ŽILINSKÁ UNIVERZITA V ŽILINE
FAKULTA RIADENIA A INFORMATIKY

Analýza rozsiahlych dát v energetických a dopravných aplikáciách

DIZERTAČNÁ PRÁCA

2020 Ing. Milan Straka

ŽILINSKÁ UNIVERZITA V ŽILINE
FAKULTA RIADENIA A INFORMATIKY

Analýza rozsiahlych dát v energetických a dopravných aplikáciách

DIZERTAČNÁ PRÁCA
28360020203008

Študijný program:	inteligentné informačné systémy
Študijný odbor:	informatika
Školiace pracovisko:	Katedra matematických metód a operačnej analýzy
	Fakulta riadenia a informatiky
	Žilinská univerzita v Žiline
Vedúci práce:	prof. Ing. Ľuboš Buzna, PhD.

Žilina 2020 **Ing. Milan Straka**

Anotácia

Typ práce:	Dizertačná práca
Akademický rok:	2019/2020
Názov práce:	Analýza rozsiahlych dát v energetických a dopravných aplikáciách
Autor:	Ing. Milan Straka
Školiteľ:	prof. Ing. Ľuboš Buzna, PhD.
Jazyk:	Slovenčina
Počet strán:	147
Počet obrázkov:	28
Počet tabuliek:	21
Počet referencií:	143
Kľúčové slová:	analýza dát elektrické vozidlá nabíjacie stanice dátová veda strojové učenie elektromobilita

Poďakovanie

Touto cestou by som sa chcel úprimne poďakovať za pomoc, odborné vedenie, cenné rady, pripomienky a množstvo času prof. Ing. Ľubošovi Buznovi, PhD., ktoré mi venoval, ako aj obrovskú trpezlivosť.

Tiež by som sa chcel veľmi pekne poďakovať Dr. Ruiovi Carvalhovi za cenné pripomienky pri spracovaní dát a písaní spoločnej publikácie.

Dr. Pasquale de Falcovi za pomoc pri používaní softvéru Matlab a používaní stromových metód.

Gijsovi van der Poelovi za cenné rady v oblasti elektromobility. Firme ElaadNL za poskytnutie dát z nabíjacej infraštruktúry.

Dr. Robertovi van den Hoedovi za umožnenie pobytu na Amsterdam University of Applied Sciences, kde som mohol pracovať s teamom pracujúcim v oblasti elektromobility.

Doc. Ing. Jaroslavovi Sivákovi, PhD. za cenné rady, kritiku a pripomienky pri tvorbe práce.

Ing. Michalovi Gregorovi, PhD. za cenné rady a pripomienky pri tvorbe práce, ako aj pri práci s LaTeX-om.

Takisto ďakujem aj kolegom z Katedry matematických metód a operačnej analýzy za ich cenné rady, ktoré pomohli pri tvorbe tejto práce.

Veľká vďaka patrí aj mojej rodine a priateľom za podporu počas celého štúdia a tvorby tejto práce.

Zoznam použitých skratiek

Väčšina použitých skratiek pochádza z anglického jazyka, kvôli zaužívaniu v hovorovej reči, niekedy aj z dôvodu nedostupnosti slovenských ekvivalentov. Ak sme použili iba anglický názov, dávame do zátvorky, na čo daný termín slúži.

ACF - autokorelačná funkcia (angl. autocorrelation function)

AIC - Acaicovo informačné kritérium (angl. Acaice information criteria)

AUC - plocha pod ROC krivkou (angl. Area under the ROC curve)

BSS - systémy zdieľania bicyklov (angl. bicycle sharing systems)

BIC - Bayesovo informačné kritérium (angl. Bayesian information criteria)

EV - elektrické vozidlo (angl. electric vehicle)

FP - falošne pozitívna hodnota (angl. false positive)

FN - falošne negatívna hodnota (angl. false negative)

GBRT - gradient boosted regression trees (klasifikačná metóda založená na rozhodovacích stromoch)

GIS - geografické informačné systémy

RF - random forest (klasifikačná metóda založená na rozhodovacích stromoch)

MAPE - priemerná absolútna štvorcová chyba (angl. mean absolute percentage error)

MCC - Matthewov korelačný koeficient (angl. Matthews correlation coefficient)

MSE - priemerná štvorcová chyba (angl. mean squarred error)

OLS - metóda najmenší štvorcov (angl. ordinary least squares)

PACF - parciálna autokorelačná funkcia (angl. partial autocorrelation function)

POI - bod záujmu (angl. point of interest)

PM - perzistenčný model (angl. persistence model)

ROC - Receiver operating characteristic (krivka na porovnávanie klasifikátorov)

RSS - suma štvorcov rezíduí (angl. residual sum of squares)

SARIMAX - sezónny autoregresný integrovaný model s kĺzavými priemermi a exogénnymi premennými (seasonal autoregressive model with moving averages and exogenous variables)

VIF - faktor zväčšenia rozptylu (angl. variance inflation factor)

TN - Správne pozitívna hodnota (angl. true negative)

TP - Správne negatívna hodnota (angl. true positive)

Abstrakt

STRAKA, Milan: *Analýza rozsiahlych dát v energetických a dopravných aplikáciách.* [Dizertačná práca] Žilinská Univerzita v Žiline. Fakulta riadenia a informatiky. Katedra matematických metód a operačnej analýzy. - Vedúci dizertačnej práce: prof. Ing. Ľuboš Buzna, PhD. - Žilina: FRI UNIZA, 2020, 147 s.

Elektromobilita je rozvíjajúce sa odvetvie, spadajúce do energetiky a dopravy a prináša vysoký potenciál na zníženie emisií. Jej rozvoj do značnej miery v súčasnosti ovplyvňujú politiky na rôznych úrovniach a ekonomické záujmy a spolieha sa na procesy založené na ľudskom rozhodovaní. Vyššia efektivita investícii do tejto oblasti si vyžaduje účinnú podporu pre rozhodovanie a to aj s využitím prostriedkov dátovej analýzy, čo je ústredná téma tejto dizertačnej práce.

Základným zdrojom sú dáta pochádzajúce z verejnej nabíjacej infraštruktúry, rozmiestnenej po celom území Holandska, ktoré patrí, v tomto ohľade, medzi najrozvinutejšie krajiny sveta. Na základe konzultácií s firmami pôsobiacimi v oblasti elektromobility sme identifikovali tri čiastkové ciele. Hlavným cieľom je vytvoriť metodológiu pre analýzu dát a dátové modelovanie v prostredí elektromobility, ktorá bude mať potenciál zväčšiť množinu dostupných nástrojov pre podporu rozhodovania v oblasti budovania a prevádzky nabíjacej infraštruktúry pre EV, pričom sa sústredíme na tri, nižšie uvedené ciele.

Prvým cieľom je identifikácia skupín podobných nabíjacích staníc a používateľov, za účelom identifikácie vznikajúcich segmentov, na ktoré je potrebné reagovať prostredníctvom obchodných modelov a stratégií. Tento cieľ dosahujeme aplikáciou metód zhlukovania a agregačných prístupov, na základe ktorých identifikujeme 4 interpretovateľné zhluky staníc.

Druhý cieľ sa zameriava na výber vhodných metód a množiny dát pre získanie predikcie spotreby energie nabíjacej infraštruktúry. Takéto predikcie sú aplikovateľné, napríklad, pri nákupe elektrickej energie alebo inteligentnom nabíjaní. Pomocou metód časových radov, strojového učenia a externých prediktorov, predikujeme časopriestorovo agregovanú spotrebu, kde dosahujeme presnosť MAPE približne 12 %.

V treťom cieli sa venujeme identifikácii vhodného umiestnenia nabíjacej infraštruktúry. Využívame rozsiahle GIS dáta, reprezentujúce okolie nabíjacích staníc. Jednoduchými metódami, ktoré sme navrhli, extrahujeme prediktory z GIS dát. Vplyvné prediktory získa-

vame pomocou regresných metód, určených na výber premenných. Pomocou dostupných metód sa vysporadúvame so vzájomnou závislosťou dát, multikolinearitou. V rámci identifikácie vhodného umiestnenia najskôr na základe dát predikujeme popularitu nabíjacej infraštruktúry spolu s interpretáciou vplyvu prediktorov. Následne, aj na základe dát z okolia, sa zameriavame na vysvetlenie priestorovej spotreby energie nabíjacou infraštruktúrou. V praxi môže byť navrhovaná metodológia použitá na podporu rozhodovania pri budovaní novej alebo prípadné rozširovanie existujúcej nabíjacej infraštruktúry.

Kľúčové slová: analýza dát, elektrické vozidlá, nabíjacie stanice, dátová veda, strojové učenie, elektromobilita

Abstract

STRAKA, Milan: *Large scale data analysis in energy and transport applications.* [Dissertation thesis] - University of Žilina. Faculty of Management Science and Informatics. Department of Software Technology. - Supervisor: prof. Ing. Ľuboš Buzna, PhD. - Žilina, FRI UNIZA, 2020, 147 p.

Electromobility is regarded as a modern trend, falling within energetics and transport and brings a high potential for reducing emissions. Its development is currently largely influenced by policies at various levels and economic interests, and it relies on processes based on human decision-making. Higher efficiency of investments in this area requires effective decision support based on the use of data analysis tools, which is the central topic of this dissertation.

The basic source is data from public charging infrastructure located throughout the Netherlands, which is one of the most developed countries in the world in this respect. Based on a detailed literature review and consultations with companies operating in the field of electromobility, we identified three goals. The main goal is to create a methodology for data analysis and data modelling in the electromobility environment, which will have the potential to enrich the set of tools available to support decision-making in the processes of deployment and operation of charging infrastructure for EVs, focusing on the three objectives below.

The first goal is to identify groups of similar charging stations and EV users, to identify emerging segments, that need to be addressed through incentivisation strategies and business models. We achieved this goal by applying clustering methods and aggregation approaches, based on which we identify 4 interpretable clusters of stations.

The second goal focuses on the selection of appropriate methods and data sets to obtain a prediction of the energy consumption of the charging infrastructure. Such predictions are applicable, for example, when purchasing electricity or in smart charging technologies. Using time-series methods, machine learning and exogenous predictors, we predicted spatiotemporally aggregated consumption while achieving MAPE accuracy of approximately 12 %.

In the third goal, we are motivated by the problem of identifying the appropriate locations of the charging infrastructure. We use extensive GIS data, representing the

vicinity of charging stations. Using simple methods, that we have proposed, we extract predictors from GIS data. Influential predictors are obtained using regression variable selection methods. Using available methods, we deal with the interdependence of data and multicollinearity. To characterize suitable locations, we first predict the popularity of the charging infrastructure based on data, together with the interpretation of influential predictors. Subsequently, based on GIS data, we focus on the problem of explaining the relationship between the characteristics of the environment surrounding the infrastructure and energy consumption. In practice, the proposed methodologies and new knowledge can be used in the development of decision making support tools dedicated to the problem of deploying new or expanding the existing charging infrastructure.

Key words: data analysis, electric vehicles, charging stations, data science, machine learning, electromobility

Obsah

Zoznam obrázkov **8**

Zoznam tabuliek **10**

1 Stav v problematike **13**

1.1 Základné pojmy . . . 13

1.1.1 Pojmy týkajúce sa analýzy dát . . . 13

1.1.2 Elektromobilita . . . 13

1.2 Úvod k dátovej analýze a k nástrojom na podporu rozhodovania . . . 15

1.3 Úvod do elektromobility . . . 17

1.4 Vybrané metódy analýzy dát . . . 20

1.4.1 Všeobecný prehľad metód . . . 20

1.4.2 Metódy predspracovania dát . . . 21

1.4.3 Meranie chýb modelov . . . 22

1.4.4 Validácia modelu . . . 25

1.4.5 Regresné metódy . . . 27

1.4.6 Regresné metódy na výber premenných . . . 30

1.4.7 Lineárne klasifikačné metódy . . . 32

1.4.8 Štatistická inferencia pomocou metódy bootstrap . . . 34

1.4.9 Metódy založené na rozhodovacích stromoch . . . 35

1.4.10 Analýza časových radov . . . 36

1.4.11 Zhlukovacie metódy . . . 40

1.5 Prehľad dostupných dát . . . 42

1.5.1 EVnetNL dáta . . . 42

1.6 GIS dáta . . . 44

1.6.1 Bodové dáta . . . 45

1.6.2 Dáta lomených čiar . . . 45

1.6.3 Polygónové dáta . . . 46

1.7 Rastrové dáta . . . 47

1.8 Prehľad dostupnej literatúry v oblasti elektromobility a súvisiacich oblastiach so zameraním na analýzu dát . . . 47

1.8.1 Analýza nabíjacieho správania sa používateľov EV 48
1.8.2 Analýza dopytu EV po elektrickej energii 49
1.8.3 Analýzy dát pochádzajúcich z nabíjacích staníc 53
1.8.4 Plánovanie a umiestňovanie nabíjacej infraštruktúry pre EV 54
1.8.5 Smart charging 57

2 Ciele dizertačnej práce **59**
2.1 Metodika práce 61

3 Vlastné prínosy **62**
3.1 Analýza segmentov nabíjacích staníc a zákazníkov 62
3.1.1 Používané indikátory a spracovanie dát 62
3.1.2 Použité metódy a tvorba modelov 63
3.1.3 Výsledky zhlukovania 64
3.2 Predikovanie spotreby nabíjacích staníc 69
3.2.1 Použité dáta 70
3.2.2 Modelovanie a predikcia spotreby elektrickej energie 72
3.3 Extrakcia prediktorov z GIS dát, testovanie metód na výber premenných . 76
3.3.1 Extrakcia prediktorov z GIS dát 76
3.3.2 Predspracovanie dát pre regresné a stromové metódy 79
3.3.3 Testovanie metód na výber premenných 82
3.4 Modelovanie popularity 86
3.4.1 Modelovanie popularity 86
3.4.2 Výsledky predikcie popularity nabíjacích miest 91
3.5 Modelovanie spotreby energie 97
3.5.1 Dodatočná príprava dát pre regresiu 97
3.6 Základné charakteristiky nabíjacích miest 98
3.6.1 Nastavenie parametrov 98
3.6.2 Metriky spotrebovanej energie nabíjacích miest 98
3.6.3 Modelovanie rozdelenia spotreby energie 99
3.6.4 Vysvetľovanie spotreby energie ostatnými indikátormi nabíjacích miest 101

3.6.5 Vysvetlenie spotreby energie na nabíjacích miestach pomocou GIS dát 103
3.6.6 Vplyv stratégie umiestňovania na spotrebu energie 106

Zoznam použitej literatúry **113**

Zoznam obrázkov

1 Rozdelenie tried pre krížovú validáciu 27
2 Diagram štatistickej inferencie pomocou metódy bootstrap 34
3 Základné charakteristiky nabíjacích staníc 44
4 Funkcie hustoty zvolených indikátorov nabíjacích staníc 64
5 Rozdelenie hodín príchodov a odchodov na transakcie 65
6 Funkcie hustoty indikátorov nabíjacích staníc priradených do rovnakého zhluku 67
7 Charakteristiky dát nabíjacích staníc v regióne Utrecht 71
8 Rozmiestnenie nabíjacích staníc v regióne Utrecht 71
9 Denná spotreba elektrickej energie nabíjacou infraštruktúrou v COROP regióne Utrecht 72
10 Trénovacie procedúry pre časové rady 73
11 Predpovede spotreby energie pomocou modelov 74
12 Ilustrácia prieniku kruhovej zóny a polygónov 78
13 Závislosť počtu chýbajúcich hodnôt a celkovej populácie 80
14 Priemerná vzdialenosť medzi koeficientami 84
15 Smerodajná odchýlka vzdialenosti medzi koeficientami 84
16 Presnosť, precíznosť a senzitivita pre klasifikačné metódy 90
17 100 ROC kriviek pre klasifikačné metódy 93
18 Graf vybraných koeficientov spolu so štatistickou inferenciou 94
19 Histogramy nabíjacích bodov pre nabíjacie miesta a nabíjacej kapacity pre nabíjacie body 98
20 Empirické rozdelenie pravdepodobnosti energie spotrebovanej nabíjacími miestami 100
21 Empirické rozdelenia vybraných regresných koeficientov vysvetľujúce spotrebu energie pomocou dát z okolia nabíjacej infraštruktúry spolu s hodnotami štatistickej inferencie 104
22 Empirické rozdelenia vybraných regresných koeficientov vysvetľujúce spotrebu energie, stratifikovanú podľa umiestňovacej stratégie, pomocou dát z okolia nabíjacej infraštruktúry spolu s hodnotami štatistickej inferencie . . 107

23 Grafy rezíduí modelov pre transformácie y 134
24 Vizualizácia Cookových vzdialeností získaných pre jednotlivé pozorovania . 135
25 P-P grafy pre kombinácie rozdelení a transformácií 137
26 Q-Q grafy pre kombinácie rozdelení a transformácií 138
27 Empirické rozdelenia vybraných regresných koeficientov vysvetľujúce spotrebu energie pomocou dát z okolia nabíjacej infraštruktúry a jej prediktorov spolu s hodnotami štatistickej inferencie 139
28 Empirické rozdelenia vybraných regresných koeficientov vysvetľujúce spotrebu energie, stratifikovanú podľa rollout stratégie, pomocou dát z okolia nabíjacej infraštruktúry spolu s hodnotami štatistickej inferencie 140

Zoznam tabuliek

1 Matica zámen 24
2 Stĺpce datasetu Meterreadings a ich popis 42
3 Stĺpce datasetu Transactions a ich popis 43
4 Početnosti pozorovaní v zhlukoch 65
5 Kontigenčná tabuľka zhlukov 67
6 Trénovacie množiny pre trénovacie procedúry 73
7 Výsledky predpovedí modelov v MAPE 74
8 R^2 indikátory výkonnosti vysvetľujúcich dáta 88
9 Funkčné formy θ 91
10 Hodnoty mier chýb pre zvolené prahové hodnoty θ 92
11 Reprezentácie energie spotrebovanej pre nabíjacie miesto a ich miera vysvetliteľnosti dátami 99
12 P-hodnoty Kolmogorov-Smirnov testu zhody rozdelení 101
13 Koeficienty a štatistiky jednoduchých regresných modelov energie vysvetlenej indikátormi nabíjacích miest 101
14 Zoznam atribútov datasetu Populačné jadrá. 126
15 Zoznam atribútov datasetu Susedstvá 128
16 Kategórie polygónov datasetu Využitie územia. 129
17 Atribúty datasetu Atlas energie 130
18 Zoznam atribútov z datasetu Životná úroveň 130
19 Zoznam atribútov asociovaný s datasetov Dopravné toky. 130
20 Zoznam kategórií zostavený z OSM bodov záujmu. 131
21 Tabuľka zástupcov korelovaných atribútov 132

Úvod

Žijeme v dobe, kde dáta tvoria jeden z najcennejších zdrojov, častokrát označované aj ako ropa 21. storočia [37]. Dokážeme ich zbierať takmer všade a spracovávať vo veľkých množstvách. Spracovaniu a analýze dát sa postupne venuje čoraz viac ľudí a celý svet sa, či už priamo alebo nepriamo, v každodennom živote stretáva s aplikáciami umelej inteligencie a strojového učenia, napríklad pri nákupoch na internete alebo fotografovaní svojím smartfónom, čím inteligentne využívajú pozbierané dáta.

O elektromobilite sa v súčasnosti hovorí ako o riešení, ktoré môže priniesť veľký potenciál vo viacerých smeroch, akými sú, napríklad, nižšie emisie, vyššia efektivita elektrických motorov oproti spaľovacím alebo nezávislosť od fosílnych palív. V tejto práci elektromobilitu vnímame aj ako moderný príklad prepojenia energetických a dopravných systémov. Niektoré štáty poskytujú pre elektrické vozidlá dotácie, a dokonca, niekde dochádza už k obmedzovaniu spaľovacích motorov, vďaka čomu sa podiel elektrických vozidiel (EV) neustále zvyšuje.

Proces nasadzovania elektromobility je spojený s rôznymi výzvami a komplikáciami. Napríklad, cena elektrických vozidiel je výrazne vyššia ako cena bežných vozidiel. Takisto z dôvodu ich nabíjania je potrebné najskôr vybudovať samostatnú infraštruktúru a aby elektromobilita výraznejšie zlepšovala životné prostredie, mala by elektrická energia pochádzať z obnoviteľných zdrojov. Ak bude rozhodovanie v elektromobilite vhodne podporené dátami, ktoré začínajú byť dostupné, môžeme predpokladať, že budú ušetrené nezanedbateľné finančné prostriedky. Celý proces pri budovaní dátami poháňaných rozhodnutí v prostredí elektromobility sa dá rozdeliť do viacerých fáz. Prvou fázou je vytvorenie plánu a stanovenie cieľov. Druhou fázou je zber, spracovanie, analýza a vyhodnotenie týchto dát. Posledná fáza je vytvorenie rozhodnutí na základe získaných výsledkov, alebo vykonanie, napríklad, ďalších simulačných či optimalizačných štúdií a následné vyvodenie záverov a rozhodnutí. V súlade s týmito trendmi si táto práca kladie za cieľ navrhnúť postup, akým bude možné podporiť rozhodovanie v tejto oblasti a to konkrétne využitím prostriedkov analýzy dát.

V prvej kapitole práce predstavíme súčasný stav problematiky a metódy, ktoré využijeme pri riešení stanovených cieľov. V druhej kapitole predstavíme stanovené ciele a tretia kapitola sa bude venovať vlastným prínosom a popisu navrhovaného riešenia. V

podkapitolách tretej kapitoly začneme modelovaním zákazníckych segmentov, prejdeme k predpovedaniu spotreby elektrickej energie nabíjacích staníc a budeme sa venovať zberu a spracovaniu GIS dát a identifikácii metód na výber premenných. Posledné dve podkapitoly sú venované predikovaniu popularity a vysvetľovaniu spotrebovanej energie, obe na základe GIS dát. V časti záverečné zhrnutie výsledkov uvedieme vedecké prínosy a prínosy pre prax.

1 Stav v problematike

Táto kapitola je venovaná súčasnému stavu v problematike dátových analýz a ich využitiu pri podpore rozhodovania v elektromobilite.

1.1 Základné pojmy

1.1.1 Pojmy týkajúce sa analýzy dát

Analýza dát je multidisciplinárna oblasť spájajúca niekoľko oblastí ako napr. matematika, informatika a kontext skúmanej oblasti, u nás EV. Pre prácu s dátami zadefinujeme a popíšeme niekoľko fundamentálnych pojmov, ktoré budeme používať v popise použitých postupov dátovej analýzy. *Vstupná premenná* (angl. input variable) X alebo aj vysvetľujúca premenná je náhodná premenná, často viacrozmerná. *Výstupná premenná* Y (angl. output variable) alebo aj vysvetľovaná premenná je náhodná premenná, ktorú sa snažíme vysvetliť v závislosti od vstupných premenných. Vzťah medzi výstupnou a vstupnou premennou popisujeme funkciou $Y = f(X) + \epsilon$, kde f je funkcia vstupnej premennej a ϵ je náhodná chyba [55]. Funkciu s odhadnutým tvarom alebo parametrami môžeme považovať za *model*. Za jednu z najdôležitejších úloh dátového modelovania môžeme považovať nájdenie vhodnej reprezentácie funkcie f. *Dátová matica*, ako realizácia vstupnej premennej $\boldsymbol{X}$ je tvorená $i = 1, \ldots, n$ *pozorovaniami* $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ (riadky matice $\boldsymbol{X}$) a *prediktormi* $x_j = \{x_{1j}, \ldots, x_{nj}\}$ pre $j = 1, \ldots, p$ (stĺpce matice $\boldsymbol{X}$). Vektor realizácií výstupnej premennej označíme y a budeme ho nazývať vektor výstupu. Dáta predpokladáme v určitej štandardnej forme, ktorú nazývame *dataset*. Dataset môže nadobúdať podobu databázovej tabuľky alebo dátovej matice [130]. Používať budeme aj *časové rady*, zjednodušene definované ako postupnosť dát, pozorované v čase. Časový rad môžeme považovať za realizáciu postupností náhodných premenných Y_t, kde ich realizácie označíme y_t v čase $t = 1, 2, \ldots, N$, kde N je dĺžka časového radu [72].

1.1.2 Elektromobilita

Terminológia pre elektromobilitu je na Slovensku stanovená v Zákone č. 251/2012 Z. z. Zákon o energetike a o zmene a doplnení niektorých zákonov, §2b [78], my uvádzame pre prípadné lepšie pochopenie základných termínov aj ich anglické názvy, ktoré môžu po-

môcť pochopiť zahraničnú literatúru. Definície sme dopĺňali aj z dokumentu holandských definícií pojmov elektromobility [77].

Pod pojmom *elektrické vozidlo* (angl. electric vehicle) (EV) rozumieme "motorové vozidlo, vybavené hnacou jednotkou, ktorá sa skladá minimálne z jedného neperifémeho elektrického motora ako meniča energie s nabíjateľným systémom ukladania elektriny, ktorý možno externe nabíjať"(zákon č. 251/2012 Z. z. Zákon o energetike a o zmene a doplnení niektorých zákonov, §2b, ods. 29 [78]). Pod *nabíjacím bodom* (angl. charging point) rozumieme "rozhranie, ktoré v určitom čase umožňuje nabíjanie jedného elektrického vozidla elektrinou alebo výmenu batérie"(zákon č. 251/2012 Z. z. Zákon o energetike a o zmene a doplnení niektorých zákonov, §2b, ods. 30 [78]). *Konektor* (angl. connector) je nabíjacia zásuvka, na ktorú sa môže napojiť práve jedno vozidlo. Nabíjací bod môže mať jeden alebo viac konektorov, pričom iba jeden z nich môže byť použitý súčasne. *Nabíjacia stanica* je "jeden alebo viac nabíjacích bodov, ktoré sú integrované do jedného zariadenia"(zákon č. 251/2012 Z. z. Zákon o energetike a o zmene a doplnení niektorých zákonov, §2b, ods. 31 [78]). *Verejne prístupnou nabíjacou stanicou* (angl. public charging station) je "nabíjacia stanica, ku ktorej je zabezpečený nediskriminačný prístup všetkým používateľom; nediskriminačný prístup môže zahŕňať rôzne spôsoby autentifikácie a platby"(zákon č. 251/2012 Z. z. Zákon o energetike a o zmene a doplnení niektorých zákonov, §2b, ods. 32 [78]). *Nabíjacím miestom* (angl. charging pool) budeme podľa [77] rozumieť miesto, na ktorom sa nachádza jedna alebo viac nabíjacích staníc, zvyčajne umiestnených vedľa seba. *Nabíjacou špičkou* (angl. charging peak) rozumieme čas nabíjania, kedy sú nabíjaním elektrických vozidiel dosiahnuté vrcholové hodnoty v spotrebe elektrickej energie. *Nabíjacou transakciou* (angl. charging transaction), zjednodušene len transakciou, budeme rozumieť nabíjaciu udalosť, začínajúcu pripojením a končiacu odpojením vozidla. Nabíjacie stanice s výkonom do 22 kW budeme označovať ako *pomalé nabíjacie stanice*.

Pojem *smart charging* je v [119] definovaný ako optimalizácia času, rýchlosti a smeru nabíjacej udalosti. Pod smerom nabíjania môžeme rozumieť, či je EV nabíjané, alebo je EV vybíjané a prúd ide do elektrickej siete, čo sa nazýva aj vehicle-to-grid. Často sa smart charging využíva na zvýšenie spotreby energie vygenerovanej z obnoviteľných zdrojov a takisto aj na znižovanie energetických špičiek. Čas od pripojenia EV na konektor až po jeho odpojenie budeme nazývať *časom pripojenia*. Čas, kedy sa počas pripojenia EV nabíja, budeme označovať ako *čas nabíjania*. Čas, kedy je EV pripojené, ale nenabíja

sa, budeme nazývať *nečinným časom* pripojenia alebo časom nečinnosti. V práci budeme rozlišovať najmä tieto dva typy EV: batériové elektrické vozidlá (BEV), ktoré sú poháňané iba energiou z batérie, a hybridné elektrické vozidlá s možnosťou dobíjania (PHEV), ktoré majú spaľovací motor a elektrický motor, ktorý je poháňaný batériou, dobíjateľnou z elektrickej siete. Osobu, ktorá jazdí na EV, či už firemnom alebo vlastnom, budeme jednotne označovať ako používateľa EV.

1.2 Úvod k dátovej analýze a k nástrojom na podporu rozhodovania

Žijeme v dobe, v ktorej dáta tvoria pevný základ rozhodovacieho procesu. V začiatkoch tvorby tejto práce vyšiel článok hovoriaci, že najcennejším zdrojom 21. storočia nie je ropa, ale dáta [37]. Poukazuje na fakt, že dáta zbierame takmer všade, kde je to možné, aj za cenu nízkej kvality a veľkého objemu za cieľom získania nových informácií. Algoritmy umelej inteligencie, ako napr. strojového učenia, dokážu získať hodnotné informácie z takýchto dát. Napríklad, čím viacej dát získava Tesla z autonómnych vozidiel, tým viac dokáže vylepšiť ich jazdné vlastnosti. Technologický giganti ako Amazon, Facebook a Google, ktorí ovládajú kyberpriestor nielen v USA, majú svoje podnikanie postavené na dátach. Toto je len málo z mnohých ukážok toho, že dáta sú nevyhnutnou súčasťou rozhodovania v dnešnej dobe. Avšak ako ropa, tak aj dáta spôsobujú nežiadúce vedľajšie efekty, najmä v podobe únikov osobných informácií, na čo nesmieme zabúdať. Dátová analýza sa venuje celému procesu od definovania problému, cez spracovanie dát až po užitočné finálne informácie získané z dát.

Proces analýzy dát môžeme podľa [99] rozdeliť do niekoľkých fáz:

1. formulácia problému;
2. zber dát;
3. predspracovanie dát;
4. exploratívna analýza;
5. modelovanie a prediktívna analýza;
6. prezentácia výsledkov.

Prvé dve fázy nebudeme bližšie popisovať, keďže sú pomerne všeobecne definované a prínosy tejto práce sa budú týkať hlavne zvyšných bodov.

Predspracovanie dát

Pozbierané dáta sú často nekompletné, obsahujú chybné údaje alebo nie sú vo vhodnom formáte. Kompletizácia a oprava chybných hodnôt dát sa nazýva aj čistením dát. Predspracovanie je časovo najnáročnejšia časť práce s dátami, a podľa [89], zaberá približne 60 % času práce s dátami. Prvým krokom tejto fázy je transformácia dát do vhodného formátu, napr. logy z nabíjacích staníc rôznych poskytovateľov zmeníme na jednotný formát pre všetky stanice a poskytovateľov. V takto transformovaných dátach môžeme jednotne nájsť rôzne chýbajúce alebo chybné hodnoty. Napríklad, zápornú dĺžku trvania nabíjania elektromobilu je potrebné nahradiť alebo aproximovať vhodnými hodnotami alebo odstrániť celé pozorovanie [49]. Kontrolovať môžeme, či sú hodnoty z vhodnej domény (numerické, dátumy, textové, a pod.). Takto predspracované hodnoty sú pripravené pre ďalšiu analýzu a spracovanie.

Exploratívna analýza

Keď máme dáta predspracované v použiteľnom formáte, môžeme prejsť na exploratívnu analýzu, v ktorej dáta preskúmavame. Použiť môžeme rôzne štatistiky, transformácie a vizualizácie, ktoré nám pomáhajú nájsť skryté vzťahy v dátach, lepšie pochopiť dáta a testovať rôzne hypotézy o dátach [14]. Dôležitú úlohu pri exploratívnej analýze zohráva nielen ľudský faktor, ale aj správny výber vhodných nástrojov, ktoré napríklad dokážu vybrať vhodné farebné palety do grafov, napr. ColorBrewer [28] alebo zvoliť vhodný typ grafu. Komplexnými nástrojmi na vizuálnu, ale aj pokročilejšiu analýzu sú, napríklad Tableau, Plotly [109, 88] alebo publikácie, ako napr. [129].

Modelovanie dát a prediktívna analýza

Po transformovaní dát do potrebných formátov a ich prvotnom prieskume, zvyčajne nasleduje modelovanie dát pomocou matematicko-štatistických modelov. Tieto modely môžu byť súčasťou dnes často spomínanej oblasti – strojového učenia. Strojové učenie je subdoménou umelej inteligencie, ktorá využíva učenie, ako napríklad zlepšovanie modelov na základe vlastných skúseností bez explicitného naprogramovania [75]. Vďaka modelovaniu dokážeme opísať zložitejšie vzťahy dát a objaviť nové štruktúry v dátach. Prediktívna analýza umožňuje využiť modely na predikovanie budúcich hodnôt, ale napríklad aj au-

tomatické triedenie nových pozorovaní.

Prezentácia výsledkov alebo nasadenie v praxi

Po analýzach a modelovaní, potrebujeme výsledky vhodne odprezentovať, spravidla vizualizáciou. Existuje množstvo spôsobov prezentovania výsledkov. Niekedy postačujú tabuľky obsahujúce stručný opis a číselné výsledky, inokedy sú to krátke správy obsahujúce niekoľko vizualizácií a v neposlednom rade aj interaktívne aplikácie umožňujúce komplexné vizualizácie. Vhodnou kombináciou opisu výsledkov a zrozumiteľných vizualizácií dokážeme z dát vytvoriť príbeh pochopiteľný pre ľudí, ktorým výsledky odprezentujeme a poskytneme im tak užitočné výstupy v jednoduchej a zrozumiteľnej forme. V prípadoch, keď je cieľom analýzy vytvorenie softvéru, napr. na predikovanie energie, záverečnou fázou môže byť jeho nasadenie v praxi.

1.3 Úvod do elektromobility

Parížska dohoda z dôvodov globálneho otepľovania nastavila ako cieľ obmedziť nárast globálnej priemernej teploty do hodnoty 2°C v porovnaní s predindustriálnou hodnotou. Veľká miera nárastu teploty sa pripisuje skleníkovým plynom. Odhaduje sa, že doprava je v európskych mestách zodpovedná za 30 % skleníkových emisií. Očakáva sa, že elektrifikácia dopravy môže prispieť k dekarbonizácií energetických systémov a rast elektrifikácie dopravy napreduje ruka v ruke s dekarbonizáciou energetického sektora. S potenciálne nulovými emisiami, sú elektrické vozidlá alternatívou pre vozidlá so spaľovacími motormi, pričom znižujú aj zvukové emisie [15].

Elektrifikácia cestnej dopravy prináša mnoho výhod, medzi ktoré podľa [17] patria:

- Energetická efektívnosť – EV premieňajú viac než 77 % elektrickej energie z napájacej elektrickej siete na pohon. Bežné vozidlá so spaľovacími motormi premieňajú iba 12 - 30 % energie z paliva na pohon [116].

- Energetická bezpečnosť – elektromobilita vo svete motorizmu znižuje závislosť importu ropy mnohých krajín a zároveň zvyšuje diverzitu foriem energie používanej na nabíjanie EV, ktoré môžu byť produkované z rôznych, najlepšie obnoviteľných zdrojov.

- Znečistenie ovzdušia – vďaka potenciálnym nulovým emisiám, EV znižujú znečis-

tenie ovzdušia najmä v urbánnych územiach a pozdĺž cestných sietí, kde je značný počet ľudí vystavený škodlivým emisiám z cestnej dopravy.

- Skleníkové plyny – elektromobilita v kombinácií so stúpajúcim nízko-uhlíkovým generovaním elektriny dokáže priniesť značnú redukciu skleníkových plynov. EV takisto poskytujú flexibilitu pre energetické sústavy.

- Zníženie hluku - EV sú tichšie ako vozidlá so spaľovacími motormi a tak prispievajú k nižšiemu hlukovému znečisteniu.

- Industriálny rozvoj - EV majú dôležitý význam v redukcii cien batériových technológií, keďže významne zvyšujú objem výroby a prispievajú k industriálnej súťaživosti v tejto oblasti.

EV v súčasnosti vyžadujú značnú podporu pre zvýšenie ich akceptácie a používania ľudmi. Častou prekážkou je strach používateľov EV z obmedzeného dojazdu, nízkej dostupnosti nabíjacej infraštruktúry či vysoké ceny EV, spôsobené najmä cenou batérie. Preto sa vlády a samosprávy snažia podporovať elektromobilitu rôznymi stimulmi, ktoré sa dajú rozdeliť do piatich kategórii [113]:

- regulačné stimuly – CO_2 štandardy, ciele v počte predaných EV;

- priame stimuly pre spotrebiteľov – dotácie, registračné a daňové úľavy na EV;

- nepriame stimuly pre spotrebiteľov – prístup k nízko emisným zónam, špeciálnym jazdným pruhom, zvýhodnené parkovanie;

- nabíjacia infraštruktúra – podpora pre nabíjacie stanice, podpora domácich súkromných nabíjacích staníc;

- komplementárne politiky – vzdelávanie spotrebiteľov, podpora výskumu a vývoja.

Politiky podpory EV sa v krajinách líšia a spôsobujú rozdielny pomer nákupov BEV a PHEV. Napr., v Holandsku boli odpustené registračné dane pre vozidlá s nízkymi emisiami, čo spôsobilo vyššie nákupy PHEV. Od roku roku 2016 však bola ponechaná nulová daň iba pre BEV, čo malo za následok omnoho vyšší nákup BEV v pomere k PHEV ako doposiaľ [113, 15].

Holandsko je krajina s jednou z najhustejších a najlepšie rozvinutých nabíjacích infraštruktúr. Približne 90 % populácie v Holandsku žije v urbánnych oblastiach a iba 42 % domácností má prístup k súkromnému parkovaniu, čo sa prejavuje aj vo vyššom využívaní verejných nabíjacích staníc v tejto krajine. Takisto k tomu prispieva aj fakt, že vysoký podiel EV bol registrovaný v Holandsku firmami [113].

Rozvoj nabíjacej štruktúry dokáže značne pomôcť používateľom prekonať aj strach z obmedzeného dojazdu, no vzniká tu problém sliepky a vajca, vo forme nabíjacej infraštruktúry a adopcie EV, ktorý bol identifikovaný ako dôležitá výzva, brániaca rastu EV ekosystému [118]. Vodiči sú zdržanliví voči kúpe EV, pokiaľ nie je vybudovaná dostatočná infraštruktúra a podobne aj operátori nabíjacích staníc neinvestujú do infraštruktúry, pokiaľ je počet EV nízky a neprináša im zisk. Momentálne je stále nedostatočná infraštruktúra signifikantný faktor, zabraňujúci väčšej penetrácií EV [25]. Rozvoj novej verejnej nabíjacej infraštruktúry zahŕňa odhadnutie vzorov jej návštevnosti na zaručenie najlepšieho využitia alokovaných zdrojov. A teda jedna cesta, ako podporiť rozhodovanie v tejto oblasti, sú dátovo riadené prístupy aj so schopnosťou predikcie. Takisto dokáže vhodne a dostatočne husto umiestnená nabíjacia infraštruktúra znížiť strach používateľov z dojazdu.

Na základe analýz správania používateľov EV v Nórsku sa zistilo, že nabíjanie EV nie je v súlade s tankovaním spaľovacích áut, aj keď rýchle nabíjanie je podobné tankovaniu. Používatelia elektrických vozidiel si najčastejšie nabíjajú vozidlá doma alebo v práci pomocou pomalých nabíjacích staníc. Tretia najčastejšia voľba pri nabíjaní sú verejne dostupné pomalé nabíjacie stanice, spolu s nabíjaním na komerčných miestach (cieľových miestach jázd vozidlom). Rýchle nabíjanie nie je využívané až tak často a primárne slúži na plánované zastávky pre jazdy na dlhé vzdialenosti [16]. Avšak, podľa [17], bude v roku 2030 95 % verejných nabíjacích staníc pre konvenčné EV poskytovať pomalé nabíjanie, oproti aktuálnym približne 70 %.

Nabíjanie elektrických vozidiel môže byť aj prostriedok pre zvýšenie flexibility elektrickej siete (hlavne pomalé nabíjanie), ako aj veľkým problémom pre elektrickú sieť (simultánne rýchle nabíjanie), ak nie je riadené správne. Pre garanciu možnej flexibility je potrebné inteligentné prepojenie medzi EV, nabíjacou infraštruktúrou a operátorom elektrickej siete, t.j. smart charging.

1.4 Vybrané metódy analýzy dát

Ďalšou etapou riešenia problému je modelovanie a prediktívna analýza. V tejto fáze využívame najmä metódy strojového učenia. Podľa [49], strojové učenie skúma, ako sa počítačové algoritmy dokážu učiť na základe dát. Hlavným cieľom strojového učenia je naučiť počítačové programy automaticky rozoznávať komplexné vzory a vytvárať inteligentné rozhodnutia na základe dát.

1.4.1 Všeobecný prehľad metód

Strojové učenie možno rozdeliť podľa spôsobu využitia dát na dve základné kategórie: učenie s učiteľom a učenie bez učiteľa.

Učenie s učiteľom

Pri učení s učiteľom súčasne využívame vopred definované prediktory a aj vektor výstupu. Cieľom je modelovať vzťah medzi vektorom výstupu a prediktormi čo najpresnejšie tak, aby tento vzťah mohol byť použitý na predpovedanie budúcich pozorovaní (predikcia) alebo lepšie porozumenie súvislostí medzi vektorom vstupu a prediktormi (inferencia) [55]. Typy úloh strojového učenia môžeme deliť podľa typu výstupnej premennej na úlohy klasifikačné, ktorých výstupné premenné sú diskrétne a úlohy regresné, ktorých výstupné premenné sú spojité.

Príkladom klasifikačnej úlohy môže byť predikcia, či bude alebo nebude nejaká nabíjacia stanica často navštevovaná na základe charakteristík okolia, kde ju chceme vybudovať. Príkladom regresnej úlohy je odhad budúceho vývoja týždennej spotreby elektrickej energie na základe predošlej spotreby, vonkajšej teploty a počtu nabíjacích staníc.

Učenie bez učiteľa

Pri učení bez učiteľa nemáme pre prediktory vopred definovaný vektor výstupu, takže nás nemá kto „učiť“. Snažíme sa preto hľadať vzťahy medzi prediktormi alebo medzi pozorovaniami. Príkladom takéhoto učenia je zhluková analýza, pri ktorej sa snažíme priradiť pozorovania do navzájom odlišných skupín, kde v každej skupine sú zaradené podobné objekty a vytvoriť tak zhluky pozorovaní [55]. Príkladom zhlukovania je vyhľadávanie podobných nabíjacích staníc na základe ich charakteristík.

1.4.2 Metódy predspracovania dát

Tieto metódy budú slúžiť na detekciu a vysporiadanie sa s chybnými hodnotami, ale aj na transformáciu hodnôt.

Škálovanie dát

Jednotky, v ktorých sú merané dáta, môžu ovplyvniť výsledky. Vo všeobecnosti, ak vyjadrujeme prediktor v menších jednotkách, jeho hodnoty budú vyššie a tak aj jeho rozsah bude väčší a preto môže mať taký prediktor väčší efekt na konečný výsledok v prípade, ak je metóda citlivá na škálu hodnôt. Aby sme sa vyhli závislosti na jednotkách použitých v dátach, aplikujeme normalizáciu alebo štandardizáciu, ktoré dáta preškálujú [49]. Napríklad, zmena jednotiek energie z kWh na menšie jednotky Wh pri jednom z prediktorov, dokáže do značnej miery pri niektorých metódach ovplyvniť výsledky analýzy.

Štandardizácia dát je operácia, pomocou ktorej škálujeme dáta tak, aby mali nulový priemer a boli v jednotkách smerodajnej odchýlky daného prediktora

$$x_j^s = \frac{x_j - \overline{x_j}}{std(x_j)}, \tag{1}$$

kde $\overline{x_j}$ je aritmetický priemer a $std(x_j)$ smerodajná odchýlka hodnôt prediktora x_j. Operácie sú aplikované na prediktor jednotlivo na všetky jeho pozorovania.

Pri *normalizácií* škálujeme prediktory na rozsah $\langle 0; 1\rangle$, pomocou vzťahu

$$x_j^n = \frac{x_j - min(x_j)}{max(x_j) - min(x_j)}, \tag{2}$$

kde $min(x_j)$ a $max(x_j)$ sú minimálna a maximálna hodnota prediktora. Takouto transformáciou môžeme, napríklad, pripraviť dáta pre zhlukovacie metódy, požadujúce rovnakú škálu dát.

Box-Cox transformácia je užitočnou logaritmicko-exponenciálnou transformáciou v tvare

$$w = \begin{cases} \ln(x_j) & \text{ak } \lambda = 0; \\ (x_j^\lambda - 1)/\lambda & \text{inak.} \end{cases} \tag{3}$$

Parameter λ určuje tvar transformácie hodnoty x_j. Využitie tejto transformácie je, napríklad, na stabilizáciu rozptylu časového radu [54], kde by mal časový rad mať po takejto

transformácií takmer konštantný rozptyl nemenný v čase.

Škálovanie dát môžeme samozrejme aplikovať aj na vektor výstupu y rovnakým spôsobom ako na prediktory.

Detekcia hodnôt mimo rozsah

Detekcia hodnôt mimo rozsah alebo anomálií je proces hľadania hodnôt pozorovaní, značne mimo očakávaný rozsah hodnôt. Takéto hodnoty sú problémové, pretože môžu znižovať kvalitu predikcií ale aj základných štatistík, akou je napríklad priemer. Na detekciu takýchto hodnôt sa používajú rôzne techniky, ako napríklad, medzikvartilové rozpätie, mediánové filtre [85] alebo aj DBSCAN zhlukovací algoritmus [49]. Ak takéto hodnoty objavíme, môžeme ich, napríklad, nahradiť mediánom alebo ich úplne odstrániť spolu s pozorovaniami alebo prediktormi, ku ktorým patria.

Chýbajúce hodnoty

Na dopĺňanie chýbajúcich hodnôt existuje niekoľko spôsobov, ako napríklad nahrádzanie chýbajúcich hodnôt mediánom, najpravdepodobnejšou hodnotou a pod. [49, s. 88]. Komplexný opis metód na nahrádzanie chýbajúcich hodnôt možno nájsť aj v [117].

1.4.3 Meranie chýb modelov

Fundamentálnym princípom takmer každej metódy učenia s učiteľom je meranie chyby na základe rozdielu pozorovaného výstupu (i-tej hodnoty prvku) y_i a jeho odhadu $\hat{y}_i$, ktorý sa nazýva rezíduum a označuje sa e_i.

Pre meranie chýb, ktoré úzko súvisia s porovnávaním modelov, si definujeme *nulový model*, ktorý s istou pravdepodobnosťou náhodne predpovedá dané výstupy.

Jedna zo základných mier chýb modelu, ktorá je aj minimalizovaná pri metóde najmenších štvorcov, je suma štvorcov rezíduí (angl. residual sum of squares - RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2, \tag{4}$$

kde e_i je i-ty prvok rezídua a n je počet pozorovaní.

Ďalšou často využívanou mierou chýb modelu je priemerná štvorcová chyba (angl. mean squarred error - MSE)

$$MSE = \frac{RSS}{n}. \tag{5}$$

Ak nechceme chybu vyjadriť v absolútnych hodnotách, ale v relatívnych, tak môžeme využiť priemernú absolútnu percentuálnu chybu (angl. mean absolute percentage error - MAPE)

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} |\frac{e_i}{y_i}|. \tag{6}$$

Nevýhodou MAPE je, že nadobúda extrémne hodnoty, ak sú hodnoty y_i blízke nule. Preto sa v takýchto prípadoch môžu použiť škálované miery chýb, kde napr. škálujeme priemernú chybu škálovacou konštantou [54].

Miery chýb modelov s hlavným účelom na porovnanie modelov

Akaikovo informačné kritérium (Acaice information criteria - AIC) vyhodnocuje chybovosť modelu, ale slúži primárne na porovnávanie modelov zohľadňujúc veľkosť modelu, najmä pre modely s odlišným počtom prediktorov. AIC je vyjadrené ako

$$AIC = n \log\left(\frac{RSS}{n}\right) + 2(p+2), \tag{7}$$

kde n je počet pozorovaní a p je počet prediktorov.

Schwartzovo Bayesovské informačné kritérium (Bayesian information criteria - BIC) je miera veľmi podobná AIC a vieme ho vyjadriť ako

$$BIC = n \log\left(\frac{RSS}{n}\right) + (p+2)\log(n). \tag{8}$$

Pri výbere modelov sa snažíme nájsť model s čo najnižšou hodnotou AIC alebo BIC [55, s. 78].

Meranie chýb modelov s binárnym výstupom

V prípade, ak vektor predpovedí $\hat{y}$ nadobúda binárne hodnoty, môžeme na meranie chýb využiť maticu zámen, ktorá nám ukáže aj druhy chýb, ktorých sa dopúšťame. Hodnotu 1 definujeme ako pozitívum a 0 ako negatívum. Keďže využívané klasifikačné metódy odhadujú pravdepodobnostné hodnoty, stanovíme si binárnu hodnotu $\theta \in \langle 0; 1\rangle$, kde ak $\hat{y} \geq \theta$, tak predikcia bude 1, inak 0. Pre modely binárnej klasifikácie budeme rozumieť pod nulovým modelom model, ktorý predikuje výstup s pravdepodobnosťou rovnou frekvencii hodnôt vektora výstupu v trénovacích dátach.

V tabuľke č. 1 TP (angl. true positive) je počet správne odhadnutých pozitívnych

		Predikované	
		1	0
Skutočné	1	TP	FP
	0	TN	FN

Tabuľka 1: Matica zámen pre binárne predikcie.

hodnôt, FP (angl. false positive) je počet nesprávne odhadnutých pozitívnych hodnôt, FN (angl. false negative) je počet nesprávne odhadnutých negatívnych hodnôt a TN (angl. true negative) je počet správne odhadnutých negatívnych hodnôt. Z prvkov tejto matice môžeme následne vytvárať ďalšie miery chybovosti.

Ako prvú si predstavíme presnosť, ako pomer správnych odhadov

$$\frac{TP + TN}{TP + TN + FP + FN}. \tag{9}$$

Druhou je *precíznosť* ako podiel správnych odhadov pozitív

$$\frac{TP}{TP + FP}. \tag{10}$$

Ďalšou mierou chýb modelu je *senzitivita* ako podiel správne odhadnutých pozitív

$$\frac{TP}{TP + FN}. \tag{11}$$

Fall-out je podiel nesprávne odhadnutých negatív

$$\frac{FP}{FP + TN}. \tag{12}$$

Často sa na porovnávanie modelov používa "receiver operating characteristics - ROC" krivka, ktorá má na x-ovej osi fall-out a na y-ovej osi senzitivitu, ktorých hodnoty pre celú škálu prahu θ sa nanášajú na krivku. ROC krivka zobrazuje kompromis medzi podielom presných predpovedí pozitív oproti podielu chybných predpovedí negatív pre meniaci sa parameter θ [49, s. 374].

Plocha pod touto krivkou sa nazýva "area under the ROC curve - AUC". Čím bližšie je ROC krivka k ľavému hornému rohu, tým vyššia je hodnota AUC a tým lepšie model klasifikuje hodnoty. Aj vďaka týmto vlastnostiam je ROC krivka vhodná na porovnávanie klasifikačných modelov [55, s. 147].

V niektorých prípadoch však vyššie uvedené miery nemusia postačovať. Napríklad, ak máme viac pozitív ako negatív vo vektore výstupu. Ak by takýto vektor tvorilo, napríklad, 25 % 1 a zvyšok 0, bola by presnosť pre nulový model, predikujúci hodnoty náhodne s rovnakým rozdelením pravdepodobnosti, rovná 0.75. Preto definujeme aj ďalšie miery chýb klasifikačných modelov, ktorých výsledky nie sú v takej miere ovplyvňované nerovnomernosťou zastúpenia tried vo vektore výstupu.

Miera zahrňujúca precíznosť a senzitivitu je F–skóre, ktoré je mierou počítajúcou harmonický priemer precíznosti a senzitivity, pričom jeho hodnota bližšie k nižšej z týchto dvoch hodnôt [92]

$$F\text{–skóre} = 2\frac{\text{senzitivita} \cdot \text{precíznosť}}{\text{senzitivita} + \text{precíznosť}}. \tag{13}$$

Komplexnejšou mierou je Matthewov korelačný koeficient (angl. Matthews correlation coefficient - MCC) [73]

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{14}$$

Ak je menovateľ rovný 0, MCC je rovný 0 tiež a tak predídeme deleniu nulou. Pre modely predikujúce s väčšou (menšou) úspešnosťou ako nulový model, MCC je pozitívny (negatívny). Modely predikujúce na úrovni nulového modelu majú MCC rovný 0. MCC je rovný 1 (-1), ak sú všetky hodnoty predikované správne (nesprávne).

1.4.4 Validácia modelu

Pri trénovaní modelov strojového učenia potrebujeme overiť ich natrénovanie, ako aj vybrať najlepší z modelov. To môžeme dosiahnuť kombináciou metód deľby pozorovaní a merania chýb. Základným delením je rozdelenie pozorovaní na trénovaciu a testovaciu množinu. Model sa učí na trénovacej množine dát a jeho presnosť sa meria na testovacej množine dát. Tento prístup je pri niektorých metódach závislý aj na rozdelení prvkov medzi tieto množiny, kde pri odlišnom výbere pozorovaní do týchto množín môžeme dostať značne odlišné výsledky.

Kompromis medzi vychýlením a rozptylom

Pri trénovaní modelov sa často stretávame s vychýlením a rozptylom modelu, ktoré sa snažíme minimalizovať. Pod rozptylom si môžeme predstaviť mieru zmeny výstupov modelu, ak by sme ho trénovali inou vzorkou trénovacích dát. Vychýleniu môžeme rozumieť

ako chybe, ktorá vzniká, ak použijeme príliš jednoduchý model na modelovanie komplikovanejšieho vzťahu. Všeobecne majú flexibilnejšie metódy nižšie vychýlenie. Je pravidlom, že čím flexibilnejšie metódy použijeme, tým sa viac zníži vychýlenie, na úkor zvýšenia rozptylu. Ak príliš zvyšujeme flexibilitu metódy, začne od určitého bodu prudšie narastať chyba na testovacích dátach a dôjde k pretrénovaniu modelu [55, s. 33].

Krížová validácia

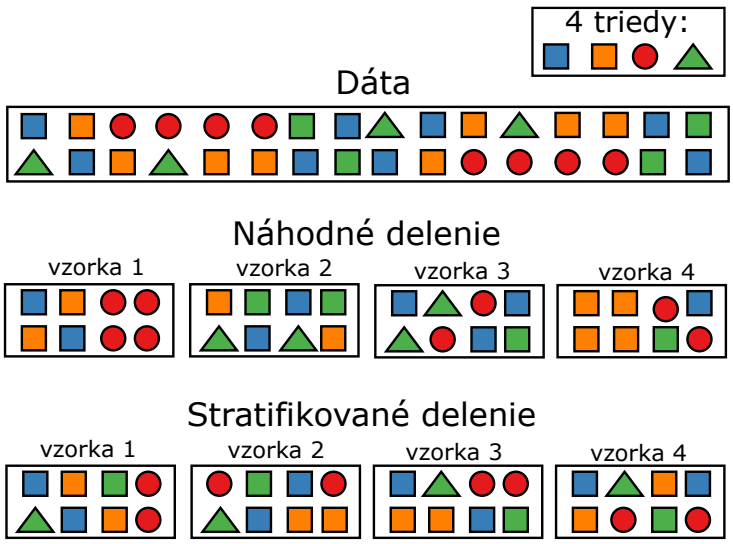
Táto metóda náhodne rozdelí pozorovania do k skupín približne rovnakej dĺžky, pričom každú množinu použije práve raz ako testovaciu a zvyšné ako trénovacie. Modely natrénované takýmto spôsobom, majú tendenciu produkovať menej variabilné výsledky oproti postupu, kedy by sme ich trénovali iba na základe rozdelenia pozorovaní na trénovaciu a testovaciu množinu. Čo sa týka vychýlenia a rozptylu platí, že modely trénované krížovou validáciou s nižším k majú vyššie vychýlenie a nižší rozptyl, ako modely, kde bolo k pri tréningu oveľa vyššie[55, s. 183].

Stratifikované náhodné delenie dát

Keď máme nerovnomerné rozdelenie tried vo vektore výstupu, môže pri validáciách nastať situácia, že sa v trénovacej množine neobjaví niektorá z tried vektora výstupu alebo početnosť tried vzorky vektora výstupu bude značne vychýlená oproti početnosti tried vektora. Toto môže spôsobiť, že sa daná trieda vyskytne so značne nižšou pravdepodobnosťou v predpovediach alebo sa v extrémnych prípadoch nevyskytne vôbec. Jedným z riešení je stratifikované delenie, kde sa udržiava približne taký pomer tried vektora predpovedí vo vzorke, aký je v originálnom vektore. Na obrázku 1 je ilustrované náhodné delenie a stratifikované delenie. Pri štandardnom náhodnom delení môžeme vidieť, že v niektorých vzorkách dochádza k absencií niektorých tried dát, čo komplikuje trénovanie modelu, zatiaľ čo stratifikované delenie si udržiava početnosť tried. Stratifikovaná krížová validácia priraďuje validačné skupiny v rámci tried, pre zachovanie početnosti tried.

Krížová validácia časových radov

Kvôli nadväznosti v poradí dát časových radov, je nevhodné pre metódy časových radov využiť štandardnú krížovú validáciu, preto sa využíva krížová validácia na základe posúvajúceho sa konca [54]. V prvom kroku použijeme k za sebou idúcich pozorovaní pre natrénovanie modelu a na $(k+1)$-tom pozorovaní model testujeme, následne sa zas posunieme o jedno časové obdobie, t.j. inkrementujeme k o 1 a zvolený počet krát opakujeme.



Obrázok 1: Ilustrácia rozdelenia tried do skupín pre krížovú validáciu pri náhodnom a pri stratifikovanom delení na 4 vzorky.

Na záver spriemerujeme hodnoty miery chýb z testovacích pozorovaní. V prípade potreby môžeme testovať, napríklad, aj na $(k + j)$-tom pozorovaní, kde j je zvolené prirodzené číslo.

Bootstrap

Bootstrap je obľúbená vzorkovacia metóda použiteľná na kvantifikáciu neistoty spojenej s odhadmi alebo metódou strojového učenia. Jej princíp spočíva v tom, že náhodne vyberáme pozorovania do novej množiny rovnakej veľkosti ako pôvodná, pričom môžeme to isté pozorovanie vybrať viackrát. Takýmto spôsobom zostavíme novú množinu pozorovaní s približne 63.2 % pôvodných pozorovaní. Z viacerých takýchto vzoriek množín môžeme následne získať, napríklad, odhad štandardnej chyby [55, s. 187].

1.4.5 Regresné metódy

Regresné metódy predpokladajú spojitú výstupnú premennú. Veľkou výhodou regresných metód je aj to, že okrem predikcií majú aj schopnosť inferencie – vyjadrenia miery vplyvu prediktorov na vektor výstupu, t.j. regresný koeficient v takomto prípade vyjadruje, o koľko sa zmení vektor výstupu ak sa prediktor zmení o jednu jednotku.

Viacnásobná lineárna regresia

Zatiaľ čo jednoduchá lineárna regresia sa zaoberá predikovaním výstupnej premennej na základe jednej vstupnej premennej, viacnásobná lineárna regresia sa zaoberá vzťahom medzi výstupnou premennou a vstupnými premennými, t.j. vzťahom medzi vektorom výstupu a prediktormi.

Vzťah viacnásobnej lineárnej regresie môžeme zapísať ako:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \tag{15}$$

kde β_0 je úrovňová konštanta, $\beta_1, \ldots, \beta_p$ sú regresné koeficienty a ϵ náhodná chyba modelu.

Ak chceme použiť maticový zápis, môžeme rozšíriť maticu prediktorov $\boldsymbol{X}$ o stĺpec tvorený 1-tkami, odpovedajúci úrovňovej konštante na maticu $\boldsymbol{X}^*$. Takáto matica bude o rozmere $(n \times (p+1))$. β bude vektor odhadovaných parametrov. Následne má vzťah (15) tvar:

$$y = \beta \boldsymbol{X}^* + \epsilon. \tag{16}$$

Odhad parametrov lineárnej regresie

Pre nájdenie odhadu parametrov β v regresnom modeli, slúži metóda najmenších štvorcov (angl. ordinary least squares - OLS), ktorá minimalizuje RSS vyjadrené ako

$$RSS = (y - \boldsymbol{X}^* \hat{\beta})^T (y - \boldsymbol{X}^* \hat{\beta}). \tag{17}$$

Derivovaním podľa β dostaneme odhad tohto parametra

$$\hat{\beta} = (\boldsymbol{X}^{*T} \boldsymbol{X}^*)^{-1} \boldsymbol{X}^{*T} y. \tag{18}$$

Meranie kvality modelu lineárnej regresie

Ak chceme pri lineárnej regresii merať, ako dobre sme odhadli vektor výstupu pomocou prediktorov, môžeme využiť koeficient determinácie - R^2. Ten získame ako

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \tag{19}$$

kde $\overline{y}$ je priemer vektora výstupu a menovateľ sa nazýva aj celková suma štvorcov. V prípade ak porovnávame dva modely trénované rovnakou metódou a druhý vznikol, napríklad, odstránením niektorých prediktorov, ten s vyšším počtom bude mať vždy vyššie alebo rovnaké R^2, pretože R^2 rastie s pridávaním prediktorov. V takomto prípade môžeme využiť adjungované R^2, ktoré penalizuje prebytočné prediktory v modeli [55, s. 213].

Potenciálne problémy lineárnej regresie

Medzi najčastejšie problémy lineárnej regresie môžeme zaradiť nasledovné tri problémy [55, 20, s. 92] :

1. *Nelinearita vzťahu medzi výstupnou premennou a vstupnými premennými* - pozorujeme pokiaľ majú rezíduá v grafe odhadnutých hodnôt a rezíduí nejaký skrytý vzor, napr. lievik, čo môže zároveň naznačovať aj nekonštantnosť rozptylu (heteroskedasticitu). Nelineárny vzťah môžeme presnejšie zachytiť napr. transformovaním vektora výstupu alebo prediktorov nelineárnymi funkciami.

2. *Vplyvné pozorovania* - pozorovanie je považované za vplyvné ak jeho odstránenie, príp. odstránenie viacerých takýchto pozorovaní spôsobí značné zmeny v odhadnutom modeli. Patria sem napríklad hodnoty mimo rozsah a pákové body [20, s. 108]. Takéto pozorovania môžeme detegovať napr. zobrazením Cookovej vzdialenosti [20, s. 111], ktorá je interpretovateľná ako vzdialenosť medzi vektorom výstupu modelu trénovaného na všetkých pozorovania a vektorom výstupu modelu trénovaného bez daného pozorovania.

3. *Multikolinearita* - Multikolinearitu definujeme ako vzájomnú závislosť prediktorov, t.j. hodnota prediktora je lineárne závislá od hodnôt iných prediktorov. Multikolinearita spôsobuje problém pri rozlišovaní efektu jednotlivých prediktorov na vektor výstupu. Multikolinearitu detegujeme pomocou faktor zväčšenia rozptylu (angl. variance inflation factor)
$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j|x_{-j}}}, \tag{20}$$
kde $R^2_{x_j|x_{-j}}$ je R^2 regresie x_j ako vektora výstupu a zvyšných prediktorov. Jednoduchú multikolinearitu, označovanú ako kolinearitu, kedy je prediktor závislý od iného prediktora, môžeme detegovať vyhodnocovaním Pearsonových korelácií medzi prediktormi. Obsiahly prehľad metód určených na vysporiadanie sa s multikolinearitou

nájdeme v [36].

Kategorické prediktory v regresii

Keďže regresia vyžaduje numerické hodnoty, kategorické hodnoty musíme vhodne reprezentovať. Ak má kategorický prediktor k kategórií, kde $k > 1$, transformujeme ho na $k-1$ prediktorov, aby sme sa vyhli úplnej multikolinearite. Tá by nastala, ak by sme kategorický prediktor kódovali k prediktormi, kde by sa každý dal napísať ako kombinácia zvyšných. Potom môžeme vplyv k-tej triedy indikovať ako nulovú hodnotu koeficientov $k-1$ prediktorov reprezentujúcich zvyšné kategórie.

Štandardizácia regresných koeficientov

Podľa [11] môžeme regresné koeficienty štandardizovať prenásobením smerodajnou odchýlkou ich prediktora, aby sme tak získali relatívny vplyv na vektor výstupu. Takto štandardizovaný koeficient nám ukáže, o koľko sa zmenia hodnoty vektora výstupu, ak sa zmení prediktor o jednu smerodajnú odchýlku.

1.4.6 Regresné metódy na výber premenných

V tejto podkapitole si predstavíme regresné metódy na výber premenných, kde vyberieme metódy krokovej selekcie a tolerantné metódy.

Metódy krokovej selekcie sú obľúbenými regresnými metódami, keďže dokážu odstránením nežiadúcich prediktorov značne zjednodušovať interpretáciu modelov a znižovať aj rozptyl týchto modelov. Tieto metódy iteratívne prehľadávajú množinu prediktorov za účelom vybrať čo najvhodnejšiu podmnožinu prediktorov. Ten prediktor, ktorý najmenej zhorší (zlepší) hodnotiace kritérium pridaný (odstránený). Dopredná a spätná selekcie sú však iba heuristickými algoritmami a negarantujú nájdenie optimálneho riešenia.

Selekcia hrubou silou

Táto regresná metóda odhadne všetky možné kombinácie prediktorov a zvolí tú najlepšiu. Jej zjavnou nevýhodou je vysoká náročnosť už pri $p > 20$ [55, s. 206], no na druhej strane garantuje nájdenie optimálneho riešenia v podobe najlepšej kombinácie prediktorov.

Dopredná selekcia

Dopredná selekcia začína prehľadávanie s prázdnou množinou prediktorov a postupne rozširuje model o prediktor, ktorý najviac zlepší hodnotiace kritérium. Z takto vytvorených

modelov potom na základe výsledkov krížovej validácie vyberie najlepší model.

Spätná selekcia

Spätná selekcia postupne odstraňuje prediktory z modelu obsahujúceho na začiatku všetky prediktory, podľa toho, ktorý najviac zhorší hodnotiace kritérium. Z takto vytvorených modelov potom na základe výsledkov krížovej validácie vyberie najlepší model.

Tolerantné metódy

Tolerantné metódy redukujú koeficienty lineárnej regresie, a poskytujú tak riešenie s väčším vychýlením no menším rozptylom. Takéto metódy možno všeobecne zapísať matematickým modelom ako minimalizáciu sumy štvorcov rezíduí a pokutovej funkcie

$$\underset{\beta}{argmin}\ RSS + \lambda g(.), \tag{21}$$

pričom $\lambda \geq 0$ je hyperparameter, ktorý kontroluje mieru redukcie $g(.)$ je pokutová funkcia, najčastejšie kombinácia alebo modifikácia l-noriem. Čím väčšiu hodnotu nadobúda λ, tým viac sú koeficienty redukované.

Elastic net

Elastic net je tolerantnou metódou využívajúca kombináciu l_1 a umocnenej l_2 normy

$$\underset{\beta}{argmin}\ RSS + \lambda \sum_{j=1}^{p} (\alpha|\beta_j| + \frac{(1-\alpha)}{2}\beta_j^2), \tag{22}$$

kde $\alpha \in \langle 0; 1\rangle$ je parameter voľby medzi vplyvom noriem. Parameter λ potom volíme na základe krížovej validácie. V prípadoch, kedy nájdeme optimálnu hodnotu parametra λ pomocou krížovej validácie v najmenšej hodnote miery chybovosti, označíme ho λ_{min}^{CV} a takto odhadnutý parameter β_j ako β_j^{CV}. Hyperparametrom α môžeme regulovať riedkosť modelu. Čím je α bližšie k 1, tým je väčšia šanca, že bude model redší, pretože l_1 norma v lasso metóde zabezpečuje, že niektoré koeficienty môžu byť presne rovné 0. Tento hyperparameter môžeme nájsť aj pomocou krížovej validácie. Metódu *lasso* získame nastavením parametra *alpha* na hodnotu 1 lasso [55, s. 219]. Toto je výhodou najmä pri vysokom počte prediktorov, kedy takýmto spôsobom lasso vykoná selekciu prediktorov a výsledkom je riedky model. Ďalší špeciálny prípad, ak je $\alpha = 0$, sa nazýva hrebeňová regresia [55, s. 215]. Hrebeňová regresia vďaka pokutovej funkcii $\lambda \sum_{j=1}^{p} \beta_j^2$ znižuje s rastom hodnoty

parametra λ rozptyl za cenu zvýšeného vychýlenia. V prípadoch, kedy nájdeme parameter λ pomocou krížovej validácie ho označíme ako λ_{min}^{CV} a takto odhadnutý parameter β_j ako β_j^{CV}.

OSCAR a PACS

Metóda octagonal shrinkage for clustering and regression (OSCAR) [8], inšpirovaná metódou elastic net, využíva kombináciu dvoch pokutových členov

$$\underset{\beta}{argmin}\ RSS + \lambda\Big[\sum_{j=1}^{p}|\beta_j| + c\sum_{1\leq j<k\leq p} max(|\beta_j|,|\beta_k|)\Big], \tag{23}$$

kombináciu l_1 normy a párovej l_∞ normy, pričom $c \in \langle 0, \infty\rangle$ kontroluje relatívnu váhu noriem a λ ich mieru vplyvu. Takáto kombinácia pokutových funkcií, ako už z názvu vyplýva, spôsobuje zhlukovanie korelovaných prediktorov. Čím vyššie nastavíme konštantu c, tým viac bude metóda prediktory zhlukovať. Zhlukovanie sa prejavuje tak, že metóda korelovaným prediktorom nastaví koeficienty s rovnakou alebo približne rovnakou absolútnou hodnotou, čo môžeme využiť práve na identifikovanie podobných prediktorov a prípadne vybrať nejakú z nich ako zástupcu ostatných.

V [93] autori publikovali rozšírenie metódy OSCAR, ktorá je výpočtovo jednoduchšia a nazvali ju pairwise absolute clustering and sparsity (PACS). Využívajú tu konvexnú pokutovú funkciu, ktorá podporuje rovnosť hodnôt koeficientov, vďaka penalizovaniu párových súčtov a rozdielov koeficientov, a takým spôsobom zhlukuje prediktory. Optimalizačnú úlohu možno napísať ako

$$\underset{\beta}{argmin}\ RSS + \lambda\Big[\sum_{j=1}^{p} w_j|\beta_j| + \sum_{1\leq j<k\leq p} w_{jk(-)}|\beta_k - \beta_j| \sum_{1\leq j<k\leq p} w_{jk(+)}|\beta_j + \beta_k|\Big], \tag{24}$$

kde w_j a w_{jk} sú nezáporné váhy. Podobnosť pokutovej metódy OSCAR a PACS je možné ukázať aj tým, že $\max|\beta_j|,|\beta_k| = \frac{1}{2}\big(|\beta_k - \beta_j| + |\beta_j + \beta_k|\big)$.

1.4.7 Lineárne klasifikačné metódy

V prípade, ak máme kategorickú, inak povedané aj kvalitatívnu, výstupnú premennú, vhodnejším spôsobom na predpovedanie je priraďovanie pozorovaní do tried - klasifikácia. Často však klasifikačné metódy predikujú najskôr pravdepodobnosť a správajú sa v tomto zmysle podobne ako regresné metódy [55]. V tejto časti si pre účely práce predstavíme

takéto metódy s využitím na binárnu klasifikáciu. Výstupnú premennú budeme modelovať ako náhodnú premennú $Y \in \{0, 1\}$. Lineárne klasifikačné metódy sú také, ktoré modelujú výstupnú premennú pomocou lineárnych vzťahov so vstupmi.

Logistická regresia

Logistická regresia je populárna metóda pre binárnu klasifikáciu, založená na optimalizačnej úlohe, ktorá sa dá riešiť metódami konvexnej optimalizácie. Logistická funkcia vychádza z pravdepodobnostného modelu a logistickej funkcie

$$P(Y = 1|\boldsymbol{X}^*) = \frac{1}{1 + e^{-(\beta \boldsymbol{X}^*)}}. \tag{25}$$

Maximálne vierohodný odhad parametrov β_0 a β v rovnici (25) nájdeme riešením optimalizačného problému

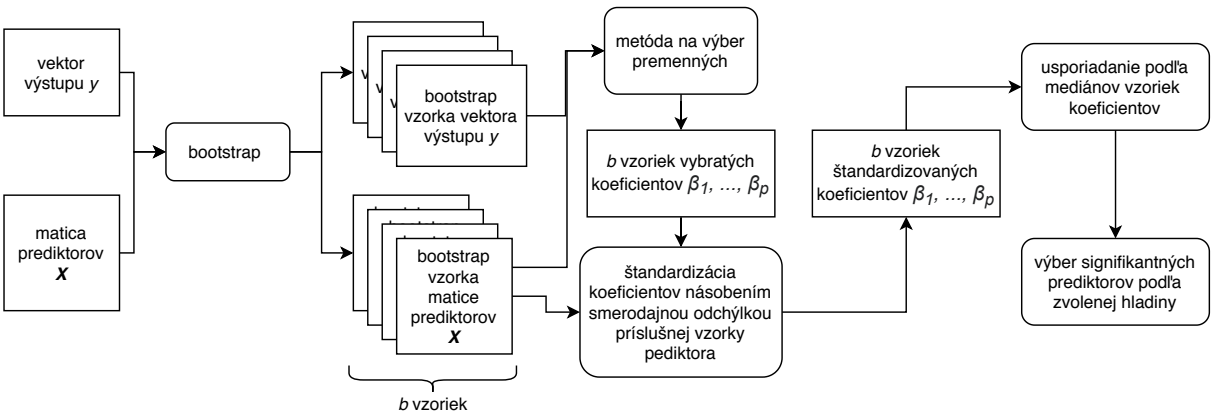
$$\underset{\beta_0,\beta}{\text{maximize}} \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^T \boldsymbol{x}_i) - \log\left(1 + e^{\beta_0 + \beta^T \boldsymbol{x}_i}\right) \right\}. \tag{26}$$

Podobne ako v prípade metódy lasso, logistickú regresiu s l_1 pokutou (LR-l_1) dostaneme pridaním l_1 regularizácie k účelovej funkcii (26) a teda formulujeme nasledovný optimalizačný problém

$$\underset{\beta_0,\beta}{\text{maximize}} \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^T \boldsymbol{x}_i) - \log\left(1 + e^{\beta_0 + \beta^T \boldsymbol{x}_i}\right) \right\} - \lambda \left\|\beta\right\|_1, \tag{27}$$

pre $\lambda \geq 0$. Pokutová funkcia l_1 umožňuje nastaviť β koeficienty menej informatívnych prediktorov z pohľadu výstupnej premennej na nulu a tak zvyšuje jednoduchosť a vysvetľovaciu schopnosť modelu. Hyperparameter λ v účelovej funkcii (27) umožňuje voliť pomer medzi kvalitou odhadu a riedkosťou (angl. sparsity - počtom prediktorov) modelu. Vzťah (27) je oproti vzťahu (22) maximalizačný, keďže vierohodnostná funkcia má záporné znamienko a to zmení maximalizáciu na minimalizáciu.

Rovnica (25) sa tiež používa na predikciu. Odhadnuté hodnoty regresných koeficientov $\hat{\beta}_0$ a $\hat{\beta}$ spolu s pozorovaniami $\boldsymbol{x}$ sa vložia do pravej strany rovnice (25) a výsledná hodnota je potom $\hat{y} = P(Y = 1)$. Binárnu hodnotu potom dosiahneme porovnaním s prahovou hodnotou θ.



Obrázok 2: Diagram štatistickej inferencie pomocou metódy bootstrap.

1.4.8 Štatistická inferencia pomocou metódy bootstrap

Hlavným cieľom pri inferencii v regresných úlohách je získať prehľad o vplyve prediktorov na vektor výstupu, ako aj informácia o tom, do akej miery prediktor súvisí s vektorom výstupu, čo indikuje p-hodnota jeho koeficienta. Klasické p-hodnoty, získané pre odhady pomocou OLS, nemusia po dátovo riadenom výbere premenných byť ďalej platné, keďže je model vybratý stochastickým procesom, čo klasická štatistická teória negarantuje [5]. Za takýto stochastický proces sa môžu považovať, napríklad, metódy výberu premenných predstavené v predošlej častiach. Nás budú hlavne zaujímať metódy inferencie po výbere premenných pri $l-1$ regularizácií.

V [112] poskytuje 6. kapitola komplexný prehľad takýchto metód, avšak väčšina z nich má konzervatívne štatistické podmienky, ktoré je ťažké splniť veľkou škálou prediktorov s rôznymi rozdeleniami alebo nevedia priamo vypočítať p-hodnoty.

Metóda uvedená v [112, s. 142] využíva prevzorkovanie dát pomocou bootstrapu a poskytuje tak empirické pravdepodobnostné rozdelenie koeficientov, vyjadrujúce ich mieru stability. Ak by koeficient často presahoval nulovú hodnotu (menil znamienko) alebo ju dosahoval, poukazuje to na jeho nespoľahlivosť – nízku vierohodnosť vplyvu prediktora na vektor výstupu. Pri metódach s $l-1$ normou môžeme využiť vlastnosť výberu premenných v kombinácii s boostrapom na získanie frekvencie výberu daného koeficientu (podiel nulových hodnôt koeficientu), ktorú môžeme použiť ako mieru signifikantnosti.

Pre lasso a LR-l_1 je odhad p-hodnôt vďaka adaptívnej procedúre komplikovaný. Aby nedošlo k prezentácií nespoľahlivých výsledkov, a aj na základe odporúčaní v podobnej štúdií [53], tak pre získanie informácie o signifikantnosti koeficientov vykonávame štatistickú inferenciu pomocou metódy bootstrap. Celý tento postup je zobrazený v Obrázku 2.

Vytvárame b vzoriek dát z pozorovaní vektora výstupu y a matice prediktorov $\boldsymbol{X}$. Na každej takto vzniknutej vzorke dát natrénujeme model danej metódy, t.j. dostaneme b modelov. Takto získame empirické rozdelenia regresných koeficientov.

Keďže vplyv prediktora na vektor výstupu rastie s absolútnou hodnotou prislúchajúceho štandardizovaného koeficientu [96, s. 372], vykonávame štandardizáciu koeficientov. Každý koeficient modelu štandardizujeme násobením smerodajnou odchýlkou prislúchajúcej vzorky prediktora. Následne koeficienty usporiadame na základe mediánov ich vzoriek. Absolútna hodnotu mediánu bootstrapovej realizácie koeficient predstavuje mieru sily vplyvu prediktora na vektor výstupu. Pozitívne (negatívne) znamienko koeficientu indikuje kladný (záporný) vplyv prediktora. Rôzne znamienka regresných koeficientov pre realizácie bootstrapu môžu byť prisúdené nízkej signifikantnosti alebo výberu korelovaných prediktorov [112, p. 144]. Preto čím konzistentnejšie sú hodnoty štandardizovaného regresného koeficientu pre bootstrapové vzorky, tým signifikantnejší je prediktor prislúchajúci regresnému koeficientu. Na základe zvolenej hladiny požadovanej signifikantnosti potom vyberieme prediktory na podľa počtu výberov modelom a ako aj opačných znamienok voči mediánu vzorky.

1.4.9 Metódy založené na rozhodovacích stromoch

Gradient boosted regression trees (GBRT) a random forest (RF) sú metódy založené na rozhodovacích stromoch. Rozhodovacie stromy teda delia hodnoty prediktorov $x_i, ..., x_p$ na $J \leq n$ neprekrývajúcich sa regiónov obsahujúcich pozorovania priraďujúc tak triedy k pozorovaniam [55, s. 306]. Pozorovania postupujú na základe podmienok v uzloch od koreňa až po list, ktorý reprezentuje región.

Random forest

RF, alebo aj náhodný les, využíva metódu baggingu (bootstrap aggregation), t.j. kombináciu bootstrapu a priemerovania (agregácie). Stručne je postup tvorby lesa nasledovný: Najskôr sa vytvorí trénovacia množina b skupín, vzorkovaním pôvodných dát pomocou bootstrapu, na ktorých sa natrénuje b stromov. Nakoniec sa všetky predikcie pre dané pozorovanie spriemerujú na jednu predikciu, za cieľom zníženia rozptylu modelu. Pri každom delení na vnútornom uzle, vytvoríme náhodnú množinu prediktorov o veľkosti m, čím znížime podobnosť stromov trénovaných na dátach získaných bootstrapom [55, s. 319]. Podľa [51, s. 588] je algoritmus trénovania pre náhodný les nasledovný:

1. Vyberieme vzorku dát pomocou metódy bootstrap.
2. Budujeme rozhodovací strom T_b na vzorke dát rekurzívnym opakovaním nasledovných krokov pre každý rozhodovací uzol stromu, pokým nedosiahneme zastavovacie kritérium.

   Vyberieme náhodne m z p prediktorov.

   Vyberieme najlepšie rozdeľovacie kritérium z vybraných prediktorov.

   Rozdelíme vrchol na dvoch potomkov.
3. Vrátime skupinu stromov ${T_b}^{B_1}$.

Predikciu pre nové pozorovanie získame:

- pre regresné stromy ako priemer predikcií zo stromov;
- pre klasifikačné stromy väčšinovým hlasovaním za výslednú triedu.

Gradient boosted regression trees

GBRT vylepšuje predikcie rozhodovacích stromov pomocou boostingu, ktorý produkuje pomocou M slabých klasifikátorov silnú komisiu. Slabé klasifikátory sú zvyčajne jednoduché stromy iteratívne trénované na rezíduách predošlých modelov, v malej miere lepšie predikujúce ako nulový model. Pre spresnenie klasifikácie sa snaží metóda minimalizovať MSE pomocou gradientu, kde rezíduum je rozdiel medzi aktuálnym a predošlým odhadom klasifikátora. V m-tej iterácií, kde $m \in 1, \ldots, M$ dostaneme predikciu rekurzívnym vzťahom

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \sum_{j=1}^{J^{(m)}} \gamma_j^{(m)} \cdot 1\{\boldsymbol{x} \in R_{j^{(m)}}\}, \tag{28}$$

kde F je predikcia slabého klasifikátora, $J^{(m)}$ je počet konečných regiónov $R_1^{(m)}, \ldots, R_{J^{(m)}}^{(m)}$ (každý prislúchajúci jednému listu) a $\gamma_1^{(m)}, \ldots, \gamma_{J^{(m)}}^{(m)}$ sú váhy odhadnuté minimalizáciou MSE pri každej iterácií m. Predikciu získame M-tou iteráciou algoritmu.

1.4.10 Analýza časových radov

Predpovedanie časových radov je dôležitá technika dátovej analýzy, využívaná ako základ pre expertné a automatické plánovanie v mnohých aplikačných doménach akými sú predaje, riadenie dopravy a riadenie energie. Oproti analýze dát z minulosti, nám

predpovedanie umožňuje náhľad na možný budúci vývoj, čo je hodnotné pri vykonávaní strategických rozhodnutí [31]. Definíciu časového radu z podkapitoly 1.1 upresníme na časový rad s pravidelnou frekvenciou, t.j. konštantným časom medzi dvomi nasledujúcimi pozorovaniami. Frekvencia je teda dĺžka úseku medzi dvomi nasledujúcimi pozorovaniami, napr., pri časových radoch s dennou frekvenciou predstavuje jedno pozorovanie jeden deň.

Komponenty časového radu

Časový rad môže vykazovať rôzne vzory, ktoré sa dajú identifikovať jeho rozdelením na viacero komponentov, každý reprezentujúci jeden vzor. Prvým takýmto vzorom je trendovo-cyklický komponent T_t, často nazývaný len trendový, ktorý zachytáva dlhodobý pokles alebo nárast. Cyklom nazývame výkyvy v dátach s nepravidelnou prióodou výskytu. Takéto fluktuácie sa často vyskytujú v ekonomických dátach, napr. ekonomické cykly. Ďalším komponentom je sezónna zložka S_t, ktorá predstavuje opakujúci sa štrukturálny vzor v určitej časovej perióde. Príkladom je ovplyvňovanie časového radu dňami v týždni – týždenná sezónnosť. Vyskytuje sa oproti cyklom s pravidelnou a známou frekvenciou výskytu [54]. Aditívnu dekompozíciu časového radu potom vieme zapísať ako

$$y_t = S_t + T_t + R_t, \tag{29}$$

kde R_t je zvyšok. Multiplikatívnu dekompozíciu získame výmenou sčítania za násobenie.

Stacionarizácia časového radu

Ak y_t je stacionárny časový rad, tak pre ľubovoľné s, pravdepodobnostné rozdelenie prvkov časového radu $y_{t+1}, \ldots, y_{t+s}$ nezávisí od času t – je v čase nemenné. Nestacionaritu spôsobujú napríklad trendová zložka, sezónnosť alebo monotónne meniaci sa rozptyl radu. Za stacionárny časový rad považujeme, napríklad, biely šum, čo je časový rad nevykazujúci závislosť na predošlých zložkách (autokoreláciu), generovaný z jedného pravdepodobnostného rozdelenia s nulovou strednou hodnotou [54]. Na odstránenie rastúceho alebo klesajúceho rozptylu hodnôt časového radu vieme využiť rôzne transformácie týchto hodnôt ako napríklad Box-Cox transformáciu.

Stacionarizácia diferenciami

Stabilizáciou rozptylu však nemusíme stabilizovať strednú hodnotu. Diferencovanie pomáha stabilizovať strednú hodnotu časového radu a oddeliť tak trend a sezónnosť, čím

vhodnejšie pripraví časový rad pre predikčné metódy. Prvú diferenciu definujeme ako

$$y'_t = y_t - y_{t-1}. \tag{30}$$

Druhá diferencia je rozdiel medzi diferenciami y'_t a y'_{t-1}.

Špeciálnym prípadom diferencií sú sezónne diferencie, definované ako rozdiel aktuálnej hodnoty a hodnoty posunutej o sezónu dĺžky m

$$y'_t = y_t - y_{t-m}. \tag{31}$$

Predikovanie hodnôt

Predpoveď hodnoty y_t budeme označovať ako $\hat{y}_t$, predpoveď zo všetkých predošlých pozorovaní $y_1, \ldots, y_{t-1}$ ako $\hat{y}_{t|t-1}$. Potom $y(T+h|T)$ definujeme ako predpoveď h období dopredu zo všetkých pozorovaní po čas T. Rezíduum definujeme ako rozdiel medzi predikciou a reálnou hodnotou.

Diagnostika rezíduí

Diagnostika rezíduí je vhodná na kontrolu, či model adekvátne zachytil všetky pravidelnosti v dátach. Výsledkom dobrej predikčnej metódy sú rezíduá s nasledovnými vlastnosťami [54]:

- Rezíduá sú nekorelované. Ak sa medzi nimi nachádza korelácia, tak model pravdepodobne nezachytil nejakú informáciu.
- Rezíduá majú nulovú strednú hodnotu. Ak majú rezíduá inú ako nulovú strednú hodnotu, výsledky sú vychýlené.

Ďalej ešte autori považujú za užitočné, ale nie nevyhnutné, zvážiť dve vlastnosti rezíduí a tými sú konštantný rozptyl a normálne rozdelenie.

Pre analýzu rezíduí slúžia rôzne testy ako napríklad Durbin-Watson test zobrazenia autokorelačnej funkcie (angl. autocorrelation function - ACF) a parciálnej autokorelačnej funkcie (angl. partial autocorrelation function - PACF) rezíduí alebo rôzne iné grafické zobrazenia, v ktorých môžeme podobne ako v regresii nájsť vzory v rezíduách [72]. ACF je autokorelačná funkcia merajúca korelácie medzi aktuálnym pozorovaním a minulými pozorovaniami. Aby sa odstránili korelácie s predošlými hodnotami a tak vyzdvihli iné

požadované korelácie, bola navrhnutá PACF. PACF meria vzťah medzi y_t a y_{t-k} tak, že odstráni efekt predošlých oneskorení $t = 1, 2, \ldots, k-1$ a odhadne tak závislosť aktuálnej hodnoty na oneskorení.

ARIMA modelovanie

Autoregressive integrated moving average (ARIMA) je druh modelovania časových radov, ktoré majú niektorú z troch ďalej popísaných zložiek. Prvou zložkou je autoregresná zložka, ktorá predstavuje lineárnu závislosť aktuálneho pozorovania y_t v čase t na p predošlých pozorovaniach, vtedy vravíme o AR(p) procese. Ďalšou zložkou je zložka kĺzavých priemerov, ktorá využíva predošlé chyby modelov. Ak aktuálna hodnota y_t je závislá na q predošlých chybách, hovoríme o MA(q) procese. ARMA (p, q) proces potom zapíšeme ako

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}. \tag{32}$$

Vo vyššie uvedenom modeli sú $\phi_1, \phi_2, \ldots, \phi_p$ parametre asociované s oneskorenými premennými, $\theta_1, \theta_2, \ldots, \theta_q$ sú parametre rezíduí q predošlých modelov (kĺzavé priemery) a $\epsilon_t, \epsilon_{t-1}, \epsilon_{t-q}$ je vektor rezíduí považovaný za biely šum. Treťou zložkou ARIMA modelu je zložka diferencií a stupeň vykonaných diferencií definuje parameter d. Príkladom modelu môže byť

$$y'_t = \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1}, \tag{33}$$

ktorý je ARIMA$(2, 1, 1)$ model, keďže má 2 autoregresné parametre ($p = 2$), použité sú diferencie prvého rádu ($d = 1$) a 1 kĺzavý priemer ($q = 1$).

SARIMA modelovanie

SARIMA modelovanie je kombináciou dvoch ARIMA modelov, kde druhý model je sezónny (angl. seasonal). Pri takýchto modeloch sa využíva značenie SARIMA$(p, d, q)(P, D, Q)_m$, kde v sezónnom modeli je m počet pozorovaní za danú sezónu. Napríklad, pri týždennej sezónnosti a dennej frekvencií dát je $m = 7$. Ak by teda vyššie uvedený ARIMA$(2, 1, 1)$ model mal týždennú sezónnosť, s tým že by aktuálna hodnota závisela od hodnoty týždeň dozadu, značili by sme tento model ako SARIMA$(2, 1, 1)(1, 0, 0)_7$.

Výber a odhad parametrov

Parametre v SARIMA modeloch a takisto sezónnosť môžeme voliť na základe ACF a PACF funkcií. Podľa grafov ACF a PACF funkcií vieme potom následne určiť AR a MA

stupeň. Pre správne určenie sezónnosti môžeme taktiež využiť grafy oneskorení, sezónne grafy a polárne sezónne grafy [54], ktoré poskytujú jednoduchú a zrozumiteľnú vizualizáciu časových radov z pohľadu sezónnosti. Pri výbere medzi modelmi s hodnotami parametrov p a q sa môžeme napríklad riadiť podľa hodnôt AIC, BIC alebo AICc [54].

Exogénne premenné

Do ARIMA modelov môžeme zahrnúť exogénne premenné, ako napríklad dni prázdnin, počasie a.p. Takéto ARIMA modely rozšírené o iný časový rad sa nazývajú ARIMAX modely. Časové rady vykazujú často nepravidelné výkyvy, ktoré môžeme odhadnúť doplnením exogénnych premenných, alebo nám exogénne premenné môžu slúžiť napríklad na odhad ďalšej sezónnej zložky. ARIMAX modely so sezónnousťou budeme označovať SARIMAX modely.

Binárne premenné

Binárne premenné nám môžu poukazovať na výskyty nejakých javov. Potom môžeme zahrnúť taký vektor ako exogénnu premennú do modelu. Napríklad, vektor prázdnin, kde 1 bude ak sú v daný deň prázdniny a 0 inak.

Fourierove rady

Alternatívou ako modelovať sezónne zložky časových radov sú Fourierove rady, ktoré dokážu pomocou matematických funkcií *sin* a *cos* aproximovať tieto zložky. Takýto rad potom pridáme do modelu ako exogénnu premennú. Ak pridáme Fourierov rad ako exogénnu premennú do ARMA modelu, nazývame takýto model F-arma [69] a môžeme ho zapísať ako

$$y_t = a + \sum_{k=1}^{K} [\alpha_k sin(2\pi \frac{kt}{m}) + \beta_k cos(2\pi \frac{kt}{m})] + N_t \tag{34}$$

kde m je počet pozorovaní v sezóne, a je vyrovnávacia konštanta, N_t je ARIMA model, K je počet Fourierových členov α_k a β_k sú parametre.

1.4.11 Zhlukovacie metódy

Pri zhlukovacích metódach sa snažíme nájsť skupiny podobných pozorovaní, pričom tieto pozorovania v skupine majú byť zároveň odlišné od pozorovaní z ostatných skupín [55, s. 385].

K-means

Metóda delí pozorovania na k zhlukov, kde sa jedno pozorovanie priradí práve jednému zhluku. Parameter k stanovujeme vopred.

Algoritmus pre riešenie k-means podľa [55, s. 388] vyzerá nasledovne:

1. Náhodne priraď číslo od 1 po k každému pozorovaniu.

2. Iteruj, pokiaľ sa priradenia k zhlukom neprestanú meniť:

   Pre každý z k zhlukov vypočítaj centrum zhluku. Centrum zhluku je priemer pozorovaní v danom zhluku.

   Každé pozorovanie priraď k najbližšiemu zhluku, ktorý je definovaný euklidovskou vzdialenosťou.

Výsledok k-means závisí však od počiatočného stanovenia bodov vyššie uvedeného algoritmu a tak tento algoritmus často nájde iba lokálne optimum. Pre voľbu vhodného k je možné využiť expertný odhad alebo metódy ako metóda lakťa (elbow method) [49, s. 486], ktorá spočíva v zobrazení vzájomnej vzdialenosti prvkov v rámci zhluku oproti počtu zhlukov. Táto vzdialenosť väčšinou s počtom zhlukov klesá a snažíme sa nájsť "lakťový bod", t.j. počet zhlukov, pri ktorom už pokles vo vzdialenosti nie je tak výrazný.

Hierarchické zhlukovanie

Možnou alternatívou pre k-means je hierarchické zhlukovanie, nevyžadujúce voľbu k vopred. Hierarchické zhlukovanie priraďuje pozorovania do hierarchického usporiadania v stromovej štruktúre. Vizualizácie tejto štruktúry sa nazývajú dendrogramy. Bližšie si popíšeme aglomeračné hierarchické zhlukovanie, využívajúce spôsob zdola nahor. Tento algoritmus zhlukovania zlučuje iteratívne pozorovania na základe vzdialenosti do zhlukov a zhluky do väčších zhlukov, pokiaľ nevytvoria všetky pozorovania jeden spoločný zhluk.

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) je algoritmus, ako už jeho názov napovedá, založený na hustote bodov. Algoritmus spája body na základe vopred stanoveného parametra hustoty $MinPts$ a parametra minimálnej vzdialenosti ϵ. Výhodou tohto algoritmu je, že dokáže nájsť zhluky rôznych tvarov, pričom predošlé dva algoritmy nachádzajú zhluky guľovitého alebo elipsoidného tvaru.

1.5 Prehľad dostupných dát

Sekcia je venovaná prehľadu rôznych druhov dát, ktoré v práci využívame. Tie delíme na dve hlavné skupiny: dáta z nabíjacej infraštruktúry a dáta geografických informačných systémov (GIS).

1.5.1 EVnetNL dáta

Spoločnosť ElaadNL nám poskytla dva datasety pochádzajúce z nabíjacej infraštruktúry v Holandsku, spoločne nazývané EVnetNL. Nabíjacie udalosti v dátach začínajú v januári 2012 a končia v marci 2016. Dáta boli počas tvorby tejto práce aktualizované, ale firma ElaadNL odovzdala značnú časť infraštruktúry samosprávam a stratila tak prístup k dátam týchto staníc. Preto sme sa zamerali na časové obdobie, ktoré predchádzalo zmene prevádzkovateľa. Prvý dataset sa nazýva Meterreading, alebo odčítania z merača umiestneného na stanici, a obsahuje vyše 32 000 000 pozorovaní z nabíjania EV, ktoré sú zaznamenávané každých 15 minút, kým je vozidlo pripojené. Popísaný je v tabuľke 2.

Názov stĺpca	Popis stĺpca
Transaction_index	Unikátny index transakcie (transakcia začína pripojením a končí odpojením EV).
Charge point	Identifikátor nabíjacej stanice.
Connector	Číslo nabíjacieho bodu (stanice majú 1 alebo viac nabíjacích bodov).
Collectedvalue	Celková spotrebovaná energia na stanici pre daný nabíjací bod vo Wh meraná v približne 15 minútových intervaloch.
AveragePower	Priemerný výkon v kW medzi dvomi po sebe nasledujúcimi odčítaniami hodnôt merača (pre jednu transakciu).
UTCTime	Časová pečiatka merania spotreby v UTC časovej zóne, vo formáte „YYYY-MM-DD hh:mm:ss“.
EnergyInterval	Energia v kWh spotrebovaná medzi dvomi po sebe nasledujúcimi odčítaniami (pre jednu transakciu).

Tabuľka 2: Stĺpce datasetu Meterreadings a ich popis.

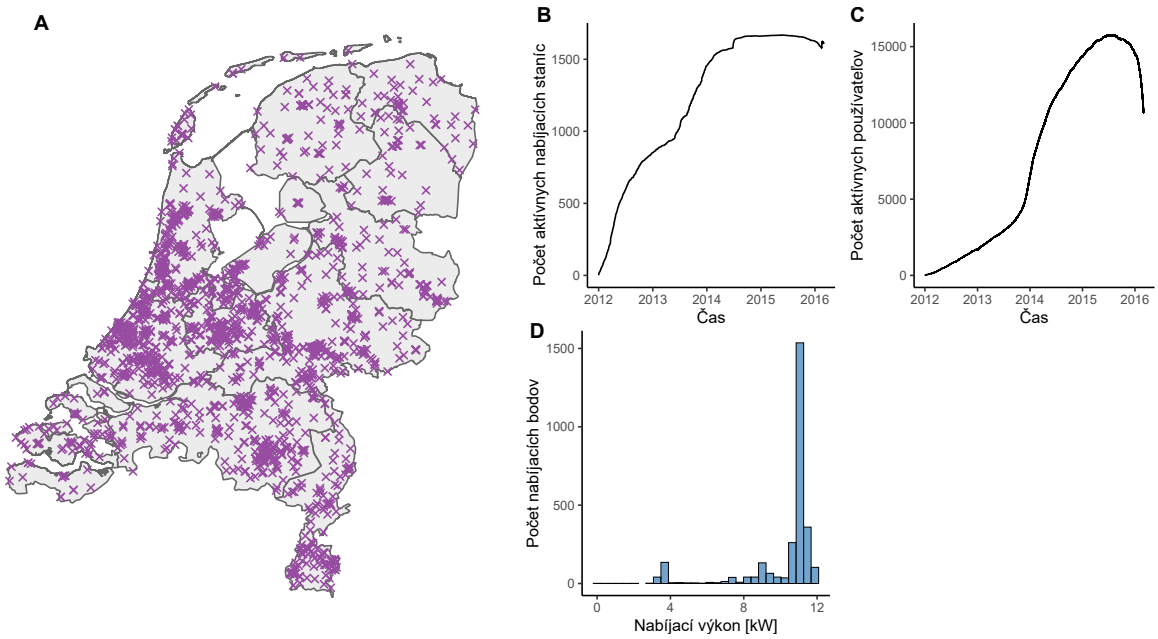
Druhý dataset *Transactions* vznikol pretransformovaním a doplnením datasetu Me-

terreadings. Tento dataset obsahuje viac ako 1 000 000 nabíjacích transakcií a obsahuje stĺpce uvedené v tabuľke 3.

Názov stĺpca	Popis stĺpca
Transaction_index	Identifikátor nabíjacej udalosti.
ChargePoint	Identifikátor nabíjacej stanice.
Connector	Číslo nabíjacieho bodu.
StartCard	RFID (radio frequency identifications) identifikátor karty, ktorá bola použitá na inicializáciu nabíjacej transakcie.
StopCard	RFID identifikátor karty, ktorou sa ukončila nabíjacia transakcia.
UTCTransactionStart	Čas začiatku nabíjacej transakcie v UTC časovej zóne.
UTCTransactionStop	Čas konca nabíjacej transakcie v UTC časovej zóne.
MeterStart	Stav merača spotreby energie na začiatku nabíjacej udalosti.
MeterStop	Stav merača spotreby energie na konci nabíjacej udalosti.
ConnectedTime	Doba pripojenia vypočítaná ako rozdiel medzi časom začiatku a konca nabíjacej udalosti.
ChargeTime	Celkový čas, kedy sa vozidlo nabíjalo.
IdleTime	Celkový čas, kedy vozidlo bolo pripojené, no nenabíjalo sa.
Latitude	Súradnica určujúca polohu stanice - zemepisná šírka.
Longitude	Súradnica určujúca polohu stanice - zemepisná dĺžka.
TotalEnergy	Celková energia spotrebovaná počas nabíjacej transakcie.
MaxPower	Maximálny nameraný výkon počas nabíjacej transakcie.

Tabuľka 3: Stĺpce datasetu Transactions a ich popis.

Celkovo sú v datasetoch údaje z 1747 pomalých nabíjacích staníc s výkonom do 12 kW (Obrázok 3**D**), pričom tieto stanice postupne pribúdali spolu s používateľmi (Obrázok 3**BC**). Tieto aktívne počty nabíjacích staníc a používateľov EV sme odhadli na základe transakcií, kde stanica aj používateľ začína byť aktívny prvou transakciou a končí poslednou transakciou. Ako môžeme vidieť na mape Obrázok 3**A**, stanice sú rozmiestnené po celom území Holandska, s vyššou koncentráciou na západe krajiny okolo väčších miest ako Utrecht, Rotterdam a Haag, avšak s chýbajúcimi dátami v hlavnom meste Amsterdam. Počet používateľov EV odhadujeme na základe identifikátorov použitých RFID kariet na približne 54 000.



Obrázok 3: **A** Rozmiestnenie ElaadNL nabíjacích staníc na území Holandska; **B** Počet aktívnych nabíjacích staníc v čase; **C** Počet aktívnych používateľov reprezentovaných RFID kartami; **D** Histogram maximálneho výkonu nabíjacích bodov.

Mimo týchto dát máme k nabíjacím staniciam aj informáciu o tom, či boli umiestnené na základe dopytu alebo na základe strategického plánovania, ako aj druh najbližšieho cestného úseku (cesta 1., 2., alebo 3. triedy).

1.6 GIS dáta

Pomocou dát z geografických informačných systémov (GIS) reprezentujeme okolie nabíjacích staníc, o ktorom predpokladáme, že vplýva na ich využívanie. GIS dáta charakterizujú zemský povrch a väčšina z nich pochádza z verejne dostupných zdrojov. Pre reprezentáciu priestorových údajov používame dva základné modely [58, s. 55]:

- rastrová reprezentácia - modeluje časti zemského povrchu pomocou buniek pravidelného tvaru, ktoré sú charakterizované hodnotami atribútov;
- vektorová reprezentácia - modeluje časti zemského povrchu a objekty na zemskom povrchu pomocou základných geometrických tvarov (body, čiary, mnohouholníky), ktoré sú charakterizované hodnotami atribútov.

Príkladom charakteristiky priestorového objektu (atribútu) je počet obyvateľov. Pre niektoré atribúty uvádzame skratky veľkými písmenami s číslom ako dolným indexom, pretože ich samotný popis je moc dlhý a čitateľ ich nájde v prílohe A.

1.6.1 Bodové dáta

Bod je element vektorovej reprezentácie, jednodznačne definovaný súradnicami v priestore [58, s. 55]. Bodové dáta budeme v práci využívať hlavne na reprezentáciu objektov v okolí stanice.

OpenStreetMap

OpenStreetMap (OSM) je jedna z najúspešnejších otvorených mapových platforiem [81]. Z OSM pre územie Holandska sme extrahovali body záujmu (angl. points of interes - POI), ktoré sa nachádzajú v 2 km $\times$ 2 km štvorcových území so stredom v mieste nabíjacieho miesta. Identifikovali sme 593 rôznych druhov POI, pričom niektoré sa vyskytovali veľmi zriedka. Z tohto dôvodu sme rozdelili POI do 15 kategórií, vymenovaných v Tabuľke 20 v Prílohe ***1***, modelujúcich miesta v okolí stanice, ktoré očakávame, že môžu navštíviť používatelia EV. Zvažovali sme aj použitie GoogleMaps [46], no tie poskytujú iba informácie o aktuálnych POI, a tak sme zvolili použitie historických OSM dát [80] z obdobia, kedy boli EVnetNL dáta zbierané.

Nabíjacie stanice 2015

Pre odhad pozície všetkých dostupných nabíjacích staníc v roku 2015, vyskladali sme dataset Nabíjacie stanice 2015 z databáz OpenChargeMap [79] a OplaadPalen [82]. Podľa dátumu, kedy stanica pribudla do databázy, sme z OpenChargeMap a OplaadPalen vyextrahovali pozície nabíjacích staníc, ktoré boli dostupné na konci roka 2015.

1.6.2 Dáta lomených čiar

Lomená čiara (polyline) je potom postupnosť úsečiek, kde koncovým bodom jednej úsečky začína druhá úsečka, okrem začiatočného a koncového bodu [58, s. 55].

Dopravné toky

Na modelovanie vplyvu nabíjacích staníc, berieme do úvahy aj dáta s dopravnými tokmi v Holandsku [114] (angl. traffic flows). Dáta sú organizované pomocou grafového modelu cestnej siete s vysokým rozlíšením, reprezentované lomenými čiarami. Pre každý cestný úsek sú k dispozícii dopravné toky pre tri typy vozidiel, osobné autá, autobusy a nákladné autá a tri časti dňa, deň, večer a noc. Podrobný opis atribútov sa nachádza v Prílohe ***1*** v tabuľke 19, pričom skratky atribútov tohto datasetu začínajú písmenami TF.

1.6.3 Polygónové dáta

Polygón je definovaný uzavretou lomenou čiarou vrátane priestoru, ktorý táto lomená čiara ohraničuje [58, s. 55]. Veľkosť a tvar polygónu sú v použitých dátach väčšinou volené na základe územného celku ktorý reprezentuje.

Susedstvá

Tento dataset [101] obsahuje údaje o susedstvách (angl. neighbourhoods) v Holandsku. Susedstvo je definované ako časť územia, ktorá je z rozvojového hľadiska alebo socio-ekonomickej štruktúry homogénne definovaná. Homogénne znamená, že jedna funkcia je dominantná, napr. rezidenčná funkcia, priemyselná funkcia alebo rekreačná funkcia, no môže existovať aj kombinácia týchto funkcií. Dataset Susedstvá je tvorený a udržiavaný Holandským štatistickým úradom a popisuje hlavne populačné dáta, firmy, sociálnu úroveň obyvateľstva, počet registrovaných áut a pod. Podrobný opis atribútov sa nachádza v Prílohe ***1***, Tabuľke 15, pričom skratky atribútov tohto datasetu začínajú písmenom N.

Atlas energie

Je dataset [39] v rozlíšení susedstiev a obsahuje informácie o ročnej spotrebe elektrickej energie a plynu domácností a firiem (na úrovni budov) spolu s počtom odberných miest. Podrobný opis atribútov sa nachádza v Prílohe ***1***, Tabuľke 17, pričom skratky atribútov tohto datasetu začínajú písmenom E.

Kvalita života

Kvalita života [74] (angl. liveability) je dataset odhadujúci kvalitu života v Holandsku na úrovni susedstiev, kombinujúci 100 faktorov do piatich dimenzií: bývanie, socio-ekonomické pozadie obyvateľstva, služby, bezpečnosť a fyzické prostredie. Podrobný opis atribútov sa nachádza v Prílohe ***1***, Tabuľke 18, pričom skratky atribútov tohto datasetu začínajú písmenom L.

Populačné jadrá

Populačné jadrá [102] (angl. Population Cores) sú súvislé územné celky alebo aj obce, v ktorých sa nachádza aspoň 50 domov alebo 25 obyvateľov. Tieto územné celky reprezentujú celistvé mestá a obce. Obsahujú podrobné dáta o obyvateľoch, ich zamestnaní, domácnostiach a nehnuteľnostiach. Podrobný opis atribútov sa nachádza v Prílohe ***1***, Tabuľke 14, pričom skratky atribútov tohto datasetu začínajú písmenami PC.

Využitie pôdy

Dataset využitie pôdy [100] (angl. Land Use) popisuje využitie územia v Holandsku pomocou polygónov. Každý polygón obsahuje hodnotu kategórie využitia pôdy. Patrí sem napr. dopravná infraštruktúra, budovy, rekreačné územia a priemyselné zóny. Z polygónových dát má tento dataset najdetailnejšie priestorové rozlíšenie. Podrobný opis atribútov sa nachádza v Prílohe ***1***, Tabuľke 16, pričom skratky atribútov tohto datasetu začínajú písmenami LC.

1.7 Rastrové dáta

V práci budeme využívať pravidelné rastrové dáta, ktoré majú presne definovaný tvar buniek.

LandScan

Využívame dataset LandScan, populačný raster odhadujúci 24-hodinový priemer počtu populácie, nazývaný aj ambientná populácia (angl. ambient population), s rozlíšením 1 $km \times 1\ km$ na území rovníka a približne 800 $m \times 800\ m$ na území Holandska. V porovnaní s datasetmi Populačné jadrá a Susedstvá zachytávajúmi rezidencnú populáciu, Landscan zachytáva práve mobilitu obyvateľov.

1.8 Prehľad dostupnej literatúry v oblasti elektromobility a súvisiacich oblastiach so zameraním na analýzu dát

S rastom počtu EV na cestách narastá aj počet štúdií používajúcich reálne dáta pochádzajúce z EV a nabíjacej infraštruktúry. V [86] môžeme nájsť komplexný prehľad literatúry venujúcej sa dátovo orientovaným štúdiám v doméne nabíjania EV. Na základe tohoto prehľadu autori uzavreli, že vo výskumnej oblasti EV a analýzy dát, chýbajú najmä vhodné dáta z oblasti EV a metodológia pre umiestňovanie nabíjacích staníc.

Mnoho štúdií skúma a predikuje vplyv EV na operáciu elektrickej distribučnej sústavy [86]. Kým niektoré vyzdvihujú pozitívny vplyv na redukciu špičkovej spotreby elektrickej energie, ostatné poukazujú na potenciálne negatívne efekty, ako napr. stupeň variability nabíjania ukázal, že je dosť vysoký na to, aby limitoval integráciu EV nabíjania a obnoviteľných zdrojov energie.

Prvotné štúdie skúmajúce EV využívali najmä dáta mobility obyvateľstva zozbierané pomocou bežných vozidiel alebo len dáta pochádzajúce od malého počtu EV. V [22, 12] použili odhad mobility národného dopravného prieskumu na odhad dopytu parkovaní, pre vylepšenie umiestnenia nabíjacích staníc. Zároveň sa objavovali aj štúdie, využívajúce dáta z menšieho počtu EV, ako napr. austrálsky test EV [98] alebo jazdy EV v [9], na základe ktorých vytvorili model pre simuláciu správania väčšieho počtu EV. V [40] analyzovali dáta z jázd BEV, pričom zistili v zime o 34 % vyššiu spotrebu a nižší dojazd ako uvádzaný výrobcom. V [61] simulovali dopyt po nabíjaní v čase a priestore na základe dopravných dát a GIS dát v Austrálií. Jedným z výsledkov boli aj špičky v nabíjaní okolo 8:30 a okolo 18:00 cez pracovné dni. Mnoho z týchto štúdií poukazuje na dôležitosť používať reálne dáta z EV, hlavne z dôvodov odlišností pri odhade z dát bežných vozidiel, akou je napríklad doba parkovania alebo využívanie primárne na jazdy v meste [128].

1.8.1 Analýza nabíjacieho správania sa používateľov EV

Množstvo dátovo orientovaných štúdií sa venuje analýze nabíjania EV. V [30] využili dáta z jázd BEV v kombinácii s dotazníkmi na skúmanie vplyvu socio-ekonomických faktorov na nabíjacie vzory a rozhodovanie používateľov BEV o nabíjaní. V [128] analyzovali dáta z skúšobných jázd flotily EV, za účelom získania vzorov využívania EV popisujúcich správanie ich používateľov. Venovali sa najmä analýzam časov začiatku a konca nabíjania, prejdenej vzdialenosti, a nabitej energii. V [62] analyzovali jazdy EV, ktoré dvomi spôsobmi zhlukovali, s cieľom poskytnúť informácie o parametroch jázd najmä výrobcom EV. Shepero a kol. [94] analyzovali nabíjacie transakcie vo Švédsku. Porovnali ich s generovanými profilmi na základe mobility. Modely generujúce profily lepšie reprezentovali často navštevované stanice oproti menej často navštevovaným. Na základe jázd 700 EV, bodov záujmu v okolí nabíjacích staníc a cien energie autori v [137] analyzovali pomocou logistickej regresie faktory vplývajúce na to, či používatelia vozidlá nabíjajú doma, v práci alebo na verejných nabíjacích staniciach, t.j. na voľbu lokácie (typu) nabíjania. Používatelia PHEV volili častejšie domáce a pracovné nabíjanie oproti verejným nabíjacím staniciam. Medzi hlavné faktory voľby nabíjania na danom type nabíjacej stanice patrili cena nabíjania, úroveň nabitia batérie pri pripojení vozidla, čas pobytu na mieste kvôli ktorému volili nabíjaciu stanicu, nabíjací výkon a hustota verejnej nabíjacej infraštruktúry. Tieto štúdie ukazujú prvotné charakteristiky EV a nabíjacej infraštruktúry najmä z

pohľadu správania, vplývajúce faktory.

1.8.2 Analýza dopytu EV po elektrickej energii

Najmä v minulosti z dôvodu nedostatku dát a obáv o preťaženie siete nabíjaním EV, boli frekventované štúdie odhadujúce dopyt EV po elektrickej energii. Model na predikovanie energie spotrebovanej nabíjacími stanicami na základe reálnych dát pomocou strojového učenia, predstavili autori v [24]. Dátovo riadený model na predikciu energeticky efektívnej trasy pre EV s využitím strojového učenia bol vytvorený v [32]. Pomocou neurónových sietí predikovali rýchlostný profil vozidla na základe meteorologických, cestných a dopravných dát. Lineárna regresia potom na základe ostatných parametrov a výstupu z neurónovej siete predikuje spotrebu energie na cestných úsekoch. Výsledky tejto štúdie odporučili aj ako vstup pre ďalšie algoritmy. V [135] predikovali dopyt EV po energii s 30 minútovou frekvenciou pomocou support vector machines a Monte Carlo, pričom support vector machines dokázala presnejšie predikovať dopyt, odhadnutý na základe dopravných dát. V [83] odhadli dopyt po nabíjaní na verejných miestach na základe času stráveného pri POI, ochoty kráčať k POI a ďalších faktorov. V [95] odhadovali dopyt po nabíjaní na parkovacích miestach na základe typov blízkych budov a ochoty kráčať. Nabíjacie stanice rozdelili do troch kategórií – domáce, pracovné a iné, pričom merali aj vrcholový výkon nabíjacej infraštruktúry, ktorý bol závislý od nabíjacieho prúdu a od časovej frekvencie pozorovaní. Odhad mobility na základe GPS dát 1.6 milióna ľudí, s využitím faktorizácie matíc pre klasifikáciu druhu využitej dopravy autori vykonali v [138]. Následne určili potenciál na adopciu EV na základe dotazníkov ale aj faktorov ako napr. odhadnutá spotreba elektrickej energie EV na jednotku vzdialenosti. Dátovo orientovaná štúdia z oblasti energetiky s väčším množstvom skutočných dát je v [124], sústreďuje sa na časopriestorovú analýzu spotreby elektrickej energie urbánnych zón. Autori analyzovali denné profily spotreby elektrickej energie v rôznych typoch zón v Holandsku. Klasifikovali spotrebu do na tri typy: rezidenčná, podniková a zmiešaná a analyzovali 3 priestorové úrovne: susedstvá, okrsky a mestské samosprávy. V závere pozdvihujú dôležitosť takých štúdií pre odhad urbánneho dopytu.

Vyššie spomenuté štúdie využívajú rôzne zdroje dát so zameraním na odhad dopytu EV po energii. Väčšinou sa jedná o jednoduchšie aplikácie metód strojového učenia využívajúce práve dáta ktoré by mohli vplývať na skutočnú spotrebu EV, z ktorých je možné

prevziať metodológiu. Pre nás sú taktiež prínosné použité dáta, ako napríklad POI, či dopravné dáta, prípadne aj použité stratifikácie urbánnych území.

Publikácie skúmajúce vzory nabíjania používateľov EV pre odhad vplyvu EV na elektrickú sieť

V [76] pomocou pravdepodobnostnej metódy kombinovali dva datasety – profily nabíjania EV a rezidenčný dopyt zo smart meračov, na odhad vplyvu nárastu EV na distribučnú sieť. Porovnávali vidiecke a urbánne územie, pričom distribučná sieť v urbánnych oblastiach dokázala zniesť vyššiu penetráciu EV ako distribučné siete vo vidieckych oblastiach. V urbánnych oblastiach dominovalo verejné nabíjanie, vo vidieckych naopak nabíjanie doma. Fischer a kol. [41] analyzovali vplyv EV pomocou parkovacích lokácií ako aj potenciál na stratégie riadenia záťaže elektrickej siete, z pohľadu širšej škály socio-ekonomických, behaviorálnych a priestorových faktorov. Jazdy EV simulujú pomocou datasetu mapujúceho mobilitu obyvateľstva. Medzi hlavné faktory, ktoré mali vplyv na nabíjacie správanie patria zamestnanie používateľov EV, pričom druh domácnosti a ekonomický status mali vplyv na ich nabíjacie vzory. Taktiež zistili, že výkon nabíjacích staníc značne vplýva na energetické špičky, najmä ak má domácnosť viac áut. Autori v [122] vytvorili dátovo-orientovanú metodológiu na identifikáciu nabíjania EV v domácnostiach, ktorá má využitie v zlepšení riadenia distribúcie elektrickej energie. Táto metodológia sa skladá z niekoľkých fáz. Prvou je príprava dát, ktorá zahŕňa čistenie dát (usporadúvanie časových radov, formátovanie dát, nahradenie chýbajúcich hodnôt). Nasleduje identifikácia metrík na izoláciu nabíjacích vzorov (energetická obálka a prahovanie hodnôt pre lepšiu klasifikáciu) a následne identifikácia parametrov metód. Potom pomocou metódy najbližších susedov, algoritmu náhodného lesa, klasifikačných a regresných stromov a CHAID algoritmu pokúšali určiť, či sa v domácnostiach nabíja EV, pričom dosiahli presnosť vyše 80 %. Publikácia [136] je jednou z prvých publikácií využívajúcich reálne dáta o nabíjaní z EV. Autori najskôr popisujú metodológiu na spracovanie a čistenie dát. Na získanie informácií o nabíjaní a jeho vplyve na elektrickú sieť používajú model založený na dátovej analýze. Ten extrahuje pomocou zhlukovania reprezentatívne profily denných dopytov, na základe korelácie určí vplyv atribútu počasia a analyzuje aj rastový koeficient dopytu EV pre danú oblasť. Tieto tri vstupy sú potom použité vo fuzzy modeli, vyhodnocujúcom riziko vplyvu EV na distribučné siete v jednotlivých oblastiach.

V tejto oblasti sa taktiež nachádza viacero publikácií využívajúcich pokročilejšie dátové

analýzy, kde možno brať inšpirácie pre časové rady ako aj zhlukovú analýzu. Vidíme tu využívanie socio-ekonomických dát, ako aj rôzne prístupy strojového učenia.

Analýza spotreby elektrickej energie vyvolanej nabíjaním EV pomocou časových radov

V literatúre môžeme nájsť aj niekoľko štúdií zameraných na analýzu elektrickej energie spotrebovanej nabíjacou infraštruktúrou, ktoré využívajú časové rady. Ďalšie užitočné štúdie, z ktorých sa možno inšpirovať môžeme nájsť aj pre podobné aplikáciách ako napr. analýza spotreby elektrickej energie v budovách. Predikovanie spotreby elektrickej energie supermarketu pomocou lineárnej regresie na základe predošlých členov časového radu a exogénnych dát vykonali autori v [10]. V [29] porovnávajú niekoľko modelov na predikovanie spotreby elektrickej energie budovy s hodinovou frekvenciou, využívajúc exogénne premenné. Porovnali základný autoregresný model s metódou najmenších štvorcov a support sector regression, pričom najlepší výsledok dosiahol autoregresný model.

Predikovanie denných profilov nabíjania EV je jeden zo spôsobov predpovedania spotreby elektrickej energie EV podobný časovým radom. Takto v [3] predpovedajú spotrebu nabíjacej infraštruktúry aj na základe rozsiahlych dopravných a meteorologických dát. Denné profily dopravy zoskupili pomocou zhlukovania do 4 skupín pre autá a 3 skupín pre autobusy. Následne pomocou šedej relačnej analýzy získali vplyvné faktory z počasia. Na základe týchto dát potom predikovali denné profily, ktoré sa líšili najmä pre dni v týždni, ročné obdobie a podľa toho či sa jedná o komerčné alebo rezidenčné nabíjanie.

Jednou z prvých publikácií zameraných na predikovanie spotreby energie a obsadenosti nabíjacích staníc v čase je [6]. Autori predikovali, či bude nabíjací bod na stanici obsadený pomocou logistickej regresie. Ako externé prediktory využili oneskorené dáta z daného ale aj iných nabíjacích bodov v okolí. Na konci navrhli predpovedať spotrebu agregovane, keďže správanie na jednom nabíjacom bod je značne náhodné.

Autori v [1] predikovali dopyt EV po elektrickej energii a aj celkovú spotrebu elektrickej energie na deň vopred pomocou ARIMA modelov pre menšie územie s hodinovou frekvenciou, za účelom zníženia prevádzkových nákladov. Na základe historických dĺžok jázd a parametrov ako napr. kapacita batérie a nabíjací výkon odhadujú dopyt EV po nabíjaní, pričom tieto odhady boli príliš optimistické a nevykazovali také fluktuácie ako naše dáta, prípadne dáta v [66]. Následne porovnali model predikujúci obe spotreby oddelene a spoločne, pričom lepší bol model s oddeleným predikciami, aj z dôvodu, že bežná

spotreba má odlišné sezónne vzory ako spotreba EV. V [71] autori vyhodnocujú ktorým zo 4 algoritmov strojového učenia (time weighted dot product based nearest neighbours, support vector regression, RF, modified pattern sequence forecasting) sú získané najlepšie predikcie spotreby elektrickej energie EV na základe dát zo stanice alebo dát z EV. Porovnanie vykonali z hľadiska bezpečnosti osobných údajov používateľov EV a výpočtovej rýchlosti. Najlepšie výsledky dosiahli algoritmy time weighted dot product based nearest neighbours a modified pattern sequence forecasting. Pričom nebol zistený rozdiel v presnosti v závislosti od toho ktoré dáta použili, dáta zo staníc považovali za bezpečnejšie z pohľadu osobných údajov, avšak dáta z EV sú rýchlejšie na spracovanie. Autor v [66] predikuje časopriestorovo agregovanú hodinovú dennú spotrebu energie EV pomocou sezónnych ARIMA modelov. Identifikoval významné rozdiely v profile spotreby medzi víkendami a pracovnými dňami. Dlhodobú spotrebu považuje za ťažšie predpovedateľnú a spojenú s nestacionaritou, kvôli ktorej odporúča pretrénovávať ARIMA modely s časom.

Modely sú v oblasti analyzujúcej spotrebu elektrickej energie budov oproti modelom predikujúcim spotrebu nabíjacej infraštruktúry značne rozvinutejšie a je možné z nich prevziať metodológiu. Existujú aj rozsiahle publikácie o predikovaní časových radov spotreby elektrickej energie energie ako napríklad [31], ktorá bola jednou z inšpirácií v našej práci. Za dva hlavné dôvody prečo nie je dostatok publikácií predikujúcich časové rady spotreby elektrickej energie spotrebovanej nabíjacou infraštruktúrou vidíme hlavne nedostatok dát ako aj vyššiu náhodnosť systému, napr. oproti relatívne stabilnej spotrebe elektrickej energie budov. V prípade, že sa na niekoľkých nabíjacích staniciach nabíja niekoľko PHEV výkonom 3 kW, a príde tam elektrické vozidlo s kapacitou batérie 80 kWh a maximálnym možným výkonom nabíjania 11 kW, môže to spôsobiť značné vychýlenie oproti bežnej spotrebe.

Predikovanie spotreby elektrickej energie budov

Ako sme už spomenuli vyššie, podobným sektorom kde je dopyt po elektrickej energii pomocou dátových analýz skúmaných vo väčšej miere, je napríklad predikovanie spotreby elektrickej energie budov. V [142] nájdeme prehľad článkov, kde sa autori zaoberajú predikciou elektrickej energie v budovách. Prehľadová publikácia [108] sa tiež venuje predikčným modelom elektrickej energie v budovách, pričom rozoberajú aj riešenia využívajúce metódy strojového učenia. V [115] autori predikovali spotrebu elektrickej energie budovy na základe metód strojového učenia s využitím exogénnych prediktorov, ako boli elektrické

spotrebiče ale aj faktory domácnosti. Porovnali tu krokovú regresiu, rozhodovací strom a neurónové siete, pričom metódy mali podobné výsledky a o málo lepší bol rozhodovací strom. Autor v [53] predikuje spotrebu elektrickej energie budov pomocou regularizovaných metód na základe GIS dát. Použil metódy: hrebeňová regresia, lasso, elastic net a hierarchic group lasso, ktoré dokáže identifikovať interakčné efekty. Na túto štúdiu nadstavujú autori v [70] a podrobnejšie popisujú metodológiu na výber premenných. Použili tu viacero metód na výber premenných, pričom najlepšie výsledky dosiahla metóda support vector regression v kombinácií s elastic net. Z posledných článkov, vidíme, že metódy na výber premenných sú už využívané v energetickom sektore na identifikáciu vplyvných faktorov. Avšak iba v [53] autor poukazuje na fakt, že by sa mala testovať významnosť regresných koeficientov v takýchto modeloch. Ak sa netestuje, môže dôjsť k prezentovaniu prediktorov, ktoré boli stochasticky vybraté a nie sú previazané na výstupnú premennú.

1.8.3 Analýzy dát pochádzajúcich z nabíjacích staníc

Publikácie využívajúce dátovú analýzu, sa často venujú štatistickému modelovaniu nabíjacích transakcií a odhadujú pravdepodobnostné rozdelenia premenných reprezentujúcich charakteristiky nabíjacej infraštruktúry, ako napríklad [34] a [60]. V tejto časti si popíšeme hlavne deskriptívne analýzy dát z nabíjacích staníc.

V [141] vyextrahovali nabíjací profil EV zo spotreby domácností pomocou NILE algoritmu. Štatisticky analyzovali nabíjacie profily, pozorujúc rozdiely medzi víkendmi a pracovnými dňami a kvantifikovali flexibilitu nabíjania EV. Autor v [60] analyzoval denné profily nabíjacích staníc v dvoch štátoch v USA. Najskôr analyzoval súvislosti medzi nabíjacími profilmi v San Diegu s cenami elektrickej energie, pričom sa viac nabíjalo v noci, kedy bol prúd najlacnejší. Následne boli analyzované štatisticky charakterizujúce jednotlivé časti denných profilov, pričom pre každú hodinu bolo odhadnuté pravdepodobnostné rozdelenie. Pre sviatočné dni boli vytvorené zvláštne modely, keďže sa profily odlišovali od bežných dní. V [23] analyzovali začiatok pripojenia, čas pripojenia a nabitú energiu a odhadli ich pravdepodobnostné rozdelenia pomocou kernel funkcii. Analyzovali aj korelácie a závislosti medzi týmito charakteristikami, pričom našli zápornú koreláciu medzi začiatkom a časom pripojenia, so silnou závislosťou v nízkych hodnotách (dolnom chvoste). Analýza priestorových vzorov tvorených spotrebou energie na nabíjacích staniciach bola prezentovaná v [68]. Autori v [42] analyzovali EVnetNL dataset a odhadli pravdepodob-

nostné rozdelenia základných identifikátorov nabíjacích staníc, ktoré závisia na správaní používateľov EV. Rozdelenie príchodov úspešne modelovali kombináciou beta rozdelení. Čas nečinnosti, nabíjania a pripojenia modelovali aj v závislosti od najbližšieho cestného úseku. Jedno z podstatných zistení bolo, že v akomkoľvek čase je 75 % staníc obsadených nabitým EV.

Deskriptívne analýzy okrem rozdelení pravdepodobností charakteristík nabíjacej infraštruktúry odhadujú aj ich vzájomný vplyv, čo pomáha lepšie pochopiť vlastnosti a vplyv týchto charakteristík.

Analýza zhlukov dát nabíjacej infraštruktúry

Budovanie modelov strojového učenia za pomoci zhlukovania zvyšuje presnosť modelov [19]. Napríklad, predpovedanie spotreby elektrickej energie určitej skupiny nabíjacích staníc s podobnými vzormi využitia, môžeme lepšie predikovať ako keď uvážime všetky nabíjacie stanice naraz. V [133] bolo zhlukovanie aplikované na zachytenie neistoty v správaní používateľov EV, a na vytvorenie hranice na predpovede energie deň vopred. Pomocou k-means algoritmu, zhlukovali denné profily nabíjania EV v [136]. Najreprezentatívnejšie ťažisko zhluku bolo využité ako vstup do rizikového modelu. V [118] autori popísali nabíjacie správanie na základe nabíjacích dát z Amsterdamu. Charakterizovali a analyzovali šesť typov používateľov EV identifikovaných z dát. Ďalší prístup aplikovaný na dáta zozbierané v Holandsku je [90]. Správanie používateľov EV je modelované časmi príchodov a odchodov z nabíjacej stanice. Takisto zhlukovaním sú analyzované aj rozdiely spôsobené víkendmi a zmenami ročných období.

1.8.4 Plánovanie a umiestňovanie nabíjacej infraštruktúry pre EV

Ďalšou rozsiahlou kategóriou publikácií je umiestňovanie nabíjacích staníc a plánovanie nabíjacej infraštruktúry EV. Aj tu sa využívajú rôzne dáta na odhad jazdných vzorov EV, ako napríklad OD matice [132], alebo dáta z obyčajných vozidiel [2, 35]. Prehľadová publikácia [50] skúmajúca preferencie používateľov EV v súvislosti s nabíjacou infraštruktúrou. Zameriava sa najmä na dôležitosť nabíjacej infraštruktúry troch hlavných kategórií (domáca, pracovná a verejná), prístup k infraštruktúre a cenu nabíjania ako aj koľko staníc je potrebných na uvedenie EV. Medzi najdôležitejšie patrí domáce nabíjanie, nasledované pracovným a verejným nabíjaním, avšak aj verejné nabíjanie považujú za dôležité hlavne pri začiatkoch rozvoja používania EV. Tu však treba pripomenúť, že domáce nabíjanie je

problémové najmä v husto zaľudnených urbánnych oblastiach, kde je potrebná verejná nabíjacia infraštruktúra. Takisto plánujú skôr budovať rýchle nabíjacie stanice. Výsledky publikácie sú trochu odlišné od scenárov v [17], ktoré predpokladajú skôr pomalú verejnú nabíjaciu infraštruktúru. Aj keď prehľadová publikácia [50] sumarizuje štúdie používajúce skutočné dáta, väčšinou používajú dáta z datasetov s násobne menším počtom transakcií a nabíjacích staníc ako dáta, ktoré používame my. Prehľad využitia dát v publikáciách týkajúcich sa umiestňovania staníc EV, je možné nájsť v aj kapitole 3.1 v [84].

V [45] umiestňujú nabíjacie stanice na základe odhadu dopyt BEV po nabíjaní v urbánnych oblastiach za pomoci populačných dát. Optimálne umiestňovanie pomalej nabíjacej infraštruktúry s odhadom dopytu po nabíjaní pomocou regresie a GIS dát nájdeme v [43]. Autori tu najskôr odhadujú nočný dopyt po nabíjaní pomocou odhadu počtu áut a jednoduchých populačných charakteristík. Denný dopyt po nabíjaní odhadujú na základe miery zamestnanosti, odhadovanej typom budov, a nabíjanie v blízkosti pracoviska považujú za hlavný zdroj denného dopytu. Pre oba dopyty preratúvajú počet EV z celkového počtu vozidiel, pomocou odhadovanej penetrácie EV. Z toho počtu EV následne odhadujú dopyt. Potom vo vybranej časti mesta umiestňujú pomocou prístupu maximálneho pokrytia nabíjacie stanice, využívajúc odhad dopytu. Na základe tejto považujeme za zaujímavé vyskúšať osobitne získať faktory pre urbánne časti nabíjacej infraštruktúry.

V [47] autori hľadali optimálne umiestnenie staníc, pričom sa snažili odhadovať dopyt po nabíjaní na základe dát z okolitých staníc, hustoty ciest a bodov záujmu v okolí. V [110] na základe dát z BEV rozširovali nabíjaciu infraštruktúru v meste Wuhan pomocou genetického algoritmu, a analyzovali výskyt nízkeho stavu nabitia batérie (pod 20 %) pre rôzne druhy finančnej podpory na budovanie staníc. Autori v [125] identifikovali model na umiestňovanie nabíjacích staníc na základe dopytu, ktorý odhadovali pomocou POI kategórií a ochoty pohybovať sa peši v blízkosti staníc. Návrh optimálneho umiestnenia staníc založeného na metaheuristikách, využívajúceho ekonomické faktory a faktory elektrickej siete nájdeme v [33]. Umiestňovanie staníc na základe mobility obyvateľov, s cieľom zníženia dojazdu za stanicami pomocou optimalizačných metód autori popísali v [121]. Ďalšia štúdia [140] optimalizuje umiestnenie nabíjacej infraštruktúru na základe požadovaných investícií, ceny prevádzky a pokrytia služieb, pomocou particle swarm optimization. Využívajú aj GIS dáta pre zahrnutie dát o doprave a elektrickej sieti.

V [26] pomocou dvojúrovňového modelu umiestňujú nabíjacie stanice, s prípadom pou-

žitia Maďarsku. V prvej úrovni určujú počet staníc v okrskoch, na základe odhadu dopytu po nich pomocou faktorov ako počet EV, výška príjmov a turizmus. V druhej úrovni vyberajú hexagóny s dĺžkou každej strany 250m, v ktorých bude umiestnená stanica. Dopyt po nabíjaní v hexagóne určujú ako kombináciu denného a nočného dopytu. Tieto dve zložky dopytu odhadujú pomocou vybraných miest záujmu v okolí ako aj vzorov správania používateľov EV získaných z dotazníkov. Pre denné nabíjanie sú použité vo väčšom rozsahu rýchle nabíjacie stanice a pre nočné nabíjanie sú použité pomalé nabíjacie stanice umiestnené najmä v rezidenčných oblastiach. Výber hexagónu pre umiestnenie zohľadňuje aj umiestnenie staníc v okolitých hexagónoch. Jedno z hlavných zistení je, že nabíjacie stanice na parkoviskách typu „park and ride“ v husto zastavaných územiach sú vhodnejšie pre obsluhu dopytu po nabíjaní ako čerpacie stanice. V [123] je dopyt po energii v piatich európskych mestách odhadovaný pomocou dát zo zdieľaných vozidiel. Na základe odhadu dopytu je vypočítaná ziskovosť staníc a tie sú umiestňované pomocou hexagonálnej siete. Na predikcie bola využitá metóda RF, pričom boli predikcie vylepšené aj bodmi záujmu získanými z OSM. Model je o 24 % lepší ako naivný model a autor ho odporúča využiť na podporu rozhodovania pre operátorov nabíjacích staníc.

Ako si môžeme všimnúť obe predošlé štúdie odporúčajú umiestňovanie nabíjacích do hexagónov, čo považujeme z vhodný spôsob umiestnenia pre optimalizačné úlohy, ktoré môžu byť nadstavbou na našu prácu.

V [87] identifikovali faktory vplývajúce na využívanie nabíjacích zón na základe dát zozbieraných z existujúcej nabíjacej infraštruktúry a jej okolia a zároveň vytvorili metodológiu na predikciu využitia nabíjacej infraštruktúry. Pre pochopenie vplyvu prediktorov najskôr aplikovali lineárnu regresiu a následne využili na predikcie metódu XGBoost. V práci využili 5 prediktorov: počet POI, počet konkurenčných nabíjacích staníc, ID nabíjacej zóny, počet RFID kariet a počet nabíjacích staníc v zóne. Nabíjacie zóny vytvorili ako priestorovo agregované stanice v okruhu približne 3km. Následne optimalizovali umiestnenie nových staníc na základe kompromisu medzi ich využitím a počtom. V [52] autori navrhli množinu ukazovateľov na porovnanie dvoch stratégií umiestňovania pre nabíjaciu infraštruktúru: dopytovo orientovaný rozvoj a strategické umiestňovanie nabíjacích staníc. Žiadna zo stratégií nie je dominantná vo všetkých metrikách a rozdiely medzi stratégiami sa zmenšujú, ako sa rozširuje nabíjacia infraštruktúra v čase. Logistická regresia s viacerými premennými bola využitá v [131] na určenie kľúčových faktorov, vysvetľujúcich

heterogenitu v trvaní nabíjania kategorizovaných nabíjacích udalostí. Premenné spojené s hodinou dňa a typom nabíjacej stanice majú najväčší vplyv. Niektoré prediktory ako typ urbánneho územia, hustota nabíjacích staníc a parkovacie možnosti boli zvažované tiež.

Z uvedeného prehľadu literatúry môžeme vidieť, že v posledných rokoch vznikajú dátovo orientované štúdie na umiestňovanie staníc alebo na vyhodnotenie umiestnenia staníc najmä pomocou ich okolia. Medzi používané faktory patria najmä POI v blízkosti nabíjacích staníc, ale aj urbánne indikátory.

Systémy zdieľania bicyklov

Podobnou aplikáciou, najmä z pohľadu využívania staníc, sú systémy zdieľania bicyklov (angl. bicycle sharing systems) (BSS) v mestách. V takýchto systémoch je viac dostupných dát oproti elektromobilite, z ktorých veľké množstvo tvoria voľne dostupné dáta. Analýzu dát za účelom lepšieho návrhu a expanzie BSS nájdeme napríklad v [97]. Autori tu predikovali dopyt po zdieľaných bicykloch pomocou lineárnej regresie. Využili dáta o taxíkoch, počasí a dáta o populácií, pričom využili aj transformácie prediktorov. Pri priestorovej agregácii na úroveň susedstiev dosiahli vyššiu presnosť. Štatistickú analýzu jázd bicyklov krátkych vzdialeností s cieľom vylepšenia rozhodovaní v oblastiach ako predikcia dopravy, umiestňovanie staníc a realokácie bicyklov nájdeme v [18]. Zhou v [143] analyzoval jazdy bicyklov pomocou zhlukovej analýzy pre získanie podobných profilov jázd a zamerali sa na rozdiely v mobilite medzi pohlaviami, ako aj identifikovali rôzne jazdné vzory a trendy. Počet voľných miest na BSS staniciach na základe reálnych dát vo forme časových radov pomocou ARMA modelov predikovali autori v [56].

1.8.5 Smart charging

Veľa štúdií sa v súčasnosti venuje inteligentnému nabíjaniu, t.j. smart chargingu, či už za cieľom znížiť energetické špičky, znížiť ceny nabíjania alebo efektívnejšie využiť obnoviteľných zdrojov elektrickej energie. V [65] predikujú pomocou dynamického programovania a niekoľkých metód strojového učenia čas, kedy počas pripojenia nabíjať EV. V [7] analyzovali dáta z nabíjacích staníc a nabíjacie profily a iné atribúty nabíjacích staníc. Takisto skúmali vplyv obyčajného nabíjania na elektrickú sieť a testovali aj oneskorené nabíjanie pomocou centralizovaných a decentralizovaných stratégií pre redukciu energetických špičiek. V [133] vytvorili zhluky používateľov, ktorým na základe neurónovej siete vedia priradiť nových používateľov EV, pre využitie v smart chargingu. Schopnosť metód

strojového učenia XGBoost, RF a Gradient Boosting na predikciu času nečinnosti EV, ktorý má široké využitie v smart chargingu, bola vyhodnotená v [67]. Ako najdôležitejšie prediktory pre predikcie vyšli čas dňa, nabitá energia, čas nabíjania a maximálny výkon nabíjacej stanice, pre ktoré evaluovali aj polynomiálny tvar vplyvu na čas nečinnosti. Najlepšie výsledky dosiahla metóda XGBoost, s $R^2 = 0.63$. V [57] analyzovali nabíjacie transakcie a vplyv územne agregovaných staníc na elektrickú sieť. Analyzovali potenciál smart chargingu z hľadiska cien ako aj vyrovnávania energetickej špičky, pričom dokázali špičku v jednotlivých mesiacoch znížiť na 30 až 42% ale ze predpokladu, že časy ukončenia transakcie sú vopred známe. Rozsiahla štúdia [134] kombinujúca viacero druhov dát a to konkrétne dáta z mobilov obyvateľov zachytávajúcich mobilitu, nabíjacie transakcie z nabíjacích staníc mimo obytných oblastí a prieskumy týkajúce sa využívania konvenčných a elektrických vozidiel. Dáta z nabíjacích staníc analyzovali a odvodili vzory nabíjania. Za pomoci týchto dát vylepšili simulácie EV vytvorené pomocou dát z mobility a dotazníkov. Cieľom simulácie bolo znížiť energetickú špičku, pomocou plánovanie časov príchodov a odchodov EV k nabíjacím staniciam. V [90] analyzovali pomocou EVnetNL datasetu flexibilitu nabíjania EV, ktorú definovali ako maximálnu záťaž, čo možno posunúť na isté trvanie kedykoľvek počas dňa. Najskôr identifikovali 3 zhluky transakcií na základe časov príchodov a odchodov. Pre zhluky potom analyzovali rozdiely v časoch príchodov, odchodov a nečinnosti pre ročné obdobia a víkendy.

Ako môžeme vidieť, existuje viacero uplatnení dátovej analýzy v elektromobilite, no stále je málo komplexných dátovo orientovaných štúdií, čo prisudzujeme najmä nedostatku dát v tejto oblasti. Takisto sa dá čerpať z podobných aplikácií, medzi ktorými sme identifikovali najmä analýzu spotreby elektrickej energie v budovách a bike-sharing.

2 Ciele dizertačnej práce

Dáta môžu pomôcť pri identifikácii správnych rozhodnutí, avšak na to ich potrebujeme vhodne spracovať a vyhodnotiť. Na tento účel slúžia rôzne nástroje a metódy, ktorých je v dnešnej dobe veľké množstvo. Tieto nástroje a metódy je potrebné vedieť správne používať, prípadne rozšíriť existujúce metódy alebo vytvoriť vhodné kombinácie metód pre dosiahnutie čo najpriaznivejších výsledkov. Na základe rokovaní s expertmi zo spoločnosti ElaadNL [38], holandského znalostného a inovačného centra v oblasti nabíjacej infraštruktúry a spoločnosti Greenway [48], ktorá sa venuje budovaniu a prevádzke nabíjacích staníc na Slovensku a okolitých krajinách, ako aj prehľadu dostupnej literatúry a prác v oblasti elektromobility, sme formulovali tri ciele pre dizertačnú prácu. Hlavným cieľom je vytvoriť metodológiu pre analýzu dát a dátové modelovanie v prostredí elektromobility, ktorá bude mať potenciál zväčšiť množinu dostupných nástrojov pre podporu rozhodovania v oblasti budovania a prevádzky nabíjacej infraštruktúry pre EV, pričom sa budeme sústrediť na tri, nižšie uvedené ciele.

Cieľ 1: Vytvorenie segmentov nabíjacích staníc a zákazníkov

Na začiatku má o elektromobilitu záujem iba určitá skupina ľudí – tzv. skorí inovátori a postupne vznikajú nové kategórie zákazníkov, ktoré je potrebné identifikovať, sledovať a reagovať ponukou služieb na ich dopyt. Podobne nabíjacie stanice sú využívané rôznym spôsobom, nočné nabíjanie v blízkosti domova, denné nabíjanie v blízkosti práce a pod. A teda segmenty zákazníkov príp. nabíjacích staníc sú charakterizované tým, že vykazujú iný vzorec nabíjania elektrických vozidiel.

Našou úlohou je na základe vizuálnych analýz dát a zhlukovacích algoritmov identifikovať zhluky a porovnať ich s literatúrou. Výstupom má byť jednak lepšie pochopenie nabíjacieho správania a takisto nájdené segmenty môžu byť využité na stratifikáciu dát v ďalších analýzach, ako napr. predikcia energie alebo predikcia využívania určitej skupiny staníc.

Cieľ 2: Predikovanie spotreby elektrickej energie na nabíjacích staniciach

Z hľadiska prevádzkovateľa staníc, ale aj prevádzkovateľa elektrickej siete, je potrebné predikovať budúci dopyt po elektrickej energii, či už z krátkodobého alebo aj dlhodobého hľadiska. Tu sa jedná najmä o vytvorenie podpory pre operatívne rozhodnutia, ako sú

napríklad nákup elektrickej energie pre budúce obdobie alebo aj predpoveď spotreby na ďalší deň pre potreby smart chargingu. Otestujeme možnosti vytvoriť krátkodobé a dlhodobé predpovede spotreby elektrickej energie na nabíjacích staniciach a uvážime rôznu mieru agregácie staníc.

Spotrebu chceme predpovedať pomocou odhadových metód určených pre časové rady a metód strojového učenia. Modely sa pokúsime vylepšiť pomocou rôznych externých premenných, ako napríklad počasie v okolí stanice. Okrem priestorovej agregácie, vyskúšame aj agregáciu staníc na základe nabíjacích vlastností, čo by mohlo zlepšiť presnosť predpovedí. Výsledkom môžu byť okrem vhodnej metodiky na predpovedanie agregovanej spotreby staníc aj informácie o variabilite spotreby a miery, do akej vieme túto spotrebu predpovedať.

Cieľ 3: Identifikácia ukazovateľov vhodného umiestenia nabíjacích staníc

Jedným zo základných prvkov v elektromobilite sú nabíjacie stanice. Ich umiestňovanie je súčasťou strategických rozhodnutí, ktoré sú spojené s vyššími finančnými nákladmi a dlhodobou platnosťou týchto rozhodnutí. Z prehľadov literatúry vieme, že stanice sa často umiestňujú na základe empiricky podložených alebo čiastočne empiricky podložených rozhodnutí, pričom je nedostatok literatúry venujúcej sa charakteristikám vhodného umiestnenia takýchto staníc.

Miesta vhodné pre umiestnenia nabíjacích staníc a najmä ich charakteristiky chceme určovať na základe existujúcich dát o nabíjaní vozidiel, spolu so socio-ekonomickými ukazovateľmi a rôznymi inými druhmi dát. Na odhad vplyvu plánujeme využiť inferenčnú schopnosť lineárnej regresie a jej modifikácií. Z týchto modifikácií využijeme najmä tolerantné a selekčné metódy, pretože sa dá očakávať, že prediktory odvodené z dát budú často podobné a závislé, s čím sa tieto metódy vedia čiastočne vysporiadať.

Očakávame, že na základe analýz budeme vedieť identifikovať faktory, ktoré potenciálne vplývajú na nabíjanie na staniciach. Modely využívajúce tieto faktory ako prediktory by mali byť schopné určiť vhodné umiestnenia staníc a teda výsledky bude možné použiť pre rozšírenie lokačných analýz.

2.1 Metodika práce

V každej podkapitole kapitoly 4 budeme používať nasledovný postup. Najskôr uvedieme problematiku, predspracovanie a úvodnú analýzu dát. Potom popíšeme použitie metód, tvorbu modelov, popíšeme výpočtové experimenty a na záver vyhodnotíme výsledky.

V podkapitole 3.1 vykonávame zhlukovú analýzu dát so zameraním na segmenty nabíjacích staníc pomocou dvoch rôznych pohľadov na agregáciu a zhlukovanie dát.

V podkapitole 3.2 popisujeme predpovedanie časovo a priestorovo agregovanej spotreby nabíjacích staníc pomocou metód určených pre časové rady a metód založených na rozhodovacích stromoch. Na vylepšenie predpovedí využívame externé prediktory a tri rôzne procedúry na trénovanie.

V podkapitole 3.3 popisujeme metódy ako extrahovať prediktory z rôznych GIS atribútov a následne sa venujeme porovnaniu metód pre selekciu premenných.

Identifikátorom vhodného umiestnenia nabíjacej infraštruktúry sa venujeme v podkapitolách 3.4 a 3.5. Najskôr identifikujeme a definujeme ukazovatele úspešnosti nabíjacích miest. Následne predikujeme popularitu, ktorá bolo identifikovaná ako najlepšie vysvetliteľný ukazovateľ na základe iných faktorov charakterizujúcich okolie nabíjacej infraštruktúry. Následne sa venujeme analýze heterogenity energie spotrebovanej nabíjacou infraštruktúrou na základe identifikácie signifikantných faktorov okolia.

3 Vlastné prínosy

Táto kapitola je venovaná opisu navrhnutých riešení a vlastných prínosov v súlade so stanovenými cieľmi uvedenými v predošlej časti.

3.1 Analýza segmentov nabíjacích staníc a zákazníkov

Ako sme uviedli v prehľade literatúry, zhlukovanie v dostupných štúdiách slúžilo najmä na nájdenie profilov nabíjania alebo na analýzu segmentov zákazníkov a popis nabíjacieho správania. My sa pozrieme na analýzu segmentov nabíjacích staníc, ktorá je v literatúre pokrytá v podstatne menšom rozsahu ako segmentácia zákazníkov. Poznanie nabíjacích staníc s podobným správaním napomáha k lepšiemu porozumeniu využívania nabíjacej infraštruktúry. Okrem samotnej segmentácie porovnávame aj dva prístupy na získanie zhlukov (segmentov) staníc z dát vo forme transakcií, líšiacich sa najmä v agregácií. Prvý prístup sumarizuje nabíjacie transakcie a následne zhlukuje nabíjacie stanice, druhý prístup najskôr zhlukuje nabíjacie transakcie a následne sumarizuje výsledky podľa prislúchajúcich nabíjacích staníc.

3.1.1 Používané indikátory a spracovanie dát

Na základe vytvoreného prehľadu literatúry sme identifikovali tri hlavné triedy indikátorov na charakterizovanie výkonnosti nabíjacích staníc a transakcií: návštevnosť stanice, využitie stanice a časové vzory používania. Pre každú triedu sme vybrali minimálnu množinu indikátorov. Pri transakciách uvažujeme:

- Návštevnosť:

  N_{tran} - počet transakcií na stanici, kde prebehla transakcia.

- Využívanie:

  r_{tran} - relatívny čas nabíjania, t.j. pomer medzi časom nabíjania a časom pripojenia.

- Časový vzor používania:

  TS_{tran} - čas začiatku transakcie v hodinách;

  TE_{tran} - čas konca transakcie v hodinách.

Pri nabíjacích staniciach uvažujeme:

- Návštevnosť:

  N_{stat} - počet transakcií na stanici.

- Využívanie:

  r_{stat} - relatívny čas nabíjania, t.j. pomer medzi časom nabíjania a časom pripojenia vypočítaný zo všetkých transakcií, ktoré na stanici prebehli.

- Časový vzor používania:

  TS_{stat} - priemerný čas začiatku všetkých transakcií v hodinách a

  TE_{stat} - priemerný čas konca všetkých transakcií v hodinách, ktoré na nabíjacej stanici prebehli.

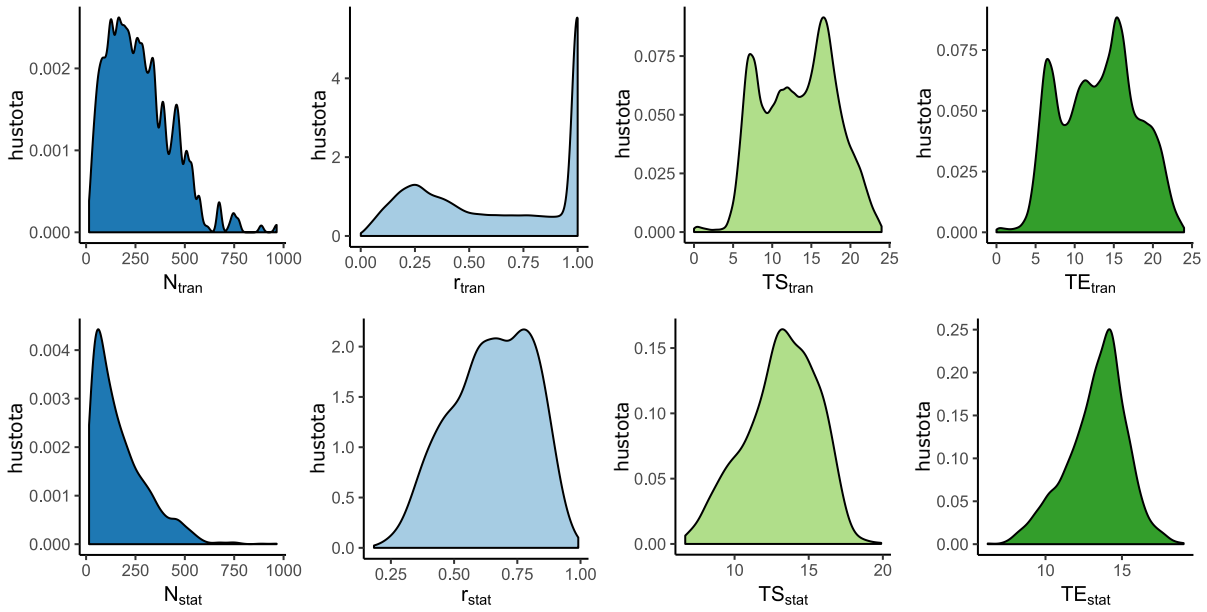
Spracovanie dát pre výpočty

Na výpočet indikátorov sme použili iba dáta medzi 1. januárom 2015 a 31. decembrom 2015, kedy bol počet transakcií pomerne stály, a obdobie 1 roka nie je tak ovplyvňované sezónnosťou, ako by to mohlo byť napr. pri pätnástich mesiacoch. Aby sme získali reprezentatívne dáta, neuvažovali sme stanice, ktoré mali menej ako 30 transakcií. Funkcie hustoty pravdepodobnosti jednotlivých indikátorov zobrazujeme na obrázku 4. Na základe rozdielov medzi spodným a vrchným obrázkom, najmä v prítomnosti multimodality, si môžeme všimnúť možné straty informácií spôsobené agregáciou transakcií do staníc najmä prí časových indikátoroch. Multimodalita na druhej strane ukazuje, že by zhlukovacie algoritmy mohli vrátiť rozlíšiteľné segmenty staníc.

Indikátory staníc a transakcií môžu nadobúdať hodnoty z rôzneho rozsahu a využívané zhlukovacie algoritmy vyžadujú škálované vstupy. Z toho dôvodu škálujeme všetky indikátory na rovnaký interval $\langle 0, 1\rangle$ pomocou vzťahu (2) na normalizáciu dát.

3.1.2 Použité metódy a tvorba modelov

Na zhlukovanie využívame metódy k-means, aglomeratívne hierarchické zhlukovanie a DBSCAN. Kvôli rôznemu počtu transakcií na staniciach vzniká otázka, ako správne reprezentovať stanice pomocou dát z transakcií. Preto navrhujeme a porovnávame dva nasledovné prístupy. V prvom prístupe získavame z transakcií sumárne štatistiky pre stanice,



Obrázok 4: Funkcie hustoty zvolených indikátorov, kde vrchný riadok reprezentuje hustoty indikátorov transakcií, spodný hustoty indikátorov nabíjacích staníc.

ktoré zhlukujeme. V druhom najskôr zhlukujeme transakcie a potom vzniknuté zhluky priradíme k staniciam.

Agregácia prvá - zhlukovanie druhé

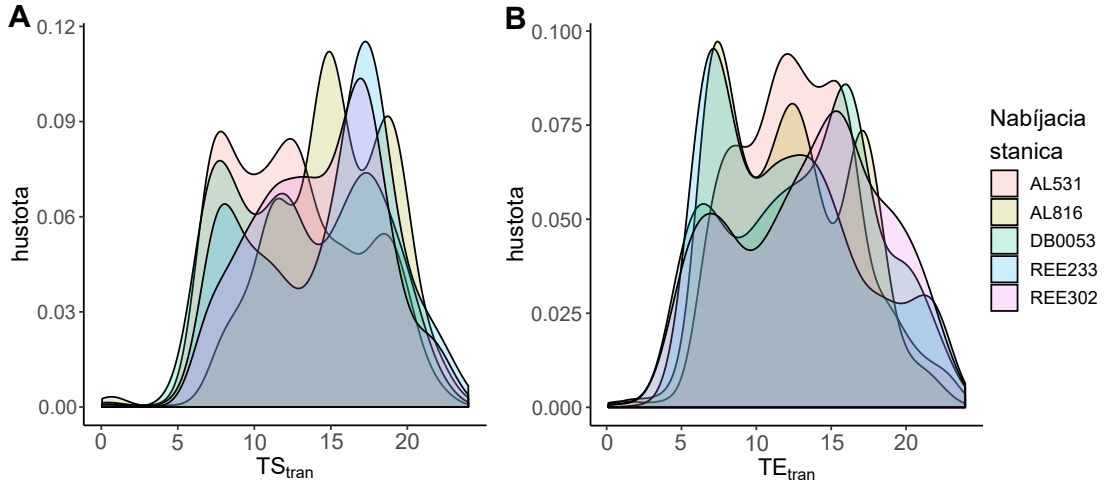
Tento prístup najskôr sumarizuje indikátory transakcií pre stanice, spočítaním priemerov indikátorov transakcií, ktoré prebehli na stanici. V ďalšom kroku aplikuje zhlukovací algoritmus na takto sumarizované dáta staníc. Nevýhodou tohto prístupu je, že priemerná hodnota nie vždy dobre reprezentuje všetky transakcie stanice, obzvlášť ak ich hustota je multimodálna (obrázok 5).

Zhlukovanie prvé - kategorizácia druhá

Prístup najskôr zhlukovacími algoritmami nájdeme zhluky transakcií a potom priradí stanice k zhlukom, v ktorých majú najvyšší počet transakcií. Toto priradenie je asociované s veľkou neistotou ak je pre stanicu podobný počet transakcií v dvoch alebo viacerých zhlukoch.

3.1.3 Výsledky zhlukovania

Pre *agregáciu prvú - zhlukovanie druhé* porovnávame veľkosti zhlukov v tabuľke 4. DBSCAN algoritmus našiel dva malé zhluky a jeden veľký zhluk, pričom nechal veľa staníc nepriradených žiadnemu zo zhlukov. S nárastom parametra ϵ sú skoro všetky pozorovania



Obrázok 5: Rozdelenie hodín príchodov (**A**) a odchodov (**B**) na transakcie piatich vybraných často navštevovaných staniciach.

priradené do jedného zhluku. Tieto výsledky napovedajú, že metóda nie je vhodná na takýto typ dát, vzory v dátach pravdepodobne nevyhovujú algoritmu, a tak sme ju vylúčili z ďalších analýz.

zhluk/metóda	DBSCAN	k-means	hierarchické
bez zhluku	767	0	0
zhluk 1	549	297	186
zhluk 2	36	234	99
zhluk 3	69	334	256
zhluk 4	0	556	880

Tabuľka 4: Početnosť pozorovaní v zhlukoch pre zhlukovacie metódy. Parametre metód boli nastavené tak, aby sme dostali 4 zhluky v každej metóde.

Výsledky získané prístupom agregovanie prvé - zhlukovanie druhé

Obe zostávajúce metódy, k-means a hierarchické zhlukovanie očakávajú nastavenie parametrov určujúce výsledný počet zhlukov. Na základe metódy lakťa, nastavujeme hodnotu parametra k v metóde k-means na hodnotu 4. V aglomeratívnom hierarchickom zhlukovaní bola použitá miera kompletného prepojenia [49, s. 462]. Hierarchické zhlukovanie našlo veľmi podobné zhluky ako k-means, preto v nasledujúcom texte interpretujeme iba výsledky k-means zhlukovania.

Prvý identifikovaný zhluk je charakterizovaný priemerným r_{stat}, nízkym N_{stat}, skorým TS_{stat} a neskorým TE_{stat}, ktoré pravdepodobne reprezentujú nabíjanie pri zamestnaní, kde používatelia zapoja EV ráno a odpoja poobede alebo večer. Druhý zhluk má stredné hodnoty r_{stat}, vysoké hodnoty N_{stat}, doobedné až poobedné TS_{stat} a poobedné TE_{stat},

čo môže reprezentovať populárne nabíjacie stanice umiestnené v blízkosti obchodov alebo voľnočasových aktivít. V treťom zhluku vidíme stanice charakterizované nízkym r_{stat} a N_{stat}, neskorým TS_{stat} a s TE_{stat} odpovedajúcim ranným hodinám, z čoho vyplývajú dlhé časy pripojenia. Takéto nabíjacie stanice sú pravdepodobne umiestnené blízko rezidenčných zón, kde nechávajú používatelia EV cez noc aby sa nabili. Štvrtý zhluk obsahuje krátke transakcie s TS_{stat} okolo obeda, TE_{stat} okolo poobedia, s vysokým r_{stat} a nízkym N_{stat} odpovedajúcim menej populárnym staniciam, ktoré používatelia opúšťajú po plnom nabití. Grafy hustôt jednotlivých charakteristík staníc priradených k rovnakému zhluku metódou agregácia prvá - zhlukovanie druhé sú zobrazené vo vrchnom rade obrázku 6.

Výsledky získané prístupom zhlukovanie prvé kategorizácia druhá

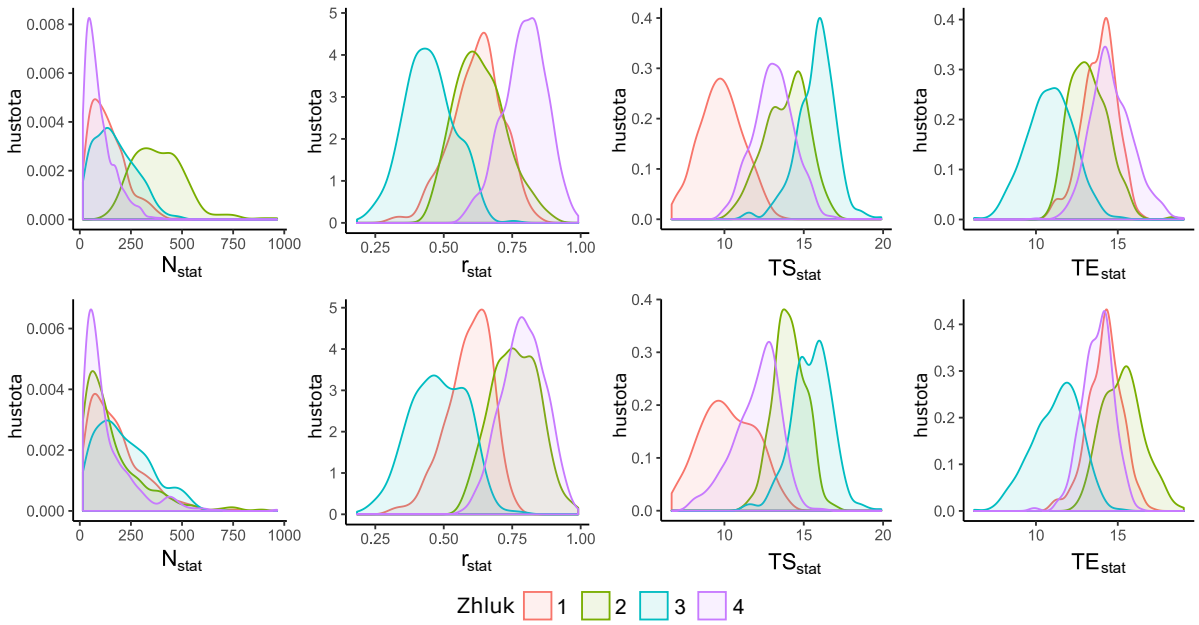
Berúc do úvahy fakt, že hierarchické zhlukovanie a algoritmus DBSCAN vyžadujú výpočet matice vzdialeností, ktorý je by bol teraz výpočtovo veľmi náročný, kvôli vysokému počtu transakcií, na zhlukovanie používame v tejto časti iba metódu k-means. Hodnota parametra k bola opäť nastavená na 4 na základe metódy lakťa.

Prvý zhluk obdržaný týmto prístupom obsahuje nabíjacie stanice, ktoré sa pravdepodobne nachádzajú blízko miesta zamestnania, kde používatelia pripájajú EV ráno a odpájajú ho poobede. Druhý zhluk tvoria stanice kde EV používatelia nabíjajú EV poobede a r_{stat} je vysoký. V treťom zhluku sú nabíjacie stanice, ktoré sú využívané najmä v noci, čo je pravdepodobne zodpovedné za nízke hodnoty r_{stat}. V poslednom zhluku sú nabíjacie stanice charakterizované nízkym r_{stat}, nabíjanie začína okolo obeda a končí poobede. Návštevnosť staníc nevykazuje výrazný vplyv na zaradenie staníc do zhlukov. Grafy hustoty indikátorov staníc pre tento prístup sú zobrazené v spodnom rade obrázku 6.

Porovnanie prístupov

Pre porovnanie prístupov sme si zvolili nasledovný spôsob a značenie. Vektor a obsahuje indexy zhlukov, do ktorých boli stanice zaradené prístupom *zhlukovanie prvé - kategorizácia druhá*. Podobne hodnoty vektora b odpovedajú indexom zhlukov do ktorých boli zaradené stanice prístupom *agregácia prvá-zhlukovanie druhé*. Výsledky získané oboma prístupmi porovnávame výpočtom kontigenčnej tabuľky 5 pomocou vzťahu

$$C_{ij} = \sum_{l=1}^{|b|} \sum_{o=1}^{|a|} q(b_l, i) q(a_o, j), \tag{35}$$



Obrázok 6: Funkcie hustoty indikátorov nabíjacích staníc priradených do rovnakého zhluku. Vrchný rad prislúcha k zhlukom nabíjacích staníc obdržaných prístupom agregácia prvá - zhlukovanie druhé. V spodnom rade sú zobrazené grafy hustoty indikátorov staníc priradených k zhlukom prístupu zhlukovanie prvé-kategorizácia druhá. V oboch prípadoch bol na identifikáciu zhlukov použitý algoritmus k-means.

kde q je funkcia vracajúca hodnotu 1 ak sú oba argumenty rovné a 0 inak. Hodnoty koeficientov C_{ij} sú uvedené v Tabuľke 5. Predpokladáme, že pozorované rozdiely v priradení staníc k zhlukom sú spôsobené multimodalitou rozdelenia identifikátora r_{char}. Zhluky získané metódou *agregácia prvá zhlukovanie druhé* sú ľahšie interpretovateľné aj vďaka hodnotám N_{trans}, ktoré sa viacej variujú oproti druhej metóde.

C_{ij}	$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	211	0	5	81
$i=2$	36	55	98	45
$i=3$	3	1	329	1
$i=4$	20	210	14	312

Tabuľka 5: Kontigenčná tabuľka pre zhlukovanie nabíjacích staníc. Zhluky získané prvým prístupom, zhlukovaním nabíjacích staníc, sú v riadkoch, a zhluky získané druhým prístupom, zhlukovaním nabíjacích transakcií, sú v stĺpcoch.

Na záver podkapitoly môžeme konštatovať, porovnanie oboch prístupov ukázalo, že obe metódy identifikovali podobné skupiny staníc. Z dvoch porovnaných metodologických prístupov, považujeme za lepší prístup *agregovanie prvé zhlukovanie druhé*, na základe jednoduchšej výpočtovej zložitosti dát a lepšej interpretovateľnosti výsledkov.

Podstatná časť tejto kapitoly vychádza z našej publikácie [105]. Na spracovanie dát

bola využitá knižnica *dplyr*. Funkcie *kmeans* a *hclust* boli použité pre algoritmy k-means a hierachické zhlukovanie. Funkcia *dbscan* z rovnomennej knižnice bola použitá ako implementácia algoritmu DBSCAN.

3.2 Predikovanie spotreby nabíjacích staníc

Táto podkapitola sa venuje predikovaniu časopriestorovo agregovaného dopytu nabíjacích staníc po elektrickej energii pomocou časových radov.

Pre prevádzkovateľov siete nabíjacích staníc, ale aj pre dodávateľov energie je výhodné mať čo najlepšie informácie o energii, ktorá bude spotrebovaná nabíjacou infraštruktúrou. Vďaka vopred známemu profilu a objemu elektrickej energie môžu prevádzkovatelia energiu vopred nakúpiť výhodnejšie, ale aj vylepšiť technológie smart chargingu. Predikovanie spotreby elektrickej energie, či už pre domácnosti alebo priemysel, je pomerne dobre prebádaná oblasť, o čom svedčí veľké množstvo literatúry pre túto oblasť. Počet prác zaoberajúcich sa dátovo orientovaným predikovaním spotreby elektrickej energie je v prostredí elektromobility ešte stále nízky, čo je pravdepodobne spôsobené najmä nízkou dostupnosťou rozsiahlejších dát z nabíjacích staníc. Z prehľadu literatúry (podkapitola 1.8) a najmä časti "Analýza spotreby elektrickej energie vyvolanej nabíjaním EV pomocou časových radov".

Predikciu elektrickej energie môžeme rozdeliť podľa dĺžky predpovedaného časového intervalu na krátkodobú (obdobie niekoľkých hodín až jedného týždňa s frekvenciou dát v minútach až hodinách) a dlhodobú (predpovede na nadchádzajúce mesiace príp. aj roky, s frekvenciou v dňoch alebo aj mesiacoch). Krátkodobé predikovanie energie súvisí najmä s nákupom elektrickej energie, ale taktiež môže byť relevantné pre aplikácie zamerané na efektívnejšie využitie obnoviteľných zdrojov energie [91]. Dlhodobé predikcie môžeme použiť na aproximáciu správania systému, odhad záťaže distribučnej sústavy, ale aj plánovanie inteligentných elektrických sietí.

Ako už bolo spomenuté v prehľade literatúry, autori sa v [6] snažili krátkodobo predpovedať obsadenosť a spotrebu energie jednotlivých staníc, pričom nedosiahli uspokojivé výsledky a odporučili stanice priestorovo agregovať. Predpovede pre jednotlivé stanice sú obtiažne najmä preto, že je potrebné predpovedať typ pripojeného vozidla a obsadenosť stanice. Ak je daná stanica obsadená, používateľ EV pravdepodobne použije stanicu v blízkom okolí. Takáto zmena sa na agregovanej spotrebe niekoľkých staníc neprejaví. Priestorová agregácia staníc môže byť zaujímavá aj z pohľadu inteligentných elektrických sietí, keďže sieť zásobuje viacero nabíjacích staníc, ktoré môžu tvoriť značný podiel dopytu po elektrickej energii.

Najskôr predstavíme použité dáta vrátane exogénnych prediktorov. Potom popíšeme

modely použité na predikovanie spotrebovanej energie a skombinujeme ich s tromi trénovacími procedúrami, ktoré boli navrhnuté za účelom vylepšenia kvality predikcií.

3.2.1 Použité dáta

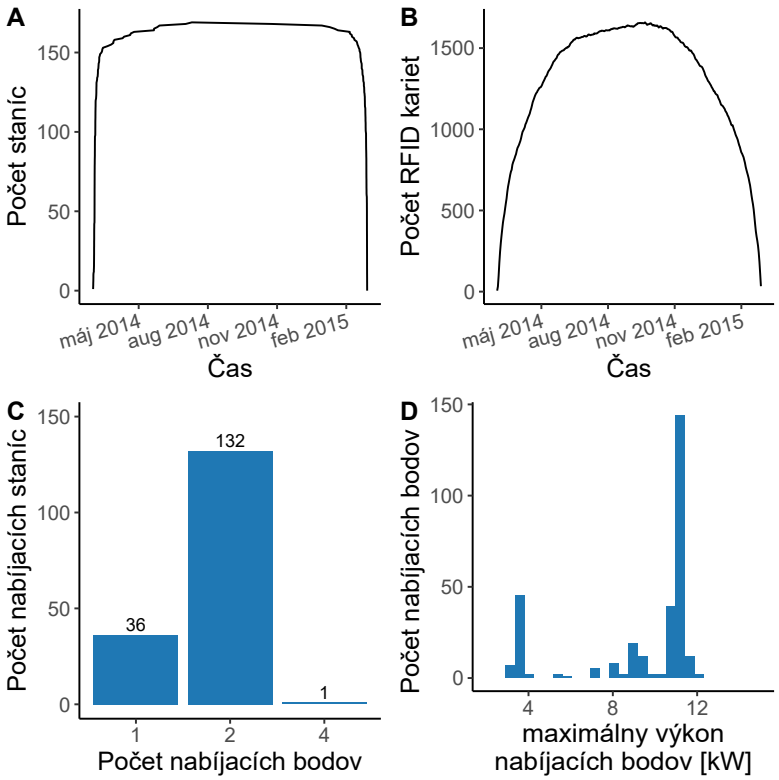
Pre potreby predpovedania spotreby elektrickej energie agregujeme spotrebu nabíjacích staníc z časového a priestorového hľadiska. Pre priestorovú agregáciu nabíjacích staníc, sme použili COROP regióny rovnako ako v [4], deliace Holandsko na 40 častí. Časovo sme dáta agregovali na dennú frekvenciu. Výrazne vyššiu ako aj stabilnejšiu spotrebu, najmä z pohľadu sezónnosti, mal v porovnaní s inými COROP regiónmi región Utrecht. Toto prisudzujeme aj faktu, že v tomto regióne EVnetNL databáza obsahuje najviac nabíjacích staníc (169). Pre trénovanie predikčným modelov uvažujeme obdobie medzi 3. marcom 2014 a 1. marcom 2015 (364 dní), pre testovanie obdobie medzi 2. marcom 2015 a 6. decembrom (280 dní). V tomto období bol počet staníc už relatívne stabilný (obrázok 7A) a počet aktívnych používateľov EV prekračuje 1600 (obrázok 7B). Pre toto obdobie a geografický región sme vyextrahovali 1 665 409 odčítaní z meračov uložených v datasete Meterreadings a prislúchajúcich k 51 642 nabíjacím transakciám z datasetu Transactions. Hlavným získaným časovým radom bola agregovaná spotreba energie pre región Utrecht s dennou frekvenciou zobrazená na obrázku 9.

V časovom rade spotreby energie na obrázku 9 sú zreteľné sezónne vzory. Na identifikáciu sezónnych vzorov na krátkej časovej škále sme využili ACF a PACF, ktoré indikovali silnú týždennú sezónnosť.

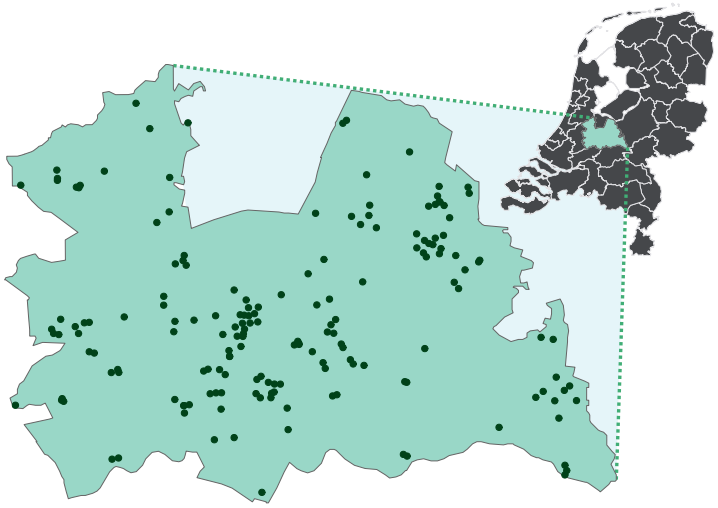
Externé dáta

Pre vylepšenie predpovedí uvažujeme aj externé prediktory, ktoré môžeme vytvoriť z externých dát, pri ktorých očakávame vplyv na spotrebovanú energiu. Priemerná spotreba energie cez víkendy je oveľa nižšia (v priemere o 30 %) ako cez pracovné dni. Z tohto dôvodu sme pridali binárny prediktor na rozlíšenie víkendov a pracovných dní. Spotreba cez víkendy je podobná sviatkom a dňom medzi sviatkami a víkendmi, preto sme pridali binárny prediktor aj pre sviatky.

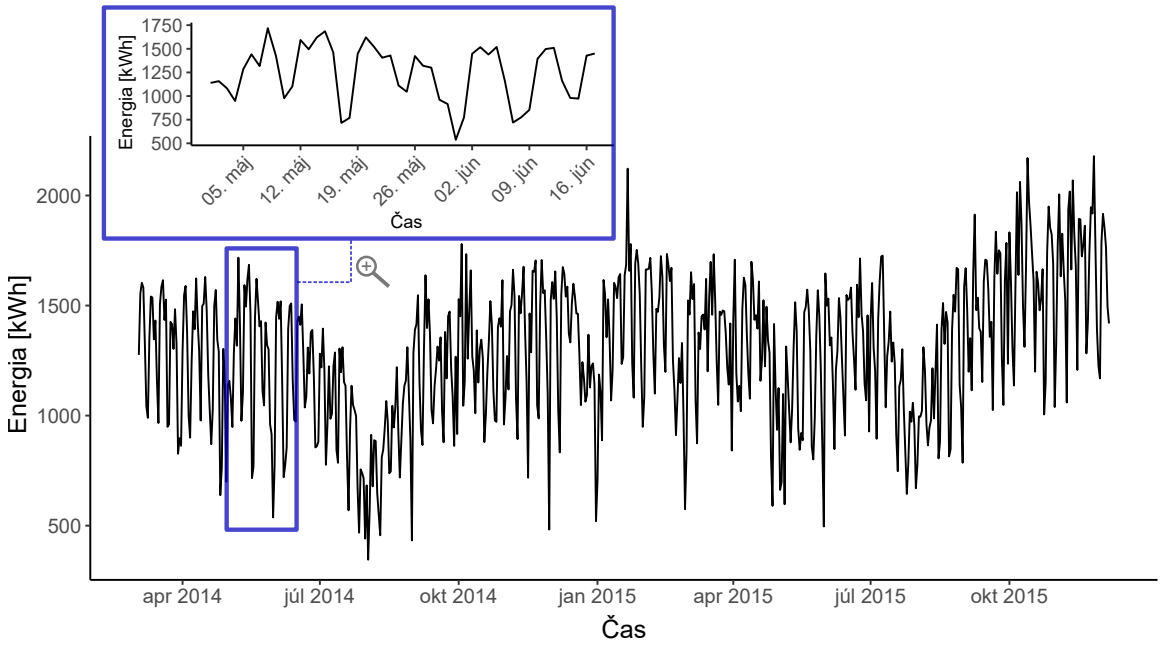
Spotreba energie často vykazuje viacnásobnú sezónnosť [31] a pozorovaním dát sme identifikovali pokles spotrebovanej energie cez prázdniny a ku koncu septembra, ktorý bol prítomný aj vo všetkých 4 rokoch úplných dát pre daný región, preto sme analyzovali spotrebovanú energiu 364 dní späť (ten istý deň predošlý rok) ako exogénnu premennú.



Obrázok 7: **A** Počet aktívnych staníc v COROP regióne Utrecht, vo zvolenej časovej perióde od 3. marca 2014 do 1. marca 2015. **B** Počet aktívnych používateľov odhadnutý z počtu RFID kariet, identifikovaný ich časom prvého a posledného použitia. **C** Počet nabíjacích staníc s daným počtom nabíjacích bodov. **D** Histogram maximálneho výkonu nabíjacích staníc. Hodnoty boli odhadnuté z odčítaní merača spotreby.



Obrázok 8: Geografické pozície 169 nabíjacích staníc v regióne Utrecht a pozícia regiónu v Holandsku.



Obrázok 9: Denná spotreba elektrickej energie nabíjacou infraštruktúrou v COROP regióne Utrecht. Modrý štvorec je priblížením úseku časového radu pre lepšie zobrazenie týždennej sezónnosti.

Skúmaním vzťahu medzi spotrebou jednotlivých používateľov a celkovou spotrebou energie sme zistili, že 300 z viac ako 5000 všetkých používateľov spotrebovalo približne 50 % energie. Počet týchto používateľov má korelačný koeficient 0.91 so spotrebovanou energiou a nazveme ich super používatelia. Počet super používateľov bol signifikantnejší pre predpovedanie spotrebovanej energie ako energia spotrebovaná pred 364 dňami. Toto prisudzujeme pravidelnejším vzorom v počte super používateľov ako energii spotrebovanej všetkými používateľmi. Pre tieto dôvody sme vybrali ako exogénny prediktor počet super používateľov pred 364 dňami namiesto energie spotrebovanej pred 364 dňami. Pre lepšie odchytenie sezónnosti sme do modelov pridali aj Fourierove rady (rovnica (34)) s frekvenciou časového radu a 1 až 3 členmi.

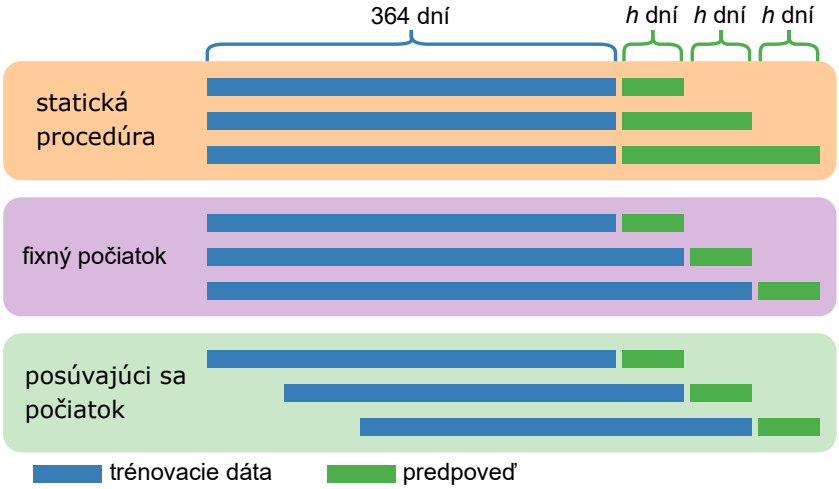
3.2.2 Modelovanie a predikcia spotreby elektrickej energie

Spotrebovanú energiu na staniciach sme sa rozhodli modelovať pomocou SARIMAX modelov s exogénnymi prediktormi, RF a GBRT modelmi a chybu budeme merať pomocou MAPE.

Ako možné vylepšenie sme vytvorili striedavý model, skladajúci sa z dvoch SARIMAX modelov. Prvý modeluje pracovné dni a druhý víkendy. Pre zahrnutie informácie o dňoch v týždni vo víkendovej časti striedavého modelu, sme ako prediktor pridali predikovanú

Procedúra	Trénovacia množina
statická	$y_i,\ y_{i+1},\ \ldots,\ y_{i+d}$
fixný počiatok	$y_i,\ y_{i+1},\ \ldots,\ y_{i+d+j}$
posúvajúci sa počiatok	$y_{i+j},\ y_{i+1+j},\ \ldots,\ y_{i+d+j}$

Tabuľka 6: Trénovacie množiny pre jednotlivé procedúry, kde i je čas začiatku dát, j je čas začiatku predpovede a d je veľkosť trénovacej množiny.



Obrázok 10: Trénovacie procedúry pre časové rady.

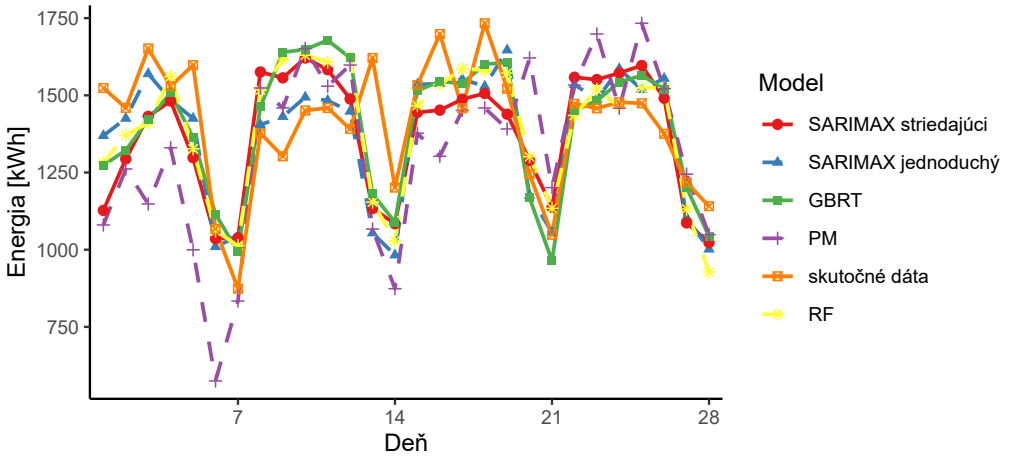
spotrebu za pracovný týždeň, po ktorom nastupuje daný víkend, ako exogénnu premennú.

S plynúcim časom rastie počet nabíjacích staníc a počet EV, preto sme navrhli tri rôzne procedúry na trénovanie modelov. Princíp procedúr je zobrazený v obrázku 10 a prehľad trénovacích množín pre jednotlivé procedúry je zobrazený v tabuľke 6.

Statická procedúra využíva stále rovnaké dáta na trénovanie, netrénuje modely nanovo s plynúcim časom a predlžuje iba dĺžku predpovede o novú periódu. Procedúra s fixným počiatkom vylepšuje predpovede trénovaním aj na nových dátach a model predpovedá vždy iba novú periódu. Procedúra s posúvajúcim sa počiatkom uvažuje na trénovanie dáta z intervalu konštantnej dĺžky, ktorý sa posúva v čase, t.j. trénuje vždy na dátach z posledného 364-dňového úseku. Model predpovedá jednu periódu rovnako ako procedúra s fixným počiatkom.

Na porovnanie výsledkov sme použili perzistenčný model (PM), čo je nulový model využívajúci sezónnosť, predikujúci periódy cieľového časového horizontu replikovaním poslednej pozorovanej periódy.

Modely sme testovali na predpovediach 7, 14 a 28 dní dopredu. Na testovanie používame 280 dní a tak nám vyjdú nasledovné počty množín predpovedí, $280/7 = 40$, $280/14 = 20$, $280/28 = 10$, v závislosti od dĺžky predpovede.



Obrázok 11: Ilustrácia predpovedí spotreby energie pomocou jednotlivých modelov predpovedajúcich 7 dní vopred pre typické 28 dňové obdobie. Skutočné dáta sú vyznačené oranžovou čiarou.

MAPE modelu (%)	Dĺžka predpovede		
	7 dní	14 dní	28 dní
SARIMAX jednoduchý	**12.02**	12.35	12.97
SARIMAX striedavý	12.18	**12.15**	**12.60**
RF	12.58	12.55	13.28
GBRT	12.68	13.22	13.45
PM	16.40	16.86	17.98

Tabuľka 7: Výsledky pre jednotlivé modely a dĺžky predpovedí v percentách MAPE. Najlepšie výsledky pre jednotlivé dĺžky predpovedí sú zvýraznené tučným písmom.

Aby sme identifikovali najlepšiu predikčnú architektúru, testovali a porovnali sme každú z troch trénovacích procedúr na všetkých modeloch. Výsledky predpovedí uvádzame pre kombináciu metód a trénovacích procedúr v tabuľke 7. Ilustrácia získaných 7 dňových predpovedí pre 28-dňové obdobie je zobrazená v obrázku 11. Hodnoty MAPE ukazujú, že výsledky sú ovplyvnené použitou trénovacou procedúrou a dĺžkou predpovede. GBRT model dosahuje najlepší výkon pre procedúru s posúvajúcim sa počiatkom. SARIMAX a RF dosiahli najlepší výkon v kombinácii s trénovacou procedúrou s fixným počiatkom. Striedavý SARIMAX model prekonal jednoduchý SARIMAX model na 14 a 28 dňových predpovediach.

Pre 7 dňovú dĺžku predpovedí dosahuje najlepšie výsledky jednoduchý SARIMAX model s hodnotu MAPE o 1.3 menšou v porovnaní s druhým najlepší modelom a s chybou o 26.7 % menšou v porovnaní s perzistenčným modelom. Pre 14 a 28 dňovú predpoveď dosahuje najlepšie výsledky striedavý SARIMAX model a to o 28 % a 30 % lepšie v porovnaní s PM modelom.

Na základe týchto výsledkov môžeme usúdiť, že na zvolenej časopriestorovej úrovni vieme s relatívne dobrou presnosťou predpovedať spotrebovanú energiu. Dosiahnuté výsledky sú približne o 30 % lepšie ako poskytuje sezónny model PM.

Podkapitola vychádza z vlastných publikácií [104, 13]. Pre implementáciu SARIMAX modelu sme použili knižnicu *auto.arima* balíka *forecast* v prostredí jazyka R, kde funkcia automaticky vyhľadáva parametre a rád SARIMAX modelu. Pre RF a GBRT modely sme použili prostredie MATLAB s automatickou optimalizáciou hyperparametrov a použili sme 5-skupinovú krížovú validáciu na výber prediktorov.

3.3 Extrakcia prediktorov z GIS dát, testovanie metód na výber premenných

Nasledujúce dve podkapitoly sa venujú analýze indikátorov vhodného umiestnenia nabíjacej infraštruktúry pomocou regresných úloh. V tejto kapitole uvedieme opis metodológie, ktorá je spoločná pre obe podkapitoly. Najskôr sa budeme zaoberať tvorbou prediktorov z GIS dát a následne predstavíme výsledky porovnania metód určených na výber premenných (angl. variable selection).

Nabíjacie miesta

Ak sú stanice príliš blízko seba, napr. stanice v podzemných parkoviskách, nedokážeme ich poriadne rozlíšiť v priestorovej analýze a tak sme vytvorili z nich nabíjacie miesta nasledovným spôsobom. Ak bola vzdušná vzdialenosť medzi stanicami menšia ako 50 metrov, agregovali sme ich a vytvorili z nich jedno nabíjacie miesto, reprezentované súradnicami prostrednej zo staníc. V prípade ak bol párny počet staníc, bola z prostredných vybratá tá s prvým označením stanice v abecede.

3.3.1 Extrakcia prediktorov z GIS dát

Pre vhodnú reprezentáciu (modelovanie) okolia nabíjacieho miesta pomocou dátovej matice $\boldsymbol{X}$, ktorá je vstupom do metód strojového učenia, máme k dispozícií niekoľko možností. Jedným z nich sú aj kernel funkcie, priraďujúce váhy bodom v priestore. Typický príklad je Gaussovský kernel [64]. Ďalším príkladom modelovania je obalová zóna (angl. buffer), definovaný indikačnou funkciou, ktorá priraďuje hodnotu 1 reprezentáciám objektov v danej vzdialenosti od bodu reprezentujúceho lokáciu nabíjacieho miesta. Vzdialenosť môžeme, napríklad, merať ako vzdušnú vzdialenosť, ktorá tvorí kruhovú zónu alebo cestnú vzdialenosť, ktorá vytvorí zónu nepravidelného tvaru [111]. Kvôli obmedzenej dostupnosti dát a nižšej výpočtovej zložitosti sme použili kruhové zóny, centrované na bode reprezentujúcom nabíjacie miesto a definované polomerom r.

Pre nabíjacie miesto chceme z jeho okolia získať jednu skalárnu hodnotu - pozorovanie daného prediktora, napr. počet ľudí žijúcich v okolí, ktorú budeme extrahovať z atribútov objektov patriacich do okolia nabíjacieho miesta reprezentovaného kruhovou zónou. GIS dáta sú reprezentované objektami, ku ktorým je priradený jeden alebo viac atribútov. Ak uvažujeme ako objekty polygóny, tak každá kruhová zóna môže mať prienik s jedným

alebo viacerými polygónmi. Teda hodnotu atribútu potrebujeme odhadnúť z polygónov, ktoré majú prienik s danou kruhovou zónou. Pre jednoduchosť predpokladáme, že hodnoty atribútu sú rovnomerne rozdelené po celej ploche polygónu. Keďže naše odhadové procedúry závisia na type atribútu, najskôr zosumarizujeme hlavné typy atribútov podľa [27, s. 12]:

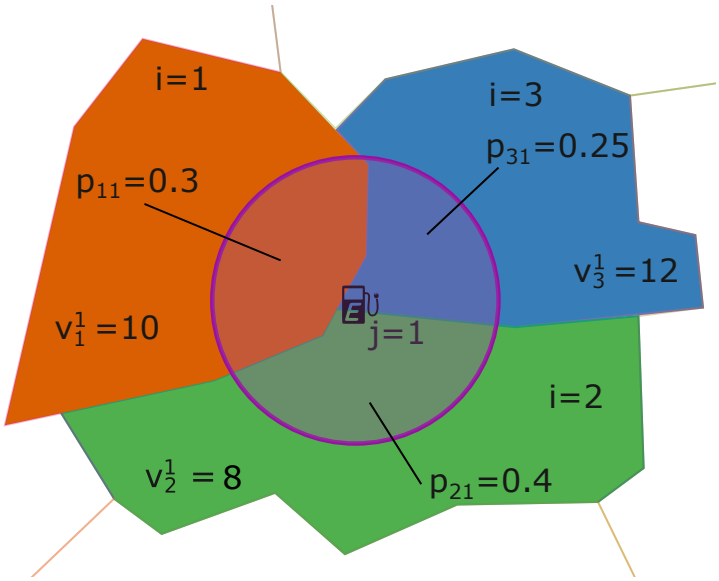
- *Atribút vyjadrujúci počet* - numericky reprezentuje počet objektov, napr. populáciu v polygóne.
- *Priemerný atribút* - je numerický atribút vyjadrujúci priemernú hodnotu, napr. priemerný plat obyvateľov územia reprezentovaného polygónom.
- *Percentuálny atribút* - nadobúda percentuálne hodnoty, napr. percento ženskej populácie v polygóne.
- *Kategorický atribút* - nadobúda hodnoty z konečnej diskrétnej množiny hodnôt. Kategorické atribúty môžeme rozdeliť na dva druhy podľa usporiadania hodnôt:
  - *Nominálne atribúty* - nemajú vnútorné usporiadanie hodnôt. Tento druh atribútu je najviac využívaný na popísanie pokrytia alebo využitia územia, napr. či je územie obytné územie, park alebo voda.
  - *Poradový atribút* - kategorické hodnoty atribútu, ktoré sa dajú usporiadať, napr. hustota adries vyjadrená číslami od 1 do 5, kde 1 je najnižšia a 5 je najvyššia hustota.

Pre zjednodušenie popisu nasledovných výpočtov, budeme spoločným názvom *relatívne atribúty* označovať priemerné, percentuálne a kategorické poradové atribúty.

Odhad hodnôt atribútov pre kruhové zóny

Keďže chceme reprezentovať okolie prediktormi a zvolili sme kruhové zóny ako reprezentáciu okolia nabíjacích miest, tak potrebujeme odvodiť vzťahy pre získanie odhadu hodnoty prediktora (kruhovej zóny) z atribútov priestorových objektov prenikajúcich s kruhovou zónou.

Aby sme formálne popísali náš prístup, uvádzame nasledovné značenie. Prienik i-teho polygónu a j-tej kruhovej zóny značíme indexom ij. Symbol v_i^l označuje hodnotu atribútu l v i-tom polygóne, v_{ij}^l je hodnota atribútu l prieniku ij a $\bar{v}_j^l$ je hodnota atribútu l asociovaná



Obrázok 12: Schematická ilustrácia prieniku kruhovej zóny $j = 1$ s tromi polygónmi $i = 1, 2$ a 3.

s j-tou kruhovou zónou. Takéto značenie sme zvolili aj preto, lebo väčšina výpočtov hodnôt pre kruhovú zónu vychádza z aritmetického priemeru. Podobne, hodnota s_i je plocha i-teho polygónu, s_{ij} je plocha prieniku ij a $\bar{s}_j$ je plocha j-tej kruhovej zóny. Hodnota p_{ij} je pomer plochy prieniku ij voči ploche polygónu i, t.j. $p_{ij} = s_{ij}/s_i$. M_j je množina indexov polygónov, ktoré majú nenulový prienik s j-tou kruhovou zónou.

Pre relatívny atribút l môže existovať *súvisiaci atribút* vyjadrujúci počet označený m, spresňujúci výpočet relatívneho atribútu. Toto vychádza z predpokladu, že pri výpočte relatívneho atribútu nemusí úlohu zohrávať len veľkosť územia, ale hlavne atribút vyjadrujúci počet. Napríklad, pre percento žien je súvisiaci celkový atribút celková populácia, pomocou ktorej potom môžeme presnejšie odhadnúť hodnotu pre kruhovú zónu. Napríklad, predstavme si, že kruhová zóna má rovnako veľkú plochu prieniku s dvomi polygónmi. V prvom prieniku žije 20 % žien a 100 ľudí a v druhom 40 % žien a 900 ľudí, odhad podľa plochy by bol 30 %, no v skutočnosti je to 38 %. Nasledovný zoznam obsahuje vzťah pre výpočet hodnoty atribútu pre oblasť kruhovej zóny pre každý typ atribútu.

a) Atribút vyjadrujúci počet - hodnotu atribútu l vyjadrujúceho počet odhadneme ako sumu jeho hodnôt v polygónoch v kruhovej zóne

$$\bar{v}_j^l = \sum_{i \in M_j} v_{ij}^l. \tag{36}$$

b) Relatívny atribút so známou hodnotou súvisiaceho atribútu - ak je dostupná hodnota súvisiaceho atribútu v_i^m, tak môže byť využitá a hodnotu kruhovej zóny odhadneme ako

$$\bar{v}_j^l = \frac{\sum_{i \in M_j} v_i^l v_i^m p_{ij}}{\sum_{i \in M_j} v_i^l p_{ij}}. \tag{37}$$

c) Relatívny atribút s neznámou hodnotou súvisiaceho atribútu - odhad môžeme vypočítať podľa

$$\bar{v}_j^l = \frac{\sum_{i \in M_j} v_i^l s_{ij}}{s_j}. \tag{38}$$

d) Kategorický nominálny atribút - tu odhadujeme hodnotu atribútu $\bar{v}_j^{lc}$ pre každú kategóriu c ako

$$\bar{v}_j^{lc} = \sum_{i \in M_{jc}} \frac{s_{ij}}{s_j}, \tag{39}$$

kde M_{jc} je množina indexov polygónov kategórie c, čo majú neprázdny prienik s kruhovou zónou j.

e) Kategorický poradový atribút - je odhadnutý rovnakým spôsobom ako relatívny atribút s neznámou hodnotou súvisiaceho atribútu (vzťah (38)).

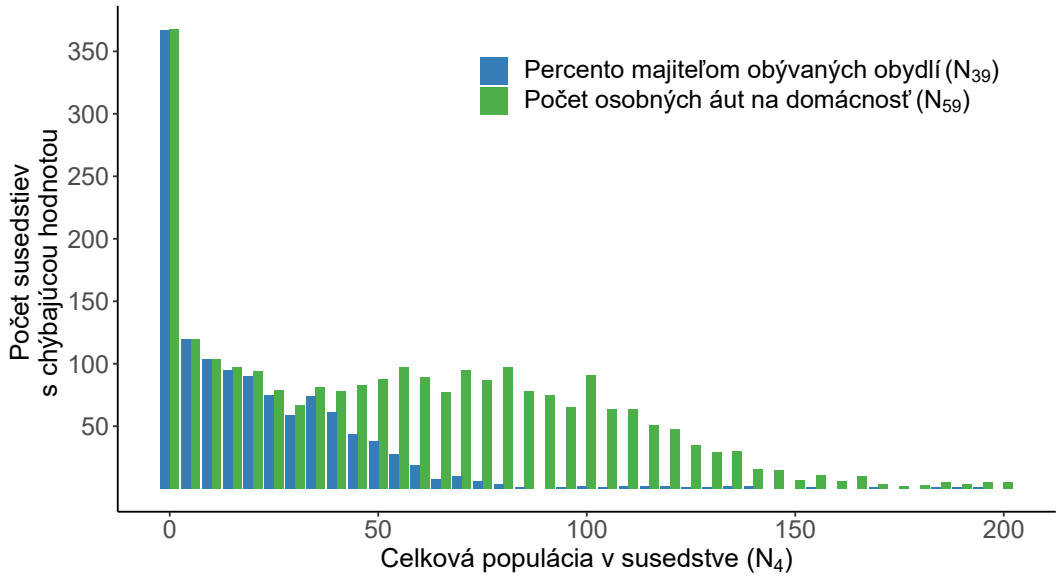
Tieto odhady následne aplikujeme na atribúty, pre ktoré sme ručne určili ich typ, prípadne aj ich súvisiace atribúty.

3.3.2 Predspracovanie dát pre regresné a stromové metódy

Keďže isté druhy dát vyžadujú ešte predspracovanie najmä pre regresné úlohy, uvedieme si postup, ktorým tieto dáta spracovávame.

Chýbajúce hodnoty

Mnoho polygónov v GIS dátach obsahovalo atribúty s chýbajúcimi hodnotami. Pri vizualizácii takýchto polygónov s chýbajúcimi hodnotami atribútov sme zistili, že príčinou boli najmä oblasti reprezentujúce vodu a územia s nízkym počtom obyvateľov. Obrázok 13 zobrazuje pre dva atribúty závislosť počtu chýbajúcich hodnôt od počtu obyvateľov. Polygóny reprezentujúce vodu sme z dát vylúčili, okrem pokrytia územia (dataset Využitie pôdy), keďže všetky nabíjacie miesta sú umiestnené na suchom území a vplyv takejto vody nepovažujeme za významný. Ak chýbajú hodnoty GIS atribútu iba pre jedno alebo



Obrázok 13: Ilustrácia závislosti medzi počtom chýbajúcich hodnôt pre dva atribúty pochádzajúce z datasetu Susedstvá a celkovej populácie. Pre oba vybrané atribúty je percento vlastníkom obývaných nehnuteľností (N_{39}) a počet áut v domácnosti (N_{59}). Ako si môžeme všimnúť, počet chýbajúcich hodnôt v dátach klesá s populáciou.

malú skupinu nabíjacích miest, buď sa z dát odstráni daný atribút alebo celé nabíjacie miesto. So zvyšnými chýbajúcimi hodnotami sa vysporadúvame v záverečných fázach odhadu hodnôt pre kruhovú zónu.

Pomocou analýzy dát sme vytvorili pravidlá na doplnenie chýbajúcich dát, uvedených v Prílohe ***A-3***.

Výber polomeru kruhovej zóny

Polomer kruhovej zóny r sme vybrali z intervalu, ktorý sme zvolili na základe očakávanej vzdialenosti, ktorú sú ochotní vodiči vozidiel prejsť pešo z parkovacieho miesta (čo je v našom prípade nabíjacie miesto) do požadovaného cieľa cesty [120]. Konkrétnu hodnotu polomeru r sme vybrali ako hodnotu pre ktorú bola dosiahnutá najlepšia vysvetliteľnosť, daná hodnotou R^2, pri odhade metódou najmenších štvorcov vektora výstupu prediktormi.

Extrahovanie prediktorov z GIS atribútov

Pre väčšinu GIS dát bolo možné prediktory z atribútov vyextrahovať priamo pomocou kruhových zón. Pre tie, kde sa nehodilo tento prístup uplatniť, si popíšeme procedúru extrahovania jednotlivo.

Vysoký stupeň korelácií sme identifikovali medzi atribútmi datasetu Dopravné toky,

charakterizujúceho priemerný tok áut ako funkciu času dňa. Z tohto dôvodu sme vytvorili tri prediktory (jeden pre každý dopravný mód) agregovaním tokov pre rôzne časy dňa: počet áut za deň ($TF_{cars} = TF_1 + TF_2 + TF_3$), počet autobusov za deň ($TF_{buses} = TF_4 + TF_5 + TF_6$) a počet nákladných vozidiel za deň ($TF_{trucks} = TF_7 + TF_8 + TF_9$). Vysvetlivky k skratkám atribútov sa nachádzajú v Prílohe ***A-1***, Tabuľke 19

Pre každý nabíjací bod sme identifikovali najbližší cestný segment a použili jeho tok ako prediktor. Okrem toho sme pre každú kruhovú zónu vypočítali hustotu dopravy vynásobením dĺžky všetkých ciest ich tokom dopravy a podelili plochou kruhovej zóny. Hustota ciest bola vypočítaná podelením dĺžky ciest plochou kruhovej zóny a použitá tiež ako prediktor [63]. Pomocou kódovania kategorických premenných v regresii, sme kódovali typ najbližšieho cestného segmentu (kategorický atribút s hodnotami: rezidenčný, primárny, sekundárny alebo terciárny) využitím 3 binárnych prediktorov.

Z bodových datasetov OpenStreetMap a Nabíjacie stanice 2015 sme extrahovali prediktory dvomi spôsobmi. Použili sme vzdušnú vzdialenosť k najbližšiemu bodu a počet bodov v kruhovej zóne, ak tam sa nejaké body nachádzali, inak boli obe hodnoty nastavené na 0. Z EVnetNL datasetu sme vyextrahovali ako prediktory počet nabíjacích bodov na nabíjacom mieste, maximálny výkon nabíjacieho miesta, zemepisnú šírku a dĺžku a umiestňovaciu stratégiu nabíjacieho miesta.

Na záver sme vyhodnotili chýbajúce hodnoty pre všetky prediktory. Ak nejakému prediktoru chýbalo menej ako 1.5 % hodnôt, chýbajúce hodnoty boli nahradené mediánom dostupných hodnôt, ináč bol prediktor odstránený.

Predspracovanie GIS prediktorov

Po extrakcii prediktorov je potrebné ich ďalej analyzovať a pripraviť pre použitie štatistických metód. Aby sme sa vyhli dátam, ktoré prinášajú malé množstvo užitočnej informácie, analyzovali sme frekvenciu hodnôt vo všetkých prediktoroch. Našli sme prípady, kde bolo viac ako 95 % hodnôt rovných nule. Desať takýchto prediktorov, deväť týkajúcich sa pokrytia územia a jeden odpovedajúci typu cestných úsekov, sme odstránili, pretože môžu spôsobovať problémy pri deľbe dát na trénovacie a testovacie a majú iba malú vysvetľovaciu schopnosť [59].

Keďže zdrojové dáta vykazujú vysokú mieru kolinearity, našli sme silné závislosti aj medzi niektorými prediktormi. Ak bola absolútna hodnota korelačného koeficientu medzi skupinou prediktorov väčšia ako 0.95, zvolili sme medzi nimi zástupcu a ostatné prediktory

zo skupiny sme vylúčili z analýzy. Tieto prediktory zobrazuje Príloha ***A-2***, Tabuľka 21.

Použitie nelineárnych výrazov v regresii môže vylepšiť kvalitu modelu. Množinu prediktorov sme rozšírili pridaním prediktorov získaných aplikovaním funkcií $\log(\cdot+1)$, kde $\sqrt{\cdot}$ a $(\cdot)^2$ na všetky prediktory okrem binárnych. Podobne sme pridali aj násobky všetkých dvojíc prediktorov (interakčné členy) v súlade s tým, ako to je často odporúčané v literatúre [55, s. 93]. Následne sme rozšírenou maticou prediktorov modelovali vektory výstupu a merali MSE v krížovej validácií. Keďže takáto úprava nepriniesla výraznejšie zlepšenie kvality modelu, tak sme tieto nelineárne členy nezahrnuli.

3.3.3 Testovanie metód na výber premenných

Predspracované dáta z okolia nabíjacej infraštruktúry vykazujú vysokú mieru multikolinearity a tak treba preveriť, ako sa s ňou metódy na výber premenných vedia vysporiadať.

Keďže sa zameriamave hlavne na inferenciu, metódy ako analýza hlavných komponentov alebo čiastočná metóda najmenších štvorcov [51, s. 67, 81] na redukciu kolinearity, nie sú príliš vhodné, keďže lineárne kombinácie prediktorov, ktoré vytvárajú, sú ťažko interpretovateľné. Preto sa v tejto časti zameriame hlavne na regresné metódy s krokovou selekciou a tolerantné metódy.

Porovnávať budeme metódu najmenších štvorcov, nasledovné metódy krokovej selekcie: výberová regresia hrubou silou, dopredná a spätná regresia a tolerantné metódy: hrebeňovú regresiu, lasso, elastic net a PACS.

Testovacie úlohy

Na porovnávacie testy sme vygenerovali umelé dáta so závislými prediktormi. Ich úlohou je simulovať rôzne miery multikolinearity a preveriť schopnosť metód vysporiadať sa s multikolinearitou. Každý dataset má 16 prediktorov a 1000 pozorovaní. Prediktory sú organizované do 4 skupín po 4 prediktory. Dve skupiny tvoria korelované prediktory so vzájomnými koreláciami so stupňom $c_1 \in \{0.7, 0.8, 0.9\}$, jednu multikolineárne a jednu nekorelované prediktory. Prediktory boli generované z viacrozmerného normálneho rozdelenia $N(0, \Sigma)$, s nulovým priemerom a kovariančnou maticou Σ. Prvky matice Σ boli nastavené tak, aby spĺňali požadované stupne korelácie v skupinách prediktorov.

Vo štvrtej skupine prediktorov bola multikolinearita pridaná ako lineárna kombinácia prediktorov z ostatných skupín spolu s prídavkom šumu s nulovým priemerom a rozptylom $v_2 \in \{0.1, 0.2, 0.5\}$: $x_9 = \frac{1}{3}x_1 + \frac{1}{3}x_2 + \frac{1}{3}x_3 + N(0, v_2)$, $x_{10} = \frac{1}{3}x_{13} + \frac{1}{3}x_{14} + \frac{1}{3}x_{15} + N(0, v_2)$,

$x_{11} = \frac{1}{4}x_1 + \frac{1}{4}x_2 + \frac{1}{4}x_5 + \frac{1}{4}x_6 + N(0, v_2)$ a $x_{12} = \frac{1}{4}x_1 + \frac{1}{4}x_2 + \frac{1}{4}x_{13} + \frac{1}{4}x_{14} + N(0, v_2)$.

Pri budovaní vektora výstupu y boli koeficienty β_i, kde $i \in 1 \dots p$, nastavené na 1, kde i bolo párne číslo, ináč boli nastavené na 0, čím sa eliminoval vplyv prediktory daných prediktorom na odpoveď, t.j. stali sa falošne vplývajúcimi. K vektoru výstupu sme ešte pridali náhodný šum s normálnym rozdelením $N(0, v_1)$, kde $v_1 \in \{1.4, 1.8, 2.2\}$. Datasety pre numerické experimenty boli vytvorené kombináciou hodnôt parametrov v_1, v_2 a c_1, pričom pre každú z 9 kombinácií sme použili v experimentoch 100 náhodných inštancií.

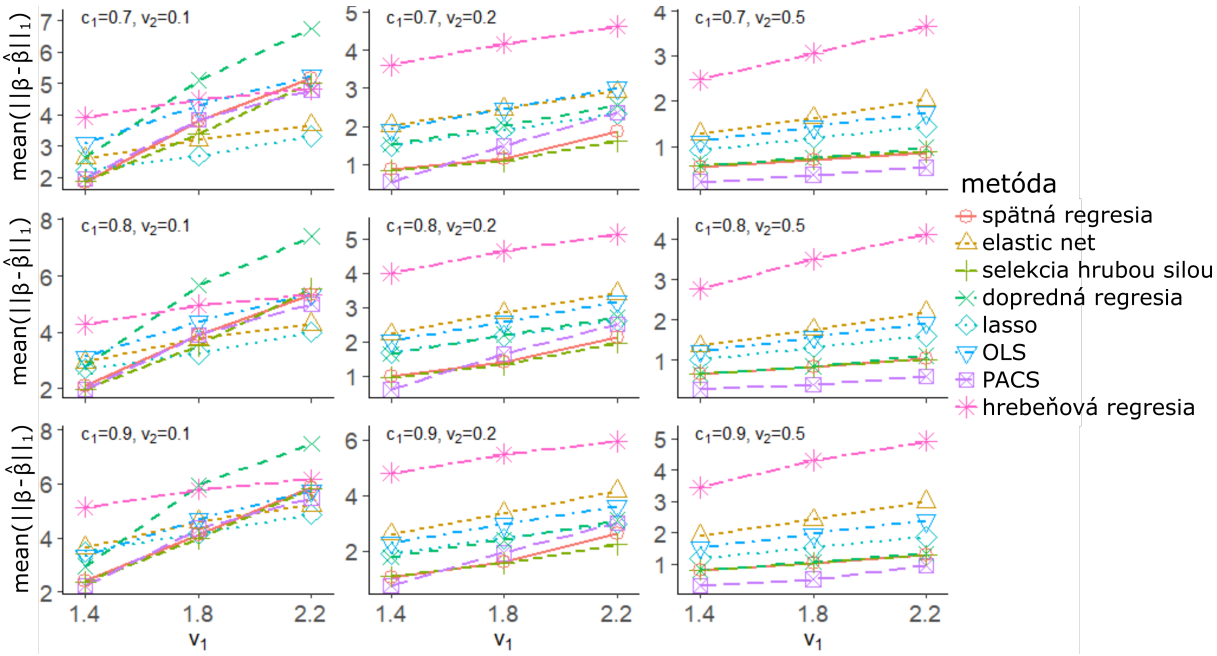
Na trénovanie metód bola využitá krížová validácia s $k = 10$ a ako kritérium pre výber parametrov modelu bola použitá najmenšia hodnota MSE. Pre krokové metódy sa týmto spôsobom vyberali prediktory, pre tolerantné metódy sa vyberal parameter λ_{min}^{CV}, pričom bola uvažovaná postupnosť hodnôt 10^i pre i v intervale od -4 po 0 v krokoch po 0.02. Hodnota hyperparametra α v metóde elastic net bola vybraná tak, aby tiež odpovedala najmenšej hodnote MSE dosiahnutej na testovacích dátach.

Výsledky porovnania metód

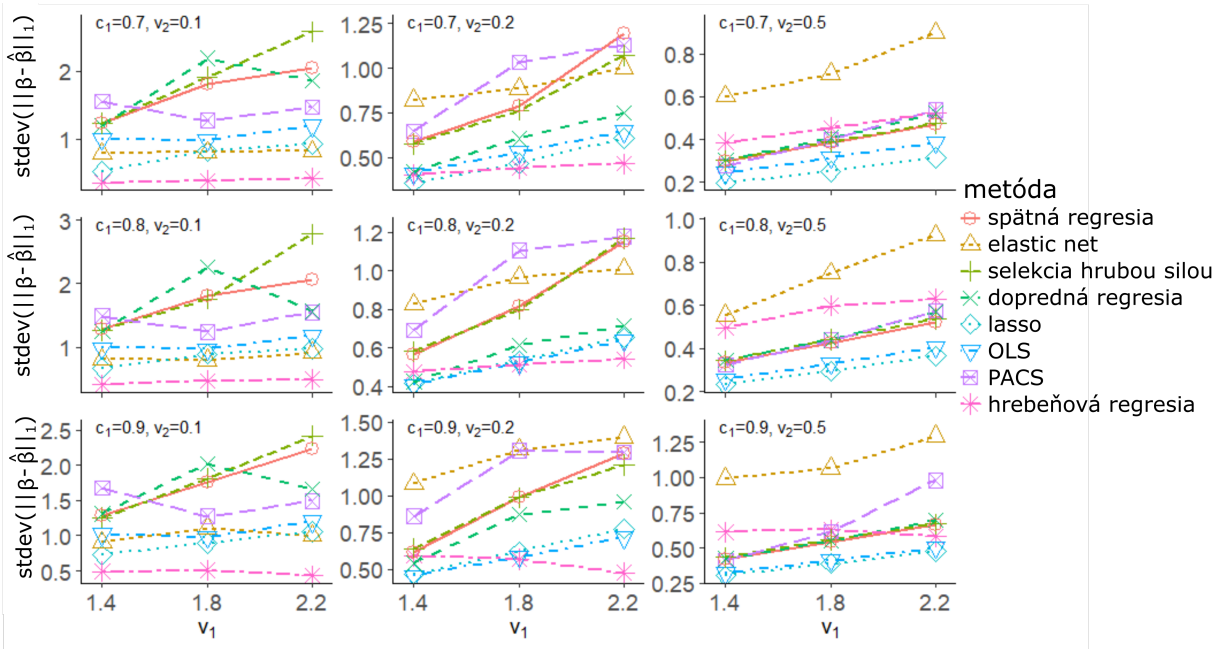
Na evaluáciu schopnosti metód vyberať premenné sme použili Manhattanskú vzdialenosť [49, s. 72] $\|\beta - \hat{\beta}\|_1$ medzi vektorom β skutočných hodnôt použitých pri generovaní dát a vektorom $\hat{\beta}$ odhadnutým metódami. Priemerné hodnoty Manhattanskej vzdialenosti sú zobrazené v Obrázku 14 a smerodajná odchýlka vzdialeností je ukázaná v Obrázku 15.

Metóda PACS má nízke hodnoty Manhattanskej vzdialenosti pre nižšie stupne multikolinearity, no má vysoký rozptyl hodnôt. Najlepší výsledok medzi všetkými modelmi pre strednú multikolinearitu majú spätná selekcia a selekcia hrubou silou, ale obe metódy majú vysokú smerodajnú odchýlku indikujúcu vysoký rozptyl v schopnosti výberu premenných. Metódy lasso a dopredná regresia majú podobné výsledky. Pre nižší stupeň multikolinearity, dopredná regresia dosiala trochu nižšie hodnoty Manhattanskej vzdialenosti, ale lasso má oveľa nižší rozptyl, čo ukazuje jeho väčšiu robustnosť. Metóda najmenších štvorcov dosahuje podobné výsledky ako elastic net, ale s rastúcimi koreláciami má metóda najmenších štvorcov menší priemer a smerodajnú odchýlku Manhattanskej vzdialenosti. Hrebeňová regresia dosiahla systematicky horšie výsledky s vysokou hodnotou priemernej Manhattanskej vzdialenosti. Ľahko povšimnuteľné je, že meniaci sa stupeň korelácie medzi prediktormi ovplyvňuje schopnosť výberu premenných metódami, ale nie až v takej miere ako multikolinearita.

Celkovo dosiahli najlepšie výsledky metódy PACS a lasso, pričom metóda PACS má



Obrázok 14: Priemerná hodnota Manhattanskej vzdialenosti medzi skutočnými a odhadnutými hodnotami regresných koeficientov.



Obrázok 15: Smerodajná odchýlka Manhattanskej vzdialenosti medzi skutočnými a odhadnutými hodnotami regresných koeficientov.

vyššiu variabilitu výsledkov a aj značne vyššiu výpočtovú náročnosť. Preto pre ďalšie použitie považujeme ako najlepšieho kandidáta metódu lasso, ktorá je aj populárnejšia a lepšie preskúmaná.

Spracovanie GIS dát v tejto podkapitole vychádzalo z vlastnej publikácie [106] a porovnávanie metód na výber premenných z publikácie [103]. Pre generovanie dát sme použili balík *mvtnorm* a funkciu *rmvnorm*, pre metódu lasso, hrebeňovú regresiu a elastic net bola použitá knižnica glmnet s rovnomennou funkciou, pre metódu PACS upravená funkcia *fitPacs* knižnice *clere* a pre tolerantné metódy.

3.4 Modelovanie popularity

Táto podkapitola je zameraná na predikciu a inferenciu popularity nabíjacích miest na základe ich okolia, ako jedného z hlavných indikátorov výkonnosti nabíjacej infraštruktúry.

Ako sme si mohli všimnúť v prehľade literatúry, žiadna zo štúdií sa nevenuje čisto vplyvu okolia stanice ale zahŕňa aj prediktory reprezentujúce správanie, ako napríklad časy príchodov, ktoré nemusia byť vopred známe pri budovaní staníc. Preto sa zameriavame najmä na vplyv okolia a preskúmavame aj potenciálne ukazovatele výkonnosti nabíjacej infraštruktúry.

Na základe aspektov vyjadrujúcich výkonnosť nabíjacej infraštruktúry popíšeme ukazovatele na meranie výkonnosti nabíjacích miest, kde sme sa zamerali na popularitu nabíjacích miest reprezentovanú počtom unikátnych používateľov EV na nabíjacom mieste. Podľa popularity zoradené nabíjacie stanice reprezentujeme binárnou premennou. Predikcie vykonávame logistickou regresiou regularizovanou $l-1$ normou a stromovými metódami na základe prediktorov nabíjacej infraštruktúry a jej okolia.

3.4.1 Modelovanie popularity

Identifikovali sme niekoľko širších aspektov spojených s nabíjacou infraštruktúrou, ktoré môže byť užitočné zohľadniť pri jej plánovaní a nasadzovaní:

- Z pohľadu vlády a samospráv, usilujúcich sa o inováciu cestnej dopravy, je jedným z hlavných cieľov, pri rozvoji nabíjacej infraštruktúry, stimulácia väčšieho využívania EV ako aj efektívna a férová investícia, keďže sú použité verejné zdroje.

- V súčasnosti môže operátorov elektrickej siete znepokojovať stabilita distribučných systémov a bezproblémová integrácia obnoviteľných zdrojov energie. Technológie ako smart charging môžu pomôcť využiť potenciál EV v dosiahnutí týchto cieľov. Toto vyžaduje súhru medzi elektrickou energiou vygenerovanou pomocou obnoviteľných zdrojov a nabíjaním EV.

- Pri budovaní alebo rozširovaní siete nabíjacích miest je jedným z hlavných cieľov operátorov nabíjacej infraštruktúry zisk, čo vyžaduje umiestnenie nabíjacích miest v lokalitách, ktoré majú potenciál byť dostatočne výnosné.

Rozsiahla množina indikátorov vytvorená za účelom porovnať umiestňovacie stratégie nabíjacej infraštruktúry medzi jednotlivými kontinentami, krajinami a regiónmi bola navrhnutá v [68]. S ohľadom na rozličné uhly pohľadu expertov, samospráv, operátorov elektrických systémov a operátorov nabíjacej infraštruktúry a zvažujúc dostupnosť dát sme vybrali nasledovné ukazovatele na meranie výkonnosti nabíjacích miest:

- *Spotrebovaná energia:* Na základe predplateného programu, používatelia EV platia pravidelne fixnú sumu a majú nelimitovaný prístup k nabíjacím službám alebo im je zarátaný poplatok pri pripojení vozidla a zaplatia stanovenú sumu za jednotku spotrebovanej energie. Preto je spotrebovaná energia veľmi úzko spätá so ziskovosťou a navyše aj indikuje, ako zložité je integrovať nabíjacie miesto s elektrickou sieťou.

- *Počet nabíjacích transakcií:* Čím viac vozidiel navštevuje nabíjacie miesto, tým vyšší je potenciálny zisk. Vysoký počet transakcií síce viac zaťažuje elektrickú sieť, no môže priniesť aj viac príležitostí pre smart charging.

- *Popularita nabíjacieho miesta:* Schopnosť nabíjacieho miesta prilákať rozsiahlu skupinu používateľov EV môže byť aproximovaná počtom unikátnych RFID kariet, použitých na nabíjacom mieste. Populárne nabíjacie miesta sú menej náchylné na náhodné fluktuácie v používaní v porovnaní s nabíjacími miestami, ktoré sú často používané, no iba malou skupinou používateľov EV. Je to najmä z dôvodu vyššej variability používateľov a tým pádom nižšieho vplyvu nových alebo jednorázových návštevníkov, čo prídu nabíjať. Navyše verejné investície do populárnych nabíjacích miest môžu byť považované za sociálne férovejšie.

- *Doba nabíjania:* Doplnkovou informáciou o využívaní nabíjacieho bodu je doba nabíjania, t.j. čas požadovaný na prenos energie medzi nabíjacou infraštruktúrou a vozidlom. Dlhá doba nabíjania indikuje vysoké využitie nabíjacieho miesta, ale môže taktiež indikovať, že výkon nabíjacieho miesta by mal byť navýšený.

- *Relatívny čas nabíjania:* V prípade, že používatelia EV nechávajú vozidlo pripojené dlhšie, ako doba potrebná na nabitie, nabíjacie miesto nie je efektívne využívané. Takéto správanie môže byť motivované parkovaním zadarmo na nabíjacích miestach a limituje prístup k nabíjacím miestam pre ostatných používateľov EV. Relatívny čas nabíjania je číslo medzi 0 a 1, ktoré je podielom času nabíjania a času pripojenia. Na jednej strane môže nabíjacie miesto s vysokým relatívnym časom nabíjania

Indikátory výkonnosti nabíjacích staníc	R^2
Spotrebovaná energia [kWh]	0.44
Počet nabíjacích transakcií	0.50
Popularita (počet unikátnych RFID kariet)	0.60
Doba nabíjania [hodiny]	0.48
Pomer nabíjania (doba nabíjania podelená dobou pripojenia)	0.40
Relatívna obsadenosť nabíjacieho miesta (doba nabíjania podelená celkovým časom dostupnosti)	0.42
Relatívne využitie kapacity nabíjacieho miesta (spotrebovaná energia podelená maximálnou menovitou energiou)	0.38

Tabuľka 8: R^2 indikátorov výkonnosti vysvetlených metódou OLS pomocou prediktorov charakterizujúcich okolie a ľudské aktivity v blízkosti nabíjacích miest.

dosiahnuť vyšší zisk, keďže je viac využívané používateľmi EV. Avšak na druhej strane, nízky relatívny čas nabíjania znamená vyšší potenciál pre smart charging.

- *Relatívna obsadenosť nabíjacieho miesta:* Ak je vysoká obsadenosť nabíjacieho miesta, ostatní používatelia EV sú často prinútení hľadať iné alternatívy pre nabitie svojho EV, čo znižuje vnímanú kvalitu poskytovaných služieb. Relatívna obsadenosť nabíjacieho miesta je daná podielom dĺžky doby, kedy je nabíjacie miesto obsadené dĺžkou pozorovanej periódy.

- *Relatívne využitie kapacity nabíjacieho miesta:* To, ako veľmi nabíjacie miesto zaťažuje elektrickú sieť môže byť odhadnuté pomerom spotrebovanej energie a nominálnou kapacitou nabíjacieho miesta [68]. Nominálna kapacita nabíjacieho miesta je teoretická hodnota spotrebovanej energie, keby sa na danom mieste nabíjali EV na maximálnom možnom výkone počas celej pozorovanej periódy. Tento indikátor môže informovať operátorov elektrickej siete o miere využitia kapacity siete, ktorá by eventuálne mohla byť využitá ostatnými spotrebiteľmi. Okrem toho je aj jeden z indikátorov, ktorý môže pomôcť operátorom nabíjacej infraštruktúry lepšie pochopiť využívanie nabíjacích miest.

Hodnoty všetkých vybraných indikátorov výkonnosti nabíjacích miest boli vypočítané z EVnetNL datasetu berúc do úvahy rok 2015. Pre analýzu miery vysvetlenia navrhovaných indikátorov výkonnosti prediktormi získanými z GIS dát, aplikujeme OLS metódu na každý z indikátorov oddelene. V tabuľke 8 sú zobrazené hodnoty R^2. Najvyššia hodnota R^2 a teda najvyšší potenciál pre analýzu dát má popularita nabíjacích miest (vyjadrená unikátnym počtom RFID kariet, ktorými sa iniciuje nabíjanie). Preto sústredíme ďalšie

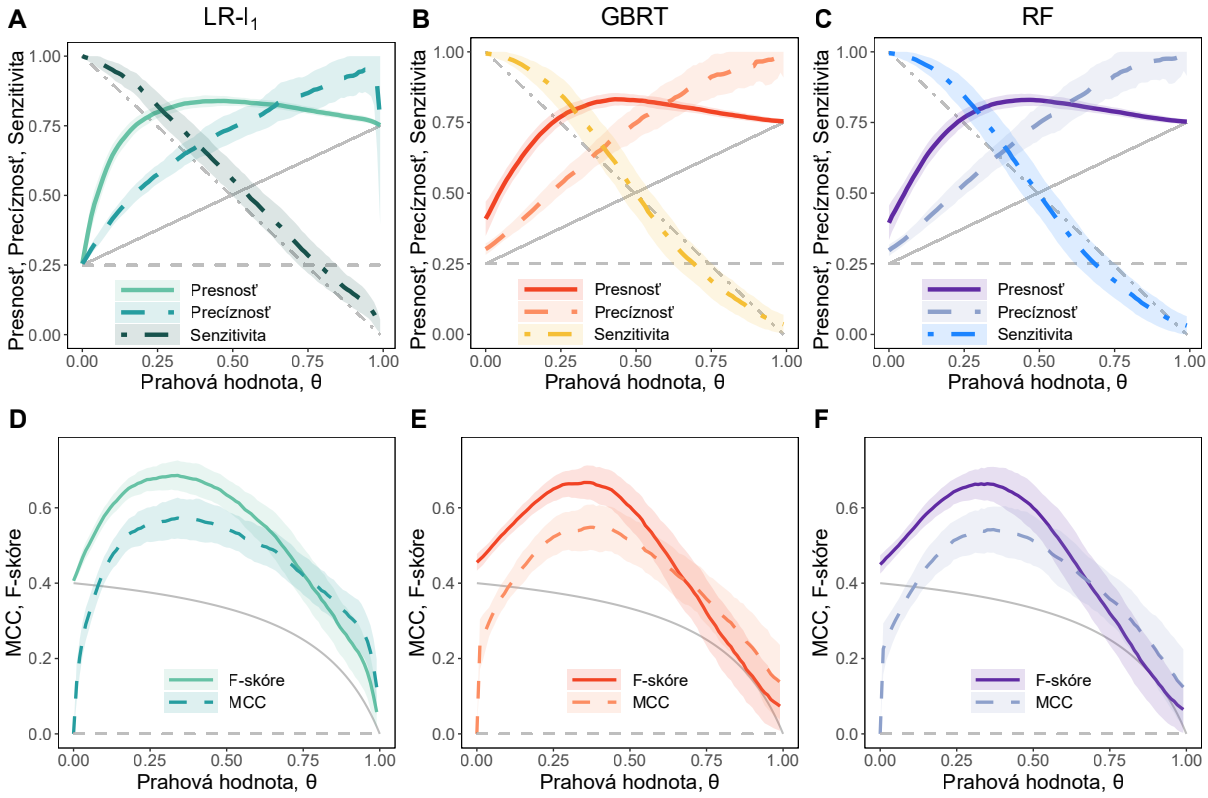
analýzy na tento indikátor. Pri plánovaní umiestnenia novej infraštruktúry potrebujeme často vyberať z konečnej množiny kandidátov umiestnenia (lokácií, kde je prípustné nainštalovať nabíjaciu infraštruktúru s ohľadom na vlastníctvo pôdy, dodávku energie, potenciál na prilákanie dostatočného dopytu atď.). V takej situácií nie je nevyhnutné, na ohodnotenie kandidátov umiestnenia, odhadnúť presný počet používateľov EV, ktorých dané nabíjacie miesto priláka, ale postačuje aj predpovedať, či bude dané nabíjacie miesto populárne. Preto modelujeme popularitu ako binárnu veličinu, a z tohto dôvodu, redukujeme problém, ktorému sa budeme venovať v tejto podkapitole, na binárnu klasifikáciu. Na jej riešenie použijeme logistickú regresiu regularizovanú l_1 normou, GBRT a RF. Keďže zvolené metódy vracajú pravdepodobnostnú predpoveď, môžeme použiť hodnotu prahového parametra θ pre získanie výsledných binárnych hodnôt. Ak je $\hat{y} \geq \theta$, predikcia bude 1, inak 0.

Nastavenie parametrov

Polomer kruhovej zóny, reprezentujúcej okolie nabíjacieho miesta, sme vyberali z množiny {100, 150, 200, 250, 300, 350, 400, 450, 500} metrov, kde prediktory vysvetľovali popularitu, pričom najlepšia miera vysvetlenia vyjadrená hodnotou R^2 bola dosiahnutá pre $r = 350$.

V experimentoch sme do trénovacej a testovacej množiny priradili 80 % a 20 % pozorovaní v takomto poradí, s použitím stratifikovaného delenia, podľa hodnôt vektora výstupu. Aby sme dosiahli spoľahlivejšie výsledky, vyhodnocujeme variabilitu metód merania chýb analyzovaním výstupov 100 modelov, trénovaných na 100 rôznych vzorkách trénovacej a testovacej množiny. Pri vyhodnocovaní mier chybovosti, meníme prahovú hodnotu θ v rozsahu od 0 do 0.99 v krokoch po 0.01. Hyperparameter $\lambda_m^{CV} in$ zo vzťahu (27) bol určený pomocou stratifikovanej krížovej validácie z množiny hodnôt 10^i pre i v intervale od -4 po 0 v krokoch po 0.02, ako hodnota, kde bola dosiahnutá najmenšia hodnota AUC. Stratifikovanú krížovú validáciu aplikujeme aj na stromové metódy, kde je potrebné určiť hodnoty hyperparametrov, ako miera zmršťovania, minimálna dĺžka listov a maximálny počet rozdelení. Hodnoty mier chybovosti boli vyčíslené aplikovaním trénovaných modelov na testovacie dáta. Počet vzoriek dát bootstrapom pri štatstickej inferencii sme nastavili na $b = 500$.

Popularitu sme zakódovali do binárneho vektora y. Najskôr sme vyextrahovali popularitu nabíjacích bodov, t.j. počet použitých unikátnych RFID kariet. V druhom kroku sme



Obrázok 16: Priemerná hodnota presnosti, precíznosti a senzitivity (**A**-**C**) a F–skóre a MCC (**D**-**F**) vyhodnotená na testovacích dátach ako funkcia prahovej hodnoty θ. Panely **A**, **D** odpovedajú metóde LR-l_1, **B**, **F** metóde GBRT a **C**, **G** metóde RF. Každá miera je zobrazená iným štýlom čiar. Hrubé čiary odpovedajú priemerným hodnotám pre skupinu 100 rôznych trénovacích a testovacích rozdelení dát a plochy reprezentujú jednu smerodajnú odchýlku odhadnutú z dát. Tenké čiary ukazujú hodnoty mier pre nulový model predpovedajúci populárne nabíjacie miesta náhodne s pravdepodobnosťou 0.25.

zoradili nabíjacie miesta podľa popularity zostupne. Po tretie sme vrchných $z = 25$ % umiestnených nabíjacích označili hodnotou 1, zvyšných 75 % hodnotu 0. Vykonali sme aj experimenty s hodnotami z 15 %, 20 %, 30 % a 35 %, ale obdržali sme veľmi podobné výsledky ako pre hodnotu 25 %.

Pri výpočtoch sme použili l_1 regularizovanú logistickú regresiu implementovanú funkciou *cv.glmnet* v balíku R *glmnet* [44]. GBRT a RF sú implementované v MATLAB prostredí Statistic and Machine Learning Toolbox použitím funkcie *fitrensemble* na trénovanie modelov a funkcie *cvpartition* pre krížovú validáciu. Funkcia *fitrensemble* je použitá na nájdenie optimálneho rozdelenia priestoru prediktorov pre maximalizovanie prínosu štandardnej klasifikácie a deliacej techniky pre regresné stromy.

Miera	Funkcia
Presnosť	$0.25 + 0.5\theta$
Precíznosť	$1 - \theta$
Senzitivita	0.25
F–skóre	$\frac{1-\theta}{2.5-2\theta}$
MCC	0

Tabuľka 9: Funkčné formy popisujúce strednú hodnotu mier chybovosti nulového modelu ako funkciu prahovej hodnoty θ.

3.4.2 Výsledky predikcie popularity nabíjacích miest

Obrázok 16 zobrazuje priemernú presnosť, precíznosť a senzitivitu (**A**-**C**), F–skóre a MCC (**D**-**E**) pre všetky tri metódy. Kvalitu predikcie môžeme posúdiť porovnaním výsledkov s nulovým modelom, ktorý predikuje populárne nabíjacie miesta náhodne s pravdepodobnosťou 0.25, odpovedajúcej zvolenej hodnote z. Odvodené funkčné vzťahy vyjadrujúce hodnoty mier chybovosti nulového modelu sú uvedené v Tabuľke 9 a sú zobrazené v Obrázku 16 tenkými čiarami.

Všetky tri metódy prekonali nulový model v mierach presnosti a precíznosti na celom rozsahu hodnôt θ. Senzitivita nevyhnutne klesá s prahovou hodnotou θ. Pre hodnoty θ väčšie ako 0.5, senzitivita klesá pod 0.5, čo môže mať za dôsledok nízku aplikovateľnosť predikcií, keďže by bol nízky podiel populárnych miest predikovaných správne. Ak θ rastie, senzitivita môže dosiahnuť nulové hodnoty aj pre ideálny model. Ak je θ vyššia ako 0.5, pre metódy rozhodovacích stromov je senzitivita menšia ako nulový model, čo potvrdzuje, že taká prahová hodnota je príliš vysoká.

V literatúre existuje množstvo prístupov na výber prahovej hodnoty θ. Napríklad, vhodný postup ako stanoviť rovnováhu medzi presnosťou, precíznosťou a senzitivitou môžeme nájsť v [59]. My sme vybrali dve doplnkové miery, MCC a F–skóre, ktoré sú vyhodnotené v obrázku 16. Obe miery dosahujú jedno maximum, ktoré je tiež globálne maximum. V tomto maxime miery ukazujú na zaujímavú kombináciu hodnôt, hlavne pre aplikáciu riešení do praxe. Prahové hodnoty prislúchajúce maximu označujeme ako $\theta_{MCC_{max}}$ a $\theta_{F\text{–}skóre_{max}}$. Hodnoty presnosti, precíznosti a senzitivity pre $\theta_{MCC_{max}}$ a $\theta_{F\text{–}skóre_{max}}$ sú uvedené v Tabuľke 10.

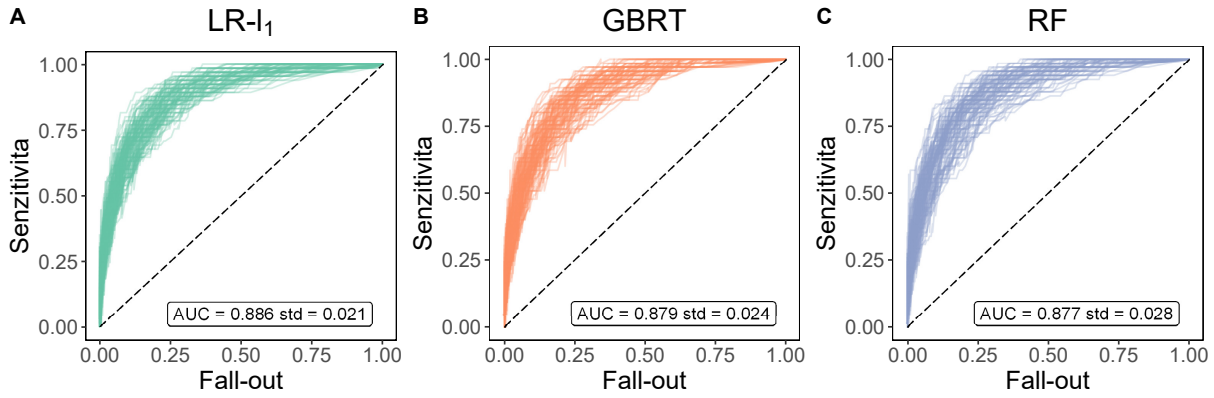
Pri rozhodovaní sa o rozšírení existujúcej nabíjacej infraštruktúry, je často potrebné

	LR-l_1	GBRT	RF
$\theta_{MCC_{max}}$	0.34	0.35	0.37
Presnosť	0.829	0.823	0.819
Precíznosť	0.648	0.644	0.632
Senzitivita	0.728	0.694	0.701
MCC	0.571	0.548	0.542
$\theta_{\text{F–skóre}_{max}}$	0.34	0.35	0.35
Presnosť	0.829	0.813	0.813
Precíznosť	0.633	0.614	0.615
Senzitivita	0.750	0.736	0.727
F–skóre	0.685	0.667	0.664

Tabuľka 10: Priemerné hodnoty presnosti, precíznosti a senzitivity z Obrázka 16 pre všetky tri metódy odpovedajúce maximálnej hodnote MCC miery a maximálnej hodnote F–skóre.

zobrať do úvahy protichodné faktory. Záujmové osoby, ako napr. prevádzkovatelia nabíjacej infraštruktúry a samosprávy, si vyberú prahovú hodnotu θ na základe ich očakávaní a postoja k riziku. Čím nižšia je použitá hodnota θ, tým pravdepodobnejšia je úspešná identifikácia populárnych lokácií, pričom sa ale zvyšuje riziko umiestnenia nabíjacieho miesta do nepopulárnych oblastí. Naopak čím vyššia je použitá hodnota θ, tým vyššia je záruka identifikácie skutočne populárnych nabíjacích miest s nevýhodou možnosti prehliadnutia potenciálne populárnych lokácií. Preto je potrebné zvoliť vhodný kompromis medzi veľkosťou parametra θ na základe požadovaného výsledku. Na základe pozorovaných hodnôt mier chybovosti odporúčame prahovú hodnotu θ nastaviť v rozsahu od 0.3 po 0.45, kde sú precíznosť aj senzitivita relatívne vysoké. Porovnaním nulového modelu a mier uvedených v tabuľke 10 môžeme uzavrieť, že faktory okolia a charakteristiky nabíjacích miest obsahujú prediktívnu silu umožňujúcu čiastočne vysvetliť popularitu nabíjacích staníc. Vo všeobecnosti to, ktoré hodnoty mier úspešnosti modelu sa dajú považovať za dobré, závisí aj od aplikačnej oblasti [55, s .70]. Uvažujúc dostupné dátové analýzy, týkajúce sa ľudského rozhodovania v podobných oblastiach, ako napr. systémy zdieľania bicyklov [139, 21], môžu byť hodnoty presnosti presahujúce hodnotu 0.8, kým sú precíznosť a senzitivita väčšie ako 0.65, považované za priaznivé.

Už jednoduchý náhľad na Obrázok 16 indikuje, že všetky tri metódy dosahujú veľmi podobné výsledky. Na vyhodnotenie, či sú výsledky štatisticky odlišné, testujeme rozdiely vo vzorkách AUC (plocha pod ROC krivkou). ROC krivky prislúchajúce k 100 rozdielnym trénovacím a testovacím rozdeleniam dát zobrazujeme v Obrázku 17. V štatistických testoch uvažujeme hladinu významnosti $\alpha = 0.01$.



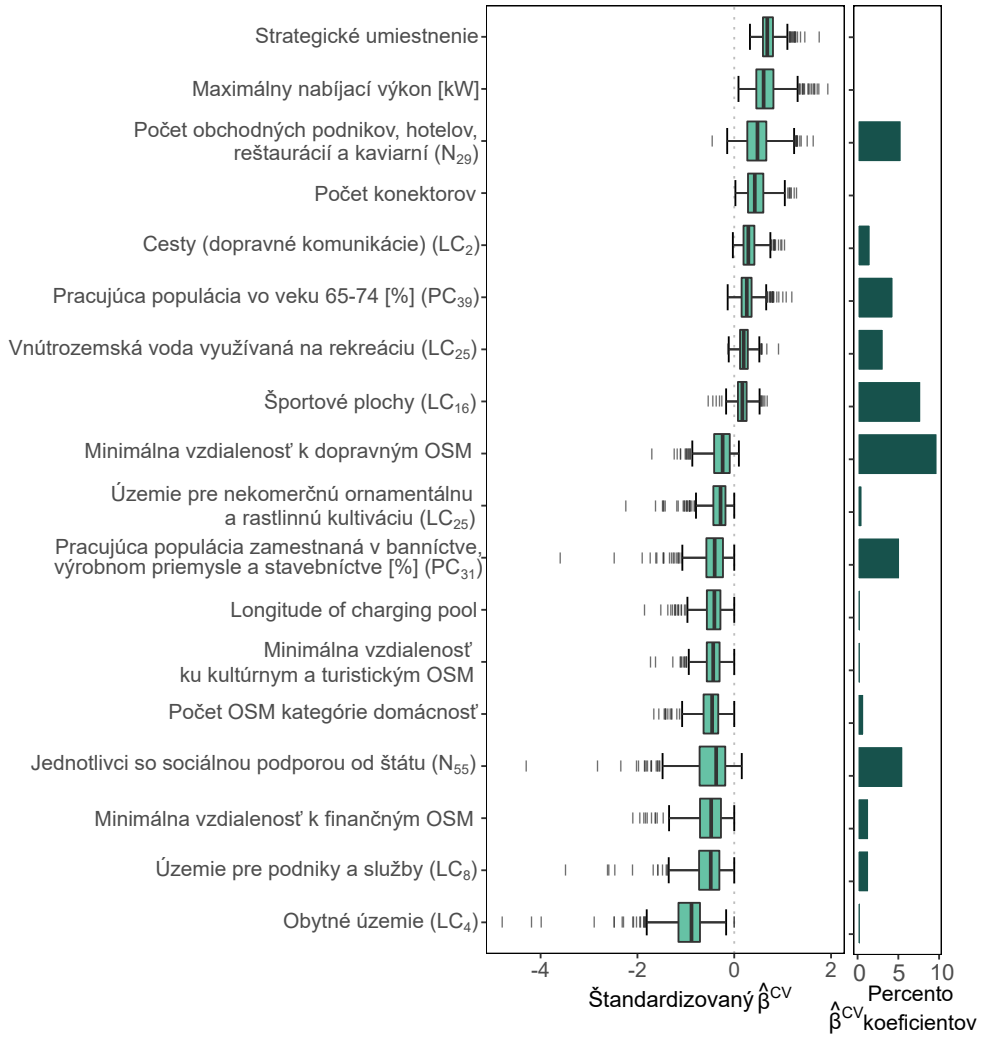
Obrázok 17: Vzorka 100 ROC kriviek, každá prislúchajúca rozličnému rozdeleniu dát na trénovaciu a testovaciu množinu. Čiarkovaná čiara zobrazuje výsledok odpovedajúci náhodnému klasifikátoru. Hodnota plochy pod krivkou ROC (AUC) a prislúchajúce smerodajné odchýlky (std) sú zobrazené v rámiku v spodnej časti grafu.

V prvom kroku, sme testovali pomocou F–testu rovnosť rozptylov medzi výsledkami všetkých metód. Štatisticky signifikantný rozdiel sme identifikovali iba pre LR-l_1 a RF metódy, kde LR-l_1 má menší rozptyl ($p = 0.0017$). Na porovnanie stredných hodnôt, pre prípad, kedy nie sú rozptyly štatisticky rozdielne, sme použili t–test, v opačnom prípade sme použili Welchov t–test [127]. Priemer AUC LR-l_1 je väčší ako priemer RF ($p = 0.0094$) a nie je odlišný od priemeru GBRT ($p = 0.0379$). Priemerné hodnoty AUC metód RF a GBRT nie sú štatisticky odlišné ($p = 0.5091$). Na základe hodnôt AUC a štatistických testov je metóda LR-l_1 stabilnejšia ako RF a prekonáva túto metódu v priemerných hodnotách AUC. Na hladine významnosti $\alpha = 0.05$ má LR-l_1 signifikantne vyššie priemerné hodnoty AUC ako metódy GBRT a RF.

Charakteristiky ovplyvňujúce popularitu verejných nabíjacích miest

Pokutová funkcia l_1 použitá v metóde LR-l_1 umožňuje výber významných prediktorov. Toto je výhodou v porovnaní s rozhodovacími stromami, ktoré v našom prípade vracajú modely zahŕňajúce v priemere 169 prediktorov, čo robí tieto modely ťažko interpretovateľné. Z týchto dôvodov ďalej interpretujeme iba výsledky LR-l_1 metódy.

Pomocou metódy štatistickej inferencie podkapitoly 1.4.8 sme potom vytvorili 500 LR $l - 1$ modelov a získali z nich koeficienty. Takto získané koeficienty prezentujeme v Obrázku 18, pričom sme vybrali iba koeficienty, ktoré mali nenulovú hodnotu, t.j. prislúchajúce prediktory boli vybraté aspoň v 90 % z 500 modelov, t.j. boli nulové maximálne v 10 % vzoriek. Stĺpcový graf ukazuje percento modelov, v ktorých bol daný koeficient rovný hodnote 0.



Obrázok 18: V ľavej časti grafu je ze zobrazený Tukey box-plot štandardizovaných koeficientov $\hat{\beta}_j^{CV}$ získaný na základe 500 $LR-l1$ modelov. Koeficienty sú usporiadané na základe mediánov ich hodnôt získaných jednotlivými modelmi. Zobrazujeme iba prediktory, ktoré boli vybrané najmenej v 90 % modelov. Stĺpcový graf vpravo zobrazuje percento modelov, v ktorých vyšiel štandardizovaný koeficient $\hat{\beta}_j^{CV}$ nulový, t.j. j-ty prediktor nebol vybratý.

Prediktory vybrané pomocou LR-$l-1$ modelov a štatistickou inferenciou môžu byť kategorizované do troch skupín:

- **Funkcia geografického územia** (v blízkosti nabíjacieho miesta): je signifikantný počet podnikov zameraných na obchod, počet hotelov, reštaurácií a kaviarní, vnútrozemské vodné plochy určené na rekreáciu, športové plochy a cesty, všetky s pozitívnym vplyvom. Minimálna vzdialenosť ku kultúrnym a turistickým a finančným OSM objektom mala negatívny koeficient, čo znamená, že čím je taký prediktor vzdialenejší od nabíjacieho miesta, tým klesá popularita nabíjacieho miesta. Ak je takýto objekt v blízkosti nabíjacieho miesta, tak môže zvyšovať jeho popularitu.

  Naopak, obytné územia, územia s nekomerčnou okrasnou a záhradnou kultiváciou a prítomnosť OSM prediktorov súvisiacich s domácnosťou indikujú znižovanie popularity nabíjacích miest. Tieto zistenia súvisia s intuíciou, že v obytných územiach sú nabíjacie miesta navštevované homogénnejšou skupinou ľudí ako v rušných urbánnych územiach, pretože v blízkosti obydlí nabíjajú zvyčajne rovnakí ľudia, narozdiel od, napríklad, rušnejších mestských centie. Takisto, najpravdepodobnejšie vysvetlenie negatívneho vplyvu firemného a industriálneho územia je pracovné nabíjanie [90], t.j. nabíjanie flotily firemných áut alebo zriedkavé využívanie nabíjacích miest (malou skupinou) zamestnancov prichádzajúcich do práce.

- **Charakteristiky populácie** (žijúcej v okolí nabíjacieho miesta): Populačná skupina pozitívne spätá s popularitou nabíjacích miest sú pracujúci starší ľudia vo veku od 65 do 74 rokov, ktorí sú považovaní za častých vlastníkov EV. Naopak, popularita je negatívne spätá s percentom obyvateľstva pracujúcom v baníctve, výrobnom priemysle a stavebnom sektore, ako aj s počtom osôb závislých na sociálnej podpore od štátu. Tieto výsledky naznačujú, že ekonomická prosperita v blízkosti nabíjacích staníc ovplyvňuje návštevnosť nabíjacích bodov.

- **Charakteristiky nabíjacích miest**: populárne nabíjacie miesta sú pravdepodobne rozmiestnené na základe strategického umiestnenia, majú vyšší maximálny výkon a viac nabíjacích bodov. Negatívny vplyv zemepisnej dĺžky môže byť vysvetlený geografiou Holandska, kde je v západnej časti krajiny vyššia urbanizácia keďže sa tam nachádza väčšina veľkých miest.

V tejto podkapitole sme na základe aspektov vyjadrujúcich výkonnosť nabíjacej in-

fraštruktúry definovali sedem indikátorov výkonnosti nabíjacích staníc. Ako cieľ sme si stanovili úlohu pokúsiť sa predpovedať popularitu nabíjacích staníc, ktorú sme reprezentovali binárnou premennou. Použili sme tri metódy na predikovanie, pričom najlepšie výsledky, no nie výrazne odlišné od ostatných metód, dosiahla metóda $LR - l_1$. Takisto sme interpretovali prediktory, ktoré potencionálne vplývajú na popularitu. Podkapitola vychádza z vlastnej publikácie [107].

3.5 Modelovanie spotreby energie

Táto podkapitola je primárne venovaná analýze rozdelenia spotreby elektrickej energie medzi jednotlivé nabíjacie miesta, pomocou ich okolia, za účelom získania prediktorov vysvetľujúcich toto rozdelenie.

Môžeme si všimnúť podobnosť s predošlou kapitolou, avšak táto sa primárne zameriava na spotrebu elektrickej energie, ako aj využíva stratifikáciu dát pre lepšie charakterizovanie vybratých skupín staníc.

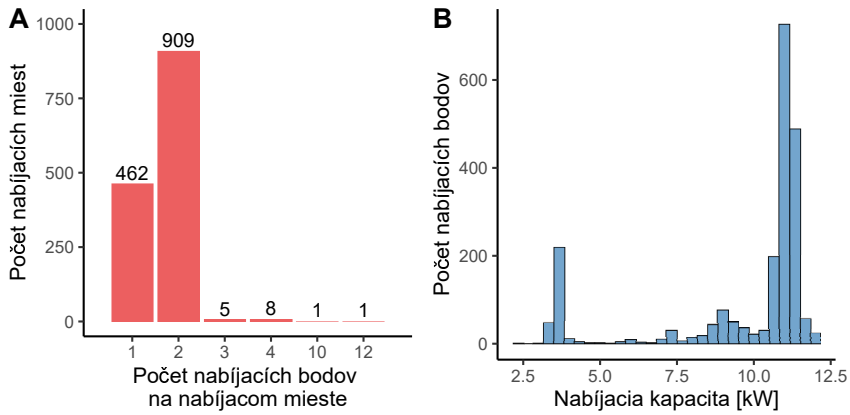
Najskôr si ukážeme vysporiadanie sa s možnými problémami regresie ako multikolinearita a vplyvné pozorovania (kap. 1.4.5). Potom vyberieme vhodnú reprezentáciu spotreby energie pre nabíjacie miesta. Následne analyzujeme pravdepodobnostné rozdelenie spotreby elektrickej energie a vplyv charakteristík nabíjacích miest na spotrebu. Pomocou metód na výber premenných analyzujeme prediktory vplývajúce na spotrebu energie, dodatočne analyzujeme vplyv prediktorov pre energiu spotrebovanú na nabíjacích miestach. Následne skúšame aj vplyv prediktorov na rôzne podskupiny staníc.

3.5.1 Dodatočná príprava dát pre regresiu

Pre lepšie vysvetlenie pomocou okolia stanice, ako aj kvôli faktu, že charakteristiky nabíjacej infraštruktúry nie sú vopred známe, budeme uvažovať za maticu prediktorov dáta bez EVnetNL dát nabíjacej infraštruktúry. Experimentami sme zistili, že v tomto prípade je lasso citlivejšie na multikolinearitu oproti LR-$l-1$ metóde. Preto sme navyše k odstráneniu korelácií v podkapitole 3.3 časti "Predspracovanie dát pre regresné a stromové metódy", zmiernili aj vplyv multikolinearity, ktorú sme detegovali pomocou VIF. Iteratívne sme odstraňovali prediktory s najvyššou hodnotou VIF, ktorá sa zakaždým prepočítala, až kým nemali všetky VIF hodnotu menšiu 10. Podrobnejší popis nájdeme v prílohe ***A-4***.

Testovali sme aj, či jednoduché transformácie ($\sqrt{y}$, y^2, $\log(y)$) a Box-Cox transformácia vektora výstupu y môžu vylepšiť lineárny model odhadnutý OLS pomocou $\boldsymbol{X}$. Na základe vzoru v rezíduách sme vybrali $log(y)$ transformáciu. Preto budeme v modelovaní odteraz uvažovať logaritmicky transformovanú energiu, ak nebude povedané inak. Podrobnejší popis nájdeme v prílohe ***A-5***.

Vplyvné pozorovania, stručne povedané nabíjacie miesta s nežiadúcou veľkosťou vplyvu hodnôt prediktorov a pozorovaní na regresiu, detegujeme pomocou Cookovej vzdialenosti.



Obrázok 19: **A** Počet nabíjacích miest v prevádzke 1. januára 2015 s daným počtom nabíjacích bodov, odhadnutý pomocou prvej a poslednej transakcie nabíjacieho miesta. **B** Histogram nabíjacej kapacity nabíjacích bodov odhadnutý z 15-minútových odčítaní z merača energie nabíjacej stanice (Meterreadings datasetu).

Na základe hodnôt tejto vzdialenosti sme odstránili 8 pozorovaní a MSE lineárneho modelu kleslo o 2.4 %.

Nakoniec sme po predspracovaní získali 1259 pozorovaní a 119 prediktorov z GIS dát a 5 prediktorov EVnetNL dát. Pri aplikácií OLS metódy na dáta po tomto predspracovaní dostaneme $R^2 = 0.435$. Výsledná matica $\boldsymbol{X}$ je teda o rozmere $n = 1259$, $p = 119$, kde pre lepšie zachytenie vplyvu neuvažujeme EVnetNL prediktory.

3.6 Základné charakteristiky nabíjacích miest

Prehľad charakteristík nabíjacej infraštruktúry po očistení dát zobrazuje obrázok 19.

3.6.1 Nastavenie parametrov

Veľkosť kruhovej zóny sme volili z množiny hodnôt od 100 metrov inkrementujúcich sa o 50 metrov do 500 metrov. Najlepšie vysvetliteľný, na základe R^2, v regresii y a $\boldsymbol{X}$ bola kruhová zóna s polomerom $r = 350$ metrov. Pre krížovú validáciu používame $k = 10$. Pre hyperparameter λ, ktorého hodnotu hľadáme krížovou validáciou, sme stanovili hodnoty 10^i pre i v intervale od -4 po 0 v krokoch po 0.02.

3.6.2 Metriky spotrebovanej energie nabíjacích miest

Nabíjacie miesta sa líšia v maximálnej kapacite a v počte nabíjacích bodov (19), čo môže viesť k rozdielom v spotrebe energie ako je ilustrované na obrázku 20A, nabíjacie miesta s vyšším počtom nabíjacích bodov zvyknú mať vyššiu spotrebu energie. Navyše, počet

Vektor výstupu charakterizujúci nabíjacie miesto	R^2
Spotrebovaná energia	0.435
Spotrebovaná energia na nabíjací bod	0.428
Maximálna spotrebovaná energia na nabíjací bod	0.398
Spotrebovaná energia na jednotku nabíjacej kapacity	0.391

Tabuľka 11: Rôzne vektory výstupu charakterizujúce spotrebovanú energiu na nabíjacom mieste. Koeficient determinácie, R^2, bol získaný odhadom metódou OLS logaritmicky transformovaného vektora výstupu a maticou prediktorov.

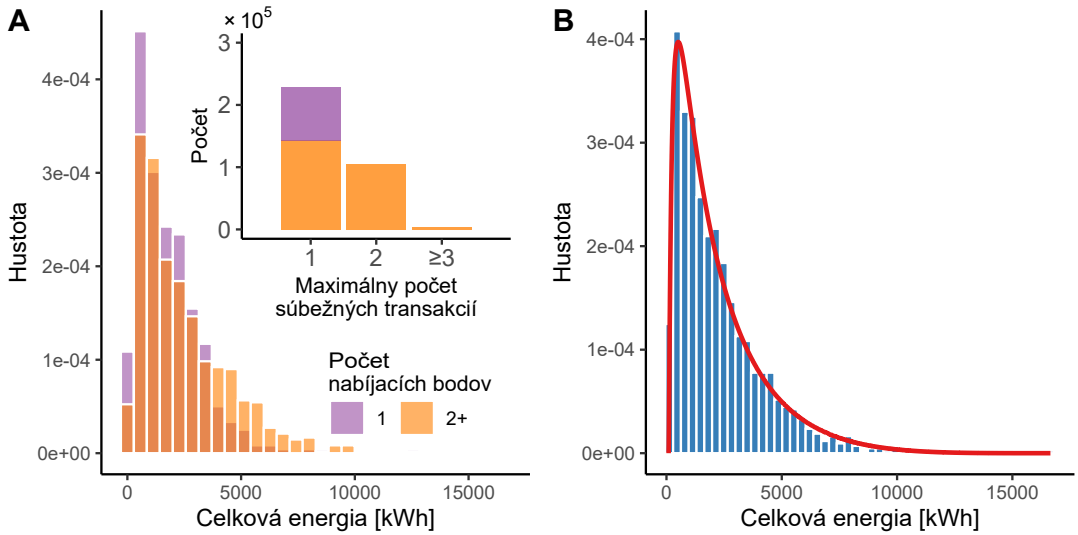
prípadov, kedy dve transakcie bežia na nabíjacom mieste paralelne nie je zanedbateľný (stĺpcový graf 20(A)). Aby sme analyzovali, do akej miery je potrebné počítať s týmito efektami, definujeme v tabuľke 11 niekoľko jednoduchých metrík spotrebovanej energie na nabíjacích miestach. Aby sme zistili, či niektoré metriky vieme lepšie vysvetliť pomocou matice prediktorov, používame metódu OLS. Hodnoty R^2 uvádzame v tabuľke 11. Najvyššie R^2 dostávame pre spotrebovanú energiu na nabíjacom mieste. Vysokú podobnosť hodnôt R^2 prisudzujeme vysokej korelácií medzi vektormi výstupu. Pearsonov koeficient korelácie pre všetky páry vektorov výstupu má rozsah od 0.86 po 0.98. Pre ďalšie analýzy preto ostáva spotrebovaná energiu na nabíjacom mieste ako vektor výstupu y.

3.6.3 Modelovanie rozdelenia spotreby energie

Histogram hustoty indikuje, že spotrebovaná energia na nabíjacích miestach je heterogénna a pozitívne vychýlená (viď obrázok 20B). Vysokú spotrebu pozorujeme na malej skupine nabíjacích miest a nízku spotrebu pozorujeme na veľkej skupine nabíjacích miest.

Nech Y je náhodná premenná s hustotou pravdepodobnosti $f_Y(y)$, modelujúca rozdelenie spotreby energie na nabíjacích miestach. Pre odhad hustoty $f_Y(y)$ uvažujeme transformáciu $g(y)$ vektora výstupu y (funkcia $g()$ je aplikovaná na y po prvkoch) a definujeme náhodnú premennú $Z = g(Y)$ s funkciou hustoty $f_Z(z)$. Potom náhodná premenná Z modeluje vektor $z = g(y)$. Pre odhadnutie rozdelenia vektora z pomocou funkcie $f_Z(z)$, uvažujeme tri bežné pravdepodobnostné rozdelenia (Weibullove, beta a gama). Na odhad používame metódu maximálnej vierohodnosti a metódu momentov. Metóda momentov poskytla lepší odhad parametrov pre beta a gama rozdelenie, zatiaľ čo pre Weibullove rozdelenie bola lepšia metóda maximálnej vierohodnosti. V tabuľke 12 zobrazujeme p–hodnoty Kolmogorov-Smirnov testu dobrej zhody aplikované na $f_Y(y)$ a y, využívajúc jednoduché transformácie $g(y)$ (y^2, y^3, $\sqrt{y}$, $\sqrt[3]{y}$ a $log(y)$).

Na prebádanie, ktoré hodnoty vektora výstupu y sú uspokojivo vysvetlené hustotou



Obrázok 20: Empirické rozdelenie pravdepodobnosti energie spotrebovanej nabíjacími miestami. **A** Zobrazujeme dve rôzne pravdepodobnostné rozdelenia, fialové pre nabíjacie miesta s jedným nabíjacím bodom a oranžové pre nabíjacie miesta s viac ako jedným nabíjacím bodom. Graf v pravej hornej časti tohto obrázku zobrazuje vrstvený stĺpcový graf počtu EV nabíjajúcich sa paralelne s ostatnými EV na nabíjacom mieste. **B** Empirické rozdelenie pravdepodobnosti celkovej energie spotrebovanej na nabíjacích miestach. Červená čiara reprezentuje distribučnú funkciu rozdelenia s rovnicou (40).

pravdepodobnosti $f_Y(y)$, je odporúčané kombinovať Kolmogorov-Smirnov test s grafickými metódami [126], napríklad P-P a Q-Q grafmi. Na obrázkoch 25 a 26 v prílohe ***B-2*** zobrazujeme P-P a Q-Q grafy získané pomocou y and $f_Y(y)$ pre kombinácie rozdelení a transformácií, ktoré mali v tabuľke 12 štatisticky signifikantný test.

Žiadne z rozdelení nezachytáva dáta najlepšie v celom rozsahu spotrebovanej energie. Najväčšie p-hodnoty boli nájdené pre $g(y) = \sqrt{y}$ a $g(y) = \sqrt[3]{y}$ v kombinácií s beta rozdelením a $g(y) = \log(y)$ v kombinácií s Weibullovým rozdelením. Tieto tri kombinácie nájdeme v paneloch A, B a I obrázkov 25 a 26 (prílohy ***B-2***). Q-Q obrázky indikujú, že beta rozdelenie vysvetľuje lepšie malé hodnoty ako Weibullovo rozdelenie. Niekoľko veľkých hodnôt nie je dobre zachytených ani beta ani Weibullovým rozdelením. Aj keď má $g(y) = \sqrt{y}$ kombinované s beta rozdelením vyššiu p-hodnotu ako $g(y) = \sqrt[3]{y}$, Q-Q graf indikuje, že horšie odhaduje nabíjacie miesta s nízkou spotrebou energie. Preto, berúc do úvahy p-hodnoty, Q-Q a P-P grafy, považujeme za najlepší odhad beta rozdelenie kombinované s transformáciou $g(y) = \sqrt[3]{y}$. Funkcia hustoty rozdelenia je odvodená v prílohe B a má tvar:

		Rozdelenie		
		Weibull	beta	gama
Transf., $g(y)$	y	0.093	0.000	0.026
	y^2	0.013	0.000	0.000
	y^3	0.019	0.000	0.000
	$\sqrt{y}$	0.093	0.924	0.163
	$\sqrt[3]{y}$	0.093	0.582	0.095
	$log(y)$	0.506	0.081	0.000

Tabuľka 12: P-hodnoty Kolmogorov-Smirnov testu dobrej zhody získané po aplikovaní testu na vektor výstupu y a funkcie hustoty $f_Y(y)$. Uvažujúc transformáciu $Z = g(Y)$, hustota $f_Y(y)$ je získaná z $f_Z(z)$. Hustota $f_Z(z)$ bola odhadnutá pre spotrebu energie štandardnými rozdeleniami (Weibullove, beta and gama). V niekoľkých prípadoch sme našli štatisticky významné výsledky (p-hodnota > 0.05). Testovali sme aj exponenciálne, normálne a log-normálne rozdelenia, ale metódy pre ne nenašli signifikantný odhad parametrov pre rozdelenie y.

Model	$\hat{k}$	R^2	priemer	stdev	kv
$y = kn$	8.10	0.90	8.68	3.64	0.42
$y = kt$	895.88	0.59	897.40	765.28	0.89
$y = kp$	632.94	0.56	678.85	589.68	0.87
$y = k(t \circ p)$	245.44	0.59	269.27	228.24	0.85
$y = k(n \circ p)$	2.49	0.95	2.52	0.63	0.25
$y = k(n \circ t)$	3.25	0.94	3.45	0.92	0.42

Tabuľka 13: Jednoduché regresné modely pre spotrebovanej energie na nabíjacích miestach. V stĺpcoch sú odhady regresného koeficientu $\hat{k}$; odpovedajúce hodnoty R^2 získané metódou OLS; priemerná hodnota kvantity reprezentovaná koeficientom k počítaná pre nabíjacie miesta (priemer); prislúchajúca smerodajná odchýlka (stdev) a koeficient variácie (kv) vypočítaný podelením smerodajnej odchýlky priemerom. Symbol $\circ$ označuje násobenie vektorov po prvkoch (Hadamardov súčin).

$$f(y, \alpha, \beta) = \frac{\left(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\right)^{\alpha-1} \left(1 - \frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\right)^{\beta-1}}{B(\alpha, \beta) 2 (y_{max} - y_{min}) y^{\frac{2}{3}}}. \tag{40}$$

Symbol $B(\alpha, \beta)$ označuje Beta funkciu a odhady parametrov nadobúdajú nasledovné hodnoty: $\alpha = 2.576$, $\beta = 4.528$, $y_{min} = 91.55$ kWh a $y_{max} = 16\ 649.40$ kWh.

3.6.4 Vysvetľovanie spotreby energie ostatnými indikátormi nabíjacích miest

Zaujímavé náhľady na charakteristiky nabíjacej infraštruktúry získame analyzovaním vzťahu medzi spotrebou energie a ostatnými charakteristikami nabíjacích miest tvoriacich spotrebu energie. Spotreba energie na nabíjacom mieste i môže byť dekomponovaná

na násobok troch iných indikátorov, t.j.

$$y_i = n_i t_i p_i, \tag{41}$$

kde n_i je počet nabíjacích transakcií, vykonaných na nabíjacom mieste i, t_i je priemerný čas nabíjania na transakciu nabíjacieho miesta i a p_i je priemerný nabíjací výkon nabíjacieho miesta i. Pre všetky nabíjacie miesta tieto kvantity organizujeme ako vektory n, t a p. Aby sme odhadli, akým spôsobom tieto tri faktory prispievajú k heterogenite spotrebovanej energie nabíjacími miestami, skúmame šesť modelov prezentovaných v tabuľke 13. Modely sú založené na rovnici (41), kde jeden alebo násobok dvoch indikátorov, reprezentované regresným koeficientom k, sú považované pre nabíjacie miesta za nemenné.

Odhad $\hat{k}$ koeficienta k sme získali pomocou metódy OLS. Medzi modelmi vysvetľujúcimi spotrebovanú energiu pomocou jedného indikátora má najvyššiu hodnotu R^2 model vysvetľujúci spotrebovanú energiu počtom nabíjacích transakcií. Podobne majú vysoké hodnoty R^2 aj modely, vysvetľujúce spotrebu energie z párov indikátorov zahŕňajúcich počet transakcií. Teda hlavný faktor spojený s rozdelením spotrebovanej energie medzi nabíjacie miesta je počet transakcií. Fluktuácie v nabíjacích vzoroch (t.j. priemerný nabíjací čas a nabíjací výkon) zohrávajú omnoho menšiu rolu.

V tabuľke 13 sme spočítali priemer, smerodajnú odchýlku a koeficient variácie pre všetky nabíjacie miesta pre kvantity reprezentované koeficientom k. Napríklad v modeli $y = kn$, koeficient k nahrádza v rovnici (41) výraz $t_i p_i$, ktorý môže byť interpretovaný ako priemerná spotrebovaná energia na transakciu. Priemerná energia spotrebovaná na transakciu je 8.68 kWh, priemerný čas nabíjania na transakciu je 2.52 hodín a priemerný výkon dosahuje 3.45 kW. Tieto čísla indikujú, že väčšina nabíjaných EV sú PHEV. Nabíjacia kapacita, ktorá je pre väčšinu nabíjacích bodov okolo 11 kW (viď obrázok 19), je značne nevyužitá. Hodnoty koeficientu variácie sú vysoké v prípadoch, kde je počet transakcií zahrnutý v analyzovaných kvantitách, potvrdzujúc, že najvyšší rozptyl je asociovaný s počtom transakcií.

Vzhľadom na to, že počet nabíjacích transakcií je úzko spojený so spotrebou energie na nabíjacích miestach, považujeme za kľúčové, lepšie pochopiť spôsob rozhodovania používateľov EV o možnostiach nabíjania. Predpokladáme, že niektoré exogénne faktory, charakterizujúce prostredie v okolí nabíjacích miest, majú vplyv na spotrebu energie na nabíjacích miestach, čo je predmetom skúmania v nasledujúcej sekcii.

3.6.5 Vysvetlenie spotreby energie na nabíjacích miestach pomocou GIS dát

V tejto časti aplikujeme metodológiu spracovania dát na maticu prediktorov $\boldsymbol{X}$ odvodenú z GIS dát.

Pomocou metódy štatistickej inferencie podkapitoly 1.4.8 sme potom vytvorili $b = 10\ 000$ lasso modelov a získali z nich koeficienty. Vyšší počet modelov oproti predošlej kapitole je potrebný najmä z dôvodou vyššej variability výsledkov pre túto metódu.

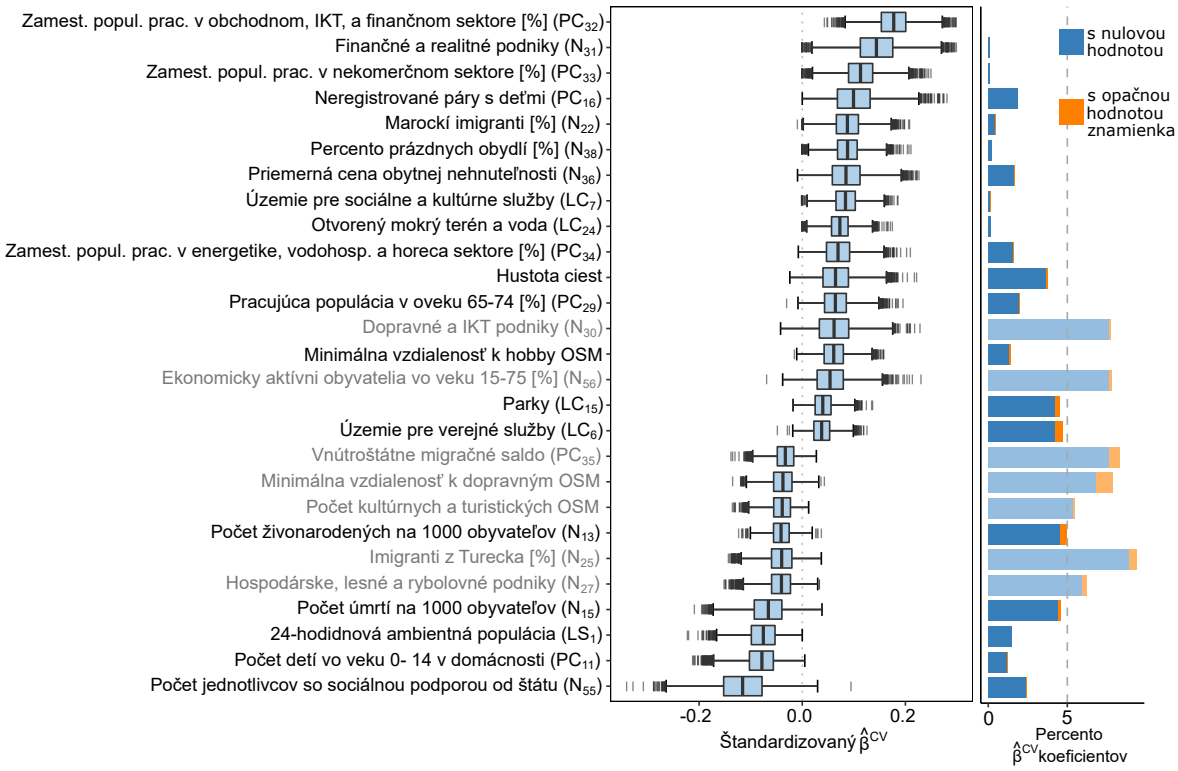
Pre podobnosť s p-hodnotou, považujeme za signifikantné tie prediktory, kde je počet bootstrapových vzoriek s nulovou hodnotou koeficientu menší ako 5 % a počet vzoriek s opačným znamienkom koeficienta je zanedbateľný. Aby sme ale poskytli širší náhľad na výsledky, ukazujeme v obrázku 21 empirické rozdelenia štandardizovaných regresných koeficientov, ktoré dosiahli nulovú hodnotu maximálne v 10 % realizácií bootstrapu. Pohľad na Tukeyho box-plot potvrdzuje potrebu použitia bootstrapu alebo všeobecnejšie, potrebu používať metódy štatistickej inferencie. Väčšina prezentovaných koeficientov mala v niekoľkých vzorkách nulovú hodnotu alebo výnimočne hodnotu s opačným znamienkom, čo by mohlo viesť k vynechaniu týchto koeficientov alebo ich interpretovaniu zlým spôsobom, ak by sme evaluovali iba jeden beh lasso metódy.

Zvolené regresné koeficienty môžu byť asociované s tromi rôznymi priestorovými škálami. Niektoré popisujú blízke okolie nabíjacích miest, napr. počet finančných a realitných podnikov (N_{32}), ostatné prislúchajú k obciam, napr. prediktory odvodené z datasetu Populačné jadrá ako napr. percento obyvateľov zamestnaných v nekomerčnom sektore (PC_{33}). Posledná skupina regresných koeficientov má potenciál na charakterizovanie lokácie nabíjacích miest na úrovni krajiny, napr. zemepisná šírka a dĺžka.

Regresné koeficienty zobrazené v obrázku 21 sú zoradené zostupne od najvyššej po najnižšiu hodnotu mediánu. Pre prehľadnosť sme usporiadali signifikantné prediktory s pozitívnym znamienkom mediánu do štyroch skupín:

- **Fyzické prostredie(+):** Územie pre sociálne a kultúrne služby (LC_7); Otvorený mokrý terén a voda (LC_{24}); Hustota ciest, Parky (LC_{15}); Územie, kde sú poskytované verejné služby (LC_6).

- **Populácia(+):** Populácia zamestnaná vo veľkoobchode a maloobchode, doprave, skladníctve, IKT službách a finančných službách [%] (PC_{32}); Populácia zamestnaná v nekomerčnom sektore [%] (PC_{33}); Slobodné páry s deťmi (PC_{16}); Imigranti



Obrázok 21: Empirické rozdelenia štandardizovaných regresných koeficientov získaných lasso metódou kombinovanou s 10 zložkovou krížovou validáciou aplikovanou na 10 000 vzorkách bootstrapovaných dát. Zobrazujeme iba prediktory, kde hodnota regresného koeficientu bola nastavená na nulu maximálne v 10 % vzoriek. Koeficienty sú zoradené zostupe od najvyššej hodnoty mediánu po najnižšiu. Ľavý panel zobrazuje Tukeyho box-plot koeficientov. Na pravo je umiestnený vrstvený stĺpcový graf, kde je zobrazené percento vzoriek, kedy bol regresný koeficient $\hat{\beta}^{CV}$ nastavený na nulu a počet vzoriek dosiahol opačné znamienko ako medián. Za signifikantné považujeme tie prediktory (indikované tmavo modrou farbou), kde je počet vzoriek s nulovým koeficientom menší ako 5 % a počet vzoriek s opačným znamienkom je nízky. Svetlo šedá čiarkovaná čiara indikuje 5 % prahovú hodnotu. *Zamest. popul. prac. - zamestnaná populácia pracujúca.

z Maroka [%] (N_{22}); Populácia zamestnaná v energetike, vodnom a odpadovom hospodárstve a horeca (hotely/reštaurácie/kaviarne) sektore [%] (PC_{34}); Pracujúca populácia vo veku 65-74 rokov [%] (PC_{29}).

- **Služby a podniky(+):** Finančné a realitné podniky (N_{31}); Minimálna vzdialenosť k hobby OSM.

- **Budovy(+):** Percento prázdnych obydlí [%] (N_{38}); Priemerná cena obytnej nehnuteľnosti v tisícoch eur (N_{36}).

Podobne sme organizovali aj signifikantné prediktory s negatívnou hodnotou mediánu:

- **Populácia(-):** Počet jednotlivcov poberajúcich sociálnu pomoc od štátu (N_{55}); Počet osôb vo veku 0 - 14 vo viacčlennej domácnosti s deťmi (PC_{11}); Ambientná populácia (LS_1); Počet úmrtí v roku 2015 na tisíc obyvateľov (N_{15}); Počet živonarodených detí v roku 2015 na tisíc obyvateľov (N_{13}).

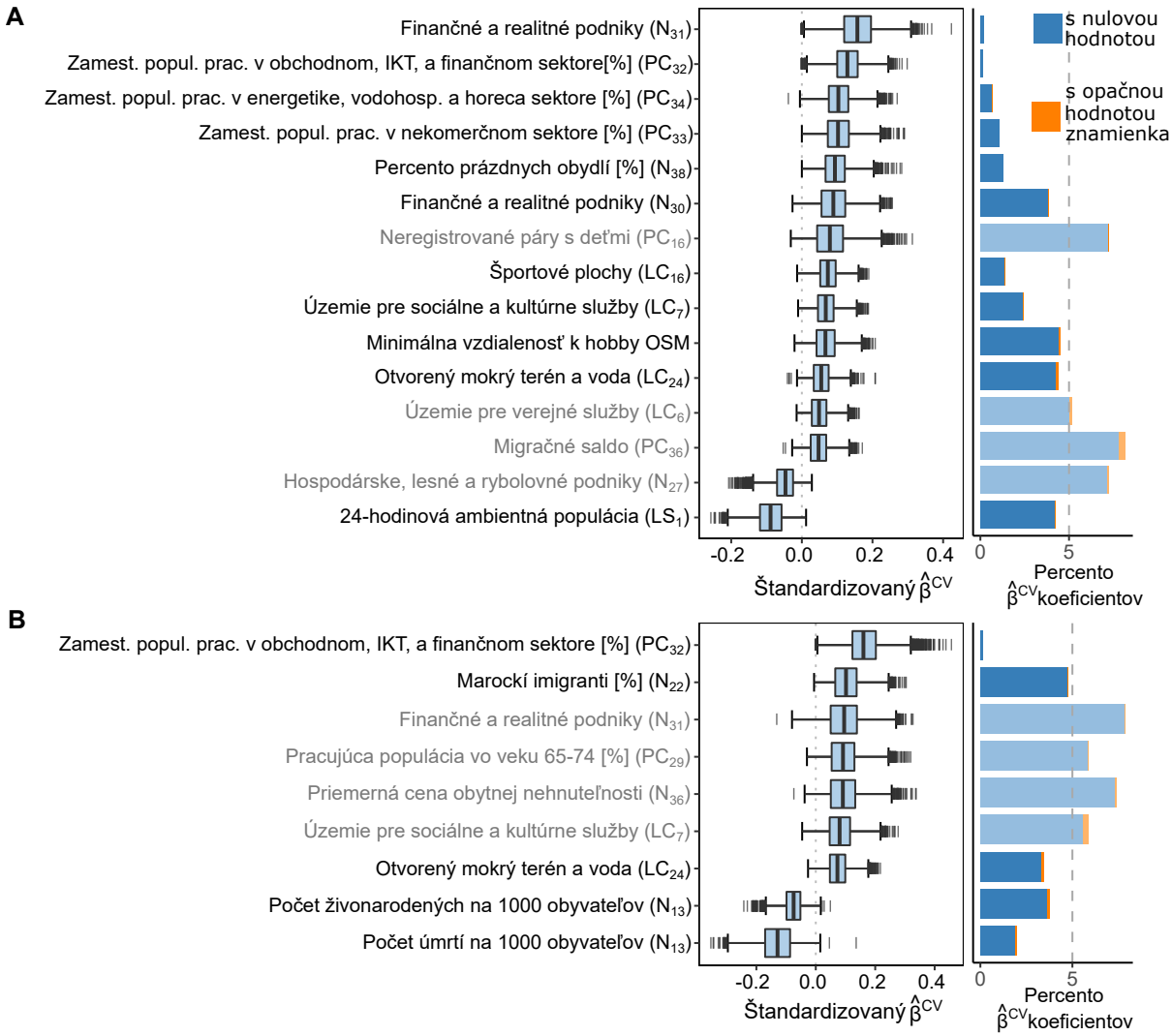
Najvyšší počet signifikantných prediktorov sme našli v skupine populácia, k čomu pravdepodobne prispieva aj najvyšší počet prediktorov v tejto skupine. Mnoho signifikantných prediktorov ukazuje na jeden faktor. Najväčšia skupina prediktorov, PC_{32}, PC_{29}, N_{36}, (N_{55}, N_{46}) indikuje, že vysoký (nízky) príjem a bohatstvo (chudoba) sú pozitívne (negatívne) spojené s množstvom spotrebovanej energie na nabíjacích miestach. Najpravdepodobnejšiou príčinou sú vysoké ceny EV, čo ich robí dostupné pre lepšie ekonomicky zabezpečených obyvateľov a podniky. Pravdepodobne z toho istého dôvodu sú niektoré signifikantné prediktory (PC_{11}, N_{13}) spojené s deťmi a mládežou a so staršou populáciou na dôchodku (N_{15}), t.j. so sociálnymi skupinami, ktoré zvyčajne nemajú vysoké príjmy a často sú odkázané na starostlivosť iných osôb. Vysoké percento prázdnych obydlí (N_{38}) a vysoká cena obytných nehnuteľností (N_{36}) sú asociované s vysokou spotrebou energie na nabíjacích miestach. Toto môže reprezentovať novo postavené a nie úplne obývané oblasti s vyšším štandardom života, vyjadreným cez vyššie hodnoty nehnuteľností. Ak má prediktor reprezentujúci minimálnu vzdialenosť k objektu pozitívny koeficient, potom spotreba energie rastie s rastúcou vzdialenosťou od daného objektu a naopak. A teda blízkosť bodov záujmu spojených s hobby je negatívne asociovaná so spotrebou energie na nabíjacích miestach. Navyše hustota ciest je tiež medzi prediktormi, ktoré sú pozitívne prepojené so spotrebou energie na nabíjacích miestach, čo naznačuje, že dobrý prístup k nabíjacím miestam prispieva k vyššej spotrebe energie.

Po zahrnutí piatich EVnetNL prediktorov (počet nabíjacích bodov, maximálny výkon, zemepisná šírka, zemepisná dĺžka a umiestňovacia stratégia) do matice prediktorov, sme analýzu opakovali. Výsledky sú zobrazené v obrázku 27 v prílohe ***B-3***. Vo všeobecnosti sú signifikantné prediktory podobné ako v obrázku 21; aj keď počet signifikantných prediktorov je nižší, čo pripisujeme nahradeniu niektorých prediktorov EVnetNL prediktormi. Napríklad, mokrý terén a voda (LC_{24}), prediktor signifikantný v modeli bez EVnetNL prediktorov, môže byť vyjadrený zemepisnou dĺžkou. Negatívny vplyv zemepisnej dĺžky môže byť vysvetlený geografickou polohou Holandska, kde je západná časť krajiny viac urbanizovaná a nájdeme tu väčšinu holandských miest. Zároveň je v tejto oblasti veľa povrchovej vody, keďže je zväčša situovaná pod úrovňou mora. Signifikantnosť niektorých prediktorov zo skupín Fyzické prostredie(+) a Populácia(-) bola redukovaná. Všetky EVnetNL prediktory sú signifikantné a maximálny nabíjací výkon, počet nabíjacích bodov a zemepisná dĺžka majú silný vplyv indikujúc potenciálny vplyv parametrov nabíjacích miest na spotrebu energie.

3.6.6 Vplyv stratégie umiestňovania na spotrebu energie

Väčšina nabíjacích miest bola umiestnená použitím jedným z dvoch (strategický alebo dopytovo orientovaný) spôsobov umiestňovania [52]. Strategicky umiestnené nabíjacie miesta sú umiestnené v blízkosti verejných lokalít, kde je intuitívne očakávané nabíjanie EV. Nabíjacie miesta umiestnené dopytovo orientovaným spôsobom sú budované na základe žiadosti od používateľov EV, zvyčajne v blízkosti ich obydlí. V tejto sekcii skúmame, či stratégia umiestňovania spôsobuje rozdiel vo faktoroch asociovaných so spotrebou energie. Informáciu o umiestňovacej stratégii z EVnetNL datasetu používame na rozdelenie nabíjacích miest do dvoch skupín. Lasso metódu aplikujeme na každú skupinu separátne. Vybraté prediktory v obrázku 22 sa do veľkej miery zhodujú s prediktormi vybratými pre kompletný dataset (viď obrázok 21).

Energia, spotrebovaná na strategicky umiestnených nabíjacích miestach (obrázok 22A), je pozitívne prepojená s pracovným sektorom obyvateľstva a fyzickým prostredím, t.j. s určitým typom lokalít v blízkosti nabíjacích miest. Pracovný sektor obyvateľov reprezentovaný prediktormi PC_{32}, PC_{33} a PC_{34} indikuje prevládajúce podniky v obciach, pozitívne prepojený so spotrebou energie. Navyše, vybrané prediktory pre strategické umiestňovanie staníc odkazujú na určité podniky a lokácie (športoviská, socio-kultúrne miesta), ktoré



Obrázok 22: Zobrazujeme iba prediktory, kde hodnota regresného koeficienta bola nastavená na nulu maximálne v 10 % vzoriek. Empirické rozdelenia štandardizovaných regresných koeficientov získaných lasso metódou kombinovanou s 10 zložkovou krížovou validáciou aplikovanou na 10 000 vzoriek bootstrapovaných dát. **A** Strategicky umiestnené nabíjacie miesta. **B** Dopytovo orientované umiestnené nabíjacie miesta. Zobrazujeme iba prediktory, kde hodnota regresného koeficienta bola nastavená na nulu maximálne v 10 % vzoriek. Koeficienty sú zoradené zostupe od najvyššej hodnoty mediánu po najnižšiu. Ľavý panel zobrazuje Tukeyho box-plot koeficientov. Napravo, vrstvený stĺpcový graf ukazuje percento vzoriek, kedy bol regresný koeficient $\hat{\beta}^{CV}$ nastavený na nulu a počet vzoriek, kedy dosiahol opačné znamienko ako medián. Za signifikantné považujeme tie prediktory (indikované tmavo modrou farbou), kde je počet vzoriek s nulovým koeficientom menší ako 5 % a počet vzoriek s opačným znamienkom je nízky. Svetlo šedá čiarkovaná čiara indikuje 5 % prahovú hodnotu. *Zamest. popul. prac. - zamestnaná populácia pracujúca.

môžu byť asociované s príležitostným nabíjaním.

Pre nabíjacie miesta s dopytovo orientovanou stratégiou umiestňovania, (obrázok 22B), negatívny koeficient počtu úmrtí na 1000 obyvateľov a živonarodených detí na 1000 obyvateľov indikujú, že územia s vyššou pôrodnosťou a úmrtnosťou sú negatívne spojené so spotrebou energie, ukazujúc na územia obývané sociálne slabšími skupinami.

Testovali sme aj ďalšie stratifikácie, podľa ktorých sme delili stanice do neprekrývajúcich sa skupín, napr. na základe počtu nabíjacích bodov, miery obytných zón (reprezentovanej podielom obytného územia v kruhovej zóne nabíjacieho miesta), administratívneho rozdelenia Holandska na 12 provincií a počtu obyvateľov v obciach.

Okrem posledného kritéria, sme obdržali nízky počet vybraných prediktorov. Delením nabíjacích miest na dve skupiny na základe populácie obce, v ktorej sú umiestnené, zvažujúc prahovú hodnotu 50 000 obyvateľov, sme obdržali dve, približne rovnako veľké skupiny. Zaujímavým zistením je, že nabíjacie miesta umiestnené v obciach s viac ako 50 000 obyvateľmi spotrebovali o 48 % viac elektrickej energie ako nabíjacie miesta druhej skupiny. Vyšší počet signifikantných prediktorov, spojených s populáciou v obciach, finančnými a realitnými podnikmi a fyzickým prostredím, nachádzame vo výsledkoch pre nabíjacie miesta umiestnené v obciach s nižším počtom obyvateľov (viď obrázok 28 v prílohe C).

Táto podkapitola sa venovala analýze rozdelenia spotreby elektrickej energie medzi nabíjacie miesta. Spotreba energie má pre nabíjacie miesta po transformácií beta rozdelenie a je najviac ovplyvňovaná počtom transakcií spomedzi základných identifikátorov tvoriacich spotrebu nabíjacích miest. Pri analýze vplyvu okolia sa prejavila najviac ekonomická prosperita obyvateľstva a firiem, či už negatívne alebo pozitívne. Takisto sme ukázali aj faktory vplývajúce na nabíjacie miesta podľa umiestňovacích stratégií. Tieto faktory môžu slúžiť najmä ako vstup pre optimalizácie. Ukázaná metodológia je vhodná najmä na identifikáciu faktorov vplývajúcich nielen na spotrebu elektrickej energie, ale aj na iné indikátory, pomocou ich okolia. Takisto ponúka vyhodnotenie signifikantnosti výsledkov po výbere premenných, čo je v literatúre často opomínané.

Záverečné zhrnutie výsledkov

V tejto práci sa zaoberáme skúmaním možností, ako využiť dátovú analýzu na podporu rozhodovania v oblasti elektromobility.

V prvej kapitole analyzujeme súčasný stav v problematike. Na začiatok sme definovali základné pojmy, motivovali elektromobilitu, ktorá je prepojením energetických a dopravných systémov a dátovú analýzu ako pilier pre rozhodovanie. Ďalej sme stručne opísali používané metódy na analýzu dát a použité dáta. Kapitolu sme uzavreli prehľadom literatúry v oblasti elektromobility so zameraním na analýzu dát.

V druhej kapitole sme priblížili tri hlavné ciele a metodiku práce. Pre prvý cieľ sme poskytli riešenie segmentáciou nabíjacích staníc na základe navrhnutých identifikátorov, využitím troch zhlukovacích metód, kde sme porovnávali aj dva prístupy zhlukovania kombinovaného s agregáciou dát. Podarilo sa nám identifikovať 4 zhluky staníc, ktoré sme interpretovali. Lepší z prístupov k zhlukovaniu bol ten, čo najskôr agregoval dáta a následne ich zhlukoval. Zhlukovanie nabíjacích staníc má potenciál prispieť k lepšiemu poznaniu vzorov využívania nabíjacích staníc a môže napomôcť vylepšiť výsledky modelovania indikátorov nabíjacej infraštruktúry pre pokročilejšie učiace sa algoritmy.

Druhý cieľ sme naplnili zostavením, otestovaním a vyhodnotením metód umožňujúcich predpovedať časopriestorovo agregovanú spotrebu nabíjacích staníc. Na základe prechádzajúcich výsledkov a odporúčaní v literatúre sme spotrebu priestorovo agregovali. Uvažovali sme dáta s dennou frekvenciou. Predbežná analýza odhalila v dátach sezónne vzory a aj z toho dôvodu sme využili SARIMAX, GBRT a RF metódy, ktoré sme vylepšili externými prediktormi. Modely sme trénovali tromi procedúrami odlišujúcimi sa v dĺžke trénovacích množín. Najlepšie predpovede dosiahla metóda SARIMAX, natrénovaná postupne pribúdajúcimi dátami, s presnosťou MAPE okolo 12 %, pričom všetky modely výrazne prekonali náhodný sezónny model.

Spôsob akým sme naplnili tretí cieľ sme opísali v troch podkapitolách. Prvá z týchto podkapitol sa venuje extrakcii prediktorov z GIS dát, reprezentujúcich okolie nabíjacích miest, a analyzuje schopnosti regresných metód, vyberať premenné v prostredí multikolinearity. Využili sme dve hlavné triedy takýchto metód: regularizované a krokové regresné metódy. Variovali sme stupeň multikolinearity a skúmali, ako sa s ňou jednotlivé metódy vedia vysporiadať. Najlepšie výsledky dosiahli metódy PACS a lasso, pričom metóda lasso

mala nižší výpočtový čas a aj nižšiu variabilitu výsledkov. Aj z tohto dôvodu sme metódu lasso a $l-1$ regularizáciu odporučili pre ďalšie použitie v rámci dizertačnej práce.

Predikovanie popularity nabíjacej infraštruktúry je jedným z hlavných bodov tretieho cieľa. Popularitu sme predikovali na základe charakteristík nabíjacích miest a okolia nabíjacích miest reprezentovaného GIS dátami. Na základe literatúry a v spolupráci s expertmi v oblasti elektromobility sme identifikovali tri hlavné aspekty pri nasadzovaní a plánovaní nabíjacej infraštruktúry, pomocou ktorých sme vybrali sedem indikátorov výkonnosti nabíjacej infraštruktúry. Medzi nimi mala popularita, kvantifikovaná ako počet unikátnych používateľov EV, čo sa nabíjali na nabíjacom mieste, podstatný význam a najlepšiu vysvetliteľnosť dátami. Popularitu sme pre jednoduchosť a aplikovateľnosť kódovali ako binárnu premennú a využili sme prahovú hodnotu θ pre klasifikáciu pravdepodobnostnej hodnoty, s využitím ako miera prístupu k riziku. Okrem metódy LR-l_1 sme využili aj stromové metódy RF a GBRT. Všetky použité modely dokázali popularitu predikovať lepšie ako náhodný model, pričom najvyššiu presnosť (0.829) mala metóda LR-l_1. Okrem predikovania sme analyzovali aj vplyvné prediktory, kde sme uprednostnili metódu LR-l_1 pred stromovými metódami, ktoré využívali veľký počet prediktorov. Na analýzu vplyvných prediktorov metódy LR-l_1 sme využili štatistickú inferenciu pomocou metódy bootstrap. Vplyvné prediktory sme rozdelili do troch skupín: funkcia geografického územia, charakteristiky populácie a charakteristiky nabíjacích miest, kde vplyvom dominovali hlavne charakteristiky nabíjacích miest.

V poslednej podkapitole, spadajúcej pod tretí cieľ, opisujeme analýzu závislosti medzi rozdelením spotrebovanej elektrickej energie medzi nabíjacie miesta a charakteristikami okolia nabíjacieho miesta. Najskôr sme analyzovali rozdelenie spotreby energie, ktorá sa dá modelovať po jednoduchej transformácií beta rozdelením. Z indikátorov, ktoré úzko súvisia so spotrebovanou energiou, spotrebu energie najviac ovplyvňuje počet nabíjacích transakcií. Pomocou metódy lasso sme identifikovali prediktory okolia nabíjacej infraštruktúry, potenciálne vplývajúce na spotrebu energie. Využitím prevzorkovacej metódy bootstrap sme indikovali štatistickú spoľahlivosť výsledkov. Prediktory, ktoré významne ovplyvňujú spotrebovanú energiu poukazujú najmä na ekonomickú prosperitu. Napríklad, obyvatelia a firmy s vysokým (nízkym) príjmom, situované v blízkosti nabíjacích miest majú pozitívny (negatívny) vplyv na spotrebu energie. Podobne ekonomická prosperita je vyjadrená drahším novopostaveným bývaním pozitívne prepojeným na spotrebovanú

energiu. Takisto významný vplyv má pracovný sektor obyvateľstva v samosprávach, ako aj počet finančných a realitných podnikov. Najväčší negatívny vplyv majú obyvatelia závislí na sociálnej pomoci. Stratifikácia nabíjacích miest, podľa stratégie použitej pri jej umiestňovaní, viedla k rozdeleniu regresných koeficientov na dve skupiny. Typy podnikov, pracovný sektor obyvateľstva a verejné miesta záujmu sú prepojené na vyššiu spotrebu energie na nabíjacích miestach umiestnených strategicky. Populačné charakteristiky, napr. počet pôrodov a úmrtí na 1000 obyvateľov, sú prepojené na spotrebu energie na nabíjacích miestach umiestnených na základe dopytu.

Primárne využitie výsledkov vidíme v podpore pri zostavovaní optimalizačných, simulačných ako aj učiacich sa modelov na základe nami vybratých dát, vhodne popisujúcich okolie nabíjacích staníc. Takto získané modely dokážu pomôcť efektívnejšie rozmiestniť nabíjaciu infraštruktúru a prispieť tak k energetickej efektívnosti a ekologickejšej doprave. Zostavená metodológia na stratifikáciu staníc, popísaná v podkapitole 3.5, môže byť použitá na vylepšenie umiestňovacích stratégií nabíjacej infraštruktúry, zameraných na určité skupiny nabíjacích staníc, napr. stanice určené na pracovné nabíjanie. Takisto zostavená metodológia v podkapitolách 3.2 a 3.5 môže byť použitá na plánovanie kapacity elektrickej siete a to zvážením energetického dopytu nabíjacej infraštruktúry.

Za hlavné teoretické ako aj praktické prínosy pokladáme výsledky dátových analýz, ktoré priniesli nové poznatky o správaní sa obslužného systému, skladajúceho sa s veľkého množstva nabíjacích staníc. Boli to napríklad pravdepodobnostné rozdelenie spotreby energie, charakteristiky časových radov, správanie popísané jednotlivými zhlukmi nabíjacích staníc ako aj samotné interpretácie koeficientov, ktoré odpovedajú prediktorom potenciálne vplývajúcim na popularitu a spotrebu energie. Takisto medzi dôležité prínosy práce patrí porovnanie metód, ktoré bolo možné použiť pri riešení jednotlivých problémov (napr. predpovedanie časových radov, určenie zhlukov staníc, predikcia popularity atď.). Určili sme, ktorá z uvažovaných metód alebo postupov je pre daný problém najvhodnejšia.

Na základe tohoto súhrnu považujeme ciele práce za splnené. Uvedené výsledky výskumu prinášajú nové možnosti aplikácie a nové poznatky pre vednú oblasť inteligentných informačných systémov a spolu s použitou metodológiou môžu nájsť uplatnenie v oblasti elektromobility.

Odporúčania pre ďalšiu prácu

V budúcnosti by sme sa chceli zamerať na tvorbu segmentov zameranú na časové využívanie nabíjacej infraštruktúry. V porovnaní s [119] by sme chceli dosiahnuť spresnenie predpovedania časov pripojenia k nabíjacím staniciam. Zaujímavé výsledky môže priniesť aj zopakovanie analýzy identifikátorov vplývajúcich na prevádzku nabíjacej infraštruktúry samostatne pre veľké mestské zóny ako Amsterdam alebo Rotterdam. Takisto ďalej spolupracujeme s autormi [107] na pravdepodobnostných predikciách spotreby energie s hodinovou frekvenciou.

Zoznam použitej literatúry

[1] Amini, M.H., Kargarian, A., Karabasoglu, O.: 'Arima-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation', *Electric Power Systems Research*, vol. 140, pp. 378–390, 2016

[2] Andrenacci, N., Ragona, R., Valenti, G.: 'A demand-side approach to the optimal deployment of electric vehicle charging stations in metropolitan areas', *Applied Energy*, vol. 182, pp. 39–46, 2016

[3] Arias, M.B., Bae, S.: 'Electric vehicle charging demand forecasting model based on big data technologies', *Applied energy*, vol. 183, pp. 327–339, 2016

[4] Belgrado, P.F., Buzna, L., et al.: 'Evaluating the predictability of future energy consumption-application of statistical classification models to data from ev charging points.', *VEHITS*, pp. 617–625, 2018

[5] Berk, R., Brown, L., et al.: 'Valid post-selection inference', *The Annals of Statistics*, vol. 41, no. 2, pp. 802–837, 2013

[6] Bikcora, C., Refa, N., et al.: 'Prediction of availability and charging rate at charging stations for electric vehicles', *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, p. 1–6, 2016, doi:10.1109/PMAPS.2016.7764216

[7] Blasius, E., Wang, Z.: 'Effects of charging battery electric vehicles on local grid regarding standardized load profile in administration sector', *Applied energy*, vol. 224, pp. 330–339, 2018

[8] Bondell, H.D., Reich, B.J.: 'Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar', *Biometrics*, vol. 64, no. 1, pp. 115–123, 2008

[9] Brady, J., O'Mahony, M.: 'Modelling charging profiles of electric vehicles based on real-world electric vehicle charging data', *Sustainable Cities and Society*, vol. 26, pp. 203–216, 2016

[10] Braun, M., Altan, H., Beck, S.: 'Using regression analysis to predict the future energy consumption of a supermarket in the uk', *Applied Energy*, vol. 130, pp. 305–313, 2014

[11] Bring, J.: 'How to standardize regression coefficients', *The American Statistician*, vol. 48, no. 3, pp. 209–213, 1994

[12] Brooker, R.P., Qin, N.: 'Identification of potential locations of electric vehicle supply equipment', *Journal of Power Sources*, vol. 299, pp. 76–84, 2015

[13] Buzna, L., De Falco, P., et al.: 'Electric vehicle load forecasting: A comparison between time series and machine learning approaches', *2019 1st International Conference on Energy Transition in the Mediterranean Area (SyNERGY MED)*, pp. 1–5, IEEE, 2019

[14] Cady, F.: *The data science handbook*, John Wiley & Sons, 2017

[15] Cazzola, P., Gorner, M., et al.: 'Global ev outlook 2016', Tech. rep., International Energy Agency, France, 2016

[16] Cazzola, P., Gorner, M., et al.: 'Global ev outlook 2017: Two million and counting', Tech. rep., International Energy Agency,, 2017

[17] Cazzola, P., Gorner, M., et al.: 'Global ev outlook 2019: Scaling-up the transition to electric mobility', Tech. rep., International Energy Agency,, 2019

[18] Chang, X., Shen, J., et al.: 'Statistical patterns of human mobility in emerging bicycle sharing systems', *PloS one*, vol. 13, no. 3, 2018

[19] Chaouch, M.: 'Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves.', *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 411–419, 2014

[20] Chatterjee, S., Hadi, A.S.: *Regression analysis by example*, John Wiley & Sons, 2015

[21] Chen, L., Ma, X., et al.: 'Understanding bike trip patterns leveraging bike sharing system open data', *Frontiers of computer science*, vol. 11, no. 1, p. 38–48, 2017

[22] Chen, T.D., Kockelman, K.M., Khan, M.: 'Locating electric vehicle charging stations: Parking-based assignment method for seattle, Washington', *Transportation Research Record*, vol. 2385, no. 1, p. 28–36, 2013

[23] Chen, Z., Zhang, Z., et al.: 'An analysis of the charging characteristics of electric vehicles based on measured data and its application', *IEEE Access*, vol. 6, pp. 24475–24487, 2018

[24] Cheon, S., Kang, S.J.: 'An electric power consumption analysis system for the installation of electric vehicle charging stations', *Energies*, vol. 10, no. 10, p. 1534, 2017

[25] Coffman, M., Bernstein, P., Wee, S.: 'Electric vehicles revisited: a review of factors that affect adoption', *Transport Reviews*, vol. 37, no. 1, p. 79–93, 2017

[26] Csiszár, C., Csonka, B., et al.: 'Urban public charging station locating method for electric vehicles based on land use approach', *Journal of Transport Geography*, vol. 74, pp. 173–180, 2019

[27] Cuesta, H., Kumar, S.: *Practical Data Analysis*, Packt Publishing Ltd, 2016

[28] Cynthia Brewer, Mark Harrower and The Pennsylvania State University: 'Colorbrewer', `http://colorbrewer2.org/`, 2019, accessed: 2019-09-12

[29] Dagnely, P., Ruette, T., et al.: 'Predicting hourly energy consumption. can you beat an autoregressive model', *Proceeding of the 24th Annual Machine Learning Conference of Belgium and the Netherlands, Benelearn, Delft, The Netherlands*, vol. 19, 2015

[30] Daina, N., Polak, J.W., Sivakumar, A.: 'Patent and latent predictors of electric vehicle charging behavior', *Transportation Research Record*, vol. 2502, no. 1, pp. 116–123, 2015

[31] Dannecker, L.: *Energy time series forecasting: efficient and accurate forecasting of evolving time series from the energy domain*, Springer, 2015

[32] De Cauwer, C., Verbeke, W., et al.: 'A data-driven method for energy consumption prediction and energy-efficient routing of electric vehicles in real-world conditions', *Energies*, vol. 10, no. 5, p. 608, 2017

[33] Deb, S., Tammi, K., et al.: 'Charging station placement for electric vehicles: A case study of guwahati city, india', *IEEE Access*, vol. 7, pp. 100270–100282, 2019

[34] Divshali, P.H., Evens, C.: 'Behaviour analysis of electrical vehicle flexibility based on large-scale charging data', *2019 IEEE Milan PowerTech*, pp. 1–6, IEEE, 2019

[35] Dong, J., Liu, C., Lin, Z.: 'Charging infrastructure planning for promoting battery electric vehicles: An activity-based approach using multiday travel data', *Transportation Research Part C: Emerging Technologies*, vol. 38, pp. 44–55, 2014

[36] Dormann, C.F., Elith, J., et al.: 'Collinearity: a review of methods to deal with it and a simulation study evaluating their performance', *Ecography*, vol. 36, no. 1, pp. 27–46, 2013

[37] Economist, T.: 'The world's most valuable resource is no longer oil, but data', *The Economist: New York, NY, USA*, 2017

[38] ElaadNL: 'Elaadnl', `https://www.elaad.nl/`, 2018, accessed: 2019-06-21

[39] EnergieAtlas, N.: 'Energy atlas', `https://www.pdok.nl/downloads?articleid=1951681`, 2015, accessed: 2018-10-16

[40] Fetene, G.M., Kaplan, S., et al.: 'Harnessing big data for estimating the energy consumption and driving range of electric vehicles', *Transportation Research Part D: Transport and Environment*, vol. 54, pp. 1–11, 2017

[41] Fischer, D., Harbrecht, A., et al.: 'Electric vehicles' impacts on residential electric local profiles–a stochastic modelling approach considering socio-economic, behavioural and spatial factors', *Applied energy*, vol. 233, pp. 644–658, 2019

[42] Flammini, M.G., Prettico, G., et al.: 'Statistical characterisation of the real transaction data gathered from electric vehicle charging stations', *Electric Power Systems Research*, vol. 166, p. 136–150, 2019, ISSN 0378-7796, doi:10.1016/j.epsr.2018.09.022

[43] Frade, I., Ribeiro, A., et al.: 'Optimal location of charging stations for electric vehicles in a neighborhood in lisbon, portugal', *Transportation Research Record*, vol. 2252, no. 1, pp. 91–98, 2011

[44] Friedman, J., Hastie, T., Tibshirani, R.: 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software*, vol. 33, no. 1, pp. 1–22, 2010

[45] Giménez-Gaydou, D.A., Ribeiro, A.S., et al.: 'Optimal location of battery electric vehicle charging stations in urban areas: A new approach', *International Journal of Sustainable Transportation*, vol. 10, no. 5, pp. 393–405, 2016

[46] Google: 'Google maps', Accessed: 2018-11-02

[47] Gopalakrishnan, R., Biswas, A., et al.: 'Demand prediction and placement optimization for electric vehicle charging stations', *arXiv preprint arXiv:1604.05472*, 2016

[48] GreenWay: 'Greenway', `https://greenway.sk`, 2018, accessed: 2019-06-24

[49] Han, J., Pei, J., Kamber, M.: *Data mining: concepts and techniques*, Elsevier, 2011

[50] Hardman, S., Jenn, A., et al.: 'A review of consumer preferences of and interactions with electric vehicle charging infrastructure', *Transportation Research Part D: Transport and Environment*, vol. 62, pp. 508–523, 2018

[51] Hastie, T., Tibshirani, R., Friedman, J.: *The elements of statistical learning: data mining, inference and prediction*, Springer, 2nd ed., 2009

[52] Helmus, J., Spoelstra, J., et al.: 'Assessment of public charging infrastructure push and pull rollout strategies: The case of the netherlands', *Energy policy*, vol. 121, p. 35–47, 2018

[53] Hsu, D.: 'Identifying key variables and interactions in statistical models of building energy consumption using regularization', *Energy*, vol. 83, pp. 144–155, 2015

[54] Hyndman, R.J., Athanasopoulos, G.: *Forecasting: principles and practice*, OTexts, 2018

[55] James, G., Witten, D., et al.: *An introduction to statistical learning*, vol. 112, Springer, 2013

[56] Kaltenbrunner, A., Meza, R., et al.: 'Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system', *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455–466, 2010

[57] Kara, E.C., Macdonald, J.S., et al.: 'Estimating the benefits of electric vehicle smart charging at non-residential locations: A data-driven approach', *Applied Energy*, vol. 155, pp. 515–525, 2015

[58] Koťka, V., Koreň, M., Ižvoltová, J.: *Geoinformatika pre inžinierov*, Edis, Žilina, 215

[59] Kuhn, M., Johnson, K.: *Applied predictive modeling*, vol. 26, Springer, 2013

[60] Leou, R.C., Teng, J.H., Su, C.L.: 'Modelling and verifying the load behaviour of electric vehicle charging stations based on field measurements', *IET Generation, Transmission & Distribution*, vol. 9, no. 11, pp. 1112–1119, 2015

[61] Li, M., Lenzen, M., et al.: 'Gis-based probabilistic modeling of bev charging load for australia', *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3525–3534, 2018

[62] Li, X., Zhang, Q., et al.: 'A data-driven two-level clustering model for driving pattern analysis of electric vehicles and a case study', *Journal of cleaner production*, vol. 206, pp. 827–837, 2019

[63] Liu, S.V., Chen, F.L., Xue, J.: 'Evaluation of traffic density parameters as an indicator of vehicle emission-related near-road air pollution: A case study with nexus measurement data on black carbon', *Int J Environ Res Public Health*, vol. 14, no. 12, p. 1581, 2017

[64] Lloyd, C.D.: *Local models for spatial analysis*, CRC press, 2010

[65] López, K.L., Gagné, C., Gardner, M.A.: 'Demand-side management using deep learning for smart charging of electric vehicles', *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2683–2691, 2018

[66] Louie, H.M.: 'Time-series modeling of aggregated electric vehicle charging station load', *Electric Power Components and Systems*, vol. 45, no. 14, pp. 1498–1511, 2017, doi:10.1080/15325008.2017.1336583

[67] Lucas, A., Barranco, R., Refa, N.: 'Ev idle time estimation on charging infrastructure, comparing supervised machine learning regressions', *Energies*, vol. 12, no. 2, p. 269, 2019

[68] Lucas, A., Prettico, G., et al.: 'Indicator-based methodology for assessing ev charging infrastructure using exploratory data analysis', *Energies*, vol. 11, no. 7, 2018, ISSN 1996-1073, doi:10.3390/en11071869

[69] Ludlow, J., Enders, W.: 'Estimating non-linear arma models using fourier coefficients', *International Journal of Forecasting*, vol. 16, no. 3, pp. 333–347, 2000

[70] Ma, J., Cheng, J.C.: 'Estimation of the building energy use intensity in the urban scale by integrating gis and big data technology', *Applied Energy*, vol. 183, pp. 182–192, 2016

[71] Majidpour, M., Qiu, C., et al.: 'Forecasting the ev charging load based on customer profile or station measurement?', *Applied energy*, vol. 163, pp. 134–141, 2016

[72] Marček, D., Marček, M.: *Analỳza, modelovanie a prognózovanie časovỳch radov s aplikáciami v ekonomike*, Žilinská univerzita, 2001

[73] Matthews, B.W.: 'Comparison of the predicted and observed secondary structure of t4 phage lysozyme', *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, p. 442–451, 1975

[74] Ministry of the Interior and Kingdom Relations: 'Liveability meter', `https://data.overheid.nl/data/dataset/leefbaarometer-2-0---meting-2016`, 2015, accessed: 2018-10-15

[75] Mitchell, T.M., et al.: 'Machine learning. 1997', *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997

[76] Neaimeh, M., Wardle, R., et al.: 'A probabilistic approach to combining smart meter and electric vehicle charging data to investigate distribution network impacts', *Applied Energy*, vol. 157, pp. 688–698, 2015

[77] Netherlands Enterprise Agency: 'Electric vehicle charging - definitions and explanation', `https://bit.ly/2LehwPk`, 2019, accessed: 2019-06-19

[78] NRSR: 'Zákon o energetike a o zmene a doplnení niektorých zákonov', https://www.zakonypreludi.sk/zz/2012-251, 2019, accessed: 2019-12-15

[79] OpenChargeMap: https://openchargemap.org, 2015, accessed: 2019-01-10

[80] OpenStreetMap: Accessed: 2018-12-02

[81] OpenStreetMap: https://www.openstreetmap.org, 2015, accessed: 2019-02-13

[82] Oplaadpalen: https://www.oplaadpalen.nl/, 2015, accessed: 2019-02-20

[83] Pagany, R., Marquardt, A., Zink, R.: 'Electric charging demand location model—a user-and destination-based locating approach for electric vehicle charging stations', *Sustainability*, vol. 11, no. 8, p. 2301, 2019

[84] Pagany, R., Ramirez Camargo, L., Dorner, W.: 'A review of spatial localization methodologies for the electric vehicle charging infrastructure', *International Journal of Sustainable Transportation*, vol. 13, no. 6, pp. 433–449, 2019

[85] Pearson, R.K., Neuvo, Y., et al.: 'Generalized hampel filters', *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–18, 2016

[86] Pevec, D., Babic, J., Podobnik, V.: 'Electric vehicles: A data science perspective review', *Electronics*, vol. 8, no. 10, p. 1190, 2019, ISSN 2079-9292, doi:10.3390/electronics8101190

[87] Pevec, D., Babic, J., et al.: 'A data-driven statistical approach for extending electric vehicle charging infrastructure', *International Journal of Energy Research*, vol. 42, no. 9, pp. 3102–3120, 2018, doi:10.1002/er.3978

[88] Plotly: 'Plotly', https://www.plot.ly/, 2019, accessed: 2019-10-11

[89] Press, G.: 'Data scientists spend most of their time cleaning data, 2016', https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data, 2016, accessed: 2019-05-03

[90] Sadeghianpourhamami, N., Refa, N., et al.: 'Quantitive analysis of electric vehicle flexibility: A data-driven approach', *International Journal of Electrical Power & Energy Systems*, vol. 95, p. 451–462, 2018, ISSN 0142-0615, doi:10.1016/j.ijepes.2017.09.007

[91] Sandén, B.A., Wallgren, P.: *Systems Perspectives on Electromobility 2017*, Chalmers University of Technology, 2017

[92] Sasaki, Y., et al.: 'The truth of the f-measure', 2007

[93] Sharma, D.B., Bondell, H.D., Zhang, H.H.: 'Consistent group identification and variable selection in regression with correlated predictors', *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 319–340, 2013

[94] Shepero, M., Munkhammar, J.: 'Data from electric vehicle charging stations: Analysis and model development', 2018

[95] Shepero, M., Munkhammar, J.: 'Spatial markov chain model for electric vehicle charging in cities using geographical information system (gis) data', *Applied Energy*, vol. 231, pp. 1089–1099, 2018

[96] Siegel, A.: *Practical business statistics*, Academic Press, 2016

[97] Singhvi, D., Singhvi, S., et al.: 'Predicting bike usage for new york city's bike sharing system', *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015

[98] Speidel, S., Bräunl, T.: 'Driving and charging patterns of electric vehicles for energy usage', *Renewable and Sustainable Energy Reviews*, vol. 40, pp. 97–110, 2014

[99] Springboard: *Plug in electric vehicles in smart grids*, 2018, accessed: 2018-6-11

[100] Statistics Netherlands: 'Cbs land cover', `https://www.pdok.nl/downloads?articleid=1951731`, 2015, accessed: 2016-09-14

[101] Statistics Netherlands: 'Neighbourhoods dataset 2015', `https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische%20data/wijk-en-buurtkaart-2015`, 2015, accessed: 2018-08-20

[102] Statistics Netherlands: 'Population cores in the netherlands', `https://www.cbs.nl/nl-nl/achtergrond/2014/13/bevolkingskernen-in-nederland-2011`, 2015, accessed: 2017-09-14

[103] Straka, M.: 'Robustness of variable selection regression methods', *Proceedings of the MIST conference 2018*, pp. 1–6, CreateSpace Independent Publishing Platform, 2018

[104] Straka, M., Buzna, L.: 'Use cases and introductory analysis of the dataset collected within the large network of public charging stations', *International Conference on Reliability and Statistics in Transportation and Communication*, pp. 203–213, Springer, 2018

[105] Straka, M., Buzna, L.: 'Clustering algorithms applied to usage related segments of electric vehicle charging stations', *Transportation Research Procedia*, vol. 40, pp. 1576–1582, 2019

[106] Straka, M., Buzna, L.: 'Preprocessing of gis data for electric vehicle charging stations analysis and evaluation of the predictors significance', *Transportation Research Procedia*, vol. 40, pp. 1583–1590, 2019

[107] Straka, M., De Falco, P., et al.: 'Predicting popularity of electric vehicle charging infrastructure in urban context', *IEEE Access*, vol. 8, pp. 11315–11327, 2020

[108] Suganthi, L., Samuel, A.A.: 'Energy models for demand forecasting—a review', *Renewable and sustainable energy reviews*, vol. 16, no. 2, pp. 1223–1240, 2012

[109] Tableau: 'Tableau software', `https://www.tableau.com/`, 2018, accessed: 2018-8-25

[110] Tao, Y., Huang, M., Yang, L.: 'Data-driven optimized layout of battery electric vehicle charging infrastructure', *Energy*, vol. 150, pp. 735–744, 2018

[111] Thornton, L.E., Pearce, J.R., Kavanagh, A.M.: 'Using geographic information systems (gis) to assess the role of the built environment in influencing obesity: a glossary', *International Journal of Behavioral Nutrition and Physical Activity*, vol. 8, no. 1, p. 71, 2011

[112] Tibshirani, R., Wainwright, M., Hastie, T.: *Statistical learning with sparsity: the lasso and generalizations*, Chapman and Hall/CRC, 2015

[113] Tietge, U., Mock, P., et al.: 'Comparison of leading electric vehicle policy and deployment in europe', Tech. rep., The International Council on Clean Transportation, 2016

[114] Traffic flows: 'The database was provided for reserach purposes by the national institute for public healt and environment.', `http://www.rivm.nl/`, 2015, accessed: 2019-01-07

[115] Tso, G.K., Yau, K.K.: 'Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks', *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007

[116] U.S. Department of Eenergy: 'All-electric vehicles', `https://www.fueleconomy.gov/feg/evtech.shtml`, 2020, accessed: 2020-05-10

[117] Van Buuren, S.: *Flexible imputation of missing data*, Chapman and Hall/CRC, 2018

[118] Van den Hoed, R., Helmus, J., et al.: 'Data analysis on the public charge infrastructure in the city of amsterdam', *World Electric Vehicle Journal*, vol. 6, no. 4, pp. 829–838, 2013, doi:10.3390/wevj6040829

[119] van den Hoed, R., Maase, S., et al.: *E-mobility getting smart with data*, Amsterdam University of Applied Sciences, 2019

[120] van der Waerden, P., Timmermans, H., de Bruin-Verhoeven, M.: 'Car drivers' characteristics and the maximum walking distance between parking facility and final destination', *Journal of Transport and Land Use*, vol. 10, no. 1, p. 1–11, 2015, ISSN 1938-7849, doi:10.5198/jtlu.2015.568

[121] Vazifeh, M.M., Zhang, H., et al.: 'Optimizing the deployment of electric vehicle charging stations using pervasive mobility data', *Transportation Research Part A: Policy and Practice*, vol. 121, pp. 75–91, 2019

[122] Verma, A., Asadi, A., et al.: 'A data-driven approach to identify households with plug-in electrical vehicles (pevs)', *Applied energy*, vol. 160, p. 71–79, 2015

[123] von Rueden, J., Kahlen, M., Belo, R.: 'Charging ahead-predicting optimal charging station locations across multiple cities', 2017

[124] Voulis, N., Warnier, M., Brazier, F.M.: 'Understanding spatio-temporal electricity demand at different urban scales: A data-driven approach', *Applied Energy*, vol. 230, pp. 1157–1171, 2018

[125] Wagner, S., Götzinger, M., Neumann, D.: 'Optimal location of charging stations in smart cities: A points of interest based approach', pp. 1–18, 2013

[126] Warren-Hicks, W.J., Hart, A.: *Application of uncertainty analysis to ecological risks of pesticides*, CRC Press, 2010

[127] Welch, B.L.: 'The generalization ofstudent's' problem when several different population variances are involved', *Biometrika*, vol. 34, no. 1/2, p. 28–35, 1947, doi: 10.2307/2332510

[128] Weldon, P., Morrissey, P., et al.: 'An investigation into usage patterns of electric vehicles in ireland', *Transportation Research Part D: Transport and Environment*, vol. 43, pp. 207–225, 2016

[129] Wickham, H.: *ggplot2: elegant graphics for data analysis*, Springer, 2016

[130] Wikipedia: 'Data set, 2016', `https://en.wikipedia.org/wiki/Data_set`, 2019, accessed: 2019-11-04

[131] Wolbertus, R., Kroesen, M., et al.: 'Fully charged: An empirical study into the factors that influence connection times at ev-charging stations', *Energy Policy*, vol. 123, p. 1–7, 2018, ISSN 0301-4215, doi:10.1016/j.enpol.2018.08.030

[132] Xiang, Y., Liu, J., et al.: 'Economic planning of electric vehicle charging stations considering traffic constraints and load profile templates', *Applied Energy*, vol. 178, pp. 647–659, 2016

[133] Xiong, Y., Wang, B., et al.: 'Electric vehicle driver clustering using statistical model and machine learning', *arXiv preprint arXiv:1802.04193*, 2018

[134] Xu, Y., Çolak, S., et al.: 'Planning for electric vehicle needs by coupling charging profiles with urban mobility', *Nature Energy*, vol. 3, no. 6, pp. 484–493, 2018

[135] Xydas, E., Marmaras, C., et al.: 'Forecasting electric vehicle charging demand using support vector machines', *2013 48th International Universities' Power Engineering Conference (UPEC)*, pp. 1–6, IEEE, 2013

[136] Xydas, E., Marmaras, C., et al.: 'A data-driven approach for characterising the charging demand of electric vehicles: A uk case study', *Applied energy*, vol. 162, pp. 763–771, 2016

[137] Yun, B., Sun, D.J., et al.: 'A charging location choice model for plug-in hybrid electric vehicle users', *Sustainability*, vol. 11, no. 20, p. 5761, 2019

[138] Zhang, H., Song, X., et al.: 'Battery electric vehicles in japan: Human mobile behavior based adoption potential analysis and policy target response', *Applied Energy*, vol. 220, pp. 527–535, 2018

[139] Zhang, J., Pan, X., et al.: 'Bicycle-sharing system analysis and trip prediction', *2016 17th IEEE international conference on mobile data management (MDM)*, vol. 1, p. 174–179, IEEE, 2016

[140] Zhang, Y., Zhang, Q., et al.: 'Gis-based multi-objective particle swarm optimization of charging stations for electric vehicles', *Energy*, vol. 169, pp. 844–853, 2019

[141] Zhao, H., Yan, X., Ren, H.: 'Quantifying flexibility of residential electric vehicle charging loads using non-intrusive load extracting algorithm in demand response', *Sustainable Cities and Society*, vol. 50, p. 101664, 2019

[142] Zhao, H.x., Magoulès, F.: 'A review on the prediction of building energy consumption', *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012

[143] Zhou, X.: 'Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in chicago', *PloS one*, vol. 10, no. 10, 2015

PRÍLOHY

Príloha A

Táto príloha uvádza v prvej časti zoznam a stručný popis použitých atribútov GIS dát, ako aj tabuľky súvisiace s ich spracovaním v druhej a tretej časti.

A-1 Tabuľky s vysvetlivkami a skratkami použitých atribútov

Uvažujeme niekoľko typov dát ako v kapitole 3.3.2: priemerné, počtové, pomerové, kategorické a percentuálne.

Populačné jadrá

Identifikátor	Popis	Typ atribútu
PC_1	Priemerný vek populácie na území populačného jadra 1. januára 2011	priemerný
PC_2	Počet osôb vo veku 15 - 24 rokov v jednočlennej domácnosti.	počtový
PC_3	Počet osôb vo veku 25 - 44 rokov v jednočlennej domácnosti.	počtový
PC_4	Počet osôb vo veku 45 - 64 rokov v jednočlennej domácnosti.	počtový
PC_5	Počet osôb starších ako 65 rokov v jednočlennej domácnosti	počtový
PC_6	Počet osôb vo veku 0 - 14 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí.	počtový
PC_7	Počet osôb vo veku 15 - 24 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí.	počtový
PC_8	Počet osôb vo veku 25 - 44 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí.	počtový
PC_9	Počet osôb vo veku 45 - 64 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí.	počtový
PC_{10}	Počet osôb starších ako 65 rokov vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí.	počtový
PC_{11}	Počet osôb vo veku 0 - 14 vo viacčlennej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi.	počtový
PC_{12}	Počet osôb vo veku 15 - 24 vo viacčlennej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi.	počtový
PC_{13}	Počet osôb vo veku 25 - 44 vo viacčlennej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi.	počtový
PC_{14}	Počet osôb vo veku 45 - 64 vo viacčlennej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi.	počtový
PC_{15}	Počet osôb vo veku 45 - 64 vo viacčlennej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi.	počtový
PC_{16}	Počet jednotlivcov žijúcich ako slobodný pár s biologickými, nevlastnými alebo adoptovanými deťmi v spoločnej domácnosti.	počtový
PC_{17}	Počet jednotlivcov žijúcich ako slobodný pár bez biologických, nevlastných alebo adoptovaných detí v spoločnej domácnosti.	počtový

PC_{18}	Počet jednotlivcov v uzavretom manželstve alebo registrovanom partnerstve žijúcich v spoločnej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi.	počtový
PC_{19}	Počet jednotlivcov v uzavretom manželstve alebo registrovanom partnerstve žijúcich v spoločnej domácnosti bez biologických, nevlastných alebo adoptovaných detí.	počtový
PC_{20}	Počet jednotlivcov v spoločnej domácnosti jedného rodiča žijúcich s najmenej jedným dieťaťom.	počtový
PC_{21}	Počet jednotlivcov žijúcich v spoločnej domácnosti, ktorí nie sú rodičmi, deťmi alebo v partnerskom vzťahu s ostatnými.	počtový
PC_{22}	Počet jednotlivcov obývajúcich inštitučné zariadenia ako sú domovy sociálnej starostlivosti, domovy dôchodcov, detské domovy, krízové centrá, rehabilitačné centrá a väzenia.	počtový
PC_{23}	Počet jednotlivcov, ktorých rodičia sa obaja narodili v Holandsku bez ohľadu na ich krajinu narodenia.	počtový
PC_{24}	Počet jednotlivcov s aspoň jedným rodičom narodeným v Európe (okrem Turecka), Severnej Amerike, Oceánii, Indonézii alebo Japonsku.	počtový
PC_{25}	Počet jednotlivcov s aspoň jedným rodičom narodeným v Afrike, Latinskej Amerike, Ázií (okrem Indonézie a Japonska) alebo Turecku.	počtový
PC_{26}	Percentuálny podiel pracujúcej populácie vo veku 25-44 rokov.	percentuálny
PC_{27}	Percentuálny podiel pracujúcej populácie vo veku 45-54 rokov.	percentuálny
PC_{28}	Percentuálny podiel pracujúcej populácie vo veku 55-64 rokov.	percentuálny
PC_{29}	Percentuálny podiel pracujúcej populácie vo veku 65-74 rokov.	percentuálny
PC_{30}	Percentuálny podiel pracujúcej populácie zamestnanej v poľnohospodárstve, lesníctve a rybolove.	percentuálny
PC_{31}	Percentuálny podiel pracujúcej populácie zamestnanej v baníctve, výrobnom priemysle a stavebníctve.	percentuálny
PC_{32}	Percentuálny podiel pracujúcej populácie zamestnanej v komerčných službách patriacich do kategórií: veľkoobchod a maloobchod, doprava a skladníctvo, informačno-komunikačné služby, prenájom hnuteľností, finančné služby a veterinárne služby.	percentuálny
PC_{33}	Percentuálny podiel pracujúcej populácie zamestnanej v nekomerčných službách patriacich do kategórií: verejná správa a verejné služby, vzdelávanie, zdravie a starostlivosť, kultúra, zábava a rekreácia, služby prislúchajúce potrebám domácnosti, extrateritoriálne organizácie a ostatné služby.	percentuálny
PC_{34}	Percentuálny podiel pracujúcej populácie zamestnanej v nasledovných kategóriách: energetika, vodné a odpadové hospodárstvo, hotely, reštaurácie a kaviarne a špecializované služby okrem veterinárstva (kategórie nezahrnuté v PC_{30}-PC_{33})	percentuálny

PC_{35}	Vnútroštátne migračné saldo vyjadrené ako počet jednotlivcov, ktorí sa prisťahovali sa do zemepisnej oblasti populačného jadra mínus počet jednotlivcov, ktorí sa presťahovali mimo populačného jadra v Holandsku v období 1. januára 2001 až 1. januára 2011.	počtový
PC_{36}	Migračné saldo vyjadrené ako počet jednotlivcov, ktorí imigrovali do populačného jadra mínus počet jednotlivcov, ktorí emigrantov z populačného jadra v období 1. januára 2001 až 1. januára 2011.	počtový
PC_{37}	Počet domácností s dvomi jednotlivcami.	počtový
PC_{38}	Počet domácností s tromi jednotlivcami.	počtový
PC_{39}	Počet domácností so štyrmi jednotlivcami.	počtový
PC_{40}	Počet domácností s piatimi jednotlivcami.	počtový
PC_{41}	Počet domácností so šiestimi a viacerými jednotlivcami.	počtový
PC_{42}	Priemerná obytnej nehnuteľnosti v eurách.	priemerný
PC_{43}	Priemerná cena obytnej nehnuteľnosti, obývanej majiteľom, v eurách.	priemerný
PC_{44}	Priemerná cena nájomnej obytnej nehnuteľnosti v eurách.	priemerný
PC_{45}	Počet rekreačných nehnuteľností v území populačného jadra.	počtový

Tabuľka 14: Zoznam atribútov datasetu Populačné jadrá.

Susedstvá

Identifikátor	Popis	Typ atribútu
N_1	Percento adries v susedstve s najčastejším PSČ. Najčastejšie PSČ je PSČ priradené najväčšiemu počtu adries v susedstve. Tento atribút je organizovaný v šiestich kategóriách podľa percenta adries zdieľajúcich rovnaké PSČ. Prvá kategória: viac ako 90 % adries, druhá kategória: 81-90 %, tretia kategória: 71-80 %, štvrtá kategória: 61-70 % , piata kategória: 51-60 %, šiesta kategória: menej ako 50 % adries zdieľa rovnaké PSČ.	kategorický
N_2	Priemerná hustota adries v susedstve. Pre každú adresu v susedstve, uvažujeme kruhové územie s polomerom 1 kilometer centrované na adresu. Ako prvú určujeme hustotu adries v každom kruhovom území. Ako druhý vypočítame priemer hustoty adries uvažujúc všetky kruhové územia. Tento atribút využívame na odhad stupňa urbanity územia.	priemerný
N_3	Urbánna trieda susedstva, založená na hustote nehnuteľností (päť tried)	kategorický
N_4	Celková populácia.	počtový
N_5	Percento populácie vo veku 15 - 24 rokov.	percentuálny
N_6	Percento populácie vo veku 25 - 44 rokov.	percentuálny
N_7	Percento populácie vo veku 45 - 54 rokov.	percentuálny
N_8	Percento populácie vo veku 65 a viac rokov.	percentuálny
N_9	Percento slobodných obyvateľov.	percentuálny

N_{10}	Percento obyvateľov v manželskom zväzku alebo registrovanom partnerstve.	percentuálny
N_{11}	Percento ovdovených obyvateľov.	percentuálny
N_{12}	Počet živonarodených v roku 2015.	počtový
N_{13}	Počet živonarodených v roku 2015 na tisíc obyvateľov.	počtový
N_{14}	Počet úmrtí v roku 2015.	počtový
N_{15}	Počet úmrtí v roku 2015 na tisíc obyvateľov.	počtový
N_{16}	Počet domácností.	počtový
N_{17}	Percento jednočlenných domácností.	percentuálny
N_{18}	Percento viacčlenných bezdetných domácností	percentuálny
N_{19}	Priemerný počet členov domácnosti.	priemerný
N_{20}	Percento imigrantov so západným pôvodom (Európa (okrem Turecka), Severná Amerika, Oceánia, Indonézia alebo Japonsko), relatívne k celkovej populácií.	percentuálny
N_{21}	Percento imigrantov s nezápadným pôvodom (Afrika, Latinská Amerika, Ázia (okrem Indonézie a Japonska) alebo Turecka), relatívne k celkovej populácií.	percentuálny
N_{22}	Percento imigrantov s marockým pôvodom, relatívne k celkovej populácií.	percentuálny
N_{23}	Percento imigrantov s marockým pôvodom v Holandských Antilách a Arube, relatívne k celkovej populácií.	percentuálny
N_{24}	Percento imigrantov s pôvodom zo Surinamu, relatívne k celkovej populácií.	percentuálny
N_{25}	Percento imigrantov z Turecka, relatívne k celkovej populácií.	percentuálny
N_{26}	Percento imigrantov s iným nezápadným pôvodom, relatívne k celkovej populácií, $N_{26} = N_{21} - (N_{22} + N_{23} + N_{24} + N_{25})$.	percentuálny
N_{27}	Počet hospodárskych, lesných a rybolovných podnikov.	počtový
N_{28}	Počet priemyselných a energetických podnikov.	počtový
N_{29}	Počet obchodných podnikov, hotelov, reštaurácií a kaviarní.	počtový
N_{30}	Počet dopravných, informačných a komunikačných podnikov.	počtový
N_{31}	Počet finančných služieb a podnikov s nehnuteľnosťami.	počtový
N_{32}	Počet podnikových služieb.	počtový
N_{33}	Počet kultúrnych, rekreačných a ostatných služieb nezahrnutých v N_{27}-N_{32}.	počtový
N_{34}	Celkový počet podnikov.	počtový
N_{35}	Počet obytných nehnuteľností s aspoň jednou obytnou funkciou a prípadne aj s inými funkciami.	počtový
N_{36}	Priemerná cena obytnej nehnuteľnosti v tisícoch eur.	priemerný
N_{37}	Počet viacgeneračných obydlí ako percento celkového trhu s rezidentnými nehnuteľnosťami.	percentuálny
N_{38}	Percento prázdnych obydlí.	percentuálny
N_{39}	Percento majiteľom obývaných obydlí.	percentuálny
N_{40}	Percento nájomných obydlí.	percentuálny
N_{41}	Percento nájomných obydlí vlastnených ubytovacou organizáciou (napr. bytovým družstvom).	percentuálny

N_{42}	Percento nájomných obydlí vlastnených inými vlastníkmi ako ubytovacie asociácie (e.g. obydlia prenajímané osobou, ktorá je majiteľom).	percentuálny
N_{43}	Percento obydlí postavených v roku 2000 alebo neskôr, vypočítané z celkového počtu obydlí.	percentuálny
N_{44}	Počet jednotlivcov poberajúcich príjem.	počtový
N_{45}	Priemerný príjem v tisícoch eur na jednotlivca poberajúceho príjem.	priemerný
N_{46}	Priemerný príjem v tisícoch eur na obyvateľa.	priemerný
N_{47}	Percento jednotlivcov žijúcich v domácnostiach patriacich medzi celonárodných 40 % osôb s najnižším osobným príjmom.	percentuálny
N_{48}	Percento jednotlivcov žijúcich v domácnostiach patriacich medzi celonárodných 20 % osôb s najvyšším osobným príjmom	percentuálny
N_{49}	Percento domácností patriacich medzi celonárodných 40 % domácností s najnižším osobným príjmom.	percentuálny
N_{50}	Percento domácností patriacich medzi celonárodných 20 % domácností s najnižším osobným príjmom.	percentuálny
N_{51}	Percento domácností s nízkou kúpnou silou.	percentuálny
N_{52}	Percento domácností pod alebo blízko sociálneho minima, okrem študentských domácností.	percentuálny
N_{53}	Počet jednotlivcov pobrajúcich dávky z dôvodu nezamestnanosti.	počtový
N_{54}	Počet jednotlivcov poberajúcich výhody zo zákonného poistenia v nezamestnanosti.	počtový
N_{55}	Počet jednotlivcov poberajúcich sociálnu pomoc od štátu.	počtový
N_{56}	Percento ekonomicky aktívnych jednotlivcov vo veku medzi 15 a 75 rokmi.	percentuálny
N_{57}	Počet jednotlivcov poberajúcich starobný dôchodok.	počtový
N_{58}	Počet motorových vozidiel na cestnú osobnú dopravu, okrem mopedov a motocyklov, do deväť miest na sedenie (vrátane vodiča).	počtový
N_{59}	Počet osobných áut na domácnosť.	počtový
N_{60}	Počet dodávok, nákladných áut, traktorov, vybraných špeciálnych vozidiel (napr. požiarne autá, čistiace autá, odťahové autá) a autobusov.	počtový
N_{61}	Počet motocyklov, skútrov a motorových invalidných vozidiel s motocyklovou registráciou.	počtový
N_{62}	Počet áut starých šesť rokov a viac.	počtový
N_{63}	Počet benzínových áut.	počtový

Tabuľka 15: Zoznam atribútov datasetu Susedstvá.

Využitie územia

Identifikátor	Popis	Typ atribútu
LC_1	Železnica.	nominálny
LC_2	Cesta.	nominálny

LC_3	Letisko.	nominálny
LC_4	Obytné územie.	nominálny
LC_5	Územie pre maloobchod, hotely, reštaurácie a kaviarne.	nominálny
LC_6	Územie pre verejné služby.	nominálny
LC_7	Územie pre sociálne a kultúrne služby.	nominálny
LC_8	Územie pre podniky a služby.	nominálny
LC_9	Smetisko.	nominálny
LC_{10}	Vrakoviská a šrotoviská.	nominálny
LC_{11}	Cintorín.	nominálny
LC_{12}	Bane, ropné a plynové polia.	nominálny
LC_{13}	Stavenisko.	nominálny
LC_{14}	Čiastočne spevnený terén (nezastavaný pustatina, nezarastené násypy a nepoužívané koľajnice).	nominálny
LC_{15}	Parky.	nominálny
LC_{16}	Športové plochy.	nominálny
LC_{17}	Územie pre nekomerčnú ornamentálnu a rastlinnú kultiváciu (parcely a školské záhrady).	nominálny
LC_{18}	Rekreačný terén používaný na jednodňovú rekreáciu (jednodňový kemping, zoo, safari parky, zábavné parky, prístavy a múzeá pod holým nebom).	nominálny
LC_{19}	Obytné územie využívané na viacdňovú rekreáciu (kemping, prázdninové domy a mládežnícke ubytovne).	nominálny
LC_{20}	Územie pre skleníkové záhradníctvo.	nominálny
LC_{21}	Ostatné poľnohospodárske územie.	nominálny
LC_{22}	Les.	nominálny
LC_{23}	Otvorený suchý prírodný terén.	nominálny
LC_{24}	Voda (Otvorený mokrý terén a voda).	nominálny
LC_{25}	Voda využívaná na rekreáciu (voda v parkoch a golfových ihriskách, veslovacie dráhy a rekreačné bazény).	nominálny

Tabuľka 16: Kategórie polygónov datasetu Využitie územia.

Spotreba energie

Identifikátor	Popis	Typ atribútu
EC_1	Priemerná spotreba plynu obytných nehnuteľností [m^3].	priemerný
EC_2	Ročná spotreba plynu obytných nehnuteľností [m^3].	počtový
EC_3	Počet obytných nehnuteľností kde bola meraná spotreba plynu.	počtový
EC_4	Priemerná spotreba elektrickej energie obytných nehnuteľností [kWh].	priemerný
EC_5	Ročná spotreba elektrickej energie obytných nehnuteľností [kWh].	počtový
EC_6	Počet obytných nehnuteľností kde bola meraná spotreba elektrickej energie.	počtový
EC_7	Priemerná spotreba plynu podnikmi [m^3].	priemerný
EC_8	Ročná spotreba plynu podnikmi [m^3].	počtový
EC_9	Počet podnikov kde bola meraná spotreba plynu.	počtový
EC_{10}	Priemerná spotreba elektrickej energie podnikmi [kWh].	priemerný

EC_{11}	Ročná spotreba elektrickej energie podnikmi [kWh].	počtový
EC_{12}	Počet podnikov kde bola meraná spotreba elektrickej energie.	počtový

Tabuľka 17: Atribúty datasetu Atlas energie a ich popis.

Životná úroveň

Identifikátor	Popis	Typ atribútu
L_1	Index životnej úrovne 2016 - podkategória obydlia, zahŕňajúca fakory ako typ, rok postavenia a vlastníctvo obydlí (odchýlka od národného priemeru).	numeric
L_2	Index životnej úrovne 2016 - podkategória socio-ekonomické pozadie obyvateľstva, zahŕňajúce typy rodín, migračné zázemie, mieru nezamestnanosti a podobné populačné charakteristiky (odchýlka od národného priemeru).	numeric
L_3	Index životnej úrovne 2016 - podkategória služby, zostavená hlavne z dostupnosti rôznych služieb v pešej vzdialenosti (odchýlka od národného priemeru)	numeric
L_4	Index životnej úrovne 2016 - podkategória bezpečnosť, vyhodnocujúca kriminalitu rôzneho druhu (odchýlka od národného priemeru).	numeric
L_5	Index životnej úrovne 2016 - podkategória fyzické prostredie, zahŕňajúca blízkosť rôznych objektov, napr. lesy, továrne, veterné turbíny, diaľnice, atď. (odchýlka od národného priemeru).	numeric
L_6	Liveability index 2016 - komplexný index zohľadňujúci kvalitu života.	kategorický poradový

Tabuľka 18: Zoznam atribútov z datasetu Životná úroveň

Dopravné toky

Identifikátor	Popis	Typ atribútu
TF_1	Počet áut za hodinu (7–19 h).	počtový
TF_2	Počet áut za hodinu (19–23 hr).	počtový
TF_3	Počet áut za hodinu (23–7 hr).	počtový
TF_4	Počet autobusov za hodinu (7–19 hr).	počtový
TF_5	Počet autobusov za hodinu (19–23 hr).	počtový
TF_6	Počet autobusov za hodinu (23–7 hr).	počtový
TF_7	Počet nákladných vozidiel za hodinu (7–19 hr).	počtový
TF_8	Počet nákladných vozidiel za hodinu (19–23 hr).	počtový
TF_9	Počet nákladných vozidiel za hodinu (23–7 hr).	počtový

Tabuľka 19: Zoznam atribútov asociovaný s datasetov Dopravné toky.

OSM dáta

Pätnásť kategórií OSM dát.

Identifikátor	Popis	Počet agregovaných OSM kategórií	Typ atribútu
OSM_1	Zdravie.	39	bodový
OSM_2	Zábava.	147	bodový
OSM_3	Kultúra a turizmus.	26	bodový
OSM_4	Financie.	10	bodový
OSM_5	Móda.	49	bodový
OSM_6	Jedlo.	42	bodový
OSM_7	Doprava.	36	bodový
OSM_8	Práca.	26	bodový
OSM_9	Domácnosť.	76	bodový
OSM_{10}	Vzdelanie.	17	bodový
OSM_{11}	Verejné objekty.	29	bodový
OSM_{12}	Hobby.	37	bodový
OSM_{13}	Šport.	39	bodový
OSM_{14}	Ubytovanie.	11	bodový
OSM_{15}	Rodina.	9	bodový

Tabuľka 20: Zoznam kategórií zostavený z OSM bodov záujmu.

A-2 Tabuľka zástupcov korelovaných atribútov

Tabuľka 21 zobrazuje zástupcov korelovaných atribútov.

Reprezentatívny prediktor	Vylúčený prediktor
Počet podnikov kde bola meraná spotreba plynu (EC_9)	Number of Počet podnikov kde bola meraná spotreba elektrickej energie (EC_{12})
Percento majiteľom obývaných obydlí (N_{39})	Percento nájomných obydlí (N_{40})
Percento domácností s nízkou kúpnou silou (N_{51})	Percento domácností pod alebo blízko sociálneho minima, okrem študentských domácností (N_{52})
Celková populácia (N_4)	Počet rezidentných nehnuteľností s aspoň jednou rezidentnou funkciou a prípadne aj s inými funkciami (N_{35}); Počet domácností (N_{16}); Počet jednotlivcov poberajúcich príjem (N_{44}); Ročná spotreba elektrickej energie obytných nehnuteľností (EC_5); Počet obytných nehnuteľností kde bola meraná spotreba plynu. (EC_3); Počet obytných nehnuteľností kde bola meraná spotreba elektrickej energie. (EC_6)
Priemerný počet členov domácnosti (N_{19})	Percento jednočlenných domácností (N_{17})
Priemerný príjem v tisícoch eur na obyvateľa (N_{46})	Priemerný príjem v tisícoch eur na jednotlivca poberajúceho príjem (N_{45})

Počet osôb vo veku 0 - 14 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí (PC_6)	Počet osôb vo veku 25 - 44 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí (PC_8)
Počet osôb vo veku 15 - 24 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí (PC_7)	Počet osôb vo veku 45 - 64 vo viacčlennej domácnosti bez biologických, nevlastných alebo adoptovaných detí (PC_9)
Počet jednotlivcov žijúcich ako slobodný pár bez biologických, nevlastných alebo adoptovaných detí v spoločnej domácnosti (PC_{17})	Počet osôb vo veku 25 - 44 vo viacčlennej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi (PC_{13})
Počet jednotlivcov v uzavretom manželstve alebo registrovanom partnerstve žijúcich v spoločnej domácnosti bez biologických, nevlastných alebo adoptovaných detí (PC_{19})	Počet osôb vo veku 45 - 64 vo viacčlennej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi (PC_{14})
Počet jednotlivcov žijúcich v spoločnej domácnosti, ktorí nie sú rodičmi, deťmi alebo v partnerskom vzťahu s ostatnými (PC_{21})	Počet osôb vo veku 25 - 44 rokov v jednočlennej domácnosti (PC_3)
Počet domácností s piatimi jednotlivcami (PC_{40})	Počet jednotlivcov v uzavretom manželstve alebo registrovanom partnerstve žijúcich v spoločnej domácnosti s biologickými, nevlastnými alebo adoptovanými deťmi (PC_{18})
riemerná obytnej nehnuteľnosti v eurách (PC_{42})	Priemerná cena obytnej nehnuteľnosti, obývanej majiteľom, v eurách (PC_{43}); Priemerná cena nájomnej obytnej nehnuteľnosti v eurách (PC_{44})

Tabuľka 21: Každý riadok prislúcha k skupine prediktorov so vzájomnou hodnotou Parsonovho korelačného koeficientu väčšieho ako 0.95. Prediktory v prvom stĺpci sú vybraté ako reprezentatívne pre skupinu prediktorov vypísaných v druhom stĺpci.

A-3 Dopĺňanie hodnôt chýbajúcich atribútov

Na základe analýzy dát sme vytvorili nasledovné pravidlá, podľa ktorých sme niektoré chýbajúce dáta doplnili:

a) Ak má územie reprezentované polygónom nula obyvateľov, chýbajúce hodnoty atribútov informujúce o populácií (napr. počet jednotlivcov poberajúcich sociálne dávky alebo iné využívajúcich výhody pre telesne postihnutých) boli nastavené na nula.

b) Ak je počet jednotlivcov poberajúcich príjem na území predstavovanom polygónom rovný nule, potom sa chýbajúca hodnota atribútov spojených s populačnými skupinami, ktoré majú istý druh alebo stupeň príjmu nastavila na nula.

c) Ak je počet obydlí na území predstavovanom polygónom nulový, tak sme nastavili na hodnotu nula atribúty informujúce o špecifickom type obydlí (napr. priemerná cena obývateľnej nehnuteľnosti).

d) Ak bol počet domácností na území predstavovanom polygónom nulový, potom boli chýbajúce dáta o domácnostiach nastavené na hodnotu nula.

e) Ak chýba trieda urbanizácie územia predstavovaného polygónonom polygóne, potom je nastavená hodnota tohto atribútu na najnižší stupeň urbanizácie - neurbánnu zónu.

f) Ak chýba percento adries na území s rovnakým poštovým smerovým číslom, potom je nastavená najnižšia hodnota percent ($< 50\%$ adries).

g) Ak je počet budov kde sa merala hodnota elektriny alebo plynu menší ako 5, chýbajúce hodnoty elektriny alebo plynu sú nastavené na hodnotu nula.

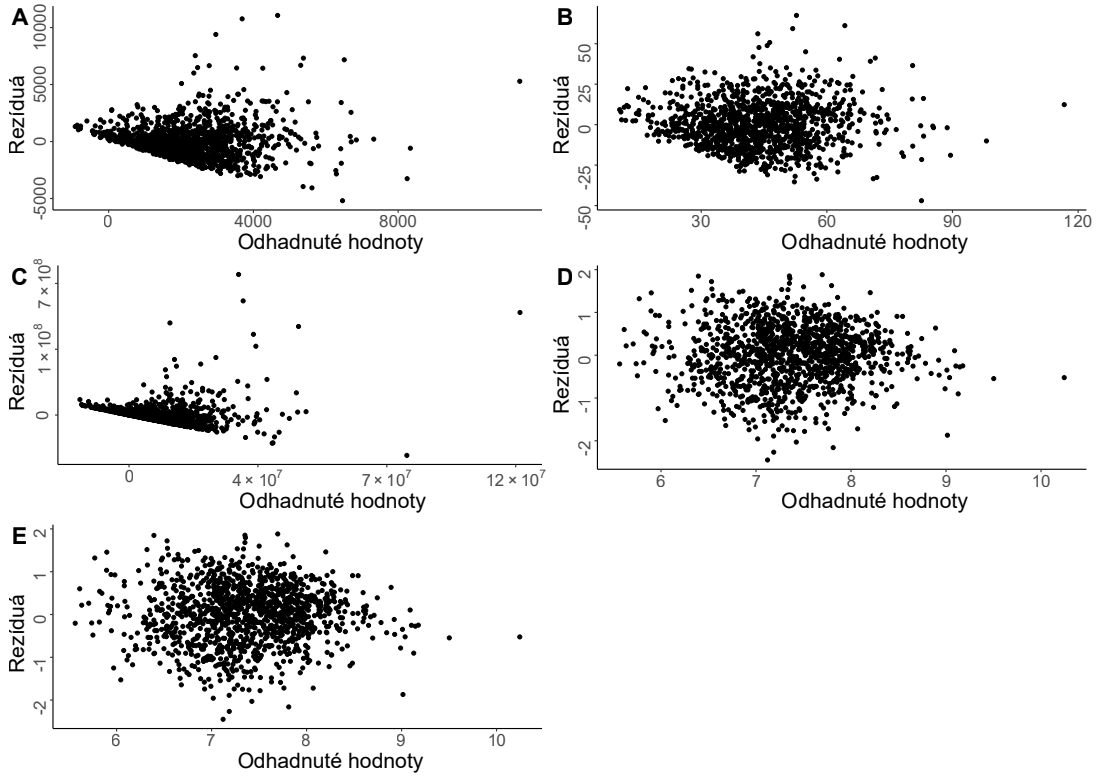
Ak to bolo možné, chýbajúce hodnoty atribútov sme odhadli pomocou známych hodnôt ostatných atribútov, napr. chýbajúce hodnoty priemernej veľkosti domácnosti sme nahradili počtom obyvateľov podeleným počtom domácností. Ak nebolo možné uplatniť žiadne z vyššie uvedených pravidiel, tak bola ponechaná chýbajúca hodnota atribútu.

A-4 Odstraňovanie prediktorov pomocou hodnoty VIF

Pre zmiernenie mutlikolinearity prítomnej v GIS dátach, ktorá bola problémom pre metódu lasso sme využili dvoj krokový algoritmus postupného odstraňovania prediktorov. V prvom kroku vypočítame hodnoty VIF koeficientov, v druhom kroku vylúčime prediktor s najvyššou VIF hodnotou, ak je väčšia alebo rovná ako 10, inak algoritmus zastavíme. Tento proces eliminoval nasledovných 47 prediktorov: PC_{37}, PC_{23}, N_{34}, PC_6, N_{21}, N_{63}, PC_{38}, N_9, N_{39}, PC_{20}, PC_{19}, N_4, PC_{17}, N_8, N_{62}, PC_{39}, PC_4, N_{48}, N_{49}, LC_4, PC_{21}, PC_{31}, PC_5, N_{10}, PC_{26}, N_{32}, N_{46}, N_6, N_2, N_{54}, PC_7, N_{47}, L_6, hustotu dopravy autobusov, PC_{25}, N_{29}, N_{33}, N_{19}, N_{58}, PC_{40}, PC_{24}, EC_2, PC_{12}, N_{53}, N_{50}, rezidenčné cestné segmenty, L_2. Opis týchto skratiek atribútov možno nájsť v prílohe A. Na redukciu multikolinearity na prípustnú mieru poukázali aj podľa hodnôt mier odvodených z vlastných hodnôt korelačnej matice [20, s. 252].

A-5 Nelineárne transformácie vektora výstupu

Testovali sme aj, či transformácie: $\sqrt{y}$, y^2, $\log(y)$ a Box-Cox transformácia [59, p. 32] vektora výstupu y môžu vylepšiť lineárny model, získaný pomocou OLS a matice prediktorov $\boldsymbol{X}$. Na obrázku 23 sú ukázané grafy rezíduí týchto modelov. Pre vhodný model by nemali grafy rezíduí javiť rozlíšiteľný vzor, indikujúc že model vhodne vysvetľuje vektor výstupu. Naopak prítomnosť vzoru v grafe rezíduí môže indikovať potrebu nelineárneho vzťahu pre vylepšenie modelu. Grafy rezíduí v paneloch (D) a (E) Obrázku 23 indikujú, že transformácia $\log(y)$ a Box-Cox transformácia s hodnotami exponentu $\lambda = 0.1$ vedú k najnáhodnejším vzorom rezíduí. Keďže Box-Cox funkcia sa limitne blíži k *log* funkcii,

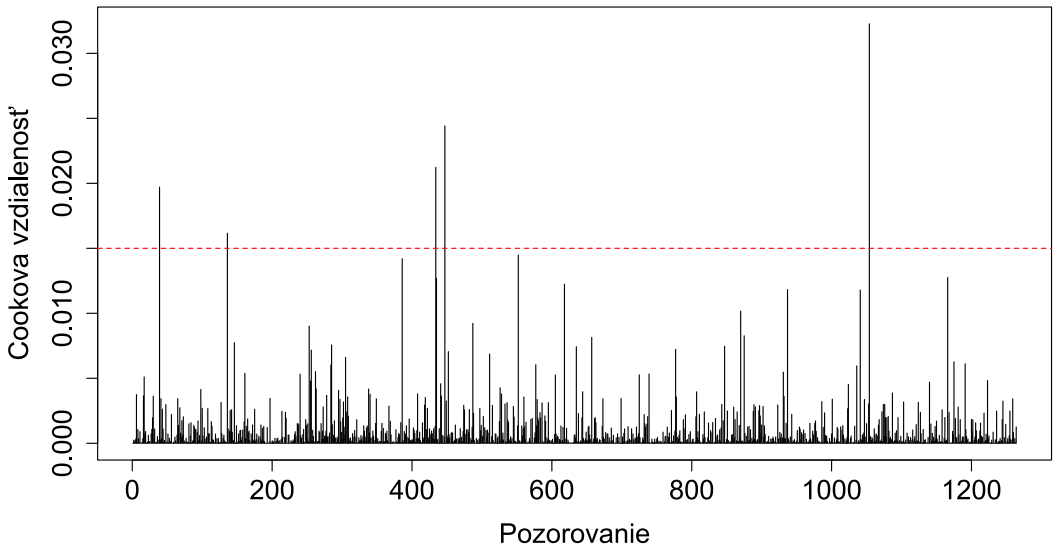


Obrázok 23: Grafy rezíduí. (A) Graf rezíduí získaný lineárnym modelom získaným aplikáciou OLS na $\boldsymbol{X}$ (po predspracovaní) a vektora výstupu y, (B) transformácia $\sqrt{y}$, (C) transformácia y^2, (D) transformácia $\log(y)$ a (E) Box-Cox transformácia (with $\lambda = 0.1$). Rezíduá vyzerajú náhodnejšie po aplikácií $\log(y)$ a Box-Cox transformácií.

keď sa λ blíži k nule, tak sú si tieto transformácie pri $\lambda = 0.1$ veľmi podobné. Preto aplikujeme ďalej transformáciu $\log(y)$.

A-6 Analýza vplyvných pozorovaní

Vplyvné pozorovania, stručne povedané nabíjacie miesta s nežiadúcou veľkosťou vplyvom na regresiu, detegujeme pomocou Cookovej vzdialenosti. V obrázku 24 sú zobrazené Cookove vzdialenosti všetkých pozorovaní, kde sa sústredíme na vzdialenosti vyčnievajúce spomedzi ostatných Podľa odporúčania [20] nastavujeme prahovú hodnotu na 0.015 a nachádzame 8 pozorovaní, kde Cookova vzdialenosť presahuje prahovú hodnotu. Tieto pozorovania sú považované za vplyvné pozorovania a tak sme ich vyradili z y aj $\boldsymbol{X}$. Vylepšila sa aj kvalita odhadu, napr. MSE OLS modelu kleslo z 0.523 na 0.512, t.j. o 2.4 % oproti situácií bez týchto pozorovaní.



Obrázok 24: Vizualizácia Cookových vzdialeností získaných pre jednotlivé pozorovania. Čiarkovaná čiara indikuje prahovú hodnotu aplikovanú na označenie vplyvných pozorovaní (t.j. tých čo prekročili prahovú hodnotu).

Príloha B - Spotreba elektrickej energie

B-1 Odvodenie funkcie hustoty pre spotrebu elektrickej energie s transformovanou funkciou hustoty beta rozdelenia

Zvažujúc Kolmogorov-Smirnov test dobrej zhody spolu s prihliadnutím na Q-Q a P-P grafy [126] sme uzavreli, že najuspokojivejší odhad rozdelenia spotrebovanej energie na nabíjacích miestach získame transformovaním dát s funkciou $g(y) = \sqrt[3]{y}$ (funkcia je aplikovaná na vektor po prvkoch) a využitím beta rozdelenia (viď podkapitolu 3.6.3). Beta rozdelenie je definované na intervale $\langle 0, 1\rangle$. Po preškálovaní je beta rozdelenie vhodné na reprezentáciu náhodnej premennej medzi minimálnou hodnotou y_{min} a maximálnou hodnotou y_{max}. Preto modelujeme beta rozdelením náhodnú premennú

$$Z = \frac{\sqrt[3]{Y} - y_{min}}{y_{max} - y_{min}}, \tag{42}$$

kde náhodná premenná Y modeluje spotrebu energie. Použitím rovnice (42) ustanovujeme vzťah medzi hustotou rozdelenia Y a Z ako

$$\begin{aligned} F_Y(y) = P(Y < y) = \\ P((Z(y_{max} - y_{min}) + y_{min})^3 < y) = \\ P\Big(Z < \frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big) = F_Z\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big). \end{aligned} \tag{43}$$

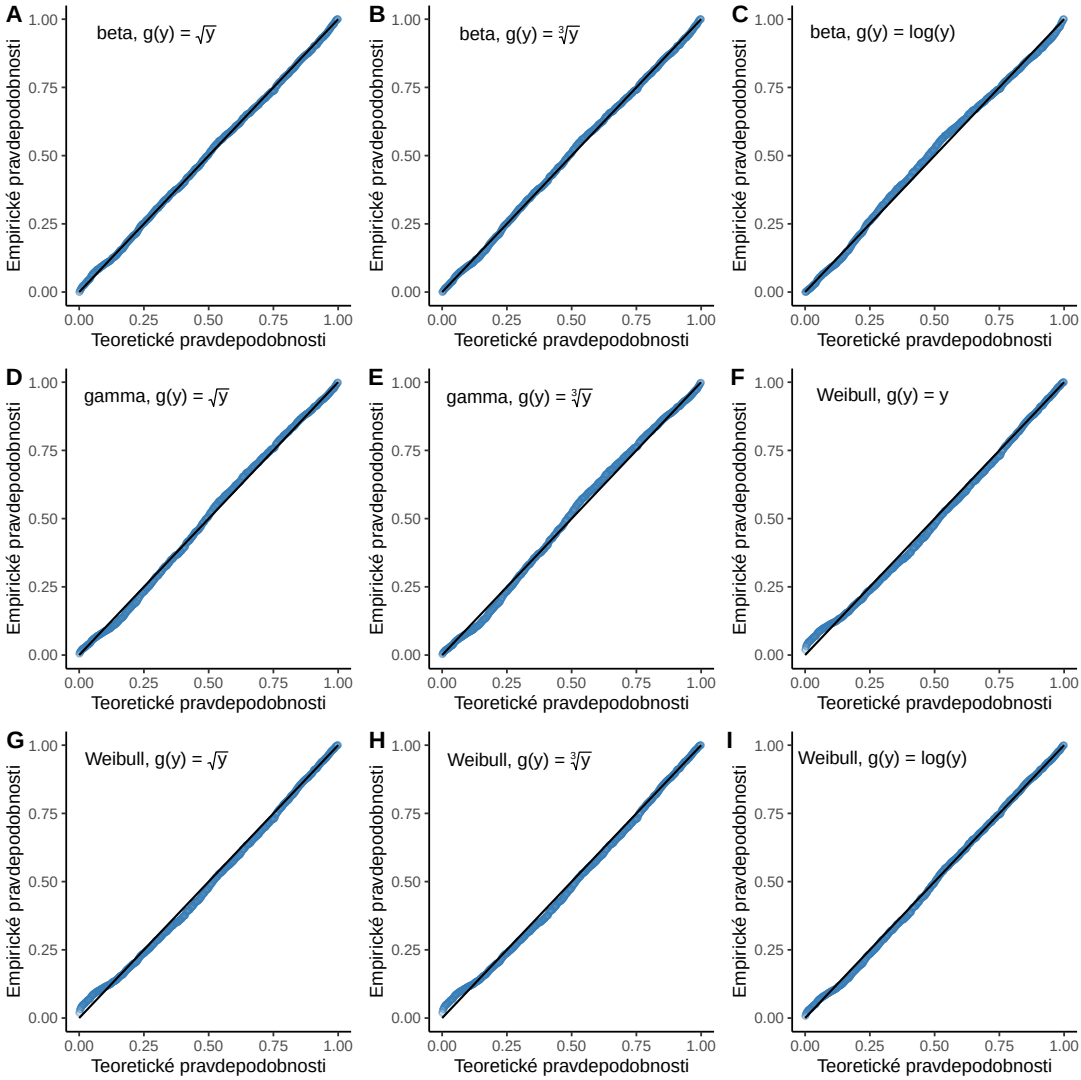
V dôsledku toho je funkcia hustoty charakterizujúca náhodnú premennú Y

$$\begin{aligned} f_Y(y) = \big[F_Y(y)\big]' = \left[F_Z\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)\right]' = \\ F_Z'\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)\left[\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\right]' = \\ f_Z\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)\frac{1}{3(y_{max} - y_{min})y^{\frac{2}{3}}}. \end{aligned} \tag{44}$$

Keďže $f_Z(\boldsymbol{z})$ je funkcia hustoty pravdepodobnosti beta rozdelenia, dostaneme nasledovnú funkciu hustoty pre spotrebu energie

$$f_Y(y, \alpha, \beta) = \frac{\Big(\frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)^{\alpha-1}\Big(1 - \frac{\sqrt[3]{y} - y_{min}}{y_{max} - y_{min}}\Big)^{\beta-1}}{B(\alpha, \beta)3(y_{max} - y_{min})y^{\frac{2}{3}}}, \tag{45}$$

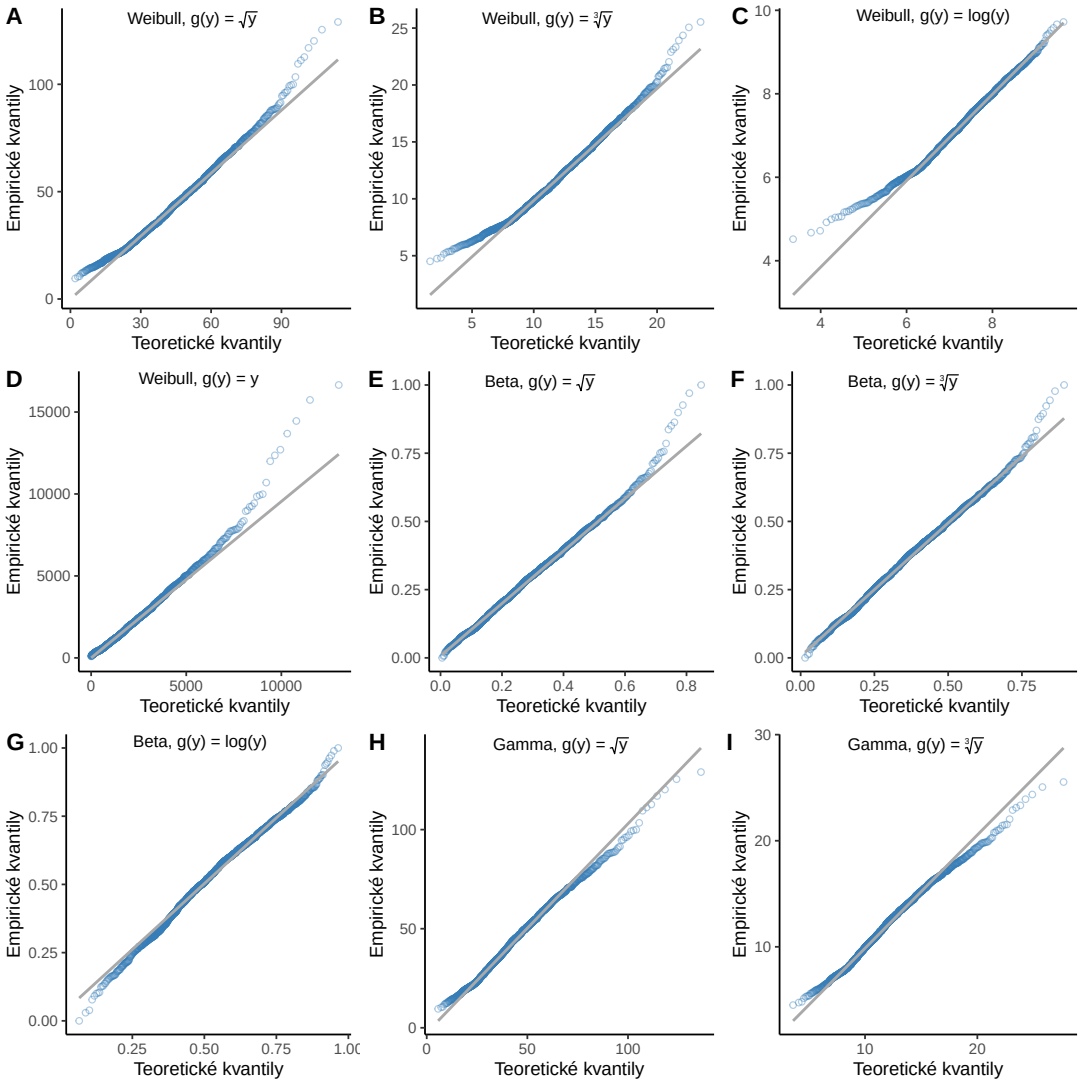
kde $B(\alpha, \beta)$ je Beta funkcia s parametrami α a β.



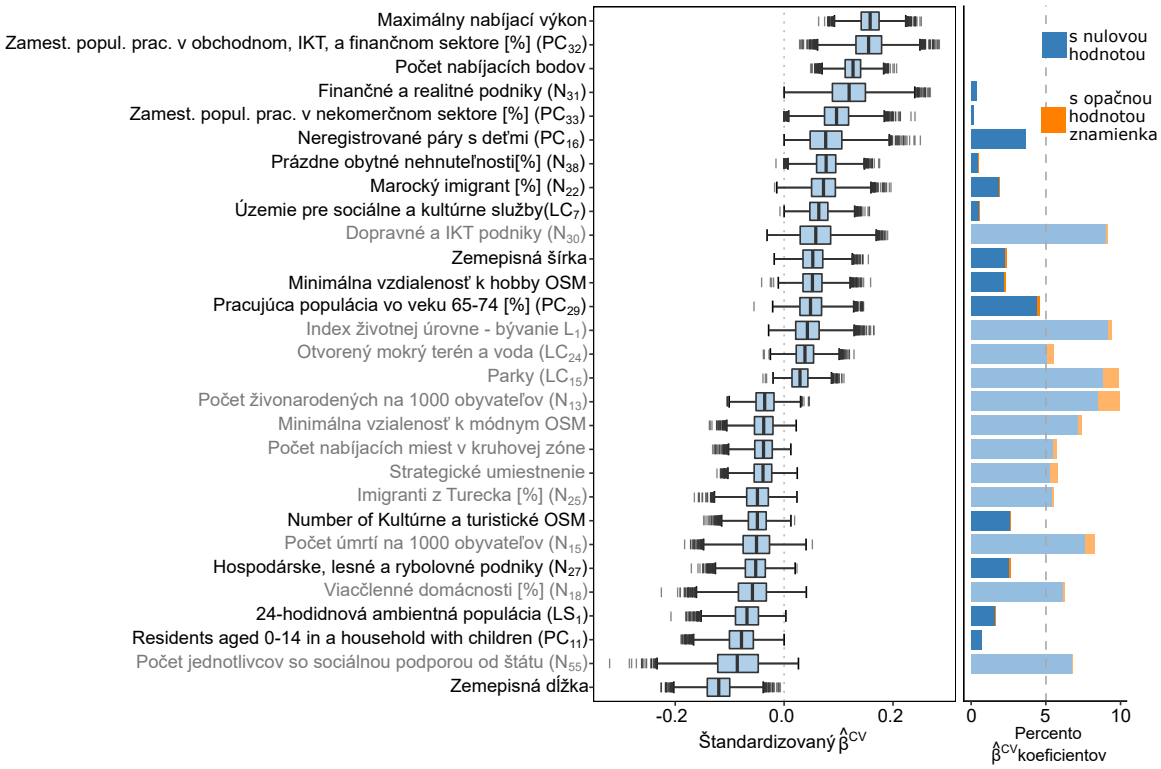
Obrázok 25: P-P grafy pre štatisticky signifikantné kombinácie štandardných pravdepodobnostných rozdelení s transformáciou $g(y)$ z tabuľky 12. V ľavom hornom rohu je použité pravdepodobnostné rozdelenie a transformačná funkcia $g(y)$. Čím bližšie sú body k diagonálnej čiare, tým lepšie funkcia hustoty pravdepodobnosti $f_Y(y)$ aproximuje vektor výstupu y.

B-2 Grafy použité pri odhade rozdelenia pravdepodobnosti spotreby energie

V tejto prílohe zobrazujeme aj Q-Q a P-P grafy odhadov pravedpodobnostných rozdelení spotrebovanej energie na nabíjacích miestach, ktoré sme neuvádzali v hlavnej práci z priestorových dôvodov.



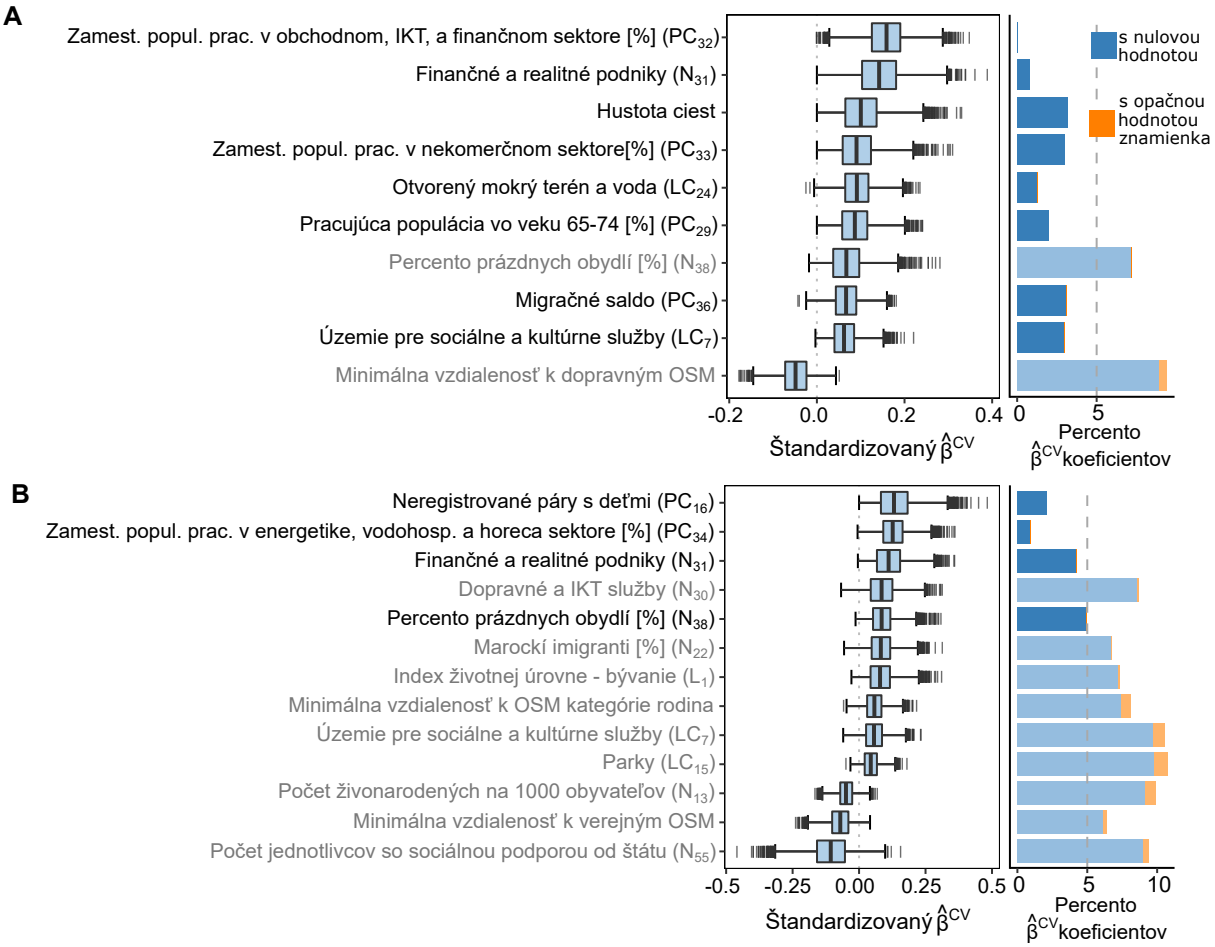
Obrázok 26: Q-Q grafy (obe osy sú transformované logaritmom) pre štatisticky signifikantné kombinácie štandardných pravdepodobnostných rozdelení s transformáciou $g(y)$ z tabuľky 12. V ľavom hornom rohu je použité pravdepodobnostné rozdelenie a transformačná funkcia $g(y)$. Čím bližšie sú body k diagonálnej čiare, tým lepšie funkcia hustoty pravdepodobnosti $f_Y(y)$ aproximuje vektor výstupu y.



Obrázok 27: Empirické rozdelenia štandardizovaných regresných koeficientov získaných lasso metódou kombinovanou s 10 zložkovou krížovou validáciou aplikovanou na 10 000 vzoriek bootstrapovaných dát. Zobrazujeme iba prediktory kde hodnota regresného koeficienta bola nastavená na nulu maximálne v 10 % vzoriek. Pre vysvetlenie spotrebovanej energie boli použité GIS dáta spolu s prediktormi EVnetNL dát. Koeficienty sú zoradené zostupe od najvyššej hodnoty mediánu po najnižšiu. Ľavý panel zobrazuje Tukeyho box-plot koeficientov. Na pravo, vrstvený stĺpcový graf ukazuje percento vzoriek, kedy bol regresný koeficient $\hat{\beta}^{CV}$ nastavený na nulu a počet vzoriek kedy dosiahol opačné znamienko ako medián. Za signifikantné považujeme tie prediktory (indikované tmavo modrou farbou), kde je počet vzoriek s nulovým koeficientom menší ako 5 % a počet vzoriek s opačným znamienkom je nízky. Svetlo šedá čiarkovaná čiara indikuje 5 % prahovú hodnotu. *Zamest. popul. prac. - zamestnaná populácia pracujúca

Príloha C - Výsledky štatistickej inferencie

Táto sekcia ukazuje grafy, ktoré sa nedostali do práce z dôvodu stručnosti a maximálneho využitia miesta. Obrázok 27 zobrazuje výsledok modelu vysvetľujúci spotrebu energie nabíjacej infraštruktúry pomocou GIS dát s prediktormi nabíjacích miest. Hlavný rozdiel medzi obrázkami 21 a 27 je, že niektoré prediktory prvého obrázka boli nahradené prediktormi nabíjacích miest v druhom obrázku, poukazujúc na možný vplyv prediktorov nabíjacej infraštruktúry na spotrebu energie. Obrázok 28 zobrazuje výsledky modelu vysvetľujúce spotrebu energie využitím stratifikácie podľa populácie obcí, s nabíjacími miestami umiestnenými v obciach s menej (**A**) a viac (**B**) ako 50 000 obyvateľmi.



Obrázok 28: Empirické rozdelenia štandardizovaných regresných koeficientov získaných lasso metódou kombinovanou s 10 zložkovou krížovou validáciou aplikovanou na 10 000 vzoriek bootstrapovaných dát. Zobrazujeme iba prediktory kde hodnota regresného koeficienta bola nastavená na nulu maximálne v 10 % vzoriek. V **A** sú nabíjacie miesta umiestnené v obciach s menej ako 50 000 obyvateľmi a v **B** nabíjacie miesta umiestnené v obciach s viac ako 50 000 obyvateľmi. Koeficienty sú zoradené zostupe od najvyššej hodnoty mediánu po najnižšiu. Ľavý panel zobrazuje Tukeyho box-plot koeficientov. Na pravo, vrstvený stĺpcový graf ukazuje percento vzoriek, kedy bol regresný koeficient $\hat{\beta}^{CV}$ nastavený na nulu a počet vzoriek kedy dosiahol opačné znamienko ako medián. Za signifikantné považujeme tie prediktory (indikované tmavo modrou farbou), kde je počet vzoriek s nulovým koeficientom menší ako 5 % a počet vzoriek s opačným znamienkom je nízky. Svetlo šedá čiarkovaná čiara indikuje 5 % prahovú hodnotu.*Zamest. popul. prac. - zamestnaná populácia pracujúca.