# *Informační Bulletin*

*České statistické společnosti* *číslo 3, ročník 19, 1. července 2008*

## JAK NA ROZHODOVACÍ STROMY

**Marta Žambochová**

**Abstract**: The tree structure is a popular instrument of the information presentation in many spheres of common life. It finds its use in data analysis on account of its siplicity and its clarity. The expanded group of trees in data modelling and simulation is the group of assorted decision trees. We can solve the classification and prediction tasks by means of decision trees. The paper deals with a comparison of some algorithms for a creation of decision trees. This article shows a way how and why to include decisions trees in an education on faculties of economics. It shows using the trees for finding the resolution of some economic problems.

*Key words*: The decision tree, CART, QUEST, an education on faculties of economics, facultative subjects.

### Úvod

Velmi rozšířenou skupinou stromů, kterých se využívá v datových modelech, jsou různé typy rozhodovacích stromů. Rozhodovací stromy jsou struktury, které rekurzivně rozdělují zkoumaná data dle určitých rozhodovacích kritérií. Kořen stromu reprezentuje celý populační soubor. Vnitřní uzly stromu reprezentují podmnožiny populačního souboru. V listech stromu můžeme vyčíst hodnoty vysvětlované proměnné.

Rozhodovací strom se vytváří rekurzivně dělením prostoru hodnot prediktorů. Máme-li strom s jedním listem, hledáme otázku (podmínku větvení),

která nejlépe rozděluje prostor zkoumaných dat do podmnožin. Takto nám vznikne strom s více listy. Nyní pro každý nový list hledáme otázku, která množinu dat náležící tomuto listu co nejlépe dělí do podmnožin.

Proces dělení se zastaví, pokud bude splněno kritérium pro zastavení. Omezení obsažená v kritériu pro zastavení mohou být např. „hloubka“ stromu, počet listů stromu, stupeň homogennosti množin dat v listech, ...

Dalším krokem algoritmů je prořezávání stromu (prunning). Je nutno určit „správnou“ velikost stromu (příliš malé stromy dostatečně nevystihují všechny zákonitosti v datech, příliš veliké stromy zahrnují do popisu i nahodilé vlastnosti dat). Vygenerují se podstromy stromu vzniklého algoritmem a porovnává se jejich kvalita generalizace (jak dobře vystihují data).

Postup může být takový, že se rozhodovací stromy nejdříve vytváří na tzv. trénovacích datech a poté se jejich kvalita ověří na tzv. testovacích datech.

Jiným způsobem je křížová validace (cross validation), kdy k vytváření stromu a jeho podstromů použijí všechna data. Poté se data rozdělí na několik disjunktních, přibližně stejně velkých částí a postupně se vždy jedna část dat ze souboru vyjme. Pomocí vzniklých souborů dat se ověřuje kvalita stromu a jeho podstromů.

Vybere se takový podstrom, který má nejnižší odhad skutečné chyby. Pokud existuje více podstromů se srovnatelným odhadem skutečné chyby, vybírá se ten nejmenší.

Jednotlivé algoritmy vytváření rozhodovacích stromů se liší následnými charakteristikami:

- pravidlo dělení (splitting rule)
- kritérium pro zastavení (stopping rule)
- typ podmínek větvení
  - multivariantní (testuje se několik prediktorů)
  - univariantní (v daném kroku se testuje pouze jeden z prediktorů)
- způsob větvení
  - binární (každý z uzlů, kromě listů, se dělí na dva následníky)
  - k-ární (některý z uzlů se dělí na více než dvě části)
- typ výsledného stromu, popis obsahu listů
  - klasifikační stromy (v každém listu je přiřazení třídy)
  - regresní stromy (v každém listu je přiřazení konstanty – odhad hodnoty závislé proměnné)

- typ prediktorů
  - kategoriální
  - ordinální

## Algoritmy pro vytváření rozhodovacích stromů

Pro vytváření rozhodovacích stromů bylo vyvinuto velké množství algoritmů. Nejvíce používané jsou CART (L. Breiman, J. H. Freidman, R. A. Olshen, C. J. Stone, 1984), ID3 (J. R. Quinlan, 1975), C4.5 (J. R. Quinlan, 1993), AID (J. N. Morgan a J. A. Sonquist, 1963), CHAID (G. V. Kass, 1980) a QUEST (W. Y. Loh and Y. S. Shih, 1997).

Ve článku budou stručně zmíněny dva ze jmenovaných algoritmů, CART a QUEST, které jsou základem algoritmů v SW produktu STATISTICA, jenž je použit ke zpracování dat v motivačním příkladu uvedeném ve článku.

## Algoritmus CART

Algoritmus poprvé popsali autoři L. Breiman, J. H. Freidman, R. A. Olshen, C. J. Stone v roce 1984 ve článku „Classification and Regression trees".

Algoritmus je použitelný v případě, že máme jednu nebo více nezávislých proměnných. Tyto proměnné mohou být buď spojité nebo kategoriální (ordinální i nominální). Dále máme jednu závislou proměnnou, která také může být kategoriální (nominální i ordinální) nebo spojitá.

Výsledkem jsou binární stromy, protože jsou zde přípustné pouze otázky (podmínky dělení), na které je možno odpovědět ano/ne (Je věk menší než 30 let? Je pohlaví mužské? ...).

Algoritmus dělení je různý pro klasifikační stromy a pro stromy regresní.

Klasifikační stromy používáme v případě, že je závislá proměnná kategoriální. To znamená, že se soubor původních dat snažíme v závislosti na nezávislých proměnných rozdělit do skupin, přičemž, v ideálním případě, každá skupina má přiřazení ke stejné kategorii závislé proměnné.

Homogenita uzlů-potomků je měřena pomocí tzv. funkce znečištění (impurity function) $i(t)$. Maximální homogenita vzniklých dvou potomků je počítána jako maximální změna (snížení) znečištění $\Delta i(t)$.

Algoritmus CART řeší pro každý uzel maximalizační problém pro funkci $\Delta i(t)$ přes všechna možná dělení uzlu, to znamená, že hledá dělení, které přináší maximální zlepšení homogenity dat.

Regresní stromy se používají v případě, že závislá proměnná není kategoriální. Každá její hodnota může být v obecnosti různá.

V tomto případě algoritmus hledá nejlepší dělení na základě minimalizace součtu rozptylů v rámci jednotlivých dvou vzniklých uzlů-potomků. Algoritmus pracuje na základě algoritmu minimalizace součtu čtverců.

## Algoritmus QUEST

Metoda je popsána ve článku z roku 1997 W. Y. Loh and Y. S. Shih: „Split selection methods for classification trees". Algoritmus je použitelný pouze pro nominální závislou proměnnou. Obdobně, jako v případu CART, jsou vytvářeny binární stromy. Na rozdíl od metody CART, provádí metoda QUEST v průběhu budování stromu odděleně výběr proměnné pro štěpení uzlu a výběr dělícího bodu. Metoda QUEST (for Quick, Unbiased, Efficient, Statistical Tree) odstraňuje některé nevýhody algoritmů používajících vyčerpávající hledání (např. CART), jako je náročnost zpracování, snížení obecnosti výsledku, a podobně.

Tato metoda je vylepšením algoritmu FACT, který popsali v roce 1988 autoři W. Y. Loh a N. Vanichsetakul. V prvním kroku algoritmus převede všechny kategoriální nezávislé proměnné na „ordinální" pomocí CRIMCORD transformace.

Dále je v každém listovém uzlu pro každou proměnnou prováděn ANOVA F-test. Pokud největší ze vzniklých F-statistik je větší než předem daná hodnota $F_0$, pak příslušná proměnná je vybrána pro dělení uzlu. Pokud tomu tak není, je pro všechny proměnné proveden Leveneův F-test. Pokud je největší Leveneova F-statistika větší než $F_0$, pak je pro dělení uzlu vybrána tato proměnná. Jinak (tzn. není žádná ANOVA F-statistika ani Levenoeva F-statistika větší než $F_0$) je pro dělení vybrána proměnná s největší ANOVA F-statistikou. Pro dělení uzlu je tedy vybrán ten prediktor, který je se závislou proměnnou nejvíce asociován.

Pro hledání dělícího bodu pro vybranou nezávislou proměnnou je využívána metoda Kvadratické diskriminační analýzy (QDA), na rozdíl od algoritmu FACT, kde je využívána metoda Lineární diskriminační analýzy (LDA).

Postup je rekurzivně opakován až do zastavení (na základě kritéria pro zastavení).

## SW pro zpracování rozhodovacích stromů

- velké statistické SW balíky
  - výhody
    - přehlednost

        * dobrá dostupnost
        * relativně dobrá dokumentace
    - nevýhody
        * vysoké pořizovací náklady
        * nutno kupovat několik modulů
        * velké nároky na HW
        * nepřesný popis použitých metod
- ostatní (komerční a nekomerční)
    - výhody
        * malé (resp. žádné) pořizovací náklady
        * je možno koupit samostatně modul na tvorbu rozhodovacích stromů
        * relativně malé nároky na HW
    - nevýhody
        * mnohdy nedostatečná dokumentace
        * většinou chybí jakákoliv zmínka o použitých metodách

## Motivační příklad

Studovali jsme vzorek studentů střední a vysoké školy a zkoumali jsme, zda lze z určitých hledisek životního stylu vyvodit váhovou kategorii získanou na základě BMI (Body Mass Index) vypočítaného ze zjištěné váhy a výšky osoby.

Ze sledovaných položek jsme za nezávislé proměnné vybrali počet hodin denně strávených u počítače a u televize, průměrný počet hodin sportu za týden, průměrný počet hodin spánku, průměrný počet jídel během dne, převažující druh stravování (fast food, stravování v jídelně resp. restauraci, domácí strava, studená kuchyně).

Jako závislou proměnnou jsme zvolili kategoriální proměnnou nabývající hodnot „podváha“, „normální váha“, „nadváha“ a „obezita“.

Ve statistickém SW STATISTICA jsme použili různé možnosti sestrojení rozhodovacího stromu, vypovídajícího o struktuře sledovaného vzorku studentů. Jednak jsme vytvořili klasifikační strom pomocí algoritmu C&RT vyčerpávajícího prohledávání (viz Obr. 1, Tab. 1), jednak pomocí metody založené na principu QUEST (viz Obr. 2, Tab. 2). Dále jsme vytvořili strom pomocí standardní metody C&RT z modulu Data-Mining, včetně V-fold

Crossvalidation – metody na výběr optimálního stromu (viz Obr. 3, Tab. 3). Shrneme-li výsledky, dostáváme:

- Global CV cost = 0,13636; s.d. CV cost = 0,03272 (vyčerpávající C&RT)
- Global CV cost = 0,22727; s.d. CV cost = 0,03996 (QUEST)
- Global CV cost = 0,081818; s.d. CV cost = 0,026133 (standard C&RT)

Z tohoto hlediska se tedy jeví jako optimální strom vytvořený posledním způsobem.

Pokud rozhodovací strom převedeme na pravidla, pak se dostáváme k závěru, že největší vliv (ze sledovaných hledisek) na váhovou kategorii má druh stravy. Nejvíce ohroženy jsou osoby stravující se ve „fast foodech“ a jídelnách, resp. restauracích.

Ve skupině osob ohrožených vysokou hmotností se oddělují tři skupiny, a to na základě množství sportu. Sport je tedy druhým faktorem ovlivňujícím váhu osob. Osoby trpící obezitou se projevují velkým nedostatkem sportu. Osoby s nadváhou sportují pouze málo a osoby s normální vahou mají nejvyšší intenzitu sportování.

Skupina stravujících se jiným způsobem se dělí na dvě skupiny odlišné množstvím spánku. Obě tyto skupiny se dále dělí na základě počtu denních jídel. Délka spánku a počet denních jídel tedy také ovlivňují váhovou kategorii.

Ze struktury stromu je zřejmé, jak se výše zmíněné faktory projevují na zařazení osob do váhové kategorie.

## Zařazení problematiky rozhodovacích stromů do výuky

V posledních letech se většina ekonomicky zaměřených vysokých škol potýká s problémem snižování hodinových dotací tzv. kvantitativních předmětů (matematika, statistika, ...). Náhrada povinných předmětů nepovinnými však není vždy jednoduchá.

Jedním z problémů je zajištění potřebné návaznosti jednotlivých předmětů. Zavedení výběrových seminářů s matematickou či statistickou tématikou se také potýká s malým zájmem ze strany studentů, kteří mají většinou z obdobných předmětů strach.

Pomoci by mohlo zavedení předmětů, které studenty na první pohled neodradí. Částečným řešením může být i pouze vhodná volba názvu předmětu. To ale není dlouhodobě příliš účinné. Účinnější možností je zavádění předmětů, které názorně ukazují řešení reálných situací z oborů blízkých studijnímu zaměření studentů za skrytého použití matematických či statistických metod. Pokud se podaří studenty zaujmout problematikou, budou více

přístupni přijmout vysvětlení metod, na kterých je řešení založeno, i kdyby se jednalo o metody matematické či statistické.

Výklad látky pak neprobíhá standardním způsobem hierarchicky od nejjednoduššího postupně stále ke složitějšímu, ale naopak se začne cílovým stupněm, to znamená nejsložitějším, postupně se pak osvětlují potřebné informace na nižším stupni složitosti a to až do úplného porozumění problematice.

Rozhodovací stromy jsou problematikou, která splňuje předchozí požadavky. Na první pohled jsou rozhodovací stromy velmi názorné, i laik se velmi rychle zorientuje v grafice stromové struktury a je schopen vyčíst potřebné informace. Využití rozhodovacích stromů je v ekonomickém oboru velice široké, takže je možno studenty seznámit s konkrétními příklady z ekonomického života.

V teorii rozhodovacích stromů jsou jednak využity základní poznatky z oblasti teorie grafů a jednak poněkud širší poznatky ze statistiky. Studenti se učí jednotlivým informacím s vědomím jejich využitelnosti a důležitosti pro tvorbu rozhodovacích stromů. Tím odpadá potřeba přesvědčovat studenty o potřebě daného učiva.

Celkově si myslím, že zavedení tématu rozhodovacích stromů (a jiných podobných témat) do výuky ve formě výběrových seminářů, je velmi dobrou cestou k rozšíření látky s matematickým zaměřením.

## Literatura


[1] Antoch J.: Klasifikace a regresní stromy. Sborník ROBUST 88.

[2] Bentley, J. L.: Multidimensional Binary Search Trees Used for Associative Searching. Comm. ACM, vol. 18, pp. 509-517, 1975.

[3] Berikov, V., Litvinenko, A.: Methods for statistical data analysis with decision trees,
`http://www.math.nsc.ru/AP/datamine/eng/decisiontree.htm`

[4] Loh, W.-Y. and Shih, Y.-S., Split selection methods for classification trees, Statistica Sinica, vol. 7, pp. 815-840, 1997.

[5] Savický, P., Klaschka, J., a Antoch J.: Optimální klasifikační stromy. Sborník ROBUST 2000.

[6] Dostupné po registraci na: SPSS.com – White paper – AnswerTree Algorithm Summary.

[7] Timofeev R.: Classification and Regression Trees (CART) Theory and Applications, CASE – Center of Applied Statistics and Economics, Humboldt University, Berlin, 2004.

[8] Wilkinson, L.: Tree Structured Data Analysis: AID, CHAID and CART – Sun Valley, ID, Sawtooth/SYSTAT Joint Software Konference, 1992.

[9] Žambochová M.: Použití stromů ve statistice – Sborník, 2006, Ústí n. L., ISBN 80-7044-795-8.

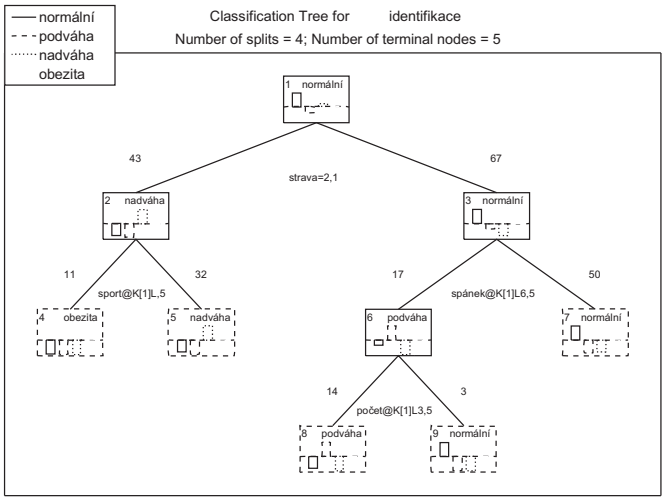
[10] Classification Trees: `www.statsoft.com/textbook/stclatre.html`

[11] Classification and Regression Trees (C&RT): `http://www.fmi.uni-sofia.bg/fmi/statist/education/textbook/ENG/stcart.html`

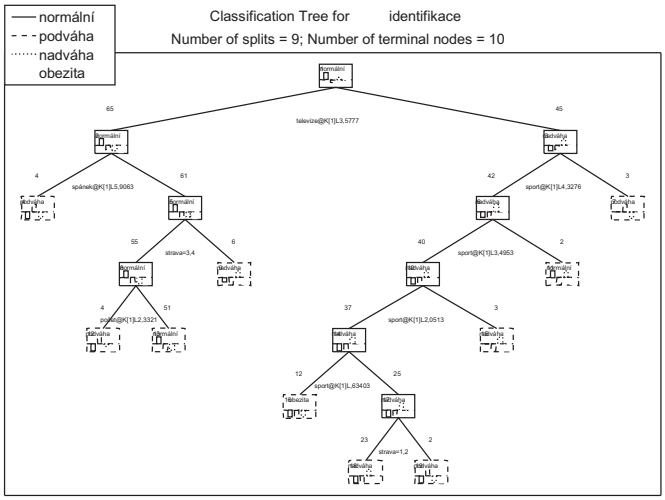
*Adresa*: RNDr. Marta Žambochová, Univerzita J. E. Purkyně v Ústí n. L., Fakulta sociálně ekonomická, Katedra matematiky a statistiky

*E-mail*: `zambochova@FSE.UJEP.cz`

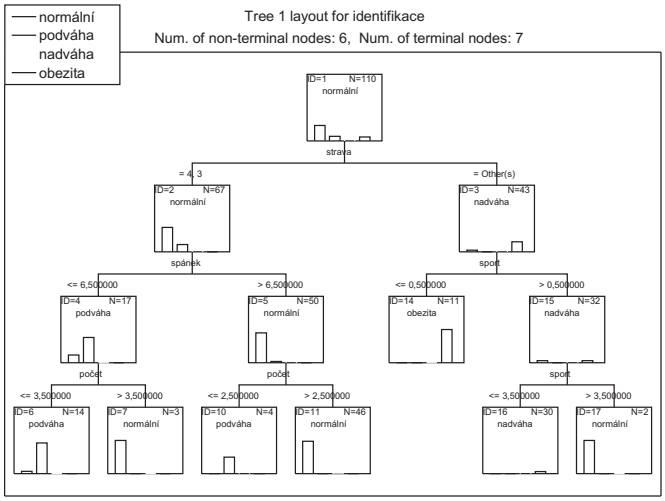
## Obrazové a tabulkové přílohy




Obrázek 1: Klasifikační strom vytvořený pomocí algoritmu C&RT vyčerpávajícího prohledávání




Obrázek 2: Klasifikační strom vytvořený na základě algoritmu QUEST




Obrázek 3: Strom vytvořený pomocí standardní metody C&RT z modulu Data-Mining, včetně V-fold Crossvalidation – metody na výběr optimálního stromu

| Node | Node branch | Left branch | Right normální | n in cls podváha | n in cls nadváha | n in cls obezita | n in cls Predict. | Split | Split | Split | Split |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 3 | 51 | 15 | 31 | 13 | normální | | strava | 2 | 1 |
| **2** | 4 | 5 | 2 | 0 | 28 | 13 | nadváha | −0,5 | sport | | |
| **3** | 6 | 7 | 49 | 15 | 3 | 0 | normální | −6,5 | spánek | | |
| **4** | | | 0 | 0 | 0 | 11 | obezita | | | | |
| **5** | | | 2 | 0 | 28 | 2 | nadváha | | | | |
| **6** | 8 | 9 | 4 | 13 | 0 | 0 | podváha | −3,5 | počet | | |
| **7** | | | 45 | 2 | 3 | 0 | normální | | | | |
| **8** | | | 1 | 13 | 0 | 0 | podváha | | | | |
| **9** | | | 3 | 0 | 0 | 0 | normální | | | | |

Tabulka 1: Popis uzlů klasifikačního stromu vytvořeného pomocí algoritmu C&RT vyčerpávajícího prohledávání

| Node | Node branch | Left branch | Right normální | n in cls podváha | n in cls nadváha | n in cls obezita | n in cls Predict. | Split | Split | Split | Split |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 3 | 51 | 15 | 31 | 13 | normální | −3,57773 | televize | | |
| **2** | 4 | 5 | 48 | 10 | 7 | 0 | normální | −5,90627 | spánek | | |
| **3** | 6 | 7 | 3 | 5 | 24 | 13 | nadváha | −4,32758 | sport | | |
| **4** | | | 0 | 4 | 0 | 0 | podváha | | | | |
| **5** | 8 | 9 | 48 | 6 | 7 | 0 | normální | | strava | 3 | 4 |
| **6** | 10 | 11 | 2 | 3 | 24 | 13 | nadváha | −3,49525 | sport | | |
| **7** | | | 1 | 2 | 0 | 0 | podváha | | | | |
| **8** | 12 | 13 | 47 | 6 | 2 | 0 | normální | −2,33213 | počet | | |
| **9** | | | 1 | 0 | 1 | 0 | normální | | | | |
| **10** | 14 | 15 | 1 | 3 | 23 | 13 | nadváha | −2,05128 | sport | | |
| **11** | | | 1 | 0 | 1 | 0 | normální | | | | |
| **12** | | | 0 | 3 | 1 | 0 | podváha | | | | |
| **13** | | | 47 | 3 | 1 | 0 | normální | | | | |
| **14** | 16 | 17 | 0 | 3 | 21 | 13 | nadváha | −0,634035 | sport | | |
| **15** | | | 1 | 0 | 2 | 0 | nadváha | | | | |
| **16** | | | 0 | 1 | 0 | 11 | obezita | | | | |
| **17** | 18 | 19 | 0 | 2 | 21 | 2 | nadváha | | strava | 1 | 2 |
| **18** | | | 0 | 0 | 21 | 2 | nadváha | | | | |
| **19** | | | 0 | 2 | 0 | 0 | podváha | | | | |

Tabulka 2: Popis uzlů klasifikačního stromu vytvořeného na základě algoritmu QUEST

| Node | Left branch | Right branch | Size of node | N in class normální | N in class podváha | N in class nadváha | N in class obezita | Selected | Split | Split | Split | Split |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 3 | 110 | 51 | 15 | 31 | 13 | normální | strava | | 4 | 3 |
| **2** | 4 | 5 | 67 | 49 | 15 | 3 | 0 | normální | spánek | 6,5 | | |
| **4** | 6 | 7 | 17 | 4 | 13 | 0 | 0 | podváha | počet | 3,5 | | |
| **6** | | | 14 | 1 | 13 | 0 | 0 | podváha | | | | |
| **7** | | | 3 | 3 | 0 | 0 | 0 | normální | | | | |
| **5** | 10 | | 11 | 50 | 45 | 2 | 3 | normální | počet | 2,5 | | |
| **5** | 10 | | 11 | 50 | 45 | 2 | 3 | normální | počet | 2,5 | | |
| **10** | | | 4 | 0 | 2 | 2 | 0 | podváha | | | | |
| **11** | | | 46 | 45 | 0 | 1 | 0 | normální | | | | |
| **3** | 14 | 15 | 43 | 2 | 0 | 28 | 13 | nadváha | sport | 0,5 | | |
| **14** | | | 11 | 0 | 0 | 0 | 11 | obezita | | | | |
| **15** | 16 | 17 | 32 | 2 | 0 | 28 | 2 | nadváha | sport | 3,5 | | |
| **16** | | | 30 | 0 | 0 | 28 | 2 | nadváha | | | | |
| **17** | | | 2 | 2 | 0 | 0 | 0 | normální | | | | |

Tabulka 3: Popis uzlů klasifikačního stromu vytvořeného na základě standardní metody C&RT z modulu Data Mining

# MODELOVÁNÍ ROČNÍHO CHODU UVB ZÁŘENÍ V ANTARKTIDĚ

**Marie Budíková a Ladislav Budík**

**Abstract**: One of the issues resolved within the Czech-Ukrainian scientific cooperation implemented on the Vernadsky Station (formerly the British Faraday Station) in Antarctica since 2002 is the measurement of different radiation fluxes (i.e. global solar radiation intensity and UV radiation intensity). UVB radiation measurements were extended also by the erythemally effective radiation. This paper submits some results of modelling intensities of erythemally effective UVB radiation at the level of daily sums in relation to the ozone concentrations, extraterrestrial and surface intensities of global solar radiation and to the extraterrestrial intensities of UVB radiation.

Two types of UVB radiation absorption models are compared in this contribution. Linear and nonlinear (with hyperbolic principle of UVB radiation absorption) regression models are examined and their quality and limitations are discussed.

## Úvod

Téma přednášky vychází z naší účasti v grantu „Vliv atmosférických faktorů na režim UV záření v prostoru Antarktického poloostrova“, jehož řešitelem je prof. RNDr. Pavel Prošek, CSc. z Geografického ústavu PřF MU. Zpracovávaná data pocházejí z ukrajinské polární stanice Vernadskij. Nyní již máme k dispozici i data z nově vybudované české polární stanice J. G. Mendela.

Stanice Vernadskij je umístěna na 65°15’S, 64°16’W na ostrově Galindéz, který je součástí Argentinského souostroví na západním pobřeží Antarktického poloostrova. Měří se zde jak množství ozónu, tak globální záření a UV záření.

## UV záření a jeho vlastnosti

Sluneční záření, které dopadá na zemský povrch, se z hlediska vlnové délky dělí do čtyř skupin: radiové záření má největší vlnovou délku, od 1 mm výše. Infračervené záření má délky od 730 nm do 1000 nm, viditelné záření od 400 nm do 730 nm a UV záření pak od 200 do 400 nm. Slunce v menší míře emituje i rentgenové záření a gama záření, avšak to se díky absorpci atmosférou nedostane až na zemský povrch.

UV záření tvoří asi 7 % celkového slunečního záření. Je výrazně pohlcováno kyslíkem a ozónem v atmosféře, na zemský povrch ho dopadne jenom asi třetina původního množství. Z hlediska vlnové délky rozdělujeme UV záření na UVA, UVB a UVC záření.

UVA záření proniká hluboko do kůže a způsobuje její předčasné stárnutí. Proniká sklem. Podílí se na tvorbě volných radikálů, narušuje činnost imunitního systému.

UVB záření je zhoubné pro živé organizmy. Je schopno rozkládat nebo narušovat životně důležité organické sloučeniny a podílet se na vzniku rakoviny kůže. Má velmi negativní vliv na zrak, dokáže poničit tyčinky a čípky, gangliové buňky a nervová zakončení v rohovce (tzv. „sněžná slepota“). Jednobuněčné organizmy dokáže zničit zcela. Proniká i vodou, ale jen do hloubky několika metrů (kde je však soustředěna většina podvodních organizmů). UVB záření též negativně ovlivňuje růst zelených rostlin, účinnost fotosyntézy, ale i třeba celkovou plochu jejich listů.

UVC záření je zatím zcela pohlcováno v atmosféře, tudíž na zemský povrch nedopadá. V důsledku zeslabování ozónové vrstvy stoupá obava z jeho negativních vlivů na organizmy.

V důsledku zeslabování ozónové vrstvy stoupá obava z negativních vlivů UV záření na organizmy.

## Základní údaje o ozónu

Ozón je přirozenou součástí atmosféry, je to tříatomová molekula kyslíku. Ve stratosféře chrání život na Zemi před zhoubným zářením, v troposféře má však toxický vliv na živé organizmy, poškozuje dýchací orgány živočichů a rostlin.

Množství ozónu v atmosféře závisí na rovnováze procesů, které ozón produkují s procesy, které ho v atmosféře ničí. V poslední čtvrtině 20. století bylo zjištěno, že právě stratosférického ozónu ubývá a naopak troposférického ozónu přibývá.

Celkové množství ozónu ve vertikálním sloupci atmosféry nad zemským povrchem, zkráceně celkový ozón, se udává v tzv. Dobsonových jednotkách. Jedna DU je definována jako tloušťka tisíciny milimetru vrstvy ozónu při tlaku 1013 hPa a teplotě 273 K.

Rozložení průměrných hodnot celkového ozónu v atmosféře závisí na zeměpisné šířce, ročním období a výšce nad zemským povrchem. Hodnoty celkového ozónu rostou od rovníku směrem k vyšším šířkám, mají jednoduchý roční chod s maximem na jaře a minimem na podzim. Rovnováha stratosférického ozónu je narušována antropogenní činností, kdy se do stratosféry

dostávají molekuly chlóru a brómu, které rozkládají ozón. Extrémním případem zeslabení ozonové vrstvy je tzv. ozonová anomálie, která byla poprvé pozorována nad Antarktidou.

## Proces vytváření lineárního regresního modelu

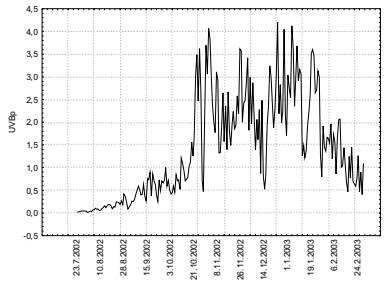
Do prvého kontaktu s daty jsem se dostala v r. 2004, kdy se na mě obrátili kolegové prof. Prošek a dr. Láska z Geografického ústavu naší fakulty se žádostí, abych vytvořila regresní model, který by umožnil predikovat hodnoty UVB záření na povrchu Země v závislosti na hodnotách UVB záření na horní hranici atmosféry, oblačnosti a ozónu. Data pocházela ze stanice Vernadskij z let 2002 a 2003. (Zde se budeme zabývat hodnotami měření od 23. 7. 2002 do 28. 2. 2003.) Výsledky tohoto zpracování byly publikovány jako kapitola The Regime of Total and Biological Effective Ultraviolet Radiation at Vernadsky Station (Argentine Islands, Antarctica) and the Impact of Ozone and Cloudiness in 2002 and 2003 v knize Czech Geography at the Dawn of Milenium, kterou vydala v r. 2004 Palackého univerzita v Olomouci.

Označení sledovaných proměnných:

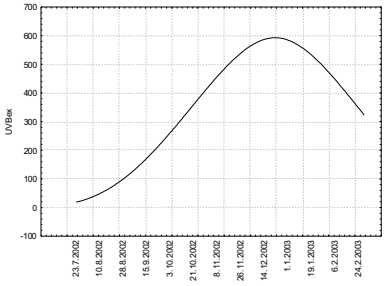
- $\mathrm{UVB}_p$ ... denní úhrnná hodnota UVB záření na povrchu Země (v $\mathrm{kJ.m^{-2}.d^{-1}}$)
- $\mathrm{UVB}_{ex}$ ... denní úhrnná hodnota UVB záření na horní hranici atmosféry (v $\mathrm{kJ.m^{-2}.d^{-1}}$)
- Ozon ... průměrné denní množství ozónu (v Dobsonových jednotkách, pozemní měření)
- Obl ... průměrné denní pokrytí oblohy oblačností (v osminách oblohy)

Podívejme se na časový průběh sledovaných proměnných, v období mezi 23. 7. 2002 – 28. 2. 2003.

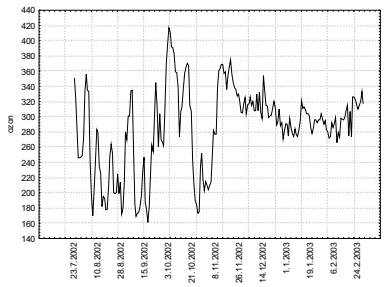
Průběh $\mathrm{UVB}_p$



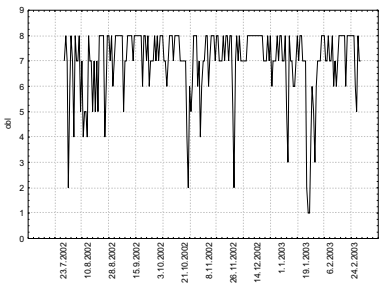

Průběh $\mathrm{UVB}_{ex}$



Průběh ozónu (pozemní měření)

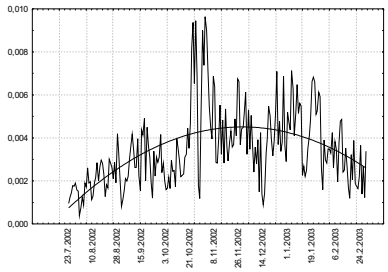



Průběh oblačnosti

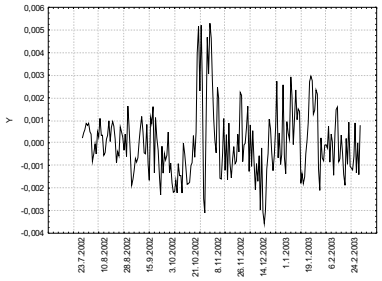



Vzhledem k tomu, že proměnná $\mathrm{UVB}_p$ vykazuje výraznou heterskedasticitu, jeví se vhodnější modelovat nikoliv přímo proměnnou $\mathrm{UVB}_p$, ale podíl $\mathrm{UVB}_p/\mathrm{UVB}_{ex}$. Průběh tohoto podílu však má zřetelný parabolický trend. Ten odstraníme, získáme tím závisle proměnnou veličinu $Y$ a budeme se zabývat lineárním modelem $Y = \beta_0 + \beta_1 \cdot \mathrm{Ozon} + \beta_2 \cdot \mathrm{Obl} + \varepsilon$.

Podíl $\mathrm{UVB}_p/\mathrm{UVB}_{ex}$ s parabolickým trendem




$Y = \mathrm{UVB}_p/\mathrm{UVB}_{ex}$ mínus parabolický trend



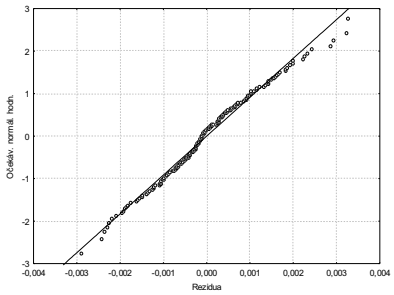

Odhady parametrů s 95% intervaly spolehlivosti (v tisícinách)

| Parametr | Odhad | Dolní mez | Horní mez | $t(218)$ | $p$-hodnota |
|---|---|---|---|---|---|
| $\beta_0$ | 6,866 | 5,928 | 7,804 | 14,4165 | <0,0001 |
| $\beta_1$ | −0,012 | −0,009 | −0,015 | −7,8893 | <0,0001 |
| $\beta_2$ | −0,550 | −0,654 | −0,446 | −10,2940 | <0,0001 |

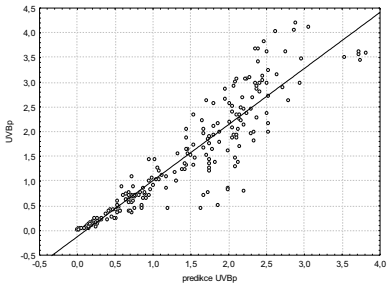
Ve všech případech zamítáme na hladině významnosti 0,05 hypotézy o nevýznamnosti regresních parametrů, odpovídající $p$-hodnoty jsou vždy velmi blízké 0. Podíl rozptylu závisle proměnné veličiny $\mathrm{UVB}_p$ je modelem vysvětlen z 49,8 %.

Kvalitu modelu posoudíme jednak pomocí normálního pravděpodobnostního grafu reziduí a jednak grafem závislosti hodnot $\mathrm{UVB}_p$ na predikovaných hodnotách $\mathrm{UVB}_p$.

N-P graf reziduí




Závislost $\mathrm{UVB}_p$ na predikci $\mathrm{UVB}_p$




Vidíme, že model má značné nedostatky, které jsou zvláště patrné na dvourozměrném bodovém diagramu. Zřetelně se zde projevuje heteroskedasticita.

## Proces vytváření nelineárního regresního modelu

V této fázi vstoupil do zpracování dat můj manžel, který je původním povoláním fyzik, ale již od začátku 90. let pracuje jako hydrolog v brněnské pobočce ČHMÚ.

Jeho první návrh spočíval v tom, že hodnoty ozónu je třeba přepočítat na hodnoty tzv. efektivního ozónu, neboť je nutné vzít do úvahy střední denní úhlovou výšku Slunce nad obzorem. Výsledná hodnota efektivní koncentrace ozónu se získá řešením problému průchodu paprsku kulovými vrstvami (řešení v rovině, které je jednodušší, dává při malých výškách Slunce výrazně vyšší hodnotu koncentrace):

$$\mathrm{E\_O} = \frac{-2r \cdot \sin\alpha + \sqrt{(2r \cdot \sin\alpha)^2 + 4[(r+h)^2 - r^2] \cdot (1 + \sin^2\alpha)}}{2(1 + \sin^2\alpha)} \cdot \frac{\mathrm{Ozon}}{h},$$

kde E_O ... je efektivní množství ozónu v Dobsonových jednotkách,
$r$ ... je poloměr zeměkoule v km,
$h$ ... je střední výška ozonosféry nad zemským povrchem v km,

$\alpha$ ... je střední denní úhlová výška Slunce nad obzorem v radiánech,
Ozon ... je množství ozónu v Dobsonových jednotkách.

Druhý návrh se již týkal samotného modelu. Hodnoty pozemního UVB záření se vysvětlují pomocí nelineárního modelu s hyperbolickým útlumem UV záření:

$$\mathrm{UVB}_p = \left(\frac{\mathrm{GZ}_p}{\mathrm{GZ}_{ex}}\right)^{\beta_0} \frac{\mathrm{UVB}_{ex}}{1 + \beta_1 \cdot \mathrm{E\_O}^{\beta_2}},$$

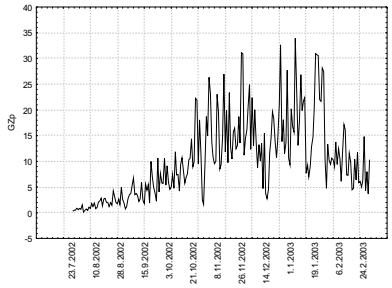
kde $\mathrm{GZ}_p$ ... denní úhrnná hodnota celkového slunečního záření na povrchu Země (v $\mathrm{MJ.m^{-2}.d^{-1}}$)
$\mathrm{GZ}_{ex}$ ... denní úhrnná hodnota celkového slunečního záření na horní hranici atmosféry (v $\mathrm{MJ.m^{-2}.d^{-1}}$)

Oproti původnímu modelu, kde vystupovalo průměrné denní pokrytí oblohy oblačností, zde vystupuje podíl $\mathrm{GZ}_p/\mathrm{GZ}_{ex}$. Díky použití globálního záření model již vlastně zahrnuje oblačnost.

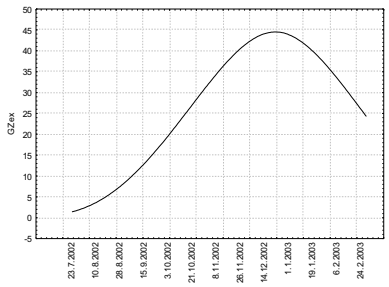
V modelu figurují regresní parametry $\beta_0, \beta_1, \beta_2$. Parametr $\beta_0$ aproximuje odlišnost absorpce globálního záření a UVB záření v atmosféře, parametry $\beta_1, \beta_2$ vyjadřují efektivitu ozonosféry při absorpci a rozptylu UVB záření.

Časový průběh nově použitých proměnných $\mathrm{GZ}_p$ a $\mathrm{GZ}_{ex}$ vidíme na následujících dvou grafech:

Průběh $\mathrm{GZ}_p$




Průběh $\mathrm{GZ}_{ex}$




Parametry $\beta_0, \beta_1, \beta_2$ nelineárního hyperbolického regresního modelu odhadujeme pomocí systému STATISTICA Levenbergovou-Marquardtovou metodou za použití metody nejmenších čtverců s počátečními aproximacemi všech tří parametrů rovnými 0,1.

Odhady parametrů s 95% intervaly spolehlivosti

| Parametr | Odhad | Dolní mez | Horní mez | $t(218)$ | $p$-hodnota |
|---|---|---|---|---|---|
| $\beta_0$ | 0,6704 | 0,6459 | 0,6949 | 53,9454 | <0,0001 |
| $\beta_1$ | 0,1055 | 0,0618 | 0,1492 | 4,7548 | <0,0001 |
| $\beta_2$ | 1,0920 | 1,0267 | 1,1573 | 32,9573 | <0,0001 |

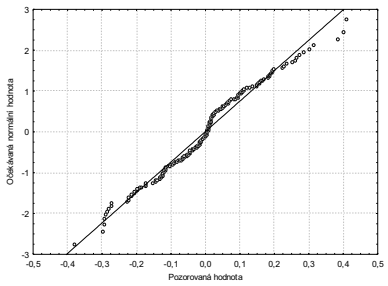
Ve všech případech zamítáme na hladině významnosti 0,05 hypotézy o nevýznamnosti regresních parametrů, odpovídající $p$-hodnoty jsou vždy velmi blízké 0. Podíl rozptylu závisle proměnné veličiny $\mathrm{UVB}_p$ je modelem vysvětlen z 98,7 %. Oproti původnímu lineárnímu modelu vidíme obrovské zlepšení – v něm byla variabilita závisle proměnné veličiny vysvětlena jenom z necelých 50 %.

Výsledky nelineárního regresního modelu s hyperbolickým útlumem UVB záření byly prezentovány v srpnu 2007 na konferenci TIES 2007 v Mikulově ve formě posteru.

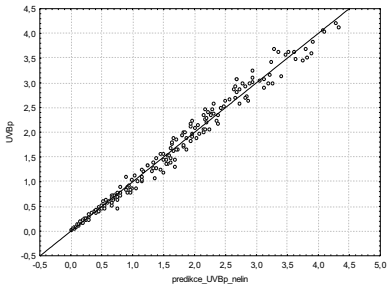
Co se týká rozložení reziduí, je stále dosti vzdálené od normálního rozložení. Rozložení reziduí je kladně zešikmeno a jeho špičatost je podstatně větší než by odpovídalo normálnímu rozložení. Všechny tři testy normality, které poskytuje systém STATISTICA, zamítají hypotézu o normalitě na hladině významnosti 0,05.

Graf závislosti hodnot $\mathrm{UVB}_p$ na predikovaných hodnotách $\mathrm{UVB}_p$ vypadá podstatně lépe než u lineárního modelu, heteroskedasticita je jenom nepatrná.

N-P graf reziduí




Závislost $\mathrm{UVB}_p$ na predikci $\mathrm{UVB}_p$




## Výhledy do budoucna

V následujícím období bychom se rádi soustředili na vylepšení modelu pomocí transformace dat před použitím metody nejmenších čtverců. Máme také

v úmyslu kromě pozemních měření ozónu použít družicová měření, která provádí družice TOMS.

Jak plyne ze zaměření grantu, budeme rovněž zpracovávat data z nové české stanice J. G. Mendela a dosažené výsledky porovnávat s výsledky pro stanici Vernadskij.

Jedním z úkolů je také sestavit model pro predikci $UVB_p$, kde místo poměru extraterestriálního globálního záření a pozemního globálního záření budou vystupovat jednotlivé atmosférické jevy (např. plocha a druh oblačnosti) a regresní rovnice bude vysvětlovat vztah mezi $UVB_p$ a těmito meteorologickými prvky.

*Adresa*: Marie Budíková, Ústav matematiky a statistiky PřF MU, Janáčkovo nám. 2a, 602 00 Brno, email: budikova@math.muni.cz.
Ladislav Budík, ČHMÚ, pobočka Brno, Kroftova 43, 616 67 Brno, email: budik@chmi.cz.

# ÚLOHA A VÝZNAM SCIENTOMETRIE V HODNOCENÍ VĚDECKÉ PRÁCE

**Pavel Drábek**

V dnešní době se ve sdělovacích prostředcích můžeme seznámit se statistikami všeho druhu. Typickým příkladem jsou nejrůznější průzkumy veřejného mínění, nebo statistiky zaměřené do konkrétních oblastí jako je ekonomika, kultura, sport atd. Současný stav výpočetní techniky a její bouřlivý rozvoj nám umožňuje zpracovávat stále větší množství dat a získávat z nich nejrůznější informace. Během fotbalového přenosu se tak například můžeme dozvědět, kolikrát v loňské sezóně Pavel Nedvěd vystřelil na branku, kolik těchto střel brankář kryl a kolikrát Nedvěd skóroval. Na základě spolehlivých statistik je možné zmapovat vývoj událostí v minulosti, předpovědět tendence vývoje v budoucnosti a potom případně budoucnost žádoucím směrem ovlivnit.

Pokud se s takovými statistikami nakládá rozumně a s vědomím všech jejich předností, ale i nedostatků, pak mohou být velmi prospěšné a mohou ušetřit hodně peněz i lidské námahy. Na druhou stranu, vyvozování závěrů ze statistických dat bez dostatečné analýzy jejich vypovídajících schopností, může být nejen velmi zavádějící, ale ve svých důsledcích i nebezpečné. Typickým příkladem je celosvětová snaha umět změřit účinnost vědecké a výzkumné práce tak, aby se neplýtvalo finančními prostředky, které na ni jsou

vynakládané. V dobré víře tento proces maximálně zobjektivizovat, byl vytvořen systém ukazatelů, jako je například impaktní faktor, medián oboru, citační index atd. Na základě těchto ukazatelů je možné zjistit, v jaké intenzitě příslušný vědec publikuje, v jak kvalitních časopisech publikuje nebo jak často je citován a kolik citací připadá v průměru na jednu jeho publikovanou práci. Na základě toho se pak měří jeho takzvaný vědecký výkon, případně výkon celého vědeckého týmu, celé univerzity či ústavu nebo, dokonce, celého státu. Není pochyb o tom, že tyto údaje jsou cennými informacemi pro všechny, kdo jsou zodpovědní za účelné vynakládání finančních prostředků na vědu, zejména pokud jde o peníze daňových poplatníků. Každý „zodpovědný“ by si však měl být zároveň vědom toho, že s těmito informacemi je třeba zacházet s maximální opatrností. V mnoha vědních oborech se skutečná účinnost nedá měřit v krátkodobých horizontech volebních období politiků a vědecké výsledky, které později přinesou velký užitek nelze dopředu naplánovat. Nelze ani spolehlivě zajistit to, že míra peněz do různých oblastí vědy a výzkumu investovaných se adekvátně odrazí v míře užitku z těchto oblastí plynoucích. Investice do vědy a výzkumu vždycky byla, je a bude dost riskantním podnikem. Jde o to taková rizika v maximální míře minimalizovat, ale přitom takovým způsobem, abychom v dobré víře nenapáchali více škody než užitku.

Fundamentalistický způsob prosazování kvantifikovaného systému hodnocení vědy a výzkumu má řadu negativních důsledků. Člověk je tvor přizpůsobivý a tak jsme dnes svědky toho, že vědečtí pracovníci ve snaze o získání finančních prostředků publikují výsledky své práce ne proto, že chtějí odborné veřejnosti něco nového sdělit, ale proto, aby získali patřičný počet hodnotících bodů. Pokud autor dospěje k výsledku, obvykle se jej logicky snaží prodat za co největší počet bodů a místo jedné publikace výsledek opublikuje ve více článcích. Aby se nedopustil plagiátorství své vlastní práce, výsledek je publikován po částech, což stěžuje orientaci potenciálních čtenářů. Celkově jsme svědky rychlého nárůstu vědeckých prací, které nikdo nečte a nemá čas je ani podrobně zrecenzovat. Jsem přesvědčen o tom, že je to jeden z důsledků nyní tak široce uplatňovaného kvantifikovaného systému hodnocení vědy a výzkumu.

Vedle výhod, které nám scientometrie přináší a které spočívají v tom, že na jejím základě je možné „odfiltrovat“ pracovníky a pracoviště, kteří evidentně nic neprodukují a přitom čerpají finanční prostředky, je třeba vidět i výše uvedená úskalí i rizika plynoucí z příliš jednostranného způsobu její implementace ve vědecké sféře. Snad nejhorší může být bodovací tabulka, ve které jsou vědci či jejich týmy seřazeni do lineárního pořadí podle počtu získaných bodů a na základě tohoto pořadí se rozhodne, kdo je lepší a kdo je

horší. Bylo by zajímavé zjistit, jak by si v současném systému hodnocení vědecké výkonnosti vedl například Kurt Gödel, který publikoval poměrně málo prací, ty však podstatně ovlivnily celou matematiku dvacátého století. Je zřejmé, že scientometrické hodnocení je pouze jednou složkou komplexního pohledu na vědeckou výkonnost a může mít v různých oborech různou váhu a vypovídající schopnost. Musí však být vždy důsledně konfrontováno s názorem obecně uznávaných „autorit" jednotlivých oborů. Co se týká výhledu do budoucnosti, jsem však spíše skeptik, a to z následujících důvodů:

1. Pro poskytovatele finančních prostředků je pohodlné zdůvodnit svoje rozhodnutí na základě „učeně se tvářících" a „objektivně vypadajících" scientometrických údajů. Kromě toho jsou podobné systémy užívány hojně i ve světě, mnoho lidí se tím zaštiťuje a někteří upřímně věří v jejich spolehlivost.
2. Pro mnoho lidí, ve vědě působících, je pohodlnější soustředit se na sbírání bodů než na výzkum samotný.
3. Pro mnoho vedoucích vědeckých pracovníků je kvantifikované hodnocení vítaným nástrojem řízení. Z vlastní zkušenosti vím, jak je těžké a nepříjemné svému podřízenému sdělit, že s jeho prací nejsem spokojen. Je naopak mnohem snazší říci, že „to" vyplývá z objektivních kritérií.
4. Řada výkonných vědců, zejména z oborů, kde scientometrie skutečně poskytuje poměrně přesné informace o vědecké zdatnosti, bohužel prosazuje její užívání „plošně", a to i v oborech, kde je její vypovídající hodnota přinejmenším pochybná. Jejich přirozená autorita tím nahrává lidem uvedeným v bodech 1, 2 a 3 a tím jen prohlubuje problémy, které z toho pramení.

Na závěr bych se rád vrátil k přirovnání s fotbalem. Legendárního Franze Beckenbauera jistě nebude počet zápasů na jeden vstřelený gól řadit na přední místa fotbalových statistik. Možná ani počet zákroků, kterými soupeři zabránil gól vstřelit. Přesto si vysloužil přezdívku „Kaiser Franz". Bylo to proto, že když byl v akci, divák viděl fotbalovou krásu a umění. On totiž tvořil hru! Jsem přesvědčen o tom, že vedle scientometrických údajů, které o současné vědě hodně vypovídají, je třeba vzít velmi vážně v potaz také to, že podstatný vliv na vývoj vědy mají lidé, kteří „tvoří hru", ale možná nejsou na předních místech scientometrických statistik.

V Plzni dne 21. 2. 2007.

*Poznámka.* Tento článek vyšel v Bulletinu AV číslo č. 4 v roce 2007.

# KONTROLA ROZSÁHLÝCH DOTAZNÍKŮ Z HLEDISKA LOGICKÝCH VAZEB

**Zdeněk Hlávka, Daniel Hlubinka a Michal Kulich**

**Abstrakt**: Na velkém datovém souboru si ukážeme, jak automatizovat kontrolu logických vazeb v dotazníku. Vytvoříme pomocné databáze, z nichž bude jasné, zda nějaké pozorování je chybějící (nebo naopak je vyplněno, přestože má být přeskočeno). Ukážeme si, že ve velkém dotazníku můžeme očekávat netriviální komplikace a jak si s nimi lze poradit. Na závěr pochválíme tazatele za dobře odvedenou práci.

## 1. Shoda struktury dotazníku a vyplněných údajů

### 1.1. Základní úkol

Budeme se zabývat otázkou, jak zjistit, zda vyplněné údaje a vnitřní struktura dotazníku jsou v logické shodě. Jako jednoduchý příklad uveďme dvojici otázek

1. Váš rodinný stav (svobodný/á, ženatý/vdaná, rozvedený/á, ovdovělý/á)
2. Datum sňatku (den, měsíc a rok)

Zjevně nemá smysl odpovídat na druhou otázku, jestliže je odpověď na první otázku svobodný/á. V takovém případě potřebujeme označit „chybějící“ pozorování jako *správně přeskočené*. Na druhou stranu, bude-li datum sňatku přesto vyplněno, pak toto nechybějící pozorování musí být označeno jako *neočekávaně vyplněné*. V tuto chvíli pomineme budoucí či jinak zjevně nemožná data sňatků a soustředíme se pouze na strukturu dotazníku a shodu s daty.

Někdy bývá zvykem případně doplnit chybějící odpověď, je-li z ostatních odpovědí možné chybějící údaj zjistit. Všimněme si, že to v předchozím příkladu nelze; datum sňatku může vyplnit každý ženatý, ovdovělý i rozvedený. Nicméně zdůrazněme hned na začátku, že naším úkolem není hádat či určovat, co je správně vyplněné, případně čím by mohlo být chybějící pozorování logicky doplněno. To případně můžeme zkusit u dalších dílčích studií prováděných na této databázi. Původní databázi však měnit svévolně nesmíme, bude s ní pracovat ještě (doufejme) mnoho dalších lidí a musí dostat data tak, jak byla posbírána.

Běžný dotazník má zásadní výhodu, že je vyplňován „lineárně“, tedy odpověď na pozdější otázku již nemůže ovlivnit fakt, zda předchozí otázka má,

nebo nemá být zodpovězena. Tato zdánlivá trivialita značně zjednoduší řešení našeho úkolu. Pro dotazník s jen několika otázkami lze samozřejmě postupovat přímo a zkoumat otázku po otázce. Snad se nám podaří čtenáře během čtení článku přesvědčit, že v našem případě již přímý postup téměř není možný a co nejobecnější a naprogramovatelný přístup je téměř jediný možný.

## 1.2. Hlavní myšlenka řešení

Myšlenka řešení je do značné míry uzpůsobena zvolenému prostředí R. Mějme datový rámec (data frame) `data` obsahující $k$ proměnných a $n$ měření. Prvním krokem je vytvoření „stínových booleovských dat“ o stejném rozsahu. Vytvořme nejprve matici (rámec) `shadow` o $k$ sloupcích a $n$ řádcích obsahující samé hodnoty `TRUE`. Zjistíme-li během procházení databáze `data`, že nějaká hodnota má být z nějakého důvodu přeskočena, na jejím místě ve stínových datech vložíme `FALSE`.

Porovnáním naměřených a stínových dat pak okamžitě zjistíme, zda

1. hodnota má být (stínová hodnota je `TRUE`) a je vyplněna (to je správný stav), nebo
2. hodnota nemá být (stínová hodnota je `FALSE`) a není vyplněna (to je také správně), nebo
3. hodnota má být, ale není vyplněna (chybějící pozorování), nebo
4. hodnota nemá být, ale je vyplněna (neočekávané pozorování).

Tyto čtyři možnosti jsou disjunktní.

## 1.3. Komplikace

Zatím nevíme, jak efektivně a správně automatizovat překlápění hodnot `TRUE/FALSE` ve stínových datech. Tomu se budeme věnovat o něco později. Teď si však musíme všimnout nepříjemné komplikace. O tom, zda nějakou hodnotu v datech očekáváme, nebo nikoliv, možná rozhoduje některá z předchozích proměnných; a ta může chybět! Někdy bychom sice mohli hodnotu této chybějící rozhodovací proměnné „vypočítat“ z ostatních, ale tento svévolný zásah do dotazníku jsme si zakázali.

Z tohoto důvodu zavádíme další pomocné stínové proměnné `shadna` (shadow – not available), které indikují fakt, že o dané hodnotě nejsme schopni rozhodnout, zda má, nebo nemá být přítomna. Nejprve celou $n \times k$ matici (rámec) `shadna` vyplníme hodnotami `TRUE`. Pochopitelně musíme ošetřit skutečnost, že rozhodovací proměnná může být správně vynechána (její stínová

hodnota je `FALSE`). V tom případě by ale i zkoumaná proměnná měla být přeskočena (ve „stínu“ také `FALSE`), jinak bychom měli logicky nekonzistentní dotazník. Je-li tedy stínová hodnota kontrolované proměnné `TRUE` (zatím nebyl důvod k jejímu přeskočení), ale rozhodovací proměnná chybí, nastavíme do tabulky `shadna` hodnotu `FALSE`.

Tím se nám ovšem čtyři možnosti rozrostou na osm, z nichž v pořádku jsou jen první dvě. Jedná se o stav, kdy

1. hodnota má být a je vyplněna (v pořádku), nebo
2. hodnota nemá být a není vyplněna (v pořádku), nebo
3. hodnota má být, ale není vyplněna (chybějící pozorování), nebo
4. hodnota nemá být, ale je vyplněna (neočekávané pozorování), nebo
5. rozhodovací proměnná není vyplněna, hodnota je vyplněna a není jiný důvod k jejímu přeskočení, nebo
6. rozhodovací proměnná není vyplněna, hodnota není vyplněna a není jiný důvod k jejímu přeskočení, nebo
7. rozhodovací proměnná není vyplněna, hodnota je vyplněna a je jiný důvod k jejímu přeskočení, nebo
8. rozhodovací proměnná není vyplněna, hodnota není vyplněna a je jiný důvod k jejímu přeskočení.

Zdálo by se, že poslední dva body nemohou nastat, ale někdy jsou dotazníky natolik komplikované, že, kupodivu, mohou. Je to ale dáno určitým „porušením pravidel“, kdy přeskakované bloky nejsou do sebe vnořeny, ale mohou se ve výjimečných případech jen částečně překrýt. Na to by si sice měl sestavovatel dotazníku dát pozor, nicméně to nemusí být tak snadné uhlídat.

## 2. Postup při řešení

Naším úkolem teď je vytvořit program, který k dotazníku s nějakou vnitřní strukturou dokáže vytvořit dva soubory stínových proměnných. Tyto stínové proměnné mohou být využity k vyhledávání problematických míst v dotazníku. Hlavním účelem je ale nakonec tvorba přehledných tabulek, které popíšeme v závěru článku.

Jednotlivé postupy a potíže budeme ilustrovat na skutečném, velmi podrobném, dotazníku. Účelem dotazníku je zejména zjistit, zda a jak se dotazovaný chová rizikově ve vztahu k HIV/AIDS a také jaké má povědomí o této chorobě.

Jak jsme již uvedli, využijeme jednosměrné vyplňování dotazníku v čase. Vycházíme z toho, že jestliže nějaká, takzvaná *rozhodovací*, proměnná nabývá určitých hodnot, pak několik následujících proměnných nebude vyplněno.

*Příklad* 1. Otázky na konzumaci alkoholu:

Q1. Pil jste ve svém životě alkohol? ANO/NE
Q2. V kolika dnech z posledních 30 jste pil alkohol? UVEĎTE POČET
Q3. V kolika dnech z posledních 30 jste pil alkohol až do opilosti? UVEĎTE POČET

Je jasné, že při záporné odpovědi na otázku 1 jsou obě následující přeskočeny, při odpovědi 0 na druhou otázku je přeskočena otázka třetí. Ve všech případech se pak dostáváme k otázce čtvrté, na kterou by mělo být odpovězeno v každém případě.

Q4. Užil jste ve svém životě drogu? ANO/NE

Rozhodli jsme se, že nakonec vyjdeme z tabulek, které budou udávat: *rozhodovací proměnnou; první nepřeskočenou proměnnou; podmínky pro přeskočení.*

Vraťme se k příkladu 1. Odpověď ANO/NE je faktorová proměnná, úroveň faktoru 0 odpovídá NE, úroveň 1 odpovídá ANO. Naše tabulka závislostí tedy bude mít první dva řádky následující.

```
Q1   Q4   0
Q2   Q4   0
```

Takto ošetříme všechny jednoduché typy závislostí. Všimněme si, že jsme schopni jednotně zpracovat jak faktorové, tak i numerické proměnné.

## 2.1. Rozhodování na základě více odpovědí

Při zkoumání dotazníku brzy zjistíme, že naše fantazie nedokáže podchytit všechny možné komplikace předchozího postupu. První je rozhodování o dalším postupu na základě více odpovědí. Zde je příklad.

*Příklad* 2. Otázky na užívání drog (část)

Q7. Užil jste ve svém životě heroin injekčně? ANO/NE/NEVÍM
Q7a. V posledních 6-ti měsících? ANO/NE
Q7b. Kolikrát v posledních 30-ti dnech? UVEĎTE POČET DNÍ
Q7c. Ve dnech, kdy užíváte drogu, typicky kolikrát denně? UVEĎTE KOLIKRÁT DENNĚ
Q8. Užil jste ve svém životě amfetaminy? ANO/NE/NEVÍM

Q8a. V posledních 6-ti měsících? ANO/NE
Q8b. Kolikrát v posledních 30-ti dnech? UVEĎTE POČET DNÍ
Q8c. Ve dnech, kdy užíváte drogu, typicky kolikrát denně? UVEĎTE KOLIKRÁT DENNĚ
Q9. Užil jste ve svém životě amfetaminy injekčně? ANO/NE/NEVÍM
...
Q10. Užil jste v životě jinou drogu než výše uvedenou? ANO/NE
...
INSTRUKCE: Pokud dotazovaný neužíval drogy injekčně (odpovědi na Q7 a Q9 jsou NE), přeskočte na Q15

Kódy faktorových proměnných jsou 1–ANO, 2–NE, 3–NEVÍM. Závislosti vypadají následovně:

```
Q7   Q8   2,3
Q7a  Q8   2
Q8   Q10  2,3
Q8a  Q9   2
Q9   Q10  2,3
Q9a  Q10  2
```

Po odpovědi na otázku jakou jinou drogu dotazovaný užíval je nutné rozhodnout, zda se máme věnovat otázkám týkajícím se užívání jehel. Instrukce hovoří jasně, ovšem jak něco takového můžeme vtělit do našich tabulek, nechceme-li ještě více zobecnit náš postup?

Připomeňme, že budeme procházet otázky jednu po druhé a ve chvíli, kdy narazíme na rozhodovací proměnnou, podíváme se na její hodnotu a rozhodneme, zda následujících několik otázek má, nebo nemá být vyplněno.

Řešení, které volíme, je založeno na pomocné proměnné vložené na správné místo dotazníku. Vytvoříme odpoveď na otázku

A1. Užil jste někdy drogy injekčně? ANO/NE

Zdánlivě je to snadné. Pokud dotazovaný sdělí, že nikdy žádnou drogu nebral, pak všechny odpovědi jsou chybějící a i odpověď na A1. musí chybět. Jinak dáme odpověď NE, pokud obě odpovědi na Q7 i Q9 jsou NE. Drobná potíž, kterou musíme ošetřit je ta, že odpověď na Q9 je podmíněna kladnou odpovědí na Q8. Hodnota A1. je proto NE (úroveň faktoru je 2), pokud

```
Q7 == Ne & (Q9 == Ne | Q8 == Ne | Q8 == Nevim)
```

a do naší tabulky závislostí přidáme na vhodné místo

```
A1   Q15  2
```

Chybějící hodnota A1 je možná pouze v případě, že dotazovaný uvedl, že nikdy žádnou drogu nebral a tento případ je ošetřen již dříve při odpovědi na otázku Q4, kdy máme v tabulce závislostí

```
Q4   Q15  2
```

## 2.2. Rozhodování na základě dřívější odpovědi

Zavedení pomocné proměnné je při námi zvoleném řešení nezbytné také ve chvíli, kdy o zodpovězení otázky rozhodujeme na základě nějaké dřívější otázky. Uveďme opět příklad.

*Příklad* 3. V dotazníku následují otázky na pohlavní život dotazovaného. Potvrdí-li dotazovaný pohlavní styk ve svém dřívějším životě a také v posledních šesti měsících, následují tyto otázky:

Q20. S kolika různými lidmi jste měl v posledních šesti měsících pohlavní styk? UVEĎTE POČET.

Nyní následují otázky týkající se obecně sexuálního života dotazovaného a poté

Q24. Jste ochoten/ochotna mi říci o vašich partnerech více detailů? ANO/NE
INSTRUKCE: Je-li počet partnerů uvedený v otázce Q20 větší než 5, vyberte pět nejposlednějších. Je-li počet nejvýše 5 uveďte všechny. Postupujte od posledních k dřívějším partnerům.
Q25a. Jaké jsou iniciály partnera č. 1
Q26a. Pohlaví partnera 1. MUŽ/ŽENA
...
Q29a. Kolikrát jste se s partnerem milovali v posledních 30 dnech. UVEĎTE POČET.
INSTRUKCE: Je-li počet 0 přejděte na dalšího partnera. Není-li další partner přejděte na Q31.
Q30a. Kolikrát jste při pohlavním styku užili kondom? UVEĎTE POČET.
Q25b. Jaké jsou iniciály partnera č. 2
...
Q31. Byl/a jste v posledních 6 měsících nucen/a k pohlavnímu styku? ANO/NE

Je jasné, že na otázky následující po Q20 nebudeme odpovídat, pokud uvedeme, že jsme neměli žádného milence/milenku v posledním půlroce. Na otázky Q25 až Q30 také nemůžeme čekat odpověď, pokud nám dotazovaný odmítne odpovídat (Q24=NE, což je úroveň faktoru 2). Takže dostáváme závislosti

```
Q20   Q31   0
Q24   Q31   2
```

Prokoušeme-li se až k otázce Q25a, víme, že mají být zodpovězeny otázky Q25a až Q29a. Je-li Q29a $> 0$, pak přejdeme na Q30a. Potom se musíme rozhodnout na základě hodnoty uvedené v Q20, jestli pokračujeme otázkou Q25b (je-li Q20 $> 1$), nebo přejdeme na Q31. Jestli ale máme Q29a $= 0$, pak se musíme rozhodnout okamžitě na základě Q20, co dál. Tento problém jsme vyřešili vložením dalších pomocných proměnných A3a až A3d za odpovědi Q30a až Q30d. Hodnoty A3x jsou shodné s hodnotou Q20. Závislostní tabulka je nyní následující.

```
Q29a   A3a   0
A3a    Q31   1
Q29b   A3b   0
A3b    Q31   2
Q29c   A3c   0
A3c    Q31   3
```

a tak dále. V každém případě, po vyplnění informací o každém milenci/milence narazíme na pomocnou otázku A3x, na jejímž základě rozhodneme, zda pokračovat ve vyplňování informací o dalším partnerovi, nebo zda zbytek této tabulky přeskočíme. Na tuto otázku narazíme nehledě na hodnotu vyplněnou v otázce Q29x.

## 2.3. Aby toho nebylo dost ...

Když jsme úspěšně zvládli výše uvedené potíže, snadno dostaneme pocit, že už nás nic nepřekvapí. Ale překvapí.

Vzhledem k tématu dotazníku chtějí mít tazatelé jasno o sexuálně rizikovém chování dotazovaných. Projeví se to například v důkladném dotazu na dosavadní pohlavní život. V dotazníku tak čelíme následující čtveřici dotazů:

Q15. Měl/a jste v životě vaginální pohlavní styk? ANO/NE/NEVÍM
Q15a. Odpověděl/a jste na otázku 15, že nikoliv, nebo nevíte. Chceme si být naprosto jisti, že vaše odpověď je správná. Váginální pohlavní styk znamená [...]. Měl/a jste kdykoliv v minulosti vaginální pohlavní styk? ANO/NE
Q16. Měl/a jste v životě anální pohlavní styk? ANO/NE/NEVÍM
Q16a. Odpověděl/a jste na otázku 16, že nikoliv, nebo že nevíte. Chceme si být naprosto jisti, že vaše odpověď je správná. Anální pohlavní styk znamená [...]. Měl/a jste kdykoliv v minulosti anální pohlavní styk? ANO/NE

Tak a teď co? Instrukce říkají, že v případě kladné odpovědi na otázku 15 je přeskočena otázka 15a (to už umíme) a v případě kladné odpovědi na otázku 16 je přeskočena otázka 16a. Ovšem pokračování v dotazníku je podmíněno tím, že kladná je odpověď na kteroukoliv z předchozích čtyř otázek (jinak přeskakujeme polovinu dotazníku). Neboli za člověka, který nikdy neměl žádný pohlavní styk považujeme pouze toho, kdo na obě otázky 15a a 16a odpověděl ne. Nezbývá, než vytvořit další pomocnou proměnnou, která nám řekne, zda přeskočíme polovinu dotazníku, nebo pokračujeme dále. Musíme si samozřejmě dobře rozmyslet, jak to uděláme, protože v naprosté většině je jedna z otázek 15a, 16a přeskočena, tudíž chybí, a zoufalý programátor dostane tisíce chybějících pozorování—když se pokusí konstruovat novou proměnnou jako logický součin 15a=NE i 16a=NE. Poté, co zvládneme i tento logický krok, zjistíme, že nemáme ošetřeno, zda nějaká odpověď opravdu nechybí.

Takže nakonec naši pomocnou proměnnou vytváříme dvoukrokově. V prvním kroku vyrobíme hodnoty NE tím, že na otázky 15, 15a, 16, 16a jsou všechny odpovědi negativní. Tím dostaneme spoustu chybějících pozorování, protože při konstrukci

```
(Q15 == Ne | Q15 == Nevim) & (Q16 == Ne | Q16 == Nevim)
& Q15a == Ne & Q16a == Ne
```

stačí, aby jedna z odpovědí chyběla a celá proměnná je prohlášena za chybějící. Musíme teď naopak změnit spoustu v tuto chvíli chybějících pozorování na odpověď ANO. To uděláme pomocí

```
Q15 == Ano | Q15a == Ano | Q16 == Ano | Q16a == Ano
```

Teď už naopak čekáme jakoukoliv komplikaci, nezklameme se, ale naštěstí už je poslední závažná, které čelíme. V dotazníku je několik otázek s možností výběru z více nabídek. U některých je možné vybrat právě jednu možnost, u některých se zaškrtne každá, která je vhodná a u některých se musí vybrat alespoň jedna z možností.

První varianta, kdy se vybírá právě jedna možnost, je snadná. V databázi se tato proměnná vyskytuje jako jedna proměnná, jejíž hodnota udává zvolenou možnost. Snadno zjistíme, zda chybí, nechybí a jakou má hodnotu. Druhá a zejména třetí jsou složitější. Podívejme se opět na příklad. Máme zjistit sociální postavení dané domácnosti a jedním z ukazatelů v Zimbabwe je střešní krytina částí domu, například kuchyně. Zjistíme-li, že domácnost má samostatnou kuchyň, ptáme se na její střešní krytinu:

Q21a Má vaše domácnost kuchyň? A má-li, jaká je střešní krytina vaší kuchyně? (možnosti: tráva, azbest, plech, jiná)

Dotazovaný může zvolit více možností. Ty jsou v databázi uvedeny jako otázky 21b, 21c, 21d, 21e a jejich hodnota je 0/1. My máme zjistit, zda odpověděl v pořádku na celou otázku 21, má-li jejich domov kuchyň. Teď už ale jsme zkušení a rovnou vytvoříme pomocnou proměnnou, která je součtem odpovědí na otázky 21b–e. Malinká potíž je v tom, že hodnota 0 znamená, že okénko je nezaškrtnuté, což může znamenat jak chybějící pozorování, tak fakt, že tuto střešní krytinu kuchyně nemá. Nakonec jsme to vyřešili tak, že má-li domácnost kuchyň a součet odpovědí na otázky 21b–e je nulový, považujeme toto pozorování za chybějící. Jinak je buď v pořádku vyplněné (domácnost má kuchyň a součet je nenulový), nebo správně vynechané (domácnost nemá kuchyň a součet je nulový).

## 3. Zpracování a závěr

Samotné zpracování dotazníků je pak už celkem snadné. Dostaneme dva stínové datové rámce. Odstraníme pomocné proměnné. Stínová data si schováme (v budoucnu pro ně najdeme uplatnění). Podíváme se, zda tam, kde odpověď má být, je, a tam, kde nemá být, není, a jak to vypadá u otázek, kde nevíme, zda odpověď být má, nebo nemá. Tím dostaneme jednoznačnou hodnotu (1 až 8), která určuje zda hodnota (ne)uvedená v dotazníku je (ne)uvedená v souladu s ostatními odpovědmi. Pak už jen vytvoříme tabulky, ve kterých shrneme jednotlivé počty hodnot 1 až 8 u jednotlivých otázek. Jen pro úplnost uveďme, že textové výstupy z R jsme zpracovali programem gawk do TEXového vstupu a výsledné tabulky máme ve formátu pdf.

Je nutné říci, že tazatelé v jednotlivých zemích odvedli dobrou práci. I v sociálně slabých oblastech dokázali dosáhnout vysoké úspěšnosti vyplnění dotazníků bez většího počtu chybějících či přebývajících pozorování. V Thajsku se jedná o desetiny procent, v ostatních zemích u výjimečných otázek o několik málo procent pozorování. Vzhledem k tomu, že jde o přibližně 300, občas docela intimních, otázek, nezbývá než vyslovit uznání celé organizaci zajišťující sběr dat.

**Poděkování autorů:** Příspěvek vznikl s podporou výzkumného záměru MSM 0021620839. Tento příspěvek byl přednesen v rámci Mikulášského statistického dne. Autoři děkují organizátorům této zdařilé akce.

# VELKOVÝROBA TABULEK POMOCÍ AWK

**Zdeněk Hlávka**

## 1. Úvod

Zkratka AWK znamená Aho, Weinberger, Kernighan a označuje poněkud svérázný programovací jazyk navržený pro zpracovávání textových souborů.

První verze tohoto programovacího jazyka vznikla již v roce 1977. V roce 1985 byl tento programovací jazyk dále rozšířen [1, 2] a tato nová implementace byla vydána pod volnou licencí v roce 1996. Další informace lze nalézt na stránce Briana Kernighana: `http://cm.bell-labs.com/cm/cs/awkbook/`.

GNU awk (nebo gawk) je jiná implementace jazyka AWK, která je (téměř jistě) obsažená v každé distribuci Linuxu. Další distribuce existují pod názvy xgawk, mawk nebo TAWK (verze pro DOS a Windows).

## 2. Příklad

Princip jazyka AWK si ukážeme na jednoduchém příkladě. Předpokládejme, že výsledky rozsáhlé simulační studie jsou uloženy v následujícím značně nepřehledném textovém souboru `priklad.txt`:

```
TABULKA.AA "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L"
"1" 1 2.0599209764691 1.1747333245979 1.03814385890582 2.0599209746917
1.10206315043 27 1.00304826425488 1.16483656626828 1.04785974714491
1.0215094731292 1.1656626828 1.03386899720460 1.00215433687505
"2" 1 1.2111642320962 1.1785509411592 1.09217031887314 1.21116423209622
1.168397750859 1.0225654118313 1.1664016471655 1.05402221156164
1.00649949971005 1.1664064716555 1.02731120933558 1.00184132881134
"3" 1 1.3220358323917 1.0628344035499 1.01922692920646 1.32203958323917
1.052451742835 1.00663451863578 1.0622641328610 1.0229410009790
1.01245982055420 1.0622661328608 1.02234509043763 1.00537251503339
"4" 1 3.0406725862321 1.12150481270856 1.0758055731613 3.04367252319
1.1156698237625 58 1.00211380906327 1.2487169003298 1.102381764515
1.01161646944822 1.24871679003298 1.10077794969210 1.00206064501333
```

Programovací jazyk AWK umožňuje tento textový výstup jednoduše přeformátovat do podoby LaTeXové tabulky. Následující postup funguje se standardní Linuxovou implementací `gawk`.

Nejprve vytvoříme soubor `tabulka.awk`, který obsahuje všechny potřebné instrukce:

```
BEGIN{
    getline;
    print "\\begin{tabular*}%";
    print "{\\textwidth}{l|*{",NF-1,"}{@{\\extracolsep\\fill}r}}";
    row=1
}
```

```
{
    printf "%.0f",row;
    for (i=2;i<=NF/2;i++) {
        printf "&%.1f",$(i+NF/2);
        printf "&{\\scriptsize (%.3f)}",$i;
    }
    print "\\\\";
    row=row+1
}

END{
    print "\\end{tabular*}";
}
```

Úvodní část programu, za příkazem `BEGIN`, obsahuje příkazy, které se provedou pouze na počátku. Zde vytiskneme hlavičku naší tabulky. Proměnná `NF` zde vytiskne počet záznamů na prvním řádku vstupního textového souboru (musíme odečíst 1, protože výsledná tabulka bude mít o jeden sloupeček méně). Vytvoříme proměnnou `row`, která bude obsahovat číslo řádku. Příkaz `getline` „přeskočí“ nezajímavý první řádek.

Hlavní část programu nalezneme v dalších složených závorkách. Tyto příkazy se vykonají postupně pro každý řádek vstupního textového souboru. Nejdříve vytiskneme číslo řádku a pak postupně tiskneme čísla z druhé poloviny textového souboru (velkým písmem a s přesností na jedno desetinné místo) a za ně vždy příslušná čísla z první poloviny textového souboru (menším písmem, v závorce a s přesností na tři desetinná místa). Vše doplňujeme potřebnými oddělovači a na konec každého řádku nezapomeneme vytisknout dvě obrácená lomítka.

Ve třetí části programu, za příkazem `END`, již vytiskneme `\end{tabular*}` a ukončíme tím tabulku.

Z příkazové řádky spustíme náš program:

```
gawk -f tabulka.awk < priklad.txt > priklad.tex
```

a získáme pěkně zformátovanou LaTeXovou tabulku v souboru `priklad.tex`, který obratem vložíme do našeho příspěvku pomocí `\input{priklad.tex}` příkazu:

```
\begin{table}
\input{priklad.tex}
\caption{Pěkně zformátovaná tabulka.}
\end{table}
```

Výsledek našeho snažení se vzápětí objeví v pěkně zformátované Tabulce 1.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.2 | (1.000) | 1.0 | (2.060) | 1.0 | (1.175) | 1.2 | (1.038) | 1.0 | (2.060) | 1.0 | (1.130) |
| 2 | 1.2 | (1.000) | 1.1 | (1.211) | 1.0 | (1.179) | 1.2 | (1.092) | 1.0 | (1.211) | 1.0 | (1.162) |
| 3 | 1.1 | (1.000) | 1.0 | (1.322) | 1.0 | (1.063) | 1.1 | (1.019) | 1.0 | (1.322) | 1.0 | (1.058) |
| 4 | 1.2 | (1.000) | 1.1 | (3.040) | 1.0 | (1.122) | 1.2 | (1.018) | 1.1 | (3.040) | 1.0 | (1.110) |

Tabulka 1: Pěkně zformátovaná tabulka.

Hlavní výhoda tohoto postupu je fakt, že pro změnu ve formátování Tabulky 1 stačí tuto změnu provést pouze na jednom místě v `tabulka.awk`. Naprosto triviální je například změna počtu desetinných míst pro všechna čísla v Tabulce 1.

Celý postup lze samozřejmě ještě více automatizovat a zjednodušit. Například, program `gawk` můžeme zavolat přímo z našeho statistického software už při ukládání výsledků. Pokud jsme nedočkaví, tak můžeme tabulku zároveň i přeložit LaTeXem do PDF nebo PS. Lze si představit i jiné použití v kombinaci s příkazem `Sweave()` ve statistickém programovacím prostředí `R` dostupném na: `http://www.r-project.org/`.

## 3. Závěr

Je zřejmé, že AWK můžeme využít i k jiným, náročnějším úkolům. Díky své rychlosti se hodí k provádění jednoduchých operací na rozsáhlých datových souborech, k automatizaci psaní opakujících se počítačových programů, nebo k získávání potřebných informací z jasně zformátovaných rozsáhlých LaTeXových souborů. Jako příklad si lze představit automatické odesílání emailů autorům příspěvků v tomto sborníku, vždy s automaticky vyplněnou správnou emailovou adresou, automaticky vyplněným správným jménem autora a automaticky vyplněným správným názvem příspěvku.

Fantazii se zkrátka žádné meze nekladou.

## Literatura


[1] Alfred V. Aho, Brian W. Kernighan, and Peter J. Weinberger. *The AWK programming language.* Addison-Wesley Publishing Company, 1988.

[2] D. Dougherty and A. Robbins. *sed & awk, Second Edition.* O'Reilly Media, 1997.

# ZE SVĚTA TEXu...

## Poznámky od redakce

Novinky lze shrnout do těchto oblastí:

**Čeština v IL2 i přes Babel.** Po instalaci TEX distribuce a spuštění skriptu cshbabel z `cstug.cz` můžete řádit s nádhernou (Olšákovskou a spol.) češtinou, slovenštinou přes Babel balíček.

**Latin Modern fonty.** Open Source fonty (lmodern.sty) začleněna v mezinárodních TEX distribucích. Je uvažováno o rozšíření matematických sad. `http://www.gust.org.pl/projects/e-foundry/latin-modern`

**TEX GYRE fonty.** Open Source fonty reimplementující základních osm PostScriptových písem. K dispozici v testovacích verzích. `http://www.gust.org.pl/projects/e-foundry/tex-gyre`

**METAPOST knihovna.** Opravují se chyby z dřívějška, implementují se nové partie až po potenciální 3D podporu. `http://tug.org/metapost`

**XeTEX.** [zítech] Rozšíření TEXu s potenciálem práce s PostScript, TrueType i OpenType fonty. `http://scripts.sil.org/XeTeX`

**LuaTEX** Tento pojem [luahtech] si dobře uložte do paměti. Tudy jde TEX budoucnost. Je to založeno nad pdfTEXem a následuje aktivní rozvoj LuaTEXu. Shrneme-li poznámky od tvůrců, tak zběžný výčet vlastností:

- Skriptovací objektový jazyk Lua vnořený do TEXu.
- Zpřístupnění datových struktur TEXu.
- Přímé načítání OpenType fontů.
- Dynamická alokace paměti.
- Řízení zlomu odstavce a stránky – v plánu.
- Vnoření METAPOST knihovny – v plánu 2008.
- Vytvoření rozhraní pro pluginy – v plánu 2010.

`http://www.luatex.org/` nebo jsou soubory pro Windows na `fsci.fuk.kindai.ac.jp/kakuto/win32-ptex/web2c75-e.html`

**Nejbližší TEX konference:**

- TUG 2008: 21. 7. – 24. 7. 2008, University College Cork, Irsko.
- ConTEXt User Meeting: 20. 8. – 25. 8. 2008, Bohini, Slovinsko.
- 1. ročník TEXperience 2008: 26. 9. – 28. 9. 2008, Rusava, Česká republika. Více informací a přihláška pro zájemce bude na `cstug.cz`.

## OBSAH BULLETINU



**ISSN 1210 – 8022. Informační Bulletin** České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo.

**Předseda společnosti:** Doc. RNDr. Gejza Dohnal, CSc., ÚTM FS ČVUT v Praze, Karlovo náměstí 13, 121 35 Praha 2, e-mail: gejza.dohnal@fs.cvut.cz

**Ediční rada:** Prof. Ing. Václav Čermák, DrSc. (předseda), Prof. RNDr. Jaromír Antoch, CSc., Doc. Ing. Josef Tvrdík, CSc., RNDr. Marek Malý, CSc., Doc. RNDr. Jiří Michálek, CSc., Doc. RNDr. Zdeněk Karpíšek, CSc. a Prof. Ing. Jiří Militký, CSc.

**Techničtí redaktoři:** Doc. RNDr. Gejza Dohnal, CSc., gejza.dohnal@fs.cvut.cz a Ing. Pavel Stříž, Ph.D., striz@fame.utb.cz

**Pokyny autorům:** <http://www.statspol.cz/bulletiny/sablony.htm>

**FTP:** exp.uis.fame.utb.cz; uživatel: csts; heslo: csts

**WEB server:** <http://www.statspol.cz/>