KARSZTFEJLŐDÉS XI.
Szombathely, 2006. pp. 15-24.

A FÖLDTANI ADATOK ADATELEMZÉSÉNEK NEHÉZSÉGEI

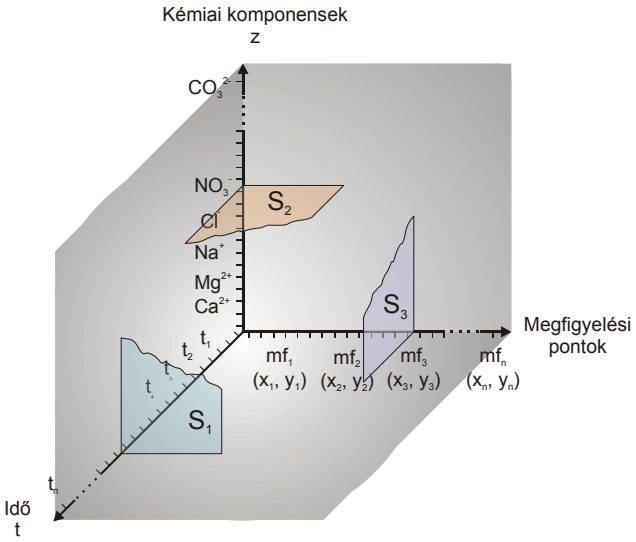
KOVÁCS JÓZSEF[1] – KOVÁCSNÉ SZÉKELY ILONA[2]

[1]ELTE, Földrajz- és Földtudományi intézet, 1116 Budapest PázmányPéter sétány1/C, kevesolt@geology.elte.hu
[2]BGF, KVIFK, Módszertani Intézet, 1054 Budapest, Alkotmány u. 9-11, iszekely@geology.elte.hu

Abstract: Using the results of a geological analysis and with the help of a four-dimensional modell we tried to show the notion of the spatiotemporal sample and some of its basic characteristics. On the basis of these considerations we give the definition of the spatiotemporal sample in order to be satisfactory from both the theoretical and the practical points of view. We propose the following definition: In practical sense the values of a parameter of certain phenomenon that can be associated with x,y,z,t coordinates and either measured in situ, analysed or computed are called a spatiotemporal sample. The sample in the practical sense corresponds to one element of the mathematical sample, with the difference that it is associated with a space-time unit with a volume bigger than zero.

Bevezetés

A földtani, hidrogeológiai és környezetföldtani gyakorlatban is egyre inkább jellemző az olyan mennyiségű mérési eredmény, adat megjelenése, ami nehezen átlátható. Matematikai eszköztár alkalmazása nélkülözhetetlen és ilyen módon a feldolgozás a szubjektív értékelési módszerekről az objektívek felé tolódik el. A vizsgálati módszerek között egyre inkább jut jelentős szerephez az adatfeldolgozás és vele párhuzamosan a sztochasztikus szemlélet. Napjainkban már a determinisztikus egyenletekkel leírható hidrogeológia tér modellezésében jártas szakemberek között is többen vallják: „*A jövő mindenképpen a sztochasztikus modellezésé, a kérdés, hogy megtaláljuk-e a bizonytalanságok közvetlen jogi kezelésének módját, vagy megkapják-e a földtani és vízföldtani szakemberek azt a lehetőséget, hogy valószínűségelméleti alapon számított eredmények értékelésével a hatályos jogszabályok szellemében járjanak el*” (*KOVÁCS – SZANYI* 2005). Ahhoz azonban, hogy a földtudományok művelői sikerrel vegyék fel alkalmazható eszköztárukba a geomatematika által nyújtott lehetőségeket, kellő áttekintéssel kell rendelkezniük azokról. Ennek azonban van néhány feltétele. Elsősorban meg kell fogalmazni milyen adathalmaz áll rendelkezésére a földtudományi szakmának és mikor milyen eszköz alkalmazható az adatelemzés módszereiből. Fontos szempont az is, hogy a szakmai köztudatba beépüljön az a tény, miszerint a mintából számított statisztikák valószínűségi változók.



1.ábra: A három dimenziós adathalmaz
Fig. 1: The three-dimensional mass of data

Milyen adathalmazból kell dolgoznunk?

Egy földtani folyamatot leggyakrabban egy időpontban az állapotjellemzők megadásával írunk le. Ha a folyamat állapotának változásait is követni kívánjuk, idősorokkal van dolgunk. Tekintsük át egy kicsit részletesebben a fentieket és induljunk ki az *1. ábra* S_1 síkjából. Ekkor, mint a földtudományokban előforduló megoldandó feladatok jelentős részében, különböző térbeli pontokon mért paraméterek vizsgálatára van szükség. Tekintsünk meg erre az esetre egy példát. A Bakonyban egy napon (matematikai értelemben azonos időpontban) a főkarsztra szűrőzött megfigyelő kutakból vízmintákat veszünk és azokat kémiai analízissel több kémiai komponensre megvizsgáljuk. Az így kapott eredményeket táblázatban rögzítjük, olyan módon, hogy az egyes oszlopok egy-egy kémiai paraméternek, míg az egyes sorok egy-egy karsztvízmegfigyelő kútnak felelnek meg. A vizsgálatokhoz – az egyváltozós statisztikai elemzésen túl – a sokváltozós adatelemző módszerek adnak lehetőségeket. A számtalan eszköz közül megemlítjük a leggyakrabban használatos klaszter-, diszkriminancia-, faktor- és főkomponens analízist, valamint a sokdimenziós skálázást. Ezeket a módszereket alkalmazhatjuk megfigyelési pontjainkra, amikor azok közötti kapcsolati viszonyok feltárására van szükség. Ilyen feladat például, ha arra vagyunk kíváncsiak, mely mintavételi pontok kémiai karaktere hasonlít legjobban egymás-

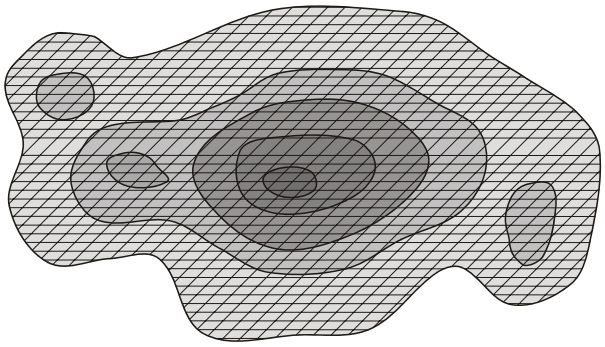
hoz. De lehet cél a változók (kémiai komponensek) magyarázata is. Ekkor ugyanazt a módszert használjuk csak az adatmátrixnak nem a soraira, hanem az oszlopaira végzünk vizsgálatokat.

Adataink nagyon gyakran tartalmazzák a 3. dimenziót, az időt. Abban az esetben, ha (az előbbi példánál maradva) több karsztvíz megfigyelő kútban, több időpontban – legjobb, ha azonos időközönként – mérünk egy paramétert, például a vízszintet, az S_2 síkban vagyunk. Ebben az esetben, ha arra vagyunk kíváncsiak milyen háttértényezők befolyásolják a vízszintek időbeli fluktuációját, akkor dinamikus faktoranalízist alkalmazunk. Alkalmazása az utóbbi években kezdődött el, eddig elért és a várható eredmények jelentős segítséget adnak a környezetvédelemnek (*KOVÁCS* et al., 2004). Ha egy rögzített megfigyelési pontban, egyetlen karsztvízmegfigyelő kútban, több paraméter (például kémiai komponens) időbeli változásait figyeljük meg, akkor az S_3 síkban dolgozunk. Ilyenkor gyakori alkalmazás a „*klasszikus idősoros*" vizsgálat, ami magában foglalja az egyes paraméterek tartós irányzatának – trend – és periodikus viselkedésének meghatározását, ami nagyon gyakran felmerülő igény. Egy folyamatban megállapított periódust és trendet gyakran használnak fel előrejelzésre. A jövőre nézve azonban csak akkor vonhatóak le következtetések, ha bizonyosak vagyunk abban, hogy az idősort alakító hatások a jövőben is fent fognak maradni.

A földtudományokban gyakran fontos az egyes pontok térbeli helyzete, amelyeket a megfigyelési pontok tengelyén x_i és y_i koordinátákkal jelöltünk. Ha vizsgálatainknál figyelembe kell vennünk, a mintavételi pontok térbeli elhelyezkedését is, akkor a geostatisztika eszköztára segíthet bizonyos problémák megoldásában. Ilyen eset például egy paraméter térképének elkészítése egy adott időpillanatban. Példaként szolgálnak a VITUKI gondozásában megjelent, a Dunántúli-középhegység karsztvízszint térképei (*LORBERER* 1978-2001). Ha egy paraméter időbeli változásait kívánjuk vizsgálni, mint például a Dunántúli-középhegységben a karsztvízszint felszín időbeli változását, a megfigyelési pontok térbeli struktúrájának figyelembe vétele mellett, a geostatisztikának kevés eszköz áll rendelkezésére. Néhány módszer esetében született már megoldás. Ilyen az empirikus variogram függvény háromdimenziós esetben (*DRYDEN* et al. 2005).

Az adatelemző módszerektől csak akkor várhatunk jó eredményt, ha az adott folyamat vizsgálatára olyan jellemzőket veszünk figyelembe, amelyek kellően pontosan írják le azt. Ennek eldöntése mindig az adott szaktudomány feladata és felelőssége. Lényeges követelmény az adatelemző módszerek szempontjából, az alapvető statisztikai stabilitás miatt, $n>>m$, vagyis a mintavételi pontok száma jelentősen nagyobb legyen, mint a vizsgált paramétereké.

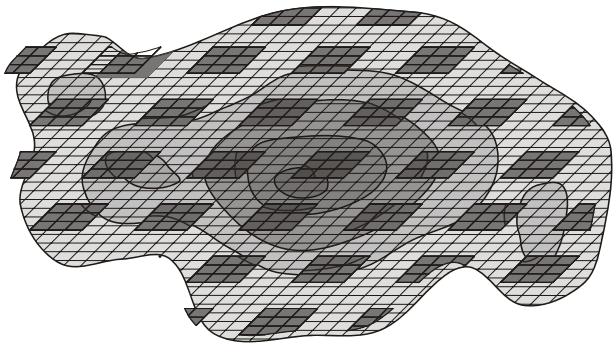
Gyakorlati alkalmazásoknál, több esetben problémát jelent a minta térfogata. Belátható, hogy valamely időben és térben változó természeti jelenség adott időponthoz rendelhető háromdimenziós metszete, elméletileg végtelen számú ($N=\infty$), „*nulla térfogat*”-ú ($V=0$) elemi részre osztható. Ebből következően egy jelenség kutatása során valamely földtani paraméter vonatkozásában egyetlen olyan minta realizáció állítható elő, melynek elemszáma végtelen. Elméletileg ez az adathalmaz tekinthető a vizsgált sokaságnak. A végtelen elemszámú sokaság szemléltetéséhez tekintsük meg a *2. ábrát*.



2. ábra: Egy elméleti jelenség adott paraméterének izovonalas képe. A jelenség „terület”-ét végtelen számú „nulla térfogat”-ú elemi részre osztottuk fel.
Fig. 2: Isoline figure of a “parameter” of a theoretical phenomenon

Valamely jelenség kutatása során – főként anyagi okok miatt – meglehetősen ritkán adódik arra lehetőség, hogy olyan (esetenként több ezer) elemszámú mintát vegyünk, amelyből a statisztikai jellemzők nagy pontossággal számíthatók.

A gyakorlatban egy másik problémával is szemben találjuk magunkat. Ez pedig az, hogy a minta elemi részeinek „*térfogat*”-a nem nulla, hanem nullánál nagyobb és mindenképpen mérhető nagyságú. A helyzet az, hogy $V>>0$, de a jelenség egészéhez képest $V\approx 0$, azaz a minta elemek térfogata és a szórás közötti kapcsolatot figyelmen kívül lehet hagyni, de $N<<\infty$. Ekkor felmerül a kérdés, hogy vajon az ilyen minta reprezentatívnak tekinthető-e, azaz a minta valóban híven tükrözi a sokaságot, amelyből származik. A *3. ábra* egy lehetséges minta realizációt mutat $N<\infty$ és $V>0$ esetére. Az esetek zömében nincs arra lehetőség, hogy közel végtelen elemszámú mintát vegyünk, vagy a mintavételt kisebb elemszám mellett többször megismételjük.



3. ábra: Egy lehetséges minta realizáció, $N<\infty$ és $V>0$ esetén
Fig. 3: A possible sample realisation for $N<\infty$ and $V>0$

A minta fogalma és néhány tulajdonsága egy példa tükrében

A statisztikai minta az X valószínűségi változóra vonatkozó véges számú független megfigyelés eredménye $X=(X_1,X_2, \ldots ,X_n)$, ahol $X_1,X_2, \ldots ,X_n$ egymástól független, azonos eloszlású valószínűségi változók. A minta elemeinek eloszlása megegyezik a sokság eloszlásával és a mintaelemek várható értéke $E(X_i)=m$ és szórása $D(X_i)=\sigma$ ahol $i=1, 2, \ldots n$.
A minta realizációja a megfigyelések számszerűsített értékei $(x_1,x_2,\ldots,x_n)$, ha egy konkrét mintavételnél $X_1=x_1$, $X_2=x_2, \ldots , X_n=x_n$ adódik.
Ez a megfogalmazás nehezen érthető, különösen nem matematikusok számára. Induljunk ki ezért egy másfajta megközelítésből.

A gyakorlatban szinte megoldhatatlan anyagi nehézségekbe ütközik olyan nagy elemszámú minta előállítása, ami be tudja tölteni a vizsgált sokaság szerepét. Ezért a minta fogalmának tisztázására egyfajta közelítést alkalmaztunk. A Jósva-patak vizéből (*MAUCHA* 1998) 1981.01.05 és 1983.12.26 között 200 alkalommal vett vízminta 16 vízkémiai paraméterét vizsgálták meg, mérték továbbá a patak vízhozamát és hőmérsékletét. Ezekből három paramétert választottunk ki: vezetőképességet, pH-t, kalciumot, amelyek a statisztikai modellben valószínűségi változók.

A mérési eredményeket rendre statisztikai sokaság elemeinek tekintjük. Jelen esetben az elemszám 197, 167 és 189 volt. Ezt elég nagynak tartjuk ahhoz, hogy sokaságként tekintsük és felhasználjuk a tárgyalt statisztikai fogalmak szemléltetésére anélkül, hogy a matematikai elmélet követelményei jelentősen sérülnének.

Mivel a valóságban véges sokaságot kaptunk meghatározhatóvá váltak a valószínűségi változók várható értékei és szórásai. Ezt az *I. táblázat* mutatja be.

I. táblázat
Table I.

A sokaság paraméterei
The parameters of the manifold

Változók	*Mintaszám [db]*	*Átlag [mg/kg]*	*Szórás [mg/kg]*	*Relatív szórás*
Vezetőképesség	197	530,4	37,85	0,07
Ph	167	7	0,14	0,02
Ca	189	111	9,51	0,09

A vezetőképesség, pH és kalcium sokaságaiból annak bemutatására, hogy a minta elemei valószínűségi változók, véletlenszerűen 10, 30 és 100 elemű mintát vettünk, 1000 - szer. A *II. táblázat* vezetőképességre vonatkozó minta realizációiból mutat be részleteket. Jól követhető, hogy a 30 elemű minták realizációi mintáról mintára változnak.

II. táblázat
Table II.

A minták realizációi
Realisations of conductivity–samples

Vezetőképesség	*Minta realizáció*						
	X_1	X_2	X_3	X_4	…	X_{29}	X_{30}
1. minta	581	521	531	508	…	560	566
2. minta	465	564	526	543	…	516	576
.							
.							
.							
1000. minta	511	542	478	516	…	558	507

A *II. táblázat* mintáiból alapstatisztikák számíthatók, amelyek közül az egyik legfontosabb, a mintaátlag kerül bemutatásra a *III. táblázatban*. A táblázatot két további valószínűségi változóval bővítettük, a pH-val és a kalciummal. A táblázat adatai szemléletesen láttatják azt az állítást, hogy a mintaátlag is valószínűségi változó, mintáról mintára változik és értékei szóródnak a sokasági átlag körül (*I. táblázat*).

III. táblázat
Table III.

A mintaátlagok realizációi
Realisations of the sample means

Minták	*Mintaátlag*		
	Vezetőképesség	*pH*	*Ca*
1. minta	538,63	7,163	110,57
2. minta	526,73	7,161	108,90
.			
.			
.			
1000. minta	542,26	7,141	111,47

IV. táblázat
Table IV.

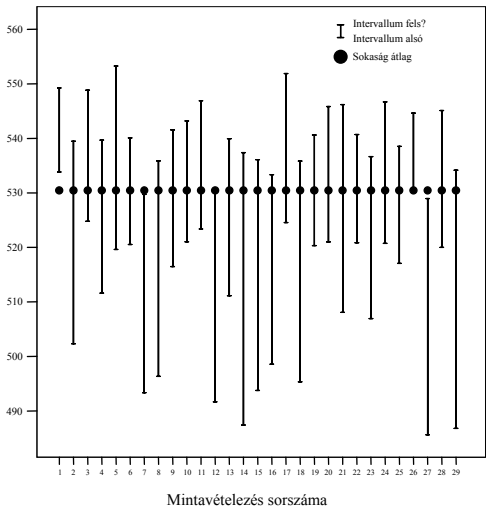
A mintaátlagok átlagai és standard hibái
Averages and standard errors of the sample means

Valószínűségi változó-mintarealizáció	*Mitavételezés száma*	*Átlagok átlaga*	*Átlagok standard hibája*
Vezetőképesség-10	1000	530,85	11,79
Vezetőképesség-30	1000	530,42	5,3
Vezetőképesség-100	1000	530,35	3,8
pH-10	1000	7,158	0,047
pH-30	1000	7,159	0,021
pH-100	1000	7,160	0,014
Ca-10	1000	110,87	2,940
Ca-30	1000	110,77	1,327
Ca-100	1000	110,81	0,912

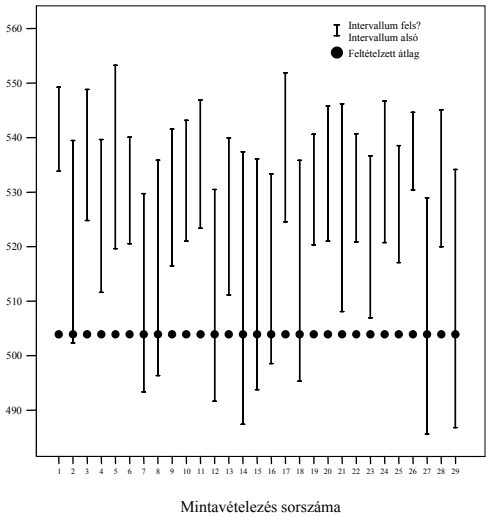
Természetesen az összes lehetséges mintaátlag átlaga adja a sokasági átlagot, azaz a várható étéket (m). Ez a tulajdonság a becslés torzítatlanságát jelenti: $E(\overline{X}) = m$

Ezt az elméleti megállapítást csak akkor lehetne bemutatni, ha az összes lehetséges mintaátlagot figyelembe vennénk. Ez azonban nehézségekbe ütközik, mivel például a vezetőképesség esetében az általunk sokaságnak tekintett 197 mintaelemből, $1{,}90*10^{47}$ módon lehet 30 elemű mintát kiválasztani. Más szavakkal: ennyi féle 30 elemű minta realizációt vagyunk képesek ebből a sokaságból előállítani és következésképpen ennyi különböző átlagot (100 elemű minta kiválasztására $1{,}11*10^{58}$ lehetőség van.). Ennek teljesítése gyakorlatilag lehetetlen. Ezért csak annak bemutatására lehet vállalkozni, hogy példánkon mutassuk be: a mintaátlag jól közelíti a sokasági

átlagot és hibája csökken a minta elemszámának növelésével. Ezt a következőképpen valósítottuk meg. A sokaságokból 10, 30, 100 elemű mintákat vettünk, szintén 1000-szer. Kiszámítottuk a mintaátlagok átlagát és rendre az átlagok hibáit Az eredményekből néhányat a *IV. táblázat* tartalmaz. Az adatok a gyakorlatban is meggyőznek a fenti állításunkról.



4. ábra: Konfidencia intervallum és elsőfajú hiba
Fig. 4: Confidence intervall and type I error



5. ábra: Konfidencia intervallum és másodfajú hiba
Fig. 5: Confidence intervall and type II error

A *III. táblázatban* megadtuk az összes mért érték átlagát. Tekintsük ezt a változók várható értékének. A gyakorlatban a várható érték nem ismert. Ha erről az értékről van egy sejtésünk, ezt az úgynevezett t-statisztikai próbával tesztelhetjük. Feltételezzük, hogy a vezetőképesség paraméter normális eloszlású. Esetünkben a sokaság átlaga 530,4 (μS/cm) volt. Arról döntünk, hogy ez az érték elfogadható-e a sokaság átlagának. 1000 véletlenszerűen kiválasztott 30 elemű minta alapján 95%-os megbízhatósági szintű konfidencia intervallumokat konstruálunk a sokaság átlagára. Ez azt jelenti, hogy az esetek 95%-ban tartalmazzák ezt. A *4. ábra* bemutatja a sokaság átlagát és 30 konfidencia intervallum véletlen elhelyezkedését, továbbá szemlélteti, hogy néhány intervallum nem tartalmazza a sokasági átlagot. Szoros kapcsolat van a hipotézis vizsgálat és a konfidencia intervallum között. Mi döntésünket a gyakorlatban mindig csak egy minta alapján hozzuk meg. Ha például az ábrán levő egyes sorszámú minta alapján döntünk, akkor elutasítjuk azt a feltevést, hogy a sokasági átlag 530,4(μS/cm), holott ez igaz. Ebben az esetben elsőfajú hibát követünk el.

A valódi átlagot módosítottuk 5%-al. Így arról szeretnénk dönteni, hogy az átlagos vezetőképesség 503,9 (μS/cm) elfogadható-e sokasági átlagnak. Az *5. ábra* szemlélteti, hogy több intervallum tartalmazza ezt az értéket. Ha ezen intervallumok egyike alapján hozzuk meg döntésünket, akkor előfordulhat, hogy elfogadjuk sokasági átlagként az 503,9(μS/cm) értéket, ebben az esetben másodfajú hibát követünk el, mert elfogadunk egy olyan feltevést, ami nem igaz (*DÉVÉNYI – GULYÁS* 1988).

Összefoglalás

Az adatelemző módszerek alkalmazási lehetőségeit egy „*négydimenziós*” modellen mutattuk be. Egy karsztos terület patakjának elemzési eredményeinek felhasználásával szemléltettük a mintát, annak néhány tulajdonságát. Definiáljuk a mintát úgy, hogy ez a meghatározás mind elméleti, mind gyakorlati szempontból kielégítő legyen. A következő meghatározást javasoljuk. A gyakorlati élet mintának nevezi valamely vizsgált jelenség adott paraméterének x,y,z,t koordinátákhoz, vagy azok intervallumához köthető, in situ mért, elemzett, vagy az előbbiekből számított értékét. A gyakorlati értelemben vett minta, a matematikai minta egy elemének felel meg, azzal a különbséggel, hogy vonatkozatási térfogata nagyobb, mint nulla.

IRODALOM

DÉVÉNYI D. – GULYÁS O. (1988): Matematikai statisztikai módszerek a meteorológiában, - Tankönyvkiadó, Budapest

DRYDEN, I. L. – MÁRKUS L. – TAYLOR C. C. – KOVÁCS J.(2005): Non-stacionary spatio-temporal analysis of karst water levels, - Applied Statics Vol. 54, Part 3, p. 1-18.

J. KOVÁCS – L. MÁRKUS - G. HALUPKA (2004): Dynamic Factor Analysis for Quantifying Aquifer Vulnerability, - Acta Geol. Hung. Vol. 47, p. 1-17.

KOVÁCS B. – SZANYI J. (2005): Hidrodinamikai és transzportmodellezés II., - Miskolci Egyetem Szegedi Egyetem és Gáma–Geo Kft, Szeged

MAUCHA L. (1998): Az Aggteleki-hegység karszthidrológiai kutatási eredményei és zavartalan hidrológiai adatsorai (1958-1993), - A VITUKI Rt. Hidrológiai Intézete, Kézirat

LORBERER Á. (szerkesztő, 1978-2001): A Dunántúli-középhegység éves karsztvízszint-állapottérképe, M= 1:200 000, - VITUKI, Budapest