Aikasarjat

Tilastotieteessä aikasarja tarkoittaa yleensä sarjaa, jossa peräkkäisten havaintojen aikaväli on aina sama.

Aikasarja on laajassa mielessä stationäärinen (wide sense stationary, WSS), jos odotusarvo on riippumaton origon valinnasta:

$$E(t) = E(t + \tau).$$

Aikasarja on tiukasti stationäärinen (strict sense stationary, SSS), jos mitattavan suureen todennäköisyysjakauma on riippumaton origon valinnasta.

Aikasarja-analyysin perusvaiheet:

- Havaintojen tasoittaminen tai muu suodattaminen.
- Hitaasti muuttuvien vaihtelujen eli trendin poistaminen. Tätä vaihtelua voidaan kuvata esimerkiksi suoralla tai korkeamman asteen polynomilla.
- Jaksollisen vaihtelun selvittäminen. Tilastotieteessä jakso on usein vuodenaikojen mukainen kausivaihtelu tai muu tunnettu jakso.
- Tulevien arvojen ennustaminen ja niiden todennäköisyysjakauman arvioiminen.

Aineiston tasoittaminen

Yksinkertaisin keino: havaittu arvo korvataan painotetulla keskiarvolla

$$g_i = \alpha f_i + (1 - \alpha) g_{i-1},$$

missä f_i on uusi havaittu arvo, g_{i-1} viimeisin jo muunnettu arvo ja α jokin vakio $0 < \alpha < 1$.

Mitä lähempänä ykköstä α on, sitä tarkemmin lopputulos seuraa alkuperäistä käyrää. Kun painoa pienennetään, saadaan tasaisempi käyrä, mutta samalla amplitudi pienenee ja vaihteluissa tapahtuu vaihesiirto.

Sopii myös reaaliaikaisiin sovelluksiin, koska tasoitukseen käytetään kullakin hetkellä vain jo havaittuja arvoja.

Liukuva keskiarvo:

$$g_i = \sum_{k=-K}^{K} w_k f_{i+k},$$

missä painojen summa $\sum w_k = 1$.

Ei aiheuta vaiheen muuttumista. Jotta saataisiin samantasoinen tasoitus kuin edellisellä menetelmällä, tarvitaan yleensä enemmän pisteitä. Tasoitettuja arvoja ei voi tallettaa alkuperäisten päälle.

Koska tarvitaan sekä menneitä että tulevia arvoja, ei sovellu ennustamiseen.

Liukuva keskiarvo voidaan laskea myös pelkästään aikaisemmista arvoista

$$g_i = \sum_{k=0}^{2K+1} w_k f_{i-k},$$

jolloin mukaan tulee taas vaihesiirto.

Regressio- ja autoregressiomalli

Systemaattista vaihtelua voidaan kuvata mallilla, jossa muuttujana on pelkästään aika. Malli voidaan yleistää myös muotoon

$$f(t) = \sum_i a_i \phi_i(t, \mathbf{x}),$$

missä kantafunktiot ϕ_i ovat täysin mielivaltaisia funktioita. Ne voivat riippua paitsi ajasta myös mistä tahansa muista suureista $\mathbf{x}$.

Kantafunktioiden argumentit voivat olla peräisin myös muista aikasarjoista. Silloin kysymyksessä on regressiomalli.

Regressiomalli soveltuu, jos ilmiöiden välillä on todella jokin muukin syy-yhteys kuin pelkästään tilastollinen. (Vrt. jäätelönmyynnin ja hukkumisonnettomuuksien korrelaatio.)

Jos jokin kantafunktio riippuu toisen aikasarjan arvosta samalla hetkellä (tai tulevaisuudessa), sen laskemiseksi on pystyttävä ennustamaan toisenkin aikasarjan kehitys. (Jotta jäätelökioskin pitäjä osaisi tilata sopivan jäätelövaraston huomista varten, hänen pitäisi ennustaa ensin, kuinka monta henkeä huomenna tulee hukkumaan.)

Jos kantafunktiot riippuvat ennustettavan muuttujan aikaisemmista arvoista, kyseessä on autoregressiomalli, esimerkiksi

$$f(t+1) = 2f(t) - f(t-1),$$

Jos aineistossa esiintyy jakso T, voitaisiin kaikkein yksinkertaisemmassa tapauksessa käyttää mallia

$$f(t+1) = f(t+1-T).$$

Jos systemaattinen vaihtelu ei ole sinimuotoista, sitä voi olla vaikea esittää muutamalla Fourier'n sarjan termillä. Sen sijaan autoregressiomallin mukainen ennuste voidaan laskea samalla tavoin, olipa vaihtelu minkä muotoista tahansa.

Periodin määrittäminen

Ongelmia:

- Etsittävän jaksollisuuden lisäksi aineistossa esiintyy muutakin vaihtelua.
- Aineistossa on aukkoja tai havaintojakso on lyhyt etsittävään jaksoon verrattuna.
- Aineistossa on useita riippumattomia jaksoja, jotka joissakin tapauksissa saattavat johtaa virheellisiin tuloksiin.
- Jakson pituus ei ole vakio, vaan muuttuu ajan mittaan.
- Vaihe muuttuu ajan mukana.
- Mittausten ajoissa esiintyvä jaksollisuus saattaa aiheuttaa ylimääräisiä näennäisiä jaksoja. Esimerkiksi muuttuvaa tähteä havaitaan vain muutama tunti keskimäärin yhden tähtivuorokauden välein.

Fourier'n muunnos

Muunnos esittää aineiston eri taajuuskomponentteja, joten maksimi jonkin taajuuden kohdalla on osoitus vastaavasta jaksollisuudesta.

Jos vaihtelu noudattaa sinikäyrää, Fourier'n muunnoksessa näkyy vain yksi maksimi. Jos muoto poikkeaa sinikäyrästä, aineiston kuvaamiseen tarvitaan myös korkeampia taajuuksia ja muunnokseen ilmestyy piikkejä myös perustaajuuden monikertojen kohdalle.

Joissakin tilanteissa saatetaan saada virheellisiä jaksoja, jos aineistossa esiintyy kaksi tai useampia lähekkäisiä jaksoja.

Autokorrelaatio

Nykyisen havainnon ja k aika-askelta myöhemmän havainnon välistä riippuvuutta kuvaa autokorrelaatio

$$R(k) = \frac{1}{N-k-1} \sum_{i=1}^{N-k} f_i f_{i+k},$$

missä N on havaintojen määrä. Kun tämä lasketaan viiveen k eri arvoilla, $k = 1, \ldots, N-2$, saadaan autokorrelaatiofunktio. Jotta tämä olisi mielekästä, aikasarjan on oltava oltava ainakin laajassa mielessä stationäärinen.

Ennen autokorrelaatiofunktion laskemista aineistosta kannattaa vähentää trendi.

Etsityn jakson monikerrat näkyvät aina autokorrelaatiofunktiossa.

Jos havainnot ovat puhdasta valkoista kohinaa, peräkkäiset arvot eivät ole lainkaan korreloituneita. Jos aineistossa esiintyy ajan suhteen jatkuvaa vaihtelua, peräkkäiset arvot ovat lähellä toisiaan, ja voimakkain korrelaatio vallitsee peräkkäisten pisteiden ($k = 1$) välillä. Pieniä k:n arvoja vastaavat suuret korrelaatiot eivät kuitenkaan johdu mistään jaksollisuudesta. Vasta seuraava paikallinen maksimi on osoitus ilmiön jaksollisuudesta.

Rakennefunktiot

Autokorrelaatio mittaa suureiden $x(t+\tau)$ ja $x(t)$ välistä riippuvuutta niiden tulon odotusarvon avulla. Tulojen sijasta vodaan tutkia erotusta $x(t+\tau) - x(t)$.

Ensimmäisen asteen rakennefunktio (structure function) on

$$D(\tau) = \frac{1}{N} \sum \left(x(t+\tau) - x(t) \right)^2 .$$

tai

$$D(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} \left(x(t+\tau) - x(t) \right)^2 dt.$$

Korkeamman asteen rakennefunktiot:

$$D_n(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} \left(\sum_{k=1}^{n} (-1)^k \binom{n}{k} x(t+k\tau) \right)^2 dt.$$

Toisen asteen rakennefunktio tunnetaan Allanin varianssina, ja sitä käytetään laajalti kuvaamaan systeemien stabiilisuutta.

Riittää, että erotukset ovat stationäärisiä. Rakennefunktio voidaan laskea, vaikka autokorrelaatiofunktio viiveen funktiona ei olisi mielekäs.

Rakennefunktio on määritelty aikasarjalle myös, jos keskiarvo ja varianssi eivät ole määriteltyjä. Rakennefunktioita voidaan siis soveltaa useampaan aikasarjaan kuin autokorrelaatiofunktioita.

Vaihedispersiomenetelmä

verrataan oletetun jakson T päässä olevia arvoja toisiinsa. Jos ne ovat lähellä toisiaan, saadaan pieni arvo poikkeamaa kuvaavalle mitalle

$$d(T) = \frac{1}{S} \sum_i \sum_j w(t_i, t_j)|y_i - y_j|^2,$$

missä

$$S = \sum_i \sum_j w(t_i, t_j)$$

Funktio w valitaan siten, että se on likimain nolla muualla paitsi aikaeron ollessa lähellä jakson pituutta tai sen monikertaa, ts. $w(t_i, t_j)$ on selvästi positiivinen vain, kun $|t_i - t_j| \approx kT$, $k = 0, 1, 2, \ldots$. Kun dispersio d piirretään jakson T funktiona, saadaan käyrä, jonka minimit vastaavat aikasarjan jaksoja.

Epätasavälisyys

Tähtitieteessä havainnot eivät yleensä ole tasavälisiä.

Mikäli kyseessä on vain muutama puuttuva piste ja muutokset eivät ole suuria, puuttuvat pisteet voidaan interpoloida. Yleisemmässä tapauksessa tämä ei useinkaan ole järkevää.

Mikäli datassa olevat aukot ovat pitkiä, lyhyitä ja pitkiä aikaskaaloja voi joutua analysoimaan erikseen tasaväliseen dataan soveltuvin menetelmin. Havaintotiheyttä voi pyrkiä tasoittamaan luokittelemalla mittaukset sopivan mittaisiin aikaväleihin ja laskemalla ko. aikavälin mittaiset keskiarvot kuten allaolevista esimerkeistä ilmenee.

Autokorrelaatiofunktio: (Edelson ja Krolik, Ap. J., 333, 646, 1988). Jokaiselle havaitulle aikaviiveelle on vain yksi mittaus, joten autokorrelaatiofunktiolle ei saada kunnon arviota. Muodostetaan sopivan mittaisia aikaluokkia, ja kunkin aikaluokan korrelaatiofunktion arvo on tämän luokan keskimääräinen korrelaatio ja luokan sisäisestä hajonnasta saadaan virhearvio.

Rakennefunktioiden laskemista epätasavälisessä datassa ei ole paljoa tutkittu. Periaatteessa rakennefunktioille voidaan käyttää samanlaista luokittelua kuin käytettiin autokorrelaatiofunktioille Edelsonin ja Krolikin menetelmässä.

Fourier'n muunnos: Deeming: Fourier analysis with unequally-spaced data, *Astrophysics and Space Science* **36** 137–158 (1975).

Tehospektri: Scargle, Ap.J., 263, 835, 1982; Lomb, Ap. Space Sci., 39, L147, 1976.

$$P(\omega) = \frac{1}{2}\left(\frac{\left[\sum x_i \cos(\omega(t_i - t_0))\right]^2}{\sum \cos^2(\omega(t_i - t_0))} + \frac{\left[\sum x_i \sin(\omega(t_i - t_0))\right]^2}{\sum \sin^2(\omega(t_i - t_0))}\right),$$

missä

$$t_0 = \frac{1}{2\omega} \arctan\left(\frac{\sum \sin 2\omega t_i}{\sum \cos 2\omega t_i}\right).$$

```
program lomb
 implicit none
 integer, parameter :: n=100
 real, parameter :: pi=3.141592654
 real, dimension(n) :: t, x, noise, f, P
 real :: w, k=4.0
 integer i
 read (*,*) w
 call random_number(t)
 call random_number(noise)
 x=sin(k*2*pi*t)+w*noise
 do i=1,n
  f(i)=0.1*i
  P(i)=period(t, x, n, f(i))
  write(*,*) f(i),P(i)
 end do
contains

real function period(t, x, n, f)
! Lombin periodogrammi taajuudella f
! t(1:n)=aika, x(1:n)=havaitut arvot
 integer, intent(in) :: n
 real, dimension(n), intent(in) :: t, x
 real, intent(in) :: f
 real, parameter :: pi=3.141592654
 real s1, s2, s3, s4, mean, var, tau, omega
 integer i

 omega=2*pi*f
 ! x:n keskiarvo ja varianssi
 mean = sum(x)/n
 var= sum((x-mean)**2)/(n-1)
 ! tau
 tau = sum(sin(2*omega*t))/sum(cos(2*omega*t))
 tau = atan(tau)/(2*omega)
 ! P(omega)
 s1 = sum((x-mean)*cos(omega*(t-tau)))
 s2 = sum(cos(omega*(t-tau))**2)
 s3 = sum((x-mean)*sin(omega*(t-tau)))
 s4 = sum(sin(omega*(t-tau))**2)
 period = (s1**2/s2 + s3**2/s4)/(2*var)
end function
end program
```