OPTIMALIZÁCIA VYBRANÝCH METÓD PRIESTOROVEJ ŠTATISTIKY V TROJROZMERNOM PRIESTORE S VYUŽITÍM KD-STROMU

Eva STOPKOVÁ[1]

[1] Katedra geodetických základov, Stavebná fakulta, Slovenská technická univerzita v Bratislave, Radlinského 11, 813 68, Bratislava, Slovenská republika
eva.stopkova@stuba.sk

Abstrakt

Článok sumarizuje vývoj a testovanie analytických nástrojov, založených na metódach priestorovej štatistiky (3D variogram a 3D analýza najbližšieho suseda), s aplikáciou *kd-stromu* na zefektívnenie vyhľadávania najbližších susedov. Nástroje boli vyvíjané ako moduly v prostredí open source softvéru *GRASS GIS 7* (GRASS GIS Development Team, 2013) s pripojenou knižnicou *Point Cloud Library* (Open Perception, Inc., 2012), ktorá okrem iného zahŕňa aj funkcie na tvorbu *kd-stromu*. Moduly boli overené numerickým porovnaním s výsledkami predchádzajúcich verzií bez použitia *kd-stromu*, pričom testovaná bola aj doba trvania výpočtu.

Abstract

The article summarizes the development and testing of analytical tools, which have been based on the methods of spatial statistics (3D variogram and 3D Nearest Neighbour Analysis), with usage of *kd-tree* that makes selection of the nearest neighbours more effective. The tools have been developed as modules in the environment of open source software *GRASS GIS 7* (GRASS GIS Development Team, 2013) with included library *Point Cloud Library* (Open Perception, Inc., 2012) that contains also functions for *kd-tree* creation. The verification of the modules involved numerical comparing with the results of previous versions without using *kd-tree* and also the process time have been tested.

Kľúčové slová: *kd-strom*; 3D analýza najbližšieho suseda; 3D kriging.

Keywords: *kd-tree*; 3D Nearest Neighbour Analysis; 3D kriging.

ÚVOD

Aj úplne odlišné metódy priestorovej štatistiky môže spájať podobný problém, napr. vyhľadávanie najbližšieho suseda alebo ďalších bodov v blízkom okolí. Túto úlohu, síce jednoduchú, ale výpočtovo a časovo náročnú hlavne pri veľkých súboroch bodov, možno zefektívniť použitím priestorových indexov. Pre 3D dáta sú v súčasnosti dostupné priestorové indexy napr. *kd-strom* a *oct-strom* v knižnici *Point Cloud Library – PCL* (Open Perception, Inc., 2012). Táto práca pojednáva o použití *kd-stromu* pri zefektívnení vyhľadávania najbližšieho suseda bodu, pri zefektívnení výpočtu variogramu a interpolácie hodnôt v trojrozmernom priestore.

1 PRIESTOROVÉ ANALÝZY V TROJROZMERNOM PRIESTORE

Prostredie 3D GIS vyjadruje tretí rozmer priestoru – výšku pomocou súradníc, ktoré sú súčasťou geometrického vyjadrenia geoobjektov rovnako ako rovinné súradnice. Tento prístup umožňuje modelovať a analyzovať rôzne javy vyjadrením pomocou funkcie polohy v trojrozmernom priestore:

$q = f(x, y, z)$

Hoci sa tzv. 2,5D GIS zameriava predovšetkým na prácu s geoobjektmi v priemete do roviny (ako v klasickej mape), tiež môže pristupovať k javom v trojrozmernom priestore podobne ako 3D GIS. Zásadný rozdiel však spočíva v tom, že ak aj súradnice vyjadrujúce výšku vstupujú do priestorovej analýzy, je to len

formou atribútu (Raper, 1992) – nie sú súčasťou geometrickej zložky geoobjektu. Výškové súradnice v 2,5D prostredí slúžia predovšetkým na trojrozmernú vizualizáciu objektov.

Priestorové analýzy v trojrozmernom priestore sa od dvojrozmerných líšia matematickým aparátom, ktorý však môže vychádzať z teórie pre dvojrozmerný priestor, ale aj špecifickými dátovými typmi a topologickými pravidlami.

2 KD-STROM

Pri práci s veľkými súbormi dát sa môžu uplatniť priestorové indexy, ktoré každému riadku tabuľky v priestorovej databáze priradia jedinečnú hodnotu a zároveň dáta usporiadajú do vyhľadávacieho stromu. Podľa (PostGIS, 2013) urýchľujú proces vyhľadávania tým, že prechádzajú dátami podľa danej štruktúry miesto toho, aby ju „skenovali“ celú.

Existuje viacero typov priestorových indexov s množstvom variácií, avšak nie každý z nich je vhodný na indexovanie 3D dát. V tejto práci bol použitý tzv. *kd-strom* z knižnice *PCL* (Open Perception, Inc., 2012).

Kd-strom (Bentley, 1975) predstavuje binárny vyhľadávací index, ktorého uzly sú tvorené jednotlivými bodmi vo vzorke. Každý uzol obsahuje smerníky (*pointer*) na ďalšie uzly vo vzorke podľa podmienok, ktoré závisia od rozmeru *k*. V trojrozmernom priestore sa body usporadúvajú od mediánu v závislosti od veľkosti súradnicových rozdielov v smere každej z osí podľa (O'Leary).

Inú dostupnú alternatívu k nemu predstavuje *oct-strom* z knižnice *PCL* (Open Perception, Inc., 2012).

3 PRAKTICKÁ APLIKÁCIA

V rámci predchádzajúcej práce boli vyvinuté dva moduly na prácu s dátami v trojrozmernom priestore v prostredí open source softvéru *GRASS GIS 7* (GRASS Development Team, 2013).

3.1 Vybrané metódy priestorovej štatistiky

Prvý z modulov, *v.kriging* (Stopková, 2013a), umožňuje interpolovať dáta v trojrozmernom priestore upravenou metódou *bežný kriging* (Ordinary kriging). Úpravy interpolačnej metódy zahŕňali predovšetkým zmeny vo výpočte experimentálneho a teoretického variogramu, ale aj použitie dátových typov určených na prácu s 3D dátami.

Ďalší modul, *v.nn_spatial_stat* (Stopková, 2013b), je založený na *analýze najbližšieho suseda* (Nearest Neighbour Analysis), t.j. na štatistickom testovaní porovnania skutočnej priemernej vzdialenosti medzi najbližšími susediacimi bodmi r_A s hodnotou priemernej vzdialenosti najbližších susedov r_E, ktorá by bola očakávaná v prípade, že body vo vzorke dát sú rozmiestnené náhodne (viď Tab. 1). Modul umožňuje posúdiť náhodnosť, resp. zhlukovitosť či pravidelnosť rozmiestnenia bodov v danej vzorke. Zásadný rozdiel oproti analogickej analýze v dvojrozmernom priestore spočíva hlavne v rozdielnom vyjadrení priemernej vzdialenosti najbližších susediacich bodov očakávanej pri ich náhodnom rozmiestnení pomocou Poissonovej exponenciálnej funkcie.

3.2 Vymedzenie najbližšieho okolia bodu

Spoločným menovateľom oboch modulov je práca s bodmi v najbližšom okolí každého bodu zo vzorky. Kým v analýze najbližšieho suseda je potrebné určiť vzdialenosť najbližšieho bodu ku každému z bodov zo vzorky, t.j. posudzuje sa vzťah dvoch bodov, pri výpočte variogramu v rámci interpolácie je potrebné vymedziť celé okolie výpočtového bodu. Toto okolie by malo byť relevantné z hľadiska vzájomného vplyvu na hodnoty javov, ktoré sú lokalizované na výpočtovom bode a na okolitých bodoch. Rovnako pri samotnom výpočte interpolovanej hodnoty nie je potrebné brať do úvahy hodnoty, ktoré vzhľadom na veľkú vzdialenosť medzi bodmi nemajú vplyv na veľkosť hodnoty lokalizovanú na výpočtovom bode.

Najjednoduchším, no zároveň veľmi neefektívnym spôsobom ako vyhľadať najbližší bod k výpočtovému bodu, resp. nájsť body v jeho najbližšom okolí, je v každom kroku cyklu vypočítať vzdialenosti medzi bodmi a určiť z nich najmenšiu (alebo ostatné relevantné).

Ak uvážime, že vo vzorke sa nachádza *n* bodov, takýto postup vyžaduje len pri výpočte vzdialeností medzi nimi *0.5·n·(n-1)* operácií. Navyše, ak pri interpolácii používame sieť interpolovaných bodov *nx·ny·nz*, kde jednotlivé premenné vyjadrujú počet riadkov *nx*, stĺpcov *ny* a hladín *nz* 3D siete, počet operácií pri vyhľadávaní relevantného okolia vzrastie na *(nx·ny·nz)·n*, pretože je potrebné počítať vzdialenosť každého výpočtového bodu od všetkých vstupných bodov. Ďalším krokom práce preto bolo zefektívnenie vyhľadávania bodov v susedstve aplikáciou priestorových indexov v trojrozmernom priestore.

3.3 Aplikácia *kd-stromu*

V tejto práci bol použitý *kd-strom* z knižnice *Point Cloud Library – PCL* (Open Perception, Inc., 2012), ktorá obsahuje množstvo funkcií na spracovanie 2D/3D obrazových a bodových dát.

Pri vyhľadávaní najbližších susedov pre výpočet ich priemernej vzdialenosti bola použitá funkcia *nearestKsearch*, pričom počet hľadaných bodov *K = 2.* Prvý bod v zozname *K* najbližších susedných bodov je totiž identický s bodom, pre ktorý prebieha vyhľadávanie. Výstupom funkcie je okrem ich indexov aj štvorec ich vzdialenosti od výpočtového bodu. Tieto hodnoty sú však v ďalších krokoch nepoužiteľné, pretože knižnica *PCL* (Open Perception, Inc., 2012) používa len presnosť *float* (PCL Users mailing list, 2012). Hoci elegantnejším riešením by bolo pracovať len s bodmi vo formáte *PCL::PointCloud*, kvôli presnosti výsledkov je zatiaľ výhodnejšie prevziať len určené indexy a používať vlastné štruktúry bodov s presnosťou *double*.

Pre účely variogramu v moduli *v.kriging* boli body v najbližšom okolí vyhľadávané pomocou funkcie *radiusSearch* s polomerom vyhľadávania daným preponou trojuholníka daného maximálnou horizontálnou vzdialenosťou a maximálnym prevýšením. Pre takúto vzdialenosť je ešte relevantné používať body na výpočet variogramu. Vhodnejšie by bolo použiť anizotropnú funkciu, avšak táto sa v knižnici zatiaľ nenachádza, preto boli vybrané body vytriedené dodatočne systémom podmienok.

4 TESTOVANIE

Oba moduly doplnené o funkcionalitu *kd-stromu* s následným vyhľadávaním blízkych bodov boli testované porovnaním výsledkov s výstupmi pôvodných verzií modulov, overených v predchádzajúcich prácach (Stopková, 2013a, b). Výsledky sú zosumarizované v ukážkach Tab. 1 a Tab. 2.

Pri prvom testovaní modulu *v.nn_spatial_stat* boli použité syntetické dáta (Stopková, 2013b). Tab. 1 obsahuje výsledky testovania na vzorke náhodne vygenerovaných bodov v trojrozmernom priestore pomocou modulu *v.random* (McCauley & Landa, 2010).

Tab. 1 Porovnanie 3D analýzy najbližšieho suseda dvomi verziami modulu *v.nn_spatial_stat* (2 000 bodov)

Počet bodov	2000			
	bez *kd-stromu*	s *kd-stromom*	rozdiel	
r_A [m]	346.071782	346.071782	0.000	mm
r_E [m]	323.531486	323.531486	0.000	mm
R = rA / rE	1.069670	1.069670	0.000000	
Testovacia štatistika	0.191691	0.191691	0.000000	
Obj. najmenšieho ohraničujúceho kvádra [m^3]	398423031180.489	398423031180.489	0.000000	m^3
Čas trvania výpočtu [s]	0.093	0.053	0.040	s

Vzorka 2000 bodov je pre testovanie trvania výpočtu nedostatočná, keďže časové rozdiely pri výpočtoch s použitím priestorového indexu a bez neho sú pri takýchto súboroch dát zanedbateľné. Preto pri ďalšom testovaní boli použité body gravimetrického mapovania (Kubeš et al., 2001) s počtom 211 631 (Tab. 2).

Modul bol testovaný na počítači s parametrami:

Procesor: Intel(R) Core(TM) i5 CPU, 2.80 GHz, 4 jadrá
Operačná pamäť: 4.00 GB
Operačný systém: Ubuntu 12.04

Tab. 2 Porovnanie 3D analýzy najbližšieho suseda dvomi verziami modulu *v.nn_spatial_stat* (211631 bodov)

Počet bodov	211631			
	bez *kd-stromu*	s *kd-stromom*	rozdiel	
r_A [m]	351.333082	351.333120	-0.038	mm
r_E [m]	516.156602	516.156602	0.000	mm
$R = r_A / r_E$	0.680671	0.680672	-0.000001	
Testovacia štatistika	-0.878612	-0.878612	0.000000	
Obj. najmenšieho ohr. kvádra [m^3]	171193993672530.281	171193993672530.281	0.000000	m^3
Čas trvania výpočtu [s]	962.408	388.749	573.659	s

Modul *v.kriging* bol testovaný interpoláciou matematicky vygenerovaných hodnôt normálneho tiažového zrýchlenia γ na vybraných bodoch gravimetrického mapovania (Kubeš et al., 2001). Výsledky testovania sumarizuje Tab. 3.

Tab. 3 Porovnanie 3D interpolácie (*bežný kriging*) dvomi verziami modulu

	γ s kd-stromom	γ bez kd-stromu	rozdiel
Počet hodnôt	20163	20163	0
Minimum [mGal]	980604.466076	980604.408343	0.05773
Maximum [mGal]	980799.764585	980799.791156	-0.02657
Stredná hodnota [mGal]	980702.115331	980702.099750	0.01558
Disperzia [$mGal^2$]	3355.604297	3359.751172	-4.14688
Štandardná odchýlka [mGal]	57.927578	57.963361	-0.03578
Čas trvania výpočtu [s]	4.145	4.261	-0.116

V charakteristikách presnosti interpolácie hodnôt normálneho tiažového zrýchlenia, ktorá bola prevedená s použitím *kd-stromu* a bez neho, sú badateľné numerické rozdiely. Tieto boli pravdepodobne spôsobené tým, že experimentálny variogram $\gamma(h, \Delta z)$ bol s použitím priestorového indexovania určený na základe množiny bodov, ktorá sa mierne líšila od výberu v pôvodnom výpočte, čo bude predmetom ďalšieho skúmania. Tým sa zmenili aj koeficienty teoretického variogramu (Tab. 4). Premenná *h* predstavuje horizontálnu vzdialenosť bodov a Δz je ich prevýšenie.

Tab. 4 Koeficienty teoretického variogramu s použitím *kd-stromu* a bez neho

$\gamma(h,\Delta z) = a \cdot h^2 + b \cdot \Delta z^2 + c$	a	b	c
bez *kd-stromu*	0.043714	0.000020	-387.254400
s *kd-stromom*	0.043301	0.000026	-499.359364

MOŽNOSTI ĎALŠIEHO VÝVOJA

V budúcej práci bude prioritou ďalšie testovanie modulu *v.nn_spatial_stat* a hľadanie spôsobov, ako zefektívniť výpočet objemu najmenšieho ohraničujúceho kvádra vzorky, ktorý je potrebný pri určení hustoty bodov v danej oblasti.

V prípade modulu *v.kriging* bude potrebné optimalizovať výpočet variogramu rozdelením záujmového územia na menšie celky. Táto úprava je potrebná nato, aby bolo možné aplikovať zúženie výberu bodov aj pri samotnej interpolácii, ktorá zatiaľ prebieha bez indexovania. Pri pokuse aplikovať priestorový index boli totiž výsledky zaťažené neprijateľne veľkou numerickou chybou a nezodpovedali reálnym očakávaniam. Ku kvalite testovania určite prispeje aj kontrola interpolovaných hodnôt pomocou krížovej validácie (Wackernagel, 2003), ktorá v moduli zatiaľ nie je implementovaná.

ZÁVER

V prípade modulu *v.nn_spatial_stat*, ktorý je založený na analýze najbližšieho suseda, sa použitie priestorového indexu *kd-tree* ukázalo byť jednoznačne priaznivým krokom, ktorý viedol k výraznému zefektívneniu modulu. Zároveň však bude potrebné urýchliť výpočet objemu najmenšieho ohraničujúceho kvádra.

Pri použití *kd-tree* v moduli *v.kriging* sa neprejavilo žiadne výrazné urýchlenie výpočtu. Príčina sa však môže skrývať v skutočnosti, že zatiaľ bol pomocou priestorového indexu optimalizovaný len výpočet variogramu. Numerické výsledky po aplikácii priestorového indexu do výpočtovo najnáročnejšej časti interpolácie (samotný odhad neznámych hodnôt na základe vzdialeností vstupných bodov od výpočtového s použitím teoretického variogramu) boli natoľko vzdialené očakávaniam (hodnote skutočného zrýchlenia, cca. 9.80 $m \cdot s^{-2}$), že optimalizácia odhadu interpolovaných hodnôt bude predmetom ďalšej práce. Hlavný dôraz bude pritom kladený na kritériá vymedzenia relevantného okolia pre výpočet variogramu a následnú interpoláciu bodov.

LITERATÚRA

Bentley, J. L. (1975) Multidimensional binary search trees used for associative searching. In: Communications of the ACM, Vol. 18. Numb. 9, pp. 509-517. Dostupné na: http://delivery.acm.org/10.1145/370000/361007/p509-bentley.pdf?ip=147.175.19.167&id=361007&acc=ACTIVE%20SERVICE&key=C2716FEBFA981EF18403C4491FB72FD9EE41502081180C7C&CFID=374232331&CFTOKEN=67470855&__acm__=1383032253_2c41de508d7636dfc7665a2f0d137eaa

GRASS Development Team (2013) Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Foundation Project. Dostupné na: http://grass.osgeo.org

Kubeš, P., Grand, T., Šefara, J., Pašteka, R., Bielik, M., Daniel, S. (2001) Atlas geofyzikálnych máp a profilov : Záverečná správa geologickej úlohy, Časť gravimetria. Štátny geologický ústav Dionýza Štúra, Bratislava.

McCAULEY, J. D., LANDA, M. (2010). *v.random* [Počítačový program]. Dostupné na: http://trac.osgeo.org/grass/wiki/DownloadSource#GRASS7

O'Leary, G. KdTree Search [online]. Tutorial. Dostupné na: http://pointclouds.org/documentation/tutorials/index.php

Open Perception, Inc. (2012) *Point Cloud Library*. Dostupné na: http://www.pointclouds.org/

PCL Users mailing list (2012) c++ float type and precision issues. Dostupné na: http://www.pcl-users.org/c-float-type-and-precision-issues-td3959221.html

PostGIS (2013) Using PostGIS: Data Management and Queries. In PostGIS 1.5.3 Manual. Dostupné na: http://postgis.net/docs/using_postgis_dbmanagement.html#id471477

Raper, J. F., 1992. Key 3D modelling concepts for geoscientific analysis. In TURNER, A. K. Three-Dimmensional Modeling with Geoscientific Information Systems. Dordrecht : Kluwer Academic Publishers, 1992. ISBN 0-7923-1550-2. S. 215 – 232.

Stopková, E. (2013a) Interpolácia dát v 3D priestore rozšírenou metódou kriging. In: GIS Ostrava 2013, Ostrava, 21.-23. január 2013. Dostupné na: http://gis.vsb.cz/GIS_Ostrava/GIS_Ova_2013/sbornik/papers/gis20135076d889cc867.pdf

Stopková, E. (2013b) Extension of mathematical background for Nearest Neighbour Analysis in three-dimensional space (abstrakt). In: Geoinformatics 2013. Praha, 11.-12. júl 2013. Dostupné na: http://geoinformatics.fsv.cvut.cz/gwiki/Nearest_Neighbour_Analysis_in_three-dimensional_space

Wackernagel, H. (2003) Multivariate Geostatistics: An Introduction with Applications. Heidelberg : Springer-Verlag. 2003. 387 s. ISBN 3-540-44142-5.