Mika Antola ja Janne Koponen

Havaintoaineiston sovittaminen logistiseen yhtälöön

Matemaattinen biologia 2
seminaari
29.1.2004

Jyväskylän yliopisto
Matematiikan- ja tilastotieteen laitos

Sisältö

1 Logistinen malli

Tarkastellaan aluksi logististisen kasvuyhtälön ratkaisun

$$N(t) = \frac{K}{1 + e^{-rt}(\frac{K}{N_0} - 1)} \tag{1.1}$$

joitain ominaisuuksia.

Lause 1.1 *Olkoon* $L \neq 0 \neq K$

$$N(t) = \frac{K}{1 + e^{-rt}(\frac{K}{N_0} - 1)} \text{ ja } M(t) = \frac{L}{1 + e^{-st}(\frac{L}{M_0} - 1)} \tag{1.2}$$

logistisen mallin ratkaisuja. Tällöin

$$N(t) = M(t + C) \; \forall t \in \mathbb{R} \; \Leftrightarrow \; M_0 = \frac{K}{e^{rC}(1 + \frac{K}{N_0} - 1)} \;, s = r \text{ ja } L = K.$$

Todistus: "$N(t) = M(t + C) \Rightarrow L = K$":
Oletetaan tunnetuksi $lim_{t \to \infty} N(t) = K$. Tällöin

$$L - K = \lim_{t\to\infty} M(t) - \lim_{t\to\infty} N(t) = \lim_{t\to\infty} N(t) - \lim_{t\to\infty} N(t+C) = \lim_{t\to\infty} N(t) - \lim_{t\to\infty} N(t) = 0.$$

"$N(t) = M(t + C) \Rightarrow s = r$":
Koska $L = K \neq 0$, niin ehdosta $N(t) = M(t + C)$ seuraa

$$e^{-rt} \underbrace{(\frac{K}{N_0} - 1)}_{vakio} = e^{-s(t+C)}(\frac{K}{M_0} - 1) = e^{-st} \underbrace{e^{-sC}(\frac{K}{M_0} - 1)}_{vakio}.$$

Tämän on pädettävä kaikilla t:n arvoilla, joten ainoa vaihtoehto on, että $s = t$.

"$N(t) = M(t + C) \Rightarrow M_0 = \ldots$:"
Edellisestä saadaan ehto

$$\frac{K}{N_0} - 1 = e^{-rC}(\frac{K}{M_0} - 1).$$

Tästä saadaan suoralla laskulla

$$M_0 = \frac{K}{e^{rC}(\frac{K}{N_0} - 1) + 1}.$$

"$\Leftarrow$:" Sijoitetaan annetut M_0, L ja s $M(t + C)$:hen ja lasketaan. □

Lause 1.2 *Olkoon* L, K, N *ja* M *kuten yhtälössä (1.2). Tällöin*

$$N(t) = M(Ct) \; \forall t \in \mathbb{R} \; \Leftrightarrow \; M_0 = N_0 \;, s = \frac{r}{C} \text{ ja } L = K.$$

Todistus: "$\Rightarrow$:"
Suoria laskuja, kun $t = 0$, $t \to \infty$ ja kun sijoitetaan näissä todistetut L ja M_0 yhtälöön $N(t) = M(Ct)$.

"$\Leftarrow$:"
Sijoitetaan annetut M_0, L ja s $M(Ct)$:hen ja lasketaan. $\square$

Lause 1.3 *Olkoon L, K, N ja M kuten yhtälössä (1.2). Tällöin*

$$N(t) = CM(t)\ \forall t \in \mathbb{R} \iff M_0 = \frac{N_0}{C}\ ,\ s = r \text{ ja } L = \frac{K}{C}.$$

Todistus: "$\Rightarrow$:"
Suoria laskuja, kun $t = 0$, $t \to \infty$ ja kun sijoitetaan näissä todistetut L ja M_0 yhtälöön $N(t) = CM(t)$.

"$\Leftarrow$:"
Sijoitetaan annetut M_0, L ja s $CM(t)$:hen ja lasketaan.

$\square$

2 Käyränsovitus

2.1 Esimerkki

Käydään aluksi läpi esimerkki, jonka avulla voidaan selventää mistä käyränsovituksessa on kyse. Oletetaan että tarkkailtava prosessi käyttäytyy yhtälön

$$m(t) = at + b \tag{2.3}$$

mukaan. Tässä t on aika ja m jokin prosessista mitattu suure, esimerkiksi populaation koko. Mallissa on tuntemattomat vakiot a ja b. Mallin avulla pyritään ennustamaan prosessin käyttäytymistä, joten tarvitaan arvot a:lle ja b:lle. Oletetaan, että meillä on kokeellista tietoa järjestelmästä. Merkitään näitä havaintoja M:llä ja havaintoajankohtia T:llä. Olkoon havaintoja n kpl. Tällöin $M = [M_1, M_2, \ldots, M_n] \in \mathbb{R}^n$ ja $T = [T_1, T_2, \ldots, T_n] \in \mathbb{R}^n$.

Muodostetaan edellisestä yhtälöryhmä

$$\begin{cases} M_1 &= aT_1 + b \\ M_2 &= aT_2 + b \\ & \vdots \\ M_n &= aT_n + b, \end{cases}$$

josta yritämme ratkaista a:n ja b:n. Huomaamme kuitenkin, että jos $n > 2$, ei tällä yhtälöryhmällä ole ratkaisua kuin erikoistapauksissa. Tämän perusteella emme voi vaatia, että saatu ratkaisu kulkisi kaikkien mittapisteiden kautta.

Haluamme kuitenkin sovittaa mittausarvot malliin "mahdollisimman hyvin". Matemaattisessa mielessä tämä tarkoittaa jonkin virhefunktion minimoimista. Hyvin yleinen menetelmä on käyttää neliöllistä virhefunktiota

$$E(M, T, m) = \sqrt{\sum_{i=1}^{n} (m(T_i) - M_i)^2},$$

jolloin puhutaan pienimmän neliösumman menetelmästä (PNS-menetelmä). Tässä esimerkissä siis

$$E(M, T, a, b) = \sqrt{\sum_{i=1}^{n} (aT_i + b - M_i)^2}. \tag{2.4}$$

Pyritään nyt minimoimaan E:n arvo. On selvää, että E:n minimi on samassa pisteessä, kuin E^2 minimi. Ainoat muuttujat, joita ei ole kiinnitetty ovat a ja b. Koska E^2 on toisen asteen polynomi a:n ja b:n suhteen ja E^2:n kuvaaja aukeaa ylöspäin, pätee E^2:n minimissä

$$\begin{cases} \frac{\partial}{\partial a} E^2(M, T, a, b) = 0 \\ \frac{\partial}{\partial b} E^2(M, T, a, b) = 0. \end{cases}$$

Lasketaan edelliset osittaisderivaatat, eli

$$\begin{aligned}
0 &= \frac{\partial}{\partial a} E^2(M, T, a, b) \\
&= \frac{\partial}{\partial a} \sum_{i=1}^{n} (aT_i + b - M_i)^2 \\
&= \frac{\partial}{\partial a} \sum_{i=1}^{n} M_i^2 - 2bM_i - 2aT_iM_i + b^2 + 2abT_i + a^2T_i^2 \\
&= \sum_{i=1}^{n} -2T_iM_i + 2bT_i + 2aT_i^2
\end{aligned}$$

ja

$$\begin{aligned}
0 &= \frac{\partial}{\partial b} E^2(M, T, a, b) \\
&= \frac{\partial}{\partial b} \sum_{i=1}^{n} (aT_i + b - M_i)^2 \\
&= \frac{\partial}{\partial b} \sum_{i=1}^{n} M_i^2 - 2bM_i - 2aT_iM_i + b^2 + 2abT_i + a^2T_i^2 \\
&= \sum_{i=1}^{n} -2M_i + 2b + 2aT_i.
\end{aligned}$$

Kun jaetaan molemmat yhtälöt 2:lla, niin saadaan yhtälöpari

$$\begin{cases} \sum_{i=1}^{n} -T_i M_i + bT_i + aT_i^2 = 0 \\ \sum_{i=1}^{n} -M_i + b + aT_i = 0. \end{cases} \tag{2.5}$$

Tästä osaamme ratkaista a:n ja b:n (HT). Näin saatu suora on PNS-mielessä tutkittua prosessia parhaiten kuvaava suora.

2.2 Käyränsovitus logistisen mallin kohdalla

Logistisen mallin kohdalla saamme virhefunktion

$$E(M, T, N_0, K, r) = \sqrt{\sum_{i=1}^{n} \left(\frac{K}{1 + e^{-rT_i}(\frac{K}{N_0} - 1)} - M_i\right)^2}.$$

Käsitellään tässäkin tapauksessa tuon varsinaisen virhefunktion neliötä E^2. On siis etsittävä N_0, k ja r siten että funktio

$$E^2(M, T, N_0, K, r) = \sum_{i=1}^{n} \left(\frac{K}{1 + e^{-rT_i}(\frac{K}{N_0} - 1)} - M_i\right)^2$$

minimoituu. Vaikka tämän derivointi onnistuukin, niin derivaattojen nollakohdista on vaikea ratkaista haluttuja parametreja suljetussa muodossa. On siis turvauduttava numeeriseen laskentaan.

Esimerkeissä on optimoinnissa, eli virhefunktio E:n minimoinnissa, käytetty ns. multistart-menetelmää. Menetelmässä pyritään löytämään globaali minimi aloittamalla minimin haku satunnaisesti valituista pisteistä. Näin saaduista minimiehdokkaista valitaan sitten pienin. Lisäksi tässä yhteydessä mittausajankohdat siirretään alkamaan 0:sta ja alkuperäisen tehtävän parametrit lasketaan minimin löytämisen jälkeen lauseen 1.1 avulla. Varsinaista minimointialgoritmia ei tässä yhteydessä käydä tarkempaan läpi. Esimerkkitapaukset on laskettu R-ohjelmistolla käyttäen `optim`-funktiota 50:llä satunnaisella alkupisteellä.

Koska yliopistolla järjestetään useita optimointia käsitteleviä kursseja, ei tässä puututa optimoinnin problematiikkaan tämän tarkemmin.

3 Mallin validiteetti

Edellä saatiin määrättyä PNS-mielessä parhaat logistisen yhtälön kertoimet suhteessa tunnettuun havaintoaineistoon. Nyt voidaan sijoittaa kertoimet logistiseen kasvuyhtälöön, jolloin saadaan malli havaitun suureen käyttäytymisestä ajan funktiona. Koska mallin antamat arvot (yleensä) poikkeavat tunnetuista havainnoista niitä kutsutaan mallin antamiksi estimaateiksi. Jotta mallilla ja sen antamilla estimaateilla olisi merkitystä, pitäisi pystyä arvioimaan mallin/estimaattien luotettavuutta.

PNS-menetelmässä minimoitiin (logistisen kasvuyhtälön) virhefunktiota $E(M, T, N_0, K, r)$. Saatiin PNS-mielessä paras estimaatti, jolle voidaan laskea keskihajonta

$$\sigma = \frac{E(M, T, N_0, K, r)}{\sqrt{n-p}}, \tag{3.6}$$

missä n on tunnettujen arvojen lukumäärä ja p niiden parametrien lukumäärä, joiden suhteen virhefunktiota minimoitiin ($n - p$ on vapausasteiden lukumäärä). On selvää, että laajempi havaintoaineisto antaa pienemmän virheen.

Tämä ei kuitenkaan vielä riitä puuttuvan tai ennustettavan arvon estimaatin validiteetin arvioimiseen, koska se estimoi keskiarvoa ja todellinen havaintoaineisto sisältää aina myös satunnaiskomponentin. Tästä syystä estimaatille yritetään määrittää luottamusväli, väli jolla estimoitavan suureen arvo on annetulla todennäköisyydellä (esim. 95% eli todennäköisyydellä 0.95). Tällöin puhutaan myös estimaatin merkitsevyydestä (vastaavasti 5% tai 0.05).

Luottamusvälin määrittäminen edellyttää tietoa (tai arviota) siitä kuinka satunnaisuuden aiheuttama (sekä laskenta- ym.) poikkeama on jakautunut. Esimerkiksi, jos oletuksena on normaalijakautuneisuus, niin luottamusväli saadaan Studentin t-jakaumasta (mikäli pitkittäisaineisto on laaja, niin vastaavasti suoraan normaalijakaumasta):

$$p\text{-luottamusväli} = [\hat{y} - t_{1-\alpha/2, n-p}\hat{\sigma}, \hat{y} + t_{1-\alpha/2, n-p}\hat{\sigma}], \tag{3.7}$$

missä

$\hat{\sigma}$ $=\sigma/\sqrt{n}=$keskiarvon keskivirhe.

$\hat{y}$ =mallista laskettu estimaatti,

α =merkitsevyys,

$n-p$ =vapausasteiden lukumäärä ja

t = Studentin jakauman vastaava kerroin. [tilastotieteen perus- ja menetelmäkurssit]

Toisaalta voitaisiin tutkia funktion yleistä virhettä: yhden muuttujan tapauksessa tämä tarkoittaa sitä, että lasketaan funktion derivaatta argumentin suhteen, jolloin

$$\sigma(f) = \frac{df}{dx}\sigma(x), \tag{3.8}$$

missä $\sigma(f)$ on funktion virhe ja $\sigma(x)$ argumentin virhe.

Esimerkki
Kuution särmä on 240 mm ja epätarkkuus 1.5 mm. Mikä on kuution tilavuus ja sen epätarkkuus? (13.8 $\pm$ 0.26 dm3)

Vastaavasti useamman muuttujan funktiosta lasketaan osittaisderivaatat ja lasketaan neliösumman neliöjuuri. Näin tehdään, koska kokonaisdifferentiaalin suora soveltaminen saattaisi johtaa liian pieneen virhearvioon, jos osittaisdifferentiaalit (merkkien johdosta) sattuisivat kumoamaan toisiaan. Toisaalta niiden itseisarvojen ynnääminen veisi virheen maksimiarvioon.

$$\sigma(f) = \sqrt{\sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i}\sigma(x_i)\right]^2}, \tag{3.9}$$

Logistisen yhtälön kohdalla on siis laskettava osittaisderivaatat N_0:n, K:n ja r:n suhteen, joihin sitten sovelletaan kaavaa (3.9). Ongelmana on se, että pitäisi tietää muuttujien N_0, K ja r virhe.

4 Esimerkkejä

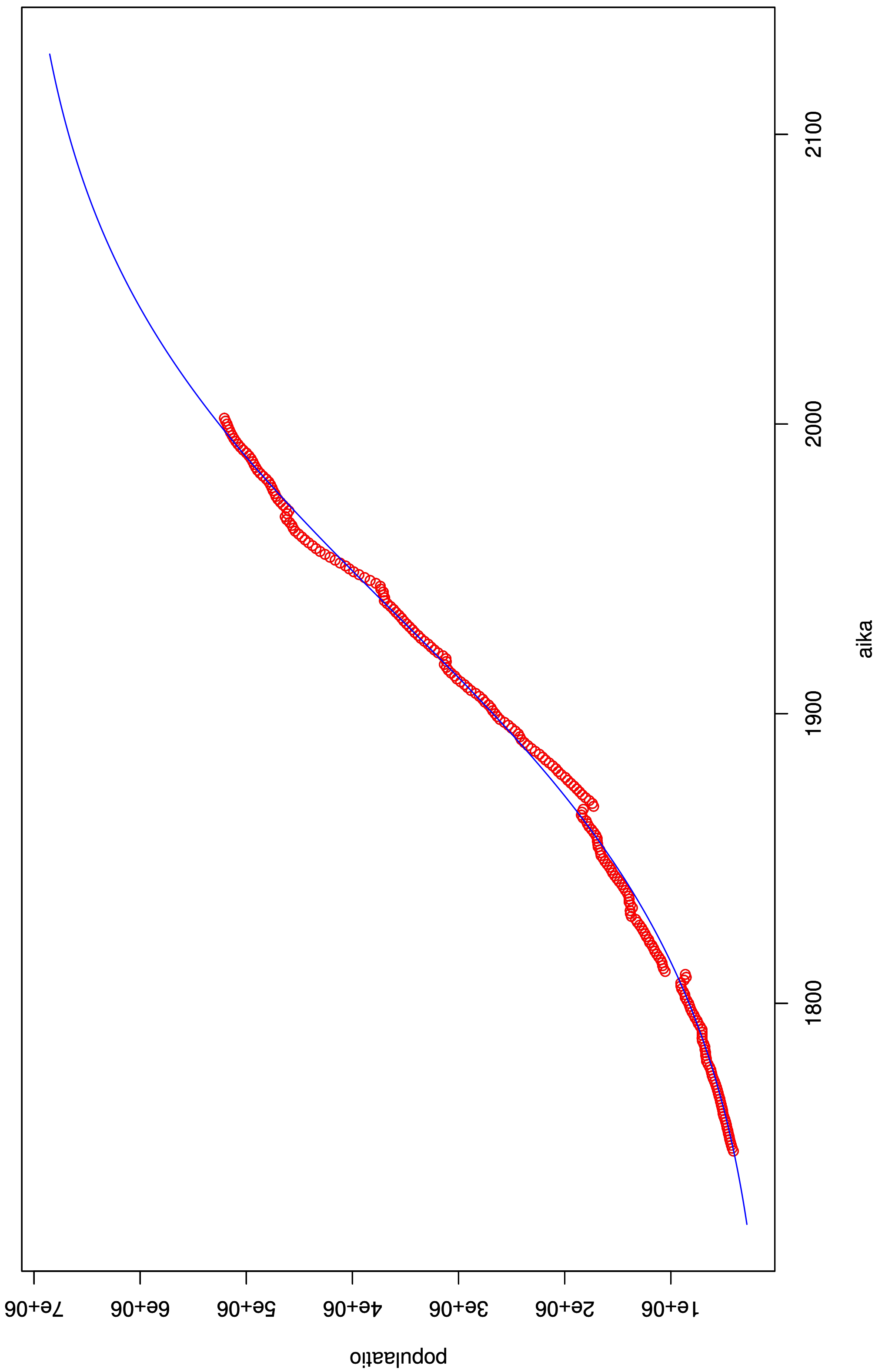
4.1 Suomen väkiluku

Suomen väkiluvun kehityksestä viimeisen noin 250 vuoden aikana löytyy tilastokeskuksen sivuilta [1]. Tätä tilastoa käyttäen voidaan sovittamaa logistista mallia Suomen väkiluvun kehitykseen. Kuvassa 4.1 on esitetty suomen väkilukuun sovitettu käyrä. Tällöin on logistisen mallin parametreiksi saatu $r = 0.015$, $K = 7.22{\cdot}10^6$ ja $N_0 = 1.38{\cdot}10^{-6}$. Vastaavasti jos sovitus tehdään vain vuosien 1900-2002 mukaan, kuva 4.3, saadaan parametreiksi $r = 0.019$, $K = 6.36 \cdot 10^6$ ja $N_0 = 1.49 \cdot 10^{-9}$. Näistä kiinnostavin havainto lienee K:n arvon pieneneminen, joka on tulkittavissa siten että Suomen kantokyky on pienentynyt! Tämän seurauksena onkin syytä miettiä mallin sopivuutta tähän tapaukseen. Suoritetaan vielä vertailua tilastojen sataan ensimmäiseen vuoteen. Tämä tapaus on esitetty kuvassa 4.2. Tällöin parametreiksi saadaan $r = 0.018$, $K = 4.59 \cdot 10^6$ ja $N_0 = 1.72 \cdot 10^{-8}$.

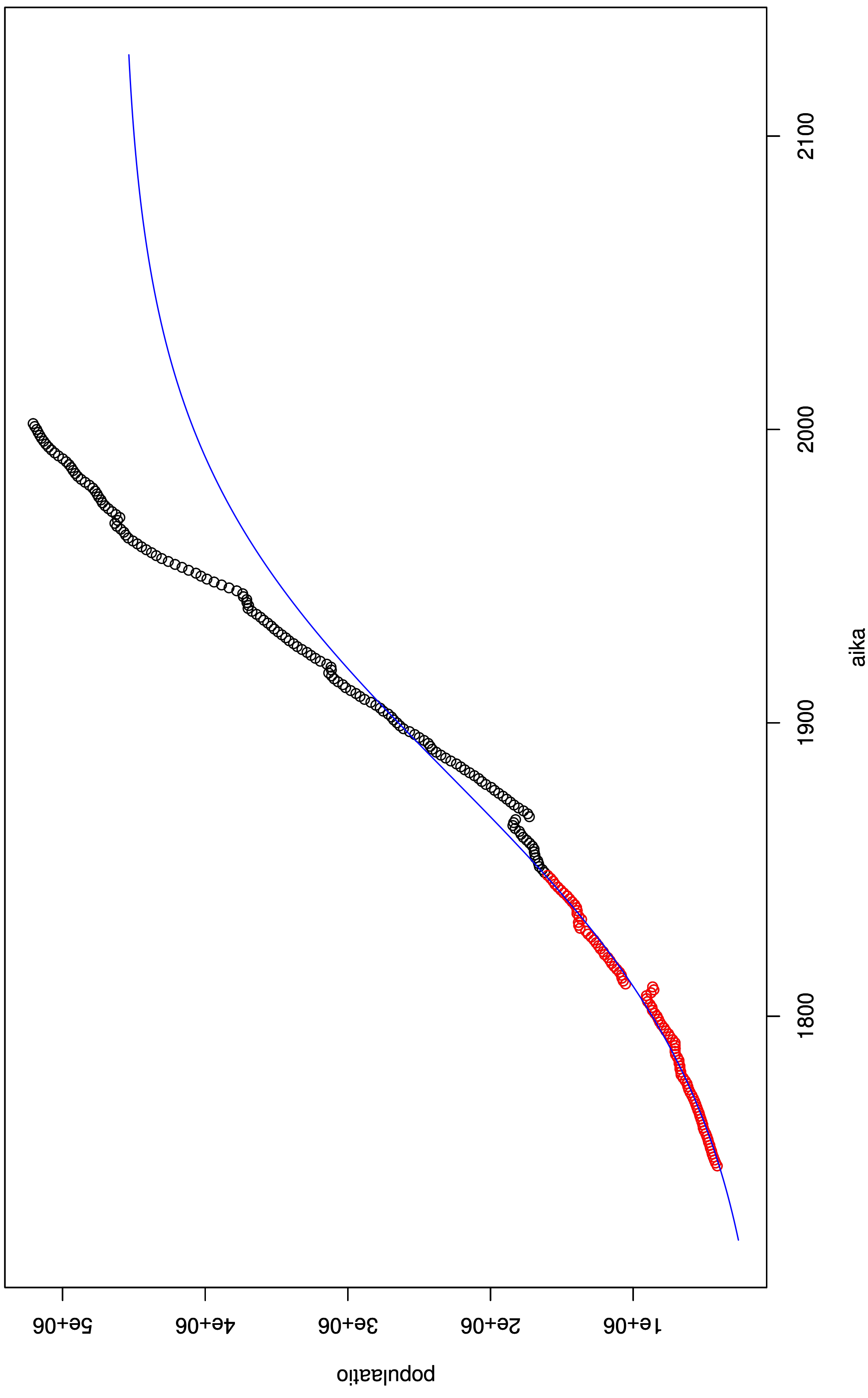
5 Havaintoaineiston suodattaminen

Mallin validiteettia edellä arvioitaessa tehtiin se oikeastaan yksittäisen, mallin avulla laskettavan estimaatin mielessä. Kuitenkin väkiluvun kaltaisia pitkittäisiä aikasarjoja tutkittaessa mielenkiinto saattaa kohdistua enemmänkin yleiseen trendiin kuin yksittäisiin estimaatteihin.

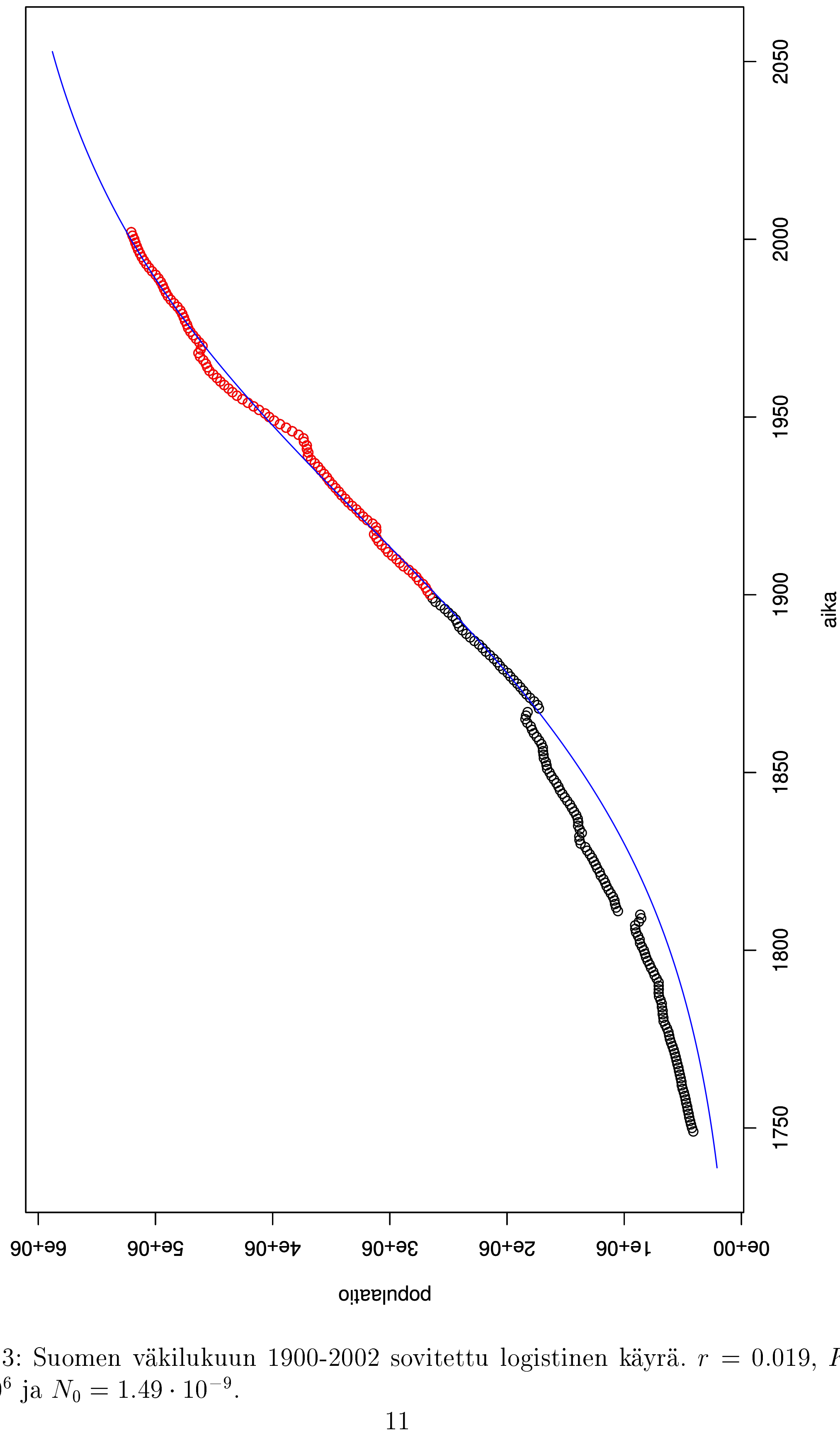
Aikasarja voidaan jakaa karkeasti neljään komponenttiin: kausivaihtelu, muu syklinen vaihtelu, trendi ja muu epäsäännöllinen vaihtelu eli kohina (esim. satunnaisvaihtelu, mittausvirheet). Vuosittaisen väkiluvun aikasarjassa kyseeseen tulee lähinnä trendi ja



Kuva 4.1: Suomen väkiluku 1749-2002 ja siihen sovitettu logistinen käyrä. $r = 0.015$, $K = 7.22 \cdot 10^6$ ja $N_0 = 1.38 \cdot 10^{-6}$.



Kuva 4.2: Suomen väkilukuun 1749-1849 sovitettu logistinen käyrä. $r = 0.018$, $K = 4.59 \cdot 10^6$ ja $N_0 = 1.72 \cdot 10^{-8}$.



Kuva 4.3: Suomen väkilukuun 1900-2002 sovitettu logistinen käyrä. $r = 0.019$, $K = 6.36 \cdot 10^6$ ja $N_0 = 1.49 \cdot 10^{-9}$.

kohina.

Mitä voitaisiin tehdä väkilukuesimerkissämme, jotta trendi saataisiin paremmin esiin ja siten käyrä sovitettua aineistoon paremmin? Ensimmäisenä mieleen tulee vaihtaa sovitusmenetelmää: neliösumman käyttö korostaa kohinan vaikutusta (varsinkin virheeseen). Helpoin muutos olisi tutkia neliön sijasta itseisarvoa (normia).

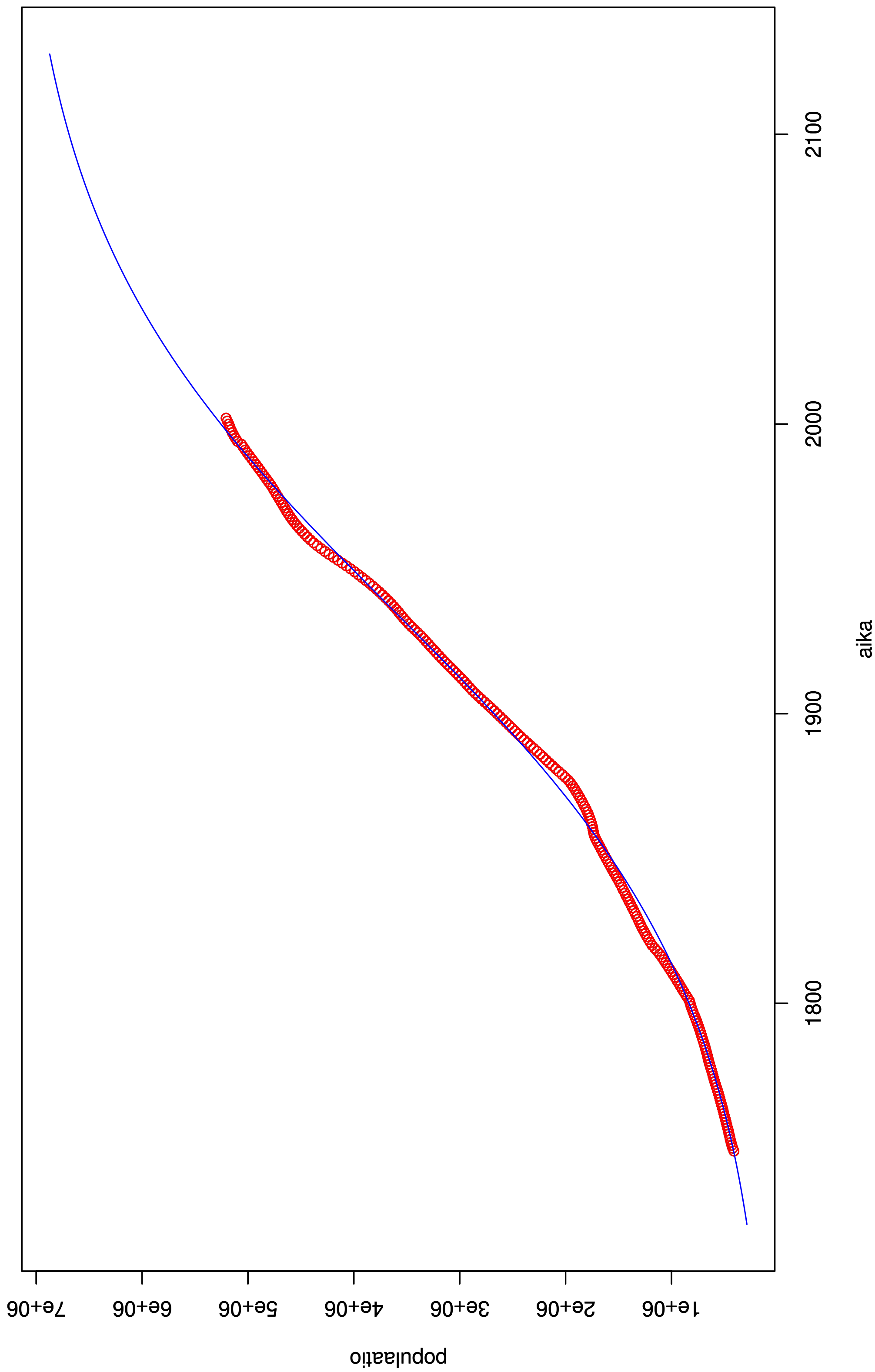
Tutkitaan kuitenkin yksinkertaista lineaarista suodatusta.

$$\hat{y}_t = \frac{1}{2q+1} \sum_{i=-q}^{q} x_{t+r} \tag{5.10}$$

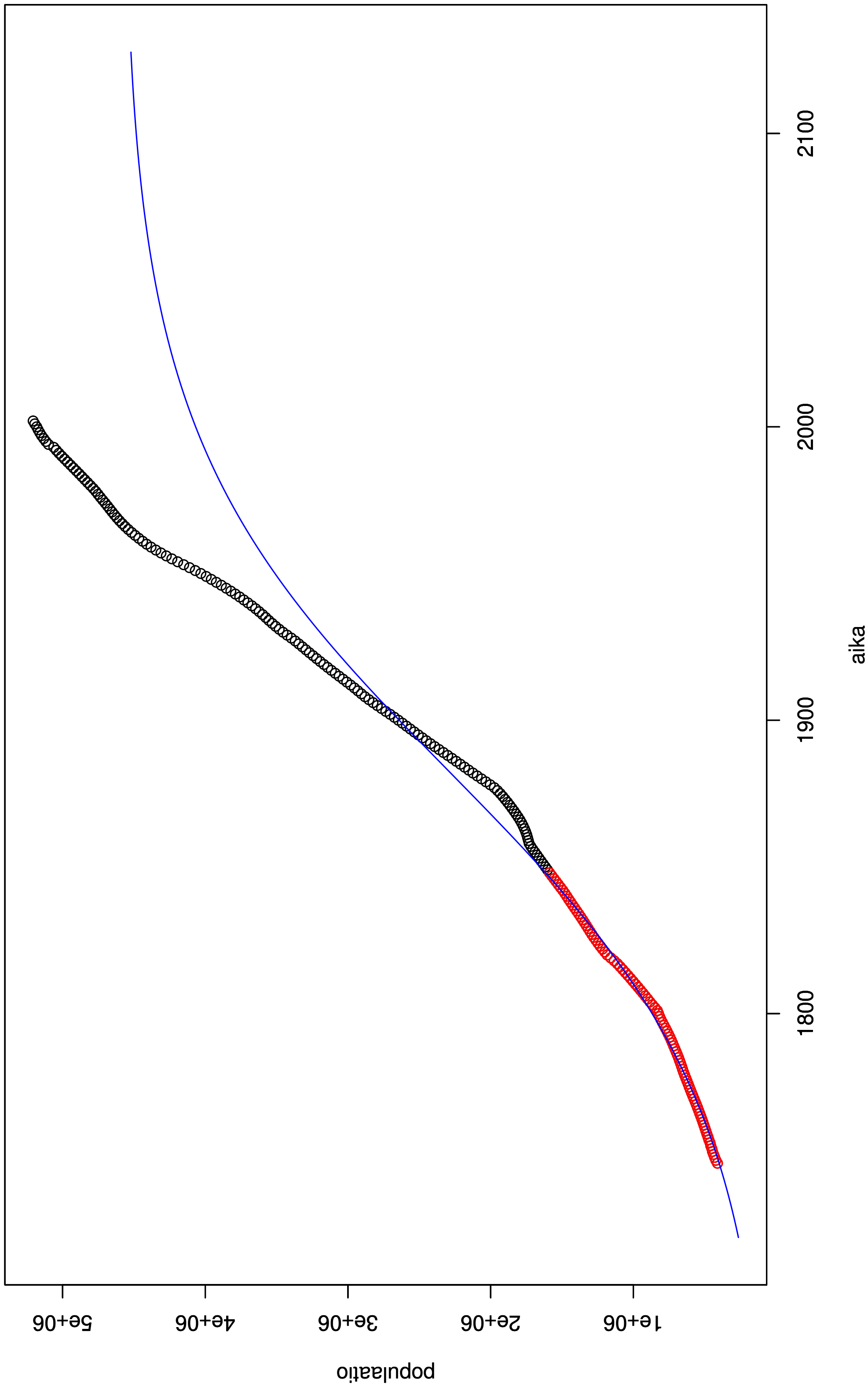
Tässä lasketaan aikasarjasta x_t liukuva keskiarvo, jossa huomioidaa x_t:n ympäriltä q kappaletta havaintoja. Tämä tasoittaa aineistosta pahimmat “piikit”. Uuden suodatetun aineiston odotusarvo on sama kuin alkuperäisellä, mutta varianssi on pienempi. Ennustamisen kannalta olisi tietysti parempi huomioida suodatuksessa vain edeltäviä havaintoja. Taulukossa **??** on esitetty suodatuksen (5.10) vaikutus virheeseen, kun $q = 10$.

aika	σ suodattamaton	σ suodatettu
1749-2002	5195	3928
1749-1849	6409	4803
1900-2002	6117	4612

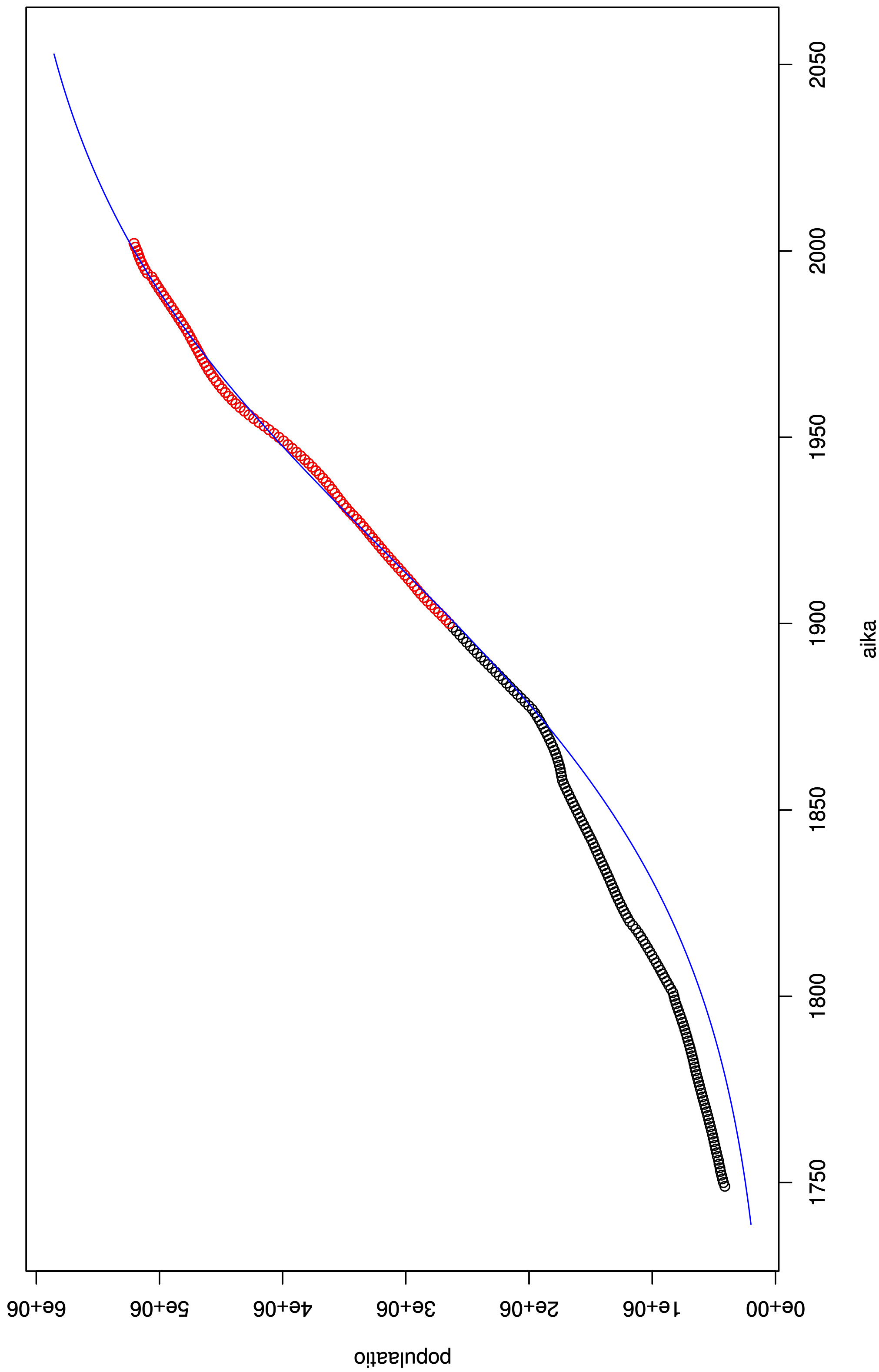
Taulukko 5.1: Lineaarisen suodatuksen vaikutus virheeseen



Kuva 5.4: Suomen suodatettuun väkilukuun 1749-2002 sovitettu logistinen käyrä. $r = 1.51 \cdot 10^{-2}$, $K = 7.21 \cdot 10^{6}$ ja $N_0 = 1.5 \cdot 10^{-6}$



Kuva 5.5: Suomen suodatettuun väkilukuun 1749-1849 sovitettu logistinen käyrä. $r = 1.76 \cdot 10^{-2}$, $K = 4.57 \cdot 10^{6}$ ja $N_0 = 1.78 \cdot 10^{-8}$



Kuva 5.6: Suomen suodatettuun väkilukuun 1900-2002 sovitettu logistinen käyrä. $r = 1.86 \cdot 10^{-2}$, $K = 6.37 \cdot 10^6$ ja $N_0 = 1.81 \cdot 10^{-9}$

Harjoituksia

1. Ratkaise a ja b yhtälöparista (2.5).

2. Ennusta m:n arvo hetkellö $t = 100$ kun oletetaan (2.3):n malli ja $M = [0, 7, 9, 14]$ sekä $T = [0, 20, 50, 80]$. Käytä virhefunktiona kaavan (2.4) määrittelemää fuktiota.

3. Osoita, että virhefunktion E minimi saa arvon nolla, jos oletetaan malli (2.3) ja virhefunktio (2.4) sekä $n = 2$ ja $T_1 \neq T_2$.

4. Tutustu johonkin numeerisen laskennan ohjelmistoon ja sen optimointimahdollisuuksiin. Ratkaise tällä ohjelmalla tehtävä 2.

5. Olkoon L, K, N ja M kuten yhtälössä (1.2). Oletetaan lisäksi, että $N(t) = M(t) + C$. Osoita, että $C = 0$.

6. Olkoon z_1, z_2, z_3 riippumattomia muuttujia $Ez_i = \mu$, $Var z_i = \delta_z^2$. $Y_2 = \frac{1}{3}(z_1 + z_2 + z_3)$, liukuvan keskiarvon suodatus. Näytä että $EY_2 = Ez_2$ ja $Var Y_2 \leq Var z_2$.

Viitteet

[1] `www.stat.fi/tk/he/vaesto_vaki1749.xls`, viitattu 17.1.2004