

VYUŽITÍ DAT DATABÁZE EUROSTAT A URBAN ATLAS VE VÝUCE PŘEDMĚTU DATA MINING

USE OF EUROSTAT AND URBAN ATLAS DATA IN DATA MINING COURSE

Zdena Dobešová

Abstrakt

Magisterské studium Geoinformatika a kartografie má zařazen předmět Data Mining. V rámci praktického procvičování témat byla využita data statistického úřadu Evropské unie Eurostat a dále data programu Copernicus Land Monitoring Service – Urban Atlas. Jedná se o volně dostupná data, která jsou průběžně aktualizována. Zpracování dat přináší studentům znalost o evropských geografických tématech, která souvisí se zaměřením jejich studia. Pro výuku shlukové analýzy byla použita data průměrné týdenní pracovní doby a podíl pracovníků v jednotlivých odvětvích. Analýza časových řad je procvičována na datech železniční dopravy. Neuronové sítě nebo frekventované sady lze aplikovat na data Urban Atlas o využití území Evropských měst.

Klíčová slova: Eurostat, Urban Atlas, Data Mining, geoinformatika

Abstract

The Master's degree in Geoinformatics and Cartography includes Data Mining course. The data of the Statistical Office of the European Union Eurostat and the data of the Copernicus Land Monitoring Service - Urban Atlas were used in practical lectures. Mentioned data are freely available. Data processing provides students with knowledge of European geographical topics related to their study. The data on average weekly working hours and the share of workers in each industry were used for learning cluster analyses. Time series analysis is practiced on rail transport data. Neural networks or frequent datasets can be applied to Urban Atlas data on land use of European cities.

Keywords: Eurostat, Urban Atlas, Data Mining, Geoinformatics

Úvod

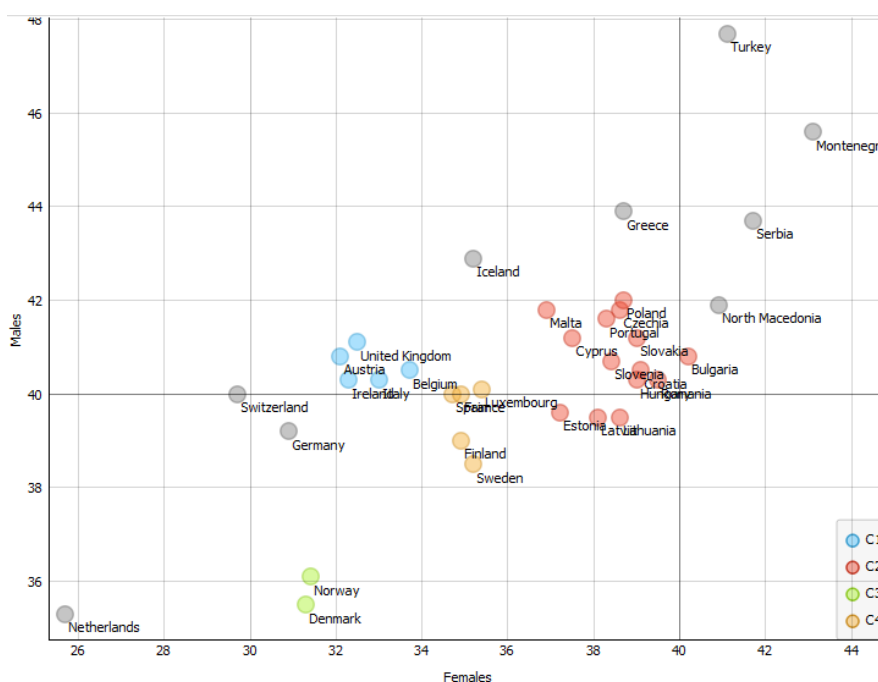
Praktická výuka na konkrétních datech je nezbytnou součástí výuky a napomáhá k pochopení teoretického výkladu jednotlivých metod. Při hledání vhodných aktuálních dat pro předmět Data Mining byly vytipována statistická databáze Evropské Unie Eurostat a její jednotlivé statistiky jako např. přeprava cestujících (Eurostat 2022). Druhým vhodným zdrojem z hlediska oboru Geoinformatika a kartografie, kam je předmět Data Mining zařazen, se jevila databáze projektu Copernicus – Urban Atlas (UA) (Copernicus Programme 2020), kde jsou dostupná data o využití území v městských funkčních oblastech. Vzhledem k tomu, že výuka předmětu Data Mining je převážně realizována v softwaru Orange, tak byla pro praktickou část cvičení nachystána učebnice s ukázkou praktických příkladů (Dobešová 2022). Některé příklady v učebnici právě čerpají data z výše uvedených zdrojů. Navíc je dostupná i cvičebnice pro samostatné procvičování úloh, která i vysvětluje postup a získávání dat v uvedených dvou

databází. Obě učebnice jsou ke stažení na webu projektu <http://UrbanDM.upol.cz>. V textu tohoto příspěvku budou nastíněny některé vybrané úlohy, které jsou procvičovány ve výuce a vychází z dat ze zmíněných databází.

1 Použití dat z databáze Eurostat

Pro základní úlohy inspekce dat, jako je zjišťování minimálních, maximálních, průměrných, mediánových hodnot, četností, chybějících hodnot, odlehlých hodnot lze použít řadu dat z databáze Eurostat. Každá datová sada má unikátní označení kódem a tak lze data aktualizovaná data opakovaně vyhledávat, odkazovat a stahovat, popř. upravovat výběry.

Jednou z vhodných datových sad je sada s kódem LFSA_EWHUN2, která obsahuje *průměrný počet odpracovaných hodin v týdnu* v každé zemi. Počty pracovních hodin jsou uvedeny samostatně pro ženy, muže a celkem. Na těchto datech lze dobře demonstrovat odlehlé hodnoty, kdy v Turecku mají muži nejdelší pracovní týden a to skoro 48 hod. Pomocí krabicového grafu lze zobrazit rozpětí hodnot, zajímavé je i použití houslového grafu, pro základní představu o datech. V rámci předmětu Data mining se jako typická úloha procvičuje shlukování. Na těchto datech lze dobře vyšetřit shluky podobných států v Evropě, které se podobají počtem pracovních hodin jak pro muže, tak pro ženy. Prakticky byla použita shlukovací metoda k-Means a DBSCAN (Obr. 1). Studenti nejen procvičují postupy zpracování dat ale hlavně i interpretaci výsledků a vhodnost použitých metod a nastavení jejich parametrů.



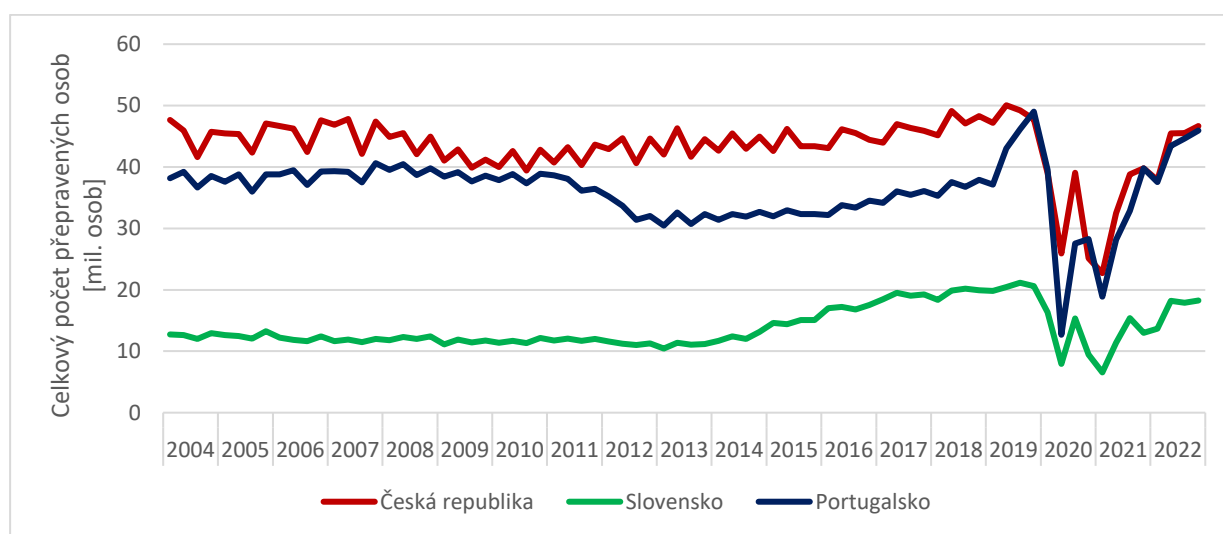
Obrázek 1: Shlukování metodou DBSCAN států EU podle průměrného počtu pracovních hodin v týdnu mužů a žen (parametry Core point Neighbors=2, Distance=1)

Druhý příklad dobrého zdroje jsou data Eurostatu o *počtu pracovníků v jednotlivých odvětvích* (NACE) s kódem NAMA_10_A64_E. Státy Evropy se výrazně liší počtem pracovníků v jednotlivých odvětvích. Zde už se jedná o multidimenzionální data, kdy na první úrovni členění je každý stát popsán počtem pracovníků v 21 typech odvětví (Masopust et al. 2021). Zde si studenti musí nejprve absolutní počty pracovníků přepočítat na relativní hodnoty podle

počtu obyvatel v produktivním věku. Tato úprava spadá pro předpřípravy dat, která je také teoreticky přednášena v úvodních hodinách předmětu. Na tato data lze dobře aplikovat analýzu hlavních komponent. Zajímavé jsou korelace, kdy např. pozitivně korelují počty zaměstnanců v odvětví „školství“ a odvětví „umění, zábava, rekreace“. Díky hierarchickému shlukování lze určit státy, které jsou si podobné počtem pracovníků ve stejných odvětvích. Naopak interpretačně zajímavé jsou odlehlé státy jako je Lucembursko nebo Rumunsko (Dobešová et al. 2022).

Databáze Eurostat poskytuje statistiku *počtu cizinců žijících v jednotlivých zemích EU*. Statistika je z roku 2011. Pro každý stát jsou uvedeny počty cizinců podle kontinentu, odkud do konkrétní země přišli. Na těchto datech lze vyšetřovat podobné země podle složení cizinců.

Pro téma zpracování časových řad lze v databázi Eurostat nalézt řadu vhodných dat. Ve výuce byla použita data o *počtu přepravených osob po železnici*, která je uváděna po čtvrtletích. V datech je dobře viditelný propad přepravy v pandemickém roce 2020 a 2021 (Obr. 2).



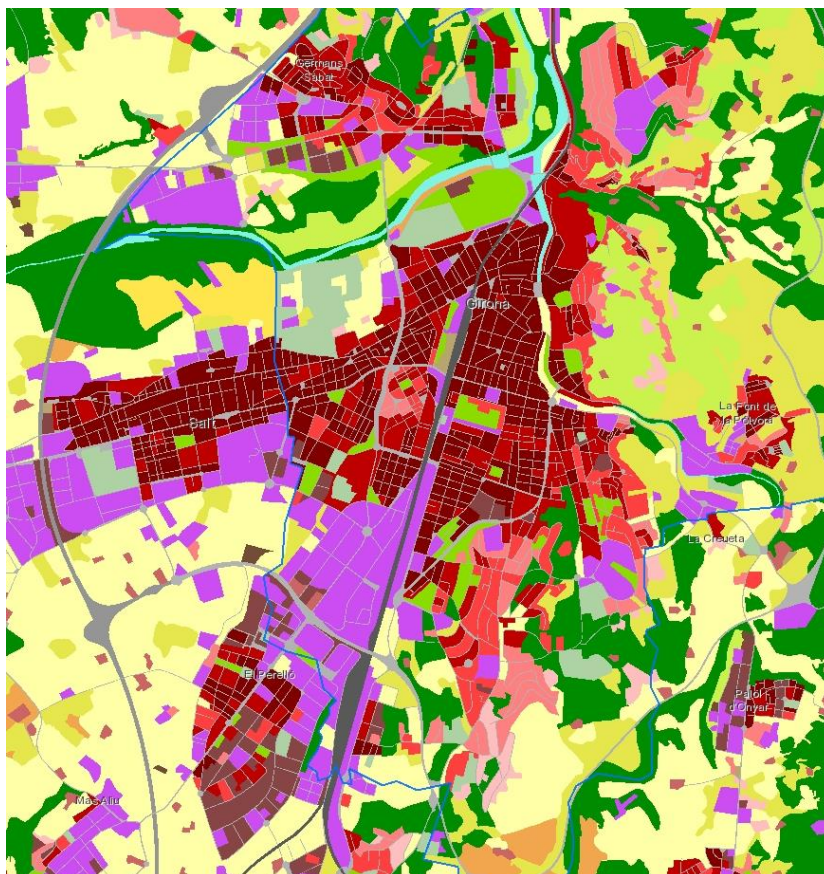
Obrázek 2: Vývoj počtu přepravených osob od roku 2004 do roku 2022 v České republice, na Slovensku a v Portugalsku (Eurostat 2022)

2 Použití dat Urban Atlas

Projekt Urban Atlas poskytuje data u využití území ve skoro 800 městských funkčních oblastech (tzv. FUA) v rámci celé Evropské Unie. Jedná se o prostorová data, kde je pro verzi 2018 vymezeno 27 kategorií využití území, jako jsou různé plochy s různou hustotou zástavby, silnice, plochy zeleně, vodní plochy, komerční průmyslové a vojenské plochy, různé druhy ploch zemědělsky využívaných až po plochy letišť (Obr. 4).

Na těchto datech lze nachystat řada úloh. Od statistických ukazatelů o počtu, rozloze jednotlivých ploch až po využití neuronových sítí pro zjištění prostorových vzorů. Data byla použita pro demonstraci nalezení podobných měst podle prostorového uspořádání ploch. Jako vstup byly použity kruhové mapové výřezy center měst předchystané v ArcGIS softwaru (Janoušek 2019). Software Orange disponuje doplňkem, který nabízí natrénované neuronové sítě, kde nejlepší výsledky pro tato data přinášela síť Painters. Studenti jsou tímto příkladem uvedeným v učebnici (Dobešová 2022) seznámeni s výsledky zkoumání prezentovaným v publikovaných vědeckých článcích (Dobesova 2019) (Dobesova 2020).

Data byla využita i pro semestrální úlohu, kdy každý student zpracoval několik vybraných měst.



Obrázek 3: Zdrojová data Urban Atlas pro město Girona ve Španělsku

Data Urban Atlas byla použita i na zjišťování frekventovaných sad sousednosti ploch s různým využitím. Data mining metoda asociačních pravidel a frekventovaných sad je v rámci předmětu také přednášena. Frekventované sady, které vycházejí z prostorové sousednosti byly řešeny v rámci diplomové práce (Novák 2023). Jednalo se o aplikaci analýzy nákupního košíku (MBA) a generování transakcí sousednosti polygonů. Předpokládám zařadit postup pro zjišťování frekventovaných sad navrženého v diplomové do výuky v nadcházejícím akademickém roce.

Závěr

Použití dat Eurostat a Urban Atlas ve výuce přináší několik výhod. Data jsou volně dostupná ke stažení. Databáze jsou průběžně aktualizované a jsou tak dostupná aktuální data. V případě časových řad dopravy se jedná např. o čtvrtletní data od roku 2004. Další výhodou je, že se jedná o reálná data na rozdíl od „umělých“ trénovacích datasetů dodávaných často k učebnicím či softwarům. Výhodou je velký rozsah dat, kdy lze zadávat samostatné práce jednotlivým studentům ze cvičení, či ve formě semestrálních úloh, kdy každý student pracuje se svým zadaným výběrem dat. Stažená data obsahují metadatový popis o významu dat, časovém rozsahu atd. Což nepřímo edukuje studenty mimo náplň předmětu, kdy pracují se správným metadatovým popisem dat (např. souřadnicový systém u dat Urban Atlas). Použití těchto dat přináší povědomí o Evropském regionu a v řadě socioekonomických témat. Přínosná pro rozhled studentů je i interpretace dat, kdy na základě zjištěných faktů mohou kombinovat různé informace. Studenti tak získávají nejen dovednosti v použití metod data mining, ale rozšiřují se jejich znalosti v geografii a geoinformatice. Další zdroje vhodných dat lze nalézt v knize (Pászto et al. 2020). Doporučit lze YouTube kanál sw Orange (Pretnar 2016).

Poděkování

Výuka je podpořena projektem Erasmus+ Programme of the European Union, Jean Monnet Module, Project No. 620791-EPP-1-2020-1-CZ-EPPJMO-MODULE "Data mining and analyzing of urban structures as contribution to European Union studies".

Literatura

- Copernicus Programme (2020). Urban Atlas. [5-5-2022] Dostupné z: <https://land.copernicus.eu/local/urban-atlas>.
- Dobesova Z. (2019). The Similarity of European Cities Based on Image Analysis. In: Silhavy R., Silhavy P. (ed) Advances in Intelligent Systems and Computing. Springer, Cham, 341–348.
- Dobesova Z. (2020). Experiment in Finding Look-Alike European Cities Using Urban Atlas Data. ISPRS Int J Geo-Information 9:406. <https://doi.org/10.3390/ijgi9060406>
- Dobešová Z. (2022) ORANGE, Praktický návod do cvičení předmětu Data Mining. Univerzita Palackého v Olomouci, Olomouc, Dostupné z: <https://www.vydavatelstviupol.cz/cz/978-80-244-6086-4>
- Dobešová Z., Macků K., Kučera M. (2022) Výuka geoinformatických předmětů na příkladech dat Evropské unie. In: Sympozium GIS Ostrava 2022. Ostrava
- Eurostat (2022) Passengers transported (detailed reporting only) - (quarterly data). [12-10-2022] Dostupné z: https://ec.europa.eu/eurostat/databrowser/product/page/RAIL_PA_QUARTAL.1
- Janoušek M. (2019). Porovnání urbánního prostoru pomocí kruhových výsečí. Univerzita Palackého v Olomouci, Diplomová práce, [7-21-2023]. Dostupné z: <https://theses.cz/id/w91ot7/>
- Masopust J., Dobesova Z., Macků K. (2021). Utilisation of EU Employment Data in Lecturing Data Mining Course BT - Artificial Intelligence in Intelligent Systems. In: Silhavy R (ed). Springer International Publishing, Cham, 601–616
- Novák P. (2023) Porovnání sousednosti využití území evropských měst pomocí frekventovaných sad. Univerzita Palackého v Olomouci, Diplomová práce. [7-21-2023]. Dostupné z: <https://theses.cz/id/h6kc79/>
- Pászto V., Redecker A., Macků K, et al. (2020). Data Sources. In: Pászto V., Jürgens C., Tominc P., Burian J. (eds) Spationomy: Spatial Exploration of Economic Data and Methods of Interdisciplinary Analytics. Springer International Publishing, Cham, 3–38
- Pretnar A (2016). All I See is Silhouette. [11-9-2021] Dostupné z: <https://orangedatamining.com/blog/2016/03/23/all-i-see-is-silhouette/>.

Kontaktní údaje

doc. Ing. Zdena DOBEŠOVÁ, Ph.D.
Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého
17. listopadu 50, 771 46 Olomouc
e-mail: zdena.dobesova@upol.cz