Samuli Kotiranta
AB8349

Tehtävä 3 – Datan valmistelu

Datan valmistelussa keskeistä on valita saatavilla olevasta datasta sellaiset osat, jotka ovat tutkimuskysymyksen kannalta merkityksellisiä. Tässä vaiheessa siis korostuu alussa tehdyn tutkimuskysymyksen ja liiketoimintasuunnitelman ymmärtäminen sekä tavoitteen määrittelyn selkeys.

Pohdittaessa kiinnostavia datan osia, voisi tässä vaiheessa olla järkevää rajata esimerkiksi useimmiten puuttuvat aikavälit tarkastelun ulkopuolelle. Se helpottanee[1] käytännön työskentelyä myöhemmin. Ajallista rajausta tehdessä on luontevaa ajatella, että tuoreemmalla datalla on luonnollista painoarvoa enemmän kuin vanhemmalla eli mahdollista katkaisua kannattaa soveltaa ennemmin alku kuin loppupäähän.

Datan tarkastelu kiinteistönumeron funktiona auttaa löytämään erityisen runsaasti vettä kuluttavia kiinteistöjä, ja triviaalisti mahdollistaa kiinteistökohtaisen aikasarjaseurannan. Aikasarja kokonaiskulutuksessa alueittain tai kiinteistöittäin paljastaa nopeasti putkiston vuototilanteet. Jälkimmäisen toteuttaminen koneellisesti mahdollistaa vaikkapa vuotavan vessanpöntön paikantamisen. On siis järkevää pitää muuttujina vähintäänkin aika ja kiinteistönumero.

Kiinteistökohtaista vaihtelua vedenkulutuksessa on luonnollisesti henkilöluvun mukaan. Tiedostossa ”water consumer habits” on saatavilla laajalti kuluttajaprofiileja, jotka eivät (luonnollisestikaan) ole yhdistettävissä yksittäiseen kiinteistöön. Sen avulla voi kuitenkin tuottaa tilastollisesti tyypillisiä käyttöprofiileja, joita voidaan myöhemmin käyttää kiinteistökohtaisten tilastolliseen vakiointiin (ja edelleen esimerkiksi kiinteistöjen klusterointiin).

Formaatin osalta tutkittava data on annettu CSV-muodossa. Koska kyse on yksinkertaisesta alfanumeerisesta aineistosta CSV-muoto on tarkoitukseen aivan riittävä ja käytännöllinen.

[1] Itselläni on taustaa tähtitieteellisen datan käsittelystä. Siellä aukkoinen ja heteroskedastinen data on enemmän sääntö kuin poikkeus, sillä havainto-olosuhteet eivät yleensä ole valittavissa. Alalla käytetyt tilastolliset menetelmät ovat yleensä hyvin tolerantteja datan epäsäännölliselle jakautumiselle. Erityisesti aikasarjojen kohdalla ongelma on usein enemmän laskentamenetelmän valinnassa kuin aukoissa sinänsä.