Pekka Hjelt

Aikasarjamallit

Aikasarja koostuu järjestyksessä olevista havainnoista, ja yleensä se on tasavälinen ja diskreetti eli havaintopisteet ovat erillisiä. Lisäksi aikasarjassa on yleensä autokorrelaatiota eli havaintojen keskinäistä riippuvuutta, minkä vuoksi otoksiin tarkoitetut tilastolliset menetelmät eivät sovellu aikasarjojen käsittelyyn.

G.E.P. Box ja G.M. Jenkins julkaisivat 1970 kirjan ARIMA-malleista, ja se aloitti uuden aikakauden aikasarja-analyysissa. Heidän esittämänsä mallit oli tunnettu jo vuosikymmeniä, mutta vasta nyt saatiin systemaattinen kokonaisesitys siitä, miten nämä mallit tunnistetaan, estimoidaan niitten tuntemattomat parametrit, testataan mallin kelvollisuus ja laaditaan ennusteita. Tietokoneiden tuo yleiseen käyttöön mahdollisti menetelmien soveltamisen käytäntöön.

ARIMA-malleissa lähdetään siitä, että havaitun aikasarjan on synnyttänyt jokin stokastinen eli satunnaisprosessi. Näitä prosesseja on kolmea tyyppiä: autoregressiiviset (AR), liukuvakeskiarvo- (MA) ja yhdistetyt autoregressiiviset liukuvakeskiarvoprosessit (ARMA). Ennalta ei tiedetä, mistä prosessista on kyse, mutta kun jokaisella tyypillä on oma autokorrelaatiorakenteensa, aikasarjan havaitut autokorrelaatiot antavat lähtökohdan mallin tunnistamiseen. Kun malliin sisältyvät parametrit on estimoitu, pystytään tilastotieteen teorian keinon päättelemään, kelpaako valittu malli. Jos se ei kelpaa, kokeillaan uutta mallia niiden tietojen pohjalta, jotka saadaan ensimmäisen mallin tuloksista. Näin jatketaan, kunnes löydetään sopiva malli.

Mahdollisia malleja ei käytännössä ole kovin runsaasti, koska ehdottomana periaatteena on, että valitaan **niukkaparametrinen, riittävä malli.** Ei siis tavoitella kuten eräillä aloilla täydellistä mallia. Tähän on ainakin kolme pääsyytä:

1) tieteessä yleisen taloudellisuuden ja varovaisuuden periaatetta noudattaen tyydytään mieluummin yksinkertaiseen kuin monimutkaiseen selitykseen;
2) koska parametrien estimointi on varsin vaativa matemaattinen tehtävä, se onnistuu parhaiten, kun estimoitavia parametreja ei ole kovin monta;
3) tulevaisuuden ennustaminen tapahtuu parhaiten yksinkertaisella mallilla.

ARIMA-malleja voi käyttää vain stationaarisiin aikasarjoihin. Stationaarisuutta voidaan kuvata siten, että aikasarjan keskiarvo ja varianssi ovat likimain samat aikavälin jokaisessa kohdassa. Koska etenkin taloudelliset aikasarjat ovat yleensä epästationaarisia ”trendin” vuoksi, ne on ennen mallin sovittamista stationarisoitava. Keskiarvon osalta se tapahtuu käyttämällä havaintojen differenssejä eli erotuksia ja varianssin stationarisointi tapahtuu yleensä logaritmimuunnoksen avulla.

ARIMA(p,d,q) tarkoittaa mallia, jossa on p kappaletta autoregressiiviparametreja, q liukuvakeskiarvoparametria ja d ilmoittaa differenssiasteen. Yleensä nämä luvut ovat pieniä, 0 ja 1 ovat tavallisia, d=2 harvinainen ja 3:ea suurempi p:n tai q:n arvo ei esiinny kovinkaan usein. Lisäksi mukana voi olla ns. vakiotermi, jonka olemassaolo liittyy ”trendiin”.

Monia ilmiöitä havainnoidaan useammin kuin kerran vuodessa, jolloin aikasarjassa ilmenee kausivaihtelua. Se voidaan mallintaa samalla tavalla kuin ns. perusvaihtelukin, ja ne yhdistetään yleiseksi kerrannaiseksi kausivaihtelumalliksi ARIMA(p,d,q)$\times$(P,D,Q)$_s$. s tarkoittaa kausivaihtelujakson pituutta (kuukausisarjassa 12, neljännesvuosisarjassa 4). Täydellinen malli voidaan kirjoittaa:

$$(1-B)^d(1-B^s)^D Y_t = C+[(1-\theta_1 B-...-\theta_q B^q)(1-\Theta_1 B-\ldots-\Theta_Q{}^{Qs})] / [(1-\phi_1 B-\ldots-\phi_p B^p)(1-\Phi_1 B-...-\Phi_P{}^{Ps})]a_t.$$

Siinä $(1-B)^d$ tarkoittaa d:ttä differenssiä, $(1-B^s)^D$ D:ttä kausidifferenssiä; d ja D ovat yleensä 0 tai 1. Vastaavasti B:t ovat viiveoperaattoreja eli $BY_t = Y_{t-1}$, $B^pY_t = Y_{t-p}$.
Olennaisia päättelyn kannalta ovat ns. **vakio C, liukuvakeskiarvoparametrit $\theta_1,\ldots,\theta_q$, $\Theta_1,\ldots,\Theta_Q$, autoregressiiviset parametrit $\phi_1,\ldots,\phi_p$, $\phi_1,...,\phi_P$ sekä jäännöshajonta σ_a.** Versaalit tarkoittavat aina kausivaihteluosaa, gemenat ns. perusosaa. Kun ARIMA-malli esitetään edelliseen tapaan osamäärämuotoisena eli molemmat puolet jaetaan autoregressiiviset termit sisältävällä hakasulkulausekkeella, **C ilmoittaa suoraan aikasarjan trendin.** Ellei sen estimaatti ole tilastollisesti merkitsevä ja d = 0 ja D = 0, C:n arvo on likimain sama luin sarjan keskiarvo, jonka ympärillä heilahtelu tapahtuu. Jos d =1 ja D = 0, C antaa keskimääräisen muutoksen kuukaudesta tai vuosineljänneksestä toiseen, ja jos d = 0 ja D = 1, C ilmoittaa keskimääräisen vuosimuutoksen. Muut tapaukset ovatkin harvinaisempia. Ilmastoaikasarjat ovat joko jo alun alkaen stationaarisia eli differenssit nollia tai tarvitaan kausidifferenssiä (D = 1), ja jos siinä tilanteessa C poikkeaa merkitsevästi nollasta, sarjalla on jokin suunta.

Mallin valinta on iteratiivinen prosessi. Ensimmäisessä vaiheessa tutkitaan aikasarjan autokorrelaatiorakennetta, sillä jokaisella kyseeseen tulevalla stokastisella prosessilla on tyypillinen autokorrelaatiorakenteensa. Koska kyse on empiirisistä havainnoista, sen havaitseminen ei ole aina helppoa, mutta jokin alustava malli voidaan aina löytää. Vanhemmista menetelmistä poiketen valinta ei ole siis mielivaltainen, vaan sarjan oma historia on aina lähtökohtana. Parametrien estimoinnin jälkeen nähdään, täyttääkö malli vaaditut ehdot, ja ellei täytä, muutetaan mallia ensimmäisten tulosten viitoittamaan suuntaan. Monissa nykyisissä tietokoneohjelmissa mallin tunnistaminen ja alustava estimointi on pitkälle automatisoitu, mutta tulos ei ole välttämättä hyväksyttävä.

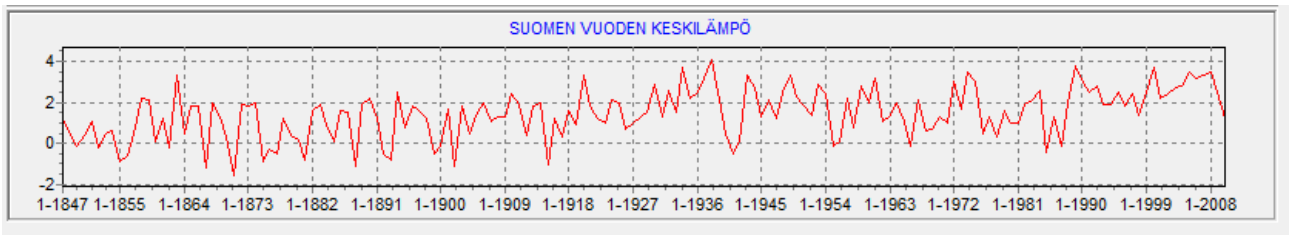
Valinnassa käytetään useita tilastotieteen teoriaan pohjautuvia kriteerejä:

- AR-kerrointen on täytettävä stationaarisuusehdot, MA-kerrointen käännettävyysehdot;
- kerrointen väliset korrelaatiot eivät saa olla liian suuria; usein käytetään vaatimusta, että niiden itseisarvo on alle puolen;
- kerrointen on oltava tilastollisesti merkitseviä; todennäköisyystasona käytetään 0,95:tä tai tätä korkeampaa arvoa, jos aikasarja on pitkä;
- jäännössarjan on oltava riittävästi satunnainen eli siinä ei saa esiintyä merkitseviä autokorrelaatioita ainakaan tärkeillä viiveillä (1 ja 2; s ja 2s), mutta muuten huomiota ei kiinnitetä niinkään yksittäisiin kertoimiin, vaan koko rakenteeseen;
- jäännössarjan keskiarvo ei saa poikkea merkitsevästi nollasta:
- näiden lisäksi voidaan kiinnittää huomiota myös jäännössarjan normaalisuuteen ja sen etumerkkien satunnaisuuteen ym., mutta usein tällaiset tarkastelut sivuutetaan.

Parametrien estimointi on vaativa matemaattinen minimointitehtävä, joka edellyttää tietokoneen käyttöä. Käytössä on jossain määrin toisistaan poikkeavia menetelmiä, jotka voivat johtaa hiukan erilaisiin tuloksiin. Sen lisäksi kerrointen etumerkit voivat olla vastakkaiset menetelmästä riippuen, mutta analyysin tärkeimmässä tehtävässä eli ennustamisessa ne tietenkin toimivat oikein.

SUOMEN KESKILÄMPÖ

Vuosisarja 1847 - 2010



Riittävä malli on yksinkertainen ARIMA(0,1,1) eli IMA(1,1). Stationarisointi edellyttää siis ensimmäistä differenssiä. Estimaatit ovat:

$\theta_1 = 0{,}91$, jonka keskivirhe on 0,33 eli se on sekä tilastollisesti merkitsevä että täyttää käännettävyysehdon (estimaatti + 2 kertaa keskivirhe on pienempi kuin 1). Jäännössarjan keskiarvo ei poikkea merkitsevästi nollasta (t = 1,44), jäännöshajonta $\sigma_a = 1{,}10$ eikä malliin sisälly vakiotermiä C. Tämä tarkoittaa, ettei sarjalla ole trendiä, vaan kaikki *ennusteet vuodesta 2011 alkaen ovat 2,49 astetta.* Kun jäännössarjassa ei ole myöskään merkitsevää autokorrelaatiota, tämä malli osoittautuu riittäväksi.

Vuosisarja 1909 - 2010

Malliksi saadaan ARIMA(1,0,0) eli AR(1). Differenssiä ei ole. Estimointitulokset:

$\phi_1 = 0{,}29$, **keskivirhe 0,095** eli estimaatti on merkitsevä ja täyttää stationaarisuusehdon, joka on ihan samanlainen kuin MA-kertoimen kohdalla edellä. $C = 1{,}84$, jonka ympärillä sarja siis vaihtelee ja joka on myös ennuste tulevaisuuteen. Jäännöshajonta $\sigma_a = 1{,}03$ astetta, jäännöskeskiarvo ei merkitsevä (t = 0,51). Kaikki jäännösautokorrelaatiokertoimet ovat pieniä. Siis: Suomen keskilämpö ei osoita viimeisen sadan vuoden ajalta kehitystä mihinkään suuntaan. Tulos on oleellisesti sama, vaikka jätän pois kylmät vuodet 2009 ja 2010.

Vuodenaikasarja 1847 - 2011

Tämän 656 havainnon aikasarjan malliksi tulee ARIMA(1,0,0)×$(0,1,1)_4$, siis kausidifferenssi mukana ja kaksi kerrointa. $\phi_1 = 0{,}22$, $\Theta_1 = 0{,}95$. Molemmat ovat tilastollisesti merkitseviä ja täyttävät stationaarisuus- ja käännettävyysehdon. Tähänkään malliin ei tule vakiotermiä, mutta jäännöskeskiarvo on aika lähellä merkitsevyysrajaa (t = 1,91). Jäännöshajonta $\sigma_a = 1{,}82$ on tietenkin suurempi kuin vuosisarjojen kohdalla, koska lämpötilan vaihtelu vuodenaikojen kesken on paljon laajempaa. Autokorrelaatiokertoimet täyttävät satunnaisuusvaatimuksen, joskin viiveen 3 kerroin ylittää merkitsevyysrajan, vaikka sen arvo onkin vain 0,085. Pitkissä aikasarjoissa ilmenee helposti merkitsevyyksiä, vaikka absoluuttiset luvut ovatkin pieniä.

Vuodenaikaennusteet tulevaisuudessa ovat

talvi -8,5 **kevät** 1,0 **kesä** 14,0 **syksy** 2,6 astetta.

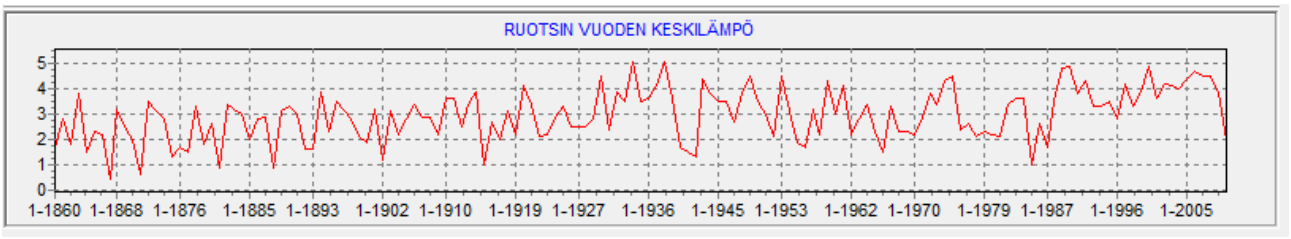
Talvi sisältää edellisen vuoden joulukuun ja tilastovuoden tammi- ja helmikuun, kevät maalis-, huhti- ja toukokuun, kesä kesä-, heinä- ja elokuun sekä syksy syys-, loka- ja marraskuun.

Jos malliin ”pakotetaan” vakio C, sen arvoksi tulee 0,012 eli hiukan enemmän kuin asteen sadasosa, ja se on tilastollisesti merkitsevä. Kertoimet pysyvät liki samoina, mutta kausivaihtelun MA-kerroin kohoaa hyvin lähelle käännettävyysehdon rajaa. Jäännöshajonta on 1,81 ja jäännössarja liki samanlainen kuin edellä, ja ennusteet osoittavat hidasta suuntaa lämpimämpään.

Vuodet 1909 - 2011

Malli ARIMA(1,0,0)×(1,1,1)$_4$, kertoimet ϕ_1 = ***0,18,*** $\phi_1 = 0,14, \Theta_1$ = ***0,98,*** jotka täyttävät muut ehdot, mutta kausivaihtelun MA-kerroin ylittää käännettävyysrajan. Vakiokerroin ei ole merkitsevä eikä jäännöskeskiarvo (t = 1,52) myöskään, ja autokorrelaatiokertoimet riittävän pienet. Jäännöshajonta on ***1,77.*** Tämä on paras vähäparametrinen, jonka vikana on tuo MA-kerroin. Sen korkea arvo kaikissa Suomen ilmastomalleissa johtuu sään voimakkaasta kausiriippuvuudesta. Vuodenaikaennusteet ovat kauttaaltaan 2 - 4 kymmenesosa-astetta alemmat kuin koko jakson mallissa.

RUOTSIN KESKILÄMPÖ 1860 - 2010

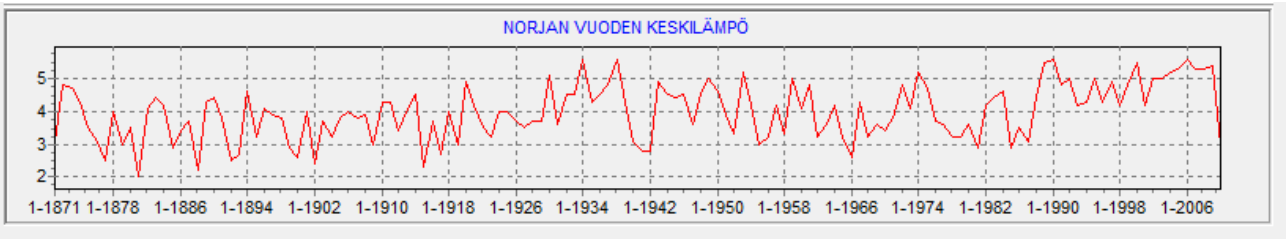


ARIMA(0,1,1); θ_1 = ***0,94,*** täyttää ehdot. Vakio ***C = 0,011*** ja juuri ja juuri merkitsevä (t = 2.08). Jäännöshajonta ***0,93*** astetta, jäännöskeskiarvo ei ole merkitsevä (t = -0,87), jäännösautokorrelaatiot riittävän pienet. Ennuste vuodeksi 2011 on ***3,82*** astetta ja se kasvaa jokseenkin yhden sadasosa-asteen vuodessa, joten mitään äkillistä ja valtavaa ilmastonmuutosta ei ole näkyvissä.

Vuodet 1909 - 2010

ARIMA(1,0,0), ϕ_1 **= 0,32,** täyttää ehdot. ***C* = *3,2,*** ja koska ei ole differenssejä, se ilmoittaa keskiarvotason ja samalla vakioennusteen. Jäännöshajonta on ***0,93*** astetta, jäännössarjan keskiarvo ei poikkea nollasta (t = 1,03), jäännösautokorrelaatiokertoimet ovat pieniä.

NORJAN KESKILÄMPÖ 1871 - 2010

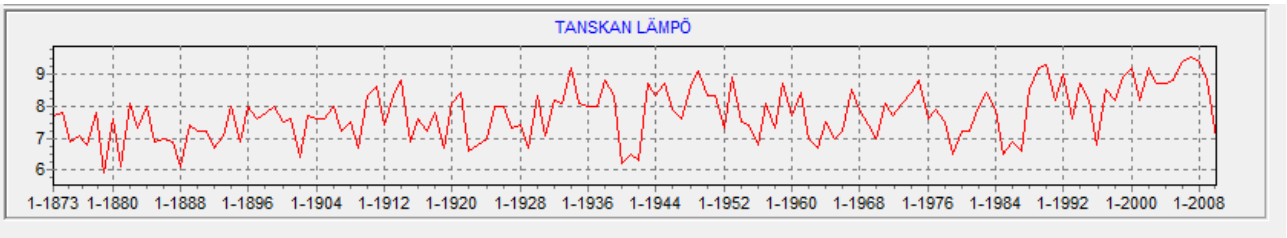


ARIMA(0,1,1), ***θ_1* = *0,87,*** täyttää ehdot. Vakio ei ole merkitsevä (t = 1,17), jäännöskeskiarvo ei myöskään (t = 0,88). Jäännöshajonta on ***0,77*** astetta ja jäännösautokorrelaatiot matalia. Ennuste on koko ajan ***4,8*** astetta.

Vuodet 1909 - 2010

ARIMA(1,0,0), ***ϕ_1* = *0,36,*** täyttää ehdot. ***C* = *4,1,*** ja koska ei ole differenssejä, se ilmoittaa keskiarvotason ja samalla vakioennusteen. Jäännöshajonta on ***0,77*** astetta, jäännössarjan keskiarvo ei poikkea nollasta (t = 1,40), jäännösautokorrelaatiokertoimet ovat riittävän matalia.

TANSKAN KESKILÄMPÖ 1873 - 2010



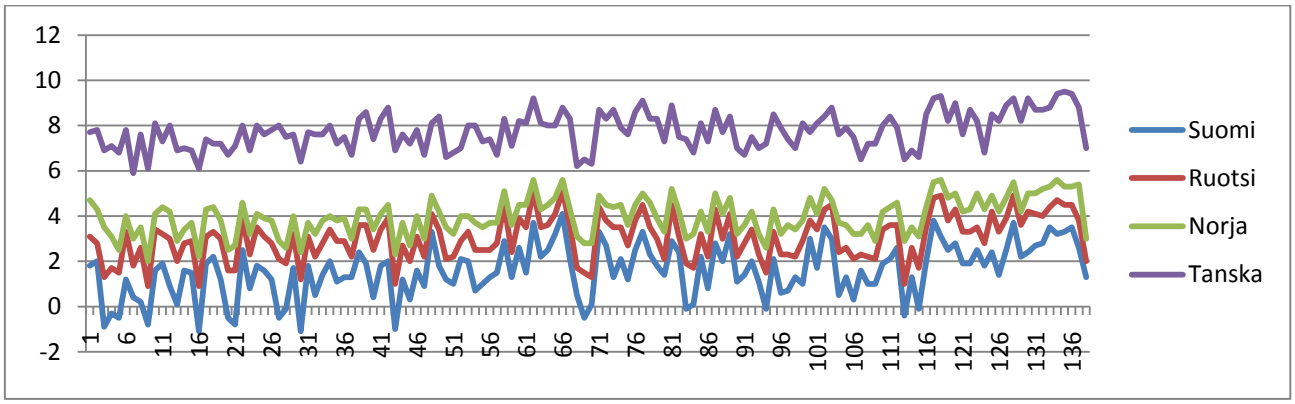
ARIMA(0,1,1), ***θ_1* = *0,82,*** täyttää ehdot. Vakio ei ole merkitsevä (t = 1,05), jäännöskeskiarvo ei myöskään (t = 0,74). Jäännöshajonta on ***0,72*** astetta ja jäännösautokorrelaatiot matalia. Ennuste on koko ajan ***8,6*** astetta.

Vuodet 1909 - 2010

ARIMA(1,0,0), $\phi_1 = 0{,}36$, täyttää ehdot. $C = 7{,}9$, ja koska ei ole differenssejä, se ilmoittaa keskiarvotason ja samalla vakioennusteen. Jäännöshajonta on ***0,77*** astetta, jäännössarjan keskiarvo ei poikkea nollasta (t = 1,48), jäännösautokorrelaatiokertoimet ovat riittävän matalia.

Tulokset

Missään pohjoismaassa ei keskilämpö ole muuttunut tilastollisesti merkitsevästi ainakaan viimeisen sadan vuoden aikana, vaan vaihdellut keskiarvonsa ympärillä, ja siksi myöskään ennusteet tulevaisuuteen eivät sisällä Ilmatieteen laitoksen uhkailemaa monen asteen nousua vuoteen 2100 mennessä. Kaikkien maiden sarjojen kuvaajissa näkyy sama muoto ja selvää noin 65 vuoden sykli. 2000-luvun alkukymmenen lämpimät vuodet eivät olleet mitään poikkeuksia, vaan vastaavaa on koettu jo 1930-luvulla (kuvassa numeroiden 58 ja 67 välillä suunnilleen).



Lähteenä olevat aikasarjat ovat kyseisten maiden ilmatieteen laitosten julkaisemia. Analyysiohjelmana on ollut Espanjan keskuspankin tekemä **Tramo/Seats**, jota Eurostat suosittelee EU:n jäsenvaltioille ja joka onkin käytössä keskuspankeissa, valtiovarain- ja muissa ministeriöissä, tilastovirastoissa, tutkimuslaitoksissa ym. kautta maailman.