



JUNALIIKENTEEN VIIVÄSTYSTEN STOKASTINEN MALLINTAMINEN

Lappeenrannan–Lahden teknillinen yliopisto LUT

Laskennallisen tekniikan koulutusohjelma, Kandidaatintyö

2025

Amanda Seppinen

Tarkastaja: Tomás Soto

Tiivistelmä

Lappeenrannan–Lahden teknillinen yliopisto LUT

School of Engineering Science

Laskennallinen tekniikka

Amanda Seppinen

Junaliikenteen viivästysten stokastinen mallintaminen

Kandidaatintyö

2025

24 sivua, 2 kuvaa, 1 taulukko ja 1 liite

Tarkastaja: Tomás Soto

Avainsanat: junaliikenteen aikataulut, stokastinen mallinnus, parametriestimointi, satunnaismuuttujat

Junaliikenteen täsmällisyys on keskeinen tekijä niin matkustajien tyytyväisyyden kuin liikennejärjestelmän toimivuuden kannalta. Tässä kandidaatintyössä tarkastellaan junaliikenteen viivästysten mallintamista stokastisten menetelmien avulla. Työn tavoitteena on arvioida, miten satunnaiset viivästykset käyttäytyvät yksittäisen aseman näkökulmasta ja soveltaa parametriestimointia näiden viivästysten mallintamiseen.

Aineistona käytettiin Fintrafficin Digitraffic-rajapinnan kautta haettua toteumatietoa Helsingin päärautatieaseman junista valituilta yksittäisiltä päiviltä vuoden 2025 aikana. Aineistosta laskettiin viivästysten jakaumia ja simuloitiin junien saapumisaikoja satunnaismuuttujina. Simulaatioilla pyrittiin arvioimaan viiveiden todennäköisyysjakamaa sekä ymmärtämään viivästysten kasaantumista yksittäisen päivän aikana.

Tulokset osoittivat, että pienet viivästykset ovat yleisiä, mutta suuria viivästyksiä esiintyy satunnaisesti ja ne voivat aiheuttaa merkittäviä poikkeamia aikatauluista. Mallinnus antoi realistisen kuvan viivästysten vaihtelusta ja soveltuu jatkossa käytettäväksi esimerkiksi riskienhallinnan ja kapasiteetinsuunnittelun tukena.

Abstract

Lappeenranta-Lahti University of Technology LUT
LUT School of Engineering Science
Degree Programme in Computational Engineering

Amanda Seppinen

Stochastic modeling of train delays

Kandidaatintyö

2025

24 pages, 2 figures, 1 table and 1 appendix

Examiner: Tomás Soto

Keywords: train timetables, stochastic modelling, parametric estimation, random variables

Punctuality in rail transport is a key factor affecting both passenger satisfaction and the overall efficiency of the transportation system. This bachelor's thesis explores the modeling of train delays using stochastic methods. The aim is to analyze how random delays behave from the perspective of a single station and to apply parameter estimation to model these delays.

The dataset consists of realized train traffic data retrieved via Fintraffic's Digitraffic API for selected days in 2025 at Helsinki Central Station. Delay distributions were calculated from the data, and train arrival times were simulated as random variables. The simulations aimed to estimate the probability distribution of delays and to understand the accumulation of delays over the course of a day.

The results indicate that small delays are common, while larger delays occur sporadically and can lead to significant deviations from scheduled times. The modeling provided a realistic view of the variability in delays and can be applied in future for supporting risk management and capacity planning.

Symboli- ja lyhenneluettelo

Kreikkalaiset aakkoset

α	(alfa)	1/K
β	(beta)	
Γ	(Gamma)	
θ	(theta)	
ξ	(xi)	

Lyhenteet

MLE suurimman uskottavuuden estimointi (maximum likelihood estimation)

Sisällysluettelo

Tiivistelmä	
Abstract	
Symboli- ja lyhenneluettelo	
Sisällysluettelo	5
Kuva- ja taulukkoluetelo	7
1 Johdanto	8
1.1 Junaliikenteen merkitys Suomessa	8
1.2 Työn tavoitteet	8
1.3 Työn rakenne	9
2 Viivästykset ja aiempi tutkimus	10
2.1 Viivästysten taustatekijät	10
2.2 Aiemmat tutkimukset	10
3 Stokastinen mallinnus	12
3.1 Stokastiset menetelmät	12
3.2 Parametristimointi	12
3.3 Jakauman valinta	13
4 Mallinnus	15
4.1 Data	15
4.2 Jakaumien sovittaminen dataan	15
4.3 Yksinkertainen matemaattinen malli	16
4.3.1 Gammajakauma	16
4.3.2 Vinonormaalijakauma	17
4.3.3 Mallinnusprosessi	17
4.4 Toteutus ja simulointi	17
4.5 Validointi ja tarkastelu	18
5 Tulokset	21
6 Johtopäätökset	22

Lähdeluettelo

23

Liitteet

Liite A Algoritmit Julia-ohjelmointikielellä

Kuvaluettelo

- 1 Vinonormaalijakaumaan sovitettut datapisteet
- 2 Gammajakaumaan sovitettut datapisteet

Taulukkoluetelo

- 1 Simuloitujen viiveiden keskiarvo ja hajonta

1 Johdanto

1.1 Junaliikenteen merkitys Suomessa

Junaliikenne on keskeinen osa Suomen liikennejärjestelmää, erityisesti matkustaja- ja tavara-liikenteessä. Se tarjoaa ympäristöystävällisen vaihtoehdon kaupunkien väliseen matkustamiseen ja on tärkeä osa logistiikkaketjua pitkien etäisyyksien tavarakuljetuksissa. Vuonna 2023 rautateiden henkilöliikenteessä tehtiin yli 82 000 matkaa, ja matkamäärät ovat kasvaneet koronapandemian jälkeen (Väylävirasto, 2023a). Suomen rataverkko on laaja, 5915 kilometriä pitkä, ja sen sujuvuus on tärkeää yhteiskunnan ja talouden kannalta (Väylävirasto, 2023b). Viivästykset ovat yksi suurimmista haasteista junaliikenteessä, sillä niillä on vaikutusta talouteen, matkustajakokemukseen ja ympäristöön.

1.2 Työn tavoitteet

Tämän kandidaatintyön tavoitteena on analysoida junaliikenteen viivästyksiä Suomessa ja mallintaa niitä stokastisin menetelmin hyödyntäen parametriesitystä. Työssä lähdetään tutkimaan miten hyvin stokastinen mallintaminen kuvaa Helsingin päärautatieasemalla junien havaittuja saapumisviiveitä. Mallinnuksessa hyödynnetään kahta erilaista todennäköisyysjakaumaa – gamma- ja vinonormaalijakaumaa – jotka soveltuvat kuvaamaan viivästysten satunnaisvaihtelua. Näiden jakaumien avulla muodostetaan kaksi yksinkertaista matemaattista mallia, joiden pohjalta voidaan simuloida junien saapumisaikoja ja arvioida viivästysten todennäköistä esiintymistä, sekä niiden mahdollisia kasaantumisasiäviä päivän mittaan. Työssä pyritään arvioimaan kumpi käytetyistä jakaumista soveltuu mallintamaan käytettyä aineistoa paremmin.

Työn tarkastelun kohteena on Helsingin päärautatieasema, ja analyysi perustuu Fintrafficin tarjoamaan toteumatietoon satunnaisesti valikoiduilta päiviltä vuoden 2025 alkupuolelta. Viivästyksiksi määritellään kaikki poikkeamat aikataulun mukaisista saapumisajoista riippumatta niiden suuruudesta. Työssä painotetaan simulaatiomenetelmien soveltamista ja niiden avulla saadun mallin tulkintaa.

1.3 Työn rakenne

Työ koostuu kolmesta pääluvusta, johdannosta, saaduista tuloksista ja johtopäätöksistä. Luvussa 2 käsitellään viivästysten taustatekijöitä ja aiempia tutkimuksia. Luvussa 3 esitellään stokastisten menetelmien teoria ja niiden soveltaminen junaliikenteen viivästysten analysoinnissa. Luvussa 4 rakennetaan matemaattinen malli ja tarkastellaan sen käytännön toteutusta.

2 Viivästykset ja aiempi tutkimus

2.1 Viivästysten taustatekijät

Junaliikenteen viivästyksiä voivat aiheuttaa useat eri tekijät, jotka voidaan jakaa karkeasti infrastruktuurista, operatiivisista syistä ja ulkoisista tekijöistä johtuviin viiveisiin (Goverde, 2010). Infrastruktuuripohjaiset viivästykset johtuvat esimerkiksi ratatöistä, kapasiteetti- rajoitteista tai teknisistä vioista. Suomessa erityisesti harva rataverkko ja pitkät yksiraiteiset osuudet lisäävät häiriöiden vaikutusherkkyyttä (Väylävirasto, 2022).

Operatiiviset viivästykset liittyvät junakaluston teknisiin ongelmiin, henkilöstön saatavuuteen tai liikenteenohjauksen päätöksiin. Myös liikenteen tiheys voi aiheuttaa ruuhkautumista ja puskuroituvia viiveitä erityisesti pääkaupunkiseudun lähiliikenteessä. Ulkoiset tekijät, kuten sääolosuhteet (esimerkiksi runsas lumi, myrskyt) ja mahdolliset kolari- tai onnettomuus- tilanteet, voivat johtaa äkillisiin ja merkittäviin viivästyksiin.

Viivästysten analysointi ja mallintaminen edellyttää siten monipuolista ymmärrystä eri viiveiden taustamekanismeista. Usein viiveitä tarkastellaan jakaumina, mutta ne poikkeavat normaalijakaumasta: suurin osa junista saapuu ajallaan tai lähes ajallaan, mutta pieni osa kärsii suurista poikkeamista, mikä tuottaa vinon ja pitkäpyrstöisen jakauman (Huisman, Boucherie & Dijk, 2004).

2.2 Aiemmat tutkimukset

Junaliikenteen täsmällisyyttä ja viivästyksiä on tutkittu laajasti eri näkökulmista, kuten operatiivisen simuloinnin, kapasiteetinhallinnan ja asiakastytyväisyyden kannalta. Suomessa ja muualla Euroopassa tutkimus on keskittynyt erityisesti junaliikenteen luotettavuuteen, viiveiden syntymekanismeihin sekä mahdollisuuksiin ennakoida tai lieventää myöhästymisiä.

Tilastolliset ja stokastiset mallit ovat muodostuneet tärkeäksi työkaluksi viivästysanalyysissä. Esimerkiksi Goverde (2005) esitteli tilastollisen menetelmän, jossa junien saapumisajat mallinnettiin satunnaismuuttujina ja ennusteet viivästymisestä johdettiin jatkuvasta jakaumasta (Goverde, 2005). Samankaltaisia lähestymistapoja on sovellettu myös Monte Carlo –menetelmien avulla, erityisesti tilanteissa, joissa junaliikenteen tilaa halutaan arvioida

usean toistettavan simulaation perusteella (Fischetti, Grosso & Toth, 2009).

Tutkimuksissa on havaittu, että viiveiden jakauma ei usein seuraa normaalijakaumaa, vaan on epäsymmetrinen, raskashäntäinen tai muuten vinoutunut. Tämä havainto on johtanut siihen, että gammajakauma, lognormaalijakauma ja vinonormaalijakauma ovat nousseet suosituiksi vaihtoehtoiksi viiveiden tutkinnassa (Carey & Kwienciński, 1995; Kecman & Goverde, 2015). Näillä jakaumilla on kyky kuvata viivästysten stokastista luonnetta realistisemmin erityisesti tilanteissa, joissa poikkeuksellisen suuria viiveitä esiintyy harvakseltaan, mutta niiden vaikutus on merkittävä.

Lisäksi viivästysennustuksissa on sovellettu koneoppimismenetelmiä, mutta ne vaativat usein huomattavasti laajempaa tietomäärää ja tulkitsevuus voi jäädä tilastollisia malleja heikommaksi (Cacchiani, Caprara & Toth, 2014). Tämän työn rajauksen ja käytössä olevien resursien vuoksi painopiste on kuitenkin klassisissa stokastisissa malleissa, joissa teoreettinen rakenne tukee mallin tulkintaa ja parametrien estimointi on mahdollista suoraan havaintoaineistosta.

Aiempi kirjallisuus osoittaa, että viivästysten stokastinen mallintaminen on sekä perusteltua että käytännössä sovellettavissa. Työssä käytettävät mallit ovat siten osa jo vakiintunutta tutkimuslinjaa, mutta niiden soveltaminen suomalaisen junaliikenteen aineistoon on ollut rajallista, mikä perustelee tutkimuksen relevanssin.

3 Stokastinen mallinnus

3.1 Stokastiset menetelmät

Stokastiset menetelmät ovat tilastollisia työkaluja, joilla mallinnetaan järjestelmiä, joissa satunnaisuus ja epävarmuus ovat olennainen osa ilmiötä. Junaliikenteessä viivästyksiin vaikuttaa monia satunnaisia tekijöitä, kuten sääolosuhteet, tekniset viat, matkustajamäärät ja muiden junien vaikutukset. Näiden yhteisvaikutus voidaan kuvata stokastisena prosessina, jossa esimerkiksi saapumisaika on satunnaismuuttuja (Ross, 2014; Carey & Kwienciński, 1995; Goverde, 2005).

Stokastinen mallintaminen perustuu sopivan todennäköisyysjakauman valintaan ilmiön mukaan. Junaviiveitä voidaan kuvata esimerkiksi gamma-, lognormaali- tai eksponentiaalijakaumilla, riippuen datan muodosta ja vinoudesta (Yuan & Hansen, 2007). Tärkeää on jakauman parametrien estimointi historiallisesta aineistosta, jolloin mallista saadaan realistinen ja sovellettavissa oleva.

Tässä työssä hyödynnetään kahta todennäköisyysjakaumaa: gammajakaumaa ja vinonormaalijakaumaa. Gammajakauma $\text{Gamma}(\alpha, \beta)$ soveltuu positiivisille muuttujille ja kuvaa viiveiden vaihtelua muotoparametrin α ja skaalausparametrin β avulla. Vinonormaalijakauma on puolestaan normaalijakauman laajennus, jossa lisäparametri ohjaa vinoutta (Azzalini, 1985). Näiden jakaumien avulla voidaan analysoida viivästysten jakautumista, todennäköisyyksiä ja herkkyksiä, mikä tukee junaliikenteen suunnittelua ja päätöksentekoa (Goverde & Kecman, 2015).

3.2 Parametriestimointi

Parametriestimointi on olennainen osa stokastista mallinnusta, sillä todennäköisyysjakauman muoto määräytyy sen parametrien perusteella. Jotta malli kuvaisi ilmiötä realistisesti, on parametrit estimoitava tarkasti havaintoaineiston perusteella. Tyypillisiä menetelmiä ovat suurimman uskottavuuden estimointi (maximum likelihood estimation, MLE) ja momenttiestimointi (method of moments), joista MLE on yleisimmin käytetty sen tehokkuuden vuoksi (Casella & Berger, 2002).

Tässä työssä käytetään MLE-menetelmää gammajakauman parametrien estimoimiseksi, hyödyntäen Julia-kielen (Julia Computing, 2023) `fit`-funktiota. Vinonormaalijakauman osalta parametrien estimointi toteutetaan numeerisesti minimoimalla negatiivinen log-todennäköisyysfunktio optimointimenetelmän avulla käyttäen Julia-kielen kirjastoa `Optim.jl` (JuliaNLSolvers Developers, 2023). Näin saadaan arviot jakaumien parametreille, jotka parhaiten selittivät havaittua viivedataa.

Estimoinnin tuloksia hyödynnetään simuloitaessa viiveiden jakaumaa ja verrattaessa simuloitua aineistoa havaittuun. Tämä mahdollistaa mallin uskottavuuden arvioinnin ja viivästysten todennäköisyysjakauman tarkastelun, mikä on tärkeää liikennejärjestelmän analysoinnissa ja kehittämisessä.

3.3 Jakauman valinta

Jakauman valinta on keskeinen osa stokastista mallinnusta, sillä väärä jakaumaoletus voi johtaa harhaanjohtaviin johtopäätöksiin simuloinnin ja mallin validoinnin osalta. Tässä työssä tarkastellaan kolmea erilaista todennäköisyysjakaumaa: normaalijakaumaa, gammajakaumaa ja vinonormaalijakaumaa (Skew Normal Distribution). Valintaa ohjaa sekä teoreettinen sopivuus että empiirinen soveltuvuus havaittuun junien viivästysdataan.

Normaalijakauma on yksi yleisimmin käytetyistä jakaumista stokastisessa mallinnuksessa sen matemaattisen yksinkertaisuuden ja laajan sovellettavuuden vuoksi (Law & Kelton, 2000). Se soveltuu hyvin tilanteisiin, joissa havaintojen jakauma on symmetrinen keskiarvon ympärillä. Junaliikenteen viivästykset eivät kuitenkaan tyypillisesti ole symmetrisiä, vaan ne painottuvat usein pieniin positiivisiin arvoihin ja voivat satunnaisesti saada myös huomattavan suuria arvoja. Lisäksi viiveet eivät voi olla negatiivisia, mikä tekee normaalijakaumasta huonosti soveltuvan mallin tällaiseen ilmiöön.

Gammajakauma taas soveltuu erityisen hyvin positiivisesti vinoutuneisiin, ei-negatiivisiin ilmiöihin, kuten odotus- tai palveluaikoihin (Johnson, Kotz & Balakrishnan, 1994). Koska junaviivästykset ovat luonteeltaan ei-negatiivisia ja useimmiten pieniä, mutta sisältävät myös harvinaisia suuria viiveitä, gammajakauma tarjoaa luontevan vaihtoehdon niiden stokastiseen mallintamiseen.

Kolmas tarkasteltu vaihtoehto on vinonormaalijakauma, joka laajentaa normaalijakauman muotoa sallimalla vinoutumisen (skewness) säätelyn. Se on hyödyllinen, mikäli havaittu data muistuttaa normaalijakaumaa, mutta osoittaa systemaattista epäsymmetriaa (Azzalini, 1985). Tämä tekee siitä joustavan mallin erityisesti tilanteissa, joissa viiveitä esiintyy sekä negatii-

visella että positiivisella puolella, mutta jakauma on selvästi vino.

4 Mallinnus

4.1 Data

Tässä työssä käytetty aineisto perustuu Fintrafficin tarjoamaan Digitraffic –rajapintaan (Finttraffic, 2025), josta haettiin junaliikenteen toteutumätietoa yksittäisiltä päiviltä vuoden 2025 puolelta. Käytetyt päivät valittiin satunnaisesti ja niiden tarkoituksena oli tarjota esimerkkejä junaliikenteen viivästysten vaihtelusta normaalissa liikenteessä. Päiviksi valikoituivat perjantai 28.2.2025, maanantai 10.3.2025 ja maanantai 31.3.2025. Päivämäärien data saatiin lisäämällä rajapinnan URL osoitteen perään päivämäärä muodossa YYYY-MM-DD. Lisäksi tutkiessa on poimittu erityisesti Helsingin päärautatieasemaa (tunniste "HKI") koskevat junat. Tällä rajauksella pyrittiin keskittymään vilkkaaseen liikenteen solmukohtaan, jossa viivästysten dynamiikka on erityisen mielenkiintoista ja olosuhteet haastavat täsmällisyyttä.

Digitraffic -rajapinnasta saatava data sisältää muun muassa junan numeron ja tyyphin (esimerkiksi InterCity, Pendolino, lähijuna), junien aikataulunmukaiset ja toteutuneet saapumisajat Helsingin päärautatieasemalle, junan poikkeama aikataulusta minuutteina, sekä mahdolliset junien peruutukset. Poikkeustilanteita ei erikseen suodatettu pois (lakot tai suuremmat häiriöt).

Rajallisten resurssien vuoksi aineisto rajattiin pieneen määrään päiviä ja keskityttiin Helsingin päärautatieasemaan. Tämän myötä mallinnus ja simulaatio kohdistettiin kuvaamaan nimenomaan tällaisen suuren liikennekeskuksen arkipäiväistä satunnaisuutta, ei koko maanlaajuista junaliikennettä.

4.2 Jakaumien sovittaminen dataan

Havaittujen viiveiden perusteella jakaumien sovittaminen suoritetaan kahdella tavalla: parametrisella sovituksella gammajakaumaan ja numeerisella optimoinnilla vinonormaalijakaumaan. Gammajakauman sovitus toteutetaan käyttämällä `Distributions.jl`-kirjaston `fit`-funktioita, joka arvioi jakauman parametrit (muoto- ja skaala-arvot) maksimoimalla aineiston uskottavuuden (MLE, maximum likelihood estimation). Tämä menetelmä on tehokas ja yleisesti käytetty jatkuvien jakaumien sovituksessa (Casella & Berger, 2002).

Vinonormaalijakauman sovitus puolestaan toteutetaan määrittelemällä negatiivisen log -todennäköisyysfunktion `nll` ja minimoimalla se `Optim.jl`-kirjaston avulla. Tämä lähestymistapa mahdollistaa myös vinousparametrin (α) estimoinnin, mikä tekee siitä soveltuvan silloin, kun jakauman muoto poikkeaa symmetrisyydestä. Sovituksen alkuarvauksina käytetään havaittujen viiveiden keskiarvoa ja hajontaa sekä vinoudelle arvoa nolla.

Tulosten perusteella molemmat jakaumat tuottavat realistisia malleja, mutta niiden sovituksen laatua arvioidaan vertaamalla havaittujen ja simuloitujen viiveiden jakaumien visuaalista vastaavuutta histogrammien avulla. Tämän lisäksi voidaan jatkossa hyödyntää tilastollisia goodness-of-fit -testejä (esimerkiksi Kolmogorov-Smirnovin testiä) sovituksen kvantitatiiviseksi arvioimiseksi.

4.3 Yksinkertainen matemaattinen malli

Tässä työssä viivästysten stokastista luonnetta lähestytään yksinkertaistetun todennäköisyysmallin avulla, jossa havaittua dataa käytetään sovittamaan sopiva tilastollinen jakauma. Malli rakentuu olettamuksesta, että viivästykset noudattavat tiettyä jatkuvaa todennäköisyysjakamaa.

Mallin keskiössä on satunnaismuuttuja X , joka kuvaa junan viivettä (minuuteissa). Oletetaan, että $X \sim F(\theta)$, missä F on joko gammajakauma tai vinonormaalijakauma ja θ on jakauman parametrivektori.

4.3.1 Gammajakauma

Gammajakauma on jatkuva jakauma, jonka tiheysfunktio on muotoa:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, \quad x > 0$$

Missä $\alpha > 0$ on muotoparametri ja $\beta > 0$ skaala. Gammajakauma soveltuu erityisesti sellaisiin viiveisiin, jotka ovat aina positiivisia ja voivat sisältää pitkähäntäisiä jakautumia.

4.3.2 Vinonormaalijakauma

Vinonormaalijakauma (Skew Normal Distribution) on laajennus normaalijakaumasta ja sen tiheysfunktio on:

$$f(x; \xi, \omega, \alpha) = \frac{2}{\omega} \phi\left(\frac{x - \xi}{\omega}\right) \Phi\left(\alpha \cdot \frac{x - \xi}{\omega}\right),$$

Missä ϕ ja Φ ovat normaalijakauman tiheys- ja kertymäfunktio, $\xi \in R$ on sijaintiparametri, $\omega > 0$ hajonta, ja $\alpha \in R$ vinousparametri. Tämä jakauma mahdollistaa epäsymmetrisen mallinnuksen ja soveltuu tilanteisiin, joissa viivejakauma ei ole symmetrinen.

4.3.3 Mallinnusprosessi

Mallinnusprosessi lähtee käyntiin datan esikäsittelyllä. Tässä vaiheessa havaituista junaviiveistä poistetaan puuttuvat ja virheelliset arvot. Gammajakaumaa varten suodatetaan vain positiiviset viiveet.

Jakaumaa sovittaessa valitulle jakaumalle sovitetaan parametrit kerätystä datasta. Gammajakauman parametrit voidaan estimoida esimerkiksi momentti- tai maksimiestimaattorilla. Vinonormaalijakauman parametrit sovitetaan maksimoimalla log-todennäköisyys. Mallinnuksen kulku vaiheittain on esitetty liitteenä olevassa algoritmossa (Liite 1).

Tällainen yksinkertaistettu malli ei pyri kuvaamaan viiveiden kaikkia syitä tai dynamiikkaa, vaan tarjoaa tilastollisen välineen arvioida viiveiden jakaumaa ja riskiä. Mallin käyttökelpoisuus riippuu sen soveltuvuudesta havaittuun aineistoon – esimerkiksi jakauman vinous, huipukkuus ja hännät voivat ohjata jakauman valintaa (Law & Kelton, 2007).

4.4 Toteutus ja simulointi

Simulointi toteutetaan parametriestimoinnilla, jossa mallista generoidaan suuri määrä (10 000) satunnaisviivettä, joita verrataan havaittuun aineistoon. Tämä mahdollistaa stokastisen mallin tuottaman viivejakauman arvioinnin ja sen vertailun havaittuihin viiveisiin.

Simulaatio toteutetaan Julia-ohjelmointikielellä, erityisesti hyödyntäen `Distributions.jl`, `Optim.jl`, `DataFrames.jl` ja `Plots.jl` -kirjastoja. Simulaation ydinprosessi lähtee liikkeelle valitsemalla sopiva jakauma ja sovittamalla parametrit jakaumaan. (ks. Luku 4.2). Seuraavaksi sovitetusta jakaumasta otetaan satunnaisotanta ja jakaumat visualisoidaan his-

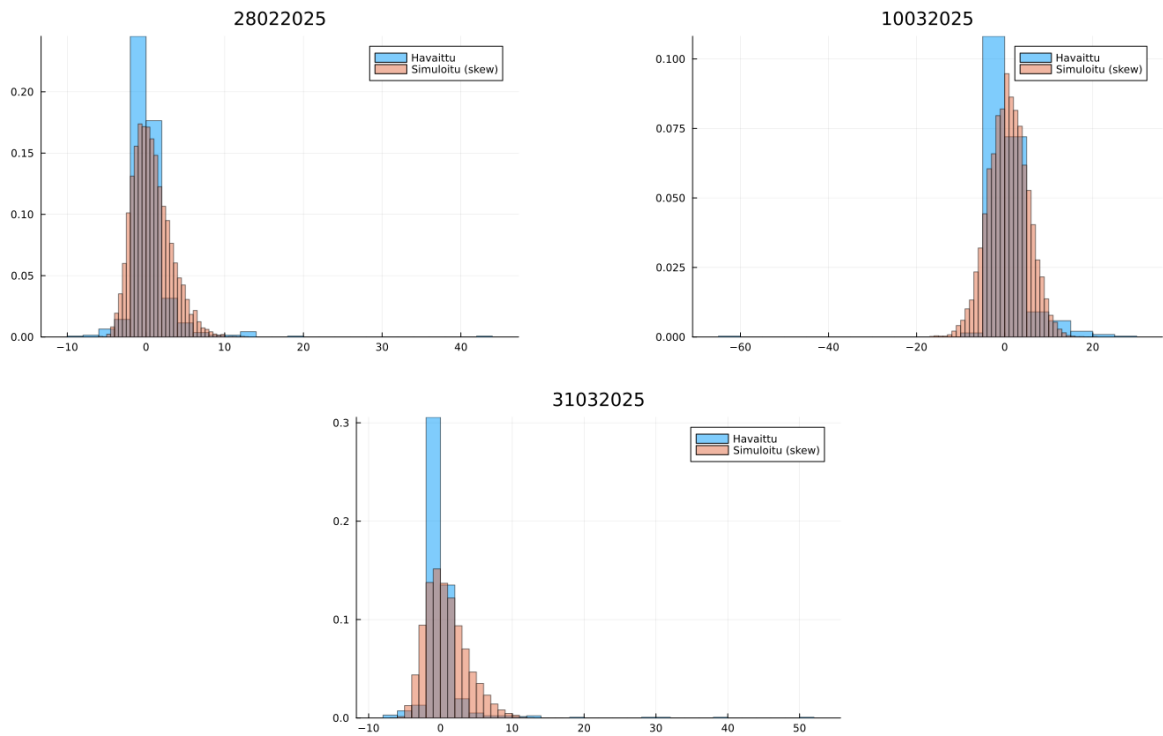
togrammin avulla. Tämän jälkeen simuloituja viiveitä tarkastellaan tilastollisesti (keskiarvo ja hajonta).

Tämä vaihe toimii keskeisenä siltana empiirisen datan ja matemaattisen mallin välillä. Visualisoinnit mahdollistavat intuitiivisen tarkastelun mallin toimivuudesta suhteessa todellisiin havaintoihin.

4.5 Validointi ja tarkastelu

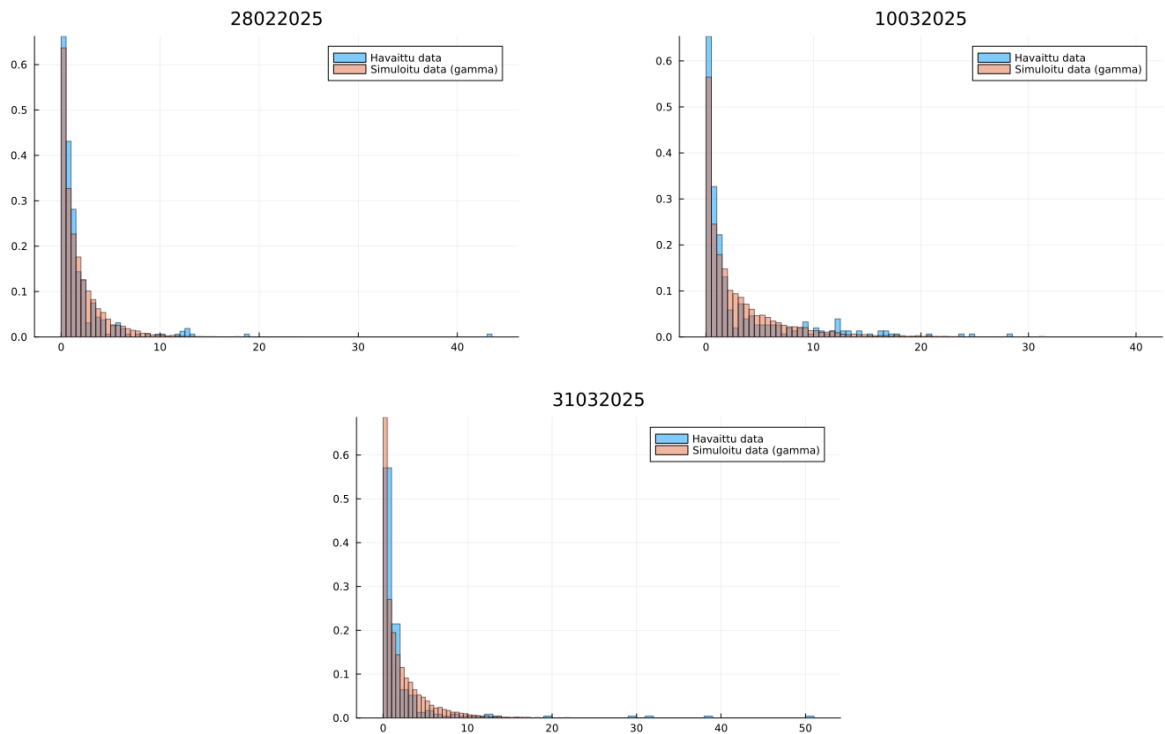
Mallinnuksen luotettavuuden arvioimiseksi suoritetaan tulosten vertailua havaittujen ja simuloitujen viiveiden jakaumien välillä. Simulaatioita toteutetaan käyttäen kahta eri todennäköisyysjakaumaa: gammajakaumaa ja vinonormaalijakaumaa. Tavoitteena on tunnistaa, kumpi näistä kuvaa kerättyä havaintoaineistoa paremmin.

Vertailun perusteella molemmat jakaumat tuottavat realistisia tuloksia, mutta niiden soveltuvuus vaihtelee riippuen siitä, sisällytetäänkö käytettävään dataan myös nollaviiveet. Gammajakauma soveltuu erityisesti positiivisten viiveiden mallintamiseen, kun taas vinonormaalijakauma mahdollistaa myös nollaviiveiden huomioimisen jakauman vasemmassa laidassa. Kummankin jakauman avulla saadaan simuloitua viiveiden tiheysjakaumia, jotka visuaalisesti muistuttavat havaittua dataa, mutta eksakteja goodness-of-fit-mittareita, kuten Kolmogorov-Smirnovin testiä, voitaisiin käyttää tarkempaan arviointiin jatkotyössä.



Kuva 1: Vinonormaalijakaumaan sovitetut datapisteet

Vinonormaalijakaumaan sovitetuista kuvaajista (Kuva 1) nähdään, että jakaumalla pystytään osittain kuvantamaan viiveitä, mutta havaittu ja simuloitu data eroavat kuitenkin melko suuresti toisistaan. Lisäksi nähdään, että vaikka havaitussa datassa on mukana myös yksittäisiä suurempia viivästyksiä, vinonormaalijakauma ei ota näitä huomioon kuvaajassa. Yksittäiset suuremmat viivästyksset ovat harvinaisempia, mutta niiden vaikutus junaliikenteeseen on kuitenkin merkittävä. Yksikin suurempi viivästys voi vaikuttaa muiden junien aikatauluihin esimerkiksi siten, että aikataulun mukaisesti asemalle saapuva juna joutuu odottamaan, että raide vapautuu myöhässä olleelta junalta.



Kuva 2: Gammajakaumaan sovitetut datapisteet

Gammajakauman sovitus kuvaa viiveitä paremmin kuin vinonormaalijakauman sovitus (Kuva 2). Havaittu ja simuloitu data eivät ole aivan samanlaiset, mutta niiden eri ei ole valtava.

Kuvallisen vertailun perusteella havaitaan, että gammajakauma pystyy kuvaamaan viiveiden hajontaa paremmin, erityisesti tapauksissa, joissa jakauma oli selvästi vinoutunut. Tämä korostaa mallin valinnan merkitystä ja tukee tarvetta tehdä perusteltuja jakaumaoletuksia havaintodatan muodon perusteella.

5 Tulokset

Tuloksissa keskitytään vertailemaan kahden eri jakauman – gamma- ja vinonormaalijakauman – soveltuvuutta mallintamaan havaittua viiveaineistoa.

Visualisoinnin perusteella molemmat jakaumat tarjoavat laadukkaita approksimaatioita, mutta niiden erot ilmenevät jakauman muodon yksityiskohdissa. Gammajakauma vastaa hyvin tilanteisiin, joissa havaittu data sisältää pelkästään ei-negatiivisia viiveitä ja näyttää oikealle vinoa jakaumaa. Vinonormaalijakauma tarjoaa joustavuutta tapauksissa, joissa jakauma on asymmetrinen, mutta sisältää sekä nollaviiveitä että pieniä negatiivisia arvoja, joita saattaa esiintyä datassa esimerkiksi ajoissa saapuneiden junien muodossa.

Simuloitujen viiveiden keskiarvot ja hajonnat ovat molemmilla jakaumilla lähellä havaittujen viiveiden vastaavia arvoja:

Taulukko 1: Simuloitujen viiveiden keskiarvo ja hajonta

Jakauma	Simuloitu keskiarvo (min)	Simuloitu hajonta (min)
Gamma	3.21	5.78
Vinonormaali	2.98	5.64

Tämä viittaa siihen, että kumpikin jakauma kykenee mallintamaan viiveiden perusluonteen, mutta jatkotarkastelu esimerkiksi goodness-of-fit –testein on suositeltavaa, mikäli mallin käyttöä halutaan laajentaa esimerkiksi riskianalyysiin tai kapasiteetilaskentaan. Simulaatiot osoittavat lisäksi, että havaittujen viiveiden jakauma ei ole symmetrinen, mikä puoltaa normaalijakauman hylkäämistä yksinkertaisena mallina tässä sovelluksessa. Tämä tukee valintaa käyttää joustavampia jakaumia, kuten gammaa tai vinonormaalia.

Mallinnettujen kuvien pohjalta (Kuva 1 & Kuva 2) nähdään, että gammajakauma kykenee mallintamaan havaittuja viiveitä huomattavasti paremmin kuin vertailussa ollut vinonormaalijakauma. Vaikka mallinnus gammajakaumalla on parempi, se ei kuitenkaan vastaa läheskään täydellisesti havaittua dataa, mikä jättää puolestaan runsaasti tilaa jatkotutkimukselle. Mikäli jatkotutkimuksissa saataisiin mallinnettua täydellisesti sopiva malli, joka vastaa havaittua dataa, voitaisiin tutkimustulokset raportoida eteenpäin Fintraffic-yhtiölle junaliikenteen aikataulujen parantamiseksi.

6 Johtopäätökset

Tässä kandidaatintyössä tarkasteltiin junaliikenteen viivästyksiä stokastisena ilmiönä ja kehitettiin yksinkertainen parametriestimointiin perustuva malli viiveiden kuvaamiseksi. Työssä kerättiin havaintoaineistoa avoimen Digitraffic-rajapinnan kautta ja mallinnettiin viiveiden jakaumaa kahden todennäköisyysjakauman – gammajakauman ja vinonormaalijakauman – avulla.

Tulokset osoittavat, että molemmat jakaumat voivat mallintaa junaliikenteen viiveitä käyttökelpoisesti, mutta niiden soveltuvuus riippuu datan rakenteesta. Gammajakauma sopii erityisesti tilanteisiin, joissa viiveet ovat aina positiivisia, kun taas vinonormaalijakauma mahdollistaa myös nollaviiveiden sisällyttämisen malliin. Kuitenkin gammajakauma soveltuu junaliikenteen viivästysten mallintamiseen paremmin, sillä poikkeama todellisen datan ja simuloitun datan välillä on huomattavasti pienempi. Mallien avulla tuotettiin simuloituja jakauksia, jotka vastasivat hyvin todellisia havaintoja. Niiden avulla voidaan simuloida järjestelmän toimintaa ilman koko liikenneverkon tarkkaa mallia.

Työ osoittaa, että yksinkertaiset stokastiset mallit voivat tarjota hyödyllistä informaatiota junaliikenteen häiriöiden ennustamiseen ja analysointiin. Jatkotutkimuksessa voitaisiin hyödyntää edistyneempiä tilastollisia menetelmiä, kuten bayesilaisia menetelmiä tai koneoppimispohjaisia lähestymistapoja, jotka voivat ottaa huomioon useampia muuttujia samanaikaisesti. Lisäksi olisi perusteltua käyttää useita datalähteitä ja tarkastella viivästysten vaikutuksia esimerkiksi matkustajakokemukseen tai suunniteltujen junayhteyksien katkeamiseen.

Lähdeluettelo

- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12(2), s. 171–178.
- Cacchiani, V., Caprara, A. & Toth, P. (2014). Models and algorithms for train timetable optimization. *Transportation Science* 48(2), s. 161–175.
- Carey, M. & Kwienciński, A. (1995). Stochastic approximation to the effects of headway control on punctuality. *Transportation Research Part B: Methodological* 29(4), s. 281–295.
- Casella, G. & Berger, R. L. (2002). *Statistical Inference*. 2. painos. Duxbury.
- Computing, J. (2023). *Julia Programming Language*. URL: <https://julialang.org/>.
- Developers, J. (2023). *Optim.jl: Julia package for mathematical optimization*. URL: <https://github.com/JuliaNLSolvers/Optim.jl>.
- Fintraffic (2025). *Digitraffic API: Junaliikenteen tietorajapinta*. <https://rata.digitraffic.fi/api/v1/trains>. Saatavilla: <https://rata.digitraffic.fi/api/v1/trains/>.
- Fischetti, M., Grosso, A. & Toth, P. (2009). A probabilistic optimization model for real-time train delay management. *Transportation Science* 43(3), s. 321–335.
- Goverde, R. M. P. (2005). Punctuality of railway operations and timetable stability analysis. *Transportation Research Part A: Policy and Practice* 39(6), s. 495–513.
- Goverde, R. M. P. (2010). Railway timetable stability analysis using max-plus system theory. *Transportation Research Part B: Methodological*.
- Govere, R. M. P. & Kecman, P. (2015). Online data-driven train delay prediction. *IEEE Transactions on Intelligent Transportation Systems* 16(1), s. 465–474.
- Huisman, T., Boucherie, R. J. & Dijk, N. M. V. (2004). A stochastic model for railway delays. *Operations Research Spectrum*.
- Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1*. Wiley.
- Kecman, P. & Goverde, R. M. P. (2015). Online data-driven train delay prediction. *IEEE Transactions on Intelligent Transportation Systems* 16(1), s. 465–474.
- Law, A. M. & Kelton, W. D. (2000). *Simulation Modeling and Analysis*. 3. painos. McGraw-Hill.
- Law, A. M. & Kelton, W. D. (2007). *Simulation Modeling and Analysis*. 4. painos. McGraw-Hill Education.
- Ross, S. M. (2014). *Simulation*. 5. painos. Academic Press.
- Väylävirasto (2022). *Rautatieverkon tilannekatsaus*. Technical Report. Suomi: Väylävirasto.
- Väylävirasto (2023b). *Rataverkko*. Verkkosivu. Saatavilla: <https://vayla.fi/vaylista/rataverkko>.
- Väylävirasto (2023a). *Rautateiden tavaraliikenne 1990-2023*. Verkkosivu. Saatavilla: https://vayla.fi/documents/25230764/55126781/Rautateiden+tavaraliikenne_1990-2023.

pdf/85ae9984-988a-77fe-52e7-264ea6f1b906/Rautateiden+tavaraliikenne_1990-2023.pdf?t=1714049000925.

Yuan, J. & Hansen, I. A. (2007). Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B*.

Liite A Algoritmit Julia-ohjelmointikielellä

Algorithm 1 Junaviiveiden stokastinen mallinnus ja parametriestimointi

```
1: Alusta tyhjä DataFrame viiveaineiston tallennukseen
2: for jokainen  $pvm \in \{28.2.2025, 10.3.2025, 31.3.2025\}$  do
3:   Lataa junadata Digitraffic-rajapinnasta annetulle päivämäärälle
4:   for jokainen juna do
5:     for jokainen pysähdysasemariivi do
6:       if asema = HKI ja tyyppi = ARRIVAL then
7:         Laske viive:  $actual\_time - scheduled\_time$ 
8:         Tallenna viive DataFrameen
9:       end if
10:    end for
11:  end for
12:  Tallenna DataFrame CSV-tiedostoon
13: end for
14: for jokainen CSV-tiedosto do
15:   Lue viiveaineisto tiedostosta
16:   Poista puuttuvat ja virheelliset arvot
17:   Valitse sopiva jakauma (esim. gamma tai vino-normal)
18:   Estimoi jakauman parametrit maksimoimalla todennäköisyys
19:   Generoi satunnaisia viiveitä sovitetusta jakaumasta
20:   Piirrä histogrammi: havaittu vs. simuloitu jakauma
21:   Tallenna kuva
22: end for
```
