PUBLIKOVÁNÍ PROSTOROVÝCH DAT A VÝSLEDKŮ JEJICH ANALÝZ POMOCÍ INTERNETOVÝCH MAPOVÝCH SERVERŮ

Jitka Komárková, Miloslav Hub, Renáta Máchová, Martin Ulrich
Ústav systémového inženýrství a informatiky, FES, Univerzita Pardubice

***Abstrakt**: Běžnou součástí každodenního života pracovníků mnoha organizací i občanů se stalo využívání prostorových dat a výsledků jejich analýz. Data jsou interaktivně zpřístupňována pomocí internetových geografických informačních systémů. Po počátečním bouřlivém rozvoji jsou dnes i na tento typ informačních systémů kladeny obvyklé požadavky související s kvalitou poskytovaných služeb. Mezi ně patří i rychlost odezvy a kvalita poskytovaných dat. V obou případech může být kvalita poskytovaných služeb výrazně ovlivněna komprimací prostorových dat. Byly proto studovány možnosti komprimace prostorových dat a rychlost odezvy mapových serverů v závislosti na velikosti zpracovávaných dat.*

***Abstract**: Utilization of geographic information has become an inherent part of everyday life. Data are interactively provided by means of Internet Geographic Information Systems (GIS). Nowadays, the same quality characteristics are applied to Internet GIS as to the other information systems. Server response time and a quality of provided data belong to them. Possibility of spatial data compression and server response times of selected Internet GIS were studied.*

***Klíčová slova:** prostorová data, komprimace, internetový GIS, internetové mapové servery, kvalita služeb*

***Key words**: spatial data, compression, Internet GIS, Internet map server, quality of service*

1. Úvod

Prostorová data – pro mnohé uživatele v podstatě informace o území, sice patří k významnému druhu informací, bez kterých se dnes málokdo obejde, ale plnohodnotná práce s nimi je neustále vyhrazena především specializovaným odborníkům. Rovný přístup všech k informacím bez diskriminace určitých skupin obyvatelstva (např. věkových) však patří k cílům informační společnosti budované jak u nás, tak v zemích Evropské unie (viz např. Státní informační politika ČR [1], iniciativa eEurope [2] a jejich novelizace). S rostoucí počítačovou gramotností přitom roste i poptávka po těchto informacích. Zároveň se díky grafické podobě jedná o i informace s vysokou vypovídací schopností.

Bouřlivý rozvoj proto zaznamenávají technologie, jejichž cílem je zpřístupňovat prostorové informace laickým koncovým uživatelům – dnes obvykle nazývané internetové geografické informační systémy (GIS). Jejich účelem je interaktivně zpřístupňovat prostorová data a výsledky jejich analýz v intuitivně ovladatelném prostředí v prostředí Internetu či intranetu. Zároveň mohou být nabízeny také tzv. geo-služby. Nejčastějším klientem je WWW prohlížeč, který je z hlediska laických koncových uživatelů nejpřívětivější, ale lze používat i specializované WWW prohlížečky či těžké klienty [3, 4].

Nejvyužívanější architekturou internetových GIS systémů je architektura klient/server, v třívrstvé, resp. vícevrstvé, podobě. Aplikační GIS logika je představována mapovým serverem, jehož úkolem je zpracovat požadavek klienta a vrátit mu zpět odpověď – nejčastěji ve formě obrázků znázorňujících výsledek. Dále jsou samostatnými vrstvami systému WWW server, datový server a klient [5].

Významným problémem, se kterým se dnes potýkají všechny geografické informační systémy, nejen internetové, je otázka zpracování a zpřístupňování velkých objemů dat, převážně rastrového charakteru, která pocházejí především z dálkového průzkumu Země. Existence standardů zajišťujících interoperabilitu jednotlivých systémů a široce respektovaných odbornou komunitou je významnou skutečností, která usnadňuje využívání principů paralelních a distribuovaných systémů [6].

Internetová GIS řešení jsou dnes chápána jako specifický druh informačního systému, který by měl poskytovat dostatečně kvalitní služby, mezi něž patří rychlost odezvy a kvalitní informace s dostatečnou vypovídací hodnotou.

2. Rychlost odezvy mapového serveru

Jeden z faktorů významně ovlivňujících kvalitu služeb poskytovaných internetovými mapovými servery. Pomalá odezva je negativně vnímána uživateli informačního systému.

2.1 Formulace problému

Měření rychlost odezvy serveru je potřeba provádět s vyšší přesností, než s jakou lze běžně získat potřebné informace z logovacích souborů jednotlivých operačních systémů, firewallů, atd. Je potřeba dosáhnout přesnosti na tisíciny sekund, zatímco běžně jsou události probíhající v daném operačním systému zaznamenávány s přesností na celé sekundy. Bylo proto potřeba najít vhodný způsob, jak vůbec měření realizovat.

2.2 Návrh řešení

V literatuře nebyl nalezen žádný doporučený postup pro měření rychlosti odezvy internetového mapového serveru. Proto byl navržen a vytvořen podpůrný nástroj pro měření rychlosti odezvy serveru s přesností na tisíciny sekund. Nástroj byl vytvořen v prostředí operačního systému Linux, konkrétně v prostředí Debian GNU/Linux 3.0r1 s jádrem 2.4.19 a následně byl použit v prostředí RedHat 7.3 s jádrem 2.4.18-3.

Podstatou nástroje je úprava syslogd daemona operačního systému tak, aby zapisoval čas s požadovanou přesností. Program syslogd daemon byl proto rozšířen o následující dva fragmenty v části funkce „void vsyslog“:

```
struct timeval cas;
struct timezone zona;
struct tm *p_tm;


zona.tz_minuteswest=0;
zona.tz_dsttime=0;

void)gettimeofday(&cas, &zona);
p_tm=localtime(&cas.tv_sec);
void)sprintf(tbuf,"$%#02d:%#02d:%#02d.%#02ld ", p_tm->tm_hour, p_tm->tm_min,
              p_tm->tm_sec, (cas.tv_usec)/1000);
```

Volání jádra gettimeofday() zaznamenává do svého prvního parametru čas s přesností na mikrosekundy. Funkce localtime slouží pro převod času do formátu, který je bližší lidem (hh:mm:ss.ttt). Funkce sprintf zapíše čas do parametru tbuf v určeném formátu ($hh:mm:ss.tt).

Rozšíření bylo provedeno v souboru syslog.c, následně musel být daemon nově přeložen a nainstalován. Celý změněný výsek souboru syslog.c je uveden v příloze č. 1.

Čas příchodu požadavku na server a odchodu odpovědi ze serveru byl zaznamenáván s využitím další služby jádra systému - filtrování paketů. Příslušná pravidla se nastavují pomocí iptables.

Pro zachytávání spojení na port 80 (http) sloužila tato pravidla:

```
iptables -A INPUT --proto tcp --destination-port 80 --syn -j LOG --log-level info

iptables -A OUTPUT --proto tcp --source-port 80 --tcp-flags FIN FIN -j LOG --log-level info

iptables -A INPUT --proto tcp --destination-port 13 --syn -j LOG --log-prefix="ZNACKA " --log-level info
```

Pravidla byla vytvořena tak, aby byl zaznamenán paket:

- s příznakem SYN (tento paket začíná spojení), který přichází port na 80,
- s nastaveným příznakem FIN (ukončení spojení), který odchází z portu 80,
- přicházející na port 13 (daytime), který sloužil jako značka splnění požadavku.

Instalace nástroje vyžaduje administrátorská práva.

V prostředí operačních systémů typu Linux pak probíhala další měření. Z důvodu vyloučení vlivu zátěže počítačové sítě a klientské stanice byla zvolena jako první varianta měření rychlosti odezvy na serveru. Měření byla prováděna na dedikovaném serveru.

2.3 Naměřená data a jejich statistické zpracování

Měření bylo prováděno na dedikovaném serveru, který byl nainstalován v prostředí VMware jako virtuální server.

Konfigurace serveru - „hostitele“: procesor CPU AMD ATHLON XP 2000+, základní deska Microstar MS-K7T266 Pro 2 RU, Socket A, VIA266A, pevný disk Wester Digital 1200JB 120 GB UATA/100 7200 RPM - 8MB buffer, síťová karta 3COM FAST ETHERLINK XL PCI-TX 10/100, paměť 1024 MB: DIMM DDR 32M X 64,512MB,3.3V,266MHZ KINGMAX CL2. Operační systém: Linux RedHat 7.3 s jádrem 2.4.19.

Konfigurace testovacího serveru: stejná jako u fyzického serveru s následujícími rozdíly: paměť 256 MB, pevný disk 4 GB. Operační systém: Linux RedHat 7.3 s jádrem 2.4.18-3.

Každé měření bylo provedeno celkem 12x, z čehož 3x byla použita funkce posun, 3x funkce zvětšit měřítko a 3x funkce zmenšit měřítko.

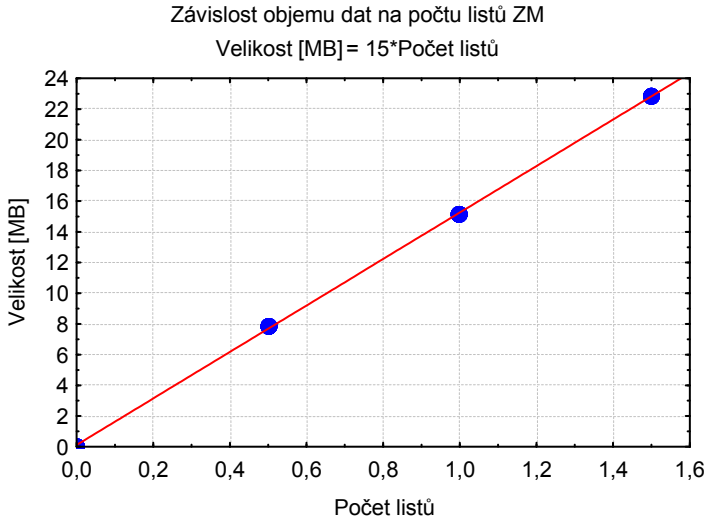
V rámci statistického zpracování naměřených dat bylo úkolem stanovit závislost doby zpracování dat (závislé proměnné) na objemu a způsobu zobrazení těchto dat a typu dat (nezávislých proměnných).

Doba zpracování dat byla měřena v sekundách s přesností na mikrosekundy. Objem zpracovaných dat byl měřen zároveň v informatickém smyslu objemu dat, tedy v MB (megabajtech) a zároveň v počtu zobrazených stránek. Způsob zobrazení geografických dat může být dlaždicově nebo bez použití dlaždic. Zpracovávaná data mohou být dvou typů, a to rastrová nebo vektorová.

Jak však bylo předpokládáno, objem zpracovávaných dat v MB je závislý na počtu zobrazených stránek a to dokonce čistě lineárně (přímá úměra), neboť výběrový koeficient korelace mezi objemem dat v MB a v počtu stránek daný vztahem (1), definovaným takto [7]:

$$r_{x,y} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{x}^2} * \sqrt{\sum_{i=1}^{n} Y_i^2 - n\bar{y}^2}} \qquad (1)$$

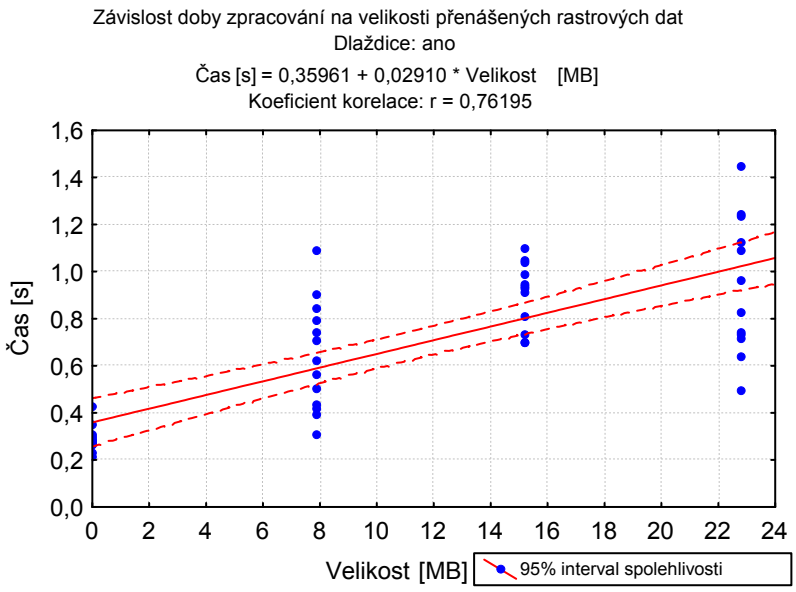
je roven 1, a to jak v případě dlaždicového zobrazení, tak i v zobrazení bez dlaždic. Tato přímá lineární závislost je znázorněna v následujícím grafu:



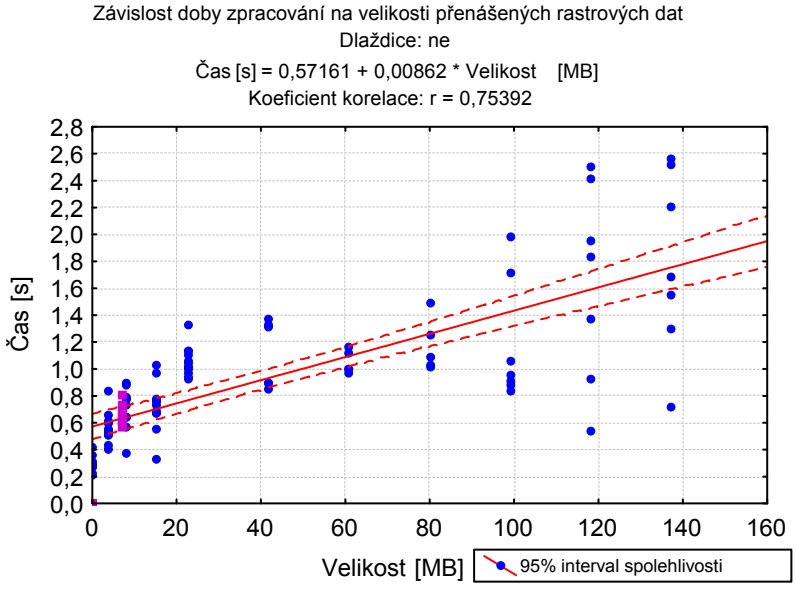
Obr. č. 1: Závislost objemu rastrových dat na počtu listů základní mapy odvozené 1:5000

Z těchto důvodů není třeba velikost objemu dat v zobrazených stránkách dále uvažovat a za nezávislé proměnné budou považovány pouze velikost dat v MB, typ dat a způsob zobrazení (který nebude třeba také uvažovat, jak bude prokázáno později).

V rámci určení závislosti doby zpracování rastrových dat na objemu těchto dat byly měřeny tyto doby několikrát pro různé stránky a to vždy zvlášť pro dlaždicové zobrazení a pro zobrazení bez dlaždic. Výsledky měření jsou znázorněny v následujících grafech a jsou proloženy přímkou prostřednictvím metody nejmenších čtverců. Zároveň je znázorněn i 95% interval spolehlivosti pro odhad parametrů jednoduchého lineárního regresního modelu, ve kterém jsou příslušné parametry odhadovány metodou nejmenších čtverců.



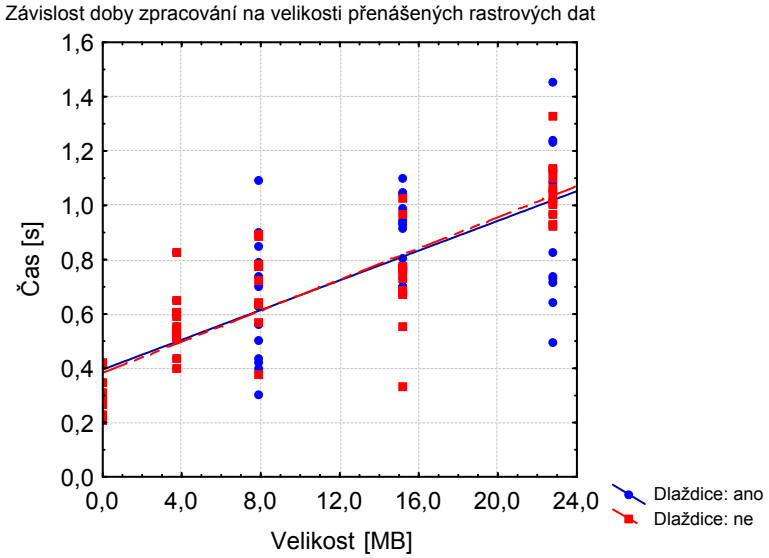
Obr.č. 2: Závislost doby zpracování na velikosti rastrových dat uspořádaných do dlaždic - proklad přímkou



Obr.č. 3: Závislost doby zpracování na velikosti rastrových dat neuspořádaných do dlaždic - proklad přímkou

Z obou grafů (a zároveň i z matematického popisu přímek) je patrné, že odhad parametru a přímky f(x) = a + b.x je kladný. Tedy i v případě, že výstupem nejsou žádná data, zpracování požadavku zabere nějaký čas.

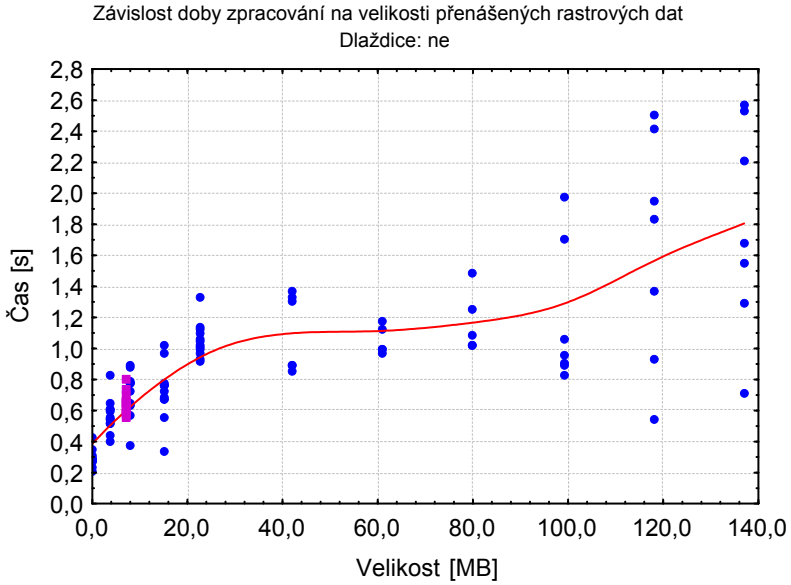
Porovnáme-li však regresní model pro naměřená rastrová data do velikosti 24 MB pro dlaždicové zobrazení a zobrazení bez dlaždic, zjistíme, že obě přímky jsou téměř totožné. Ke stejnému závěru lze dojít i po provedení znaménkového testu [8], při kterém je testována hypotéza, že doba zpracování při dlaždicovém zobrazení je shodná s dobou při zobrazení bez dlaždic. Tato hypotéza není v rámci výběrového souboru vyvrácena a tudíž není důvod uvažovat, že je doba zpracování dat při dlaždicovém a nedlaždicovém zpracování odlišná.



Obr. č. 4: Závislost doby zpracování na velikosti rastrových dat s využitím uspořádání do dlaždic a bez uspořádání do dlaždic

Výběrový koeficient korelace mezi objemem zpracování dat v MB a doby zpracování je však přibližně 0,75. Po provedení testu významnosti pro koeficient korelace, kdy si jako nulovou hypotézu stanovíme, že koeficient korelace je roven 0 dojdeme k závěru, že tuto hypotézu musíme zamítnout, a že tedy tyto veličiny jsou korelované. Tedy že mezi nimi existuje statisticky významná lineární závislost.

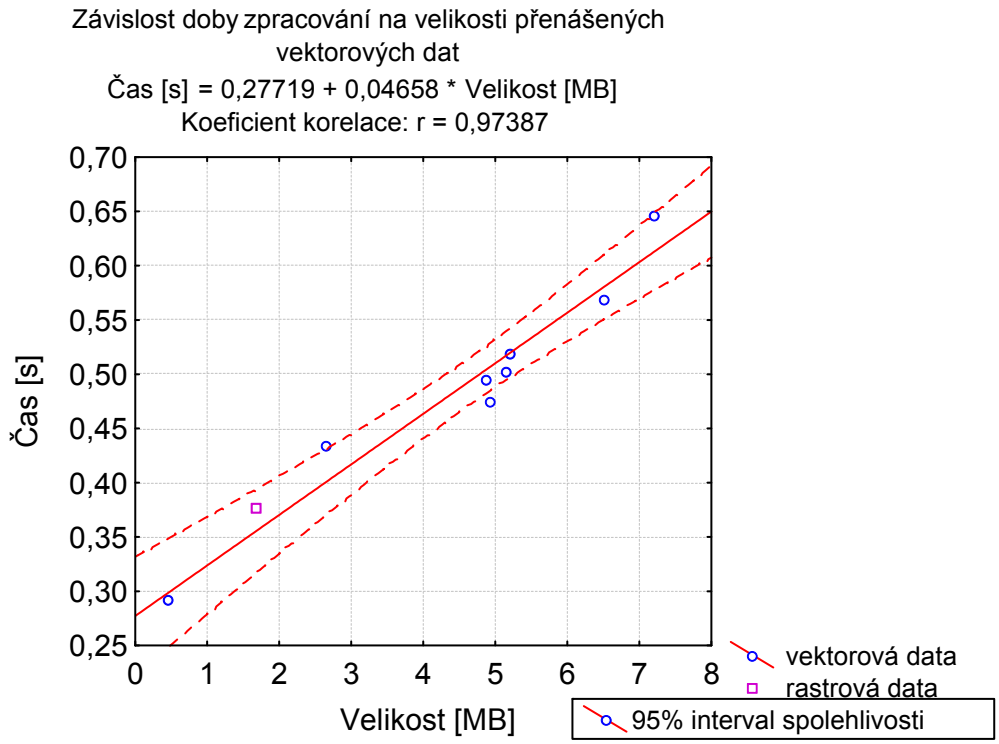
V tomto okamžiku však nezbytně vystává otázka, zda je jednoduchý model lineární regrese pro modelování této závislosti tím nejvhodnějším. Proložíme-li naměřená data křivkou prostřednictvím metody vážené vzdálenosti nejmenších čtverců [9], získáme následující průběh:



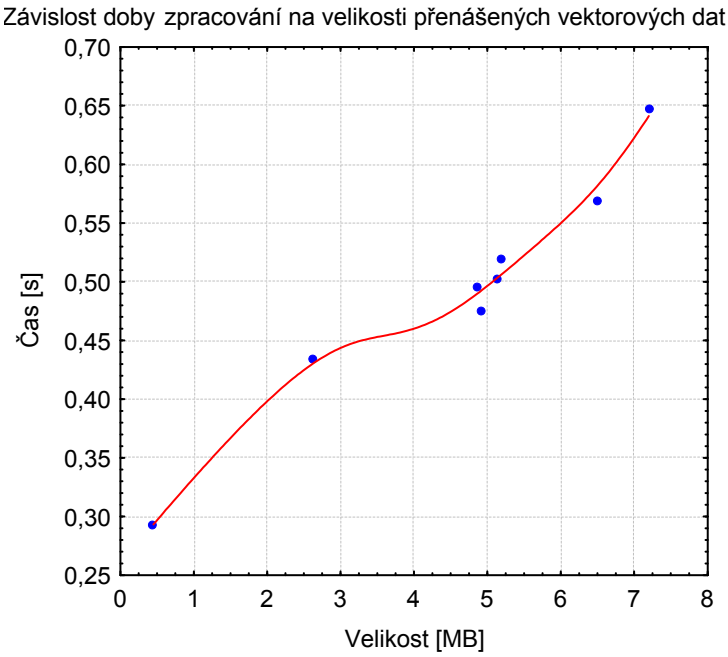
Obr. č. 5: Závislost doby zpracování na velikosti rastrových dat – proklad křivkou (metoda vážené vzdálenosti nejmenších čtverců)

Průběh této funkce je nejprve konkávní a plynule přechází v konvexní křivku. Inflexní bod se nachází přibližně v hodnotě 55 MB.

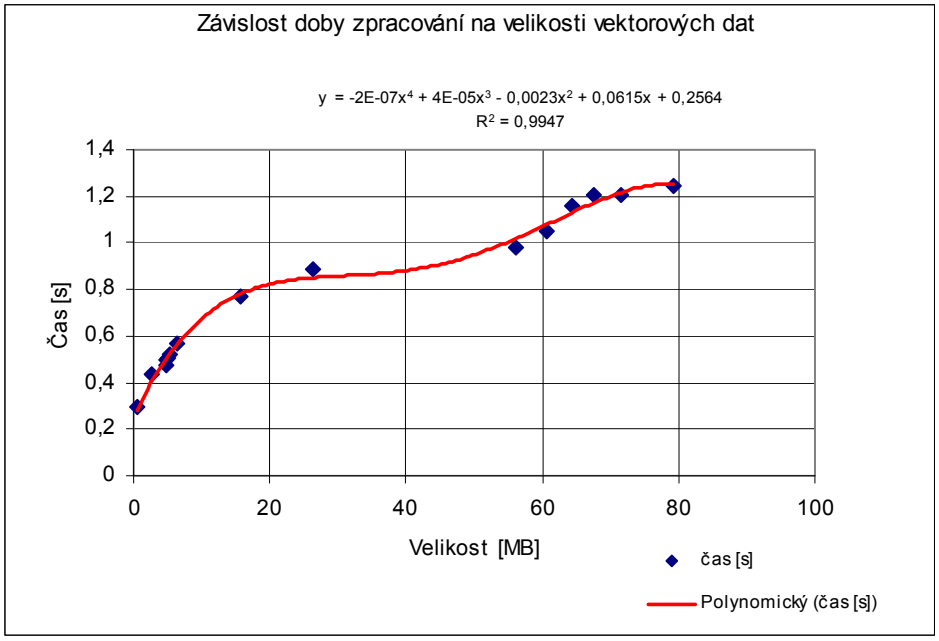
Podobné průběhy bychom získali, i kdybychom narozdíl od rastrových dat analyzovali data vektorová (jejichž objem však bývá zpravidla menší, než v případě rastrových dat).



Obr. č. 6: Závislost doby zpracování na velikosti vektorových dat – proklad přímkou a odhad 95% intervalu spolehlivosti



Obr. č. 7: Závislost doby zpracování na velikosti vektorových dat – proklad křivkou (metoda vážené vzdálenosti nejmenších čtverců)



Obr. č. 8: Závislost doby zpracování na velikosti vektorových dat – proklad křivkou (polynom 4. stupně)

3. Komprimace prostorových dat

Metody komprimace prostorových dat umožňují významně zmenšit objem přenášených dat na jedné straně, na druhé straně mohou snížit kvalitu výstupu tak, že výstup není využitelný.

3.1 Metody využívané v praxi

Vzhledem ke stále vzrůstajícím objemům se dnes již zcela běžně provádí komprimace rastrových prostorových dat. Za nejvhodnější metody komprimace pro prostorová rastrová data lze vzhledem k dnešním objemům rastrových dat považovat metody MrSID (Multi-resolution Seamless Image Database) a ECW (Enhanced Compressed Wavele), které například odstraňují limity velikosti jednotlivých bloků ve formátech TIFF a JPEG. Jde o ztrátové vlnkové komprese, které poskytují zhruba stejné výsledky, přičemž nejlepší výsledky dávají při použití ještě nezkomprimovaných obrazů, tj. použít např. již jednou zkomprimovaný JPG soubor. Vzhledem ke ztrátovosti technologií je obecně lepší komprimovat původní data a komprimovaná data použít pouze v případě, kdy není potřebná kvalita výchozích dat. Mezi přednosti vlnkových metod komprimace dat patří možnost dekomprimovat pouze vybranou část snímku či odstranit komprimaci po blocích závislou na paměti serveru. Podporují i tvorbu mozaik [10, 11].

Při práci s prostorovými daty se naopak vůbec neprovádí komprimace vektorových dat. Bylo proto prověřováno, zda je vůbec možné prostorová vektorová data jako taková komprimovat. Testována byla vlnková komprimace a komprimace Fourierovou metodou. Jedná se o ztrátové komprimace, které jsou uplatnitelné pouze v určitých situacích, např. při generalizaci, zmenšování měřítka zobrazené mapy, a pouze pro velké objemy dat s polygonovými nebo liniovými prvky s velkým počtem úseků [12].

3.2 Experimentální výsledky a závěry

Vlnkové komprimace jsou ztrátové metody, které mohou vést při nevhodném použití až ke znehodnocení dat. Maximální vhodný cílový komprimační poměr pro zpřístupňování

barevných rastrových dat na Internetu je 40:1. Skutečný komprimační poměr závisí na charakteru snímku – snímky obsahující velké plochy stejného druhu (např. velká pole) jsou komprimovány s větším poměrem, řádově cca 60:1, zatímco snímky s malými různorodými plochami (např. hustě osídlené území s malými různě využívanými pozemky) jsou ve skutečnosti komprimovány v menším poměru, řádově cca 20-30:1 Při cílovém komprimačním poměru 40:1 však již nelze data následně využít například pro tvorbu kvalitních tiskových výstupů. Pro takovéto úkoly jsou akceptovatelné komprimační poměry do 25:1. V případě černobílých dat jsou všechny poměry nižší (do 15:1 pro tisk a do 30:1 pro zpřístupnění na Internetu).

Následující Tabulka č. 1 dokumentuje výsledky získané při ECW komprimaci barevného leteckého snímku velikosti ¼ mapového listu ZMO5, Tabulka č. 2 dokumentuje ECW komprimaci barevného reliéfu ČR, kde je řada stejných ploch.

Tabulka č. 1 - komprimační poměry barveného leteckého snímku

cílový poměr	aktuální poměr	výstupní velikost	čas [s]	rychlost [MB/s]
5 :1	5,3 :1	3 MB	7	2,29
10 :1	9,2 :1	1,7 MB	6	2,38
15 :1	12,3 :1	1,3 MB	5	2,92
20 :1	16 :1	1007,4 KB	5	2,98
25 :1	19,2 :1	843,5 KB	5	2,97
30 :1	22,3 :1	723,6 KB	5	3,09
35 :1	25,3 :1	638,1 KB	5	3,05
40 :1	29 :1	557,5 KB	5	3,12
45 :1	31,8 :1	508,4 KB	5	2,85
50 :1	35,3 :1	458 KB	5	3,18
60 :1	41,6 :1	338,6 KB	5	3,23
70 :1	47 :1	343,8 KB	5	3,29
80 :1	52,3 :1	308,7 KB	5	2,23

Tabulka č. 2 – komprimační poměry barveného reliéfu ČR

cílový poměr	aktuální poměr	výstupní velikost	čas [s]	rychlost [MB/s]
10 :1	27,2 :1	666,4 KB	5	3,38
20 :1	43,6 :1	416,2 KB	5	3,58
30 :1	53,6 :1	338,4 KB	5	3,55
40 :1	61,8 :1	293,5 KB	5	3,61
50 :1	68,1 :1	266,6 KB	5	3,54
60 :1	74 :1	245,1 KB	5	3,22
70 :1	78,7 :1	230,6 KB	5	3,44
80 :1	83,3 :1	217,9 KB	5	3,58

Komprimace vektorových dat se jeví v případě GIS jako technika využitelná pouze ve vybraných situacích - především v případě, kdy dochází ke zmenšování měřítka a tím generalizaci zobrazovaných prvků, čímž ovšem dochází ke zbytečnému znehodnocení drahých dat. V praxi jsou přitom k dispozici značně levnější, již generalizované datové sady. Vzhledem k tomu, že všechny internetové mapové servery podporují pouze vybrané formáty vektorových dat, vyžaduje dopředu předpřipravené datové sady v jím podporovaných formátech, ale žádný z nich zatím komprimaci a dekomprimaci vektorových dat nepodporuje, je vhodnější využít levnější generalizované sady - i z hlediska poměru cena/výkon.

4. Závěr

Internetové geografické informační systémy prošly v uplynulých několika letech obdobím bouřlivého rozvoje. Dnes jsou považovány za součást informačních systémů a musí tedy poskytovat služby odpovídající kvality. Mimo jiné musí zpracovat s dostatečnou rychlostí požadavek klienta a poskytnout mu jím požadované informace v odpovídající kvalitě.

Jak bylo prokázáno v rámci této studie, rychlost odezvy serveru výrazně závisí kromě faktorů typu zatížení počítačové sítě či klienta také na velikosti zpracovávaných dat. Rychlost odezvy mohou zkrátit v případě velkoobjemových rastrových dat vlnkové komprimační technologie – MrSID a ECW. Na druhé straně mohou tyto technologie při nevhodném použití – zvolení příliš vysokého komprimačního poměru, snížit kvalitu poskytovaných dat.

6. Použitá literatura:

[1] *Vláda ČR: usnesení č. 525. Státní informační politika – cesta k informační společnosti*. 31. 5. 1999. [on-line]. [cit. 2006-01-05]. URL <http://www.vlada.cz>

[2] *The European Commission, eEurope - An Information Society For All*. For the Special European Council on Employment, Economic reforms and Social Cohesion - Towards a Europe based on Innovation and Knowledge. Lisbon, Portugal, 23.-24.3.2000, [on-line]. [cit. 2006-01-05]. URL <http://europa.eu.int>

[3] LONGLEY, P. at al.: *Geographic information systems and science*. John Wiley & Sons, Chichester, England, 2001

[4] LOWE J. W., The anatomy of Web map services. *Geospatial Solutions*, Vol. 12, No. 9, 2002, pp. 42-45

[5] PENG, Z.-R., TSOU, M.-H., *Internet GIS: distributed geographic information services for the internet and wireless networks*. John Wiley & Sons, 2003.

[6] HAWICK, A., K., CODDINGTON, P., D., JAMES H., A., Distributed frameworks and parallel algorithms for processing large-scale geographic data. *Parallel Computing*, Vol.29, 2003, pp. 1297–1333

[7] PEARSON, K. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London*, Ser. A, *187*, (1896), 253-318.

[8] DODD, B. Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, *11*, (1979), 478-484.

[9] NETER, J., WASSERMAN, W., KUTNER, M. H.. *Applied linear statistical models: Regression, analysis of variance, and experimental designs.* Homewood, IL: Irwin, 1985

[10] Image compression: ER Mapper 6.0 ECW v2.0 versus MrSID 1.3 [on-line]. [cit. 2006-01-05]. URL <http://www.ermapper.com/document/doc.aspx?doc_id=35>

[11] HALLAM, A. *Tips on ECW Image Compression - Digital Earth Pty Ltd, Australia*, [on-line]. [cit. 2006-01-05]. URL <http://www.digitalearth.com.au/pubs_ecw_tips.php>

[12] ČAPEK, J. Komprimace vektorových geodat. *Ekonomie a Management*, Roč. VII, č. 4 (2004), s. 120 – 128

Kontaktní adresy:

Ing. Jitka Komárková, Ph.D., Ing. Miloslav Hub, Ph.D., Ing. Renáta Máchová, Ph.D.,
Ing. Martin Ulrich
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní, Univerzita Pardubice
Studentská 84, 532 10 Pardubice
jitka.komarkova@upce.cz
miloslav.hub@upce.cz
renata.machova@upce.cz

Příloha č. 1 – změněná část souboru syslog.c

```
void
vsyslog(pri, fmt, ap)
                        int pri;
                        const char *fmt;
                        va_list ap;
{
                        register int cnt;
                        register char *p;

        //P R I D A N O
                        struct timeval cas;
                        struct timezone zona;
                        struct tm *p_tm;
        //K O N E C

                        //time_t now;
                        int fd, saved_errno;
                        char tbuf[2048], fmt_cpy[1024], *stdp = (char *) 0;

                        saved_errno = errno;

                        /* see if we should just throw out this message */
                        if (!LOG_MASK(LOG_PRI(pri)) || (pri &~ (LOG_PRIMASK|LOG_FACMASK)))
                             return;
                        if (LogFile < 0 || !connected)
                             openlog(LogTag, LogStat | LOG_NDELAY, LogFacility);

                        /* set default facility if none specified */
                        if ((pri & LOG_FACMASK) == 0)
                             pri |= LogFacility;

                        /* build the message */
                        //(void)time(&now);
                        //(void)sprintf(tbuf, "<%d>%.15s ", pri, ctime(&now) + 4);

        // P R I D A N O

                        zona.tz_minuteswest=0;
                        zona.tz_dsttime=0;

                        (void)gettimeofday(&cas, &zona);
                        p_tm=localtime(&cas.tv_sec);

                        (void)sprintf(tbuf,"$%#02d:%#02d:%#02d.%#02ld ", p_tm->tm_hour, p_tm-
>tm_min, p_tm->tm_sec, (cas.tv_usec)/1000);

        //K O N E C
```