NÉHÁNY STATISZTIKAI MÓDSZER A METEOROLÓGIÁBAN

Egyetemi doktori értekezés

Faragó Tibor

1977

Tartalomjegyzék

BEVEZETŐ

Alig 50 éve, hogy egy "álmodozó" meteorológus, Richardson felvetette egy "emberi számológépekből" álló "időjárás-gyár" gondolatát, melyben tobb tizezer ember alkotta volna a mai számitógépek aritmetikáját és operativ memóriáját. Alig 30 év telt el az első számitógép szerkesztése óta, melyet két meteorológussal, Charney-val és Fjörtofttal karöltve a magyar származásu von Neumann éppen egy időjárás-előrejelzési modellen próbált ki. Ma pedig már szinte elképzelhetetlennek tünik, hogy a meteorológiai kutatási és operativ tevékenységek dontő tobbsége ne éppen számitógépre tervezett programok illetve rendszrek formájában realizálódjon. Különosképpen igaz ez a megállapitás a staisztikai kutatásokra, melyek például nagyobb skáláju meteorológiai mezők elemzése esetén már meglehetősen memóriaigényesek.Éppen ezért az általánosabb statisztikai vizsgálatok mellett nem egy vizsgálatban fontos szerep jut a többé-kevésbé optimális adattárolásnak, valamint a fizikailag és statisztikailag is alátámasztott adatzsugoritási eljárásoknak.

A meteorológia területén a statisztikai módszerek felhasználásának két alapvető területét különithetjük el. Az első terület a klimatológia, melynek keretében megadott meteorológiai elemek vagy azok függvényei átlagos viselkedésére nézve kivánunk következtetéseket levonni. E téren gyakran találkozhatunk a szakirodalomban meteorológiai állapothatározók rövidebb-hosszabb távu átlagainak statisztikai-fizikai elemzésével, eloszlásbecslésekkel és -illesztésekkel, e becslések torzitottságának vizsgálatával, statisztikai jellegü osztálybasorolásokkal stb. Itt csak egyetlen összefoglaló münkára hivatkozunk ebből a témakörből, amely a különféle meteorológiai mezők statisztikai szerkezetéről ad áttekintést Czelnai,Gandin és Zaharjev szerkesztésében /1976/. A statisztikai prognosztika területe sokszor nem különithető el élesen a klimatológiai feldolgozások technikájától, a kutatások célja azonban alapvetően más. E téren az egyik legujabb eljárásként a klasszikus dinamikai /hidro- és termodinamikai/ modellekkel

szoros kapcsolatban álló sztochasztikus-dinamikus modelleket kell megemliteni. Freiberger és Grenander /1965/ ötletadó munkája nyomán több szerző foglalkozott a dinamikus rendszerek ilyen általánositásával, igy tobbek között Gleeson/1966/, Epstein/1969/, Tatarszkij/1969/, Fleming/1971/, Szonyecskin et al./1976/ és Faragó/1977/. Egy másik kutatási ág keretében a vizsgált fizikai folyamatot valamilyen speciális sztochasztikus folyamattal modellezik, melyet statisztikai eljárások segitségével körvonalaznak és utóbb paramétereznek. Többen is kisérleteket végeztek például a csapadékos és nem csapadékos napok sorozatának Markov-folyamattal való szimulálására, illetőleg más vonatkozásban a stacionárius és a Gauss-folyamatok elméletének alkalmazására. Ide sorolható a sztochasztikus automaták elméletének felhasználása is. Ez utóbbival kapcsolatban Kapovicsné-Maller-Titkos /1976/ dolgozatára utalunk. Korábban magam is foglalkoztam ennek az elméletnek prognosztikai alkalmazhatóságával /Faragó,1973,1974 /. Igen átfogó, elsősorban szovjet irodalommal rendelkezik a sztochasztikus lineáris elmélet meteorológiai tárgyu felhasználása. Ilyen jellegü eljárásnak tekinthetjük a meteorológus berkekben "természetes empirikus ortogonális sorfejtésnek"nevezett Karhunen-Loéve sorfejtés alkalmazását, a kanonikus korreláció becslését vagy a különféle regressziós modelleket /Bagrov,1966; Thom, 1966; WMO,1966; Gandin, 1967; Craddock-Colgate,1974 stb/. A regresszió tipusu extrapolációval már korábban foglalkoztunk általános és meteorológiai alkalmazási megközelitésben /Faragó-Gulyás,1974/. A teljesség igénye nélkül itt végül megemlitjük még a sürüségfüggvények illetve feltételes sürüségfüggvények közvetlen vagy közvetett becslését. Ezek vizsgálatának problémája kétirányu: az esetek tulnyomó többségben egyrészt nem áll rendelkezésre a jó hatásfoku becsléshez szükséges kellő menynyiségü adat, másrészt viszont bizonyos feladatoknál szinte egyedül ez a közelités látszik eredményesnek. E kutatásokhoz sorolható Bagrov diszkriminancia-analizisre támaszkodó, a havi csapadékösszeg előrejelezhetőségét vizsgáló modellje, vagy például a svéd Nyberg /1976/ kontingenciákat alkalmazó modellje.

A becslési és illeszkedési eljárásokkal foglalkozott egy korábbi, konkrét számitástechnikai vonatkozásokat is felölelő tanulmányunk /Faragó-Gulyás,1975/. A "közvetett" becslések csoportjának egyik legegyszerübb, de a meteorológiai prognosztikában a leggyakrabban hasznositott tipusa a "legközelebbi társ" eljárások köre. Ezzel kapcsolatban többek között Craddock/1958/ vagy Koppány/1974/ dolgozatára hivatkozhatunk. Az alapelveket és néhány operativ modellt tárgyal Faragó-Kaba-Gulyás/1975/.

A Központi Meteorológiai Intézetben távprognosztikai célokra egy olyan k-NN / k legközelebbi társat felhasználó, azaz k-Nearest Neighbour/ alapu statisztikai modell került kidolgozásra /Faragó et al.,1975; Faragó-Kaba,1976 /, melyben prediktorként egy speciális /diszkrét/ mintatér elemeit alkalmaztuk. Mindenekelőtt a modell "előrejelző képességére" voltunk kiváncsiak, amit az empirikus Bayes-rizikó közvetett becslésével jellemeztünk. A becslések elvégzésére Wagner /1971/ egy tételének általánositása adott módot. E témával foglalkozik e dolgozat 1. fejezete.

A nagydimenziószámu statisztikai becsléseknél jelentős problémát jelent az előzetes adatzsugoritás. Ezzel kapcsolatban lényeges kérdés, hogy ez a transzformáció miként változtatja meg a kérdéses statisztikai becslési feladatot: folytonosan viselkedik-e az optimális /bayesi/ becslésen a kitüzött veszteségfüggvény? Korábban speciálisan a sorfejtéses adatredukciókkal foglalkoztunk /Faragó-Gulyás,1973a,1973b/, utóbb pedig a vizsgálat tárgya volt a becslés folytonossága /Faragó,1975a/,valamint általánosabb vetületben az un. hiba-torzitás függvény viselkedése /Faragó-Györfi,1974,1975/.

A sorfejtéses adatredukció alkalmazása jelenleg egy multiregressziós extrapolációs feladat kapcsán merült fel. Mig egy korábbi dolgozatban /Faragó,1975b/ egydimenziós mintapéldán a regressziós becslés javitásának olyan lehetőségét mutattuk be, amelyik a mintatér particionálására épül, addig ebben a feladatban a nagydimenziószámu minták előzetes adatredukciót tesznek szükségessé. Az optimálisnak számitó Karhunen-Loéve sorfejtés azonban ismét csak a nagy dimenziószám miatt nem végezhető el közvetlenül és igy még egy ezt is megelőző - spektrális - adatredukciót alkalmaztunk.

A 2. fejezetben bemutatjuk a sorfejtéses adatzsugoritásokkal összefüggő elméleti vizsgálatokat, a lineáris multiregresszió alapjait, valamint az ezeket felhasználó meteorológiai modell számitástechnikai tervét. Végül bemutatunk a szuperponált kódolással kapcsolatos néhány konkrét eredményt.

Mint már fentebb utaltunk rá, az "analógiás" / NN / eljárások igen elterjedtek a meteorológiai alkalmazásokban. Osztályozási szabályok statisztikai tanulására, azaz "alakfelismerésre" számos módszer ismeretes a statisztikai szakirodalomból és ezek sorában is az NN-módszerek többféle verziója található meg. Nagyszámu osztály /hipotézis/ és nagyobb dimenziószám mellett általában más algoritmusokkal nem is próbálkozhatunk. Bizonyos, a légköri mozgásrendszerek strukturáját leiró un. makroszinoptikus kódok felismerésével kapcsolatban éppen ilyen probléma volt adott. Ennek megfelelően memória- és müvelettakarékos szelektiv NN-eljárás került kidolgozásra. Ennek egy empirikus vizsgálatát mutatja be a 3. fejezet.

A statisztikai kutatások sorában végül maximumértékek becslésével kapcsolatos vizsgálatot mutatunk be/ Faragó,1977/. A vizsgálat célja annak eldontése, hogy a statisztika alapjául milyen időegységre adjuk meg a maximumértékeket adott mintaszám esetén. A kérdés konzisztencia-vizsgálattal dönthető el. Elemezzük a gyenge függőség és a paraméteres statisztika esetét is. Az ezt tárgyaló 4. fejezet egy konkrét analizissel zárul.

1. "LEGKÖZELEBBI TÁRS" /NN/ BECSLÉSI ELJÁRÁS ÉS A BAYES-DÖNTÉS ÁTLAGOS VESZTESÉGÉNEK BECSLÉSE DISZKRÉT MINTATÉR ÉS FOLYTONOS PARAMÉTERTÉR ESETÉN

1.1. Egy meteorológiai feladat

Hess és Brezowsky /1952/ makroszinoptikus kódok olyan rendszerét dolgozta ki, melyekkel napról-napra tobbé-kevésbé jól jellemezhető az Európai Szinoptikai Körzet /ESzK/ fölötti levegő troposzférába eső részének áramlási strukturája, a markáns légnyomásképződmények elhelyezkedése és közvetve e régió időjárása. E katalógus 30 elemet tartalmaz. Ha egy adott időszak minden napjához hozzárendeljük a megfelelő kódot /a továbbiakban HB-kódot/, akkor az igy nyert sorozat az ESzK fölötti légcirkuláció menetét irja le nagy vonalakban. E kódok hosszu időre visszanyuló archivuma létezik számitógépes adathordozón, amely lehetővé teszi, hogy pl. egy adott hónap x HB-kódsorozatához hasonlókat keressünk a multból prognosztikai céllal. Két tetszőleges HB-kód osszehasonlitására adott egy S nyereségmatrix /"scoring"-táblázat/. Ennek segitségével definiálható egy $\Delta(x,\tilde{x})$ hasonlósági mérőszám a kódsorozatok között, amelyik kvázimetrika:

$$\Delta(x,\tilde{x}) \geqslant 0$$

$$\Delta(x,\tilde{x}) = 0 \iff x = \tilde{x}$$

Δ - szimmetrikus

Most azt a feladatot tüzzük ki, hogy "analógiás" eljárással megbecsüljük az x_o aktuális mintaelemet /megfigyelést/ követő hónap ϑ_o budapesti középhőmérsékletét. Az analógiakeresés a $T_N = \{(x_k, \vartheta_k)\}_{k=1}^{N}$ adott megfigyelési sorozatra /archivumra/ épül, ahol ϑ_k az azt a hónapot követő budapesti havi középhőmérsékleti érték, amelyikre x_k vonatkozik.

1.2. A probléma matematikai megközelitése

Legyen adott az (x_o, ϑ_o) valószinüségi változópár,

$x_o \in X$ - metrikus mintatér Δ metrikával, $\vartheta_o \in \Theta$ a paramétertér. $T_N = \{ (x_k, \vartheta_k) \}_{k=1}^{N}$ független, egyforma eloszlásu megfigyelések sorozatát jelöli az (x_o, ϑ_o) párra. Döntésfüggvénynek nevezzük a mintatér mérhető leképezését a paramétertérre:

$$d: \quad X \longrightarrow \Theta$$
$$d(x_o) = \hat{\vartheta} \quad ,$$

ahol $\hat{\vartheta}$-t a ϑ-nek a d döntésfüggvénnyel származtatott becslésének tekintjük. E becslés jóságának mérésére bevezetjük a

/1/ $$W(\vartheta_o, \hat{\vartheta}) = (\vartheta_o - \hat{\vartheta})^2$$

veszteségfüggvényt, itt már kikötve, hogy $\Theta \subset R_1$.
Az átlagos veszteség:

/2/ $$R(d) = E\, W(\vartheta_o, d(x_o)) \quad .$$

A bayesi döntésfüggvény definiciószerüen a legkisebb átlagos veszteségü döntés az összes döntésfüggvények kozül:

$$d^* : \quad R^* = R(d^*) \leq R(d) \qquad \forall d \quad .$$

Ismeretes, /Rényi, 1970/, hogy ebben az esetben a bayesi döntésfüggvény éppen a feltételes várhatóérték /ha az létezik/

/3/ $$d^*(x) = E(\vartheta / x) \quad .$$

A feladat minél optimálisabb megoldása, azaz ϑ_o minél pontosabb becslése a Bayes-döntés közvetlen közelitésével érhető el. Ez azonban általában nem kivitelezhető, mert bár létezik nem egy, például szekvenciális eljárás többek között a megfelelő feltételes sürüségfüggvény becslésére, melyek aszimptótikusan kedvező tulajdonságokkal rendelkeznek, de a gyakorlatban a feladatok többsége viszonylagosan kismintás.

A "legközelebbi társ" /a továbbiakban NN/ eljárást a következő döntésfüggvény irja le:

$$d_N: \quad d_N(x) = \vartheta_k \quad ,$$

ahol ϑ_k annak az x_k mintának a paramétere, melyre

$$\Delta(x,x_k) \leq \Delta(x,x_\ell) \qquad \ell=1,2,\ldots,N \ ,$$
$$\Delta(x,x_k) < \Delta(x,x_\ell) \qquad \ell<k \ .$$

A továbbiakban a

$$d_N(x_o) = \vartheta_{o(N)} \quad ; \quad R_N = R(d_N)$$

jelöléseket fogjuk alkalmazni. Feltéve, hogy $N \to \infty$ esetén létezik az NN-döntés átlagos veszteségének határértéke, jelölje $R=\lim_{N\to\infty} R_N$ az NN-döntés aszimptótikus veszteségét.

"Nulla-egy" veszteségfüggvényre

/3a/ $$W(\vartheta_o,\hat{\vartheta}) = \begin{cases} 0 \ , & \text{ha } \vartheta_o = \hat{\vartheta} \\ 1 \ , & \text{egyébként} \end{cases} \ ,$$

amikoris $R(d)$ a $d(x) = \hat{\vartheta}$ döntés hibavalószinüségét jelenti, Wagner/1971/ bebizonyitotta, hogy az

/4/ $$S_N = \frac{1}{N}\sum_1^N W(\vartheta_k, \vartheta_{k(N)})$$

statisztika aszimptótikusan torzitatlan becslése R -nek, sőt $S_N \to R$ sztochasztikusan. Itt $\vartheta_{k(N)}$ a ϑ_k NN-becslése a $T_N - \{(x_k, \vartheta_k)\}$ mintasorozatból. Márpedig ha eképpen R közelithető egy véges minta segitségével, akkor Cover/1969/ alapján az adott modell előrejelzőképességét kifejező R^* bayesi átlagos veszteség is becsülhető:

$$R^* \in \left[\left(1 - \sqrt{1 - 2R_N}\right) \cdot \frac{1}{2} \ , \ R_N \right] \ .$$

Kevésbé kedvező a helyzet az /1/ veszteségfüggvény esetén: $R^* = 2R$.

Ahhoz tehát, hogy az /1/ négyzetes veszteségfüggvény esetén R^* -t megbecsülhessük, szükséges a Wagner-tétel általánositása. Ehhez mindenekelőtt az NN-módszerek alapvető konvergenciatételére hivatkozunk.

<u>1. Tétel</u> /Cover-Hart,1967/: Ha az X mintatér szeparábilis és

metrikus a Δ metrikával és $x_1, x_2, \ldots, x_N, x_o$ független, egyforma eloszlásu valószinüségi változósorozat, akkor $x_{o(N)} \to x_o$ 1 valószinüséggel.

Ez a tétel igaz marad olyan diszkrét mintatérre is, melyen csak egy Δ kvázimetrika adott. Most már megfogalmazható az általánositott Wagner-tétel, /Faragó-Kaba,1976/.

2. Tétel: Legyen adott az X mintatér olyan Δ /metrikával vagy/ kvázimetrikával, hogy a független, egyforma eloszlásu $x_1 x_2, \ldots, x_N, x_o$ mintákra teljesüljön az $x_{o(N)} \to x_o$ /m.m./ konvergencia. Feltéve, hogy a Θ paramétertér korlátos, valamint $d^*(x) = E(\vartheta_o / x)$ és $\sigma(x) = E(\vartheta_o^2 / x)$ feltételes várható értékek folytonosak /m.m./, igaz az $S_N \to R$ négyzetes középben vett konvergencia.

Bizonyitás. Irjuk fel a megfelelő négyzetes középben vett eltérést és bontsuk fel a zárójelet:

$$/5/ \qquad E(S_N - R)^2 = E(S_N^2) - 2R\,E(S_N) + R^2 = E(S_N^2) - 2RR_{N-1} + R^2 ,$$

ahol kihasználtuk, hogy S_N torzitatlan becslés R_{N-1} -re:

$$E(S_N) = \frac{1}{N}\sum_{k=1}^{N} E(\vartheta_k - \vartheta_{k(N)})^2 = \frac{1}{N}\sum_{k=1}^{N} R_{N-1} = R_{N-1} .$$

Cover/1968b/ nyomán ismeretes, hogy $R_N \to R$, következésképpen most azt kell kimutatnunk, hogy $E(S_N^2) \to R^2$. Itt

$$E(S_N^2) = \frac{1}{N^2} \sum_{k,l} E\{(\vartheta_k - \vartheta_{k(N)})^2 (\vartheta_l - \vartheta_{l(N)})^2\} =$$

$$= \frac{1}{N^2} \sum_{k=1}^{N} E(\vartheta_k - \vartheta_{k(N)})^4 + \frac{1}{N^2} \sum_{k \neq l} E\{(\vartheta_k - \vartheta_{k(N)})^2 (\vartheta_l - \vartheta_{l(N)})^2$$

$$= \frac{1}{N} E(\vartheta_k - \vartheta_{k(N)})^4 + \frac{N-1}{N} E\{(\vartheta_k - \vartheta_{k(N)})^2 (\vartheta_l - \vartheta_{l(N)})^2\} ,$$

ahol az utóbbi kifejezés az $(x_1, \vartheta_1), (x_2, \vartheta_2), \ldots, (x_N, \vartheta_N)$ párok függetlensége és egyforma eloszlása miatt tetszőleges k és l indexekre értendő. A jobboldal első tagja

Ⓗ korlátossága miatt tart nullához, tehát ha igazolást nyer, hogy

$$/6/ \qquad E(\vartheta_k - \vartheta_{k(N)})^2 (\vartheta_l - \vartheta_{l(N)})^2 \xrightarrow[N\to\infty]{} R^2 ,$$

akkor $E(S_N^2) \to R^2$ is teljesülni fog. /6/ két tényezője nem független, ezért bevezetjük ϑ_k és ϑ_l NN-becslését a $T_N - \{(x_k, \vartheta_k), (x_l, \vartheta_l)\}$ mintasorozatból, melyet $\tilde{\vartheta}_k$ ill. $\tilde{\vartheta}_l$ -lel jelölünk. Az igy nyert két becslés aszimptótikusan felcserélhető, ugyanis a

$$K = \left| (\vartheta_k - \vartheta_{k(N)})^2 (\vartheta_l - \vartheta_{l(N)})^2 - (\vartheta_k - \tilde{\vartheta}_{k(N)})^2 (\vartheta_l - \tilde{\vartheta}_{l(N)})^2 \right|$$

jelöléssel K csak az $A = \{\vartheta_{k(N)} = \vartheta_l \cup \vartheta_{l(N)} = \vartheta_k\}$ eseményen különbözhet nullától, tehát

$$E(K) = E(K/A)\, P(A) \leq C \frac{2}{N-1} \xrightarrow[N\to\infty]{} 0 ,$$

ahol a feltételes várhatóérték Ⓗ korlátossága miatt korlátos.

Tekintsük most /6/ baloldalát a módositott NN-becslésekkel - átalakitva :

$$/7/ \qquad E\left[E\{(\vartheta_k - \tilde{\vartheta}_{k(N)})^2 (\vartheta_l - \tilde{\vartheta}_{l(N)})^2 / x_k, \tilde{x}_{k(N)}, x_l, \tilde{x}_{l(N)}\}\right] =$$

$$= E\left[E\{(\vartheta_k - \tilde{\vartheta}_{k(N)})^2 / x_k, \tilde{x}_{k(N)}\}\, E\{(\vartheta_l - \tilde{\vartheta}_{l(N)})^2 / x_l, \tilde{x}_{l(N)}\}\right] =$$

$$= E\left[L(x_k) \cdot L(x_l)\right] ,$$

ahol

$$L(x_k) = \sigma(x_k) - 2d^*(x_k) \cdot d^*(\tilde{x}_{k(N)}) + \sigma(\tilde{x}_{k(N)}) .$$

A Cover-Hart - tétel alapján $\tilde{x}_{k(N)} \to x_k$, $\tilde{x}_{l(N)} \to x_l$ /m.m./, tehát d^* és σ folytonossága miatt

$$L(x_k) \to 2\,\sigma(x_k) - 2\left[d^*(x_k)\right]^2 =$$

$$= 2\, E\{(\vartheta_k - d^*(x_k))^2 / x_k\} \equiv 2r^*(x_k)$$

és hasonlóképpen

$$L(x_1) \longrightarrow 2\,E\left\{(\vartheta_1 - d^*(x_1))^2/x_1\right\} \equiv 2r^*(x_1) \quad .$$

Mivel a paramétertér korlátos, alkalmazható a Lebesque-tétel és igy a /7/ várhatóérték, azaz /6/ baloldala megfelelőképpen konvergál a

$$4\,E\left\{r^*(x_k)\,r^*(x_1)\right\} = 4\,Er^*(x_k)\,Er^*(x_1) = 4\,(R^*)^2$$

értékhez. /Itt kihasználtuk x_k és x_1 függetlenségét./ Cover/1968b/ tétele alapján $R=2R^*$ tehát valóban fennáll a /6/ konvergencia és ezzel a tételt igazoltuk. □

E tétel lehetőséget nyujt R^* becslésére, bár a konvergencia sebessége ismeretlen. Csupán végesdimenziós /diszkrét/ Θ -ra lelhetők fel ezzel kapcsolatos eredmények pl. Cover/1968a/ és Fritz/1975/ munkáiban.

Ha R^* közelitése kellő alapot nyujt a $\hat{\vartheta}$ becslés pontositására, akkor az NN-eljárásról érdemes áttérni a k-NN - eljárásra. Cover/1968b/ vizsgálatai alapján ugyanis ennek aszimptótikus átlagos vesztesége

$$R^{(k)} = \left(1 + \frac{1}{k}\right) R^* \quad .$$

Ebben az esetben a becslés ϑ_o-ra:

$$\hat{\vartheta} = \frac{1}{k} \sum_{j=1}^{k} \vartheta^{(j)}_{o(N)} \quad ,$$

ahol $(x^{(j)}_{o(N)}, \vartheta^{(j)}_{o(N)}) \in T_N$ és $x^{(1)}_{o(N)}, x^{(2)}_{o(N)}, \ldots, x^{(k)}_{o(N)}$ az x_o minta k legközelebbi társa.

1.3. A meteorológiai feladat megoldása

Olyan programot készitettünk, mely az 1.1 pontban vázolt meteorológiai feladatban egy aktuális x_o kódsorozathoz független, egyforma eloszlásunak tekinthető T_N megfigyelési sorozatból megadott számu analógiát keres ki. N=93 esetén megbecsültük az S_N empirikus átlagos veszteséget akkor, amikor x_o a január havi HB-kódsorozatot, ϑ_o

pedig a február havi budapesti átlaghőmérsékletet jellemezte: S_N=11,7. Ennek alapján $R^* \approx 5,9$. ϑ_0 szórásnégyzete σ_ϑ =9,0 , tehát a Bayes-döntés már jelentékeny információnyereséget nyujthat. A többé-kevésbé optimális k=9 esetére S_N= 8,4 ,ami számottevő javulást jelent. /Rögzitett N mellett persze k növekedésével az S_N becslés hatékonysága romlik./ Mindamellett σ_ϑ-val összevetve a kapott S_N még tul nagy, de ezzel a k-NN -technikával csak ugy érhetnénk el jobb eredményt, ha k -t tovább növelhetnénk. Ennek viszont gátat szab a rögzitett N érték.

Az 1.,2. mellékletek az "analógia-kereső" program egy olyan ujabb variánsát mutatják be, ahol az adott x_0 -hoz már nemcsak a megelőző kerek naptári hónapokból keresünk analógiákat, hanem általánosabban x_k minta lehet a naptári hónapot csak részben fedő hónapnyi hosszuságu időszak kódsorozata is. Ezzel a fogással tovább javult a becslések minősége, a numerikus értékek további konkrét bemutatásától itt azonban eltekintünk.

2. ADATZSUGORITÁS A STATISZTIKAI BECSLÉSELMÉLETBEN

2.1. A lineáris regresszióról a meteorológiában és egy előrejelzési feladat

A dinamikus meteorológia elméletének legnagyobb vivmánya a légkör hidro- és termodinamikáját leiró un. primitiv egyenletek rendszerének megalkotása. A nagyteljesitményü számitógépek megjelenése - mint már a bevezetőben emlitettük - lehetővé tette, hogy e primitiv egyenletek segitségével és necsak egy elméletileg erősen szürt egyenletrendszerrel numerikusan is modellezni lehessen a légköri folyamatokat. A numerikus modellezésnél akár differencia-módszereket akár spektrális feprezentációt alkalmaznak, lényegében az egyenletrendszer analitikus megoldásának közvetlen approximációjáról van szó.

Korábban még maga von Neumann /Smagorinsky, 1969 / három csoportba osztályozta a légköri jelenségek előrejelzéseit. Az azóta eltelt időszak tapasztalatái őt igazolták: alapjaiban eltérő technikával kell közelitenünk a rövid /legfeljebb kb. 1-2 hétre szóló/, a hosszutávu /1-12 hónapos/ és az ultrahosszutávu előrejelzéseket. Mig a rövidtávu előrejelzéseknél uralkodóbb a dinamikai alapu modellezés, addig a hosszutávuaknál szinte kizárólagosan statisztikai eljárásokkal találkozhatunk a szakirodalomban; végül az ultrahosszu távu vizsgálatok ismét csak dinamikai /a termo- hidrodinamikai egyenletrendszer megoldására épülő/ módszerekre támaszkodhatnak. A statisztikai eljárások is akkor lehetenek jobb hatásfokuak ha pl. az egyszerü statisztikai analógiakeresésben kifejezett implicit prediktor-prediktandusz relación tulmenően a dinamikai eljárásokhoz hasonlóan minél közvetlenebbül "analitikusan" próbálják közeliteni a megoldást - az emlitett reláció paraméteres alakjának minél jobb megadásával. A legegyszerübb esetben lineáris multiregressziót választanak e célra és nem ritkán a vele szoros kapcsolatban álló lineáris maximálkorrelációt, az un. kanonikus korrelációt is megbecsülik /Judin,1967; Bagrov, 1968; Gandin,1967; Glahn,1968; Dujceva-Pegy,1970; Bagrov-Mjakiseva,1969/. Ennek általános elméletével, számitástechnikai realizációjával és meteorológiai alkalmazhatóságával fog-

lalkoztunk már korábban /Faragó-Gulyás,1974/. Még hatásosabb lehet a mintatér fizikai vagy statisztikai meggondolásokra épülő particionálása és az egyes partíciókon való lineáris regresszióe kapcsolat létesitése. Ennek egy alkalmazását Faragó /1975/ mutatja be; az optimális particionálás egy speciális esetével Ter-Mkrtcsan/1970 / foglalkozott.

A hidro- és termodinamikai egyenletrendszerből származtatható legegyszerübb verzió, amely azonban még nagy vonalakban modellezi a légköri folyamatok fejlődését, az un. divergenciamentes barotrop modell. Ez az a modell, melyen von Neumann és társai a bevezetőben emlitett klasszikus vizsgálataikat elvégezték. Ennek révén vált a meteorológiában kitüntetetté az 500 mb-os /közepes vagy barotrop szint/ topografikus mezeje, mely azóta a statisztikai modellezésben is jelentős szerepet játszik. Ezen a szinten ugyanis már kellőképpen lassuak a változások ahhoz, hogy hosszabb érvényességi időtartamu extrapolációra is gondolhassunk, másfelől viszont még elég jelentős a kapcsolat e szint és a talajfelszinhez közeli időjárás között.

Ezért - irodalmi előtanulmányok után - olyan számitástechnikai rendszer tervét fogalmaztuk meg, mely lehetőséget teremt /pl./ az 500 mb-os topográfia előrejelzésére. Figyelembevéve a légköri mozgások átlagos sebességrendjét a hosszabbtávu előrejelzések már feltétlenül hemiszférikus nagyságu mezők figyelembevételét teszik szükségessé. Magaslégöri megfigyelések már több mint 20 éve folynak rendszeresen és nekünk rendelkezésünkre áll az ezek alapján készült, egy reguláris rácshálózatra interpolált napi megfigyelések archivuma az emlitett 500 mb-os topgrafikus mezőkből. E reguláis rács a szélességek mentén 10°-os, a hosszuságok mentén 5°-os osztásu /a pólus közelében ritkitott és csak a 20° é.sz.-ig terjed/.

Láthatóan a prediktor és a prediktandusz mintatere itt egybeesik és mintegy 500 dimenziós. Ekkora dimenziószámu mintákkal a regressziós extrapoláció feladata gyakorlatilag megoldhatatlan: előzetes adatredukcióra van szükség.

2.2. Sorfejtéses adatzsugoritások és azok szuperponálása

Az 1.2. ponthoz hasonlóan legyen $(x,\vartheta) \in (X,\Theta)$ a vizsgált valószinüségi változópár /a zérus indexeket elhagytuk/,

W a veszteségfüggvény, d^* pedig a minimális $R^* = R(d^*) = EW(\vartheta, d^*(x))$ átlagos veszteséget adó /bayesi/ becslés.

A mintatér valamilyen mérhető

/1/ $$F: \quad X \longrightarrow Y$$
$$F(x) = y$$

transzformációja nyomán - ahol $\{X, \mathfrak{X}\}$, $\{Y, \mathfrak{Y}\}$ mérhető terek - az eredeti feladat helyett most ϑ-nak y-mérhető becslését keressük. /Egészen pontosan, ha $\{\Omega, \mathcal{A}, P\}$ jelöli a valószinüségi mezőt, akkor $\mathcal{A} \supset \mathcal{A}_x \supset \mathcal{A}_y$ és most $\mathcal{A}_y$-mérhető $\hat{\vartheta}$-becslést keresünk. Itt $\mathcal{A}_x$, $\mathcal{A}_y$ az x illetve $y=F(x)$ valószinüségi változó által generált σ-algebra./ Erre feltétlenül igaz az

/2/ $$R^* \equiv EW(\vartheta_o, d^*(x_o)) \leq EW(\vartheta_o, d_F^*(F(x_o))) \equiv R_F^*$$

reláció.

Most bebizonyitjuk, hogy például $\Theta = \{0,1\}$ esetén, vagyis a legegyszerübb döntési feladatban bizonyos értelemben R^* folytonosan változik a mintatér transzformációinak függvényében. Ebben az esetben

/3/ $$d^*(x) = \text{sign}\, E(\vartheta/x) = \text{sign}\big(P(\vartheta=1/x) - P(\vartheta=-1/x)\big)$$

és természetszerüen W -nek az 1.1.-ben /3a/ alatt bemutatott 0-1 veszteségfüggvényt választjuk. Ekkor az átlagos veszteség a hibavalószinüséget jelenti. Jelölje $\delta(x)$ a feltételes várható értéket:

$$\delta^*(x) = E(\vartheta/x) \quad ,$$

akkor tetszőleges d döntésfüggvényre

/4/ $$R(d) = \frac{1}{2}\left(1 - E\{d(x)\delta^*(x)\}\right) \quad .$$

<u>Lemma:</u> Legyen F_1, F_2 két mérhető leképzés X -ből Y-ba és

$$D \equiv \{ F_1(x) \neq F_2(x) \} \in \mathcal{A} \quad , \quad P\{D\} < \varepsilon \quad ,$$

akkor

/5/ $$|R(d_{F_1}) - R(d_{F_2})| < 2\varepsilon \quad .$$

Bizonyitás./A /Faragó,1975a/-ban adottnál sokkal triviálisabb./

$$|R(d_{F_1}) - R(d_{F_2})| \leq |P(d_{F_1}(x) \neq \vartheta/\bar{D}) - P(d_{F_2}(x) = \vartheta/\bar{D})| P\{\bar{D}\} +$$
$$+ |P(d_{F_1}(x) \neq \vartheta/D) - P(d_{F_2}(x) \neq \vartheta/D)| P\{D\} \leq 2\varepsilon \quad . \square$$

1. Tétel /Faragó,1975a/: Legyen az X mintatér szeparábilis metrikus mintatér a Δ metrikával és tegyük fel, hogy a /3/ döntésfüggvény folytonos a következő értelemben:

$$\forall \varepsilon > 0 \quad \exists \delta > 0$$

/6/ $$Q\{ x: d(x') = \text{const, ha } \Delta(x,x') < \delta \text{ /m.m./} \} \geq 1 - \varepsilon \quad ,$$

ahol Q az x valószinüségi változó által generált valószinüségi mérték X-n , $Q\{B\} = P\{ x \in B \}$. Legyen továbbá adott a mintatér mérhető transzformációja önmagába, $F: X \to X$, melyre

/7/ $$\Delta(x, F(x)) < \delta/4 \quad \text{/m.m./} \quad .$$

Akkor a /2/-ben definiált átlagos veszteségek eltérésére:

$$\lambda(F) \equiv |R_F^* - R^*| < \varepsilon \ .$$

Bizonyitás. /4/-ből

/8/ $$\lambda(F) \leq \int_{\{d_F^*(x) = d^*(x)\}} |\delta^*(x)| \, dP \leq Q\{ d_F^*(x) = d^*(x) \} \quad .$$

Bevezetjük a következő jelöléseket:

$$X_0 = \{ x;\ d^*(x') = \text{const}, \Delta(x,x') < \delta \text{ /m.m./} \} \ ,$$
$$U_r(x) = \{ x' : \Delta(x,x') < r \} \quad , \quad r > 0 \quad ,$$
$$A = \{ d_F^* \neq d^* \} \in \mathcal{X} \ .$$

Válasszunk egy tetszőleges $x \in X_0$ pontot és ennek $U_{\delta/2}(x)$ környezetét. Akkor /7/ következtében

$$V_x \equiv F^{-1}(F\{U_{\delta/2}(x)\}) \subset U_\delta(x) \quad ,$$

továbbá /6/-nak és X_0 definiciójának figyelembevételével, látható, hogy d^* konstans $U_\delta(x)$ -n. Az általánosság csorbitása nélkül kiköthetjük, hogy $d^*(x') = 1$ m.m. $x' \in U_\delta(x)$ -re. Mivel V_x eleme az F által generált $\sigma_F \subset \mathfrak{X}$ σ-algebrának, felirható

$$\int_{V_x \cap \{d_F^* = -1\}} E(\delta^*/F)\, dQ = \int_{V_x \cap \{d_F^* = -1\}} \delta^* dQ \quad .$$

Ha itt az integrálási tartomány Q-mértéke nullától különböző, akkor a baloldal negativ, a jobboldal viszont pozitiv előjelü. Következésképpen minden $x \in X_0$ -ra $d^*(x') = d_F^*(x')$ m.m. $x' \in U_{\delta/2}(x)$ elemre, hiszen $U_{\delta/2} \subset V_x$. Tehát X szeparabilitásából azonnal következik, hogy $d^*(x) = d_F^*(x)$ m.m. $x \in X_0$ -ra, azaz $Q\{X_0 A\} = 0$. Igy /8/-ból:

$$\lambda(F) \leq Q\{\overline{X}_0 A\} \leq Q\{\overline{X}_0\} < \varepsilon \quad ,$$

amit éppen bizonyitani kellett. □

Akkor is igaz marad a tétel állitása, ha a transzformáció csak átlagosan kismértékben torzitja az eredeti mintaelemeket.

<u>2. Tétel</u> /Faragó, 1975a/: Ha a fenti tétel feltételei között /7/ helyett azt tesszük fel, hogy valamilyen valós $\delta_1 > 0$ -ra, melyre $\delta_1 \leq \varepsilon\delta/8$, fennáll az

/9/ $\quad E\Delta(x, F(x)) < \delta_1$

egyenlőtlenség, akkor $\lambda(F) < 2\varepsilon$.

Bizonyitás. Alkalmazzuk a Csebisev-egyenlőtlenséget:

$$Q\{\Delta(x, F(x)) < \frac{\delta}{4}\} > 1 - 4\frac{E\Delta(x, F(x))}{\delta} > 1 - \frac{4\delta_1}{\delta} > 1 - \frac{\varepsilon}{2} \quad .$$

Legyen $X_1 = \{\Delta(x, F(x)) < \delta/4\}$ és vezessük be az

$$F_1(x) = \begin{cases} F(x) & , \quad x \in X_1 \\ x & , \quad x \notin X_1 \end{cases}$$

operátort. Akkor az lemma alapján

$$\left|R(d_F^*) - R(d_{F_1}^*)\right| < \varepsilon$$

és F_1 kielégíti az előző tétel feltételeit, következésképpen

$$\lambda(F) \leq \lambda(F_1) + \left|R(d_F^*) - R(d_{F_1}^*)\right| < 2\varepsilon$$

és ezzel a tételt igazoltuk. □

A fenti tételek körülményeinek megfelelő tipikus transzformációk a csonkitott sorfejtések. Legyen x_t, $t \in T$ véges szórásu folyamat, $x_t \in L_2(\Omega, \mathcal{A}, P)$ és $\varphi_k(t)$ egy teljes ortonormált rendszer az $L_2(T)$ térben. Ha most a folyamat sorbafejthető /m.m./

$$x_t = \sum_{k=1}^{\infty} c_k \varphi_k(t)$$

akkor e sorfejtés kellő számu első együtthatója felfogható a folyamat reprezentációjaként és ha a mintatér megfelelő

$$F: \quad x_t \longrightarrow \sum_{k=1}^{K} c_k \varphi_k(t)$$

transzformációjára teljesülnek valamelyik fenti tétel feltételei, akkor remélhetjük, hogy az eredeti statisztikai feladat is csak kismértékben torzul.

A bizonyos értelemben optimális sorfejtéses adatredukciót az un. Karhunen-Loéve sorfejtés adná meg /Fu,1960; Watanabe, 1972; Fritz,1971/, de elvégzéséhez többek között ugyancsak a mintaanyag statisztikai vizsgálata /nevezetesen a kovariancia-függvény becslése/ szükséges, ami nagy dimenziószám mellett igen körülményes /memória- és gépidőigényes/.

A hemiszférikus /és más/ meteorológiai mezők dinamikai és statisztikai vizsgálataiban igen elterjedtek a spektrális módszerek. Ezek alapján jól ismert, és a magassági /topografikus/ mezőkre is jellemző, hogy a mozgásformák közül a prognosztikában kiemelkedő szerepet játszanak az egészen nagy hullámhosszak /az un. Rossby-hullámok/, mert ezek egyben a legnagyobb megmaradási hajlamuak is. A hidrodinamikai egyenletrendszerek egyik szürési kritériuma is éppen az, hogy az un. meteorológiailag káros /kisebb hullámhosszuságu, meteorológiailag érdektelen, pl.

gravitációs/ hullámokat elkülönitik. Az ezek elhagyásával megadott approximáció tehát nem feltétlenül eredményez analitikusan kitünő becslést, de fizikailag jól megalapozott és várhatóan elég jó előrejelezhetőséget biztosit.

Mindezek alapján kézenfekvő, hogy a kérdéses mintaelemeket /lásd 2.1. pont/ előbb Fourier-sorba fejtsük.

Jelöljön $G(\varphi, \theta)$ egy a gömbfelületen gömbi koordinátákban adott véletlen függvényt, $0 \leq \varphi \leq 2\pi$, $0 \leq \theta \leq \pi$. Az un. tesszerális szférikus harmonikusok /TSH/ teljes ortonormált rendszerét a

/10/
$$U_n^m(\varphi,\theta) = \sqrt{\frac{2n+1}{2\pi}\,\frac{(n-m)!}{(n+m)!}}\; P_n^m(\cos\theta)\cos m\varphi$$
$$V_n^m(\varphi,\theta) = \sqrt{\frac{2n+1}{2\pi}\,\frac{(n-m)!}{(n+m)!}}\; P_n^m(\cos\theta)\sin m\varphi$$

függvények alkotják, ahol P_n^m, $n=0,1,\ldots$ $m=0,1,\ldots,n$ az n -edfoku, m -rendü Legendre-függvény.

Ujabban egyre intenzivebben hasznositják a TSH -k szerinti sorfejtéses reprezentációt a differenciamódszerek helyett is a hidrodinamikai differenciálegyenlet-rendszerek megoldásában /Jefimov,1976; Machenhauer,1974/ többféle lezárási /sorcsonkitási/ eljárást ajánlva. Mi itt a legtermészetesebbet választva a G mező approximációját a következőképpen adjuk meg:

/11/
$$G(\varphi,\theta) = \sum_{n=0}^{M}\left[\tfrac{1}{2}a_{n,0}U_n^0(\varphi,\theta) + \sum_{m=1}^{n} a_{n,m}U_n^m(\varphi,\theta) + b_{n,m}V_n^m(\varphi,\theta)\right]$$
$$a_{n,m} = \int_0^{2\pi}\int_0^{\pi} G(\varphi,\theta)\,U_n^m(\varphi,\theta)\sin\theta\,d\theta\,d\varphi;\quad b_{n,m} = \int_0^{2\pi}\int_0^{\pi} G(\varphi,\theta)\,V_n^m(\varphi,\theta)\sin\theta\,d\theta\,d\varphi.$$

Hemiszférikus mezőinket az Egyenlitőre szimmetrikusan egészitjük ki, ezáltal mindazon együtthatók zérusok lesznek, melyekre $n+m$=páratlan .

A kapott $C=$ [illegible] valószinüségi vektorváltozó most már /csökkentett dimenziószáma révén gyakorlatilag is/ Karhunen-Loéve sorba fejthető:

/12/
$$C = \sum_{k=1}^{2M} \sqrt{\lambda_k}\, f_k \varphi_k \quad ,$$

ahol $\varphi_k \in R_{2M}$, λ_k a C kovarianciamátrixa(Γ)sajátérték-

egyenletének megoldásai:

$$\Gamma\varphi_k = \lambda_k \varphi_k \qquad \text{és} \qquad f_k = \frac{1}{\sqrt{\lambda_k}} C \varphi_k \ .$$

E sorfejtés optimális tulajdonságait itt nem részleżezzük.

Összefoglalva, nagydimenziószámu minták esetén két egymástkövető /szuperponált/ sorfejtéses adatredukció alkalmazható, melyek eredményeképpen végülis az x_t folyamat reprezentációját az $f = (f_1, f_2, \ldots, f_K)$ valószinüségi vektorváltozó adja meg valamilyen alkalmas K esetén.

Konkrét feadatunkban lineáris multiregressziós extrapolácó kiépitése a cél, pontosabban a

$$W(f,f') = \sum_k (f_k - f'_k)^2$$

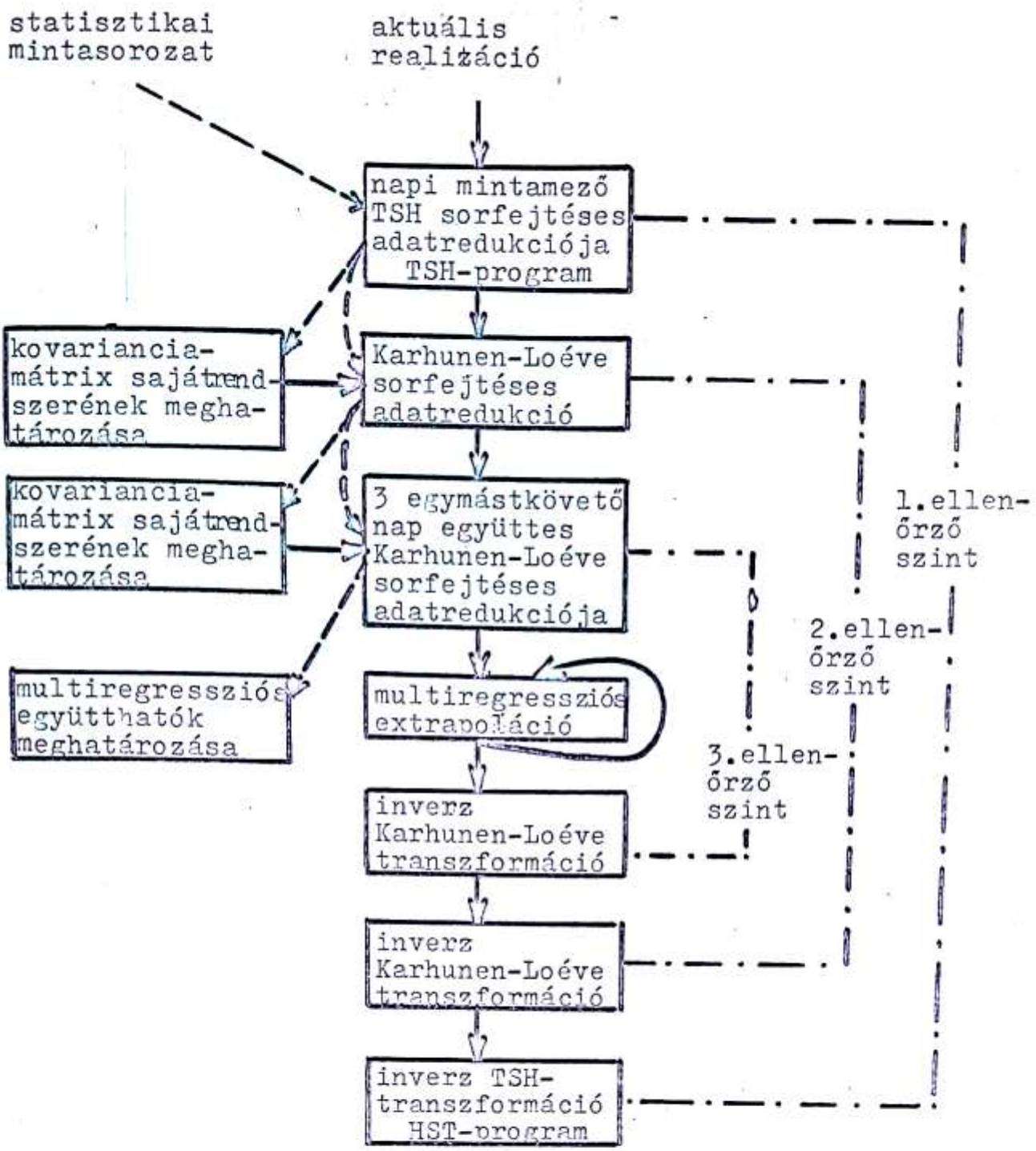
négyzetes veszteségfüggvény mellett - ahol f' az f-t követő időszakhoz tartozó transzformált minta -, az optimális lineáris f' becslés előállitása a feladat.

2.3. Konkrét példa a szuperponált adatzsugoritásra egy multiregressziós extrapolációs modell esetén

A 2.2.-ben vázolt adatredukciós eljárások egy olyan modellben nyernek alkalmazást, melynek célja az 500 mb-os /abszolut/ topográfia multiregressziós bázisu előrejelzése. A modellt a egy programrendszer realizálja; melyet az 1.ábra mutat be.

E fejezetben az adatzsugoritásokkal foglalkoztunk és ennek megfelelően a 3,4,5. mellékletekben bemutatjuk a spektrális ill. Karhunun-Loéve sorfejtéssel összefüggő TSH, HST és KL1 programokat. egy-egy futási eredménnyel illusztrálva.

A mintavételezés az 1967-68 -as évekből történt. A TSH-HST programpárral az 1. ellenőrző szinten vizsgálhatjuk meg konkrét realizációkon a spektrális sorfejtéses adatredukció minőségét a z egyszerü normabecslésen tulmenően. A példák elemzését a mellékletben adjuk meg.



1. ábra : Szuperponált adatredukciót megvalósító multiregressziós extrapolációs modellt realizáló programsorozat

3. EGY "SZELEKTIV LEGKÖZELEBBI TÁRS" STATISZTIKAI BECSLÉSI ELJÁRÁS EMPIRIKUS VIZSGÁLATA

3.1. Makroszinoptikus kódok statisztikai felismerésének problémája

Az 1.1. pontban emlitett Hess-Brezowsky-féle makroszinoptikus kódok /HB-kódok/ kapcsán adott a következő alakfelismerési feladat.

Az európai-atlanti térség fölött bevezetünk egy 5x7 -es rácshálózatot és ennek rácspontjaiban megadjuk a /megfelelőképpen interpolált/ talajszinti nyomási értéket egy-egy kiválasztott napon. Ilyen 35-dimenziós multbeli mintákból tetszőlegesen hosszu sorozat készithető ugy, hogy mindegyik mező mellett ismeretes az arra a napra szinoptikusok által megadott HB-kód. A feladat ennek a hozzárendelésnek a statisztikai tanulása.

A meteorológiai szempontból ugyancsak szerény 35 rácspont a matematikai modell kitüzésénél már igen komoly problémát jelent. Ez annál is inkább igaz, mivel az osztályok száma ebben a konkrét feladatban 30. Ezért annak ellenére, hogy számos eljárás ismeretes a feltételes eloszlások közvetett becslésére, szeparáló függvényeket előállitó tanuló-felismerő algoritmusokra, mégis csupán "Nearest Neighbour" /NN/ -tipusu eljárásokra támaszkodhatunk. Ebben az esetben a legnagyobb gond a tárolandó mintaelemek nagy száma és ezzel összefüggésben az analógiák várhatóan nagy keresési ideje. Éppen ezért olyan módszerre van szükség, mely eleve csökkenti az archiválásra kerülő minták számát.

3.2. Egy szelektiv NN-eljárás /SNN/

Legyen $T_N=\{(x_k,\vartheta_k)\}_{k=1}^{N}$, (x_k,ϑ_k), $(x_o,\vartheta_o)\in(X,\Theta)$ független, azonos eloszlásu mintasorozat; $X\subset R_M$ -metrikus mintér a Δ metrikával, Θ pedig véges /diszkrét/ paramétertér /ld. 1.2. pont/.

A közönséges NN-eljárásra támaszkodó alakfelismerési modell-

ben, a "tanulási fázisban" minden beérkező (x_k, ϑ_k) mintapárt megőrzünk és ennek megfelelően T_N mintasorozat mellett N távolságot kell kiszámitatunk x_0 és $x_1, x_2, \ldots, x_N$ között, illetve a legközelebbi társ kikereséséhez $O(N^2)$ összehasonlitást kell tennünk. Itt N rendszerint elég nagy értékeket érhet el és ezért egy-egy osztályozás meglehetősen "költséges".

E gondolat jegyében az NN-módszer különféle javitásai láttak napvilágot /Hart,1968; Gates,1972/.

Tipikus az olyan módositás, mikoris a teljes T_N mintaanyagot önmagán tesztelik és ennek nyomán bizonyos elemeit elhaggyák. Mi most egy olyan eljárást mutatunk be, melyben az archivum számossága általában véve nem növekszik olyan gyorsan, mint N.

Minden mintapár mellé hozzá fogunk rendelni egy ν_k természetes számot, amelyet osztályozási gyakoriságnak nevezünk. Az algoritmust a következőképpen adjuk meg: az első lépésben az archivum egyetlen elemet tartalmaz

$$S_1 = \{(x_1, \vartheta_1)\} \quad , \quad \nu_1 = 1 \quad ;$$

a /k+1/-edik $(x_{k+1}, \vartheta_{k+1}) \in T_N$ minta módositja a k-adik lépésben adott S_k archivumot

$$/1/ \qquad S_{k+1} = \begin{cases} S_k \cup \{(x_{k+1}, \vartheta_{k+1})\} & \text{ha } \vartheta_{k+1} \neq \vartheta \\ [S_k - \{(x, \vartheta)\}] \cup \{(\tilde{x}, \tilde{\vartheta})\} & \text{ha } \vartheta_{k+1} = \vartheta \end{cases}$$

és megfelelőképpen az uj mintaelem osztályozási gyakorisága $\nu_{k+1} = 1$ lesz, ha $\vartheta_{k+1} \neq \vartheta$, illetve a módositott (x, ϑ) minta osztályozási gyakorisága is változik eggyel $\tilde{\nu} = \nu + 1$, amikor $\vartheta_{k+1} = \vartheta$. Itt $(x_{k+1}, \vartheta_{k+1})$ legközelebbi társa az S_k mintasorozatból (x, ϑ), illetve ν jelöli ennek osztályozási gyakoriságát; $\vartheta_{k+1} = \vartheta$ esetén a módositott minta:

$$/2/ \qquad \tilde{x} = \frac{\nu \cdot x + x_{k+1}}{k+1} \quad , \qquad \tilde{\vartheta} = \vartheta \quad .$$

Másszóval, ha az eddigi archivum alapján az uj minta osztályozása helytelen, akkor azt felvesszük az archivumba $\nu_{k+1} = 1$ -

gyel; ha az osztályozás helyes, akkor csak az ezt osztályozó mintát korrigáljuk /2/ szerint, valamint az osztályozási gyakoriságát növeljük eggyel. Nyilvánvaló, hogy olyan szeparálható osztályok esetén, amikoris tetszőleges $x \in X$ esetén legfeljebb egy olyan $\vartheta \in \Theta$ létezik, hogy az $f(\vartheta/x)$ feltételes sűrűségfüggvény pozitiv és az $X_\vartheta = \{x: f(\vartheta/x) > 0\}$ jelöléssel $\forall \vartheta$ -ra és $\forall x \in X_\vartheta$ -re teljesül a

$$\forall x' \in X_\vartheta \quad \forall \vartheta'' \neq \vartheta \quad \forall x'' \in X_{\vartheta''} \quad \Delta(x,x') < \Delta(x,x'')$$

feltétel, akkor S_N összesen legfeljebb card(Θ) mintapárból fog állni és ezek első tagjai /2/ miatt rendre konvergálnak a megfelelő $E(x/X_\vartheta)$ feltételes várható értékhez. /Hiszen ekkor minden NN-osztályozás korrekt lesz /m.m./ és igy minden X_ϑ -ra külön-külön alkalmazható a nagy számok tétele./ Ettől az egészen speciális esettől eltekintve, emnek az SNN-eljárásnak a konvergenciájára nézve semmilyen elméleti eredmény nem ismeretes. A klasszikus NN-konvergenciák érvénytelenek, hiszen itt az archivum elemei változnak N növekedésével. Ezért egy egyszerü empirikus vizsgálat hivatott az SNN-eljárás egy előzetes elemzésére. Ennek során megvizsgáljuk azt a verziót is, amikor /2/ a következőképpen módosul: $\tilde{x} = x$, vagyis az archivum elemei nem változnak meg. Nevezzük ezt módositott NN-eljárásnak /MNN/. Ennek konvergenciája triviálisan következik a közönséges NN-módszerek konvergenciájából.

Legyen $\Theta = \{0,1\}$ és legyen $f(x/\vartheta)$ gaussi sűrűségfüggvény $\vartheta = 0$ esetén $m_0 = 0$, $\vartheta = 1$ esetén pedig $m_1 = 2$ várható-értékkel, valamint mindkét esetben $\sigma_0 = \sigma_1 = 1$ szórással. Legyen továbbá $P\{\vartheta = 0\} = P\{\vartheta = 1\} = 1/2$. Ebben a statisztikai felismerési feladatban az elméleti Bayes-döntést nyilvánvalóan a

$$/3/ \qquad \vartheta^* = \begin{cases} 0 & , \text{ ha } x_0 < 1 \\ 1 & , \text{ ha } x_0 \geqslant 1 \end{cases}$$

valószinüségi változó adja meg. A Bayes-hiba

$$/4/ \qquad P\{\vartheta_0 \neq \vartheta^*\} = P\{\vartheta_0 = 0, \ \vartheta^* = 1\} + P\{\vartheta_0 = 1, \ \vartheta^* = 0\} =$$

$$= P\{x_0 \geqslant 1/\vartheta_0 = 0\} P\{\vartheta_0 = 0\} + P\{x_0 < 1/\vartheta_0 = 1\} P\{\vartheta_0 = 1\} =$$

$$= \frac{1}{2}\left\{\left[1-\int_{-\infty}^{1}\varphi(x;0,1)dx\right]+\int_{-\infty}^{1}\varphi(x;2,1)dx\right\} = 0,16.$$

ahol φ a normális eloszlás sürüségfüggvénye az adott paraméterekkel.N=100 esetén egy konkrét numerikus példában

	NN	MNN	SNN
az archivum számossága	100	32	28

valamint 100 ujabb /véletlen gaussi számgenerátorral előállitott/ mintával számitva az empirikus hibavalószinüségek

	NN	MNN	SNN
az empirikus hibavalószinüség	0,20	0,18	0,18 .

A mintaanyagból párhuzamosan megbecsültük a Bayes-hibát is és eredményül 0,13-t kaptunk, ami igen jól közeliti a /4/ alatti elméleti értéket. /A realizáló programot és a három futási eredményt a 6. melléklet mutatja be./

3.3. Az SNN-eljárás alkalmazása

A 3.2.-ben vázolt szelektiv NN-módszer első közelitésben sikerrel vizsgázott a 3.1.-ben leirt feladat kapcsán. A 30 osztály közül összesen 26-ból gyüjtöttünk mintákat a becsült a priori valószinüségekkel arányos számban. A szekvenciális tanulást megelőzően mindegyik osztályból kisorsoltunk 1-1 elemet és előre archiváltuk. Ezt követően 110 mintával "tanitottunk": 37 esetben az uj minta paramétere egybeesett annak NN-becslésével a megelőző mintákból. /Mivel a szinoptikai gyakorlatból ismeretes, hogy némely HB-osztályok elég nagy hasonlóságot mutathatnak, tehát feltehetően a Bayes-hiba eleve meglehetősen nagy, ezért figyeltük, hogy a második legközelebbi társ, illetve a szinoptikus szempontból hasonló NN-döntések száma mekkora. Ebben az értelemben még további 28 esetről volt elmondható, hogy elég jó becslésnek minősült./ Természetesen a módszer konvergenciája és minőségének becslése még további vizsgálatokat igényel.

4. MAXIMUMÉRTÉKEK ELOSZLÁSBECSLÉSÉNEK IDŐEGYSÉG-PROBLÉMÁJA

4.1. Egy "rövid" szélsebesség-maximum megfigyelési sorozat

Az extremális értékek statisztikai vizsgálata igen fontos szerepet játszik a klimatológiában. A csapadék-maximumokkal foglalkozott többek között Hershfield/1973/, mig mások a szélsebesség maximális értékeinek eloszlását elemezték /Péczely,1965; Bozó, 1964/. A téma kapcsán általában a legnagyobb gond a kellő számu, statisztikailag homogén minta hiánya.

Miskolcról például 1955 óta rendelkezünk olyan szélsebességi adatokkal, melyek a kérdéses statisztikai vizsgálat alapjául szolgálhatnak. A tervezésben /pl. épületek, kémények stb. tervezésében/ általában az a kérdés merül fel, hogy mekkora a valószinüsége egy-egy meghatározott küszöbértéket tullépő szélsebességnek adott időszak" - pl. két év - alatt. Tehát ha az 1955-től 1974-ig terjedő időszakból éves maximumokat képezünk a szokásos módon, akkor összesen csak 20 minta áll rendelkezésünkre a kérdéses statisztika becslésére. Mivel általában az emlitett küszöbértékek az eloszlás "szélén" helyezkednek el, ez azt jelenti, hogy viszonylag kis valószinüségü eseményekről van szó. Ennek fényében méginkább belátható, hogy a megfigyelések emlitett száma meglehetősen kicsi.

A kétéves maximum eloszlásnak becslése azonban nem feltétlenül csak éves maximumértékekből végezhető el. Ha statisztikailag homogén marad a mintasorozat, akkor lecsökkenthetjük ezt az elemi időszakot például félévre. A fő kérdés most az, miként változik meg eközben a becslés minősége.

4.2. A maximumeloszlás az időegység függvényében

Legyen $X_1, X_2, \ldots, X_M, \ldots$ független, azonos eloszlásu valószinüségi változók sorozata. Célunk a

$$/1/ \qquad p_T = P\left\{ \max_{t:\, t-M_1=1,2,\ldots,T} X_t < x \right\}$$

valószinüség becslése valamilyen rögzitett x, M_1 és T -re.

/A gyakorlatban $M_1 > M$, de ennek a modell szempontjából nincs különösebb jelentősége./ Tegyük fel, hogy a T "időintervallum" w egyenlő, egymástkövető τ hosszuságu részre osztható, akkor

$$/2/ \qquad p_T = \prod_{k=0}^{w-1} P\left\{ \max_{\substack{t:\, t-M_1=k\tau+1,\\ k\tau+2,\dots,k\tau+\tau}} X_t < x \right\} = p_\tau^w \quad .$$

Tegyük fel továbbá, hogy M számu megfigyelésünk van /M rögzitett/ és az egyszerüség kedvéért M is osztható τ-val, $M=N\tau$, akkor p_τ becslése:

$$/3/ \qquad \hat{p}_\tau = \frac{1}{N} \sum_{k=0}^{N-1} I_k$$

ahol

$$/4/ \qquad I_k = \begin{cases} 1 & \text{, ha } \max\{X_{k\tau+1},\dots,X_{k\tau+\tau}\} < x \\ 0 & \text{, egyébként .} \end{cases}$$

Következésképpen p_T becslése

$$/5/ \qquad \hat{p}_T = \hat{p}_\tau^w \quad .$$

Mindenekelőtt belátjuk az /5/ becslés aszimptótikus torzitatlanságát /Faragó,1977/.

<u>1.tétel</u>: Fennáll a következő aszimptótikus reláció:

$$/6/ \qquad E\hat{p}_T = p_T + \frac{1}{N}\left\{ -\frac{w(w-1)}{2} p_\tau^w + \frac{(w-1)^2}{2} p_\tau^{w-1} \right\} + O(N^{-2}) \; .$$

<u>Bizonyitás</u>. /3/-ból

$$/7/ \qquad E\hat{p}_T = \frac{1}{N^w} E\left\{ \sum_{\Sigma k_i = w} I_1^{k_1} I_2^{k_2} \dots I_N^{k_N} \frac{w!}{k_1!k_2!\dots k_N!} \right\} =$$

$$= \frac{1}{N^w} \sum_{\ell=1}^{w} J_\ell \, p_\tau^\ell \quad ,$$

hacsak $N \geq w$. Itt

$$J_\ell = \sum_{\substack{\exists k_{i_1},\dots,k_{i_\ell} \neq 0 \\ \text{és } k_i=0 \text{ egyébként,} \\ \text{hogy } k_{i_1}+\dots+k_{i_\ell}=w}} \frac{w!}{k_1!k_2!\dots k_N!} =$$

$$=\binom{N}{\ell}\sum_{\substack{\sum k_{i_j}=w\\ k_{i_j}\geq 1}}\frac{w!}{k_{i_1}!k_{i_2}!\dots k_{i_\ell}!}\ .$$

Speciálisan

/8/ $$J_w=\binom{N}{w}\cdot w!\ ,\qquad J_{w-1}=\binom{N}{w-1}(w-1)\frac{w!}{2}\ .$$

/7/-ben csak a leglassabban konvergáló első két tagot megtartva, majd /8/-ból a behelyettesitéseket elvégezve

$$E\hat{p}_T=\frac{1}{N^w}\left[J_w p_\tau^w+J_{w-1}p_\tau^{w-1}+O(N^{w-2})\right]=$$
$$=p_\tau^w+\frac{1}{N}\left[-\frac{w(w-1)}{2}p_\tau^w+\frac{(w-1)^2}{2}p_\tau^{w-1}\right]+O(N^{-2}),$$

amit /2/ figyelembevételével éppen bizonyitani kellett. □

A fenti tételbőlazonnal következik a p_T becslés aszimptótikus torzitatlansága.

A következő tétel e becslés aszimptótikusan erős konzisztenciáját mutatja meg.

<u>2.tétel:</u> Az /5/ becslés szórására teljesül, hogy

/9/ $$D\hat{p}_T=\frac{1}{N}p_\tau^{2w-1}\left(w^2-w^2p_\tau-\frac{1}{2}\right)+O(N^{-2})\ .$$

<u>Bizonyitás.</u> Fejtsük ki a szórást:

$$D\hat{p}_T=E\left\{\frac{1}{N}\sum_{k=1}^{N}I_k\right\}^{2w}-\left(E\hat{p}_T\right)^2\ .$$

Az első tagra alkalmazható az 1.tétel, ha $N\geq 2w$:

$$E\hat{p}_T^2=p_T^2+\frac{1}{N}\alpha(2w;\ p_\tau)+O(N^{-2})$$

az $$\alpha(w;\ p_\tau)=-\frac{w(w-1)}{2}p_\tau^w+\frac{(w-1)^2}{2}\cdot p_\tau^{w-1}$$

jelöléssel. Ugyancsak az 1.tétel alapján

$$\left(E\hat{p}_T\right)^2 = p_T^2 + \frac{2}{N}\alpha(w;p_\tau)\cdot p_T + O(N^{-2}).$$

Következésképpen

$$D\hat{p}_T = \frac{1}{N}\left[\alpha(2w;p_\tau) - 2\,\alpha(w;p_\tau)\cdot p_T\right] + O(N^{-2}),$$

ami α behelyettesitésével éppen /9/-t adja. □

Vizsgáljuk most meg /9/ függését az időegységtől. Pontosabban, mivel /9/ aszimptótikáját a

$$/10/ \qquad \beta(\tau;\, p_1,T) = p_\tau^{2w-1}\left(w^2 - w^2 p_\tau - \frac{1}{2}\right)$$

együttható határozza meg, helyettesitsük be ide az

$$N=M/\tau \quad , \quad w=T/\tau \quad , \quad p_\tau = p_1^\tau$$

kifejezéseket. Akkor

$$/11/ \qquad \beta(\tau;\, p_1,T) = \tau\cdot p_1^{2T-\tau}\left[\left(1-p_1^\tau\right)\frac{T^2}{\tau^2} - \frac{1}{2}\right]$$

egyenlőség adja meg a vizsgálandó függvénykapcsolatot.

Abban a speciális esetben, amikor a $\tau_1=1$ és $\tau_2=2$ időegységeket kivánjuk összehasonlitani, azt kapjuk, hogy

$$/12/ \qquad |\beta(2;p_1,T)| - |\beta(1;p_1,T)| = p_1^{2T-2}\left[2\left|(1-p_1^2)\frac{T^2}{4} - \frac{1}{2}\right| - p_1\left|(1-p_1)T^2 - \frac{1}{2}\right|\right].$$

Ennek most csupán előjelét vizsgálva a szögletes zárójeles kifejezés értékeit $T=2,3,4$ esetben az 1.ábrán vázoltuk fel. Ebből azt a következtetést vonhatjuk le, hogy $\tau_1=1$ mellett aszimptótikusan jobb az /5/ becslés, mint $\tau_2=2$ -re egy x_1 küszöbérték alatt, illetve egy x_2 / $x_1<x_2<1$ /küszöb felett.

4.3. Gyengén függő minta és egy paraméteres becslés esete

A /4/-ben bevezetett indikátorfüggvényeket /illetve a megfelelő eseményeket/ erősen keverőnek mondjuk /Rosenblatt,1956;

Ibragimov-Rozanov,1970/, ha minden $q>0$ -ra

/13/ $$\rho_n(q) = \sup_{\substack{n_1,\dots,n_q \\ n(n_1,\dots,n_q)=n}} \left| E\left(\prod_{i=1}^{q} I_{n_i}\right) - \prod_{i=1}^{q} EI_{n_i} \right| \xrightarrow[n\to\infty]{} 0 ,$$

ahol $n=n(n_1,n_2,\dots,n_q)=\min\limits_{1\le i,j\le q} |n_i-n_j|$.

<u>3.tétel:</u> Feltéve, hogy

/14/ $$\frac{1}{N}\sum_{n=1}^{N} \rho_n(w) \longrightarrow 0$$

az /5/ becslés aszimptóti kusan torzitatlan és konzisztens marad.

<u>Bizonyitás</u>. Jelölje itt $\hat{r}_T$ az /5/ becslést a /13/-mal adott függő mintára, akkor

$$E\hat{r}_T - E\hat{\rho}_T = \frac{1}{N^w}\left\{ \sum_{n=1}^{\left[\frac{n-1}{w-1}\right]} \sum_{\substack{n_1,\dots,n_w \\ n_i\neq n_j \\ n(n_1,\dots,n_w)=n}} \left[E\left(\prod_{i=1}^{w} I_{n_i}\right) - p_\tau^w\right] w! + \right.$$

$$\left. + \sum_{\nu=1}^{w-1} \sum_{\substack{n_1,\dots,n_\nu \\ n_i\neq n_j \\ k_i\geq 1,\ \Sigma k_i=w}} \left[E\left(\prod_{i=1}^{\nu} I_{n_i}^{k_i}\right) - p_\tau^\nu\right] \frac{w!}{k_1!k_2!\dots k_\nu!} \right\} .$$

Az egyszerüség kedvéért tegyük fel, hogy $N-1$ osztható $(w-1)$ -gyel, $L(w-1) = N-1$, valamint A_n jelölje az összes olyan elrendezést az $1 \le n_1,n_2,\dots,n_w \le N$ $(n_i\neq n_j)$ számokból, melyekre $n(n_1,n_2,\dots,n_w)\geq n$

$$A_n = \binom{w+(N-1-(w-1))n}{w}$$

és B_n azokat, melyekre $n(n_1,n_2,\dots,n_w)=n$

$$B_n=A_n-A_{n+1}=\frac{1}{w!}(w-1)\,w\,N^{w-1} + O(N^{w-2}) \qquad n\le L-1$$

$$B_L=A_L=1 \quad .$$

Ekkor

$$|E\hat{r}_T-E\hat{\rho}_T| \le \frac{1}{N^w}\cdot w!\sum_{n=0}^{L} B_n\,\rho_n(w) \le$$

$$\leq \frac{1}{N}(w-1)\, w \sum_{n=0}^{L} \varrho_n \leq \frac{1}{N}(w-1)\, w \sum_{n=0}^{N} \varrho_n \to 0 .$$

Az 1.tétel és /14/ alapján ebből már közvetlenül belátható $\hat{r}_T$ aszimptótikus torzitatlansága. Részben a 2.tétel bizonyitására támaszkodva a konzisztencia hasonlóképpen igazolható. □

Következésképpen az /5/ becslés alkalmazható a /13/-szerint gyengén függő mintákra is, de az erősen keverő minták definiciója tul általános ahhoz, hogy itt konkrétan megvizsgálható legyen az időegység-probléma.

Paraméteres becslés esetén már többet mondhatunk. Legyen az $X_1, X_2, \ldots, X_M, \ldots$ mintaelemek eloszlása a priori ismerten gaussi m_1 várható értékkel és σ_1 szórással. A paraméteres becslés p_T -re:

$$/15/ \qquad \hat{p}_T = \hat{p}_\tau^w = \left[\int_{-\infty}^{x} \varphi(z;\, \hat{m}_\tau ,\, \sigma_\tau)\, dz \right]^w$$

feltéve itt, hogy σ_τ előre ismert, $\hat{m}_\tau$ pedig a megfelelő empirikus várhatóérték:

$$\hat{m}_\tau = \frac{1}{N} \sum_{k=0}^{N-1} \max \left\{ X_{k\tau+1}, \ldots, X_{k\tau+\tau} \right\} .$$

/ φ a normális sürüségfüggvény $\hat{m}_\tau$, σ_τ paraméterekkel./

<u>4.tétel:</u> A /15/ becslés szórásának aszimptótikus alakja:

$$/16/ \qquad D\hat{p}_T = \frac{1}{M}\, p_1^{(T-\tau)}\, T^2 \frac{1}{2\pi\tau} \exp\left\{ -\frac{(x-m_\tau)^2}{2\sigma_\tau^2} \right\} + O(N^{-2}) .$$

<u>Bizonyitás.</u> Fejtsük Taylor-sorba a sürüségfüggvényt a várhatóérték paramétere szerint:

$$\varphi(z;\hat{m}_\tau ,\sigma_\tau) = \varphi(z;m_\tau ,\sigma_\tau) - \varphi(z;m_\tau,\sigma_\tau)\frac{z-m_\tau}{\sigma_\tau^2}(\hat{m}_\tau - m_\tau) +$$
$$+ \frac{1}{2}\varphi(z;m_\tau,\sigma_\tau)\frac{1}{\sigma_\tau^2}\left[\frac{(z-m_\tau)^2}{\sigma_\tau^2} + 1\right](\hat{m}_\tau - m_\tau)^2 + O(\hat{m}_\tau - m_\tau)^3 .$$

Ezt /15/-be helyettesitve:

$$E\hat{p}_T = p_T + \left\{ w p^{w-1} \frac{1}{2} \int_{-\infty}^{x} \varphi(z;m_\tau,\sigma_\tau)\frac{1}{\sigma_\tau^2}\left[\frac{(z-m_\tau)^2}{\sigma_\tau^2} + 1\right] \right. +$$

$$+\binom{w}{2} p_\tau^{w-2}\left[\int_{-\infty}^{x} \varphi(z;m_\tau,\sigma_\tau)\,\frac{(z-m_\tau)}{\sigma_\tau^2}\right]^2\Bigg\}\,\frac{\sigma_\tau^2}{N} + O(N^{-2}) .$$

Jelölje $\alpha(w;p_\tau)$ ismét az N^{-1} tag együtthatóját. Vegyük észre, hogy

$$\int_{-\infty}^{x} z\,\varphi(z;m_\tau,\sigma_\tau)\,dz = -\sigma_\tau^2\,\varphi(x;m_\tau,\sigma_\tau) + m_\tau p_\tau$$

és igy

$$D\hat{p}_\tau = \frac{1}{N}\left[\alpha(2w;p_\tau) - 2\,\alpha(w;p_\tau)\,p_\tau + O(N^{-2})\right. =$$

$$= p_\tau^{2w-2}\,w^2\,\varphi^2(x;m_\tau,\sigma_\tau)\frac{\sigma_\tau^2}{N} + O(N^{-2}) =$$

$$= \frac{1}{M}\,p_1^{2(T-\tau)}\,T^2\frac{1}{2\pi\tau}\exp\left\{-\frac{(x-m_\tau)^2}{2\sigma_\tau^2}\right\} + O(M^{-2}) .$$

Ezzel a tételt bizonyitottuk.

A 3. tételből azonnal következik a /15/ becslés aszimptótikus torzitatlansága és könnyen levezetheztő ennek aszimptótikus konzisztenciája is.

Jelölje /16/-ban M^{-1} együtthatóját $\beta(\tau;p_1,T)$. Megvizsgálva a $\tau_1=1$, $\tau_2=2$ speciális esetet, azt kapjuk, hogy

$$|\beta(2;p_1,T)| - |\beta(1;p_1,T)| \geqslant 0$$

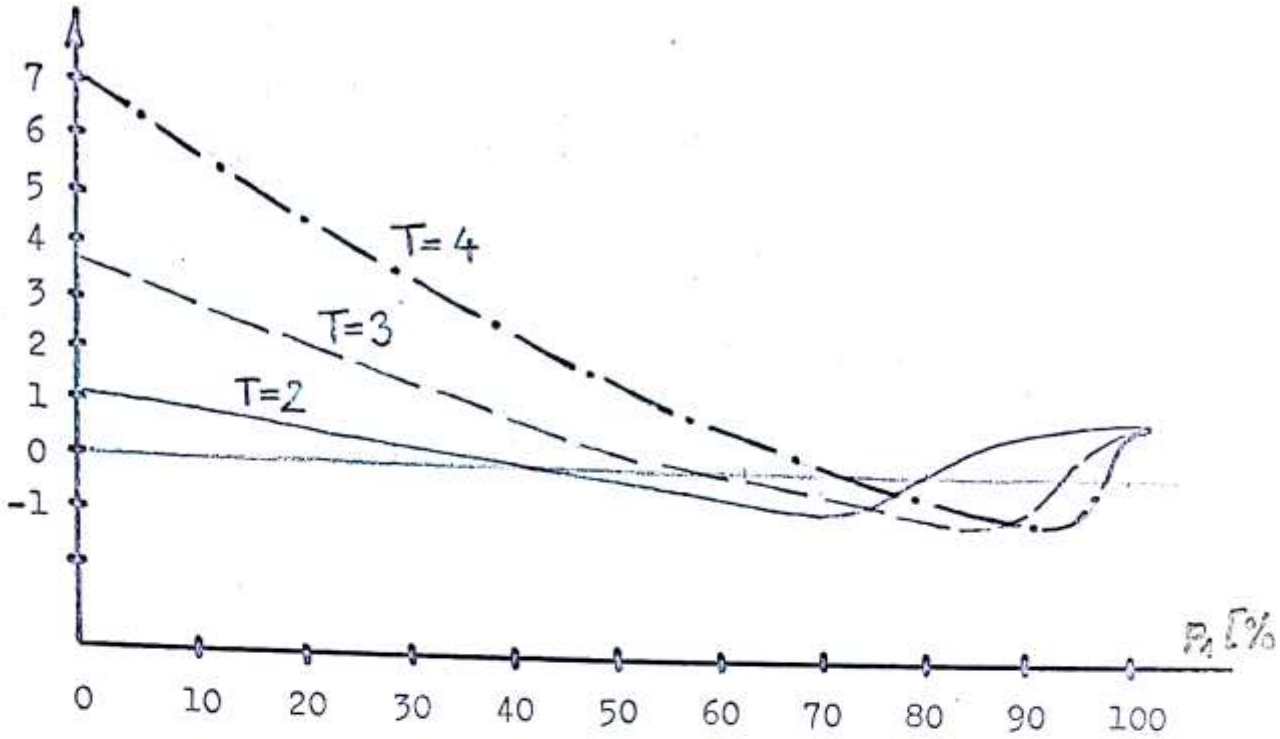
/felhasználva az $m_2=m_1+\sigma_1/\sqrt{\pi}$ és $\sigma_2^2=(1-1/\pi)\,\sigma_1^2$ relációkat/, ami azt jelenti, hogy általában a kisebbik időegységü paraméteres becslés jobb aszimptótikusan.

4.4. Egy szélsebességi mérési sorozat analizise féléves elemi maximumértékek alkalmazásával

A 4.1.-ben emlitett mintasorból féléves maximumértékeket számitottunk a szokásos éves értékek helyett. A 4.2. és 4.3. pontban bemutatott eredmények alapján helyesebb ennek az időegységnek az alkalmazása, ha az $N_1=20$ éves érték helyett kapott $N=40$ féléves minta továbbra is statisztikailag homogén marad.

Ennek érdekében /Faragó,1977/ -ben először az első illetve második félévi értékekből álló kétdimenziós mintasorozat normalitását igazoltuk statisztikailag /Hald,1967/, majd a korrelációs együtthatóra t-statisztikát számitottunk / $\hat{\rho}$ =0,22 ; t=0,96; 18 szabadsági fok/, ami alapján 95%-os szignifikanciaszinten elfogadható volt a korrelálatlanság, azaz a függetlenség hipotézise. /Az itt alkalmazott statisztikai eljárásokkal kapcsolatosan Lehmann/1966/, Vincze/1973/ és Prékopa/1972/ munkáira hivatkozunk./

A homogenitás ellenőrzése érdekében előbb F-teszt következett /F=1,93/, majd az empirikus várható értékek egyenlőségének igazolására ismét Student-próba /t=1,07 ; a szabadsági fokok száma 38/, amelynek következtében 90 illetve 95%-os szinten elfogadtuk a két szórásnégyzet illetve a két várhatóérték azonosságának hipotézisét. Mindezek alapján a féléves időegységekre kiszámitottuk az empirikus várhatóértéket és szórást és a továbbiakban az e paraméterekkel adott normális eloszlásunak feltételezett féléves időegységü szélmaximum képezte a tetszőleges számu év alatt a tetszőleges küszöbérték tullépési valószinüségek becslésének alapját.



1.ábra

ZÁRÓMEGJEGYZÉSEK

A meteorológiai megfigyelések rendszeres archiválása, a meteorológiai szolgáltatások összessége, a kutatások sorában pedig elsősorban a statisztikai vizsgálatok széles köre számitógépes meteorológiai adatbank és az ezzel kapcsolatban álló adatbankszervező és feldolgozói programok rendszerének létrehozását igényli. Ebbe a témakörbe vágó kérdésekkel foglalkozik többek között a korábbi helyzetről átfogó képet adó WMO/1964/ kiadvány, vagy például Boriszenkov-Romanov/1969/ és Craddock /1970/ munkája. Az emlitett rendszer megteremtésével kapcsolatos alapelveket, illetve az eddig megvalósult lépéseket egy tanulmányban foglaltuk össze /Faragó et al.,1976/.

Az adott tipusu feldolgozás szerint orientált adatbankok általános ismérvein - gyors adathozzáférhetőségen, az adatok különféle rendezettségü elérésének egyszerü szerkesztésén, állománykiegészitési tevékenység kiépitésén stb. - tulmenően a staisztikai természetü általános /például alkalmazott klimatológiai/ és egyedi /általában klimatológián vagy prognosztikán belüli kutatási/ célok egyéb szempontokat is előirnak. A legfontosabb ezek sorában, hogy különféle - többé-kevésbé szekvenciális - mintavételezéseket könnyen, gyorsan és mi több valamilyen speciális "lekérdező" nyelv segitségével egyszerüen megadhatóan végezhessünk el. Emlitésreméltó ezen tulmenően a "származtatott" állományok problémája is. Nagykiterjedésü adatmezők, vektorok esetén, mint például a 2. fejezetben láttuk, a statisztikai vizsgálatokat az eredeti minták előzetes transzformációja /adatredukció/ előzheti meg. Ez pedig gyakorlatilag azt jelenti, hogy a megadott "eredeti" adatállományok mellett az adatbankba felvesszük ezek bizonyos változatait is.

A statisztikai feldolgozás-orientált /pl. távprognosztikai, klimatológiai/ adatbank mellett várhatóan más jellegü meteorológiai adatbankok is kidolgozásra kerülnek majd az Országos Meteorológiai Szolgálat néhány éven belül esedékes nagykapacitásu R-kategóriáju számitógépén. A jelenlegi előkészitő szakaszban,

korábban az SzKI IBM 370/125 tipusu gépén dolgoztunk DOS operációs rendszreben, ujabban pedig a SzÜV 370/145 -ös gépén OS rendszerben. A statisztikai feldolgozások szempontjából elsősorban a rendelkezésünkre álló SSP /Scientific Subroutin Package/ tudományos programkönyvtár érdemel külön emlitést, mely egy sor statisztikai becslési és hipotézisvizsgálati eljárást tartalmaz számitógépes /Fortran/ szubrutin formájában. Például a 2.3. pontban vázolt modell messzemenően hasznositja az SSP-ben adott különféle eljárásokat.

A jelenleg folyó statisztikai vizsgálatok szemszögéből is már alapvető jelentőségü a számitógépes feldolgozási lehetőség, a későbbiek során várható számitógép installálás után pedig a jelenleg sokszor csak elméletileg kidolgozott, vagy más kül- és belföldi kutatók által analizált eljárások tág köre lesz nagyobb adatseregre alkalmazható, illetve adaptálható.

A távprognosztikával kapcsolatos kutatások az Országos Meteorológiai Szolgálatnál jelenleg a Távprognosztikai Csoport keretében folynak, melyek szerves részét képezik az 1.1., 2.1. és 3.1. pontban leirt feladatok. E vizsgálatokat Czövek Istvánné tudományos munkatárssal, meteorológussal közösen végezzük; a globális regressziós modell kidolgozása és számitástechnikai realizálása során Szlachányi Kornélné tudományos segédmunkatárs is bekapcsolódott a kutatásokba. A programozási /főként PL1-nyelvü programszerkesztési / munkákban nagy szerepe van Salamon Lászlóné rendszerszervezőnek. Az adatzsugoritásokkal kapcsolatos elmélet és a regressziószámitási alkalmazások terén korábban Gulyás Ottó kandidátussal, a Módszertani Csoport vezetőjével végeztünk közösen kutatásokat. A 4.1. pontban leirt feladatot Szalma Jánosné tudományos munkatárs, az Éghajlati Tájékoztató Osztály tagja vetette fel.

MELLÉKLETEK

Hivatkozások

Bagrov, N.A.,1959: Analiticseszkoje predsztavlenyije poszledo vatyelnosztyi meteorologicseszkih polej poszredsztvom jesz jesztvennih ortogonalnix szosztovljajuscsih. Trudi CIP, 74

Bagrov, N.A.,1966: Predszkazanyije meszjacsnovo kolicsesztva oszadkov. Meteorologija i Gidrologija,7.

Bagrov, N.A.,1968: O nyekotorih oszobennoszttyah korrelacion-novo analiza i ih primenyenyija k prognozam pogodi, Meteor logija i Gidrologija,1.

Bagrov, N.A., Mjakiseva,N.N.,1969: Prognoz tyemperaturi na meszjac. Trudi GMC, 44.

Boriszenkov, E.P., Romanov,M.A.,1969: Algoritmi i programma sztatiszticseszkoj obrabotki informacii na EVM. Gidrometeo-izdat,Leningrad.

Bozó P.,1974: A szélmaximum eloszlásának becsléséről. Kézirat. Orsz. Meteorológiai Szolgálat, Budapest.

Cover,T.M.,Hart,P.E.,1967: Nearest neighbour pattern classification. IEEE Trans. on Inf. Theory, 1.

Cover,T.M.,1968a: Rates of convergence of nearest neighbour decision procedures. First Annual Hawaii Intern. Conf. on Systems Theory.

Cover,T.M.,1968b: Estimation by the NN rule. IEEE Trans. on Inf. Theory,1.

Cover,T.M.,1969: Learning in pattern recognition. Metodologies of pattern recognition, S.Watanabe ed. Academic Press, New York.

Craddock,J.M.,1958: Research in the Meteorological Office concerned with long-range forecasting. Meteorological Magazine, 3.

Craddock,J.M.,1970: Work in synoptic climatology with a digitized data bank. Meteorological Magazine, 2.

Craddock,J.M., Colgate,M.G.,1974: The use of eigenvectors for smoothing and prediction. Bull. of the Inst. of Math. and its Appl., 5-6.

Czelnai R.,Gandin,L.S.,Zaharev,V.I.,1976: Sztatiszticseszkaje sztruktura meteorologicseszkih polej. Orsz. Meteorológiai Szolgálat, Budapest.

Dujceva,M.A.,Pegy,D.A.,1970: O mnogoparametricseszkoj szheme prognoza anomalii szrednyej meszjacsnoj temperaturi vozduha. Trudi GMC, 64.

Epstein,E.S.,1969: Stochastic dynamic prediction. Tellus,6., 739-759.

Faragó,T.,1973: K posztanovke zadacsi o predszkazivanyii sz pomoscsju konyecsnovo verojatnosztnovo avtomata.Vics.tehnika i voproszi kibernetiki, 9.,49-61.

Faragó T.,1974: Kresenyiju zadacsi o predszkazivanyii sz pomoscsju konyecsnovo verojatnosztnovo avtomata. Vics.tehnika i voproszi kibernetiki, 11.,109-130.

Faragó T.,Gulyás O.,1973a: Some method of expansion type feature extraction in pattern recognition. Proc. of Prague Symp. on Asymptotic Statistics, 69-87.

Faragó T.,Gulyás O.,1973b: The feature extraction in pattern recognition. Proc. of Conf. on Information Theory, Tallin.

Faragó T.,Gulyás O.,1974: A regressziós tipusu extrapoláció. Meteorológiai Tanulmányok, 2., Országos Meteorológiai Szolgálat,Budapest.

Faragó T.,Gulyás O.,1975: A valószinüségi sürüségfüggvény becs és illesztése. Meteorológiai Tanulmányok, 3., Országos Meteorológiai Szolgálat, Budapest.

Faragó T.,1975: A mintatér bizonyos transzformációiról. Problemi peredacsi informacii, 1.,102.107.

Faragó T. et al.,1975: Az analógia elvén alapuló prognosztikai módszerek matematikai modellje. Időjárás, 3., 166-176.

Faragó T.,1975: Egy kombinált regressziós becslésről. Időjárás 6.,366-371.

Faragó T. et al.,1975: A hosszabb érvényességi idejü prognózisokkal kapcsolatos kutatások és azok számitástechnikai vonatkozásai, "Az OMSz adatfeldolgozó és adattároló rendszer alapjainak kialakitása R-50-kategóriáju számitógépre" c. OMFB-tanulmányban.

Faragó T.,Kaba M.,1976: Sztatiszticseszkij prognoz meszjacznih szrednyih temperatur i ocenka Bajeszova riszka. Időjárás, 6., 313-325.

Faragó et al,1976: Távprognosztikai adatbank és kezelőrendszer jelene, az alkalmazói programokkal való kapcsolatai és átfogó lános kiépitésének alapvető szempontjai. OMFB-tanulmány, Orsz. Meteorológiai Szolgálat, Budapest.

Faragó T., 1977: Introduction of characteristic function in stochastic-dynamic prediction. beadva a Tellus-hoz.

Faragó T.,1977: On the estimation of the probability distribution of maximum values and the statistical analysis of a wind velocity sample. Időjárás, 1., 27-39.

Fleming,R.J.,1971: On stochastic dynamic prediction. Monthly Weather Rev., 11.,851-872.

Freiberger,W.,Grenander,U.,1965: On the formulation of statistical meteorology. Rev. of Intern. Stat. Institute. 1., 59-68.

Fritz J.,1971: A Karhunen-Loéve sorfejtésről. Szemináriumi Közl., Távközl. Kutató Intézet, Budapest.

Fritz J.,1975: Distribution free exponential error bound for nearest neighbour pattern classification. IEEE Trans. on Information Theory, 1.

Fu,K.S.,1968: Sequential methods in pattern recognition and machine learning. Acad.Press, New York.

Gandin, L.S.,1967.: O primenyenyii metoda kanonocseszkih korrelacij v meteorologii. Trudi GGO, 203., Leningrad.

Gates, G.W.,1972: The reduced NN-rule. IEEE Trans. on Information Theory, 18.

Glahn,H.,1968: Canonical correlationand its relationship to discriminant analysis and multiple regression. J. Atmos. Sci., 1.

Gleeson,T.A.,1966: A causal relation for probabilities in synoptic meteorology. J.Appl.Meteorology, 3.,365-368.

Hart,P.E.,1968: The condensed NN-rule. IEEE Trans.on Information Theory, 3.

Hershfield,D.M.,1973: On the probability of extreme rainfall events. Bull.Amer. Meteorological Soc.,1013-1022.

Hess,P.,Brezowsky,H.,1952: Katalog der Grosswetterlagen Europas. Berichte das Deutschen Wetterdienstes in der US Zone, 33.

Ibragimov,I.A.,Rozanov,J.A.,1970: Gausszovszkije szlucsajnije processzi. Nauka, Moszkva.

Jefimov,A.A.,1972: K raszcsetu obobscsennik szfericseszkih funkcij. Trudi GGO, 380.

Judin,M.I.,1967: Fiziko-sztatiszticseszkije metodi prognozov pogodi i vozmozsnosztyi ih vnyedrenyija. Meteorologija i Gidrologija, 11.

Hald,A.,1967: Statistical theory with engineering applications. Wiley, New York.

Kapovits R. et al.,1975: A légköri folyamatok szimulációján alapuló középtávu előrejelzési modell. Reprint. Orsz. Meteorológiai Szolgálat, Budapest.

Koppány Gy.,1974: Az analógiák extrapolációjának felhasználása a havi és évszakos előrejelzések készitéséhez. Időjárás, 2.,97-108.

Lehmann,E.L.,1966: Testing Statistical Hypotheses. Wiley, New York.

Mackenhauer,B.,1974: On the present state of spectral methods in numerical integrations of global atmospheric models. Rep.Intern.Symp.on Spectral Methods in Numerical Weather Prediction; the GARP Programme on Num. Exp., 7.

Nyberg, A.,1975: An experiment in forecasting monthly mean temperature in Stockholm. Tellus, 1.

Péczely,Gy.,1957: Grosswetterlagen in Ungarn. Orsz. Meteorológiai Szolgálat Kisebb Kiadványai, 30., Budapest.

Prékopa,A.,1972: Valószinüség elmélet. Müszaki Kiadó,Budapest.

Rényi,A.,1970: Valószinüségszámitás. Tankönyvkiadó,Budapest.

Rosenblatt,M.,1956: A central limit theorem and the strong mixing condition. Proc.Nat.Acad.Sci. USA.

Smagorinsky,J.,1969: Problems and promises of deterministic extended range forecasting. Bull.Amer.Meteor.Soc.,5., 286-312.

Szonyecskin, D.M.,1976: Obosznovanyije sztatisztiki v meteorológii. Trudi GMC, 181.

Tatarskij, V.I.,1969: Iszledovanyije dinamicsoszkih uravnyonyij pri verojatnoszthom prognoze baricseszkovo polja. Izv. AN SzSzSzR, Fizika atmoszferi i okeana, 3.

Ter-Mkrtcsan,M.G.,1970: O primenyenyii diszkriminantnovo analiza dlja ulucssenyija sztatiszticseszkih prognozov po metodu mnozsesztvennoj regresszii. Trudi GMC,64.

Thom,H.C.S.,1966: Some methods of climatological analysis. WMO Techn. Note, 81.

Vincze,I.,1975: Matematikai statisztika .Müszaki Kiadó,Budapest.

Wagner,T.J.,1971: Convergence of the NN-rule. IEEE Trans. on Information Theory, 5.

Watanabe, S.,1972: Karhunen-Loéve expansion and factor analysis. Trans.of the Fourth Prague Conf on Inf. Theory.

WMO Techn. Note, 71.,1966: Statistical analysis and prognosis in meteorology.

WMO Techn. Note, 74.,1967: Data processing by machine methods.

Faragó T.,Györfi L.,1974: Error distortion function in two - hypothesis decision problem. Proc. of IFAC Conference, Budapest.

Faragó T.,Györfi L.,1975: On the continuity of the error distortion function for multiple hypothesis decisions. IEEE Trans. on Information Theory, 6., 458-460.