TAMPEREEN YLIOPISTO
Pro gradu -tutkielma

Kalle Suominen

Leen-Carterin malli

Matematiikan, tilastotieteen ja filosofian laitos
Matematiikka
Syyskuu 2006

Tampereen yliopisto

Matematiikan, tilastotieteen ja filosofian laitos

Suominen, Kalle: Leen-Carterin malli

Pro gradu -tutkielma, 44 s.

Matematiikka

Syyskuu 2006

Tiivistelmä

Vuonna 2007 tulee voimaan TyEL-eläkelaki, joka yhdistää yksityisen sektorin palkansaajat yhden eläkelain piiriin. Tässä tutkielmassa sovelletaan Leen-Carterin mallia yli 65-vuotiaiden eläkkeellä olevien TyEL-eläkejärjestelmään kuuluvien henkilö- ja rahakuolleisuuteen. Aineisto on vuosilta 1986–2004, ja siinä on eriteltynä ikä- ja vuosiluokkakohtaisesti kuolleet, elävät, kuolleiden keskimääräinen eläke ja elävien keskimääräinen eläke. Leen-Carterin mallin parametrit muodostetaan singulaariarvohajotelman, pienimmän neliösumman ja suurimman uskottavuuden menetelmillä. Leen-Carterin mallista saatujen tulosten perusteella tehdään ennuste 65-vuotiaiden elinajanodotteista vuoteen 2035 asti.

Sisältö

Johdanto

Vuonna 2007 tulee voimaan TyEL-eläkelaki, joka yhdistää yksityisen sektorin palkansaajat yhden eläkelain piiriin. Tämän tutkielman tarkoitus on soveltaa Leen-Carterin mallia yli 65-vuotiaiden eläkkeellä olevien TyEL-eläkejärjestelmään kuuluvien kuolleisuuteen. Löydetyn Leen-Carterin mallin perusteella tehdään ennuste 65-vuotiaiden TyEL:aisten elinajanodotteen kehittymisestä tulevaisuudessa. Lisäksi tehdään ennuste 65-vuotiaiden TyEL:aisten rahakuolleisuudesta. Käytännössä rahakuolleisuuden elinajanodotteella tarkoitetaan sitä aikaa, kun raha kiertää eläkejärjestelmän läpi. Tosiasia on, että suurempaa eläkettä saavat elävät yleensä kauemmin kuin pientä eläkettä saavat. Tästä syystä rahakuolleisuudesta saadut elinajanodotteet voivat olla suurempia kuin lukumääräkuolleisuudesta saadut luvut. Jatkossa teoriaosien kohdalla puhutaan pelkästään kuolleisuudesta, mutta samat ehdot ja mallit pätevät myös rahakuolleisuuteen.

Ensimmäisessä luvussa esitellään aineisto ja kerrotaan, millä tavalla kuolleisuusluvut muodostetaan. Toisessa luvussa esitellään Leen-Carterin malli ja kolme menetelmää, joilla pystytään estimoimaan mallin parametrit. Kolmannessa luvussa sovelletaan menetelmiä ja tutkitaan kuinka hyvin estimoidut mallit sopivat havaintoaineistoihin. Tämän jälkeen käydään läpi hieman aikasarja-analyysin perusteita ja tehdään ennusteet kuolleisuuksien kehittymiselle. Luvussa neljä tehdään Expost-ennuste ja katsotaan, olisiko menneisyydessä tehty ennuste toteutunut. Viimeisessä luvussa pohditaan virhetermin vaikutusta.

Pääasiallinen lähde on Marie-Claire Koissin lisensiaattityö *Fitting and Forecasting Mortality Rate with the Lee-Carter Model*. Perustana aikasarja-analyysissä on käytetty Arto Luoman kirjoittamaa Tampereen yliopiston luentorunkoa *Aikasarja-analyysi I*. Kuvat on tehty Excel- ja R-ohjelmilla ja ne on muutettu pdf-muotoon ohjelmalla WMF2EPS. Mallien estimoinnit on tehty R- ja Mathematica-ohjelmilla.

1 Aineiston esittely

Olemme kiinnostuneita yli 65-vuotiaista eläkkeellä olevista henkilöistä. Aineistona käytämme Eläketurvakeskuksen (ETK) tietokannasta saatuja matriiseja, joista näemme sekä eläkeläisten että kuolleiden lukumäärät vuosina 1986–2004. Näistä matriiseista muodostamme kuolemanvaaraluvut erikseen miehille sekä naisille. Lisäksi käytössämme on matriisit, joista näemme sekä elävien että kuolleiden keskimääräiset eläkkeet. Näiden tietojen avulla pystymme muodostamaan rahakuolleisuuden kuolemanvaaraluvut. On olemassa monia tapoja kuolemanvaaralukujen muodostamiseen. Tässä työssä käytämme yksinkertaista kaavaa

$$q_{x,t} = \frac{\text{Kuolkm}_{x,t}}{\text{Elkm}_{x,t}},$$

jossa $q_{x,t}$ tarkoittaa kuolleisuutta ja $\text{Kuolkm}_{x,t}$ niitä ihmisiä, jotka ovat kuolleet x-ikäisenä vuonna t ja vastaavasti $\text{Elkm}_{x,t}$ tarkoittaa x-ikäisenä elossa olleita vuonna t.

Rahakuolleisuudessa kuolemanvaaraluvut muodostetaan kaavalla

$$q_{x,t} = \frac{\text{KKE}_{x,t}\text{Kuolkm}_{x,t}}{\text{EKE}_{x,t}\text{Elkm}_{x,t}},$$

jossa $\text{KKE}_{x,t}$ tarkoittaa kuolleiden keskieläkettä, $\text{Kuolkm}_{x,t}$ kuolemien lukumäärää, $\text{EKE}_{x,t}$ vastaavasti elossa olevien eläkeläisten keskieläkettä ja $\text{Elkm}_{x,t}$ elossa olevien lukumäärää. Näistä kuolemanvaaraluvuista muodostamme erikseen matriisit naisten ja miesten lukumäärä- ja rahakuolleisuuksille.

2 Leen-Carterin malli

2.1 Alustus

Väestönmuutokset kiinnostavat kansantaloudellisesti monia eri tahoja. Yhteiskunnan päättäjät tarvitsevat väestöennusteita suunnitellessaan esimerkiksi erilaisten koulujen tarpeita tai sairaaloiden keskittämistä. Me tutkimme yli 65-vuotiaiden TyEL:aisten kuolleisuuden muutosta, jonka tunteminen on tärkeää eläkejärjestelmää kehittäville tahoille.

Ongelmamme on löytää mahdollisimman hyvin sopiva malli iässä x ja ajassa t tapahtuvalle kuolleisuudelle. Voisimme kirjoittaa kuolleisuusmallin muodossa

$$f(q_{x,t}) = \alpha_x + \beta_t + \epsilon_{x,t}. \tag{2.1}$$

Tässä mallissa $q_{x,t}$ on x-ikäisen kuolevuuden taso ajanhetkellä t. Mallin parametrit α_x ja β_t ovat kuolleisuuden tasot ikäluokalle x ajanhetkellä t. Virhetermissä $\epsilon_{x,t}$ on mallissa tapahtuva virhe. Yhtälössä (2.1) on kuitenkin puutteita, koska sillä ei ole yksikäsitteistä ratkaisua. Tämän voimme helposti todeta kirjoittamalla yhtälön (2.1) muotoon

$$\begin{aligned} f(q_{x,t}) &= \alpha_x + \beta_t + \epsilon_{x,t} \\ &= (\alpha_x + h) + (\beta_t - h) + \epsilon_{x,t} \\ &= \alpha_x^{'} + \beta_t^{'} + \epsilon_{x,t}. \end{aligned}$$

Näemme, että $\alpha_x + \beta_t$ ja $\alpha_x^{'} + \beta_t^{'}$ molemmat toteuttavat yhtälön. Yhtälön (2.1) kaltainen malli ei siis ole hyvä, joten meidän täytyy jollakin tavalla yhdistää toisiinsa ikämuuttuja x ja aikamuuttuja t.

2.2 Leen-Carterin malli

Leen-Carterin mallia käytetään yleisesti kuolleisuuksia tutkittaessa. Malli on suhteellisen yksinkertainen, mutta siitä huolimatta sillä voidaan mallintaa kuolleisuutta erittäin hyvin. Lisäksi voimme tehdä ennusteen tulevaisuuteen mallista saatujen parametrien avulla.

Leen-Carterin mallissa kuolleisuusmatriisi $(q_{x,t})$ kirjoitetaan muodossa

$$\ln(q_{x,t}) = a_x + b_x k_t + \epsilon_{x,t}. \tag{2.2}$$

Vektorissa a_x on havaintoaineiston x-vuotiaiden kuolleisuuden keskiarvo. Vektorissa b_x on aineiston x-vuotiaiden kuolleisuuden muuttumisen voimakkuus. Vektori k_t kuvaa kuolleisuuden vuosikohtaista tasoa verrattuna aineiston keskimääräiseen kuolleisuuden tasoon. Estimoinnissa tapahtuva virhe on virhetermissä $\epsilon_{x,t}$. Identifioituvuusehtojen ehtojen vuoksi on oltava $\sum_{x=1}^{X} b_x^2 = 1$ ja $\sum_{t=1}^{T} k_t = 0$. Identifioituvuusehdoista seuraa, että vektorissa a_x on jokaisen

ikäluokan logaritmoitu keskikuolleisuus. Tämän voimme todistaa ottamalla summan yli ajan yhtälön (2.2) molemmilta puolilta (yksinkertaistamisen vuoksi jätämme virhetermin käsittelemättä), jolloin saamme

$$\begin{aligned}\sum_{t=1}^{T} \ln(q_{x,t}) = \sum_{t=1}^{T} (\hat{a}_x + \hat{b}_x \hat{k}_t) &= \sum_{t=1}^{T} \hat{a}_x + \sum_{t=1}^{T} \hat{b}_x \hat{k}_t \\ &= T\hat{a}_x + \hat{b}_x \sum_{t=1}^{T} \hat{k}_t \\ &= T\hat{a}_x,\end{aligned}$$

joten

$$\hat{a}_x = \frac{1}{T} \sum_{t=1}^{T} \ln(q_{x,t}).$$

Yhtälöketjun toiseksi viimeinen vaihe seuraa siitä, että identifioituvuusehtojen mukaan $\sum_{t=1}^{T} \hat{k}_t = 0$.

2.2.1 Singulaariarvohajotelma

Alkuperäisessä työssään Lee ja Carter muodostivat parametrien estimaatit singulaariarvohajotelman avulla. Singulaariarvohajotelma on erittäin luonnollinen valinta, sillä sen tuloksesta saamme poimittua kaksi vektoria, joiden avulla saamme sovitettua pienimmän neliösumman sovitteen aineistoon.[2, s. 11]

Määritelmä 2.1. Olkoon matriisi A kokoa $m \times m$. Lukua $\lambda \in \mathbf{C}$ sanotaan matriisin A *ominaisarvoksi*, jos on olemassa sellainen vektori $x \neq 0$, että

$$Ax = \lambda x.$$

Ehdon $Ax = \lambda x$ toteuttavia vektoreita x sanotaan ominaisarvoa λ vastaaviksi *ominaisvektoreiksi*.

Määritelmä 2.2. Olkoon matriisi Y kokoa $m \times n$. Matriisin Y *singulaariarvot* ovat matriisin $Y^T Y$ ominaisarvojen neliöjuuret.

Määritelmä 2.3. Olkoon $m \times n$ matriisi Y astetta r. *Singulaariarvohajotelma* (SVD) matriisille Yon muotoa

$$Y = UDV^T = [U, D, V], \tag{2.3}$$

missä U, V ja D tarkoittavat seuraavia matriiseja:

- D on diagonaalimatriisi, jonka alkiot saadaan matriisin Y^TY ominaisarvojen neliöjuurista
$$\sigma_{i,i} = \sqrt{\lambda_i}.$$
- V:n sarakkeet ovat matriisin Y^TY ortonormaalit ominaisvektorit.
- U:n sarakkeet ovat matriisin YY^T ortonormaalit ominaisvektorit.

Merkitään matriisin U sarakkeita vektoreilla $u_1, u_2, \ldots, u_m$. Merkitään vektoreilla $v_1, v_2, \ldots, v_n$ matriisin V sarakkeiden transpooseja. Matriisin D singulaariarvoja merkitsemme muuttujilla $d_1, d_2, \ldots, d_r$, missä $d_1 > d_2 > \cdots > d_r$.

Esimerkki 2.1. Matriisille A kokoa $m \times n$ tehty singulaariarvohajotelma voidaan kirjoittaa muodossa

$$\begin{aligned}
\mathrm{SVD}(A) &= [U, D, V] \\
&= \begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1m} \\ u_{21} & u_{22} & \ldots & u_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \ldots & u_{mm} \end{pmatrix} \begin{pmatrix} d_1 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & d_2 & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & d_r & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \ldots & v_{1n} \\ v_{21} & v_{22} & \ldots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \ldots & v_{nn} \end{pmatrix}^T \\
&= u_1 d_1 v_1 + u_2 d_2 v_2 + \cdots + u_m d_r v_n,
\end{aligned}$$

jossa U on $m \times m$ -matriisi, D on $m \times n$ -matriisi ja V on $n \times n$ -matriisi.

Nyt voimme estimoida kaavan (2.2) parametrit seuraavasti:

1. Vektorissa $\hat{a}_x$ on logaritmoidut ikäluokkakohtaiset keskiarvot.
2. Muodostetaan matriisi $Z = (\ln(q_{x,t}) - \hat{a}_x)$, missä $x = 1, 2, \ldots, X$ ja $t = 1, 2, \ldots, T$.

3. Lasketaan singulaariarvohajotelma matriisille Z, siis $\mathrm{SVD}(Z) = [U, D, V]$.

4. Vektori b saadaan yhtälöstä $b = u_1$, jolloin $\sum_x b_x^2 = 1$.

5. Vektori k muodostetaan yhtälöstä $k = d_1 v_1$.

On otettava huomioon, että muodostamalla sovitteen kuolleisuusmatriisille edellä mainitulla tavalla käytämme hyväksemme oikeastaan vain murto-osaa singulaariarvohajotelman tuloksesta. Saisimme tarkemman estimoinnin, jos ottaisimme mukaan useamman termin. Mallia tehtäessä voisimme ottaa huomioon esimerkiksi kaksi ensimmäistä termiä ($u_1 d_1 v_1$ ja $u_2 d_2 v_2$), mutta tällöin parametrien estimointi ei olisi enää niin yksinkertaista. Vain ensimmäisen $u_1 d_1 v_1$ termin käyttämistä voimme perustella myös sillä, että siihen liittyvä singulaariarvo d_1 on huomattavasti suurempi kuin termiin $u_2 d_2 v_2$ liittyvä singulaariarvo d_2 ja näin ollen termi $u_2 d_2 v_2$ ei tuo enää merkittävää lisäarvoa estimointiin.[5, s. 3]

Todistamme seuraavaksi, että tekemällä singulaariarvohajotelman matriisille Z ja muodostamalla vektorin $\hat{k}_t$ kohdan 5 mukaisesti saamme $\sum_{t=1}^{T} \hat{k}_t = 0$.

Apulause 2.1. *Olkoon matriisi A kokoa $p \times n$. Jos matriisi A toteuttaa ehdon $\sum_{i=1}^{n} a_{t,i} = 0$, kun $t = 1, 2, \ldots, p$, niin matriisi $A^T A$, joka on kokoa $n \times n$, toteuttaa ehdon $\sum_{i=1}^{n} a_{i,t} = 0$, kun $t = 1, 2, \ldots, n$.*

Todistus. Olkoon matriisi A muotoa

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \ldots & a_{pn} \end{pmatrix}.$$

Silloin

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \ldots & a_{p1} \\ a_{12} & a_{22} & \ldots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \ldots & a_{pn} \end{pmatrix}.$$

Matriisin A^TA i. sarake on

$$\begin{pmatrix} a_{11}a_{1i} + a_{21}a_{2i} + \cdots + a_{p1}a_{pi} \\ a_{12}a_{1i} + a_{22}a_{2i} + \cdots + a_{p2}a_{pi} \\ \vdots \\ a_{1i}a_{1i} + a_{2i}a_{2i} + \cdots + a_{ni}a_{pi} \\ \vdots \\ a_{1n}a_{1i} + a_{2n}a_{2i} + \cdots + a_{pn}a_{pi} \end{pmatrix}.$$

Laskemalla yhteen i. sarakkeen alkiot ja käyttämällä oletusta $\sum_{i=1}^{n} a_{t,i} = 0$ saamme tulokseksi

$a_{1i}(a_{11} + a_{12} + \cdots + a_{1n}) + a_{2i}(a_{21} + a_{22} + \cdots + a_{2n}) + \cdots + a_{pn}(a_{p1} + a_{p2} + \cdots + a_{pn}) = 0.$ □

Lause 2.2. *Jos matriisin A rivien summa on nolla, niin silloin jokaisen matriisista A^TA muodostetun ortonormaalin ominaisvektorin summa on nolla.*

Todistus. Merkitsemme ortonormaalia ominaisvektoria vektorilla v_1, jolloin matriisin V sarakkeina ovat vektorit v_1,v_2, …,v_n. Matriisin V sarakkeet saadaan yhtälöstä $Bx = \lambda x$, jossa $B = A^TA$. Olkoon matriisi B muotoa

$$B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{pmatrix}.$$

Ongelmamme on siis ratkaista yhtälöryhmä

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + \cdots + b_{1n}x_n &= \lambda x_1 \\ b_{21}x_1 + b_{22}x_2 + \cdots + b_{2n}x_n &= \lambda x_2 \\ &\vdots \\ b_{n1}x_1 + b_{n2}x_2 + \cdots + b_{nn}x_n &= \lambda x_n. \end{aligned}$$

Laskemalla yhtälöt puolittain yhteen ja ottamalla termit x_i yhteisiksi tekijöiksi saamme yhtälöryhmän muotoon

$$x_1(b_{11} + b_{21} + \cdots + b_{n1}) + x_2(b_{12} + b_{22} + \cdots + b_{n2}) + \cdots + x_n(b_{1n} + b_{2n} + \cdots + b_{nn}) \\ = \lambda(x_1 + x_2 + \cdots + x_n).$$

Nyt voimme käyttää hyväksi apulausetta 2.1, jonka mukaan siis $\sum_{t=1}^{n} b_{t,i} = 0$, kun $i = 1, 2, \ldots, n$. Saamme

$$\lambda(x_1 + x_2 + \cdots + x_n) = 0.$$

□

2.2.2 Suurimman uskottavuuden estimaatti

Voimme pitää kuolemaa itsenäisenä tapahtumana, mikäli unohdamme tarttuvat sairaudet, luonnonkatastrofit ja sodat. Tästä syystä kuolemien lukumäärä populaatiossa x ajanhetkellä t noudattaa Poisson-jakaumaa parametrein $\lambda_{x,t}$. Merkitsemme kuolleiden lukumäärää satunnaismuuttujalla $D_{x,t}$. On siis likimain voimassa, että

$$\lambda_{x,t} = q_{x,t} E_{x,t}. \tag{2.4}$$

Yhtälössä (2.4) muuttuja $q_{x,t}$ tarkoittaa x-ikäisten kuolleisuutta vuonna t ja $E_{x,t}$ vuoden t alussa elossa olevien x-ikäisten ihmisten lukumäärää. Poisson-jakautuneen satunnaismuuttujan X todennäköisyysfunktio on muotoa

$$L_{x,t} = P(D_{x,t} = d_{x,t}) = \frac{\lambda_{x,t}^{d_{x,t}} e^{-\lambda}}{d_{x,t}!}, \tag{2.5}$$

jossa merkintä $P(D_{x,t} = d_{x,t})$ tarkoittaa todennäköisyyttä, että vuonna t kuolee täsmälleen muuttujan $d_{x,t}$ verran ihmisiä.

Suurimman uskottavuuden määrittämiseksi logaritmoimme funktion (2.5), jolloin saamme

$$\ln(L_{x,t}) = d_{x,t} \ln(\lambda_{x,t}) - \lambda_{x,t} - \ln(d_{x,t}!).$$

Oletamme, että kuolemat ovat toisistaan riippumattomia tapahtumia, jolloin voimme muodostaa logaritmoidun uskottavuusfunktion l kaavalla

$$l = \sum_{x,t}(\ln(L_{x,t})) = \sum_{x,t}[d_{x,t}\ln(\lambda_{x,t}) - \lambda_{x,t} - \ln(d_{x,t}!)]. \tag{2.6}$$

Nyt voimme maksimoida funktion (2.6) muuttujan λ suhteen. Termi $\ln(d_{x,t}!)$ ei riipu muuttujasta λ, joten voimme jättää sen pois maksimia määrittäessämme. Siispä l saavuttaa maksiminsa muuttujan λ suhteen silloin ja vain silloin, kun lauseke

$$\sum[d_{x,t}\ln(\lambda_{x,t}) - \lambda_{x,t}]$$

saavuttaa maksiminsa muuttujan λ suhteen. Saamme yhtälöiden (2.2) ja (2.4) perusteella ongelman muotoon

$$\sum_{x,t}[d_{x,t}\ln(e^{\hat{a}_x+\hat{b}_x\hat{k}_t}E_{x,t}) - E_{x,t}e^{\hat{a}_x+\hat{b}_x\hat{k}_t}].$$

Merkitsemme

$$f(\hat{a},\hat{b},\hat{k}) = \sum_{x,t}[d_{x,t}\ln(e^{\hat{a}_x+\hat{b}_x\hat{k}_t}E_{x,t}) - E_{x,t}e^{\hat{a}_x+\hat{b}_x\hat{k}_t}], \tag{2.7}$$

jolloin tehtävämme on siis maksimoida funktio $f(\hat{a},\hat{b},\hat{k})$. Yhtälössä (2.7) merkintä $\sum_{x,t}$ tarkoittaa kaksoissummaa $\sum_{x=1}^{X}\sum_{t=1}^{T}$, joten voimme kirjoittaa yhtälön (2.7) muotoon

$$\begin{aligned} f(\hat{a},\hat{b},\hat{k}) &= \sum_{x,t}[d_{x,t}(\hat{a}_x + \hat{b}_x\hat{k}_t) + d_{x,t}\ln E_{x,t} - E_{x,t}e^{\hat{a}_x+\hat{b}_x\hat{k}_t}] \\ &= \sum_{x=1}^{X}(\hat{a}_x\sum_{t=1}^{T}d_{x,t}) + \sum_{x=1}^{X}(\hat{b}_x\sum_{t=1}^{T}d_{x,t}\hat{k}_t) + \sum_{x,t}d_{x,t}\ln E_{x,t} - \sum_{x,t}E_{x,t}e^{\hat{a}_x+\hat{b}_x\hat{k}_t}. \end{aligned}$$

Funktion f osittaisderivaatat muuttujien $\hat{a}_x$, $\hat{b}_x$ ja $\hat{k}_t$ suhteen ovat

$$\begin{aligned} \frac{\partial f}{\partial \hat{a}_x} &= \sum_{t=1}^{T}d_{x,t} - \sum_{t=1}^{T}E_{x,t}(e^{\hat{a}_x+\hat{b}_x\hat{k}_t}), && x = 1,2,\ldots,X, \\ \frac{\partial f}{\partial \hat{b}_x} &= \sum_{t=1}^{T}\hat{k}_t d_{x,t} - \sum_{t=1}^{T}\hat{k}_t E_{x,t}(e^{\hat{a}_x+\hat{b}_x\hat{k}_t}), && x = 1,2,\ldots,X, \\ \frac{\partial f}{\partial \hat{k}_t} &= \sum_{x=1}^{X}\hat{b}_x d_{x,t} - \sum_{x=1}^{X}\hat{b}_x E_{x,t}(e^{\hat{a}_x+\hat{b}_x\hat{k}_t}), && t = 1,2,\ldots,T. \end{aligned}$$

Nyt meillä on kolme yhtälöä ja kolme tuntematonta muuttujaa. On monia tapoja ratkaista tämäntyyppisiä yhtälöryhmiä. Voisimme asettaa yhtälöt nolliksi ja ratkaista muuttujat $\hat{a}_x$, $\hat{b}_x$ ja $\hat{k}_t$. Tässä työssä käytämme hyväksi R-ohjelman valmista funktiota nlm, joka minimoi käyttäjän antaman funktion. Funktiota nlm ei voi suoraan käyttää yhtälöön (2.7), sillä ongelmana on maksimoida yhtälön funktio. Yhtälön merkkiä vaihtamalla saamme muunnettua maksimointiongelman minimointiongelmaksi ja voimme käyttää nlm-funktiota.

2.2.3 Painotetun pienimmän neliösumman menetelmä

Painotetun pienimmän neliösumman menetelmän tarkoitus on mallia tehtäessä painottaa jotakin ulkopuolista muuttujaa. Painotuksina voidaan käyttää muun muassa kuolleiden lukumäärää tai kuolleilta vapautuneita pääomia. Tässä työssä lukumääräkuolleisuutta tutkittaessa painotusmatriiseina käytämme kuolleiden lukumääriä. Käytämme samoja painotusmatriiseja soveltaessamme mallia rahakuolleisuuteen. Tehtävämme on siis minimoida funktio

$$g(\hat{a},\hat{b},\hat{k}) = \sum_{x,t} d_{x,t}(\ln(q_{x,t}) - \hat{a}_x - \hat{b}_x\hat{k}_t)^2 = \sum_{x=1}^{X}\sum_{t=1}^{T} d_{x,t}[\epsilon_{x,t}]^2. \tag{2.8}$$

Matriisissa $d_{x,t}$ on x-ikäisenä kuolleiden lukumäärä vuonna t ja virhetermi $\epsilon_{x,t} = \ln(q_{x,t}) - \hat{a}_x - \hat{b}_x\hat{k}_t$. Samat rajoitusehdot kuin yhtälössä (2.2) ovat voimassa eli $\sum_t \hat{k}_t = 0$ ja $\sum_x \hat{b}_x^2 = 1$.

Laskemme aluksi virhetermin $\epsilon = \ln(q_{x,t}) - \hat{a}_x - \hat{b}_x\hat{k}_t$ osittaisderivaatat muuttujien $\hat{a}_x$, $\hat{b}_x$ ja $\hat{k}_t$ suhteen. Saamme, että

$$\frac{\partial\epsilon_{x,t}}{\partial\hat{a}_x} = -1, \qquad \frac{\partial\epsilon_{x,t}}{\partial\hat{b}_x} = -\hat{k}_t \qquad \text{ja} \qquad \frac{\partial\epsilon_{x,t}}{\partial\hat{k}_x} = -\hat{b}_x.$$

missä $x = 1, 2, \ldots, X$ ja $t = 1, 2, \ldots, T$.

Osittaisderivoimalla kaava (2.8) muuttujien $\hat{a}$, $\hat{b}$ ja $\hat{k}$ suhteen saamme

$$\frac{\partial g}{\partial \hat{a}_x} = 2\sum_{t=1}^{T} d_{x,t}\epsilon_{x,t}\frac{\partial \epsilon}{\partial \hat{a}_x} = -2\sum_{t=1}^{T} d_{x,t}[\ln(q_{x,t}) - \hat{a}_x - \hat{b}_x\hat{k}_t], \qquad \forall x \in \{1,2,\ldots,X\},$$

$$\frac{\partial g}{\partial \hat{b}_x} = 2\sum_{t=1}^{T} d_{x,t}\epsilon_{x,t}\frac{\partial \epsilon}{\partial \hat{b}_x} = -2\sum_{t=1}^{T} d_{x,t}\hat{k}_t[\ln(q_{x,t}) - \hat{a}_x - \hat{b}_x\hat{k}_t], \qquad \forall x \in \{1,2,\ldots,X\},$$

$$\frac{\partial g}{\partial \hat{k}_t} = 2\sum_{x=1}^{X} d_{x,t}\epsilon_{x,t}\frac{\partial \epsilon}{\partial \hat{k}_t} = -2\sum_{x=1}^{X} d_{x,t}\hat{b}_x[\ln(q_{x,t}) - \hat{a}_x - \hat{b}_x\hat{k}_t], \qquad \forall t \in \{1,2,\ldots,T\}.$$

Asetamme nämä osittaisderivaatat nolliksi ja ratkaisemme yhtälöt muuttujien $\hat{a}_x$, $\hat{b}_x$ ja $\hat{k}_t$ suhteen, jolloin saamme minimoitua yhtälön (2.8) funktion. Siis

$$\hat{a}_x = \frac{1}{\sum_{t=1}^{T} d_{x,t}}\sum_{t=1}^{T} d_{x,t}[\ln(q_{x,t}) - \hat{b}_x\hat{k}_t], \qquad \forall x \in \{1,2,\ldots,X\},$$

$$\hat{b}_x = \frac{1}{\sum_{t=1}^{T} d_{x,t}\hat{k}_t^2}\sum_{t=1}^{T} d_{x,t}[\ln(q_{x,t}) - \hat{a}_x], \qquad \forall x \in \{1,2,\ldots,X\},$$

$$\hat{k}_t = \frac{1}{\sum_{x=1}^{X} d_{x,t}\hat{b}_x^2}\sum_{x=1}^{X} d_{x,t}[\ln(q_{x,t}) - \hat{a}_x], \qquad \forall t \in \{1,2,\ldots,T\}.$$

Tässä työssä käytämme R-ohjelman valmista funktiota nlm minimoidessamme yhtälöä (2.8).

3 Menetelmien tulokset

Olemme sovittaneet Leen-Carterin mallia neljään eri aineistoon. Jokaiseen aineistoon on sovitettu pienimmän neliösumman sovite (PNS), suurimman uskottavuuden estimaatti (SUE) ja singulaariarvohajotelma (SVD). Seuraavaksi vertailemme eri menetelmillä saatuja vektoreita $\hat{a}_x$, $\hat{b}_x$ ja $\hat{k}_t$ saadaksemme selville, onko menetelmien tuloksissa eroja.

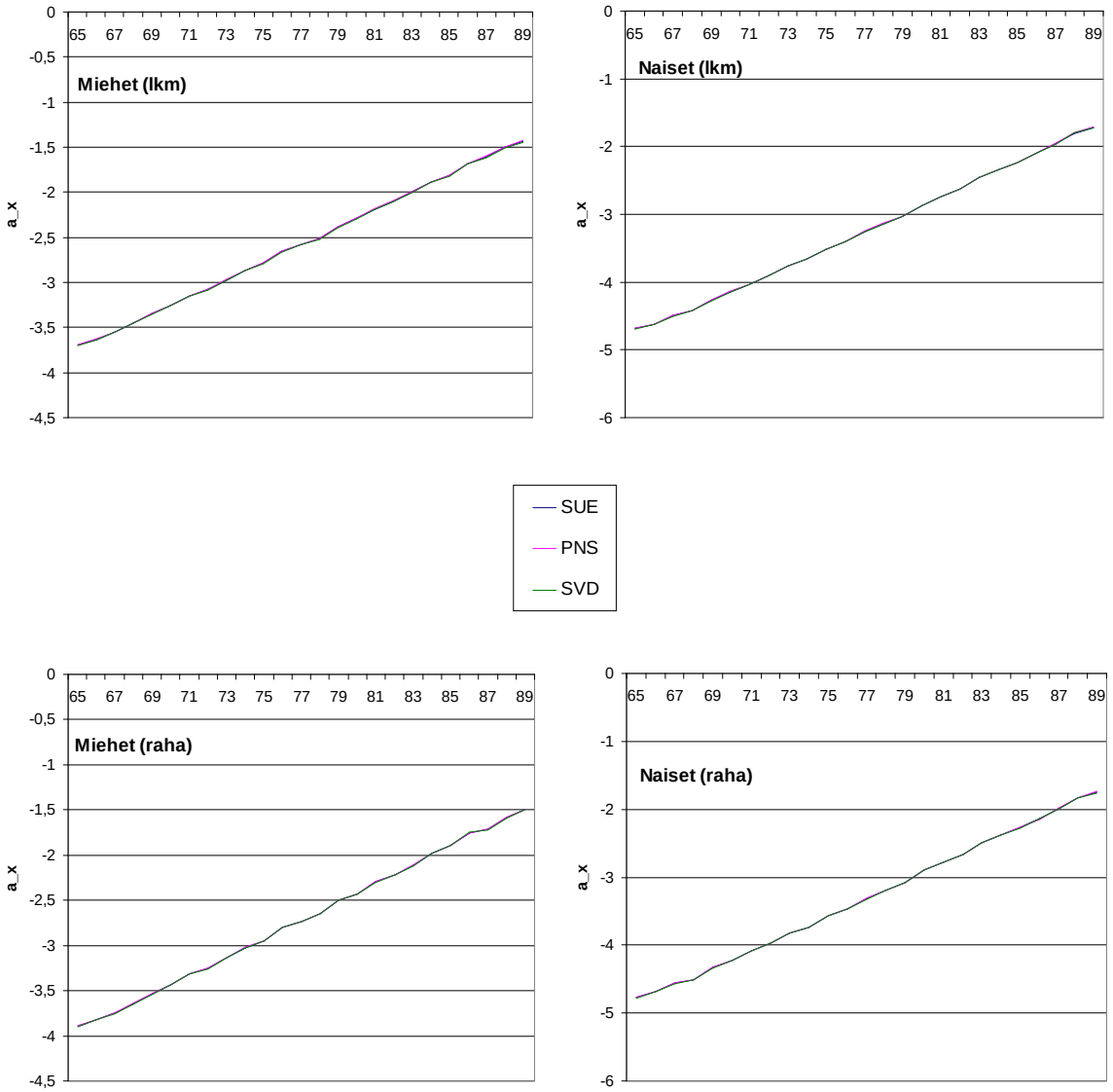
Kuvassa 1 on piirretty vektorin $\hat{a}_x$ kuvaajat. Näemme, ettei menetelmissä synny vektoreiden $\hat{a}_x$ välille suurta eroa. Tämä on täysin ymmärrettävää,

sillä vektorissa $\hat{a}_x$ on ikäluokkakohtaiset logaritmoidut kuolleisuudet. Singulaariarvohajotelman tapauksessa kuolleisuus on laskettu suoraan aineistosta, joten sitä ei estimoida lainkaan. Myös muut menetelmät antavat hyvin tarkasti samanlaisen vektorin $\hat{a}_x$ kuin singulaariarvohajotelma.

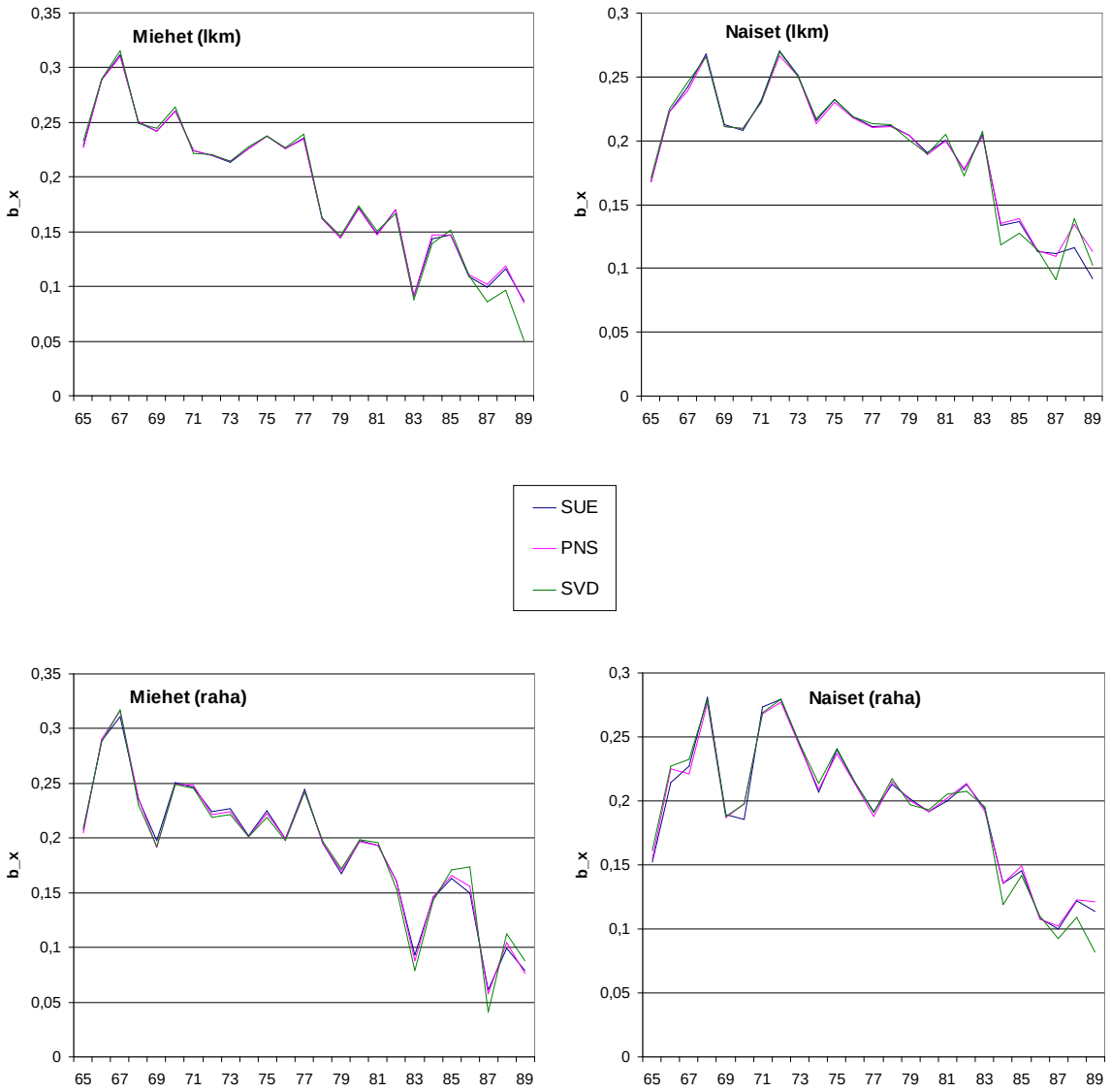
Kuvaan 2 on piirretty vektorin $\hat{b}_x$ kuvaajat. Mallien välillä ei ole havaittavissa säännöllisiä eroavaisuuksia. Ainoastaan vanhimmissa ikäluokissa esiintyy pieniä eroja. Ei kuitenkaan voida sanoa, että jollakin menetelmällä saataisiin selvästi muista poikkeava vektori $\hat{b}_x$. Laskeva trendi vektorissa $\hat{b}_x$ kertoo siitä, että kuolleisuuden muutosnopeus hidastuu siirryttäessä vanhempiin ikäluokkiin.

Vektori $\hat{k}_t$ on mielenkiintoisin estimoiduista vektoreista, sillä sen pohjalta teemme myöhemmin ennusteen tulevaisuudesta. Itse asiassa lineaarisessa ennusteessa otamme huomioon vain vektorin $\hat{k}_t$ alku- ja loppupisteiden arvot. Tässäkään tapauksessa erot menetelmien välillä eivät kuitenkaan ole suuria, kuten kuvasta 3 näemme.

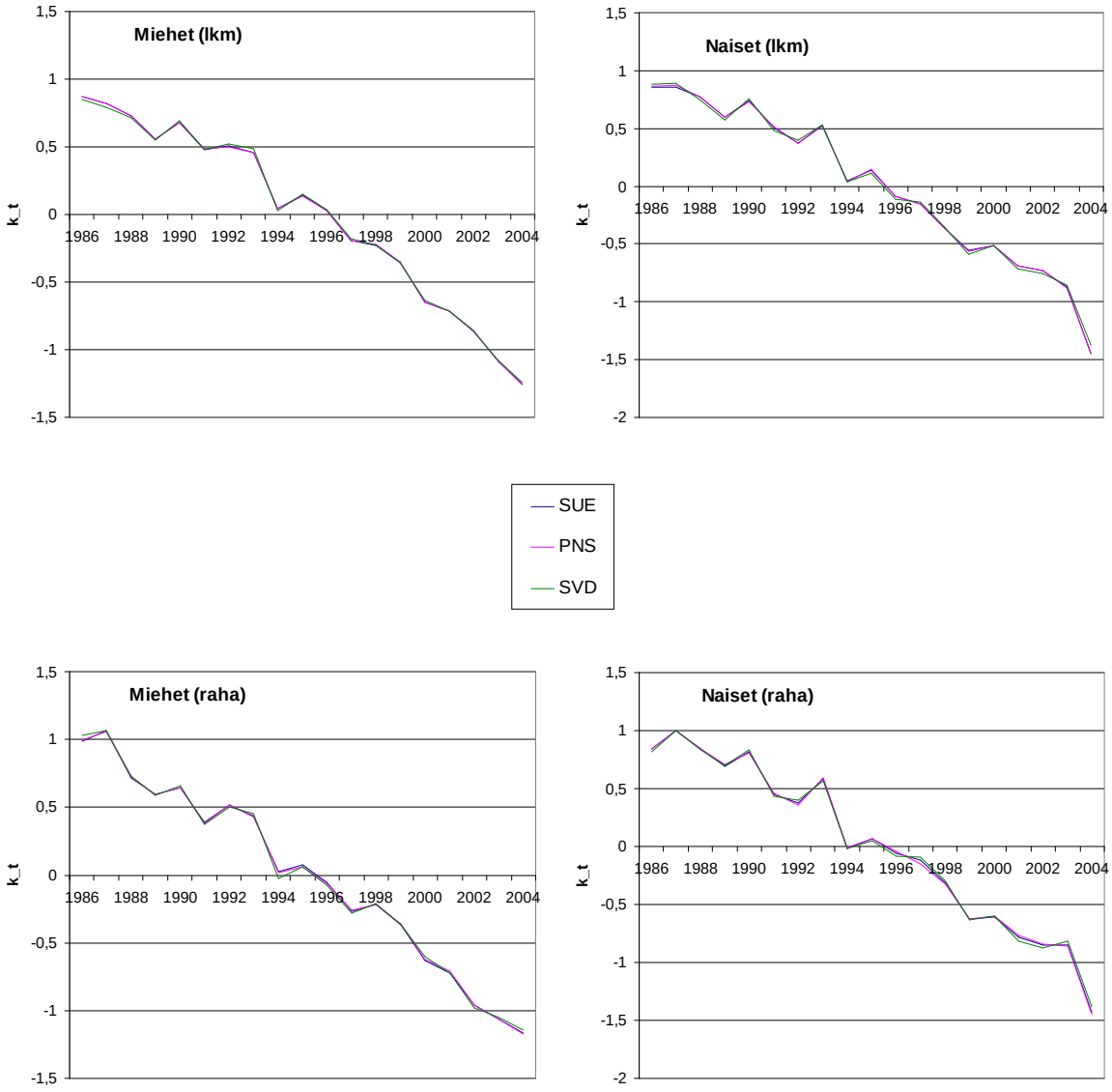
Naisten ja miesten välinen vertailu ei ole kovin mielekästä, sillä naisten kuolleisuus on jokaisessa ikäluokassa matalempaa kuin miesten. Tästä huolimatta voimme todeta, että kuvaajat muistuttavat toisiaan paljon.



Kuva 1: Vektorin $\hat{a}_x$ kuvaaja ikäluokille 65-89.



Kuva 2: Vektorin $\hat{b}_x$ kuvaaja ikäluokille 65-89.

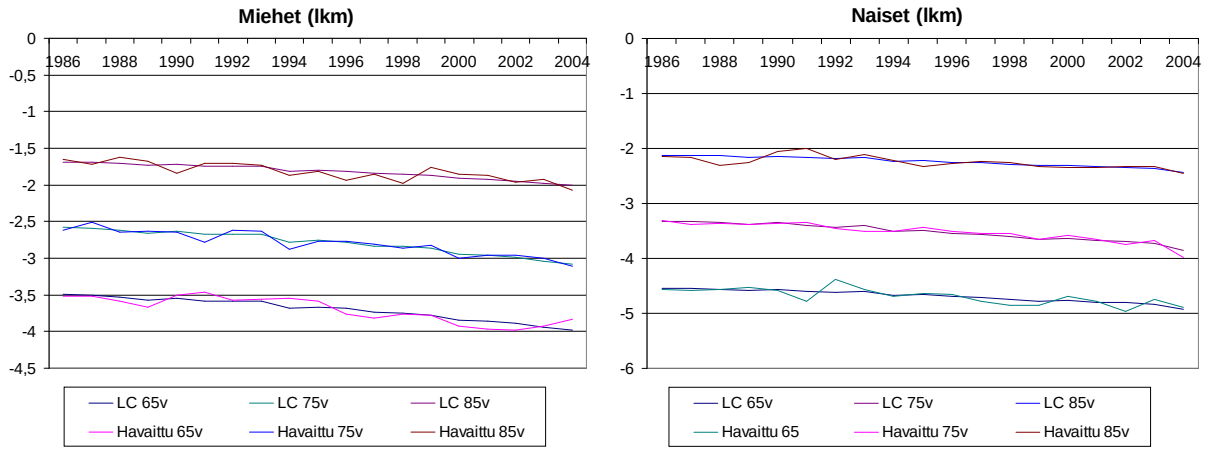


Kuva 3: Vektorin $\hat{k}_t$ kuvaaja vuosina 1986-2004.

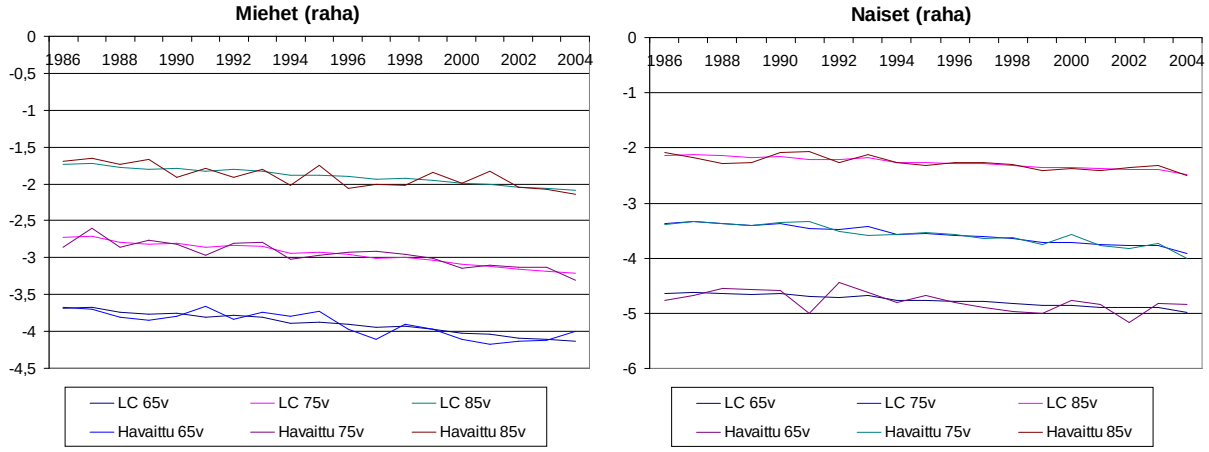
3.1 Leen-Carterin mallin sopivuus alkuperäiseen aineistoon

On tietenkin tärkeää tutkia, kuinka hyvin Leen-Carterin mallista saatu matriisi vastaa oikeaa havaintomatriisia. Seuraavaksi havainnollistamme sitä kuvien avulla ja myöhemmin tutkimme mallien sopivuutta normeja käyttäen. Piirrämme nyt samaan kuvioon suurimman uskottavuuden menetelmän avulla estimoimamme kuolleisuusmatriisin sekä alkuperäisen havaintomatriisimme. Teemme siis erikseen sovitteet sekä miesten ja naisten henkikuolleisuusmatriiseille että miesten ja naisten rahakuolleisuusmatriiseille.

Leen-Carterin malli sopii erittäin hyvin havaintomatriiseille, kuten kuvista 4 ja 5 näemme. (Huom. Kuolleisuusluvut ovat logaritmoituja.) Kuvissa 4 ja 5 on piirretty havaitut kuolleisuusluvut sekä suurimman uskottavuuden mallilla muodostetut kuolleisuusluvut kolmelle eri ikäluokalle. Havaituissa aineistossa on nähtävissä pientä vaihtelua, mutta päätrendi on jokaisessa ikäluokassa kuitenkin laskeva. Tämä tarkoittaa sitä, että kuolleisuus on vähentynyt kaikissa kolmessa ikäluokassa vuosina 1986–2004.



Kuva 4: Suurimman uskottavuuden menetelmällä saatu Leen-Carterin malli henkikuolleisuudelle.



Kuva 5: Suurimman uskottavuuden menetelmällä saatu Leen-Carterin malli rahakuolleisuudelle.

3.2 Normeista

Leen-Carterin mallilla estimoidun matriisin sopivuutta havaintomatriisiin voidaan mitata niin sanotulla Frobeniuksen normilla, joka on yleisin käytetty matriisinormi. Frobeniuksen normi on itse asiassa Eukleideen normi, mutta matriiseille sovellettuna puhutaan usein Frobeniuksen normista. Toinen havaintomatriisin ja estimoidun matriisin sopivuutta mittaava normi on l_1-normi, joka määritellään yleensä vain vektoreille, mutta sovellamme sitä nyt myös matriiseille. Laskemme lisäksi residuaalien summan, jonka arvo kertoo, ovatko estimoimamme kuolleisuusluvut sijoittuneet mahdollisesti yli vai alle havaittuiden kuolleisuuslukujen.
HUOM. Kuolleisuusluvut ovat logaritmoituja.

Muodostetaan kaikille estimoiduille matriiseille virhematriisi E, joka saadaan kaavalla

$$E = H - L = (h_{i,j} - l_{i,j}),$$

missä matriisissa H on havaitut kuolleisuusluvut ja matriisissa L Leen-Carterin mallilla estimoidut kuolleisuusluvut.

Määritelmä 3.1. Matriisille A yleinen $l_p - normi$ määritellään kaavalla

$$\|A\|_p = \left(\sum_{i,j} |a_{i,j}|^p \right)^{1/p}.$$

Määritelmä 3.2. Frobeniuksen normi on erikoistapaus l_p-normista, kun $p = 2$. Määritellään *Frobeniuksen normi* matriisille E kaavalla

$$\|E\|_2 = \left(\sum_{i,j} (e_{i,j})^2 \right)^{1/2}.$$

[4, s. 291]

Seuraavassa taulukossa on laskettu Frobeniuksen normi jokaiselle virhematriisille.

Henkik.	SVD	PNS	SUE	Rahak.	SVD	PNS	SUE
Miehet	1,43	1,45	1,44	Miehet	1,76	1,77	1,77
Naiset	1,43	1,45	1,45	Naiset	1,86	1,88	1,88

Taulukko 1: Frobeniuksen normit

Määritelmä 3.3. l_1-normin on erikoistapaus l_p-normista, kun $p = 1$. Määrittelemme matriisille E $l_1 - normin$ kaavalla

$$\|E\|_1 = \sum_{i,j} |e_{i,j}|.$$

[4, s. 291]

Henkik.	SVD	PNS	SUE	Rahak.	SVD	PNS	SUE
Miehet	24,01	23,89	23,89	Miehet	30,60	30,71	30,78
Naiset	24,57	24,58	24,57	Naiset	31,92	32,06	31,94

Taulukko 2: l_1-normit

Määritelmä 3.4. Määrittelemme estimoitujen arvojen poikkeaman havaitusta matriisista kaavalla

$$\sum_{i,j} e_{i,j}. \tag{3.1}$$

Henkik.	SVD	PNS	SUE	Rahak.	SVD	PNS	SUE
Miehet	0,00	2,54	1,58	Miehet	0,00	1,58	1,20
Naiset	0,00	2,46	1,00	Naiset	0,00	2,71	2,01

Taulukko 3: Poikkeamat

On erittäin hankalaa sanoa, mikä menetelmä tuottaisi parhaan sovitteen. Taulukosta 1 näemme, että Frobeniuksen normin arvot ovat lähes identtiset kaikilla menetelmillä. Oli myös odotettavissa, että rahakuolleisuudelle lasketut normit ovat suurempia kuin henkikuolleisuudelle lasketut. Yksi syy tähän

voi olla se, että rahakuolleisuuden havaintoaineistossa on suurempia vaihteluita kuin henkikuolleisuusaineistossa, ks. esim. kuva 5. L_1-normin arvoissakaan ei ole juuri eroa. Sen sijaan eroja löytyy residuaalien summista. Singulaariarvohajotelmalla (SVD) tuotettu sovite minimoi residuaalien summan nollaksi. Voimme havaita myös, että suurimman uskottavuuden menetelmällä saadut sovitteet antavat pienemmät residuaalien summat kuin pienimmän neliösumman menetelmällä estimoidut sovitteet.

4 Leen-Carterin mallin ennustaminen

Leen-Carterin mallin vahvuus on siinä, että ainoastaan kuolleisuuden muutosta kuvaavaa vektoria $\hat{k}_t$ ennustetaan. Vektori $\hat{a}_x$, joka kuvaa keskiarvokuolleisuutta, pysyy vakiona kuten myös vektori $\hat{b}_x$. Tehtävämme on siis löytää vektorille $\hat{k}_t$ mahdollisimman sopiva ARIMA-malli ja tehdä löydetyn mallin pohjalta ennuste. Lineaarista ennustetta pidetään yleisesti parhaana, mutta myös muunlaisia malleja voidaan kokeilla. Käymme seuraavaksi läpi hieman aikasarja-analyysiä ja sen jälkeen teemme jokaiselle mallillemme lineaarisen ennusteen ARIMA(0,1,0)-mallilla. Lisäksi yritämme etsiä, voisiko jokin muu ARIMA-malli tulla kysymykseen ennusteessa.

4.1 Aikasarja-analyysin perusteita

Yleensä aikasarja pyritään erilaisilla muunnoksilla saattamaan sellaiseen muotoon, että sen voi katsoa olevan realisaatio stationaarisesta prosessista. Stationaarista prosessia voidaan yleensä mallintaa tilastollisesti niin kutsutun ARMA-prosessin avulla. Alkuperäinen aikasarja voidaan sitten kuvata käyttäen stationaarista aikasarjaa ja sellaisia muunnoksia, joilla siitä saadaan alkuperäinen sarja.
[3, s. 1]

Määritelmä 4.1. Aikasarjan $\{X_t\}$ sanotaan olevan (*heikosti*) *stationaarinen*, mikäli se täyttää seuraavat kolme ehtoa:

1. $E\left|X_t\right|^2 < \infty$, $\forall t \in N$,
2. $EX_t = \mu$, $\forall t \in N$,

3. $Cov(X_r, X_s) = Cov(X_{r+t}, X_{s+t}), \qquad \forall t, r, s \in N.$

Kovarianssifunktion määritelmän mukaan $Cov(X_r, X_s) = E[(X_r - E(X_r))(X_s - E(X_s))]$. Funktiota $Cov(X_t, X_{t+h}) = \gamma$ sanotaan autokovarianssifunktioksi viiveellä h.

Määritelmä 4.2. Korreloimattomien satunnaismuuttujien jonoa $\{X_t\}$, jossa muuttujilla on odotusarvo 0 ja varianssi σ^2, sanotaan valkoisen kohinan prosessiksi ja sitä merkitään $X_t \sim WN(0, \sigma^2)$.

Määritelmä 4.3. Stationaarisen aikasarjan $\{X_t\}$ sanotaan olevan AR(p)-prosessi, mikäli se on muotoa

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t, \tag{4.1}$$

jossa $\epsilon_t \sim WN(0, \sigma^2)$ ja luvut ϕ_i ovat vakioita.

Määritelmä 4.4. Stationaarisen aikasarjan $\{X_t\}$ sanotaan olevan MA(q)-prosessi, mikäli se on muotoa

$$X_t = \theta_0 \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}, \tag{4.2}$$

jossa $\epsilon_t \sim WN(0, \sigma^2)$ ja luvut θ_i ovat vakioita.

Määritelmä 4.5. Kahden havainnon x_i ja y_j välinen *korrelaatio* määritellään kaavalla

$$\rho_{i,j} = \frac{Cov(x_i, y_j)}{\sqrt{Var(x_i) Var(y_j)}}. \tag{4.3}$$

HUOM. Mikäli satunnaismuuttujat ovat riippumattomia, ovat ne myös korreloimattomia, mutta korreloimattomat satunnaismuuttujat voivat kuitenkin olla riippuvia.

Määritelmä 4.6. Tutkittaessa aikasarjan käyttäytymistä on usein hyödyllistä tutkia sen autokorrelaatiofunktiota. *Autokorrelaatiofunktio* viiveellä h määritellään aikasarjalle $\{X_t\}$ kaavalla

$$\rho(h) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t) Var(X_{t+h})}}.$$

Esimerkki 4.1. MA(q)-prosessin $X_t = \theta_0\epsilon_t - \theta_1\epsilon_{t-1} - \cdots - \theta_q\epsilon_{t-q}$ autokorrelaatiofunktio on muotoa

$$\rho_x(h) = \begin{cases} 1, & \text{kun } h = 0, \\ \frac{\sum_{i=0}^{q-h}\theta_i\theta_{i+h}}{\sum_{i=0}^{q}\theta_i^2}, & \text{kun } h = 1, 2, \ldots, q, \\ 0, & \text{kun } h > q. \end{cases}$$

Käytännössä pystymme siis autokorrelaatiofunktion arvoista päättelemään voisiko sarjaa mallintaa MA(q)-prosessina. Mikäli viiveellä h lasketun autokorrelaatiofunktion arvo menee yli 95%:n luottamusvälin, niin silloin voimme yrittää sovittaa aineistoon MA(h)-prosessia.

Esimerkki 4.2. MA(1)-prosessi. Oletetaan, että jono $\{Z_t, t = 0, t = \pm 1, t = \pm 2\}$ on valkoista kohinaa. Määritellään $X_t = Z_t + \theta Z_{t-1}$, missä θ on reaaliluku. Tällöin $\mu_x(t) = 0$ ja

$$\begin{aligned} EX_t^2 &= E(Z_t + \theta Z_{t-1})^2 = E(Z_t^2 + 2Z_t\theta Z_{t-1} + \theta^2 Z_{t-1}^2) \\ &= E(Z_t^2) + E(2Z_t\theta Z_{t-1}) + E(\theta^2 Z_{t-1}^2) \\ &= \sigma^2 + \sigma^2\theta^2 \\ &= \sigma^2(1 + \theta^2) < \infty. \end{aligned}$$

Lisäksi prosessin autokovarianssifunktio on

$$\gamma_x(t + h, t) = \begin{cases} \sigma^2(1 + \theta^2), & \text{kun } h = 0, \\ \sigma^2\theta, & \text{kun } h = \pm 1, \\ 0, & \text{kun } |h| > 1. \end{cases}$$

Koska $\mu_x(t)$ ja $\gamma_x(t+h, t)$ eivät riipu t:stä, kyseessä on stationaarinen prosessi. Autokorrelaatiofunktio prosessille $\{X_t\}$ on

$$\rho_x(t + h, t) = \begin{cases} 1, & \text{kun } h = 0, \\ \frac{\theta}{1+\theta^2}, & \text{kun } h = \pm 1, \\ 0, & \text{kun } |h| > 1. \end{cases}$$

Esimerkki 4.3. AR(1)-prosessi. Oletetaan, että $\{X_t\}$ on stationaarinen aikasarja, joka toteuttaa yhtälön

$$X_t = \theta X_{t-1} + Z_t, \qquad t = 0, \pm 1, \pm 2 \ldots \tag{4.4}$$

missä $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$, $|\theta| < 1$ ja Z_t on korreloimaton satunnaismuuttujien X_s kanssa, kun $s < t$. Tällöin sanotaan, että $\{X_t\}$ noudattaa autoregressiivistä prosessia viiveellä yksi (eli AR(1)-prosessia). Ottamalla odotusarvon yhtälön (4.4) molemmilta puolilta saamme prosessin odotusarvoksi $EX_t = 0$, sillä

$$\begin{aligned}\mu = E(X_t) &= E(\theta X_{t-1} + Z_t) \\ &= E(\theta X_{t-1}) + E(Z_t) \\ &= \theta\mu\end{aligned}$$

eli

$$\mu = 0.$$

Autokovarianssifunktion määrittämiseksi kerromme yhtälön (4.4) puolittain X_{t-h}:lla, missä $h > 0$, ja otamme odotusarvon puolittain, jolloin saamme

$$\begin{aligned}E(X_t X_{t-h}) &= \theta E(X_{t-1} X_{t-h}) + E(Z_t X_{t-h}) \\ \Leftrightarrow Cov(X_t, X_{t-h}) &= \theta Cov(X_{t-1}, X_{t-h}) + 0 \\ \Leftrightarrow \gamma_x(h) &= \theta\gamma_x(h-1).\end{aligned}$$

Rekursiivisesti voimme päätellä, että $\gamma_x(h) = \theta^h\gamma_x(0)$. Koska $\gamma_x(h) = Cov(X_{t+h}, X_t) = Cov(X_t, X_{t-h}) = \gamma_x(-h)$, positiivisille ja negatiivisille viiveille h soveltuva autokovarianssin kaava on $\gamma_x(h) = \theta^{|h|}\gamma_x(0)$. Autokorrelaatio viiveellä h on puolestaan $\rho_x(h) = \gamma_x(h)/\gamma_x(0) = \theta^{|h|}, h = 0, \pm 1, \pm 2, \ldots$ Voidaksemme määrittää, mikä on $\gamma_x(0)$, toteamme ensin, että

$$\begin{aligned}Cov(X_t, Z_t) &= Cov(\theta X_{t-1} + Z_t, Z_t) \\ &= \theta Cov(X_{t-1}, Z_t) + Cov(Z_t, Z_t) = Cov(Z_t, Z_t) \\ &= \sigma^2,\end{aligned}$$

sillä oletuksen mukaan $Cov(X_{t-1}, Z_t) = 0$. Täten

$$\begin{aligned}\gamma_x(0) = Cov(X_t, X_t) &= Cov(X_t, \theta X_{t-1} + Z_t) \\ &= \theta Cov(X_t, X_{t-1}) + Cov(X_t, Z_t) \\ &= \theta^2\gamma_x(0) + \sigma^2,\end{aligned}$$

josta saamme ratkaistua, että

$$\gamma_x(0) = \frac{\sigma^2}{1-\theta^2}.$$

Prosessin autokovarianssifunktio on

$$\gamma_x(t+h,t) = \begin{cases} \frac{\sigma^2}{1-\theta^2}, & \text{kun } h = 0, \\ \frac{\sigma^2\theta^{|h|}}{1-\theta^2}, & \text{kun } h \neq 0. \end{cases}$$

Koska $\mu_x(t)$ ja autokovarianssifunktio $\gamma_x(t+h,t)$ eivät riipu t:stä, kyseessä on stationaarinen prosessi. Prosessin autokorrelaatiofunktio on

$$\rho_x(t+h,t) = \begin{cases} 1, & \text{kun } h = 0, \\ \theta^{|h|}, & \text{kun } h \neq 0. \end{cases}$$

Määritelmä 4.7. Aikasarjan $\{X_t\}$ sanotaan olevan ARMA(p,q)-prosessi, mikäli aikasarja on stationaarinen ja muotoa

(4.5)
$$X_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p k_{t-p} + \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \cdots - \theta_q\epsilon_{t-q},$$

jossa vakioita $\{\phi_1, \phi_2, \ldots, \phi_p\}$ kutsutaan AR-parametreiksi ja vakioita $\{\theta_1, \theta_2, \ldots, \theta_q\}$ MA-parametreiksi.

4.2 Aikasarjan saattaminen stationaariseksi

Differointi viiveellä yksi hävittää sarjasta lineaarisen trendikomponentin. Jos sarja voidaan esittää muodossa $X_t = a + bt + Z_t$, missä $E(Z_t) = 0$, niin differoitu sarja on

$$\begin{aligned} X_t - X_{t-1} &= (a + bt + Z_t) - (a + b(t-1) + Z_{t-1}) \\ &= b + Z_t - Z_{t-1}. \end{aligned}$$

Aikasarjan stationaarisuutta voidaan testata usealla eri tavalla. Yksi tapa on tutkia sarjasta laskettuja autokorrelaatio- ja osittaisautokorrelaatiofunktioita.

ARIMA(p,1,q)-prosessilla tarkoitetaan lähes samaa kuin ARMA(p,q)-prosessilla, mutta nyt mallinnettava sarja differoidaan ensin viiveellä 1. Yleisemmin voidaan puhua myös ARIMA(p,d,q)-prosessista, jossa sarja differoidaan d kertaa.

Valkoisen kohinan prosessista laskettujen autokorrelaatio- ja osittaisautokorrelaatioiden (ACF ja PACF) arvojen pitäisi vähentyä eksponentiaalisesti kohti nollaa, sillä määritelmän mukaan $\rho(0) = 0$, kun $h \neq 0$. Itse asiassa 95% ACF- ja PACF-arvoista pitäisi pysyä rajojen $\pm\frac{1.96}{\sqrt{n}}$ sisäpuolella. Lausekkeessa $\pm\frac{1.96}{\sqrt{n}}$ muuttuja n tarkoittaa havaintojen lukumäärää. [3, s. 4,32]

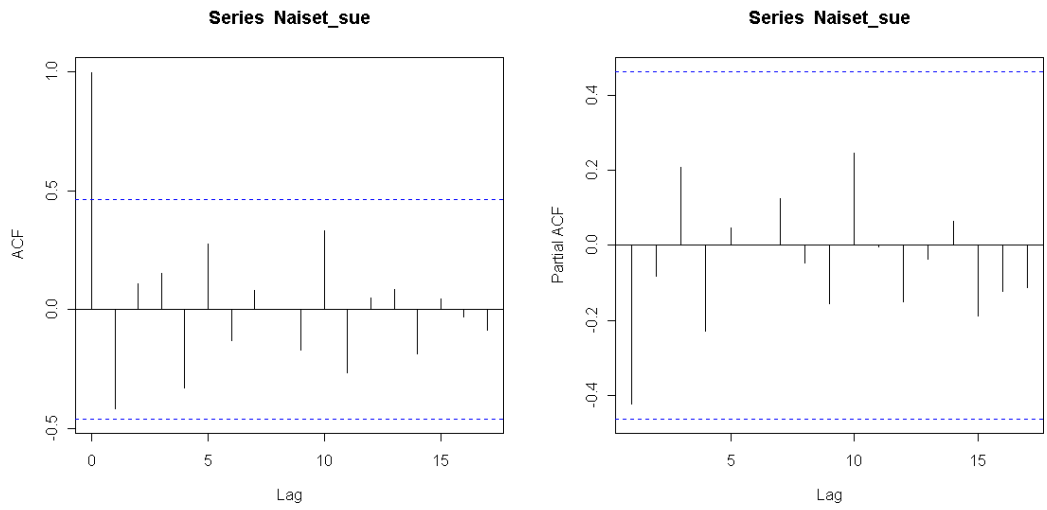
4.3 Vektorin k_t ennustaminen

Näemme kuvasta 3, että kaikista aineistoista estimoidut vektorit $\hat{k}_t$ ovat selvästi laskevia. Voimme eliminoida tämän laskevan trendin differoimalla vektorin $\hat{k}_t$ viiveellä 1. Piirtämällä autokorrelaatio- ja osittaisautokorrelaatiofunktioiden kuvaajat sarjasta $\{\nabla\hat{k}_t^*\} = \{\hat{k}_t - \hat{k}_{t-1}\}$ voimme tutkia sarjan $\{\nabla\hat{k}_t^*\}$ stationaarisuutta. Stationaarisuuden testaamiseen on olemassa myös erilaisia testejä, mutta tässä työssä testaamme stationaarisuutta vain ACF- ja PACF- kuvien avulla.

Otamme esimerkkinä naisten SUE-menetelmällä saadun vektorin $\hat{k}_t$ ennustamisen. Kuten kuvasta 6 näemme, differoidun sarjan $\{\nabla\hat{k}_t^*\}$ ACF- ja PACF-arvot ovat 95%:n luottamusvälien sisäpuolella. Voimme siis olettaa, että sarja on stationaarinen, ja näin ollen muodostaa sille ennusteen ARIMA$(0, 1, 0)$-mallilla. Käytännössä mallinnamme sarjaa siis satunnaiskävelyprosessina, jolloin mallimme on

$$\hat{k}_t = \hat{k}_{t-1} + \hat{c} + \varepsilon_t,$$

jossa $\hat{c} = E(\nabla\hat{k}_t)$ ja virhetermi $\varepsilon_t \sim N(0, \sigma^2)$.



Kuva 6: Naisten henkikuolleisuusaineistosta lasketun $\nabla\hat{k}_t$ ACF:n ja PACF:n kuvaajat.

Seuraavaksi laskemme esimerkkinä muuttujan $\hat{c}$ arvot miesten ja naisten henkikuolleisuusaineistosta lasketuille vektoreille $\hat{k}_t$:

Miehet

(SVD) $\hat{k}_t \approx \hat{k}_{t-1} - 0,11618,$

(SUE) $\hat{k}_t \approx \hat{k}_{t-1} - 0,11819,$

(PNS) $\hat{k}_t \approx \hat{k}_{t-1} - 0,11836,$

Naiset

(SVD) $\hat{k}_t \approx \hat{k}_{t-1} - 0,1251,$

(SUE) $\hat{k}_t \approx \hat{k}_{t-1} - 0,1278,$

(PNS) $\hat{k}_t \approx \hat{k}_{t-1} - 0,1282.$

Sarjalle $\hat{k}_t$ tekemämme ennusteen luottamusvälit muodostamme kaavalla

$$[\hat{k}_t - 1.96 * std(\hat{k}_t), \hat{k}_t + 1.96 * std(\hat{k}_t)], \tag{4.6}$$

jossa $std(\hat{k}_t)$ on sarjan $(\hat{k}_t)$ hajonta. Kuvassa 7 on piirretty PNS-menetelmällä saatu sarja $\hat{k}_t$ ja 95%:n luottamusvälit ennustetuille arvoille.

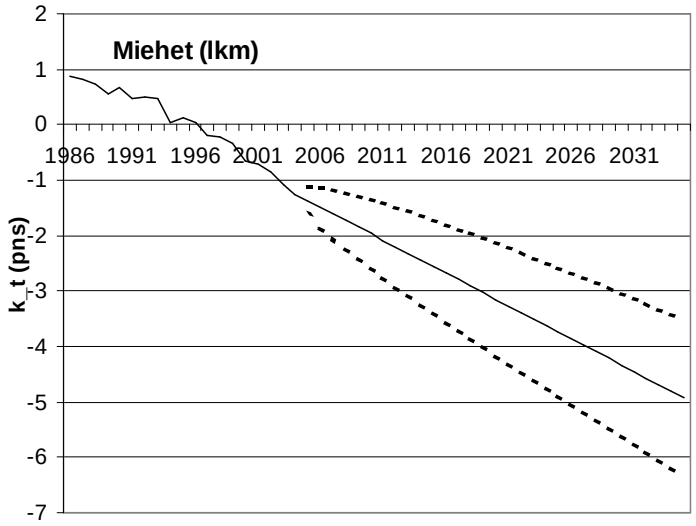
Esimerkki 4.4. Otamme esimerkkinä miesten lukumääräkuolleisuusaineistoon PNS-menetelmällä sovitetun vektorin $\hat{k}_t$ ennustamisen ja muodostamme ennusteelle luottamusvälit. Ennuste on tehty ARIMA$(0,1,0)$-mallilla, joten se on lineaarinen. Oletamme lisäksi, että virhetermi $e_t \sim N(0, \sigma^2)$. Aluksi muodostamme luvun $\hat{c} = E(\nabla \hat{k}_t)$, joka on nyt

$$\frac{k_{2004} - k_{1986}}{18} = \frac{-2,31046}{18} = -0,11836.$$

Seuraavaksi muodostamme yhden vuoden ennusteen eli $k_{2005} = k_{2004} - 0,11836$. Keskihajonta vektorin k_t vuoden 2005 arvolle on vektorin $\nabla \hat{k}_t$ $(t = 2005)$ hajonta. Merkitään sitä $std(e_1)$. Muodostetaan sarja x_t differoimalla vektori k_t viiveellä 1 ja lasketaan $std(e_1)$ kaavalla

$$\begin{aligned} std(e_1) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &\approx \sqrt{\frac{1}{18} * 0,30109} \\ &\approx 0,129334. \end{aligned}$$

Seuraavan vuoden keskihajonta on $\sqrt{2}std(e_1)$ ja kolmannen vuoden $\sqrt{3}std(e_1)$, jne. Haluttu 95%:n luottamusväli saadaan kertomalla keskihajonta luvulla ± 1.96. Arvon k_{2035} luottamusvälin yläraja on $3,519$, alaraja $6,337$ ja keskiennuste $4,925$.



Kuva 7: Miesten henkikuolleisuusaineistosta estimoitu ja ennustettu k_t, 95%:n luottamusvälit katkoviivoilla.

4.4 Elinajanodotteen laskeminen

Elinajanodotteen muutos on yksi tärkeimmistä tunnusluvuista, kun halutaan kuvata väestön ikäkehityksessä tapahtuvaa muutosta. Elinajanodote määritetään yleensä vastasyntyneille, mutta samalla periaatteella voimme laskea myös 65-vuotiaiden elinajanodotteet. Tässä tutkielmassa keskitymme vain 65-vuotiaiden elinajanodotteeseen. Laskemme elinajanodotteet Tilastokeskuksen esittämällä tavalla. Laskenta etenee seuraavasti:

1. Tehdään taulu, jossa on jokaisen ikäluokan kuolemanvaaraluvut halutulta vuodelta.

2. Oletetaan, että 65 vuotta täyttäneitä ihmisiä on alussa 100 000.

3. Käyttämällä 65-vuotiaiden kuolemanvaaralukua lasketaan, kuinka moni heistä kuolee 65-vuotiaana ja vähennetään kuolleiden lukumäärä alkuperäisestä 100 000 ihmisen alkupopulaatiosta.

4. Seuraavaksi lasketaan kuinka moni jäljelle jääneistä henkilöistä kuolee 66-vuotiaina. Saatu tulos vähennetään 66-vuotiaiden alkupopulaatiosta.

5. Näin jatkamalla saamme vektorin, jossa on ikäluokkakohtaisesti elossa olevien lukumäärät.

6. Seuraavaksi lasketaan yhteen elossa olevien henkilöiden määrä kumulatiivisesti vanhimmasta ikävuodesta alkaen. Saadussa vektorissa on elettävänä olevien vuosien kokonaislukumäärä.

7. Lopuksi muodostetaan eri ikäluokkien elinajanodote jakamalla elettävänä olevien vuosien kokonaismäärä saman iän saavuttaneiden lukumäärällä.

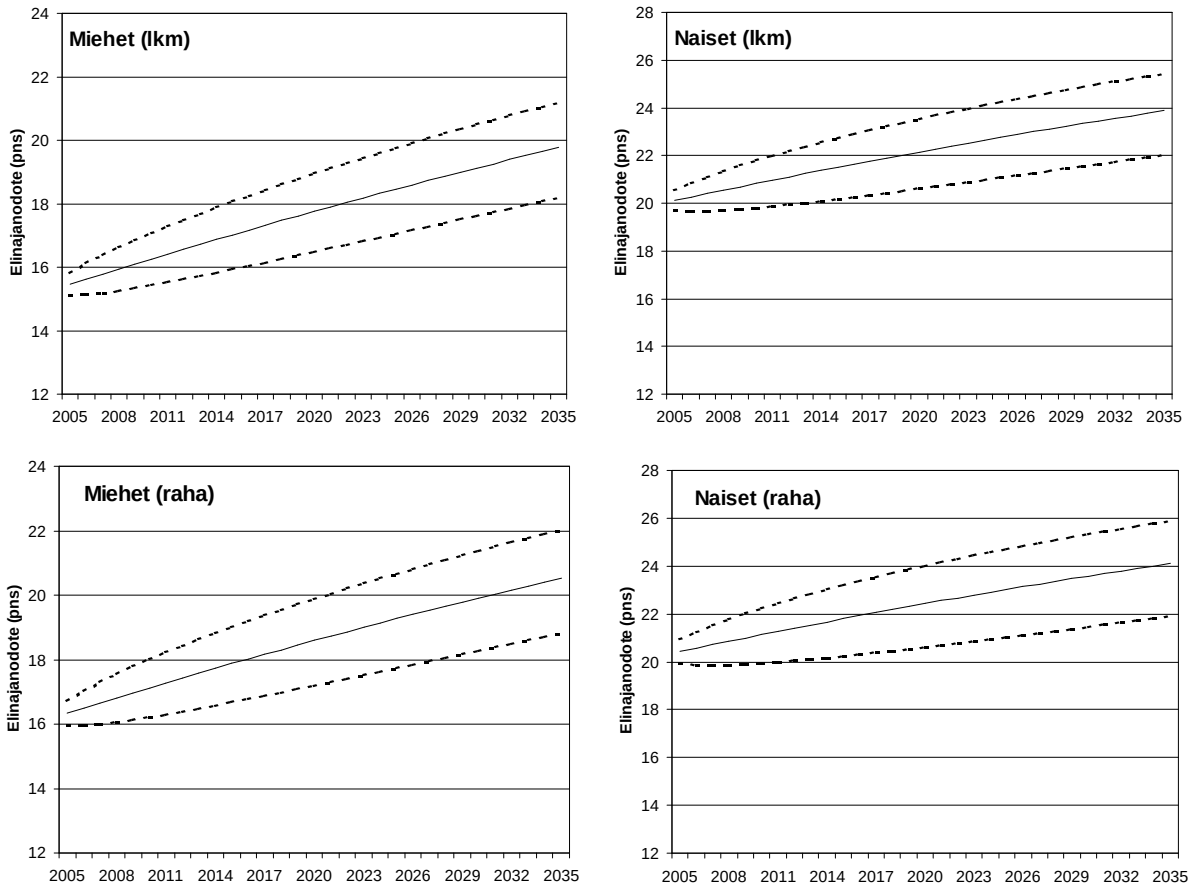
4.5 Elinajanodotteen ennustaminen

Kuolleisuuslukuja käyttäen voimme laatia ennusteen elinajanodotteista. Taulukossa 4 on verrattu eri menetelmien antamia elinajanodotteen ennusteita 65-vuotiaille. Laskiessamme elinajanodotteita 65-vuotiaille tarvitsemme myös yli 90-vuotiaiden kuolleisuuslukuja. Tästä syystä jatkamme aineistoamme ETK:n tekemillä ennusteilla. Käytännössä siis skaalaamme ETK:n yli 90-vuotiaiden ennusteen siten, että se jatkaa omia kuolleisuuslukujamme. Lisäksi teemme oletuksen, että jokainen kuolee viimeistään 100 vuoden iässä.

Henkik.	Vuosi	SVD	PNS	SUE	Rahak.	Vuosi	SVD	PNS	SUE
Miehet	2005	15,47	15,47	15,48	Miehet	2005	16,34	16,35	16,35
	2015	16,96	17,02	17,03		2015	17,88	17,88	17,88
	2025	18,33	18,46	18,46		2025	19,29	19,27	19,27
	2035	19,56	19,77	19,77		2035	20,56	20,53	20,54
Naiset	2005	20,07	20,11	20,11	Naiset	2005	20,33	20,43	20,41
	2015	21,38	21,51	21,46		2015	21,58	21,80	21,76
	2025	22,58	22,76	22,67		2025	22,69	23,02	22,96
	2035	23,65	23,88	23,73		2035	23,66	24,11	24,03

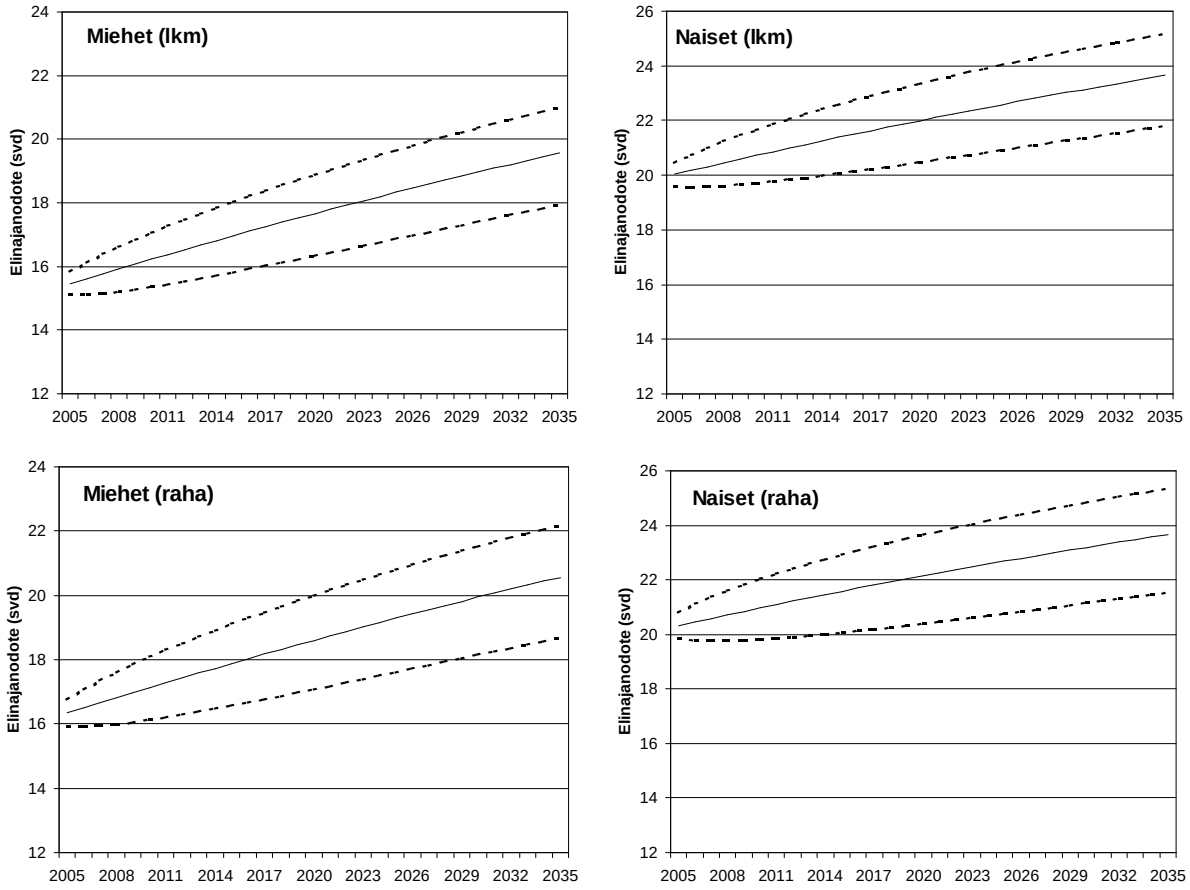
Taulukko 4: 65-vuotiaiden elinajanodotteiden ennusteet vuosille 2015, 2025 ja 2035

Taulukosta 4 voidaan havaita, että eri menetelmät antavat hyvin samantyyppisen ennusteen tulevaisuudesta. On mielenkiintoista havaita, että ennusteen mukaan naisten ja miesten elinajanodotteet lähestyvät tulevaisuu-

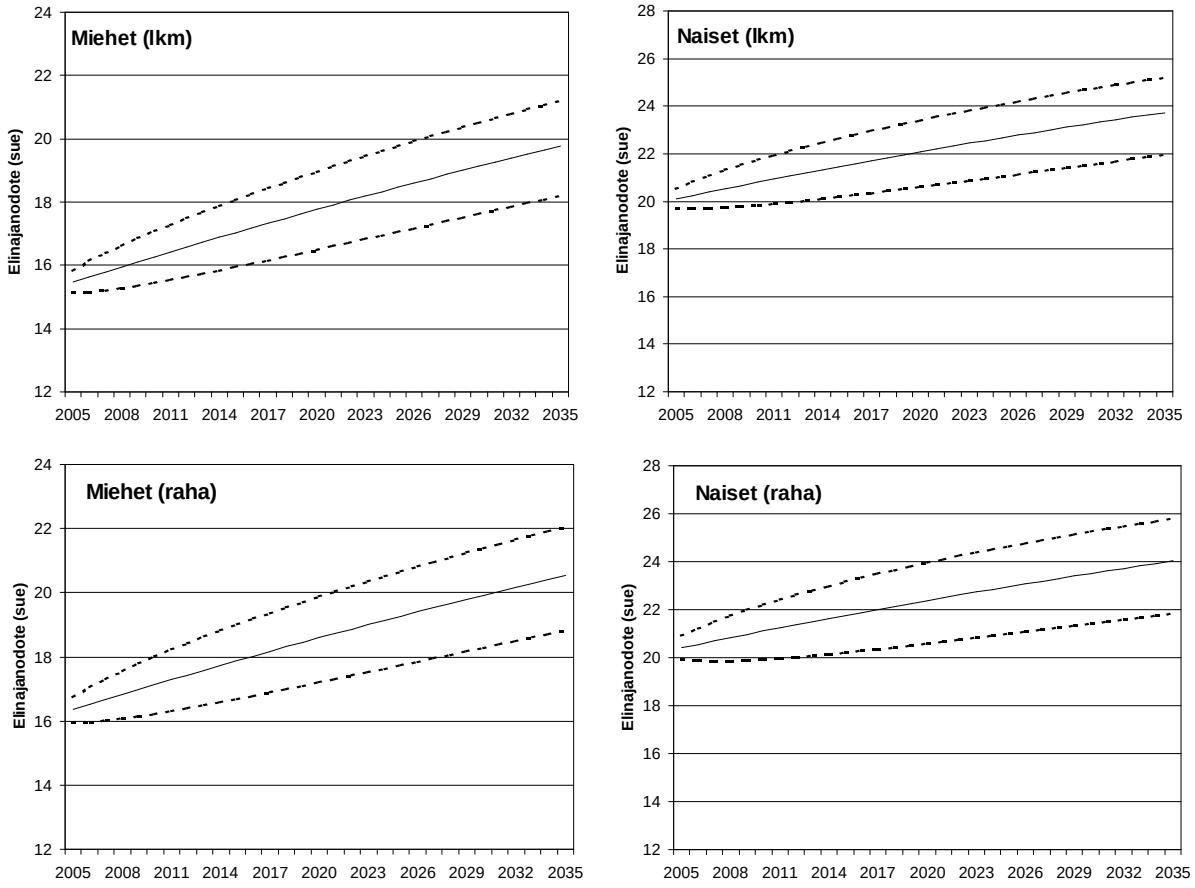


Kuva 8: PNS-menetelmällä tehty elinajanennuste, katkoviivoilla 95%:n luottamusvälit.

dessa toisiaan. Vuonna 2015 naisten elinajanodote on noin $4,45$ vuotta suurempi kuin miesten, mutta vuonna 2035 ero on supistunut noin $4,05$ vuoteen.



Kuva 9: SVD-menetelmällä tehty elinajanennuste, katkoviivoilla 95%:n luottamusvälit.

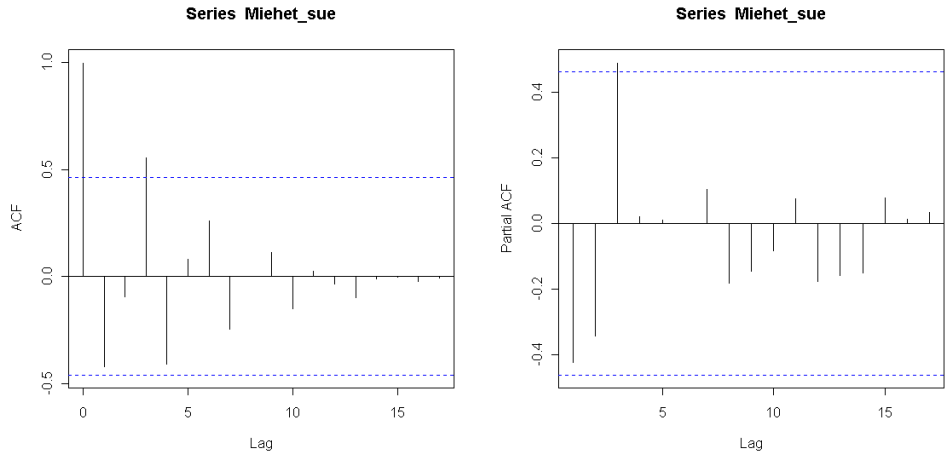


Kuva 10: SUE-menetelmällä tehty elinajanennuste, katkoviivoilla 95%:n luottamusvälit.

5 Toisenlaisen ARIMA-mallin etsiminen

Mallin etsinnässä voimme käyttää hyväksi autokorrelaatiofunktion ja osittais-autokorrelaatiofunktion kuvaajia. Seuraavaksi etsimme miesten (lkm) suurimman uskottavuuden estimaatilla saamallemme vektorille $\hat{k}_t$ toisenlaisen ARIMA-mallin.

Aluksi differoimme sarjan $\hat{k}_t$ viiveellä yksi ja piirrämme differoidun sarjan ACF:n ja PACF:n kuvaajat. Kuten kuvasta 11 näemme, sekä ACF:n että PACF:n arvot ovat melko hyvin 95%:n luottamusvälien sisällä.



Kuva 11: ACF:n ja PACF:n kuvaajat.

Näistä kuvaajista voimme päätellä, että MA(3)- tai AR(3)-mallia voisi sovittaa aineistoon, sillä nyt sekä ACF:n että PACF:n arvot menevät viiveellä kolme yli 95%:n luottamusvälin. Voisimme myös sovittaa aineistoon ARIMA(3,1,3)-mallia, mutta silloin mallissa olisi turhan paljon parametreja. Teemme sovitteen AR(3)-mallilla, jolloin mallimme on muotoa

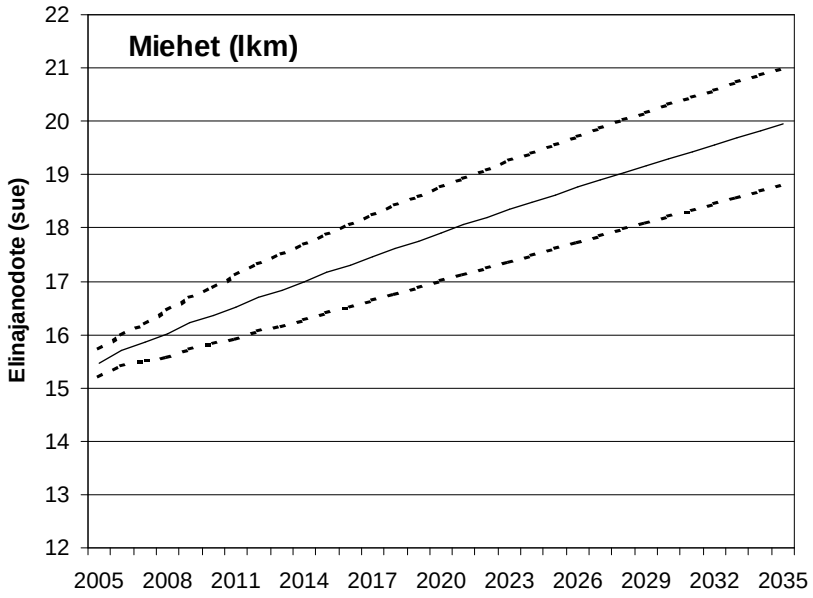
$$k_t = \phi_1 k_{t-1} + \phi_2 k_{t-2} + \phi_3 k_{t-3} + \epsilon_t,$$

jossa $\epsilon_t \sim WN(0, \sigma_t^2)$ ja kertoimet ovat

$$k_t = -0,4034 k_{t-1} - 0,0639 k_{t-2} + 0,4908 k_{t-3} + \epsilon_t.$$

Tekemällä ennusteen vektorille $\hat{k}_t$ käyttäen ARIMA(3,1,0)-mallia voimme laskea elinajanodotteen ja 95%:n luottamusvälin ennustetulle elinajanodotteelle. Tulos sinänsä ei poikkea juurikaan ARIMA(0,1,0)-mallilla tehdystä

ennusteesta. Molemmat mallit antavat noin 20 vuoden elinajanodotteen 65-vuotiaille vuonna 2035. Sen sijaan 95%:n luottamusväli on nyt hieman kapeampi vuonna 2035.



Kuva 12: ARIMA(3,1,0)-mallilla piirretty elinajanennuste, katkoviivoilla 95%:n luottamusvälit.

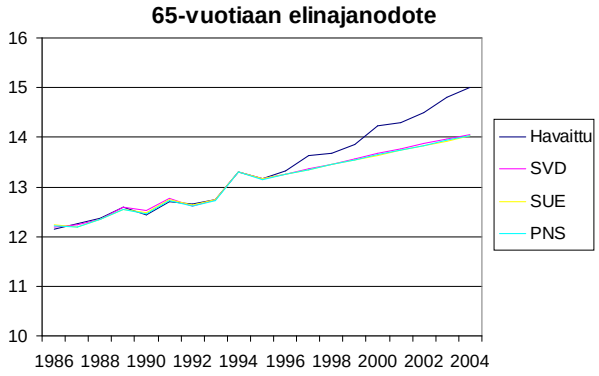
6 Menneestä ajasta tulevaisuuteen

Expost-ennusteella tarkoitus on tutkia, olisiko menneisyydessä tehty ennuste toteunut nykypäivänä. Aineistossamme on havaintovuosia vain 19, joten sovellamme Leen-Carterin mallia vuosiin 1986–1995 ja sen jälkeen teemme ennusteen vuosille 1996–2004. Lisäksi oletamme, että kaikki ihmiset kuolevat täytettyään 90 vuotta. Joudumme tekemään näin, sillä aineistossamme ei ole kuolleisuuslukuja kuin 89. ikävuoteen asti. Emme siis pysty täysin vertaamaan Expost-ennusteen tuloksia virallisiin elinajanodotelukuihin, mutta voimme laskea alkuperäisestä ainestostamme elinajanodotteet ja tutkia, olisiko vuonna 1996 tehty ennustus ollut hyvä. Käytämme aineistona miesten henkikuolleisuuslukuja. Elinajanodotteet laskemme luvussa 4.4 esitetyllä tavalla.

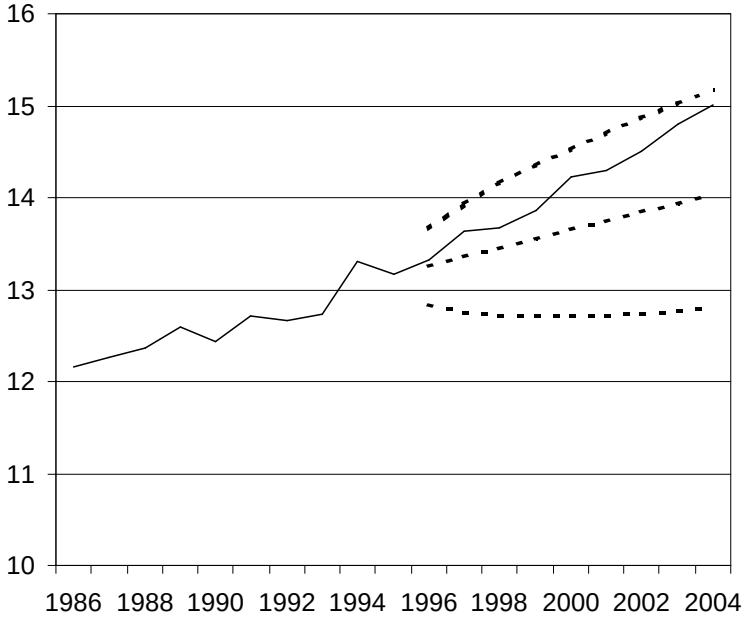
Kuvassa 14 on aluksi sovellettu pienimmän neliösumman menetelmää. Sen jälkeen vektorille k_t on tehty lineaarinen ennuste ARIMA(0,1,0)-mallilla. Kuvassa 15 on puolestaan käytetty ARIMA(1,1,0)-mallia ennustetta tehtäessä. Molempiin kuviin on lisäksi piirretty 95%:n luottamusvälit. Luottamusvälit on muodostettu vektorin k_t hajonnan avulla. Kuten havaitsemme, ARIMA(0,1,0)-malli olisi ollut tuolloin parempi valinta, sillä toteutunut elinajanodote sisältyy luottamusväliin. Toisaalta ARIMA(1,1,0)-mallilla tehty luottamusväli on hieman pienempi kuin ARIMA(0,1,0)-mallilla muodostettu.

Elinajanodotteen kasvu on ollut melko ripeää vuosina 1986–2004. Kun vuonna 1986 65-vuotiaan elinajanodote oli $12,15$ vuotta, niin vuonna 2004 65-vuotiaan elinajanodote oli jo $15,01$ vuotta. Keskimääräinen vuosilisäys on ollut noin $0,1586$ vuotta eli noin $1,90$ kuukautta. Oletuksena siis oli, että kaikki kuolevat viimeistään täytettyään 90 vuotta.

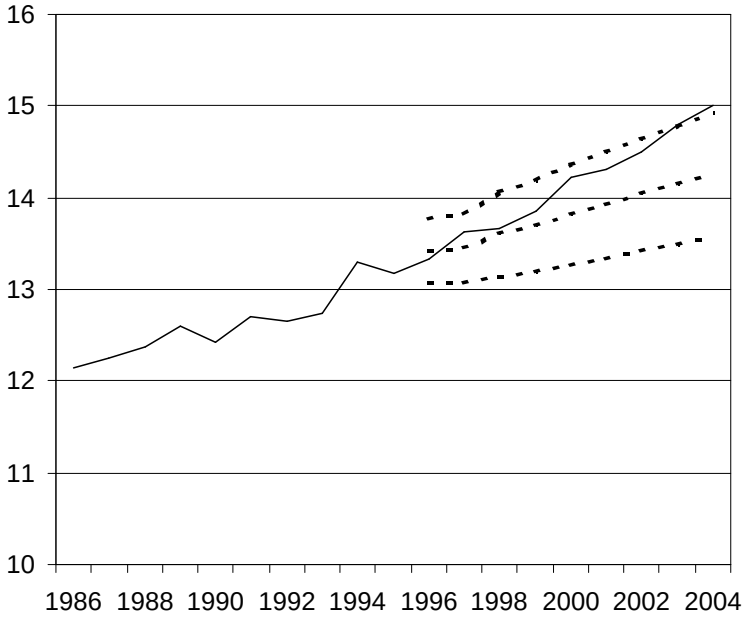
Elinajanodotteen kasvu on ollut lähes lineaarista vuosina 1986–1995, kuten voimme havaita kuvasta 13. Tätä samaa lineaarisuutta jatkaa myös ennusteemme. Todellisuudessa kuitenkin vuoden 1996 jälkeen elinajanodotteen keskimääräinen vuosilisäys on ollut suurempaa kuin vuosien 1986–1995 datasta voisi ennustaa. Tästä syystä lineaarinen ennusteemme jää reilusti havaitusta elinajanodotteesta. Toisaalta ennuste on tehty vain 10 vuoden havaintojen perusteella, joten kaiken kaikkiaan voidaan sanoa, että ennuste on onnistunut.



Kuva 13: Elinajanodote 65-vuotiaalle, kaikki ennusteet tehty ARIMA(0,1,0)-mallilla.



Kuva 14: Miesten SUE:lla tehty ARIMA(0,1,0)-ennuste ja 95 %:n luottamusvälit ennusteelle.



Kuva 15: Miesten SUE:lle ARIMA(1,1,0)-ennuste ja 95 %:n luottamusvälit ennusteelle.

7 Pohdintaa ennustamiseen liittyvästä virheestä

7.1 Ennustamisessa tapahtuva virhe

Ennustamamme vuoden s logaritmoidun kuolleisuusluvun ja toteutuneen kuolleisuusluvun välinen yhteys voidaan ilmaista yhtälöllä

$$\ln(q_{x,s}) = (\hat{a}_x + \alpha_x) + (\hat{b}_x + \beta_x)(\hat{k}_s + u_s) + \epsilon_{x,s}, \tag{7.1}$$

jossa

- $\hat{a}_x$ ja $\hat{b}_x$ ovat estimoidut a_x ja b_x,
- $\hat{k}_s$ on ennustettu k_s,
- α ja β ovat vektoreiden a_x ja b_x estimoinnissa tapahtuneet virheet,
- u_s on vektorille k_t tehdyn ennusteen virhe vuonna s.

Muodostamme kuolemanvaaraluvut kaavalla

$$\widehat{\ln(q_{x,s})} = \hat{a}_x + \hat{b}_x\hat{k}_s. \tag{7.2}$$

Kokonaisvirhe $E_{x,s}$ ennustetulle logaritmoidulle kuolleisuusluvulle voidaan nyt laskea yhtälöiden (7.1) ja (7.2) erotuksena, jolloin

$$\begin{aligned} E_{x,s} &= (\hat{a}_x + \alpha_x) + (\hat{b}_x + \beta_x)(\hat{k}_s + u_s) + \epsilon_{x,s} - (\hat{a}_x + \hat{b}_x\hat{k}_s) \\ &= \alpha_x + \epsilon_{x,s} + (\hat{b}_x + \beta_x)(\hat{k}_s + u_s) - \hat{b}_x\hat{k}_s \\ &= \alpha_x + \epsilon_{x,s} + u_s(\hat{b}_x + \beta_x) + \beta_x\hat{k}_s. \end{aligned} \tag{7.3}$$

Voimme siis kirjoittaa oikean ja ennustetun logaritmoidun kuolleisuusluvun seuraavanlaisesti

$$\ln(q_{x,s}) = \widehat{\ln(q_{x,s})} + E_{x,s},$$

joten pienille kokonaisvirheille $E_{x,s}$ on voimassa seuraava yhtälö,

$$\begin{aligned} q_{x,s} &= \widehat{q_{x,s}} * \exp(E_{x,}) \\ &\approx \widehat{q_{x,s}}(1 + E_{x,s}). \end{aligned}$$

Näin ollen

$$q_{x,s} - \widehat{q_{x,s}} = \widehat{q_{x,s}} E_{x,s}.$$

Toisin sanoen kuolleisuusluvun $q_{x,s}$ virhe on $\widehat{q_{x,s}} E_{x,s}$, jossa $E_{x,s}$ on logaritmoidun kuolleisuusluvun $\ln(q_{x,s})$ virhe. Lee ja Carter merkitsivät termiä $\widehat{q_{x,s}} E_{x,s}$ muuttujalla $\theta_{x,s}$ ja väittivät, että mikäli yliarvioidaan kuolleisuutta $q_{x,s}$ termin $\theta_{x,s}$ verran se vähentää elinajanodotetta ajanhetkellä s termin $\theta_{x,s} T_{x,s}$ verran. Muuttuja $T_{x,s}$ edustaa elinaikataulussa henkilön elämiä vuosia. Kts. [1, s. 22].

Oletetaan, että virhetermit eivät korreloi keskenään. Tällöin Lee ja Carter näyttivät, että elinajanodotteen varianssi ennustettaessa s vuotta eteenpäin voidaan ilmaista yhtälöllä

$$\begin{aligned} Var(\hat{e}_0) &= \sigma_{us}^2 (\sum b_x \hat{q}_{x,s} T_{x,s})^2 \\ &\quad + \sum (\hat{q}_{x,s} T_{x,s})^2 \times (\sigma_{\alpha x}^2 + \sigma_{\varepsilon x,s}^2 + \sigma_{\beta x}^2 k_s^2 + \sigma_{\beta x}^2 \sigma_{us}^2). \end{aligned} \tag{7.4}$$

Yhtälössä (7.4) termit α_x, β_x ja k_t ovat termien $\hat{a}_x$, $\hat{b}_x$ ja $\hat{k}_t$ virheitä ja ne voidaan arvioida numeerisesti.

Parametrin $\hat{a}_x$ virhe

Parametrin $\hat{a}_x = \ln(q_{x,t})/T$ estimoinnissa tapahtunut virhe α_x voidaan määrittää kaavalla

$$Var(\alpha_x) = \frac{Var \ln(q_{x,t})}{T},$$

jossa T on aika-akselin pituus.[1, s. 42]

Parametrin $\hat{b}_x$ virhe

Parametrin $\hat{b}_x$ estimoinnissa tapahtunut virhe β_x onkin huomattavasti vaikeammin estimoitavissa. Voimme sanoa, että parametrin b_x estimoinnissa tapahtuneen virheen vaikutus on todella pieni kokonaisvirheeseen verrattuna. Unohdamme siis kokonaan parametriin $\hat{b}_x$ kohdistuvan virheen. [1, s.42]

Parametrin $\hat{k}_t$ virhe

Voimme helposti todistaa, että ennustettaessa s vuotta eteenpäin parametria $\hat{k}$, niin parametrin $\hat{k}$ keskivirhe kasvaa ennustettavien vuosien s neliöjuuren suhteessa.

Todistus. Oletetaan, että parametri $\hat{k}$ on mallinnettu satunnaiskävelyllä, siis

$$\hat{k}_t = \hat{k}_{t-1} + c + \varepsilon_t. \tag{7.5}$$

Asetetaan $k_0 = \hat{k}_t$, jossa $\hat{k}_0$ on estimoimamme vektorin $\hat{k}_t$ viimeisin arvo. Nyt yhden vuoden ennusteen keskivirhe on

$$\begin{aligned} se_1 &= std(k_1 - \hat{k}_1) \\ &= std(k_0 + c + \varepsilon_1 - \hat{k}_0 - \hat{c}) = se(\varepsilon_1). \end{aligned}$$

Saamme vuoden n kuluttua keskivirheeksi

$$\begin{aligned} se_n &= std(k_n - \hat{k}_n) \\ &= \sqrt{n} * std(\varepsilon_1). \end{aligned}$$

□

[1, s. 42]

7.2 Simuloinnit

Seuraavaksi tutkimme vektorille k_t tehdyn ennusteen luottamusvälin riittävyyttä. Käytännössä simuloimme vektorille ennusteen kaavalla

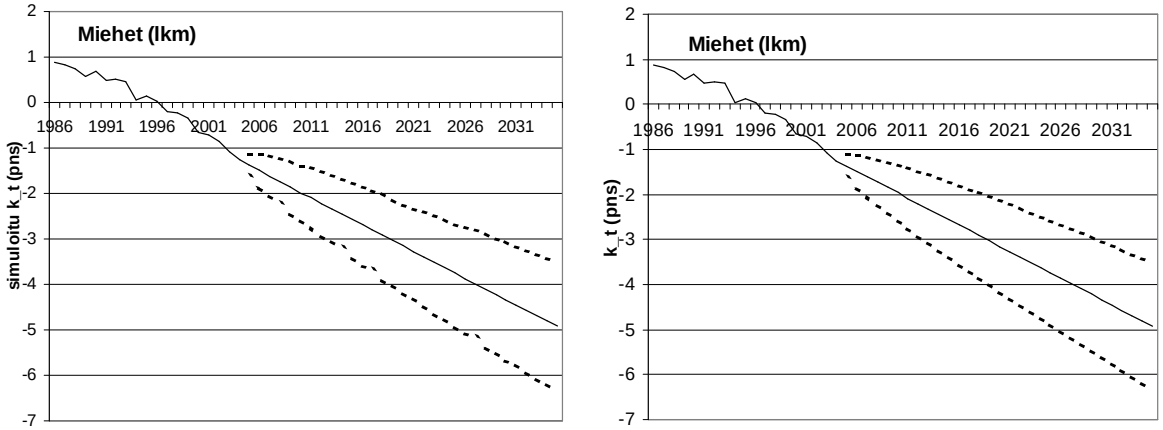
$$k_{t+1} = k_t + \hat{c} + \varepsilon_t,$$

jossa $\hat{c} = E(\nabla \hat{k}_t)$ ja $\varepsilon_t \sim N(0, \sigma^2)$. Suurin ongelma on muuttujan ε_t varianssin määrittäminen.

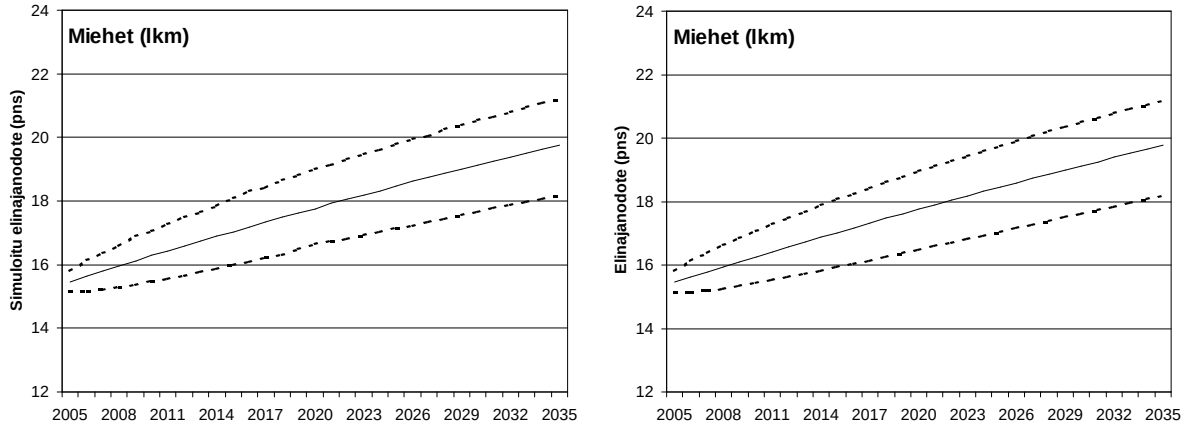
Esimerkkinä otamme miesten (pns) henkikuolleisuusaineistosta saadun vektorin k_t ennusteen simuloinnin. Käytämme varianssina lukua $0,13$, joka on päätelty vektorin $\nabla \hat{k}_t$ varianssista. Simuloimme mallin tuhat kertaa ja vertaamme tuloksia aiemmin muodostamiimme luottamusväleihin.

Lajittelemme sarjat vuosikohtaisesti ja piirrämme samaan kuvaan luottamusvälien ylärajan, alarajan sekä keskiennusteen. Muodostamme 95% luottamusvälin poistamalla jokaisesta vuodesta 25 suurinta ja 25 pienintä arvoa.

Emme siis välttämättä poista samoja sarjoja joka vuosi. Kuvasta 16 näemme, että simuloitu ennuste muistuttaa paljon ARIMA(0,1,0)-mallilla tehtyä ennustetta. Simuloinnin perusteella voimme pitää ennustetulle vektorille k_t käyttämiämme luottamusvälejä riittävinä. Kuvassa 17 on piirretty simuloinnin ja ARIMA(0,1,0)-mallin avulla saadut elinajanodotteet. Elinajanodotteetkin ovat nyt lähes yhtenäiset. Näin pitää ollakin, sillä ennustettaessa elinajanodotetta käytämme ainoastaan vektorin $\hat{k}_t$ ennustettuja arvoja, muut parametrit tulevat alkuperäisestä mallista.



Kuva 16: Simuloinnilla saadut luottamusvälit ja ARIMA(0,1,0)-mallilla saadut luottamusvälit.



Kuva 17: Simuloinnilla saadut luottamusvälit ja ARIMA(0,1,0)-mallilla saadut luottamusvälit.

Viitteet

[1] Marie-Claire Koissi. *Fitting and Forecasting Mortality Rate with the Lee-Carter Model*, Lic. Thesis Åbo Akademi University, 2003.

[2] Mika Mäkinen. *Referenssikuolevuus henkivakuutusyhtiöille*, SHV - harjoitustyö 26.2.2004.

[3] Arto Luoma, *Aikasarja-anyysi I*, Tampereen yliopisto, 2005, URL *http://www.uta.fi/ al18853/luentod.pdf*

[4] Roger A. Horn and Charles R. Johnson. *Matrix analysis* Cambridge University Press, 1985.

[5] Nico Keilman and Dingh Quang Pham, *Prediction intervals for Lee-Carter-based mortality forecasts*, June 2006, URL *http://epc2006.princeton.edu/download.aspx?submissionId=60211 #search=%22prediction%20intervals%20for%20lee-carter-based%22*