Informační Bulletin
Zvláštní vydání č. 1
České Statistické Společnosti
Toto zvláštní číslo je věnováno konferenci COMPSTAT a statistickému software

OHLÉDNUTÍ ZA COMPSTATEM 90

Jaromír ANTOCH

Letošní, v pořadí již devátý ročník sympozia *COMPSTAT* se konal ve dnech 9.-15.9.1990 v Dubrovníku, Jugoslávie. Pod záštitou *Mezinárodní asociace pro výpočetní statistiku* (*IASC*) jej zorganizovala Universita v Záhřebu a zúčastnilo se jej 273 odborníků z celého světa. Mezi oficiálními hosty byl též Prof. G. Kulldorff (S), předseda *International Statistical Institute* (*ISI*). Přehled účastníků dle zemí podal v minulém čísle Zpravodaje T. Havránek spolu s jejich charakterizací, takže jej není třeba opakovat.

V tomto krátkém příspěvku bych rád shrnul některé své dojmy a alespoň krátce tak přiblížil *COMPSTAT 90* těm kolegům, kteří se jej neměli možnost zúčastnit. Hlavní pozornost bude soustředěna na odborný program. Důvodem je jak omezený rozsah Zpravodaje, tak to, že na zhodnocení programů jež jsme měli možnost vidět a zároveň si vyzkoušet slíbili se zaměřit další českoslovenští účastníci sympozia.

Odborný program

Záběr sympozia byl, jak už je na *COMPSTAT*ech tradičně zvykem, značně široký. Celkově bylo v programu pamatováno na 8 zvaných přednášek, 2 výukové semináře, 43 sdělení, 82 krátkých sdělení, 20 ukázek programů účastníků a panelovou diskusi. Bohužel, vzhledem k nepřítomnosti některých autorů nebylo nakonec vše presentováno. Vedle toho probíhala neustálá předvádění produktů velkých komerčních firem, tj. *BMDP*, *NAG*, *SAS* a *SPSS*. Stručný přehled od-

borné problematiky shrnutý dle témat lze nalézt v tabulce 1.

Struktura a obsah příspěvků jsou poměrně dobrým indikátorem rozložení současného zájmu v oblasti výpočetní statistiky. Podobně jako v minulých letech, i *COMPSTAT* 90 byl dominován především příspěvky z oblasti analýzy mnohorozměrných dat, pracemi zaměřenými na rozvoj algoritmů a statistických programů, jakož i problematikou expertních systémů a umělé inteligence. Mnohem menší pozornost než se čekalo byla věnována problematice databází.

Po tomto stručném úvodu mi dovolte, abych se u jednotlivých programových celků zastavil poněkud podrobněji.

Zvané přednášky

Na sympoziu odeznělo sedm z osmi zvaných přednášek. Nejen mně, ale převážné většině účastníků s nimiž jsem o tom hovořil, se z pozvaných řečníků nejvíce líbilo vystoupení T. Havránka na téma *O metodách pro vyhledávání modelů*. Bylo tomu především proto, že se jednalo o jedinou vyžádanou přednášku, jež si kladla za cíl globálně shrnout základní myšlenky určité obecné metodologie. Konkrétněji, postupy vhodné pro vyhledávání modelů v různých oblastech matematické statistiky. Při výkladu přitom autor neopomněl:

- ukázat spojení s nejnovějšími výsledky metod umělé inteligence;
- popsat prostor, který pro realizaci a podstatné urychlení popisovaných postupů otevřely paralelní počítače, metody paralelního zpracování dat a s nimi spojené výsledky informatiků ;
- zmínit řadu doposud otevřených problémů a možných aplikací, jež na své vyřešení doposud čekají ;
- na jedné straně stručně připomenout historii popisované metodologie, včetně zmínky o nezdarech a slepých uličkách jež stály v cestě, a na druhé straně účastníkům přiblížit alespoň některé úspěšně implementované a navzájem velmi různorodé aplikace.

S touto koncepcí ostře kontrastovaly jinak velmi dobře připravené a přednesené přednášky P.J. Rousseeuwa (B) *Rychle spočitatelné odhady v regresi s vysokým bodem zvratu* a I.G. Žurbenka (SU) *Spektrální analýza nestacionárních časových řad*, v nichž se oba autoři zaměřili na určitou velmi úzkou oblast svého zájmu. Přesto, že se ji snažili rozebrat do detailů i přes existenci od-

povídajícího programového vybavení a konkrétní aplikace mi připadá, že tyto přednášky neměly ten dopad jaký by mohly mít na specializovaných konferencích (důvodem bylo především značně různorodé složení posluchačů). Ale tam by asi oba autoři museli poněkud přidat, neboť by jim "více kolegů vidělo do talíře". Navíc oběma přednáškám chybělo hlubší zasazení popisovaných metod v rámci robustní statistiky, resp. časových řad.

Přednáška K. Neumanna (DDR) na téma *Kooperativní zpracování dat - výzva pro statistiku a organizování databází* přinesla řadu velmi obecných myšlenek, na jejichž základě by se mohlo přistupovat ke konstrukci oficiálních statistických informačních systémů. Autor se ostatně touto problematikou zabývá již řadu let. To, co celému přístupu asi nejvíce chybí je především mravenční práce desítek lidí, kteří by tyto jinak zajímavé myšlenky uvedli v život a ukázali jejich skutečné výhody při konstukci nejen oficiálních (statistických) informačních systémů a databází.

Na rozdíl od předchozího řečníka se T.S. Arthanari (IND) ve své přednášce *Optimalizace ve statistice - současné trendy* zaměřil na presentaci použití výsledků matematického programování pro řešení statistických problémů v kontrole kvality, isotónní mediánové regresi, stratifikaci či Taguchiho přístupu k plánování experimentů. Tato přehledně uspořádaná přednáška se i pro nespecialistu v oboru po téměř celou dobu výkladu dobře sledovala. Jak však bylo poznamenáno v koreferátu a následné diskusi, jednalo se spíše o pěkný přehled již dávno známých metod než o současné trendy.

Ti, kteří se v pátek ráno sešli na poslední vyžádané přednášce, a bylo jich více než 200, aby si vyslechli nejen jednotící pohled na výpočetní prostředí pro statistiku v uplynulých třiceti letech, ale především určitou vizi (alespoň blízké) budoucnosti, odešli vesměs zklamáni. To s čím nás J.M. Chambers (USA) seznámil byly vesměs notoricky známé informace. O budoucnosti přitom nepadlo téměř ani slovo. A tak i když ani ve sborníku účastníci nenašli byť jeden řádek, nelze se příliš divit jejich poznámkám, že si po čtvrteční závěrečné večeři měli raději pospat.

Sdělení a krátká sdělení

Jak vyplývá z tabulky 1, pro *COMPSTAT* 90 bylo programovým výborem vybráno 43 půlhodinových sdělení a 82 krátkých čtvrthodinových sdělení. Početně silné zastoupení mezi nimi měly především přednášky účastníků z France, Holandska, Itálie, Spojených států a Velké Británie. Zde je zajímavé si všimnout, že např. z 12 účastníků z USA jich 11 mělo příspěvek. Takovéto "úspěšnosti" nedosáhli statistici z žádné jiné země, z níž se zúčastnilo alespoň 5 zástupců. To svědčí nejenom o kvalitě příspěvků, ale též o silném tlaku v USA nedovolit (si) účast na byť sebedůležitějším konferenci není-li co presentovat.

Prohlédneme-li si podrobněji tabulku 1, vidíme, že nejvíce referátů bylo z oblasti klasifikace a analýzy mnohorozměrných dat. Zastoupena zde byla především francouzská a italská škola. První z nich zaměřená především na různé typy projekcí mnohorozměrných dat za účelem získání jejich jednodušších representací, resp. na speciální úlohy a problémy analýzy hlavních komponent, zatímco druhá zabývající se spíše problematikou škálování, analýzou mnohorozměrných kontingenčních tabulek apod. Poměrně vysoký byl též počet příspěvků týkajících se expertních systémů.

Mezi sdělení byly po delší diskusi (v rámci programového výboru) zařazeny i dva příspěvky z oblasti komerčního programového vybavení, tj. informace o verzi 4.0 systému *GLIM* a o tvorbě statistických modelů v jazyce *S+*. Důvodem byl velký význam, který oba tyto systémy sehrály při rozvoji a ovlivnění současných trendů ve výpočetní statistice. O zájmu mezi účastníky *COMPSTAT*u nejlépe svědčily přeplněné přednáškové místnosti. K chvále přednášejících je třeba zdůraznit, že obě přednášky byly věnovány pouze statistice a poskytly posluchačům řadu velmi zajímavých a podnětných informací.

Vysoký byl též počet krátkých sdělení věnovaných tématu algoritmy a statistické programové vybavení. Bylo tomu především proto, že řada autorů se ve svém sdělení snažila "udělat reklamu" programovému produktu, který zpravidla souběžně presentovala v předváděcí sekci.

Výukové semináře (tutorials)

I na tomto *COMPSTAT*u se pokračovalo v tradici výukových seminářů (tutorials). K jejich přípravě byli vyzváni E. Diday (F) pro téma *Klasifikace* a A.J. Westlake (GB) pro téma *Řízení statistických databází a relační modely*. Je třeba říci, že oba semináře byly více než hojně navštíveny. Bohužel, tento velký zájem poněkud poškodil řečníky v souběžně se konající sekci.

Zastavme se nejprve alespoň krátce u prvního ze seminářů. E. Diday se v něm především pokusil popsat obecnou metodologii representací dat a klasifikátorů umožňující jednotný pohled na řadu postupů užívaných v klasifikaci. Úvodem přednášející zavedl pojem *klasifikační prostor*, dovolující jednoduše popsat vztahy mezi jednotlivými shluky a pojem *representační prostor*, poskytující nástroj k jejich representaci. Dále byla podrobně popsána *metoda dynamických mraků* vhodná především pro situace, kdy pracujeme s velkou populací pro niž sice neexistuje statistický model pro celek, je však vhodný a velmi přesný pro určité subpopulace. Druhá část přednášky byla věnována problematice klasifikace v situaci, kdy má uživatel alespoň částečnou apriorní představu o typech shluků které by rád získal a je-li schopen předem zadat alespoň částečnou hierachii mezi shluky. Poslední část přednášky byla věnována grafickým representacím shluků. Autor se zaměřil především na tzv. *pyramidy*, umožňující grafickou representaci částečně se překrývajících shluků typických například pro chaotická data, a s nimi úzce spojené pojmy *ultrametrika* a *index nepodobnosti*.

V druhém semináři se přednášející zaměřil na úplně odlišnou oblast, totiž na *relační databázové systémy*. Po zavedení pojmu *databáze*, jejích vlastností a potřebné terminologie se autor podrobně zaměřil na *relační modely*. Po krátkém shrnutí jejich historie byla podána formální definice relačních modelů a blíže osvětlena role operátorů a klíčů. Dále se A.J. Westlake zaměřil na jazyky typu *SQL*. Po tomto spíše teoretickém úvodu následoval podrobný popis relačních systémů. V třetí části se přednášející podrobněji zaměřil na problematiku relačních databází, jejich integritu, potřebu normalizace a na relační modelování. Zároveň byly presentovány některé statistické aplikace. Nezapomnělo se ani na shrnutí určitých stávajících omezení relačních databází.

Panelová diskuse

Místo času pro zaplavání si před středečním obědem bylo programovým výborem pamatováno na panelovou diskusi na téma *Jak by mělo vypadat statistické programové vybavení budoucnosti ?* Připravil ji Y. Dodge (CH), který k přednesení krátkých úvodních vstupů majících za úkol diskusi rozproudit pozval J.M. Chamberse (USA), D. Handa (UK), P. Naeva (D), R.W. Payna (GB) a P.J. Rousseuwa (B) jako representanty poměrně širokého spektra názorů. Po jejich vystoupení bylo slovo dáno též dalším účastníkům sympozia. Diskuse mimo jiné potvrdila známý fakt, že převážná většina uživatelů v praxi používá především ty metody, pro které má k dispozici potřebné programové vybavení. Řada vystupujících v této souvislosti ostře kritizovala velké komerční softwareové společnosti za malou pružnost při rozšiřování svých programových celků o známé a teoreticky dobře prozkoumané postupy, jež jsou bohatě zdokumentovány v literatuře a k nimž již existuje programové vybavení, pocházející vesměs od autorů jednotlivých metod. Jako typický příklad byly uváděny např. robustní metody. Především právě tato část diskuse mi připomněla známé spory z *ROBUST*ů o zneužívání statistiky lékaři a jim podobnými uživateli. A tak i když se převážná většina diskutujících ve svých názorech shodla, ve větším počtu chyběli především *oni zneuživatelé*, tj. producenti a tvůrci velkých programových systémů. Na druhé straně je třeba uvést, že již 27.9.1990 jsem obdržel poměrně dlouhý dopis od firmy *SAS*, ve kterém se výpadům jež zazněly během panelu brání, a na řadě konkrétních příkladů dokazují co vše v uplynulém období na tomto poli jejich společnost udělala.

Řada dalších řečníků se zaměřila úplně jiným směrem, tj. na otázku vytváření optimálního *výpočetního prostředí pro statistiku*. To jest, velmi zjednodušeně řečeno, celek skládající se z počítačů a instalovaného programového vybavení, který by umožnil nejen (statistickou) analýzu dat, ale i plnění každodenních běžných úkolů - počínaje psaním dopisů a zpráv, a konče výběrem informací z velkých databází či komunikací s kolegy apod.

Do obou oblastí částečně zapadla i vystoupení účastníků hovořících o svých snech. Reálnou odpovědí jim bylo zpravidla zdůraznění faktu, že *převážná většina podstatných vylepšení statistic-*

kého výpočetního prostředí byla a je především výsledkem pokroku v obecném počítání a v technickém rozvoji, nikoliv výsledkem nejnovějšího vývoje moderní matematické statistiky. Zdá se přitom, že tomu tak bude i v budoucnu.

Výstava komerčního programového vybavení

Jak jsem uvedl již v úvodu, *COMPSTAT*u 90 se z velkých komerčních softwareových společností zúčastnily firmy *BMDP*, *NAG*, *SAS* a *SPSS*. Jelikož na rozdíl od minulosti tentokrát nebylo zrealizováno napojení na sálové počítače, místo toho měli jednotliví zástupci s sebou ať již personální počítače (*IBM PC* či *PS/2* a *Macintosh*), resp. pracovní stanici *HP* (*SAS*).Tím též byly předurčeny programové produkty, jimiž se vystavovatelé presentovali :

BMDP .... verzí 90 pro *IBM PC*;

NAG .... verzemi 3.77 a 4.0 *GLIM*u *IBM PC*,
*GENSTAT*em pro *IBM PC*,
programem *MLP* pro nelineární statistické modely,
programem *TSA* pro analýzu časových řad,
knihovnami fortranských podprogramů;

SAS .... verzí 6.06 pro *IBM PC* & *PS/2* a pro *Macintosh*,
verzí 6.07 pro pracovní stanici *HP*;

SPSS .... verzí 3.1 pro *IBM PC* & *PS/2*,
verzí 4.0 pro *Macintosh*.

Samozřejmě, každá z firem s sebou měla více či méně kompletní sadu manuálů nejen pro vystavované verze, ale i pro verze pracující pod *UNIX*em na počítačích vyšších tříd apod., jakož i podrobné informační materiály o jednotlivých verzích.

U všech čtyřech vystavovatelů bylo možné pozorovat důraz především na následující vylepšení stávajících systémů :

- snahu po maximálním zjednodušení obsluhy - zvláště markantní je tomu u *BMDP* a *SAS*u ;
- úpravu programů ve verzích pro *PC* tak, aby bylo možné *automaticky* využívat rozšířené paměti *EMS* pro zvětšení rozsahu zpracovávaných dat (má-li ji uživatel k dispozici) ;
- doplnění o nové programy a moduly.

Co se na druhé straně bohužel téměř vůbec nezměnilo k lepšímu je nevyhovující kvalita manuálů, nikoliv po tiskové stránce.

Zvláště u *SAS*u a *SPSS* sice uživatel dostane k dispozici několik tisíc stran textu, ale zajímá-li ho jak jsou počítána studentizovaná residua či jaká metoda a jak byla implementována pro nelineární regresi apod., musí si to vesměs odvodit metodou pokusu a omylu sám. Osobně to považuji za velmi závažný nedostatek. Nejlépe mi pak z tohoto hodnocení stále vycházejí dva *štíhlé manuály*, resp. *hubené sešity* s řešenými příklady, firmy *BMDP*.

Jelikož přehledu programů pro statistiku, ať již od velkých komerčních firem či jednotlivých účastníků a jejich zhodnocení by měl být v tomto čísle Zpravodaje věnován článek J. Militkého a J. Tvrdíka, nebudu se u tohoto bodu dále zdržovat.

Nicméně na závěr odstavce bych rád dodal, že výrazně vzrostl zájem velkých firem o prodej svých výrobků u nás. I když ne úplně, tak alespoň částečně odpadla řada omezujících opatření, mimo jiné ze strany nechvalně známého výboru *COCOM* apod. Obzvláště velkou aktivitu s cílem získat rozhodující vliv na našem trhu vyvíjí *SAS*, připravující otevření svého zastoupení v Bratislavě, a později snad i v Praze.

Pár slov závěrem

Na schůzi vedení evropské sekce *IASC* bylo rozhodnuto, že příští *COMPSTAT* se uskuteční ve dnech 24.-28.8.1992 v Neuchatelu, Švýcarsko. Přípravou byl pověřen Y. Dodge. Jubilejní, jedenáctý *COMPSTAT* se po dvaceti letech (1974-1994) vrátí na sklonku srpna 1994 do Vídně, kde se jeho organizace ujala tamní *Technická universita*. Jak nám sdělili P.Sint a W.Grossmann, počítá se s přenesením středečního dopoledního programu do areálu *MFF UKo* v Mlynské dolině.

K velké lítosti je třeba konstatovat, že na letošní *COMPSTAT* byly přihlášeny pouze čtyři referáty z Československa, což je mnohem méně než v minulosti. Mělo by totiž být v zájmu celé naší statistické veřejnosti se nejenom účastnit, ale též presentovat své výsledky na takovémto representativním fóru. Přitom na tématický zájezd který připravovala odborná skupina pro výpočetní statistiku, se přihlásilo 14 osob. Zájezd sice vzhledem k probíhajícím změnám v celé naší společnosti bohužel padl, ale až v dubnu t.r. zatímco výběr přednášek proběhl již v únoru.

Dále mi dovolte apelovat na všechny naše autory, aby při zasílání příspěvků věnovali pozornost též časopisům *Computational Statistics and Data Analysis* (*CSDA*) a *Computational Statistics Quarterly*. V obou případech se jedná o recenzované časopisy s poměrně krátkou dobou mezi zasláním příspěvku a otištěním, samozřejmě v případě přijetí. *CSDA* je navíc pokryt *Citation Indexem*. Naším zástupcem v redakčních radách obou časopisů je T. Havránek, jemuž je možné zasílat rukopisy. Navíc, časopis *CSDA* se stane od příštího roku též členským časopisem *IASC*. Zahrne přitom vedle stávajících sekcí pokrývajících metodologii, aplikace a výpočetní aspekty též dosavadní členský časopis Asociace, tj. *Statistical Software Newsletter*. Ten tím pádem po patnácti letech úspěšné existence přestane vycházet.

Na úplný závěr mého domácího cvičení z pilnosti mi dovolte zvolat spolu s klasikem

COMPSTAT 90 JE MRTEV - AŤ ŽIJE COMPSTAT 92

Tabulka 1. Přehled přednášek podle výběru programového výboru

Téma	Zvané přednášky	Sdělení	Krátká sdělení	Výukové semináře
*algoritmy a statistické programové vybavení	1	5	22	-
*analýza mnohorozměrných dat a tvorba modelů	1	10	10	-
*časové řady	1	4	7	-
*databázové systémy a statistika	1	1	10	1
*expertní systémy ve statistice	1	7	5	-
*klasifikace	1	4	15	1
*optimalizace a nelineární modely	1	1	3	-
*robustní statistika a její výpočetní aspekty	1	5	7	-
*různé	-	6	3	-
celkem	8	43	82	2

Konference COMPSTAT a umělá inteligence

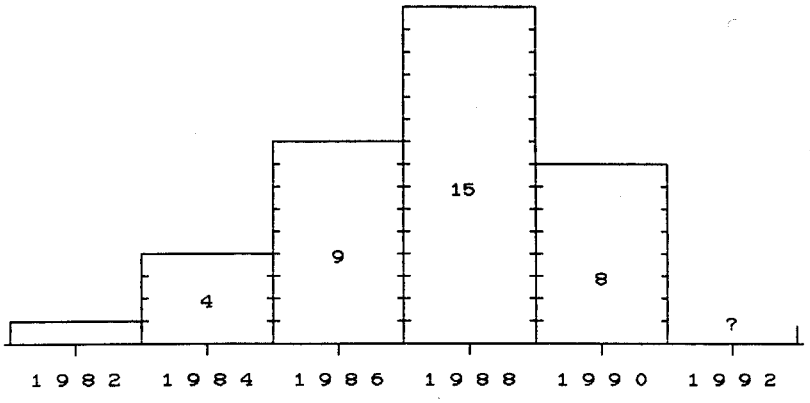
Tomáš Havránek

Umělá inteligence /AI/ jako jedna z oblastí informatiky /computer science/ ovlivnila v posledních letech silně obsah konferencí COMPSTAT. Pokusme se nastínit tento vývoj a některé jeho podstatné rysy.

V roce 1978 /Leiden/ přednesl J. A. Nelder přednášku "The future of statistical software", ve které není ještě o umělé inteligenci zmínka; věnoval se především klasičtějším technikám informatiky /struktury dat, operační systémy, statistické jazyky/. V r. 1980 /Edinburgh/ je věnována celá sekce interaktivnímu počítání /interactive computing/ s příspěvky typu "Interactive or batch?", ale o umělé inteligenci opět nic /pomíjíme oblast rozpoznávání obrazců a shlukové analýzy, která se někdy také přičítá k umělé inteligenci/. V roce 1982 /Toulouse/ se objevuje vyžádaná přednáška P. Hájka a J. Ivánka "Artificial inteligence and data analysis", která referuje o dvou věcech - použití AI technik pro řízení analýzy dat jako kognitivní aktivity /ambiciozní, ale nerealizovaný projekt GUHA 80/ a o reálné možnosti použít technik expertních systémů k podpoře rozhodování při volbě parametrů běhů statistických programových systémů. V témže roce je také první přednáška o paralelním zpracování dat /Sylwestrowicz/ zejména pro generování pseudonáhodných výběrů. Zato v r. 1984 /Praha/ je AI technikám věnována celá sekce /Inteligent software in statistical data analysis/ se čtyřmi přednáškami: W. Gale a D. Pregibon "Constructing expert system for data analysis by working example" /vyžádaná přednáška/, D. Pregibon a W. Gale "REX - an expert system for regression analysis", A. Kowalski, A. Shutt "Analysis of suppositions" a T. Pukkila, S. Puntanen a O. Stenman "On the possibilities to automate the study of statistical dependence between two variables". Významné jsou především první dvě přednášky; druhá z nich referuje o pracujícím experimentálním systému pro podporu rozhodování v regresní analýze, první pak otevírá problém získávání znalostí, které by statistický expertní systém měl využívat a navrhuje i určité řešení. V roce 1986 /Řím/ je již explicitně sekce expertní systémy s devíti přednáškami. Většina

přednášek pojednává o projektech expertních systémů pro podporu rozhodování při statistické analýze dat, zejména pro podporu uživatelů statistického programového vybavení; projekty jsou více či spíše méně rozpracované či implementované. V r. 1988 je opět věnována celá sekce expertním systémům /15 článků/. Témata přednášek se začínají diferencovat; klasické téma jsou projekty zmiňované výše. U nich jsou dvě třídy - pracující, užitečná, ale málo inteligentní /např. P. Hietala referuje o systému ESTES pro analýzu časových řad/ a inteligentní ale zatím nepracující. Druhé téma je otázka získávání znalostí pro expertní systémy /obecně/ na základě statistických dat /Jiroušek a Kříž, Iványek a Stejskal, Norbotten/. Třetí téma je izolované - proč expertní systémy vůbec pracují tak jak pracují a je mu věnován příspěvek P. Hájka "Towards a probabilistic analysis of MYCIN-like expert systems". Předváděn byl program GLIMSE, který v konverzaci s uživatelem řídí práci známého systému GLIM pro zobecněné lineární modely. Pro úplnost podotkněme, že v Kodani byl o další výukovou přednášku o technikách umělé inteligence požádán W. gale. Zajímavé je, že se objevila jedna přednáška o využití neurosítí. Vyžádaná přednáška o paralelních metodách v lineární algebře /G. W. Stewart/ zůstala izolována. Přednášky v Kodani vyvolaly pravděpodobně reakci B. Steitberga "On the non-existence of expert systems" /Statistical Software Newsletters, 19 /1988/, 55-62/, který uzavírá, že nepotřebujeme umělou inteligenci, ale "less stupid systems" /autor byl členem programového výboru konference COMPSTAT 80/. Jako kuriositu uvádím, že Streitberg cituje v textu jako podstatný článek /Nelder, 1977/ a v seznamu literatury má Nelder J. A. /1978/: Inteligent programs, the next stage in statistical computing, COMPSTAT 78, Physica-Verlag. Přitom ve sborníku COMPSTAT 78 je zcela jiný, výše zmíněný Nelderův článek. V diskusi, která v SSN následuje, si toho nikdo nevšiml, včetně Neldera. Konečně letos /Dubrovnik/ byla na programu sekce expertní systémy ve statistice s osmi příspěvky; s expertními systémy měla z nich něco společného asi polovina. Jednotlivé přednášky byly: Wittkowski, K. M.: Statistical knowledge based systems - critical remarks and requirements for approval, Kuzmenkov, V. V., Terskin, O. I.: New approach to Guha-method from the realiability view-

point, Hebrail, G., Suchard, M.: Classifying documents: a discriminant analysis and an expert system work together, Gale, W. A., Church, K. W.: Estimation procedures for language context: poor extimates are worse than none, van den Berg, G. M., Visser, R. A.: Knowledge modelling for statistical consultation systems; two empirical studies, Gebhardt, F.: An expert system strategy for selecting interesting results, Gottee, M. J.: Computer assisted interpretation of conditional independence graphs, Dorda, W., Froeschl, K. A., Grossmann, W.: Wamastex - heuristic guidance for statistical analysis. Je patrný vývoj zpět k širšímu chápání problematiky AI ve statistice a odklon od "pouhých" jednotlivých projektů expertních systémů.



Celkový obrázek vývoje počtu článků, věnovaných AI na konferencích COMPSTAT v minulých letech, mi něco připomíná. Možná, že zůstane zajímavé, ale náročné jádro k dalšímu výzkumu i aplikacím?

Změny v syntaxi SPSS/PC+ 3.# oproti 2.#

Pro uživatele, který je zvyklý pracovat s SPSS/PC+ je nejpodstatnější vnější změnou nutnost od verze 3.1 **udávat alespoň první tři písmena obou slov** v případě víceslovních příkazů.

Drobné kosmetické úpravy byly provedeny v REVIEW, za podstatné lze pokládat důsledné ovládání pomocí kombinace **Alt-písmeno**:

Alt-Z zvětší okno na celou obrazovku,
Alt-P připojení bloku k určenému souboru (případně jej vytvoří)
Alt-D vymaže aktuální řádek (při editování),
Alt-U obnoví právě vymazaný řádek,
F1+Alt-G vyvolá glosář nápovědy.

Z verze 2.0 lze podle našich zkušeností bez problémů použít SPSS Data Entry II, ale nikoliv jiné doplňkové systémy (Trends).

Při popisu jazyka budeme používat následujících symbolů:

[ ] výraz uzavřený v hranatých závorkách je nepovinný,
{ } množinové závorky znamenají, že je nutno použít jeden z vyjmenovaných prvků,
číslo je default hodnota, kterou není nutno uvádět a na jeho místě lze uvést jinou hodnotu příslušného typu,
| svislá čára odděluje jednotlivé alternativní prvky,
. tečka ukončuje príkaz jazyka SPSS.

a) Nové procedury v BASE

Výrazným pokrokem je pro statistiku EXAMINE, pro práci s daty potěší FLIP a lékaři ocení RANKm který umožňuje rozdělit do stejných skupin podle kvantilů.

aa) EXAMINE

Procedura je nástrojem explorační analýzy dat. Produkuje základní statistiky, jednorozměrné grafy a robustní charakteristiky polohy.

Typické minimální použití je

EXAMINE VARIABLES=jméno_proměnné.

Výsledkem jsou v tomto případě
- základní jednorozměrné statistiky
- vertikální krabičkový graf (boxplot) a

- číslicový histogram (stem_and_leaf).

Úplná syntaxe příkazu EXAMINE je

EXAMINE VARIABLE=seznam_proměnných [BY seznam_proměnných]

za BY jsou uvedeny kategoriální proměnné,

[/COMPARE={ GROUP | VARIABLE }] slouží k řízení způsobu zobrazování krabicového grafu. Použitím dílčího příkazu se kreslí do jednoho grafu :

GROUP default, všechny skupiny BY,

VARIABLE všechny proměnné,

[/SCALE={ PLOTWISE | UNIFORM }] slouží k řízení měřítek grafů:

PLOTWISE znamená pro každý graf vlastní měřítko,

UNIFORM znamená grafy ve stejném měřítku,

[/ID= { $CASENUM | jméno_proměnné }] umožňuje identifikovat jednotlivá pozorování pomocí uvedené proměnné (default $CASENUM jsou pořadí pozorování),

[/FREQUENCIES [FROM(počáteční_hodnota)] **[BY(přírůstek)**]

generuje tabulku intervalového rozdělení četností se zadanou počáteční hodnotou a krokem (defaultem FROM je minimum, defaultem BY je přírůstek použitý v stem_and_leaf grafu),

[/PERCENTILES= **způsob výpočtu kvantilů,**

{ HAVERAGE default, diskrétní interpolace,

WAVERAGE spojitá interpolace,

ROUND pozorování nejbližší k teor.kavntilu,

EMPIRICAL empirická distribuční funkce,

AEMPIRICAL zprůměrovaná distribuční funkce,

NONE žádný kvantil (potlačení defaultu),

}

[(seznam_percentilů)] defaultem jsou 5,10,25,50,75,90,95.

[/PLOT= **grafické znázornění**

{ [STEMLEAF] [BOXPLOT] krabicový graf a číslicový histogram (defaulty),

NPPLOT normální graf (vizuální metoda pro testování normality), Shapiro-Wilksův a Lillienforsova modifikace Kolmogo- rov-Smirnovova testu normality,

SPREADLEVEL(hodnota) graf poloha-rozptyl (logaritmus mediánu versus logaritmus kvartilové odchylky), který umožňuje
- testovat homogenitu rozptylů pomocí Levenova testu,
- odhadnout parametr exponenciální transformace (na základě jednotkového přírůstku regresní přímky v grafu),

HISTOGRAM histogram,
ALL všechny grafy,
NONE potlačení výstupu grafů (defaultů),
}
[/STATISTICS= **jednorozměrné charakteristiky**
{ **DESCRIPTIVES** default, (průměr, medián, směr. odchylka, rozptyl, šikmost a její směrodatná odchylka, špičatost a její směrodatná odchylka, kvartilová odchylka, rozsah, minimum, maximum a průměr s vyloučením 5% odlehlých hodnot),
EXTREME(5) obecně n nejmenších a největších hodnot,
ALL statistiky i extrémní pozorování,
NONE žádný výstup statistik,
}
[/MESTIMATOR **robustní estimátory polohy,**
NONE žádný (default),
ALL všechny s default hodnotami,
HUBER [(1.339)] Huberův M-odhad,
ANDREW [(1,34)] Andrewův M-odhad,
HAMPEL [(1.7, 3.4, 8.5)] Hampelův M-odhad,
TUKEY [(4.685)] Tukeyho M-odhad.

Robustní odhady polohy jsou vhodné pro odhady polohy ve výběrech s extrémními a vybočujícími pozorováními. Zahrnuté čtyři odhady se liší v průběhu váhové funkce:
- malé odchylky mají jednotkovou váhu u Huberova a Hampelova odhadu,
- velké odhylky mají nulovou váhu u Tukeyho, Hampelova a Andrewsova odhadu.

Volba je poměrně obtížná a je možno měnit i nastavené hodno-

ty, na nichž závisí konkrétní průběh váhové funkce.

[/MISSING= **způsob práce s chybějícími pozorováními,**
{LISTWISE default, jsou vyloučeny případy, kde chybí některá hodnota,
REPORT chybějící pozorování vytvářejí zvláštní kategorii,
PAIRWISE pokud je možno provést konkrétní výpočet, provede se, i když pro jiné výpočty s tímto pozorováním některá hodnota chybí,
}
[INCLUDE]] pozorování s uživatelskými chybějícími hodnotami jsou zahrnuta do zpracování.

ab) AUTORECODE

Procedura slouží k rekódování hodnot kategoriálních proměnných do posloupnosti celých čísel. Tím je umožněno jednak používat tyto nové proměnné ve statistikých procedurách, jednak je možno racionalizovat výstupy. Typické minimální použití je
RECODE proměnná /INTO nová_proměnná.

Úplná syntaxe je
AUTORECODE VARIABLE=seznam_proměnných
/INTO seznam_nových_proměnných
[/DESCENDING] číslování od konce,
[/PRINT]. tisk tabulky, která obsahuje přiřazení čísel k jednotlivým názvům.

ac) FLIP

Transpozice datové matice. Minimální syntaxe je
FLIP.

V tomto případě se provede záměna řádků a sloupců datové matice. Jména sloupců se uloží do proměnné CASE.LBL, nová jména sloupců budou VAR001 atd, v případě, že data neobsahovaly proměnnou CASE.LBL, jinak budou uvedeny názvy v CASE.LBL (jména sloupců lze určit také pomocí NEWNAMES).

Transponovat lze maximálně 500 sloupců (proměnných) a 499 řádků (limitem je také kapacita operační paměti, potřeba je (r*c+r+c)*8 bytů).

Úplná syntaxe je

FLIP
[[VARIABLES=] {seznam_proměnných | ALL }]
[/NEWNAMES=proměnná]. proměnná obsahující nová jména sloupců,

ad) RANK

Procedura slouží pro nahrazení původních hodnot jejich pořadími. To je výhodné například při různých děleních výběru, při statistických testech vyžadujících pořadí a podobně.

Typické minimální použití je

RANK jméno_proměnné.

Jeho výsledkem je

- vytvoření pořadí proměnné a jeho uložení do proměnné *Rjméno_proměnné*,
- výstup tabulky uspořádaných hodnot a jejich pořadí.

Procedura dále umožňuje

- vypočítavat pořadí různými algoritmy pro práci se stejnými hodnotami,
- rozdělovat data do kategorií podle pořadí (NTILES),
- počítat řadu pořadových statistik.

Úplná syntaxe je

RANK [VARIABLES=] seznam_proměnných
[({A|D})] vzestupné (A, default), nebo sestupné (D) řazení,
[BY seznam_proměnných]
[/TIES= způsob práce se stejnými hodnotami
MEAN default, průměrné pořadí,
LOW první pořadí všem stejným hodnotám,
HIGH poslední pořadí všem stejným hodnotám,
CONDENSE řadí se pouze různé hodnoty.
]
[/FRACTION= odhad výběrové distibuční funkci pro funkce **NORMAL** a **PROPORTION**
BLOM (r-3/8)/(N+1/4)
TUKEY (r-.5)/N kde N je počet pozorování
VW (r-1/3)/(N+1/3) r je pořadí, které je
RANKIT r/(N+1) z intervalu 1..N.
]

[/PRINT= ovládání výstupu
YES default, tisk přehledu,
NO potlačení tisku,
]
[/MISSING= práce s chybějícími hodnotami
EXCLUDE default, vyloučení všech chybějících hodnot,
INCLUDE zahrnutí uživatelských chyb.hodnot.
]
[/pořadová_funkce [INTO jméno_proměnné]] vytváření pořadových funkcí.

INTO určuje jméno nově vytvořené proměnné. Pokud není uvedeno, vytvoří se automaticky ve formě
první_písmeno_pořadové_funkce + prvních_7_písmen_jména_proměnné

Pořadové funkce jsou

RANK default, pořadí
RFRACTION relativní pořadí (v případě RANK=MEANS nebo HIGH výběrová distribuční funkce)
NORMAL normální skóry, odhady výběrové distribuční funkce dané ve FRACTION (defaultem je BLOM) jsou nahrazeny příslušnými kvantily normálního rozdělení,
PERCENT relativní pořadí v procentech,
PROPORTION odhad distribuční funkce (dané FRACTION),
N N počet pozorování,
SAVAGE Savageovy exponenciální skóry,
NTILES(n) nová proměnná obsahuje hodnoty od 1 do n, které rozdělují soubor podle pořadí do n skupin. S výhodou se používá při rozdělování souboru do několika stejných částí a jejich srovnávání prostřednictvím BY.

b. Nové procedury v Advanced statistics

Pro řadového pracovníka pokroku je důležitá zejména procedura NLR, proto se nebudeme dalšími inovacemi zabývat. Asi by bylo vhodné jim věnovat speciální článek od odborníka.

ba) NLR

Proceduru NLR lze pokládat za odpověď SPSS na zdařilý modul NONLIN známého paketu SYSTAT, ale není tak elegantní.

Procedura nelineární regrese používá Marquardtův algoritmus. Model se zadává v symbolické formě. Je třeba zadat počáteční odhady parametrů, čímž se zároveň rozliší, co jsou parametry a co jsou proměnné. Není třeba (ale je možno) zadávat derivace podle parametrů.

Minimální zadání obsahuje dílčí příkazy

```
MODEL PROGRAM proměnná=hodnota [proměnná=hodnota ...].
 transformace.
NLR závisle_proměnná WITH nezávisle_proměnné.
```

V první části se zadávají počáteční odhady parametrů.

Druhá část obsahuje zadání teoretické rovnice, kde je vhodné použít defaultové hodnoty PRED (z predicted, teoretická hodnota). Pokud použijeme jiné jméno, musíme je opět přiřadit proměnné PRED v části NLR.

Třetí část specifikuje pozorování teoretických hodnot a hodnoty nezávisle proměnných, které jsou odděleny příkazem WITH.

Příklad: Uvedeme výpočet logistické funkce.

```
DATA LIST FREE /y,i.
BEGIN DATA.
5 0
12 1
31 2
62 3
110 4
END DATA.
MODEL PROGRAM a=6 b=-.3 c=200.
COMPUTE PRED=c/(1+EXP(a+b*i)).
NLR y WITH i.
```

Vyhledem k tomu, že ve druhé části lze použít veškerých možností transformací jazyka SPSS, lze řešit např. úlohy z oblasti segmentové regrese (použitím podmíněných příkazů IF).

Úplná syntaxe NLR je

MODEL PROGRAM proměnná=hodnota [proměnná=hodnota ...] .

V této části jsou uvedeny

- počáteční hodnoty parametrů formou
  jméno_parametru=hodnota

transformace. funkční předpis pro výpočet teoretických hodnot ve formě
PRED=výraz nebo
jméno_proměnné=výraz (pak je ale nutno vymezit jméno proměnné příkazem PRED=)

[DERIVATIVES transformace]. funkční předpisy pro výpočet derivací podle jednotlivých parametrů tvaru
d.parametr=výraz.

NLR závisle_proměnná WITH nezávisle_proměnné. Vymezení proměnných.

[/OUTFILE=soubor] uložení odhadů do souboru (parametrů, součtu čtverců a počet pozorování)

[/FILE=soubor] čtení počátečních odhadů ze souboru

[/PRED=jméno_proměnné] jméno proměnné, která odbsahuje teoretické hodnoty (pokud je jiné než PRED)

[/SAVE uložení dočasných proměnných

- **[PRED]** teoretické hodnoty
- **[RESID[(jméno_proměnné)]** rezidua
- **[DERIVATIVES]]** derivace

[/CRITERIA= kritéria konvergence (uvedeny jsou defaultové hodnoty)

- **[ITER 100]** počet iterací
- **[CKDER .5]** kritická hodnota derivací (0 vypíná kontrolu)
- **[SSCON 1E-8]** změna součtu čtverců
- **[PCON 1E-8]** změna hodnoty parametrů
- **[RCON 1E-8]** změna korelace mezi rezidui a derivacemi

Co je MIM ?

Marta Horáková

MIM je program pro osobní počítače kompatibilní s IBM PC. Jeho autory jsou David Edwards (Novo-Nordisk CNS Division, Novo-Nordisk A/5, Sydmarken 5, DK 2860 Søborg, Denmark) a Marta Horáková (Ústav systematické a ekologické biologie ČSAV, Květná 8, 603 65 Brno). Informace můžr podat také Tomáš Havránek (SVT ČSAV, Pod vodárenskou věží 2, 182 07 Praha 8).

Poslední verze z roku 1989 je zdokonalením verzí starších v posloupnosti: Edwards MIM 1987, Horáková MIMAS 1987, Edwards MIM 1989, Horáková MIMAS 1989.

Název programu MIM je z anglického 'Mixed Interaction Models' a napovídá, že se jedná o softwarové vybavení pro práci se smíšenými interakčními modely studovanými zejména v [2],[4] a [5].

Písmena AS (Automatical Selection) v názvu dřívějších verzí znamenají, že možnosti práce s jedním modelem popisujícím strukturu závislosti diskrétních a spojitých náhodných veličin byly doplněny o automatické vyhledávání množiny modelů, které připadají v úvahu jako alternativní vysvětlení struktury závislosti veličin na základě informace obsažené v datech. Algoritmus výběru je omezen na tzv. grafové modely a odpovídá postupům z [3]. Možnosti vyhledávání jsou samozřejmě v poslední verzi programu MIM zahrnuty také.

Vstupní data programu MIM jsou realizací náhodného výběru ze smíšeného rozložení pravděpodobností. Jsou zadávána buď po řádcích, z nichž každý odpovídá jednomu objektu, nebo ve formě četností tříd určených kombinací hodnot diskrétních veličin, průměrů a výběrové kovarianční matice spojitých veličin v každé třídě.

Model struktury závislosti je zadáván tzv. generující sentencí modelu skládající se ze tří částí oddělených lomítkem: diskrétních generátorů, lineárních generátorů a kvadratických generátorů ([2]).

Jsou-li A, B diskrétní náhodné veličiny, X, Y spojité veličiny, pak např. AB/ABX,Y/ABXX,XY,YY je správně utvořená generující sentence popisující model, který při zápisu obdobném zvyklostem v analýze rozptylu vyjádříme takto:

$$\log f(i,j,x,y) = \alpha + \alpha^{A}_{(i)} + \alpha^{B}_{(j)} + \alpha^{AB}_{(i,j)} +$$

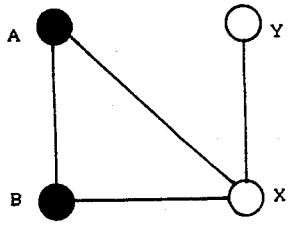
$$+ (\beta^{X}+\beta^{AX}_{(i)}+\beta^{BX}_{(j)}+\beta^{ABX}_{(i,j)}).x + (\beta^{Y}).y -$$

$$- \tfrac{1}{2}(\delta^{XX}+\delta^{AXX}_{(i)}+\delta^{BXX}_{(j)}+\delta^{ABXX}_{(i,j)}).x.x -$$

$$- \tfrac{1}{2}(\delta^{YY}).y.y - (\delta^{XY}).x.y \quad ,$$

kde f je hustota tzv. CG-rozložení, α, β, δ jsou diskrétní, lineární a kvadratické interakce.

Každému smíšenému interakčnímu modelu, resp. generující sentenci, lze přiřadit graf. V našem případě je to graf



Tomuto grafu odpovídá více modelů, např. model s generující sentencí AB/ABX,Y/XY,XX,YY. Pokud však každému grafu přiřadíme generující sentenci, pro niž platí:

- její diskrétní generátory jsou právě kliky podgrafu určeného právě diskrétními veličinami,
- pro každou spojitou veličinu : lineární generátory jsou právě kliky podgrafu určeného právě touto veličinou a diskrétními veličinami,
- pro každou spojitou veličinu a dvojici spojitých veličin: kvadratické generátory jsou právě kliky podgrafu určeného právě těmito veličinami a diskrétními veličinami,

pak hovoříme o tzv. 'grafových modelech'. Z uživatelského hlediska je třída grafových modelů příjemná zejména proto, že každá chybějící hrana v grafu odpovídá jisté podmíněné nezávislosti a interpretaci lze snadno vyčíst z grafu. Pro grafový model se zde uvedeným grafem platí, že náhodné veličiny A, B jsou podmíněně nezávislé s Y při daném X, neboť každá cesta z {A,B} do {Y} vede přes {X}.

Své požadavky zadává uživatel prostřednictvím příkazů jazyka MIM. Jejich podrobný popis popis syntaxe, sémantiky i jednoduché

příklady použití jsou součástí systému HELP. Zde je také podrobný seznam adekvátní literatury i přehled omezení při práci s MIMem. První dialog s MIMem obvykle začíná jeho spuštěním zadáním příkazu

```
12:00:00 C:> MIM
```

a volbou nápovědy

```
MIM -> HELP
```

Podstatným výstupem je výsledek testování uživatelem definovaného modelu, případně vyhledání množiny modelů odpovídajících daným datům.

Literatura:

[1] Edwards,D.: A guide to MIM. Research report (Statistical research unit university of Copenhagen, 1989).

[2] Edwards,D.: Hierarchical mixed interaction models. JRSS Serie B, 52 (1990), 3-20.

[3] Edwards,D., Havránek,T.:A fast model selection procedure for large families of models. JASA 82 (1987), 205-213.

[4] Lauritzen,S.L.: Mixed graphical association models. Scandinavian Journal of Statistics (1989).

[5] Lauritzen,S.L., Wermuth,N.: Graphical models for associations between variables, some of which are quantitative and some qualitative. Annals of Statistics 17 (1989), 31-57.

[6] Horáková,M.: Smíšené interakční modely. Sborník ROBUST 88 (1988)

[7] Horáková,M.: Grafové modely a vyhledávání modelu. Sborník ROBUST 90 (1990).

[8] Whittaker,J.: Graphical models for in applied multivariate statistics. Wiley, Chichester (1990).

Na rozhraní ledna a února se bude konat

*** 1. Výroční konference České statistické společnosti ***

Přesný termín a program budou upřesněny ve vánočním čísle IB

*

Toto (experimentální) zvláštní číslo IB je vydané při příležitosti COMPSTATu 90 a je věnované oblasti, která se velmi rychle mění a vyvíjí - statistickému software. Informace z této oblasti poměrně rychle zastarávají a proto je třeba reagovat co nejrychleji. Rádi bychom znali vaše názory na podobné akce, případně náměty na další zvláštní (monotematická) čísla IB.

**

Článek doc. Militkého a ing. Tvrdíka, na nějž odkazuje dr. Antoch, jsme do uzávěrky tohoto čísla nedostali. Pokud článek dostaneme do uzávěrky vánočního čísla IB, zařadíme jej tam. (Do)

***

Nikoli nevýznamným problémem při používání různých programových produktů je obecně absence přehledné česky psané dokumentace. Přitom na různých místech vznikají různé překlady, úvody, návody apod. Myslíme si, že práce vložená do takovýchto textů by mohla sloužit více uživatelům při vhodně zvolené formě distribuce. Jeden návrh (z VŠE) zde předkládáme a zároveň prosíme o vaše názory :

Na VŠE vznikají - zejména pro potřeby výuky - úvody do práce s některými důležitými systémy. Úroveň těchto popisů není přehnaně vysoká, nicméně jsou psány česky v textovém editoru TEXT602, jsou k dispozici ve formě souborů a lze je tedy distribuovat na disketách v relativně krátké době. Jedna z možností, jak toto realizovat by byla : poslat disketu a 50,- Kčs (hotově, nebo složenkou na účet Společnosti) a obratem dostat současnou verzi zvoleného systému v T602. Tento text však nepředávat dál a komerčně jej nevyužívat. Druhou možností je vydávat skripta a poslat skriptum. Tímto způsobem můžeme v současné době nabídnout např. **STATGRAPHICS, SYSTAT, MINITAB, SPSS** (kromě jiného).

****

Připomínky, náměty a příspěvky posílejte na adresu tajemníka:
dr. Gejza Dohnal, Jeronýmova 7, 13000 Praha 3