ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA DOPRAVNÍ

Bc. Oliver Pulda

Statistické modelování dopravních parametrů pomocí klastrovacích metod

Diplomová práce

2023

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta dopravní
děkan
Konviktská 20, 110 00 Praha 1

K620..... Ústav dopravní telematiky

ZADÁNÍ DIPLOMOVÉ PRÁCE
(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení studenta (včetně titulů):

Bc. Oliver Pulda

Studijní program (obor/specializace) studenta:

navazující magisterský – IS – Inteligentní dopravní systémy

Název tématu (česky): **Statistické modelování dopravních parametrů pomocí klastrovacích metod**

Název tématu (anglicky): Statistical modeling of traffic parameters using clustering methods

Zásady pro vypracování

Při zpracování diplomové práce se řiďte následujícími pokyny:

- Popis relevantních naměřených dat z dopravních detektorů a parametrů měřených lokací
- Příprava a filtrace datových vzorků k analýze
- Volba relevantních klastrovacích metod pro datovou analýzu
- Implementace zvolených klastrovacích metod na připravené datové vzorky
- Tvorba statistického modelu dopravních parametrů na základě integrace výsledků klastrovacích metod
- Validace modelů, rozbor a stanovení výsledků

Rozsah grafických prací:

Rozsah průvodní zprávy: minimálně - doplňte počet stran (BP = 35; DP = 55) - stran textu (včetně obrázků, grafů a tabulek, které jsou součástí průvodní zprávy)

Seznam odborné literatury: D. T. Larose. Discovering Knowledge in Data. An Introduction to Data Mining. Willey, 2005.

I. Nagy, E. Suzdaleva. Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components, Springer, 2017.

Vedoucí diplomové práce: **Ing. Patrik Horažďovský, Ph.D.**

doc. Ing. Evženie Uglickich, CSc.

Datum zadání diplomové práce: **30. června 2022**
(datum prvního zadání této práce, které musí být nejpozději 10 měsíců před datem prvního předpokládaného odevzdání této práce vyplývajícího ze standardní doby studia)

Datum odevzdání diplomové práce: **15. května 2023**
a) datum prvního předpokládaného odevzdání práce vyplývající ze standardní doby studia a z doporučeného časového plánu studia
b) v případě odkladu odevzdání práce následující datum odevzdání práce vyplývající z doporučeného časového plánu studia

L. S.

Ing. Zuzana Bělinová, Ph.D.
vedoucí
Ústavu dopravní telematiky

prof. Ing. Ondřej Přibyl, Ph.D.
děkan fakulty

Potvrzuji převzetí zadání diplomové práce.

Bc. Oliver Pulda
jméno a podpis studenta

V Praze dne ..... 3. července 2022

Poděkování

Chtěl bych poděkovat svým oběma vedoucím Ing. Patriku Horažďovskému Ph.D. a doc. Ing. Evženie Uglickich CSc. za jeji pomoc a podporu

Čestné prohlášení

Předkládám tímto k posouzení a obhajobě diplomovou práci, zpracovanou na ČVUT v Praze Fakultě dopravní. Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných pracích. Nemám žádný důvod proti užití tohoto školního díla ve smyslu § 60 Zákona Č. 121/2000 Sb., o právu autorském, o právech související s právem autorským a o změně některých zákonů.

V Praze, 15.5.2023

Oliver Pulda

Vaše jméno

ABSTRAKT

Cílem této práce je pomocí metod strojového učení nalézt vlastnosti lokali, které mohly mít vliv na chování řidiče. Budou využity klastrovací metody, K-means, Fuzzy c-means, DBSCAN, Spektrální klastrování a Hierarchické klastrování. Následně bude jedna z variant vybrána pro validaci pomocí Naivního Bayese, který pomůže definovat ideální k, a sním ideální mix rozdělení dat. Na základě informací z metod učení bez učitele a učení s učitelem, bude nalezena ideální kombinace lokalit, ve kterých se naleznou propojující prvky.

ABSTRACT

The goal of this study is to identify features of locations that could influence driver behavior using machine learning methods. Clustering methods such as K-means, Fuzzy c-means, DBSCAN, Spectral clustering, and Hierarchical clustering will be utilized. Subsequently, one of the variants will be selected for validation using Naive Bayes, which will help define the ideal k, and with it, the ideal mix of data distribution. Based on the information from unsupervised and supervised learning methods, the ideal combination of locations will be found where connecting elements can be identified.

Obsah

Úvod

Prvotní myšlenka k této práci je navázána na zdroj dat, který byl za účelem zkoumání rychlosti a dopravních intenzit v okolí. Zadavatelem těchto měření byla nejčastěji obec, která chtěla znát dopravní realitu v obci na daném úseku. Kolegové na pracovišti v Děčíně si všimli, že při měření různých oblastí se dopravní charakteristiky liší. Byla vedena diskuse, z jakého důvodu tato realita nastává. Měřené oblasti byly různého druhu, jednalo se o malé obce, velké obce, obce v blízkosti okresního města anebo v odlehlých oblastech. Při bližším zkoumání bylo možné si všimnout nejen rozdílů, ale také podobností. Na některých místech je možné si všimnout, že se třeba vyskytuje osvětlený přechod pro chodce, a jinde ne. Na některých místech bylo možné identifikovat jako společný prvek svodidla a směrové oblouky. Úvaha byla, jestli tyto parametry komunikace ovlivňují řidiče a tím pádem chování vozidla, a jak takovou provázanost odhalit.
Metoda, kterou si řešitel vybral pro odhalování společných charakteristik, jsou metody strojového učení, konkrétně s učitelem (Naivní Bayes) a bez učitele, konkrétně: K-measn, DBSCAN, Fuzzy c-means, Spektrální klastrování a Hierarchické klastrování. Metody bez učitele pomůžou nalézt skupiny oblastí, podle které jsou nějakým způsobem datově podobné. Metody s učitelem následně použijí tyto výstupy a pokusí se je validovat. Výstupem budou skupiny dat, které podle dané metody spolu nějak souvisí. Zároveň tento pohled bude konfrontován s dopravně odbornými fakty, se snahou nalézt průnik řešení. Výsledkem by mělo být nalezení nejlepší klastrovací metody a obecných prvků v obcích, které jsou pro vybrané lokality společné. Důležité je doplnit, že data samotná nebyla měřena pro tento typ analýzy, ani se neočekávalo, že takový typ vyhodnocení bude zpracován. Pro zpracování a vyhodnocení dat byly použity programovací jazyky Python a Scilab.

Měřené oblasti

Pracoviště ČVUT FD v Děčíně na základě žádosti municipalit nasbíralo v 16 různých obcích a jejich částech dopravní data, v některých případech bylo provedeno měření dvakrát. Data měření stejného místa dvakrát nebyla spojována kvůli předpokladu, že pokud oblasti mají působit nějakým stylem na řidiče, tak na ně budou působit stejně i při opakovaném měření. Sloužily tudíž jako ukazatel, jak moc opravdu oblasti ovlivňují chování řidiče.

2.1 Bynov

Místo měření bylo prováděno v obci Děčín v lokalitě historicky známé jako Bynov. Zařízení se nalézalo v ulici Teplická mezi křižovatkami Bynovská a Široká. Jedná se o hlavní komunikaci, která je vedena Děčínem až k řece Labe. Komunikace patří mezi významné dopravní tepny obce, je využívána jak místními řidiči tak tranzitní dopravou. Je proto možné očekávat vysoké dopravní hustoty během špiček. Měřená oblast má v blízkosti obchod s potravinami, do kterého je vyčleněn samostatný pruh pro odbočení doleva. Stejně tak je i samostatný pruh pro odbočení doleva při vjezdu do ulice Bynovská. Komunikace je obousměrná, přehledná, dobře osvětlená a rozhledové podmínky jsou dostatečné. Měření proběhlo v období 13.12.2022 - 19.12.2022.

2.2 Bynovec

Obec se nalézá blízko CHKO Labské pískovce v Ústeckém kraji, v Děčínském okrese. Jedná se o obec malé velikosti. Měřená komunikace spadá do kategorie III. třídy. Silnice není příliš osvětlená, pokud pojedou proti sobě široká vozidla, bude problematické projet. Místo měření je blízko výjezdu z lesa do obce. Komunikaci nekříží chodníky, jen účelové komunikace. Až na druhém konci obce se dělí na další směry. Měření proběhlo v období 14.8.2021 - 20.8.2021 a 14.10.2021 - 20.10.2021.

2.3 Libouchec

Obec se nalézá blízko CHKO České středohoří v Ústeckém kraji, v Děčínském okrese. Silnice spadá do kategorie I. třídy. Měření probíhalo v blízkosti křížení účelové komunikace, které je osazeno návěstidly. Na silnici kvůli častým směrovým obloukům není možné předjíždět, bude nejčastěji využívána pro tranzitní dopravu. Měření proběhlo v období 4.1.2022 - 10.1.2022.

2.4 Krásná Lípa

Obec se nalézá v Ústeckém kraji, v Děčínském okrese. Místo měření se nachází na komunikaci II. třídy v ulici Varnsdorfská, blízko konce obce. I když spadá komunikace do druhé kategorie, je poměrně úzká a v měřené oblasti má minimální prostor po stranách. Naopak, jsou zde svodidla, která případné míjení širokých vozidel zhoršují. Na měřenou silnici jsou primárně připojeny účelové komunikace, které slouží k obsluze domů v obci. Silnice obsahuje mnoho směrových oblouků, které znemožňují dobrý rozhled pro řidiče vozidla. Komunikace bude využívána převážně místním obyvatelstvem a pro tranzit. Měření proběhlo v období 12.6.2021 - 18.6.2021 a 12.7.2021 - 18.7.2021.

2.5 Staré Křečany

Obec se nalézá v Ústeckém kraji, v Děčínském okrese. Obec se nachází poblíž města Rumburk, kam místní dojíždějí do práce nebo služeb. Měření se provádělo v oblasti zvané Panský, u konce obce směrem na Brtníky. Lokace je řídce zastavěná, neosvětlená a okolí komunikace je nebezpečné pro vysokou jízdní rychlost. V přilehlém prostoru je mnoho vzrostlých stromů a sloupů vysokého napětí. Silnice spadá do kategorie III. třídy. Měření proběhlo v období 21.6.2021 - 27.6.2021 a 11.9.2021 - 17.9.2021.

2.6 Rohatce

Rohatce jsou část obce Hrobce v Ústeckém kraji v Děčínském okrese. Jedná se o obec s 88 domy a 271 obyvateli. Měřená oblast se nacházela na komunikaci III. třídy, která se v jižní části obce rozděluje do 3 směrů. Jak již bylo uvedeno, obec je řídce zastavěná, kolem vozovky je dostatek místa pro vyhnutí se dvou nadrozměrných vozidel. Silnice protíná obec směrovým obloukem, na který se napojují účelové komunikace. V obci je nové moderní osvětlení, které je poměrně hustě rozmístěno. Měření proběhlo v období 2.8.2021 - 8.8.2021.

2.7 Nebočany

Nebočany byly připojeny k obci Děčín, nacházejí se v jeho jižní části. Je zde vedena komunikace II. třídy, která prochází zmíněnou částí Děčína. Silnice propojuje město s Ústí nad Labem, komunikace bude hodně využívána transitně a pro místní obyvatele. Silnice je dobře osvětlená, je obklopena zástavbou a na několik místech je křižována vedlejšími komunikacemi. Měření proběhlo v období 3.2.2021 - 9.2.2021.

2.8 Děčín Ústecká

Děčín Ústecká je silnice na levém břehu Labe, která propojuje město s Ústím nad Labem po pravém korytu řeky. Jedná se o komunikaci II. třídy, dobře osvětlenou s dobrými rozhledovými poměry. Kolem komunikace je hodně prostoru a v některých případech je úplně vyčleněn odbočující pruh doleva. Na silnici v místě měření je zúžení ostrůvkem. Měření proběhlo v období 20.1.2022 - 26.1.2022.

2.9 Děčín Teplická

Měřený úsek se nachází na stejné komunikaci jako v případě měření Bynov. Toto měření bylo ale umístěno blíže centru, mezi ulice Na Hrachách a Osecká. Úsek je přímý, vedený kolem mírné zástavby, před kterou je chodník pro chodce. Silnici protínají dva přechody pro chodce s vlastním osvětlením. K ulici jsou napojeny místní a obslužné komunikace. Silnice je dobře osvětlená a přehledná. Bude využívána pro cesty do centra a dál na východ nebo na sever. Měření proběhlo v období 12.1.2022 - 18.1.2022.

2.10 Česká Kamenice Děčínská

Zkoumaný úsek je silnice č. 13, komunikace II. třídy. Na vybraném místě se téměř nenacházejí stavby. Ty jsou až dále do obce, za směrovým obloukem. V blízkém okolí silnice je možné vidět hustou vegetaci a svah. Komunikace je mírně osvětlena starými výbojkovými lampami. V místě není možné provádět úkon předjíždění, a na silnici nesmí v měřeném místě vjíždět zemědělská vozidla a povozy tažené koňmi. Rozhledové a prostorové podmínky jsou dobré, vozidla vidí daleko vpřed. Měření proběhlo v období 23.10.2021 - 29.10.2021.

2.11 Česká Kamenice Líska

Místo měření v České Kamenici Líska bylo umístěno v blízkosti křížení komunikací. Jedná se o silnici II. třídy, komunikace je jak před, tak i za místem měření ve směrových obloucích. Ve vybraném místě není možné předjíždět ostatní vozidla. Okolí silnice je spíše mírně prostorné, v některých případech dokonce ohraničeno svodidly a hustou vegetací, zároveň je v některých místech méně osvětlená. Komunikace je dostatečně široká pro nákladní vozidla. Měřený úsek se nachází v blízkosti vjezdu do obce. Měření proběhlo v období 26.4.2021 - 2.5.2021.

2.12 Česká Kamenice Dukelských Hrdinů

Jedná se o stejnou silnici jako v případě Česká Kamenice Děčínská. Ovšem měření bylo prováděno blíže k centru u obchodu s potravinami. V přilehlé blízkosti silnice se nacházejí budovy a chodníky, silnice je křižována přechodem. V celé délce komunikace se nacházejí tři křižování s vedlejšími komunikacemi. Vozovka je dobře osvětlena a jsou na ní dobré rozhledové poměry. Budou ji využívat místní a lidé z okolí, kteří jedou za službami a transit. Měření proběhlo v období 10.11.2021 - 16.11.2021.

2.13 Hrobce

Silnice III. třídy vedena obcí Hrobce je přímá, má spoustu prostoru v blízkém okolí. V místě měření není vedena hustou zástavbou. Je dobře osvětlena. Úsek, kde bylo prováděno měření se nachází hned u vjezdu do obce. Komunikace je dostatečně široká i pro nákladní vozidla. Měření proběhlo v období 23.7.2021 - 30.7.2021.

2.14 Huntířov

Vybraný úsek je měřen na komunikaci II. třídy, která je vedena obcí Huntířov. Specifikum oblasti je, že délka úseku, který musí vozidlo ujet aby se vjelo a vyjelo z obce a zároveň zůstalo na silnici č. 13, je velmi krátká. Vjezd do obce je vždy ze směrového oblouku po kvalitní komunikaci, která je v jednom případě až tříproudá. Poblíž míst měření je přechod pro chodce, který je zvýrazněn a ze jednoho směru je před něj položena značka "ochranný pás". V místě přechodu se nalézá i křižení. Měření proběhlo v období 22.5.2022 - 28.5.2021.

2.15 Huntířov Nová Oleška

Nová Oleška je oddělená část Huntířova, kterou je vedena komunikace III. třídy. Lokalita je prostorově problematická, silnice je úzká a nemá prostor v přilehlém okolí. V některých případech je dokonce ohraničena betonovými sloupky, které mají zabraňovat pádu do vodní plochy. Místo měření je ve směrovém oblouku, který je mírně neosvětlený v celé své délce. Obecně lze říct, že celá komunikace obsahuje mnoho směrových oblouků. K silnici jsou napojeny účelové komunikace. Infrastrukturu budou využívat převážně místní. Měření proběhlo v období 31.5.2021 - 6.6.2021 a 3.7.2021 - 9.7.2021.

2.16 Labská Stráň

Měřená oblast je specifická tím, že silnice III. třídy zde končí a nepokračuje nijak dál, pokud se nebudou počítat nezpevněné polní a lesní cesty. Místo měření je nedaleko autobusové zastávky a vjezdu do obce. Komunikace je široká s prostorem v přilehlém okolí, ačkoliv následně je vedena do hustší části obce. Silnice je mírně osvětlená. Měření proběhlo v období 23.8.2021 - 29.8.2021.

Strojové učení

Strojové učení v dopravě zásadně přispívá k bezpečnosti, efektivitě a udržitelnosti dopravy. Dokáže predikovat kongesce, dokáže optimalizovat dopravní cesty a je nedílnou součástí autonomních vozidel. Podílí se na preventivní údržbě vozidla pomocí prediktivní analýzy, dokáže plánovat VHD a řídit dopravu. Základní metody, se kterými se čtenáři v práci setkají, se dělí na dvě elementární skupiny.

S nárůstem objemu generovaných dat je zapotřebí nástrojů a metod, které tato data automaticky a efektivně zpracují. Strojové učení tuto problematiku řeší a nabízí metody, které pomáhají detekovat v datech struktury a vzory, podle kterých pak je schopno tvořit predikce. [1]

3.1 Bez učitele

Je dána množina $X = (x_1, ..., x_n)$ reprezentující body, pro které platí. $x_i \in X$ pro všechna $i \in [n] := 1, ..., n$. Učení bez učitele má za úkol nalézt vzor v datové sadě X. Metody učení bez učitele jsou například: odhad kvantilu, shlukování, redukce dimenzionality atd. [2]

3.2 S učitelem

Pracuje-li se s metodou *učení s učitelem*, je úkolem dosáhnout zobrazení x na y, jenž je definováno tréninkovými daty, která obsahují (x_i, y_i). Výstupní hodnota y_i (nebo též známá jako instance). [2]

3.3 Klastrování a klastrovací metody v dopravě

Klastrovací metody patří mezi metody strojového učení a slouží k lepšímu porozumění datových výstupů. Příklady užití jsou i v dopravě.

Jednou z možností použití klastrovacích metod byla analýza VHD z hlediska efektivity (*Analysis of Public Transportation for Efficiency*). Pomocí klastrovacích metod se výzkumníci z Turecka snažili identifikovat neefektivní trasy a navrhnout zlepšení. V práci se snažili identifikovat podobné trasy a zastávky z hlediska dopravní cesty. Jako koeficienty efektivity byly bráni cestující, na jednotku kilometru. Vytvořila se tabulka vzdáleností, obsahují autobusové trasy v řádkách ve sloupcích. Buňky nabývaly hodnot 0 až 1, kdy 0 byly absolutní shoda a 1 maximální rozdílnost. Pro určení shluků byla použita Hierarchická metoda klastrování. Výsledný graf zobrazoval, které linky autobusu je možné propojit v jednu a zlepšit tak efektivitu dopravy. [3]

Ve studii *A Clustering-Based Framework for Understanding Individuals' Travel Mode Choice Behavior* se

řešitelé pomocí GPS dat snažili identifikovat skupiny lidí podle přepravy. Bylo hledáno chování jednotlivců a jejich dopravních návyků, které se měnily v čase a na které bylo zapotřebí brát ohled. Pomocí Hierarchického klastrování řešitelé nalezli základní dopravní skupiny lidí, podle času a volby dopravy. [4]

Další z aplikací klastrovacích metod v dopravě byla předvedena v práci *Clustering Methods for Determination of Optimal Locations of Container Storage and Distribution Centers*. Díky klastrovacím metodám bylo možné propojit parametry produkce železniční sítě a různé cíle a nalézt nejlepší stanice pro kontejnerovou dopravu. V tomto případě byla vybrána metod k-means, které reflektovalo 9 kritérií, ty byly: přítomnost mezinárodních koridorů, připravenost infrastruktury, kapacity u přilehlé stanice atd. [5]

V článku *Development of an application using a clustering algorithm for definition of collective transportation routes and times* se autoři zabývají vývojem aplikace, která pracuje s DBSCAN klastrovací metodou. Cíle bylo optimalizovat cestovní čas pro velké skupiny lidí tak, že bylo buď přiděleno větší vozidlo nebo více spojů. Díky tomu se povedlo průměrně redukovat 45,80% času tráveného lidmi v dopravě. [6]

Příklady výše ukazují, jak je možné využívat klastrovací analýzu v dopravě. Proto je podle řešitele tento nástroj žádoucí pro bližší zkoumání.

Klastrovací metody

Aktuální svět nabízí mnoho možností, jak sbírat data. Téměř každé elektronické zařízení je schopno sbírat a sdílet data. S rostoucím osazováním vyhodnocovací techniky do různých zařízení, od domácích spotřebičů po vojenské zařízení, se bude tento trend zvětšovat.

V dopravě mají metody bez učitele dvojí význam.

1. Již existující data v dopravě
2. Nová data v dopravě

Je žádoucí, aby se doprava řídila a budovala efektivně. Klatrovací metody nabízí možnosti, jak lépe rozpoznat struktury, tendence a trendy v existujících datových sadách. Ty je potom možné aplikovat do reálného provozu na stávající infrastrukturu a tím jí udělat bezpečnější.

Zároveň s nástupem autonomních vozidel bude nutné data zpracovávat a vyhodnocovat v reálném čase, například v aplikaci C-ITS. Klastrovací metody znovu nabízí možnost, jak taková data v reálném čase zpracovávat a vyhodnocovat pro aktuální potřeby dopravy a tím jí udělat bezpečnější, ekologičtější a efektivnější.

S enormním nárůstem vyhodnocovatelných dat se jejich manuální analýza stává složitější a finančně náročnější. Proto jsou dnes používány nástroje, které pomáhají tuto činnost automatizovat. Jedním z těchto nástrojů jsou i klastrovací nástroje, anglicky ***clustering methods***. Tato metoda je schopná dříve nerozdělená data označit a rozdělit od společných klastrů, které se od sebe liší. I přes dané usnadnění pomocí klastrovacích metod, je zásadní znát kontext a význam vyhodnocovaných dat. [7] Klastrovací analýza vytváří shluky dat, která jsou nějakým stylem podobná nebo rozdílná pro ostatní shluky dat. Každá takové množina obsahuje svůj středový bod, podle kterého se následně určuje, zdali body v okolí budou či nebudou náležet do stejné množiny. Analyzovat data pomocí různých klastrovacích metod má za následek existenci různého počtu nalezených klastrů. Proto je také nutné určit kvalitu jednotlivých možností, které nabízí klastrovací metody. [8]

Pro určování klastrů je zásadní metoda, jak se bude rozhodovat o přiřazení jednotlivých datových bodů k samotnému klastru. Nejčastěji se používá ***matice podobností*** nebo ***vzdálenost mezi body***. Koeficienty podobnosti určují, jak moc se jednotlivé body podobají. Je nutné, aby množina variací $\boldsymbol{p}$ byla pro oba

porovnávané prvky stejná. Samotných koeficientů podobností je několik, je to dáno přístupem, jak zapracovat do výsledku nepodobnost prvků.

V případě určování klastru podle vzdálenosti bodů je používána funkce ***d(x,y)*** pro jednotlivé dvojice bodů v datové sadě $\boldsymbol{E}$, pokud platí:

$$d(x, y) \geq 0; d(x, y) = 0 \text{ if } x = y \tag{4.1}$$

$$d(x, y) = d(y, x) \tag{4.2}$$

$$d(x, y) + d(y, z) \geq d(x, y) \tag{4.3}$$

Pro výpočet vzdáleností se nejčastěji používá Euklidovská metrika. [3] Kde pro vzdálenost mezi body i a j platí:

$$d(x, y) = \left[\sum_{j=1}^{d} (X_{ik} - X_{jk})^2 \right]^{\frac{1}{2}} \tag{4.4}$$

[9]

Stejně tak jako je několik koeficientů podobností, tak existuje i mnoho variací pro počítání vzdáleností, díky čemuž se určení správné metody komplikuje. Euklidovské počítání vzdáleností je nejpoužívanější, ale nemusí být vždy správné. Pokud je struktura dat předem známá, může být výběr výpočtu vzdáleností jednodušší. [10]

Určování vhodné počtu klastrů k

Součástí klastrovací analýzy je i ověření správného určení klastrů danou metodou. Jedním z důvodů je i fakt, že s nárůstem klastrovacích metod roste i počet přístupů jak jednotlivé klastry určit. Tyto postupy se mohou pro každou z metod diametrálně lišit. Pro K-means, kde se počítá vzdálenost od středového bodu, bude postup validace jiný než u metody DBSCAN, která používá poloměr vzdálenosti pro každý bod.

Pro validaci klastru je významné rozeznávat uskupení, která mají potenciál být nenáhodná.

5.1 Koheze a separace

Koheze a separace patří mezi základní nástroje validace klastru. Jako celkovou validitu můžeme označit váženým součtem validit jednotlivých klastrů z množiny K.

$$\text{celková validita} = [\sum_{i=1}^{K}(w_i)\text{validita } (C_i)] \tag{5.1}$$

Kde C_i je i-tý klastr.

Koheze, separace nebo kombinace obojího mohou být použity jako funkce validity, váhy jednotlivých klastrů jsou odvíjeny např. od jejich velikosti, náročnosti nebo můžou mít pouze váhu 1. Čím větší jsou hodnoty koheze, tím je klastr kvalitnější a čím jsou hodnoty separace nižší, tím je klastr kvalitnější. [11]

Koheze a separace může použita dvojím způsobem:

1. Graph-Based View

2. Porototype-Based View

Pro první případ, kdy používáme kohezi a separaci graph-based klastrů, může být definice koheze dána jako vážený součet vazeb uvnitř klastru. Naopak separace může být definovaná jako suma vzdáleností bodů z klastru C_i k bodům klastru C_j. [11]

$$koheze(C_i) = \sum_{x \in C_i, y \in C_i} proximity(x, y) \tag{5.2}$$

[5]

$$separace(C_i, C_j) = \sum_{x \in C_i, y \in C_j} proximity(x, y) \tag{5.3}$$

[5]

Pro typy klastrů prototype-based je možné definovat kohezi jako vážený součet vzdáleností od středu klastru. Separace je zase vzdálenost center c_i kalstru C_i od centra c_j klastru C_j. [11]

$$koheze(C_i) = \sum_{x \in C_i} proximity(x, c_i) \tag{5.4}$$

$$separace(C_i, C_j) = proximity(c_i, c_j) \tag{5.5}$$

5.2 Silhouette koeficient

Metoda koheze a separace je využívána v metodě ***The Silhouette Coefficient***, hodnoty koeficientu se určuje následovně:

1. Spočítají se vzdálenosti pro každý bod i ke všem ostatním bodům z klastru a vytvoří se z nich průměrná vzdálenost a_i.

2. Pro každý i se vypočítá vzdálenost od všech ostatních bodů v ostatních klastrech (k nimž nenáleží), a nalezne se nejkratší průměrnou vzdálenost b_i ke všem ostatním klastrům.

3. Koeficient pro každý bod i se spočítá jako $s_i = (b_i - a_i)/\max(a_i, b_i)$.

Hodnoty koeficientu mohou nabývat -1 až 1, je to dáno tím, že a_i může být větší než b_i. Tj. pokud průměrná vzdálenost v klastru je větší než minimální průměrná vzdálenost v druhém klastru. Klastr je kvalitnější, čím více se a_i přibližuje k 0, tím pádem se s_i více přibližuje k 1.[11]

5.3 Davies-Bouldin Index (DBI)

Jedna z další možností, jak určit, validovat klastry, je ***The Davies-Bouldin Index***, který je definován takto:

Mějme prvky $(x_1, x_2...x_n) \in E_p$ klastru C kde E_p je p-dimenzionální euklidovský prostor.[12]

$$S(x_1, x_2...x_n) \geq 0 \tag{5.6}$$

$$S(x_1, x_2...x_n) = 0 \; if \; x_i = x_j \; \forall x_i, x_j \in C \tag{5.7}$$

Výstupem je nástroj na separování klastrů, $R(S_i, S_j, M_{i_j})$ díky čemuž je možné určit průměrnou podobnost každého klastru s každým klastrem.[12] Úroveň podobnosti lze měřit pokud platí:

$$R(S_i, S_j, M_{i_j}) \geq 0 \tag{5.8}$$

$$R(S_i, S_j, M_{i_j}) = R(S_j, S_i, M_{j_i}) \tag{5.9}$$

$$R(S_i, S_j, M_{i_j}) = 0 \; if \; S_i = S_j = 0 \tag{5.10}$$

$$if S_j = S_k \; and \; M_{i_j} < M_{i_k} \; then \; R(S_i, S_j, M_{i_j}) > R(S_i, S_k, M_{i_k}) \tag{5.11}$$

$$if M_{i_j} = M_{i_k} \; and \; S_j > S_k \; then \; R(S_i, S_j, M_{i_j}) > R(S_i, S_k, M_{i_k}) \tag{5.12}$$

M_{i_j} je vzdálenost mezi vektory, které reprezentují klastr i a j, a S_i a S_j jsou míry rozptyly klastru i a i.[12]

Definice implikuje základní limitace R, ty jsou:

1. Funkce podobnosti R je kladná nebo rovna 0.

2. Je symetrická.

3. Podobnost mezi klastry je 0, pokud mají míru rozptylu 0.

4. Pokud vzdálenost mezi shluky se zvětší a míra rozptylu je konstantní, podobnost shluků klesá.

5. Při konstantní vzdálenosti a zvyšujícím se míry rozptylu, se podobnost zvyšuje.

R_{ij} definováno jako:

$$R_{ij} \equiv \frac{S_i + S_j}{M_{ij}} \tag{5.13}$$

$\bar{R}$ je definováno jako:

$$\overline{R_{ij}} \equiv \frac{1}{N} \sum_{i=1}^{N} R_i \tag{5.14}$$

kde platí, že $R_i \equiv \max_{i \neq j} R_{ij}$.

Metoda určení klastru neovlivňuje použití metody na výpočet míry podobnosti $\bar{R}$.[12]

5.4 Calinski-Harabasz index

Mějme n bodů ve v-dimenzionálním Euklidovském prostoru, $P_1, P_2, ..., P_n$. Pokud původní matice $v \times x$ bude označena jako X, kde řádky jsou pozorované proměnné a sloupce dané jedince, pak můžeme říct, že: $X = (x_1, x_2, ..., x_n)$, kde v souřadnic obsahují sloupcový vektor x_i bodů P_i.[13]
Jsou-li souřadnice uváděny na pravoúhlé osy v euklidovském prostoru, pak vzdálenost d_{ij} mezi P_i a P_j bude funkce vypadat:

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j), \quad i, j = 1, 2, ...n \tag{5.15}$$

Podobný předpis bude platit i pro vzdálenost mezi body a jejich centrem. Míra rozptylu n bodů je určována součtem čtvercových vzdáleností bodů od jejich centra (viz Gower, 1967). Daný součet je roven Stopě matici

R, nicméně může být i nalezeno v páru vzdáleností d_{ij} při použití vzorce Stopy:

$$R = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^2 \tag{5.16}$$

Předpis neztrácí platnost i v případě, kdy jsou souřadnice bodů převedeny na šikmé Euklidovské osy s danou vzdálenostní funkcí. Včetně bodů, které reprezentují vzorky namísto jedinců, jejich vzdálenosti průměrů definovány Mahalanobisovou generalizovanou vzdáleností D^2.[13]

Případě zkoumání rozdělení n bodů do skupin, jenž obsahují n_1, n_2, n_k bodů ($n_1 + n_2 + n_k = n$), pak WGSS je vypočítá pomocí pravé strany rovnice 20, výsledky pro každý klastr jsou vytvořeny zvlášť a následně sečteny. Pokud by matice R byla rozdělena na části rozptylu uvnitř klastru a na rozptyly mezi klastry $R = BSS + WSS$, bylo by možné dosáhnout stejného výsledku analýzou variance matice R v euklidovském prostoru, a poté výpočtem stopy matice WSS.(viz.. Friedman and Rubin, 1967,p.1163).[13]
Díky tomu je možné, že:

$$WGGS = Stopa\ W = Stopa\ R_1 + Stopa\ R_2 + ...Stopa\ R_k \tag{5.17}$$

kde,

$$\text{Stopa } R_g = {n_g}^{-1}(d_{12(g)}^2 + d_{13(g)}^2 + \cdots + d_{(n_{(g)}-1)n_g(g)}^2) \tag{5.18}$$

Metoda lze použít i v případech, kdy body P_i a P_j jenž mají d_{ij} v klastru g, kde ($g = 1, 2, ..., k$), nejsou ve standardním euklidovském prostoru a disperzní matice $R, BSSaWSS(= R_1 + ... + R_K)$ nemusí mít dopad. Pak se použije běžné značení $WGSS$ pro Stopu matice WSS a pro $BGSS$ se použije Stopa BSS a TSS (total sum of squares) pro R. Při dodržení podmínky minimálního rozptylu se rozhodujeme o tom, jak rozdělit nejkratší vzdálenost do k klastrů, pro které je WGSS minimální. Pokud není znám počet klastrů k, je vytvořena iterace, kde $k = 1$, $k = 2$ až $k = n$, nalezneme se nejlepší ***rozdělení součtu čtverců*** dendritu, kde se vypočítá jak nejmenší WGSS, tak maximální BGSS a kritérium poměru rozptylu (VRC). [13]

$$VRC = \frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}} \tag{5.19}$$

Tuto hodnotu je možné brát jako přibližný ukazatel počtu klastrů. Dané kritérium je analogické k F-statistice, v jednorozměrné analýze. I když absentuje pravděpodobnostní teorie, jenž by odůvodňovala užití VRC (23), má dané kritérium pozitivní matematické vlastnosti.[13] V případě, že $\overline{d^2}$ bude označeno jako obecný průměr ze všechny $\frac{n(n-1)}{2}$ druhých mocnin vzdáleností d_{ij}^2, a $\overline{d_g^2}$ jako průměr ze všech vzdáleností $\frac{n_g(n_g-1)}{2}$ čtvercových mocnin vzdáleností v klastru g-té skupiny (g = 1, 2, ..., k) pak z (20):

$$TSS = \frac{1}{2}(n-1)d^2 \tag{5.20}$$

$$WGSS = \frac{1}{2}((n_1 - 1)\overline{d}_1^2 + (n_2 - 1)\overline{d}_2^2 + \cdots + (n_k - 1)\overline{d}_k^2) \tag{5.21}$$

a

$$BGSS = \frac{1}{2}((k - 1)\overline{d}^2 + (n - k)A_k), \tag{5.22}$$

kde

$$A_k = \frac{1}{n-k}\left((n_1 - 1)(\overline{d}^2 - \overline{d}_1^2) + (n_2 - 1)(\overline{d}^2 - \overline{d}_2^2) + \cdots + (n_k - 1)(\overline{d}^2 - \overline{d}_k^2)\right), \tag{5.23}$$

[13] je váženým průměrem mezi čtvercovými vzdálenostmi klastrů a průměrem čtvercových vzdáleností bodů v klastru, pak může být napsáno:

$$VRC = \frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}} = \left(\frac{\overline{d}^2 + \frac{n-k}{k-1}A_k}{\overline{d}^2 - A_k}\right). \tag{5.24}$$

[13] V případech kdy se všechny vzdálenosti všech páru rovnají, pak hodnota A_k nabývá 0 a VRC 1. Kritérium $WGSS$ minimalizuje A_k pro dané k. Funkce A_k může být také použita k porovnání různých rozdělení získaných s různým počtem klastrů, rozdíl $A_k - A_{k-1}$ naznačí průměrný zisk kompaktnosti v klastru, který vychází ze změny $k - 1$ na k klastrů. Chování A_k jakožto funkce k může být citlivé na přítomnost klastrů. [13]

$$\frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}} = \left(\frac{1 + \frac{n-k}{k-1}a_k}{1 - a_k}\right). \tag{5.25}$$

kde

$$a_k = \frac{A_k}{\overline{d}^2}. \tag{5.26}$$

[13] Protože $WGSS$ je minimální, je a_k mezi 0 a 1, když je $a_k = 0$, jsou vzdálenosti mezi všemi dvojicemi bodů stejné. Pokud je $a_k = 1$ je klastr ideální, bez variance uvnitř klastru. Když jsou body v prostoru rozprostřeny rovnoměrně, bude se hodnota a_k blížit k 1 s rostoucím počtem k. S rostoucí k a konstantním a_k bude mít hodnota VRC tendenci klesat, zároveň, pokud bude a_k růst, hodnota VRC i s rostoucím k může neklesat.

Pokud jsou body seskupeny do k_O přirozených klastrů, kde vnitřní variance je malá, pak změna z $k_O - 1$ na k_O způsobí významný nárůst a_k a tím růst VRC.[7]

Určování koeficientu VRC může být užitečné při hledání optimálního počtu klastrů a to tak, že optimální počet klastrů k, je taková, kdy hodnota VRC dosahuje globálního, či lokálního maxima. Pokud lokálních maxim existuje více, je nejlepší volba vybrat minimální k. Prosto stačí dosáhnout prvního lokálního maxima pro získání hodnoty k. Pokud jsou hodnoty VRC při postupných iteracích k kontinuálně rostoucí, neexistuje takové optimální k, které by nebylo ničím jiným, než vytvořením ze všech bodů jednotlivé klastry.[7]

5.5 K-means

Mějme datovou sadu obsahující $x_1, x_2, ..., x_n$ skládající se z N pozorování náhodné D-rozměrné euklidovské proměnné x. Rozdělme datovou sadu do K shluků, kde K je prozatím pevně daná hodnota. Přirozeně se

dá očekávat postup, že se shluk je ***skupina bodů, jejichž vzdálenost mezi body ve shluku je menší, než k bodů mimo shluk***. Mějme tedy D-rozměrné vektory μ_k kde $k = 1, 2, ..., K$, přičemž μ_k je nazýván prototypem (průměrným bodem) klastru k. Cílem je nalézt minimální hodnoty součtů čtvercových vzdáleností všech ostatních bodů k nejbližšímu vektoru μ_k.[14]

Zavede-li se formální notace pro přiřazení jednotlivých datových bodů ke shlukům. Pro každý bod x_n se zavede odpovídající sadu binárních indikátorů s odpovídajícím indikátorovými proměnnými $r_{n_k} \in \{0, 1\}$, kde $k = 1, 2, ..., K$, jenž popisují příslušnost bodu x_n k určitému shluku K.[14] Náleží-li datový bod x_n konkrétnímu shluku k, pak $r_{n_k} = 1$ a $r_{n_j} = 0$ pro $j \neq k$, tento způsob zápisu se nazývá ***1-of-K***. Následně je možné definovat funkci, které je taky známá pod názvem ***distortion measure***:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} |x_n - \mu_k|^2 \tag{5.27}$$

[14] Předpis popisuje součet čtverců vzdáleností datového bodu x_n k vektoru μ_k. K dosažení nejmenší hodnoty J je zapotřebí nalézt ideální hodnoty r_{nk} a μ_k. Pomocí iterativního postupu lze takového výsledku dosáhnout. Iterace bude obsahovat dva kroky, které optimalizují hodnoty r_{nk} a μ_k.[14]

V kroku 0 se přiřadí počáteční hodnota pro μ_k. Následně v 1. kroku se minimalizuje J s ohledem na r_{nk}, přičemž hodnoty μ_k zůstávají stanovené. V 2. kroku se znovu minimalizuje J, ale tentokrát s ohledem na μ_k, přičemž hodnoty r_{nk} zůstávají pevně stanovené. Iterace se opakuje, dokud není dosaženo konvergence.[14]

Vzorec (23) ukazuje, že J je lineární funkcí r_{nk}, a tak může být optimalizace snadno provedena a dává uzavřené řešení. Protože podmínky zahrnující různá n jsou nezávislé, může být optimalizace provedena pro každé n tak, že $r_{nk} = 1$ pro takové k, které dává minimální hodnotu. To znamená, že každý n-tý datový bod je přiřazen ke shluku k nejbližšímu středu.[14] To lze zapsat jako:

$$r_{nk} = \begin{cases} 1 & \text{pokud } k = \arg\min_j |x_n - \mu_j|^2 \\ .0 & \text{jinak} \end{cases} \tag{5.28}$$

[14]

Dalším krokem je optimalizace μ_k se pevně daným r_{nk}. Pro μ_k je funkce J kvadratická, tudíž může být minimalizována tak, že se pomocí derivace podle μ_k nalezne 0, což je:

$$2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \tag{5.29}$$

přičemž výsledek je:

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{5.30}$$

[14] Jmenovatel ve výrazu (34) je roven počtu bodů náležících ke shluku k, díky tomu je výsledek snadno

interpretovatelný. μ_k se nastaví rovno průměru všech datových bodů x_n náležících shluku k. Proto se dané metodě říká ***K-means***. Opakované přepočítávání středů shluků a přiřazování datových bodů v obou fázích se provádí dokud nedojde k další změně přiřazení, nebo pokud není dosažen maximální počet opakování.[14]

5.6 Fuzzy c-means

Fuzzy c-means je jednou z dalších možností, jak rozdělit datovou sadu X na různé shluky c, jenž jsou neprázdné a disjunktní. Shluky z datasetu X jsou v tomto případě nazývány *tvrdými c-rozděleními* X (hard c-partition). Důležitou stránkou ve Fuzzy c-means přístupu je odlišnost chápání příslušnosti bodu ke shluku. Podle Zadaha (1965) je reprezentace podobnosti bodu s každým shlukem pomocí funkce členství, jenž nabývá hodnot od nuly do jedné. Bod má členství ve všech shlucích, je-li podobnost vysoká, hodnota se přibližuje jedné, obráceně, čím menší podobnost tím se více hodnota přibližuje nule.[15]

Fuzzy c-rozdělení X, definuje příslušnost pomocí funkce členství každého bodu ke každému shluku od nuly do jedné.[15] Nechť $Y = Y_1, Y_2, ..., Y_n$ je množina N pozorování v R_n, Y_k je vektor vlastnosti k. Y_{kj} je j-tý vlastnost Y_k. Je-li C celé číslo $2cn$, pak konvenční c-rozdělení Y je c-násobek $(Y_1, Y_2, ..., Y_c)$ podmnožiny Y, který splňuje podmínky:

$$Y_i \neq \phi^{\mathrm{T}} \qquad 1 \leq i \leq c \tag{5.31}$$

$$Y_i \cap Y_j = \phi \qquad i \neq j \tag{5.32}$$

$$\bigcup_{i=1}^{c} Y_i = Y \tag{5.33}$$

[15]

V předcházejících rovnicích je symbol ϕ jako symbol pro prázdnou množinu a $\cap, \cup$ značí průnik a sjednocení. Dále množiny Y_1 budou nazývány *shluky v* Y.[15]

Samotný postup *Fuzzy c-means* jde popsat změnou formulace podmínek (5.31,5.32,5.33) v maticových termínech.[15] Nechť U je matice obsahující reálná čísla o rozměrech $c \times N$, $U = [u_{ij}]$. Maticové dělení Y_i rovnic (5.31,5.32,5.33) je reprezentováno U:

$$u_i(y_k) = u_{ik} = \begin{Bmatrix} 1, & y_k \in Y_i \\ 0, & \text{jinak} \end{Bmatrix} \tag{5.34}$$

$$\sum_{i=1}^{N} u_{ik} > 0 \qquad \forall i \tag{5.35}$$

$$\sum_{i=1}^{N} u_{ik} = 1 \qquad \forall k \tag{5.36}$$

[15]

u_i je funkcí v rovnici; u_i: $u_i : Y \to 0, 1$. V tradičních modelech je u_i reprezentováno jako charakteristická funkce u_i. To znamená, že u_i a Y_i se vzájemně definují, což umožňuje označit u_i jako i-tý pevný dílčí soubor rozdělení. Přestože je tento způsob označení nezvyklý, je klíčový pro pochopení pojmu "fuzzy množina". U je pevné c-rozdělení, protože podmínky u rovnic (5.31,5.32,5.33) a (5.34,5.35,5.36) jsou ekvivalentní. U je označováno jako fuzzy c-rozdělení Y, když prvky U nabývají hodnot v rozmezí $[0, 1]$, avšak stále vyhovují požadavkům daným rovnicemi (5.35) a (5.36). Definující c funkce $u_i : Y \to [0, 1]$ které nabývají hodnot $U_i(Y_k) \in [0, 1]$ ty jsou interpretovány jako příslušnost y_{kS} k fuzzy podmnožinách u_i z Y.[15] Množiny všech pevných a fuzzy c-rozdělení Y označujeme jako:

$$M_c = \{U_{c\times N} | u_{ik} \in [0, 1]\} \tag{5.37}$$

$$M_{fc} = \{U_{c\times N} | u_{ik} \in [0, 1]\} \tag{5.38}$$

Na rozdíl od fuzzy rozdělení, kdy každý bod určitou mírou náleží do každého shluku, existuje pevné rozdělení, kdy je bod jednoznačně přiřazen danému shluku. M_c je součástí M_{fc}, to implikuje, že fuzzy shlukování může být pevné c-rozdělení, naopak c-rozdělení nejde provést při pevných algoritmech.[15]

$$J_m(U, v) = \sum_{k=1}^{N}\sum_{i=1}^{N}(u_{ik}^m ||y_k - v_i||_A^2) \tag{5.39}$$

Další proměnné jsou:

$$Y = Y_1, Y_2, \ldots, Y_N \subset \mathbb{R}^n = \text{datové body}, \tag{5.40}$$

$$c = \text{počet klastrů } Y; \qquad 2 \leq c < n, \tag{5.41}$$

$$m = \text{váhy exponent}; \qquad 1 \leq m < \infty, \tag{5.42}$$

$$U = \text{fuzzy } c\text{-rozdělení } Y; \qquad U \in M_{fc}, \tag{5.43}$$

$$v = (v_1, v_2, \ldots, v_c) = \text{středové vektory}, \tag{5.44}$$

$$v_i = (v_{i1}, v_{i2}, \ldots, v_{in}) = \text{centra klastrů} \quad i, \tag{5.45}$$

$$||\quad||_A = \text{indukovanou } A\text{-normu } \mathbb{R}^n, \tag{5.46}$$

A je pozitivně definitní váhová matice o rozměrech $n \times n$. (5.47)

[15]

Vzdálenost na druhou mezi y_k a v_i, která je vyjádřena rovnicí (5.39), je vypočítána v A-normě.[15] Ke každé chybě na druhou je přiřazena váha, takže chyba je $(u_{ik})^m$ m-tá mocnina příslušnost y_k v shluku i. Jako středy shluků, nebo těžiště dělících množin jsou určeny vektory v_i v rovnici (5.45). V případě, že $m = 1$, pouze pro pevné hodnoty $U \in M_c$ lze prokázat, že Jm dosahuje minima a související v_i jsou přesně geometrická těžiště Y_{iS}.[15] Na základě těchto pozorování lze J_m rozkládat na elementární prvky a zjistit, jakou vlastnost měří Y_k.[15]

V rovnici J_m se nachází dva parametry A a m. Pro každou čtvercovou chybu d_{ik}^2 ovlivňuje relativní váhy exponent m, čím více se m přibližuje hodnotě 1, tím více se zmenšuje J_m a rozdělení s stává tvrdším. Pro m blížící se k nekonečnu je každé optimální U pro J_m blíž k $\frac{1}{c}$. To znamená, že náležitost datových bodů ve shlucích se stává více nejednoznačným. Zvyšování m zmenšuje příslušnost k nejmlhavějšímu stavu. Pro každý výběr m, s předpokladem, že ostatní parametry zůstávají neměnné, je definována jedna varianta algoritmu FCM. Vhodné rozpětí m se nachází v intervalu $1,5 \leq m \leq 3,0$.[15]

Dalším důležitým parametrem je váhová matice A. V prostou R^n kontroluje tvar předpokládaného shluku. Jelikož každá norma v prostoru R^n je odvozena od skalárního součinu pomocí vzorce:

$$\langle x, y \rangle_A = x^\top A_y \tag{5.48}$$

existuje nekonečné množství A-norem, které lze použít v rovnici J_m. Obecně je používáno jen pár těchto norem.[15] FCM lze použít tyto tři normy:

$$A = I \sim \text{-Euclidean Norm}, \tag{5.49}$$

$$A = D_y^{-1} \sim \text{Diagonal Norm} \tag{5.50}$$

$$A = C_y^{-1} \quad \sim \text{Mahalonobis Norm} \tag{5.51}$$

[15]

Když je A rovno jednotkové matice (I), identifikuje Jm shluky ve tvaru hypersféry; pro jakékoliv jiné A mají shluky základní tvar hyperelipsoidu, přičemž osy jsou proporcionální k vlastním hodnotám matice A. Při použití diagonální normy jsou rozměry efektivně škálovány prostřednictvím vlastních hodnot. Optimální řešení pro fuzzy shlukování Y je dáno dvojicemi (U, V), které dosahují lokálního minima hodnoty J_m. Je obtížné dosáhnout stavu aby $m = 1$, ačkoliv jsou potřebné podmínky známé, jenže těžko aplikovatelné, jelikož M_c je diskrétní a velké. Pokud $m > 1$ a y_k se nerovná $\hat{v}_j$ pro každé j a k, pak může být $(\hat{U}, \hat{v})$ lokálně

optimální pro J_m pouze za následujících podmínek:

$$\hat{v}_i = \frac{\sum_{k=1}^{N}(\hat{u}_{ik})^m y_k}{\sum_{k=1}^{N}(\hat{u}_{ik})^m}; \quad l \leq i \leq c; \tag{5.52}$$

$$U_{ik} = \left(\sum_{j=1}^{c} (\frac{\hat{d}_{ik}}{\hat{d}_{jk}})^{2/(m-1)} \right)^{-1}, \quad l \leq k \leq N; \quad l \leq i \tag{5.53}$$

kde $\hat{d}_{ik} = ||y_k - \hat{v}_i||_A$.[15] V rovnici (xy) uvedené podmínky jsou nutné, avšak nikoliv postačující. Díky jednoduché Picardově iteraci lze J_m optimalizovat. Proces zahrnuje opakovaný přechod mezi rovnicemi (5.52) a (5.53) až do bodu, kdy se iterace projevují pouze malými změnami v hodnotách $\hat{U}$ a $\hat{v}$.[15] Přístup je formalizován:

(A1) Stanov konstanty c, m, A, $|\cdot|A$. Vyber jako výchozí bod matici $U^{(0)}$ patřící do množiny M_{fc}. Poté pokračuj krokem k, kde k se rovná 0, 1, ..., až do L_{MAX}.

(A2) Spočítej průměrné hodnoty $\delta^{(k)}$, pro i = 1, 2, ..., c, v souladu s rovnicí (11a).

(A3) Na základě rovnice (5.53) vypočítej novou verzi matice příslušnosti $U^{(k+1)} = [u_{ij}^{(k+1)}]$.

(A4) Proveď porovnání mezi $U^{(k+1)}$ a $U^{(k)}$ využitím vhodné normy pro matice. Pokud je rozdíl $|U^{(k+1)} - U^{(k)}|$ menší než ε, ukonči proces. V opačném případě přiřaď $U^{(k)}$ hodnotu $U^{(k+1)}$ a pokračuj opět od kroku (A2).

[15]

Mezi elementární algoritmické přístupy pro Fuzzy c-means patří rovnice A1-A4. Studium teoretické konvergence posloupnosti $\hat{U}^{(k)}, \mathbf{v}^{(k)}, k = 0, 1, ...$, generované pravidly (A1) až (A4), bylo provedeno (Bezdek, 1981). Numerická konvergence nastává nejčastěji mezi 10 až 25 iteracemi. Není zaručeno, že lokální minima J_m vždy odpovídají dobrým shlukům datové sady Y, protože i při globálním minimu J_m mohou být některé esteticky nevýrazné segmenty. Jako prevence před tímto problémem, se počítá u každého $\hat{U}$ z Fuzzy c-means algoritmu určité množství druhů funkcionálů validity shluku.[15] Často používané koeficienty jsou koeficienty dělení a entropie:

$\hat{U} \in M_{fc}$: Rovnice (5.54) obsahuje logaritmický základ a $a \in (1, \infty)$. Charakteristiky F_c a H_c, které se používají pro ověřování platnosti, zahrnují:

$$F_c(\hat{U}) = sum_{k=1}^{N} \sum_{k=1}^{c} \frac{(\hat{u}_{ik})^2}{N} \tag{5.54}$$

$$H_c(\hat{U}) = -\sum_{k=1}^{N} \sum_{k=1}^{c} \frac{(\hat{u}_{ik} \log_a \hat{u}_{ik})}{N} \tag{5.55}$$

[15] Entropie H_c je citlivější na lokální změny v kvalitě rozdělení než F_c. Postup fuzzy c-means nejdříve vypočítá hodnoty F_c, H_c a $(1 - F_c)$, v případě, že U je součástí M_c a $a = e = 2,71...$, nerovnost $(1 - F) < H$

platí, což ovlivňuje poslední veličinu. Obecně se J_m dokáže více přizpůsobit na mnohem větší množství tvarů dat, než FCM. Nicméně FCM je stále jednou z nejvýznamnějších fuzzy shlukovacích metod.[15]

5.7 DBSCAN

Hlavní myšlenka algoritmu ***DBSCAN*** spočívá v tom, že pro každý bod shluku z datového zdroje D musí existovat nějaké okolí o určitém poloměru, ve kterém se nachází alespoň minimální počet bodů, takže bude dosažena určitá hustota. Tvar okolí je nastaven podle volby vzdálenostní funkce pro body p a q, kterou již nazýváme $dist(p, q)$.[16]

Definice: Eps-okolí bodu

Nechť *Eps*-okolí bodu p značíme jako $N_{Eps}(p)$.

$$N_{Eps}(p) = q \in D \mid \text{dist}(p, q) \leq \text{Eps} \tag{5.56}$$

[16]

První, co se nabízí, je přístup, že pro každý bod ve shluku existuje *Eps*-okolí, kde se nachází alespoň minimální počet bodů $MinPts$. Tento postup je problematický a špatně aplikovatelný z důvodu existence dvou typů bodů v klastru. První typ bodu patří do shluku (jádrové body), druhý typ bodu se nachází na rozhraní shluku (hraniční body). V tu chvíli nastává fakt, že v *Eps*-okolí hraničního bodu se nachází méně bodů než v *Eps*-okolí jádrového bodu. V tu chvíli by minimální hodnota pro počet bodů rovnala velmi nízké hladině, aby byly zahrnuty všechny body pro stejný shluk. Nicméně tato hodnota by nebyla typická pro daný shluk a byla by výrazně ovlivněna přítomností šumu. Proto je nutné, aby ve shluku C existoval bod q pro každý bod p v C tak, že p patří do *Eps*-okolí q a $N_{Eps}(p)$ musí obsahovat aspoň $MinPts$ bodů.[16]

Druhá definice říká, že bod p je přímo dostupný z bodu q z hlediska hustoty s ohledem na Eps a $MinPts$, pokud splňuje následující podmínky:

$$1. \quad p \in N_{Eps}(q) \tag{5.57}$$

$$2. \quad \mid N_{Eps}(q) \mid \geq MinPts \tag{5.58}$$

[16] Pro jádrové body je patrná hustotní symetričnost, obecně symetrická ale není. Pokud jsou zahrnuty jádrové a hraniční body, dochází k asymetrii.

Def.: dosažitelnost podle hustoty

Bod p je dosažitelný z bodu q s ohledem na Eps a $MinPts$ vzhledem k hustotě, pokud existuje sekvence bodů $p1, ..., p_n$, kde $p1$ je rovno q a p_n je rovno p, takže každý bod p_{i+1} je přímo dosažitelný na základě hustoty od bodu p_i.[16]

K ***přímé dosažitelnosti podle hustoty*** se přidává elementární rozšíření, ***dosažitelnost podle hustoty***. Tento vztah je tranzitivní, avšak obecně není symetrický. Pro jádrové body ovšem platí symetrie.[16] ***Dosažitelnost podle hustoty*** nemusí platit pro každou dvojici okrajových bodů ze shluku C, protože podmínka pro jádrové body nemusí pro oba body platit. Avšak, v rámci shluku C musí existovat jádrový bod, od kterého jsou oba hraniční body C ***dosažitelné na základě hustoty***.[16]

Def.: propojenost hustotou

Body p a q jsou ***propojené hustotou*** bez zanedbání Eps a $MinPts$, existuje-li bod o, takže body p a q jsou ***dosažitelné podle hustoty*** z o bez zanedbání Eps a $MinPts$. ***Propojení hustotou*** je symetrický a pro body ***dosažitelností podle hustoty*** i reflexivní vztah.[16]

Definice shluku je taková, že shluk je množina bodů, které jsou ***propojené hustotou***, což je silně podmíněno ***dosažitelností podle hustoty***. Šum je bod, který nenáleží k žádnému shluku.[16]

Def.: shluk

Nechť C je shluk v datasetu D, který má minimální počet bodů $MinPts$ a Eps je poloměr jeho Eps-okolí. Potom shluk C se nazývá silný shluk, pokud platí:

1. Pro každý bod $p \in C$ existuje cesta, která spojuje bod p s jiným bodem $q \in C$, přičemž každé dva sousedící body na této cestě jsou vzdáleny od sebe nejvýše Eps.
2. Shluk C je maximální množina bodů, která splňuje první podmínku. To znamená, že k shluku nelze přidat další bod, který by splňoval podmínky silného shluku.

[16] Silný shluk je tedy tvořen jedním nebo více sousedícími kompaktními oblastmi, které jsou vzdáleny od sebe nejvýše Eps a mají minimální počet bodů $MinPts$. Pokud existuje více takových oblastí, jsou všechny považovány za součást stejného shluku.[16]

Def.: šum

Nechť $C_1, C_2..., C_k$ jsou shluky z datasetu D bez zanedbání Eps a $MinPts_i$, $i = 1, 2, ..., k$. Definice šumu pak zní: Šum je množina bodů v datasetu D, jenž nenáleží do C_i,

$$\text{šum} = p \in D \mid \forall i : p \notin C_i. \tag{5.59}$$

[16] Existuje alespoň $MinPts$ počet datových bodů ve shluku C s ohledem na Eps a $MinPts$. Jelikož C obsahuje minimálně jeden bod p, p musí být hustotně propojen sám se sebou prostřednictvím nějakého bodu o (který může být totožný s p). Tudíž minimálně o musí splňovat podmínku jádrových bodů, pak v Eps-okolí bodu o bude alespoň $MinPts$ bodů.[16]

Lemmata ověřují správnost shlukovacího algoritmu.
Ve dvou iteracích je možné nalézt shluk, který zohledňuje parametry Eps a $MinPts$. První kroku se určí bod z datasetu, který splňuje podmínky jádrového bodu, jakožto počáteční bod. Ve druhém kroku, propojíme všechny body podle definice ***dosažitelností podle hustoty*** z počátečního bodu datasetu, tím se vytvoří shluk, ke kterému náleží i počáteční bod.[16]

Lemma 1:
Nechť bod p v D a $|N_{\text{Eps}}(p)| \geq$ MinPts. Pak množina $O = \{o \mid o \in \mathcal{D}$ a o je dosažitelný podle hustoty z p s ohledem na Eps a M
[16] se stává shlukem s ohledem na Eps a $MinPts$. Není zprvu patrné, že kterýkoliv jádrový bod určuje shluk C s ohledem na Eps a $MinPts$. Shluk C obsahuje všechny body, které jsou ***dosažitelností podle hustoty*** z jakéhokoliv jádrového bodu C, a proto obsahuje právě ty body, které jsou v C ***dosažitelné podle hustoty***. [16]

Lemma 2:
Mějme shluk C s ohledem na Eps a $MinPts$, kde v C je libovolný bod p, jenž $|N_{\text{Eps}}(p)| \geq$ MinPts. Pak C je rovno množině $O = \{o \mid o$je dosažitelný podle hustoty z p s ohledem na Eps a MinPts$\}$.[16]

5.8 Spectrall analysis

Spektrální klastrování patří k atypickým metodám klastrování. Výsledky ze spektrálního klastrování často nabízí více výhod než tradiční přístupy, zavedení a je snadné a lze efektivně řešit pomocí metod lineární algebry.[17]

Nechť $x_1, x_2, ..., x_n$ je množina kde pro každou datovou dvojici x_i a x_j existuje koncepce podobnosti $s_{ij} \geq 0$, přirozený a očekávatelný postup je takový, že body které shlukujeme jsou si podobné tak, že vytváří samostatná uskupení, která se oddělují od odlišných bodů. Graf podobnosti $G = (V, E)$ je vhodným nástroje pro zobrazení dat, je-li jediná dostupná informace o vztahu mezi daty jejich podobnost. Jednotlivé datové body x_i reprezentují vrcholy v_i v grafu G. Mezi body x_i a x_j je vytvořena hrana, které má váhu definovanou podle s_{ij} je-li podobnost datových bodů x_i a x_j větší než nulová nebo větší než určitá mez. Tzn. Hledá se přeformulování problému klastrizace podle grafu podobnosti, který hledá rozdělení grafu tak, že: obsahuje hrany mezi různými shluky s nízkými vahami, a hrany uvnitř shluků mají vysoké váhy.[17]

Je dána množina vrcholů $V = v_1, ..., v_n$ v grafu $G = (E, V)$. Graf G obsahuje nezáporné váhy $w_ij \geq 0$. Matice sousednosti obsahující váhy je definována jako $W = (w_{ij})_{i,j=1,...,n}$. Vrcholy nejsou v_i a v_j spojené hranou právě tehdy když $w_{ij} = 0$. V grafu G platí $w_{ij} = w_{ji}$, protože je neorientovaný.[17] Definice stupně vrcholu $v_i \in V$:

$$d_i = \sum_{j=1}^{n} wij \tag{5.60}$$

Suma je prováděna pouze pro vrcholy, které jsou přímo propojené s v_i, jelikož pro všechny ostatní vrcholy v_j

je váha $w_{ij} = 0$. Následně je definována matice stupňů D, jenž je diagonální a obsahuje stupně $d_1, ..., d_n$. Podmnožina vrcholů Av V má doplněk V/A jako -A, a indikátorový vektor $\mathbb{1}_A = (f_1, \ldots, f_n)' \in R^n$ jenž obsahuje položky $f_i = 1$, je-li $v_i \in A$ pokud ne, je $f_i = 0$.[17]

Zkrácený zápis $i \in A$ pro množinu indexů indexů $i|v_i \in A$ hlavně v součtech typu $\sum_{i \in A} w_{ij}$. Pro ne nutně disjunktní množiny $A, B \subset V$, je definováno:

$$W(A, B) := \sum_{i \in A, j \in B} w_{ij}. \tag{5.61}$$

Existují dva způsoby, jak určovat velikost podmnožiny $A \subset V$:

$|A|$ měří velikost A podle počtu vrcholů, a $vol(A)$ určuje velikost A sumou přes váhy hran připojených k vrcholům v A. Jsou-li dva různé body v A tak, že všechny okolní body taktéž leží v A, pak podmnožina $A \subset V$ je spojena. Pokud není možné nalézt spojení mezi vrcholy v A a A a pokud je podmnožina A spojená, pak je A a nazýváno souvislá komponenta. Pokud $A_i \cap A_j = \emptyset$ a $A_1 \cup \cdots \cup A_k = V$ pak grafový rozklad je tvořen množinami $A_1, ..., A_k$.[17]

Přístupů jak propojovat body $x_1, ..., x_n$ pomocí dvojic vzdálenostní d_{ij} nebo pomocí dvojic podobností s_{ij} a tvořit z nich graf je několik. Modelování místních vazeb mezi datovými body vytváří grafy podobnosti.

V metodě spektrálního klastrování se používá několik přístupů, jak vytvořit graf podobnosti. Typ propojení bodů $x_1, ..., x_n$ a zobrazení grafu podobnosti má vliv na výsledek spektrálního klastrování. Jsou využívány metody epsilon-okolí, k-nejbližších sousedů a úplně propojený graf.[17]

Laplace

Laplacovi matice grafů patří mezi hlavní nástroje pro spektrální klastrování. Neexistuje jediná správná matice graf, každý autor svůj přístup nazývá maticí grafem Laplace. Nechť G je neorientovaný vážený graf s maticí vah W, kde $w_{ij} = w_{ji} \geq 0$. Budou-li používány vektory z vlastní matice, nebude vždy dané, že jsou normalizované. Kupříkladu pokud je dán konstantní vektor 1 a a nějaký násobek $a1$, kde a není rovno nule, budou oba považovány za shodné vlastní vektory. Respektujíce násobost, budou vlastní čísla v každém případě seřazena vzestupně. Bude-li uvedeno *prvních* $\boldsymbol{k}$ *vlastních vektorů*, pak jsou myšlena nejmenší vlastní čísla vlastních vektorů.[17]

Nenormalizovaná Laplaceova matice grafu je zavedena jako $L = D - W$. Popíšeme základní fakta, která jsou pro spektrální klastrování elementární.[17]

Věta 1: Vlastnosti L

Matice L má charakteristiky typu:

1. Pro všechny vektory $f \in \mathbb{R}^n$ platí: $f^\top L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$.
2. L je pozitivně semidefinitní a symetrická.
3. 0 je nejmenší vlastní číslo matice L a konstantní jednotkový vektor je odpovídající vlastní vektor.
4. Počet záporných vlastních čísel matice L je n: $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.[17]

Nenormalizovaný Laplacovský graf není na diagonálních prvcích matice sousednosti W závislý. Jeden stejný nenormalizovaný Laplacovský graf je společný pro každou matici sousednosti, která se shoduje s W mimo diagonálu. Zejména nezmění Laplacovský graf hrany, které vedou v grafu do stejného vrcholu. Údaje jako vlastní čísla a vlastní vektory nenormalizovaného Laplacosvkého grafu lze využít k popisu vlastní grafů.[17]

Věta 2: L spektrum a počet souvislých komponent

Neorientovaný graf G obsahuje kladné, nebo nulové váhy. Potom souvislé komponenty $A1, ..., A_k$ grafu, se rovnají násobnosti k vlastního čísla 0 z matice L. Indikátorové vektory $1_{A_1}, ..., 1_{A_k}$ generují vlastní prostor vlastního čísla 0.[17]

Existují dvě matice, které mají označení normalizované Laplacovské grafy. Obě matice jsou úzce provázané, jejichž definice je:

1. $L_{\text{sym}} := D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$,
2. $L_{\text{rw}} := D^{-1} L = I - D^{-1} W$.

První matice je nazývána L_{sym} jenž je symetrická, druhá matice je L_{rw} a přímo souvisí s náhodnou procházkou (random walks).[17]

Věta 3: Vlastnosti L_{sym} a L_{rw}

Vlastnosti normalizovaných Laplaceových matic jsou následující:

1. Pro každé $f \in \mathbb{R}^n$ platí $f^T L_{\text{sym}} f = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2$.
2. Je-li λ vlastní hodnota L_{rw} jehož vlastní vektor je u, právě tehdy když λ je rovněž vlastní hodnota L_{sym} s vlastním vektorem $w = D^{\frac{1}{2}} u$
3. Je-li λ vlastní hodnota L_{rw} jehož vlastní vektor je u, pak λ a u naplní podmínky zobecněného vlastního problému (the generalized eigenproblem) $Lu = \lambda Du$.
4. 0 je vlastní hodnota L_{rw} s kontaktním jednotkovým vektorem, jako vlastním vektorem. Zároveň 0 je vlastní hodnota L_{sym} jehož vlastní vektor je $w = D^{\frac{1}{2}}$⊮

5. L_{rw} a L_{sym} obsahují n nezáporných reálných vlastních hodnot $0 = \lambda_1 \leq \cdots \leq \lambda_n$ a jsou pozitivně semidefinitní.[17]

Stejně tak jako v případě nenormalizovaného Laplaceova operátoru grafu, násobnost vlastní hodnoty 0 je i u normalizovaného Laplaceova operátoru grafu provázána s množstvím souvislých komponent:

Věta 4 Spektrum L_{rw}, L_{sym} a jejich počet souvislých komponent
G je neorientovaný graf, jenž obsahuje nulové a kladné váhy. Potom počet souvislých komponent $\nVdash_{A_1}, ..., \nVdash_{A_k}$ je roven násobku k vlastní hodnoty 0 pro L_{rw} a L_{sym}. V případě L_{sym} je vlastní prostor pro 0 utvářen vektory $D^{\frac{1}{2}}1_{A_i}$ a v případě L_{rw} je vlastní prostor utvářen indikátorovými vektory $\nVdash_{A_i}$.[17]

Algoritmus

Nechť $x_1, ..., x_n$ je soubor datových bodů. Pomocí některé funkce podobnosti, se měří dvojice podobnosti $s_{ij} = s(x_i, x_j)$. Funkce je symetrická, kladná nebo nulová a $S = (s_ij)$ kde $(i, j = 1, ..., n)$ je její matice příslušnosti. Podle toho, jaký ze dvou normalizovaných Laplaceanových grafů je použit, takové bude spektrální klastrování.[17]

Unnormalized spectral clustering

Prvotní vstup je matice $S \in \mathbb{R}^{n \times n}$ a počet vytvářených shluků k. Vytvoří se graf podobnosti a W jako matice sousednosti s váhami. Následně se spočítá Laplaceův operátor L a prvních k vlastních vektorů $u_1, ..., u_k$ z L. Dále je dána $U \in \mathbb{R}^{n \times k}$ matice, jenž obsahuje $u_1, ..., u_k$ ve sloupcích. Pro $i = 1, ..., n$ kde $y_i \in \mathbb{R}^k$ je vektor, který odpovídá i-tému řádku matice U. Body $y_i \in \mathbb{R}^k$ $i = 1, ..., n$ se shlukují podle k-means metody do jednotlivých shluků.[17]

Normalized spectral clustering according to Shi and Malik

Prvotní vstup je matice $S \in \mathbb{R}^{n \times n}$ a počet vytvářených shluků k. Vytvoří se graf podobnosti a W jako matice sousednosti s váhami. Následně se spočítá Laplaceův nenormalizovaný operátor L. Z obecného vlastního problému $Lu = \lambda Du$ se vypočte k obecných vlastní vektorů $u_1, ..., u_k$. Dále je dána $U \in \mathbb{R}^{n \times k}$ matice, jenž obsahuje $u_1, ..., u_k$ ve sloupcích.Pro $i = 1, ..., n$ kde $y_i \in \mathbb{R}^k$ je vektor, který odpovídá i-tému řádku matice U. Body $y_i \in \mathbb{R}^k$ $i = 1, ..., n$ se shlukují podle k-means metody do jednotlivých shluků.[17]

Normalized spectral clustering according to Ng et al.

Prvotní vstup je matice $S \in \mathbb{R}^{n \times n}$ a počet vytvářených shluků k. Vytvoří se graf podobnosti a W jako matice sousednosti s váhami. Následně se spočítá Laplaceův normalizovaný operátor L_{sym}. Z L_{sym} se vypočte k obecných vlastní vektorů $u_1, ..., u_k$. Dále je dána $U \in \mathbb{R}^{n \times k}$ matice, jenž obsahuje $u_1, ..., u_k$ ve sloupcích. Sestavte matici $T \in \mathbb{R}^{n \times k}$ z matice Unormalizací řádků tak, že budou mít normu 1, tedy definujte $t_{ij} = \frac{u_{ij}}{(\sum_k u_{ik}^2)^{\frac{1}{2}}}$. Pro $i = 1, ..., n$ kde $y_i \in \mathbb{R}^k$ je vektor, který odpovídá i-tému řádku matice T. Body $y_i \in \mathbb{R}^k$ $i = 1, ..., n$ se shlukují podle k-means metody do jednotlivých shluků.[17]

Ve všech případech je různě využíván grafický Laplaceovský operátor. V algoritmech se mění reprezentace abstraktních datových bodů x_i do bodů $y_i \in R^k$, při použití Laplacových operátorů je změna užitečná.[17]

5.9 Hierarchical clustering

Na rozdíl od jiných metod (k-means, k-medoids, DBSCAN atd.), nejsou shluky v hierarchické metodě vytvořeny v jednom kroku. V určitém časovém intervalu se vytvoří 1 až n shluků nebo n až 1, z nichž každý obsahuje alespoň jeden bod z datasetu. Hierarchická metoda shlukování se dělá pomocí dvou základních přístupů:

1. Aglomerativní - počáteční shluky jsou všechny jednotlivé body datasetu. Při postupné iteraci se dané shluky zmenšují (tj. snižuje se počet shluků), ale zvětšuje se jejich objem, tedy shluky se rozšiřují o okolní body.

2. Dílčí - naopak počáteční shluk se rozpadá na menší shluky.

[18]

Oba přístupy se snaží najít optimální krok v definovaném smyslu a využívají matici vzdáleností. Při tvorbě nebo rozpadu shluků hierarchické metody není možné shluk již zpětně spojit nebo nerozdělit. Veškeré aglomerativní metody redukují data na jeden shluk obsahující všechny datové body, proto bude pozorovatel určit v jaké fázi iterace ukončí, aby dosáhl "optimálního"počtu shluků.[18]

Oba přístupy je možné zobrazovat graficky, a to pomocí dendogramu, což je dvojrozměrný graf reprezentující sloučení nebo rozklad v každé iteraci analýzy.[18]

5.9.1 Aglomerativní

Jedná se o nejčastěji používané metody v hierarchických metodách. Datové body o počtu n z datasetu jsou nejdříve považovány za samostatné shluky o velikosti jedna. V poslední iteraci jsou naopak všechny n body shlukovány do jednoho shluku. Fundamentální postup těchto aglomerativních přístupů je podobný. Následující ukázka bude obsahovat příklad o jednoduché vazbě (*single linkage*), nicméně existuje také postup, kdy se využívá vazby na centroid (*centroid linkage*). V každé iteraci se datový bod nebo shluk datových bodů sloučí, podle toho, jestli jsou nejblíže (nebo nejpodobnější). Různorodost metod se liší v přístupu, jak je definován nejbližší soused/sousední shluk (nebo nejpodobnější).[18]

Jedna z technik je technika nejbližšího souseda (*the nearest-neighbour technique*). Při posuzování vzdálenosti mezi skupinami se berou v úvahu pouze ty páry, které se skládají z jednoho člena z každé zúčastněné skupiny.[18]

Matice vzdálenosti M_1 je dána:

$$M_1 = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0.0 & 3.0 & 7.0 & 11.0 & 10.0 \\ 2 & 3.0 & 0.0 & 6.0 & 10.0 & 9.0 \\ 3 & 7.0 & 6.0 & 0.0 & 5.0 & 6.0 \\ 4 & 11.0 & 10.0 & 5.0 & 0.0 & 4.0 \\ 5 & 10.0 & 9.0 & 6.0 & 4.0 & 0.0 \end{bmatrix} \tag{5.62}$$

Nenulový prvek, který je zároveň minimální, je 1 a 2, a tyto body se propojí do nového shluku.[18] Vzdálenosti mezi novým shlukem a zbytkem datových bodů jsou dány následovně:

Minimální hodnota vzdálenosti v matici M_2 mezi samostatnými datovými body je mezi 4 a 5.Tyto prvky vytvoří nový shluk, který bude mít nové vzdálenosti.

$d_{1,2,3} = 6.0$ jako $(1,2),(4,5)$ $d_{1,2,4} = \min(d_{1,4}, d_{1,5}, d_{2,4}, d_{2,5}) = d_{2,5} = 9.0$jako $(4,5),3$ $d_{1,2,5} = \min(d_{1,3}, d_{1,4}, d_{2,4}, d_2$

Nově vytvořená matice M_3 pak bude mít následující podobu:

$$M_3 = \begin{bmatrix} & (1,2) & 3 & (4,5) \\ (1,2) & 0.0 & 5.0 & 8.0 \\ 3 & 5.0 & 0.0 & 4.0 \\ (4,5) & 8.0 & 4.0 & 0.0 \end{bmatrix} \tag{5.63}$$

[18]

Nejmenší vzdálenost $d((4,5),3)$ je mezi $(4,5)$ a datovým bodem 3, a ten je přidán do shluku obsahujícího 4, 5 a 3. Poslední vytvořený shluk pak vznikne spojením shluku s body 1 a 2 a shluku 4, 5 a 3.

Existují i další metody pro aglomerativní přístup klastrování, například úplné propojení (*complete linkage*), které je opačné k jednoduché vazbě (*single linkage*) a používá vzdálenost mezi nejvzdálenějšími páry jednotlivých datových bodů. Další metodou je průměrné skupinové propojení (*group average linkage*), také známé jako UPGMA, kde je vzdálenost dvou shluků definována jako průměrnou vzdáleností jednotlivých dvojic datových bodů, které se skládají vždy z jednoho bodu z každého shluku. Z alternativ se využívá metoda nevážené dvojice bodů pomocí centroidů (*the unweighted pair-group method using the centroid approach*), známá jako UPGMC. Zde se namísto matice vzdáleností používá matice dat (*data matrix*), a spojují se shluky s nejvíce podobnými středními vektory.[18]

5.9.2 Divisivní

Divisivní metody mají opačný postup než aglomerativní, počáteční shluky nejsou jednotlivé datové body, ale všechny body jsou v jednom shluku. Tento shluk se postupně rozpadá na menší shluky. Tyto metody jsou výpočetně náročné, pokud se při každém kroku zkoumá všechno možných $2^{k-1} - 1$ rozdělení do dvou podskupin shluku obsahujícího k objektů. Nicméně existují pro data o p binárních proměnných jednoduché a výpočetně efektivní divisní metody. Tyto metody určují přítomnost nebo nepřítomnost p proměnných a dělí podle nich shluky, což znamená, že na každém kroku shluky obsahují prvky s určitými vlastnostmi, které jsou buď všechny přítomné nebo všechny chybějící. Vstupem je tedy binární matice.[18]

5.9.3 Monothetické dělící metody

Proměnná podle monothetických metod je závislá na optimalizaci kritéria homogenity shluků nebo asociaci s ostatními proměnnými. Díky tomu je možné minimalizovat počet potřebných rozdělení. Jeden z kritérií homogenity je informační obsah, označovaný jako C (což zde zastupuje nepořádek či chaos), který je definován prostřednictvím p proměnných a n objektů (podle Lance a Williamse, 1968):

$$C = pn \log n - \sum_{k=1}^{p} [f_k \log f_k - (n - f_k) \log(n - f_k)] \tag{5.64}$$

kde f_k je počet bodů s k-tou vlastností. Má-li se skupina X rozdělit na dvě skupiny A a B, pak se C sníží tak, že $C_X - C_A - C_B$. Optimální skupina shluků by měla obsahovat prvky se stejnými vlastnostmi a hodnota C by měla být nulová. Z tohoto důvodu se na každém kroku shluky rozdělují podle atributu, jehož použití přináší největší pokles hodnoty C.[18]

Další z přístupů ke zkoumání homogenity je analýza asociací (Williams a Lambert, 1959). Ten zkoumá atributy dat a shluků a vybírá ty, které mají největší význam při rozdělování dat.[18] Pro dvojici V_i a V_i nabývající hodnot 0 a 1 s četností sledování:

$$jad \cdot bcj \tag{5.65}$$

$$(ad \cdot bc)^2 \tag{5.66}$$

$$\frac{(ad \cdot bc)^2 \cdot n}{(a+b)(a+c)(b+d)(c+d)} \tag{5.67}$$

$$\sqrt{\frac{(ad \cdot bc)^2 \cdot n}{(a+b)(a+c)(b+d)(c+d)}} \tag{5.68}$$

$$\frac{(ad \cdot bc)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{5.69}$$

[18]

Hledá se na každé úrovni největší asociace atributu pro správné rozdělení. Kritéria 30 a 31 jsou dobrá, pokud jsou periferní součty nulové. Trojice posledních kritérií je propojená s chi-kvadrátovou statistikou,

druhou mocninou a Pearsonovým korelačním koeficientem.[18]

Monothetické metody nabízí poměrně dobré vlastnosti při klasifikaci nových členů a případů, kdy absentují data. Dále je díky monothetickým metodám snadné určit, který faktor nebo vlastnost je klíčová při rozdělování shluků při každém kroku. Nevýhodou ovšem může být, že pokud shluk nebo data mají určitý raritní nebo vzácný atribut, následující rozdělování daných shluků může vést nevhodným směrem.[18]

5.9.4 Polytetické dělící metody

Polytetické dělící metody mohou využívat matici vzdálenosti a zároveň využívat všechny proměnné současně. Problematické může být, že postup MacNaughton-Smitha et al. (1964) nezohledňuje všechna možná rozdělení. Nově vytvářený shluk začíná tak, že se vybere objekt, který je nejdále od ostatních v rámci původního shluku. Poté se vybírají a přesouvají objekty, které jsou blíže k vybranému objektu. Tento krok se opakuje, přičemž se vybírá další shluk k rozdělení na základě největšího průměru. Průměr je zde definován jako největší vzdálenost mezi jakýmikoli dvěma objekty ve skupině.[18]

Mějme matici L_1:

$$L_1 = \begin{bmatrix} 0 & 15 & 9 & 35 & 32 & 45 & 50 & 12 \\ 15 & 0 & 11 & 28 & 30 & 39 & 41 & 17 \\ 9 & 11 & 0 & 24 & 27 & 36 & 41 & 8 \\ 35 & 28 & 24 & 0 & 9 & 18 & 21 & 29 \\ 32 & 30 & 27 & 9 & 0 & 13 & 19 & 31 \\ 45 & 39 & 36 & 18 & 13 & 0 & 6 & 40 \\ 50 & 41 & 41 & 21 & 19 & 6 & 0 & 46 \\ 12 & 17 & 8 & 29 & 31 & 40 & 46 & 0 \end{bmatrix} \tag{5.70}$$

Bod, který má největší průměrnou vzdálenost ke všem ostatním body, je bod 1, a bude proto vybrán jako počáteční bod pro nový shluk. Shluk bude tedy rozdělen na dva shluky: (1) a $(2, 3, 4, 5, 6, 7)$. Poté se porovná průměrná vzdálenost jedinců v hlavní skupině s jedinci ve vedlejší skupině s průměrnou vzdáleností jedinců v hlavní skupině k ostatním jedincům ve stejné skupině. Zjistí se rozdíl mezi oběma průměry. [18]

Jedinec v hlavní skupině	Průměrná vzdálenost k dělící skupině (A)	Průměrná vzdálenost k hlavní skupině (B)	B - A
2	15.0	28.0	13.0
3	9.0	26.3	17.3
4	35.0	18.0	-17.0
5	32.0	19.7	-12.3
6	45.0	23.0	-22.0
7	50.0	25.8	-24.2
8	12.0	25.3	13.3

Nejvyšší rozdíl mezi oběma průměry je mezi $17, 3$ pro bod 3. Dělící skupina je proto obohacena o tento bod, což dává $(1, 3)$ a $(2, 4, 5, 6, 7, 8)$.[18]

Jedinec v hlavní skupině	Průměrná vzdálenost k vedlejší skupině (A)	Průměrná vzdálenost v hlavní skupině (B)	B - A
2	8.5	29.5	21.0
4	25.5	13.2	-12.3
5	25.5	15.0	-10.5
6	34.5	16.0	-18.5
7	39.0	18.7	-20.3

[18]

Nyní se datový bod 2 přesune a vznikne shluk $(1, 2, 3)$ a $(4, 5, 6, 7)$. Poté se opakuje výpočet a získá se průměrná vzdálenost k vedlejší skupině a v hlavní skupině pro každý bod v hlavní skupině. Výsledky jsou uvedeny v následující tabulce:

Jedinec v hlavní skupině	Průměrná vzdálenost k vedlejší skupině (A)	Průměrná vzdálenost v hlavní skupině (B)	B - A
4	24.3	10.0	-14.3
5	25.3	11.7	-13.6
6	34.3	10.0	-24.3
7	38.0	13.0	-25.0

[18]

Tím, že jsou všechny hodnoty již záporné, může se proces rozdělování zaměřit na oba shluky odděleně a vytvářet nové shluky.[18]

5.9.5 Matematické vlastnosti

Hierarchické klastrování má několik matematických vlastností, které mohou být pro danou metodu definující. Jednou z vlastností je *ultrametrická vlastnost*, kterou poprvé představili Hartigton (1967), Jardine et al. (1967) a Johnson (1967). Ultrametrická vlastnost říká, že:

$$h_{ij} \leq \max(h_{ik}, h_{jk}) \quad \text{pro všechny } i, j \text{ a } k, \tag{5.71}$$

kde h_{ij} je definována jako vzdálenost mezi shluky i a j. Jedním z možných popisů je, že pro jakékoliv tři objekty mají dvě vzdálenosti stejnou hodnotu. Tato podmínka nemusí platit pro prvky matice vzdálenosti, ale platí pro délky h_{ij} bodů, které náleží jednomu shluku. Nedodržení ultrametrické vlastnosti může vést k vzniku inverzí nebo zvratů.[18]

Další vlastností hierarchických metod je schopnost deformovat prostor. Například navazující řetězový efekt jednoduché vazby (*single linkage*), který do jednoho shluku napojuje různorodé shluky, může být ukázkou takového zkreslení prostoru. Naopak dilatace prostoru, kdy shluky mají tendenci se při úplném propojení (*complete linkage*) táhnout k sobě, může být příkladem opaku k předchozímu zkreslení.[18]

Metody šetřící prostorem pří průměrném skupinovém propojením(*group average linkage*) mají nerovnost:

$$d_{iuv} \leq d_{i(uv)} \leq D_{iuv} \tag{5.72}$$

Pro objekt i a shluky u a v je d_{iuv} definována jako minimální vzdálenost mezi objektem i a shluky u a v, zatímco D_{iuv} je definována jako maximální vzdálenost mezi objektem i a shluky u a v. Vzdálenost mezi objektem i a sloučením shluků u a v je definována jako $d_{i(uv)}$. To znamená, že vzdálenosti k nově vytvořeným shlukům se nacházejí mezi vzdálenostmi, které existovaly mezi objekty a původními nesloučenými shluky.

Fisher a Van Nesse (1971) zavedli mnoho admissibilních vlastností, které usnadňují výběr vhodné metody pro shlukování. Jedním z příkladů je dobře strukturovaná admissibilita, která je spojena s parametry Lancea a Williamse, jak je definuje Mirkin (1996), v kontextu přijatelnosti.[18]

Vlastnost podle Mirkina: Shluková přijatelnost: existuje takové shlukování, že veškeré vzdálenosti mezi shluky jsou větší než všechny vzdálenosti uvnitř shluku.[18] Shluková přijatelnost a uchovávání prostoru podle Mirkina jsou rovny podmínkám níže pro libovolné x a y takové, že $0 < x < 1$ a $y > 0$:

$$a(x, y) + a(1 - x, y) = 1 \tag{5.73}$$

$$b(x, 1 - x, y) = 0 \tag{5.74}$$

$$|g(y)| \leq a(x, y) \tag{5.75}$$

kde a, b a g jsou parametry v Lance-Williamsově rekurentním vzorci, reprezentované jako funkce velikostí shluků, s $x = n_k/n_+$, $y = n_i/n_+$ a $z = n_j/n_+$, kde n_+ představuje součet n_i, n_j a n_k.

ance-Williamsovy parametry jsou spojovány s vlastností uchovávání prostoru i Ohsumim a Nakamurou (1989) a Chenem a Van Nessem (1996). Další užitečné vlastnosti jsou:

Konvexní admissibility: pokud lze objekty zobrazit v euklidovském prostoru, konvexní obaly částí se nemohou křížit.

Proporcionální admissibility bodů: zdvojení bodů neovlivní hranice částí.

Monotonní admissibility: monotónní konverze prvků v matici podobnosti neovlivní výsledné shlukování.

Konvexní admissibility je definována pro euklidovský prostor a zabraňuje narušení jednoho shluku druhým. Pokud není jasná struktura shluků, je lepší se řezání shluků shlukem vyvarovat.

Proporcionální admissibility bodů je relevantní v situacích, kdy datová sada obsahuje duplicitní pozorování. Tato vlastnost přispívá k robustnosti výsledného shlukování při mírných rozdílech mezi pozorováními.

Monotonní admissibility zaručuje, že monotónní konverze prvků v matici podobnosti neovlivní výsledné shlukování.[18]

5.10 Bayes

Mějme náhodný vektor naměřených dat $y = (y_1, y_2, ..., y_n)$, kde pravděpodobností rozdělení je dáno funkce $f(x|\theta)$, která je podmíněnou hustotou pravděpodobnosti dat x za předpokladu parametrů θ(funkce pravděpodobnosti). Parametr θ je konstanta nebo vektor neznámých konstant. Data y jsou sbírána proto, aby bylo možné parametr θ modelovat a určit. Θ označuje množinu všech možných hodnot, kterých může nabývat θ. Nejčastěji se dané hodnoty nachází v euklidovském prostoru R^p nebo jeho podmnožině. [19]

Bayesovské odhadování Při popisu náhodných parametrů náhodné veličin se používá hustota pravděpodobnosti (hp). Ta je nazývána apriorní, je-li použita jen předběžná znalost problematiky, pokud se pro další zpracování využijí i naměřená data, stává se popis aposteriorním. Prvotní znalost parametrů se nachází v apriorní hustotě pravděpodobnosti $f(\Theta)$, která tyto parametry popisuje. Nechť T je maximální časová hodnota odhadování, pak data y_t se měří v časech $t = 1, 2, 3, \ldots, T$. Aposteriorní hustota pravděpodobnosti je pak přímý důsledek zpřesňování popisů parametrů z apriorní hustoty pravděpodobnosti na základě informacích z naměřených dat. $f(\theta) \to f(\theta|d(1)) \to f(\theta|d(2)) \to ... \to f(\theta|d(T))$ [20] **Bayesův vztah** Nechť ψ_t je regresní vektor dat obsahující y_t a kde $d_1, d_2, \ldots, d_t$ jsou naměřená data, Byesův vztah upřesňuje parametry hustoty pravděpodobnosti podle znalostí z měřených $y_1, y_2, \ldots, y_t$.

$$f(\Theta|d(t)) \propto f(y_t|\psi_t, \Theta) f(\Theta|d(t-1)) \tag{5.76}$$

Iterace je prováděna podle $t = 1, 2, 3, \ldots, T$, který má počátek v apriorní hustotě pravděpodobnosti $f(\Theta|d(0)) = f(\Theta)$. [**?**]

Obecné odvození Bayesova vzorce:

Nechť A,B a C jsou náhodné veličiny, a A,B má podmíněnou hustotu pravděpodobnosti podmíněnou C, pak:

$$f(A,B \mid C) = \begin{cases} f(A \mid B,C)f(B \mid C) & \text{pravá strana,} \\ f(B \mid A,C)f(A \mid C) & \text{levá strana.} \end{cases} \tag{5.77}$$

Pokud se výrazy dají do rovnosti, vznikne:

$$f(A,B \mid C)f(B \mid C) = f(B \mid A,C)f(A \mid C). \tag{5.78}$$

Bayesův vzorec vznikne po vyjádření $f(B \mid A,C)$ nebo ekvivalentně $f(A \mid B,C)$:

$$f(B \mid A,C) = \frac{f(A \mid B,C)f(B \mid C)}{f(A \mid C)} \tag{5.79}$$

Jak již bylo dříve zmíněno, pomocí Bayesova vzorce lze přepočítávat apriorní hustotu pravděpodobnosti $f(B \mid C)$ na aposteriorní hustotu pravděpodobnosti $f(B \mid A,C)$. Aposteriorní hustota pravděpodobnosti čerpá informace z náhodné veličiny A prostřednictvím hustoty pravděpodobnosti $f(A \mid B,C)$ a apriorní hustota popisuje náhodnou veličinu B pouze v závislosti na náhodné veličině C.

$$p(\theta|y) = \frac{p(\theta,y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{5.80}$$

Pro $p(y) = \sum_\theta p(\theta)p(y|\theta)$ součet je definován přes všechny hodnoty θ, je-li*thet* spojité, pak $p(y) = \int p(\theta)p(y|\theta)d\theta$. Pokud $p(y)$ je nezávislé na θ a je konstantní, pak se může předcházející vzorec zapsat:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{5.81}$$

Člen $p(y|\theta)$ má význam funkce θ. Cílem byesovského modelu je dosáhnout výrazu $p(\theta,y)$, a pomocí kalkulací zjistit co nejvíce z podmíněné pravděpodobnosti $p(\theta|y)$.[21]

5.10.1 Naivní Bayes

Pomocí Naivního Bayese se budou validovat výstupy z K-means a zároveň hledat nejlepší možné k.
Nechť X prostor konečných vektorů $x = [x_1, x_2, \ldots, x_n]$, které reprezentují body v datovém prostoru a je předpokládáno, že body jsou nějakým stylem seskupeny. Jinak taky nazývány shluky. Každý takový shluk má vlastní identifikační hodnotu, ukazatel, který nebývá hodnot $c = 1, 2, \ldots, c_n$. Určený vektor x je cílem klasifikace. Ukazovátko c bude popisováno pravděpodobností náhodnou proměnou $f(c|x)$, kde x je hledaný klasifikovaný vektor.Každá skupina má své vlastní modely, značené jako $f(x|c) = fc(x)$. Tyto modely poskytují pravděpodobnostní popis toho, jak vektory x patří do konkrétní třídy c.

Modely

V kontextu úloh týkajících se shlukování a klasifikace pracujeme se dvěma modely, přičemž v tuto chvíli předpokládáme, že jejich parametry jsou nám známé.

Model datových bodů $f(x|c)$

Model pro klasifikaci $f(c|x)$

Tyto dva modely jsou vzájemně propojeny prostřednictvím Bayesova pravidla

$$f(c|x) = \frac{f(x|c)f(c)}{f(x)}. \tag{5.82}$$

Samotný Naivní Bayes pouze rozšiřuje myšlenku v tom, že se předpokládá nezávislost datových položek x. To znamená:

$$f(x \mid c) = \prod_{i=1}^{n} f(x_i; t \mid c) \tag{5.83}$$

Je předpokládáno, že datové body v jednom shluku se liší pouze v šumu a tudíž nezávislé.

Analýza dat

V této části se budou jednotlivá naměřená data zpracovávat pomocí metod s a bez učitele, jak bylo výše popsáno.

6.1 Úvodní hypotéza a předpoklady

Základní poznatek měřící skupiny na pracovišti v Děčíně byl, že při měření rychlosti vozidel na různých místech se vozidla pohybují různě rychle. Přesněji řečeno, vozidla neměla v každé oblasti stejnou průměrnou rychlost. Předpoklad zní, že rychlost vozidla může ovlivňovat i charakter vozovky a prvky, které přiléhají ke komunikaci v bezprostřední blízkosti. Pomocí klastrovacích metod bude tedy zkoumáno, jaké oblasti se vyskytují v různých klastrech a jak spolu mohou souviset. Bude hledán průnik jednotlivých časových intervalů a dopravních parametrů, jehož výsledkem bude finální skupina lokalit, které jsou si v určitých rysech podobné.

6.2 Postup práce

Kapitola shrnuje postup, který je detailně rozepsán na další stranách, kapitolou *Datová struktura* počínaje.

Celý proces vyhodnocení lze rozdělit na několik částí.

1. První pohled na nasbíraná data.
2. Výstupy z metod strojového učení bez učitele.
3. Výstupy metody strojového učení s učitelem.
4. Porovnání výstupů metod strojového učení bez učitele.

První pohled na nasbíraná data Do této kapitoly spadají kapitoly.

1. Datová struktura
2. Příprava dat
3. Testy

Datová struktura

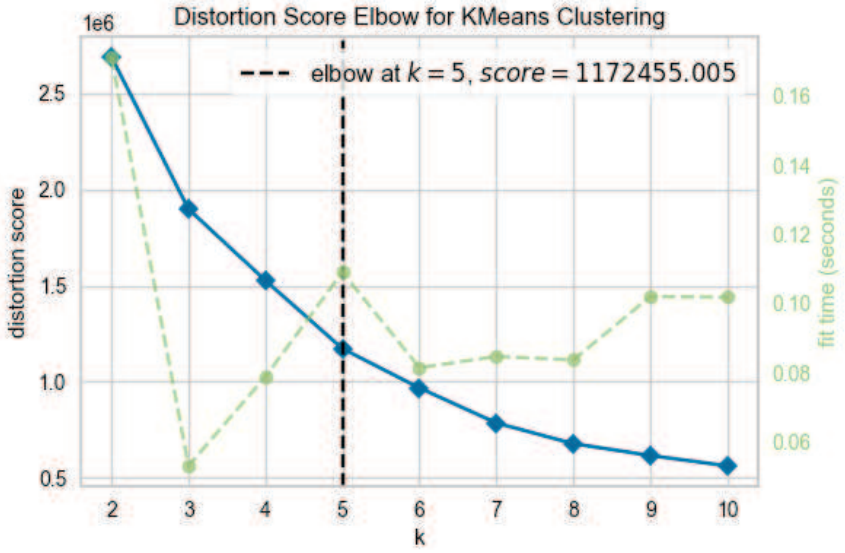
Kapitola stručně popisuje, jaký byl počet dat v datových zdrojích, jakého typu a jakých hodnot nabývaly.

Příprava dat

Při práci se surovými daty bylo zapotřebí data předpřipravit, a tato část je zaměřená na to jakým způsob a jaká data byla upravena.

Testy

Jedná se o první pohledy na datovou strukturu. Vyhodnocují se základní statické veličiny a sumarizují se hodnoty z datových zdrojů (maxima, minima, průměry).



Obrázek 6.1:

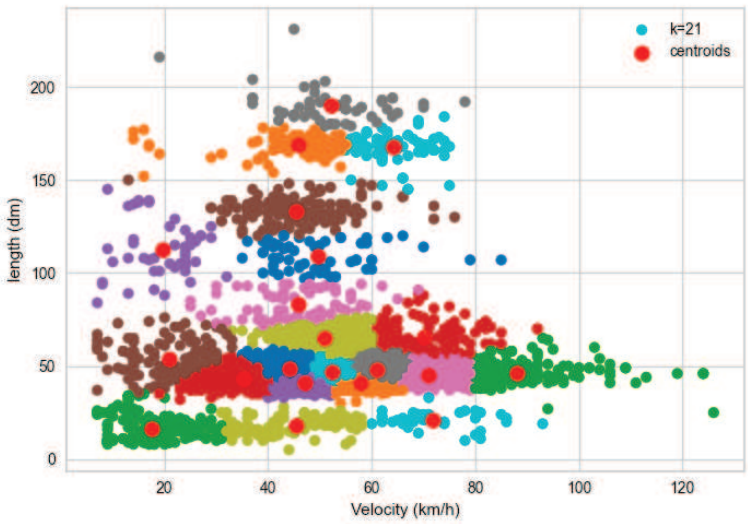
Výstupy z metod strojového učení bez učitele

Jedná se o užití několika klastrovacích metod, kterými jsou:

1. K-means
2. DBSCAN
3. Fuzzy c-meams
4. Spectral clustering
5. Hierachical clustering

Výstupy z těchto metod jsou obecně grafy vyhodnocování počtu klastrů. Tj. hledání ideálního klastru. Graf párů veličin:

1. Length (dm) - Sound (dB)
2. Length (dm) - Velocity (km/h)
3. Sound (dB) - Velocity (km/h)



Obrázek 6.2:

První obrázek 6.1 zobrazuje určení ideálního klastru dle matematického postupu, určí se hodnota k. Tato hodnota se použije v klastrovacím algoritmu, který je na obrázku 6.2. Obrázek zobrazuje dvě veličiny a rozložení dat v 2D prostoru. Každý bod náleží nějakému klastru, který je barevně vyznačen. Ovšem z grafického vyjádření není možné rozpoznat, jaké obce a jak moc se nacházejí v jednotlivých klastrech. Proto je každé hodnotě přiřazeno ID a jméno obce. Následně se spočítá pro každé ID, do jakého klastru nejvíce spadá. Výstup je generován ve formátu .csv, jenž obsahuje čtyři údaje:

1. ID
2. Obec
3. cluster
4. percentage

Nejdůležitější jsou hodnoty Obec, cluster a percentage. Obec slouží k identifikaci vyhodnocovatelem, cluster informuje, do jakých všech klastrů obec spadá a percentage vyjadřuje z kolika procent do definovaného klastru spadá 7.1 např.:

ID	OBEC	CLUSTER	PERCENTAGE
1	BYNOVEC	1	60%
1	BYNOVEC	2	30%
1	BYNOVEC	3	7%
1	BYNOVEC	4	3%

Tabulka 6.1: Vzor výstupu csv souboru klastrů

Sečtou-li se procenta ve sloupci percentage, výsledkem je 100%, což je procentuální vyjádření všech dat pro obec Bynovec. Všech 100% dat je rozloženo mezi klastry 1,2,3,4, které reprezentuje graf výše, kde jsou barevně vyznačeny. Každá obec má své procentuální zastoupení v jednotlivých klastrech, které se bude sdružovat a porovnávat. Příklad výstupu, se kterým se nejčastěji bude pracovat. Tabulka obsahuje tři sloupce: Klastr, Název obce, počet obcí.

Klastr	Název obce	Počet obcí
1,2,3	Bynovec, Staré Křečany	2
0,2,3	Nová Oleška	1

Tabulka 6.2: Vzorový výstup klastrů

Klastr

Klastr obsahuje jednotlivé klastry do kterých spadají dané obce, na příklad Bynovec a Staré Křečany. Na pořadí čísel záleží. Jedná o sekvenci klastrů podle velikosti do kterých obec náleží, viz tabulka **Vzor výstupu csv souboru klastrů**.

Název obce

Zde jsou vyjmenovány všechny obce, které mají stejnou klastrovou sekvenci, jinak také nazýváno klastrovým mixem.

Počet obcí

Tento údaj pouze vyjadřuje počet obcí pro daný klastrový mix.

Jednotlivé výsledky z klastrovacích metod budou validovány Naivním Bayesem, tem vezme část dat, pokusí se nalézt strukturu, a na zbytku se pokusí vytvořit predikci. Následně se do validace vloží i jiné hodnoty pro k, které ale první algoritmus neodhalil, a zjistí se jejich úspěšnost.

Nakonec se vezme nejlepší hodnota validace, podle které se vyberou skupiny lokalit/obcí a navrhne se obecné řešení, jak problematiku rychlosti a překračovaní imisí řešit.

6.3 Datová struktura

Zdrojové soubory byly již ze zmíněných 20 měření. Objem jednotlivých datových souboru se výrazně lišil, pohyboval se v rozmezí od 202 kB pod 3793 kB. Objem souborů odráží počet uskutečněných měření, je poměrné zřejmé, že čím více vozidel bylo naměřeno tím více má soubor dat. Jednotlivé datové soubory byly ve formátu .xlsx. Samotné datové soubory obsahují 12 sloupců: *Device-ID*,*Date*,*Direction*,*Entry velocity*,*Exit velocity*,*Length (dm)*,*Class*, *Class description*, *Sound (dB)*, *distance (cm)*, *Velocity (km/h)*.

Device-ID

Device-ID označuje ID měřícího zařízení, slouží k jasnému propojení dat se senzorem.

Date

Sloupec Date obsahuje datum. Jedná se o týdenní interval. Ve sloupci se nachází hodnoty dne, měsíce a roku. Zároveň obsahuje i časový údaj, hodiny, minuty a vteřiny. Tzn. formát DD:MM:YYYY HH:MM:SS, příklad: 11.09.2021 0:12:22.

Direction

Sloupec Direction obsahuje údaje o směru jízdy záznamu vozidla. Údaj může nabývat hodnot 1 a 2. Hodnota 1 určuje směr k měřícímu zařízení, hodnota 2 od měřícího zařízení.

Entry velocity

Údaj by měl obsahovat rychlost na vstupním senzoru. V tomto případě sloupec obsahuje pouze hodnoty 0, jelikož senzor nebyl umístěn v páru.

Exit velocity

Pro koncovou/výstupní rychlost platí to samé co pro *Entry velocity*. Hodnoty jsou pouze 0, jelikož senzor nebyl umístěn v páru.

Length (dm)

Hodnoty délky vozidla jsou uváděny v jednotkách dm, obsahuje tedy pouze číselné údaje. Sloupec nabývá hodnot v intervalu od 10 dm (1 m) po cca 200 dm (20 m).

Class

Class description	Class
carwt	2
truck	3
bus	5
unc.mv	6
car	7
truckwt	8
semi truck	9
MCyc6	10
LGV	11
bicykle	230
vph	250

Tabulka 6.3: Přiřazení číselných údajů k popisku

Class - Třída, obsahuje informaci o typu vozidla. Jedná se o číselný údaj, který je se pohybuje v intervalu od 1 do 11 a dva speciální číselné údaje 230 a 250.

Class description

Kadý číselný údaj ze sloupce *Class* má vlastní popis. Slouží k jednodušší identifikaci. Jednotlivé popisy jsou:

Sound (dB)

Ve sloupci je zaznamenaná hlasitost projíždějících vozidel. Zvukový záznam byl zaznamenán na škále od 0dB po 95* zjistit maximum dB.

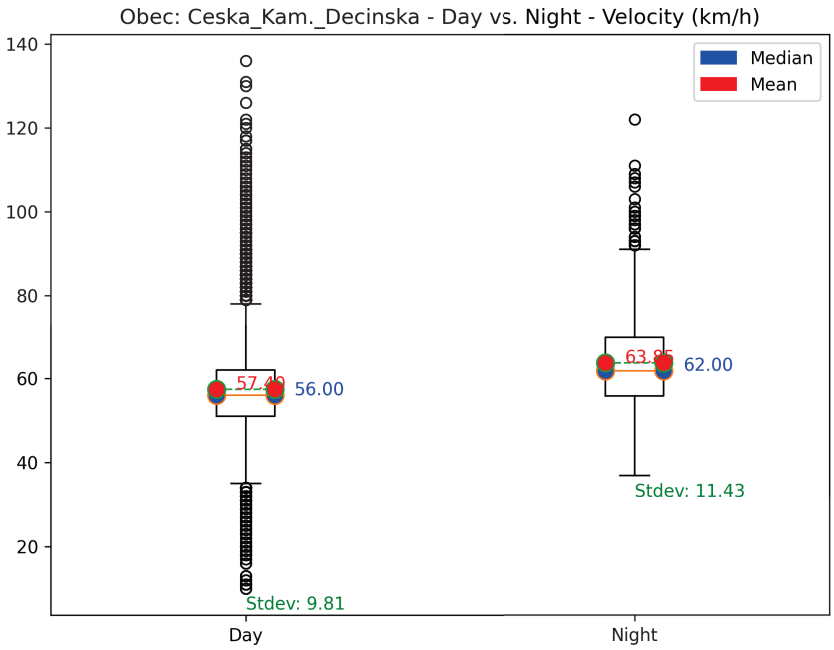
distance (cm)

Distance (cm) obsahuje informaci o vzdálenosti naměřeného vozidla od senzoru. Číselné rozpětí se pohybuje od 50 cm do 615 cm* zjistit maximum.

6.3.1 Příprava dat

Před samostatnou analýzou dat bylo zapotřebí data opravit, filtrovat nebo případně změnit datové typy. Již výše bylo zmíněno jaké dopravní údaje byly měřeny. Některé údaje nabývaly hodnot, které byly chybné. Například, když záznam vozidla obsahoval zvuk *Sound (dB)* roven 0. Je patrné, že pokud se hodnota jakéhokoliv údaje bude rovnat 0, nedává smysl. Proto můžou být všechny nestatické údaje podrobeny této podmínce. Další úprava je změna formátu *Date*, respektive separace údaje o datu a času.

Postup byl takový, nejdříve se obsah sloupce *Date* převedl do formátu `%d.%m.%Y %H:%M:%S`. Následně se jednotlivé údaje jako datum čas a hodina rozdělí do jednotlivých sloupců. Při výběru vhodných dat bylo zapotřebí vybrat *běžné pracovní dny* TP 189, pro správnou kvalitu dat.[26] To jsou dny, úterý, středa a čtvrtek. Byl proto vytvořen nový sloupec `weekday`, který obsahuje informaci o dnu v týdnu podle data. Následně se vyfiltrují pouze určené dny. Další problematikou je časové rozdělení měřeného intervalu. Senzor měřil celých 24 hodin, nicméně očekává se, že řidiči se budou obecně chovat jinak v denních hodinách, a v nočních hodinách. Pokud se bude posuzovat chování řidičů, je nutné tyto dva segmenty oddělit. Tyto časy byly určeny již podle existující definice z dokumentu TP 219.[26] Jednotlivé datové sety byly rozděleny na noční segmenty `df_N` a denní segmenty `df_D`. Jak již bylo zmíněno, datové soubory měly různou velikost, celkem všechny soubory měly více než 500 tisíc řádků. Enormní objem dat nebylo možné na dostupné technice kompletně vyhodnotit, proto byla zapotřebí redukce. Data byla redukována na 2000 řádků pro každý datový set. Pokud datový set obsahoval méně řádků, byl vložen celý. Aby se ale neztratil význam dat, je



Obrázek 6.3: Boxplot - Česká Kamenice Děčínská - Den/Noc - rychlost (km/h)

potřeba jednotlivé řádky vybírat náhodně, to je zajištěno pomocí funkce `df.sample(n=2000, random_state=1)`, a přímo pomocí parametru `random_state=1`.

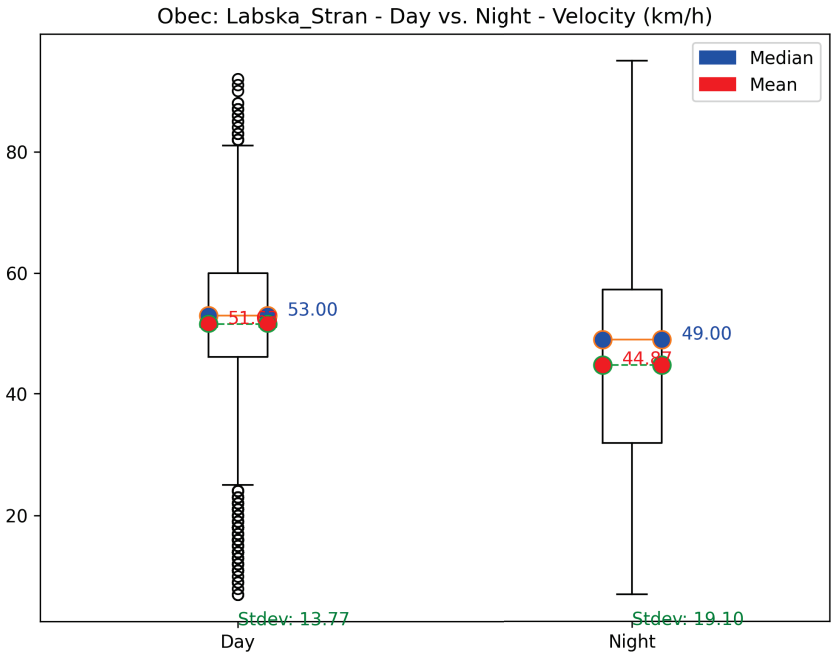
6.3.2 Testy

Pro snazší identifikaci struktury dat byly vytvořeny histogramy (obrázek 6.5 a 6.6 a 6.7), boxploty (obrázek 6.3 a 6.4) a Q-Q ploty, které snadno odhalovaly, zda-li data mají normální rozdělení nebo ne a kde se nachází průměr.(viz příloha histogramy, boxploty a Q-Q ploty) Jak bylo uvedeno, data byla rozdělena na noční a denní část kvůli možnému zkreslení výsledku chování řidičů. Pokud data zobrazíme, můžeme zjistit značné rozdíly. V grafu 1, kde je vidět Česká Kamenice - Děčínská, je patrný rozptyl hodnot v období Den a v období Noc. Během dne vozidla sice dosahují vyšších rychlostí, ale také naopak projíždí i výrazně nižší. Proto průměr a medián vykazuje pro den a noc odchylku až 6 km/h. Některých případech může nastat opačná tendence, jako například v Labské Stráni na obrázku 6.4. Nicméně ze 20 ti měření nastal tento jev pouze 3 krát, a to v případě již zmíněné Labské Stráně v Huntířově Nové Olešné a v Hrobcích.

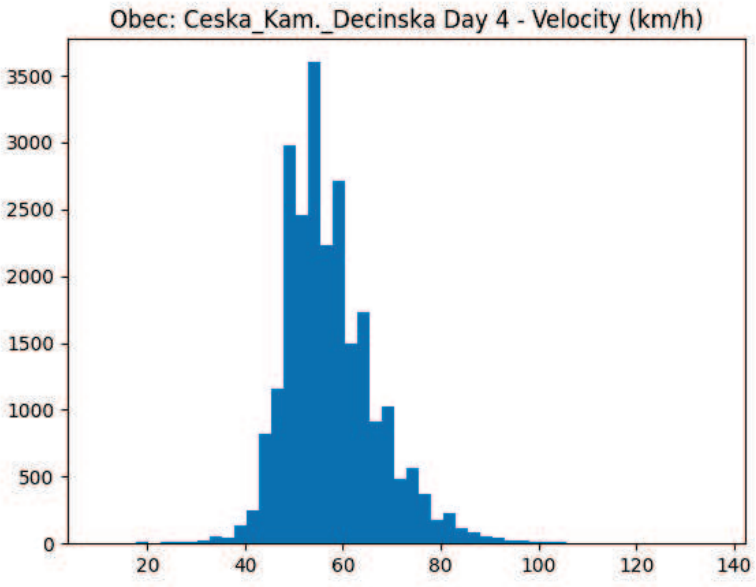
Pokud se sumarizují extrémní hodnoty z veličin *Sound (dB)*, *Length (dm)* a *Velocity (km/h)*. Nejvyšší průměrnou (i mediánovou) rychlost přes den mají Staré Křečany 2, činí 59 km/h a v noci stoupá o cca 1 km/h pro průměr i medián. vozidla zde v průměru dosahují délky 4,9 m přes den a v noci 5,2 m, mediánově pouze 4,5 m. Nejedná se o extrémní hodnoty. Co se týče zvuku, dosahují Staré Křečany hodnot kolem 78-79 dB pro průměr a pro medián cca 79-79,5 dB pro den a noc. (viz grafické přílohy) Při podrobnějším zkoumání histogramu *Length (dm)* jsou patrné tři věci.

1. Nejpočetnější hodnoty se nezachází v intervalu 4,5-6 m. .
2. Nachází se zde i nezanedbatelné množina vozidel kratších než 2,5 m.
3. Vozidel větších než 7,5 m je nejméně, ale také ne zanedbatelně, hodnoty jsou totiž až trojnásobné oproti průměru.

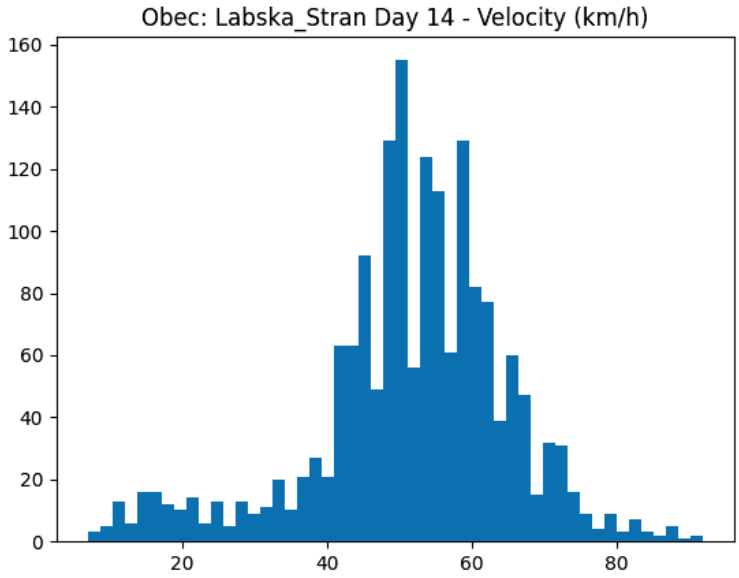
Bod 2. a bod 1. mohou mít velký vliv na rychlost a akustické zatížení. Vozidla kratší než 2,5 m mohou být motocykly, které mohou překračovat maximální povolenou rychlost. Vozidla delší než 7,5 m budou nákladní automobily, které naopak generují větší zvukovou zátěž. Stále zde ovšem dominují osobní automobily s délkou 4,5 m až 5,5 m, které budou hlavními pachateli překračování rychlosti a tím i generátory hluku.



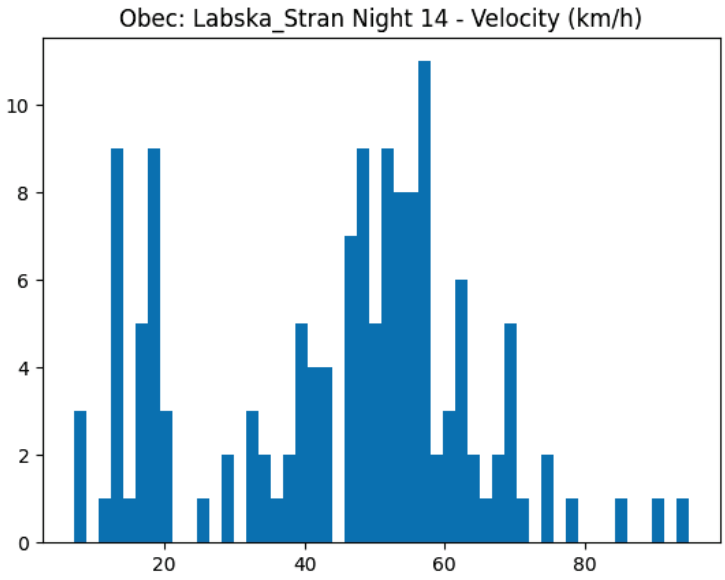
Obrázek 6.4: Boxplot - Labská stráň - Den/Noc - rychlost (km/h)



Obrázek 6.5: Česká Kamenice - Děčísnká: Den - ryhclost (km/h)



Obrázek 6.6: Labská Stráň: Den - ryhclost (km/h)



Obrázek 6.7: Labská Stráň: NOc - ryhclost (km/h)

Naopak nejnižší průměrnou i mediánovou rychlost má Nová Oleška 2, od 42 km/h do 46 km/h. Ovšem pro průměr rychlost v nočních hodinách klesá a pro medián roste. Průměry Sound (dB) a Length (dm) jsou všechny nejmenší hodnoty ze všech. Pouze v případě noci a délky vozidla se průměr zvětšuje z 4,5 m na 5,5 m. V tu chvíli se Nová Oleška pro noční délku vozidla řadí do průměru všech hodnot, nikoliv do extrému. Pro mediánové hodnoty se Nová Oleška přesouvá s délkou vozidla na průměrných 4,5 m a s akustickou emisí na 70 dB, obojí pro den i noc. Při bližší zkoumání histogramů Sound (dB) a Length (dm) je možné povšimnout si určitých detailů, které ovlivňují průměr a medián.

pro délku:

1. Nachází se zde nezanedbatelná množina vozidel kratších než 2,5 m.
2. Vozidel větších než 7,5 m je nejméně, ale také ne zanedbatelně, s vyšší koncentrací při hodnotách 12,5 m.
3. V nočních hodinách výrazně ubyde vozidel s délkou pod 2,5 m a 4 m až 5 m ku nárůstu vozidel delších než 11 m.

pro akustickou zátěž:

1. Během dne jsou patrné nízké zvukové imise 40 dB až 60 dB, ty mohou být způsobeny nízkou denní rychlostí, která klesá až rychlosti 10 k
2. Hodnoty kolem nočního průměru 70 dB výrazně klesají, ale jsou zase kompenzovány výrazným nárůstem kolem 72-75 dB.

Ze sumarizací je patrné, že se rychlosti, akustické zatížení a délky vozidel v různých četnostech nacházejí v naměřených oblastech. Nicméně v tuto chvíli nelze pomocí histogramů obrázky(5-8), nebo boxplotů (obrázky 3-4) přesněji určit, jaké jsou podobnosti či vztahy mezi všemi oblastmi a jejich veličinami.

6.4 Klastrování

Na rozdíl od předchozího přístupu se v klastrovacích metodách vyhodnocují veličiny v páru. Vytvoří se hluky a budou se vyhodnocovat jednotlivé náležitosti měřených oblastí k daným klastrům. Byly vytvořeny páry:

1. *Sound (dB)*, *Length (dm)*
2. *Sound (dB)*,*Velocity (km/h)*
3. *Length (dm)*, *Velocity (km/h)*

6.5 K-means

Před samostatným určováním klastrů se určuje nejlepší počet klastrů, které je možné v daném páru dopravních veličin určit. Po definování správného počtu k klastrů se jednotlivé datové body přiřadí do správné skupiny.

Metoda	**Počet klastrů**	
	Den	**Noc**
Calinski-Harabasz	2	2
Davies-Bouldin	5	6
Elbow	2	2

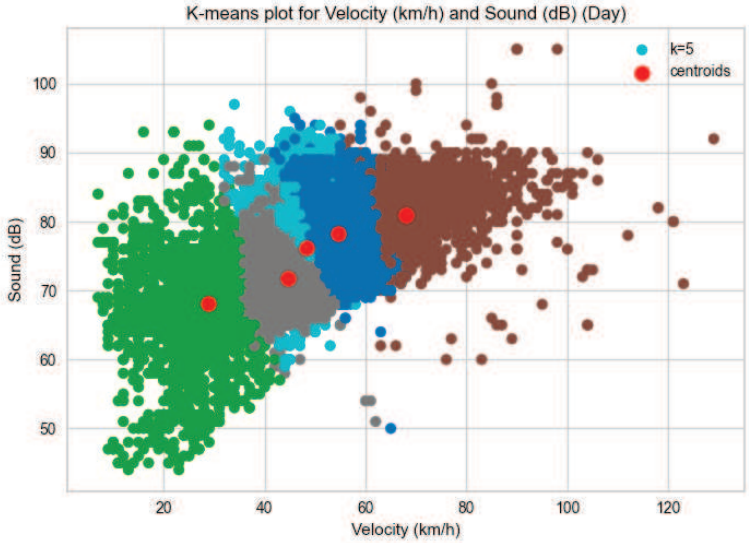
Tabulka 6.4: Počty klastrů pro inicializaci

Ty byly vybrány jako 2,5,6. Hodnota 2 se v práci vyloučí, jelikož je příliš hrubá pro určování skupin lokalit/obcí. Proto se bude v celé práci pracovat přibližně s hodnotou 5,6 v k-means a 5 v ostatních případech (Obrázek 6.8 a 6.9).

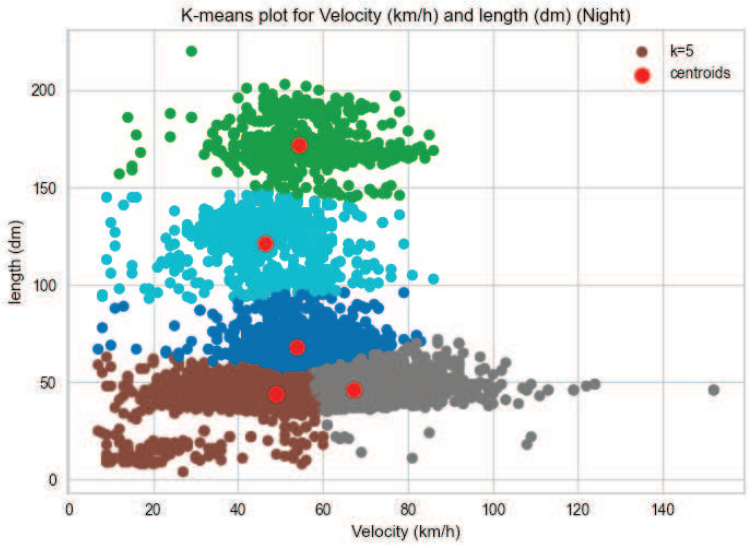
6.5.1 Sound (dB) - Length (dm)

Počáteční inicializace počtu k byla položena v intervalu od 2 do 20. Minimální počet klastrů, při kterém je možné hledat společné vlastnosti oblastí je 2 a maximální je 20, to se rovná počtu měření.

Je patrné, že volba $k = 2$ je algoritmů nejlepší, nicméně dosazovat se budou obě hodnoty.



Obrázek 6.8: K-measn: Velocity (km/h) - Sound (dB) k=5 - denní interval



Obrázek 6.9: K-measn: Velocity (km/h) - Sound (dB) k=5 - noční interval

Barevně rozlišení vymezuje prostor jednotlivých klastrů a červený bod definuje střed klastru. Při prvním zkoumání grafu *Dne*(6.8) je vidět, že vozidla jedoucí do 40 km/h (zelený klastr) vytváří akustické znečištění skoro na celé intervalu *Sound dB*. Naopak vozidla jedoucí nad 65 km/h (hnědý klastr), se pohybují pouze v akustické hladině 70 dB a výše. V obrázku 6.9

je naopak vidět, že vozidla mezi 100 dm a 150 dm jezdí maximálně 85 km/h. Zastoupení obcí v klastrech Samotné grafické vyjádření dat (Obrázek 6.8 a 6.9) slouží pouze pro orientaci, aby bylo možné více specifikovat charakteristiky oblastí, je nutné data číselně interpretovat. Výstupem jsou dvě tabulky (Tabulka 6.5 a 6.6). První obsahuje procentuální zastoupení obce v daném klastru, druhá vyjadřuje v jak je daná obec v klastrech rozprostřená .

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Obec	**Klastr**
Bynov	1,2,4
Bynovec1	1,2,0
Bynovec2	1,2,0
Ceska Kam. Decinska	1,2,4
Ceska Kam. Duk. Hrdinu	1,2,4
Ceska Kamenice Liska	1,2,4
Decin Teplicka	1,2,0
Decin Ustecka	1,2,0
Hrobce	1,2,5
Huntirov	1,2,3
Huntirov Nov. Oles.1	1,5,2
Kasna Lipa2	3,2,5
Krasna Lipa1	3,2,5
Labska Stran	3,2,5
Libouchec	3,2,0
Nebocany	3,2,4
NovaOleska2	3,5,2
Rohatce	3,2,0
Stare Krecany1	3,2,5
Stare Krecany2	3,2,5

Tabulka 6.5: K-menas: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Sound (dB) - Length (dm) neseskupená

Pořadí čísel klastrů (Tabulka 6.5) je dáno jejich výskytem pro danou obec. Na první pohled jsou patrné dvě věci. Nejvíce zastoupené klastry na prvních místech jsou 1 a 3. Druhým nejčastějším klastrem je 2. Díky tomu je možné seskupit obce na pět základních množin. Ty budou tvořeny klastry 1,2 a 3,2. Toto jsou hlavní dvě skupiny. Důvodem, proč hodnoty 1, 2 a 3 jsou nejčastěji zastoupené, je dáno polohou klastrů. Nacházejí se nejblíže průměrné a mediánové hodnotě dat, kde se také nachází podle histogramu. Proto je také nutné sledovat příslušnost i k dalším klastrům.

V tabulce 6.6 je patrné, jak K-means algoritmus rozdělil jednotlivé obce do různých kategorií. Sloupec *Klastr* obsahuje pořadí klastrů, ve kterých se dané obce vyskytují nejvíce (na pořadí klastrů záleží), od největšího zastoupení po nejmenší. Na první pohled je patrné, že se data podle všeho podobají pro Krásnou Lípu 1 a Krásnou Lípu 2, v prvním řádku, a pro Bynovec 1 a Bynovec 2, ve třetím řádku. Podle K-means mají data stejnou strukturu, což odpovídá skutečnosti, jelikož se jedná o stejná místa, liší se jen datum měření.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Název obce	Počet obcí
3,2,5	Kasna Lipa2, Krasna Lipa1, Labska Stran, Stare Krecany1, Stare Krecany2	5
1,2,4	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Ceska Kamenice Liska	4
1,2,0	Bynovec1, Bynovec2, Decin Teplicka, Decin Ustecka	4
3,2,0	Libouchec, Rohatce	2
3,5,2	Nova Oleska2	2
1,2,3	Huntirov	1
1,2,5	Hrobce	1
1,5,2	Huntirov Nov. Oles.1	1
3,2,4	Nebocany	1

Tabulka 6.6: K-menas: Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Sound (dB) - Length (dm)

Obec	**Klastr**
Bynov	0,1,3
Bynovec1	0,2,3
Bynovec2	0,2,3
Ceska Kam. Decinska	0,1,2,3
Ceska Kam. Duk. Hrdinu	0,1,2,3
Ceska Kamenice Liska	0,1,2,3
Decin Teplicka	0,1,2,3
Decin Ustecka	0,1,2,3
Hrobce	0,1,2,3
Huntirov	0,1,2,3
Huntirov Nov. Oles.1	0,2,3
Kasna Lipa2	0,2,3
Krasna Lipa1	0,2,3
Labska Stran	0,1,2,3
Libouchec	0,2,3
Nebocany	0,1,2,3
Nova Oleska2	0,2,3
Rohatce	0,2,3
Stare Krecany1	0,2,3
Stare Krecany2	0,2,3

Tabulka 6.7: K-menas: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Sound (dB) - Length (dm) neseskupená

Případ nočního intervalu ukazuje (Tabulka 6.7), že klastr 0 je pro všechny měření společný, na rozdíl od denní intervalu. V případě druhého klastru už shoda nepanuje a příslušnosti se dělí na hlavní děv možnosti 1 a 2. Pro lepší přehlednost budou data seskupena do stejných klastrových mixů.

Klastr	Název obce	Počet obcí
0,1,2	Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Labska Stran, Nebocany	9
0,2,3	Huntirov Nov. Oles.1, Kasna Lipa2, Krasna Lipa1, Libouchec, Nova Oleska2, Rohatce, Stare Krecany1, Stare Krecany2	8
0,1,3	Bynov, Bynovec1, Bynovec2	3

Tabulka 6.8: K-menas: Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Sound (dB) - Length (dm)

Největší skupina obsahuje 9 obcí (Tabulka 6.8) a čítá většinu zaznamenaného měření. Druhá největší skupina je ale pouze o jedno měření menší. Nejmenší uskupení čítá pouhé 3 měření, Bynov, Bynovec1 a Bynovec2. Je patrné, že všechna párová měření jsou ve stejných klastrových mixech, což může naznačovat homogenitu při generování dat měřenou oblastí.

6.5.2 Sound (dB) - Velocity (km/h)

V případě páru *Sound (dB) - Velocity (km/h)* byl k-means dost odlišný, než v případě *Sound (dB) - Length (dm)*.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Název obce	Počet obcí
0,3,2	Ceska Kam. Decinska, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka	4
3,0,1	Bynov, Bynovec1, Bynovec2, Ceska Kam. Duk. Hrdinu	4
4,2,1	Hrobce, Labska Stran, Nebocany	3
2,4,1	Rohatce, Stare Krecany1, Stare Krecany2	3
0,4,3	Huntirov	1
4,0,1	Krasna Lipa1	1
4,1,0	Libouchec	1
4,1,2	NovaOleska2	1
4,1,3	Kasna Lipa2	1

Tabulka 6.9: K-means: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Sound (dB) - Velocity (km/h)

V tomto případě (Tabulka 6.8) algoritmus oddělil Krásnou Lupu 1 a Krásnou Lípu 2, je nutné poukázat na skutečnost, že se liší pouze v pořadí druhého a třetího klastru, kterému nejvíce náleží. Je patrné, že převládá klastr 4, následovaný 0.
Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
1,0,2	Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka	3
0,1,3	Hrobce, Huntirov	2
0,4,1	Kasna Lipa2, Krasna Lipa1	2
2,1,3	Bynov, Bynovec2	2
0,1,4	Libouchec, Nebocany	2
1,0,3	Stare Krecany1	1
1,3,0	Stare Krecany2	1
2,1,0	Ceska Kam. Duk. Hrdinu	1
0,4,2	Huntirov Nov. Oles.1	1
4,0,1	Labska Stran	1
2,1,4	Bynovec1	1

Tabulka 6.10: K-means: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Sound (dB) - Velocity (km/h)

Tabulka 6.9 je mnohem více fragmentovaná. Je možné, že je tento jev způsoben větším rozptylem dat a tím i různorodější kombinací klastrů. Všechny případy sekundárního měření v obcích (Bynovec 1,2 atd.) se nedostaly do stejných tří klastrů. Je ovšem vidět, že se jedná o malé výchylky. Například Staré Křečany 1 a Staré Křečany 2 se liší pouze v pořadí klastrů 0 a 3. Jinak náleží ke stejným klastrům.

6.5.3 Length (dm) - Velocity (km/h)

Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
0,1,3	Bynov, Bynovec1, Bynovec2, Decin Teplicka, Decin Ustecka, Huntirov, Huntirov Nov. Oles.1, Krasna Lipa1, Libouchec, Nebocany, Nova Oleska2, Rohatce	12
1,0,3	Ceska Kam. Decinska, Ceska Kamenice Liska, Labska Stran, Stare Krecany1, Stare Krecany2, Hrobce, Kasna Lipa2	7
0,3,2	Ceska Kam. Duk. Hrdinu	1

Tabulka 6.11: K-means: Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Length (dm) - Velocity (km/h)

Nejpočetnější skupina je zde tvořena dvanácti měřeními (Tabulka 6.10), druhá největší je tvořena sedmi měřeními. Poslední skupina tvoří pouze jednu obec, Česká Kamenice Dukelských Hrdinů. Zároveň jsou všechny zdvojená měření ve stejných skupinách, to může naznačovat, že výstupy z oblasti se neliší nehledě na datum měření.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* Znovu se projevila vysoká fragmentace klastrového mixu. Sekudárně měřené obce se dokonce rozdělily do různých skupin. Stejně jako v předchozím nočním případě je zde hodně oddělených skupin.

Klastr	Názvy obcí	Počet obcí
2,3,0	Bynov, Bynovec1, Bynovec2, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Libouchec	9
2,3,4	Krasna Lipa1, Labska Stran, Nebocany, Rohatce	4
2,4,3	Kasna Lipa2, Nova Oleska2	2
3,2,1	Ceska Kam. Decinska	1
3,2,0	Ceska Kamenice Liska	1
2,4,0	Huntirov Nov. Oles.1	1
3,2,4	Stare Krecany1	1
3,2,3	Stare Krecany2	1

Tabulka 6.12: K-means: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Length (dm) - Velocity (km/h)

Noční interval (Tabulka 6.11) vytvořil mnohem větší rozmanitost klastrových skupin. Je zde více jednotkových uskupení, než bylo možné vidět v předchozích případech. Dokonce ve třech případech rozdělil stejnou oblast od sebe a byla zařazena do jiného klastrového mixu.s

6.5.4 Shrnutí k-means

Algoritmus přinášel různé počty shluků pro denní a noční dobu. Hodnoty se pohybovaly pro k od 2 do 6, data během dne vykazovala menší tendenci se štěpit do různých klastrových mixů, než v nočních datech. Noční data naopak vykazovala větší variabilitu ve vytváření klastrových kombinací. Vytvoří-li se ucelený výstup z jednotlivých klastrových tabulek, budou finální skupiny obcí vypadat takto:

Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Skupina	Obce
Skupina 1	Bynov, Bynovec1, Bynovec2, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Libouchec
Skupina 2	Ceska Kam. Decinska, Labska Stran
Skupina 3	Stare Krecany1, Stare Krecany2, Ceska Kamenice Liska

Tabulka 6.13: K-menas: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Skupina	Obce
Skupina 1	Bynov, Bynovec1, Bynovec2, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Libouchec, Ceska Kamenice Liska
Skupina 2	Ceska Kam. Decinska, Labska Stran
Skupina 3	Stare Krecany1, Stare Krecany2, Ceska Kamenice Liska

Tabulka 6.14: K-means: Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Length (dm) - Velocity (km/h)

Výstupy jsou nakonec pro obě varianty stejné (Tabulka 6.13 a 6.14), odlišnosti jsou totiž velmi nepatrné. První klastr, který je vždy u kombinací uveden je pro danou obec největší, nejvíc datových bodů spadá právě do něj. Tudíž, pokud se liší hlavě v posledním/předposledním klastru, není to tak významný rozdíl, jako když se liší v prvním.

6.5.5 Dopravní realita podle K-means

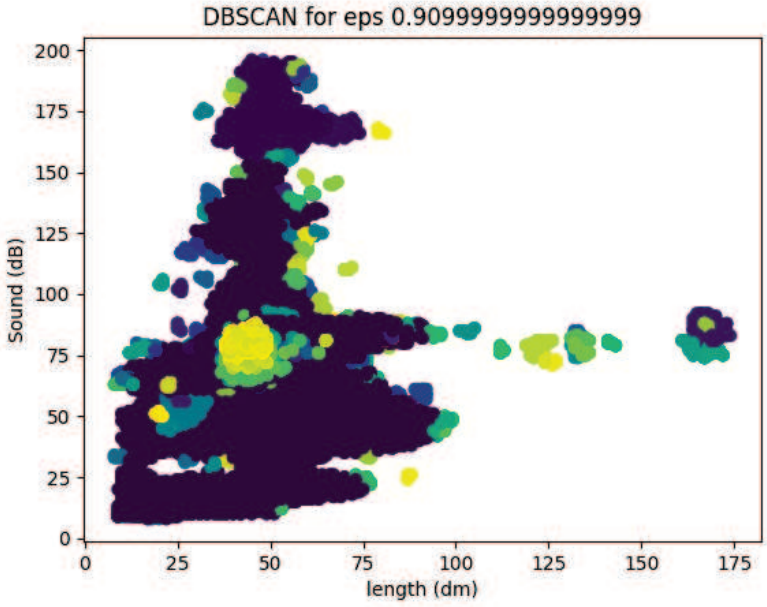
Skupina 1: Pro skupinu je typická hlavní komunikace, dlouhý rovný úsek. Komunikace jsou v dobrém stavu, s absencí světelných návěstidel. Komunikace jsou ve většině případů dobře osvětlené.

Skupina 2: Jedná se o úsek s komunikací v dobrém stavu, v mírné zástavbě a je zde veden obousměrný provoz. Nekříží se v místě žádné další komunikace a není zde přítomna světelná signalizace.

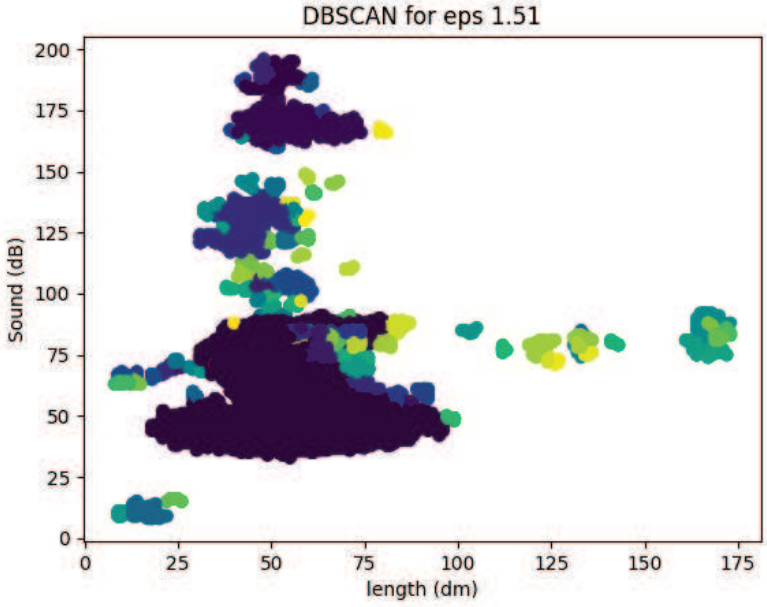
Skupina 3: Oblast má obousměrný provoz, komunikace jsou rovný dlouhý úsek v mírné zástavbě, mírně osvětlené. V oblasti je málo míst v blízkosti komunikace pro nouzové zastavení.

6.6 DBSCAN

Další z alternativ při tvorbě klastrů je DBSCAN (Obrázek 6.10 a 6.11). V tato metoda se liší přístupem přiřazování bodů do shluků od předchozí metody k-means, nepracuje se vzdáleností od pomyslného středu, ale s Epsilon okolím. Nacházejí-li se body v daném okolí, jsou zařazeny do klastrů. Pokud ne, jedná se o šum Grafy ukazují příklady volby epsilon. V prvním případě je epsilon zvoleno jako 0.9099, ve druhém jako 1,51.(Ostatní výsledky viz. příloha DBSCAN)



Obrázek 6.10: DBSCAN: Velocity km/h - length (dm) denní interval - Eps:1.6



Obrázek 6.11: DBSCAN: Velocity km/h - length (dm) denní interval - Eps:0.7

Tato metoda generovala pro dvojice dopravních dat následují klastry.(Grafické vyobrazení viz příloha DBSCAN)

6.6.1 Sound (dB) - Velocity (km/h)

Pro zvuk a rychlost jsou DBSCAN vygeneroval strukturu.
Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
0,-1,33	Bynov, Bynovec2	2
0,-1,17	Nova Oleska2, Rohatce	2
0,-1,35	Ceska Kam. Duk. Hrdinu, Decin Teplicka	2
0,-1,32	Bynovec1	1
0,-1,34	Ceska Kam. Decinska	1
0,-1,36	Ceska Kamenice Liska	1
0,-1,38	Decin Ustecka	1
0,-1,40	Hrobce	1
0,-1,1	Huntirov	1
0,-1,5	Huntirov Nov. Oles.1	1
0,-1,6	Kasna Lipa2	1
0,-1,12	Krasna Lipa1	1
0,-1,13	Labska Stran	1
0,-1,14	Libouchec	1
0,-1,16	Nebocany	1
0,-1,20	Stare Krecany1	1
0,-1,22	Stare Krecany2	1

Tabulka 6.15: DBSCAN: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Sound (dB) - Velocity (km/h)

Ačkoliv je algoritmus nastaven na epsilon 1,51, sále existuje mnoho bodů, které jsou zařazeny do klastru -1, jako šum (Tabulka 6.15). Nejpočetnější je shluk 0, ten odpovídá nejběžněji přítomným hodnotám, které se vyskytují ve všech klastrech. Zajímavější je třetí hodnota klastru. DBSCAN určil jen několik dvojic, těmi jsou:

1. Bynov, Bynovec2
2. Nova Oleska2, Rohatce.
3. Ceska Kam. Duk. Hrdinu, Decin Teplicka.

Je zajímavé, že dvojice naměřených stejných oblastí nemá stejný klastr, nicméně je možné si všimnout, že nejsou daleko od sebe. Například Bynovec 2 má třetí klastr 33 a Bynovec 1 má třetí klastr 32.(Tabulka 6.15)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
0,-1,18	Bynovec1, Bynovec2	2
0,-1,24	Decin Teplicka, Decin Ustecka	2
0,-1,4	Kasna Lipa2, Krasna Lipa1	2
0,-1,7	Libouchec, Nebocany	2
0,-1,15	Stare Krecany1, Stare Krecany2	2
0,-1,17	Bynov	1
0,-1,20	Ceska Kam. Decinska	1
0,-1,21	Ceska Kam. Duk. Hrdinu	1
0,-1,22	Ceska Kamenice Liska	1
0,-1,26	Hrobce	1
0,-1,1	Huntirov	1
0,-1	Huntirov Nov. Oles.1	1
0,-1,5	Labska Stran	1
-1,0,12	Nova Oleska2	1
0,-1,14	Rohatce	1

Tabulka 6.16: DBSCAN: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Sound (dB) - Velocity (km/h)

V nočních datech (Tabulka 6.16) je vidět větší provázanost, ačkoliv stéle DBSCAN určuje mnoho lokací, které jsou solitérní. V tomto případě už ale určil stejné páry měření v jedné oblasti jako jeden klastr, viz. Bynovec 1 a 2, Staré Křečany 1 a 2. Jsou zde dva výjimečné případy, Huntířov Nová Oleška a Nová Oleška. Tyto dvě obce mají úplně odlišnou skladbu klastrů. Huntířov Nová Oleška přísluší jen dvěma klastrům, 0 a -1, zatímco Nová Oleška nejvíce spadá pod klastr -1. To může značit velký rozptyl jednotlivých dat mimo hlavní klastr a zároveň rozestupy mezi body větší než je epsilon. Pro ostatní obce je také možné dosáhnout -1 jako hlavní klastr, ale děje se tak při menších - epsilon hodnotách.

6.6.2 Length (dm) - Sound (dB)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
0,-1,12	Huntirov Nov. Oles.1, Kasna Lipa2, Krasna Lipa1, Labska Stran	4
0,-1,56	Bynovec1, Bynovec2	2
0,-1,81	Decin Teplicka, Decin Ustecka	2
0,-1,27	Libouchec, Nebocany	2
0,-1,48	Stare Krecany1, Stare Krecany2	2
0,-1,55	Bynov	1
0,-1,68	Ceska Kam. Decinska	1
0,-1,71	Ceska Kam. Duk. Hrdinu	1
0,-1,75	Ceska Kamenice Liska	1
0,-1,84	Hrobce	1
0,-1,1	Huntirov	1
0,-1,43	Nova Oleska2	1
0,-1,40	Rohatce	1

Tabulka 6.17: DBSAN: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Length (dm) - Sound (dB)

Délka a zvuk již vykazují větší propojenost dat (Tabulka 6.17). Sice stále převažují hodnoty klastrů 0 a -1, ale jak třetí klastr je patrná větší variabilita. DBSCAN na rozdíl od předchozího páru veličin určil všechny dvojice stejných měření v obci stejným klastrem (Bynovec 1 a 2, Krásná Lípa 1 a 2, Staré Křečany 1 a 2).

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
0,-1,4	Decin Teplicka, Decin Ustecka, Hrobce	3
0,-1,11	Kasna Lipa2, Krasna Lipa1	2
0,-1,24	Bynovec1, Bynovec2	2
0,-1,14	Libouchec, Nebocany	2
0,-1,17	Stare Krecany1, Stare Krecany2	2
0,-1,18	Bynov	1
0,-1,25	Ceska Kam. Decinska	1
0,-1,28	Ceska Kam. Duk. Hrdinu	1
0,-1,36	Ceska Kamenice Liska	1
0,-1,2	Huntirov	1
0,-1,1	Huntirov Nov. Oles.1	1
0,-1,13	Labska Stran	1
0,-1	Nova Oleska2, Rohatce	2

Tabulka 6.18: DBSCAN: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Length (dm) - Sound (dB)

I v tomto (Tabulka 6.18) případě jsou všechny dvojice stejného měření společně v jednom klastru, ale jsou ovšem úplně vyčleněny od ostatních obcí. Tentokrát se Nová Oleška a Rohatce stávají lokacemi, které spadají jen do dvou klastrů. Znovu do 0 a -1. Největší skupina obsahuje 3 lokality, Děčín Teplická, Děčín Ústecká a Hrobce.

6.6.3 Length (dm) - Velocity (km/h)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
0,-1,1	Bynovec2, Bynovec1, Ceska Kamenice Liska, Libouchec, Nova Oleska2, Rohatce, Stare Krecany2, Stare Krecany1	8
0,1,-1	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Nebocany	6
-1,0,1	Hrobce	1
1,-1,0	Huntirov	1
0,-1,9	Huntirov Nov. Oles.1	1
0,-1,13	Kasna Lipa2	1
0,-1,7	Krasna Lipa1	1
0,-1,27	Labska Stran	1

Tabulka 6.19: DBSVAN: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Length (dm) - Velocity (km/h)

Délka - rychlost (Tabulka 6.19) vykazuje největší variabilitu a vytváří nejpočetnější skupinu o osmi obcích a druhou nejpočetnější o šesti. V podstatě dává dohromady všechny dvojice stejného měření až na Krásnou Lípu 1 a 2, kterou úplně vyčleňuje. Zajímavé je, že druhá nejpočetnější skupin má až na třetím místě klastrového zastoupení klastr -1, šumový: Obvykle bývá první nebo druhý v pořadí. Naopak Hrobce mají většinu dat v klastru šumu a Huntířov jako jediný má nejpočetnější klastr 1.

Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
0,8,-1	Bynovec1, Bynovec2, Decin Teplicka, Decin Ustecka, Huntirov Nov. Oles.1, Nova Oleska2, Stare Krecany1, Stare Krecany2	7
0,-1,2	Ceska Kam. Decinska, Ceska Kamenice Liska, Huntirov	4
0,-1,8	Libouchec, Nebocany, Rohatce	3
0,-1,19	Hrobce, Labska Stran	2
0,2,-1	Bynov, Ceska Kam. Duk. Hrdinu	2
0,-1,29	Kasna Lipa2	1
0,-1,33	Krasna Lipa1	1

Tabulka 6.20: DBSCAN: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Length (dm) - Velocity (km/h)

Nejpočetnější skupina (Tabulka 6.20) čítající sedm obcí má zároveň jednu z nejzajímavějších skladeb klastrů, se skupinou Bynov, Česká Kamenice Dukelských Hrdinů má až na třetím místě klastr šumu (-1). V tomto případě se také Krásná Lípa 1 a 2 rozdělily dvě samostatné uskupení.

6.6.4 Shrnutí DBSCAN

Určitý typ dat má největší zastoupení napříč obcemi (řidiči jezdí nejvíce rychlostí od 48-50 km/h nebo vozidla mají nejčastěji délku do 5 m). Proto se stávalo, že všechny obce primárně spadaly do jednoho klastru (mimo výjimky jako byla délka-rychlost obec Huntířov atd.t). Naopak existovala data, která byla už tak rozptýlená, že je epsilon parametr nezachytil. Pokud ale obráceně byl nastavený příliš velký epsilon parametr, data se shlukla do dvou klastrů a regionální rozdíly zmizely. Kvůli tomu byl raději nastaven menší než větší parametr s větším výskytem šumových dat. I tak se ale stávalo, že se jedinečnost obce projevovala až v posledním (třetím) klastru, ke kterému obec příslušela. Nejčastěji obce spadaly do klastrů 0 a -1, kdy alrgoritmus určil nejpočetnější skupinu, která činila osm obcí. Jednalo se o dvojici parametrů délka - rychlost v nočních hodinách. Nejčastěji byly zastoupeny skupiny o počtu 1. I přes velkou citlivost na parametr epsilon, algoritmus zařadil v pěti případech ze šesti dvojic měření ve stejných obcích (např.: Nová Oleška 1 a 2).

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Skupina 1	Bynovec1, Bynovec2
Skupina 2	Decin Teplicka, Decin Ustecka, Ceska Kam. Duk. Hrdinu
Skupina 3	Libouchec, Nebocany
Skupina 4	Stare Krecany1, Stare Krecany2
Skupina 5	Kasna Lipa2, Krasna Lipa1
Skupina 6	Bynov
Skupina 7	Ceska Kam. Decinska
Skupina 8	Ceska Kamenice Liska
Skupina 9	Hrobce
Skupina 10	Huntirov
Skupina 11	Huntirov Nov. Oles.1
Skupina 12	Labska Stran
Skupina 13	Rohatce,Nova Oleska2

Tabulka 6.21: DBSCAN: Denní interval fúze

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Skupina 1	Bynovec1, Bynovec2
Skupina 2	Decin Teplicka, Decin Ustecka
Skupina 3	Libouchec, Nebocany
Skupina 4	Stare Krecany1, Stare Krecany2
Skupina 5	Kasna Lipa2, Krasna Lipa1
Skupina 6	Bynov
Skupina 7	Ceska Kam. Decinska
Skupina 8	Ceska Kam. Duk. Hrdinu
Skupina 9	Ceska Kamenice Liska
Skupina 10	Hrobce
Skupina 11	Huntirov
Skupina 12	Huntirov Nov. Oles.1
Skupina 13	Labska Stran
Skupina 14	Nova Oleska2
Skupina 15	Rohatce

Tabulka 6.22: DBSCAN: Noční interval fúze

Je patrné, že DBSCAN určil pouze dvě skupiny jak pro noc tak i pro den (Tabulka 6.21 a 6.22). Pro den: Nová Oleška, Rohatce a Česká Kamenice Dukelských Hrdinů, Děčín - Teplická, Děčín - Ústecká. Pro noc: Děčín - Teplická, Děčín - Ústecká a Libouchec, Nebočany.

Jediný průnik je tedy v Děčíně Teplická/Ústecká pro den i noc. Nicméně, pro definování společných infrastrukturních prvků na komunikaci, se bude řešit noc a den zvlášť.

6.6.5 Dorapvní realita podle DBSCAN

Smysl má vyhodnocovat jen pouze skupiny, které jsou alespoň ve dvojicích.
Skupina je stejná jak pro den, tak pro noc.

Skupina 3: Libouchec, Nebočany Místa měření jsou si podobná v charakteru oblasti. Vozidlo jede v obou případech po hlavní obousměrné komunikaci, která následně kříží několik vedlejších komunikací. Dopravní prostředky projíždí hustou zástavbou s úzkými chodníky končící domy. Řidiči museli překonat mírný směrový směrový oblouku v obou případech.

Dopravní realita obcí/lokalit pro interval *6:00 h až 22:00 h* Skupina 2: Děčín Teplická, Děčín Ústecká, Česká Kamenice Dukelských Hrdinů Všechna tři místa, kde byla prováděna měření, se nacházejí na přímém úseku, který je i hlavní silnicí. Komunikace je přehledná, osvětlená s chodníky, či širším prostorem na krajích komunikace. Infrastruktura se nekříží s vedlejšími komunikacemi, pouze s přechodem pro chodce, který je ale jen v jednom případě oddělen ostrůvkem a v druhém speciálně označen.

Dopravní realita obcí/lokalit pro interval *22:00 h až 6:00 h* Skupina 2: Děčín Teplická, Děčín Ústecká Silnice je v obou případe přímá bez křížení, obsahuje přechod pro chodce bez značení. Oblast je dobře osvětlena a v jednom případě navazuje na komunikace, kde je povolena rychlost 70 km/h.

6.7 Fuzzy C-means

Dalším z klastrových přístupů je fuzzy c-means.(Viz grafická příloha Fuzzy c-means) k Bylo vybráno jako páté a jedná se o podobné hodnoty jako byly u k-means, tudíž bude jednodušší výsledky porovnávat.:

6.7.1 Length (dm) - Sound (dB)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Nejpočetnější skupina obsahuje sedm členů (Tabulka 6.25), a nejmenší dva. Všechny páry měření ve stejné obci mimo Krásnou Lípu jsou ve stejných skupinách klastrů. Ani v jednom případě nenastává osamocení obce v klastrové kombinaci, jako tomu bylo v předcházející metodě.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Na rozdíl ode dne se zde již vyskytují osamocené lokality v klastrové kombinaci. Celkem se v tabulce vyskytují tři osamocené varianty. Největší skupina čítá osm lokalit. Druhá největší a nejčastěji se vyskytující se hodnota je dvě. Lokalit v páru jsou čtyři, když se nebude počítat dvakrát provedené měření v jedné obci (Staré Křečany 1 a 2). Jediná taková dvojice, která nemá samostatné klastrové seskupení ani není ve větší skupině, je obec Bynovec 1 a 2. Nejčastěji se vyskytuje v hodnotách klastr 1 a 3. V obou případech jako dominantní.

6.7.2 Velocity (km/h) - Sound (dB)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
0,1,2	Bynov, Bynovec2, Bynovec1, Ceska Kam. Duk. Hrdinu, Decin Teplicka	5
3,1,2	Kasna Lipa2, Krasna Lipa1, Libouchec, Nova Oleska2	4
3,2,1	Labska Stran, Nebocany, Rohatce	3
0,2,1	Ceska Kam. Decinska, Ceska Kamenice Liska	2
2,3,1	Stare Krecany2, Stare Krecany1	2
0,1,3	Decin Ustecka	1
1,3,0	Huntirov Nov. Oles.1	1
2,0,3	Hrobce	1
3,0,2	Huntirov	1

Tabulka 6.25: Fuzzy c-means: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Velocity (km/h) - Sound (dB)

V případě rychlosti a délky (Tabulka 6.26) se vyskytují všechny hodnoty skupin od jedné do pěti, kde skupiny o jedné lokalitě jsou čtyři a skupina o pěti je jedna. Dvakrát provedené měření v jedné obci se vyskytuje v každém případě v jedné skupině klastrových kombinací. Nejčastěji se vyskytují klastry 0 a 3.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
3,0,2	Bynov, Bynovec2, Bynovec1, Ceska Kamenice Liska, Ceska. Kam. Deciska	5
1,0,2	Huntirov, Libouchec, Nebocany, Rohatce	4
0,2,1	Ceska Kam. Duk. Hrdinu, Decin Teplicka	2
1,0,3	Kasna Lipa2, Krasna Lipa1	2
3,1,0	Labska Stran, Nova Oleska2	2
2,0,1	Stare Krecany2, Stare Krecany1	2
0,1,2	Decin Ustecka	1
1,2,0	Hrobce	1
1,3,0	Huntirov Nov. Oles.1	1

Tabulka 6.26: Fuzzy c-means: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Velocity (km/h) - Sound (dB)

Výsledek (Tabulka 6.27) je podobný jako pro denní data. Lokality, které jsou osamocené, mají společné prvky Děčín Ústecká a Huntířov Nová Oleška. Klastrové kombinace jsou velmi rozmanité, střídají se variace klastrů 3,2,1,0. Nejpočetnější skupiny obsahují pět lokalit.

6.7.3 Velocity (km/h) - length (dm)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
2,0,3	Ceska Kam. Decinska, Ceska Kamenice Liska, Hrobce, Labska Stran, Stare Krecany2, Stare Krecany1	6
0,3,2	Decin Teplicka, Decin Ustecka, Huntirov, Krasna Lipa1, Nebocany	5
3,0,2	Huntirov Nov. Oles.1, Kasna Lipa2, Libouchec, Nova Oleska2	4
0,2,3	Bynovec2, Bynovec1, Rohatce	3
3,0,1	Bynov, Ceska Kam. Duk. Hrdinu	2

Tabulka 6.27: Fuzzy c-means: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Velocity (km/h) - length (dm)

Tabulka 6.28 rychlost - délka ukazuje větší celky, než v předchozích případech. Sice neobsahuje větší, nebo stejně velkou skupinu jako v první dvojici dopravních veličin, ale zase neobsahuje žádné solitérní lokality. Nejnižší hodnota pro skupinu jsou dvě, a největší skupina čítá šest lokalit. Největší zastoupení mají klastry 0 a 3. Třikrát se vyskytuje i klastr 2 a to v dominantní pozici v největší skupině lokalit. Jednou se v klastrových variacích vyskytuje klastr 1, konkrétně ve variaci 3,0,1.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
1,2,0	Bynovec2, Bynovec1, Decin Teplicka, Decin Ustecka, Hrobce, Kasna Lipa2, Krasna Lipa1, Labska Stran, Nebocany, Rohatce	10
1,2,3	Bynov, Ceska Kam. Duk. Hrdinu, Huntirov, Libouchec	4
2,1,3	Ceska Kam. Decinska, Ceska Kamenice Liska	2
1,0,2	Huntirov Nov. Oles.1, Nova Oleska2	2
2,1,0	Stare Krecany2, Stare Krecany1	2

Tabulka 6.28: Fuzzy c-means: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Velocity (km/h) - length (dm)

V nočním intervalu (Tabulka 6.29) nalézáme největší skupinu obcí, kterých je pro klastrovou kombinaci 1,2,0, celkem deset. Druhá nejnižší hodnota počtu obcí je dva. Všechny párová měření se nachází ve společných klastrech. Nejčastěji dominantní klastr je číslo 1 a 2, často střídají první a druhou pozici v sestupné klastrové kombinaci.

6.7.4 Shrnutí Fuzzy C-means

Pro algoritmus Fuzzy c-means byla vybrána hodnota $k - 5$. Jelikož bylo již v k-means pracováno s podobnou hodnotou, bude srovnání jednotlivých přístupů jednodušší. Metoda generuje poměrně širokou škálu kombinací, nicméně nemá tendence oddělovat opakovaná měření v jedné obci od sebe. V několika případech vytvořila solitérní klastrovou kombinaci, ale vždy i vícenásobnou. Fuzzy c-means také vytvořil jednu z největších skupin, obsahující 10 lokalit. Va každém měřeném páru pro den a noc bylo možné nalézt skupinu o velikosti alespoň čtyř lokalit.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Skupina 1	Bynov, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Bynovec2, Bynovec1
Skupina 2	Rohatce, Nebocany
Skupina 3	Libouchec, Nova Oleska2
Skupina 4	Decin Ustecka, Huntirov
Skupina 5	Krasna Lipa1, Krasna Lipa2, Huntirov Nov. Oles.1
Skupina 6	Krecany2, Stare Krecany1
Skupina 7	Ceska Kam. Decinska, Ceska Kamenice Liska
Skupina 8	Labska straň

Tabulka 6.29: Fuzzy c-measn: Denní interval fúze

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Skupina 1	Libouchec, Huntirov
Skupina 2	Stare Krecany2, Stare Krecany1
Skupina 3	Huntirov Nov. Oles.1, Nova Oleska2
Skupina 4	Ceska Kamenice Liska, Ceska Kam. Deciska, Bynovec2, Bynovec1, Bynov
Skupina 5	Nebocany, Rohatce, Labska Stran
Skupina 6	Decin Teplicka, Decin Ustecka, Kasna Lipa2, Krasna Lipa1
Skupina 7	Hrobce
Skupina 8	Ceska Kam. Duk. Hrdinu

Tabulka 6.30: Fuzzy c-measn: Noční interval fúze

I přes to, že Fuzzy c-means generoval velké skupiny obcí a minimum solitérních, průnik denních a nočních hodnot (Tabulka 6.29 a 6.30) stále vytváří 8 skupin. Jediný průnik obou tabulek se nachází v oblasti Bynovec 1 a 2, Bynov. Následně Česká Kamenice Liška, Česká Kamenice Děčínská a ve všech párech měření v jedné obci (Krásná Lípa 1 a 2, Staré Křečany 1 a 2 atd.)

6.7.5 Dopravní realita podle Fuzzy C-means

Dopravní realita obcí/lokalit pro interval *6:00 h až 22:00 h* Skupina 1: Bynov, Česká Kamenice Dukelských Hrdinů, Děčín Teplická, Bynovec
Všechny oblasti mimo Bynovce křižují přechod. Jedná se o rovné přímé úseky s dobrými rozhledy, kde komunikace je hlavní komunikací.

Skupina 2: Rohatce, Nebočany
Vozidla v obou případech musí projíždět směrovým obloukem, který je v obou obcích hlavní komunikací. Komunikaci v místě měření křižují vedlejší komunikace a místní. V místech nejsou přechody a v některých částech ani chodníky.

Skupina 3: Libouchec, Nová Oleška
Společný charakter je zde postaven na směrových vlastnostech. V obou případech vozidla jedou po komunikaci, která má více směrových oblouků, ve kterých se nedá bezpečně předjíždět pomalé vozidlo. Zároveň je v některých místech absence krajnice, buď jsou instalována svodidla, nebo staré betonové sloupky.

Skupina 4: Děčín Ústecká, Huntířov
V obou případech se nachází v místě široká hlavní komunikace, která měla na jedné straně změnu maximální rychlosti z 70/90 km/h na 50 km/h. Na komunikacích nelze předjíždět.

Skupina 5: Krásná Lípa, Huntířov Nová Oleška
Lokace neumožňují předjíždění vozidel. Místo měření se nachází v blízkosti vjezdu do obce, kde byl razantní přechod z 90 km/h na 50 km/h.

Skupina 7: Ceska Kam. Decinska, Ceska Kamenice Líska
Vozidla projíždějí nepříliš hustou zástavbou, komunikace je dlouhá, široká a obsahuje několik mírných směrových oblouků. V okolí komunikace se nacházejí svodidla a absentuje vetší krajnice pro bezpečnější vyhýbání vozidel nebo odstavení.

Dopravní realita obcí/lokalit pro interval *22:00 h až 6:00 h* Skupina 4: Ceska Kamenice Liska, Ceska Kam. Deciska, Bynovec2, Bynov Na komunikaci jsou dobré rozhledové poměry, vyskytují se zde minimální počty křížení komunikací a pokud, tak primárně napojení účelových komunikací.

Skupina 5: Nebocany, Rohatce, Labska Stran Všech případech se jedná o užší prostor bez krajnice, a bez chodníků. Ve dvou případech nejsou dopravní pruhy vyznačeny.

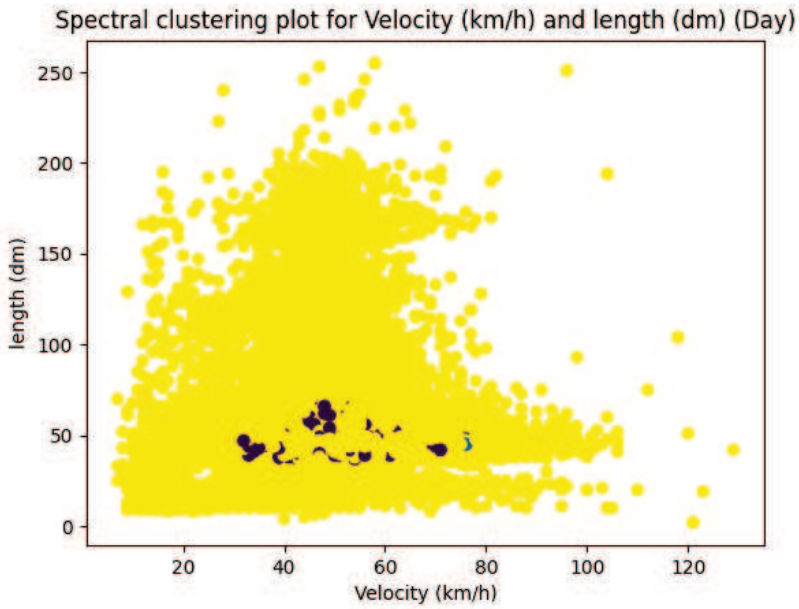
Skupina 6: Decin Teplicka, Decin Ustecka, Kasna Lipa V místě měření není umožněno vozidlům předjíždět, komunikace je mírně osvětlená. V oblasti je několik směrových oblouků a výjezd z obce.

6.8 Spektrální klastrování

Obrázek 6.12 ukazuje příklad, jak takový graf ze Spektrální analýzy může vypadat. Spektrální analýza klastrů při volně $k = 5$ vygenerovala tyto výsledky (Viz grafické přílohy Spektrální klastrování):

6.8.1 Length (dm) - Sound (dB)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*



Obrázek 6.12:

Klastr	Názvy obcí	Počet obcí
0,4	Bynov, Ceska Kam. Duk. Hrdinu, Hrobce, Huntirov, Kasna Lipa2, Krasna Lipa1, Labska Stran, Libouchec, Rohatce, Stare Krecany2	10
0,4,1	Bynovec2, Bynovec1, Decin Teplicka	3
0,1,3,4	Ceska Kamenice Liska, Stare Krecany1	2
0	Huntirov Nov. Oles.1, Nova Oleska2	2
0,1,2,4	Ceska Kam. Decinska	1
0,4,1,3	Decin Ustecka	1
0,2,1,4	Nebocany	1

Tabulka 6.31: Spektrální klastrvání: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Length (dm) - Sound (dB)

Výstupy (Tabulka 6.31) pro Spektrální klastrování jsou dosti odlišné od ostatních metod. Prní tabulka zobrazuje skupinu deseti měření. Téměř všechna párová měření byl algoritmus schopný začlenit k sobě. Vytvořil i dvě solitérní skupiny Děčín Ústecká a Nebočany. Hlavní klastr byl 0 a pouze Nová Oleška a Huntířov Nová Oleška spadaly pouze pod klastr 0.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
0,2,1	Bynovec1, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka, Hrobce, Labska Stran, Libouchec, Nebocany, Stare Krecany1	9
0,2	Bynovec2, Huntirov Nov. Oles.1, Kasna Lipa2, Krasna Lipa1, Nova Oleska2, Rohatce, Stare Krecany2	7
0,1,2	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Huntirov	4

Tabulka 6.32: Spektrální klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Length (dm) - Sound (dB)

Pro noční data (Tabulka 6.32) už algoritmus vytvořil méně skupin, než vyšle pro denní interval. Největší skupina čítá devět měření a liší se od druhé pouze o třetí klastr, který je pro druhou skupinu rozšířen o klastr 1. Třetí skupina má základní odlišnost ve druhém klastru, který je zde určen jako klastr číslo 2.

6.8.2 Velocity (km/h) - Sound (dB)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
0,3,1	Bynovec2, Decin Ustecka, Kasna Lipa2, Krasna Lipa1, Libouchec	5
0,3	Huntirov, Huntirov Nov. Oles.1, Labska Stran, Nova Oleska2, Rohatce	5
0,3,1,2	Bynov, Bynovec1, Ceska Kam. Duk. Hrdinu, Decin Teplicka	4
0,3,4	Ceska Kam. Decinska, Ceska Kamenice Liska, Nebocany	3
3,0,1	Hrobce, Stare Krecany1	2
3,0	Stare Krecany2	1

Tabulka 6.33: Spektrální klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Velocity (km/h) - Sound (dB)

Denní interval (Tabulka 6.33) pro rychlost a zvuk jsou skupiny středně velké a pouze Staré Křečany 2 se úplně oddělily od větších skupin. Dvě největší skupiny jsou o počtu pěti lokalit.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
1,0	Bynovec2, Huntirov, Huntirov Nov. Oles.1, Kasna Lipa2, Krasna Lipa1, Labska Stran, Libouchec, Nebocany, Rohatce	9
0,1	Ceska Kam. Decinska, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka, Hrobce, Stare Krecany2, Stare Krecany1	7
1,4,0	Bynov	1
1,0,4	Bynovec1	1
1,0,3,2	Ceska Kam. Duk. Hrdinu	1
1	Nova Oleska2	1

Tabulka 6.34: Spektrální klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Velocity (km/h) - Sound (dB)

Naopak noční interval (Tabulka 6.34) rovnoměrnost ztratil. Jsou zde patrné dvě dominantní skupiny, které střídají pořadí klastru 0,1 nebo 1,0. Skupin lokalit s pořadím klastrů 1,0 je devět a skupiny s pořadím klastrů 0,1 je sedm. Pak se už vyskytují pouze solitérní lokality.

6.8.3 Velocity (km/h) - length (dm)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
2,3,1,0	Bynov, Bynovec1, Ceska Kam. Duk. Hrdinu, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Krasna Lipa1, Labska Stran, Nova Oleska2, Stare Krecany2, Stare Krecany1	13
2,3,0,1	Huntirov Nov. Oles.1, Kasna Lipa2, Nebocany	3
3,2,1,0	Bynovec2, Rohatce	2
1,2,3,0	Ceska Kam. Decinska	1
0,2,3,1	Libouchec	1

Tabulka 6.35: Spektrální klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Velocity (km/h) - length (dm)

Dvojice rychlost - délka (Tabulka 6.35) zobrazuje ještě extrémnější dominanci jedno klastrového mixu. Bylo nutné jemnějšího rozdělení než v jiných případech, aby bylo možné jednoznačně lokality rozdělit. Největší skupina obsahuje 13 lokalit, nejmenší 1. Solitérní obce jsou ve dvou případech, Libouchec a Česká Kamenice Děčínská.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
0,2,1	Bynovec1, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka, Hrobce, Labska Stran, Libouchec, Nebocany, Stare Krecany1	9
0,2	Bynovec2, Huntirov Nov. Oles.1, Kasna Lipa2, Krasna Lipa1, Nova Oleska2, Rohatce, Stare Krecany2	7
0,1,2	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Huntirov	4

Tabulka 6.36: Spektrální klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Velocity (km/h) - length (dm)

Na rozdíl (Tabulka 6.36) od denní případu se lokality rozdělily spíš do tří skupin než do jedné. V některých případech nebyl schopný algoritmus propojit dvojice měření, jako je Bynovec 1 a 2, nebo Staré Křečany 1 a 2. Největší skupina činní 9 lokalit a nejmenší 4.

6.8.4 Shrnutí Spektrálního klastrování

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* pro všechny dvojice:

Skupina	Obce
Skupina 1	Libouchec, Krasna Lipa2, Krasna Lipa1
Skupina 2	Ceska Kam. Decinska
Skupina 3	Huntirov Nov. Oles.1, Nova Oleska2, Huntirov, Labská stráň
Skupina 4	Nebocany, Hrobce, Stare Krecany2, Stare Krecany1
Skupina 5	Bynov, Bynovec1, Bynovec2, Ceska Kam. Duk. Hrdinu, Decin Ustecka, Decin Teplicka
Skupina 6	Ceska Kamenice Liska, Rohatce

Tabulka 6.37: Spektrální klastrování: Denní interval fúze

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* pro všechny dvojice:

Skupina	Obce
Skupina 1	Hrobce, Stare Krecany2, Stare Krecany1
Skupina 2	Bynov, Bynovec1, Bynovec2
Skupina 3	Ceska Kam. Decinska, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka, Stare Krecany2, Stare Krecany1
Skupina 4	Libouchec, Nebocany, Rohatce
Skupina 5	Huntirov, Huntirov Nov. Oles.1, Kasna Lipa2, Krasna Lipa1, Labska Stran

Tabulka 6.38: Spektrální klastrování: Denní interval fúze

Spektrální klastrování tvořilo v některých případech extrémně velké skupiny, obecně ale skupiny byl rovnoměrné a minimum z nich obsahovalo solitérní lokality. Bylo ovšem nutné zkoumání a vyhodnocování klastrových mixů prozkoumávat až do jemných rozdílů. Naopak někdy Spektrální klastrování určilo pouze jeden jediný klastr pro skupinu, což je v celé práci unikátní.

6.8.5 Dopravní realita podle Spektrálního klastrování

Společné případy (Tabulka 6.37 a 6.38):

Dopravní realita obcí/lokalit pro interval *6:00 h až 22:00 h* Skupina 1: Libouchec, Krásná Lípa Komunikace jsou v obou případe se směrovými oblouky, nejsou zde ideální rozhledové poměry a okolí komunikace neposkytuje příliš prostoru.

Skupina 3: Nová Oleška, Huntířov, Labská Stráň
Dvě ze tří oblastí byly měřeny blízko u vjezdu do obce.

Skupina 4: Nebočany, Hrobce, Staré Křečany
V místě měření jsou rovné, ale dále jedním směrem jsou buď křížené nebo je zde směrový oblouk. Dvě lokality mají málo prostoru v přilehlém okolí silnice.

Skupina 5: Bynov, Bynovec, Česká Kamenice Dukelský Hrdinů, Děčín Ústecká, Děčín Teplická
Ve většině případů jsou úseky dlouhé, rovné, dobře osvětlené. Jedná se o hlavní komunikace a kříží je vedlejší nebo účelové komunikace. Ve většině případů se zde nachází přechody nebo zúžení.

Skupina 6: Česká Kamenice Líska, Rohatce

Silnice se nachází v malé obci, která je v místě měření křížena vedlejší komunikací, která je místního charakteru. Oblast je dobře osvětlená a v přilehlé oblasti komunikace je hodně prostoru pro případné vyhnutí se nákladních vozidel.

Dopravní realita obcí/lokalit pro interval *22:00 h až 6:00 h*

Skupina 1: Hrobce, Staré Křečany
Společným prvkem je dlouhá rovná komunikace, méně osvětlená, měření bylo prováděno v blízkosti vjezdu do obce.

Skupina 2: Bynov, Bynovec
Vozovka je vedena méně hustou zástavbou.

Skupina 3: Česká Kamenice Děčínska, Česká Kamenice Líska, Děčín Teplicka, Děčín Ústecká, Staré Krečany
Ve většině případů jsou úseky dobře osvětleny a křižují je vedlejší komunikace. Všechny jsou přímé a je možné vidět již z dálky protijedoucí vozidlo. V okolí se spíš nedá sjet z vozovky, v některých případech pouze na chodník.

Skupina 4: Libouchec, Nebočany, Rohatce
Na všech místech měření je v bezprostřední vzdálenosti křížení. Komunikace je dobře osvětlená a není vhodná pro předjíždění vozidel.

Skupina 5: Huntířov, Nová Oleška, Krásná Lípa
Dva případy jsou měřeny ve směrovém oblouku(Nová Oleška, Krásná Lípa), a poslední případ je výjezdem ve měrovém oblouku. Komunikace je ve dvou případech úzká a přilehlý prostor neumožňuje vyhnutí se vozidel.

6.9 Hierarchické klastrování

Poslední klastrovací se liší od předchozích tím, že neexistuje "ideální"klastr, ale v čase se vytváří nové klastry podle určitého přístupu. Začíná se s maximálním počtem klastrů = počet datových bodů a končí se s jedním klastrem, který obsahuje jen jeden klastr. V tomto případě vyextrahujeme ze sekvence dobu, kdy algoritmus nalezne pět klastrů. Výsledky se budou zkoumat níže.

6.9.1 Velocity (km/h) - Sound (dB)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
3,2,1,5,4	Bynov, Bynovec1, Bynovec2, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Huntirov Nov. Oles.1, Kasna Lipa2, Labska Stran, Libouchec, Nova Oleska2, Rohatce, Stare Krecany1, Stare Krecany2	17
2,3,1,4,5	Ceska Kamenice Liska, Krasna Lipa1	2
2,3,1	Nebocany	1

Tabulka 6.39: Hierarchické klastrování: Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Velocity (km/h) - Sound (dB)

Algoritmus vygeneroval druhou největší skupinu (Tabulka 6.39) v celé práci, čítající 17 obcí. Následuje skupina se dvěma obcemi: Česká Kamenice Liška, Krásna Lípa1 a solitérní Nebočany. Je zajímavé, že se sedmnáct obcí shoduje až do pátého klastru. I přes tak vysokou přesnost nebyl schopný algoritmus přiřadit Krásnou Lípu 1 ke Krásné Lípě 2. Ovšem liší se pouze v pořadí posledního a předposledního klastru, takže rozdíly jsou malé.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
3,2,1,5,4	Bynov, Bynovec1, Bynovec2, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Huntirov Nov. Oles.1, Kasna Lipa2, Labska Stran, Libouchec, Nova Oleska2, Rohatce, Stare Krecany1, Stare Krecany2	17
2,3,1,4,5	Ceska Kamenice Liska, Krasna Lipa1	2
2,3,1	Nebocany	1

Tabulka 6.40: Hierarchické klastrování: Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Velocity (km/h) - Sound (dB)

V nočním intervalu (Tabulka 6.40) již tak velkou skupinu není možné nalézt. Ovšem, stále zde jsou početné skupiny o 5 až 4 obcích, pouze Labská stráň a Česká Kamenice Liška nejsou do žádné skupiny zařazeny. Je na první pohled zřejmé, že skupiny se dělí dvě zaklání, ty které začínají klastrem 1 a ty které začínají klastrem 2.

6.9.2 length(dm) - Sound (dB)

Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
3,2,1,5,4	Bynov, Bynovec2, Bynovec1, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Ceska Kamenice Liska, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Huntirov Nov. Oles.1, Kasna Lipa2, Labska Stran, Libouchec, Nebocany, Nova Oleska2, Rohatce, Stare Krecany2, Stare Krecany1	19
2,3,1,5,4	Krasna Lipa1	1

Tabulka 6.41: Hierarchické klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - length(dm) - Sound (dB)

Největší vygenerovaný klastr (Tabulka 6.41) obsahuje devatenáct obcí. Algoritmus znovu určil schodu až na 5 klastrů hluboko. Znovu se nepodařilo zařadit Krásnou Lípu 1, která se liší na rozdíl od minulé neshody v prvním dominantním klastru.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
1,2,3,4,5	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Hrobce, Decin Teplicka, Decin Ustecka	5
1,2,3,5	Kasna Lipa2, Libouchec, Nebocany, Huntirov, Nova Oleska2	5
1,2,5,3	Bynovec1, Bynovec2, Krasna Lipa1, Rohatce	4
2,1,3,5	Huntirov Nov. Oles.1, Stare Krecany1, Stare Krecany2	4
2,1,3,4,5	Ceska Kamenice Liska	1
2,1,3,4	Labska Stran	1

Tabulka 6.42: Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - length(dm) - Sound (dB)

Znovu se v nočních datech (Tabulka 6.42) vyskytuje větší variabilita klastrů a jejich kombinací. Solitérní obce jsou pouze dvě: Česká Kamenice a Labská Stráň. V tabulce je možné zaznamenat hned dvě skupiny o počtu 5 a 4 obcí. Znovu se klastrový mix dělí na dvě základní skupiny s dominantními klastry 1 nebo 2. Například Česká Kamenice Liška se liší pouze v pořadí dominantního klastru. Ovšem, je důležité říct, že do prvního klastru vždy obec spadá nejvíce. Jedná se o sestupné pořadí náležitosti obcí k jednotlivým klastrům.

6.9.3 Velocity (km/h) - length(dm)

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Klastr	Názvy obcí	Počet obcí
3,2,1,5,4	Bynov, Bynovec1, Bynovec2, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Huntirov Nov. Oles.1, Kasna Lipa2, Labska Stran, Libouchec, Nova Oleska2, Rohatce, Stare Krecany1, Stare Krecany2	17
3,2,1,4,5	Ceska Kamenice Liska, Nebocany	2
2,3,1,4,5	Krasna Lipa1	1

Tabulka 6.43: Hierarchické klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h* - Velocity (km/h) - length(dm)

Poměrně zajímavý výsledek (Tabulka 6.43) skýtá i rychlost-délka, pokud se porovná s tabulkou rychlost-zvuk, jsou výsledky stejné, až na obec Krásná Lípa 1. Ta se ze skupiny s Českou Kamenicí Liškou prohodila s Nebočanami, a stala se soliterní obcí. Znovu je zde patrný jedna velká skupiny o sedmnácti obcích, stejná jako u dvojice veličin rychlost-zvuk.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Klastr	Názvy obcí	Počet obcí
1,2,3,4,5	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Hrobce, Decin Teplicka, Decin Ustecka	5
1,2,3,5	Kasna Lipa2, Libouchec, Nebocany, Huntirov, Nova Oleska2	5
1,2,5,3	Bynovec1, Bynovec2, Krasna Lipa1, Rohatce	4
2,1,3,5	Huntirov Nov. Oles.1, Stare Krecany1, Stare Krecany2	4
2,1,3,4,5	Ceska Kamenice Liska	1
2,1,3,4	Labska Stran	1

Tabulka 6.44: Hierarchické klastrování: Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h* - Velocity (km/h) - length(dm)

Výstup je identický s výstupem pro dvojici délka-zvuk (Tabulka 6.44). To je jediný případ, kdy jsou tabulky shodné. V pozdějším vyhodnocení je zapotřebí na to brát zřetel.

6.9.4 Shrnutí hierarchického klastrování

Pro hierarchické klastrování se vyhodnocovalo $k = 5$. Metoda vytvářela největší skupiny obcí a to sedmnáct až devatenáct. Jen minimálně bylo možné zaznamenat solitérní obce, a to v případech, když byl klastrový mix definován poměrně jemně. Nejčastějšími solitéry byla Labská Stráň, Česká Kamenice Liška a někdy i Krásní Lípa 1. Metoda při určování byla mnohem konzistentnější než předešlé přístupy.

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *6:00 h až 22:00 h*

Skupina	Obce
Skupina 1	Bynov, Bynovec1, Bynovec2, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Decin Teplicka, Decin Ustecka, Hrobce, Huntirov, Huntirov Nov. Oles.1, Kasna Lipa2, Labska Stran, Libouchec, Nova Oleska2, Rohatce, Stare Krecany1, Stare Krecany2, Krasna Lipa1
Skupina 2	Ceska Kamenice Liska

Tabulka 6.45: Hierarchické klastrování: Denní interval fúze

Výstupní tabulka mixů klastrů a obcí/lokalit pro interval *22:00 h až 6:00 h*

Skupina	Obce
Skupina 1	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Hrobce, Decin Teplicka, Decin Ustecka
Skupina 2	Kasna Lipa2, Krasna Lipa1, Libouchec, Nebocany, Huntirov, Nova Oleska2
Skupina 3	Bynovec1, Bynovec2, Rohatce
Skupina 4	Huntirov Nov. Oles.1, Stare Krecany1, Stare Krecany2
Skupina 5	Ceska Kamenice Liska
Skupina 6	Labska Stran

Tabulka 6.46: Hierarchické klastrování: Denní interval fúze

První tabulka 6.45 ukazuje, že denní data jsou mnohem více konzistentní, než noční (Tabulka 6.46). V obou případech sjednocení se Česká Kamenice Liška oddělila od větší skupiny dat a ani neměla tendence vytvářet dvojice, jako to bylo v jiných případech klastrování. Stojí za povšimnutí, že ostatní měření z České Kamenice zůstalo vždy ve stejném klastrovém mixu.

6.9.5 Dopravní realita podle hierarchického klastrování

Data pro denní dobu (Tabulka 6.45) není možno dobře interpretovat, jelikož je skupina příliš velká.

Výstupní tabulka mixů klatrů a obcí/lokalit pro interval *22:00 h až 6:00 h* Skupina 1: Tato početná skupina je velice rozmanitá společným prvkem je dlouhý rovný úsek, který je v blízkosti měření ohraničen komunikací s vyšší maximální povolenou rychlostí, než je 50 km/h.

Skupina 2: Oblasti jsou si spíše nepodobné, jediná možná souvislost je dána zakřivením vozovky. Nejedná se o přímé rovné úseky, je možné zde nalézt spíše směrové oblouky, ve který se dopravní veličiny měřily nebo ze kterých vozidla vyjíždí. Dalším společným prvkem je upozornění, nebo spíše charakter komunikace vyzývající ke klidnější jízdě, například přechod, křížení či tabulka upozorňující na výskyt chodců na vozovce.

Skupina 3: Úseky jsou si podobné, jedná se o místa s širší komunikací bez zvýraznění jednotlivých pruhů. Silnice je vedena mírně hustou zástavbou, ovšem se možností vozidlo odstavit mimo komunikaci. Rozhledové poměry narušuje zakřivení a stoupání/klesání komunikace.

Skupina 4 Lokality jsou si velice podobné, jedná se o úsek v malé obci. Komunikace je úzká a bez krajnice nebo možnosti

vozidlo bezpečně odstavit či se bezpečně vyhnout širším vozidlům. Komunikace kříží vedlejší účelové komunikace, silnice je neosvětlená.

6.10 Bayes

V této kapitole budou data z předchozích metod použit ve strojovém učení s učitelem. Díky označení dat jednotlivými klastry bude možné naučit model strukturu a následně sám predikovat. Data se rozdělí na 80% a 20%. Budou testovány jednotlivé varianty $k = 2, 4, 6, 7, 8, 9$ a porovnávat se jejich hodnoty

6.10.1 K-means inicializace algoritmu

Pro K-means by zvolena hodnoty k jako 5 a 6. Naivní Bayes měl schopnost odhadovat data následovně.
Pro denní data:
Velocity (km/h) - Sound (dB):
Grafické znázornění odhadování klastrů se zobrazuje jako úspěšné, až na klastr číslo 5. Zde predikce výrazně zvýšila počet predikovaných dat. Graf zobrazující data samotná ukazuje, že Bayes nebyl schopný predikovat data nad 75 dB. Úspěšnost predikce činí 30.120482%

length (dm) - Sound (dB):
Ukazovátko poměrně úspěšně odhaduje klastry 1, 2 a 4, ovšem 5 a 3 téměř vůbec. Stejně tak hodnoty pod 75 dB a 120 dm. Dá se předpokládat, že tyto hodnoty náleží do klastru 3 a 5. Nicméně přesnost činní 18,143161%

Velocity (km/h) - length (dm):
Grafické zobrazení ukazovátka nijak nezobrazuje schopnost správně nebo špatně odhadnuté hodnoty. Graf zobrazující data rychlost - délka ukazuje, že naivní Bayes nebyl schopný odhadnout data od 10 km/h do 120 km/h a od 50 dm po 100 dm. Přesnost činní 36,658847%

Pro noční data: Pro noční data vychází dost rozdílné výsledky, než pro denní data.
Velocity (km/h) - Sound (dB):
Ukazovátko znovu poměrně dobře ukazuje, že Bayes byl schopný odhadovat většinu klastrů, jen klastr 3 nebyl schopný vůbec zaznamenat. Klastr číslo 2 odhadoval, ale neúplně. Pokud se ale porovná s grafem hodnot, schopnost predikovat se razantně snižuje, pouze hodnoty nad 80 dB měly úspěšnou predikci. To se projevuje i v úspěšnosti, která dosahuje pouze 14,501312%

length (dm) - Sound (dB):
Zobrazení predikce klastrů zobrazuje evidentní potíže s jejich odhadem. Klastry 1,3 a 4 nejvíce zobrazují vychýlenost algoritmu, 1 a 3 jsou obsahují predikce bez reálných dat a naopak 4 obsahuje minimum až žádné predikce. Nepredikovaná, nebo špatně predikovaná data, se nachází v intervalu od 82 db do 90 dB a od 40 dm do 200 dm. Přesnost činní 15,616798%

Velocity (km/h) - length (dm):
Odhad klastrů graficky je úspěšný, uazovátku nemělo problém rozpoznat všechny klastry, nicméně hustota dat není znemožňuje přesnější popis pomocí grafu. Druhý graf obsahující data již tak přesný není. Bayes měl problém identifikovat data ležící pod hranicí délky 45 dm a na celém intervalu rychlosti. Úspěšnost je ovšem poměrně vysoká, dosahuje 40,563953%

6.10.2 K-means inicializace vyhodnocovatele

Hodnoty pro $k = 5$ pro denní data a $k = 6$ pro noční data, hodnoty kolísaly od 14% od 40%. Nastává otázka, zda-li existují taková k, která mají lepší výsledky, a jak se pak takové výstupy liší od původních, algoritmem určených k. Do Bayese postupně bude vstupovat inicializace k-meas s hodnotami $2, 3, 4, 7, 8, 9$.

Velocity (km/h) - Sound (dB):

Při testech různých k se ukazuje (Tabulka 6.47), že i pro větší hodnoty než 5, se může zvedat úspěšnost predikce. Největší hodnota 71% je jen pro $k = 2$, jedná příliš malý počet klastrů pro efektivní rozdělení obcí. Ovšem pro $k = 3$ je hodna vyšší než pro původních $k = 5$.

k	Úspěšnost v %
2	71.810773%
3	44.755493%
4	36.782424%
5	14,501312%
8	16.176471%
9	21.810773%

Tabulka 6.47: Naivní Bayes: Velocity (km/h) - Sound (dB): pro interval *6:00 h až 22:00 h*

length (dm) - Sound (dB):

I přes malé hodnoty k není schopný algoritmus dosáhnout lepších hodno úspěšnosti než 19,383416% Tuto hodnotu algoritmus dosáhne pouze $k = 2$ (Tabulka 6.48). Jenže jak už bylo naznačeno, ideální počet skupin podle dopravně expertního pohledu by měl být 4. takže musí být aspoň $k = 2$ klastry. V tu chvíli se pohybuje úspěšnost kolem 15,060241% pro $k = 4$ až 18,887314% pro $k = 3$. Jedná se o jediný případ, kdy větší hodnoty k dosahují nižší hodnot predikce.

k	Úspěšnost v %
2	19.383416%
3	18.887314%
4	15.060241%
5	15,616798%
7	11.658398%
8	8.4691708%
9	8.6109142%
10	10.347271%

Tabulka 6.48: Naivní Bayes: length (dm) - Sound (dB) pro interval *6:00 h až 22:00 h*

Velocity (km/h) - length (dm):

Poměrně zajímavé výsledky (Tabulka 6.49) se nachází v kapitole rychlost - délka. V tabulce je vidět, jak hodnoty k mají různou úspěšnost pro předpověď Bayesem. Pro $k = 2$ je úspěšnost až 92%, ovšem znovu, $k = 2$ je příliš malá hodnota. Ostatní výsledky, které vygeneroval Bayes pro hodnoty 3 a 4 stojí za prozkoumání.

k	Úspěšnost v %
2	92.62872%
3	59.238795%
4	54.262872%
5	40,563953%
7	48.85171%
8	51.69774%
9	52.975676%

Tabulka 6.49: Naivní Bayes: Velocity (km/h) - length (dm) pro interval *6:00 h až 22:00 h*

Pro noční data:

Stejný postup bude proveden i v nočním intervalu, zde bylo ovšem jako ideální počet klastrů k zvoleno číslo 6.

Velocity (km/h) - Sound (dB):

Na rozdíl od denního případu není v tomto případě hodna nejnižší z intervalu vybraných k (Tabulka 6.50). Ale stále jsou zde vyšší a to i pro větší k hodnoty. Největší jsou pro $k = 2$ a $k = 3$, hodnota 2 původně vybrána také algoritmem ale zavržena pro přílišnou hrubost při rozdělování obcí.

k	Úspěšnost v %
2	62.532808%
3	38.910761%
4	27.165354%
6	14,501312%
7	12.204724%
8	31.036745%
9	26.246719%

Tabulka 6.50: Naivní Bayes: Velocity (km/h) - Sound (dB) pro interval *22:00 h až 6:00 h*

length (dm) - Sound (dB): Algoritmus nevybral v případě rychlost - zvuk (Tabulka 6.51) nejhorší hodnotu, jako v případě denního intervalu. Jedná se o třetí nejlepší výsledek. Hned po $k = 2$ a $k = 3$. První hodnotu algoritmus určil, ale byla příliš hrubá pro určování charakterů obcí.

k	Úspěšnost v %
2	33.595801%
3	17.125984%
4	10.761155%
6	15,616798%
7	9.5800525%
8	9.0551181%
9	8.4645669%

Tabulka 6.51: Naivní Bayes: length (dm) - Sound (dB) pro interval *22:00 h až 6:00 h*

Velocity (km/h) - length (dm): Rychlost a délka (Tabulka 6.52) vykazuje největší předpoklady pro možnost predikce. Pokud by bylo $k = 2$, dostává se Bayes až k hranici 98%. Ani v tomto případě není vybrané $k = 6$ nejnižší, ale stále dosahuje pouhých 40%

k	Úspěšnost v %
2	98.037656%
3	93.21135%
4	55.034032%
6	40,563953%
7	30.283744%
8	30.151154%
9	28.710333%

Tabulka 6.52: Naivní Bayes: Velocity (km/h) - length (dm) pro interval *22:00 h až 6:00 h*

6.10.3 Obce k-means podle Bayese

Byly vybrány hodnoty k podle předchozí kapitoly. Volba byla taková, aby k bylo větší než 2 a zároveň aby mělo hodnotu *Úspěšnost v %* větší než 66%

Pro noční variantu:

1. Velocity (km/h) - length (dm): k=3

Den:
Velocity (km/h) - Sound (dB) pro k=3

Klastr	Názvy obcí	Počet obcí
0,2,1	Bynovec1, Decin Teplicka, Decin Ustecka, Hrobce, Labska Stran, Nebocany, Rohatce, Stare Krecany1	8
0,1,2	Bynov, Ceska Kam. Decinska, Ceska Kam. Duk. Hrdinu, Ceska Kamenice Liska, Huntirov, Libouchec	6
0,2	Bynovec2, Huntirov Nov. Oles.1, Kasna Lipa2, Krasna Lipa1, Nova Oleska2, Stare Krecany2	6

Tabulka 6.53: K-means podle Naivního Byese k=3

Tabulka 6.53 zobrazuje jaký je klastrový mix, pro $k = 3$ a pro hodnotu Naivního Bayese větší než 66%

Shrnutí metod

Každá metoda přistupovala k datům různými způsoby, podle předem dané teorie. Nicméně porovnejme výstupy podle reálných dopravní předpokladů. Skupiny budou určeny následovně:

Skupina	Obce
Skupina 1	Krásná Lípa, Staré Křečany, Nová Oleška
Skupina 2	Labská stráň, Bynovec, Rohatce,
Skupina 3	Hrobce, Nebočany, Česká Kamenice Líska, Huntířov, Libouchec
Skupina 4	Děčín Teplická, Děčín Ústecká, Česká Kamenice Děčínská, Huntířov, Česká Kamenice Dukelských Hrdinů, Bynov

Tabulka 7.1: Finální tabulka

Toto rozdělení (Tabulka 7.1) je nejkonzistentnější a nejvíce odpovídá reálné podobě komunikace a jejímu okolí.

7.0.1 K-means

V případě prvního algoritmu se shoduje pohled řešitele na skupinu obcí: Česká Kamenice Dukelských Hrdinů, Děčín Teplická, Děčín Ústecká, Huntířov. Metoda k-means nabízela nejlepší vlastnosti při testu Naivním Bayese, proto byla použita i jako validační. Hodnoty ideálního k byly určeny jako $2, 5, 6$.

7.0.2 DBSCAN

DBSCAN se protíná s pohledem vyhotovitele v obcích: Děčín Teplická, Děčín Ústecka, Česká Kamenice Dukelských Hrdinů a Libouchec, Nebočany. Ostatní obce rozdělil do samostatných skupin. Podobný výsledek se dá nalézt i v nočním intervalu, tam v něm jsou dvojice: Děčín Teplická, Děčín Ústecká a Libouchec, Nebočany. Volba Epsilon, které ovlivňovalo výsledek, byla velmi složitá. Velké Epsilon vytvořilo pouze pár shluků, které nebylo možné nějak efektivně rozdělit do skupin lokalit. Naopak příliš malé Epsilon nutilo algoritmus definovat mnoho hodnot jako šum.

7.0.3 Fuzzy c-means

Fuzzy c-means se shoduje v případech: Bynov, Česká Kamenice Dukelských Hrdinů, Děčín Teplická, Libouchec, Děčín Ústecká, Huntířov. Noční interval měl shodné lokality Libouchec, Huntířov a Děčín Teplická, Děčín Ústecká. Algoritmus měl poměrně jednoduché možnosti nastavení a výsledky nevytvářely problematické interpretace, jako u DBSCAN. Úspěšně definoval klastry, díky čemuž bylo možné dobře rozlišit skupiny obcí.

7.0.4 Spektrálního klastrování

Čtvrtá metoda se shoduje s vyhotovitelem pouze v případ Nebočany, Hrobce pro denní data a pro noční skupiny jsou to Česká Kamenice Děčínska, Děčín Teplická, Děčín Ústecká, Česká Kamenice Děčínská pak pro Libouchec, Nebočany Spektrální

klastrování bylo nejnáročnější na tvorbu výstupů, export grafů a tabulek trvala někdy i dlouhé hodiny. Zároveň výstupy nejsou extrémně odlišné od výstupů jiných metod.

7.0.5 Hierarchical

Pro denní data je výstup příliš hrubý, než aby se dal interpretovat. Skupina obsahovala téměř všechny obce. Rozdíl oproti ostatním metodám spočíval v principu, jak tvořil shluky. Nejlépe je postup patrný na grafické znázornění. Algoritmus nejdříve bral všechny datové body jako shluky a následně je spojuje do větší uskupení. Postup se opakuje, dokud není každý datový bod v jedno klastru. V noční intervalu se shoduje Děčín Teplická, Děčín Ústecka, Bynov, Česká Kamenice Děčínská a Bynovec, Rohatce

7.0.6 Naivní Bayes

Jakožto strojového učení byl Naivní Bayes nejdříve inicializován k-menas výsledky. Následně vybrán případ, kdy Naivní Bayes dosáhl úspěšnosti větší než 66% a zároveň mělo inicializovaný vzorek počet klastrů $k > 2$. Díky tomu bylo jednodušší nalézt kombinace klastrových mixů pro jednotlivé obce. Naivní Bayes se s vyhodnocovatelem shoduje na následujících uskupení obcí:

1. Děčín Teplická, Děčín Ústecka
2. Krásna Lípa, Nová Oleška, Staré Křečany
3. Dobře osvětlený úsek
4. Bynov,Česká Kamenice Dečínská, Česká Kamenice Dukelských Hrdinů, Huntířov

7.0.7 Seřazení metod podle úspěšnosti

Jedním z cílů bylo nalezení nejlepší metody, která by efektivně dokázala pracovat danými dopravními daty.

1. K-means
2. Fuzzy c-means
3. Hierarchické klastrování
4. Spektrální klastrování
5. DBSCAN

Návrh dopravních opatření dle výsledků

Za základě výstupů ze všech metod, je možné definovat obecné prvky/vlastnosti komunikace či oblasti. Hlavní dopravní parametr, rychlost vozidla a s ní spojené vysoké hlukové imise se nejčastěji vyskytují v pěti případech a jejich kombinacích.

1. Malá rozloha obce/nízká hustota zástavby
2. Dlouhý rovný úsek
3. Dobře osvětlený úsek
4. Blízké okolí vjezdu do obce
5. Široká komunikace a prostorné okraje komunikace

Pokud obec bude chtít zklidňovat nebo měřit kritická místa komunikaci, jsou typové lokace, kde je největší pravděpodobnost porušování dopravních předpisů a tudíž největší riziko nehody.

8.1 SWOT

Jednoduchá SWOT analýza popíše a reflektuje klady, zápory, příležitosti a hrozby výstupu z klastrové analýzy.

1. Silné stránky

Výstup je na základě matematických metod, které jsou variabilní, takže poskytují široký pohled na danou problematiku. Pokud jednotlivé metody naleznou společný bod, nabývá shodnost charakteru obcí vyšší váhy.

2. Slabé stránky

Výstup je obecný a proto nenabízí přímá a konkrétní fakta. Bylo by zapotřebí data měřit ještě dalšími způsoby, aby bylo možné detailně určit prvek po prvku. Jedná se spíše o pravděpodobnost vlivu prvků a oblasti na řidiče.

3. Příležitosti

Díky seznámení se s problematikou a nalezení dobrý predikčních k, je možné zkoumat tyto vlastni dál (jak vlastnosti k-means a Naivního Bayese v dopravě a hledání prvků, tak vlastnosti obcí/komunikací) a zpřesňovat či rozšiřovat škálu výstupů.

4. Hrozby

Při aplikaci výstupů dojde ke špatnému porozumění významu výsledků metody a jejího místa v dopravně-inženýrském řešení.

podpoření Příležitostí Rozvoj vybraného tématu může pomoci k lepšímu pochopení souvislostí. Je třeba se zaměřit na jednu metodu a tu dále rozvíjet, jelikož každá z použitých metod má ještě několik variací, které je možné aplikovat. Díky základnímu přehledu je vytvořen styčný prvek pro navazující práce. Klastrovací metody jsou hojně využívány při zpracování dat a proto mají své místo i v dopravě. Jako další směr může být vypracovat práci obráceně, nalézt konkrétní prvky v obcích, naměřit dopravní veličiny a vyhodnotit, zdali mají vliv.

Závěr

Klastrovací metody jsou moderní a komplexní matematicko-statistickou disciplínou, jinak známou jako strojové učení, které má i v dopravě své místo. V této práci se vyhotovitel prezentoval několik odlišných přístupů klastrové analýzy na dopravních datech, s cílem nalézt společné vlastnosti měřených lokalit.
Nejdříve bylo zapotřebí data předpřipravit, provést filtraci, naformátovat data a přidat nové sloupce. Následně proběhlo první vyhodnocení, kde se zkoumaly základní parametry, průměry, minima maxima. Jako další krok bylo zpracování dat v pěti klastrovacích metodách, jejichž výstupem byly grafy. Ty byly převedeny do .csv formátu aby bylo možné je zpracovat do čitelného výstupu. Spočítal se obsah lokalit v klastrech a vytvořil se základní výstup klastrových mixů a lokalit. Jak bylo zmíněno, byly vygenerovány i grafy, nicméně bylo jich enormní množství, takže nebylo možné je vložit přímo do práce.

Každá metoda měla výstup určitým způsobem odlišný, a proto bylo zapotřebí dopravně vzdělaného vyhotovitele, aby nejasné případy rozdělil sám. Díky tomu bylo možné nalézt skupinu obcí, která je fúzí dopravně-vzdělané osoby, metody učení s učitelem a metody učení bez učitele. Jak se očekávalo, metody bez učitele dávaly různé výstupy podle vlastního přístupu k datovým skupinám. Ovšem v mnoha případech se výsledky podobaly a při validování výsledků se jen několik výstupů dostalo nad hranici 66%. Společné prvky pro jednotlivé oblasti na základě metod strojového učení s učitelem a bez učitele byly nalezeny. Jako další rozvoj se nabízí cílené hledání prvků a jejich měření, vytváření obecné predikce na rychlost vozidle a zvukové imise.

Tento výstup je obecný, a proto umožňuje širší pohled na celou dopravní situaci ve zkoumaný oblastech. Důležité je doplnit, že tyto výstupy jsou sice obecně platné, a že se mohou brát jako vodítko při nalézání důvodů, proč v dané obci je zvýšená rychlost vozidel a nejsou dodržovány hlukové imisní normy. Nicméně stále je nutné brát v potaz jedinečný charakter a socioekonomické vztahy dané oblasti k okolí. I tyto parametry ovlivňují chování řidičů, ale už hůře se napojují na dopravní data.

Literatura

[1] MURPHY, Kevin P. Machine Learning: A Probabilistic Perspective. Cambridge, MA: The MIT Press, 2012. ISBN 978-0-262-01802-9.

[2] CHAPELLE Olivier, SCHÖLKOPF Bernhard a ZIEN Alexander. Semi-Supervised Learning. Cambridge, MA: The MIT Press, 2006. ISBN 978-0-262-03358-9.

[3] Autor: ÖZGÜN Kamer, GÜNAY Melih, BASARAN Barış Doruk a KALEMSIZ Melisa. Analysis of Public Transportation for Efficiency. In: Trends in Data Engineering Methods for Intelligent Systems, Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2020). July 2021. DOI: 10.1007/978-3-030-79357-9_63.

[4] ZHAO Pengxiang, BUCHER Dominik, MARTIN Henry a RAUBAL Martin. A Clustering-Based Framework for Understanding Individuals' Travel Mode Choice Behavior. In: Geospatial Technologies for Local and Regional Development. June 2019. DOI: 10.1007/978-3-030-14745-7_5.

[5] MOSKVICHEVA Oleg, MOSKVICHEVA, Elena a BULATOV Andrei. Clustering Methods for Determination of Optimal Locations of Container Storage and Distribution Centers. Předneseno na: TransSiberia 2020 Conference, Samara State University of Railway Transport, ul. Svobody 2V, 443066, Samara, Russia.

[6] ANDRADE Thiago C., PEREIRA Marconi A. a WANNER Elizabeth F. Development of an application using a clustering algorithm for definition of collective transportation routes and times. Předneseno na: XV Brazilian Symposium on Geoinformatics, Campos do Jordão/SP, Brazil, November 2014.

[7] AGGARWAL Charu C. Data Clustering: Algorithms and Applications. CRC Press LLC, 2013. str. 32. ISBN 1-4665-5821-0, 978-1-4665-5821-2. DOI: 10.1201/9781315373515.

[8] ALJARAH Ibrahim. Evolutionary Data Clustering: Algorithms and Applications. Springer, 2021. str. 3. ISBN 981-3341-90-4, 978-981-3341-90-6.

[9] EVERITT, Brian. Cluster Analysis. 5. vydání. Chichester: Wiley, 2011. 330 s. ISBN 978-0-470-74991-3 (váz.); 978-0-470-97780-4 (e-book); 978-0-470-97781-1 (e-book); 978-0-470-97844-3 (e-book).

[10] LANTZ, Brett. Machine Learning with R: Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R. 2. vydání. Birmingham: Packt Publishing, 2015. ISBN 978-1-78439-390-8.

[11] TAN Pang-Ning, STEINBACH Michael a KUMAR Vipin. Introduction to Data Mining, (First Edition). Addison-Wesley, 2018. ISBN 978-0-13-312890-1.

[12] DAVIES David L. a BOULDIN Don. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224-227, 1979. DOI: 10.1109/TPAMI.1979.4766909.

[13] CALIŃSKI Tadeusz a HARABASZ JA. A Dendrite Method for Cluster Analysis. Communications in Statistics, str. 1-27, January 1974. DOI: 10.1080/03610927408827101.

[14] BISHOP, Christopher M. Pattern Recognition and Machine Learning. Springer, 2006. ISBN 978-0-387-31073-2, 0-387-31073-8.

[15] BEZDEK, James C. Fuzzy C-means cluster analysis. Scholarpedia, January 2011. DOI: 10.4249/scholarpedia.2057. Zdroj: DBLP.

[16] ESTER Martin, KRIEGEL Hans-Peter, SANDER Jörg a XU Xiaowei. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon. KDD'96, str. 226-231. AAAI Press, 1996. DOI: 10.5555/3001460.3001507.

[17] von LUXBURG Ulrike. A tutorial on spectral clustering. Statistics and Computing, svazek 17, str. 395-416, 22. srpna 2007. Springer. DOI: 10.1007/s11222-007-9033-z.

[18] EVERITT Brian S., LANDAU Sabine, LEESE Morven a STAHL Daniel. Cluster Analysis. 5th Edition. John Wiley & Sons, 2011. ISBN 978-0-470-74991-3.

[19] GANTER Bernhard a WILLE Rudolf. Formal Concept Analysis: Mathematical Foundations. Springer, 1999. ISBN 978-3-540-62771-5.

[20] ROBERT Christian P. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Second Edition. Springer, 2007. ISBN 978-0-387-71598-1.

[21] GELMAN Andrew. Bayesian Data Analysis. Third Edition (with errors fixed as of 15 February 2021). Chapman & Hall/CRC, 2013. ISBN 978-1-439-84055-8.

[22] GHOSH Jayanta K., DELAMPADY Mohan a SAMANTA Tapas. An Introduction to Bayesian Analysis: Theory and Methods. Springer, 2006. ISBN 978-0-387-23027-9.

[24] NAGY Ivan a SUZDALEVA Evgenia. Algorithms and Programs of Dynamic Mixture Estimation: Unified Approach to Different Types of Components. 1. vydání. Springer International Publishing, 2017. ISBN 3-319-64671-0. DOI: 10.1007/978-3-319-64671-8.8

[25] BEZDEK, James C. Fuzzy C-means cluster analysis. Scholarpedia, January 2011. DOI: 10.4249/scholarpedia.2057. Zdroj: DBLP.

[26] MARTOLOS, Jan (Ing., Ph.D.) a BARTOŠ, Luděk (Ing., Ph.D.). Technické podmínky – TP 189 Stanovení intenzit dopravy na PK. Třetí vydání. Plzeň: EDIP s.r.o., 2012. ISBN 978-80-87394-06-9. Počet stran: 70.

[27] TECHNICKÉ PODMÍNKY – TP 219 Dopravně inženýrská data pro posuzování vlivů dopravy na životní prostředí. 1. vydání. EDIP, 2009. ISBN 978-80-87394-00-7. Schválilo: Ministerstvo dopravy. Zpracovatel: EDIP s.r.o., Pařížská 1230/1, Plzeň. Zpracovatelé: Ing. Jan Martolos, Ph.D., Ing. Luděk Bartoš, Ph.D. Počet stran: 45. Tech. redakční rada: Ing. Radek Kropelnický (Ředitelství silnic a dálnic ČR), Ing. Pavel Hudler (Ředitelství silnic a dálnic ČR), Ing. Michal Uhlík, Ph.D. (ČVUT v Praze, Fakulta stavební), Ing. Libor Ládyš (EKOLA group, spol. s r.o.), RNDr. Miloš Liberko (ENVICONSULT), Mgr. Jan Karel (ATEM), doc. RNDr. Petr Anděl, CSc. (EVERNIA s.r.o.). Zástupce koordinátora: Ing. Veronika Říhová (Ředitelství silnic a dálnic ČR).

[28] UGLICKICH, Evzenie. [Online]. [cit. 17.2.2023]. Dostupné z: ://staff.utia.cas.cz/uglickich/kontakt.html