Maalitodennäköisyyksien mallintaminen jääkiekossa

Tilastotieteen pro gradu -tutkielma

17.1.2019

Jani Pellinen
Matematiikan ja tilastotieteen laitos
Jyväskylän yliopisto

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Pellinen, Jani: Maalitodennäköisyyksien mallintaminen jääkiekossa

Tilastotieteen pro gradu -tutkielma (44 sivua)
17.1.2019

Tiivistelmä

Data-analyysin hyödyntäminen urheilulajien analysoinnissa on yleistynyt tekniikan kehityksen myötä. Luultavasti tunnetuin läpimurto on tapahtunut baseballissa, joka sopii toistokokeiden tapaiselta tyyliltään erinomaisesti mallinnettavaksi. Jääkiekossa data-analyysi on vasta tekemässä nousuaan, sillä ottelutapahtumien tilastointi on ollut varsin niukkaa verrattuna moneen muuhun lajiin. Pelkkien maalimäärien analysoinnissa ongelmana on maalien vähäinen määrä ottelukohtaisesti. Kaikkien laukausten sisällyttäminen analyyseihin kasvattaa ottelukohtaista otoskokoa paljon, ja laukaisukartoista saatavien sijaintitietojen hyödyntäminen tuo lisäarvoa. SM-Liigan laukaisukarttoja on saatavilla syksystä 2014 alkaen.

Tässä työssä sovitetaan maalitodennäköisyysmalli jääkiekon SM-Liigan laukaisudataan, joka on saatavilla SM-Liigan verkkosivuilta. Maalitodennäköisyydellä tarkoitetaan yksittäisen laukauksen maaliin menemisen todennäköisyyttä. Kyseistä todennäköisyyttä voidaan mallintaa hyödyntämällä laukaisukartoista löytyviä tietoja sijainneista ja pelaajista sekä erikseen laskettavista lisämuuttujista. SM-Liigan aineistoon tehtyä maalitodennäköisyysmallia ei ole aiemmin julkaistu.

Maalitodennäköisyysmallit sovitetaan käyttämällä yleistettyjä logistisia sekamalleja, joiden avulla laukovan pelaajan ja torjuvan maalivahdin vaikutus maalitodennäköisyyteen voidaan huomioida. Mallien validoinnissa keskitytään enimmäkseen mallien kalibraatioon, eli mallin tuottamien todennäköisyysestimaattien laatuun. Mallien kalibraatiota tutkiessa saadaan käsitys siitä, vastaako todennäköisyysestimaatti todellisuutta ja onko se luotettava. Työssä tehtyjen kalibraatiotarkastelujen perusteella mallit vaikuttavat toimivilta. Työssä käsitellään myös raakadatan muokkaamista mallintamisen kannalta sopivaan muotoon ja mallintamisessa käytettävien muuttujien laskentaa sekä esitetään tapoja sille, miten maalitodennäköisyyksiä voidaan hyödyntää pelaajien ja joukkueiden suoritusten arvioinnissa.

Avainsanat: Urheiluanalytiikka, Ennustaminen, Sekamalli, Jääkiekko, Maalitodennäköisyys, Kalibraatio

Sisältö

1 Johdanto

Jääkiekossa maalimäärät ovat melko pieniä siihen nähden, kuinka paljon yksittäisessä ottelussa esiintyy laukauksia ja maalintekotilanteita. Vaikka maalit määrittävät ottelun voittajan, ne eivät itsessään kerro kovinkaan paljon ottelun etenemisestä yleisesti. Joukkueiden todellinen taso välittyy usein siitä, kumpi joukkue hallitsee ottelua ja ottelutapahtumia. Lähestymistapoja ottelun hallintaan on useita. Puolustussuuntautunut joukkue pyrkii usein hallitsemaan peliä estämällä vastustajan hyökkäyspeliä parhaansa mukaan. Vastaavasti hyökkäyssuuntautunut joukkue usein hallitsee kiekkoa ja pyrkii rakentamaan mahdollisimman paljon hyviä maalipaikkoja. Kaikkien lähestymistapojen tavoitteena on luoda enemmän laadukkaita maalipaikkoja kuin vastustaja ja voittaa ottelu tekemällä enemmän maaleja. Urheilun hienous on kuitenkin siinä, että paremmin pelannut joukkue ei aina voita. Tämän takia on tärkeää analysoida muutakin kuin pelkkiä maaleja. Laukaisumäärät ja kiekonhallinta antavat jo laajemman kokonaiskuvan ottelutapahtumista. Jääkiekon korkeimman ammattilaissarjan NHL:n osalta on havaittu, että laukaisumäärät korreloivat ajallisen kiekonhallinnan kanssa (Barnes, 2008). Laukaussuhdetta ja muita laukaisumääriin perustuvia tunnuslukuja kutsutaankin usein virheellisesti kiekonhallintaluvuiksi. Laukaussuhde lasketaan jakamalla joukkueen laukaisumäärä ottelun kokonaislaukaisumäärällä.

Pelkissä laukaisumäärissä ja kiekonhallinnassa on myös omat heikkoutensa. Ne voivat olla tehottomia, eli laukaisutilaston hallitseminen ja kiekon pitäminen eivät välttämättä johda laadukkaisiin maalipaikkoihin vastustajan onnistuessa puolustuspelissään. Pelkkien maalipaikkojen laskeminen voi olla intuitiivisesti hyvän oloinen vaihtoehto. Subjektiivisesti peliä katsomalla tehty maalipaikkalaskenta on siinä määrin ongelmallista, että niiden määrittäminen historiallisesta datasta on mahdotonta ilman koko aineiston läpikäyntiä, joka käytännössä vaatisi ottelukoosteiden katsomista ja se veisi paljon aikaa. Laukauksista on kuitenkin olemassa SM-Liigan keräämää sijaintidataa, jota voidaan hyödyntää analyysityössä. Tässä vaiheessa voidaan ottaa käyttöön maalitodennäköisyydet, joiden tarkoituksena on antaa objektiivisesti mahdollisimman hyvä arvio laukauksen lopputulokselle. Ideana on siis estimoida todennäköisyys maalin syntymiselle. Laukaisumääriin verrattuna maalitodennäköisyyksien laskennassa laadukkaille maalipaikoille annetaan suurempi painoarvo.

Maalitodennäköisyyksien huomioiminen lajin analysoinnissa on tärkeää, sillä käyt-

töön on tulossa teknologiaa, joka mahdollistaa aiempaa laadukkaamman datan keräämisen. Esimerkiksi älykiekkoteknologian avulla laukausten ja pelaajien sijainnit saadaan määriteltyä tarkasti. Tällaisen informaation hyödyntäminen todennäköisesti parantaa maalitodennäköisyysmalleja entisestään. Tällä hetkellä käytössä on ainoastaan laukaisukarttoja, joiden avulla saadaan tehtyä jo varsin hyödyllisiä malleja. Esimerkiksi laukausta edeltävien syöttöjen ja fyysisten muuttujien, kuten kiekon nopeuden, huomioiminen on tämän hetken datalla mahdotonta.

Maalitodennäköisyysmalleja on aiemmin tehty NHL:n laukaisudataan. Maalitodennäköisyyksiin liittyviä akateemisia julkaisuja on hyvin niukasti, sillä analyysityötä tehneet henkilöt ovat julkaisseet mallejaan lähinnä blogiteksteissä ja jääkiekon tilastoanalytiikkaan erikoistuneilla verkkosivuilla. Brian Macdonaldin (2012) konferenssijulkaisu käsittelee maaliodotusarvoja ottelun tasolla. Kyseisessä työssä mallinnetaan ottelun maalimääriä joukkueittain käyttämällä selittäjänä ottelupöytäkirjan tilastoja, kuten laukaisumääriä ja aloitusvoittoja, jolloin yksittäisille laukauksille ei kuitenkaan saada maalitodennäköisyysestimaatteja. Tämän työn terminologiassa maalitodennäköisyydellä tarkoitetaan yksittäisen laukauksen maaliin menemisen todennäköisyyttä. Maaliodotusarvolla sen sijaan tarkoitetaan maalitodennäköisyyksiin perustuvaa odotettua maalimäärää yli jonkun tietyn ajan, esimerkiksi ottelun osalta. Yksittäisten laukausten todennäköisyysmalleissa, kuten esimerkiksi Hockey Graphs -sivuston (2015) julkaisemassa mallissa, käytetään yleensä laukausten sijaintitietoja, aikoja ja niistä johdettavia muuttujia. Mallit sovitetaan usein logistisena regressiona, ja myös neuroverkkojen käyttäminen on yleistä. Edellä mainitussa artikkelissa laukova pelaaja on huomioitu käyttämällä yhtenä selittäjänä pelaajan laukausten viimeistelyprosenttia, jota on regressoitu kohti kaikkien pelaajien keskiarvoa. Viimeistelyprosentilla tarkoitetaan maaliin menneiden laukausten osuutta pelaajan kaikista laukauksista.

Tämän työn pääpaino on itse mallintamisprosessissa. Aluksi tarkastellaan tarjolla olevaa laukaisudataa, millaisessa muodossa aineisto on ja miten se saadaan hankittua käyttöön. Seuraavaksi siirrytään analysoimaan aineistoa, jolloin tarkastellaan aineistosta johdettavissa olevia muuttujia, ja millaisia muokkauksia ja lisätietoja aineistoon voidaan hakea. Tämän jälkeen käsitellään lyhyesti yleistetyn logistisen sekamallin teoriaa ja estimointia. Mallissa hyödynnettävien splinien teoriaa myös sivutaan lyhyesti. Ennen mallien tulosten tarkastelua käydään läpi tapoja, joilla mallien sopivuutta voi-

daan arvioida. Perinteinen luokitteluvirheeseen perustuva ennustetarkkuus ei ole nyt mielenkiinnon kohteena, vaan kiinnostuksen kohteena on todennäköisyysarvion tarkkuus, jota arvioitaessa puhutaan mallin kalibraatiosta. Loppuluvussa käsitellään työn edetessä esiin nousseiden seikkojen lisäksi mahdollisia sovelluskohteita, joissa maalitodennäköisyyksiä voidaan hyödyntää käytännössä.

2 Aineisto

Työssä käytettävä aineisto on hankittu SM-Liigan verkkosivuilta (https://liiga.fi/). Otteluiden seurantasivuilla on kartta laukausten sijainneista kauden 2014–2015 alusta lähtien. Toimitsijat keräävät laukausten sijainnit ja ajat ottelun kuluessa, ja laukaisukartta päivittyy seurantasivuilla lähes reaaliaikaisesti. Jokaiselle laukaukselle merkitään sijainnin lisäksi laukoja ja laukauksen lopputulos. Laukauksella on neljä lopputulosvaihtoehtoa: maali, ohi, maalivahdin torjunta ja blokki. Blokilla tarkoitetaan puolustavan joukkueen kenttäpelaajan torjumaa laukausta. Maalin tolppaan osuneet laukaukset tulkitaan ohilaukaukseksi. Seurantasivulla julkaistaan myös ottelupöytäkirja, josta on johdettavissa lisää attribuutteja laukauksille. Maaliin menneille laukauksille saadaan pöytäkirjasta korkeintaan kaksi syöttäjää. Jokaiselle laukaukselle saadaan määritettyä pöytäkirjan avulla torjuva maalivahti. Joukkueiden kenttäpelaajien lukumäärät voidaan laskea perustuen jäähyjen aikoihin ja maalivahtien läsnäoloon. Kenttäpelaajien lukumäärästä voi päätellä, onko kyseisen laukaus tullut erikoistilanteen aikana. Erikoistilanteella tarkoitetaan yli- ja alivoimapeliä. Sijaintien ja aikojen avulla voidaan laukauksille johtaa lisää muuttujia, joita tarkastellaan myöhemmin. Tarvittavat tiedot löytyvät otteluiden seurantasivujen lähdekoodista. Laukaisukartan sisältäviä ottelutietoja on saatavilla kauden 2014–2015 alusta alkaen. Kirjoitushetkellä keväällä 2018 aineisto kattaa neljä kokonaista kautta, 1960 ottelua ja laukauksia aineistossa on yhteensä 184136 kappaletta. Mallintamista varten aineisto jaetaan opetus- ja testiaineistoon. Opetusaineistoksi määritetään aineiston kolme ensimmäistä kautta ja testiaineistoksi viimeisin neljäs kausi (2017–2018). Opetusaineisto sisältää 1960 ottelua ja 137040 laukausta. Testiaineisto vastaavasti 496 ottelua ja 47096 laukausta. Mallit sovitetaan opetusaineistoon ja testiaineistoa käytetään mallien vertailuun.

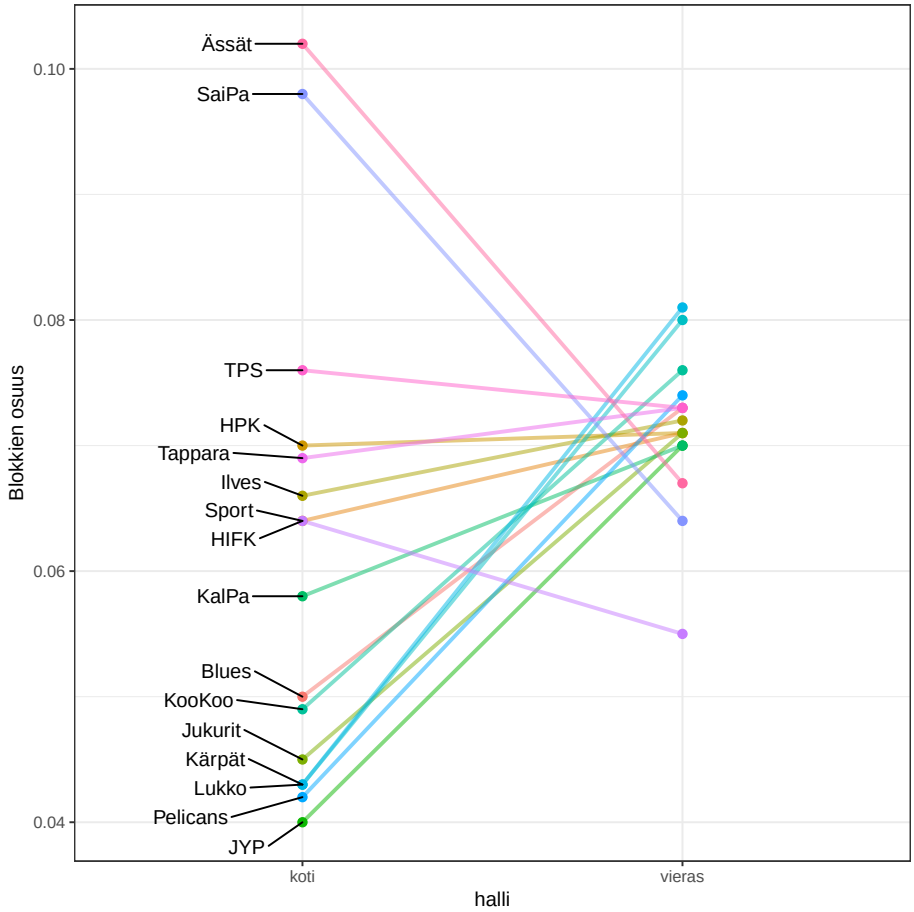
2.1 Hakuprosessi

Koko aineiston hakeminen alkaa kausien otteluohjelmien lukemisella. Jokaisen kauden otteluohjelmat löytyvät omilta sivuiltaan, joissa on taulukoituna kaikki ottelut ja kuhunkin otteluun liittyvät alasivut. Alasivuista tarvitaan ottelun pelaajatilastosivua ja seurantasivua. Näille sivuille johtavat linkit luetaan otteluohjelmista. Itse aineiston hakemisessa kaikki linkit käydään läpi ottelu kerrallaan. Yksittäisen ottelun osalta ensimmäisenä käsitellään pelaajatilastosivu, johon kummankin joukkueen kokoonpanossa olevien pelaajien ja maalivahtien ottelutilastot ovat taulukoituna. Taulukossa on myös jokaisen pelaajan pelaajasivulle johtava linkki, joka sisältää yksilöllisen tunnisteen pelaajalle. Seuraavaksi käsitellään seurantasivu. Seurantasivulta löytyy eräänlainen ottelupöytäkirja, johon on tilastoitu mm. maalit ja jäähyt. Ottelupöytäkirjassa ovat myös ajankohdat, jolloin maalivahti on vaihdettu tai otettu pois maalilta. Työn kannalta olennaisin asia on seurantasivulla oleva laukaisukartta, johon laukausten sijainnit on merkattu pisteinä kenttäkuvan päälle (kuva 3). Yksittäisestä laukauksesta saa näkyville lisätietoja siirtämällä osoittimen kyseisen laukaisupisteen päälle, jolloin näkyviin tulee lisätietoja sisältävä tekstilaatikko. Nämä tiedot löytyvät luettavassa muodossa seurantasivujen lähdekoodista koodin 1 mukaisessa muodossa. Pelaajatilastosivun ja seurantasivun lähdekoodi luetaan R-ympäristöön (R Core Team, 2018), ja tarvittavat tiedot haetaan html-koodista hyödyntämällä `rvest`-paketin funktioita (Wickham, 2016b). Hakuprosessissa ottelutietoja tallennetaan yksi ottelu kerrallaan, joten prosessin voi suorittaa päivittämällä jo olemassa olevaa tietokantaa tai vaihtoehtoisesti hakemalla kaikki tiedot alusta alkaen uudestaan. Aineiston käsittelyä ja muuttujien johtamista tarkastellaan lähemmin kappaleessa 3.

Koodi 1: Esimerkki laukaisukartan pisteestä. Laukauksen sijainti on kirjattu ensimmäisen div-solmun style-attribuuttiin. Toisesta solmusta löytyy lisätietoja.

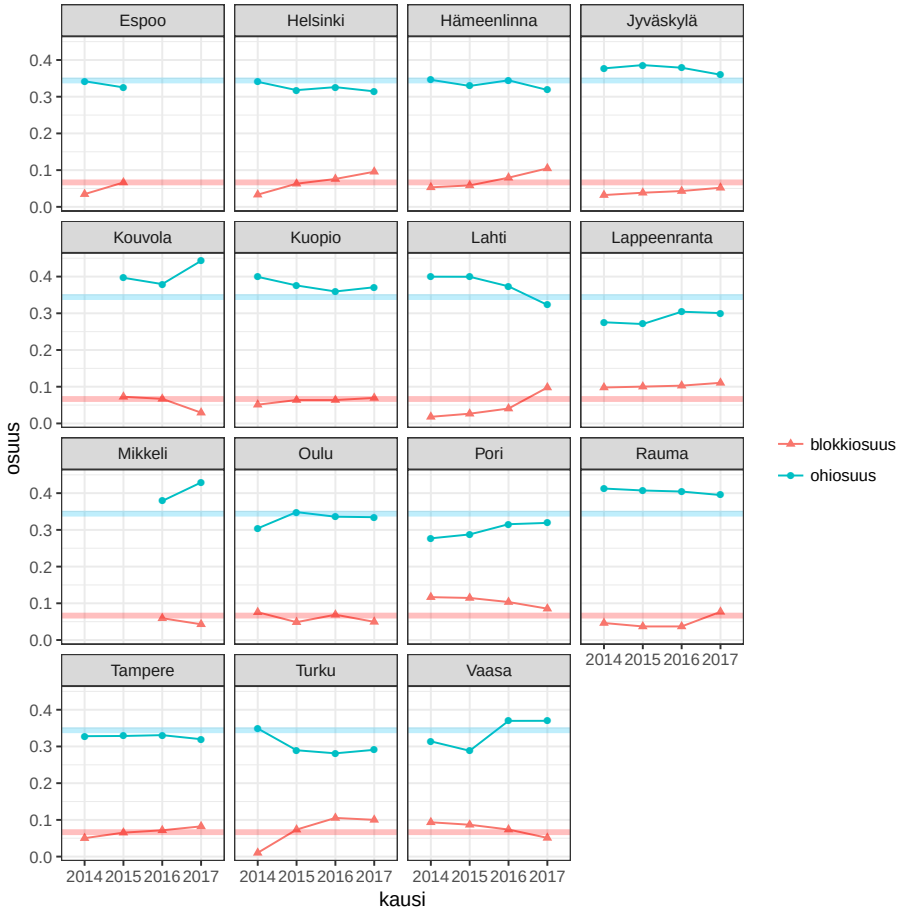
```
<div class="shot home period-2 event-goal player-29885743"
    data-tooltipid="347809" style="top: 41.190661478599225%; left:
    85.1328125%;"></div>
    <div class="shot-tooltip tooltip-347809"
        style="top:45.190661478599225%; right:14.8671875%; " >
      Laukoja: Kristian Vesalainen<br>
      Joukkue: Karpat<br>
      Aika: 32:22<br>
      Maali
    </div>
```

2.2 Paikkakuntien väliset erot



Kuva 1: Joukkueiden blokattujen laukausten osuuksia koti- ja vieraspeleissä.

Pelipaikkakuntien välillä on eroja laukausten kirjaamisessa. Selvin ero on blokattujen laukausten ja ohilaukausten määrissä. Erot voisivat johtua pelipaikkakunnalla pelaavan kotijoukkueen pelityylistä, mutta vertailu saman joukkueen koti- ja vierasotteluiden välillä paljastaa, että tietyissä kotihalleissa blokkien osuus kaikista laukauksista poikkeaa selvästi vierasotteluista, jotka jakaantuvat kaikkiin muihin halleihin (kuva 1). Vastaavasti ohi menneiden laukausten osuuksia tarkasteltaessa huomataan, että niillä paikkakunnilla, jossa blokkausten osuudet ovat korkeammat, ohi menneiden laukausten osuudet ovat matalammat. Kyseessä saattaa olla tulkinnallinen ero kirjaajien välillä. Kuvan 2 mukaan erot ovat korjaantuneet kausien kuluessa joillain paikkakunnilla. Tämä asia on syytä huomioida blokkeja ja ohi menneitä laukauksia koskevassa mallintamisessa (tarkemmin kappaleessa 4).



Kuva 2: Paikkakuntien eroja blokattujen ja maalin ohi menneiden laukausten osuuksissa kausittain. Vaakaviivat kulkevat koko aineiston keskiarvojen kohdalla. Esimerkiksi Lappeenrannassa blokkeja kirjataan keskimääräistä enemmän ja Raumalla vähemmän. Lahdessa ohilaukausten osuus on vähentynyt.

2.3 Datan laatu

Sijaintidata ei ole täysin tarkkaa, sillä sijainnit kerätään silmämääräisesti ottelun aikana, jolloin inhimilliset virheet ovat hyvin mahdollisia. Sijainnit ovat kuitenkin pääosin oikein, ja voitaneen olettaa, ettei lähes kahdensadantuhannen laukauksen aineistossa tästä aiheudu suurta haittaa. Joissain yksittäisissä otteluissa on kirjattu laukauksia virheellisesti useita kymmeniä samalle sekuntiluvulle. Esimerkiksi 19.11.2016 Rauman Lukon ja Mikkelin Jukureiden välillä pelatussa ottelussa on kirjattu 35 laukausta toisen erän viimeiselle sekunnille aikaan 39:59. Vastaavia tapauksia on aineistosta muutamia, ja kyse on luultavasti jonkinlaisesta pelikelloon tai kirjaamiseen liittyvästä viasta. Yhteensä aineistosta löytyy kuusi ottelua, joissa on merkattuna yli kymmenen laukausta yhdelle sekunnille. Aineistoa on onneksi melko paljon, joten yksittäisten

otteluiden poistamisesta ei aiheudu kovin suurta haittaa. Tällaisia virheitä sisältävät ottelut on poistettu aineistosta kokonaan, sillä virheelliset ajat vaikuttavat olennaisesti rebound-laukauksiin ja esimerkiksi laukaisukulman muutosnopeuteen. Käytännössä on mahdollista, että esimerkiksi maalin edessä olevassa kahakassa voidaan useita yksittäisiä sohaisuja tulkita laukaisuyritykseksi. Usean saman sekunnin laukauksen tilanteita on kuitenkin melko vähäinen määrä, joten virheelliset tilanteet voidaan löytää tarkastelemalla graafisesti laukausten sijainteja. Viiden samanaikaisen laukauksen tilanteet ovat kaikki selvästi mahdottomia sijaintien perusteella. Neljän laukauksen tapauksia on selvästi enemmän, joten virheellisen tapauksen kynnysarvoksi on valittu viisi samanaikaista laukausta. Tämän johdosta aineistosta poistetaan ennen mallintamisprosessia yhteensä 14 ottelua. Muutamassa kauden 2014–2015 ottelussa on kirjattu laukauksia väärään päätyyn eli käänteisesti verrattuna kuvaan 3. Tämä on tapahtunut koko ottelun tai yksittäisen erän aikana. Nämä tapaukset on paikannettu ja sijainnit on korjattu peilaamalla väärään päätyyn merkatut laukaukset takaisin oikeaan päätyyn.

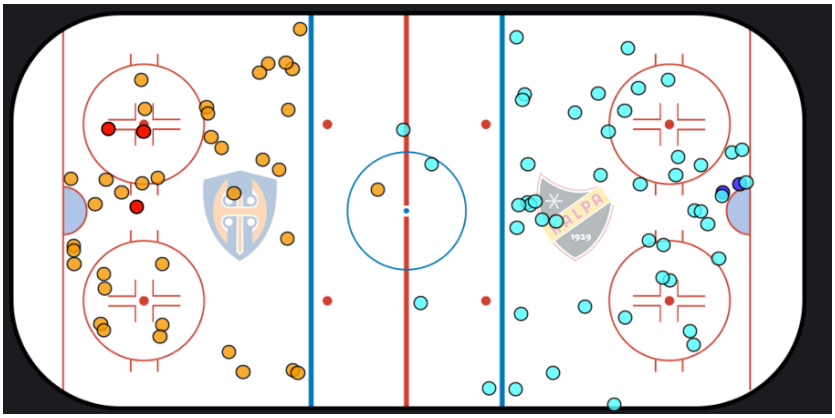
Rangaistuslaukaukset ovat hyvin poikkeuksellisia tilanteita, joita ei kuitenkaan voida erotella pöytäkirjan ja laukaisuaikojen perusteella riittävän hyvin. Rangaistuslaukaus kirjataan sille sekunnille, jolla tuomari on viheltänyt pelin poikki rangaistuslaukaukseen johtavan rikkeen tapahduttua. Rikkeen tapahtumisen ja pelikatkon välillä kuitenkin peli jää usein käymään, ja tilanteesta voi syntyä laukaus, jonka takia peli vihelletään poikki ja rangaistuslaukaus suoritetaan. Tämä viimeinen laukaus kuitenkin kirjataan samalle sekunnille itse rangaistuslaukauksen kanssa, jolloin nämä on mahdotonta erottaa toisistaan. Tästä syystä rangaistuslaukauksia ei ole huomioitu erikseen. Pelin aikaisia rangaistuslaukauksia tapahtuu kuitenkin melko harvoin, joten tästä valinnasta ei todennäköisesti aiheudu suurta harhaa.

3 Muuttujat

Seuraavassa luvussa tarkastellaan SM-Liigan sivuilta haetun raakadatan käsittelyä, mitä muuttujia maalitodennäköisyysmallissa voidaan käyttää, miten muuttujat määritellään ja mitä muita tekijöitä sivujen tiedoista on saatavilla. Sijaintitietojen lisäksi käytettävissä on laukausten ajankohdat ja ottelupöytäkirjasta on johdettavissa mm. pelaajamäärät.

3.1 Sijaintimuuttujat

Laukausten koordinaatit kirjataan otteluiden aikana pisteinä kaukalon kuvaan. Raakadatassa tieto sijainnista on ilmoitettu koodin 1 mukaisesti kaukalokuvan dimensioihin perustuvana prosenttiosuutena, joka tarkoittaa etäisyyttä kuvan reunasta. Esimerkiksi koodissa 1 oleva sijainti ”left: 85%” tarkoittaa sitä, että vaaka-akselin koordinaatti on 85 % kuvan leveydestä. Ensiksi raakadatan luvut muunnetaan sellaiseen muotoon, jossa kaukalon keskipiste on kohdassa $x = 0.50, y = 0.50$. Tämän jälkeen sijainnit voidaan skaalata metreiksi käyttämällä yleisiä SM-Liigan jääkiekkokaukaloiden mittoja, jolloin skaalaamalla saadut koordinaatit kuvaavat etäisyyttä vasempaan päätyyn ja alalaitaan. Kaukaloiden dimensioissa on pieniä eroja jäähallien välillä, mutta olennaista on se, että laukausten kannalta kriittiset etäisyydet ovat likimain samat jokaisessa kaukalossa. Tämä tarkoittaa sitä, että kaukaloiden mitat eroavat enimmäkseen reunoilla ja keskialueella. Muuttujien laskennan kannalta olennaisten sijaintien, kuten maalien ja maaliviivan, koordinaatit voidaan selvittää kuvasta piirtämällä koordinaatisto kaukalokuvan päälle. Laukausten sijainnit on kirjattu joukkueittain eri päätyihin kuten kuvassa 3.



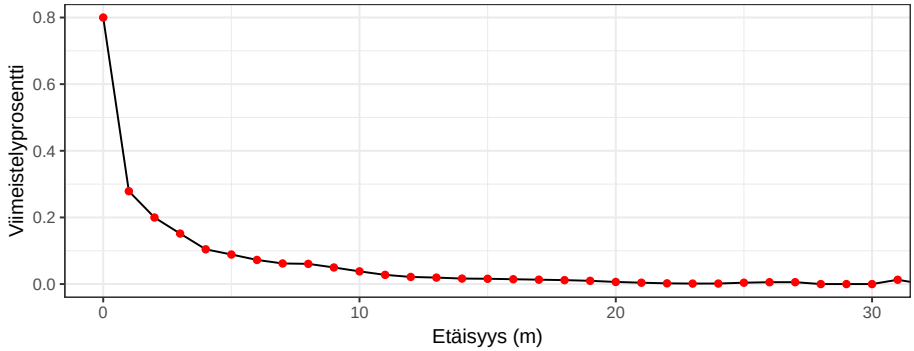
Kuva 3: Laukaisukartta kauden 2016-2017 1. finaalista Tappara-KalPa. Laukausten sijainnit ovat merkitty karttaan erivärisillä pisteillä. Maalit on merkitty tummalla värillä.
Lähde: http://liiga.fi/ottelut/2016-2017/playoffs/6619/seuranta/ (10.10.2017)

Laukaisupisteen etäisyys maalista lasketaan kahden pisteen välisenä etäisyytenä:

$$d = \sqrt{(x - x_{\text{maali}})^2 + (y - y_{\text{maali}})^2} \tag{1}$$

Kaavassa x ja y ovat laukauksen koordinaatit, ja maaliviivan keskipisteen sijainti

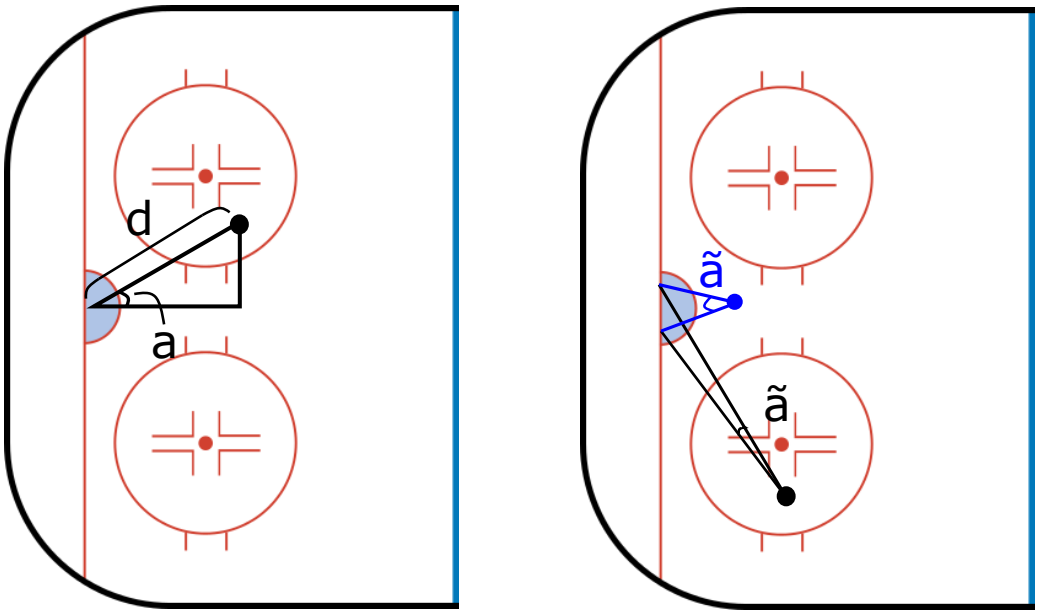
$(x_{\text{maali}}, y_{\text{maali}})$ tunnetaan. Etäisyys maalista on luultavasti tärkein yksittäinen tekijä maalitodennäköisyyden kannalta. Kuvassa 4 on piirretty laukausten viimeistelyprosentti etäisyyden suhteen. Viimeistelyprosentilla tarkoitetaan maaliin menneiden laukausten osuutta kaikista laukauksista. Hyvin läheltä maalia lauottaessa viimeistelyprosentti on huomattavasti korkeampi. Muuttujien välinen epälineaarisuus on syytä ottaa huomioon mallintamisessa. Laukauksen keskilinjakulma a lasketaan kuvan 5 va-



Kuva 4: Laukausten viimeistelyprosentti etäisyyden suhteen. Vaaka-akselin etäisyydet on pyöristetty alaspäin. Esimerkiksi ensimmäinen piste vastaa alle metrin etäisyydeltä tulleita laukauksia. Tyhjään maaliin lauottuja laukauksia ei ole huomioitu.

semman puolen mukaisesti kulmana kentän keskilinjan ja laukauksen välillä maalin keskeltä katsottuna. Tuolloin keskeltä tulleiden laukausten kulma on noin nolla ja päätyviivan läheltä noin 90 astetta. Laukaisukulma voidaan laskea myös kuvan 5 oikean puolen tapaan pelaajasta katsottuna. Tuolloin muodostetaan maalin tolppien ja laukaisupisteen kautta kulkeva kolmio, jolloin kulma lasketaan laukaisupistettä vastaavalle kärjelle. Tähän tapaan laskettu kulma $\tilde{a}$ muuttuu myös etäisyyden mukaan. Pelaajasta katsottuna kulma vastaa sitä, kuinka suurena maali näkyy. Kulmaa $\tilde{a}$ kutsutaan visuaaliseksi kulmaksi. Mallien eri tasoilla käytetään tilanteen mukaan eri kulmamuuttujia. Malleja tarkastellaan luvussa 4.

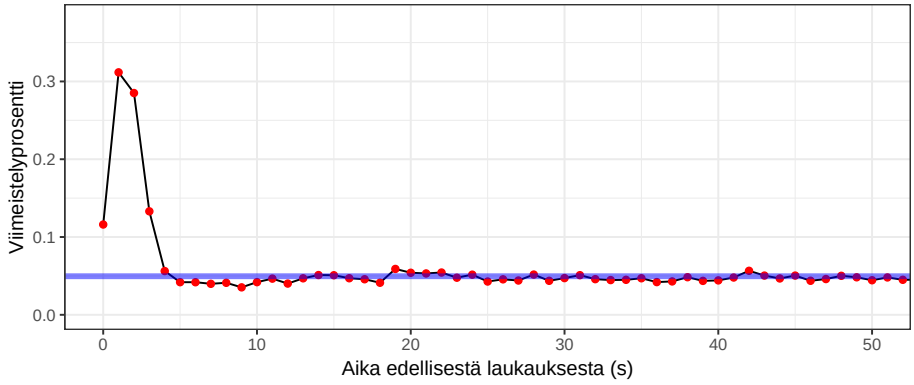
Aineistosta voidaan laskea myös laukaisukulman muutos Δa joukkueen edelliseen laukaukseen verrattuna. Koska kuvan 5 vasemman puolen mukaisesti laskettu kulma on yhtä suuri vasemmalla ja oikealla puolella, muutoksen Δa laskemista varten maalivahdista katsottuna oikean kenttäpuoliskon laukausten kulma muutetaan negatiiviseksi. Muutos on mielekästä laskea maalivahdista katsottuna, sillä suuret muutokset laukausten välillä ovat maalivahdeille haastavia. Erityisesti seuraavaksi käsiteltävien rebound-laukauksien yhteydessä kulman muuttuminen kannattaa huomioida.



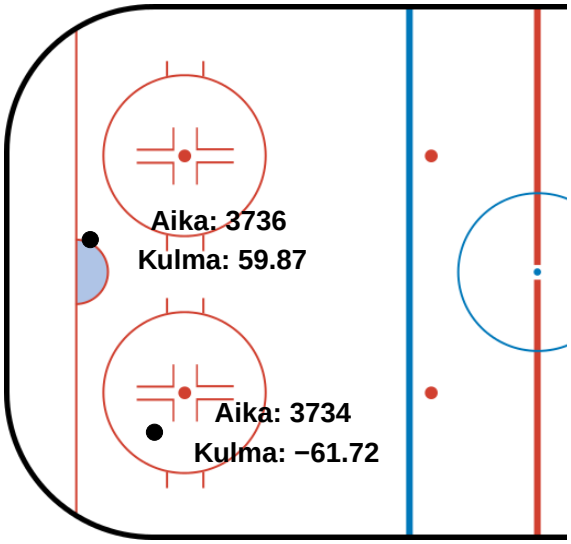
Kuva 5: Havainnollistus etäisyyden d, keskilinjakulman a ja visuaalisen kulman $\tilde{a}$ laskemisesta.

3.2 Aikaan liittyvät muuttujat

Rebound-laukauksella tarkoitetaan epäonnistuneen laukaisuyrityksen kimmokkeista, vastajoukkueen pelaajan blokkauksesta tai maalivahdin torjunnasta syntyvää uutta laukausta. Nämä ovat usein haastavia tilanteita puolustavalle joukkueelle, sillä niitä on vaikea ennakoida. Rebound-laukaukset voidaan tunnistaa datasta ajan perusteella, sillä ajallinen ero Δt joukkueen edelliseen laukaukseen on pieni. Rebound-laukaukset voidaan kategorisoida jonkin tietyn kynnysarvon perusteella, esimerkiksi $\Delta t \leq 3$ (s) voi toimia indikaattorina rebound-laukaukselle. Mallissa Δt voidaan pitää myös jatkuvana muuttujana, jolloin vältytään käyttämästä mielivaltaista kynnysarvoa. Kuvassa 6 on tarkasteltu viimeistelyprosentteja edellisestä laukauksesta kuluneen ajan suhteen. Kuvan perusteella vaikuttaa siltä, että alle kolmen sekunnin eroilla maalinteko onnistuu tavallista paremmin. Kolmen sekunnin jälkeen viimeistelyprosentti ei juurikaan eroa yleisestä viimeistelyprosentista. Joillain suurilla eroilla löytyy poikkeavia lukuja, mutta se on lähinnä pienen otoksen sattumaa. Tämän perusteella malleissa voisi käyttää selittäjänä rebound-laukausta kolmen sekunnin kynnysarvolla.



Kuva 6: Viimeistelyprosentti edellisestä laukauksesta kuluneen ajan suhteen. Vaakaviiva kulkee kaikkien laukausten viimeistelyprosentin kohdalla.



Kuva 7: Rebound-laukaus voidaan tunnistaa laukausten välisen ajan avulla.

3.3 Muut muuttujat

Ylivoimapeli on syytä huomioida mallissa, sillä laukausten viimeistelyprosentti on ylivoimatilanteissa korkeampi verrattuna tavalliseen peliin tasakentällisin. Tasakentällisin maalien osuus kaikista laukauksista on 0.044, yhden pelaajan ylivoimalla 0.067 ja kahden pelaajan ylivoimalla 0.099. Erityisesti kahden pelaajan ylivoimalla puolustava joukkue on selvästi epäedullisessa asemassa. Ylivoimatilanteisiin liittyy myös omia erityispiirteitään, joita ei tämän hetken datalla voida huomioida. Hyökkäävät pelaajat pyrkivät usein häiritsemään maalivahdin näkökenttää sijoittumalla maalin edustalle. Tuolloin kaukaa lähtevät laukaukset ovat tavallista vaarallisempia. Hyökkää-

jät voivat myös yrittää ohjata kaukolaukauksia. Tietoa ohjauksesta ei ole saatavilla, mutta laukaisuyritys kuitenkin merkitään ohjauksen sijaintiin. Lisäksi ylivoimapelissä joukkueet pyrkivät luomaan laukauksia suoraan poikittaissyötöistä tai maalivahdin kannalta vaikeasti arvioitavista laukaisukulmista. Esimerkiksi maalin takaa lähtevät syötöt maalin edustalle ovat maalivahdin kannalta erittäin haastavia. Syötöistä ei ole saatavilla tietoja, joten niitä ei voi huomioida maalitodennäköisyysmalleissa. Mallintamisessa voidaan käytännön pohjalta olettaa suuremman ylivoiman olevan aina vaarallisempi, eli kahden ja kolmen pelaajan ylivoimat ovat yhden pelaajan ylivoimia vaarallisempia. Mallissa käytetään ylivoimatilanteiden muuttujana kenttäpelaajien määrän erotusta: $n_{\text{yv}} = n_{\text{hyök}} - n_{\text{puol}}$.

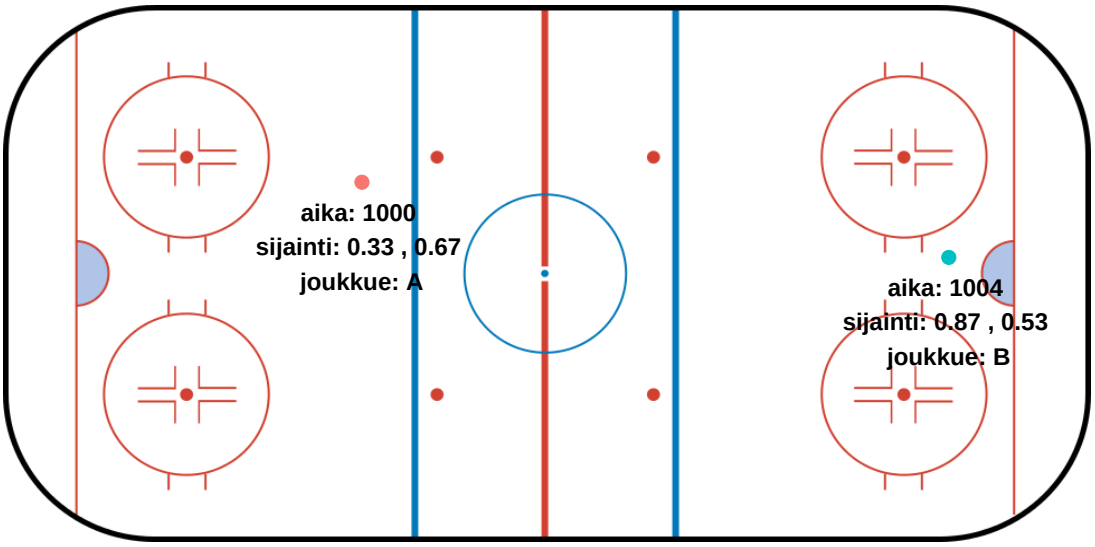
Läpiajot ja ylivoimaiset hyökkäykset ovat tavallista laukausta vaarallisempia maalipaikkoja. Läpiajolla tarkoitetaan tilannetta, jossa hyökkäävä pelaaja kohtaa vastustajan maalivahdin ilman, että välissä on puolustavia pelaajia. Ylivoimaisella hyökkäyksellä tarkoitetaan yksittäistä pelitilannetta, jossa hyökkäävällä joukkueella on tilanteessa enemmän pelaajia kuin puolustavalla joukkueella, esimerkiksi kaksi hyökkääjää yhtä puolustajaa vastaan. Tällaisia tilanteita ei kuitenkaan voida tunnistaa pelkästään laukausten sijaintijen ja aikojen perusteella. Osa läpiajoista ja nopeista hyökkäyksistä saadaan kuitenkin tunnistettua nopeusmuuttujan avulla. Jos esimerkiksi kotijoukkue laukoo omalla hyökkäysalueellaan ja pienen ajan kuluttua vierasjoukkue laukoo omalla hyökkäysalueellaan, on hyökkäys ensinnäkin ollut varsin nopea ja kyseessä on saattanut olla läpiajo. Tilannenopeus s laukaukselle i määritellään:

$$s_i = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}{t_i - t_{i-1}}, \tag{2}$$

eli lasketaan laukauksen etäisyys edelliseen laukaukseen jaettuna niiden välisellä ajalla. Nopeusmuuttuja s lasketaan eräkohtaisesti, jolloin erien ensimmäisten laukausten tilannenopeuksia ei lasketa yli erätauon vertaamalla edellisen erän viimeiseen laukaukseen. Pisteeksi (x_0, y_0) on luontevaa määritellä kaukalon keskipiste, josta jokainen erä aloitetaan, jolloin t_0 on erän alku.

3.4 Muokkaukset

Edellisissä kappaleissa käsitellyt muuttujat on johdettu laukausten sijainneista ja ajankohdista. Laukauksille on tarjolla myös muita tietoja, joita voi hyödyntää malleis-



Kuva 8: Esimerkki mahdollisesta läpiajosta.

sa. Yleensä otteluiden loppuhetkillä, toisen joukkueen johtaessa peliä yhdellä tai kahdella maalilla, häviöllä oleva joukkue yrittää hakea lisämaaleja suuremmalla riskillä käyttämällä maalivahdin sijasta ylimääräistä kenttäpelaajaa. Tuolloin johtoasemassa oleva joukkue pääsee laukomaan tyhjään maaliin. Tilanne on tavanomaiseen peliin verrattuna niin poikkeuksellinen, että maalitodennäköisyysmalleissa tyhjää maalia kohti lauotut laukaisuyritykset kannattaa jättää huomioimatta, sillä tavoitteena on mallintaa tyypillisten pelitilanteiden maalitodennäköisyyksiä. Tyhjien maalien sisällyttäminen tuo mukanaan harhaa esimerkiksi alivoimalla suoritettujen laukausten todennäköisyyksiin, sillä ilman maalivahtia pelaava joukkue saavuttaa yleensä ylivoimatilanteen. Tuolloin alivoimalla pelaava joukkue laukoo ”tyhjiin”. Tyhjät maalit voivat myös sekoittaa malleissa etäisyyden vaikutusta todennäköisyyksiin.

Joukkueiden aloittavat maalivahdit ja tiedot maalivahtien vaihdoista on sisällytetty seurantasivun ottelupöytäkirjaan, joten maalivahtien vaihtokartan muodostamista varten voidaan kirjoittaa algoritmi, jonka avulla jokaiselle laukaukselle saadaan määritettyä torjuva maalivahti. Laukauksen suorittanut pelaaja ja pelaajan tunnistenumero on kirjattuna laukaisupisteiden yhteydessä koodin 1 tapaan. Pelaaja ja maalivahti ovat mukana malleissa satunnaisefekteinä. Tätä käsitellään tarkemmin luvussa 4. Pelaajien kätisyys haetaan aineistoon erillisestä tiedostosta. Kätisyys kertoo laukooko pelaaja oikealta vai vasemmalta puolelta. Malleja tarkasteltaessa huomattiin, että kä-

tisyyden sijasta on parempi tarkastella kätisyyttä suhteessa laukauksen sijaintiin, eli tuleeko laukaus kätisyyden suhteen sisä- vai ulkokaistalta. Esimerkiksi vasenkätisen (vasen käsi alhaalla pelaavan) pelaajan laukoessa pelaajasta katsottuna kentän oikealta puolelta kyseessä on sisäkaistalaukaus, jossa lapa on lähempänä keskilinjaa. Tällainen laukaus voi olla helpompi tähdätä maalin kumpaankin kulmaan. Sen sijaan ulkokaistalaukaus, jossa vasenkätinen laukoo kentän vasemmalta puolelta, voi olla maalivahdille helpompi arvioida. Toisaalta, keskelle sijoittuneen puolustajan on hankalampi häiritä ulkokaistalaukausta.

Ylivoimalaukausten määrittämiseksi tarvitaan tieto joukkueiden pelaajamääristä laukausten tapahtumahetkellä. Pelaajamääriä ei ole suoraan tarjolla, mutta seurantasivun ottelupöytäkirjaa voidaan käyttää apuna. Kenttäpelaajien määrän selvittämiseksi on kirjoitettu algoritmi, joka selvittää kummankin joukkueen pelaajamäärän siihen vaikuttavien tapahtumien, eli jäähyjen ja ylivoimamaalien, perusteella. Algoritmissa hyödynnetään ottelupöytäkirjassa ilmoitettuja jäähyjen alkamisaikoja ja ylivoimamaaleja. Henkilökohtaiset rangaistukset eivät vaikuta pelaajamääriin, joten niitä ei tarvitse käsitellä. Lisäksi valtaosa toistensa kumoavista yhtäaikaisista rangaistuksista voidaan sivuuttaa. Jäähyjen päättymisaikoja ei ole ilmoitettu pöytäkirjassa lainkaan, joten ne on pääteltävä alkamisaikojen ja ylivoimamaalien perusteella. Kun päättymisajat on selvitetty, on tiedossa kaikki ajanhetket, jolloin kenttäpelaajien määrä muuttuu. Itse algoritmissa nämä muutospisteet käydään läpi siten, että samalla ylläpidetään joukkueen ”jäähyaitiota”, jolloin aitiossa olevien pelaajien määrän perusteella saadaan kenttäpelaajien määrä selville.

Ottelupöytäkirjan data ei kuitenkaan ole täydellistä, joten algoritmissa joudutaan tekemään oletuksia. Tavallisten kahden ja viiden minuutin rangaistusten käsittely onnistuu, mutta ongelmia tuottavat 2+2 minuutin tuplajäähyt. 2+2 minuutin jäähy on käytännössä kaksi peräkkäistä kahden minuutin jäähyä, joista jälkimmäinen alkaa ensimmäisen päätyttyä. Niitä tuomitaan yleensä korkealla mailalla pelaamisesta rikkeen ollessa vakava ja nk. isojen rangaistusten yhteydessä. Pöytäkirjassa 2+2 minuutin jäähyt ilmoitetaan kahtena erillisenä kahden minuutin jäähynä samalle pelaajalle samaan aikaan. Niitä ei siis voida yksiselitteisesti erottaa kahdesta samaan aikaan tuomitusta erillisestä jäähystä, jolloin jäähyt kärsitään samaan aikaan ja joukkue pelaa kahden pelaajan alivoimalla. Kuitenkin korkeasta mailasta johtuneet 2+2

minuutin jäähyt voidaan tunnistaa sillä oletuksella, että kyseessä ei ole kaksi erillistä rikettä. Algoritmi ei siis ole tästä johtuen aivan täydellinen, mutta valtaosa pelaajamääristä menee kuitenkin oikein, ja virheitä tapahtuu vain muutamassa harvinaisessa tilanteessa. Vertailukohtana NHL:n tarjoamassa vapaassa datassa on saatavilla pelaajien vaihtokartat, joiden perusteella saadaan selville, mitkä pelaajat ovat olleet milläkin ajanhetkellä kentällä. Tämän perusteella oikeat pelaajamäärät saadaan selville helposti ilman erityistä algoritmia. SM-Liigan vaihtokartat eivät kuitenkaan ole vapaasti saatavilla.

Taulukko 1: Muuttujataulukko

Muuttuja	**Symboli**	**Selite**
Etäisyys	d	Laukauksen etäisyys maaliviivan keskikohdasta (kuva 5) .
Keskilinjakulma	a	Laukauksen ja keskilinjan välinen kulma maalivahdista katsottuna (kuva 5).
Visuaalinen kulma	$\tilde{a}$	Maalin näkymiskulma pelaajasta katsottuna (kuva 5).
Kulman muutos	Δa	Kulman muutos joukkueen edelliseen laukaukseen
Rebound	$r = I(\Delta t \leq 3)$	Edellisestä laukauksesta kulunut aika sekunneissa (kuva 7).
Ylivoima	n_{yv}	Monenko pelaajan ylivoima. Negatiivinen alivoimille.
Tilannenopeus	s	Etäisyys edelliseen laukaukseen jaettuna ajalla (kuva 8).
Kätisyys/kaista	w	Ulkokaista vai sisäkaista
Pelaaja	P	Laukova pelaaja
Maalivahti	G	Torjuva maalivahti
Paikkakunta	R	Pelipaikkakunta

4 Mallit

Seuraavassa luvussa keskitytään mallintamisprosessiin. Aluksi tarkastellaan erilaisia tapoja maalitodennäköisyyksien mallintamiseen. Työssä käytetyn yleistetyn lineaarisen sekamallin teoriaa ja splinien hyödyntämistä mallinnusprosessissa käydään läpi lyhyesti. Lopuksi tarkastellaan työssä sovitettavan mallin osamalleja ja niiden muuttujia kullakin tasolla. Mallin rakenteen tarkoituksena on ottaa huomioon laukaisutilanteen tapahtumaketjumainen rakenne: maalille asti päästäkseen on ensin vältettävä

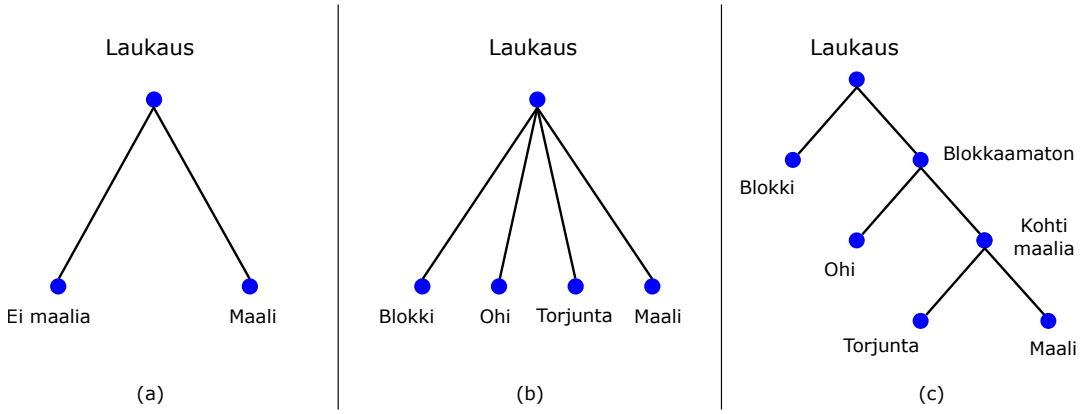
blokkaus ja ohilaukominen.

4.1 Erilaisia tapoja mallintaa maalitodennäköisyyksiä

Maalitodennäköisyyksiä voidaan mallintaa ilman varsinaisia tilastollisia malleja käyttämällä esimerkiksi erilaisia ruudukkomenetelmiä, joissa kenttä jaetaan alueisiin ja maalitodennäköisyyden estimaattina käytetään alueen laukausten viimeistelyprosenttia. Aluejako voi perustua joihinkin tiettyihin sektoreihin, tai jos laukaisupisteitä on riittävän paljon, voidaan kentästä muodostaa tiheä hilakuvio. Aluejakoihin perustuvien menetelmien ongelmana ovat rajatapaukset, joissa esimerkiksi kaksi hyvin samankaltaista ja lähekkäistä laukausta saavat toisistaan liialti poikkeavat todennäköisyysestimaatit aluerajan kulkiessa juuri näiden pisteiden välissä. Tuolloin hyvin minimaalinen sijainnin siirto vaikuttaa turhan karkeasti ennusteeseen. Alueiden välisellä tasoittelulla tältä ongelmalta kuitenkin vältytään. Reuna-alueilla dataa ei kuitenkaan ole tarpeeksi luotettavien ennusteiden muodostamiseksi. Kyseisessä menetelmässä jatkuvia muuttuja – tässä tapauksessa etäisyys ja kulma – on yksinkertaisuuden vuoksi kategorisoitu. Parametrisissa malleissa nämä muuttujat voidaan pitää jatkuvina, jolloin vältytään kategorisoinnin ongelmilta. Lisäksi yksinkertaisilla menetelmillä ei välttämättä pystytä huomioimaan sijainnin lisäksi muita taustamuuttujia ilman ylimääräistä kategorisointia. Ruudukkomenetelmän etu on se, että se mahdollistaa monenlaiset interaktiot sijainnin ja kulman välillä. Yksinkertaisin lähestymistapa maalitodennäköisyyksien tilastolliseen mallintamiseen on käyttää logistista regressiomallia, jossa vasteena käytetään indikaattoria maalille. Logistisen regression sijasta voidaan käyttää myös jotain muuta todennäköisyyksien mallintamiseen soveltuvaa menetelmää. Esimerkiksi multinomiaalisella logistisella regressiolla (Hosmer et al., 2013, s. 260) saadaan estimoitua todennäköisyydet myös muille laukauksen lopputuloksille, eli maalin lisäksi torjunnalle, ohilaukaukselle ja blokille. On myös mahdollista soveltaa neuroverkkoja ja muita todennäköisyysperusteisia luokittelumenetelmiä tai sovittaa malli bayesiläisittäin.

4.2 Mallin rakenne

Tässä työssä malli rakennetaan kuvassa 9(c) esitetyn graafin mukaisesti. Tällä tavalla tehtynä mallintaminen etenee samaan tapaan kuin oikea pelitilanne. Ensiksi mallinnetaan todennäköisyys blokatuksi tulemiselle. Tämän jälkeen mallinnetaan todennä-



Kuva 9: Kolme erilaista tapaa mallintaa maalitodennäköisyyksiä.

köisyys ohilaukaukselle ja lopulta maalille. Sovitetaan siis kolme erillistä mallia. Ensimmäisessä mallissa käytetään kaikkia laukauksia ja vastemuuttujana indikaattoria blokatulle laukaukselle, jolloin mallin ennusteena saadaan blokkaamisen todennäköisyys. Tämän jälkeen karsitaan aineistosta blokatut laukaukset pois ja rakennetaan seuraava malli. Toisessa vaiheessa mallinnetaan ohilaukomisen todennäköisyyttä aineistoon, joka sisältää blokkaamattomat laukaukset. Vasteena siis käytetään indikaattoria ohilaukaukselle ja ennusteena saadaan ohilaukauksen todennäköisyys. Lopulta aineistosta karsitaan myös ohilaukaukset pois, jolloin jäljellä ovat vain maalia kohti menneet laukaukset, eli maalivahdin torjumat laukaukset ja maalit. Kolmannessa mallissa mallinnetaan maalin todennäköisyyttä aineistoon, joka sisältää laukaukset maalia kohti. Tällaisen mallintamistavan etuna on se, että koko mallin eri tasoilla voidaan huomioida eri muuttujia. Lopullinen maalitodennäköisyys p lasketaan tulona

$$p = \Pr(\text{ei blokata}) \Pr(\text{ei ohi}|\text{ei blokata}) \Pr(\text{maali}|\text{ei blokata}, \text{ei ohi}). \tag{3}$$

4.3 Epälineaarisuudet

Kuten kuvasta 4 nähdään, laukaisupisteen etäisyyden ja viimeistelyprosentin yhteys on hyvin epälineaarinen. Epälineaarisuudet voidaan huomioida käyttämällä regressiosplinejä (Wegman & Wright, 1983). Regressiosplini on solmupisteiden perusteella paloittain määritelty polynomi, joka on sileä solmupisteissä. Luonnollinen splini määritellään lineaariseksi muuttujan arvoalueen ulkopuolella. Tekniikan ideana on korvata alkuperäinen muuttuja x epälineaarisilla kantafunktioilla $f_i(x)$, $i = 1, \ldots, K+1$, jossa K on solmupisteiden lukumäärä. Muuttujan x arvoalue jaetaan osiin solmupisteiden perusteella ja kuhunkin osaan sovitetaan oma funktio f_i. Solmupisteiden valintaan

on useita menetelmiä. Manuaalisen valitsemisen lisäksi voidaan käyttää esimerkiksi muuttujan kvantiilipisteitä tai ristiinvalidointia hyödyntäviä tekniikoita pisteiden valitsemiseksi (Hastie et al., 2001). Opetusaineistoon tehtyjen kokeilujen perusteella kaksi kvantiilipistettä ($\frac{1}{3}$ ja $\frac{2}{3}$) solmupisteinä on riittävän hyvä valinta AIC-kriteerin (Akaike, 1974) ja kalibraatioiden perusteella. Muunnoksia käytetään malleissa etäisyyteen d ja kulmiin a ja $\tilde{a}$. Mallien sovituksessa hyödynnetään R-paketin `splines` funktiota `ns()`. Splinien perusteoriaa käsitellään tarkemmin lähteessä (de Boor, 1978) ja tilastotieteen sovelluksia mm. teoksissa (Wegman & Wright, 1983) sekä (Hastie et al., 2001).

4.4 Logistinen sekamalli

Olkoon vektori $\boldsymbol{y} = (y_1, \ldots, y_n)$ mallin vastemuuttuja, joka on kussakin osamallissa indikaattorimuuttuja laukauksen lopputulostyypille (blokki, ohi, maali). Odotusarvovektori vastemuuttujalle on $\boldsymbol{\pi} = E(\boldsymbol{y})$. $\boldsymbol{X}$ on mallin selittäjät sisältävä $(n \times q)$-dimensioinen design-matriisi. Kokonaisluku q tarkoittaa mallin parametrien määrää. Parametri $\boldsymbol{\beta}$ on regressiokertoimet sisältävä $(q \times 1)$-vektori. Kaksiarvoisen muuttujan todennäköisyyksien mallintamiseen sopiva tavallinen logistinen regressiomalli määritellään

$$\text{logit}(E(\boldsymbol{y})) = \text{logit}(\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{4}$$

jossa linkkifunktiona on logit-funktio $\text{logit}(\boldsymbol{\pi}) = \log(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}})$.

Logistisessa sekamallissa huomioidaan yksilöiden tai klustereiden vaikutus satunnaisefektien $\boldsymbol{u}$ avulla. Aluksi oletetaan, että havainnot y_i noudattavat Bernoulli-jakaumaa ehdolla satunnaisefektit u_i. Malli voidaan kirjoittaa

$$\text{logit}(E(\boldsymbol{y}|\boldsymbol{u})) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tag{5}$$

jossa $\boldsymbol{Z}$ on satunnaisefektien design-matriisi. Oletetaan myös, että satunnaisefektit ovat multinormaalijakautuneita: $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. Havaintojen ja satunnaisefektien yhteistiheysfunktio $f(\boldsymbol{y}, \boldsymbol{u})$ voidaan kirjoittaa ehdollisten jakaumien avulla

$$f(\boldsymbol{y}_i, \boldsymbol{u}_i) = f(\boldsymbol{y}_i|\boldsymbol{u}_i) f(\boldsymbol{u}_i). \tag{6}$$

Mallin kiinteät parametrit $\boldsymbol{\beta}$ estimoidaan maksimoimalla $\boldsymbol{y}$:n marginaalinen uskottavuusfunktio, joka lasketaan integroimalla marginaalitiheydestä latentit satunnaise-

fektit pois.

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{u}) f(\boldsymbol{u}) d\boldsymbol{u}. \tag{7}$$

Integraalia ei saada ratkaistua eksaktisti yleistetyille lineaarisille sekamalleille, joissa käytetään epälineaarista linkkifunktiota, joten ratkaisussa hyödynnetään erilaisia approksimaatioita. Menetelmiä parametrien estimointiin käsitellään Pinheiron ja Batesin artikkelissa (Pinheiro & Bates, 1995), jossa tarkastellaan mm. Laplacen approksimaatiota ja AGQ-menetelmää (*adaptive Gaussian quadrature*). Menetelmät on implementoitu R-paketissa `lme4` (Bates et al., 2015), jota käytetään tämän työn mallien sovittamisessa. AGQ-menetelmä on toteutettu vain yhden satunnaisefektin malleille, joten jokaisessa osamallissa hyödynnetään Laplacen approksimaatioon perustuvaa laskentaa. Algoritmi koostuu kolmesta osasta:

1. Estimoidaan satunnaisefektien ehdolliset moodit $\hat{\mathbf{u}}$ käyttämällä PIRLS-menetelmää (*Penalised Iteratively Reweighted Least Squares*, (Bates, 2011)).

2. Approksimoidaan kaavan 7 integraalia käyttämällä Laplacen approksimaatiota $\hat{\mathbf{u}}$:n ympäristössä.

3. Sijoitetaan saatu approksimaatio mallin logaritmiseen uskottavuusfunktioon, joka optimoidaan $\boldsymbol{\beta}$:n ja $\boldsymbol{\Sigma}$:n suhteen.

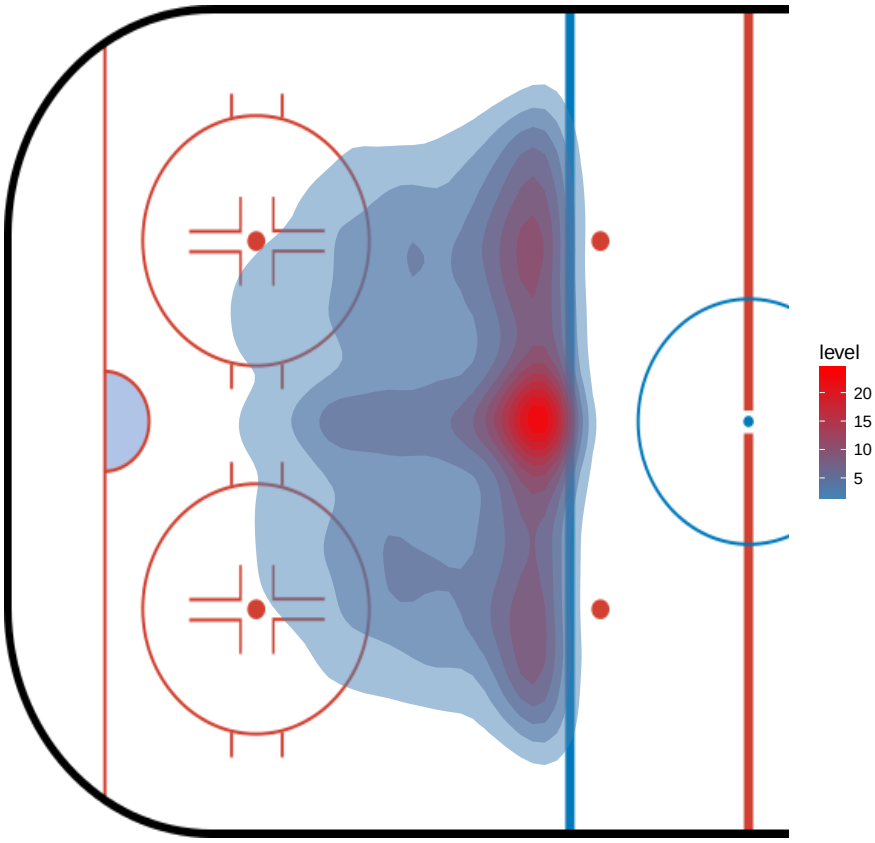
Algoritmin teknisiä yksityiskohtia käsitellään tarkemmin edellä mainitussa artikkelissa (Pinheiro & Bates, 1995).

4.5 Osamallit

Seuraavaksi tarkastellaan jokaisen osamallin rakentamista ja muuttujien valintaa mallin eri tasoille.

4.5.1 Malli blokkauksille

Ensimmäisenä mallinnetaan laukausten blokkaamisen todennäköisyyttä käyttäen aineistoa, jossa on mukana kaikki laukaisuyritykset. Lähtökohtana blokkitodennäköisyyksien mallintamiseen toimii yksinkertainen logistinen regressiomalli. Kuvassa 10 on piirretty lämpökartta blokattujen laukausten sijainneista hyödyntämällä kaksiulotteista ydinestimointia, joka on toteutettu R-paketin `ggplot2` (Wickham, 2016a) funktiolla `stat_density2d`. Huomataan, että blokkiin jää enimmäkseen kaukaa maalista



Kuva 10: Lämpökartta blokattujen laukausten sijainneista. Kuva on muodostettu suorittamalla kaksiulotteinen ydinestimointi blokattujen laukausten sijainteihin.

ja erityisesti keskeltä lauottuja laukauksia. Etäisyyden d ja keskilinjakulman a käyttäminen selittäjänä vaikuttaa kannattavalta kuvan 10 perusteella, sillä sijainnit ovat keskimäärin kaukana maalista ja keskellä. Ylivoimatilanteen huomioiminen mallissa on aiheellista, sillä puolustavan joukkueen toiminta poikkeaa tavallisesta pelitilanteesta ja pelaajia on muutenkin vähemmän kentällä. Rebound-laukaukset ovat myös kiinnostavia, sillä ne ovat puolustavan joukkueen kannalta vaikeasti ennakoitavissa. Kappaleessa 3 mainitut erot paikkakuntien välisissä tulkinnoissa kannattaa ottaa huomioon malleissa. Pelipaikkakunta voidaan sisällyttää malliin vakiotermiin kiinnitettävänä satunnaiskomponenttina, jolloin logistinen regressiomalli laajennetaan yleistetyksi lineaariseksi sekamalliksi. Voidaan myös tarkastella, olisiko puolustavan tai hyökkäävän joukkueen sisällyttäminen malliin hyödyllistä. Seuraavassa kaavassa $\Pr(b_i)$ tarkoittaa todennäköisyyttä sille, että laukaus i blokataan. Malliyhtälö kirjoitetaan

$$\Pr(b_{ij}) = \text{logit}^{-1}\Big(\boldsymbol{\beta}_1\, \hat{\mathbf{f}}(d_i) + \boldsymbol{\beta}_2\, \hat{\mathbf{g}}(a_i) + \beta_3\, n_{\text{yv},i} + \beta_4\, r_i + \alpha_j^R\Big)$$
$$\alpha_j^R \sim N(\mu_R, \sigma_R^2),$$

missä i käy läpi kaikki pelipaikkakunnan laukaukset ja j pelipaikkakunnat. Luonnollisen kuutiosplinin mukaisia kantafunktioita merkitään $\hat{\mathbf{f}}(.)$ ja $\hat{\mathbf{g}}(.)$. Splinille määritetään kaksi solmupistettä, joten kantafunktioita on yhteensä kolme kappaletta. Tuolloin aineiston etäisyysmuuttuja d korvataan kantafunktioiden arvoilla. Esimerkiksi laukaukselle i $\hat{\mathbf{f}}(d_i) = \big(\hat{f}_1(d_i), \hat{f}_2(d_i), \hat{f}_3(d_i)\big)^T$ ja myös regressiokertoimia on kolme kappaletta $\boldsymbol{\beta}_1 = \big(\beta_{1,1}, \beta_{1,2}, \beta_{1,3}\big)$. Vastaavat merkinnät pätevät kulmamuuttujalle. Termit n_{yv} ja r ovat taulukon 1 mukaisesti ylivoiman määrä ja rebound-indikaattori. Satunnaisefekti α^R kiinnitetään mallin vakiotermiin pelipaikkakunnan mukaan.

4.5.2 Malli ohilaukauksille

Seuraavaksi aineistosta poistetaan blokatut laukaukset, jolloin jäljelle jäävät maalia kohti menneiden laukausten lisäksi ohilaukaukset. Olennainen selittäjä ohilaukomiselle on laukauksen visuaalinen kulma $\tilde{a}$, joka kertoo laukaisukohteen "koon" pelaajasta katsottuna. Lisäksi tutkitaan mm. ylivoimatilanteen ja rebound-laukausten vaikutusta. Myös ulko- ja sisäkaistalta laukominen huomioidaan mallissa (w, taulukko 1). Laukausten blokkaamisen yhteydessä tarkastellut erot paikkakuntien välillä vaikuttavat ainakin osittain heijastuvan myös ohilaukauksiin (kuva 2). On mahdollista, että tulkinnoissa on eroja paikkakunnittain. Esimerkiksi puolustavan pelaajan mailasta maalin ohi ohjaantuneet laukaukset saatetaan tulkita jossain blokiksi ja toisaalla ohilaukaukseksi. Pelipaikkakunta sisällytetään malliin satunnaisefektinä samaan tapaan kuin blokkausmallissa. Myös laukovan pelaajan vaikutus otetaan tässä vaiheessa mukaan malliin satunnaisefektinä. Regressiosplinien ja etäisyysmuuttujan käyttäminen ohilaukausten mallissa osoittautui tarpeettomaksi. Tästä kerrotaan lisää pohdintaosiossa. Malliyhtälö kirjoitetaan

$$\Pr(o_{ijk}) = \text{logit}^{-1}\Big(\beta_1\, \tilde{a}_i + \beta_2\, n_{\text{yv},i} + \beta_3\, r_i + \beta_4\, w_i + \alpha_j^R + \alpha_k^P\Big)$$
$$\alpha_j^R \sim N(\mu_R, \sigma_R^2)$$
$$\alpha_k^P \sim N(\mu_P, \sigma_P^2),$$

missä i käy läpi blokkaamattomat laukaukset, j paikkakunnat ja k pelaajat. Ohilaukauksen indikaattori on o. Satunnaisefekti kiinnitetään mallin vakiotermiin pelipaikkakunnittain ja pelaajittain (α^R ja α^P).

4.5.3 Malli maaleille

Kolmannessa vaiheessa aineistosta karsitaan blokkien lisäksi ohilaukaukset pois. Jäljellä ovat vain maaliin menneet ja maalivahdin torjumat laukaukset. Mallissa käytetään edelleen selittäjänä etäisyyttä d, visuaalista kulmaa $\tilde{a}$ ja kaistaa w. Ylivoimatilanne huomioidaan mallissa kenttäpelaajien määrän erotuksena n_{yv}, kuten kappaleessa 3 määriteltiin. Rebound-laukaukset ovat myös mukana mallissa. Rebound-laukaukset, joissa kulma muuttuu paljon alkuperäisestä laukauksesta, ovat haastavia tilanteita maalivahdille, joten reboundien ja keskilinjakulman muutoksen Δa yhteisvaikutusta kannattaa tutkia. Laukova pelaaja sisällytetään malliin satunnaisefektinä, jolloin esimerkiksi säännöllisesti yli odotusarvojen viimeisteleville pelaajille estimoituu korkeampi vakiotermi. Samaan tapaan voidaan huomioida torjuva maalivahti. Maalivahdin sisällyttäminen malliin on mielekästä tehdä tässä vaiheessa, sillä aiemmissa laukauksissa maalivahti ei ole joutunut torjumaan laukauksia. Kuvassa 8 esitetty nopeusmuuttuja jätettiin mallista pois, sillä regressiokerroin estimoituu negatiiviseksi. Se on kyseisen muuttujan tarkoituksen kannalta käänteinen, sillä tavoitteena on tunnistaa nopeita vastaiskuja ja läpiajoja, joiden pitäisi olla tavallista vaarallisempia maalipaikkoja. Muuttujan laskennassa hyödynnetään aikaa edelliseen laukaukseen, joka on mukana esimerkiksi kulman muutoksen laskennassa ja reboundin määrittelyssä. Intuitiivisesti väärän merkkinen kerroin voi siis johtua kollineaarisuudesta. Malli voidaan esittää muodossa:

$$
\begin{aligned}
\Pr(m_{ikl}) = \text{logit}^{-1}\Big(&\boldsymbol{\beta}_1\, \hat{\mathbf{f}}(d_i) + \boldsymbol{\beta}_2\, \hat{\mathbf{g}}(\tilde{a}_i) + \beta_3\, n_{\text{yv},i} + \beta_4\, r_i + \beta_5\, w_i + \\
&\beta_6\, \Delta a_i + \beta_7\, (\Delta a_i \times r_i) + \alpha_k^P + \alpha_l^G\Big) \\
\alpha_k^P &\sim N(\mu_P, \sigma_P^2) \\
\alpha_l^G &\sim N(\mu_G, \sigma_G^2),
\end{aligned}
$$

missä i käy läpi laukaukset, k pelaajat ja l maalivahdit. Δa symboloi laukauksen keskilinjakulman muutosta. Satunnaisefektit kiinnitetään jälleen mallin vakiotermiin pelaajan (α^P) ja maalivahdin osalta (α^G). Merkintä $\Delta a \times r$ tarkoittaa kulman muu-

toksen ja reboundin interaktiota. Osamallien muuttujat on koottu taulukkoon 2.

Taulukko 2: Osamallien muuttujat.

malli	kiinteät efektit
blokki	etäisyys, keskilinjakulma, ylivoima, rebound
ohi	visuaalinen kulma, ylivoima, rebound, kaista
maali	etäisyys, visuaalinen kulma, ylivoima, kaista, (rebound × kulman muutos)

malli	satunnaisefektit
blokki	paikkakunta
ohi	paikkakunta, pelaaja
maali	pelaaja, maalivahti

5 Menetelmiä mallien sopivuuden arviointiin

5.1 Log-tappio

Mallien vertailussa erityisen kiinnostuksen kohteena on todennäköisyysestimaatin tarkkuus. Eräs tapa tarkastella estimaattien tarkkuutta on laskea log-tappio (*log-loss*)

$$\text{log-loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \tag{8}$$

jossa N on havaintojen määrä, y_i on arvoja 0 ja 1 saava muuttuja ja $\hat{y}_i$ on mallin ennustama todennäköisyys onnistumiselle arvolle 1. Kyseessä on siis oikeiden vaihtoehtojen todennäköisyyksien logaritmisumman keskiarvo, joka on skaalattu positiiviseksi. Log-tappio on pienempi, eli rankaisee vähemmän, jos oikealle vaihtoehdolle annetaan korkea todennäköisyys. Malleja vertailtaessa pienempi arvo on parempi. Log-tappiota kutsutaan myös ristientropiatappioksi, sillä se voidaan ilmaista havaintojen ja ennusteiden välisenä ristientropiana, ja se liittyy läheisesti havaitun ja ennustetun jakauman väliseen Kullback-Leibler-divergenssiin (Buja et al., 2005) seuraavalla tavalla. Olkoon a havaintojen ja b ennusteiden jakauma,

$$a \in \{y, 1 - y\}, \qquad b \in \{\hat{y}, 1 - \hat{y}\}.$$

Jakaumien a ja b välinen ristientropia H voidaan kirjoittaa seuraavaan tapaan:

$$H(a,b) = -\sum_i a_i \log b_i = -y \log(\hat{y}) - (1-y)\log(1-\hat{y}), \tag{9}$$

joka on sama kuin yleinen log-tappio. Toisaalta ristientropia voidaan määritellä myös seuraavasti:

$$H(a,b) = H(a) + D_{\mathrm{KL}}(a||b), \tag{10}$$

jossa $D_{\mathrm{KL}}(a||b) = \sum_i a(i) \log \frac{a(i)}{b(i)}$ on havaitun ja ennustetun jakauman välinen Kullback-Leibler-divergenssi, joka mittaa jakaumien a ja b välistä eroa. Log-tappiolle saadaan siis tulkinta havaitun ja ennustetun jakauman välisenä erona, joka poikkeaa Kullback-Leibler-divergenssistä havaintojen entropian $H(a)$ verran. $H(a)$ on vakio ennusteiden b suhteen. Mallien sopivuutta arvioitaessa voidaan laskea testiaineistolle log-tappio, jota voidaan hyödyntää mm. muuttujajoukkojen valinnassa. Eräs viitteellinen arvo log-tappiolle saadaan laskemalla "kolikonheittomallin" log-tappio. Tällä tarkoitetaan mallia, joka antaa jokaiselle havainnolle todennäköisyyden $\hat{y} = \frac{1}{2}$. Tuolloin log-tappioksi saadaan aina $\log(2) \approx 0.693$, sillä oikeilla havainnoilla y_i ei ole tappion kannalta mitään väliä todennäköisyysestimaattien ollessa samat kummallekin vaihtoehdolle:

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \log(\frac{1}{2}) + (1-y_i)\log(\frac{1}{2}) = -\frac{1}{N}\sum_{i=1}^{N} \log(\frac{1}{2}) = \log(2).$$

Vertailu arvoon $\log(2)$ kertoo siis siitä, onko mallin ennusteista hyötyä verrattuna täysin satunnaiseen arvaukseen. Yleisesti log-tappiota käytetään pistearvona mallien välisessä vertailussa ja ristiinvalidoinnissa vaihtoehtoisena tunnuslukuna ennustevirheelle. Log-tappiota käytetään myös ennustekilpailujen pisteytyksessä. Esimerkiksi ennustekilpailuja järjestävän Kagglen (Becker) kilpailuissa ennustemallit pisteytetään yleensä log-tappion perusteella.

5.2 Mallin kalibraation tarkastelu

Mallin kalibraatiolla tarkoitetaan sitä, kuinka hyvin mallin tuottamat todennäköisyydet pitävät paikkaansa. Toteutuuko arvio 0.6 oikeasti noin 60 kertaa sadassa toistossa? Usein ennustetarkkuudella tarkoitetaan luokittelun onnistumista, jolloin kiinnostuksen kohteena on luokittelun tulos, joka on tässä tapauksessa maali tai ei maalia. Luokittelun tulos määräytyy yleensä sen perusteella, ylittääkö mallin tuottama todennäköisyysestimaatti arvon 0.5. Ennusteista voidaan esimerkiksi laskea testiaineiston

oikein luokiteltujen laukausten osuus kaikista testiaineiston laukauksista, mutta tässä työssä ei kyseistä tapaa käytetä. Kiinnostuksen kohteena ei varsinaisesti ole ennustetarkkuus, vaan ennemmin ennusteen taustalla olevan todennäköisyysarvon laatu. Kalibraatiota tarkasteltaessa verrataan todennäköisyysestimaatteja todellisiin osuuksiin. Oletetaan esimerkkitilanne, jossa n kappaletta havaintoja luokitellaan kahteen luokkaan, ja oletetaan myös malli, joka on ennustanut kaikki havainnot oikein: se on antanut todennäköisyysestimaatiksi jokaiselle oikealle luokalle esimerkiksi 0.75. Tällaisen mallin kalibraatio ei ole kovin hyvä, sillä ennusteet 0.75 eivät vastaa toteutunutta osuutta 1.0, vaikka kaikki ennusteet ovat luokittelun kannalta oikein. Mallien kalibraation tarkastelussa edetään Hosmer–Lemeshow-testin tapaan (Hosmer et al., 2013, s. 147). Ensin havainnot järjestetään ennustetun todennäköisyyden perusteella suuruusjärjestykseen. Tämän jälkeen havainnot jaetaan k kvantiiliryhmään siten, että kussakin ryhmässä on $\frac{1}{k}$ osuus kaikista havainnoista. Esimerkiksi tapauksessa $k = 10$ jokainen ryhmä sisältää 10 %:n osuuden havainnoista. Ryhmät ovat siis likimain yhtä suuria. Jokaiselle ryhmälle lasketaan todennäköisyysestimaattien vaihteluväli ennustetuista arvoista ja lisäksi toteutunut osuus. Kalibraatiota voidaan tarkastella vertaamalla ryhmän toteutunutta osuutta ryhmän todennäköisyyksien vaihteluväliin. Hyville kalibraatioille toteutuneet osuudet osuvat kvantiiliryhmän todennäköisyyksien vaihteluvälille. Kappaleen 6.1 kuvissa tulokset on esitetty graafisesti.

5.3 Tunnuslukujen ennustekyky

Prediktiivisyydellä tarkoitetaan yleisesti ennustekykyä. Urheilusarjoissa joukkuetason tunnuslukujen prediktiivisyydellä tarkoitetaan sitä, kuinka jokin tunnusluku korreloi tulevaisuuden tunnuslukujen kanssa. Yleensä kiinnostuksen kohteena ovat korrelaatiot tulevaisuuden voittoihin tai maaleihin. Ennustekyvyn tarkasteluissa joukkueen kausi jaetaan kahteen osaan jostain tietystä ajanhetkestä, esimerkiksi 20 pelatun ottelun kohdalta. Joukkueille lasketaan tunnusluvut molemmille kauden osille, jonka jälkeen tarkastellaan alku- ja loppukauden tunnuslukujen välisiä korrelaatioita. Yleensä kiinnostuksen kohteena ovat korrelaatiot joukkueen tulevaisuuden maaleihin. Edellä mainitun jakopisteen valinta on mielivaltainen, joten korrelaatiot voidaan laskea vaikka kaikille mahdollisille jakopisteille. Tuolloin hyödynnetään graafisia tarkasteluja. Joukkueet kannattaa käsitellä kausittain, sillä joukkueiden kokoonpanot, valmentajat ja suoritustasot vaihtelevat kausien välillä. Kaudesta 2014-2015 kauteen 2017-2018 saadaan siis yhteensä 59 joukkuekautta. Valitaan jakopisteiksi ottelut 15, 20, 25 ..., 45.

Tarkastellaan vain runkosarjapelejä. Prosessi etenee seuraavasti:

1. Jaetaan jokainen joukkuekausi jakopisteen kohdalta kahteen osaan: alku- ja loppukauteen.
2. Lasketaan molemmille osille tunnusluvut.
   (a) Viidellä viittä vastaan tehdyt maalit
   (b) Viidellä viittä vastaan laukaukset
   (c) Viidellä viittä vastaan tehty maaliodotusarvo
3. Lasketaan korrelaatiot edellisen kohdan alkukauden lukujen ja loppukauden viidellä viittä vastaan tehtyjen maalien välille.
4. Toistetaan kohdat 1–3 kaikille jakopisteille.

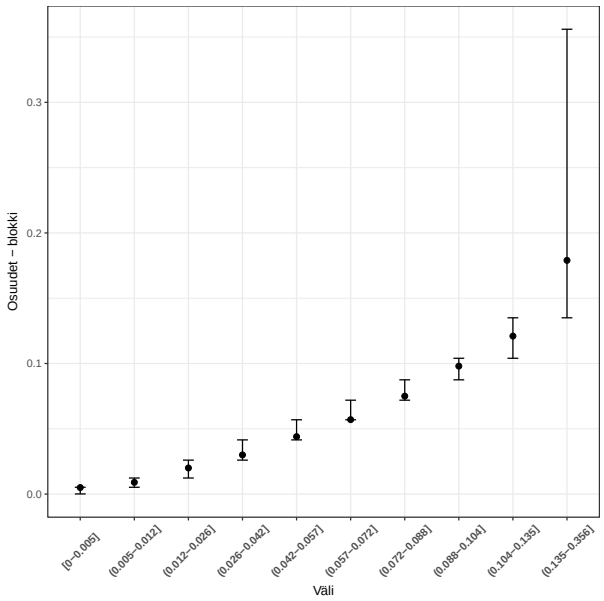
Tämän prosessin tuloksena saadaan hyökkäyspeliä kuvaavien tunnuslukujen korrelaatioita tulevaisuuden tehtyihin maaleihin. Sama toistetaan vastaaville puolustuspeliä kuvaaville tunnusluvuille, jolloin kohdan 3 korrelaatiot lasketaan loppukauden viidellä viittä vastaan päästettyihin maaleihin. Tarkastelu tehdään myös maalisuhteelle, laukaussuhteelle ja maaliodotusarvosuhteelle, jotka kuvaavat enemmän joukkueen kokonaissuoritusta. Esimerkiksi joukkueen maalisuhde lasketaan jakamalla tehdyt maalit joukkueen otteluiden kokonaismaalimäärällä. Laukaus- ja maaliodotusarvosuhteen laskenta tapahtuu vastaavasti. Hyökkäyslukujen korrelaatiot tulevaisuuden maaleihin, puolustuslukujen korrelaatiot tulevaisuuden päästettyihin maaleihin ja alkukauden suhdelukujen korrelaatiot loppukauden suhdelukuihin on esitetty luvussa 6.4.

6 Tulokset

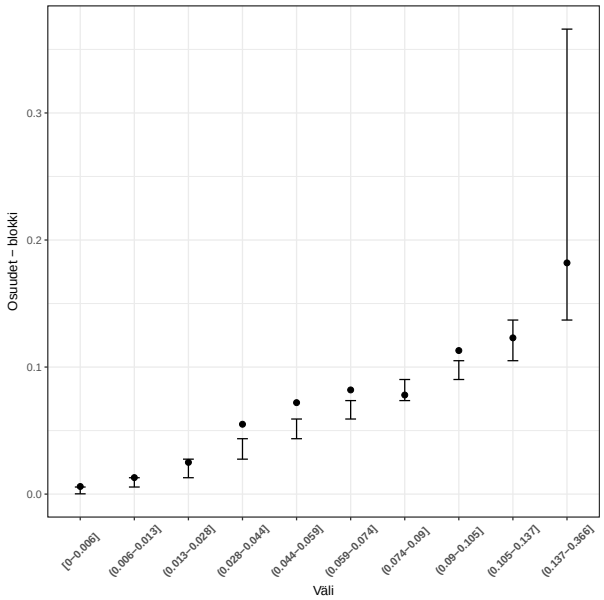
6.1 Mallien sopivuus

Seuraavassa kappaleessa arvioidaan mallien toimivuutta. Kalibraatioita tarkastellaan graafisesti opetus- ja testidataan tehtyjen kalibraatiokuvaajien avulla. Mallien muuttujajoukkojen valinnassa on hyödynnetty graafisten tarkastelujen lisäksi log-tappiota. Kalibraatiokuvaajissa kunkin kvantiiliryhmän ennusteiden vaihteluväli on merkitty pystyviivalla. Välien toteutuneet osuudet on merkitty pisteellä vastaavan välin kohdalle. Todennäköisyys kulkee kuvaajan pystyakselilla ja vaaka-akselille on eroteltu kvantiiliryhmät. Kuvaajien avulla saadaan selville, kuinka hyvin mallien ennustamat arvot vastaavat toteutuneita osuuksia. Esimerkiksi pisteen ollessa välin yläpuolella, ovat kyseisen välin ennusteet todellisia arvoja pienempiä. Kuvissa 11–16 on esi-

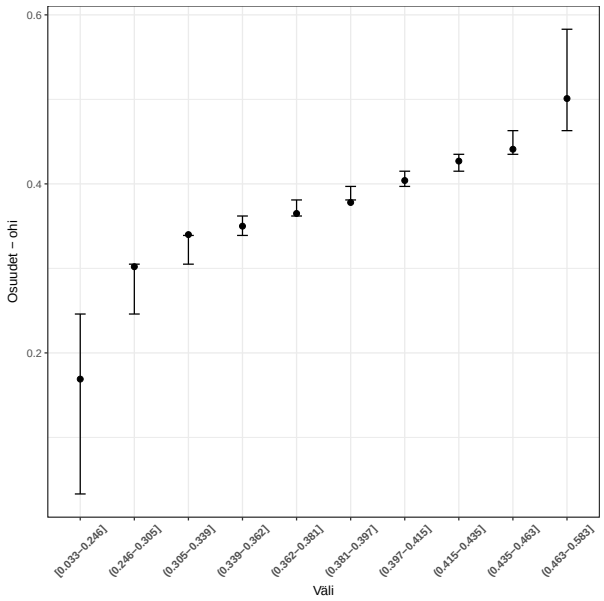
tetty kunkin osamallin kalibraatiokuvaajat opetus- ja testiaineistoon. Blokkausmallin opetusaineistoon tehty kalibraatiokuvaaja näyttää varsin hyvältä toteutuneiden osuuksien sijaitessa oikeilla väleillä. Testiaineistossa huomataan epätarkkuutta keskimmäisillä väleillä, joissa toteutuneet blokkiosuudet ovat ennustettua korkeampia. Kuvaajan mukaan välillä 0.028–0.074 testiaineiston ennusteet ovat hieman liian matalia. Ohilaukausten mallissa testiaineiston kalibraatio näyttäisi olevan kohtuullinen. Opetusaineistolle kuitenkin saadaan hieman liian korkeita ennusteita arvojen 0.3–0.4 läheisyydessä. Vastaavasti viimeisillä väleillä on hieman liian matalia ennusteita. Maalimallissa kalibraatiokuvaajat näyttävät varsin hyviltä. Ennusteiden kvantiilivälit ovat melko kapeita. Käytännössä tämä tarkoittaa sitä, että valtaosa maalia kohti lauotuista laukauksista on todennäköisyydeltään melko pieniä ja varsin pieni osuus on erityisen vaarallisia.



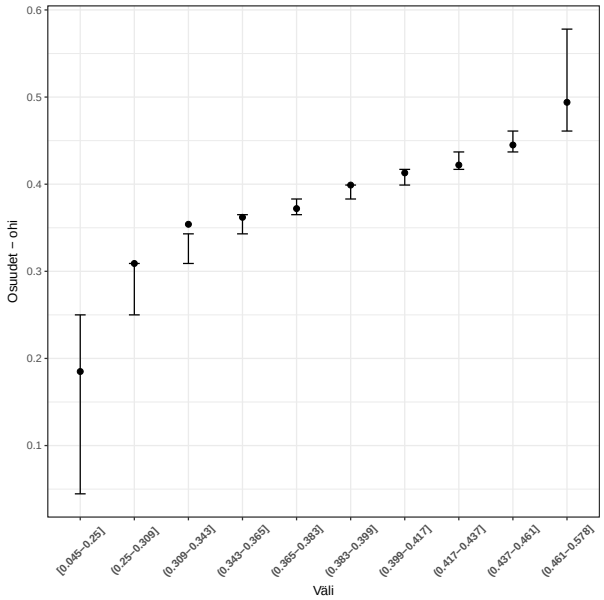
Kuva 11: Blokkausmallin kalibraatiokuvaaja opetusaineistoon. Pystyviiva kuvaa kvantiiliryhmän ennusteiden vaihteluväliä ja piste kyseisen välin ennusteiden toteutunutta osuutta.



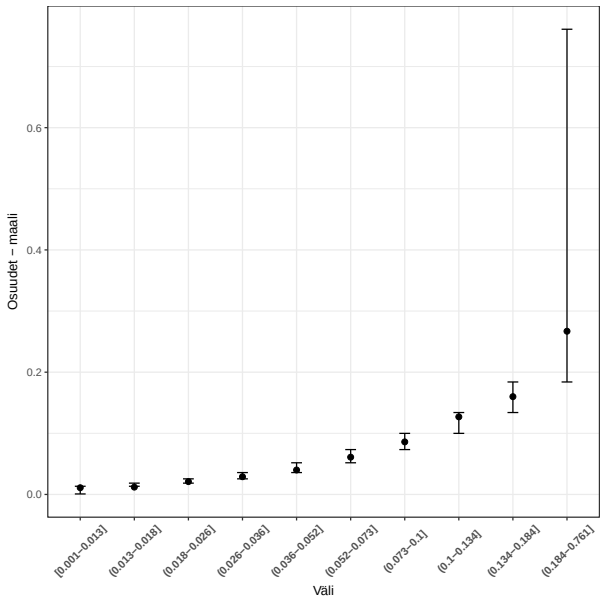
Kuva 12: Blokkausmallin kalibraatiokuvaaja testiaineistoon. Pystyviiva kuvaa kvantiiliryhmän ennusteiden vaihteluväliä ja piste kyseisen välin ennusteiden toteutunutta osuutta.



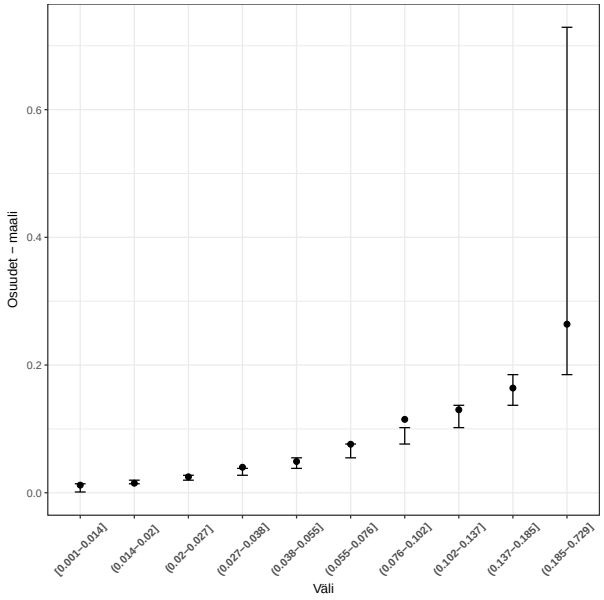
Kuva 13: Ohilaukausten mallin kalibraatiokuvaaja opetusaineistoon. Pystyviiva kuvaa kvantiiliryhmän ennusteiden vaihteluväliä ja piste kyseisen välin ennusteiden toteutunutta osuutta.



Kuva 14: Ohilaukausten mallin kalibraatiokuvaaja testiaineistoon. Pystyviiva kuvaa kvantiiliryhmän ennusteiden vaihteluväliä ja piste kyseisen välin ennusteiden toteutunutta osuutta.



Kuva 15: Maalimallin kalibraatiokuvaaja opetusaineistoon. Pystyviiva kuvaa kvantiiliryhmän ennusteiden vaihteluväliä ja piste kyseisen välin ennusteiden toteutunutta osuutta.



Kuva 16: Maalimallin kalibraatiokuvaaja testiaineistoon. Pystyviiva kuvaa kvantiiliryhmän ennusteiden vaihteluväliä ja piste kyseisen välin ennusteiden toteutunutta osuutta.

6.2 Mallien rakenteiden vertailu

Seuraavaksi vertaillaan erilaisten rakenteiden välisiä eroja. Opetusaineistoon sovitetaan kuvan 9 graafien (a) ja (b) mukaiset mallit, jonka jälkeen lasketaan ennusteet ja

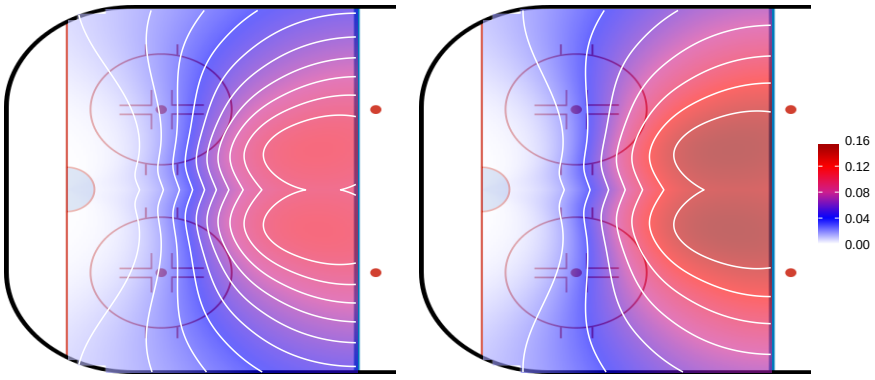
log-tappiot testiaineistolle. Graafia (a) vastaavaa malli sovitetaan logistisena regressiomallina käyttämällä samoja muuttujia kuin mallin (c) maalimallissa. Muuttujat löytyvät taulukosta 2. Vertailun vuoksi sovitetaan myös vastaava sekamalli, jossa satunnaisefektinä käytetään laukovaa pelaajaa ja torjuvaa maalivahtia. Graafia (b) vastaava malli sovitetaan käyttämällä multinomiaalista logistista regressiota samaan muuttujajoukkoon. Erot mallien log-tappioissa ovat melko pieniä.

Taulukko 3: Kuvassa 9 esitettyjen rakenteiden mukaisten mallien log-tappiot testiaineistossa.

Malli	Log-tappio testiaineistossa
9(a) kiinteät efektit	0.1306
9(a) sekamalli	0.1305
9(b)	0.1309
9(c)	0.1311

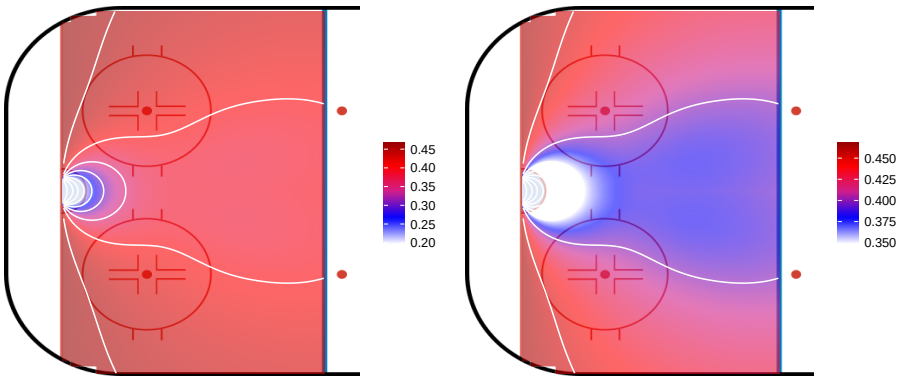
6.3 Mallin visualisointi

Osamalleja ja kokonaismallia voidaan visualisoida lämpökarttakuvien avulla. Karttakuvissa hyökkäysalueelle on asetettu tiheä ruudukko pisteitä, joille lasketaan sijaintimuuttujien arvot, eli etäisyydet ja kulmat kuvan 5 mukaisesti. Muille muuttujille annetaan kiinteät arvot. Tämän jälkeen jokaiselle pisteelle lasketaan todennäköisyysestimaatti, jonka perusteella kyseisen ruudun väri määräytyy. Lopuksi lämpökarttaa tasoitetaan lineaarisella interpolaatiolla lähekkäisten ruutujen välillä. Kuvassa 17 on visualisoitu laukauksen blokkaamisen todennäköisyydet tavallisen laukauksen eri sijainneille. Vasemmassa kuvassa on sijaintien todennäköisyydet tasakentällisin ja toisessa kuvassa ylivoimatilanteessa. Pelipaikkakuntaan liittyvä satunnaisefekti on asetettu nollaksi. Kuvasta nähdään, että kaukaa ja keskeltä lauottaessa blokatuksi tuleminen on todennäköisempää. Ylivoimatilanteissa blokkaukset ovat todennäköisempiä verrattuna peliin tasakentällisin.



Kuva 17: Lämpökarttakuva laukausten sijaintien blokkaustodennäköisyyksistä. Vasemmalla peli tasakentällisin ja oikealla ylivoimapeli.

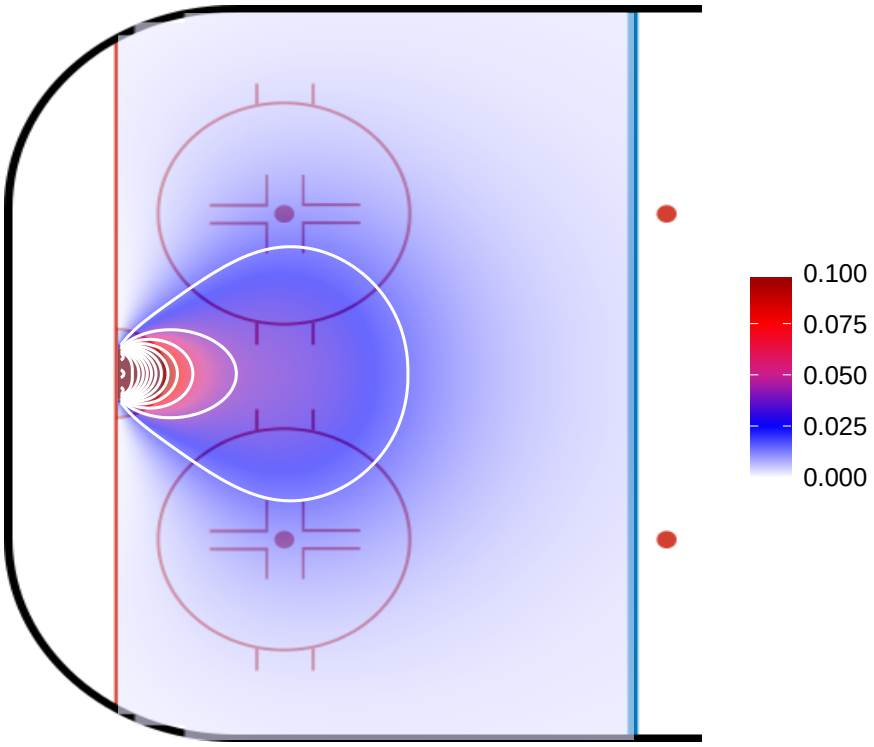
Kuvassa 9(c) esitetyn kokonaismallin toisen portaan mallin visualisaatio on esitetty kuvassa 18. Kyseessä on siis eri sijainneille ohilaukomisen ehdolliset todennäköisyydet, jossa ehtona on se, että laukausta ei blokata. Erot ovat yleisesti ottaen melko pieniä, mutta maalin lähellä ohilaukominen on luonnollisesti epätodennäköisempää. Toisessa kuvassa on havainnollisuuden vuoksi tehty sama kuva pienemmällä väriskaalalla, jotta erot alueiden välillä näkyisivät paremmin. Kuvasta huomataan, että kentän keskikaistalta, josta maali näkyy hyvin, ohilaukominen on epätodennäköisempää. Sen sijaan pienistä kulmista osuminen on vaikeampaa. Myös kauemmas siirryttäessä ohilaukominen käy hieman todennäköisemmäksi.



Kuva 18: Lämpökarttakuva ohilaukauksen todennäköisyydestä ehdolla, että laukausta ei blokata. Oikeanpuoleisessa kuvassa on piirretty erottuvuuden takia sama lämpökartta tiheämmällä väriskaalalla. Hyvin lähellä maalia todennäköisyydet putoavat nopeasti nollaa kohti.

Mallin lopulliset maalitodennäköisyydet tasakentällisin on visualisoitu kuvassa 19.

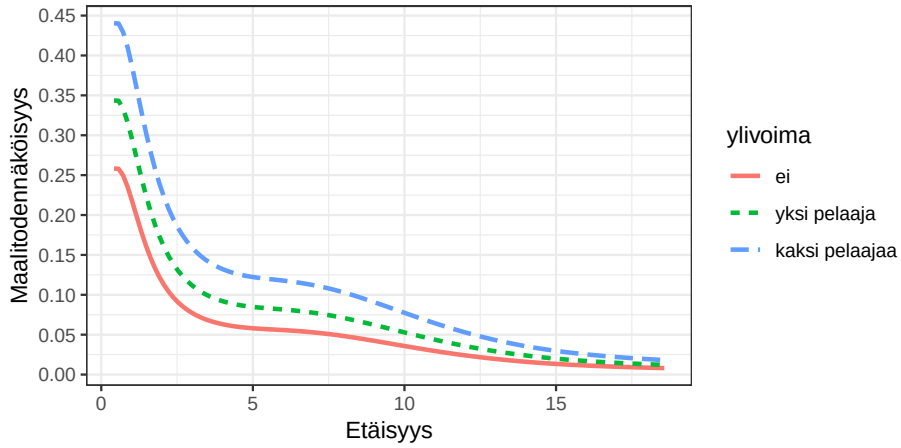
Visualisointi on hieman haastavaa, sillä estimaatit ovat yleisesti ottaen melko pieniä ja lähellä maalia arvot nousevat voimakkaasti. Kuvassa 19 kaikki 0.1:tä suuremmat todennäköisyydet on kuvattu tummanpunaisella värillä. Maalin lähellä tavallisten laukausten maalitodennäköisyydet nousevat noin 0.3:een. Ylivoimatilanteissa ja reboundlaukauksissa sekä kulman muuttuessa paljon maalitodennäköisyydet ovat korkeimmillaan lähes 0.8. Mainituilla lisämuuttujilla ei ole mallissa yhteisvaikutusta kulmaan ja etäisyyteen, joten kuvan 19 todennäköisyyskentän muoto säilyy likimain samana, mutta arvot muuttuvat regressiokertoimien mukaisesti.



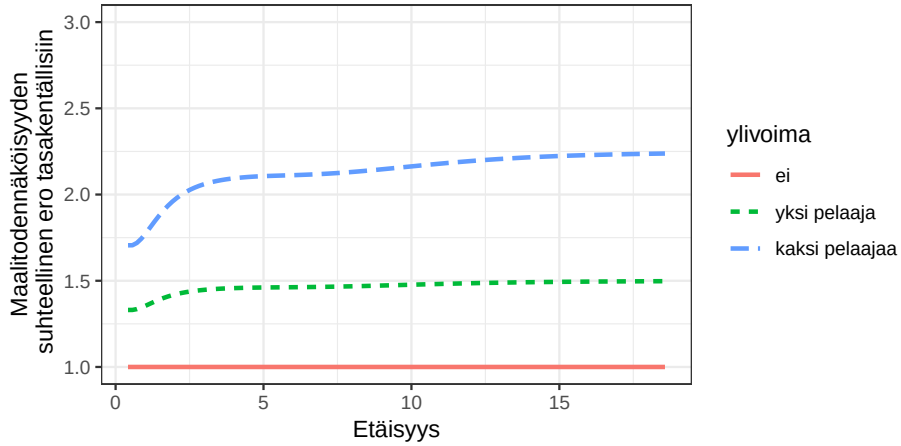
Kuva 19: Lämpökarttakuva mallin lopullisista maalitodennäköisyyksistä ehdolla, että laukausta ei blokata ja se ei mene ohi.

Kuvassa 20 esitetään laukauksen pelkän maalimallin, eli kokonaisrakenteen kolmannen portaan mallin, tuottamia todennäköisyysennusteita eri etäisyyksille ja ylivoimatilanteille. Tilanteessa oletetaan laukauksen meneminen maalivahdille asti. Laukaisupiste on määritelty kentän keskilinjalle ja etäisyys mitataan keskilinjaa pitkin. Etäisyyden lisäksi laukauksen visuaalinen kulma pienenee etäisyyden kasvaessa. Kuvaajasta huomataan, että ylivoimalaukaukset ovat selvästi vaarallisempia pienillä etäisyyksillä. Absoluuttinen ero on huomattava noin kymmeneen metriin saakka. Suhteellinen

ero on tämän jälkeenkin selvä (kuva 21).



Kuva 20: Visualisaatio etäisyyden vaikutuksesta ylivoimatilanteissa laukauksen suuntautuessa maalia kohti.

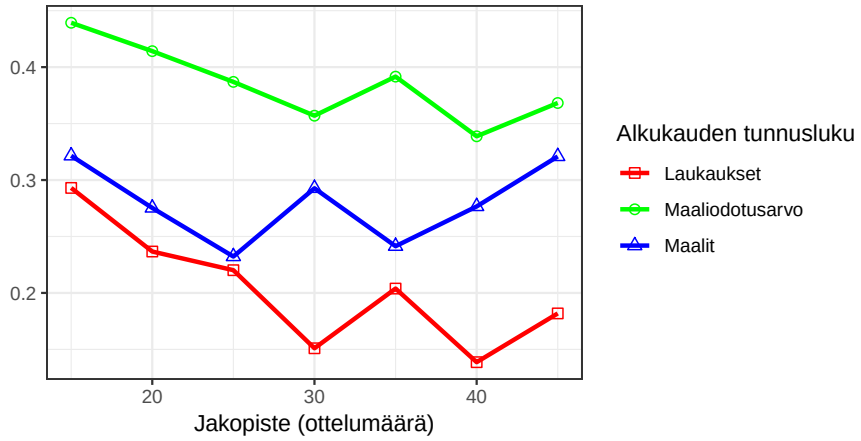


Kuva 21: Kuvan 20 maalitodennäköisyyksien suhteelliset erot tasakentällisiin.

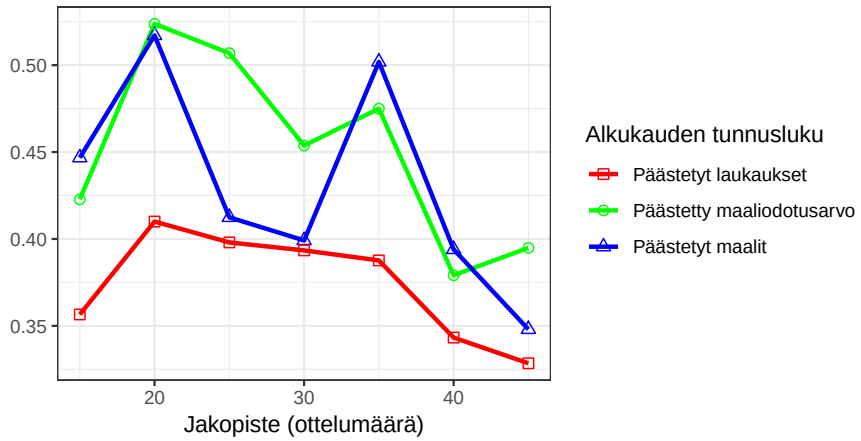
6.4 Maaliodotusarvojen ennustekyky

Ennustekykyä tutkittaessa opetus- ja testiaineisto on yhdistetty kokonaiseksi aineistoksi. Hyökkäyslukujen kuvaajassa 22 maaliodotusarvojen korrelaatio tulevaisuuden maaleihin on lähes jokaisessa jakopisteessä muita tunnuslukuja korkeampi. Huomio kiinnittyy myös siihen, kuinka alkukauden tehdyt maalit korreloivat loppukauden maalien kanssa vahvemmin kuin laukaukset. Puolustuspeliin liittyvissä tunnusluvuissa (kuva 23) maaliodotusarvot korreloivat enimmäkseen paremmin verrattuna laukauksiin perustuviin lukuihin, mutta ero muihin tunnuslukuihin ei ole niin selvä verrataes-

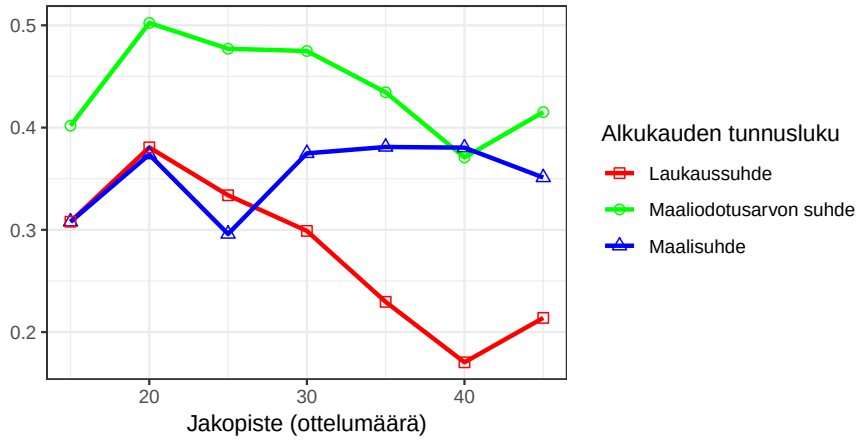
sa hyökkäyspelin tunnuslukuihin. Päästetyt maalit korreloivat likimain yhtä vahvasti tulevaisuuden päästettyjen maalien kanssa 35 ottelun jälkeen. Kuvan 24 suhdelukujen korrelaatiosta huomataan, että maaliodotusarvosuhde korreloi parhaiten tulevaisuuden maalisuhteen. Ennustekyvyn arviointia voitaisiin tehdä myös todennäköisyysmallien avulla. Tuolloin muodostettaisiin aineisto otteluista ja kuhunkin otteluun osallistuvan joukkueen sen hetken tunnusluvuista, ja ottelun lopputulosta selitettäisiin mallissa joukkueiden tunnusluvuilla. Malleja voidaan tämän jälkeen vertailla mm. log-tappion avulla.



Kuva 22: Alkukauden tunnuslukujen korrelaatiot loppukauden tehtyihin maaleihin per ottelu eri jakopisteillä. Kaikki tunnusluvut on laskettu viidellä viittä vastaan pelatulta ajalta.



Kuva 23: Alkukauden tunnuslukujen korrelaatiot loppukauden päästettyihin maaleihin per ottelu eri jakopisteillä. Kaikki tunnusluvut on laskettu viidellä viittä vastaan pelatulta ajalta.



Kuva 24: Alkukauden tunnuslukujen korrelaatiot loppukauden maalisuhteeseen eri jakopisteillä. Kaikki tunnusluvut on laskettu viidellä viittä vastaan pelatulta ajalta.

7 Pohdintaa

7.1 Maaliodotusarvot ottelun tasolla

Yksittäisen laukauksen tasolla puhutaan laukauksen maalitodennäköisyydestä. Ottelun tai joukkueen maalimäärälle voidaan laskea odotusarvo summaamalla maalitodennäköisyyksiä. Maaliodotusarvojen laskennassa on kuitenkin syytä huomioida muutama käytännön tekijä. Rebound-laukaukset vaativat toteutuakseen aina sen, että edeltävä laukaus ei ole mennyt maaliin. On olemassa tilanteita, joissa hyvin pienen ajan sisään sama joukkue laukoo useita perättäisiä rebound-laukauksia. Tuolloin maalitodennäköisyydet ovat hyvin korkeita, jolloin ne saattavat summautua yksittäisen tilanteen osalta yli yhden. Se ei ole käytännössä järkevää, sillä myöhemmät rebound-laukaukset tapahtuvat aina sillä ehdolla, että aikaisemmat eivät ole menneet maaliin. Summattaessa yli ajan on parempi käyttää rebound-korjattuja lukuja. Rebound-korjatut maalitodennäköisyydet $\tilde{p}$ lasketaan ehdollisena todennäköisyytenä:

$$\tilde{p_i} = \begin{cases} p_i\,(1 - p_{i-1}), & \text{jos rebound} \\ p_i, & \text{muuten.} \end{cases} \tag{11}$$

Tällä korjauksella pelitilanteen maaliodotusarvosta saadaan realistisempi arvio, sillä todennäköisyydet eivät summaudu yli yhden. Kyseinen ilmiö kuitenkin yleistyy muihinkin pelitilanteisiin. Esimerkiksi toisen joukkueen pitkät hallintajaksot vastustajan hyökkäysalueella saattavat tuottaa paljon maalipaikkoja. Tällaiselle pitkälle jaksolle

laskettu maaliodotusarvo voi mennä yli yhden, vaikka todellisuudessa hallintajakso etenee aina sillä ehdolla, että edellinen laukaus ei ole tuottanut maalia. Samoin tavallisen kahden minuutin ylivoimapelin aikana joukkue voi tehdä korkeintaan yhden maalin, jonka jälkeen ylivoimatilanne päättyy. Maaliodotusarvoihin on siis mahdollista tehdä yhä pidemmälle vietyjä korjauksia perustuen siihen, että yhdestä kiekonhallinnasta voidaan tehdä korkeintaan yksi maali.

Ottelun tilanne vaikuttaa yleensä pelitapahtumiin monella tapaa. Joukkueen johtaessa ottelua se saattaa usein pelata pienemmällä riskitasolla, jolloin pelaajien tekemät valinnat ja joukkueen taktiikka ovat usein enemmän puolustusorientoituneita verrattuna peliin tasatilanteessa. Vastaavasti tappioasemassa pelaava joukkue saattaa hakea tasoittavaa tai kaventavaa maalia suuremmalla riskitasolla. Tämän takia pelin kuva saattaa usein vääristyä, jos ajatellaan siten, että paremman joukkueen kuuluisi enimmäkseen hallita pelitapahtumia. Mainittu efekti heijastuu myös tilastolukemiin. Kyseessä on tilanne-efekti (*score-effect*), jota voidaan hallita ottelun tilanteen mukaan määräytyvillä korjauskertoimilla. Tällaisia korjauksia kutsutaan tilannekorjauksiksi (*score adjustment*), joiden metodologiaa on selitetty mm. McCurdyn (2014) blogikirjoituksessa. Menetelmän ideana on asettaa laukauksille painokertoimet pelitilanteen mukaan. Kyseinen menetelmä huomioi myös efektin erot koti- ja vieraskentän välillä. Käytännössä tappioasemassa pelaava joukkue on usein edullisessa asemassa johtavan joukkueen passivoituessa, joten tappiolla olevan joukkueen laukauksille asetetaan pienentävä painokerroin. Johtavan joukkueen laukauksia vastaavasti korostetaan. Tällaisen korjausmenetelmän käyttö yleisesti ottaen parantaa lukemien ennustekykyä tehtäessä kappaleessa 5.3 esitettyjä prediktiivisyystarkasteluja.

7.2 Mallien toimivuus ja parannukset

Graafisten tarkasteluiden perusteella mallien kalibraatioon voi olla tyytyväinen. Kuten kalibraatiokuvaajista nähdään, maalitodennäköisyyksien kannalta erityisen laadukkaita maalipaikkoja on vain pieni osa. Kymmenen prosentin kvantiiliryhmien vaihteluvälien pituudet ovat yleensä melko pieniä, kuten esimerkiksi kuvasta 15 nähdään. Mallin tarkkuutta suurilla todennäköisyyksillä voitaisiin tarkastella lähemmin tekemällä kalibraatiokuvaaja pelkästään viimeisestä kvantiiliryhmästä. Tämän työn mallia voisi mahdollisesti parannella yksinkertaistamalla hieman rakennetta. Taulukossa 3 esitettyjen log-tappioiden perusteella yksinkertaistamisesta voisi olla hyötyä, sil-

lä yksinkertaisemmilla rakenteilla tappiopistemäärät ovat hieman parempia. Täytyy kuitenkin muistaa, että erot log-tappioissa ovat hyvin pieniä. Yksi mahdollisuus olisi yhdistää blokatut laukaukset ja ohilaukaukset omaksi ryhmäkseen, jolloin kokonaismallista poistuisi yksi taso. Tuolloin myös vältyttäisiin ongelmalta blokkien ja ohilaukausten määritelmässä, jossa on kappaleen 2.2 mukaisesti selviä eroja paikkakuntien välillä. Toisaalta ohilaukomiseen liittyvät yksilövaikutukset sekoittuvat tässä tapauksessa blokattuihin laukauksiin, jolloin pelaajaefektin tulkinta olisi selvästi erilainen. Käytettävissä olevalla aineistolla ei luultavasti saada kovin suuria parannuksia mallin tarkkuuteen. Isompia muutoksia voidaan mahdollisesti saada sisällyttämällä malliin mm. syöttödataan perustuvia lisämuuttujia. Poikittaissyötöillä ja maalin takaa tulevilla syötöillä on luultavasti positiivinen vaikutus maalitodennäköisyyksiin.

Alun perin ohilaukausten malliin oli tarkoitus ottaa etäisyysmuuttuja ja regressiosplinit mukaan. Kappaleessa 6.3 tehdyn mallin visualisoinnin mukaan tuossa tapauksessa kentän reuna-alueille estimoituu hyvin matalia ohilaukomisen todennäköisyyksiä. Tämä on käytännön kannalta epäloogista. Piirretyn todennäköisyyskentän mukaan keskellä hyökkäysaluetta, paikalla, josta maali näkyy eniten, ohilaukominen olisi paljon todennäköisempää kuin reunoilla. Syynä tähän on luultavasti reunoilta tulevien laukausten vähäinen määrä, jolloin ekstrapolaatio aiheuttaa virheitä. Lisäksi ohilaukomisen todennäköisyys pienenee siirryttäessä lähemmäs siniviivaa. Eräs mahdollinen syy on etäisyyden ja kulman $\tilde{a}$ riippuvuus toisistaan: kauemmas siirryttäessä myös $\tilde{a}$ pienenee. Mallien kalibraatioissa ja muissa tunnusluvuissa ei tapahtunut suurta muutosta splinien ja etäisyyden poistamisen seurauksena. Samalla kuitenkin lämpökarttakuva muuttui käytännön kannalta järkevämmäksi.

Ohilaukausten ja maalien mallit eivät aivan täyttäneet sovituksessa käytettyä konvergenssikriteeriota. Tämä johtuu mahdollisesti siitä, että satunnaisefektien keskihajonnat ovat varsin pieniä. Käytännössä tämä tarkoittaa sitä, että pelaajien väliset efektit maalitodennäköisyyksiin ovat suurimmalta osin vähäisiä. Konvergenssikriteerin täyttyminen on mallin sovituksessa melko pienestä kiinni. Ongelman merkityksellisyyttä voidaan tutkia tekemällä vertailuja eri estimointimenetelmiin. Malli voidaan esimerkiksi sovittaa menetelmällä, joka on rakenteeltaan sama kuin luvussa 4.4 esitetty algoritmi, mutta mallin kiinteät efektit $\boldsymbol{\beta}$ estimoidaan jo vaiheen 1. PIRLS-algoritmissa. Menetelmä ei ole eksakti ja sen käyttö ei ole suositeltua (Bates et al.,

2018), mutta se on huomattavasti nopeampi verrattuna esitettyyn algoritmiin. Lisäksi konvergenssiongelmia ei esiinny. Erot estimoiduissa parametreissa ovat hyvin pieniä menetelmien välillä, jonka perusteella voitaneen olettaa, ettei konvergenssiongelma ole kovin vakava. Samat tulokset voidaan siis saavuttaa käyttämällä ns. nopeaa algoritmia, joka ei ole täysin eksakti, tai korrektimmalla menetelmällä, joka ei aivan täytä konvergenssikriteeriä.

7.3 Tunnuslukujen ennustekyky

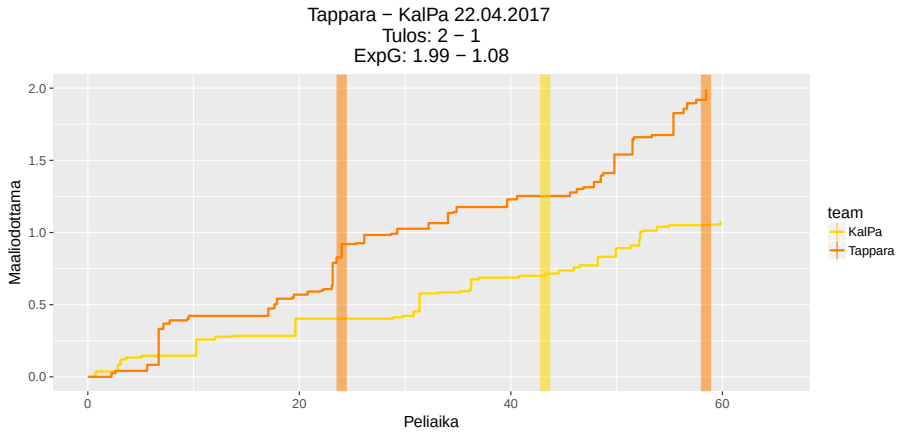
Kuvissa 22–24 esitettyjen korrelaatioiden mukaan tehdyt ja päästetyt maaliodotusarvot sekä maaliodotusarvosuhteet ovat tulevaisuuden menestyksen kannalta parhaat yksittäiset tunnusluvut kuvaamaan joukkueen suoritusten tasoa, sillä niiden korrelaatiot tulevaisuuden maaleihin ovat näistä luvuista korkeimmat. On kuitenkin huomioitava näihin lukuihin liittyvä epävarmuus. Korrelaatioille Fisherin Z-muunnoksen avulla lasketut luottamusvälit ovat hyvin leveitä. Esimerkiksi kuvassa 22 kolmenkymmenen ottelun kohdalla maalien korrelaatio loppukauden maaleihin on noin 0.29 95 %:n luottamusvälillä (0.04, 0.51) ja maaliodotusarvon korrelaatio loppukauden maaleihin on noin 0.36 vastaavalla luottamusvälillä (0.11, 0.56). Tämä asia on sivuutettu useissa tunnuslukujen prediktiivisyyttä käsittelevissä blogikirjoituksissa. Esimerkiksi johdannossa mainitun NHL:n otteludataan perustuvan maalitodennäköisyysmallin (Hockey Graphs, 2015) tuottamat maaliodotusarvot on testattu vastaavaan tapaan tarkastelemalla korrelaatioita tulevaisuuden maaleihin ja maalisuhteisiin, mutta lukujen epävarmuutta ei ole huomioitu. Kyseisessä artikkelissa, ja monissa muissa NHL:n dataan tehdyissä vastaavanlaisissa analyyseissa, on myös havaittu laukausten korreloivan maaleja paremmin tulevaisuuden maaleihin. SM-Liigan aineistossa maalien korrelaatio on kuitenkin korkeampi. Syy voi olla sarjojen välisissä eroissa, mutta ennen suurempia johtopäätöksiä on kuitenkin huomioitava se, että SM-Liigan osalta korrelaatiot on laskettu vain 29 joukkuekaudesta. Sarjojen pelityyleissä on eroja laukaisutiheyden osalta. NHL:ssä laukaisumäärän ottelukohtainen keskiarvo on 112.8 laukausta ja SM-Liigassa 93.9 laukausta. NHL:n keskiarvo on laskettu kauden 2009–2010 alusta kauden 2017–2018 loppuun ja SM-Liigan keskiarvo kauden 2014–2015 alusta kauden 2017–2018 loppuun. Estimaattien 95 %:n luottamusvälit ovat (112.5, 113.1) ja (93.4, 94.5). Ottelukohtainen laukaisumäärä on siis NHL:ssä SM-Liigaa korkeampi ja ero on tilastollisesti merkitsevä.

Tunnuslukujen prediktiivistä arvoa voidaan myös tarkastella ennustemallien avulla. Eräs yksinkertainen tapa rakentaa ennustemalli on koota aineisto kaikkien otteluiden lopputuloksista ja yhdistää kuhunkin otteluun osallistuvien joukkueiden sen hetkiset tunnusluvut. Ennustemallissa lopputulosta selitetään koti- ja vierasjoukkueen tunnusluvuilla. Lopputulos voi olla esimerkiksi kolmiluokkainen (kotivoitto, tasapeli, vierasvoitto), jolloin se määritellään varsinaisen peliajan lopputuloksen mukaan. Vaste voidaan myös määrittää kaksiluokkaiseksi, jolloin kyseessä on koko ottelun lopputulos, jossa tasapelin ratkaiseva jatkoaika ja rangaistuslaukauskilpailu on huomioitu. Tunnuslukujen prediktiivistä arvoa voi tutkia vertailemalla eri selittäjäjoukkoja käyttävien mallien log-tappioita.

7.4 Käytännön hyödyt

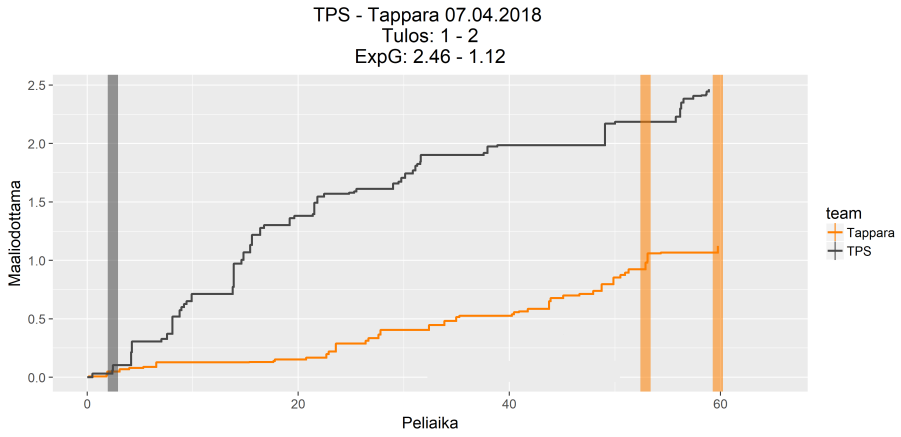
Maalitodennäköisyyksiä ja -odotusarvoja voidaan hyödyntää lajin analysoinnissa monella tapaa ottelun, joukkueen, pelaajien ja pelin yleisessä analysoinnissa.

Yksittäisistä otteluista voidaan tehdä esimerkiksi graafisia tarkasteluja. Kuvissa 25 ja 26 on esitetty, kuinka ottelun maaliodotusarvo on kertynyt joukkueittain maalitodennäköisyyksien summana. Kuvaaja hyppää ylöspäin laukauksen kohdalla ja hypyn suuruus määräytyy maalitodennäköisyydestä. Tällaisesta grafiikasta voidaan havaita, millä tavalla ottelu on edennyt laukausten ja maalipaikkojen laadun kannalta. Kuvioon voidaan lisäksi merkitä ne ajanjaksot, jolloin toinen joukkue on pelannut ylivoimaisena. Eräs olennainen pelin osa-alue on ns. *matchup*-pelaaminen, eli valmennuksen tekemä tiettyjen ketjujen peluuttaminen säännöllisesti toisiaan vastaan. Tämä voidaan toteuttaa pelikellon käydessä, tai kotijoukkueen toimesta pelikatkolla, sillä kotijoukkueella on oikeus viimeiseen vaihtoon pelikatkon aikana. Kotijoukkue voi esimerkiksi asettaa parhaan kentällisen aina vierasjoukkueen heikompia kentällisiä vastaan tai jollain muulla tavalla hakea itselleen suotuisia kentällisasetelmia. Eräs taktiikka on peluuttaa erityisen hyvin puolustavaa kentällistä vastustajan vaarallisinta hyökkäysketjua vastaan. Maaliodotusarvojen avulla voidaan tutkia eri kentällisasetelmien toimivuutta tarkastelemalla esimerkiksi maaliodotusarvojen erotusta tiettyjen kentällisasetelmien aikana. Graafinen tarkastelu on tähän käytännöllisin vaihtoehto. Tämän toteuttaminen SM-Liigan dataan vaatisi vaihtokarttojen käyttämistä, mutta ne eivät ole vapaasti käytettävissä, kuten aiemmin ylivoima-algoritmin yhteydessä todettiin.



Kuva 25: Maaliodotusarvon kertyminen maalitodennäköisyyksien summana joukkueittain Tapparan ja KalPan välisessä finaaliottelussa keväällä 2017. Maaliin johtaneet laukaukset on korostettu pystyviivalla. Tässä tapauksessa toteutunut lopputulos vastaa hyvin maaliodotusarvoja.

Myös pelaajien suorituksia voi analysoida hyödyntämällä maalitodennäköisyyksistä laskettuja maaliodotusarvoja. Sekamalliin satunnaisefektinä sisällytetty pelaajaefekti käytännössä kertoo pelaajan vaikutuksen laukausten maalitodennäköisyyksiin. Ilman satunnaisefektejä voidaan yksinkertaisesti laskea pelaajan odotusarvoinen maalimäärä summaamalla maalitodennäköisyyksiä ja verrata tätä toteutuneeseen maalimäärään. Satunnaisefektin etuna on se, että ennuste huomioi pienet otoskoot. Esimerkiksi pelaaja, joka on laukonut hyvin vähäisen määrän laukauksia, mutta onnistunut niissä verrattain usein, saa todennäköisesti efektikseen populaatiokeskiarvoa lähellä olevan luvun, sillä laukausten määrä on vielä varsin pieni. Tätä voidaan ajatella eräänlaisena regularisointina. Vasta suuremmalla otoskoolla pelaajan efekti poikkeaa enemmän populaatiokeskiarvosta. Maalivahtien satunnaisefektin avulla voidaan maalivahteja analysoida vastaavaan tapaan: hyville maalivahdeille estimoituu negatiivinen efekti maalitodennäköisyyteen. Ilman satunnaisefektiä voidaan laskea odotettu päästettyjen maalien määrä maalitodennäköisyyksien summana ja verrata sitä toteutuneeseen. Myös torjuntaprosentille saadaan odotusarvo, jota voidaan verrata toteutuneeseen. Pelaajia voidaan myös profiloida tarkastelemalla esimerkiksi maali- ja laukaisumäärien ohella laukausten keskimääräisiä maalitodennäköisyyksiä. Pelaaja, jonka laukausten keskimääräinen maalitodennäköisyys on hyvin korkea, laukoo, ja olennaisesti pääsee laukomaan, laadukkailta paikoilta. Joku pelaaja voi laukoa määrällisesti paljon, mutta laukausten maalitodennäköisyydet saattavat olla pieniä. Pelaaja, joka saavuttaa suuria maalimääriä huolimatta heikoista maalitodennäköisyyksistä, on



Kuva 26: Maaliodotusarvon kertyminen maalitodennäköisyyksien summana joukkueittain TPS:n ja Tapparan välisessä pudotuspeliottelussa keväällä 2018. Maaliin johtaneet laukaukset on korostettu pystyviivalla. TPS on ottelun alkupuolella kehitellyt selvästi enemmän maaliodotusarvoa, mutta maalit ovat jääneet puuttumaan. Vastaavasti Tappara on saanut tehtyä kaksi maalia kolmanteen erään melko vähäisellä maaliodotusarvolla.

luultavasti erittäin taitava laukoja. Näissä analyyseissa täytyy kuitenkin muistaa, että viimeistelyprosentit vaihtelevat hyvin paljon, jolloin johtopäätösten tekeminen pienistä otoskoista ei ole suositeltavaa. Tässä on myös yksi mahdollinen sovelluskohde klusterointianalyysille, jonka avulla voidaan löytää esimerkiksi erilaisia laukojatyyppejä. Klusterointia on sovellettu pelaajatyyppien erotteluun aiemmin perinteisten tilastolukemien avulla (Vincent & Byron, 2009).

Syöttödatan kerääminen tuo todennäköisesti paljon uutta jääkiekon data-analyysiin. Maalitodennäköisyysmallien parantamisen lisäksi saadaan käyttöön lisää hyödyllisiä tunnuslukuja pelaajien analysointiin. Edellä mainittu pelaaja-analyysi suosii pelaajia, jotka laukovat paljon. Esimerkiksi taitavat peliä rakentavat pelaajat ja syöttäjät eivät välttämättä loista laukomiseen liittyvissä tunnusluvuissa. Jos jokaiselle laukaukselle kirjattaisiin syöttäjä, voitaisiin pelaajille laskea maalitodennäköisyyksien avulla myös syöttöpisteiden odotusarvo. Tämän avulla saataisiin tunnistettua pelaajia, jotka osaavat rakentaa laadukkaita maalipaikkoja muille kenttäpelaajille. Lisäksi syöttödatan avulla laukaukseen johtaneelle syöttöketjulle ja siihen osallistuneille pelaajille saadaan laskettua maalitodennäköisyysarvo, joka huomioi laajemmin yksittäisen maalipaikan rakentamiseen osallistuneiden pelaajien osuuksia.

Viitteet

Akaike Hirotugu. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Barnes Tim. Zone Time, VHockey 2008 (blogi). URL `https://web.archive.org/web/20140713155423/http://vhockey.blogspot.com/2008/08/zone-time.html`. luettu: 15.1.2019.

Bates Douglas. Computational Methods for Mixed Models. Technical report, University of Wisconsin, 2011. URL `https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf`.

Bates Douglas, Mächler Martin, Bolker Ben & Walker Steve. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Bates Douglas, Mächler Martin, Bolker Ben, Walker Steven, Bjoesen Christensen Rune Haubo, Singmann Henrik, Dai Bin, Scheipl Fabian, Grothendieck Gabor, Green Peter & Fox John. Linear Mixed-Effects Models using 'Eigen' and S4, 2018. URL `https://cran.r-project.org/web/packages/lme4/lme4.pdf`. luettu: 7.6.2018.

Becker Dan. Kaggle: What is log loss? URL `https://www.kaggle.com/dansbecker/what-is-log-loss`.

Buja Andreas, Stuetzle Werner & Shen Yi. Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications. 11 2005. URL `http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf`.

de Boor Carl. *A Practical Guide to Splines*, volume 27. Springer-Verlag New York, 1978.

Hastie Trevor, Tibshirani Robert & Friedman Jerome. *The Elements of Statistical Learning*, volume 2. Springer, 2001.

Hockey Graphs. Expected Goals Are a Better Predictor of Future Scoring Than Corsi, Goals, 2015. URL `https://hockey-graphs.com/2015/10/01/expected-goals-are-a-better-predictor-of-future-scoring-than-corsi-goals`. luettu: 11.2.2018.

Hosmer David W, Lemeshow Stanley & Sturdivant Rodney X. *Applied Logistic Regression*, volume 398. John Wiley & Sons, 2013.

Macdonald Brian. *An Expected Goals Model for Evaluating NHL Teams and Players*. Proceedings of the 2012 MIT Sloan Sports Analytics Conference, 2012. URL http://www.hockeyanalytics.com/Research_files/NHL-Expected-Goals-Brian-Macdonald.pdf.

McCurdy Micah Blake. Better Way to Compute Score-Adjusted Fenwick, 2014. URL http://hockeyviz.com/txt/senstats. luettu: 12.7.2018.

Pinheiro José C & Bates Douglas M. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4(1):12–35, 1995.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

SM-Liigan verkkosivu. https://liiga.fi/.

Vincent Claude B & Byron Eastman. Defining the Style of Play in the NHL: An Application of Cluster Analysis. *Journal of Quantitave Analysis in Sports*, 5(10), 2009. doi: https://doi.org/10.2202/1559-0410.1133.

Wegman Edward J & Wright Ian W. Splines in Statistics. *Journal of the American Statistical Association*, 78(382):351–365, 1983. URL http://www.jstor.org/stable/2288640.

Wickham Hadley. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016a. ISBN 978-3-319-24277-4. URL http://ggplot2.org.

Wickham Hadley. *rvest: Easily Harvest (Scrape) Web Pages*, 2016b. URL https://CRAN.R-project.org/package=rvest. R package version 0.3.2.