

JOENSUUN YLIOPISTO
TIETOJENKÄSITTELYTIETEEN LAITOS
Raporttisarja A

**Puutteiden lukumäärän estimointi
Norden/Rayleigh-jakauman ja
Gamma-jakauman
avulla**

Matti Niemi

Report A-2001-5
(Korjattu)

ISSN 0789-7316
ISBN 952-458-084-5

Puutteiden lukumäärän estimointi Norden/Rayleigh-jakauman ja Gamma-jakauman avulla

Matti Niemi

Tietojenkäsittelytieteen laitos

Joensuun yliopisto

PL 111, 80101 Joensuu

(Korjattu versio)

Tiivistelmä

Ohjelmistotuotannossa voidaan estimoida puutteiden lukumäärätietoja erilaisilla matemaattisilla malleilla. Tutkimusraportissa sovitetaan Norden/Rayleigh-jakauman ja Gamma-jakauman mukaiset käyrät pienimmän neliösumman menetelmän ja logaritimuunnoksen avulla havaintoaineistoon. Aineistona on kirjallisuudesta saadut, todellisissa ohjelmistoprojekteissa kerätyt puutetiedot. Saatuja malleja verrataan keskenään projektitietojen avulla. Lisäksi selvitetään, kuinka olemassa olevaa mallia voidaan muuttaa ottamalla huomioon käynnissä olevan projektin puutetiedot. Tulokset osoittavat jakaumien soveltuvan lähes yhtä hyvin mallintamaan puutetietoja. Molempien jakaumien tuottamien mallien käyttäytyminen on kuitenkin riippuvainen havaintoaineistosta.

Avainsanat: Gamma-malli, Norden/Rayleigh-malli, puutteiden lukumäärän estimointi.

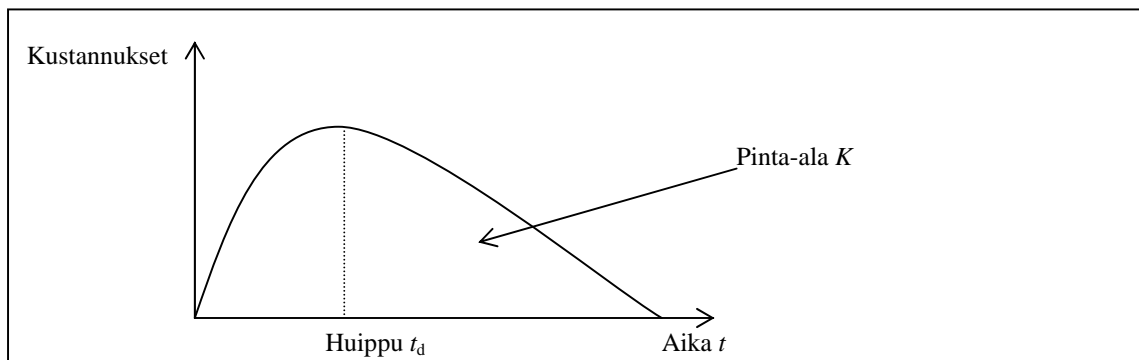
1 Johdanto

Ohjelmiston ja ohjelmistoprosessin mittaamiseksi pitää kerätä tietoja, joilla helpotetaan aikataulun, kustannusten ja laadun valvontaa. On tärkeää johdonmukaisesti määritellä ja

mitata suoraan mitattavissa olevia peruskohteita, kuten ohjelmiston koko, puutteet, työ-
määrä ja aika.

Putnam (1978) kehitti arviointimenetelmän, nk. Putnamin menetelmän eli SLIM-mene-
telmän (Software Life Cycle Management), ohjelmointikustannusten arvioimiseksi. Mene-
telmä perustuu näkemykseen, että ohjelmistoprojektien kustannukset seuraavat hyvin mää-
riteltyä kaavaa. Mallin tutkimustiedot perustuvat yli 200 ohjelmiston kehittämisprojektista,
pääasiallisesti yhdysvaltalaisen US Army Computer Systems Command organisaatiosta,
kerättyihin tietoihin (Londeix, 1987).

Putnamin menetelmä on dynaaminen, monen muuttujan malli ja se perustuu Norden/Ray-
leigh-mallin yhtälöstä saatavaan käyrään. Putnamin menetelmä mallintaa aikaisempien, sa-
man kaltaisten ohjelmistoprojektien historiatietoja käyttäen projektimallin, jonka mukaan
voidaan tulevia projektia arvioida. Mallista havaitaan, että kustannusten määrä on kasvava
ennen käyrän huippukohtaa ja pienenee tämän jälkeen. Käyrän periaate on esitetty kuvassa
1.



Kuva 1: Norden/Rayleigh-käyrä.

Pillai ja Nair (1997) ovat vertailleet Rayleigh-jakaumaa Gamma-jakauman kanssa ohjel-
mistoprojektin miehityksen estimoinnissa. Tutkimuksessaan he vertailivat jakaumia yhdes-
tä projektista saatujen miehitystietojen avulla ja totesivat jakaumien olevan lähes ekviva-
lenteja. Kitchenham ja Linkman (1997) ovat käyttäneet Rayleigh-jakaumaa työmäärän es-
timointiin.

Kanin (1995) mukaan Norden/Rayleigh-käyrää voidaan käyttää yleisenä laadunvalvonnan mallina, jolloin käyrä voisi kuvata puutteita ajan funktiona. Fentonin et al. (1999) mukaan puutteiden lukumäärän ennustaminen on eräs laadun ennustamisessa sovellettava käytäntö. Puute (defect) yleisesti voidaan määritellä suunnittelemattomaksi piirteeksi, joka heikentää alkion hyödyllisyyttä tai arvoa, tai muunlaiseksi vajavaisuudeksi, epätäydellisyydeksi tai riittämättömyydeksi (Florac et al. 1992). Ohjelmistossa, siihen liittyvässä muussa tuotoksessa tai ohjelmistoprosessissa, puute on mikä tahansa vika tai epätäydellisyys.

Tässä tutkimuksessa Rayleigh-jakaumaa sovitetaan viiteen esimerkkiaineistoon puutteiden lukumäärän estimoimiseksi. Norden/Rayleigh-käyrän antamia tuloksia verrataan Gamma-jakauman antamiin tuloksiin. Lisäksi selvitetään uudesta projektista kerättyjen puutteiden vaikutusta olemassa olevaan vanhaan malliin. Tavoitteena tällöin on selvittää parantaako uudesta projektista saadut puutetiedot ennustettavuutta projektin aikana.

2 Norden/Rayleigh-käyrän sovittaminen

Norden/Rayleigh-yhtälö on muotoa (Putnam, 1978):

$$p(t) = 2Kate^{-at^2} \quad (1)$$

Kaavassa (1) K tarkoittaa yhtälön kuvaaman käyrän rajoittamaa pinta-alaa, a on positiivinen vakio siten, että $a = 1/2t_d^2$, missä t_d on käyrän maksimikohta ja t tarkoittaa ajankohtaa. Kaavaa (1) vastaava kertymäfunktio on muotoa

$$P = K(1 - e^{-at^2}) \quad (2)$$

Jotta kaavaa (1) voidaan soveltaa ohjelmistoprojekteissa, pitää se sovittaa havaintoaineistoon. Sovittamisen jälkeen Norden/Rayleigh-yhtälö voidaan ohjelmoida käytössä olevaan laadunvalvontavälineeseen (Niemi, 2002). Sovittaminen voidaan tehdä pienimmän neliö-

summan avulla käyttäen logaritimuunnosta. Logaritmuunnos kaavasta (1) saadaan seuraavasti:

$$\begin{aligned}
 \ln p(t) &= \ln(2Kae^{-at^2}) \\
 \ln p(t) &= \ln 2 + \ln K + \ln a + \ln t - at^2 \\
 \ln p - \ln t &= \ln 2 + \ln K + \ln a - at^2 \\
 \ln(p/t) &= \ln 2 + \ln K + \ln(1/2t_d^2) - t^2/2t_d^2 = \ln(K/2t_d^2) + \ln 2 + (-1/2t_d^2)t^2 \\
 &= \ln(K/t_d^2) - \ln 2 + \ln 2 + (-1/2t_d^2)t^2 \\
 &= \ln(K/t_d^2) + (-1/2t_d^2)t^2
 \end{aligned} \tag{3}$$

Asettamalla $y = \ln(p/t)$ ja $x = t^2$ saadaan kaava (3) lineaariseen muotoon:

$$y = mx + c \tag{4}$$

missä kulmakerroin m saadaan ratkaisemalla yhtälö

$$m = -1/2t_d^2 \tag{5}$$

ja y-akselin leikkauspiste c ratkaisemalla yhtälö

$$c = \ln(K/t_d^2) \tag{6}$$

Kun m ja c tunnetaan voidaan kaavasta (5) ratkaista t_d

$$t_d = (-1/2m)^{(-1/2)} \tag{7}$$

ja kaavasta (6) voidaan ratkaista K

$$c = \ln K - \ln(t_d^2)$$

$$\ln K = c + \ln(t_d^2)$$

$$K = e^{(c + \ln(t_d^2))}$$

Kulmakerroin m ja leikkauspiste c ratkaistaan pienimmän neliösumman menetelmällä käyttäen kaavoja (Spiegel, 1961):

$$c = \left(\left(\sum Y \right) \left(\sum X^2 \right) - \left(\sum X \right) \left(\sum XY \right) \right) / \left(N \sum X^2 - \left(\sum X \right)^2 \right)$$

$$m = \left(\left(N \sum XY \right) - \left(\sum X \right) \left(\sum Y \right) \right) / \left(N \sum X^2 - \left(\sum X \right)^2 \right)$$

missä X on aikavektori ($x_i = t_i^2$) ja Y on vastaava projektikohtainen puutevektori ($y_i = \ln(p_i / t_i)$).

3 Gamma-mallin sovittaminen

Pillai ja Nair (1997) ovat osoittaneet, että kaavan (4) mukainen linearisointi voidaan suorittaa myös asettamalla $x = t$ ja $y = \ln(p / t^2)$ vaikuttamatta Putnamin mallin informaatioisältöön. Tätä voidaan soveltaa mallinnettaessa puutteita Gamma-jakauman avulla, jonka tiheysfunktio on muotoa (Musa et al., 1987):

$$f(t | \alpha, \beta) = 1 / (\Gamma(\alpha) \beta^\alpha) \left(t^{(\alpha-1)} \right) e^{(-t / \beta)} \quad (8)$$

Linearisointiyhtälö (4) kirjoitetaan muotoon

$$\ln(p / t^2) = \gamma + \mu t \quad (9)$$

missä μ on kulmakerroin ja γ on y-akselin leikkauspiste. Poistamalla kaavassa (9) logarit-
mifunktiot molemmilta puolilta saadaan

$$\begin{aligned}\ln(p) - \ln(t^2) &= \gamma + \mu t \\ \ln(p) &= \gamma + \mu t + \ln(t^2) \\ p &= (e^\gamma)(e^{\mu t})(e^{\ln(t^2)}) = e^\gamma e^{\mu t} t^2\end{aligned}\tag{10}$$

Yhtälö (10) on Gamma-jakauman (8) erikoistapaus merkitsemällä

$$\begin{aligned}1/(\Gamma(\alpha)\beta^\alpha) &= e^\gamma \\ \alpha - 1 &= 2 \\ -1/\beta &= \mu\end{aligned}\tag{11}$$

josta saadaan $\alpha = 3$. Ottamalla derivaatta yhtälöstä (8) saadaan

$$\begin{aligned}df(t | \alpha, \beta) / dt &= 1/(\Gamma(\alpha)\beta^\alpha) / (\Gamma(\alpha)\beta^\alpha) ((\alpha - 1)t^{(\alpha-2)})(e^{(-t/\beta)}) + (-1/\beta)(e^{(-t/\beta)})(t^{(\alpha-1)}) \\ df(t | \alpha, \beta) / dt &= e^{(-t/\beta)} / (\Gamma(\alpha)\beta^\alpha) ((-t^{(\alpha-1)})/\beta) + (\alpha - 1)(t^{(\alpha-2)})\end{aligned}\tag{12}$$

Derivaatan huippukohta saadaan kaavoista (11) ja (12):

$$t^{(\alpha-1)}(-1/\beta) + (\alpha - 1)(t^{(\alpha-2)}) = 0\tag{13}$$

johon sijoittamalla $\alpha = 3$ saadaan

$$\begin{aligned}t^2(-1/\beta) + 2t &= 0 \\ 2t &= t^2(1/\beta) \\ 2 &= t(1/\beta)\end{aligned}$$

$$t = 2\beta$$

Norden/Rayleigh-käyrässä huippukohta saavutetaan ajan hetkellä

$$t_d = 2\beta$$

josta saadaan

$$\beta = t_d / 2 \tag{14}$$

Integroimalla kaavan (8) Gamma-jakauman Gamma-funktio (Musa et al., 1987) nolasta äärettömään saadaan, kun $\alpha = 3$ (Tuominen ja Norlamo, 1983):

$$\begin{aligned} \Gamma(\alpha) &= \int t^{(\alpha-1)} e^{-t} dt = \\ (\alpha - 1)! &= (3 - 1)! = 2 \end{aligned} \tag{15}$$

Ottamalla huomioon Norden/Rayleigh-käyrän rajoittama pinta-ala on K saadaan Gamma-jakauman (8) ja kaavan (14) avulla, kun $\alpha = 3$:

$$\begin{aligned} f(t | \alpha, \beta) &= \left(8K / (\Gamma(3) t_d^3) \right) \left(t^2 e^{-2t/t_d} \right) = \\ & \left(4K / t_d^3 \right)^2 e^{-2t/t_d} \end{aligned} \tag{16}$$

Yhtälöllä (16) on vastaava muoto kuin Norden/Rayleigh-käyrällä, mutta Gamma-malli on yleisempi ja helpompi sovittaa kuin Putnamin malli (Pillai ja Nair, 1997). Gamma-mallin sovittamiseksi suoritetaan linearisointi logaritmuunnoksella kaavasta (8) ottamalla huomioon Putnamin mallin pinta-ala K :

$$\begin{aligned} p &= K / (\Gamma(\alpha) \beta^\alpha) \left(t^{(\alpha-1)} \right) e^{-(t/\beta)} \\ \ln p &= \ln(K / (\Gamma(\alpha) \beta^\alpha)) + \ln(t^{(\alpha-1)}) + \ln(e^{-(t/\beta)}) \end{aligned}$$

$$\ln(p/t^{(\alpha-1)}) = \ln(K/(\Gamma(\alpha)\beta^\alpha)) - (1/\beta)t \quad (17)$$

Kaava (17) linearisoidaan muotoon:

$$y = \gamma + \mu x, \text{ missä}$$

$$y = \ln(p/t^{(\alpha-1)}) = \ln(p/t^2) \text{ ja } x = t \quad (18)$$

Kulmakerroin μ saadaan ratkaisemalla yhtälö

$$\mu = -1/\beta = -(\alpha - 1)/t_d \quad (19)$$

ja y-akselin leikkauspiste γ ratkaisemalla yhtälö

$$\gamma = \ln(K/(\Gamma(\alpha)\beta^\alpha)) \quad (20)$$

Kun $\alpha = 3$ ja γ sekä μ tunnetaan pienimmän neliösumman menetelmän avulla, voidaan ratkaista t_d ja K . Arvo t_d saadaan suoraan kaavasta (19):

$$t_d = -2/\mu \quad (21)$$

Arvo K lasketaan kaavasta (20):

$$e^\gamma = K/\Gamma(\alpha)\beta^\alpha$$

$$K = (\Gamma(\alpha)\beta^\alpha)e^\gamma = 2\beta^3 e^\gamma = (2t_d^3)/(2^3)e^\gamma = (t_d^3/4)e^\gamma \quad (22)$$

Kitchenham (1997) ehdottaa parametrin α arvoksi $\alpha = 2$ työmääräkustannuksia mallinnettaessa. Tällöin jakauma on etupainotteisempi verrattuna arvoon $\alpha = 3$. Vastaavasti kaavat (14), (15), (16), (18), (19), (21) ja (22) muuttuvat muotoon:

$$\beta = t_d \quad (14')$$

$$\Gamma(\alpha) = \int t^{(\alpha-1)} e^{-t} dt = 1 \quad (15')$$

$$f(t | \alpha, \beta) = \left(8K / t_d^3 \right) t^2 e^{(-2t/t_d)} \quad (16')$$

$$y = \ln(p/t) \text{ ja } x = t \quad (18')$$

$$\mu = -1/t_d \quad (19')$$

$$t_d = -1/\mu \quad (21')$$

$$K = t_d^2 e^\gamma \quad (22')$$

4 Soveltaminen esimerkkiaineistoon

Taulukon 1 esimerkkiaineisto perustuu IBM Watson Research Center –tutkimuslaitoksessa erään puuteluokittelun, Orthogonal Defect Classification (ODC), soveltamista varten viidestä eri projektista hankittuihin puuteluetteloihin (Lyu, 1995). Luetteloissa olleiden puutteiden määrä on suhteutettu kahteentoista yhtä pitkään ajankohtaan. Projektia ODC4 ei ole otettu mukaan tähän tutkimukseen, koska se keston ja puutteiden lukumäärän suhteen oli täysin erilainen suhteessa muihin projekteihin.

Taulukko 1: Esimerkkiaineisto havaituista puutteista.

Projekti	Ajankohta											
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
ODC1	37	66	95	139	70	174	119	136	109	113	128	21
ODC2	51	45	3	16	20	6	8	4	20	40	20	22
ODC3	26	14	33	40	20	38	19	16	40	54	67	33
ODC5	11	9	52	35	14	121	73	45	8	4	6	8
ODC6	4	5	5	20	21	27	16	16	16	21	21	21

4.1 Norden/Rayleigh-mallin ja Gamma-mallin arviointi projekteittain

Todellisen tiedon sekä mallin ennustaman tiedon poikkeamasta käytetään ilmaisua tietojoukon jäännös (residual). Jäännöksen perusteella laskettua standardipoikkeamaa ja keskiarvoa voidaan käyttää perustana mallin hyvyyden ja kahden projektin samankaltaisuuden arvioin-

nissa. Havaittujen ja ennustettujen arvojen eroavaisuutta voidaan ajatella ns. ennustevirheenä (prediction error). Keskimääräisestä ennustevirheestä käytetään nimitystä ennustevirhe (prediction bias) ja sen standardipoikkeamaa käytetään usein ennustevaihtelun (variation) mittarina. Ennustevirheen neliön keskiarvon neliöjuuri (Root Mean Square Prediction Errors, RMSPE) on mittari siitä tarkkuudesta, kuinka malli ennustaa todellista tietoa (Pillai ja Nair, 1997).

Kaavat (23) – (26) esittävät mittarin RMSPE laskennan.

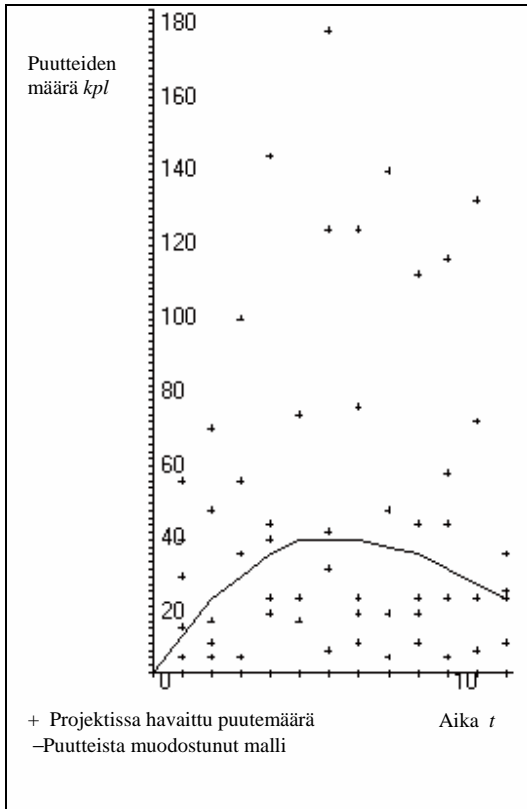
$$RMSPE = \sqrt{Bias^2 + Variation^2} \quad (23)$$

$$PE_i = (Actual_i - Predicted_i) \quad (24)$$

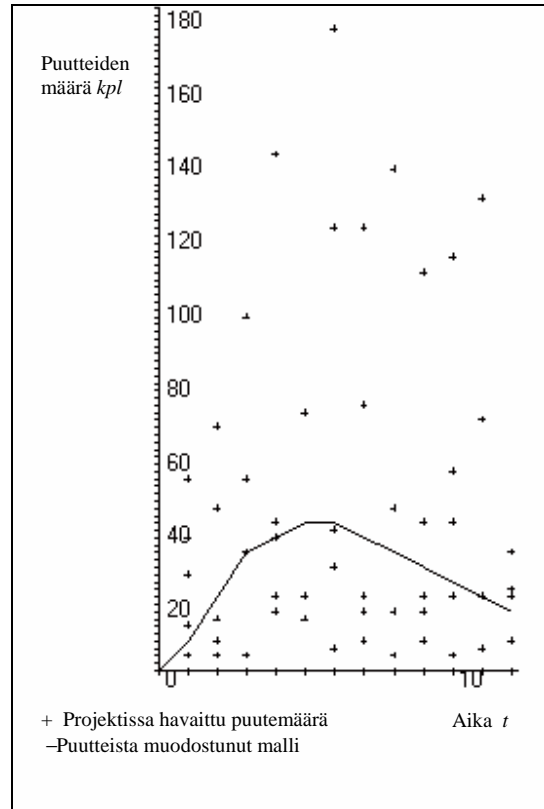
$$Bias = \sum_{i=1}^N \frac{PE_i}{N} \quad (25)$$

$$Variation = \sqrt{\frac{\sum_{i=1}^N (PE_i - Bias)^2}{N - 1}} \quad (26)$$

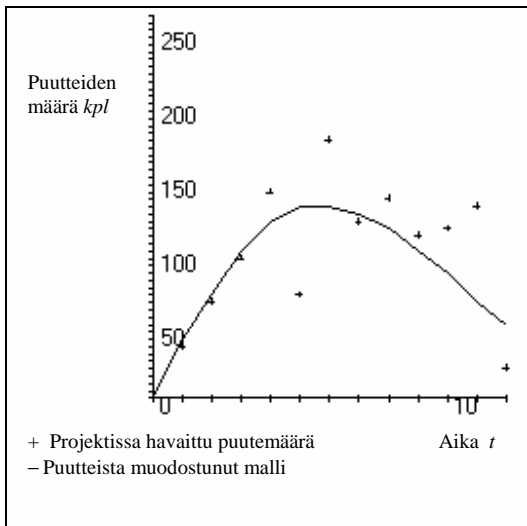
Soveltamalla luvussa 3 esitettyä käyrän sovittamista saadaan taulukon 1 aineiston perusteella projektien puutteista kuvien 2 ja 3 mukaiset mallit ajan t suhteen. Näitä malleja voidaan soveltaa ennustettaessa saman tyyppisten projektien puutemääriä projektien eri vaiheissa. Kun tunnetaan puutteiden korjaamiseen tarvittava keskimääräinen työmäärä, voidaan projektia suunniteltaessa hyödyntää mallia resursseja varattaessa. Taulukkojen 2 ja 3 viimeisillä riveillä on esitetty kuvien 2 ja 3 mallien ja todellisten puutemäärien erot.



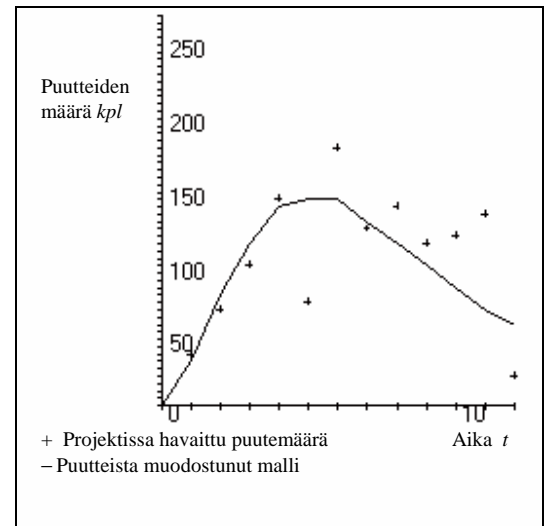
Kuva 2. Norden/Rayleigh-malli.



Kuva 3. Gamma-malli.



Kuva 4. Projektin ODC1
Norden/Rayleigh-malli.



Kuva 5. Projektin ODC1 Gamma-malli
($\alpha = 3$).

Kustakin projektista voidaan laatia yksilöllinen malli, joka perustuu projektin omiin puute-määriin. Tämän perusteella voidaan arvioida kuinka hyvin Norden/Rayleigh-malli ja Gamma-malli pystyvät mallintamaan yksittäistä projektia. Kuvissa 4 ja 5 on esitetty taulukon 1 projektin ODC1 perusteella lasketut mallit.

Taulukoissa 2 ja 3 on esitetty projektikohtaisen Norden/Rayleigh-mallin sekä Gamma-mallin ja projektissa havaittujen puutteiden arvojen erot taulukon 1 viidelle projektille. Pillain ja Nairin (1997) tutkimuksessa tarkastellaan yhdestä projektista kerättyjä miehitystietoja. He havaitsivat, että Gamma-malli antaa hieman pienemmän vinoutuman verrattuna Norden/Rayleigh-malliin. Ennustevaihtelu ja RMSPE olivat Norden/Rayleigh-mallissa hieman pienempiä kuin Gamma-mallissa. Taulukkojen 2 ja 3 perusteella vinoutuma on saman suuntainen Pillain ja Nairin tulosten kanssa, paitsi projektien ODC5 ja ODC6 osalta, joissa Norden/Rayleigh-mallin antama tulos on hieman pienempi kuin Gamma-mallin antama tulos.

Myöskin vaihtelu ja RMSPE vastaavat Pillain ja Nairin saamia tuloksia, paitsi projektin ODC6 osalta, jossa Gamma-mallin antama tulos on pienempi kuin Norden/Rayleigh-mallin antama tulos. Mallien käyttäytyminen näyttää siis osaksi olevan riippuvainen havaintoaineistosta ¹. Tämä näkyy myös taulukkojen 2 ja 3 viimeisille riveille lasketuista arvoista. Taulukon 4 perusteella nähdään, että Gamma-malli arvolla $\alpha = 2$ mallintaa selvästi huonommin taulukon 1 projekteja kuin Gamma-malli arvolla $\alpha = 3$.

Taulukko 2: Norden/Rayleigh-mallin ja havaittujen puutearvojen erot.

Projekti	Vinoutuma	Vaihtelu	RMSPE
ODC1	6,250	32,407	33,004
ODC2	4,500	20,039	20,538
ODC3	0,500	16,828	16,835
ODC5	10,000	30,704	32,291
ODC6	0,333	4,716	4,728
Kaikki projektit	13,267	39,963	42,108

¹ Kun tässä tutkimuksessa käytetyllä ohjelmistolla (Niemi, 2002) muodostetaan taulukon (taulukko2) projekteille mallit ja käytetään näitä malleja lähdetietoina tulee Bias ja siten myös RMSPE nolaksi. Tämä tarkoittaa sitä, että ero tunnusluvuissa (Pillai ja Nair, 1997) johtuu lähdeaineiston eroavaisuudesta.

Taulukko 3: Gamma-mallin ($\alpha=3$) ja havaittujen puutearvojen erot.

Projekti	Vinoutuma	Vaihtelu	RMSPE
ODC1	3,750	35,079	35,279
ODC2	4,167	21,238	21,643
ODC3	-0,333	19,685	19,688
ODC5	10,167	31,365	32,972
ODC6	-0,417	4,399	4,419
Kaikki projektit	12,600	40,150	42,080

Taulukko 4: Gamma-mallin ($\alpha=2$) ja havaittujen puutearvojen erot.

Projekti	Vinoutuma	Vaihtelu	RMSPE
ODC1	-123,750	65,187	139,870
ODC2	-17,333	28,864	33,669
ODC3	-49,167	31,951	58,637
ODC5	-12,750	37,414	39,527
ODC6	-18,833	17,012	25,379
Kaikki projektit	-25,150	42,638	49,503

Gamma-mallin antaman ennusteen poikkeama (prediction error deviation, PED) prosentteina Norden/ Rayleigh-mallista voidaan laskea kaavalla

$$PED = \frac{Gamma_{rmspe} - Norden / Rayleigh_{rmspe}}{Norden / Rayleigh_{rmspe}} * 100 \quad (27)$$

Ennustepoikkeamat taulukoiden 2 ja 3 perusteella on laskettu taulukossa 5. Tulos on saman suuntainen Pillain ja Nairin (1997) saaman tuloksen kanssa, eli mallit ovat lähes ekvivalentit tunnusluvun RMSPE avulla tulkittuna.

Taulukko 5. Gamma-mallin ($\alpha=3$) ennustepoikkeama prosentteina Norden/Rayleigh-mallista.

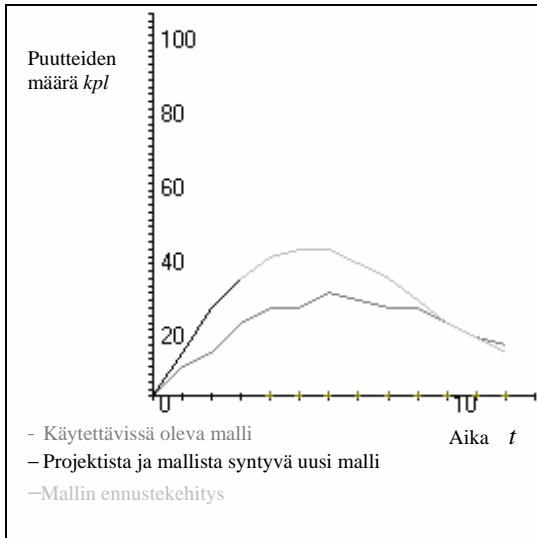
Projekti	PED
ODC1	6,89
ODC2	5,38
ODC3	10,95
ODC5	2,11
ODC6	-6,53
Kaikki projektit	-0,66

4.2 Norden/Rayleigh-mallin ja Gamma-mallin laskeminen projektin edetessä

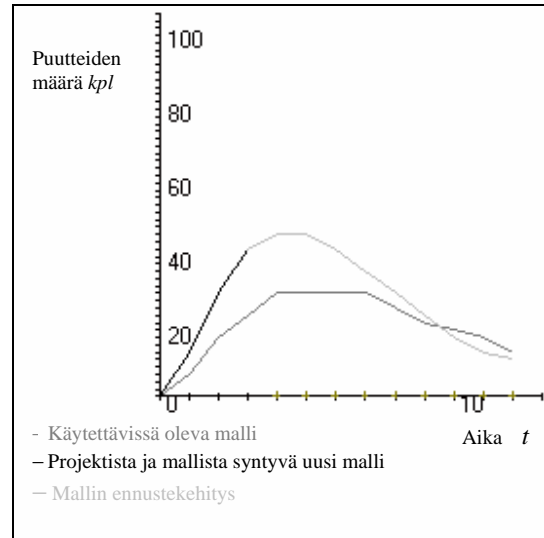
Historiatiedon pohjalta laadittua mallia voidaan käyttää estimoitaessa uuden projektin puutetietoja. Periaatteessa prosessin pitäisi parantua ajan funktiona eli mittaustuloksista pitäisi oppia tekemään työ paremmin. Eräänä tavoitteena voidaan pitää puutteiden löytämistä mahdollisimman aikaisin projektin edetessä (Kan, 1995).

Seuraavassa pyritään selvittämään, pystyykö uudesta projektista kerättyjen puutetietojen perusteella vaikuttamaan aiempien projektien puutetiedoista muodostettuun malliin uuden projektin edetessä. Uuden mallin laskennassa otetaan huomioon uuden projektin huomioidut puutetiedot nykyhetkeen saakka ja vanhan mallin muodostaman käyrän pisteet koko elinkaaren ajalta. Uutta mallia verrataan vanhaan malliin neljän virstanpylvään T3, T6, T9 ja T12 kohdalla. Virstanpylväät jakavat projektin elinkaaren siis neljään yhtä pitkään ajanjaksoon.

Kuvissa 6 ja 7 projektin ODC1 tuottamien puutetietojen avulla ajanhetkeen T3 mennessä on laskettu uudet mallit. Vanhat mallit on laskettu taulukon 1 projektitietojen perusteella pois lukien projekti ODC1. Tässä tapauksessa muutos vanhaan malliin on selvä. Intuitiivisesti ajatellen vaikutus on kuitenkin riippuvainen siitä, kuinka paljon uusi projekti eroaa aiemmista projekteista. Ero aiheutuu myös, mikäli uudessa projektissa puutteet pystytään löytämään aiemmin eli on pystytty parantamaan prosessia.



Kuva 6. Projektin ODC1 Norden/Rayleigh-malli ajankohdassa T3.



Kuva 7. Projektin ODC1 Gamma-malli ($\alpha=3$) ajankohdassa T3.

Taulukoissa 6 ja 8 on laskettu muutokset projektin edetessä kullekin taulukon 1 projektille. Taulukoiden arvot osoittavat, että uuden projektin vaikutus malliin on projektista riippuva. Toisaalta tuloksista nähdään, että vanhan ja uuden mallien ero muuttuu projektin edetessä. Myös tämä muutos on projekteista riippuvaista, mutta useimmissa tapauksissa vinoutuman itseisarvo, vaihtelu sekä RMSPE pienenevät projektin edetessä. Tämä tarkoittaa sitä, että projektin uudet arvot pystyvät parantamaan uutta mallia. Tulokset ovat samansuuntaisia molemmilla jakaumilla.

Taulukoissa 7 ja 9 on laskettu kullekin projektille tunnusluvut vanhan mallin perusteella projektin päättyessä. Vanha malli on muodostettu muiden kuin tarkasteltavan projektin puutetietojen perusteella. Taulukoiden RMSPE-arvoja ja taulukoiden 6 ja 8 ajankohdan T12 RMSPE-arvoja vertaamalla voidaan havaita, että uusi malli pystyy ennustamaan kunkin projektin puutetietoja vanhaa mallia paremmin. Ainoa poikkeus ilmenee projektin ODC5 kohdalla. Vanhan ja uuden mallin ennustamiskykyä on tarkemmin vertailtu taulukoissa 10 ja 11, joissa on laskettu uuden mallin ennustepoikkeama vanhasta mallista kaavan (27) PED-luvun avulla. Esimerkiksi taulukon 10 PED-arvo ajanhetkellä T12 projektin ODC1 osalta saadaan taulukoista 6 ja 7 laskemalla: $(68,137 - 90,572) / 90,572 * 100 = -24,770$.

Taulukoista 10 ja 11 nähdään, että ainoastaan Gamma-jakauma antaa projektille ODC5 positiivisen PED-arvon projektin päättyessä, jolloin vanha malli siis on parempi kuin uusi malli. Hyvin yleinen trendi on myöskin PED-tunnusluvun paraneminen projektin edetessä uuden mallin kannalta katsoen. Esimerkiksi projektin ODC1 osalta ajanhetkellä T3 uusi malli antaa Norden/Rayleigh-jakauman osalta noin 7 % paremman ennusteen kuin vanha malli. Vastaavasti ajanhetkellä T6 ero on noin 11 % uuden mallin hyväksi. Jakaumien välillä ei näytä olevan merkittäviä eroja PED-tunnusluvun avulla tulkittuna.

Taulukko 6: Norden/Rayleigh-malli projektin edetessä: uusi malli vanhan mallin avulla.

Projekti	Vinoutuma				Vaihtelu				RMSPE			
	T3	T6	T9	T12	T3	T6	T9	T12	T3	T6	T9	T12
ODC1	74,200	68,823	62,658	57,819	39,417	39,044	37,532	36,052	84,417	79,127	73,038	68,137
ODC2	-9,001	-4,125	0,477	-0,865	24,284	22,700	21,650	21,782	25,901	23,072	21,655	21,800
ODC3	6,212	6,873	8,089	4,728	20,078	19,822	19,629	17,820	21,017	20,980	21,230	18,436
ODC5	2,546	1,875	3,489	7,990	32,643	32,773	31,985	30,685	32,742	32,773	32,175	31,708
ODC6	-8,516	-7,817	-5,347	-5,322	5,284	5,140	5,217	5,194	10,022	9,356	7,470	7,436

Taulukko 7: Norden/Rayleigh-malli kullekin projektille muiden projektien avulla.

Projekti	Vinoutuma	Vaihtelu	RMSPE
ODC1	81,025	40,476	90,572
ODC2	-9,092	24,205	25,856
ODC3	7,713	19,012	20,517
ODC5	3,104	32,910	33,056
ODC6	-15,165	9,64	17,970

Taulukko 8: Gamma-mallin ($\alpha = 3$) ja projektissa havaittujen puutearvojen erojen muutokset projektin edetessä.

Projekti	Vinoutuma				Vaihtelu				RMSPE			
	T3	T6	T9	T12	T3	T6	T9	T12	T3	T6	T9	T12
ODC1	73,867	68,760	61,923	56,556	40,528	40,376	38,203	35,913	84,254	79,738	72,759	66,995
ODC2	-9,729	-3,892	0,341	-1,549	26,110	24,060	22,791	23,241	27,864	24,373	22,793	23,292
ODC3	5,597	6,768	7,712	3,88	21,683	21,282	21,110	20,009	22,394	22,332	22,475	20,382
ODC5	1,774	1,793	3,132	7,424	30,660	30,821	30,603	30,584	30,711	30,873	30,763	31,472
ODC6	-9,526	-8,076	-5,549	-6,000	6,730	6,330	6,505	6,305	11,663	10,261	8,550	8,704

Taulukko 9: Gamma-malli ($\alpha=3$) kullekin projektille muiden projektien avulla.

Projekti	Vinoutuma	Vaihtelu	RMSPE
ODC1	80,431	40,012	89,834
ODC2	-10,010	25,965	27,828
ODC3	6,953	20,383	21,536
ODC5	2,289	30,888	30,973
ODC6	-16,118	12,316	20,285

Taulukko 10: Norden/Rayleigh-malli projektin edetessä: uuden mallin ennustepoikkeama vanhasta mallista PED-tunnusluvulla mitattuna.

Projekti	T3	T6	T9	T12
ODC1	-7,2346672727	-12,6363220015	-19,3590625880	-24,7701430813
ODC2	0,1742336624	-10,7679462917	-16,2481288054	-15,6868635655
ODC3	2,4361684247	2,2559016811	3,4749687142	-10,1413744899
ODC5	-0,9509722975	-0,8553010963	-2,6658858354	-4,0773801044
ODC6	-44,2255839090	-47,9386370678	-58,4282263864	-58,6170888501

Taulukko 11: Gamma-malli projektin edetessä: uuden mallin ennustepoikkeama vanhasta mallista PED-tunnusluvulla mitattuna.

Projekti	T3	T6	T9	T12
ODC1	-6,2108655373	-11,2380459171	-19,0065210893	-25,4233211383
ODC2	0,1272891167	-12,4177805421	-18,0921782435	-16,3003292116
ODC3	3,9832334430	3,6968410223	4,3569689560	-5,3604706851
ODC5	-0,8447654558	-0,3216651620	-0,6772397526	1,6120195425
ODC6	-42,5029609444	-49,4152845173	-57,8475169174	-57,0914308819

5 Yhteenveto

Matemaattisia malleja käytettäessä puutteiden lukumäärää estimoitaessa oletetaan aina, että ohjelmistoprojekteissa havaittujen puutteiden lukumäärät noudattavat mallien perustana olevia todennäköisyysjakaumia. Jakauman muotoon on mahdollista vaikuttaa jonkin verran parametrien avulla, kuten tässä tutkimuksessa tehtiin Gamma-jakauman α -parametrin osalta.

Sovellettaessa Norden/Rayleigh-jakaumaa ja Gamma-jakaumaa samaan aineistoon, havaittiin, että ne tuottavat lähes ekvivalentit tulokset. Ongelmana on se, että ohjelmistoprojektit ovat usein erilaisia, jolloin sama jakauma ei välttämättä sovellu eri projekteille estimointimallin perustaksi. Ohjelmistoprojekteista riippuen erot todellisten puutetietojen ja mallin

ennustamien puutetietojen välillä saattavat tulla suuriksi. Mallia on kuitenkin mahdollista parantaa projektin aikana päivittämällä mallia dynaamisesti projektin edetessä.

Viitteet:

Fenton N.E., Neil M.A. 1999. Critique of Software Defect Prediction Models, *IEEE Transactions on Software Engineering*, 25(5), 675-689.

Florac W.A. 1992. *Software Quality Measurement: A Framework for Counting Problems and Defects*, Technical Report CMU/SE-92-TR-22, Software Engineering Institute, USA.

Kan S.H. 1994. *Metrics and models in Software Quality Engineering*. Addison-Wesley Publishing Company. ISBN 0-201-63339-6.

Kitchenham B., Linkman S. 1997. Estimates, Uncertainty, and Risk, *IEEE Software*, May/June, 69-74.

Londeix B. 1987. *Cost Estimation for Software Development*, Addison-Wesley Publishing Company. ISBN 0-201-17451-0.

Lyu M.R. (toim). 1995. *Handbook of Software Reliability Engineering*, The McGraw-Hill Companies, Inc. ISBN 0-07-039400-8.

Musa J.D., Iannino A., Okumoto K. 1987. *Software Reliability Measurement, Prediction, Application*, McGraw-Hill Book Company, USA. ISBN 0-07-044093-X.

Niemi M.A. 2002. *PUTTE – Puutteiden estimointijärjestelmä*, Joensuun yliopisto, Tietojenkäsittelytieteen laitos (tekeillä).

Pillai K., Nair S.V.S. 1997. A Model for Software Development Effort and Cost Estimation, *IEEE Transactions on Software Engineering*, 23(8), 485-497.

Putnam L.H. 1978. A General Empirical Solution to the Macro Software Sizing and Estimation Problem, *IEEE Transactions on Software Engineering*, 4, July, 345-361.

Spiegel M.R. 1961. *Theory and Problems of Statistic*, McGraw-Hill, New York.

Tuominen P., Norlamo P. 1983. *Todennäköisyyslaskenta*, Limes Ry, Helsinki. ISBN 951-745-080- X.