VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STAVEBNÍ
ÚSTAV GEODÉZIE

FACULTY OF CIVIL ENGINEERING
INSTITUTE OF GEODESY

# METODIKA ŘEŠENÍ MASIVNÍCH ÚLOH V GIS

METHODOLOGY FOR THE SOLUTION OF MASSIVE TASKS IN GIS

TEZE DISERTAČNÍ PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE — Ing. IRENA OPATŘILOVÁ
AUTHOR

VEDOUCÍ PRÁCE — doc. Ing. DALIBOR BARTONĚK, CSc.
SUPERVISOR

BRNO 2015

**Abstrakt**

Disertační práce se zabývá problematikou řešení masivních úloh v GIS. Tyto úlohy zpracovávají geografická data velkých objemů a různých formátů. Práce popisuje teoretický rozbor složitosti úloh a možnosti optimalizace dílčích procesů, které vedou k přijatelnému řešení. Zamýšlí se nad možností využití paralelismu v GIS, čímž lze zrychlit zpracování velkého objemu geodat. Navrhuje také způsob optimalizace procesů prostřednictvím algoritmu, který stanoví počet nutných prostředků k úspěšnému vyřešení úlohy v zadaném čase a k přiřazení procesů těmto prostředkům. Dále je zde navržen algoritmus pro optimalizaci při přípravě dat rozsáhlých GIS projektů. Algoritmy byly ověřeny v rámci výzkumného projektu, jehož cílem byla analýza povrchů terénu nad plynovody na území ČR vyjma dvou krajů. Hlavní metodou analýzy byla klasifikace obrazu ortofota, která byla dále zpřesněná filtrací z vrstev ZABAGED. Proto se práce zabývá i možnostmi zpřesnění výsledků klasifikace obrazu s využitím nástrojů GIS a stanovením chybovosti výsledků analýzy. Výstupy analýzy jsou nyní využívány pro strategické plánování údržby a rozvoje plynárenských zařízení v ČR. Výsledky práce mají obecný význam pro řešení stejné třídy úloh v GIS.

**Abstract**

This doctoral thesis deals with the issue of solving massive tasks in GIS. These tasks process large volumes of geographic data with different formats. The thesis describes a theoretical analysis of the complexity of tasks and the possibilities to optimize sub-processes which lead to an acceptable solution. It considers the possibility of using parallelism in GIS, which leads to an acceleration in the processing of large volumes of geographic data. It also proposes a method for the optimization of processes through an algorithm which determines the number of means necessary for the successful solution of a task at a specified time and assigns processes to these means. Additionally, there is a proposed algorithm for the optimization of the preparation of data for extensive GIS projects. The algorithms have been validated by the results of a research project, the aim of which was to analyse the terrain surface above a gas line in the Czech Republic. The primary method of analysis was the classification of an orthophoto image, which was further refined through filtration using the ZABAGED layers. Therefore, the thesis deals with the possibility of improving the results of image classification using GIS instruments as well as dealing with the determination of the error rate in analysis results. The results of the analysis are now used for the strategic planning of maintenance and the development of gas facilities in the Czech Republic. The results of the work have general importance regarding the performance of other operations of the same class in GIS.

**Klíčová slova:** Geografický informační systém, optimalizace, masivní úlohy, klasifikace obrazu, prostorové analýzy

**Keywords:** Geographic information system, optimization, massive tasks, image classification, spatial analysis

# Obsah

# 1 ÚVOD

Tematicky je oblast disertační práce zaměřena na problematiku řešení masivních úloh v geografických informačních systémech (GIS). Úlohy v GIS se vyznačují tím, že převážná část operací probíhá nad geodatabází obsahující územně vázaná data, tzv. geodata. Právě rozsah modelovaného území a typ vstupních dat výrazně odlišují procesy GIS oproti procesům, které probíhají v běžných informačních systémech. V GIS je potřeba v některých případech pracovat s velkým objemem prostorových dat, přičemž se tento objem díky dynamicky se rozvíjející dostupnosti dalších užitečných datových sad a zdokonalováním metod jejich pořizování neustále zvětšuje. Rozsáhlé a složité úlohy vedou k masivním výpočtům, které jsou velmi náročné na přípravu dat, strojový čas, kapacitu paměti a výkonnost použité výpočetní techniky. Disertační práce se zabývá teoretickým rozborem složitosti úloh a možnostmi optimalizace dílčích procesů, které vedou k přijatelnému řešení.

Z hlediska značných datových objemů je téměř vyloučeno pracovat s celkovým objemem dat najednou vzhledem ke spolehlivosti, kapacitě a výkonnosti hardware a software. Proto je nutné vstupní datové sady předzpracovat a vhodně je rozčlenit na menší části. Problém spočívá v nalezení varianty takového rozčlenění, aby poměr strojočasu potřebného na zpracování úloh a času režijního (přípravné a pomocné operace vykonávané zpravidla lidským faktorem) byl co největší a celý proces v GIS probíhal efektivně, v nejkratším čase a se zachováním požadované kvality výstupů. Proto disertační práce řeší metodiku předzpracování datových sad v rámci rozsáhlých GIS projektů s cílem dosáhnout optimálního zpracování projektu v rámci dané informační infrastruktury (hardware, software) s uvážením lidského faktoru. V práci je navržen vhodný způsob optimalizace rozdělení dat do menších částí.

V disertaci není řešena jen optimalizace přípravy dat, ale práce se zabývá i obecnou optimalizací zpracování velkého objemu geografických dat. Podstatou metody je hierarchický rozklad množiny procesů na elementární procesy a přiřazení prostředků těmto procesům. Dále je v práci analyzována časová náročnost masivních úloh v závislosti na typu vstupních dat a měřítkovém koeficientu, který je definován jako poměr rozsahu modelované oblasti k rozměru nejmenšího detailu ve vstupních datech.

Teoretické předpoklady metodiky řešení složitých úloh v GIS byly ověřeny na výzkumném projektu datové analýzy uložení plynárenských zařízení pod určitými typy terénních povrchů na území ČR. Projekt byl řešen pro společnost RWE za účelem stanovení reprodukčních hodnot plynárenských zařízení (plynovodů) a ocenění nákladů, které by bylo nutné vynaložit na vybudování nových sítí. Výsledky analýzy jsou nyní využívány pro strategické plánování údržby a rozvoje plynárenských zařízení v ČR. Řešení tohoto projektu lze považovat za masivní úlohu, která byla zpracována v prostředí GIS a jehož hlavní metodou byla klasifikace obrazu ortofota s následnou filtrací. Proto se část disertační práce zabývá možnostmi zpřesnění klasifikace obrazu a zvýšení technické vytěžitelnosti informací z rastrového obrazu prostřednictvím nástrojů GIS. Dále je v práci diskutována kvalita dosažených výsledků a popsána problematika stanovení chybovosti při automatizované klasifikaci obrazu se zpřesněním výsledků pomocí filtrace.

# 2 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

V tabulce 1 je výčet a hodnocení nejčastěji používaných metod týkající se problematiky disertační práce. Konkrétně jsou zde hodnoceny metody pro dvě základní oblasti, a to obecné řešení složitých úloh a optimalizace přípravy dat rozsáhlých projektů. Hodnocení je zde uvedeno ze tří hledisek – složitost úlohy, její účinnost a vhodnost pro GIS. Složitost úlohy je hodnocena slovně, ostatní dva parametry jsou porovnány číselným stupněm 1 až 3, přičemž čím vyšší číselná hodnota, tím je metoda účinnější/vhodnější pro GIS. Z tabulky vyplývají následující závěry k jednotlivým metodám, u kterých je uvedena související literatura.

| **Oblast problému** | **Metoda (princip) řešení** | **Kritéria hodnocení řešení** | | |
|---|---|---|---|---|
| | | **Složitost metody** | **Účinnost, výkonnost** | **Vhodnost pro GIS** |
| **Obecné řešení složitých úloh** | Využití paralelismu | Velmi vysoká | 3 | 3 |
| | Zjednodušení rozdělením | Vysoká | 1 | 3 |
| | Speciální algoritmy | Střední | 1 | 2 |
| | Organizace procesů | Velmi vysoká | 2 | 2 |
| **Optimalizace přípravy dat rozsáhlých projektů** | Heuristika | Velmi vysoká | 2 | 3 |
| | Stromové struktury (hierarchie jevů nebo GO) | Střední | 1 | 1 |
| | Shluková analýza | Střední | 2 | 3 |
| | Hrubé množiny | Střední | 1 | 1 |
| | Dolování dat | Vysoká | 1 | 3 |
| | Speciální algoritmy | Střední | 3 | 2 |

Tab. 1 – Hodnocení metod řešení

Obecné řešení složitých úloh

- *Využití paralelismu* – Cílem metody je rozdělit vstupní úlohu na nezávislé procesy. Toto rozdělení je nejobtížnějším článkem metody, protože jde o NP-úplný problém, který dosud nebyl vyřešen efektivním algoritmem. Pokud se úlohu podaří rozdělit, je výkonnost metody vysoká. V územně vázaných systémech je tohle rozdělení jednodušší než v obecných systémech, protože úloha se dá rozdělit podle územního členění (pokud se areály nepřekrývají, tvoří tak nezávislé oblasti pro zpracování). [1]
- *Zjednodušení rozdělením* – Složité entity je vhodné rozdělit (vyhlazením, generalizací) na menší z důvodu jednoduššího zpracování úlohy vzhledem k možnostem současného HW a SW. Složitost úlohy pak vychází z míry zjednodušení. [2]
- *Speciální algoritmy* – Snahou je rozdělit obrovské množství vstupních dat tak, aby zpracovatelský segment byl na nejrychlejším médiu (paměť počítače, SSD rychlý disk). Ze zdrojových dat se tak vybírají jen data relevantní, přičemž dochází ke snížení objemu dat. [3]
- *Organizace procesů* – Snahou je vhodně uspořádat dílčí procesy tak, aby celková doba zpracování kompletní úlohy byla co nejkratší. Algoritmicky jde o velmi obtížný problém, jehož podstatou je přidělování prostředků procesům. [4]

Optimalizace přípravy dat rozsáhlých projektů

- *Heuristika* – Metoda vybírá z nadbytečného množství dat jen ta data, která budou postačovat ke kvalitnímu výsledku zpracování. Jde o kombinatorickou úlohu, při níž dochází k prohledávání velkého stavového prostoru. Jednou z oblastí využití je např. laserové skenovaní a zpracování mračen bodů. [5]
- *Stromové struktury* – Vstupní data jsou hierarchicky rozdělena, např. podle dat reprezentující geografické objekty (GO). Ze vstupních dat lze vybírat podle stromové struktury v závislosti na úrovni detailu, tzn., čím je vyšší detail, tím je úloha složitější a výkonnost úlohy nižší. [6]
- *Shluková analýza* – Využití této metody je např. pro územní analýzy, kde je cílem zjednodušit vyhodnocení složitých jevů nebo geografických objektů (sociologické jevy, výskyt epidemií aj.). Slouží tak nejen ke zpřehlednění výstupů, ale i k objasnění příčin jevů. [7]
- *Hrubé množiny* – Souvisí s přesností a aproximací kvality výstupů. Podle toho jsou pak vybírány kvalitní vstupy. Řeší se zde stanovení kritéria, zda úlohu znovu zpracovat, anebo je výstup postačující. [8]
- *Dolování dat* – Při této metodě dochází k doplnění chybějící informace na základě vstupních dat. V mnoha případech je zde využití predikce, a tudíž výsledek má pravděpodobnostní charakter. Spolehlivost metody je nižší, protože nová doplněná informace je získaná nepřímo a účinnost metody je ovlivněná kvalitou predikce. [9]
- *Speciální algoritmy* – V tomto případě jde o procesy, které pro danou (dílčí) úlohu vyberou veškerá potřebná data pro další zpracování. [10]

Z hodnot v tabulce 1 a ze stručného slovního hodnocení metod vyplývají následující závěry. Mezi nejúčinnější metodu pro řešení složitých úloh v GIS bude patřit paralelismus, jenž spočívá v rozdělení úlohy do pokud možno nezávislých procesů, které lze zpracovávat současně na několika prostředcích (počítačích). K tomu bude nejprve zapotřebí stanovit optimální dílčí zpracovatelskou jednotku z hlediska rozsahu území a potažmo i objemu vstupních dat. K jejímu určení bude vhodné použít metody experimentu nebo heuristiky. Dále bude zapotřebí navrhnout a využít speciální algoritmy pro optimalizaci přípravy dat tak, aby v následujících krocích zpracování úlohy byla využívána pouze vybraná potřebná data.

V oblasti obecného řešení složitých úloh a optimalizace příravy dat rozsáhlých projektů jsou v mnoha případech aplikovány metody, které nevykazují optimální poměr cena/výkon/doba řešení. Ve své práci se pokusím tuto situaci zlepšit návrhem nového řešení.

**Cílem disertační práce** je navrhnout vhodné nebo inovovat stávající metody pro řešení výše uvedených problémů a ověřit jejich efektivitu v konkrétní aplikaci v GIS. Verifikace metodiky je realizována na projektu pro společnost RWE, jehož cílem bylo zpracovat analýzu povrchů nad průběhem plynovodů na celém území ČR vyjma dvou krajů. Podstatou řešení projektu bylo navrhnout novou metodu na bázi klasifikace obrazu s velmi vysokým rozlišením na rozsáhlém území s cílem dosáhnout co nejlepšího výsledku. Charakteristikou tohoto procesu je vytěžení informací a převod rastru na vektorový tvar obrazu, který umožňuje přesnější výsledky prováděných operací.

# 3 TEORETICKÝ ZÁKLAD

Tato kapitola je kvůli přehlednosti rozdělena do pěti částí. První část popisuje obecný teoretický základ u řešení masivních úloh v GIS. Druhá část se zabývá předzpracováním geografických dat. Třetí část kapitoly řeší algoritmus pro optimalizaci velkého množství geodat. Čtvrtá část zkoumá problematiku zpřesnění klasifikace obrazu a pátá část kapitoly obsahuje návrh na stanovení chybovosti výsledků klasifikace.

## *3.1 Řešení úloh v GIS*

Řešení úloh v GIS je možné matematicky popsat touto formulí:

$$R = (I, S, Q, Y, \delta, \varphi, \psi), \tag{1}$$

kde

$$I \subseteq (A_1 \times A_2 \times \cdots \times A_l), \tag{2}$$

$I$ … množina vstupních datových sad, $A_i$ jsou dílčí datové sady
$l$ … počet dílčích datových sad
$S$ … množina vnitřních stavů GIS
$Q$ … množina přípustných výsledků (výsledných vrstev), které vyhovují požadované kvalitě
$Y$ … množina výstupů z GIS
$\delta$ … koeficient územní podrobnosti daný vztahem:

$$\delta = \frac{P_{tot}}{P_{det}}, \tag{3}$$

kde $P_{tot}$ … celková plocha modelovaného území, které se řeší v rámci dané úlohy
$P_{det}$ … plocha detailu, který je popsán ve vstupních vrstvách $I$

$\varphi$ … zobrazení (vstupní predikát), které má charakter přípravy vstupních dat:

$$\varphi : I \rightarrow S, \tag{4}$$

$\psi$ … zobrazení (výstupní predikát):

$$\psi : S \cap Q \rightarrow Y. \tag{5}$$

Operace $\gamma$: $S \cap Q$ představuje proces kvality výstupu (kartografického díla).

Časová náročnost dané úlohy v GIS z teoretického hlediska je dána vztahem:

$$T = f(\delta, \varphi, \psi, l). \tag{6}$$

V této rovnici $\delta$ představuje měřítko zobrazení z reality do digitální geodatabáze, $\varphi$ reprezentuje složitost přípravy vstupních dat, tj. transformaci vstupních vrstev do vnitřních stavů GIS, $\psi$ má charakter vnitřních procesů v GIS, tj. nástrojů pro územní operace, analýzy, editace, kontrolu kvality $\gamma$ apod. a $l$ je počet typů vstupních vrstev – viz vztah (2).

Pokusme se nyní o rozbor časové náročnosti úloh v GIS podle vztahu (6). Parametry v této rovnici můžeme rozdělit podle míry ovlivnění ze strany uživatele do 2 skupin:

1. vstupní $\varphi$ a výstupní $\psi$ predikáty jsou dány charakterem úlohy a uživatel je může ovlivnit jen z malé části
2. koeficienty $\delta$ a $l$ naopak můžeme ovlivnit dosti značně. Parametr $n$ je dán úspěšností výstupu a jeho kvalitou, na koeficientu $\delta$ závisí nejen doba zpracování, ale i spolehlivost vnitřních procesů $\psi$.

Z hlediska dob trvání jednotlivých procesů můžeme rovnici (6) přepsat také do tohoto tvaru:

$$T = T_{Ia} + T_{Im} + T_{Oa} + T_{Om}, \tag{7}$$

kde $T_{Ia}$ … je čas přípravy dat automatizovaně
$T_{Im}$ … je doba přípravy dat, která probíhá poloautomaticky nebo ručně (vstupní kontroly)
$T_{Oa}$ … je čas zpracování dat a výpočtu výsledků automatizovaně
$T_{Om}$ … je čas zpracování dat a výpočtu výsledků poloautomaticky nebo ručně (výstupní kontroly).

V rámci experimentálních výsledků disertační práce se podařilo empiricky stanovit hodnotu funkce v rovnici (7).

## *3.2 Příprava dat*

Hlavním cílem této fáze je upravit vstupní data tak, aby byla zajištěna řešitelnost a spolehlivost všech procesů, které budou v GIS probíhat. Pokud jsou vstupní data příliš rozsáhlá (objemná), není možné výše uvedené požadavky zajistit. V tomto případě je zapotřebí datové sady rozdělit na menší části podle vhodně stanovených kritérií.

V práci byl navržen hierarchický rozklad množiny vstupních dat podle:

1) administrativního uspořádání území v první hierarchické úrovni
2) věcného uspořádání, tj. podle existence geografických objektů v území ve druhé hierarchické úrovni.

Na základě koeficientu $\delta$, definovaného v předchozí podkapitole, bude rozhodnuto o tom, podle jakých dílčích územních celků v rámci administrativního uspořádání ČR (tj. rozklad v 1. hierarchické úrovni) budou rozděleny vstupní datové sady projektu.

## *3.3 Algoritmus pro optimalizaci zpracování velkého množství geodat*

Rozdělením vstupních dat do menších celků proces přípravy dat zcela nekončí. V další fázi musíme zajistit, aby celý projekt byl vyřešen v zadaném čase (termínu). Vhodnou metodou je optimalizace všech procesů, které budou v GIS následovat. Z hlediska efektivity je užitečné využít možného paralelismu v GIS a přidělit dílčí datové sady dostupným prostředkům (počítačům).

Úlohou je sestavit optimální rozvrh přiřazení procesů prostředkům tak, aby celková doba řešení $T_{tot}$ byla minimální. Účelová funkce, která se bude optimalizovat, je dána vztahem:

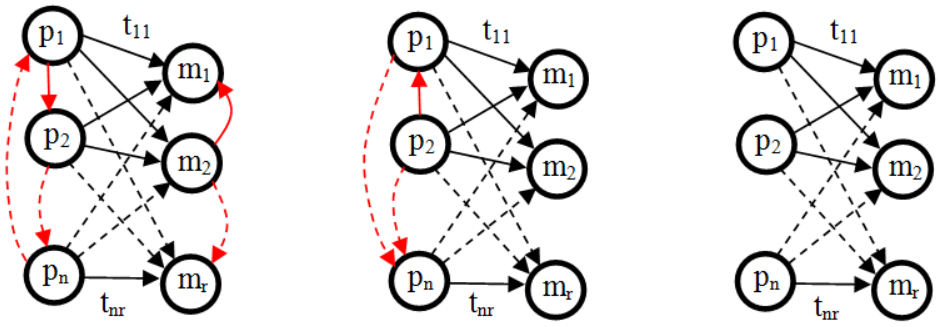
$$\sum_{i=1}^{n} t_{ij} \leq T_{tot} \,, \tag{8}$$

kde $t_{ij}$ je doba trvání procesu $p_i$ na prostředku $m_j \in M$, $j = 1, 2, \ldots, r$.

$T_{tot}$ je celková doba řešení úlohy, která se skládá ze všech procesů $p_i \in P$, $i = 1, 2, \ldots, n$.

Obtížnost optimalizace je dána stupněm závislosti mezi procesy a prostředky. Prakticky mohou nastat tyto varianty závislosti, které jsou znázorněné v grafu doby zpracování procesů prostředky na obr. 1 (hrany, reprezentující závislost, jsou v grafu označeny červeně):

a) procesy i prostředky jsou navzájem závislé (obr. 1a)
b) procesy jsou navzájem závislé, prostředky jsou nezávislé (obr. 1b)
c) procesy i prostředky jsou navzájem nezávislé (obr. 1c).




Obr. 1 – Graf doby zpracování procesů prostředky a) procesy i prostředky jsou navzájem závislé, b) procesy jsou navzájem závislé, prostředky jsou nezávislé, c) procesy i prostředky jsou navzájem nezávislé (časy jsou v grafu pro přehlednost uvedeny jen u krajních hran)

Je zřejmé, že nejjednodušší varianta je ad c), nejsložitější je varianta ad a). Tento případ lze zjednodušit např. tím, že pokud jsou procesy nebo prostředky navzájem závislé, sloučí se všechny závislé procesy nebo prostředky do skupin a na tyto skupiny se aplikuje optimalizační algoritmus. V rámci závislých skupin se optimalizace řeší sekvenčně (postupně). Algoritmus byl navržen především pro nezávislé skupiny i procesy, viz obr. 1 c). Tento případ se v systémech GIS vyskytuje častěji než v jiných oborech.

Na základě grafu na obr. 1 sestavíme mapu procesů $M$, jejíž řádky odpovídají procesům, sloupce prostředkům a prvky matice $m_{ij}$ reprezentují dobu zpracování. Ukázka obecné mapy procesů je v tab. 2.

| ***Procesy/prostředky*** | $m_1$ | $m_2$ | ... | $m_r$ |
|---|---|---|---|---|
| $p_1$ | $t_{11}$ | $t_{12}$ | ... | $t_{1r}$ |
| $p_2$ | ... | ... | ... | |
| ... | ... | ... | ... | |
| $p_n$ | $t_{n1}$ | $t_{n2}$ | ... | $t_{nr}$ |

Tab. 2 – Mapa procesů

**Popis optimalizačního algoritmu**

Nejdříve definujeme procesy, které se v systému vyskytují, a určíme prostředky, na nichž je možné dané procesy zpracovávat. Potom je zapotřebí stanovit doby zpracování $t_{ij}$ *i*-tého procesu $p_i$ zpracovávaném na *j*-tém prostředku $m_j$. Tyto doby se zjistí buď kvalifikovaným odhadem, nebo experimentálně.

Dále se zjišťuje vzájemná závislost mezi procesy a prostředky. Tento vztah lze zjistit např. orientovaným grafem (viz obr. 1). Pokud mezi procesy nebo prostředky existují závislosti, lze tento problém vyřešit sloučením závislých uzlů tak, že mezi sloučenými a ostatními uzly už nejsou žádné závislosti. Pokud jsou procesy a prostředky navzájem nezávislé, můžeme sestavit mapu procesů a prostředků. Záhlaví řádků této mapy obsahuje jednotlivé procesy $p_i$, záhlaví sloupců prostředky $m_j$, do buněk pak zapíšeme doby trvání procesu $t_{ij}$ na daném prostředku (viz tab. 2). V této mapě uspořádáme řádky podle součtu časů sestupně a sloupce podle součtů časů ve sloupcích vzestupně tak, aby platilo:

$$\sum_{j=1}^{r} t_{1j} \geq \sum_{j=1}^{r} t_{2j} \geq \ldots \geq \sum_{j=1}^{r} t_{nj}, \sum_{i=1}^{n} t_{i1} \leq \sum_{i=1}^{n} t_{i2} \leq \ldots \leq \sum_{i=1}^{n} t_{ir} \ . \qquad (9)$$

Prakticky to znamená, že časově náročnější proces má v mapě nižší index řádku než proces, který trvá kratší dobu. Výkonnější prostředek pak má nižší index sloupce než prostředek méně výkonný. Smyslem tohoto uspořádání je, aby nejnáročnější proces byl řešen pokud možno na nejvýkonnějším prostředku. Na tomto principu je založen optimalizační algoritmus.

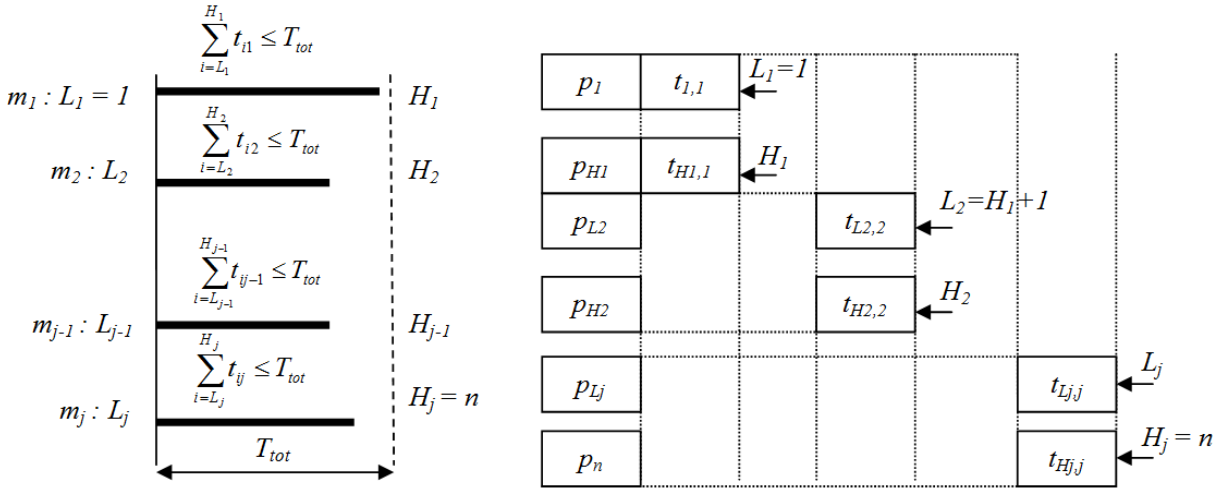
Po sestavení mapy ji musíme otestovat, zda je korektní. Může nastat případ, kdy některý proces $p_i$ nepůjde řešit ani na jednom z prostředků, které jsou v mapě uvedeny. V tomto případě se do příslušného pole zapíše hodnota $t_{ij} = \infty$, tzv. strojového nekonečna, což je maximální číslo, které lze v počítači zobrazit. Před vlastním algoritmem musíme zkontrolovat, aby každý řádek obsahoval alespoň jeden údaj, který je menší než toto strojové nekonečno. Pokud některý řádek obsahuje samá strojová nekonečna, musíme přidat do mapy takový prostředek, který daný proces vyřeší v konečném čase, tj. platí: $t_{ij} < \infty$.

Jádrem optimalizační procedury je dvojnásobný cyklus, kde vnější cyklus probíhá po řádcích (index *i*) a vnitřní cyklus po sloupcích (index *j*). V každém *j*-tém sloupci se stanoví dolní mez $L_j$ a horní mez $H_j$. V těchto mezích se po řádcích v daném sloupci sčítají hodnoty $t_{ij}$ *i*-tého procesu $p_i$ a při každém součtu se testuje, zda součet mezi $L_j$ a $H_j$ nepřesahuje hodnotu celkové doby $T_{tot}$ řešení úlohy. Pokud ne, pokračuje se v součtu dob $t_{ij}$ v daném sloupci. Pokud ano, přejde se na další sloupec *(j = j + 1)*, tj. musí se použít další prostředek $m_j$.

Nejdříve se testuje, zda index *j* nepřekročil hodnotu $m_r$, tj. počet prostředků, které máme k dispozici. Pokud ano, pak musíme přidat nový prostředek a zařadit jej do mapy procesů. Pak se spustí dvojitý cyklus znovu. Pokud ne, tak se nastaví dolní a horní mez $L_j$ a $H_j$ řádku v dalším *j*-tém sloupci a postupně se sčítají hodnoty $t_{ij}$ mezi těmito mezemi, dokud součet nepřesáhne hodnotu $T_{tot}$. Dojdeme-li na konec řádků, tj. $L_j = n$, algoritmus končí a aktuální index *j* (číslo sloupce v matici procesů a prostředků) udává optimální počet prostředků, který je postačující pro vyřešení celé úlohy.

Protože předpokládáme, že procesy jsou na sobě navzájem nezávislé, můžeme je spouštět na určených prostředcích paralelně. Situace je znázorněna na obr. 2. Časová náročnost

algoritmu je v nejhorším případě (tj. kdy index sloupce *j* dosáhne hodnoty *r*) úměrná součinu *n x r*, tj. *O(nr)*. Algoritmus byl ověřen v rámci experimentálních výsledků disertační práce.




Obr. 2 – Paralelní zpracování procesů na prostředcích

## 3.4 *Zpřesnění výsledků klasifikace obrazu*

V této části je navržena metoda pro zlepšení kvality výsledků klasifikace a určení klíčových faktorů, které tuto kvalitu významně ovlivňují. K posouzení kvality výsledků klasifikace ortofota využijeme teorii hrubých množin. Příslušnost prvku k hrubé množině je zprostředkována pomocí speciální relace ekvivalence, tzv. relace nerozlišitelnosti. Původní definice hrubých množin je předložena v publikaci [11].

Nechť množina *X* reprezentuje klasifikační třídu ortofoto. Pak každý pixel ortofota musíme s určitou jistotou zařadit do nějaké třídy (neexistují nezařazené pixely). Pomocí hrubé množiny můžeme definovat přibližnou přesnost $\alpha_{RE}(X)$, s jakou nalezená aproximace reprezentuje vybranou množinu *X*:

$$\alpha_{RE}(X) = \frac{card(Posi_{RE}(X))}{card(Poss_{RE}(X))}, \tag{10}$$

kde $Posi_{RE}(X)$ je dolní aproximace, tj. popis objektů, které s jistotou patří do podmnožiny *X* a $Poss_{RE}(X)$ je horní aproximace, tj. množina prvků, které mohou patřit do podmnožiny *X*.

Aproximace ve vztahu (10) je v našem případě měřítkem kvality klasifikace, tj. do jaké míry odpovídá zařazení daného pixelu do příslušné třídy. Tato přesnost je ovlivněna především těmito faktory: kvalitou vstupních dat (v našem případě ortofota), zvolenou metodou automatické klasifikace, kvalitou trénovací množiny (pokud jde o řízenou klasifikaci) a doplňkovými zpřesňujícími operacemi (filtrace, kontrola apod.).

## 3.5 *Stanovení chybovosti analýzy*

Tato podkapitola řeší vyhodnocení přesnosti výsledků klasifikace obrazu ortofota. V GIS je pojem přesnosti posuzován podle tří hlavních hledisek:

1. polohová přesnost geografických objektů
2. časová přesnost (souvisí s aktualizací dat)
3. tematická (atributová) přesnost.

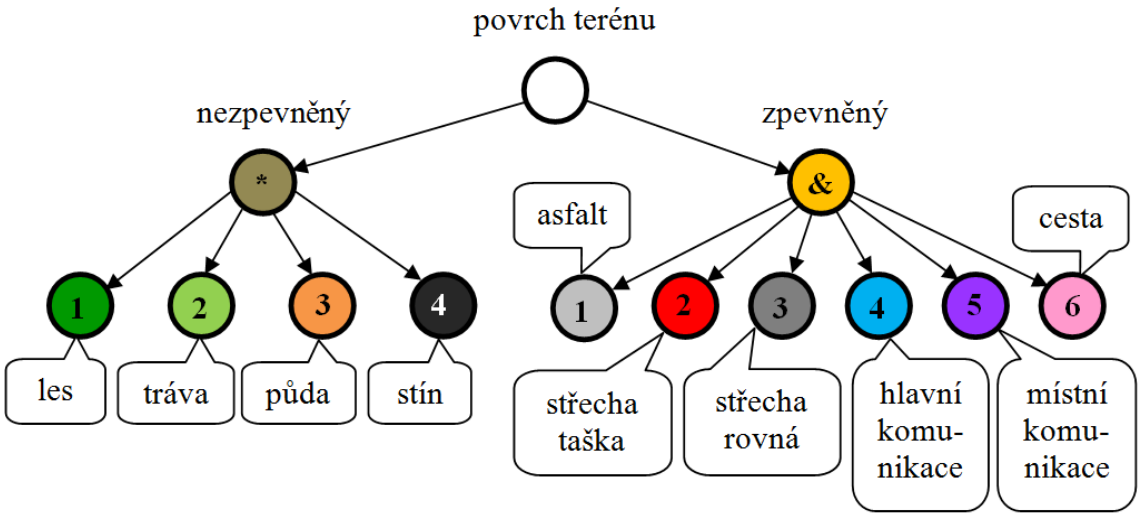
Polohová a časová přesnost se modeluje pomocí metrických prostorů [12]. Tyto přesnosti jsou dány použitím nejaktuálnějších dostupných dat od poskytovatele. V našem případě bude podrobněji řešena tematická přesnost.

Každé klasifikační třídě $X_i$ můžeme přiřadit název podle druhu povrchu, který reprezentuje (např. les, půda, asfalt apod.). Na množině $X$ pak definujeme tematickou (atributovou) metriku $d_a$ takto:

$$d_a(x_i, x_j) = 0, \text{ jsou-li } x_i, x_j \text{ ze stejné třídy}$$
$$d_a(x_i, x_j) = k, \qquad (11)$$

kde $k \in N$, jsou-li $x_i$, $x_j$ z různé třídy, $N$ je množina přirozených čísel.

Prvek $x_i \in X_i$ můžeme v našem případě považovat za pixel ortofota. Vzdálenost $d_a$ mezi pixely, které patří do různých tříd, pak vypočítáme pomocí klasifikačního stromu. Strom má 1 kořen, který se hierarchicky větví, přičemž počet hierarchických stupňů není omezen. Listy stromu reprezentují jednotlivé klasifikační třídy $X_i$. Strom si sestavuje uživatel účelově podle charakteru úlohy, kterou řeší. V našem případě je strom znázorněn na obr. 3.




Obr. 3 – Klasifikační strom

Vnitřní vrcholy stromu, které se dále větví, mají přiřazen zvláštní symbol (*, &, …). Listy stromu jsou očíslovány přirozenými čísly v pořadí hran, které vycházejí ze stejného uzlu. Na základě klasifikačního stromu můžeme pro každou třídu odvodit jednoznačný řetězcový kód, s jehož pomocí je možné klasifikační strom rekonstruovat. Např. pro třídu *půda* je řetězcový kód **3*, pro třídu *cesta* kód *&6*. Atributová metrika je v rámci podstromu rovna *1*. Například v obr. 3 mají třídy *les*, *tráva*, *půda*, *stín* vzájemnou vzdálenost *1*. Vzdálenost tříd přes *n* vnitřních uzlů je rovna *n*, např. *půda-asfalt*, nebo *tráva-střecha rovná*, mají vzdálenost *2*. Smyslem zavedení tematické metriky spočívá v tom, že pokud identifikujeme např. místo *trávy les*, jde o menší chybu, než kdybychom místo *trávy* identifikovali např. *asfalt*. Tato metrika byla použita pro analýzu chybovosti klasifikace – viz podkapitola 4.5 a 5.6.

# 4 METODIKA ŘEŠENÍ

Teoretická východiska disertační práce byla ověřena na projektu zabývajícím se klasifikací údajů o uložení plynárenských zařízení pod určitými typy povrchů terénu. Tento projekt je specifický svým územním rozsahem, objemem zpracovávaných dat, relativně krátkou dobou řešení a snahou o co největší přesnost výstupu. Rozsah zpracovávaného území byl 68312 $km^2$ a pokrýval téměř celé území našeho státu. Objem zpracovávaných dat se pohyboval okolo 1 TB. Doba řešení byla stanovena na 16 týdnů. Přesnost výsledku byla očekávána v řádech maximálně několika jednotek metru, maximální chybovost byla stanovena do 5 %. Celý tento proces bylo potřeba co nejvíce automatizovat. Tyto charakteristiky řadí projekt mezi masivní úlohy se závazným termínem dokončení. Vzhledem k těmto vlastnostem bylo nutné navrhnout a ověřit metodiku řešení masivních úloh v GIS.

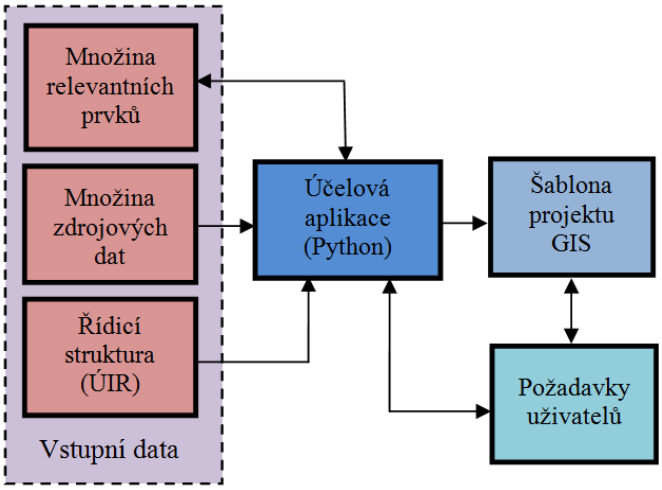
Při prvotní fázi zkoumání vhodných dat pro analýzu druhů povrchů musel být brán zřetel na rozsah zpracovávaného území, které pokrývá téměř celé území státu. Proto bylo potřeba uvažovat za vhodná taková data, která se vyskytují po celé ČR. Z tohoto důvodu bylo vybráno za hlavní zdrojovou datovou sadu pro analýzu **ortofoto**, z kterého se dá pomocí klasifikace obrazu získat informace o typu povrchu. Po důkladnější analýze bylo shledáno ortofoto za jediný zdroj dat jako nedostačující, zejména proto, že z něj nelze spolehlivě získat informaci o hlavní a místní komunikaci (požadováno zadavatelem projektu). Proto byl za druhou zdrojovou datovou sadu vybrán **ZABAGED**, a to konkrétně liniový zákres komunikací, který zpřesňuje výsledky klasifikace obrazu nad ortofotem a navíc poskytuje důležitou informaci o druhu komunikace. Za základní metodu řešení lze tedy považovat klasifikaci obrazu ortofota s rozlišením 0,25 m/pixel, která je následně zpřesněna vybranými vrstvami ZABAGED. Celý proces je řešen pomocí řady funkčních nástrojů v GIS tvořící analytický celek.

## *4.1 Příprava dat*

Příprava dat hraje velmi důležitou roli v řešení masivních úloh. Zdrojová data mohou být různých datových typů (vektor, rastr, tabulková databáze) a je nutné je upravit před hlavním zpracováním. Tím se rozumí odstranit topologické chyby a jiné nedostatky, ale také vhodně rozdělit velké množství dat na menší segmenty pro zpracování. Proto je hlavním cílem kvalitní přípravy datové struktury projektu rozdělení a uspořádání datových sad do vhodných menších celků a jejich uložení do vhodné projektové struktury prostředí GIS. Na obr. 4 je znázorněno obecné schéma přípravy dat. V případě projektu bylo jádrem řešení vytvoření účelové softwarové aplikace ve skriptovacím jazyku Python s podporou knihoven ArcGIS ESRI.

Tato softwarová aplikace třídí a vybírá zdrojová data do speciální datové struktury šablony projektu, která je předem vytvořena v prostředí ArcGIS. Aplikace tak realizuje zobrazení $\varphi: X \rightarrow Y$, kde množina $X$ reprezentuje zdroj vstupních dat a množina $Y$ digitální geodatabázi v rámci šablony projektu. Zdrojová data softwarová aplikace rozděluje na základě řídicí struktury – územně identifikačního registru (ÚIR). V rozsáhlých projektech mnohdy nastanou případy, kdy toto rozdělení je vhodné upravit vypuštěním redundantních dat. Např. optimalizaci pokrytí území ortofotem je výhodné realizovat pouze na datově společných (vzájemně překrytých) oblastech. Proto je účelné dané dílčí územní části dále rozložit podle

tematického hlediska. Dostáváme tak hierarchický rozklad vstupní množiny datových sad, kde kritérium druhého stupně rozkladu je množina relevantních prvků.




Obr. 4 – Schéma předzpracování vstupních dat projektu

## 4.2 *Výkonnost a řešitelnost složité úlohy*

Výkonnost technologie a řešitelnost složité úlohy v GIS je ovlivněna především těmito faktory:

- parametry technického vybavení – hardware (HW)
- výpočetní mohutnost vybraného programového vybavení – software (SW)
- zvolená metodika zpracování celé úlohy
- organizace celého zpracovatelského procesu a podmínky, v nichž probíhá řešení (úroveň informační infrastruktury)
- počet geografických objektů ve vstupních datech a jejich geometrická a topologická složitost.

Míra vlivu faktorů na dílčí doby zpracování podle níže uvedeného vztahu (12) je uvedena v tabulce 3. Tato míra vlivu je ohodnocena třemi stupni: nízký, střední a vysoký vliv.

| Faktory/doba | HW | SW | Metodika | Organizace | Počet GO |
|---|---|---|---|---|---|
| $T_{rež}$ | střední | střední | střední | střední | střední |
| $T_{pa}$ | vysoký | vysoký | nízký | nízký | vysoký |
| $T_{pm}$ | střední | střední | střední | střední | střední |

Tab. 3 – Vliv faktorů na dobu řešení úloh s enormním objemem dat

Celková doba řešení úlohy je dána vztahem:

$$T_c = T_{rež} + T_{pa} + T_{pm} \tag{12}$$

kde $T_{rež}$ … režijní čas (příprava dat)

$T_{pa}$ … doba automatizovaného zpracování

$T_{pm}$ … doba manuálního zpracování.

Dalším významným faktorem, který sice zásadně neovlivňuje výkonnost, ale souvisí s řešitelností úlohy, je spolehlivost. Dlouhá doba řešení $T_c$ zvyšuje riziko nespolehlivosti a má značné nároky na parametry HW a SW. Podle tabulky 3 lze tuto dobu zkrátit kvalitním HW, SW a snížením počtu geografických objektů. Protože HW a SW mají svoje technická omezení, jeví se jako nejvhodnější krok, vedoucí ke zkrácení celkové doby řešení a současně zvýšení spolehlivosti procesů, rozdělit vstupní data do menších částí. Z hlediska efektivnosti je zapotřebí, aby doba režie nepřesáhla dobu zpracování, tj. aby poměr doby processingu $T_p = T_{pa} + T_{pm}$ k době režie byl větší než 1:

$$\eta = \frac{\sum_{i=1}^{n} T_{p_i}}{\sum_{i=1}^{n} T_{rez_i}} > 1 \qquad (13)$$

kde $T_{rezi}$ … čas režie

$T_{pi}$ … doba zpracování dílčího bloku vstupních dat

$n$ … počet bloků, na které je objem vstupních dat rozdělen.

Hodnoty časů $T_{rezi}$ a $T_{pi}$ byly empiricky stanoveny v rámci disertační práce.

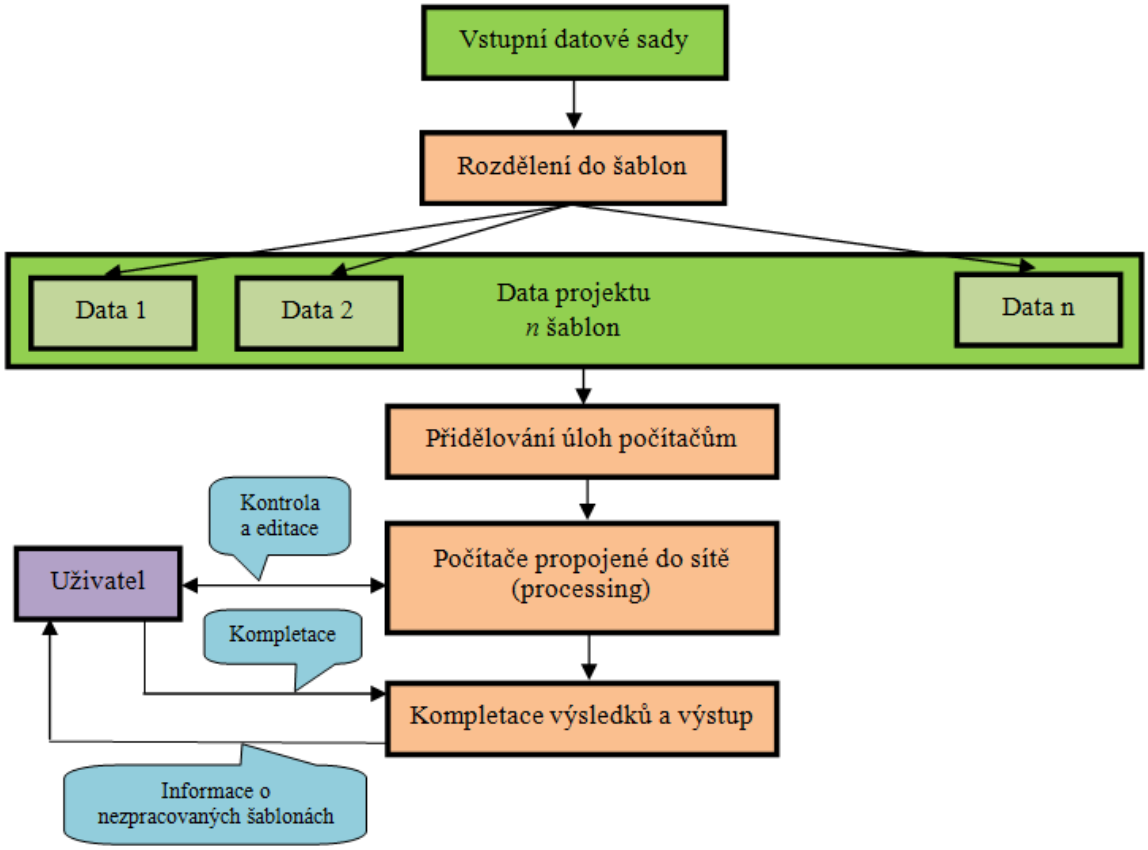
## *4.3 Navržená technologie zpracování velkého množství geografických dat*

Realizovaná technologie zpracování značného objemu dat projektu, blíže popsaná v experimentálních výsledcích práce, je založena na poloautomatickém principu. Blokové schéma postupu zpracování je na obr. 5. Podstatou je rozdělení dat ze vstupních datových sad do samostatných bloků, tzv. šablon, podle územního a tematického principu. Územní rozdělení je provedeno podle hranic administrativního členění ČR, podkladem pro tematické rozdělení dat je zákres průběhu plynovodu.

Toto rozdělení vstupních dat se provádí automatizovaně podle účelové procedury v jazyku Python s podporou knihoven ESRI. V této fázi je nejprve nutné stanovit optimální územní jednotku zpracování jako prvek šablony projektu (viz *Data i* na obr. 5). Datové rozdělení podle územního principu je na základě koeficientu $\delta$ (viz vztah 3). Velikost jednotlivých bloků dat je diskutována v kapitole 5.1.

Tímto způsobem se data celého projektu rozdělí do dílčích šablon, které se pak zpracovávají nezávisle v počítačích propojených do sítě. Přidělování dílčích úloh jednotlivým počítačům je vyřešeno algoritmem, který je blíže popsán v podkapitole 3.2. Díky němu se určí počet nutných prostředků (počítačů) k vyřešení úlohy ve stanovené lhůtě a k těmto prostředkům se přidělí jednotlivé dílčí úlohy ke zpracování (tzv. processing šablon).

Hlavním procesem je klasifikace obrazu ortofota pro účel určení druhu povrchu v místě plynárenských zařízení s následnou filtrací vybranými vrstvami vektorové datové sady ZABAGED. Tento proces probíhal zcela automatizovaně v účelové proceduře vytvořené v jazyku Python. Po skončení tohoto procesu následuje vizuální kontrola uživatelem. Na závěr se dílčí výsledky z jednotlivých datových projektů kompletují do společného výstupu - filtrovatelné kontingenční tabulky. Tento proces probíhá automatizovaně opět v proceduře v jazyku Python.




Obr. 5 – Harmonogram polo-automatických procesů ve velkém GIS projektu

### *4.4 Zpřesnění klasifikace obrazu*

Vstupní datové sady projektu byly uloženy na serveru a obsahovaly:

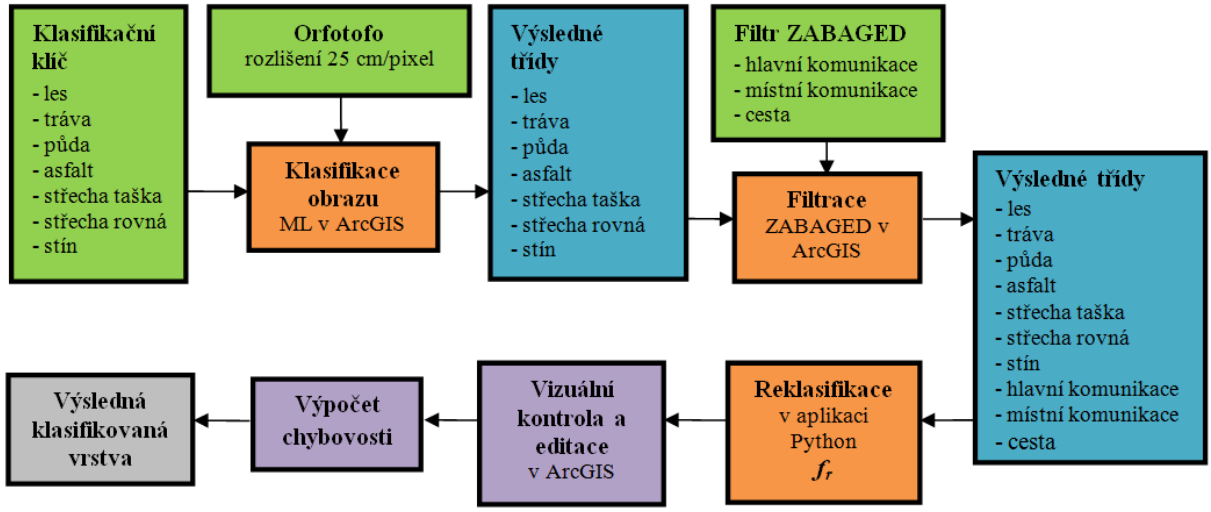
- rastrové ortofoto snímky ČR v měřítku 1: 5 000 s rozlišením 0,25 m/pixel (TIFF)
- vektorový průběh plynového vedení (DGN formát)
- vektorový průběh os vybraných typů komunikací ze ZABAGED (SHP formát)
- vektorové průběhy hranic území podle administrativního členění státu (SHP).

V rámci přípravy dat se ze vstupních dat zkopírovaly datové sady do šablony projektu, která byla členěná podle územního principu. Vlastní zpracování dat probíhalo v prostředí ArcGIS. Základní metodou řešení datové analýzy v projektu byla klasifikace obrazu ortofota s rozlišením 0,25 m/pixel, která probíhala v několika fázích. V těchto fázích docházelo k postupnému zpřesňování výsledků, viz obr. 6:

1. automatická klasifikace metodou Maximum Likelihood (ML) v ArcGIS
2. automatická filtrace výsledku vybranými vrstvami ZABAGED v ArcGIS
3. automatická reklasifikace pomocí účelové aplikace v jazyku Python
4. manuální kontrola výsledků a výpočet chybovosti.

Výsledkem klasifikace obrazu metodou Maximum Likelihood bylo 7 tříd povrchů. Protože průběh plynovodu v mnoha případech zasahuje do komunikací, byl jako další krok zpřesnění výsledků zvolen filtr ZABAGED s vrstvami komunikací. Filtr byl realizován obalovou zónou kolem os komunikací. Šířky obalových zón byly určeny na základě přímého měření vybraných

vzorků komunikací v každé projektové datové územní jednotce v ArcGIS. Filtrací byly rozšířeny třídy z klasifikace obrazu o třídy žádaných komunikací (hlavní a místní) a nezanedbatelným přínosem bylo také zpětné zpřesnění třídy zpevněného povrchu, konkrétně povrchu typu asfalt, protože hlavní i místní komunikace jsou tvořeny právě tímto povrchem. Následně proběhla automatizovaná reklasifikace výstupu na základě charakteristiky typů plynovodů v daném prostředí a další mírné zpřesnění výsledků klasifikace bylo dosaženo až závěrečnou vizuální kontrolou s následnou ruční editací v prostředí ArcGIS.




Obr. 6 – Postupné zpřesňování výsledků klasifikace obrazu ortofota

## 4.5 *Stanovení chybovosti analýzy*

Součástí závěrečných fází projektu byla vizuální kontrola výsledků v mapovém okně ArcGIS a stanovení absolutní chybovosti klasifikace. Chybovost určíme na základě poznatků z podkapitoly 3.5 a s využitím následujících úvah vyplývající z obrázku 6.

Nechť $C$ je množina prvků ve výsledné vrstvě klasifikace. Prvkem zde rozumíme rozlišitelnou jednotku, např. pixel ortofota (v rastrové vrstvě) nebo jednotkovou délku plynového vedení (ve vektorové vrstvě). Chybovost pak vypočítáme podle vztahu:

$$e = \frac{\sum_{i=1}^{k} n_i d_a(c_i, c_j)}{card(C)}, \tag{14}$$

kde

$d_a$ … atributová vzdálenost prvků $c_i$, $c_j$ (viz podkapitola 3.5)

$n_i$ … počet chybně zařazených prvků $c_i \in C$ (po vizuální kontrole byl místo $c_i$ jako správný určen prvek $c_j \in C$)

$k$ … počet segmentů (skupin prvků), kde byla nalezena chyba

$card(C)$ … počet prvků množiny $C$.

# 5 EXPERIMENTÁLNÍ VÝSLEDKY

Experimentální výsledky disertační práce souvisí s výše zmíněným projektem klasifikování druhů povrchů terénu nad průběhem trasy plynového vedení. Analýza povrchů měla být provedena ve třech variantách pro průběh plynovodu typu vysokotlak, hlavní řad a přípojky. Byly požadovány 3 klasifikační kategorie výstupů pro každý typ plynovodu:

A. 2 třídy druhu povrchu: *zpevněný* a *nezpevněný*
B. 5 tříd druhu povrchu: *asfalt*, *hlavní komunikace*, *místní komunikace, nevázaný* (štěrk, zelená plocha a další nezpevněné povrchy), *neznámý*
C. 10 tříd druhu povrchu: *les*, *travní porost*, *holá půda*, *asfalt*, *střecha – taška*, *střecha – rovná*, *stín*, *hlavní komunikace*, *místní komunikace*, *cesta*. To znamená analýza úseků plynovodu pod všemi typy povrchů, které jsou z datových sad klasifikovatelné.

Výstupem je strukturovaná tabulka ve formátu XLS s klasifikovanými údaji typů povrchů nad průběhem trasy plynovodu a grafický výstup strukturovaných klasifikovaných údajů ve formátu SHP pro vysokotlak, hlavní řad a přípojky.

## *5.1 Datová analýza povrchů terénu*

Řešitelnost úlohy byla nejprve otestována v rámci pilotního projektu [13]. Výsledky projektu stanovily výběr vhodných datových sad pro zpracování úlohy (viz úvod předchozí kapitoly), dílčí zpracovatelskou jednotku, vhodný software a hardware, a nakonec byl zkonkretizován kompletní postup jednotlivých fází řešení úlohy (viz obr. 7).

Základní zpracovatelská jednotka

Rozdělení vstupních dat v GIS musí být nejen v souladu s rovnicí (13), ale také by mělo logicky respektovat územní a tematické členění. Podle pilotního projektu bylo na základě koeficientu $\delta$ (viz vztah 3) stanoveno optimální rozdělení dat na datové jednotky projektu o objemu ~2,2 GB, které odpovídají rozsahu území o rozloze ~368 km$^2$. Taková rozloha odpovídá administrativnímu členění ČR na tzv. obce s rozšířenou působností (ORP). Zpracovávané území analýzy pokrývalo celkem 188 ORP.

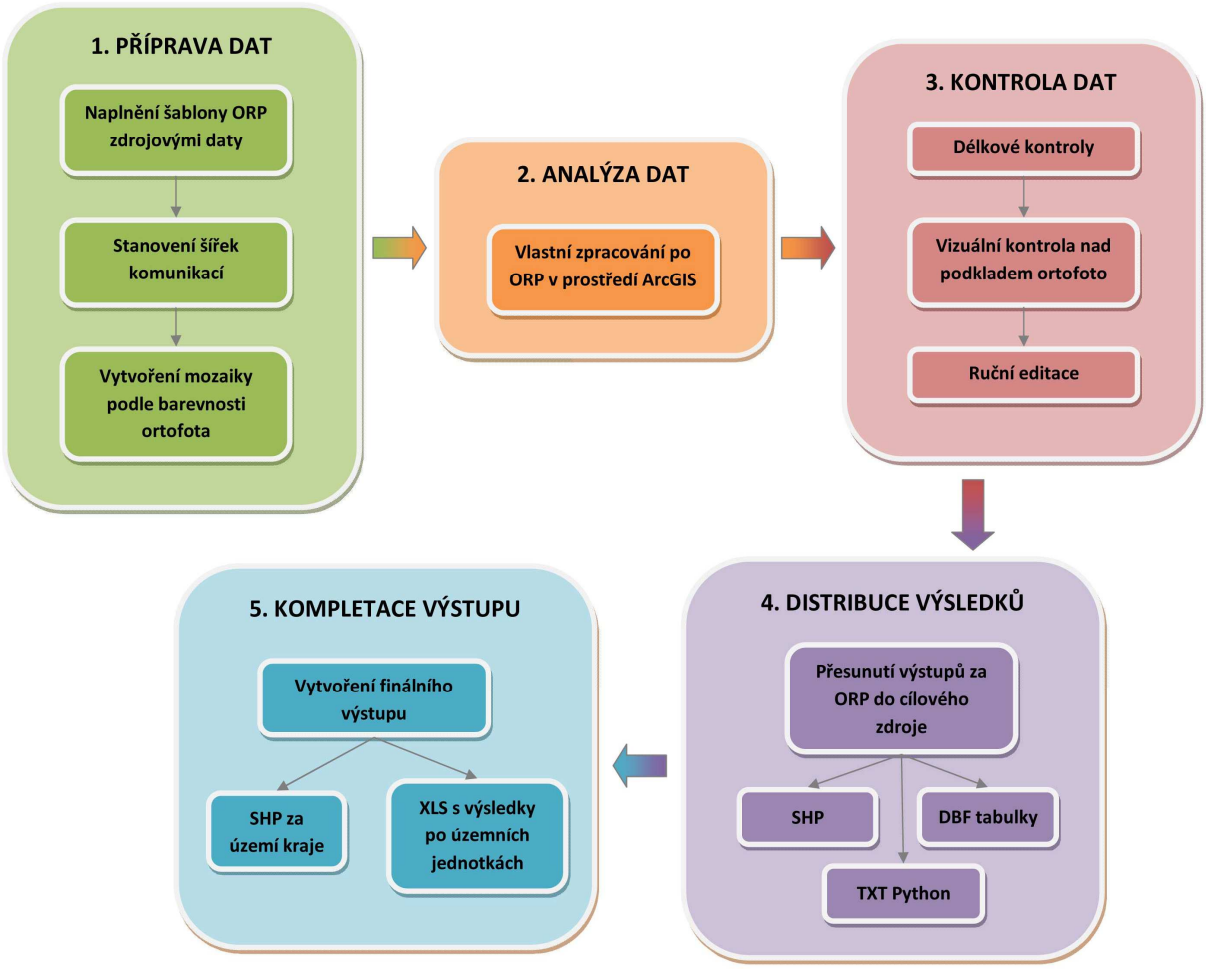
Výběr vhodného software

Vzhledem k tomu, že plynárenská zařízení jsou geografické objekty s územně vázanými informacemi, nestačila k úspěšnému řešení jen prostá klasifikace obrazu, ale bylo zapotřebí i aplikace prostorových analýz na vstupní data. Z tohoto hlediska se jako optimální programové prostředí jevilo prostředí GIS s integrovaným modulem pro klasifikaci obrazu. Z tohoto důvodu byl využit pro realizaci projektu software ArcGIS verze 10.0 od firmy ESRI. Protože bylo potřeba jednotlivé kroky úlohy co nejvíce automatizovat, byl dále využíván aplikační software Python 2.6 pro tvorbu skriptů.

Výběr vhodného hardware

Pro úspěšné zpracování úlohy s velkým objemem dat bylo potřeba vytvořit počítačovou síť s výkonnými jednotkami. K určení počtu vhodných prostředků pro řešení úlohy a k přiřazení

procesů jednotlivým prostředkům byl využit výše uvedený optimalizační algoritmus (viz podkapitola 3.2). Aplikací algoritmu bylo zjištěno, že minimální počet prostředků pro celkovou dobu řešení je použití 3 výkonných PC a 1 notebooku. Notebooků bylo nakonec použito více, ale sloužily spíše jen jako záložní zdroje, nebo jako prostředek ke vzdálenému přístupu k PC.




Obr. 7 – Fáze pracovního postupu řešení úlohy

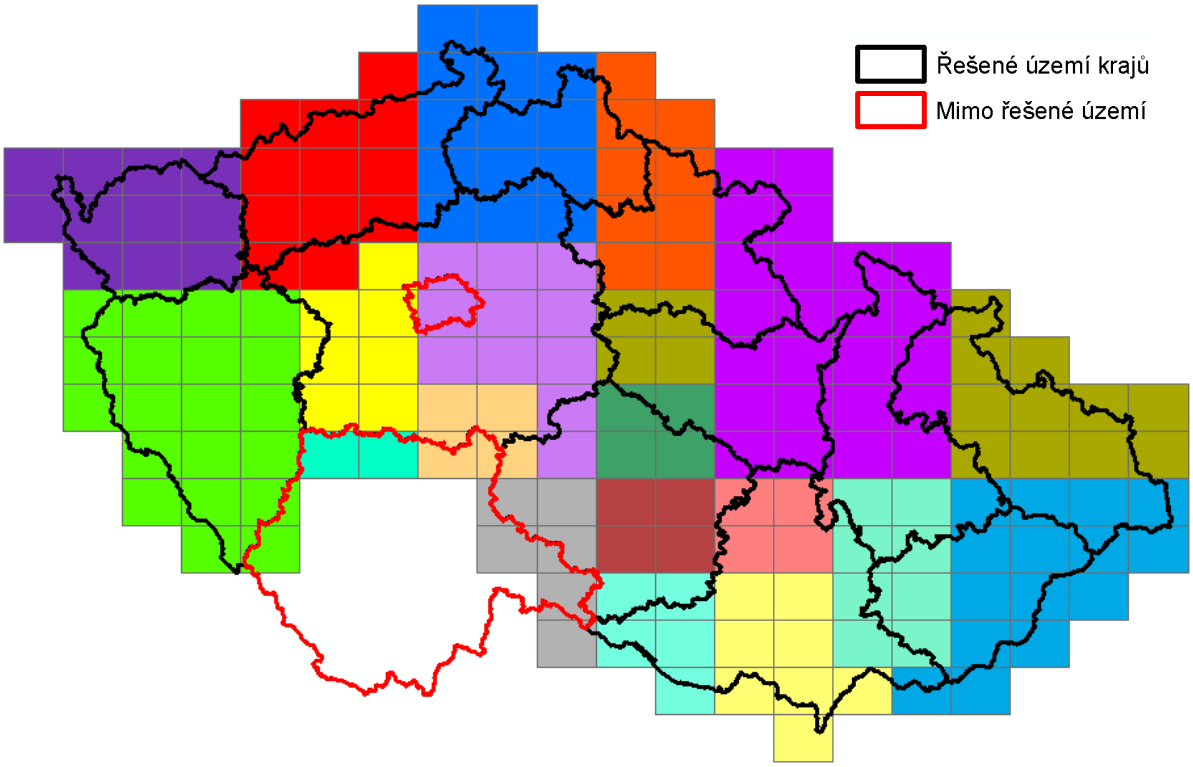
## 5.2 *Příprava dat*

Příprava dat je velmi důležitou fází při řešení masivních úloh před hlavním zpracováním dat. Tato fáze se skládala ze dvou kroků. Po získání veškerých potřebných dat pro analýzu bylo nejprve potřeba tato data upravit a dále je nachystat pro následné zpracování.

Úprava zdrojových dat

Nejnáročnější práce byla s datovou vrstvou *ortofoto*. Po zhlédnutí získaných snímků, které pokrývaly celé zpracovávané území, bylo zjištěno, že nejsou stejné škály barevnosti. Tato situace zkomplikovala řešení úlohy, protože nemohlo být použito univerzálního klasifikačního klíče při klasifikaci obrazu ortofota. Po detailnějším zkoumání bylo nad zájmovým územím nalezeno 20 odlišných oblastí barevnosti, které však byly různých rozměrů (viz obr. 8). Dále bylo zjištěno, že hranice těchto oblastí kopírují hranice kladu listů Státní mapy v měřítku

1: 50 000 (SM50). Z tohoto kladu je odvozen klad mapových listů ortofoto v měřítku 1: 5 000. Pro těchto 20 oblastí barevnosti byly v prostředí ArcGIS manuálně vytvořeny trénovací množiny pro každou z těchto oblastí a z nich byly následně spočítány klasifikační klíče.




Obr. 8 – Oblasti barevnosti ortofota podle kladu mapových listů SM50

Příprava datových zpracovatelských jednotek

Jakmile byla provedena kompletní úprava zdrojových dat a byly vytvořeny klasifikační klíče, mohlo se přistoupit k přípravě datových zpracovatelských jednotek před vlastním zpracováním. Nejprve bylo potřeba vytvořit strukturu adresářů po územních jednotkách kraje, okresu a ORP. Tato struktura sloužila ke snadnější orientaci při manuální práci s jednotlivými zpracovatelskými jednotkami.

Následovaly tří hlavní kroky přípravy datových jednotek:

1. Naplnění šablony ORP zdrojovými daty
2. Stanovení skutečných šířek komunikací
3. Vytvoření mozaiky podle barevnosti ortofota.

*Naplnění šablony ORP zdrojovými daty*

Řešení předzpracování vstupních dat s cílem naplnit šablonu zdrojovými daty je blíže popsáno v podkapitole 4.1. Do adresáře vybraného ORP se nakopírovala šablona prázdného GIS projektu, do kterého se následně zkopírovala a upravovala zdrojová data. Všechny výše uvedené operace této fáze probíhaly automatizovaně a bylo poměrně složité je naprogramovat.

*Stanovení skutečných šířek komunikací*

Bylo zjištěno, že velikosti šířek komunikací nejsou ve všech ORP stejné a tudíž je nelze jednoduše parametrizovat. Rozdíly ve velikosti defaultních hodnot bufferů kolem komunikací oproti skutečnosti by vnesly chyby do analýzy. Proto bylo rozhodnuto, že musí proběhnout v rámci procesu přípravy dat zjištění skutečných šířek komunikací v daném ORP. Šířky byly poté stanoveny jako průměrné hodnoty z manuálního odměřování nad zákresem linií ZABAGED. Jako podklad pro tuto činnost sloužila vytvořená mozaika z ortofota.

*Vytvoření mozaiky podle barevnosti ortofota*

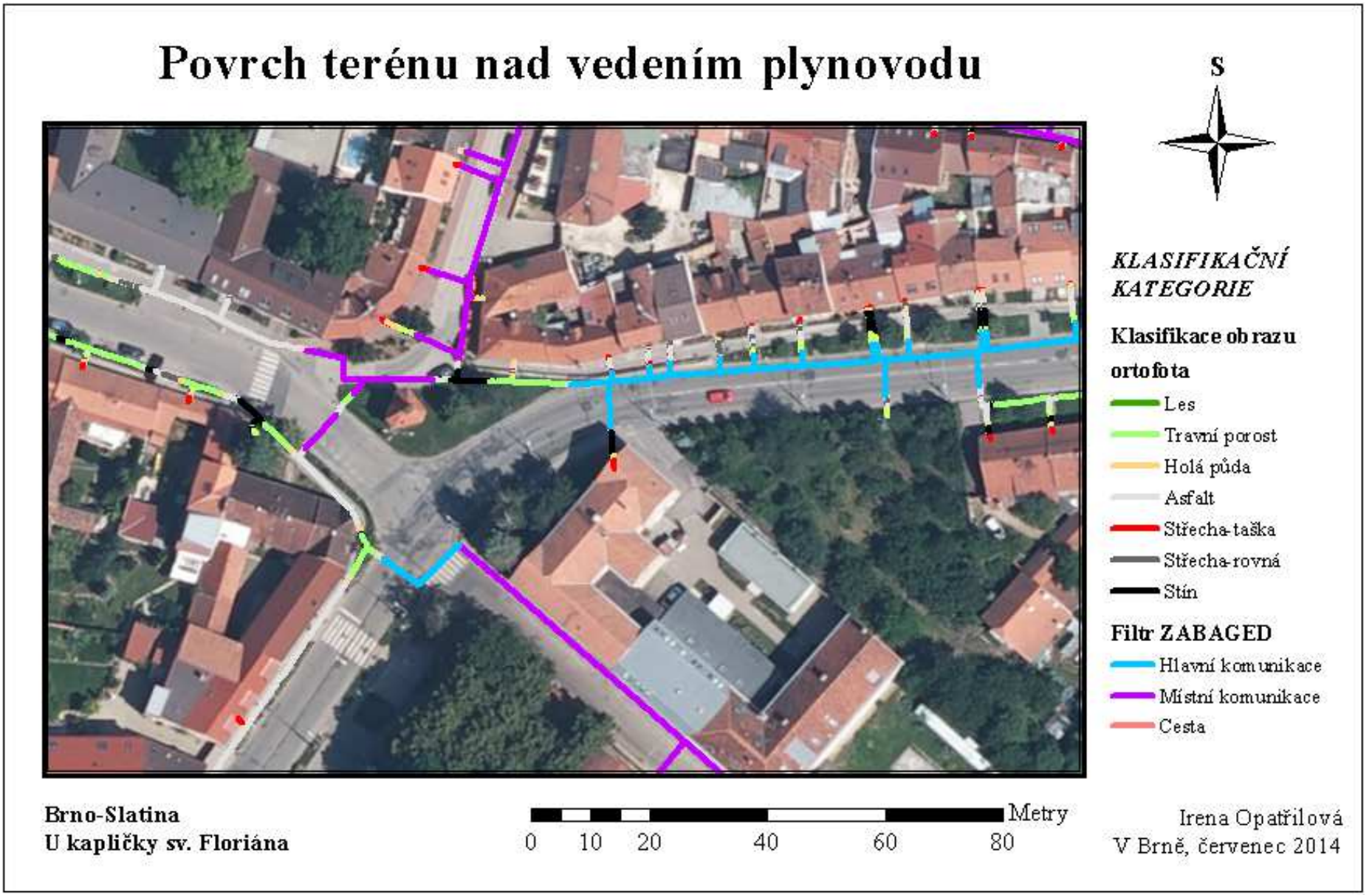
Z důvodu nestejné barevnosti ortofota bylo potřeba v dalším kroku přípravy vytvořit mozaiky spojených rastrů se stejnou barevností nad daným ORP. Jako poklad pro výběr ortofota k tvorbě mozaik sloužil upravený klad SM50 ve formátu SHP. Poté byly zkopírovány ze vstupních datových sad potřebné vybrané klasifikační klíče do šablony ORP a dále byla smazána původní celistvá mozaika z ortofota nad celým ORP bez rozlišení barevnosti. Celý tento proces probíhat automatizovaně. Nejvíce byla zastoupena skupina s jedním klasifikačním klíčem (108 ORP), objevila se zde však i taková ORP, která zasahovala do čtyř oblastí barevností ortofota.

## *5.3 Zpracování dat v GIS*

V této fázi dochází k vlastnímu zpracování analýzy nad územím příslušného ORP. Vše probíhalo automatizovaně v prostředí ArcGIS. Řešení hlavní analýzy je stručně popsáno v následujícím textu. Nad mozaikami z ortofota stejné barevnosti, které pokrývaly pouze oblast vedení plynovodu, byla provedena klasifikace obrazu. Konkrétně byl zvolen za typ řízené klasifikace *Maximum Likelihood*. Při klasifikaci obrazu byl automatizovaně přiřazen konkrétní mozaice, která spadala do dané oblasti barevnosti, příslušný klasifikační klíč. Proces klasifikace obrazu byl jeden z časově nejnáročnějších analýz v rámci celé úlohy. Průměrně tento proces trval pro území jednoho ORP 50 minut. Je důležité poznamenat, že provádět klasifikaci obrazu pro rozlišení 0,25 m/pixel v tak malém měřítku jako je rozsah téměř celého území státu, navíc s různou barevností vstupních dat, je úkol specifický a není běžnou součástí úloh dálkového průzkumu Země.

Protože samotný výsledek z klasifikace obrazu by nebyl dostatečně přesný (např. koruny stromů zejména ve městech často zakrývají komunikace, kde jsou tato místa chybně určena jako vegetace) a protože bylo potřeba do analýzy zavést i atribut rozdělení komunikací na hlavní a místní, upravené vrstvy vybraných komunikací ze ZABAGED tvořily tzv. filtr pro vrstvu z klasifikace obrazu. Proces filtrace výstupu z klasifikace obrazu vedl ke zlepšení přesnosti analýzy.

V závěrečných fázích analýzy byl výsledek upraven na územní podrobnost částí obcí a vytvořen výstup pro 3 typy plynovodů. Výstupem byly statistické tabulky oklasifikovaného plynovodu pro jednotlivé typy podrobnosti klasifikace (A, B a C) a grafický výstup ve formátu SHP. Ukázka grafického výstupu analýzy dat je na obr. 9.




Obr. 9 – Ukázka grafického výstupu analýzy dat

## *5.4 Kontrola dat*

Kontrola dat byla přítomna ve všech fázích úlohy. Při přípravě dat se jednalo o vizuální kontrolu správně naplněné šablony, vytvořených bufferů komunikací vůči skutečným šířkám, či správně vytvořených mozaik pro stejné barevnosti ortofoto. Nejdůležitější část kontroly však byla po hlavním zpracování. Ta se skládala ze dvou fází, a to z kontrolních součtů délek plynu v DBF tabulkách a vizuální kontroly grafického výstupu v ArcMap.

## *5.5 Distribuce výsledků a kompletace výstupu*

Pokud byl výstup za ORP zkontrolován a byl korektní, bylo potřeba z něj vyexportovat výsledky do adresáře s výstupy. Kompletací výstupu vzniklo několik SHP souborů o celkové velikosti 8,8 GB a XLS soubor s kontingenční tabulkou pro lehké procházení a filtrování hodnoty podle různých územních celků od krajů až po jednotlivé části obcí.

## *5.6 Stanovení chybovosti výstupu analýzy*

Cílem projektu bylo klasifikovat povrch nad liniovým průběhem vedení plynovodu s maximální chybovostí do 5 %. Řízená klasifikace rastrového obrazu metodou Maximum Likelihood v prostředí ArcGIS vykazovala průměrnou úspěšnost cca **72 %**, což bylo pro potřeby projektu nedostačující. Předběžný průzkum ukázal, že největší chybovost rastrové klasifikace obrazu je v místě komunikací, kde okolní vegetace nebo budovy vrhají stín na zemský povrch, který se tak stává neklasifikovatelný a tím je zkreslen výsledek klasifikace. Proto byl výsledek automatizované klasifikace zpřesněn cca o **25 %** následnou reklasifikací použitím filtru vektorové datové sady komunikací z databáze ZABAGED.

Výsledek byl dále zpřesněn přetříděním klasifikovaných prvků v extravilánu a intravilánu pomocí vzájemných logických podmínek klasifikovaných skupin implementovaných do účelově vytvořené aplikace v jazyku Python (úprava atributové tabulky výstupu). Poté proběhla vizuální kontrola výsledků v mapovém okně ArcGIS, ruční editace chybných úseků u hlavního řadu a přegenerování výsledků. Ruční editací bylo odstraněno už jen cca **1 %** chyb. Na závěr bylo provedeno stanovení chybovosti klasifikace.

Mezi příčiny chybovosti výsledku patří následující:

1. různá kvalita barevnosti ortofota
2. variabilita šířek komunikací stejného typu
3. variabilita barevnosti téhož povrchu
4. výskyt zastíněných prostor v důsledku vegetace nebo budov
5. nesoulad průběhu komunikací v ortofoto a v ZABAGED.

Absolutní chybovost analýzy

Tabulka 4 uvádí absolutní chybovost analýzy dat určenou statistickým vyhodnocením dle vztahu (14) z výběrového souboru pokrývajících 21 % řešeného území (tj. 39 vybraných ORP). Vzorky ORP byly vybírány účelově pro podchycení výše uvedených příčin chybovosti, zejména různobarevnosti ortofoto. Chybné úseky špatně oklasifikovaného plynovodu byly proměřovány ručně v ArcGIS mezi kategoriemi povrchu zpevněný, nezpevněný a stín.

| | Absolutní chybovost analýzy [%] | | |
|---|---|---|---|
| Typ plynovodu | Hlavní řad | Přípojky | Vysokotlak |
| Průměr | 1,9 | 1,5 | 0,5 |
| Standardní odchylka | 1,4 | 1,1 | 0,4 |
| Maximální odchylka | 5,4 | 4,6 | 2,6 |
| Minimální odchylka | 1,4 | 1,1 | 0,4 |

Tab. 4 – Absolutní chybovost analýzy

Chybovost v závislosti na barevnosti ortofoto

Účelový výběr vzorků ORP pro určení absolutní chybovosti analýzy byl podmíněn tomu, aby vzorky reprezentovaly jednu z těchto zvolených kategorií:

- krajské město
- ORP, kde byly tvořeny trénovací množiny a spočítán klasifikační klíč pro danou oblast barevnosti ortofoto
- ORP v rámci jedné oblasti barevnosti, které bylo nejvzdálenější od místa, kde byl vytvořen klasifikační klíč
- ORP, které pokrývá více oblastí barevnosti ortofoto.

Statistické výsledky analýzy chybovosti v závislosti na těchto čtyřech kategoriích vzorků byly zpracovány pro hlavní řad. Chybovost automatizace zpracování byla získána součtem změn po ruční editaci a absolutní chybovosti v daném ORP a vyjadřuje tak přesnost řešení úlohy bez zásahu ruční kontroly a úpravy výsledků.

Průměrné výsledky v rámci jedné skupiny chybovosti jsou v rozmezí do **1,6 %**, tzn., že významná závislost přesnosti analýzy dat na charakteru ORP se nepotvrdila. Chybovost automatizace zpracování dat měla očekávanou největší hodnotu v případě kategorie ORP, které bylo nejvzdálenější od vytvořeného klasifikačního klíče, tj. **4,6 %**. Naopak nejlepší přesnost automatizace analýzy (**3,2 %**) dosáhla kategorie krajských měst, kde zřejmě hrála velkou roli filtrace pomocí komunikací ZABAGED.

## *5.7 Vyhodnocení optimalizace přípravy dat*

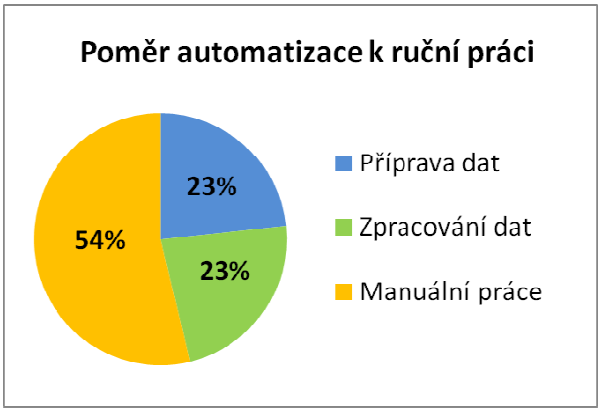
Ze zdrojových dat byly automatizovaně vygenerovány dílčí datové projekty pro území ORP. Automatizované vygenerování dílčích datových projektů představovalo nezanedbatelný časový nárok a probíhalo na výkonných počítačových strojích typu PC. Průměrná doba zpracování jednotky ORP byla 1 hod 19 min, přičemž průměrný objem dat v jednotce činil 2,2 GB. Celkový objem dat pro celé zpracovávané území byl po datové přípravě 420 GB.

Strojočas, potřebný na přípravu dat, je ovlivněn členitostí průběhu plynovodu, což souvisí s množstvím grafických prvků. Úsporu času lze dosáhnout zmenšením množství dat. Celkové množství dat pokrývající územní jednotku bylo redukováno pomocí tematické selekce podle věcného uspořádání na data optimálně pokrývající pouze průběh vedení plynovodu. Po této optimalizaci byl snížen objem dat z celkového množství (100 % = 880 GB) na množství **48 %** (420 GB), tedy prakticky o polovinu, což znamená i cca poloviční úsporu strojočasu potřebného na přípravu dat.

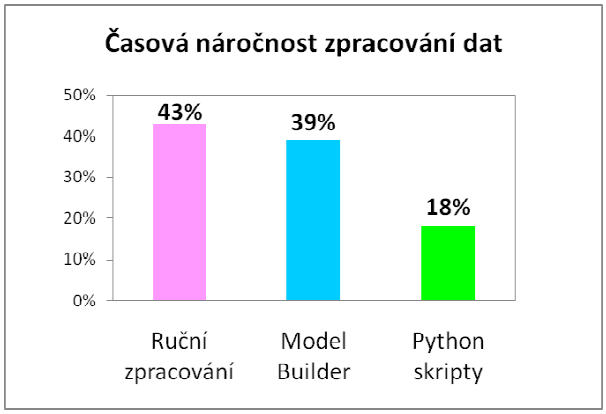
## 5.8 *Vyhodnocení poměru automatizace k ruční práci*

Jednotlivé procesy úlohy zpracování dat včetně datové přípravy bylo snahou co nejvíce zautomatizovat, a tím maximálně zefektivnit celou práci. Průměrná doba automatizovaného zpracování jednoho ORP byla 1 hod 47 min, průměrná velikost zpracovaného projektu ORP byla 3 GB. Celkový objem dat pro celé zpracovávané území byl po hlavní analýze 560 GB.

Na obr. 10 je znázorněn procentuální podíl dílčích automatizovaných procesů datové přípravy ($T_{Ia}$) a zpracování dat ($T_{Oa}$) a procesu manuální práce ($T_{Im}$ + $T_{Om}$) na celku pro řešené území o celkové rozloze 68312 km$^2$, viz vztah (7). Manuální práce trvaly na 1 datové jednotce ORP průměrně 4,5 hodiny, přičemž tyto práce zahrnovaly zejména ruční odměřování skutečných šířek komunikací ve fázi přípravy dat, dále kontrolu a editaci dat po hlavním zpracování.




Obr. 10 – Podíly dílčích automatizovaných procesů a lidské práce na celku




Obr. 11 – Podíly časové náročnosti různých typů zpracování dat

Přestože podíl lidské práce je časově poměrově nadpoloviční, vysoká automatizace přípravy dat a zpracování dat formou Python skriptů vysoce zefektivňuje celý proces vyhodnocení. Při porovnání časové náročnosti výhradně ručního zpracování dat, zpracování prostřednictvím nástroje Model Builder a zpracování prostřednictvím skriptu Python je procentuální poměr časové náročnosti znázorněn na obr. 11. Ručním zpracováním úlohy se myslí postupné manuální zadávání parametrů do nástrojů z ArcToolbox v prostředí ArcMap. Tato forma je zcela bez automatizace a je časově nejnáročnější. Podobně časově náročné je zpracování s využitím aplikace Model Builder v prostředí ArcMap, které může být poloautomatizované. Jednotlivé procesy pomocí Model Builderu trvaly přibližně o čtvrtinu déle, než když běžely při samostatném ručním spouštění. V porovnání s ručním zpracováním je zde však ta výhoda, že parametry jsou již nastaveny a nemusí se manuálně zadávat. Tím se šetří čas. Lidská práce je potřeba pouze u spouštění jednotlivých modelů. Modelů muselo být vytvořeno více, protože zpracování hlavní analýzy bylo velmi složité a vytvořit tak jeden jednodušší kompletní model bylo nemožné. Nejrychlejší, a tím nejefektivnější, je zpracování dat pomocí Python skriptů, které je plně automatizované.

# 6 ZÁVĚR

Hlavním cílem disertační práce bylo navrhnout a ověřit metodiku řešení masivních úloh v GIS. Ověření metodiky bylo realizováno na výzkumném projektu pro společnost RWE. Předmětem projektu byla klasifikace povrchů terénu nad plynovodním vedením. Hlavní metodou řešení byla klasifikace obrazu ortofota s následným zpřesněním pomocí filtrace vybranými datovými vrstvami ZABAGED.

V práci byly řešeny tyto konkrétní problémy:

- posouzení řešitelnosti masivních úloh v GIS pomocí koeficientu $\delta$ (viz vztah 3)
- možnosti využití paralelismu v GIS metodou hierarchického rozkladu daného území
- návrh algoritmu pro optimalizaci úloh přidělování procesů jednotlivým prostředkům
- zpřesnění výsledků klasifikace ortofota filtrací využitím vhodných datových sad
- návrh metody pro stanovení chybovosti datové analýzy.

Výsledky zpracovávané analýzy lze charakterizovat těmito základními vlastnostmi [14]:

- datová analýza vykazuje velmi vysokou vypovídací schopnost s nízkou chybovostí
- technologie datové analýzy umožňuje opakovatelnost bez vlivu lidského faktoru
- technologie se vyznačuje rychlostí získání výsledků datové analýzy
- efektivnost reprocessingu se odvíjí od aktualizace klíčových zdrojových sad, např. ortofota nebo od většího množství změn v území nad hodnotu absolutní chybovosti datové analýzy, která činí 2 %
- technologie datové analýzy má do budoucna kvalitativní potenciál dalšího zpřesnění datové analýzy s možnostmi budoucího využití nových, kvalitativně vyšších, zdrojových datových sad.

**Výsledky disertace mohou být obecně zhodnoceny podle přínosů pro teorii a praxi:**

Za přínos pro teorii je možné považovat:

- koncepce řešení složitých úloh v GIS, využití paralelismu metodou hierarchického rozkladu modelovaného území do tříd na územním a věcném (tematickém) principu
- návrh algoritmu pro optimalizaci úloh s pevným termínem řešení. Algoritmus pracuje na principu přidělování dílčích úloh, na které lze rozdělit hlavní úlohu jednotlivým prostředků, které má uživatel k dispozici.
- návrh hodnocení chybovosti výsledků klasifikace na základě metriky, odvozené ze struktury klasifikačního stromu.

Přínosem pro praxi jsou tyto poznatky:

- ověření technologie zpracování rozsáhlých projektů v GIS na konkrétní úloze
- metoda zpřesnění výsledků klasifikace obrazu použitím relevantních datových sad
- zjištění úskalí a kritických míst v navržené technologické lince způsobených kvalitou vstupních dat, spolehlivostí některých funkcí prostorových analýz v ArcGIS a lidským faktorem
- stanovení reálné limitní hranice kvality výsledků klasifikačního procesu (2 %).

Technologie popsaná v disertační práci má obecný charakter a může být využita pro klasifikaci povrchu nad liniovými inženýrskými sítěmi, jako jsou např. vodovody, produktovody, kanalizace, elektrické energetické rozvody, sdělovací vedení apod.

## PROJEKTY SOUVISEJÍCÍ S DISERTAČNÍ PRACÍ

Pilotní projekt byl řešen v rámci:

- standardního projektu FAST-S-13-2069 Specifického výzkumu pro rok 2013, s názvem *Dolování geo-prostorových dat z disponibilních standardních datových zdrojů prostřednictvím GIS*
- smluvního výzkumu HS123570212101 s názvem *Klasifikace údajů o uložení plynárenských zařízení pod určitými typy povrchů terénu.*

Problematikou tematické oblasti disertační práce se zabývá:

- standardní projekt FAST-S-14-2298 Specifického výzkumu pro rok 2014, s názvem *Řešení masivních úloh v GIS*
- smluvní výzkum HS12357021212200 s názvem *Analýza povrchů nad plynovody RWE na území ČR.*

## PRODUKTY SOUVISEJÍCÍ S DISERTAČNÍ PRACÍ

Ověřená technologie s názvem *Klasifikace údajů o uložení inženýrských sítí a zařízení pod určitými typy povrchů terénu* z roku 2013 je zapsána v databázi RIV s označením RIV/00216305:26110/13:PR27262.

Ověřená technologie s názvem *Technologie zpracování velkého množství geografických dat*, která souvisí s výše uvedeným smluvním výzkumem *Analýza povrchů nad plynovody RWE na území ČR*, je připravená k podání do databáze RIV.

# PUBLIKACE SOUVISEJÍCÍ S DISERTAČNÍ PRACÍ

***Publikace v impaktovaném časopise***

BARTONĚK, D.; BUREŠ, J.; OPATŘILOVÁ, I. Optimization of Pre-Processing of Extensive Projects in Geographic Information Systems. *Advanced Science Letters*, 2014, vol. 20, no. 10/11/12, pp. 2026-2029, ISSN 1936-6612, DOI: 10.1166/asl.2014.5664.

***Kapitola v knize***

BARTONĚK, D.; BUREŠ, J.; OPATŘILOVÁ, I. The Solution of Massive Tasks in GIS Exemplified by Determining Terrain Surface Types above Gas Pipelines in the Czech Republic. *Thematic Cartography for the Society*, Springer, 2014, pp. 95-104, ISBN 978-3-319-08180-9, DOI: 10.1007/978-3-319-08180-9_8.

***Konferenční články vedené v databázi Conference Proceedings Citation Index***

BARTONĚK, D.; BUREŠ, J.; OPATŘILOVÁ, I. Technology of processing of enormous amounts of geographical data. *Proceedings of 14th GeoConference on Informatics, Geoinformatics and Remote Sensing, Albena 19. 6 – 25. 6. 2014,* SGEM2014, vol. 3, pp. 917-924, ISBN 978-619-7105-12-4, DOI: 10.5593/SGEM2014/B23/S11.116.

BARTONĚK, D.; BUREŠ, J.; OPATŘILOVÁ, I.; VITULA, A. Method of error assessment in image classification. *Proceedings of 14th GeoConference on Informatics, Geoinformatics and Remote Sensing, Albena 19. 6 – 25. 6. 2014,* SGEM2014, vol. 3, pp. 745-752, ISBN 978-619-7105-12-4, DOI: 10.5593/SGEM2014/B23/S11.095.

***Publikace v neimpaktovaném časopise ve světových databázích***

BARTONĚK, D.; BUREŠ, J.; OPATŘILOVÁ, I. Possibilities of Improvement of Image Classification via GIS Tools. *Proceedings of 5th International Conference on Cartography and GIS, Riviera, Bulgaria 15. – 21. 6. 2014*, pp. 96-102, ISSN 1314-0604.

***Publikace přijaté na mezinárodních konferencích, které nebyly dosud publikovány***

BARTONĚK, D.; OPATŘILOVÁ, I. Optimization of Processing of Enormous Amounts of Geographical Data. *The 2nd Global Conference on Computer Science, Software, Networks and Engineering, Kuşadası, Turkey 6. – 8. 11. 2014*, 11 p.

BARTONĚK, D.; BUREŠ, J.; OPATŘILOVÁ, I. Enhancement of Image Classification via GIS. *The 2nd International Conferences on Computer Graphics, Visualization, Computer Vision, and Game Technology, Bandung, Indonesia 29. – 30. 10. 2014*, 6 p.

BARTONĚK, D.; BUREŠ, J.; OPATŘILOVÁ, I. Workflow for Analysis of Enormous Amounts of Geographical Data. *The 2nd International Conference on Advances in Intelligent Systems in Bioinformatics, Chem-Informatics, Business Intelligence, Social Media and Cybernetics, Jakarta, Indonesia 27. – 28. 9. 2014*, 8 p.

# POUŽITÁ LITERATURA


[1] GUAN, X.; LIESMARS, H., W.; LIESMARS, L., L. A Parallel Framework for Processing Massive Spatial Data with a Split–and–Merge Paradigm. *Transactions in GIS*, 2012, vol. 16, no. 6, pp. 829-843.

[2] WU, H.; PAN, M.; YAO, L.; et all. A partition-based serial algorithm for generating viewshed on massive DEMs. *International Journal of Geographical Information Science*, 2010, vol. 21, no. 9, pp. 955-964.

[3] RICHTER, R.; DŐLLNER, J. Concepts and techniques for integration, analysis and visualization of massive 3D point clouds. *Computers, Environment and Urban Systems*, 2013, vol. 45, pp. 114-124.

[4] ARGE, L.; CHASE, J.; S., HALPIN, P.; TOMA, L. Efficient Flow Computation on Massive Grid Terrain Datasets. *GeoInformatica*, 2003, vol. 7, no. 4, pp. 283-313.

[5] THIELE, M.; BADER, A.; LEHNER, W. Multi-objective scheduling for real-time data warehouses. *Computer Science - Research and Development*, 2009, vol. 24, no. 3, pp. 137-151.

[6] ARGE, L.; VENGROFF, D., E.; VITTER, J., S. External-Memory Algorithms for Processing Line Segments in Geographic Information Systems. *Algorithmica*, 2007, vol. 47, pp. 1-25.

[7] POLAT, K. A novel data preprocessing method to estimate the air pollution ($SO_2$): neighbor-based feature scaling (NBFS). *Neural Computer and Applications*, 2011, vol. 21, no. 8, pp. 1987-1994.

[8] LI X.-w.; QI Y.-f. A Data Preprocessing Algorithm for Classification Model Based On Rough Sets. *Physics Procedia*, 2012, vol. 25, pp. 2025-2029.

[9] McCUE, C. Operationally Relevant Preprocessing. *Data Mining and Predictive Analysis*, 2007, pp. 93-115.

[10] LI, L.; XU, Z., X. A Preprocessing Program for Hydrologic Model—A Case Study in the Wei River Basin. *Procedia Environmental Sciences*, 2012, vol. 13, pp. 766-777.

[11] PAWLAK, Z. Rough sets. *International Journal of Parallel Programming*, 1982, vol. 11, no. 5, pp. 341-356.

[12] SEARCÓID, M, O. Metric Spaces, *Springer Undergraduate Mathematics Series*, 2007, 304 p., ISBN 1-84628-369-8.

[13] BUREŠ, J.; BARTONĚK, D.; OPATŘILOVÁ, I. *Klasifikace údajů o uložení plynárenských zařízení pod určitými typy povrchů terénu*, 2013, Závěrečná zpráva o řešení pilotního projektu č. 9413000195 (HS123570212101), FAST VUT v Brně, 32 str.

[14] BUREŠ, J.; BARTONĚK, D.; OPATŘILOVÁ, I. *Analýza povrchů nad plynovody RWE na území ČR*, 2014, Závěrečná zpráva o řešení projektu AdMaS ED2.1.00/03.0097 (HS1235702121200), FAST VUT v Brně, 30 str.