Idősorelemzési módszerek gyakorlati alkalmazásának előkészítése

Edelényi Márton[1], Pödör Zoltán[2], Jereb László[1]
[1] Nyugat-magyarországi Egyetem, Faipari Mérnöki Kar
[1] Nyugat-magyarországi Egyetem, Erdőmérnöki Kar

Döntéseink sokszor hosszú távon meghatározhatják egy-egy gyakorlati probléma, terület további kezelését, alakulását. Rendkívül fontos, hogy ezek a döntések megalapozottak, tudományosan is alátámasztottak, és így igazolhatóak legyenek. Igaz ez a klímaváltozás által az ember és az ökológia szempontjából gyakorolt hatás vizsgálata során, de igaz más faipari, erdészeti, vagy gazdasági szempontból is.

A megfelelő döntések meghozatalának alapja a már rendelkezésre álló adatok elemzése és ezen eredmények felhasználása a jövőt meghatározó döntések kialakításában. Manapság számos cikk, kutatás keres a mért paraméterek között összefüggéseket, függvényesíthető kapcsolatokat. Jelen cikkünkben egy olyan módszer formalizált változatát mutatjuk be, amellyel ilyen jellegű vizsgálatokat támogathatunk azáltal, hogy a szokásos összefüggés elemzéshez képest mélyebb és teljesebb összefüggésekhez is juthatunk.

A cikkünkben nem célunk a vizsgálathoz szükséges minden lépés, említett módszer (adatok előkészítése, korrelációszámítás, szignifikancia vizsgálat) teljes körű tárgyalása.

IDŐSOROK ELŐKÉSZÍTÉSE

Az elemzések megkezdése előtt egy rendkívül fontos lépés az adatok megfelelő előkészítése, ami a következő elemeket foglalhatja magában:

- adathiányok kezelése, pl. hiányzó elem időbeli környezetében mért adatok alapján, egyéb, közel elhelyezkedő mérési hely adatainak felhasználása,
- esetleges adathibák felderítése, javítása,
- a különböző vagy akár egyetlen paraméterre vonatkozó eltérő mérési gyakoriságú adatok kezelése.

Mindezek gondot jelenthetnek, hiszen az idősorok elemzése során lényeges azok időbelisége, a mérések eredményeként adódó idősorok hosszának egyezése és teljessége.

A nyers mérési adatokból egymással összevethető idősorokat előállítani nem minden esetben triviális feladat. Sok esetben nem a teljes idősorokat akarjuk egymással összevetni, hanem adott időegységre lebontott paraméterek közötti kapcsolatokat szeretnénk elemezni. Így egyrészt meg kell határozni egy időegységet (nap, hét, hónap, stb.), amelyre nézve képezzük idősorainkat, illetve képezni kell a paraméterek adott időegységre vonatkozó értékeit. Az előállítás sokszor egyszerű. Például hőmérsékleti adatok esetén képezzük a havi mérési adatok átlagát, vagy összegét a feladattól függően. Vannak esetek, amikor ez nem ennyire egyértelmű. Például, ha csapadékot mérünk, akkor egy rögzített esemény kezdődhet adott hónap végén és befejeződhet a következő hónap elején. Azaz, a mért csapadék mennyiséget valamilyen módon el kell osztani a két hónap között.

KORRELÁCIÓANALÍZIS ÉS A SZIGNIFIKANCIA TESZT

Az idősorok elemzésének egyik leggyakrabban használt eszköze a korreláció-analízis, amely alapvetően két adatsor közti lineáris kapcsolat erősségét vizsgálja:

$$\mathrm{r} = \mathrm{korrel}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

A számított korrelációs érték gyakorlatilag azt mutatja meg, hogy a két változó vonatkozásában felírt regressziós egyenesre mennyire jól illeszkednek az adatsorok pontpárjai. Ahhoz, hogy a kapott eredmény szignifikanciáját ellenőrizzük, vizsgálatot kell végezni egy $m - 2$ szabadságfokú t-statisztika felhasználásával:

$$\mathrm{t} = \mathrm{r}\sqrt{\frac{\mathrm{m} - 2}{1 - \mathrm{r}^2}}$$

Abban az esetben, ha a kapott t érték nagyobb, mint a t-tábla megfelelő szignifikanciájú eleme, akkor a kapott eredményt valós lineáris kapcsolatnak minősítjük.

Ez a Pearson-féle korrelációszámítás, melynek eredményét nagyban ronthatják a kiugró adatok, hiszen azok távol esnek regressziós egyenestől. Részben ennek kiküszöbölésére szolgál a rangkorreláció, amely alapvetően nem azt vizsgálja, hogy a két adatsor pontpárjai mennyire illeszkednek a regressziós egyenesre, hanem azt, hogy a két adatsor mennyire változik együtt.

Rangkorreláció mérésére használhatjuk Spearman és Kendall módszerét is. A Spearman-féle rang korreláció a kapcsolat szorosságának mérésére a két változó rangszámainak különbségét használja fel, azaz a két adatsor összetartozó adatpárjainak rangjai közti különbséget:

$$r_{rang} = korrel_{rang}(x, y) = \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2 - 1)}$$

A Pearson-féle korrelációs együtthatóhoz hasonlóan itt is végezhetünk szignifikancia tesztet.

Sok esetben azonban fontos lehet, hogy két adatsor közti késleltetett hatásokat is vizsgáljuk, azaz az egyik adatsor elemeit időben eltoljuk a másikhoz képest és így vizsgáljuk a köztük fennálló korrelációt. Illetve egy adott adatsor és ugyanazon adatsor időbeni eltoltja közti kapcsolat vizsgálata, aminek segítségével például a periodicitás kérdése vizsgálható. Előbbi mérésére a kereszt-, utóbbira az autokorreláció szolgál.

Jelölje k az adatsor eltolásának mértékét, ahol $0 \leq k \leq adatsor\ hossza - 2$. Amennyiben egy adott k érték választása mellet a korrelációs együttható értéke 1-hez (vagy -1-hez) közelít, akkor ez azt mutatja, hogy időben egy k egység késleltetésű hatás jelenik meg a függő adatsorban a független változóhoz képest.

FORMALIZÁLÁS

Tegyük fel, hogy kiindulásként adott két vektor, melyek nem feltétlenül azonos hosszúságúak:

$$\bar{\mathrm{u}} = \begin{pmatrix} \mathrm{u}_1 \\ \vdots \\ \mathrm{u}_\mathrm{p} \end{pmatrix} \text{ és } \bar{\mathrm{v}} = \begin{pmatrix} \mathrm{v}_1 \\ \vdots \\ \mathrm{v}_\mathrm{q} \end{pmatrix}.$$

A korábban már említett, feladat megoldásához illeszkedő, előkészítő lépések elvégzése után a feladatnak megfelelő periodicitású (napi, heti, havi, éves, stb.) adatsorokat képzünk ezekből. Így adódnak az

$$\bar{y} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} \text{ és } \bar{x} = \begin{pmatrix} x_0 \\ \vdots \\ x_n \end{pmatrix}$$

adatsorok. Az egyszerűsítés kedvéért éljünk azzal a feltételezéssel, hogy az adatsor a választott ciklus első elemével kezdődik és az utolsó elemével ér véget (pl. január, december) és így $c|(n+1)$, azaz a vektor dimenziója osztható az egyes ciklusok elemszámával (c-vel). Amennyiben ez nem teljesül, akkor rövidítsük le az adatsort, hogy megfeleljen ennek a kritériumnak. Ez legfeljebb $2c-2$ rekord veszteséget jelenthet. Célunk az $\bar{x}$ és $\bar{y}$ adatsorok közötti kapcsolatok feltérképezése.

Bemutatunk egy módszert, amely egyrészt alkalmas a korábban ismertetett korreláció-elemzések megvalósítására, az $\bar{x}$ és $\bar{y}$ eredeti komponensei közti lineáris-, auto- és keresztkorreláció számítására. Másrészt ennél mélyebb összefüggéseket is képesek lehetünk felfedni, mivel kereshetjük azt is, hogy az alapállapothoz képest adott távolsággal eltolt és adott szélességű időablakba (pl. január-február-március havi adatösszeg) eső $\bar{y}$ vektorkomponensekből valamilyen matematikai művelet alkalmazásával képzett új vektorkomponensek milyen kapcsolatot mutatnak az $\bar{x}$ vektorból megfelelő módon képzett $\bar{x}'$ vektorral. A matematikai műveletre – az alkalmazott feladat jellegéből adódóan – az összegzés mellett további példa lehet az a szorzat, átlag, minimum/maximum érték. A művelet megvalósítható az adott feladatban értelmes módon definiálható, összes lehetséges eltolásra és szélességre. Cikkünkben összegzést és lineáris korrelációszámítást fogunk alkalmazni.

Módszerünk bemeneti paraméterei $\bar{x}$ és $\bar{y}$ azonos időbélyeggel ellátott vektorok, valamint I, J és c egész számok, illetve egy α valós szám. A vektorok komponenseinek számozását az utolsó időbélyegtől tekintjük, azaz 0 értékkel jelöljük az időben utolsó elem, n-el az időben elem indexét. Az I szám fogja jelölni az eltolás maximális értékét, azaz azt, hogy meddig mehetünk vissza időben a kiinduló ponthoz képest. Ennek megfelelően I a 0 értéktől értelmezhető, maximális értéke a problémától függ, illetve úgy kell megválasztani, hogy a korrelációszámítás elvégezhető legyen. A J paraméter az ablak maximális szélességét jelöli, maximális értéke n és I függvénye. A c szám az általunk definiált ciklushossz (pl. 12, amennyiben havi adatokról van szó), azaz azt jelöli, hogy a vektorban egy ciklus hány komponenst jelent. Az α egy szignifikancia szintet jelöl.

Az eljárás a kapott paraméterek függvényében az $\bar{x}$ és $\bar{y}$ vektorokból származtat megfelelő $\bar{x}'$ és $\bar{y}'$ vektorokat. A származtatások számát – mint korábban említettük – az I és J értékek határozzák meg. A kapott $\bar{x}'$ és $\bar{y}'$ vektorokra számítjuk a korreláció mértékét.

Legyen i és j két ciklusváltozó, amelyekre igaz, hogy $0 \leq i \leq I-1$ és $1 \leq j \leq J$. Adott i és j értékek esetén jelölje KSZ az aktuálisan képzett vektorokra a komponensek számát. Az egyes ciklusléptékben értékét a következő formula szolgáltatja:

$$KSZ = \frac{n+1}{c} - \left\lfloor \frac{i+j}{c} \right\rfloor + \left\lfloor \frac{i}{c} \right\rfloor.$$

Így adott i-re az $\bar{x}'_i$ KSZ komponensű származtatott oszlopvektor meghatározása az $\bar{x}$ vektor komponenseiből a következő szerint adódik:

$$\bar{x}'_i = \begin{pmatrix} x_i \\ x_{i+c} \\ x_{i+2c} \\ \vdots \\ x_{i+c(KSZ-1)} \end{pmatrix}$$

Illetve az $\bar{y}'_{i,j}$ oszlopvektor komponensei a következő képlet szerint adottak:

$$\bar{y}'_{i,j} = \begin{pmatrix} \sum_{l=i+c*0}^{i+j-1+c*0} y_l \\ \sum_{l=i+c*1}^{i+j-1+c*1} y_l \\ \vdots \\ \sum_{l=i+c*(KSZ-2)}^{i+j-1+c*(KSZ-2)} y_l \\ \sum_{l=i+c*(KSZ-1)}^{i+j-1+c*(KSZ-1)} y_l \end{pmatrix}$$

Az így előállított $\bar{x}'_i$ és $\bar{y}'_{i,j}$ származtatott vektorokra számítjuk ki a korreláció mértékét, jelölje ezt $r_{i,j} = \mathrm{korrel}(\bar{x}'_i, \bar{y}'_{i,j})$.

Módszerünk lényege, hogy az összes lehetséges i (eltolás maximális mértéke) és j (az eltolt ablak maximális szélessége) értékek esetén kiszámítjuk a korreláltság mértékét. Ez $i * j$ darab értékét fog jelenteni, melyeket mátrixokban tárolunk. Mivel a korreláció szignifikanciáját is vizsgáljuk így az eljárás kimenete c db, $i * j$ méretű mátrixpár. A mátrixpárok alakjai a következők:

$$R = \begin{pmatrix} r_{0,1} & r_{0,2} & \cdots & r_{0,j} \\ r_{1,1} & r_{1,2} & \cdots & r_{1,j} \\ \vdots & \vdots & \ddots & \vdots \\ r_{i-1,1} & r_{i-1,2} & \cdots & r_{i-1,j} \end{pmatrix}, \qquad S = \begin{pmatrix} s_{0,1} & s_{0,2} & \cdots & s_{0,j} \\ s_{1,1} & s_{1,2} & \cdots & s_{1,j} \\ \vdots & \vdots & \ddots & \vdots \\ s_{i-1,1} & s_{i-1,2} & \cdots & s_{i-1,j} \end{pmatrix}$$

A párok közül az egyik mátrix (R) értékei megfelelnek az egyes ciklusrészekre adott eltolással (i) és ablakszélességgel (j), $\bar{x}'_i$ és $\bar{y}'_{i,j}$ vektorokra kapott korrelációs értékekkel. A másik (S) egy logikai mátrix, ami az α paramétertől függő szignifikancia vizsgálat eredményét tartalmazza, ahol 1-es érték jelöli azt, hogy a korrelációs érték szignifikáns, 0 azt, hogy nem.

ÖSSZEGZÉS

Úgy gondoljuk, hogy a megalapozott döntések meghozatalának az egyik legfontosabb alapfeltétele az adott terület megismerése, összefüggéseinek felfedése, mely az adatok megfelelő előkészítését követő elemzésén alapul.

Az általunk kidolgozott eljárás mindenképpen alkalmas a paraméterek közti kapcsolatok felfedésére, késleltetett hatások vizsgálatára. Az időablakos technikának köszönhetően olyan összefüggések is kinyerhetőek, melyeket eddig egyáltalán nem, vagy csak nem a maguk teljességében vizsgáltak (minden lehetséges szélességgel, minden lehetséges eltolással). A módszer minden olyan területen felhasználható, ahol két adat/idősor közötti kapcsolat feltárására vagy a függvényesíthető kapcsolat felírására van szükség, sőt a származtatott vektorok általánosításával még összetettebb kapcsolatok vizsgálatát is lehetővé teszi.