Tampereen yliopisto

Joonas Kari

KLUSTEROINNIN HYÖDYNTÄMINEN JAKELUVERKONHALTIJOIDEN VALVONTATIETOJEN ANALYSOIMISESSA

Informaatioteknologian ja viestinnän tiedekunta
Diplomityö
maaliskuu 2019

TIIVISTELMÄ

Joonas Kari: Klusteroinnin hyödyntäminen jakeluverkonhaltijoiden valvontatietojen analysoimisessa
Diplomityö
Tampereen yliopisto
Sähkötekniikka, DI
Maaliskuu 2019

Diplomityön taustalla on Energiaviraston tavoite pyrkiä löytämään yhteyksiä sähkönjakeluverkonhaltijoiden valvonnan yhteydessä toimitettujen tunnuslukujen ja yhtiöiden taloudellisen menestyksen väliltä. Tämän työn tavoitteena on tutkia klusteroinnin hyödyntämistä valvontatietojen tutkimisessa ja rakentaa klusterointityökalu helpottamaan tietojen analysoimista.

Tässä työssä tutkittiin Energiaviraston viranomaisvalvonnassa keräämän sähkönjakeluverkonhaltijoiden valvontadatan analysointia käyttämällä k-means ja k-medoids klusterointia. Työssä arvioitiin klusteroinnin hyötyjä aineiston analysoimisessa. Matlabilla rakennettiin klusterointityökalu, jossa käytettiin Matlabin sisäisiä k-means ja k-medoids funktioita. Klusterointityökalun parametrit, kuten k-arvo ja käytettävä etäisyydenmittaus menetelmä määriteltiin. Rakennettua klusterointityökalua testattiin klusteroimalla 0,4 kV, 1-70 kV ja 110 kV maakaapelointiasteita vuodelta 2017. Testin perusteella menetelmät toimivat kolmella muuttujalla.

Aineistona käytettiin sähkönjakeluverkonhaltijoiden teknisiä tunnuslukuja ja kohtuullisen hinnoittelun laskelmia vuodelta 2017. Teknisten tunnuslukujen esittäessä sähköverkonhaltijoiden fyysisiä ominaisuuksia, kohtuullinen tuotto esittää taloudellista onnistumista. Molemmista aineistoista tehtiin suppeammat versiot. Tekniset tunnusluvut jaettiin kolmeen osaan ja kohtuullisen hinnoittelun laskelmat supistettiin 9 muuttujaan. Lisäksi todettiin klusterointityökalun soveltuvan universaalin dataan.

Klusterointituloksia analysoitiin etsimällä selittäviä tekijöitä klustereiden muodostumiseen ja aineistojen välisiin samankaltaisuuksiin. Selviä yksittäisiä tekijöitä ei löytynyt. Teknisiä tunnuslukuja selitti eniten yhtiöiden käyttöpaikoilla ja siirretyllä energialla selittyvät kokoon perustuvat ominaisuudet. Lisäksi selittäviä tekijöitä jaottelun perusteella olivat keskijänniteverkon keskeytysluvut, suurin siirretty tuntikeskiteho ja 0,4 kV johtopituus. Kohtuullisen hinnoittelun laskelmien tunnuslukuja yhdisti vahva korrelaatio verkon nykykäyttöarvoon. Aineistojen välillä ei löytynyt klusterointituloksista yhteisiä tekijöitä.

Avainsanat: Energiavirasto, klusterointi, k-means, k-medoids, sähköverkonhaltija, regulaatio, Matlab

ABSTRACT

Joonas Kari: Utilization of clustering in the supervision of Finnish electricity distribution system operators
Master of Science Thesis
Tampere University
Master's Degree Programme in Electrical Engineering
March 2019

Background for this master's thesis is the goal of Finnish Energy Authority to find new connections between indicators and economic success of electricity distribution system operators. Goals of this thesis are to analyze usefulness of clustering in research of the collected data and to develop a clustering tool to help the analysis.

In this thesis the data collected by the Finnish Energy Authority in the supervision of electricity distribution system operators was analyzed using the k-means and k-medoids clustering. The work consisted of evaluating benefits of clustering and developing a clustering tool. A clustering tool was built in Matlab, by using Matlab's internal k-means and k- medoids functions. The parameters of the clustering tool, such as the k-value and the distance measurement method were defined. The developed clustering tool was tested by clustering the 0.4 kV, 1-70 kV and 110 kV ground cable percents of the year 2017. Clustering tool did work with three-dimensional test data.

The two databases consisted of the technical indicators of the electricity distribution system operators and the calculations of the reasonable return for 2017. The technical indicators represent the physical characteristics of the distribution system operators and the reasonable return represents economic success. Limited versions of both databases were made. The technical indicators were divided into three parts and the calculations of reasonable return were reduced to 9 variables. Clustering tool operation for universally selected data were tested and verified.

The results of clustering were analyzed by seeking explanatory factors for clusters formation and similarities between databases. No obvious individual factors were found. The most significant common variables of clusters were size-based properties such as number of customers and the transferred energy. Second common variables were the interruptions of the medium voltage network, the highest transmitted hourly power and 0.4 kV cable total length by the basis of the breakdown. Variables of the reasonable return limited database had a strong correlation with current network value. Clustering results of the limited and the non-limited databases were close to equal. There were no common clustering results between the technical indicators and reasonable return.

Keywords: Clustering, distribution system operator, Finnish Energy Authority, regulation, k-means, k-medoids, Matlab

ALKUSANAT

Työ tehtiin Energiavirastolle ja aloitettiin Verkot-ryhmässä vuoden 2018 lokakuun alussa Tampereen teknillisen yliopiston aikana ja saatiin päätökseen Tampereen yliopistolle vuoden 2019 toukokuussa. Haluan kiittää työn tarkastajina toimineita professori Pertti Järventaustaa ja tutkijatohtori Antti Mutasta hyvistä kehitysideoista ja kommenteista. Järventaustan raudanluja kokemus auttoi opinnäytetyön raamien ja sisällön muodostamisessa. Työn aihe oli haastava ja itselle työn alussa melko tuntematon ja Mutasen asiantuntijuus klusteroinnista mahdollisti lähtökohtaisesti koko työn. Molemmille vielä kiitokset kovasta työstä tämän diplomityön tiimoilta.

Energiaviraston puolelta haluan kiittää erityisesti työn ohjaajaani verkkoinsinööri Joel Seppälää koko työn ajan saadusta ohjauksesta ja tuesta työhön liittyen. Kiitokset myös koko Verkot-ryhmälle, jolta sain moneen kysymykseen vastaukset ja jaksamista vääntää diplomityötä hyvän porukan tukemana. Työn tekemiseen sain riittävästi aikaa ja työrauhaa, jotka molemmat olivat itselle hyvin tärkeitä asioita.

Erityiskiitos vanhemmilleni työelämän neuvoista ja oikeastaan kaikesta muustakin, sekä Mimosalle rakkaudesta, kärsivällisyydestä, ja ymmärtämisestä työn viemään aikaan ja jaksamiseen. Lisäksi kolmet (3) kiitokset ystävilleni diplomityön aikaisesta tuesta ja vapaa-ajan seurasta. Työ tehtiin opiskeluaikojen Joopen Sähkövoiman hengessä, pidetään lippu korkealla!

Hämeenlinnassa, 19.6.2019

Joonas Kari

SISÄLLYSLUETTELO

LYHENTEET JA MERKINNÄT

ABC-aineisto	aineisto, jossa ovat mukana volyymi-, panos- ja toimitusvarmuus-muuttujat. Katso liite A
AMR	engl. Automatic Meter Reading, automatisoitu mittariluenta
CAIDI	engl. Customer Avarage Interruption Duration Index, kuluttajan kokema keskimääräinen keskeytysaika
CAIFI	engl. Customer Avarage Interruption Frequency index, kuluttajan kokema keskimääräinen keskytys tiheys
KAH	keskeytyksestä aiheutunut haitta
KOPEX	kontrollotavissa olevat operatiiviset kustannukset
NaN	engl. Not a Number, tyhjä arvo
PAM	engl. Partition Around Medoids, Osittelu medoidien ympärillä
PCA	engl. Principal Component Analysis, komponenttianalyysi
SSE	engl. Sum of Squared Error, virheen neliösumma
StoNED	engl. Stochastic Non-smooth Envelopment of Data
WACC	engl. Weighted Avaraged Cost of Capital, pääoman keskimääräinen kustannus

A, B	skalaarinen vakio
CH	Calinski-Hrabasz arvo
$d(\bar{p}, \bar{q})$	vektorin $\bar{p}$ euklidinen etäisyys vektorista $\bar{q}$
dmin,t	pienin etäisyyden keskiarvo havaintovektorille $\bar{t}$
dka,t	havaintovektorin $\bar{t}$ keskimääräinen etäisyys klusterin muihin havaintovektoreihin
$d_{t,s}$	havaintovektoreiden $\bar{t}$ ja $\bar{s}$ keskinäinen etäisyys toisistaan
$E_n(log(W_k))$	odotusarvo Monte Carlo -näytteistykseen referenssijakaumasta
Gap(k)	Gap-analyysin arvo klusterin lukumäärälle *k*
i	muuttujan indeksi
j	klusterin indeksi
K	klusterien kokonaismäärä
μ_i	muuttujan *i* keskiarvo
μ	keskiarvo datasta
N	havaintovektoreiden kokonaislukumäärä (yhtiöiden lukumäärä)
n_i	muuttujan *i* arvojen lukumäärä
σ_i	muuttujan *i* keskihajonta
p_{ti}	havaintovektorin *p* vektorialkio muuttujalle *t* klusterissa *j*
p(A,B)	Pearsonin korrelaatiokerroin luvuille *A* ja *B*
q_{ji}	klusterin *j* keskipistevektorin alkio muuttujalle *i*
R	Davies-Bouldin indeksi
$R_{i,j}$	Davies-Bouldin indeksi havaintovektoreille *i* ja *j*
S_i	klusterin *i* hajanaisuutta kuvaava muuttuja
S_j	klusterin *j* hajanaisuutta kuvaava muuttuja
SI	Siluetti arvo
SS_b	klusterien välinen varianssi
SS_w	klusterin jäsenten välinen varianssi
t	klusterin jäsen
$\bar{t}$	havaintovektori klusterin jäsenelle *t*
x_{ti}	havaintovektorin $\bar{t}$ alkio muuttujalle *i*

$x_{i,norm}$	normalisoitu x_{ti}
y_{ji}	klusterin *j* keskipistevektorin alkio muuttujalle *i*

1. JOHDANTO

Robotiikka, automaatio, koneoppiminen ja Big data. Teknologia on tuonut mukanaan uusia keinoja kerätä dataa ja mahdollisuuksia tutkia sitä tarkemmin kuin ennen. Tietoa myös digitalisaation myötä kerätään aikaisempaa enemmän. Monipuolinen ja kasvava tietomäärä lisää luonnollisesti myös tietojen analysoimiseen vaadittavaa aikaa ja osaamista. Tietotekniikan avulla pystytään käsittelemään rutiininomaisia tehtäviä ja vähentämään käsin tehtävää työtä. (Ventä et al. 2018) Keinoja hyödyntää uusia teknologioita kehitetään jatkuvasti, tehostamisen ja taloudellisen hyödyn toivossa (Freddi 2018).

Sähkömarkkinat, erityisesti sähkön siirto- ja jakeluverkkotoiminta ovat olleet vahvassa viranomaisvalvonnassa markkinoiden avauduttua vuonna 1995. Näitä valvoo ja kehittää Työ- ja elinkeinoministeriön alainen virasto, nykyiseltä nimeltään Energiavirasto. Valvonnan toteuttamisessa on tärkeässä roolissa verkkoyhtiöiltä kerätyt valvontatiedot, joita kerätään useista eri osa-alueista. Tietoja on kerätty paljon ja niitä on pääsääntöisesti käytetty alkuperäiseen keräämistarkoitukseensa, monipuolisen hyödyntämisen ollessa haasteellista suuren määrän takia. Analysoimalla tietoja automatisoidusti saadaan lisää tietoa valvontamenetelmien hyödyistä ja vaikutuksista, mikä auttaa itse valvonnassa ja sen kehittämisessä. Viraston tulee jatkuvasti kehittää valvontamenetelmiään muuttamatta kuitenkaan jo hyväksi havaittuja vakiintuneita käytäntöjään (Energiavirasto 2015).

Sähkönjakeluverkkotoiminta on monopolitoimintaa, jossa korostuu kuluttajien oikeudenmukainen kohtelu, sekä jakeluverkkoyhtiöiden välinen tasavertainen ja yhdenmukainen viranomaistoiminta. Vertailemalla yhtiöiden suoriutumista erilaisissa toimintaympäristöissä voidaan saada uutta tietoja jakeluverkkoyhtiöiden suoriutumisesta monopolialalla valvonnan näkökulmasta. Yhtiöiden toiminnan parempi ymmärtäminen auttaisi virastoa kehittämään ja parantamaan valvontamenetelmiään. Kerätyistä tiedoista on mahdollista saada lisää tietoa jakeluverkkoyhtiöistä, niiden onnistumisesta valvonnan näkökulmasta ja onnistumisen menestystekijöistä. Tähän tarvittaisiin työkaluja, joita ei ole yleisesti näin pienelle markkinasegmentille tarjolla kaupallisesti kohtuullisella hinnalla.

1.1 Tutkimuksen tavoitteet

Tässä diplomityössä tutkitaan koneoppimisessakin käytettävän dataklusteroinnin hyödyntämistä jakeluverkonhaltijoiden vertailemiseen ja avaintekijöiden tunnistamiseen teknisistä tunnusluvuista. Klusteroinnilla tarkoitetaan tässä tilastoidun tiedon automaattista ja koneellista jaottelua samankaltaisten arvojen tai annettujen parametrien perusteella.

Klusteroinnilla tutkitaan Energiaviraston kohtuullisen hinnoittelun valvonnassa käytettyjä valvontatietoja. Tekemällä klusterointeja useilla tunnusluvuilla voidaan tunnistaa tietyn muuttujan tai usean muuttujan suhteen samantyyppiset jakeluverkonhaltijat.

Työn taustalla olevana tehtävänä on luoda uutta tietoa tunnuslukujen ja yhtiöiden taloudellisten tulosten välisistä yhteyksistä. Tähän pyritään rakentamalla työkalu helpottamaan avaintekijöiden tuomista esille erilaisista aineistoista. Mikäli yhteyksiä tiettyjen tietojen välillä löydetään, voidaan uusia tietoja hyödyntää valvonnan kehittämisessä ja tutkimuksessa. Oikealla työkalulla pystytään myös tutkimaan esimerkiksi kannustimien suhteellisia vaikutuksia yhtiöihin. Käytännössä tarvitaan täysin uusi työkalu, jota ei Energiavirastolla ole, ja työn tärkeimpänä tavoitteena on saada aikaan toimiva prototyyppi, josta kehitetään tutkimuksen edetessä toimiva työkalu. Työkalu suunnitellaan ja rakennetaan Matlab -ohjelmistolla yleiskäyttöiseksi apuvälineeksi isojen tai monimutkaisten Excel taulukoiden nopeaan tulkitsemiseen ja tunnuslukujen automaattiseen järjestelyyn ryhmiksi eli klustereiksi. Työkalun klusterointiin käyttämät tunnusluvut ovat valmiiksi kerätyistä valvontatiedoista valittuja.

Klusterointia voidaan hyödyntää eri tarkoituksiin kerättyyn dataan ja eri vuosilta oikeilla klusterointimenetelmillä. Samat klusterointimenetelmät eivät ole välttämättä sopivia eri tyyppisiin aineistoihin ja yhtä universaalia menetelmää ei ole. Yksinkertaisuuden vuoksi tämän työn klusteroinneissa käytetään pelkästään numeerisesti saatavilla olevia lähtötietoja. Tutkimuksessa käytettävä aineisto on kvantitatiivinen ja valmiiksi kerätty osana viranomaisvalvontaa. Rakennettavassa työkalussa tullaan hyödyntämään muun muassa julkisestikin saatavilla olevia sähköverkon teknisiä tunnuslukuja, jotka ovat saatavilla Energiaviraston internetsivuilta. Lisäksi aihetta rajataan pelkästään sähkönjakeluverkkoyhtiöihin, eikä tässä työssä hyödynnetä suurjännitteisistä sähkönjakeluverkkoyhtiöistä kerättyä dataa.

Virastolla on kerättynä teknisiä ja taloudellisia tunnuslukuja vuodesta 1996 lähtien ja tässä työssä hyödynnetään vuosia 1997-2017. Tarkempi tarkastelu tullaan tekemään vuodelle 2017 ja työkalun tuottamat tärkeimmät tulokset käydään työssä läpi. Teknisistä tunnusluvuista pyritään löytämään yhteisiä tekijöitä yhtiöiden välillä. Teknisten tunnuslukujen klusterointituloksia verrataan kohtuullisen hinnoittelun laskelmiin. Kohtuullisen hinnoittelun laskelmia käytetään myös testaamaan muuttujien valintaa korrelaation avulla. Tekniset tunnusluvut kuvaavat fyysistä verkkoa ja sen ominaisuuksia, kohtuullisen hinnoittelun laskelmien perustuessa pitkälti taloudellisiin ominaisuuksiin. Mikäli fyysisesti samantyyppiset yhtiöt käyttäytyvät eri tavoin kohtuullisen tuoton näkökulmasta, voitaisiin saada arvokasta uutta tietoa.

1.2 Tutkimuksen lähtökohdat

Jakeluverkkoyhtiöt ovat kauppa- ja teollisuusministeriön työryhmän tekemän selvityksen perusteella jaoteltu keskeytysherkkyydeltään loogisesti kolmeen luokkaan, perustuen

maakaapelointiasteeseen. Vuodesta 2005 lähtien luokkina ovat toimineet: kaupunki (käytetään myös ”city” -nimeä) (maakaapelointiaste > 75 %), taajama (30-75 %) ja maaseutu (< 30 %). (Hirvonen et al. 2006, s 29) Selvityksen perusteella esimerkiksi keskeytysten määrän ajatellaan olevan osittain riippuvainen maakaapelointiasteesta. Samantyyppistä jaottelua on alalla yleisesti hyödynnetty monissa eri yhteyksissä. Ilman klusterointia tehty jaottelu kolmeen ryhmään loogisten ominaisuuksien kautta viittaisi fyysisisiin jaotteluperusteisiin ja klusteroinnin olisi siten mahdollista löytää teknisistä tunnusluvuista fyysisiin ominaisuuksiin perustuvaa jaottelua, esimerkiksi maakaapelointiasteen perusteella.

Tutkimuksen ydin muodostuu klusterointityökalun muodostamisesta ja soveltuvien parametrien valitsemisesta. Klusterien välillä voidaan tehdä vertailuja esimerkiksi, ovatko samantyyppiset yhtiöt hyötyneet samalla tavalla valvontamenetelmien kannustimista. Klusteroinnilla voidaan myös etsiä yhtäläisyyksiä aineistojen väliltä tai ulkoisista aineistoista, esimerkiksi maantieteellisten sijainnin ja keskeytystilastojen välillä tai velkaantumisasteen ja keskeytyksestä aiheutuneen haitan (KAH) kehittymisen välillä. Vaikka tässä tutkimuksessa saadaankin oheistuotteena uutta tietoa yhtiöiden valvontaan, on työn pääpaino edelleen itse työkalun kehittämisessä ja soveltumisessa viraston toimintaan.

Viraston nykyinen menetelmä tietojen analysointiin perustuu käsin tarkastettaviin tilastoihin. Tämän tyyppisessä tarkastelussa korostuu tarkastavan henkilön substanssin tuntemus ja ajankäyttö. Valvontatietoja on kerätty paljon ja nykyinen menetelmä saattaa johtaa tilanteeseen, jossa voidaan tehdä virheellisiä päätelmiä ja inhimillisiä virheitä. Käsin tehty analysointi ja tarkastus on hidasta ja tämän takia analysointeja tiedoista tehdään välttämättömien lisäksi vain vähän.

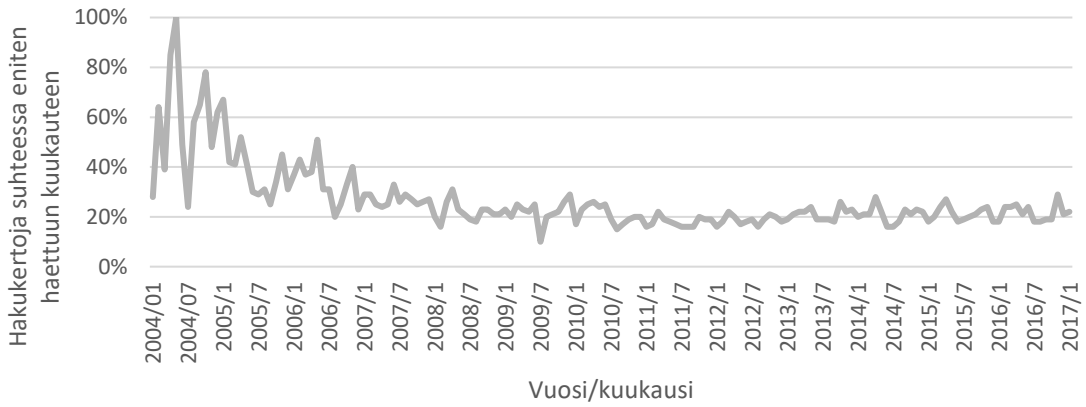
Tutkimuksessa käytettäviä klusterointimenetelmiä ei ole hyödynnetty ennen verkkovalvonnassa. Klusteroinnin soveltuvuutta valvontaan ei tunneta etukäteen ja se on keskeinen osa tätä tutkimusta. Vähäisen tiedon takia tutkimus aloitetaankin tutkimalla kirjallisuutta, jotta pystytään rakentamaan kokonaiskuva klusteroinnista. Kirjallisuusselvityksen perusteella pyritään valitsemaan soveltuvat klusterointimenetelmät. Esiselvityksen perusteella rakennetaan toimivat prototyypit ja testataan niiden toiminta. Näin menettelemällä pyritään löytämään parhaiten soveltuva klusterointimenetelmä. Vaikeuksia aiheuttaa lähtöaineistossa valvontatietojen muuttuminen vuosien varrella muun muassa yritysfuusioiden ja kerättyjen valvontatietojen suhteen.

Klusterointimenetelmiä on olemassa huomattava määrä, joista vain osa sopii aineiston ja halutun lopputuloksen kannalta käytettäväksi tässä tutkimuksessa. Matemaattisesti klusterointia on tutkittu jo kymmeniä vuosia. Pelkästään k-means menetelmää on aloitettu tutkimaan jo 1950-luvulla (Jain 2010). Klusterointiin ja klusterointimenetelmiin, sekä niiden hyödyntämiseen löytyy teoreettisten tieteellisten julkaisujen lisäksi myös menetelmien soveltamista useilta eri aloilta muun muassa sähkövoimatekniikasta, lääketieteestä ja tietotekniikasta.

Klusterointia on hyödynnetty sähkövoimatekniikassa esimerkiksi sähkönkäyttäjien luokittelemiseen Suomessa (Mutanen 2018; Räsänen et al. 2010). Matlabia tutkimuksissaan ovat hyödyntäneet Mutanen ja Koivisto et al. (2013), Koiviston tutkimuksessa on käytetty k-means ja PCA menetelmiä Matlabin sisäisistä toolboxeista. Mutasen tekemässä väitöskirjassa (Mutanen 2018) klusterointia hyödynnetään AMR-mittarien tuntimittaustietojen avulla mallintamaan verkostolaskennassa hyödynnettävien käyttäjäryhmien kuormitusprofiileja. Tässä työssä klusterointiin hyödynnetään Matlabin sisäisiä algoritmeja Statistics and Machine learning -toolboxista, kuten myös Koivisto et al. (2013) tutkimuksessa on hyödynnetty, edellyttäen niiden sopivan aineistoon.

Sähkönkäyttäjien luokitteluun soveltuvia menetelmiä on vertailtu Chicco et al. (2006) tutkimuksessa, jossa oli mukana muun muassa k-means ja dimensioiden redusointia PCA menetelmällä. PCA:n hyödyntämistä k-means klusterointimenetelmän toimivuuden parantamisessa on tutkittu Ding:n ja He:n tutkimuksessa (Ding & He 2004). Klusteroinnin hyödyntämistä sähköverkon toimitusvarmuuteen on tutkittu muun muassa Intiassa (Kalyani & Swarup 2011), sekä sähkönjakeluverkon ennakoivaan huoltoon Kiinassa (Cui et al. 2016). Lisäksi klusterointia on käytetty hyväksi Brasiliassa CAIDI:n ja CAIFI:n jakeluverkonhaltijoiden valtakunnallisen vertailutason määrittämisessä (Tanure et al. 2006).

Google Scholarilla hakemalla otsikoista hakusanat ”data clustering” ja rajaamalla tulokset ennen vuotta 2017 löytyy 13 500 tulosta. Rajaamalla vuoteen 2007 asti tuloksia löytyi 3 250 kappaletta. Google trendin mukaan samat hakusanat ovat saaneet kerätyn aineiston perusteella vuonna 2004 huippunsa kuvan 1 mukaisesti ja tämän jälkeen hakumäärät ovat laskeneet ja tasaantuneet noin 20 % paikkeille. Kuvassa 1 hakumääriä verrataan huippumäärään eli vuoden 2004 toukokuuhun.



Kuva 1. *Google Trend:n perusteella vuosien 2004-2017 hakukerrat sanoille "data clustering".*

Google Scholarin ja kuvassa 1 olevan Google Trend:n testien perusteella klusterointi ja sen hyödyntäminen ovat herättäneet kiinnostusta tasaisesti ja klusterointi on viimeisen 10 vuoden aikana vakiinnuttanut paikkansa tieteellisten työkalujen joukossa.

1.3 Työn rakenne

Kirjallisuusselvityksen avulla tutustutaan klusterointimenetelmiin ja valitaan parhaiten soveltuvat menetelmät, joita hyödynnetään klusterointityökalussa. Tutkimuksen tavoitteet, rakenne ja lähtökohdat esitellään ensimmäisessä luvussa. Toisessa luvussa käsitellään Energiaviraston tehtävät valvovana viranomaisena, sekä valvonnan nykytilanne. Neljännen valvontajakson valvontamenetelmät käydään lyhyesti läpi ja käsitellään tarkemmin kohtuullisen tuoton määräytyminen ja aineistojen vaikutus niissä.

Luvussa 3 tutustutaan k-means ja k-medoids klusterointiin, sekä tässä työssä käytettävään etäisyyden mittausmenetelmään. Klusterointiin liittyviin ongelmiin, kuten eriytymiin (engl. *outlier)*, ulottuvuuksien määrään ja klusterien lukumäärän valintaan liittyviin analyyseihin perehdytään myös siltä osin, kuin tässä työssä niitä tullaan hyödyntämään.

Matlabilla tehtyä klusterointityökalua ja tarkempia parametreja käsitellään neljännessä luvussa. Aluksi tutustutaan Matlabin klusterointiominaisuuksiin ja käytettyihin funktioihin. Lisäksi klusteroinnille ominaisten ongelmien ratkaisuja tässä työssä käydään läpi. Tämän jälkeen suunnitellaan ja rakennetaan työkalun prototyyppi.

Suunnitteluvaiheessa arvioidaan tarve lähtöaineiston käsittelylle klusterointien onnistumiseksi. Ensimmäisessä prototyypissä ei käytetä kaikkia tunnuslukuja. Prototyyppiin käytetään maakaapelointiasteita kaikilta jakeluverkkoyhtiöiltä. Suunnittelussa on tärkeää huomioida, mitä lähtötietoja käytetään prototyypissä ja mitä tietoja näistä saadaan ulos. Lähtötietojen ja parametrien asettamisen jälkeen kirjoitetaan Matlab -koodi. Prototyyppien testauksen jälkeen prototyyppiä rakennetaan työkaluksi klusteroimaan kaikkia tunnuslukuja kerralla ja tämän jälkeen laajennetaan klusteroimaan useita vuosia.

Viidennessä luvussa esitellään tulokset ja tehdään päätelmiä klusteroinnin toimivuudesta valvontatietojen analysoimisessa. Tuloksia analysoidaan lisää ja pohditaan klusteroinnin hyötyjä ja haittoja, sekä ongelmia.

2. VALVONTATIETOJEN HYÖDYNTÄMINEN NYKYÄÄN JA LÄHITULEVAISUUDESSA

Laissa Energiavirastosta on kuvattu viraston tehtävät yleisesti seuraavasti: ”Sähkö- ja maakaasumarkkinoiden valvontaa ja seurantaa, sähkö- ja maakaasumarkkinoiden toimivuuden, energiatehokkuuden ja uusiutuvan energian käytön edistämistä sekä energiapolitiikan, kasvihuonekaasujen päästökaupan ja energiatehokkuuden toimeenpanotehtäviä varten on Energiavirasto.”. Lain mukaan Virasto edustaa ja vastaa valtion etua hoitamissaan tehtävissä erilaisissa viranomaistahoissa, kuten esimerkiksi tuomioistuimissa. (Laki Energiavirastosta 2015) Viraston nykyinen toimivalta perustuu hallituksen esitykseen 20/2013, jossa määritellään muun muassa sähkö- ja maakaasunjakelun valvontaviranomaisen tehtäviä. Virasto toimii itsenäisesti työ- ja elinkeinoministeriön alaisuudessa ja sillä on oma talousbudjetti. (Hallituksen esitys 2013)

Energiavirastosta määrätyn lain mukaisten tehtävien suorittamiseksi, Energiaviraston tulee sähkömarkkinalain 27 §:n mukaan saada verkonhaltijalta tarpeelliseksi katsovansa tunnusluvut (Sähkömarkkinalaki 2013), jotka verkonhaltijan on valvontalain 30 §:n perusteella velvollisuus toimittaa (Laki sähkö- ja maakaasumarkkinoiden valvonnasta 2013). Jättämättömiä tietoja vaaditaan jättämään ja tätä pystytään tarvittaessa tehostamaan esimerkiksi uhkasakoilla valvontalain 31 §:n perusteella (Laki sähkö- ja maakaasumarkkinoiden valvonnasta 2013). Energiaviraston keräämä tieto on julkisuuslain mukaisesti julkisesti saatavilla. Poikkeuksia ovat muun muassa asiakirjat, jotka sisältävät tietoja yksityisestä liike- tai ammattisalaisuudesta 24 §:n 1 momentin mukaisesti (laki viranomaisten toiminnan julkisuudesta 1999).

Energiaviraston suorittamaa valvontaa on lähtökohtaisesti kahta eri tyyppiä. Välitön valvonta tapahtuu yksittäisten tapausten ja sitä koskevien yksittäisten tietojen perusteella. Välillinen valvonta tapahtuu valvontamenetelmillä, joiden toiminta perustuu kerättyihin valvontatietoihin. Kerättävät tiedot ovat määritelty valvontamenetelmissä, tunnuslukumääräyksessä ja kauppa- ja teollisuusministeriön sähköliiketoimintojen eriyttämisestä annetussa asetuksessa. Valvontamenetelmien mukaan valvonnan toteutuminen edellyttää valvontatietojen olevan oikein ja jätetty ajallaan, sekä siinä muodossa kuin ne ovat määritelty. (Energiavirasto 2015)

Energiavirastossa tiedot käydään inhimillisten virheiden varalta läpi ja tarvittaessa pyydetään korjaamaan tai täydentämään annettuja tietoja. Annettujen tietojen oikeellisuus perustuu pääosin yhtiöiden juridiseen ja moraaliseen vastuuseen, jota täydennetään viraston tekemillä tarkastuskäynneillä ja pistokokeilla. Julkisten tietojen lisäksi virastolla on kerättynä myös luottamuksellista valvontatietoa, jota hyödynnetään viranomaisvalvonnassa ja sisäisessä tutkimuksessa.

Kerättäviä tietoja ovat muun muassa verkon rakennetiedot, tekniset ja taloudelliset tunnusluvut, tilinpäätöstiedot ja kehittämissuunnitelmat. Verkonrakennetiedot jätetään vuosittain maaliskuun loppuun mennessä, eivätkä ne ole julkisia. Tekniset ja taloudelliset tunnusluvut sekä eriytetyn tilinpäätöksen tiedot tulee jättää toukokuun loppuun mennessä. Tekniset ja taloudelliset tunnusluvut, kuten myös tilinpäätöstiedot julkaistaan. Vuosittain kerättävien tietojen lisäksi kerätään myös muita tietoja. Esimerkiksi toimitusvarmuuden parantamisen seurannassa käytetään joka toinen vuosi kerättäviä kehittämissuunnitelmia. Kehittämissuunnitelmat ovat osittain julkisia. (Energiavirasto 2015)

Tässä tutkimuksessa käytetty tieto on julkista teknisten tunnuslukujen osalta. Kohtuullisen hinnoittelun laskelmat eivät ole julkisia, eikä niitä käytetä tämän työn osalta muuta kuin vertailuaineistona teknisille tunnusluvuille ja ulottuvuuksien vähentämisen testaamiseen teknisistä tunnusluvuista eriävänä näennäisesti riippumattomana aineistona.

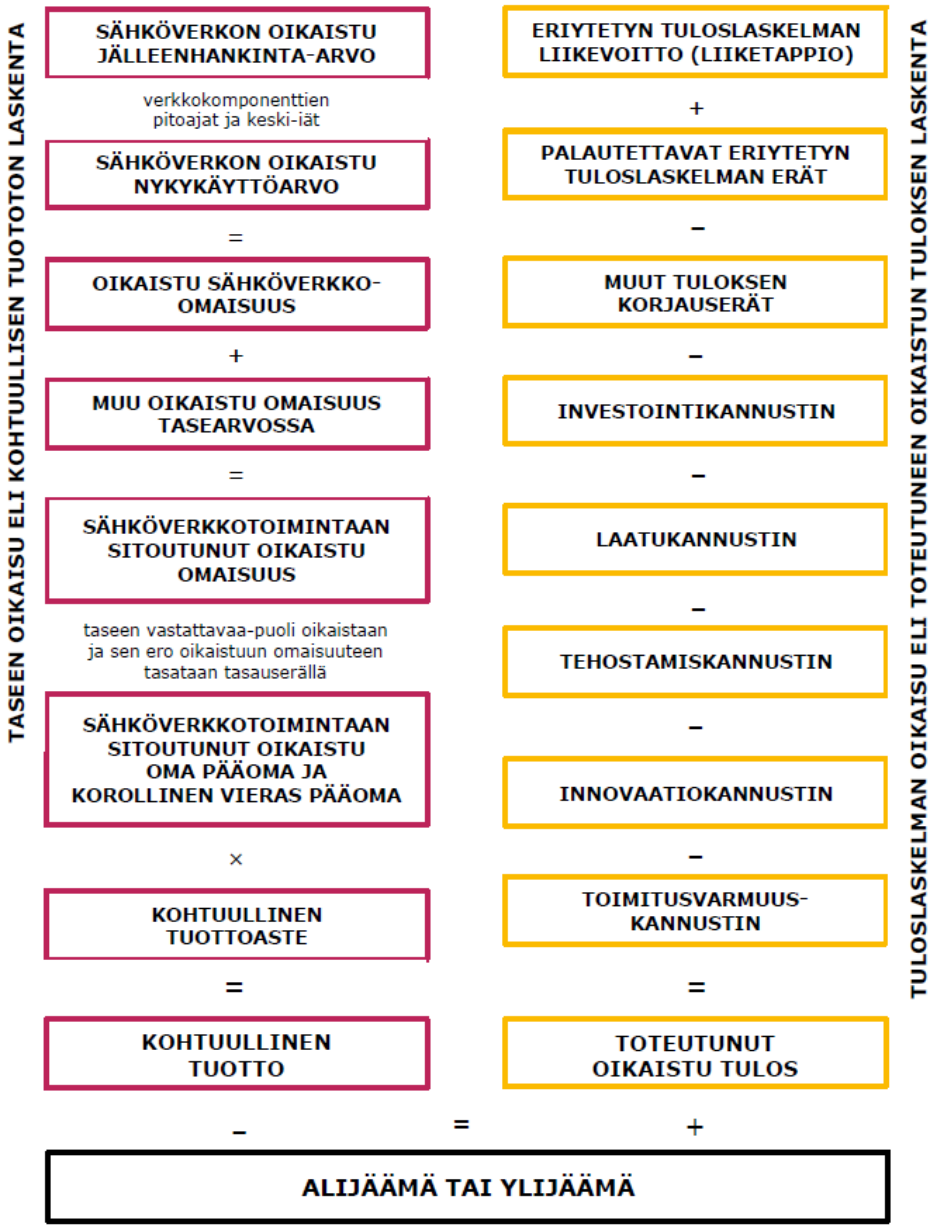
Valvontatiedot kerätään ja tarkastetaan internet-pohjaisen valvontatietojärjestelmän avulla. Nykyisessä valvontatietojärjestelmässä on käytettävissä tiedot alkaen vuodesta 2005. Tietojen käsitteleminen vaatii ulkoisia työkaluja, joista yhtenä työkaluna käytettävään Microsoft Accessiin on kerätty viraston toiminnan aikana vuodesta 1996 lähtien jätetyt tekniset ja taloudelliset tunnusluvut. Valvontatietojärjestelmästä tiedot saadaan tuotua Microsoft Excel taulukko-ohjelmistoon, jossa tietoja pystytään käsittelemään tai viemään edelleen muihin Excelin kanssa yhteensopiviin ohjelmistoihin. Tarkastettaessa tietoja verrataan yhtiöiden aikaisempiin tietoihin ja muihin loogisesti samankaltaisiksi ajateltuihin yhtiöihin, esimerkiksi perustuen maakaapelointiasteisiin tai siirrettyyn energiaan. Vertailut tehdään joko suoraan valvontatietojärjestelmässä tai apuohjelmissa yksinkertaisilla funktioilla ja tuloksia vertaillaan visuaalisesti. Mahdolliset tietojen analysoinnit toteutetaan samoilla menetelmillä.

Otettaessa huomioon vain sähkönjakeluverkkoyhtiöt ilman suurjännitteisiä- tai maakaasunjakeluverkkoyhtiöitä, kerätään tietoja huomattavat määrät. Valvontatiedot ovat lähtökohtaisesti yksityiskohtaisia ja kerätään vuosittain 77 (vuonna 2017) sähkönjakeluverkkoyhtiöltä. Tietoja on kerätty vuodesta 1996 lähtien. Jakeluverkonhaltijoiden määrät ovat muuttuneet vuosien aikana. Osa yhtiöistä ei ole enää jakeluverkkoyhtiöitä tai ovat yritysfuusioiden kautta sulautuneet toisiin yhtiöihin ja kadonneet yhtiölistalta. Teknisiä tunnuslukuja on kerättynä 159 kappaletta vuodelle 2017 ja siten yksittäisiä arvopisteitä muodostuu $77 \cdot 159$ eli yhteensä 12 243 arvopistettä per vuosi. Näiden lisäksi esimerkiksi pelkästään verkon rakennetiedoista muodostuu todella suuri tietomäärä, sisältäen muun muassa verkkokomponenttien määrät, keski-iät, nykykäyttöarvot ja tasapoistot kaikilta valvonnan piiriin kuuluvilta yhtiöiltä. Tutkimuksessa käytettävät tunnusluvut ja niitä koskeva tunnuslukumääräys on saatavilla Energiaviraston internet-sivuilta.

2.1 Jakeluverkonhaltijan tuoton määräytyminen

Valvontapäätöksellä vahvistetaan vuosittain tehtävät laskelmat ja laskelmien perusteella verkonhaltijoiden valvontajakson kohtuullisen tuoton alittavan tai ylittävän rahamäärän. Valvontajakson sisällä havaituista virheistä voidaan kommentoida viimeistään valvontajakson lopussa annettavasta valvontapäätöksen luonnoksesta. Hyväksyttyyn valvontapäätökseen voidaan jälkikäteen hakea lakiteitse muutoksia. Valvontapäätöksen alijäämä voidaan yhtiön niin halutessa tasoittaa seuraavan valvontajakson aikana. Ylijäämä on tasoitettava seuraavan valvontajakson aikana. Painavin syin voidaan virastolta anoa ali- ja ylijäämien tasoittamiseen lisäaikaa. Ali- ja ylijäämät vaikuttavat kuluttajien maksamaan siirtohintaan. Alijäämä tarkoittaa yhtiön perineen siirtomaksua vähemmän kuin olisi ollut mahdollista ja antaa mahdollisuuden korottaa maksuja tasattavaksi. Ylijäämä velvoittaa tasaamaan kuluttajille siirtomaksuissa perityt liiat maksut. (Energiavirasto 2015)

Jakeluverkonhaltijan tulee eriyttää muu liiketoiminta sähköverkkotoiminnastaan kohtuullisen tuoton laskennassa. Eriyttämisestä on ohjeistettu viraston suosituksessa. (Sähkömarkkinalaki 2013; Energiavirasto 2015). Energiavirasto laskee toimitetuista valvontatiedoista muun muassa sallitun tuoton määrittelemisessä käytettävät kannustimet. Kuva 2 esittää yhteenvetona sähköjakeluverkon kohtuullisen tuoton määritteleminen ja toteutuneen oikaistun tuloksen laskenta. Kuvasta on tärkeää huomioida eri komponenttien etumerkit, jotka kuvaavat onko komponentin vaikutus tuottoon bonusta vai sanktiota.



Kuva 2. *Kohtuullisen tuoton määräytyminen (Energiavirasto 2013).*

Kuvan 2 vasen puoli kuvaa yhtiön eriytettyihin tilinpäätöstietoihin perustuvia kustannuksia ja niistä johdettua kohtuullista tuottoa. Oikea puoli kuvaa toteutunutta tulosta perustuen tuloslaskelman oikaistuun tulokseen. Inflaatio otetaan huomioon molemmin puolin. Vasemmalla puolella inflaatio huomioidaan kertomalla sähköverkkotoimintaan sitoutunut oikaistu oma ja korollinen vieras pääoma nimellisellä kohtuullisella tuottoasteella, joka sisältää inflaatio odotuksen eli WACC -%. WACC-mallin parametrit on määritelty käyttäen apuna Energiaviraston teettämää lausuntoa (Ernst & Young Oy 2014). Oikealla puolella inflaatio korjataan kuluttajahintaindeksillä investointi-, laatu- ja tehostamiskannustimessa.

Pääsääntöisesti valvontamenetelmät ovat samat jakeluverkon- ja suurjännitteisen jakeluverkonhaltijan kesken. Erot löytyvät laatu-, tehostamis- ja toimitusvarmuuskannustimien laskennassa.

2.2 Kohtuullinen tuotto

Sähköverkon oikaistu jälleenhankinta-arvo saadaan verkkokomponenttien määrien ja yksikköhintojen perusteella. Oikaistusta jälleenhankinta-arvosta lasketaan nykykäyttöarvo todellisten komponenttien keski-ikien ja pitoaikojen perusteella. Taselaskennan mukaisesti pysyvänä vastaavana nykykäyttöarvon lisäksi oikaistun taseen puolella ovat: muut pysyvät vastaavat tasearvossaan, oikaistut vaihtuvat vastaavat tasearvossaan, vaihto-omaisuus ja myyntisaamiset vastaavasti tasearvossaan. Yhdessä näistä muodostuu sähköverkkotoimintaan sitoutunut oikaistu omaisuus ja taselaskennan vastaavaa puoli. Taselaskennan vastattavaa puolelle saadaan verkkotoimintaan sitoutunut oikaistu pääoma laskemalla yhteen oikaistu oma pääoma, oikaistu korollinen ja koroton vieras pääoma, sekä oikaistun taseen tasauserä. Alla olevassa taulukossa esitetään kuva 2:n vasen puoli tase vertailuna vastattavan ollessa oikealla ja vastaavan vasemmalla.

Taulukko 1. *Kohtuullisen oikaistun tuoton taselaskelma.*

Vastaavaa	Vastattavaa
• Verkon nykykäyttöarvo • Muut pysyvät vastaavat tasearvossaan • Vaihto-omaisuus tasearvossaan • Myyntisaamiset tasearvossaan	Oikaistu oma pääoma: • oikaistu oma pääoma tasearvossaan • annetut ja saadut konserniavustukset vähennettynä laskennallisella verovelalla • vapaaehtoiset varaukset • oikaistun taseentasauserä Oikaistu vieras pääoma: • korolliset ja korottomat velat tasearvossaan • pääomalainat tasearvossaan • annetun, mutta maksamattoman korollisen ja korottoman konserniavustuksen oman pääoman osuus • pakolliset varaukset tasearvossaan • muiden kuin sähköverkon hyödykkeiden laskennallisen verovelan osuus

Vastattavaa puoli kerrotaan kohtuullisen tuottoasteen kertoimella. Kohtuullinen tuottoaste lasketaan pääoman painotetun keskikustannuksen eli WACC-mallin perusteella. Kertolaskun lopputuloksesta saadaan kohtuullinen tuotto. (Energiavirasto 2015)

Verkon nykykäyttöarvon määrittelemiseen käytetään verkon rakennetietoja. Rakennetiedot sisältävät sähköverkon yleisimmät komponentit. Jakeluverkonhaltijat täyttävät tietoihin komponenttien lukumäärät, investoinnit, korvausinvestoinnit, poistot, ostot, myynnit, sekä keski-iän. Rakennetietojen avulla määritetään myös jälleenhankinta-arvo.

2.3 Toteutunut oikaistu tulos

Toteutuneen oikaistun tuloksen laskennassa pohjatietona on eriytetyn tuloslaskelman liikevoitto tai -tappio. Palautettavat eriytetyn tuloslaskelman erät sisältävät palautuskelpoisten liittymismaksujen nettomuutoksen, maksetut verkkovuokrat, suunnitelman mukaiset poistot liikearvosta, muihin tuottoihin ja kuluihin kirjatut verkonosuuden myyntivoitot ja -tappiot, sekä suunnitelman mukaiset poistot ja arvonalentumiset sähköverkon hyödykkeistä. Muihin tuloksen korjauseriin merkitään rahoitusomaisuuden kohtuulliset kustannukset.

Innovaatiokannustimen ajatuksena on tukea yhtiöitä kehittämään, tutkimaan ja käyttöönottamaan uusia toimintamenetelmiä ja tekniikkaa, kuten älykkäitä sähköverkkoja. Tutkimus- ja kehitystoiminnalla on ominaista kulujen syntyminen ennen tuottoja. Oikaistussa tuloksessa huomioidaan enintään kaikkien valvontajakson vuosien verkkotoimintojen liikevaihdon summasta 1 % innovaatiokannustimeen. Yksittäisen vuoden kohtuulliset tutkimus- ja kehityskustannukset voidaan ylittää tai alittaa ilman vaikutusta kannustimeen. Kannustimessa tarkastellaan valvontajaksoa kokonaisuutena. (Energiavirasto 2015)

2.3.1 Investointikannustin

Komponenttien yksikköhinnat on määritelty energiaviraston vuosien 2014-2015 tekemän kyselyn pohjalta. Kysely suoritettiin kaikkien sähkön jakeluverkonhaltijoiden ja suurjännitteisten jakeluverkonhaltijoiden kesken ja kyselyssä pyydettiin tietoja komponenttien toteutuneista investointikuluista. Yksikköhinnat ovat keskihajonnalla ja investointimäärillä painotettuja tuloksia. Kaikkien komponenttien yksikköhintoja ei ole pystytty vähäisen otannan takia määrittelemään edellä mainitulla tavalla ja muita menetelmiä on joidenkin komponenttien kohdalla käytetty. Muita määrityksessä avuksi käytettyjä keinoja ovat komponentin suhteelliset kustannusrakennetiedot, verkonhaltioilta pyydetyt kustannusarviot ja saman verkkokomponenttiryhmän muiden komponenttien tiedot. Nykyiset yksikköhinnat ovat voimassa kuluvalla neljännellä valvontajaksolla ja vuonna 2020 alkavalla viidennellä valvontajaksolla. (Energiavirasto 2015)

Yksikköhintojen avulla määritetään verkolle jälleenhankinta-arvo, joka perustuu käytössä olevien komponenttien todellisiin ikätietoihin. Kuvan 1 vasemmalla puolella oleva oikaistu verkon nykykäyttöarvo perustuu yksikköhintoihin, verkkokomponenttien pitoaikoihin ja ikätietoihin. Mikäli yhtiö on investoinut keskimääräisesti yksikköhintoja halvemmalla, verkon oikaistu jälleenhankinta-arvo kasvaa toteutuneita kustannuksia korkeammaksi. Oikaistusta jälleenhankinta-arvosta lasketaan vuosittainen tasapoisto jakamalla

jälleenhankinta-arvo verkonhaltijan valvontajakson alussa valitsemalla pitoajalla. Laskennallinen tasapoisto hyväksytään täysmääräisenä koko komponentin tosiasiallisen käyttöajan, eikä pitoajan ylittäminen vaikuta laskennalliseen tasapoistoon. Pitoajan ylittäminen laskee kuitenkin komponentin nykykäyttöarvon nollaan ja siten ohjaa korvausinvestointeihin. Tasapoisto ja verkonnykykäyttöarvo ohjaavat yhdessä jakeluverkonhaltijoita tekemään korvausinvestointeja kustannustehokkaasti ja riittävästi. Tasapoisto ohjaa yhtiötä pitämään komponentti toiminnassa yhtiön pitoaika-väliltä valitseman pitoajan keston. Investointikannustimen inflaatio korjataan kuluttajahintaindeksillä vuosittain jälleenhankinta-arvosta tehtäviin tasapoistoihin. (Energiavirasto 2015)

Investointikannustimen jälleenhankinta-arvo määritetään rakennetietojen perusteella ja laskennallinen tasapoisto määritetään edelleen jälleenhankinta-arvosta. Investointikannustin perustuu siten täysin rakennetietoihin ja viraston määrittelemiin yksikköhintoihin.

2.3.2 Laatukannustin

Laatukannustimen tarkoituksena on kannustaa verkonhaltijaa parantamaan sähkön jakelun laatua lain vaatimaa alimmaistasoa paremmaksi. Kannustimessa lasketaan tarkasteluvuoden toteutuneet keskeytyskustannukset ja vähennetään se vertailutasosta, joka määräytyy kyseisen yhtiön kahden edellisen valvontajakson toteutuneiden keskeytyskustannusten (KAH) keskiarvosta. Vertailutaso 4. valvontajaksolle on määritelty vuosien 2008-2014 keskijänniteverkon keskeytyskustannusten perusteella. Viidennellä valvontajaksolla huomioidaan myös suurjännitteisen jakeluverkon keskeytykset laatukannustimessa.

Virasto on kerännyt vuodesta 2013 lähtien suurjännitteisen jakeluverkon keskeytystilastoja viidennettä valvontajaksoa varten. Vertailutasosta ei poisteta suurhäiriötilanteista aiheutuneita keskeytyksiä, vaan näillä korvataan verkonhaltijalle häiriöistä aiheutuneet kustannukset. Keskeytyskustannusten laskentaa varten tarvitaan tieto keskeytysten lukumääristä ja ajoista. Kerättyjen valvontatietojen ja keskeytysten yksikköhintojen perusteella lasketaan tarkasteluvuodelle keskeytyskustannukset.

Keskeytykset ovat jaoteltu tyyppinsä mukaan odottamattomiin ja suunniteltuihin keskeytyksiin, sekä aika- ja pikajälleenkytkentöihin. Eri keskeytysten syntytavoilla on eri arvo asiakkaalle. Tämän takia keskeytyksille on laadittu taulukon 2 mukaiset yksikköhinnat vuoden 2005 rahanarvossa. Tarkasteluvuoden rahanarvoon yksikköhinnat korjataan kertomalla kuluttajahintaindeksillä. Yksikköhinnat ovat itsessään riippumattomia vuosittaisesta vaihtelusta ja ne kerrotaan yhtiön vuosienergioilla painotetuilla keskeytysajoilla, sekä yhtiön siirretyllä vuotuisella energialla tuntia kohden.

Taulukko 2. *Keskeytyskustannusten yksikköhinta vuoden 2005 rahanarvossa. (Energiavirasto 2015)*

Yksikkö	Odottamaton	Suunniteltu	Aikajälleenkytkentä	Pikajälleenkytkentä
€/kW	1,1	0,5	1,1	0,55
€/kWh	11,0	6,8	-	-

Laatukannustimen vaikutusta esimerkiksi suurhäiriöihin kohtuullistetaan muiden kannustimien tavoin toteutuneen oikaistun tuloksen laskennassa huomioimalla kannustinsanktio tai -bonus, joka on enintään 15 % vuoden kohtuullisesta tuotosta. Kannustimen inflaatiokorjaus tehdään kuluttajahintaindeksillä vuoden 2005 rahanarvossa oleviin keskeytysten yksikköhintoihin. Korjaus tehdään yhtiöiden referenssitasoon ja toteutuneisiin keskeytyskustannuksiin. (Energiavirasto 2015)

Laatukannustimen KAH-luvut lasketaan teknisten tunnuslukujen keskeytystietojen perusteella. Vuosittaiset keskeytysluvut kerätään osana teknisiä tunnuslukuja. Näihin kuuluvut muun muassa eri jännitetasojen pikajälleenkytkennät, aikajälleenkytkennät, odottamattomat ja suunnitellut keskeytykset.

2.3.3 Tehostamiskannustin

Tehostamiskannustimen tarkoituksena on ohjata yhtiötä toimimaan kustannustehokkaasti. Kannustimen vaikutus perustuu referenssitasoon ja toteutuneiden kustannusten erotukseen. Itse kannustimen määrityksessä käytetään 6 eri osatekijää:

- yleinen tehostamistavoite
- yrityskohtaisen tehokkuuden mittaamisen muuttujat
- yrityksen tehostamistavoite
- yrityksen tehostamiskustannusten vertailutaso
- yrityksen toteutuneet tehostamiskustannukset
- tehostamiskannustin toteutuneen oikaistun tuloksen laskennassa.

Näistä ensimmäistä käytetään Energiaviraston valvonnassa myös korjaamaan uusien toimintatapojen ja mahdollisten lisääntyneiden velvollisuuksien lisäämää työmäärää nostamalla tai laskemalla tasoa ja on yhteinen perustaso kaikille verkonhaltijoille.

Viraston teettämän selvityksen (Kuosmanen et al. 2014) mukaan sopiva tehostamistaso kaikille verkkotoiminnoille olisi 2 % pitkän aikavälin tuottavuuskehityksen perusteella. Virasto soveltaa 4 ja 5 valvontajaksolla yleiselle tehostamistavoitteelle arvoa 0 % uusien toimintatapojen käyttöönoton seurauksena. Nyt käytettävä alhainen prosentti aiheutuu muutosten tarkkojen kustannusten luotettavan arvioinnin vaikeudesta. Yleisen tehostamistavoitteen käyttö monopolialoilla on yleistä. Yrityskohtaista tehokkuutta mitataan panos-, tuotos- ja toimintaympäristömuuttujilla.

Taulukossa 3 on eriteltynä yrityskohtaista tehokkuutta mittaavat muuttujat. Samoja muuttujia käytetään myös yrityskohtaisen tehokkuustavoitteen määrittämiseen.

Taulukko 3. *Yrityskohtaista tehokkuutta mittaavat muuttujat.*

Panosmuuttujat (€)	**Tuotosmuuttujat (yksikkö)**	**Toimintaympäristö-muuttujat (-)**
• Kontrolloitavissa olevat operatiiviset kustannukset, KOPEX • Sähköverkon jälleenhankinta arvo	• Siirretyn energian määrä (GWh) • Sähköverkon kokonaispituus (km) • Käyttöpaikkamäärä (kpl) • Keskeytyskustannukset, KAH (€)	• Liittymät (kpl) / käyttöpaikat (kpl)

Keskeytyskustannuksia pidetään ei toivottuna tuotosmuuttujana ja toiminnan sivutuotteena ne eivät ole välttämättömiä yritykselle. Toimintaympäristömuuttuja kuvaa kuinka suuri osuus liittymistä on haja-asutusalueella ja kuinka paljon taajamissa tai kaupungeissa. (Energiavirasto 2015)

Tehostamiskannustimen tunnusluvut määritetään verkon rakennetiedoista ja teknisistä tunnusluvuista. Taulukon 3 muuttujista sähköverkon jälleenhankinta-arvo määritetään rakennetiedoista. Tuotosmuuttujat ja toimintaympäristömuuttujat saadaan teknisistä tunnusluvuista, toimintaympäristömuuttujan muodostuessa käyttöpaikka- ja liittymämääristä.

2.3.4 Toimitusvarmuuskannustin

Vuonna 2013 säädettiin myrsky- ja lumikuormien aiheuttamien pitkien sähkökatkojen takia toimitusvarmuutta velvoittava lisäys sähkömarkkinalakiin (hallituksen esitys 2013; Sähkömarkkinalaki 2013), joka velvoittaa yhtiöitä vähentämään myrskyjen ja lumikuormien vaikutusta sähkön keskeytyksiin haja-asutusalueella 6 tuntiin ja niiden ulkopuolella 36 tuntiin vuoteen 2028 mennessä. Yhtiöt ovat voineet hakea virastolta lisäaikaa vaatimusten täyttämiseen ja aikaa on myönnetty osalle vuoteen 2036 asti (Sähkömarkkinalaki 2013).

Toimitusvarmuuskannustimen tarkoitus on varmistaa kustannustehokkaasti sähkömarkkinalain mukaiset toimitusvarmuusvaatimukset kaikille yhtiöille. Yhtiö voi hakea kannustinta perustellusti, mikäli toimitusvarmuusvaatimuksiin ei voida päästä yhtiön normaalien kunnossapitotoimien ja korvausinvestointien kautta. Kannustin jakautuu kahteen osaan: ennenaikaisiin korvausinvestointeihin, sekä kunnossapito- ja varautumistoimenpiteisiin.

Verkonhaltijan tulee rakennetietojen yhteydessä jättää myös selvitys ja perustelut haettaessa toimitusvarmuuskannustinta. Selvityksen perusteella virasto hyväksyy tai hylkää haetut kustannukset. Toimitusvarmuuskannustimeen ei huomioida suurjännitteisiä jakeluverkonhaltijoita.

Hankkeet käsitellään tapauskohtaisesti ja mikäli korvausinvestoinnit kuuluvat verkonhaltijan normaalin kehittämisvelvollisuuden piiriin ei niitä huomioida kannustimessa. Korvausinvestointeihin hyväksytään vain 0,4 kV ja 20 kV ilmajohdot ja näiden pylväsmuuntamot, sekä 20 kV ilmajohtoverkon erottimet ja katkaisijat. Toimitusvarmuuteen hyväksyttyjä investointeja ei hyväksytä muihin kannustimiin. (Energiavirasto 2015)

Toimitusvarmuuskannustimen alaskirjaukset toimitetaan rakennetietojen yhteydessä ylimääräisinä liitteinä. Ne eivät ole osa kerättäviä verkonrakennetietoja itsessään, eikä niitä tulla tämän työn osalta erikseen käyttämään. Toimitusvarmuuskannustimen vaikutus on kuitenkin osa kohtuullisen hinnoittelun laskelmia ja siten myös tämän työn aineistossa.

3. DATAKLUSTEROINTI

Klusterointi on luokittelua ja kaikki luokittelu voidaan jakaa joko valvottuun tai valvomattomaan luokitteluun. Valvotussa luokittelussa äärettömän monta uutta havaintoa lisätään ennalta määriteltyjen ominaisuuksien tai funktioiden perusteella äärellisiin luokkiin ja valvomattomassa äärelliset havainnot luokitellaan äärellisesti ennalta määrättömiin luokkiin. Luokittelu itsessään on vanha keksintö. Kreikkalainen filosofi Plato esitti noin 400 BC luokittelun konseptin, jota Aristoteles noin 50 vuotta myöhemmin itse pohti (Masulli & Rovetta 2015). Klusteroinnin tavoitteena on luokitella dataa vähällä tai olemattomalla lähtötiedoilla aineistosta (Xu & Wunch 2009). Klusterointia käytetään aineistojen parempaan ymmärtämiseen, kuten poikkeuksien ja keskeisten piirteiden etsimiseen (Celebi et al. 2012). Matemaattinen teoria on menetelmäkohtaista, mutta yhteistä näillä on arvopisteiden vertailu toisiinsa välillisesti tai välittömästi. Luokkien eli klusterien määritelmiä on olemassa useita ja esimerkiksi Xu & Wunch (2009) mukaan klusterin sisällä olevat havainnot ovat keskenään samankaltaisia ja muiden klustereiden ominaisuuksiin nähden erilaisia. Havainnon muodostuessa useammasta kuin yhdestä muuttujasta syntyy havaintovektori, jonka vektorialkiot kuvaavat eri muuttujia.

Klusteroinnin yhteydessä saatetaan käyttää isojen ja moniulotteisten tietojen muokkaamisessa datan faktorointia. Faktoroinnilla tarkoitetaan ulottuvuuksien määrän tai tietomäärän vähentämistä hukkaamatta kuitenkaan samassa suhteessa kokonaisuuden antamaa informaatiota. Moniulotteisella datalla tarkoitetaan yleensä tilannetta, jossa havaintojen lukumäärä on pienempi kuin käytettävien muuttujien (Xu & Wunch 2009, s 237). Klusterointi löytää klustereita myös aineistoista, joissa luonnollista ryhmittelyä ei esiinny. Tämän takia aineisto tulisi etukäteen arvioida klusteroinnin hyödyllisyyden suhteen. (Jain 2010)

Klusterointi voidaan jakaa kovaan ja pehmeään klusterointiin. Pehmeästä klusteroinnista käytetään myös englanninkielistä termiä *fuzzy*. Pehmeässä klusteroinnissa datapisteille annetaan numeerinen painoarvo, jonka perusteella datapiste liitetään yhteen tai useampaan klusteriin. Kovassa klusteroinnissa jokainen datapiste kuuluu vain yhteen klusteriin. (Xu & Wunch 2009, s 84)

Menetelmät voidaan jakaa myös usealla muulla tavalla, kuten esimerkiksi hierarkkiseen ja osittelevaan. Hierarkkisissa menetelmissä klusterit voidaan muodostaa asettamalla kaikki arvopisteet omiksi klustereikseen ja sen jälkeen yhdistelemällä samankaltaisia klustereita tai sitten asettamalla kaikki arvopisteet yhdeksi klusteriksi ja jakamalla sitä iteroivasti pienemmiksi klustereiksi. Osittelevassa klusteroinnissa klusterit löydetään samanaikaisesti (Jain 2010). Osittelevissa menetelmissä tarvitaan lähtötietoina klusterien lukumäärä, joka voi olla vaikea päätellä etukäteen. Mikäli klusterointimenetelmä ei vaadi

klusterien lukumäärää lähtötietona, on niillä yleensä muita vaikeasti määriteltäviä parametreja. (Mutanen 2018)

Erilaisia klusterointimenetelmiä on useita ja niiden soveltuvuus ratkaistavaan tehtävään vaihtelee, eikä täysin yleispätevää menetelmää ole. Esiteltävistä menetelmistä k-means on vanhimpia ja käytetyimpiä. Kyseisestä menetelmästä on lukuisia muunnelmia ja parannuksia, joista yksi on k-medoids.

3.1 K-means ja k-medoids

K-means on käytetyimpiä ja yksinkertaisempia klusterointimenetelmiä. K-means kuuluu osittteleviin klusterointimenetelmiin ja sitä on käytetty jo 1950-luvulla (Celebi et al. 2012). Perinteinen k-means on kova menetelmä, jonka ajatuksena on minimoida iteroimalla arvopisteiden etäisyydet lähimpään klusteriin. Etäisyyden mittaukseen voidaan käyttää eri menetelmiä. Tästä menetelmästä on useita muunnoksia ja esimerkiksi pehmeä variaatio on myös olemassa. K-means sisältyy sisäänrakennettuna Matlabin kirjastoihin. K-meansillä pyritään minimoimaan tavoitefunktiota, joka yleensä on kaavan 1 mukainen virheen neliösumma:

$$\sum_{j=1}^{K} \sum_{t \in j} (\bar{p}_t - \bar{q}_j)^2, \tag{1}$$

Missä klusterin indeksi on *j*, klusterien kokonaismäärä *K,* yksittäinen klusterin jäsen on *t* ja $\bar{p}_t$ on jäsenen *t* muuttujien muodostama vektori, vastaavasti $\bar{q}_j$ on sen klusterin ominaisvektori, johon $\bar{p}_t$ kuuluu. (Celebi et al. 2012)

Kaavaa 1 yleisesti käytetään osittelevien klusterointialgoritmien tavoitefunktiona ja englannin kielisessä kirjallisuudessa siitä käytetään nimitystä SSE (Sum of Squared Error) (Celebi et al. 2012). Alla on k-means klusterointimenetelmän yleinen toimintaperiaate (Steinley 2006):

1. Jokaiselle klusterille määritetään aloitusvektori, joka sisältää jokaiselle muuttujalle kyseisen klusterin ominaisarvon.
2. Jokaisen havainnon vektorietäisyys mitataan jollain etäisyyden mittaamismenetelmällä jokaisen klusterin vektoriin ja asetetaan havainto pienimmän etäisyyden saamaan klusteriin.
3. Lasketaan eli päivitetään jokaiselle klusterille uudet ominaisvektorit, laskemalla klusterin havaintojen vektorien keskiarvo.
4. Toistetaan 2. ja 3. kohtaa, kunnes lopetuskriteeri täyttyy.

K-meansissä klusterien lukumäärä on tiedettävä alussa. Lukumäärän määrittelee käytetty aineisto ja klusteroinnin taustalla oleva tehtävä. Klusterien lukumäärälle voidaan asettaa

teoreettinen maksimi ja minimi, jotka ovat triviaaleja. Teoriassa pienin k-arvo on 1, jolloin kaikki havaintovektorit ovat samassa klusterissa. Teoreettinen maksimi on havaintovektorien lukumäärä, jolloin jokainen havaintovektori on omana klusterinaan. Klusterien määrä voidaan tietää etukäteen, mutta yleensä näin ei ole. Määrä voidaan määrittää myös erilaisilla määräanalyysi-menetelmillä.

K-means menetelmässä klusterointi aloitetaan kohdan 1 mukaisesti päättämällä klusterien ominaisarvot jokaiselle muuttujalle. Muuttujien ominaisarvot eli vektorialkiot muodostavat yhdessä aloitusvektorin. K-meansin heikkous on aloitusvektori, sillä eri klusterien alkuarvoilla voidaan saada eri tuloksia. Aloitusvektorit voidaan asettaa manuaalisesti käsin tai luoda algoritmien avulla. Käytetyimpiin tapoihin kuuluu aloitusvektorien luonti satunnaisalustuksella. Tällöin aloitusvektorin alkiot luodaan satunnaisesti perustuen vastaavien vektorialkioiden pienimpien ja suurimpien arvojen luomaan arvoväliin.

Menetelmän tuloksena saadaan lokaali minimi, joka saattaa olla globaali minimi. Globaalin minimin löytämiseksi on k-meansillä menetelmä ajettava kaikilla eri lähtövektoreilla, mikä ei ole käytännössä järkevää toteuttaa. Mahdollisimman montaa satunnaista lähtövektoria käyttämällä voidaan päästä globaaliin minimiin (Jain 2010). Tämä on käytännössä epävarma tapa, sillä keskikokoisella aineistolla tulisi sattumanvaraisia lähtöarvauksia laskea tuhansia ja liian pienellä määrällä lähtöarvauksia lopullinen tulos saattaa edelleen olla lokaali minimi (Steinley 2006, s6). Asettaessa kiinteät aloitusvektorit, eivät tulokset muutu ajokertojen välillä. Kiinteillä esiasetetuilla aloitusvektoreilla tulisi käytön kannalta olla perusteltu syy. Syynä voi olla klusteroinnin tarkoitus etsiä esimerkiksi klusteroitavasta aineistosta tiedettyjä ominaisuuksia tai yhteyksiä ulkoisiin lähteisiin. Tämä ei kuitenkaan takaa globaalia minimiä, jos aloitusvektorien määrittely on klusteroinnin kannalta virheellinen.

Matlabissa kohtaa 2 voidaan muuttaa niin, että jokainen havaintovektori testataan vaihtamalla havaintovektori jokaiseen klusteriin ja laskemalla klusterin keskipistevektori uudestaan (MathWorks inc. 2019a). Kokonaisvirhesumman laskiessa vaihto pidetään ja muutoin vaihdetaan takaisin. Tällöin tulokseksi saadaan varmasti kyseisen aloitusvektorin lokaali minimi.

Aloitusvektorien asettamisen jälkeen kohdassa 2 jokaista havaintovektoria verrataan klusterien keskipistevektoreihin. Riippuen minkälaista variaatiota k-meansistä käytetään, vaihtelevat vertailumenetelmä ja riippuen aineistosta, mahdollisesti myös tulokset. Tavallisesti vertailussa käytetään yksinkertaisuuden vuoksi Euklidinista etäisyyttä, joka on lyhin etäisyys kahden pisteen välillä n-ulotteisessa avaruudessa. Tässä työssä käytetään k-means -menetelmässä vertailemiseen Euklidinista etäisyyttä kaavan 2 mukaisesti (Korenius et al. 2007):

$$d(\bar{p}, \bar{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}, \quad (2)$$

Missä $\bar{p}$ ja $\bar{q}$ ovat vektoreita, vektorin alkio on n ulottuvuudessa p_n. Vastaavasti q_n on tässä klusterin keskipistevektorin vektorialkio vastaavassa ulottuvuudessa. Vektorin $\bar{p}$ lyhin etäisyys klusterin $\bar{q}$ keskipistevektorista on $d(\bar{p}, \bar{q})$.

Käytettäessä euklidinista etäisyyttä k-means luo klusterit pallomaisiksi. (Jain 2010). Tavallisessa euklidisessa etäisyydessä kaavan 2 mukaisesti lauseke on neliöjuuressa, mutta yleisesti klusteroinnissa neliöjuuri jätetään pois laskentatehon parantamiseksi. Klusteroinnin vertaillessa havaintovektoreita keskenään neliöjuuren puuttuminen ei muuta tuloksia. Matlab käyttää oletuksena juuretonta etäisyyttä k-means ja k-medoids menetelmissä (MathWorks inc. 2019a; MathWorks inc. 2019f). Tässä työssä aineisto on suhteellisen pieni klusteroinnin kannalta, eikä klusterointiin kuluva aika aiheuta toimenpiteitä. Neliöjuuretonta kaavaa 2 käytetään jokaiselle havaintovektorille, jokaisen klusterin suhteen. Tämä tarkoittaa, että lasketaan vektorin $\bar{p}$ etäisyys kaikista klustereista 1…j, jonka perusteella asetetaan arvopiste pienimmän tuloksen saaneeseen klusteriin.

Kun kaikki havaintovektorit 1…t on lisätty johonkin klusteriin, päivitetään klusterin keskipistevektori kohdan 3 mukaisesti. Uusi keskipiste voidaan laskea esimerkiksi aritmeettisella keskiarvolla kyseisen päivitettävän klusterin sen hetkisistä havainnoista. Keskiarvo voi olla myös painotettu.

Keskipistevektorin päivittämisen jälkeen verrataan uudestaan jokaista havaintovektoria päivitettyihin keskipistevektoreihin ja asetetaan havaintovektori lähimpään klusteriin. Tämän jälkeen jatketaan uudelleen kolmanteen vaiheeseen. Toista ja kolmatta vaihetta voidaan jatkaa useita kierroksia haluttuun lopetuskriteeriin saakka. Raja-arvoksi voidaan asettaa esimerkiksi haluttu iterointikierrosmäärä, jonka jälkeen klusterointi on valmis tai siihen saakka, kunnes kahden peräkkäisen iteraatiokierroksen päivitetyt keskipistevektorit eivät muutu. (Olszewski 2012)

K-medoids toimii samankaltaisesti k-meansin kanssa. Näkyvin ero näiden kahden menetelmän lopputuloksien välillä syntyy klusterin keskipistevektorin määrityksessä. K-means klusterin keskipistevektori voidaan olettaa olevan klusterin havaintovektoreiden esimerkiksi keskiarvo, kun taas k-medoids menetelmän kohdalla keskipistevektoriksi valikoituu klusterin havaintovektori, jonka etäisyys klusterin muiden havaintovektoreiden etäisyyksien keskiarvosta on pienimmillään. (Xu & Wunch 2009 s,73)

K-meansin tavoin k-medoids voi käyttää laajasti eri etäisyyden mittaamismenetelmiä. Samat rajoitukset koskevat molempia menetelmiä etäisyyden määrittelemisessä. K-medoids menetelmässä iterointimenetelmän valintaan vaikuttaa aineiston koko. Menetelmiä on muun muassa PAM (Partitioining Around Medoids), jonka ero k-meansiin tapahtuu k-meansin ensimmäisen ja kolmannen vaiheen muutoksella. PAM:ssa klusterien keskipis-

tevektorit ovat aineiston havaintovektoreita ja keskipistevektoriksi valittua havaintoa kutsutaan myös medoidiksi. K-meansin vaiheessa 3 klusterille asetetaan sen omista alkioista medoidi, jolla klusterin virhesumma on pienimmillään. Tämän jälkeen vaihdetaan medoidi johonkin toiseen ei medoidiin, asetetaan havaintovektorit lähimpään klusteriin ja lasketaan virhesumma uudestaan. Mikäli virhesumma ei laskenut, vaihdetaan tulos takaisin. Menetelmä on tunnetusti tehokkain pienillä ja keskikokoisilla aineistoilla, mutta sen tehokkuus huomattavan suurilla aineistoilla huononee klusterointiin tarvittavan laskenta-ajan takia. (Park & Jun 2008).

3.2 Eriytymät ja etäisyyden mittausmenetelmät

Yleensä ennen klusterointia tiedoista poistetaan niin kutsutut eriytymät eli täysin poikkeavat arvopisteet, jotka ovat tavalla tai toisella virheellistä informaatiota. (Hautamäki et al. 2005). Virhesummaa minimoivissa menetelmissä eriytymät saattavat jakaa isoja klustereita virheellisesti pienempiin klustereihin. (Xu & Wunch 2009, s 64) Pienien aineistojen kohdalla klusteroinnissa eriytymien vaikutus tuloksiin on suurempi. Havaintovektorin jäädessä yksittäiseksi klusterikseen arvioidaan, onko kyseinen klusterin jäsen eriytymä vai aineistoon kuuluva muista poikkeava havaintovektori.

Eriytymä voidaan käsitellä muutamalla eri tavalla. Yleisin tapa on poistaa ne lähtöaineistosta, mutta osa klusterointimenetelmistä pystyy havaitsemaan eriytymät klusteroinnin yhteydessä. Havaitsemiseen vaikuttaa menetelmän toimintaperiaate. Esimerkiksi DBSCAN perustaa klusterin niistä havaintovektoreista, jotka ovat tarpeeksi lähekkäin. Tällöin etukäteen on määritelty montako jäsentä tarvitaan klusteriin ja kuinka lähekkäin niiden tulee olla toisiinsa nähden. Yksittäiset eriytymät eivät siis DBSCAN:ssa vaikuta lopputuloksiin. Myös eriytymille herkillä menetelmillä voi olla variaatioita, joilla eriytymiä voidaan havaita ja poistaa. (Hautamäki et al. 2005)

K-means klusteroinnissa voidaan käyttää etäisyyden mittaamiseen useita eri menetelmiä. Kuten myös klusteroinnissa käytettävän menetelmän valinnan suhteen, myös etäisyyteen käytettävän menetelmän valinta tulisi perustua käytettävään aineistoon. Klusteroinnissa käytettäviä etäisyyksiä ovat euklidisen etäisyyden lisäksi muun muassa Kosini ja Manhattan. Edellä mainittuja etäisyyksiä voidaan käyttää myös Matlabin kirjastoista (MathWorks inc. 2019a). Kosini etäisyyttä käytetään muun muassa tekstiaineistojen klusteroinnissa (Steinbach et al. 2004).

K-medoids toimii paremmin kuin k-means eriytymiä sisältävissä aineistoissa. Keskipistevektorin muodostuessa eriytymän ja jonkin toisen jäsenen välille, k-meansissa vektori muodostuu näiden välille. K-medoidsissa keskipistevektori voi muodostua eriytymään tai eriytymää lähimpään olevaan jäseneen. K-meansissa keskipistevektorin synnyttyä näiden jäsenten välille mahdollisten muiden havaintovektoreiden etäisyys klusterin keskipis-

teestä saattaa kasvaa ja klusterointi ei välttämättä löydä oikeita jäseniä. Eriytymien vaikutus k-medoidsissa on siis pienempi johtuen edellä mainitusta ominaisuudesta (Manjoro et al. 2016**)**.

3.3 Moniulotteisen datan käsittely ja ongelmat

Moniulotteinen aineisto aiheuttaa vaikeuksia osalle klusterointimenetelmistä. Yleisesti ulottuvuuksien määrän kasvaessa myös klusteroinnin tarkkuus huononee. Tämä johtuu satunnaisuudesta, joka pienentää eroja eri klusterien kesken, jolloin klusterit sekoittuvat keskenään. Satunnaisuus saattaa pahimmassa tapauksessa luoda näennäisesti havaintovektoreita yhdistäviä tekijöitä. Käytännössä tämä saattaa johtaa virheellisiin tulkintoihin. (Xu & Wunch 2009, s 238) Moniuloitteisuudesta aiheutuneita ongelmia kutsutaan englanniksi ”*Curse Of Dimensionality*”. Ongelmia pyritään ratkaisemaan esimerkiksi käyttämällä moniulotteisiin aineistoihin toimivia menetelmiä, tai vähentämään ennen klusterointia ulottuvuuksien määrää. (Steinbach et al. 2004).

Aineistosta puuttuvat arvot täytyy huomioida klusteroinnissa. Tiedoista voidaan poistaa puuttuvia arvoja sisältäviä havaintovektoreita tai arvioida ja täyttää puuttuvat arvot muiden arvojen perusteella. Aineiston koon kasvaessa mahdollisuus puuttuviin tietoihin kasvaa. (Klawonn et al. 2015)

Vahvasti keskenään negatiivisesti tai positiivisesti korreloivat muuttujat saavat suuremman painoarvon, johtuen saman taustavaikuttajan vaikutuksesta useampaan kuin yhteen lukuun. Painotuksen määrä riippuu korreloivien muuttujien lukumäärästä ja etumerkistä. (Dormann et al. 2012).

Moniulotteista aineistoa voidaan klusteroida muutamalla eri periaatteella riippuen aineiston luonteesta. Aineisto voidaan yrittää klusteroida aliulottuvuuksissa (Assent 2015) tai käyttää menetelmiä, jotka soveltuvat moniulotteiseen dataan. Ulottuvuuksien vähentämiseen voidaan myös käyttää aineistoa projektoivia menetelmiä (Bouveyron & Brunet-Saumard 2012 s, 57). Yksinkertaisimmillaan ulottuvuuksien määrästä aiheutuvat ongelmat voidaan ratkaista poistamalla klusteroitavia ulottuvuuksia. Poistamalla ulottuvuuksia menetetään kuitenkin niiden sisältämä data ja huolimattomilla valinnoilla tulosten laatu huononee (Bouveyron & Brunet-Saumard 2012 s, 59).

Klusteroimalla eri suuruusluokan ulottuvuuksia tulee suuremman etäisyyden omaavalle muuttujalle isompi painoarvo käytettäessä euklidinista etäisyyttä (Bora & Gupta 2014). Tämän vaikutusta voidaan pienentää asettamalla jokaisen ulottuvuuden arvot samalle arvovälille. Normalisoinnin vaikutusta käsitellään vielä luvussa 4. Normalisointi voi joko parantaa tai huonontaa klusterointituloksia (Strehl et al. 2000). Strehl et al. tekemän tutkimuksen mukaan metriset etäisyydet kuten euklidinen etäisyys eivät sovellu käytettä-

väksi isojen ja harvojen ulottuvuuksien klusteroinnissa. Tutkimus oli toteutettu klusteroimalla tekstiaineistoa, eikä se välttämättä kuvaa hyvin tässä työssä käytettäviä numeerisia aineistoja.

3.4 Lukumäärän analyysimenetelmät ja korrelaation laskeminen

K-arvon valintaan käytetään usein erilaisia lukumäärän analyysimenetelmiä, jos k-arvoa ei tiedetä etukäteen ja käytetään osittelevia menetelmiä. Matlabin kirjastoissa on esimääriteltynä muun muassa: Davies-Bouldin -, Siluetti-, Gap- ja Calinski-Harabasz -analyysit. Lukumääräanalyysien avulla voidaan tietyn tyyppisestä aineistosta saada k-arvo suoraan. (Tibsirani et al. 2001) Analyysien käytettävyys riippuu kuitenkin käytettävän aineiston homogeenisuudesta ja tulokset vaihtelevat käytettäessä k-meansia. Matlabin sisäisissä kirjastoissa analyysit tuottavat samalla aineistolla ja parametreilla eri tuloksia ajokertojen kesken. Mutasen tutkimuksessa (Mutanen 2018) todetaan tulosten eriävän ajokertojen välillä. Tämä johtuu k-meansin tulosten riippuvuudesta aloitusvektoreihin. Siluetti- ja Calinski-Harabasz -analyysiin voidaan syöttää kiinnitetyt aloitusvektorit.

Virheen neliösumma -käyrä muodostetaan klusteroimalla aineistoa usealla k-arvolla, virheen neliösumman ollessa y-akselilla ja k-arvon x-akselilla. K-arvoa kasvatetaan 1 tai 2 lähtien yhdellä haluttuun määrään asti, muodostaen siten diskreetin virheen neliösummakäyrän. Mutasen tutkimuksen perusteella paras arvo voidaan löytää myös suoraan virheen neliösumma -käyrästä esimerkiksi visuaalisesti etsimällä käyrän polvikohta. Polvikohdassa virhesumma ei enää pienene huomattavasti klusterien lukumäärän kasvaessa. Matemaattisesti polvikohta on suurin toisen derivaatan arvo, mutta ongelmaksi matemaattisille malleille muodostuu virheen neliösumma -käyrän diskreettisyys. (Mutanen 2018)

Davies-Bouldin -indeksi kuvaa klusterin sisäisten keskimääräisten havaintovektoreiden etäisyyksien suhdetta klusterien välisiin etäisyyksiin. Mikäli indeksi on optimitilanteessa alhainen klusterin sisäisten etäisyyksin keskiarvot ovat pieniä ja klusterien väliset etäisyydet suuria. Matlabin kirjaston Davies-Bouldin -indeksi perustuu (MathWorks inc. 2019b) euklidiniseen etäisyyteen (kaava 2) klusterien välillä ja on erityistilanne kaavasta 3:

$$\bar{R} \equiv \frac{1}{N}\sum_{i=1}^{K} maksimi\ \frac{S_i + S_j}{d_{i,j}}\ (kun\ i \neq j), \tag{3}$$

jossa K on klusterien lukumäärä ja kaavan 3 mukaisesti $d_{i,j}$ kuvaa klusterien i ja j keskipistevektoreiden etäisyyttä ja kyseisten klustereiden jäsenten hajanaisuutta kuvaavat S_i ja S_j.

Mikäli klusterin *i* ja *j* sisällä etäisyydet pysyvät vakiona ja etäisyydet kyseisten klusterien välillä kasvaa, näiden samankaltaisuuksien suhde laskee. Davies-Bouldin indeksissä voidaan käyttää klustereiden etäisyyden mittaamiseen muitakin menetelmiä kuin euklinidista etäisyyttä. (Bouldin & Davies 1979)

Siluetti analyysi saa arvoja -1 ja 1 välillä. Mikäli siluetti arvo on lähellä arvoa -1 havaintovektori ei ole oikeassa klusterissa. Lähellä arvoa 1 oleva havaintovektorin arvo on oikeassa klusterissa. Arvoa 0,8 voidaan pitää jo todella hyvänä tuloksena, vaikka yleisesti määriteltyä raja-arvoa hyvälle tulokselle ei ole. Määritelmän mukaan:

$$SI = \frac{d_{min,t} - d_{ka,t}}{maksimi(d_{min,t}, d_{ka,t})}, \tag{4}$$

jossa $d_{min,t}$ on pienin etäisyyden keskiarvo havaintovektorille $\bar{t}$ verrattuna muihin klustereihin kuin omaansa, keskimääräinen etäisyys saman klusterin muihin havaintoihin on $d_{ka,t}$.

Rousseleewin mukaan pelkkää huonoa siluetti arvoa tarkastelemalla ei tulisi k-arvon valintaa tehdä, vaan aineistosta tulisi poistaa eriytymät ja kokeilla muuttuvatko tulokset. (Rousseleew, 1987)

Gap-analyysi etsii suurinta arvoa vertailemalla aineistosta tehtyä käyrää odotuskäyrään kaavan 5 mukaisesti:

$$Gap(k) = E_n^*(\log(W_k)) - \log(W_k), \tag{5}$$

jossa k on k-arvo, $E_n^*(\log(W_k))$ on odotusarvo, joka perustuu Monte Carlo -näytteistykseen referenssijakaumasta.

Optimaalinen klustereiden lukumäärä saadaan k-arvolla, jolla kaavan 5 toinen tekijä on kauimpana odotuskäyrästä. Kaavassa 5 esitetty termi W_k määritetään

$$W_k = \sum_{j=1}^{k} \frac{1}{2n_r} \sum_{t,s \in k} d_{t,s}, \tag{6}$$

jossa $d_{t,s}$ on havaintovektoreiden $\bar{t}$ ja $\bar{s}$ keskinäinen etäisyys, ensimmäinen summa lausekkeen n_r on näytekoko, testattavaa k-arvoa merkitään k:lla.

Analyysiä voidaan käyttää kaikilla etäisyyksien mittausmenetelmillä. Aineiston ollessa selvästi jakautunut k klusteriin referenssijakauma laskee hitaammin kuin sen vertailukohta ja Gap-arvon tulisi silloin olla suurimmillaan. (Tibshiran et al. 2001)

Calinski-Harabasz -kriteeri soveltuu k-arvon määrittelemiseen parhaiten käytettäessä k-means klusterointiin euklidinista etäisyyttä. Kriteeri määritellään kaavassa 7 esitetyllä tavalla:

$$CH = \frac{SS_b}{SS_w} \times \frac{M-k}{k-1}, \tag{7}$$

jossa M on havaintovektoreiden lukumäärä. Klusterien välistä varianssia kuvataan SS_b ja vastaavasti klusterin jäsenten välistä varianssia SS_w.

Kaavan 7 mukaista klusterien välistä varianssia SS_b kuvataan kaavan 9 mukaan:

$$SS_b = \sum_{j=1}^{k} n_j \, d(\overline{q_j}, \overline{m}), \tag{8}$$

jossa n_j on klusterin j jäsenten määrä, klusterin j keskipistevektori on $\overline{q_j}$ ja $\overline{m}$ on keskiarvo datasta. Termi $d(\overline{q_j}, \overline{m})$ on kyseisten vektoreiden euklidinen etäisyys toisistaan.

Kaavan 8 mukaisesti koko aineistolle saadaan kaikkien havaintovektoreiden klusterin sisäinen varianssi SS_w kaavan 9 mukaisesti.

$$SS_w = \sum_{j=1}^{k} \sum_{\bar{t} \in k} d(\bar{t}\ , \overline{q_j})\ , \tag{9}$$

Jossa $\bar{t}$ on klusterin j havaintovektori.

Kaavan 8 mukaisesti myös kaavassa 9 on euklidinen etäisyys. Mikäli klusterien jäsenet ovat lähekkäin klustereissaan SS_w saa pieniä arvoja ja klustereiden olleessa selvästi erillään SS_b saa suuria arvoja. CH-indeksin saadessa suurimman arvonsa klusterit ovat parhaiten erillään ja klustereiden jäsenet lähekkäin toisiaan. (MathWorks inc. 2019d)

4. KLUSTEROINTITYÖKALU JA SEN RAKENNE

Työkalu suunnitellaan tehtävänannon mukaisesti mahdollisimman yleiskäyttöiseksi ja tämän jälkeen sitä testataan tutkimalla vuoden 2017 teknisiä tunnuslukuja ja kohtuullisen hinnoittelun laskelmia. Klusterointimenetelmien testaamiseen ja luvussa 5 tehtävään aineiston tarkasteluun tarvitaan Matlab-koodi, joka lukee ja valmistelee lähtöaineiston, klusteroi sen ja palauttaa tulokset luettavaan muotoon. Koodin tulee olla myös tarpeeksi mukautuva, jotta klusterointimenetelmää voidaan muuttaa tai lisätä rinnakkaistarkasteluja. Toimiva työkalu tulee rakentaa huomioimaan aineistojen erilainen muotoilu mahdollisimman hyvin. Hyvän ohjelmointitavan mukaisesti itse koodin dokumentointi tulee olla asianmukainen ja riittävä. Aineiston tulee olla muotoiltu siten että, yhtiöt tai jaoteltavat ovat riveittäin ja halutut muuttujat ovat sarakkeittain. Lisäksi joissain tarkasteluissa, kuten takautuvassa fuusiossa ensimmäinen sarake on varattu yhtiötunnukselle ja vasta toinen yhtiön nimelle.

Matlab -koodi on kirjoitettu varmuuskopioinnin takia eri Matlab-tiedostoihin. Kaikkien käytettyjen koodien rakenne on seuraavanlainen:

1. Luetaan lähtöaineisto ja ladataan se Matlabiin.
2. Tehdään mahdolliset muokkaukset aineistoon Matlabissa. Muun muassa normalisointi halutulle välille [0,1] tai [-1,1].
3. Klusterointi ja sen parametrien asettaminen.
4. Tulosten ulkoasun asettaminen ja tallentaminen tai/ja esittäminen.

Työkalu klusteroi vuoden kerrallaan ja klusteroinnit ovat toisistaan riippumattomia. Parametrit ja aineistot voidaan asettaa eriäviksi vuoden perusteella, mikäli on tarvetta. Tässä työssä parametrit pidetään samoina vuosien välillä ja k-arvo pääsääntöisesti samana myös aineistojen välillä.

4.1 Työkalun sisäänmenot ja ulostulot

Työkalun sisäänmenona on Excel-tiedoston taulukko, jonka rivit muodostavat havainnot ja sarakkeet käytettävät muuttujat. Yhdessä näistä muodostuvat havaintovektorit. Tässä työssä havaintovektori alkaa yhtiön nimi -sarakkeen jälkeen ja sitä ennen on erilaisia täydentäviä tietoja, kuten yhtiötunnuksia. Mikäli täydentäviä tietoja on, tarvitsee työkalun asetuksiin kirjoittaa tämä, muuten indeksit eivät täsmää. Matlab suorittaa klusteroinnin olettamalla ensimmäisen sarakkeen muodostavan luokiteltavat havainnot, eli tässä työssä jakeluverkkoyhtiöt ja sen jälkeen klusterointiin käytettävät muuttujat, joita tässä työssä ovat muun muassa tekniset tunnusluvut. Työkalun ja klusteroinnin kannalta ei ole väliä

mistä sarakkeet muodostuvat, kunhan muuttujat ovat numeerisia. Yhtiöiden tilalla voitaisiin käyttää esimerkiksi voimalaitosrekisterissä olevia voimalaitoksia ja muuttujina niiden nimellistehoa.

Työkalun ulostulona on Matlabin taulukkoja ja matriiseja, joissa on:

- yhtiön nimi
- yhtiön klusterin indeksi
- yhtiön etäisyys muiden ja oman klusterin keskipistevektorista
- jokaiselle klusterille niiden jäsenten yhteenlaskettu virheen neliösumma

Usean vuoden klusteroinneissa tiedot tallennetaan jokaiselle vuodelle erikseen. Tallennetuista tiedoista voidaan kerätä työkalun funktiolla yhtiöiden klusterin indeksi eri vuosilta samaan taulukkoon. Samalla periaatteella voidaan myös muut tiedot kerätä, mutta erillistä koodia ei sitä varten työkaluun sisällytetty.

Klusteroinnin jälkeen halutut tiedot kerätään yhteen ja haluttaessa tallennetaan Excel-tiedostoon. K-arvon määrittelemiseen käytettävät määrän analyysimenetelmät ja virheen neliösumman käyrät piirretään haluttaessa myös. Tiedostoon kirjoituksen ja analyysien piirtämisen valintaan käytetään työkalussa globaaleja muuttujia ja *if*-lauseita. Tietoja voidaan tarkastella suoraan Matlabissa ja myös sieltä taulukoita tai tietoja viedä Windowsin kopioi ja liitä toiminnoilla esimerkiksi Exceliin.

4.2 Matlab ja sen hyödynnettävät ominaisuudet

Matlabin sisältämässä Machine learning -toolboxissa on mukana muun muassa k-means ja k-medoids. Lisäksi Matlabin toolboxit sisältävät analyysityökaluja, kuten määräanalyysejä. Klusterointimenetelmien matemaattisen luonteen pohjalta voidaan rakentaa funktiot Matlabilla myös ilman valmiita kirjastoja ja samaten klusterointia voidaan tehdä eri laskentaohjelmilla ja ohjelmointikielillä. Matlabin k-means klusterointi suorittaa euklidinisen etäisyyden sisäisesti ja se on oletus etäisyyden mittausmenetelmä (MathWorks inc. 2019a).

Matlabin sisäisessä funktiossa on mahdollisuus ajaa klusterointi useilla eri lähtöarvauksella eli replikaatilla ja näistä pienimmän tavoitefunktion saanut tulos valitaan lopulliseksi tulokseksi. (MathWorks inc. 2019a; MathWorks inc. 2019f) Tämä pienentää lopullisen tuloksen virheen neliösummaa ja minimoi alkuarvauksen vaikutuksen tuloksiin (Jain 2010). Aineistossa ilman esiasetettuja aloitusvektoreita tehtävissä klusteroinneissa käytettiin 500 replikaattia. Klusteroidessa ilman asetettuja aloitusvektoreita Matlab määrittelee aloitusvektorit oletuksena k-means ++ menetelmällä. Tämä menetelmä suosii selviä erillään olevia arvopisteitä ja ei siten sovi eriytymiä sisältävään aineistoon (Mutanen 2018).

Klusterointiparametrien vaikutusten vertailemiseksi k-means- ja k-medoids -menetelmillä oli kiinnitettävä aloitusvektori, jotta sen aiheuttama variaatio ei vaikuttaisi tuloksiin. Tätä varten Matlabiin kirjoitettiin koodi, joka jakaa k-arvon perusteella tasaväleihin klusterien aloitusvektorialkiot yhden muuttujan tapauksessa. Koodi huomioi muuttujien lukumäärän ja monistaa sen perusteella aloitusvektoriin alkioita luoden aloitusvektorit jokaiselle muuttujalle. Ensimmäisen klusterin aloitusvektori on pelkästään nollia sisältävä vektori. Nollavektori valittiin mukaan tunnuslukujen saadessa normalisoinnin jälkeen arvoja [0,1] ja tiettyjen tunnuslukujen fyysisen luonteen vuoksi. Tunnusluvuissa on esimerkiksi maksettuja vakiokorvauksia ja suurjännitteistä verkkoa, jota ei kaikilla ole, mutta jolla saattaa olla merkitystä tuloksiin. Myöhemmin klusteroidessa kohtuullisen hinnoittelun laskelmia todettiin kyseisen aloitusvektoriluonnin olevan käyttökelvoton kyseiseen aineistoon. Tämä johtuu aineistossa olevista negatiivisista arvoista.

Seuraavia Matlabin vähemmän tunnetumpia funktioita käytettiin apuna: kmeans, kmedoids, evalcluster ja corrcoef. Evalclusters -funktiolla pystytään ajamaan k-arvon määrittelemiseen käytettäviä menetelmiä. Funktioon pystytään asettamaan muutamia parametreja, kuten klusterointimenetelmä. Menetelmät eivät sisällä k-medoidsia, mutta sisältävät k-meansin. Corrcoef -funktio laskee korrelaatiokertoimen ja palauttaa korrelaatiomatriisin. Oletuksena se laskee käyttäen Pearsonin korrelaatiokertoimen jokaisen muuttujan suhteen. Matlab palauttaa klusteroinnissa havaintovektorin klusterin indeksin, sen etäisyyden kaikkiin keskipistevektoreihin, kaikkien klusterien keskipistevektorit ja klusterin alkioiden etäisyyksien summan omaan keskipistevektoriin. Viimeisintä voidaan hyödyntää esimerkiksi eriytymien sijoittamiseen tai rajatapausten havainnointiin. Matlabillä pystytään visualisoimaan ja luomaan taulukoita tuloksista ja lähtötiedoista. (MathWorks inc. 2019e)

4.3 Aineiston valmistelu

Käytetyt klusterointimenetelmät vaativat onnistuneen klusteroinnin toteuttamiseksi tietojen normalisointia. Ilman normalisointia tiedoissa korostuisivat isot absoluuttiset arvot ja esimerkiksi prosenttiluvut jäisivät triviaaleiksi klusteroinnin tuloksiin. Tällaisia arvoja käytetyissä teknisissä tunnusluvuissa ovat esimerkiksi vuodessa siirretty energia (GWh) ja maakaapelointiaste. Lisäksi useissa absoluuttisissa tunnusluvuissa taustatekijänä näkyy yhtiön koko. Esimerkiksi verkkopituus ja siirretty energia ovat suuremmissa yhtiöissä huomattavasti suurempia verrattuna pienempiin yhtiöihin. Tässä työssä aineisto normalisoidaan Matlabissa. Aineisto muuttuu kahdessa eri suunnassa. Vuosien välillä yhtiöiden määrät vaihtelevat ja muuttavat rivimääriä. Kerättyjen tunnuslukujen osalta tiedot vaihtelevat jaksoittain: 1996-2004, 2005-2012, 2013-2015 ja 2016 eteenpäin.

Nykyisin käytössä olevassa valvontatietojärjestelmässä ei ole saatavilla vuosien 1996-2004 tietoja. Tämän puutteen vuoksi aputyökaluna on käytetty erillistä MS Access-poh-

jaista tietokantaa. Tietokannassa on kerättynä valvonnassa käytetyt tekniset ja taloudelliset tunnusluvut vuosilta 1996-2017 ja sitä päivitetään tarpeen mukaan. Käyttämällä lähtöaineistona Accessia vähennetään tässä työssä mahdollisia inhimillisiä virheitä.

Lähtötiedot voidaan tarvittaessa korjata vertailukelpoisiksi vuosien välillä. Kerättyjen tietojen vastaavuus on tarkastettu ja tutkimuksessa käytettävien aineistojen erot vuosien kesken ovat tiedossa. Tiedot ovat korjattu Access -työkalun avulla. Yhtiöt voidaan takautuvasti fuusioida tiedoista vastaamaan vuoden 2017 tilannetta yhdistämällä aiemmin yksittäisinä yhtiöinä toimivat yhtiöt nykyisin toimivana yhtiönä. Yhtiöiden määrän pysyessä samana koko tarkasteluvälillä, fuusioista aiheutuneet muutokset vuosien välillä eivät aiheuta suuria muutoksia. Klusterointi on pääasiassa toteutettu vuositason tarkasteluina ja eri vuosien tiedot eivät vaikuta keskenään. Teknisille tunnusluvuille ja kohtuullisen hinnoittelun laskelmille tehtävät tarkemmat tarkastelut tehdään vuoden 2017 aineistoille.

Matlabin ohjelmointi tarkastaa yleisesti aineiston arvojen tyypin soveltuvuuden funktioon. Koodi keskeytetään löydettäessä epäsopiva muuttujatyyppi, kuten tyhjä arvo eli esimerkiksi NaN (engl. Not a Number). Excel -aineistossa on valvontatietojärjestelmän luomia otsikkosarakkeita, joiden yhtiöille antamat arvot ovat tyhjiä. Valvontatietojärjestelmän täyttö mahdollistaa myös tyhjän arvon jättämisen, jolloin useat yhtiöt ovat jättäneet nollan sijasta tyhjän arvon. Tämän työn aineisto on viranomaisvalvonnassa tarkastettu puuttuvien tietojen varalta ja siten tässä työssä käytettävässä aineistossa ei ole puuttuvia arvoja tai puuttuva arvo kuvaa nollaa. Matlab tarkastaa klusteroitaessa aineiston muiden kuin numeeristen arvojen varalta. NaN arvoja löytyessä klusterointityökalussa Matlab keskeyttää koodin ajamisen virheilmoituksella. Aineiston luonteen takia ei tyhjiä arvoja sisältäviä havaintovektoreita voida poistaa, eikä sille ole tarvetta. Aineisto tarkastetaan ennen klusterointia tyhjien solujen varalta ja löydettäessä asetetaan nollaksi. Mikäli valvontatietojärjestelmään on jätetty tyhjä tietue merkkaamaan nollaa, aineisto korjaantuu tällä tarkastelulla.

Keskeisille tunnusluvuille tehtiin vastaavat tunnukset, joiden avulla pystytään tunnistamaan keskenään eri vuosina toisiaan vastaavat tunnusluvut. Näitä hyödynnettiin tässä työssä, jotta voidaan valita pelkästään muuttumattomat tai halutut tunnusluvut klusteroinnin aineistoksi. Kerättyjen tietojen lukumäärä on kasvanut huomattavasti vuodesta 1997. Mikäli käytettäisiin vain niitä lukuja, joita on kerätty koko tarkastelujaksolta, olisi tunnuslukumäärät pieniä.

Yhtiöt fuusioitiin vuodesta 1997 asti vastaamaan vuoden 2017 yhtiöitä. Käytännössä tämä toteutetaan ajamalla Accessista kysely ja käyttämällä kyselyn tuloksia. Accessiin on tehty tietue, jolla pystytään yhdistämään yhtiöt toteutuneiden yritysfuusioiden mukaan. Klusteroinnit toteutetaan fuusiossakin vuosi kerrallaan ja tämän vuoksi tuloksiin vaikuttaa esimerkiksi kyseisen vuoden myrskyt ja niistä mahdollisesti aiheutuneet keskeytykset. Fuusiossa absoluuttiset tunnusluvut, kuten johtopituudet summattiin yhteen. Suhteellisten lukujen kohdalla yhtiöistä otettiin käyttöpaikkapainotettu keskiarvo.

4.3.1 Aineistojen normalisointi

Tässä työssä normalisoinnilla tarkoitetaan aineiston muuttujien arvojen muuttaminen vertailukelpoisiksi suhteessa toisiinsa. Aineiston muuttamiseksi välille [0,1] tai kohtuullisen hinnoittelun laskelmissa välille [-1,1] käytettiin absoluuttisten lukujen jakamiseen kahta eri menetelmää. maksiminormalisointia ja Euklidinista normalisointia.

Tässä työssä teknisisiin tunnuslukuihin käytetty menetelmä on maksimi normalisointi. Kyseisessä normalisoinnissa jokaiselle arvopisteelle etsitään kyseisen muuttujan isoin arvo ja jaetaan jokaisen yhtiön kyseinen muuttuja sillä. Tällöin absoluuttisten erojen ja normalisoitujen lukujen suhde pysyy lineaarisena. Normalisoinnin jälkeen tunnusluvun suurin arvo tulee olemaan 1 ja pienin 0 tai suurempi. Suurimman arvon ollessa nolla, jakajana toimii nolla itse. Tämä aiheutti ongelmia kohtuullisen hinnoittelun laskelmilla.

Kohtuullisen hinnoittelun laskelmissa negatiivisten ja nolla arvojen takia aloitusvektorien määrittely välille [0,1] huonontaa klusterointituloksia negatiivisten arvojen huomioidessa vain pienimmän keskipisteen omaavan klusterin, joka on normalisoinnin luonteen takia indeksiltään 1. Lisäksi aineistossa on muuttujia, joiden suurin arvo on nolla. Maksimilla jako -koodia oltaisiin voitu muokata huomiomaan poikkeustilanteita. Lukujen etumerkki kuvaa onko esimerkiksi toimitusvarmuuskannustimesta saatu sanktiota vai bonusta, jolloin etumerkki itsessään kuvaa tärkeää informaatiota ja negatiivisia lukuja ei voida vain muuttaa positiivisiksi. Tämän takia kohtuullisen hinnoittelun laskelmissa käytettiin PAM ja k-medoids -menetelmää ja normalisointi toteutettiin välille [-1,1] käyttäen euklidinista potenssinormalisointia kaavan 10 mukaisesti.

$$x_{i,norm} = \sqrt{\frac{x_i^2}{\sum_{i=1}^{n} x_i^2}}, \tag{10}$$

jossa n on arvojen lukumäärä ja i on muuttujan indeksi ja x_i on arvo indeksillä i

Normalisoinnin tapahtuessa yllä olevan kaavan mukaisesti hukataan tieto negatiivisista arvoista. Matlabissa on muodostettu etumerkin sisältävä matriisi ja kyseisen matriisin avulla palautetaan etumerkki. Mahdolliset arvot normalisoidussa matriisissa ovat välillä [-1 1].

Normalisoinnin takia tässä työssä myös Matlabin palauttamat keskipistevektorit ovat välillä [-1,1]. Tiedot palautetaan absoluuttisiin arvoihin keskipistevektoreista, jos halutaan tietoa todellisista arvoista. Tämä voidaan toteuttaa esimerkiksi ottamalla klusterien lähimpänä keskipistevektoria oleva havaintovektori ja käyttämällä sen absoluuttisia arvoja. K-medoidsissa keskipistevektorin absoluuttiset arvot ovat samat medoidi-havaintovektorin arvojen kanssa. Vaihtoehtoisesti klusterin absoluuttiset keskipistevektorin alkioiden arvot saadaan summaamalla tai keskiarvoistamalla klusterin havaintovektorin alkioiden absoluuttiset arvot.

4.3.2 Eriytymien käsittely

Klusteroinnin matemaattisen luonteen takia k-means pyrkii tekemään eriytymistä klustereita jo alhaisilla k-arvoilla. Tämän työn aineistojen kohdalla eriytymät on tunnistettava ulkoisten tietojen avulla. Aineistossa on useita yhtiöitä, joiden tunnusluvut ovat hyvin poikkeavia muista yhtiöistä ja hakeutuvat herkästi yksittäisiksi tai muutaman yhtiön klustereiksi. Eriytymät voivat olla tässä työssä esimerkiksi yhtiöitä, jotka ovat tyypiltään enemmän teollisuus- tai kiinteistöverkkoja kuin perinteisiä jakeluverkkoyhtiöitä. Käytettävässä aineistossa on vain yksi eriytymä, Karhu Voima Oy. Tunnusluvuiltaan esimerkiksi maantieteellisesti laajat ja pienet yhtiöt ovat tunnusluvuiltaan hyvin erilaisia ja aiheuttavat eriytimien kaltaista käyttäytymistä.

Tarkasteltaessa vuosittain tulee arvioida klusteroinnin syytä. Aineisto on sen mukainen, mitä jakeluverkkoyhtiöitä on sinä vuonna toiminut sähkönjakeluverkkoyhtiöinä. Tunnistettu entinen jakeluverkkoyhtiö Karhu Voima Oy on eriytymä verrattaessa aikaisempia vuosia vuoteen 2017. Yhtiöiden tietojen ollessa toisistaan riippumattomia, voidaan eriytymä poistaa. Myöhemmin puhuttaessa eriytymistä, tarkoitetaan havaintovektoreita, jotka luovat yksittäisiä klustereita. Näitä eriytymien kaltaisesti käyttäytyviä yhtiöitä ei haluta tiedoista kokonaan poistaa niiden ollessa aineistoon kuuluvia tärkeitä havaintovektoreita.

Tässä työssä eriytymiä ja niiden kaltaisia havaintovektoreita käsitellään toisen tai korkeamman asteen klusteroinneilla. Tällä tarkoitetaan klusteroitavien yhtiöiden valitsemista aikaisemman klusteroinnin perusteella. Aineistoon tehdään klusterointi ja tämän perusteella valitaan aineistosta ne yhtiöt, jotka ovat suurimmassa klusterissa ja klusteroidaan uudelleen kyseiset yhtiöt, mikäli yhtiömäärä on suurempi kuin valitun k-arvon. Tässä toisen asteen klusteroinnilla saadaan tasaväli aloitusvektorilla eroteltua suurin osa eriytymistä, ensimmäisellä asteella aloitusvektoreiden löytäessä eriytymät ja muodostaessa niistä yksittäisiä tai pieniä klustereita. Tässä työssä suuret klusterit jaettiin samalla k-arvolla, kuin edellisellä tasolla. Toisella asteella tapahtuva klusterointi klusteroi aineiston ilman outlierien vaikutusta tulokseen. Asteittaista klusterointia voitaisiin jatkaa seuraaville asteille eli kaikkiin k-arvoa suurempiin klustereihin.

Esimerkiksi k-arvolla 9 saadaan lopulta yhteensä 18 klusteria. Mikäli klusteroitaisiin käyttämällä tasavälialoitusvektoreita kasvattamalla klusterien lukumäärää ei aloitusvektorit välttämättä muodostuisi optimaalisiin paikkoihin ja eriytymien vaikutus näkyisi edelleen, verrattuna kahdessa osassa tehtävään klusterointiin. Tässä työssä yksittäisten alkioiden ja suuret yli 20 alkion klusterit eivät ole toivottuja jatko-analysoinnin kannalta. Tämä menettely mahdollistaa kiinteän aloitusvektorien asettamisen ja optimoinnin lokaaliin minimiin klusteroinnin perusteella.

4.3.3 Työn aineistojen ominaisuudet

Klusteroitaessa teknisiä tunnuslukuja absoluuttisilla arvoilla ja tarkastelemalla yhtä vuotta kerrallaan, havaittiin klusteroinnin perustuvan yhtiöiden kokoon. Tämä johtuu teknisten tunnuslukujen monimuotoisuudesta ja eri käyttötarkoituksista. Klusterointi-parametreillä ei ollut suurta merkitystä tässä tarkastelussa erojen kannalta. Luvut ovat pääosin todellisia arvoja, joita käytetään kyseisen yhtiön laskelmiin, eikä lukuja suhteuteta muihin yhtiöihin.

Koska tutkimuksen taustalla on tehtävänä löytää yhtiöiden suorituskykyjen eroja, eikä niiden kokoeroja, tiedot tulee normalisoida myös yhtiöiden koon mukaan. Yhtiöiden suhteuttamattomat tunnusluvut jaettiin omilla käyttöpaikkamäärillään. Tällöin yhtiöiden kokoerot tasoittuvat tunnusluvuista ja klusterointi ei korosta pelkästään yhtiön kokoa. Käyttöpaikkamäärän sijasta oltaisiin painotuksessa voitu käyttää esimerkiksi siirrettyä energiaa. Useiden tunnuslukujen korreloidessa yhtiön kokoon, ei haluttu lähteä karsimaan tunnuslukuja lukumäärällisesti. Tällöin oltaisiin menetetty suuri määrä tietoa.

Tutkimuksessa käytettävissä aineistossa on useita ulottuvuuksia ja ulottuvuuksien määrä aiheuttaa ongelmia luvun 3.3 mukaisesti. Klusterointituloksia pyritään tässä työssä parantamaan normalisoimalla aineisto ja poistamalla ulottuvuuksia. Ulottuvuuksien poistaminen perustuu aineistosta tehtyyn korrelaatiomatriisiin ja aineiston kuvaamiin fyysisiin ominaisuuksiin.

Teknisten tunnuslukujen kohdalla ulottuvuuksien aiheuttamaa ongelmaa lähestyttiin kahdella tavalla. Ensimmäinen tapa oli tutkia korrelaatioita. Tunnuslukujen korrelaatiosta ei päädytty haluttuun lopputulokseen. Matriisin luvuista vähän korreloivat olivat toisarvoisia, yksityiskohtaisia lukuja tai kaikilla yhtiöillä ei ollut arvoa kyseisessä luvussa, esimerkiksi suurjännitteisen jakeluverkon keskeytyksiä ei ole kaikilla yhtiöillä. Toisena tapana hyödynnettiin Virastossa tehtyä tunnuslukujen jaottelua. Jaottelulla tunnusluvut ovat kolmessa ryhmässä: Volyymia kuvaavat muuttujat- (A), panos- (B) ja toimitusvarmuusmuuttujat (C). Volyymia kuvaavia muuttujia, kuten siirrettyä energiaa, kutsutaan tästä eteenpäin volyymimuuttujina. Tarkempi tutkimus päätettiin toteuttaa kyseisen ryhmittelyn perusteella. Liitteessä A on esiteltynä myöhemminkin esille tuleva tunnusluvuista valikoitu aineisto, ”ABC-aineisto”, jossa on merkittynä myös jaotellut tunnusluvut ja käyttöpaikkapainotukset.

Kohtuullisen hinnoittelun laskelmien klusteroinnissa käytettiin viraston viranomaisvalvonnassa kerättyä aineistoa. Myös kohtuullisen hinnoittelun laskelmissa oli havaittavissa kokoeroista johtunutta klusteroitumista ja kohtuullisen hinnoittelun laskelmat painotettiin käyttöpaikkamäärillä. Painotus tehtiin kaikille muuttujille. Aineistoa päädyttiin käyttämään etsittäessä yhtäläisyyksiä fyysisiä ominaisuuksia kuvaavien teknisten tunnuslukujen ja taloudellista suoriutumista kuvaavien kohtuullisen hinnoittelun laskelmien välillä. Toisin kuin teknisten tunnuslukujen kohdalla kohtuullisella hinnoittelulla ei ole valmiiksi

yhtiöiden välisiä fuusioitumisista tehtyä tunnusta. Kohtuullisen hinnoittelun laskelmista tehtiin pienempi aineisto perustuen korrelaatiomatriisiin ja sen klusterointitulosta verrattiin koko aineistoon.

4.3.4 Korrelaatio

Ulottuvuuksien vähentäminen voidaan tehdä poistamalla yhtäläisesti korreloivia tunnuslukuja. Korrelointia varten Matlabistä löytyy sisäisenä funktiona *corrcoef* -funktio, joka kuvaa muuttujien riippuvuutta toisistaan korrelaatiomatriisin avulla. Korrelaatiomatriisi tehtiin teknisille tunnusluvuille ja kohtuullisen hinnoittelun laskelmille Pearsonin korrelaatiokertoimella. Matriisia tulkittiin visuaalisesti käyttämällä *heatmap* -toimintoa Excelissä. Euklidinisella normalisoinnilla ei ollut vaikutusta korrelaation tuloksiin suhteiden pysyessä lukujen välillä samana. Tämä tarkastettiin tekemällä normalisoitu ja normalisoimaton korrelaatiomatriisi ja vertailemalla tuloksia.

Korrelaatiomatriisista pystytään nopeasti havaitsemaan muuttujille korrelaatiot muihin muuttujiin. Jokaisella skalaariselle muuttujalle voidaan laskea korrelaatiokerroin eli kerroin paljonko muuttuja korreloi toiseen muuttujaan. Korrelaatiokertoimen arvot ovat arvovälillä [-1 1]. Mikäli arvo on -1, vallitsee negatiivinen korrelaatio, 0 viittaa korrelaation puuttumiseen ja 1 on positiivinen korrelaatio. Pearsonin korrelaatiokerroin skalaarisille luvuille *A* ja *B* saadaan kaavan 11 mukaisesti.

$$p(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i - \overline{\mu_A}}{\sigma_A}\right)\left(\frac{B_i - \overline{\mu_B}}{\sigma_B}\right), \tag{11}$$

Jossa *A* ja *B* ovat satunnaisia skalaarisia vakioita, *N* on havaintovektoreiden kokonaislukumäärä, $\overline{\mu_A}$ on muuttujan *A* keskiarvo ja vastaavasti $\overline{\mu_B}$ on muuttujan *B* keskiarvo, muuttujan *A* keskihajonta on σ_A ja vastaavasti σ_B on muuttujan *B* keskihajonta.

Kaavaa 11 käytetään kaikille muuttujille ja näistä tuloksista muodostetaan matriisi. Muuttujien lukumäärän ollessa *i*, tulee korrelaatiomatriisista $i \times i$ kokoinen. Jokainen muuttuja on matriisissa kahteen kertaan. Kaavan 11 merkinnöillä, mikäli *A* ja *B* ovat yhtä suuret on Pearsonin korrelaatiokerroin 1. Tästä johtuen korrelaatiomatriisin diagonaaliarvot ovat aina 1. (MathWorks inc. 2019c)

Muuttujan korreloidessa hyvin toiseen muuttujaan Pearsonin korrelaatiokerroin on lähellä -1 tai 1. Tällöin muuttujien välillä on arvojen vähenemisen tai kasvamisen välillä riippuvuus. Korrelaatiokertoimen ollessa nolla muuttujat ovat täysin toisistaan riippumattomia. Tässä työssä korrelaatio on hyvä, mikäli $|p(A,B)| > 0{,}75$ toteutuu.

Teknisistä tunnusluvuista tehdyn korrelaatiomatriisin sijasta päädyttiin jakamaan aineisto valmiiksi määriteltyihin tuotos-, panos- ja toimitusvarmuusmuuttujiin kuten luvussa 4.2.3

todettiin. Teknisissä tunnusluvuissa ei hyödynnetty korrelaatiomatriisia, sillä korrelaatiomatriisia tutkiessa huomattiin korrelaatioiden vaihtelevan huomattavasti. Osa korreloi hyvin positiivisesti tai negatiivisesti, osa heikosta ja osa ei ollenkaan. Kohtuullisen hinnoittelun laskelmien kohdalla päädyttiin ulottuvuuksien redusointiin hyödyntää korrelaatiomatriisia, koska suurin osa muuttujista korreloi verkon nykykäyttöarvoon joko hyvin positiivisesti tai negatiivisesti. Korrelaatio oli seuraavilla muuttujilla nykykäyttöarvoon verrattuna heikko ($p(A,B) < |0{,}45|$):

- vaihto-omaisuus tasearvossa
- poistoerä muista kuin verkon hyödykkeistä, oman pääoman osuus
- saadut konserniavustukset oman pääoman osuus
- oikaistun taseen tasauserä
- pääomalainat tasearvossa
- annetut, mutta maksamattomat, korolliset konserniavustukset
- annetut, mutta maksamattomat, korottomat konserniavustukset
- pakolliset varaukset tasearvossa
- muiden kuin verkon hyödykkeiden poistoeron laskennallisen verovelan osuus
- muihin kuluihin kirjattu verkonosuuden myyntitappio
- muihin tuottoihin kirjattu verkonosuuden myyntivoitto
- taseen liittymismaksukertymän nettomuutos

Muutaman muuttujan kohdalla korrelaatio oli kohtalainen ($0{,}45 < |p(A,B)| < 0{,}75$):

- annettujen konserniavustuksen oman pääoman osuus
- edellisen ja tämän valvontajakson yli/alijäämän määrä

Lopuilla korrelaatiokerroin oli nykykäyttöarvoon nähden korkea. Pienempään aineistoon valittiin osittain korrelaation perusteella 9 muuttujaa:

- sähköverkko oikaistussa nykykäyttöarvossa
- vaihto-omaisuus tasearvossa, poistoerä muista kuin verkon hyödykkeistä oman pääoman osuus
- saadut konserniavustukset oman pääoman osuus
- annetut mutta maksamattomat korolliset konserniavustukset
- muiden kuin verkon hyödykkeiden poistoeron laskennallisen verovelan osuus
- taseen liittymismaksukertymän nettomuutos
- innovaatiokannustin
- toteutunut oikaistu tulos

Osa ei-korreloivista muuttujista valikoitui pois erilaisista ulkoisista syistä. Esimerkiksi verkkojen ostojen ja myyntien korrelaatiokerroin on melko lähellä nollaa verrattuna verkon nykykäyttöarvoon, koska tiedot koskevat vain harvoja yhtiöitä.

4.4 Klusterien lukumäärä ja yhtiöiden kokoerojen kompensointi

Kuten luvussa 3.1 todettiin, k-means ja k-medoids menetelmät vaativat klusterien lukumäärän etukäteen. Työn alussa päädyttiin valitsemaan k-arvo välille 2-20. Tämän perusteella tutkittiin mitkä arvot soveltuisivat käytettäväksi. K-arvot 2-4 todettiin olevan edelleen liian pieniä ja arvot lähempänä 20 ovat ei toivottuja tämän työn myöhemmän tarkastelun kannalta. Tarkoituksena on etsiä alkioiden perusteella yhtäläisyyksiä ja klusterien sisältäessä yhden alkion tai vain pienen määrän alkioita, on vaikeampi löytää yhteisiä tekijöitä. Liian suurella alkioiden määrällä ongelmaksi muodostuu liian useat satunnaiset yhteiset tekijät ja klusteroinnin hyöty analysoinnin apuna pienenee klusteroinnin löytäessä olemattomia ryhmiä.

K-arvoa arvioitiin prototyyppivaiheessa tutkimalla muodostettuja klustereita eri arvoilla ja pääteltiin yhdelle vuodelle lukumäärä. Lukumäärää pyrittiin edelleen etsimään siluetti -menetelmällä Matlabin avulla. Klusterointiasetusten vaihtamisella todettiin olevan vaikutusta, sekä k-meansin ominaisella satunnaisuudella. Klusterien optimaalisen lukumäärän etsimiseen siluetti -menetelmä ei pystynyt määrittämään järkeviä lukuja ja eri ajokertojen tulokset eivät olleet yhdenmukaisia keskenään. Tämän lisäksi vuosien välillä oli eroja. Kuten Mutasen tutkimuksessa (Mutanen 2018), myös tässä tutkimuksessa siluetti antoi k-means klusteroinnille parhaan k-arvon ollessa 2. Tämä on tutkittavan aineiston ja tehtävän tavoitteen kohdalla liian pieni k-arvo.

Teknisten tunnuslukujen k-means ja k-medoids tuloksia vertailemalla todetaan tulosten eriävän. K-arvona käytettiin arvoa 12 ja aloitusvektorina k-meansissä aikaisemmin mainittua tasa-arvoaloitusta. K-medoidsin kohdalla käytettiin 500 replikaattia. Johtuen sattumanvaraisesta aloitusvektorista k-medoidsin klusterien indeksit eroavat k-meansista. Samassa tarkastelussa todettiin k-medoidsin saavan eri ajokerroilla samat klusterin jäsenet, vaikka klusterien indeksit erosivat.

Yhtiöiden kokoa huomioimattomat tunnusluvut jaettiin käyttöpaikkamäärillä. Tällä pyrittiin kompensoimaan yhtiöiden keskinäisiä kokoeroja. Teknisiä tunnuslukuja korjattiin vuosittain jakamalla 0,4-110 kV käyttöpaikkojen summalla kaikki absoluuttiset arvot. Prosenttilukuja tai jo suhteutettuja arvoja ei korjattu. Vaikka kaikki muuttujat eivät ole suhteutettu samalla arvolla, kokoero yhtiöiden välillä tasoittuu. Kohtuullisen hinnoittelun laskelmia kompensoitiin tuomalla käyttöpaikkamäärät teknisistä tunnusluvuista ja jakamalla niillä kaikki aineiston muuttujat.

Käyttöpaikkamäärillä jaettuna tunnuslukuihin voi tulla virhettä yksittäisten yhtiöiden kohdalla. Tämä voi tapahtua erittäin suurille tai erittäin pienille yhtiöille, johtuen koon määrittelemisen moninaisuudesta ja yksinkertaisesta painotusperiaatteesta. Kokoerot tulee kuitenkin kompensoida ja kompensoinnista syntyvä mahdollinen virhe oletetaan pieneksi. Mahdollisen virheen havainnointi yksittäisestä yhtiöstä on manuaalisesti aikaa

vievä prosessi ja lopputuloksen tarkkuudesta ei ole varmuutta. Klusterointitulokset muuttuvat riippuen siitä käytetäänkö käyttöpaikkakompensaatiota vai ei, jolloin painotusta tulee käyttää ainoastaan, kun taustalla oleva tehtävä sitä vaatii.

5. TULOKSET JA JOHTOPÄÄTÖKSET

Klusterointituloksissa syntyi eriytymiä ja paras k-arvo vaihteli aineistojen välillä. Työn edetessä huomattiin klusteroitavan aineiston olevan suurin tuloksiin vaikuttava tekijä. Vuoden 2017 teknisissä tunnusluvuissa oli eriytymien kaltaisia yhtiöitä ja muuttujat eivät olleet samoissa yksiköissä. Vuosien välillä havaittiin myös muutoksia niin yhtiömäärissä kuin tunnusluvuissakin. Osittain tunnuslukujen arvot vaihtelivat vuosien välillä huomattavasti.

Tutkimuksessa klusterointitulosten analysointi vuoden 2017 aineistoilla voitiin jakaa 3 osaan. Ensimmäinen osa koostui yhden aineiston klusterointitulosten analysoinnista lähtöaineiston avulla, toinen osa aineiston ja siitä luodun aliaineiston klusteroinneista ja tulosten vertailuista, sekä viimeinen kahdesta toisistaan näennäisesti riippumattoman aineiston tulosten vertailemisesta. Ensimmäinen ja toinen osa suoritettiin teknisille tunnusluvuille ja viimeinen osa teknisille tunnusluvuille ja kohtuullisen hinnoittelun laskelmille. Lisäksi demonstroitiin täysin ulkoisen aineiston hyödyntämistä sisäisen aineiston avulla.

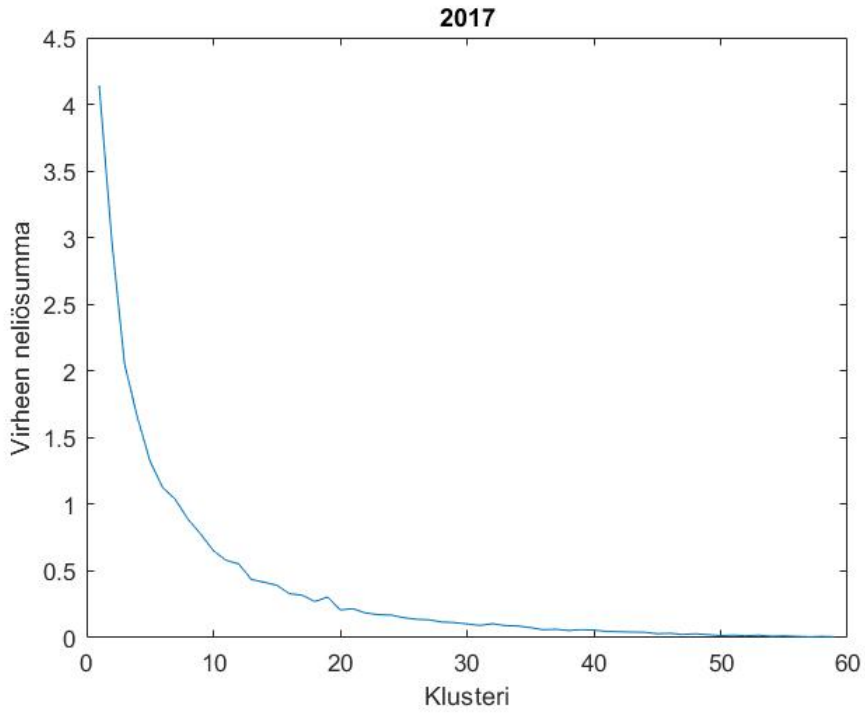
Käyttämällä euklidinista etäisyyttä oletetaan klustereiden muodostuvan pallomaisiksi ja tasaisen suuriksi. Käytetyistä aineistoista ei voida sanoa tulosten perusteella, toteutuuko tämä ja ovatko klusterit muodostuneet todellisuutta vastaaviksi.

5.1 K-arvon määritteleminen vuoden 2017 aineistoille

Kuvaajat teknisille tunnusluvuille määräanalyyseistä käytettäessä koko aineistoa ovat liitteessä D. Analyysit suoritettiin Matlabin evalclusters -funktiolla käyttäen väliä [2,20] ja k-meansia. K-medoidsia ei evalclusters -funktiossa ole. (Mathworks inc. 2019e)

Yleisesti todettiin yhden alkion klusterien olevan pienempi ongelma kuin liian suurien klusterien. Yksittäisiä yhtiöitä voidaan tutkia mahdollisten erityisten ominaisuuksien varalta, mikä selittäisi ryhmittymisen. Syynä tähän oli alkuperäinen tehtävä selittävien tekijöiden löytämiseksi. Arvojen perustuessa uusimpiin käytettävissä oleviin tunnuslukuihin niiden vastaavuus lyhyellä aikajänteellä toimii uusien vuosien klusteroinnissa. Tämän lisäksi mahdolliset huomiot aineistosta kuvaavat nykytilannetta, sekä tämän hetkistä valvontajaksoa.

Maakaapelointiasteen-testiaineistosta tehtiin lukumääräanalyysit (liite B) ja virheen neliösumma -käyrä laskettiin k-meansilla k-arvon määrittelemiseksi. K-arvo määräytyi suurimmalta osin aineistosta tehdystä virheen neliösumma -käyrästä Kuva 3 perusteella.



Kuva 3. *Maakaapelointiasteiden virheen neliösumma -käyrä k-means.*

Lukumäärän analyysit antoivat osittain eriävät tulokset. Siluetti ja Calinski-Harabasz -analyysit ehdottivat k-arvoksi 2. Siluetin perusteella seuraavaksi paras k-arvo olisi 4 ja sitten 19. Calinski-Harabasz -analyysi ehdotti toissijaisiksi samoja arvoja, mutta eri järjestyksessä. Gap-analyysin perusteella paras arvo olisi 20. Davies-Bouldin indeksin perusteella k-arvon tulisi olla 19.

Analyysien perusteella klusterien lukumäärä voisi tarkasteluvälillä olla 2, 4 tai 19. Tarkastelemalla tuloksia tehtävän kannalta, k-arvo 2 on liian pieni, sillä klustereiden muodostamilla yhtiöillä on liian paljon yhteisiä ja eriäviä tekijöitä, jolloin on hankalaa tai mahdotonta päätellä selittäviä tekijöitä. Myöskään virheen neliösumma -käyrä ei tue k-arvoa 2. K-arvolla 19 klustereita on liikaa yksittäisinä klustereina, yhteensä 5 yhtiötä. Tuloksista nähdään kuitenkin selkeät rajat ja klusterien arvot ovat klusterin sisällä lähempänä toisiaan kuin alemmilla k-arvoilla. Tämän perusteella voidaan todeta liian suuren k-arvon olevan parempi tulosten analysoinnin kannalta kuin liian pienen. Testiaineiston kohdalla paras k-arvo olisi 4. Analyysien tulosten perusteella 4 on lähellä parasta arvoa. Klustereita ei muodostu yksittäisiksi ja selkeät jakoperusteet ovat nähtävillä vertaillessa klusterointituloksia absoluuttisiin arvoihin.

Taulukossa 5 on merkittynä k-arvon määrittämiseen käytettyjen määränanalyysimenetelmien (Liite C) perusteella tehdyt arviot k-arvoiksi teknisisille tunnusluvuille. Ensimmäinen lokaali paras tarkoittaa joko suurinta tai pienintä analyysissä saatua arvoa. Toinen lokaali paras tarkoittaa seuraavaksi parasta arvoa.

Taulukko 4. *Teknisten tunnuslukujen koko aineistosta tehty määräanalyysien tulokset.*

Vuosi	1. lokaali paras k-arvo				• 2. lokaali paras k-arvo			
	Siluetti	**Calinski-Harabasz**	**Gap**	**Davies-Bouldin**	**Siluetti**	**Calinski-Harabasz**	**Gap**	**Davies-Bouldin**
2015	6	3	16	18	18	10	18	11
2016	8	5	15	18	12	12	18	11
2017	6	3	16	17	14	9	20	14

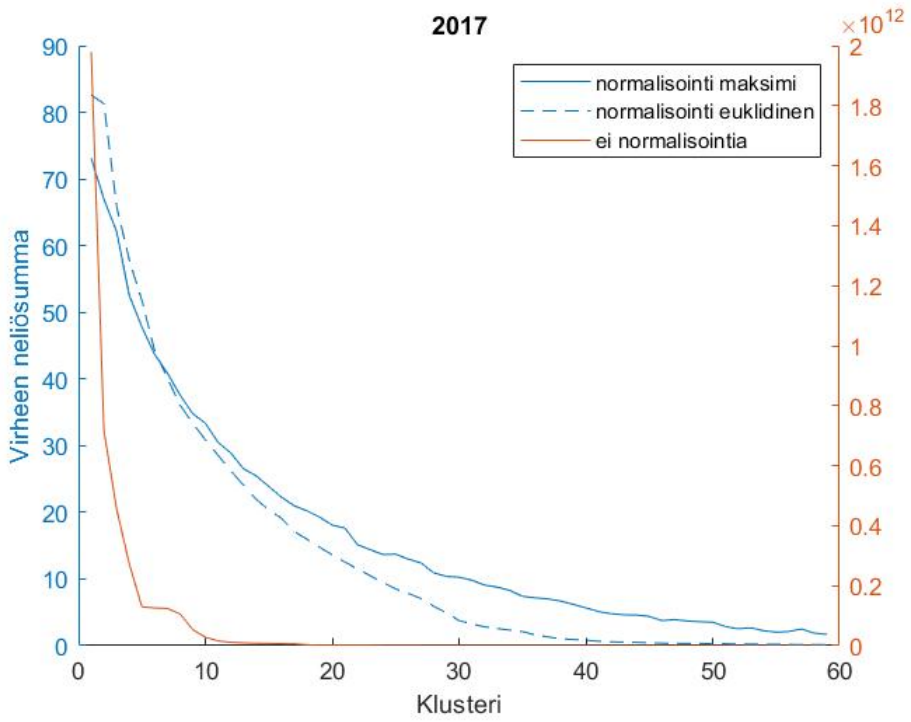
Vaikka taulukon 5 mukaiset arvot antavat viitteitä, niin myös pienemmillä k-arvoilla saadaan lokaaleja hyviä tuloksia. Teknisille tunnusluvuille siluetti -analyysi antoi parhaaksi k-arvoksi 2 jokaiselle vuodelle. K-arvo 2 hylättiin liian pienenä tarkasteluvälin valinnassa. Siluetilla k-arvot 12-16 loivat puolikaaren, lokaalin maksimin ollessa 14. Calinski-Harabasz -analyysin mukaan vuoden 2017 toisen lokaalin maksimin jälkeen arvo ei laske huomattavasti klusterien lukumäärän kasvaessa. Myös vuosien 2015 ja 2016 Calinski-Harabasz -arvoille käy samoin. Gap-analyysien tulos nousee klusterien lukumäärän kasvaessa. Gap-analyysi saisi parasta arvoa alemman lokaalin maksimin jo arvolla 12, mutta arvo pysyy kuitenkin 16 klusterin jälkeen 12 korkeammalla. Davies-Bouldin -analyysi antaa 1. lokaalin minimin jälkeen myös alemmilla k-arvoilla hyviä tuloksia ja varsinkin vuoden 2017 kohdalla ero ensimmäisen lokaalin minimin ja toisen lokaalin minimin välillä on pieni.

Tulokset vaihtelivat paljon ajokertojen välillä. Vertaillessa käyttäjäpainotuksen vaikutusta k-arvoon arvioitiin käyttäjäpainotuksen pienentävän virhesummaa absoluuttisten arvojen pienentyessä. Tämä ei automaattisesti johda tulosten parantumiseen. Klusterit vaikuttaisivat muodostuvan epätasaisen kokoisiksi ja selvää k-arvoa ei voida näin monilla epätasaisesti muodostuville klustereille yksiulotteisesti löytää siluetti-analyysillä. Käyttämällä muita analyysejä saadaan vastaavanlaisia tuloksia, tulosten muuttuessa ajokertojen välillä. Kasvattamalla klusterien lukumäärää ja lähestyttäessä maksimia tulokset paranevat, sillä klusterin havaintovektoreiden summavirhe korreloi negatiivisesti klusterien määrän kasvun kanssa.

Määräanalyysimenetelmien jälkeen k-arvon valitsemisessa käytettiin niin kutsuttua virheen neliösumman polvimenetelmää. Menetelmässä arvioidaan käyrästä kohta, missä virhesumma ei enää pienene huomattavasti klusterien lukumäärän kasvaessa. Visuaalisesti tutkimalla virheen neliösummakäyrää teknisten tunnuslukujen aineistolla ei selvää käännekohtaa löydetty normalisoiduille teknisille tunnusluvuille alle 20 klusterin.

Kuvassa 4 nähdään teknisten tunnuslukujen virheen neliösumma -käyristä normalisoinnin vaikutukset virhesummaan. Kuvan 4 perusteella normalisoimattomalla aineistolla k-arvo olisi 5 tai halutessa virhesummaa pienemmäksi, seuraava potentiaalinen k-arvo olisi noin 12. Maksiminormalisointi näyttäisi tukevan 12, vaikka selvää taittumakohtaa ei

näyttäisi olevan. Euklidinisella normalisoinnilla selvä polvikohta nähdään k-arvon 30 jälkeen.



Kuva 4. *Virheen neliösumma -käyrä teknisistä tunnusluvuista käyttöpaikkapainotuksella.*

Teknisille tunnusluvuille päädyttiin käyttämään virheen neliösumma -käyrän perusteella k-arvoa 12. Määräanalyysit osittain tukivat valintaa. K-arvo määriteltiin välille [2,20] ja tällä välillä 12 on kompromissi yksittäisten klusterien ja liian suurten klusterien välillä.

Käyttöpaikkapainotuksella ja euklidinisella normalisoinnilla kohtuullisen hinnoittelun virheen neliösumma -käyrästä voidaan polvikohta löytää selvästi k-arvolla 11. Ilman käyttöpaikkapainotusta käyrä silottuu taitekohdistaan ja polvikohta siirtyy kauemmaksi, k-arvoon 16. Kohtuullisen hinnoittelun laskelmista tehdyn pienen aineiston virheen neliösumma -käyrästä voidaan huomata selkeä polvikohta 4 klusterilla. Edelleen suurimman klusterin virheen neliösumma -käyrä viittaa 4 klusteriin.

Vaikka virheen neliösumma -käyrän perusteella k-arvon tulisi olla 11 kohtuullisen hinnoittelun laskelmilla, käytettäväksi k-arvoksi valittiin 12. Valinta tehtiin virheen neliösumma -käyrien perusteella ja osittain analyysien tukemana. Kahdentoista valinnassa suurimmaksi tekijäksi nousi teknisten tunnuslukujen käyttöpaikkamäärillä painotetut vuoden 2017 virheen neliösumma -käyristä tehdyt arvioit optimaalisesta arvosta. Tämän lisäksi vuosien 2015-2017 siluetti-, Calinski-Harabasz- ja gap- analyysien perusteella voidaan teknisille tunnusluvuille pitää kohtalaisena kompromissina arvoa 12. Vuosien välillä oli pieniä eroja ja k-arvon valinnassa huomioitiin kaikki vuodet. Suurin painoarvo annettiin kuitenkin vuodelle 2017, koska tarkemmat tarkastelut tehtiin käyttäen vuotta 2017.

Aineiston vaihtuessa ja aineiston sisällä vuosien välillä jokin toinen k-arvo saattaisi soveltua paremmin käytettäväksi. Aineistojen ja vuosien välillä päädyttiin käyttämään samaa klusterien lukumäärää, jotta nähtäisiin muodostuvatko klusterit samalla tavalla eri aineistoilla.

5.2 Työkalun testaus

Prototyyppivaiheessa ja testiaineistona käytetään jokaisen yhtiön muutamaa tunnuslukua yhdeltä vuodelta. Prototyypin aineistona on käytetty yksinomaan teknisiä tunnuslukuja ja niiden variaatioita vuodelta 2017. Valmista työkalua testattiin vuoden 2017 maakaapelointiasteilla. Tämän jälkeen klusterointia laajennetaan asteittaisesti koskemaan kaikkia tunnuslukuja ja kaikkia tarkasteluvälin vuosia. Työkalussa voidaan valita erikseen klusteroitavat vuodet.

Klusterointia ja työkalua testattiin klusteroimalla teknisissä tunnusluvuissa mukana olevia maakaapelointiasteita vuodelta 2017 (Liite D). Aineistona oli siis 0,4 kV, 1-70 kV ja 110 kV maakaapelointiasteet. Tällöin aineiston muodostama matriisi oli 77 × 3. Klusterointi suoritettiin tietoisena 110 kV maakaapelointiasteen osittaisesta virheestä. Kaikilla verkonhaltijoilla ei ole ollenkaan 110 kV verkkoa, jolloin myös verkon maakaapelointiaste on 0 %. Aineistossa on mukana myös yhtiöt, joiden SJ-verkko muodostuu kokonaisuudessaan ilmajohdosta. Klusterointi tehtiin kahteen kertaan k-arvolla 12. Ensimmäisellä kerralla käytettiin k-medoidsia 500 replikaatilla ja euklidinista normalisointia. Toisella kerralla käytettiin k-meansia, tasavälialoitusta ja maksiminormalisointia. Tuloksia verrattiin absoluuttiseen aineistoon. Normalisoinnin vaikutusta tutkittiin myös k-meansilla.

K-medoidsin luomissa klustereissa on selvästi havaittavissa erottelua. Kolme klusteria muodostui yksittäisiksi klustereiksi. Samat yhtiöt muodostuivat myös k-meansilla yhden jäsenen klustereiksi. Yhteistä näillä yhtiöillä oli korkea SJ-verkon kaapelointiaste, 100 %, 62 %, ja 40 %. PJ- ja KJ-verkon puolesta kyseiset yhtiöt sopisivat muihin klustereihin. Soveltuvuutta ja SJ-verkon vaikutusta testattiin k-meansilla poistamalla aineistosta SJ-verkon luvut ja yhtiöt sulautuivat muihin klustereihin, muodostamatta ainuttakaan yhden yhtiön klusteria. Klustereiden jäsenten PJ- ja KJ-kaapelointiasteen ollessa kohtalaisen samoja, erottavaksi tekijäksi muodostui juurikin SJ-verkon kaapelointiaste.

Esimerkiksi Helen Sähköverkko Oy:n ja Rovaniemen Verkko Oy:n ollessa kahdestaan samassa klusterissa, olisi JE-Siirto Oy ilman 15 % eroa SJ-kaapeloinnissa luultavimmin samassa klusterissa. Huomioitavaa on myös SJ-verkon pituuksien ero Helenin ja Rovaniemen välillä. Suhteellisen kaapelointiasteen ollessa prosentteina yhtä suuri, absoluuttisissa pituuksissa on SJ-verkolla eroa 175 km. Yleisesti pieni SJ-kaapelointiprosentti ja nolla eivät saaneet eri painoarvoa. Vasta huomattavat erot kaikkien muuttujien kohdalla erottelivat yhtiötä eri klustereihin. K-medoidsin perusteella klusterointityökalu toimii halutulla tavalla. Klusterointi ajettiin kahteen kertaan ja klustereissa ei tullut muutoksia.

K-means klusteroinnilla ja euklidinisella normalisoinnilla tulokset olivat samankaltaisia kuin k-medoidsilla. Yksittäisiä klustereita oli kolme, jotka sulautuivat muihin klustereihin jättäessä SJ-kaapelointiasteen pois. Myös Helen Sähköverkko Oy ja Rovaniemen Verkko Oy olivat kahdestaan samassa klusterissa. Osa pienistä muutaman yhtiön klustereista yhdistyi ja isoin klusteri hajosi kahteen verrattuna k-medoidsiin. Lisäksi yksittäisiä muutoksia k-medoidsiin havaittiin. Pääsääntöisesti klusterointitulokset olivat edelleen loogisia ja samat havainnot k-medoidsista toistuivat.

Vaihdettaessa normalisointia k-meansissa, tulokset eroavat aikaisempaan. Euklidinisella normalisoinnilla alun perin saatu ensimmäinen klusteri hajoaa kolmeen erilliseen klusteriin maksiminormalisoinnilla. Kaksi yhtiötä vaihtavat myös klusteria. Edelleen seuraava klusteri hajoaa kahteen erilliseen klusteriin. Maksimi normalisoinnilla muodostuu ainoastaan yksi yksittäinen klusteri Rauman Energia Sähköverkko Oy:stä. Oletuksena Rauman Energia Sähköverkko Oy:n eriytymiselle oli 100 % SJ-verkon kaapelointiaste. Yhtiön lähin etäisyys oli oman klusterin jälkeen klusteriin 12, jolla on nollasta eroava SJ-verkon keskipiste. Tämän jälkeen tulivat klusterit 10 ja 8, joilla ei ole nollasta eroavaa SJ-verkon keskipistettä. Taulukossa 4 on esitelty keskipisteet molempien menetelmien tapauksessa.

Taulukko 5. *Muuttujien keskipisteet maksimi normalisoidulla k-meansilla ja euklidinisella normalisoidulla k-medoidsilla.*

Klusterin indeksi	K-means			K-medoids		
	PJ-verkon keskipiste	KJ-verkon keskipiste	SJ-verkon keskipiste	PJ-verkon keskipiste	KJ-verkon keskipiste	SJ-verkon keskipiste
1	0,335	0,243	0,014	0,142	0,194	0,167
2	0,483	0,041	0,000	0,174	0,149	0,276
3	0,430	0,159	0,000	0,093	0,083	0,003
4	0,285	0,323	0,026	0,037	0,017	0,000
5	0,315	0,152	0,000	0,147	0,186	0,689
6	0,361	0,204	0,000	0,161	0,160	0,069
7	0,385	0,157	0,000	0,169	0,235	0,000
8	0,401	0,235	0,000	0,156	0,201	0,424
9	0,462	0,120	0,000	0,121	0,055	0,000
10	0,354	0,098	0,000	0,153	0,154	0,000
11	0,371	0,116	0,000	0,191	0,257	0,241
12	0,455	0,068	0,008	0,081	0,037	0,000

Vaihtamalla k-arvo neljään, huomattiin klusterointituloksissa selvä suuruusjärjestys, osittain tasavälialoitusvektoreiden takia. Ensimmäinen klusteri muodostui vähäisistä PJ- ja KJ-kaapeloinneista, toisessa klusterissa oli hieman enemmän, kolmannessa edelleen suuremmat kaapelointiasteet ja neljännessä klusterissa arvot olivat samoja kuin kolmannessa tai suurempia. SJ-kaapelointiasteet olivat kolmessa ensimmäisessä klusterissa alle 30 %.

Kaikissa paitsi neljännessä klusterissa oli yhtiöitä, joilla SJ-kaapelointi oli 0 %. SJ-kaapelointi yleistyi ja kasvoi yhtiöillä klusterin indeksin kasvaessa.

Molemmat klusterointimenetelmät tuottivat maakaapelointiasteilla toimivia tuloksia ja työkalu huomioi jokaisen tunnusluvun arvot. Normalisoinnilla oli vaikutusta tuloksiin. Tulosten perusteella vaikuttaisi, ettei yhden muuttujan huomattava ero vaikuta ratkaisevasti klustereihin muiden arvojen ollessa tarpeeksi lähekkäin aineistoon nähden liian isoilla k-arvoilla. Testi aineiston kohdalla k-arvo 12 oli mahdollisesti väärä pakottaen maksimi normalisoinnilla Rauman Energian yksittäiseksi klusteriksi SJ-verkon kaapelointiasteen perusteella. Testiaineistolla ei testattu käyttäjäpainotuksen vaikutusta klusterointituloksiin. Maakaapelointiasteet olivat prosenttilukuina suhteutettu verkkopituuksiin.

5.3 Kaikkien teknisten tunnuslukujen ja ABC-aineiston klusterointi

Teknisiä tunnuslukuja klusteroitiin useilla eri aineistovariaatioilla. Näitä olivat: Koko aineisto, panosmuuttujat, toimitusvarmuusmuuttujat, volyymimuuttujat, tunnuslukukerrallaan ja ”ABC-aineisto”. Absoluuttisissa teknisissä tunnusluvuissa oletettiin esiintyvän taustatekijänä yhtiöiden koko. Tämä oletus perustui tunnuslukujen absoluuttisiin arvoihin ja klusteroinnin teoriaan. Klusterien indekseillä viitataan liitteessä E oleviin klusterien indekseihin. Tunnuslukujen lähemmässä tarkastelussa hyödynnettiin Excelin sort- ja heat-map -toimintoja.

Vertailemalla ilman käyttöpaikkapainotusta tehtyä klusterointia ja siirrettyä energia- ja käyttöpaikkamäärällä, klusterointituloksessa esiintyi yhtiön kokoon perustuvaa jakautumista. Caruna Oy, Elenia Oy ja Helen Sähköverkko Oy olivat yksinäisinä klustereina koko aineistolla klusteroitaessa. Edellä mainitut kolme yhtiötä olivat käyttöpaikkamääriltään ja siirretyiltä energioiltaan suurimmat yhtiöt. Lisäksi yksittäisenä klusterina oli Jeppo Kraft Andelslag, jonka käyttöpaikkamäärä vuonna 2017 on kaikista yhtiöistä pienin: noin tuhannesosa Caruna Oy:n käyttöpaikkamäärästä.

Tämä kokoon perustuva huomio ei ollut täysin yksiselitteinen ja lisätietoa etsittiin vertailemalla klusterointituloksia keskenään ja absoluuttisiin arvoihin. Esimerkiksi käyttöpaikkapainottomalla ”ABC-aineistolla” tehdyllä klusteroinnilla Lappeenrannan Energiaverkko Oy kuuluu ensimmäiseen klusteriin Kuopion Sähköverkko Oy:n kuuluessa indeksiltään toiseen klusteriin. Lappeenrannan Energiaverkko Oy:n siirretty energia ja käyttöpaikkamäärä ovat enemmän kuin Kuopion Sähköverkko Oy:llä. Mikäli energia tai käyttöpaikkamäärä olisivat ainoat yksiselitteiset yhteiset tekijät, näiden yhtiöiden paikat vaihtuisivat keskenään.

5.3.1 ABC-aineiston klusterointi käyttöpaikkapainotuksella

Klusteroimalla A, B ja C aineistojen muodostamaa yhtenäistä kokonaisuutta k-meansilla, käyttöpaikkapainotuksella ja maksimi normalisoinnilla, ei löydetä yksittäisiä tunnuslukuja kuvaamaan kokonaisuutta. Vertailemalla näiden kolmen ryhmän klusterointeja ABC-aineistoon huomataan suurimmasta klusterista tehdyn toisen klusteroinnin tulosten jakautuvan osittain toimitusvarmuusmuuttujista tehtyyn klusterointiin. Kaksi suurinta klusteria sisälsivät 61 yhtiötä (79 %). Suurimman klusterin toisen asteen klusteroinnin jakamat yhtiöt ovat toimitusvarmuus-aineistossa samassa klusterissa.

Liitteen E mukaisten klusterointien keskipistevektorien lähin yhtiö ja sen etäisyys muista klustereista on esitetty Taulukko 6.

Taulukko 6. *Klusterin keskipistevektoria lähin yhtiö ja sen etäisyys muiden klustereiden keskipistevektoreihin.*

Yhtiö	Yhtiön etäisyys muista klustereista ja (yhtiöiden lukumäärä klusterissa)											
	1 (1)	**2 (27)**	**3 (34)**	**4 (2)**	**5 (1)**	**6 (1)**	**7 (3)**	**8 (1)**	**9 (1)**	**10 (2)**	**11 (2)**	**12 (2)**
Jeppo Kraft Andelslag	**0,00**	6,69	6,10	9,05	18,20	12,18	8,27	8,74	8,89	8,24	8,09	7,41
Vantaan Energia Sähköverkot Oy	7,12	**0,10**	1,78	4,03	15,63	5,66	2,91	5,86	4,01	4,82	2,74	2,54
Koillis-Satakunnan Sähkö Oy	6,69	1,40	**0,15**	3,05	13,42	7,08	2,11	4,68	2,72	2,35	2,26	2,10
KSS Verkko Oy	9,19	3,25	3,37	**1,15**	9,09	8,83	4,36	5,35	5,56	6,11	4,80	4,33
PKS Sähkönsiirto Oy	18,20	15,04	12,74	6,67	**0,00**	19,67	13,21	10,26	14,09	14,92	15,57	14,41
Herrfors Nät-Verkko Oy Ab	12,18	5,70	6,46	8,84	19,67	**0,00**	6,08	10,97	9,65	8,78	7,20	8,32
Caruna Oy	9,27	3,51	2,93	4,48	13,50	6,79	**1,05**	6,39	5,84	5,33	4,48	4,67
Sipoon Energia Oy	8,74	5,39	4,14	4,39	10,26	10,97	5,78	**0,00**	6,40	7,20	6,78	5,29
Haukiputaan Sähköosuuskunta	8,89	3,57	2,84	5,31	14,09	9,65	4,66	6,40	**0,00**	5,77	4,35	4,43
Muonion Sähköosuuskunta	8,86	4,53	2,84	5,62	15,37	9,14	4,51	7,64	6,30	**0,95**	4,23	5,09
Koillis-Lapin Sähkö Oy	9,38	3,99	2,95	5,05	15,63	7,60	2,78	7,71	5,00	3,55	**0,96**	4,98
Imatran Seudun Sähkönsiirto Oy	8,86	2,57	2,55	5,28	16,65	8,93	4,50	7,04	5,13	4,98	4,13	**0,78**

Taulukko 6 avulla pystytään yhden jäsenen klusterit liittämään isompiin klustereihin haluttaessa. Esimerkiksi PKS Sähkönsiirto Oy voitaisiin siirtää tämän perusteella klusterista 5 klusteriin 4, jolloin päästäisiin yhden jäsenen klusterista eroon. Taulukosta voidaan myös huomata klusterin 12 eniten edustavimman jäsenen eroavan klusterin 10 vastaa-

vasta. Tämän perusteella pystyttäisiin tutkimaan Imatran Seudun Sähkönsiirto Oy:n tietoja ja verrata niitä Muonion Sähköosuuskunnan tietoihin, jolloin saataisiin parempi kuva klustereiden eroavaisuuksista ja niiden ominaisuuksista. Taulukkoa voitaisiin laajentaa kaikkiin klusteroinnissa olleisiin 77 yhtiöön ja verrata yksittäisen yhtiön etäisyyksiä muihin klustereihin. Jokaisen aineiston klusterointituloksesta voidaan muodostaa vastaava taulukko.

Muutamat yhtiöt ovat yksittäisinä klustereina ”ABC-aineistolla”, mutta volyymi ja tuotos klusteroinneissa useamman yhtiön klustereissa. Kyseiset yhtiöt ovat toimitusvarmuus-aineistossa kuitenkin yhden jäsenen klustereina. Tämä viittaisi toimitusvarmuusmuuttujien vaikuttavan olennaisesti ”ABC-Aineiston” kokonaistuloksiin.

Suurin klusteri pääsääntöisesti muodostui volyymimuuttuja-aliklusteroinnin yhden klusterin sisään. Tämän lisäksi havaittiin yhtäläisyyksiä toimitusvarmuus-klusterointiin Toiseksi suurimman klusterin klusterointituloksissa on yhtäläisyyksiä volyymimuuttujista tehtyyn klusterointiin, kaikkien yhtiöiden kuuluessa volyymimuuttuja-klusteroinnin kolmeen klusteriin. Yksi näistä klustereista vastaa täysin 11 yhtiötä ja näiden yhtiöiden muodostamat klusterit sisältyvät kokonaisuudessaan kyseiseen klusteriin. Saman klusterin yhtiöt muodostavat panosmuuttuja-aliklusteroinnin kaksi klusteria, kolmea poikkeusta lukuun ottamatta. Poikkeukset kuuluvat indeksiltään seuraavaan klusteriin.

Lopuilla 16 yhtiöllä vastaavuudet volyymi-, panos- ja toimitusvarmuusklusterointeihin vaihtelivat. Yhtä klusteria yhdisti volyymi- ja panos -klusteroinnit ja toimitusvarmuusmuuttujissa erosivat omiksi yksittäisiksi klustereiksi. Erään klusterin ainoa yhteinen tekijä oli toimitusvarmuus klusteroinnin tulos. PKS Sähkönsiirto Oy oli yksittäisenä klusterina, mutta olisi soveltunut volyymi- ja panosklusterointien perusteella suurimpaan klusteriin. PKS Sähkönsiirto Oy:n kohdalla erottava tekijä tuli toimitusvarmuusklusteroinnin tuloksessa, jossa yhtiö oli omana yksittäisenä klusterinaan.

5.3.2 Teknisten tunnuslukujen fuusio

Tekniset tunnusluvut fuusioitiin ja klusteroitiin. Klusteroinnissa käytettiin maksiminormalisointia ja tasavälialoitusvektoreita, k-arvon ollessa 12, sekä käyttöpaikkapainotusta. Mikäli kyseisenä vuonna käyttöpaikkamääriä ei ole kerätty, käytettiin asiakasmääriä. Yhtiöt ovat fuusioitu takautuvasti vastaamaan vuoden 2017 yhtiöitä. Fuusion ansiosta yhtiöt ovat vertailukelpoisia vuosien välillä. Tarkastelussa kuitenkin aiheutuu virhettä fuusioinnista joidenkin muuttujien ollessa keskiarvoja usean yhtiön tunnusluvuista. Tulokseksi saatiin liitteen F mukainen taulukko, jossa on merkittynä vuosittaiset klusterointitulokset yhtiöittäin teknisille tunnusluvuille vuosilta 1997-2017.

Valvontajaksojen välillä kerätyt tunnusluvut vaihtelevat ja klusteroitaessa vuosi kerrallaan tulokset eivät ole yhtiöiden ominaisuuksien kannalta vertailukelpoisia suoraan valvontajaksojen välillä. Haluttaessa tutkia tiettyä ominaisuutta tulisi valita tarkasteluväli,

jonka aikana kerätyt tiedot eivät ole muuttuneet tai vaihtoehtoisesti valita sellaiset tunnusluvut, joita on kerätty koko halutulla tarkasteluvälillä. Tämän tarkastelun tarkoituksena oli testata, voidaanko työkalua hyödyntää fuusioimalla yhtiöitä takautuvasti ja vertailla tuloksia keskenään.

Liitteessä F esitellyistä tuloksista nähdään tiettyjen yhtiöiden toistuvan yksittäisenä klusterina useina vuosina, kerättyjen teknisten tunnuslukujen muuttuessa vuosien välillä. Huomionarvoisesti ensimmäistä kertaa vuonna 2016 ja 2017 on Jeppo Kraft Andelslagin muodostama yksittäinen klusteri. Vuoden 2015 ja 2016 kerätyt tunnusluvut eroavat toisistaan, kun taas 2016 ja 2017 ovat samoilla tiedoilla kerätty. Fuusioitu Vantaan Energia Sähköverkot Oy on vuosina 1997-2013 yksittäisenä klusterina, kun taas siitä eteenpäin se on osa isompaa klusteria. Tämän perusteella klusterointituloksiin vaikuttaisivat vahvasti kerätyt tunnusluvut ja vuosien välillä muuttuneet tiedot vaikuttavat klusterointituloksiin. Yhtiöiden takautuva fuusio vaikuttaisi toimivan klusterointitulosten muuttuessa kerättyjen tunnuslukujen vaihtuessa.

Useiden tunnuslukujen keskiarvoistaminen ja summaaminen ei kuvaa täysin oikein yhtiöiden fuusiota ja parempiin tuloksiin päästäisiin valittaessa yksittäisiä muuttujia ja tekemällä tarkemmat yhdistämisperiaatteet. Tässä työssä fuusiossa käytettiin kaikkia sinä vuonna kerättyjä tunnuslukuja ja fuusion toimivuus perustui kerättyjen valvontatietojen muutoksiin vuosien välillä samassa suhteessa klustereiden muutoksiin.

5.4 Volyymi-, panos- ja toimitusvarmuusmuuttuja klusteroinnit

Volyymi-, panos- ja toimitusvarmuusmuuttujien luomille aineistoille tehtiin klusteroinnit ja näitä verrattiin jokaisesta tunnusluvusta tehtyyn yksittäiseen klusterointiin. Teknisten tunnuslukujen jakautuessa edellä mainittuihin kolmeen alueeseen pohdittiin mahdollisuutta löytää jokaista osa-aluetta parhaiten kuvaava tunnusluku. Klusteroimalla parhaiten kuvaavia tunnuslukuja pyrittiin päästä samaan lopputulokseen kaikkien tunnuslukujen klusterointitulosten kanssa.

Parhaiten kuvaava tunnusluku etsittiin klusteroimalla jokaista yksittäistä tunnuslukua omana aineistonaan ja vertailemalla sitä omaan tyyppiaineistonsa, volyymi-, panos-, tai toimitusvarmuusaineiston klusterointituloksiin. Vastaavuudella tarkoitetaan tässä työssä yhtiöiden klusterointitulosten samankaltaisuutta, 100 % tarkoittaessa kaikkien 77 yhtiön olevan samoin klusteroituna kahden aineiston välillä. Klusteroinneissa käytetiin tasavälialoitusta ja maksiminormalisointia. Vertailu toteutettiin visuaalisesti Excelissä. Vertailemalla tunnuslukujen tuloksia aineistoon päädyttiin valitsemaan parhaiten jaottelua vastaava tunnusluku. Klusterointitulokset ja klusterien indeksit ovat esitetty liitteessä E.

5.4.1 Panosmuuttuja-aineiston klusterointi

Panosmuuttujia oli vain kolme kappaletta ja tarkastelu muistutti maakaapeloinnista tehtyä testitarkastelua. Panosmuuttujat kuvasivat käyttöpaikkamäärillä jaettuja johtopituuksia 0,4 kV, 1-70 kV ja 110 kV verkoissa.

Pääsääntöisesti klusterit muodostuivat PJ-verkon pituudesta. Tätä jaottelua KJ- ja SJ-verkot täydensivät ja niiden vaikutus oli selvä. Viisi klusteria (klusterien indeksit 1,2,3,6 ja 8) pääsääntöisesti muodostuivat PJ-verkosta. Lopuista klustereista oli kaksi isompaa klusteria (4 ja 12), joissa 22 yhtiötä sekoittuivat vielä keskenään. Näistä kahdesta klusterista KJ-verkon pituus erotteli klusterit omiksi klustereikseen. Klusterit 1, 2 ja 3 pysyivät muutamia poikkeuksia lukuun ottamatta muuttumattomina KJ-verkon pituuksissa. Yksi klusteri muodostui KJ- ja SJ-verkon johtopituuksista. Yhden yhtiön klustereita oli kolme, joilla yhteiseksi tekijäksi muodostui SJ-verkon pituus. Yhtiöt eivät PJ- ja KJ-pituuksien erotessa muiden yhtiöiden pituuksista päätyneet isompiin klustereihin.

Vertailemalla jokaisesta tunnusluvusta tehtyä klusterointia panosmuuttujien klusterointitulokseen, 0,4 kV jakeluverkon pituus (B1) kuvasi parhaiten panosmuuttujia yhtiöiden ollessa 79 % samoissa klustereissa.

5.4.2 Volyymimuuttuja-aineiston klusterointi

Volyymimuuttuja-aineistosta tehtyjä klusterointituloksia yhdisti jokainen yksittäinen muuttuja absoluuttisesta aineistosta. Volyymimuuttuja aineistolle tunnuslukuja on 25 kappaletta ja toisin kuin panosaineiston kohdalla selvää yhteyttä klusterointituloksen ja yksittäisen tunnusluvun välillä ei ollut. Ainoa tunnusluku ilman käyttäjäpainotusta volyymimuuttuja-aineistossa on suurin siirretty tuntikeskiteho (MWh/h) (A13).

Suurimmalla klusterilla (3) kyseinen tunnusluku A13 tuottaa paljon hajontaa, muilla klustereilla kyseisen luvun arvot ovat vähemmän hajanaisesti. Kuten maakaapeloinnin testauksessakin klusterit vaikuttaisivat muodostuvan eri jännitetasojen keskinäisistä suhteista tasaisesti kaikista tunnusluvuista. Ainoat selkeästi enemmän klustereita yhdistävät tunnusluvut ovat liittymien lukumäärä ja aineistossa nämä ovat jaettuna käyttöpaikkamäärillä.

Parhaiten kuvaava tunnusluku yksittäisten muuttujien klusterointeja verrattaessa volyymiaineiston klusterointiin oli tunnusluku A13. Tunnusluku kuvasi volyymimuuttujissa yksittäisiksi klustereiksi jääneet yhtiöt myös yksittäisinä ja muutaman pienemmän klusterin yhteneväisesti. Yhtiöistä 37 täsmäsi keskenään täysin (51 %) ja muutamien kohdalla tuli vielä jaottelua. Esimerkiksi volyymiklusteri 10 jakaantui kolmeen 3 yhtiön ja yhteen yksittäiseen klusteriin. Tunnusluku A13 klusteroi kuitenkin volyymiaineiston klusterit 5, 6, 7, 8 ja 9 tunnusluvun ensimmäiseen klusteriin.

5.4.3 Toimitusvarmuusmuuttuja-aineiston klusterointi

Toimitusvarmuusaineistosta parhaiten aineistoa kuvaavan tunnusluvun etsintä oli visuaalisesti vaikeinta, verrattaessa volyymi- ja panosaineistoihin. Syynä tähän oli tunnuslukujen kasvu volyymi-aineiston 25 tunnusluvusta toimitusvarmuusaineiston 43 tunnuslukuun. Tarkastelu aloitettiin etsimällä yhteisiä tekijöitä absoluuttisista ja käyttäjäpainotetuista luvuista.

Suurimman klusterin (1) ja toiseksi suurimman toimitusvarmuusklusterin (2) kohdalla suurimmat yhtäläisyydet löytyivät KJ-verkon keskeytysluvuista. Näiden perusteella KJ-keskeytyksillä vaikuttaisi oleva suuri vaikutus toimitusvarmuusklusterointiin. Muissakin luvuissa yhteneväisyyksiä esiintyi, mutta hajonta oli suurempaa yhtiöiden välillä verrattaessa suurimpaan klusteriin. KJ-keskeytysluvut eivät yksiselitteisesti selittäneet ryhmiä ja hajontaa tapahtui. Lisäksi klusterilla 9 hajonta oli kaikissa tunnusluvuissa laajempaa kuin muissa klustereissa. Yhtiöitä klusterissa 9 on kolme. Tämä saattaa viitata liian suureen k-arvoon ja eriytymien vaikutukseen.

Klusteri 3 sisälsi 9 yhtiötä ja kyseisillä yhtiöllä siirtämättä jäänyt energia (C70), KJ-verkon keskeytysluvut ja vakiokorvaukset ovat selvästi yhtenevät. Muiden toimitusvarmuuttakuvaavien lukujen kohdalla hajonta on suurempaa. Kaikkia klustereita parhaiten erottelevat luvut olivat KJ-verkon keskeytyslukuja.

Yksittäisten klusterien tarkempi tarkastelu viittasi toimitusvarmuus muuttujissa KJ- ja PJ-verkon energiapainotettujen keskeytyslukujen eroihin. Tarkastelussa tutkittiin yksittäisten klusterien toisistaan eniten eroavia tunnuslukuja. Kyseisiä tunnuslukuja ei käyttöpaikkapainotettu energiapainotuksen takia. Tämän perusteella toimitusvarmuus-aliaineistossa vaikuttaisi yksittäisiin klustereihin korostuvan PJ-verkon keskeytysluvut, vaikka suuremmissa klustereissa korostuivat KJ-verkon keskeytysluvut. Lisäksi sähkömarkkinalain vakiokorjaukset vaihtelivat yhtiöiden välillä. Jeppo Kraft oli yksittäisistä klustereista ainoa, jonka luvut olisivat sopineet isompiin klustereihin ja syytä eriytymiselle ei suoraan tunnusluvuista löydetty. Tämä saattaisi edelleen viitata liian suureen k-arvoon tälle aineistolle.

Hyvin kuvaavaa yksittäisistä tunnusluvuista tehtyä klusterointitulosta ei löydetty kuvaamaan toimitusvarmuus-aineistosta tehtyä klusterointia. Parhaan tuloksen antoi KJ-verkon pikajälleenkytkennät painotettuna energiamäärillä (C15). Kyseisellä tunnusluvulla oli 42 yhtiötä (62 %) vastaavissa klustereissa. Yhtiöt eivät kuitenkaan kahden edellisen aineiston parhaiten kuvaavien tunnuslukujen tavoin olleet selvästi ryhmittäytyneitä, vaan hajanaisesti jakautuneena eri klustereihin.

5.5 Havainnot teknisistä tunnusluvuista

Ilman painotuksia tehtyjä klusterointituloksia vertailtaessa havaittiin eniten selittävimmän ominaisuuden olevan tavalla tai toisella riippuvainen yhtiön koosta. Pääsääntöisesti siirretty energia ja käyttöpaikkamäärät korreloivat positiivisesti keskenään ja selvää erittelyä ei voida näiden kahden välillä tehdä. Tulosten perusteella käyttöpaikkapainotus vaikuttaisi toimivan.

Eri suureita sisältävän ”ABC-aineiston” klusterien muodostuminen on usean muuttujan summa. Klusterin sisällä voi jakautuminen riippua useasta eri muuttujasta, mikä voisi viitata liian vähäiseen k-arvoon tai liian monen muuttujan aiheuttamaan satunnaisuuteen. Suurimman klusterin sisällä klusterin jäseniä yhdistävät volyymimuuttujat, joiden samankaltaisuutta erottelee toimitusvarmuus- ja panosmuuttujat. Toiseksi suurimman klusterin samankaltaisuus vaikuttaisi tulevan toimitusvarmuusmuuttujista ja erottelevat tekijät volyymimuuttujista.

Klusterit vaikuttaisivat muodostuvan moniulotteisesti samankaltaisuuden perusteella ja klusterien välillä ei välttämättä ole globaalisti vallitsevaa samankaltaisuutta. Tulosten perusteella aineistoa yhdistävät tekijät ovat johtopituudet, KJ-verkon keskeytysluvut, liittymien lukumäärät ja suurin siirretty tuntikeskiteho. Kohtalaisen hyvä määrä tulkittavia muuttujia aineistolla todettiin olevan 25 kappaletta. Tällöin pystytään jo visuaalisesti tulkitsemaan tuloksia kohtuullisessa ajassa ja muodostamaan helposti kokonaiskuvan.

Volyymi-, panos- ja toimitusvarmuusmuuttuja-aineistojen klusterointituloksien ja jokaisen yksittäisen tunnusluvun klusterointien perusteella valittiin yksittäisistä tunnusluvuista kolme tunnuslukua klusteroitavaksi. Taulukko 7 on koottuna aineistoja kuvaavat luvut.

Taulukko 7. *Volyymi-, panos- ja toimitusvarmuusaineiston parhaiten kuvaavin tunnusluku ja sen vastaavuus omaan aineistoon.*

	Volyymiaineisto	**Panosaineisto**	**Toimitusvarmuus-aineisto**
Parhaiten kuvaava tunnusluku	Suurin siirretty tuntikeskiteho (A13)	0,4 kV jakeluverkon pituus (B1)	KJ-verkon pikajälleenkytkennät (C15)
Vastaavuus aineistoon	51 %	79 %	62 %

Yhdistämällä parhaiten kuvaavat tunnusluvut tTaulukko 7 ja klusteroimalla näitä kolmea tunnuslukua saatiin 17 yhtiötä (22 %) samoihin klustereihin verrattuna koko aineistosta tehtyyn klusterointiin. Vastaavuudet olivat joko suurimmasta kahdesta klusterista tai yksittäisten yhtiöiden klustereita. Tämän perusteella teknisiä tunnuslukuja ei voida vähentää vain kolmeen muuttujaan, ilman merkittävää muutosta aineiston ominaisuuksissa. Käytännössä valitut tunnusluvut eivät yksinään kuvaa kuin vähemmistöä kaikista yhtiöistä.

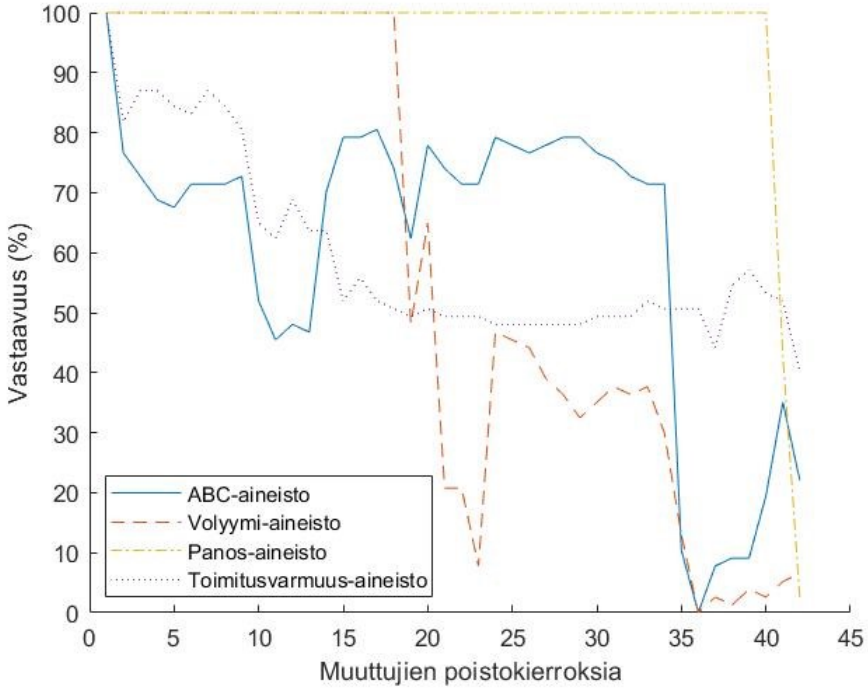
5.6 ABC-aineiston muuttujien vähentäminen ja vastaavuus alkuperäiseen aineistoon

Todettaessa A-, B- ja C-aineistojen erillään vastaavan yhdistettyä ”ABC-aineistoa” heikosti, päätettiin kokeilla, kuinka monta muuttujaa voidaan aineistosta poistaa, ennen kuin tulosten samankaltaisuus ”ABC-aineistoon” verrattuna laskee samalla tasolle (22 %), kuin kolmella muuttujalla luvussa 5.5 todettiin. Tarkoituksena on poistaa tunnusluku kerrallaan muuttujia, kunnes jäljellä ovat enää A-, B- ja C-aineistojen parhaiten kuvaavat muuttujat. Aineistojen muuttujamäärien erotessa päädyttiin poistamaan C-aineiston muuttujia, kunnes niitä on yhtä paljon kuin A-aineistolla. Tämän jälkeen molempia poistetaan, kunnes päädytään enää 3 muuttujaan jokaisesta aineistosta. Poistamisessa pyritään aloittamaan vähiten merkitsevistä muuttujista ja kierros kierrokselta merkitsevyys kasvaa. Hypoteesina oletetaan vastaavuuden laskevan, kunnes päädytään kolmella muuttujalla 22 % verrattuna koko aineistoon.

Muuttujien oletetaan yksinkertaistuksen takia olevan toisistaan riippumattomia. Klusteroinnin ollessa riippuvainen aineistosta ja jokaisesta muuttujasta, on jokainen aineistoon lisättävä tai vähennettävä muuttuja valittava joka kierroksella tarkkaan. Kuten klusteroinnin tavoitefunktion minimin löytämisessä, myös poistettavilla muuttujilla on paras järjestys. Aineiston muuttujien poisto tarvitsisi ajaa jokaisella permutaatiolla ja valita parhaimman vastaavuuden saanut yhdistelmä. Muuttujat poistetaan klusteroinnin kannalta tärkeysjärjestyksessä aloittaen toimitusvarmuusmuuttujista. Muuttujia vähennetään, kunnes toimitusvarmuusmuuttujia on enää 25 eli saman verran kuin volyymimuuttujia. Tämän jälkeen aloitetaan toimitusvarmuusmuuttujien lisäksi poistamaan myös volyymimuuttujia. Lopulta päästään jokaisen aineiston kohdalla 3 muuttujaan ja poistetaan vielä kahden kierroksen ajan muiden lisäksi myös panosmuuttujia. Klusterointitulosten vastaavuus olisi parhaimmassa tapauksessa laskeva funktio, jonka viimeinen arvo saisi parhaimman vastaavuuden pienimmällä muuttujien määrällä.

Tässä työssä pienin muuttujien lukumäärä valittiin kolmeksi. Erilaisia kombinaatioita saadaan aineistolla yhteensä $1{,}2800 \cdot 10^{14}$. Parasta järjestystä ei lähdetty etsimään ja poistettavien muuttujien valinnat suoritettiin ajan säästämiseksi loogisilla valinnoilla. Toimitusvarmuusmuuttujista poistettiin ensimmäiseksi SJ-verkon keskeytyksiä, sillä kyseistä verkkoa ei kaikilla ole ja se on pääsääntöisesti vähävikaista. Valinnoilla pyrittiin pitämään KJ-keskeytyksiä mukana mahdollisimman pitkään. Viimeinen jäljelle jäänyt muuttuja oli KJ-verkon pikajälleenkytkennät (C15). Volyymi-aineiston kohdalla pyrittiin jättämään kaikkien jännitetasojen verkkopalveluasiakkaille siirretty energia aineistoon mahdollisimman pitkään. Tunnusluku A13 jäi viimeiseen vaiheeseen.

Vertailu tehtiin automaattisesti Matlabilla ja esimerkiksi eriytymien kohdalla vastaavuus syntyy vain klusterin indeksin ollessa samat. Eriytymienkin tulee olla samalla indeksillä, jotta vastaavuus löytyy. Kuvassa 5 on esitelty eräällä valinnalla saadut vastaavuudet alkuperäisiin omiin aineistoihinsa muuttujien lukumäärän vähentyessä.



Kuva 5. *A-, B-, C- ja ABC-aineistojen vastaavuudet täyteen aineistoon verrattuna, kun muuttujia poistetaan vähitellen.*

Kuvassa 5 volyymi- ja panosaineistojen muuttujien vähentäminen alkaa vasta 18 toimitusvarmuusmuuttujan jälkeen ja panosmuuttujien vähentäminen vasta 40 toimitusvarmuusmuuttujan ja 12 volyymimuuttujan vähentämisen jälkeen. ABC-aineiston vastaavuus vastaa melko hyvin toimitusvarmuus-aineiston käyrää. Kuvasta nähdään myös A-, ja C-aineistojen muuttujien vähenemisen vaikutukset kokonaistulosten vastaavuuteen. Volyymiaineiston laskiessa 18-22 muuttujalla, kokonaisvastaavuudessa tulee vastaavanlainen pudotus ja pieni nousu.

Tämän perusteella aineistosta voitaisiin poistaa yhteensä 50 tunnuslukua ja 71 % (55 yhtiötä) asettuu edelleen samoihin klustereihin kuin ”ABC-aineiston” klusteroinneissa. Vastaavasti 6 muuttujalla saataisiin 35 % (27 yhtiötä) vastaamaan 156 muuttujan aineistoa. Kolmella muuttujalla A13, B1 ja C15 vastaavuus oli enää 22 % (17 yhtiötä). Poistamisjärjestys vaikuttaa klusterointituloksiin ja sitä muuttamalla ja testaamalla päästäisiin matemaattisesti lähemmäksi klusterointiin eniten vaikuttavia muuttujia.

5.7 Kohtuullisen hinnoittelun laskelmien klusterointi

Kohtuullisen hinnoittelun laskelmia klusteroitiin vuoden 2017 tiedoista. Valvontatietojärjestelmästä kohtuullisen hinnoittelun tiedot saadaan vuodesta 2005 lähtien. Kuten muissakin aineistoissa, myös kohtuullisen hinnoittelun laskelmissa korostuvat yhtiöiden kokoerot. Tässä klusteroinnissa kohtuullisen hinnoittelun tunnusluvut jaettiin käyttöpaik-

kamäärillä. Tarkoituksena oli alun perin klusteroida vuodesta 2006 lähtien. Tästä luovuttiin kokoerojen kompensoinnissa olleiden käytännön ongelmien takia. Kohtuullisen hinnoittelun laskelmat jouduttiin ottamaan valvontajärjestelmästä ja samalla ei saatu tuotua tietoja kyseisten yhtiöiden asiakasmääristä. Käytössä oleva aineisto eroaa yhtiönimiltään ja indekseiltään teknisiin tunnuslukuihin nähden, jolloin kompensointia ei pystytä toteuttamaan automaattisesti ilman käsityötä.

Klusterointia varten tehty kokoerojen kompensointi toteutettiin käsin käyttämällä käyttöpaikkalukumääriä teknisisistä tunnusluvuista ja klusterointia rajoitettiin vuoteen 2017. Klusterointi suoritettiin k-medoidsilla ja PAM-menetelmällä ilman aloitusvektorien syöttöä antaen k-means++:n luoda aloitusvektorit 500 replikaatille. Aineisto normalisoitiin käyttäen euklidinista normalisointia.

Kohtuullisen hinnoittelun laskelmissa verrattiin korrelaatiomatriisin avulla luotua 9 muuttujan aineistoa ja täyttä kohtuullisen hinnoittelun laskelmaa vuodelta 2017. Tarkastelussa oltiin mahdollisimman kriittisiä. K-medoidsin muuttaessa klusterien nimeämistä satunnaisesti, tarkastelussa klusterin indeksi vastasi sitä indeksiä, jonka yhtiöt vastasivat enemmistönä parhaiten toisiaan. Yksittäisten havaintojen kohdalla todettiin vastaavuus vain, mikäli havainto oli molemmissa klusteroinneissa yksittäinen. Taulukko 8 on esitelty osa huomiosta täyden ja pienen aineiston välillä.

Taulukko 8. *Kohtuullisen hinnoittelun täyden ja pienen aineiston klusterointitulosten vertailu.*

Aineisto	Suurimman klusterin jäsenten vastaavuus (kpl)	Suurimman klusterin jäseniä (kpl)	Jäseniä samoissa klustereissa (%)	Jäseniä samoissa klustereissa (kpl)	Yhden jäsenen klustereita (kpl)
Täysi aineisto	50	59	-	77	9
Pieni aineisto	21	52	91	65	8

Molempiin aineistoihin muodostui yksi huomattavan suuri klusteri. Täydellä aineistolla suurimmassa klusterissa oli 59 yhtiötä, verrattuna pienemmän aineiston suurimman klusterin 52 yhtiöön. Ero muodostuu pienemmän aineiston vähäisemmällä muuttujamäärällä ulottuvuuksien aiheuttaman satunnaisuuden vähentäessä eroja. Molempiin syntyi yhden alkion klustereita. Täyteen aineistoon 9 kappaletta ja pienempään aineistoon 8 kappaletta. Kohtuullisen hinnoittelun laskelmissa verrattaessa täyttä aineistoa ja pienempää aineistoa, pienempi aineisto täsmäsi 91 %. Suurimpaan klusteriin kuuluvat yhtiöt klusteroitiin vielä uudestaan. Tällöin klusterointitulokset eivät vastanneet enää yhtä hyvin. Huomioitaessa vain molemmissa esiintyneet yhtiöt vastaavuus 50 yhtiöstä oli 21 yhtiötä (42 %).

Täydellä kohtuullisen hinnoittelun laskelmilla useat keskenään korreloivat muuttujat ovat mukana aineistossa ja siten vaikuttaisivat muuttavan tuloksia. Pienen aineiston kohdalla on esimerkiksi suurimman klusterin toisella klusteroinnilla homogeenisempi jakauma yhtiöillä klustereihin, kuin koko aineistolla, jolla suurin osa yhtiöistä oli edelleen samassa

klusterissa. Aineistojen suurimpien klustereiden koot eivät olleet samoja. Kohtuullisen hinnoittelun laskelmissa selittävät tekijät vaikuttaisivat olevan pienempään aineistoon valitut muuttujat.

Käyttöpaikkapainotuksella tehtyjen teknisten tunnuslukujen ja kohtuullisen hinnoittelun laskelmien klusterointituloksia verrattiin myös toisiinsa. Tuloksissa ei ollut selviä yhtäläisyyksiä teknisten tunnuslukujen käyttöpaikkapainotetuille koko aineiston ja ”ABC-aineiston” välillä. Kohtuullisen hinnoittelun laskelman tuloksia verrattiin myös volyymi-, panos- ja toimitusvarmuusmuuttujien klusterointeihin. Myöskään näiden kesken ei muodostunut yhtäläisyyksiä klustereissa.

5.8 Universaalin aineiston hyödyntäminen klusteroinnissa

Keskeytyslukuja ja ilmatieteen laitoksen aineistoja yhdistettiin ja klusteroitiin. Tarkoituksena on demonstroida työkalun käytettävyyttä valvontatietojen yhdistämistä ulkoisiin tietolähteisiin. Hypoteesina on, että tuulisten alueiden tunnusluvut korreloivat positiivisesti vikataajuuteen ja ryhmittyvät siten samoihin klustereihin. Tätä korrelaatiota vähentäisi tai poistaisi maakaapelointiaste, joka jaottelisi samantyyppisissä toimintaympäristöissä olevia yhtiöitä

Klusteroitavaksi aineistoksi valittiin ilmatieteenlaitoksen tiedoista kerätyt vuoden 2017 sateisten päivien (päivän sadekertymä > 1 mm) ja tuulisten päivien lukumäärät (päivän aikana yhden tunnin keskimääräinen tuulen nopeus > 10 m/s), sekä vuoden aikana satanut sademäärä kyseisten yhtiöiden toimintapaikkakunnista valittuna. Aineistoon lisättiin teknisistä tunnusluvuista KJ-verkon keskeytysten lukumäärät, jaettuna KJ-verkon johtopituudella (kpl/km) eli vikataajuudella.

Ilmatieteenlaitoksen tietojen saatavuudessa oli ongelmia, jolloin parikymmentä yhtiötä sai lähimmän sääaseman tiedot ja monella yhtiöllä sääasema oli sama. Tiedot on kerätty yhdeltä alueella sijaitsevalta sääasemalta, jolloin laajojen jakeluverkkojen kohdalla tiedot eivät välttämättä kuvaa koko alueella vallinneita olosuhteita. Sääasemien valinnoilla on pyritty kuvaamaan parhaiten yhtiön toimintaympäristön sääoloja.

Klusterointimenetelmänä käytettiin k-meansia, euklidinista potenssietäisyyttä ja aineiston normalisoinnissa maksiminormalisointia. Virheen neliösumma -käyrän perusteella päädyttiin käyttämään k-arvona 8. Kyseisellä k-arvolla ei syntynyt yksittäisiä klustereita, mutta sitä korkeammilla arvoilla yhden jäsenen klustereita muodostui. Taulukko 9 on esitelty klustereiden absoluuttisista arvoista tehtyjä havaintoja.

Taulukko 9. *Ilmatieteenlaitoksen ja KJ-vikataajuuden luoman aineiston klustereiden ominaisuuksia.*

Klusterin indeksi	yhtiöitä (kpl)	Kaapelointiasteen keskiarvo (%)	Vikataajuus keskiarvo (kpl/km)	Kaapelointiasteen minimiarvo-maksimiarvo (%)	Vikataajuus minimiarvo-maksimiarvo (kpl/km)	Tuulisten päivien keskiarvo vuodessa (kpl)	Tuulisten päivien minimi-maksimi -lukumäärä (kpl)
1	2	31,52	0,15	3,08-59,96	0,10-0,19	25	24-25
2	7	32,87	0,4	18,57-74,08	0,26-0,58	2	0-4
3	7	14,72	0,26	5,57-31,82	0,22-0,29	16	11-20
4	20	47,55	0,14	2,70-95,56	0,02-0,20	2	0-6
5	11	60,32	0,1	1,66-99,75	0,02-0,17	8	3-12
6	14	18,14	0,27	1,48-56,59	0,22-0,35	3	0-7
7	5	31,09	0,45	4,35-75,27	0,36-0,58	13	10-16
8	11	17,73	0,41	2,07-70,44	0,33-0,60	1	0-4

Taulukko 9 mukaisesti saatiin yhtiöiden maakaapelointiasteen avulla luotua kuva KJ-verkon maakaapelointiasteen vaikutuksesta KJ-verkon keskeytyksiin ja tuulisiin päiviin. Klusterin 2 yhtiöiden keskeytystiheys on keskimäärin korkeampi kuin klusterin 4, vaikka molempien tuulisten päivien lukumäärät ovat alhaisia. Tutkimalla maakaapelointiasteita huomataan klusterin 2 kaapelointiasteen keskiarvon olevan 14,7 % korkeampi. Klusterin 3 keskeytystiheys on maakaapelointiasteeseen nähden pieni, vaikka tuulisia päiviä on klusteria 2 enemmän. Alin keskeytystiheys saadaan klusterille 5, jolla on korkein maakaapelointiaste ja tuulisten päivien keskiarvo on kohtalainen.

Vertailemalla vikataajuutta tuulisiin päiviin huomataan positiivista korrelaatiota. Korkea maakaapelointiaste tulosten perusteella vähentäisi vikataajuutta tuulisuuteen verrattuna. Klusterin 3 kohdalla suhteellisen matalalla kaapelointiasteella on päästy kohtalaisen matalaan vikataajuuteen huomioiden tuulisten päivien lukumäärät. Näillä yhtiöillä vaikuttaisi olevan muita keinoja vähentää vikataajuutta kuin maakaapelointiaste. Klusterin 5 havainnosta tulee huomioida mahdollinen virhe, sillä klusterissa 5 yhtiötä sai saman sääaseman tiedot.

Vertailemalla klusteria 6 ja 8 huomataan keskimäärin samoilla maakaapelointiasteilla saatavan erilaiset vikataajuuden arvot. Tutkimalla tarkemmin klusterin numero 8 yhtiöiden tietoja, huomataan yhtiöiden keskimääräisen vikataajuuden olevan toiseksi suurin, vaikka tuulisia päiviä ei ole kuin keskimäärin yksi. Klusterista 8 tehdyn havainnon taustalla voi olla heikko verkko, vikojen johtuvan muusta kuin käytetyistä tuulisuustiedoista tai tuulisuustietojen virheellinen kuvaus yhtiön toimintaympäristöstä.

5.9 Työkalun soveltuvuus ja käytettävyys

Työkalu soveltuu yleisesti kaikkeen numeerisesti tilastoituun tietoon ja sen hyödynnettävyys tarkemmin määriteltyihin tarkasteluihin on hyvä. Tässä työssä käytettiin pääsään-

töisesti teknisiä tunnuslukuja ja kohtuullisen hinnoittelun laskelmia. Vertailemalla pienempien aineistojen klusterointituloksia koko aineistojen tuloksiin, voidaan todeta työkalun soveltuvan huomattavasti paremmin pieniin muuttujamääriin. Aineistona voidaan käyttää täysin toisistaan riippumattomia muuttujia.

Klusteroitaessa teknisiä tunnuslukuja ja kohtuullisen hinnoittelun laskelmia todettiin tulosten tulkitsemiseen tarvittavan substanssiosaamista tai aineiston karsimista. Pelkän käytetyn aineiston perusteella ei voida yksikäsitteisesti selittäviä tekijöitä yhtiön koon lisäksi löytää. Esimerkiksi keskeytyslukujen taustalla saattaa olla myrskyisenä vuotena alueella vallinneet tuuliolosuhteet. Tekijän tietäessä esimerkiksi ulkoisten aineistojen kautta lukujen taustalla vallinneista olosuhteista tai syistä, voidaan päästä paremmin selville klusterointitulosten määräytymisestä ja sitä kautta yhtiöiden ominaisuuksista. Laajoilla aineistoilla huomattavalla muuttujamäärällä ei välttämättä yksittäisiä selittäviä tekijöitä löydetä, klusteroinnin huomioidessa jokaisen muuttujan ja yhden muuttujan ero kahden havaintovektorin välillä ei välttämättä riitä muuttamaan klusterointitulosta moniulotteisilla aineistoilla.

Havaintovektorien määrän ollessa alle 100 kappaletta ja yhdellä tai kahdella muuttujalla klusteroinnista saatu hyöty on pieni, ellei olematon verrattuna taulukko-ohjelmistoon. Muuttujien lukumäärän kasvaessa analysoinnin haastavuus kasvaa ilman taulukko-ohjelmistoja parempia työkaluja. Havaintovektorien määrän kasvaessa pienillä muuttujamäärillä päästäisiin käytettäviin tuloksiin taulukko-ohjelmistoillakin. Klusteroinnista tehdyt ryhmittelyt voidaan perustella yhtiöiden toimittamien tunnuslukujen matemaattisella jaottelulla. Matlabilla klusteroinnin hyöty saadaankin tekijästä riippumattomista tuloksista, jotka perustuvat täysin matemaattisiin arvoihin ja jokaiselle havaintovektorille sekä klusterille saadaan matemaattiset yksiselitteiset tulokset. Matlabin palauttaessa tulosten lisäksi klusterien keskipistevektorit ja jokaisen havaintovektorin etäisyyden kaikkiin klustereiden keskipistevektoreihin, saadaan näiden avulla arvokasta informaatiota klustereiden muodostumisesta. Informaation avulla voidaan saada vihjeitä yhdistävistä tekijöistä tai rajattua yhdistäviä tekijöitä suuristakin aineistoista.

Alalla yhtenä yleisenä jaotteluna sähkönjakeluverkonhaltijoille on vallinnut jako kolmeen klusteriin: kaupunki-, taajama- ja maaseutuyhtiöihin. Teknisten tunnuslukujen ja testiaineistojen kaapelointiasteiden klusterointien perusteella tämä saattaa olla vuoden 2017 tietojen perusteella vanhentunut jaottelu yhtiöiden kaapeloidessa ilmajohtoja toimitusvarmuusvaatimusten myötä. Kaapelointiin perustuvalla aineistolla tehty klusterien lukumääräanalyysi ehdotti k-arvoksi 4. Tämä voisi viitata yleisen loogisesti tehdyn jaottelun olevan jokseenkin perusteltu, mutta ei täysin yhtiöiden toimintaa selittävä.

Klusteroinnilla voitaisiin saada luotua fyysisiin ominaisuuksiin perustuva jaottelu sähkönjakeluverkonhaltijoille teknisten tunnuslukujen avulla. ”ABC-aineistolla” paras k-arvo kasvoi testiaineistoa suuremmaksi. ”ABC-aineisto” klusteroitiin k-arvolla 12 jolloin muodostui muutamia yhden jäsenen klustereita ja kaksi suurempaa klusteria. Yhtiöiden

luokittelu tulisi olla näiden perusteella vähintään 4 ryhmää ja enintään 12 klusteria ”ABC-aineistolla”. Analysointiin k-arvo 12 toimi, mutta luokitteluun alempi arvo voisi sopia paremmin.

Työkalun testaamisessa maakaapelointiasteella todettiin klusteroinnin tekevän matemaattisen tarkastelun sille määritellystä aineistosta. Liian suppealla aineistolla ei välttämättä päästä haluttuun lopputulokseen, jos aineistosta puuttuu selittäviä tekijöitä. Maakaapelointiasteen klusteroinnissa oltaisiin saatettu päästä enemmän fyysisiä yhtiöitä kuvaaviin tuloksiin lisäämällä aineistoon johtopituudet tai suhteuttamalla prosenttiluvut johtopituuksiin. SJ-verkkoa ei kaikilla ole, jolloin sen kaapelointiaste on 0. Jos SJ-verkon kaapelointiasteella on suuri painoarvo, eivät klusterointitulokset kuvaisi fyysisiä ominaisuuksia. Ilman klusteroitavan aineiston muokkaamista samaan johtopäätökseen päästään myös ulkoisten aineistojen ja tekijän substanssiosaamisen perusteella.

Aineisto voidaan valita eri lähteistä. Luvussa 5.7 käytettiin hyväksi ilmatieteenlaitoksen säätietoja, mutta tämän tilalla oltaisiin voitu käyttää mitä vain numeerista aineistoa rivimäärän vastatessa havaintovektoreiden määriä. Klustereiden muodostumisen taustatekijöitä tutkittiin hyödyntämällä aineiston ulkopuolista tietoa. Selittäviä tekijöitä klustereiden muodostumiseen löydettiin maakaapelointiasteesta, joka oletettiin selittäväksi tekijäksi ennen klusterointia. Tarkastelun perusteella todettiin klusteroinnin soveltuvan eri lähteistä olevien aineistojen väliseen tutkimukseen ja hypoteesimaiseen tarkasteluun.

Vaikka itse klusterointi on tekijästä riippumaton, tulosten tulkinta ja aineiston määrittely eivät ole. Tärkein klusteroinnissa vaikuttava tekijä on aineiston valinta ja ymmärtäminen. Kuten luvun 3 johdannossa todettiin (Jain 2010), tulisi aineistosta tarkastaa, onko sieltä löydettävissä jakautumista. Jos tarkastelua ei voida tehdä etukäteen, tulisi tuloksia tutkiessa pyrkiä löytämään loogisia perusteluita tuloksille. Tulokset voivat olla satunnaisuuden tuotosta ja näennäisesti yhdistävät tekijät eivät ole todellisia.

Mahdollisuuksien mukaan aineistosta tulisi poistaa tai painottaa pois jo havaitut yhteiset luokittelevat ominaisuudet, jotta seuraavan tason jakoperusteet tulisivat esiin. Yksinkertaisin tapa on poistaa muuttujat, jotka eniten vaikuttavat klusterien muodostumiseen. Tässä työssä aineistossa yhtiöiden koko näkyy suurimmassa osassa tunnuslukuja ja muuttujien poistaminen ei ollut haluttu tapa löytää uusia yhdistäviä tekijöitä. Turvallisin tapa saada riippumattomia tuloksia on jättää painotukset ja muokkaukset kokonaan tekemättä tai tehdä niitä mahdollisimman vähän, mikäli muokkausten vaikutuksia ei tarkkaan tiedetä tai ne eivät ole yksiselitteisiä. Painotuksia tehdessä tulisi ne perustella ja tuloksia tarkastella kriittisemmin. Tämän työn klusterointien perusteella yhtiöiden koko on taustalla monessa tunnusluvussa, joka on luonnollista huomioiden tunnuslukujen keräämistarkoituksen. Kokoero yhtiöiden välillä on uuden tiedon löytämiseksi poistettava aineistosta, mutta yhtiön koko ei ole yksikäsitteinen ja käyttöpaikkapainotus ei universaalisti kuvaa sitä. Tästä aiheutuu virhettä, jonka vaikutus oletetaan klusterointituloksissa pieneksi.

6. YHTEENVETO

Työssä tarkasteltiin klusterointityökalujen soveltuvuutta sähkönjakeluverkonhaltijoiden kohtuullisen hinnoittelun määräytymisessä käytettäviin valvontatietoihin. Lisäksi tutustuttiin k-means ja k-medoids menetelmien teoriaan ja käytettäviin parametreihin kirjallisuuden pohjalta. Teoriaa sovellettiin valittuihin vuoden 2017 aineistoihin ja niille määriteltiin klusterointiparametrit molemmille menetelmille. Työssä testattiin k-means ja k-medoids klusterointimenetelmien toimivuus Matlabillä tehdyllä työkalulla hyödyntäen Matlabin sisäisiä funktioita. Hyödynnettävyys muutamaan keskeiseen viranomaisvalvonnassa käytettyjen aineistojen analysoimiseen testattiin. Aineistot muodostuvat monipuolisesti useista eri tekijöistä riippumatta täysin yhdestäkään muuttujasta. Tutkimuksessa huomattiin yhtiöiden koon vaikuttavan monien lukujen taustalla ja ilman kokoerojen huomioimista ryhmät muodostuivat vain koon mukaan.

Klusteroinnin tulokset ovat käyttäjästä riippumattomia ja perustuvat puhtaasti matemaattisiin ominaisuuksiin. Klusterointi itsessään ei ota kantaa fyysisiin tekijöihin aineiston taustalla, vaan fyysiset tekijät tulevat esille tarkasteltaessa tuloksia ja käytettyä aineistoa. Klusteroinnin käyttäjän on otettava aineisto huomioon ja ymmärrettävä klusterointituloksen merkitys. Aineisto tulisi ymmärtää tarpeeksi hyvin jo klusterointimenetelmää valittaessa.

K-arvon löytäminen on vaikea ja tärkeä vaihe osittelevassa klusteroinnissa. K-arvon eli ryhmien määrän löytämiseksi klusterointia voidaan tehdä eri k-arvoilla ja löytää sitä kautta aineistoon toimiva k-arvo. Aineistoon kannattaa tehdä määräanalyysejä ja virheen neliösummakäyrä tuomaan lisätietoja tehtävään sopivasta k-arvosta. Eri k-arvoilla tehtyjä klusterointeja vertailemalla pystytään löytämään rajatapauksia ja eriytymiä. Mikäli aineisto ja sen käyttäytyminen tunnetaan tarpeeksi hyvin, voidaan ryhmistä löytää poikkeavia yksilöitä ja näin saada itse ryhmästä uutta tietoa.

Työkalua testattiin maakaapelointiasteilla ja aineistoon tehdyn k-arvon määrittelyiden perusteella maakaapelointiasteilla muodostuisi neljä klusteria. Klusterien yhdistävien ja erottelevien tekijöiden löytäminen oli helppo nähdä verrattaessa tuloksia absoluuttisiin arvoihin. Käytettyjen tunnuslukujen kuvatessa samaa ilmiötä, ei tulosten analysoinnilla saatu tuotettua uutta tietoa yhtiöistä. Alalla yleisesti käytetty staattinen jako kolmeen ryhmään maakaapelointiasteen perusteella voi olla liian pieni ryhmämäärä. Ottaen huomioon lainsäädännöstä johtuvat huomattavat muutokset yhtiöiden toiminnassa, sähköjakeluverkonhaltijat voitaisiin jakaa fyysisten ominaisuuksien perusteella klusteroinnilla paremmin kuvaamaan vuosittain muuttuvaa todellista tilannetta.

Yksittäisiä koko aineistoa kuvaavia tunnuslukuja ei löytynyt teknisistä tunnusluvuista. Silti parhaiten klustereita selittäviä tekijöitä olivat KJ-verkon keskeytysluvut, suurin siirretty tuntikeskiteho, liittymien lukumäärä ja johtopituudet. Kyseiset luvut ovat jaettuna käyttöpaikkamäärillä. Jakeluverkonhaltijat jakautuvat ilman kokoerojen kompensointia kahteen isoon klusteriin sisältäen noin 70 % kaikista yhtiöistä. Näiden yhtiöiden tunnusluvut ja tulokset ovat samankaltaisia. Käyttäjämääräpainotuksella 61 yhtiötä (79 %) muodostivat kaksi suurinta klusteria.

Kohtuullisen hinnoittelun laskelmien kohdalla hyödyntämällä korrelaatiomatriisia pystyttiin 9 muuttujalla toistamaan 45 muuttujan klusterointitulokset. Kohtuullisen hinnoittelun laskelmissa pienempään aineistoon valitut muuttujat vaikuttaisivat olevan täyden aineiston klusterointitulosten taustalla. Teknisten tunnuslukujen ja kohtuullisen hinnoittelun laskelmien välillä ei löydetty suoraan klusterointituloksista yhteisiä klustereita.

Työkalun toimivuutta takautuvan fuusion kautta tehtäviin tarkasteluihin testattiin lyhyesti. Klusterointitulokset muuttuivat vuosien välillä kerättyjen tunnuslukujen vaihtuessa. Takautuvaa vertailua nykyisiin yhtiöihin pystytään tämän perusteella tekemään ja tunnuslukujen tutkimista vuosien vaihdellessa.

Universaalin aineiston hyödyntämistä testattiin käyttämällä ilmatieteenlaitoksen tuulisuus ja sadetietoja, yhdistettynä teknisisistä tunnusluvuista saatuihin keskeytystiheys tietoihin. Hypoteesina oli klusterien muodostuvan tuulisuuden ja keskeytystietojen perusteella. Tästä eriytyisi ne yhtiöt, joilla maakaapelointiasteen ollessa korkea, tuulisuus ei vaikuttaisi keskeytysten nousuun.

Tarkastelujen ja analysointien perusteella klusterointia voidaan hyödyntää isojen aineistojen kohdalla tuomaan esille vihjeitä yhteyksistä ja pienien aineistojen kohdalla tarkempaa tarkastelua. Aineistona voidaan k-means ja k-medoids menetelmillä käyttää kaikkea numeerista dataa.

Pienillä 1-3 muuttujan ja alle 100 havaintovektorin aineistoilla Excelin avulla päästään samaan lopputulokseen klusteroinnin kanssa. Klusterointi kuitenkin helpottaa analysointia ja tuo lisäinformaatiota, kuten havaintovektorien etäisyyksiä klustereiden keskipisteistä. Matlabilla toteutetun klusteroinnin keskeinen hyöty on matemaattinen laskentakapasiteetti, jonka avulla voidaan käyttää suuria data-aineistoja sekä paljon ulottuvuuksia.

Tässä työssä tehty tarkastelu on hyvin suppea ja käsittää vain yhden klusterointimenetelmän ja sen variaation. Klusterointituloksia tarvitsee tulkita käytetty aineisto huomioiden ja pohtia aineiston soveltuvuutta klusterointiin. Klusteroinnin onnistuminen on helppo havaita keinotekoisissa aineistoissa, joissa tiedetään aineiston soveltuvuus ja mahdolliset parametrit. Oikean maailman aineistosta ei voida välttämättä tulosten oikeellisuutta tietää ilman substanssiosaamista ja tarkempaa analysointia.

Klusterointi on tämän työn perusteella oiva apuväline suuren datamäärän analysointiin. Näin ollen tutkimusta voisi jatkaa vertailemalla tuloksia myös muihin aineistoihin, kuten esimerkiksi ilmatieteenlaitoksen tuottamiin säätietoihin tai viranomaisen keräämiin tietoihin, kuten sähkönjakeluverkonhaltijoiden kehittämissuunnitelmiin tai rakennetietoihin.

Klusteroinnilla löydettiin aineistosta eriytymiä ja tätä ominaisuutta voitaisiin hyödyntää mahdollisesti tietojen tarkastamisessa virheen etsintätyökaluna myös muihin viraston keräämiin aineistoihin. Työkalun rakentaminen omaksi itsenäiseksi ohjelmistoksi riippumattomaksi Matlabin asennuksesta on edellytys laajempaan käyttöönottoon virastolla.

Työssä kehitettyä klusterointityökalua voitaisiin myös edelleen kehittää, testaamalla esimerkiksi muita klusterointimenetelmiä. Painotus toisella muuttujalla tai menetelmällä käyttöpaikkamäärällä jakamisen sijaan saattaisi parantaa tuloksia. Usean vuoden tarkasteluissa käyttämällä keskeytyslukuja, pystyttäisiin esimerkiksi tarkastelemaan mihin yhtiöihin myrskyjen vaikutukset näkyvät ja mihin ei. Aineiston tutkimiseen voitaisiin testata faktorointimenetelmiä, joita useasti käytetään tuomaan muuttujien välisiä yhteyksiä esiin ja pienentämään ulottuvuuksien lukumäärää menettämättä kuitenkaan paljoa informaatiota verrattuna muuttujien poistamiseen.

LÄHTEET

Assent, I. (2015). Efficient Density-Based Subspace Clustering in High Dimensions. Springer-Verlag Berlin Heidelberg. CHDD 2012, LNCS 7627, S. 34-49. DOI: 10.1007/978-3-662-48577-4_3

Bouldin, D. Davies, D. (1979). A Cluster Separation Measure. IEEE. Transactions On Pattern Analysis And Machine Intelligence. Vol PamI-1, no. 2. S. 224-227.

Bouveyron, C. Brunet-Saumard, C. (2012). Model-based clustering of high-dimensional data: A review. Elsevier B.V. Computational Statistics and Data Analysis 71 (2014). S. 52–78. DOI:10.1016/j.csda.2012.12.008

Bora, D. Gupta, A. (2014). Effect of Different Distance Measures on the Performance of K-Means Algorithm: AnExperimental Study in Matlab. International Journal of Computer Science and Information Technologies, Vol. 5. S. 2501-2506.

Celebi, M,E. Kingravi, H, A. Vela, P ,A. (2012). A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications 40 (2013) S. 200–210.

Chicco, G. Napoli, R. Piglione, F (2006). Comparisons Among Clustering Techniques for Electricity Customer Classification. IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 21, NO. 2, S. 933-940.

Cui, Y. Su, J. Ma, L. Liu, Y. Chen, H. Lin, J. Wang, J. Yuan, S. (2016). The Mining Analysis of Distribution Network Operation Efficiency Based on Big Data. China International Conference on Electricity Distribution (CICED 2016). Xi'an, 10-13 Aug, 2016

Ding, C. He, X. (2004). K-means Clustering via Principal Component Analysis. Proceedings of the 21 st International Conference on Machine Learning, Banff, Canada, 2004.

Dormann, C, F. Elith, J. Bacher, S. Buchmann, C. Carl, G. Carré, G. ,Jaime R. Garc í a Marqu é z , Bernd Gruber , Bruno Lafourcade , Pedro J. Leit ã o , Tamara M ü nkem ü ller ,Colin McClean , Patrick E. Osborne , Reineking, B. Schröder, B. Skidmore, A, K. Zurell, D. Lautenbach, S. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36. S. 27–46. DOI: 10.1111/j.1600-0587.2012.07348.x

Energiavirasto (2015). Valvontamenetelmät neljännellä 1.1.2016 – 31.12.2019 ja viidennellä 1.1.2020 – 31.12.2023 valvontajaksolla. Energiavirasto. Saatavilla sähköisesti: (viitattu 18.6.2019) https://energiavirasto.fi/hinnoittelun-valvonta

Energiavirasto (2017). Määräys sähköverkkotoiminnan tunnusluvuista ja niiden julkaisemisesta. Energiavirasto. Dokumenttinumero 2167/002/2016. Saatavilla sähköisesti: (viitattu 18.6.2019) https://energiavirasto.fi/maaraykset

Energiavirasto (2014). Määräys sähkönjakeluverkon kehittämissuunnitelmasta. Energiavirasto. Dokumenttinumero 823/002/2013. Saatavilla sähköisesti: (viitattu 18.6.2019) https://energiavirasto.fi/maaraykset

Energiavirasto (2015). Energiaviraston suositus. Sähkö- ja maakaasuliiketoimintojen laskennallinen ja oikeudellinen eriyttäminen. Energiavirasto. Dokumenttinumero 2449/421/2015

Ernst & Young Oy. (2014). Kohtuullisen tuottoasteen määrittäminen sähkö- ja maakaasuverkkotoimintaan sitoutuneelle pääomalle.

Freddi, D. (2018). AI & Society. Springer Science & Business Media. vol.33(3). S. 393 – 403 DOI: 10.1007/s00146-017-0740-5

Hallituksen esitys, (2013). 20/2013.

Hautamäki, V. Cherednichenko, S. Kärkkäinen,I. Kinnunen,T. Fränti, P. (2005). Improving K-Means by Outlier Removal. Speech and Image Processing Unit, Department of Computer Science, University of Joensuu, P.O. Box 111, FI-80101, Joensuu, Finland

Hirvonen, R. Jauhiainen, M. Kinnunen, M. Lehtinen, H. Lehtisalo, T. Sandholm, P. Seppälä, P. Turkki, J. Turunen, T. Öhman, L. (2006). Sähkönjakelun toimitusvarmuuden kehittäminen. Kauppa- ja teollisuusministeriö. Diaarinumero KTM 1/070/2006

Jain, A. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31. S. 651–666. DOI: 10.1016/j.patrec.2009.09.011

Kalyani, S. Swarup, K.S. (2011). Particle swarm optimization based K-means clustering approach for security assessment in power systems. Expert Systems with Applications 38 S.10839–10846. DOI:10.1016/j.eswa.2011.02.086

Klawonn, F. Höppner, F. Jayram, B. (2015). What are Clusters in High Dimensions and are they Difficult to Find?. Springer-Verlag Berlin Heidelberg. CHDD 2012, LNCS 7627, S. 14–33. DOI: 10.1007/978-3-662-48577-4 2

Koivisto, M. Heine, P. Mellin, I. Lehtonen, M. (2013). Clustering of Connection Points and Load Modeling in Distribution Systems. IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 28, NO. 2,

Korenius, T. Laurikala, J. Juhola, M. (2007). On principal component analysis, cosine and Euclidean measures in information retrieval, Department of Computer Sciences, 33014 University of Tampere, Kanslerinrinne 1, Tampere, Finland

Kuosmanen, T. & M. Kortelainen (2012). Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. Journal of Productivity Analysis 38(1), S. 11-28. DOI: 10.1007/s11123-010-0201-3

Kuosmanen, T. Saastamoinen, A. Keshvari, A. Johnson, A. Parmeter, C. (2014). Yleinen tehostamistavoite sähkön ja maakaasun siirto- ja jakeluverkkotoiminnan valvontamal-

leissa sekä tehostamiskannustimen arviointi: Ehdotus Energiaviraston soveltamien menetelmien kehittämiseksi neljännellä valvontajaksolla 2016 – 2019, Sigma-Hat Economics Oy

Laki Energiavirastosta. (2013). L 870/2013. Saatavilla sähköisesti: (viitattu 8.5.2019) https://www.finlex.fi/fi/laki/ajantasa/2013/20130870

Laki sähkö- ja maakaasumarkkinoiden valvonnasta. (2013). L 590/2013. Saatavilla sähköisesti: (viitattu 8.5.2019) https://www.finlex.fi/fi/laki/ajantasa/2013/20130590

Laki viranomaisten toiminnan julkisuudesta. (1999). L 621/1999. Saatavilla sähköisesti: (viitattu 8.5.2019) https://www.finlex.fi/fi/laki/ajantasa/1999/19990621

Mutanen, A (2018). Improving Electricity Distribution System State Estimation With AMR-based Load Profiles. Tampere: Tampereen teknillinen yliopisto, 2018.

MathWorks inc. (2019a) Verkkosivu saatavissa (viitattu 30.4.2019): https://se.mathworks.com/help/stats/kmeans.html

MathWorks inc. (2019b) Verkkosivu saatavissa (viitattu 30.4.2019): https://se.mathworks.com/help/stats/evalclusters.htmlhttps://se.mathworks.com/help/stats/clustering.evaluation.daviesbouldinevaluation-class.html?s_tid=doc_ta

MathWorks inc (2019c) Verkkosivu saatavissa (viitattu 30.4.2019): https://se.mathworks.com/help/matlab/ref/corrcoef.html

MathWorks inc (2019d) Verkkosivu saatavissa (viitattu 30.4.2019): https://se.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation-class.html

MathWorks inc. (2019e) Verkkosivu saatavissa (viitattu 30.4.2019): https://se.mathworks.com/help/stats/evalclusters.html

MathWorks inc. (2019f) Verkkosivu saatavissa (viitattu 30.4.2019): https://se.mathworks.com/help/stats/kmedoids.html

Olszewski, D. 2012. k-Means Clustering of Asymmetric Data. Springer, Berlin, Heidelberg. Hybrid Artificial Intelligent Systems. HAIS 2012. vol 7208. S. 243-254. Online ISBN: 978-3-642-28942-2

Park, H-S. Jun, C-H. (2008). A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications 36. S. 3336-3341.

Steinley, D. (2006). K-means clustering: A half-century synthesis. British Journal of Mathematical & Statistical Psychology. 59, ProQuest

Strehl, A. Ghosh, J. Mooney, R. (2000). Impact of Similarity Measures on Web-page Clustering. American Association for Artificial Intelligence. Technical Report WS-00-01. S. 58-64.

Sähkömarkkinalaki (2013). L 588/2013 Saatavilla sähköisesti: (viitattu 8.5.2019) https://www.finlex.fi/fi/laki/ajantasa/2013/20130588

Räsänen, T. Voukantsis, D. Niska, H. Karatzas, K. Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data, Applied Energy 87. S. 3538–3545

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Elsevier Science Publishers B.V. Journal of Computational and Applied Matehmatics 20. S. 53-65.

Manjoro, W,S. Dhakar, M. Chaurasia, B,K. (2016). IEEE. Operational analysis of k-medoids and k-means algorithms on noisy data. International Conference on Communication and Signal Processing, April 6-8, 2016, India S. 1500-1505

Steinbach, M. Ertöz, L. Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. Springer, Berlin, Heidelberg. Wille L.T. (eds) New Directions in Statistical Physics. DOI: https://doi.org/10.1007/978-3-662-08968-2_16

Masull, F. Rovetta, S. (2015). (workshop) Clustering High-Dimensional Data. First International Workshop, CHDD 2012 Naples, Italy, May 15, 2012. S. 1-14 DOI: 10.1007/978-3-662-48577-4

Tibshirani, R. Walther, G. Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 63, No. 2. 1369-7412/01/63411. S. 411-423

Tanure, J, E, P, S. Tahan, C, M, V. Marangon Lima, J, W. (2006) Establishing Quality Performance of Distribution Companies Based on Yardstick Regulation. IEEEE Transactions on power systems. vol. 21. no. 3. AUGUST 2006. DOI: 10.1109/TPWRS.2006.879283

Ventä, O. Honkatukia, J. Häkkinen, K. Kettunen, O. Niemelä, M. Airaksinen, M. Vainio, T. (2018). Robotisaation ja automatisaation vaikutukset Suomen kansantalouteen 2030. Valtioneuvoston kanslia, DOI: http://urn.fi/URN:ISBN:978-952-287-484-9

Xu, R. Wunch, D.C. (2009) Clustering. John Wiley & Sons, Inc., Hoboken, New Jersey, ISBN: 978-0-470-27680-8

LIITE A: TEKNISTEN TUNNUSLUKUJEN JAOTTELU A-, B- JA C -AINEISTOIHIN

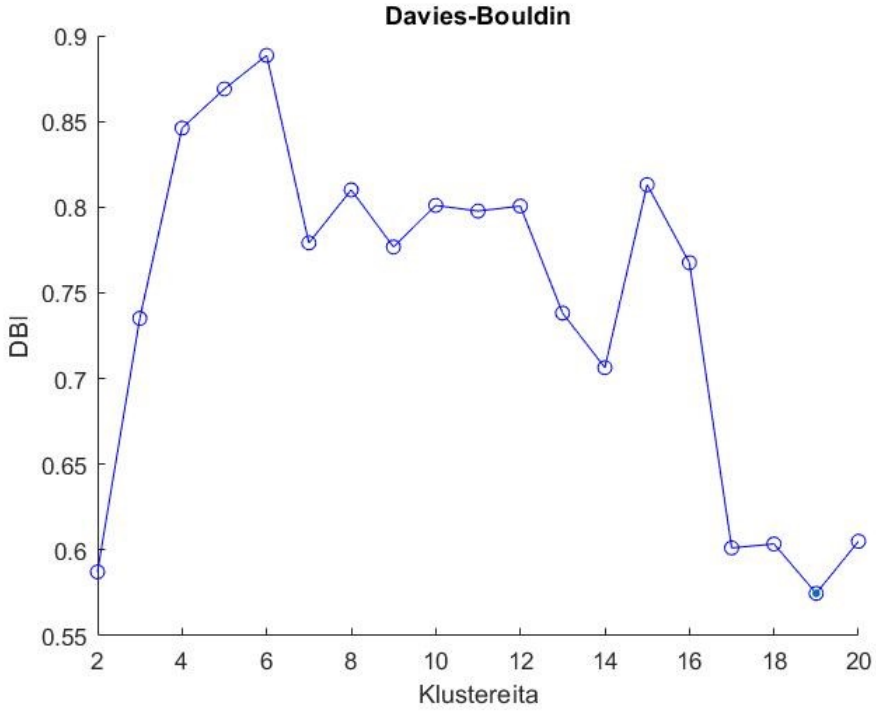
Tummennetut tunnukset ovat jaettu käyttöpaikkamäärillä, mikäli käyttöpaikkapainotusta on käytetty. A-aineistoon kuuluvat kaikki ”Volyymi” -ryhmän tunnusluvut, B-aineistoon kaikki ”Panos” -ryhmän tunnusluvut ja C-aineistoon kaikki ”Toimitusvarmuus” -ryhmän tunnusluvut. Kaikki liitteen tunnusluvut yhdessä muodostavat ABC-aineiston.

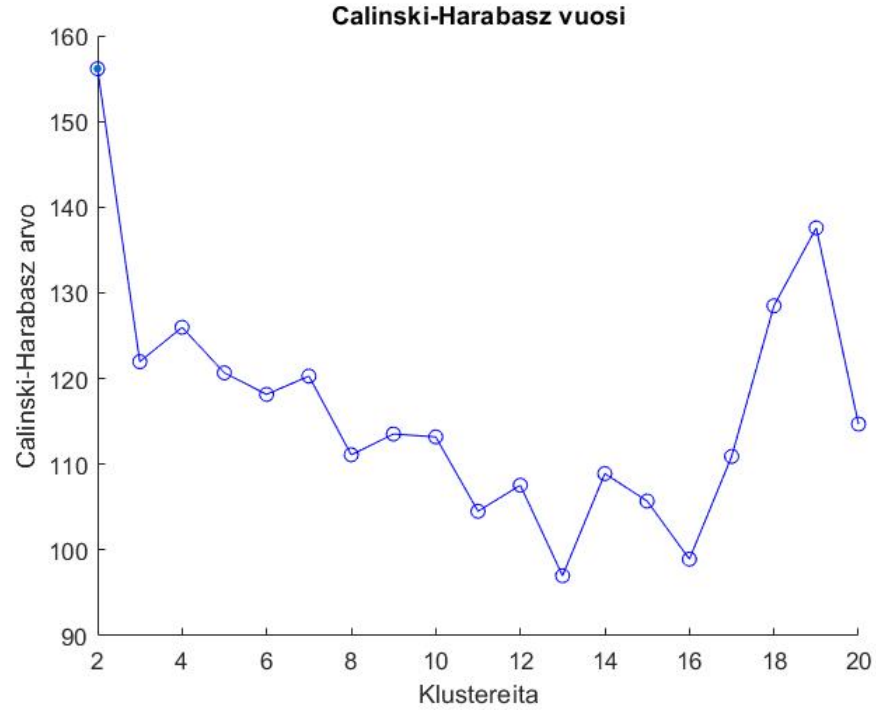
Ryhmä	Tunnus	Teknisten tunnuslukujen nimitys
Volyymi	**A1**	PJ-verkosta verkkopalveluasiakkaille siirretty sähköenergia, GWh
Volyymi	**A2**	KJ-verkosta verkkopalveluasiakkaille siirretty sähköenergia, GWh
Volyymi	**A3**	SJ-verkosta verkkopalveluasiakkaille siirretty sähköenergia, GWh
Volyymi	**A7**	PJ-verkkoon verkkopalveluasiakkailta vastaanotettu sähköenergia, GWh
Volyymi	**A8**	KJ-verkkoon verkkopalveluasiakkailta vastaanotettu sähköenergia, GWh
Volyymi	**A9**	SJ-verkkoon verkkopalveluasiakkailta vastaanotettu sähköenergia, GWh
Volyymi	**A4**	PJ-verkosta toisille verkonhaltijoille siirretty sähköenergia, GWh
Volyymi	**A5**	KJ-verkosta toisille verkonhaltijoille siirretty sähköenergia, GWh
Volyymi	**A6**	SJ-verkosta toisille verkonhaltijoille siirretty sähköenergia, GWh
Volyymi	**A10**	PJ-verkkoon toisilta verkonhaltijoilta vastaanotettu sähköenergia, GWh
Volyymi	**A11**	KJ-verkkoon toisilta verkonhaltijoilta vastaanotettu sähköenergia, GWh
Volyymi	**A12**	SJ-verkkoon toisilta verkonhaltijoilta vastaanotettu sähköenergia, GWh
Volyymi	A13	Suurin verkkoon vastaanotettu tuntikeskiteho, MWh/h
Volyymi	**A50**	PJ-käyttöpaikkojen lukumäärä, kpl
Volyymi	**A51**	KJ-käyttöpaikkojen lukumäärä, kpl
Volyymi	**A52**	SJ-käyttöpaikkojen lukumäärä, kpl
Volyymi	**A40**	PJ-verkon liittymät, kpl
Volyymi	**A41**	KJ-verkon liittymät, kpl
Volyymi	**A42**	SJ-verkon liittymät, kpl
Volyymi	**A23**	PJ-verkko, kulutuksen verkkopalvelusopimukset, kpl
Volyymi	**A24**	PJ-verkko, tuotannon verkkopalvelusopimukset, kpl
Volyymi	**A25**	KJ-verkko, kulutuksen verkkopalvelusopimukset, kpl
Volyymi	**A26**	KJ-verkko, tuotannon verkkopalvelusopimukset, kpl
Volyymi	**A27**	SJ-verkko, kulutuksen verkkopalvelusopimukset, kpl
Volyymi	**A28**	SJ-verkko, tuotannon verkkopalvelusopimukset, kpl
Toimitusvarmuus	**C50**	Omasta verkosta alkunsa saaneiden SJ-verkon odottamattomien pysyvien keskeytysten lukumäärä, kpl
Toimitusvarmuus	**C51**	Toisen verkonhaltijan verkoista alkunsa saaneiden SJ-verkon odottamattomien pysyvien keskeytysten lukumäärä, kpl

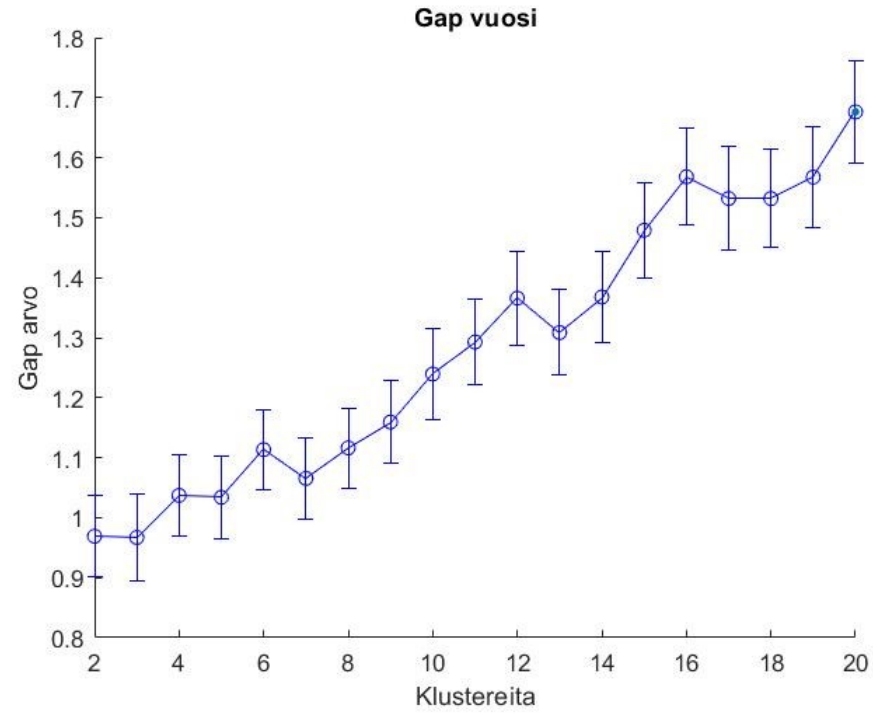
Toimitusvarmuus	**C52**	Verkonhaltijan SJ-verkon suunniteltujen keskeytysten lukumäärä, kpl
Toimitusvarmuus	**C53**	Verkonhaltijan SJ-verkon aikajälleenkytkentöjen lukumäärä, kpl
Toimitusvarmuus	**C54**	Verkonhaltijan SJ-verkon pikajälleenkytkentöjen lukumäärä, kpl
Toimitusvarmuus	C60	Omasta verkosta alkunsa saaneiden SJ-verkon odottamattomien pysyvien keskeytysten keskeytysaika, h/a
Toimitusvarmuus	C61	Toisen verkonhaltijan verkoista alkunsa saaneiden SJ-verkon odottamattomien pysyvien keskeytysten keskeytysaika, h/a
Toimitusvarmuus	C62	Verkonhaltijan SJ-verkon suunniteltujen keskeytysten keskeytysaika, h/a
Toimitusvarmuus	C63	Verkonhaltijan SJ-verkon aikajälleenkytkentöjen keskeytysaika, h/a
Toimitusvarmuus	**C70**	SJ-verkossa siirtämättä jäänyt energia, MWh
Toimitusvarmuus	**C21**	odottamattomat pysyvät keskeytykset, kpl
Toimitusvarmuus	**C22**	suunnitellut keskeytykset, kpl
Toimitusvarmuus	**C23**	pikajälleenkytkennät, kpl
Toimitusvarmuus	**C24**	aikajälleenkytkennät, kpl
Toimitusvarmuus	C10	KJ-verkon odottamattomista pysyvistä keskeytyksistä asiakkaille aiheutunut, vuosienergioilla painotettu keskeytysaika, h/v
Toimitusvarmuus	C11	KJ-verkon odottamattomista pysyvistä keskeytyksistä asiakkaille aiheutunut, vuosienergioilla painotettu keskeytysmäärä, kpl/v
Toimitusvarmuus	C13	KJ-verkon suunnitelluista keskeytyksistä asiakkaille aiheutunut, vuosienergioilla painotettu keskeytysmäärä, kpl/v
Toimitusvarmuus	C14	KJ-verkon aikajälleenkytkennöistä asiakkaille aiheutunut, vuosienergioilla painotettu keskeytysmäärä, kpl
Toimitusvarmuus	C15	KJ-verkon pikajälleenkytkennöistä aiheutunut, vuosienergioilla painotettu keskeytysmäärä, kpl
Toimitusvarmuus	**C20**	PJ-verkossa tapahtuneiden kaikkien odottamattomien ja suunniteltujen keskeytysten vuosittainen lukumäärä, kpl
Toimitusvarmuus	**C80**	odottamattomat pysyvät keskeytykset, kpl
Toimitusvarmuus	**C81**	suunnitellut keskeytykset, kpl
Toimitusvarmuus	C82	PJ-verkon odottamattomista pysyvistä keskeytyksistä aiheutunut vuosienergioilla painotettu vuosittainen keskeytysaika h/v
Toimitusvarmuus	C83	PJ-verkon odottamattomista pysyvistä keskeytyksistä aiheutunut vuosienergioilla painotettu vuosittainen keskeytysmäärä, kpl/v
Toimitusvarmuus	C84	PJ-verkon suunnitelluista pysyvistä keskeytyksistä aiheutunut vuosienergioilla painotettu vuosittainen keskeytysaika, h/v
Toimitusvarmuus	C85	PJ-verkon suunnitelluista pysyvistä keskeytyksistä aiheutunut vuosienergioilla painotettu vuosittainen keskeytysmäärä, kpl/v
Toimitusvarmuus	**-**	Sähkömarkkinalain (588/2013) 100 § mukaisia vakiokorvauksia maksettu, euroa
Toimitusvarmuus	**C30**	a) 12-24 tuntia, euroa
Toimitusvarmuus	**C31**	b) 24-72 tuntia, euroa
Toimitusvarmuus	**C32**	c) 72-120 tuntia, euroa
Toimitusvarmuus	**C34**	d) 120-192 tuntia, euroa
Toimitusvarmuus	**C35**	e) 192-288 tuntia, euroa
Toimitusvarmuus	**C36**	f) yli 288 tuntia, euroa

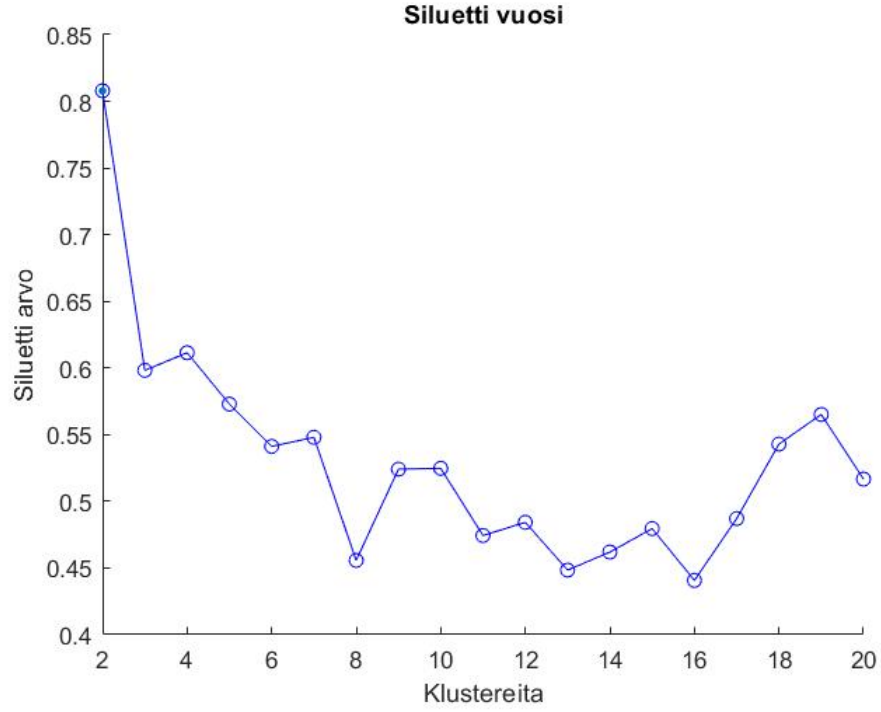
Toimitusvarmuus	**-**	Sähkömarkkinalain (588/2013) 100 § mukaisia vakiokorvauksia saaneiden asiakkaiden lukumäärä, kpl
Toimitusvarmuus	**C37**	a) 12-24 tuntia, kpl
Toimitusvarmuus	**C38**	b) 24-72 tuntia, kpl
Toimitusvarmuus	**C39**	c) 72-120 tuntia, kpl
Toimitusvarmuus	**C41**	d) 120-192 tuntia, kpl
Toimitusvarmuus	**C42**	e) 192-288 tuntia, kpl
Toimitusvarmuus	**C43**	f) yli 288 tuntia, kpl
Toimitusvarmuus	**C90**	Niiden käyttöpaikkojen lukumäärä, joilla sähkömarkkinalain (588/2013) 51 § mukainen toimitusvarmuustaso ei ole täyttynyt
Toimitusvarmuus	**C91**	Asemakaava-alueella sijaitsevien käyttöpaikkojen lukumäärä, joissa sähkömarkkinalain 51 § mukainen toimitusvarmuustaso ei ole täyttynyt, kpl
Toimitusvarmuus	**C92**	Asemakaava-alueen ulkopuolella sijaitsevien käyttöpaikkojen lukumäärä, jossa sähkömarkkinalain 51 § mukainen toimitusvarmuustaso ei ole täyttynyt, kpl
Toimitusvarmuus	**C93**	Niiden käyttöpaikkojen määrä, joissa verkonhaltijan paikallisiin olosuhteisiin määrittämä toimitusvarmuustaso ei ole täyttynyt, kpl
Panos	**B1**	0,4 kV:n verkkopituus
Panos	**B2**	1-70 kV:n verkkopituus
Panos	**B3**	110 kV:n verkkopituus

LIITE B: LUKUMÄÄRÄANALYYSIT TESTI-AINEISTOLLE (VUODEN 2017 PK-, KJ- JA SJ-MAAKAAPELOINTIASTEET)



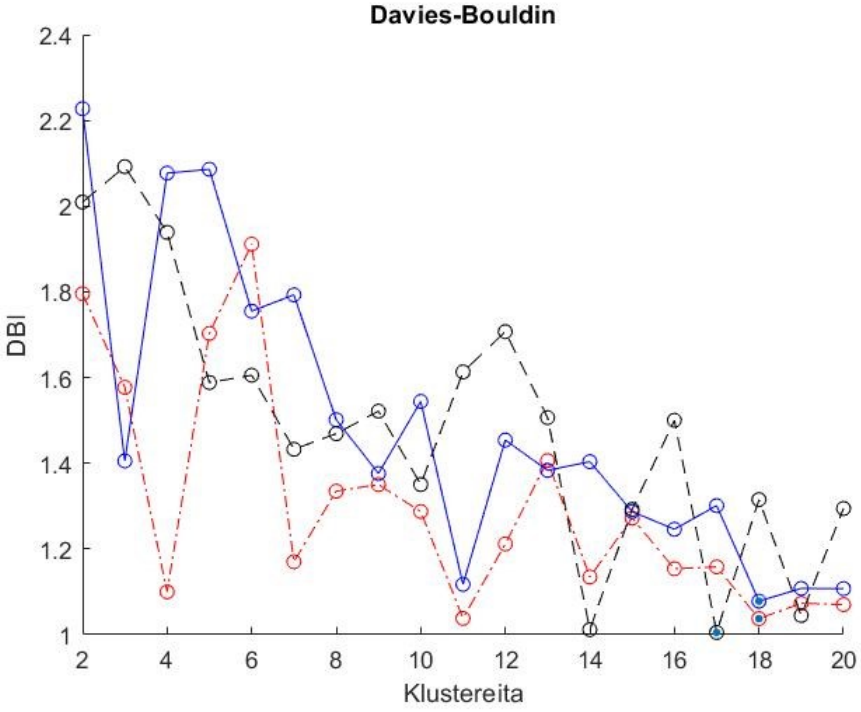


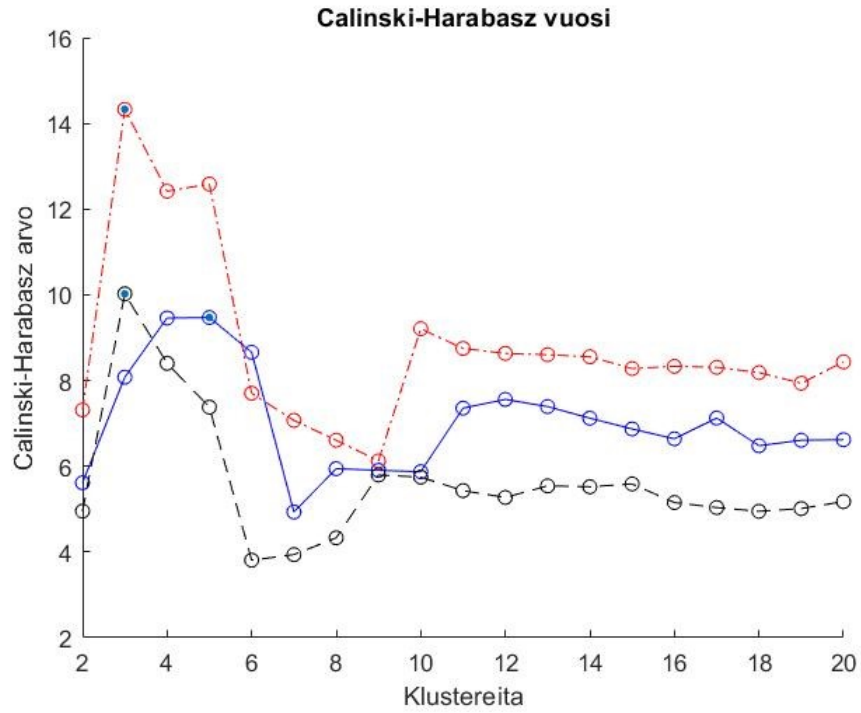

Gap vuosi
Gap arvo
Klustereita
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
2
4
6
8
10
12
14
16
18
20

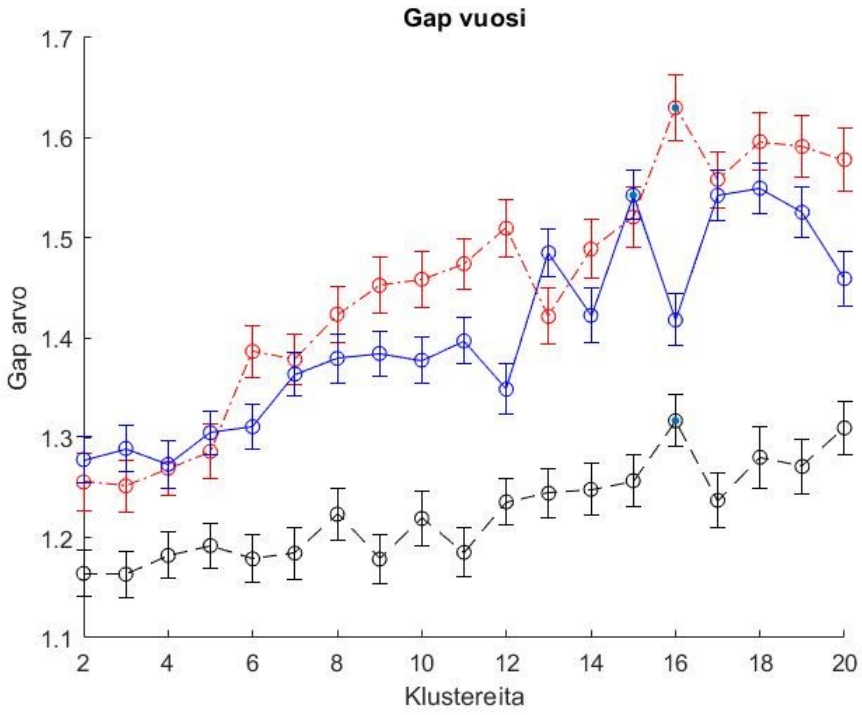

Siluetti vuosi
Siluetti arvo
Klustereita
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
2
4
6
8
10
12
14
16
18
20

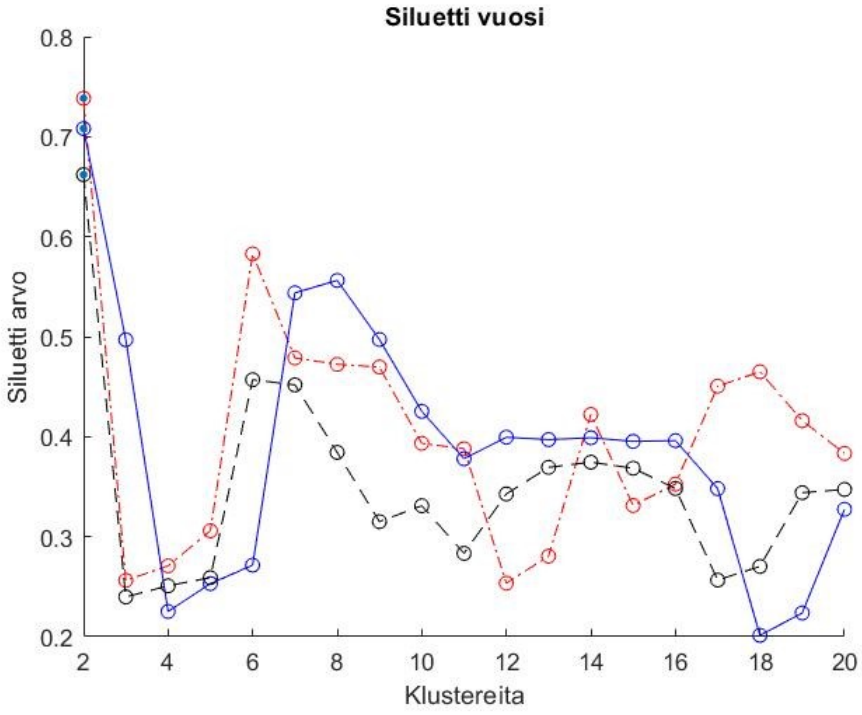
LIITE C: LUKUMÄÄRÄANALYYSIT TEKNISISTÄ TUNNUSLUVUISTA MAKSIMINORMALISOINNILLA 2015, 2016 JA 2017

Kuvaajissa punaisella piste-viiva -käyrällä on vuoden 2015 tiedot, sinisellä jatkuvalla viivalla 2016 ja mustalla viiva -käyrällä 2017. Analyysien aineistot ovat käyttöpaikkapainotuksella ja maksiminormalisoinnilla syötettynä.






Gap vuosi
Gap arvo
Klustereita
1.1
1.2
1.3
1.4
1.5
1.6
1.7
2
4
6
8
10
12
14
16
18
20


Siluetti vuosi
Siluetti arvo
Klustereita
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2
4
6
8
10
12
14
16
18
20

LIITE D: MAAKAAPELOINTIASTEEN KLUSTEROINTI

Kaapelointi-asteen väritys kuvaa luvun suuruutta, pienien arvojen ollessa punaisia, valkoisten arvojen keskitasoa ja vihreiden ollessa korkeita. Alla olevassa taulukossa tulosten selitteet.

Tunnus	Selite	Aloitusvektorien luonti	Normalisointi	K-arvo
A	Klusterointi k-means	Tasaväli	Euklidinen	12
B	Klusterointi k-medoids	500 replikaattia k-means++	Euklidinen	12
C	Klusterointi k-means	Tasaväli	Maksimi	12
D	Klusterointi k-means	Tasaväli	Maksimi	4
PJ	PJ-kaapelointi -%	-	-	-
KJ	KJ-kaapelointi -%	-	-	-
SJ	SJ-kaapelointi -%	-	-	-

Punaisella taustavärillä on merkitty A, B ja C sarakkeisiin yhden jäsenen klusterit. D sarakkeessa taustavärillä merkitty ne yhtiöt, joilla ei ole ollenkaan SJ-verkkoa.

Yhtiö	A	B	C	D	PJ	KJ	SJ
'Alajärven Sähkö Oy'	1	12	2	1	24,73	12,37	0,00
'Jeppo Kraft Andelslag'	1	2	3	1	30,90	15,15	0,00
'Keuruun Sähkö Oy'	1	2	3	1	37,70	15,62	0,00
'Koillis-Satakunnan Sähkö Oy'	1	2	3	1	34,93	8,92	0,00
'Kokemäen Sähkö Oy'	1	2	3	1	34,93	20,27	0,00
'Lankosken Sähkö Oy'	1	12	2	1	18,71	2,07	0,00
'Lehtimäen Sähkö Oy'	1	12	1	1	9,17	4,35	0,00
'Rantakairan Sähkö Oy'	1	12	1	1	6,76	7,42	0,00
'Tenergia Oy'	1	12	2	1	20,10	3,08	0,00
'Tunturiverkko Oy'	1	12	2	1	29,81	3,08	0,00
'Vakka-Suomen Voima Oy'	1	2	3	1	35,87	20,37	0,00
'Vetelin Energia Oy'	1	12	2	1	27,82	5,57	0,00
'Oulun Seudun Sähkö Verkkopalvelut Oy'	1	2	3	1	36,37	11,59	0,00
'Sipoon Energia Oy'	2	5	3	1	27,96	32,25	2,64
'Enontekiön Sähkö Oy'	1	12	1	1	7,86	1,66	0,00
'Muonion Sähköosuuskunta'	1	12	1	1	5,56	1,48	0,00
'Nivos Energia Oy'	1	2	3	1	32,84	24,23	1,38
'Koillis-Lapin Sähkö Oy'	1	12	2	1	28,59	2,70	0,00
'Kymenlaakson Sähköverkko Oy'	1	12	2	1	24,78	18,57	0,00
'PKS Sähkönsiirto Oy'	1	12	2	1	18,95	6,62	0,00
'Tornionlaakson Sähkö Oy'	1	12	2	1	22,58	7,31	0,00
'Järvi-Suomen Energia Oy'	1	12	2	1	25,67	10,33	0,06
'Savon Voima Verkko Oy'	1	2	3	1	34,43	10,66	0,00
'Esse Elektro-Kraft Ab'	1	2	3	2	47,30	4,13	0,00
'Kronoby Elverk Ab'	1	2	3	2	45,11	11,44	0,00
'Leppäkosken Sähkö Oy'	1	2	3	2	42,32	17,53	0,00
'Vatajankosken Sähkö Oy'	1	2	3	2	42,82	6,39	0,00
'Verkko Korpela Oy'	1	2	3	2	45,43	12,42	0,00

'Imatran Seudun Sähkönsiirto Oy'	2	8	5	2	55,02	20,65	0,00
'Jylhän Sähköosuuskunta'	2	8	5	2	67,89	9,72	0,00
'Keminmaan Energia Oy'	2	8	5	2	57,92	19,32	0,00
'Vimpelin Voima Oy'	2	8	5	2	76,31	13,89	0,00
'Nykarleby Kraftverk Ab'	2	2	3	2	39,32	23,40	0,00
'Valkeakosken Energia Oy'	2	5	4	2	56,99	36,47	0,00
'Iin Energia Oy'	1	2	3	2	42,52	15,46	0,00
'Parikkalan Valo Oy'	2	2	5	2	54,33	10,33	0,00
'Lammaisten Energia Oy'	2	8	5	2	66,23	30,64	0,00
'Paneliankosken Voima Oy'	2	5	4	2	47,95	26,19	0,00
'Köyliön-Säkylän Sähkö Oy'	2	5	4	2	60,24	34,56	0,00
'Kuoreveden Sähkö Oy'	2	5	4	2	48,94	28,44	0,00
'Haukiputaan Sähköosuuskunta'	2	8	5	2	62,34	21,19	0,00
'Nurmijärven Sähköverkko Oy'	2	8	5	2	57,84	26,29	0,00
'Sallila Sähkönsiirto Oy'	2	8	5	2	67,10	25,69	0,00
'Porvoon Sähköverkko Oy'	2	5	4	2	42,74	33,83	0,00
'Outokummun Energia Oy'	1	2	3	2	41,67	14,48	0,00
'Vaasan Sähköverkko Oy'	2	8	5	2	64,81	31,82	0,00
'KSS Verkko Oy'	2	2	4	2	48,49	20,82	0,00
'Rovakaira Oy'	1	2	3	2	45,38	5,91	0,17
'Lappeenrannan Energiaverkot Oy'	2	8	4	2	54,08	23,05	5,83
'Loiste Sähköverkko Oy'	1	2	3	2	45,49	8,03	2,24
'Herrfors Nät-Verkko Oy Ab'	2	8	4	2	52,07	23,46	0,00
'Elenia Oy'	2	5	4	2	47,55	32,07	0,41
'Caruna Oy'	2	5	4	2	44,95	41,77	1,20
'Raahen Energia Oy'	9	1	9	3	91,01	95,56	0,00
'Haminan Energia Oy'	11	7	6	3	61,61	74,08	0,00
'Kemin Energia ja Vesi Oy'	11	7	7	3	82,96	72,50	0,00
'Äänekosken Energia Oy'	10	7	6	3	78,45	59,96	0,00
'Naantalin Energia Oy'	12	4	8	3	81,08	70,44	18,92
'Ekenäs Energi Ab'	11	1	7	3	94,66	75,73	0,00
'Forssan Verkkopalvelut Oy'	11	7	7	3	87,04	70,33	0,00
'JE-Siirto Oy'	7	4	10	3	97,14	86,91	21,55
'Kokkolan Energiaverkot Oy'	10	7	6	3	70,83	56,56	0,00
'Keravan Energia Oy'	9	1	9	3	86,02	90,35	4,00
'Tornion Energia Oy'	10	7	6	3	73,88	44,62	0,00
'ESE-Verkko Oy'	3	9	7	3	82,96	62,13	10,03
'Oulun Energia Siirto ja Jakelu Oy'	3	9	7	3	87,72	70,31	7,74
'Vantaan Energia Sähköverkot Oy'	9	1	9	3	86,93	91,29	0,00
'Tampereen Sähköverkko Oy'	7	4	8	3	70,77	66,86	28,80
'Pori Energia Sähköverkot Oy'	10	7	6	3	72,91	56,59	0,53
'Turku Energia Sähköverkot Oy'	12	9	8	3	71,41	67,05	13,26
'LE-Sähköverkko Oy'	12	9	8	3	75,92	56,29	17,02
'Caruna Espoo Oy'	7	4	8	3	73,09	75,27	24,30
'Rauman Energia Sähköverkko Oy'	5	6	11	4	75,39	72,46	100,00
'Seiverkot Oy'	6	11	12	4	89,21	57,92	40,00
'Kuopion Sähköverkko Oy'	8	10	12	4	79,95	77,99	61,51
'Rovaniemen Verkko Oy'	4	3	10	4	92,54	89,11	38,97
'Helen Sähköverkko Oy'	4	3	10	4	97,98	99,75	35,02

LIITE E: TEKNISTEN TUNNUSLUKUJEN ABC-KLUSTEROINTI

Taulukossa yhtiöiden klusterin indeksi näkyy taulukossa. Tyhjä arvo tarkoittaa, että kyseinen yhtiö ei ollut mukana isommassa klusterissa ja siten ei saanut uutta klusteria asteittaisessa klusteroinnissa. Taulukossa on ensimmäisissä kuudessa sarakkeessa käyttöpaikkapainotuksella tehdyt klusteroinnit ja tämän jälkeen käyttöpaikkapainottomat klusterien indeksit yhtiöille.

Tunnus	Selite	käyttöpaikkapainotus
A1	ABC-aineiston klusterin indeksi	Kyllä
B1	Toisen asteen klusteroinnin indeksi	Kyllä
C1	Kolmannen asteen klusteroinnin indeksi	Kyllä
D1	Volyymi-aineisto klusterin indeksi	Kyllä
E1	Panos-aineisto klusterin indeksi	Kyllä
F1	Toimitusvarmuus-aineisto klusterin indeksi	Kyllä
A2	ABC-aineiston klusterin indeksi	Ei
B2	Toisen asteen klusteroinnin indeksi	Ei
D2	Volyymi-aineisto klusterin indeksi	Ei
E2	Panos-aineisto klusterin indeksi	Ei
F2	Toimitusvarmuus-aineisto klusterin indeksi	Ei

Yllä olevassa taulukossa on selitetty alla tuloksissa olevat tunnukset. Punaisella taustavärillä on merkitty yhden jäsenen klusterit.

Yhtiö, maksiminormalisointi, k-means, K12, aineisto TeTu 2017 (ABC)	A1	B1	C1	D1	E1	F1	A2	B2	D2	E2	F2
'Jeppo Kraft Andelslag'	1	-	-	2	4	5	8	-	11	1	4
'Haminan Energia Oy'	2	-	1	11	2	1	1	1	1	1	1
'Kemin Energia ja Vesi Oy'	2	-	2	11	2	1	1	1	1	1	1
'Lappeenrannan Energiaverkot Oy'	2	-	2	11	3	1	1	1	2	2	1
'Naantalin Energia Oy'	2	-	2	11	2	1	1	1	1	1	1
'Nurmijärven Sähköverkko Oy'	2	-	2	3	2	1	1	1	1	1	1
'Tornion Energia Oy'	2	-	2	3	2	1	1	1	1	1	1
'Vaasan Sähköverkko Oy'	2	-	2	11	3	1	2	-	2	2	1
'Äänekosken Energia Oy'	2	-	2	3	2	1	1	1	1	1	1
'Valkeakosken Energia Oy'	2	-	3	3	2	1	1	3	5	1	1
'Helen Sähköverkko Oy'	2	-	4	4	1	1	4	-	6	10	1
'Forssan Verkkopalvelut Oy'	2	-	5	11	3	1	1	1	2	1	1
'Pori Energia Sähköverkot Oy'	2	-	6	4	2	1	2	-	2	2	1
'Caruna Espoo Oy'	2	-	7	4	1	1	2	-	3	10	12
'Keravan Energia Oy'	2	-	8	11	1	1	1	1	1	1	1
'Kokkolan Energiaverkot Oy'	2	-	8	11	2	1	1	9	1	1	1
'Raahen Energia Oy'	2	-	8	11	2	1	1	9	1	1	1
'Rauman Energia Sähköverkko Oy'	2	-	8	11	1	1	1	1	1	1	1
'Turku Energia Sähköverkot Oy'	2	-	8	11	1	1	2	-	3	2	1

'Oulun Energia Siirto ja Jakelu Oy'	2	-	9	4	1	1	2	-	3	2	1
'JE-Siirto Oy'	2	-	10	4	1	1	2	-	2	1	1
'Kuopion Sähköverkko Oy'	2	-	10	4	1	1	2	-	2	1	1
'Rovaniemen Verkko Oy'	2	-	10	4	1	1	1	1	1	1	1
'Vantaan Energia Sähköverkot Oy'	2	-	10	4	1	1	2	-	3	2	1
'ESE-Verkko Oy'	2	-	11	4	1	1	1	1	1	1	1
'LE-Sähköverkko Oy'	2	-	11	4	2	1	2	-	2	2	1
'Tampereen Sähköverkko Oy'	2	-	11	4	1	1	2	-	3	2	1
'Seiverkot Oy'	2	-	12	11	1	1	1	1	1	1	1
'Tenergia Oy'	3	1	-	7	6	3	1	8	1	1	2
'Alajärven Sähkö Oy'	3	2	-	3	4	2	1	8	1	1	2
'Iin Energia Oy'	3	2	-	3	4	2	1	5	11	1	1
'Jylhän Sähköosuuskunta'	3	2	-	3	4	2	1	1	1	1	1
'Keminmaan Energia Oy'	3	2	-	3	4	1	1	4	9	1	1
'Keuruun Sähkö Oy'	3	2	-	3	4	2	1	1	1	1	1
'Koillis-Satakunnan Sähkö Oy'	3	2	-	3	6	2	1	1	1	2	1
'Kokemäen Sähkö Oy'	3	2	-	3	4	1	1	1	1	1	1
'Kronoby Elverk Ab'	3	2	-	3	6	2	1	1	1	1	1
'Leppäkosken Sähkö Oy'	3	2	-	3	4	2	1	10	1	2	1
'Oulun Seudun Sähkö Verkkopalvelut Oy'	3	2	-	3	3	1	1	5	9	2	1
'Sallila Sähkönsiirto Oy'	3	2	-	3	4	1	1	1	1	2	1
'Tunturiverkko Oy'	3	2	-	3	5	1	1	1	1	1	1
'Vatajankosken Sähkö Oy'	3	2	-	11	6	2	1	1	1	2	1
'Verkko Korpela Oy'	3	2	-	3	4	2	1	1	1	2	1
'Vimpelin Voima Oy'	3	2	-	3	6	2	1	1	1	1	1
'Esse Elektro-Kraft Ab'	3	3	-	3	6	3	1	8	1	1	2
'Järvi-Suomen Energia Oy'	3	3	-	3	8	3	10	-	2	9	8
'Lankosken Sähkö Oy'	3	3	-	3	6	3	1	7	1	1	2
'Parikkalan Valo Oy'	3	3	-	3	6	3	1	8	1	1	2
'Savon Voima Verkko Oy'	3	3	-	3	12	3	10	-	12	5	11
'Vetelin Energia Oy'	3	4	-	3	4	1	1	1	1	1	1
'Paneliankosken Voima Oy'	3	5	-	10	4	1	1	1	1	1	1
'Nykarleby Kraftverk Ab'	3	6	-	8	4	2	1	9	1	1	1
'Rantakairan Sähkö Oy'	3	7	-	7	5	2	1	6	1	1	2
'Lammaisten Energia Oy'	3	8	-	9	3	1	1	1	1	1	1
'Köyliön-Säkylän Sähkö Oy'	3	9	-	10	4	1	1	1	1	1	1
'Kymenlaakson Sähköverkko Oy'	3	10	-	3	3	2	3	-	2	3	12
'Lehtimäen Sähkö Oy'	3	11	-	3	6	3	1	6	1	1	2
'Kuoreveden Sähkö Oy'	3	12	-	3	8	2	1	1	1	1	1
'Nivos Energia Oy'	3	12	-	3	12	2	1	1	1	1	1
'Outokummun Energia Oy'	3	12	-	3	12	2	1	1	1	1	1
'Rovakaira Oy'	3	12	-	3	5	2	1	1	1	2	1
'Vakka-Suomen Voima Oy'	3	12	-	3	4	2	1	1	1	2	1
'KSS Verkko Oy'	4	-	-	11	3	10	3	-	2	2	12
'Loiste Sähköverkko Oy'	4	-	-	3	12	4	3	-	1	3	3
'PKS Sähkönsiirto Oy'	5	-	-	11	6	6	5	-	9	4	6
'Herrfors Nät-Verkko Oy Ab'	6	-	-	6	11	1	11	-	4	6	1

'Caruna Oy'	7	-	-	5	12	2	7	-	10	12	5
'Elenia Oy'	7	-	-	5	12	9	6	-	7	11	5
'Tornionlaakson Sähkö Oy'	7	-	-	1	9	2	1	2	2	7	2
'Sipoon Energia Oy'	8	-	-	3	4	7	1	12	1	1	2
'Haukiputaan Sähköosuuskunta'	9	-	-	3	3	8	1	11	1	1	1
'Enontekiön Sähkö Oy'	10	-	-	12	10	3	12	-	8	1	2
'Muonion Sähköosuuskunta'	10	-	-	3	7	3	1	7	1	1	2
'Ekenäs Energi Ab'	11	-	-	11	2	9	12	-	1	1	10
'Koillis-Lapin Sähkö Oy'	11	-	-	3	9	9	12	-	1	8	10
'Imatran Seudun Sähkönsiirto Oy'	12	-	-	3	3	11	9	-	1	1	9
'Porvoon Sähköverkko Oy'	12	-	-	3	3	12	9	-	2	2	7

LIITE F: TEKNISET TUNNUSLUVUT FUUSIOLLA 1997-2017

Toisesta sarakkeesta lähtien on merkitty yhtiöittäin teknisten tunnuslukujen klusterointitulokset vuosilta 1997-2017. Taustavärillä on merkittynä yhden jäsenen klusterit

yhtiö, käyttäjäpainotus, maksimi normalisointi tetu	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
'Jeppo Kraft Andelslag'	4	4	3	4	2	7	3	3	2	2	3	3	3	2	3	3	3	3	3	8	1
'Caruna Espoo Oy 3)'	2	11	11	11	11	12	9	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'ESE-Verkko Oy'	3	4	4	12	4	9	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Forssan Verkkopalvelut Oy'	4	4	3	4	5	4	4	3	12	12	10	2	2	12	2	2	2	2	2	2	2
'Haminan Energia Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Helen Sähköverkko Oy'	3	5	7	12	5	6	8	7	12	12	10	2	2	12	11	2	2	2	2	2	2
'JE-Siirto Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Kemin Energia Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	2	2	2	2	2	2	2
'KENET Oy'	4	4	3	4	2	3	4	3	2	2	10	2	2	12	2	2	2	2	2	2	2
'Keravan Energia Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Kuopion Sähköverkko Oy'	3	5	4	12	4	9	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'LE-Sähköverkko Oy'	3	4	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	10	2	2	12	2	2	2	2	2	2	2
'Naantalin Energia Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	2	2	2	2	2	2	2
'Oulun Energia Siirto ja Jakelu Oy'	8	11	11	11	9	12	9	8	7	10	8	9	8	9	9	7	2	2	2	2	2
'Pori Energia Sähköverkot Oy'	3	4	4	7	5	4	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Raahen Energia Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	2	2	2	2	2	2	2
'Rauman Energia Sähköverkko Oy'	3	5	4	12	4	4	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Rovaniemen Verkko Oy'	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	10	2	2	12	11	2	2	2	2	2	2
'Seiverkot Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Tampereen Sähköverkko Oy'	3	5	12	12	5	6	8	7	12	12	10	2	2	12	11	2	2	2	2	2	2
'Tornion Energia Oy'	4	4	3	4	3	3	4	3	12	12	10	2	2	12	2	2	2	3	2	2	2
'Turku Energia Sähköverkot Oy'	3	4	4	12	4	4	4	3	12	12	10	2	2	12	11	2	2	2	2	2	2
'Valkeakosken Energia Oy'	4	4	3	4	3	3	3	3	2	2	2	3	2	2	2	3	2	3	3	2	2
'Vantaan Energia Sähköverkot Oy'	5	1	2	1	7	5	11	10	4	8	1	6	11	1	1	9	1	2	2	2	2
'Etelä-Suomen Energia Oy'	4	4	3	4	3	3	3	3	2	2	7	3	3	2	3	3	3	11	3	3	3
'Iin Energia Oy'	4	4	3	3	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	3
'Imatran Seudun Sähkö Oy'	4	4	3	4	2	3	3	4	2	3	3	3	3	2	2	3	3	3	3	3	3
'Jylhän Sähköosuuskunta'	4	12	3	7	2	3	3	4	2	2	2	3	3	2	3	3	3	3	3	3	3
'Keminmaan Energia Oy'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	2	3	3	3	3	3	3
'Keuruun Sähkö Oy'	4	4	3	4	3	3	7	4	2	3	2	3	3	3	3	12	3	3	3	9	3
'Kokemäen Sähkö Oy'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	3
'KSS Energia Oy'	12	9	11	11	11	11	9	12	7	10	9	11	9	9	2	3	2	3	2	2	3
'Köyliön-Säkylän Sähkö Oy'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	3
'Lammaisten Energia Oy'	3	4	3	2	3	3	3	3	2	2	2	3	3	2	2	3	3	3	3	3	3
'Lappeenrannan Energiaverkot Oy'	2	11	11	11	11	11	9	12	7	6	9	11	2	2	2	3	3	3	2	3	3
'Leppäkosken Sähkö Oy'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	3
'Nivos Energia Oy 5)'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	3
'Nurmijärven Sähköverkko Oy'	4	4	3	4	3	3	3	3	11	2	2	3	3	2	2	3	3	3	3	3	3
'Oulun Seudun Sähkö Verkkopalvelut Oy'	4	4	3	4	3	3	7	3	2	2	2	3	3	2	3	3	3	3	3	3	3
'Paneliankosken Voima Oy'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	3
'Porvoon Sähköverkko Oy'	9	7	9	11	3	3	3	3	2	2	3	3	3	2	2	3	3	3	3	3	3
'Sallila Sähkönsiirto Oy'	4	3	3	3	3	3	3	11	2	2	2	3	3	2	3	1	3	3	3	3	3
'Vaasan Sähköverkko Oy'	2	11	11	8	6	10	1	1	1	9	9	11	9	9	10	7	2	3	2	2	3
'Vakka-Suomen Voima Oy'	4	4	3	4	3	3	3	4	2	2	7	3	3	2	3	3	3	3	3	3	3
'Vatajankosken Sähkö Oy'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	1	3	3	3	3	3
'Äänekosken Energia Oy 1)'	4	4	3	4	3	9	3	3	12	12	10	2	2	12	2	2	2	2	2	2	3
'Loiste Sähköverkko Oy'	4	4	3	7	3	3	7	4	3	3	3	1	3	3	3	3	3	7	3	3	4
'PKS Sähkönsiirto Oy'	2	7	3	7	3	3	7	4	3	3	5	1	12	3	3	11	3	6	4	11	4
'Herrfors Nät-Verkko Oy Ab'	7	7	8	9	12	4	7	3	12	12	10	7	7	8	8	6	7	10	10	6	5
'Ekenäs Energi Ab'	1	5	4	12	4	4	4	3	12	12	10	2	2	12	2	2	2	2	8	2	6
'Tenergia Oy'	4	3	3	3	3	3	7	4	2	2	3	12	3	3	3	3	9	3	11	3	7
'Caruna Oy 4)'	10	2	5	5	9	11	8	3	9	5	6	5	1	6	4	10	4	9	7	12	8
'Elenia Oy'	6	6	6	5	12	2	9	12	8	11	12	8	10	10	7	4	4	8	7	12	8
'Järvi-Suomen Energia Oy'	4	4	3	4	3	3	7	4	3	3	5	1	3	6	4	12	10	3	3	4	8
'Rovakaira Oy'	4	3	3	3	3	3	3	11	2	3	3	3	3	2	3	3	3	3	3	3	8
'Savon Voima Verkko Oy'	12	9	11	11	11	11	12	6	5	11	12	8	5	5	6	5	8	4	4	4	8
'Tornionlaakson Sähkö Oy'	2	9	11	11	8	8	6	2	6	7	11	10	6	11	5	8	6	12	1	9	8
'Koillis-Lapin Sähkö Oy'	4	3	3	10	10	7	10	11	2	3	3	3	3	3	3	3	11	3	3	7	9
'Muonion Sähköosuuskunta'	4	3	3	3	10	3	10	11	10	3	3	3	3	3	3	11	5	1	5	1	9
'Vetelin Energia Oy'	4	4	3	4	3	3	3	4	2	2	2	3	3	2	3	3	3	3	12	3	9
'Alajärven Sähkö Oy'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	10
'Haukiputaan Sähköosuuskunta'	4	4	3	3	3	9	3	3	2	2	2	3	3	2	2	3	3	3	3	3	10
'Koillis-Satakunnan Sähkö Oy'	4	3	3	3	10	3	7	4	2	2	2	3	3	3	3	11	3	3	3	3	10
'Kuoreveden Sähkö Oy'	4	10	3	10	10	7	10	11	2	2	2	3	3	2	3	3	3	3	3	3	10
'Kymenlaakson Sähkö Oy'	2	7	9	11	11	10	12	3	3	3	7	5	3	2	3	3	3	3	3	3	10
'Lehtimäen Sähkö Oy'	4	3	3	10	3	7	10	11	2	2	2	3	3	3	3	11	3	3	3	3	10
'Outokummun Energia Oy'	4	8	3	3	3	3	3	3	2	2	2	3	3	2	3	3	3	3	3	3	10
'Parikkalan Valo Oy'	4	12	3	7	3	3	7	4	10	3	3	3	3	7	3	12	10	3	3	3	10
'Verkko Korpela Oy'	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2	3	3	2	3	3	10	6	3	3	10
'Vimpelin Voima Oy'	4	4	3	4	3	3	3	4	2	2	2	3	3	3	3	3	3	3	3	3	10
'Enontekiön Sähkö Oy'	4	10	3	10	10	7	10	9	10	1	3	12	3	3	3	11	10	5	6	5	11
'Esse Elektro-Kraft Ab'	4	12	3	7	3	3	7	4	2	3	3	3	3	3	3	11	3	7	9	10	11
'Kronoby Elverk Ab'	4	3	3	3	3	3	2	9	2	2	2	3	3	2	3	3	3	7	3	3	11
'Lankosken Sähkö Oy'	4	12	3	7	3	3	7	4	11	3	3	3	3	3	12	12	10	3	3	3	11
'Nykarleby Kraftverk Ab'	4	4	3	4	3	3	3	3	2	2	2	3	3	2	3	11	12	3	10	3	11
'Tunturiverkko Oy'	11	9	10	6	1	1	5	5	6	4	4	4	4	4	3	3	3	3	3	3	11
'Rantakairan Sähkö Oy'	4	12	3	7	3	3	7	4	2	2	3	3	3	3	3	11	3	3	11	3	12