VÝUKA GEOINFORMATICKÝCH PŘEDMĚTŮ NA PŘÍKLADECH DAT EVROPSKÉ UNIE

Zdena DOBEŠOVÁ[1], Karel MACKŮ[2], Michal KUČERA[3]

[1,2,3] Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého, 17. listopadu 50, 771 46 Olomouc, Česká republika
zdena.dobesova@upol.cz, *karel.macku@upol.cz*

[3] Správa železnic, Jeremenkova 103/23, 779 00 Olomouc, Česká republika
KuceraMic@spravazeleznic.cz

DOI: https://doi.org/10.31490/9788024846071-153

[1] ORCID: 0000-0002-3989-5951

[2] ORCID: 0000-0002-5899-2626

[3] ORCID: 0000-0002-8633-4380

Příspěvek prošel odbornou recenzí

Abstrakt

Předměty Data Mining a Pokročilé zpracování geodat jsou povinnými předměty pro magisterské studium programu Geoinformatika a kartografie na Univerzitě Palackého. V praktických cvičeních se využívají data poskytovaná evropskými úřady. V obou předmětech jsou to data poskytovaná statistickým úřadem Evropské unie Eurostat a dále data poskytovaná Evropskou agenturou životního prostředí v programu Copernicus Land Monitoring Servise – Urban Atlas. Přínosem praktických cvičení je nejen procvičení různých metod analýzy, ale i seznámení se zdroji dat Evropské unie. Praktické příklady tak zvyšují znalost o geografických evropských tématech a možnostech získávání dat z volně dostupných zdrojů.

V rámci předmětu Data Mining jsou procvičována témata korelace, analýza hlavních komponent, hierarchické a nehierarchické shlukování na datech zaměstnanosti podle číselníku ekonomické aktivity NACE Level1. Analýza časových řad je procvičována na datech železniční dopravy v zemích EU. Zajímavé je zjišťování trendu osobní dopravy v letech 2004 až 2021 v jednotlivých evropských zemích. Dále lze dobře zjistit kvartální změny z dat dopravy v důsledku pandemie COVID-19 v roce 2020 a 2021. Postupy hledání podobnosti jsou ukázány na datech Urban Atlasu. Navíc je procvičováno použití natrénovaných neuronových sítí pro zjištění podobnosti evropských měst podle landuse center měst. Závěrečný semestrální projekt je také postaven na evropských datech. Ve cvičení je používán data miningový software Orange, který vizuálním programovacím jazykem.

Předmět Pokročilé zpracování geodat se zaměřuje především na témata související s prostorovou statistikou – nejprve provádí studenty metodami pokročilé exploratorní analýzy, dále jsou představeny prostorově vážené metody, na které navazuje využití prostorových regresních modelů. Druhou část sylabu tvoří využití metod geocomputation, ve které se studenti seznámí s tématy fuzzy logiky, teorie informace a fraktální geometrie a jejich využití v prostoru. Pro cvičení jsou využívány regionální statistiky NUTS2 z databáze Eurostat a databáze OECD.

Výukové texty jako je učebnice software Orange (Dobešová, 2022) a příklady k samostatnému procvičení nad daty EU jsou volně dostupné včetně zdrojových dat a programových kódů na stránce projektu http://urbandm.upol.cz/ v sekci *Výukové materiály*.

Abstract

Data Mining and Advanced Geodata Processing are compulsory courses for the Master's degree in Geoinformatics and Cartography at Palacký University. The practical exercises use data provided by European authorities. In both courses, these are data provided by Eurostat, the statistical office of the European Union, and data provided by the European Environment Agency under Copernicus Land Monitoring Service - Urban Atlas. The benefit of the practical exercises is not only to practise different methods of analysis but also to become familiar with the sources of European Union data. Thus, the practical examples increase the

knowledge of European geographical topics and the possibilities of obtaining data from freely available sources.

In the Data Mining course, the topics of correlation, principal component analysis, hierarchical and non-hierarchical clustering are practiced on employment data according to the NACE Level1 economic activity code. Time series analysis is practiced on rail traffic data in EU countries. Of interest is the identification of passenger and freight traffic trends from 2005 to 2021 in each European country. Furthermore, the quarterly changes in traffic due to the covid-19 pandemic in 2020 and 2021 can be well identified from the data. Similarity search procedures are shown on Urban Atlas data. In addition, the use of trained neural networks is practiced to find the similarity of European cities according to land use centres of cities. The semester assignment is also based on European data. In the exercise, the data mining software Orange is used with a visual programming language.

The Advanced Geodata Processing course focuses primarily on topics related to spatial statistics - first, it guides students through advanced exploratory analysis methods, then spatially weighted methods are introduced, followed by the use of spatial regression models. The second part of the syllabus consists of the use of geocomputation methods, in which students are introduced to the topics of fuzzy logic, information theory and fractal geometry and their applications in space. Regional NUTS2 statistics from Eurostat and OECD database are used for the exercises.

Educational texts such as the Orange software textbook and exercise book for self-practice on EU data are freely available, including source data and program codes, on the project website http://urbandm.upol.cz/ in the Teaching Materials section.

Klíčová slova: analýza; Eurostat; OECD; prostorová data; regionální statistika; Urban Atlas; výuka

Keywords: analysis; Eurostat; OECD; spatial data; Urban Atlas; regional statistics; lecturing

1. ÚVOD

V zimním semestru magisterského studia Geoinformatika a kartografie je vyučován povinný předmět Data Mining. Dotace předmětu je 3 hodiny přednášek a 3 hodiny cvičení. V letním semestru je vyučován pro stejný studijní program povinný předmět Prostorové zpracování dat s dotací 2 hodiny přednášek a 2 hodiny cvičení. V evaluaci ze strany studentů se objevil mj. požadavek, aby praktická témata byla procvičována na geografických tématech místo učebnicových příkladů pro Data Mining jako je databáze kosatců Iris nebo příklady z bankovnictví, pojišťovnictví nebo lékařství (Berka, 2005) (Šarmanová, 2012). Garantka předmětu reagovala na tuto studentskou připomínku a rozhodla se obohatit výuku a připravit praktické příklady s použitím dat Evropské statistické databáze Eurostat (Eurostat, 2021a), dat projektu Copernicus Urban Atlas (Copernicus Programme, 2020) nebo dat OECD. Inovace předmětu Data Mining a zavedení nového předmětu Prostorové zpracování dat se zároveň podařilo podpořit projektem ERASMUS+ Jean Monnet Module s názvem UrbanDM. Příklady použití dat a zkušenosti v prvním roce nasazení do výuky přináší tento příspěvek.

2. VYUŽITÍ DAT DATABÁZE EUROSTAT

Data z evropské databáze Eurostat jsou použita v úvodních tématech předmětu Data Mining, mezi která patří preprocesing dat a základní exploratorní datová analýza. Dále jsou procvičována témata korelace, analýza hlavních komponent, hierarchické a nehierarchické shlukování, klasifikace, rozhodovací stromy.

Databáze Eurostat poskytuje data rozdělená podle témat (Eurostat, 2021b). Témata jsou dále dělena na kategorie a podkategorie, ve kterých jsou publikována data v různé administrativní podrobnosti. Interaktivní rozhraní umožňuje vybírat parametry v rozsahu výběru jednotlivých států (geopolitické entity), nebo souhrnných statistik za celou EU (27 resp. 28 států). Dále lze vybírat časové rozmezí nebo konkrétní rok. Na data lze nahlížet formou tabulek, spojnicových nebo sloupcových grafů nebo i map. Pro cvičení je tak na výběr nepřeberné množství zajímavých témat. Data lze stahovat v různých formátech, stažená data obsahují kromě vlastních dat i na samostatných listech *Summary* a *Structure* metadata, kde jsou údaje o zdroji dat, času

stažení dat, poslední aktualizaci dat, časovém rozsahu dat, o jednotkách dat apod. Seznámení se s formátem a obsahem metadat je pro studenty dobrým a inspirativním výukovým příkladem předávání metadat k samotným datům. Další informace o databázi Eurostat a dalších datových zdrojích užitečných z pohledu prostorových dat EU lze nalézt v kapitole knihy Spationomy (Pászto, Redecker, et al., 2020).

Praktická část cvičení předmětu Data Mining je vedena ve volně dostupném softwaru Orange (*Orange*, 2021). Výhodou tohoto softwaru je grafická forma programování tj. sestavování postupu zpracování. Postup se sestavuje jako grafické workflow z jednotlivých uzlů (widgetů), workflow lze díky tomu sestavit uživatelsky velmi jednoduše. Studenti se tak můžou soustředit na volbu parametrů jednotlivých metod, jejich pochopení a zejména na porozumění a interpretaci dat. Pro cvičení je dostupná jednak samotná originální dokumentace softwaru, dále pak řada YouTube návodů, poskytnutá vývojovým týmem softwaru Orange na Univerzity of Ljubljana ve Slovinsku (*Orange Visual Programming Documentation*, 2021).

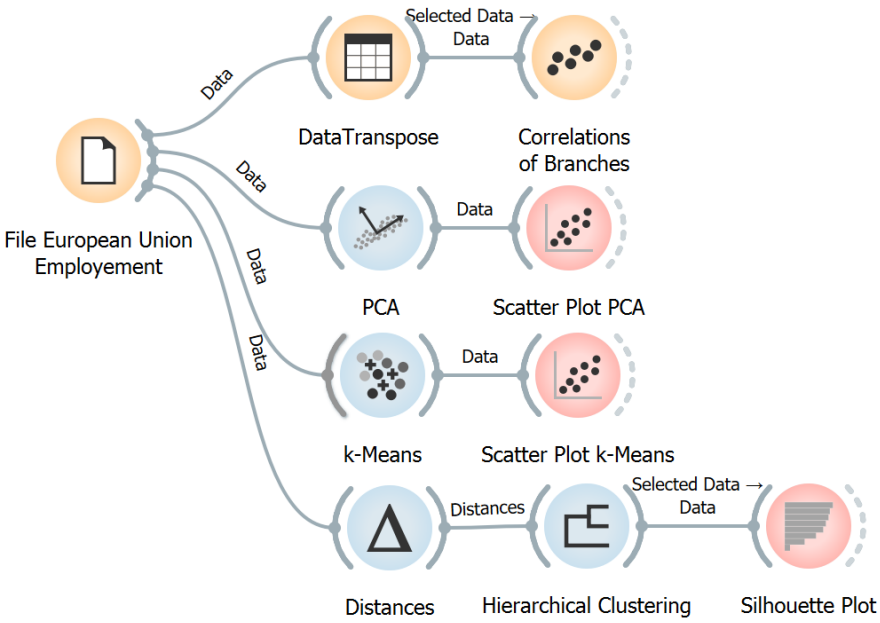
Pro studenty vznikla nově učebnice ORANGE Praktický návod do cvičení předmětu Data Mining (Dobešová, 2022), která částečně využívá evropská data. Dále vznikla Cvičebnice s praktickými příklady k samostatnému zpracování dat Eurostatu. Učebnice pro ORANGE je určena pro společnou práci studentů na cvičeních na počítačové učebně. Krok za krokem je vysvětlen a studenty vypracován postup řešení jednotlivých úloh, kdy jsou výsledky průběžně kontrolovány a zejména je vedena diskuze nad interpretací výsledků. V průběhu cvičení jsou procvičována i variantní řešení či různá nastavení parametrů metod. Cvičení lze zvládnout částečně i formou samostudia díky učebnici, což je využíváno nově v rámci kombinovaného studia. Pro zcela samostatnou práci je určena Cvičebnice. V průběhu semestru je zadán projekt, kdy každý student zpracovává jiná zdrojová data. Semestrální projekt je prezentován studenty ve formě závěrečné zprávy.

2.1 Analýza dat zaměstnanosti v EU

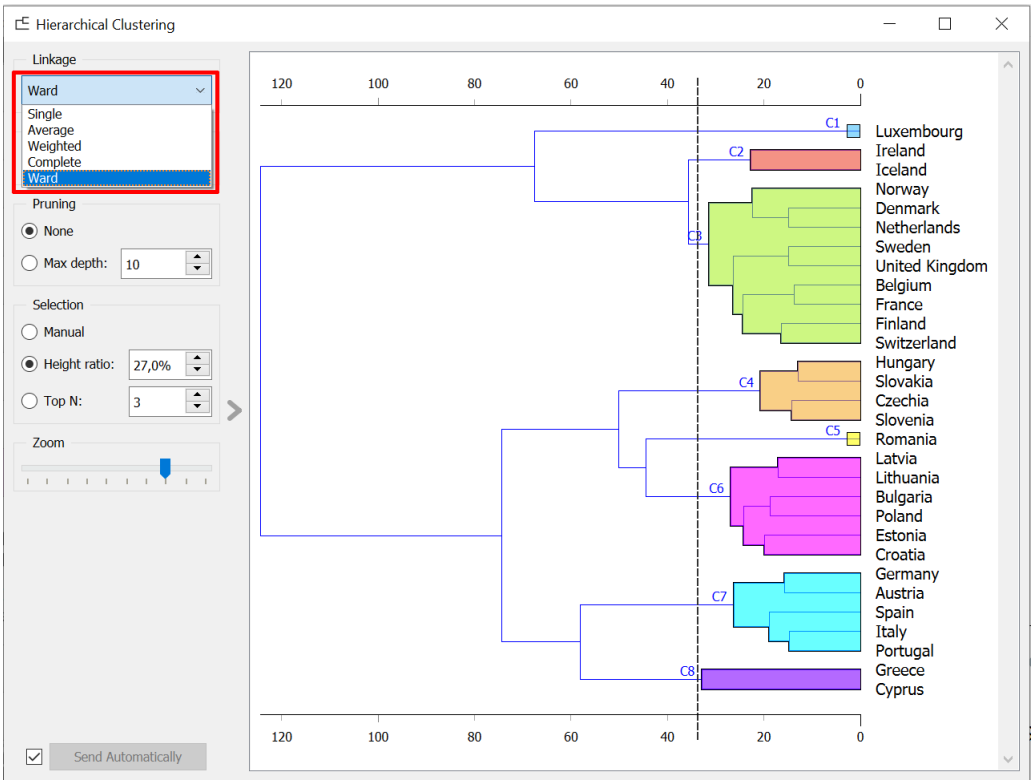
Jedním z komplexních a zajímavých témat je zaměstnanost pracujících podle druhu ekonomické aktivity v jednotlivých státech EU. Data databáze Eurostat poskytují počty zaměstnanců v jednotlivých odvětvích podle klasifikace číselníku NACE verze 2 (Eurostat, 1990). Stažená data jsou do cvičení zúžena na 21 kategorií číselníku (Level 1) a přepočítaná z absolutních na relativní hodnoty (Eurostat, 2022a).

Z hlediska probíraných témat preprocesingu, je zajímavé, že data nejsou ve všech rocích úplná a obsahují chybějící údaje nebo jsou označena jako *p* (provisional). Je možné data dopočítat (imputovat) některou z vyučovaných strategií: průměrem hodnot, nejfrekventovanější hodnotou, optimistickou, pesimistickou strategií nebo metodu 1-NN (1-nearest neighbor) či náhodnou veličinou (Šarmanová, 2012). Nakonec byla použita data z roku 2018, kdy některé státy byly z datasetu odstraněny z důvodu většího množství chybějících dat (Malta, Bosna a Hercegovina). Základním úkolem cvičení je zjistit pomocí analytických metod, jak se liší odvětvová zaměstnanost v jednotlivých zemích EU a zda je rozdíl mezi státy západní a střední/východní Evropy. Dále je úkolem zjistit, zda pokračuje orientace států střední/východní Evropy na primární a sekundární sektor, tzn. na zemědělství a průmyslový sektor. Studenti jsou tak vedeni k intepretaci zjištěných korelací, shluků a podobností.

Na Obr. 1 je ukázka workflow, kde je vypočítána korelace hodnot, analýza hlavních komponent (PCA), shlukování metodou k-Means, hierarchické shlukování a posouzení shluků pomocí grafu siluety. Jedním ze zajímavých zjištění je negativní korelace počtu zaměstnanců v *Profesionálních, vědeckých a technických aktivitách* oproti kategorii *Výroba* (manufacturing). Dalším zajímavým zjištěním je výrazná odlišnost Rumunska, které zaměstnává podstatně více lidí v zemědělství na rozdíl od všech ostatních státu Evropy. Francie, Velká Británie, Švédsko, Nizozemí, Belgie, Finsko, Švýcarsko a Dánsko tvoří shluk podobných států a jsou si tedy podobné strukturou zaměstnanců v odvětvích. V případě shlukové analýzy mohou studenti experimentovat se způsoby nastavení výpočtu vzdálenosti (euklidovská, cosinová, Manhattan) a nastavení mezishlukové vzdálenosti spojení shluků (linkage) v dendrogramu a následně určovat shluky (Obr. 2). Vzájemná podobnost Řecka a Kypru není překvapující, další shluk podobných post-komunistických zemí tvoří Slovensko, Slovinsko, Česko a Maďarsko. Postup zpracování, který se studenti učí na cvičení, byl již prezentován formou článku na on-line konferenci CSOC 2021 (Masopust et al., 2021).



Obr. 1 Workflow v programu Orange pro zpracování dat zaměstnanosti podle odvětví v EU



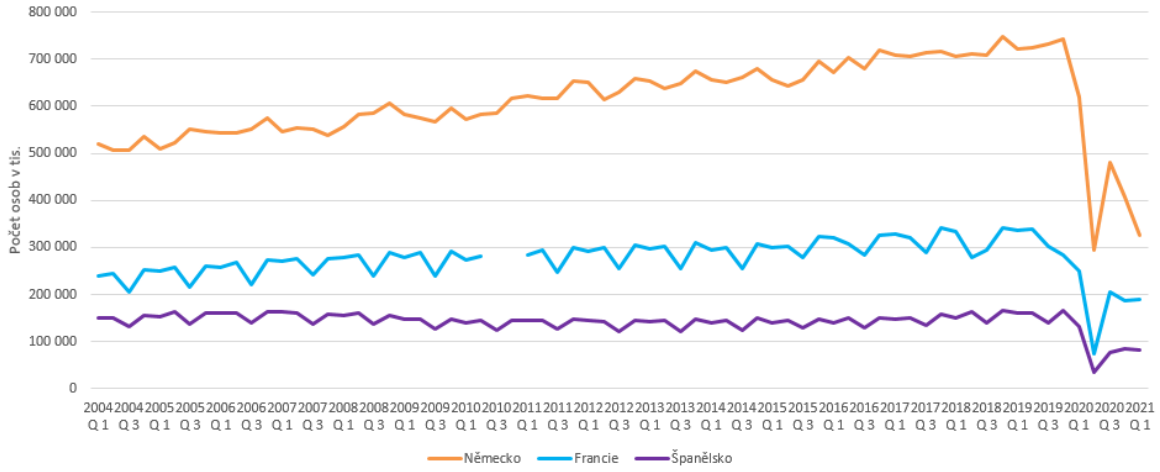
Obr. 2 Dendrogram s nastavením spojení shluků (linkage) a výběrem počtu shluků v programu Orange

2.2 Zpracování časových řad dopravy na železnici

Data databáze Eurostat lze využít i pro výuku tématu časových řad a jejich zpracování. Například data o přepravě osob na železnici jsou dostupná od roku 2004 až do Q2 roku 2021. Data jsou poskytována v čtvrtletních úhrnech (Q) pro osobní dopravu v jednotlivých státech EU (Eurostat, 2022b). Údaje za některá období a státy chybí; například Švýcarsko poskytuje data až od roku 2008, Severní Makedonie od roku 2009; chybí data pro Velkou Británii od čtvrtého čtvrtletí roku 2020, dále např. dva údaje z roku 2010 pro Francii (viz Obr. 3). Také publikování nejnovějších dat Eurostatu má určité zpoždění oproti zveřejňování dat v národních cenzech, které obsahují novější data, často však předběžného charakteru. V čase psaní článku zatím chybí data za Q3 a Q4 roku 2021.

Při analýze časových řad je zajímavé sledovat vývoj časové řady, její trend či sezónní a cyklickou složku. V případě přepravy osob na železnici jsou patrné poklesy počtu v roce 2020 a 2021 způsobené celosvětovou pandemií nemoci COVID-19. Mobilita osob výrazně poklesla vlivem různých omezení vyhlášenými jednotlivými státy v průběhu roku 2020 a 2021. Poklesla jak dojížďka za prací, tak také volnočasová mobilita (Pászto, Burian, et al., 2020). Je zajímavé, jak se pokles přepravených cestujících projevil v jednotlivých státech EU. Na Obr. 3 je průběh časových řad počtu přepravených osob v Německu, Francii a Španělsku. Ve všech třech státech měl počet přepravených osob od roku 2004 rostoucí trend. V roce 2020 nastal prudký pokles počtu přepravených osob a zejména v druhém čtvrtletí dosáhl absolutního minima. Je zajímavý vývoj v dalších čtvrtletích roku 2020, kdy v Německu po nárůstu v Q3 nastal opět pokles v Q4 a Q1 2021. U Francie a Španělska po nárůstu v Q3 2020 již nenastal takový pokles přepravy cestujících jako v Německu a stejně tak se oproti Německu liší i hodnoty v Q1 2021.

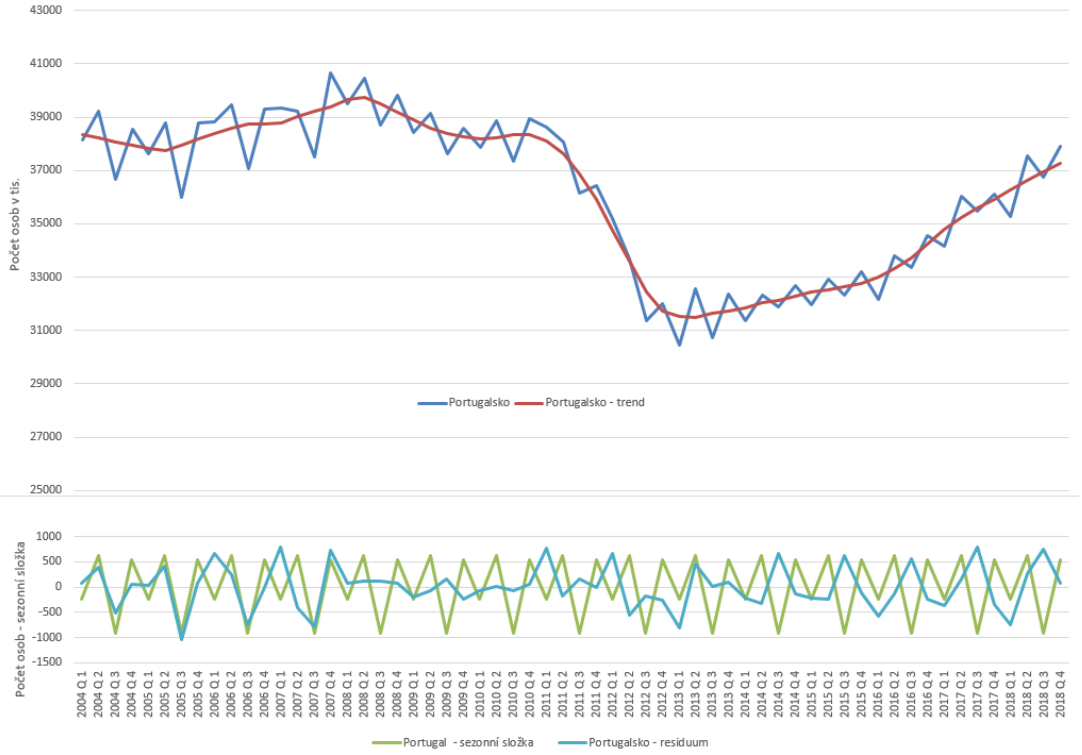
Při vykreslování časových řad je třeba dbát na rozsah svislé osy *y* pro různý rozsah zdrojových hodnot. Uvedené tři státy vykazují přepravu osob ve stovkách tisíc cestujících. Při porovnávání s menšími zeměmi EU jako je Slovensko, Dánsko, Maďarsko atd., kde se počty přepravených osob pohybují v desítkách tisíc osob, je lepší vykreslit data do samostatného grafu s odpovídajícím rozsahem hodnot na ose *y*. Studenty je třeba upozornit na fakt, že data se pohybují v různém rozsahu, seznámit je s tímto rozsahem a podle toho volit více samostatných grafů s různým rozsahem a měřítkem svislé osy. Lepší vyhodnocení poskytnou relativní hodnoty tempa růstu, kdy se počítá podíl hodnot jednotlivých čtvrtletí dvou po sobě následujících roků. Relativní hodnoty se potom pohybují kolem hodnoty 1 a poslouží k lepšímu srovnání čtvrtletí. Hodnota nad 1 znamená růst, hodnota pod 1 znamená pokles růstu. Zde se tedy studentům nabízí další možnosti praktického procvičování zpracování časových řad.



Obr. 3 Časová řada počtu přepravených osob na železnici v Německu, Francii a Španělsku (2004-2021)

Typickou úlohou časových řad je rozklad na složky časové řady. Na Obr. 4 je ukázán aditivní rozklad časové počtu cestujících řady v Portugalsku. Zpracování v softwaru Orange nabízí widget *Seasonal*, který automaticky provede rozklad vstupní časové řady. Nutností je zadat typ rozkladu: aditivní nebo multiplikativní a délku periody. V případě této čtvrtletní řady je hodnota periody 4. Při rozkladu časové řady je zjištěn trend časové řady, který je vykreslen v grafu nahoře společně s originální časovou řadou. Z původní časové řady z dat Eurostatu byla zpracována zkrácená řada, a to do roku 2018, protože zejména výkyvy z roku 2020 ovlivňují identifikaci sezónní složky. Sezonní a reziduální složka je vykreslena samostatně v dolním grafu Obr. **4** z důvodu odlišného rozsahu svislé osy. Vyšetření trendu lze také pomocí klouzavých průměrů s různě dlouhým časovým oknem. Také se nabízí výpočet centrovaného klouzavého průměru.

Poslední z typických úloh analýzy časových řad je predikce časové řady. U uvedených časových řad přepravy cestujících na železnici vlivem pandemických dat nedává smysl provádět predikci a bylo by lepší vybrat jiná vhodná data, třeba z oblasti demografie, která nejsou tak zatížena výkyvy vlivem pandemie.



Obr. 4 Aditivní rozklad časové řady počtu cestujících v Portugalsku od roku 2004 do roku 2018

3. VYUŽITÍ DAT COPERNICUS URBAN ATLAS PRO HLEDÁNÍ PODOBNOSTI EVROPSKÝCH MĚST

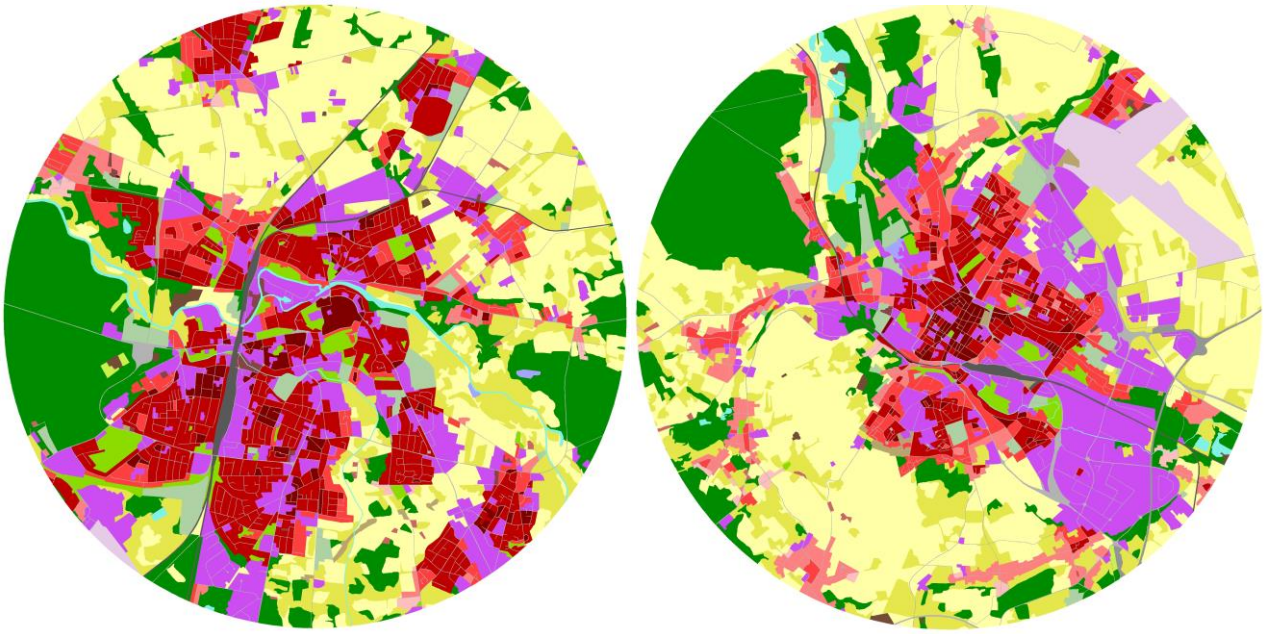
Evropská agentura životního prostředí zaštiťuje projekt Copernicus Land Monitoring Servise – Urban Atlas (UA). V rámci tohoto projektu jsou shromažďována a následně volně poskytována data o využití území pro velké územní zóny okolo velkých měst, označené jako Functional Urban Area (FUA) (Eurostat, 2020).

Jednou ze základních úloh Data Mining je hledání podobných objektů a vytváření shluků. V rámci výuky jsou studenti seznamování i s bio-inspirovanými algoritmy jako jsou neuronové sítě, genetické algoritmy či optimalizace pomocí mravenčích kolonií. Po seznámení s teorií a architekturou neuronové sítě je vhodné ukázat praktický příklad nasazení neuronové sítě. Software Orange disponuje doplňkem Image Analytics, který poskytuje embedery (připojení) k pěti různým natrénovaným neuronovým sítím (Biolab, 2019). Pro hledání podobnosti měst byla použita natrénovaná neuronová síť Painters, která je trénovaná na obrazech malířů.

Z databáze Urban Atlas bylo v rámci předchozího výzkumu nachystáno 787 map center evropských měst. Použita byla barevná legenda Urban Atlas a vyexportována sada obrázků. Podobné typy landuse mají blízké barevné vyjádření. Například různé hustoty zástavby jsou vyjádřeny různým tónem červené barvy. Zeleň a lesní porosty odstíny jsou vyjádřeny odstíny zelené barvy. Účelem bylo najít města s podobným vzorem landuse pomocí neuronové sítě. Hledání podobnosti evropských měst pomocí neuronové sítě bylo realizováno v rámci vědeckého výzkumu (Dobesova, 2020). Výsledky a praktický experiment byl následně zařazen i do výuky studentů. Studenti dostávají k dispozici vzorek obrázků 100 měst Urban Atlasu o velikosti mezi 50 až 100 tis. obyvatel (Dobesova, 2019). Následně použijí natrénovanou neuronovou síť, která vrací pro každý obrázek popisný vektor 2 048 číselných hodnot. Dále studenti mohou experimentovat s různými metodami shlukování a hledání podobných měst pomocí nejbližší vzdálenosti a dendrogramu. Nejzajímavější částí je intepretace nalezených podobností a prostorových vzorů landuse. Ukázka podobných měst je na Obr. 5, kde je město Celle v Německu poblíž Hannoveru a město Beauvais na severu Francie (75 km od Paříže). Města jsou si podobná poměrně malou plochou husté zástavby v centru (tmavě červená barva, hustota zástavby nad 80 %) a naopak převažující středně hustou zástavbou (světlá červená). Typické je pro obě města hodně roztříštěných průmyslových, komerčních, vojenských a soukromých ploch na území celého města (fialová barva). Obě města mají na okraji jedno rozsáhlé souvislé území lesního porostu. Tvarem jsou si podobné i

pravoúhlé plochy orné půdy a zeleně v okolí města. Na podobnost vzoru mají i vliv přímé silniční a železniční komunikace s viditelným železničním nádražím v obou městech (šedá barva).

Na datech Urban Atlasu je postavena i semestrální práce, kdy studenti hledají podobná měst a podobná FUA na základě popisných dat kategorií landuse, jako je počet jednotlivých druhů ploch. Zdroje získaných dílčích dat navzájem sdílí a potom hledají podobné dvojice, či trojice měst.



Obr. 5 Podobná dvojice měst Celle (vlevo) a Beauvais (vpravo) podle landuse zjištěná pomocí natrénované neuronové sítě Painters

4. VÝUKA PŘEDMĚTU POKROČILÉ ZPRACOVÁNÍ GEODAT

V rámci výuky předmětu *Pokročilé zpracování geodat* studenti prohlubují své analytické schopnosti zpracování prostorových dat, a to ve dvou hlavních směrech: první polovina semestru se zaměřuje především na aplikaci statistických metod na prostorová data. Již z dřívějších kurzů základního bakalářského studia by studenti měli být seznámeni se základními principy prostorové statistiky, a to především s metodami pro hodnocení prostorové autokorelace a geostatistickými metodami pro interpolování (kriging). Hlavním posláním úvodní části předmětu Pokročilé zpracování geodat je proto směřováno na rozšíření přehledu o těchto metodách a prohloubení znalostí prostorové statistiky, a to na úrovni pochopení významu této celé disciplíny a následné aplikace vhodných metod. Studenti se postupně poučí o pokročilých metodách exploratorní analýzy dat, jsou hlouběji popsány základní metody pro hodnocení prostorové autokorelace a prostorových vzorů, pro zkoumání prostorové nestacionarity jsou představeny základní prostorově vážené metody a závěrem je demonstrováno prostorové regresní modelování lineární a logistickou regresí.

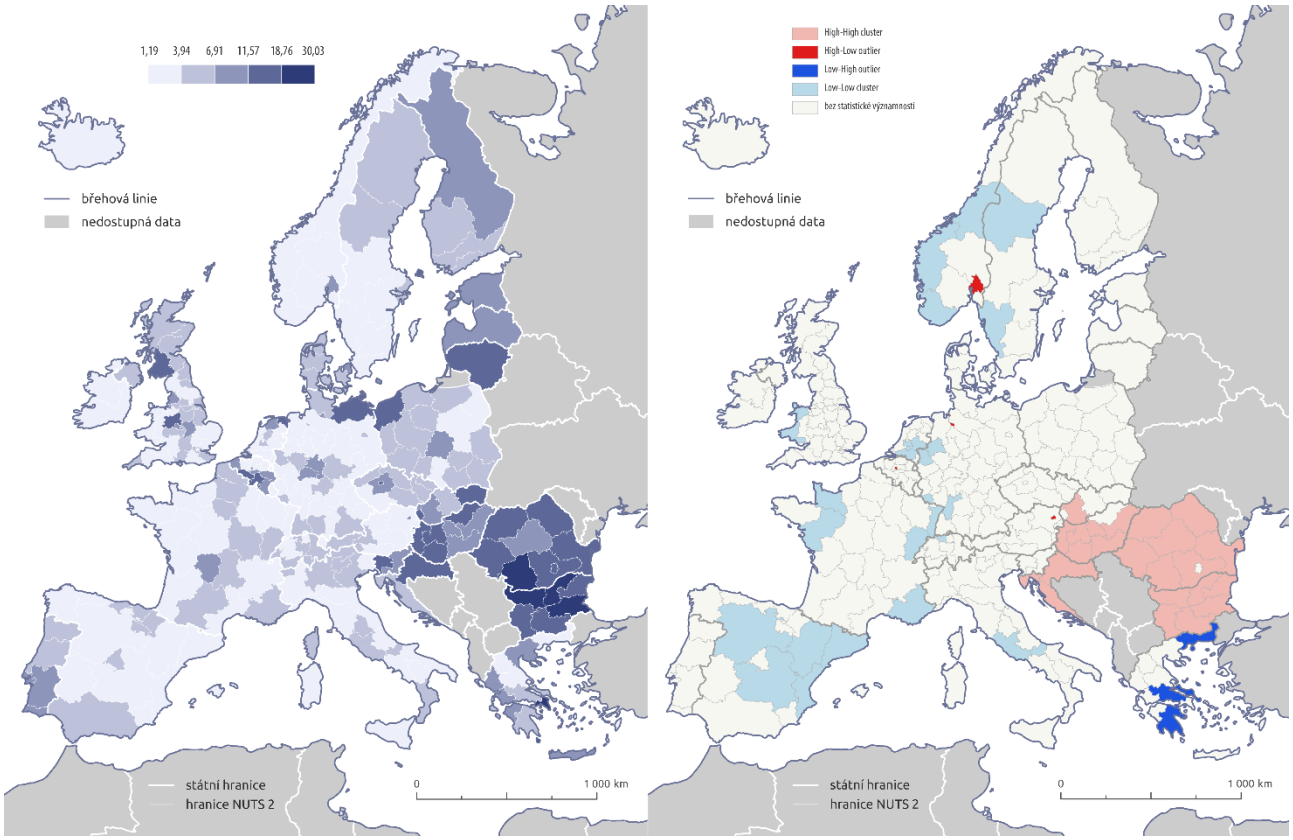
Druhá část semestrální výuky je zaměřená na představení vybraných metod zpracování prostorové informace výpočetními silami, někdy označované jako geocomputation (Openshaw & Abrahart, 1996). Z této skupiny metod jsou studentům představeny základní principy a poznatky z teorie informace, fraktální geometrie, teorie fuzzy množin a z výpočtů tvarových prostorových metrik. U každé z těchto disciplín je kromě samotné podstaty také demonstrováno možné využití pro geografické a geoinformační úkoly.

Většina praktických cvičení je řešena v prostředí software R, který nabízí širokou nabídku specializovaných balíčků, včetně těch určených pro práci s prostorovými daty. Další motivací k volbě tohoto software je snaha podpořit studenty k dalšímu osvojování si práce se skriptovacím jazykem, a rozšířit tak obecně jejich zdatnosti práce s kódem, které jsou v současném technologickém rozvoji nezbytné. Kromě R je dále ve cvičení používán software ArcGIS Pro, nebo GeoDa.

4.1 Regionální statistiky NUTS2 a jejich použití v příkladech

Pro praktická cvičení předmětu Pokročilé zpracování geodat je využita komplexní datová sada 24 vybraných statistických indikátorů roztříděných do ekonomické, zdravotní, sociální, environmentální a vzdělanostní dimenze. Datová sada byla původně sesbírána pro potřeby disertační práce vyučujícího (Macků, 2020). S využitím zdrojů především databáze Eurostat (doplněný o údaje z OECD Regional Database, Eumetsat, Deutscher Wetterdiens, Copernicus Land Monitoring Service, Evropská agentura pro životní prostředí a z jednotlivých národních statistických úřadů) ji tvoří 281 administrativních jednotek na regionální úrovni danou klasifikaci NUTS 2. Data jsou převážně aktuální k roku 2015, některé indikátory jsou platné k letem 2011, 2014 a 2018. Nad představenými daty se studenti postupně seznamují s vybranými prostorově-statistickými metodami. Část z nich bude nyní představena v krátkém popisu.

Nejprve studenti rozšíří své podvědomí o vybrané pokročilé metody exploratorní analýzy s důrazem kladeným na multidimenzionální povahu zpracovaných dat. Studenti si často nejsou vědomi vlivů, které odlehlé hodnoty mají na výsledky libovolných analýz. S touto motivací jsou v úvodu výuky představeny různé metody identifikace odlehlých hodnot. Pomocí výpočtu Mahalanobisovy vzdálenosti (Mahalanobis, 1936) se studenti učí jednotlivé odlehlé hodnoty (outliery) identifikovat a s uvědoměním si jejich vlivu vhodně postupovat v následných analýzách. Pro demonstraci metody byla ve cvičení vybrána data indikátorů o zdraví v Evropě (střední délka života, kojenecká úmrtnost, úmrtnost důsledkem rakoviny, úmrtnost důsledkem nemocí oběhové soustavy, kapacity zdravotnických zařízení a zdravotnického personálu), a pomocí Mahalanobisovy vzdálenosti byly hledány takové regiony, které se svými hodnotami nejvíce odlišují od průměrného, typického evropského regionu. Bylo identifikováno 23 regionů jako odlehlých, s maximální hodnotou Mahalanobisovy vzdálenosti 30,03 v bulharském regionu BG34 – Yugoiztochen. Samotná metrika umožňuje pouze identifikovat odlehlost, nijak ji neinterpretuje. K tomuto navazujícímu úkolu lze použít např. metodu *Deviating Data Cells* (Rousseeuw & Bossche, 2018), která dokáže identifikovat konkrétní odlehlé indikátory v kontextu všech vztahů mezi indikátory (tzv. *cell-wise* přístup). S doplněním o vhodné neprostorové vizualizace (např. metoda paralelních os a heatmap) se pak studenti snaží interpretovat příčiny odlehlosti jednotlivých regionů. V případě zmíněného regionu BG34 odlehlost pravděpodobně nejvíce způsobují nízká střední délka života, vysoká kojenecká úmrtnost a úmrtnost důsledkem nemocí oběhové soustavy.



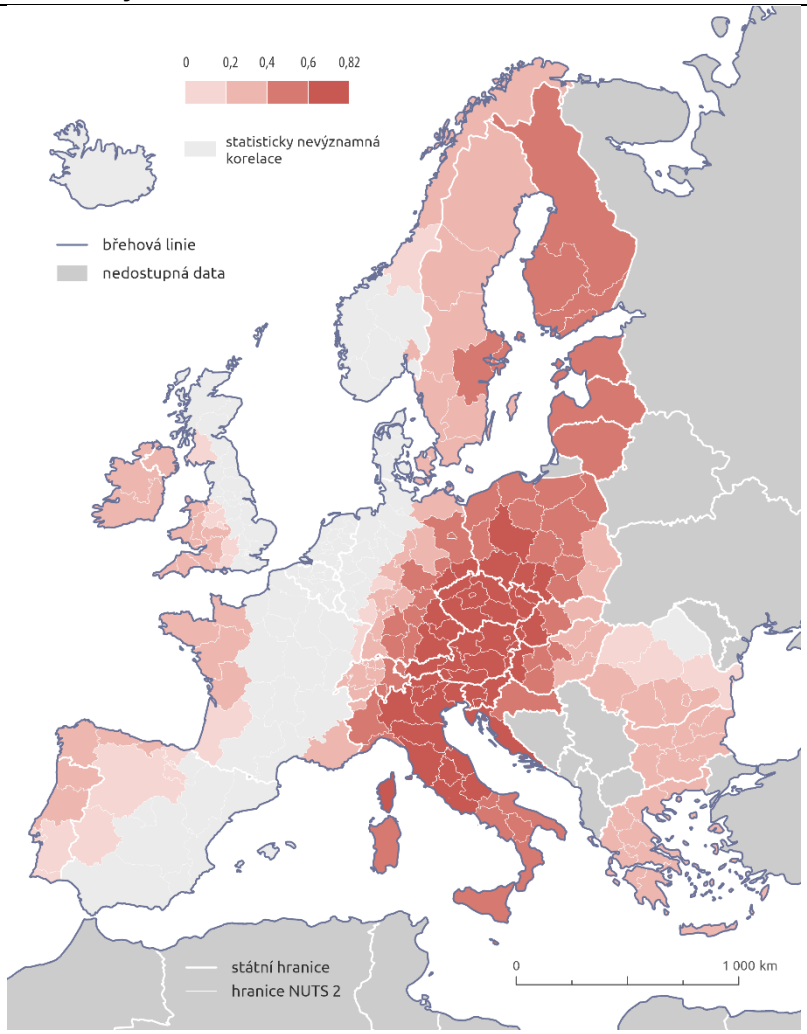
Obr. 6 Mahalanobisova vzdálenost (vlevo), identifikace prostorových shluků metodou LISA (vpravo)

Identifikace odlehlých hodnot Mahalanobisovou vzdáleností nijak nepracuje s geografickým prostorem. V další části výuky je proto tento ukazatel zobrazen v mapě, a následně je nad ním provedena analýza prostorové autokorelace, která umožní vyhledat statisticky signifikantní shluky vysokých hodnot. Ty lze pak interpretovat jako lokality, které se vymykají ostatním (výrazně se odlišují od průměru) a zároveň vykazují také prostorové seskupování. Samotná autokorelace by studentům měla být známa již z bakalářského studia, v rámci prohloubení znalostí jsou však blíže představeny např. postupy určení míry významnosti pomocí simulací metodou Monte Carlo. Při pohledu na prostorovou vizualizaci Mahalanobisovy vzdálenosti (Obr. 6 vlevo) je patrná dominance odlehlých hodnot v oblasti východní Evropy, která je podobně jako region BG34 charakteristická ne příliš dobrým zdravotním stavem obyvatelstva. S pomocí hodnocení lokální prostorové autokorelace metodou LISA je do analýzy zahrnut také prostorový aspekt, který potvrzuje vizuální inspekci a identifikuje statisticky významné prostorové shluky atributově odlehlých hodnot (Obr. 6 vpravo).

Další skupinou vyučovaných metod jsou prostorově vážené analýzy. Úvodem tématu zaznívá motivace popisující specifika prostorových dat, kdy v důsledku prostorové nestacionarity mohou být výsledky klasické inferenční statistiky lokálně zatížené chybou, případně neplatné. Jsou uvedeny základní teoretické pojmy jako nestacionarita a efekty I. a II. řádu. K úvodní demonstraci vlivu nestacionarity na výsledky statistických analýz je využita jednoduchá regresní analýza, kde po vizualizace reziduí v modelu můžeme často pozorovat tendenci k prostorovému shlukování. Po úvodní motivaci a objasnění základních aspektů prostorově vážených metod, a to tedy způsobu vymezování sousedství (topologické, vzdálenostní, fixní a adaptivní jádra) jsou představeny konkrétní metody: deskriptivní prostorově vážená statistika (prostorově vážené charakteristiky náhodných veličin, např. vyhlazené průměry), prostorově vážená korelace (Gollini et al., 2015; Kalogirou, 2012), prostorově vážená PCA (Harris et al., 2011), a samostatná lekce je pak věnována metodě prostorově vážené regrese (Brunsdon et al., 1996), která zasluhuje větší pozornost. Jako velmi zajímavá (a zároveň relativně jednoduchá) je mezi studenty hodnocena prostorově vážená korelace. Metoda hezky demonstruje lokální variabilitu, kdy např. v globálním modelu korelace můžeme nacházet statisticky nevýznamné hodnoty korelace, avšak v postupném lokálním průzkumu lze odhalit regiony, kde se charakter vzájemného vztahu dvou proměnných od globálního odlišuje (viz. Simpsonův paradox) a nacházíme zde statisticky významné hodnoty korelace.

Příkladem takové situace může být vyjádření vztahu mezi úmrtností důsledkem rakoviny a úmrtností důsledkem nemocí oběhové soustavy, který Spearmannův korelační koeficient zachycuje s hodnotou 0,2, která se dle (De Vaus, 2002) dá označit jako nízká. V lokálním pohledu však může být odhalena prostorová variabilita hodnot korelačního koeficientu. Veškeré prostorově vážené metody jsou samozřejmě velmi citlivé na nastavení vymezení sousedství, testování vlivu velikosti sousedství na výsledek by mělo být vždy součástí analýzy. Pro rychlou ukázku bylo použito adaptivní sousedství vymezené polovinou všech dostupných administrativních jednotek (180). Takovým nastavením je zahrnuto relativně velké území, výsledek je tedy shlazený, velký počet zahrnutých jednotek zároveň zajišťuje dostatečnou spolehlivost pro kvalitu lokálního odhadu. Z vizualizace na Obr. 7 lze pozorovat střední až silné hodnoty korelace mezi úmrtností důsledkem rakoviny a úmrtností důsledkem nemocí oběhové soustavy v pásu regionů táhnoucím se z České republiky na sever Itálie, ve všech směrech od tohoto jádra postupně síla korelace slábne. Zároveň je v západní části Evropy jasně zřetelný pás statisticky nevýznamných hodnot, kde se korelační koeficient blíží 0. Je vhodné upozornit, že samotné hodnoty korelace nemusí mít žádnou příčinnou souvislost, mohou zobrazovat prostý fakt vztahu hodnot sledovaných veličin v konkrétním geografickém prostoru.

Metoda lokální korelace je jednoduchým a intuitivním nástrojem zachycující vliv prostorové nestacionarity, při práci s ní je však nutno obezřetně testovat nastavení sousedství a model postupně kalibrovat. Je nezbytné mít na paměti zásadní vliv sousedství, výsledek je silně ovlivněn velikostí, tvarem, polohou regionů, hraničním efektem, nebo už samotným vymezením sousedství (adaptivní a fixní jádro, potažmo samotná jádrová funkce) který pracuje pouze s adaptivní vzdáleností, nebere proto v potaz členitý charakter území, který je typický pro evropské regiony. Bližší rozbor metod a postupů není předmětem tohoto příspěvku, zájemci však mohou zabrousit do prací (Gollini et al., 2015; Kalogirou, 2012).



Obr. 7 Hodnoty prostorově váženého korelačního koeficientu mezi úmrtností důsledkem rakoviny a úmrtnosti důsledkem nemocí oběhové soustavy

Představený text pouze stručně popisuje vybrané aktivity z výuky předmětu Pokročilé zpracování geodat. Cílem této demonstrace je přiblížit čtenáři snahu o inovaci výuky, která má svou první úroveň samozřejmě v samotných vyučovaných metodách, které rozšiřují znalosti studenta nad rámec běžného GIS operátora a podporují jeho analytické myšlení ve schopnosti více si uvědomovat prostorové závislosti různých jevů popsaných statistickými daty. Druhá úroveň podporuje myšlenku projektu ERASMUS+ Jean Monnet Module ve snaze přinést do výuky více reálných dat a zvýšit takto regionální geopolitické povědomí a poznání studentů o evropském prostředí, v jehož centru se Česko nachází.

5. ZÁVĚR

Příspěvek popisuje první zkušenosti s nasazením evropských dat do výuky dvou geoinformatických předmětů. Prezentované příklady z výuky demonstrují hlavní přínosy implementace těchto dat do výuky. Především se jedná o reálná geografická data evropského kontinentu, kde studenti žijí, která navíc svými tématy překračují národní rámec a rozvíjejí pochopení vzájemných rozdílů a podobností mezi jednotlivými evropskými regiony. Díky tomu se zvyšuje regionální geopolitické povědomí a poznání studentů o evropském prostředí. Další výhodou je fakt, že data jsou volně ke stažení v libovolných formátech. Přínosem je také seznámení se s formou metadat. Další výhodou je aktuálnost dat, kdy kontinuálně přibývají nová data. Lze tak porovnávat různá časová období (např. Urban Atlas 2012 a 2018; nové hodnoty časových řad Eurostat). V neposlední řadě je i výhodou, že studenti mohou na dalších datech procvičovat v rámci domácích, či semestrálních úloh a nejsou tak limitováni na prezentované datové sady ve cvičení. Autoři předpokládají další rozšíření a přidání nových praktických příkladů, tak aby evropská geografická data byla přínosem pro studovaný obor Geoinformatika a kartografie. Předpokládáme, že znalosti a zkušenosti studentů se promítnou i do zpracování

témat diplomových prací. V plánu je zkoumat sousednost ploch různého využití území evropských měst z Urban Atlasu pomocí frekventovaných sad a asociačních pravidel. Dále plánujeme zjišťování kolokačních vzorů nad evropskými meteorologickými daty z databáze European Severe Weather Database. Vzhledem k tomu, že prezentované poznatky a odkazované články jsou volně ke stažení na webu projektu UrbanDM.upol.cz, tak je možné je využít ve studiu, samostudiu na jiných vysokých školách či ve vzdělávání odborné veřejnosti.

6. PODĚKOVÁNÍ

Inovace předmětů o zpracování dat poskytovaných Evropskou unií je podpořena projektem ERASMUS+ Jean Monnet Module No. 620791-EPP-1-2020-1-CZ-EPPJMO-MODULE, Data mining and analyzing of urban structures as contribution to European Union studies.

Článek byl také podpořen v rámci projektu „Analýza, modelování a vizualizace prostorových jevů pomocí geoinformačních technologií“ (IGA_PrF_2022_027) za podpory interní grantové agentury Univerzity Palackého v Olomouci.

7. LITERATURA

Berka, P. (2005). *Dobývání znalostí z databází*. Academia.

Biolab. (2019). *Orange3 Image Analytics Documentation*.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

CopernicusProgramme. (2020). *Urban Atlas*.

De Vaus, D. (2002). *Analyzing social science data*. SAGE Publications Ltd.

Dobesova, Z. (2019). The Similarity of European Cities Based on Image Analysis. In *Advances in Intelligent Systems and Computing* (Vol. 1046). https://doi.org/10.1007/978-3-030-30329-7_31

Dobesova, Z. (2020). Experiment in Finding Look-Alike European Cities Using Urban Atlas Data. *ISPRS International Journal of Geo-Information*, *9*(6), 406. https://doi.org/10.3390/ijgi9060406

Dobešová, Z. (2022). ORANGE Praktický návod do cvičení předmětu Data Mining. Univerzita Palackého v Olomouci.

Eurostat. (1990). *NACE Rev. 1 - Statistical classification of economic activities in the European Union*.

Eurostat. (2020). *Statistics explained, Glossary: Functional urban area*. Eurostat.

Eurostat. (2021a). *Eurostat database*.

Eurostat. (2021b). *Eurostat database*.

Eurostat. (2022a). *National accounts employment data by industry*.

Eurostat. (2022b). *Passengers transported (detailed reporting only) - (quarterly data)*.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). GWmodel : an R package for exploring spatial heterogeneity. *Journal of Statistical Software*, *63*(17), 1–50. https://doi.org/10.1080/10095020.2014.917453

Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, *25*(10), 1717–1736. https://doi.org/10.1080/13658816.2011.554838

Kalogirou, S. (2012). Testing local versions of correlation coefficients. *Jahrbuch Für Regionalwissenschaft*, *32*(1), 45–61. https://doi.org/10.1007/s10037-011-0061-y

Macků, K. (2020). *Multidisciplinární hodnocení kvality života v Evropě na regionální úrovni*. Univerzita Palackého v Olomouci. https://doi.org/10.5507/prf.20.24458410

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences*, *2*.

Masopust, J., Dobesova, Z., & Mack\ru, K. (2021). Utilisation of EU Employment Data in Lecturing Data Mining Course. In R. Silhavy (Ed.), *Artificial Intelligence in Intelligent Systems* (pp. 601–616). Springer International Publishing.

Openshaw, S., & Abrahart, R. J. (1996). Geocomputation. *Proceedings 1st International Conference on Geocomputation*.

Orange. (2021). Orange, Data Mining Fruitful and Fun; University of Ljubljana.

Orange Visual Programming Documentation. (2021).

Pászto, V., Burian, J., & Macků, K. (2020). COVID-19 data sources : evaluation of map applications and analysis of behavior changes in Europe ' s population. *GEOGRAFIE*, *125*(2), 38. https://doi.org/https://doi.org/10.37040/geografie2020125020171

Pászto, V., Redecker, A., Mack\ru, K., Jürgens, C., & Moos, N. (2020). Data Sources. In V. Pászto, C. Jürgens, P. Tominc, & J. Burian (Eds.), *Spationomy: Spatial Exploration of Economic Data and Methods of Interdisciplinary Analytics* (pp. 3–38). Springer International Publishing. https://doi.org/10.1007/978-3-030-26626-4_1

Rousseeuw, P. J., & Bossche, W. Van Den. (2018). Detecting Deviating Data Cells. *Technometrics*, *60*(2), 135–145. https://doi.org/10.1080/00401706.2017.1340909

Šarmanová, J. (2012). *Metody analýzy dat*. Vysoká škola báňská Technická univerzita Ostrava.