Original scientific paper | Received: 20-08-2019 | Accepted: 13-10-2019

Solving Big GIS Projects on Desktop Computers

Dalibor BARTONĚK

Brno University of Technology, Faculty of Civil Engineering; Veveří 331/95, 602 00 Brno, Czech Republic
bartonek.d@fce.vutbr.cz

Abstract: We are witnessing great developments in digital information technologies. The situation encroaches on spatial data, which contain both attributive and localization features, and this determines their position unequally within an obligatory coordinate system. These changes have resulted in the rapid growth of digital data, significantly supported by technical advances regarding the devices which produce them. As technology for making spatial data advances, methods and software for big data processing are falling behind. Paradoxically, only about 2% of the total volume of data is actually used (Čerba 2017). Big data processing often requires high computation performance hardware and software. Only a few users possess the appropriate information infrastructure. The proportion of processed data would improve if big data could be processed by ordinary users. In geographical information systems (GIS), these problems arise when solving projects related to extensive territory or considerable secondary complexity, which require big data processing.

This paper focuses on the creation and verification of methods by which it would be possible to process effectively extensive projects in GIS supported by desktop hardware and software. It is a project regarding new quick methods for the functional reduction of the data volume, optimization of processing, edge detection in 3D and automated vectorization.

Keywords: geographic information system (GIS), big data, parallel computing, classification, optimization, filtration, point cloud

1 Introduction

1.1 Motivation

We are currently witnessing massive development in digital information technology. This trend can be observed at all levels – individual, institutional, state administration and local self-government, etc. The process described has a static side – in almost all areas there is extensive digitization of analogue data and dynamic part communication links. As a result of these changes, there has been a sharp increase in digital data, which is also supported by the fact that the technological advancement of equipment for producing this data is increasing significantly. These devices include a number of built-in features that simplify operation and streamline deployment. Today, almost every user, after the necessary training, can use data collection equipment in the field. At the same time, as the technology of spatial data acquisition technology has increased, the methods and software for processing large volumes of data have stagnated somewhat.

The ratio between the total volume of data and the amount processed would improve if large data users could also process large data. Practically, this means designing methods and workflows so that large data can be processed with the support of common technical and software resources.

In geographic information systems (GIS), the problems described are encountered when dealing with projects with extensive territorial links or significant secondary complexity that ultimately lead to the processing of large data. Secondary complexity in this case means that the volume of primary data is relatively small, but the solution depends on the combination of input data. An example might be searching for the Hamiltonian path in a graph where it is necessary to search a large state space.

Izvorni znanstveni rad | Primljeno: 20-08-2019 | Prihvaćeno: 13-10-2019

Provođenje velikih projekata u GIS-u na stolnim računalima

Dalibor BARTONĚK

Brno University of Technology, Faculty of Civil Engineering; Veveří 331/95, 602 00 Brno, Češka Republika
bartonek.d@fce.vutbr.cz

Članak je predan na engleskom jeziku. Na hrvatski ga je preveo V. Lapaine.
The paper was submitted in English. It was translated into Croatian by V. Lapaine.

Sažetak: Svjedočimo velikim razvojnim promjenama u digitalnim informacijskim tehnologijama. Ta situacija zadire u prostorne podatke koji sadrže atributna i lokalizacijska obilježja, a to nejednako određuje njihov položaj unutar obaveznog koordinatnog sustava. Te su promjene uzrokovale ubrzan rast digitalnih podataka, što u velikoj mjeri podržava tehnološki napredak uređaja koji takve podatke stvaraju. Dok se tehnologija dobivanja prostornih podataka razvija, metode i softver za obradu velikih podataka zaostaju. Paradoksalno je da se upotrebljava samo 2% ukupne količine podataka (Čerba 2017). Obrada velikih podataka često zahtijeva moćan hardver i softver i samo malen broj korisnika posjeduje odgovarajuću informatičku infrastrukturu. Razmjer obrađenih podataka povećao bi se kad bi obični korisnici mogli obrađivati velike podatke. U geografskim informacijskim sustavima (GIS) takvi problemi nastaju pri provođenju projekata koji pokrivaju velik teritorij ili koji imaju značajnu sekundarnu složenost, što zahtijeva obradu velike količine podataka.
Ovaj rad ima u žarištu stvaranje i provjeru metoda kojima bi se stolnim računalima i softverom mogli obrađivati podatci dobiveni u velikim GIS-projektima. Riječ je o novim brzim metodama za funkcionalno smanjenje količine podataka, optimizaciju obrade, otkrivanje rubova u tri dimenzije te automatsku vektorizaciju.

Ključne riječi: geografski informacijski sustav (GIS), veliki podatci, paralelno računanje, klasificiranje, optimiranje, filtriranje, oblak točaka

1. Uvod

1.1. Motivacija

Svjedočimo velikim razvojnim promjenama u digitalnim informacijskim tehnologijama. Taj je trend moguće uočiti na svim razinama društvenog života – individualnom, institucionalnom, ali i na razini državne administracije i vlade. Opisani proces ima statičnu stranu – u gotovo svim područjima prisutna je opsežna digitalizacija analognih podataka i dinamički dio komunikacijskih veza. Te su promjene uzrokovale veliko povećanje količine digitalnih podataka, što potvrđuje i činjenica da je uvelike tehnološki napredovala oprema za dobivanje takvih podataka. Ti uređaji sadrže velik broj ugrađenih mogućnosti koje pojednostavnjuju rad. Gotovo svaki korisnik, nakon neophodne obuke, može upotrebljavati takvu opremu za dobivanje podataka na terenu. Dok je tehnologija dobivanja prostornih podataka zabilježila velik rast, metode i softver za obradu velikih količina podataka u određenoj mjeri stagniraju.

Omjer između količine obrađenih podataka i njihove ukupne veličine povećao bi se kad bi korisnici velikih podataka također mogli obrađivati takve podatke. U praksi bi to značio razvoj metoda i toka rada kojima bi se veliki podatci mogli obrađivati uz pomoć uobičajene tehničke i softverske opreme.

U geografskim informacijskim sustavima (GIS) na opisane probleme nailazimo kad imamo posla s projektima koji obuhvaćaju velik teritorij ili su visoke sekundarne složenosti, što u krajnjem slučaju dovodi do obrade velikih podataka. U ovom se kontekstu sekundardna složenost odnosi na pojavu kad je količina primarnih podataka u okviru projekta relativno mala, no rezultat rješenja ovisi o kombinaciji ulaznih podataka. Kao primjer možemo uzeti traženje hamiltonskog puta na grafu, gdje je potrebno pretražiti veliki prostor.

1.2 Related works

The following methods are used for processing large data:

1. *Optimum process control (intelligent management)*

   This is, for example, data preprocessing in the form of special-purpose filtering with the aim of reducing the amount of data, or using an intelligent interface to select relevant data from large databases (Arra 2003).

2. *Application of special hardware or software accelerators*

   This is a special architecture of HW or SW that is adapted to the task (Allombert et al. 2014).

3. *Network solution (cloud computing, Hadoop, Apache Spark, MapReduce)*

   An example of this approach is (Bajcsy et al. 2013).

4. *Virtualization of hardware or software resources*

   This method consists of assigning tasks to free resources in a given system (Kim et al. 2015).

5. *Using ontology, the semantic Web*

   This is a special approach mainly in design and conceptual modeling (Dengel 2009).

6. *Use of parallelism with next concurrent processing*

   Using parallelism to speed up computation is one of the oldest and most widely used methods. The biggest problem in this case is the allocation of tasks to resources (processing units) see (Dixon 2015).

7. *Use of machine learning and artificial intelligence, e.g. classification of input data*

   Machine learning is another way to simplify the processing of large data volumes. It is mainly used for the purpose-specific selection of input data, which reduces their volume for further processing (Elliott et al. 2015).

8. *Deployment of supercomputers (High Performance Computing)*

   Perhaps the easiest way to ensure quick calculation is to use special supercomputers, that is, high-performance computers (Elster 2002).

9. *Using optimization methods to reduce processing complexity*

   Optimization methods can be used either for the purposeful selection of input data, thus reducing their volume, or at the processing stage, when selecting a suitable algorithm can speed up the calculation – see (Byun et al. 2016).

10. *Using an n-dimensional approach to a given task (problem transformation)*

    The method consists of transforming the problem into a lower dimension, solving the problem in that dimension and transforming the result into the original (higher) dimension. This method reduces the dimensions of the data structure and the subsequent processing can then be performed using linear algorithms (She et al. 2011).

11. *Pipeline processing*

    This approach is used to solve a large number of tasks of the same or similar structure. The essence of this method is the division of tasks into sub-parts, which can be of different characters, and the subsequent processing of each part by a specialized functional unit. The functional units then operate in parallel, thereby overlapping the parts in the processing chain over time and speeding up the overall solution time (Gopu et al. 2016).

12. *Combination of above mentioned approaches*

    Combinations of methods that streamline large-volume data processing are probably the largest group. The most typical method is described in (Dehsangi et al. 2015).

2 Methods of Solution

When solving large-scale projects in GIS, we mainly encounter the following basic characteristics of tasks:

1. Processing of analyses in a large territory, e.g. the entire territory of the Czech Republic (CR), which assumes the use of large data sets, and the results in each locality are relatively independent of each other
2. Tasks with primarily large volumes of input data with high redundancy or with data that are relatively interdependent - for example, terrestrial or aerial scanning
3. Projects with a relatively small amount of primary data but a complex solution algorithm, which is mostly based on searching a large state space consisting of a combination of input data – see task 2 – and finding the Hamiltonian path in the graph.

Based on the above characteristics, we propose the following methods for solving large projects in GIS:

(In the case of 1., where the project domain is a large area, this area will be suitable.)

- Division of input data into segments (territorially bound systems) and using parallelism for further processing. The method was used in the project "Classification of surface over gas pipelines in the CR" – see section 3.1
- If this division is not possible and the whole set of input data has to be worked (e.g. for network analysis), then implement purpose reduction or simplification (optimization) of input data
  - a) Data filtration (this method was used in the case of laser scanning – section 3.2)

1.2. Srodni radovi

Za obrađivanje velikih podataka primjenjuju se sljedeće metode:

1. *Kontrola optimalnih procesa (inteligentno upravljanje)*

   To je, primjerice, predobrada podataka u obliku posebnog filtriranja kako bi se smanjila količina podataka ili upotrebljavanje inteligentnog sučelja za odabir relevantnih podataka iz velikih baza podataka (Arra 2003).

2. *Primjena posebnih ubrzivača hardvera ili softvera*

   To je posebna arhitektura hardvera i softvera prilagođena zadatku (Allombert i dr. 2014).

3. *Mrežno rješenje (računanje u oblaku, Hadoop, Apache Spark, MapReduce)*

   Primjer takvog pristupa opisali su Bajcsy i dr. (2013).

4. *Virtualizacija hardverskih ili softverskih resursa*

   Metoda se sastoji od pridruživanja zadataka slobodnim resursima u danom sustavu (Kim i dr. 2015).

5. *Upotreba ontologije, semantički web*

   Riječ je o posebnom pristupu, koristi se uglavnom u dizajnu i konceptualnom modeliranju (Dengel 2009).

6. *Upotreba paralelizma sa sljedećom istovremenom obradom*

   Upotreba paralelizma za ubrzanje računanja jedna je od najstarijih i najviše primjenjivanih metoda. U tom je slučaju najveći problem raspodjela zadataka resursima, odnosno jedinicama obrade (Dixon 2015).

7. *Upotreba strojnog učenja i umjetne inteligencije, npr. klasifikacija ulaznih podataka*

   Strojno je učenje još jedan način pojednostavljivanja obrade velikih količina podataka. Ponajviše se primjenjuje za odabir ulaznih podataka za određenu svrhu, što smanjuje njihovu količinu za daljnju obradu (Elliot i dr. 2015).

8. *Upotreba superračunala (računanje visokih performansi)*

   Vjerojatno je najjednostavniji način osiguravanja brzog računanja upotreba superračunala, tj. računala visokih performansi (Elster 2002).

9. *Upotreba metoda optimizacije za smanjenje složenosti obrade*

   Metode optimizacije mogu se primijeniti za odabir ulaznih podataka i smanjenje njihovog obima ili u fazi obrade kad odabir prikladnog algoritma može ubrzati računanje (Byun i dr. 2016).

10. *Upotreba n-dimenzionalnog pristupa zadatku (transformacija problema)*

    Metoda se sastoji u pretvaranju problema u nižu dimenziju odgovarajućom transformacijom, rješavanjem problema u toj dimenziji i transformiranjem rezultata u početnu (višu) dimenziju. Time se smanjuju dimenzije strukture podataka, a naknadnu je obradu moguće provesti uz pomoć linearnih algoritama (She i dr. 2011).

11. *Obrada cjevovoda*

    Taj se pristup primjenjuje za rješavanje velikog broja zadataka slične strukture. Zadatci se dijele na manje dijelove koji mogu biti različitog karaktera i daljnju obradu svakog dijela vrši specijalizirana funkcionalna jedinica. Te funkcionalne jedinice djeluju paralelno i preklapaju dijelove u lancu obrade, što skraćuje ukupno vrijeme obrade (Gopu i dr. 2016).

12. *Kombinacija opisanih pristupa*

    Vjerojatno najveću grupu metoda čine one koje kombiniraju metode koje pojednostavnjuju obradu velikih količina podataka. Najtipičniji su primjeri opisani u radu Deshangija i dr. (2015).

2. Metode rješavanja

Pri izvođenju velikih projekata u GIS-u uglavnom možemo utvrditi sljedeća osnovna svojstva zadataka:

1. obrada analiza na velikom teritoriju (npr. cijeli teritorij Češke), što podrazumijeva upotrebu skupova velikih podataka pri čemu su rezultati na svakoj lokaciji relativno nezavisni jedni od drugih
2. zadatci s prvenstveno velikim količinama ulaznih podataka velike redundantnosti ili s podatcima koji su relativno međuovisni, na primjer terestričko ili zračno snimanje
3. projekti s relativno malo primarnih podataka, ali sa složenim algoritmom rješenja koji se uglavnom temelji na traženju velikog prostora ulaznih podataka (vidi prethodno opisan zadatak – traženje hamiltonskog puta na grafu).

Na temelju navedenih svojstava predlažemo sljedeće metode izvođenja velikih projekata u GIS-u:

(U slučaju 1, kad je domena projekta veliko područje, to će područje biti prikladno.)

- Podjela ulaznih podataka u dijelove (teritorijalno vezane sustave) i upotreba paralelizma za daljnju obradu. Ta je metoda primijenjena u projektu “Klasifikacija površine iznad plinovoda u Češkoj” (vidi poglavlje 3.1.).
- Ako takva podjela nije moguća i potrebno je obraditi cijeli skup ulaznih podataka (npr. za analizu mreže), tada se smanjenju ili pojednostavnjuju (optimiraju) ulazni podatci:
  a) filtriranje podataka (Ta je metoda primijenjena u slučaju laserskog skeniranja, vidi poglavlje 3.2.)
  b) smanjivanje dimenzije problema u manju dimenziju i potom transformacija natrag u veću dimenziju (Ta je metoda primijenjena pri traženju optimalne staze mjerene s pomoću metode RTK (Real Time Kinematic), vidi poglavlje 3.3.)

b) Reducing the dimension of the problem, i.e. the solution in the lower dimension, and then transforming it back to the higher dimension. This method was applied in search of the optimum track when measured by RTK (Real Time Kinematic) – see section 3.3.

The situation is illustrated in Figure 1. Both methods require a test task (pilot project) in order to make a qualified estimate of the relevant parameters (solution time, resource capacity and research team).

3 Experimental Projects

3.1 Classification of the gas-pipeline surface area in the CR

3.1.1 Task formulation

The aims of the project were:

a) To analyze surfaces over RWE gas pipelines in the Czech Republic (CR), i.e. process classification of data on gas facility storage under certain types of terrain surfaces in order to determine the reproduction values of gas facilities (gas pipelines) and estimate the costs of building new networks

b) To perform the surface analysis in three variants for high pressure gas pipelines (HP), main pipelines and domestic connections

b1) 2 classes of surface type: paved and unpaved

b2) 5 classes of surface type: asphalt, main road, local road, unbound, unknown

b3) 10 classes of surface type: forest, grassland, bare soil, asphalt, roof-tile, roof - straight, shadow, main road, local road, road

c) To produce a structured table in XLSx format with classified data of surface types above the pipeline route and graphical output of structured classified data - SHP format for high pressure gas pipelines, main lines of local networks and domestic connections

d) To analyze the 12 regions of the CR, i.e. a total of 188 municipalities with extended powers (ORP).

3.1.2 Theoretical foundations

Let the non-empty $U \neq \emptyset$ be the universe set and X the subset ($X \subseteq U$). Set U represents in our case the whole solved area while subset X is a part of the solved area. The equivalence relation R divides set U into subsets $U / R = \{X_1, X_2, ..., X_n\}$ so that for each i, j holds:

1) $X_i \subseteq U$, $X_i \neq \emptyset$ (all subsets are non-empty)

2) $X_i \cap X_j \neq \emptyset$ (the intersection of all subsets is empty)

3) $\cup_{i=1,2,...,n} X_i = U$ (the union of all subsets is just the whole set U).

The territorial division of the CR according to its administrative arrangement, which divides the territory into lower administrative units (regions, districts, etc.), can be considered as the equivalence relation R. If the conditions under 1) to 3) are fulfilled, each sub-territory has the character of class X_i in terms of set theory.

Suppose there is another relation $S \subseteq R$ such that it defines the decomposition of subsets $X_i \subseteq X$ into equivalent classes X_{ij} with the same properties in reference to 1) to 3):

$$X_i / S = \{X_{i1}, X_{i2}, ..., X_{im}\} \quad \text{for all } i = 1, 2,.., n\,. \tag{1}$$

Then the system of relations $\{R, S\}$ with sets U, X represents a hierarchical decomposition of set U. In our case, relation S defines a further division of territorial units, represented by classes X_i into lower territorial units according to the subject (GO) principle. The criterion for this division is whether geographic objects (GO) are relevant in the given area, i.e. elements that are the main subject of processing in the given GIS project (in our case, gas-pipelines).

Relation S thus divides the areals into subclasses X_{ij} for j = 1, 2,..., m, according to formula (1). Hierarchical decomposition of $X \subseteq U$ into equivalent classes is the result of a combination of distribution criteria according to:

1) the administrative structure of the territory at the first hierarchical level

2) subject arrangement, i.e. according to the existence of GO at the second hierarchical level.

Let X be a set of GOs in the real world and set Y an image of set X in digital geo-database. Then the mapping φ: $X \rightarrow Y$ must have the following properties:

1) φ is unambiguous

2) φ is continuous

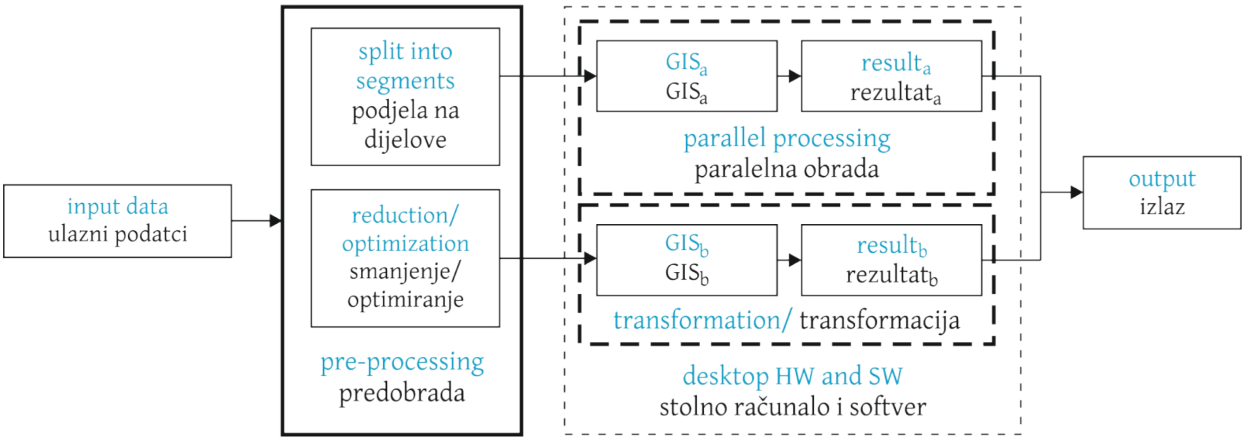
3) there is an inverse mapping φ^{-1}: $Y \rightarrow X$, which is also continuous.

Then the mapping φ is homeomorphism, which is a basic property of topological relationships between GOs. In our case, the GO (set X) of the input data set and set Y are images of those data sets in the project data warehouse.

Now let us define the solution time T_{tot}: for the entire project

$$T_{tot} = T_{rez} + T_{pa} + T_{pm} \ , \tag{2}$$

where T_{rez} is the overhead time, T_{pa} is the time of the automated and T_{pm} time of the manual processing.



Slika 1. Metode obrade velikih podataka s pomoću stolnog računala i softvera.
Fig. 1 Methods of big data processing on desktop hardware and software.

Ta je situacija prikazana na slici 1. Obje metode zahtijevaju probni zadatak (pilot projekt) kako bi se mogli dobro procijeniti relevantni parametri (potrebno vrijeme, kapacitet resursa i istraživački tim).

3. Eksperimentalni projekti

3.1. Klasifikacija površine plinovoda u Češkoj

3.1.1. Formulacija zadatka

Cilj projekta bio je:

a) analizirati površine iznad plinovoda tvrtke RWE u Češkoj, tj. obrada klasifikacije podataka o skladištenju plinskih postrojenja ispod određenih tipova površina kako bi se moglo utvrditi reprodukcijske vrijednosti plinskih postrojenja (plinovoda) i procjena troškova potrebnih za izgradnju novih mreža
b) analizirati tri vrte površina za visokotlačni plinovod (high pressure – HP), glavne plinovode i veze s kućama:
   1) dvije klase površina – asfaltirana i neasfaltirana
   2) pet klasa površina – asfalt, glavna cesta, lokalna cesta, nevezana, nepoznata
   3) deset klasa površina – šuma, pašnjak, golo tlo, asfalt, crijep, ravan krov, sjena, glavna cesta, lokalna cesta, cesta.
c) izraditi strukturiranu tablicu u formatu XLSx s klasificiranim podatcima tipova površina iznad plinovoda i grafički izlaz strukturiranih klasificiranih podataka – format SHP za visokotlačne plinovode, glavne linije lokalnih mreža i kućnih veza
d) analizirati 12 regija u Češkoj, tj. ukupno 188 općina s proširenim ovlastima (municipalities with extended powers – ORP).

3.1.2. Teorijske osnove

Neka je $U \neq \varnothing$ neprazni univerzalni skup i X podskup ($X \subseteq U$). Skup U je u našem slučaju čitavo riješeno područje, a podskup X je dio riješenog područja. Relacija ekvivalencije R dijeli skup U u podskupove $U / R = \{X_1, X_2, ..., X_n\}$ tako da za svaki i, j vrijedi:

1) $X_i \subseteq U$, $X_i \neq \varnothing$ (svi skupovi su neprazni)
2) $X_i \cap X_j \neq \varnothing$ (presjek svih podskupova je prazan)
3) $\cup_{i=1,2,...,n} X_i = U$ (unija svih podskupa je cijeli skup U).

Teritorijalna podjela Češke prema upravi dijeli teritorij na niže administrativne jedinice (regije, okruge ...) i može se smatrati relacijom ekvivalencije R. Ako su ispunjeni uvjeti 1) – 3), svaki podteritorij ima karakter razreda X_i u smislu teorije skupova.

Neka je druga relacija $S \subseteq R$ takva da definira dekompoziciju podskupova $X_i \subseteq X$ na ekvivalentne razreda X_{ij} s istim svojstvima 1) – 3):

$$X_i / S = \{X_{i1}, X_{i2}, ..., X_{im}\} \quad \text{za sve } i\text{=1, 2,.., } n \ . \qquad (1)$$

Tada sustav relacija $\{R, S\}$ zajedno sa skupovima U, X predstavlja hijerarhijsku dekompoziciju skupa U. U našem slučaju relacija S definira daljnju podjelu teritorijalnh jedinica X_i u niže teritorijalne jedinice prema principu geografskih objekata. Kriterij za tu podjelu jest relevantnost geografskih objekata (GO) u danom području, tj. jesu li to elementi koji su glavni predmet obrade u danom projektu GIS-a (u našem je slučaju riječ o plinovodu).

Dakle, relacija S dijeli područja u podrazrede X_{ij} za $j = 1, 2,..., m$ prema formuli (1). Hijerarhijska dekompozicija $X \subseteq U$ u ekvivalentne razrede rezultat je kombinacije kriterija raspodjele u:

Assuming the division of the whole project into *n* sub-parts (equation 1), then equation (2) has the form:

$$T_{tot} = T_{rez} + \sum_{i=1}^{n} (t_{pai} + t_{pmi}) = \min \ . \tag{3}$$

For efficient processing, the overhead time must not be greater then the sum of the automated and manual processing time, i.e. the ratio $T_{pa} + T_{pm}$ and T_{rez} should be at the maximum (at least greater then 1):

$$\eta = \frac{\sum_{i=1}^{n} (t_{pai} + t_{pmi})}{T_{rez}} = \max \ . \tag{4}$$

Equations (3) and (4) represent the purpose functions to be optimized. On the basis of this coefficient, it will be decided how the input data sets of the project will be divided according to the sub-territorial units within the administrative structure of the CR (relation R, i.e. decomposition in the 1st hierarchical level).

Next, let us define the coefficient of surface detail:

$$\delta = \frac{P_{tot}}{P_{\mathrm{det}}}, \tag{5}$$

where P_{tot} is the total area of sub-territory (the decomposition to the equivalence of relation *R*) and P_{det} is the area of the smallest detail contained in the input data. This coefficient affects the selection of suitable hardware and software in terms of computing performance.

Let *S* be the dataset of project results, *R* is datasets of the same type as *S* but *R* attributes are of higher quality then *S* attributes. Then we can refine the results by rewriting the values of the *S* attributes according to the attributes of the *R* set:

$$\text{if } S \cap R \neq \emptyset \ \text{ then } a_r \rightarrow a_s \, , \tag{6}$$

where $\cap$ is spatial overlap of layers, a_s is the attribute $s \in S$ and a_r is attribute $r \in R$.

3.1.3 Method of solution

The basic approach was to divide the whole project into individual parts and use parallelism to speed up processing.

- Optimal processing unit (subproject) (equation 1) in terms of efficiency and reliability
- Coefficient δ, which determines the data extent of the project (equation 2)
- Overhead time value, automated and manual processing times, and efficiency (equations 3, 4, 5)
- Selecting a reference data set *R* to improve the quality of the results (equation 6)
- Selection of suitable software
- Selection of suitable available technical means (number, configuration, etc.)

These parameters were determined by a pilot project of the same type as the main GIS project in a limited area. Parameters for the main project were then determined by extrapolation based on the ratio of the extent of the modeled areas.

To ensure the maximum degree of automation of the whole process, it is necessary to use either existing software or create your own application. In our case, a set of Python scripts was created.

An essential prerequisite for success is the optimum composition of the team. In practice, the following minimum composition of 3 categories of researchers has proven useful:

- 1 manager to manage the whole project
- 1 GIS programmer who creates the above procedures for automated processing
- 1 person to inspect all results manually and visually.

3.1.4 Experimental results

The proposed method was tested in the project for the classification of surfaces under which gas facilities are stored in the CR. The authors solved this project for GasNet Ltd., which is part of the RWE Group in the CR. The input data were orthophoto data sets with a resolution of 25 cm/pixel, road layers ZABAGED ČR (fundamental base of geographic data) and vector files of underground utility lines. The territorial coverage of the CR with an area of 64,350 km^2 involved large data processing with a total data volume of 500 GB.

For a qualified estimation of parameter values in equations (1) to (6), a pilot project on a modeled territory of the city of Brno with an area of approximately 250 km^2 was solved. The following strategy was set out in this test project:

- The optimum territorial unit (segment) of the municipality with extended competence (ORP) was chosen - see equation (1), whose average area was 368 km^2, while the average amount of data for 1 ORP was 2.5 GB. Relationships (3) and (4) also affected this decision.
- The optimum ratio η (equation 4) for the total number of data x = 500 GB was calculated from the values: n = 200, x_i = 2.5 GB.
- The optimization of data according to the topic was performed by trimming the orthophoto layer with a

1) administrativnu strukturu teritorija na 1. hijerarhijskoj razini
2) uređenje predmeta, tj. prema postojanju GO na 2. hijerarhijskoj razini.

Neka je X skup GO-ova u stvarnom svijetu, a skup Y slika skupa X u digitalnoj bazi geopodataka. Tada preslikavanje φ: $X \rightarrow Y$ mora imati sljedeća svojstva:

1) φ je nedvosmisleno (*unambiguous*)
2) φ je neprekidno
3) postoji inverzno preslikavanje φ^{-1}: $Y \rightarrow X$ koje je također neprekidno.

Tada je preslikavanje φ homeomorfizam, što je osnovno svojstvo topoloških odnosa među GO-ovima. U našem su slučaju GO-ovi (skup X) elementi skupa ulaznih podataka, a skup Y su slike tih skupova podataka u skladištu projektnih podataka.

Definirajmo vrijeme rješenja T_{tot} za cijeli projekt:

$$T_{tot} = T_{rez} + T_{pa} + T_{pm} , \tag{2}$$

gdje je T_{rez} ukupno vrijeme (*overhead time*), T_{pa} vrijeme automatizirane i T_{pm} vrijeme ručne obrade.

Pod pretpostavkom podjele cijelog projekta u n manjih dijelova (jednadžba 1), jednadžba (2) ima sljedeći oblik:

$$T_{tot} = T_{rez} + \sum_{i=1}^{n} (t_{pai} + t_{pmi}) = \min . \tag{3}$$

Za učinkovitu obradu nužno je da ukupno vrijeme nije veće od zbroja vremena automatizirane i ručne obrade, tj. da je omjer $T_{pa} + T_{pm}$ i T_{rez} najveći mogući (barem veći od 1):

$$\eta = \frac{\sum_{i=1}^{n} (t_{pai} + t_{pmi})}{T_{rez}} = \max . \tag{4}$$

Jednadžbe (3) i (4) predstavljaju funkcije koje je potrebno optimirati. Na temelju tog koeficijenta bit će odlučeno prema kojim će podteritorijalnim jedinicama unutar administrativne strukture Češke (relacija R, tj. raspad na 1. hijerarhijskoj razini) biti podijeljeni ulazni skupovi podataka projekta.

Definirajmo sada koeficijent površinskih detalja:

$$\delta = \frac{P_{tot}}{P_{\det}}, \tag{5}$$

gdje je P_{tot} ukupna površina podteritorija (dekompozicija u relaciji ekvivalencije R), a P_{det} površina najmanjeg detalja u ulaznim podatcima. Taj koeficijent utječe na izbor odgovarajućeg hardvera i softvera u smislu računskih performansi.

Neka je S podskup rezultata projekta, R skupovi podataka istog tipa kao S, ali atributi R-a su atributi veće kvalitete od atributa S. Tada možemo precizirati rezultate tako da ponovno napišemo vrijednosti atributa S u skladu s atributima R:

$$\text{ako je } S \cap R \neq \emptyset, \text{ onda je } a_r \rightarrow a_s , \tag{6}$$

gdje je $\cap$ prostorno preklapanje slojeva, a_s atribut $s \in S$ i a_r atribut $r \in R$.

3.1.3. Metoda rješavanja

Osnovni pristup bio je podjela cijelog projekta na dijelove i upotreba paralelizma za ubrzanje obrade:

- jedinica optimalne obrade (podprojekt, jednadžba 1) u smislu efikasnosti i pouzdanosti
- koeficijent δ koji određuje raširenost podataka projekta (jednadžba 2)
- vrijednost ukupnog vremena, vremena automatizirane i ručne obrade i efikasnost (jednadžbe 3, 4, 5)
- izbor referentnog skupa podataka R za poboljšanje kvalitete podataka (jednadžba 6)
- izbor odgovarajućeg softvera
- izbor odgovarajućih tehničkih sredstava (broj, konfiguracija...).

Ti su parametri određeni u pilot projektu istoga tipa kao glavni projekt GIS-a u ograničenom području. Potom su parametri glavnog projekta određeni ekstrapolacijom na temelju omjera opsega modeliranih područja.

Kako bi se osigurao najviši stupanj automatizacije cijelog procesa, potrebno je primijeniti postojeći softver ili vlastitu aplikaciju. U našem smo slučaju osmislili vlastite skripte u Pythonu.

Optimalni sastav tima osnovni je preduvjet za uspjeh. U praksi se korisnim pokazao sljedeći sastav triju kategorija istraživača:

- jedan menadžer cijelog projekta
- jedan programer GIS-a koji programira navedeni postupak za automatsku obradu
- jedna osoba za ručni i vizualni pregled svih rezultata.

3.1.4. Eksperimentalni rezultati

Predložena metoda provjerena je u projektu klasifikacije površina ispod kojih se nalaze plinska postrojenja u Češkoj. Autori su proveli ovaj projekt za GasNet Ltd. koji je dio tvrtke RWE Group u Češkoj. Ulazni podatci bili su skupovi ortofota razlučivosti od 25 cm po pikselu, slojevi ceste ZABAGED ČR (temeljna osnova geografskih podataka) i vektorskih

1 m buffer zone along the gas pipeline in consultation with the RWE company. This reduced the amount of data from 500 GB to 230 GB.

- After each partial automated ORP processing, the partial results were checked manually.
- Overhead, automated and manual processing times were determined experimentally according to the pilot project and after substituting the values in relation (4) we obtained relation (7):

$$\eta = \frac{\sum_{i=1}^{n} 0.02e^{0.4x_i}}{\sum_{i=1}^{n} 0.002x_i^2 + 0.003x_i + 0.04} \qquad (7)$$

$$T_{rez} = 0.002x^2 + 0.003x + 0.04 \ , \ T_{pa} = 0.02e^{0.4x}$$

- Based on relation (7), the duration of the whole project was estimated to be 4 months.
- The ratio of manual to automatic processing was 55% : 45%.
- Implementation of the project was performed on PC Integra 7025 computers (parameters: Intel Core i5, 3.8 GHz, 16 GB RAM, NVIDIA GTX650, 2 GB, HD SDD and VelociRaptor), which were connected to the 100 Mb/s computer network. The main criterion for this decision was the coefficient value $\delta = 10^9$. In the test project, it was found that a higher coefficient value δ would mean the use of more powerful computers.
- Data analysis was supported with Python's own procedures using ESRI libraries in ArcGIS 10.0.
- The maximum likelihood (MLC) method in ArcGIS 10.0 was selected for raster image classification (orthophoto). However, this classification only showed an average success rate of 70%, which was insufficient for the contracting authority. Therefore, the input data was supplemented with a data set of 9 communication layers, which was part of ZABAGED ČR. Using these reference layers, the classification result was refined (equation 6, Figure 2).
- The results showed a high efficiency of technology and low classification error rate in the range of 2% - 3% in the whole modeled area of the CR.

3.1.5 Conclusions

A new methodology was proposed for solving large GIS projects and confirmed the possibility of processing the using common hardware and software resources. The methodology was based on data-driven procedures implemented in the Python scripting language with the support of ESRI libraries.

The main features of the proposed methodology are:

- Optimal division of the project into sub-projects according to the territorial principle
- Parallel processing of partial projects using suitable hardware and software
- Optimizing data volume by topic
- Ensuring the quality of results by checking the outputs of individual sub-projects manually
- Increasing the quality of results (reduction of error rate) by means of a reference data set - in our case ZABAGED ČR.

The prerequisites for a successful project are:

- Project quality management, coordination of subproject phases, interoperability and interchangeability of the research team in solving individual processing tasks
- Solving a model pilot project of the same type as the main project in a limited space to obtain the necessary parameter values to determine the strategic process.

The technology is of a general nature and can be used to classify the surfaces of utilities such as water, gas, pipelines, power grids, etc. The project has been described in publications (Bartoněk et al. 2014, 2015).

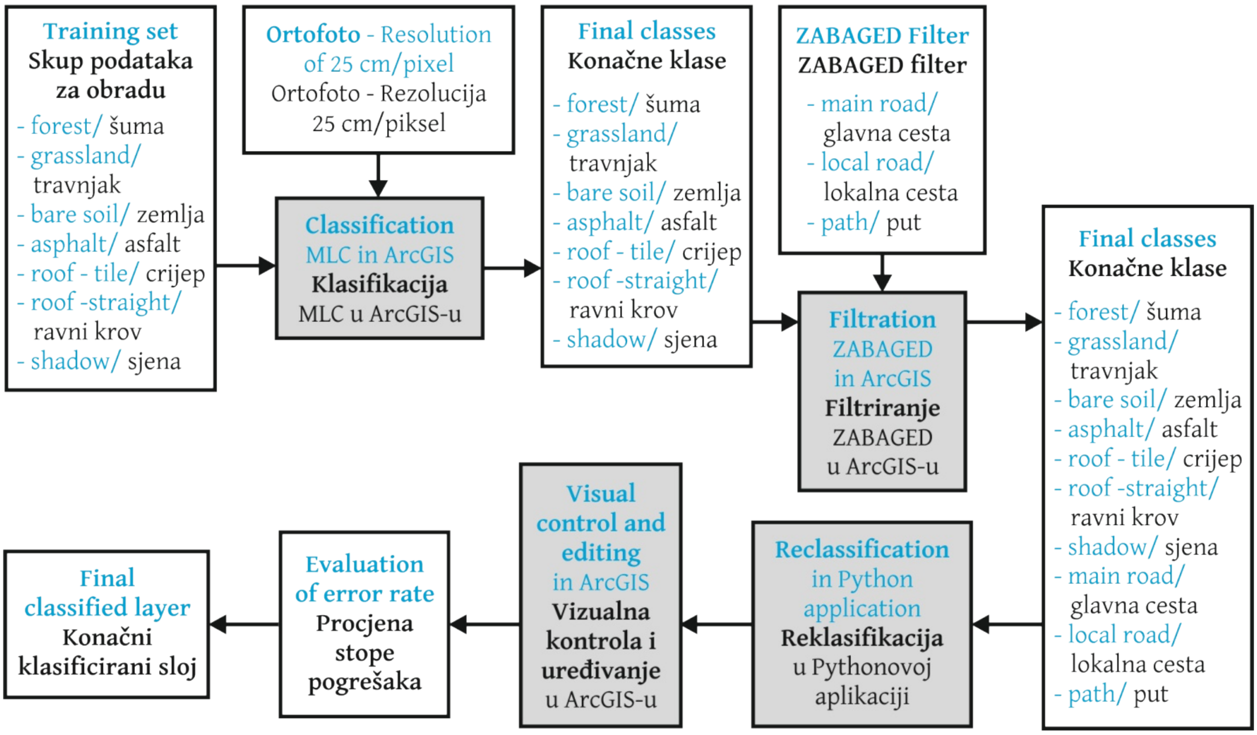
3.2 Processing Cloud Points From Laser Scanning

3.2.1 Characteristics of the problem and method of solution

In this project, edge detection from ground-based laser scanning data was solved. The essence of the task is to select a set of points that positively determine the contours of buildings at the point of contact with the terrain (ground plan).

The proposed evaluation procedure is based on parametric filtering of the input cloud over objects on the updated digital form of the cadastral map, with the aim of significantly reducing the volume of data and the subsequent detection of floor plan edges of buildings and other objects.

The aim is to propose a suitable method for the automated updating of selected elements of technical and cadastral maps. The selected elements are buildings and structures. The essence of the update is to identify differences in the ground plans of buildings or structures between the actual state and how they are drawn on the map. The input data are a cloud of points representing the current state of the terrain and a map in digital form for the location. In case of non-compliance, there are two possibilities:


Training set
Skup podataka za obradu
- forest/ šuma
- grassland/ travnjak
- bare soil/ zemlja
- asphalt/ asfalt
- roof - tile/ crijep
- roof -straight/ ravni krov
- shadow/ sjena
Ortofoto - Resolution of 25 cm/pixel
Ortofoto - Rezolucija 25 cm/piksel
Classification MLC in ArcGIS
Klasifikacija MLC u ArcGIS-u
Final classes
Konačne klase
- forest/ šuma
- grassland/ travnjak
- bare soil/ zemlja
- asphalt/ asfalt
- roof - tile/ crijep
- roof -straight/ ravni krov
- shadow/ sjena
ZABAGED Filter
ZABAGED filter
- main road/ glavna cesta
- local road/ lokalna cesta
- path/ put
Filtration ZABAGED in ArcGIS
Filtriranje ZABAGED u ArcGIS-u
Final classes
Konačne klase
- forest/ šuma
- grassland/ travnjak
- bare soil/ zemlja
- asphalt/ asfalt
- roof - tile/ crijep
- roof -straight/ ravni krov
- shadow/ sjena
- main road/ glavna cesta
- local road/ lokalna cesta
- path/ put
Reclassification in Python application
Reklasifikacija u Pythonovoj aplikaciji
Visual control and editing in ArcGIS
Vizualna kontrola i uređivanje u ArcGIS-u
Evaluation of error rate
Procjena stope pogrešaka
Final classified layer
Konačni klasificirani sloj

Slika 2. Pojedini koraci za poboljšanje klasifikacije ortofota.
Fig. 2 Individual degrees of the orthophoto classification refinement.

datoteka podzemnih vodova. Teritorijalna pokrivenost Češke s površinom od 64 350 km^2 uključivala je obradu velikih podataka ukupnog obujma od 500 GB.

Pilot projekt na modeliranom teritoriju grada Brna s površinom od otprilike 250 km^2 izveden je kako bi se mogli dobro procijeniti parametri u jednadžbama (1)–(6). U tom je projektu primijenjena sljedeća strategija:

- Kao optimalna teritorijalna jedinica (segment) odabrana je općina s proširenom nadležnošću (municipality with extended competence – ORP), vidi jednadžbu (1), čija je prosječna površina 368 km^2, prosječna količina podataka za 1 ORP je 2,5 GB. Relacije (3) i (4) također su utjecale na ovu odluku.
- Optimalni omjer η (jednadžba 4) za ukupan broj podataka x = 500 GB izračunan je iz vrijednosti: n = 200, x_i = 2,5 GB.
- Optimizacija podataka prema temi postignuta je rezanjem sloja ortofota s tampon zonom od 1 m uzduž plinovoda. Vrijednost je izabrana u konzultaciji s tvrtkom RWE. Time je količina podataka smanjena s 500 GB na 230 GB.
- Nakon svake djelomične obrade općine s proširenom nadležnošću, djelomični su rezultati provjereni ručno.
- Ukupno, automatizirano i ručno vrijeme obrade određeni su eksperimentalno prema pilot projektu. Nakon uvrštavanja vrijednosti u jednadžbu (4), dobivamo jednadžbu (7):

$$\eta = \frac{\sum_{i=1}^{n} 0{,}02e^{0{,}4x_i}}{\sum_{i=1}^{n} 0{,}002x_i^{\,2} + 0{,}003x_i + 0{,}04} \qquad (7)$$

$$T_{rez} = 0{,}002x^2 + 0{,}003x + 0{,}04 \ , \ T_{pa} = 0{,}02e^{0{,}4x}$$

- Na temelju jednadžbe (7) trajanje cijelog projekta procijenjeno je na četiri mjeseca.
- Omjer ručne i automatske obrade je 55% : 45%.
- Projekt je izveden na računalima PC Integra 7025 (Intel Core i5, 3.8 GHz, 16 GB RAM, NVIDIA GTX650, 2 GB, HD SDD i VelociRaptor) spojenima na mrežu brzine 100 Mb/s. Glavni je kriterij za ovu odluku bila vrijednost koeficijenta δ = 10^9. U testnom je projektu dobiveno da bi viša vrijednost koeficijenta δ zahtijevala upotrebu jačih računala.
- Obrada podataka provedena je postupcima u Pythonu uz upotrebu knjižnica ESRI u programu ArcGIS 10.0.

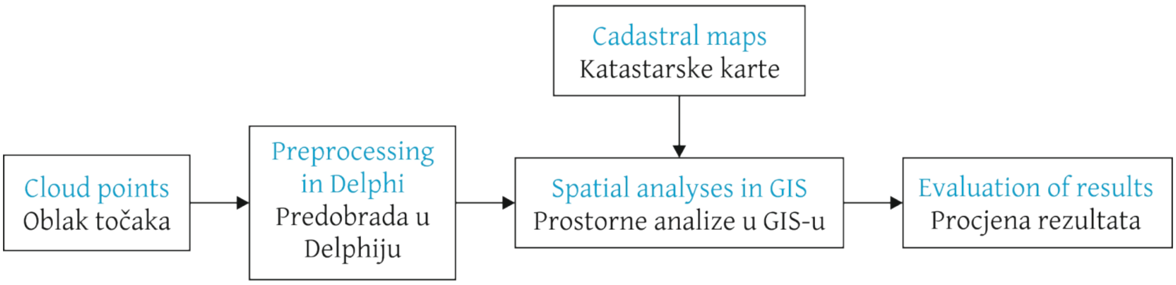


Fig. 3 Concept of detection of changes of selected objects in technical and cadastral maps.
Slika 3. Koncept utvrđivanja razlika označenih objekata na tehničkim i katastarskim kartama.

1. The actual state has more objects than the drawing on the map
2. The actual state has fewer objects than the digital map.

In the first case, we can describe the change detection process by the mathematical model:

$$R = (C \to D) - (S \to B), \tag{8}$$

where

$$D = f_r(C) \tag{9}$$
$$B = f_b(S), \tag{10}$$

where R is the resulting set of objects, set C represents the cloud of points from laser scanning and set S objects on the digital map. Set D is obtained from the point cloud by preprocessing the f_r functions, and by analogy, we generate set B by applying a buffer (function f_b) to set S. The f_r function performs edge detection (contours of buildings and constructions), while function f_b (buffer) is the precision criterion according to Czech standard ČSN 01 3410 (e.g. for the 3rd class of map accuracy the standard deviation of the determination of the breakpoint of the boundary is 14 cm).

In the second case, the change detection process can be described by a similar mathematical model with the same symbology:

$$R = [(C \to D) \to B] - S, \tag{11}$$

where

$$D = f_r(C) \tag{12}$$
$$B = f_b(D). \tag{13}$$

If $R = \{\varnothing\}$, then no changes are detected, if $R \neq \{\varnothing\}$, then there are differences between the current state on the cadastral map and the actual state in the field. A flowchart of the whole process is shown in Figure 3.

In Figure 3, f_b is provided by the ‘spatial GIS analysis’ block, while the f_r function is implemented in the ‘Delphi preprocessing’ block. While the f_b buffer function is well known from spatial analysis, finding a suitable f_r function required a number of experiments. Edge detection based on central discrete convolution has proved to be unsuitable due to its time-consuming demands. The use of point cloud aggregation according to the position in the plane (x, y) with the attribute of the frequency of points in the z-axis (height of the object) and with filtration proved to be more advantageous:

$$\begin{aligned}&\textit{Select } x, y, \textit{count}(z) \textit{ from list group by } x, y\\ &\textit{having count}(z) > p,\end{aligned} \tag{14}$$

where *list* is a list of point cloud coordinates from laser scanning and p is a parameter that specifies the height of the object (according to cloud resolution) in which we run an imaginary plane of cut that separates buildings or structures from other objects in the terrain. Practically we choose $p \geq 2$.

For this purpose, a special program was created in Delphi, which serves to prepare input data – see Figure 3. The program has the following functions:

1. It splits the point input cloud into segments of smaller data volumes so that the data can be processed on common technical and desktop computers
2. It filters point clouds according to (14) in each segment
3. It transforms the output data file into a suitable format for further GIS processing.

3.2.2 Experimental results

The method proposed in the previous section was verified in a selected locality in Brno (Czech Republic). Figure 4a shows a cloud of points around the building;

- Za klasifikaciju rasterskih slika (ortofoto) izabrana je metoda najveće vjerojatnosti (Maximum likelihood method – MLC) u programu ArcGIS 10.0. Međutim, uspješnost te klasifikacije bila je tek 70%, što za naručitelja nije bilo dovoljno. Stoga su ulazni podatci dopunjeni skupom podataka od 9 komunikacijskih slojeva, što je dio ZABAGED ČR. Time je rezultat klasifikacije poboljšan (jednadžba 6, slika 2).
- Rezultati su pokazali visoku učinkovitost tehnologije i nisku stopu pogreške klasifikacije u rasponu 2%–3% na cijelom modeliranom području Češke.

3.1.5. Zaključci

Predložena je nova metodologija za izvođenje velikih projekata GIS-a i provjerena je mogućnost njihove obrade uz pomoć uobičajenih hardverskih i softverskih resursa. Metodologija se temelji na postupcima ugrađenima u jezik Python uz podršku knjižnica ESRI.

Glavna su obilježja predložene metodologije:

- optimalna podjela projekta u podprojekte prema teritorijalnom principu
- paralelna obrada djelomičnih projekata uz pomoć odgovarajućeg hardvera i softvera
- optimizacija obujma podataka prema temi
- osiguranje kvalitete rezultata ručnom provjerom izlaza pojedinih podprojekata
- povećanje kvalitete rezultata (smanjenje stope pogrešaka) upotrebom referentnog skupa podataka – u našem je slučaju to bio ZABAGED ČR.

Preduvjeti su za uspješan projekt:

- upravljanje kvalitetom projekta, koordinacija faza podprojekata, interoperabilnost i zamjenjivost istraživačkog tima u rješavanju pojedinih zadataka obrade
- provedba pilot projekta istog tipa kao glavni projekt kako bi se dobile potrebne vrijednosti parametra i odredilo strateški proces.

Tehnologija je opće prirode i može se primijeniti za klasifikaciju površine postrojenja kao što su vodovodi, plinovodi, cjevovodi, električne mreže... Projekt je opisan u radovima Bartoněka i dr. (2014, 2015).

3.2. Obrada točaka oblaka iz laserskog skaniranja

3.2.1. Svojstva problema i metoda rješavanja

Ovaj se projekt bavio otkrivanjem rubova iz podataka dobivenih laserskim skeniranjem na tlu. Zadatak se svodi na izbor točaka koje određuju obrise zgrada na mjestu gdje one dolaze u kontakt s tlom (tlocrt).

Predloženi se postupak temelji na parametarskom filtriranju ulaznog oblaka preko objekata ažurirane digitalne inačice katastarske karte kako bi se znatno smanjio obujam podataka s naknadnim utvrđivanjem rubova tlocrta zgrada i drugih objekata.

Cilj je predložiti odgovarajuću metodu za automatizirano ažuriranje izabranih elemenata tehničkih i katastarskih karata. Izabrani su elementi zgrade i strukture. Ažuriranje se temelji na utvrđivanju razlika u tlocrtu zgrada ili struktura između stvarnog stanja i onoga što je na karti. Ulazni podatci su: oblak točaka koje prikazuju trenutačno stanje terena i digitalna karta tog mjesta. U slučaju njihovog neslaganja, moguće su dvije mogućnosti:

1. stvarno stanje pokazuje više objekata nego karta
2. stvarno stanje ima manje objekata od digitalne karte.

U prvom slučaju možemo opisati proces utvrđivanja razlike matematičkim modelom:

$$R = (C \to D) - (S \to B), \tag{8}$$

gdje je

$$D = fr(C) \tag{9}$$

$$B = fb(S), \tag{10}$$

gdje je R izlazni skup objekata, skup C predstavlja oblak točaka laserskog skeniranja, a skup S predstavlja objekte na digitalnoj karti. Skup D dobiva se iz točaka oblaka predobradom funkcijom f_r, a po analogiji, skup B dobivamo primjenom koridora (funkcija f_b) na skup S. Funkcija f_r otkriva rubove (obrise zgrada i drugih objekata), funkcija f_b (koridor) je kriterij preciznosti prema češkom standardu ČSN 01 3410 (npr. za 3. razred točnosti karte standardna devijacija određivanja prijelomnih točaka je 14 cm).

U drugom slučaju postupak utvrđivanja promjene može se opisati sličnih matematičkim modelom s istima simbolima:

$$R = [(C \to D) \to B] - S, \tag{11}$$

gdje je

$$D = fr(\mathrm{C}) \tag{12}$$

$$B = fb(\mathrm{D}). \tag{13}$$

Ako je R = {∅}, tada nema promjena, a ako je R ≠ {∅}, tada postoje promjene između katastarske karte i stvarnog stanja na terenu. Blok dijagram cijelog procesa prikazan je na slici 3.

Figure 4b shows cloud points after filtration with cadastral map.

The process is automated to the maximum possible extent in 2 applications:

1. A special program in Borland Delphi
2. Any GIS program (Geomedia, ArcGIS, QGIS).

The proposed method is relatively fast, low on hardware and software and can be used to detect planimetric changes on any digital map in conjunction with laser scanning. The results can be used to update the relevant digital map work. The project was published in an article by (Bureš et al. 2018).

3.3 Optimum GNSS RTK Measurement

3.3.1 Problem description

This project is not primarily about processing large data, but rather solving tasks with large state space. In short, this means that for the resulting solution, it is necessary to search an extensive set of possible combinations of input data. The solution time according to a 'robust' algorithm would take a disproportionately long time on special supercomputers.

In this case, it is not possible to use methods of division of input data according to the territorial principle or to use data filtering in order to reduce their volume. For a correct solution we have to work with all the input data. An example of this problem is finding a minimum path between all points in a given location to be measured by the GNSS - RTK method.

3.3.2 Task formulation

Assignment:

There is a set of spatially distributed objects to measure. The following object attributes are available: ID, position, observation time including tolerance, and set of intervals (date, time) at which it is possible to measure. The position is given by mutual distances between individual objects in length and time units. At each point, the necessary time interval to the next measurement is entered. The measurement can be repeated, or several measurement sets of instruments may be available for measurement.

Task:

Create a measurement itinerary on all objects in terms of:

1. Subject (process of measurement)
2. Economy (process efficiency)

Re. 1. Create an optimum measurement schedule at all points so that the delay due to moving between points or waiting for a suitable measurement interval at given points is minimal.

Re. 2. Determine the total cost of measurement, i.e. the calculation of travel by vehicle and on foot, hourly rate of meters, and rental of instruments. From the calculated cost of measurement, then derive the campaign's effectiveness, i.e. the optimum number of instruments, the number of meters and the appropriate time for measurement.

3.3.3 Theoretical model and method of solution

Let A be a set of objects. Each object $a \in A$ has the following attributes:

- $a.p$ is the position of the object in the given coordinate system
- $a.t$ is the time interval during which an event can occur (even repeatedly) at a given object
- $a.s$ is the time difference from the last event. During this time interval, no event can occur at the object.

Let us take a set of B time intervals, during which events can run at objects $a \in A$ and set of metric values D, whose elements express the distance between individual objects $a \in A$ in length or time units.

Let us assign function f: $A \rightarrow C$, $C \subseteq B$ and function g: $\mathrm{A} \times \mathrm{A} \rightarrow \mathrm{D}$.

Function f assigns to objects $a \in A$ a set of suitable intervals $a.t \in C$, during which events can occur at objects; function g assigns to each pair of objects a_i, $a_j \in A$ metric $d_{ij} \in D$.

The task is to optimize this function:

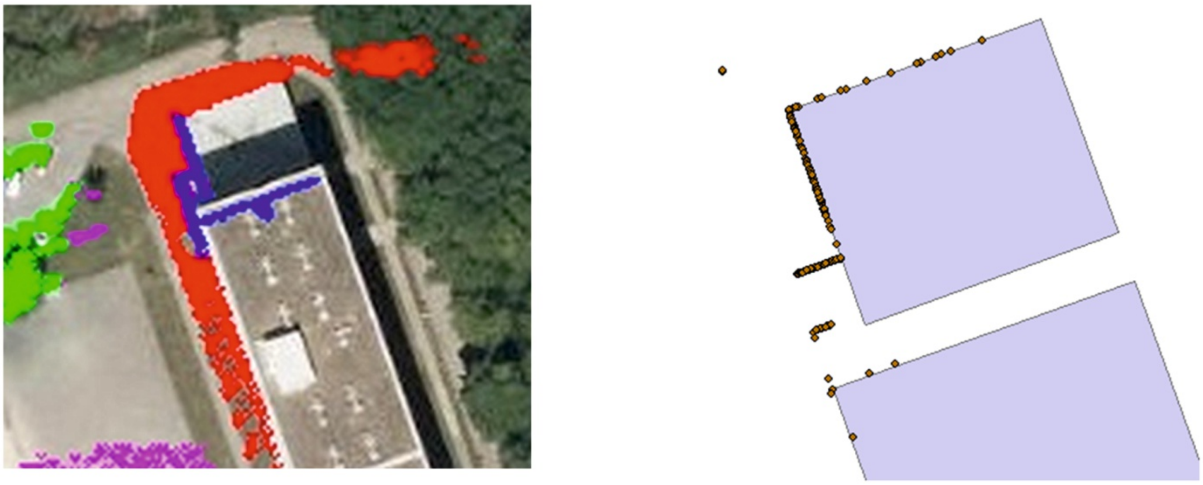
$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} a_i.t_j + \sum_{i=1}^{n}\sum_{j=1}^{m_i-1} a_i.s_j + \sum_{i,j=1,i\neq j}^{n-1} d_{ij} = \min \qquad (15)$$

for $\forall a \in A$,

where $a_i.t_j$ is the duration of the j. event on the i. element (j denotes number of repetitions of events) d_{ij} is the distance between objects a_i and a_j in time units, n is number of objects, m_i is number of events, that will run at the object a_i – see Figure 6.

3.3.4 Usage in practice

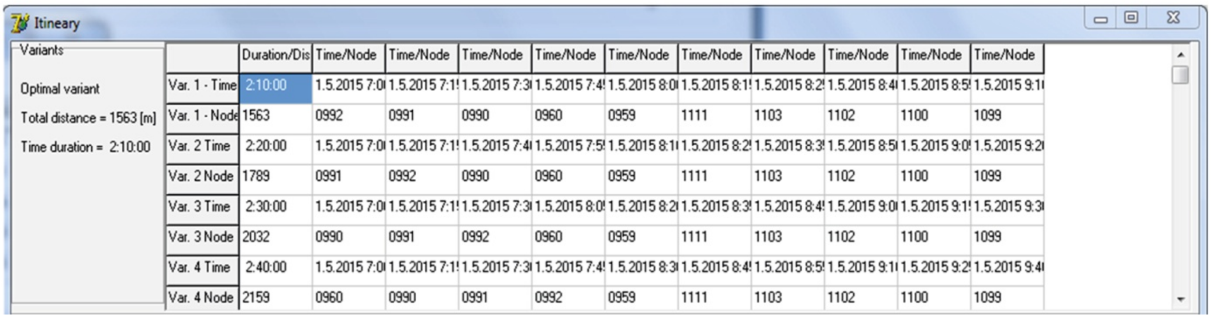
The proposed GNSS RTK measurement method was applied on several railway lines in the Czech Republic (Beroun to Rakovník, Opava to Krnov, the border to Valašské Meziříčí). The examples in the following text are from the section of the railway



a) building with points cloud
a) oblak točaka oko zgrade

b) cloud after filtration with cadastral map
b) oblak točaka nakon filtriranja s pomoću katastarske karte

Slika 4. Primjer otkrivanja razlika na katastarskoj karti.
Fig. 4 Example of detecting changes in the cadastral map.



Itineary

Variants

Optimal variant
Total distance = 1563 [m]
Time duration = 2:10:00

	Duration/Dis	Time/Node	Time/Node	Time/Node	Time/Node	Time/Node	Time/Node	Time/Node	Time/Node	Time/Node	Time/Node
Var. 1 - Time	2:10:00	1.5.2015 7:0	1.5.2015 7:1	1.5.2015 7:3	1.5.2015 7:4	1.5.2015 8:0	1.5.2015 8:1	1.5.2015 8:2	1.5.2015 8:4	1.5.2015 8:5	1.5.2015 9:1
Var. 1 - Node	1563	0992	0991	0990	0960	0959	1111	1103	1102	1100	1099
Var. 2 Time	2:20:00	1.5.2015 7:0	1.5.2015 7:1	1.5.2015 7:4	1.5.2015 7:5	1.5.2015 8:1	1.5.2015 8:2	1.5.2015 8:3	1.5.2015 8:5	1.5.2015 9:0	1.5.2015 9:2
Var. 2 Node	1789	0991	0992	0990	0960	0959	1111	1103	1102	1100	1099
Var. 3 Time	2:30:00	1.5.2015 7:0	1.5.2015 7:1	1.5.2015 7:3	1.5.2015 8:0	1.5.2015 8:2	1.5.2015 8:3	1.5.2015 8:4	1.5.2015 9:0	1.5.2015 9:1	1.5.2015 9:3
Var. 3 Node	2032	0990	0991	0992	0960	0959	1111	1103	1102	1100	1099
Var. 4 Time	2:40:00	1.5.2015 7:0	1.5.2015 7:1	1.5.2015 7:3	1.5.2015 7:4	1.5.2015 8:3	1.5.2015 8:4	1.5.2015 8:5	1.5.2015 9:1	1.5.2015 9:2	1.5.2015 9:4
Var. 4 Node	2159	0960	0990	0991	0992	0959	1111	1103	1102	1100	1099

Slika 5. Varijante rasporeda u programu.
Fig. 5 Itinerary variants in the program.

Na slici 3 f_b je prikazan blokom "spatial GIS analysis", a funkcija f_r prikazana je blokom "Delphi preprocessing". Dok je funkcija f_b dobro poznata iz prostorne analize, dobivanje odgovarajuće funkcije f_r zahtijevalo je više eksperimenata. Otkrivanje rubova na temelju centralne diskretne konvolucije pokazalo se neprikladnim zbog vremenske zahtjevnosti. Upotreba skupljanja (*aggregation*) točaka oblaka prema položaju u ravnini (x, y) s atributom učestalosti točaka na osi z (visina objekta) i s filtriranjem pokazalo se povoljnijim:

Select x, y, count(z) from list group by x, y (14)
having count(z) $> p$,

gdje je *list* popis koordinata točaka oblaka dobivenih laserskim skeniranjem, a p je parametar koji određuje visinu objekta (prema razlučivosti oblaka) u koji stavljamo zamišljenu ravninu presjeka koja dijeli zgrade ili strukture od drugih objekata na terenu. U praksi biramo $p \geq 2$.

U tu je svrhu u Delphiju osmišljen poseban program koji priprema ulazne podatke (vidi sliku 3). Taj program ima sljedeće funkcije:

1. dijeli oblak ulaznih točaka u dijelove manjeg obujma podataka kako bi se podatci mogli obraditi s pomoću uobičajenih tehničkih i stolnih računala
2. filtriranje točaka oblaka prema (14) u svakom segmentu
3. transformiranje datoteke izlaznih podataka u odgovarajući format za daljnju obradu u GIS-u.

3.2.2. Eksperimentalni rezultati

Metoda predložena u prethodnom poglavlju provjerena je na odabranom mjestu u Brnu (Češka). Slika 4a

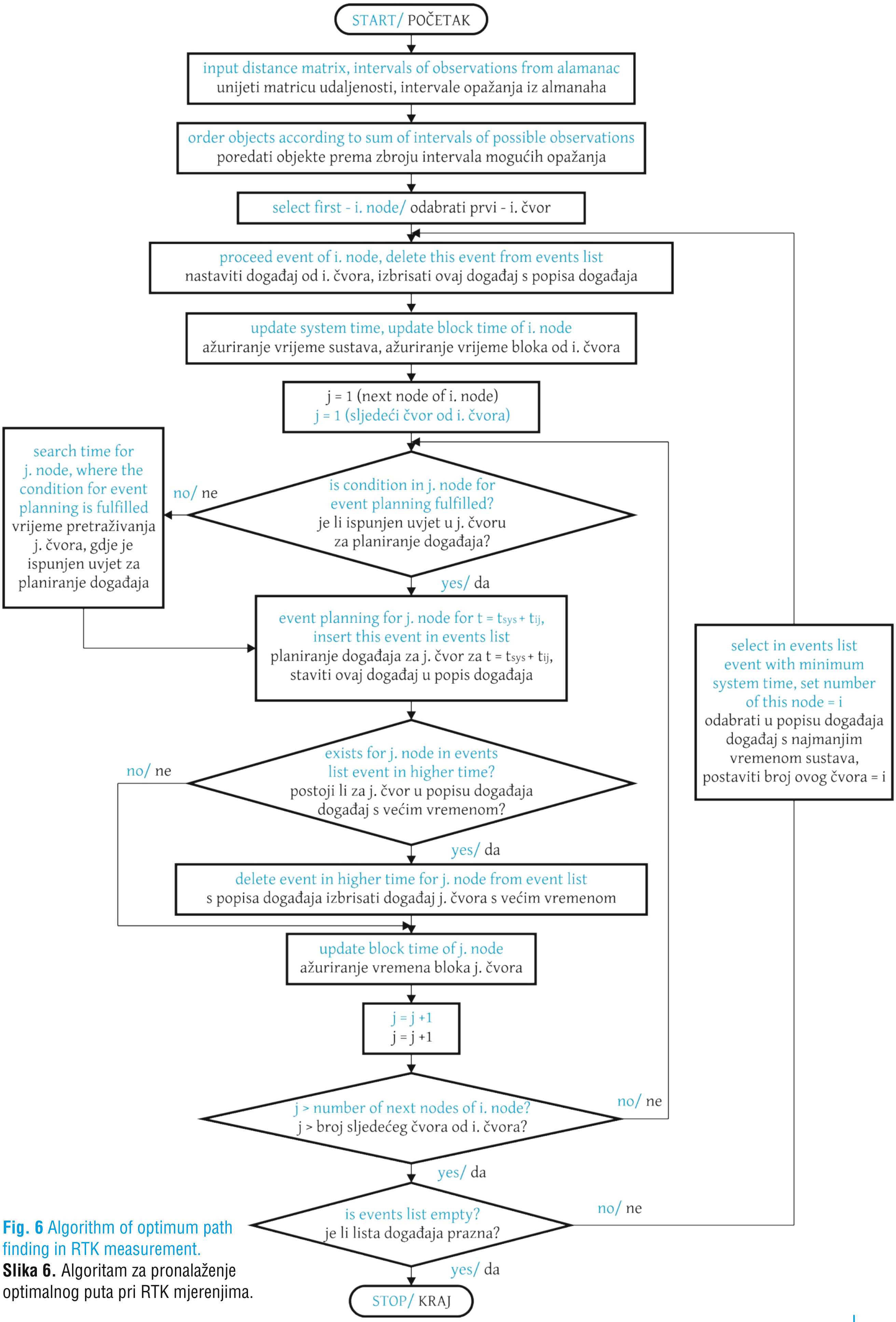

START/ POČETAK
input distance matrix, intervals of observations from alamanac
unijeti matricu udaljenosti, intervale opažanja iz almanaha
order objects according to sum of intervals of possible observations
poredati objekte prema zbroju intervala mogućih opažanja
select first - i. node/ odabrati prvi - i. čvor
proceed event of i. node, delete this event from events list
nastaviti događaj od i. čvora, izbrisati ovaj događaj s popisa događaja
update system time, update block time of i. node
ažuriranje vrijeme sustava, ažuriranje vrijeme bloka od i. čvora
j = 1 (next node of i. node)
j = 1 (sljedeći čvor od i. čvora)
is condition in j. node for event planning fulfilled?
je li ispunjen uvjet u j. čvoru za planiranje događaja?
no/ ne
search time for j. node, where the condition for event planning is fulfilled
vrijeme pretraživanja j. čvora, gdje je ispunjen uvjet za planiranje događaja
yes/ da
event planning for j. node for $t = t_{sys} + t_{ij}$, insert this event in events list
planiranje događaja za j. čvor za $t = t_{sys} + t_{ij}$, staviti ovaj događaj u popis događaja
exists for j. node in events list event in higher time?
postoji li za j. čvor u popisu događaja događaj s većim vremenom?
no/ ne
yes/ da
delete event in higher time for j. node from event list
s popisa događaja izbrisati događaj j. čvora s većim vremenom
update block time of j. node
ažuriranje vremena bloka j. čvora
j = j +1
j = j +1
j > number of next nodes of i. node?
j > broj sljedećeg čvora od i. čvora?
no/ ne
yes/ da
is events list empty?
je li lista događaja prazna?
no/ ne
yes/ da
STOP/ KRAJ
select in events list event with minimum system time, set number of this node = i
odabrati u popisu događaja događaj s najmanjim vremenom sustava, postaviti broj ovog čvora = i

Fig. 6 Algorithm of optimum path finding in RTK measurement.

Slika 6. Algoritam za pronalaženje optimalnog puta pri RTK mjerenjima.

pokazuje oblak točaka oko zgrade, slika 4b pokazuje oblak točaka nakon filtriranja s pomoću katastarske karte.

Proces je u najvećoj mjeri automatiziran s pomoću dviju aplikacija:

1. posebnog programa u okviru jezika Borland Delphi i
2. bilo kojeg programa za GIS (Geomedia, ArcGIS, QGIS).

Predložena je metoda relativno brza, nema velike hardverske i softverske zahtjeve i njome je moguće otkriti planimetrijske promjene na bilo kojoj digitalnoj karti u kombinaciji s laserskim skeniranjem. Rezultate je moguće primijeniti za ažuriranje odgovarajućih digitalnih karata. Projekt je objavljen u članku Bureša i dr. (2018).

3.3. Traženje optimalnog načina za mjerenja GNSS RTK

3.3.1. Opis problema

Ovaj se projekt ne odnosi prvenstveno na obradu velikih podataka, već na rješavanja zadataka s velikim prostorom stanja (*large state space*). To znači da je za izlaznu razlučivost potrebno pretražiti opsežan skup mogućih kombinacija ulaznih podataka. Vrijeme rješavanja prema tzv. robusnom algoritmu zahtijevalo bi nesrazmjernu količinu vremena na posebnim superračunalima.

U tom slučaju nije moguće primijeniti metode podjele ulaznih podataka prema teritorijalnom principu ili primijeniti filtriranje podataka kako bi se smanjio njihov obujam. Kako bismo dobili ispravno rješenje, moramo uzeti u obzir sve ulazne podatke. Primjer takvoga problema je pronaći najkraći put između svih točaka na danom mjestu koje je potrebno izmjeriti metodom GNSS – RTK.

3.3.2. Formulacija zadatka

Postavljanje zadatka:

Postoji skup objekata raspoređenih u prostoru na kojima treba mjeriti. Dostupna su sljedeća svojstva objekata: ID, položaj, vrijeme opažanja uključujući toleranciju, skup intervala (datum, vrijeme) u kojima je moguće mjeriti. Položaj je određen zajedničkom udaljenosti između pojedinih objekata u jedinicama duljine i vremena. U svakoj se točki unosi potrebno vrijeme do sljedećeg mjerenja. Mjerenje je moguće ponoviti ili je dostupno nekoliko skupova instrumenata za mjerenje.

Zadatak:

Stvoriti raspored mjerenja na svim objektima u smislu:

1. predmeta (proces mjerenja)
2. ekonomičnosti (učinkovitost procesa).

1. S obzirom na predmet potrebno je stvoriti optimalan raspored mjerenja na svim točkama tako da se na najmanju moguću mjeru svede odgoda zbog premještanja s jedne točke na drugu ili čekanje za odgovarajući interval mjerenja na danim točkama.

2. U smislu ekonomičnosti potrebno je odrediti ukupnu cijenu mjerenja, tj. računanje putovanja vozilima i pješačenje, cijenu mjerenja po satu, najam instrumenata. Iz cijene mjerenja potrebno je dobiti učinkovitost procesa, tj. optimalni broj instrumenata, broj mjerenja i odgovarajuće vrijeme za mjerenje.

3.3.3. Teorijski model i metoda rješavanja

Neka je A skup objekata. Svaki objekt $a \in A$ ima sljedeća svojstva:

- $a.p$ je položaj objekta u danom koordinatnom sustavu
- $a.t$ je vremenski interval tijekom kojega se događaj na danom objektu može dogoditi (i s ponavljanjem)
- $a.s$ je vremenska razlika od posljednjeg događaja. Tijekom tog vremena na objektu se ne može dogoditi nijedan drugi događaj.

Neka postoji skup vremenskih intervala B za vrijeme kojih se događaji mogu pokrenuti na objektima $a \in A$ te skup metrijskih vrijednosti D čiji elementi izražavaju udaljenost između pojedinih objekata $a \in A$ u jedinicama duljine ili vremena.

Neka su dane funkcije f: $A \to C$, $C \subseteq B$ i g: $A \times A \to D$. Funkcija f pridružuje objektima $a \in A$ skup odgovarajućih intervala $a.t \in C$ za vrijeme kojih se na objektima mogu evidentirati događaji, funkcija g pridružuje svakom paru objekata $a_i, a_j \in A$ metriku $d_{ij} \in D$.

Zadatak je optimirati sljedeću funkciju:

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} a_i.t_j + \sum_{i=1}^{n}\sum_{j=1}^{m_i-1} a_i.s_j + \sum_{i,j=1,i\neq j}^{n-1} d_{ij} = \min \qquad (15)$$

za $\forall a \in A$,

gdje je $a_i.t_j$ trajanje događaja j na elementu i (j označava broj ponavljanja događaja), d_{ij} je udaljenost između objekata a_i i a_j u jedinicama vremena, n je broj objekata, m_i je broj događaja koji će se pokrenuti na objektu a_i (vidi sliku 6).

3.3.4. Upotreba u praksi

Predložena metoda mjerenja GNSS RTK primijenjena je na nekoliko željezničkih linija u line between Hranice and Valašské Meziříčí. Figure 5 shows the output of an application created in Borland Delphi. It contains variants of the itinerary with the order of the points at which to measure. The optimum variant is on the first line. The total length of the path in this variant is 1,563 m and the total measuring time is 2 h 10 min.

The event-based method virtually represents a transformation from two domains (space and time) into one domain – time. This procedure is suitable for the approximate linear distribution of the points to be measured. There are 2 more algorithms for finding the Hamiltonian path on the graph. One of them is robust (Bartoněk 2015), which assumes a low degree of connection of nodes by edges, while the other is heuristic (Bartoněk et al. 2017), which gives fast (but only approximate) results even when configuring a complete graph.

4 Discussion and Conclusion

The projects presented covered 3 basic types of tasks (see section 2):

1. Processing analyses of a large area (e.g. the entire Czech Republic). In this case, it is recommended to divide the area into sub-areas and process them in parallel. The following conditions must be met:
    - The results of the component parts must be relatively independent of each other
    - The analysis results can be easily assembled into one common result in the form of a suitable data set
    - In order to establish basic decisions, a pilot project in the same or similar territory with the same task and results as the main project should be performed
    - The number of sub-segments is determined according to formulas (3) and (4), so that the processing is effective (ratio of processing time and overhead time)
    - Relation (7) can be used to estimate the time required for the overall project
    - Territorial details can be used to select the appropriate hardware and software - see relation (5)
    - It is recommended to use a maximum of 1 reference dataset to refine the results of spatial analyses - see relation (6). According to the author's experience, the use of more datasets puts a disproportionate burden on resources, extends the calculation and does not have a major influence on the quality of results
    - After completion of the solution in each sub-segment, it is recommended to check the results and, at the end, to assemble all the results into a whole
    - Good organization of the work by all members of the research team is the prerequisite for a successful solution.
2. Tasks with primarily large volumes of input data but with high redundancy or with data that are relatively interdependent. This section mainly concerns data acquired by ground or aerial laser scanning. The input data contain a high degree of redundancy, so the basic method to process them using conventional devices is by purpose filtering. The aggregation function was used in the project in section 3.2, reducing the data volume about 100 times. Furthermore, the application of automated or semi-automated vectorization is contemplated. This problem has not yet been satisfactorily solved.
3. Projects with a relatively small amount of primary data but a complex algorithm based on a large state space search, which is a combination of input data. Due to its primary complexity, this task class is solved by optimization or heuristic methods. In many cases, it means practically a transformation from a higher to a lower dimension to merge dimensions. In section 3.3 it meant a transformation from a spatial and temporal domain to a temporal domain, i.e. a merge of domains. The procedure is suitable for approximate linear arrangement of elements (geographical objects) in the terrain. For the general configuration of the elements, there are two more algorithms: a robust one giving accurate results, and a heuristic one, which provides only an approximate solution, but is very fast. However, tasks with complex state space are very rare in the real world because many combinations in a given task are either irrelevant or do not occur at all. Therefore, the practical complexity of combinatorial problems is always lower than the theoretical complexity and therefore each combinatorial problem can be solved automatically (on a computer) by applying a suitable method.

Češkoj (Beroun–Rakovník, Opava–Krnov, granica–Valašské Meziříčí). Primjeri u daljnjem tekstu uzeti su iz dijela željezničke linije Hranice–Valašské Meziříčí. Slika 5 pokazuje izlaz aplikacije osmišljene u jeziku Borland Delphi. Ona sadrži varijante rasporeda s redoslijedom točaka koje je potrebno izmjeriti. Optimalna varijanta nalazi se u prvom redu. Ukupna duljina puta u toj varijanti iznosi 1563 m, a ukupno vrijeme mjerenja jest 2 h i 10 min.

Ta metoda, koja se temelji na događajima, zapravo je transformacija iz dviju domena (prostor i vrijeme) u jednu – vrijeme. Taj je postupak prikladan za približno linearnu raspodjelu točaka koje treba izmjeriti. Postoje još dva algoritma za pronalaženje hamiltonskog puta na grafu. Jedan je od njih robustan (Bartoněk 2015), što pretpostavlja nizak stupanj povezanosti čvorova rubovima, dok je drugi heurističan (Bartoněk i dr. 2017) i daje brze (ali samo približne) rezultate, čak i u slučaju potpunog grafa.

4. Rasprava i zaključak

Prikazani projekti pokrivaju tri osnovna tipa zadataka (vidi drugo poglavlje):

1. Obrada analiza na velikom području (npr. cijela Češka). U tom se slučaju preporučuje podjela na podpodručja i njihova paralelna obrada. Moraju biti ispunjeni sljedeći uvjeti:
   - rezultati pojedinih dijelova moraju biti relativno nezavisni jedni o drugima
   - rezultat analize može se jednostavno sastaviti u zajednički rezultat u obliku prikladnog skupa podataka
   - kako bi se donijele osnovne odluke, potrebno je provesti pilot projekt na istom ili sličnom teritoriju
   - broj poddijelova određuje se na temelju formula (3) i (4) tako da je obrada učinkovita (omjer vremena obrade i ukupnog vremena)
   - pomoću odnosa (7) može se procijeniti potrebno vrijeme za cijeli projekt
   - detalji teritorija mogu se primijeniti za odabir odgovarajućeg hardvera i softvera – vidi jednadžbu (5)
   - predlaže se upotreba ne više od jednog referentnog skupa podataka kako bi se pročistilo rezultate prostornih analiza – vidi jednadžbu (6); prema autorovom iskustvu, upotreba drugih skupova podataka stavlja nesrazmjeran teret na resurse, proširuje računanje, a nema velik utjecaj na kvalitetu podataka
   - nakon dobivanja rješenja u svakom poddijelu, predlaže se provjeriti rezultate i naposljetku ih sastaviti u zajedničku cjelinu
   - uspješno rješenje uvjetovano je dobrom organizacijom rada svih članova istraživačkog tima.
2. Zadatci s prvenstveno velikim obujmom ulaznih podataka visoke redundantnosti ili s podatcima koji su relativno međuovisni. To se uglavnom odnosi na podatke dobivene skeniranjem na tlu ili uz pomoć zračnog lasera. Ulazni podatci sadrže visok stupanj redundantnosti pa je filtriranje osnovna metoda njihove obrade uz pomoć konvencionalnih uređaja. Funkcija sakupljanja (*aggregation*) primijenjena je u projektu opisanom u poglavlju 3.2, čime je obujam podataka smanjen otprilike 100 puta. Nadalje, razmotrena je primjena automatizirane ili poluautomatizirane vektorizacije. Taj problem zasad nije riješen na zadovoljavajući način.
3. Projekti s relativno malom količinom primarnih podataka, ali sa složenim algoritmom koji se temelji na pretraživanju velikih podataka (*large state space search*), što je kombinacija ulaznih podataka. Zbog svoje primarne složenosti taj tip zadataka rješava se metodama optimiranja ili heurističkim metodama. U mnogim je slučajevima to praktično transformacija iz više u nižu dimenziju. U poglavlju 3.3. opisana je transformacija iz vremenske i prostorne domene u vremensku domenu, tj. spajanje domena. Postupak je prikladan za približno linearni uređaj elemenata (geografskih objekata) na terenu. Postoje još dva algoritma za opću konfiguraciju elemenata: jedan robustan, koji daje točne rezultate, a drugi heurističan, koji daje samo približno rješenje, ali je vrlo brz. Međutim, zadatci s prostorom složenog stanja (*complex state space*) vrlo su rijetki u stvarnom svijetu zato što su mnoge kombinacije u danom zadatku ili bitne ili se uopće ne pojavljuju. Dakle, praktična je složenost kombinatoričkih problema uvijek niža od teorijske složenosti i stoga se svaki kombinatorički problem može riješiti automatski (na računalu) primjenom odgovarajuće metode.

References / Literatura

Allombert V, Michea D, Dupros F, Bellier Ch, Bourgine B. Aochi H, Jubertie S. (2014) An out-of-core GPU approach for accelerating geostatistical interpolation. Edited by Abramson D, Lees M, Krzhizhanovskaya V et al. International Conference on Computational Science, Book Series: Procedia Computer Science, vol. 29, 888–896

Arra T (2003) Knowledge, process and product information management to improve decision-making support at your pulp and paper mill. Appita Journal, vol. 56, issue 4, 259–261

Bajcsy P, Vandecreme A, Amelot J, Nguyen P, Chalfoun J, Brady M (2013) Terabyte-sized Image Computations on Hadoop Cluster Platforms. Edited by Hu X, Lin TY, Raghavan V et al. IEEE International Conference on Big Data, 125–131

Bartoněk D (2015) Algorithm for Travelling Salesman Problem. In: The Role of Service in the Tourism & Hospitality Industry. Edited by Ford Lumban Gaol and Fonny Hutagalung. CRC Press, Chap. 18, pp. 107–111. Print ISBN: 978-1-138-02736-7, eBook ISBN: 978-1-315-68852-7, 2015.

Bartoněk D, Bureš J, Opatřilová I (2014) Optimization of Pre-Processing of Extensive Projects in Geographic Information Systems. Advanced Science Letters, vol. 20, no. 10-11, 2026–2029

Bartoněk D, Bureš J, Opatřilová I (2015) Enhancement of Image Classification via GIS. Advanced Science Letters vol. 21, no. 12, ISSN 1936-6612, 3737–3740

Bartoněk D, Bureš J, Švábenský O (2017) Optimized GNSS RTK measurement planning for effective point occupation via heuristic analysis. Engineering Computations, vol. 34, issue 1, 90–104

Bureš J, Bartoněk D, Bárta L, Švábenský O (2018) Automated updating of selected technical and cadastral map components. Proceedings, 7th International Conference on Cartography and GIS, Sozopol, Bulgaria, ISSN: 1314-0604, Edited by Bandrova T, Konečný M, 888–895

Byun H R, Park J H, Jeong Y-S (2016) Optional Frame Selection Algorithm for Adaptive Symmetric Service of Augmented Reality Big Data on Smart Devices. Symmetry-Basel, vol. 8, issue 5, article 37

Čerba O (2017) Identické vazby propojených prostorových dat (Identical realtions of connected spatial data). Post-doctoral work. Brno University of Technology, Faculty of Civil Engineering, 211 pp. (in Czech)

Dehsangi M, Asyabi E, Sharifi M, Azhari S V (2015) cCluster: A Core Clustering Mechanism for Workload-Aware Virtual Machine Scheduling. Book authors: Awan I, Younas M, Mecella M. 3rd International Conference on Future Internet of Things and Cloud (FICLOUD) and International Conference on Open and Big Data (OBD), 248–255

Dengel A (2009) The semantic desktop as a means for personal information management. Edited by Fred A, KDIR Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, 125–135

Dixon M (2015) A Pattern Oriented Approach for Designing Scalable Analytics Applications. Edited by Jann J, Moreira J, Kumar M. 2nd Workshop on Parallel Programming for Analytics Applications (PPAA 2015), 4–8

Elliott J, de Souza R S, Krone-Martins A, Cameron E, Ishida EOO, Hilbe J for the OCIN colaboration (2015) The overlooked potential of Generalized Linear Models in astronomy-II: Gamma regression and photometric redshifts. Astronomy and Computing, vol. 10, 61–72

Elster A C (2002) High-performance computing: Past, present, and future. Edited by Fagerholm J, Haataja J, Jarvinen J et al. Applied Parallel Computing: Advanced Scientific Computing. Book Series: Lecture Notes in Computer Science, vol. 2367, 433–444

Gopu A, Hayashi S, Young M D, Kotulla R, Henschel R, Harbeck D (2016) Trident: Scalable Compute Archives Workflows, Visualization, and Analysis. Edited by Chiozzi G, Guzman JC. Software and Cyberinfrastructure for Astronomy IV, Book Series: Proceedings of SPIE, vol. 9913

Kim H-W, Park J H, Jeong Y-S (2016) Human-centric storage resource mechanism for big data on cloud service architecture. Journal of Supercomputing, vol. 72, issue 7, 2437–2452

Kim H-W, Park J H, Majigsuren D, Jeong Y-S (2015) Efficient Sustainable Operation Mechanism of Distributed Desktop Integration Storage Based on Virtualization with Ubiquitous Computing. Sustainability, vol. 7, issue 6, 7568–7580

She B, Boulanger P, Noga M (2011) Real-Time Rendering of Temporal Volumetric Data on a GPU. Edited by Banissi E, Bertschi S, Burkhard R et al. 15th International Conference on Information Visualisation, Book Series: IEEE International Conference on Information Visualization, 622–631