VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ

FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY

INSTITUTE OF MANUFACTURING TECHNOLOGY

STATISTICKÁ ANALÝZA ROZSÁHLÝCH DAT Z PRŮMYSLU

STATISTICAL ANALYSIS OF BIG INDUSTRIAL DATA

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. PETR ZAMAZAL

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. RADOVAN ŠOMPLÁK, Ph.D.

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav matematiky
Student:	**Bc. Petr Zamazal**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**Ing. Radovan Šomplák, Ph.D.**
Akademický rok:	2020/21

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Statistická analýza rozsáhlých dat z průmyslu

Stručná charakteristika problematiky úkolu:

Statistické zpracování dat z průmyslu hraje klíčovou roli při zvyšování efektivity provozu a zkvalitňování služeb. Student se bude zabývat sestavením regresních modelů v kombinaci s dalšími metodami operačního výzkumu. Modely budou popisovat vazby z reálného fungování distribučních a logistických systémů. Téma diplomové práce je spojené s řešením projektu dlouhodobé mezisektorové spolupráce STEPEP (Strategic Partnership for Environmental Technologies and Energy Production) řešeného na Ústavu procesního inženýrství. Související teoretické otázky z oblasti statistiky budou konzultovány s odborníky z Ústavu matematiky.

Cíle diplomové práce:

Prohloubení znalostí vybraných oblastí matematické statistiky.
Provedení analýzy historických dat z reálného provozu.
Sestavení statistického modelu k popsání vztahů v rámci analyzovaného řetězce.
Vyhodnocení výsledků modelu a stanovení doporučení dalšího vývoje.

Seznam doporučené literatury:

ANDĚL, J. Základy matematické statistiky. Praha: MatfyzPress, 3. vyd., 2011. ISBN 978-80-7378--62-0.

WILLIAMS, H. P. Model Building in Mathematical Programming. London School of Economics, UK: Wiley, 5th ed., 432., 2013. ISBN 978-1-118-44333-0.

MELOUN, M. a MILITKÝ, J. Statistické zpracování experimentálních dat. Praha: Academia, 2. vyd. upr. rozš., 953, 2004. ISBN 80-200-1254-0.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2020/21

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt
Práce se zabývá zpracováním reálných dat svozu odpadu. Jsou v ní popsány vybrané poznatky o statistických testech, identifikaci odlehlých hodnot, korelační analýze a lineární regresi. Tyto teoretické znalosti jsou použity za pomocí programovacího jazyka Python k zpracování dat do podoby vhodné k tvorbě lineárního regresního modelu. Výsledné modely pro dobu svozu v obci popisují mezi 70 % až 85 % variability. Na základě informací získaných při zpracování dat jsou stanovena doporučení pro svozovou společnost.

Summary
This thesis deals with processing of real data regarding waste collection. It describes select parts of the fields of statistical tests, identification of outliers, correlation analysis and linear regression. This theoretical basis is applied through the programming language Python to process the data into a form suitable for creating linear models. Final models explain between 70 % and 85 % variability. Finally, the information obtained through this analysis is used to specify recommendations for the waste management company.

Klíčová slova
lineární regrese, svoz komunálního odpadu, statistické zpracování reálných dat, korelační analýza, odlehlé hodnoty

Keywords
linear regression, municipal waste collection, statistical processing of real data, correlation analysis, outliers

ZAMAZAL, P.*Statistická analýza rozsáhlých dat z průmyslu.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2021. 51 s. Vedoucí diplomové práce doc. Ing. Radovan Šomplák, Ph.D..

Prohlašuji, že tato práce je mým původním dílem, zpracoval jsem ji samostatně pod vedením Ing. Radovana Šompláka Ph.D. a s použitím zdrojů uvedeným v seznamu literatury.

Bc. Petr Zamazal

Rád bych zde poděkoval Ing. Radovanu Šomplákovi Ph.D. za odborné vedení, konzultace a trpělivost. Dále bych chtěl poděkovat Ing. Vlastimírovi Nevrlému Ph.D. a Ing. Veronice Smejkalové za konzultace a užitečné podněty k práci.

Bc. Petr Zamazal

Obsah

1. Úvod

Diplomová práce se zabývá analýzou dat a tvorbou statistických modelů z oblasti monitoringu svozu a sběru odpadů. Tento prvek zpracovatelského řetězce odpadů představuje významnou část z celkových finančních nákladů. Práce byla tvořena ve spolupráci s Technickými službami Malá Haná (TSMH) s r. o. (viz. [11]), která se aktuálně stará o nakládání s komunálními odpady pro více jak 50 obcí. Tato společnost vznikla v prosinci 2017 a je vlastněna svazkem obcí, které jsou součástí kolektivního sběru a nakládání s odpady. Po překotném začátku, kdy bylo třeba pořídit základní vozový park a naplánovat svoz tzv. na koleně, se postupně do provozu zavádějí moderní systémy. Příkladem je monitoring svozu, který obsahuje data s GPS souřadnicemi o pohybu vozidla a vážící mechanismus pro každý výsyp sběrné nádoby. Data začala být sumarizována od roku 2019. Z logiky věci si systém sběru a správy dat prochází dynamickým vývojem, kdy se postupně odstraňují chyby a doplňují potřebné funkcionality.

Jeden z dílčích cílů této práce spočívá v ověření použitelnosti dat a identifikaci chyb v databázové struktuře. Výstupem je doporučení vhodných úprav pro zkvalitnění datové sady správcům systému. Vzniklá databáze má celou řadu uplatnění. Lze zmínit např. návrh sběrné infrastruktury s vazbou na naplněnost nádob nebo identifikaci potenciálu pro zvýšení množství separovaného komunálního odpadu. Je možné vyhodnocovat vhodnost způsobu sběru – pytlový vs. nádobový sběr aj. Další důležitou oblastí je samotný svoz odpadu. Naplánovat efektivně svozové trasy s vazbou na produkci odpadu, silniční infrastrukturu, frekvenci svozu, různé typy odpadu aj. je bez sofistikovaných nástrojů velmi obtížná úloha. Smysluplná řešení vznikají u svozových firem řadu let, kdy se obvykle čerpá ze zkušenosti z terénu. V případě nově vzniklého svazku obcí, který se navíc dynamicky každoročně rozrůstá, se jedná o oblast, kde je možné ušetřit velké množství finančních zdrojů. Do oblasti optimalizace, která se touto problematikou zabývá, patří tzv. „vehical routing problém“ (VRP) nebo „arc routing problém“ (ARP), viz [9]. Aby tyto úlohy dávaly smysluplná řešení, je nezbytné je naplnit kvalitními vstupními daty. Především se jedná o informace o produkci odpadu, silniční infrastruktuře včetně doby přejezdů a v neposlední řadě informace o vlastním sběru a jeho časové náročnosti. Právě tato data lze získat z monitoringu svozových vozidel, kterým disponují TSMH s r. o. Je třeba zmínit, že podobných spolků přibývá a tendence sumarizovat provozní data je patrná napříč všemi subjekty. Hlavním výstupem této práce je dostupná data podrobit důslednému pre-processingu. Tzn. identifikovat nesmyslné záznamy a anomálie v datech, vyřešit chybějící údaje a agregovat data do potřebné podoby pro další analýzy. Následně je možné přistoupit k

tvorbě modelů popisujících potřebná vstupní data do VRP a ARP úloh. Nedílnou součástí je shromáždění dat mimo databázi, které hrají zásadní roli při tvorbě modelů. Jde o data, která zásadním způsobem ovlivňují klíčové faktory, jako např. charakter obce (velikost, rozloha atd.). Významné faktory lze identifikovat pomocí korelační analýzy. Posledním krokem už je sestavení samotného regresního modelu a otestování jeho významnosti.

V následujícím textu je nejdříve popsán teoretický aparát potřebný ke zpracování dat, kap. 2. Následuje představení použitého softwaru, který byl využit pro implementaci zvolených přístupů a realizaci zpracování dat, kap. 3. Od kap. 4 dále je text věnován zpracování konkrétních dat, která byla poskytnuta pro vypracování této diplomové práce. Tato data jsou preprocesována, analyzována a agregována do podoby vhodné pro tvorbu modelu. Následně jsou doplněna o socio-ekonomická data. Z výsledného souboru je vytvořen lineární regresní model pro predikci doby svozu odpadu jednotlivých obcí. Lineární regrese byla vybrána, protože jde o dobře prozkoumaný statistický nástroj s pevnými matematickými základy.

2. Teoretický aparát

V této kapitole bude představen použitý aparát pravděpodobnosti a matematické statistiky. Základním pojmem je náhodná veličina, která je funkcí zajišťující přechod ze základního prostoru na množinu reálných čísel přiřazením číselné hodnoty jednotlivým jevům základního prostoru. Tento přístup umožňuje aplikaci obecných postupů na různé typy problémů. V podstatě tak veškeré náhodné veličiny odpovídající náhodným jevům (tj. pozorování, která nejsou deterministická) lze rozdělit do dvou kategorií: zaprvé spojité náhodné veličiny a zadruhé diskrétní náhodné veličiny. Diskrétní náhodnou veličinou se rozumí situace, kdy její obor hodnot obsahuje nejvýše spočetně mnoho prvků. V opačném případě se mluví o spojité náhodné veličině. Oba tyto koncepty jsou však úzce spojené a obvykle jsou pro ně aplikovány různé postupy spíše z technických důvodů, ačkoliv mají analogický význam (např. sčítání u diskrétních náhodných veličin vs integrace u spojitých). Na náhodnou veličinu navazují další pojmy jako distribuční funkce náhodné veličiny, pravděpodobnostní funkce či hustota pravděpodobnosti nebo třeba číselné charakteristiky (střední hodnota, rozptyl, šikmost, špičatost, apod.). Zmíněné pojmy a mnohé další umožňují lépe poznat a uchopit zkoumaný jev a přiřadit mu tzv. rozdělení pravděpodobnosti (funkce přiřazující jednotlivým jevům pravděpodobnosti). Při zkoumání různých dějů se navíc přišlo na to, že ne každý jev je zcela unikátní, ale objevuje se mezi nimi jistá podobnost. Chování dějů podobného charakteru pak lze často zobecnit do určitých speciálních typů rozdělení pravděpodobnosti (např. normální, binomické, atd.), což ještě více usnadňuje práci s náhodnou veličinou.

2.1. Testy hypotéz

Statistickou hypotézu lze chápat jako předpoklad o parametrech rozdělení náhodné veličiny. Hypotézu je možné pomocí různých statistických testů ověřit a výsledkem je tvrzení, zda je hypotéza zamítnuta, či nikoliv. Testovaná hypotéza se označuje jako nulová hypotéza H_0. Testování hypotézy probíhá pomocí testovací statistiky, která je za účelem daného testu sestrojena. Testovací statistika je porovnána s kritickou hodnotou. Pravděpodobnost, že se zamítne nulová hypotéza, ačkoliv je platná, se nazývá hladina významnosti α. Obvykle se hladina významnosti volí 0,05 nebo 0,01 [2].

2.1.1. Test normality

Důležitost normálního rozdělení reprezentuje centrální limitní věta. Tato věta zjednodušeně říká, že za určitých podmínek průměr výběru náhodné veličiny s konečnou střední hodnotou a konečným rozptylem je sám náhodnou veličinou, jejíž rozdělení konverguje k normálnímu. Z centrální limitní věty tedy plyne, že jevy skládající se z mnoha nezávislých náhodných procesů mívají v praxi často normální rozdělení. Příkladem takových náhodných procesů jsou náhodné chyby měření. Kvůli platnosti centrální limitní věty a také kvůli tomu, že normální rozdělení náhodné veličiny je předpokladem mnoha statistických metod, je třeba normalitu dat zkoumat. V případě nesplnění předpokladu normálního rozdělení je třeba dále hledat důvod tohoto nesplnění, kterým může být metodika sběru dat, kombinace dat z více souborů, přítomnost odlehlých hodnot apod. Bez odstranění příčiny či bez vhodné transformace dat nelze některé statistické metody použít.

Pro kontrolu předpokladu normality lze použít grafické či výpočetní metody.
Nejjednodušší grafickou metodou je vytvoření histogramu, u něhož je jeho tvar porovnáván s grafem normálního rozdělení. Přesnějšími variantami jsou Q-Q graf a P-P graf. V případě Q-Q grafu se na jednu osu vynáší kvantily hypotetického normálního rozdělení a na druhou kvantily zkoumaného souboru. U P-P grafu se pracuje s kumulativní hodnotou rozdělení. V obou případech by body grafu měly ležet na přímce.

Pro výpočetní kontrolu normality existuje řada testů vycházejících z různých vlastností normálního rozdělení. Tyto testy jsou obvykle citlivější než diagnostické grafy, zejména pro větší soubory dat. Při zamítnutí normality je nutné data hlouběji analyzovat, odchylka od normality může být způsobena např. odlehlými hodnotami.

Testy na základě šikmosti a špičatosti

Testy normality na základě šikmosti a špičatosti vycházejí z faktu, že normální rozdělení má nulovou šikmost i špičatost. Při testování normality jsou tedy testovány nulové hypotézy, že šikmost a špičatost jsou rovny nule.

Další možností je test kombinující výběrovou šikmost a špičatost. Jde o test normality založený na testovacím kritériu definovaném jako

$$C_1 = \frac{\hat{g}_1^2}{\mathrm{D}(\hat{g}_1)} + \frac{[\hat{g}_2 - \mathrm{E}(\hat{g}_2)]^2}{\mathrm{D}(\hat{g}_2)}, \tag{2.1}$$

kde $\hat{g}_1$ a $\mathrm{D}(\hat{g}_1)$ značí výběrovou šikmost a její rozptyl a $\hat{g}_2$, $\mathrm{E}(\hat{g}_2)$ a $\mathrm{D}(\hat{g}_2)$ značí výběrovou špičatost, její střední hodnotu a její rozptyl. Vzorce pro výpočet těchto veličin lze najít

v [2]. Pro nulovou hypotézu H_0 o normalitě rozdělení výběru má kritérium C_1 asymptoticky rozdělení χ^2 s dvěma stupni volnosti. Pokud tedy platí $C_1 > \chi^2_{1-\alpha}(2)$, hypotézu H_0 zamítáme.

Andersonův-Darlingův test

Andersonův-Darlingův test ověřuje normální rozdělení pomocí analýzy empirické distribuční funkce. Testována je hypotéza H_0 o rovnosti distribučních funkcí výběru a teoretického rozdělení. Testovací kritérium pro ověření normality výběru je následující

$$AD = -\frac{\sum_{i=1}^{n}(2i-1)(lnz_i + ln(1-z_{n-i+1}))}{n} - n, \tag{2.2}$$

kde z_i jsou hodnoty distribuční funkce standardizovaného normálního rozdělení. Nulová hypotéza o normalitě je zamítnuta, pokud $AD > D_{1-\alpha}$. Kritické hodnoty lze najít například v [6].

Shapirův-Wilkův test

Shapirův-Wilkův test se doporučuje zejména pro malé výběry $3 \leq n \leq 50$. Testovací statistika se používá ve tvaru

$$SW = -\frac{[\sum_{i=1}^{n} a_i x_i]^2}{\sum_{i=1}^{n}(x_i - x)}, \tag{2.3}$$

hodnoty koeficientů a_i a jejich odvození lze nalézt v literatuře [5]. Hodnota statistiky SW je posouzena s kritickou hodnotou $W^*_{1-\alpha}$. Pokud platí $W < W^*_{1-\alpha}$, hypotéza o normálním rozdělení výběru se zamítá.

Z dalších testů normality dat lze zmínit testy dobré shody, které pracují na principu porovnání distribuční funkce sledované náhodné veličiny s distribuční funkcí normovaného normálního rozdělení. Dále Kolmogorovův-Smirnovův test, který posuzuje distribuční funkce empirického a teoretického rozdělení, které nemusí být nutně normální.

2.1.2. Dvouvýběrové testy hypotéz

Častou úlohou v praxi je porovnávání dvou výběrů $\{x_i\}, i = 1, ..., n_1$, a $\{y_j\}, j = 1, ..., n_2$. Předpokladem pro zde uvedené testy jsou podmínky:

1. Výběry $\{x_i\}$ a $\{y_j\}$ jsou vzájemně nezávislé
2. Oba výběry patří do normálního rozdělení $x_i \sim N(\mu_x, \sigma_x^2)$ a $y_j \sim N(\mu_y, \sigma_y^2)$

Informace o dvouvýběrových testech jsou čerpány z [2].

Studentův t-test pro shodnost středních hodnot

Studentův t-test je určený pro testování hypotézy $H_0 : \mu_x = \mu_y$ oproti alternativní hypotéze $H_A : \mu_x \neq \mu_y$. Tento test se dělí na varianty podle splnění rovnosti rozptylů.

1. Pro případ $\sigma_x^2 = \sigma_y^2$ používáme testovací kritérium

$$T_1 = \frac{|\bar{x} - \bar{y}|}{\sqrt{(n_1 - 1)\, s_x^2 + (n_2 - 1)\, s_y^2}} \sqrt{\frac{n_1 n_2 \,(n_1 + n_2 - 2)}{n_1 + n_2}}. \tag{2.4}$$

A použijeme pro ni Studentovo rozdělení s $\nu = n_1 + n_2 - 2$ stupni volnosti. Hypotézu H_0 zamítáme na hladině významnosti α, pokud platí $T_1 > t_{1-\alpha/2}(\nu)$.

2. V případě nerovnosti $\sigma_x^2 \neq \sigma_y^2$ používáme kritérium ve tvaru

$$T_2 = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}}. \tag{2.5}$$

V tomto případě má Studentova statistika stupně volnosti

$$\nu = \frac{\left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2}{\frac{s_x^4}{n_1^2(n_1-1)} + \frac{s_y^4}{n_2^2(n_2-1)}} \tag{2.6}$$

Pokud platí $T_2 > t_{1-\alpha/2}(\nu)$, hypotézu H_0 zamítáme na hladině významnosti α.

3. Pokud se oba výběry odchylují od normálního rozdělení, je možné použít modifikovené kritérium

$$T_3 = \frac{|\bar{x} - \bar{y}| + C + D\,(\bar{x} - \bar{y})^2}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}}, \tag{2.7}$$

kde

$$C = \frac{1}{6} \frac{\frac{\hat{g}_{1x}}{n_1^2} \frac{s_x^3}{\sqrt{n_1}} - \frac{\hat{g}_{1y}}{n_2^2} \frac{s_y^3}{\sqrt{n_2}}}{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}, \tag{2.8}$$

$$D = \frac{1}{3} \frac{\frac{\hat{g}_{1x}}{n_1^2} \frac{s_x^3}{\sqrt{n_1}} - \frac{\hat{g}_{1y}}{n_2^2} \frac{s_y^3}{\sqrt{n_2}}}{\left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2}, \tag{2.9}$$

kde $\hat{g}_{1x}$ a $\hat{g}_{1y}$ značí výběrové šikmosti. Kritérium T_3 má za předpokladu hypotézy H_0 Studentovo rozdělení se stupni volnosti odpovídajícími vztahu (2.6). Test pomocí kritéria T_3 je robustní vůči sešikmení výběrového rozdělení i vůči heteroskedasticitě. Není u něj třeba

ani požadavek na shodnost rozptylů. Vůči odchýlení od normálního rozdělení ve špičatosti jsou robustní všechna tři kritéria T_1, T_2 i T_3. V případě normality výběrů rozhodujeme mezi kritérii T_1 a T_2 pomocí F-testu, pokud normalita splněna není, je třeba použít kritéria T_2, nebo T_3, protože F-test je na normalitu výběrů citlivý a nelze ho použít.

F-test pro shodu rozptylů

F-test slouží k ověření hypotézy $H_0 : \sigma_x^2 = \sigma_y^2$ oproti alternativě $H_A : \sigma_x^2 \neq \sigma_y^2$. K jeho ověření používáme testovací kritérium ve tvaru

$$F = max\left(\frac{s_x^2}{s_y^2}, \frac{s_y^2}{s_x^2}\right) \tag{2.10}$$

Platí-li $s_x^2 > s_y^2$, používáme F-rozdělení s $\nu_1 = n_1 + 1$ a $\nu_2 = n_2 + 1$ stupni volnosti. Pořadí stupňů volnosti závisí na nerovnosti rozptylů. Tento test je velmi citlivý na normalitu rozdělení. V případě špičatosti odlišné od normálního rozdělení je možné použít alternativní stupně volnosti

$$\nu_1 = \frac{n_1 - 1}{1 + \frac{\hat{g}_{2c}}{2}}, \tag{2.11}$$

$$\nu_2 = \frac{n_2 - 1}{1 + \frac{\hat{g}_{2c}}{2}}, \tag{2.12}$$

kde

$$\hat{g}_{2c} = \frac{2\left(n_1 + n_2\right)\left(\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^4 + \sum_{i=1}^{n_2}\left(y_i - \bar{y}\right)^4\right)}{\left(\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^2 + \sum_{i=1}^{n_2}\left(y_i - \bar{y}\right)^2\right)} - 3 \tag{2.13}$$

2.1.3. Chyby měření

Chybou měření se rozumí rozdíl mezi skutečnou a naměřenou hodnotou, informace o skutečné hodnotě však není k dispozici. Aby bylo možné určit přesnost naměřených hodnot, je nutné odhadnout chybu měření pomocí statistických metod nebo analýzou metodiky měření a měřícího náčiní. Chyby měření jsou buď náhodné, nebo systematické. Náhodná chyba je nepřesnost měření způsobená vlivy, které nepříliš ovlivňují měřené hodnoty. Těchto vlivů je velké množství. Sumu těchto jevů lze aproximovat normálním rozdělením. Pokud měření neodpovídají normálnímu rozdělení, má alespoň jeden z vlivů na hodnoty dat příliš významný dopad. V takovém případě se jedná o systematickou chybu a je vhodné ji studovat a zjistit její přičinu. V závislosti na přičině je možné se se systematickou chybou vypořádat buď změnou měření, nebo jejím zohledněním při nastavení parametrů měření. Například pokud se prokáže, že je měřená hodnota významně závislá na teplotě

prostředí, je vhodné provádět měření za stálé teploty. Pokud není možné provádět opakování měření, je nutné systematickou chybu identifikovat a taková data odstranit. Další možností je zahrnutí této chyby do modelu, pokud je soubor dat dostatečně velký a poskytuje potřebné informace o chybě. Je potřeba uvážit, že se měří rychlost produkce odpadu namísto samotného produkovaného množství odpadu. S časem se bude pracovat tak, že se uváží potenciální závislost dat na čase. Chyby v datech o množství odpadu mohou vznikat z důvodu chyb měření na základě různých meřících přístrojů. Dále se v datech očekává výskyt chyb, pro které nejsou k dispozici detailnější informace (např. vliv tajné skládky). V praxi není možné všechny vlivy vyloučit opakovaným měřením za stejných podmínek. Je nutné sestavit obecný model, pro jehož vytvoření jsou zapotřebí měření za různých podmínek, přičemž jsou zaznamenány údaje i o těchto podmínkách. Problém nastává při přítomnosti neznámého jevu v datech. Jeho přítomnost ztěžuje zkoumání známých jevů a sestavení modelu pro ně. Bez vizualizace či použití vhodných robustních metod může tato situace vést k vytvoření chybných závěrů.

2.2. Odlehlé hodnoty

Odlehlé hodnoty jsou takové, které mají vzhledem k odhadnutému rozdělení souboru příliš nízkou pravděpodobnost výskytu, a jsou tedy pravděpodobně projevem chyb v měření. Případně se může jednat o měření, která jsou z jiného statistického souboru, než je zbytek studovaných dat. V určitých případech může jít i o data správná, ale o chybnou úvahu, například chybně předpokládané rozdělení. Odlehlé hodnoty je tedy třeba identifikovat a rozhodnout o jejich příčině a následném vypořádání se s nimi. Pro identifikaci odlehlých hodnot v jednorozměrném statistickém souboru jsou nejznámějšími metodami *Grubbsův test* a *Dean-Dixonův* test.

2.2.1. Grubssův test

Předpokladem Grubbsova testu je, že náhodný výběr $X_1, ..., X_n$ je z normálního rozdělení. Tento test identifikuje vždy maximálně jednu odlehlou hodnotu. Lze ho použít rekurzivně vícekrát, vždy po odstranění předchozí odlehlé hodnoty. Vzhledem k tomu, že se v každé iteraci přepočítává kritérium, je třeba dávat pozor na malý počet vzorků. Může se totiž stát, že test přestává fungovat a je schopen označit všechny hodnoty za odlehlé. Jako hranici, pod kterou by se nemělo dostat, se uvádí 6 vzorků. Grubssův test je definován

pro hypotézu H_0 : Ve výběru není žádná odlehlá hodnota. Statistika je pro něj definována jako

$$G = \frac{\underset{i=1,\dots,n}{max} \left|X_i - \bar{X}\right|}{s_X}, \tag{2.14}$$

kde $\bar{X}$ je výběrový průměr a s_X je výběrová směrodatná odchylka. Hypotézu o neexistenci odlehlé hodnoty zamítáme na hladině významnosti α v případě, že platí

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n)}(n-2)}{n-2+t^2_{\alpha/(2n)}(n-2)}}, \tag{2.15}$$

kde $t^2_{\alpha/(2n)}(n-2)$ je kvantil Studentova rozdělení s $n-2$ stupni volnosti.

2.2.2. Dean-Dixonův test

Dean-Dixonův test stejně jako Grubssův test vyžaduje podmínku normality výběru. Dean--Dixonův test se doporučuje používat v omezené míře, někdy dokonce pouze pro identifikaci nejvýše jedné odlehlé hodnoty z výběru. Je vhodný hlavně pro výběry s malým množstvím prvků. Pro potřeby tohoto testu jsou seřazeny hodnoty výběru podle velikosti $X_1X_2...X_{n-1}X_n$. Test je definován pro hypotézu H_0 : Zkoumaná hodnota není odlehlá. V tu chvíli se vypočte statistika pro ověření, zda je největší hodnota X_n odlehlá, následovně

$$Q_{max} = \frac{X_n - X_{n-1}}{X_n - X_1}. \tag{2.16}$$

Obdobným způsobem se vypočte statistika pro nejmenší hodnotu

$$Q_{min} = \frac{X_2 - X_1}{X_n - X_1}. \tag{2.17}$$

Pokud statistika překročí kritickou hodnotu Q závisící na počtu vzorků n a hladině významnosti α, zamítáme nulovou hypotézu. Kritické hodnoty Q jsou počítány numerickými metodami a jsou tabelovány.

2.2.3. Robustní přístup

Rizikem klasických metod pro hledání odlehlých hodnot je, že jsou samy ovlivněny odlehlými hodnotami. To může vést až k tomu, že metody odlehlé hodnoty neodhalí, nebo dokonce označí za odlehlé ty hodnoty, které jsou v pořádku. Dalším důvodem pro po-

užívání robustních metod je odolnost vůči odchylkám od normality. V [7] je navrhnuto robustní kritérium

$$Q = \frac{\left(X_i - \underset{j=1,\dots,n}{median}\,(X_j)\right)}{k \cdot \underset{i=1,\dots,n}{median}\left|X_i - \underset{j=1,\dots,n}{median}\,(X_j)\right|}, \tag{2.18}$$

kde k je korekční konstanta, která je pro normální rozdělení rovna $1,483$. Za kritickou hodnotu Q je ve zdroji zvolena hodnota $\pm 2,5$. Z rovnice je patrné, že kritickou hodnotu Q a konstantu k by bylo možné substitučně nahradit jednou konstantou. Tedy že kritickou hodnotu pro Q je třeba měnit i podle rozdělení. Přesná hodnota by však neměla mít zásadní vliv, protože Q pro odlehlé hodnoty je obvykle výrazně vyšší.

2.3. Korelace

Mějme náhodné veličiny X a Y s konečnými druhými momenty. Pro měření lineární závislosti těchto veličin se používá *Pearsonova korelačního koeficientu*. Předpokládejme, že platí $(varX) > 0$ a $(varY) > 0$, potom je korelační koeficient ve tvaru

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{varX varY}}. \tag{2.19}$$

Platí, že $-1 \leq \rho_{X,Y} \leq 1$. Pokud jsou X a Y lineárně nezávislé, tak $\rho_{X,Y} = 0$.

2.3.1. Výběrový korelační koeficient

Pokud se pracuje s náhodným výběrem $(X_1, Y_1)', (X_2, Y_2)', ..., (X_1, Y_1)'$, tak se zavádí *výběrový korelační koeficient* za předpokladu

$$S_X^2 > 0, S_Y^2 > 0$$

jako

$$r = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}, \tag{2.20}$$

kde S_{XY} je výběrová kovariace a S_X^2 a S_Y^2 jsou výběrové směrodatné odchylky. Pro testování hypotézy o nulovosti Pearsonova korelačního koeficientu veličin X a Y se používá kritérium

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}, \tag{2.21}$$

které má Studentovo rozdělení s $n-2$ stupni volnosti. Předpokladem pro použití tohoto kritéria je ale, že se pracuje s výběrem z dvojrozměrného normálního rozdělení.

2.3.2. Spearmanův korelační koeficient

Alternativou k Pearsonovu koleračnímu koeficientu je *Spearmanův korelační koeficient*, který je robustní vůči odlehlým hodnotám a obecně odchylkám od normality, jelikož je založen na pořadí hodnot. Na rozdíl od Pearsonova koeficientu nepopisuje lineární závislost X a Y, ale popisuje, jak moc jejich vztah odpovídá monotónní funkci.
Předpokládejme, že náhodný výběr $(X_1, Y_1)', ..., (X_1, Y_1)'$ je ze spojitého rozdělení. Označme pořadí veličin $X_1, ..., X_n$ jako $r_X 1, ..., r_X n$ a pořadí veličin $Y_1, ..., Y_n$ jako $r_Y 1, ..., r_Y n$, potom Spearmanův korelační koeficient má tvar

$$r_S = \frac{\sum_{i=1}^{n} r_{Xi} r_{Yi} - n\bar{r}_X \bar{r}_Y}{(n-1) s_{rX} s_{rY}}, \tag{2.22}$$

kde $\bar{r}_X$ a $\bar{r}_Y$ jsou průměrná pořadí a s_{rX} a s_{rY} jsou směrodatné odchylky pořadí. Pro odhad r_S se používá často i tohoto vzorce

$$r_S = 1 - \frac{6}{n\left(n^2 - 1\right)} \sum_{i=1}^{n} \left(r_{Xi} - r_{Yi}\right)^2. \tag{2.23}$$

Druhý uvedený vzorec je citlivý na opakující se hodnoty, není tedy vhodný vždy.

2.4. Regrese

2.4.1. Lineární model

V této kapitole jsem čerpal hlavně z [4] doplněné o poznatky z [8, 2] Mějme nekorelované veličiny $Y_1, ..., Y_n$, u nichž předpokládáme, že jejich střední hodnoty jsou lineární funkcí $k + 1$ neznámých parametrů

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik},\ pro\ i = 1 \dots n, \tag{2.24}$$

kde x_{ij} jsou známé konstanty. Dále pro všechna i předpokládáme, že $varY_i = c$. σ je zde neznámým parametrem. Známé konstanty x_{ij} se uspořádají do matice o n řádcích a $k+1$ sloupcích

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \tag{2.25}$$

jejíž hodnost $h(\mathbf{X}) = r$ a platí $n > r$. Náhodný vektor $\mathbf{Y}$ má potom střední hodnotu $\mathbf{X}\boldsymbol{\beta}$ a variační matici $\sigma^2\mathbf{I}$. To znamená, že střední hodnota je prvkem lineárního obalu matice

$\mathbf{X}$ a jednotlivé složky vektoru $\mathbf{Y}$ mají stejný rozptyl a jsou nekorelované. Pro uvedené předpoklady se používá zápis $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. V případě $k = 1$ se jedná o *jednoduchou* lineární regresi a pro $k > 1$ o regresi *mnohonásobnou*. Pokud $\mathbf{Y_1}$ je vektorová veličina na místo skaláru Y_1, jedná se o regresi *mnohorozměrnou*. Zaveďme nyní matici $\mathbf{Q}$, jejíž sloupce tvoří ortonormální bázi *regresního prostoru* $span(\mathbf{X})$. Dále zaveďme matici $\mathbf{N}$, jejíž sloupce doplňují $\mathbf{Q}$ na ortonormální bázi prostoru $\mathbb{R}^n$. Tím vzniká ortonormální matice $\mathbf{P} = (\mathbf{Q}, \mathbf{N})$, pro kterou platí, že $span(\mathbf{X}) = span(\mathbf{Q})$ a $\mathbf{PP'} = \mathbf{P'P} = \mathbf{I}_n$. Z ortonormality sloupců matice $\mathbf{P}$ plynou tyto vztahy

$$\mathbf{QQ'} + \mathbf{NN'} = \mathbf{I}_n \quad \mathbf{Q'Q} = \mathbf{I}_n \quad \mathbf{N'N} = \mathbf{I}_{n-r}.$$

Označme $\mathbf{H} = \mathbf{QQ'}$ a $\mathbf{M} = \mathbf{NN'}$. Tyto matice jsou symetrické a idempotentní. Z výše uvedených vztahů vyplývá i vztah

$$\mathbf{y} = \mathbf{Hy} + \mathbf{My}.$$

Jelikož platí $\mathbf{H} + \mathbf{M} = \mathbf{0}$, jsou vektory na pravé straně uvedeného vztahu navzájem ortogonální. Jde tedy o průměty obecného vektoru $y \in \mathbb{R}^n$ do regresního prostoru $span(\mathbf{X})$ a *reziduálního prostoru* $span(\mathbf{X})^\perp$. Z vlastností projekce plyne jednoznačnost těchto průmětů, a tedy i matic $\mathbf{H}$ a $\mathbf{M}$.

2.4.2. Odhad vektoru středních hodnot

Nejprve bude pozornost věnována odhadu vektoru středních hodnot $\mu = \mathbf{X}\boldsymbol{\beta}$. V prostoru $span(\mathbf{X})$ se najde nejbližší vektor k náhodnému vektoru $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ a označí se $\hat{\mathbf{Y}}$.

Věta 1 (Gaussova-Markovova) *V modelu* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ *je* $\hat{\mathbf{Y}}$ *nejlepším nestranným lineárním odhadem vektoru* $\mathbf{X}\boldsymbol{\beta}$.

Důkaz je uveden v [4].

2.4.3. Rezidua

Nyní bude pozornost zaměřena na průmět vektoru $\mathbf{Y} \sim \mathbf{X}\boldsymbol{\beta}$ do prostoru reziduí $span(\mathbf{X})^\perp$ a zavede se nestranný odhad rozptylu σ^2. Vektor $\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}}$ se označí jako *vektor reziduí*. Dále se definuje *reziduální součet čtverců* jako $RSS = \|u^2\| = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$, který udává kvadrát vzdálenosti vektorů $\mathbf{Y}$ a $\hat{\mathbf{Y}}$, popisuje tedy jedním číslem jejich rozdílnost. Dále je ještě nutné zavést *reziduální rozptyl*, jako $S^2 = RSS/(n - r)$.

Věta 2 (O reziduích) *V lineárním modelu* $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ *platí*

$$\mathbf{u} = \mathbf{M}\mathbf{Y} = \mathbf{M}\mathbf{e}, \tag{2.26}$$

$$\mathbf{u} \sim (\mathbf{0}, \sigma^2\mathbf{M}), \tag{2.27}$$

$$RSS = \mathbf{e}'\mathbf{M}\mathbf{e}, \tag{2.28}$$

$$\mathrm{E}RSS = (n-r)\sigma^2, \tag{2.29}$$

$$\mathrm{E}S^2 = \sigma^2, \tag{2.30}$$

$$\mathbf{X}'\mathbf{u} = \mathbf{0}. \tag{2.31}$$

Důkaz uvedené věty lze nalézt v [4] Vektor $\mathbf{u}$ lze chápat jako odhad náhodné složky modelu $\mathbf{e} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. Z uvedených vztahů plyne také, že S^2 je nestranným odhadem rozptylu σ^2.

2.4.4. Normální rovnice

V této části textu bude tvořen odhad vektoru $\boldsymbol{\beta}$ vyjadřujícího střední hodnotu náhodného vektoru $\mathbf{Y}$ jako lineární kombinaci sloupců matice $\mathbf{X}$. Předpokládá se lineární nezávislost sloupců matice $\mathbf{X}$. Jako $\mathbf{b}$ se označí řešení soustavy

$$\mathbf{X}\mathbf{b} = \widehat{\mathbf{Y}}. \tag{2.32}$$

Vektor $\mathbf{b}$ zde potom představuje hledané koeficienty lineární kombinace. Skutečnost, že

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{u}$$

je ortogonálním rozkladem, je ekvivalentní s požadavkem na ortogonalitu vektoru reziduí $\mathbf{u}$ vůči regresnímu prostoru $span(\mathbf{X})^{\perp}$. Tento požadavek lze zapsat jako

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{0},$$

což je dále ekvivalentní s takzvanou *normální rovnicí* pro $\mathbf{b}$

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{2.33}$$

Tato rovnice je díky přítomnosti lineární kombinace řádků matice $\mathbf{X}$ na obou stranách rovnice (2.33) řešitelná. Jednoznačnost však zaručena není.

2.4.5. Koeficient determinace

Koeficient determinace R^2 se definuje jako

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

Popisuje, jak velkou část výchozí variability závisle proměnné se podařilo vysvětlit. Čím bližší je R^2 jedné, tím je tato část větší.

2.4.6. Korigovaný koeficient determinace

Hodnota R^2 roste s počtem regresorů, což zdánlivě zvyšuje přesnost modelu. Proto se zavádí *korigovaný koeficient determinace*, který je definovaný jako

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2),$$

kde k je počet sloupců *regresní matice*. Podobně jako pro R^2 platí, že čím bližší je $\bar{R}^2$ k hodnotě jedna, tím těsnější je regresní závislost.

2.4.7. Inflační faktor

Inflační faktor, zvaný také VIF (z anglického Variance Inflation Factor), je definován jako

$$VIF_j = \frac{1}{1-R_j^2}, \tag{2.34}$$

kde R_j^2 je koeficient determinace regresní závislosti j-tého sloupce matice $\mathbf{X}$ na ostatních jejích sloupcích. Popisuje tedy, kolik variability tohoto regresoru je vysvětleno regresory ostatními. VIF se proto používá pro zkoumání multikolinearity. Nabývá hodnot od 1, kdy je daný regresor nezávislý na ostatních, až do libovolné hodnoty. V [8] je však uvedena hraniční hodnota 5, nad kterou jde již o problematickou míru multikolinearity.

3. Použitý software

Zpracovávaná data jsou rozsáhlá a obsahují mnoho dílčích kategorií, je tedy nutné data zpracovávat a vyhodnocovat strojově. Znamená to omezené používání grafických metod a nutnost zaměření se na robustní číselné metody. Velikost úlohy také vznáší nároky na použité softwarové nástroje, aby bylo možné výpočty provádět s přijatelnými požadavky na výpočetní čas.

Pro analýzu, zpracování a vizualizaci dat v této práci se využívá programovací jazyk Python ve verzi 3.8. (Dokumentaci k němu je možno nalézt v [12].) Python je interpretovaný programovací jazyk umožňující více programovacích paradigmat, jako objektové, procedurální a funkcionální. Programovací jazyk Python je zvolen hned z několika důvodů. Jedná se o jazyk, který sice vznikl již v roce 1991, ale v dnešní době je velmi oblíbený a v praxi velmi používaný. Z tohoto důvodu má silnou uživatelskou základu a kvalitní dokumentaci. Dalším důvodem je, že jde o jazyk *open source*, tedy je spolu s většinou knihoven dostupný zdarma. Podstatnou výhodou jazyka Python je také to, že jsou k dispozici knihovny optimalizované na práci s maticemi a kvalitní statistické knihovny. Není tedy nutné většinu operací programovat od základu.

Analýza dat v jazyce Python je v této práci realizována v prostředí PyCharm Community Edition. Jde o freeware variantu prostředí vyvíjeného českou firmou JetBrains. Pro práci s daty je využíván také Microsoft Excel, data jsou ukládána s příponou .xlsx. Výhodou využití kombinace jazyka Python a editoru Microsoft Excel je jejich snadná vzájemná komunikace a možnost rychlého náhledu do dat. Pro práci s datovými tabulkami byla použita knihovna Pandas, která disponuje optimalizovanými nástroji pro práci s rozsáhlými daty. Tato knihovna byla nezbytná, protože i s jejím použitím trvaly některé operace na surových datech (jde o datový soubor s 358 379 záznamy) v řádu až desítek minut. V případě práce s ještě rozsáhlejšími daty by bylo vhodné převést data do databázové podoby.

Využité knihovny programovacího jazyka Python:

- Pandas – umožňuje práci s dvourozměrnými maticemi a časovými řadami. Při práci s maticemi se operuje s pojmenovanými řádky a sloupci. Knihovna Pandas je velmi dobře optimalizovaná a umožňuje rychle provádět operace s rozsáhlými

daty. Také je provázaná s dalšími knihovnami, takže jsou vzájemně kompatibilní. Pomocí knihovny Pandas lze provádět také nahrávání a ukládání .xlsx souborů.

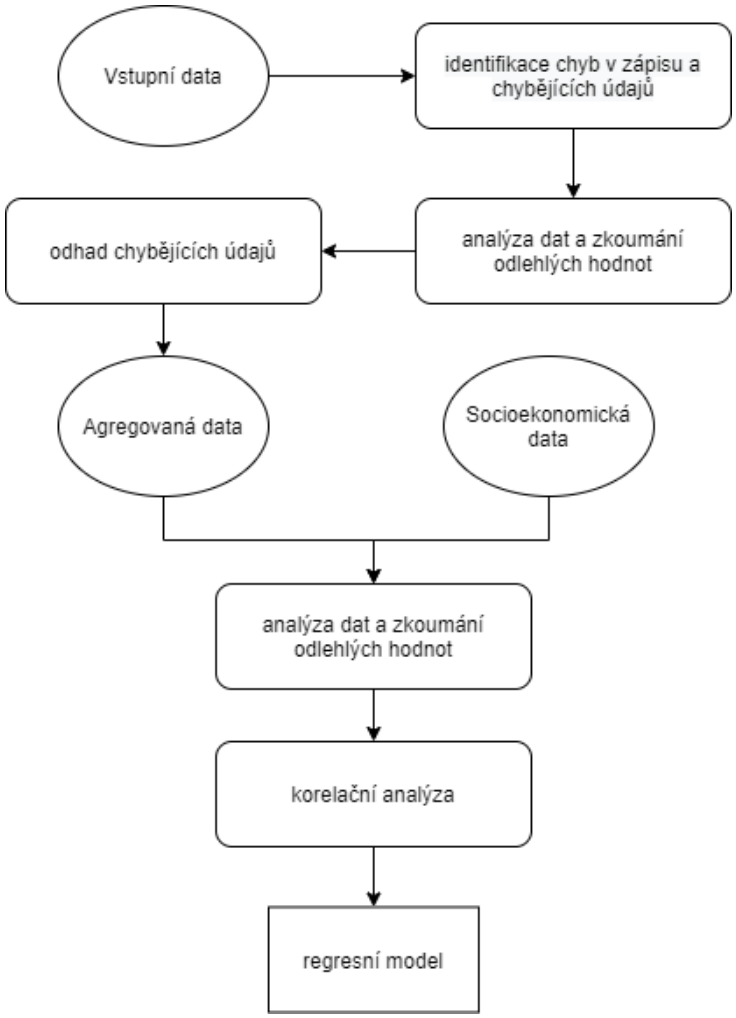
- NumPy – umožňuje také práci s vícerozměrnými maticemi. Tato knihovna obsahuje mnoho metod pro maticové operace, či optimalizované maticové varianty běžných funkcí. Knihovna NumPy využívá pro vícerozměrné matice operace pro menší úlohy zahrnuté v již zmíněné knihovně Pandas.
- Matplotlib – je knihovna sloužící k vykreslování grafických výstupů.
- math – obsahuje základní matematické funkce a konstanty.
- SciPy – zahrnuje mnoho dílčích modulů pro různé vědní obory. Pro tuto práci je zásadní statistický modlu, který umožňuje mimo jiné provádění statistických testů.
- datetime – slouží pro práci s daty a časem. Knihovna datetime umožňuje provádět výpočty na časových datech a pracovat s různými formáty času. Knihovna Pandas obsahuje její variantu optimalizovanou pro rozsáhlá maticová data.
- statsmodels – obsahuje nástroje pro tvorbu statistických modelů a jejich pohodlné vyhodnocování.

4. Analýza dat společnosti Technické služby Malá Haná

Společnost TSMH v roce 2019 opatřila svozová auta monitorovacím systémem. Díky tomu je možné detailně evidovat pohyb vozidel a množství sváženého odpadu z jednotlivých lokalit. Data takto získaná ze svozových aut byla v této studii analyzována za účelem dalšího využití při optimalizaci svozových plánů a vozového parku.

4.1. Popis dat

Pro řešení této diplomové práce poskytla společnost TSMH s r. o. data o svozu komunálních odpadů ze 46 obcí v letech 2019, 2020 a také data ze začátku roku 2021. Jedná se převážně o obce z okresů Blansko a Svitavy. Data obsahují celkem 358 379 záznamů o zastávkách u jednotlivých popelnic. Cílem zpracování těchto dat je vytvořit model pro dobu svážení v jednotlivých obcích, což umožní dále optimalizovat svozové trasy. Původní data je tedy nutné agregovat. Obdržená data jsou zaznamenána bez jakéhokoliv statistického zpracování. Vzhledem k tomu, že evidence dat probíhala po dobu dvou let více různými pracovníky, lze očekávat podstatné riziko výskytu chybových hodnot. V surových datech navíc část sledovaných údajů není uvedena u všech záznamů. Z těchto důvodů je tedy třeba před samotnou tvorbou modelu provést průzkumovou a statistickou analýzu jak agregovaných, tak neagregovaných dat. V schématu na obrázku 4.1 je struktura postupu zpracování dat.



Obrázek 4.1: Schéma postupu zpracování dat

Data poskytnutá společností TSMH s r. o. obsahují záznamy s následujícími atributy o svozu komunálních odpadů:

- ID - Identifikační číslo přiřazené jednotlivým záznamům.
- DATUM - Datum a čas svozu na dané zastávce.
  Uváděn je pouze jeden časový údaj. Není tedy zřejmé, jestli se jedná o čas odjezdu, či příjezdu na zastávku. Z tohoto důvodu není možné pracovat odděleně s dobou cest mezi zastávkami a s dobou manipulace s odpadem.
- NADOBA - Typ svážené nádoby.
  Označení začíná zkráceným názvem typu odpadu, na který je nádoba určena, pokračuje objemem nádoby a končí barvou. Některá sběrná hnízda jsou místo objemu a barvy nádoby označena jako sběrné hnízdo, tzn. že na tomto místě je více druhů komunálních odpadů.

- ODPAD - Typ sváženého odpadu (biologicky rozložitelný odpad, papír, sklo, …).
- NAZEV STANOVISTE – Adresa stanoviště, nebo název obce a text „sběrné hnízdo“. Všechna sběrná hnízda v dané obci mají stejný název, tedy názvy nejsou unikátní.
- ULICE - Ulice umístění stanoviště.
  Jedná se o zdvojení informace oproti „NAZEV STANOVISTE“. V případě sběrných hnízd je hodnota vynechána.
- EVIDENCNI CISLO - Obdoba čísla popisného v obcích s málo adresami.
- CISLO POPISNE - Číslo popisné umístění stanoviště.
  Jedná se o zdvojení informace oproti „NAZEV STANOVISTE“. V případě sběrných hnízd je uvedena hodnota „Sběrné hnízdo“.
- MNOZSTVI - Množství odpadu nabraného při této zastávce v kilogramech.
- OBEC – Název obce umístění stanoviště.
- ID NADOBY - Jedinečné číslo přiřazené svážené nádobě.
  Hodnota je ve formátu pětimístného čísla.
- LAT, LON - Zeměpisná šířka a délka polohy stanoviště.
  Z důvodu nedostatečné přesnosti nejsou tyto hodnoty dále využívány.
- SPZ - Státní poznávací značka vozidla.

Při zpracování dat je důležité rozlišovat stanoviště, nádoby a zastávky. Stanoviště je dané adresou, či odpovídá sběrnému místu. Na jednom stanovišti se však může nacházet více nádob, a to jak pro jeden typ odpadu, tak pro typy různé. Vyskytují se také záznamy, kdy jedna nádoba přísluší více adresám, to je ale zohledněno v názvu stanoviště. Jako zastávky jsou uvažována zastavení vozidla za účelem sběru odpadu. Jednotlivé zastávky tedy korespondují se záznamy ve zpracovávaných datech.

Kategorie sváženého odpadu jsou Směsný komunální odpad, Biologicky rozložitelný odpad, Papír a lepenka, Plasty a Sklo. V grafech a tabulkách pro ně bude využíváno označení SKO, BIO, PAP, PLA a SKL.

V datech se vykytuje velké množství chybějících údajů, a to zejména u atributů NADOBA, NAZEV STANOVISTE, ULICE, EVIDENCNI CISLO, CISLO POPISNE a ID NADOBY.

Důvodem je, že tyto údaje nebyly původně uváděny, a navíc kvůli chybějícímu identifikačnímu číslu nádoby je nelze zpětně doplnit. Doba, kdy se tyto údaje začaly uvádět, se liší i podle obcí. Tento nedostatek dostupných dat je možné částečně vyřešit v průběhu agregace, zbytek dat bude možné použít pouze pro některé modely. V datech se vyskytují i další chyby a nelogické záznamy. Cílem je chyby odhalit a odstranit či zajistit, aby jejich vliv na výsledek byl co nejmenší.

V průběhu tvorby práce bylo zjištěno, že důvodem pro nedostupnost některých dat je nepřítomnost monitorovacího systému na všech popelnicích. Podobně identifikační číslo mají pouze některé nádoby.

4.2. Analýza vstupních dat

4.2.1. Základní předpoklady

Pro další práci s daty se vychází z následujících předpokladů:

- Jedno vozidlo během daného dne absolvuje pouze jeden okruh. (Tento předpoklad je později vyvrácen. Okruhů může být za den i více.) V rámci tohoto okruhu toto vozidlo sváží pouze jeden typ odpadu.
- Svoz odpadu (a i nádoby) lze rozdělit na základní dva typy. Svážení odpadu od domu k domu, kdy se jedná o svoz menších popelnic náležícím k jednotlivým adresám. Dále tento typ svozu bude značen DBD (z anglického „door by door“). Druhou variantou je svoz sběrných hnízd, kde se jedná o velké popelnice nacházející se na veřejných místech. Nejvýznamnější sběrná hnízda jsou pro tříděný odpad. Tento typ svozu bude dále značen HS (hnízdový svoz).
- V rámci jednoho okruhu je sběr prováděn pouze jedním způsobem, tzn. DBD nebo HS. (Později je zjištěno, že tento předpoklad neplatí zcela vždy. Někdy jsou svezeno i několik nádob druhého typu.)

Dále je nutné upozornit, že vozidlo při jednom okruhu nemusí zastavit u všech nádob v dané obci. Některé obce se totiž pravděpodobně kvůli trase a množství odpadu svážejí v rámci více okruhů po částech. Okruhy by teoreticky měly mít stálou frekvenci svozu a složení navštívených obcí. V praxi to ale platit nemusí, dochází totiž k častým změnám a jednorázovým výjezdům. Je tedy třeba pravidelnost svozu prozkoumat.

4.2.2. Chybějící hodnoty

V prvním kroku jsou identifikovány chybějící hodnoty v datech. V tabulce 4.1 jsou uvedeny chybějící hodnoty pro jednotlivé atributy.

Tabulka 4.1: Chybějící hodnoty u jednotlivých atributů (celkový počet záznamů je 358 379)

	počet chybějících záznamů
ID	0
DATUM	0
NADOBA	208 951
ODPAD	1
NAZEV STANOVISTE	208 951
ULICE	314 137
EVIDENCNI CISLO	357 320
CISLO POPISNE	210 290
MNOZSTVI	0
OBEC	0
ID NADOBY	208 951
LAT	1 197
LON	1 197
SPZ	177

Z tabulky 4.1 je patrné, že základní údaje jako jsou DATUM, ODPAD, OBEC a MNOZSTVI jsou až na jednu výjimku pro údaj ODPAD úplné. Tento jeden záznam s chybějící informací ODPAD má uvedený záznam pro atribut NADOBA. Pro další práci je tedy chybějící záznam pro ODPAD doplněn podle záznamu NADOBA, který typ odpadu jednoznačně stanovuje. Atributy ULICE, EVIDENCNI CISLO a CISLO POPISNE nejsou pro tuto analýzu důležité. Stejnou informaci totiž obsahuje NAZEV STANOVISTE, který je přítomen vždy, kdy je přítomen kterýkoliv z výše zmíněných atributů. Dále se s těmito atributy pracuje, pouze pokud je potřeba ověřit pravdivost hodnot v atributu NAZEV STANOVISTE. LAT a LON nejsou v této práci využity z důvodu nedostatečné přesnosti. Atributy NADOBA, NAZEV STANOVISTE a ID NADOBY mají stejný počet chybějících hodnot a tyto hodnoty chybí u stejných záznamů. Celkem je k dispozici 358 379

záznamů, tedy zmiňované atributy chybí v 58,30 % záznamů. Vzhledem k množství takových záznamů by jejich odstranění znamenalo ztrátu podstatné části dat, a navíc možné znehodnocení zbývajících dat. Situace je řešena snahou chybějící hodnoty odhadnout a využitím i záznamů s nekompletními údaji v agregaci. U SPZ chybí 0,05 % záznamů, neměl by tedy být problém tyto záznamy vynechat, pokud nepůjde hodnotu SPZ odhadnout z dat. Dále bude problematika chybějících hodnot řešena v 4.2.4.

4.2.3. Chyby v datech

Jak už bylo zmíněno v kap. 4.1, v datech lze očekávat výskyt chybných hodnot. Cílem je chybné hodnoty identifikovat, aby jejich přítomnost nezpůsobovala zkreslení modelů. Pokud se po identifikaci anomálií v datech (chybných hodnot) přistoupí k jejich vynechání, pracuje se dále s hodnotami jako s chybějícími, viz kap. 4.2.2.

V prvním kroku jsou identifikovány chybné hodnoty pro kategorické atributy: ID, NADOBA, ODPAD, NAZEV STANOVISTE, OBEC, ID NADOBY a SPZ. K dispozici není seznam správných kategorií, tedy hodnot, kterých by měly záznamy nabývat. Kontrola je tedy provedena ručně, a to výpisem kategorií s počtem jejich výskytu. Kontrolováno je případné zdvojení kategorií či pravopisné chyby. Chybou v datech může být také kategorie s příliš nízkým počtem výskytů. Výstupy ruční kontroly pro jednotlivé atributy jsou následující:

- V případě ID je kontrolována unikátnost záznamů. Potvrdilo se, že každý záznam má vlastní hodnotu ID, nebyla tedy nalezena chybná hodnota.
- U atributu NADOBA bylo nalezeno znásobení kategorií z důvodu nejednotných zápisů, např. využití spojovníku namísto pomlčky, chybějící či přebývající mezery, záměny slov (’Plasty- 240 l žlutá’ a ’Plast - 240 l žlutá’, ’Papír 1 100 kontejner’ a ’Papír - 1100 l modrá’ a podobně). Chybné hodnoty tohoto typu byly opraveny.
- U atributu ODPAD se opět vyskytují znásobené kategorie, např. ’Papírové a lepenkové obaly’ a ’Papír a lepenka’. Nejednotnost záznamů tohoto typu byla ručně opravena.
- Atribut NAZEV STANOVISTE má 4 092 kategorií, takže ruční kontrola není příliš spolehlivá. Přesto tato kontrola byla provedena a nebyly nalezeny hodnoty, které by byly označeny za chybné. Vyskytovalo se zde relativně mnoho kategorií s jednou

nebo velmi málo hodnotami. Může jít ale o místa, která mají nízkou frekvenci svozu, nebo se tato místa začala svážet až ke konci měřeného období.

- Atribut ID NADOBY se chová podobně jako NAZEV STANOVISTE. Tento atribut zahrnuje dokonce 10 613 kategorií. Nebyly nalezeny žádné hodnoty, které by se na první pohled jevily jako chybné. ID NADOBY jsou pravděpodobně stanovena vnitřně samotnou svozovou firmou vzestupně pro nové nádoby. Tento předpoklad potvrzuje to, že nejvyšší čísla ID NADOBY mají nejmenší počet záznamů. Pro další analýzy není ID NADOBY důležité, a nebude tedy tomuto atributu věnována pozornost.

- U atributu SPZ se objevuje znovu různý počet mezer a záměna 'O' za '0'. Dále se zde nacházejí podezřelé kategorie, dvě z těchto kategorií mají pouze po dvou záznamech. Po podrobnějším zkoumání bylo zjištěno, že u obou SPZ se jedná pouze o jeden ranní a jeden odpolední záznam směsného komunálního odpadu. Navíc tento ranní a odpolední svoz v obou případech vykazuje stejné množství odpadu, u jedné z SPZ se jedná o velmi vysoké množství. Obec je také stejná ráno i odpoledne a ze zbylých údajů chybí všechny atributy kromě ID. Tyto záznamy byly pro navazující analýzy odstraněny. Dále má jedna kategorie název, který neodpovídá zákonnému formátu poznávacích značek. Po kontrole příslušných záznamů se zdá, že se jedná o úplný záznam svozu z jednoho dne, není tedy nutné propojovat tuto kategorii s jinou.

Dále budou prozkoumány rozpory v datech. Vzhledem k tomu, že některé informace jsou v datech přítomny ve více atributech, je možné ověřit správnost informace. V prvním případě toto nastává u ODPAD a NADOBA, kde lze vyčíst typ sváženého odpadu i z typu nádoby. Z kontroly vyplývá, že si tyto údaje odporují v 1 895 případech. Další násobná informace je evidována u atributů NAZEV STANOVISTE, ULICE, EVIDENCNI CISLO a CISLO POPISNE. V atributu NAZEV STANOVISTE by měly být zahrnuty informace z ostatních tří atributů (ULICE, EVIDENCNI CISLO A CISLO POPISNE). Ukázalo se, že formát zápisu v NAZEV STANOVISTE není vždy jednotný. V některých případech je zahrnut název obce a název ulice, někdy tyto informace chybí. Dále se vyskytují záznamy, u nichž není uvedeno číslo popisné obce, pokud je NAZEV STANOVISTE vztažen na sběrné hnízdo. Znásobení informace není přítomné ve všech záznamech. Žádný přímý rozpor ale nebyl nalezen.

4.2.4. Analýza a příprava dat

V předchozí kapitole jsem provedl základní identifikaci chybějících a chybných hodnot. Zde se budu snažit objevené problémy dále analyzovat a řešit. Budu také pokračovat v hledání zvláštností v datech.

Předpoklad o jednom okruhu za den

Nejdříve byla ověřena platnost předpokladu, že jedno vozidlo sváží během jednoho dne pouze jeden okruh a jeden typ odpadu. Z dat byly podle jednotlivých dnů a vozidel vypsány záznamy o typech odpadů. Z 1 151 svozů jich 306 obsahovalo záznamy s více druhy odpadu. Těchto 306 záznamů bylo dále detailně prostudováno. Tímto způsobem byly nalezeny záznamy o svezení Směsného komunálního odpadu, které mají chybějící atributy NADOBA, NAZEV STANOVISTE a ID NADOBY a současně všechny atributy kromě ID a ODPAD jsou shodné. Bez ID NADOBY nelze jednoznačně říci, že se jedná o zdvojení. V praxi by mohlo jít o znečištění nějakého typu odpadu, tedy ve výsledku jeho zařazení pod Směsný komunální odpad. Není ale žádoucí takový záznam uvažovat dvakrát. Může nastat situace, kdy hmotnost odpadu v tomto záznamu neodpovídá ani původnímu odpadu, ani Směsnému komunálnímu odpadu. Vytvoření nové kategorie by ale pouze dále rozmělnilo data a nepřineslo novou informaci. Aby bylo zachováno pravidlo stejného odpadu v jednom svozu, byl zachován záznam pouze s původním odpadem. Lze totiž předpokládat, že hmotnost znečištěného odpadu je blízká hmotnosti původního odpadu.

Ostatní záznamy, v nichž se vyskytují různé typy odpadu v jednom svozu, jsou způsobené časově oddělenými skupinami různého odpadu. Jinými slovy předpoklad o jednom okruhu denně byl mylný. Z toho vyvstává několik rizik. Později budou data agregována po obcích v jednotlivých okruzích, viz kap. 4.3. Porušením zmiňovaného předpokladu nastává problém s identifikací těchto okruhů. Kombinace rozdělení podle odpadu a obce nemusí stačit, protože může nastat situace, kdy se v jednom dni sváží různé části jedné obce v různých okruzích. Tato možnost zní nepravděpodobně, ale je možná. Je tedy nutné při agregaci zohledňovat časové rozdíly mezi záznamy. Mezi dvěma různými návštěvami jedné obce by měl být rozdíl několika hodin. Avšak z důvodu později objevených zvláštností v datech je i toto kritérium založené na čase problematické, viz kap. 4.2.4.

Dále byly řešeny nesrovnalosti v typu odpadu mezi ODPAD a NADOBA objevené již dříve. Aby bylo možné určit, kterou informací se řídit, byla analyzována data v kontextu časově blízkých záznamů. V okruzích, kde je dostatek informací odpovídá zbytku okruhu

informace z ODPAD. U několika okruhů se liší ODPAD a NADOBA ve všech záznamech, ve kterých jsou hodnoty obou atributů uvedeny. V těchto okruzích ale existují záznamy, kde je uveden pouze ODPAD. Vzhledem k tomu, že ODPAD je atribut, jehož hodnota je uvedena u všech záznamů, a i v místech, kde to lze ověřit, je správný, uvažuje se dále, že se jedná o lepší zdroj informací než NADOBA. Informace tedy bude převzata z atributu ODPAD. V těch případech, kde je možné provést kontrolu zbytkem svozu, se zdá, že objem nádoby koresponduje s typem svozu, který je možné vyčíst z NAZEV STANOVISTE.

Základní rozdělení na typy svozu

V této části budou záznamy rozděleny podle typu svozu (DBD nebo HS, viz kap. 4.2.1). Typ svozu je možné určit dvěma způsoby, a to z atributů NAZEV STANOVISTE, ve kterém bývají sběrná hnízda uvedena, a z NADOBA, kde DBD je svoz menších nádob (120 l nebo 240 l) a HS je svoz velkých nádob (1 100 l nebo označených „sběrné hnízdo“). Identifikace podle těchto dvou způsobů se liší pouze v 58 případech. Po krátké kontrole těchto 58 případů je patrné, že přesnější je informace z NADOBA, protože v atributu NAZEV STANOVISTE u některých velkých popelnic chybí popis „sběrné hnízdo“. Jak už bylo zjištěno dříve z rozporu v atributech ODPAD a NADOBA, viz kap. 4.2.3, některé informace o typu odpadu jsou chybné. Typ nádoby ale v rozporu nebyl, a navíc se tu informace z NAZEV STANOVISTE a NADOBA o typu svozu liší v malém počtu případů, které neodpovídají těm rozporným z NADOBA a ODPAD. Lze tedy předpokládat, že informace o objemu nádoby je bezpečná. Pro určení typu svozu byla využita informace z NADOBA. Zajímavostí je, že pro Sklo se v datech vyskytují pouze velké nádoby, stejně tak v NAZEV STANOVISTE je vždy uvedeno označení „sběrné hnízdo“. Předpokládá se tedy, že všechny záznamy o svážení skla jsou typu HS. V tabulce 4.2 jsou uvedeny počty identifikovaných záznamů. Za povšimnutí stojí nízký počet záznamů s určeným typem svozu u Směsného komunálního odpadu.

Tabulka 4.2: Počty identifikovaných typů svozů podle jednotlivých typů odpadu

	BIO	PAP	PLA	SKL	SKO
počet záznamů DBD svozu	52 486	30 511	54 286	0	3 870
počet záznamů HS svozu	994	1 595	4 213	1 353	80

Pokud je platný předpoklad o jednom typu svozu v rámci jednoho okruhu, každý typ svozu bude vyžadovat jiný model.

Frekvence svozu na základě nádob

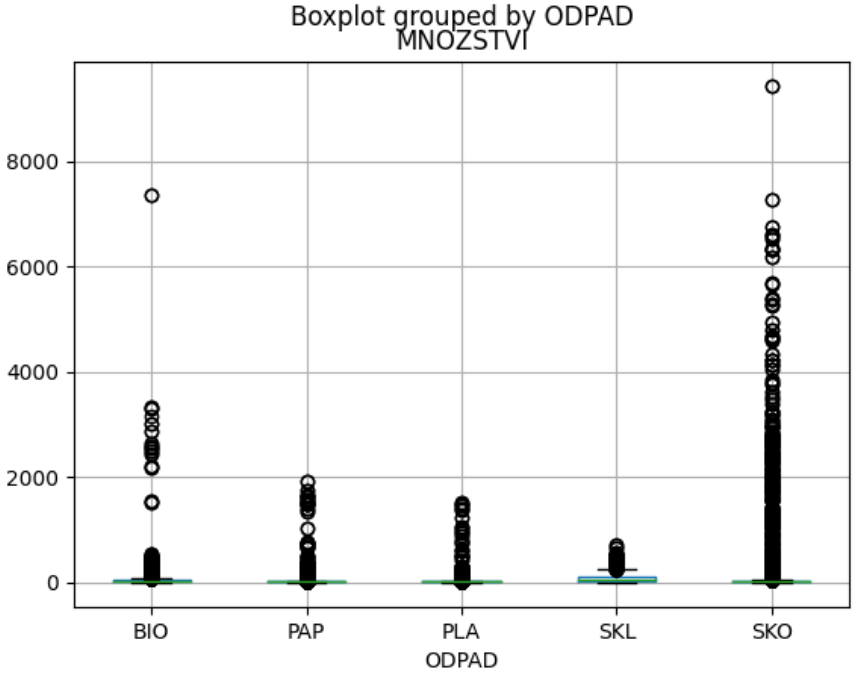
Na základě úvahy, že množství odpadu na konkrétní zastávce by mělo být závislé na době, jakou byla nádoba plněna, je zájem tuto délku zjistit. To by mělo být možné z rozdělení dat podle nádob a následně výpočtem časových rozdílů svozů. Problém ale nastává se záznamy bez uvedeného ID NADOBY. Nejen jejich dobu od posedního svozu nelze určit, ale také vytvářejí časové mezery. U ostatních záznamů potom vznikají vyšší hodnoty. Tento jev je natolik silný, že znehodnocuje tuto veličinu. Výpočet frekvence mezi svozy tedy tímto způsobem není možný.

Grafická analýza dat

Aby bylo možné posoudit základní charakteristiky rozdělení zkoumaných dat a přítomnost odlehlých hodnot, je nutné nejdříve provést základní vizualizaci. Nejdříve byla vizualizace provedena pomocí krabicového grafu 4.2. Z něj je patrná výrazná nesymetrie rozdělení, a také podezření na odlehlé hodnoty, protože hodnoty překračující 2 000 kg jsou nepravděpodobné.

V analýze odlehlých hodnot i rozdělení dat by měl být zohledněn typ svozu a obcí, protože obojí může mít vliv na parametry rozdělení. Třídění podle typu svozu je ale problematické, protože se vyskytuje mnoho záznamů, které roztřídit nelze. Pro lepší pochopení vztahu mezi odlehlými hodnotami a typy svozu, byly tyto dva parametry zaznamenány do tabulky 4.3. Z tabulky je patrné, že i v označených datech je prostor pro analýzu odlehlých hodnot. Protože u některých typů odpadu přesahuje maximální hodnota pro DBD hodnotu pro HS, což je podezřelé. Podstatnější ale je, že při porovnání maximálních hodnot s grafem 4.2 je patrné, že se vyskytuje velké množství nezařazených hodnot, která přesahují zmiňovaná maxima, a to i několikanásobně. Nejdříve se tedy předpokládá, že označené hodnoty jsou správné a že odlehlé hodnoty se vyskytují pouze u těch neoznačených. Identifikace odlehlých hodnot nebude tedy dělena podle typu svozu.

Z krabicových grafů 4.2 je patrné, že data pravděpodobně nemají normální rozdělení. V dalším kroku byly vykresleny histogramy a grafy empirických distribučních funkcí pro záznamy s hodnotou MNOZSTVI nižší než 800 kg, aby nebyly grafy příliš zkresleny vysokými hodnotami. U všech typů odpadů jsou grafy podobného charakteru, dále na obrázku 4.3 je tedy ukázka pro Biologicky rozložitelný odpad.



Obrázek 4.2: Krabicové grafy množství v kg jednotlivých typů odpadu

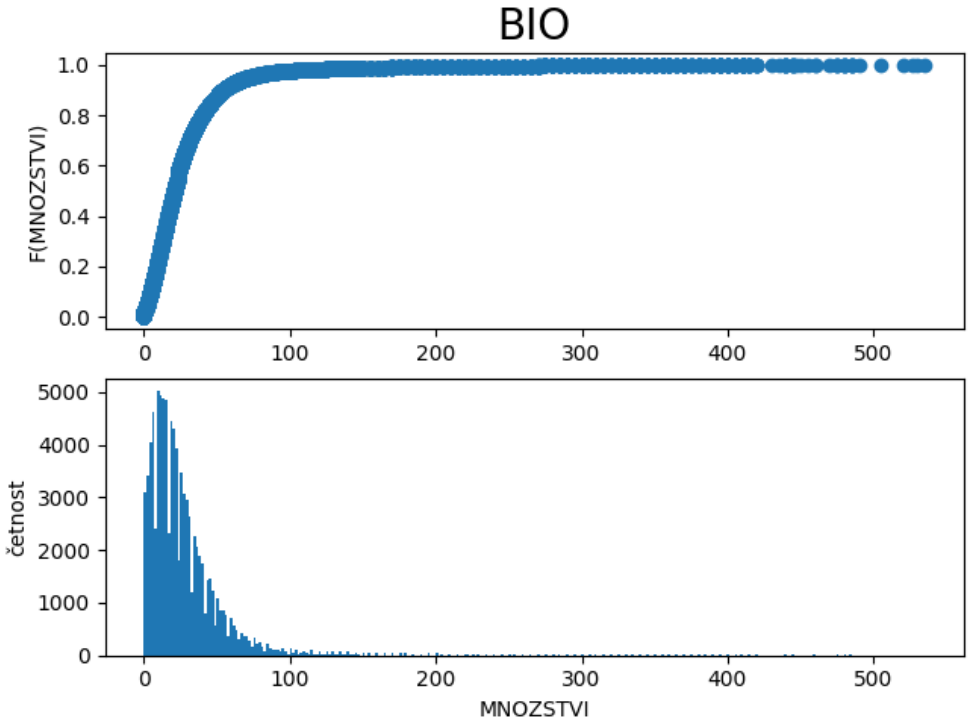
Tabulka 4.3: Maximální hodnoty množství naloženého odpadu v kg na jedné zastávce rozdělené podle typu odpadu a typu svozu.

	BIO	PAP	PLA	SKL	SKO
maximální hodnota MNOZSTVI pro DBD svoz [kg]	152	153	76	-	266
maximální hodnota MNOZSTVI pro HS svoz [kg]	485	150	153	510	230

Transformace dat

Z grafů 4.3 je patrné, že MNOZSTVI nemá normální rozdělení. Graf je podobný log--normálnímu rozdělení. Toto rozdělení je běžné u jednostranně ohraničených veličin, což hmotnost je (nemůže nabývat záporných hodnot). Pro odstranění šikmosti je možné použít logaritmickou transformaci a přiblížit tím data normálnímu rozdělení. Pro použití některých statistických nástrojů je normalita žádoucí, proto je pro ně logaritmická transformace použita. Vzhledem k přítomnosti nulových hodnot je použita transformace v podobě $X_i' = \ln(X_i + 1)$.

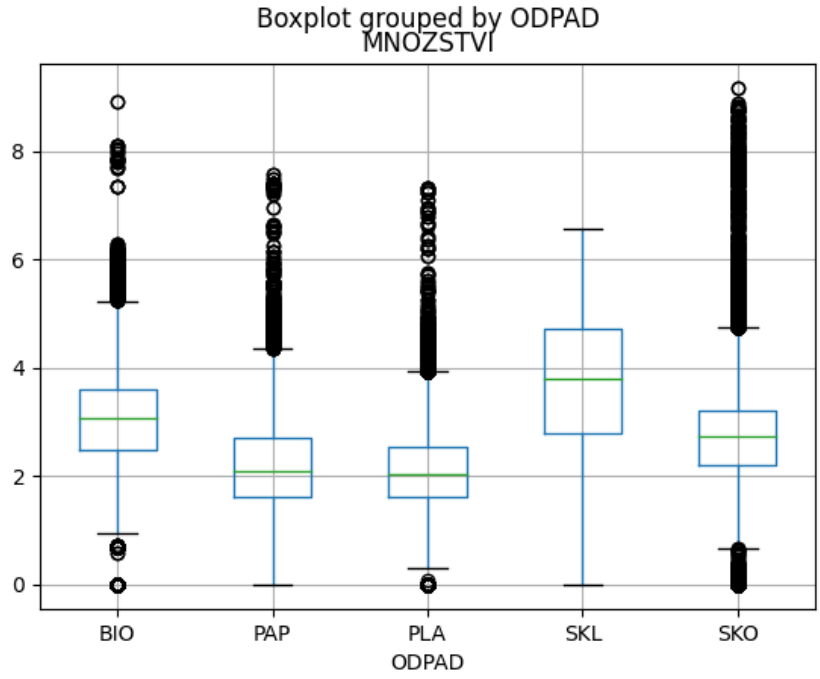
Po transformaci dat jsou pro lepší orientaci znovu vykresleny krabicové grafy. V 4.4 je vidět, že i po transformaci zůstalo mnoho hodnot mimo tzv. vousy krabicového grafu (vousy zde odpovídají $1,5$ IQR od prvního a od třetího kvartilu, kde IQR je rozdíl mezi



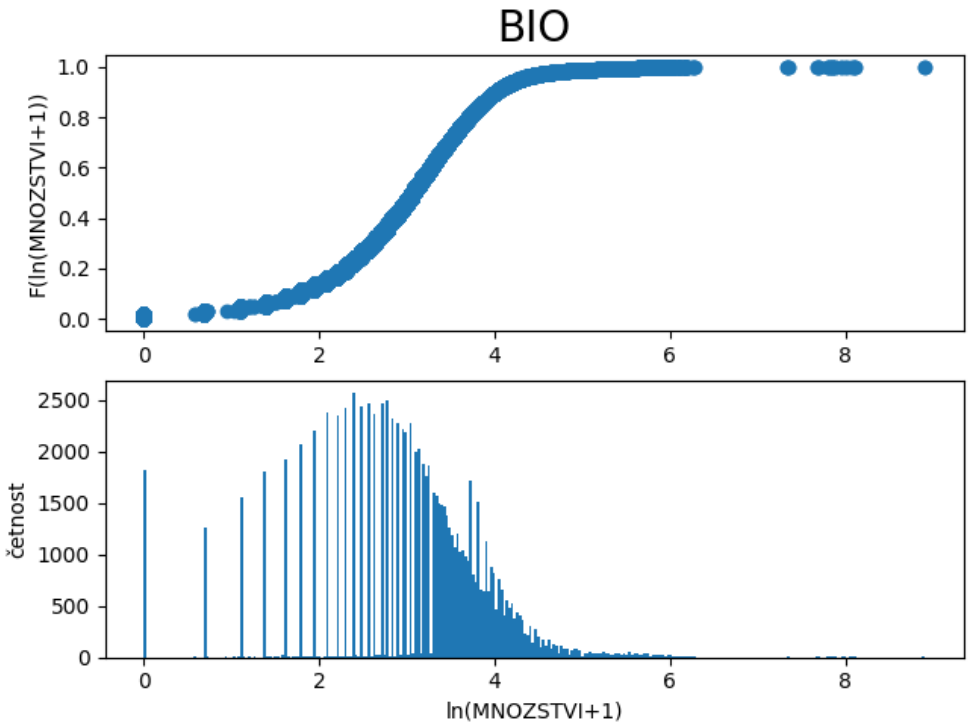
Obrázek 4.3: Histogram a graf empirické distribuční funkce pro množství v kg svezeného Biologicky rozložitelného odpadu, kde jsou použity pouze hodnoty menší 800 **kg**

prvním a třetím kvartilem). Některé takové hodnoty se vyskytují i pod grafy, což by nemělo nastat (hodnoty blízké nule jsou očekávané). Důvodem může být velké množství odlehlých hodnot, které posunou i medián. Dalším důvodem pro vysoký počet hodnot podezřelých z odlehlosti je, že pracujeme dohromady se svozy typu DBD a HS. Data jsou tedy směsí dvou rozdělení, a krabicové grafy jsou určeny pro homogenní normální rozdělení. Samotná střední část krabicového grafu již vypadá symetricky. Testovat normalitu dat před vyšetřením přítomnosti odlehlých hodnot nemá smysl.

Znovu jsou vykresleny histogramy a empirické distribuční funkce, tentokrát pro všechny hodnoty. Znovu jsou uvedeny pouze grafy pro Biologicky rozložitelný odpad. Odhlédneme-li od odlehlých hodnot, tak oba grafy 4.5 již připomínají normální rozdělení. Je to patrné hlavně na grafu empirické distribuční funkce. Histogram ukazuje nehomogenitu dat. Důvodem je zaokrouhlování. Většina záznamů obsahuje hodnoty MNOZSTVI zaokrouhlené na jednotky. Tento efekt je u nízkých hodnot velmi zvýrazněn logaritmickou transformací a naopak potlačen u vysokých hodnot. Tato nehomogenita se může projevit mimo jiné i u testů normality.



Obrázek 4.4: Krabicové grafy zlogaritmovaného množství jednotlivých typů odpadu



Obrázek 4.5: Histogram a graf empirické distribuční funkce pro zlogaritmované množství svezeného Biologicky rozložitelného odpadu

Identifikace odlehlých hodnot

Před tím, než byly strojově aplikovány testy pro identifikaci odlehlých hodnot, byly nejvyšší hodnoty ručně prozkoumány. Bylo odhaleno časté zdvojení vysokých hodnot s dalšími podezřelými vlastnostmi. Po dotazu na společnost TSMH s r. o. byla doplněna informace, že některá vozidla nemají výbavu pro průběžné měření hmotnosti svezeného odpadu. Při takových svozech je odpad zvážen pouze jednou za celou obec. Tato hmotnost je zaznamenána do dvou záznamů, které se liší pouze v ID a DATUM, kde v prvním je zaznamenán čas začátku celého svozového okruhu a v druhém jeho konec. Časy jsou pravděpodobně jen odhadnuté, protože bývají uváděny v celých půlhodinách. Tyto záznamy nenesou tedy informaci o době svozu v dané obci ani o počtu zastávek v ní. Navíc tyto hodnoty vnášejí další rozdělení do zpracovávaných dat. Než bude možné pokračovat s identifikací odlehlých hodnot, je nutné vyloučit tyto záznamy ze souboru dat. Díky dříve uvedeným vlastnostem lze problematické záznamy identifikovat.

Po identifikaci bylo zjištěno, že takto agregované jsou všechny záznamy, u kterých chybí hodnota SPZ. Tato vlastnost se také opakuje pro některé konkrétní SPZ. V součtu jde ale pouze o 957 záznamů.

Dalším krokem je odstranění očividně chybných hodnot. Na základě konzultace jsem pro toto odstranění zvolil hodnoty 300 kg pro BIO, 350 kg pro PAP, 100 kg pro PLA, 350 kg pro SKL a 500 kg pro SKO. Takto bylo odstraněno 494 hodnot.

Na takto očištěných datech je provedena statistická analýza odlehlých hodnot. Kvůli již zmiňovaným problémům s normalitou dat je použito robustní kritérium (2.2.3). Nejdříve je proveden výpočet Q na transformovaných datech, protože hodnota $1,483$ je vypočtena pro normální rozdělení a kritická hodnota by se v tom případě měla pohybovat řádově kolem hodnoty 2. Výpočet je proveden i na netransformovaných datech, protože by v nich měla být jasnější hranice mezi běžnými a odlehlými hodnotami. V obou případech není v datech přítomna skoková změna, která by pomohla identifikovat kritickou hranici. Hodnoty Q pro transformovaná data sice dosahují hodnot až kolem 5, ale vzhledem k rovnoměrnému zvyšování a kvůli problémům s normalitou tuto hodnotu Q nezavrhujeme. Podobný výsledek dávají netransformovaná data, pouze zde rostou hodnoty Q rychleji. Zdá se tedy, že všechny odlehlé hodnoty byly odstraněny expertním odhadem. Toto potvrzuje výpočet Q na datech bez odstranění hodnot na základě expertního odhadu, na základě něhož lze kritické hranice nalézt. Budu se tedy řídit expertním odhadem a další

analýza odlehlých hodnot již nebude provedena.

V tabulce 4.4 jsou zaznamenány počty záznamů odstraněných pomocí dříve uvedených způsobů.

Tabulka 4.4: Množství záznamů, které byly odstraněny, podle jednotlivých odpadů

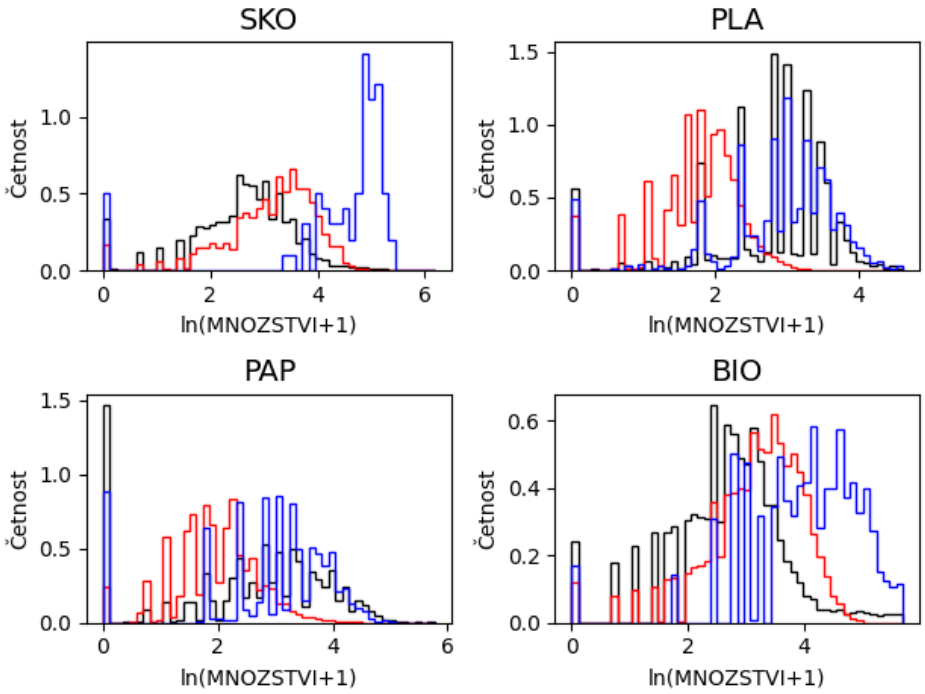
	BIO	PAP	PLA	SKL	SKO
počet hodnot odstraněných na základě identifikovaného jevu	28	299	51	70	508
počet hodnot odstraněných na základě expertního odhadu	295	3	102	81	13

Identifikace typů svozu u neoznačených záznamů

V kapitole 4.2.4 je provedena identifikace svozů odpadu. Zbývá ale stále ještě 207 472 hodnot, pro které není typ svozu určený. Informace o typech svozu ale může být užitečná u tvorby modelu. Pokusím se tedy pomocí statistických metod odhadnout typ svozu u části dat.

Nejdříve jsou vytvořeny histogramy dat rozdělených podle typů svozu a odpadu. Pro lepší vizualizaci jsou použity zlogaritmované hodnoty a také jsou normalizovány plochy histogramů (v každém grafu zvlášť), aby byly grafy různých typů svozu srovnatelné. Graf SKL není uveden, protože celý svoz probíhá v režimu HS. Výsledné grafy jsou v obrázku 4.6 a neobsahují již dříve odstraněné odlehlé hodnoty. Z grafů je patrné, že ani po transformaci a rozdělení podle typů svozu nesplňují data normalitu. Toto je ověřeno i testem normality (použit byl Andersonův-Darlingův test 2.1.1). V datech je vidět, že se grafy DBD a HS překrývají. (Tento efekt je logaritmizací vizuálně zvýrazněn, ale i v netransformovaných datech je výrazný.) Nelze tedy jednoduše stanovit číselnou hranici pro rozdělení typů svozu. V histogramech pro PLA a PAP se zdá, že graf nezařazených hodnot odpovídá grafu pro HS svoz, což naznačuje, že většina nezařazených hodnot bude tohoto typu. U BIO a SKO je naopak střed grafu nezařazených hodnot až mírně nalevo od grafu DBD svozu. Důvodem může být, že DBD svoz se skládá ze sváženích nádob objemu 120 l a 240 l. Pokud tedy v nezařazených hodnotách je více svozů 120l nádob, vysvětlovalo by to nižší hodnoty svezeného množství.

Ideou, na základě které bude provedena identifikace nezařazených hodnot, je předpoklad, že během jednoho svozového okruhu jde o sváženích pouze jednoho typu svozu a že

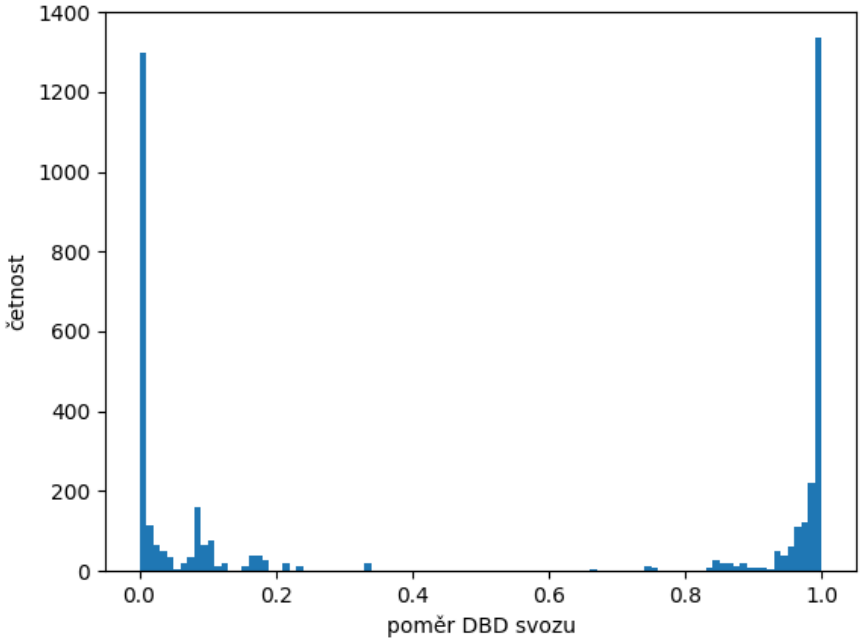


Obrázek 4.6: Histogramy transformovaných množství odpadu podle typů svozu a podle typů odpadu (červená značí svoz DBD, modrá svoz HS a černá značí nezařazené záznamy)

rozdělení jednotlivých typů svozu v jednotlivých obcích je stálé. Pro kontrolu tohoto předpokladu je vykreslen v histogramu poměr počtu DBD záznamů vůči součtu obou typů pro jednotlivé svozy. Není zohledněno množství nezařazených hodnot v těchto svozech. Z obrázku 4.7 je vidět, že přestože je většina svozů jednoho typu, vyskytují se i případy, kdy jsou do svozu přibrány nádoby druhého typu.

Pro identifikaci je použito dvouvýběrových testů. Z důvodu problémů s normalitou není možné použít F-testu pro kontrolu shody rozptylů. Kritériem se proto stane Studentův t-test ve variantě 2.7, která je robustní vůči odchylkám od normality a nevyžaduje předpoklad rovnosti rozptylů. I přes použití robustního testu jsou data před testováním logaritmicky transformována, protože zvolený test má pro normální data větší sílu. Srovnávána jsou takto nezařazená data z jednotlivých svozů v jednotlivých obcích s referenčními výběry, kterými jsou všechny záznamy dané obce pro jednotlivé typy svozu. Velikosti srovnávaných výběrů jsou proto podstatně rozdílné. To zvyšuje pravděpodobnost zamítnutí hypotézy o rovnosti středních hodnot daným testem. Soubory dat z obcí, kde je jen malý počet hodnot či nejsou nezařazené hodnoty většinou přítomných záznamů, testovat nebudu.

Tabulka 4.5 udává, kolik kterého svozu bylo identifikováno u jednotlivých svozů. Tyto



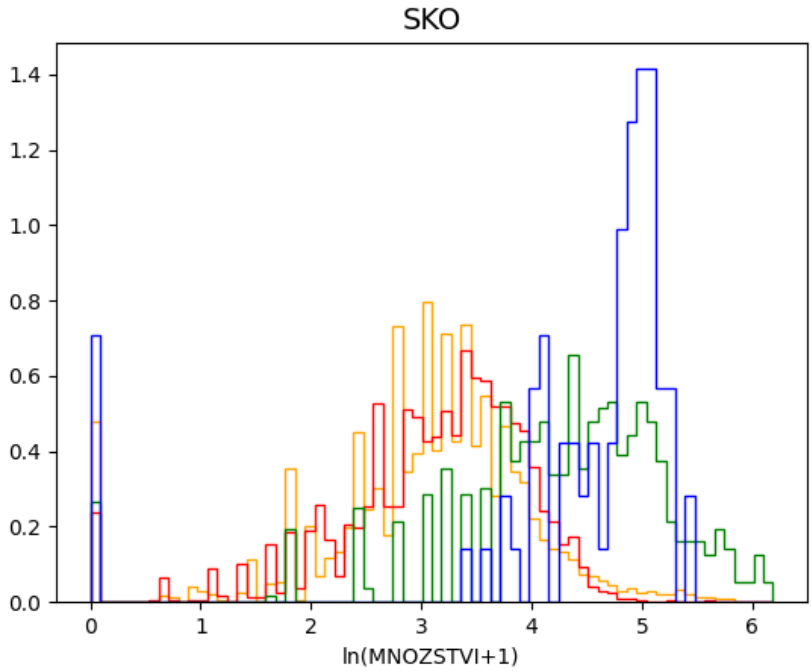
Obrázek 4.7: Histogram poměru počtu DBD svozů vůči počtu obou svozů v jednotlivých svozových okruzích

počty jsou výrazné zvlášť u HS svozu, kde jsou srovnatelné nebo větší než původní počty tohoto svozu. U Směsného komunálního odpadu je vysoký i počet nově identifikovaných DBD svozů, a dokonce dvojnásobně převyšuje původní hodnotu.

Tabulka 4.5: Počty nově identifikovaných typů svozu

	BIO	PAP	PLA	SKO
počet nových DBD záznamů	963	759	808	8788
počet nových HS záznamů	417	2085	7738	641

Nově identifikované hodnoty jsou v histogramech porovnány se starými. Kvůli nepřehlednosti takových grafů je uveden pouze jako příklad histogram pro SKO 4.8. V ostatních grafech jsou nesrovnalosti maximálně srovnatelné s těmi v tomto grafu. Nové rozdělení pro DBD svozy celkem odpovídá tomu starému. U HS svozů už je rozdíl větší, zvláště ve špičatosti. Důvodem je, že kritérium mělo posuzovat rovnost středních hodnot. Jelikož však nebyly srovnávány tyto celkové výběry, ale výběry podle obcí, jsou výsledky testu silnější, než by se zdálo z grafu. Nové identifikace typů odpadu budou proto akceptovány.



Obrázek 4.8: Histogram transformovaných množství odpadu podle typů svozu(červená značí staré záznamy DBD svozu, oranžova nové záznamy DBD svozu, modrá staré záznamy HS svozu a zelená nové HS záznamy)

4.3. Agregace

Aby mohla být data použita pro popis doby a množství sváženého odpadu je třeba je agregovat. Data jsou agregována podle ODPAD, dnů (získaných z DATUM), SPZ a OBEC. Takto vznikne nová datová sada. Z původních dat jsou přeneseny atributy SPZ, OBEC a ODPAD, jelikož přes ně probíhá agregace. Podobně se do nového atributu DATUM přenese jen informaci o dni. Nově vzniklé atributy jsou:

- ID - Identifikační číslo mnou přiřazené agregovaným záznamům.
  Toto ID zapíši i do původních dat jako ID SVOZU OBEC, abych mohl podle něj identifikovat data, z kterých nový agregovaný záznam vznikl.
- ID OKRUH - Identifikační číslo mnou přiřazené jednotlivým svážecím okruhům.
  Lze podle něj identifikovat záznamy jednoho svozového okruhu.
- MNOZSTVI - Součet hmotnosti v kg ze všech záznamů během svozu v dané agregaci.
- DOBA SVOZU - Jde o dobu jakou vozidlu trval svoz v dané agregaci.
  Vypočtena je jako rozdíl časů posledního a prvního záznamu v dané agregaci. Je

zavedena i DOBA SVOZU V MIN, která je pouze počtem minut svozu. Takto číselná hodnota se snáze strojově zpracovává.

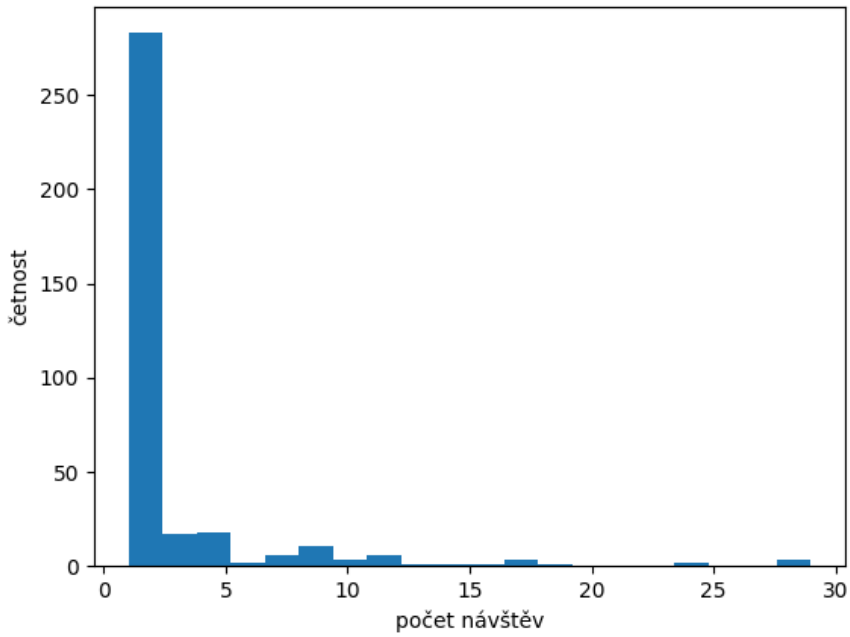
- POCET ZASTAVEK - Počet záznamů, ze kterých vznikl agregovaný záznam.
- ZACATEK SVOZU - Atribut DATUM prvního záznamu v dané agregaci.
- KONEC SVOZU - Atribut DATUM posledního záznamu v dané agregaci.
- POMER DBD BEZ NAN - Poměr počtu záznamů DBD vozu vůči součtu DBD i HS svozů v dané agregaci.
  Nezařazené záznamy tu nejsou zohledněny.
- POMER CHYB HODNOT - Poměr počtu záznamů s neurčeným typem svozu vůči celkovému počtu záznamů v dané agregaci.

4.3.1. Svozové okruhy a frekvence svozu

Při agregaci dat byly zkoumány svozové okruhy. Důvodem tohoto zkoumání byla mimo jiné snaha o zjištění frekvence svozu, o což byl již neúspěšný pokus s neagregovanými daty, viz. 4.2.4. Předpoklad, kterého by šlo použít pro identifikaci této frekvence je, že pro daný typ svozu a odpadu by měla být její hodnota stálá. Po průzkumu dat a snaze o vytvoření vhodného kritéria je ale třeba konstatovat, že jen na základě těchto informací nelze frekvence svozu vypočíst bez velkého rizika chyb. Důvodem jsou sváženi obcí po částech, relativně časté změny frekvence sváženi a jednorázové návštěvy obce. Bez ruční kontroly tedy není možné frekvenci svozu vypočíst bez silného zatížení chybami. Pro ilustraci situace je v uveden histogram četnosti počtu výskytů konkrétních kombinací obcí v okruzích. Je zde vidět že velké množství kombinací bylo použito pouze jednou. To naznačuje časté změny ve svozových plánech.

4.3.2. Přidání socioekonomických dat

Za účelem získání více informací, které by mohly mít vliv v modelu, byl ze stránek Českého statistického úřadu (ČSÚ) [10] stažen soubor s tzv. Územně analytickými podklady (ÚAP), což jsou vybrané ukazatele k jednotlivým obcím, které mají sloužit k územnímu plánování. Data za minulý kalendářní rok jsou aktualizována k 30. červnu. Nejnovější datový soubor mám tedy za rok 2019. Jelikož meziroční změna většiny obsažených ukazatelů je malá, a zkoumání časového vývoje není předmětem této práce, tak jsou tato data použita i pro roky 2020 a 2021. Z dat byly vybrány, jako pro model zajímavé, údaje o počtu



Obrázek 4.9: Histogram popisující četnost počtu výskytů konkrétních kombinací obcí v okruzích

obyvatel, zastavěné ploše v hektarech, rozloze v hektarech a průměrném věku. Tyto údaje jsou přidány k záznamům agregovaných dat jako nové atributy POCET OBYVATEL, ZASTAVENE PLOCHY V HA, ROZLOHA V HA a PRUMERNY VEK.

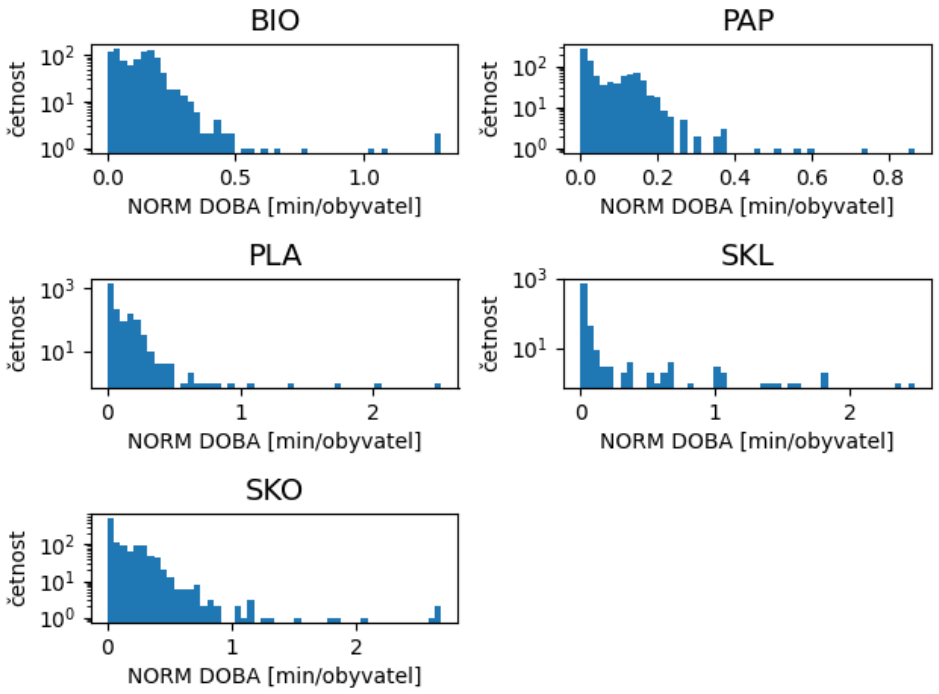
4.3.3. Analýza agregovaných dat

Jak už bylo zmíněno, máme více kvantitativních atributů, které lze zkoumat. Údaje z ČSÚ více analyzovat nemá smysl, protože není jak ověřovat jejich pravdivost. Také jsou tyto údaje stejné pro všechny záznamy obce, takže vyhodnocení hodnoty jako chybné by vedlo k odstranění dat z celé jedné obce, což není žádoucí. Z ostatních atributů jsou pro budoucí model zajímavé MNOZSTVI, DOBA SVOZU, POCET ZASTAVEK a POMER DBD BEZ NAN. POMER CHYB HODNOT by šlo použít pro model popisující data, účelem je ale vytvořit model predikční, u kterého tato veličina ztrácí smysl, protože není vlastností popisující svoz, ale vlastností dat. U MNOZSTVI, DOBA SVOZU, POCET ZASTAVEK je třeba si uvědomit, že jde o veličiny, které jsou závislé na velikosti obce. Tedy pro různé obce mají rozdělení s výrazně různými parametry. U MNOZSTVI z původního datového souboru, šlo toto zanedbat, protože množství odpadu na jednotlivých zastávkách nemá důvod se příliš lišit na základě velikosti obce a je navíc omezeno

kapacitou nádob. U agregovaných veličin toto říci nemůžeme. Je tedy třeba abychom si při průzkumu dat tuto závislost na velikosti obce uvědomovali a identifikaci odlehlých hodnot prováděli po jednotlivých obcích, nebo na normalizovaných variantách veličin. Na základně kvalifikovaného odhadu byla zvolena varianta normalizovaných atributů a byly zavedeny atributy NORM DOBA, která je podílem DOBA SVOZU V MIN a POCET OBYVATEL, a MNOZSTVI NA ZASTAVKU, které je podílem MNOZSTVI a POCET ZASTAVEK.

Histogramy pro NORM DOBA je vidět v 4.10. Pro četnost bylo použito logaritmické měřítko, aby byly snáz viditelné i nízké četnosti potenciálně odlehlých hodnot. Bohužel to naopak utlumí rozdíly v průběhu histogramů. I tak jsou ale u většiny odpadů patrné dva vrcholy. Důvodem jsou patrně dva typy svozu, kde každý typ v dané obci má různě dlouhou trasu a tedy dobu svozu, ale oba jsou děleny stejným počtem obyvatel.

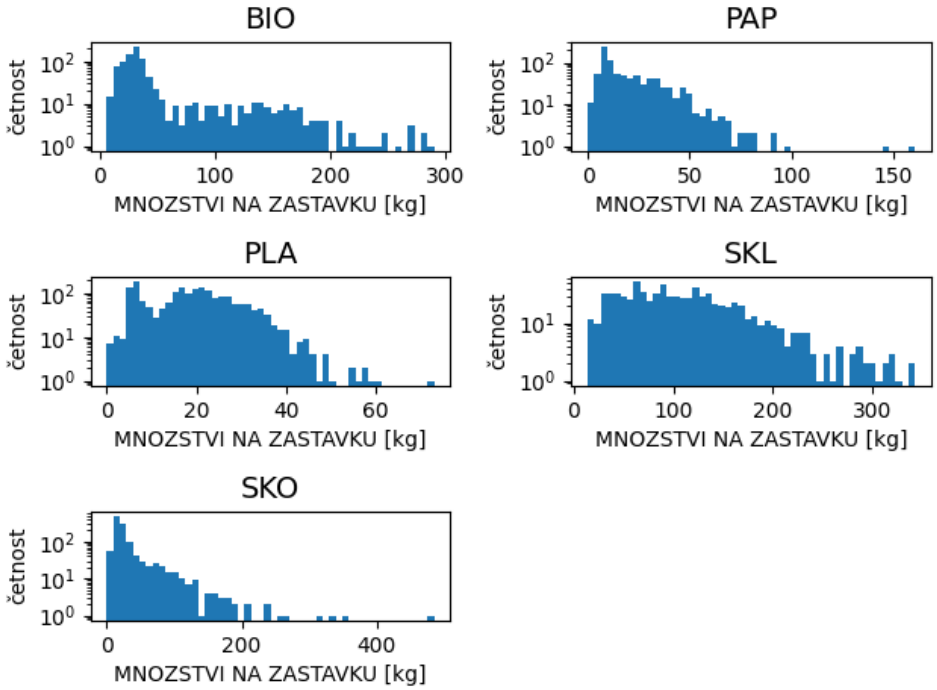
V 4.11 je vidět obdobné histogramy pro MNOZSTVI NA ZASTAVKU. Jde o veličinu



Obrázek 4.10: Histogramy veličiny NORM DOBA v minutách na obyvatele po jednotlivých odpadech. Osa četnosti je logaritmická, aby byly snáz vidět i nízké hodnoty četnosti.

přímo vázanou na MNOZSTVI v neagregovaných datech. Odlehlé hodnoty zde mohly vzniknout například na základě jevu popsaného v 4.2.4, který se z nějakého důvodu nepodařilo zachytit, ani nebyl odfiltrován na základě odlehlosti. Jsou zde vidět také již zmiňované dva vrcholy.

U obou uvedených veličin jsou hodnoty, které jsou podezřelé z odlehlosti. Jelikož rozdě-



Obrázek 4.11: Histogramy veličiny MNOZSTVI NA ZASTAVKU v kg po jednotlivých odpadech. Osa četnosti je logaritmická, aby byly snáz vidět i nízké hodnoty četnosti.

lení ani jedné není na první pohled normální, bude znovu použit robustní přístup z 2.2.3. V hodnotách NORM DOBA jsou jasněji patrné odlehlé hodnoty a kontrola jednotlivých záznamů to potvrzuje. Identifikace odlehlých hodnot je proto provedena u této veličiny jako první. Na základě vypočtených Q hodnot a na základě konzultace byla vybrána kritická hodnota 5. Tím bude označeno 395 hodnot. Poté je proveden podobný postup u veličiny MNOZSTVI NA ZASTAVKU. Zde je hodnota kritické hranice stanovena na 7. Tím je označeno dalších 217 hodnot. Hodnoty kritického Q se od doporučené hranice $2,5$ liší z důvodu odchýlení od normálního rozdělení. Takto identifikované odlehlé hodnoty jsou pro další zpracování dat vyloučeny. V tabulce 4.6 jsou uvedeny počty záznamů podle

jednotlivých typů odpadu, které zbývají.

Atribut POMER DBD BEZ NAN popisuje jaká část z agregovaného záznamu je DBD

Tabulka 4.6: Počty zbylých záznamů pro jednotlivé typy odpadu

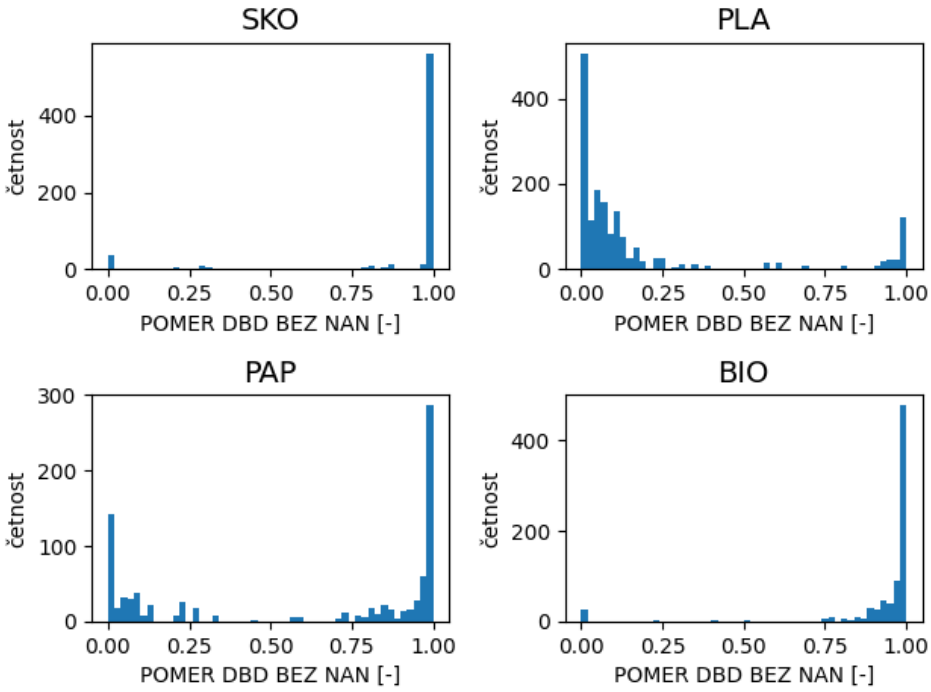
	BIO	PAP	PLA	SKL	SKO
počet záznamů	818	863	1697	753	1020

svozu. Pro hodnotu 1 jsou to všechny původní záznamy se známým typem svozu, a pro hodnotu 0 jsou všechny typu HS. V tomto ale nejsou zahrnuty neoznačené záznamy. Což znamená, že přesnost tohoto atributu pro různé agregované záznamy se liší. To by měl popisovat atribut POMER CHYB HODNOT, ten není použitelný pro předpovědi, nebude tedy do modelu zahrnut. V 4.12 jsou uvedeny histogramy této veličiny. PAP není uveden, protože má svozy pouze HS typu. Je vidět, že hodnoty veličiny jsou blízko hodnoty 0 nebo 1. V modelu se tedy bude chovat podobně jako kategoriální proměnná.

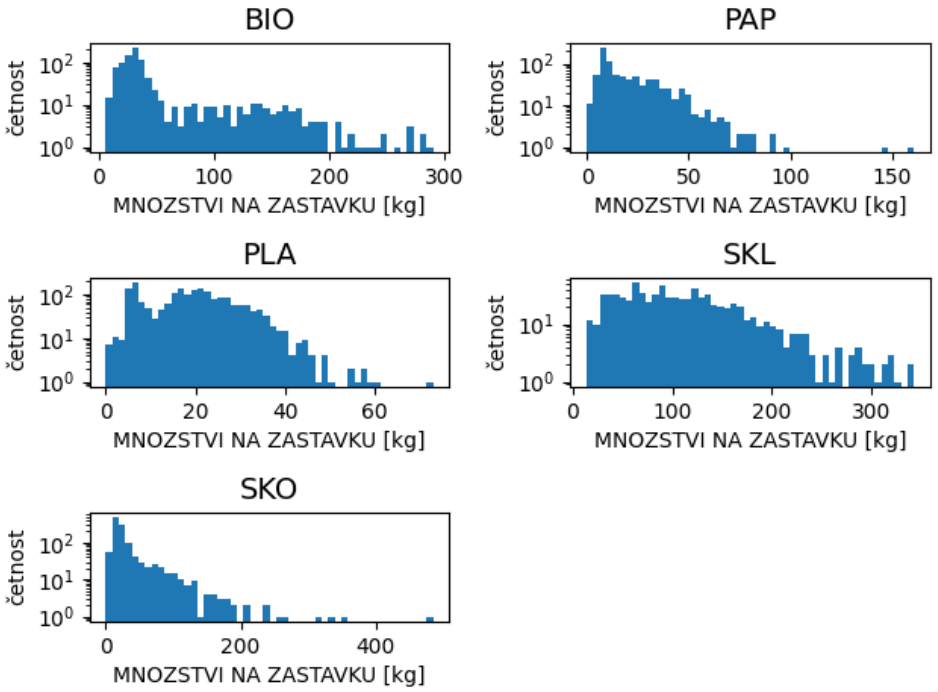
Jelikož v modelu bude jako závislá proměnná zkoumána DOBA SVOZU V MIN, jsou její histogramy také vykresleny. Na první pohled je vidět, že veličina nemá normální rozdělení. Ani se nepodařilo najít vhodnou transformaci.

4.3.4. Korelační analýza

Za účelem zjištění vztahů mezi proměnnými budou vytvořeny korelační matice pro jednotlivé typy odpadu. Jelikož není splněn předpoklad normality, tak není korektní použít Pearsonův korelační koeficient. Je tedy použit Spearmanův pořadový korelační koeficient. V 4.14 je jako příklad uvedena matice pro BIO odpad. Za povšimnutí skupina atributů DOBA SVOZU V MIN, MNOZSTVI a POCET ZASTAVEK a skupina POCET OBYVATEL, ZASTAVENE PLOCHY V HA a ROZLOHA V HA, kdy jsou prvky každé ze skupin mezi sebou silně korelované. Celkově jsou hodnoty korelace vysoké, což naznačuje riziko multikolinearity. Podobné závislosti se objevují u všech typů odpadu.



Obrázek 4.12: Histogramy veličiny POMER DBD BEZ NAN po jednotlivých odpadech.



Obrázek 4.13: Histogramy veličiny DOBA SVOZU V MIN v kg po jednotlivých odpadech.

	DOBA SVOZU V MIN	MNOZSTVI	POCET ZASTAVEK	POMER DBD BEZ NAN	POCET OBYVATEL	ZASTAVENE PLOCHY V HA	ROZLOHA V HA	PRUMERNY VEK	ODPAD NA ZASTAVKU
DOBA SVOZU V MIN	1	0.8632	0.9146	0.1509	0.743391	0.735871	0.696002	0.340916	-0.37452
MNOZSTVI	0.8632	1	0.9608	0.1892	0.588613	0.552416	0.507444	0.187163	-0.17053
POCET ZASTAVEK	0.9146	0.9608	1	0.2045	0.664933	0.636394	0.59416	0.225854	-0.38634
POMER DBD BEZ NAN	0.1509	0.1892	0.2045	1	0.110463	0.027732	-0.02157	-0.2611	-0.18968
POCET OBYVATEL	0.7434	0.5886	0.6649	0.1105	1	0.906451	0.864093	0.380434	-0.32516
ZASTAVENE PLOCHY V HA	0.7359	0.5524	0.6364	0.0277	0.906451	1	0.930813	0.565806	-0.32521
ROZLOHA V HA	0.696	0.5074	0.5942	-0.022	0.864093	0.930813	1	0.597738	-0.30593
PRUMERNY VEK	0.3409	0.1872	0.2259	-0.261	0.380434	0.565806	0.597738	1	-0.01653
ODPAD NA ZASTAVKU	-0.375	-0.171	-0.386	-0.19	-0.32516	-0.32521	-0.30593	-0.01653	1

Obrázek 4.14: Korelační matice pro Spearmanův korelační koeficient pro BIO odpad.

4.4. Model

V této části bude popsán proces tvorby lineárních regresních modelů pro predikci doby svážení odpadu v obcích. Tento predikční model má sloužit jako vstup pro VRP a ARP úlohy.

Jako první byly na základě korelace v 4.3.4 stanoveny základní proměnné. Je vždy vybrána proměnná s nejvyšší hodnotou korelace se závislou proměnnou, kterou je DOBA SVOZU V MIN. Na základě takto vybrané proměnné jsou vybrány další, jejichž korelace k ní nepřesahuje hodnotu $0,6$. Pokud v těchto jsou některé dvě, které mají vzájemnou korelaci vyšší než $0,6$, je z nich ta s nižší korelací vůči DOBA SVOZU V MIN zavrhnuta. Tímto způsobem dostáváme základní regresory pro jednotlivé typy odpadu. Vypsané jsou v 4.7.

Z těchto regresorů spolu s jejich interakcemi a druhými mocninami jsou vytvořeny modely pro každý odpad. Tyto modely jsou rozsáhlé a obsahují nevýznamné regresory. Postupně jsou tedy na základě sestupného výběru [4] odstraňovány regresory s nejvyšší p-hodnotou, ale překračující hodnotu $0,05$. Takto dostanu nové sady regresorů. Vzhledem k použití interakcí a k relativně vysokému počtu regresorů je zde podezření na multikolinearitu. Ta by mohla být vážným problémem v interpretaci modelu. Regresory korelované s jinými nesou stejnou informaci s jinými, což vede k zvýšení váhy této informace a až k fiktivní vysoké hodnotě korelačního koeficientu. Jsou proto vypočteny i hodnoty VIF 2.4.7 pro jednotlivé regresory. V tabulce 4.8 jsou uvedeny maximální hodnoty VIF pro jednotlivé modely. Doporučovaná hranice 5 je zde výrazně překročena. V modelu se tedy vyskytuje

Tabulka 4.7: Proměnné vybrané pro jednotlivé typy odpadu jako základní

typy odpadu	regresory
BIO	POCET ZASTAVEK, POMER DBD BEZ NAN, ROZLOHA V HA, PRUMERNY VEK, MNOZSTVI NA ZASTAVKU
PAP	POCET ZASTAVEK, POMER DBD BEZ NAN, ROZLOHA V HA, PRUMERNY VEK, MNOZSTVI NA ZASTAVKU
PLA	POCET ZASTAVEK, POMER DBD BEZ NAN, POCET OBYVATEL, PRUMERNY VEK, MNOZSTVI NA ZASTAVKU
SKL	POCET ZASTAVEK, ZASTAVENE PLOCHY V HA, PRUMERNY VEK, MNOZSTVI NA ZASTAVKU
SKO	POCET ZASTAVEK, POMER DBD BEZ NAN, POCET OBYVATEL, PRUMERNY VEK, MNOZSTVI NA ZASTAVKU

multikolinearita a je proto třeba snížit dále počet regresorů.

Další snižování počtu regresorů bude založeno na hodnotách VIF, které popisují, jestli

Tabulka 4.8: Počty zbylých regresorů, korigované koeficienty determinace po sestupném výběru a nejvyšší hodnotay VIF

	BIO	PAP	PLA	SKL	SKO
počet regresorů	13	6	12	7	10
korigovaný koeficient determinace	0,92	0,82	0,74	0,77	0,86
nejvyšší hodnota VIF	8 448	18	5 837	24	35

je informace obsažená v daném regresoru obsažena i ve zbytku modelu. Bude vždy odstraněna proměnná s nejvyšší hodnotou VIF a model bude přepočítán. Pokud v některé iteraci překročí některá p-hodnota hranici $0,05$, je příslušná proměnná odstraněna přednostně. Postup skončí, jakmile se všechny hodnoty dostanou pod hranici 3. Údaje o takto získaných modelech jsou v 4.9.

Výsledné modely budou otestovány na průměrech veličin jednotlivých obcí. Z takto vzniklých průměrů jsou napočítány i potřebné interakce a druhé mocniny. Tato průměrovaná data jsou dosazena do modelu a výsledné predikce jsou porovnány s průměrnou hodnotou DOBA SVOZU V MIN. Výsledek je vidět na obrázku 4.15. Z grafů je patrné, že nejlepší odhad má model pro BIO, PAP a SKO. PLA a SKL jsou méně přesné. Je ale

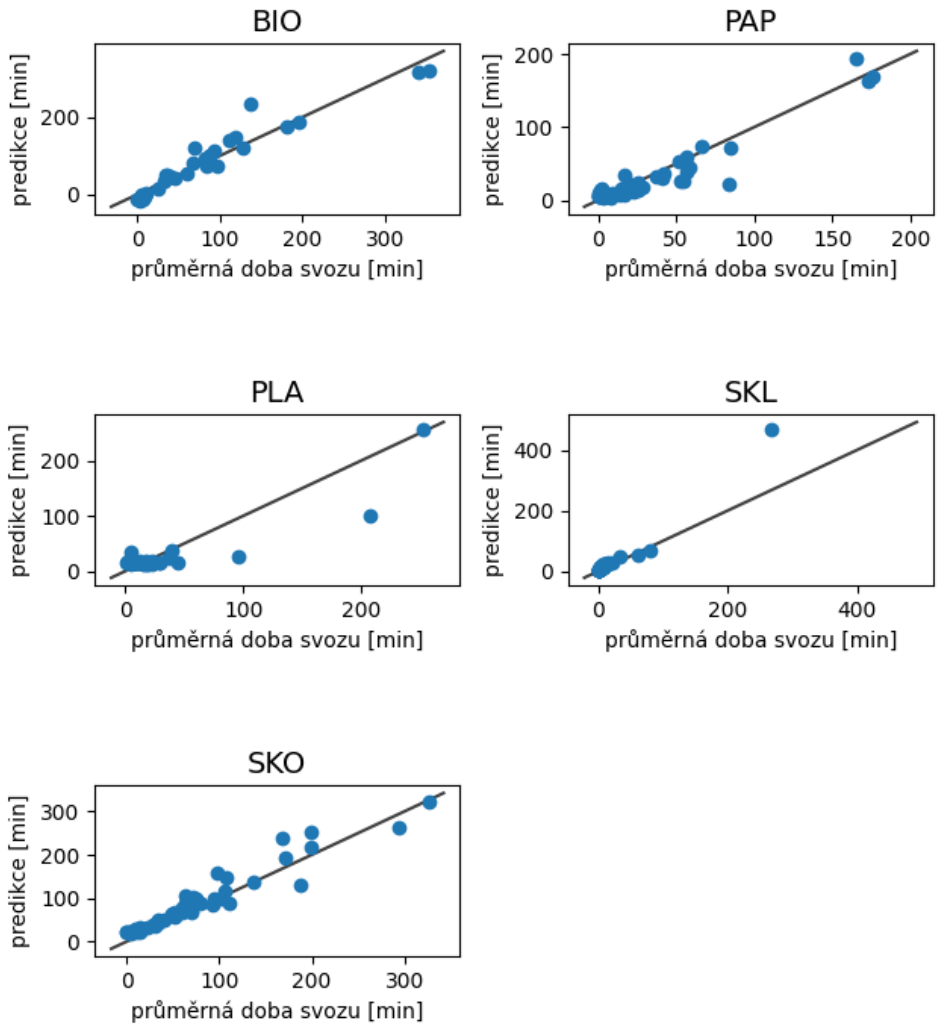
Tabulka 4.9: Regresory, korigované koeficienty determinace a nejvyšší hodnoty VIF pro výsledné modely

	BIO	PAP	PLA	SKL	SKO
konstanta	-1281,71	-398,865			
POCET ZASTAVEK				5,490131	0,67851
PRUMERNY VEK			0,306941		0,339499
ROZLOHA V HA	1077,732				
('POCET ZASTAVEK', 'POCET ZASTAVEK')	0,00069		0,00199		
('POCET ZASTAVEK', 'MNOZSTVI NA ZASTAVKU')	0,017318				
('PRUMERNY VEK', 'MNOZSTVI NA ZASTAVKU')	-0,00332				
('POCET ZASTAVEK', 'ROZLOHA V HA')		0,746649			
('POMER DBD BEZ NAN', 'ROZLOHA V HA')		-18,9114			
('ROZLOHA V HA', 'ROZLOHA V HA')		277,7194			
('ZASTAVENE PLOCHY V HA', 'PRUMERNY VEK')				-0,22075	
('POCET OBYVATEL', 'POCET OBYVATEL')			7,33E-06		2,11E-06
('POMER DBD BEZ NAN', 'POCET OBYVATEL')			-0,04402		
korigovaný koeficient determinace	0,85	0,7	0,7	0,71	0,85
nejvyšší hodnota VIF	2,36	1,44	1,71	1,94	1,38

vidět, že PLA a SKL mají nejnerovnoměrně rozložené datové soubory. Průměrovaná data v podstatě vychází z dat, na kterých se modely učily, takže je logické, že zde vidíme horší data u méně přesných modelů.

4.4.1. Analýza modelu

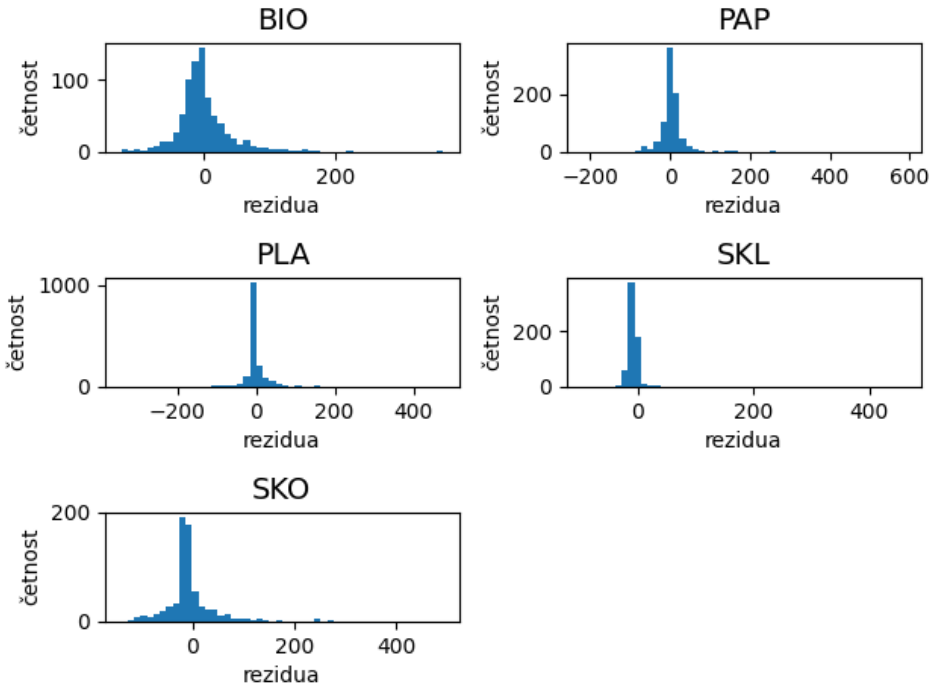
Dosud byly modely studovány pouze z hlediska korigovaného koeficientu determinace, p-hodnot a hodnot VIF. Tedy z hlediska míry vysvětlení variability závislé proměnné a z hlediska přítomnosti multikolinearity. Dalším důležitým pohledem je studium reziduí. V něm lze nalézt pozorování odlehlá od predikce modelu. Histogramy standardizovaných reziduí jsou uvedeny v 4.16 a Q-Q grafy v 4.17. Z grafů je patrné odchýlení od normálního rozdělení a také přítomnost odlehlých hodnot. Odstraněním těchto odlehlých záznamů by došlo ke zlepšení koeficientu determinace a přiblížení reziduí k normalitě. Toto je možné udělat pouze po prouzkoumání záznamů a jejich identifikaci jako chybných. V opačném případě by to vedlo k přizpůsobování dat modelu. Po analýze zmiňovaných záznamů se



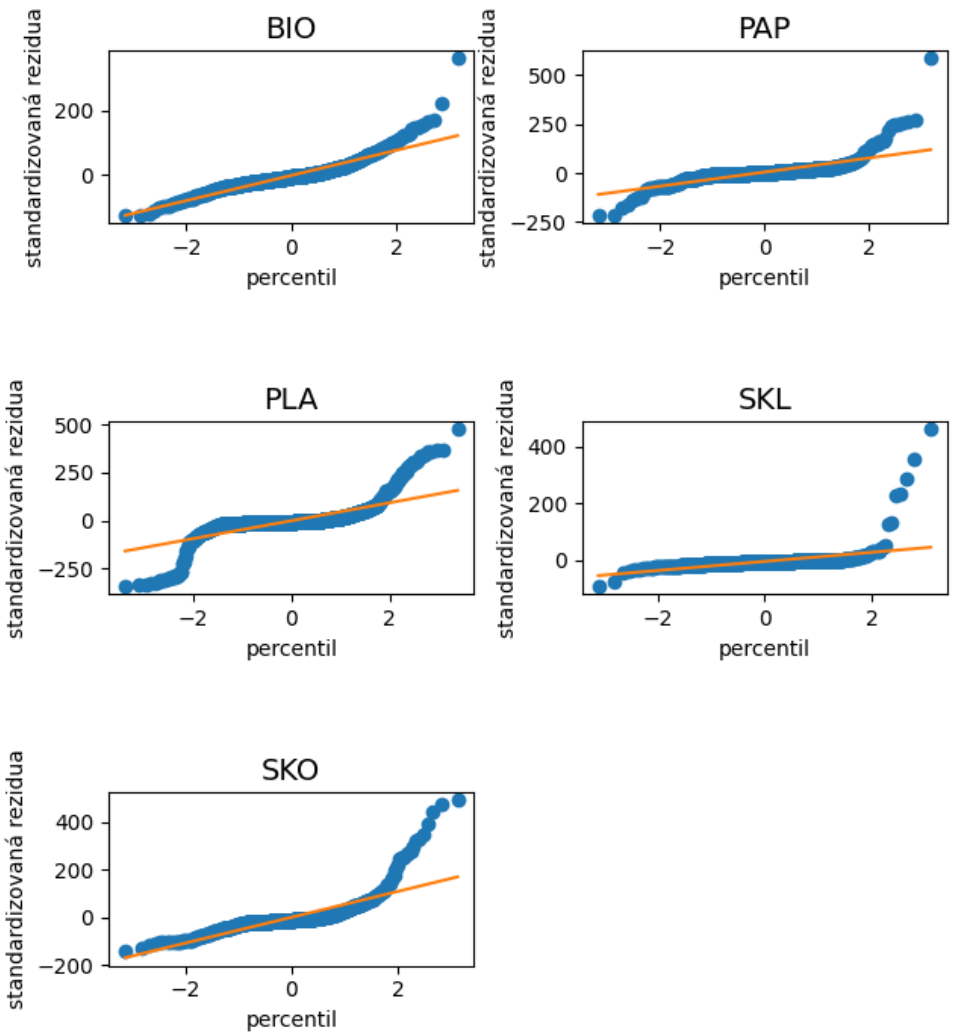
Obrázek 4.15: Graf srovnávající predikce vůči průměrným dobám svozu. Úsečkou je naznačen graf $y = x$

jeví, že jsou v pořádku. Na základě dostupných údajů je nelze vyhodnotit jako vadné. Může jít o lokální nenáhodné jevy, které z dat lze obtížně identifikovat.

Důsledkem nesplnění normality reziduí je, nekorektnost použití intervalových odhadů. Dále to vede k problémům s korektností testů významnosti jednotlivých koeficientů. Je možné se ale na lineární regresi dívat z pohledu numerického, z něhož mají modely s vysokými koeficienty determinace jistě vypovídající hodnotu. Výpočet koeficientu determinace totiž nevyžaduje normální rozdělení. Cílem je zde vytvořit bodový odhad s důrazem na použitelnost, numerický pohled je tedy dostačující.



Obrázek 4.16: Histogramy standardizovaných reziduí pro jednotlivé typy odpadu.



Obrázek 4.17: Q-Q grafy pro normální rozdělení standardizovaných reziduí pro jednotlivé typy odpadu.

4.5. Doporučení

Obdržena byla surová reálná data, na kterých před tím nebyla prováděna analýza dat ani preprocesing hodnot. Tato data nebyla z technických důvod homogenní v informacích. Cílem bylo tedy využít co nejvíce dostupná data a případně chybějící informace odhadnout. Na základě tohoto zpracování byly vytvořeny návrhy pro změnu metodiky zápisu dat, které by do budoucna mohly usnadnit zpracování sbíraných dat.

- Vytvořit identifikační číslo pro jednotlivé cesty vozidel. Toto by umožnilo výrazně snazší agregaci dat z jednotlivých svozů. Na toto identifikační číslo by mohly být

navázany informace, které jsou stejné pro celý svoz, například typ odpadu a typ svozu. Přestože jsou někdy během jedné cesty sbírány svozy obou typů, tak je ve většině případů jeden dominantní. A takto by tuto informaci bylo možné mít k dispozici i pro svozy, na kterých se nezapisuje NAZEV STANOVISTE.

- Id cest by mohlo být nahrazeno systémem zaznamenávání poslední zastávky cesty. To by umožnilo jednoznačně rozdělit záznamy podle jednotlivých cest. Id cest by v tu chvíli mohlo být zavedeno i dodatečně.

- Zaznamenávání frekvence svozu. Jde o údaj, který nebylo možné z dat přesně strojově vyčíst, ale přitom může jít o významnou proměnnou do modelu. Tuto informaci jde mnohem snáze zaznamenat při plánování svozu než poté vyčíst z dat.

- Rozdělení dat podle modernizace vozidel. V jednom případě jde o data zaznamenávána po zastávkách, v druhém ale už jde o agregovaná data, která navíc nemají uvedený údaj o čase. Je tedy vhodné tyto dvě skupiny oddělit parametrem nebo databázovými tabulkami.

- Informace o počtu nádob každého typu v obci. Z dat není zcela jasné, kolik je v obci nádob a kterého typu jsou. Není totiž vždy svážena celá obec a u části nádob chybí identifikační číslo, aby je bylo možné rozlišit. Tento údaj ale může být také pro model významný. Umožňuje pracovat s informací, jak je mezi který typ svozu produkce odpadu obce rozložena. Taky by dávalo smysl, že by počet obyvatel na nádobu ovlivňoval rychlost jejího zaplnění.

- Zavést databázi kategorií. V zápisu dat se sem tam objevují překlepy, mohlo by tedy pomoct mít databázi správných podob zápisu, na základě které by se hodnoty kontrolovaly. Předešlo by se potom rozdělení kategorie na víc fiktivních.

5. Závěr

V první kapitole je shrnuta situace společnosti TSMH s r. o. a nastíněno směřování práce.

V kapitole druhé následuje soubor teoretických poznatků v práci potřebných. Hlavně tedy testy hypotéz, analýza odlehlých hodnot, korelace a zavedení lineárního regresního modelu.

V další kapitole jsou popsány použité softwarové nástroje a důvody pro jejich použití.

V kapitole čtvrté je již samotné zpracování dat. Nejdříve jsou data popsána a vysvětlena. Poté přichází preprocesing dat, ve kterém jsou opraveny překlepy v kategoriích a jsou identifikovány chybějící hodnoty. Následně je již možné provést sofistikovanější analýzu jevů v datech, analýzu odlehlých hodnot a odhad části chybějících hodnot. Tyto data jsou agregována do svozů za obce a jsou doplněna o socioekonomické údaje ze stránek Českého statistického úřadu. Tento výsledný soubor dat je podroben analýze odlehlých hodnot. Na pročištěných datech následuje korelační analýza, na základě které jsou vybrány významné proměnné pro model. Z nich jsou vytvořeny první lineární regresní modely. Ty jsou dále vylepšovány. Výsledné modely popisují alespoň 70 % variability a některé dosahují až 85 %. Na závěr jsou zkoumána rezidua modelů. V té je bohužel zjištěno, že není splněn předpoklad normality reziduí, není možné tedy korektně provést některé testy a intervalové odhady. Přesto je model použitelný pro predikci. Budeme zde chápat lineární regresi jako numerický nástroj. Na závěr jsou sepsány postřehy a doporučení ke sběru dat.

Kromě doporučení ke sběru dat a lineárního regresního modelu je výstupem této práce i postup zpracování dat, ze kterého bude možno vyjít i v budoucnu při vylepšování modelů na základě nových dat. Část doporučení již byla společností TSMH zohledněna. Jestli budou použita zbylá a jestli bude kvalita modelu dostčující pro VRP a ARP modely, ukáže až budoucnost.

Literatura

[1] HEBÁK, Petr a Jiří HUSTOPECKÝ. *Průvodce moderními statistickými metodami.* Praha: Státní nakladatelství technické literatury, 1990

[2] MELOUN, Milan a Jiří MILITKÝ. *Statistická analýza experimentálních dat.* Vyd. 2., upr. a rozš. Praha: Academia, 2004. ISBN 80-200-1254-0.

[3] ANDĚL, Jiří. *Základy matematické statistiky.* Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

[4] ZVÁRA, Karel. *Regrese.* Vyd. 1. Praha: Matfyzpress, 2008. ISBN 978-80-7378- 041-8

[5] HEBÁK, Petr, Jiří HUSTOPECKÝ, Eva JAROŠOVÁ a Ivana MALÁ. *Vícerozměrné statistické metody.* Praha: Informatorium, 2005. ISBN 80-7333-025-3.

[6] Marsaglia, John & Marsaglia, George. (2004). *Evaluating the Anderson-Darling Distribution.* Journal of Statistical Software. 09. 10.18637/jss.v009.i02.

[7] Rousseeuw, Peter & Hubert, Mia. (2011). *Robust statistics for outlier detection.* Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery. 1. 73-79. 10.1002/widm.2

[8] JAMES, Gareth. *An introduction to statistical learning: with applications in R.* New York: Springer, 2013. Springer texts in statistics. ISBN 978-1-4614-7137-0.

[9] NEVRLÝ, Vlastimír. Modely a metody pro svozové úlohy. Brno, 2016. Diplomová práce. Vysoké učení technické v Brně. Vedoucí práce RNDr. Pavel Popela, Ph.D.

[10] *ČSÚ a územně analytické podklady. Český statistický úřad* [online]. 2021 [cit. 2021-5-20]. Dostupné z: https://www.czso.cz/csu/czso/csu_a_uzemne_analyticke_podklady

[11] *Technické služby Malá Haná* [online]. Boskovice, 2021 [cit. 2021-5-20]. Dostupné z: https://www.tsmh.cz/

[12] *Python Software Foundation. Python Language Reference, 3.6.* Available at http://www.python.org