VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ

FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY

INSTITUTE OF MATHEMATICS

STATISTICKÁ ANALÝZA DAT PRO SVOZOVÉ ÚLOHY

STATISTICAL ANALYSIS OF DATA FROM ROUTING PROBLEMS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — **LENKA MAZLOVÁ**

AUTHOR

VEDOUCÍ PRÁCE — **Ing. RADOVAN ŠOMPLÁK, Ph.D.**

SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadaní bakalářské práce

Ústav:	Ústav matematiky
Studentka:	**Lenka Mazlová**
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 bakalářské práce:

Statistická analýza dat pro svozové úlohy

Stručná charakteristika problematiky úkolu:

Zpracování provozních dat z oblasti odpadového hospodářství představuje důležitou roli pro výpočty logistických řetězců. Zásadní oblastí řetězce nakládání s odpadem je jeho svoz. Ten může představovat polovinu i více nákladů souvisejících se zpracováním odpadu. Svozové úlohy se zabývají problematikou logistiky v nejnižší možné úrovni detailu. Výsledky modelů jsou však velice citlivé na vstupní údaje. Jejich přesnost a určení reálných hodnot tak přestavuje klíčovou část příprav před samotným výpočtem. Práce se bude zabývat vyhodnocením variability logistických dat a analýzou citlivosti výsledků na vstupní údaje. K analýze budou využity intervalové odhady, testování statistických hypotéz, ověřování nutných předpokladů a další statistické metody. Postup je rozdělen na přípravnou fázi zahrnující zpracování dat v rozsáhlých databázích, které jsou získávány při monitoringu reálného provozu svozových vozidel. Druhá část práce se bude zabývat komplexním pohledem na jednotlivé svozové trasy – doba transportu, časy obsluhy, naplněnost sběrných nádob, kapacitní vytížení vozidel a další parametry související se svozovými úlohami.

Cíle bakalářské práce:

Identifikování klíčových parametrů, které ovlivňují výsledky svozových úloh.
Rozšíření znalostí ve vybraných oblastech matematické statistiky.
Analýza historických dat z reálného monitoringu svozových vozů.
Sestavení statistických modelů a testování jejich významnosti.

Seznam doporučené literatury:

WILLIAMS, H. P. Model building in mathematical programming. 5th ed. New York: John Wiley and Sons. 411 s., 2013. ISBN 978-1-118-44333-0.

NEVRLÝ, V. Modely a metody pro svozové úlohy. Diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, Ústav Matematiky, 48, Vedoucí práce RNDr. Pavel Popela, Ph.D., 2016. Dostupné z: https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=128582

LAPORTE, G. Fifty Years of Vehicle Routing. Transportation Science, 43(4), 2009. DOI: https://doi.org/10.1287/trsc.1090.0301.

Termín odevzdání bakalářské 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
Táto bakalárska práca je zameraná na štatistickú analýzu dát získaných z reálnych zvozov odpadu. Jej hlavným cieľom je oboznámiť sa s testovaním hypotéz a tvorbou modelov typickým pre matematickú štatistiku a určiť kľúčové parametre ovplyvňujúce zvozové úlohy. Jedným z výstupov tejto práce je súbor programu Microsoft Excel, v ktorom sú uskutočnené štatistické testy popísané v texte.

Summary
This bachelor thesis focuses on statistical analysis of data retrieved from real waste collection. The main goal is to become familiar with statistical hypothesis testing and model creating typical for mathematical statistics and to determine key parameters that affect routing problems. One of the outputs of this thesis is a Microsoft Excel file that contains statistical tests described in following text.

Klíčová slova
zber odpadu, biologicky rozložiteľný odpad, zvozové úlohy, testovanie hypotéz, Dixonov test, kĺzavý priemer, analýza rozptylu - ANOVA

Keywords
waste collection, biodegradable waste, routing problems, hypothesis testing, Dixon's test, moving average, Analysis of Variance - ANOVA

MAZLOVÁ, L. *Statistická analýza dat pro svozové úlohy.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2021. 48 s. Vedoucí Ing. Radovan Šomplák, Ph.D.

Prehlasujem, že svoju bakalársku prácu na tému *Statistická analýza dat pro svozové úlohy* som spracovala samostatne pod vedením Ing. Radovana Šompláka, Ph.D. s použitím materiálov uvedených v zozname použitej literatúry.

Lenka Mazlová

Poďakovanie patrí môjmu vedúcemu Ing. Radovanovi Šomplákovi, Ph.D. za odborné vedenie a rady pri spracovávaní mojej bakalárskej práce a hlavne za dodávanie motivácie v priebehu celého zostavovania tejto práce. Takisto by som sa chcela poďakovať Ing. Vlastimírovi Nevrlému, Ph.D. za pomoc pri spracovaní praktickej časti. V neposlednom rade by som sa chcela poďakovať spoločnosti Technické služby Malá Haná s.r.o. za poskytnutie dát skúmaných v tomto texte.

Lenka Mazlová

Obsah

Úvod

Rozvoj spoločnosti so sebou prináša nielen úžitok, ale aj negatívne dopady na životné prostredie. S každou ľudskou činnosťou je spojená produkcia odpadu. Tlak na efektívne nakladanie s odpadmi sa stupňuje predovšetkým vo vyspelých krajinách. Cieľom je, aby bol potenciál ukrytý v odpadoch využitý, čo vedie k zníženiu dopytu po primárnych zdrojoch.

Na prvom mieste je obmedzenie produkcie odpadu, tzv. prevencia pred vznikom. Keď už je odpad vyprodukovaný, tak preferovanou formou nakladania s ním je jeho materiálové využitie a ak to nie je možné, tak jeho energetické využitie [3]. V rámci snahy o šetrnejšie nakladanie s komunálnym odpadom (ďalej len KO) voči životnému prostrediu vznikol v rámci Európskej únie balíček k obehovému hospodárstvu, ktorý obsahuje štyri smernice. Tieto smernice sa zameriavajú na zvyšovanie podielu materiálového využitia KO ako celku [24], na obalové odpady [25], v smernici [23] je obmedzenie množstva KO ukladaného na skládku. Posledná smernica [22] sa venuje problematike autovrakov. Tieto smernice sa postupne prepisujú do legislatívy jednotlivých členských štátov. Česká republika nie je výnimkou.

V roku 2021 bol prijatý nový zákon o odpadoch [4], ktorý stanovuje nové pravidlá v oblasti odpadového hospodárstva. Tieto zmeny so sebou prinášajú potenciálne výrazné navýšenie nákladov pre obce. Logickým dôsledkom sú snahy o znižovanie prevádzkových nákladov. Jednou z najväčších častí na celkovom nákladovom koláči je zber odpadu a jeho preprava [9]. Aktuálne sa v Českej republike začínajú tvoriť nové zväzky obcí s cieľom komplexného plánovania pre zaistenie efektívnejšieho systému zberu a nakladania s odpadom pre zníženie ekonomickej záťaže na obce. V tejto, pre obce novej situácii je potrebné vytvoriť úplne nové zvozové plány, rozmiestniť nové zberné nádoby (zvýšenie separácie alebo pridávanie nových frakcií triedeného KO) a spravodlivo prerozdeliť náklady pre jednotlivých členov zväzku. Z týchto dôvodov sa zvyšuje dopyt po monitoringu spracovateľského reťazca, kam spadá váženie zberných nádob a áut, odhad naplnenosti zberných nádob, dopravné informácie o vozidle, dáta o produkcii odpadov v priebehu roku a iné.

Motivácia pre túto bakalársku prácu vznikla na základe komunikácie s konkrétnym zväzkom obcí Technické služby Malá Haná s.r.o. Tento zväzok poskytol pre štatistickú analýzu prevádzkové dáta (od roku 2019). Dynamickému vývoju v počiatkoch fungovania zväzku odpovedá aj rastúca kvalita dát z monitoringu. Jadrom tejto bakalárskej práce je teda analyzovať dostupné dáta od približne 50 obcí so zameraním na testovanie rôznorodosti produkcie jednotlivých frakcií KO v priebehu roku. Cieľom je získať väčší náhľad do väzieb v dátach, ktorý bude možné využiť k ďalšiemu zlepšeniu v strategickom a prevádzkovom plánovaní zväzku.

Bakalárska práca je štruktúrovaná nasledovne. Prvú kapitolu bakalárskej práce tvorí matematický aparát. V druhej kapitole je predstavený softvér, v ktorom sú poskytnuté dáta spracované. V tretej kapitole sú popísané dáta, ktoré sú predmetom nasledujúcich štatistických analýz v kapitole štvrtej. Piata kapitola obsahuje zhrnutie dosiahnutých výsledkov. V tejto kapitole sú diskutované najvýznamnejšie postrehy a odporúčania pre ďalší rozvoj tejto práce. Posledná kapitola sa venuje záverečnej rekapitulácii celej bakalárskej práce.

1 Matematický aparát

V tejto časti budú definované základné pojmy matematického aparátu, ktoré budú následne využívané v rámci časti riešiacej praktické úlohy spojené s produkciou odpadu a jeho logistikou (viď kapitola 4).

1.1 Úvod do pravdepodobnosti

Pojmy predstavené v tejto podkapitole boli čerpané z [13], [15] a [18].

Pod pojmom *náhodný jav* rozumieme výsledok pokusu (t. j. realizáciu určitého systému podmienok), ktorého charakteristickým rysom je, že môže, ale nemusí nastať.

Definícia 1.1. Jav A nazveme *elementárnym javom*, ak neexistujú dva rôzne javy, ktorých zjednotením dostaneme jav A – tento jav je teda najjednoduchším výsledkom náhodného pokusu a značíme ho ω alebo ω_i.

Definícia 1.2. Množinu všetkých elementárnych javov, ktoré môžu nastať ako výsledok daného náhodného pokusu, nazývame *základný priestor* alebo taktiež priestor všetkých elementárnych javov a značíme ho Ω.

Definícia 1.3. Nech Ω je základný priestor priradený náhodnému pokusu. Neprázdny systém podmnožín $\mathcal{A}$ základného priestoru Ω, ktorý spĺňa:

1. $\Omega \in \mathcal{A}$,
2. $A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A}$,
3. $A_1, A_2, ... \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$,

sa nazýva *javová σ-algebra* a množina $A \in \mathcal{A}$ sa nazýva *jav*. Dvojica $(\Omega, \mathcal{A})$ sa nazýva *javové pole*.

Definícia 1.4 (Kolmogorovova axiomatická definícia pravdepodobnosti)**.** Nech $(\Omega, \mathcal{A})$ je javové pole príslušné uvažovanému pokusu. Potom zobrazenie P, ktoré každému javu $A \in \mathcal{A}$ priraďuje číslo $P(A)$ nazveme pravdepodobnosťou na javovom poli $A \in \mathcal{A}$, ak toto zobrazenie vyhovuje nasledujúcim axiómom:

1. $P(A) \geq 0$ pre každý jav $A \in \mathcal{A}$ (nezápornosť),
2. $P(\Omega) = 1$ (normovanosť),
3. Ak je $A_1, A_2, A_3, ...$ konečná (resp. spočítateľná) postupnosť po dvoch disjunktných javoch z $\mathcal{A}$ (t. j. $A_i \in \mathcal{A}$, $A_i \cap A_j = \emptyset$, $i \neq j$; $i, j = 1, 2, ...$), tak pre konečnú postupnosť javov $A_1, A_2, A_3, ..., A_n$ platí (aditivita):

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i),$$

resp. pre spočítateľnú postupnosť javov $A_1, A_2, A_3, ...$ platí (σ-aditivita):

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Pre daný jav A nazývame hodnotu $P(A)$ *pravdepodobnosťou javu* A. Trojicu $(\Omega, \mathcal{A}, P)$ nazývame *pravdepodobnostný priestor*.

Poznámka. Systém všetkých podmnožín $\mathbb{R}^n$ obsahuje nemerateľné množiny, na ktorých nemôžeme zaviesť pravdepodobnosť. Preto sa zavádza *borelovské pole* $\mathcal{B}_n$, ktorého prvky B nazývame *n-rozmerné borelovské množiny*. Dvojica $(\mathbb{R}^n, \mathcal{B}_n)$ sa nazýva *merateľný priestor*.

1.2 Náhodná veličina

Definície, vlastnosti a poznámky k náhodným veličinám boli prebraté z [13] [15] a [18].

Definícia 1.5. Nech $(\Omega, \mathcal{A}, P)$ je pravdepodobnostný priestor. Zobrazenie $X : \Omega \longrightarrow \mathbb{R}$ sa nazýva *náhodná veličina vzhľadom k* $\mathcal{A}$ práve vtedy, keď pre $\forall x \in \mathbb{R}$ platí:

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}.$$

Obraz $X(\omega)$ sa nazýva *číselná realizácia náhodnej veličiny* X *príslušná možnému výsledku* ω. *Obor hodnôt* náhodnej veličiny X značíme M.

Poznámka. Vieme dokázať, že zobrazenie $X : \Omega \longrightarrow \mathbb{R}$ je náhodnou veličinou vzhľadom k $\mathcal{A}$ práve vtedy, keď úplný vzor ľubovoľnej borelovskej množiny $B \in \mathcal{B}$ v zobrazení X je náhodný jav vzhľadom k $\mathcal{A}$, t. j. keď pre $\forall B \in \mathcal{B}$ platí:

$$X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}.$$

Definícia 1.6. Nech $(\Omega, \mathcal{A}, P)$ je pravdepodobnostný priestor a X je náhodná veličina definovaná na javovom poli $(\Omega, \mathcal{A})$. Potom funkciu $F_X(x) = P(X \leq x)$ definovanú pre každé $x \in \mathbb{R}$ nazývame *distribučnou funkciou náhodnej veličiny* X.

Vlastnosti:

1. $0 \leq F(x) \leq 1$, $\forall x \in \mathbb{R}$,
2. $F(x)$ je neklesajúca funkcia,
3. $F(x)$ je sprava spojitá funkcia,
4. $\lim_{x\to\infty} F(x) = 1$ a $\lim_{x\to-\infty} F(x) = 0$,
5. pre ľubovoľné $x_1, x_2 \in \mathbb{R}$, $x_1 < x_2$ platí $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$,
6. pre každé $x \in \mathbb{R}$ platí $P(X = x) = F(x) - \lim_{y\to x_-} F(y)$.

Poznámka. Distribučná funkcia F_X náhodnej veličiny X na $(\Omega, \mathcal{A}, P)$ má najviac spočítateľne veľa bodov nespojitosti.

Definícia 1.7. Nech $(\Omega, \mathcal{A}, P)$ je pravdepodobnostný priestor a nech X je náhodná veličina vzhľadom k $\mathcal{A}$. Funkcia $P_X : \mathcal{B} \to \mathbb{R}$, definovaná vzťahom $P_X(B) = P(X \in B)$, $\forall B \in \mathcal{B}$, sa nazýva *rozdelenie pravdepodobnosti náhodnej veličiny* X.

Definícia 1.8. Náhodná veličina X je *diskrétna* a hovoríme, že má *diskrétne rozdelenie pravdepodobnosti*, ak nadobúda najviac spočítateľne veľa hodnôt x. Jej *pravdepodobnostná funkcia* je postupnosť, ktorá pre $\forall x$ spĺňa:

$$p(x) = P(X = x) > 0.$$

Vlastnosti:

1. $p(x) \geq 0$, $\forall x \in \mathbb{R}$ (nezápornosť),
2. $\sum_x p(x) = 1$ (normovanosť),
3. $F(x) = \sum_{t<x} p(t)$, $\forall x \in \mathbb{R}$,
4. $P(X \in M) = \sum_{x \in M} p(x)$ pre ľubovoľnú množinu reálnych čísel M.

Poznámka. Distribučná funkcia diskrétnej náhodnej veličiny má „schodovitý“ tvar.

Definícia 1.9. Náhodná veličina X je *spojitá* a hovoríme, že má *spojité rozdelenie pravdepodobnosti*, ak má spojitú distribučnú funkciu $F(x)$ pre $\forall x \in \mathbb{R}$. Jej *hustota pravdepodobnosti* je taká nezáporná, po častiach spojitá funkcia $f(x)$, že pre $\forall x \in \mathbb{R}$ platí:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Vlastnosti:

1. $f(x) \geq 0$, $\forall x \in \mathbb{R}$ (nezápornosť),
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$ (normovanosť),
3. $f(x) = F'(x)$, ak derivácia existuje,
4. $P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = \int_a^b f(x)\, dx = F(b) - F(a)$ pre ľubovoľné $a, b \in \mathbb{R}$, kde $a \leq b$,
5. $P(X = c) = 0$ pre ľubovoľné reálne číslo c.

1.2.1 Číselné charakteristiky náhodnej veličiny

Definícia 1.10. Nech X je náhodná veličina na $(\Omega, \mathcal{A}, P)$ s distribučnou funkciou F_X. Potom *stredná hodnota náhodnej veličiny* X je

$$E(X) = \int_{-\infty}^{\infty} x\, dF(x),$$

ak tento integrál existuje a je konečný. Ak integrál nie je konečný alebo neexistuje, hovoríme, že stredná hodnota $E(X)$ neexistuje.

Poznámka. Stredná hodnota charakterizuje polohu rozdelenia pravdepodobnosti náhodnej veličiny X. Ak je X diskrétna náhodná veličina s pravdepodobnostnou funkciou $p(x)$, tak $E(X) = \sum_{x \in M} x \cdot p(x)$ – ak rad konverguje absolútne. Ak je X spojitá náhodná veličina s hustotou $f(x)$, tak $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$ – ak integrál konverguje absolútne.

Definícia 1.11. Nech X je náhodná veličina na $(\Omega, \mathcal{A}, P)$. *Rozptyl náhodnej veličiny* X je

$$D(X) = E[X - E(X)]^2$$

za predpokladu, že stredná hodnota existuje. Druhá odmocnina z rozptylu sa nazýva *smerodajná odchýlka náhodnej veličiny* X.

Poznámka. Rozptyl vyjadruje mieru kolísania hodnôt náhodnej veličiny X okolo jej strednej hodnoty. Ak je X diskrétna náhodná veličina s pravdepodobnostnou funkciou $p(x)$, tak $D(X) = \sum_{x \in M}(x - E(X))^2 \cdot p(x) = \sum_{x \in M} x^2 \cdot p(x) - (E(X))^2$ – ak rad konverguje. Ak je X spojitá náhodná veličina s hustotou $f(x)$, tak $D(X) = \int_{-\infty}^{\infty}(x - E(X))^2 \cdot f(x)\,dx = \int_{-\infty}^{\infty} x^2 \cdot f(x) - (E(X))^2\,dx$ – ak integrál konverguje.

1.2.2 Normálne rozdelenie

Zápisom $X \sim N(\mu, \sigma^2)$ rozumieme, že náhodná veličina X má *normálne rozdelenie* s parametrami $\mu, \sigma^2 \in \mathbb{R}$, $\sigma^2 > 0$. Hustota tohto rozdelenia je popísaná funkciou:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ - \frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R},$$

kde číselné charakteristiky sú $E(X) = \mu$ a $D(X) = \sigma^2$.

Poznámka. Toto najčastejšie využívané rozdelenie, nazývané taktiež *Gaussovo rozdelenie*, má množstvo významných teoretických vlastností a z hľadiska aplikácií býva vhodné k vyjadreniu náhodných veličín, ktoré sme schopní interpretovať ako aditívny výsledok veľkého množstva nezávislých vplyvov.

1.3 Náhodný vektor

Pri spracovávaní tejto podkapitoly sme čerpali z [2], [13] a [15].

Definícia 1.12. Nech $X_1, ..., X_n$ sú náhodné veličiny definované na $(\Omega, \mathcal{A}, P)$. Potom vektor $\boldsymbol{X} = (X_1, ..., X_n)^T$ nazývame *náhodný vektor*.

Veta 1.13. Vieme dokázať, že zobrazenie $\boldsymbol{X} : \Omega \longrightarrow \mathbb{R}^n$ sa nazýva náhodný vektor práve vtedy, keď pre $\forall \boldsymbol{x} \in \mathbb{R}^n$ platí:

$$\{\omega \in \Omega : \boldsymbol{X}(\omega) \leq \boldsymbol{x}\} \in \mathcal{A},$$

alebo ekvivalentne, keď pre $\forall B \in \mathcal{B}_n$ platí:

$$\{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\} \in \mathcal{A}.$$

Definícia 1.14. Nech $\boldsymbol{X} = (X_1, ..., X_n)^T$ je náhodný vektor na $(\Omega, \mathcal{A}, P)$. Funkcia $P_{\boldsymbol{X}} : \mathcal{B}_n \to \mathbb{R}$ daná predpisom $P_{\boldsymbol{X}}(B) = P(\boldsymbol{X} \in B), \forall B \in \mathcal{B}_n$, sa nazýva *rozdelenie náhodného vektoru* $\boldsymbol{X}$ *príslušné pravdepodobnosti* P, prípadne simultánne (združené) rozdelenie náhodných veličín $X_1, ..., X_n$.

Definícia 1.15. Nech $\boldsymbol{X} = (X_1, ..., X_n)^T$ je náhodný vektor na $(\Omega, \mathcal{A}, P)$. Potom funkciu danú predpisom

$$\forall \boldsymbol{x} \in \mathbb{R}^n : F_{\boldsymbol{X}}(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x}),$$

(t. j. $\forall (x_1, ..., x_n)^T \in \mathbb{R}^n : F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$) nazývame *distribučnou funkciou náhodného vektoru* $\boldsymbol{X}$, prípadne simultánnou (združenou) distribučnou funkciou náhodných veličín $X_1, ..., X_n$.

Poznámka. Marginálnou distribučnou funkciou náhodnej veličiny X_i nazývame funkciu danú predpisom

$$F_{X_i}(x_i) = \lim_{x_1 \to \infty, ..., x_{i-1} \to \infty, x_{i+1} \to \infty, ..., x_n \to \infty} F(x_1, ..., x_n), \quad \forall i = \{1, ..., n\}.$$

Pojmy *diskrétny* a *spojitý náhodný vektor* sú definované podobne ako pri náhodnej veličine s rozdielom viacrozmernosti.

Veta 1.16. Nech $F(x_1, ..., x_n)$ je simultánna distribučná funkcia náhodného vektora $\boldsymbol{X} = (X_1, ..., X_n)^T$ a $F_{X_i}(x_i)$, $i = 1, ..., n$ je marginálna distribučná funkcia náhodnej veličiny X_i. Náhodné veličiny $X_1, ..., X_n$ sú nezávislé práve vtedy, keď:

$$\forall \boldsymbol{x} \in \mathbb{R}^n : F(x_1, ..., x_n) = F_{X_1}(x_1) \cdot ... \cdot F_{X_n}(x_n).$$

Dôkaz. Táto veta je dokázaná v [15].

1.3.1 Číselné charakteristiky náhodného vektora

Definícia 1.17. Nech $\boldsymbol{X} = (X_1, ..., X_n)^T$ je náhodný vektor na $(\Omega, \mathcal{A}, P)$. *Stredná hodnota* $E(\boldsymbol{X})$ *náhodného vektora* $\boldsymbol{X}$ sa definuje ako vektor stredných hodnôt zložiek náhodného vektora (t. j. $E(\boldsymbol{X}) = (EX_1, ..., EX_n)$).

Definícia 1.18. Nech $\boldsymbol{X} = (X_1, ..., X_n)^T$ je náhodný vektor na $(\Omega, \mathcal{A}, P)$. Vzájomný vzťah zložiek X_i a X_j náhodného vektora $\boldsymbol{X} = (X_1, ..., X_n)^T$ vyjadruje číselná charakteristika s názvom *kovariancia*:

$$C(X_i, X_j) = E([X_i - E(X_i)][X_j - E(X_j)]) = E(X_i X_j) - E(X_i)E(X_j).$$

Z jednotlivých kovariancií sa zostavuje *variančná matica* náhodného vektora:

$$var(\boldsymbol{X}) = [C(X_i, X_j)], \quad i = 1, ..., n, \quad j = 1, ..., n,$$

ktorá je symetrická s rozptylmi $D(X_1), ..., D(X_n)$ na hlavnej diagonále.

Definícia 1.19. Nech $\boldsymbol{X} = (X_1, ..., X_n)^T$ je náhodný vektor na $(\Omega, \mathcal{A}, P)$. Mierou závislosti náhodných veličín X_i a X_j je ich *korelačný koeficient*:

$$\rho(X_i, X_j) = E\left[\frac{X_i - E(X_i)}{\sqrt{D(X_i)}} \cdot \frac{X_j - E(X_j)}{\sqrt{D(X_j)}}\right] = \frac{C(X_i, X_j)}{\sqrt{D(X_i)} \cdot \sqrt{D(X_j)}}.$$

Z koeficientov korelácie sa zostavuje *korelačná matica* náhodného vektora:

$$corr(\boldsymbol{X}) = [\rho(X_i, X_j)], \quad i = 1, ..., n, \quad j = 1, ..., n,$$

ktorá je symetrická s jednotkami na hlavnej diagonále.

1.4 Bodové a intervalové odhady

Predpokladajme, že náhodný vektor $\boldsymbol{X} = (X_1, ..., X_n)^T$ má hustotu $f(\boldsymbol{x}, \boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T$ je neznámy parameter, prípadne parametrická funkcia. Na základe vektora X je potrebné získať čo najlepší odhad parametra $\boldsymbol{\theta}$, o ktorom dopredu vieme len toľko, že patrí do nejakého parametrického priestoru $\Theta \subset \mathbb{R}_m$. Tieto odhady rozdeľujeme na bodové a intervalové [2].

Najprv definujeme pojmy, pomocou ktorých budeme môcť definovať tieto odhady. Tieto pojmy sme čerpali z [2], [13] a [15].

Definícia 1.20. Náhodný vektor $\boldsymbol{X} = (X_1, ..., X_n)^T$, ktorého zložky sú nezávislé náhodné veličiny s rovnakým rozdelením pravdepodobnosti ako skúmaná náhodná veličina X, sa nazýva *náhodný výber rozsahu n* príslušný štatistickému znaku X alebo taktiež výber z rozdelenia s distribučnou funkciou $F_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$.

Definícia 1.21. Ľubovoľnú transformáciu náhodného výberu $T = T(X_1, ..., X_n)$ nazývame *štatistika*.

Definícia 1.22. Nech $\boldsymbol{X} = (X_1, ..., X_n)^T$ je náhodný výber rozsahu n príslušný štatistickému znaku X. Štatistiku

$\overline{X}_n = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ nazývame výberový priemer,

$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ nazývame výberový rozptyl,

$S_n = \sqrt{S_n^2}$ nazývame výberová smerodajná odchýlka.

Odhadom T parametra $\boldsymbol{\theta}$ je štatistika $T = T(X_1, ..., X_n)$, ktorá na celom parametrickom priestore nadobúda hodnoty blízke parametru $\boldsymbol{\theta}$. Využívame najmä nasledujúce tri odhady:

1. nestranný – odhad, ktorého stredná hodnota $E(T) = \boldsymbol{\theta}$ (prípadne stranný – odhad, ktorého stredná hodnota $E(T) \neq \boldsymbol{\theta}$),
2. najlepší nestranný – nestranný odhad, ktorého rozptyl je najmenší z rozptylov všetkých nestranných odhadov daného parametra $\boldsymbol{\theta}$,
3. konzistentný – ak pre ľubovoľné reálne číslo $\varepsilon > 0$ platí $\lim_{n \to \infty} P(|T - \boldsymbol{\theta}| < \varepsilon) = 1$.

Definícia 1.23. *Bodový odhad* parametra $\boldsymbol{\theta}$ je pozorovaná hodnota $t = T(x_1, ..., x_n)$ odhadu T na štatistickom súbore $(x_1, ..., x_n)$, ktorá čo najlepšie aproximuje hodnotu $\boldsymbol{\theta}$.

Definícia 1.24. *Intervalový odhad* parametra $\boldsymbol{\theta}$ je interval alebo iná vhodná množina, ktorá s dostatočne veľkou pravdepodobnosťou pokrýva hodnotu $\boldsymbol{\theta}$.

Definícia 1.25. *Interval spoľahlivosti* pre parameter $\boldsymbol{\theta}$ so spoľahlivosťou $1 - \alpha$, kde $\alpha \in \langle 0, 1 \rangle$, je dvojica štatistík $D = D(X_1, ..., X_n)$ a $H = H(X_1, ..., X_n)$, ktoré spĺňajú:

$$P(D \leq \boldsymbol{\theta} \leq H) = 1 - \alpha.$$

Interval $\langle D, H \rangle$ nazývame $100(1 - \alpha)\%$ *interval spoľahlivosti* pre parameter $\boldsymbol{\theta}$.

Poznámka. Jednostranné odhady definujeme nasledovne: Ak $P(D \leq \boldsymbol{\theta}) = 1 - \alpha$, tak štatistiku D nazývame *dolným odhadom* parametra $\boldsymbol{\theta}$ so spoľahlivosťou $1 - \alpha$. Ak $P(\boldsymbol{\theta} \leq H) = 1 - \alpha$, tak štatistiku H nazývame *horným odhadom* parametra $\boldsymbol{\theta}$ so spoľahlivosťou $1 - \alpha$.

1.5 Testovanie hypotéz

Táto podkapitola bola zostavená s pomocou [26].

Definícia 1.26. Nech $\boldsymbol{X} = (X_1, ..., X_n)^T$ je náhodný vektor so združenou distribučnou funkciou $F_{\boldsymbol{X}}(\boldsymbol{x})$. Tvrdenie o rozdelení určenom touto distribučnou funkciou nazveme *hypotéza*.

V štatistike formulujeme *nulovú hypotézu* H_0 a *alternatívnu hypotézu* H_1, ktorá je väčšinou doplnkom (negáciou) nulovej hypotézy. Sformulovanú nulovú hypotézu môžeme *zamietnuť* alebo *nezamietnuť*. Toto rozhodnutie vykonáme na základe realizácie náhodného vektora, takže nemusí byť bezchybné.

Definícia 1.27. Ak zamietneme nulovú hypotézu, hoci platí, nastáva *chyba prvého druhu*. Ak naopak nulovú hypotézu nezamietneme, keď neplatí, tak sa dopúšťame *chyby druhého druhu*.

Rozhodovanie prebieha tak, že najprv určíme *kritický obor* W_α, ktorý je množinou výsledkov pokusu, pri ktorých budeme hypotézu zamietať. Veľkosť kritického oboru zvolíme tak, aby sme platnú hypotézu zamietali nanajvýš s pravdepodobnosťou α.

Definícia 1.28. Maximálna dovolená pravdepodobnosť chyby prvého druhu sa nazýva *hladina významnosti testu* a označujeme ju α.

Definícia 1.29. Najmenšia hladina, pri ktorej by sme hypotézu ešte zamietli sa nazýva *p-hodnota* (anglicky *p-value*).

Pre stanovenie kritického oboru môžeme použiť vierohodnostnú funkciu. Uvažujme rozdelenie dané hustotou závislou na parametri $\boldsymbol{\theta}$. Pri nulovej hypotéze je rozdelenie náhodného vektora $\boldsymbol{X}$ dané hustotou $f(\boldsymbol{x}, \boldsymbol{\theta_0})$, pri alternatívnej hypotéze hustotou $f(\boldsymbol{x}, \boldsymbol{\theta_1})$. Do týchto hustôt následne dosadíme skutočne realizované hodnoty náhodných veličín $X_1 = x_1, ..., X_n = x_n$. Ak bude $f(\boldsymbol{x}, \boldsymbol{\theta_0})$ výrazne väčšia ako $f(\boldsymbol{x}, \boldsymbol{\theta_1})$, nezamietame hypotézu H_0, v opačnom prípade hypotézu H_0 zamietame.

Veta 1.30. Nech k danému $\alpha \in (0, 1)$ existuje také $c > 0$, že pre množinu $W_c = \{\boldsymbol{x} : f(\boldsymbol{x}, \boldsymbol{\theta_1}) \geq c f(\boldsymbol{x}, \boldsymbol{\theta_0})\}$ platí:

$$\int_{W_c} f(\boldsymbol{x}, \boldsymbol{\theta_0})\, d\boldsymbol{x} = \alpha.$$

Potom pre každú merateľnú množinu W takú, že $\int_W f(\boldsymbol{x}, \boldsymbol{\theta_0})\, d\boldsymbol{x} = \alpha$, platí:

$$\int_{W_c} f(\boldsymbol{x}, \boldsymbol{\theta_1})\, d\boldsymbol{x} \geq \int_W f(\boldsymbol{x}, \boldsymbol{\theta_0})\, dx.$$

Dôkaz. Dôkaz tejto vety môžeme nájsť v [26].

Poznámka. Tvrdenie vety 1.30 porovnáva pre dva uvažované kritické obory W a W_c pravdepodobnosť, s ktorou zamietneme nulovú hypotézu, keď platí alternatívna hypotéza (tzv. *sila testu*). Keďže ako W vystupuje akýkoľvek kritický obor s hladinou významnosti α, W_c je najsilnejší možný spomedzi kritických oborov s danou hladinou významnosti α.

1.6 Testy na odľahlé pozorovania

Pri skúmaní reálnych dát je obvyklé, že skúmaný náhodný výber obsahuje extrémne hodnoty, ktoré môžu ovplyvniť výsledky niektorých štatistických testov a preto je nutné im venovať pozornosť. Takéto hodnoty nazývame *odľahlé pozorovania* (anglicky *outliers*). Existuje veľa testov, ktoré slúžia na odhalenie a vyeliminovanie takýchto hodnôt ako napr. Grubbsove testy, ktoré boli predstavené v práci [10] alebo Dixonov test publikovaný v práci [6].

1.6.1 Dixonov test na jedno odľahlé pozorovanie

V tomto texte čerpáme z [6] a [14].

Jedným z predpokladov využitia Grubbsovho testu je normalita súboru, z ktorého dáta čerpáme. Normalita dát však nie je vždy zaručená a preto sa v tejto práci obraciame na Dixonov test, pri ktorom môže mať súbor ľubovoľné rozdelenie. Pred testovaním je potrebné usporiadať náhodný výber vzostupne: $x_{(1)} < x_{(2)} < ... < x_{(n)}$.

Testujeme hypotézu H_0, ktorá tvrdí, že sa v náhodnom výbere nenachádza odľahlé pozorovanie, oproti alternatívnej hypotéze H_1, ktorá tvrdí, že sa v náhodnom výbere odľahlé pozorovanie nachádza.

Základná testová štatistika pre testovanie odľahlosti pozorovania $x_{(1)}$ je daná ako pomer medzi vzdialenosťou dvoch najmenších pozorovaní a rozsahom náhodného výberu:

$$r_{10} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}.$$

Obdobne je určená aj testová štatistika pre testovanie odľahlosti pozorovania $x_{(n)}$:

$$r'_{10} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}.$$

Ak nedokážeme určiť, či je odľahlou hodnotou v náhodnom výbere pozorovanie $x_{(1)}$ alebo pozorovanie $x_{(n)}$, vypočítame obe štatistiky, porovnáme ich a ďalej pracujeme s väčšou z týchto štatistík.

V niektorých prípadoch však môže hodnota, ktorá je blízka skúmanému odľahlému pozorovaniu, skresľovať výsledok testu a preto takúto hodnotu z testu vylúčime a namiesto štatistík r_{10} a r'_{10} používame nasledujúce štatistiky:

$$r_{11} = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}, \qquad \text{resp.} \qquad r'_{11} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}},$$

ktorá testuje odľahlosť $x_{(1)}$ pri vylúčení vplyvu $x_{(n)}$, resp. odľahlosť $x_{(n)}$ pri vylúčení vplyvu $x_{(1)}$.

$$r_{21} = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}, \qquad \text{resp.} \qquad r'_{21} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}},$$

ktorá testuje odľahlosť $x_{(1)}$ pri vylúčení vplyvu $x_{(2)}$ a $x_{(n)}$, resp. odľahlosť $x_{(n)}$ pri vylúčení vplyvu $x_{(n-1)}$ a $x_{(1)}$.

$$r_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}, \qquad \text{resp.} \qquad r'_{22} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}},$$

ktorá testuje odľahlosť $x_{(1)}$ pri vylúčení vplyvu $x_{(2)}$, $x_{(n-1)}$ a $x_{(n)}$, resp. odľahlosť $x_{(n)}$ pri vylúčení vplyvu $x_{(n-1)}$, $x_{(1)}$ a $x_{(2)}$.

Hypotézu H_0 zamietneme a pozorovanie eliminujeme, ak je vypočítaná testová štatistika väčšia ako kritická hodnota nachádzajúca sa v tabuľkách. Tabuľky kritických hodnôt môžeme nájsť v [7] a v [21]. Pre výbery s rozsahom 3 až 7 pozorovaní sa používajú štatistiky r_{10} a r'_{10}, pre výbery s rozsahom 8 až 10 zas r_{11} a r'_{11}, pre výbery s rozsahom 11 až 13 to sú štatistiky r_{21} a r'_{21} a pre výbery s väčším rozsahom sa používajú štatistiky r_{22} a r'_{22}.

1.7 Analýza rozptylu – ANOVA

Jednou z metód matematickej štatistiky je aj analýza rozptylu. Informácie o tejto metóde sme získali z [2], [11],[17] a [26]. Viac informácií o analýze rozptylu a jej metódach je možné dohľadať v publikácii [2].

1.7.1 Predpoklady

Analýza rozptylu (anglicky *Analysis of variance – ANOVA*) má ako väčšina štatistických metód svoje predpoklady, ktoré musí spĺňať, inak ju nie je možné využiť. Týmito predpokladmi sú:

1. nezávislosť pozorovaných hodnôt,
2. normalita hodnôt jednotlivých náhodných výberov – tento predpoklad je nutné korektne overiť pomocou príslušného testu alebo aspoň pomocou grafických metód ako sú napr. histogram alebo krabicový graf,
3. rovnaký rozptyl hodnôt vo všetkých porovnávaných skupinách – tento predpoklad takisto overujeme pomocou adekvátneho testu (napr. Bartlettov test) alebo aspoň pomocou grafických metód spomínaných v predchádzajúcom bode.

Existuje veľa testov, ktorými môžeme testovať normalitu dát. Jedným z týchto testov je aj *Kolmogorovov – Smirnovov test.*

Veta 1.31. Testujeme hypotézu, ktorá tvrdí, že náhodný výber $\boldsymbol{X} = (X_1, ..., X_n)^T$ pochádza z rozdelenia s distribučnou funkciou $\Phi(x)$. Nech $F_n(x)$ je výberová distribučná funkcia. Testovou štatistikou je štatistika:

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi(x)|.$$

Nulovú hypotézu zamietame na hladine významnosti α, keď $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je kritická hodnota nachádzajúca sa v tabuľkách. Pre $n \geq 30$ môžeme $D_n(\alpha)$ aproximovať výrazom:

$$D_n(\alpha) \approx \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}.$$

Poznámka. Nulová hypotéza musí špecifikovať distribučnú funkciu celkom presne, vrátane všetkých jej prípadných parametrov.

Ak nepoznáme jeden alebo viac parametrov daného rozdelenia a musíme ich odhadovať z náhodného výberu, tak nemôžeme použiť štandardný Kolmogorov-Smirnovov test a jeho kritické hodnoty uvedené v tabuľkách. Vtedy pristupujeme k *Lillieforsovej variante* tohto testu.

V prípade testovania normality dát využijeme odhady strednej hodnoty a rozptylu, ktorými sú výberový priemer a výberový rozptyl uvedené v podkapitole 1.4. To znamená, že $\tilde{\mu} = \overline{X}$ a $\tilde{\sigma}^2 = S_n^2$. Testovou štatistikou je štatistika:

$$D = \max_x |F^*(x) - S_N(x)|,$$

kde $S_N(x)$ je výberová distribučná funkcia a $F^*(x)$ je distribučná funkcia normálneho rozdelenia $N(\tilde{\mu}, \tilde{\sigma}^2)$. Nulovú hypotézu, ktorá tvrdí, že náhodný výber $\boldsymbol{X} = (X_1, ..., X_n)^T$ pochádza z normálneho rozdelenia zamietame na hladine významnosti α, keď $D \geq D(\alpha)$, kde $D(\alpha)$ je kritická hodnota nachádzajúca sa v tabuľkách. Tabuľka s kritickými hodnotami je uvedená v [17].

1.7.2 Analýza rozptylu jednoduchého triedenia

Majme k dispozícii $k \geq 2$ nezávislých výberov z normálneho rozdelenia s rovnakým rozptylom, t. j. nech:

$$\begin{aligned} Y_{11}, ..., Y_{1n_1} &\sim N(\mu_1, \sigma^2), \\ Y_{21}, ..., Y_{2n_2} &\sim N(\mu_2, \sigma^2), \\ &... \\ Y_{k1}, ..., Y_{kn_k} &\sim N(\mu_k, \sigma^2). \end{aligned}$$

Uvedený model označujeme ako model *analýzy rozptylu jednoduchého triedenia* (*jednofaktorovej analýzy rozptylu*). Budeme testovať nulovú hypotézu $H_0 : \mu_1 = ... = \mu_k$ (spoločnú hodnotu označíme μ) proti alternatívnej hypotéze H_1, že aspoň dva výbery majú rôzne stredné hodnoty.

Keď usporiadame hodnoty $Y_{11}, ..., Y_{1n_1}, Y_{21}, ..., Y_{2n_2}, ..., Y_{k1}, ..., Y_{kn_k}$ do vektora, môžeme úlohu zapísať ako lineárny model:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n} \\ Y_{21} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}.$$

Ako odhady stredných hodnôt μ_i dostaneme priemery

$$\overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Potom je zrejmé, že $\widehat{Y}_{ij} = \overline{Y}_{i\bullet}$, takže reziduálny súčet štvorcov (viď [26]) je rovný súčtu súčtov štvorcov odchýlok od priemeru v jednotlivých výberoch:

$$S_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2.$$

Ak platí testovaná hypotéza (t. j. ak je vo všetkých výberoch rovnaká stredná hodnota), platí vlastne podmodel, v ktorom majú všetky zložky náhodného vektora $\boldsymbol{Y}$ rovnakú strednú hodnotu:

$$Y_{ij} \sim N(\mu, \sigma^2), \quad i = 1, ..., k, \quad j = 1, ..., n_i.$$

Odhadom tejto spoločnej strednej hodnoty je celkový priemer:

$$\overline{Y}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{Y}_{i\bullet}$$

a reziduálnym súčtom štvorcov (pri hypotéze) je výraz:

$$S_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\bullet\bullet})^2.$$

Keďže model popisuje stredné hodnoty k nezávislých parametrov a v podmodeli je len jediný, testová štatistika má tvar:

$$F = \frac{(S_T - S_e)/(k-1)}{S_e/(n-k)} \sim F(k-1, n-k).$$

Celkovú variabilitu (celkový súčet štvorcov) môžeme vyjadriť ako súčet variability vysvetlenej modelom (t. j. $S_T - S_e$) a variability nevysvetlenej modelom (reziduálny súčet štvorcov S_e).

1.7.3 Analýza rozptylu dvojného triedenia

Model z úvodu podkapitoly 1.7.2 môžeme zapísať nasledovne:

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \quad 1 \leq i \leq k, \quad 1 \leq j \leq n_i,$$

kde $e_{ij} \sim N(0, \sigma^2)$ sú nezávislé náhodné veličiny. Namiesto k nezávislých parametrov, ktoré popisujú stredné hodnoty náhodných veličín Y, tu máme celkom $k+1$ parametrov. Vzájomne jednoznačný vzťah dostaneme, keď na *efekty* α_i kladieme reparametrizačnú podmienku $\sum \alpha_i = 0$. Pretože α_i sú pevné (nenáhodné) neznáme konštanty, hovoríme o *pevných efektoch*, ktoré vyjadrujú možné odchýlky jednotlivých stredných hodnôt od spoločnej strednej úrovne μ. Nulovú hypotézu H_0 potom formulujeme ako požiadavku $\alpha_1 = ... = \alpha_k$. Vzhľadom k reparametrizačnej podmienke je táto spoločná hodnota nutne nulová. Všetky úvahy, najmä rozklad súčtu štvorcov sú uvedenou parametrizáciou nezmenené. Znak, podľa ktorého triedime veličiny Y do jednotlivých výberov (skupín), sa nazýva *faktor*. V našom prípade má tento faktor k úrovní, ktorým odpovedajú efekty α_i.

Vyjadrenie $Y_{ij} = \mu + \alpha_i + e_{ij}$ môžeme zovšeobecniť na zložitejšie modely, v ktorých sledujeme vplyv niekoľkých faktorov. Pri *analýze rozptylu dvojného triedenia* (*dvojfaktorovej analýze rozptylu*) využívame model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad 1 \leq i \leq k, \quad 1 \leq j \leq m,$$

kde $Y_{ij} \sim N(0, \sigma^2)$ sú opäť nezávislé náhodné veličiny. Podobne ako v prípade modelu jednoduchého triedenia použijeme reparametrizačné podmienky a teda predpokladáme, že platí $\sum \alpha_i = 0$ a súčasne $\sum \beta_j = 0$, kde α_i a β_j sú riadkové alebo stĺpcové pevné efekty.

Tentokrát sa celkový súčet štvorcov rozkladá do troch sčítancov $S_T = S_A + S_B + S_e$, kde:

$$S_T = \sum_{i=1}^{k}\sum_{j=1}^{m}(\overline{Y}_{ij} - \overline{Y}_{\bullet\bullet})^2,$$

$$S_A = \sum_{i=1}^{k}\sum_{j=1}^{m}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2,$$

$$S_B = \sum_{i=1}^{k}\sum_{j=1}^{m}(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})^2,$$

$$S_e = \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet})^2.$$

Súčet S_A má $k-1$ stupňov voľnosti a popisuje variabilitu riadkových priemerov $\overline{Y}_{i\bullet}$ a používame ho k testovaniu nulovej hypotézy $\alpha_1 = ... = \alpha_k = 0$. Podobne S_B má $m-1$ stupňov voľnosti a popisuje variabilitu stĺpcových priemerov $\overline{Y}_{\bullet j}$ a používame ho k testovaniu nulovej hypotézy $\beta_1 = ... = \beta_m = 0$. V oboch prípadoch porovnávame tzv. priemerný štvorec $MS_A = S_A/(k-1)$ resp. $MS_B = S_B/(m-1)$ s reziduálnym rozptylom $s^2 = S_e/((k-1)(m-1))$. Tieto medzivýsledky a odpovedajúce F-štatistiky:

$$F_A = \frac{MS_A}{MS_e}, \quad F_B = \frac{MS_B}{MS_e},$$

sa zapisujú do tabuľky analýzy rozptylu a slúžia k rozhodovaniu o tom, či riadkový alebo stĺpcový faktor ovplyvňujú strednú hodnotu vysvetľovanej premennej Y.

1.7.4 Kruskalov-Wallisov test

Ak nie je splnený aspoň jeden z predpokladov uvedených v podkapitole 1.7.1, musíme pristúpiť ku *Kruskalovmu-Wallisovmu testu*. Tento test je neparametrickou obdobou analýzy rozptylu jednoduchého triedenia a je zovšeobecnením dvojvýberového Wilcoxonovho testu, ktorý je opísaný napr. v [2]. Používame ho najmä vtedy, ak ide o výbery z rozdelení, ktoré sa významne líšia od rozdelenia normálneho.

Nech $Y_{i_1}, ..., Y_{in_i}$ je výber z nejakého rozdelenia so spojitou distribučnou funkciou F_i, $i = 1, ..., I$. Nech všetky tieto výbery sú na sebe nezávislé. Budeme testovať nulovú hypotézu $H_0 : F_1(x) = ... = F_I(x)$, $\forall x$ oproti alternatívnej hypotéze H_1, že H_0 neplatí. Všetky veličiny Y_{ij} tvoria dohromady združený náhodný výber s rozsahom $N = n_1 + ... + n_I$. Tieto veličiny následne usporiadame do rastúcej postupnosti a určíme poradie R_{ij}

každej veličiny Y_{ij} v združenom náhodnom výbere. Ďalej označíme súčet poradí v i-tom náhodnom výbere ako $T_i = \sum_{j=1}^{n_i} R_{ij}$. Celkový súčet všetkých poradí je teda $T_1 + ... + T_I = N(N+1)/2$. Ako testová štatistika sa použije:

$$Q = \frac{12}{N(N+1)} \sum_{i=1}^{I} \frac{T_i^2}{n_i} - 3(N+1).$$

Vieme ukázať, že testová štatistika Q má χ^2 rozdelenie s $k-1$ stupňami voľnosti. Nulovú hypotézu H_0 teda zamietame na hladine významnosti α, ak $Q \geq \chi^2_\alpha(k-1)$, pričom kvantily χ^2 rozdelenia je možné dohľadať v štatistických tabuľkách.

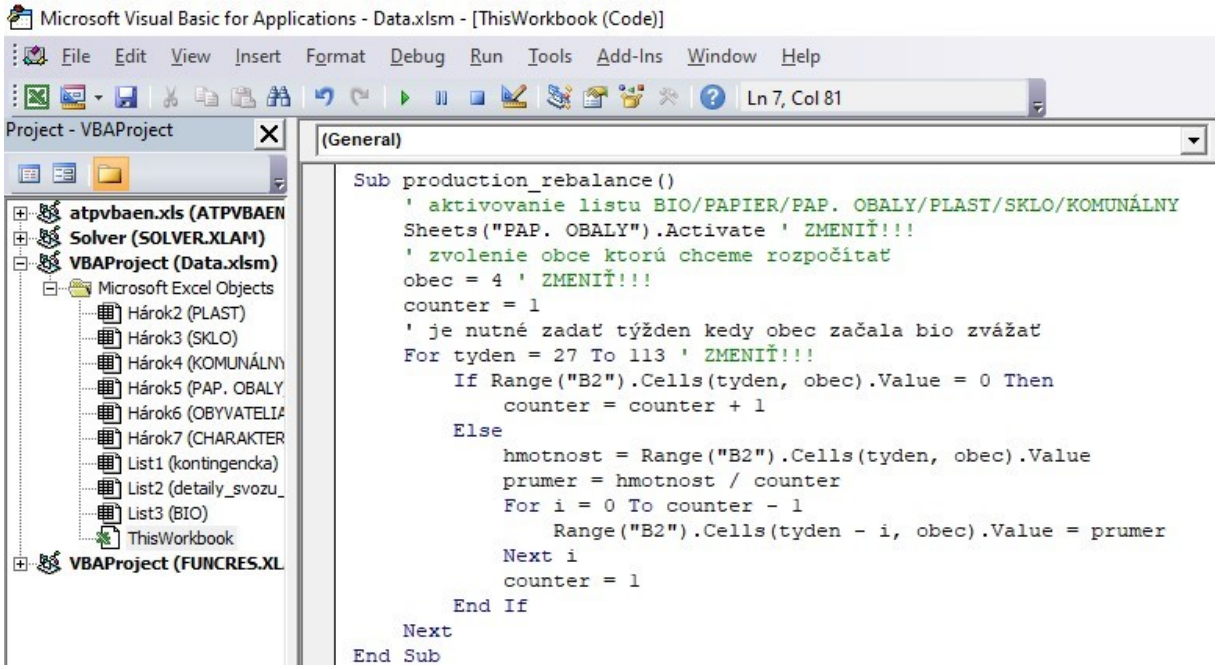
2 Použitý software – Microsoft Excel

Táto kapitola sa zaoberá opisom softvéru použitého pri spracovávaní dát. Jej text bol písaný s využitím [16].

Softvér, ktorý sme využili pri práci s poskytnutými dátami, sa nazýva Microsoft Excel. Excel je tabuľkový procesor od firmy Microsoft, dostupný pre dva operačné systémy – MS Windows a MAC OS. Jeho hlavnými cieľmi je triediť, skúmať, analyzovať a zobrazovať akékoľvek dáta. Všetky dáta môžeme upravovať a spracovávať pomocou množstva funkcií, ktoré sú súčasťou Excelu, no môžeme si vytvoriť aj svoje vlastné. Opísané činnosti sa však dajú aj zautomatizovať a to vďaka programovaciemu jazyku Visual Basic for Application.

2.1 Visual Basic for Application

Visual Basic for Application (ďalej len VBA) je objektovo orientovaný programovací jazyk, ktorý je súčasťou produktov MS Office. Tento jazyk slúži na zautomatizovanie, zrýchlenie, spresnenie a uľahčenie opakujúcej sa práce. Využitie nachádza v procedúrach, ktoré nazývame makrá. Editor jazyka VBA spolu so zoznamom a zaznamenávaním makier môžeme nájsť v karte Vývojár. Po otvorení editora jazyka VBA, ktorý je zobrazený na obr. 2.1, môžeme začať písať kód.

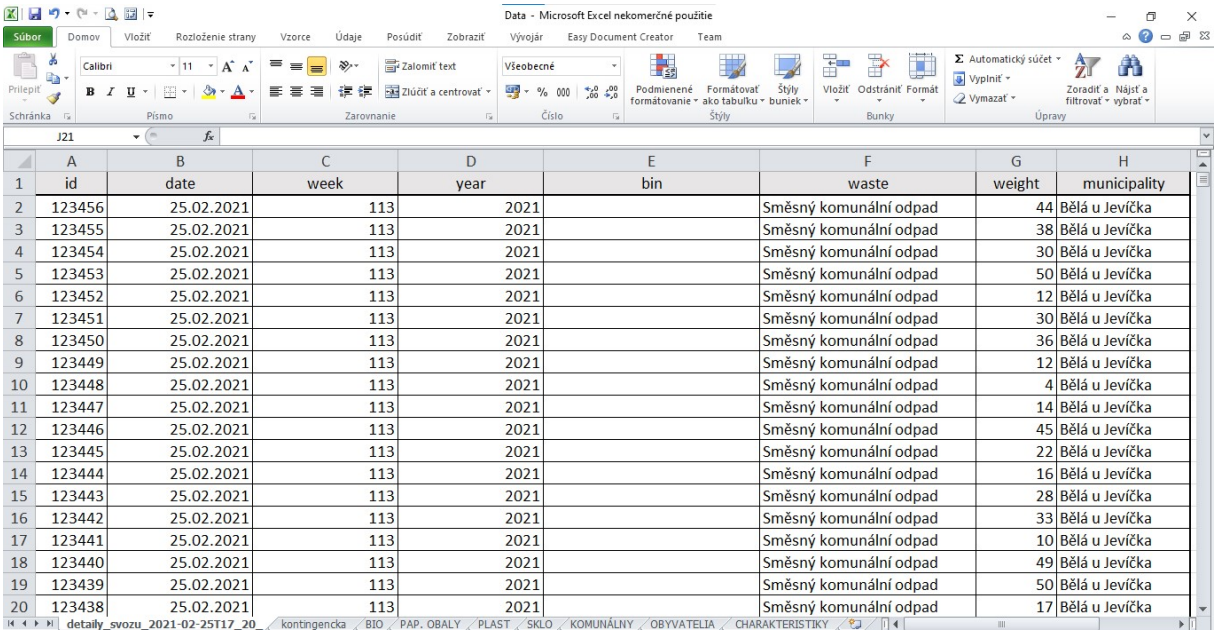


Obr. 2.1: Ukážka vývojového prostredia VBA

3 Popis a príprava dát

Úkony vykonávané v tejto a nasledujúcej kapitole boli popísané vo vývojovom diagrame priloženom v prílohe A na konci tejto práce.

Zásadnú časť tejto bakalárskej práce tvorí analýza historických dát z reálneho monitoringu zvozových vozidiel, ktoré nám boli poskytnuté spoločnosťou Technické služby Malá Haná s.r.o. (ďalej len TSMH). Tieto dáta boli dodané vo forme tabuľky, ktorú bolo potrebné spracovať a upraviť do podoby vhodnej na štatistické testy a analýzu. Pôvodný súbor obsahoval 373 603 zvozov, prvý zvoz bol zaznamenaný v apríli 2019. Časť pôvodnej tabuľky je na obr. 3.1.



	A	B	C	D	E	F	G	H
1	id	date	week	year	bin	waste	weight	municipality
2	123456	25.02.2021	113	2021		Směsný komunální odpad	44	Bělá u Jevíčka
3	123455	25.02.2021	113	2021		Směsný komunální odpad	38	Bělá u Jevíčka
4	123454	25.02.2021	113	2021		Směsný komunální odpad	30	Bělá u Jevíčka
5	123453	25.02.2021	113	2021		Směsný komunální odpad	50	Bělá u Jevíčka
6	123452	25.02.2021	113	2021		Směsný komunální odpad	12	Bělá u Jevíčka
7	123451	25.02.2021	113	2021		Směsný komunální odpad	30	Bělá u Jevíčka
8	123450	25.02.2021	113	2021		Směsný komunální odpad	36	Bělá u Jevíčka
9	123449	25.02.2021	113	2021		Směsný komunální odpad	12	Bělá u Jevíčka
10	123448	25.02.2021	113	2021		Směsný komunální odpad	4	Bělá u Jevíčka
11	123447	25.02.2021	113	2021		Směsný komunální odpad	14	Bělá u Jevíčka
12	123446	25.02.2021	113	2021		Směsný komunální odpad	45	Bělá u Jevíčka
13	123445	25.02.2021	113	2021		Směsný komunální odpad	22	Bělá u Jevíčka
14	123444	25.02.2021	113	2021		Směsný komunální odpad	16	Bělá u Jevíčka
15	123443	25.02.2021	113	2021		Směsný komunální odpad	28	Bělá u Jevíčka
16	123442	25.02.2021	113	2021		Směsný komunální odpad	33	Bělá u Jevíčka
17	123441	25.02.2021	113	2021		Směsný komunální odpad	10	Bělá u Jevíčka
18	123440	25.02.2021	113	2021		Směsný komunální odpad	49	Bělá u Jevíčka
19	123439	25.02.2021	113	2021		Směsný komunální odpad	50	Bělá u Jevíčka
20	123438	25.02.2021	113	2021		Směsný komunální odpad	17	Bělá u Jevíčka

Obr. 3.1: Ukážka pôvodných dát

V hlavičke tabuľky je zapísané, aké informácie sú zaznamenávané o každom jednom zvoze. Týmito informáciami sú:

1. id - jedinečné číslo zvozu,
2. date - dátum a čas, v ktorom bol zvoz uskutočnený,
3. week - doplnená hodnota, podľa ktorej boli dáta delené v kontingenčnej tabuľke,
4. year - rok zvozu,
5. bin - informácie o zbernej nádobe,
6. waste - typ odpadu, ktorý bol vážený,
7. weight - odvážená hmotnosť pri každom zvoze,
8. municipality - obec, v ktorom bol zvoz uskutočnený.

Ďalším krokom bolo zostrojiť kontingenčnú tabuľku. Kontingenčná tabuľka slúži na sprehľadnenie vzťahu medzi dvoma údajmi (v našom prípade medzi obcou a týždňom zvozu), kde sa v každom zo stĺpcov nachádza jedna obec a v každom z riadkov číslo týždňa. V bunke, ktorá tieto dve hodnoty spája, sa nachádza celková hmotnosť odpadu, ktorá bola zvozená z danej obce v daný týždeň. Kontingenčnú tabuľku sme chceli vytvoriť ako prehľad od začiatku roka 2019 a preto sme pôvodnú tabuľku museli doplniť o pár nulových hodnôt. Po tomto kroku sme už mohli prejsť k tvorbe kontingenčnej tabuľky, ktorá nám rozdelila dáta do jednotlivých frakcií, ktorými sú biologicky rozložiteľný odpad, papierové a lepenkové obaly, plasty, sklo a zmesový komunálny odpad.

Takto vytvorené tabuľky samostatných frakcií boli skopírované do nových hárkov v Exceli. Keďže každá z obcí má inú frekvenciu zvozu odpadu, produkcia musela byť rozpočítaná aj do týždňov, v ktorých sa odpad nezvážal. Na toto sme použili makro vytvorené vo VBA, ktoré po zadaní obce a počiatku zvozu rozpočítalo odpad nasledovne: týždne, v ktorých bol vynechaný zvoz, boli doplnené priemerom z nasledujúceho zvozu podľa počtu týždňov, ku ktorým zodpovedala táto produkcia. Takto upravené dáta, ktoré sa nachádzajú v priloženom súbore programu Microsoft Excel, už boli pripravené na štatistickú analýzu.

3.1 Ciele analýzy dát

Zvozové trasy sa vytvárajú na základe skúsenosti prevádzkovateľov spoločností, ktoré majú zvozy na starosti. Pri novovzniknutých zväzkoch obcí, ktoré si chcú riešiť zvoz sami, je to však problém. Jedným z príkladov sú TSMH, ktoré začali svoju činnosť na konci roku 2017 a od roku 2019 už zbierajú dáta zo zvozov. Cieľom zberu dát je zvýšenie efektivity zvozu a zníženie nákladov na odpadové hospodárstvo.

V rámci Ústavu procesního inženýrství Vysokého učení technického v Brně je dlhodobo vyvíjaný výpočtový nástroj NERUDA Street, ktorý je zameraný na plánovanie zvozov medzi obcami, ale aj ulicami vo väčších mestách. Kľúčovou podmienkou pre použiteľnosť výsledkov nástroja NERUDA Street sú kvalitné vstupné dáta o produkcii odpadu a dopravnej infraštruktúre vrátane dopravných dát (hustota premávky a i.).

Táto práca sa zameriava na spracovanie a analýzu dát z reálneho monitoringu. Zaznamenávané sú dáta o hmotnosti zberných nádob z každého zvozu a časové informácie v rámci celého dopravného reťazca. Kľúčovým výstupom tejto práce je vývoj produkcie odpadu v priebehu roku. Tieto informácie sú podstatné pre plánovanie zvozových oblastí a taktiež pre vozový park, ktorý sa musí neustále prispôsobovať zmenám. Týmito zmenami je zvyšovanie frakcií separátne zbieraných odpadov a zvyšujúci sa počet obcí patriacich do zväzku. V prípade TSMH bolo na konci roku 2017 vo zväzku len 17 obcí, zatiaľ čo v roku 2021 je ich vo zväzku už viac ako 50.

4 Analýza dát

Reálne dáta je možné pozorovať, porovnávať a testovať rôznymi spôsobmi. Pri opätovnej práci s dátami rovnakého typu dokážeme predpovedať isté závery. Dáta z monitoringu zvozov nie sú výnimkou a môžeme pri nich sledovať napr. sezónnosť. Jedným z prístupov spracovania sezónnych dát je tvorba a testovanie tzv. *SARIMA* modelov (skratka anglického *Seasonal Autoregressive Integrated Moving Average*, teda sezónny autoregresný integrovaný kĺzavý priemer). Túto metódu spracovania však nevyužijeme, kvôli nedostatočnému množstvu dát (k dispozícii sme mali dáta od apríla 2019 do februára 2021, pričom tieto dáta neboli vždy úplné). O SARIME sa môžeme dočítať napríklad v [12]. Naše spracovanie malo niekoľko krokov.

4.1 Štatistické testy

4.1.1 Dixonov test

Prvým krokom štatistickej analýzy bol Dixonov test, spomínaný v podkapitole 1.6. Na dáta rozdelené do frakcií bolo pustené makro vytvorené vo VBA. Toto makro vyberalo podľa rozsahu náhodného výberu z testov r_{10}, r_{11}, r_{21} a r_{22} (resp. ich obdôb r'_{10}, r'_{11}, r'_{21} a r'_{22}). Po zvolení frakcie a obce, pre ktorú chceme tento test uskutočniť sme spustili makro, ktoré:

1. vyhľadalo potenciálne odľahlé pozorovanie (minimálnu alebo maximálnu hodnotu náhodného výberu),
2. vypočítalo príslušné testové štatistiky podľa rozsahu náhodného výberu a vybralo tú, ktorá bola kritickejšia,
3. rozhodlo o vylúčení odľahlého pozorovania po porovnaní s kritickou hodnotou nachádzajúcou sa v tabuľke.

Ak bolo rozhodnuté, že pozorovanie je naozaj odľahlé a musí byť odstránené, makro elimináciu vykonalo, bunku s odstránenou hodnotou označilo farebne a nahradilo priemerom hodnôt z okolitých týždňov, aby pri vykreslení grafu produkcie nedošlo k nespojitosti krivky.

Pri niektorých obciach je možno pozorovať viac extrémnych hodnôt ako jednu. Keďže Dixonov test je test na jedno odľahlé pozorovanie, tieto anomálie boli potlačené použitím kĺzavého priemeru, ktorý opíšeme v sekcii 4.1.4.

4.1.2 Štandardizácia

Aby sme mohli medzi sebou porovnávať údaje z obcí s rôznym počtom obyvateľov, musíme dáta štandardizovať (teda prepočítať na jednu osobu). Vďaka jednoduchému makru, v ktorom stačí prepísať typ odpadu a počet obcí v hárku, sa pôvodné hodnoty predelili počtom obyvateľov v príslušnej obci a za pár desiatok sekúnd sme mali nové hodnoty, ktoré už boli vhodné na porovnávanie. Najaktuálnejšie demografické údaje o obciach združených v TSMH sme našli na stránkach Českého statistického úřadu [5].

4.1.3 Variačný koeficient

Po štandardizovaní dát sme prešli k výpočtu charakteristík, ktoré nám pomôžu pri porovnávaní obcí medzi sebou. Už po výpočte výberového priemeru a výberovej smerodajnej odchýlky je možné pozorovať, ako sa líšia zozbierané dáta v jednotlivých obciach a môžeme sa zamýšľať, čím je to zapríčinené. Je to spôsobené počtom obyvateľov žijúcich v obci? Veľkosťou obce? Nadmorskou výškou? Dĺžkou obdobia monitoringu? Tieto dve hodnoty (priemer a smerodajnú odchýlku) dáva do súvisu variačný koeficient [20]. Variačný koeficient predstavuje relatívnu mieru variability dát vzhľadom k priemeru a je daný vzorcom:

$$\text{variačný koeficient} = \frac{\text{smerodajná odchýlka}}{\text{priemer}}$$

Vypočítané výberové priemery, výberové smerodajné odchýlky aj variačné koeficienty sme zoskupili do tabuliek, v ktorých je možné porovnávať tieto charakteristiky medzi obcami a medzi frakciami navzájom. Tieto tabuľky sú súčasťou priloženého súboru programu Microsoft Excel (hárok „CHARAKTERISTIKY“). Malá hodnota variačného koeficientu znamená nízku variabilitu (premenlivosť) dát, zatiaľ čo veľká hodnota (t. j. hodnota väčšia ako 0,5) vyjadruje vysokú variabilitu dát. Bunky označené červenou farbou boli v procese post-processingu opísanom v podkapitole 4.1.5 vylúčené z dôvodu nedostatočného počtu dát, ktorý zapríčinil skreslenie vypočítaných charakteristík.

4.1.4 Kĺzavý priemer

Jednou z možností dokázania sezónnosti dát je použitie kĺzavého priemeru a sledovanie jeho prieniku s priemerom aritmetickým (tam detekujeme prechodovú oblasť). Kĺzavý priemer [8] (anglicky *Moving Average*) je výpočet, ktorý nám pomáha analyzovať dáta využitím série aritmetických priemerov rôznych podskupín náhodného výberu. Slúži na vyhladenie dát vytvorením neustále sa meniaceho priemeru a určenie trendov v sérii dát. Výpočtom kĺzavého priemeru zaručujeme vyhladenie občasných skokov v dátach.

V našej práci sme na stanovenie trendov využili najjednoduchšiu formu kĺzavého priemeru – jednoduchý kĺzavý priemer (anglicky *Simple Moving Average* – SMA). Jednostranný kĺzavý priemer sa počíta ako séria aritmetických priemerov z posledných k hodnôt náhodného výberu rozsahu n, kde $k \in \mathbb{N}$:

$$SMA_i = \frac{x_i + x_{i-1} + ... + x_{i-k+1}}{k}, \quad i = k, k+1, ..., n.$$

My však využijeme centrovaný (alebo dvojstranný) kĺzavý priemer, ktorý sa počíta ako séria aritmetických priemerov z $k = 2m + 1$ hodnôt, kde $m \in \mathbb{N}$:

$$SMA_i = \frac{x_{i-m} + x_{i-m+1} + ... + x_i + ... + x_{i+m-1} + x_{i+m}}{k}, \quad i = j, j+1, ..., n-j,$$

kde $j = \frac{k+1}{2}$. Našou voľbou bolo $k = 7$, ktoré vyhladilo krivku kĺzavého priemeru dostatočne, narozdiel od $k = 5$, kde sme mohli pozorovať výkyvy. Voľba $k = 9$ tiež nie je vhodná, keďže v takomto prípade berieme hodnoty z deviatich týždňov, čo je príliš dlhá doba na zistenie presného zlomu trendu v produkcii odpadu.

Vyhladenie dát je najjasnejšie po vykreslení hodnôt a kĺzavého priemeru do jedného grafu. Pri piatich frakciách s desiatkami obcí v každej z nich by ale bolo obtiažne

porovnávať každé dva grafy. Preto sme znovu využili makro, ktoré po zadaní frakcie, obce, prvého týždňa zvozu a posledného týždňa zvozu vypočíta kĺzavý priemer podľa vzorca:

$$SMA_i = \frac{x_{i-3} + x_{i-2} + x_{i-1} + x_i + x_{i+1} + x_{i+2} + x_{i+3}}{7}, \quad i = 4, 5, ..., n-4.$$

Ak hodnota kĺzavého priemeru presiahla v niektorom z týždňov hodnotu aritmetického priemeru, bunka sa zafarbila a ďalšie bunky boli zafarbené až dovtedy, kým hodnota kĺzavého priemeru neklesla pod hodnotu aritmetického priemeru. Prechod z nesfarbenej bunky na sfarbenú alebo naopak (teda prienik dvoch priemerov) značí prítomnosť prechodovej oblasti.

Po vykonaní tohto kroku v každej frakcii môžeme vidieť, že jedinou frakciou, v ktorej je možné pozorovať sezónnosť, je biologicky rozložiteľný odpad. Toto je zapríčinené tým, že biologicky rozložiteľný odpad sme schopní rozdeliť na kuchynský a záhradkársky odpad. Kuchynský odpad je v domácnostiach produkovaný v letnej aj zimnej sezóne, zatiaľ čo záhradkársky odpad iba v letnej sezóne. Preto po skončení letnej sezóny pozorujeme pokles produkcie biologicky rozložiteľného odpadu, zatiaľ čo na začiatku zas jej nárast. Pre túto frakciu bolo teda v Exceli (hárok „BIO“) vytvorené makro, ktoré po vybratí obce a stlačení tlačidla vykreslí graf s pôvodnými hodnotami, aritmetickým a aj kĺzavým priemerom. Vykreslené grafy slúžia na lepšiu predstavu sezón definovaných prienikom priemerov.

4.1.5 Post-processing

V poslednej fáze (alebo podľa potreby aj medzi fázami predchádzajúcimi), ktorú nazývame tzv. post-processing, je potrebné ručne vyladiť nedostatky, ktoré tabuľkový procesor sám nezvládne. Takýmito nedostatkami sú:

- *vylúčenie obcí s jediným zvozom*,
- *vylúčenie obcí s malým počtom nazbieraných dát* - dôvodom je možné skreslenie štatistických modelov,
- *dofarbenie oblastí pri výpočte kĺzavého priemeru* - pri niektorých obciach sa stalo, že kĺzavý priemer v určitých týždňoch mierne osciloval okolo aritmetického a preto ostali bunky nesfarbené.

4.1.6 Ukážkový príklad

Uvedený teoretický postup sa môže zdať nejasný, preto vymenované kroky opíšeme na konkrétnej obci a frakcii. Obcou, ktorú sme si zvolili, je **Borotín** a frakciou je **biologicky rozložiteľný odpad**.

1. **Dixonov test**
   Zvozy boli zaznamenané od 26. týždňa až po 110. týždeň a to znamená, že celkový počet pozorovaní v náhodnom výbere je 85. Testovou štatistikou pre 85 pozorovaní je r_{22} alebo r'_{22}. Po vypočítaní oboch testových štatistík sme ako väčšiu určili $r'_{22} = 0,52383$. Kritickou hodnotou v tabuľkách pre 85 pozorovaní je $r_{tab} = 0,265$. Keďže $r'_{22} > r_{tab}$, môžeme potvrdiť existenciu odľahlého pozorovania v 86. týždni. Toto pozorovanie je následne vyeliminované a nahradené priemerom okolitých hodnôt: $(7,20418 + 5,2761)/2 = 6,24014$.

2. **Štandardizácia**
   Hmotnosť odpadu v každom týždni bola predelená počtom obyvateľov obce Borotín - 431.

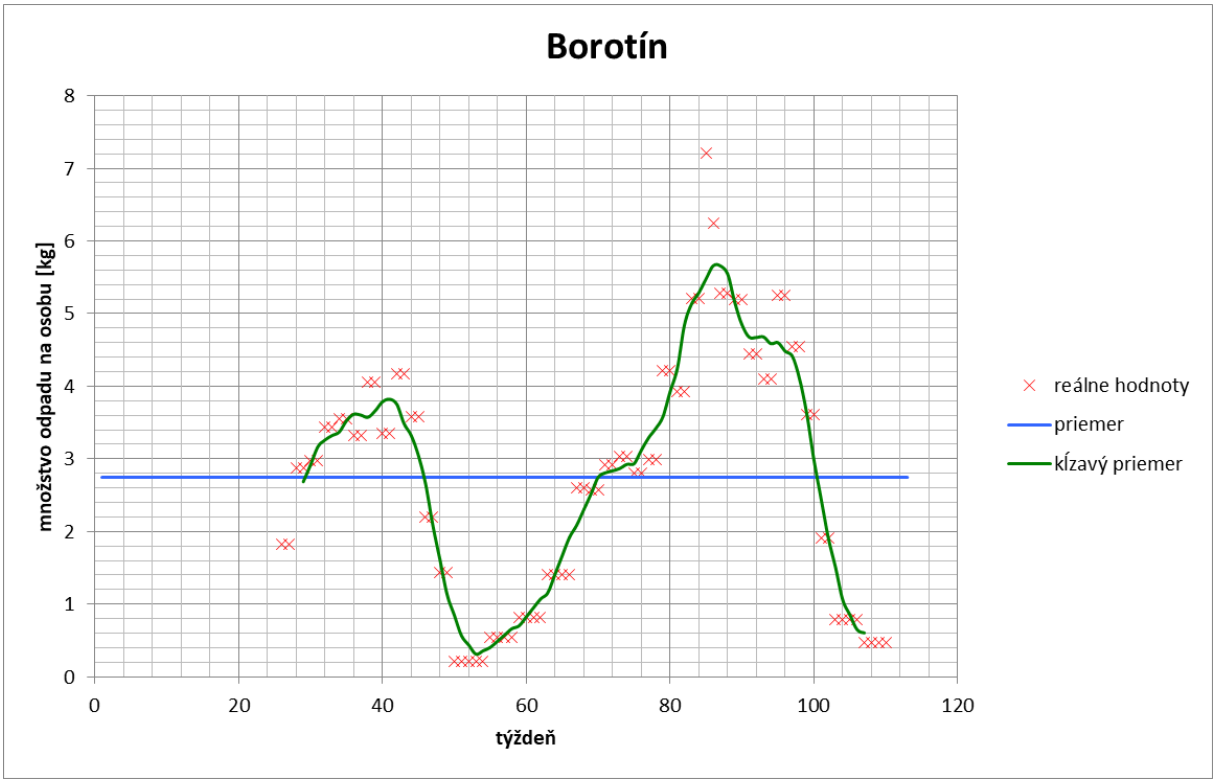
3. **Odhady základných charakteristík dát**

   – výberový priemer: $\overline{X}_n = 2,80$ kg,

   – výberová smerodajná odchýlka: $S_n = 1,85$ kg,

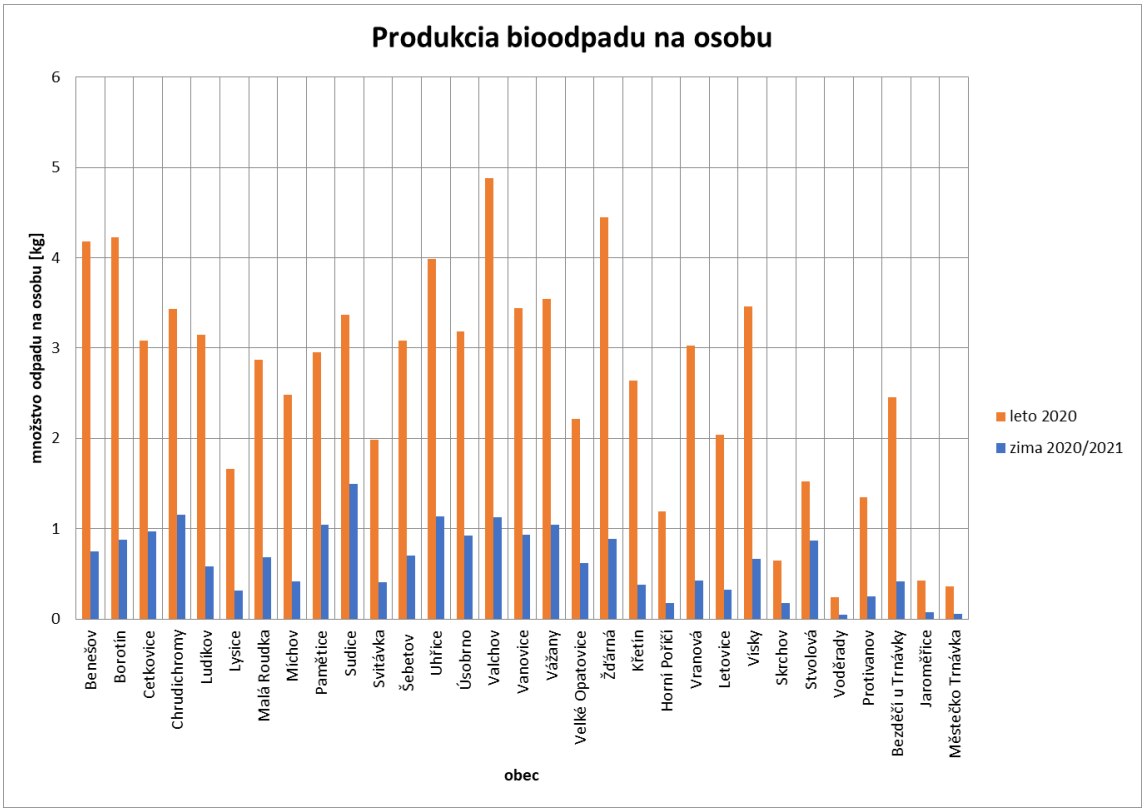
   – variačný koeficient: $\frac{S_n}{\overline{X}_n} = \frac{1,85}{2,80} = 0,66 \Rightarrow$ vysoká variabilita hmotnosti odpadu.

4. **Kĺzavý priemer a určenie sezón**
   Z obr. 4.1 je jasné, že krivka kĺzavého priemeru preťala aritmetický priemer medzi 29. a 30. týždňom (stred júla 2019), medzi 45. a 46. týždňom (stred novembra 2019), medzi 69. a 70. týždňom (koniec apríla 2020) a medzi 100. a 101. týždňom (koniec novembra 2020). Tieto prieniky sú vyvolané prudkými nárastmi a poklesmi v produkcii odpadu (spôsobené sezónnosťou), ktoré nastali napr. medzi 66. a 67. týždňom (začiatok apríla 2020), kde prudko stúpla produkcia odpadu, čo značí začiatok letnej sezóny. Z obrázku je taktiež vidno, že aj po eliminovaní odľahlej hodnoty Dixonovým testom v tomto výbere ostala extrémna hodnota v 86. týždni (stred augusta 2020). Túto hodnotu sa ale podarilo vyhladiť kĺzavým priemerom - zelenou krivkou.



Obr. 4.1: Produkcia biologicky rozložiteľného odpadu v obci Borotín

Podobne ako pri obci Borotín boli sezóny produkcie biologického odpadu identifikované aj v ostatných obciach. Dôkazom sezónnosti je aj stĺpcový graf 4.2, ktorý zobrazuje produkciu na osobu (v kilogramoch) pre každú obec počas dvoch sezón určených prienikom dvoch priemerov. Po porovnaní dvoch zo štyroch dostupných sezón (tieto dve sezóny budú následne testované a modelované v podkapitolách 4.2.2 a 4.2.3) sme zistili, že množstvo odpadu v letnej sezóne je niekoľkonásobne väčšie ako v zimnej.



Obr. 4.2: Porovnanie produkcie biologicky rozložiteľného odpadu v dvoch sezónach

Sezóny vyobrazené v stĺpcovom diagrame boli určené pre každú obec samostatne, no po porovnaní pomocou histogramov v prílohe B vidíme, že prechody medzi sezónami nastávajú v približne rovnakých týždňoch. Za týždeň prechodu medzi zimnou a letnou sezónou môžeme považovať 70. týždeň zvozu, čo predstavuje 18. kalendárny týždeň roku 2020 (t. j. prelom apríla a mája). Za týždeň prechodu medzi letnou a zimnou sezónou zas považujeme 99. týždeň, čo predstavuje 47. kalendárny týždeň roku 2020 (t. j. stred novembra).

4.2 Štatistické modely

Ďalším krokom pri zisťovaní kľúčových parametrov, ktoré ovplyvňujú množstvo odpadu vyprodukovaného a zvozeného z konkrétnych obcí je tvorba štatistických modelov. Z teórie regresnej analýzy vieme, že existuje veľa spôsobov, ktorými môžeme vytvárať lineárne modely a testovať ich významnosť. V tejto časti teoreticky opíšeme využitie analýzy rozptylu, ukážeme využitie jej neparametrickej obdoby a ďalej budeme pracovať s už existujúcim modelom NERUDA Street, ktorý je popísaný napr. v [19].

4.2.1 Analýza rozptylu

Pred začiatkom tvorby modelu analýzy rozptylu musíme skontrolovať splnenie predpokladov tejto metódy – normalitu a homoskedasticitu (predpoklad rovnakých rozptylov všetkých náhodných výberov) dát. Normalitu dát je možné overiť Lillieforsovou variantou Kolmogorovho-Smirnovho testu, spomínanou v podkapitole 1.7.1. Homoskedasticitu zas môžeme overiť Bartlettovým alebo Leveneovým testom opísaným v [11]. Ak nie je splnený aspoň jeden z týchto predpokladov, automaticky prechádzame ku neparametrickej alternatíve analýzy rozptylu – Kruskalovmu-Wallisovmu testu – spomenutému v podkapitole 1.7.4.

Ak sú splnené všetky predpoklady analýzy rozptylu, môžeme pristúpiť k tvorbe modelov. Postupujeme podľa návodu v podkapitolách 1.7.2 a 1.7.3. Nezávislým výberom z normálneho rozdelenia s rovnakým rozptylom v našom prípade rozumieme súbor hmotností zvozeného odpadu z jednotlivých týždňov. Pri jednofaktorovej analýze, kde za jediný faktor považujeme obec, z ktorej odpad pochádza, testujeme hypotézu H_0, že priemerné hmotnosti odpadu na osobu sú v každej z obcí rovnaké proti alternatívnej hypotéze H_1, že priemerná hmotnosť odpadu sa aspoň v jednej obci líši. Pomocou vzorcov uvedených v 1.7.2 alebo pomocou doplnku Data Analysis nachádzajúceho sa v Exceli v karte Údaje je možné vypočítať testovú štatistiku, ktorú následne porovnávame s kritickým oborom. Na základe výsledku testu zvolenú hypotézu zamietame alebo nezamietame.

Pri dvojfaktorovej analýze, kde za faktor nepovažujeme len výber obce, ale aj rok (resp. týždeň), v ktorom bol odpad zvozený, testujeme hypotézu H_0, že hmotnosť odpadu na osobu nezávisí na obci, z ktorej je odpad zvážaný ani na roku, v ktorom bol zvozený, proti hypotéze H_1, že hmotnosť odpadu je závislá na jednom z týchto faktorov. Keďže druhý faktor bol pridaný kvôli navyšujúcemu sa trendu v dátach, ktorý bol pravdepodobne spôsobený pridávaním zberných nádob v jednotlivých obciach, tak predpokladáme, že závislosť na roku bude potvrdená. Tento test môžeme znovu uskutočniť podľa vzorcov z kapitoly 1.7.3 alebo s pomocou doplnku Data Analysis, kde je postup analogický s jednofaktorovou analýzou rozptylu.

4.2.2 Kruskalov-Wallisov test

V stĺpcovom grafe 4.2 môžeme okrem sezónnosti pozorovať aj rozdiel produkcie biologicky rozložiteľného odpadu na osobu medzi obcami a to počas oboch vykreslených sezón. Toto tvrdenie sme sa rozhodli podporiť pomocou Kruskalovho-Wallisovho testu opísaného v podkapitole 1.7.4. Túto variantu sme zvolili z dôvodu zamietnutia jedného z predpokladov jednofaktorovej analýzy rozptylu - normality dát. Vybraný test sme aplikovali osobitne na letnú a zimnú sezónu.

K tomuto testu sme si vybrali obec Svitávka a obec Vanovice. Už na prvý pohľad môžeme zo stĺpcového grafu vidieť, že produkcia je v oboch sezónach v obci Svitávka nižšia. Najprv sme pristúpili k spomínanému testu normality. Tento test sme uskutočnili v softvéri MATLAB, v ktorom je implementovaná funkcia `lillie.test`, ktorá rozhoduje o normalite príslušných náhodných výberov pomocou Lillieforsovej varianty Kolmogorovho-Smirnovho testu, vysvetlenej v podkapitole 1.7.1. Uvažované náhodné výbery sú zobrazené v tabuľke v prílohe C. Nulovú hypotézu H_0, že náhodný výber pochádza z normálneho rozdelenia sme zamietli pre všetky štyri stĺpce tabuľky.

Po vylúčení možnosti použitia analýzy rozptylu sme prešli k samotnému Kruskalovmu-Wallisovmu testu. Testovali sme nulovú hypotézu $H_0 : F_1(x) = F_2(x), \forall x$, kde $F_1(x)$ je distribučná funkcia náhodného výberu z obce Svitávka a $F_2(x)$ je distribučná funkcia náhodného výberu z obce Vanovice, oproti alternatívnej hypotéze H_1, že H_0 neplatí. Pomocou postupu uvedeného v podkapitole 1.7.4 sme spočítali príslušné charakteristiky a nakoniec aj testové štatistiky:

leto 2020	zima 2020/2021
$n_1 = 17$	$n_1 = 11$
$n_2 = 32$	$n_2 = 13$
$N = n_1 + n_2 = 49$	$N = n_1 + n_2 = 24$
$T_1 = 217$	$T_1 = 88$
$T_2 = 1008$	$T_2 = 212$
$Q = 19,087$	$Q = 8,225$

V oboch prípadoch sme testovú štatistiku Q porovnávali s hodnotou $\chi^2_{0,05}(1) = 3,841$, ktorú sme dohľadali v štatistických tabuľkách. Keďže obe Q boli väčšie ako daný kvantil χ^2 rozdelenia, hypotézu H_0, že náhodné výbery majú rovnaké distribučné funkcie, sme zamietli na hladine významnosti $\alpha = 0,05$.

Skutočnosť, že náhodné výbery majú rozdielnu distribučnú funkciu znamená, že rozdielny je aj výberový priemer. Rozdiel v priemernej produkcii na obyvateľa v obciach Svitávka a Vanovice je s najväčšou pravdepodobnosťou spôsobený začiatkom zvozu biologicky rozložiteľného odpadu v týchto obciach. V obci Svitávka bol prvý zvoz zaznamenaný v júli 2020, zatiaľ čo zvoz v obci Vanovice prebiehal už od júna 2019. Rozdiely kvôli rôznemu začiatku zvozu môžeme pozorovať pri väčšine obcí zobrazených na grafe 4.2. Jednou z výnimiek je ale napríklad dvojica obcí Vanovice a Vážany, v ktorých zvoz začal v tom istom období a priemerná produkcia v oboch sezónach je približne rovnaká.

4.2.3 Model NERUDA Street

Reálne dáta spracované v tejto práci boli zatiaľ len upravované a štatisticky testované. Je však vhodné ukázať, kde a ako môžeme takto spracované dáta využiť. Túto ukážku vykonáme pomocou existujúceho modelu NERUDA Street. Ukážka bude predstavovať optimálne trasy zvolené spomínaným modelom na základe množstva odpadu vyzbieraného v každej obci. Pomocou modelu je možné tvoriť návrhy ľubovoľných zvozov, ktoré sa môžu dynamicky meniť s ohľadom na produkciu v danom období a tým ušetriť celkové náklady. Graf 4.2 popisuje priemernú produkciu biologicky rozložiteľného odpadu na osobu pre dve sezóny. Medzi týmito sezónami pozorujeme veľký rozdiel v produkcii v každej jednej obci. Kvôli tomuto zisteniu je vhodné pre ďalšie plánovanie zberu a zvozu odpadu uvažovať dve varianty - letnú a zimnú. Vytvoríme teda návrh zvozu biologicky rozložiteľného odpadu v letnej sezóne 2020 a zimnej sezóne 2020/2021.

Pre tento zvoz sme zvolili vozidlo s hmotnostnou kapacitou 10 ton, kompostáreň – miesto, kde je zvozený odpad vykladaný – v obci Krhov a depo – miesto, odkiaľ vyrážajú vozidlá – v obci Sudice. Vo výpočte nebolo uvažované časové obmedzenie a to znamená, že podľa doby zvozu by museli byť jednotlivé trasy rozdelené v rámci dostupného vozového parku. V tabuľkách 4.1 a 4.2 sú popísané zvolené optimálne trasy vypočítané modelom

NERUDA Street po zadaní vstupných parametrov popísaných v tabuľke D.1 nachádzajúcej sa v prílohe D.

V rámci letnej sezóny 2020 bolo zvolených celkovo 7 výjazdov s celkovou dĺžkou 433,773 km a celkovou hmotnosťou odpadu 61 062 kg. Týmito výjazdmi bolo spojených 31 obcí. Po prezretí dát sme narazili na problém s obcou Letovice, v ktorej množstvo vyzbieraného odpadu za letnú sezónu presahovalo kapacitu nami vybratého vozidla. Túto obec sme rozdelili do dvoch zvozov a preto je v tabuľke D.1 uvedených 32 obcí + depo. Ku každej trase popísanej pomocou ID v tabuľke 4.1 prislúcha jedna farebne vyznačená trasa na obr. D.1 v prílohe D.

Pre bližší opis si zvolíme napríklad výjazd č. 1. Výjazd č. 1 začal ako každý z výjazdov v depe v obci Sudice. Následne bol odpad vyzbieraný z nasledujúcich obcí v tomto poradí: Sudice (11), Vážany (18), Šebetov (13), Vanovice (17), Pamětice (10), Vísky (26) a Chrudichromy (5). Odpad s celkovou hmotnosťou 9 230 kg bol zvozený do kompostárne v obci Krhov. Celková dĺžka tejto trasy bola 38,153 km. Na obr. D.1 je výjazd č. 1 označený svetlozelenou krivkou. Analogicky boli vytvorené aj ostatné trasy.

č. výjazdu	trasa	náklad [kg]	naplnenosť vozu [%]	prejdená vzdialenosť [km]
1	11-18-13-17-10-26-5	9 230	92,30	38,153
2	12-9-25-27-28-22-21	9 719	97,19	61,680
3	24	10 000	100,00	30,094
4	33-31-32-15-14-4-2	9 112	91,12	106,141
5	6-16-20-30-3	9 876	98,76	82,619
6	8-19-23	9 844	98,44	78,146
7	29-7	3 281	32,81	36,940

Tabuľka 4.1: Návrh zvozu za leto 2020

V rámci zimnej sezóny 2020/2021 boli zvolené celkovo 2 výjazdy s celkovou dĺžkou 246,035 km a celkovou hmotnosťou odpadu 13 537 kg. Tak ako v prípade letnej sezóny bolo týmito výjazdmi spojených 31 obcí. Ku každej trase popísanej pomocou ID v tabuľke 4.2 prislúcha jedna farebne vyznačená trasa na obr. D.2 v prílohe D. Opis trás zimnej sezóny je analogický s opisom trás pre letnú sezónu.

č. výjazdu	trasa	náklad [kg]	naplnenosť vozu [%]	prejdená vzdialenosť [km]
1	11-18-10-26-24-23-21-22-28-27-13-4-14-15-32-31-33-19-8-3-17-9	9 803	98,03	150,082
2	29-12-5-2-30-20-16-6-7	3 734	37,34	95,953

Tabuľka 4.2: Návrh zvozu za zimu 2020/2021

5 Zhrnutie

Podstatnú časť praktickej časti tejto bakalárskej práce tvorila príprava dát. Poskytnuté dáta bolo potrebné premeniť do podoby vhodnej na ďalšie spracovanie. To sme docielili prerozdelením do týždňov a frakcií a následným eliminovaním odľahlých pozorovaní.

Ako prvé sme skúmali odhady základných charakteristík dát, ktoré nám umožnili prvý náhľad do problematiky. Pomocou týchto charakteristík sme dokázali porovnať ktorékoľvek dve obce a zistiť, ako sa produkcia odpadu v týchto obciach líši. Jednou z vecí, ktorá nás zaujala bola sezónnosť produkcie. Po spracovaní dát kĺzavým priemerom sa však ukázalo, že jediná frakcia, u ktorej sezónnosť v priebehu roku môžeme pozorovať, je biologicky rozložiteľný odpad. Pri ostatných typoch odpadu, ktoré boli spracované kĺzavým priemerom sa ukazoval nepravidelný charakter zmeny produkcie - dáta oscilovali okolo aritmetického priemeru. V týchto frakciách teda nemôžeme pozorovať letnú a zimnú sezónu. Na štatistické testy a modelovanie zvozových trás sme si preto zvolili biologicky rozložiteľný odpad. Ostatné typy odpadu sme netestovali a predpokladali sme konštantnú produkciu v priebehu celého roka. Pri týchto dátach môžeme porovnávať len priemernú produkciu na osobu zobrazenú v tabuľke v prílohe E. Prázdne bunky znamenajú, že obec tento typ odpadu nezvážala, alebo že sme hodnoty v priebehu post-processingu z tejto tabuľky vylúčili, lebo boli skreslené.

Pri pohľade na stĺpcový graf v podkapitole 4.1.4 je viditeľný rozdiel v produkcii biologicky rozložiteľného odpadu počas letnej a zimnej sezóny. Túto skutočnosť sme preto prehlásili za zrejmú a pustili sme sa do skúmania rozdielu v produkcii tohto odpadu v rámci jednej sezóny. Zo spomínaného grafu môžeme usúdiť, že rozdiel medzi niektorými obcami je významný. Toto tvrdenie sme chceli podložiť štatistickým testom. Po vykonaní testu medzi dvomi konkrétnymi obcami sme dospeli k záveru, že priemerná produkcia skúmaného odpadu sa líši. Ako dôvod sme uviedli začiatok zvozu odpadu, ktorý bol v jednej obci posunutý o približne rok. Toto je spôsobené tým, že TSMH sú relatívne nový zväzok, ktorý je len v procese vývoja a nám poskytnuté dáta neboli teda tak objektívne ako by boli napr. po desiatich rokoch fungovania zväzku. To nám ale nebránilo využiť zistené poznatky pri plánovaní reálneho zvozu.

Na plánovanie reálneho zvozu sme využili existujúci nástroj NERUDA Street. Aj v tomto prípade sme teda pracovali s dátami, rozdelenými na sezóny a preto sme zvozy rozdelili do dvoch období, ktoré korešpondujú s týmito sezónami. Menšia produkcia odpadu v zimnom období znamenala menší počet uskutočnených výjazdov. Navrhnuté výjazdy však boli len prvou ukážkou, ako sa dajú spracované dáta využiť na optimalizáciu trasy zvozu.

V ďalšom postupe by sme mohli uvažovať optimalizáciu obmedzenú viacerými parametrami, ako sú napr. čas zvozu, odhad času a dĺžky zvozu v rámci každej obce, no taktiež by sa mohla uvažovať dostupnosť vozového parku a kapacita dostupných vozidiel, ktoré by sme chceli využiť. Prepracovanejšia optimalizácia zvozových trás na základe spracovaných dát, ktorá by viedla k zníženiu nákladov za zber a zvoz odpadu pre obce, môže byť predmetom nadväzujúcej diplomovej práce.

Záver

Cieľom tejto práce bolo spracovanie a analýza množstva reálnych dát, pochádzajúcich z reálnych zvozov odpadu vykonaných spoločnosťou Technické služby Malá Haná s.r.o. v priebehu takmer dvoch rokov.

Prvá kapitola bola venovaná opisu teórie z pravdepodobnosti a matematickej štatistiky, v ktorej sme sa oboznámili s pojmami ako sú náhodná veličina, náhodný výber či bodový a intervalový odhad. Taktiež sme tu uviedli postupy testovania hypotéz a tvorby modelov pomocou analýzy rozptylu. V druhej kapitole sme predstavili softvér, ktorý sme využívali pri zostavovaní celej praktickej časti tejto práce. Tretia kapitola sa venovala príprave dát, ktorá začala rozdelením zvozov do týždňov, vytvorením kontingenčnej tabuľky a rozdelením zvozov do frakcií. V týchto frakciách potom pokračovala finálna úprava, ktorá pozostávala z rozpočítania produkcie do týždňov, v ktorých sa nekonal zvoz. V nasledujúcej kapitole boli využité predpripravené dáta z tretej kapitoly pri štatistických testoch, ktoré viedli k identifikácii kľúčových parametrov ovplyvňujúcich výsledky zvozov. Prvým z týchto testov bol Dixonov test, slúžiaci na elimináciu hodnôt, ktoré sa príliš odlišujú od ostatných a ktoré nazývame odľahlé pozorovania. Po štandardizácii nasledoval výpočet variačného koeficientu, podľa ktorého je možné porovnať variabilitu množstva zvozeného odpadu medzi obcami. Posledným krokom bolo využitie kĺzavého priemeru na rozpoznanie sezónnosti pri určitých frakciách. V post-processingu sa už len ručne dolaďovali detaily, ktoré makrá v Exceli nedokázali spracovať. Druhú polovicu tejto kapitoly tvoril opis tvorby štatistických modelov a testovanie hypotéz známe pod názvom analýza rozptylu. Tento opis bol čisto teoretický, doplnený testovaním hypotézy pomocou Kruskallovho-Wallisovho testu a praktické využitie bolo ukázané na existujúcom modeli NERUDA Street. Posledná kapitola slúžila na zhrnutie celej práce a boli v nej spomenuté návrhy, ako je možné ju ďalej rozvíjať.

Výstupom z tejto práce sú bližšie poznatky o zbere a zvoze dát. V prípade biologicky rozložiteľného odpadu, ktorý sme skúmali najpodrobnejšie, sme zistili významné rozdiely produkcie v priebehu roku. V rámci všetkých frakcií sme identifikovali veľkú variabilitu v produkcii odpadu medzi skúmanými obcami, čo mohlo byť spôsobené rozdielnym typom zvozu v týchto obciach. Zatiaľ čo v niektoých funguje zvoz len zo zberných miest, v iných je to už zvoz „door to door“, teda z každej domácnosti samostatne. Variabilita mohla byť však zapríčinená aj dobou zvozu v jednotlivých obciach, keďže v obciach s dlhším zvozom sú občania už zvyknutý separovať odpad. Detailnejšia analýza, v ktorej by sme sa zamerali na produkciu odpadu v jednotlivých zberných miestach, by nám mohla priniesť ešte presnejšie dáta pre zvozové úlohy.

Literatúra

[1] ANDĚL, Jiří. *Matematická statistika*. Praha: SNTL, 1985, 352 s.

[2] ANDĚL, Jiří. *Základy matematické statistiky*. Praha: Matfyzpress, 2005. ISBN 80-867-3240-1.

[3] ČESKO. *Hierarchie způsobů nakládání s odpady*, Zákon č. 185/2001 Sb. ze dne 15. května 2001 o odpadech a o změně některých dalších zákonů. In: *Sbírka zákonů České republiky*. Dostupné z: https://www.zakonyprolidi.cz/cs/2001-185

[4] ČESKO. Zákon č.541/2020 Sb. o odpadech. In: *Sbírka zákonů České republiky*. Dostupné z: https://www.zakonyprolidi.cz/cs/2020-541

[5] ČESKÝ STATISTICKÝ ÚŘAD [ČSÚ]. *Databáze demografických údajů za obce ČR*. [online]. 30. 4. 2020 [cit. 2021-4-26]. Dostupné z: https://www.czso.cz/csu/czso/databaze-demografickych-udaju-za-obce-cr

[6] DIXON, Juan W. Analysis of Extreme Values. *The Annals of Mathematical Statistics*. 1950, **21**(4), 488-506.

[7] DIXON, Juan W. Ratios involving extreme values. *The Annals of Mathematical Statistics*. 1951, **22**(1), 68-78.

[8] FERNANDO, Jason. Moving Average (MA). *Investopedia* [online]. New York City: Dotdash, 17. 1. 2021 [cit. 2021-4-26]. Dostupné z: https://www.investopedia.com/terms/m/movingaverage.asp

[9] GREGOR, J., R. ŠOMPLÁK a M. PAVLAS. Transportation Cost as an Integral Part of Supply Chain Optimisation in the Field of Waste Management. *Chemical Engineering Transactions* [online]. 2017, 20. 3. 2017, Vol 56, 1927-1932 [cit. 2021-4-26]. ISSN 2283-9216. Dostupné z: https://www.cetjournal.it/index.php/cet/article/view/CET1756322

[10] GRUBBS, Frank E. Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*. 1950, **21**(1), 27-58.

[11] HOLČÍK, Jiří, KOMENDA, Martin (eds.) a kol. *Matematická biologie: e-learningová učebnice* [online]. 1. vydání. Brno: Masarykova univerzita, 2015 [cit. 2021-4-26]. ISBN: 978-80-210-8095-9. Dostupné z: https://portal.matematickabiologie.cz/

[12] HYNDMAN, R.J. a G. ATHANASOPOULOS. *Forecasting: principles and practice* [online]. 2nd edition. Melbourne, Australia: OTexts, 2018 [cit. 2021-5-12]. Dostupné z: OTexts.com/fpp2

[13] KARPÍŠEK, Zdeněk. *Matematika IV: statistika a pravděpodobnost*. Vyd. 2., dopl. Brno: Akademické nakladatelství CERM, 2003. ISBN 80-214-2522-9.

[14] KOTLORZ, Lukáš. *Testy normality*. Praha, 2012. Bakalárska práca. Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, Katedra pravděpodobnosti a matematické statistiky. Vedoucí bakalářské práce prof. RNDr. Jiří Anděl, DrSc.

[15] KUNDEROVÁ, Pavla. *Základy pravděpodobnosti a matematické statistiky*. Olomouc: Univerzita Palackého, 2004. ISBN 80-244-0813-9.

[16] LASÁK, Pavel. *Jak na Microsoft Office (Excel,...)* [online]. [cit. 2021-4-26]. Dostupné z: https://office.lasakovi.com/

[17] LILLIEFORS, Hubert W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*. 1967, **62**(318), 399-402.

[18] MICHÁLEK, Jaroslav. *Pravděpodobnost a statistika*. Preprint. Brno, 2006. Dostupné z: https://adoc.pub/jaroslav-michalek-a-statistika.html

[19] NEVRLÝ, V. *Komplexní modely svozu odpadu*. Brno: Vysoké učení technické, v Brně, Fakulta strojního inženýrství, 2020. 76 s. Školitel: prof. Ing. Petr Stehlík CSc., dr. h.c.

[20] Popisná statistika - míry variability. *Statsoft ACADEMY* [online]. 15/10/2012 [cit. 2021-5-12]. Dostupné z: http://www.statsoft.cz/o-firme/archiv-newsletteru/newsletter-15102012/

[21] QUIROZ RUIZ, Alfredo a Surendra P. VERMA. Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. *Revista mexicana de ciencias geológicas*. 2006, **23**(2), 133-161. ISSN 2007-2902.

[22] Směrnice Evropského parlamentu a Rady (EU) 2018/849 ze dne 30. května 2018, kterou se mění směrnice 2000/53/ES o vozidlech s ukončenou životností, 2006/66/ES o bateriích a akumulátorech a odpadních bateriích a akumulátorech a 2012/19/EU o odpadních elektrických a elektronických zařízeních. In: *Úřední věstník*. Dostupné z: https://eur-lex.europa.eu/legal-content/cs/TXT/?uri=CELEX%3A32018L0849

[23] Směrnice Evropského parlamentu a Rady (EU) 2018/850 ze dne 30. května 2018, kterou se mění směrnice 1999/31/ES o skládkách odpadů. In: *Úřední věstník*. Dostupné z: https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=celex:32018L0850

[24] Směrnice Evropského parlamentu a Rady (EU) 2018/851 ze dne 30. května 2018, kterou se mění směrnice 2008/98/ES o odpadech. In: *Úřední věstník*. Dostupné z: https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX%3A32018L0851

[25] Směrnice Evropského parlamentu a Rady (EU) 2018/852 ze dne 30. května 2018, kterou se mění směrnice 94/62/ES o obalech a obalových odpadech. In: *Úřední věstník*. Dostupné z: https://eur-lex.europa.eu/legal-content/cs/TXT/?uri=CELEX%3A32018L0852

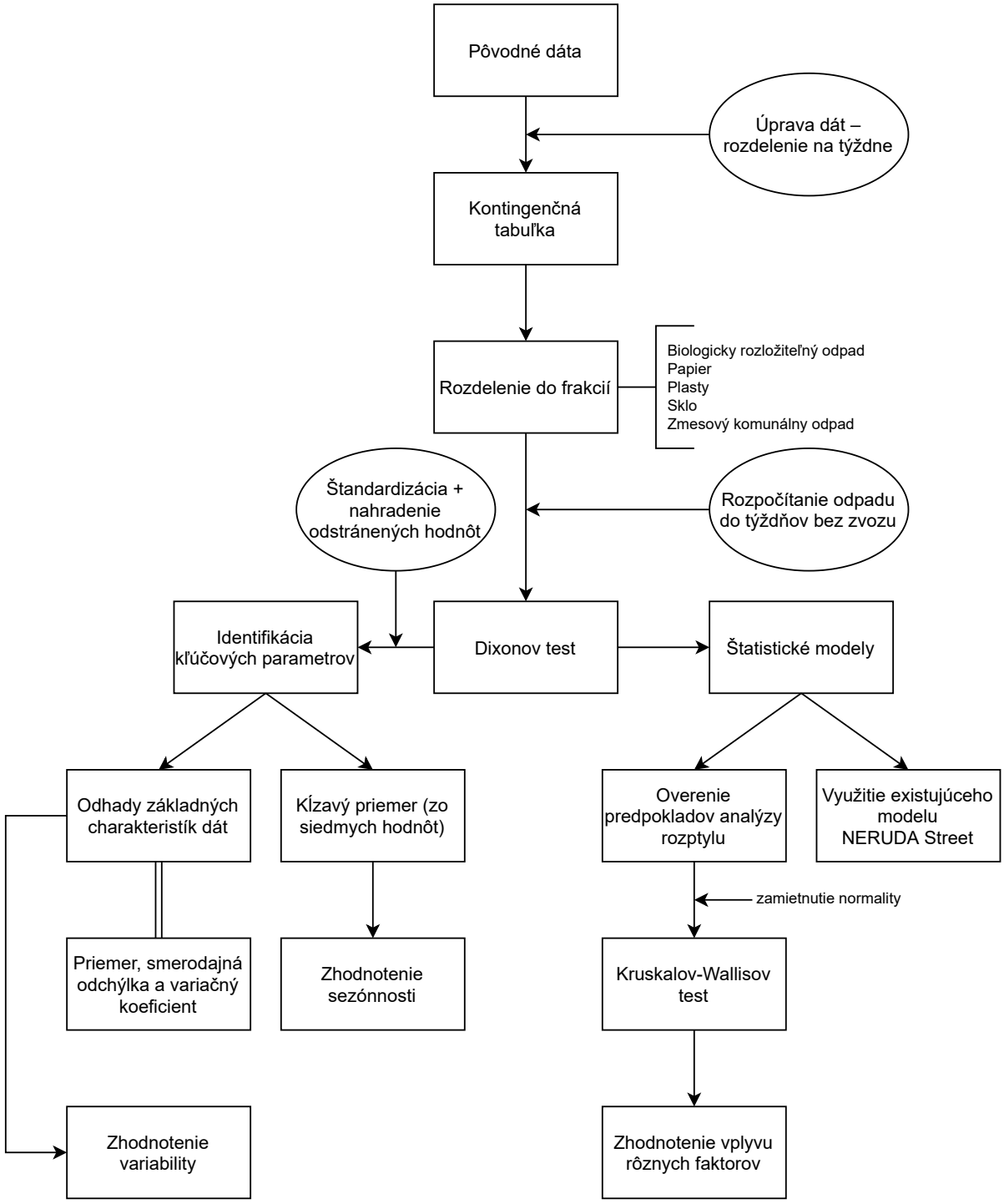
[26] ZVÁRA, Karel a Josef ŠTĚPÁN. *Pravděpodobnost a matematická statistika*. Vyd. 3. Praha: Matfyzpress, 2002. ISBN 80-858-6393-6.

Zoznam skratiek

ANOVA	analýza rozptylu (z anglického *Analysis of Variance*)
EÚ	Európska únia
ID	identifikačné číslo
KO	komunálny odpad
MS	Microsoft
SARIMA	sezónny autoregresný integrovaný kĺzavý priemer (z anglického *Seasonal Autoregressive Integrated Moving Average*)
TSMH	Technické služby Malá Haná s.r.o.
VBA	Visual Basic for Application

Prílohy

A Vývojový diagram postupu spracovania dát



B Histogramy

Obr. B.1: Histogram prechodu zo zimnej sezóny na letnú (70. týždeň = 28.4.–4.5.2020)

Obr. B.2: Histogram prechodu z letnej sezóny na zimnú (99. týždeň = 17.–23.11.2020)

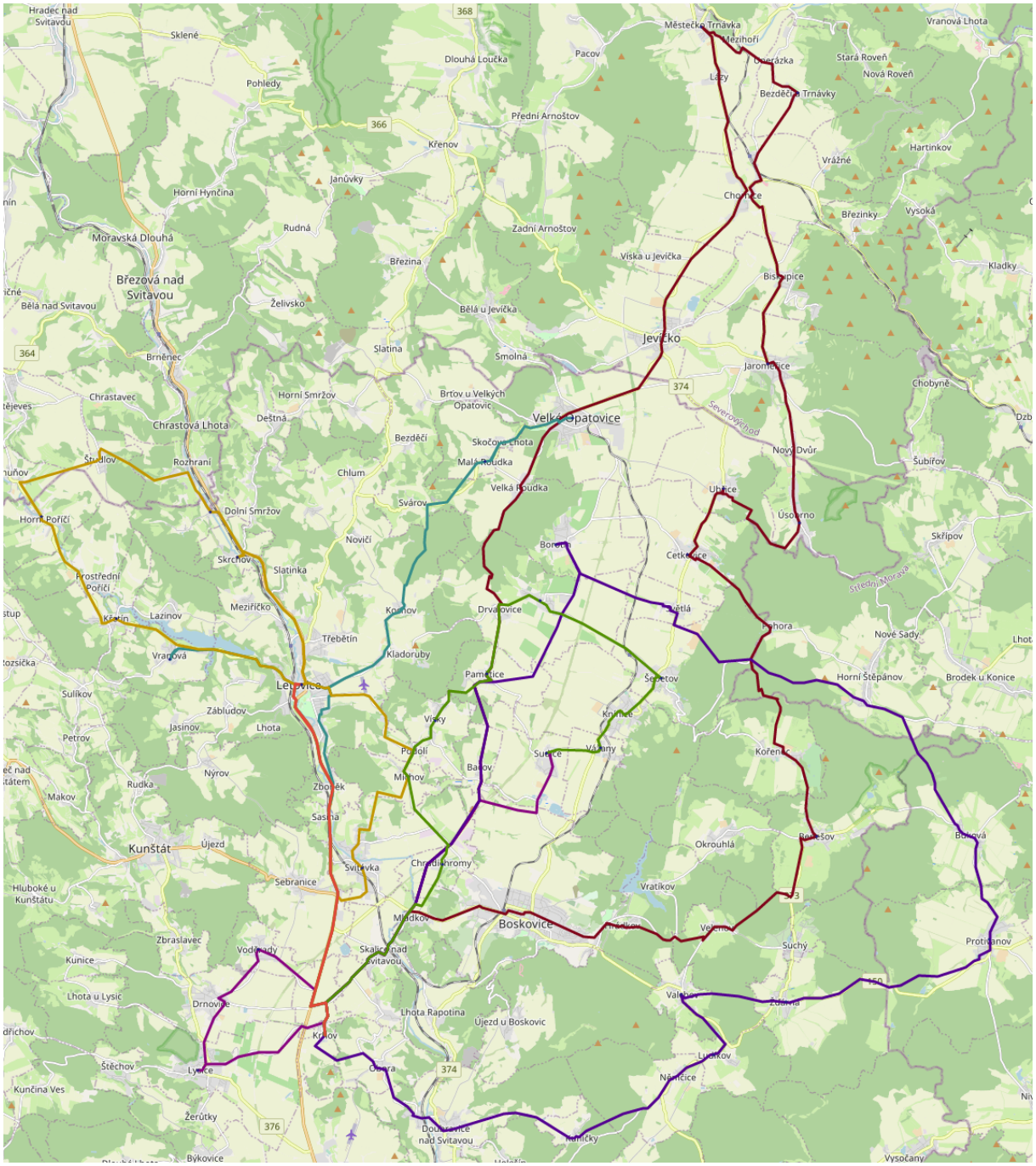
C Náhodné výbery pre Kruskallov-Wallisov test

leto 2020 [kg/týždeň]		zima 2020/2021 [kg/týždeň]	
Svitávka	Vanovice	Svitávka	Vanovice
1,572057	2,58046	1,163366	2,074713
1,572057	2,19636	0,665567	0,937261
1,993125	2,19636	0,665567	0,937261
1,993125	3,284483	0,245943	0,937261
0,983086	3,284483	0,245943	0,937261
0,983086	2,407088	0,245943	0,178161
0,983086	2,407088	0,245943	1,09834
0,983086	3,67433	0,245943	1,09834
1,737349	3,67433	0,245943	1,09834
1,737349	3,035441	0,245943	0,70546
3,657316	3,035441	0,245943	0,70546
3,657316	3,780651		0,70546
3,619912	3,780651		0,70546
3,619912	3,381226		
1,709296	3,381226		
1,709296	4,118774		
1,163366	4,118774		
	4,454981		
	4,454981		
	4,175287		
	4,175287		
	3,734674		
	3,734674		
	3,848659		
	3,848659		
	3,616858		
	3,616858		
	3,85728		
	3,85728		
	3,239464		
	3,239464		
	2,074713		

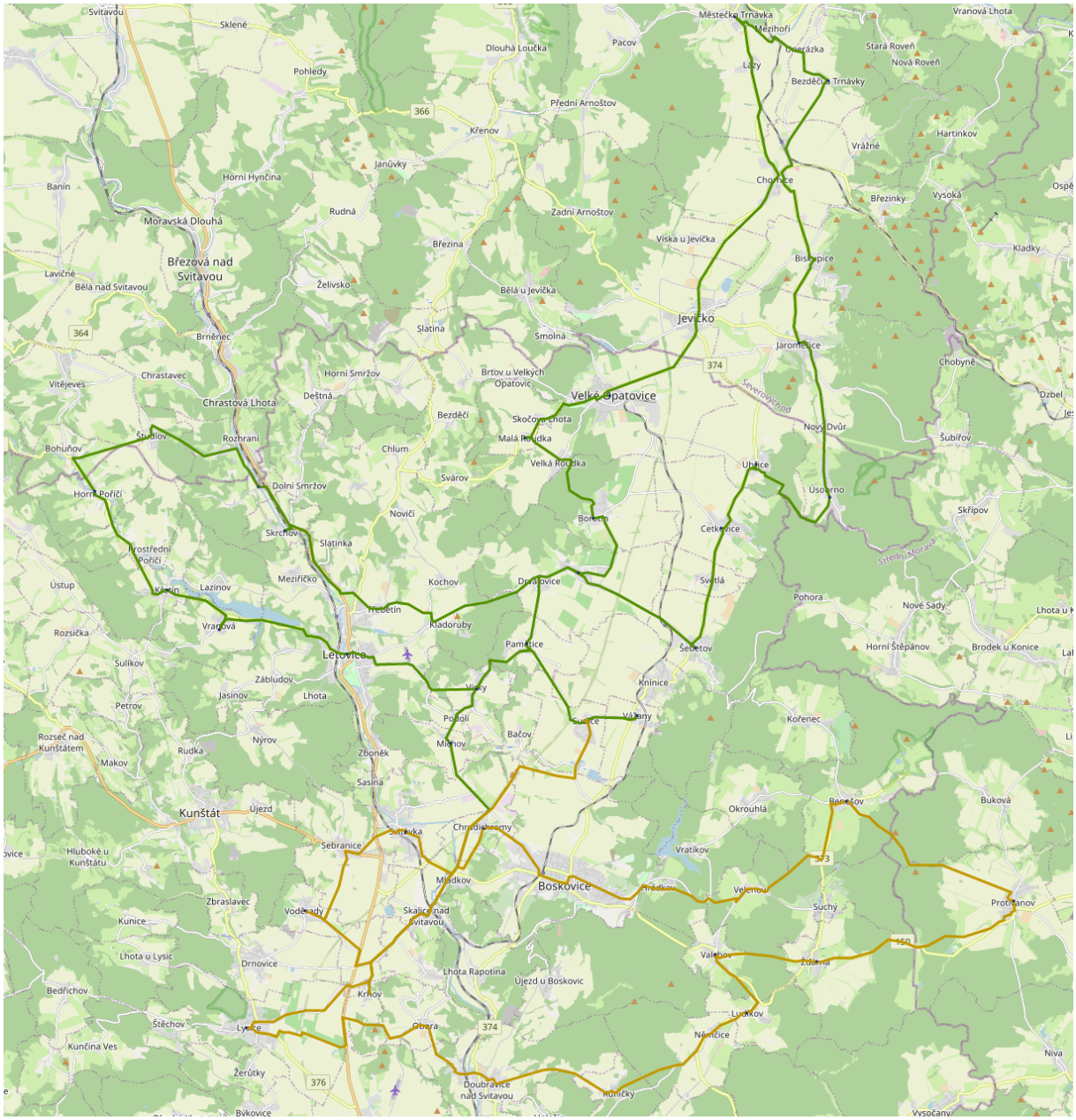
D Vstupy a výstupy modelu NERUDA Street

ID	názov obce	produkcia - leto 2020 [kg]	produkcia - zima 2020/21 [kg]
0	Krhov - kompostáreň	0	0
1	Sudice - depo	0	0
2	Benešov	2 716	486
3	Borotín	1 822	381
4	Cetkovice	2 345	736
5	Chrudichromy	705	236
6	Ludíkov	1 060	198
7	Lysice	3 151	605
8	Malá Roudka	611	146
9	Míchov	424	72
10	Pamětice	756	267
11	Sudice	1 609	715
12	Svitávka	3 601	737
13	Šebetov	2 662	607
14	Uhřice	1 225	348
15	Úsobrno	1 360	393
16	Valchov	2 197	506
17	Vanovice	1 799	487
18	Vážany	818	242
19	Velké Opatovice	8 075	2 247
20	Žďárná	3 412	684
21	Křetín	1 291	188
22	Horní Poříčí	330	48
23	Vranová	1 158	163
24	Letovice	10 000	2 171
25	Letovice	3 754	0
26	Vísky	881	170
27	Skrchov	70	20
28	Stvolová	249	143
29	Voděrady	130	26
30	Protivanov	1 385	256
31	Bezděčí u Trnávky	461	78
32	Jaroměřice	498	94
33	Městečko Trnávka	507	87

Tabuľka D.1: Vstupné dáta pre model NERUDA Street



Obr. D.1: Mapa navrhovaného zvozu za leto 2020



Obr. D.2: Mapa navrhovaného zvozu za zimu 2020/2021

E Priemerná produkcia odpadu na osobu za týždeň

	PAPIER [kg]	PLAST [kg]	SKLO [kg]	ZMESOVÝ [kg]
Bělá u Jevíčka	0,374978	0,566089	0,304762	2,410673
Benešov	0,406029	0,575065	0,264978	1,879535
Bezděčí u Trnávky	0,318896	0,500179	0,288124	2,121135
Borotín	0,487949	0,52845	0,356708	2,87452
Cetkovice	0,435	0,607346	0,282294	2,155646
Drnovice	0,071008	0,319131		2,983538
Horní Poříčí	0,039444	0,278172	0,332758	2,063947
Chrastavec	0,051241	0,299931	0,263256	
Chrudichromy	0,521704	0,626487	0,341532	3,300702
Jaroměřice	0,147763	0,34326	0,174017	1,403611
Knínice	0,11938	0,281211	0,149806	2,56066
Křetín	0,008252	0,274248	0,303795	2,138843
Lazinov				3,974804
Letovice	0,285	0,318426	0,282527	0,989362
Ludíkov	0,289659	0,461301	0,301274	2,446849
Lysice	0,319985	0,39294	0,310624	1,156951
Malá Roudka	0,400745	0,489768	0,310391	3,553854
Městečko Trnávka	0,367753	0,583101	0,355716	2,580822
Míchov	0,226946	0,278211	0,285958	3,275777
Nýrov	0,066498	0,297281	0,22314	0,092685
Pamětice	0,407545	0,582006	0,27815	1,786423
Petrov	0,027251	0,196287	0,244709	0,447348
Prostřední Poříčí	0,015056	0,19986	0,330266	2,090474
Protivanov	0,019455	0,302495		
Rozhraní	0,083857	0,300146	0,318322	
Rozsíčka	0,088028	0,386717	0,41189	2,391321
Rozstání		0,452751	0,21997	0,226383
Sebranice	0,020162	0,379785	0,099603	
Skrchov	0,219145	0,126966	0,25557	2,899557
Stvolová	0,049119	0,272349	0,29528	
Sudice	0,506496	0,620836	0,305688	2,827513
Sulíkov	0,026586	0,179927	0,240683	2,639684
Světlá	0,211289	0,227211	0,39745	4,285351
Svitávka	0,355922	0,242411	0,291769	1,534524
Šebetov	0,249135	0,416252	0,265525	1,895726
Študlov	0,097739	0,389873	0,451317	
Uhřice	0,429849	0,583228	0,19544	2,638383
Úsobrno	0,476055	0,673567	0,415144	2,224562
Valchov	0,3748	0,473646	0,251724	2,284401
Vanovice	0,452613	0,631797	0,415709	2,669571
Vážany	0,352392	0,5331	0,146859	2,355728
Velké Opatovice	0,38654	0,355593	0,243799	2,527368
Vísky		0,224038	0,195352	3,769804
Voděrady	0,223305	0,265731		3,418437
Vranová	0,054395	0,283709	0,298481	2,797191
Žďárná	0,336182	0,555412	3,937779	3,010152