TAMPEREEN
YLIOPISTO

# Apteeraustulosten vertaileminen


Tilastotiede

Pro Gradu

Matematiikan, tilastotieteen ja

filosofian laitos

Tampereen yliopisto

Toukokuu 2006

Jarkko Penttinen

TAMPEREEN YLIOPISTO

Matematiikan, tilastotieteen ja filosofian laitos

PENTTINEN, JARKKO: Apteeraustulosten vertaileminen

Pro Gradu –tutkielma, 87s., 22 liites.

Tilastotiede

Toukokuu 2006

---

Tiivistelmä

Metsästä saatavat tulot ovat metsän omistajalle tärkeä asia. Suurille metsän omistajille, kuten Metsähallitukselle, prosentuaalisesti pienetkin parannukset metsästä saatavaan tuloon voisivat tuottaa vuositasolla jopa satojatuhansia euroja. Nykyisin, kun puiden runkojen katkonta osiin eli apteeraus tapahtuu melkein kokonaan koneellisesti, on herätty etsimään uudenlaisia katkontatuloksia parantavia keinoja. Erityisen kiinnostuneita ollaan katkontatulosten vertailusta eri hakkuutyömaiden välillä. Aikaisemmin ei ole ollut olemassa tilastollisesti pätevää tapaa todeta eroja hakkuutyömaiden välillä ja tähän ongelmaan tutkielma yrittää etsiä vastauksen.

Tutkimusaineisto kerättiin hakkuukoneilta vuoden 2005 ensimmäisellä puoliskolla ja varsinaiset datan analyysit suoritettiin seuraavien muutaman kuukauden aikana. Datan muokkaamiseen ja analysointiin käytettiin pääosin SPSS- ja R-ohjelmistoja. Koska hakkuiden todelliset tulokset ovat riippuvaisia tietyistä alueellisesti ja yksilöllisesti sovituista hinnoista, tutkimuksen analyysit perustuvat tuloksiin, jotka on laskettu puutavaralajeille luodun yhtenäisen hinnoittelun pohjalta katkottujen runkojen tilavuuksien perusteella.

Datan kuvailevan vertailun perusteella aineistoa voidaan jaotella sekä alueellisiin että läpimittaluokittaisiin ryhmiin ja yksinkertaisten tunnuslukujen laskemisen avulla voidaan todeta etelän läpimittaluokittaisten tulosten olevan pohjoisten tuloksia parempia.

Regressioanalyysien perusteella vahvimmaksi selittäjäksi kaikissa läpimittaluokkien ryhmissä nousi rungon halkaisija rinnan korkeudelta (1.3 metriä) mitattuna. Mallien korjatut selitysasteet olivat suhteellisen suuria ja vaihtelivat männyllä välillä 84–96 % ja kuusella välillä 72–96 %.

Analyysitulosten pohjalta voidaan rakentaa toteutuneen katkonnan seuraamisen väline. Tehtävien johtopäätösten perustana suosittelen käytettäväksi sekä regressioanalyysin tuottamia malleja että yksinkertaisia tunnuslukuja (keskiarvot ja kvantiilit).

**Hakusanat:** apteeraus, puiden katkonta, metsän hakkuu, regressioanalyysi

**Sisällysluettelo:**

# 1 Johdanto

Tutkimuksen lähtökohtana oli Metsähallituksen tarve seurata entistä paremmin puuraaka-aineen hyödyntämisen tehokkuutta. Perimmäisenä tarkoituksena oli rakentaa tilastollisesti pätevä johtamisen, seurannan ja palautteen antamisen järjestelmä, jolla voitaisiin parantaa metsistä saatuja tuloja. Tämän tarkoituksen toteuttamista varten kerättiin hakkuukoneiden tietokoneilta tietoa toteutuneesta apteerauksesta eli runkojen katkonnasta vuoden 2005 alkupuoliskolta. Lopullinen aineisto oli koossa kesäkuun alkupuolella ja sen pääasiallisena analyysimenetelmänä käytettiin toimeksiantajan toivomuksen mukaisesti regressioanalyysiä.

## 1.1 Tutkimuksen aihe

Toteutuneen apteerauksen vertailemisessa tavoitteena on yksinkertaisesti kerättävissä olevien muuttujien avulla tuottaa luotettavaa ja helposti vertailtavissa olevaa informaatiota apteerauksen tuloksesta suhteessa tavoitteeseen. Analyysimenetelmänä käytettävän regressioanalyysin avulla etsitään malli tai mallit, joiden avulla tavoite voidaan asettaa. Malliin otettavien muuttujien tulee olla mahdollisimman yksinkertaisesti kerättävissä olevia, koska tuloja maksimoitaessa on yleensä järkevää olla kasvattamatta kuluja, joita lisätoimet uusien muuttujien keräämiseksi aiheuttaisivat. Muodostetaan vertailut jo olemassa olevista tiedoista, joten tehtävä ei vaadi lisätoimia normaalin apteeraustapahtuman lisäksi.

## 1.2 Aiheen merkitys

Tieteellisenä tutkimuksena aiheen merkitys korostuu siksi, ettei vastaavanlaista tutkimusta ole saatavilla. Apteeraustuloksen optimoinnista on toki tehty paljon tutkimuksia (Korvenoja & Nummi 1994; Liski & Nummi 1994; Nummi 1998; Nummi & Möttönen 2001), jotka pääasiassa keskittyvät runkokäyrän ennustamiseen ja tämän ennusteen hyödyntämiseen. Tämän tutkimuksen avulla tarkoituksena on pystyä toteamaan, onko metsäurakoitsijoiden työn tehokkuudessa eroja. Hyvin läheltä tätä tutkimusta sivuaa Tampereen yliopistossa tehty tutkimus (Nummi, Sinha & Koskela 2004), jossa apteerauksen lopputulosta pyritään mittaamaan hinta- ja kysyntämatriisien avulla. Tässä tutkimuksessa näitä matriiseja ei ole mahdollista ottaa tarkasteluun mukaan sellaisenaan, joten niiden sisältämä informaatio on lähinnä typistetty matriisien sarakkeiden ja rivien lukumääriin. Esimerkiksi, kun kysyntämatriisin mukaan asiakas on halukas ostamaan seitsemää eripituista tukkia, saa tukin pituusluokkia kuvaava muuttuja arvon seitsemän. Koska ostajan intresseissä painottuvat myös puutavaralajien läpimitat, on muodostettu myös muuttuja, joka kertoo kyseessä olevan puutavaralajin hyväksytyn minimikatkontaläpimitan. Samoin mukaan on otettu muuttuja, joka kuvaa lyhyimmän katkontavaihtoehdon pituutta.

## 1.3 Tutkimuksen käytännön toteutus

Tutkimus toteutettiin toimeksiantona Tampereen yliopiston tilastotieteen laitokselle, jossa se ohjautui allekirjoittaneelle. Tuomo Vuorenpää teki suuren työn ohjelmoimalla ohjelmat ja suorittamalla niillä tarvittavat poiminnat relevanteista tiedostoista, jotta data saatiin analysoitavaan muotoon. Hän on muutenkin ollut suurena apuna toimiessaan tutkimuksen ohjaajana Metsähallituksen puolelta. Tutkin tilastollisilla ohjelmilla

(SPSS, R ja Excel) toimitettuja datoja, kunnes data saatiin lopulliseen muotoonsa. Varsinaiset analyysit, läpimittaluokittaisia regressiomalleja lukuun ottamatta suoritettiin R-ohjelmalla, joka on ilmainen ja saatavilla internetistä osoitteesta: http://www.r-project.org. Kiitän myös seuraavia henkilöitä rakentavista kommenteista työn suhteen: Tore Högnäs (Metsähallitus), Olli Väihkönen (Metsähallitus), Erkki Liski (Tampereen yliopisto) ja Tapio Nummi (Tampereen yliopisto).

## 1.4 Tutkimuksen rakenne

Aluksi on syytä todeta, että tutkimuksen empiirisen luonteen vuoksi suurin osa analyysin taustalla olevasta teoriasta on sijoitettu liiteosaan. Viittausten avulla asiasta kiinnostunut lukija voi tarvittaessa tarkastella liiteosasta kulloinkin kyseessä olevan asian teoreettista taustaa. Luvussa 1 johdatellaan lukija tutkimuksen aiheeseen sekä siihen, mistä tutkimuksen tarve nousee. Luvussa 2 esitetään tutkimuksen kannalta olennaiset käsitteelliset lähtökohdat ja tutkimuksen keskeinen termistö, jotta tutkimuksen seuraaminen alaa tuntemattomallekin olisi mahdollista. Luvussa 3 esitellään tutkimusaineisto ja sen sisältämät muuttujat. Jotkut muuttujat on muodostettu toisten muuttujien avulla, joten esitetään myös, miten nämä muuttujat muodostetaan. Luku 4 tarkastelee dataa kuvailevan tilastotieteen menetelmin. Ensin on tarkasteltu mäntyä ja seuraavaksi kuusta. Luvun 4 lopuksi suoritetaan alustavat tulosten vertailut laskemalla läpimittaluokittaiset keskiarvot ja kvantiilit. Tarkempien vertailutuloksien saavuttamiseksi aineistot ryhmitellään maantieteellisen sijainnin mukaan ja muodostetaan uudelleen läpimittaluokittaiset keskiarvot sekä kvantiilit. Lopuksi tehdään yhteenveto kuvailevasta osuudesta. Luvussa 5 suoritetaan aineistolle regressioanalyysit. Luku 5.1 sisältää analyysit männylle ja 5.2 kuuselle. Ensin rakennetaan yksinkertaiset regressiomallit ja siitä edetään kohti lopullisia malleja. Luvussa 6 tarkastellaan tulosten soveltamista käytäntöön eli siinä suoritetaan analyysitulosten perusteella esimerkkivertailu sattumanvaraisesti valitulle työmaalle. Vertailu pitää sisällään regressiomallien estimaattien laskemisen ja ennustevälien muodostamisen, joita verrataan toteutuneisiin tuloksiin sekä niistä laskettuihin keskiarvoihin ja kvantiileihin. Vertailut esitetään myös graafisessa muodossa. Luvussa 7 tehdään yhteenveto tuloksista ja tarkastellaan tätä tutkimusta yleensä.

# 2 Käsitteelliset lähtökohdat ja keskeisten termien määrittely

Tämän luvun tarkoituksena on esitellä tutkimuksen seuraamista helpottava termistö sekä valottaa hiukan käsitteellisiä lähtökohtia alaa ennestään tuntemattomalle lukijalle. Esittelyt on pääosin pyritty pitämään mahdollisimman lyhyinä, joten asiasta enemmän kiinnostuneita kehotetaan tutustumaan alan kirjallisuuteen, erityisesti Tapion taskukirjaan, joka käsittelee metsäalaa erittäin laajasti. Koska apteeraustapahtuma on tutkimuksen kannalta oleellinen asia, on sitä luvun lopuksi käsitelty hiukan laajemmin.

Suomessa, kuten muuallakin Skandinaviassa, puun rungon pilkkominen puutavaralajeiksi tapahtuu metsän hakkuun yhteydessä. Hakkuu suoritetaan yleensä koneellisesti katkontaohjeen mukaisesti. Suomessa ja Ruotsissa koneellisen puunkorjuun osuus on noin 70–90 prosenttia (Salakari & Heimonen 1998, s. 13). Koneellisen hakkuun nopeus on omaa luokkaansa verrattuna metsurin suorittamaan hakkuuseen. Koneellisessa hakkuussa harvesteri eli metsätyökone karsii ja katkoo eli apteeraa rungot kuljettajan avustamana eri puutavaralajeiksi. Harvesterin tietokone tallentaa katkontatapahtuman tiedot PRD-tiedostoon. Katkotut rungot kerätään korjuuautolla metsästä ja kuljetetaan ostajalle.

## 2.1 PRD-tiedosto

PRD-tiedosto eli hakkuukoneen tuotantotiedosto sisältää suuren määrän apteerauksesta kertynyttä tietoa. Kaikkea tätä tietoa ei ole järkevää tai edes mahdollista lähteä tutkimaan, vaan tiedostosta poimitaan olennaiset muuttujat tarkempaa tarkastelua varten. Tutkimukseen mukaan otettavien muuttujien valinta on pääosin suoritettu Metsähallituksen toimesta.

## 2.2 Katkontaohje

Katkontaohje on valitun puuston hakkuuta varten laadittu suunnitelma. Se sisältää tiedot, joiden perusteella katkonta leimikossa suoritetaan. Näitä tietoja ovat esimerkiksi tilatut puutavaralajit ja niiden määrät (jakauma- eli tavoitematriisi) sekä hintamatriisi, joka luonnollisesti sisältää eri puutavaralajeista tilaajan maksamat hinnat. Tiedot lähetetään langattomasti hakkuukoneen tietokoneelle.

## 2.3 Hakkuumuodot

Hakkuu on puutavarantekoa, johon sisältyvät puun kaato, karsinta ja katkonta puutavaralajeiksi sekä katkotun puutavaran kasaus. Hakkuun voi suorittaa metsuri tai hakkuukone. Taulukossa 2.1 on listattu erilaiset hakkuutavat niistä datassa käytettyine koodeineen. Päätehakkuulla tarkoitetaan uudistamiskypsän metsän hakkuuta ja harvennushakkuulla metsän harventamista jäljelle jäävien puiden kasvuolosuhteiden parantamiseksi. Siis edellisessä metsän hedelmät kerätään tarkasti talteen ja jälkimmäisessä lähinnä hoidetaan metsää. Kaistaleavohakkuu on metsän uudistamismenetelmä, jossa metsää uudistetaan kaistaleittain ja pystyyn jäävän reunametsän puut siementävät hakatun alueen. Ensiharvennus on metsikössä tehtävä ensimmäinen hakkuu, joka tuottaa jonkin verran myös myyntikelpoista puutavaraa. Avohakkuu on metsänuudistamisessa käytettävä hakkuumenetelmä, jossa puusto poistetaan kokonaisuudessaan säästöpuita lukuun ottamatta istuttaen tai kylväen tapahtuvaa uudistamista varten. Harvennushakkuu on hakkuutapa, jossa tiheänä kasvavasta metsästä hakataan osa puustosta, jotta jätettävällä terveellä ja hyvälaatuisella puustolla on riittävästi kasvutilaa. Siemenpuuhakkuu tarkoittaa uudistuskypsän metsän hakkaamista siten, että pystyyn jätetään uudistamiseen tarvittavat siemenpuut. Tarkempaa tietoa hakkuumenetelmistä löytyy esimerkiksi Tapion taskukirjasta (2002).

**Taulukko 2.1** Hakkuutavat koodeineen

| **Koodi** | **Hakkuutapa** |
|---|---|
| **11** | Ensiharvennus |
| **12** | Muu harvennus tai väljennyshakkuu |
| **13** | Erirakenteisen metsän hakkuu |
| **14** | Siemen- tai ylispuiden poisto |
| **15** | Avohakkuu |
| **16** | Kaistaleavohakkuu |
| **17** | Siemen- tai suojuspuuhakkuu |
| **18** | Poimintahakkuu |
| **19** | Muu puuston käsittely |
| **20** | Uudistusalan raivaus |

## 2.4 Puutavaralajit

Rungon eri osat luokitellaan puutavaralajeiksi niiden ominaisuuksien ja arvojen mukaisesti. Karkein jaottelu puukaupan kannalta on jaottelu tukkeihin ja kuitupuuhun. Tukit menevät yleensä sahatavaraksi ja kuitupuut paperin raaka-aineeksi. Tukit ovat siis luonnollisesti arvokkaampia kuin kuitupuut. Taulukossa 2.2 on Metsähallituksen käyttämät puutavaralajit ja niiden koodit.

**Taulukko 2.2** Puutavaralajit koodeittain

| Nro | Tavaralaji | Nro | Tavaralaji | Nro | Tavaralaji |
|---|---|---|---|---|---|
| 110 | Mäntyraakki | 131 | Mäntypolttopuu | 315 | Erikoiskoivutukki |
| 111 | Mäntytukki | 210 | Kuusiraakki | 320 | Koivuhylky |
| 112 | Mäntypikkutukki | 211 | Kuusitukki | 321 | Koivukuitu |
| 113 | Pylväs | 212 | Kuusipikkutukki | 331 | Koivupolttopuu |
| 114 | Ratapuuaihe | 215 | Erikoiskuusitukki | 410 | Haaparaakki |
| 115 | Erikoismäntytukki | 220 | Kuusihylky | 411 | Haapatukki |
| 116 | Hirsitukki | 221 | Kuusikuitu | 420 | Haapahylky |
| 117 | Sahamänty | 223 | Kuusisellu | 421 | Haapakuitu |
| 118 | Kelotukki | 231 | Kuusi polttopuu | 621 | Muu lehtikuitu |
| 120 | Mäntyhylky | 310 | Koivuraakki | 631 | Muu polttopuu |
| 121 | Mäntykuitu | 311 | Koivutukki | 711 | Lehtikuusitukki |
| 129 | Muu havukuitu | 312 | Koivupikkutukki | 839 | Metsähake |

## 2.5 Läpimittaluokka

Läpimittaluokalla tässä tutkimuksessa tarkoitetaan rinnankorkeusläpimittaa eli puun rungon halkaisijaa 1.3 metrin korkeudelta rungon tyvestä mitattuna. Harvesterin koura mittaa läpimitan ns. kolmiomenetelmällä (Uusitalo 2003, s. 152). Tätä menetelmää käytetään, koska runko ei yleensä ole tasaisen pyöreä vaan pikemminkin soikea, joten näin saadaan tarkempi mittaustarkkuus.

## 2.6 Rungon viat

Rungossa olevat viat laskevat puusta saatavaa hintaa. Järeydeltään tukiksi kelpaava tukki joudutaan luokittelemaan vikojen vuoksi kuitupuuksi, polttopuuksi tai jopa lumpiksi. Liian käyrä tukki eli lenko luokitellaan kuitupuuksi. Polttopuuksi tukki voi joutua, jos se on laho eli sen tiheysominaisuudet eivät vastaa vaadittavia. Kaikkein huonolaatuisin tukki eli lumppi ei kelpaa edes polttopuuksi. Sen arvo on tällöin luonnollisesti nolla ja se joudutaan heittämään pois.

## 2.7 Apteeraus

Puiden runkojen katkontaa määrämittaisiksi osiksi kutsutaan apteeraukseksi tai pölkyttämiseksi (Tapion taskukirja 2002; Uusitalo 2003) ja tällä tarkoitetaan puun rungon katkaisukohtien määrittämistä mitta- ja laatuvaatimusten mukaisesti eri puutavaralajeiksi. Apteeraus voidaan vielä jakaa kahteen osaan: arvoapteeraus ja jakauma-apteeraus. Arvoapteerauksessa puutavaralajeihin jaon perusteena käytetään pelkästään hintamatriisia, joka kertoo jokaisen puutavaralajin eri pituus- ja halkaisijaluokittaiset arvot. Arvoapteeraus ei siis ota huomioon tavaralajikohtaisia tavoitejakaumia, vaan se pyrkii maksimoimaan yksittäisen puun rungon arvon (Uusitalo 2003). Jakauma-apteerauksessa otetaan lisäksi huomioon jakauma- eli tavoitematriisi, joka sisältää ostajan tarvitsemat kappalemäärät eri puutavaralajeja. Tällöin uuden rungon apteerauksessa huomioidaan siis myös jo kertyneet tukkimäärät. Yleensä hakkuukoneen tietokone optimoi katkontaa jakauma-apteeraus-periaatteen mukaan. Apteerauksen toteutumisen eroja voidaan kuvailla sekä käytännössä että teoriassa.

Teoriassa apteeraus etenee seuraavanlaisesti: harvesteri katkaisee rungon puun tyvestä ja sen jälkeen mittalaite mittaa noin 10 cm:n välein rungon läpimittaa 3–4 metrin matkalta. Näiden mittausten sekä korjuuohjeeseen sisältyvien hinta- ja kysyntämatriisien perusteella tietokone tekee ennustuksen optimaalisista katkaisukohdista maksimoiden koko rungon arvon. Ensimmäisen katkaisun jälkeen harvesterin koura etenee kohti latvaa ja tekee samalla edelleen mittauksia rungon läpimitoista. Sen pohjalta korjataan tietokoneen ensimmäistä ennustetta ja luodaan ennuste rungon loppuosalle. Toisen katkaisun jälkeen samaa prosessia toistetaan, kunnes runko kokonaisuudessaan on käsitelty.

Käytännössä kuljettajan työpanoksella on suuri merkitys apteerauksen onnistumisessa. Hänellä on käytännössä oltava samanlaiset valmiudet kuin metsureilla ja lisäksi on vielä hallittava koneen käsittely ja kunnossapito (Salakari & Heimonen 1998, s. 23). Hänen täytyy esimerkiksi tarkastaa, onko tukkirungossa vikoja, kuten lenko, laho, mutka, koro tai latvan vaihto, ja siten havaita, onko pölkky lajiteltava kuituun, vaikka tietokone ehdottaisi tukkia. Lenkouden tarkkailussa on hyvä ottaa myös tukin pituus huomioon, koska se saattaa vaikuttaa siihen, että tukki voidaan tukkina eikä sitä tarvitse luokitella kuitupuuksi. Rungon alaosan lahoutta täytyy myös tarkkailla. Yleensä varsinkin kokeneempi kuljettaja osaa etsiä lahoudesta kertovia merkkejä jo ennen kuin harvesterin koura tarttuu puuhun. Yleensä kuitenkin runkoa käsiteltäessä se täytyy kääntää niin, että pohja näkyy, ja jos se on musta keskustan kohdalta, rungosta sahataan pieni pätkä tyvestä pois ja katsotaan uudestaan. Jos se on edelleen musta, niin taas sahataan pieni pätkä pois. Jos lahoa ei enää ole, niin apteerataan normaalisti, muuten koko lahorunko menee polttopuuksi. Kuljettaja antaa tietokoneelle tiedon puulajista ennen kuin tietokone tekee ehdotuksen rungon jakamisesta mahdollisiin erikokoisiin eri puutavaralajeihin. Jokaisen sahauksen kohdalla kuljettaja joko hyväksyy tietokoneen tekemän ehdotuksen tai sitten hylkää sen. Yleensä huonolaatuisessa metsässä tietokoneen antamaa arviota ei voida käyttää, vaan se joudutaan lenkouden takia usein hylkäämään ja tekemään vaihtoehtoinen parhaaksi katsottu päätös sahauksen toteuttamiseksi. Hyvälaatuisessa metsässä ja harvennushakkuussa koneen antamaa katkaisuehdotusta voidaan yleensä noudattaa.

Jotta työn vaativuus tulisi oikein esille, esitetään vielä hakkuukoneenkuljettajalta vaadittavat ominaisuudet, kuten Salakari ja Heimonen ovat kirjassaan (1998, s. 24) esittäneet harvesterin kuljettajalta vaadittavat ominaisuudet:

- oltava tehokas (hyvissä olosuhteissa on hakattava 200 $m^3$ vuorossa)
- osattava tukin apteeraus ja harvennusmetsän käsittely oikeaan tiheyteen
- osattava suunnitella ajourat
- harvennustyön jäljen eli korjuuvauriot ja jäävän puuston tiheyden oltava kohdallaan
- hallittava ympäristöasiat
- osattava pitää kone kunnossa ja korjata pikkuvikoja
- osattava tehdä päivittäiset huollot
- hallittava koneen mittalaitteen käyttö ja kunnossapito sekä katkonnan ohjaus
- kyettävä tekemään pitkää vuoroa
- osattava työskennellä sekä yksin että tiimissä.

# 3 Tutkimusaineisto

Tässä luvussa hahmotellaan tutkimukselle alun perin asetetut tavoitteet, kerrotaan kuinka ja mistä muuttujista tutkimusaineisto on muodostettu. Koska tutkimusmenetelmäksi on valittu regressioanalyysi, esitellään miten selitettävä muuttuja eli tulos on muodostettu ja minkä muuttujien avulla tulosta on tarkoitus selittää.

Tavoitteena on löytää funktio tai funktiot, jotka kuvaavat puulajeittain sekä keskimääräisen että tavoitteellisen rungon arvon (€/runko) läpimittaluokittain. Hakkukoneiden tietokoneilta saatua aineistoa tutkimalla etsitään parhaiten tulosta selittävät muuttujat ja suoritetaan toteutuneen apteerauksen vertailu. Tavoite on hahmoteltu seuraavien kolmen olennaisen kysymyksen kannalta:

1. Mitkä muuttujat selittävät parhaiten toteutunutta tulosta rungon eri läpimittaluokissa?
2. Kuinka muodostaa tavoitteellinen tulos eli vertailupohja?
3. Kuinka vertailu toteutuneen ja tavoitteellisen tuloksen välillä voidaan esittää yksinkertaisesti ja selkeästi?

## 3.1 Datan saattaminen analysoitavaan muotoon

### 3.1.1 Aineisto

Tutkimuksen tavoite saavutetaan metsäkoneilta saatavan datan (PRD-tiedosto) ja korjuuohjeen datan analysoinnin avulla. PRD-tiedosto sisältää todella paljon eri tietoja katkontatapahtumasta, mutta niistä on mukaan tutkimukseen valittu vain ne, jotka etukäteen vaikuttavat olennaisilta.

### 3.1.2 Datamatriisi

Datan saattaminen datamatriisimuotoon analysointien suorittamiseksi oli ensimmäinen tehtävä ennen aineiston tutkittavaksi saamista. Alkuperäinen data sijaitsi hakkuuohjeessa, PRD-tiedostossa ja varastotiedoissa. Kaikki nämä tiedot ovat saatavilla Metsähallituksen tietojärjestelmästä ja näistä tiedoista muodostettiin datamatriisi. Matriisin muoto esitettiin toimeksiantajalle, joka hoiti sen teknisen toteuttamisen. Tämä oli tarpeen, sillä PRD-tiedoston käsittely vaatii sen erittäin yksityiskohtaista tuntemista ja ilman asiantuntemusta pelkkä datan kerääminen siitä olisi vienyt koko työhön varattuun aikaan nähden suhteettoman paljon aikaa ja voimavaroja.

## 3.2 Selitettävä muuttuja eli vaste

Tutkimuksen tarkoituksena on selittää apteerauksen tulosta euroa per runko. Läpimittaluokittainen tulos muodostetaan katkottujen puutavaralajien tilavuuksien arvojen summana. Eri puutavaralajeille on tutkimusta varten muodostettu taulukon 3.1 mukaiset arvot. Käytetyt referenssihinnat sisältävät myös kuljetuskustannukset, joten ne ovat normaaleja tehdashintoja huomattavasti suuremmat. Vielä täsmentääkseni asiaa lainaan tutkimuksen ohjaajaa Metsähallituksen puolelta, Tuomo Vuorenpäätä: ”Arvot ovat suhteellisia, ja tavaralajien välisiä hintaeroja on jonkin verran korostettu. Nämä arvot toimivat ainoastaan välineenä apteerauksen tason määrittämisessä, jonka yllä esitetyt arvotasot ohjaavat parhaiten halutun suuntaiseen tulokseen, huolelliseen apteeraukseen.”

**Taulukko 3.1** Puutavaralajien hinnat ja koodit

| TVL-koodi | Hinta: €/m3 | TVL-nimi |
|---|---|---|
| 110 | 90 | Mäntyraakki |
| 111 | 300 | Mäntytukki |
| 112 | 120 | Mäntypikkutukki |
| 113 | 300 | Pylväs |
| 114 | 120 | Ratapuu |
| 115 | 350 | Erikoismäntytukki |
| 116 | 250 | Hirsitukki |
| 117 | 110 | Sahamänty |
| 118 | 200 | Kelotukki |
| 120 | 50 | Hylky |
| 121 | 100 | Mäntykuitu |
| 131 | 50 | Polttopuu |
| 210 | 90 | Kuusiraakki |
| 211 | 300 | Kuusitukki |
| 212 | 120 | Kuusipikkutukki |
| 215 | 350 | Erikoiskuusitukki |
| 220 | 50 | Kuusihylky |
| 221 | 120 | Kuusikuitu |
| 223 | 100 | Kuusisellu |
| 231 | 50 | Kuusipolttopuu |

Työmaan tulos jokaisessa läpimittaluokassa lasketaan kaavan (3.1) mukaisesti:

(3.1) $$Tulos = V_1 * H_1 + V_2 * H_2 + ... + V_n * H_n,$$

missä $V_i$ on tavaralajin tilavuuskertymä työmaalla ja $H_i$ vastaavan tavaralajin referenssihinta. Jotta tulos saataisiin vertailukelpoiseksi työmaiden välillä, se skaalataan relevantissa läpimittaluokassa katkottujen runkojen määrällä (kaava (3.2))

(3.2) *Tulos per runko = Tulos / Runkojen lkm.*

Koska tutkimuksessa tuloksesta puhuttaessa tarkoitetaan yleensä kaavan (3.2) mukaista tulosta eli tulosta per runko, on syytä pitää tämä asia mielessä tutkimusta läpi käytäessä (erikseen mainitaan, mikäli tuloksella tarkoitetaan jotain muuta).

## 3.3 Selittävät muuttujat

**Läpimittaluokka** on ehkä muuttujista tärkein, sillä se on rungon paksuus rinnankorkeudelta eli 1.3 metrin korkeudelta puun tyvestä mitattuna. Muuttuja antaa tietoa ainakin puun kasvuolosuhteista ja iästä. Salakarin ja Heimosen (1998, s. 15) mukaan hakkuun tuotokseen voimakkaimmin vaikuttava tekijä on rungon koko ja läpimittaluokka mittaa parhaiten tätä ominaisuutta tässä aineistossa. Heidän mukaansa myös hakkuutapa (harvennus- vai päätehakkuu), hehtaarikohtainen kertymä, mahdollinen alikasvusto, puulaji ja maastotekijät ovat tuotokseen vaikuttavia tekijöitä.

Säädeltävissä olevia muuttujia ovat **tavaralajin minimiläpimitta**, **ensimmäisen katkontavaihtoehdon** (tarkoittaa lyhyintä hyväksyttyä puutavaralajin katkontapituutta) **pituus** ja **pituusluokkien lukumäärä** sekä **tavaralajien lukumäärä**. Säädeltävyys tarkoittaa, että kyseisten muuttujien arvoja voidaan muutella korjuuohjetta luodessa. Tavaralajien lukumäärä lasketaan luonnollisesti niiden tavaralajien summana, mitä leimikolta on korjuuohjeen mukaan mahdollista kerätä. Siitä, onko kyseessä olevalta leimikolta kerätty tiettyä tavaralajia, kertovat männyn tapauksessa **tavaralajimuuttujat** pikkutukki (PT), hirsitukki (HT), ratapuuaihio (RP) ja sahamänty (SM) ja kuusen tapauksessa kuusipikkutukki (KPT), erikoistukki (KET) ja sellu (KSE). Muuttujat ovat dikotomisia ja saavat arvon yksi, jos kyseistä tavaralajia on kerätty ja arvon nolla, jos sitä ei ole kerätty. Myös **hakkuutapa** on yhtälailla säädeltävissä oleva muuttuja. Leimikosta tiedossa olevien tietojen perusteella voidaan valita sille järkevin hakkuutapa.

**Pohjapinta-alamediaaniläpimitta** on pohjapinta-alalla painotettu mediaanipuun läpimitta kyseisellä leimikolla. Tämä tarkoittaa läpimittaa, jota suuremmissa ja pienemmissä puissa on 50 % pohjapinta-alasta. Muuttuja on laskettu toteutuneen katkonnan perusteella siten, että ensin lasketaan kaavalla (3.3) työmaan läpimittaluokittaisten pohjapinta-alojen summan puoliväli.

(3.3) $$S_h = \frac{\sum_{1}^{i}\left(\frac{\pi}{4} * d_{1.3i}^2 * n_i\right)}{2},$$

missä $d_{1.3i}$ on $i$. rinnankorkeusläpimitta eli läpimittaluokka ja $n_i$ on $i$:nnen läpimittaluokan runkojen lukumäärä. Seuraavaksi summataan läpimittaluokittaisia pohjapinta-aloja, kunnes kumulatiivinen summa on yhtä suuri kuin $S_h$. Se läpimittaluokka, jossa $S_h$ täyttyy, ilmaisee luokan millä pohjapinta-alamediaaniläpimitta sijaitsee. Merkitään tätä läpimittaluokkaa $LPM_{Sh}$:lla. Seuraavaksi lasketaan, monennenko $LPM_{Sh}$:n rungon kohdalla $S_h$ täyttyy. Merkitään tätä $R_{Sh}$:lla. Lopuksi oletetaan, että puut ovat tasajakautuneet läpimittaluokan sisällä ja interpoloidaan mediaaniläpimitta lineaarisesti:

(3.4) $$DBA = LPM_{Sh} + \frac{R_{Sh}}{R_{TOTAL}} * 1cm,$$

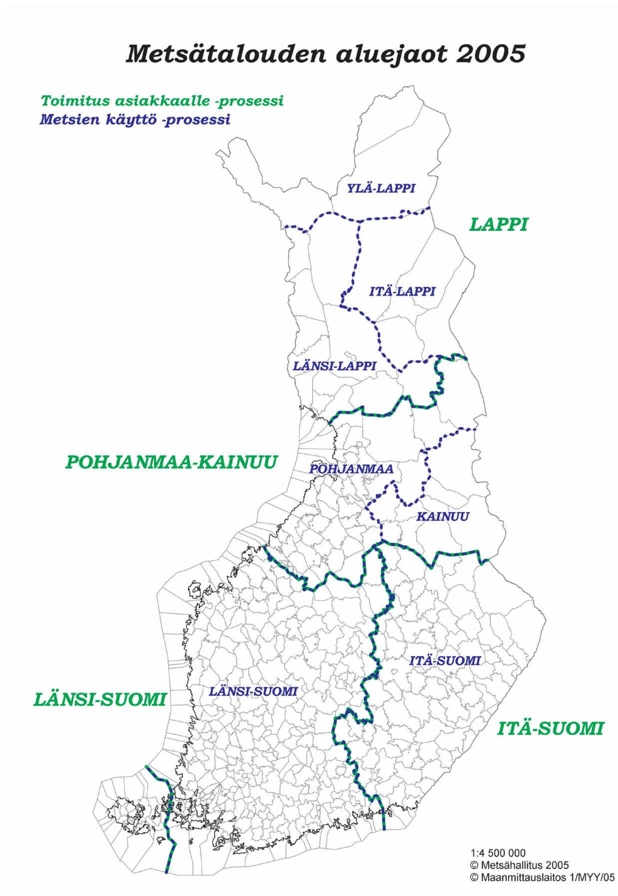
missä $R_{TOTAL}$ on $LPM_{Sh}$:n runkojen lukumäärä ja DBA on pohjapinta-alamediaaniläpimitta.

Leimikon ominaisuuksia edellisen lisäksi kuvastaa myös **pääpuulaji**. Aineistossa tästä kertoo männyn suhteellinen osuus leimikon katkotuista kuusista ja männyistä. Muuttuja on laskettu kaavan (3.5) mukaisesti.

(3.5) $$Mosuus = \frac{V_M}{V_M + V_K},$$

missä $V_M = \sum_{1}^{n} V_{M_i}$ ja $V_K = \sum_{1}^{m} V_{K_j}$ ovat katkottujen mäntyjen ja kuusien käyttöosuuksien kokonaistilavuuksien summat.

**Aluemuuttuja** on tutkimuksessa yksinkertaisuuden ja aineiston koon vuoksi valittu siten, että se saa arvon nolla, jos kyseessä on etelä ja arvon yksi, jos kyseessä on pohjoinen. Jako etelään ja pohjoiseen perustuu Metsähallituksen aluejakoon (kuvio 3.1). Etelän muodostavat yhdessä Länsi- ja Itä-Suomi, loput alueet katsotaan kuuluvaksi pohjoiseen.




**Kuvio 3.1** Metsätalouden aluejako

Myös **yksikkö**, joka muodostuu tietyn alueen työmaista, kertoo leimikon maantieteellisestä sijainnista. Yksiköiden käyttöä tähän tarkoitukseen rajoittaa kuitenkin niiden pienestä koosta johtuva aineiston epätasainen kattavuus. Joiltain yksiköiltä on enemmän havaintoja ja toisilta ei juuri lainkaan. Metsän laadusta kertovana muuttujana mukana on **lumppien suhteellinen osuus** leimikolta katkotusta puutavarasta. On hyvin vaikeaa varmasti todentaa, johtuuko tietyllä leimikolla lumppien suurempi osuus huonommasta puustosta vai hakkuukoneen kuljettajan taidon puutteesta. Jos on epäilys, että kuljettaja tekee työnsä tässä mielessä huonommin kuin muut, voidaan tarkastella kyseisen kuljettajan lumppiosuuksien historiaa ja verrata sitä muiden kuljettajien vastaavaan.

Erittäin olennainen muuttuja läpimitan lisäksi on **rungon käyttöosan keskipituus** (kaava (3.6)), joka on laskettu erikseen sekä männylle että kuuselle.

$$Kpituus = \frac{JM}{r}, \qquad (3.6)$$

missä *JM* on katkotun puutavaran juoksumetrimäärä ja *r* on runkojen lukumäärä. Tämä tunnus kuvastaa työmaalta katkottujen puiden käyttöosuuksien keskimääräistä pituutta ja mitä suurempi tämä arvo on sitä enemmän mahdollisia katkontavaihtoehtoja.

Taulukossa 3.2 on listattu tutkimuksen muuttujat niistä käytettyine koodeineen.

**Taulukko 3.2** Selittävät muuttujat

| **Muuttuja** | **Tunnus** |
|---|---|
| Läpimittaluokka | LPM |
| Minimiläpimitta (tukki) | Minlpm |
| Ensimmäisen katkontavaihtoehdon pituus (tukki) | M1 |
| Pituusluokkien lukumäärä (tukki) | Pituuslkt |
| Minimiläpimitta (kuitu) | MinlpmK |
| Ensimmäisen katkontavaihtoehdon pituus (kuitu) | M1K |
| Pituusluokkien lukumäärä (kuitu) | PituuslktK |
| Pohjapinta-alamediaaniläpimitta | DBAM (mänty) tai DBAK (kuusi) |
| Männyn suhteellinen osuus kokonaistilavuudesta | Mosuus |
| Lumppien suhteellinen osuus | MLosuus (mänty) tai KLosuus (kuusi) |
| Tavaralajit | Mtvl (mänty) tai Ktvl (kuusi) |
| Pikkutukki | PT |
| Hirsitukki | HT |
| Ratapuuaihio | RP |
| Sahamänty | SM |
| Kuusipikkutukki | KPT |
| Erikoistukki | KET |
| Sellu, kuusi | KSE |
| Hakkuutapa | FH |
| Yksikkö | Yksikko |
| Alue | Alue |
| Käyttöosan keskipituus | Kpituus |

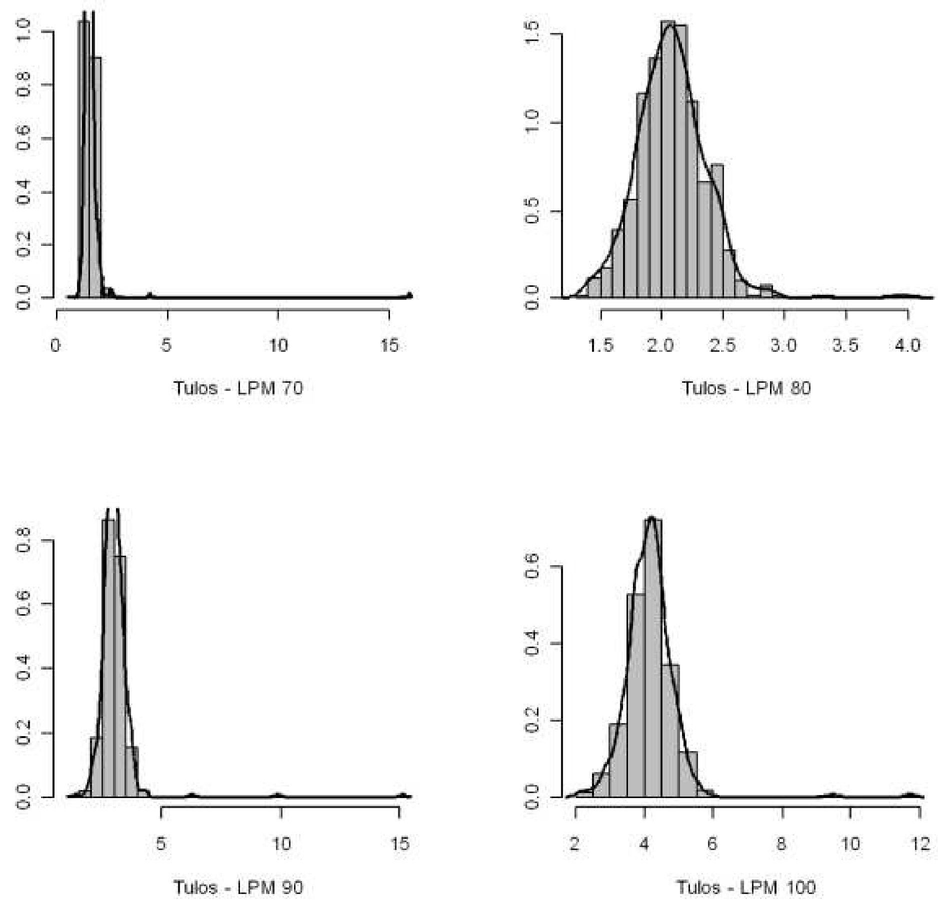
# 4 Datan kuvailevaa tarkastelua

Datan alustava tarkastelu suoritetaan kuvailevan tilastotieteen keinoin. Tarkoituksena on saada muuttujista tietoa, jota voidaan käyttää hyväksi tarkempien analyysien suorittamisessa. Tämä vaihe on erityisen tärkeä myös siksi, että mahdolliset aineiston virheet saadaan karsittua pois ja tuloksista saadaan luotettavampia. Kuvailu pitää sisällään selitettävän ja selittävien muuttujien keskinäisten riippuvuuksien, muuttujien arvojen jakaumamuotojen, luokitteluasteikollisten muuttujien ryhmien välisten erojen, riippuvuuksien muotojen ja yksinkertaisten tunnuslukujen tarkasteluja. Mänty ja kuusi on tarkasteltu erikseen, sillä niiden arvotuksessa sekä ominaisuuksissa on selkeitä eroja.

## 4.1 Mänty

### 4.1.1 Tulosten jakaumia rinnankorkeusläpimittaluokittain

Tulosten läpimittaluokittaisia jakaumia voidaan tarkastella esimerkiksi histogrammien sekä varsi ja lehti –esitysten avulla. Näin saamme kuvan siitä, kuinka havainnot ovat eri läpimittaluokissa jakautuneet. Jälkimmäinen tarkastelutapa on myös tarpeen, sillä histogrammeista (kuvio 4.1) ei ole yhtä helppoa identifioida epämääräisiä havaintoja, kuten vertailemalla kuvioita 4.1 ja 4.2 helposti havaitaan.




**Kuvio 4.1** Tuloksen läpimittaluokittaisia histogrammeja (mänty)

Esimerkiksi läpimittaluokassa 70 tuloksien arvot keskittyvät lähes kokonaan välille [1,2]. Muutama arvo on välillä [2,3] ja kaksi havaintoa on jokseenkin muista poikkeavia. Lähes varmasti voidaan sanoa, että havainto, jonka arvo on 15.1, johtuu aineiston virheestä. Kun taas havaintoa, jonka arvo on 4.2, ei välttämättä voida tuomita samalla tavoin.

```
Desimaalipilkku kuuluu kohtaan |

   1 | 4899
   2 | 00000111222222222233333333344444444444444444445555555555555555555555555555+176
   3 | 0000000000000000000000000000000000000000000000000000000001111111111111111+182
   4 | 01233
   5 |
   6 | 3
   7 |
   8 |
   9 | 9
  10 |
  11 |
  12 |
  13 |
  14 |
  15 | 1
```

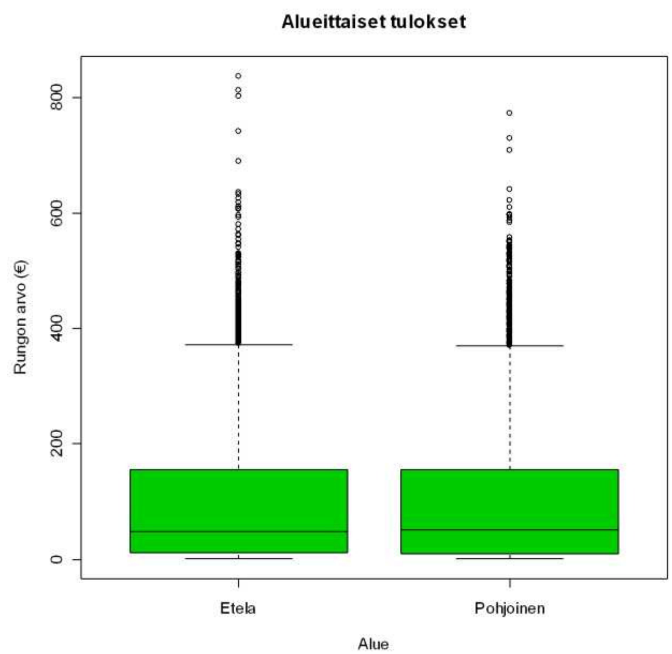
**Kuvio 4.2** Varsi ja lehti –esitys, LPM 90 (mänty)

Varsi ja lehti –esityksestä nähdään, että läpimittaluokassa 90 on joitain selkeästi muista saman läpimittaluokan havainnoista poikkeavia arvoja. Histogrammeista ei samaa yhtä selkeästi voida havaita.. Ei ole tarpeen esittää kaikkien läpimittaluokkien histogrammeja tai varsi ja lehti (*stem and leaf*) -esityksiä, sillä jo näistä käy esille, että joissain läpimittaluokissa on poikkeavia havaintoja. Tällaiset havainnot on syytä eliminoida aineistosta ennen regressiomallien rakentamista. Olisi tietysti järkevää analysoida jokainen havainto erikseen ja etsiä mahdollisia syitä poikkeavuudelle. Usein juuri tällainen yksittäisten havaintojen analysointi saattaa tuottaa mielenkiintoista informaatiota. Tähän ei tässä tutkimuksessa kuitenkaan ole mahdollista paneutua, sillä tällaisia havaintoja on koko aineistossa paljon. Poistetaan aineistosta poikkeavat havainnot ja oletetaan suurimman osan niistä olevan aineiston virheitä, jotka voivat johtua monesta syystä. Mahdollisesti ne ovat seurausta hakkuukoneen mittaus- tai rekisteröintivirheistä.

### 4.1.2 Alueittainen tuloksen tarkastelu

Tarkastellaan seuraavaksi, miten työmaiden jako pohjoiseen ja etelään näkyy tuloksissa. Aineistossa on etelästä saatuja läpimittaluokittaisia tuloksia 4508 ja pohjoisesta saatuja 11691 kappaletta. Tarkastelussa käytetään hyväksi laatikko-jana –kuviota (*Box-plot* tai *Box-and-whiskers -plot*) sekä ehdollistamiskuviota, jonka avulla voidaan havaita, mikäli vasteen arvot riippuvat eri tavalla selitettävän muuttujan eri arvoista.

Alueittaisissa tuloksissa ei laatikko-jana –kuvaajan valossa näyttäisi olevan suuria eroja (kuvio 4.3).

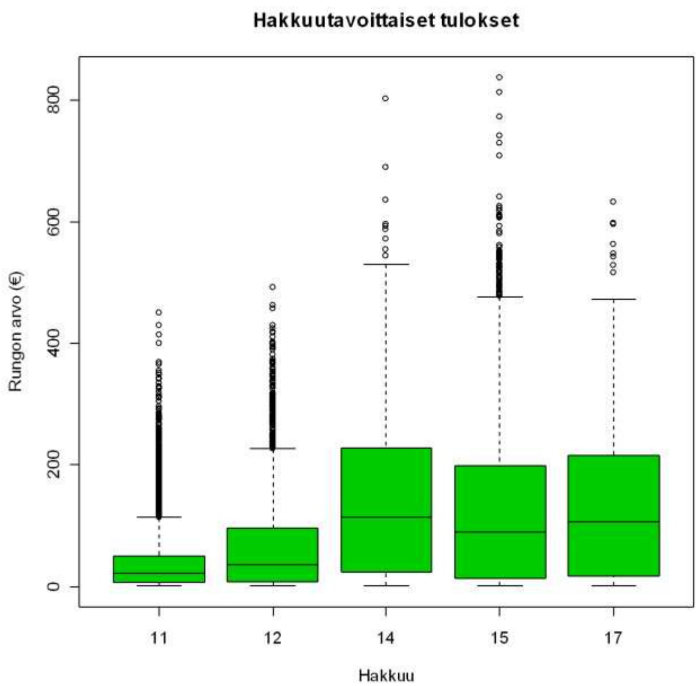



**Kuvio 4.3** Tuloksen jakauma läpimitoittain maantieteellisen sijainnin mukaan (mänty)

### 4.1.3 Hakkuutavoittainen tuloksen tarkastelu

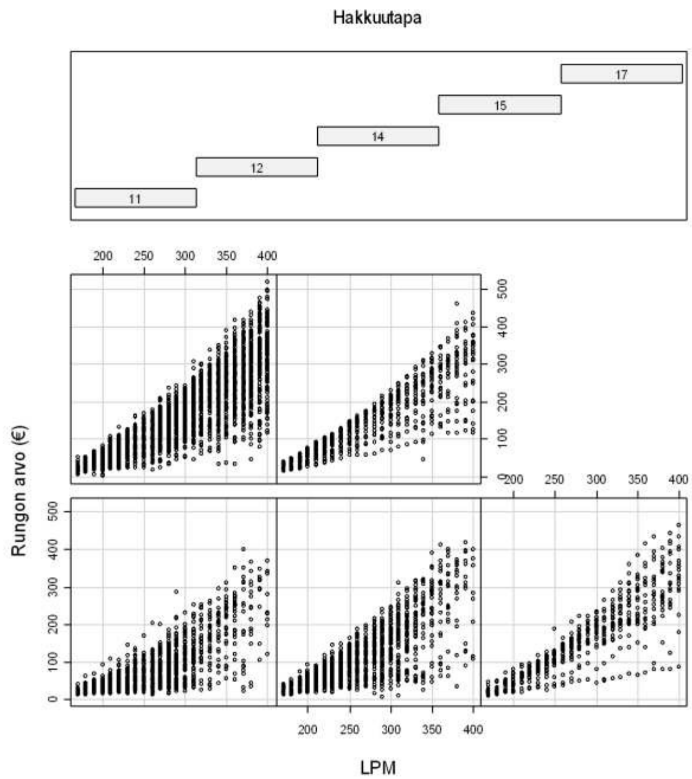
Hakkuutapa kertoo työmaan katkonnan perusteista. Vaikka taustalla olisikin oletuksena, että hakkuutapa ei vaikuta läpimittaluokittaiseen tulokseen, sillä saman rinnanympärysläpimitan omaavista puista pitäisi periaatteessa saada sama tulos välittämättä hakkutavasta, on syytä ottaa tarkasteluun myös tämä muuttuja. Hakkuutapa voi nimittäin tuoda aineistoon ja vertailupohjan rakentamiseen leimikosta tietoa, jota ei muuten ole käytössä. Tällaista tietoa voi tulla esimerkiksi epäsuorasti tietynlaisten leimikoiden valikoitumisesta tietyille hakkuutavoille. Luvun 2.3 taulukossa 2.1 on esitelty mahdolliset vaihtoehdot hakkuille.

Analysoitavaan aineistoon on otettu viisi eri hakkuutapaa: ensiharvennus (11), muu harvennus tai väljennyshakkuu (12), siemen- tai ylispuiden poisto (14), avohakkuu (15) ja siemen- tai suojuspuuhakkuu (17). Kuviossa 4.4 on tarkasteltu hakkuutavoittaisia tulosten jakaumia. Näyttäisi, että hakkuutavoilla 11 ja 12 tulosten hajonta on pienempää, kuten ovat myös tuloksien mediaanit.




**Kuvio 4.4** Hakkuutavoittaiset tulokset (mänty)

Tarkastellaan, miten tulokset ovat jakautuneet hakkuutavoittain eri läpimittaluokissa (kuvio 4.5). On havaittavissa, että jakaumissa on eroja. Esimerkiksi hakkuutavoilla 11 ja 12 on suhteellisen pieniä tuloksia myös sellaisissa läpimittaluokissa, mistä muilla hakkuutavoilla on saatu parempaa tulosta (kuvion vasemmassa alareunassa on hakkuutapa 11, hakkuutapa 12 on alhaalla keskellä, jne.). Lisäksi hakkutavan 15 hajontakuvio (vasen yläkulma) on selkeästi leveämpi kuin hakkuutavan 17 hajontakuvio (oikea yläkulma).




**Kuvio 4.5** Tulosten ja läpimittojen hajontakuviot hakkuutavoittain (mänty)

### 4.1.4 Tavaralajeittaisia tuloksen tarkasteluja

Tavaralajimuuttujiksi tutkimuksessa kutsutaan muuttujia, jotka ilmaisevat, onko työmaalta kerätty jotain tiettyä tavaralajia. Muuttuja saa arvon yksi, jos tavaralaji on kerätty. Luonnollisesti muuttujan nimi ilmaisee, mistä tavaralajista on kyse.

Tarkastellaan, onko kerättävillä puutavaralajeilla vaikutusta tulokseen. Työmaalta voidaan joko kerätä tiettyä tavaralajia tai jättää se keräämättä ja kerätä sen sijasta jotain muuta. Pikkutukin tai sahamännyn keräämisellä tai keräämättä jättämisellä ei ole suurta merkitystä tulokseen (liite A, kuvio A.1). Hirsitukin ja ratapuuaihion suhteen näyttää, että niiden keräämisellä olisi saavutettu mediaaniltaan hiukan parempi tulos.

Tarkastellaan vielä, miten tulokset ovat jakautuneet eri läpimittaluokissa riippuen siitä, onko työmaalta kerätty kyseistä tavaralajia vai ei (liite A, kuviot A.2-A.5). Ratapuuaihiolla ja hirsitukilla on pienempi tulosten hajonta ja niillä tuloksen kasvu läpimitan kasvaessa on terveemmällä pohjalla. Tämä tarkoittaa sitä, että järeimpien läpimittaluokkien runkojen arvo saadaan tarkemmin talteen kuin pienempien luokkien runkojen arvo. Tämä tulos saattaa johtua leimikon valikoitumisesta katkonnan suunnitteluvaiheessa.

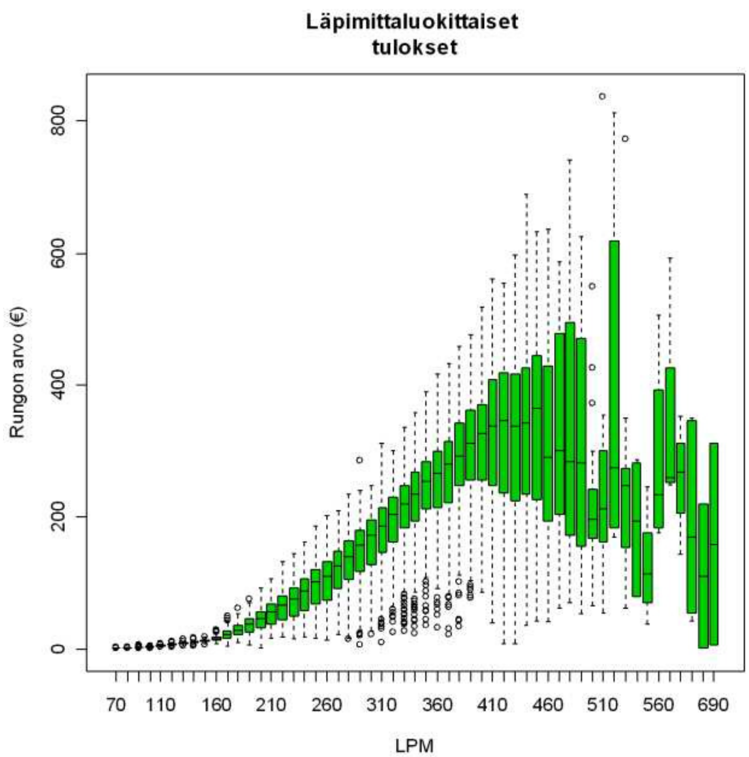
### 4.1.5 Parametrien tarkastelu

Työmaakohtaisesti säädeltäviä parametreja tutkimuksessa ovat tukki- ja kuitutavaralajeille erikseen tavaralajin minimiläpimitta, lyhyimmän katkonvaihtoehdon pituus ja käytössä olevien pituusmittojen lukumäärä. Tukkitavaralajina tässä pidetään vain tavaralajia 111 eli mäntytukkia ja kuitutavaralajina tavaralajia 121 eli mäntykuitua. Ongelmaksi muodostui näiden parametrien määrittely. PRD-tiedostoista poimittu data sisälsi yleensä samalla työmaalla saman läpimittaluokan sisällä monta eri vaihtoehtoa parametreiksi. Tukkitavaralajia saattoi olla jopa neljää eri luokitusta ja kaikilla eri arvot parametreina. Käytännössä oli meneteltävä siten, että minimiläpimitoista ja lyhyimmistä katkontavaihtoehtojen pituuksista valittiin minimit ja pituusmittojen lukumääristä maksimit. Tämä on tietysti omiaan vähentämään parametrien painoarvoja analyyseissä, sillä näin tehtäessä menetämme informaatiota. Parametrien kuvaileva tarkastelu ei tuonut esille mitään erikoista ja niiden kuvaajat ovat liitteenä (liite A, kuviot A.6 ja A.7).

### 4.1.6 Tuloksen läpimittaluokittainen tarkastelu

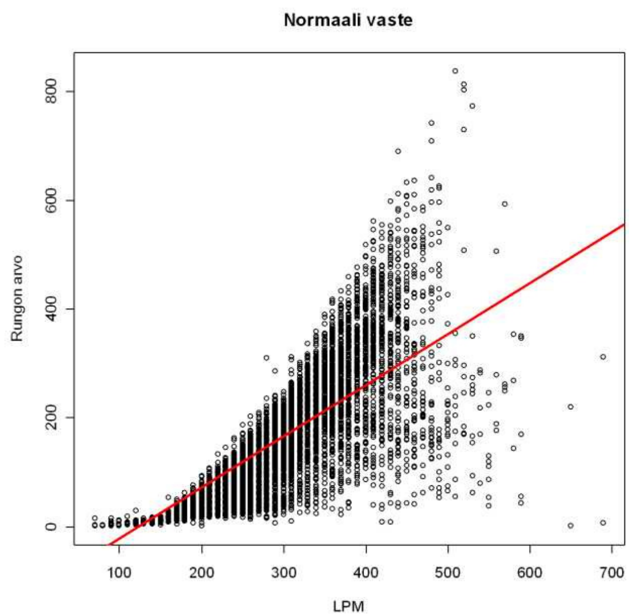
Läpimittaluokka on muodostettu mittaamalla puun rungon halkaisija millimetreinä 1,3 metrin korkeudelta eli rinnankorkeudelta ja luokittelemalla nämä mittaukset kymmenen mm:n levyisiin luokkiin. Esimerkiksi läpimittaluokkaan 140 kuuluvat kaikki rungot, joiden halkaisijan pituudet ovat vähintään 140 mm ja alle 150 mm.

Tarkastellaan läpimittaluokittaisten tulosten jakaumaa (kuvio 4.6). Tuloksen hajonta kasvaa selkeästi läpimittaluokan kasvaessa.

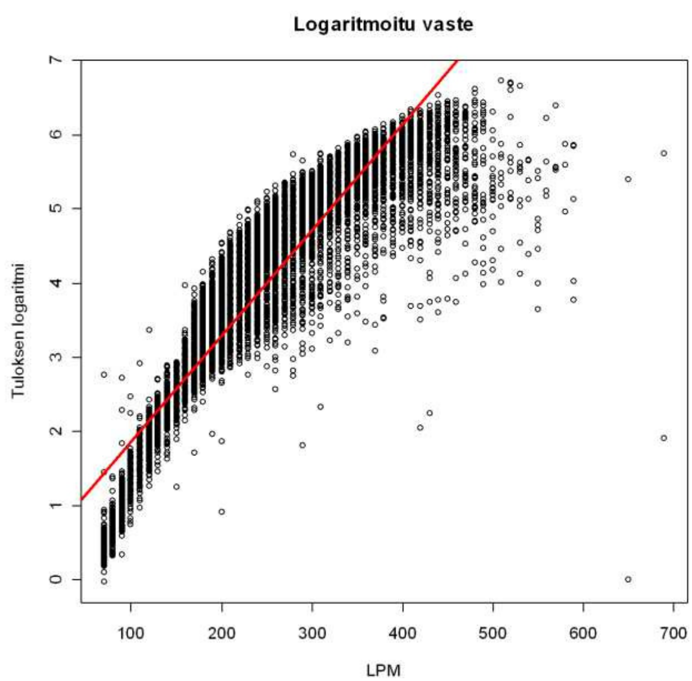



**Kuvio 4.6** Läpimittaluokittaiset laatikko-jana –kuviot (mänty)

Haluttaessa tarkastella onko läpimittaluokan ja tuloksen välillä lineaarista riippuvuutta, voidaan sovittaa kyseisten muuttujien väliseen hajontakuvioon lineaarinen regressiosuora (kuvio 4.7). Kuvion perusteella voidaan todeta, että riippuvuus ei ole täysin lineaarista.



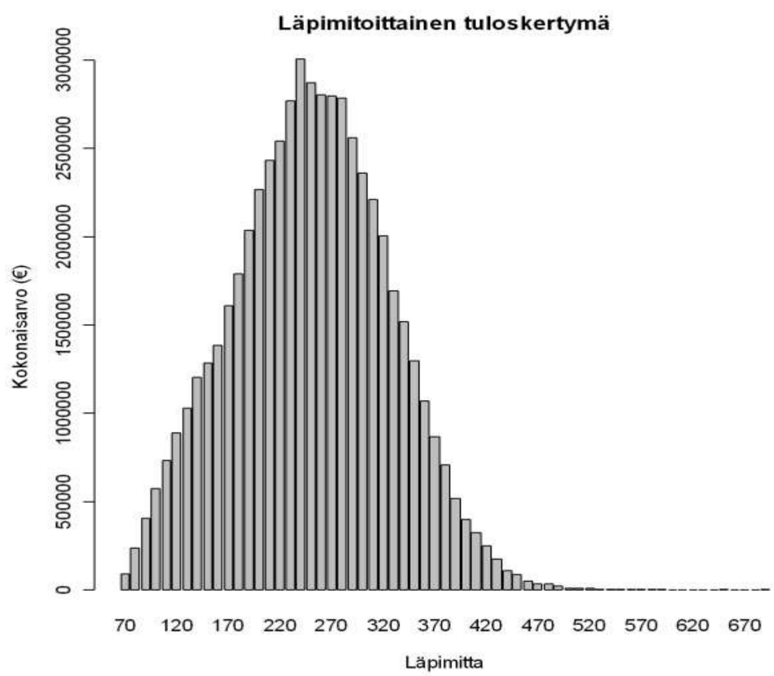

**Kuvio 4.7** Tuloksen ja läpimittaluokan hajontakuvio sekä siihen lisätty regressiosuora (mänty)




**Kuvio 4.8** Logaritmoidun tuloksen ja läpimittaluokan hajontakuvio sekä siihen lisätty regressiosuora (mänty)

Tehdään tulokselle logaritmimuunnos ja tarkastellaan onko riippuvuus nyt lineaarista (kuvio 4.8). Logaritmoidulle tulokselle voisi kuvion mukaan ehkä sovittaa kaksi erillistä lineaarista regressiosuoraa. Läpimittaluokat voitaisiin jakaa kahteen osaan luokan 170 kohdalta, koska tuloksen hajonta näyttäisi olevan suurempi sitä suuremmilla läpimittaluokilla.

Tarkastellaan seuraavaksi (kuvio 4.9), miten kaikkien leimikkojen eli työmaiden yli summattujen läpimittaluokittaisten tulosten arvot ovat jakautuneet läpimittaluokittain.



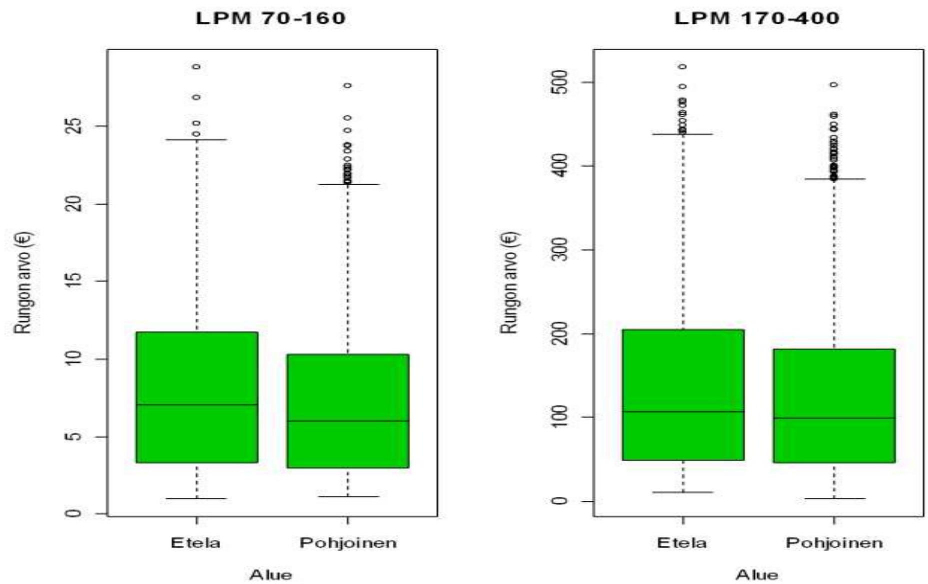

**Kuvio 4.9** Tuloskertymät läpimittaluokittain (mänty)

Suurimmat kokonaistulokset on saavutettu läpimittaluokissa 240–290, joiden molemmin puolin kokonaistulos pienenee siirryttäessä näistä luokista pois päin. Tämän tarkastelun perusteella voidaan myös rajata vähemmän tärkeitä läpimittaluokkia tarkempien analyysien ulkopuolelle, sillä kuviosta nähdään, ettei läpimittaluokilla 400:sta eteenpäin ole suurta merkitystä kokonaistuloksen kannalta.

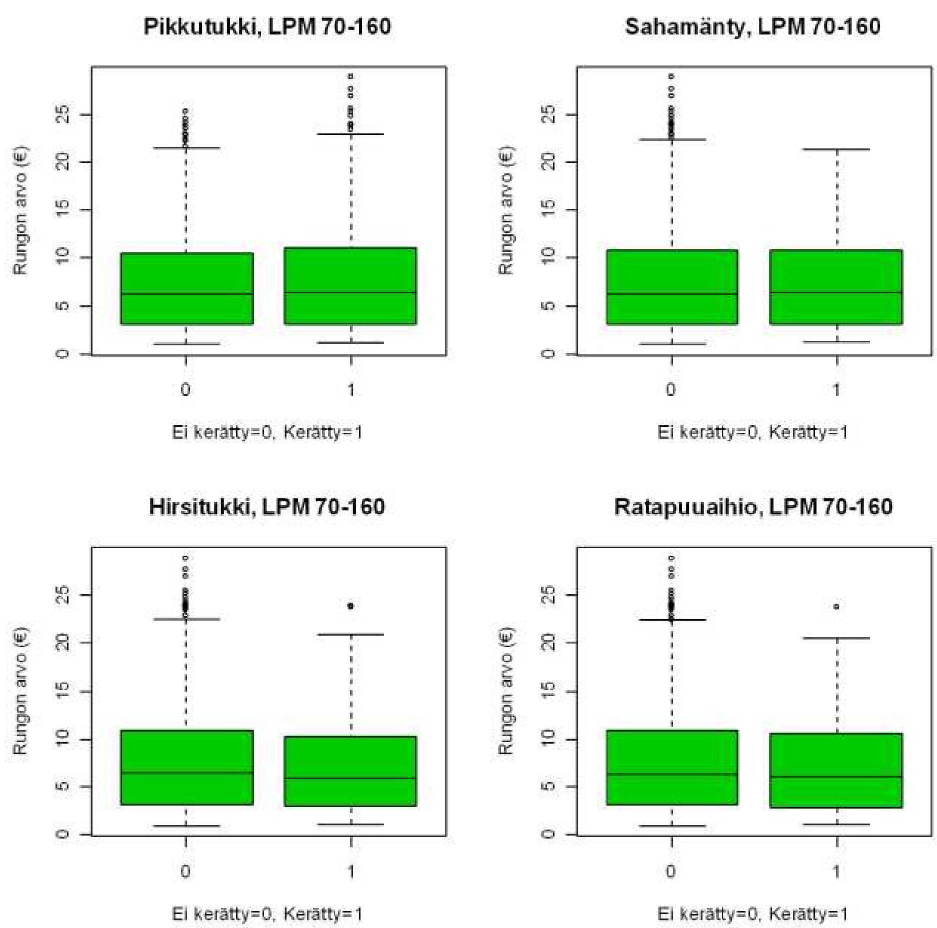
Tarkastelujen perusteella aineisto on syytä ehkä jakaa kahteen ryhmään: läpimittaluokat alle 170 ja yli 170 omiksi ryhmikseen. Perusteluja tälle jaolle löytyy myös muualta; rinnankorkeusläpimitaltaan noin 170 mm:n rungot ovat niitä, joista on jo mahdollista saada tukkitavaralajeja. Sen alapuolisissa läpimittaluokissa kuitupuu on pääasiallisesti saatava puutavaralaji. Jos koko aineistolle rakennettaisiin malli, tarkoittaisi se sitä, että merkittäviksi osoittautuvat muuttujat vaikuttaisivat tulokseen jokaisessa läpimittaluokassa. On kuitenkin perusteltua olettaa, että kuituluokissa tulokseen voivat vaikuttaa eri muuttujat kuin tukkiluokissa. Suurimman tutkittavan läpimittaluokan valitsemiseen ei ole täysin yksioikoista perustetta. Valinta osuu läpimittaluokan 400 kohdalle, koska siitä eteenpäin tuloksen hajonta kasvaa jokseenkin selkeämmin kuin aiempien läpimittaluokkien välillä ja kuten edellä tuli mainittua, näiden luokkien kontribuutio kokonaistulokseen ei ole kovin suuri. Myös aineiston jaolle vielä pienempiin osiin voi löytyä perusteita, mutta tässä tutkimuksessa päädytään tähän.

### 4.1.7 Läpimittaluokkien ryhmittelyn kuvailevaa tarkastelua

Tarkastellaan valittua läpimittaluokkien jakoa kahteen osaan tutkimalla muuttujien kuvaajia. Kokonaisaineistolle tehtyyn tarkasteluun verrattuna alueiden erot näkyvät osa-aineistoissa selkeämmin (kuvio 4.10). Tulosten mediaanit ovat etelässä hiukan suuremmat.



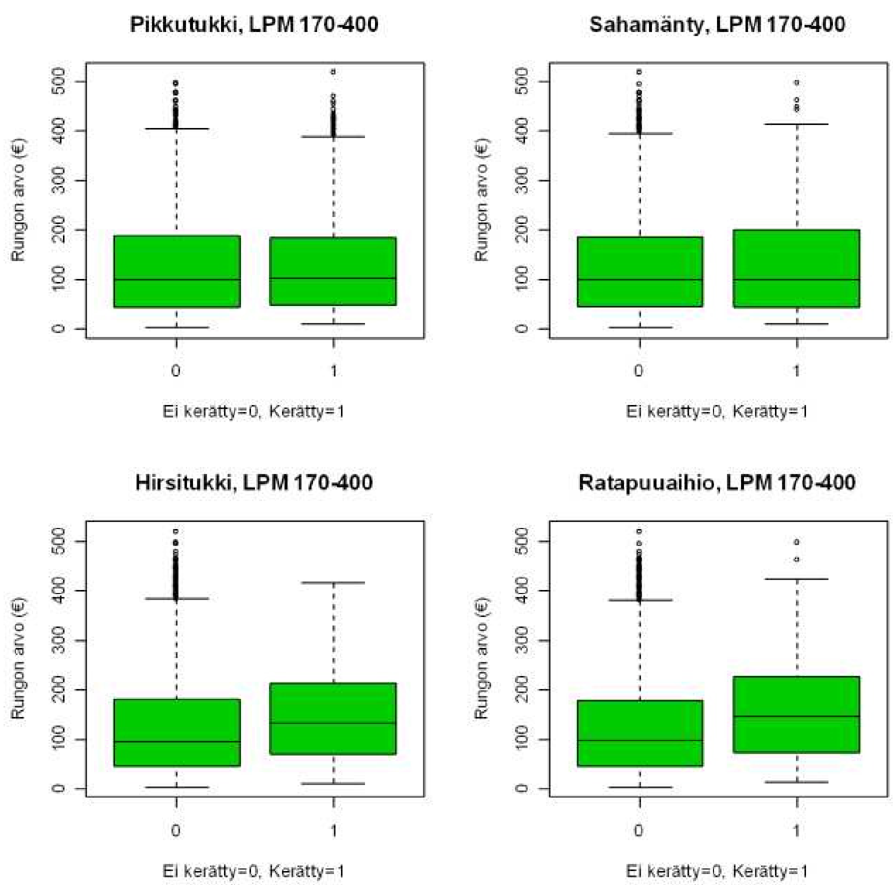

**Kuvio 4.10** Alueittaisten tulosten laatikko-jana –kuviot (mänty)




**Kuvio 4.11** Tavaralajimuuttujien laatikko-jana –kuviot (mänty)

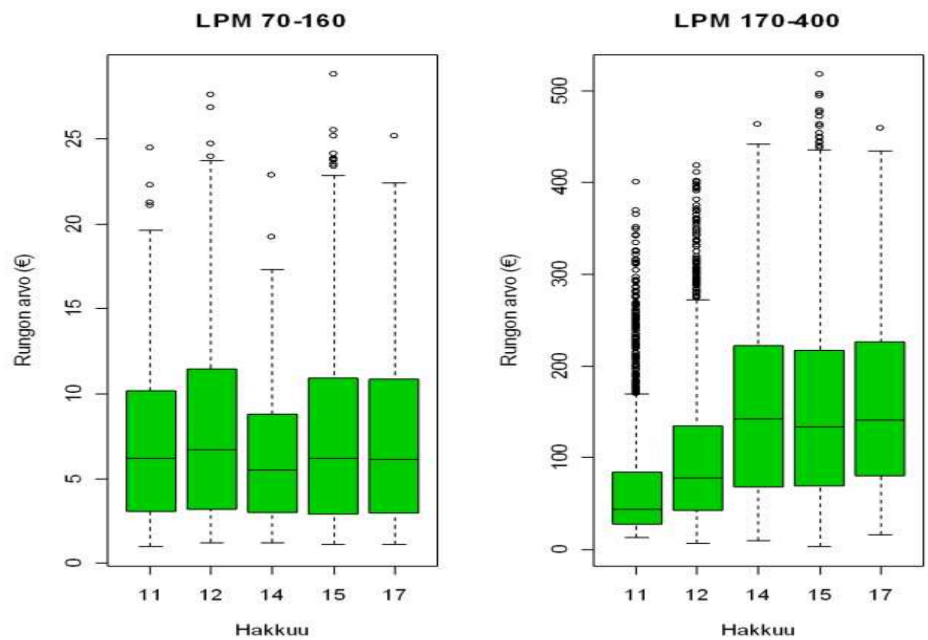
Tavaralajien suhteen kuituluokissa näyttäisi siltä, että ei juuri ole merkitystä, mitä tavaralajeja työmaalta kerättäisiin (kuvio 4.11). Tässä on syytä huomioida, että ratapuuaihion minimiläpimitta aineistossa on 275 mm. Kuituluokista ei siis ole mahdollista kerätä ratapuuaihioita, mutta se on otettu tarkasteluun mukaan siltä varalta, että esille tulisi mahdollisesti tämän tavaralajin valikoituminen tietynlaisille työmaille, koska ratapuuaihiomuuttuja on työmaakohtainen eikä läpimittaluokkakohtainen.

Tukkiluokissa (kuvio 4.12) on havaittavissa jonkinlaista tuloksen paremmuutta tapauksissa, joissa työmaalta on kerätty hirsitukkia tai ratapuuaihiota.




**Kuvio 4.12** Laatikko-jana –kuviot tavaralajimuuttujille (mänty)

Hakkuutavan tarkastelu kuitu- ja tukkijaolle ilmaisee, että hakkuutavoilla näyttäisi olevan selkeitä eroja tukkiluokissa (kuvio 4.13, oikeanpuoleinen kuvio).




**Kuvio 4.13** Hakkuutavoittaiset laatikko-jana –kuviot (mänty)

## 4.2 Kuusi

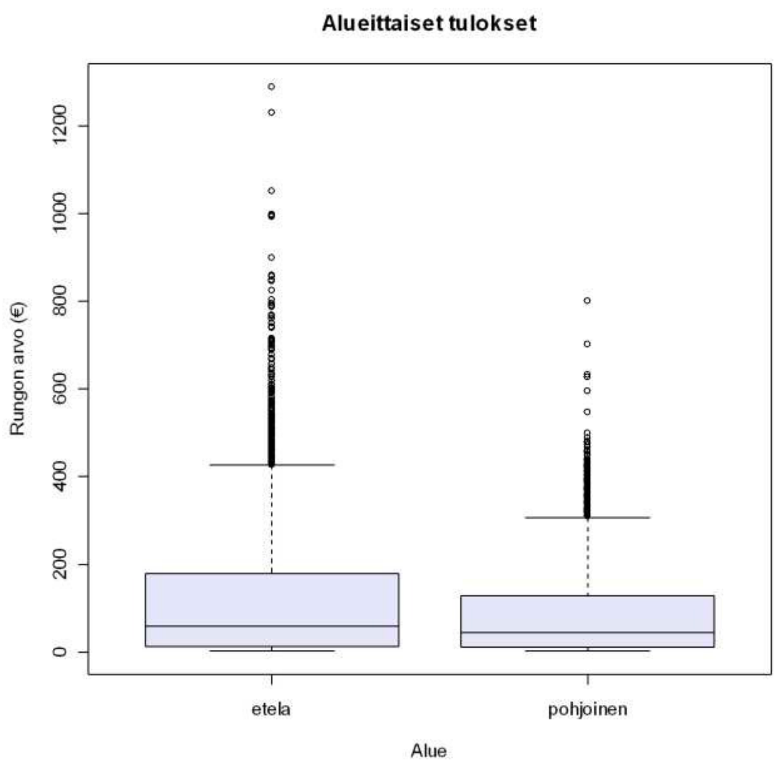
Tehdään seuraavaksi vastaavat tarkastelut kuuselle. Joitain muuttujien kuvailevia tarkasteluja on jätetty tarkemmin kommentoimatta, sillä kommentit olisivat hyvin samansuuntaisia kuin männyn tapauksessa. Asiasta kiinnostunut lukija voi kuvioita tutkimalla itse varmistua asiasta.

### 4.2.1 Tulosten jakaumia rinnankorkeusläpimittaluokittain

Läpimittaluokittaisessa tarkastelussa datasta tulee esille samankaltaisia piirteitä kuin männyn tapauksessa. Läpimittaluokissa 50 ja 60 on selkeästi virheitä, joten nämä luokat jätetään analyysin ulkopuolelle. Tarkastelujen perusteella poistettiin virheellisiksi katsottavia havaintoja aineiston eri läpimittaluokista ennen analyysejä.

### 4.2.2 Alueittainen tuloksen tarkastelu

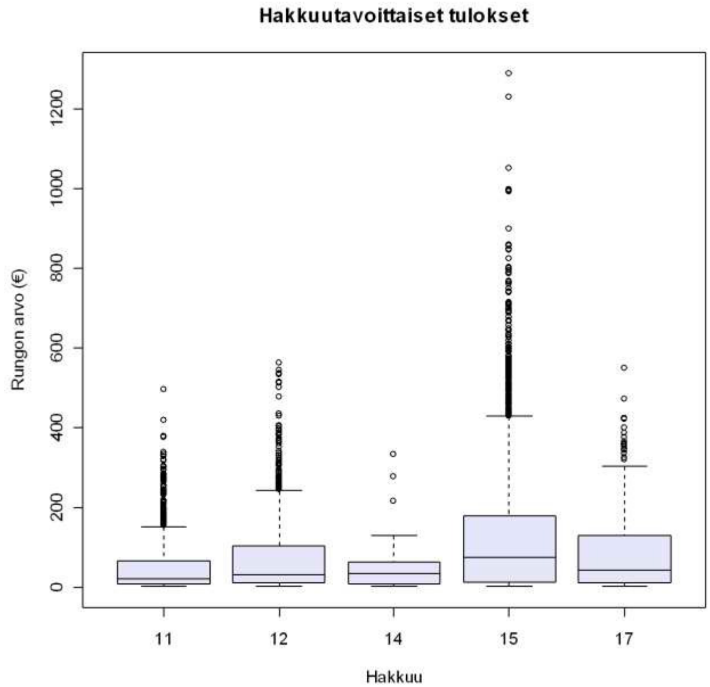
Tulosten jakaumat alueittain sisältävät jonkin verran eroja (kuvio 4.14). Etelässä tuloksen mediaani näyttää vähän suuremmalta kuin pohjoisen. Etelän tulosten hajonta on huomattavasti laajempi. Kuvion perusteella voidaan olettaa tuloksen olevan keskimäärin parempi etelässä kuin pohjoisessa.




**Kuvio 4.14** Tulosten jakaumat alueittain (kuusi)

### 4.2.3 Hakkuutavoittainen tulosten tarkastelu

Avohakkuulla näyttää olevan jonkin verran suurempi tuloksen mediaani kuin muilla hakkuutavoilla (kuvio 4.15).




**Kuvio 4.15** Tuloksien jakaumia hakkuutavoittain (kuusi)

### 4.2.4 Tavaralajeittaisia tuloksen tarkasteluja

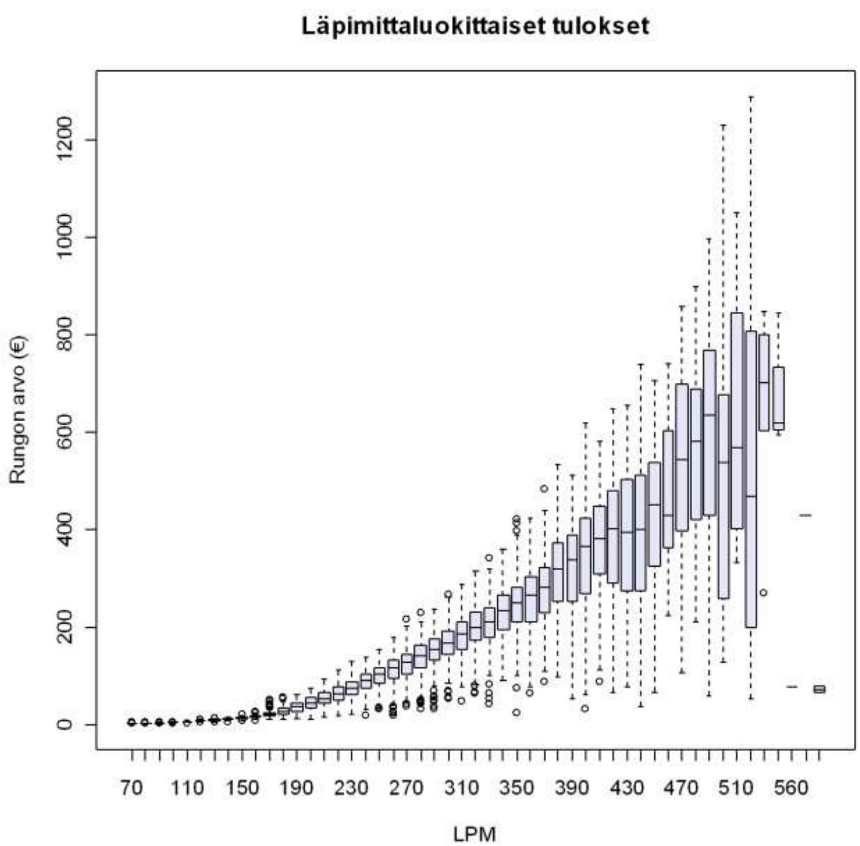
Tarkastellaan vielä, miten tulokset ovat jakautuneet eri läpimittaluokissa riippuen siitä, onko työmaalta kerätty kyseistä tavaralajia vai ei (liite A, kuviot A.8-A.10). Tulos näyttäisi olevan keskimäärin parempi niillä työmailla, joilta on kerätty erikoistukkia, ja huonompi niillä, joilta on kerätty pikkutukkia. Sellun talteenotto taas näyttäisi tuottavan tulokseltaan edullisemman jakauman. Laatikko-jana –kuvioista (liite A, kuvio A.11) on helppo tehdä sama johtopäätös edellisten kanssa: erikoistukin kerääminen vaikuttaa positiivisesti tulokseen.

### 4.2.5 Parametrien tarkastelu

Parametrien jakaumien tarkastelujen kautta saadaan myös aineiston virheitä karsittua. Kovin tarkkoihin tarkasteluihin ei kuusenkaan tapauksessa ole lähdetty, osittain siksi, että parametrien määrittelyiden takia niistä tehtyjen päätelmien painoarvo ei voi olla kovin suuri. Liitteen A kuviossa A.12 on kuitenkin kuvattu parametrien histogrammit, joista käy selville, miten parametrien arvot ovat jakautuneet.

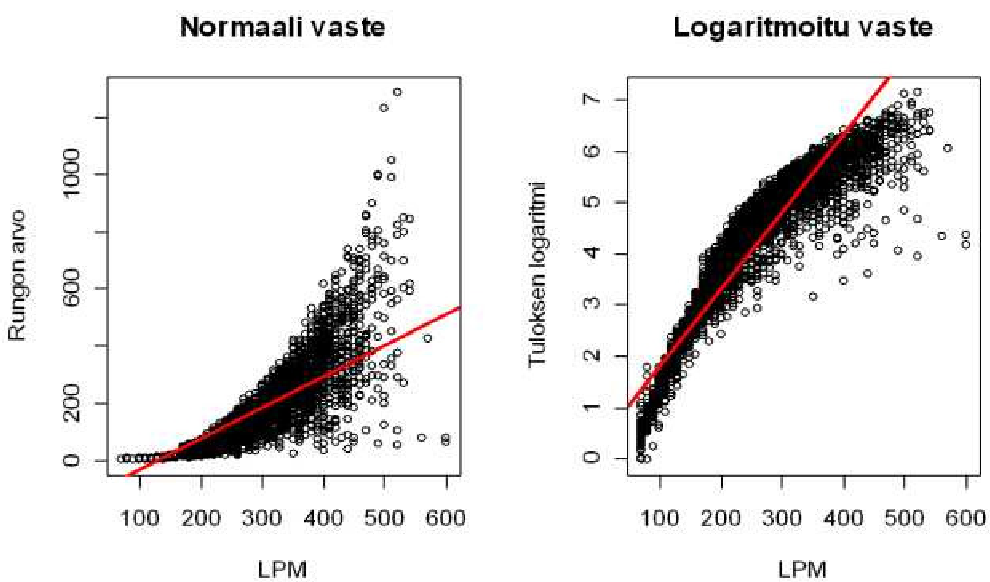
### 4.2.6 Tuloksen läpimittaluokittainen tarkastelu

Tuloksen läpimittaluokittainen mediaani ja hajonta kasvavat läpimittaluokan kasvaessa (kuvio 4.16). Verrattaessa sitä männyn vastaavaan kuvioon (kuvio 4.6), huomataan, että katkottujen kuusien läpimittaluokittaiset tulokset loppuvat aiemmin ja siksi kuvio näyttää kauniimmalta.



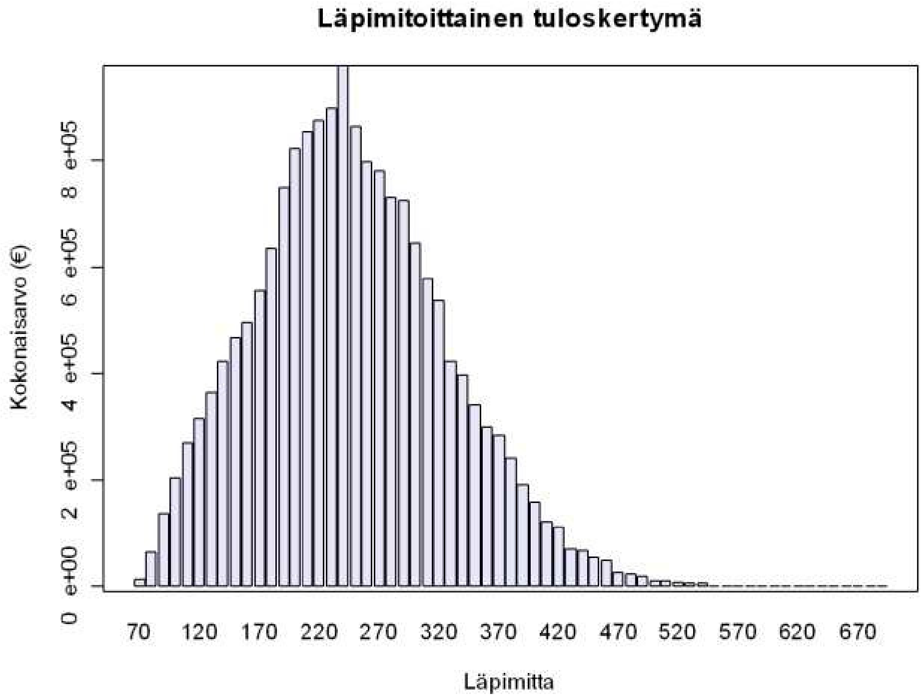

**Kuvio 4.16** Läpimittaluokittaiset tuloksien laatikko-jana – kuvaajat (kuusi)

Tutkittaessa hajontakuvioiden avulla rungon arvojen sekä näiden logaritmien jakautumista läpimittaluokittain (kuvio 4.17) ja lisäämällä näihin kuvioihin regressiosuorat, voimme katsoa, onko muuttujien välinen riippuvuus lineaarista. Logaritmointi näyttäisi sopivan tarkoituksiin paremmin, joskin silloinkin on syytä sopivasti ryhmitellä läpimittaluokkia.




**Kuvio 4.17** Tuloksen ja läpimittaluokan hajontakuvio (vasen) sekä siihen lisätty regressiosuora ja logaritmoidun tuloksen ja läpimittaluokan hajontakuvio (oikea) sekä siihen lisätty regressiosuora (kuusi)

Tarkasteltaessa läpimittaluokittaisia tuloskertymiä (kuvio 4.18) havaitaan, että jakauman huippu on läpimittaluokan 240 kohdalla. Karkeasti ottaen voidaan todeta, että suurimmat kokonaistulokset on saavutettu läpimittaluokissa 190–290. Kuten männynkin tapauksessa, voidaan tämän tarkastelun perusteella rajata vähemmän tärkeitä läpimittaluokkia tarkempien analyysien ulkopuolelle. Tarkastelujen perusteella voidaan esimerkiksi rajata tarkasteltavista läpimittaluokista pois läpimittaluokat 420:sta eteenpäin.

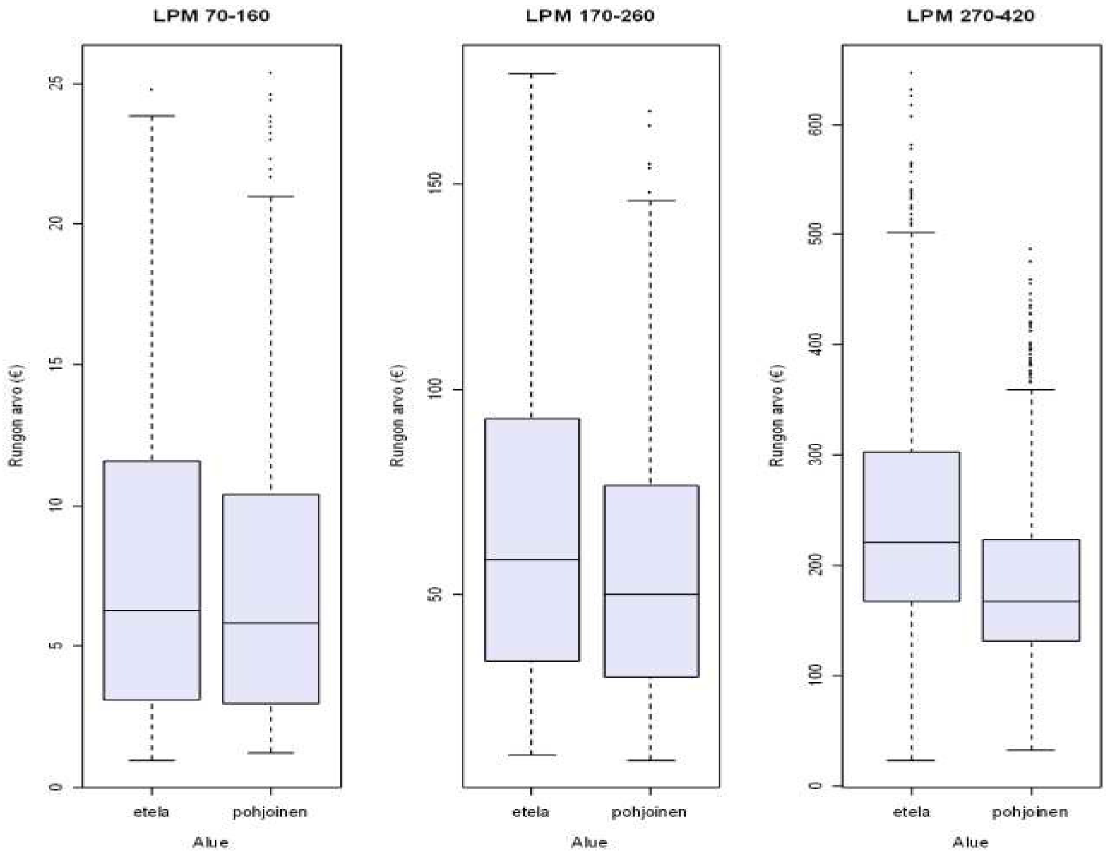



**Kuvio 4.18** Läpimittaluokittaiset tuloskertymät (kuusi)

### 4.2.7 Läpimittaluokkien ryhmittelyn kuvailevaa tarkastelua

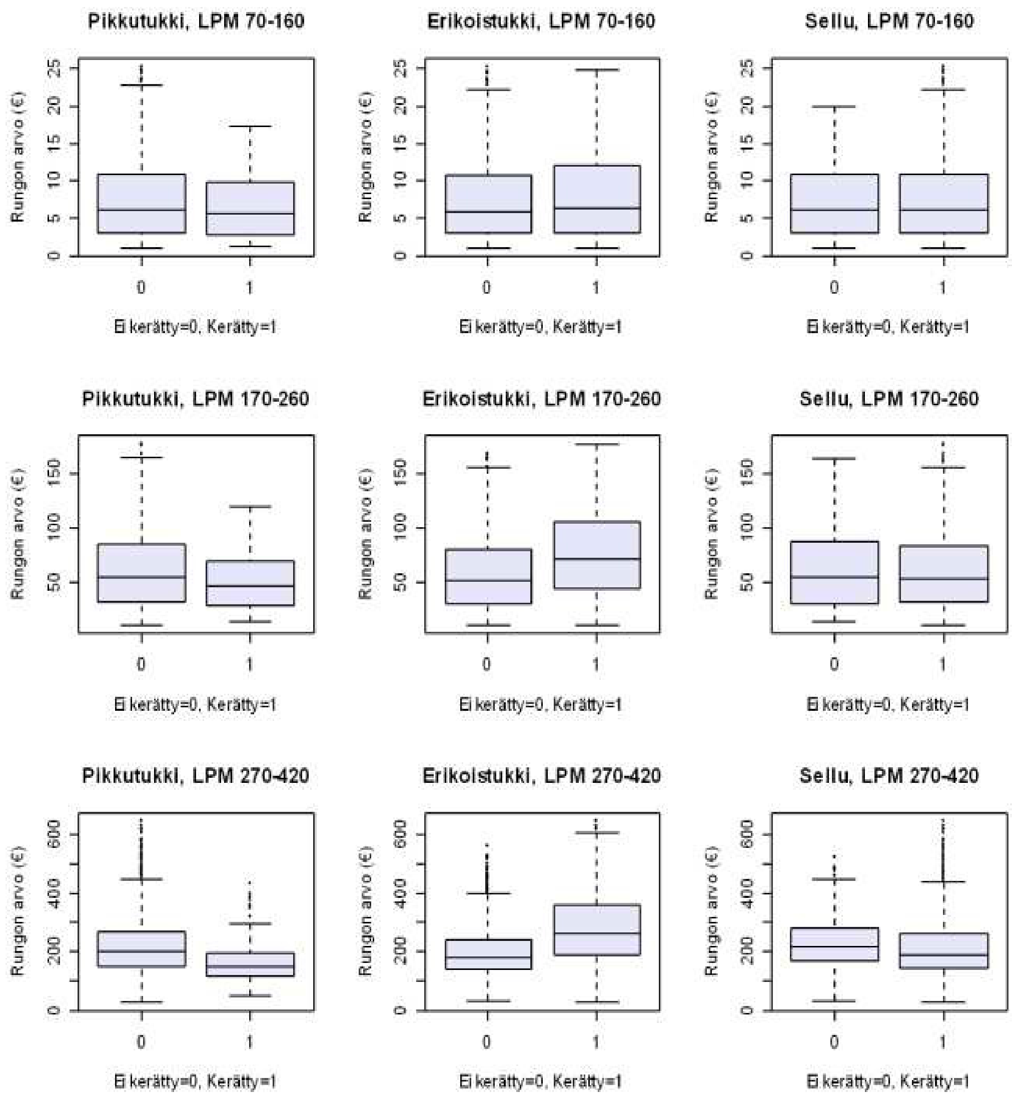
Edellä tehtyjen tarkastelujen perusteella kuusiaineisto jaetaan kolmeen eri läpimittaluokkien muodostamaan ryhmään. Ensimmäiseen ryhmään tulevat läpimittaluokat 70–160, toiseen 170–260 ja kolmanteen 270–420. Tarkastellaan seuraavaksi eri muuttujia tällä ryhmäjaolla.

Alueittaisten tulosten mediaanien erot etelän hyväksi kasvavat läpimittaluokan kasvaessa (kuvio 4.19). Varsinkin osituksessa 270–390 etelän tulos näyttää selkeästi paremmalta.



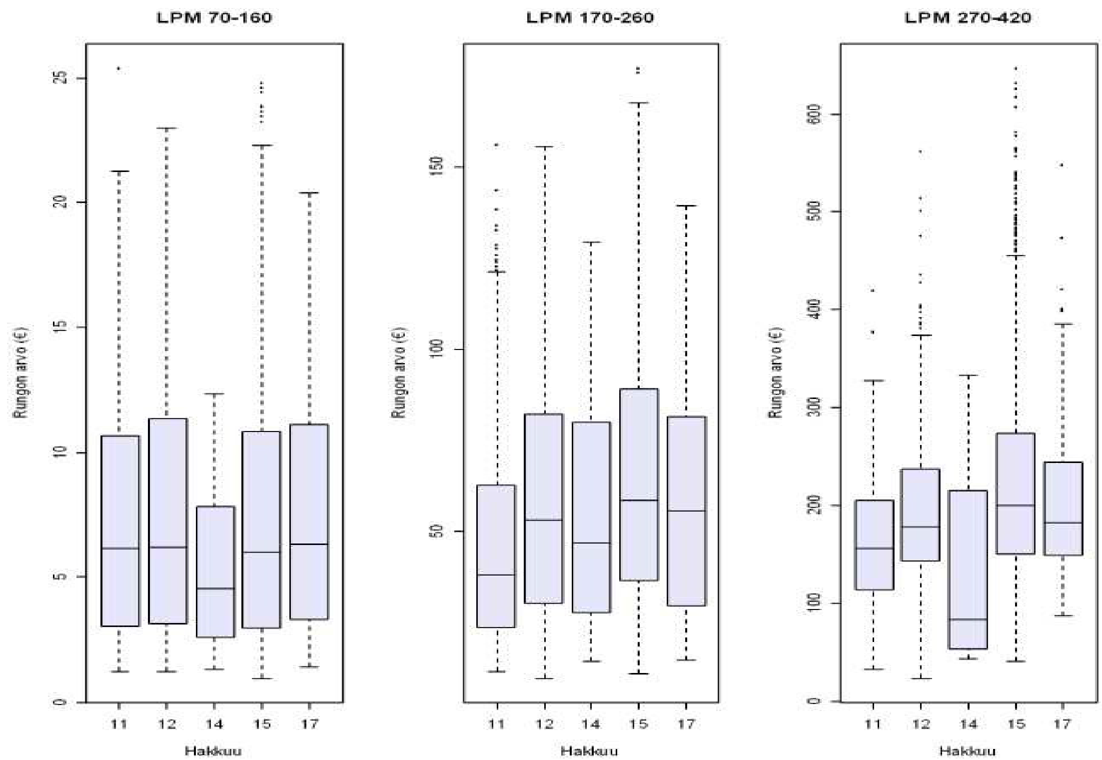

**Kuvio 4.19** Alueittaiset jakaumat eri läpimittajaoille (kuusi)

Tavaralajeittaisissa tarkasteluissa voidaan havaita joitain eroja (kuvio 4.20). Esimerkiksi työmailla, joilla on kerätty erikoistukkia, on läpimittaluokissa 170–260 ja 270–420 mediaaniltaan korkeampi tulos kuin työmailla, joilta sitä ei ole kerätty. Näyttää mielenkiintoiselta, että niiltä työmailta, joilta on kerätty pikkutukkia, on saatu heikompi tulos kuin työmailta, joilta sitä ei ole kerätty. Tässä voi olla kysymys siitä, että pikkutukkia kerättäisiin vain leimikoilta, joilla runkojakauma painottuu ohuempirunkoisiin kuusiin.




**Kuvio 4.20** Tavaralajitarkasteluja eri läpimittajaoille (kuusi)

Tarkasteltaessa tulosten jakaumia eri hakkuutavoilla (kuvio 4.21) huomataan, että alimmassa läpimittaluokkaosituksessa eroja ei juuri ole, mutta kahdessa muussa osituksessa hakkuutapa 11 näyttäisi olevan tulokseltaan jonkin verran heikompi kuin hakkuutavat 12, 15 ja 17. Hakkuutapa 14 eroaa myös selkeästi muista, mutta havaintoja tältä hakkuutavalta on vähän, joten ero voi johtua havaintojen lukumäärästä.



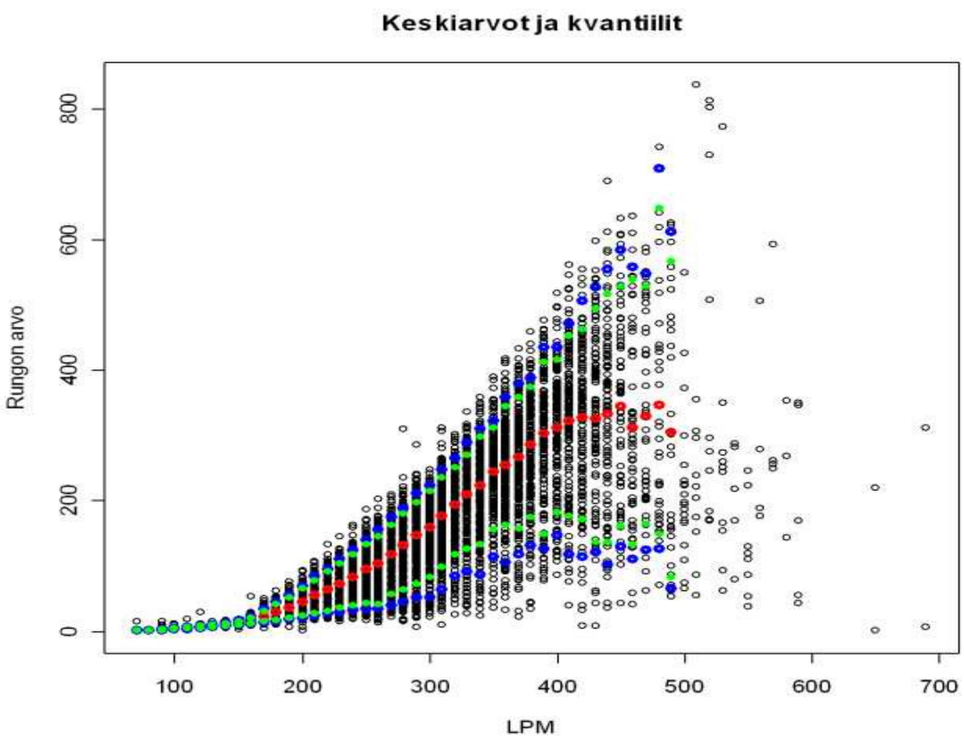

**Kuvio 4.21** Tuloksien jakaumat hakkuutavoittain eri läpimittaluokkajaoilla (kuusi)

Johtopäätöksenä edellisistä tarkasteluista voidaan todeta, että läpimittaluokkien jako näyttäisi tuottavan selkeämpiä eroja tarkasteltavien muuttujien välille. Tämä on seurausta onnistuneesta aineiston osittamisesta, jonka tavoitteena on ryhmien sisäisen varianssin pienentäminen ja ryhmien välisen varianssin kasvattaminen.

## 4.3 Läpimittaluokittaisten jakaumien käyttö tuloksien vertailuissa

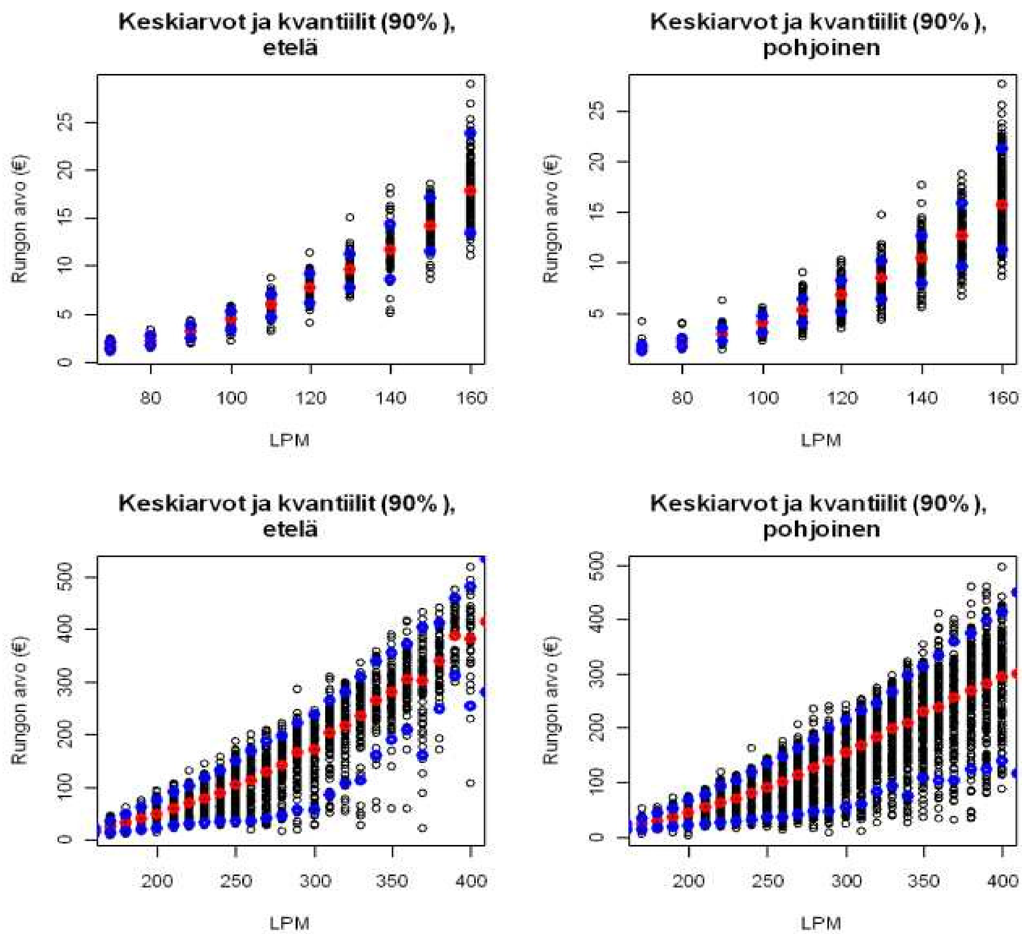
### 4.3.1 Mänty

Tutkimuksessa ei ole mahdollista käyttää optimaalisia tuloksia työmaittaisten tulosten vertailuissa. Jotta päästäisiin edes lähelle tällaista asetelmaa, täytyisi käyttää STM-tiedostojen runkokäyriä ja simuloida katkontatulokset relevanteilla parametrien arvoilla. Näistä lähtökohdista johtuen vertailu täytyy toteuttaa perustuen toteutuneiden tulosten jakaumien vertailuun. Yksinkertaisin tapa toteuttaa tämä on laskea läpimitta-luokittaisille tuloksille keskiarvot ja sopivat kvantiilit. Tällöin vältetään myös aineiston muuntamiset, mikä helpottaa tulosten tulkintaa. Tutkimusaineiston kanssa samankaltaisista lähtökohdista saatavia uusia tuloksia verrataan laskettuihin tunnuksiin.




**Kuvio 4.22** Keskiarvot ja kvantiilit (80 % ja 90 %) kokonaisaineistolle (mänty)

Kuviossa 4.22 on tuloksen ja läpimittaluokan hajontakuvio ja siihen on lisätty läpimittaluokittaiset keskiarvot sekä 80 %:n ja 90 %:n kvantiilit. Jotta tällaiset vertailut saataisiin tarkemmiksi, tarkastellaan, miltä näyttäisi, jos aineiston kokonaisvaihtelua saataisiin pienennettyä ryhmittelemällä aineistoa. Ryhmitellään aineisto alueiden mukaan ja tarkastellaan erikseen kuitu- ja tukkiluokkia (kuvio 4.23).




**Kuvio 4.23** Alueittaisia tunnuksia jaetulle aineistolle (mänty)

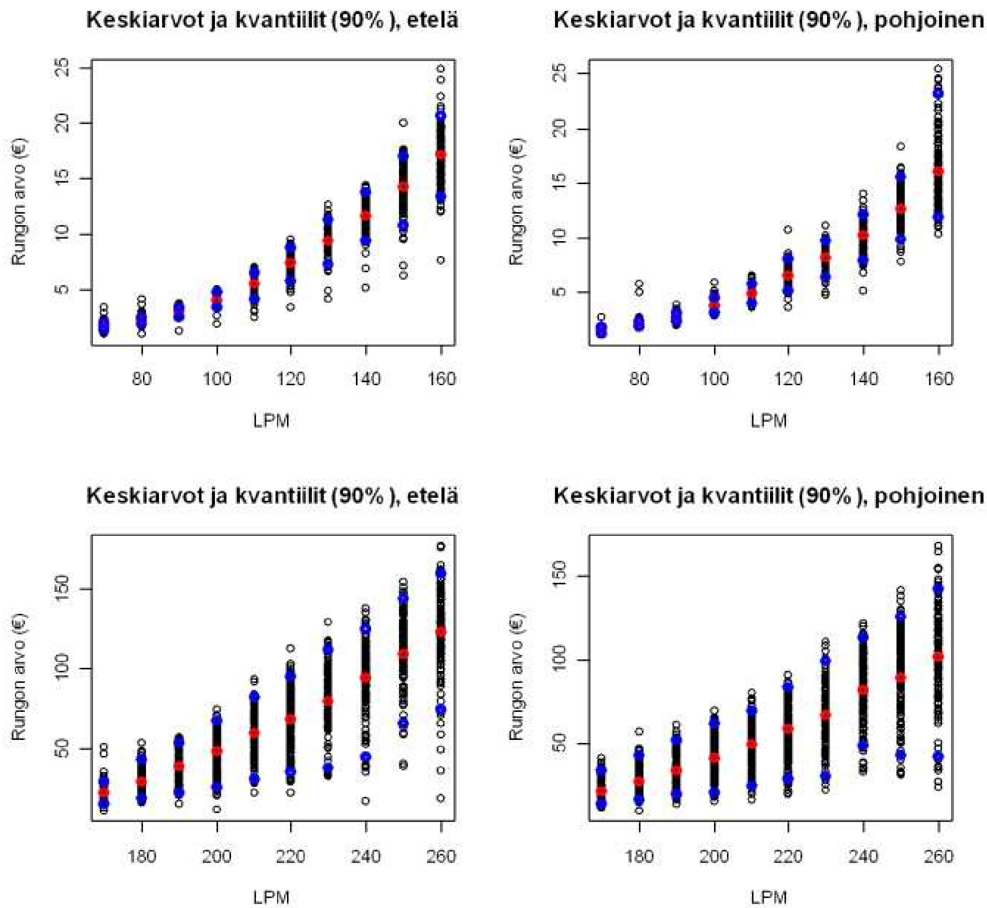
Silmämääräisesti arvioiden aluejako näyttäisi tuottavan tarkemman vertailupohjan kuin jos jakoa ei tehtäisi. Lasketaan alueille vielä läpimittaluokittaisten tulosten keskiarvot ja niille erotukset (taulukko 4.1). Tulosten keskiarvojen erotus etelän eduksi kasvaa läpimittaluokan kasvaessa. Taulukoitujen tulosten etuna on se, että tulosten erot ovat selkeämmin nähtävissä.

**Taulukko 4.1** Läpimittaluokittaisten tulosten keskiarvot alueittain sekä niiden erotus (mänty)

| LPM | Etelä | Pohjoinen | Erotus |
|---|---|---|---|
| **70** | 1,59 | 1,49 | 0,10 |
| **80** | 2,21 | 2,03 | 0,17 |
| **90** | 3,15 | 2,90 | 0,26 |
| **100** | 4,46 | 3,98 | 0,48 |
| **110** | 5,97 | 5,29 | 0,69 |
| **120** | 7,75 | 6,75 | 1,00 |
| **130** | 9,63 | 8,42 | 1,21 |
| **140** | 11,70 | 10,45 | 1,25 |
| **150** | 14,21 | 12,65 | 1,56 |
| **160** | 17,78 | 15,73 | 2,04 |
| **170** | 24,47 | 21,86 | 2,61 |
| **180** | 32,12 | 28,40 | 3,72 |
| **190** | 40,18 | 35,51 | 4,68 |
| **200** | 49,69 | 43,33 | 6,36 |
| **210** | 60,31 | 52,34 | 7,98 |
| **220** | 70,10 | 61,25 | 8,84 |
| **230** | 79,58 | 69,90 | 9,68 |
| **240** | 90,10 | 80,13 | 9,97 |
| **250** | 104,58 | 91,01 | 13,57 |
| **260** | 112,81 | 99,53 | 13,28 |
| **270** | 128,34 | 113,98 | 14,36 |
| **280** | 142,58 | 127,33 | 15,26 |
| **290** | 165,92 | 138,94 | 26,98 |
| **300** | 172,92 | 154,06 | 18,85 |
| **310** | 202,81 | 167,47 | 35,35 |
| **320** | 217,27 | 184,14 | 33,13 |
| **330** | 236,88 | 198,28 | 38,60 |
| **340** | 265,24 | 210,29 | 54,96 |
| **350** | 280,54 | 230,08 | 50,46 |
| **360** | 304,13 | 238,30 | 65,83 |
| **370** | 303,28 | 254,49 | 48,79 |
| **380** | 339,54 | 269,76 | 69,78 |
| **390** | 387,88 | 280,66 | 107,22 |
| **400** | 382,44 | 294,73 | 87,71 |
| **410** | 413,61 | 300,96 | 112,65 |
| **420** | 411,81 | 309,13 | 102,68 |
| **430** | 431,40 | 309,05 | 122,35 |
| **440** | 502,11 | 305,80 | 196,31 |
| **450** | 545,74 | 315,05 | 230,69 |
| **460** | 334,79 | 309,06 | 25,72 |
| **470** | 438,02 | 306,62 | 131,40 |
| **480** | 544,98 | 323,56 | 221,42 |
| **490** | 610,05 | 282,72 | 327,33 |

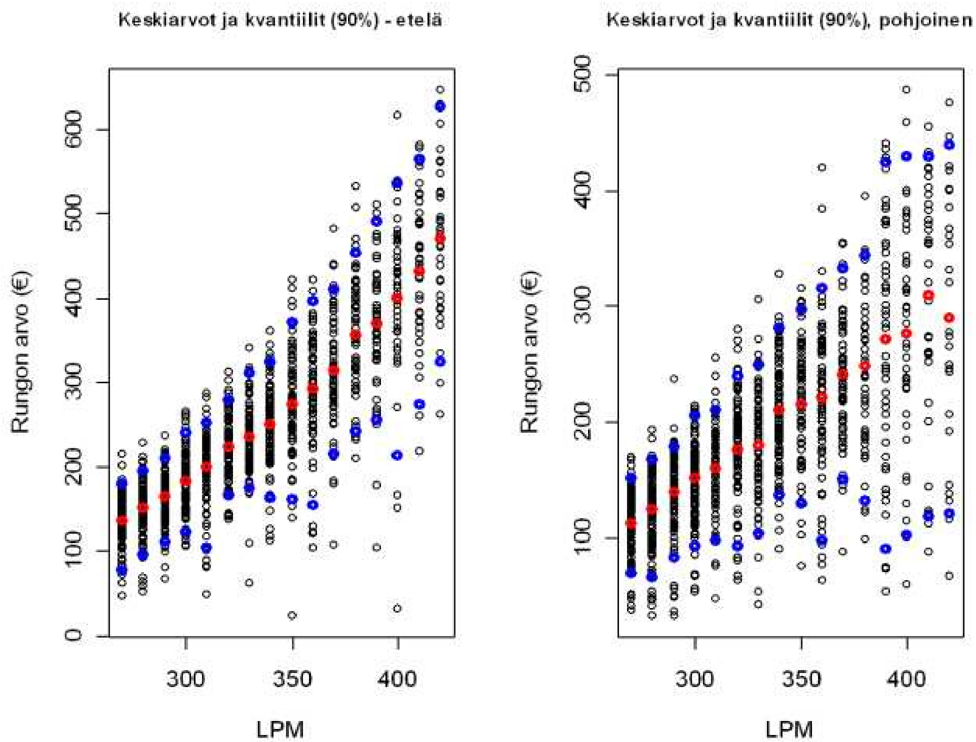
### 4.3.2 Kuusi

Kuten männynkin tapauksessa, tarkastellaan keskiarvojen ja kvantiilien hyväksikäyttöä tulosten vertailun välineenä eri alueiden välisessä tarkastelussa. Tässä ei ole lainkaan suoritettu tarkastelua kokonaisaineistolle, vaan on suoraan siirrytty ryhmäjakojen tarkasteluun (kuvio 4.24). Silmämääräisesti kuvioista on vaikea havaita mahdollisia eroja. Tarkkasilmäinen voi kuitenkin havaita, että etelän tulokset ovat yleisesti ottaen parempia.




**Kuvio 4.24** Tuloksien keskiarvot ja 90 %:n kvantiilit eri läpimittaluokkajaoille (70–160 ja 170–260) alueittain (kuusi)

Kuviosta 4.25 havaitaan selkeät erot etelän ja pohjoisen välillä. Kuviossa on syytä huomioida, että hajontakuvioiden pystyakseleilla on eri asteikko. Etelän tulos on silmämääräisesti arvioiden parempi kaikissa läpimittaluokissa ja erot suurenevat etelän hyväksi läpimittaluokan kasvaessa. Kuvioista nähdään myös, että pohjoisten alueiden tuloksissa on suurempi hajonta.




**Kuvio 4.25** Tuloksien keskiarvot ja 90 %:n kvantiillit eri läpimittaluokkajaolle 270–390 alueittain (kuusi)

## 4.4 Kuvailevan osan johtopäätökset

Sekä männyn että kuusen aineistossa on selkeitä virheitä, ei tosin mitään merkittäviä määriä, mutta ne on poistettava ennen varsinaisten analyysien suorittamista. Joitain läpimittaluokkia on syytä jättää kokonaan analyysien ulkopuolella, koska niiden vaikutus kokonaistuloksen kannalta on pieni, ne saattavat sisältää paljon virheitä tai niissä on vain muutamia havaintoja. Läpimittaluokittaisten tulosten jakaumien tutkimisen avulla löydetyt selvästi poikkeavat havainnot jätetään myös analyysien ulkopuolelle, sillä niiden läsnäolo vääristäisi tuloksia. Aineistosta poistetaan luonnollisesti myös havainnot joilla tilavuuskertymiä ei ollut.

Läpimittaluokat on syytä männyn tapauksessa jakaa kahteen osaan. Ensimmäiseen osaan tulevat läpimittaluokat 70–160 ja toiseen läpimittaluokat 170–400. Läpimittaluokat 50 ja 60 on pudotettu pois, koska niistä löytyi paljon virheitä tai epämääräisyyksiä. Näiden läpimittaluokkien vaikutus kokonaistulokseen on joka tapauksessa hyvin pieni. Kuusen tapauksessa läpimittaluokat jaetaan kolmeen ryhmään. Ensimmäiseen ryhmään tulevat läpimittaluokat 70–160, toiseen 170–260 ja kolmanteen 270–420. Molemmissa tapauksissa ensimmäinen ryhmä muodostaa ns. kuituluokkien ryhmän. Tällä tarkoitetaan, että näistä läpimittaluokista saadaan pääosin kuitupuuta, jonka hinta on huomattavasti tukkia alhaisempi. Loput läpimittaluokkaryhmät ovat ns. tukkiluokkien ryhmiä. Näillä jaoilla saadaan aineistoa ryhmiteltyä siten, että saadut analyysitulokset ovat tarkempia.

Läpimittaluokkien ryhmien tarkastelun perusteella voidaan todeta, että tukkiluokissa hakkuutavalla näyttäisi olevan vaikutusta tulokseen. Etelässä suoritettujen hakkuiden tulos näyttäisi olevan pohjoista parempi. Tämä näkyy erityisesti kuusen tapauksessa. Eri puutavaralajien keräämisen vaikutuksena esille nousi männyn tapauksessa tulosta parantava vaikutus, kun kerätään hirsitukkeja ja ratapuuaihioita. Luonnollisesti niitä voidaan kerätä vain männyn tukkiluokissa. Kuusen ollessa kysymyksessä pikkutukin kerääminen tukkiluokissa näyttäisi vähentävän tulosta, kun taas erikoistukin kerääminen näyttäisi nostavan tulosta. Sellun kohdalla tulokset eivät ole yhtä selkeitä. Näyttäisi kuitenkin, että sellun kerääminen läpimittaluokissa 260–420 vähentää tulosta.

Hakkuutapaa 13 on käytetty kahdella työmaalla ja hakkuutapaa 18 yhdellä. On syytä poistaa hakkuutapojen 13 ja 18 aineisto, koska havaintoja on läpimittaluokittain hyvin vähän. Ne voisivat tällöin vaikuttaa hakkuutavan tulkintoihin. Jäljelle jäävät hakkuutavat muutetaan faktoreiksi, jotta käytettävät tilasto-ohjelmat ymmärtävät käsitellä niitä luokkamuuttujina. Hakkuutavoista ensiharvennus eli hakkuutapa 11 näyttäisi tuottavan mediaaniltaan heikomman tuloksen kuin muut hakkuutavat. Lisäksi pohjoisessa suoritettujen hakkuiden tuloksien jakauma on laajempi kuin etelässä suoritettujen. Tämä tarkoittaa, että tulosten erot etelässä ovat pienemmät kuin pohjoisessa.

Lopuksi on syytä vielä painottaa, että kuvaileva osuus on vain suuntaa-antava. Vasta regressioanalyysien perusteella voimme todeta, selittävätkö aineiston muuttujat tilastollisesti merkittävästi vastetta ja minkä suuntaisia mahdolliset vaikutukset ovat. Kuvailevan osuuden perusteella voimme siis jakaa tutkimusaineiston eri läpimittaluokkien muodostamiin ryhmiin ja siirtyä seuraavaksi suorittamaan regressioanalyysit.

# 5 Regressioanalyysit

Aiemmissa luvuissa on esitelty tutkimuksen lähtökohtia ja alan termistöä sekä tutkittu aineiston muuttujia. Näiden tietojen pohjalta on helpompi ymmärtää tässä luvussa suoritettujen regressioanalyysien tavoitteita ja tuloksia. Aluksi kerrotaan, kuinka regressiomallien rakentaminen tapahtuu ja muodostetaan ensimmäinen malli. Ensimmäisten mallien yhteydessä on yritetty yksityiskohtaisesti selittää mallien rakentamista ja seuraavien mallien kohdalla nämä selitykset on toiston välttämiseksi sivuutettu. Selkeyden vuoksi sekä männyn että kuusen regressiomallien rakentamisen jälkeen on tulokset vielä koottu yhteen erillisiin alalukuihin (alaluvut 5.1.4 ja 5.2.4).

Regressiomallien rakentamisen apuna on käytetty kuvailevasta osasta saatuja tietoja. Ensin rakennetaan osa-aineistoille yksinkertaiset regressiomallit. Sitten suoritetaan analyysit malleille, joissa ovat aluksi kaikki selittäjät mukana. Malleista pudotetaan muuttujia mallin valinnan ohjeistamana. Regressiomallien valinta suoritetaan Akaiken informaatiokriteeriin (*AIC*) perustuen, käyttäen R-ohjelmiston step-funktiota. Koska kysymyksessä on tutkimus, jossa on tarkoitus saada yksinkertainen ja käytäntöön sopiva malli, lopullisten mallien valintaan vaikuttavat AIC:n antaman lopputuloksen lisäksi myös mallien käytön ja tulkinnan vaatimukset. Käytännössä tämä tarkoittaa, että kaikki tilastollisesti merkitsevät muuttujat eivät välttämättä päädy lopulliseen malliin.

Koska kyseessä on empiirinen tutkimus, on tutkimukselle olennaiset teoreettiset tiedot esitelty pääosin liiteosassa. Tällöin teoria ei pääse katkaisemaan tutkimuksen johdonmukaista etenemistä. Täten niissä kohdissa, missä on nähty tarpeelliseksi selventää asiaa tarkemmin, on viitattu asiaan kuuluvaan liitteeseen.

## 5.1 Regressioanalyysit männylle

### 5.1.1 Yksinkertaiset mallit

Regressiomalleja rakennettaessa lähestymistapoja on monia. Tässä tutkimuksessa on ensin tutkittu aineiston muuttujia kuvailevin keinoin. Tällaisen pohjatiedon keruun hyötynä on, että lopullisia malleja rakennettaessa ja valittaessa voidaan käyttää kertynyttä omaa näkemystä eikä malleja tarvitse valita täysin sokeasti.

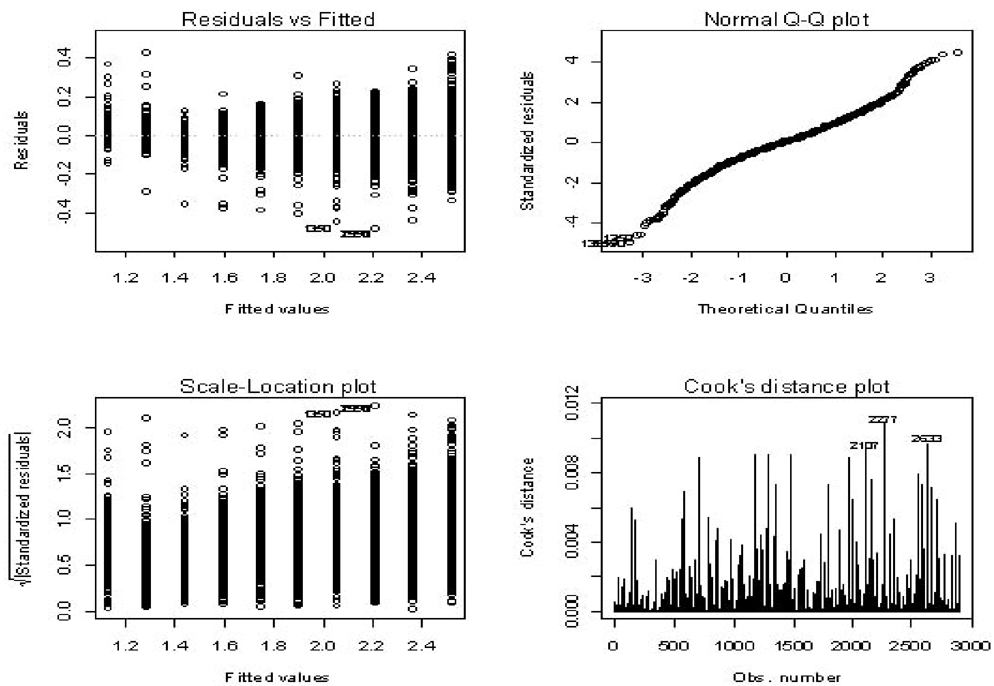
Tarkoituksena on rakentaa mahdollisimman yksinkertaiset regressiomallit. Tämä tarkoittaa, että malleihin otetaan mukaan vain selkeimmät selittäjät. Aineistolle tehdään tarvittaessa sopivia muunnoksia tai käytetään muita kuin pienimmän neliösumman menetelmää analyysien suorittamiseksi. Vain lopullisille malleille on tarpeen muodostaa ennustevälit sekä suorittaa mallien ja kertoimien tulkinnat, joten niitä ei mallien rakennusvaiheessa juuri esitellä.

Tarkasteluissa kokonaisaineisto sivuutetaan ja keskitytään vain osa-aineistojen analyyseihin. Läpimittaluokat jaetaan siis kahteen ryhmään: läpimittaluokat alle 170 ja läpimittaluokat 170 ja sitä suuremmat. Ryhmäjaolle on jo aiemmin esitetty kriteerit ja perustelut. Tarkastellaan aluksi läpimittaluokkia alle 170 mm.

### *5.1.1.1 Läpimittaluokat 70–160*

Ensimmäiseen malliin valitaan selittäjäksi ainoastaan LPM ja sillä selitetään tulosta. Kyseessä on siis yhden selittäjän regressiomalli. Malli on muotoa $y = \beta_0 + \beta_1 * LPM_i + \varepsilon,$ missä $y$ on tulos, $\beta_0$ ja $\beta_1$ ovat mallin estimoitavia parametreja, LPM on läpimittaluokka ja $\varepsilon$ on mallin virhetermi. Yhden selittäjän regressiomallia on tarkasteltu teoreettisesti tarkemmin liitteessä B.1.

Regressioanalyysin oletusten (liite B.2) tarkistaminen kertoo virhetermin heteroskedastisuudesta (kuvion 5.2 vasen alanurkka), koska sovitettujen arvojen kasvaessa kasvavat myös standardoidut residuaalit. Kuvio 5.2 sisältää myös muita oletusten tarkistamisessa hyödyllisiä kuvioita, mutta niitä ei tässä erikseen esitellä. Korjataksemme heteroskedastisuusongelman voimme joko suorittaa analyysin pienimmän neliösumman menetelmän (*PNS* tai *Ordinary Least Squares, OLS*) (liite B.3) sijasta painotetun neliösumman menetelmällä (*Weighted Least Squares, WLS*) (liite B.4) tai kokeilla datalle erilaisia muunnoksia (liite B.5). Aika usein käytetty muunnos datalle on logaritmointi. Tässä on päädytty logaritmoimaan tulos. Tällöin malli on muotoa $\log(y) = \beta_0 + \beta_1 * LPM_i + \varepsilon.$




**Kuvio 5.1** Mänty LPM 70–160

Aineistolle sovitetaan mallin mukainen regressiosuora ja testataan pitääkö paikkansa hypoteesi, jonka mukaan kaikki mallin kertoimet nollia (liite B.6, hypoteesien testaaminen). Jos mallin kaikki kertoimet eivät ole nollia, tarkastellaan mallin tuottamien p-arvojen avulla yksittäisten muuttujien merkitsevyys. Suoritetaan analyysit sekä painotetun neliösumman menetelmällä että logaritmoimalla tulos. Painotetun pienimmän neliösumman menetelmän painoiksi saatiin käänteiset läpimittaluokan kuudennet potenssit. Tulosten perusteella olemme molemmissa tapauksissa päässeet eroon sekä normaalisuus- että heteroskedastisuusongelmasta. Logaritmointi tuottaa tässä tapauksessa n. 7 %:a paremman selitysasteen (liite B.7, mallin sopivuus), joten lopulliseksi malliksi valitaan malli (5.1), jossa logaritmoitua tulosta selitetään läpimittaluokalla.

(5.1) $$\log(\hat{y}) = -1.281 + 0.02604 * LPM$$

### *5.1.1.2 Läpimittaluokat 170–400*

Suoritetaan analyysit seuraavaksi toiselle läpimittaluokkien ryhmälle, johon kuuluvat luokat 170–400. Otetaan aluksi mukaan selittäjiksi vain muuttujat LPM ja Kpituus.

Selitettäessä tulosta jäännöstermit ovat jälleen heteroskedastisia. Kokeillaan sekä painotetun neliösumman menetelmää että vasteen logaritmointia. WLS-mallissa on käytetty painoina samoja painoja kuin alle 170 läpimittaluokkien ryhmän tapauksessa. Tälläkin kertaa ne löydettiin kokeilemalla. Molempien menetelmien tuottamat tulokset ovat hyvin samansuuntaisia.

Kokeillaan vielä ottaa mukaan toisen asteen termi, koska jäännöstermien ja sovitettujen arvojen kuvaajassa näkyy hiukan käyryyttä. Toisen asteen termin ottaminen mukaan logaritmoituun malliin lisää selitysastetta 3.5 %:a, mutta virhetermi näyttää tällöin jokseenkin heteroskedastiselta. Päädytään kaavan (5.2) malliin.

$$\log(\hat{y}) = 0.8747 + 0.01055 * LPM + 0.07303 * Kpituus \qquad (5.2)$$

## 5.1.2 Usean selittäjän mallit

### *5.1.2.1 Läpimittaluokat 70–160*

Seuraavaksi lisätään malleihin selittäjiä. Ensimmäiselle osa-aineistolle tehtävään malliin otetaan mukaan seuraavat muuttujat (ks. muuttujaluettelo s. 18): LPM, Alue, DBAM, Mosuus, Kpituus, M1K, MinlpmK, MLosuus, PituuslktK, PT, SM, HT, ja FH. Selitettävänä on logaritmoitu tulos. Tällöin malli on muotoa

$$y = \beta 0 + \beta 1 * LPM + \beta 2 * Alue + \beta 3 * DBAM + \beta 4 * Mosuus + \beta 5 * Kpituus + \beta 6 * M1K$$
$$+ \beta 7 * MinlpmK + \beta 8 * MLosuus + \beta 9 * PituuslktK + \beta 10 * PT + \beta 11 * SM + \beta 12 * HT$$
$$+ \beta 13 * FH12 + \beta 14 * FH14 + \beta 15 * FH15 + \beta 16 * FH17 + \varepsilon,$$

missä y on tulos, $\beta_i$:t ovat ennalta tuntemattomia regressiomallin parametreja ja $\varepsilon$ on tuntematon virhetermi. Katso usean selittäjän regressiomalli (liite B.8).

Regressiomallin valinta voidaan suorittaa monella eri tavalla. Koska tässä tutkimuksessa mallin valintamenetelmäksi on valittu Akaiken informaatiokriteeri (*AIC*), käsitellään tässä vain sitä. AIC on hierarkkisten mallien vertailuun tarkoitettu kriteeri, joka pohjautuu uskottavuusfunktioon ja estimoitujen kertoimien (parametrien) lukumäärään *k*. Akaiken informaatiokriteeri lasketaan kaavalla: $AIC = 2k - 2\ln(L)$, jossa $L$ uskottavuusfunktio. Yleensä kuitenkin oletetaan virhetermien olevan normaalisti jakautuneita ja tällöin käytetään kaavaa $AIC = 2k + n\ln(\frac{SSE}{n})$, missä $n$ on havaintojen lukumäärä ja *SSE* virhetermien neliösumma. Perusideana Akaiken informaatiokriteerissä on, että se rankaisee parametrien määrästä. Hierarkkisessa mallin rakentamisessa valitaan usein se malli, jonka informaatiokriteerin arvo on pienin.

Seuraavaksi demonstroidaan mallin valintaa käytännössä. Aluksi rakennetaan malli, jossa on kaikki muuttujat mukana. Tarkastellaan tämän mallin tulostusta (taulukko 5.1). Lähes kaikki mallin muuttujat ovat tilastollisesti merkitseviä 5 %:n riskitasolla, vain Mosuus ja HT eivät ole. Korjattu selitysaste ei kuitenkaan ole parantunut kuin reilun prosentin verran verrattuna malliin, jossa oli pelkästään LPM selittäjänä. Saattaisi olla täysin perusteltua hyväksyä lopulliseksi malliksi tuo yksinkertainen malli jo tässä vaiheessa, mutta tarkastellaan mallin valintaa kuitenkin vielä eteenpäin. Tarkastelemalla varianssianalyysin tulostusta (taulukko 5.2), voimme arvioida muuttujien selitysvoimaa mallissa. Neliösumman mukaan LPM on selkeästi vahvin selittäjä, toiseksi paras on Alue ja kolmas on Kpituus.

**Taulukko 5.1** Regressioanalyysin tulokset, läpimittaluokat 70–160, mänty

| Muuttuja | Kerroin | Keskivirhe | t-arvo | p-arvo |
|---|---|---|---|---|
| (Intercept) | -0.9004419 | 0.0599560 | -15.018 | < 2e-16 |
| LPM | 0.0260335 | 0.0000762 | 341.634 | < 2e-16 |
| Alue | -0.0581673 | 0.0073139 | -7.953 | 2.21e-15 |
| DBAM | -0.0011911 | 0.0001094 | -10.885 | < 2e-16 |
| Mosuus | 0.0014282 | 0.0102655 | 0.139 | 0.889357 |
| FH12 | 0.0555398 | 0.0066691 | 8.328 | < 2e-16 |
| FH14 | -0.0789180 | 0.0160701 | -4.911 | 9.34e-07 |
| FH15 | 0.0205900 | 0.0088282 | 2.332 | 0.019723 |
| FH17 | 0.0455056 | 0.0126412 | 3.600 | 0.000321 |
| Kpituus | 0.0229154 | 0.0014914 | 15.365 | < 2e-16 |
| M1K | -0.0002564 | 0.0001033 | -2.484 | 0.013039 |
| MinlpmK | -0.0038697 | 0.0006012 | -6.437 | 1.33e-10 |
| MLosuus | -2.1946472 | 0.4334201 | -5.064 | 4.26e-07 |
| PituuslktK | -0.0047732 | 0.0013634 | -3.501 | 0.000467 |
| PT | 0.0311963 | 0.0052169 | 5.980 | 2.38e-09 |
| SM | 0.0499887 | 0.0088848 | 5.626 | 1.94e-08 |
| HT | -0.0023647 | 0.0072386 | -0.327 | 0.743922 |
| Korjattu selitysaste: 95.76 % | | | | |
| Residuaalin keskivirhe: 0.1569 | | | | |

**Taulukko 5.2** Varianssianalyysi, läpimittaluokat 70–160, mänty

| **Varianssianalyysi: Vaste: log(RarvoHT250)** | | | | | |
|---|---|---|---|---|---|
| | Df | Neliösumma | Keskineliösumma | F arvo | p-arvo |
| LPM | 1 | 2881.07 | 2881.07 | 1.1697e+05 | < 2.2e-16 |
| Alue | 1 | 13.80 | 13.80 | 5.6022e+02 | < 2.2e-16 |
| DBAM | 1 | 0.60 | 0.60 | 2.4168e+01 | 9.098e-07 |
| Mosuus | 1 | 1.37 | 1.37 | 5.5738e+01 | 9.645e-14 |
| FH | 4 | 6.02 | 1.51 | 6.1106e+01 | < 2.2e-16 |
| Kpituus | 1 | 8.39 | 8.39 | 3.4081e+02 | < 2.2e-16 |
| M1K | 1 | 0.04 | 0.04 | 1.5739e+00 | 0.2096995 |
| MinlpmK | 1 | 1.05 | 1.05 | 4.2478e+01 | 7.819e-11 |
| MLosuus | 1 | 0.31 | 0.31 | 1.2779e+01 | 0.0003537 |
| PituuslktK | 1 | 0.48 | 0.48 | 1.9569e+01 | 9.903e-06 |
| PT | 1 | 0.43 | 0.43 | 1.7283e+01 | 3.273e-05 |
| SM | 1 | 0.78 | 0.78 | 3.1567e+01 | 2.026e-08 |
| HT | 1 | 0.0026 | 0.0026 | 1.0670e-01 | 0.7439220 |
| Residuals | 5228 | 128.77 | 0.02 | | |

Koska mallissa on mukana paljon muuttujia, on syytä tarkastella kollineaarisuusoletuksen (liite B.2) paikkaansa pitävyyttä. Tätä varten on laskettu liitteessä tarkemmin esitetyn varianssi-inflaattorin (VIF) arvot kaikille mallin muuttujille (taulukko 5.3).

**Taulukko 5.3** VIF-arvot, läpimittaluokat 70–160, useamman selittäjän malli, mänty

| | | | | | | |
|---|---|---|---|---|---|---|
| **LPM** | **Alue** | **DBAM** | **Mosuus** | **FH12** | **FH14** | **FH15** |
| 1.001428 | 3.038773 | 8.245861 | 1.668399 | 1.606204 | 1.613227 | 4.066683 |
| **FH17** | **Kpituus** | **M1K** | **MinlpmK** | **MLosuus** | **PituuslktK** | **M1** |
| 1.875377 | 5.067028 | 1.643873 | 1.662002 | 1.206116 | 1.764501 | 2.302404 |
| **Minlpm** | **Pituuslkt** | **PT** | **SM** | **HT** | **RP** | |
| 1.062748 | 1.674391 | 1.478601 | 1.380189 | 1.599978 | 1.774121 | |

Kun VIF-arvot ovat suurempia kuin 5, mallissa esiintyy mahdollisesti kollineaarisuusongelma (ks. liite B, s.99). Koska muuttujat DBAM ja Kpituus saavat suurimmat arvot, perustetaan ongelman korjaustarkastelu niihin. Rakennetaan uudet mallit siten, että ensimmäisestä jätetään pois DBAM, toisesta Kpituus ja kolmanteen sijoitetaan molempien tilalle niistä muodostettu uusi muuttuja. Kaikista malleista saadaan nyt ongelma poistettua ja parhaimman selitysasteen tuottaa malli, josta on DBAM poistettu.

Edellä esiteltiin, miten voidaan toimia, mikäli mallissa esiintyy kollineaarisuutta. Jatketaan nyt kuitenkin mallin kehittelyä tästä huolimatta siten, että DBAM pidetään mukana ja tarkastellaan lopuksi kollineaarisuutta uudestaan. Käytetään R:n step-funktiota mallin redusoinnissa. Tarkasteltaessa redusoidun mallin tulostusta (taulukko 5.4) huomataan, että siinä on paljon muuttujia mukana. Selitysaste ei kuitenkaan juuri muutu, vaikka mallista poistetaan manuaalisesti pienemmän kontribuution muuttujia. Mallin yksinkertaistamiseksi on vielä syytä redusoida mallia. Tähän saamme apua tarkastelemalla varianssianalyysin tuloksia (taulukko 5.5), joista saamme kuvan eri muuttujien ”painoarvoista” mallissa.

**Taulukko 5.4** Regressioanalyysin tulokset, läpimittaluokat 70–160, mänty

| Muuttuja | Kerroin | Keskivirhe | t-arvo | p-arvo |
|---|---|---|---|---|
| (Intercept) | -8.993e-01 | 5.901e-02 | -15.240 | < 2e-16 |
| LPM | 2.603e-02 | 7.618e-05 | 341.740 | < 2e-16 |
| Alue | -5.802e-02 | 6.928e-03 | -8.375 | < 2e-16 |
| DBAM | -1.205e-03 | 9.804e-05 | -12.290 | < 2e-16 |
| FH12 | 5.547e-02 | 6.649e-03 | 8.342 | < 2e-16 |
| FH14 | -7.859e-02 | 1.581e-02 | -4.970 | 6.90e-07 |
| FH15 | 2.061e-02 | 8.827e-03 | 2.335 | 0.019606 |
| FH17 | 4.560e-02 | 1.242e-02 | 3.670 | 0.000245 |
| Kpituus | 2.303e-02 | 1.414e-03 | 16.288 | < 2e-16 |
| M1K | -2.518e-04 | 1.022e-04 | -2.465 | 0.013752 |
| MinlpmK | -3.867e-03 | 6.009e-04 | -6.434 | 1.35e-10 |
| MLosuus | -2.201e+00 | 4.301e-01 | -5.118 | 3.20e-07 |
| PituuslktK | -4.760e-03 | 1.362e-03 | -3.495 | 0.000479 |
| PT | 3.100e-02 | 5.108e-03 | 6.069 | 1.38e-09 |
| SM | 5.009e-02 | 8.757e-03 | 5.720 | 1.12e-08 |
| Korjattu selitysaste: 95.76 % | | | | |
| Residuaalin keskivirhe: 0.1569 | | | | |

**Taulukko 5.5** Varianssianalyysi, läpimittaluokat 70–160, mänty

| Varianssianalyysi: Vaste: log(RarvoHT250) | | | | | |
|---|---|---|---|---|---|
| | Df | Neliösumma | Keskineliösumma | F arvo | p-arvo |
| **LPM** | 1 | 2881.07 | 2881.07 | 1.1701e+05 | < 2.2e-16 |
| **Alue** | 1 | 13.80 | 13.80 | 5.6042e+02 | < 2.2e-16 |
| **DBAM** | 1 | 0.60 | 0.60 | 2.4177e+01 | 9.058e-07 |
| **FH** | 4 | 5.50 | 1.38 | 5.5873e+01 | < 2.2e-16 |
| **Kpituus** | 1 | 10.20 | 10.20 | 4.1444e+02 | < 2.2e-16 |
| **M1K** | 1 | 0.04 | 0.04 | 1.5459e+00 | 0.2137996 |
| **MinlpmK** | 1 | 1.06 | 1.06 | 4.3093e+01 | 5.724e-11 |
| **MLosuus** | 1 | 0.34 | 0.34 | 1.3927e+01 | 0.0001921 |
| **PituuslktK** | 1 | 0.48 | 0.48 | 1.9343e+01 | 1.114e-05 |
| **PT** | 1 | 0.44 | 0.44 | 1.7915e+01 | 2.349e-05 |
| **SM** | 1 | 0.81 | 0.81 | 3.2721e+01 | 1.123e-08 |
| **Residuals** | 5230 | 128.77 | 0.02 | | |

Tietyn muuttujan malliin tuoman kontribuution lisäksi on syytä hahmotella muuttujien taustalla olevia tekijöitä. Halutessamme verrata tuloksia eri työmaiden välillä olisi hyvä, jos mallissa olisi mukana muuttujia, joiden voisi olettaa kuvaavan työmaiden mahdollisia peruseroja. Tällaisia muuttujia ovat maantieteellistä sijaintia kuvaava alue (Alue), rungon käyttöosan keskipituus (Kpituus) ja tietynlaisen leimikon valikoituminen hakkuutavan mukaan (FH). Näiden muuttujien kontribuutio malliin onkin merkittävä, kun taas hakkuuohjeista tulevat vaikutukset tuntuvat olevan hyvin pieniä. Niiden mahdollinen vaikutus tulokseen ”jää suurimmaksi osaksi muiden selittäjien alle”, ainakin kun tarkastellaan osa-aineistoille muodostettuja malleja. Kuidun minimiläpimitta on niillä rajoilla, kannattaako sitä ottaa jatkomalliin mukaan. Tässä tapauksessa päätän ottaa sen vielä mukaan. Hakkuuohjeiden vaikutuksella tarkoitetaan hakkuuohjeen parametreja, jotka määrittelevät katkottavien pölkkyjen mitat. Myös työmaalta kerättävien puutavaralajien (esimerkiksi kerätäänkö pikkutukkia vai ei) kontribuutio tulokseen jää kovin pieneksi.

Rakennetaan siis vielä malli, jossa selittäjinä ovat LPM, Alue, Kpituus, FH sekä MinlpmK. Mallin selitysaste on 95.53 %, joten se on vain alle prosentin parempi kuin malli, jossa on selittäjänä vain LPM. Kuriositeetin vuoksi katsotaan vielä mallin VIF-arvot (taulukko 5.6). Kuten voidaan havaita, mallissa ei esiinny kollineaarisuutta. Loppujen lopuksi järkevin valinta lopulliseksi malliksi on yksinkertaisempi malli (5.1), koska sen selitysvoima ei juuri eroa monimutkaisemmasta mallista.

**Taulukko 5.6** VIF-arvot, läpimittaluokat 70–160, useamman selittäjän malli, mänty

| LPM | DBAM | FH12 | FH14 | FH15 | FH17 | Kpituus |
|---|---|---|---|---|---|---|
| 1.089341 | 5.114297 | 1.684263 | 1.948345 | 4.022464 | 1.846181 | 3.160686 |

### *5.1.2.2 Läpimittaluokat 170–400*

Tästä eteenpäin käytetään mallien redusoinnissa ja oletuksien tarkasteluissa samanlaisia menetelmiä kuin edellisen tarkastelun yhteydessä, joten niitä ei erikseen tarkemmin esitellä, ellei siihen ole nähty erityistä syytä. Mallien välivaiheita ei myöskään taulukoin esitellä. Jälkimmäiselle osa-aineistolle tehtävään analyysiin otetaan mukaan seuraavat muuttujat: LPM, Alue, DBAM, Mosuus, Kpituus, M1K, MinlpmK, MLosuus, PituuslktK, M1, Minlpm, Pituuslkt, PT, SM, HT, RP ja FH.

Selitettäessä tulosta virhetermit ovat jälleen kerran heteroskedastisia. Ongelman korjaamiseksi kokeillaan sekä painotetun neliösumman menetelmää (*Weighted Least Squares, WLS*) että vasteen eli tuloksen logaritmointia, mikä on usein käytetty menetelmä vastaavissa tapauksissa.

Painotettu pienimmän neliösumman menetelmä tuottaa selitysasteeltaan lähes saman tuloksen selitettävän logaritmoinnin kanssa. WLS on virhetermin normaalisuuden suhteen paremman oloinen valinta, mutta heteroskedastisuuden suhteen logaritmointi näyttää hiukan paremmalta. Tässä vaiheessa on syytä todeta, että malleissa esiintyy kollineaarisuutta. Tämä ilmenee VIF-arvoista. Jatkan kuitenkin mallin jatkokehittelyä logaritmoidun vasteen kanssa ja otan kantaa tähän ongelmaan vasta myöhemmin.

AIC tuottaa kaavan (5.3) mukaisen mallin.

$$
\begin{aligned}
\log(\hat{y}) = {} & \hat{\beta}_0 + \hat{\beta}_1 * LPM + \hat{\beta}_2 * DBAM + \hat{\beta}_3 * Mosuus + \hat{\beta}_4 * FH12 + \hat{\beta}_5 * FH14 \\
& + \hat{\beta}_6 * FH15 + \hat{\beta}_7 * FH17 + \hat{\beta}_8 * Kpituus + \hat{\beta}_9 * M1K + \hat{\beta}_{10} * MinlpmK \\
& + \hat{\beta}_{11} * MLosuus + \hat{\beta}_{12} Pituuslkt + \hat{\beta}_{13} * PT + \hat{\beta}_{14} * SM + \hat{\beta}_{15} * RP
\end{aligned}
\tag{5.3}
$$

Mallissa on kollineaarisuusongelman lisäksi mukana useita regressioneliösummaan vähän vaikuttavia muuttujia, joiden kontribuutio selitysasteeseen on hyvin pieni. On syytä vähentää selittäviä muuttujia ja toivoa, että saisimme tällöin myös ongelman ratkaistua. Aiemmin analysoituun osa-aineistoon verrattuna nyt Alue ei ole merkittävä selittäjä. Otetaan seuraavaan malliin mukaan Kpituus, FH sekä DBAM.

Edelleen ongelmana on kollineaarisuus. Varianssianalyysin perusteella voitaisiin poistaa mallista DBAM, koska sen kontribuutio malliin on pienempi kuin muuttujan Kpituus. Tehdään poisto siten, että malli rakennetaan kokonaan alusta ilman muuttujaa DBAM ja katsotaan, millaiseen malliin loppujen lopuksi päädytään. Mallissa, jossa on kaikki mahdolliset selittäjät paitsi DBAM, ei enää esiinny ongelmaa. Lopulliseksi malliksi saadaan kaavan (5.4) mukainen malli.

$$\begin{aligned}\log(\hat{y}) &= 0.8189 + 0.01054 * LPM + 0.2148 * FH12 + 0.1286 * FH15 \\ &\quad + 0.1425 * H17 + 0.06826 * Kpituus\end{aligned} \tag{5.4}$$

### 5.1.3 Lopulliset mallit

Yhteenvetona analyyseistä esitetään läpimittaluokkajaoille mallit, joihin on päädytty. Samalla tulkitaan mallien kertoimet ja muodostetaan malleille ennustevälit.

#### *5.1.3.1 Läpimittaluokat 70–160*

Vakiotermi ja läpimittaluokka ovat erittäin merkitseviä selittäjiä (p-arvo<$2*10^{-16}$) (taulukko). Selitysaste on hyvin lähellä ykköstä eli pelkkä läpimittaluokka selittää tuloksen logaritmiä lähes täysin.

**Taulukko 5.7** Regressioanalyysin tulokset, läpimittaluokat 70–160, mänty

| Muuttuja | Kerroin | Keskivirhe | t-arvo | p-arvo |
|---|---|---|---|---|
| (Intercept) | -1.281 | 1.018e-02 | -125.9 | < 2e-16 |
| LPM | 2.604e-02 | 8.543e-05 | 304.8 | < 2e-16 |
| Korjattu selitysaste: 94.66 % | | | | |
| Residuaalin keskivirhe: 0.1761 | | | | |

Läpimittaluokan muutoksen kontribuutio tuloksen odotusarvoon on $e^{0.02604}$ läpimittaluokan arvon kasvaessa yhdellä. Koska läpimittaluokan muutos on käytetyistä yksiköistä johtuen aina vähintään kymmenen yksikköä, niin läpimittaluokan arvon kasvaessa kymmenellä, kasvaa tulos $e^{0.2604}$ -kertaiseksi eli noin 1.3 -kertaiseksi.

**Taulukko 5.8** Yhteenveto männyn läpimittaluokkien 70–160 regressioanalyyseistä

| **Malliyhtälöt:** |
|---|
| $\log(y) = \beta_0 + \beta_1 X_1 + \varepsilon$ |
| $\log(\hat{y}) = -1.281 + 0.02604 * \text{LPM}$ |
| $\hat{y} = \exp(-1.281 + 0.02604 * \text{LPM})$ |
| **Ennustevälit:** |
| 90 % $\hat{y} = \exp(-1.281+0.02604*\text{LPM} \pm 1.645*0.1761)$ |
| 95 % $\hat{y} = \exp(-1.281+0.02604*\text{LPM} \pm 1.96*0.1761)$ |
| 99 % $\hat{y} = \exp(-1.281+0.02604*\text{LPM} \pm 2.576*0.1761)$ |

Ennustevälillä voidaan tarkastella, kuinka hyvin uusi tulos sijoittuu analyysien perustana oleviin tuloksiin nähden. Ennusteväleistä lisää liitteessä B.9.

### *5.1.3.2 Läpimittaluokat 170–400*

Erittäin merkittäviä tuloksen logaritmin selittäjiä ovat läpimittaluokka, käyttöosan keskipituus sekä hakkuutavat 12, 15 ja 17. Periaatteessa hakkuutavat 15 ja 17 voitaisiin yhdistää samaksi hakkuutavaksi, sillä hakkuutavan 15 kerroin sisältyy FH17:n luottamusvälille. Tässä päätän kuitenkin pitää hakkuutavat erillään. Selitysasteeksi saatiin 83.81 %. Mallissa olisi ollut myös mukana pohjapinta-alamediaaniläpimitta, mutta sen ja käyttöosan keskipituuden välinen vahva korrelaatio aiheutti aineistoon kollineaarisuusongelman. Ongelma saatiin ratkaistua pudottamalla mallista toinen näistä muuttujista pois. DBAM pudotettiin, koska varianssianalyysin mukaan muuttujan Kpituus kontribuutio regressioneliösummaan oli suurempi.

**Taulukko 5.9** Regressioanalyysin tulokset, läpimittaluokat 170–400, mänty

| Muuttuja | Kerroin | Keskivirhe | t-arvo | p-arvo |
|---|---|---|---|---|
| (Intercept) | 8.189e-01 | 1.750e-02 | 46.804 | <2e-16 |
| LPM | 1.054e-02 | 5.443e-05 | 193.675 | <2e-16 |
| FH12 | 2.148e-01 | 1.074e-02 | 19.995 | <2e-16 |
| FH15 | 1.286e-01 | 9.570e-03 | 13.437 | <2e-16 |
| FH17 | 1.425e-01 | 1.530e-02 | 9.314 | <2e-16 |
| Kpituus | 6.826e-02 | 1.171e-03 | 58.299 | <2e-16 |
| Korjattu selitysaste: 83.81 % | | | | |
| Residuaalin keskivirhe: 0.3422 | | | | |

Koska myös tässä mallissa selitetään logaritmoitua tulosta, voidaan vastaavalla tavalla laskea yksittäisten muuttujien kontribuutiot tulokseen, kun muiden muuttujien arvot pysyvät muuttumattomina. Tällöin läpimittaluokan arvon kasvaessa kymmenellä tulos kasvaa 1.11-kertaiseksi. Keskipituuden kasvaessa yhdellä yksiköllä tulos kasvaa 1.07-kertaiseksi. Hakkuutapoja 12, 15 ja 17 verrataan hakkuutapoihin 11 ja 14, jotka muodostavat ns. hylkyluokan. Kun kyseessä on hakkuutapa 12, kasvaa tulos 1.24-kertaiseksi. Jos kyseessä on hakkuutapa 15, kasvaa tulos 1.14-kertaiseksi. Hakkuutavan 17 kyseessä ollessa on tulos 1.15-kertainen. Huomioi, että tuloksen kertaistuminen on yhtä pitävää prosentuaalisen kasvun kanssa. Siis, jos tulos 1.11-kertaistuu, tarkoittaa se, että tulos kasvaa 11 %.

**Taulukko 5.10** Yhteenveto männyn läpimittaluokkien 170–400 regressioanalyyseistä

| | |
|---|---|
| **Malliyhtälöt:** | |
| $\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$ | |
| $\log(\hat{y}) = 0.8189 + 0.01054 * \mathrm{LPM} + 0.2148 * \mathrm{FH12} + 0.1286 * \mathrm{FH15} + 0.1425 * \mathrm{FH17} + 0.06826 * \mathrm{Kpituus}$ | |
| $\hat{y} = \exp(0.8189 + 0.01054 * \mathrm{LPM} + 0.2148 * \mathrm{FH12} + 0.1286 * \mathrm{FH15} + 0.1425 * \mathrm{FH17} + 0.06826 * \mathrm{Kpituus})$ | |
| **Ennustevälit:** | |
| 90 % | $\hat{y}$ = exp(0.8189+0.01054*LPM+0.2148*FH12+0.1286*FH15+0.1425*FH17 +0.06826*Kpituus ± 1.645*0.3422) |
| 95 % | $\hat{y}$ = exp(0.8189+0.01054*LPM+0.2148*FH12+0.1286*FH15+0.1425*FH17 +0.06826*Kpituus ± 1.96*0.3422) |
| 99 % | $\hat{y}$ = exp(0.8189+0.01054*LPM+0.2148*FH12+0.1286*FH15+0.1425*FH17 +0.06826*Kpituus ± 2.576*0.3422) |

## 5.2 Regressioanalyysit kuuselle

Aluksi muodostetaan, kuten männynkin tapauksessa, ensimmäiset yksinkertaiset regressiomallit. Seuraavaksi otetaan mukaan kaikki mahdolliset selittäjät ja redusoidaan mallit merkitseviin, yksinkertaisiin ja sopiviin muotoihin. Lopullisille malleille suoritetaan tulkinnat ja rakennetaan ennustevälit.

### 5.2.1 Yksinkertaiset mallit

Tarkoituksena on rakentaa mahdollisimman yksinkertaiset mallit. Tämä tarkoittaa, että malleihin otetaan mukaan vain parhaimmat selittäjät. Ensin tutkitaan koko aineistoa ja sitten osa-aineistoja. Tarvittaessa tehdään datalle muunnoksia tai käytetään muita kuin pienimmän neliösumman menetelmää analyysien suorittamiseksi.

Jaetaan läpimittaluokat nyt kolmeen ryhmään A, B ja C ja tehdään ryhmille erilliset regressioanalyysit. Ryhmäjaolle on jo aiemmin esitetty kriteerit ja perustelut. Tarkastellaan aluksi läpimittaluokkia alle 170 mm. Ensimmäiseen malliin valitaan selittäjäksi ainoastaan LPM ja sillä selitetään tulosta.

#### *5.2.1.1 Läpimittaluokat alle 170*

Kuten männynkin kohdalla, virhetermin vakiovarianssisuusoletus ei täyty, ellei ongelmaa käsitellä jotenkin. Männyn tapauksessa päädyttiin logaritmoimaan tulos. Tällä kertaa otetaan mukaan lisäksi tuloksen kuutiojuurimuunnos. Se on valittu muunnokseksi Boxin ja Coxin menetelmän (liite B.5) avulla. Selitetään sekä tuloksen kuutiojuurta että logaritmoitua tulosta pelkällä LPM:lla. Tarkastellaan regressiomallien oletusten paikkaansa pitävyyttä (liite B.2) ja huomataan, että kuutiojuurimuunnoksen tapauksessa virhetermit ovat heteroskedastisia eli niiden varianssi ei ole vakio. Logaritmimuunnoksen tapauksessa heteroskedastisuus ei näytä olevan ongelma. Molemmissa tapauksissa virhetermien normaalijakautuneisuus voidaan karkeasti ottaen hyväksyä. Vertailtaessa mallien hyvyyttä voidaan todeta vasteen logaritmoinnin olevan selitysasteeltaan reilun puoli prosenttia paremman. Korkeampi selitysaste 95.89 % ja virhetermin homoskedastisuusoletuksen parempi täyttyminen johtavat valinnan logaritmimuunnoksen suuntaan, joten lopullinen malli on muotoa $\log(y) = \beta_0 + \beta_1 * LPM$ , jossa vakiokertoimen $\beta_0$:n estimaatti $\hat{\beta}_0$ saa arvon 0.04817 ja regressiokertoimen $\beta_1$:n estimaatti $\hat{\beta}_1$ arvon 0.01545.

#### *5.2.1.2 Läpimittaluokat 170–260*

Suoritetaan analyysit seuraavaksi ryhmälle B, johon siis kuuluvat luokat 170–260. Otetaan aluksi mukaan selittäjäksi vain LPM. Selitysaste pelkällä läpimitalla on 74.53 % ja se on tässä tapauksessa selkeästi pienempi kuin ryhmän A tapauksessa. Se on kuitenkin edelleen huomattavan suuri.

#### *5.2.1.3 Läpimittaluokat 270–420*

Viimeiseen ryhmään C kuuluvat läpimittaluokat 270–420. Myös tällä kertaa selitetään tuloksen kuutiojuurta LPM:lla, koska se vaikuttaisi Boxin ja Coxin menetelmän perusteella sopivimmalta muunnokselta. Selitysasteeksi saadaan nyt 52.34 %. Selitysaste on pudonnut nyt huomattavasti edellisten läpimitta-luokittaisten ryhmien mallien selitysasteista. Selitysasteeseen yritetään saada parannusta muodostamalla mallit, joihin otetaan mukaan lisää selittäjiä.

### 5.2.2 Usean selittäjän mallit

#### *5.2.2.1 Läpimittaluokat alle 170*

Seuraavaksi lisätään malleihin selittäjiä. Osa-aineistolle A tehtävään malliin otetaan mukaan seuraavat muuttujat: LPM, Alue, DBAK, Mosuus, Kpituus, M1K, MinlpmK, KLosuus, PituuslktK, KPT, KSE, ja FH. Selitettävänä on logaritmoitu tulos.

Selitysaste kasvaa noin prosentin verrattuna yksinkertaiseen malliin, jossa selittäjänä oli pelkkä LPM. Jo tällä perusteella voisimme pysäyttää analyysin tähän ja todeta, että alle 170 läpimittaluokissa LPM selittää lähes kaiken ja malliin on turha lisätä muuttujia. Jatkan kuitenkin tilanteen tarkastelua, jotta saisimme selville, mitkä jäljellä olevista selittäjistä selittävät vahvimmin jäljelle jäävää osaa.

Automaattinen mallin valinta valitsee kaavan (5.5) mukaisen mallin.

$$\begin{aligned}\log(\hat{y}) = {} & \hat{\beta}_0 + \hat{\beta}_1 * LPM + \hat{\beta}_2 * Alue + \hat{\beta}_3 * Kpituus + \hat{\beta}_4 * KLosuus_5 + \hat{\beta}_5 * DBAK \\ & + \hat{\beta}_6 * MinlpmK + \hat{\beta}_7 * Mosuus + \hat{\beta}_8 * PituuslktK + \hat{\beta}_9 * FH14\end{aligned} \tag{5.5}$$

Merkittäviä selittäjiä on läpimittaluokka, Alue, käyttöosan keskipituus, kuusilumppien osuus, DBAK, kuidun minimiläpimitta, männyn/kuusen osuus leimikon kaadetusta puustosta, kuidun pituusluokkien määrä ja hakkuutapa. Mallia ei vielä tässä vaiheessa voida hyväksyä, vaan täytyy tarkastella, esiintyykö mallissa kollineaarisuutta. VIF (*Variation Inflation Rate*) ilmaisee, että mallissa on todennäköisesti kollineaarisuutta. Helpoin tapa käsitellä tätä ongelmaa on poistaa mallista tämän aiheuttaja tai aiheuttajat. Voidaan havaita, että ongelma on Kpituuden ja DBAK:n välinen. Pudotetaan mallista pois Kpituus, koska varianssianalyysin mukaan DBAK on selittäjän parempi eli sillä on suurempi vaikutus mallin kokonaisneliösummaan. Nyt tarkasteltaessa tilannetta, huomataan, että ongelma on poistunut.

Tässä olisi taas yksi mahdollisuus lopettaa mallin kehittely, mutta koska mallissa on edelleen mukana hyvin pienivaikutteisia selittäjiä, muokataan mallia vielä hiukan. Pudotetaan muuttujista pois muut paitsi LPM, Alue ja Mosuus.

Mallin selitysaste putosi edellisestä vain 0.06 prosenttia. Lopullisena mallina voidaan pitää tätä, automaattisen mallin valinnan valitsemaa mallia, korjattuna kollineaarisuudesta tai pelkän LPM:n mallia. Kyse on siitä minkä mallin haluaa valita tai mikä palvelee tarkoitusta parhaiten. Malli on siis nyt muotoa $\log(y) = \beta_0 + \beta_1 * LPM + \beta_2 * Alue + \beta_3 * Mosuus$, missä $\beta_0$:n estimaatti $\hat{\beta}_0$ saa arvon -1.17, $\beta_1$:n estimaatti $\hat{\beta}_1$ arvon 0.02592, $\beta_2$:n estimaatti $\hat{\beta}_2$ arvon -0.08266 ja $\beta_3$:n estimaatti $\hat{\beta}_3$ arvon -0.08683.

### *5.2.2.2 Läpimittaluokat 170–260*

Osa-aineistolle B tehtävään malliin otetaan mukaan seuraavat muuttujat: LPM, Alue, DBAM, Mosuus, Kpituus, M1K, MinlpmK, MLosuus, PituuslktK, M1, Minlpm, Pituuslkt, PT, SM, HT, RP ja FH.

Selitysaste kasvaa noin 12 prosenttia verrattuna malliin, jossa on vain läpimittaluokka selittäjänä. Automaattisen mallin valinnan tuottaman mallin oletukset ovat muuten kohdallaan, mutta mallissa on taas kollineaarisuutta. Anovan tulostuksen perusteella poistetaan tällä kertaa DBAK, koska Kpituus näyttäisi selittävän tulosta huomattavasti paremmin. Muuttujan poiston jälkeen mallissa ei enää näytä olevan kollineaarisuusongelmaa. Kehitellään mallia kuitenkin vielä hiukan. Jätetään malliin muuttujat LPM, Alue, Kpituus, KLosuus ja FH. Nyt malli on kaavan (5.6) mukainen.

$$\sqrt[3]{\hat{y}} = -1.54578 + 0.022306 * LPM - 0.125775 * Alue + 0.065965 * Kpituus - 6.833098 * KLosuus + 0.195181 * FH12 + 0.158485 * FH15 + 0.168622 * FH17 \tag{5.6}$$

### *5.2.2.3 Läpimittaluokat 270–420*

Viimeiseksi tarkastellaan läpimittaluokkia 270–420. Edellisen ryhmän tavoin selitettävänä muuttujana on tuloksen kuutiojuurimuunnos. Mallin, jossa kaikki muuttujat ovat mukana, selitysaste on 74.81 %. Kehittelemällä mallia aiemmin esille tullein tavoin saadaan lopulliseksi malliksi

$$\sqrt[3]{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 * LPM + \hat{\beta}_2 * Alue + \hat{\beta}_3 * Kpituus + \hat{\beta}_4 * Mosuus ,$$

jossa $\hat{\beta}_0$ on 0.0463876, $\hat{\beta}_1$ on 0.0146630, $\hat{\beta}_2$ on -0.3218390, $\hat{\beta}_3$ on 0.1010912 ja $\hat{\beta}_4$ on 0.3410067. Tämän mallin selitysaste on 72.33 % eli se on vain noin 2.5 % pienempi kuin mallin, jossa kaikki muuttujat ovat selittäjinä.

## 5.2.3 Lopulliset mallit

Myös kuusen tapauksessa malleissa esiintyy kollineaarisuutta, kun selittäjinä on sekä Kpituus että DBAK. Ongelma on ratkaistu pudottamalla näistä regressioneliösummaan vähäisemmän kontribuution antava. Muodostetuista malleista lopullisiksi malleiksi eri läpimittaluokkaryhmille valitsen seuraavat mallit:

### *5.2.3.1 Läpimittaluokat alle 170*

Vakiotermi, läpimittaluokka, alue ja männyn osuus ovat erittäin merkitseviä selittäjiä (p-arvo<$2*10^{-16}$) (taulukko 5.9). Selitysaste on hyvin lähellä ykköstä (96.41 %) eli tuloksen logaritmista saadaan kyseisillä muuttujilla selitettyä erittäin paljon.

**Taulukko 5.11** Regressioanalyysin tulokset, läpimittaluokat 70–160, kuusi

| Muuttuja | Kerroin | Keskivirhe | t arvo | p-arvo |
|---|---|---|---|---|
| (Intercept) | -1.170e+00 | 1.234e-02 | -94.77 | <2e-16 |
| LPM | 2.592e-02 | 9.345e-05 | 277.37 | <2e-16 |
| Alue | -8.266e-02 | 5.443e-03 | -15.19 | <2e-16 |
| Mosuus | -8.683e-02 | 8.924e-03 | -9.73 | <2e-16 |
| Selitysaste: 96.41 % | | | | |
| Residuaalin keskivirhe: 0.1412 | | | | |

Yksittäisten muuttujien kontribuutiot tulokseen voidaan kuvata seuraavasti. Läpimittaluokan muutoksen kontribuutio tuloksen odotusarvoon on $e^{0.02592}$ läpimittaluokan arvon kasvaessa yhdellä. Tällöin tulos siis kasvaa 1.03 -kertaiseksi. Koska läpimittaluokka muutos on käytetyistä yksiköistä johtuen aina vähintään kymmenen yksikköä, niin läpimittaluokan arvon kasvaessa kymmenellä, kasvaa tulos $e^{0.2592}$ eli noin 1.3 -kertaiseksi. Pohjoisen alueella tulos on noin 0.92 -kertainen etelään verrattuna eli pohjoisen tulos on noin 92 % etelän tuloksesta. Männyn osuuden kasvu leimikossa 10 %:lla, vähentää tulosta noin yhdellä prosentilla.

**Taulukko 5.12** Yhteenveto kuusen läpimittaluokkien 70–160 regressioanalyyseistä

| **Malliyhtälöt:** |
|---|
| $\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ |
| $\log(\hat{y}) = -1.17 + 0.02592 * LPM - 0.08266 * Alue - 0.08683 * Mosuus$ |
| $\hat{y} = \exp[-1.17 + 0.02592 * LPM - 0.08266 * Alue - 0.08683 * Mosuus]$ <br> $= e^{[-1.17+0.02592*LPM-0.08266*Alue-0.08683*Mosuus]}$ <br> $= e^{-1.17} * e^{0.02592*LPM} * e^{-0.08266*Alue} * e^{-0.08683*Mosuus}$ <br> $= 0.31037 * e^{0.02592*LPM} * e^{-0.08266*Alue} * e^{-0.08683*Mosuus}$ |
| **Ennustevälit:** |
| 90 % $\hat{y}$ =exp(-1.17+0.02592**LPM*-0.08266**Alue*-0.08683**Mosuus* ± 1.645*0.1412) |
| 95 % $\hat{y}$ = exp(-1.17+0.02592**LPM*-0.08266**Alue*-0.08683**Mosuus* ± 1.96*0.1412) |
| 99 % $\hat{y}$ =exp(-1.17+0.02592**LPM*-0.08266**Alue*-0.08683**Mosuus* ± 2.576*0.1412) |

### *5.2.3.2 Läpimittaluokat 170–260*

Merkittäviä tuloksen kuutiojuuren selittäjiä ovat läpimittaluokka, alue, käyttöosan keskipituus, lumppien osuus sekä hakkuutapa (taulukko 5.10). Selitysaste on 85.47 %.

**Taulukko 5.13** Regressioanalyysin tulokset, läpimittaluokat 170–260, kuusi

| Muuttuja | Kerroin | Keskivirhe | t arvo | p-arvo |
|---|---|---|---|---|
| (Intercept) | -1.545780 | 0.046134 | -33.506 | < 2e-16 |
| LPM | 0.022306 | 0.000188 | 118.635 | < 2e-16 |
| Alue | -0.125775 | 0.011969 | -10.509 | < 2e-16 |
| Kpituus | 0.065965 | 0.002388 | 27.622 | < 2e-16 |
| KLosuus | -6.833098 | 0.494189 | -13.827 | < 2e-16 |
| FH12 | 0.195181 | 0.018217 | 10.714 | < 2e-16 |
| FH15 | 0.158485 | 0.016890 | 9.383 | < 2e-16 |
| FH17 | 0.168622 | 0.030102 | 5.602 | 2.33e-08 |
| Selitysaste: 85.47 % | | | | |
| Residuaalin keskivirhe: 0.2847 | | | | |

Tässä tapauksessa, kun selitetään tuloksen kuutiojuurta, ei ole tarjolla samanlaisia yksittäisten kertoimien tulkintoja. Tässä tapauksessa tulkinnat menevät muuten normaaliin tapaan, mutta on muistettava, että selityksen kohteena on tuloksen kuutiojuuri.

Läpimittaluokan kasvaessa kymmenellä kasvaa tuloksen kuutiojuuri 0.22 euroa. Pohjoisen alueen tuloksen kuutiojuuri on 0.13 euroa pienempi kuin etelän. Käyttöosan keskipituuden kasvaessa yhdellä yksiköllä kasvaa tuloksen kuutiojuuri noin 0.07 euroa. Lumppiosuuden kasvaessa yhdellä prosentilla pienenee tuloksen kuutiojuuri 0.07 euroa. Kun kyseessä on hakkuutapa 12, kasvaa tuloksen kuutiojuuri noin 0.2 euroa verrattuna hakkuutapoihin 11 ja 14. Hakkuutapa 15 parantaa edellisen lailla tulosta noin 0.16 euroa ja hakkuutapa 17 noin 0.17 euroa.

**Taulukko 5.14** Yhteenveto kuusen läpimittaluokkien 170–260 regressioanalyyseistä

| **Malliyhtälöt:** |
|---|
| $\sqrt[3]{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$ |
| $\sqrt[3]{\hat{y}} = -1.54578 + 0.022306 * LPM - 0.125775 * Alue + 0.065965 * Kpituus$ $- 6.833098 * KLosuus + 0.195181 * FH12 + 0.158485 * FH15 + 0.168622 * FH17$ |
| $\hat{y} = \begin{bmatrix} -1.54578 + 0.022306 * LPM - 0.125775 * Alue + 0.065965 * Kpituus \\ -6.833098 * KLosuus + 0.195181 * FH12 + 0.158485 * FH15 + 0.168622 * FH17 \end{bmatrix}^3$ |
| **Ennustevälit:** |
| 90 %: $\hat{y} = \begin{bmatrix} -1.54578 + 0.022306 * LPM - 0.125775 * Alue + 0.065965 * Kpituus - 6.833098 * KLosuus \\ +0.195181 * FH12 + 0.158485 * FH15 + 0.168622 * FH17 \pm 1.645 * 0.2847 \end{bmatrix}^3$ |
| 95 %: $\hat{y} = \begin{bmatrix} -1.54578 + 0.022306 * LPM - 0.125775 * Alue + 0.065965 * Kpituus - 6.833098 * KLosuus \\ +0.195181 * FH12 + 0.158485 * FH15 + 0.168622 * FH17 \pm 1.96 * 0.2847 \end{bmatrix}^3$ |
| 99 %: $\hat{y} = \begin{bmatrix} -1.54578 + 0.022306 * LPM - 0.125775 * Alue + 0.065965 * Kpituus - 6.833098 * KLosuus \\ +0.195181 * FH12 + 0.158485 * FH15 + 0.168622 * FH17 \pm 2.576 * 0.2847 \end{bmatrix}^3$ |

### *5.2.3.3 Läpimittaluokat 270–420*

Merkittäviä tuloksen kuutiojuuren selittäjiä näissä läpimittaluokissa ovat vakiokertoimen lisäksi läpimittaluokka, alue, käyttöosan keskipituus ja männyn osuus leimikosta (taulukko 5.11). Selitysaste on 72.33 %.

**Taulukko 5.15** Regressioanalyysin tulokset, läpimittaluokat 270–420, kuusi

| Muuttuja | Kerroin | Keskivirhe | t arvo | p-arvo |
|---|---|---|---|---|
| (Intercept) | 0.0463876 | 0.0874903 | 0.53 | 0.596 |
| LPM | 0.0146630 | 0.0002398 | 61.13 | <2e-16 |
| Alue | -0.3218390 | 0.0185478 | -17.35 | <2e-16 |
| Kpituus | 0.1010912 | 0.0034836 | 29.02 | <2e-16 |
| Mosuus | 0.3410067 | 0.0314534 | 10.84 | <2e-16 |
| Selitysaste: 72.33 % | | | | |
| Residuaalin keskivirhe: 0.7238 | | | | |

Kertoimien tulkinnat ovat: LPM:n kasvusta kymmenellä seuraa, että tulos kasvaa 0.15 euroa. Pohjoisen hakkuilla tulos on 0.32 euroa pienempi kuin etelän. Keskipituuden kasvu yhdellä lisää tulosta 0.1 euroa. Männyn osuuden kasvu 10 %:lla kasvattaa tulosta 0.03 euroa.

**Taulukko 5.16** Yhteenveto kuusen läpimittaluokkien 270–420 regressioanalyyseistä

| **Malliyhtälöt:** |
|---|
| $\sqrt[3]{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ |
| $\sqrt[3]{\hat{y}} = 0.0463876 + 0.014663 * LPM - 0.321839 * Alue + 0.1010912 * Kpituus + 0.3410067 * Mosuus$ |
| $\hat{y} = \begin{bmatrix} 0.0463876 + 0.014663 * LPM - 0.321839 * Alue + 0.1010912 * Kpituus \\ +0.3410067 * Mosuus \end{bmatrix}^3$ |
| **Ennustevälit:** |
| 90 %: $\hat{y} = \begin{bmatrix} 0.0463876 + 0.014663 * LPM - 0.321839 * Alue + 0.1010912 * Kpituus \\ +0.3410067 * Mosuus \pm 1.645 * 0.7238 \end{bmatrix}^3$ |
| 95 %: $\hat{y} = \begin{bmatrix} 0.0463876 + 0.014663 * LPM - 0.321839 * Alue + 0.1010912 * Kpituus \\ +0.3410067 * Mosuus \pm 1.96 * 0.7238 \end{bmatrix}^3$ |
| 99 %: $\hat{y} = \begin{bmatrix} 0.0463876 + 0.014663 * LPM - 0.321839 * Alue + 0.1010912 * Kpituus \\ +0.3410067 * Mosuus \pm 2.576 * 0.7238 \end{bmatrix}^3$ |

# 6 Apteeraustulosten vertailu

Koska tutkimuksen tarkoituksena on vertailla suoritettuja apteerauksia, on tässä luvussa esimerkin kautta suoritettu vertailu edellisessä luvussa rakennettujen mallien sekä aineistosta laskettujen keskiarvojen ja kvantiilien avulla. Esimerkissä on valittu sattumanvarainen työmaa ja tälle suoritettu vertailu. Sama voidaan toistaa mille tahansa aineiston kanssa samoista lähtökohdista otetulle työmaalle.

Toteutuneiden apteeraustulosten vertailu suoritetaan yksittäisille työmaille seuraavalla tavalla:

- ✓ valitaan tarkasteltava työmaa,
- ✓ lasketaan ennusteet lopullisten regressiomallien avulla,
- ✓ lasketaan keskiarvot ja kvantiilit,
- ✓ verrataan toteutuneita tuloksia keskiarvoihin, kvantiileihin ja ennusteisiin sekä
- ✓ tehdään johtopäätökset.

Tarkastellaan käytännön esimerkin avulla, kuinka vertailu tehdään. Valitaan ensin sattumanvaraisesti tarkasteltavaksi jokin työmaa. Valituksi osui työmaa, jonka korjuunumero on 32367. Seuraavaksi lasketaan saatujen regressiomallien avulla estimaatit tulokselle sekä männyn että kuusen tapauksessa. Koska jotkut muuttujat saavat eri arvot riippuen siitä, onko kysymyksessä mänty vai kuusi, niin esitetään korjuunumeron muuttujien arvot erikseen molemmissa tapauksissa. Ensin tarkastellaan mäntyä. Aineiston muuttujat saavat nyt seuraavat arvot: DBAM=288.62, Hakkuu=15, Mosuus=0.1300905, Mtvl=13, PT=0, RP=0, HT=0, SM=0, Pituuslkt=7, M1=370, Minlpm=150, PituuslktK=8, M1K=200, MinlpmK=70, Kpituus=19.18184, MLosuus=0.004152183 ja alue=etelä. Aineistossa on lisäksi tulokset läpimittaluokittain ja ne esitetään samalla, kun regressiomallin mukaiset tuloksetkin.

## 6.1 Männyn läpimittaluokat 70–160

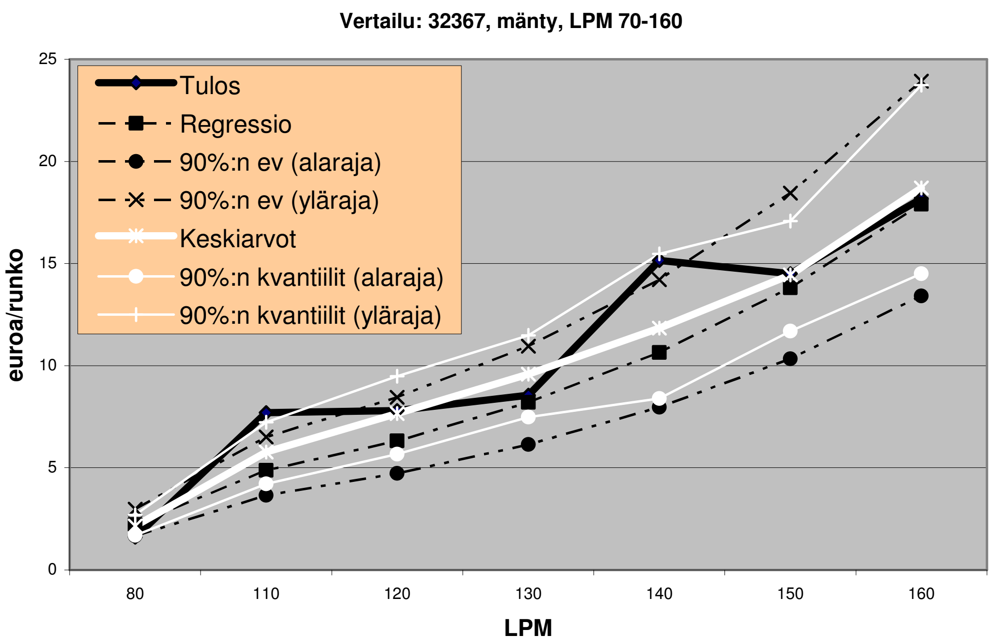
Ensimmäiseksi tarkastellaan männyn läpimittaluokkia 70–160. Regressioyhtälö on tällöin muotoa $\log(\hat{y}) = -1.281 + 0.02604 * \mathrm{LPM}$. Kaava tuloksen estimaatille saadaan ottamalla edellisestä yhtälöstä luonnollisen logaritmin käänteisoperaatio puolittain, jolloin se tulee muotoon $\hat{y} = \exp(-1.281 + 0.02604 * \mathrm{LPM})$. Kaavan avulla lasketaan estimaatit eri läpimittaluokille sijoittamalla siihen tarkasteltavan työmaan läpimittaluokat 70–160 yksi kerrallaan. Regressiomallin mukainen 90 %:n ennusteväli lasketaan käyttäen taulukon 5.8 vastaavaa yhtälöä. Lopuksi lasketaan vielä keskiarvot ja kvantiilit työmailta, jotka sijaitsevat etelässä ja joille on toteutettu hakkuu tavalla 15.

Taulukossa 6.1 on esitelty läpimittaluokittain työmaan toteutuneet tulokset (Tulos), regressiomallin mukaiset tulokset (Ennuste), 90 %:n ennusteväli (Ev90A ja Ev90Y), keskiarvot (Karvo) sekä 90 %:n kvantiilit (q05 ja q95). Läpimittaluokissa 70, 90 ja 100 kyseisellä työmaalla ei ole ollut tuloksia tai ne ovat olleet virheellisiä. Toisessa sarakkeessa on toteutunut tulos. Kolmannessa sarakkeessa on regressiomallin mukainen tuloksen estimaatti. Sarake A kertoo, onko toteutunut tulos suurempi (+) vai pienempi (-) kuin estimaatti. Sarake B kertoo, sijoittuuko toteutunut tulos kyseiselle luottamusvälille (=) vai ei (- tai +). Sarake C kertoo onko toteutunut tulos pienempi vai suurempi kuin laskettu keskiarvo. Sarake D kuvastaa sitä, sijoittuuko toteutunut tulos 90 %:n kvantiilivälille vai ei. Taulukon viimeinen sarake n ilmaisee läpimittaluokittain työmaata vastaavien aineistossa olevien tulosten lukumäärän.

**Taulukko 6.1** Vertailutulokset männyn läpimittaluokille 70–160

| LPM | Tulos | Ennuste | A | Ev90A | Ev90Y | B | Karvo | C | q05 | q95 | D | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 80 | 1,60 | 2,23 | - | 1,67 | 2,98 | - | 2,15 | - | 1,70 | 2,69 | - | 59 |
| 110 | 7,70 | 4,87 | + | 3,65 | 6,51 | + | 5,77 | + | 4,21 | 7,22 | + | 68 |
| 120 | 7,80 | 6,32 | + | 4,73 | 8,44 | = | 7,66 | + | 5,67 | 9,48 | = | 70 |
| 130 | 8,55 | 8,20 | + | 6,14 | 10,96 | = | 9,59 | - | 7,49 | 11,49 | = | 70 |
| 140 | 15,15 | 10,64 | + | 7,96 | 14,21 | + | 11,84 | + | 8,39 | 15,48 | = | 69 |
| 150 | 14,50 | 13,80 | + | 10,33 | 18,44 | = | 14,45 | + | 11,69 | 17,07 | = | 68 |
| 160 | 18,17 | 17,91 | + | 13,41 | 23,93 | = | 18,70 | - | 14,50 | 23,73 | = | 68 |

Kuviossa 6.1 esitetään taulukon 6.1 mukaiset tulokset graafisessa muodossa.




**Kuvio 6.1** Vertailutulokset graafisesti esitettynä männyn läpimittaluokille 70–160

Tulkittaessa taulukkoa 6.1, huomataan, että läpimittaluokan 80 tulos ei sijoitu 90 %:n luottamusvälille vaan on pienempi kuin alarajan arvo. Voidaan siis todeta, että regressioanalyysin mukaan kyseinen tulos on huonompi kuin 90 % odotetuista tuloksista. Vastaavalla tavalla voidaan taulukko käydä kokonaisuudessaan lävitse, mutta kaikkea taulukon sisältämää informaatiota ei ole tähän syytä purkaa.

Kuviosta 6.1 voimme havaita, että keskiarvo- ja kvantiilikäyrät ovat hyvin lähellä regressioanalyysin tuottamien tulosten avulla muodostettuja käyriä. Keskiarvot kuitenkin antavat lisäinformaatiota lopullisten johtopäätösten vetämistä varten.

## 6.2 Männyn läpimittaluokat 170–420

Seuraavaksi siirrytään tutkimaan männyn läpimittaluokkia 170–420 eli ns. tukkiluokkia. Regressioyhtälö on nyt muotoa

$$\log(\hat{y}) = 0.8189 + 0.01054 * \mathrm{LPM} + 0.2148 * \mathrm{FH12} + 0.1286 * \mathrm{FH15} + 0.1425 * \mathrm{FH17} + 0.06826 * \mathrm{Kpituus}.$$

Estimaatti tulokselle saadaan yhtälön

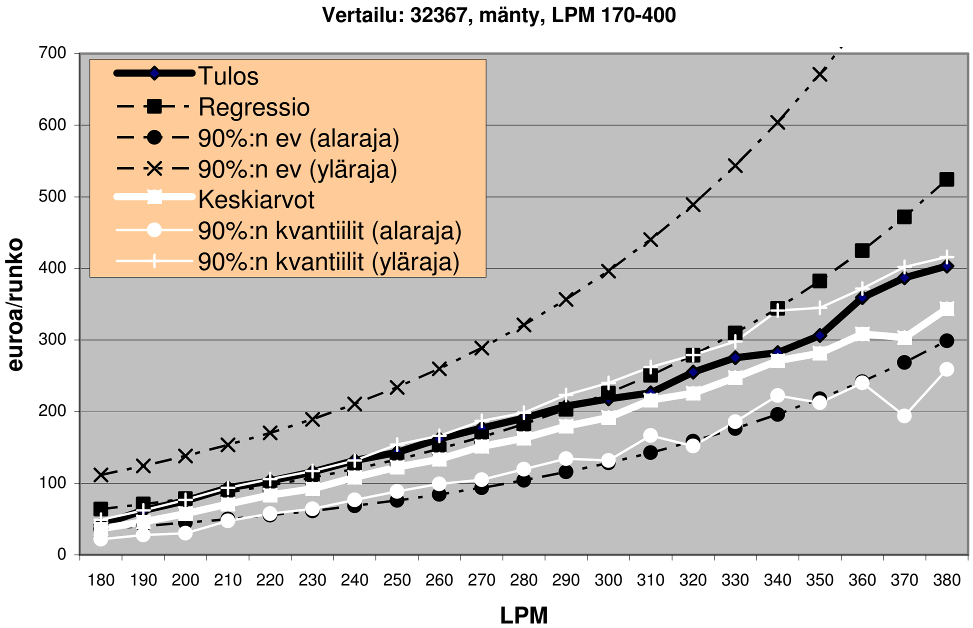
$$\hat{y} = \exp(0.8189 + 0.01054 * \mathrm{LPM} + 0.2148 * \mathrm{FH12} + 0.1286 * \mathrm{FH15} + 0.1425 * \mathrm{FH17} + 0.06826 * \mathrm{Kpituus})$$

avulla. Lasketaan tulokselle estimaatit ja ennustevälit aiemmin esitetyllä tavalla. Tulokset on esitelty taulukossa 6.2. Sarake A ilmaisee, että läpimittaluokissa 180–210 ja 300–380 tulos on ollut ennustetta pienempi ja läpimittaluokissa 220–290 suurempi. Sarakkeesta B voidaan havaita, että tutkittavien läpimittaluokkien tulosten arvot sijoittuvat kaikki 90 %:n ennustevälille, joten tulosten voidaan olettaa olevan odotusten rajoissa. Verratessa toteutuneita tuloksia läpimittaluokittaisiin keskiarvoihin, havaitaan, että jokaisessa läpimittaluokassa tulos on ollut keskimääräistä parempi. Tulokset sijaitsevat yhtä lukuun ottamatta 90 %:n kvantiilivälillä.

Kuviossa 6.2 on edellinen taulukko esitetty graafisessa muodossa. Graafisen esityksen etuna on, että yhdellä nopealla vilkaisulla saadaan hyvä yleiskuva vertailusta. Kuviosta voidaan havaita, että toteutuneet tulokset keskittyvät keskiarvojen ja kvantiilien perusteella regressiomallin tuottamiin tuloksiin nähden suhteellisen matalalle tasolle.

**Taulukko 6.2** Vertailutulokset männyn läpimittaluokille 170–400

| LPM | Tulos | Ennuste | A | Ev90A | Ev90Y | B | Karvo | C | q05 | q95 | D | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 180 | 41,87 | 63,69 | - | 36,28 | 111,83 | = | 36,49 | + | 22,00 | 49,26 | = | 69 |
| 190 | 61,91 | 70,77 | - | 40,31 | 124,26 | = | 46,29 | + | 27,57 | 62,19 | = | 67 |
| 200 | 75,35 | 78,64 | - | 44,79 | 138,07 | = | 57,46 | + | 30,03 | 76,59 | = | 71 |
| 210 | 93,64 | 87,38 | - | 49,77 | 153,42 | = | 70,74 | + | 47,56 | 93,77 | = | 71 |
| 220 | 103,33 | 97,09 | + | 55,30 | 170,47 | = | 83,62 | + | 57,43 | 105,96 | = | 70 |
| 230 | 115,80 | 107,88 | + | 61,44 | 189,42 | = | 92,19 | + | 64,18 | 116,96 | = | 70 |
| 240 | 131,91 | 119,88 | + | 68,27 | 210,48 | = | 108,18 | + | 76,73 | 131,83 | + | 70 |
| 250 | 143,06 | 133,20 | + | 75,86 | 233,87 | = | 122,40 | + | 88,72 | 153,88 | = | 70 |
| 260 | 162,03 | 148,01 | + | 84,30 | 259,87 | = | 133,43 | + | 98,83 | 166,40 | = | 68 |
| 270 | 176,95 | 164,46 | + | 93,67 | 288,75 | = | 151,53 | + | 105,05 | 187,04 | = | 70 |
| 280 | 190,99 | 182,74 | + | 104,08 | 320,85 | = | 162,81 | + | 119,60 | 198,44 | = | 68 |
| 290 | 207,91 | 203,05 | + | 115,65 | 356,51 | = | 179,78 | + | 134,41 | 223,13 | = | 71 |
| 300 | 217,67 | 225,62 | - | 128,50 | 396,14 | = | 191,63 | + | 131,80 | 239,94 | = | 66 |
| 310 | 226,21 | 250,70 | - | 142,78 | 440,17 | = | 215,92 | + | 166,73 | 262,35 | = | 68 |
| 320 | 255,05 | 278,57 | - | 158,66 | 489,10 | = | 225,77 | + | 151,99 | 279,27 | = | 64 |
| 330 | 275,27 | 309,53 | - | 176,29 | 543,47 | = | 247,73 | + | 185,85 | 298,02 | = | 65 |
| 340 | 282,28 | 343,93 | - | 195,89 | 603,88 | = | 271,24 | + | 222,63 | 341,17 | = | 58 |
| 350 | 306,20 | 382,16 | - | 217,66 | 671,00 | = | 281,66 | + | 212,39 | 345,11 | = | 60 |
| 360 | 359,40 | 424,64 | - | 241,85 | 745,59 | = | 308,32 | + | 240,14 | 371,65 | = | 57 |
| 370 | 387,10 | 471,85 | - | 268,74 | 828,46 | = | 303,53 | + | 193,83 | 402,11 | = | 48 |
| 380 | 403,20 | 524,29 | - | 298,61 | 920,55 | = | 343,57 | + | 258,93 | 415,76 | = | 41 |




**Kuvio 6.2** Vertailutulokset graafisesti esitettynä männyn läpimittaluokille 70–160

Suoritetaan kuuselle sama kuin männylle edellä. Aineiston muuttujat saavat nyt seuraavat arvot: DBAK=287.613, Hakkuu=15, Mosuus=0.1300905, Mtvl=13, KPT=0, KET=1, KSE=1, Pituuslkt=9, M1=370, Minlpm=160, PituuslktK=3, M1K=280, MinlpmK=70, Kpituus=14.57274, KLosuus=0.01976457 ja Alue=etelä.

## 6.3 Kuusen läpimittaluokat 70–160

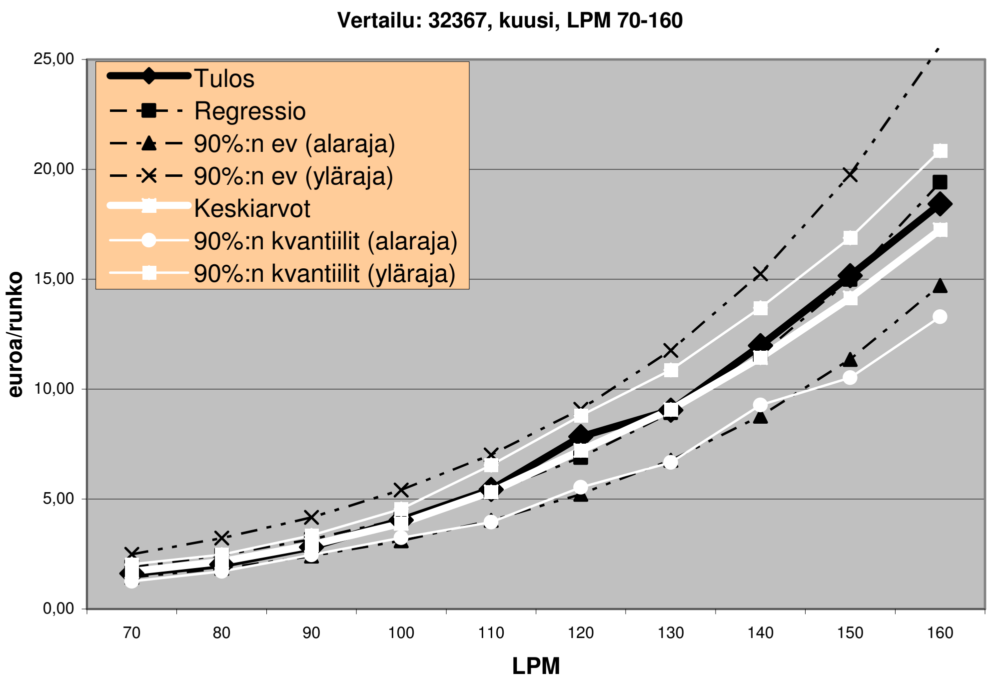
Ensin käsittelyyn otetaan kuusen läpimittaluokat 70–160 eli kuituluokat. Regressio-yhtälö on muotoa

$\log\left(\hat{y}\right) = -1.17 + 0.02592 * LPM - 0.08266 * Alue - 0.08683 * Mosuus$ ja ennuste sekä ennusteväli lasketaan taulukon 5.1 kaavojen avulla.

Taulukko 6.3 ja kuvio 6.3 pitävät sisällään vertailusta saadun informaation. Lyhyenä yhteenvetona voidaan todeta, että kaikki toteutuneet tulokset sisältyvät sekä 90 %:n ennustevälille että kvantiilivälille, joten näillä kriteereillä niiden voidaan olettaa olevan kunnossa.

**Taulukko 6.3** Vertailutulokset kuusen läpimittaluokille 70–160

| LPM | Tulos | Ennuste | A | Ev90A | Ev90Y | B | Karvo | C | q05 | q95 | D | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 70 | 1,61 | 1,88 | - | 1,49 | 2,38 | = | 1,71 | - | 1,26 | 2,04 | = | 62 |
| 80 | 2,02 | 2,44 | - | 1,93 | 3,08 | = | 2,16 | - | 1,71 | 2,47 | = | 68 |
| 90 | 2,81 | 3,16 | - | 2,51 | 3,99 | = | 2,91 | - | 2,46 | 3,35 | = | 69 |
| 100 | 4,05 | 4,10 | - | 3,25 | 5,17 | = | 3,89 | + | 3,24 | 4,55 | = | 71 |
| 110 | 5,44 | 5,31 | + | 4,21 | 6,70 | = | 5,32 | + | 3,95 | 6,54 | = | 71 |
| 120 | 7,85 | 6,88 | + | 5,46 | 8,68 | = | 7,21 | + | 5,54 | 8,79 | = | 71 |
| 130 | 9,04 | 8,92 | + | 7,07 | 11,25 | = | 9,07 | - | 6,66 | 10,88 | = | 71 |
| 140 | 11,99 | 11,56 | + | 9,16 | 14,58 | = | 11,44 | + | 9,27 | 13,70 | = | 69 |
| 150 | 15,17 | 14,98 | + | 11,87 | 18,90 | = | 14,15 | + | 10,52 | 16,89 | = | 72 |
| 160 | 18,43 | 19,41 | - | 15,39 | 24,49 | = | 17,25 | + | 13,30 | 20,85 | = | 70 |




**Kuvio 6.3** Vertailutulokset graafisesti esitettynä kuusen läpimittaluokille 70–160

## 6.4 Kuusen läpimittaluokat 170–260

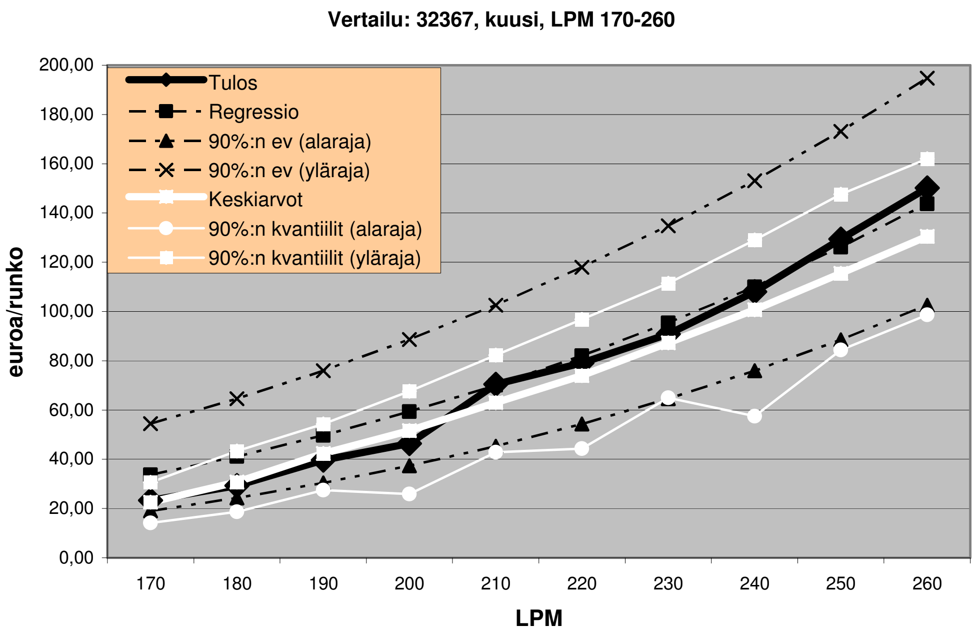
Siirrytään tutkimaan kuusen läpimittaluokkia 170–260. Regressioyhtälö on

$$\sqrt[3]{\hat{y}} = -1.54578 + 0.022306 * LPM - 0.125775 * Alue + 0.065965 * Kpituus$$
$$-6.833098 * KLosuus + 0.195181 * FH12 + 0.158485 * FH15 + 0.168622 * FH17$$

ja vertailutulokset sekä ennusteväli lasketaan taulukon 5.14 kaavojen avulla.

**Taulukko 6.4** Vertailutulokset kuusen läpimittaluokille 170–260

| LPM | Tulos | Ennuste | A | Ev90A | Ev90Y | B | Karvo | C | q05 | q95 | D | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 23,34 | 33,73 | - | 21,08 | 50,62 | = | 22,56 | + | 14,20 | 30,62 | = | 71 |
| 180 | 29,26 | 41,21 | - | 26,62 | 60,34 | = | 30,72 | - | 18,69 | 43,34 | = | 70 |
| 190 | 39,68 | 49,72 | - | 33,04 | 71,24 | = | 42,36 | - | 27,50 | 54,29 | = | 70 |
| 200 | 46,36 | 59,33 | - | 40,42 | 83,37 | = | 51,60 | - | 25,85 | 67,68 | = | 71 |
| 210 | 70,51 | 70,10 | + | 48,82 | 96,80 | = | 62,98 | + | 42,84 | 82,28 | = | 70 |
| 220 | 79,18 | 82,10 | - | 58,32 | 111,60 | = | 74,01 | + | 44,31 | 96,81 | = | 71 |
| 230 | 90,87 | 95,40 | - | 68,97 | 127,85 | = | 87,32 | + | 65,00 | 111,42 | = | 70 |
| 240 | 108,06 | 110,07 | - | 80,85 | 145,59 | = | 100,77 | + | 57,53 | 129,11 | = | 70 |
| 250 | 129,40 | 126,16 | + | 94,02 | 164,91 | = | 115,53 | + | 84,38 | 147,59 | = | 70 |
| 260 | 150,14 | 143,75 | + | 108,54 | 185,86 | = | 130,53 | + | 98,69 | 162,00 | = | 67 |




**Kuvio 6.3** Vertailutulokset graafisesti esitettynä kuusen läpimittaluokille 170–260

Tulokset on esitetty numeerisesti taulukossa 6.4 ja graafisesti kuviossa 6.4. Toteutuneet tulokset sijoittuvat sekä luottamusvälille että kvantiilivälille, joten voimme todeta tuloksen olevan kunnossa.

## 6.5 Kuusen läpimittaluokat 270–420

Viimeiseksi suoritetaan vertailut kuusen läpimittaluokille 270–420. Regressioyhtälö on

$$\sqrt[3]{\hat{y}} = 0.0463876 + 0.014663 * LPM - 0.321839 * Alue + 0.1010912 * Kpituus + 0.3410067 * Mosuus,$$

jolloin regressioyhtälön antama estimaatti tulokselle lasketaan seuraavasti:

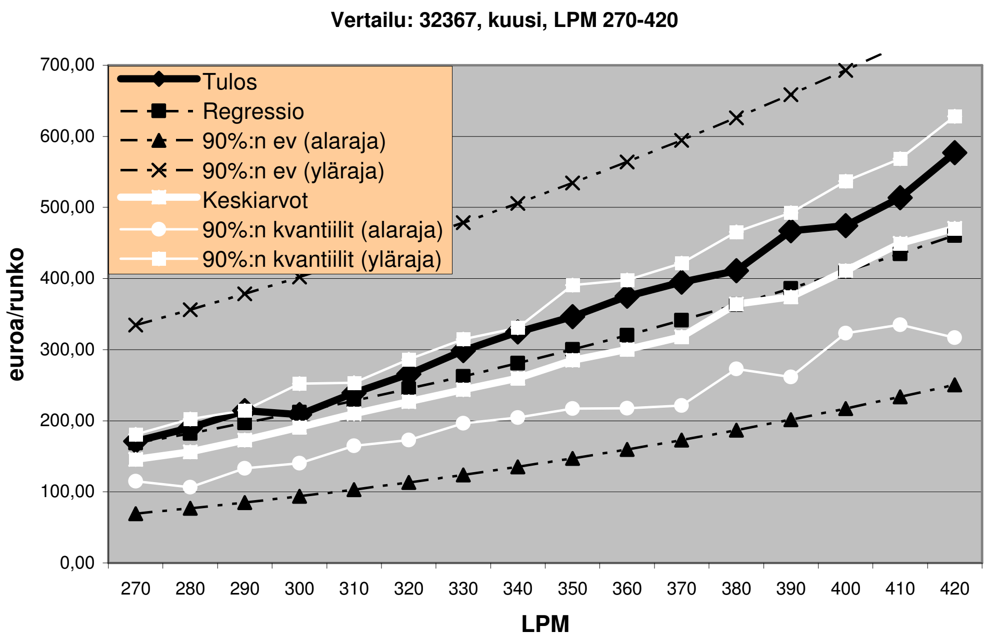
$$\hat{y} = \begin{bmatrix} 0.0463876 + 0.014663 * LPM - 0.321839 * Alue + 0.1010912 * Kpituus \\ +0.3410067 * Mosuus \end{bmatrix}^3.$$

Ennusteväli lasketaan taulukon 5.16 kaavan avulla. Taulukossa 6.5 on esitetty saadut tulokset. Kaikki tulokset sijoittuvat sekä ennuste- että kvantiilivälille. Tällä kertaa huomion arvoista on, että kaikissa muissa läpimittaluokissa paitsi luokassa 300, toteutunut tulos on suurempi kuin regressioyhtälön antama estimaatti. Keskiarvot ovat jokaisessa läpimittaluokassa pienemmät kuin toteutuneet tulokset, joten voidaan todeta, että työmaan tulos on onnistunut keskimääräistä paremmin.

**Taulukko 6.5** Vertailutulokset kuusen läpimittaluokille 270–420

| LPM | Tulos | Ennuste | A | Ev90A | Ev90Y | B | Karvo | C | q05 | q95 | D | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 270 | 171,11 | 168,47 | + | 81,31 | 302,60 | = | 145,66 | + | 114,99 | 181,00 | = | 65 |
| 280 | 189,93 | 182,24 | + | 89,85 | 322,86 | = | 156,20 | + | 106,51 | 202,26 | = | 68 |
| 290 | 214,16 | 196,75 | + | 98,97 | 344,01 | = | 172,89 | + | 132,90 | 214,47 | = | 66 |
| 300 | 208,57 | 212,01 | - | 108,68 | 366,06 | = | 190,91 | + | 140,10 | 252,01 | = | 64 |
| 310 | 239,07 | 228,04 | + | 119,01 | 389,03 | = | 210,16 | + | 164,78 | 253,42 | = | 67 |
| 320 | 265,89 | 244,85 | + | 129,97 | 412,95 | = | 226,83 | + | 172,90 | 286,27 | = | 63 |
| 330 | 298,53 | 262,48 | + | 141,59 | 437,83 | = | 243,95 | + | 196,59 | 314,52 | = | 61 |
| 340 | 324,70 | 280,93 | + | 153,88 | 463,68 | = | 259,74 | + | 204,27 | 330,74 | = | 58 |
| 350 | 346,21 | 300,22 | + | 166,86 | 490,54 | = | 285,25 | + | 217,13 | 390,57 | = | 57 |
| 360 | 375,04 | 320,38 | + | 180,55 | 518,41 | = | 300,42 | + | 217,41 | 398,01 | = | 56 |
| 370 | 394,88 | 341,42 | + | 194,97 | 547,32 | = | 318,10 | + | 221,44 | 421,52 | = | 55 |
| 380 | 410,84 | 363,36 | + | 210,14 | 577,28 | = | 363,95 | + | 272,86 | 465,44 | = | 51 |
| 390 | 467,12 | 386,23 | + | 226,07 | 608,32 | = | 374,35 | + | 261,42 | 492,73 | = | 41 |
| 400 | 474,49 | 410,03 | + | 242,79 | 640,45 | = | 411,23 | + | 322,97 | 536,79 | = | 45 |
| 410 | 513,83 | 434,79 | + | 260,32 | 673,70 | = | 449,69 | + | 334,89 | 568,81 | = | 35 |
| 420 | 577,03 | 460,53 | + | 278,67 | 708,07 | = | 470,59 | + | 316,78 | 628,01 | = | 31 |

Kuviossa 6.5 esitetään tulokset vielä graafisesti.




**Kuvio 6.4** Vertailutulokset graafisesti esitettynä kuusen läpimittaluokille 270–420

## 6.6 Yhteenveto vertailuista

Yhteenvetona voidaan todeta, että kyseisen työmaan tulokset sijaitsevat sekä männyn että kuusen tapauksessa regressiomallien tuottamien ennustevälien ja 90 %:n kvantiilivälien sisällä. Johtopäätöksenä voidaan siis todeta, että mikäli työmaan tavoitetasot olisi asetettu ennuste- ja 90 %:n kvantiilivälien perusteella, työmaan tulos olisi tavoitteiden mukainen. Yleisesti ottaen työmaan tulokset ovat keskimääräistä paremmat, kun tarkastellaan toteutuneita tuloksia suhteessa vastaavien työmaiden tuottamiin keskiarvoihin. Verratessa regressioennusteita ja keskiarvoja, huomataan, että erityisesti männyn kaikkien ja kuusen tukkiluokkien 170–260 tuloksissa on jonkin verran parantamisen varaa, kun kyseessä on hakkutapa 15 ja alue etelä. Laajemman vertailukokonaisuuden suorittaminen antaisi varmasti arvokasta lisätietoa tutkimuksen jatkokehittelyä varten, mutta tässä ei ole mahdollista lähteä siihen.

# 7 Johtopäätökset ja suositukset

Uutta apteeraustulosta voidaan verrata aiemmin toteutuneisiin monella eri tavalla. Tässä tutkimuksessa muodostettiin ensin vertailupohja mahdollisimman yksinkertaisella tavalla eli laskettiin havaintoaineistosta muutamat läpimittaluokittaiset tunnusluvut, joihin uusia tuloksia on mahdollista verrata. Menetelmän etuna on sen yksinkertaisuus ja puutavaralajien arvojen muunneltavuus. Heikkoutena on, että pienillä havaintomäärillä yksittäiset havainnot pääsevät vaikuttamaan merkittävästi tunnuslukuihin. Erityisesti tämä voidaan havaita, kun lasketaan keskiarvoja aineistoille, joissa havaintoja on vähän. Tällöin ongelman ratkaisemiseksi on mahdollista korvata keskiarvo vertailulukuna mediaanilla. Alustavissa keskiarvo- ja kvantiilivertailuissa läpimittaluokittaiset tulokset on jaoteltu hyvin karkeasti eli ainoastaan kahteen osaan: etelään ja pohjoiseen. Perusteena on käytetty työmaan sijaintia. Vertailuja olisi mahdollista saada tarkemmiksi jakamalla aineisto useampiin toisensa pois sulkeviin ryhmiin siten, että ryhmien sisäiset hajonnat ovat pienemmät kuin kokonaishajonta. Näiden vertailujen perusteella sekä männyllä että kuusella, etelän tulokset ovat olleet jokaisessa läpimittaluokassa parempia.

Tutkimusmenetelmänä on käytetty lineaarista regressioanalyysia. Lähtökohtana tällöin on ollut saada aineistosta enemmän informaatiota irti kuin mitä yksinkertaisilla tunnusluvuilla on saatavissa. Koska tutkimuksen lähtökohtana oli saada muodostettua yksi tai korkeintaan muutama regressiomalli apteeraustulosten vertailemiseksi, on tutkielmassa jätetty esittämättä yksittäisille läpimittaluokille muodostetut mallit.

Niiden etuna oli virhetermin vakiovarianssisuus, mallien parametrien yksinkertaiset tulkinnat sekä ennustevälien symmetrisyys. Kuitenkin tällöin malleja tulisi yhtä monta kuin on läpimittaluokkia ja työmaan vertailun suorittaminen ei olisi niin yksinkertaisesti tulkittavissa kuin mitä se on vähemmällä mallien määrällä. Molemmissa on etunsa ja tehtäessä aiheesta alustavaa tutkimusta, on ehkä syytä käyttää regressiomalleja läpimittaluokkien ryhmille. Mallien käytäntöön soveltaminen tuottaa varmasti tietoa mallien jatkokehittämistä varten ja silloin tarvittaessa voidaan perusteellisesti rakentaa mallit eri läpimittaluokille, mikäli se nähdään tarpeelliseksi.

Regressioanalyysejä varten pudotettiin ensin pois läpimittaluokat 50 ja 60, koska niistä löytyi paljon virheitä. Mäntyaineisto jaettiin kahteen eri läpimittaluokkien muodostamaan osaan, joita kutsuttiin tukki- ja kuituluokiksi. Ensimmäiseen osaan eli kuituluokkiin kuuluivat luokat 70–160 ja toiseen eli tukkiluokkiin luokat 170–400. Esiin nousseen heteroskedastisuusongelman vuoksi aineistoon kokeiltiin erilaisia muunnoksia, joista logaritmointi osoittautui käyttökelpoisimmaksi. Myös painotetun neliösumman menetelmää kokeiltiin, mutta sen selitysaste jäi vasteen logaritmisen muunnoksen selitysastetta heikommaksi.

Regressioanalyysit kuituluokille tuottivat selitysasteeksi noin 95 %, kun logaritmoitua tulosta selitettiin pelkällä läpimittaluokalla. Tukkiluokille selitysasteeksi saatiin noin 84 %, kun logaritmoitua tulosta selitettiin läpimittaluokalla, keskipituudella sekä hakkuutavalla.

Kuusiaineisto jaettiin kolmeen eri läpimittaluokkien muodostamaan osaan. Ensimmäiseen tulivat läpimittaluokat 70–160, toiseen luokat 170–260 ja kolmanteen luokat 270–420. Regressioanalyysi kuituluokille tuotti selitysasteeksi noin 96 %, kun logaritmoitua tulosta selitettiin läpimittaluokalla, alueella ja mäntypuun osuudella leimikosta. Läpimittaluokille 170–260 selitysasteeksi saatiin noin 85 %, kun tuloksen kuutiojuurta selitettiin läpimittaluokalla, alueella, keskipituudella, lumppiosuudella sekä hakkuutavalla. Läpimittaluokille 270–420 selitysasteeksi saatiin noin 72 %, kun tuloksen kuutiojuurta selitettiin läpimittaluokalla, alueella, keskipituudella ja mäntypuun osuudella. Vakiokerroin ei ollut mallissa tilastollisesti merkittävä, mutta koska vakion pois jättämiselle ei ole teoreettisia perusteita, on syytä pitää se mukana mallissa. Tällöin selitysastetta ei tarvitse erikseen laskea ja voidaan käyttää ohjelman laskemaa arvoa.

Mallien selitysasteet olivat erittäin korkeita johtuen läpimittaluokan vahvasta selitysluonteesta. Läpimittaluokan suuri selitysvoima olikin odotettavissa, sillä on luonnollista, että rinnan korkeudelta mitatun rungon paksuuden kasvaessa tilavuus kasvaa ja rungosta saatavien eri puutavaralajien joukko laajenee. On tietysti myös selvää, että järeämmistä rungoista maksetaan yleensä enemmän, mikä myös vaikuttaa tulosta kasvattavasti. Huomioitavana erona oli, että männyn regressiomalleissa ei alue noussut niin merkittäväksi selittäjäksi, että se olisi syytä sisällyttää lopullisiin malleihin.

Säädeltävissä olevien parametrien (minimiläpimitat, pituusluokkien määrät, jne.) vaikutukset jäivät hyvin pieniksi eikä niitä ei lopullisiin malleihin selittäviksi muuttujiksi valikoitunut yhtään. Yhtenä syynä tähän varmasti on se, että muuttujien arvoja ei voitu sellaisinaan ottaa mukaan tutkimukseen, vaan niitä jouduttiin yleistämään.

Tulosten vertailuissa suosittelen käytettävän sekä regressioanalyysin tuloksia että keskiarvo- ja kvantiilivertailuita. Regressioanalyysin tuloksien heikkoutena on niiden muunneltavuus. Vertailussa voi olla tarpeen aika ajoin muunnella tavaralajien referenssihintoja, eikä se tällöin ole mahdollista ilman uusien regressiomallin rakentamista. Keskiarvo- ja kvantiilivertailut eivät ehkä kerro mahdollisesta optimista paljoakaan, mutta ajanevat tarkoituksensa yksinkertaisuutensa ja tavaralajien hintojen erittäin helpon muunneltavuuden ansiosta. Kun järjestelmässä on molemmat, epäilyttäviin tuloksiin voidaan puuttua varmemmalta pohjalta, kuten tutkimuksen vertailuosuudesta voidaan havaita.

Jatkotutkimusaiheena voisi aluksi olla laajempien vertailuiden suorittaminen ja niiden tulosten pohjalta sitten miettiä miten saada vertailuja tarkemmiksi. Tarpeen olisi myös selvittää kuinka muunneltavissa olevat parametrit saataisiin mukaan tutkimukseen siten, ettei niiden sisältämää informaatiota menetettäisi yhtä paljon kuin mitä tässä tutkimuksessa on menetetty. Myös tarkempien leimikkotietojen (puuston ikä, leimikon koko, yms.) mukaan ottaminen olisi varmasti hyödyllistä, koska tällöin saataisiin leimikoiden ryhmittelylle parempi pohja.

# Kirjallisuus


Chatterjee, Samprit & Hadi, Ali S. & Price, Bertram (2000), *Regression analysis by example,* 3rd edition, John Wiley & Sons, Inc.

Faraway, Julian J. (2002), *Practical Regression and Anova in R.* Kirja on saatavilla internet-osoitteesta: http://www.stat.lsa.umich.edu/~faraway/book/pra.pdf.

Korvenoja, Hannu & Nummi, Tapio (1994), *Sekamalli rungon kapeneman ennustamisessa metsätyökoneelle*, Tampereen yliopiston julkaisuja, C 55.

Lappi, Juha (1993), *Metsäbiometrian menetelmiä, Silva Carelia 24*, Joensuun yliopisto.

Liski, Erkki & Nummi, Tapio (1994), *Prediction of Tree Stems to Improve Efficiency in Automatized Harvesting of Forests*, Tampereen yliopiston julkaisuja, A 279.

Malinen, Jukka (2003), *Prediction of Characteristics of Marked Stand and Metrics for Similarity of Log Distribution for Wood Procurement Management*, Joensuun yliopisto.

Maltamo, Matti ja Laukkanen, Sanna (toim.), (2001) *Metsää kuvailevat mallit, Silva Carelia 36*, Joensuun yliopisto.

Nummi, Tapio (1998), *Estimation and Prediction for Stem Data Under Measurement Errors*, Tampereen yliopiston julkaisuja, A 316.

Nummi, Tapio (1998), *On Prediction of Stem Charasteristics for Pine Trees*, Tampereen yliopiston julkaisuja, A 313.

Nummi, Tapio & Sinha, Bikas K & Koskela, Laura (2004), *A Statistical Study of the Goodness of the Bucking Outcome*, Tampereen yliopiston julkaisuja, A 348.

Nummi, Tapio & Möttönen, Jyrki (2001), *Prediction of Stem Measurement of Scots Pine*, Tampereen yliopiston julkaisuja, A 338.

Puntanen, Simo (1999a) *Regressioanalyysi I*. Tampereen yliopisto, Matematiikan ja tilastotieteen ja filosofian laitos, B: 48. Tampereen yliopisto, Tampere.

Puntanen, Simo (1999b) *Regressioanalyysi II*. Tampereen yliopisto, Matematiikan ja tilastotieteen ja filosofian laitos. B: 49. Tampereen yliopisto, Tampere.

Salakari, Hannu & Heimonen, Raija (1998), *Koneellinen puunkorjuu*, Kustannusosakeyhtiö Metsälehti.

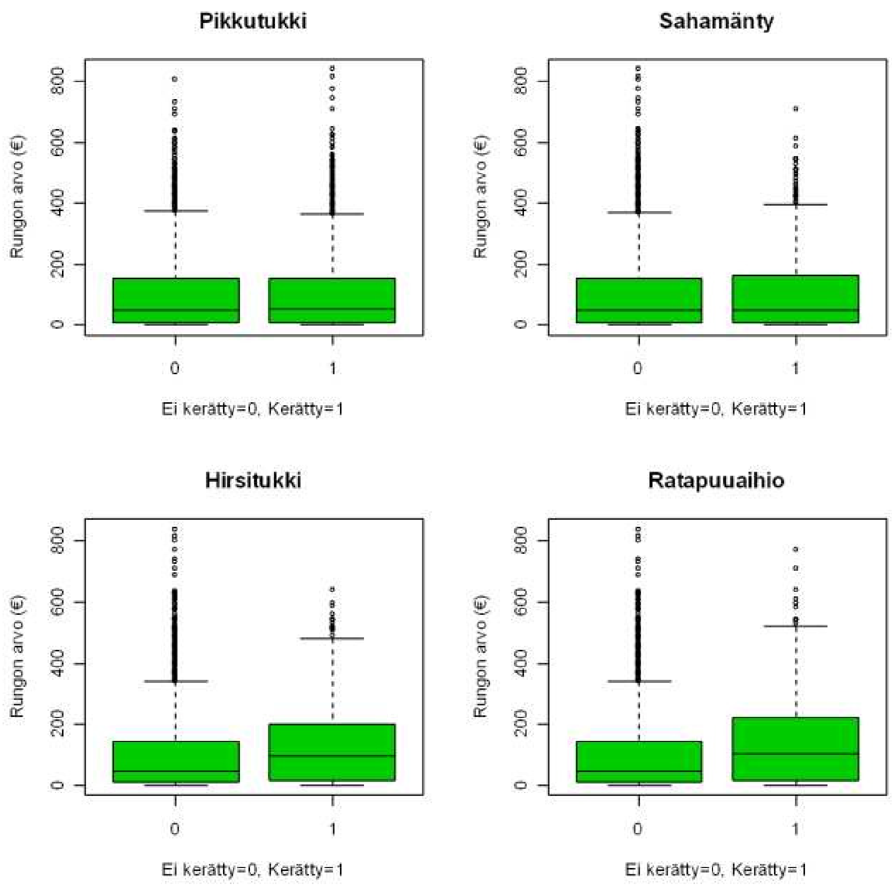
Seber, George A. F. & Lee, Alan J. (2003), *Linear regression analysis*, John Wiley & Sons, Inc.

*Tapion taskukirja* (2002), Metsätalouden kehittämiskeskus Tapio, Kustannusosakeyhtiö Metsälehti.

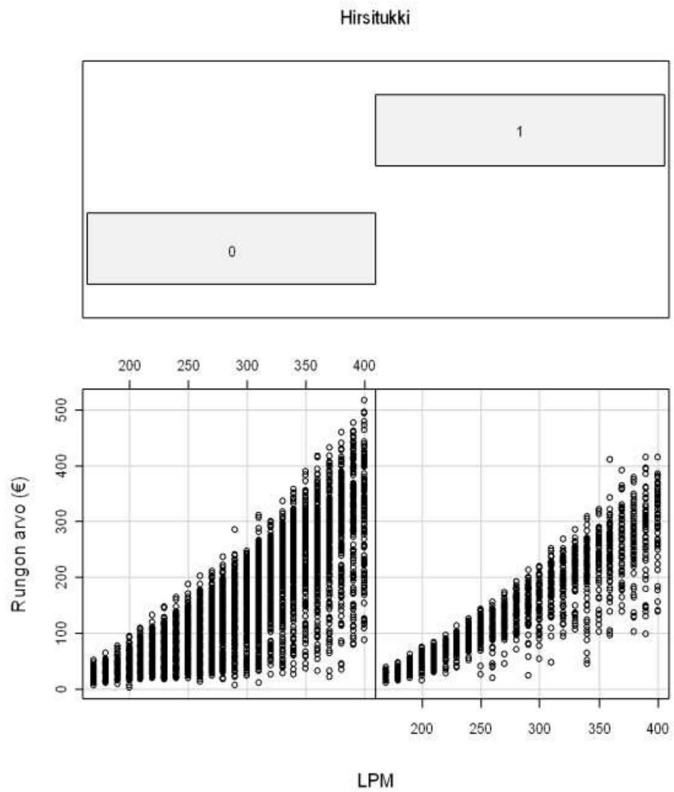
Uusitalo, Jori (2003), *Metsäteknologian perusteet*, Karisto Oy, Hämeenlinna.

# LIITEET

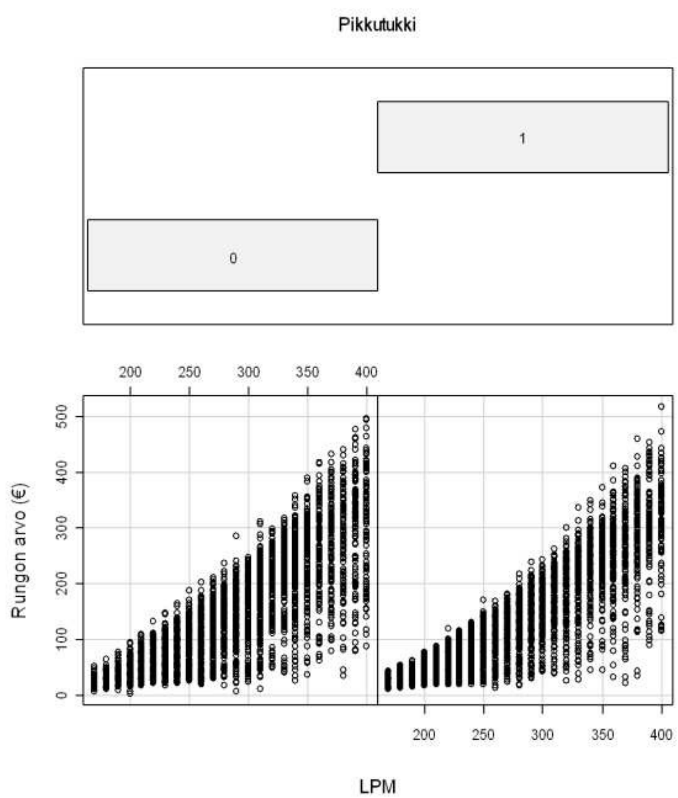
# Liite A: Kuviot



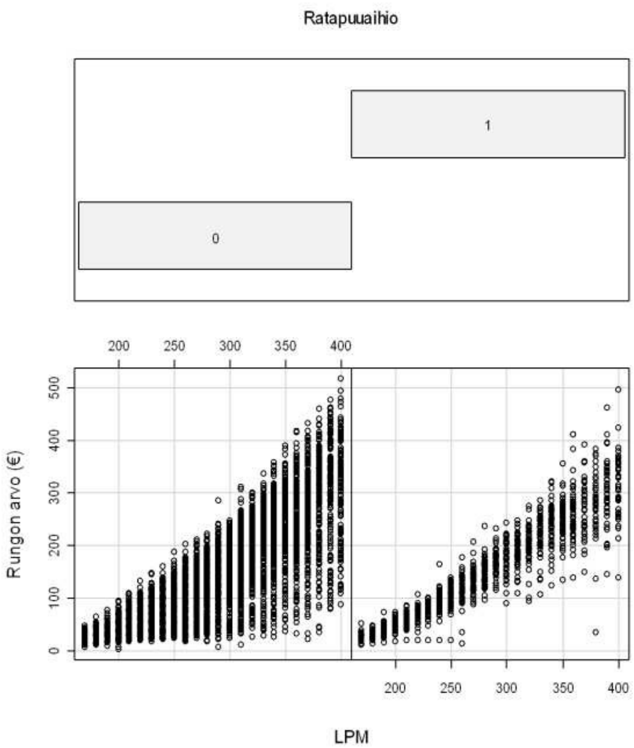

**Kuvio A.1** Tavaralajimuuttujien laatikko-jana –tarkastelu (mänty).



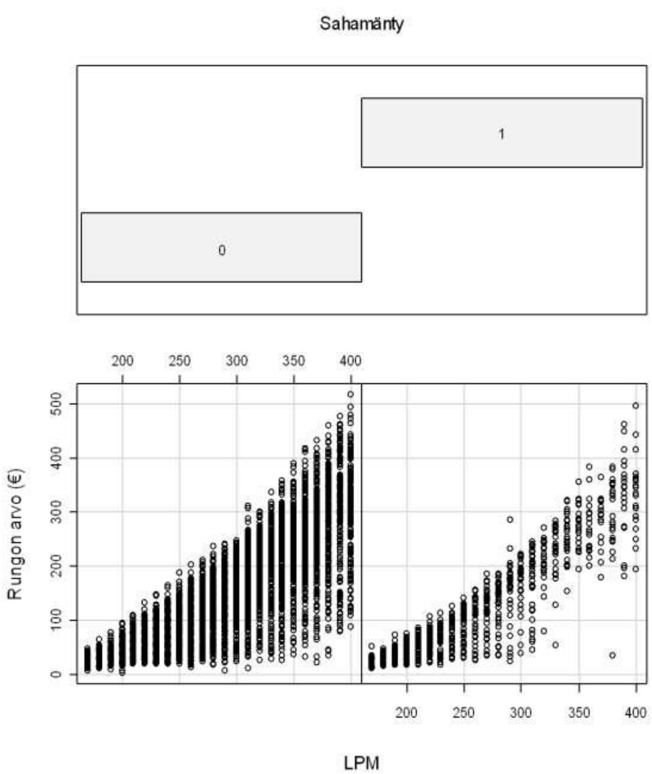

**Kuvio A.2** Tavaralajimuuttujan hirsitukki tulosjakauman tarkastelu (mänty).



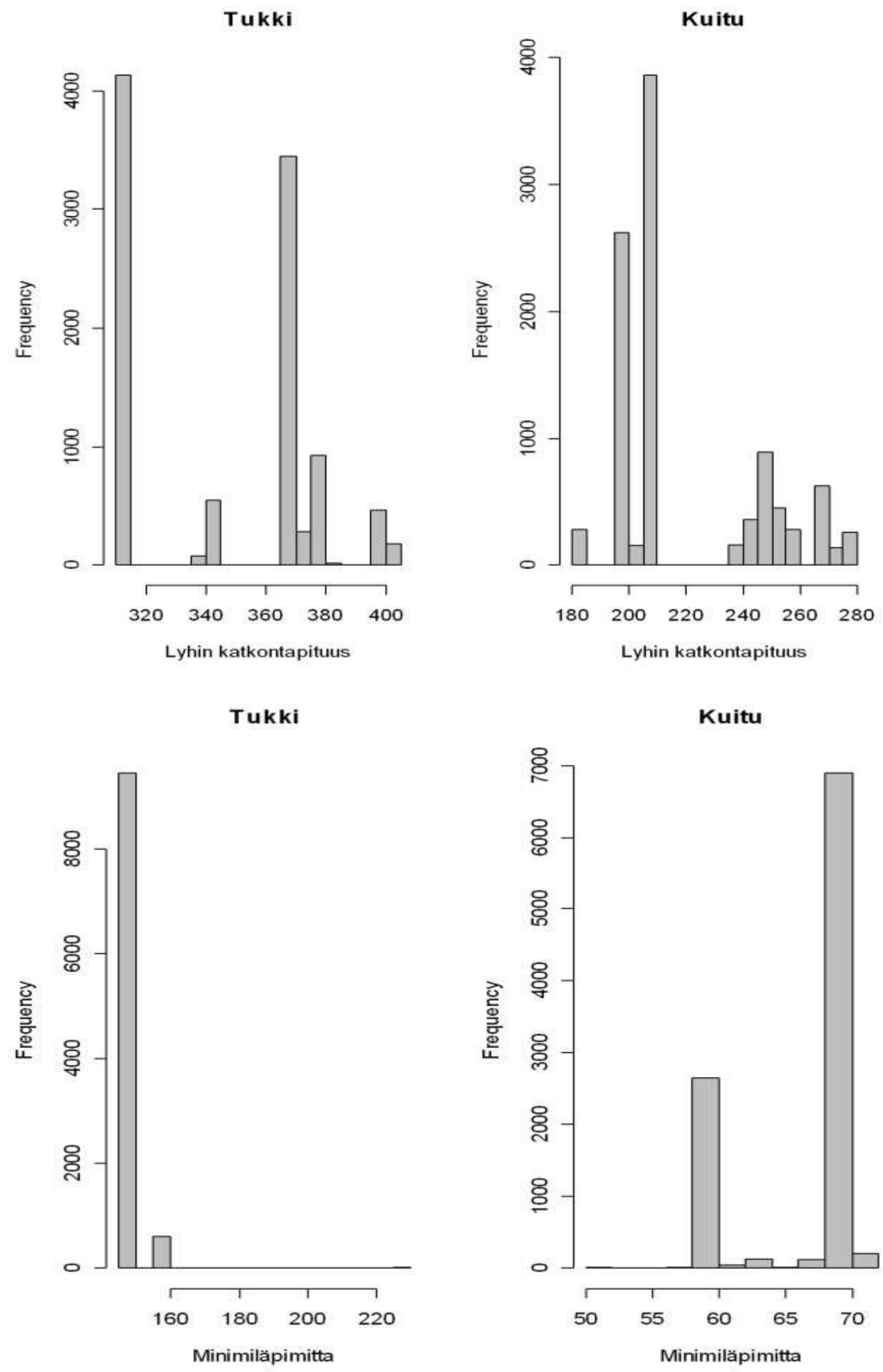

**Kuvio A.3** Tavaralajimuuttujan pikkutukki tulosjakauman tarkastelu (mänty).



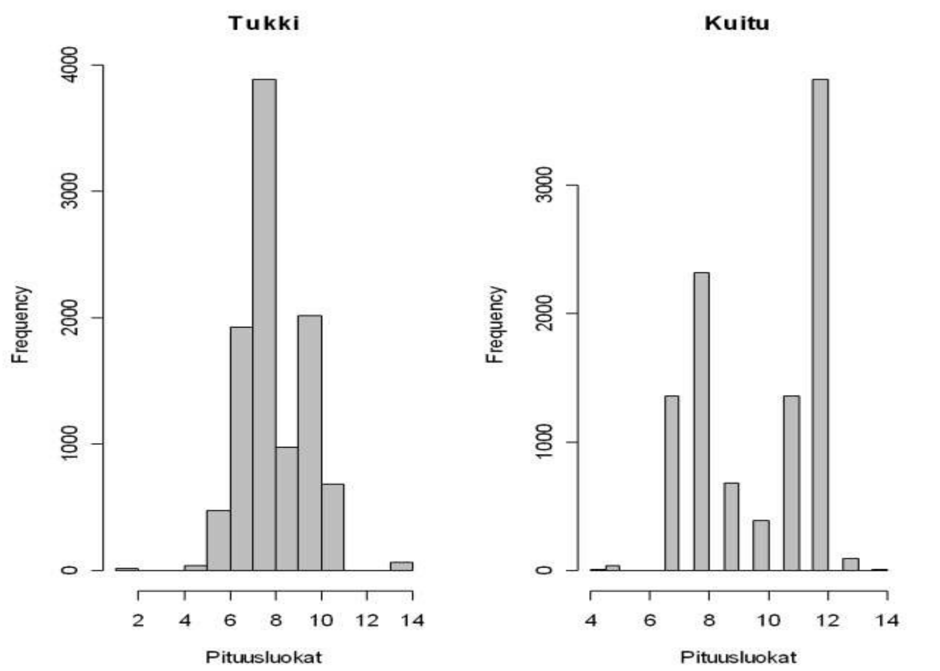

**Kuvio A.4** Tavaralajimuuttujan ratapuuaihio tulosjakauman tarkastelu (mänty).



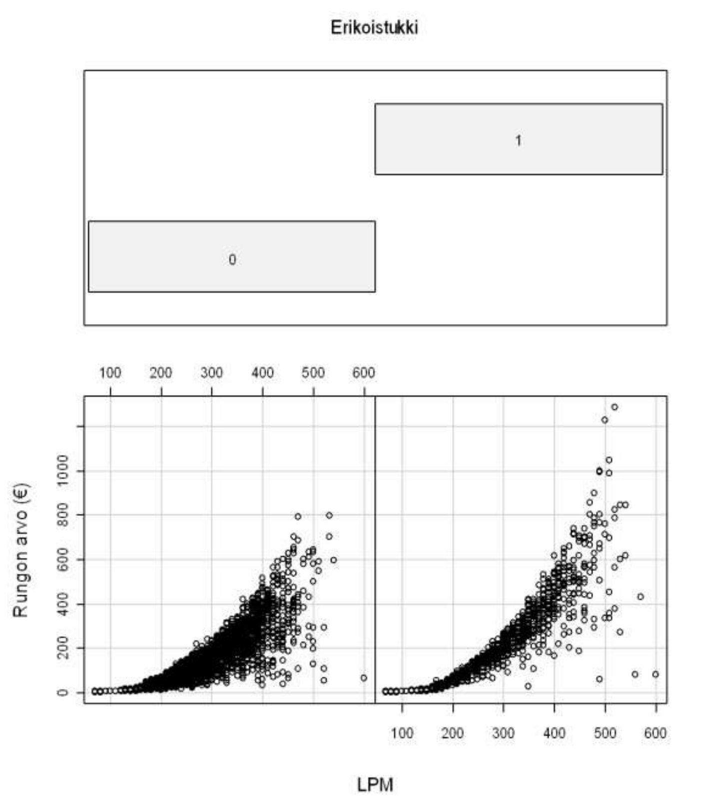

**Kuvio A.5** Tavaralajimuuttujien ratapuuaihio ja sahamänty tulosjakaumien tarkastelu (mänty).



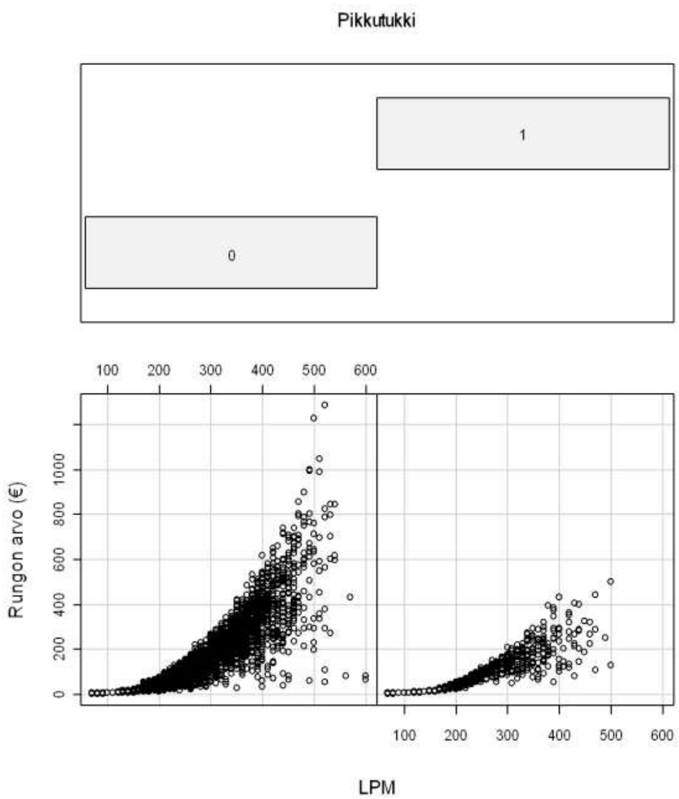

**Kuvio A.6** Lyhyimmän katkontapituuden ja minimiläpimitan histogrammit (mänty).



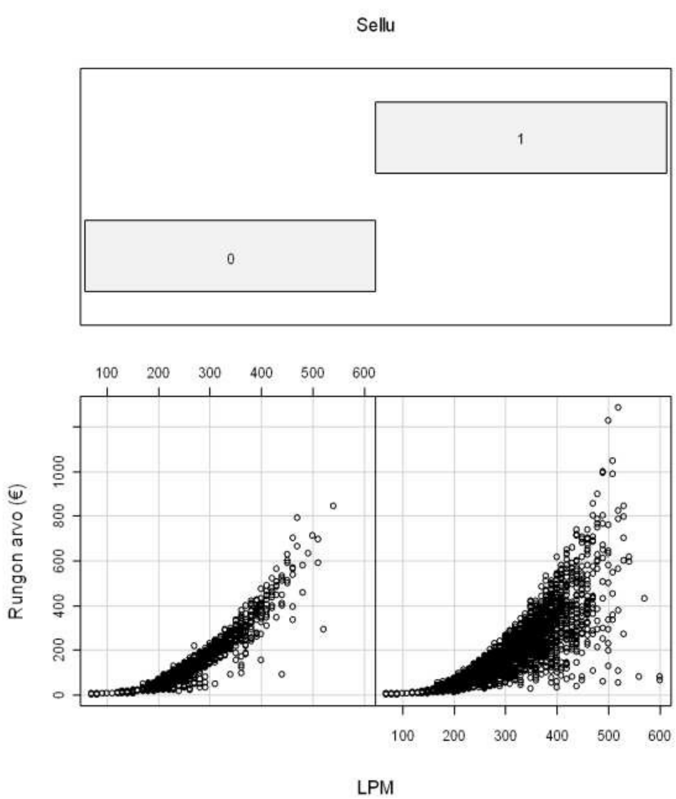

**Kuvio A.7** Pituusluokkien lukumäärien histogrammit (mänty).



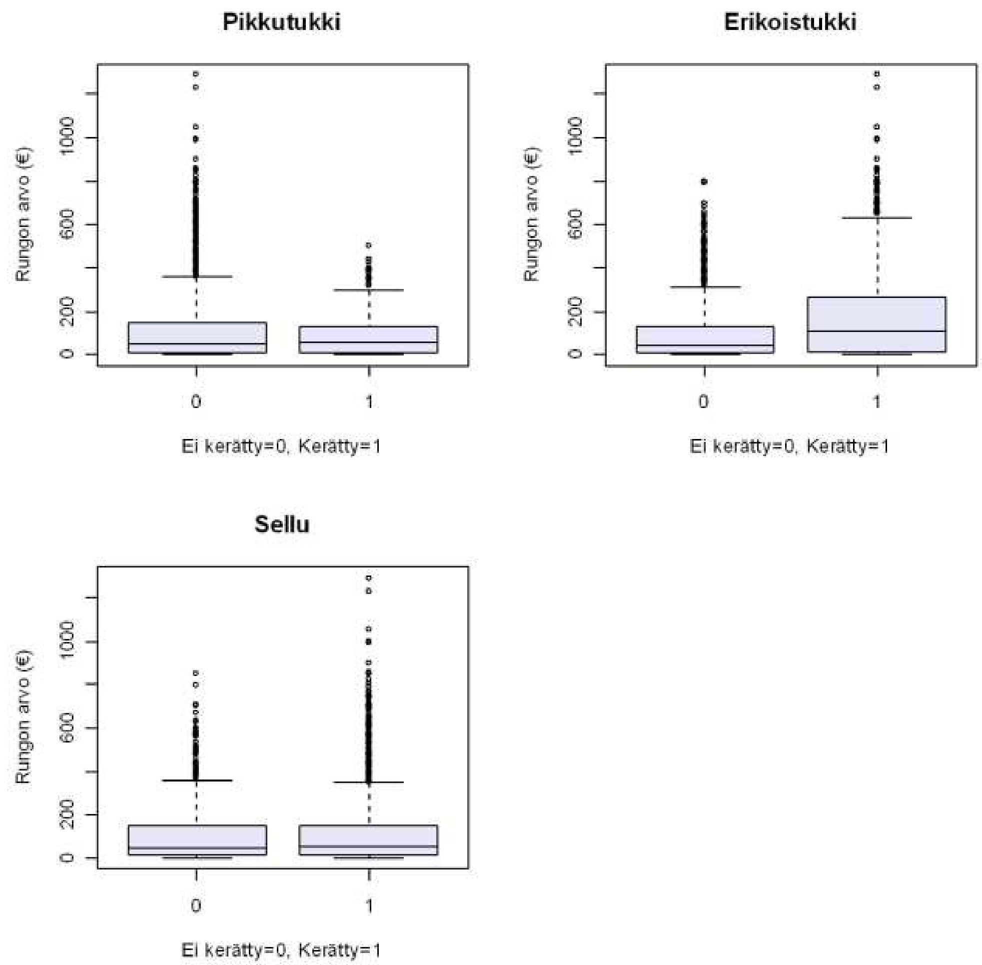

**Kuvio A.8** Tavaralajimuuttujan erikoistukki tulosjakaumien tarkastelu (kuusi).



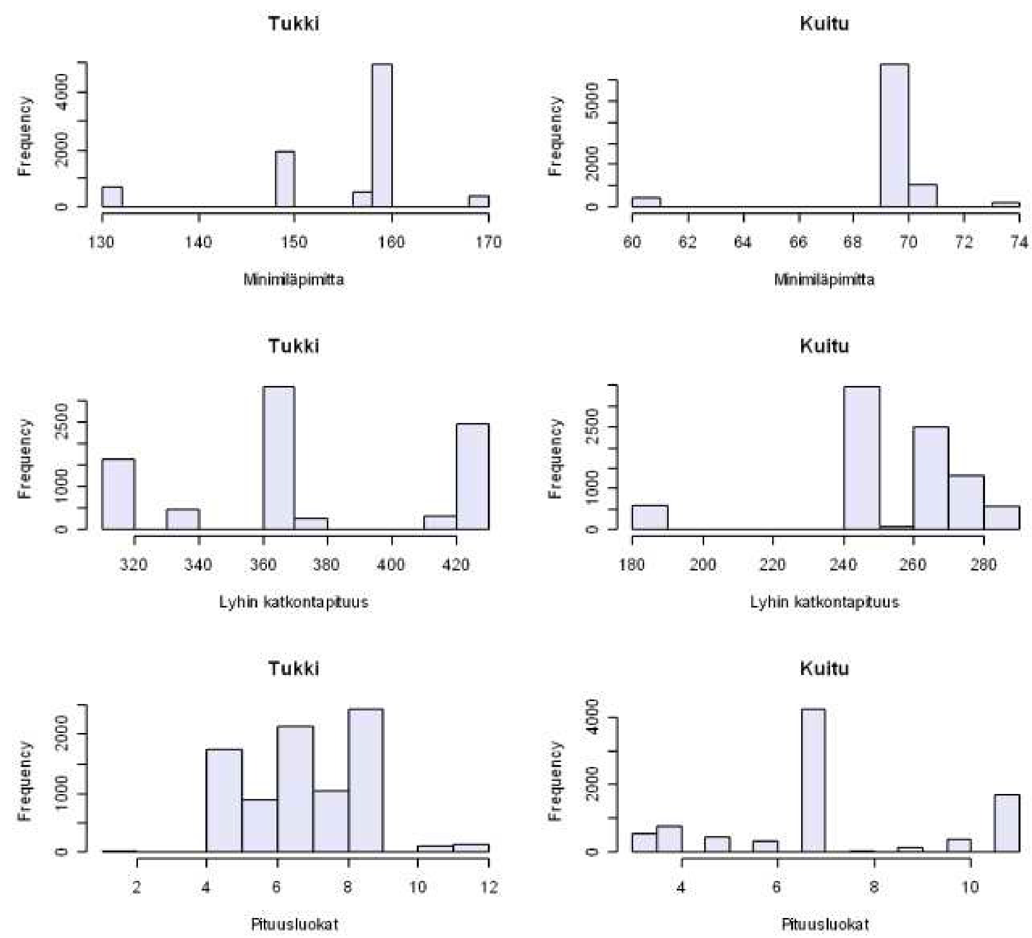

**Kuvio A.9** Tavaralajimuuttujan pikkutukki tulosjakauman tarkastelu (kuusi).




**Kuvio A.10** Tavaralajimuuttujan sellu tulosjakauman tarkastelu (kuusi).




**Kuvio A.11** Tavaralajimuuttujien laatikko-jana –tarkastelu (kuusi).




**Kuvio A.12** Parametrien histogrammit (kuusi).

# Liite B: Lineaarinen regressioanalyysi

Tämän liitteen lähteenä on käytetty pääosin teosta: Chatterjee & Hadi & Price (2000). Apuna on käytetty myös teoksia: Puntanen (1999a-b), Seber & Lee (2003) sekä Faraway (2002). Tarkoituksena on esitellä regressioanalyysin teoriaa pääpiirteittäin, jotta tutkimuksen seuraaminen olisi helpompaa. Tutkielman empiirisen luonteen vuoksi teorian esitteleminen on jätetty pintapuoliseksi.

## B.1 Yhden selittäjän lineaarinen regressio

Yhden selittäjän lineaarinen regressiomalli on muotoa

(B.1) $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i=1,2,...,n,$$

missä y on selitettävä muuttuja, josta usein käytetään myös nimitystä vaste (*response*), x on selittävä muuttuja (*predictor*), $\beta_0$ ja $\beta_1$ ovat ennalta tuntemattomia regressiokertoimia (*regression coefficients*) eli parametreja ja $\varepsilon$ on tuntematon virhetermi (*random disturbance* tai *error*).

Oletetaan, että lineaarinen yhtälö (B.1) tuottaa hyväksyttävän arvion *y*:n ja *x*:n todellisesta suhteesta. Toisin sanoen, *y* on likimääräinen *x*:n lineaarinen funktio ja $\varepsilon$ mittaa tämän likimääräisyyden tarkkuutta. Erityisesti voidaan todeta, että $\varepsilon$ ei sisällä mitään systemaattista informaatiota, joka ei jo sisälly *x*:n antamaan informaatioon *y*:tä ennustettaessa. Kerrointa $\beta_1$ kutsutaan kulmakertoimeksi (*slope*) ja se voidaan tulkita *y*:n muutoksen suuruudeksi, kun *x* muuttuu yhden yksikön. Kerrointa $\beta_0$ kutsutaan vakiokertoimeksi (*intercept* tai *constant*) ja se on *y*:n ennustettu arvo *x*:n saadessa arvon nolla. Tavallisesti virhetermien $\varepsilon_i$ oletetaan olevan toisistaan riippumattomia, $E(\varepsilon_i) = 0$ ja $Var(\varepsilon_i) = \sigma^2$, $i$=1,2,...,$n$. Testauksia varten usein oletetaan myös, että kukin $\varepsilon_i$ noudattaa normaalijakaumaa odotusarvolla 0 ja varianssilla $\sigma^2$ eli $\varepsilon_i \sim N(0,\sigma^2)$.

Lineaarinen regressiomalli voidaan esittää myös matriisimuodossa

(B.2) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

missä **y** on n×1 havaintovektori, **X** on n×p matriisi, **ε** on residuaalien muodostama n×1 vektori ja **β** on p×1 parametrivektori. Matriisia **X** kutsutaan myös mallimatriisiksi. Kirjoitettaessa kaava (B.2) auki saadaan

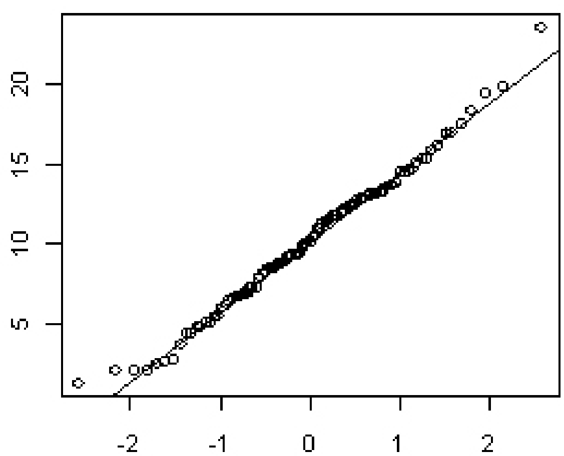
(B.3) $$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

## B.2 Lineaarisen regressiomallin oletukset

Regressiomallin oletukset kohdistuvat mallin muotoon, virhetermien ominaisuuksiin, selittäjiin ja havaintoihin. Mallin muotoon kohdistuva oletus lineaarisessa regressiossa viittaa selitettävän muuttujan lineaariseen relaatioon parametrien suhteen. Selittäjiin liittyvien oletusten mukaan selittäjät ovat epäsatunnaisia, ne eivät saa sisältää mittausvirheitä eivätkä ne saa olla toisistaan lineaarisesti riippuvia. Viimeisen oletuksen täyttymistä tarvitaan, jotta pienimmän neliösumman ratkaisu olisi yksikäsitteinen. Mikäli oletus ei täyty, puhutaan kollineaarisuudesta. Tätä ongelmaa käsitellään myöhemmin tässä alaluvussa.

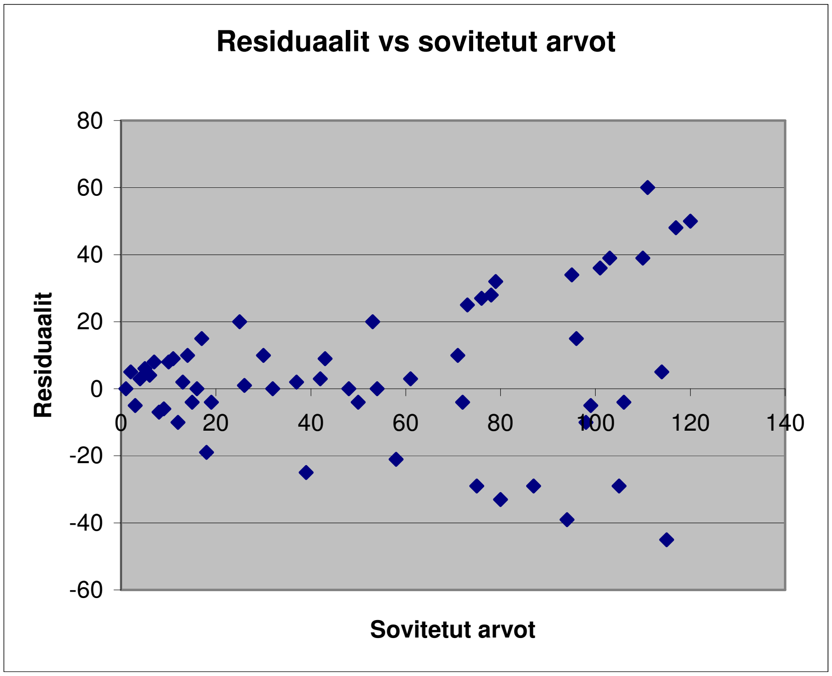
Jos virhetermit ovat toisistaan aikariippuvia, niiden arvot riippuvat aikaisemmista virhetermien arvoista. Tällöin kyse on **autokorrelaatiosta**. Autokorrelaatio on yleensä aikasarja-aineistoissa esiintyvä ongelma. Koska tässä tutkimuksessa ei autokorrelaatiota esiintynyt, sitä ei tässä tarkemmin käsitellä.

**Normaalijakautuneisuusoletuksen** paikkaansa pitävyyttä tarkastellaan yleensä piirtämällä kvantiili-kvantiili-kuvio (*QQ plot*), joka on graafinen testi, jossa virhetermit järjestetään ja näiden järjestettyjen virhetermien tulisi jotakuinkin noudattaa suoraa linjaa, mikäli ne ovat normaalisti jakautuneita. Kuviossa B.1 on havainnollistettu tilannetta, jossa havainnot ovat normaalisti jakautuneita.




**Kuvio B.1** Normaalisti jakautuneen muuttujan kvantiili-kvantiili-kuvio

Jotta mallin oletukset täyttyisivät, tulee mallin virhetermien $\varepsilon_i$ olla **vakiovarianssisia eli homoskedastisia**. Jos tämä oletus ei täyty, sanotaan virhetermien olevan heteroskedastisia. Tämä ongelma havaitaan yleensä tutkimalla hajontakuviota, jossa vaaka-akselilla on mallin sovitetut arvot ja pystyakselilla mallin residuaalit, kun vastetta $y$ on selitetty $x$:llä. Kuviossa B.2 on havainnollistettu heteroskedastisuutta edellä mainitun hajontakuvion avulla. On havaittavissa, että virhetermin varianssi kasvaa $x$:n arvojen kasvaessa eli virhetermien varianssi ei ole vakio.




**Kuvio B.2** Heteroskedastiset residuaalit

Jos vakiovarianssisuusoletus ei ole voimassa eikä tilannetta korjata, pienimmän neliösumman menetelmän (alaluku B.3) soveltaminen dataan tuottaa teoreettisessa mielessä epätarkkoja estimaatteja regressiokertoimiksi. Kertoimien estimoidut hajonnat ovat usein aliarvioituja, jolloin saadaan väärä kuva niiden tarkkuudesta.

Heteroskedastisuusongelma voidaan ratkaista monella tavalla. Tällaisia tapoja ovat erilaiset dataan kohdistuvat muunnokset sekä painotetun neliösumman menetelmä. Näitä tarkastellaan myöhemmin hiukan tarkemmin (alaluvut B.4 ja B.5).

Kun regressiomalliin lisätään muuttuja ja mallissa jo aiemmin olevien muuttujien regressiokertoimien arvot muuttuvat merkittävästi, on syytä tarkastella esiintyykö mallissa **kollineaarisuutta**. Tämä ongelma esiintyy, kun kaksi tai useampi selittävä muuttuja korreloivat voimakkaasti keskenään. Tällöin regressiokertoimien estimaateista tulee epäluotettavia ja niiden tulkinta hankaloituu. Ongelma voidaan ratkaista muodostamalla korreloivista muuttujista yhdistetty muuttuja tai yhdistettyjä muuttujia, jättämällä mallista pois jokin voimakkaasti muiden kanssa korreloiva muuttuja, muodostamalla uusia riippumattomia muuttujia esimerkiksi pääkomponenttianalyysin avulla tai käyttämällä toisenlaista regressioanalyysiä (esim. harjaregressio).

Yleensä pelkkä muuttujien välisten korrelaatiokertoimien tarkasteleminen ei riitä kollineaarisuuden toteamiseen, vaan tarvitaan muita tunnuslukuja. Tähän tarkoitukseen käytetään VIF-tunnuslukua (*Variance Inflation Factor*). Se lasketaan kaavalla

(B.4) $$VIF_j = \frac{1}{1-R_j^2},\ j = 1,\ldots,p\,,$$

missä $R_j^2$ on yhteiskorrelaatiokertoimen neliö, kun muuttujaa $X_j$ selitetään kaikilla muilla selittäjillä ja $p$ viittaa selittävien muuttujien lukumäärään. Jos muuttuja $X_j$ korreloi voimakkaasti muiden selittäjien kanssa, $R_j^2$ on lähellä ykköstä ja tällöin $VIF_j$ on suuri. Suorittamalla yksinkertainen Google-haku hakusanoilla vif ja values, voidaan todeta, että suurin osa tutkijoista käyttää raja-arvona arvoa 10, mutta myös arvoa 5 käytetään. Olen valinnut raja-arvoksi arvon 5, joten sitä suuremmat VIF-arvot ilmaisevat kollineaarisuuden tässä tutkimuksessa.

## B.3 Pienimmän neliösumman menetelmä (PNS)

Pienimmän neliösumman menetelmän ominaisuutena on, että vähäiset oletusten rikkomiset eivät merkittävästi pilaa datasta tehtyjä tulkintoja ja johtopäätöksiä

(Chatterjee & Hadi & Price, s.88). Kuitenkin, mikäli samanaikaisesti rikotaan useita oletuksia, voi tulokset vääristyä merkittävästi.

Tutkittavalle datalle halutaan estimoida tuntemattomat parametrit, jotka yhden selittäjän tapauksessa ovat $\beta_0$ ja $\beta_1$. Tämä voidaan ilmaista haluna löytää suora, joka antaa parhaan sovituksen vasteen ja selittävän muuttujan hajontakuvion pisteille. Regressiosuora voidaan sovittaa havaintoaineistolle monella eri menetelmällä. Yleisin tapa on pienimmän neliösumman menetelmä (*OLS,* O*rdinary Least Squares*). Menetelmässä ratkaistaan regressiomallin parametrien $\beta_0$:n ja $\beta_1$:n arvot sovittamalla regressiosuora havaintoaineistoon minimoiden pisteiden vertikaalisten etäisyyksien neliöiden summa sovitettavaan suoraan nähden. Nämä etäisyydet vastaavat selittäjän virhettä. Virheet saadaan ratkaisemalla $\varepsilon$ kaavasta (B.1).

(B.5) $$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \qquad i = 1, 2, ..., n.$$

Näiden etäisyyksien (virheiden) neliösumma voidaan kirjoittaa seuraavasti

(B.6) $$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Kertoimien $\beta_0$ ja $\beta_1$ arvot, jotka minimoivat neliösumman S($\beta_0, \beta_1$):n saadaan seuraavasti

(B.7) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

ja

(B.8) $$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Kaavoissa (B.7) ja (B.8), $\overline{y}$ on y:n keskiarvo ja $\overline{x}$ on x:n keskiarvo. Estimaatteja $\hat{\beta}_0$ ja $\hat{\beta}_1$ kutsutaan parametrien $\beta_0$ :n ja $\beta_1$ :n pienimmän neliösumman estimaateiksi.

## B.4 Painotetun neliösumman menetelmä

Virhetermin oletukset riippumattomuudesta ja identtisestä jakautuneisuudesta odotusarvonaan nolla ja vakiovarianssisuudesta (*homoscedasticity*) eivät aina pidä paikkaansa. Syynä voi olla regressioyhtälön väärä spesifiointi, tärkeitä muuttujia saattaa puuttua tai datassa on joitain äärihavaintoja. Painotettua pienimmän neliösumman menetelmää käytetään, kun vakiovarianssisuusoletus ei täyty eli virhetermi on heteroskedastinen. Regressioyhtälön parametrit saadaan estimoitua minimoimalla painotettu virhetermien neliösumma käyttäen painoja, jotka ovat kääntäen verrannollisia virheen varianssiin. Oletetaan kaavan (B.1) malli, jossa virhetermit ovat normaalisti jakautuneita, odotusarvonaan nolla ja niiden varianssi on verrannollinen selittäjän toiseen potenssiin eli $\varepsilon_i \sim N(0,\sigma^2 x_i^2)$. Asettamalla $\tilde{y}_i = \frac{y_i}{x_i}$ *ja* $\tilde{x}_i = \frac{1}{x_i}$, malli tulee muotoon

(B.9) $$\tilde{y}_i = \beta_0 \tilde{x}_i + \beta_1 + \varepsilon_i,$$

jossa $\varepsilon_i \sim N(0,\sigma^2)$ . Mallin sovittaminen on yhtäpitävää kuin yhtälön

(B.10) $$\sum_{i=1}^{n} \left( \frac{1}{x_i} \right)^2 (y_i - \beta_0 - \beta_1 x_i)^2$$

minimointi. Tällöin painotetun neliösumman yhtälö voidaan kirjoittaa muotoon

(B.11) $$S(\beta, w) = \sum_{i=1}^{n} \frac{1}{w_i} (y_i - \beta_0 - \beta_1 x_i)^2,$$

missä $w_i = \frac{1}{x_i^2}$.

## B.5 Vastemuuttujan muunnokset

Tämä luku perustuu pääosin teokseen: Faraway (2000). Suoritettaessa analyysejä saattaa syystä tai toisesta tulla tarve tehdä datalle muunnoksia. Usein syynä on regressiomallin virhetermin vakiovarianssisuusoletuksen rikkoon-tuminen. Tällöin on mahdollista yrittää korjata tilanne etsimällä vasteelle sopiva muunnos ja sopiva keino etsiä ratkaisu ongelmaan on käyttää Boxin ja Coxin menetelmää.

Menetelmä on tarkoitettu ainoastaan vasteille, joiden arvot ovat positiivisia. Se tuottaa sopivan muunnoksen etsimällä datalle parhaan sovitteen. Boxin ja Coxin menetelmä muuntaa vasteen, jossa $\lambda$:lla indeksoitu muunnosperheen joukko on

$$t_\lambda(y) = \begin{cases} \frac{y_\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}.$$

Logaritmimuunnos on yksi regressioanalyysin yleisimmin käytetyistä muunnoksista. Muunnos on erityisen hyödyllinen, kun tarkasteltavan muuttujan keskihajonta on verrattain suuri keskiarvoon nähden (Chatterjee & Hadi & Price). Logaritmimuunnos yleensä pienentää vaihtelua ja epäsymmetrisyyttä. Tämä muunnos on erittäin tehokas heteroskedastisuuden tapauksessa. Yhden selittäjän tapauksessa vastemuuttujan logaritmointi tuottaa seuraavan yhtälön

(B.12) $$\log(y_i) = \beta_0 + \beta_1 x_1 + \varepsilon.$$

Muunnettaessa yhtälö muotoon, jossa vaste on normaalimuodossa, saadaan

(B.13) $$y = \exp(\beta_0 + \beta_1 x)\exp(\varepsilon) = e^{\beta_0} e^{\beta_1 x} e^{\varepsilon}.$$

Mallissa (B.13) virhetermit tulevat malliin multiplikatiivisesti eivätkä additiivisesti, kuten yleensä. Normaalien regressiomenetelmien käyttö logaritmoidulle vasteelle vaatii, että uskomme virhetermien tulevan malliin multiplikatiivisesti. Käytännössä emme usein kuitenkaan tiedä, kuinka virhetermit malliin tulevat. Yleinen lähestymistapa on kokeilla erilaisia muunnoksia ja tutkia sitten, pitävätkö virhetermeiltä vaaditut oletukset paikkansa.

Yleensä ennusteet pitää ilmaista alkuperäisellä skaalalla. Regressiokertoimet täytyy tulkita transformoidun skaalan mukaisesti. Ei ole olemassa suoraviivaista tapaa muuntaa ne takaisin arvoiksi, jotka voidaan tulkita alkuperäisellä skaalalla.

Logaritmoitaessa vaste regressiokertoimilla on erityinen tulkinta:

(B.14) $$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

(B.15) $$\hat{y} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_1} \cdots e^{\hat{\beta}_p x_p}$$

$x_1$:n arvon kasvu yhdellä yksiköllä vastaa alkuperäisen skaalan ennustetun vasteen kertomista $e^{\hat{\beta}_1}$:lla. Siksi logaritmista skaalaa käytettäessä regressiokertoimet pitää tulkita multiplikatiivisesti sen sijaan, että ne tulkittaisiin additiivisella tavalla.

## B.6 Hypoteesien testaaminen

Muuttujan $x$ käyttökelpoisuutta vasteen $y$ selittäjänä voidaan tarkastella korrelaatiokertoimen sekä $x$:n ja $y$:n hajontakuvion avulla. Muodollisempi tapa käyttökelpoisuuden tarkastelemiseksi on asettaa jokin regressiokerrointa koskeva hypoteesi ja tutkia sen paikkaansa pitävyyttä. Yleensä asetetaan hypoteesi $\beta_1 = 0$ eli $y$:n ja $x$:n välillä ei ole lineaarista riippuvuutta. Tämän hypoteesin testaaminen vaatii, että residuaalit $\varepsilon_i$ ovat toisistaan riippumattomia, normaalisti jakautuneita keskiarvolla 0 ja vakiovarianssilla $\sigma^2$ (ks. alaluku (B.2)). Näiden oletusten voimassaolo takaa regressiokertoimien $\beta_0$ ja $\beta_1$ harhattomuuden[1] (Chatterjee & Hadi & Price, s.33). Normaalisuusoletuksen ollessa voimassa sopiva testisuure testaamaan nollahypoteesia $H_0$: $\beta_1 = 0$ vaihtoehtoista hypoteesia $H_1$: $\beta_1 \neq 0$ vastaan on t-testi

(B.16) $$t_1 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}.$$

Kaavassa (B.16), kuten jatkossakin, *s.e.* tarkoittaa keskivirhettä (*standard error*). Testisuure $t_1$ noudattaa Studentin t-jakaumaa vapausastein $n-2$. Testiä suoritettaessa asetetaan haluttu merkitsevyystaso $\alpha$. Nollahypoteesi $H_0$ tulee hylätyksi merkitsevyystasolla $\alpha$, mikäli

(B.17) $$|t_1| \geq t_{(n-2,\alpha/2)}.$$

Missä $|t_1|$ tarkoittaa $t_1$:n itseisarvoa ja tällöin hyväksytään vaihtoehtoinen hypoteesi $H_1$. Yhtäpitävä peruste kaavalle (B.17) on verrata t-testin p-arvoa ja hylätä $H_0$, mikäli $p(|t_1|) \leq \alpha$, missä $p(|t_1|)$ on p-arvo eli todennäköisyys, että Studentin t-jakaumaa vapausastein $n-2$ noudattava satunnaismuuttuja saa suuremman arvon kuin $(|t_1|)$.

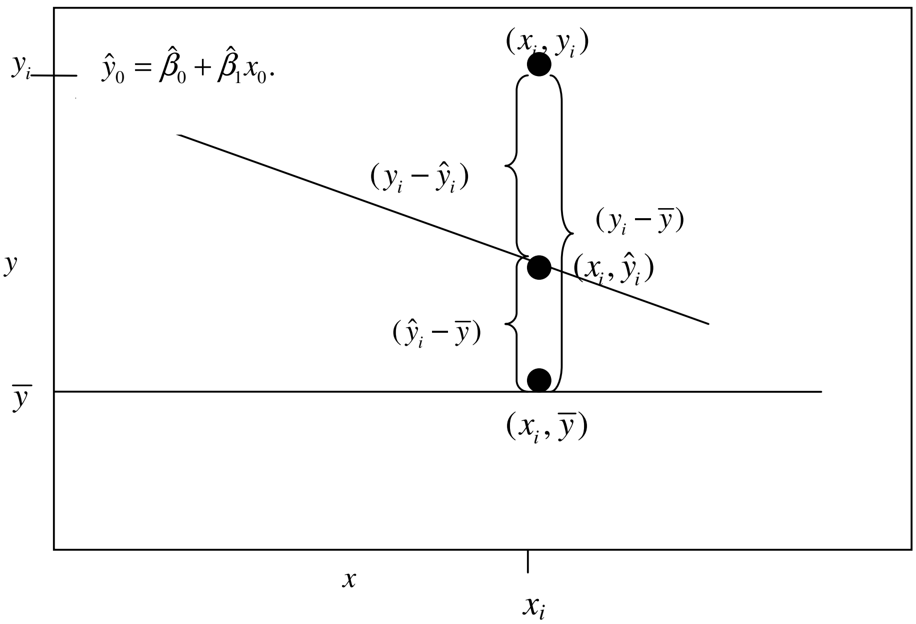
## B.7 Mallin sopivuuden mittaaminen

Sen lisäksi, että olemme kiinnostuneita siitä, onko vasteen ja selittäjän välillä lineaarista riippuvuutta, olemme luonnollisesti kiinnostuneita myös mallin selitysvoimasta.

[1] Estimaatin $\hat{\theta}$ sanotaan olevan parametrin $\theta$ harhaton estimaatti, mikäli $\hat{\theta}$:n odotusarvo on yhtä suuri kuin $\theta$.

Yleinen tapa mitata sitä, kuinka hyvin malli selittää selitettävän ja selittäjän välistä riippuvuutta, on laskea selitysaste. Tätä varten datasta lasketaan ensin seuraavat neliösummat:

$SST = \sum (y_i - \overline{y})^2$, $SSR = \sum (\hat{y}_i - \overline{y})^2$, $SSE = \sum (y_i - \hat{y}_i)^2$, joista *SST* tarkoittaa kokonaisneliösummaa, *SSR* regressioneliösummaa ja *SSE* virheneliösummaa. Kuvio B.3 havainnollistaa tilannetta, kun tarkastellaan vain yhtä satunnaisesti valittua datan pistettä $(x_i, y_i)$.




**Kuvio B.3.** Etäisyyksien laskeminen

Kokonaisneliösumma muodostuu regressioneliösumman ja virheneliösumman summana eli *SST=SSR+SSE*. Koska SSR kuvaa sitä, kuinka hyvin *x* selittää *y*:tä, voidaan muodostaa suhdeluku *SSR/SST*, joka kertoo, kuinka paljon *x*:n avulla voidaan selittää *y*:n vaihtelua. Tätä suhdelukua merkitään $R^2$.

(B.18) $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Selitysasteen huonona puolena on se, että se kasvaa, kun malliin lisätään muuttujia. Siksi onkin syytä käyttää nk. korjattua selitysastetta (*adjusted* $R^2$). Tämä saadaan, kun edelliseen tehdään vapausastekorjaus. Korjattu selitysasteen $R_{adj}^2$ laskentakaava on

(B.19) $$R_{adj}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}.$$

Yleensä on siis perustellumpaa käyttää korjattua selitysastetta normaalin selitysasteen sijasta.

## B.8 Usean selittäjän lineaarinen regressio

Useamman selittäjän regressiomalli, kun data sisältää *n* havaintoa selitettävästä muuttujasta *y* ja *p* selittävää muuttujaa $x_1, x_2, ..., x_p$, voidaan esittää seuraavasti:

(B.20) $$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon_i, \qquad i=1,2,...,n,$$

missä $\beta_1, \beta_2, \beta_3, ..., \beta_p$ ovat regressiokertoimia ja $\varepsilon_i$ on virhetermi.

Useamman selittäjän regressiomalli on matriisimuodossa sama kuin yhtälössä (B.2), mutta nyt termit ovat seuraavanlaiset:

(B.21) $$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \textit{ ja } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

ja missä $x_{i0} = 1$ kaikilla *i*.

Virhetermiä koskevat oletukset pienimmän neliösumman menetelmässä ovat, kuten alaluvussa B.1 on esitetty: $E(\boldsymbol{\varepsilon}) = 0$ ja $Var(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \mathbf{I}_n$. Virhetermit ovat siis riippumattomia, virhetermin odotusarvo on nolla ja niiden varianssi on vakio. Tästä seuraa, että $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$.

Regressiokertoimien vektorin $\boldsymbol{\beta}$ pienimmän neliösumman estimaattori $\hat{\boldsymbol{\beta}}$ saadaan minimoimalla neliösumma, kun havaintojen poikkeamat odotusarvoistaan lasketaan yhteen. Täten pienimmän neliösumman estimaattorit saadaan minimoimalla $S(\boldsymbol{\beta})$, missä

(B.22) $$S(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

$S(\boldsymbol{\beta})$ :n minimointi johtaa yhtälöryhmään

(B.23) $$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}.$$

Oletetaan, että $(\mathbf{X}^T\mathbf{X})$ on kääntyvä eli sillä on olemassa käänteismatriisi. Silloin pienimmän neliösumman estimaatit voidaan kirjoittaa muodossa

(B.24) $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Pienimmän neliösumman estimaattorilla on seuraava ominaisuus: $\hat{\boldsymbol{\beta}}$ on $\boldsymbol{\beta}$ :n harhaton estimaattori varianssi-kovarianssimatriisilla

(B.25) $$Var(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Kaikista $\beta$:n harhattomista lineaarisista estimaattoreista pienimmän neliösumman estimaattorilla on pienin varianssi. Tästä syystä $\hat{\beta}$ :a kutsutaan $\beta$:n parhaaksi lineaariseksi harhattomaksi estimaattoriksi (*best linear unbiased estimator, BLUE*).

## B.9 Ennustevälit

Usein sovellusten näkökulmasta katsottuna on tarpeen tutkia, kuinka uusi havainto sijoittuu regressioanalyysin pohjana olevaan dataan verrattuna. Uuden datapisteen sijainnin tarkastelussa käytetään regressiomallista saatuja parametreja ja mallin keskivirhettä sekä t-jakauman taulukoituja arvoja. 100(1-$\alpha$) %:n ennustevälit uusille havainnoille muodostetaan seuraavasti. Ennustettu arvo on

(B.26) $$\hat{y}_0 = \beta_0 + \beta_1 x_0.$$

Tämän ennusteen keskivirhe on

(B.27) $$s.e.(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum (x_i - \overline{x})^2}}.$$

Edellisten kaavojen (B.26) ja (B.27) sekä taulukoidun t-arvon avulla voidaan muodostaa 100(1-$\alpha$) %:n ennusteväli (kaava(B.28)).

(B.28) $$\hat{y}_0 \pm t_{(n-2,\alpha/2)} s.e.(\hat{y}_0).$$

Usean selittäjän tapauksessa ennusteväli lasketaan seuraavasti:

(B.29) $$\hat{y}_0 \pm t_{(n-p,\alpha/2)} \hat{\sigma}\sqrt{1 + x_0^T (X^T X)^{-1} x_0}\ .$$