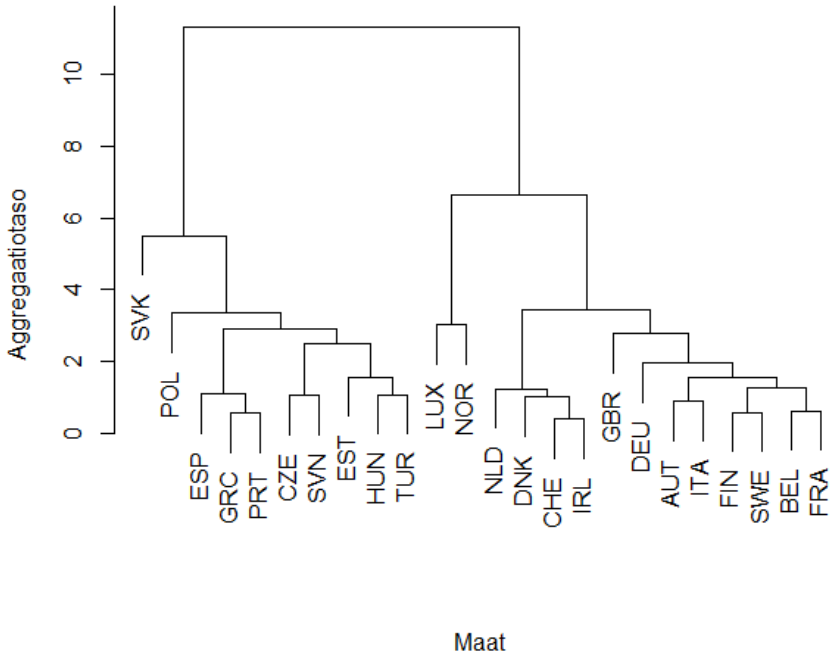
Kuva 6: Klusterit kartalla

tokannassa on vuodelta 2009, joka on epäideaalinen analyysin kannalta, mutta se on kuitenkin ajallisesti lähempänä vuotta 2006 kuin vuotta 2014. Puuttuvan tiedon välttämiseksi käytämme tätä tietoa siinä uskossa, että kyseinen tieto on lähempänä todellista, ei-havainnoitua vuoden 2006 arvoa.

Suoritettaessa vuoden 2006 aineistolla täysin identtinen klusterianalyysi kuin vuoden 2014 aineistolla tehtiin edellä, saadaan tuloksena seuraava kuvassa 7 esitettävä dendrogrammi.

Verrattaessa tulosta kuvassa 4 olevaan vuoden 2014 dendrogrammiin, havaitaan niissä niin samankaltaisuutta kuin eroavaisuuksiakin. Tarkastelleen jälleen syntyviä klustereita joillakin eri katkaisutasoilla.

Valittaessa katkaisutaso siten, että klustereita jää jäljelle kaksi, on vuoden 2006 aineistossa havaittavissa samankaltainen kahtiajakautuminen kuin vuoden 2014 tapauksessa. Kun tarkastellaan näin syntyviä kahta erillään toisistaan olevaa suurklusteria tarkemmin, havaitaan, että ne muodostuvat yhtä poikkeusta lukuun ottamatta täysin samoista valtioista kuin vuoden 2014 aineistoa käytettäessä. Ensimmäisessä suurklusterissa ovat jälleen tarkastelun sisältämät Etelä-ja Itä-Euroopan maat, toisessa taas vauraina pidetyt Länsi- ja Pohjois-Euroopan maat. Yksi poikkeus kuitenkin on havaittavissa: Vuoden 2006 tapauksessa Italia on yllättäen eri klusteriin kuin vuonna 2014. Se on siis vuonna 2006 ollut aiemmin lähempänä toista, vauraiden valtioiden suurklusteria, kun uusimman aineiston perusteella se kuuluisi pikemmin-



Kuva 7: Dendrogrammi vuoden 2006 aineistolla

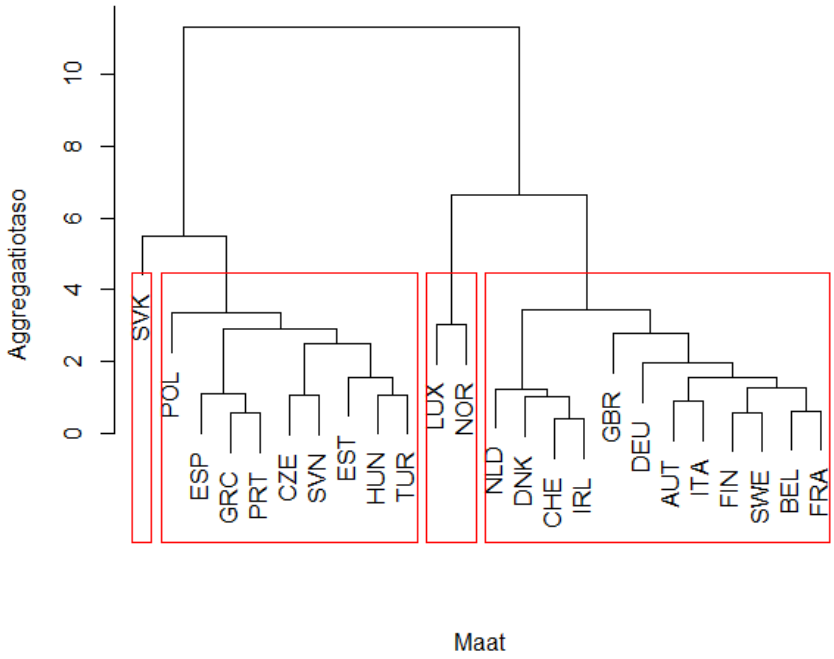
kin muiden Etelä-ja Itä-Euroopan muodostamaan suurklusteriin. Tulos voi viitata siihen, että Italian taloustilanne on tarkasteluvälillä huonontunut suhteessa samaan klusteriin vuonna 2006 kuuluneisiin valtioihin, jonka takia sen osalta on tapahtunut siirtymä suurklusterista toiseen.

Valtioiden lähimmissä pareissa on toki havaittavissa eroavaisuutta: Esimerkiksi Suomi on vuoden 2006 aineistossa lähimpänä Ruotsia, toisin kuin vuoden 2014 tapauksessa, jossa Suomea taloudellisesti lähimpänä oli Ranska. Korkean aggregaatiotason (k=2) klustereissa ei kuitenkaan ole havaittavissa dramaattisia eroavaisuuksia eri vuosien aineistoa tarkastellessa. Klusterointi ei näyttäisi kuitenkaan olevan täysin stabiili ajan suhteen, vaan yksittäisiin valtioihin liittyviä muutoksia on ajassa tapahtunut. Tämä havaitaan kahden klusterin tapauksessa Italian siirtymisenä, mutta vieläkin selkeämmin suuremman klusterimäärän tapauksessa.

Kun taas katkaisutaso valitaan siten, että klustereiden lukumäärä on neljä, eivät tulokset ole aivan samanlaisia kuin vuoden 2014 aineistolla, kuten kuvasta 8 käy hyvin ilmi. Tarkastellaan seuraavaksi ryhmittymisessä tapahtuneita muutoksia. Klusterointiin liittyvät ryhmäkeskiarvot esitetään taulukossa 5.

Ryhmä 4: Slovakia.

Vuoden 2006 aineistossa Slovakialla on muiden ryhmien keskiarvoihin nähden moninkertaisesti suurempi finanssiyritysten velkaantumisaste, jopa 13.6. Slovakialla on myös selvästi muihin ryhmiin nähden suurempi harmonisoitu työttömyysprosentti,



Kuva 8: Klusterien lkm = 4, vuoden 2006 aineistolla

Taulukko 5: Ryhmäkeskiarvot,k = 4, vuoden 2006 aineistolla

	1	2	3	4
DE-ratio	3.23	3.85	3.15	13.6
GDPcap	37174.57	22129.19	65971.95	18748.37
GDPh	53.55	30.25	80.61	30.04
HUR	6.41	8.4	4	13.47
Net trade	4.1	-4.45	23.65	-3.99
PLI	117	71.78	125.5	56

noin 13.47. Myös hintatasoltaan, joka on vain 56% OECD-keskiarvosta, on valtio muiden ryhmien keksiarvoja pienempi. Nettovienniltä, työn tuottavuudelta ja BKT per capitalta ryhmä on hyvin lähellä ryhmää 2. Slovakia erottuu siis vuoden 2006 aineistossa erityisesti työttömyysprosentin ja velkaantumisasteensa perusteella.

Ryhmä 3: Norja ja Luxemburg.

Tällä ryhmällä on huomattavasti suurempi BKT per capita, työn tuottavuus sekä nettovienti kuin muilla ryhmillä, sekä hieman pienempi työttömyysprosentti. Velkaantumisasteelta ja hintatasolta ryhmä on hyvin lähellä ryhmää 1.

Luxemburg oli vuoden 2014 aineistolla selkeästi poikkeava havainto, mutta vuoden 2006 aineistolla se muodostaakin klusterin yhdessä Norjan kanssa. Vuoden

2014 aineistossa Norja oli sen sijaan lähempänä muista vauraista maista koostuvien maiden klusteria. Vaikuttaakin siltä, että Norjan asema on vuonna 2006 ollut erilainen muihin Länsi- ja Pohjois-Euroopan maihin nähden vuoteen 2014 verrattuna.

Ryhmä 2: Puola, Espanja, Kreikka, Portugali, Tsekki, Slovenia, Viro, Unkari, Turkki.

Kyseinen ryhmä on hyvin samanlainen vuoden 2014 aineistosta saatuun ryhmään, joskin Espanja ja Kreikka olivat tällöin erillisessä ryhmässä. Ryhmä erottuu sen matalalla elintasolla, työn tuottavuudella sekä suhtellisen matalalla hintatasolla ja korkealla työttömyysprosentilla. Toisin kuin vuoden 2014 aineistolla, on ryhmän nettovienti ollut vuonna 2006 hieman negatiivinen.

Ryhmä 1 Alankomaat, Tanska, Sveitsi, Irlanti, Iso-Britannia, Saksa, Itävalta, *Italia*, Suomi, Ruotsi, Belgia, Ranska.

Tämä ryhmä on Italiaa ja Norjaa lukuun ottamatta sama kuin vuoden 2014 aineistosta saatu ryhmä. Ryhmä profiloituu myös samalla tavalla kuin vuoden 2014 tapauksessa: Sillä on tällä kertaa Norjasta ja Luxemburgista muodostuvaa poikkeusryhmää lukuun ottamatta suurin BKT per capita, työn tuottavuus sekä hintataso. Työttömyysprosentiltaan ja velkaantumisasteeltaan se on ryhmien 2 ja 3 välimaastossa, ja nettovienniltään samaa tasoa kuin vuonna 2014.

Vuoden 2014 aineistossa havaittiin Espanjan, Kreikan, Italian sekä Slovakian muodostama ryhmä. Kyseistä ryhmää profiloi keskitasoinen BKT per capita ja työn tuottavuus, sekä korkea finanssiyritysten velkaantumisaste ja työttömyysprosentti. Vuoden 2006 aineistossa tätä ryhmää ei kuitenkaan ole ollenkaan havaittavissa. Espanja ja Kreikka ovat kyllä edelleen indikaattoriensa perusteella melko lähellä toisiaan, mutta kaksi ryhmän muuta jäsentä ovat näistä hyvinkin kaukana. Aiemmin Espanjan ja Kreikan kanssa samaan klusteriin ryhmittynyt Slovakia muodostaa vuoden 2006 aineistossa kokonaan oman ryhmänsä. Voitanee kysyä, onko selkeitä muutoksia tapahtunut Slovakian osalta, vai liittyvätkö muutokset pikemminkin muihin samaan klusteriin liityneisiin valtioihin. Samoja kysymyksiä voidaan asettaa myös aiemmin Kreikan ja Espanjan kanssa ryhmittyneelle Italialle. Tarkastellaan kyseisten valtioiden talousindikaattoreita erikseen molemmmilta vuosilta. Nämä esitetään taulukoissa 6 ja 7.

Taulukko 6: SVK,ESP,GRC ja ITA vuoden 2006 aineistossa

Valtio	DE-ratio	GDPcap	GDPh	HUR	Net trade	PLI
SVK	13.60	18748.37	30.04	13.47	-3.99	56
ESP	3.76	30886.32	42.14	8.46	-5.92	92
GRC	2.96	28272.63	34.40	9.03	-10.50	88
ITA	2.88	31797.93	47.11	6.79	-0.84	105

Mitä kyseisille valtioille on tapahtunut vuoden 2006 jälkeen? Slovakian tapauksessa BKT per capita on noussut huomattavan paljon kuluneiden vuosien aikana, ja se onkin saavuttanut huomattavasti taulukon maita tarkasteluvälillä. Samalla sen

Taulukko 7: SVK,ESP,GRC ja ITA vuoden 2014 aineistossa

Valtio	DE-ratio	GDPcap	GDPh	HUR	Net trade	PLI
SVK	9.00	28326.97	36.68	13.21	3.65	66.6
ESP	4.95	33637.59	47.50	24.45	2.49	90.1
GRC	11.32	26710.42	32.93	26.55	-2.56	82.3
ITA	8.56	35459.18	47.01	12.65	3.03	101.0

hintataso ja työn tuottavuus ovat nousseet, joskaan eivät aivan yhtä merkittävästi. Finanssiyritysten velkaantumisaste on pienentynyt arvoon 9, joka sekin on edelleen melko suuri arvo. Työttömyysprosentissa ei ole tapahtunut merkittäviä muutoksia. Muiden valtioiden indikaattoreita tarkastelemalla selviää, miksi ne ovat ryhmittyneet yhteen vuonna 2014, kun vuonna 2006 ne olivat vielä erillään. Sekä Italiassa, Espanjassa että Kreikassa vuoteen 2006 nähden on työttömyys noussut erittäin suuriin lukemiin. Myös finanssiyritysten velkaantumisasteet ovat nousseet selkeästi aiempaa korkeammalle, erityisesti Kreikan ja Italian tapauksessa. Kyse ei ole siis niinkään Slovakiassa tilanteessa tapahtuneeista muutoksista, vaan enemmänkin Espanjan, Kreikan ja Italian yritysten velkaantumisesta ja työttömyyden rajusta kasvusta.

Kyseisille havainnoille löytyy mahdollinen selitys Euroalueen finanssikriisistä. Kreikka, Espanja sekä Italia kaikki kuuluvat nimittäin Euroopan 2010-luvun kriisin pahimpiin kärsijöihin. Espanja ja Kreikka ovat kaksi yhteensä viidestä kriisin seurauksena valtionvararikon tehneestä valtiosta.

Kreikka voidaan mieltää Euroalueen finanssikriisiin käynnistäneenä valtiona: Vuonna 2009 selvisi, että Kreikka oli useamman vuoden ajan vääristellyt taloustilastojaan. Tällöin paljastui, että Kreikan valtion velka ja budjettialijäämä olivat selvästi alakanttiin arvioituja. Tämä puolestaan horjutti sijoittajien luottamusta Kreikan talouteen, minkä takia Kreikan oli maksettava yhtä korkeampaa korkoa lainoistaan. Tästä käynnistynyt noidankehä pahensi Kreikan tilannetta entisestään velkakustannusten noustessa, ja lopulta Kreikalla ei ollut enää mahdollisuutta maksaa valtavaksi paisunutta velkaansa. [23] Kriisin seuraukset näkyvät vuoden 2014 aineistossa edelleen Kreikan taloudessa: työttömyys on hyvin korkeissa lukemissa, ja finanssiyritykset korviaan myöten veloissa.

Espanjassa kriisin aiheuttajana ei sen sijaan ollut valtion velkaisuus, vaan pääsyynä oli asuntojen hintakupla ja siihen liittyvä kestämätön BKT:n nousu. Rakennussektorin sekä kiinteistöinvestointien ajama verotulojen kasvu riittivät pitämään Espanjan budjetin ylijäämäisenä kulutuksen lisääntymisestä huolimatta. Lopulta tämä hintakupla kuitenkin puhkesi, jonka jälkeen asuntojen hinnat lähtivät laskuun ja rakennusalan tuotanto romahti. Kriisin seurauksena Espanjan talous kääntyi voimakkaaseen laskuun, ja sitä kautta puolestaan työttömyys jyrkkään kasvuun. [24]

Myös Italia on ollut kiintessä yhteydessä Euroalueen finanssikriisiin: Kreikan tavoin suuresti velkaantuneena valtiona, mutta taloudeltaan selkeästi suurempana myös Italian vakavaraisuus on herättänyt sijoittajien keskuudessa pelkoa.[25] Kriisin seuraukset näyttäisivät edelleen vaikuttavan Italian talouteen: työttömyysprosentti ja finanssiyritysten velkaantumisaste ovat molemmat selvästi kohonneella tasolla.

Vuoden 2014 aineistosta tehty klusterointi näyttäisikin erottelevan jokseenkin omaksi ryhmäkseen Euroalueen finanssikriisistä pahiten kärsineitä maita. Vuonna 2006 finanssikriisi ei ollut vielä alkanut, minkä takia siitä kärsineet maat (Kreikka, Espanja ja Italia) eivät erottuneet omaksi ryhmäkseen. Toisaalta on mielenkiintoista huomata, että muut finanssikriisistä paljon kärsineet valtiot (Irlanti, Portugali) eivät juurikaan erotu ryhmittelyltään vuosien 2014 ja 2006 tapauksissa.

Tarkastellaan vielä Norjan ja Luxemburgin havaintoja vuosilta 2006 ja 2014. Nämä esitetään taulukoissa 8 ja 9.

Taulukko 8: NOR,LUX vuoden 2006 aineistolla

Valtio	DE-ratio	GDPcap	GDPh	HUR	Net trade	PLI
LUX	0.63	77257.99	80.22	4.58	30.36	115
NOR	5.67	54685.91	81.00	3.43	16.94	136

Taulukko 9: NOR,LUX vuoden 2014 aineistolla

Valtio	DE-ratio	GDPcap	GDPh	HUR	Net trade	PLI
LUX	0.67	98110.11	79.28	6.05	32.39	121
NOR	3.83	65705.17	78.51	3.53	9.18	151

Huomataan, että Luxemburgin BKT per capita on selvästi noussut ja työttömyysprosentti jonkin verran noussut vuodesta 2006, mutta muilta osin muuttujien arvot ovat samantasoisia. Norjan tapauksessa taas velkaantumisaste on hieman laskenut, eli maiden välinen ero on tässä muuttujassa jopa hieman kaventunut. Työttömyysprosentti ja työn tuottavuus ovat pysyneet jokseenkin samoina, kun taas BKT per capita ja hintataso ovat selvästi nousseet, sekä nettovienti vähentynyt. Vuoden 2014 aineistossa Norjan erot Luxemburgiin nähden ovat kasvaneet vuoteen 2006 verrattuna: molemmilla mailla on BKT per capita kasvanut selkeästi, mutta Luxemburgilla se on kasvanut enemmän. Valtioiden hintatasojen ero on myös kasvanut Norjan hintatason kasvun selkeästi ylittäessä Luxemburgin vastaava. Myös työttömyysprosenttien välinen ero on kasvanut, kun Luxemburgin työttömyysprosentti on hieman kasvanut Norjan vastaavan pysyessä lähes ennallaan. Myös valtioiden nettoviennin osuudessa vuosien välinen ero on kasvanut.

Kumpikaan kyseisistä valtioista ei ole käytettyjen talousindikaattorien perusteella kärsinyt samoin tavoin Euroopan talouskriisistä kuin edellä mainitut Etelä-Euroopan valtiot. Vuosien 2006 ja 2014 aineistossa eri tavoin syntyvä ryhmittely ei siis synny näiden valtioiden tapauksessa yhtä selkeästi tietyn reaalitalouden ilmiön seurauksena, vaan on pikemminkin yhdistelmä pienemmistä, yksittäisissä indikaattoreissa tapahtuneista maiden välisten erojen kasvusta, joiden summana Norjan ryhmittely muuttuu eri vuosien aineistoa tarkastellessa.

5 Yhteenveto

Klusterianalyysin perusteella oli mahdollista ryhmitellä Euroopan maita taloudellisesti koko Eurooppaa yhtenäisempiin ryhmiin. Näissä ryhmittelyn havaittiin olevan jokseenkin yhteydessä maiden maantietelliseen sijaintiin Euroopan sisällä: Tarkastellut Länsi- ja Pohjois-Euroopan maat ovat molempien tarkasteltujen vuosien perusteella lähempänä toisiaan, samoin kuin tarkastellut Itä- ja Keski-Euroopan maat. Vuoden 2014 aineistosta pystyttiin klusteroinnin avulla myös erottelemaan omaksi ryhmäkseen korkeasta työttömyydestä ja finanssiyritysten velkaantumisasteesta kärsivien maiden ryhmä, joista valtaosa on ollut kärsivänä osallisena Euroopan 2010-lukua hallinneessa talouskriisissä.

Vaikka valtaosa maista ryhmittyi vuosien 2006 ja 2014 tarkastelussa samalla tavalla, ei maiden välinen ryhmittely tutkimuksen perusteella ole kuitenkaan aivan staattinen, vaan yksittäisten maiden ryhmittely voi muuttua. Tämä taas voi aiheutua joko radikaalien reaalitalouden muutoksien seurauksena, kuten Kreikan, Italian ja Espanjan tapauksessa, tai monien pienempien muutosvaikutusten summana, kuten Norjan tapauksessa.

Analyysin jatkokehittelyä ajatellen siihen mukaan otettavat muuttujat voisivat olla sellaisia, että niillä katetaan talouden kenttää hieman kokonaisvaltaisemmin. Tässä työssä valtaosa muuttujista on tavalla tai toisella kytköksissä BKT:een, mikä on havaittavissa esimerkiksi tarkasteltaessa aineiston korrelaatiorakennetta. Kenties BKT:n voisi jättää kokonaan analyysista pois, ja ottaa sen tilalle jotain muuta, tai vaihtoehtoisesti miettiä muiden muuttujien tilalle BKT:sta vähemmän riippuvaisia muuttujia. Toisaalta BKT:n luonne on hyvin yleismaailmallinen, joten tämä voi osoittautua jokseenkin haasteelliseksi.

Tämän lisäksi voisi olla mielenkiintoista toistaa samankaltainen analyysi aineistolla, joka kattaa pelkän taloudellisen hyvinvoinnin lisäksi myös laajemmin valtion ja sen asukkaiden hyvinvointia kuvaavia muuttujia. Tällaisia voisivat olla esimerkiksi terveyteen, ympäristöön tai onnellisuuteen liittyvät mittarit. Näitä käyttämällä saatasiin kenties kokonaisvaltaisempi kuva siitä, millaisia ryhmittymiä valtiot muodostavat.

Olisi mielenkiintoista myös toistaa analyysi siten, että pyrkisi ottamaan aikaulottuvuuden vielä tarkemmin huomioon. Aineistoa on saatavilla usealta vuodelta vähintään vuositasolla, ja näistä eri muuttujien muodostamista aikasarjoista voisi havaita ajan suhteen tapahtuneita muutoksia valtioiden välisessä ryhmittymisessä, ja sitä kautta etsiä näille muutoksille mahdollisia tulkintoja esimerkiksi reaalimaailman tapahtumista ja ilmiöistä.

Analyysin voisi pyrkiä laajentamaan myös Euroopan ulkopuolelle. Esimerkiksi voisi olla mielenkiintoista tutkia Aasian maita, joissa talous lienee kehittyvän jokseenkin eri tavalla kuin Euroopassa.

Viitteet

[1] Jones, B. *Why is unity so important to Europe?* CNN, Verkkolehti, Päivitetty 6.9.2012, Viitattu 16.4.2016. Saatavissa: http://edition.cnn.com/2011/11/04/world/europe/european-unity-explainer/

[2] *European Union.* Verkkodokumentti. Encyclopedia of Management, 2009. Viitattu 16.4.2016. Saatavissa: http://www.encyclopedia.com/doc/1G2-3273100096.html

[3] Simar, L. *An Introduction to Multivariate Data Analysis.*Université Catholique de Louvain Press, 2008, Belgia, Kappale 6.

[4] Norušis, M. *IBM SPSS Statistics 19 Statistical Procedures Companion*. Prentice Hall, 2012, New Jersey, United States. Kappale 17.

[5] Everitt, B. *Cluster Analysis.* 5. painos. Wiley, 2011, Chichester, West Sussex, UK.

[6] MacKay, D. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2003. Kappale 20. ISBN 0-521-64298-1. MR 2012999.

[7] Parkin, M. *Economics: European edition.* 9. painos. Pearson, 2010.

[8] OECD (2016), Gross domestic product (GDP) (indicator). doi: 10.1787/dc2f7aec-en. Haettu 1.3.2016.

[9] *How Do We Measure Standard of Living?*. Verkkodokumentti. The Federal Bank of Boston. Saatavissa: https://www.bostonfed.org/education/ledger/ledger03/winter/measure.pdf

[10] Schreyer, P. , Koechlin, F. *Purchasing power parities- measurement and uses.* OECD Statistics Brief, 2002, No. 3, sivu 7.

[11] OECD (2016), Price level indices (indicator). doi: 10.1787/c0266784-en. Haettu 1.3.2016.

[12] OECD Manual: *Measuring Productivity; Measurement of Aggregate and Industry-Level Productivity Growth*, OECD, 2001.

[13] OECD (2016), GDP per hour worked (indicator). doi: 10.1787/1439e590-en. Haettu 1.3.2016.

[14] Dey-Chowdhury, S., Goodridge, P., Wallis, G. *The ONS Productivity Handbook: A Statistical Overview and Guide* . Office for National Statistics, Palgrave Macmillan, 2007, Kappale 5.

[15] Brealey, A. Myers, S. Allen, F. *Principles of Corporate Finance*, 11.painos. McGraw-Hill Irwin, 2014, New York.

[16] OECD (2016), Financial corporations debt to equity ratio (indicator). doi: 10.1787/a3108a99-en. Haettu 1.3.2016.

[17] Forte, S. *Capital structure: Optimal leverage and maturity choice in a dynamic model.* Universidad Carlos III de Madrid, February 2004

[18] Mankiw, G. *Brief Principles of Macroeconomics.* 6. painos. South-Western Cengage Learning, United States, 2012.

[19] OECD (2016), Trade in goods and services (indicator). doi: 10.1787/0fe445d9-en. Haettu 1.3.2016.

[20] OECD (2016), Harmonised unemployment rate (HUR) (indicator). doi: 10.1787/52570002-en. Haettu 1.3.2016.

[21] Luxembourg (09/24/10) (arkistoitu versio) U.S. Bilateral Relations Background Notes. US Department of State. United States, 2010.

[22] The World Bank Database (2016), Unemployment, total (% of total labor force) (indicator). The World Bank Group. Saatavissa: http://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?

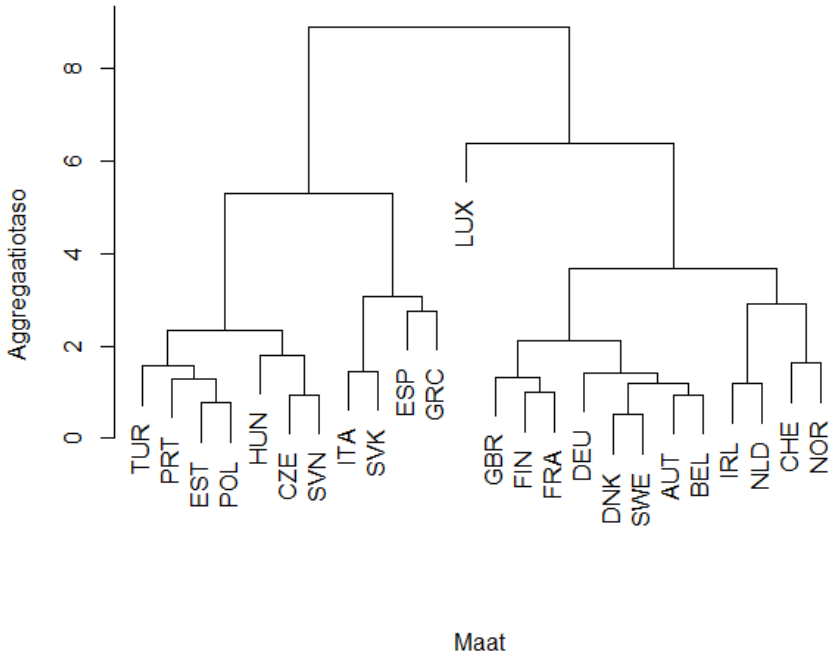
[23] Higgins, M. Klitgaard, T. *Saving Imbalances and the Euro Area Sovereign Debt Crisis.* Current Issues in Economics and Finance 17, Federal Reserve Bank of New York, United States, 2011. Saatavissa: https://www.newyorkfed.org/medialibrary/media/research/current_issues/ci17-5.pdf

[24] Knight, L. *Spanish economy: What is to blame for its problems?*. BBC News, Verkkolehti, Päivitetty 18.5.2012, Viitattu 31.3.2016. Saatavissa: http://www.bbc.com/news/business-17753891

[25] Weissman, J. *4 Reasons Why Italy's Economy Is Such a Disaster.* The Atlantic, Verkkolehti, Päivitetty 10.11.2011, Viitattu 31.3.2016. Saatavissa: https://web.archive.org/web/20120601044644/http://www.theatlantic.com/business/archive/2011/11/4-reasons-why-italys-economy-is-such-a-disaster/248238/

A Muut etäisyysmitat

Tässä liitteessä esitetään muihin kuin työssä pääasiallisesti käytettyyn etäisyysmittaan liittyvät dendrogrammit valtioiden klusterianalyysista.

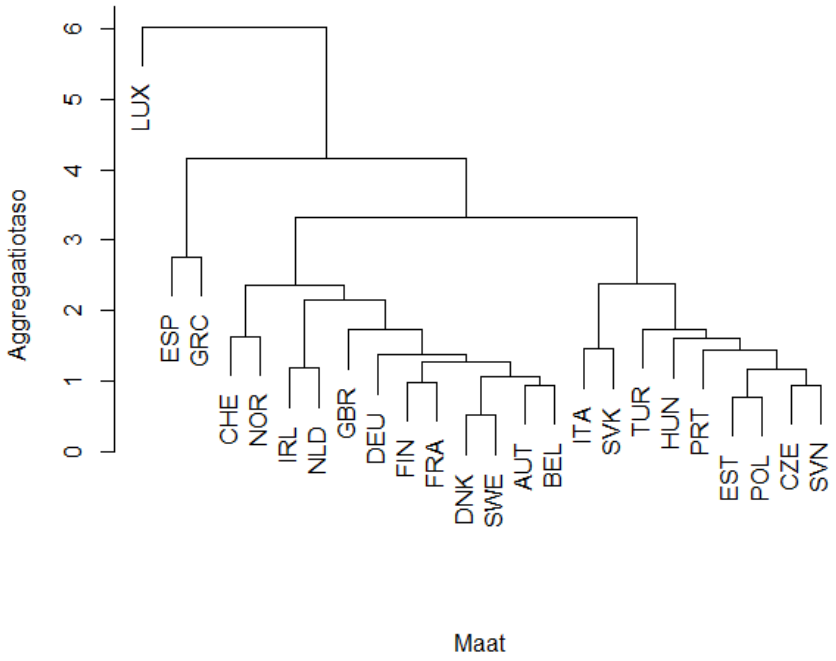


Kuva A1: Dendrogrammi maksimietäisyydellä (CL)

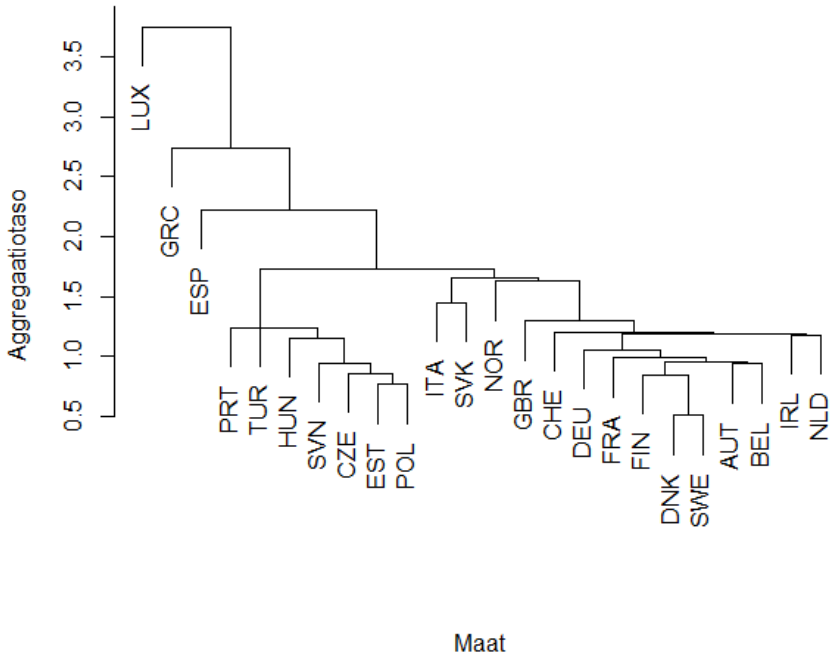
Maksimietäisyyksiä käytettäessä (kuva A1) syntyvä klusterihierarkia on lähes identtinen kuin Wardin kriteeriä käytettäessä. Vaikuttaa siltä, että käytetyn aineiston tapauksessa klusterien pisteiden maksimietäisyys ja ryhmien välisen vaihtelun vähenemisen minimointi tuottavat hyvin samankaltaisia tuloksia.

Myös keskiarvoetäisyyksiä käytettäessä (kuva A2) ovat klusterit hyvin samankaltaisia, joskin tässä tapauksessa Espanja ja Kreikka erottuu selkeästi omaksi ryhmäkseen. Muut ryhmät ovat samankaltaisia kuin Wardin tapauksessa: Itä-ja Etelä-Europaan valtiot sekä Länsi-,Keski- ja Pohjois-Euroopan valtiot omissa ryhmissään. Keskiarvoetäisyyksien tapauksessa nämä suuret ryhmät ovat jopa lähempänä toisiaan kuin Espanjan ja Kreikan muodostama ryhmä.

Jopa minimetäisyyksiä käytettäessä (kuva A3) on ryhmissä samankaltaisuutta, mutta tässä Luxemburgin lisäksi myös Kreikka ja Espanja ovat muista poikkeavia havaintoja, jotka ryhmitellään viimeisinä yksittäisinä havaintoina kaikista muista valtioista koostuvaan klusteriin. Italia ja Slovakia näyttävät minimietäisyyden perusteella olevan lähempänä vauraita valtioita, mikä puolestaan hankaloittaa tulosten tulkintaa.



Kuva A2: Dendrogrammi keskiarvoetäisyydellä (AL)



Kuva A3: Dendrogrammi minimietäisyydellä (SL)