JOENSUUN YLIOPISTO
TIETOJENKÄSITTELYTIETEEN LAITOS
Raporttisarja A

Puutteiden lukumäärän estimointi toisen asteen polynomin avulla

Matti Niemi

Report A-2003-5

ACM D.2.
ISSN 0789-7316
ISBN 952-458-406-9

Puutteiden lukumäärän estimointi toisen asteen polynomin avulla

Matti Niemi

Tietojenkäsittelytieteen laitos
Joensuun yliopisto
PL 111, 80101 Joensuu

Tiivistelmä

Ohjelmistotuotannossa voidaan estimoida puutteiden lukumäärätietoja erilaisilla matemaattisilla malleilla. Tutkimusraportissa sovitetaan toisen asteen polynomin mukainen käyrä regressioanalyysin avulla havaintoaineistoon. Aineistona on kirjallisuudesta saadut, todellisissa ohjelmistoprojekteissa kerätyt puutetiedot. Saatua mallia arvioidaan F-testin avulla sekä verrataan aiemmin saatuihin tutkimustuloksiin. Lisäksi selvitetään, kuinka olemassaolevaa perusmallia voidaan muuttaa ottamalla huomioon käynnissä olevan projektin puutetiedot eli muodostamalla dynaaminen malli. Tulokset osoittavat polynomisen mallin soveltuvan melko huonosti tässä tutkimuksessa käytettyyn puuteaineistoon. Aiemman tutkimuksen perusteella saadut tulokset Norden/Rayleigh-jakaumalla ovat parempia dynaamisen mallin osalta.

Avainsanat: toisen asteen polynominen malli, regressioanalyysi, F-testi, puutteiden lukumäärän estimointi

1 Johdanto

Tutkimuksen lähtökohtana on tutkia erilaisten laskentamallien sopivuutta ohjelmistotuotannossa ilmenevien puutteiden mallintamiseen. Mallinnus jakaantuu perusmallinnukseen ja dynaamiseen mallinnukseen Niemen (2002) mukaisesti.

Toisen asteen polynominen malli perustuu tilastotieteellisen tutkimusmenetelmän, regressioanalyysin, matemaattiseen rakenteeseen. Regressioanalyysin (Tilastokeskus, 2003) avulla voidaan ennustaa yhden tai useamman muuttujan vaikutusta johonkin muuhun muuttujaan. Regressioanalyysissa ensimmäinen askel on yhden tai useamman muuttujan nimeäminen riippumattomaksi ja yhden muuttujan nimeäminen riippuvaksi muuttujaksi. Regressiomallin käyttö perustuu olettamukseen syy-seuraussuhteesta riippumattomien ja riippuvan muuttujan välillä. Riippumattomat muuttujat selittävät riippuvan muuttujan vaihtelua. Riippuvan muuttujan arvot vaihtelevat riippumattomien muuttujien arvojen mukaan. Kun riippuva ja riippumattomat muuttujat on nimetty, voidaan laskea regressiosuora tai -käyrä. Sen avulla voidaan ennustaa riippumattomien muuttujien muutosten vaikutus riippuvassa muuttujassa.

Kahden muuttujan lineaarinen regressio on muotoa $y = a + bx$ ja toisen asteen polynominen regressio on muotoa $y = a + b_1x + b_2x^2$. Kerroin b_i kuvaa muuttujien välistä muuntosuhdetta: kuinka paljon y muuttuu, jos x muuttuu yhden yksikön verran; termi a puolestaan ilmoittaa, missä kohdassa suora tai käyrä leikkaa y-akselin.

Kun havaintopisteet asetetaan koordinaatistoon, ne eivät yleensä sijaitse samalla suoralla tai käyrällä satunnaisvaihtelun vuoksi, vaan hajaantuvat sen ympärille. Matemaattinen ratkaisu regressiosuoralle perustuu pienimmän neliösumman menetelmälle (Spiegel, 1961; Milton ja Arnold, 1995; Niemi, 2001). Sen avulla löydetään regressiosuora tai -käyrä, joka on keskimääräisesti kaikkein lähimpänä havaintopisteitä.

Tämä tutkimus on jatkoa Niemen (2001) esittämälle mallinnukselle. Tutkimuksissa käytetään samaa aineistoa, johon nyt sovelletaan regressioanalyysia toisen asteen polynomin avulla[1]. Luvussa 2 esitetään yleinen polynominen malli ja sitä vastaavien normaaliyhtälöiden ratkaiseminen matriisilaskennan avulla. Esimerkin avulla esitetään mallinnus toisen asteen polynomin avulla. Luvussa 3 esitetään, kuinka mallia analysoidaan F-testin avulla. Luvussa 4 toisen asteen polynomi sovitetaan kirjallisuudesta saatuun aineistoon perusmallin ja dynaamisen mallin muodostamiseksi. Saatuja tuloksia arvioidaan ja verrataan Niemen (2001) saamiin tuloksiin.

2 Polynominen malli

2.1 Yleinen malli

Yleinen polynominen *p*-asteinen regressiomalli ilmaisee riippuvan muuttujan *Y* (havaintopiste) odotusarvon riippumattoman muuttujan *X* (selittäjä) polynomisena funktiona (Puntanen, 1999; Draper ja Smith, 1981; Sen ja Srivastava, 1990). Tämä voidaan ilmaista muodossa (Milton ja Arnold, 1995)

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p, \tag{1}$$

jossa p on positiivinen kokonaisluku ja suurempi kuin yksi.

Asettamalla $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_4 = x^4$,…, $x_p = x^p$ malli voidaan uudelleen kirjoittaa yleisenä lineaarisena mallina

$$\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{2}$$

Parametreja β_0, β_1, β_2,..., β_p estimoitaessa pienimmän neliösumman menetelmällä muodostetaan polynominen malli

$$Y \mid x = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + E, \tag{3}$$

jossa $Y|x$ on riippuva muuttuja, kun riippumattoman muuttujan arvo on x ja E on *satunnaisvirhe* (jäännös, residuaali) (Milton ja Arnold, 1995; Puntanen, 1999), joka on muuttujan $Y|x$ ja sen odotusarvon $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p$ välinen poikkeama.

Otoksen, jonka koko on *n*, havaintopisteet ovat muotoa

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + E_i, \tag{4}$$

[1] Tutkimuksessa hyödynnetään SPSS 10.1 for Windows -ohjelmaa.

kun $i = 1, 2, \ldots, n$.

Vastaava mallin käyrä saadaan

$$\hat{y} = \hat{\mu}_{Y|x} = b_0 + b_1 x + b_2 x^2 + \ldots + b_p x^p , \tag{5}$$

jossa $\hat{y}$ on *ennustearvo* ja b_0, b_1, b_2,..., b_p ovat parametrien β_0, β_1, β_2,..., β_p pienimmän neliösumman estimaatteja. Jotta estimaattiarvot saadaan, täytyy kaavan (6) esittämä neliöiden summa eli *jäännösneliösumma SSE* minimoida.

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \left(b_0 + b_1 x_i + b_2 x_i^2 + \ldots + b_p x_i^p \right) \right]^2 , \tag{6}$$

Residuaali e_i tarkoittaa havaintopisteen y_i ja estimaatin $\hat{y}_i$ välistä eroa. Minimointi tuottaa normaaliyhtälöt:

$$\begin{aligned}
& nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 + \ldots + b_p \sum_{i=1}^{n} x_i^p = \sum_{i=1}^{n} y_i \\
& b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 + \ldots + b_p \sum_{i=1}^{n} x_i^{p+1} = \sum_{i=1}^{n} x_i y_i \\
& \quad . \\
& \quad . \\
& \quad . \\
& b_0 \sum_{i=1}^{n} x_i^p + b_1 \sum_{i=1}^{n} x_i^{p+1} + b_2 \sum_{i=1}^{n} x_i^{p+2} + \ldots + b_p \sum_{i=1}^{n} x_i^{p_i^{2p}} = \sum_{i=1}^{n} x_i^p y_i .
\end{aligned} \tag{7}$$

Tässä tutkimuksessa tarkasteltavan toisen asteen polynominen malli saadaan kaavasta (1) ja voidaan kirjoittaa muotoon:

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{8}$$

Mallia (8) vastaavat normaaliyhtälöt voidaan johtaa yhtälöistä (7) ja kirjoittaa muotoon:

$$\begin{aligned}
& nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\
& b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \\
& b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i
\end{aligned} \tag{9}$$

2.2 Matriisiesitys

Normaaliyhtälöiden ratkaisemisessa voidaan hyödyntää matriisilaskentaa. Tällöin on laadittava sopiva *mallimatriisi X* (model specification matrix) ja *havaintovektori Y* (observed responses vector) (Puntanen, 1999; Milton ja Arnold, 1995), jotka perustuvat yleiseen kaavaan (3). Mallia määrittävät yhtälöt ovat:

$$\begin{aligned} \mathrm{Y}_1 &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots + \beta_p x_1^p + E_1 \\ \mathrm{Y}_2 &= \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \ldots + \beta_p x_2^p + E_2 \\ &\vdots \\ \mathrm{Y}_n &= \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \ldots + \beta_p x_n^p + E_n \end{aligned} \qquad (10)$$

Yhtälöistä (10) voidaan päätellä mallimatriisi *X*, jonka kunkin rivin ensimmäinen alkio on ykkönen, ja havaintovektori *Y* seuraavanlaisiksi (Milton ja Arnold, 1995):

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ . & & & & \\ . & & & & \\ . & & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{bmatrix} \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}$$

Normaaliyhtälöiden matriisiesityksen löytämiseksi tarkastellaan matriisia $X'X$, jossa X' on mallimatriisin *X* transpoosi. Näin saadaan

$$X'X = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \\ x_1^2 & x_2^2 & \ldots & x_n^2 \\ . & & & \\ . & & & \\ . & & & \\ x_1^p & x_2^p & \ldots & x_n^p \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ 1 & x_3 & x_3^2 & \ldots & x_3^p \\ . & & & & \\ . & & & & \\ . & & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \ldots & \sum_{i=1}^n x_i^p \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \ldots & \sum_{i=1}^n x_i^{p+1} \\ . & & & & \\ . & & & & \\ . & & & & \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^{p+1} & \sum_{i=1}^n x_i^{p+2} & \ldots & \sum_{i=1}^n x_i^{p+2} \end{bmatrix}$$

Matriisiesityksen löytämiseksi mallin parametrien β_0, β_1, β_2,..., β_k pienimmän neliösumman estimaatit ratkaistaan yhtälöllä:

$$\left(X'X\right)b = X'y \tag{11}$$

Yhtälöstä (11) saadaan pienimmän neliösumman estimaatit laskemalla:

$$\hat{\beta} = b = \left(X'X\right)^{-1}X'y \tag{12}$$

2.3 Esimerkki

Toisen asteen mallin ratkaisua havainnollistetaan käyttäen taulukon 1 esimerkkiaineiston (Milton ja Arnold, 1995) havaintoarvoja.

Taulukko 1: Esimerkkiaineisto.

Havaintokohta x	Havainnon arvo y
5	14,0
5	12,5
10	7,0
10	5,0
15	2,1
15	1,8
20	6,2
20	4,9
25	13,2
25	14,6

Taulukon 1 arvoilla saadaan normaaliyhtälöiden (9) alkioiksi:

$$n = 10 \qquad \sum x^2 = 2750 \qquad \sum x^4 = 1223750 \qquad \sum xy = 1228$$

$$\sum x = 150 \qquad \sum x^3 = 56250 \qquad \sum y = 81{,}3 \qquad \sum x^2 y = 24555$$

jotka sijoitettuna antavat seuraavat normaaliyhtälöt

$$10b_0 + 150b_1 + 2750b_2 = 81{,}3$$

$$150b_0 + 2750b_1 + 56250b_2 = 1228$$

$$2750b_0 + 56250b_1 + 1223750b_2 = 24555$$

Näin toisen asteen polynomisen mallin käyrän kertoimiksi muodostuvat

$$b_0 = 27{,}3\,,\ b_1 = -3{,}313 \text{ ja } b_2 = 0{,}111$$

jotka sijoittamalla kaavaan (5) saadaan toisen asteen käyrän yhtälö esimerkkitapauksessa:

$$\hat{y} = 27{,}3 + \left(-3{,}3123x\right) + 0{,}111x^2$$

Matriisilaskentaa soveltaen taulukon 1 esimerkki voidaan ratkaista[2]:

$$X = \begin{bmatrix} 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 10 & 100 \\ 1 & 10 & 100 \\ 1 & 15 & 225 \\ 1 & 15 & 225 \\ 1 & 20 & 400 \\ 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 25 & 625 \end{bmatrix} \qquad y = \begin{bmatrix} 14,0 \\ 12,5 \\ 7,0 \\ 5,0 \\ 2,1 \\ 1,8 \\ 6,2 \\ 4,9 \\ 13,2 \\ 14,6 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 10 & 150 & 2750 \\ 150 & 2750 & 56250 \\ 2750 & 56250 & 1223750 \end{bmatrix}$$

jolloin kertoimiksi saadaan

$$b = \left(X'X\right)^{-1} X'y = \begin{bmatrix} 27,3000 \\ -3,3130 \\ 0,1110 \end{bmatrix},$$

jotka ovat samat kuin normaaliyhtälöiden ratkaisuina saadut kertoimet.

3 Mallin analysointi F-testillä

Toisen asteen polynomisen mallin käyrä voidaan kirjoittaa yhtälön (5) perusteella muotoon

$$\hat{y} = b_0 + b_1 x + b_2 x^2 \tag{13}$$

Tämän käyrän mukaisen mallin sopivuutta aineistoon voidaan testata esimerkiksi F-testillä tai t-testillä (Milton ja Arnold, 1995). Kummatkin testimallit antavat saman tuloksen yhden selittäjän mallissa (Draper ja Smith, 1981; Korpela, 2002). Koska tässä tutkimuksessa tarvitaan vain yhden selittäjän mallia, käydään seuraavassa läpi vain F-testi, joka perustuu tilastotieteen huomattavan kehittäjän R. A. Fisherin mukaan nimettyyn teoreettiseen jakaumaan (Tilastokeskus, 2003).

Kunkin havaintopisteen y_i poikkeama havaintopisteiden keskiarvosta $\bar{y}$ voidaan ilmaista muodossa

[2] Ratkaisu on suoritettu Matlab v. 6.1-ohjelmistolla.

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \tag{14}$$

josta saadaan

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{15}$$

Yhtälö (15) voidaan tulkita: kokonaisvaihtelusumma *SST* = jäännösneliösumma *SSE* + selitettyneliösumma *SSR*.

Näin mallin selittämä osuus voidaan laskea

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{16}$$

Tästä R^2-arvosta käytetään nimitystä *mallin selitysaste*[3].

Mitä enemmän R^2 poikkeaa nollasta, sitä parempi kyseinen malli on eli sitä paremmin x selittää y:tä. Jos $R^2 = 1$, x:llä voidaan täysin ennustaa y:tä, eli kaikki otoksen sijaitsevat mallin käyrällä. Mallin selitysaste paranee, kun selittävien muuttujien joukkoon lisätään mikä hyvänsä uusi muuttuja. Mallin selitysastetta voi siis kasvattaa lisäämällä malliin tarpeeksi paljon 'turhia' muuttujia. Tästä johtuen mallien selitysasteita ei voi verrata keskenään, kun malleissa on eri määrä selittäviä muuttujia.

Selitysasteen R^2 sijasta onkin usein parempi tarkastella *korjattua selitysastetta* R_a^2 (Adjusted R Square, Adjusted R^2), koska tämä huomioi selittävien muuttujien määrän. Korjaus on tehty ottaen huomioon mallissa olevien selittäjien sekä havaintojen lukumäärä. Korjattuja selitysasteita voi vertailla keskenään, mutta korjattu selitysaste voi laskea, jos selitettäväksi muuttujaksi lisätään huonosti selittävä muuttuja. Korjattu selitysaste R_a^2 voidaan laskea (Sen ja Srivastava, 1990) seuraavasti:

$$R_a^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n - v - 1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2 / (n - 1)} \tag{17}$$

[3] Selitysaste kertoo, kuinka monta prosenttia malli selittää riippuvan muuttajan vaihtelusta. Ennustamisessa vaaditaan erittäin korkeata selitysastetta, vähintään luokkaa 0,6 (Mauranen et al., 1993). Yhden selittävän muuttujan tapauksessa selitysaste on yhteneväinen korrelaatiokertoimen neliön kanssa: $R^2 = r_{x.y}^{\ 2}$.

Mallin hyvyyttä testattaessa F-testillä voidaan käyttää vakiintuneena esittämistapana ANOVA–taulukkoa.

Taulukko 2: ANOVA–taulukko (Draper ja Smith, 1981; Korpela, 2002).

Vaihtelulähde	Neliösumma	Vapausaste[4]	Varianssiestimaatti	*F*-arvo
Selitetty	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	v	$MSR = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{v}$	$F = \frac{SSE}{s_{x.y}^2}$
Jäännös	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	n-v-1	$s_{x.y}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - v - 1}$	
Kokonais	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	n-1		

Tutkimuksen hypoteesin testauksen yleisenä ideana on, että muotoillaan hypoteesi, joka on vastoin alkuperäistä oletusta ja sen jälkeen tutkitaan, voidaanko tämä hypoteesi kumota empiirisen aineiston perusteella. Tätä alkuperäisen oletuksen vastaista hypoteesia kutsutaan nimellä *nollahypoteesi* (null hypothesis). Nollahypoteesia on tapana merkitä H_0. Nollahypoteesin lisäksi tarvitaan *vastahypoteesi* (alternative hypothesis), joka hyväksytään, jos nollahypoteesi pystytään kumoamaan. Tätä hypoteesia merkitään H_1 (Milton ja Arnold, 1995).

F-testin hypoteesit ovat (Korpela, 2002; Milton ja Arnold, 1995):

$H_0 : R^2 = 0$ eli x ei selitä y:n vaihtelua eli $\beta_1 = \beta_2 = 0$,
$H_1 : R^2 > 0$ eli x selittää y:n vaihtelua eli $\beta_i \neq 0$ ainakin yhdelle i, kun $i = 1, 2$.

Hypoteesien testaamiseksi lasketaan F-jakauman arvo $\frac{SSR / v}{SSE / (n - v - 1)}$. Jos F-arvoksi saadaan suuri arvo suhteessa F-jakauman raja-arvoon ja p-arvoksi[5] pieni arvo suhteessa merkitsevyystasoon α, H_0 hylätään ja todetaan x:n selittävän y:n vaihtelua. Vastaavasti, jos F-arvo on pieni ja p-arvo on iso, niin selitetyn määrän voidaan katsoa mahdollisesti syntyneen sattumasta johtuen.

[4] Vapausasteella tarkoitetaan nk. vapaiden havaintojen lukumäärää. Tilastollisen tunnusluvun vapausasteiden lukumäärä on havaintojen lukumäärä n ja varatut vapausasteet v on tunnusluvun estimoimista varten aineistosta laskettujen parametrien lukumäärä (Puranen, 2003). Esimerkiksi kaavaa (13) vastaavalle mallille $v = 2$, koska x ja x^2 tulkitaan eri parametreiksi kaavan (2) mukaisesti.

[5] Jokaisen tilastollisen testin tuloksena saadaan ns. p-arvo, joka on pienin taso, joksi α voitaisiin asettaa H_0-hypoteesin hylkäämiseksi. Jos $p \leq \alpha$, niin H_0-hypoteesi voidaan hylätä merkitsevyystasolla α (Milton ja Arnold, 1995).

F-arvot on taulukoitu ainakin luottamusväleille[6] 90 %, 95 % ja 99 %.

Taulukko 3: Malli F-jakauman taulukosta $P(F_{v,n-v-1} \leq f) = 0{,}95$ (Draper ja Smith, 1981; Milton ja Arnold, 1995).

n-v-1 \ v	1	2	3	...	60	...	120
1	161,448	199,500	215,707	...	252,191	...	253,252
2	18,513	19,000	19,164	...	19,478	...	19,487
3	10,128	9,552	9,277	...	19,463	...	19,464
...	...	...	...	...	...	...	...
60	4,001	3,150	2,758	...	1,534	...	1,467
...	...	...	...	...	...	...	...
120	3,9301	3,072	2,6681	...	1,429	...	1,352

Taulukon 1 esimerkkiaineiston perusteella voidaan johtaa taulukon 5 ANOVA-taulukko, välivaiheena taulukon 4 laskelmat.

Taulukko 4: ANOVA-taulukon edellyttämät laskelmat.

Havaintokohta x	Odotusarvo $\hat{y}$	Keskiarvo $\bar{y}$	Selitetty $(\hat{y}_i - \bar{y})^2$	Jäännös $(y_i - \hat{y})^2$	Kokonais $(y_i - \bar{y})^2$
5	13,5100	8,1300	57,8888	1,2602	53,5538
10	5,2700	8,1300	16,3592	3,0658	11,0738
15	2,5800	8,1300	61,6050	0,8388	76,4298
20	5,4400	8,1300	14,4722	0,8692	14,1578
25	13,8500	8,1300	65,4368	0,9850	67,5658

Taulukko 5: ANOVA–taulukko.

Vaihtelulähde	Neliösumma	Vapausaste	Varianssiestimaatti	F-arvo
Selitetty	SSR=215,76200	2	MSR=107,881000	107,58897
Jäännös	SSE=7,01900	7	$s^2_{x.y} = 1{,}002714$	
Kokonais	SST=222,78100	9		

Täten F–testin perusteella hypoteesi $H_0 : R^2 = 0$ voidaan hylätä 95 %:n luottamusvälillä, koska taulukosta 5 saatava F-arvo 107,58897 on suurempi kuin raja-arvo $F_{2,7} = 4{,}737$ ja $p = 0{,}000 < 0{,}5$. Samoin kaavaa (16) soveltamalla selitysasteeksi saadaan $R^2 = \frac{215{,}76200}{222{,}78100} = 0{,}98849$. Koska selitysaste on lähes yksi, mallia voidaan käyttää myös ennustamiseen.

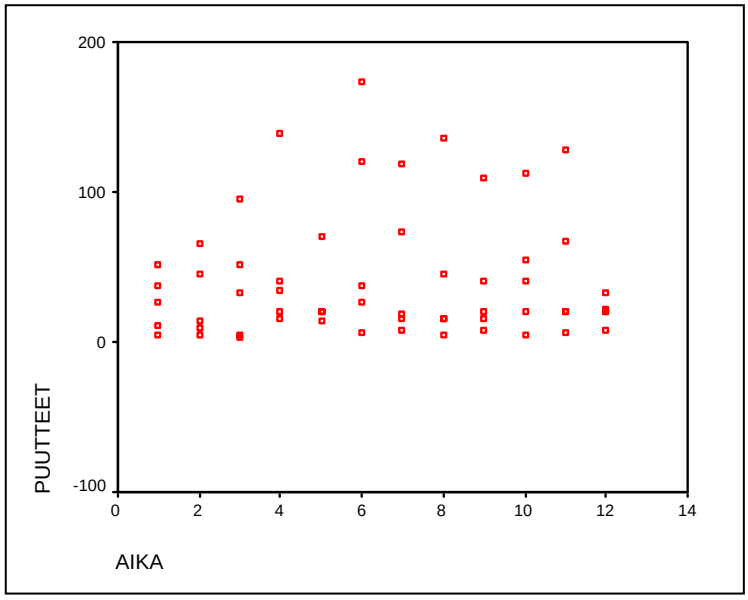
[6] 100(1-α) % luottamusväli parametrille θ on väli [L_1, L_2] siten, että P [$L_1 \leq \theta \leq L_2$] = 1-α. Yksipuoleisen F-testin tapauksessa tarkastellaan todennäköisyyttä $P(F_{v,n-v-1} \leq f) = 1 - \alpha$, missä f on F-jakauman raja-arvo (Milton ja Arnold, 1995).

4 Soveltaminen esimerkkiaineistoon

Puutteiden mallintamiseksi toisen asteen polynomin avulla havaintoaineistona käytetään taulukon 6 esimerkkiaineistoa, joka perustuu IBM Watson Research Center –tutkimuslaitoksessa erään puuteluokittelun, Orthogonal Defect Classification (ODC), soveltamista varten viidestä eri projektista hankittuihin puuteluetteloihin (Lyu, 1995). Esimerkkiaineiston hajontakuvio on esitetty kuvassa 1. Hajontakuviosta nähdään hyvin, että suhteellisen pieniä arvoja paljon, jotka hajautuvat tasaisesti eri ajankohtiin. Suuri osa havainnoista on kuitenkin satunnaisesti hajautuneita.

Taulukko 6: Esimerkkiaineisto havaituista puutteista (Niemi, 2001).

	Ajankohta											
Projekti	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



Kuva 1: Projektiaineiston puutehavaintojen hajontakuvio.

4.1 Perusmalli

Soveltamalla taulukon 6 aineistoon luvussa 3 esitettyä F-testiä saadaan taulukon 7 mukaiset arvot 95 %:n luottamusvälillä[7]. Taulukossa on esitetty *F*-arvo, raja-arvo, *p*-arvo ja selitysaste erikseen kullekin projektille ja kaikille projekteille. Taulukosta nähdään, että yksittäisillä projekteilla selitysaste on selvästi suurempi kuin nolla, mutta ainoastaan projekteja ODC1 ja ODC6 voitaisiin käyttää ennustamiseen. Ottamalla malliin mukaan kaikki projektit selitysaste on melkein nolla, eli puutteiden lukumäärää ei voida selittää tässä esimerkkitapauksessa ajan

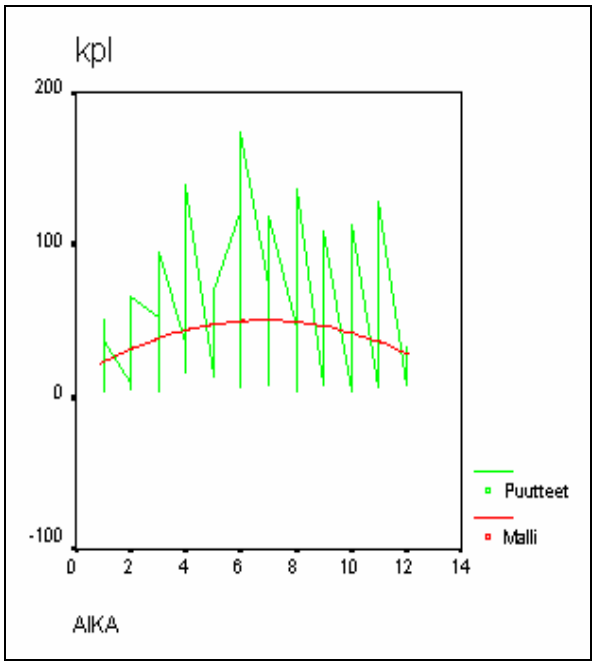
[7] Laskenta on suoritettu SPSS 10.1 for Windows –ohjelmalla.

funktiona. Taulukosta nähdään myös, että merkitsevyystasolla α = 0,05 hypoteesi H_0 hylättäisiin vain projektien ODC1, ODC2 ja ODC6 osalta.

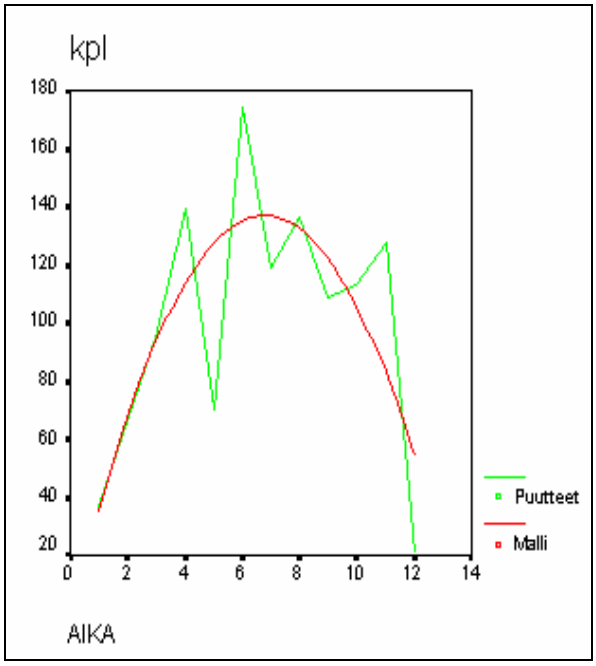
Taulukko 7. SPSS -ohjelman laskemia tunnuslukuja esimerkkiprojekteille.

Projektit / Tunnusluvut	*F*-arvo	Raja-arvo $F_{v,\,n\text{-}v\text{-}1}$	*p*-arvo	R^2
ODC1	6,30585	4,256	0,0194	0,58356
ODC2	4,43455	4,256	0,0427	0,49634
ODC3	2,01688	4,256	0,1889	0,30949
ODC5	2,81530	4,256	0,1123	0,38485
ODC6	7,21021	4,256	0,0135	0,61572
Kaikki projektit	1,39435	3,160	0,2563	0,04664

Kuvassa 2 on mallia vastaava toisen asteen polynomin mukainen käyrä esimerkkiaineistolle kaikkien projektien osalta ja kuvassa 3 on vain yhdelle projektille eli projektille ODC1.



Kuva 2. Kaikkien projektien malli.



Kuva 3. Projektin ODC1 malli.

Tulosten vertailemiseksi Niemen (2001) Norden/Rayleigh-mallin tuottamien tulosten kanssa on taulukkoon 8 laskettu Pillain ja Nairin (1997) esittämät ennustettavuuden hyvyyttä osoittavat tunnusluvut projekteittain. Verrattaessa taulukon 8 vinoutumasarakkeen arvoja vastaaviin Niemen (2001) Norden/Rayleigh-jakauman avulla saamiin arvoihin nähdään, että SPSS-ohjelmisto pystyy paremmin sijoittamaan käyrän pistejoukkoon kuin Niemen (2002) esittelemä puutteiden estimointiohjelmisto. Sen sijaan vaihtelu ja RMSPE ovat taulukossa 8 samaa suuruusluokkaa kuin Niemen (2001) Norden/Rayleigh-jakaumalle saamat arvot.

Taulukko 8: Mallin ja havaittujen puutearvojen erot.

Projekti	Vinoutuma	Vaihtelu	RMSPE
ODC1	-3,3333E -06	2,8843E+01	2,8843E+01
ODC2	0,0000E+00	1,1468E+01	1,1468E+01
ODC3	-8,3333E -07	1,3217E+01	1,3217E+01
ODC5	0,0000E+00	2,8066E+01	2,8066E+01
ODC6	1,6667E -06	4,6649E+00	4,6649E+01
Kaikki projektit	7,7716E -16	3,9937E+01	3,9937E+01

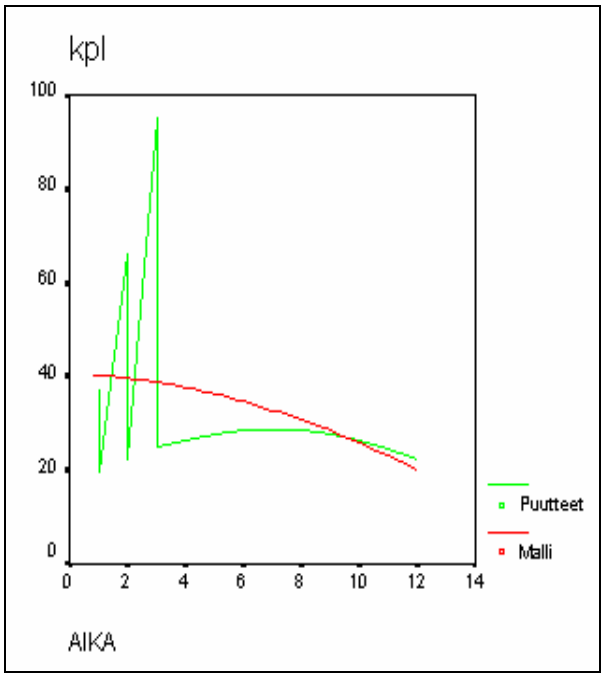
4.2 Dynaaminen malli

Dynaamisen mallin tarkoitus on Niemen (2002) mukaan hyödyntää olemassa olevan, aiempien projektien perusteella lasketun, mallin pisteitä sekä käynnissä olevasta projektista nykyhetkeen saakka kerättyjä puutetietoja. Taulukkoon 9 on laskettu kullekin taulukon 6 projektille F-testin mukaiset arvot ajanhetkillä T3, T6, T9 ja T12 ottamalla huomioon muiden projektien puuteaineiston avulla lasketun mallin pisteet ja tarkasteltavan projektin puutetiedot tarkasteltavaan ajankohtaan mennessä projektin edetessä. Taulukosta nähdään, että tyypillisesti selitysaste R^2 pienenee projektin edetessä, eli uudet puutetiedot heikentävät mallilla ennustettavuutta. Vastaavasti useimmissa projekteissa uusien puutetietojen huomioiminen mallia laskettaessa kasvattaa *p*-arvoa ja pienentää *F*-arvoa.

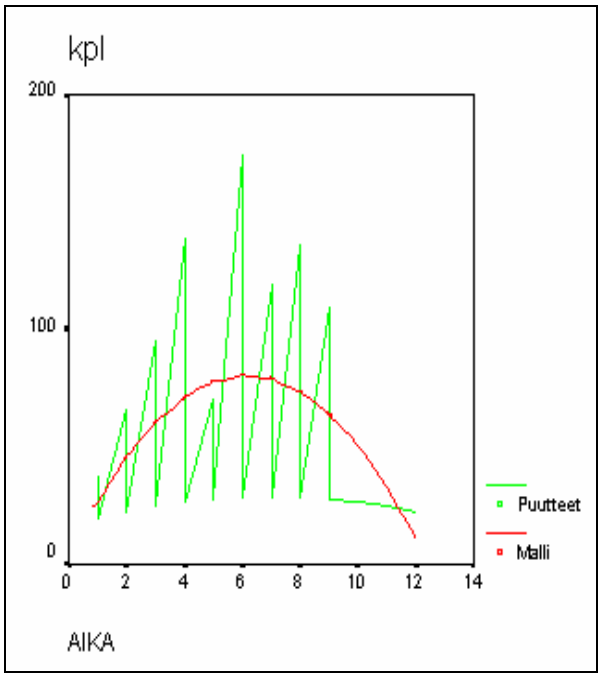
Taulukko 9. SPSS -ohjelman laskemia tunnuslukuja esimerkkiprojekteille projektin edetessä.

Projektit ajankohtana T/ Tunnusluvut	*F*-arvo	Raja-arvo $F_{v, n-v-1}$	*p*-arvo	R^2
ODC1 / T 3	0,75274	3,885	0,4921	0,11147
ODC1 / T 6	1,35278	3,682	0,2883	0,15281
ODC1 / T 9	2,03330	3,555	0,1599	0,18429
ODC1 / T 12	1,64527	3,467	0,2169	0,13547
ODC2 / T 3	3,46188	3,885	0,0650	0,36588
ODC2 / T 6	0,67205	3,682	0,5254	0,08224
ODC2 / T 9	0,04275	3,555	0,9583	0,00473
ODC2 / T 12	0,06884	3,467	0,9337	0,00651
ODC3 / T 3	32,72287	3,885	0,0000	0,84505
ODC3 / T 6	8,58739	3,682	0,0033	0,53380
ODC3 / T 9	2,58526	3,555	0,1030	0,22315
ODC3 / T 12	2,36458	3,467	0,1185	0,18381
ODC5 / T 3	13,01790	3,885	0,0010	0,68451
ODC5 / T 6	3,25715	3,682	0,0669	0,30279
ODC5 / T 9	2,70743	3,555	0,0938	0,23126
ODC5 / T 12	3,82423	3,467	0,0383	0,26698
ODC6 / T 3	10,52682	3,885	0,0023	0,63695
ODC6 / T 6	5,49985	3,682	0,0162	0,42307
ODC6 / T 9	1,93648	3,555	0,1731	0,17707
ODC6 / T 12	2,03291	3,467	0,1559	0,16221

Kuvissa 4 ja 5 on havainnollistettu projektin ODC1 vaikutusta malliin ajanhetkillä T3 ja T9.



Kuva 4. Projektin ODC1 vaikutus malliin ajankohdassa T3.



Kuva 5. Projektin ODC1 vaikutus malliin ajankohdassa T9.

Taulukossa 10 on laskettu mallin ja havaittujen puutearvojen erot projektin edetessä. Pillain ja Nairin (1997) esittämien tunnuslukujen avulla. Taulukosta nähdään, että vinoutuma on likimain nolla kullakin ajanhetkellä. Vaihtelu ja RMSPE kasvavat useimmissa tapauksissa projektin edetessä, eli vaikutus on saman suuntainen kuin SPSS-ohjelman tuottamien tunnuslukujen ilmaisema vaikutus. Nämä tulokset ovat huonommat verrattuna Niemen (2001) saamiin tuloksiin Norden/Rayleigh-jakauman avulla.

Taulukko 10: Mallin ja havaittujen puutearvojen erot. projektin edetessä.

	Vinoutuma				**Vaihtelu**				**RMSPE**			
Projekti	T3	T6	T9	T12	T3	T6	T9	T12	T3	T6	T9	T12
ODC1	1,3E-06	-5,6E-08	-9,5E-07	-3,2E-15	18,9725	40,7463	29,4983	45,7591	18,9725	40,7463	29,4983	45,7591
ODC2	1,9E-06	0,0E+00	9,5E-08	1,2E-07	13,4310	18,2863	20,3004	19,3296	13,4310	18,2863	20,3004	19,3296
ODC3	6,7E-08	-5,6E-08	-2,4E-07	-1,3E-07	5,28137	8,85540	11,8676	13,0931	5,28137	8,85540	11,8676	13,0931
ODC5	0,0E+00	-1,1E-07	4,8E-08	8,3E-08	7,61811	20,0274	21,3922	22,0920	7,61811	20,0274	21,3922	22,0920
ODC6	6,7E-08	-1,1E-07	-4,8E-08	-4,2E-08	11,9112	14,4100	17,1216	16,5280	11,9112	14,4100	17,1216	16,5280

5 Yhteenveto

Tässä tutkimuksessa käytetyn aineiston hajontakuviosta (kuva 1) voidaan havaita, että aika ja puutteiden lukumäärä eivät korreloi hyvin eli lineaarinen riippuvuus on pieni. Funktionalista toisen asteen riippuvuutta aineistossa on havaittavissa jonkin verran, mutta varsinkaan ennustamiseen aineistosta laskettua toisen asteen polynomista mallia ei voi suositella. Kuvan 1 perusteella aineisto ei myöskään sisällä sellaisia *poikkeavia havaintoja* (outlier), jotka pois-

tamalla mallia voitaisiin saada paremmaksi. Saadut tulokset ovat puutteiden mallinnuksen kannalta huonompia kuin Niemen (2001) saamat tulokset Norden/Rayleigh-jakaumalla dynaamisen mallin osalta.

Viitteet:

Draper N.R., Smith H.,1981. *Applied Regression Analysis*, John Wiley & Sons, New York, ISBN 0-471-02995-5.

Korpela E., 2002. *Regressioanalyysi –luentosarja v. 2002,* Joensuun yliopisto, Tilastotieteen laitos, Joensuu. http://joyx.joensuu.fi/~ek/regr/regr.html. (10.11.2003).

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

Mauranen K., Halonen P., Jokela V., 1993. *SPSS-opas*, Kuopion yliopisto, ATK-keskus, Kuopio.

Milton J.S., Arnold J.C., 1995. *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*, The McGraw-Hill, Inc., New York, ISBN 0-07-113535-9.

Niemi, M., 2001. *Puutteiden lukumäärän estimointi Norden/Rayleigh-jakauman ja Gamma-jakauman avulla*, Joensuun yliopisto, Tietojenkäsittelytieteen laitos, Joensuu, ISBN 952-458-084-5. (Korjattu versio saatavissa osoitteesta: ftp://ftp.cs.joensuu.fi/pub/Reports/A-2001-5.pdf)

Niemi M., 2002. *PUTTE-Puutteiden estimointijärjestelmä*, Joensuun yliopisto, Tietojenkäsittelytieteen laitos, Joensuu, ISBN 952-458-219-8.

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.

Puntanen S., 1999. *Regressioanalyysi I*, Tampereen yliopisto, Tampere, ISBN 951-44-4489-2.

Puranen, J., 2003. *Tilastotieteen sanastoa*, Helsingin yliopisto, Tilastotieteen laitos, Helsinki. http://noppa5.pc.helsinki.fi/uudet/da1htm/sanasto.html (12.11.2003).

Sen A.K., Srivastava M., 1990. *Regression Analysis: Theory, Methods, and Applications,* Springel-Verlag, New York, ISBN 0-387-97211-0.

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

Tilastokeskus, 2003. *Johdatus tilastolliseen ajatteluun*, Helsinki.
http://www.stat.fi/tk/tp/verkkokoulu/vk/tt/index.html. 3.6.2003 (10.10.2003).