PROGRAMY PRO STATISTICKOU ANALÝZU DAT

Josef Tvrdík

Abstrakt: Příspěvek se zabývá programovým vybavením pro analýzu dat. Je podán stručný přehled rozšířených nebo dostupných programových prostředků a porovnány jejich vlastnosti. Ze zkušeností v oblasti statistického software jsou vyvozena některá obecná doporučení pro vývoj přenositelných aplikačních programů. Na příkladu vývoje interaktivního statistického systému INSTA jsou ukázány některé obtíže, které zvláště u nás doprovázejí vývoj přenositelných programů.

1. Proč se na programátorském semináři zabývat statistickým softwarem?

a) Programové vybavení pro statistickou analýzu dat má za sebou mnohaletou tradici a přitom jeho vývoj stále pokračuje v mnoha směrech, v některých s rostoucí intenzitou.

b) Statistický software (dále SSW) byl jedním z prvních produktů aplikačního programového vybavení určeného pro mnohonásobné opakované využití na různých instalacích, počítačích a v rozdílných souvislostech. Dlouholetá činnost úspěšných výrobců SSW (např. SPSS, BMDP, SAS) vedla k tomu, že počty licencí těchto produktů se počítají na desetitisíce (SAS v r. 1987 byl implementován na 22 000 instalací) a počty uživatelů jsou do milionů. Z této historie je užitečné odvodit poznatky i pro vývoj jiných druhů aplikačních programů.

c) Rozšíření mikropočítačů v 80. létech přineslo další mocný impuls rozvoji SSW s kladnými i zápornými důsledky. Nyní po světě existuje mnoho set různých programových balíků a systémů pro analýzu dat. Tyto produkty se liší v mnoha

ohledech - obsahu, architektuře, ceně, spolehlivosti, kvalitě uživatelského rozhraní atd. Tyto rozdíly si zaslouhují alespoň zamyšlení.

d) Přes rozbujelost a údánlivý nadbytek SSW se velmi často setkáváme se situacemi "mnoho dat - málo analýzy".

e) Současný vývoj SSW je v úzkém vztahu dalších zajímavých oblastí aplikací počítačů jako je grafika, databázové systémy, jazyky vysoké úrovně a umělá inteligence (zejména expertní systémy).

2. Jaké jsou druhy SSW?

Historicky nejstarším druhem SSW jsou knihovny podprogramů. Jako příklad lze uvést letitý SSP (scientific subroutine package). V současné době jsou nejrozšířenější knihovny NAG a IMSL. U nás je pro počítače SMEP dostupná knihovna SPMSV, nápadně připomínající jednu ze starších verzí knihovny IMSL. Vzhledem k tomu, že k užití knihoven podprogramů je nutná znalost programování a že jsou již pohodlněji použitelné programové produkty, ztrácejí knihovny podprogramů jako SSW pro koncového uživatele na významu. Jsou však nepostradatelné pro výrobce SSW.

Některé jednoduché metody analýzy dat jsou součástí jiných druhů programových produktů, které spojuje zpravidla jednotný způsob ovládání (např. jednotná syntaxe řídícího jazyka) a jednotná struktura tzv. savefilu, t.j. souboru, přes který si programy mohou vyměňovat data, případně i výsledky k další analýze. Přitom však kterýkoli program paketu je možno užívat samostatně. Příklady paketů statistických programů jsou BMDP nebo GUHA.

Za programově nejdokonalejší produkt SSW jsou nyní považovány programové systémy pro analýzu dat. Ve svých funkcích jsou podobné programovým paketům, ale z pohledu uživatele se

jeví jako jeden program. Touto architekturou SSW lze obohatit funkce systému, zjednodušit uživatelské rozhraní (a tím i ovládání) programového systému a zejména snížit nároky programů na vnější paměť (společné moduly nejsou opakovány v každém programu jako u paketu). To je výhodné zvláště pro řešení SSW na mikropočítačích (např. STATGRAPHICS, SYSTAT).

3. Co je společné pro SSW?

Základní sjednocující vlastností všech produktů SSW (a prakticky všech metod analýzy dat) je zpracovávání stejné jednoduché vstupní datové struktury. Touto datovou strukturou je dvourozměrná matice číselných údajů - tabulka. Často se jí také říká datová matice. Řádky této tabulky odpovídají jednotlivým měřeným případům (objektům, statistickým jednotkám, anglicky většinou case, terminologie není zcela ustálená) a sloupce jednotlivým měřeným proměnným (veličinám, znakům, angl. variable).

Produkty SSW zpravidla dovolují uživateli tyto skutečnosti zaznamenat, t.j. pojmenovat veličiny, případné objekty a usnadnit si tak interpretaci výsledků statisticko-analytické úlohy. Některé produkty umožňují různé operace nad vstupní tabulkou, jako naplnění tabulky daty, aktualizaci jednotlivých údajů, výběr řádků nebo sloupců (vytvoření podtabulky), přidávání sloupců tabulky aritmetickými nebo logickými operacemi, spojování tabulek apod.

Těmto operacím se říká předzpracování nebo správa dat (data handling, data manipulation, data preprocessing). Dále statistické pakety a systémy obsahují několik až několik desítek procedur pro analýzu dat a výpočty statistických charakteristik.

Společné pro prakticky všechny současné produkty SSW je to, že zcela ponechávají na uživateli

a) jak vstupní datové matice vyjadřuje zkoumanou část reálného světa

b) volbu statisticko-analytických metod k řešení úlohy

c) interpretaci výsledků statistických výpočtů při řešení úlohy.

Zdá se, že tento odstup od řešení konkrétních úloh spolu s jednoduchostí vstupní datové struktury byly nutné podmínky k tomu, aby obecný přenositelný SSW bylo vůbec možno vytvořit.

4. Jak se orientovat ve struktuře a funkcích SSW?

Základní představa pro orientaci ve SSW vychází ze vstupní datové struktury - tabulky dat. Pro uživatelský pohled na SSW není důležité, jakým způsobem je v konkrétním produktu tato struktura implementována (nejčastěji to bývá sekvenční soubor, ve kterém věta odpovídá případu). Podobně není důležité, zda některé produkty dovolují zpracovávat i jiné vstupy, odvozené z obecně vstupní datové struktury (kontingenční tabulky, korelační matice apod.). Produkty SSW mají funkce pro předzpracování vstupní tabulky a funkce statisticko-analytické.

Ty jsou obvykle děleny podle skupin metod na:

- deskriptivní jednoduché statistiky (+ grafika)
- neparametrické metody
- regresní metody
- analýza rozptylu a kovariance
- kontingenční tabulky
- mnohorozměrné metody (faktorová analýza, hlavní komponenty, kanonické korelace, diskriminační analýzu atd.)
- shluková analýza
- analýza časových řad
- analýza přežití (selhání)
- případně další specializované metody.

Dalším hlediskem pro orientaci ve statistických funkcích SSW je to, kolika výběrů (skupin) se metoda nebo úloha týká. Podle tohoto hlediska se dělí metody do tří skupin: problém jednoho výběru, problémy dvou výběrů (porovnání mezi dvěma skupinami) a problémy více výběrů (zde je už záležitost vzájemného porovnání komplikovanější).

5. Čím se liší produkty SSW?

Především obsahově - jsou systémy obecné, ve kterých jsou zastoupeny všechny skupiny metod z předcházejícího odstavce (např. BMDP, SAS, SPSS, GENSTAT, SYSTAT, STATGRAPHICS) a specializované (RATS pro analýzu časových řad, GLIM pro obecný log-lineární model, GUHA pro explorativní analýzu kategoriálních dat, CLUSTAN pro metody shlukové analýzy atd.). Mezi obecnými systémy jsou podstatné rozdíly v obsahu, zejména podle toho, zda jsou určeny statisticky kvalifikovanému či nekvalifikovanému uživateli. Obsahové rozdíly jsou často způsobeny i subjektivními příčinami - řešitelskými kapacitami a erudicí výrobců. To se týká zejména SSW pro mikropočítače.

Odlišnosti mezi systémy jsou i ve způsobech komunikace s uživatelem - dávkově nebo interaktivně, u interaktivních systémů jsou značné rozdíly ve formě i obsahu dialogu.

Podstatné rozdíly mezi systémy jsou v možnostech předzpracování dat. Úspěšné kvalitní systémy umožňují velmi pružný vstup dat z různých typů souborů, z databází, poskytují množství operací nad datovou tabulkou a dovolují snadnou výměnu dat mezi různými statistickými systémy.

Produkty SSW se mohou podstatně lišit i spolehlivostí. Mezi stovkami produktů jsou některé velmi nekvalitní, s numerickými i metodickými chybami. Obecně platí, že čím je systém rozšířenější a služby výrobce širší, tím je menší počet chyb. Bezchybný SSW asi neexistuje, například dokonce i v BMDP "přežila" ve třech verzích, t.j. šest let, chyba ve výpočtu významnosti velmi často užívané t - statistiky.

6. Jak získávat SSW?

Před léty rozdělil kolega Pokorný způsoby, jak získat SSW na 1) čestné, 2) nečestné, 3) vlastním přičiněním, t.j. programováním - s tím, že první je lepší než druhý a druhý lepší než třetí. Nevýhody a rizika druhého způsobu stále vzrůstají - žádné služby a záruky k SSW, počítačové viry, možnosti soudního postihu neoprávněného uživatele. Přesto je první způsob u nás málo rozšířený. Uživatelských licencí je pár (např. BMDP a SBSS v SVT ČSAV) a převážná většina ekonomických šéfů nepovažuje SSW za zboží hodné devizových investic. Některé produkty navíc podléhají vývozním omezením západních zemí.

Snadněji dostupné jsou tuzemské produkty SSW, i když výběr je dosti omezený. Pro počítače s OS IBM/360 je např. SASD (statistická analýza sociálních dat), která obsahuje základní statistické metody a analýzu kontingenčních tabulek, nebo GUHA, což je specializovaný paket pro explorativní analýzu kategoriálních a smíšených dat. Pro minipočítače je určen systém STADIA (statistická analýza experimentálních dat) a INSTA (interaktivní statistický systém pro minipočítače SMEP). Pro mikropočítače s operačním systémem CP/M jsou dostupné systémy DAISY-83 a ABSTAT. Pro šestnáctibitové mikropočítače třídy IBM PC je možné získat SSW z volně kopírovatelné public domain (např. MYSTAT, MLIB).

7. Jaký bude další vývoj v oblasti SSW?

Především bude dále pokračovat vývoj už nyní pozorovatelné kategorizace softwaru podle druhu uživatelů. Jiné programové prostředky jak co do funkcí, tak co do uživatelského rozhraní potřebuje uživatel pro běžné aplikace, jiné profesionální statistik, zabývající se vývojem metod analýzy dat, jiný SSW je vhodný pro výuku atd.

SSW pro široké použití směřuje k jednoduchosti z pohledu uživatele. Týká se to především způsobu ovládání (interakce, možnost účinných HELPů), automatizovaného záznamu postupu řešení úlohy (protokol, funkce DIARY) a interpretace a prezentace výsledků (užití pokročilých grafických prostředků ve výstupech, možnosti editování textových i grafických výstupů a jejich rychlá příprava k publikaci). Probíhají pokusy s využitím expertních systémů pro konzultace užití SSW i pro interpretaci výsledků. SSW se stává součástí databázových systémů a vzniká integrovaný software, poskytující uživateli bohatost potřebných funkcí snadno použitelných v rámci jednoho vnějšího mýtu.

8. Jaké poučení poskytuje SSW pro jiné aplikační programy?

Pro obecné aplikační programy mnohonásobného opakovaného užití jsou důležité následující věci (bez nároku na úplnost či neomylnost):

- dostatek potenciálních uživatelů
- jednoduchá a všeobecně přijímaná struktura vstupních dat
- představa o závažnosti jednotlivých funkcí softwaru je společná pro většinu uživatelů
- propagací a publikacemi lze uživatele připravit, vzdělávat, případně i získat
- je nutné věnovat velkou péči vnějšímu mýtu, jak komunikaci programu s uživatelem, tak ostatním složkám (manuály, učebnice atd.)
- pro úspěšnost a uživatelskou popularitu je důležitá spolehlivost produktu a služby s produktem poskytované
- úspěšnost produktu je podmíněna jeho dlouhověkostí, tzn. průběžným vývojem, proto je důležitá modulární architektura produktu, zejména důsledné oddělení komunikačních od výpočetních modulů
- i více produktů stejné kategorie může být úspěšných, navzájem si nejen konkurují, ale mohou se i doplňovat.

9. Jak byl vyvíjen systém INSTA?

Stručně lze odpovědět - jak je v kraji zvykem. Vývoj je poznamenán řadou chyb, které by neměly být opakovány. Záměr vznikl v r. 1984. V té době byla orientace na minipočítače jediná u nás přijatelná možnost. Pak řešení pokračovalo jen občasnými diskusními semináři. Těch se zúčastňovalo okolo 15 lidí - programátorů, statistiků i kvalifikovaných uživatelů. Většina z nich se tvářila, že se chce zúčastnit realizace projektu. Důkladné posouzení záměru je jistě důležité, ale v tomto případě bylo neúměrně dlouhé. Realizace projektu začala vlastně až v r. 1987. Zde již početní účast byla podstatně nižší, prakticky jen tři lidé a to ještě každý jen asi z jedné čtvrtiny své pracovní kapacity.

Protože se diskusemi postup řešení zdržel, začalo se pracovat překotně - návrh architektury systému, návrh uživatelského rozhraní a programování probíhaly paralelně. Tím se sice ochnal nějaký ztracený čas, ale zase jiný ztratil, neboť některá dílčí řešení bylo nutno zahodit a udělat znovu. Programové prostředí pro řešení vznikalo postupně až v průběhu programování. Podobně programový prostředek (v tomto případě SVS) umožňující jednotný návrh a efektivní implementaci uživatelského rozhraní nezávislého na typu terminálu byl získán až v r. 1988.

Přes tuto minimální koncentraci řešitelských kapacit a ostatních zdrojů v řešení projektu i řadu evidentních chyb v postupu je dokončena verze INSTA 88/1, která obsahuje dosti úplné funkce předzpracování dat, základní statistiky, jednoduché metody analýzy rozptylu, regrese, analýzu kontingenčních tabulek a některé další funkce. Systém je interaktivní, s HELPY a stručnou tištěnou příručkou. Tato verze je použitelná v dravém prostředí různých uživatelů, s nimiž komunikuje česky (bez čárek a háčku).

Vývoj systému INSTA pokračuje, a autoři doufají, že v projektu je vytvořeno prostředí pro podstatné rozšíření funkcí systému. Po těchto zkušenostech asi mnohého napadne, jaké

programové produkty mohly už u nás vzniknout, kdyby byly prostředky a kapacity soustřeďovány na projekty pro mnohonásobné opakované využití (kdyby nebyla tak vysoká "rozestavěnost") a nedělaly se evicentní chyby v řešení.

Literatura:

1. ABSTAT, uživatelská příručka, AK Slušovice, 1988
2. Bébr R., Na úrovni, Programování 88, DT ČSVTS, Ostrava, 1988
3. Dixon W.J./ed./, BMDP Statistical Software, Manual, Univ. of California Press, 1985
4. Francis I., Statistical Software. A Comparative Review, North Holland, 1961
5. Hájek P./ed./, Programové systémy pro analýzu dat, sborník semináře ČSVTS Matematického ústavu ČSAV, Praha, 1988
6. Havlová H., Špitálský J., Vošvrda M., User's Manual [illegible], ÚTIA ČSAV, Praha, 1986
7. Hevránek T./ed./, GUHA - Package of Programs for Exploratory Data Analysis, Reference Manual, version 2.3, SVT ČSAV, Praha, 1988
8. Jiroušek R., Kříž O., Řehák J., SASD - příručka pro uživatele, DT ČSVTS, Pardubice, 1984
9. Osecký P., Principy, přehled a počítačové využití statistických metod, sborník SOFSEM 88, 1988
10. Řezanková H., Žváček J., Pakety statistických programů, Statistika č. 5, 1987
11. SMEP - soubor podprogramů pro MVS, Datasystém, 1986
12. Tvrdík J., Liška M., Řídící jazyky programových systémů pro analýzu dat, Programování 82, DT ČSVTS Ostrava, 1982
13. Vanek M., SVS, používateľská príručka, Kovoprojekta, Bratislava, 1987
14. Zgydlar K., Tvrdík J., Ettlerová E., INSTA - uživatelská příručka, SVT LF UK Hradec Králové, 1989
15. Žváček J., Řezanková H., Statistické programovací jazyky, sborník letní školy ROBUST '88, JČSMF Praha, 1988
16. Žváček J., Řezanková H., Statistické výpočetní prostředí, Statistika, č. 11, 1988