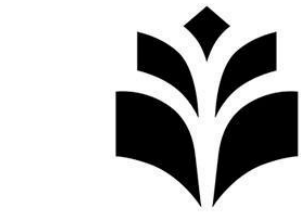


Kaukolämpöasiakkaiden ryhmittely kulutusprofiilien perusteella

Vili Lavikainen
Tietojenkäsittelytieteen koulutusohjelma
Itä-Suomen yliopisto
Luonnontieteiden ja metsätieteiden tiedekunta
Tietojenkäsittelytieteen laitos / tietojenkäsittelytiede
03.12.2023

Itä-Suomen yliopisto, Luonnontieteiden ja metsätieteiden tiedekunta

Tietojenkäsittelytieteen laitos

Tietojenkäsittelytieteen koulutusohjelma

Lavikainen, Vili: Kaukolämpöasiakkaiden ryhmittely kulutusprofiilien perusteella

Pro gradu –tutkielma, 51 sivua

Pro gradu –tutkielma, professori Pasi Fränti, Reima Lassila

Joulukuu 2023

Asiasanat: kaukolämpö, klusterointi, kulutusprofiili, k-shape

ACM-luokat (ACM Computing Classification System, 2012 versio): CCS → Computing methodologies → Machine learning → Learning paradigms → Unsupervised learning → Cluster analysis; CCS → Hardware → Power and energy → Energy distribution

Kaukolämpöasiakkaiden kulutuksen syvällinen ymmärtäminen on tärkeää tehokkaan kaukolämpöverkon operoinnin ja hallinnan kannalta. Etäluettavat mittarit ovat mahdollistaneet kaukolämpöasiakkaiden kulutuksen tarkemman analysoimisen, mistä hyötyy asiakas sekä energiayhtiö. Energia-alan yrityksillä on suuri tarve kehittää, ja optimoida omaa toimintaa ilmastonmuutoksen, sekä taloudellisten syiden takia. Datan hyödyntäminen on avainroolissa energiantuotannon ja -jakelun kehittämisessä. Tässä tutkielmassa analysoitiin kuopiolaisten kaukolämpöasiakkaiden vuoden 2021 kulutusta. Kaukolämpöasiakkaille muodostettiin kaksi kulutusprofiilia kuvaamaan niiden keskimääräistä vuorokauden kulutusta arkipäivinä sekä viikonlopun päivinä, ilman ulkolämpötilan vaikutusta. Kulutusprofiilit perustuvat lineaariseen regressiomalliin. Kulutusprofiilit edustavat asiakaskohtaista lämmöntarvetta, mikä riippuu rakennuksen käytöstä. Muodostetut kulutusprofiilit klusteroitiin k-Shape-algoritmilla, ja niistä muodostettiin neljä yleistä kulutusta kuvaavaa ryhmää. Muodostettujen ryhmien keskimääräiset kulutusprofiilit osoittivat selviä eroja ryhmien välillä, ja niistä pystyi tunnistamaan karkeasti rakennuksen käytön.

University of Eastern Finland, Faculty of Science and Forestry

School of Computing

Lavikainen, Vili: Clustering District Heating Customers Based on Load Profiles

Master's Thesis, 51 pages

Supervisors: professor Pasi Fränti, Reima Lassila

December 2023

Keywords: district heating, clustering, load profile, k-shape

CR Categories (ACM Computing Classification System, 2012 version): CCS → Computing methodologies → Machine learning → Learning paradigms → Unsupervised learning → Cluster analysis; CCS → Hardware → Power and energy → Energy distribution

A deep understanding of the consumption of district heating customers is important for the operation and management of an efficient district heating network. Remotely readable meters have made it possible to analyze the consumption of district heating customers more precisely, which benefits both the customer and the energy company. Companies in the energy sector have a great need to develop and optimize their own operations due to climate change and economic reasons. The utilization of data plays a key role in the development of energy production and distribution. In this study, the consumption of Kuopio's district heating customers in 2021 was analyzed. Two consumption profiles were created, one for district heating customers to describe their average daily consumption on weekdays, and the other one for weekend days, without the effect of outside temperature. Consumption profiles are based on a linear regression model. The consumption profiles represent customer-specific heat demand, which depends on the use of the building. The formed consumption profiles were clustered with the k-Shape algorithm, and four groups describing general consumption were formed from them. The average consumption profiles of the formed groups showed clear differences between the groups, and from them it was possible to roughly identify the use of the building.

Alkusanat

Haluan kiittää Jami Miettistä siitä, että sain vapaasti valita itseäni kiinnostavan aiheen, ja hyödyntää siihen oikeaa mittausdataa. Kiitos Reima Lassila kommenteista, ja kiitos koko Kuopion Energia Oy tuesta, mikä mahdollisti opintojen suorittamisen loppuun aikataulussa töiden ohella.

Suuri kiitos tämän työn pääohjaajalle Pasi Fräntille tarkoista kommenteista, ja hyvistä ehdotuksista. Autoit parantamaan kriittistä tiedonhakua ja -ajatteluani yhä enemmän.

Lopuksi suuri kiitos kaikille ihmisille, joiden kanssa olen saanut kokea ichi-go ichi-e -hetkiä, sekä erityisesti ystävilleni. Teidän merkitystänne en voi korostaa liikaa.

Kuopio, Joulukuu 2023
Vili Lavikainen

Lyhenteet

CHP	Combined heat and power; Sähkön ja lämmön yhteistuotanto
MSCONS	Metered service consumption report; Sanomapalvelu
MDLP	Modified daily load profile; Päivittäinen kuormitusprofiili
CHI	Calinski-Harabasz-indeksi
DBI	Davies-Boulding-indeksi
SC	Silhuette Coefficient; Siluetti-indeksi
SSD	Sum of Squared Distances; Neliöetäisyyksien summa
GS	Gap statistic

Symbolit

K	Klustereiden lukumäärä
T_{sh}	Käännepiste
H	Lämmön kokonaiskulutus yhdessä aikavälissä
A	Aikavälin indeksi
T	Keskiulkolämpötila aikavälissä A
β_0^x	Mallin x vakiotermi
β_1^x	Mallin x keskimääräinen kulutuksen muutos ulkolämpötilasta
β_2^x	Mallin x keskimääräinen kulutuksen muutos aikavälissä A
$\hat{\beta}_y^x$	Arkipäivän mallin x termin y estimaatti
$\tilde{\beta}_y^x$	Viikonlopun päivän mallin x termin y estimaatti
$\hat{u}_i$	Tyypillinen arkipäivän lämmönkäyttö aikaindeksissä i
$\tilde{u}_i$	Tyypillinen viikonlopun päivän lämmönkäyttö aikaindeksissä i
$mean\{u\}$	Vektorin u keskiarvo
$std\{u\}$	Vektorin u keskihajonta

Sisältö

1 Johdanto

Prosessia siirtyä fossiilisista polttoaineista uusiutuviin energiajärjestelmiin kutsutaan energiamurrokseksi (energy transition), mikä on ratkaisevan tärkeä ilmastonmuutoksen hillitsemisessä. Vuonna 2022 energiasektori oli Suomessa suurin kasvihuonekaasujen päästölähde, joka muodosti 72 % kokonaispäästöistä (Suomen virallinen tilasto 2022), lisäksi kohonneet päästöoikeuksien hinnat ovat luoneet merkittävää ympäristöllistä- ja taloudellista painetta siirtyä uusiutuviin energiajärjestelmiin, ja optimoida nykyisiä järjestelmiä toimimaan yhä tehokkaammin. Energia-alalla on myös uusia haasteita. Energian tuottajien ja -kuluttajien välinen ero ei ole enää selvä. Rakennusten sähkön omatuotanto on lisääntymässä yhä enemmän, kun aurinkopaneelien hinnat ovat laskeneet. Lisäksi energiantuotannossa tuotannon kohdistaminen tietyille ajanhetkille on tullut yhä tärkeämmäksi, kun uusiutuvia energianlähteitä hyödynnetään yhä enemmän, mikä lisää suurempaa vaihtelua tuotannon määrissä. Erilaiset energianvarastointimenetelmät ovat yleistymässä tasoittamaan tätä vaihtelua. Se luo suuria optimointimahdollisuuksia, joiden positiiviset vaikutukset ovat merkittäviä. Myös etäluettavat mittarit ja niiden pienentynyt mittausväli mahdollistaa energiankäytön syvällisemmän ymmärryksen, ja paremman mahdollisuuden puuttua vikatilanteisiin. Lisäksi nämä mahdollistavat erilaisten tekoälymenetelmien hyödyntämisen.

Kaukolämpöjärjestelmän tehokkaan hallinnan ja toiminnan kannalta lämmöntarpeen tunnistaminen on elintärkeää, mikä voi auttaa analysoimaan lämmöntarpeen ominaisuuksia ja diagnosoimaan toimintahäiriöitä (Gadd & Werner 2013). Lämmön kulutusprofiilien analyysin lisäksi tarkka tuntikohtainen lämmön kulutusennuste on tunnistettu kaukolämpöjärjestelmän toiminnan säätelyn perustaksi, mikä auttaa energian kulutuksen vähentämisessä (Cho et al. 2009).

Klusterointi (ryhmittely) on luokiteltu suosituimmaksi kuvaavaksi tiedonlouhintatekniikaksi (Lu et al. 2019). Klusterointialgoritmeja on käytetty menestyksekkäästi rakennusten energiajärjestelmien analysoinnissa suurista tietomassoista (Yu et al. 2012). Se on osoittautunut tehokkaaksi tekniikaksi kulutusprofiilien tunnistamiseen (Xiao & Fan 2014), sekä muiden

tiedonlouhintatekniikoiden esikäsittelyvaiheeksi, erityisesti energiankäytön ennustamiseen (Goia et al. 2010; Duan et al. 2011).

Tässä tutkielmassa analysoin kuopiolaisten kaukolämpöasiakkaiden lämmönkäyttödataa vuodelta 2021, sekä muodostan jokaisesta asiakkaasta kaksi kulutusprofiilia, mitkä edustavat asiakkaan keskimääräistä kulutusta vuorokauden aikana arkipäivinä ja viikonlopun päivinä. Lisäksi muodostetut profiilit klusteroidaan, joista saadaan esille yleisimmät asiakkaiden keskimääräiset kulutusprofiilit. Lopuksi analysoin muodostetut klusterit.

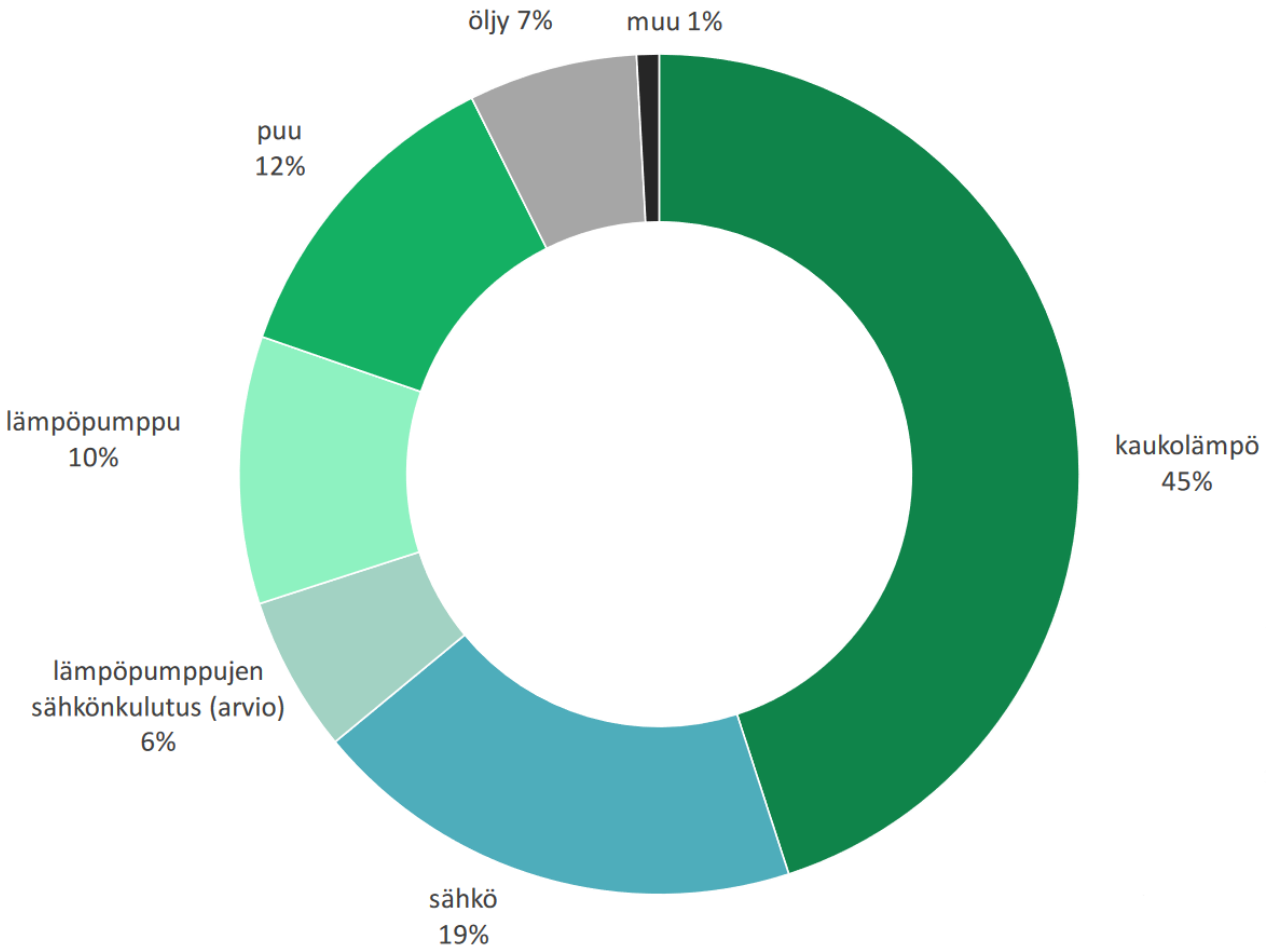
Luvussa 2 esittelen kaukolämpöä, ja kaukolämpöjärjestelmän toimintaa. Lisäksi esittelen yleisimpiä haasteita, mitä kaukolämpöyhtiöt kohtaavat, sekä miten digitalisaatio ja data-analyysi voi auttaa kulutuksen analysoinnilla.

Luvussa 3 esittelen klusterointia, jonka jälkeen aliluvussa 3.1 kuvailen tutkielmassa käytettyä tietojoukkoa. Aliluvussa 3.2 tietojoukolle suoritetaan esiprosessointi, missä tunnistetaan virheellisiä- ja puuttuvia lukemia, sekä käsitellään datassa olevat puutteet. Aliluvussa 3.3 kaukolämpöasiakkaille muodostetaan kulutusprofiilit hyödyntäen Wang et al. (2019) ehdottamaa menetelmää. Aliluvussa 3.4 valitaan klusterointimenetelmä suorittamaan klusterointi. Aliluvussa 3.5 arvioidaan klusterointia eri klusterien lukumäärillä, ja valitaan lopullisessa klusteroinnissa käytettävä klusterien lukumäärä.

Luvussa 4 analysoidaan muodostetut klusterit. Analysoinnissa tarkastellaan klustereiden edustajia (sentroideja), ja miten ne eroavat keskenään. Analysoinnissa tarkastellaan myös kaukolämpöasiakkaiden rakennustyyppien jakautumista klusterissa, sekä klusteriin nimettyjen asiakkaiden yhteenlaskettua energiankäyttöä vuoden aikana

2 Kaukolämpö

Kaukolämpö on yleisin lämmitysmuoto Suomessa. Vuonna 2020 45 % Suomen asuin- ja palvelurakennusten lämmitysenergiasta tuotettiin kaukolämmöllä (Kuva 1), ja sillä on 155 000 asiakasta (Energiateollisuus 2022; Motiva 2022). kaukolämmitys voidaan määritellä siten, että se on rakennusten ja käyttöveden lämmittämiseen tarvittavan lämmön keskitettyä tuotantoa ja julkista jakelua asiakkaina oleville kiinteistöille. Kaukolämmölle on myös ominaista, että sen organisoitu toiminta toteutetaan liiketoiminnan muodossa (Koskelainen et al. 2006).



Kuva 1. Lämmityksen markkinaosuudet asuin- ja palvelurakennuksissa vuonna 2020. (Energiateollisuus 2021)

Kaukolämmön perusidea on käyttää paikallisia polttoaineita tai muuten hukkaan meneviä lämpöresursseja paikallisten asiakkaiden lämmitystarpeiden tyydyttämiseksi käyttäen lämmönjakeluverkkoa paikallisena markkinapaikkana. Kaukolämpöön käytetty lämpö voi olla hukkalämpöä lämpövoimalan tuotannosta sähkön ja lämmön yhteislaistoksista (Combined Heat and Power, lyh. CHP), teollisuuden hukkalämpöä, kotitalousjätteen poltosta syntyvää kierrätyslämpöä, muita huonolaatuisia polttoaineita sekä biomassaa esimerkiksi metsätalouden jätepuuta. Kaukolämpölaitoksissa lämpöä voidaan myös tuottaa käyttämällä aurinkolämpöä, maalämpöä, tuulivoimalla toimivaa sähkölämpöä sekä mahdollisesti ydinvoimaa. (Frederiksen & Werner 2013; Tulkki 2022) Kaukolämpö on ollut ja on teknisesti erinomainen ratkaisu, ja se mahdollistaa keskitetyn lämmöntuotannon. Keskitetyssä energiantuotannossa hyötysuhde on korkea, ja sillä on etenkin tiheästi asutetuissa ja viileissä oloissa selviä etuja muihin lämmitysmuotoihin nähden (Salo 2021).

Suomessa kaukolämpö perustuu vielä paljon polttoon perustuviin lämmönlähteisiin (Kuva 2). Fossiiliset polttoaineet ovat yleisesti olleet kustannustehokkain lähde kaukolämmön tuotantoon (Salo 2021). Fossiilisten polttoaineiden polttamista pyritään vähentämään. Euroopan unionin ilmastopolitiikalla pyritään vähentämään fossiilisten päästöjen määrää, mikä näkyy Suomessa verotuksessa ja päästöoikeuksissa. Lisäksi polttoaineiden hankintakustannukset nousivat, kun Venäjä aloitti Ukrainassa hyökkäyssodan, ja energiapuun tuonti Venäjältä lopetettiin. Nämä asiat ovat lisänneet kaukolämmön kustannuksia merkittävästi. Jotta ilmastotavoitteiden saavuttaminen ja kaukolämpöyhtiöiden kustannusten vähentäminen on mahdollista, toimintaa tulee kehittää.

Tärkein tekijä tällaisten järjestelmien tehokkuuden lisäämisessä on jakelulämpötilojen alentaminen siten, että energian tarjonnan ja kysynnän laatu paranee (Gadd & Werner 2014). Matalien lämpötilojen saavuttaminen verkossa vaatii älykkäitä ohjausjärjestelmiä, ja kehitettyjä strategioita tunnistaa jatkuvasti korkeita paluulämpötiloja aiheuttavia toimintavirheitä. Tällaisten strategioiden suunnittelussa on ensiarvoisen tärkeää tuntea asiakkaat syvällisesti, ja ymmärtää paremmin heidän lämmönkäyttöään, sillä yhdelläkin lämpöasemalla voi olla merkittävä vaikutus järjestelmän globaaliin tehokkuuteen. (Calikus et al. 2019)

Kaukolämmön rooli on merkittävä, mutta kaukolämpöteknologiaa on kehitettävä edelleen verkkohäviöiden vähentämiseksi, synergioiden hyödyntämiseksi ja sitä kautta järjestelmän matalalämpöisten tuotantoyksiköiden tehokkuuden lisäämiseksi. Uusiutuva energia yhdessä energiansäästön ja CHP-tuotannon kanssa on olennainen tekijä ilmastonmuutoksen torjunnassa Euroopassa ja monilla muilla alueilla. (Lund et al. 2014)

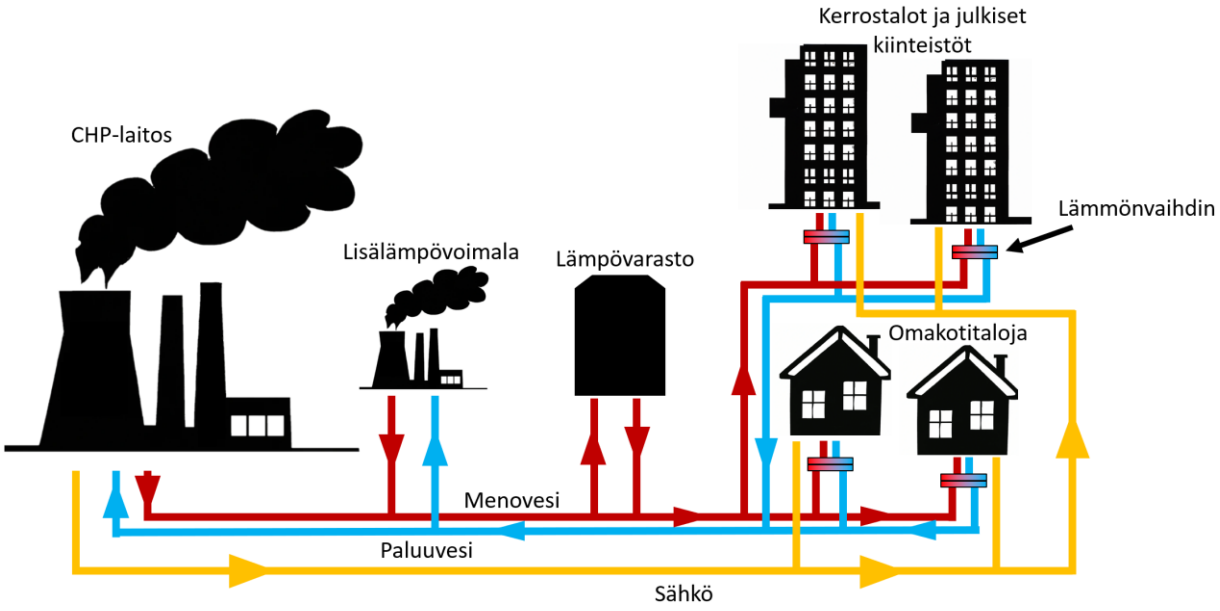
Kaukolämmön kilpailukyky syntyy lämmöntuotannon ja lämmönjakelun ehtojen yhdistelmästä. Eräs tärkeä lämmönjakelun ehto on, että lämmöntarve on keskitettävä jakelukustannusten ja lämpöhäviöiden minimoimiseksi (Frederiksen & Werner 2013). Matala lämpötiheys harvoin asutetuilla alueilla johtaa suhteellisesti korkeampiin jakelukustannuksiin ja hävikkiin (Nilsson et al. 2008; Reidhav & Werner 2008). Strategisia pidemmän aikavälin valintoja tehtäessä olennaista on se, että suuret kaupungit ovat riittävän tiheitä kestämään asiakkaiden lämmöntarpeen merkittävää vähenemistä menettämättä kaukolämmön yleistä kilpailukykyä (Persson & Werner 2011). Vastauksena erilaisiin kaukolämmön haasteisiin kaukolämpö on siirtymässä älykkäämpään lämpöjärjestelmään, jota kutsutaan neljännen sukupolven kaukolämpöjärjestelmäksi (Lund et al. 2014). Tämä järjestelmä sisältää älykkään lämpöjärjestelmän ja älykkään kaukolämpöverkon vuorovaikutuksen (Levih 2017; Sameti & Haghighat 2017).

Seuraavissa aliluvuissa selitetään kaukolämmön perustoimintaa, mitä haasteita kaukolämmössä on ja miten näitä haasteita voidaan ratkaista data-analytiikan avulla.

2.1 Kaukolämmön toiminta

Kaukolämmön toimintaperiaate on, että yhdessä tai useammassa lämpövoimalassa lämmitetään vesi, joka kuljetetaan asiakkaille kaukolämpöverkossa veden tai höyryn avulla (Kuva 2). Tyypillisiä kaukolämpöasiakkaita ovat: asuintalot, teollisuus-, liike- ja julkiset rakennukset (Koskelainen et al. 2006). Lämpö siirretään tyypillisesti lämmönvaihtimen kautta lämmönjakokeskuksessa asiakkaan omaan lämmitysjärjestelmään, ja viilentynyt vesi ohjataan takaisin lämmitettäväksi. Nykyisin lämpövaraajat ovat harvinaisia rakennuksissa, joten lämmönkäyttö kohdistetaan suoraan lämpöverkkoon. Lämmönsyöttöä ohjataan paine-eron

säätimellä ja menoveden lämpötilalla. Kaukolämmöllä lämmitetään pääsääntöisesti asiakkaan tilat ja käyttövesi. Joitain teollisia sovelluksia on, mutta useimmissa tilanteissa toimitettava lämpö kaukolämpöverkoissa on liian pieni käytettäväksi teollisissa prosesseissa (Gadd & Werner 2013.) Lämmöntuotannon tasaamiseksi ja CHP-laitoksen sähköntuotannon optimoimiseksi lämmitetty vesi voidaan ohjata lämpövarastoon, missä lämpöenergia varastoidaan muun muassa veteen. Lämpövarastoa hyödyntämällä voidaan lisätä lämmöntuotannon joustavuutta ja potentiaalia vähentämällä lämmöntuotannon kysyntää huippukuormituksen aikana, sekä tasapainottamalla kysynnän ja tarjonnan välistä eroa lyhyellä ja pitkällä aikavälillä (Skytte & Olsen 2016; Schweiger et al. 2017; Razmara et al. 2017; Li & Wang 2014).

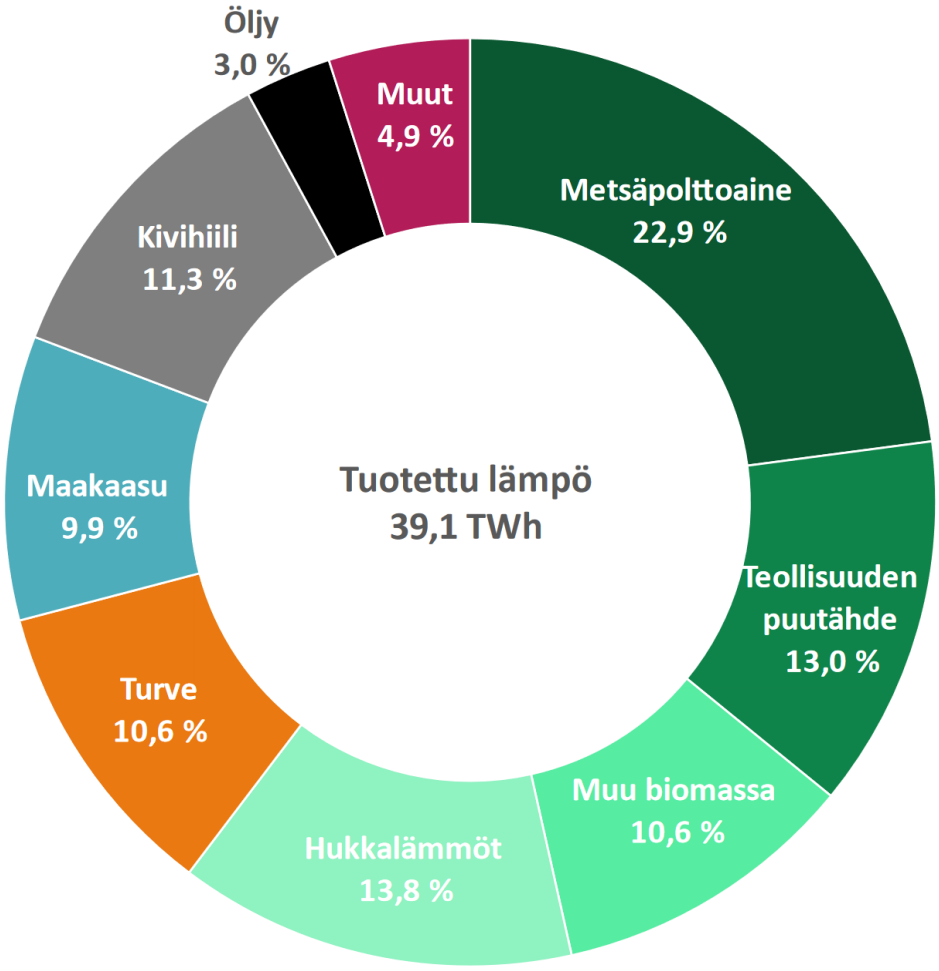


Kuva 2. Kaukolämpöverkko.

Fossiiliset polttoaineet ovat yleisesti olleet kustannustehokkain kaukolämmön tuotantolähde ja se on useimmissa tapauksissa vahvasti sidoksissa sähkön hintaan. Fossiilisten polttoaineiden kallistuminen (osittain kasvihuonepäästökustannusten nousun vuoksi), ja kansainväliset ilmastopolitiikat ovat tarjonneet suotuisat markkinaolosuhteet uusiutuvien energialähteiden käyttöönotolle. Lisäksi hiilineutraali lämmitys on tärkein yksittäinen kasvihuonepäästöjen vähentämistekijä useimmissa Pohjoismaiden kaupungeissa. Nämä markkinatrendit voivat myös edistää tehokkaampaa kaukolämpöjärjestelmää. Uusiutuvaa lämpöä voidaan tuottaa muun

muassa biomassa-, aurinko-, geoterminen tai sähkönlähteet, joilla on alhainen hiilidioksidivaikutus. (Salo 2021)

Suomessa vuonna 2021 kaukolämmön pääasiallinen polttoaine on puu (kuva 3). Muut suosituimmat polttoaineet kaukolämmössä ovat muut energianlähteet, kivihiili, maakaasu ja turve. Yhdessä nämä muodostavat 88,2 % kaikista käytetyistä polttoaineista (Suomen virallinen tilasto 2022.)



Kuva 3. Kaukolämmön hankinnan energialähteet Suomessa vuonna 2021. (Energiateollisuus, 2021)

2.2 Kaukolämmön haasteita

Kaukolämpöjärjestelmät kohtaavat useita erilaisia haasteita. Koska kaukolämpö toteutetaan pääosin liiketoiminnan muodossa, kustannuksiin kiinnitetään paljon huomiota.

Lämmöntuotannon muuttuvia tuotantokustannuksia ovat: polttoainekustannukset, valmistevero, päästöoikeuskustannukset ja muut muuttuvat kustannukset (Känkänen et al. 2017).

Kaukolämpöjärjestelmän pääasiallinen lisäkustannus paikalliseen lämmöntuotantovaihtoehtoon verrattuna on väistämättömät lämmönjakelun kustannukset. Tämä kustannus sisältää sekä alkuperäisten verkkoinvestointikustannusten vuotuisen takaisinmaksun, sekä ylimääräisiä käyttökustannuksia lämmönjakoon liittyvien lämpötila- ja painehäviöiden kompensoimiseksi. Pysyäkseen kilpailukykyisenä kaukolämmön kokonaiskustannusten on oltava alhaisemmat kuin minkään paikallisen lämmöntuotantovaihtoehdon kustannukset. (Persson & Werner 2011) Näiden lisäksi on myös ilmastonmuutoksen torjumiseen liittyviä kustannuksia, jotka liittyvät käytettyyn polttoaineeseen ja sen määrään.

Euroopan unionin päästökauppajärjestelmä on EU:n ilmastopolitiikan tärkein tapa päästöjen vähentämiseen kustannustehokkaasti, ja se on maailman ensimmäinen ja laajin hiilidioksidimarkkina. Päästöoikeuksien ilmaisjaolla ja hintakehityksellä on vaikutusta kaukolämmön tuotantokustannuksiin, ja sitä kautta kaukolämmön hintaan Suomessa. Päästökaupan kustannukset kaukolämmöntuotannolle riippuvat voimakkaasti käytettävistä polttoaineista. Fossiilisia polttoaineita käyttävillä laitoksilla polttoainekustannusten ja verojen osuus on korkea, minkä vuoksi päästöoikeuden hinnan osuus kokonaiskustannuksista ei ole yhtä korkea kuin turvetta polttavalla laitoksella, joilla verojen suuruus on pienempi. (Känkänen et al. 2017)

Venäjän Ukrainassa aloittaman hyökkäyssodan seurauksena energiantuotannossa käytetyn puun hinta on noussut, koska puun tuonti Venäjältä lopetettiin. Energiapuun nousseen hinnan sekä heikomman saatavuuden takia, turpeen käyttöä on jouduttu nostamaan. Turpeen päästöoikeuksien takia sen hinta on noussut merkittävästi. Lisäksi lisääntynyt sähköntuotanto Suomessa on alentanut sähkön hintaa, mikä vähentää CHP-tuotannosta saatavia sähkön myyntituloja. Suomessa nousseet polttoainekustannukset, ja laskenut sähköstä saatava tuotto on huonontanut kaukolämpöyhtiöiden taloudellista tilannetta.

2.3 Kulutuksen analysointi

Etäluettavien mittareiden käyttöönotto automaattiseen kulutuksen mittaamiseen on mahdollistanut kulutuksen tallentamisen ennenäkemättömällä tarkkuudella. Mittarit siirtyivät puolivuosittain manuaalisista lukemista automaattisiin tuntikohtaisiin lukemiin (Tureczek 2019). Etäluettavat mittarit kykenevät mittaamisen lisäksi lähettämään mitatun tiedon suoraan kaukolämpöyhtiölle. Koska mittaustietoja saadaan usein, se mahdollistaa asiakkaiden lämmönkäytön analysoinnin ja paremman verkon hallinnan. Mbiydzenyuy et al. (2021) ja Darby (2010) ovat listanneet, että kaukolämpösektorin digitalisaatio tarjoaa erilaisia mahdollisuuksia kuten:

- Energian huippukäytön alentaminen
- Sisätilojen lämpötilan optimointi
- Energiatuotannon optimointi
- Lämmönjakokeskusten ja kaukolämpöverkon ympärivuorokautinen valvonta
- Petosten vähentäminen
- Tarkka laskutus
- Veden säästäminen

Älymittareiden avulla voidaan myös kehittää tarkempia ennustemalleja ja yksityiskohtaisia analyyseja rakennusten energiankulutuksen vaikuttajista (Kipping & Trømborg 2016).

Asiakkaiden lämmönkäytön parempi ymmärtäminen yksittäisen lämmönjakokeskuksen tasolla voi olla merkittävä vaikutus koko järjestelmän tehokkuuteen (Calikus et al. 2019). Lämpökuormakaaviot (heat load pattern) edustavat tyypillisimpiä käyttäytymismalleja kaukolämpöverkossa, ja antavat tieto siitä, miten eri asiakasryhmät käyttävät lämpöä. Tällaisten kaavioiden analysointi on olennaista tehokkaan kaukolämpöjärjestelmän toiminnan ja -hallinnan kannalta (Noussan et al. 2017). Asiakasanalyysin avulla kaukolämpöyritykset voivat optimoida toimintaansa, ottaa käyttöön uusia ohjausstrategioita ja personoida kysynnän hallintaa tietyille asiakasryhmille (Calikus et al. 2019).

Toinen tärkeä asia asiakkaiden lämmönkäytön analysoinnissa on tunnistaa poikkeava kulutus, koska yksikin ongelmallinen asiakas voi vaikuttaa kaukolämpöverkon yleiseen suorituskykyyn (Calikus et al. 2019). Tyypillisten ja poikkeavien lämmönkäyttöjen löytäminen on monimutkainen tehtävä erityisesti kaukolämpöjärjestelmissä, joissa on monia asiakkaita, ja joilla on erilaiset ominaisuudet. Lämmöntarpeeseen voivat vaikuttaa useat eri tekijät. Lämmöntarve määräytyy muun muassa ulkolämpötilan, sisälämpötilan, rakennuksen materiaalien, rakennuksen rakenteen, säätilan ja yksilöllisen käyttäytymisen mukaan (Ma et al. 2014). Yksilöllinen käyttäytyminen on jokaiselle rakennukselle omaa, ja siihen voi vaikuttaa moni eri tekijä. Näitä eri tekijöitä voivat olla muun muassa toinen lämmitysjärjestelmä kuten takka tai ilmalämpöpumppu, lämpimän käyttöveden kulutus, rakennuksessa olevien ihmisten määrä, erilaiset lämmittävät sähkölaitteet, rakennuksen oma ilmanvaihto sekä rakennukseen kohdistuva tuuli. Koskelainen et al. (2006) arvioi, että asuinkiinteistön vuosittainen energiankulutus jakaantuu siten että, huonetilojen lämmitykseen kuluu 40 %, ilmanvaihdon lämmitykseen kuluu 35 % ja käyttöveden lämmitykseen kuluu 25 %.

Kaukolämpöasiakkaiden kulutuksen ymmärtäminen, optimointi ja seuranta vaatii, että asiakkaille muodostetaan keskimääräistä kulutusta edustava profiili. Tämän avulla kaukolämpöjärjestelmän toimintaa voidaan tehostaa koko kaukolämpöverkon sekä yksittäisen rakennuksen tasolla. Muodostetuista profiileista voidaan muodostaa ryhmiä, mitkä edustavat ryhmään kuuluvien kaukolämpöasiakkaiden tyypillistä lämmönkäyttöä. Näitä ryhmiä voidaan hyödyntää kaukolämmön tuotannon ja -jakelun optimoinnissa, poikkeavan kulutuksen tunnistamisessa sekä hinnoittelun kehittämisessä.

3 Asiakkaiden klusterointi

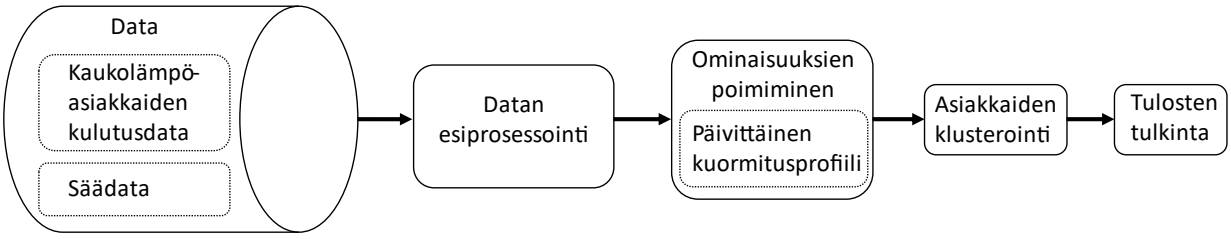
Objektien klusterointi (ryhmittely) on yhtä vanha, kuin ihmisen tarve kuvata ihmisten ja objektien keskeisiä ominaisuuksia, ja tunnistaa niille oma tyyppi. Siksi se kattaa useita tieteenaloja: matematiikasta ja tilastotieteestä biologiaan ja genetiikkaan, joista jokainen käyttää eri termejä kuvaamaan klusteroinnin avulla muodostettuja topologioita. Biologisista ”taksonomioista” lääketieteellisiin ”oireyhtymiin”, ja geneettisistä ”genotyypeistä” tuotannon ”ryhmäteknologioihin”. Kaikilla ongelma on identtinen: kategorioiden muodostaminen entiteeteille, ja yksilöiden osoittaminen oikeisiin ryhmiin (Rokach & Maimon 2005). Jain (2010) jakoi klusteroinnin kolmeen päätarkoitukseen:

- Taustalla oleva rakenne: saada tietoa datasta, luoda hypoteeseja, havaita poikkeavuuksia ja tunnistaa keskeisiä piirteitä.
- Luonnollinen luokittelu: tunnistaa lajien tai organismien samankaltaisuutta (fylogeneettinen suhde).
- Tiivistys: menetelmä tietojen järjestämiseen ja yhteenvetoon klusteriprototyyppien avulla.

Klusterointi ja luokittelu on yksi tärkeimmistä keinoista suuren tietomäärän hallinnoinnissa ja analysoinnissa (Xu & Wunsch 2005). Klusterointi ja luokittelu ovat molemmat tiedonlouhinnan perustehtäviä. Luokittelua käytetään enimmäkseen ohjattuna oppimismenetelmänä, ja klusterointia ohjaamatonta oppimista varten. Klusteroinnin tavoite on kuvailla dataa, ja luokittelun tavoite on ennustaa dataa (Veyssieres & Plant 1998.) Koska ryhmittelyn tavoitteena on löytää uusia kategorioita, uudet ryhmät ovat itsessään kiinnostavia ja niiden arviointi on luontaista. Luokittelutehtävissä tärkeä osa arvioinnista on kuitenkin ulkoista, koska ryhmien tulee heijastaa joitain luokkien referenssijoukkoa (Rokach & Maimon 2005.)

Tässä tutkielmassa pyritään muodostamaan kaukolämpöasiakkaille ominaisia kulutustottumuksia vuorokauden aikana, ja niiden perusteella muodostetaan ryhmiä (klustereita), missä samankaltaiset käyttäjät ovat samassa ryhmässä ja erilaiset toisessa.

Asiakkaiden klusteroinnin prosessi esitetään kuvassa neljä. Ensimmäiseksi analysoidaan käytössä olevaa kaukolämmön kulutusdataa. Data esiprosessoidaan puutteiden korjaamiseksi. Tämän jälkeen datasta poimitaan ominaisuudet, minkä perusteella klusterointi suoritetaan. Ominaisuuksien muodostamisen jälkeen asiakkaat klusteroidaan, jonka jälkeen analysoidaan muodostetut klusterit.



Kuva 4. Klusteroinnin prosessi.

3.1 Kaukolämmön kulutusdata

Tässä tutkielmassa käytetään kuopiolaisten kaukolämpöasiakkaiden kaukolämmön kulutusdataa vuodelta 2021, sekä Ilmatieteenlaitoksen tarjoamat ulkolämpötila-arvot Kuopion Savilahden alueelta. Mittaukset on otettu tunnin välein jokaisesta kohteesta, ja ne ilmaisevat lämpöenergian käyttöä megawattitunteina. Datassa olevia kaukolämpöasiakkaita on yhteensä 6089, mutta kun suodattaa pois epävarmat ja epäonnistuneet lukemat, asiakkaiden määräksi tulee 6084, ja niiden jakauma rakennustyypeittäin esitetään taulukossa 1.

Kaukolämpömittauksia on yhteensä 54 657 082 kappaletta. Jokaisessa mittauksessa on statusarvo, joka kertoo mittauksen laadusta. Statukset ovat MSCONS (metered service consumption report) -sanomapalvelun mukaisilla koodeilla, joista datassa esiintyvä ”136” tarkoittaa onnistunutta lukemaa, ”Z02” tarkoittaa epävarmaa lukemaa ja ”Z03” tarkoittaa, että lukemaa ei ole saatu (Ediel forum 2010). Lukemien jakautuminen eri statuksille esitetään taulukossa 2. Koska mittaukset tehdään tunnin välein, niin jokaiselle käyttöpaikalle tulisi olla 24 energiankäyttö lukemaa jokaisena päivänä. Datasta kuitenkin löytyi 733 885 puuttuvaa lukemaa, joita ei ole datassa ollenkaan.

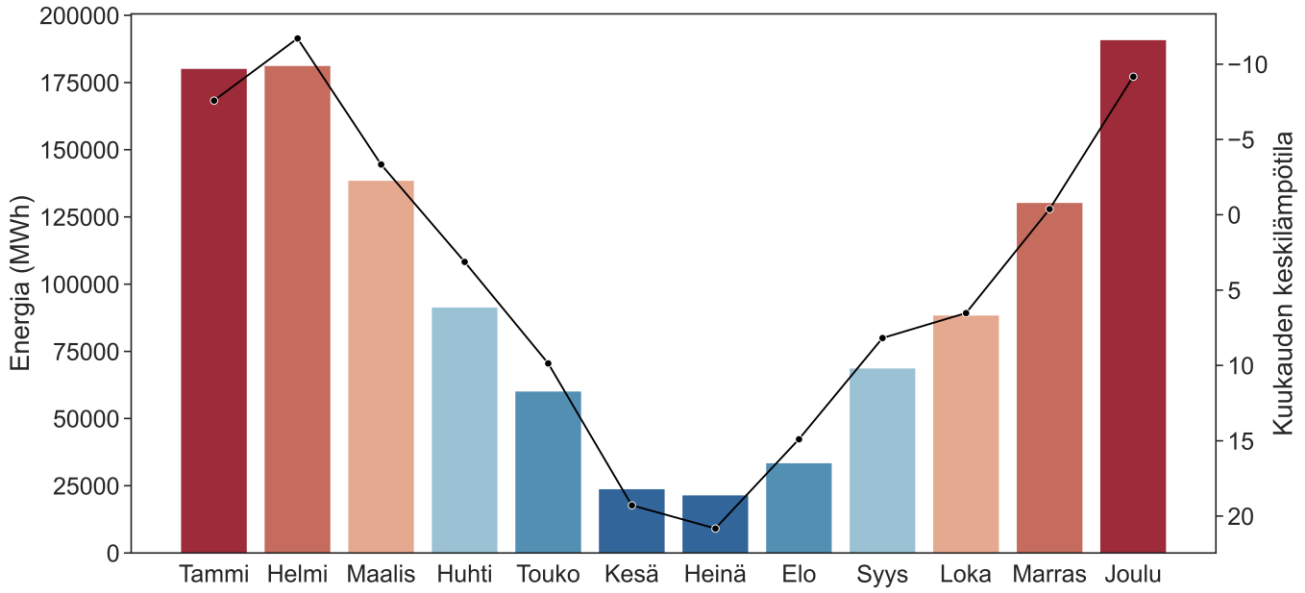
Taulukko 1. Kaukolämpöasiakkaiden jakauma rakennustyypeittäin.

Rakennustyyppi	**Kohteiden määrä**
Pientalo	3777
Rivi- ja kerrostalot	1602
Palvelukiinteistö	324
Julkiset palvelut	213
Teollisuusrakennukset	131
Kuljetusalan kiinteistö	20
Muu	17
Yhteensä	6 084

Taulukko 2. Kaukolämpölukemien määrä eri statuksilla.

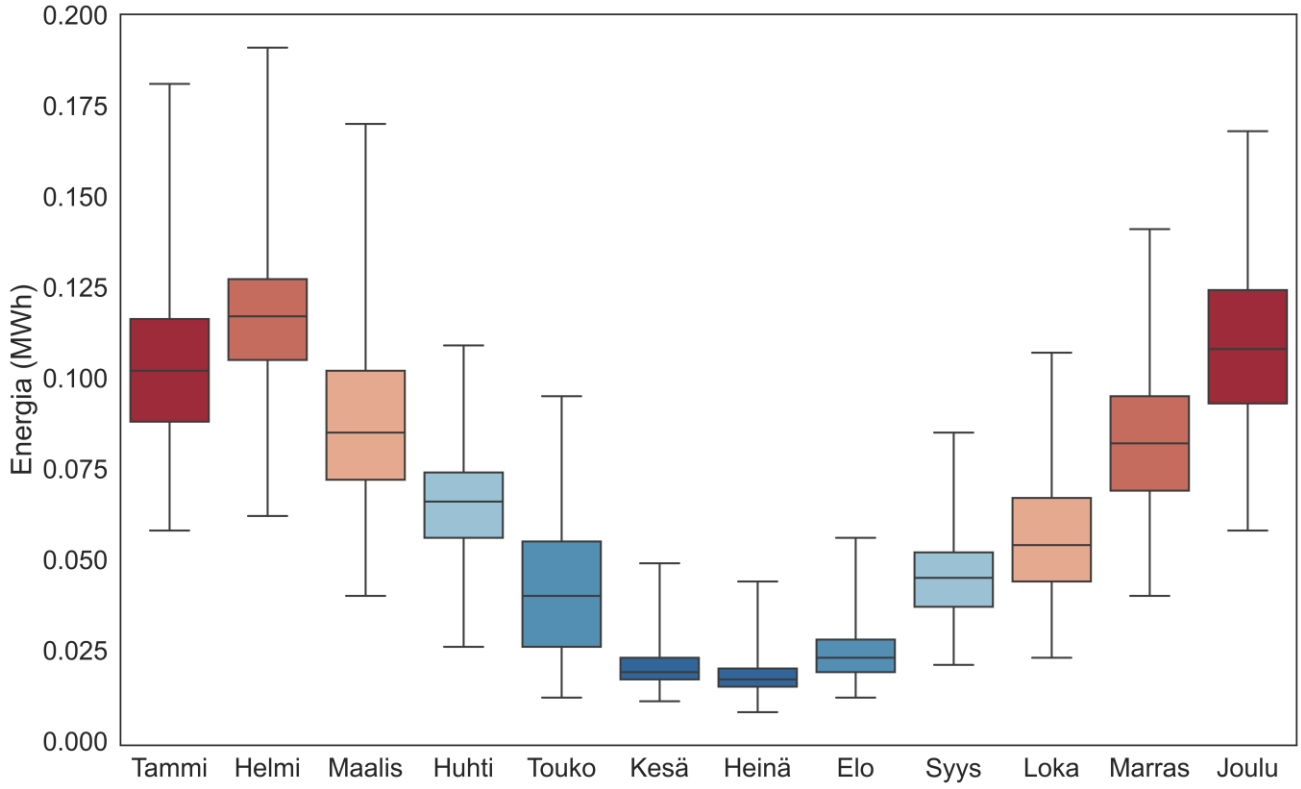
Status	**Lukemien määrä**
136	54 557 377
Z02	48
Z03	9 965
puuttuvia	733 885
Yhteensä	55 301 275

Kuvasta 5 nähdään kaukolämmön kokonaiskäyttö kuukausittain, sekä kuukausien keskilämpötila. Arvot korreloivat keskenään hyvin, koska lämmönkäytön määrä seuraa keskimääräistä ulkolämpötilaa. Kuvaajasta nähdään, että lämpöä käytetään myös kuumina kesäkuukausina. Kesäkuukausien keskilämpötilat ovat noin 20 °C, ja silti kaukolämpöä käytetään. Tämä lämpö ei mene tilojen lämmittämiseen vaan lämpimän käyttöveden lämmittämiseen. Koska lämmintä käyttövettä käytetään keskimäärin yhtä paljon ulkolämpötilasta riippumatta, voidaan nähdä sen lämmittämiseen kuluva energia kuvan 5 kesä- ja heinäkuun käytön määrästä.



Kuva 5. Kaukolämmön käyttö vuoden 2021 aikana kuukausittain, sekä ulkolämpötilan kuukauden keskiarvo. Energian käyttö on ilmoitettu pylväinä ja ulkolämpötilan keskiarvo on mustalla viivalla.

Kuvasta 6 nähdään yksittäisen asiakkaan energiankäyttö vuoden aikana eri kuukausina. Kuukauden kulutus esitetään viiksilaatikkoina (box plot), josta nähdään tilastollisia tunnuslukuja. Laatikon sisällä oleva viiva kertoo kuukauden kulutuksen mediaanin, laatikon ylä- ja alareuna näyttää ylä- ja alakvartaalin. Laatikosta lähtevät viivat näyttävät kuukauden minimi- ja maksimikäytön. Kuvasta 6 nähdään, että kuukausien energiankäyttö noudattaa hyvin samanlaista kaavaa kokonaisenergiankulutuksen kanssa (Kuva 5). Lisäksi nähdään, että energiankäytön hajonta on merkittävästi pienempää kuukausina, milloin ei lämmitetä rakennusta (kesäkuukaudet) verrattuna lämmityskuukausiin.



Kuva 6. Yksittäisen kohteen energiankäyttö kuukausittain. Yksittäisen kuukauden energiankäyttö on esitetty viiksilaatikolla.

3.2 Esiprosessointi

Datan esiprosessointi on yksi tiedonlouhinnan vaiheista, mikä sisältää datan valmistelun ja muuntamisen sopivaan muotoon. Huolellinen datan esiprosessointi mahdollistaa parempien lopputulosten saamisen, ja voi olla ehto oikean tuloksen saamiseksi, varsinkin luokittelutehtävissä.

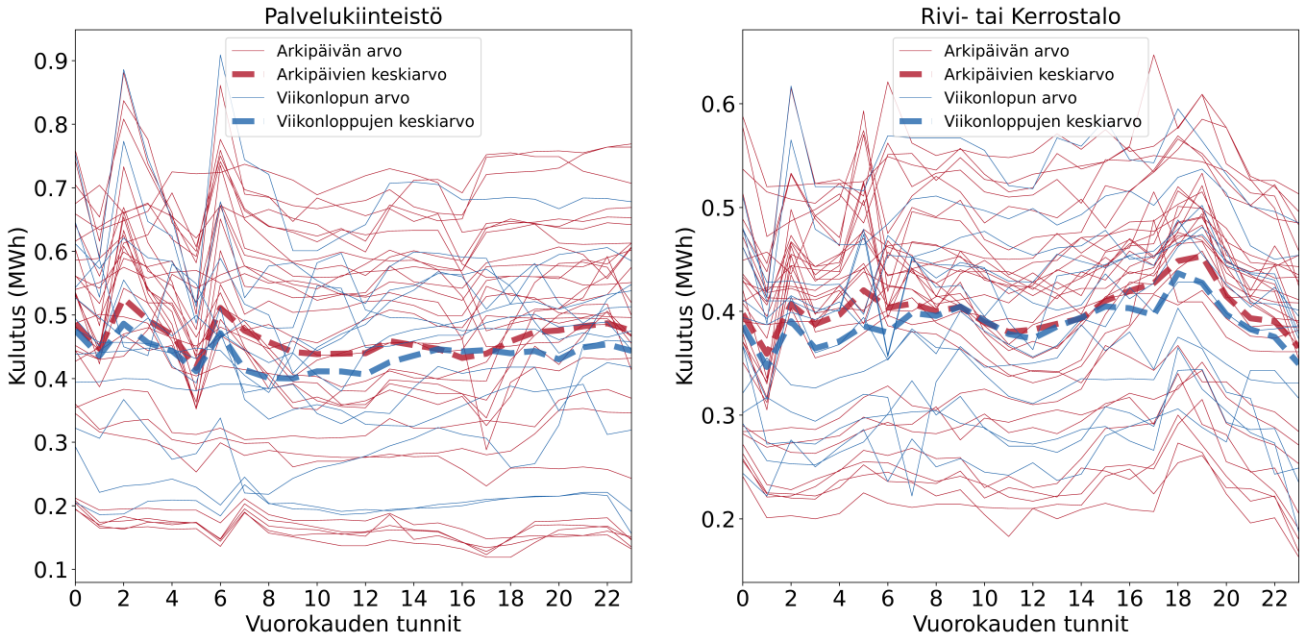
Kaukolämpödata sisältää suhteellisen suuren määrän puuttuvia arvoja, jotka vaativat käsittelyä. Datasta löytyi 23 eri ajanhetkeä, joista puuttui ulkolämpötila-arvo, mikä on yhteensä 139 817 puuttuvaa ulkolämpötila-arvoa. Luvut näille puuttuville ajanhetkille haettiin ja lisättiin tietojoukkoon ilmatieteen laitoksen avoimen datan palvelusta (Ilmatieteen laitos 2023). Tietojoukossa oli yksi virheellinen energialukema, joka oli epärealistisen suuri. Se johtuu todennäköisimmin mittarivirheestä. Tämä lukema suodatettiin pois. Tarkastelussa keskitytään vuorokauden sisäisiin lämpötilan vaihteluihin, joten on tärkeää, että data sisältää vuorokauden

sisäiset lukemat ilman merkittäviä puutteita. Datasta poistetaan niiden vuorokausien lukemat, joissa asiakkaan tiedoista puuttuu enemmän kuin neljä lukemaa. Muissa tilanteissa puuttuvat lukemat luodaan lineaarisella interpolaatiolla.

3.3 Ominaisuuksien valitseminen

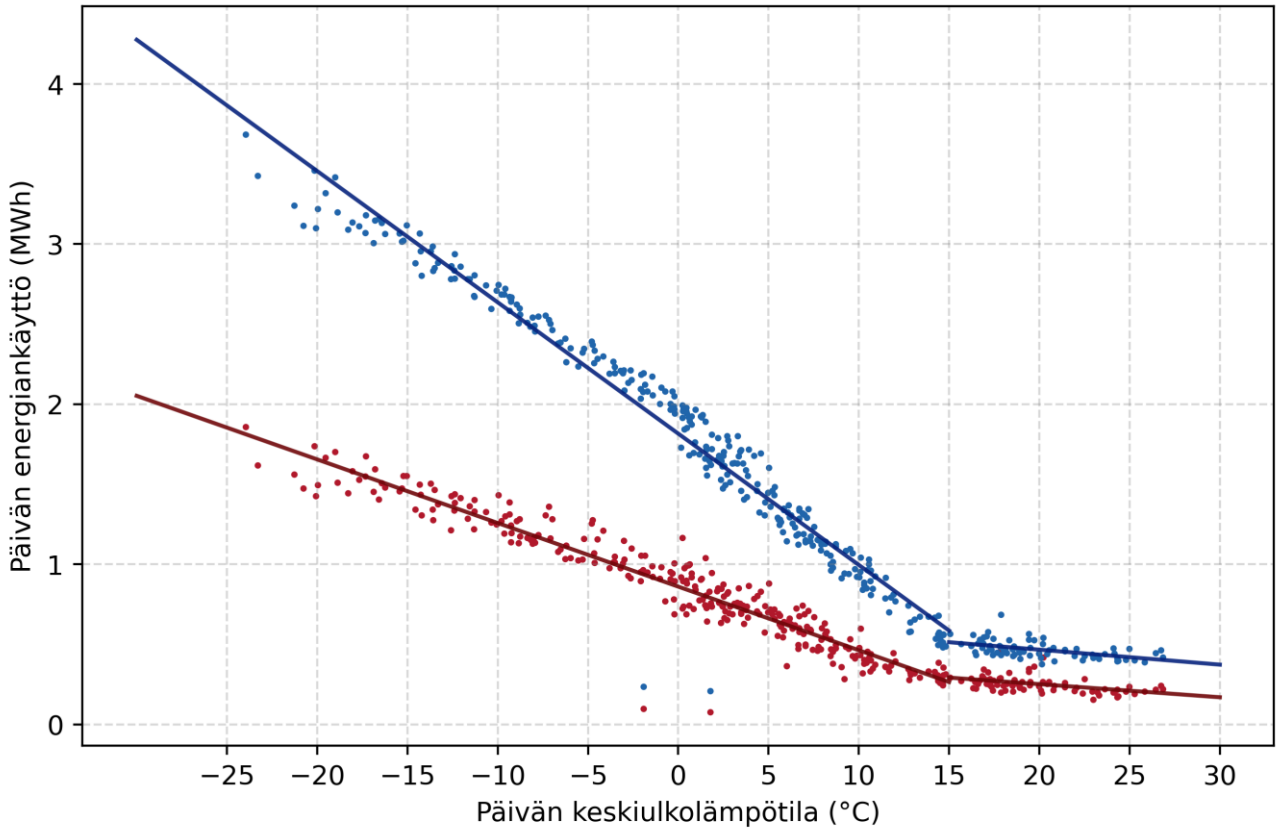
Tässä tutkielmassa pyritään muodostamaan kaukolämpöasiakkaille vuorokauden aikana tapahtuvaa keskimääräistä kulutusta, ilman ulkolämpötilan vaikutus. Tämän takia ominaisuuksien valitsemisessa tiedoista tulee poimia ulkolämpötiloista riippumaton käyttö.

Kuvasta 7 nähdään kahden eri rakennustyypin asiakkaan päivittäinen lämmönkäyttö kuutena lämmityskuukautena. Kuvassa arkipäivät ja viikonlopun päivät on merkitty eri väreillä (arki punaisella ja viikonloppu sinisellä), ja näiden päivien keskiarvo on merkattu saman värisellä paksummalla viivalla. Kuvasta 7 nähdään selkeästi asiakkaiden erilainen lämmönkäytön tapa. Palvelukiinteistön asiakkaalla lämmityspiikit ovat useimmiten tuntien kaksi ja kuusi aikaan, kun taas rivi- tai kerrostaloasiakkaalla havaitaan kolme merkittävämpää lämmityshetkeä tuntien kaksi, viisi ja 19 aikaan. Aamulla kulutus on rivi- tai kerrostaloasiakkaalla vaihtelevampaa, eikä aamun piikit ole yhtä selviä, kuin palvelukiinteistön asiakkaalla. Kummallakin asiakkaalla käyttö on hieman pienempää viikonlopun päivien aikana, ja käyttö eroaa hieman arkipäivien käytöstä. Kummankin asiakkaan lämmönkäytön määrä vaihtelee paljon eri vuorokausien välillä, koska vuorokausien ulkolämpötilat vaihtelevat. Kylminä päivinä lämmitys on suurempaa, kuin lämpöisempinä päivinä. Vuorokausien käytöissä on kuitenkin havaittavissa toistuvia kaavoja.



Kuva 7. Kahden eri käyttäjän päivittäinen kulutus tunneittain kuutena lämmityskuukautena. Jokainen ohut viiva esittää yhden päivän kulutusta. Päivien kulutukset on jaettu arkipäiviin (punainen) ja viikonlopun päiviin (sininen). Paksu viiva kuvaa näiden päivien kulutusten keskiarvoa.

Vuorokausitasolla kuvasta 8 nähdään että, kun vuorokauden keskilämpötila on matala, niin asiakkaan lämmönkäyttö korreloi melko lineaarisesti ulkolämpötilan kanssa. Lämmönkäyttö vähenee ulkolämpötilan kasvaessa, ja kun keskiulkolämpötila on yli 15 °C rakennuksen lämmitys lopetetaan, jonka jälkeen käyttö koostuu vain käyttöveden lämmityksestä, ja joissain kohteissa kosteiden tilojen jatkuvasta lattialämmityksestä. Kun keskiulkolämpötila on yli 15 °C, lämmön käyttö ei ole kuitenkaan vakio, mitä voisi selittää se, että rakennusten asukkaat käyttävät yhä vähemmän lämmintä vettä suihkussa, kun ulkona on hyvin lämmin.

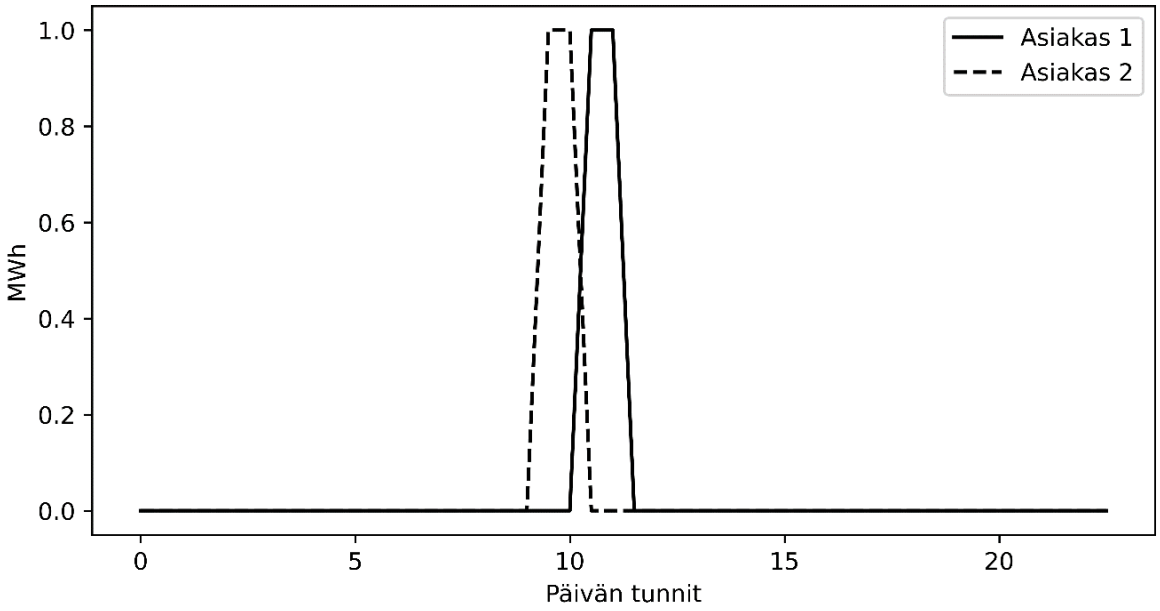


Kuva 8. Kahden eri asiakkaan päivän energiankäyttö suhteessa päivän keskiulkolämpötilaan, ja neljä sovitettua regressiosuoraa.

Kuvassa 8 olevista asiakkaista voitaisiin esimerkiksi muodostaa yksinkertainen malli lämmönkäytöstä vuorokausitasolla keskiulkolämpötilan suhteen. Malli olisi muotoa $Energia = Vakio + x * Keskiulkolämpötila$, missä $Vakio$ on muuttuja, joka edustaa lämmönkäytön vakiomäärää asiakkaalla, eli kuvan 8 viivan korkeutta, ja x edustaa ulkolämpötilan vaikutuksen suuruutta, eli kuvan 8 viivan jyrkkyyttä. Mitä suurempi x:n arvo on, sitä suurempi lämmönkäyttö ulkolämpötilojen laskiessa. Yhdelle asiakkaalle täytyy luoda kaksi eri mallia. Toinen niihin tilanteisiin, missä keskiulkolämpötila on alle 15 °C, ja toinen niihin tilanteisiin, missä keskiulkolämpötila on yli 15 °C, koska käyttö on erilaista näissä tilanteissa. Esimerkiksi kuvan 8 sinisellä merkityn asiakkaan matalan ulkolämpötilan sovitettu malli olisi $Energia = 1{,}8163 - 0{,}0820 * Keskiulkolämpötila$, ja korkean ulkolämpötilan sovitettu malli olisi $Energia = 0{,}6540 - 0{,}0093 * Keskiulkolämpötila$. Tällöin esimerkiksi, jos vuorokauden keskilämpötila olisi 5 °C käytettäisiin matalan lämpötilan mallia, ja arvioitu vuorokauden energiankäyttö olisi 1,8163 – 0,0820 * 5 = 1,4063 MWh.

Vuorokauden sisäiseen vaihteluun Wang et al. (2019) ehdotti menetelmän kuvata jokaiselle asiakkaalle tyypillinen päivittäinen lämmönkäyttö ilman ulkolämpötilan vaikutusta. Tätä lämmönkäytön ominaisuutta Wang et al. (2019) kutsui nimellä muutettu päivittäinen kuormitusprofiili (modified daily load profile, lyh. MDLP). MDLP edustaa kaukolämpöasiakkaan keskimääräistä lämmönkulutusta vuorokauden aikana ilman ulkolämpötilan vaikutusta. Se paljastaa rakennukselle yksilöllisiä kulutustottumuksia, mihin vaikuttaa esimerkiksi lämpimän käyttöveden käyttö sekä ilmanvaihto.

Jokaisen päivän lämmönkäyttö katsotaan koostuvan kahdesta pääosasta. Ensimmäinen vastaa tilan lämmitystä. Kun ulkolämpötila alittaa tietyn kynnyksen T_{sh}, kuluu merkittävä määrä lämpöä, joka on karkeasti lineaarisesti verrannollinen ulkolämpötilaan. Toinen osa lämmönkulutuksesta koostuu säännöllisistä käyttäjäkohtaisista toiminnoista tilojen lämmityksen lisäksi. Yleisin tällainen toiminto käyttöveden lämmittäminen. Kulutuksen toisen osan määrä oletetaan liittyvän vuorokaudenaikaan, ja kulutuskäyttäytyminen voi olla samanlaista eri päivinä. (Wang et al. 2019)



Kuva 9. Double-penalty-ongelma.

Eri käyttäjillä kulutus voi olla samanlaista, mutta se voi tapahtua hieman eri aikaan yksilöllisten rytmien takia. Esimerkiksi ihmiset heräävät hieman eri aikoihin, ja sen seurauksena käyvät suihkussa hieman eri aikoihin, mikä vaikuttaa lämmön käytön määrään. Tätä kutsutaan double-penalty-ongelmaksi (Kuva 9). Sen vaikutuksen vähentämiseksi jaetaan jokainen päivä n aikaväliin, joista jokainen sisältää $\mu = \frac{24}{n}$ tuntia. Käytetään kvantitatiivista muuttujaa H, joka kuvaa lämmön kokonaiskulutusta yhdessä aikavälissä n. Käytetään kvantitatiivista muuttujaa T ilmaisemaan kyseisen aikavälin keskimääräinen ulkolämpötila, ja käytetään n-tason kategorista muuttujaa A ilmaisemaan kyseisen aikavälin indeksiä. Käytetään muuttujaa T_{sh} kuvaamaan sitä ulkolämpötilaa, minkä yli ei enää lämmitetä rakennuksen tiloja. Kun vuorokauden keskimääräinen ulkolämpötila $T \leq T_{sh}$, H:n, T:n ja A:n välinen suhde mallinnetaan käyttämällä monen selittäjän lineaarista regressiomallia:

$$\mathrm{H} = \beta_0^l + \beta_1^l (T_{sh} - T) + \beta_2^l \mathrm{A} + \epsilon^l, \quad (1)$$

missä regressiokerroin β_1^l tarkoittaa keskimääräistä kulutuksen muutosta, kun lämpötila T muuttuu yhden asteen, β_0^l tarkoittaa matalan lämpötilan mallin vakiotermiä, β_2^l kvantifioi A:n ja H:n välisen yhteyden, ja ϵ^l on virhetermi, mikä heijastaa tuntemattoman tekijän vaikutusta. (Wang et al. 2019)

Kun vuorokauden keskimääräinen ulkolämpötila $T > T_{sh}$ sisätilan lämmitys loppuu, kuten kuvasta 8 voidaan nähdä, että korkeissa lämpötiloissa lineaarinen riippuvuus käytön H ulkolämpötiloissa vähenee merkittävästi. H:n ja A:n välinen suhde mallinnetaan käyttämällä yksinkertaista regressiomallia:

$$\mathrm{H} = \beta_0^h + \beta_2^h \mathrm{A} + \epsilon^h, \quad (2)$$

missä β_0^h tarkoittaa korkean lämpötilan mallin vakiotermiä, β_2^h kvantifioi A:n ja H:n välisen yhteyden, ja ϵ^hon virhetermi. (Wang et al. 2019)

Kategoristen arvojen sisällyttämiseksi lineaariseen regressioon, malleissa muuttuja A esitetään n:llä valebinäärimuuttujalla (dummy variable): A asetetaan n-ulotteiseksi vektoriksi $[\alpha_1,$

$\alpha_2, \dots, \alpha_n]^T$, missä $\alpha_i \in \{0, 1\}$ ja $\sum_{i=1}^{n} \alpha_i = 1$. i. alkio $\alpha_i = 1$ tarkoittaa, että aikaindeksi on i. Vastaavasti matalan lämpötilan mallissa $\beta_2^l = [\beta_{2,1}^l, \beta_{2,2}^l, \dots, \beta_{2,n}^l]$ ja korkean lämpötilan mallissa $\beta_2^h = [\beta_{2,1}^h, \beta_{2,2}^h, \dots, \beta_{2,n}^h]$ ovat kukin n-ulotteisia vektoreja, jonka i. elementti heijastaa aikaindeksillä i käyttäjäkohtaisen käyttäytymisen vaikutusta lämmönkulutukseen. (Wang et al. 2019)

Jokaiselle käyttäjälle on yhteensä $2n$ tuntematonta kerrointa malleissa (1) ja (2), $\beta_0^l, \beta_1^l, \beta_{2,1}^l \dots, \beta_{2,n}^l$, $\beta_0^h, \beta_1^h, \beta_{2,1}^h \dots, \beta_{2,n}^h$, jotka tulee oppia datasta. Kuten Chelmis et al. (2015) tutkimuksessa erotetaan työ- ja lomapäivien mallit toisistaan. Työpäiviä ovat arkipäivät ja lomapäiviä ovat viikonlopun päivät. Käytetään $\hat{T}_{sh} = 15\ °C$, joka on päätelty kuvasta 8, koska siinä pisteessä asiakkaan tilojen lämmitys loppuu. Tämän jälkeen mallit (1) ja (2) sovitetaan kaikkien työpäivien lämmönkäytön ja ulkolämpötilojen kanssa. Mallin sovittamisessa käytetään R-ohjelmointikielen (R Development Core Team 2005) lm-funktiota, joka sovittaa mallin hyödyntäen QR-matriisihajotelmaa. Regressiokertoimien estimaatit on merkitty $\hat{\beta}_0^l$, $\hat{\beta}_1^l$, $\hat{\beta}_{2,i}^l, \dots, \hat{\beta}_{2,n}^l, \hat{\beta}_0^h, \hat{\beta}_{2,1}^h, \dots, \hat{\beta}_{2,n}^h$.

Poistetaan ne lämmönkäytöt, jotka noudattavat lineaarisesti ulkolämpötilaa. Kun ulkolämpötila on alempi kuin $\hat{T}_{sh}$ arvot $\hat{\beta}_0^l + \hat{\beta}_{2,i}^l$ esittävät keskimääräistä kulutusta ajan hetkellä i ($i \in \{1, 2, \dots, n\}$) jokaisella päivällä ilman ulkolämpötilan lineaarista vaikutusta. Kun ulkolämpötila on suurempi kuin $\hat{T}_{sh}$ kulutusarvot ovat $\hat{\beta}_0^h + \hat{\beta}_{2,i}^h$. Siten tyypillinen arkipäivien käyttö kaukolämpöasiakkaalla hetkellä i merkitään $\hat{u}_i$, joka voidaan ottaa matalan- ja korkean lämpötilan kulutuksen summana:

$$\hat{u}_i = \hat{\beta}_0^l + \hat{\beta}_{2,i}^l + \hat{\beta}_0^h + \hat{\beta}_{2,i}^h. \quad (3)$$

Samalla lailla lineaariregressiomallit (1) ja (2) sovitetaan viikonlopun päivien lämmönkäytön ja ulkolämpötilojen kanssa, ja saadaan regressiokertoimet
$\tilde{\beta}_0^l$, $\tilde{\beta}_1^l$, $\tilde{\beta}_0^l$, …, $\tilde{\beta}_{2,n}^l$, $\tilde{\beta}_0^h$, $\tilde{\beta}_{2,1}^h$, …, $\tilde{\beta}_{2,n}^h$, ja $\tilde{T}_{sh}$. Kaukolämpöasiakkaan tyypillinen viikonlopun päivien kulutus ajan hetkellä i saadaan:

$$\tilde{u}_i = \tilde{\beta}_0^l + \tilde{\beta}_{2,i}^l + \tilde{\beta}_0^h + \tilde{\beta}_{2,n}^h. \quad (4)$$

Wang et al. (2019) tutkimusta seuraten olkoon $u = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n, \tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n]$. Tällöin u edustaa käyttäjän päivittäistä kulutusmallia, kun lämpötilaan lineaarisesti korreloiva kulutus poistetaan. Määritetään MDLP u:n normalisoiduksi versioksi, jotta se kuvastaa vain kunkin käyttäjän vaihtelevaa suuntausta absoluuttisten arvojen sijaan. Normalisointi tehdään kaavalla:

$$a = \frac{u - mean\{u\}}{std\{u\}}. \quad (5)$$

Nyt a:ta voidaan pitää $2n$-ulotteisena ominaisuusvektorina käyttäjälle. Asiakkaat voidaan klusteroida niiden tyypillisen päivittäisen kulutusprofiilin mukaan, jotka on luotu a:n perusteella.

Muodostetuissa kulutusmalleissa jotkin $\hat{u}_i$ ja $\tilde{u}_i$ eivät saaneet arvoa. Nämä kulutusmallit, joissa esiintyi tyhjiä arvoja, poistettiin tietojoukosta. Poistamisen jälkeen mallien määrä tippui 6085 kappaleesta lopulliseen 6079 kappaleeseen, jotka otetaan mukaan klusterointiin.

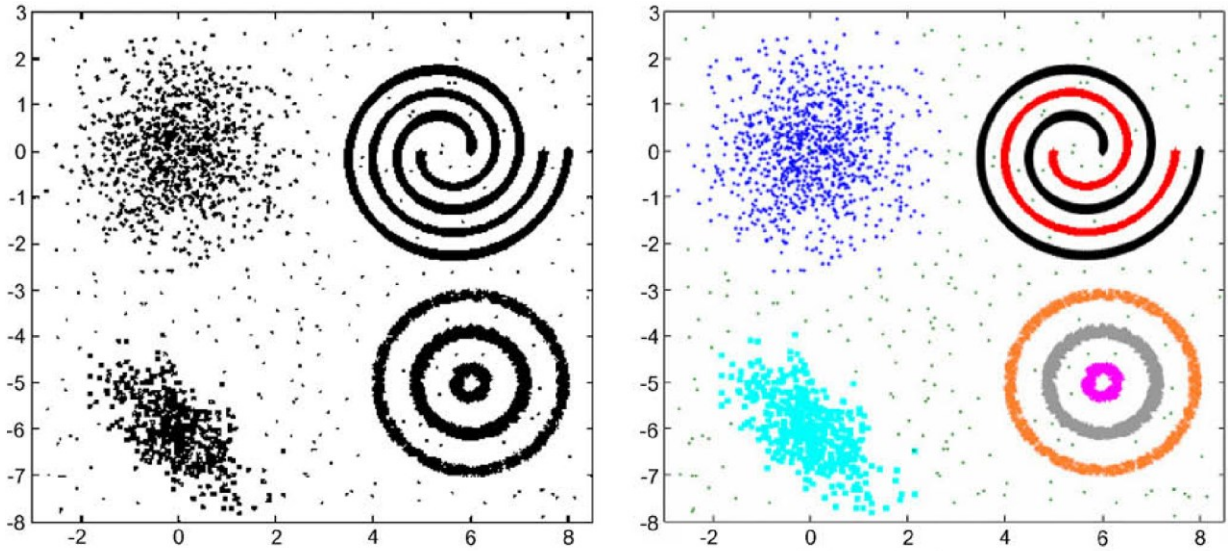
3.4 Klusterointimenetelmä

Klusteroinnin (tunnetaan myös klusterianalyysinä) tavoitteena on löytää luonnollisia ryhmiä hahmojen, pisteiden tai objektien joukosta (Jain 2010). Merriam-Webster (2023) määritteli klusterianalyysin tilastolliseksi luokittelumenetelmäksi, jolla selvitetään, että kuuluvatko populaation yksilöt eri ryhmiin tekemällä kvantitatiivisia vertailuja useista ominaisuuksista. Kuvassa 10 esitetään esimerkki klusteroinnista. Tavoitteena on löytää luonnolliset ryhmät (Kuva 10b) automaattisesti luokittelemattomasta datasta (Kuva 10a).

Klusteroinnin operationaalinen määritelmä voidaan todeta seuraavasti. N objektin esityksessä etsitään K ryhmää valitun samankaltaisuusmittarin perusteella siten, että saman ryhmän jäsenten samankaltaisuus on suuri, kun taas samankaltaisuus eri ryhmien jäseniin on pieni (Jain, 2010). Joskus tavoitteena on järjestää klusterit luonnolliseksi hierarkiaksi. Se edellyttää itse

klusterien ryhmittelyä peräkkäin siten, että jokaisella hierarkian tasolla saman ryhmän klusterit ovat samankaltaisempia toistensa kanssa kuin eri ryhmissä olevat (Hastie et al. 2009.)

Muodollisesti klusterointirakenne voidaan esittää S:n alijoukkojen joukkona $C = C_1, \dots, C_k$ siten, että: $S = U_{i=1}^{k} C_i$ ja $C_i \cap C_j = \emptyset$ kun $i \neq j$. Näin ollen mikä tahansa esiintymä S:ssä kuuluu täsmälleen yhteen ja vain yhteen osajoukkoon (Rokach & Maimon, 2005.) Kuva 10 esittää, että klusterit voivat erota toisistaan muodon, koon ja tiheyden mukaan. Datassa oleva kohina (noise) tekee klusterien havaitsemisesta vielä vaikeampaa. Ihanteellinen klusteri voidaan määritellä joukoksi pisteitä, jotka ovat kompakteja ja eristettyjä. Todellisuudessa klusteri on subjektiivinen kokonaisuus, jonka merkitys ja tulkinta vaatii toimialan tietämystä (Jain 2010.)



Kuva 10. Klusterien monimuotoisuus. Kuvan a-kohdassa seitsemän klusteria on merkitty kuvan b-kohdassa eri väreillä. Klusterit eroavat muodoltaan, kooltaan ja tiheydeltään. Vaikka nämä klusterit ovat selviä data-analyytikolle, mikään käytettävissä olevista klusterointialgoritmeista ei pysty havaitsemaan kaikkia näitä klustereita. (Jain 2010)

Klusterointialgoritmit voidaan jakaa laajasti kahteen ryhmään: hierarkkisiin ja osittaviin (Jain 2010). Hierarkkiset klusterointimenetelmät muodostavat klusterit osioiden rekursiivisesti joko ylhäältä alas tai alhaalta ylös. Nämä menetelmät voidaan jakaa agglomeratiiviseen hierarkkiseen klusterointiin ja jakautuvaan hierarkkiseen klusterointiin. Agglomeratiivisessa hierarkkisessa klusteroinnissa jokainen objekti edustaa aluksi omaa klusteriaan. Sitten klustereita yhdistetään

peräkkäin, kunnes haluttu klusterirakenne saadaan. Jakavassa hierarkkisessa klusteroinnissa kaikki objektit kuuluvat aluksi yhteen klusteriin. Sitten klusteri jaetaan aliklusteriin, jotka jaetaan peräkkäin omaan aliklusteriinsa. Tätä prosessia jatketaan, kunnes haluttu klusterirakenne on saavutettu. Hierarkkisten menetelmien tulos on dendrogrammi, joka edustaa objektien sisäkkäistä ryhmittelyä ja samankaltaisuustasoja, joilla ryhmittelyt muuttuvat. Dataobjektien klusterointi saadaan leikkaamalla dendrogrammi halutulle samankaltaisuustasolle. (Gupta & Jain 2014)

Suosittuja hierarkkisia klusterointialgoritmeja ovat Average linkage ja Wardin menetelmä, ja kaikista tunnetuin ja yksinkertaisin osittava klusterointialgoritmi on k-means. Syitä k-means algoritmin suosioon on, että: sen käyttöönotto on yksinkertaista ja helppoa, sen rajoitteet tunnetaan paremmin, kuin mahdollisesti parempien, mutta vähemmän tutkittujen menetelmien, joilla voi olla tuntemattomia tai piilotettuja rajoitteita, ja sen paikallinen hienosäätö on hyvin tehokas, sekä empiirinen menestys ovat tärkeimmät syyt sen suosiolle (Jain 2010; Fränti & Sieranoja 2019.)

k-means algoritmi voidaan selittää Jain (2010) mukaan seuraavasti. Olkoon $X = x_i,\ i = 1,\ \ldots,\ n$ on joukko n d-ulotteisia pisteitä, jotka klusteroidaan K klusterin joukkoon, $C = \{C_k,\ k = 1,\ \ldots,\ K\}$. k-means algoritmi löytää osituksen siten, että klusterien empiirisen keskiarvon ja klusterien pisteiden välinen neliövirhe minimoidaan. Olkoon μ_k klusterin c_k keskiarvo. μ_k ja klusterissa c_k olevien pisteiden neliövirhe on määritelty

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - u_k\|^2 .$$

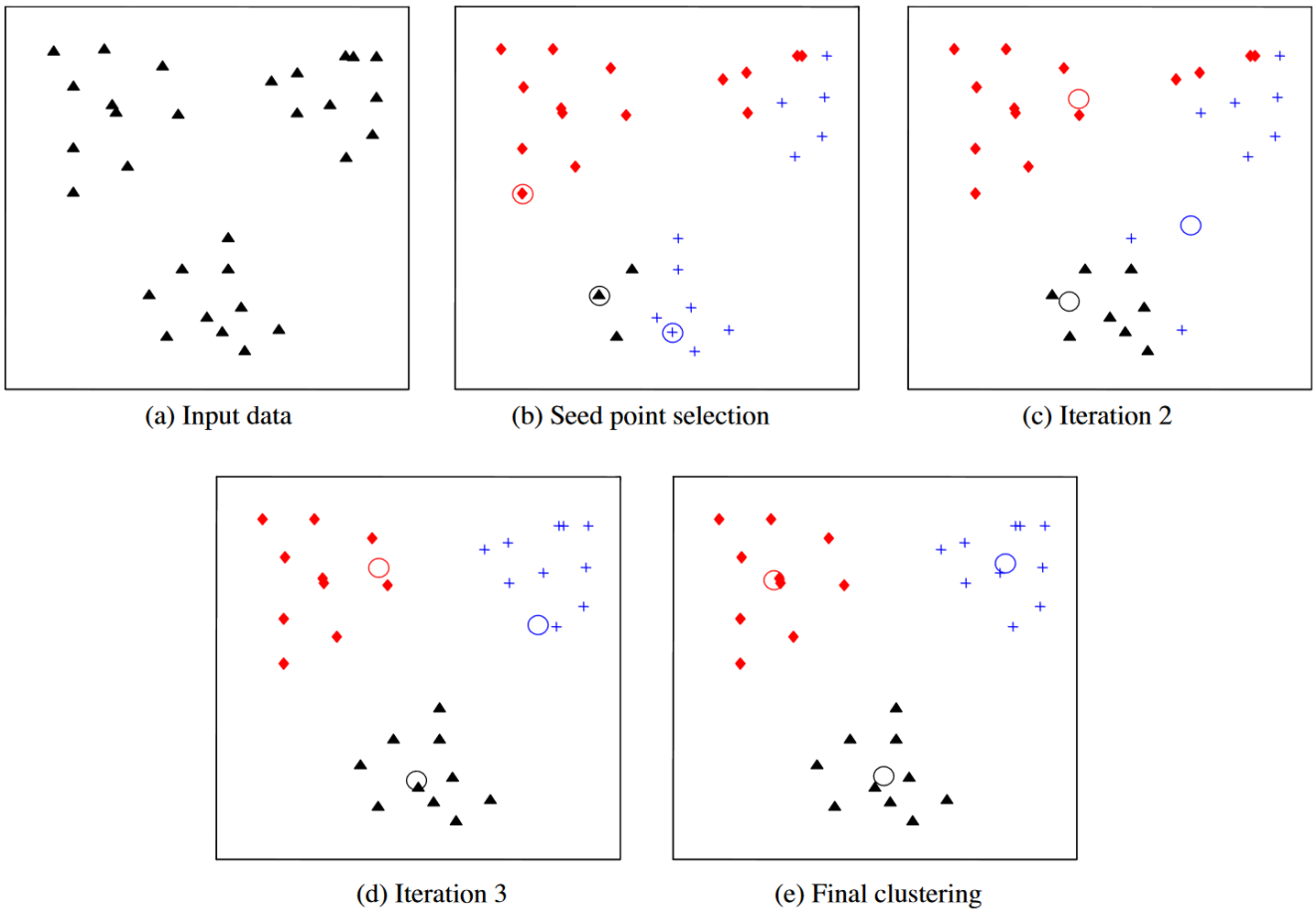
k-means:in tavoite on minimoida neliövirheen summa kaikilla K klustereilla,

$$J(C) = \sum_{k=1}^{k} \sum_{x_i \in C_k} \|x_i - u_k\|^2 .$$

k-means alkaa alkuperäisellä osituksella, jossa on K klusteria, ja määrittää klustereille pisteet nelivirheen pienentämiseksi. Koska neliövirhe pienenee aina klusterien lukumäärän K kasvaessa ($J(C) = 0$, kun $K = n$), se voidaan minimoida vain kiinteälle määrälle klustereita (Jain & Dubes 1988.) k-means-algoritmin toiminta voidaan kuvata seuraavasti:

1. Valitse K satunnaista ryhmän edustajaa.
2. Ryhmitä alkiot lähimpään edustajaansa ja laske neliövirhe.
3. Määritä edustajille uusi piste laskemalla sille nimettyjen pisteiden keskiarvo.
4. Aloita uudestaan kohdasta 2 kunnes kohdassa 3 edustajien pisteiden arvo ei muutu.

Kuvassa 11 esitetään k-means-algoritmin vaiheet kaksiulotteisessa tietojoukossa, jossa on kolme klusteria.



Kuva 11. Kuva k-means-algoritmista. (a) Kaksiulotteinen syötedata, jossa on kolme klusteria; (b) kolme siemenpistettä, jotka on valittu klusterikeskuksiksi, ja datapisteiden alkuperäinen kohdistaminen klustereihin; (c) ja (d) väli iteraatiot klusteritunnisteiden ja niiden keskusten

päivittämiseksi; (e) k-means-algoritmilla saatu lopullinen klusterointi konvergoituneena. (Jain 2010)

k-means tarvitsee kolme käyttäjän määrittämää parametria: klustereiden määrä K, klusterien alustus ja etäisyysmitta. Kaikista kriittisin valinta on K. Tyypillisesti k-means suoritetaan itsenäisesti eri K:n arvoilla ja valitaan ositus, joka vaikuttaa toimialueen asiantuntijan kannalta merkityksellisimmältä. Erilaiset alustukset voivat johtaa erilaisiin lopullisiin klustereihin, koska k-means konvergoituvat vain paikalliseen minimiin. Yksi tapa päästä yli paikallisesta minimistä on suorittaa k-means-algoritmi tietylle K:lle useilla eri aloitusosioilla ja valita osio, jolla on pienin neliövirhe. Etäisyyden mittana käytetään tyypillisesti euklidisista etäisyyttä pisteiden ja klusterin keskusten välisen etäisyyden laskemiseen, mutta muitakin metriikoita käytetään erilaisissa sovelluksissa. (Jain 2010)

Erilaisia klusterointimenetelmiä on kehitetty paljon. Toiset suoriutuvat joistain tehtävistä paremmin kuin toiset, mutta mikään olemassa olevista menetelmistä ei kykene suoriutumaan hyvin jokaisesta klusterointiongelmasta (Jain 2010). Siksi on tärkeää vertailla eri menetelmiä valittaessa käytettävää klusterointimenetelmää. Eri klusterointimenetelmät menetelmät voidaan yleisesti luokitella kuuteen ryhmään: osiointi (partitioning), hierarkkinen, ruudukkopohjainen (grid-based), mallipohjainen, tiheyspohjainen klusterointi ja monivaiheiset klusterointialgoritmit.

Tutkimuksissa, missä on klusteroitu kaukolämpöasiakkaita menetelmät ovat vaihdelleet. Esimerkiksi Ma et al. (2017) käytti tutkimuksessaan PAM-algoritmia (partition around medoids), Kiluk (2017) käytti lähimmän naapurin (Nearest-Neighbour) menetelmää, Gianniou et al. (2018) käytti k-means-algoritmia, Wang et al. (2019) Gaussin sekoitemallia (Gaussian Mixture Model) ja Calikus et al. (2019) käytti tutkimuksessaan k-Shape-algoritmia. Suurin osa töistä on käyttänyt klassista ja laajasti käytettyä k-means-algoritmia (Mbiydzenyuy et al. 2021), mutta voidaan todeta, että tehtävään ei ole vielä vakiintunutta menetelmää.

Tässä tutkielmassa ollaan kiinnostuneita kaukolämpöasiakkaiden keskimääräisistä kulutusprofiileista. Tarkoituksena on muodostaa ryhmiä, joiden kulutusprofiilit ovat keskenään

mahdollisimman samanlaisia. Kohdassa 3.5 luodut ominaisuusvektorit ovat samankokoisia ja ne ovat normalisoituja, ja ovat siksi hyvin vertailukelpoisia keskenään. Klusterointimenetelmän tulisi klusteroida kulutusprofiilit niiden muodon mukaan.

Aikasarjadatan klusterointi perustuu enimmäkseen klassisiin klusterointimenetelmiin, joko korvaamalla oletusetäisyysmitan aikasarjoille sopivammalla tai muuttamalla aikasarjat ”tasaiseksi” dataksi, jotta olemassa olevia klusterointialgoritmeja voidaan käyttää suoraan (Liao 2005). Klusterointimenetelmän valinta voi kuitenkin vaikuttaa tarkkuuteen, koska jokainen menetelmä ilmaisee klusterien homogeenisyyttä ja erottelua eri tavalla, ja tehokkuuteen, koska laskentakustannukset vaihtelevat menetelmästä toiseen (Paparrizoa & Gravano 2015). Esimerkiksi spektriklusterointi (Filippone et al. 2008) tai tietyt hierarkkisen klusteroinnin muunnelmat (Kaufman ja Rousseeuw 2009) ovat sopivampia tunnistamaan tiheyteen perustuvia klustereita kuin partitiomenetelmät, kuten k-means (MacQueen 1967) tai k-metoid (Kaufman ja Rousseeuw 2009).

k-Shape ja k-MS (k-MultipleShapes, lyh. k-MS) ovat muotopohjaisia algoritmeja aikasarjadatan klusterointiin, jotka ovat tehokkaita ja toimialueesta riippumattomia. k-Shape ja k-MS perustuvat skaalautuvaan iteratiiviseen tarkennusmenettelyyn, joka on samanlainen kuin k-means-algoritmissa, mutta merkittävällä eroavaisuudella. k-Shape ja k-MS käyttävät sekä erilaista etäisyyden mittaa ja erilaista menetelmää sentroidien laskemiseen, kuin k-means-algoritmi. k-Shape ja k-MS yrittävät säilyttää aikasarjojen muodot vertaillessaan niitä. (Paparrizos & Gravano 2017)

Tässä tutkimuksessa käytetään k-Shape-algoritmia klusterointimenetelmänä. Vaikka k-MS on merkittävästi tarkempi (Paparrizos & Gravano 2017) valitaan k-Shape-algoritmi valmiin toteutuksen ja aikarajoitteen takia. Algoritmin toteutuksena käytetään tslearn-kirjastoa (Tavenard et al. 2020). Klusteroinnissa algoritmi suoritetaan 10 kertaa eri sentroidien alkuarvoilla, ja niistä valitaan paras lopputulos keskimääräisen neliöetäisyyden perusteella. Keskimääräisen neliöetäisyyksien vaihtelun kynnyksenä pidetään arvoa 1×10^{-6}, eli jos peräkkäisten iteraatioiden välinen keskimääräisten neliöetäisyyksien vaihtelu on vähemmän kuin kynnysarvo, algoritmi konvergoituu. Iteraatioiden maksimimäärä on viisi.

3.5 klusteroinnin arviointi

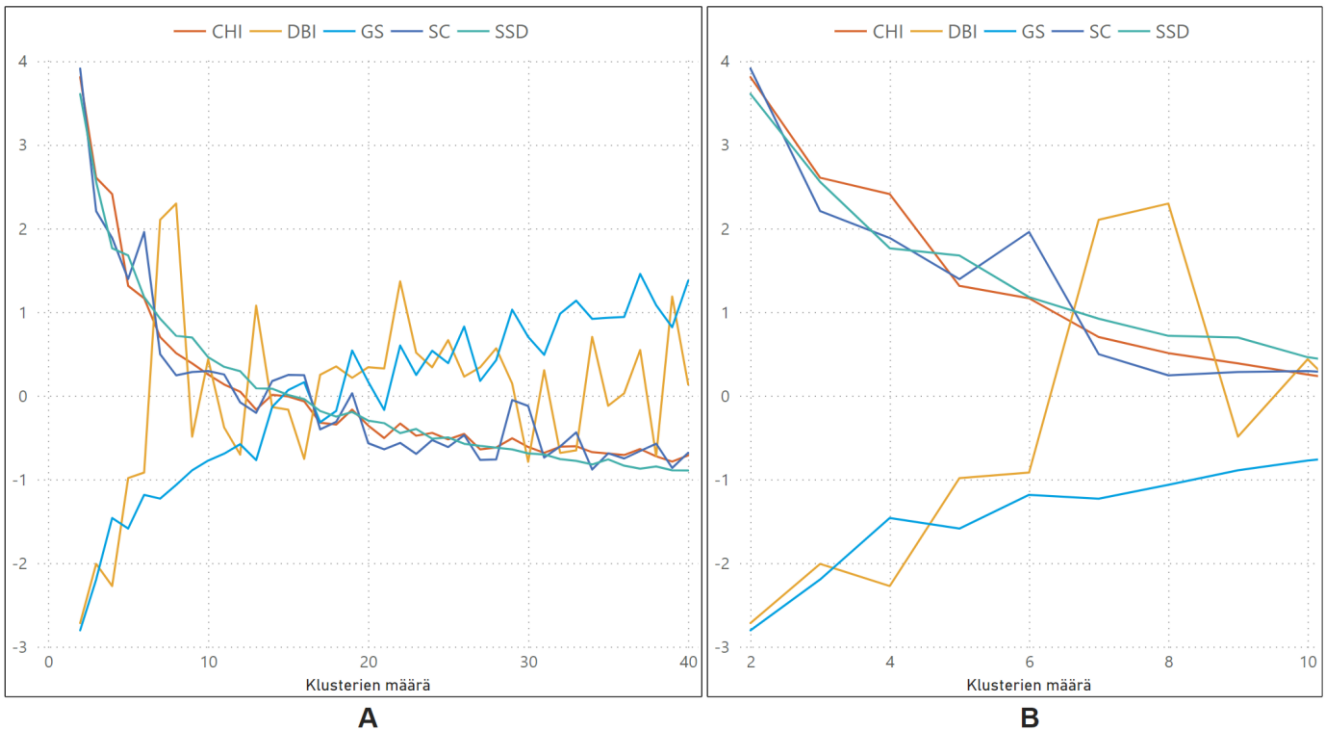
Klusterien lukumäärän valinta riippuu tavoitteesta. Datan segmentoinnissa K määritellään yleensä osaksi ongelmaa. Esimerkiksi yrityksessä voi olla K määrä myyjiä, ja tavoitteena on jakaa asiakastietokanta K-segmentteihin, yksi kullekin myyjälle siten, että jokaiselle myyjälle määrätyt asiakkaat ovat mahdollisimman samankaltaisia keskenään. Usein kuitenkin klusterianalyysiä käytetään kuvaavan tilaston saamiseksi sen selvittämiseksi, missä määrin tietokannan muodostavat havainnot jakautuvat luonnollisiin erillisryhmiin. Tällaisten ryhmien K^* lukumäärää ei tunneta ja se edellyttää, että se, kuin myös itse ryhmittely arvioidaan datasta. (Hastie et al. 2009)

Dataan perustuvat menetelmät K^*:n estimoimiseksi tutkivat tyypillisesti klusterin sisäistä eroa W_K klusterien lukumäärän K funktiona. Erilliset ratkaisut saadaan $K \in \{1, 2, \dots, K_{max}\}$. Vastaavat arvot $\{W_1, W_1, \dots, W_{K_{max}}\}$ yleensä pienenevät K:n kasvaessa. Näin on myös silloin, kun kriteeri arvioidaan riippumattomalla testijoukolla, koska suuri määrä klusterikeskuksia pyrkii täyttämään ominaisuustilan tiheästi ja siten ne ovat lähellä kaikkia datapisteitä. Siten ristiinvalidointitekniikoita, jotka ovat hyödyllisiä mallien valinnassa ohjatussa oppimisessa, ei voida käyttää tässä yhteydessä. (Hastie et al. 2009)

Lähestymistavan taustalla oleva intuitio on, että jos todella on K^* erillisiä havaintojen ryhmittelyjä (erilaisuusmitan määrittelemänä), silloin $K < K^*$:lle algoritmin palauttamat klusterit sisältävät kukin todellisten taustalla olevien ryhmien osajoukon. Toisin sanoen ratkaisu ei kohdista havaintoja samassa luonnollisesti esiintyvässä ryhmässä eri estimoituihin klustereihin. Siinä määrin kuin näin on, ratkaisukriteerin arvolla on taipumus laskea oleellisesti jokaisen peräkkäisen määrättyjen klustereiden määrän lisäyksen yhteydessä, $W_{K+1} \ll W_K$, koska luonnolliset ryhmät osoitetaan peräkkäin erillisiin klustereihin. Jos $K > K^*$ yhden arvioiduista klustereista on jaettava vähintään yksi luonnollisista ryhmistä kahdeksi alaryhmäksi. Tällä on taipumus saada aikaan pienempi kriteerin lasku, kun K kasvaa edelleen. Luonnollisen ryhmän jakaminen, jossa havainnot ovat kaikki melko lähellä toisiaan, vähentää kriteeriä vähemmän, kuin kahden hyvin erotetun ryhmän liitoksen jakaminen niiden oikeiksi osiksi. (Hastie et al. 2009)

Siinä määrin kuin tämä skenaario toteutuu, kriteeriarvojen peräkkäiset erot $W_K - W_{K+1}$ pienenevät jyrkästi, kun $K = K^*$. Eli $\{W_K - W_{K+1} \mid K < K^*\} \gg \{W_K - W_{K+1} \mid K \geq K^*\}$. Arvio $\widehat{K}^*$ K^*:lle saadaan sitten tunnistamalla ”mutka” W_K:n kuvaajassa K:n funktiona. Kuten muutkin klusterointi menettelyjen aspektit, tämä lähestymistapa on jokseenkin heuristinen. (Hastie et al. 2009)

Klusterointia verrataan eri K lukumäärällä, ja jokaisesta lukumäärästä lasketaan viisi eri mitta-arvoa. Valitut mitta-arvot ovat Calinski-Harabasz-indeksi (CHI), Davies-Boulding-indeksi (DBI), Siluetti-indeksi (Silhuette Coefficient, SC), Neliöetäisyyksien summa (Sum of Squared Distances, SSD) ja Gap statistic (GS). Eri mitat voivat ehdottaa eri klusterien lukumäärää, joten lopullinen määrä valitaan vertailemalla eri mittoja keskenään. Mitat saavat keskenään hyvin erisuuruisia arvoja, joten jokaisen mitan arvot normalisoidaan kaavan 5 mukaan, että niitä voi verrata helpommin keskenään. Kuvasta 12 nähdään eri mittojen normalisoidut arvot eri klusterien lukumäärissä, kun datajoukko on klusteroitu k-Shape-algoritmilla.



Kuva 12. Normalisoitujen mittojen arvot eri klusterimäärillä. Kuva A esittää arvot 40 klusterien määrään asti ja kuva B esittää 10 klusterien määrään asti. Oranssi viiva on normalisoitu Calinski-Harabasz-indeksi (CHI), keltainen on normalisoitu Davies-Boulding-indeksi (DBI), Sininen on

normalisoitu Gap Statistic (GS), tummansininen on normalisoitu Siluetti-indeksi (SC) ja vihreä on neliöetäisyyksien summa (SSD).

3.5.1 Neliöetäisyyksien summa

Neliöetäisyyksien summa on jokaisen datapisteen euklidinen neliöetäisyys pisteen lähimmästä klusterin sentroidista. Datapisteen u ja lähimmän sentroidin c välinen etäisyys lasketaan kaavalla:

$$d_i = \sum_{i=0}^{n_1}(u_i - c_i),$$

missä n_1 on aikalohkojen määrä, mihin vuorokauden tunnit on jaettu jokaisen asiakkaan mallissa, ja etäisyyksien d neliöiden summa lasketaan kaavalla:

$$\sum_{i=0}^{n_2} d_i^2,$$

missä, n_2 on asiakkaiden määrä. Neliöetäisyyksien summa on laskettu tslearn-kirjaston KShape.inertia_-funktiolla (Tavenard et al. 2020).

Kirjallisuudessa usein käytetään heuristista menetelmää, tunnistaa neliöetäisyyksien summista eri klusterien lukumäärillä kuvaajasta ”kyynärpää”, jonka taitekohta olisi oikea klusterien lukumäärä. Schubert (2023) kritisoi, että kyynärpää menetelmällä ei ole teoreettista tukea, ja sitä ei tulisi käyttää, koska parempia vaihtoehtoja on tunnettu kirjallisuudessa jo pitkään.

Kuvasta 12 nähdään, että neliöetäisyyksien summa laskee suhteellisen tasaisesti klusterien lukumäärän kasvaessa, eikä kuvaajassa esiinny poikkeavuuksia. Arvot laskevat eniten K:n arvoon neljä asti, jonka jälkeen lasku on tasaista. Voidaan todeta, että neliöetäisyyksien summa ei osoita, mitään valittavaa klusterien lukumäärää, ja tulisi keskittyä enemmän muiden mittojen arvoihin.

3.5.2 Davies-Bouldin-indeksi

Davies-Boulding (DB) -indeksi (Davies & Bouldin 1979) on klusterin sisäisen sironnan summan ja klusterin välisen erotuksen summan funktio. Sironta i:nnen klusterin sisällä S_i lasketaan muodossa $s_i = \frac{1}{|C_i|}\sum_{x \in C_i}\{\|x - z_i\|\}$ ja klusterin C_i ja C_j välinen etäisyys, jota merkitään d_{ij}:llä, määritetään $d_{ij} = \left\|z_i - z_j\right\|$. Tässä z_i esittää i:ttä klusterin keskustaa. DB-indeksi määritellään siten, että:

$$DB = \frac{1}{K}\sum_{i=1}^{K} R_{i,qt}\ ,$$

missä $R_{i,qt} = max_{j,j \neq i}\left\{\frac{S_{i,q}+S_{i,q}}{d_{ij,t}}\right\}$. Tavoitteena on minimoida DB-indeksi luonnollisen klusteroinnin saavuttamiseksi (Maulik & Bandyopadhyay 2002.)

Tässä tutkielmassa DB-indeksi on laskettu scikit-learn-kirjaston sklearn.metrics.davies_boulding_score-funktiolla (Pedregosa et al. 2011). Kuvasta 12 nähdään, että DB-indeksi saavuttaa pienimmän arvonsa, kun klusterien lukumäärä on kaksi. Arvot eivät pienene K:n arvon viisi jälkeen. DB-indeksin ehdottamaksi K:n arvoksi voidaan pitää kahta.

3.5.3 Calinski-Harabasz-indeksi

Calinski-Harabasz (CH) -indeksi (Caliński & Harabasz 1974), joka tunnetaan myös nimellä varianssisuhteen kriteerinä, missä n datapistettä ja K klusteria lasketaan

$$CH = \frac{[SSB\ /\ (K-1)]}{[SSW\ /\ (n-k)]}.$$

Tässä SSB on neliöiden summa klustereiden välillä ja SSW on neliöiden summa klustereiden sisällä. Hierarkian enimmäistasoa käytetään osoittamaan oikea osioiden lukumäärä datassa. SSB voidaan kirjoittaa muodossa:

$$SSB = \sum_{k=1}^{K} n_k \|z_k - z\|^2,$$

missä n_k on pisteiden lukumäärä klusterissa k ja z on sentroidi koko datajoukossa. SSW voidaan kirjoittaa muodossa:

$$SSW = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_i - z_k\|^2.$$

Siten CH, indeksi voidaan kirjoittaa muodossa:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k \|z_k - z\|^2}{K-1}\right] \Big/ \left[\frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_i - z_k\|^2}{n-k}\right].$$

Indeksi L määritetään seuraavasti:

$$L(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^p,$$

missä K on klustereiden lukumäärä. Tässä,

$$E_K = \sum_{k=1}^{K} \sum_{j=1}^{n} u_{kj} \|x_j - z_k\|,$$

ja

$$D_K = max_{i,j=1}^{K} \|x_j - z_k\|.$$

n on pisteiden kokonaismäärä datajoukossa, $U(x) = \left[u_{kj}\right]_{K \times n}$ on ositusmatriisi (partition matrix), ja z_k on k:nnen klusterin keskipiste. K:n arvoa, jolle K on maksimoitu, pidetään klustereiden oikeana määränä. (Maulik & Bandyopadhyay 2002)

Tässä tutkielmassa CH-indeksi on laskettu scikit-learn-kirjaston sklearn.metrics.calinski_harabasz_score-funktiolla (Pedregosa et al. 2011). Kuvasta 12 nähdään, että CH-indeksin arvot vähenevät klusterien lukumäärän lisääntyessä. Indeksin arvot eivät nouse merkittävästi missään kohtaan, ja suurin indeksin arvojen laskut tapahtuvat, kun klusterien määrä nousee kolmeen ja viiteen. Suurin CH-indeksin arvo on klustereiden lukumäärän ollessa kaksi, joten CH-indeksin suosittelemaksi klusterien lukumääräksi pidetään kahta. CH-indeksin arvot eivät osoita selkeästi mitään luonnollista K:n arvo, ja antaa yhtä vähän arvoa klusterien lukumäärän valinnassa, kuin neliöetäisyyksien summa.

3.5.4 Siluetti-indeksi

Siluetti-indeksi (Rousseeuw 1987) on mitta siitä, kuinka samanlainen objekti on omaan klusteriinsa (koheesio) verrattuna muihin klustereihin (erottelu). Siluetti vaihtelee -1 ja 1 välillä, jossa korkea arvo osoittaa, että kohde sopii hyvin omaan klusteriinsa ja huonosti viereiseen klusteriin. Siten positiiviset ja negatiiviset suuret siluettileveydet (SL) osoittavat, että vastaava objekti on ryhmitelty hyvin, ja vastaavasti väärin. Objektien, joiden SL-kelpoisuusindeksi on lähellä nollaa, ei pidetä selkeästi erotettavissa klusterien välillä. (Sinaga & Yang 2020)

Tässä tutkielmassa siluetti-indeksi on laskettu scikit-learn-kirjaston sklearn.metrics.silhouette_score-funktiolla (Pedregosa et al. 2011). Kuvasta 12 nähdään, että siluetti-indeksi saa suurimman arvonsa klustereiden lukumäärän ollessa kaksi, jossa siluetti-indeksin arvo on 0,13. Kuvasta 12 nähdään myös, että siluetti-indeksin arvot laskevat klusterien lukumäärän lisääntyessä. Klusterien lukumäärän ollessa kuusi, siluetti-indeksi nousee kahteen edelliseen arvoon verrattuna, ja tämän jälkeen jatkavat laskua pienellä vaihtelulla. Koska siluetti-indeksin suurin arvo on lähellä nollaa, voidaan indeksin perusteella todeta, ettei luonnollisia ryhmiä ole (Sinaga & Yang 2020).

3.5.5 Gap statistic

Gap statistic (GS) on Tibshirani et al. (2001) kehittämä menetelmä arvioimaan klusterien lukumäärää tietojoukossa. GS vertaa käyrää $\log W_K$ generoituun vertailutietojoukkoon, jolla on samanlaiset ominaisuudet kuin oikealla datalla. Se arvioi optimaalisen klusterimäärän olevan paikka, jossa kahden käyrän välinen ero on suurin. Se on automaattinen tapa paikantaa edellä mainittu ”mutka”. Se toimii myös kohtuullisen hyvin, kun tiedot jakautuvat yhteen klusteriin, ja siinä tapauksessa on taipumus arvioida optimaalisen klustereiden lukumääräksi yksi. Tämä skenaario, jossa useimmat muut kilpailevat menetelmät epäonnistuvat. (Hastie et al. 2009)

GS-arvojen laskennassa käytetään gapstat-kirjaston (Maloney 2020) gapstat_score-funktiota. Funktion parametrit pidetään oletuksena, paitsi parametrille ”clusterer” annetaan käytetty k-Shape-algoritmi. Kuvasta 12 nähdään, että GS-arvot nousevat klusterien määrän noustessa. Arvot kasvavat eniten K:n arvoon neljä asti. Kun K on seitsemän ja 12 välissä, arvot nousevat hyvin lineaarisesti, jonka jälkeen GS-arvot nousevat suuremmalla vaihtelulla. Suurin GS-arvo on lasketuista klusterien määristä K:n ollessa 37. Arvoissa ei ole merkittävää huippua missään kohtaa lasketuista klustereista. Hastie et al. (2009) mukaan Gap-estimaatti K^*on pienin K, joka tuottaa GS-arvon yhden GS-arvojen keskihajonnan sisällä kohdassa $K + 1$. Eli pienin K:n arvo, jonka jälkeen GS-arvot nousevat vähemmän pienempiin K:n arvoihin verrattuna. Silloin klusterien lukumääräksi tulisi valita neljä.

3.5.6 Klusterien lukumäärän valitseminen

Klusterien lukumäärän K valitsemisessa käytetään apuna viittä eri mittaria, jotka ovat SSD, DBI, CHI, SC ja GS. Mittareiden suosittelemat klusterien lukumäärät eroavat toisistaan. SSD-arvojen perusteella ei voida antaa tukea, millekään K:n arvolle. DBI suosittelee K:n arvoksi kaksi, mutta myös neljä on mahdollinen. CHI ei anna selkeää tukea millekään K:n arvolle, mutta voidaan todeta, että se suosittelee K:n arvoksi pienempää kuin kuusi. SC ei anna selkeää tukea millekään tietylle K:n arvolle. Suurin SC-arvo on K:n ollessa kaksi, kuitenkin arvot nousevat hetkellisesti K:n ollessa kuusi, josta arvot laskevat selvästi. Joten voisi sanoa, että SC-arvojen mukaan K olisi alle seitsemän. GS suosittelee K:n arvoksi neljää.

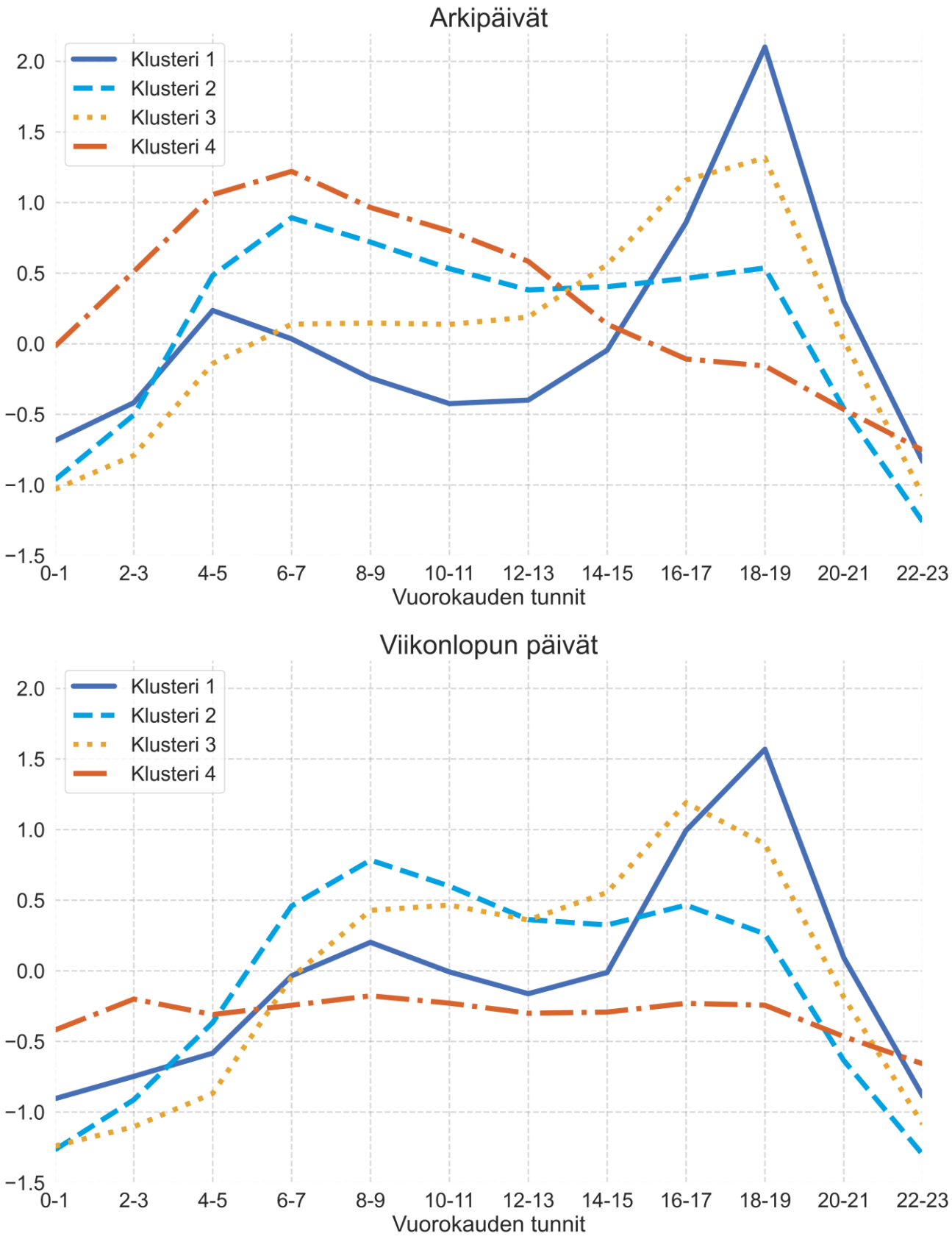
Koska klusteroinnilla muodostetaan asiakkaiden keskimääräisiä käyttöprofiileja, olisi hyvä, että muodostettavien ryhmien lukumäärä olisi hallittavan kokoinen, jotta sitä olisi helppo hyödyntää esimerkiksi erilaisiin markkinointi- ja myyntitarkoituksiin. Tarkoituksena on muodostaa yleistettäviä malleja asiakkaiden vuorokauden aikaisista käytöistä. Eri mittojen suosittelemien K:n arvojen, sekä klusteroinnin tarkoituksen huomioimisen perusteella valitaan klustereiden lukumääräksi neljä. Lopullinen klusterointi suoritetaan 1000 kertaa eri sentroidien alkuarvoilla, ja maksimi iteraatiomäärä on 2000, että klusteroinnin lopputulos olisi sovitettu mahdollisimman hyvin.

4 Tulokset ja pohdintaa

Kaikista 6084 kaukolämpöasiakkaasta muodostettiin 6079 asiakkaalle niiden keskimääräistä lämmönkäyttöä kuvaavat kulutusprofiilit vuorokauden aikana arkipäiville ja viikonlopun päiville. Nämä kulutusprofiilit edustavat keskimääräistä lämmönkäyttöä vuorokauden aikana ilman ulkolämpötilan vaikutusta. Kulutusprofiileista muodostettiin neljä klusteria edustamaan niiden jäsenten keskimääräisiä kulutusprofiileja. Kuva 13 esittää muodostettujen klustereiden keskimääräiset arvot eli sentroidit.

Kuvasta 13 nähdään, että muodostetuissa klustereissa on selkeitä eroja keskenään, mutta niissä on kuitenkin havaittavissa samankaltaisuuksia. Kaikilla lämmönkäyttö nousee aamulla ja laskee illalla. Kaikilla klustereiden sentroideilla käyrän muoto vaihtelee hieman arkipäivien ja viikonlopun päivien välillä. Klustereiden 1 ja 3 jäsenten keskimääräisissä lämmönkäytöissä on havaittavissa selkeä huippu vuorokauden aikana, kun taas klustereiden 2 ja 4 jäsenten käyttö on keskimäärin tasaisempaa. Kuvasta 14 nähdään, että muodostetut klustereiden sentroidit edustavat hyvin muodostettuja keskimääräisiä kulutusprofiileja. Kuvasta 14 nähdään, että klusterin 1 jäsenien huippukäyttö jakaantuu kolmeen eri ajanhetkeen. Tämä osoittaa selkeästi yksilöiden vaihtelua, mutta klusterin 1 jakaminen kolmeksi eri klusteriksi ei toisi lisäarvoa.

Klusterin 1 jäsenten keskimääräinen lämmönkäyttö kohdistuu arkipäivinä aamulla tuntien 4 ja 5 aikana, sitten käyttö vähenee, ja on päivällä pienimmillään tuntien 10 ja 14 välissä, jonka jälkeen lämmönkäyttö nousee, ja tuntien 18 ja 19 aikana lämpöä käytetään piikkimäisesti huomattavasti enemmän kuin aamulla. Tämä sama kaavio toistuu klusterilla 1 viikonloppuina samanlaisena, mutta käyttö alkaa aamulla myöhemmin tuntien 8 ja 9 aikana, ja on pienintä päivällä tuntien 12 ja 13 aikana, minkä jälkeen lämmönkäyttö nousee, ja on suurinta illalla tuntien 18 ja 19 aikana. Käyttö on hieman tasaisempaa ja korkeampaa arkipäivien käyttöön verrattuna, vaikka illan käyttö on selkeästi suurempaa aamuun verrattuna, kuten arkipäivien käytöissä.



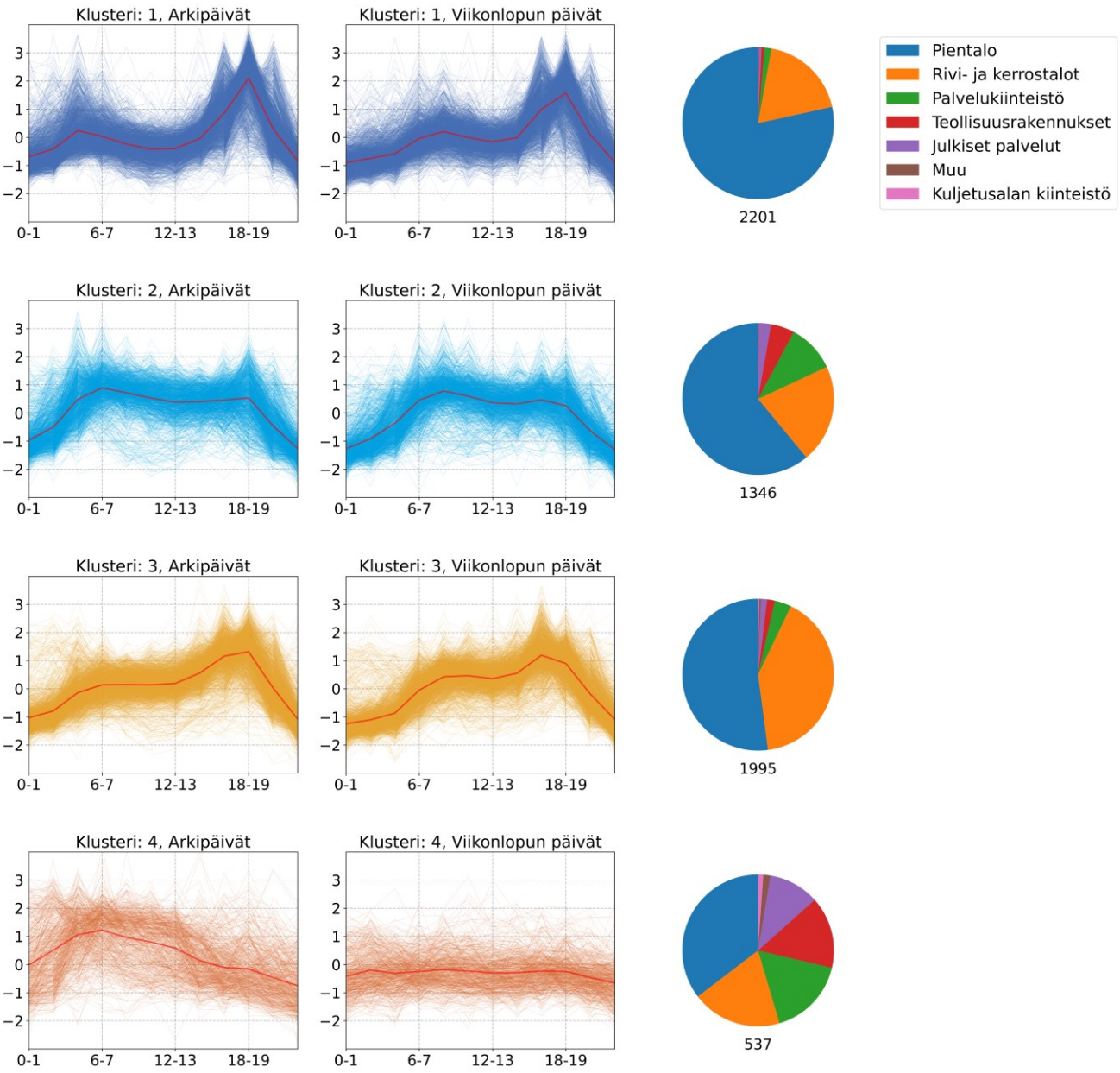
Kuva 13. Klustereiden sentroidit, kun $K = 4$.

Klusterin 2 jäsenten keskimääräinen lämmönkäyttö on suhteellisen tasaista päivän aikana. Arkipäivinä lämmönkäyttö saavuttaa huippunsa tuntien 6 ja 7 aikana, ja laskee hieman tunteihin 12 ja 13, ja lopulta lämmönkäyttö loppuu tunnin 19 jälkeen. Viikonloppuna klusterin 2 jäsenten keskimääräinen lämmönkäyttö on hyvin samanlaista, mutta kaksi tuntia myöhemmin arkipäiviin verrattuna.

Klusterin 3 jäsenten keskimääräinen lämmönkäyttö nousee arkipäivinä tuntiin 6 asti, minkä jälkeen pysyy noin vakiona tuntiin 13 asti, minkä jälkeen kulutus nousee, ja saavuttaa huippunsa tuntien 18 ja 19 aikana, josta kulutus laskee tasaisesti minimiin, mikä on tuntina 23. Viikonloppuna klusterin 3 jäsenten keskimääräinen lämmönkäyttö noudattaa hyvin samalaista kaavaa, kuin arkipäivinä. Kuitenkin viikonloppuna keskimääräinen lämmönkäyttö nousee kaksi tuntia myöhemmin kuin arkipäivinä, mutta loppuu samaan aikaan. Viikonloppuna lämmönkäyttö on myös suurempaa päivän aikana, ja huippuhetkellä se on melkein yhtä suurta.

Klusterin 4 jäsenillä on vain yksi huippu arkipäivien keskimääräisissä käytöissä. Klusterin 4 jäsenten vuorokaudenkäyttö alkaa kaikista keskimääräistä lämmönkäytöistä suurimpana, ja kasvaa arkipäivinä keskimäärin tasaisesti tunteihin 6 ja 7 asti, minkä jälkeen käyttö laskee tasaisesti vuorokauden loppuun asti. Klusterin 4 jäsenten käyttö on viikonloppuisin suhteellisen tasaista, eikä silloin ole selkeää huippukohtaa. Klusterin 4 jäsenten viikonlopun keskimääräinen lämmönkäyttö nousee hieman tuntiin 2 asti, ja pysyy suhteellisen tasaisena tuntiin 19 asti, minkä jälkeen käyttö laskee.

Kuvasta 14 nähdään kaukolämpöasiakkaille muodostetut kulutusprofiilit ja asiakkaiden rakennustyyppien jakauma klusterissa, sekä asiakkaiden määrä klusterissa. Jokaisessa klusterissa pientalot muodostavat merkittävän enemmistön jokaisessa klusterissa, joista merkittävästi eniten klusterissa 1, ja toiseksi eniten klusterissa 3. Voidaan todeta, että vaikka melkein kaikki pientaloista on asumiskäytössä, niiden joukosta löytyy erilaisia lämmönkäyttöprofiileja. Sama asia voidaan todeta rivi- ja kerrostalokiinteistöistä. Vaikka rivi- ja kerrostalokiinteistöissä asuu samassa rakennuksessa monta eri taloutta, ne muodostavat yhdessä samankaltaista käyttöä, kuin pientaloissa, joissa asuu yleensä vain yksi talous. Merkittävästi suurin osa rivi- ja kerrostalokiinteistöistä kuuluu klusteriin 3.

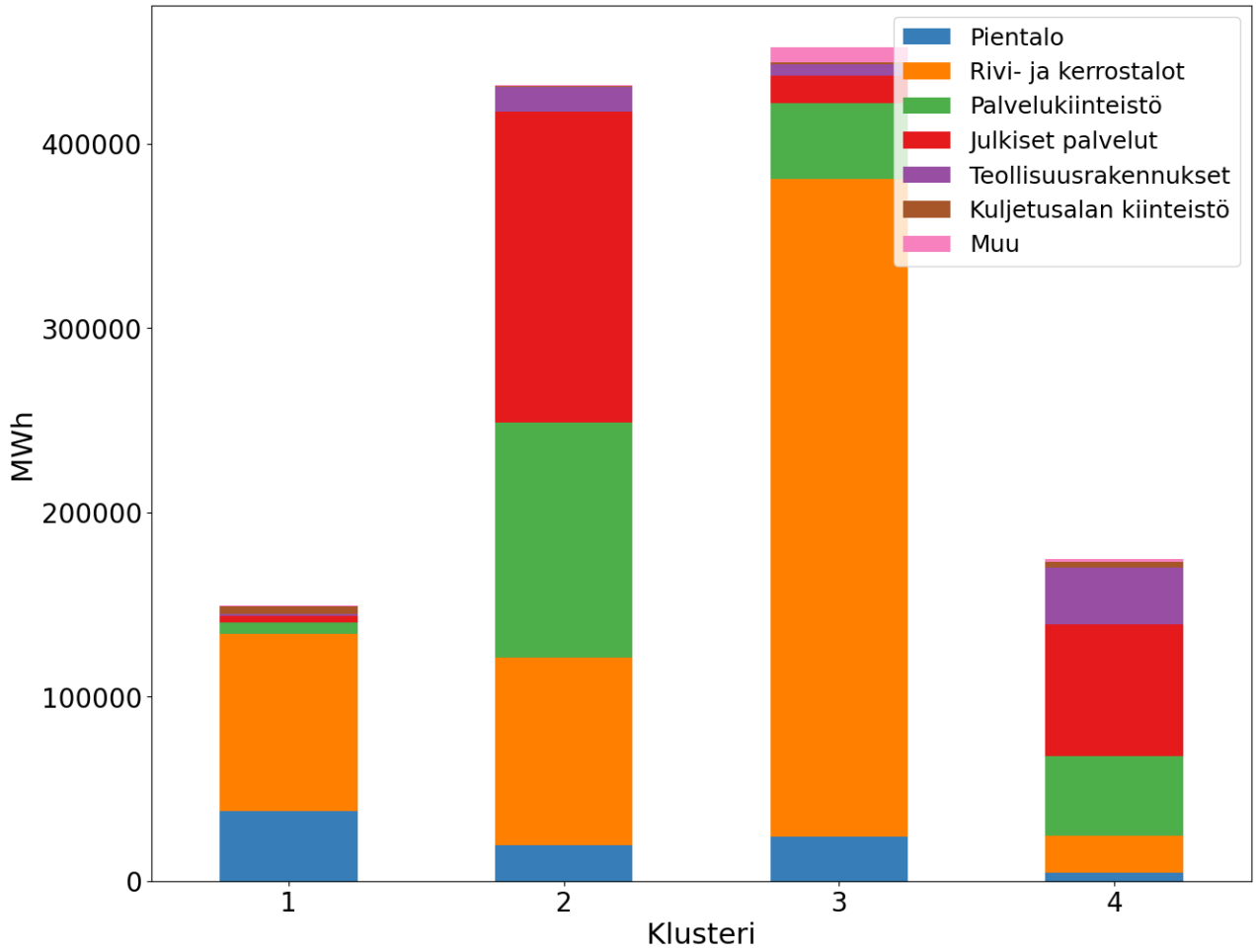


Kuva 14. Yksittäisten asiakkaiden kulutusprofiilit ja rakennustyyppien jakauma klusterissa, kun $K = 4$. Kulutusprofiilit esitetään ohuina viivoina kuvaajassa, ja klusterin sentroidi on paksumpi punainen viiva. Rakennustyyppien jakauma klusterissa esitetään ympyräkaaviona. Ympyräkaavion alla oleva numero kertoo klusteriin kuuluvien asiakkaiden määrän.

Klusterit 1 ja 3 ovat keskenään kaikista samankaltaisimpia, missä klusterissa 1 muutokset ovat korostetumpia kuin klusterissa 3. Näissä kiinteistöissä voi olettaa suurimman osan asukkaista käyvän töissä kodin ulkopuolella, koska käyttöpiikit mukailevat perinteisen työajan muokkaamaa kotona oloa. Klusterissa 3 muutokset ovat pienempiä, ja siihen kuuluu enemmän kerros- ja

rivitaloja, mikä selittäisi pienempää vaihtelua, kun rakennuksen kaikki asukkaat eivät välttämättä noudata perinteistä työrytmiä.

Klusteriin 4 muodostuu suurimmaksi osaksi rakennuksista, mitkä eivät ole todennäköisesti asumiskäytössä. Näissä kiinteistöissä suurin keskimääräinen lämmönkulutus sijoittuu arkipäivien aamuun, ja viikonloppuisin käyttö on tasaista. Voidaan olettaa, että näitä kiinteistöjä käytetään suurimmaksi osaksi arkipäivisin, ja silloin käyttö kohdistuu aamuun. Myös moni palvelu- ja teollisuuskiinteistö kuuluu klusteriin 2, missä kulutusprofiilit noudattavat hyvin selkeästi työaikaa. Klusterin 2 kulutus määräytyy todennäköisimmin ilmanvaihdon takia, koske käyttö on hyvin tasaista läpi normaalin aukioloajan.



Kuva15. Asiakkaiden energiankäyttö klustereittain ja rakennustyypeittäin, kun $K = 4$. Pylväät edustavat klusterissa olevien asiakkaiden kokonaisenergiankäyttöä vuoden aikana. Pylväiden värit kertovat energiankäytön osuuden rakennustyypeittäin klusterin sisällä.

Suurin osa kaikista asiakkaista kuuluu klusteriin 1 ja 3, silti asiakkaiden kokonaiskulutus klusterissa 1 on kaikista klustereista pienintä (Kuva 15). Kaukolämpöyhtiön näkökulmasta klusterit 2 ja 3 ovat kaikista kiinnostavimpia, koska niiden kulutus vaikuttaa eniten koko kaukolämpöverkon lämmöntarpeeseen. Vaikka julkiset- ja palvelukiinteistöt muodostavat määrällisesti pienen osan klusteriin 2 kuuluvista rakennuksista, niiden vaikutus kokonaisenergiankulutuksessa on merkittävä. Toiseksi merkittävin rakennustyyppi on rivi- ja kerrostalot, mitkä muodostavat suurimman osan lämmönkäytöstä klustereissa 1 ja 3. Rivi- ja kerrostalojen lämmönkäyttö on myös melkein yhtä suurta klusterissa 1 ja 2.

Pientalojen osuus kokonaisenergiankäytössä ei ole merkittävä missään klusterissa. Klusterissa 1 pientalot muodostavat yli neljäsosan kaikista rakennustyypeistä, ja silti niiden lämmönkäyttö on alle kolmasosan kokonaislämmönkäytöstä klusterissa. Lisäksi, vaikka klusterissa 3 on neljä kertaa vähemmän rakennuksia kuin klusterissa 1, sen jäsenet käyttävät silti enemmän lämpöä.

Nämä klusterit tuovat hyödyllistä tietoa kuopiolaisten kaukolämpöasiakkaiden yleisimmistä kulutusprofiileista. Sekä siitä, että minkälaiset kulutusprofiilit ovat suurimpia lämmönkäyttäjiä, sekä miten asiakkaiden rakennustyypit jakaantuvat eri klustereihin. Jatkossa kaukolämpöasiakkaille voisi muodostaa kulutusennustemallit, ja hyödyntää sitä älykkääseen lämmönsäätöön, mikä vähentäisi lämmönkulutusta ilman, että asiakas huomaa vaikutusta. Lisäksi keskimääräisiä kulutusmalleja voitaisiin hyödyntää poikkeamien tunnistukseen nopealla aikavälillä. Jos asiakkaan kulutus on merkittävästi suurempaa, kuin keskimääräinen käyttö, on syytä epäillä vuotoa tai jotain muuta häiriötä. Tällaisella varoitusjärjestelmällä voitaisiin pienentää mahdollisten vesivahinkojen laajuutta.

Tässä tutkielmassa ei otettu huomioon rakennuksen rakenteisiin varastoitunutta energiaa. Rakennuksen lämmöntarpeen ennustemalleissa tulisi ottaa huomioon terminen inertia, eli kuinka paljon rakennuksen rakenteet vapauttavat sisätilaan lämpöä. Rakenteet lämpenevät auringon säteilyn tai aikaisemman lämmityksen vaikutuksena, ja tämä lämpö vaikuttaa rakennuksen sisätilojen lämmöntarpeeseen.

5 Yhteenveto

Tässä tutkielmassa muodostettiin 6079 kuopiolaisille kaukolämpöasiakkaille kulutusprofiilit edustamaan asiakkaan keskimääräistä lämmönkäyttöä arkipäivien ja viikonlopun päivien aikana ilman ulkolämpötilan vaikutusta. Kulutusprofiilit muodostettiin lineaariseen regressioon perustuvalla menetelmällä. Muodostetut kulutusprofiilit klusteroitiin neljäksi klusteriksi k-Shape-algoritmilla.

Muodostetuissa klustereissa keskimääräiset kulutusprofiilit erosivat toisistaan. Klustereista oli tunnistettavissa keskimääräiset ajat, milloin rakennuksessa kulutettiin ulkolämpötilasta riippumatonta energiaa. Klustereista tunnisti eroja lämmönkulutuksissa arkipäivien ja viikonlopun päivien välillä. Lämmönkulutuksen määrässä mitattuna tunnistettiin merkittävimmät klusterit, joiden jäsenet muodostavat merkittävimmän vaikutuksen kaukolämpöverkon kokonaislämmöntarpeeseen.

6 Lähteet

Aggarwal, C. C., & Reddy, C. K. (2013). Data Clustering: Algorithms and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) (1st ed.). Chapman and Hall/CRC.

Averfalk, H., & Werner, S. (2020). Economic benefits of fourth generation district heating. Energy, 193, 116727.

Calikus, E., Nowaczyk, S., Sant'Anna, A., Gadd, H., & Werner, S. (2019). A data-driven approach for discovering heat load patterns in district heating. Applied Energy, 252. https://doi.org/10.1016/j.apenergy.2019.113409

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. Journal of statistical software, 61, 1-36.

Chelmis C, Kolte J, Prasanna VK. Big data analytics for demand response: clustering over space and time. In: IEEE international conference on big data, Santa Clara, USA, 29 October-1 November 2015.

Cho, H., Luck, R., Eksioglu, S. D., & Chamra, L. M. (2009). Cost-optimized real-time operation of CHP systems. Energy and Buildings, 41(4), 445-451.

Darby, S. (2010). Smart metering: What potential for householder engagement? Building Research and Information, 38(5), 442–457. https://doi.org/10.1080/09613218.2010.492660

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

Duan, P., Xie, K., Guo, T., & Huang, X. (2011). Short-term load forecasting for electric power systems using the PSO-SVR and FCM clustering techniques. Energies, 4(1), 173-184.

Ediel forum. (2010), Message handbook for Ediel. https://ediel.org/wp-content/uploads/2019/02/MSCONS-24E-20100215.pdf [viitattu 9.2.2023]

Elmegaard, B., Ommen, T. S., Markussen, M., & Iversen, J. (2016). Integration of space heating and hot water supply in low temperature district heating. Energy and Buildings, 124, 255-264.

Energiateollisuus ry. (2021), Kaukolämpövuosi 2021. https://www.slideshare.net/energiateollisuus/kaukolampovuosi-2021 [viitattu 20.11.2023]

Energiateollisuus ry. (2022), Energiavuosi 2021 – Kaukolämpö. https://energia.fi/uutishuone/materiaalipankki/energiavuosi_2021_-_kaukolampo.html [viitattu 16.1.2023]

Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. Pattern recognition, 41(1), 176-190.

Fränti, P., & Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats?. Pattern Recognition, 93, 95-112.

Frederiksen, S., & Werner, S. (2013). District Heating and Cooling. Studentlitteratur AB, Lund.
Gadd, H., & Werner, S. (2013). Heat load patterns in district heating substations. Applied energy, 108, 176-183.

Gadd, H., & Werner, S. (2014). Achieving low return temperatures from district heating substations. Applied energy, 136, 59-67.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: with applications in R. Spinger.

Gianniou, P., Liu, X., Heller, A., Nielsen, P. S., & Rode, C. (2018). Clustering-based analysis for residential district heating data. Energy conversion and management, 165, 840-850.

Goia, A., May, C., & Fusai, G. (2010). Functional clustering and linear regression for peak load forecasting. International Journal of Forecasting, 26(4), 700-711.

Gupta, M., & Jain, M. R. (2014). A performance evaluation of SMCA using similarity association & proximity coefficient relation for hierarchical clustering. Int. J. Eng. Trend. Technol.(IJETT), 15, 354.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of intelligent information systems, 17, 107-145.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) (2nd ed.). Springer.

Ilmatieteenlaitos. (2023), Ilmatieteen laitoksen avoin data ja lähdekoodi. https://www.ilmatieteenlaitos.fi/avoin-data [viitattu 2.3.2023]

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Känkänen, J., Patronen, J., Vilén, K., Saarela, J. (2017). Päästökauppadirektiivin uudistamisen vaikutukset Suomen energiasektoriin ja teollisuuteen. Valtioneuvoston selvitys ja tutkimustoiminnan julkaisusarja 56/2017.

Kiluk, S. (2017). Diagnostic information system dynamics in the evaluation of machine learning algorithms for the supervision of energy efficiency of district heating-supplied buildings. Energy Conversion and Management, 150, 904-913.

Kipping, A., & Trømborg, E. (2016). Modeling and disaggregating hourly electricity consumption in Norwegian dwellings based on smart meter data. Energy and Buildings, 118, 350–369.

Koskelainen, L., Saarela, R., Sipilä, K. (2006): Kaukolämmön käsikirja. Energiateollisuus ry, Helsinki.

Levihn, F. (2017). CHP and heat pumps to balance renewable power production: Lessons from the district heating network in Stockholm. Energy, 137, 670-678.

Li, H., & Wang, S. J. (2014). Challenges in smart low-temperature district heating development. Energy Procedia, 61, 1472-1475.

Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

Lu, Y., Tian, Z., Peng, P., Niu, J., Li, W., & Zhang, H. (2019). GMM clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system. Energy and Buildings, 190, 49-60.

Lund, H., Werner, S., Wiltshire, R., Svendsen, S., Thorsen, J. E., Hvelplund, F., & Mathiesen, B. V. (2014). 4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems. Energy, 68, 1-11.

Ma, Z., Li, H., Sun, Q., Wang, C., Yan, A., & Starfelt, F. (2014). Statistical analysis of energy consumption patterns on the heat demand of buildings in district heating systems. Energy and Buildings, 85, 464-472.

Ma, Z., Yan, R., & Nord, N. (2017). A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. Energy, 134, 90-102.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Maloney, J. (2020). gapstat. GitHub. https://github.com/jmmaloney3/gapstat

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on pattern analysis and machine intelligence, 24(12), 1650–1654.

Mbiydzenyuy, G., Nowaczyk, S., Knutsson, H., Vanhoudt, D., Brage, J., & Calikus, E. (2021). Opportunities for machine learning in district heating. Applied Scienc-es (Switzerland), 11(13). https://doi.org/10.3390/app11136112

Merriam-Webster. (2023). Cluster analysis. In Merriam-Webster.com dictionary. https://www.merriam-webster.com/dictionary/cluster analysis. [viitattu 3.1.2023]

Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. Applied psychological measurement, 11(4), 329-354.

Motiva Oy (2022) Kaukolämpö., https://www.motiva.fi/koti_ja_asuminen/rakentaminen/lammitysjarjestelman_valinta/lammitysmuodot/kaukolampo (13.11.2022).

Nilsson, S. F., Reidhav, C., Lygnerud, K., & Werner, S. (2008). Sparse district-heating in Sweden. Applied Energy, 85(7), 555-564.

Noussan, M., Jarre, M., & Poggio, A. (2017). Real operation data analysis on district heating load patterns. Energy, 129, 70-78.

Paparrizos, J., & Gravano, L. (2017). Fast and accurate time-series clustering. ACM Transactions on Database Systems (TODS), 42(2), 1-49.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Persson, U., & Werner, S. (2011). Heat distribution and the future competitiveness of district heating. Applied Energy, 88(3), 568–576.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Ramos, S., Duarte, J. M. M., Soares, J., Vale, Z., & Duarte, F. J. (2012, July). Typical load profiles in the smart grid context—A clustering methods comparison. In 2012 IEEE Power and Energy Society General Meeting (pp. 1-8). IEEE.

Razmara, M., Bharati, G. R., Hanover, D., Shahbakhti, M., Paudyal, S., & Robinett III, R. D. (2017). Building-to-grid predictive power flow control for demand response and demand flexibility programs. Applied Energy, 203, 128-141.

Reidhav, C., & Werner, S. (2008). Profitability of sparse district heating. Applied Energy, 85(9), 867-877.

Rokach, L., & Maimon, O. (2005). Clustering methods. In Data mining and knowledge discovery handbook (pp. 321-352). Springer, Boston, MA.

Rosen, M. A. (2021). Nuclear energy: non-electric applications. European Journal of Sustainable Development Research, 5(1), em0147.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Salo, S. (2021). Demand Response in District-Heated Buildings.

Sameti, M., & Haghighat, F. (2017). Optimization approaches in district heating and cooling thermal network. Energy and Buildings, 140, 121-130.

Sarvaranta, A., Jääskeläinen, J., Puolakka, J. & Kouri, P. (2012), Kaukolämmön hinnoittelun nykytila ja tulevaisuuden mahdollisuudet, Technical report, ÅF Consulting Oy, Espoo, Finland. Saatavilla: http://docplayer.fi/1155236-Kaukolammon-hinnoittelun-nykytila-ja-\tulevaisuuden-mahdollisuudet.html

Schubert, E. (2023). Stop using the elbow criterion for k-means and how to choose the number of clusters instead. ACM SIGKDD Explorations Newsletter, 25(1), 36-42.

Schweiger, G., Rantzer, J., Ericsson, K., & Lauenburg, P. (2017). The potential of power-to-heat in Swedish district heating systems. Energy, 137, 661-669.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.

Skytte, K., & Olsen, O. J. (2016). Regulatory barriers for flexible coupling of the Nordic power and district heating markets. In 2016 13th International Conference on the European Energy Market (EEM) (pp. 1-5). IEEE.

Sun, Q., Li, H., Wallin, F., & Zhang, Q. (2016). Marginal costs for district heating. Energy Procedia, 104, 323-328.

Suomen virallinen tilasto (SVT): Kasvihuonekaasut [verkkojulkaisu]. Helsinki: Tilastokeskus [Viitattu: 14.11.2023]. Saantitapa: https://www.stat.fi/julkaisu/cl8a46vp7vq8n0bvyqi4724gw

Suomen virallinen tilasto: Energian hankinta ja kulutus [verkkojulkaisu]. ISSN = 1799-795X. Taulukko: Kaukolämmön tuotanto Suomessa, 2000–2021. Helsinki: Tilastokeskus [viitattu 26.11.2022]. Saantitapa: https://pxdata.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__salatuo/statfin_salatuo_pxt_12b7.px

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., ... Woods, E. (2020). Tslearn, A Machine Learning Toolkit for Time Series Data. Journal of Machine Learning Research, 21(118), 1–6. Retrieved from http://jmlr.org/papers/v21/20-091.html

Theodoridis, S., & Koutroumbas, K. (2008). Pattern recognition. Academic Press.

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), 63(2), 411-423.

Tulkki, V., Arnold, M., Leppänen, J., Soppela, O., Hyvärinen, J. (2022). Ydinkaukolämpöselvitys. VTT. Saatavilla: https://www.vttresearch.com/fi/project_news/selvitys-ydinenergian-kayton-mahdollisuuksista-kaukolammon-tuotannossa

Tureczek, A. M., Nielsen, P. S., Madsen, H., & Brun, A. (2019). Clustering district heat exchange stations using smart meter consumption data. Energy and buildings, 182, 144–158.

Veyssieres, M. P., & Plant, R. E. (1998). Identification of vegetation state and transition domains in California's hardwood rangelands. University of California, 101.

Wang, C., Du, Y., Li, H., Wallin, F., & Min, G. (2019). New methods for clustering district heating users based on consumption patterns. Applied Energy, 251, 113373.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), 236-244.

Werner, S. (2017). District heating and cooling in Sweden. Energy, 126, 419-429.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

Xiao, F., & Fan, C. (2014). Data mining in building automation system for improving building operational performance. Energy and buildings, 75, 109-118.

Yu, Z. J., Haghighat, F., Fung, B. C., & Zhou, L. (2012). A novel methodology for knowledge discovery through mining associations between building operational data. Energy and Buildings, 47, 430-440.