DEBRECENI EGYETEM

Matematikai Intézet

Informatikai Kar

IDŐSORELEMZÉSI MÓDSZERTANOK ÖSSZEHASONLÍTÁSA STATISZTIKAI TANULÓ ALGORITMUSOK SEGÍTSÉGÉVEL

Témavezető:

Dr. Ispány Márton

egyetemi docens

Készítette:

Fábián László

programtervező matematikus

DEBRECEN

2008

TARTALOMJEGYZÉK

oldalszám

Bevezetés

A dolgozat keretében bepillantást kívánok nyújtani az idősorelemzési módszertanok elméleti és gyakorlati vizsgálatába. A gyakorlati vonatkozáson belül nevezetesen annak a folyamatnak egy részére, melynek keretében a 2001/2002-es év során a Központi Statisztikai Hivatalban kiválasztottuk és rendszeresítettük a TRAMO/SEATS szezonális kiigazítási módszertant az X12-ARIMA alternatívával szemben, mint egységes szezonális kiigazítási módszertant a hivatalon belül. Azért választottam ezt a témát, mert az említett munkát elvégző munkacsoportnak magam is tevékeny tagja voltam.

A KSH az 1990-es évek eleje-közepe óta a Kanadai Statisztikai Hivatalban kifejlesztett X11-ARIMA módszert használta a szezonális kiigazítások elvégzésére. Időközben a nemzetközi gyakorlatban újabb, illetve más elméleti alapon nyugvó szezonális kiigazító módszerek is megjelentek. Az 1990-es években például nemzetközi szinten egyre inkább elterjedt az X12-ARIMA és a TRAMO/SEATS. Az Eurostat felmérése szerint az EU tagállamok gyakorlatában e két módszer használata vált a legelterjedtebbé. Ehhez az elméleti és gyakorlati fejlődéshez kapcsolódóan 2000-2001 táján a KSH-ban is megnőtt az igény egy új, egységesen használható módszer és gyakorlat bevezetésére, és jóval több szezonálisan kiigazított adat publikálására. A Hivatalra több oldalról nehezedett nyomás: egyrészt a negyedéves GDP kiigazítását is el kellett kezdeni, másrészt a hazai felhasználók, különösen a Magyar Nemzeti Bank is kritikát fogalmazott meg a KSH szezonális igazításával kapcsolatosan, illetve az Európai Unió harmonizációs igénye, amely a tag és tagjelölt országok szezonális kiigazításának egységesítését támogatta, is indokolta egy új módszer bevezetését.

Összességében tehát az elméleti modellek fejlődése, az EU harmonizációs feladat, továbbá a megnövekedett felhasználói igények szükségessé tették a hivatalban alkalmazott módszertan felülvizsgálatát. Az egységes szezonális kiigazítási igény végül is az X12-ARIMA vs. TRAMO/SEATS közüli választás szükségességét eredményezte.

A választást megalapozó összehasonlítási munkánk során figyelembe vett szempontok, illetve az alkalmazott módszerek körét most egy újabb lehetőséggel, adatbányászati vizsgálat bevonásával fogom kibővíteni. Az elemzés során elvégzendő kiigazítások szempontjából előnyt jelent, hogy a szükséges alapadatok biztosítására nagyobb elemszámú idősorok váltak időközben elérhetővé.

Az idősorok viselkedését nagymértékben befolyásolják olyan tényezők, amelyek különböző évek azonos időszakaiban (pl. hónap) azonos irányban és közel azonos mértékben hatnak az idősor alakulására. Ilyenek lehetnek az évszakok, vagy a piaci, például a gazdasági konjunkturális ingadozások következményei is, esetleg adminisztratív hatások vagy különféle társadalmi tényezők. Ezeket együttesen szezonális hatásnak nevezzük. Az elemzők gyakran a folyamatok olyan jellemzőire kíváncsiak, amelyeket a nagymértékű szezonális hatás elfed, ezért szükség van ennek eltávolítására. A szezonalitás kiszűrését szezonális kiigazításnak nevezzük, melynek részletesebb tárgyalását is célul tűztem ki a jelen a dolgozat elkészítése során.

I. A szezonális kiigazítás elméleti alapjai

A szezonális kiigazítás tárgyalása előtt ki kell térni az idősorelemzés elméleti hátterének szükséges részleteire:

- Beszélni kell az idősorok elméletéről, az ARMA folyamatokról, a felhasznált statisztikai becslésekről, próbákról, valamint a szezonális kiigazítás módszereiről és az erre használható szoftverekről;

- Az elméleti ismeretek alapján meg kell határozni az elvégzendő feladatokat, mint például a szezonálisan kiigazítandó idősorok szakmai vizsgálata, trendváltozások, kiugró értékek, esetleges törések, befolyásoló tényezők azonosítása, hatásuk kimutatása, megvitatása a kiigazítást megelőző döntések előkészítése érdekében, valamint a szezonális kiigazítási eljárások (X12-ARIMA, TRAMO/SEATS) program szerinti futtatása, eredményeinek összehasonlítása adott kritériumok szerint. A kritériumok közül kiemelném a szezonális kiigazítás minőségének indexét – *sa quality index* –, amely mint mérőszám, az aktualizált alapadatokon elvégzett összehasonlításom egyik, jelen esetben kiemelt alapjául fog szolgálni a kísérleti számításokban.

Gazdasági, társadalmi folyamatok vizsgálatához gyakran idősorokat használnak, amelyek a folyamatok időbeli alakulását írják le.
Statisztikai szempontból az idősor az egyes időpontokhoz, vagy időszakokhoz rendelt valószínűségi változóértékek sokasága. Az általuk jellemzett folyamatokban számos tényező játszik szerepet, melyekről többnyire nem áll rendelkezésre kielégítő, részletes információ. Sokaságuk és komplexitásuk miatt nem lehet teljes körűen, megfelelő részletességgel figyelembe venni őket, vagyis nem lehet analitikusan kalkulálni a teljes rendszer, illetve folyamat működésére gyakorolt részhatásukat [6]. Érzékelhető viszont e tényezők összhatása közvetett módon, az időtényezőn keresztül. Az időtényező a gyűjtője a jelenséget befolyásoló tényezők sokaságának [2], ezáltal speciális sztochasztikus kapcsolatnak tekinthetők, ahol a magyarázó változó szerepét formailag az időtényező tölti be.

1. Az idősorelemzés típusai

Az idősorok vizsgálatának két fő területe van, nevezetesen a determinisztikus- és a sztochasztikus idősorelemzés, melyek közül az előbbit csak röviden, megemlítés szintjén fogom érinteni egyrészt terjedelmi okokból, másrészt mert ennek már hosszú időre visszanyúló, közismert elméleti háttere van. A kitűzött feladat szempontjából inkább a sztochasztikus idősorelemzés elméletének tárgyalására lesz szükség, annak is lehetőleg a szezonális kiigazítás elméletében valamilyen módon felhasznált témaköreire frekventálva.
A determinisztikus idősorelemzés során alkalmazott modelltípust azért nevezik determinisztikusnak, mert az idősornak csak az ún. determinisztikus komponenseit (pl. az alapirányzatból és a periodikus ingadozásból származó komponensét) veszi fegyelembe. Ezeket az összetevőket igyekszik teljesen determinisztikusan kezelni, a véletlen hatását pedig egy külön összetevőbe tömöríteni és lehetőleg minél jobban eliminálni a becslések, illetve az előrejelzések során.

A véletlen ingadozásokért felelős komponensben ugyanis valószínűségi változók sokaságának hatása játszik szerepet. Ez az összetevő mintegy azok eredőjét képezi, és hatásának eredményeképpen sztochasztikusan ingadoznak a vizsgált idősor adatai a determinisztikus módon becsült komponensekből előállított görbe körül. Ez a komponens fogja képezni a becslések maradékát — *reziduumokat* —, és egyben a becslések hibáját is a determinisztikus modellben.

2. Idősorok komponensei és kapcsolódási lehetőségei

Az additív modell esetében az idősor egyes értékei az alábbi alapvető komponensek összegződése révén állnak elő, az

$$Y(t) = X_T(t) + X_S(t) + X_C(t) + X_\xi(t) \qquad (2.1)$$

formulának megfelelően, ahol

$Y(t)$: a vizsgált idősor tényadatai,
$X_T(t)$: trend (a tartósan érvényesülő tendencia, hosszú távú alapirányzat)
$X_S(t)$: szezonális komponens (a trendtől való eltérés, az éven belüli, periodikus ingadozás mértéke)
$X_C(t)$: ciklikus komponens (a hosszabb távú ingadozás)
$X_\xi(t)$: irreguláris komponens (stacionárius, $m_\xi = 0$ várható értékű, σ_ξ szórású, Gauss eloszlású véletlen hatás).

Az idősorok összetevői két módon kapcsolódhatnak egymáshoz: összegszerűen, illetve szorzatszerűen. Az előbbi additív, az utóbbi multiplikatív modellhez vezet. Ezeken az alapmodelleken kívül megkülönböztetünk még logadditív és pszeudoadditív modelleket is. A logadditív modellben a tényezők logaritmusai kapcsolódnak összegszerűen egymáshoz. A pszeudoadditív modellt az angol statisztikai hivatalban fejlesztették ki olyan idősorokra, ahol az összekapcsolódás lényégében multiplikatív, de egyes szezon esetében az idősor rendkívül kis értékeket vesz fel[1]. Mivel az általam vizsgált iparstatisztikai idősorok esetében többnyire a multiplikatív összekapcsolódás tételezhető fel, illetve mutatható ki, ezért a program ismertetésénél használt példák is főként erre fognak vonatkozni.
Az említett idősorokra jellemző, a programban használt multiplikatív modell esetén az idősor felbontása:

$$Y_t = T_t \times S_t \times I_t \times D_t \times E_t$$

ahol T_t a trend, S_t a szezonális komponens, I_t a véletlen hatás, D_t a munkanap-hatás, E_t a húsvét-hatás.

3. ARMA folyamatok

Az idősorok elméletében és alkalmazásában az autoregresszív és mozgóátlag- (ARMA) folyamatok jelentősége az utóbbi évtizedekben rendkívül megnőtt. Ez annak köszönhető, hogy az ARMA sztochasztikus folyamatok matematikai szempontból jól

[1] Sugár András (1999): , Szezonális kisimító eljárások összehasonlítása. Gazdasági Minisztérium,' Gazdaságelemző intézet

kezelhetők, és a folyamatok egy elég általános osztályát képviselik. Emiatt a gyakorlatban előforduló, stacionárius viselkedést mutató véletlen folyamatok nagy része jól közelíthető az ARMA folyamatokkal.
Az

$$y_t \sim f_t(y_t). \tag{2.2}$$

sztochasztikus folyamat egymástól független valószínűségi változók sorozatának egy reprezentációja (egy ún. valószínűségi vektorváltozó). Az egyes megfigyelések külön-külön is egy-egy valószínűségi változó adott realizációi.
Statisztikai szempontból nézve egy adott

$$[y_1, y_2, \ldots, y_t].$$

idősor az egyes időpontokhoz, vagy időszakokhoz rendelt valószínűségi változóértékek sokasága. Így tehát az idősorokat a sztochasztikus folyamatok egy diszkrét, megfigyelt realizációjának tekintjük. Sztochasztikus folyamat minden olyan idősor, amelynek pillanatnyi alakulását saját korábbi állapotából és a véletlen hatásokból lehet magyarázni. E felfogás szerint a véletlen változó beépül a folyamatba, annak aktív alkotóeleme lesz, és a jelenség fő mozgatójává válik. A sztochasztikus idősorelemzés során tehát a véletlen hatását reprezentáló valószínűségi változót az idősor modelljének szerves részeként kezeljük.

Stacionaritás

Fontos dolog továbbá a vizsgált folyamat stacionaritása. A tárgyalásra kerülő autoregresszív folyamatatok például csak akkor becsülhetők, illetve előrejelezhetők, ha a vizsgált folyamat, illetve az arra illesztett modell kielégíti a tágabb értelemben vett stacionaritás kritériumait.

Egy stacionárius folyamat ugyanis lehet *szűkebb értelemben stacionárius* (szigorúan, vagy elsőrendűen stacionárius), illetve *tágabb értelemben stacionárius* (gyengén, vagy másodrendűen stacionárius).
A szűkebb értelem vett stacionáritás esetén az $Y(t)$ folyamat eloszlás-, ill. sűrűségfüggvényei bármely időeltolással szemben invariánsak lesznek.
Matematikai megfogalmazásban: Az $\{y_t,\ t \in T\}$ sztochasztikus folyamat szűkebb értelemben stacionárius, ha az $\{y_{t+s},\ t \in T_0\}$ vektor eloszlása a T halmaz véges T_0 részhalmazaira független s-től.
Tágabb értelemben stacionáriusnak (gyengén stacionáriusnak) nevezzük azt az $y(t)$ folyamatot, amelyre teljesül, hogy

$$M_y(t) = E[y(t)] = m_y\ , \tag{2.3}$$

azaz a várható érték minden időpontban azonos, továbbá az autokorrelációs függvényére[2] fennáll, hogy

$$R_{yy}(t_1\ ,\ t_2) = R_{yy}(t_1 - t_2) = R_{yy}(\tau)\ , \tag{2.4}$$

vagyis a korrelációfüggvény nem függ a konkrét időpontoktól, csak a két időpont különbségétől ($|t_1-t_2| = \tau$) [26].
Egy sztochasztikus folyamat stacionaritása úgy is megfogalmazható, hogy a statisztikai jellemzői időben állandók. Néhány gazdasági idősor eleve kielégíti az említett feltételeket. Amelyek viszont nem, azok esetében is biztosítható a stacionaritás, egyszerű transzformációk segítségével, amint azt később látni fogjuk.

[2] Az autokorreláció (valamint az autokovarianca) definiciójáról a későbbiekben lesz szó.

3.1. AR és MA modellek

Az *autoregresszív* (AR) jelző arra utal, hogy a folyamat részben saját múltjára vonatkozó lineáris regresszióként írható fel. A *mozgó átlag* (MA) jelző pedig azt fejezi ki, hogy a lineáris regresszió „hibatagja" az ε_t fehérzaj mozgó átlaga, azaz a jelen és a véges múlt lineáris kombinációja.
(Mivel azonban a mozgóátlagolású tagok együtthatói – paraméterei – nem szükségképpen pozitívak és összegük sem okvetlenül egységnyi, így az elnevezés tulajdonképpen pontatlan: nem valódi mozgóátlagról van szó).

Általános esetben, egy q-adrendű MA(q) folyamat alakja

$$\theta(\boldsymbol{B})\ \varepsilon_t = y_t\ ,$$

ahol $\theta(\boldsymbol{B})$ a visszaléptetés operátorának polinomja:

$$\theta(\boldsymbol{B}) = \theta_0 - \theta_1\ \mathbf{B} - \ldots - \theta_q\ \mathbf{B}^q\ .$$

Tehát egy általános q-adrendű mozgóátlag MA(q) folyamatot a következő összefüggés írja le:

$$\theta(\boldsymbol{B})\ \varepsilon_t = \theta_0\varepsilon_t - \theta_1\ \varepsilon_{t-1} - \ldots - \theta_q\ \varepsilon_{t-q} = y_t\ , \tag{3.1}$$

ahol ε_t fehérzaj (korrelálatlan sztochasztikus folyamat zérus várható értékkel és konstans varianciával).

Azáltal, hogy az y_t-t alkotó tagokban csak a fehérzaj-folyamat különböző értékei szerepelnek, és $E(y_t) = 0$, így

$$\boldsymbol{E}y_t\varepsilon_n = \begin{cases} 0, & ha \quad n < t-q \quad vagy \quad n > t \\ \theta_{t-n,} & ha \quad t-q \le n \le t \end{cases}\ . \tag{3.2}$$

Mivel (3.1) és y_{t+k} szorzatának várható értéke

$$\mathbf{E}y_{t+k}\,\varepsilon_t = \theta_0\,\mathbf{E}y_{t+k}\,\varepsilon_t - \theta_1\,\mathbf{E}y_{t+k}\,\varepsilon_{t-1} - \ldots - \theta_q\,\mathbf{E}y_{t+k}\,\varepsilon_{t-q} = y_t\ ,$$

így (3.2) felhasználásával

$$c_k = \boldsymbol{E}y_t\,y_{t+k} = \begin{cases} \theta_0\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q, & ha \quad 0 \le k \le q \\ 0, & ha \quad k > q \end{cases}\ . \tag{3.3}$$

A (3.3) összefüggés alapján a mozgóátlag folyamat mindig stacionárius, mivel az egyenlet jobb oldala véges, így a baloldal, az autokovarianciák is végesek. Egyben az autokorrelációs függvénye is véges sok nullától különböző értéket vesz fel a k késleltetési értékek függvényében.

Adott $c_0, c_1, \ldots, c_q$ *valós számokhoz akkor és csakis akkor találhatók olyan* $\theta_0, \theta_1, \ldots,$ θ_q *valós számok, amelyekre*
teljesül , ha $c_0 > 0$ *és a komplex síkon definiált*

$$\Gamma(B) = c_0 + \sum_{k=1}^{q} c_k\,(B^k + B^{-k})$$

függvénynek az egységkörön csakis páros multiplicitású gyöke van.

Wold (1954): (a) *Ha a* $c_0, c_1, \ldots, c_q$ *valós számokhoz találhatók olyan valós* $\theta_0, \theta_1, \ldots, \theta_q$ *számok, amelyekre*

$$c_k = \theta_0\,\theta_k + \theta_1\,\theta_{k+1} + \ldots + \theta_{q-k}\,\theta_q, \qquad \text{ha} \qquad 0 \le k \le q$$

teljesül, akkor van olyan y_t MA(q) *folyamat, amelyre* $\mathbf{E}y_t = 0$, és

$$c_k = \mathbf{E}y_t\,y_{t+k}, \quad \text{ha} \qquad 0 \le k \le q$$

(b) *Ha az* y_t *stacionárius Gauss-folyamatra* $\mathbf{E}y_t=0$, *és a* $c_k=\mathbf{E}y_t y_{t+k}$ *autokovarianciákra*

$$c_k = 0, \qquad \textit{ha} \qquad k>q$$

teljesül valamilyen $q \ge 0$ *mellett, akkor* y_t MA(q) *folyamat.*

A mozgóátlag folyamat θ_i paramétereinek ki kell elégíteniük az ún. invertibilitási feltételt, mely az elsőrendű MA folyamat ($q = 1$) esetében $|\theta_1|<1$. Ez azt jelenti, hogy a belőle transzformált végtelenrendű autoregresszív folyamatnak, mint végtelen sornak konvergensnek kell lennie. Ezt a transzformációt el lehet végezni szukcesszív behelyettesítésekkel, ahogy (3.12)-ből a (3.13) előállítása is történik a későbbi modellverifikáció témájával foglalkozó szakaszban.

Felhasználva az $\boldsymbol{B}$ késleltető operátort az előbbi folyamat a következő formát veszi fel:

$$y_t = (1 + \theta_1 \boldsymbol{B})\varepsilon_t,$$

Egy p-edrendű AR(p) folyamat formája általános esetben

$$\phi(\boldsymbol{B})\, y_t = \sigma\varepsilon_t,$$

ahol $\phi(\mathbf{B})$ a visszaléptetés operátorának polinomja (ill. $\boldsymbol{B}\, y_t = y_{t-1}$):

$$\phi(\boldsymbol{B}) = 1 - \phi_1 \boldsymbol{B} - \ldots - \phi_p \boldsymbol{B}^p \tag{3.4}$$

Így tehát

$$\phi(\boldsymbol{B})\, y_t = y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} = \sigma\varepsilon_t, \tag{3.5}$$

ill. átrendezve

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \sigma\varepsilon_t \qquad t \in \boldsymbol{Z},. \tag{3.6}$$

ahol ε_t most egységnyi szórású fehérzaj, és $\sigma > 0$. A σ-t sokszor 1-re választják, és ekkor ε_t már nem egységnyi, hanem σ szórású lesz. Tehát σ-t mintegy beleolvasztják a zajfolyamatba — pl. Box-Jenkins modellek.

Amennyiben az $\boldsymbol{B}$ visszaléptetés-operátorának van olyan (3.4) szerinti polinomja, amely kielégíti a (3.5) egyenletet, akkor az y_t folyamatot *p-edrendű autoregresszív* folyamatnak nevezzük.

Mivel a (3.4) polinomnak természetesen komplex gyökei is lehetnek, így az a feltétel, hogy a $\phi(\boldsymbol{B})$ operátorhoz mikor tartozik olyan stacionárius y_t folyamat, amely megoldása a (3.5) differenciaegyenletnek, a következőképpen fogalmazható meg:

Ha ε_t fehérzaj-folyamat, $\sigma>0$ és az $\phi(B) = 1- \phi_1 s- \ldots - \phi_p s^p$ polinom gyökei az egységkörön kívül helyezkednek el, akkor van olyan y_t stacionárius AR(p) folyamat , amelyre (3.6) *teljesül.*

Ha $\phi(\boldsymbol{B}) = 1+ \phi_1 B +\ldots+ \phi_p B^p$ olyan p-edfokú polinom, amelynek gyökei a komplex egységkörön kívül helyezkednek el, akkor a $\phi(\boldsymbol{B})$ operátor invertálható.

Az autoregresszív folyamatok stacionaritási feltételei két speciális esetben, az első- és másodrendű folyamatok esetén (a gyakorlatban ugyanis ezek fordulnak elő legtöbbször):

1. Az

$$\varepsilon_t = (1-\phi_1 \boldsymbol{B}) y_t ,$$

vagy másképpen

$$y_t = \phi_1 y_{t-1} + \varepsilon_t \,^3,$$

elsőrendű AR(1) folyamat esetén például

$$-1 < \phi_1 < 1$$

a stacionaritás feltétele (ahol $\boldsymbol{B}y_t=y_{t-1}$).

2. Az alábbi másodrendű autoregresszív folyamat esetén:
Az

$$y_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \sigma \varepsilon_t$$

AR(2) folyamat $\phi(B) = 1- \phi_1 B - \phi_2 B^2$ operátor polinomjának stacionaritási — ld. (3.12), és a (3.13) összefüggések — feltétele az, hogy az alábbi relációk egyaránt teljesüljenek a paraméterekre:

$$a_2 - a_1 < 1, \qquad \text{ill.} \qquad a_2 + a_1 < 1, \qquad \text{és} \qquad -1 < a_2 < 1.$$

Tetszőleges *p* rendű AR folyamatra is érvényes, hogy amennyiben van konvergens végtelen mozgóátlag előállítása, vagyis invertálható, akkor stacionárius, mivel a (3.3)-nak megfelelően minden MA folyamat stacionárius.[4][26]

---

[3] Ez a folyamat egyébként egy ún. *Markov-lánc*, ahol bármely „jövőbeli" eseménynek a feltételes valószínűsége bármely „múltbeli" esemény és az $y_{t-1}=i$ „jelen" esemény mellett független a múltbeli eseményektől, csak a rendszer jelen állapotától függ.

[4] Speciális esetben a (3.12) szerinti elsőrendű AR folyamatra, ill. a (3.15) szerinti előállítására is kimondható ugyanez.

A tetszőleges rendű AR folyamat ún. *végtelen mozgóátlag reprezentációjának* előállítása a (3.5) összefüggés, illetve a (3.4) polinom felhasználásával a következő módon történik:

$$\phi(\boldsymbol{B})\, y_t = \varepsilon_t,$$

$$y_t = \frac{1}{\phi(\boldsymbol{B})}\varepsilon_t = \frac{1}{1-\phi_1\boldsymbol{B}-\ldots-\phi_p\boldsymbol{B}^p}\varepsilon_t = \Gamma(\boldsymbol{B})\varepsilon_t \ ,$$

ahol

$$\Gamma(\boldsymbol{B}) = \left(1-\phi_1\boldsymbol{B}-\ldots-\phi_p\boldsymbol{B}^p\right)^{-1} = 1+\gamma_1\boldsymbol{B}+\gamma_2\boldsymbol{B}^2+\ldots$$

az ún. *generátorfüggvény*.

A $\Gamma(\boldsymbol{B})$ polinom γ (Wold) együtthatóinak kiszámítási módja, ill. annak rekurzív algoritmusa például a soron következő ún. *kombinált* (ARMA) modellek esetében az alábbi módszerrel állítható elő, szukcesszív behelyettesítéseket felhasználva:

$$\left.\begin{aligned}
\gamma_1 &= \phi_1 - \theta_1 \\
\gamma_2 &= \phi_1\gamma_1 + \phi_2 - \theta_2 \\
&\vdots \\
\gamma_j &= \phi_1\gamma_{j-1} + \ldots + \phi_p\gamma_{j-p} - \theta_j
\end{aligned}\right\}$$

Az ϕ_t paraméterek a $\phi(\boldsymbol{B})$, a θ_t paraméterek pedig $\theta(\boldsymbol{B})$ polinom együtthatóit jelentik.

Megj:

Amennyiben a kiszámolt

$$C(k) = \sigma^2 \sum_{j=0}^{\infty} \gamma_j\, \gamma_{j+k} \ , \qquad \gamma_0 = 1$$

Wold együtthatók autokovariancia függvényének értékei végesek — minek elégséges feltétele az együtthatók abszolút értékeiből képzett

$$\sum_{j=0}^{\infty} \left|\gamma_j\right|$$

végtelen sor konvergenciája —, akkor a folyamat stacionárius.

3.2. ARMA modellek

A kevert, vagy kombinált ún. *ARMA-modellek* mind autoregresszív, mind mozgóátlagolású tagokat tartalmaznak. Speciális esetük az elsőrendű ARMA(1, 1) folyamat, melynek képlete:

$$(1-\phi_1 \boldsymbol{B})\, y_t = (1+\theta_1 \boldsymbol{B})\varepsilon_t ,$$

illetve

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Az általános ARMA(p,q) modell formája:

$$\phi(\boldsymbol{B}) y_t = \theta(\boldsymbol{B})\varepsilon_t .$$

Abban az esetben, ha a véletlen folyamat az időeltolással szemben nem invariáns, legalább másodrendben nem stacionárius, először is megfelelő differenciaképzéssel megpróbáljuk stacionáriussá tenni a folyamatot.

A regresszióanalízis során ugyanis alapvető feltétel a megfigyelések függetlensége. A gyakorlatban azonban ritkán függetlenek egymástól a különböző időpontokban megfigyelt értékek

Mivel közelítőleg lineáris trend esetén például az egymást követő tényadatok között konstans értékű a különbség, vagyis a növekmény vagy csökkenés, így az egységintervallum szerinti differenciaképzéssel a trendhatást tekintve konstans értéken tartható a folyamat.
A periódusidő szerinti differenciálással pedig a szezonális időeltolás mellett tapasztalható tendenciális hatások szűrhetők.

Például a (3.14)-ben bemutatott bolyongási folyamat esetén — amelyről említettem, hogy nem stacionárius — látható, hogy az első differenciák sorozata az

$$\varepsilon_t = y_t - y_{t-1}$$

szerinti fehérzajt szolgáltatja, amely már stacionárius.

3.3. ARIMA modellek

A differenciálások után tehát stacionáriussá tehető a folyamat és a következő ARIMA(p, d, q) modellel írható fel reguláris esetben:

$$\phi(\boldsymbol{B})\, \nabla^d y_t = \theta(\boldsymbol{B})\, \varepsilon_t$$

ahol ∇^d a differenciaképzés operátor-polinomja.
Egységnyi késleltetés mellett végzett differenciálás esetén például:

$$\nabla^d = (1-\boldsymbol{B})^d .$$

Szezonális késleltetés esetén

$$\nabla_s^D = (1-\boldsymbol{B}^s)^D,$$

ahol az *s* a megfigyelések száma / év (s=4 – negyedéves idősorok, ill. s=12 – havi bontású idősorok esetében).

3.4. SARIMA modellek

Amennyiben a vizsgált idősor szignifikáns szezonalitást is tartalmaz, akkor (a differenciálások után már stacionárius) ún. *szezonális autoregresszív integrált mozgóátlag folyamatnak* tekintjük, és egy multiplikatív, ún. *SARIMA* (*p*, *d*, *q*) (*P*, *D*, *Q*) modelltípussal írjuk le.

Ennek általános képlete: $\phi_p(\boldsymbol{B})\Phi_P(\boldsymbol{B}^S)\nabla^d\nabla_s^D y_t = \theta_q(\boldsymbol{B})\Theta_Q(\boldsymbol{B}^S)\varepsilon_t$, (3.7)

ahol

$$\phi_p(B)\Phi_P(B^S) = (1+\phi_1 B+\phi_2 B^2+\ldots+\phi_p B^p)(1+\Phi_1 B^S+\phi_2 B^{2S}+\ldots+\phi_P B^{PS})$$

az ún. autoregresszív operátor-polinom, ill.

$$\theta_q(B)\Theta_Q(B^S) = (1+\theta_1 B+\theta_2 B^2+\ldots+\theta_q B^q)(1+\Theta_1 B^S+\theta_2 B^{2S}+\ldots+\theta_Q B^{QS})$$

az ún. mozgóátlag operátor-polinom.

Adott év t-edik időszaka és az előző év azonos időszaka közötti kapcsolatot leíró modell:

$$\Phi_P(B^S)\nabla_s^D y_t = \Theta_Q(B^S)\alpha_t \quad (3.8)$$

Az α_t, α_{t-1}, … hibatényező azonban általában még nem autokorrelálatlan (az adott év *t*-edik időszaka nemcsak az előző év azonos időszakával, hanem pl. az azt megelőző időszakokkal is kapcsolatban áll).

Tehát az α_t korrelációban állhat α_{t-1}-gyel, α_{t-2}-vel stb.

Ezt a kapcsolatot kezelő (leíró) modell:

$$\phi_p(B)\nabla^d\alpha_t = \theta_q(B)\varepsilon_t, \quad (3.9)$$

ahol az ε_t már fehérzaj-folyamat (tehát korrelálatlan és szórása, illetve várható értéke időben állandó).

Amennyiben a (3.9) modell képletét behelyettesítjük a (3.8) modellbe, akkor előáll az (3.7) általános modell.

E modell tipikus példája az ún. *Airline* modell, melynek jelölése: SARIMA (0, 1, 1) (0, 1, 1).

3.5. ARMA modellek ellenőrzése (modellverifikáció)

Az ARMA modellek struktúrájának, típusának, illetve a modell paramétereinek becslését, a modellidentifikációt, tehát a modell illesztését általában empirikus idősorok alapján végezzük. Bár a modell a valóságnak csak egy megközelítését jelenti, fontos, hogy az illeszkedés a megfigyelt adatokra minél jobb legyen. A modell jó illeszkedése elengedhetetlen feltétele a modell alapján levonható következtetések helyességének.

Az ARMA folyamatok az ún. lineáris folyamatok kategóriájába tartoznak, azaz felírhatók

$$\sum_{i=0}^{\infty} \varphi_i \varepsilon_{t-i} \quad (t = 0,1,2,\ldots)$$

formában, ahol ε_t fehérzaj-folyamat.

Mivel az alkalmazott ARIMA modellek a lineáris sztochasztikus folyamatok családjába tartoznak, így a modellek ellenőrzése során először is a standard lineáris modell kritérium-rendszerével kell foglalkozni.

3.6. A standard lineáris modell

Ahhoz, hogy a paraméterek értékeinek becslését vizsgálhassuk, illetve ellenőrizhessük, szükség van néhány feltétel teljesülésére.
Mivel az idősorok esetén az Y (eredményváltozó) sztochasztikus változó, tehát valamely adott t_i érték mellett a mintaelem Y ismérvértéke η_i valószínűségi változó, valamint a vizsgált idősorra illesztett modell ε_i reziduuma szintén valószínűségi változó, az alábbi (3.10) megszorítások nélkül a különböző t_i és t_j értékek mellett az ε_i és ε_j eloszlása, ennek következtében pedig az η_i és η_j valószínűségi változók eloszlása is különböző lehet. Az Y változónak a $t = t_i$ feltételre vonatkozó feltételes eloszlása szolgáltatja ugyanis az η_i eloszlását.
Az ún. *standard lineáris modell* feltételrendszere a következő kritériumokat támasztja a becsléssel szemben:

$$\left.\begin{array}{l} 1.\ \varepsilon_i \text{ és ezáltal } \eta_i \text{ is normális eloszlású, minden } i = 1, 2, \ldots, n \\ \qquad \text{mellett, különbségük pedig a } \beta_0 + \beta_1 t_i \text{ állandó érték;} \\ 2.\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \text{ ha } i \neq j \quad (i = 1, 2, \ldots, n\ ;\ j = 1, 2, \ldots, n); \\ 3.\ M(\varepsilon_i) = 0 \quad (i = 1, 2, \ldots, n); \\ 4.\ D(\varepsilon_i) = D(\eta_i) = \sigma \quad (i = 1, 2, \ldots, n); \quad \text{(függetlenül } t_i \text{ – től).} \end{array}\right\} \qquad (3.10)$$

3.7. A reziduumok korrelációs struktúrájána

A lineáris modell alkalmazhatóságának kritérium-rendszere szerint a modell reziduumainak autokorrelálatlannak kell lenni.

Az autokorrelációs függvény becslése az autokovarianca-függvény felhasználásával történik.
Az autokovarianciákra kétféle becslés adható. Az egyik a korrigált, a másik a korrigálatlan.
N számú megfigyelés, ill. standard idősor értékek mellett az autokovarianciák becslései:

$$\hat{C}_k = \frac{1}{N-k} \sum_{t=1}^{N-k} x_t \, x_{t+k} \qquad \text{(korrigált)}$$

illetve

$$\hat{c}_k = \frac{1}{N} \sum_{t=1}^{N-k} x_t \, x_{t+k} \qquad \text{(korrigálatlan)}$$

feltételezve, hogy az M[$x(t)$] = 0.

Megj.: Amennyiben az M[(x(t)] ≠ 0, akkor a korrigált:

$$\hat{C}_k = \frac{1}{N-k} \sum_{t=1}^{N-k} \left(x_t - \bar{x}\right) \cdot \left(x_{t+k} - \bar{x}\right)$$

ill. a korrigálatlan:

$$\hat{c}_k = \frac{1}{N} \sum_{t=1}^{N-k} \left(x_t - \bar{x}\right) \cdot \left(x_{t+k} - \bar{x}\right).$$

A $\hat{C}_k$ torzítatlan becslése az elméleti autokovarianciának.
A kétféle becslés azonban könnyen átszámítható egymásba

$$\hat{C}_k = \frac{N}{N-k} \hat{c}_k \, .$$

A korrigálatlan r_k autokorrelációs együttható becslése:

$$\hat{r}_k = \frac{\hat{c}_k}{\hat{c}_0} = \frac{\sum_{t=1}^{N-k} \left(x_t - \bar{x}\right) \cdot \left(x_{t+k} - \bar{x}\right)}{\sum_{t=1}^{N} \left(x_t - \bar{x}\right)^2} \, .$$

Ha M(x_t) = 0, akkor a korrigált R_k becslése:

$$\hat{R}_k = \frac{\sum_{t=1}^{N-k} x_t x_{t+k}}{\left(\sum_{t=1}^{N-k} x_t^2 \right)^{1/2} \left(\sum_{t=1}^{N-k} x_{t+k}^2 \right)^{1/2}} \, , \qquad k = 1, 2, \ldots, M \, .$$

Az itt szereplő *M* a maximális késleltetés.

Ha pedig $M(x_t) \neq 0$, akkor:

$$\hat{R}_k = \frac{\frac{1}{N-k}\sum_{t=1}^{N-k} x_t x_{t+k} - \frac{1}{(N-k)^2}\left(\sum_{t=1}^{N-k} x_t\right)\left(\sum_{t=1}^{N-k} x_{t+k}\right)}{\left\{\frac{1}{N-k}\sum_{t=1}^{N-k} x_t^2 - \frac{1}{(N-k)^2}\left(\sum_{t=1}^{N-k} x_t\right)^2\right\}^{1/2}\left\{\frac{1}{N-k}\sum_{t=1}^{N-k} x_{t+k}^2 - \frac{1}{(N-k)^2}\left(\sum_{t=1}^{N-k} x_{t+k}^2\right)^2\right\}^{1/2}}$$

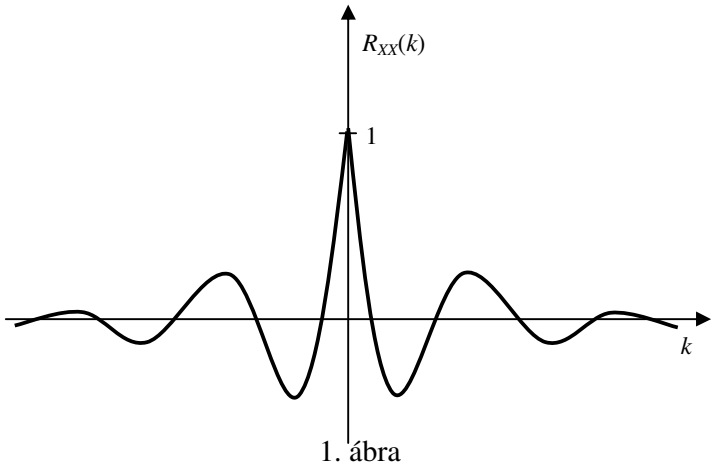
Teljesen véletlen folyamat esetén, amely egymástól független, azonos eloszlású és teljes mértékben autokorrelálatlan valószínűségi változók összessége, az autokovariancia függvény

$$C(k) = \mathrm{Cov}(X_t, X_{t+k}) = 0\,, \qquad k = \pm 1, \pm 2, \ldots$$

és az autokorrelációs függvény

$$r(k) = \begin{cases} 1, & \text{ha} \quad k = 0 \\ 0, & \text{ha} \quad k = \pm 1, \pm 2, \ \ldots \end{cases} .$$

Ez a folyamat, melyet *fehérzajnak* is hívnak, nemcsak gyenge, hanem erős értelemben is stacionárius. A fehérzaj fizikailag megvalósíthatatlan, a folyamatnak megfelelő autokorrelációs függvény az egység impulzusfüggvény (DIRAC-féle $\delta(x)$ függvény).
A gyakorlatban előforduló jellegzetes stacionárius idősorok azonban gyakran tartalmaznak rövid távú autokorrelációt. Viszonylag magas $r(1)$ érték esetén tehát az $r(2)$ és $r(3)$ is szignifikánsan eltér a nulla értéktől, de az időeltolás (k) növekedésével a burkológörbéje exponenciálisan nullához tart, ahogy ezt a következő ábra is szemlélteti.



1. ábra

A modell ellenőrzése tehát az illesztett ARMA modell szignifikanciájának vizsgálatát jelenti. Ennek a tesztelése történhet például az ún. p*ortmanteau próba* elvégzésével.[5].
Ez a próba a hibatényező korrelációs függvényének vizsgálatán alapszik.
A

$$H_0 : \text{ARMA}(p, q)$$

nullhipotézissel szemben a

$$H_1 : \text{ARMA}(p_1, q_1) \quad p_1 \geq p,\ q1 \geq q$$

alternatív hipotézist állítom.
Az ARMA(p, q) folyamatok esetén ugyanis a

$$Q_m = N\sum_{k=1}^{m}(\hat{r}_k^{\varepsilon})^2 \tag{3.11}$$

próbastatisztika χ^2_{K-p-q} eloszlású, ahol K egy megfelelően nagy érték, ahol N az idősor elemeinek száma.
Például egy 5 %-os szignifikancia szint mellett a Q_K 95 %-os megbízhatósággal kisebb értéket kell felvegyen, mint az inverz χ^2 – eloszlásfüggvény 0,95-os valószínűségi értékhez tartozó táblázati értéke. Ellenkező esetben az eltérést szignifikánsnak tekinthető és a nullhipotézist elutasítjuk.

A lineáris modell lényegi tesztelési eljárásaként először is megvizsgáljuk, hogy a reziduumok autokorrelálatlannak tekinthetők-e.
Amennyiben nem tekinthető annak, vagyis a hibatényező autokorrelált, akkor a becsült modell nem fogadható el, mert téves eredményre vezet.

Abban az esetben például, ha a hibatényező elsőrendű autokorrelációt tartalmaz, akkor a t-edik reziduális érték az alábbi módon áll elő:

$$\varepsilon_t = \begin{cases} 0 & \text{ha} \quad t = 0 \\ p\varepsilon_{t-1} + v_t & \text{ha} \quad t = 1, 2, \ldots \end{cases}, \tag{3.12}$$

ahol p az *autokorrelációs együttható*,
v_t pedig m várható értékű és konstans szórású valószínűségi változó.

Az iménti képlettel leírható folyamatokat a szakirodalmak *elsőrendű autoregresszív,* AR(1) *folyamatoknak* — ill. speciális esetben, amikor p=1 és $m \neq 0$, *bolyongási folyamatoknak* — nevezik.
A (3.12)-ből szukcesszív behelyettesítéssel élve adódik az

$$\varepsilon_t = v_t + p\,v_{t-1} + p^2\,v_{t-2} + \ldots + p^n\,v_{t-n} + \ldots\ , \tag{3.13}$$

ún. végtelenrendű *mozgóátlag*, vagy röviden *MA folyamat*. Amennyiben egy autoregresszív folyamat előállítható ily módon MA folyamat formájában, akkor a vizsgált folyamat *invertálható*.

[5] Ezt a csekély hatékonyságú próbát Box és Pierce vezette be 1970-ben, s azóta többen is foglalkoztak a javításával.

Az ún. *bolyongási folyamat* esetén például ε_t a fehér zajnak a $t=1$ időponttól — ahol a folyamat nulla értéket vesz fel — a t időpontig kumulált értéke:

$$\varepsilon_t = \varepsilon_{t-1} + v_t \,. \tag{3.14}$$

Ez a folyamat nyilvánvalóan nem stacionárius, hiszen ebben az esetben $M(\varepsilon_t) = t \cdot m$, ami idővel változó érték.

A $\boldsymbol{B}$ késleltető operátor bevezetésével azonban, melyre

$$\boldsymbol{B}^j \varepsilon_t = \varepsilon_{t-j} \quad [6] \qquad j \in N \,,$$

a (3.12) a következő módon is írható, ha $t = 1, 2, \ldots$:

$$(1-p\boldsymbol{B})\varepsilon_t = v_t$$

melyből

$$\varepsilon_t = \frac{1}{1-p\boldsymbol{B}} v_t = \left(1 + p\boldsymbol{B} + p^2\boldsymbol{B}^2 + \ldots\right) v_t \,, \tag{3.15}$$

mivel a fenti egyenlőség zárójeles kifejezésének, mint pB kvóciensű végtelen geo–metriai sornak az összege $\frac{1}{1-p\boldsymbol{B}}$.

A (3.15)-ben kapott eredmény pedig megegyezik (3.13)-mal. Látható, hogy amennyiben $|p\boldsymbol{B}| < 1$, akkor a (3.15)-ben szereplő végtelen sor konvergens és ily módon a (3.12) folyamat stacionárius.
Ha tehát egy AR folyamat előállítható konvergens végtelen MA alakban, akkor stacionárius. Ha $|p\boldsymbol{B}| \geq 1$, akkor a végtelen sor divergens, tehát ε_t minden határon túl a végtelenhez tart annak ellenére, hogy v_t várható értéke nulla. A szórása ugyanis nem az, így csak nulla körül szóródik, de nullától különböző értékeket vesz fel.

A vizsgálat során használt programcsomag (Demetra) diagnosztikái közül több is a reiduumok fenti tulajdonságaira visszavezethető próbastatisztikák alapján működik.

[6] Az **B**-vel jelölt késleltető operátor természetesen az $\mathbf{B}^j$ speciális esete, amennyiben j=1.

II. Szezonális kiigazítási módszerek

1. A szezonális kiigazítás módszereinek fejlődése – rövid történeti áttekintés[7]

A múlt század végén Angliában, az asztronómia és a meteorológia területén tevékenykedő kutatók fogalmazták meg először, hogy egy megfigyelt idősor több, meg nem figyelhető tényező összhatásának az eredménye. A korabeli kutatók úgy gondolták, hogy a két változó közötti „hamis" korrelációt a trend okozza, amelyet ezért előbb ki kell szűrni az idősorból. Poynting (1884) és Hooker (1901) a trendet és a szezonális hatásokat az árak több évre történő átlagolásával próbálta kiszűrni. Spencer (1904) és Andersen (1914) mutatta be magasabb fokú polinomok alkalmazását a trend leválasztására: A közgazdaságtan területén tevékenykedő kutatók a gazdasági ciklus kimutatása érdekében választották le a szezonális hatásokat és a trendet.
A 20-as és 30-as években kezdődött meg az igazán aktív kutatás a szezonális kiigazítás területén Person (1919) munkája következtében, amelyben felírta a híres képletet, ami szerint (multiplikatív modellt feltételezve) egy idősor kifejezhető a következő formában:

$$X_t = S_t \times T_t \times C_t \times R_t ,$$

ahol

S_t a szezonális ingadozást leíró komponens;

T_t a hosszútávú trend;

C_t a középtávú ciklus;

R_t a véletlen tényező.

Person módszere állandó szezonális tényezőket használt, annak ellenére, hogy a kor statisztikai irodalmában több publikáció szólt arról, hogy a fix szezonalitás feltételezése sok esetben nem helytálló. Sydensticker és Britten (1922) vezették be a formális szezonális kiigazító modellben a változó szezonális tényezőket. Crum 1925-ben módosította Person módszerét, hogy alkalmassá váljon a változó szezonalitás kezelésére.
A első átfogó szezonális kiigazitó rendszert Macauley (1930) dolgozta ki. A módszer három alaplépésből áll:

- Havi adatokat feltételezve 12 tagú centrírozott mozgóátlag leválasztásával és ezek átlagolása után fix szezonális tényezők előállítása;
- A trend becslése lineárisan vagy magasabb fokszámú polinom illesztésével;
- A mozgóátlagok osztása a trend becsült értékével a ciklikus komponens értékének becslését adja.

Ezt a módszert nevezik ma klasszikus dekompozíciónak, amely sok modern szezonkiigazító eljárásnak képezi alapját, így az X11-nek is.
Az 50-es évek során két fontos újítás történt. Az első az exponenciális simító eljárások elterjedése volt, amelyekkel az addigi számítások mennyiségét jelentősen csökkenteni lehetett legalább ugyanolyan jó eredmények mellett, mint a korábban használt módszerekkel. A másik fontos fejlődést a számítógépek megjelenése jelentette, ami szintén a számítások gyorsaságának növekedését segítette elő. Arra, amire korábban

[7] Björn Fischer (1995): Decomposition of Time Series Comparing Different Methods in Theory and Practice alapján

több nap kellett, most pár másodperc is elég volt. A kutatók így sokkal bonyolultabb modelleket dolgozhattak ki, hiszen az új verziókat könnyen tesztelhették nagy számú idősoron is.
A Census I. módszer 1954-ben jelent meg; ez annyival lépett túl a Macauley féle eljáráson, hogy az idősort egyszerű extrapolálással előre-hátra meghosszabbította, hogy pótolja a mozgóátlagolás során elvesztett elemeket. Ennek továbbfejlesztett változata a Census II. (1955), amely az akkor használatos technikák elektronikus verziója volt - kidolgozásában Julius Shishkin játszott úttörő szerepet.
A Census II-t több kritika is érte, mégpedig azzal kapcsolatban, hogy a módszer sok tekintetben ad hoc jellegű eljárásokra épül, amelyeket nem támaszt alá semmilyen statisztikai elmélet - a szezonális ingadozásoknak csak a kiugró részeit szűri ki, a többit benne hagyja stb. A modell folyamatos felülvizsgálatát követően alakult ki 1965-re az X11-es változat. Ennek legfőbb újdonsága a munkanap-tényező regressziós módszerrel történő kezelése. Ezen kívül az új változatban a felhasználó választhat az additív és a multiplikatív modell között és meghatározhatja, hogy milyen mozgóátlagot kíván használni.
Box és Jenkins az autoregresszív mozgóátlagolásra irányuló, 70-es években folytatott kutatásainak hatására fejlődött ki 1980-ra a kanadai statisztikai hivatalnál az X11-ARIMA változat, ami abban különbözött elődeitől, hogy az idősor előre és hátra történő meghosszabbítását az ARIMA modellezés segítségével végezte el. Ez a becslés minőségét javította. Az újabb X11-ARIMA/88 és X11-ARIMA/2000 verziók főleg diagnosztikákkal bővítették az addigi módszertant. A 90-es évek közepén jelent meg az X12-ARIMA program, amely új alapokra helyezte a munkanapok, ünnepnapok és outlierek[8] hatásának elemzését és a hiányzó adatok pótlását; valamint tovább bővítette a diagnosztikák körét.
A mozgóátlagolású technikák - amelyekről tudjuk, hogy erősen ad hoc jellegűek - mellett kifejlődtek modellszemléletű megközelítések is, amelyek között érdemes megkülönböztetni a determinisztikus és sztochasztikus jellegű modelleket. A determinisztikus modellek a trendet és a szezonalitást egy előre elrendelt pályának tekintik; amelyre a véletlen csak olyan módon gyakorol hatást, hogy eltéríti az idősor tényleges értékét ettől a pályától. A determinisztikus jellegű modellek a regressziószámításra épülnek, amely a trendet és a szezonalitást valamilyen determinisztikus módon megadott függvényekkel kezeli. Ebbe a modellcsaládba tartoznak a DAINTIES és a BV4 programok.
A sztochasztikus módszerek a véletlennek jelentős hatást tulajdonítanak, ez a modellezésben fontos szerepet játszik. Ezek története Yule autoregresszív (1927), illetve Slutsky mozgóátlagolású modelljéig (1937) nyúlik vissza. Wold alkalmazta először a mozgóátlagolású modellt valós adatokra, illetve ő dolgozta ki a vegyes ARMA modellek használatát (1954). Azonban a számítások nehézkessége miatt az ARMA modelleket csak nagyon kevesen használták, egészen a számítógépek széles körű elterjedéséig, illetve amíg Box és Jenkins meg nem fogalmazta azokat a kritériumokat, amelyekkel minden idősorra meghatározható egy konkrét típusú és fokú ARIMA modell. Ennek általános képlete szezonális idősorra:

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t,$$

ahol B a késleltetési operátor ($Bz_t = z_{t-1}$), így $\nabla z_t = (1-B)z_t$, továbbá $\phi_p(B), \Phi_P(B^s), \theta_q(B), \Theta_Q(B^s)$ a nemszezonális és szezonális autoregresszív és mozgóátlag polinomok, ε_t pedig fehér zaj folyamat.

[8] 2. ábra

A log (0,1,1)(0,1,1) modell (az ún. Airline modell, ami arról kapta a nevét, hogy egy légitársaság adatait használták a módszer bemutatására), viszonylag kevés paraméterrel jól illeszthető sok idősorra.
A modell első gyakorlati megvalósítása az Angol Nemzeti Bankban történt a '80-as években. Ezt fejlesztették tovább a Spanyol Nemzeti Bankban Augustin Maravall irányítása alatt, ennek eredménye a TRAMO/SEATS program. Annak ellenére, hogy a program egy teljesen új megközelítést takar, az utóbbi években nagyon széles körben elterjedt. [56]

2. TRAMO/SEATS

A TRAMO/SEATS teljesen modell-alapú megközelítést alkalmaz. Két programból áll, melyek egymással együttműködnek. A TRAMO végzi az idősor előfeldolgozását, linearizálását, kezeli a hiányzó megfigyeléseket és az outliereket, valamint a munkanap- és húsvét-hatást. A SEATS átveszi a linearizált idősort, és elvégzi az idősor komponensekre bontását, majd a szezonálisan kiigazított (trend-ciklus + irreguláris komponens) idősorba újra beveszi az outliereket.

2.1. TRAMO[9] (Time series Regression with ARIMA noise, Missing values and Outliers)

A TRAMO regressziót végez az idősorra ARIMA zajjal. A program a következő regressziós modellt illeszti:

$$z_t = y_t'\beta + v_t,$$

ahol $z = (z_{t_1}, \ldots, z_{t_M})$ a megfigyelések vektora, $\beta = (\beta_1, \ldots, \beta_n)'$ a regressziós együtthatók vektora, $y_t' = (y_{1t}, \ldots, y_{nt})$ az n darab regressziós változót jelöli, v_t pedig ARIMA folyamat, így a következő egyenlet teljesül rá:

$$\phi(B)\delta(B)v_t = \theta(B)a_t,$$

ahol B a backshift (visszatolás) operátor, $\phi(B)$, $\delta(B)$ és $\theta(B)$ B-ben polinomok és a_t normális fehérzaj 0 várhatóértékkel és σ_a^2 szórásnégyzettel.
A $\delta(B)$ polinom tartalmazza a differenciálás egységgyökeit (a szezonális differenciálását is), $\phi(B)$ jelöli az autoregresszív polinomot, míg $\theta(B)$ a mozgóátlag polinomot. Ezekről a polinomokról feltesszük, hogy a következő alakba felírhatóak:

$$\delta(B) = (1-B)^d (1-B^s)^D$$

$$\phi(B) = (1 + \phi_1 B + \cdots + \phi_p B^p)(1 + \Phi_1 B^s + \cdots + \Phi_P B^{s\times P})$$

$$\theta(B) = (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B^s + \cdots + \Theta_Q B^{s\times Q}),$$

ahol s jelöli az évenkénti megfigyelések számát.
A regressziós változókat megadhatja a felhasználó, vagy generálhatja a program. A program által generálható változók a munkanap változók (egy, kettő, hat vagy hét darab), a húsvét-hatás változó, valamint egyéb változók a következők szerint:

[9] A programok következő ismertetése Gómez és Maravall (1996) cikkén alapul.

a) dummy (mesterséges) változók
b) egyesek és nullák tetszőleges sorozata
c) $1/(1-\delta B)$ ráalkalmazva egyesek és nullák sorozatára, ahol $0 < \delta \leq 1$
d) $1/(1-\delta_s B^s)$ ráalkalmazva egyesek és nullák sorozatára, ahol $0 < \delta_s \leq 1$
e) $1/(1-B)(1-B^s)$ ráalkalmazva egyesek és nullák sorozatára.

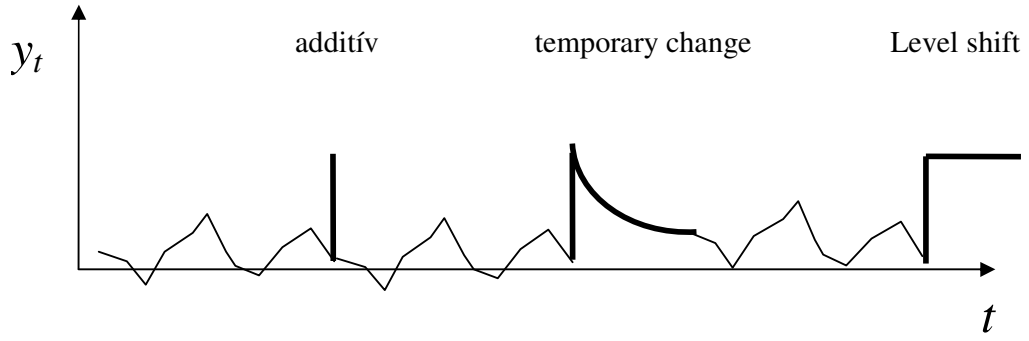
A program:

1) megbecsli a regressziós modell és az ARIMA folyamat paramétereit maximum likelihood módszerrel (vagy feltétel nélküli/feltételes legkisebb négyzetek módszerével);
2) detektálja és korrigálja a különféle típusú outliereket (2. ábra);
3) kiszámítja az idősor optimális előrejelzését és ennek átlagos négyzetes hibáját (MSE);
4) a hiányzó megfigyelések értékét optimálisan interpolálja és kiszámítja a hozzájuk tartozó átlagos négyzetes hibát
5) rendelkezik automatikus modell identifikációs és automatikus outlier kezelési lehetőséggel.

A regressziós paraméterek és az ARIMA modell paramétereinek becslése történhet az előbbiek kiemelésével a likelihood függvényből, vagy együttes becslést alkalmazva. Számos algoritmus áll rendelkezésre a likelihood függvény számítására, vagy pontosabban a nemlineáris legkisebb négyzetek minimalizálására (pl. a differenciált idősorokra Morf, Sidhu és Kailath (1974) algoritmusa). A nemdifferenciált idősorokra a közönséges Kalman szűrő, vagy négyzetgyökös verziója (lásd Anderson és Moore, 1979) alkalmazható. Ez utóbbi numerikus nehézségek felmerülése esetén megfelelő, bár határozottan lassabb.
A regressziós paramétereket kiemelve a likelihood függvényből, a becslésükhöz először az inverz hiba kovariancia mátrixra alkalmazandó a Cholesky felbontást, amivel transzformálni kell a regressziós egyenletet (a Kalman szűrővel hatékonyan számolhatók a transzformált regresszió változói). Ezután a kapott legkisebb négyzetes feladatot ortogonális mátrixfaktorizációval (QR felbontás) meg lehet oldani, amihez a Householder transzformáció használandó. Ez az eljárás hatékony és numerikusan stabil módszer, amely mátrix inverzió nélkül számítja ki a regressziós paraméterek általános legkisebb négyzetes becslését.
Az előrejelzés lehetséges a közönséges Kalman szűrővel, vagy annak négyzetgyökös verziójával. A hiányzó megfigyelések interpolációja egy egyszerűsített fix-pont simítóval történik, ami Kohn és Ansley (1986) eredményeivel azonosat ad. A likelihood függvényből kiemelve a regressziós paramétereket, az előrejelzés és az interpoláció átlagos négyzetes hibája Kohn és Ansley (1985) megközelítését követve kapható meg.
Amikor a kezdeti hiányzó értékek közül néhány becsülhetetlen (szabad paraméter), akkor a program detektáltja őket, és megjelöli azokat az előrejelzéseket illetve interpolációkat, amelyek ezektől a szabad paraméterektől függenek. A felhasználó ezután tetszőleges (tipikusan nagyon nagy vagy kicsi) értékeket rendel a szabad paraméterekhez, és újra futtatja a programot. Ezen a módon eljárva az ARIMA modell összes paramétere megbecsülhető, mivel a minimalizálandó függvény nem függ a szabad paraméterektől. Ezenfelül, nyilvánvaló lesz, hogy mely előrejelzésekre és interpolációkra hatnak ezek a tetszőleges értékek, mert a többi becsléstől erősen el fognak térni. Ennek ellenére, ha az összes ismeretlen paramétert együttesen becsüljük, lehetséges, hogy a program nem jelzi az összes szabad paramétert.

A hiányzó megfigyelések additív outlierként is kezelhetők. Ebben az esetben a likelihood függvény korrigálható, úgy hogy egybeessen a standard hiányzó megfigyeléses eset likelihood függvényével.
A program tudja detektálni az outliereket és kiiktatni a hatásukat; az outliereket beviheti a felhasználó, vagy automatikusan detektálja a program, ehhez egy Tsay (1986) és Chen és Liu (1993) munkáin alapuló megközelítést használ. Az outliereket egyesével detektálja Tsay (1986) javaslata szerint, és többszörös regressziót használ a hamis outlierek detektálására, mint Chen és Liu (1993). Az outlierek elfogadására illetve elutasítására használt eljárás hasonlít a „legjobb" regressziós egyenlet kiválasztásához használt lépésenkénti regressziós eljáráshoz. Így egy robusztusabb eljárást kapunk, mint amit Chen és Liu (1993) használ, ott ugyanis az outlierek „visszafelé menő" eliminációja történik, ami miatt az eljárás első lépésében túl sok outliert detektálhat.

Röviden, a Tramo regressziós paramétereket közönséges legkisebb négyzetes becsléssel inicializálja, és az ARMA modell paramétereit először két regresszióval becsüli, mint Hannan és Rissanen (1982). Ezután a Kalman szűrő és a QR algoritmus új regressziós paraméter becsléseket és regressziós reziduálisokat adnak. Minden egyes megfigyelésnél t-próbát végez a négy fajta outlierre, mint Chen és Liu (1993). Ha vannak olyan outlierek, amelyek t-értékei (abszolút értékben) nagyobbak, mint egy előre kiválasztott C kritikus szint, akkor az abszolút értékben legnagyobb t-értékűt választja, különben az idősor mentes az outlier-hatásoktól, és az algoritmus megáll.
Hogyha a program detektált outliert, akkor az idősort megtisztítja a hatásuktól, és az ARMA modell paramétereit újra becsli. Majd egy többszörös regressziót hajt végre a Kalman szűrő és a QR algoritmus segítségével. Ha így az outlierek közül valamelyik t-értéke abszolút értékben kisebb a kritikus C szintnél, akkor a legkisebb abszolút értékű t-értékkel rendelkező outliert kiveszi a regresszióból, majd a többszörös regressziót újra becsli. A következő lépésben a legutolsó többszörös regresszió reziduálisait használva t-próbát végez a négyféle típusú outlierre és minden egyes megfigyelésre. Ha van olyan outlier, amelynek a t-értéke abszolút értékben nagyobb a C kritikus szintnél, akkor a legnagyobb abszolút értékű t-értékkel rendelkező outliert kiválasztja, és az ARMA modell paramétereinek becslésével folytatja az iterációt. Különben az algoritmus megáll.
Ennek az algoritmusnak figyelemre méltó tulajdonsága, hogy az összes számolás lineáris regressziós technikákra épül, ami csökkenti a számítási időt. A figyelembe vett outlierek a következők: additív, innovációs (innovational), szint eltolás (level shift), csillapodó törés (temporary change).
Outlierek típusai (az innovációs ritkán kerül alkalmazásra a gyakorlatban):



2. ábra

A program az automatikus ARIMA modell identifikációra is tartalmaz eljárást. Ez két lépésben megy. Az első adja a modell nemstacionaritási polinomját, $\delta(B)$-t. Ez AR és ARMA(1,1) modellek egy sorozatán való iterációval történik. Az eljárás Tiao és Tsay (1983, 3.2 és 4.1 Tétel) valamint Tsay (1984, 2.1 Korollárium) munkáin alapul. Reguláris (nemszezonális) és szezonális differenciákat nyer, egészen a maximális $\nabla^2\nabla_s$ rendig, ellenőrizve a lehetséges komplex egység gyököket nemnulla és nemszezonális frekvenciáknál.
A második lépés azonosítja az ARMA modellt a stacionárius idősorra (outlierektől és regressziós típusú hatásoktól megtisztítva) a Hannan-Rissanen eljárást követve, egy olyan finomítással, hogy a modell innovációinak varianciájának becslésénél az első reziduálisokat nullák helyett a Kalman szűrővel számolja. Az általános multiplikatív modellre:

$$\phi_p(B)\Phi_P(B^s)x_t = \theta_q(B)\Theta_Q(B^s)a_t$$

a keresés a $0 \le (p,q) \le 3$, $0 \le (P,Q) \le 2$ tartományban történik. Ezt szekvenciálisan csinálja a program (rögzített reguláris polinomokra a szezonális polinomokat kapjuk és fordítva), és a polinomok végső rendjét a BIC kritériumnak megfelelően választja, kiegészítve néhány lehetséges feltétellel, amiknek célja a takarékosság növelése és a „kiegyensúlyozott" (balanced) modellek (az AR és MA rendek közel vannak egymáshoz) előnyben részesítése.
Végül a program kombinálja az outlierek automatikus detektálását és korrekcióját a most leírt automatikus ARIMA modell identifikációval.
Bár a TRAMO használható önmagában is, például előrejelzésre, de úgy is tekinthető, mint ami megtisztít egy „szennyezett" ARIMA sort. Azaz, adott idősorra interpolálja a hiányzó megfigyeléseket, azonosítja az outliereket és hatásukat eltávolítja, becsli a munkanap- és a húsvét-hatást, végül szolgáltat egy lineáris, tisztán sztochasztikus folyamatot (azaz az ARIMA modellt). Így a TRAMO használható úgy, mint egy előfeldolgozó a SEATS-hez, ami aztán felbontja a „linearizált" idősort sztochasztikus komponenseire. [56]

2.2. SEATS (Signal Extraction in ARIMA Time Series)

A program az idősorok nem megfigyelt komponensekre bontásának úgynevezett ARIMA modell alapú módszerei közé tartozik. Az alapvető irodalom a következő: Cleveland és Tiao (1976), Box, Hillmer, és Tiao (1978), Burman (1980), Hillmer és Tiao (1982), Bell és Hillmer (1984), Maravall és Pierce (1987). Ezek a megközelítések szoros kapcsolatban vannak egymással és a programban használttal. A SEATS-et Burmannek az Angol Nemzeti Bankban szezonális kiigazításra írt programjából (1982-es verzió) fejlesztették ki.

A program egy ARIMA modell illesztésével kezd. Jelölje x_t az eredeti sort (vagy annak logaritmikus transzformációját), és

$$z_t = \delta(B)x_t$$

jelölje a differenciált sort, ahol B a visszatolás operátor, és $\delta(B)$ jelöli a stacionaritás eléréséhez az x_t-re vett differenciálásokat. A SEATS-ben

$$\delta(B) = \nabla^d \nabla_s^D,$$

ahol $\nabla = 1 - B$ és $\nabla_s^D = (1 - B^s)^D$ jelöli az s periódusú szezonális differenciálást. A z_t differenciált sorra a modellt a következőképpen lehet kifejezni:

$$\phi(B) z_t = \theta(B) a_t + \mu,$$

ahol μ konstans, a_t normális eloszlású fehérzaj folyamat, nulla várható értékkel és σ_a^2 szórásnégyzettel, $\phi(B)$ és $\theta(B)$ autoregresszív (AR) és mozgóátlag (MA) polinomok, amiket ki lehet fejezni, mint egy reguláris (nemszezonális), B-beli polinom és egy szezonális, B^s-beli polinom szorzatát:

$$\phi(B) = \phi_r(B)\phi_s(B^s),$$
$$\theta(B) = \theta_r(B)\theta_s(B^s).$$

A fentieket visszahelyettesítve a z_t-re vonatkozó egyenletbe:

$$\phi_r(B)\phi_s(B^s)\nabla^d \nabla_s^D x_t = \theta_r(B)\theta_s(B^s) a_t + \mu.$$

A $\phi(B)$ autoregresszív polinomnak lehetnek egység gyökei, amiket tipikusan nagy pontossággal becsülnek. Például, egység gyökök jelentkezhetnek $\phi(B)$-ben, ha az idősor nemstacionárius ciklikus komponenst tartalmaz, vagy ha a sor aluldifferenciált. Úgy is megjelenhetnek, mint nemstacionárius szezonális harmonikusok.

A program a sort több különféle komponensre bontja. A dekompozíció lehet multiplikatív, vagy additív. Mivel az előbbi logaritmálással az utóbbiba alakítható, ezért jelen tárgyalásban az additív modellel foglalkozom, azaz:

$$x_t = \sum_i x_{it},$$

ahol x_{it} jelöl egy komponenst. A SEATS által figyelembe vett komponensek a következők:

x_{pt} = a TREND komponens,

x_{st} = a SZEZONÁLIS komponens,

x_{ct} = a CIKLIKUS komponens,

x_{ut} = az IRREGULÁRIS komponens.

A trend komponens képviseli az idősor hosszútávú fejlődését, és spektrális csúcsot mutat a nulla frekvenciánál. A szezonális komponens jeleníti meg a szezonális frekvenciáknál jelentkező spektrális csúcsokat, és az irreguláris komponens képviseli a szabálytalan, fehérzaj viselkedést, és így lapos spektruma van. A ciklikus komponens képviseli a szezonálisan kiigazított sor trendjére vonatkozó, a tiszta fehérzajtól különböző eltéréseket. Így a ciklust hozzáadva a SEATS irreguláris komponenséhez, kapjuk a gazdasági ciklus standard definícióját; ld. például Stock és Watson (1988). A komponenseket teljes egészében a megfigyelt sorra vonatkozó ARIMA modell struktúrájából vezetjük le, az ARIMA modell pedig közvetlenül identifikálható az adatokból. A programot leginkább a havi vagy ritkább gyakoriságú adatokkal való munkára szánták, a megfigyelések maximális száma 600 lehet.

A felbontás azt feltételezi, hogy a komponensek ortogonálisak, és mindegyik külön-külön ARIMA modellel felírható. Annak érdekében, hogy azonosítsuk a komponenseket, szükséges, hogy azok (az irreguláris komponens kivételével) zajtól mentesek legyenek. Ezt hívják kanonikus tulajdonságnak, és magában foglalja, hogy az irreguláristól különböző komponensekből nem lehet kinyerni hozzáadott fehérzajt. Az irreguláris komponens szórása ily módon maximalizált, ezzel szemben a trend, szezonális és ciklikus komponens annyira stabil, amennyire csak lehetséges. Bár önkényes feltevés, mivel bármely más elfogadható komponens kifejezhető a kanonikusnak és független fehérzajnak az összegeként, mégis ésszerű elkerülni a komponens zajjal „szennyezését", hacsak nincsenek a-priori okai annak, hogy így tegyünk.

A SEATS olyan modellt tételez fel, hogy az idősor lineáris, gaussi innovációkkal. Amikor ez a feltételezés nem teljesül, a SEATS képes együttműködni a TRAMO-val, ami eltávolítja a sorból a különleges hatásokat, azonosítja és eltávolítja a különféle típusú outliereket, és interpolálja a hiányzó megfigyeléseket. Rendelkezik egy automatikus modell identifikációs lehetőséggel is. Az ARIMA modell becslése az egzakt maximum likelihood módszerrel történik, ez megtalálható Gómez és Maravall (1994) munkájában. Legkisebb négyzetes típusú algoritmusok szintén rendelkezésre állnak.

Az AR és MA polinomok (inverz) gyökeinek mindig az egységkörben illetve az egységkör belsejében kell lenniük. Amikor egy gyök abszolút értéke konvergál egy előre beállított, 1 körüli intervallumon belül (alapbeállítás: [0.97,1]), a program automatikusan rögzíti a gyököt. Ha ez egy AR gyök, akkor az abszolút értéket 1-re állítja, ha MA gyök, akkor az alsó végpontra (alapbeállítás: 0.97). Ez az egyszerű tulajdonság a programot nagyon robusztussá teszi a túl- és aluldifferenciálással szemben.

Az ARIMA modellt használja a TRAMO által linearizált sor szűrésére, és az új reziduálisokat vizsgálja. Ezután a program rátér az ARIMA modell felbontására. Ez frekvencia-tartományban történik.
A SEATS-módszer dekompoziciós eljárása a spektrumfelbontására épül. Régóta alkalmazott függvényvizsgálati technika a Fourier-analízis, amelynél egy periodikus függvényt különböző frekvenciájú és amplitúdójú szinuszok és koszinuszok végtelen összegére bontunk fel, ahol a frekvenciák előre adottak, és a függvényt a megfelelő frekvenciához tartozó amplitúdókkal jellemezzük. Az eljárás alkalmazható nem periodikus függvények és sorozatok vizsgálatára is.
A Fourier-analízis mintájára vezették be az idősorok vizsgálatára a spekrál-analízist (vagy frekvenciatartománybeli elemzést). A spektrál-analízis azt a megközelítést alkalmazza, hogy egy stacionárius idősor autokovariancia függvényének értékeit írja fel egy ún. spektrális sűrűségfüggvény Fourier-együtthatóiként. A spektrális sűrűségfüggvényt röviden spektrumnak szokás nevezni, és a [-л, л] intervallumon értelmezett szimetrikus, nemnegatív függvényről van szó.
Spektrális sűrűségfüggvény nem minden stacionárius idősorhoz létezik, de a létezéshez elégséges, hogy az autokovarianciák abszolutértékeinek összege véges legyen. Ez a legtöbb gyakorlatban előforduló idősornál feltehető.
Ilyenkor a spektrum egyszerűen felírható a következőképpen:

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-ik\lambda} \gamma(k)$$

ahol a γ az autokovariancia függvény.

Ennek az összefüggésnek a segítségével bizonyos folyamatok spektruma egyszerűen számolható, például egy σ szórású fehérzaj spektruma $\frac{\sigma^2}{2\pi}$, azaz konstans.
A spektrumban megjelenő csúcsok az adott frekvenciához tartozó periodikus viselkedést jelzik az idősorban. A spektrum vizsgálata segítséget jelent akkor is, amikor úgynevezett időinvariáns lineáris szűrők hatását vizsgáljuk. Például az ARMA-folyamat is értelmezhető egy speciális időinvariáns lineáris szűrő outputjaként, amit a fehérzajra, mint inputra alkalmaznak. Ilyen esetben az input folyamat spektrumának és a szűrő úgynevezett négyzetes erősítésének (squared gain) szorzataként áll elő az output spektruma. Az X család és a TRAMO/SEATS szezonális kiigazításának eredménye is értelmezhető az eredeti idősorra vett valamilyen időinvariáns lineáris szűrő outputjaként, ezért a spektrális megközelítés hasznosnak bizonyult a szezonális kiigazítás különféle módszereinek összehasonlításánál is.

A spektrum előállításánál az elméleti autokovariancia függvényből indulunk ki, tehát elméleti spektrumról van szó, amelyet az idősor tényleges realizációja alapján becsülni kell. A becslésre alapvetően két mód kínálkozik. Az egyik az ún. periodogrammal (ill. simított periodogrammal) való becslés, amelynél lényegében az autokovariancia függvény becslésének segítségével történik a spektrum becslése.
A másik lehetőség, hogy az idősorra valamilyen modellt (például ARMA) becslünk és a modell elméleti spektrumának ismeretében kiszámítjuk az idősor spektrumát. Ez utóbbi megközelítést alkalmazza a TRAMO/SEATS módszer a spektrum becslésére.

A spektrumot a fentiekben stacionárius idősorokra értelmezzük, de lehetőség van bizonyos nemstacionárius idősorokra a spektrumhoz hasonló ún. pszeudo-spektrumot értelmezni. A (szezonális) ARIMA- modell pszeudo-spektrumát úgy kapjuk meg, hogy egy olyan ARMA-modell spektrumát írjuk fel, ahol a differenciaképzést beleértjük az autoregresszív részbe, azaz egységgyököket is megengedünk az autoregresszív polinomban. Ilyen módon a spektrumnak bizonyos helyeken pólusai lesznek, azaz a reguláris (sima) differenciálásnak megfelelő zérus illetve a szezonális diffrerenciálásnak megfelelő szezonális frekvenciáknál a spektrum értéke végtelen. Képlettel: ha x_t egy szezonális ARIMA-folyamat, és $\Phi(B)x_t=\Theta(B)a_t$, ahol a Φ az autóregresszív polinom, amely a differenciálást is tartalmazza, Θ a mozgó átlag polinom, a_t pedig σ szórású fehérzaj, akkor az x_t folyamat pszeudóspektruma:

$$f_x(\lambda) = \frac{\sigma^2 \left|\theta\left(e^{-i\lambda}\right)\right|^2}{2\pi \left|\phi\left(e^{-i\lambda}\right)\right|^2}.$$

A spektrumot (vagy pszeudo-spektrumot) felosztjuk a különböző komponensekhez hozzárendelt spektrumok összegére. (Ezeket főként a modell AR gyökeiből határozza meg a program). A trend, szezonális és ciklikus komponensekre vonatkozó kanonikus feltétel egyértelmű felbontást határoz meg, amiből a komponensek ARIMA modelljei megkaphatóak (beleértve a komponens innovációs varianciákat). Ha az ARIMA modell nem tesz lehetővé elfogadható dekompozíciót, akkor helyettesíti egy felbontható közelítéssel.

Egy konkrét $[x_1, x_2, \ldots, x_T]$ realizációra a program kiszámítja a komponensek Minimum Átlagos Négyzetes Hiba (Minimum Mean Square Error /MMSE/) becslését, amit úgy

számol, hogy egy Wiener-Kolmogorov típusú szűrőt alkalmaz a véges idősorra, kiterjesztve azt a „jövő” illetve a „múlt” felé (l. Burman, 1980). Minden $t = 1,\ldots,T$-re és minden i komponensre kiszámítja az $\hat{x}_{it|T}$ becslést, ami egyenlő az $E(x_{it}|x_1,\ldots,x_T)$ feltételes várható értékkel.

Amikor $T \to \infty$, akkor az $\hat{x}_{it|T}$ becslésből lesz az $\hat{x}_{it}$ „végső” becslés. (A gyakorlatban ez elég nagy $k = T - t$-re történik, és a program jelzi mekkora k-t lehet feltételezni.) $t = T$-re a párhuzamos becslést, $\hat{x}_{iT|T}$-t vesszük, ami a sor utolsó megfigyelésének becslése. Alkalmazási nézőpontból a végső és a párhuzamos becslések a leginkább érdekesek. Amikor $T - k < t < T$, $\hat{x}_{it|T}$ ad egy előzetes becslést, míg $t > T$-re egy előrejelzést. A komponensek becslése mellett a program sok év hosszúságú előrejelzést is ad, és a standard hibáit az összes becslésnek és előrejelzésnek. Az utolsó kettő és a következő két évre vonatkozólag az előzetes becslés és az előrejelzés revíziójának standard hibáját is szolgáltatja. Továbbá a program kiszámítja minden egyes komponensben lévő innovációk Minimum Átlagos Négyzetes Hibáját.

A komponensek (illetve azok stacionárius transzformáltjai) és azok MMSE becsléseik együttes eloszlását kiszámítja; a szórásnégyzetekkel, auto- és keresztkorrelációkkal karakterizálhatók. Az MMSE becslések elméleti és a tapasztalati momentumainak összehasonlítása további elemzendő adat. A program a szűrőt is kiadja, ami a súlyokat írja le, amivel a különféle a_j innovációk a megfigyelt folyamatban hozzájárulnak az $\hat{x}_{it|T}$ becsléshez. Ezek a súlyok közvetlenül adják a mozgó átlag kifejezést a revíziókra. Ezután végrehajt a trend és a szezonálisan kiigazított idősor becslési hibájára egy analízist. Jelölje

$$d_{it} = x_{it} - \hat{x}_{it},$$
$$d_{it|T} = x_{it} - \hat{x}_{it|T},$$
$$r_{it|T} = \hat{x}_{it} - \hat{x}_{it|T}$$

a végső becslési hibát, az előzetes becslési hibát és a revíziós hibát az $\hat{x}_{it|T}$ előzetes becslésben. A d_{it}, $d_{it|T}$, $r_{it|T}$ szórásnégyzeteit és autokorrelációs függvényeit kijelzi a program. Ez után megmutatja, hogy csökken a revíziós hiba szórásnégyzete a párhuzamos becslésben, ahogy több megfigyelést hozzáadunk, és így azt is megmutatja, hogy a gyakorlatban milyen gyors a konvergencia a végső becsléshez. Hasonlóan a program kiszámítja a pontosság romlását, ahogy az előrejelzés távolodik a párhuzamos becsléstől.

Amikor a TRAMO-t és a SEATS-et együtt futtatjuk, akkor a hatásokat , amiket a dekomponálás érdekében a TRAMO eltávolított az idősorból, a végső komponensekbe visszateszi. Így például a szint eltolás outliereket a trendhez rendeli, míg az időleges változás és additív outlierek az irreguláris tagba kerülnek; a szezonális komponens tartalmazni fogja a munkanap- és húsvét-hatást.

A programban két automatikus modell identifikációs eljárás áll rendelkezésre. Az egyik pontosabb de lassabb; a másik egyszerűbb és nagyon gyors, és nagyszámú idősor esetén ennek használata javasolt. Ez az egyszerűsített eljárás az alapmodellre épül, és csak akkor keres más beállítást, ha a sor tisztán eltér az alaptól. Az alapmodell az úgynevezett Airline modell, melyet Box és Jenkins (1970) vizsgált. Az Airline modell

gyakran bizonyul megfelelőnek sok sorra, és jól viselkedő becslési szűrőt ad a komponensekre. A modellt a következő egyenlet írja le:

$$\nabla\nabla_{12}x_t = (1+\theta_1 B)(1+\theta_{12}B^{12})a_t + c,$$

ahol $-1<\theta_1<1$ és $-1<\theta_{12}<0$, és x_t a sor logaritmusa. A komponensek a következő típusúak:

$$\nabla^2 x_{pt} = \theta_p(B)a_{pt},$$

$$Sx_{st} = \theta_s(B)a_{st},$$

ahol $S = 1+B+\cdots+B^{11}$, és $\theta_p(B)$ és $\theta_s(B)$ foka 2 illetve 11. Más rögzített szűrőkhöz képest a SEATS rendelkezik egy előnnyel: három paraméter határozható meg: θ_1, ami a trend komponens stabilitásával van kapcsolatban, θ_{12}, ami a szezonális komponens stabilitásával kapcsolatos, és σ_a^2, ami a sor teljes jósolhatóságának egyfajta mértéke. Így, még az alapmodellt véve is, a komponensek becsléseire vonatkozó szűrők alkalmazkodnak minden egyes sor egyéni struktúrájához.

Annak ellenére, hogy a TRAMO-SEATS modell-alapú megközelítést alkalmaz, hatékony alternatívája lehet „rutin felhasználásra" a (többé-kevésbé) rögzített szűrős eljárásoknak (lásd Fischer, 1994), míg eközben sokkal gazdagabb outputot szolgáltat, különösen, ami a rövidtávú következtetéseket illeti. [56]

3. X12-ARIMA

Az X12-ARIMA módszer dekompozíciós része lényegében megegyezik az X11-ARIMA azonos lépéseivel. Amiben a két program eltér, az a munkanaphatás és az outlierek kezelése. Az X11-ARIMA először végrehajtja a szezonális dekompozíciót, azaz leválasztja a trendet és a szezonális komponenst, majd megvizsgálja, hogy a maradék tényezőkben maradtak-e outlierek, és, hogy a munkanaphatás és a húsvéthatás kimutatható-e szignifikánsan a véletlen hatásban. Ha igen, akkor az előzetes korrekciókat ez alapján módosítja.
Az X12 ugyanakkor az eredeti idősorból kiindulva állapítja meg, hogy van-e szignifikáns munkanap-, húsvét- vagy outlier hatás. Ha ezen hatások bármelyike szignifikáns, akkor először igazítja az idősort, majd a korrigált idősorra hajtja végre a dekompozíciót.
Az X11-ARIMA-hoz hasonlóan- a mozgóátlag technika alkalmazása miatt – az X12-ARIMA esetén is szükség van arra, hogy az idősort meghosszabbítsa előre és hátra egyaránt. Az X12-nél a meghosszabbításhoz használt Arima –modell paramétereinek becslésével egyidőben történik a naptárihatások, outlierek regressziós változóinak meghatározása is.
Így a dekompozíció előtt az X12 egy regARIMA-modellt illeszt az idősorra:

$$Y_t = \sum \delta_i X_{it} + z_t,$$

ahol z_t egy szezonális ARIMA-folyamat: (p,d,q) $(P,D,Q)s$. Az egyenletet kifejezve z_t-re és ezt behelyettesítve az ARIMA modell egyenletébe, a következő formát kapjuk:

$$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D\left(Y_t-\sum\delta_i X_{it}\right)=\theta_q(B)\Theta_Q(B^s)\varepsilon_t,$$

ahol az előző felyezetben használt jelöléseknek megfelelően *B* a késleltetési operátor, $\phi_p(B), \Phi_p(B^s), \theta_q(B), \Theta_Q(B^s)$ a nemszezonális és szezonális autoregresszív és mozgóátlag polinomok, ε_t fehérzaj.

Kiválasztja a megfelelő regressziós változókat, megbecsli az együtthatókat, képezi az $Y_t-\sum\delta_i X_{it}$ változót. Ennek képezi a megfelelő differenciáit és becsli a szezonális ARIMA- modellt. Ez utóbbi az X11-ARIMA módszerhez hasonlóan arra szolgál, hogy előre-hátra meghosszabbítsa az idősort.

Az X12 regARIMA részében különböző regressziós változók alkalmazhatók, például trendkonstans, fix szezonális változó, munkanaphatás, a hónap hosszának kezelése, a szökőév kezelése, a húsvéthatás, outlierek.

Munkanaphatás

A X11-ARIMA-tól eltérően a munkanapokat nem hét, hanem hat változóval kezeli. A változók értéke: az adott nap száma – vasárnapok száma. Lehetőség van olyan opció választására is, hogy a munkanapokat egy regressziós változóval kezelje a program. Ebben az esetben minden hónaphoz hozzárendeli, azaz számszerűsíti, hogy a munkanapok száma mennyivel több, mint a hétvégi napok száma: a [hétköznapok száma-5/2 (szombatok és vasárnapok száma)] értékét rendeli hozzá.

A húsvét hatása

A húsvéthatást is egyszerűbben kezeli, mint az X11-ARIMA. A regressziós változó értéke attól függ, hogy az ünnep előtt hány napig lehet érezni a hatását. Márciusban és áprilisban 1/v, ahol v a húsvét előtt adott hónapra eső napok számát jelenti, a többi hónapban 0 az értéke.

Szökőévhatás

Az X12-ARIMA már a szökőév hatását is figyelembe veszi. A szökőév kezelésére bevezetett regressziós változó értéke 0,75, ha szökőév februárja van, és -0,25 a többi februárban, illetve 0 a többi hónapban.

Outlierek típusa

Az X12-ARIMA is az outlierek (kiugró értékek) 4 típusát különbözteti meg: additív outlier, szinteltolódás, csillapodó jellegű törés, átmeneti színváltás.

Az outlierek becslése történhet automatikus kereséssel, de a felhasználónak is lehetősége van az időpontok megadására. Az automatikus keresés két részből áll: a forward és a backward változószelekciós részből.

A program először a véletlen tényezőre számol standard hibát. Megbecsli a regressziós együtthatókat és az ARIMA-modell paramétereit, majd a véletlen tényezők mediántól vett átlagos abszolút eltérését számítja ki. Ennek 1,49-szorosát tekinti a véletlen tényező

szórásának. Ennek segítségével *t* értéket számol, és ezt hasonlítja a *t*-statisztika értékéhez. A forward eljárás keretében egy adott *t* kritikus értékhez (alapbeállításban 3,8) viszonyítja a modellbe be nem vont változók bevonása melletti *t* értékét. Kiválasztja a legnagyobb abszolút értékűt, és ha ez meghaladja a kritikus értéket, akkor bevonja a változót a modellbe, majd újra becsli a modellt. Az algoritmust addig folytatja, amíg talál a kritikus *t* értéknél nagyobb abszolút értéket.

Az előző forward eljárás során kapott változókból indul ki a backward algoritmus. A program itt lépésenként kihagyja azt a változót (az abszolút értékben legalacsonyabb *t* értékűt), amely nem felel meg az adott kritériumoknak.

Végül az utolsó lépés az ARIMA-modell típusának meghatározása. A programban 5 alapmodell található, de kiindulópontként bármilyen tetszőleges modell megadható. Az alapmodellek szezonális modell része mindig (0,1,1) alakú, a nem szezonális rész modellje pedig mindig (0,1,1)(0,1,2), (2,1,0), (0,2,2), (2,1,2) alakú lehet. A program automatikusan választ a modellek közül az Akaike-féle információs kritérium(AIC) alapján, de a diagnosztikák (autokorrelációs és parciális autokorrelációs függvények, erre vonatkozó Box Pierce- és Ljung-Box- próbák) segítségével a felhasználó is dönthet az alkalmazandó modellről. A modell illesztése után a program két kritérium teljesülését is figyeli: egyrészt az utolsó három évre az átlagos abszolút hibának az átlag százalékában kifejezett értéke ne haladja meg a 15%-ot, másrészt a Ljung- Box-statisztika értéke 0.1%-os szinten ne elegyen szignifikáns. Ezek a kritériumok azonban nem befolyásolják a modellek közti választást, csak utólagos ellenőrzésre szolgálnak.
Amennyiben ezek a feltételek egyik modellre sem teljesülnek, a program nem illeszt ARIMA- becslést, hanem az X11-ARIMA hagyományos aszimmetrikus filterét használja. [60

4. A kiigazítás minőségének ellenőrzése

M-statisztikák, Q mutató

AZ X11-ARIMA diagnosztikáit ez a program is tartalmazza. Ezek az eredmények stabilitását és a véletlen tényező szerepét értékelik.

Csúszó tartományok és változások követése elvén alapuló tesztek

Az X11-ARIMA tesztjei kiegészültek ezekkel az új diagnosztikákkal is. A csúszó tartományok (sliding spans) elve a mintavételi hiba becslésekor alkalmazható bootstrap elvnek, a változások követése (revision history) elv pedig a jackknife elvnek feleltethető meg.

A csúszó tartományok teszt is az eredmények stabilitását ellenőrzi, de úgy, hogy a szezonális kiigazítást évente csúsztatva végzi el és az azonos időszakokra vonatkozó értékeket (például a szezonális tényező értékét) hasonlítja egymáshoz. Multiplikatív modell esetén a megfigyelési tartományt mindig egy évvel csúsztatja. A program számolja minden időszakra a szezonális tényező minimális és maximális értékét, ha ezek között 0,03-nál, azaz 3 százalékpontnál nagyobb eltérés van, akkor az adott időszakra a szezonális tényezőt instabilnak tekinti. Ha ez az értékek több mint 15%-ra fennáll, az eredmények nem tekinthetők stabilnak.

A változások követésének elve az új adatok megjelenésének hatását vizsgálja és számszerűsíti. A változást általában láncindexszerűen, az előző időszakhoz viszonyítva számszerűsíti. Az új adat megjelenése az idősorban általában azzal jár, hogy a trend és a kiigazított értékek visszamenőleg is megváltoznak. Annak érdekében, hogy a stabilitást valamennyire biztosítani lehessen, az X12-ARIMA esetében is a modell rögzítését ajánlják. Ez konkrétan azt jelenti, hogy az előző év végéig vizsgálva az adatokat, számszerűsítik az outliereket, a munkanapok hatását, az ünnepeket, majd az idősor szezonális kiigazítása során megbecslik a következő év szezonális hatását. Az új adat bekerülésekor a regARIMA-modellt újrabecslik, de az előző év beállításait (ARIMA-modell, munkanaphatás) megtartják. Az így kapott tényezőket kiszűrik az idősorból, de a szezonális kiigazítást már a becsült szezonténye zőkkel végzik.

Szignifikáns szezonalitás az eredeti idősorban

Az eredeti idősor helyett a program a Hederson-féle mozgóátlagolású trendek leválasztása után hajtja végre a tesztet. A nullhipotézis szerint a trend leválasztása után kapott tényezők havi átlaga megegyezik egymással, azaz nincs az idősorban szezonalitás. Ha a véletlen tényező normális eloszlású fehérzaj, akkor a hipotézist (11, N-12) szabadságfokú F-próbával teszteli. Amennyiben a hullhipotézis elfogadható, nincs értelme a szezonális kisimításnak, sőt káros is lehet.
Ezzel a teszttel lehet azt is ellenőrizni, hogy maradt-e szezonalitás az idősorban a kisimítás után. Ilyenkor a tesztet a véletlen tényező végső becslésére kell elvégezni.

Spektrális elemzés

Szintén új diagnosztikai elem az X12-ARIMA programban a spektrális elemzés megjelenése, amely lehetővé teszi, hogy a ciklikus jellegű munkanaphatást és szezonalitást a frekvencia-tartományon történő elemzéssel is megvizsgáljuk. A program becsüli és kiírja a véletlen tényező és a szezonálisan kiigazított idősor spektrumát, valamint jelöli a naptári- és munkanap hosszakhoz tartotó frekvenciák helyeit. [56]

III. A TRAMO/SEATS és az X12-ARIMA módszertanok összehasonlítása

1. Az összehasonlítás során felhasznált idősorok

Az eredeti kísérleti elemzéseink során mintegy 176 idősort vizsgáltunk meg, illetve igazítottuk ki szezonálisan. Első lépésben a Demetra szoftver automatikus beállításokkal történő lefuttatását végeztük el minden idősorra, mind a Tramo/Seats, mind az X12-Arima opció kiválasztása mellett külön-külön. A Tramo/Seats a 176 futtatás során csak 22 alkalommal nem tudott adekvát modellt illeszteni a vizsgált idősorra a diagnosztikai eredmények kiértékelése után. Ez az esetek csupán 12,5 %-a.
Az X12-Arima ugyanakkor 60 esetben vetette el az általa illesztett modellt a kapott diagnosztikák alapján. Ez már valamivel több, mint 34 %-a az összes idősornak.

Az újabb alapadatokat, melyeken a szezonális kiigazító módszertanok összehasonlításának speciális kiegészítését elvégzem, a következő idősorok havi bontású értékei képezik:

- Ipari tevékenység belföldi árbevétele (TEÁOR alapján képzett 4 fő feletti ipari vállalkozások, 1000 Ft)
- Ipari tevékenység export árbevétele (TEÁOR alapján képzett 4 fő feletti ipari vállalkozások, 1000 Ft)
- Ipari tevékenység összes nettó árbevétele (TEÁOR alapján képzett 4 fő feletti ipari vállalkozások, 1000 Ft)
- Ipari tevékenység termelési értéke forgalmi adó nélkül, árbevételbe beszámító árkiegészítéssel (TEÁOR alapján képzett 4 fő feletti ipari vállalkozások, 1000 Ft)
- Exportértékesítés termelői árindexe, előző hó = 100,0 (százalék)
- Fogyasztóiár-index, 1990 év = 100,0 (százalék)

Az iparstatisztikai árbevétel-idősorok terjedelme: 1999.január – 2008. május;
Az iparstatisztikai termelési értéksor terjedelme: 1999.január – 2008. május;
Az exportértékesítés termelői árindex-sorok terjedelme: 1999.január – 2008. május;
Az fogyasztói árindex-sorok terjedelme: 1992.január – 2008. július;

Az iparstatisztikai idősorok kategóriák (TEÁOR) szerinti bontása:

C-E ipar összesen
C BÁNYÁSZAT
CB EGYÉB ÁSVÁNYBÁNYÁSZAT
DA ÉLELMISZER, ITAL, DOHÁNY GYÁRTÁSA
17 - 19 Textiliák, ruházati-, bőr- és szőrmetermékgyártás
20 - 22 Fa-, papír-, nyomdaipari termékgyárt., kiadói tev.
23 - 25 Vegyipar
29 - 35 Gépipar
36 - 37 Egyéb feldolgozóipar, hulladékvisszanyerés
D FELDOLGOZÓIPAR
DB TEXTÍLIA, TEXTILÁRU GYÁRTÁSA
DC BŐRTERMÉK, LÁBBELI GYÁRTÁSA
DD FAFELDOLGOZÁS
DE PAPÍRGYÁRTÁS, KIADÓI, NYOMDAI TEVÉKENYSÉG

DF KOKSZGYÁRT., KŐOLAJFELDOLGOZÁS, NUKLEÁRISFŰTŐANYAGGYÁRTÁS
DG VEGYI ANYAG, TERMÉK GYÁRTÁSA
DH GUMI-, MŰANYAG TERMÉK GYÁRTÁSA
DI EGYÉB NEMFÉM ÁSVÁNYI TERMÉK GYÁRTÁSA
DJ FÉMALAPANYAG, FÉMFELDOLGOZÁSI TERMÉK GYÁRTÁSA
DK GÉP, BERENDEZÉS GYÁRTÁSA
DL VILLAMOS GÉP, MŰSZER GYÁRTÁSA
DM JÁRMŰGYÁRTÁS
DN MÁSHOVA NEM SOROLT FELDOLGOZÓIPAR

1. táblázat

A fogyasztói árindex kategóriák (fogyasztási főcsoportok) szerinti bontása:

Mindösszesen - Kiadási főcsoportok
10–17 ÉLELMISZEREK ÖSSZESEN
18-19 SZESZES ITALOK, DOHÁNYÁRUK ÖSSZESEN
3 RUHÁZKODÁSI CIKKEK ÖSSZESEN
4 TARTÓS FOGYASZTÁSI CIKKEK ÖSSZESEN
50 HÁZTARTÁSI ENERGIA ÖSSZESEN
51-56 EGYÉB CIKKEK, ÜZEMANYAGOK ÖSSZESEN
6 SZOLGÁLTATÁSOK ÖSSZESEN

2. táblázat

Két különböző modell eredményeinek összehasonlítására a Demetra programcsomag definiál egy, a szezonális kiigazítás minőségét jelző indexet – *SA quality index* –, amely mint mérőszám fog alapjául szolgálni a módszertani összehasonlításhoz.

Az index kiszámítása során felhasználásra kerülnek az alábbi, szintén általánosan alkalmazott diagnosztikai statisztikák:

2. Kritériumok az automatikus kiigazítás minőségellenőrzéséhez

A Demetra automatikusan detektálja a nem kielégítő eredményeket (az alapértelmezett vagy felhasználó által definiált döntési szabályoknak megfelelően).

2.1. A lehetséges diagnosztikák a Tramo/Seats SA-metódushoz:

A TRAMO/SEATS tesztjei azt mérik, hogy a megbecsült és aztán felhasznált ARIMA-modell mennyire illeszkedik az idősorra. Azt a hipotézist teszteli, hogy a becsült reziduumok vajon tényleg normális fehérzajként viselkednek-e. Egyrészt azt vizsgálja meg, hogy van-e autokorreláció a reziduumokban. Ehhez ún. portmanteau típusú próbákat alkalmaz, a Ljung-Box- és a Box-Pierce-féle teszteket a reziduumokra. A reziduumok linearitásának teszteléséhez ugyanezen próbákat a reziduumok négyzetére alkalmazza. Másrészt a normalitást ellenőrzi a reziduumokra vonatkozó ferdeség és csúcsosság próbákkal és a kettő kombinálásával kapott normalitás teszttel.

Az említett próbák a következőképpen néznek ki:

Reziduumok autokorrelációján alapuló statisztikák

- Ljung-Box statisztika:

A Demetra teszteli, hogy a Ljung-Box statisztika kisebb-e mint $\chi^2_{m,\alpha}$. A $\chi^2_{m,\alpha}$ értéke az *m* és α-tól függ. Az *m* a szabadságfok (az idősorok periodicitás 2-szeresének és az ARIMA modell együtthatói számának különbsége), az α valószínűség pedig különböző értékek közül választható (10%, 5%, 2.5%, 2%, 1%, 0.5%, 0.2% ill. 0.1%), ahol az alapértelmezett érték 5%.

Például:
Ha α = 5%, az idősor pedig havi periodicitású (3.8) és az ARIMA-modell 1 paraméterrel identifikálható (m=24-1), akkor a Ljung-Box < 35,2 relációnak kell teljesülnie.

A reziduumok autokorrelációs struktúrájának ellenőrzése:
Már említettem, hogy többen is foglalkoztak a (3.11) szerinti portmanteau statisztika javításával, amely az első *m* késleltetés melletti (*m*-ed rendű) autokorreláltság szignifikanciáját teszteli a reziduumok esetében. A *m* értékének megválasztása azonban önkényes. Havi bontási idősorok esetén lehet például *m* = 24.
Fehérzaj-folyamatból származó adatokra az autokorrelációs együtthatók független normális eloszlásúak, a Q_m stasztika pedig χ^2_m eloszlású lesz.
Amikor tehát az autokorrelációk becslései legalább aszimptotikusan normális eloszlásúak 1/N varianciával, akkor alkalmazható a portmanteau-próba.
Amikor viszont az adatok az illesztett ARIMA modell reziduumaiként állnak elő, két korrekció válik szükségessé:
Először is Ljung és Box megmutatta (1978), hogy kis minták esetében az autokorrelációk varianciájának pontosabb közelítését szolgáltatja az (N-k)/N(N+2) kifejezés, mint az 1/N. A Q-statisztikát tehát ennek megfelelően korrigálva:

$$Q_m = N(N+2)\sum_{k=1}^{m} \hat{r}^2(k)/(N-k)$$

ahol *r* a reziduumok becsült autokorreláció függvénye. Ha a reziduumok egy ARMA(*p*,*q*) folyamat reziduumai, akkor a Q_m eloszlása aszimptotikusan $\chi^2_{m-(p+q)}$, *m* értékét az idősor megfigyeléseinek gyakorisága alapján választható meg, havi idősoroknál *m*=24, negyedéves idősoroknál *m*=16 a használatos.
Másrészről pedig az (*m*-*p*-*q*) szabadságfokú χ^2_{m-p-q} eloszlásfüggvény jobb közelítést ad a vizsgált autokorrelációs együtthatók véges dimenziós eloszlásaira.

A *Q*-statisztikát tovább módosították annak érdekében, hogy a speciális szezonális késleltetési értékeket is figyelembe vegyék. A 12 és 24 havi késleltetési értékhez tartozó autokorrelációs együtthatók szignifikanciája tesztelhető például az alábbi statisztika felhasználásával:

$$Q_s = N(N+2)\left[r^2(12)/(N-12) + r^2(24)/(N-24)\right].$$

Pierce (1978) megmutatta, hogy a Q_s robusztus közelítéssel χ^2_2 eloszlású.

- Box-Pierce statisztika:

A Demetra teszteli, hogy a Box-Pierce statisztika kisebb-e mint $\chi^2_{2,\alpha}$, melynek értéke csak α-tól függ. Az α valószínűség itt is különböző értékek közül választható (10%, 5%, 2.5%, 2%, 1%, 0.5%, 0.2% ill. 0.1%), és az alapértelmezett érték a 5%.

A szezonális autokorreláció vizsgálatára való a Box-Pierce-teszt[10]:

$$Q_s = N(N+2)\sum_{j=1}^{3}\frac{\hat{r}^2(js)}{N-js}$$

ahol *s* az évenkénti megfigyelések száma (azaz 12 a havi és 4 a negyedéves idősoroknál). A nullhipotézis melletti eloszlás közelítőleg χ^2_2.

Reziduumok függetlensége:

- Ljung-Box statisztika a négyzetes reziduálok autokorrelációja alapján (α alapért.: 5%).
- Box-Pierce statisztika a négyzetes reziduálok autokorrelációja alapján (α alapért.: 5%).

A reziduumok eloszlásának leírása (jellemzői):

Reziduumok Normalitása

A normalitás-vizsgálat itt azt teszteli, hogy a *normality* próbastatisztika kisebb-e mint $\chi^2_{2,\alpha}$.

Reziduumok asszimetriája

Ferdeség (Skewness) (a harmadik centrális momentum):

A Demetra teszteli, hogy a *Skewness* statisztika bele esik-e a $\left[-z_{\alpha/2}\sqrt{\frac{6}{N}},\ z_{\alpha/2}\sqrt{\frac{6}{N}}\right]$ intervallumba. A $\sqrt{\frac{6}{N}}$ a Tramo/Seats által számított standard hiba, ahol N az idősor hossza.

---

[10] A Box-Pierce-féle próbán gyakran nem a fenti szezonális autokorrelációt vizsgáló tesztet értik, hanem a Ljung-Box próbánál gyengébb autokorreláció tesztet, amelynek tesztstatisztikája:

$N\sum_{j=1}^{m}\hat{r}^2(j)$, és aszimptotikus eloszlása ugyanaz, mint a Ljung-Box- tesztnél.

A $z_{\alpha/2}$ csak az α-tól függ, ahol az α lehetséges értékei: 10%, 5%, 2.5%, 2%, 1%, 0.5%, 0.2% vagy 0.1%, és az alapértelmezett érték 5%.

Például:

Ha α=5% és az idősor hossza 80 megfigyelésből áll (N=80), akkor a |*Skewness*| < 0,537 $= 1{,}96\sqrt{\frac{6}{80}}$ relációnak kell teljesülnie.

Kurtosis (negyedik centrális momentum)

A Demetra teszteli, hogy a *Kurtosis* statisztika bele esik-e a $\left[-z_{\alpha/2}\sqrt{\frac{24}{N}}+3,\ z_{\alpha/2}\sqrt{\frac{24}{N}}+3\right]$ intervallumba. A $\sqrt{\frac{24}{N}}$ a Tramo/Seats által számított standard hiba, ahol N az idősor hossza.
A $z_{\alpha/2}$ itt is csak az α-tól függ, ahol az α lehetséges értékei: 10%, 5%, 2.5%, 2%, 1%, 0.5%, 0.2% vagy 0.1%, és az alapértelmezett érték 5%.

Outlier-ek száma

A program által detektált Outlier-ek száma nem haladhatja meg a megfigyelések számának egy bizonyos százalékát (alapértelmezés szerint 5%-át).

Megj.:

A bemutatott lineáris modellek képesek leírni az idősorok autokorrelációs viselkedését, amely a statisztikai adatsorok másodrendű momentumával kapcsolatos tulajdonság. A Wold dekompozíció által ezek a modellek egyszerűen úgy tekinthetők, mint a fehérzaj-változók lineáris kombinációja. Kérdés, hogy a fehérzaj-folyamat normális eloszlású vagy sem, mely meghatározza, hogy a reprezentáció mennyire képes megragadni az idősorok magasabb rendű momemtumait. Normális eloszlású fehérzaj-folyamat esetében nevezetesen

$$m_\alpha = T^{-1}\sum_{i=1}^{T}\varepsilon_t^\alpha \quad \alpha = 2,3,4,$$

α rendű momentum esetén:

$$\sqrt{T}m_\alpha \sim N\left(\mu_\alpha, \alpha!\sigma^{2\alpha}\right) \quad \alpha = 2,3,4,\ldots .$$

Lehetséges például ellenőrizni, hogy a becsült reziduumok eloszlása szimmetrikus-e , amint az a normális eloszlás esetében is van. Ez elvégezhető a ferdeségi teszt végrehajtása által, amely a harmadik momentum alapján számítható:

$$S = \sqrt{T}\,\frac{m_3}{\hat{\sigma}^3} \sim N(0,6).$$

A negyedik momentum alapján pedig kimutathatóak a reziduumok eloszlásának túlságosan nagy szélei: a kutosis-teszt az alábbi

$$K = \sqrt{T}\,\frac{m_4 - 3}{\hat{\sigma}^4} \sim N(0,24)$$

próbafüggvény szerint végezhető, amely szignifikáns lesz, amikor túl sok reziduum vesz fel nagy abszolút értéket. Így tehát felhasználható a reziduumokban lévő outlier-ek százalékának tesztelésére.

Végül az *S* és a *K* statisztika kombinálható a normalitás tesztelésére, melynek formája:

$$N = S^2 + K^2$$

és amennyiben az *S* és *K* becslések függetlenek, akkor az *N* statisztika χ_2^2 eloszlású.

2.2. A lehetséges diagnosztikák a X-12-Arima SA-metódushoz

Reziduumok autokorrelációján alapuló statisztikák

- Ljung-Box statisztika:
  Itt az α valószínűség alapértelmezett értéke 0.1%.

Reziduumok függetlensége:

- Ljung-Box statisztika a négyzetes reziduálok autokorrelációja alapján (α alapért.: 0.1%).

Előrejelzési hiba (Forecast Error):

A minta utolsó évére vonatkozó előrejelzés standard hibájának átlagos százalékos mértéke. A program ellenőrzi, hogy ez az érték kisebb legyen mint α, ahol az α lehetséges értékei: 20%, 15%, 10%, 5%; az alapértelmezett érték 15%. Ez a teszt használható az ARIMA-modell kiválasztására is: a legkisebb előrejelzési hibájú modell kerül kiválasztásra.

Outlier-ek száma

A program által detektált Outlier-ek száma nem haladhatja meg a megfigyelések számának egy bizonyos százalékát (alapértelmezés szerint 5%-át).

Ad-hoc minőség értékelő statisztikák:

Kombinált Q statisztika (kombinált M1 és M3 – M11).

Az M1-M11-gyel jelölt összegző statisztikák eredménye minden esetben 0 és 3 közötti szám. Az M1-M6 statisztikák a véletlen tényezőhöz, az M7-M11 statisztikák a szezonális tényezőhöz kapcsolódnak. A kiigazítás sikeresnek tekinthető, ha valamennyi M-statisztika értéke 0 és 1 közé esik. Általában, minél kisebb értéket kapunk valamely M-statisztikára, annál jobbnak tekinthető a szezonális kiigazítás.

Az M statisztikák a következők:

Ml - az eredeti idősor 3 havi változásából az irreguláris tényező részaránya;
M2 - a trend kiszűrésével nyert idősor szórásából az irreguláris komponens részaránya;

M3 - a véletlen tényező átlagos egyhavi változása összehasonlítva a trend átlagos egyhavi változásával;
M4 - az irreguláris tényező autokorrelációjának mértékét fejezi ki, vagyis azt; hogy felismerhető-e valamilyen minta az irreguláris tényezőben;
M5 - hány hónap alatt haladja meg a trend változása az irreguláris komponens változását;
M6 - a véletlen tényező egyévi változása a szezonális tényező egyévi változásához viszonyítva;
M7 - a változó és állandó szezonalitás aránya;
M8 - a szezonális tényező éves átlagos ingadozásának nagysága;
M9 - a szezonális tényező átlagos lineáris jellegű változása;
M10 - az M8 mutató, az idősor utolsó néhány évének adatára;
Ml l - az M9 mutató, az idősor utolsó néhány évének adatára;

A globális kombinált Q statisztika, mint értékelési kritérium alapján jelzett kiigazítási minőség, az M1-M11 mutatók súlyozott átlaga. A dekompozíció eredménye akkor tekinthető elfogadhatónak, ha értéke 1-nél kisebb.

3. SA Quality Index

Mindezek felhasználásával a következőképpen kerül kiszámításra a minőségi index (*SA quality index*):

$$QualityIndex = \frac{10}{M}\sum_j w_j \left| \frac{tesztStat_j - optErtek_j}{limitErtek_j - optErtek_j} \right|^{penalty}$$

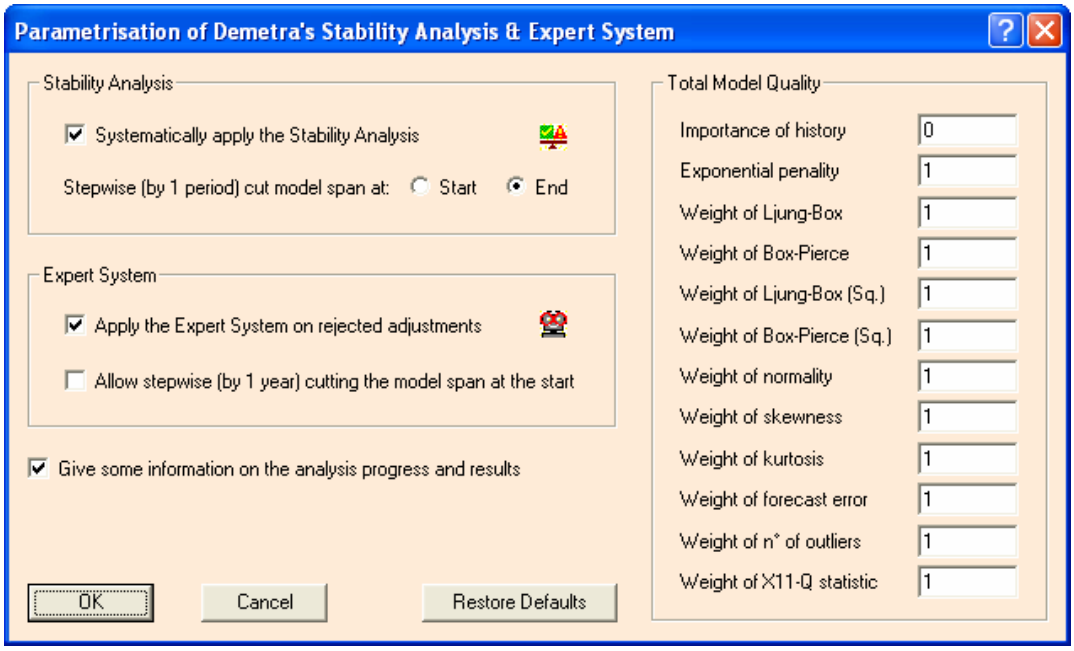
$limitErtek_j$: a konfidencia-intervallum határai az adott tesztstatisztika számára (a választható legnagyobb konfidencia szint mellett, pl. 0.1%);
$optErtek_j$: az adott tesztstatisztika optimális értéke;
$tesztStat_j$: a kiigazító metódus által szolgáltatott tesztstatisztika érték;
w_j : az egyes statisztikákhoz beállított súlyozó tényező (az alapértelmezett súlyérték 1, így alapértelmezésben nincs befolyásoló hatása az indexre), j= 1, … ,10;
penalty : a „büntető” exponens fokozza a hátrányos/előnyös tesztérték-hatást (az alapértelmezett értéke 1, így alapértelmezésben nincs befolyásoló hatása az indexre);
M : a rendelkezésre álló tesztstatisztikák száma.

A *QualityIndex* értékei 0-tól (az optimális érték) végtelenig (a legrosszabb érték) változhat. Ha az értéke nagyobb mint 10 (standard alap érték), akkor legalább egy tesztstatisztika szignifikáns kell legyen. Elfogadható kiigazítás *QualityIndex* - e kisebb mint 10.

A Demetra hozzáadott súlyozással előnyben részesítheti azokat a modelleket, amelyek gyakrabban fordulnak elő (pl. az Airline modell). Az ún. *Importance of history* súlyérték alapértelmezésben 0, mely esetben nincs hozzáadott érték engedélyezve a történeti modelleknek.

Lehetőség van a programban a különböző modell-jellemzők, illetve a diagnosztikai statisztikák, együttesen a teljes modell minőség – 4. ábra: TQM (*Total Model Quality*) –

súlyozására, amelyek befolyásolhatják a *quality index* értékét. A TQM-ben az alábbi súlyozott jellemzők és statisztikák szerepelnek (az ábrán az alapértelmezett súlyérték szerepelnek):



4. ábra

Az alapértelmezés szerinti kiigazítás Q*ualityIndex*-ének eredményei az ImportanceOfHistory_0 változó értékeit fogják képezni az összehasonlítás alapjául szolgáló osztályozási eljárás input adatai között.
Az egyes súlyok értékét változtatva a TQM-ben, a *quality index* értéke eltérő módon változik különböző idősorok esetében a választott módszertől függően. Rendre kétszeres súllyal figyelembe véve az egyes mutatókat a számított indexben az alapértelmezett súlyértékhez képest, a következő változókat lehet képezni:

ImportanceOfHistory_0: nincs hozzáadott súlyérték engedélyezve a gyakoribb előfordulású modelleknek a *QualityIndex*-ben
ImportanceOfHistory_1: van hozzáadott súlyérték engedélyezve a gyakoribb előfordulású modelleknek a *QualityIndex*-ben
ExponentialPenalty_2: a *penalty* kitevő négyzetesen kerül alkalmazásra a *QualityIndex*-ben

Az alábbi „statisztika_2” formájú változó megnevezések azt jelentik, hogy az adott tesztstatisztika 2-es súlytényezővel (duplán) kerül figyelembevételre a *QualityIndex*-ben, míg a többi csak 1-gyel (szimplán):

- Ljung-Box_2
- Box-Pierce_2
- Ljung-Box(Sq)_2
- Box-Pierce(Sq)_2
- Normality_2
- Skewness_2
- Kurtosis_2

- ForcastError_2
- NoOfOutliers_2
- X11-Qstatistic_2

CombinedWeights_1: csak a Ljung-Box ill. Kurtosis statisztikák, valamint az outlier-ek száma van figyelembe véve (amelyek mindkét módszertan diagnosztikái között szerepel);

CombinedWeights_2: mindegyik statisztika benne van, de közülük csak az előbbi három lesz egyszerre kiemelten súlyozva;

Tramo_indicator: bináris változó, amely azt fogja mutatni, hogy az adott kiigazítás melyik módszertant követve került végrehajtásra:
- Tramo/Seats : 1
- X12-arima : 0.

Cél: Az a változó, amelynek becsülni akarom az értékeit más változók alapján.

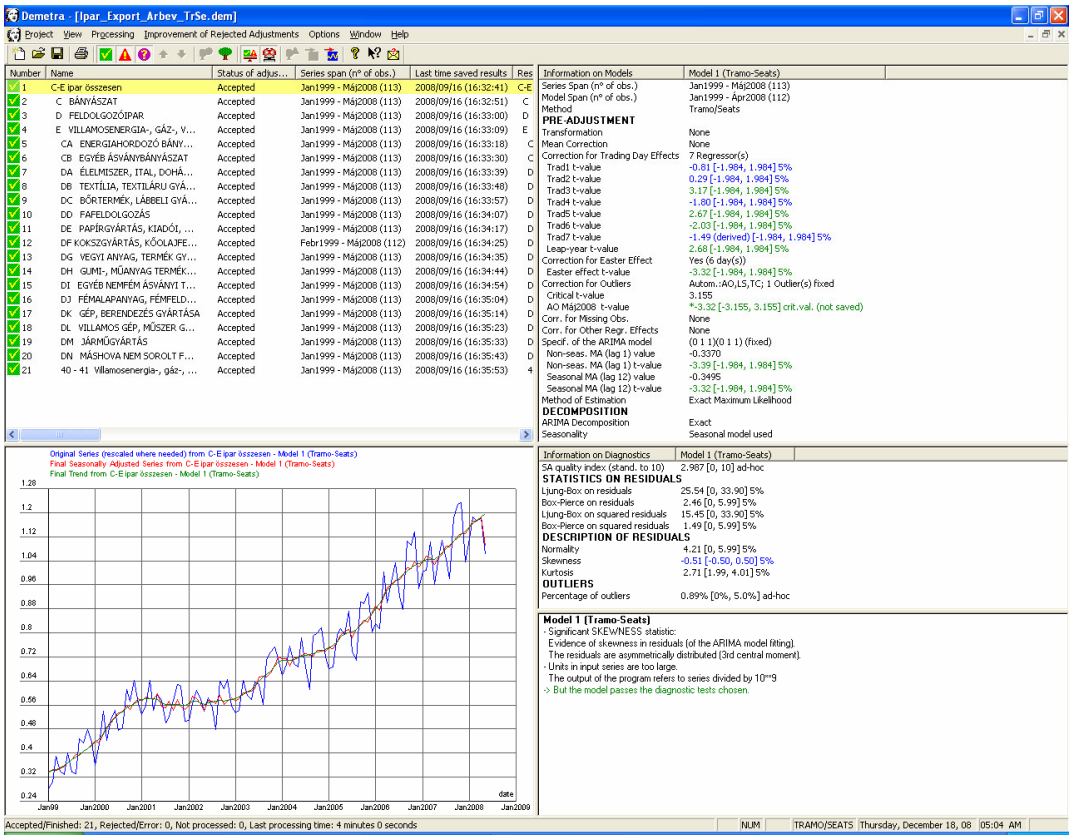
(Jelen esetben a Tramo_indicator)

Input: A célváltozó becslésére felhasznált (független magyarázó) változó.

(Jelen esetben a 5. ábra további 15 db változója)

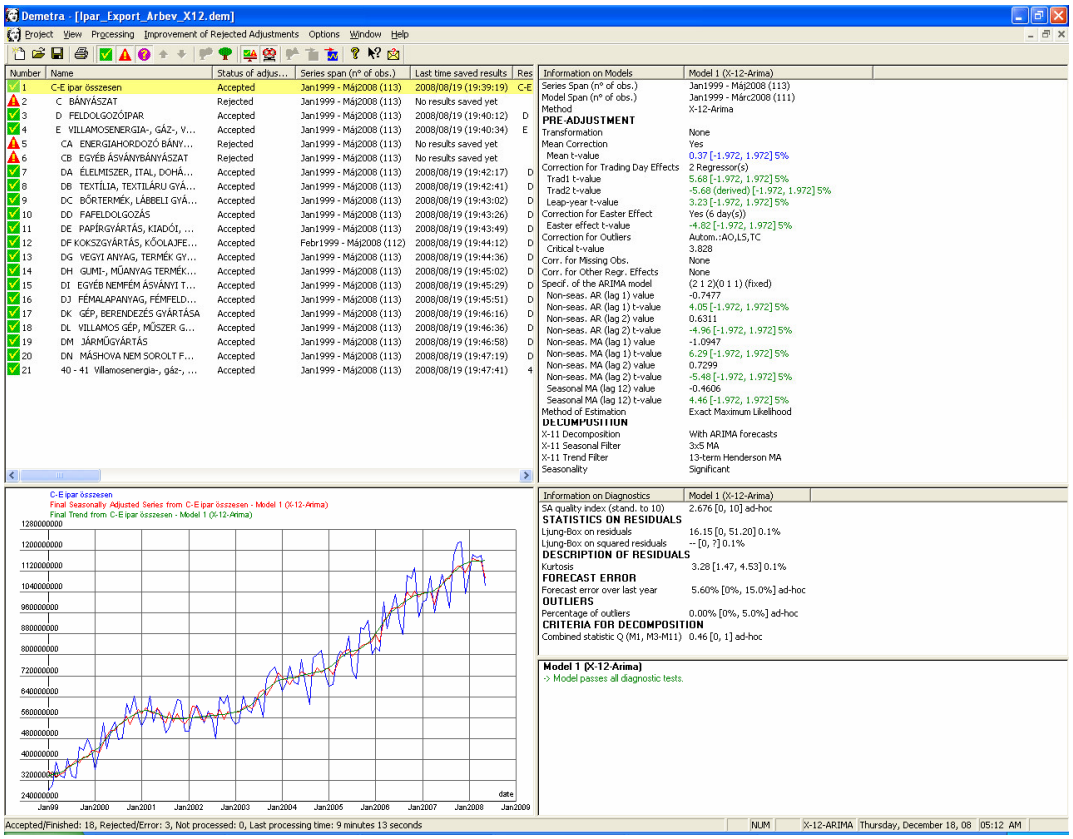
A Demetra programcsomag automatikus modulját lefuttatva az összes iparstatisztikai, illetve árindex idősorra mind a Tramo/Seats, mind az X-12-Arima módszerek kiválasztásával, illetve minden a felsorolt TQM beálltás mellett, a kapot *QualityIndex*-ek bekerülnek az input adathalmazba a megfelelő változóknak az 1. illetve 2. táblázat szerinti kategóriákba tartozó értékeiként.

Az ipari export árbevétel TRAMO/SEATS által kiigazított idősorai (példa):



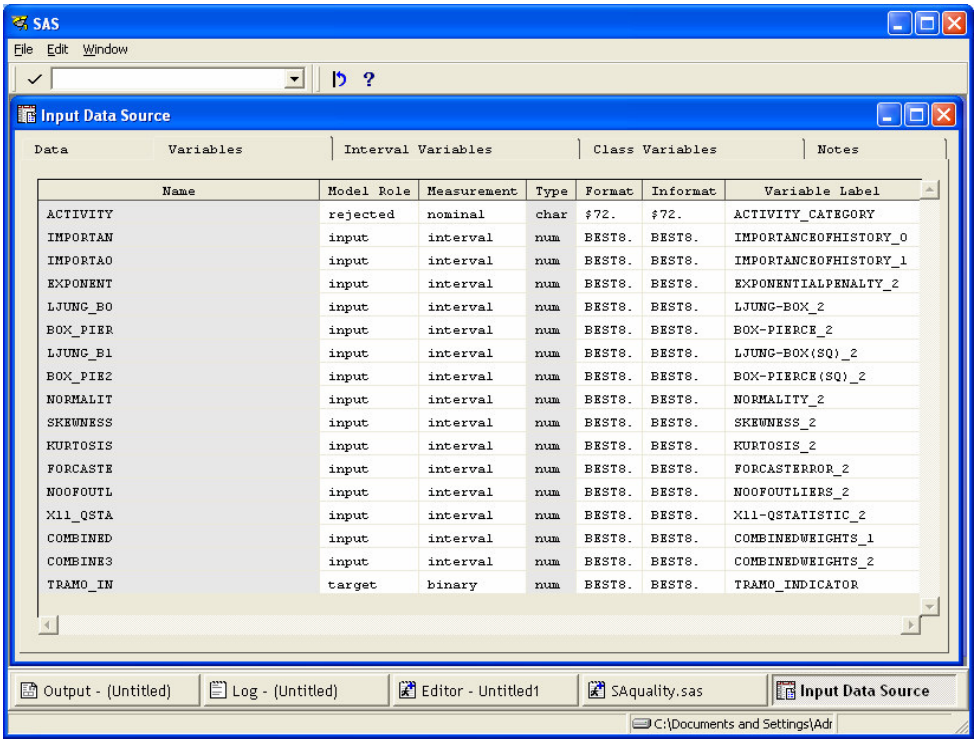
3. ábra

Az ipari export árbevétel TRAMO/SEATS által kiigazított idősorai:



4. ábra

A kiigazítások során kapott 230 db rekord pedig a SAS Enterprise Miner adatbányászati szoftveralkalmazás input adatait fogja képezni az ezután következő döntési folyamatban:



Name	Model Role	Measurement	Type	Format	Informat	Variable Label
ACTIVITY	rejected	nominal	char	$72.	$72.	ACTIVITY_CATEGORY
IMPORTAN	input	interval	num	BEST8.	BEST8.	IMPORTANCEOFHISTORY_0
IMPORTA0	input	interval	num	BEST8.	BEST8.	IMPORTANCEOFHISTORY_1
EXPONENT	input	interval	num	BEST8.	BEST8.	EXPONENTIALPENALTY_2
LJUNG_B0	input	interval	num	BEST8.	BEST8.	LJUNG-BOX_2
BOX_PIER	input	interval	num	BEST8.	BEST8.	BOX-PIERCE_2
LJUNG_B1	input	interval	num	BEST8.	BEST8.	LJUNG-BOX(SQ)_2
BOX_PIE2	input	interval	num	BEST8.	BEST8.	BOX-PIERCE(SQ)_2
NORMALIT	input	interval	num	BEST8.	BEST8.	NORMALITY_2
SKEWNESS	input	interval	num	BEST8.	BEST8.	SKEWNESS_2
KURTOSIS	input	interval	num	BEST8.	BEST8.	KURTOSIS_2
FORCASTE	input	interval	num	BEST8.	BEST8.	FORCASTERROR_2
NOOFOUTL	input	interval	num	BEST8.	BEST8.	NOOFOUTLIERS_2
X11_QSTA	input	interval	num	BEST8.	BEST8.	X11-QSTATISTIC_2
COMBINED	input	interval	num	BEST8.	BEST8.	COMBINEDWEIGHTS_1
COMBINE3	input	interval	num	BEST8.	BEST8.	COMBINEDWEIGHTS_2
TRAMO_IN	target	binary	num	BEST8.	BEST8.	TRAMO_INDICATOR

5. ábra

A döntés meghozatalához a kiigazítások eredményeinek, illetve a különböző tesztstatisztikák súlyozása mellett kapott *QualityIndex*-eknek az osztályozását fogom elvégezni döntési fa segítségével, amely egyike az eddigiekben legsikeresebbnek bizonyult tanulási algoritmusoknak. Először a cselekvő alrendszert, majd tanítását fogom megmutatni.

4. Döntési fák

Egy empirikus fa egyszerű szabályok sorozatának alkalmazásával reprezentálja az adatoknak egy szegmentálását, vagyis bonyolult összefüggéseket egyszerű döntések sorozatára vezet vissza. Minden szabály az input adatok valamely tulajdonságának értékére vonatkozó tesztnek felel meg, a csúcsból kilépő ágakat pedig a teszt lehetséges kimeneteleinek felelteti meg. Egyik szabályt a másik után alkalmazva, egymást tartalmazó szegmensek hierarchiája áll elő eredményül. Ezt a hierarchiát nevezik fának, és a szegmenseket csúcsnak. Az kezdő (originális) csúcs a teljes adathalmazt tartalmazza és a fa gyökerének, a végső csúcsokat pedig leveleknek nevezik. Minden levél egy döntést eredményez, amely alkalmazandó minden megfigyelésre a levélben.
A csúcsokban feltett kérdésekre adott válaszoknak megfelelően a fa gyökeréből kiindulva egy levélig lépkedünk lefelé a fában. A levél címkéje fogja meghatározni a döntést.
A döntési fák előnyös tulajdonsága, hogy a feltételeket összeolvasva – a gyökérből egy levélbe vezető út mentén – könnyen értelmezhető szabályokat ad a döntés meghozatalára, illetve jól érthető módon magyarázható, hogy a fa miért az adott döntést

hozta. A szabályok egyértelműek, hiszen az input adatokból tetszőleges rekord egyértelműen kerül besorolásra valamelyik levélbe, és az adott rekord csak ehhez a levélhez tartozó szabályra illeszkedik.[30]
A fa csúcsai felhasználhatók például olyan döntési fa felépítésére, amely többek között alkalmas a megfigyeléseknek egy bináris célváltozó értékein alapuló osztályozására. A jelen munka során konkrétan erre fogom alkalmazni.
A **Tree** tehát szabályoknak egy halmazát állítja elő, melyeknek SAS implementációja az intervallum típusú input adataimon alapuló többutas vágásokat[11] keres. Megválasztható a vágás kritériuma illetve annak módja, amely meghatározza a fa felépítésének metódusát.

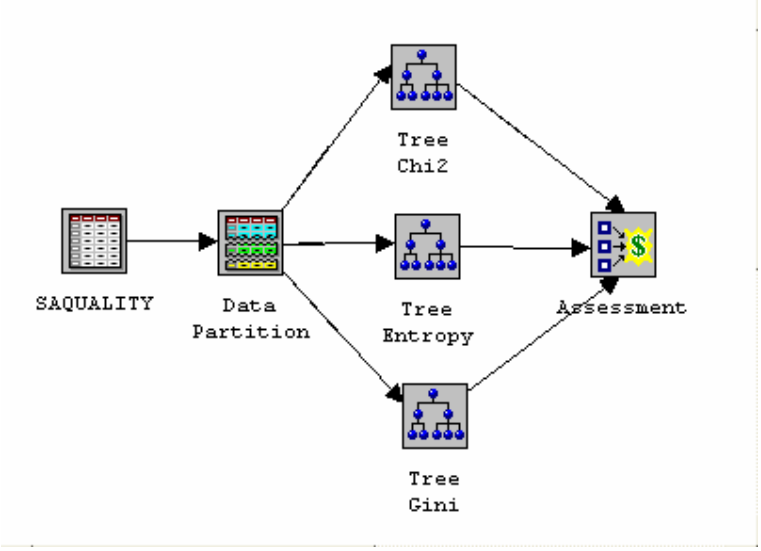
4.1. Döntési fa felépítése tanulással

A döntési fa bemenetként egy attribútumokkal leírt rekordot kap, és egy „döntést", a bemenetre adott válasz (a *Tramo_indicator*, mint osztályattribútum) jósolt értékét adja vissza eredményként. A bemeneti attribútumok jelen esetben folytonosak, a kimeneti érték pedig diszkrét, ennek megfelelően diszkrét értékkészletű függvény tanulásáról, vagyis osztályozásról (classification) van szó. Konkrétabban bináris (boolean) osztályozásról, mivel minden megfigyelés (példa) vagy igaznak, vagy hamisnak kerül besorolásra.
Ez a módszer egy tesztsorozat végrehajtása után jut el a döntéshez, ahol a fa belső csúcsai egy adott tulajdonság értékére vonatkozó tesztnek felelnek meg, a csúcsból kilépő ágakat a teszt lehetséges kimeneteivel címkézve. A levélcsúcsok megadják a levél elérésekor visszaadandó értéket, mivel minden levélhez hozzá kell rendelni a döntést (a célváltozó egy értékét), melyre használható az ún. többségi szavazás elve: amely osztályba a legtöbb tanítóminta tartozik, az lesz a döntés (belső csúcsokhoz is rendelhető ezen elv alapján döntés).

A jelen dolgozatban vizsgált probléma: eldönteni, hogy az iparstatisztikai idősorok esetében az adott TEAOR-ba tartozó ágazatokban, illetve a fogyasztói árindexnél az adott kiadási főcsoportokban kapott kiigazítások elfogadhatók-e a feltételezett módszertan alapján végrehajtott kiigazításnak. A cél: tanulással kialakítani a *Tramo_indicator* cél-predikátum (osztályattribútum) definícióját annak eldöntésére, hogy az adott kiigazítás melyik módszertan alkalmazásával történt.

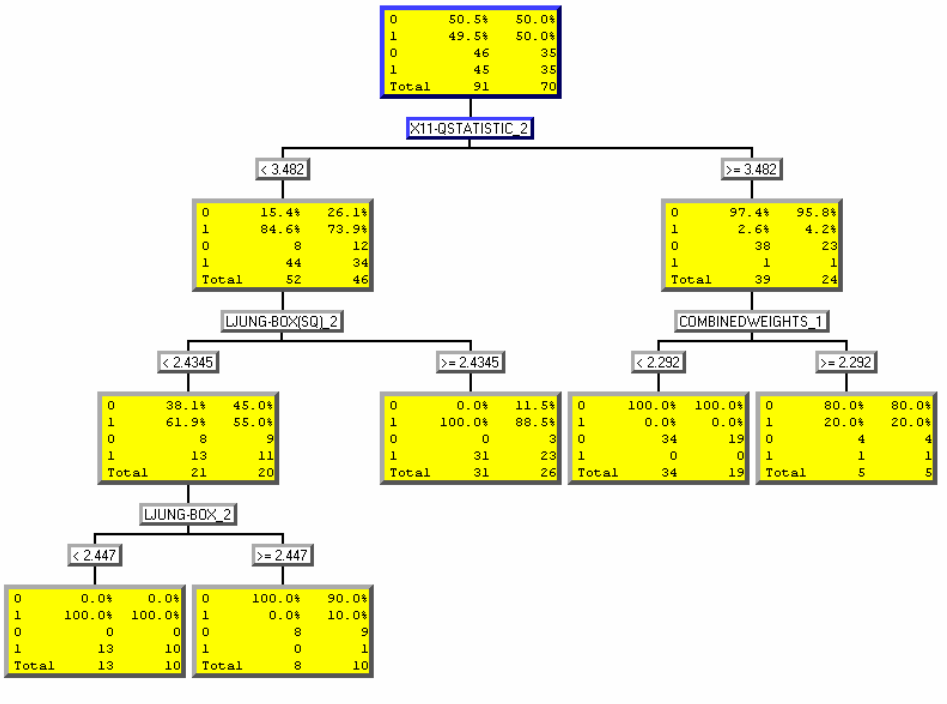
[11] Vágás: magyarázó változóknak bizonyos elv szerinti kettévágása egy változó (dimenzió) mentén, úgy hogy a levelek minél homogénebbek legyenek a célváltozó szerint.

A döntés meghozatalát támogató folyamat SAS EM (*Enterprise Miner*) ***diagramja:***



6. ábra

A SAS EM által az adott feladathoz felépített döntési fa az alábbi ábrán látható:



7. ábra.

A példákat a döntési fa a gyökérnél kezdi feldolgozni, követi a megfelelő ágakat, amíg egy levélhez el nem ér.

A döntési szabálysorozat három szabályból áll:

I. X11_QSTA<3,482 ÉS LJUNG_B1<2,4345 ÉS LJUNG_B0<2,447 → TRAMO_IN=1
II. X11_QSTA<3,482 ÉS LJUNG_B1≥2,4345 → TRAMO_IN=1
III. X11_QSTA<3,482 ÉS COMBINED≥2,292 → TRAMO_IN=1

Például az X11-Qstatistic_2<3,482 és Ljung-Box(Sq)_2<2,4345 és Ljung-Box_2<2,447 attribútumokkal jellemezhető rekord a Tramo/Seats osztályba sorolást fogja eredményezni, vagyis a Tramo_indicator = 1 lesz.

Logikai felírást használva a döntési fa a *Tramo_indicator* célpredikátum egy hipotézise, amely a következő formában felírt állításnak felel meg:

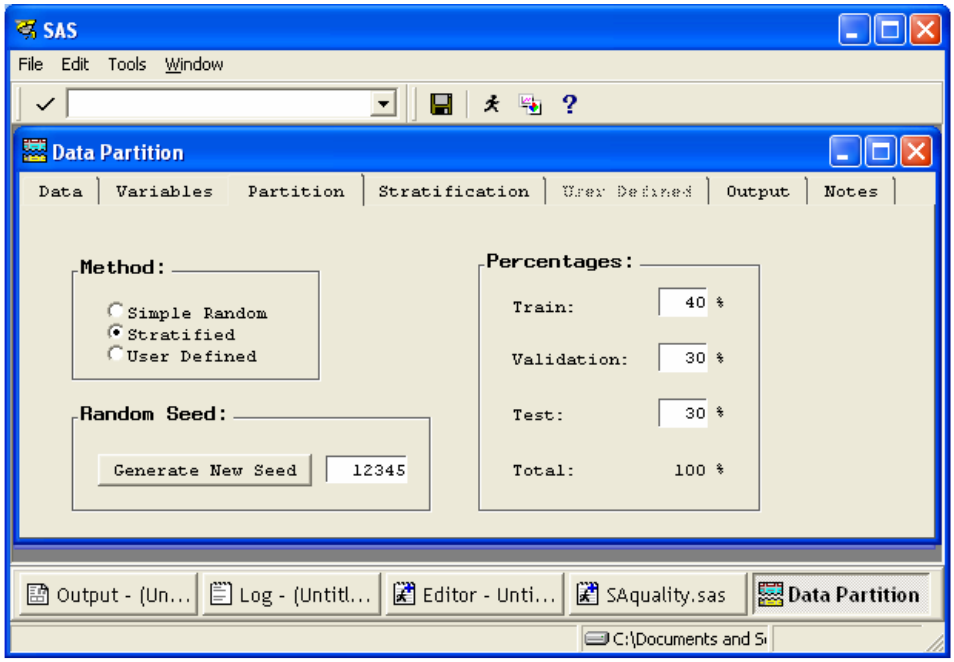
$$\begin{aligned}\text{TRAMO\_IN} \Leftrightarrow\ & (\text{X11\_QSTA}<3{,}482 \wedge \text{LJUNG\_B1}<2{,}4345 \wedge \text{LJUNG\_B0}<2{,}447) \vee \\ & (\text{X11\_QSTA}<3{,}482 \wedge \text{LJUNG\_B1}\geq 2{,}4345) \vee \\ & (\text{X11\_QSTA}<3{,}482 \wedge \text{COMBINED}\geq 2{,}292)\ ,\end{aligned}$$

ahol mindegyik feltétel azon tesztek konjunkciójának felel meg, amelyeket a gyökértől egy logikai igaz kimenetet jelentő levélig megtett út során végzünk. A döntési fákról általánosságban is elmondható, hogy teljes kifejezőképességgel bírnak az ítéletlogikai nyelvek területén, mert döntési faként tetszőleges logikai (Boole) függvény felírható. Ez triviálisan is megvalósítható a függvény igazságtáblájának minden sorát megfeleltetve a döntési fa egy útjának. Ez viszont, a vizsgált probléma minden változóját figyelembe véve, nagyon nagy döntési fához vezetne (a bevont változók számával exponenciálisan növekvő fához, mivel az igazságtábla is exponenciálisan növekvő számú sort tartalmazna). Azonban, mint a kiigazító eljárások eredményeinek összehasonlításának példája is illusztrálja, az adott esetben a vizsgált függvényt jóval kisebb fával is lehetséges reprezentálni.

Megj.:
A döntési fa exponenciálisan növekvő mérete bizonyos függvényfajtáknál valóban problémát jelent. Például a paritásfüggvénynél, amely akkor és csak akkor ad 1-et, ha páros számú bemenet 1 értékű, vagy a többségfüggvénynél, amely akkor ad 1-et, ha bemeneteinek több mint fele 1 értékű. A döntési fák tehát bizonyos függvények esetén jók, mások esetén nem. Szükség van ezért néhány megfelelő algoritmusra, hogy adott nagy térben is konzisztens hipotézist lehessen találni. [29]

A döntési fák felépítése példák alapján

A logikai döntési fa által kezelt példa a bemeneti attribútumok vektorából és egyetlen logikai kimeneti értékből áll. Azok nevezhetők pozitív példáknak amelyekben a *Tramo_indicator* értéke igaz, és azok a negatív példák, amelyekben az értéke hamis. A példák teljes halmaza a tanító halmaz, ami a SAS rétegzett (Stratified) partícionáló metódusa esetén alapértelmezésben az input adatok 40 %-át teszi ki**.**



8. ábra

Training -- az adatok 40% -a: A tanító adathalmaz, amely a kezdeti fa illesztésére használandó. Tartalmazza a fa tanítására szolgáló input és célváltozó értékeket.

Validation -- az adatok 30% -a: A validáló adathalmaz alapvetően a fa kiértékelésére szolgál, mert a döntési fák hajlamosak a túltanításra, így a program a validáló adathalmazt használja egy egyszerűbb illesztéshez való visszatérésre (amennyiben szükséges), mint amilyet kizárólag a tanító adathalmazra alapozott illesztés eredményez. Tartalmazza az input és a célváltozó azon értékeit, amelyek közvetett módon kerülnek felhasználásra a tanítás során.

Test -- az adatok 30% -a: További „fenntartott" adathalmaz a fa kiértékelésére. Tartalmazza az input és a célváltozó azon értékeit, amelyek nem kerülnek felhasználásra a tanítás során, hanem a modell hibájának, általánosítási képességének egy végső, torzítatlan becslésre használandó.

Megj.: Az alapértelmezett véletlen magot használom a particionált adathalmaz előállítására**.**

A tanító halmaznak megfelelő döntési fa megtalálásának van egy triviális megoldása, amely egy olyan fa megkonstruálását jelenti, amelyben minden példához egy külön saját út tartozik egy - a példához tartozó - levélcsúcshoz. A levélhez vezető út mentén sorra tesztelve az attribútumokat, és a megfigyeléshez tartozó tesztértéket követve a levél a megfigyelés besorolását adja. Ez azonban nem igen ad információt egyetlen más esetről sem, egyszerűen memorizálja a példákat, nem nyer ki semmilyen mintázatot a példákból. Az Ockham borotvája12 elvét alkalmazva a legkisebb döntési fa megtalálása a cél, amely konzisztens a példákkal. Némi egyszerű heurisztika bevetésével jó eredmény érhető el egy „kisebb" fa megtalálásában.

[12] Az ún. Okham borotvája elv: azt legegyszerűbb hipotézist részesítsük előnyben, amely konzisztens az adatokkal, mivel azok a hipotézisek, amelyek nem egyszerűbbek, mint maguk az adatok, nem nyernek ki semmilyen mintázatot az adatokból.

A Döntési-Fa-Tanulás algoritmus alapötlete, hogy először a legfontosabb attribútum kerül tesztelésre, ahol a „legfontosabb" attribútum alatt az értendő, amely a példák besorolásában a legnagyobb változást okozza.
A döntési fa tanulás során az attribútumok kiválasztására szolgáló eljárás arra irányul, hogy minimalizáljuk az eredményül kapott fa mélységét. Azt az attribútumot választjuk, amellyel a lehető legmesszebbre jutunk a példák pontos osztályozásában. Egy tökéletes attribútum a példákat egy csupa pozitív és egy csupa negatív példát tartalmazó halmazra osztja. Az X11_QSTA nem tökéletes attribútum, de meglehetősen jó. Egy valójában haszontalan attribútum tesztjének eredményeként a kapott halmazokban nagyjából ugyanolyan arányban lesz pozitív és negatív példa, mint az eredeti halmazban.
Mindössze arra van szükség, hogy formális mértéket találjunk arra, hogy mit jelent a „meglehetősen jó" és a „valójában haszontalan". A mérték akkor érje el maximumát, amikor az attribútum tökéletes, és akkor legyen minimális, amikor az attribútumnak egyáltalán nincs semmi haszna. Egy megfelelő mérték az attribútum által szolgáltatott információ (information) várható értéke, ahol az információt abban a matematikai értelmezésben használjuk, ahogy először Shannon és Weaver definiálta (Shannon és Weaver, 1949). Az, hogy a válaszban mennyi információ rejlik, az az előzetes a priori tudástól függ. Minél kevesebbet tudunk, annál több a szolgáltatott információ. Az információelmélet bitekben méri az információtartalmat. Egy bit információ ahhoz elég, hogy egy olyan kérdésre, amelyről semmilyen előzetes elképzelésünk sem volt igen/nem választ megadjunk. [29]
Döntési fa tanulás esetén a megválaszolásra váró kérdés az, hogy egy adott példának mi a helyes besorolása. Egy jó döntési fa választ ad erre a kérdésre.

A tanító adatbázisból rekurzívan kerül felépítésre a fa. A teljes tanító adatbázisból kiindulva egy olyan kérdést kell keresni, amelyet megválaszolva a teljes tanulóhalmaz jól szétvágható. Egy szétvágást akkor tekinthető jónak, ha a célváltozó eloszlása a keletkezett részekben kevésbé szórt, mint a szétvágás előtt. Ezután rekurzívan kell alkalmazni a fenti eljárást a részekre. A rekurzió akkor áll meg valamelyik ágban, ha az alábbi feltételek közül valamelyik teljesül:

- a levél minden eleme ugyanabba az osztályba tartozik;
- nincs több attribútum, ami alapján az elemeket tovább lehetne osztani (a levélhez tartozó osztály ekkor az lesz, amelyikhez a legtöbb tanítópont tartozik);
- az adott levélhez nem tartozik tanítópont;
- a fa mélysége elérte az előre beállított korlátot;
- nincs olyan vágás, amely javítani tudná az aktuális osztályozást. [30]

Még mielőtt egyetlen attribútum is tesztelésre kerülne, a válaszok valószínűsége becsülhető a tanító halmazban található pozitív és negatív minták arányával. Tegyük fel, hogy a tanító halmazban **p** pozitív és **n** negatív példa található. Ez esetben a program tehát a vizsgált példában indulásképpen a rekordokat pozitív és negatív példahalmazokba sorolja a TRAMO_IN bináris célváltozó logikai értékei alapján. Ezek után megkeresi, hogy melyik attribútumot teszteli először a fában. Sorban elveti az olyan attribútumokat, amelyek tesztjének a különböző kimenetelei esetén közel ugyannyi pozitív (igaz) és negatív (hamis) példa kerül eredményül a kapott halmazokba. A program szerint végül az X11_QSTA változó egy meglehetősen fontos attribútumnak bizonyult. Miután az első attribútum (X11_QSTA) tesztje csoportokra bontja a példákat, mindegyik teszteredmény egy újabb döntési fa tanulási problémát eredményez, kevesebb példával és eggyel kevesebb attribútummal:

1. Ha van mind pozitív mind negatív példa is, akkor az algoritmus a legjobb attribútumot választja a szétosztásukra. A 7. ábrán látható, hogy a Ljung-Box(SQ)_2 attribútum alkalmas a megmaradó példák osztályozására.
2. Ha valamennyi megmaradt példa pozitív, akkor a válasz Igen, ha negatív, akkor Nem.
3. Ha nem marad példa a teszt egyik kimenetele esetén, akkor nincs megfigyelve ilyen esetet, és a válasz a szülőcsúcsban többségben levő példák alapján kerül meghatározásra.
4. Ha nem maradt attribútum további tesztelésre, de mind pozitív, mind negatív példák is maradtak, akkor ezeknek a példáknak pontosan azonos jellemzőik vannak, de különböző osztályokba tartoznak. Ez a helyzet állt elő a vizsgált feladatban a III. X11_QSTA<3,482 ÉS COMBINED≥2,292 szabály alkalmazása után. Ilyen esetben például zajosak (noise) lehetnek az adatok, vagy az attribútumok nem adnak elég információt a szituáció teljes leírására, illetve a problématér valójában nemdeterminisztikus. Ilyenkor megoldás lehet a problémára a többségi szavazás.

DÖNTÉSI-FA-TANULÁS algoritmus (lásd [29]):

function DöNTÉSi-FA-TANULÁs(*példák*, *attribútumok, alapérték)* **returns** egy döntési fa
inputs: példák, a példák halmaza
attribútumok, **az attribútumok halmaza**
alapérték, **a célpredikátum alapértéke**

if *példák* üres **then return** *alapérték*
else if *példák* minden elemének azonos a besorolása **then return** a besorolás
else if *attribútumok* üres halmaz **then return TöBBSÉGl-ÉRTÉK(***példák***)**
else
    legjobb <— **ATTRIBÚTUM-VÁLASZTÁS** *(attribútumok, példák)*
    fa<— **egy új döntési fa, a gyökér a** *legjobb* **attribútum tesztje**
    m <— **TöBBSÉGl-ÉRTÉK** (*példák*)
    for each *legjobb* minden v_i- értékére **do**
        példákj <— {a *példák* azon elemei, amelyekre *legjobb* = v_i}
        részfa <— **DöNTÉSi-FA-TANULÁS(***példák*$_{i}$-, *attribútumok-legjobb, m)*
        a *fa* döntési fához adjunk egy *v,* címkéjű ágat és a *részfa* részfát
return *fa*

Ily módon remélhetőleg kisszámú teszttel helyes osztályozás kapható, vagyis a fában minden út rövid lesz, így fa mérete lecsökken.
A döntési fák nagy előnye, hogy a lényegtelen változókat automatikusan felismerik. Ha nem nyerhető információ egy változóból a célváltozóra vonatkozóan, akkor azt nem is tesztelik. Ez azért előnyös, mert így zaj jelenlétében sem romlik a fák teljesítménye. Továbbá nagyban segíti a problémamegértésünket is, mert megtudjuk, hogy mely változók fontosak, és a legfontosabb változókat a fa a gyökér közelében teszteli. Előny még az is, hogy a döntési fák hatékonyan építhetők fel nagyméretű adathalmazokra is. Lényeges tulajdonság továbbá az is, hogy egy csúcsnak mennyi gyermeke lehet, de minden olyan fa, amely csúcsainak kettőnél több gyermeke is lehet, mindig átrajzolható bináris fává. Ennél fogva a legtöbb algoritmus csak bináris fát tud állít elő. [29]

4.2. Döntési fákat felépítő algoritmusok

A döntési fák felépítésére az alábbi algoritmus családok ismertek:

I. ID3 – Interactive Dichotomizer 3 (Interaktív tagoló / felosztó) család, melynek aktuális változata a C5.0 . A tesztattribútum kiválasztásához az entrópia csökkenését alkalmazza.
II. CART – Classification and Regression Trees (Osztályozó és regressziós fák)
III. CHAID – Chi-squared Automatic Interaction Detection (Khi-négyzet alapú automatikus interakció detektálás)

Az ID3 –az egyik legismertebb osztályzó algoritmus– a tesztattribútum kiválasztásához az entrópia csökkenését alkalmazza:
Az entrópia az osztályattribútum információtartalmának mértékét, vagyis a célváltozó értékével kapcsolatos bizonytalanságot fejezi ki. A két lehetséges értéket p_1 és p_2 valószínűséggel felvevő *Tramo_indicator*, mint valószínűségi célváltozó Shannon-féle entrópiája (információ-tartalma):

$$H\,(TRAMO\_IN) = H\,(p_1, p_2) = -\sum_{i=1}^{2} p_i \log_2 p_i$$

Az entrópia a célváltozó értékével kapcsolatos bizonytalanságot fejezi ki. Jelölje valamely magyarázó változót (a 15-ből) X, és megfigyelési értékét x_i, ekkor a célváltozóval kapcsolatos bizonytalanság:

$$H\left(TRAMO\_IN \mid X = x_i\right) =$$
$$-\sum_{j=0}^{1} P\left(TRAMO\_IN = j \mid X = x_i\right) \log_2 P\left(TRAMO\_IN = j \mid X = x_i\right)$$

Így X megfigyelésével a várható bizonytalanság:

$$H\left(TRAMO\_IN \mid X\right) = \sum_i P\left(X = x_i\right) H\left(TRAMO\_IN \mid X = x_i\right)$$

Eszerint X megfigyelésének lehetősége a bizonytalanság csökkenését eredményezi:

$$I\left(TRAMO\_IN \mid X\right) = H\left(TRAMO\_IN\right) - H\left(TRAMO\_IN \mid X\right),$$

melynek értéke lesz az entrópia változás a vágás révén, vagyis X ennyi információt hordoz a célváltozóról. Az ID3 a $TRAMO\_IN$ attribútum szerinti osztályozáskor olyan X attribútumon szerint ágazik szét, amelyre $I(TRAMO\_IN\,, X)$ maximális, ill. $H\,(TRAMO\_IN\,|X)$ minimális.
Az ID3 algoritmus tehát a minimális feltételes entrópiával rendelkező attribútumot választja ki, és nem feltétlenül bináris fát hoz létre (amennyiben a cél a bináris fa felépítése, akkor a magyarázó X attribútum típusától függően kétféle feltételt kell létrehozni). Az entrópia változás a vágás révén ugyanis azoknál az attribútumoknál nagy, amelyek sok értéket vesznek fel, melynek következtében sokfelé ágazik a fa, ami terebélyes fákat eredményez. Ha például

az attribútumok között szerepelne az azonosító kód, akkor az 0 feltételes entrópiát produkálna, így az algoritmus azt választaná.
Megoldás lehet erre a problémára a nyereségarány mutató (gain ratio) használata, mint normált feltételes információ. Ez a mutató a gyerek csúcsokba kerülő tanítópontok számát figyelembe véve "bünteti" azokat az attribútumokat, amelyek túl sok gyereket hoznak létre. [30]
A nyereségarány úgy kapható, hogy a vágás révén előállt entrópia változást elosztjuk az adott attribútum entrópiájával :

$$gain\_ratio = \frac{I(TRAMO\_IN, X)}{H(X)}$$

A vágás alapja tehát a nyereségarány, de sajnos a nyereségarány sok esetben "túlkompenzál" és olyan attribútumokat részesít előnyben, amelynek az entrópiája kicsi. Általános megoldás erre a problémára, hogy azon attribútumok közül választják ki a legnagyobb nyereségarányt adó attribútumot, amelyekhez tartozó feltételes információ legalább akkora, mint az összes vizsgált attribútumhoz tartozó feltételes információk átlaga. [30]

Az ID3 családba tartozó fák kizárólag osztályozásra, a CHAID és a CART osztályozásra és előrejelzésre is alkalmazható. A C4.5 (ill. kereskedelmi, javított változata a C5.0) és a CHAID fák a döntésekhez csak egyetlen attribútumra vonatkozó egyenlő, kisebb, nagyobb relációk tesztelését alkalmazzák a csúcsokban (egyváltozós fák), vagyis a jellemzők terét téglatestekre vágják fel. A CART fák ferdén is tudnak vágni, attribútumok lineáris kombinációját is tesztelik.
A CART eljárás csak bináris döntéseket használ, egy C4.5 fa viszont annyi felé ágazik egy nominális attribútumra, ahány lehetséges értéket felvehet az attribútum.

A leglényegesebb különbség a különböző fák között, hogy mit tekintenek jó vágásnak, döntésnek. Nominális, ill. bináris célváltozó esetén a CHAID eljárás – nevének megfelelően – a χ^2-tesztet használja. A CART metodológia a Gini-indexet minimalizálja. A Gini-index alapján mindig olyan attribútumot keres, amely alapján a legnagyobb homogén osztályt tudja leválasztani.
Ha a magyarázandó célváltozó intervallum skálán mért, akkor a CART eljárás a célváltozó varianciájának csökkentésére törekszik, a CHAID pedig F-tesztet használ.
A CHAID ún. konzervatív eljárás, csak addig növeli a fát, amíg egy előre adott küszöböt meghalad a csúcsban alkalmazható legjobb szétvágás χ^2-, vagy F-teszt szerinti szignifikanciája.
A CART és C4.5 eljárások vagy felépíthetnek nagyméretű fát, amelyik a tanuló adatbázison tökéletesen működik, vagy heurisztikus leállási szabályokat alkalmaznak a fa mélységére vonatkozóan: a fa egy előre adott korlátnál egyszerűen nem lehet mélyebb, illetve egy csúcsot már nem enged szétvágni, ha kevesebb eset tartozik bele egy adott korlátnál. Mindenesetre a kialakuló fa nagy és terebélyes lesz, túl speciális, amely nem csak az alappopuláció jellemzőit, hanem a mintában előforduló véletlen sajátosságokat is modellezi. Ezért a fát meg szokták metszeni (pruning) a felépítés után egy ellenőrző adatbázist használva, elhagyva ily módon a felesleges döntéseket.
Tanácsos megvizsgálni, hogy a generált C5.0 vagy CHAID fa nem tesztel-e egymás után ismételten kevés számú (2-3) attribútum értékét. Ez arra utalhat, hogy az attribútumok valamely függvénye bír magyarázó erővel és a fa az ismételt vágásokkal ezt a kapcsolatot próbálja közelíteni.

Többféle mutatószám létezik tehát a vágási kritérium kiválasztására, melyek között nem létezik legjobb, mert mindegyikhez lehet készíteni olyan adatbázist, amelyet rosszul osztályoz az adott vágási kritériumot használó algoritmus. [30]

4.3. Vágási kritériumok

A SAS-beli **Tree** támogatja mind az automatikus, mind az interaktív tanulást. Automatikus módban futtatva automatikusan rangsorolja az input változókat az alapján, hogy mennyire meghatározó a szerepük a fa megkonstruálásában. Ez a rangsorolás használható lehet a változók alkalmas kiválogatására. Az automatikus lépések felülírhatók a vágási szabályok interaktív definiálásával, illetve csúcsok vagy részfák explicit eltávolításával. Egy vágási szabály kiértékelésére vonatkozó kritérium vagy egy statisztikai szignifikancia vizsgálaton, nevezetesen egy F-próbán vagy egy χ^2-próbán, vagy a variancia, entrópia, illetve a Gini-index mértékének csökkenésén alapul. Mindegyik kritérium lehetővé teszi részfák egy szekvenciájának a felépítését.

A választható vágási-kritériumok a célváltozó típusától függnek[13]. Mivel a jelen munkában bináris célváltozóval foglalkozom, ezért a vágási kritériumokat csak a bináris[14] típusú célváltozók esetére fogom tárgyalni.
Bináris célváltozó esetén SAS EM-ben a fenti három algoritmusnak megfelelő három vágási-kritérium használható:

1. χ^2-próba (alapértelmezett) – a célváltozó / az elágazó csúcs Pearson-féle χ^2 mértéke, az alapértelmezett 0.20 szignafikancia szint mellet. A CHAID eljárás csak addig növeli a fa méretét, amíg a csúcsban alkalmazható legjobb vágás χ^2-próba szerinti szignifikanciája meghaladja az előre megadott 0.20 küszöbértéket.
   Kontingencia táblája:

Y \ V	Bal	jobb
jó	n_{11}	n_{12}
rossz	n_{21}	n_{22}

   Y és V függetlenségének erőssége:

$$\chi^2 = \sum_{i,j=1}^{2} \frac{\left(n_{ij} - \frac{n_{i+}\ n_{+j}}{n} \right)^2}{\frac{n_{i+}\ n_{+j}}{n}}$$

$$\chi^2 = n \frac{\left(n_{11}\ n_{22} - n_{12}\ n_{21} \right)^2}{n_{1+}\ n_{2+}\ n_{+1}\ n_{+2}}$$

   Minél szorosabb legyen a kapcsolat a V vágás és a Tramo_indicator célváltozó között, ill. a kettő függőségének erőssége.

2. Entrópia redukció – a levél-szennyezettség entrópia szerinti mértékének csökkentése.

   Az *s* levél szennyezettsége:

$$I(s) = I(p, 1\text{-}p) = - p \log_2 p - (1\text{-}p) \log_2 (1\text{-}p),$$

   ahol a p, $1\text{-}p$ a bináris célváltozó eloszlása az s levélnél.

13 Az F-próba ill. a variancia-redukció intervallum típusú célváltozó esetén használható
14 Illetve a nominális típusú célváltozók esetén is az itt tárgyalt vágások alkalmazhatók

3. Gini redukció – a levél-szennyezettség Gini-index szerinti mértékének csökkentése (CART).

Az s levél szennyezettsége:

$$I(s) = I(p, 1\text{-}p) = 1 - p^2 - (1\text{-}p)^2 = 2p\,(1\text{-}p),$$

ahol a p, $1\text{-}p$ a bináris célváltozó eloszlása az s levélnél.

Az utóbbi két módszert alkalmazó algoritmus esetében a T fa hibája:

$$E(T) = \sum_{s \hat{\in} T} p(s)\, I(s),$$

ahol $\hat{\in}$ szimbólummal jelölöm, hogy s csúcsa T-nek (levéloperátor);
$p(s)$: az s levél valószínűsége (v. aránya).

$$\hat{p}(s) = \frac{\textit{levélbeli rekordok száma}}{\textit{összes rekordok száma}}$$

$$\hat{p} = \frac{\textit{az } i-\textit{edik osztályba tartozó rekordok száma } s-\textit{ben}}{\textit{összes rekordok száma } s-\textit{ben}}; \quad \hat{I}(s) = I(\hat{p}, 1-\hat{p})$$

$$\tilde{T} = (T \setminus \{s_0\}) \cup \{bal\} \cup \{jobb\}$$

$$E(\tilde{T}) = \sum_{s \hat{\in} \tilde{T}} p(s)\, I(s) = \sum_{\substack{s \hat{\in} T \\ s \neq s_0}} p(s)\, I(s) + p(b)\, I(b) + p(j)\, I(j).$$

Maximalizálandó:

$$E(T) - E(\tilde{T}) = p(s_0)\, I(s_0) - p(b)\, I(b) - p(j)\, I(j) = p(s_0)\left[I(s_0) - \frac{p(b)}{p(s_0)}\, I(b) - \frac{p(j)}{p(s_0)}\, I(j) \right]$$

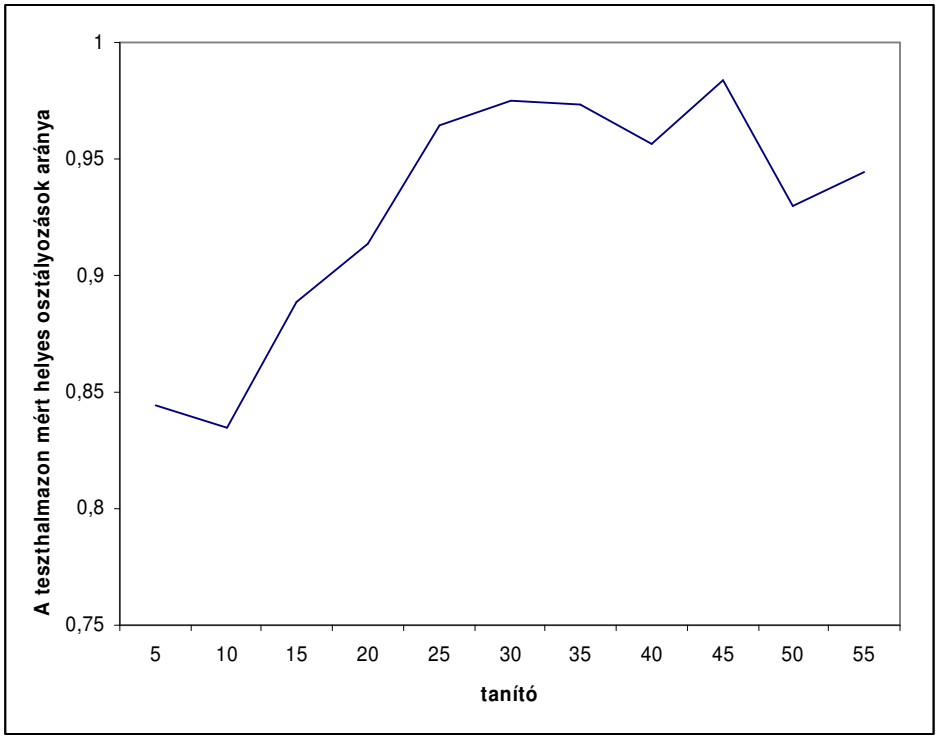
5. A tanuló algoritmus teljesítményének becslése

Amennyiben egy tanuló algoritmus olyan hipotéziseket hoz létre, amelyek az általuk előzetesen nem látott rekordok (megfigyelések) osztályba sorolását jól jósolják meg, akkor az algoritmus jónak mondható. Olyan módszertant kell tehát találni, amellyel az osztályba soroló képességet méréssel lehet becsülni, és megmondható általa, hogy mi módon lehet előre megbecsülni a jóslás minőséget. Egy hipotézis minősége megbecsülhető az ismertté vált tényleges osztályba-sorolások alapján, hiszen akkor jó egy jóslás, ha igaznak bizonyul.
Ez elvégezhető egy teszthalmaznak nevezett mintahalmaz segítségével. Nem lehet az összes rendelkezésre álló példát tanításra használni, mert akkor további adatokat kell gyűjteni a teszteléshez, amint ez már korábban is említésre került. Ezért kényelmesebb

megoldást jelent gyűjteni egy nagy példahalmazt, és ezt szétosztani diszjunkt részekre. [29]

5.1. A tanuló algoritmus tanulási görbéje

A tanító algoritmust a tanító halmazon alkalmazva generálható egy *h* hipotézis, a teszthalmazon pedig megmérhető, hogy a *h* hipotézis hány százalékára ad helyes osztályba-sorolást a tanító halmaz.
A lépések megismétlendők különböző halmazméretekre, és mindegyik mérethez véletlenszerűen kiválasztott különböző tanító halmazokra.
Az eljárás egy adathalmazt állít elő eredményül, amelynek feldolgozásával megkapható az átlagos jóslási képesség a tanító halmaz méretének függvényében. A függvényt ábrázolva megjeleníthető az adott algoritmusnak a tanulási görbéje egy adott tématerületre vonatkozóan.
A döntési-Fa-Tanulás algoritmus a vizsgált feladat (a kiigazító módszerek összehasonlítása) példáival felvett tanulási görbéje a 9. ábrán látható.



9. ábra

Téves osztályozások aránya a különböző méretű tanító halmazok esetén:

Input adatok 5%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0	0.2568807339	0.1559633028
Entropy	0	0.2568807339	0.1559633028
Gini	0	0.2568807339	0.1559633028

Input adatok 10%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.0416666667	0.2233009709	0.1650485437
Entropy	0.0416666667	0.2233009709	0.1650485437
Gini	0.0416666667	0.2233009709	0.1650485437

Input adatok 15%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.1176470588	0.2577319588	0.1313131313
Entropy	0,058824	0,226804	0,111111
Gini	0.1176470588	0.2577319588	0.1313131313

Input adatok 20%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.0652173913	0.1648351648	0.0537634409
Entropy	0,043478	0,186813	0,086022
Gini	0.0652173913	0.1648351648	0.0537634409

Input adatok 25%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.0617283951	0.1477272727	0.0588235294
Entropy	0,035088	0,125	0,035294
Gini	0.0617283951	0.1477272727	0.0588235294

Input adatok 30%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.0724637681	0.1375	0.0617283951
Entropy	0,043478	0,075	0,024691
Gini	0.0724637681	0.1375	0.0617283951

Input adatok 35%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.0506329114	0.1973684211	0.08
Entropy	0,037975	0,052632	0,026667
Gini	0.0506329114	0.1973684211	0.08

Input adatok 40%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.010989011	0.0714285714	0.0434782609
Entropy	0.010989011	0.0714285714	0.0434782609
Gini	0.010989011	0.0714285714	0.0434782609

Input adatok 45%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.0388349515	0.078125	0.0476190476
Entropy	0,009709	0,0625	0,015873
Gini	0.0388349515	0.078125	0.0476190476

Input adatok 50%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.1043478261	0.0517241379	0.0701754386
Entropy	0.1043478261	0.0517241379	0.0701754386
Gini	0.1043478261	0.0517241379	0.0701754386

Input adatok 55%-a tanító halmazban:

Tree Name	Misclassification Rate	Valid: Misclassification Rate	Test: Misclassification Rate
Chi2	0.144	0.1176470588	0.1111111111
Entropy	0.08	0.0784313725	0.0555555556
Gini	0.144	0.1176470588	0.1111111111

A tanító halmaz méretével javul a predikció minősége. Ez okból az ilyen görbéket ún. boldog görbéknek (happy graphs) is nevezik. Ez ugyanis azt jelzi, hogy az adatokban valóban van valami mintázat, és ezt a mintázatot az algoritmus felfedezi.
A tanuló algoritmusnak az előtt nem szabad „látnia" a tesztadatokat, mielőtt tesztelésre nem kerül velük a megtanított hipotézis. Sajnos azonban könnyű abba a hibába beleesni, hogy a tesztadatokra kukucskálunk (peeking) tanítás közben. Ez tipikusan a következőképpen történhet: általában a tanuló algoritmusok számos hangolási lehetőséggel rendelkeznek, amellyel változtatni lehet az algoritmus viselkedését. Számos különböző kritérium lehet például, amelyek alapján kiválasztja a következő attribútumot a döntési fa a tanulásnál. A különböző beállítások mentén több különböző hipotézist generál, majd a teszthalmazon leteszteli mindegyiket, és ezután a legjobb hipotézis szerinti predikciós eredményt tekinti a döntés eredményének. Ilyenkor azonban kukucskálás történik. Ennek az oka, hogy a hipotézist a teszthalmazon mért eredménye alapján választja ki az algoritmus, melynek következtében beszivárog a teszthalmaz által tartalmazott információ a tanuló algoritmusba. Következésképpen minden eljárásnak, amely összehasonlítja a hipotéziseknek a teszthalmazon nyújtott teljesítményét, egy új teszthalmazt kellene használnia, hogy kukucskálás nélkül mérhesse a végül kiválasztott hipotézis teljesítményét. Ez viszont túlságosan problémás, nehézkes a gyakorlatban, így továbbra is az előbb leírt módon, szennyezett adathalmazokon kerülnek futtatásra az algoritmusok. [29]

5.2. Zaj és túlilleszkedés

Ha van legalább kettő vagy több példa, amelyek az attribútumokra nézve azonos leírással rendelkeznek, de az osztályba sorolásuk eltérő, akkor nem lehet képes a Döntési-Fa-Tanulás algoritmus olyan döntési fát találni, amely konzisztens mindegyik példával.
Az igény az lenne, hogy determinisztikus osztályzás esetén minden levélcsúcs adja vissza a hozzá tartozó halmaz többségi osztályát, más esetekben pedig a relatív gyakoriságok alapján becsült osztályba tartozási valószínűségeket. Valójában elég valószínű azonban, hogy a tanuló algoritmus akkor is talál olyan döntési fát, amely

konzisztens az összes példával, ha lényeges információ hiányzik. Ez azért történhet meg, mert az algoritmus irreleváns attribútumokat is felhasználhat, amennyiben vannak, és ebből kifolyólag hamis megkülönböztetést tesz a példák közt.
Amikor a hipotézisek nagy halmaza lehetséges, akkor megnő a veszélye annak is, hogy az algoritmus értelmetlen „szabályosságot" talál az adatokban. Ez a probléma a túlilleszkedés (overfítting). Ez általános jelenség, akkor is előállhat, ha a keresett függvénynek egyáltalán nincs valószínűségi jellege. Ez a probléma nem csak a döntési fákat, hanem az összes tanulási algoritmust sújtja.
A probléma megoldására van egy egyszerű, ún. döntési fa metszés (decision tree pruning) nevű módszer. A metszés (nyesés) működése azon alapul, hogy a nem nyilvánvalóan releváns attribútumok tesztje mentén akkor is megakadályozza a mintahalmaz ismételt (rekurzív) felosztását, ha nem egyforma az adott csúcsban az adatok osztályba-sorolása. Kérdés, hogy hogyan lehet észrevenni egy attribútumról, hogy irreleváns?
Amennyiben egy irreleváns attribútumra alapozva került kettéosztásra a mintahalmaz, akkor az várható, hogy ilyenkor az eredményül kapott részhalmazokban az egyes osztályokba tartozó minták nagyjából ugyanabban az arányban fognak szerepelni, mint az eredeti halmazban. Ekkor közel nulla az információnyereség, melynek hiánya tehát jó jelzés lehet az irrelevanciára.
További kérdés, hogy mekkorának kell lennie az információnyereségnek ahhoz, hogy szétoszthtató legyen a mintahalmaz ezen attribútum mentén. Ennek eldöntésére statisztikai szignifikancia-tesztet kell használni, amely azt a nullhipotézist teszteli, hogy egyáltalán nincs közös mintázat a példákban.
Ezután az aktuális adathalmazt vizsgálva meg kell határozni annak mértékét, hogy mennyire tér el az adathalmaz a tökéletesen mintázat nélkülitől. Ha statisztikailag már valószínűtlennek tekinthető az eltérés mértéke (5% vagy ennél kisebb valószínűség jellemzi), akkor ez annak bizonyítékaként vehető, hogy jelen van egy alapvető mintázat az adatokban, ahol a valószínűség a véletlen mintavételezés esetén várható eltérések standard eloszlását feltételezve kerül kiszámításra.
Ekkor a nullhipotézis az, hogy az attribútum irreleváns, ennek következtében nulla lenne egy végtelen nagy mintahalmazra vett információnyereség. Azt kell kiszámítani a nullhipotézist feltéve, hogy milyen valószínűséggel léphet fel a megfigyelt eloszlásnak a várt pozitív és negatív eseteloszlástól való eltérése egy v méretű mintahalmazban. Az eltérés mértéke megadható a részhalmazok tényleges pozitív és negatív esetszámának a nullhipotézis fennállása esetén várt esetszámokkal való összehasonlításával.
Feltéve a nullhipotézist, az eltérés mérték v-1 szabadságfokú χ^2 eloszlást követ (az attribútum irrelevanciájának valószínűsége tehát a standard χ^2 eloszlás segítségével számítható).
A zaj, metszés alkalmazásával kezelhető. Az osztályozási hibák a predikciós hibában lineáris növekedést okoznak. A metszéssel készült döntési fák lényegesen jobb eredményt adnak, mint a metszés nélkül készültek, amennyiben az adatok nagy zajjal terheltek. Továbbá a metszéssel készült fák könnyebben érthetők, mivel gyakran lényegesen kisebbek.

A **keresztvalidáció** (cross-validation) egy másik túlilleszkedést csökkentő módszer, amely annak becslésén alapul, hogy a még nem látott esetekre adandó válaszokat az egyes hipotézisek mennyire jól fogják megjósolni. Ennek becsléséhez az ismert adatok egy részét félreteszi, és ezekkel teszteli a megmaradt adatok alapján tanulással létrehozott fa predikciós képességét. K-szoros keresztvalidációról van szó, ha k kísérletet végezve minden esetben az adatok más és más 1/k-ad része lesz félretéve

validációs tesztcélra, majd a végén az eredmények átlagolásra kerülnek. Elterjedten használt *k* érték az 5 és a 10. Speciális esetként használják a *k* = n választást, amelyet *hagyj-ki-egyet* keresztvalidációs módszernek neveznek. A keresztvalidáció tetszőleges döntési fa tanulási módszerrel együtt alkalmazható (a metszést is beleértve). Célja a jó predikciós képességgel rendelkező döntési fa kiválasztása, mely predikciós képességet egy új teszthalmazon kell mérni a kukucskálási jelenség elkerülése végett.[29]

6. Az osztályozási feladat eredményeinek értelmezése

A következő táblázat mutatja a döntési feladat 7. ábra szerinti fa alapján végrehajtott osztályozásának eredmény tábláját (az 8. ábra szerinti adatparticionálás után):

SOURCE	STAT	TRAMO_IN	==> 0	==> 1	TOTAL
TRAIN	N	0	46	0	46
TRAIN	N	1	1	44	45
TRAIN	N	+	47	44	91
TRAIN	Row%	0	100	0	100
TRAIN	Row%	1	2	98	100
TRAIN	Row%	+	52	48	100
TRAIN	Col%	0	98	0	51
TRAIN	Col%	1	2	100	49
TRAIN	Col%	+	100	100	100
TRAIN	%	0	51	0	51
TRAIN	%	1	1	48	49
TRAIN	%	+	52	48	100
VALID	N	0	32	3	35
VALID	N	1	2	33	35
VALID	N	+	34	36	70
VALID	Row%	0	91	9	100
VALID	Row%	1	6	94	100
VALID	Row%	+	49	51	100
VALID	Col%	0	94	8	50
VALID	Col%	1	6	92	50
VALID	Col%	+	100	100	100
VALID	%	0	46	4	50
VALID	%	1	3	47	50
VALID	%	+	49	51	100

3. táblázat

SOURCE jelzi, hogy a statisztika a tanító (TRAIN), vagy a validáló (VALID) adatokból származik.

STAT mutatja a megfigyelések számát (N), a sor százalékot (Row%), az oszlop százalékot (Col%) és a mindösszesen százalékot (%).

TRAMO_IN mutatja a célváltozó értékeit (jelen esetben 0 és 1); a + szimbólum reprezentálja a célváltozó mindösszesen értékét.

==>0 és **==>1** oszlopok tartalmazzák az előrejelzett célváltozó statisztikáit a tábla minden sorára.

A kapott fa alapján a célváltozónak

- 44 db ’1’ értéke került helyesen osztályozásra a tanító halmazból;
- 1 db ’1’ értéke került rosszul (0 érték szerint) osztályozásra a tanító halmazból;
- 46 db ’0’ értéke került helyesen osztályozásra a tanító halmazból;
- Egy ’0’ értéke sem került rosszul (1 érték szerint) osztályozásra a tanító halmazból.

Az sor-, oszlop-, és mindösszesen százalékok hasonló módon értelmezendők.

A TRAMO_IN célváltozóra vonatkozó levél statisztikák:

Node	Leaf	N	N * PRIORS	V N	V N * PRIORS	%V 0	%V 1	%0	%1
8	1	13	13	10	10	0.00	100.00	0.00	100.00
9	2	8	8	10	10	90.00	10.00	100.00	0.00
5	3	31	31	26	26	11.54	88.46	0.00	100.00
3	4	39	39	24	24	95.83	4.17	97.44	2.56

4. táblázat

Leaf jelzi a levélcsúcsok azonosító számát.

N a megfigyelések számát mutatja minden levélben a tanító adathalmazra vonatkozóan.

VN a megfigyelések számát mutatja minden levélben a validáló adathalmazra vonatkozóan (gondoskodva a validáló adathalmaz használatáról).

%1 és **%0** a ’1’ ill. ’0’ értékek százalékát mutatja minden levélben a tanító adathalmazra tekintettel.

%V1 és **%V0** az ’1’ ill. ’0’ értékek százalékát mutatja minden levélben a validáló adathalmazra tekintettel.

A következő táblázat a tanító és validáló adathalmazok minden részfájában helyesen osztályozott megfigyelések arányát listázza ki:

Misclassification Rate

Leaves	Training	Validation
1	0.4945	0.5000
2	0.0989	0.1857
3	0.0989	0.1857
4	0.0110	0.0714
5	0.0110	0.0714

5. táblázat

A 5. táblázat egy olyan táblát tartalmaz, amely mértéket szolgáltat arra, hogy a fa mennyire jól írja le az adatokat. Minden részfára értékelő statisztikákat listáz ki a tanító és a validáló adathalmaz szerint (a statisztika típusa függ a célváltozó típusától és a **Tree** konfigurációs interfészének ’Advanced’ lapján választható modellértékelési mértéktől).
A tábla megjeleníti az értékelést számos adatpartícó jelöltre. Ha van validáló adathalmaz, a validáló adatokon alapuló értékelés megbízhatóbb, mint a tanító adatokon nyugvó.
Alapértelmezésben a legjobb értékkelés –ezen belül– a legkevesebb levéllel rendelkező fa ki van jelölve, mely fának az értékelő grafikonbeli függőleges referencia vonal felel meg (ld. 10. ábra). Ez a vizsgált munkában a 4 levéllel rendelkező fát jelenti, mert az 5

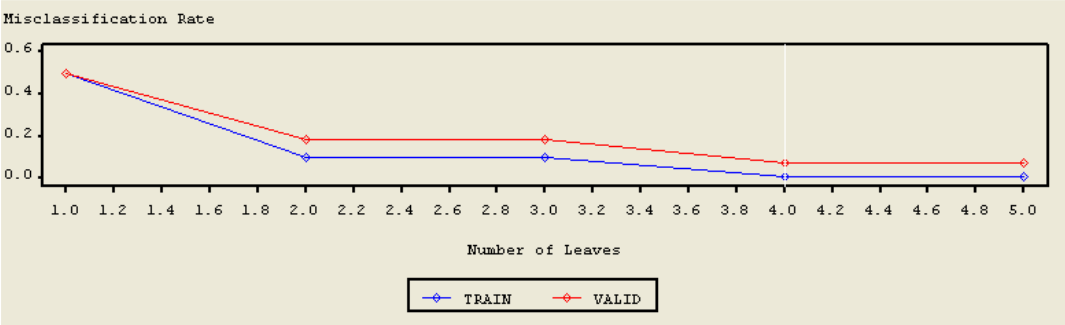
levéllel rendelkező esetében már nem javul tovább a téves osztályozási arány (Misclassification Rate).
A totális levélszennyezettség Gini index szerinti mértéke még némi javulást mutat, de nem jelentős:

Average Square Error (Gini index)

Leaves	Training	Validation
1	0.4999	0.5001
2	0.1702	0.2960
3	0.1303	0.2574
4	0.0214	0.1418
5	0.0176	0.1371

6. táblázat

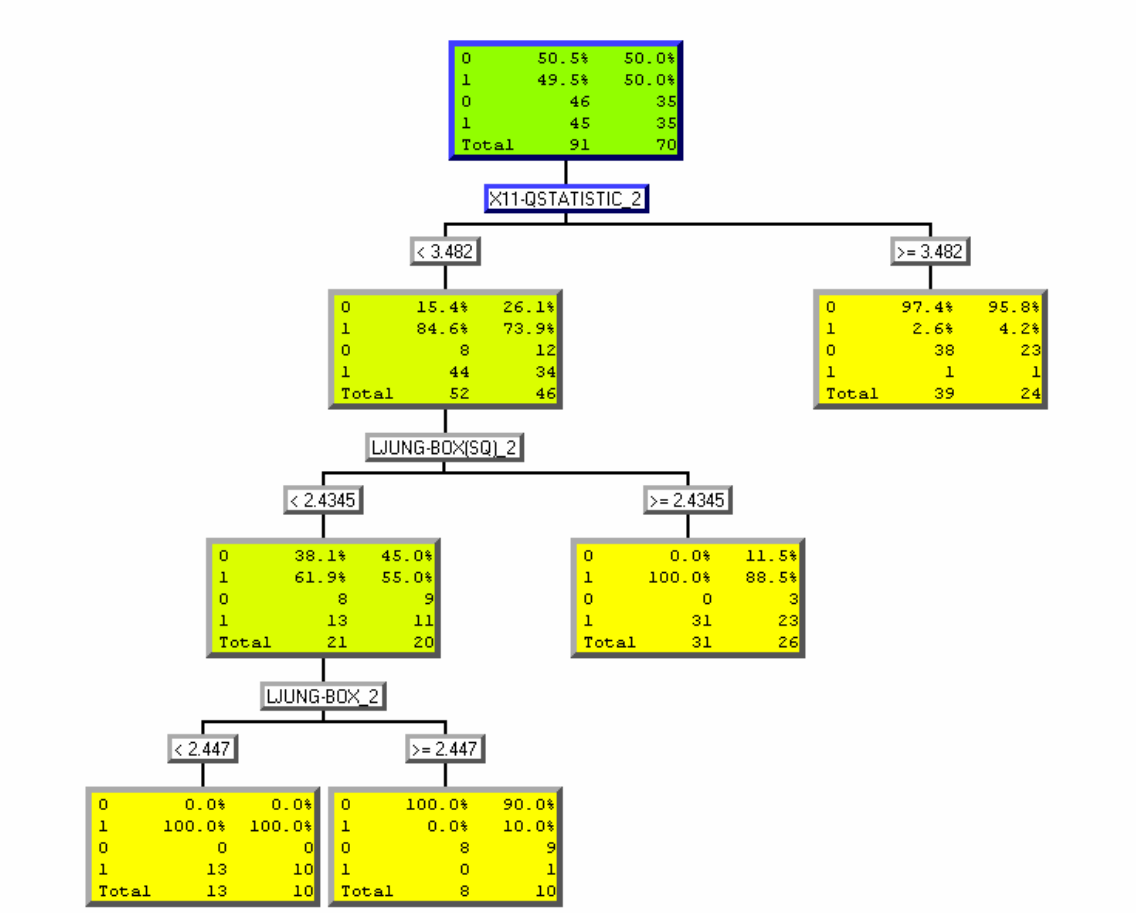
A következő ábra az eredmények grafikus megjelenítését mutatja:



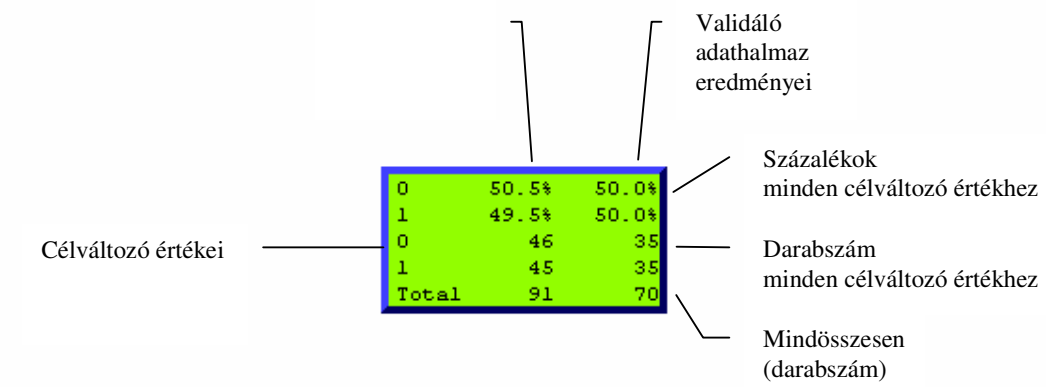
10. ábra

Az eredmények grafikus megjelenítésére szolgáló 'Plot' lap a függőleges tengelyén a különböző részfákra vonatkozó értékelés grafikonját jeleníti meg. A **Tree** automatikusan kiválasztja azt a részfát, amely optimalizálja a modellnek az 'Advanced' lapon (a **Tree** konfigurációs interfészén) választható értékelését. A függőleges fehér referencia vonal azonosítja ezt a részfát (a 4 levélszámú részfa).

A végleges, az értékelés által elfogadott fa (partíció) tehát a következőképpen néz ki:



A csúcsok az alábbi statisztikákat tartalmazzák:



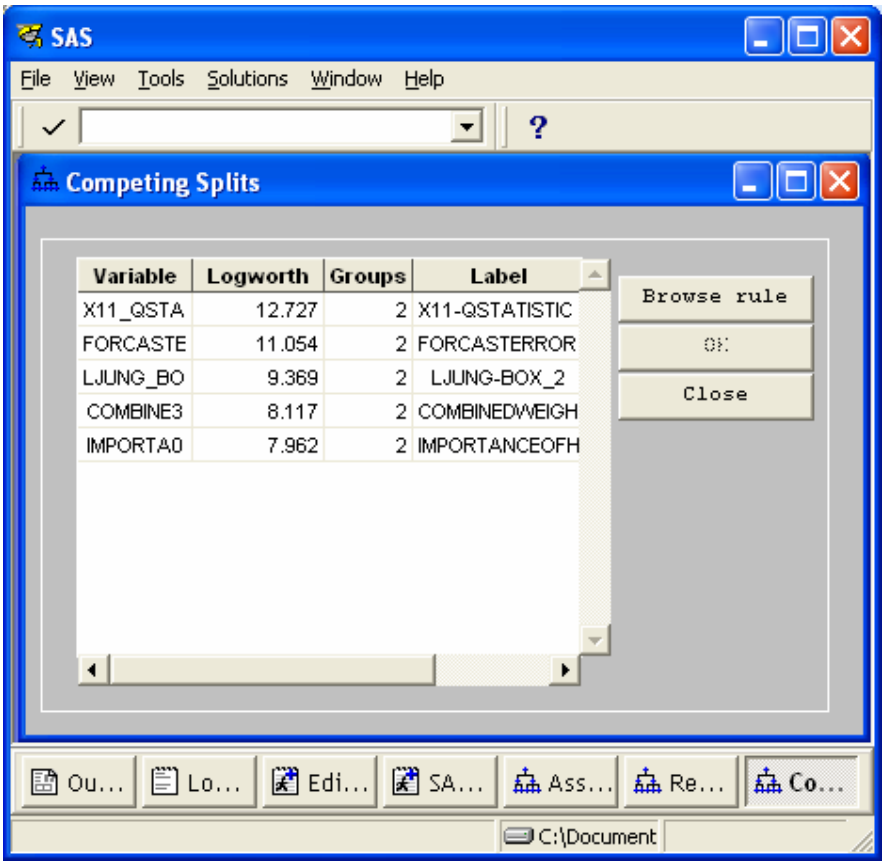
11. ábra

A **Tree** diagram a részei:

- Gyökér csúcs – az összes megfigyelést tartalmazza.
- Belső csúcs – nem befejező csúcsok, amelyek vágási szabályt tartalmaznak.
- Levél csúcs – befejező csúcsok, amelyek a végső osztályozást tartalmazzák a megfigyelések egy halmazára.

A numerikus címke közvetlenül a csúcsok fölött jelöli, hogy melyik értékpontnál találta a **Tree** szignifikánsnak a vágást (jelen esetben) a középre pozicionált cimkével jelölt intervallum típusú magyarázó változón. A jelen vizsgálat során a kezdeti vágás az X11-QSTATISTIC_2 változón volt végrehajtva. A vágás után előállt első (bal) al-szegmens csúcsa lényegesen több ’1’ osztályú megfigyelést tartalmaz, mint ’0’-t, míg a második (jobb) sokkal több ’0’ osztályút, mint ’1’-est, amelyből mindössze csak egyet. Az X11-QSTATISTIC_2 változón végrehajtott vágás tehát nagyon hatásosan szétválogatja a vizsgált adathalmaz rekordjait. A két LJUNG-BOX változó alapján következnek a további rekurzív vágások.

A következő ablakban a program azokat a vágási szabályokat listázza ki, amelyek a gyökér csúcsban szintén alkalmazhatóak lennének:



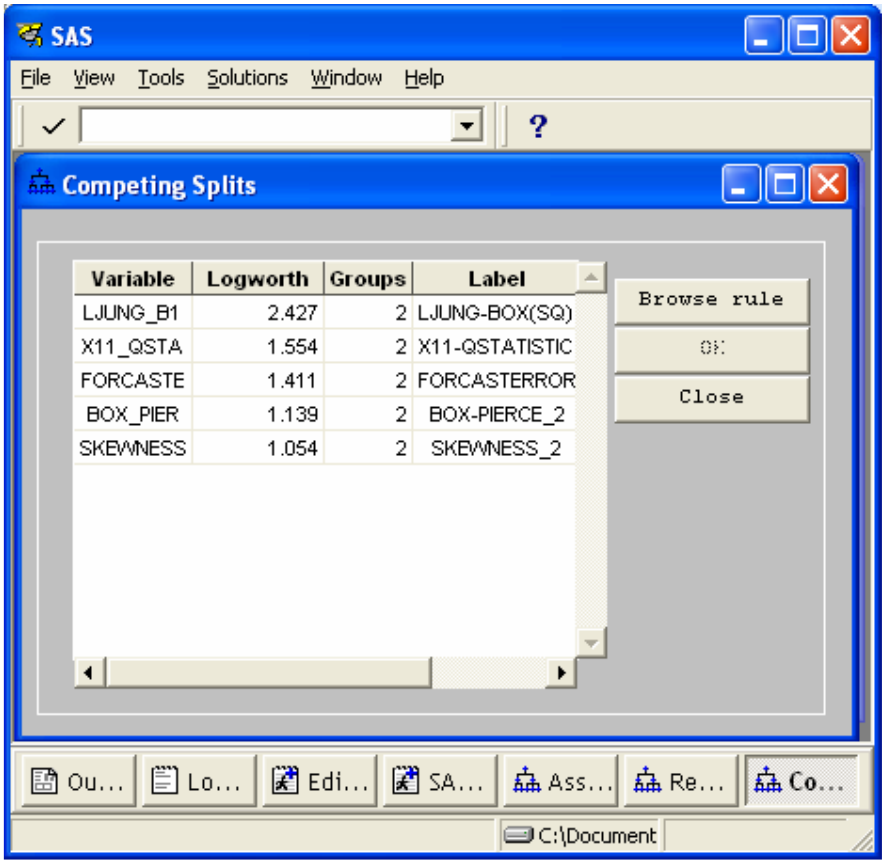
12. ábra

Tallózni lehet a vágási szabályok és értékmérői között (mennyire jó egy változó a kezdeti adathalmaz vágására). Az értékmérő jelzi, hogy mennyire jól osztja el egy

változó az adatokat az egyes osztályokba. A jó vágó változók nagy **LOGWORTH** értékkel rendelkeznek.

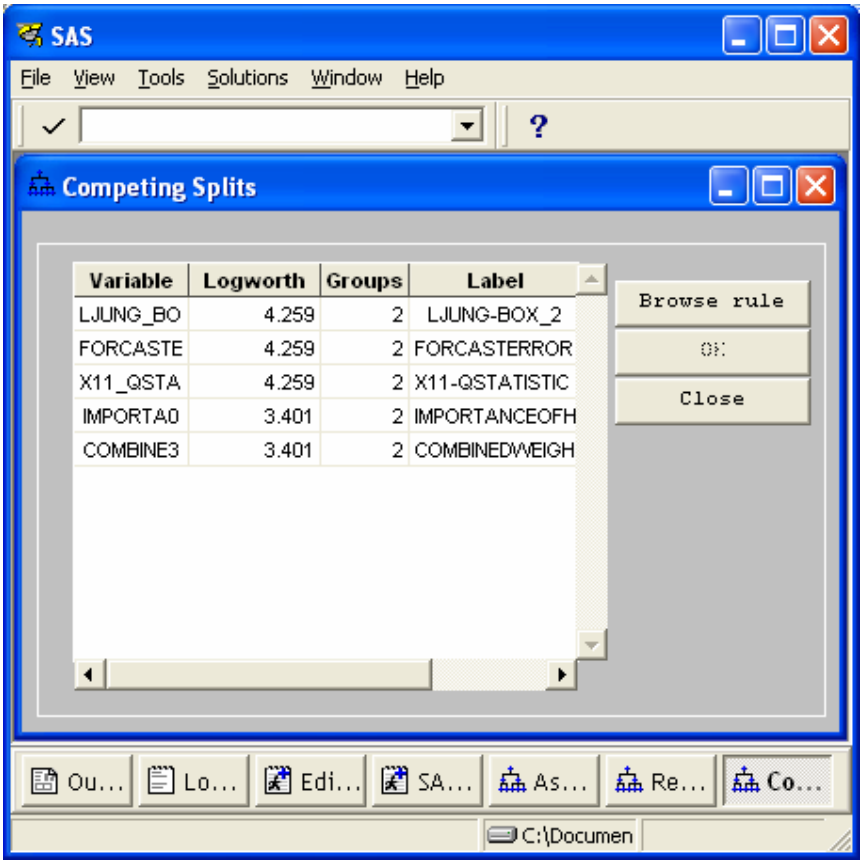
Kategoria célváltozó esetén: LOGWORTH = -log(Chi-square p-value)
Intervallum célváltozó esetén: LOGWORTH = -log(F test p-value)

A legjobbnak itt is az X11-QSTATISTIC_2 látszik, de nem sokkal gyengébb a FORCASTERROR_2 sem, vagyis nem rossz eredménnyel lehetne használni ezt a változót is a kezdeti vágás végrehajtására.
A 2-es csúcsban a LJUNG-BOX(SQ)_2 után a következő legjobb ugyancsak az X11-QSTATISTIC_2 lenne, de itt az ötös lista végén már megjelennek a BOX-PIERCE_2 és a SKEWNESS_2 változók is:



13. ábra

A 4-es csúcsban pedig az IMPORTANCEOFHISTORY_1 és a COMBINEDWEIGHTS_2 változók is szóba jöhetnének még az eddig említetteken kívül:



14. ábra

Az egyes változóknak a fa felépítésében betöltött szerepét, fontosságát mérlegelve tehát konklúziót lehet vonni, miszerint a predikciós vizsgálatok is alátámasztják-e a KSH-beli szezonális kiigazítás során alkalmazandó módszertan kiválasztására meghozott döntésünket.

IV. Összefoglalás

Megállapítható, hogy a módszertani összehasonlítás szempontjából a két legmeghatározóbb diagnosztika a Demetrában az X11-Q statistic és a LJUNG-BOX tesztstatisztika. Az előbbi az A X12-ARIMA, az utóbbi a TRAMO/SEATS tesztstatisztikája. Az X12-Arima *QualityIndex*-e azonban kedvezőtlenebb változást mutat, amennyiben a kiigazítás minőségét mutató indexet összeállító TQM-ben az X11_Q statistic diagnosztikát kiemelt (2-szeres) súllyal veszem figyelembe, mint a Tramo/Seats, amikor is LJUNG-BOX statisztikát súlyozom kiemelten.
Mindezek alátámasztják tehát –több egyéb érv mellett– azon döntésünket, mikor is a TRAMO/SEATS kiigazító módszertant tekintettük megfelelőbbnek a KSH idősorainak kiigazítására, és ennek megfelelően a 2002-es próbaévtől kezdődően a TRAMO/SEATS módszertan bevezetésére, ill. alkalmazására tettünk javaslatot ***módszertani váltás*** keretében.

MÓDSZERTANI VÁLTÁS A KSH-BAN

2002 előtt a KSH-ban csak néhány területen végeztek szezonális kiigazítást, többnyire az X11 és Xll-ARIMA módszerrel. Általában az év utolsó adatával, automatikus futtatással elvégezték a szezonális kiigazítást, majd a program által a következő időszakra előrejelzett szezonális tényezők alapján határozták meg a szezonálisan kiigazított idősor értékeit. A munkanapok és a húsvét hatását általában nem vették figyelembe, ahol azonban a munkanaphatást is számszerűsítették, ott ezt havonta határozták meg az előre becsült szezonális tényező leválasztása után. A trendet havonta újrabecsülték a program segítségével. Sok területen alkalmazták a szezonális kiigazítás mellett vagy helyett az eredeti idősorból képzett „időszak / előző év azonos időszaka" mutatót is.

Az Eurostat már jóval a csatlakozás előtt igényelte a leendő tagországok adatait is, nem csak alapadat formájában, hanem munkanappal kiigazítva, illetve szezonálisan és munkanappal kiigazítva is. Annak érdekében, hogy az egyes országok adatai összehasonlíthatóak, és aggregátumok képzésére alkalmasak legyenek, egységes elvárásokat fogalmazott meg a jelenlegi és leendő tagországok számára (Eurostat, 1998). Két kiigazítási módszert ajánlott: az X12-ARIMA és a TRAMO/SEATS módszert. A könnyebb használhatóság kedvéért kifejlesztetett egy kezelőfelületet (Demetra), amely mindenki számára ingyenesen hozzáférhető (ill. letölthető az Eurostat honlapjáról) és mindkét módszert tartalmazza.

2001 második felében a módszertani osztály munkatársaiból és az egyes szakstatisztikák szakértőiből megalakult a szezonális kiigazítás harmonizációjára egy munkacsoport, melynek magam is tagja voltam. A csoport összetételének kialakításánál alapvető szempont volt, hogy minden olyan főosztály részt vegyen a munkában, ahol évesnél nagyobb gyakoriságú adatgyűjtések feldolgozása folyik.

További szempont volt, hogy majd a kialakításra kerülő gyakorlatot szakstatisztikától független szakértők koordinálják, akik minden szakterület szempontját egyformán figyelembe tudják venni, hiszen a szezonális kiigazítás módszertana független a

szakstatisztikáktól (bár a kiigazítás végeredményét nagyban befolyásolja a szakstatisztikusok véleménye, például az outlierek magyarázhatóságáról, a munkanap- és húsvéthatás meglétéről), speciális matematikai statisztikai ismereteket igényel, és az Eurostat előírásai a fő irányelvek tekintetében minden területen azonosak. Ezért lehetővé vált, hogy ne valamelyik szakfőosztály koordinálja a szezonális kiigazítást, hanem erre specializálódott munkatársak, egy munkacsoportot alakítva végezzék el ezt a feladatot.

A munkacsoportbeli tevékenységünk során egy elemzésekre épülő, megalapozott javaslatot igyekeztünk tenni a szezonális kiigazítás új rendszerére vonatkozóan, amely mind a Hivatal, mind a partnerintézmények és a felhasználók számára elfogadható. A munkacsoport célja egy olyan egységes kiigazítási politika kialakítása volt, amely minden olyan idősorra, amelyet a KSH-ban ki kell igazítani, megfelelő minőséggel alkalmazható, ugyanakkor 2002 elejétől bevezetésre kerülhet.

A korábbi gyakorlat feltérképezése mellett az Eurostattal való összhang érdekében megvizsgáltuk az Eurostat által ajánlott XI2-ARIMA és a TRAMO/SEATS módszert.

A 176 idősoron a Demetra segítségével elvégzett vizsgálatok, tesztelések során a 11 szempont alapján értékeltük a módszereket:[15]

Értékelési szempont	TRAMO/SEATS	X12-ARIMA
Nemzetközi ajánlásoknak megfelelő módszer	+	+
Tudományosan elfogadott, korszerű módszer	++	+
Rövid idősorok kezelése	-	--
Magyar ünnepnapok kezelése	+	-
Stabil eredmények az idősor végén	++	+
Szezonalitás szűrésének hatásossága	++	+
Eredmények statisztikai diagnosztikája	+	++
Automatikus futtatás lehetősége	+	+
Nagyszámú idősor kezelése	++	+
Könnyen kezelhető input és output fájlok	+	+
Felhasználóbarát kezelőfelület	+	+

A Demetra környezetéből adódóan az automatikus futtatás lehetősége, a felhasználóbarát kezelőfelület és a könnyen kezelhető input és output fájlok tekintetében nem volt különbség a két módszer között. Szintén mindkét módszer megfelel a nemzetközi ajánlásoknak, hiszen az Eurostat ajánlásának megfelelően választotta ki a munkacsoport ezt a két szóba jöhető módszert.
A TRAMO/SEATS módszertan elmélete az idősorelemzés korszerűbb, sztochasztikus módszerein alapszik.

[15] A „+" és „-" jelek a két módszer közötti különbséget tükrözik egy szempontra vonatkozóan, de nem alkalmasak egy módszer különféle szempontok szerinti összehasonlítására, pl. a táblázatból nem következik, hogy a TRAMO/SEATS tudományosan jobban elfogadott, mint amennyire nemzetközileg ajánlott.

A rövid idősorok kezelése mindkét módszer esetében problematikus kérdés. A rövid idősorok problémája azonban az ad hoc filterek sajátosságainak következtében érzékenyebben érinti az X12-ARIMA eljárást.
A TRAMO/SEATS módszertan továbbá fel van készítve a magyar ünnepnapok kezelésére is, valamint stabilabb eredményeket szolgáltat, mivel az újabb adatok megjelenése a becslés során mérsékeltebb változást okoz az idősor végén, mint az X12-nél használt ad hoc filterek esetén.
A módszer vizsgálata mellett sor került a szezonálisan kiigazítandó idősorok matematikai statisztikai szempontból történő vizsgálatára, szakmailag adekvát modellezésére, a kiugró értékek, trendváltozások, esetleges strukturális törések, egyéb befolyásoló tényezők (pl. munkanapok, ünnepnapok) hatásának kimutatására, szignifikanciájuk vizsgálatára, megvitatásra, illetve szakstatisztikai, valamint közgazdasági szempontból történő magyarázatára.

Az összehasonlítás eredményeit összesítettük, és ezek alapján úgy ítéltük meg, hogy a hazai körülmények között a TRAMO/SEATS jól, illetve jobban alkalmazható, mint az X12-ARIMA, így a TRAMO/SEATS alkalmazását javasoltuk a Hivatalon belül. A program matematikai-statisztikai háttere jobb eredményeket tesz lehetővé, mint az alternatív program, az X12-ARIMA. A mozgóátlagolással végzett becslések csak kellően hosszú idősorok esetén képesek jó hatásfokkal kiszűrni a véletlen ingadozások hatását, így az X12-ARIMA a tartós tendenciát és a szezonálisan kiigazított értékeket is csak nagyobb ingadozással, illetve nagyobb bizonytalansággal képes becsülni. Továbbá a TRAM0/SEATS által generált trend és a szezonálisan kiigazított idősor vége sokkal kevésbé érzékeny egy újabb adat megjelenésére, ezért az időszak végére lényegesen stabilabb eredményeket produkált.
Belső egyeztetések után 2002 I. negyedévében külső szakértők (Pénzügyminisztérium, Magyar Nemzeti Bank) részvételével szakmai fórumon kerültek bemutatásra és megvitatásra az eredmények.
A belső és külső egyeztetések után elkészült a döntési javaslat a KSH vezetősége részére, amely tartalmazta a munka főbb eredményeit, és ennek alapján javaslatot tett az alkalmazandó módszerre.

Végül 2002-től, mint próbaévtől kezdődően új, egységes módszertant vezetett be a KSH[16], amely összhangban van a hazai- és az Eurostat-elvárásokkal, és teljesíti a kiigazítással szemben támasztott követelményeket.
Az új gyakorlatot a 2002. február 5-i Gazdaságstatisztikai Felhasználói Fórum elfogadta, majd a módszertani váltást a 2002. február 18-i Elnöki Értekezlet jóváhagyta. Ezután a KSH sajtóközleményben (2002. március 14.) tájékoztatta a közvéleményt a Hivatal szezonális kiigazítási módszertanában történt változásról.

[16] A próbaév tapasztalatainak kiértékelése után a szezonális kiigazítás egységes gyakorlatáról szóló szabályzat, amely a KSH-ban kiigazításra kerülő összes idősorra vonatkozik 2004 áprilisában lépett életbe.

IRODALOMJEGYZÉK

[1] A. Leonte – R. Trandafir: *A valószínűségszámítás klasszikus és aktuális problémái*.

Műszaki Könyvkiadó, Budapest, 1986.

[2] Besenyei Lajos – Gidai Erzsébet — Nováky Erszébet: *Jövőkutatás, előrejelzés a gyakorlatban*. Módszertani kézikönyv.

Közgazdasági és Jogi Könyvkiadó, Budapest, 1977.

[3] Ezekiel Mordecai: *Korreláció és regresszió-analízis*: *Lineáris és nemlineáris módszerek.*

Közgazdasági és Jogi Könyvkiadó, Budapest, 1970.

[4] Éltető Ödön – Meszéna György – Ziermann Margit: *Sztochasztikus módszerek és modellek*.

Közgazdasági és Jogi Könyvkiadó, Budapest, 1982.

[5] Dr. Ferenczy Pál: *Kommunikációs Eszközök*.

LSI Oktatóközpont, 1992

[6] Frederick S. Hillier – Gerald J. Liebermann: *Bevezetés az operációkutatásba*.

LSI Oktatóközpont, Budapest, 1994.

[7] Füstös László – Kovács Erzsébet: *A számítógépes adatelemzés statisztikai módszerei*.

Tankönyvkiadó, Budapest, 1989.

[8] George E: P. Box & Gwilym M. Jenkins: *Time series analysis forcasting and control*

Holden Day, San Francisco, Cambridge, London, Amsterdam, 1970

[9] Hajtman Béla: *Bevezetés a matematikai statisztikába*. 2. kiadás.

Akadémiai Kiadó, Budapest, 1971.

[10] Tusnády Gábor – Ziermann Margit: *Idősorok analízise.*

Műszaki Könyvkiadó, Budapest, 1986.

[11] Köves Pál – Párniczky Gábor: *Általános statisztika*.

Közgazdasági és Jogi Könyvkiadó, Budapest, 1973.

[12] Köves Pál – Párniczky Gábor: *Általános statisztika*. 2. javított kiadás.

Közgazdasági és Jogi Könyvkiadó, Budapest, 1975.

[13] LOTHAR SACHS: *Statisztikai módszerek*.

Mezőgazdasági Kiadó, Budapest, 1985.

[14] Ludwig Arnold: *Sztochasztikus differenciálegyenletek. Elmélet és alkalmazás.*

Műszaki Könyvkiadó, Budapest, 1984.

[15] Meszéna György – Ziermann Margit: *Valószínűségelmélet és matematikai statisztika*.

Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.

[16] Mundruczó György: *Alkalmazott regressziószámítás*.

Akadémiai Kiadó, Budapest, 1981.

[17] Obádovics J. Gyula: *Valószínűségszámítás és matematikai statisztika*

Scolar Kiadó, Budapest, 1997.

[18] Prékopa András: *Valószínűségelmélet műszaki alkalmazásokkal.* 2. kiadás.

Műszaki Könyvkiadó, Budapest, 1972.

[19] Reimann József – Tóth Julianna: *Valószínűségszámítás és matematikai statisztika*.

Nemzeti Tankönykiadó, Budapest, 1985.

[20] R. L. Kashyap – A. Ramachandra Rao: *Dinamic Stochastic Models from Empirical Data.*

Academic Press – New York – San Francisco – London, 1976.

[21] Samuel Karlin – Howard M. Taylor: *Sztochasztikus folyamatok.*

Gondolat Kiadó, Budapest, 1985.

[22] Shirley Dowdy and Stanley Wearden – John Wiley & Sons: *Statistcs for Research* West Virginia University

New York – Chichester – Brisbane – Toronto – Singapore, 1983

[23] Some Noises with $1/f$ Spektrum, a Bridge Between Direct Current and White Noise, THE TRANSACTIONS ON INFORMATION THEORY, VOL. IT–13, No. 2, APRIL 1967

[24] J. Morgan, John Wiley & Sons, Andrew F. Siegel: *Statistics and Data Analysis an introduction*

– University of Washington at Seattle, 1983

New York – Chichester – Brisbane – Toronto – Singapore

[25] Móri F. Tamás és Székely J. Gábor: *Többváltozós statisztikai analízis*.

Műszaki Könyvkiadó, Budapest, 1986.

[26] Várlaki Péter: *BEVEZETÉS A STATISZTIKAI RENDSZER-IDENTIFIKÁCIÓBA.*

Műszaki Könyvkiadó, Budapest, 1986.

[27] Deak István: *Véletlenszám-generátork és alkalmazásuk*. (*Az operációkutatás matematikai módszerei.*)

Akadémiai Kiadó, Budapest, 1986.

[28] W. Feller: *Bevezetés a valószínűségszámításba és alkalmazásaiba.*

Műszaki Könyvkiadó, Budapest, 1978

[29] Stuart Russell – Peter Norvig: *Mesterséges Intelligencia*, 2005 Panem Könyvkiadó

[30]web:http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf

[31] Bauer P.- Földesi E.(2003). *Észrevételek az idősor elemzési módszerek alkalmazásával kapcsolatos kérdésekhez.* Statisztikai szemle, 81. évf. 9. szám, szeptember, 826-831.p.
[32] Bauer P.- Földesi E.(2004). *A szezoális kiigazítás harmonizációja a Központi Statisztikai Hivatalban. Statisztikai szemle*, 82. évf. 8. szám, augusztus, 691-704. p.
[33] Brockwell, P.J.-Davis, R.A. (1996). *Introduction to Time Series and Forecasting*. New York: Spinger.
[34] Eustat (1998a). Seasonal Adjustment Methods – *A Comparison for Indrusty Statistics, revised version*. Luxembourg.
[35]web:http://forum.europa.eu.int/Public/irc/dsis/eurosam/library?l=/documents_methodological&vm=detailed&sb=Title. Utolsó megtekintés:2005. július 12.
[36] Eurostat (1998b). *Seasonal Adjustment Policy* – Some Eurostat Proposals. SAM 98 Seminar, 22-24 October, Bucharest.
[37]web: http://europa.eu.int/en/comm/eurostat/reserch/noris4/documents/policy/index.htm . Utolsó megtekintés:2005.július 12.
[38] Eurostat (2003). Feasibility study about Demetra in Eurostat/NSIs and ECB/NCBs. Redndelkezésünkre bocsátotta Jean-Marc Museux az Eurostat részéről.
[39] Eurostat (2004a). *Recommendations for Seasonal Adjustment in STS*. Working Party, 12 November, Luxembourg.
[40] Eurostat (2004b). *Recommendations for Working – Day Adjustment in STS*. Working Pary, 12 November, Luxembourg.
[41] Eurostat (2004c). *Seasonal Adjustment in Eurostat: new parameters. Working Party*, 12 November, Luxembourg.
[42] Fischer, B. (1995). *Decomposition of Time Series – Comparing Different Methods in Theory and Practice*. Luxembourg.
[43] web: http://europa.eu.int/comm/eurostat/research/index.htm?http://europa.eu.int/en/comm/eurostat/research/noris4/&1. Utolsó megtekintés: 2005. július 12.
[44] Gómez, V. – Maravall, A.(1996). *Programs TRAMO (Time series Regression with ARIMA noise, Missing observations, and Outliers) and SEATS (Signal Extraction in ARIMA Time Series). Instructions for the User*. Working Paper 9628, Servicio de Estudios, Banco de Espana.
[45] Gómes, V. – Maravall, A. (2001). *Automatic Modelling Methods for Univariate Series*. In: Pena, D. – Tiao, G.C. – Tsay, R.S. (eds.):A Course In Time Series Anylysis. New York: J. Wiley and Sons.
[46] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton Univercity Press.
[47] Maravall, A. (1995). *Unobserved Components in Economic Time Series*. In: Pesaran, H.- Schmidt, P.- Wickens, M. (eds.): The Handbook of Applied Econometrics. Vol. 1, Oxford:Basil Blackwell.
[48] Maravall, A. (1999.) *Short-Term Analysis of Macroeconomic Time Series*. In: Kirman, A.P. – Gérard – Varet, L-A. (eds.):Economics Beyond the Millennium. Oxford: Oxford Univercity Press.
[49] OECD (2002). *Harmonising Seasonal Adjustment Methods in European Union and OECD Countries.* Short-term Economic Statistics Expert Group – Meeting, 24-25 June, Paris.
[50] [web: http:///www.oecd.org/dataoecd/1/9/1933606.doc. Utolsó megtekintés:2005. július 12.]

[51] Planas, C.(1997). *Applied Time Series Analysis: Modelling, Forecasting, Unobserved Components Analysis and the Wiener – Kolmogorov Filter*. Luxembourg.
[52] [web: http://forum.europa.eu.int/Public/irc/dsis/eurosam/library?L=/documents_methodological&vm=detailed&sb=Title. Ulolsó megtekintés:2005 Július 12.]
[53] Sugár A.: *Szezonális kisimító eljárások összehasonlítása*. Gazdasági Minisztérium, Gazdaságelemző Intézet, (1999a).
[54] Sugár A.: *Szezonális kiigazítási eljárások (I.)*. Statisztikai Szemle, 77.évf.9.szám, szeptember, 705-721.p., (1999b)
[55] Victor Gómez and Agustín Maravall: *Demetra 2.0 User Manual*, Release Version 2.0 (Service Pack 1), May 2002
[56] Bauer Péter, Földesi Erika, Berki Natália, Fábián László: *Szezonális kiigazítás*. Statisztikai módszertani füzetek, KSH, Budapest, 2005

Köszönetnyilvánítás

A diplomamunka elkészítése egy időigényes és sok fáradságot követelő feladat.
Szakmai és formai kivitelezésében, az elkészítéséhez nyújtott erkölcsi támogatásban többen is segítségemre voltak.
Elsősorban szeretnék köszönetet mondani témavezetőmnek Dr. Ispány Márton egyetemi docens Úrnak, aki segített a munkám szakmai tökéletesítésében. Továbbá köszönöm családomnak, akik segítettek a dolgozat formai kivitelezésében, tolerálták az erre fordított időmet, türelmesek voltak velem szemben és biztosították a megfelelő körülményeket a munkához.