ČASOVÉ RADY A ICH DEKOMPOZÍCIA

Časový rad tvoria hodnoty, ktoré sú zhromažďované, zaznamenávané alebo pozorované postupne v čase (t. j. sú chronologicky usporiadané v čase). To ale neznamená, že čas je jediná nezávisle premenná, od ktorej hodnoty časového radu závisia. Vysvetľujúcou premennou (t.j. premennou, ktorej zmeny ovplyvňujú chovanie časového radu) môžu byť (a väčšinou aj bývajú) aj iné hodnoty. Napr. ako časový rad môžeme sledovať ceny pohonných hmôt. Zároveň však môžeme ceny pohonných hmôt brať ako vysvetľujúcu premennú pre časový rad, ktorý tvoria počty novozaregistrovaných osobných áut. Tieto hodnoty môžeme zase považovať za vysvetľujúcu premennú pre časový rad, ktorý tvorí počet dopravných nehôd a pod.

Analýza časových radov je hodnotným zdrojom informácií o priebehu skúmaného javu a často poskytuje základ pre úspešnú prognózu..

Ako príklad uvedieme niektoré veličiny, ktoré môžeme analyzovať pomocou časových radov v rôznych oblastiach.

Hydrológia:

- priemerné mesačné, resp. ročné prietoky
- výška hladiny podzemnej vody
- množstvo zrážok
- teplota vody a pod.

Financie:

- zdaniteľný a čistý zisk
- výnosy investícií
- zisk na jednu akciu
- vyplatené dividendy
- úroková miera a pod.

Geodézia:

- dlhodobé merania GPS v sieti permanentných staníc
- slapové efekty

- geocentrické súradnice permanentných staníc a pod.

Personalistika:

- počet zamestnancov
- počet dní PN
- priemerná hodinová mzda
- počet absencií
- počet ľudí, ktorí odišli z firmy a pod.

Výpočtové stredisko:

- počet užívateľov centrálnych (výkonných) počítačov
- počet spotrebovaných hodín CPU
- percento zamestnancov aktívne pracujúcich s osobnými počítačmi
- náklady na transakciu
- počet oneskorených alebo nepresných hlásení a pod.

Ako už bolo povedané, časový rad tvoria hodnoty, ktoré sú zaznamenávané postupne v čase (t. j. sú chronologicky usporiadané v čase). Interval medzi jednotlivými pozorovaniami pritom nemusí byť vo všeobecnosti rovnaký. My však budeme predpokladať, že časový interval medzi jednotlivými pozorovaniami je konštantný, t. j. budeme uvažovať diskrétny časový rad s ekvidistantným časovým krokom.

Diskrétny časový rad tvorí usporiadaná množina dát (x_1, ..., $x_{n, \ldots}$), získaných v rovnakom časovom intervale (sekunda, hodina, deň, ...). Budeme predpokladať, že x_t je reálne číslo. Hlavnou úlohou analýzy časových radov je pochopiť základný mechanizmus, podľa ktorého sú generované pozorované údaje a potom (podľa možnosti) predpovedať budúce hodnoty.

Predpokladajme, že generujúci mechanizmus je náhodný. Teda časový rad (x_1, ..., $x_{n, \ldots}$) je jednou realizáciou stochastického procesu (X_1, ..., $X_{n, \ldots}$), kde každá X_t je náhodná premenná so svojim rozdelením pravdepodobnosti a x_t je jedna hodnota tejto náhodnej premennej. Celkový počet hodnôt v časovom rade nazývame *dĺžka časového radu* a budeme ju označovať n.

Ak v čase t sledujeme hodnotu iba jedného znaku, (x_1, ..., $x_{n, \ldots}$), takýto časový rad nazývame *jednorozmerný časový rad*. Ak v čase t sledujeme súčasne hodnotu viacerých (napr. k > 1) parametrov, dostávame časový rad, ktorého prvky sú k-rozmerné vektory. Vtedy hovoríme o *mnohorozmernom časovom rade*.

Každý stochastický proces $\{X_t, t \geq 0\}$ môžeme opísať pomocou jeho charakteristík:

Stredná hodnota: $\mu_t = E(X_t)$

Rozptyl (druhý centrálny moment): $\sigma_t^2 = D(X_t) = E(X_t - \mu_t)^2$

Šikmosť (tretí centrálny moment): $\mu_t^{(3)} = E(X_t - \mu_t)^3$

Špicatosť (štvrtý centrálny moment): $\mu_t^{(4)} = E(X_t - \mu_t)^4$

Kovariančná funkcia : $\gamma(r,s) = E[(X_r - \mu_r)(X_s - \mu_s)]$

Korelačná funkcia: $\rho(r,s) = \dfrac{cov(X_r, X_s)}{\sqrt{D(X_r)}\sqrt{D(X_s)}}$.

Špeciálnym prípadom stochastických procesov sú stacionárne stochastické procesy. Stochastický proces $\{X_t, t \geq 0\}$ je *striktne stacionárny*, ak sa jeho štatistické vlastnosti v čase nemenia (t. j. rozdelenia pravdepodobnosti sú invariantné v čase).

Pretože striktná stacionarita sa v praxi veľmi ťažko overuje, vystačíme so slabou stacionaritou.

Hovoríme, že stochastický proces $\{X_t, t \geq 0\}$ je *slabo (kovariančne) stacionárny*, ak:

1. $\mu_t = E(X_t) = \mu$, $\sigma^2_t = D(X_t) = E(X_t - \mu_t)^2 = \sigma^2$ pre $t = 1, ..., n$
2. $\gamma(s, r) = cov(X_s, X_r) = E[(X_s - \mu_s)(X_r - \mu_r)]$ je len funkciou $(s - r)$, t. j. závisí len od toho, ako sú od seba X_s, X_r v čase vzdialené a nie od toho, na akom úseku časovej osi sa nachádzajú.

Rozdiel $s - r$ nazývame posunutie a budeme ho označovať k.

Časový rad ako jedna realizácia stochastického procesu $\{X_t, t \geq 0\}$ je teda **slabo (kovariančne) stacionárny**, ak základné štatistické charakteristiky (stredná hodnota a rozptyl) ostanú konštantné v priebehu celého časového radu. Stacionárny časový rad je rovnomerne vyvážený (t. j. s konštantným rozptylom) okolo konštantnej úrovne (t. j. má konštantnú strednú hodnotu), pričom závislosť (korelácia) medzi jeho dvoma ľubovoľnými pozorovaniami závisí len od ich vzájomnej časovej vzdialenosti (t. j. len na počte časových úsekov k, ktoré medzi nimi ležia), a nie na ich skutočnej polohe v časovom rade (t. j. na t).

Stacionárne časové rady môžeme charakterizovať pomocou nasledujúcich výberových charakteristík:

Výberový priemer (odhad strednej hodnoty): $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$

Výberový rozptyl (odhad disperzie σ^2): $\hat{s}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$

Výberová šikmosť (odhad šikmosti $\mu^{(3)}$): $\xi = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{x_t - \bar{x}}{\hat{s}}\right)^3$

Výberová špicatosť (odhad špicatosti $\mu^{(4)}$): $\kappa = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{x_t - \bar{x}}{\hat{s}}\right)^4$

Výberová autokovariančná funkcia: $\gamma(k) = cov(X_t, X_{t-k})$

Jednou špecifickou črtou časových radov je **autokorelácia** - každé pozorovanie je štatisticky závislé na predchádzajúcom, resp. predchádzajúcich pozorovaniach. Preto sa časový rad analyzuje omnoho ťažšie než bežný regresný model, v ktorom sú jednotlivé pozorovania navzájom nezávislé.

Korelácia medzi členmi časového radu posunutými o k časových jednotiek (medzi X_t a X_{t+k}) sa nazýva **autokorelácia k-teho stupňa**. Korelácia medzi vedľa sebou ležiacimi členmi (v čase t a t+1) sa nazýva **autokorelácia prvého stupňa**. Autokorelácia môže byť záporná, kladná alebo nulová. V prípade kladnej autokorelácie môžeme pri kladnej hodnote v čase t predpokladať, že aj hodnota v čase t+1 bude s vysokou pravdepodobnosťou kladná. V prípade zápornej autokorelácie majú hodnoty tendenciu meniť znamienka (t. j. po kladných hodnotách nasledujú s vysokou pravdepodobnosťou záporné a naopak). **Autokorelačný koeficient prvého stupňa** ρ_1 odhadujeme pomocou výberového autokorelačného koeficienta r_1, ktorý počítame podľa vzorca:

$$r_1 = \frac{\sum_{t=1}^{n-1}\left(x_t - \bar{x}\right)\left(x_{t+1} - \bar{x}\right)}{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2}$$

kde:

- $\bar{x}$ aritmetický priemer hodnôt časového radu
- x_t pozorovanie v čase t
- x_{t+1} pozorovanie v čase t+1.

Autokorelačný koeficient k-teho stupňa ρ_k (t. j. medzi X_t a X_{t+k}) odhadujeme pomocou výberového autokorelačného koeficienta r_k, ktorý počítame podľa vzorca:

$$r_k = \frac{\sum_{t=1}^{n-k}\left(x_t - \bar{x}\right)\left(x_{t+k} - \bar{x}\right)}{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2}.$$

Hodnota výberového autokorelačného koeficienta r_k pre ľubovoľné $k \geq 0$ je z intervalu <-1,1>. Keď je $|r_k| \geq 0.5$, hovoríme o silnej korelácii; pre $0 < |r_k| < 0.5$ o slabej korelácii a pre $r_k = 0$ o nekorelovaných veličinách.

Grafické zobrazenie autokorelačných koeficientov pre rôzne k od 1 až po L (kde $L \leq n/2$) sa nazýva **korelogram**.

O tom, či je autokorelačný koeficient ρ_1 štatisticky významne rôzny od 0, sa presvedčíme testovaním hypotézy o nulovosti autokorelačného koeficienta ρ_1:

$$H_0: \rho_1 = 0 \qquad H_1: \rho_1 \neq 0.$$

Podľa tejto hypotézy môžeme na špecifikovanej hladine významnosti považovať autokorelačný koeficient ρ_1 za nulový, ak hodnota r_1 padne do intervalu:

$$< -\frac{z}{\sqrt{n}}, \quad +\frac{z}{\sqrt{n}} >$$

kde

- z - je kritická hodnota normovaného normálneho rozdelenia pre danú hladinu významnosti
- n - je počet pozorovaní v danom časovom rade.

Vlastnosti $\gamma(k)$ a r_k:

1. $r_0 = 1$
2. $|\gamma(k)| \leq \gamma(0)$; $|r_k| \leq 1$
3. $\gamma(k) = \gamma(-k)$; $r_k = r_{-k} \quad \forall k \in Z$

DEKOMPOZÍCIA ČASOVÉHO RADU

Časové rady môžu byť rozložené na niekoľko zložiek. Sú to:

trend:	T
sezónna zložka:	S
cyklická zložka:	C
reziduálna (náhodná) zložka:	e.

Vykonanie rozkladu (**dekompozícia**) časového radu na tieto zložky je motivované nádejou, že v jednotlivých zložkách rozkladu sa ľahšie podarí identifikovať pravidelné správanie radu ako v pôvodnom nerozloženom rade.

Trend zachycuje dlhodobé zmeny v priemernom správaní sa časového radu (napr. dlhodobý rast alebo dlhodobý pokles). Trendová zložka obvykle vzniká ako dôsledok pôsobenia síl, ktoré systematicky pôsobia v rovnakom smere.

Sezónna zložka opisuje periodické zmeny v časovom rade, ktorých perióda sa rovná určitej štandardnej jednotke času (týždeň, mesiac, rok a pod.) alebo jej konštantnému násobku. Sezónna zložka teda zachycuje zmeny, ktoré sa pravidelne opakujú rok čo rok. Rozbor eliminovanej sezónnej zložky môže podstatne rozšíriť naše vedomosti o zákonitostiach správania určitého javu a prispieť ku konštrukcii dokonalejších predpovedí u uvažovaného časového radu. Ďalším dôležitým cieľom je tiež získanie *sezónne očisteného časového radu*, z ktorého bola sezónna zložka odstránená alebo aspoň potlačená na maximálne možnú mieru. Sezónne očistený časový rad zbavený sezónnych a náhodných fluktuácií umožňuje efektívnejšie štúdium dlhodobých tendencií, ktorým je priebeh časového radu podriadený. Napr. pre mesačné časové rady sezónna zložka vyjadruje zmeny radu, ktoré sa pravidelne opakujú v jednotlivých mesiacoch (pre časový rad, obsahujúci počet predaných kusov zimného oblečenia budú zrejme hodnoty na jar a v lete výrazne nižšie ako na jeseň a v zime pravidelne každý rok). Sezónna zložka môže zachycovať aj mnohé nepravidelnosti, napr. sviatky, rôznu dĺžku mesiacov a pod.

Cyklická zložka je periodická zložka, ktorej perióda nezodpovedá kalendárnym jednotkám. Odhad jej periódy môže byť niekedy veľmi ťažký (používajú sa pri tom napr. metódy spektrálnej analýzy). Ide vlastne o fluktuáciu okolo trendu, v ktorej sa strieda fáza rastu s fázou poklesu. Typickým príkladom cyklickej zložky je *ekonomický cyklus* (business cycle), ktorý charakterizuje rast a potom pokles ekonomickej aktivity. Jeho perióda sa pohybuje v rozmedzí od 5 do 7 rokov. Cyklická zložka môže mať svoje príčiny aj mimo ekonomickú oblasť. Napr. cyklické zmeny v klíme spôsobujú cyklické výkyvy v poľnohospodárskej produkcii, cyklické zmeny v móde zase vyvolávajú cyklické zmeny v odbyte rôznych odvetví odevného priemyslu. Eliminácia cyklickej zložky je zložitá jednak z vecných dôvodov, pretože je niekedy ťažké nájsť príčiny vedúce k jej vzniku, ale aj z výpočtových dôvodov, pretože charakter tejto veličiny sa môže v čase výrazne meniť.

Reziduálna zložka je zložka, ktorá ostane v časovom rade po odstránení trendu, sezónnej a cyklickej zložky. Je tvorená náhodnými pohybmi (fluktuáciami) v priebehu časového radu, ktoré nemajú (rozpoznateľný) systematický charakter. Preto sa už nepočíta medzi predchádzajúce tzv. systematické zložky časového radu. Pokrýva tiež chyby v meraní údajov, ktoré tvoria časový rad. Obvykle sa predpokladá, že reziduálna zložka je **biely šum** - nekorelované náhodné veličiny s nulovou strednou hodnotou.

Časový rad si na základe predchádzajúceho výkladu môžeme predstaviť ako trend, na ktorý sú "nabalené" periodické zložky (sezónna a cyklická) a biely šum. Označme x hodnoty časového radu. Vlastný rozklad môže byť dvojakého druhu:

Aditívna dekompozícia:

$$x = T + S + C + e.$$

Pri aditívnej dekompozícii sú jednotlivé zložky uvažované vo svojich skutočných absolútnych hodnotách a sú merané v jednotkách časového radu x.

Multiplikatívna dekompozícia:

$$x = T \,.\, S \,.\, C \,.\, e.$$

Pri tejto dekompozícii je väčšinou len trend uvažovaný vo svojej absolútnej hodnote a teda je meraný v jednotkách časového radu Y. Ostatné zložky sú uvažované v relatívnych hodnotách voči trendu a sú teda bezrozmerné.

Časové rady spravidla obsahujú reziduálnu zložku. Okrem nej môžu (ale nemusia) obsahovať aj jednu, dve alebo všetky tri ostatné zložky.

Časový rad môžeme na jednotlivé zložky rozložiť viacerými spôsobmi. Jednou z najčastejšie používaných (aj najznámejších) metód je rozklad časového radu pomocou mnohorozmernej regresie. Ďalej môžeme používať kĺzavé priemery, exponenciálne vyrovnávanie a pod.

Najčastejšie používané typy trendu sú:

Konštantný trend: $T_t = \beta_0$, $t = 1, \ldots, n.$

Konštantný trend používame, ak $x_{t+1} - x_t \approx 0$.

Lineárny trend $T_t = \beta_0 + \beta_1 t$, $t = 1, \ldots, n.$

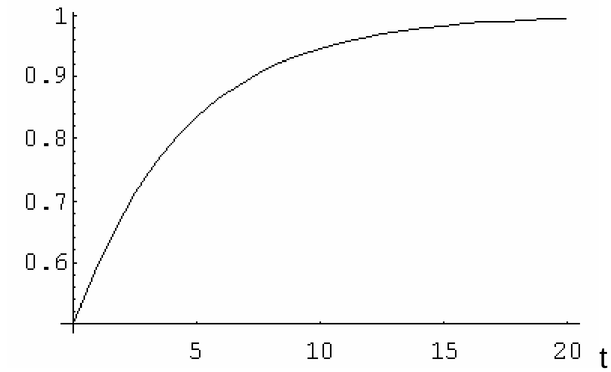
kde koeficienty β_0, β_1 odhadneme použitím lineárnej regresie. Lineárny trend používame, ak sú prvé diferencie, t. j. $x_{t+1} - x_t$, približne konštantné.

Kvadratický trend $T_t = \beta_0 + \beta_1 t + \beta_2 t^2$, $t = 1, \ldots, n.$

Koeficienty β_0, β_1, β_2 opäť odhadneme pomocou lineárnej regresie. Kvadratický trend používame, ak sú druhé diferencie, t. j. $x_{t+2} - 2\,x_{t+1} + x_t$ približne konštantné.

Exponenciálny trend $T_t = \alpha.\beta^t$, $t = 1, \ldots, n$, $\beta > 0.$

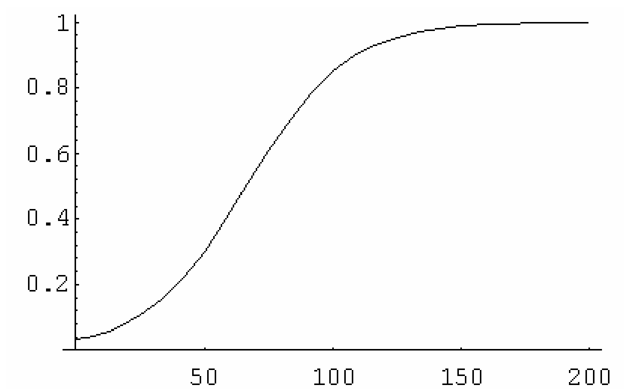
Tento typ trendu sa používa na opísanie modelu rastu nových produktov, spoločností alebo priemyslu. Životný cyklus typický pre nový produkt, spoločnosť alebo priemysel je nasledovný:



Začiatok Rast Zrelosť a nasýtenie

Ak je $\alpha > 0$, potom pre $\beta > 1$ dochádza k rastu, zatiaľ čo pre $0 < \beta < 1$ nastáva pokles. Parametre α a β odhadneme pomocou nelineárnej regresie. Exponenciálny trend používame, ak sú podiely po sebe idúcich členov časového radu x_{t+1} / x_t približne konštantné.

Logistický trend $T_t = \dfrac{\gamma}{1+\alpha.\beta^t}$, $t = 1, \ldots, n, \beta > 0, \gamma > 0,$



Kritérium pre jeho použitie je, aby podiely $(1/x_{t+2} - 1/x_{t+1}) / (1/x_{t+1} - 1/x_t)$ boli približne konštantné.

Pre elimináciu <u>sezónnej zložky</u> sa v aditívnom modeli často používajú **regresné metódy** založené na teórii lineárneho regresného modelu. Používa sa pri tom metóda kvalitatívnych premenných. Ak na každý rok pripadne L pozorovaní (sezón) príslušného časového radu (napr. pri mesačných časových radoch L = 12, pri kvartálnych L = 4), potom pri tejto metóde vyjadrujeme sezónnu zložku v tvare:

$$S_t = \alpha_2 Y_{2t} + \alpha_3 Y_{3t} + \ldots + \alpha_L Y_{L\,t}$$

kde Y_{2t}, Y_{3t}, …, $Y_{L\,t}$ sú **kvalitatívne premenné**, ktoré sú definované nasledujúcim spôsobom:

$$Y_{i\,t} = \begin{cases} 1 & \text{ak čas t odpovedá (i-1) -vému obdobiu v roku} \\ 0 & \text{v inom období v roku} \end{cases}$$

pre i = 2, 3, …, L.

Z regresných odhadov parametrov α_2, α_3, …, α_L sa ľahko určia sezónne faktory. Veľmi často sa tiež zároveň odhaduje aj priebeh trendu (ak je lineárny z hľadiska odhadovaných parametrov).

Iná regresná metóda pre elimináciu sezónnej zložky je založená na tom, že sa túto zložku snažíme odhadnúť pomocou vhodne zvolenej matematickej funkcie. Najčastejšie sa používajú goniometrické funkcie s dĺžkou periódy L: sin(2 π t/L), cos(2 π t/L). Keď predpokladáme lineárny trend, môže mať model napr. tvar:

$$x_t = \beta_0 + \beta_1 t + \beta_2 \sin(2\pi t/L) + \beta_3 \cos(2\pi t/L) + e_t .$$

Bežne používané dekompozičné metódy kladú dôraz predovšetkým na prácu so systematickými zložkami časového radu (t. j. trend, sezónna a cyklická zložka), My však budeme uvažovať aj reziduálnu zložku, ktorá môže byť tvorená korelovanými (navzájom závislými) pozorovaniami.

Autokorelačné koeficienty, počítané pre rôzne hodnoty k, pomôžu nájsť odpoveď na nasledujúce otázky:

- sú dáta náhodné
- majú tieto dáta nekonštantný trend (teda sú nestacionárne)
- sú dáta sezónne.

Ak je časový rad náhodný, korelácia medzi X_t a X_{t+k}, $\forall k \geq 1$ sa rovná 0, to znamená, že odpovedajúce hodnoty časového radu nie sú navzájom korelované.

Ak má časový rad nekonštantný trend, sú X_t a X_{t+1} vysoko korelované; výberové autokorelačné koeficienty sú štatisticky významne rôzne od nuly pre niekoľko prvých k a potom pomaly klesajú k nule s rastúcim k. Autokorelačný koeficient prvého stupňa je v tomto prípade najväčší a pomerne veľký je aj autokorelačný koeficient druhého stupňa.

Ak má časový rad sezónny charakter, ako štatisticky významné sa ukazujú autokorelačné koeficienty pre násobky k = L (kde L je perióda sezónnej zložky).

Pokiaľ by sme zistili, že ani jeden autokorelačný koeficient nemôžeme považovať za štatisticky významne rôzny od nuly, znamená to, že máme reziduálny časový rad zložený z nekorelovaných náhodných veličín.