JIŘÍ HORÁK, IGOR IVAN, MARKÉTA NÁVRATOVÁ, JIŘÍ ARDIELLI

VYHLEDÁVÁNÍ ČESKÝCH MĚST UŽIVATELI GOOGLE

HORÁK, J., IVAN, I., NÁVRATOVÁ, M., ARDIELLI, J. (2013): Searching for Czech towns by Google users, 118, No. 3, pp. 284–307. – Selected web search engines provide statistics regarding user activities according to the topic, time, and optionally, location of the search. The statistics provided by Google Insights for Search (Google Trends) for the names of Czech cities as parts of online queries within a six-year period were explored and analysed according to their frequency and associated topics. This data is calibrated using a system of etalons. The distribution of Czech city search results between resident and non-resident users is estimated using associated topics and the location of the origin of the query. The frequency of search for Brno and Olomouc provide highly above-average results. Most of the other regional centres show a slightly above-average frequency of search. Ostrava, České Budějovice and Ústí nad Labem are among the below-average searched cities. The paper introduces a new data source, recommends its appropriate processing, explains pros and cons, and comments on possible issues.
KEY WORDS: Google – web search engine – towns of Czechia.

Příspěvek byl zpracován v rámci grantového projektu GA 403/09/1720 „Industriální město v postindustriální společnosti“.

Úvod

Internet je v dnešní době významným informačním prostředím a přes všechny výhrady k relevanci některých informací není sporu o tom, že postupně vytlačuje v řadě oblastí jiná média a informační zdroje. Uživatelů internetu rychle přibývá. V roce 2005 používalo internet 32 % obyvatel Česka ve věku nad 15 let; v roce 2010 to bylo již 62 % populace, čemuž odpovídá 5,5 milionu uživatelů (ČSÚ 2010). Dle reportu společnosti NetMonitor (NetMonitor 2011) čítá internetová populace v Česku více jak 5,8 miliónů uživatelů (zpráva za červen 2011). Spolu s vyšším využíváním internetu roste také používání vyhledávačů pro získávání požadovaných informací. Již v roce 1997 používalo vyhledávače pro lokalizaci online informací nebo služeb více než 80 % uživatelů, kteří hledali informace na webu (Nielsen Media 1997), dnes bude tento podíl ještě vyšší. Méně si již uživatelé uvědomují, že tato činnost je monitorována. Poskytovatelé vyhledávacích služeb evidují četnost dotazování jednotlivých frází a další důležité informace jako čas, asociovaná témata či oblast, odkud se uživatelé dotazují. Takové informace lze vhodně využít v marketingu, mohou být ale zajímavé i pro geografy či sociology. Je to jeden z příkladů nového směru poznání reality, který přináší geoinformatika (Voženílek 2002). V roce 2008 zpřístupnil Google statistiky vyhledávání prostřednictvím webové aplikace *Google Insights for Search* (dále GI). V říjnu 2012 se *Google Insights* spojil s *Google*

Trends, nicméně všechny funkce zůstaly stejné. V článku se používá původní označení, při kterém bylo prováděno vyhledávání. Aplikace umožňuje posoudit četnost vyhledávání jistého dotazu, především vývoj vyhledávání v čase, místa častého vyhledávání a slovní spojení zadávaná s dotazem, a to počínaje rokem 2004. Tyto statistiky přináší nové možnosti pro zkoumání chování a zájmů uživatelů a do jisté míry vypovídají o využití internetu, požadovaných informacích a subjektivní charakteristice cílových objektů. Získané statistiky odráží reálné zájmy lidí a trendy ve společnosti, které bychom těžko získali z jiných pramenů. S přihlédnutím ke stále zvyšujícímu se počtu uživatelů internetu se jedná o cenný zdroj informací do budoucna, komplementární k jiným zdrojům informací. Toto se také projevilo v narůstajícím počtu publikací, které pracují s tímto datovým zdrojem. Velmi časté je využití takto získaných dat ve zdravotnictví, kde např. Breyer a kol. (2011) prokázali přítomnost korelace geografické a časové distribuce vyhledávání informací o ledvinových kamenech a faktickém známém výskytu ledvinových kamenů v populaci. Dále Brownstein, Clark, Madoff (2009) analyzují četnost vyhledávání názvu specifické bakterie salmonely a skutečného počtu případů onemocnění touto bakterií v USA a opět prokazují existující silnou vzájemnou korelaci. Ibuka a kol. (2010) využívají GI jako jeden z datových zdrojů při analýze dynamiky vnímání rizika vzhledem k hrozící pandemii viru H1N1 v roce 2009 a prokazují rychlý nárůst vyhledávání klíčových slov (prasečí chřipka, chřipka, symptomy chřipky apod.) napříč USA a následný rychlý návrat na původní nízké hodnoty pouze během 14 dní. Podobnou problematikou se zabývají například také Ginsberg a kol. (2009), kdy se autoři snaží díky analýze dotazů v GI odhadovat týdenní aktivitu viru chřipky, což umožňuje detekovat vznik epidemií v oblastech s velkou populací připojenou na internet a využívající Google jako svůj vyhledávač. Příhodné a rovněž časté je využívání těchto dat v oblasti marketingu, kdy může být jedním z důležitých zdrojů pro nejrůznější obchodní strategie (např. Clipp 2011). Webb (2009) pak prokázal korelaci mezi vyhledáváním výrazu domácí exekuce a skutečným počtem domácích exekucí v USA. Uplatnění v oblasti ekonomie představují Preis, Reith, Stanley (2010), kteří sledují v sedmileté časové řadě vazby objemu vyhledávaných dotazů a fluktuací na finančních trzích v týdenních řezech. Zjistili silnou korelaci mezi finančními transakcemi nejvýznamnějších 500 firem (dle Standard & Poor's) a objemu vyhledávání daných názvů firem. Navíc z hlediska analýzy těchto časových řad jsou patrné opakující se tendence v jejich vývoji. Rovněž v geografii či sociologii nachází tato data svoje uplatnění, kde např. Scheitle (2011) ve svém článku analyzuje možnost využití dat z GI pro studium nejrůznějších sociálních oblastí v rámci geografických jednotek a porovnává tato data s daty z jiných zdrojů. Dochází k závěru, že data získaná z GI úzce korespondují s vybranými existujícími daty (nezaměstnanost, imigrace, terorismus apod.) a vzhledem k jejich dostupnosti doporučuje využívání těchto dat i pro vědecké účely. GI se aktuálně využívá rovněž v celé řadě jiných vědních oborů. Nicméně je třeba brát na vědomí specifické aspekty tohoto datového zdroje, které jsou diskutovány dále v článku. Možností využití podobných datových zdrojů se ve své práci zabývá také Baram-Tsabari, Segev (2011) a v závěru doporučují využití těchto datových zdrojů jako další z možných zdrojů pro práce zaměřené na analyzování zájmů populace, vytváření mezinárodních komparací a objevování motivací, proč lidé vyhledávají konkrétní termíny.

Vyhledávací algoritmy společnosti Google umožňují využít dvě varianty analýz pro potřeby klasifikace např. měst či regionů (Boulton a kol. 2011). První se nazývá analýza webového obsahu *(Web Content Analysis)*, kdy jako hlavní zdroj informací slouží počet existujících odkazů, které vyhledávač Google nalezne pro zadané město a následně podle počtu existujících odkazů daná města porovnává. Jedná se však spíše o pasivní informace, které jsou publikovány na internetu a nesouvisí přímo se zájmem obyvatel, jelikož dané články či odkazy nemusí být vůbec navštíveny či přečteny. Druhou metodou je pak analýza webové aktivity *(Web Activity Analysis)*, která se zabývá již aktivním vyhledáváním daných měst internetovými uživateli. Tento typ analýz je předmětem zájmu tohoto článku, jehož cíle lze rozdělit do dvou rovin. První je rovina teoreticko-metodologická, jejímž cílem je představit nový zdroj dat a poskytnout doporučení ke zpracování těchto dat pro možnosti využití v široké škále dalších analýz. V empirické části jsou pak metodika a datový zdroj aplikovány s cílem kvantifikovat a analyzovat zájem o velká města Česka uživateli internetového vyhledávače Google, a to na základě analýzy nejčastějších témat (příležitost k zábavě, cíl pro hledání práce, sport, turistické cíle, zdravotnické či vzdělávací instituce apod.), které tito uživatelé vyhledávají spolu s jednotlivými názvy měst. Ta mohou být pro jednotlivá města typická, resp. uživatelé si je mohou s danými městy vnitřně spojovat.

Hledání geografických cílů

Obecně je možno uvažovat o důvodech, které někoho vedou k vyhledání informací o daném objektu na internetu. Lze je rozdělit na 2 základní skupiny – vyhledání pojmu, který je pro uživatele neznámý (např. nabídka na akci, která se koná na jemu neznámém místě); nebo vyhledání pojmů, o kterých má jistou představu, a kde si potřebuje ověřit připojenou informaci (pracovní doba známého obchodu či instituce, které hodlá navštívit; program kina či sportovního klubu; seznam místních zajímavostí v destinaci, kterou plánuje navštívit). Právě tato druhá skupina je spojena s jistými mentálními představami člověka, a tedy do určité míry souvisí i s mentálními či kognitivními mapami (Nižnanský 2006). Subjektivní vnímání okolí odráží i jeho individuální rysy a priority. Dotazy jsou kladeny v souladu s představami o požadovaném objektu či službě, představami o jeho pojmenování či klíčových slovech, které mohou usnadnit nalezení požadované informace. Je potřebné si uvědomit, co vlastně je předmětem hledání na internetu. Většina dotazů nesměřuje na získání základních informací o městě (velikost, historie, členění, jména zastupitelů), ale na služby typu kino, fotbal, úřad, zaměstnání, místní doprava, ubytování. Hodnotíme zájem o město prostřednictvím evidence zájmu o poskytované služby, což je ovlivněno řadou objektivních faktorů (existence služeb, jejich marketing apod.) i subjektivních faktorů (zejména individuální potřeby, hodnocení a prioritizace služeb).

Pokud máme k dispozici dostatečné množství záznamů o individuálním chování osob, stává se validní výpovědí o chování, a přeneseně i výpovědí o názorech a postojích, základní populace v daném časovém intervalu. Základní populací je v tomto případě soubor uživatelů vyhledávače Google přistupujících

na internet v Česku, která je ale dostatečně velká (odhad jeden milion vysvětlený v kapitole Diskuse) pro získání významných závěrů.

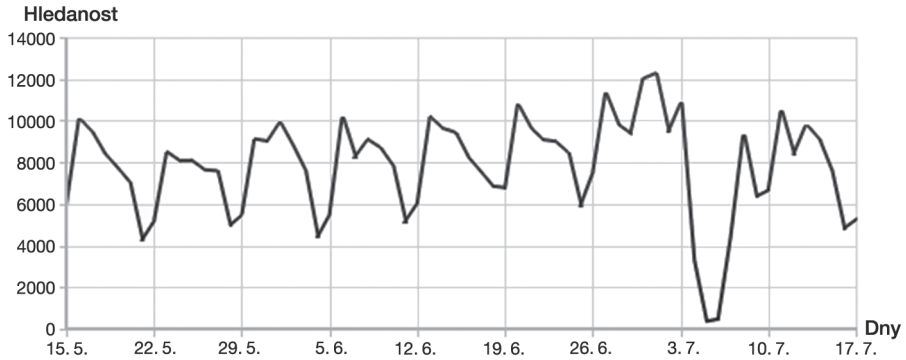
Využití statistiky vyhledávače Google je demonstrováno na vyhledávání názvů velkých měst Česka se zaměřením především na četnosti vyhledávání a na asociace termínů použitých při vyhledávání. K lepšímu hodnocení výsledků by rovněž přispělo rozčlenění objemu vyhledávání na dotazy položené místními obyvateli a ostatními.

Siwek, Bogdová (2007) rozlišují vnitřní a vnější faktor ve vnímání regionů a spojují je s mírou identity obyvatel (vnitřní faktor) a mírou percepce ve vědomí ostatních obyvatel (vnější faktor). Rozlišení vnitřního a vnějšího faktoru je v GI možné pomocí exportu statistik, kde Google mimo jiné uvádí název místa, ze kterého byl dotaz zadán. Vlastní hodnocení výsledků však vykázalo některé podivné anomálie a proto autoři doporučují výstupy z interní lokalizace Google důsledně kontrolovat a zvažovat vhodnost jejich použití. Jinou možnost hodnocení představuje posouzení typu dotazu a podle jeho tématu přiřazení k vnitřním a vnějším faktorům. S určitou nepřesností totiž lze odhadovat, že některé hledané fráze lze dobře započítat ve prospěch vnitřního faktoru (fráze ve vztahu k zapojení obyvatel do místní správy, instituce, volný čas, nákupy apod.) a jiné naopak ve prospěch vnějšího faktoru (cestování, ubytování, turistické atraktivity, lázeňství apod.), i když zůstává řada frází, u kterých je nutno počítat s přispěním k oběma faktorům (např. vzdělání).

Vyhledávače a jejich statistiky

Vyhledávače hledají zadaný dotaz „fulltextově". Zpravidla prohledávají titulek, popis, obsah webové stránky a také odkazy na tuto stránku na jiných webových adresách (Hlavenka 2004). K vyhledávačům, které poskytují informace o vyhledávání zadaných dotazů, patří v současnosti především Google a v Česku také vyhledávač Seznam. Jak uvádí Procházka (2012), ještě v letech 2010 a 2011 byl mezi českými uživateli internetu nejvíce používaným vyhledávačem Seznam.cz (téměř 60 %), nicméně v současnosti má již největší podíl na vyhledávání v Česku právě Google (v lednu 2011 51 %). Česko tedy následuje trend ve většině zemí, kde je nejpoužívanějším vyhledávačem právě Google. Např. v USA vede Google s přibližně 64 %, následovaný Bing (17 %) a Yahoo! (Consumer Search 2011). Ostatní vyhledávače používá pouze malý zlomek českých uživatelů. Z hlediska sledování statistiky vyhledávání nabízí Seznam službu, která zobrazuje absolutní hodnoty vyhledávání zadaného dotazu během posledních dvou měsíců. Google oproti tomu poskytuje již výše zmíněný nástroj *Google Insights*, resp. *Google Trends*, který je zatím ze srovnatelných produktů nejvyspělejší, a proto se jím také tento článek zabývá přednostně. Z dalších vyhledávačů je možno uvést Yahoo!, které nabízí službu zabývající se prohledáváním zpráv novin New York Times.

Statistiky vyhledávače Seznam nepřináší tak komplexní pohled na vyhledávání daného hesla jako GI. Přínosem oproti svému konkurentovi je zobrazování absolutních hodnot vyhledávání a nevýhodou pak je jeho orientace na poměrně krátký časový interval (60 dní) a chybějící export dat. Na obrázku 1 je uveden jeden z typů výstupů na dotaz *karlovy vary* ze dne 17. 7. 2011 (rozšířená shoda).



Obr. 1 – Statistika dotazu vyhledávače Seznam pro dotaz *karlovy vary* (17. 7. 2011, rozšířená shoda). Pramen: Seznam.cz.

Je sledováno období konání filmového festivalu, přesto však při porovnání s výsledky z ledna 2011 jsou četnosti dotazování podobné (denní maximum pro období 15. 5. – 17. 7. 2011 je 12 280 dotazů, zatímco pro období 22. 11. 2010 až 24. 1. 2011 bylo 10 771 dotazů). Z grafu jsou dobře patrné týdenní cykly četnosti (vrchol v pondělí, minima ve dnech volna, absolutní propad o státních svátcích) i základní trend ve vývoji. Výsledky naznačují, že vyhledávání informací se většina lidí věnuje v pracovním týdnu (a zřejmě hlavně v pracovní době).

Aplikace *Google Insights for Search*, v české verzi *Google Statistiky* vyhledávání, resp. *Google Trends*, slouží k porovnávání statistik vyhledávaných dotazů ve vyhledávači Google. Pro plné využití služeb tohoto nástroje je potřebné mít účet Google. K dispozici je relativně dlouhá časová řada počínaje rokem 2004. Při zadávání dotazů je možné využít logické operátory. Základní zápis víceslovného dotazu s mezerami *ostrava loutkové divadlo* započte dotazy, které obsahují všechna tato slova v libovolném pořadí. V případě využití logických operátorů je možné využít znak „+“, který zastupuje „nebo“ (zadáním *ústí* + *usti* jsou sečteny výsledky pro zápis s diakritikou a bez diakritiky), znak „-“ definuje rozdíl (používá se pro odečtení výsledků, které se týkají jiného významu slova, např. dotaz pro město Tábor: *tábor -letní -dětský)*. Pokud je třeba zjistit statistiku vyhledávání přesného slovního spojení, je výraz zapsán do uvozovek *„zoo dvůr králové nad labem“* (*Google Statistiky vyhledávání* 2011). Jak již bylo naznačeno, dotaz s diakritikou je považován nástrojem GI za zcela odlišný od stejného dotazu bez diakritiky. Velké a malé znaky nejsou rozlišovány. Ve filtru oblasti vyhledávání je možné vymezit zájmovou statistiku z hlediska času (délka období, členění po měsících), zdroje dat (web, obrázky, zprávy, výrobky – pro Česko zatím jen webové vyhledávání), lokality, odkud byl dotaz zadán (státy a podoblasti; v Česku směs bývalých a nových krajů, jinde zpravidla NUTS 2), tematické kategorie (27 oblastí, bohužel pro Česko toto rozdělení zatím není dostupné).

Výstupy lze získat ve formě časové řady, geografické statistiky (státy a oblasti) a tematické statistiky (nejčastější hledané fráze; obr. 2). Export výsledku ve formátu CSV umožňuje další zpracování dat. Ve všech případech jsou ale poskytovány pouze relativní hodnoty vyhledávání (RHV), což je významný rozdíl

Obr. 2 – První část výstupu aplikace GI. Vyhledávání dotazu *„karlovy vary“* a „mariánské lázně“, 1 – graf zájmu v průběhu času, 2 – relativní objemy vyhledávání dotazů, 3 – možnost předpovědi a titulků zpráv. Pramen: *Google Insights for Search.*

oproti Seznamu. Nejčetnější výskyt získává hodnotu 100 (maximální výskyt ve sledovaném časovém období či nejčetnější hledaná fráze), ostatní hodnoty jsou vyjádřeny relativně podle ní (mají např. 80% četnost). Za celé sledované období pak GI vypočte relativní objem vyhledávání (ROV) jako průměrnou hodnotu ze všech relevantních hodnot vyhledávání. Je nutné zdůraznit, že společnost Google umožňuje zobrazit statistiky vyhledávání jen pro dotazy, které překročily určitý práh počtu vyhledávání. Nicméně konkrétní hodnota tohoto prahu není zveřejněna. Více prostoru je tomuto problému věnováno v závěrečné diskuzi.

Při jednom zpracování jsou získány výsledky pro maximálně 5 dotazů, což je nedostatečné pro většinu potřeb. Proto je důležité použít vhodnou formu kvantifikace výsledků. Proces přepočtu na společnou lineární škálu hodnot lze označit za kalibraci hodnot. Ke kalibraci se používá párové srovnání statistiky hledaného výrazu (města) s etalonem. Jako etalony byly použity vybrané geografické názvy z důvodu srozumitelnosti, nicméně je možné použít i negeografické výrazy. Soustava etalonů by měla splňovat následující podmínky – název neobsahuje znaky s diakritikou, následující etalon má poskytovat desetinu hodnoty vyššího a výsledky etalonu musí být stabilní v čase. Druhá podmínka týkající se snižování zájmu na desetinu nebyla vždy zcela splněna. Třetí podmínka vychází ze skutečnosti, že se ve výsledcích vyskytuje jistá nepřesnost, resp. nestabilita v čase. Důvodem je používání aproximací pro výpočet těchto výsledků, jak dokládá sdělení na hlavní stránce *Google Statistiky vyhledávání* (2011). V rámci analýzy statistiky výsledků navrhovaných etalonů byly proto po dobu jednoho měsíce každý den opakovány párové dotazy na etalony (*praha* a *brno*, *praha* a *pardubice*, *pardubice* a *poruba*, *poruba* a *mohelnice*, *mohelnice* a *radonice*). Z výsledků relativního objemu vyhledávání (průměr za šestileté období) byly vypočteny přepočtené objemy vyhledávání (POV) a hodnotila se odchylka v POV pomocí variačního koeficientu (tab. 1). Výsledky ukazují, že variační koeficient pro přepočtený objem vyhledávání je velmi nízký (od 0,6 % pro etalon *brno* po 7,6 % pro etalon *radonice*). Lze shrnout, že variabilita hodnot, které poskytuje nástroj *Google Insights* pro zvolené etalony, je nízká a prakticky neovlivní výsledky dotazování.

Tab. 1 – Statistické ukazatele přepočteného objemu vyhledávání (POV) pro etalony (Návratová 2011)

Ukazatel	Dotaz zadaný do GI				
	brno	*pardubice*	*poruba*	*mohelnice*	*radonice*
Průměr POV	56 932,5	9 873,0	964,2	491,8	48,9
Směrodatná odchylka POV	364,3	73,7	30,0	16,4	3,7
Variační koeficient POV (%)	0,6	0,7	3,1	3,3	7,6

Tab. 2 – Srovnávací stupnice, hodnoty ke dni 4. 1. 2011

Text 1. etalonu	Text 2. etalonu	ROV 1. etalonu	ROV 2. etalonu	POV 1. etalonu	Kalibrační koeficient etalonu	Přibližný poměr statistik
praha				100 000*		
pardubice	*praha*	8	79	10 127	1 265,875	1:10
poruba	*pardubice*	8	80	1 013	126,625	1:10
mohelnice	*poruba*	37	76	493	13,324	1:2
radonice	*mohelnice*	6	58	51	8,5	1:10

Pozn.: ROV – relativní objem vyhledávání, POV – přepočtený objem vyhledávání; * zvolený výchozí bod stupnice

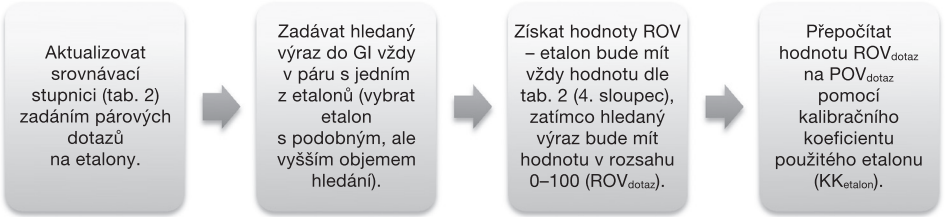
Získané průměrné výsledky četností přepočteného objemu vyhledávání pro jednotlivé etalony jsou dále využity pro kalibraci statistik hledaných výrazů. Každý dotaz je tedy díky srovnání se sadou etalonů přepočten ze svého relativního objemu vyhledávání na přepočtený objem vyhledávání podle vztahu (1). Hodnota přepočteného objemu vyhledávání nejméně vyhledávaných cílů (ale ještě zveřejněných na Google) je přibližně 1.

$$POV_{dotaz} = KK_{etalon} \times ROV_{dotaz}, \quad (1)$$

kde POV_{dotaz} je výsledná kalibrovaná hodnota objemu vyhledávání příslušného dotazu, KK_{etalon} je kalibrační koeficient etalonu, který se vypočte jako podíl POV_{etalon} a ROV_{etalon}, ROV_{dotaz} je hodnota poskytnutá aplikací GI při provedeném párovém srovnání pro hledaný dotaz, POV_{etalon} je kalibrovaná hodnota etalonu, ROV_{etalon} je hodnota etalonu zjištěná při srovnání s nadřazeným etalonem (tab. 2). Tabulku 2 je potřebné aktualizovat nejlépe v den provádění dotazování.

Vyhledávání názvů měst v Česku

Pro posouzení četnosti vyhledávání názvů měst v Česku byly použity statistiky *Google Insights for Search* (GI). Výběr názvů jsme omezili na města v Česku s počtem obyvatel 10 tisíc a více (zdroj dat ÚIR-ZSJ, stav k 1. 1. 2010). Statistiky z GI byly získávány 3. 1. a 4. 1. 2011. Aplikace GI byla nastavena místně na celé Česko a časově od začátku listopadu roku 2004 do konce října roku 2010 (celkem 6 let), aby se zamezilo sezónním vlivům. Provádělo se párové dotazování – paralelně byly položeny dva dotazy, z nichž jeden je zájmový a druhý

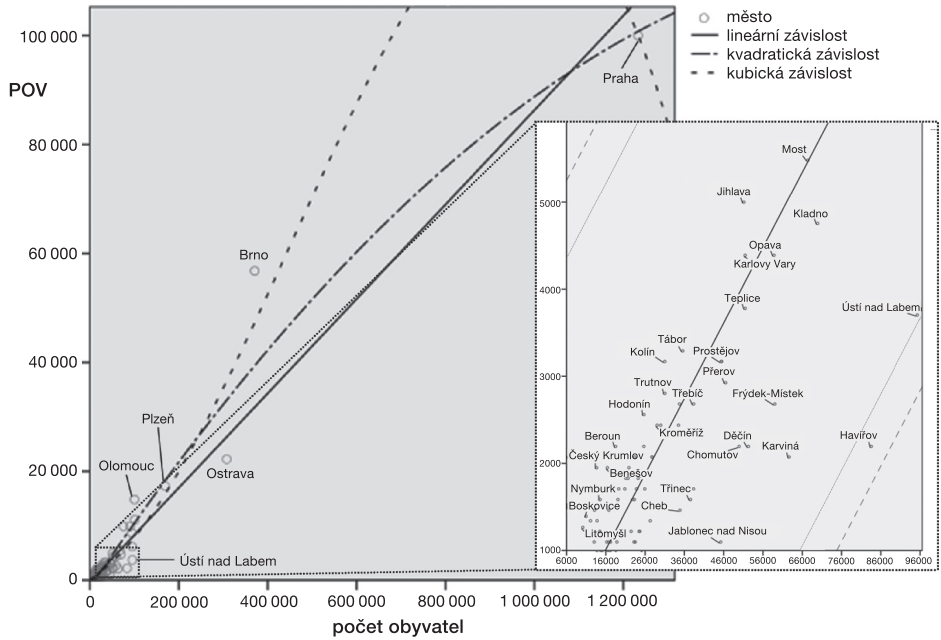


Obr. 3 – Postup stanovení přepočteného objemu vyhledávání

etalonový pro kalibraci. Následně se vypočítal přepočtený objem vyhledávání podle výše uvedeného vzorce (1) a postupu (obr. 3). Pro eliminaci problému s diakritikou byly dotazy zadávány ve tvaru s i bez diakritiky. Výjimkou bylo město Aš, kde byl zadán název pouze s diakritikou pro vyloučení záměny s anglickým slovem „as“. U názvů měst, která vznikla spojením dvou sídel jako Frýdek-Místek a Brandýs nad Labem-Stará Boleslav, byly vypsány varianty oficiálního názvu s pomlčkou, mezerou místo pomlčky, názvy dílčích částí, a to vše s diakritikou a bez diakritiky. Město Dvůr Králové nad Labem mělo při zadání celého názvu relativní objem vyhledávání několikanásobně nižší než při zadání *dvůr králové* + *dvur kralove*, proto byla zvolena tato kratší varianta. Dotazy na města s mnohoznačným názvem (synonyma) byla zapsána pomocí logického rozdílu v aplikaci GI (tj. pomocí znaku „-“), kterým byly vyloučeny z vyhledání odlišné významy slov v dotazu. K identifikaci slov odlišného významu, která se významně podílí na výsledku statistiky, byl použit seznam nejčastěji vyhledávaných výrazů, který GI poskytuje. Týkalo se to dotazů (*zábřeh -ostrava), (most -černý -cerny -wanted -karlův -karluv -nuselský), (tábor -letní -dětský -koncentrační), (hranice -smrti -čr -pozemku -věková -státní), (slaný -dort -koláč -michal -závin), (chodov -praha -obchodní -nákupní -oc -obchodni -hypernova -centrum -prague -park -mercedes), (jičín -nový -novy -starý -stary).* V případech, kdy nebylo možné mnohoznačné výrazy vhodně odečíst, byl použit opačný mechanismus, tj. logické sčítání nejčastěji vyhledávaných výrazů, které mají spojitost s daným sídlem (např. *ostrov nad ohří* + *ostrov nad ohri* + *kino ostrov* + *město ostrov* + *mesto ostrov)*. Tato neúplná kombinace hesel však snižuje relativní objem vyhledávání, s čímž musíme počítat při interpretaci.

Hodnocení objemu vyhledávání měst Česka

K hodnocení se použilo několik typů výstupů. Základní kartodiagram přepočteného objemu vyhledávání (obr. 5) zobrazuje absolutní vyjádření objemu vyhledávání jednotlivých měst, je však obtížné jej interpretovat bez vhodného modelu. Lze předpokládat, že četnost vyhledávání by měla být úměrná velikosti města. Vyšší počet obyvatel by tedy měl znamenat vyšší počet jimi generovaných dotazů (vnitřní faktor) a zároveň větší město poskytuje více služeb (instituce, obchodní a volnočasové aktivity, vzdělání) a tak na něj bude směřováno více dotazů od obyvatel z jiných sídel (vnější faktor). Proto byla jako základní teoretický model zvolena regresní závislost mezi počtem obyvatel města a četností jeho vyhledávání. Nejvyšší hodnota koeficientu determinace

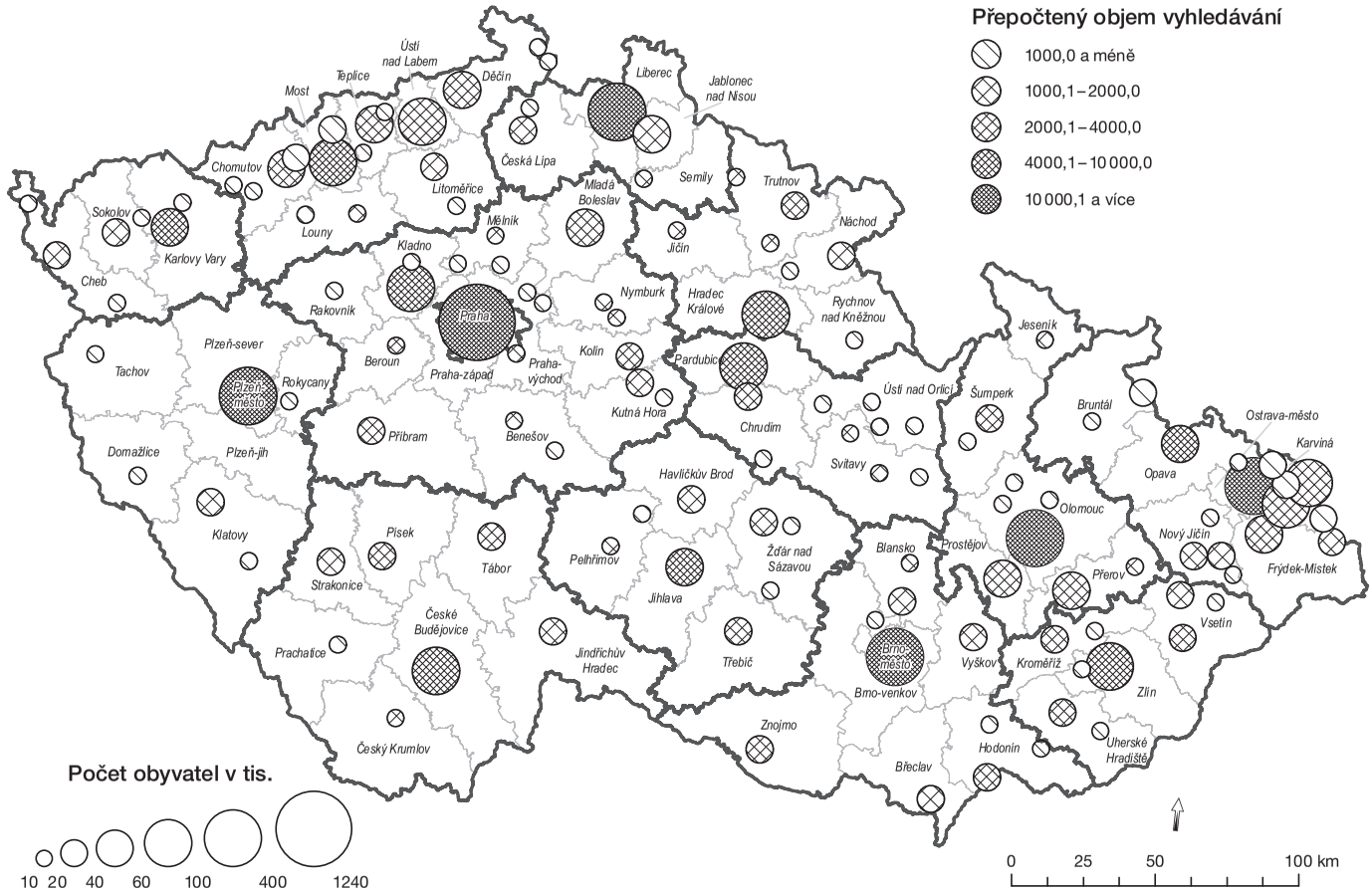


Obr. 4 – Závislost přečtených objemů vyhledávání (POV) na počtu obyvatel města (regresní křivky), vpravo detail s intervaly spolehlivosti 90 a 95 %. Pramen: přepočtená data GI, vlastní zpracování.

byla zjištěna pro kubickou závislost ($R^2 = 0{,}966$), následuje kvadratická závislost ($R^2 = 0{,}952$) a lineární závislost ($R^2 = 0{,}936$). Rozdíly však nejsou velké (obr. 4). Kubická závislost má evidentně nesmyslný průběh, kvadratická je více vychylována nejvyššími hodnotami (především Praha a Brno) a její průběh se bude více lišit v jednotlivých časových obdobích. Nejjednodušší interpretaci a použití má lineární regrese. Její rovnice je $Y = 0{,}087 \times X - 376{,}243$. Na základě tohoto regresního modelu byly vytvořeny dva typy výstupů – vlastní hodnocení pozice města v regresním grafu (obr. 4) a kartodiagram vzdáleností od regresní přímky (obr. 6) jako míry odchylky od základního regresního vztahu závislosti na počtu obyvatel. S touto vzdáleností klesá pravděpodobnost, že by dané město respektovalo obecnou závislost mezi počtem obyvatel a objemem vyhledávání.

V práci Nemec, Horák (2009) byla zjišťována závislost mezi počtem zpráv o městě v RSS kanálu ČT24 a počtem obyvatel města. Po testování několika závislostí uvádí autoři jako nejpřesnější kubickou závislost, následovanou kvadratickou závislostí a lineární závislostí s koeficientem determinace $R^2 = 0{,}752$. Pro grafické znázornění se použila lineární závislost s 95% intervalem spolehlivosti. Použití 95% a 90% intervalů spolehlivosti pro přepočtené výsledky z GI nevedlo k efektivní klasifikaci měst, proto byl větší důraz kladen na interpretaci vzdálenosti od regresní přímky.

Lineární regresní model je spjat s určitými předpoklady v datech, zdůrazňuje pozici a odchylky největších měst, zatímco rozdíly pro menší města jsou neznatelné. Proto bylo hodnocení doplněno o výpočet Spearmanova koeficientu



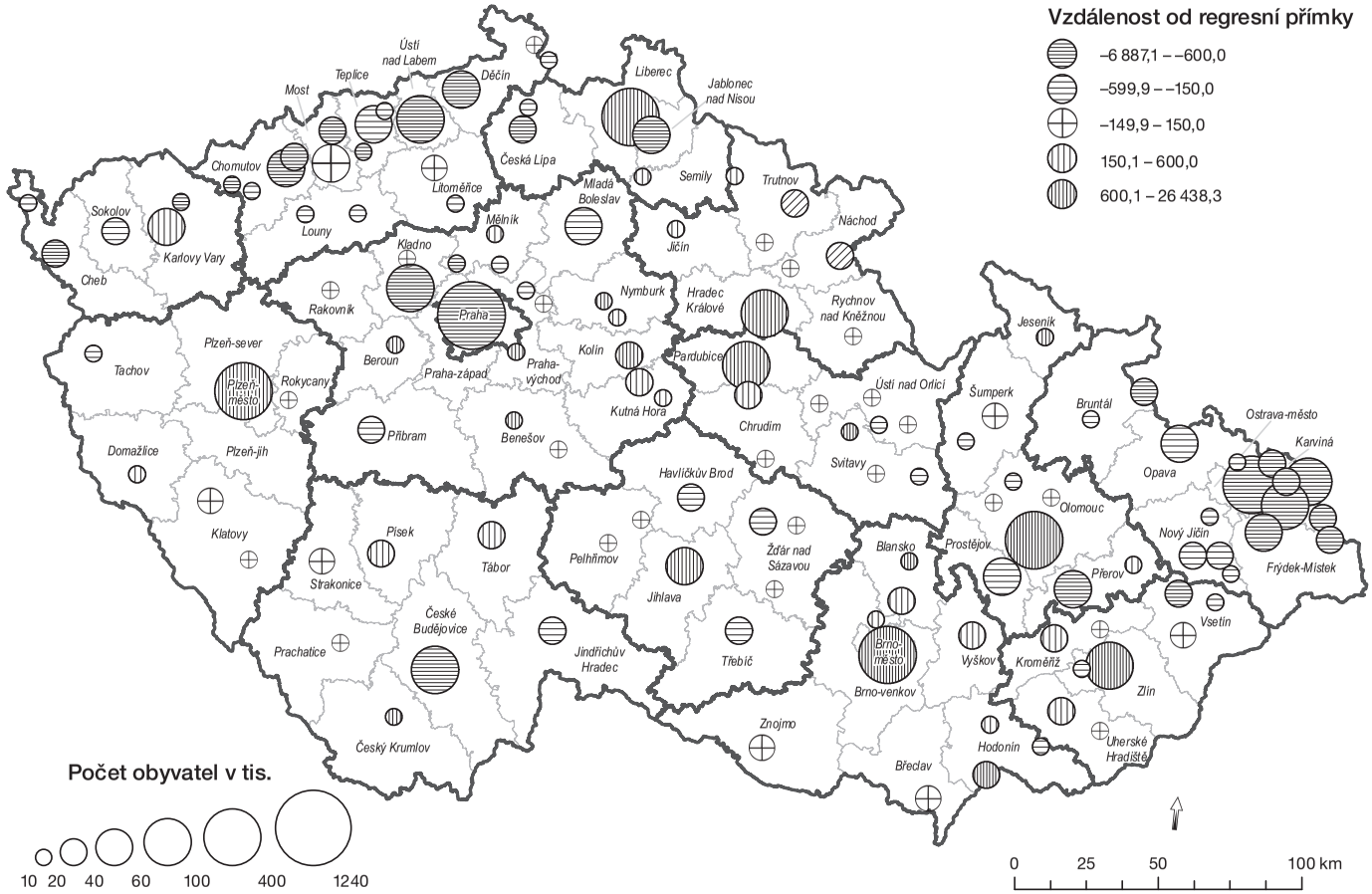
Obr. 5 – Přepočtený objem vyhledávání názvů měst nad 10 000 obyvatel ve vyhledávači Google. Pramen: přepočtená data GI, vlastní zpracování.

korelace pro pořadí podle počtu obyvatel a pořadí podle přepočteného objemu vyhledávání. Výsledný koeficient je roven 0,81 (interval <0,72; 0,87> na hladině významnosti 0,05). To potvrzuje výsledky regresní analýzy a vhodnost zvoleného teoretického modelu závislosti objemu vyhledávání na počtu obyvatel. Kartodiagram odchylek v pořadí (obr. 7) je vhodný pro hodnocení pozice menších měst, protože rozdíly v pořadí velkých měst nemohou být velké. Rozdělení do třídních intervalů rozdílu pořadí bylo provedeno s využitím směrodatné odchylky, jejíž použití je vhodné pro data s normálním rozdělením (normální rozdělení bylo ověřeno pomocí Pearsonova χ^2 testu na hladině významnosti 0,01).

Současně se hodnotí i témata nejčastějších dotazů spojených s názvy měst (ukázka vyhodnocení témat dotazování pro několik vybraných měst je na obr. 8) a dále hodnocení míst, ze kterých jsou dotazy na města položeny. Protože je autory považována identifikace lokalizace položení dotazu na název města (interní lokalizace Google) ve sledovaném období za nespolehlivou (některé výsledky nebyly uspokojivě vysvětlitelné, jako např. vysoká intenzita dotazování Zlína z Nejdku nebo Prahy ze Strakonic), byla interpretace omezena na výpočet podílu vnitřního faktoru u větších měst. Odhad podílu vnitřního faktoru se počítá jako množství dotazů položených z domovského města k množství všech ostatních známých dotazů na dané město z Česka. Z konstrukce výpočtu vyplývá, že jde o aproximaci obecně spíše přeceňující podíl vnitřního faktoru, protože některá zdrojová místa nejsou uvedena ve statistice, kterou poskytuje GI jako místa položení dotazu na název města.

Praha je samozřejmě nejvíce vyhledávané město v Česku. Její pozice se ale obtížně hodnotí, protože objem vyhledávání tvoří krajní hodnotu variační řady, ke které je vázán průběh regresní závislosti. Rozdíl v pořadí je nulový, poloha pod regresní přímkou není vypovídající. Podobně dopadlo i hodnocení statistiky zpráv regionálního kanálu ČT24. Nejčastěji vyhledávané dotazy pro Prahu se zaměřují hlavně na její městské části: *praha 4*, *praha 6*, *praha 1*, *hotel*, *praha 5*, *praha 10*, *restaurace*, *praha 2*, *mhd*, *praha 8*, *praha 3*, *praha 9*, *praha 7*, *mapa*, *metro*, *zoo*. Odhad vnitřního faktoru pak je 43 %, což znamená, že 57 % dotazů bylo provedeno z jiných míst Česka. Brno se odchyluje nejvíce ze všech měst od regresní závislosti (tedy i z intervalu spolehlivosti). I když je absolutní hodnota odchylky ovlivněna 2. pozicí v řadě, jednoznačně lze konstatovat, že Brno je výrazně nadprůměrně vyhledávané ve vyhledávači Google. Nejčastěji vyhledávané dotazy jsou spojeny se zábavou, cestováním a kulturou, následované hledáním práce, univerzitami a zdravotnictvím: *restaurace*, *hotel*, *divadlo*, *olympia*, *práce*, *kino*, *vut*, *nemocnice*, *mhd*, o několik příček dále *masarykova univerzita*, *mu*. Jak ale ukazuje odhad vnitřního faktoru (76 %), vysoký podíl dotazování je tady generován přímo obyvateli Brna (obr. 5).

Zájem o Ostravu je charakterizován jako slabě podprůměrný. Zvláště kontrastní je srovnání s Brnem. Ačkoliv má Ostrava jen o 17 % méně obyvatel než Brno, zájem o ni při vyhledávání v prostředí Google je téměř třetinový (o 61 % menší). Nejčastěji vyhledávané dotazy pro Ostravu jsou: *cinestar*, *práce*, *hotel*, *restaurace*, *colours of ostrava*, *nemocnice*, *program*, *divadlo*, *zoo*, *mhd*, *futurum,* o několik příček dále *univerzita*, *všb*. Z obrázku 8 vyplývá, že zobecněná struktura témat dotazování je podobná Brnu, kdy větší rozdíly v 8 kategoriích jsou patrné pouze v podílu nakupování a volnočasových aktivit (11 % nakupování v Ostravě proti 18 % v Brně, 40 % volnočasové aktivity proti 27 % v Brně).



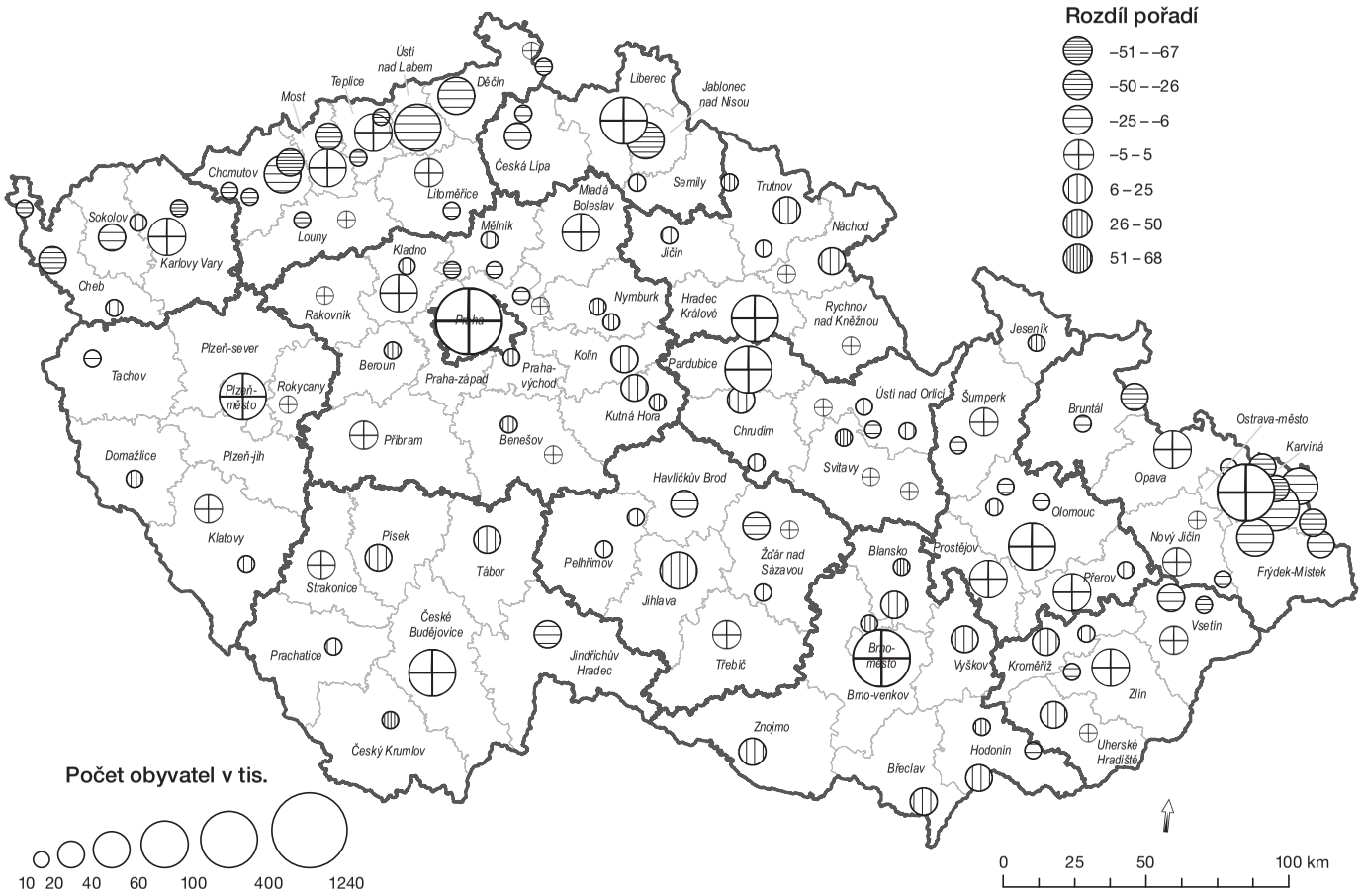
Obr. 6 – Odchylky od regresní přímky závislosti mezi počtem obyvatel a objemem vyhledávání (města nad 10000 obyvatel ve vyhledávači Google). Pramen: přepočtená data GI, vlastní zpracování.

Hlavní rozdíl mezi oběma městy je tedy v objemu dotazování. Odhad vnitřního faktoru pro Ostravu je 47 %, nižší hodnota vnitřního faktoru ve srovnání s Brnem je patrně způsobena větším počtem velkých měst v okolí Ostravy, jejichž obyvatelé často využívají služby poskytované jen na území města Ostrava (multikino, větší nabídka práce, divadla, velká nákupní centra apod.). To potvrzují také města, odkud se Ostrava významně vyhledává – Opava, Karviná a Frýdek-Místek, zatímco v případě Brna žádný takový vztah na blízké obce není, a nejsilnější město, odkud se uživatelé dotazují na Brno, je Praha. Z výsledků na obrázku 6 a 7 je navíc zřejmé, že se tato města v okolí Ostravy vyznačují podprůměrným zájmem o vyhledávání v prostředí Google. Pokud by se seskupily výsledky měst Ostrava, Havířov, Karviná, Orlová, Bohumín, Opava, Frýdek-Místek, Český Těšín a Třinec, získáme aglomeraci s celkovým počtem 689 626 obyvatel, avšak přepočtený objem vyhledávání je pouze 37 342, což je pouze 63 % teoretické hodnoty odvozené z výše uvedeného regresního modelu. Srovnání s nadprůměrným objemem vyhledávání pro Brno (56 790) je velmi kontrastní.

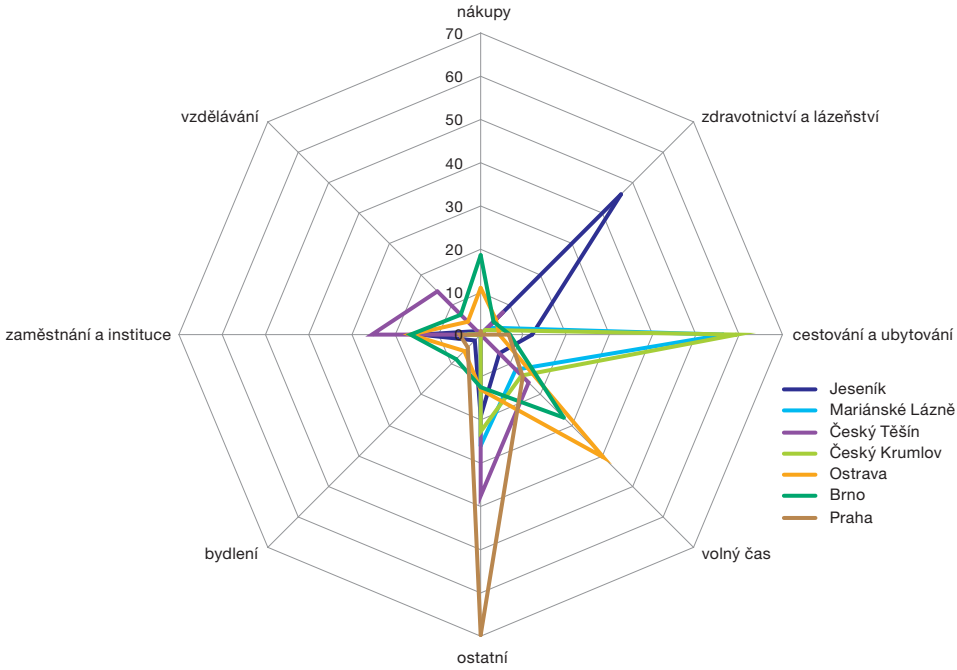
Dalším výrazně nadprůměrně dotazovaným městem je Olomouc. Vybočuje nad 95% interval spolehlivosti, v pořadí měst předbíhá Liberec (rozdíl pořadí +1). Nejčastěji vyhledávané dotazy jsou: *cinestar*, *univerzita*, *olympia*, *nemocnice*, *flora*, *knihovna*, *kino*, *divadlo*, *sigma*, o několik příček dále *up*, *univerzita palackého*. Odhad vnitřního faktoru je 59 %, zdá se tedy, že se na nadprůměrném zájmu podílí především aktivita jejích obyvatel. Podobně výrazně nadprůměrná je Olomouc i v regionálním zpravodajství ČT24 (Nemec, Horák 2009).

Do skupiny mírně nadprůměrně vyhledávaných krajských měst je možné zařadit Plzeň, Liberec, Hradec Králové, Pardubice, Zlín, Jihlavu a Karlovy Vary. Nejčastější dotazy pro Plzeň jsou: *práce*, *plaza*, *cinestar*, *město*, *mhd*, *divadlo*, *hotel*, *zoo*; odhad vnitřního faktoru 81 % svědčí o mimořádně vysokém podílu dotazování domácích obyvatel. Liberec je zadáván ve spojitosti s aquaparkem Babylon a zoologickou zahradou; odhad vnitřního faktoru je 74 %. Nejčastější slovní spojení zadávaná do vyhledávače Google s městem Hradec Králové jsou*: pardubice*, *univerzita*, *cinestar*, *nemocnice*, *práce*, *mhd*, *knihovna*; odhad vnitřního faktoru je 45 %. Pardubice jsou vyhledávány v souvislosti s hesly: *univerzita*, *kino, hc*, *nemocnice*, *práce*, *mhd*, *cinestar*; odhad vnitřního faktoru je pouze 38 % se silným druhým centrem v Chrudimi. Největší zájem o Zlín zapříčiňuje*: práce, kino, divadlo*, *nemocnice,* o několik příček dále *utb (Univerzita Tomáše Bati)*; odhad vnitřního faktoru je 27 % (vysoké zastoupení externích dotazovatelů může být silně ovlivněno anomálně vysokým podílem z Nejdku, který se po úpravě geografické lokalizace již od roku 2011 neobjevuje). Hlavní témata dotazů pro Jihlavu jsou: *zoo, city park, cinestar, vysočina, dukla, kino, práce* a *nemocnice;* odhad vnitřního faktoru je 61 %. Nejčetnější témata pro Karlovy Vary jsou v souladu s očekáváním: *festival, hotel a lázně,* odhad vnitřního faktoru je 0 % (patrně projev starší chybné interní lokalizace Google, protože nejsilnější vazby jsou na Cheb se 100 % a následují Chomutov a Most po cca 30 %).

Zbývající krajská města jsou mírně podprůměrně vyhledávaná. Nejčastější dotazy na České Budějovice zní: *práce*, *nemocnice*, *cinestar*, *úřad*, *mhd*, *magistrát;* odhad vnitřního faktoru je běžných 44 % a oproti populačnímu pořadí se město předbíhá o 2 pozice. Nejhůře ze všech krajských měst je na tom Ústí



Obr. 7 – Rozdíl pořadí měst podle počtu obyvatel a podle vyhledávání (města nad 10 000 obyvatel ve vyhledávači Google). Pramen: přepočtená data GI, vlastní zpracování.



Obr. 8 – Nejvíce vyhledávané dotazy uživateli Google pro vybraná města podle předmětu zájmu (Návratová 2011)

nad Labem, jehož pozice je výrazněji podprůměrná (pokles o 10 míst vzhledem k populační velikosti), což ale může být ovlivněno komplikovaností zadávání názvu. V souvislosti s tímto názvem jsou zadávány dotazy: *teplice*, *úřad práce*, *mhd, děčín*, *masarykova nemocnice*, *krajský úřad*. Odhad vnitřního faktoru je 55 %. Obecně se dá konstatovat, že mezi nejčastěji vyhledávaná témata ve všech krajských městech patří kina a divadla (program, rezervace lístků apod.), nemocnice, ZOO, univerzity a služby spojené s pobytem ve městě – MHD, mapy, ubytování.

Pokud bychom srovnali výsledky se zastoupením měst v regionálním zpravodajství ČT24 (Nemec, Horák 2009), vysoce nadprůměrně vystupují ve zprávách Karlovy Vary a nadprůměrné zastoupení mají i Plzeň, Liberec, Pardubice, Jihlava, České Budějovice i Ústí nad Labem.

Další velká města mají z hlediska objemu dotazování uživateli Google pozici spíše podprůměrnou s ohledem na počet obyvatel. Město Most je průměrně vyhledávané; nejčastější téma se týká autodromu. Kladno a Karviná jsou klasifikovány jako slabě podprůměrné. Kladno láká uživatele svým aquaparkem a hokejovým klubem. U Karviné se první tři nejčastěji vyhledávané dotazy týkají práce. To vypovídá o vysoké nezaměstnanosti ve městě. Havířov je výrazně podprůměrně vyhledávaný, což koresponduje s jinými hodnoceními zájmu o toto město, jeho krátkou historií a nízkou nabídkou městských funkcí. Dotazy zadávané ve spojitosti s Havířovem jsou: *kino*, *nemocnice*, *práce*, *město*, *hc, knihovna, čsad, hotel*.

Tab. 3 – Pět měst s největším kladným a pět měst se záporným rozdílem pořadí*

Název města	Rozdíl pořadí
Litomyšl	68
Boskovice	65
Český Krumlov	57
Jeseník	49
Říčany	45
Jirkov	–67
Kralupy nad Vltavou	–59
Orlová	–58
Ostrov	–58
Český Těšín	–47

Pozn.: *rozdíl pořadí města podle počtu obyvatel a pořadí podle přepočteného objemu vyhledávání v GI. Pramen: přepočtená data GI, vlastní zpracování.

Celkově lze charakterizovat, že shluk málo vyhledávaných měst v prostředí Google se nachází v sociálně problematických a strukturálně postižených regionech na severozápadě Čech a východě Moravy, kdy 75 % zástupců měst s největší zápornou odchylkou od regresního modelu leží v Moravskoslezském, Ústeckém a Karlovarském kraji, což potvrzují také mapy na obrázku 6 a 7. Jednou z mnoha příčin nezájmu o tyto funkčně periferní regiony může být i nízká úroveň sociálního kapitálu (více o problematice sociálního kapitálu např. Pileček, Jančák 2010; Jančák a kol. 2010). Tyto faktory mohou být způsobeny industriální historií těchto regionů a prokazují se jejich některé společné vlastnosti. V nedávné minulosti tyto regiony prošly procesem deindustrializace a nemohou nabídnout tak pestrou paletu lákadel ve formě přírodních či kulturních atrakcí, jako stará historická města či oblasti s širokým přírodním bohatstvím. Postindustriální města začínají pomalu zlepšovat svou pozici v atraktivitě (Hruška-Tvrdý a kol. 2010). Přestože návštěvnost řady těchto měst narůstá kvůli rostoucímu zájmu o přítomné technické památky (Heřmanová 2011), tak ty stále nedosahují takové úrovně návštěvnosti jako klasické cíle, jak vyplývá z výsledků výzkumu návštěvnosti turistických cílů (CzechTourism 2011). Tato situace se promítá do nízké frekvence vyhledávání ve vyhledávači Google.

Pro hodnocení postavení menších měst byly použity rozdíly v pořadí (obr. 7, tab. 3). Na rozdíl od větších měst již není uveden podíl vnitřního faktoru vypočtený ze zastoupení měst lokalizovaných Googlem, protože zejména u menších měst je považována přesnost této lokalizace před rokem 2011 za problematickou a před jejich použitím je nutné, aby uživatel individuálně zvážil podle skladby identifikovaných měst relevantnost výsledků. Největší kladný rozdíl pořadí podle počtu obyvatel a podle vyhledávání má město Litomyšl, na jehož pozici se podílí významně především kultura a cestování. Nejvíce vyhledávané výrazy ve spojení s tímto městem jsou: *smetanova litomyšl*, *zámek, nemocnice*, *školky, hotel*, *ubytování*, *město litomyšl*, *kotelna*, *pedagogická škola*, *kino*, *kraj*, *práce*, *knihovna*, *penzion*. Další v pořadí je město Boskovice, které je pro uživatele vyhledávače Google spojeno s dotazy: *nemocnice*, *kino*, *westernové městečko*, *festival*, *zámek*, *hrad*, *lázně*, *minerva*, *ubytování*. Stejně jako u Litomyšle bude

převládat vnější faktor. Český Krumlov je typickým zástupcem města, kde je vnitřní faktor zanedbatelný, vypovídají o tom hesla: *ubytování*, *hotel*, *penzion*, *české budějovice*, *zámek*, *divadlo*, *hluboká*, *hostel*, *weather*, *festival*, *restaurace*. Zajímavá je podobnost struktury dotazování s Mariánskými Lázněmi (obr. 8), tedy zřejmě obecné vyjádření zájmu o hlavní turistické destinace. Město Jeseník potvrdilo svou pozici známého lázeňského města, kdy se 46 % nejčastěji vyhledávaných dotazů týkalo zdravotnictví a lázeňství. Je zde ale také zastoupen významně vnitřní faktor (odhad podílu vnitřního faktoru je vysokých 88 %). Podle Pilečeka a Jančáka (2010) má Jeseník vysokou hodnotu komponenty občanské participace, která vypovídá o vysoké volební účasti obyvatel a zapojení ve veřejných aktivitách. Zájem obyvatel o město mohl významně napomoci umístění na předních příčkách podle rozdílu pořadí v počtu obyvatel a objemu vyhledávání. Uživatelské dotazy pro Jeseník jsou následující: *lázně*, *ubytování*, *město*, *hotel*, *práce*, *hrubý jeseník*, *penzion*, *kino*, *úřad práce*.

Největší záporný rozdíl pořadí podle počtu obyvatel a podle vyhledávání připadá na město Jirkov. Pouze tři nejvyhledávanější výrazy: *chomutov*, *město* a *kino* vypovídají o nízké míře identity obyvatel a téměř nulovém vlivu vnějšího faktoru. Další v pořadí, město Kralupy nad Vltavou nemá ve statistikách vyhledávání údaje o nejčastějších dotazech. Město Ostrov má mnohoznačný název a formulace dotazu zjevně způsobila nízkou hodnotu rozdílu pořadí (tab. 3). U dotazu na město Orlová výrazně převládá míra identity. Uživatelé pokládají dotazy související s životem ve městě: *karviná*, *havířov*, *město*, *gymnázium*, *nemocnice*, *kino*, *orlová lutyně*, *hc*, *bazén*, *vesmír*, *orfa*. Stejně tak je tomu i u města Český Těšín, kde nejčastější dotazy zní: *třinec*, *úřad*, *karviná*, *kino*, *práce, gymnázium*, *nemocnice*, *zš*, *knihovna*. Tato dvě města jsou typickými zástupci, kde podle typu dotazování převládá vnitřní faktor.

Diskuse

Pro využití tohoto netradičního zdroje je klíčová otázka důvěry. Jedním z možných ohrožení je možnost vychýlit poskytovanou statistiku automatizovaným generováním účelových dotazů. K tomu je možné využít aplikace pro generování dotazů na danou webovou adresu typu *WAPT, JMeter* apod. Problém jsme ověřili dvoutýdenním testováním vyhledání dotazů *„Cimrman“* a *„8gh5a7sw1r“*. Zatímco v případě nejslavnějšího Čecha jsme se snažili pouze zvýšit četnost dotazování v GI, ve druhém případě byl použit bezvýznamový řetězec, který nikdo jiný nehledá, a tedy cílem bylo založit úplně novou statistiku. V případě Cimrmana bylo provedeno 90 438 dotazů od 28. 7. do 1. 8., pro druhý řetězec bylo vygenerováno 300 000 dotazů na Google v období 2. 8.–13. 8. 2011. Google se proti cílenému dotazování bránil, takže přibližně 5/6 dotazů v obou případech nebylo korektně zpracováno. I přesto byly počty realizovaných dotazů tak vysoké, že by se měly ve statistice projevit (12 331, resp. 47 992 dokončených dotazů). Žádný nárůst pro heslo *„Cimrman“* však nenastal a druhý výraz stále nemá dostupnou žádnou statistiku, tj. vykazuje příliš malý objem hledání. Pravděpodobným vysvětlením je, že dotazy od téhož uživatele (ze stejné IP adresy) jsou odfiltrovány a započteny do denní statistiky právě jednou. Statistiky GI tedy dokumentují počet různých počítačových adres,

ze kterých bylo dotazování provedeno, bez ohledu na aktivitu jednotlivců. Tato interpretace zřejmě vyhovuje pro většinu marketingových i výzkumných aplikací. Výsledky testování tak prokázaly, že nelze jednoduše ovlivnit dostupné statistiky poskytované GI.

Po tomto testování je možné shrnout základní výhody a nevýhody statistik vyhledávačů jako zdroje informací pro geografický výzkum. K hlavním výhodám patří objektivita průzkumu. Většina uživatelů neví, že je jejich činnost monitorována. Zřejmě, ani kdyby to věděli, nezmění to jejich chování a kladení otázek při vyhledávání. Uživatel tedy není ovlivněn subjektivními pocity jako při vyplňování dotazníků, s odpověďmi nejsou spojena žádná pozitivní očekávání apod. Scheitle (2011) zdůrazňuje, že zjišťování není invazivní, nevzbuzuje obavy, nevnucuje otázky a neprovádí apriorní kategorizaci odpovědí. Dále je možné vyzvednout dostupnost a levnost (data jsou veřejně a stále dostupná), snadnost interpretace (většinou snadná čitelnost a interpretovatelnost témat dotazování), možnost kontinuálního sledování v čase, velký objem dat (resp. velký počet respondentů), aktuálnost (s minimální časovou prodlevou je možné sledovat vývoj zájmu uživatelů) a možnost studia mezinárodních aspektů (srovnávání různých cílů, kladení dotazů v různých jazycích, sledování zahraničních regionů).

Mezi nevýhody se řadí skutečnost, že šetření pokrývá pouze jistou část populace, která používá vyhledávač Google na internetu. Velikost vzorku sledované populace v roce 2010 je možné odhadnout podle dostupných údajů na 1,86 mil. osob. Výpočet vychází z následujících údajů: počet obyvatel ve věku 16+ je 8 908 817, z toho 62 % jsou uživatelé internetu (ČSÚ 2010), podíl uživatelů využívající některý vyhledávač je 80 % (Nielsen Media 1997) a podíl Google mezi vyhledávači na českém trhu je 42 % v roce 2010. Stejným způsobem lze odhadnout na základě dostupných údajů velikost zkoumané populace v roce 2006 na 255 tisíc a v roce 2008 již na 1,14 mil. osob. Průměrnou velikost vzorku za zkoumané období lze tedy odhadnout na 1 mil. osob. Naopak v roce 2012 lze očekávat nárůst na 2 mil. osob.

Výrazné rozdíly v používání internetu (ČSÚ 2010) jsou vázané na vzdělání obyvatel (se základním vzděláním 17 %, zatímco s terciérním vzděláním 89 %) a na jejich věk (95 % v kategorii 16–24 let, zatímco v kategorii 75+ pouze 6 %), které se však promítají do geografických rozdílů až druhotně. Nicméně, jak uvádí např. Wagner, Hassanein, Head (2010), stejně jak stárne průměrný věk populace ve státech světa, tak právě starší část populace představuje nejrychleji rostoucí skupinu uživatelů počítačů a internetu, i když jejich chování je od mladší populace odlišné. Informace o profilu uživatelů Google není dostupná, a proto není zřejmé, zda a jak se liší od struktury populace využívající internet, resp. složení české společnosti. Pokud jsou ale srovnávány výsledky mezi sebou tak, jak je to použito v tomto článku (časové závislosti, porovnání výskytu více témat, zastoupení v území apod.), neměly by být odchylky významné. Další nevýhodou speciálně u GI je relativní vyjádření výsledků statistik. Je nutno využívat relativní srovnání v čase nebo proti jiným tématům (soustava etalonů). Alternativou může být použití absolutních četností dotazování poskytovaných některými vyhledávači v krátkodobém horizontu jako je např. Seznam.cz. Četnost dotazování na „Karlovy Vary“ (rozšířená shoda), prezentované na obrázku 1, kolísala mezi 4 000 a 12 000 denně (v krátkém sledovaném období),

měsíční vyhodnocení (15. 5.–14 .6. 2011) ukazuje průměr 7 735 dotazů denně. Odhad stanovený úměrou ukazuje na absolutní počty hledání v rozsahu 176 tisíc pro Prahu a 49 pro Nové Město na Moravě. Lze odhadnout, že průměrná denní četnost dotazů na jedno město je 2 150 (medián počtu dotazování). Stejný výsledek byl získán i při porovnání četnosti dotazování na dvě města (Praha a Varnsdorf) v únoru 2012. Celkový počet dotazů za všechna sledovaná města za jeden den je pak přibližně 760 tisíc. Přesné přepočty nejsou možné z důvodu rozdílu mezi údaji poskytovanými Google a Seznam (v případě Seznamu jen krátkodobý aktuální průměr, nejsou odfiltrovány dotazy položené mimo Česko atd.).

Některé vyhledávače publikují statistické výsledky pro dotazy až od jisté prahové četnosti. Proto se může stát, že nelze vyhodnotit všechny zadané dotazy. Dalším vážným nedostatkem je dosavadní nízká kvalita určování místa zadání dotazů (interní lokalizace provedená Google). Lokalizace v případě pevných sítí závisí především na podrobnosti popisu konfigurace datové sítě a je zdrojem anomálií zejména pro lokalizaci na úroveň obcí. Statistiky poskytované za regiony jsou zatím spolehlivější. V případě mobilních uživatelů internetu využívající bezdrátové připojení je přesnost lokalizace výrazně lepší a v budoucnu lze očekávat výrazné zlepšení kvality lokalizace poskytované Google. U těchto uživatelů jsou možnosti mnohem přesnější lokalizace na základě technického vybavení mobilního zařízení a zprovoznění jeho komponent jako je GPS modul, lokalizace s využitím síly signálu dostupných základnových převodních stanic (BTS), případně s pomocí WiFi modulu a odposlechu SSID a MAC adres okolních sítí (Krejcar, Janckulik, Motalova 2010). Jak vyplývá z prohlášení Google, v červenci 2011 došlo ke zpřesnění geografické lokalizace, která byla zpětně aplikována na data od počátku roku 2011. Tyto změny jsou částečně patrné, a to obzvláště na výše zmíněných anomálních vazbách. Dá se tedy očekávat, že se bude lokalizace nadále zpřesňovat. V současnosti je však doporučeno výsledky lokalizace míst původu dotazů Googlem individuálně kontrolovat.

Při interpretaci by také mohla vzniknout otázka na rozdíly uvnitř Česka a jejich dopady na výsledné hodnocení. Z hlediska zastoupení uživatelů v rámci Česka se nepředpokládá velký vliv na výsledné relativní srovnání. Podíl obyvatel využívající internet v jednotlivých krajích Česka nevykazuje velké rozdíly. Dle výsledků statistického šetření ve 2. pololetí 2010 (ČSÚ 2010) se podíl populace využívající internet u většiny krajů pohybuje mezi 43 a 49 %, vyjma krajů Liberecký (40 %), Jihomoravský (51 %), Královéhradecký (53 %) a Praha (59 %). Cílem vyhledávání jsou však města nad 10 000 obyvatel. Pokud se hodnotí rozdíly v počtu uživatelů internetu jen v těchto městech (promítající se do spolehlivosti vyjádření vnitřního faktoru) je podíl uživatelů internetu 62,6 % v kategorii měst s počtem obyvatel 10 000–49 999 a v kategorii 50 000 a více 66,8 % (ČSÚ 2010), což dokládá jen malé rozdíly.

Výsledky rovněž závisí i na způsobu dotazování. Z důvodu metodické jednotnosti byly vyhledávány názvy velkých měst Česka s pouze drobnými úpravami (text bez diakritiky, zkrácení víceslovného názvu, kombinace nejvíce vyhledávaných hesel v souvislosti se jménem města apod.). Nebyly vyhledávány názvy částí obcí, městských obvodů, městských částí ani názvy významných objektů či atrakcí na území sledovaných měst. Omezené hodnocení posoudilo vliv pouze u čtyř atrakcí (v závorce relativní objem vyhledávání při párovém srovnání

s posuzovaným městem): *petřín+petrin+petřínská+petrinska* (1), *praha* (79); *grand prix* (1), *brno* (79); *babylon* (9), *liberec* (63); *vřídlo+vridlo* (0); *karlovy vary* (54). Pouze v případě Babylonu byl zaznamenán významnější objem vyhledávání ve statistice, který by mohl ovlivnit výsledky pro Liberec. Relativní objem vyhledávání pro Liberec by se teoreticky zvýšil z 63 % na 72 %, což by mělo způsobit změnu přepočtených objemů vyhledávání z 11 111 na 12 700 (pro přesnější stanovení by se musely provést párová hodnocení s etalony). Zahrnutí dalších objektů či atrakcí by tedy zvýšilo zjištěné objemy vyhledávání, na druhou stranu by však každý výběr dalších názvů způsobil nejednoznačnosti – u částí měst zvýšené riziko kolize názvů (např. Vítkovice, Vinohrady), u objektů nejednoznačnosti jejich výběru (Olympia, Cinestar apod.) a rovněž riziko kolize názvů. Tato situace je dalším argumentem pro preferenci používání relativních srovnání či srovnání s modelovou situací. Na takovém hodnocení situace Liberce by se nic nezměnilo – pořadí mezi městy zůstává stejné (6. místo, i když se poněkud snížil odstup od páté Olomouce) a rovněž klasifikovaná odchylka od regresní přímky je stejná (mírně nadprůměrná pozice).

V případě využití tohoto zdroje pro komparace se zahraničím hraje svou roli také jazyk zadávání dotazu, jak upozorňují Al-Eroud a kol. (2011), kteří zkoumali statistiky zadávání oblíbených arabských výrazů do vyhledávače Google v angličtině a naopak. Zjistili, že hodnotící algoritmus závisí více na daném jazyku dotazu než na vlastním obsahu. V současné době je preferovaná angličtina, ale v blízké budoucnosti se počítá s oddělením použitého jazyka a obsahu a hodnocení bude jazykově nezávislé. V případě vyhledávání v rámci Evropy by však podobné problémy neměly být tak výrazné.

Závěr

Používání statistik vyhledávačů na internetu představuje nový zdroj dat, který přináší výhody pro některé typy výzkumu. Použití statistik GI je zpravidla spojeno s potřebou párových dotazů a hodnocení pomocí soustavy etalonů. Je možné zkoumat časový vývoj zájmu uživatelů či nejvíce vyhledávaná témata. Využití statistiky *Google Insights for Search* za 6 let je demonstrováno na porovnání vyhledávání názvu měst Česka nad 10 tisíc obyvatel uživateli Google z Česka. Byla prokázána silná regresní závislost objemu vyhledávání na počtu obyvatel. Vedle hodnocení přepočteného objemu vyhledávání se hodnotily i pozice města podle regresní přímky a intervalů spolehlivosti, rozdíly pořadí měst dle objemu vyhledávání a počtu obyvatel, témata nejčastějších dotazů. Hodnocení podílu vnějšího a vnitřního faktoru vycházelo jednak z podrobnější statistiky míst, odkud byl dotaz položen, a jednak z klasifikace tématu dotazování. Použití informace o místu, odkud byl dotaz položen (interní lokalizace Google) je zatím nedostatečně spolehlivé a při jejím využití je nezbytné kontrolovat výsledky a obezřetně je interpretovat. U některých, zejména menších měst, vznikají těžko vysvětlitelné anomálie a i druhý postup vede pouze k jisté aproximaci. Hodnocení vnějšího a vnitřního faktoru tedy zatím vyžaduje jisté větší úsilí a individuální prověření výsledků. Ostatní části, kde Google neprovádí interpretace lokalizace, se dají použít spolehlivě, jak dokázaly výsledky testování.

Výrazně nadprůměrné výsledky v objemu vyhledávání uživateli Google dle výše uvedeného regresního modelu dosáhly Brno a Olomouc, mírně nadprůměrné Plzeň, Liberec, Hradec Králové, Pardubice, Zlín, Jihlava a Karlovy Vary. Vzory (témata) vyhledávání vybraných měst se liší v některých aspektech. Zatímco pro Prahu dominuje dotazování na městské části, Brno a Ostrava mají podobné vzory dotazování vyjma zájmu o nakupování (výrazně vyšší podíl v Brně) a volnočasové aktivity (které s podílem více než 40 % dotazů dominují hledání v Ostravě). U dalších měst se kromě obecných témat (kino, zdravotnictví, MHD, práce, úřad) projevuje zájem o místní specifické aktivity (ZOO, hokej, fotbal, akvaparky, festivaly a další atrakce). Další krajská města (České Budějovice, Ústí nad Labem) a jiná populačně silná města jsou zpravidla mírně podprůměrně vyhledávaná uživateli Google. Z velkých měst má nejhorší výsledky Havířov, který má podobný objem hledání jako např. populačně mnohem menší Beroun. Z malých měst jsou relativně nejvíce vyhledávanými cíli v prostředí Google Litomyšl, Boskovice, Český Krumlov, Jeseník a Říčany, naopak nejméně vyhledávanými jsou Jirkov, Kralupy nad Vltavou, Ostrov, Orlová a Český Těšín. Jsou také evidentní menší objemy vyhledávání bývalých či současných industriálních měst na severovýchodě a severozápadě Česka. V budoucnu mohou poskytnout zajímavé výsledky vícerozměrné statistiky (např. analýza shlukování měst, jejich kategorizace) či podrobnější studium časové řady.

Výsledky získané zpracováním statistik *Google Insights for Search*, resp. *Google Trends*, nelze zobecňovat na celou populaci či populaci používající internet. Průměrnou velikost osloveného vzorku populace ve sledovaném období lze odhadnout na 1 milion osob, ale tento vzorek je určitým způsobem vychýlen. Nejde o výsledky získané výběrovým šetřením se zajištěním náhodnosti výběru. Rovněž způsob kladení dotazů může ovlivnit výsledek šetření. Proto je doporučováno používat relativní srovnání při zpracování a interpretaci výsledků.

Literatura:

Al-EROUD, A. F., Al-RAMAHI, M. A., Al-KABI, M. N., ALSMADI, I. M., Al-SHAWAKFA, E. M. (2011): Evaluating Google queries based on language preferences. In: Journal of Information Science, 37, č. 3, s. 282–292.

BARAM-TSABARI, A., SEGEV, E. (2011): Exploring new web-based tools to identify public interest in science. In: Public understanding of science, 20, č. 1, s. 130–143.

BOULTON, A., DEVRIENDT, L., BRUNN, S. D., DERUDDER, B., WITLOX, F. (2011): City Networks in Cyberspace and Time: Using Google Hyperlinks to Measure Global Economic and Environmental Crises. In: Firmino, R. J.; Duarte, F., Ultramari, C. (eds.), ICTs for Mobile and Ubiquitous Urban Infrastructures: Surveillance, Locative Media and Global Networks, Hershey (Pennsylvania), IGI Global, s. 67–87.

BREYER, B. N., SEN, S., AARONSON, D. S., STOLLER, M. L., ERICKSON, B. A., EISENBERG, M. L. (2011): Use of Google Insights for Search to Track Seasonal and Geographic Kidney Stone Incidence, the United States, Urology, 8/2011, 78, č. 2, s. 267–271.

BROWNSTEIN, J. S., CLARK, C. F., MADOFF. L. C. (2009): Digital Disease Detection – Harnessing the Web for Public Health Surveillance. New England Journal of Medicine, 360, s. 2153–2157.

CLIPP, C. (2011): An Exploration of Multimedia Multitasking: How Television Advertising Impacts Google Search, Duke University, North Carolina, 39 s.

CONSUMER SEARCH (2011), http://www.consumersearch.com/search-engine-reviews/best-search-engines> (10. 8. 2011).

CZECHTOURISM (2011): Návštěvnost turistických cílů v ČR 2010, http://vyzkum.czechtourism.cz/ (10. 8. 2011).
ČSÚ (2010): Využívání informačních a komunikačních technologií v domácnostech a mezi jednotlivci v roce 2010, http://www.czso.cz/csu/2010edicniplan.nsf/t/E4003156C1/$File/970110.pdf (10. 8. 2011).
GINSBERG, J., MOHEBBI, M., PATEL, R., BRAMMER, M., BRILLIANT, L. (2009). Letter: Detecting influenza epidemics using search query data. Nature, 457, s. 1012–1014.
GOOGLE STATISTIKY VYHLEDÁVÁNÍ (2011): Hlavní stránka, http://www.google.com/insights/search/# (10. 8. 2011)
HEŘMANOVÁ, E. (2011): Potenciál technických památek v kontextu rozvoje cestovního ruchu a rozvoje regionů v Česku. Geografické rozhledy, 21, č. 1, s. 22–23.
HLAVENKA, J. (2004): Mistrovství ve vyhledávání na internetu. 2. vyd. Praha: Computer press, 195 s.
HRUŠKA-TVRDÝ, L. a kol. (2010): Industriální město v postindustriální společnosti, 1. díl. VŠB – Technická univerzita v Ostravě, 120 s.
IBUKA, Y., CHAPMAN, G. B., MEYERS, L. A., LI, M., GALVANI, A. P. (2010): The dynamics of risk perceptions and precautionary behavior in response to 2009 (H1N1) pandemic influenza, BMC Infectious Diseases, 10, č. 1, čl. 296.
JANČÁK, V., CHROMÝ, P., MARADA, M., HAVLÍČEK, T., VONDRÁČKOVÁ, P. (2010): Sociální kapitál jako faktor rozvoje periferních oblastí: analýza vybraných složek sociálního kapitálu v typově odlišných periferiích Česka. Geografie, 115, č. 2, s. 207–222.
KREJCAR, O., JANCKULIK, D., MOTALOVA, L. (2010): Dataflow Optimization Using of WiFi, GSM, UMTS, BT and GPS positioning in Mobile Information Systems on Mobile Devices. In 2nd International Conference on Computer Engineering and Applicatons (ICCEA 2010), 19.–21. 3. 2010, Bali Island, Indonesia, 2, IEEE Comp Soc, s. 127–131.
NÁVRATOVÁ, M. (2011): Vyhledávání českých regionů a měst uživateli Google. Bakalářská práce. VŠB-TU Ostrava, 52 s.
NEMEC, P., HORÁK, J. (2009): The Geographical and Temporal Balance of Czech TV CT24 News. In: Horák, J. a kol.: Advances in Geoinformation Technologies 2009. 1. vydání. VŠB-TU Ostrava, s. 117–131.
NETMONITOR (2011): SPIR NetMonitor. Výzkum sociodemografie návštěvníků internetu v České republice. Červen 2011, http://www.netmonitor.cz/sites/default/files/vvnetmon/2011_06_total.pdf (10. 8. 2011).
NIELSEN MEDIA (1997): Search engines most popular method of surfing the web. Commerce Net/Nielsen Media, http://www.commerce.net/news/press/0416.html (10. 8. 2011).
NIŽNANSKÝ, B. (2006): Teória mapového zobrazovania na báze empiricko-teoretického geografického a kartografického výskumu. Ružomberok, Brno, 198 s.
PILEČEK, J., JANČÁK, V. (2010): Je možné měřit sociální kapitál? Analýza územní diferenciace okresů Česka. Geografie, 115, č. 1, s. 78–95.
PREIS, T., REITH, D., STANLEY, H. E. (2010): Complex dynamics of our economic life on different scales: insights from search engine query data. In: Philosophical transactions of the Royal Society a-mathematical physical and engineering sciences, 368, č. 1933, s. 5707–5719.
PROCHÁZKA, D. (2012): SEO. Cesta k propagaci vlastního webu. GRADA Publishing, Praha, s. 152.
SCHEITLE, C. P. (2011): Google's Insights for Search: A Note Evaluating the Use of Search Engine Data in Social Research. Social Science Quarterly, 92, s. 285–295.
SIWEK, T., BOGDOVÁ, K. (2007): České kulturně-historické regiony ve vědomí svých obyvatel. Sociologický časopis, č. 5, s. 1039–1053.
VOŽENÍLEK, V. (2002): Geoinformatic literacy: Indispensability or nonsense? Geografie, 107, č. 4, s. 371–382.
WAGNER, N., HASSANEIN, K., HEAD, M. (2010): Computer use by older adults: A multidisciplinary review. In: Computers in human behavior, 26, č. 5, s. 870–882.
WEBB, G. K. (2009): Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures. Issues in Information Systems, X, č. 2, s. 82–87.

Summary

SEARCHING FOR CZECH TOWNS BY GOOGLE USERS

Several web search engines (i.e. Google, Seznam, Yahoo!) provide statistics of user activities sorted by topics, time and, optionally, location. Such sources offer large volumes of data relevant for different research activities including geographical and sociological objectives. Advantages and drawbacks of these sources are discussed in the paper. The credibility of Google Insights for Search (GI) was tested using an automated generation of thousands of queries. Results show no effect of simple artificial attempts to influence the final statistics. The system of processing provided data (statistics of GI) is based on recalculation and standardisation of relative data provided by GI. The modified data is suitable for various comparative analyses.

The statistics provided by GI for names of Czech cities (cities of Czechia) in Google in the six-year period were explored and analysed according to the frequency and their associated topics. Data was calibrated using a system of six etalons (results of searching for six selected cities – table 2).

Frequency of searching Czech cities names is highly correlated with their population. Cities are classified using 95% and 90% confidence intervals of the linear regression and also according to the distance from the regression line. Small cities are evaluated using differences in the ordered lists of population size and searching frequency. The interest in Czech cities expressed by resident (an internal factor) and non-resident Google users (an external factor) is evaluated by associated topics and the share of queries located by Google as within the target city to compared to all queries. Prague dominates these statistics; the share of internal queries is about 43%. Brno shows highly above-average results and the internal factor is very high (76%). The most frequent topics searched together with Brno by Google users are associated with leisure activities, travelling and culture. Another highly searched large city is Olomouc. Most of the other regional centres have slightly above-average frequency of being searched by Google users (Plzeň, Liberec, Hradec Králové, Pardubice, Zlín, Jihlava and Karlovy Vary). Ostrava belongs to the below-average searched cities, but the thematic pattern of searching is similar to Brno (excluding higher preferences for shopping in Brno and very high share of searching leisure time activities in Ostrava). Also České Budějovice and Ústí nad Labem demonstrate a below-average frequency of searching. Extremely low results (frequency in searching by Google users) are shown for Havířov which can be explained by a low interest in this city and a low level of its social capital.

The most frequently searched small cities by Google users are Litomyšl, Boskovice, Český Krumlov, Jeseník and Říčany, usually associated with cultural events, tourism and spa activities. On the other hand, the less frequently searched small cities are Jirkov, Kralupy nad Vltavou, Orlová, Ostrov and Český Těšín. Their positions may usually be explained by a lower external interest and an insufficient internal factor.

The analysis demonstrates the usefulness of GI statistics for research purposes. Nevertheless, these outputs cannot be applied for the whole population or the internet's population. The main drawback of used GI statistics is a low reliability of the identification of the place of origin for the queries. The location estimated by Google in Czechia shows several anomalies and indicates indicates unsatisfactory results for locations estimated by Google before January 2011. The update (in July 2011) of internal algorithms applied for geographical localization improved the situation, but an evaluation of users' locations for older data (2004–2010) requires a thorough examination and verification. Since October 2012 the GI service is provided under the name of Google Trends.

Fig. 1 – Statistics of the query *karlovy vary* provided by web search engine Seznam (17. 7. 2011, extended matching).

Fig. 2 – First part of the Google Insights application output. Statistics for the query *karlovy vary* and *mariánské lázně*, 1 – a graph of interest over time, 2 – relative volumes of search queries, 3 – the possibility for the time serie's prediction and news captions. Source: Google Insights for Search.

Fig. 3 – The method of determining the recalculated volume of search queries. Box 1 – submission of the paired requests for the etalons to update the comparison table (tab. 2). Box 2 – submission of the searched query to the GI always paired with one of the etalons (choice of an etalon with a comparable but higher volume of queries). Box 3 – the determination of the value of the relative volume of queries *(ROV)* – the value of the etalon will always be drawn from table 2 (column 4), whereas the searched item will reach the value between 0 and 100 ROV_{dotaz}. Box 4 – recalculation of the ROV_{dotaz} value to the recalculated values of search frequency (POV_{dotaz}) through the calibration coefficient of the used etalon (KK_{etalon}).

Fig. 4 – The dependence of the recalculated volume of searching *(POV)* on the population of Czech towns (regression analysis). Right: detail of the confidence intervals of 90% and 95%. In legend: city, linear dependence, quadratic dependence, cubic dependence. Source: recalculated GI data, own analysis.

Fig. 5 – Recalculated volume of queries for names of cities with populations above 10,000 through Google search. The size of the mark is directly proportional to the size of the population in thousands. Source: recalculated GI data, own analysis.

Fig. 6 – Deviations from the linear regression dependence between the population and the volume of queries (cities over 10,000 inhabitants using the Google search engine). The size of the mark is directly proportional to the size of the population in thousands. Source: recalculated GI data, own analysis.

Fig. 7 – Difference in the ranking of cities by population and by the volume of queries (cities over 10,000 inhabitants using the Google search engine). The size of the mark is directly proportional to the size of the population in thousands. Source: recalculated GI data, own analysis.

Fig. 8 – Most searched phrases of Google users for selected cities according to subject of interest (Návratová 2011). Edges of the graph from the top clockwise: shopping, health & spa, travel & accomodation, leisure, others, housing, job & institution, education.

Pracoviště autorů: Institut geoinformatiky, VŠB-TU Ostrava, 17. listopadu 15, 70833, Ostrava-Poruba; e-mail: jiri.horak@vsb.cz.

Do redakce došlo 24. 8. 2011; do tisku bylo přijato 2. 6. 2013.

Citační vzor:
HORÁK, J., IVAN, I., NÁVRATOVÁ, M., ARDIELLI, J. (2013): Vyhledávání českých měst uživateli Google. Geografie, 118, č. 3, s. 284–307.