

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-182905-73688

Bc. Matúš Cuper

**IDENTIFIKÁCIA NEŠTANDARDNÉHO
SPRÁVANIA ODBERATEĽOV
V ENERGETICKEJ SIETI**

Diplomová práca

Vedúci práce: Ing. Marek Lóderer

apríl 2019

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-182905-73688

Bc. Matúš Cuper

**IDENTIFIKÁCIA NEŠTANDARDNÉHO
SPRÁVANIA ODBERATEĽOV
V ENERGETICKEJ SIETI**

Diplomová práca

Študijný program: Inteligentné softvérové systémy
Študijný odbor: 9.2.5 Softvérové inžinierstvo a 9.2.8 Umelá inteligencia
Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového
inžinierstva, FIIT STU v Bratislave
Vedúci práce: Ing. Marek Lóderer

apríl 2019

ČESTNÉ PREHLÁSENIE

Čestne prehlasujem, že diplomovú prácu som vypracoval samostatne pod vedením vedúceho diplomovej práce a s použitím odbornej literatúry, ktorá je uvedená v zozname použitej literatúry.

V Bratislave, 30.4.2019

.....

Matúš Cuper

POĎAKOVANIE

Ďakujem vedúcemu diplomovej práce Ing. Marekovi Lódererovi za odborné vedenie, cenné rady a pripomienky pri spracovaní diplomovej práce.

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Študijný program: Inteligentné softvérové systémy
Autor: Bc. Matúš Cuper
Diplomová práca: Identifikácia neštandardného správania odberateľov
v energetickej sieti
Vedúci diplomovej práce: Ing. Marek Lóderer
apríl 2019

V práci sme sa zamerali na identifikáciu anomálií v energetických časových radoch. Anomálie môžu vznikať na základe neštandardného správania odberateľov alebo poruchy inteligentného merača spotreby elektrickej energie. Cieľom diplomovej práce je identifikovať oba takéto prípady a znížiť tak straty distribučnej spoločnosti. Zároveň je nutné identifikovať iba také prípady, kedy sa jedná o dočasnú zmenu v správaní, či už je to dôsledkom zmeny počtu obyvateľov, počasia alebo výnimočnou udalosťou. So vznikajúcimi technológiami sa postupne mení aj profil spotreby odberateľov, a preto je nutné správne identifikovať aj nové trendy v dátach.

Analyzovali sme časové rady, anomálie a používané metódy na ich identifikáciu. Opísali sme problémy, ktoré vznikajú pri identifikácii anomálií v doméne energetiky, a ktorým musí čeliť aj naša metóda. Bližšie sme sa zamerali na zhlukovanie časových radoch, ktoré prináša nové prístupy do zhlukovania vysokodimenzionálnych dát, medzi ktoré patrí aj vyhľadzovanie, redukcia dimenzií alebo selekcia atribútov. Navrhovaná metóda zlúči diskretizované vyhľadené časové rady a následne sú identifikované anomálie na základe vytvorených zhlukov a rozloženia profilu používateľa v zhlukoch.

Annotation

Slovak University of Technology in Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Intelligent Software Systems
Author: Matúš Cuper
Master's thesis: Identification of abnormal behavior of customers in the power grid
Supervisor: Marek Lóderer
April 2019

In the thesis we focused on anomaly identification in energy time series. Anomalies can be caused by abnormal behavior of customers or a failure of intelligent meter of electricity load. The aim of this master thesis is to identify these mentioned cases and reduce electricity loss of distribution companies. Also it is necessary to identify only those cases, when the behavioral change is temporal, whether it is a result of different number of residents, weather or an exceptional occasion. Nowadays, the electricity load profile of customers is changing as new technologies are involved and therefore it is necessary to correctly identify new trends in data.

We also analyzed time series, anomalies and methods used for their identification. We described problems linked to identifications of anomalies in the domain of electricity, while our method is facing these problems as well. We focused on time series clustering, which brings new approaches to clustering of multidimensional data, including smoothing, dimension reduction and attribute selection. Our proposed method clusters discretized smoothed time series and then, subsequently identifies anomalies, based on created clusters and layout of customers profiles in clusters.

Obsah

1	Úvod	1
2	Analýza problému	3
2.1	Časové rady	3
2.1.1	Analýza časových radov	3
2.1.2	Zložky časových radov	4
2.1.3	Typy modelov časových radov	6
2.1.4	Delenie časových radov	7
2.2	Detekcia anomálií	7
2.2.1	Typy anomálií	8
2.2.2	Rozsah výskytu anomálií	10
2.2.3	Prístupy k identifikácii anomálií	11
2.3	Techniky detektie anomálií	11
2.3.1	Klasifikácia	11
2.3.2	Analýza najbližšieho suseda	12
2.3.3	Zhlukovanie	12
2.3.4	Štatistické metódy	12
2.3.5	Extrémna Studentova odchýlka	13
2.4	Metódy zhlukovania časových radov	14
2.4.1	Zhlukovanie na základe dočasnej susednosti	14
2.4.2	Zhlukovanie na základe reprezentácie	16
2.4.3	Zhlukovanie na základe modelu	16
2.4.4	Ďalšie prístupy k zhlukovaniu	16
2.4.5	Metriky vzdialenosťi	16
2.5	Predspracovanie dát	19
2.5.1	Filtrovanie odberateľov	20
2.5.2	Výber atribútov	20
2.5.3	Extrakcia čít	20
2.5.4	Reprezentácia FeaClip	20
2.5.5	Agregácia dát	21
2.5.6	Redukcia dimenzií	21
2.5.7	Segmentácia časových radov	22
2.5.8	Normalizácia číselných vektorov	24
2.5.9	Vyhľadzovanie časových radov	24
2.6	Anomálie v energetických časových radoch	24
2.7	Vyhodnocovacie metriky	26
2.7.1	Zhlukovacie validačné indexy	27
2.8	Súvisiace práce v doméne energetiky a identifikácií anomálií	28
2.9	Zhodnotenie analýzy	29
3	Návrh riešenia	31
3.1	Vytvorenie zhlukov	31
3.2	Skórovanie podozrivých zhlukov a inštancií	33
3.3	Selekcia podozrivých odberateľov	34
3.4	Vyhľadzovanie časových radov podozrivých odberateľov	36
3.5	Skórovanie odberateľov metódou S-H-ESD	37

4 Experimentálne overenie	41
4.1 Výber hyperparametrov zhlukovania	43
4.2 Vyhodnotenie navrhovanej metódy	46
4.3 Porovnanie existujúcich riešení	48
4.3.1 Identifikácia anomálií metódou S-H-ESD	49
4.3.2 Identifikácia anomálií metódou FeaClip	50
5 Zhodnotenie	51
Literatúra	53
Dodatok A Technická dokumentácia	I
Dodatok B Vizualizácie experimentov pre výber hyperparametrov	III
Dodatok C Plán práce na riešení projektu	IX
C.1 Plán do zimného semestra	IX
C.2 Plán do letného semestra	X
Dodatok D Opis digitálnej časti práce	XI

Zoznam obrázkov

1	Príklad trendovej zložky časového radu.	4
2	Príklad sezónnej zložky časového radu.	5
3	Príklad reziduálnej zložky časového radu.	5
4	Príklad multiplikatívneho modelu.	6
5	Príklad aditívneho modelu.	6
6	Príklad periodických časových radov.	7
7	Príklad neperiodických časových radov.	7
8	Príklad bodových anomalií.	9
9	Príklad kontextových anomalií.	9
10	Príklad skupinových anomalií.	10
11	Anomália detegované pomocou S-ESD algoritmu.	13
12	Anomálie detegované pomocou S-H-ESD algoritmu.	14
13	Príklad reprezentácie vytvorených zhľukov pomocou dendrogramu.	15
14	Príklad porovnávania časových radov pomocou dynamickej deformácií času.	18
15	Príklad porovnávania časových radov pomocou Euklidovskej vzdialenosťi. .	19
16	Príklad časového radu bez redukcie dimenzií.	22
17	Príklad redukovaného časového radu.	23
18	Príklad redukcie dimenzií pomocou PCA.	23
19	Čiastočné zníženie spotreby elektrickej energie.	25
20	Úplné zníženie spotreby elektrickej energie.	25
21	Výsledky Shafferovho testu validačných indexov.	28
22	Stavový diagram procesu identifikácií anomalií.	32
23	Skórovanie podozrivých inštancií a zhľukov.	35
24	Reprezentácia spotrebiteľov pomocou metódy FeaClip.	36
25	Vyhľadenie časových radov pomocou metódy LOESS.	37
26	Vyhľadenie príznakov anomálnosti pomocou metódy LOESS.	37
27	Vizualizácia skóre odberateľa pred pridaním S-H-ESD.	38
28	Vizualizácia skóre odberateľa po pridaní S-H-ESD.	39
29	Krabicový graf spotreby odberateľov.	41
30	Krabicový graf spotreby odberateľov počas pracovných dní.	42
31	Krabicový graf spotreby odberateľov počas dní voľna.	42
32	Porovnanie veľkosti posuvného okna a počtu zhľukov pre Silhouetteov index.	43
33	Porovnanie veľkosti posuvného okna a počtu zhľukov pre Davies-Bouldinov index.	44
34	Porovnanie vzdialostných metrík pomocou Silhouetteovho indexu.	45
35	Porovnanie veľkostí a typov posuvných okien pomocou Silhouetteovho indexu.	45
36	Vizualizácia vytvoreného zhľukovania so zvýraznenými medoidmi.	46
37	Vizualizácia vypočítaného skóre pre odberateľa 2172.	47
38	Vizualizácia vypočítaného skóre pre odberateľa 6536.	48
39	Vizualizácia všetkých odberateľov, anomálie identifikované pomocou S-H-ESD.	49
40	Vizualizácia S-H-ESD skóre odberateľov.	50
41	Porovnanie veľkosti posuvného okna a počtu zhľukov (Silhouetteov index). .	III
42	Porovnanie veľkosti posuvného okna a počtu zhľukov (Calinski-Harabaszov index).	III
43	Porovnanie veľkosti posuvného okna a počtu zhľukov (Davies-Bouldinov index). .	IV

44	Porovnanie veľkosti posuvného okna a počtu zhlukov (Modifikovaný Davies-Bouldinov index).	IV
45	Porovnanie veľkosti posuvného okna a počtu zhlukov (Silhouetteov index).	IV
46	Porovnanie veľkosti posuvného okna a počtu zhlukov (Calinski-Harabaszov index).	V
47	Porovnanie veľkosti posuvného okna a počtu zhlukov (Davies-Bouldinov index).	V
48	Porovnanie veľkosti posuvného okna a počtu zhlukov (Modifikovaný Davies-Bouldinov index).	V
49	Porovnanie vzdialenosných metrík (Silhouetteov index).	VI
50	Porovnanie vzdialenosných metrík (Calinski-Harabaszov index).	VI
51	Porovnanie vzdialenosných metrík (Davies-Bouldinov index).	VI
52	Porovnanie vzdialenosných metrík (Modifikovaný Davies-Bouldinov index).	VII
53	Porovnanie veľkostí a typov posuvných okien (Silhouetteov index).	VII
54	Porovnanie veľkostí a typov posuvných okien (Calinski-Harabaszov index).	VII
55	Porovnanie veľkostí a typov posuvných okien (Davies-Bouldinov index).	VIII
56	Porovnanie veľkostí a typov posuvných okien (Modifikovaný Davies-Bouldinov index).	VIII

Zoznam tabuľiek

1	Matica zámen (angl. <i>Confusion matrix</i>)	26
2	Validačná matica zhľukovania časových radov	27
3	Atribúty metódy FeaClip a ich opis.	36
4	Charakteristiky polohy použitého datasetu.	41
5	Charakteristiky polohy po rozdelení datasetu.	43
6	Porovnanie výsledkov spracovania existujúcich riešení.	48
7	Použité knižnice jazyka R.	I

1 Úvod

Jedným z problémov, ktorým v súčasnosti čelia distribučné spoločnosti, je detekcia neštan-dardného správania odberateľov. Jej úlohou je identifikovať profily zákazníkov, ktorí svojím správaním porušujú stanovené podmienky a manipulujú s hodnotami nameranými meračmi za cieľom obohatenia sa. Samozrejme tiež dochádza k prípadom, kedy je presnosť meracieho zariadenia nižšia aj bez zapríčinenia zákazníka. Ďalším faktorom ovplyvňujúcim predikciu spotreby elektrickej energie je nepredvídateľné správanie zákazníka. Ide o prípady, kedy sú dodržané zmluvné podmienky a meracie zariadenie nevykazuje chybu s významnou veľkosťou. Všetky prípady sú pre distribučnú spoločnosť nežiaduce a je v záujme zníženia strát a optimalizácie distribúcie, ich čo najskôr identifikovať a dodatočne monitorovať. Odhalením neobvyklého správania je možné zvýšiť stabilitu distribučnej siete, zvýšiť presnosť predikcie distribútora, prípadne rozšíriť kapacitu siete a pripraviť ju tak na budúce trendy. Obvykle sú za účelom odhalenia vykonávané náhodné kontroly, ktoré pokrývajú iba nízky počet zákazníkov s anomálnym správaním. Na základe množstva dát získavaných z inteligentných meračov je možné modelovať správanie buď jednotlivých zákazníkov alebo aj rovnakých skupín používateľov. Distribučné spoločnosti tak žnižujú straty pri dodávke energie a zároveň dokážu preverovať iba odberateľov, ktorí svojím profilom nezapadajú medzi odberateľov so štandardným správaním.

Cieľom práce je identifikovať v časových radoch spotreby elektrickej energie také invervaly, ktoré svojou charakteristikou nezodpovedajú správaniu ostatných odberateľov v datasete. Vzhľadom na fakt, že manuálne označenie takýchto intervalov je časovo a finančne náročné, je metóda založená na učení bez učiteľa. Na druhej strane predpokladá veľké množstvo dát odberateľov, na základe ktorých sú medzi nimi identifikované podobnosti a odlišnosti. Miera odlišnosti definuje aj mieru anomálnosti daného intervalu. Spracovanie a transformácia veľkých datasetov je dosiahnutá pomocou štatistických metód a strojového učenia.

Práca je rozdelená do piatich kapitol vrátane úvodu. V kapitole 2 sa zaoberáme analýzou existujúcich riešení, rôznymi prístupmi pri identifikácii anomalií, zhlukovaním časových radoch, ale aj ich transformáciou alebo redukciami dimenzií. Osobitná pozornosť je venovaná štatistickým metódam, ktorých cieľom je identifikovať anomálie a zmeny v časových radoch. V kapitole 3 navrhujeme vlastné riešenie, ktoré na základe zhlukovania časových radoch počíta skóre anomálnosti pre každého odberateľa. Skóre je ďalej použité pri bližšej identifikácii intervalov, ktorých priebeh je neobvyklý prípadne náhodný. V kapitole 4 sú bližšie popísané vykonané experimenty a porovnanie navrhovaného riešenia s existujúcimi. V poslednej kapitole 5 sú zhodnotené výsledky našej práce.

2 Analýza problému

Tak ako je spomenuté v článku [27], straty v distribučných sieťach v niektorých krajinách tvoria až 30% z celkového objemu distribuovanej energie. Väčšinu strát vytvára svojimi vlastnosťami samotná sieť, no nezadanbateľnú časť tvoria aj nelegálne odbery. Pravidelná kontrola všetkých odberateľov by bola časovo aj finančne náročná, preto je potrebné správne identifikovať zákazníkov s neštandardnou spotrebou energie, čím sa minimalizujú náklady spojené s kontrolami. Zatiaľ čo v minulosti bola možná identifikácia nelegálnych odberov len fyzickou kontrolou, dnes vieme obmedziť okruh podozrivých aj na diaľku, keďže inteligentné merače nám poskytujú dátá v pravidelných intervaloch s minimálnou odchýlkou.

Vďaka tomu vznikajú nové možnosti identifikácie neštandardného správania využitím dátovej analytiky a strojového učenia. Zatiaľ čo väčšina algoritmov na identifikáciu anomálií pracuje s nízkorozmernými dátami, časové rady predstavujú presný opak a použité metódy sa líšia od tých klasických. Výzvou pri skupinových a kontextových anomáliách je aj vhodný výber premenných, na základe ktorých budú anomálie identifikované. Zvýšenie presnosti pri hľadaní anomálií môžeme docieliť kombinovaním rôznych zdrojov dát, či už by sa jednalo o počasie alebo údaje z inteligentných meračov iných druhov energie. Cieľom tejto kapitoly je preto analyzovať a porovnať používané metódy pri detekcii anomálií v časových radoch a zamerať sa najmä na vhodnú reprezentáciu jednotlivých odberateľov pomocou získaných dát.

2.1 Časové rady

Meranie časových radoch predstavujú množinu dátových bodov, usporiadané v chronologickom poradí. Takúto množinu môžeme definovať ako množinu vektorov $x(t)$, kde premenná x predstavuje časový rad a t čas, kedy bolo meranie vykonané. Časové rady pozostávajúce z merania jednej veličiny sa nazývajú jednorozmerné, pri meraní viacerých veličín sa jedná o viacrozmerné časové rady. Tiež ich môžeme rozdeliť na spojité a diskrétné. Spojité časové rady merajú pozorovanú veličinu v každej jednotke času. Môže sa jednať napr. o počasie, veľkosť prietoku rieky alebo koncentráciu látok pri chemických procesoch. Diskrétné časové rady sú pozorované spravidla v rovnakých časových intervaloch, napr. rokoch, dňoch či minútach. Stretnúť sa s nimi môžeme pri kurzoch mien, produkcií štátov či spotrebe elektrickej energie [1].

2.1.1 Analýza časových radoch

Časové rady môžeme reprezentovať pomocou matematického modelu, ktorého parametre sú dané nameranými dátami. Parametre sú určené na základe dátovej analýzy nazhromaždených dát. Cieľom je určiť parametre tak, aby predikcia výsledného modelu bola čo najpresnejšia. Proces analýzy a úpravy parametrov je možné opakovať pokial model nedosahuje dostatočne uspokojivé výsledky [1].

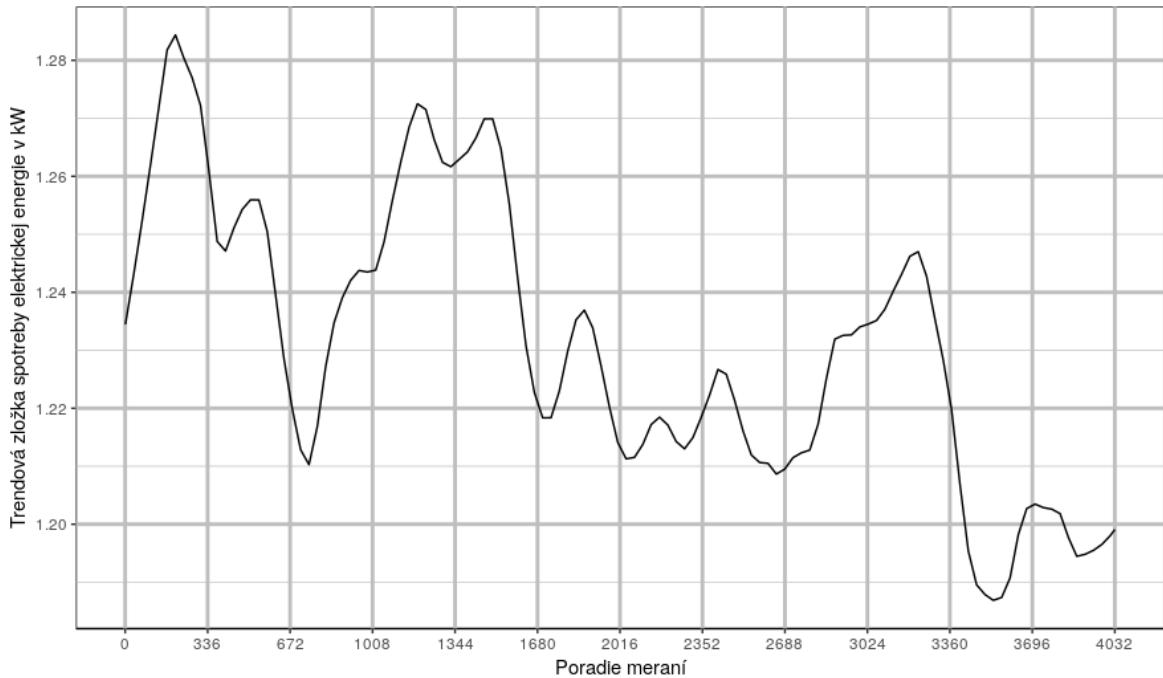
Premenná \hat{x} vo vzorci 1 predstavuje predikovanú hodnotu časového radu x . Cieľom je nájsť funkciu $f(x)$, ktorá predikuje budúce hodnoty časového radu x tak, aby boli čo najpresnejšie, konzistentné a objektívne [36].

$$\hat{x}(t + \Delta_t) = f(x(t - a), x(t - b), x(t - c), \dots) \quad (1)$$

2.1.2 Zložky časových radov

Na vývoj časových radov vplývajú ich jednotlivé komponenty, z ktorých pozostávajú. Ich vývoj je ovplyvnení rôznymi faktormi, či už ekonomickými, počasím, sviatkami alebo kultúrou. Priebehy grafov jednotlivých komponentov potom môžu byť cyklické, rastúce, klesajúce alebo stagnujúce v závislosti od toho, či existuje zmena, ktorá je trvalá alebo opakujúca. Taktiež aj veľkosť periód tohto cyklu môže byť rôzna, a to niekoľko hodín, dní, mesiacov či rokov. Keďže prostredie, v ktorom meriame predpovedanú veličinu sa vyvíja, rovnako sa vyvíja aj správanie pozorovanej veličiny. Preto je potrebné pri modelovaní správania uvažovať jednotlivé komponenty časového radu. V literatúre sa najčastejšie stretávame s rozdelením na 4 komponenty, a to trendovú, cyklickú, sezónnu a reziduálnu zložku [16].

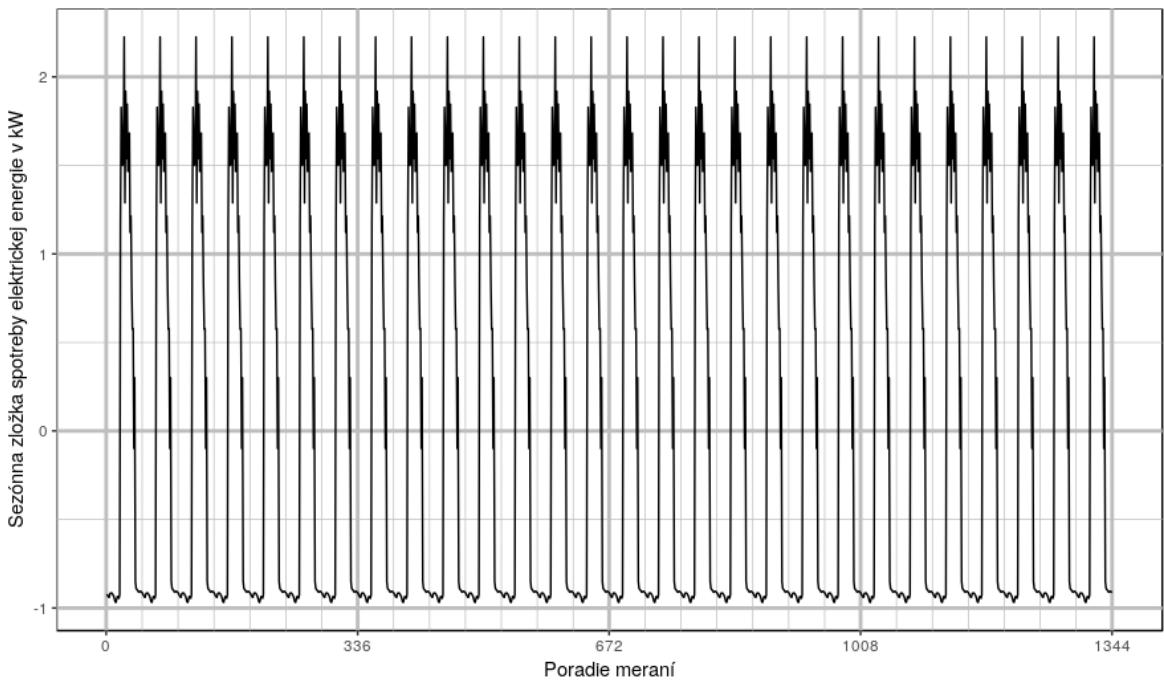
Trendová zložka zastupuje dlhodobé správanie časového radu. Ide o dlhodobé klesanie, rast alebo stagnáciu časového radu. Príkladom môže byť neustále predĺžovanie priemernej doby dožitia alebo aj rast svetovej populácie. Priebeh dekomponovanej trendovej zložky môžeme vidieť na obrázku 1 [1].



Obr. 1: Príklad trendovej zložky časového radu.

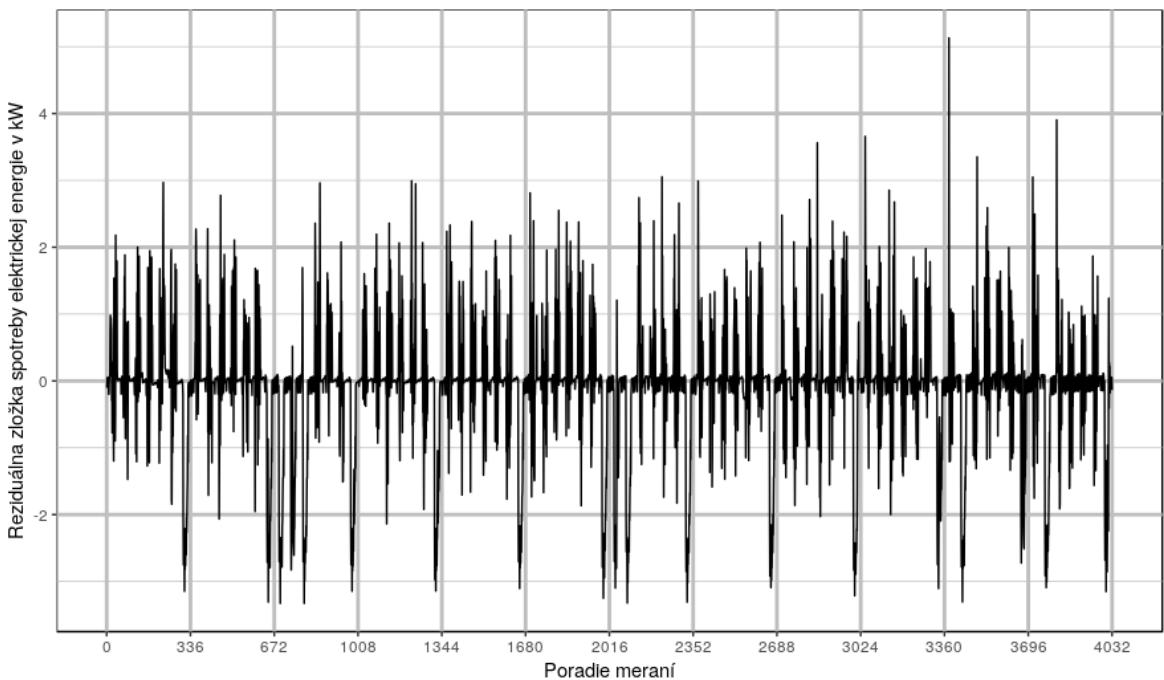
Cyklická zložka predstavuje strednodobú opakujúcu sa zmenu. Najčastejšie sa pri tom jedná o obdobie 2 a viac rokov. Táto zložka býva výrazne zastúpená pri ekonomických a finančných časových radoch. Príkladom môže byť aj podnikateľský cyklus, ktorý pozostáva zo 4 opakujúcich sa fáz [1].

Sezónna zložka sa počas roka mení a predstavuje tak striedanie ročných období. Priebeh funkcie je ovplyvňovaný najmä podnebnými podmienkami a počasím, ale aj kultúrou, náboženstvom či tradíciami. Príkladom môže byť predaj sezónnych výrobkov, ktorý sa počas roka výrazne mení. Priebeh funkcie dekomponovanej zložky môžeme vidieť na obrázku 2 [1].



Obr. 2: Príklad sezónnej zložky časového radu.

Reziduálna zložka v literatúre často označovaná aj ako náhodná zložka alebo biely šum, predstavuje nepredvídateľnú veličinu, ktorá nesystematicky ovplyvňuje pozorovaný časový rad. Metóda jej merania zatiaľ nie je v štatistike definovaná. Priebeh funkcie nemá žiadny vzor a môže vznikať na základe prírodných katastrof, ale aj nepredvídateľnej zhody náhod. Príklad priebehu môže byť aj graf znázornený na obrázku 3 [1].

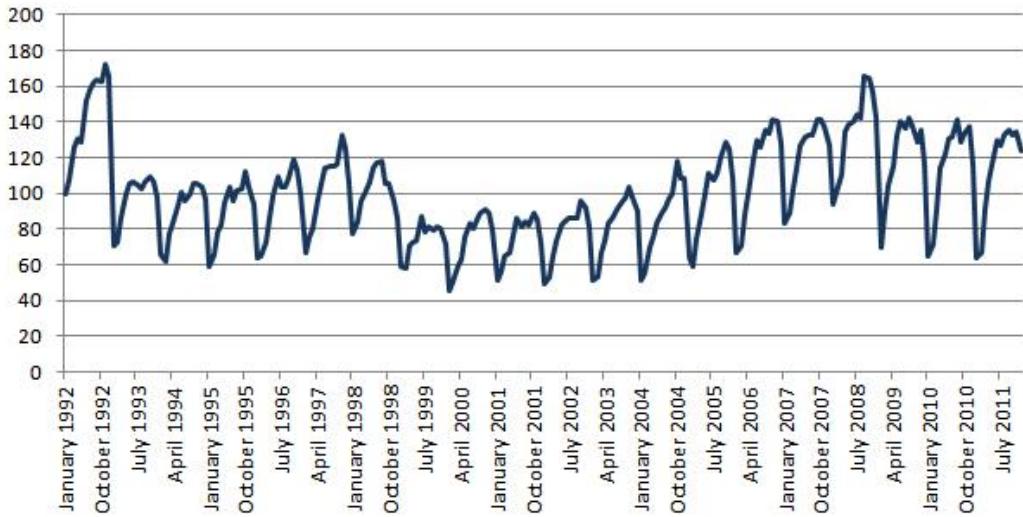


Obr. 3: Príklad reziduálnej zložky časového radu.

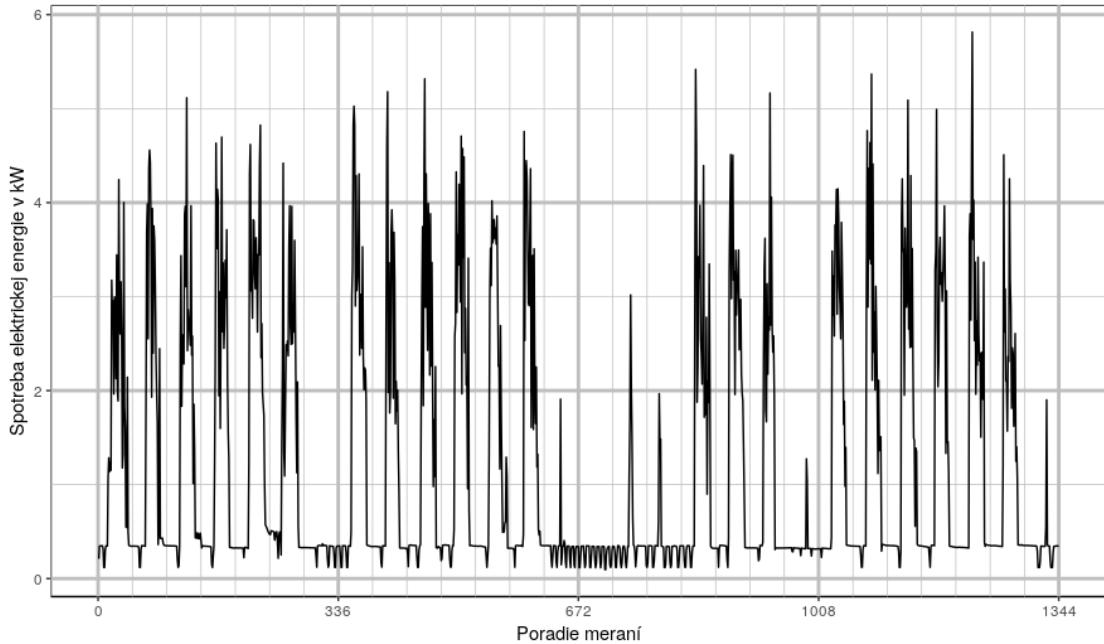
2.1.3 Typy modelov časových radov

Kombináciou komponentov časových radov identifikovaných v predchádzajúcej kapitole vznikajú 2 typy modelov, aditívny a multiplikatívny.

$$\begin{aligned} Y(t) &= T(t) \times S(t) \times C(t) \times I(t) \\ Y(t) &= T(t) + S(t) + C(t) + I(t) \end{aligned} \quad (2)$$



Obr. 4: Príklad multiplikatívneho modelu, index stavebnej produkcie Slovenska, Eurostat.



Obr. 5: Príklad aditívneho modelu, spotreba elektrickej energie v regióne, Slovensko.

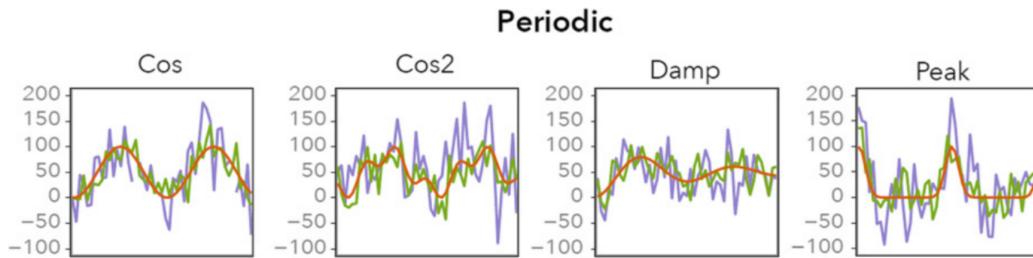
Vo vzorci 2, $Y(t)$ predstavuje meranie pozorovanej veličiny v čase t . Ostatné premenné T , S , C a I reprezentujú trendový, sezónny, cyklický a reziduálny komponent. Veličiny

multiplikatívneho modelu sa môžu vzájomne ovplyvňovať, zatiaľ čo pri aditívnom modeli predpokladáme ich nezávislosť. Multiplikatívny model je znázornený na obrázku 4 a aditívny na obrázku 5 [1].

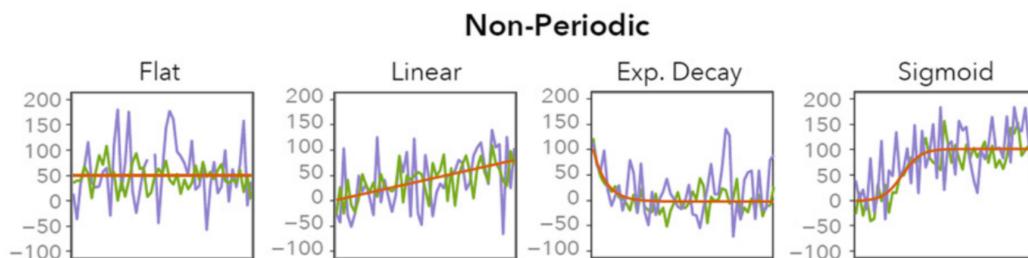
2.1.4 Delenie časových radov

Výraznými vlastnosťami časových radov sú aj synchrónnosť a periodicita, znázornená na obrázkoch 6 a 7. Vznikajú tak 4 nasledujúce kategórie [42]:

- **Periodické a synchrónne časové rady** predstavujú najjednoduchšiu kombináciu, keďže každý časový rad má konštantnú časovú periódu a zároveň sú všetky časové rady časovo zarovnané na konkrétny časový bod.
- **Neperiodické a synchrónne časové rady** nemajú žiadnu periodicitu, ale opäť sú časovo zarovnané.
- **Periodické a asynchronné časové rady** nie sú časovo zarovnané, ale obsahujú periodicitu, čiže začiatok períody v každom časovom rade je iný.
- **Neperiodické a asynchronné časové rady** predstavujú skupinu, do ktorej spadajú ostatné časové rady, ktoré neobsahujú periodicitu ani synchrónnosť.



Obr. 6: Príklad periodických časových radov [31].



Obr. 7: Príklad neperiodických časových radov [31].

2.2 Detekcia anomálií

Anomálne správanie alebo anomália je definovaná ako vzor v správaní, ktorý nezodpovedá štandardnému správaniu. Pri dátach z inteligentných meračov, anomália zodpovedá meraniu, ktoré sa nenachádza v oblasti normálnych dát.

Pri identifikácii anomálií je najskôr potrebné zamyslieť sa nad nasledovnými problémami [6]:

- **Definovanie oblasti normálnych dát** je veľmi náročné, nakoľko hranica medzi normálnymi dátami a anomáliami je nepresná a môže tak dôjsť k nesprávnemu označeniu meraní.
- **Anomálie vytvorené škodlivou činnosťou** sa javia ako normálne dáta, čo sťaže definíciu normálneho správania.
- **Evolúcia dát** spôsobuje, že definícia normálneho správania sa môže časom zmeniť.
- **Presná predstava o anomálii** je často rôzna naprieč viacerými odbormi, a preto neexistuje univerzálny spôsob na určovanie anomálií.
- **Dostupnosť označených dát** zlepšuje presnosť identifikácie anomálií, avšak často takéto dátu neexistujú alebo ich je potrebné označiť, čo spravidla býva drahé.
- **Biely šum** vyskytujúci sa v dátach má tendenciu skresľovať normálne dáta, ktorých identifikácia je následne zložitá.

Na detekciu anomálií sa bežne používajú algoritmy určené na klasifikáciu, ako je napríklad naivný Bayesovský klasifikátor (angl. *Naive Bayes*), k-najbližší susedia (angl. *k-nearest neighbors*), rozhodovacie stromy (angl. *decision tree*), náhodné lesy (angl. *random forests*), neurónové siete so spätnou propagáciou (angl. *neural networks with backpropagation*) alebo metóda podporných vektorov (angl. *support vector machine*) [9].

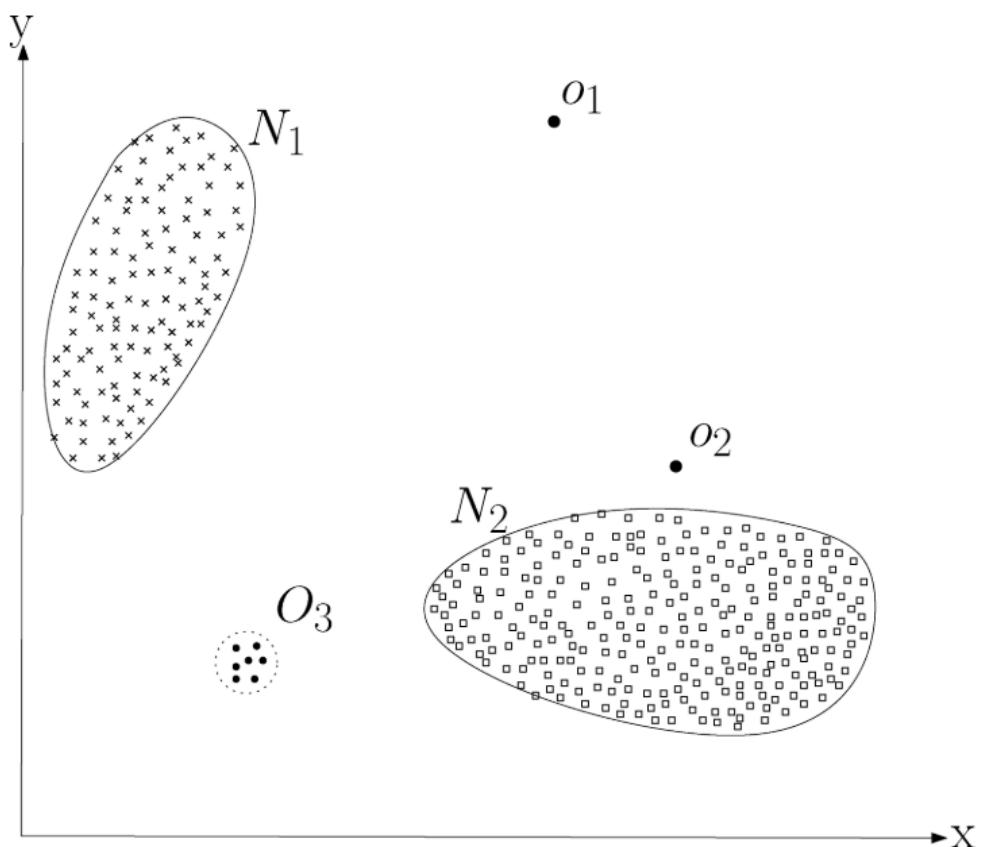
2.2.1 Typy anomálií

Dôležitým aspektom pri uplatnení detektie anomálií je charakter anomálie. Z tohto dôvodu môžeme anomálie rozdeliť do nasledujúcich troch skupín.

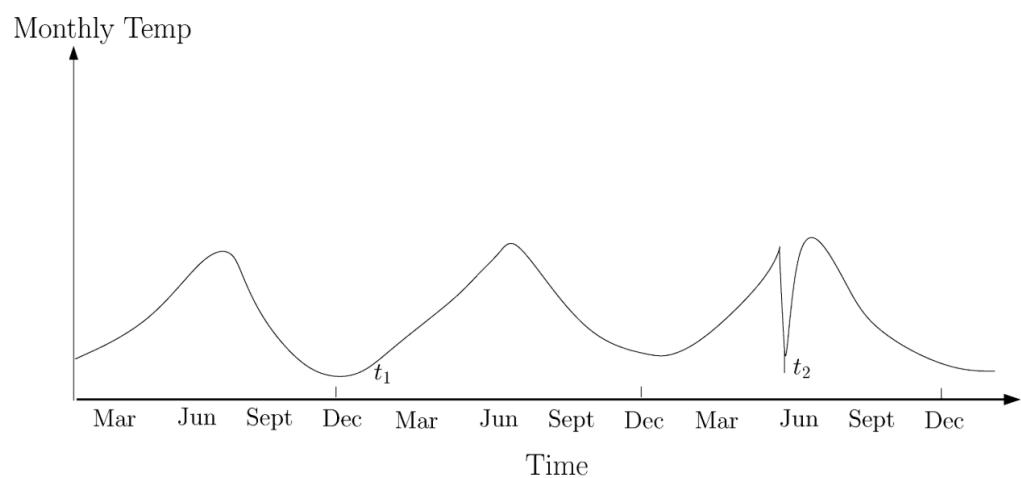
Bodové anomálie predstavujú inštancie, ktoré sa nenachádzajú v oblasti normálnych dát a je možné ich detegovať jednotlivo. Jedná sa o najjednoduchší typ anomálie a sústreďuje sa naň väčšina výskumov. Príkladom zo skutočného života môže byť detekcia podvodov s kreditnými kartami, kedy transakcia výrazne väčšieho objemu peňazí predstavuje podvod, zatiaľ čo ostatné transakcie, nachádzajúce sa v normálnom rozsahu predstavujú normálne dátu, ktoré nie sú anomáliou [6].

Kontextové anomálie predstavujú inštancie, ktoré sa nachádzajú v oblasti normálnych dát, ale v špecifickom kontexte sú považované za anomáliu. Kontext je daný kontextovými atribútmi v dátach, na základe ktorých sa určujú susedné inštancie. Nekontextové atribúty, nazývané aj behaviorálne, reprezentujú meranú veličinu. Napríklad pri meteorologických meraniach, budú informácie o polohe alebo nadmorskej výške predstavovať kontextové atribúty, zatiaľ čo množstvo zrážok alebo slnečných hodín budú behaviorálne atribúty [38].

Anomálne správanie inštancií je dané behaviorálnymi atribútmi v určitom kontexte. Čiže ak inštancia s danými behaviorálnymi atribútmi je považovaná za normálnu, iná inštancia s rovnakými behaviorálnymi, ale s rôznymi kontextovými atribútmi môže byť považovaná za anomáliu. Kontextové anomálie boli najčastejšie identifikované v časových radoch. Príkladom môžu byť opäť transakcie väčšieho objemu peňazí, ktoré sú bežné v období pred Vianocami, ale neštandardné v inom ročnom období. Zatiaľ čo v niektorých prípadoch je definovanie kontextu priamočiare, existujú domény, kde to jednoduché nie je. Dôležité je aby kontextové atribúty boli zmysluplné určené v cieľovej doméne ich aplikácie [6, 38].

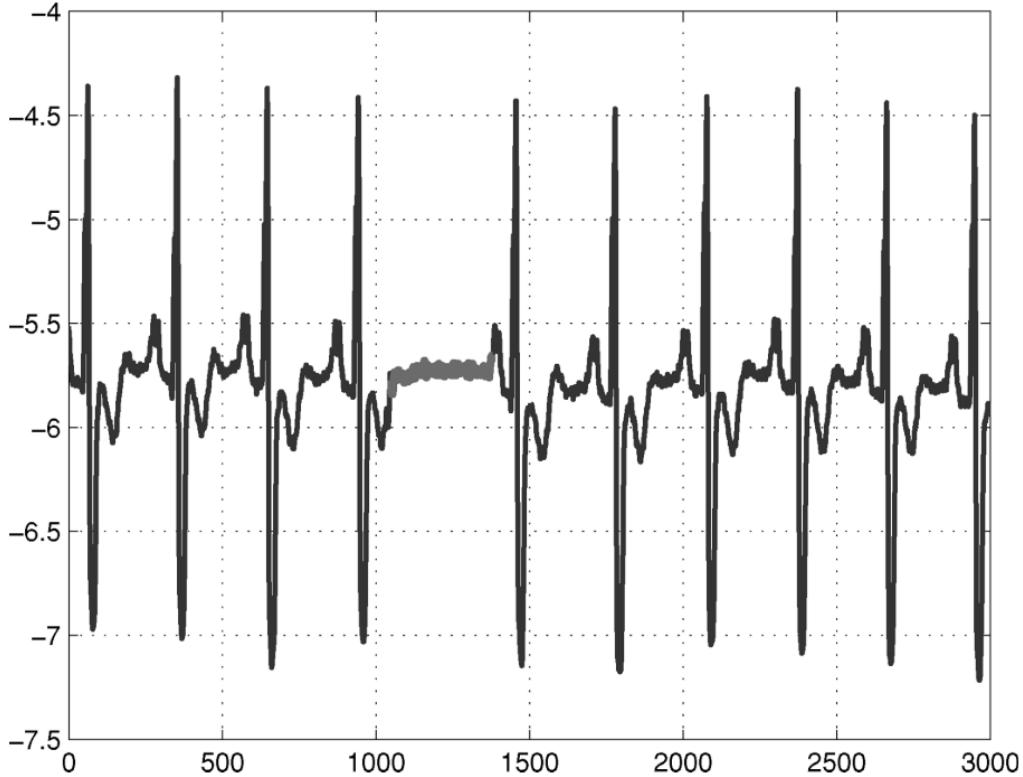


Obr. 8: Príklad bodových anomálií [6].



Obr. 9: Príklad kontextových anomálií [6].

Skupinové anomálie sa nachádzajú v oblasti normálnych dát, ale skupina týchto inštancií tvorí spolu anomáliu. Vzniknutá anomália obsahuje sekvenciu inštancií, ktorá by pri inom zoradení nepredstavovala anomáliu. Taktiež sa jednotlivé inštancie môžu nachádzať v rozsahu normálnych dát. Príkladom môžu byť systémové volania operačného systému, ktoré sú v prípade dodržania určitej postupnosti označené ako činnosť škodlivého softvéru [15].



Obr. 10: Príklad skupinových anomálií [6].

Zatiaľ čo bodové anomálie sa môžu vyskytovať v každom datasete, skupinové sa vyskytujú iba v datasetoch, kde existuje medzi inštanciami vzťah. Pri kontextových anomáliách je potrebné určiť kontextové atribúty, ktoré sa v niektorých datasetoch ani nemusia nachádzať. Problém detektie bodových a skupinových anomálií je možné transformovať na problém detektie kontextových anomálií, v prípade, že sa prihliada na kontext jednotlivých inštancií. Techniky používané pri detekcii skupinových anomálií sa značne líšia od techník používaných pri bodových a kontextových anomáliách [6, 15].

2.2.2 Rozsah výskytu anomálií

Za anomáliu v našej doméne považujeme správanie odberateľa, ktoré sa výrazne líši od ostatných odberateľov. Anomáliou môže byť celé pozorované obdobie alebo iba jeho určitá časť. Keďže datasety, ktoré máme k dispozícii obsahujú konečný počet meraní a teda nie sú spojité, anomália môže byť reprezentovaná aj jediným meraním. Anomálie môžeme taktiež rozdeliť na pozitívne a negatívne, v závislosti od toho, aké sú očakávané a reálne hodnoty. Ak je ich rozdiel kladný hovorí o pozitívnych anomáliach, inak o negatívnych [21].

Intervaly jednotlivých časových radov, ktoré metóda označí ako anomálne, môžeme ďalej rozdeliť na lokálne a globálne anomálie. Delenie vzniká na základe dekompozície časových radov, kde globálne anomálie sú porovnávané so sezónnou zložkou a lokálne anomálie sú

identifikované vnútri sezónnych vzorov. Zatiaľ čo globálne anomálie sú identifikované zväčša na základe porovnávania očakávaných a reálnych hodnôt, identifikácia lokálnych anomálií je náročnejšia ak má byť navrhované riešenie robustné. Opäť vychádzame z porovnávania očakávaných a reálnych hodnôt, no jedná sa o menšie intervale, ktorých sa môže vyskytovať rádovo viac. Robustné riešenie to musí zohľadniť a identifikovať iba signifikantné anomálie [21].

2.2.3 Prístupy k identifikácii anomálií

V praxi sa stretávame s datasetmi, ktoré sa líšia v množstve označených dát, počte typov anomálií, ktoré budeme detegovať alebo aj pomerom medzi normálnymi inštanciami a tými neštandardnými. Často je označovanie inštancií vykonávané manuálne ľudskými expertmi drahé a neefektívne. Taktiež proces späťnej väzby môže byť zdĺhavý a nepraktický. Z toho dôvodu je dôležité zvoliť správny prístup pri identifikácii anomálií. V súčasnosti existujú 3 prístupy, a to detekcia anomálií s učiteľom (angl. *supervised learning*), bez učiteľa (angl. *unsupervised learning*) a ich kombinácia (angl. *semi-supervised learning*) [6].

Detekcia bez učiteľa nepotrebuje označené trénovacie dáta, vďaka čomu je široko aplikovateľná a často používaná. Vychádza z predpokladu, že normálne inštancie majú majoritné zastúpenie v množine. Ak táto podmienka nie je splnená, môže často dochádzať k falošnému alarmu [6].

Detekcia s učiteľom potrebuje trénovacie dáta s označenými inštanciami ako normálnymi, tak aj anomálnymi. Cieľom je vytvoriť prediktívny model, ktorého úlohou je určiť triedu inštancie. Problémom je nepomer anomálnych inštancií v porovnaní s normálnymi a ich označenie ľudským expertom môže byť časovo a finančne náročné [6].

Kombinované učenie je kombináciou predchádzajúcich dvoch prístupov a počíta s označenou iba jednou triedou inštancií. Typicky sú označené normálne inštancie, keďže ich identifikácia je menej náročná. V takom prípade je vytvorený model pre normálnu triedu a identifikácia anomálií prebieha v testovacej vzorke dát [6].

2.3 Techniky detekcie anomálií

Detegovať anomálie rôznych typov môžeme niekoľkými spôsobmi, čo závisí aj od samotných dát. Ich úplnosť, množstvo a oblasť, v ktorej boli zozbierané sú kritické pre správny výber techniky, pomocou ktorej budú anomálie identifikované. Nás budú zaujímať najmä detekcie anomálií v časových radoch. Analyzované metódy sú najmä z oblasti strojového učenia a dátovej analýzy, ale pre úplnosť sú spomenuté aj iné používané metódy.

2.3.1 Klasifikácia

Pomocou naučeného modelu, nazývaného aj klasifikátor, sú rozoznávané triedy jednotlivých inštancií. Pri detekcii anomálneho správania, klasifikátor rozlišuje iba medzi dvoma triedami, triedou normálnych dát a anomálií. Vzhľadom na to, že na natrénovanie klasifikátora sú potrebné označené dáta, ide o učenie s učiteľom. Na implementovanie klasifikátora môžeme použiť techniky založené na rôznych typoch neurónových sietí, Bayesových sieťach, pravidlových systémoch či metóde podporných vektorov [6, 41].

2.3.2 Analýza najbližšieho suseda

Metóda určí na základe vzdialenosť alebo podobnosti medzi dátovými inštanciami, či sa jedná o normálnu inštanciu alebo anomáliu. To je vypočítané pomocou vzdialenosťí medzi testovanou inštanciou a všetkými bodmi, alebo iba k najbližším bodmi. Pri viacozmerných dátach je vzdialenosť určovaná pre každú dimenziu zvlášť. Metóda je založená na predpoklade, že zatiaľ čo normálne inštancie sa nachádzajú pri sebe a sú husto usporiadane, anomálie sú vzdialenejšie, prípadne na okraji vzniknutých oblastí. Aplikácia je možná pomocou techník založených na relatívnej hustote alebo vzdialosti najbližších k susedných inštancií [41].

2.3.3 Zhlukovanie

Jedná sa o učenie bez učiteľa, keďže zhluky inštancií sú vytvorené na základe ich vzdialenosťí či podobnosti. Techniky ďalej delíme do kategórií na základe predpokladu o dátových inštanciách [41].

Prvá kategória predpokladá, že normálne inštancie patria do zhluku, zatiaľ čo anomálne nepatria do žiadneho. Používané sú zhlukovacie algoritmy ako DBSCAN alebo ROCK, pri ktorých nie nutne každá inštancia musí patriť do zhluku. Nevýhodou algoritmov môže byť neoptimálne použitie pri detekcii anomálií, keďže sú primárne určené na riešenie zhlukovacích problémov [6].

Druhá kategória predpokladá, že normálne inštancie ležia v blízkosti najbližšieho centroidu a anomálne inštancie sú od neho vzdialené. Algoritmy väčšinou pozostávajú z dvoch krokov, v prvom sú inštancie pridelené do zhluku a v druhom je vypočítané ich anomálne skóre na základe vzdialenosťi od centroidu daného zhluku. Používanými algoritmami sú neurónové siete (konkrétnie SOM) alebo algoritmus k-means, ktoré sa môžu natrénovať aj pomocou kombinovaného učenia. Do rovnakej skupiny spadá aj metóda k-medoids, ktorá funguje podobne ako k-means, rozdielom je výpočet centroidov. Pri metóde k-medoids je centroidom inštancia, ktorej vzdialenosť od všetkých ostatných inštancií je minimálna. Pri k-means centroidom nemusí byť reálna inštancia [6].

Posledná kategória pracuje s predpokladom, že normálne inštancie sú súčasťou veľkých a hustých zhlukov, na druhej strane anomálie patria do malých a riedkych zhlukov. Používanými algoritmami sú napr. CBLOF (angl. *Cluster-Based Local Outlier Factor*) alebo $k\text{-}d$ stromy. V princípe algoritmy najskôr vytvárajú zhluky a až potom určujú, na základe ich hustoty, či sa jedná o normálne zhluky alebo anomálie. Zhluk je vytvorený iba v prípade, že inštancia sa nachádza mimo preddefinovaného rádiusu od centra daného zhluku [35].

2.3.4 Štatistické metódy

Jedná sa o súbor metód založených na štatistikke. K jednotlivým výpočtom sú väčšinou používané priemerné hodnoty, odchýlky, atď. V praxi sa používajú metódy kĺzavého priemeru, $3 \cdot \sigma$ pravidlo, dekompozícia časových radov, ale aj metóda extrémnej Studentovej odchýlky, ktorá je bližšie opísaná v nasledujúcej podkapitole 2.3.5. Pravidlo $3 \cdot \sigma$ je bežne používané na identifikáciu globálnych anomálií, ktoré sú detegované po prekročení trojnásobku hodnoty štandardnej odchýlky. Pri sezónnych anomáliach tento typ detekcie zlyháva, keďže odchýlka je vypočítaná nad celým pozorovaným časovým radom. Pri jeho segmentácii je metóda úspešná iba v prípade, kedy sa odchýlka nepretržite mení [18].

Metóda kĺzavého priemeru má niekoľko modifikácií, na základe ovahania jednotlivých meraní vzniká napr. metóda kĺzavého priemeru s exponencionálnym váhovaním (angl. *exponentially weighted moving average*), skrátene EWMA. Autori v práci [18] porovnávali okrem

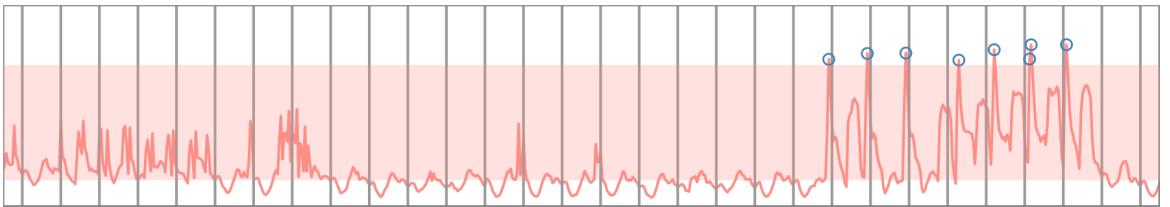
EWMA aj PEWMA (pravdepodobnostné exponencionálne váhovanie), kde v kombinácií s metódou ESD nedosiahli postačujúce výsledky. Kľavý priemer zlyhával pri identifikácii sezónnych anomalií.

2.3.5 Extrémna Studentova odchýlka

V práci budeme používať najmä metódu extrémnej Studentovej odchýlky (angl. *Extreme Studentized Deviation*) a jej ďalšie derivácie. Metóda ESD slúži na detekciu viacerých anomálnych inštancií. Jediným vstupným parametrom metódy je najväčší možný počet podozrivých inštancií v danom datasete. Generalizovaná ESD sa snaží maximalizovať odchýlku datasetu $|x_i - \tilde{x}|$ pre s inštancií. Počet inštancií, sa postupne znižuje, pokiaľ nie je dosiahnutá stanovená hranica. Pre každý odobraný počet inštancií sú overované všetky kombinácie. Tento vzťah môžeme zapísť jednoduchou rovnicou 3 definovanú pre i odobraných inštancií, ktorá je v štatistike často označovaná aj ako Grubbov test [23, 33].

$$R_i = \frac{\max_i |x_i - \tilde{x}|}{n - i} \quad (3)$$

Do rovnakej kategórie môžeme zaradiť aj sezónnu ESD (angl. *Seasonal Extreme Studentized Deviation*), ktorá rovnako využíva ESD na identifikáciu anomalií. Kľúčovým rozdielom je aplikovanie ESD až na dátu, ktoré boli dekomponované pomocou modifikovaného STL algoritmu. Vďaka tomu algoritmus deteguje globálne anomálie, ktoré sa rozpínajú mimo očakávaných sezónnych extrémov, ale aj lokálne anomálie, ktoré by inak ostali zamaskované sezónnou zložkou časových radov. Modifikácia STL algoritmu pozostáva v zamenení trendovej zložky mediánom daného časového radu. Reziduálna zložka je potom vypočítaná ako rozdiel nameranej hodnoty a súčtu sezónneho komponentu a mediánu. Zmena vzorca použitého na dekompozíciu, zabráni tvorbe falošných anomalií v reziduálnej zložke časového radu. Hlavnými obmedzením S-ESD sú datasety obsahujúce väčší podiel anomalií. Môžeme si to všimnúť na obrázku 11, kde anomálie nachádzajúce sa vo zvýraznenom regióne nie sú detegované, keďže ich množstvo ovplyvňuje ako priemer tak aj štandardnú odchýlku. Kvôli tomu algoritmus neoznačuje podozrivé pozorovania čím vzniká mnoho falošne neoznačených inštancií [18].

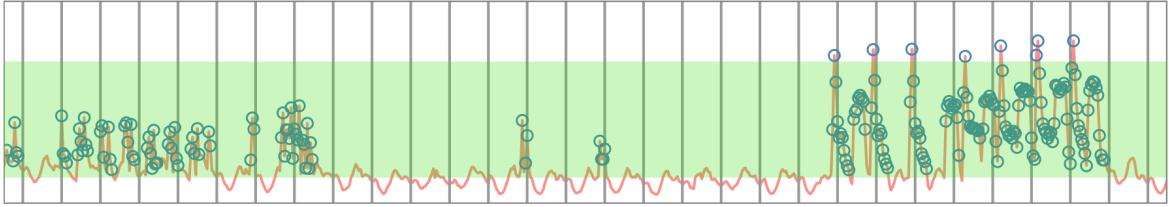


Obr. 11: Anomálie detegované pomocou S-ESD algoritmu [18].

Cieľom sezónnej hybridnej ESD (angl. *Seasonal Hybrid Extreme Studentized Deviation*) odstrániť obmedzenia, ktoré vznikajú pri S-ESD. Rovnako je použitá modifikovaná dekompozícia STL. Rozdiel je v ESD, kde namiesto priemeru a štandardnej odchýlky je použitá robustnejšia štatistická metóda, ktorá je schopná tolerovať až 50% anomalií v časovom rade. Jedná sa o absolútну odchýlku mediánu MAD (angl. *Median Absolute Deviation*), ktorú vypočítame pomocou vzorca 4 [18].

$$MAD = \text{median}_i(|X_i - \text{median}_j(X_j)|) \quad (4)$$

Cenou za to je vyššia výpočtová náročnosť metódy, keďže MAD požaduje zoradenie dát pred výpočtom ESD. Na druhej strane sú hodnoty F-skóre takmer až o 30% vyššie. V datasetoch s nízkym počtom výskytov anomálií môže byť vhodnejšie použitie metódy S-ESD [18].



Obr. 12: Anomálie detegované pomocou S-H-ESD algoritmu [18].

Dalšou metódou založenou na ESD je aj rozšírená S-H-ESD (angl. *Enhanced Seasonal Hybrid Extreme Studentized Deviation*), ktorej proces pozostáva z troch krokov, a to transformácia dát, dekompozícia časových radov a analýza reziduálnej zložky. Účelom transformácie dát je minimalizovať počet falošne identifikovaných normálnych inštancií a zároveň normalizovať vstupné časové rady pomocou Box-Coxovej transformácie, keďže parametrické štatistické testy dosahujú lepšie výsledky pri normálnom rozdelení dát. Na optimálne nastavenie parametrov normalizačnej funkcie je použitá metóda maximálnej pravdepodobnosti (angl. *Maximum Likelihood method*), ktorá je výpočtovo nenáročná a vhodná na daný problém. V procese dekompozície je použitá LOESS regresia (angl. *Locally Estimated Scatterplot Smoothing*), keďže klasická dekompozícia môže byť ovplyvnená výskytom anomálií vo vstupných dátach. Cieľom modifikovanej dekompozície je pomocou súrada vnořených cyklov a váh, robustne a iteratívne identifikovať trend a sezónnosť v danom časovom rade. Posledným procesom je samotná analýza reziduálnej zložky, kde bežná ESD metóda potrebuje k ako vstupný parameter označujúci počet anomálnych inštancií. V navrhovanej metóde je parameter vypočítaný automaticky na základe štandardnej odchýlky spracovávaného okna [45].

2.4 Metódy zhlukovania časových radov

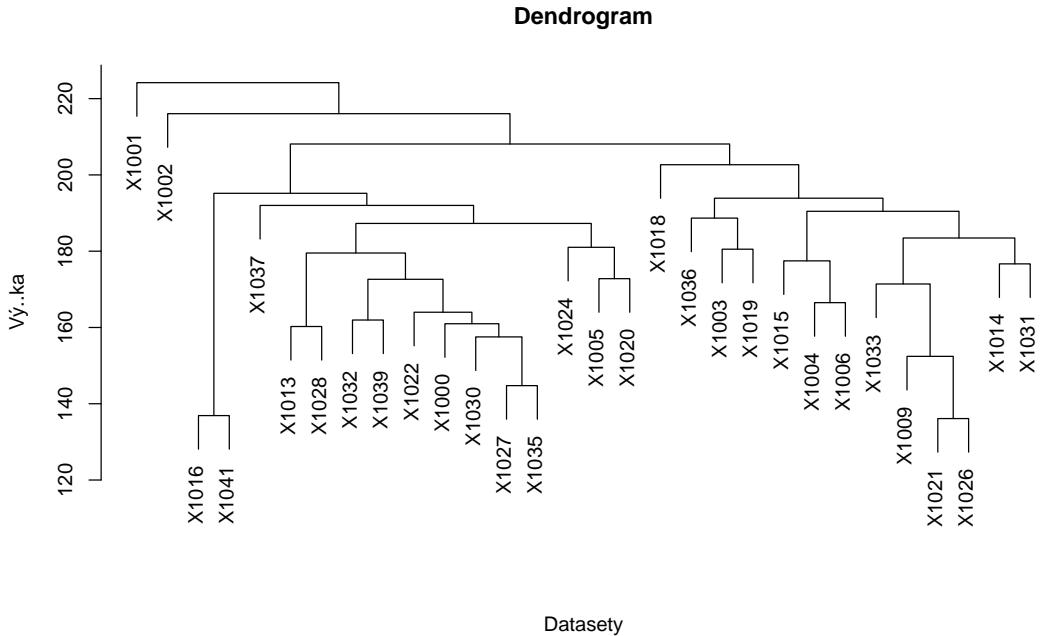
Cieľom zhlukovania je rozdeliť dátové inštancie do k zhlukov na základe spoločných črt. V prípade, že inštancie sú reprezentované nízkodimenzionálnym vektorom v Euklidovskom priestore, môžu byť na zhlukovanie použité klasické techniky spomenuté v 2.3. Ak inštancie reprezentujú časový rad, nasadenie takýchto štandardných prístupov je zriedkavé [17].

Metódy používané na zhlukovanie časových radov môžeme rozdeliť do 3 skupín, na základe reprezentácie dát, s ktorými pracujú. Prvá skupina predpokladá surové dátá, druhá pracuje s extrahovanými vlastnosťami z dát a posledná metóda pristupuje k dátam pomocou vytvoreného modelu. Prístupy sú opísané v nasledujúcich podkapitolách 2.4.1, 2.4.2, 2.4.3 a 2.4.4 [32].

2.4.1 Zhlukovanie na základe dočasnej susednosti

Metóda (angl. *Temporal-Proximity based clustering approach*) pracuje priamo so surovými, neupravenými dátami, kvôli čomu sa zvykne nazývať aj zhlukovanie na základe surových dát (angl. *Raw data based clustering approach*). Hlavným princípom je striedanie viacerých vzdialenosných alebo podobnostných metrík pre použité časové rady [32].

Hierarchické zhlukovanie produkuje vnorenú hierarchiu skupín podobných časových radov na základe vzdialenosných matíc jednotlivých inštancií. Hierarchia je graficky reprezentovaná pomocou dendrogramu, príkladom môže byť graf 13. Výhodou je, že nie je nutné zadávať počet zhlukov, ktoré ideme identifikovať. Nevýhodou je obmedzenie výpočtu iba na menšie datasety, keďže výpočtová zložitosť tejto metódy je kvadratická [13].



Obr. 13: Príklad reprezentácie vytvorených zhlukov pomocou dendrogramu.

Metóda hierarchického zhlukovania zoskupuje časové rady do stromu zhlukov. Vo všeobecnosti existujú dva typy týchto metód, agglomeratívne a deliace. Agglomeratívne metódy zo začiatku umiestňujú časové rady do samostatného zhluku, následne ich postupne spájajú do väčších zhlukov, pokiaľ neexistuje jediný zhluk, alebo nie je ukončovacou podmienkou práve k zhlukov. Deliace metódy sú pravým opakom, kedy sú jednotlivé zhluky postupne delené na menšie a umiestňované do hierarchického stromu. Na zlepšenie kvality zhlukovania pri hierarchickom zhlukovaní sú používané bežné zhlukovacie techniky [46].

Agglomeratívne zhlukovanie na základe vzdialosti medzi dvoma zhlukmi nameranej pomocou dvojice najbližších časových radov umiestnených v rôznych zhlukoch, predstavujú potenciálnych kandidátov na zlúčenie. Podobnosť môže určovať aj *Wardov algoritmus minimálnej variancie*, ktorý zlúči zhluky s najmenším nárastom variancie. V každom kroku sú tak vyskúšané všetky kombinácie dvojíc zhlukov, následne je vybrané minimum. Porovnávané časové rady nemusia mať vždy rovnakú dĺžku. Nevýhodou metódy je najmä vysoký počet operácií, ale aj neschopnosť späťne zmeniť rozhodnutie zlúčiť zhluky [46].

Deliace zhlukovanie nie je obmedzené iba na časové rady rovnakej dĺžky. Zároveň tiež nie je možné zmeniť delenie zhluku, ktoré už bolo vykonané. Na meranie vzdialenosnosti môžu byť použité metriky opísané v 2.4.5 [46].

2.4.2 Zhlukovanie na základe reprezentácie

Keďže manipulácia so súrovými dátami je často náročná a dátá navyše obsahujú nadbytočné informácie, táto metóda (angl. *Representation based clustering approach*) najskôr transformuje dátá do vektoru vlastností a až následne sú aplikované zhlukovacie algoritmy. V literatúre sa zvykne označovať aj ako zhlukovanie na základe vlastností (angl. *Feature based clustering approach*) [32].

Samoorganizované mapy predstavujú triedu neurónových sietí, kde sú neuróny usporiadane v nízkodimenzionálnej štruktúre. Trénovanie prebieha iteratívne a bez učiteľa. Proces začína pridelením náhodných hodnôt váhovým vektorom w . Každá iterácia trénovania pozostáva z 3 krokov a to náhodného výberu vstupného vektoru z trénovacej množiny, evaluácie siete a aktualizovaní váhových vektorov. Po natrénovaní je vypočítaná Euklidovská vzdialenosť medzi vstupným vzorom a váhovým vektorom. Následne je neurón s najmenšou vzdialenosťou označený ako t a ostatné váhy ostatných neurónov sú aktualizované v závislosti od vzdialenosťi od neurónu t . Nevýhodou je opäť náročné spracovanie časových radov rôznych dĺžok, keďže dĺžka časového radu definuje aj dĺžku váhového vektora w [22, 46].

2.4.3 Zhlukovanie na základe modelu

Metóda (angl. *Model based clustering approach*) predpokladá, že každý časový rad je generovaný určitým modelom alebo pravdepodobnosťou distribúciou. Časové rady sú považované za podobné ak aj modely charakterizujúce jednotlivé časové rady sú si podobné [32].

ARIMA model navrhnutý v práci [48] zhlukuje jednorozmerné časové rady. Predpokladali, že časové rady sú vygenerované k rôznymi ARIMA modelmi. Vylepšili algoritmus na maximalizáciu očakávaní (angl. *expectation maximization algorithm*) tak, že sa naučil správne určiť koeficienty a parametre jednotlivých modelov zvyšovaním počtu modelov až do momentu, kedy vznikol redundantný model. Algoritmus skonvergoval v prípade, že počet modelov neboli väčší ako aktuálny počet zhlukov. Na záver boli odstránené podobné modely, čím sa ešte zmenšil výsledný počet zhlukov k .

2.4.4 Ďalšie prístupy k zhlukovaniu

Ďalší prístup je založený na oknách fixnej veľkosti (angl. *Windows based clustering approach*). V diskretizovaných časových radoch sú následne identifikované anomálne úseky. Nevýhodou metódy je náročnosť voľby správnej veľkosti okna tak, aby zachytila anomáliu a jej výpočtová zložitosť [42].

Pristup založený na skrytých Markovovych modeloch (angl. *Hidden Markov models based approach*) je reprezentovaný výkonným konečným stavovým strojom. Vychádza z predpokladu, že existuje skrytý proces, ktorý je Markovský a zároveň generuje normálne časové rady. Nevýhodou je, že technika zlyháva v prípade, že takýto proces neexistuje. Na základe vytvoreného Markovovho modelu sú merania, skupina meraní alebo celý časový rad označené za anomálie [42].

2.4.5 Metriky vzdialenosťi

Kľúčovou záležitosťou pri zhlukovaní časových radov na základe ich podobnosti, je meranie vzdialenosťi medzi nimi. Rovnako ako pri zhlukovaní bodových inštancií je potrebné

definovať si metódy merania vzdialenosť. Najčastejšími metrikami sú Euklidovská a Manhattanská vzdialenosť. Vhodnosť aplikovania týchto klasických metód je nízka, keďže nameraná vzdialenosť zachytáva aj použitú škálu v dátach. Pri porovnávaní časových radov nás spravidla zaujíma zmena krivky časového radu a rovnaká dĺžka porovnávaných časových radov [13, 46].

Metódy používané na meranie vzdialenosť medzi časovými radmi môžeme rozdeliť do 3 skupín založených na atribútoch, na modeloch a na tvare krivky. Pri atribútových metódach je pre každý časový rad vypočítaný atribútový vektor, na základe ktorého je vypočítaná napr. Euklidovská vzdialenosť medzi jednotlivými inštanciami. Modelové techniky používajú parametrický model, do ktorého vstupujú časové rady. Vzdialenosť je potom definovaná ako vzdialenosť medzi jednotlivými modelmi. Metódy porovnávajúce tvary kriviek sa snažia prispôsobiť výsledný tvar časového radu nelineárnym rozťahovaním a kontrakciou časových osí [17].

Korelačný koeficient $r(X, Y)$ meria stupeň lineárnej závislosti medzi dvoma časovými radmi X a Y . Vyjadríme ho vzorcom 5.

$$r(X, Y) = \frac{E[(X - E[X]) \cdot (Y - E[Y])]}}{E[(X - E[X])^2] \cdot E[(Y - E[Y])^2]} \quad (5)$$

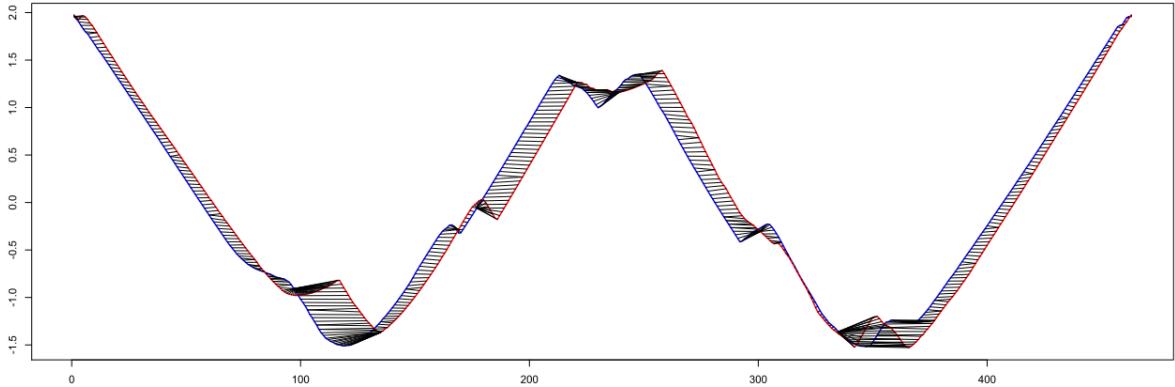
Korelacia blízka -1 znamená, že nárast kriviek časových radov je zrkadlový. Pri korelácii rovnej 0 hovoríme o rozdielnych časových radoch a pri hodnote 1 o podobných. Na základe hodnoty korelácie, potom môžeme vyjadriť vzdialenosť vzorcom 6. Nevýhodou je, ak máme k dispozícii iba malú, prípadne krátku časť datasetu. V takom prípade sa podobnosť touto metrikou určuje len fažko. Keďže korelacia zachytáva iba lineárnu podobnosť časových radov, pri aplikovaní metriky na dva nelineárne podobné časové rady, sú vyhodnotené ako vzdialené [13].

$$D_r(X, Y) = \sqrt{\frac{1}{2} \cdot (1 - r(X, Y))} \quad (6)$$

Dynamické deformovanie času predstavuje metódu (angl. *Dynamic Time Warping*), ktorá dokáže zachytiť nelineárne skreslenie medzi časovými radmi vďaka prideleniu viacerých hodnôt časového radu X druhému časovému radu Y . Takto metóda viac zodpovedá ľudskej intuícii. Na obrázku 14 si môžeme všimnúť, že sú porovnávané hodnoty, ktoré by sme intuitívne zvolili pri zarovnaní časových radov podľa tvaru krivky. D_{DTW} je vypočítané pomocou dynamického programovania, práve kvôli množstvu existujúcich kombinácií. Rekurzia je vyjadrená vzorcom 7 [13, 14, 19].

$$D_{DTW}(i, j) = \begin{cases} d(x_i, y_j) + \min \begin{cases} D_{DTW}(i-1, j) \\ D_{DTW}(i, j-1) \text{ ak } i \neq 0 \text{ a } j \neq 0 \\ D_{DTW}(i-1, j-1) \end{cases} & \\ 0 \text{ ak } i = 0 \text{ a } j = 0 & \\ \infty \text{ inak} & \end{cases} \quad (7)$$

Do rovnakej rodiny vzdialenosťných metrík patrí aj rýchle globálne zarovnanie kernelov (angl. *Fast global alignment kernels*) skrátene GAK. Cieľom metódy je znížiť veľkú časovú náročnosť DTW s dosiahnutím porovnateľných výsledkov. Podobne aj metrika založená na



Obr. 14: Príklad porovnávania časových radoch pomocou dynamickej deformácií času (časové rady sú vyznačené modrou a červenou) [26].

tvare časových radoch (angl. *Shape-based distance*) skrátene SBD, znižuje časovú náročnosť výpočtu vzdialenosť medzi časovými radmi. Naroďal od GAK nepoužíva kernely, ale štatistické metódy založené na krízovej korelácií (angl. *cross-correlation*). Bližšie sa týmito metrikami zaobrali autori v prácach [11] a [30].

Kvalitatívna vzdialenosť je metóda založená na kvalitatívnom porovnávaní tvaru dvoch časových radoch. Pre časové rady X a Y vyberieme dvojicu bodov i a j , ktoré označujú zmenu premennej v danom časovom rade. Tak vznikajú 3 možnosti, hodnoty v časovom rade rastú ($X_i < X_j$), nemenia sa ($X_i \approx X_j$) alebo klesajú ($X_i > X_j$). Vzdialenosť potom vyjadrimo vzorcom 8, pomocou ktorého spočítame počet zhôd v raste časových radoch. Práve funkcia $Diff(q_1, q_2)$ vyjadruje rozdiel v zmene rastu. Metóda nemá nevýhody, ktoré vznikali pri korelácií, na druhú stranu je aplikovateľná iba na krátke časové rady bez toho, aby sa dramaticky znížila kvalita odhadu vzdialenosť. Podobnosť tvarov kriviek je detegovaná aj v prípade, kedy neexistuje medzi časovými radmi lineárna alebo nelineárna závislosť [13].

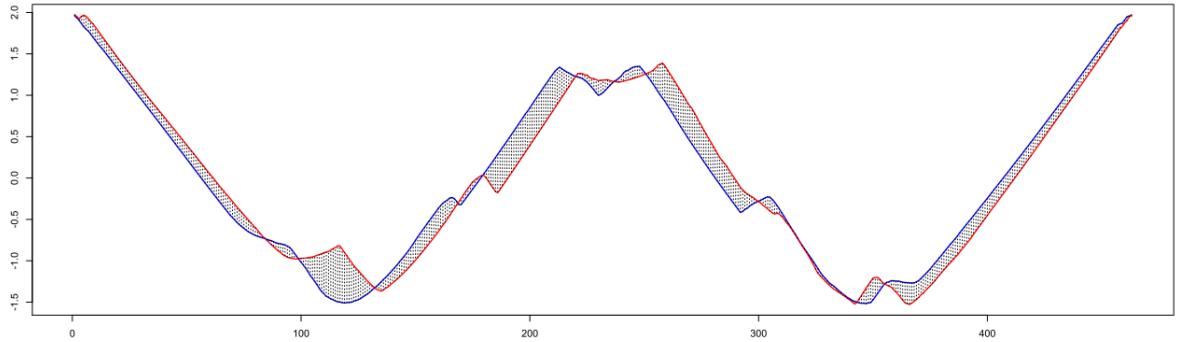
$$D_q(X, Y) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2 \cdot Diff(q(X_i, X_j), q(Y_i, Y_j))}{N \cdot (N - 1)} \quad (8)$$

Euklidovská vzdialenosť je používaná najmä pri klasických zhlukovacích problémoch. Ak zvolený časový rad má dĺžku n , vzdialenosť vypočítame vzorcom 9. Na obrázku 15 sú vždy porovnávané hodnoty vyskytujúce sa v rovnakom čase t [46].

$$D_E(X, Y) = \sqrt{\sum_{k=1}^n (X_{ik} - Y_{jk})^2} \quad (9)$$

Manhattanská vzdialenosť je rovnako ako Euklidovská vzdialenosť používaná najmä pri klasických zhlukovacích problémoch. Výpočet je tiež veľmi podobný, môžeme ho vyjadriť vzorcom 10 [10].

$$D_M(X, Y) = \sum_{k=1}^n |X_{ik} - Y_{jk}| \quad (10)$$



Obr. 15: Príklad porovnávania časových radov pomocou Euklidovskej vzdialenosť (časové rady sú vyznačené modrou a červenou) [26].

Pearsonov korelačný koeficient je používaný pri výpočte vzdialenosť, ktorá je založená na vzájomnej korelácii. Vo vzorci 11 reprezentuje \tilde{X} aritmetický priemer časového radu X . Vzdialenosť vyjadríme vzorcom 11 [46].

$$r(X, Y) = \frac{\sum_{i=1}^n (X_i - \tilde{X}) \cdot (Y_i - \tilde{Y})}{\sqrt{\sum_{i=1}^n (X_i - \tilde{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \tilde{Y})^2}} \quad (11)$$

$$D_P(X, Y) = 2 \cdot (1 - r(X, Y)) \quad (12)$$

Vzdialenosť medzi krátkymi časovými radmi je metóda (angl. *Short time series*), ktorá meria vzdialenosť ako sumu štvorcových rozdielov medzi krivkami jednotlivých časových radov. Na odstránenie nežiaducich efektov škály sa používa z standardizácia. Matematicky vzdialenosť vyjadríme vzorcom 13. Zložka t_k predstavuje čas [46].

$$D_{STS}(X, Y) = \sqrt{\sum_{k=1}^n \left(\frac{Y_{j(k+1)} - Y_{jk}}{t_{(k+1)} - t_k} - \frac{X_{i(k+1)} - X_{ik}}{t_{(k+1)} - t_k} \right)^2} \quad (13)$$

2.5 Predspracovanie dát

Pri metódach založených na dátovej analytike a strojovom určení je nesmierne dôležité zvoliť vhodnú reprezentáciu dát, vybrať atribúty, ktoré sú relevantné pre zvolený problém a často krát aj odstrániť chýbajúce alebo nekompletné časové rady. Znalosť vstupných dát a špecifickosť danej domény prináša k predspracovaniu dát ďalšie prístupy, ktoré zvyšujú správnosť použitých úprav.

Najčastejšie používanými vysvetľujúcimi premennými sú:

- geografická poloha
- voltáž distribučnej siete
- tarifná skupina
- energetická sebestačnosť
- pravidelnosť platieb

- priemerná spotreba
- používané elektrospotrebiče
- veľkosť a typ objektu

Ďalšou premennou, ktorá vysvetľuje krátkodobé zmeny v správaní jednotlivých odberateľov je počasie. To je pre viacerých odberateľov rovnaké a viaže sa na konkrétny región, v ktorom sa nachádza meteorologická stanica. Dáta z nich sú väčšinou verejne dostupné [40].

2.5.1 Filtrovanie odberateľov

Dáta z inteligentných meračov bývajú často nekompletné a s chýbajúcimi hodnotami. Väčšina algoritmov nedokáže spracovať takéto dátá a všetky časové rady musia byť rovnakej dĺžky. Rovnako sú nepoužiteľné dátá, ktoré boli poškodené pri samotnom zbere dát, nie však pri meraní. Zatiaľ čo chybné meracie zariadenia môžu spadať do detekcie anomalií a zaujímajú nás, dátá ktoré boli zduplicované alebo inak poškodené až pri ukladaní môžeme vylúčiť z datasetu. Prípady, kedy zákazník bol zapojený do siete až v priebehu meraní, musíme ošetrovať špeciálne, najčastejšie vynechaním alebo orezaním na najbližšiu menšiu dĺžku posuvného okna [28].

2.5.2 Výber atribútov

Väčšina dát pochádzajúcich z inteligentných meračov obsahuje iba stĺpce s časovou známkou a momentálnou spotrebou daného uzlu v sieti. Z týchto informácií ešte vieme určiť, mesiac, týždeň, deň prípadne deň v týždni alebo sviatok. Niektoré z extrahovaných atribútov úzko súvisia s funkciou spotreby elektrickej energie. Pri vytváraní presného modelu je preto nevyhnutné správne identifikovať takéto atribúty. Otestovanie všetkých kombinácií by bolo časovo a výpočtovo náročné. Najjednoduchším spôsobom je vytvorenie korelačnej matice jednotlivých vysvetľujúcich premenných a sledovanej veličiny [7].

2.5.3 Extrakcia črt

Ďalšou technikou používanou pri príprave dát je tvorba nových atribútov založených na pôvodných, surových dátach. V súvisiacom článku [28] ide napr. o vytvorenie hodinového prieberu pre každého zákazníka. Vzťah priemernej spotreby x_h môžeme definovať rôzne, v našom prípade ide o podiel mesačnej priemernej spotreby nasledujúceho mesiaca P_{h+1} a rozdielu dennej spotreby v aktuálnom a nasledujúcom mesiaci $D_{h+1} - D_h$, čo zapíšeme vzorcom 14.

$$x_h = \frac{P_{h+1}}{D_{h+1} - D_h} \quad (14)$$

2.5.4 Reprezentácia FeaClip

Ako bolo už spomenuté, niektoré datasety obsahujú informácie iba o meranej veličine, čo niekedy nemusí byť postačujúce. Preto vznikajú nové atribúty popisujúce sledovanú veličinu. Jednou z nich je metóda FeaClip, ktorá na základe reprezentácie dát ako bitového reťazca, vytvára nové atribúty. Nad vybraným posuvným oknom nad datasetom je aplikovaná transformácia popísaná rovnicou 15, čím sú merania s hodnotami väčšími ako priemer aktuálneho posuvného okna nahradené hodnotou 1, inak 0. Na vzniknutý reťazec je aplikované kódovanie dĺžky behu (angl. *Run-length encoding*). Beh je súvislá postupnosť jedného znaku, dĺžka

behu predstavuje počet znakov v takomto behu. Analyzované okno časového radu je transformované na osmicu čísel, a to maximum z dĺžok jednotkových behov, maximum z dĺžok nulových behov, počet jednotiek v refazci, počet prechodov medzi rôznymi behmi a počty prvých a posledných núl a jednotiek. Výhodami reprezentácie sú najmä redukcia dimenzií, zvýraznenie charakteru dát, paraleлизmus a jednoduchá interpretácia dát [25].

$$\hat{x}_i = \begin{cases} 1 & \text{ak } x_i > \mu \\ 0 & \text{inak} \end{cases}, \text{ pre } i \in (1, 2, \dots, n) \quad (15)$$

2.5.5 Agregácia dát

Dáta z meračov sú dostupné v pravidelných intervaloch. Pre jednoduchšiu manipuláciu s časovými radmi a redukcii dimenzií, môžu byť dátá agregované do väčších intervalov. Pri použití viacerých datasetov s rôznou frekvenciou zberu, je agregácia hustejšieho časového radu nutná, keďže by tak vzniklo množstvo chýbajúcich hodnôt. Agregácia dát tiež vyhľadzuje malé odchýlky v časových radoch, čo môže sťažiť identifikáciu náhlej zmeny správania odberateľov. To môže viesť k nesprávnemu označeniu správania odberateľa za neštandardné [7].

Cieľom agregácie časových radov môže byť aj redukcia na priemer, prípadne medián, dňa alebo týždňa. So zredukovanými dátami je potom možné pracovať rýchlo a efektívne, keďže ich pamäťová náročnosť je iba zlomkom oproti pôvodnej. Zároveň však vzniká priestor na stratenie informácie o anomálnej aktivite odberateľa, čo je nutné zvážiť pri konkrétnej implementácii.

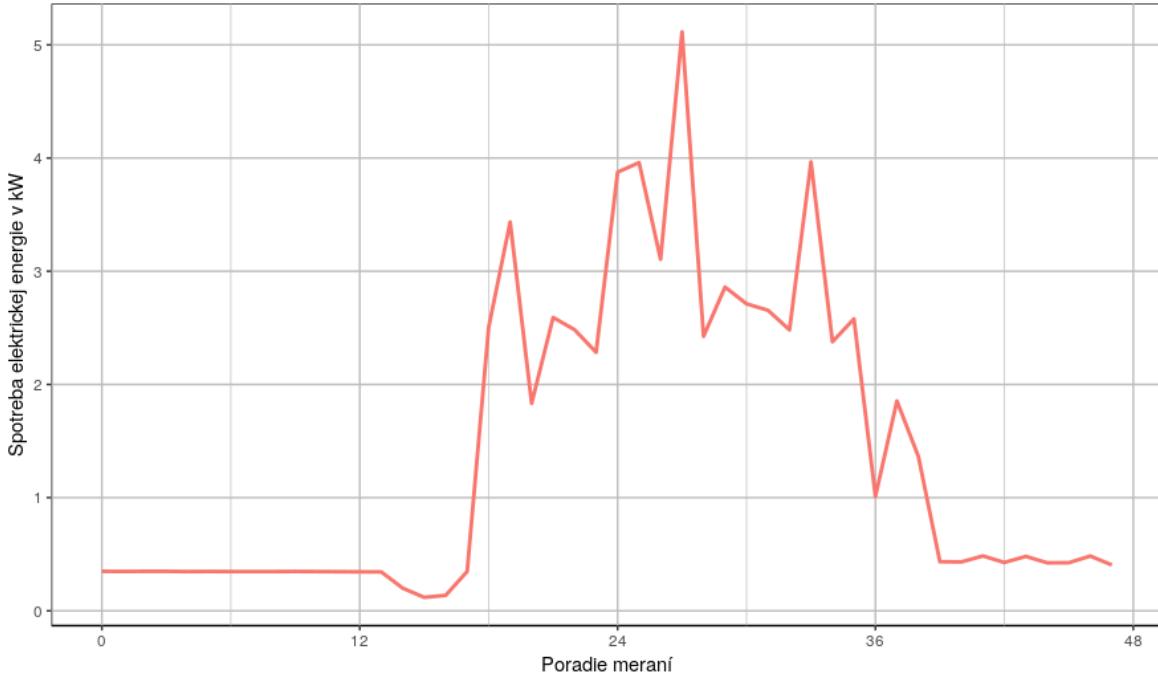
2.5.6 Redukcia dimenzií

Jednou z najjednoduchších metód používaných pri redukcii dát je práve vzorkovanie (angl. *sampling*). Parametrami sú m a n , ktoré predstavujú počet dimenzií pred a po procese vzorkovania. Vzdialenosť sa medzi jednotlivými inštanciami zväčšuje, no zároveň je rovnaká medzi všetkými inštanciami. Nevýhodou je, že tvar výsledného časového radu je oproti pôvodnému skreslený, čo môžeme vidieť na obrázkoch 16 a 17 [14].

Lepšie výsledky dostaneme ak pri vzorkovaní budeme priemerovať hodnoty vo vzniknutých intervaloch. Táto metóda sa zvykne nazývať aj po častiach agregovaná approximácia (angl. *piecewise aggregate approximation*), skrátene PAA. Vylepšenou verziou je metóda APC, kde vzniknuté intervale majú rôznu dĺžku, v závislosti od tvaru časového radu. Tiež môžeme okrem priemera použiť medián zvoleného intervalu. Obe metódy môžeme vidieť na obrázku 17 [5].

Ďalšou metódou je approximácia pomocou rovných čiar, kde hlavnými kategóriami sú lineárna interpolácia a lineárna regresia. Bežnou metódou pri interpolácii je použiť po častiach lineárnu approximáciu (angl. *piecewise linear approximation*). Algoritmus začína vytvorením odhadu časového radu, ktorý používa polovicu vytvorených intervalov. Tie sú následne zlučované, pokial nie je splnené ukončovacie kritérium, napr. celkový počet intervalov. Poradie zlučovania je určené na základe ceny zlučovania [14].

Žiaducim efektom pri redukovaní dimenzií je zachovanie charakteristických bodov. Tieto body sa zvyknú nazývať percepčne dôležité body (angl. *perceptually important points*), skrátene PIP. Algoritmus najskôr určí prvé tri body, a to prvý, posledný a bod, ktorý je od týchto dvoch najvzdialenejší. Ďalšie body sú určované na základe maximálnej vertikálnej vzdialenosť medzi dvoma susednými bodmi PIP. Proces pokračuje pokial nie sú zoradené podľa dôležitosti všetky pôvodné body. Na obrázkoch 16 a 17 môžeme vidieť, že tvar kriviek



Obr. 16: Príklad časového radu bez redukcie dimenzií.

pôvodného časového radu a redukovaného je mierne odlišný, čo je spôsobené roztahnutím alebo zúžením podintervalov v redukovanom časovom rade [14].

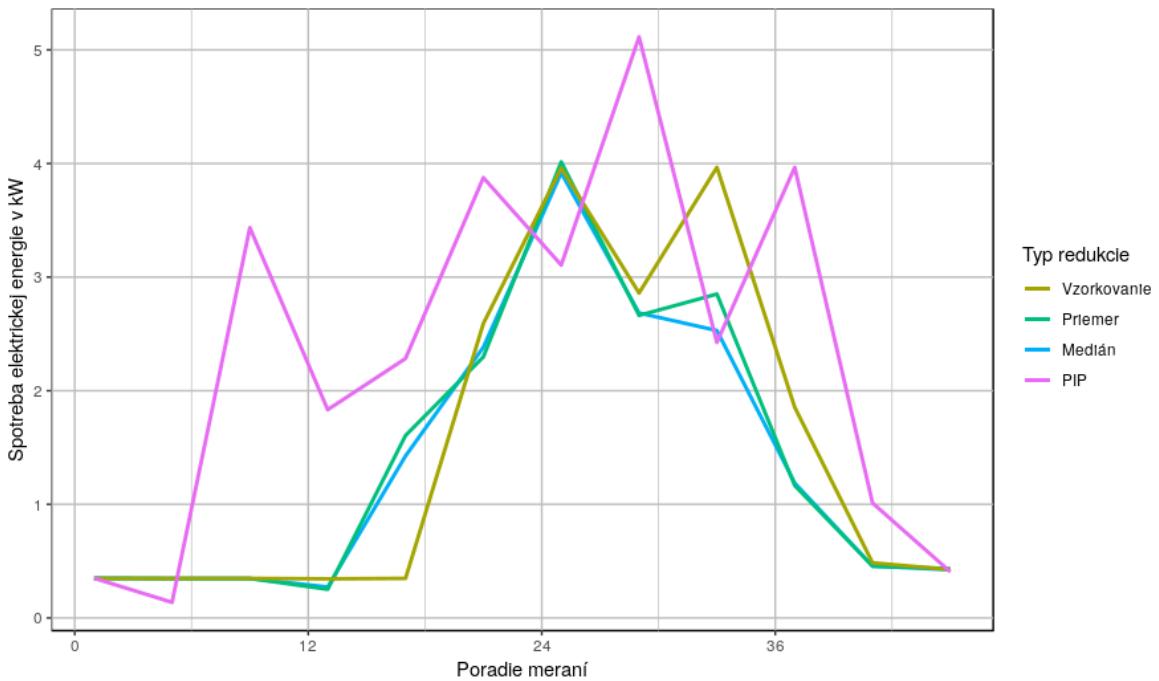
Ďalší prístup používaný pri reprezentovaní časových radov je ich konvertovanie z PAA do symbolickej formy. Najskôr sú diskretizované do intervalov, ktoré sú následne konvertované do symbolov. Táto metóda sa nazýva symbolická agregovaná aproximácia (angl. *symbolic aggregate approximation*), skrátene SAX. Algoritmus rozdelí obor hodnôt na regióny a každý z nich je namapovaný na iný symbol [14].

Ďalšou metódou je analýza hlavných komponentov (angl. *principal component analysis*), skrátene PCA. Obvykle sa PCA používa na elimináciu menej významných komponentov, čím sa znižuje dimenzionalita dát. Metóda má uplatnenie aj pri analýze či vizualizácii vysokodimenzionálnych dát. Najskôr sú vypočítané priemery pre jednotlivé dimenzie dát, z nich variancie a kovariančná matica. Na základe kovariančnej matice sú vypočítané vlastné hodnoty a vektory (angl. *eigenvalues* a *eigenvectors*), ktoré definujú rovinu, na ktorú sú pôvodné dátá premietané. Zobrazenie pri tom dosahuje najnižšiu chybu rekonštrukcie a zároveň najnižšiu vzdialenosť meraní od vzniknutej roviny. Ako príklad môže poslúžiť obrázok 18, na ktorom je vizualizovaná redukcia dvojdimenziorných dát [14, 37].

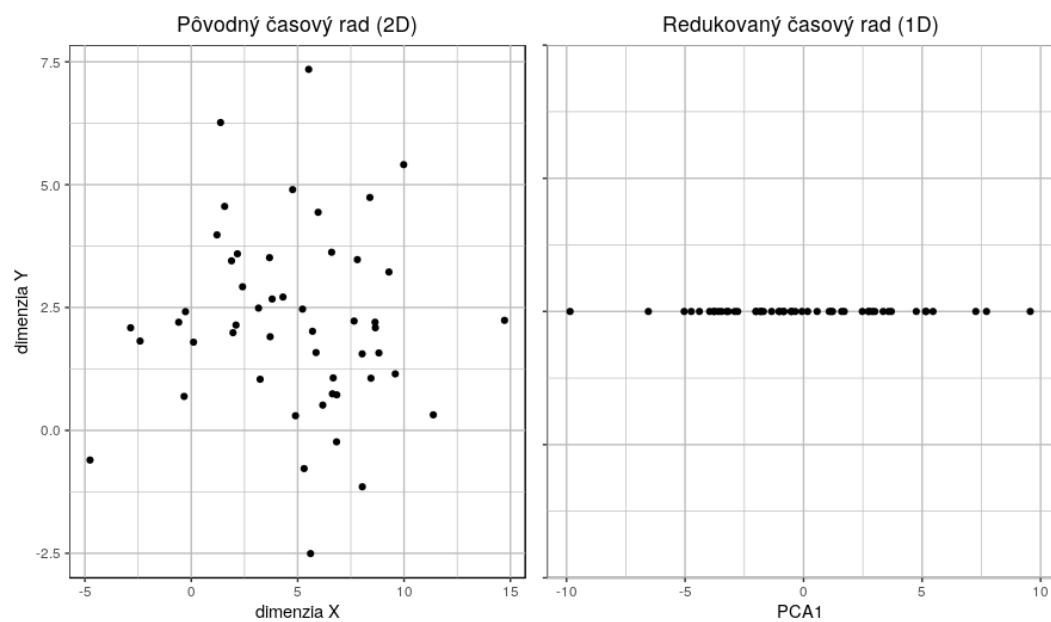
Na podobnom princípe ako DTW je založené aj hľadanie najdlhšej spoločnej podpostupnosti (angl. *longest common subsequence*), skrátene LCSS. Ide o variáciu editačnej vzdialnosti a spájania dvoch sekvencií, ktoré sa môžu natiahnuť a vynechať tak niektoré elementy bez toho, aby sa menilo ich poradie v rámci postupnosti. Narozenie od DTW, výstupy nie sú skreslené anomáliami v dátach [14].

2.5.7 Segmentácia časových radov

Časové rady sú charakteristické súvislým priebehom, a preto pri ich segmentácii je nutné celič viacerým problémom. Najjednoduchším prístupom je rozdeliť časový rad pomocou okna fixnej dĺžky do segmentov, z ktorých vznikajú jednoduché vzory. Jedinou úlohou je



Obr. 17: Príklad redukovaného časového radu.



Obr. 18: Príklad redukcie dimenzií pomocou PCA.

správne zvoliť dĺžku okna. Pri použití tejto metódy existujú dva hlavné problémy. Typické vzory môžu mať variabilnú dĺžku a ich výskyt môže byť rôzny. Práve preto je vhodnejšie použiť dynamický prístup, ktorý rozdeľuje časový rad práve v bodoch, ktoré zachovávajú cyklicky vyskytujúce sa vzory a vznikajú tak segmenty s rôznymi dĺžkami [14].

2.5.8 Normalizácia číselných vektorov

Rozsahy nameraných hodnôt inteligentnými meračmi sa môžu lísiť, pri jednotlivých odberateľoch dokonca aj rádovo. Pri zhľukovaní takýchto časových radov je preto potrebná najskôr ich normalizácia, v prípade zhľukovania na základe tvaru priebehov. Existuje viacero druhov normalizácií, no v práci budeme používať najmä štandardné skóre, nazývané aj z-skóre (angl. *z-score*). Hodnotu vypočítame ako podiel rozdielu hodnoty a priemeru a štandardnej odchýlky. Normalizáciu vyjadrimo nasledujúcim vzorcom 16 [2]

$$z = \frac{x - \mu}{\sigma} \quad (16)$$

2.5.9 Vyhladzovanie časových radov

Spracovávané dátá často obsahujú lokálne extrémy alebo anomálne meranie, ktorých vplyv je nutné pred ďalším spracovaním eliminovať. Zatiaľ čo niektoré metódy sú robustné a výsledky spracovania nie sú skreslené extrémnymi hodnotami, iné metódy sa vyznačujú extrémnou citlivosťou a je potrebné predspracovať dátá pomocou vyhladzovania. Jednou takou je aj metóda lokálne odhadovaného vyhladzovania rozptylu (angl. *locally estimated scatterplot smoothing*), skrátene LOESS. Predpokladajme, že veličina y je definovaná pomocou regresnej funkcie $g(x)$, ktorá predikuje danú veličinu s náhodnou chybou ϵ . Vzťah môžeme zapísť rovnicou 17. Ide však o neparametrickú štatistickú metódu, keďže pracujú s početnosťami súboru alebo s ich poradím určenom vo vstupných dátach. Princíp metódy spočíva v lokálnej aproximácii prediktora x_i pomocou susediacich meraní na základe veľkosti chyby voči skutočným meraniam. So zmenou veľkosti rádiusu prediktora x_i sa mení aj miera vyhladzovania. Váha jednotlivých meraní sa zmenšuje v závislosti od vzdialnosti od prediktora [8].

$$y_i = g(x_i) + \epsilon_i \quad (17)$$

2.6 Anomálie v energetických časových radoch

V distribučných sieťach vznikajú straty, ktoré vo všeobecnosti môžeme rozdeliť na technické a netechnické. Technické straty sú spôsobené vlastnosťami obvodu ako napr. odporom materiálu či únikmi cez poškodenú izoláciu a môžu sa meniť pri rôznych teplotách či počasí. Medzi netechnické straty patria najmä nelegálne odbery. V práci sa budeme zaoberať ich identifikáciou na základe anomálneho správania spotrebiteľa. Keďže je časovo a finančne náročné pravidelne kontrolovať odberateľov tak, aby sa predišlo nelegálnemu odberu, je potrebné znížiť počet podozrivých odberateľov na minimum a zároveň maximalizovať pravdepodobnosť, s ktorou budú kontrolovaní iba odberatelia s neštandardnými odbermi [9, 34].

Pri identifikácii anomalií je spravidla najskôr definovaná oblasť, ktorej inštancie považujeme za normálne. Za anomálie považujeme inštancie nachádzajúce sa mimo oblasti, alebo na jej okraji. V prípade, že na trénovanie modelu máme k dispozícii označené iba anomálne dátá, je najskôr definovaná oblasť anomálnych dát a až následne normálna oblasť. Pri identifikácii anomalií v časových radoch v doméne energetiky je takýto prístup len ľahko aplikovateľný

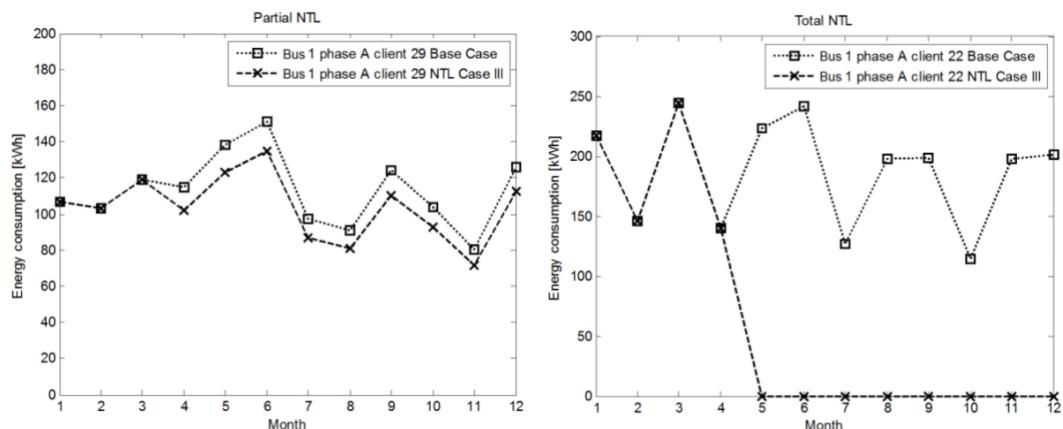
nakoľko podobné časové rady pri rôznych domácnostiach môžu, ale nemusia predstavovať normálne správanie [39].

Najčastejšími metódami používanými pri nelegálnom odbere je obídenie meračov spotreby energie či samotná manipulácia s nimi. Merače tak poskytujú nesprávne informácie o spotrebenej energie odberateľmi, čo je možné detegovať až po identifikácii celkových netechnických strát v sieti. Ďalšou populárnu metódou používanou na detekciu nelegálnych odberov je analýza spotrebiteľského profilu zákazníka, kedy je našou snahou identifikovať nepravidelné vzory v nameraných spotrebiteľských dátach [34]. Tak ako je spomenuté v práci [12], nelegálne odbery môžu prebiehať iba v určitom čase prípadne iba pri zvýšenej spotrebe. Identifikácia takýchto nelegálnych odberov je náročná a prípadná kontrola nemusí odhaliť manipuláciu s meracím zariadením.

Vďaka inteligentným meračom je možné detegovať nelegálne odbery omnoho rýchlejšie, najmä kvôli vysokej frekvencii zberania údajov. Taktôto sú identifikované aj také odbery, ktoré by sa pri klasických meraniach stratili v týždenných alebo mesačných agregáciách. Úspešnosť detektie nelegálnych odberov je výrazne vyššia najmä pri neštandardných spotrebách alebo ak sa jedná o neopakujúcu udalosť. Problém vzniká ak odberateľ systematicky mení nelegálnu spotrebu a kopíruje vzory, ktoré vznikajú v dátach pri legálnom odbere. Vtedy je potrebné mať k dispozícii väčšie množstvo dát a zároveň použiť zložitejšie algoritmy detektie anomálií, ktoré sú popísané v súvisiacej práci [29].

V súvisiacich prácach sa autori zaoberali určením netechnických strát v elektrických distribučných sieťach s použitím rôznych štatistických metód alebo strojového učenia. Dostupné dátá od distribútorov pochádzali najmä z jedného zdroja, lokality a zameriaval sa na jeden zdroj energie. Dáta, ktoré budeme mať k dispozícii disponujú podobnými vlastnosťami. V súvisiacej práci [9] boli použité viaceré zdroje dát a energie, následkom čoho bola zvýšená presnosť identifikácie anomálneho správania odberateľa. Ďalším zdrojom dát môžu byť agregované hodnoty meraní z klasických meračov, prípadne spätná väzba zo samotných kontrol odberateľov.

Typickou črtou netechnických strát je negatívny skok v spotrebe elektrickej energie. Nasleduje po poškodení inteligentného meracieho zariadenia alebo pri začatí nelegálneho odberu. Pokles môže byť zapríčinený aj zmenou počtu ľudí, miestnosti prípadne ich funkcie alebo zvýšením energetickej sebestačnosti. Následkom je nižšia nameraná spotreba energie v dlhšom horizonte. Zniženie spotreby môže byť čiastočné alebo úplné, ako môžeme vidieť na obrázkoch 19 a 20 [39, 43].



Obr. 19: Čiastočné zníženie spotreby elektrickej energie [43].

Obr. 20: Úplné zníženie spotreby elektrickej energie [43].

Z pohľadu výskytu anomálie môžu nastať nasledovné scenáre:

- Anomália vznikne neodborným pripojením odberateľa do energetickej siete alebo existuje ešte pred tým ako, nastane zber dát inteligentnými meračmi. Keďže celý časový rad pozostáva z chybných dát, odhalenie anomálie je nepravdepodobné.
- Anomália vznikne v priebehu sledovaného intervalu a zároveň je odhalená a ďalej sa už nevyskytuje.
- Anomália vznikne v priebehu sledovaného intervalu a nie je odhalená. Táto skupina je predmetom celej našej práce.

Prvý prípad anomálií je možné odhaliť iba na základe vysvetľujúcich premenných, ktoré nemusia byť pravdivé, ak sú dodané samotným odberateľom. Druhú skupinu je potrebné v dátach označiť, prípadne anomálne merania vynechať pri ďalšom klasifikovaní [39].

2.7 Vyhodnocovacie metriky

Za predpokladu, že získané dátá budú obsahovať aj označené inštancie, prípadne budú označené dodatočne na základe výpočtov, môžeme na vyhodnotenie úspešnosti použiť aj maticu zámen. V takom prípade budeme musieť predpovedať triedu jednotlivých inštancií, a teda či sa jedná o normálneho alebo anomálneho odberateľa. Jednoduchý klasifikátor označí prvých n odberateľov, ktorých miera pravdepodobnosti výskytu anomálneho odberu je najvyššia, za anomálnych. Pri vyjadrení matice zámen pomocou tabuľky 1 potom riadky predstavujú predpovedanú triedu a stĺpce skutočnú. Vznikajú tak 4 kategórie, správne označení podozriví odberatelia (angl. *true positive*), nesprávne označení podozriví odberatelia (angl. *false positive*), nesprávne označení normálneho odberatelia (angl. *true negative*) a správne označení normálneho odberatelia (angl. *false negative*). Kvalitu klasifikácie potom môžeme zmerať pomocou presnosti a pokrytie. Presnosť vypočítame vzorcom 18, kedy ide o pomer správne označených anomálií a celkový počet označených anomálií. Tým vypočítame percento odberateľov, ktorých sme správne klasifikovali ako podozrivých.

$$\text{Presnosť} = \frac{TP}{TP + FP} \quad (18)$$

Pokrytie označuje pomer správne označených anomálií a celkový počet skutočných anomálií. Vyjadríme ju pomocou vzorca 19.

$$\text{Pokrytie} = \frac{TP}{TP + FN} \quad (19)$$

Aby sa predišlo situácií, kedy sa v dátach nachádza iba malý počet anomálnych odberateľov a pre model by tak bolo výhodnejšie označovať iba tých, s ktorými si je takmer istý, je dôležité brať do úvahy aj túto metriku. Obe metriky sú vyjadrené v percentách [43, 47].

Tabuľka 1: Matica zámen (angl. *Confusion matrix*)

		skutočnosť	
		anomálna kategória	normálna kategória
predikcia	anomálna kategória	TP (true positive)	FP (false positive)
	normálna kategória	FN (false negative)	TN (true negative)

Ďalšou používanou metrikou je aj tzv. F-skóre, ktoré obsahuje informácie oboch predchádzajúcich metrík. Keďže ide o súčet metrík, tiež je vyjadrené v percentách. Cieľom práce je maximalizovať túto metriku. F-skóre vyjadríme pomocou vzorca 20, kde P predstavuje presnosť a C predstavuje pokrytie [43].

$$F = 2 \cdot (P^{-1} + C^{-1})^{-1} \quad (20)$$

2.7.1 Zhlukovacie validačné indexy

Zhlukovanie je metóda, ktorej cieľom je určiť skupinu, do ktorej spadá daná inštancia. Triedenie prebieha na základe atribútov inštancie. Keďže sa jedná o učenie bez učiteľa, je potrebná validácia výsledného zhlukovania. V praxi sa používajú validačné indexy zhlukov (angl. *cluster validity indeces*). Indexy sa delia na externé a interné, v závislosti od dostupnosti skutočných tried zhlukovaného datasetu [3].

Externé indexy zhlukov obsahujú napr. Randov, Jaccardov alebo Fowlkes-Mallowsov index. Naivným prístupom je porovnávanie zhlukov a počítanie dvojíc inštancií, ktoré sa nachádzajú v rovnakom zhluku. Maticu zámen tak môžeme prepísať do tabuľky 2. Časové rady nachádzajúce sa v rovnakom zhluku pri rôznych zhlukovaniach X a Y sa nachádzajú v kategórií *true positive* [4].

Tabuľka 2: Validačná matica zhlukovania časových radov

	Rovnaké v množine Y	Rôzne v množine Y
Rovnaké v množine X	TP (true positive)	FP (false positive)
Rôzne v množine X	FN (false negative)	TN (true negative)

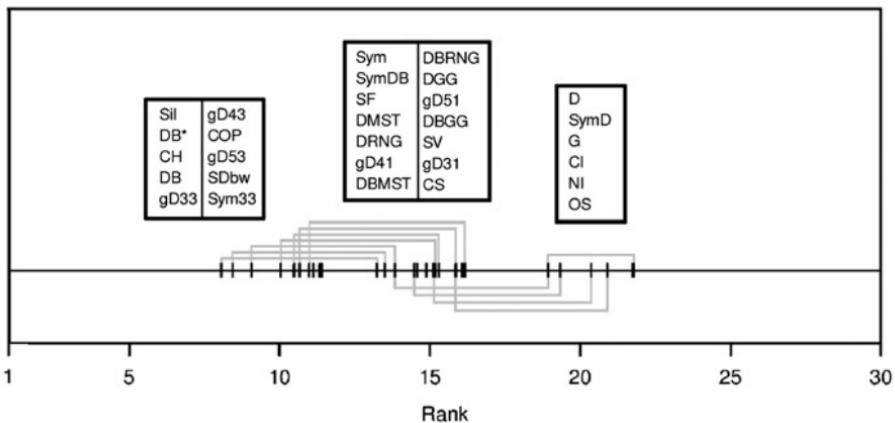
Spomínané validačné indexy môžeme vyjadriť nasledujúcimi vzorcami, a to Randov index vzorcom 21, Jaccardov index vzorcom 22 a Fowlkes-Mallowsov index vzorcom 23. Indexy sú bližšie popísané v práci [4].

$$RI = \frac{TP}{FP + FN + TP} \quad (21)$$

$$J = \frac{TP + TN}{FP + FN + TP + TN} \quad (22)$$

$$FM = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} \quad (23)$$

Interné indexy zhlukov predstavujú jedinú metriku, ktorou je možné overiť zhlukovanie pri dátach, ktoré neobsahujú skutočné triedy inštancií. Medzi používané indexy patria napr. Dunnov index, Calinski-Harabasov index, Gamma index, C-index, Davies-Bouldinov index, Silhouetteev index a mnoho ďalších. V práci [3] autori analyzovali a porovnali 30 rôznych validačných indexov na rôznych datasetoch. Na syntetických datasetoch sa najviac osvedčili Silhouetteev index, modifikovaný Davies-Bouldinov index a Calinski-Harabasov index. Pri reálnych datasetoch boli výsledky podobné, čiže indexy s horšími výsledkami dosiahnutými pri syntetických datasetoch ich dosahovali aj na reálnych dátach. Vyššie spomenuté 3 indexy dosiahli však horšie skóre ako skórovacia funkcia, generalizované Dunnove indexy a COP index.



Obr. 21: Výsledky Shafferovho testu validačných indexov so stupňom dôležitosti 10% [3].

V závere autori vyhodnotili výsledky svojich experimentov a graficky ich interpretovali pomocou Shafferovho testu 21. Nižší rank predstavuje lepšie výsledky validačného indexu na rôznych datasoch. Zároveň neexistuje výrazný štatistický rozdiel medzi jednotlivými indexami nachádzajúcimi sa v rovnakej skupine. Aj keď nie je možné jednoznačne určiť objektívne najlepší validačný index, autori odporúčajú indexy nachádzajúce sa v prvej skupine indexov a to napr. Silhouetteev index, modifikovaný Davies-Bouldinov index, Calinski-Harabaszov index, Davies-Bouldinov index, generalizovaný Dunnov index a COP index [3].

2.8 Súvisiace práce v doméne energetiky a identifikácií anomálií

V [17] bola pri zhľukovaní použitá aj kombinácia viacerých metód, konkrétnie k-means, metóda náhodnej výmeny a aglomeratívne zhľukovanie. Ako už bolo spomenuté v 2.3, úlohou algoritmu k-means namapoval existujúce inštancie do k zhľukov. Aj keď metóda náhodnej výmeny je obmedzená na zhľukovacie problémy v Euklidovskom priestore, bola použitá aj pri zhľukovaní časových radov a zabraňuje zaseknutiu zhľuku v lokálnom minime. V princípe je náhodne vybraný zhľuk, ktorý bude vymazaný a za centroid bude vybraný jeden časový rad z neho. Ak takéto riešenie je lepšie ako bez rozpustenia zhľuku je nahradené pôvodným. Ako bolo spomenuté v 2.4.1, cieľom aglomeratívneho zhľukovania je všetky časové rady označiť ako zhľuky a následne ich iteratívne zhľukovať. V momente, keď je vytvorených k zhľukov, je vypočítaný centroid zhľuku a určená hierarchia zhľukov.

V práci [7] boli pri určovaní podozrivých aktivít odberateľov úspešne aplikované rozhodovacie stromy. Po vytvorení trénovacej a testovacej množiny boli vygenerované rozhodovacie pravidlá reprezentujúce model normálnej spotreby elektrickej energie. Po predikcii boli porovnané predikované a testovacie dátá pomocou štatistickej metódy RMSE. Výsledkom experimentov je dostatočne presná predikcia spotreby energie, vypočítaná iba na základe atribútov extrahovaných z časovej známky. Prekročením stanovej hranice boli inštancie považované za anomálne. Počas experimentov boli použité M5P rozhodovacie učiace stromy.

Predmetom článku [20] bolo navrhnutie novú vlnovú techniku na reprezentovanie viacerých vlastností meraných dát. Tiež vytvorili nový model, ktorý v sebe zahrňa viacero modelov, čím je pridávanie ďalších komponentov do detekčného systému jednoduché. Navrhovaná metóda je citlivá na lokálne zmeny vo vzore dát. Taktiež dosiahli s relatívne malým množstvom meraní presnosť až 78% na trénovacej množine a 70% na testovacej množine. Metóda je citlivá na

zmeny amplitúd a frekvencií v dátach z meračov. Nevýhodou je, že model nedokáže zachytiť nevýrazné zmeny a trendy v dátach.

2.9 Zhodnotenie analýzy

Narastajúce množstvo zbieraných dát v doméne energetiky z monitorovaných systémov predstavuje množstvo skrytých znalostí. Vzniká potreba vydolovať ich a následne využiť na optimalizáciu procesov, zníženie prevádzkových nákladov alebo predpovedanie budúcej záťaže energetických sietí. Na základe nepredvídateľných udalostí alebo náhodného správania odberateľov vznikajú v datasetoch intervale, ktoré nezodpovedajú štandardnému správaniu. Tie označujeme ako intervale s výskytom anomálií. Cieľom našej práce ich bude nájsť a zmenšiť dĺžku nájdeného intervalu tak, aby bol čo najmenší, no zároveň v sebe zahrňal identifikované anomálie.

Identifikácia anomálií v časových radoch prináša so sebou viacero výziev, medzi tie najčastejšie patrí vysoká dimenzionalita dát, definícia normálneho správania, ale najmä absencia označených dát. Označenie dát je navyše náročné pre ľudského experta a taktiež sa veľmi líši definícia anomálie pri rôznych doménach. Ani normálne správanie nie je možné jednoznačne a jednoducho určiť, keďže tisíce odberateľov sa správa unikátnie. Z dostupných dát však vieme po normalizácii extrahovať vzory, ktoré po následnom zhlukovaní predstavujú rádovo menej skupín, s ktorými ďalej pracujeme ako s definíciou normálneho správania. Väčšina článkov zaobrájúca sa zhlukovaním, sa zameriava na nízkorozmerné dáta. Pri vysokodimenzionálnych dátach sú metriky podobnosti inštancií zväčša zamerané na tvary jednotlivých priebehov, než na absolútne hodnoty pozorovaní.

Cieľom našej práce je pomocou zhlukovania časových radov vhodne zadefinovať normálne správanie odberateľov a presnejšie identifikovať intervale obsahujúce anomálie. Pri zhlukovaní časových radov experimentálne overíme vhodnosť voľby hyperparametrov ako je napr. počet zhlukov, vzdialenosťná metrika alebo veľkosť použitého posuvného okna. Riedke zhluky budeme považovať za anomálne a budú podrobenej ďalšej analýze, kedy budú identifikované zlomy, lokálne a globálne anomálie.

Vzhľadom na to, že dostupné dáta neobsahujú informáciu o anomáliách, budeme pri evaluácii riešenia používať syntetický dataset, ktorý bude vytvorený na základe dostupných dát a znalostí o anomáliách.

3 Návrh riešenia

Pomocou metód strojového učenia a dátovej analytiky sa zameriame na identifikáciu anomálií v časových radoch v oblasti distribučných spoločností. Na základe dostupných dát môžu nastať dva rôzne scenáre. Ak dataset bude obsahovať iba časovú známku a spotrebu elektrickej energie daného zákazníka, zhlukovanie je možné iba na základe časového radu spotreby a výsledky budú evaluované pomocou vzdialenosí medzi jednotlivými časovými radmi vo vnútri zhlukov. Naopak, ak dataset obsahuje viaceré vysvetľujúce premenné, potom je možné vytvoriť model, ktorý bude zhlukovať odberateľov na základe týchto atribútov. Tak bude zabezpečená evaluácia pôvodného zhlukovacieho modelu. Dáta, ktoré máme k dispozícii obsahujú iba časovú známku, množstvo odoberanej elektrickej energie a príznak označujúci dni pracovného pokoja.

Z experimentov môžeme predpokladať, že zhlukovacie algoritmy vytvárajú husté a riedke zhluky. Primárne sa budeme zameriavať na analýzu časových radov, ktoré spadajú do riedkych zhlukov a už ony samotné môžu predstavovať anomálie. Cieľom je v takýchto časových radoch čo najpresnejšie identifikovať a lokalizovať intervale s neštandardným správaním odberateľa. Musíme pri tom brať ohľad najmä na cyklus dní a týždňov, no zároveň pristupovať k zvykom odberateľov jednotlivo a zvážiť ich pri označovaní anomálneho intervalu.

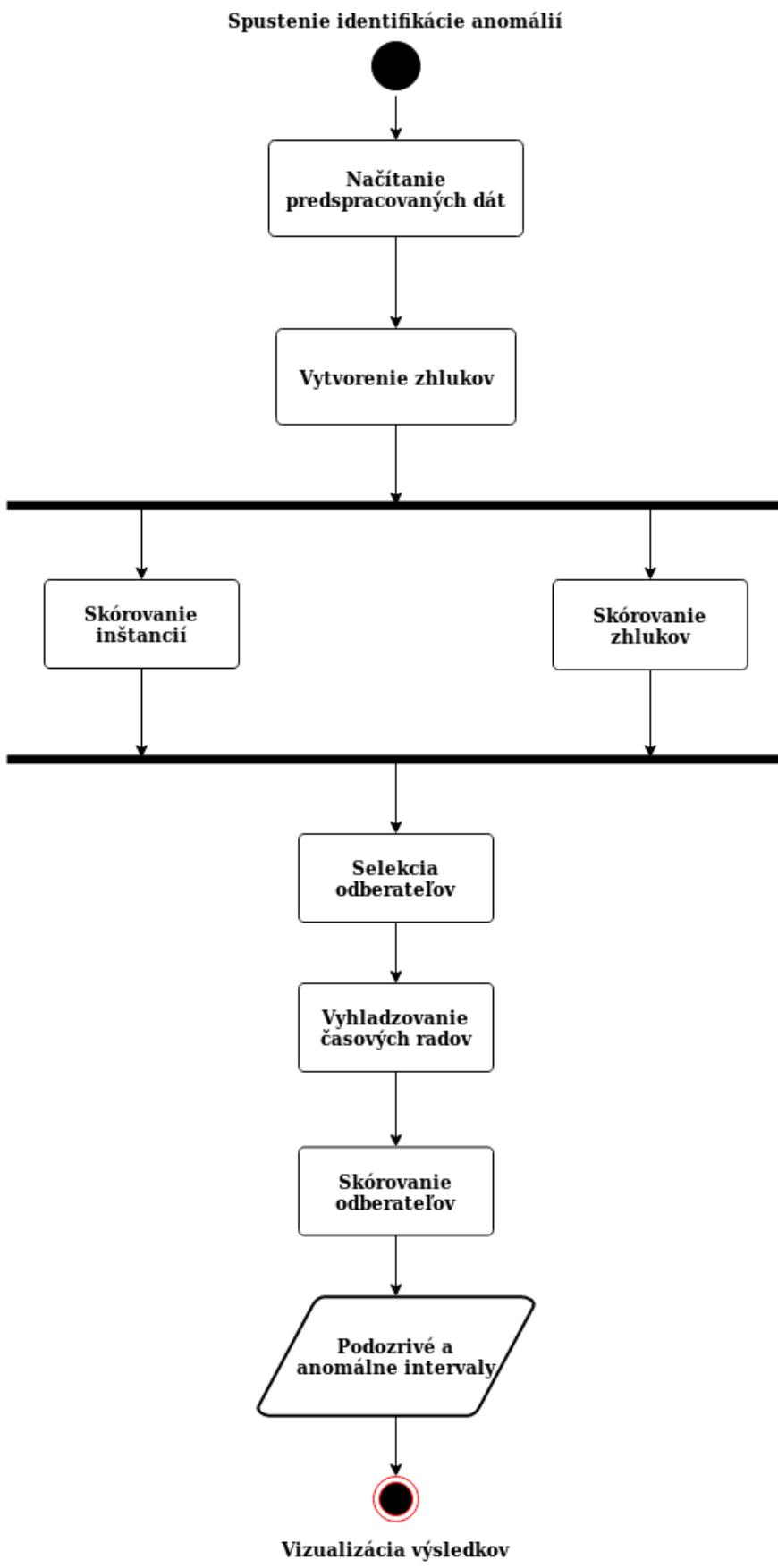
Výstupom opísaného procesu sú podezrivé a anomálne časové rady a jednotlivé merania v nich, ktoré sú taktiež považované za anomálie. Na výstupe sa môže podieľať viacero algoritmov, čo je potrebné zohľadniť pri vytváraní výsledného skóre. Na záver je potrebné zlúčiť jednotlivé merania do intervalov, ktoré svojim skóre opisujú mieru istoty, že označený interval obsahuje anomáliu. Výhodou takého spracovania je univerzálnosť riešenia, jednoduchá vizualizácia, ale najmä klasifikácia rôznych typov anomálií. Zatiaľ čo lokálne anomálie sú výsledkom krátkodobej zmeny správania odberateľa a môže sa jednať aj o výsledok náhody, globálne anomálie predstavujú výraznejšiu alebo dlhodobejšiu zmenu, ktorá môže byť predmetom záujmu distribútorov elektrickej energie.

Pre lepšie znázornenie je opísaný postup vizualizovaný stavovým diagramom na obrázku 22. Jednotlivé kroky sú ďalej rozpísané v nasledujúcich kapitolách.

Dáta sú po načítaní rozdelené do dvoch skupín. Prvá skupina obsahuje iba pracovné dni, druhá víkendy a sviatky. Cieľom je zachytiť podobné správanie odberateľov do jednej skupiny tak, aby sa neprekryvalo. Vzniknuté časové rady je nutné pred ďalším spracovaním normalizovať, napr. pomocou z-skóre. Normalizácia je potrebná kvôli použitým metrikám podobnosti časových radov, ktoré porovnávajú inštancie na základe tvaru krivky a nie ich absolútnych hodnôt ako je to napr. pri Euklidovskej. Zhluky vo vytvorenom zhlukovaní sú rozdelené na základe početnosti jednotlivých skupín na majoritné a minoritné od čoho sa odvíja hodnota skóre. Časové rady z oboch skupín sú následne analyzované pomocou metódy FeaClip a je určené skóre, na základe ktorého sú identifikované podezrivý odberatelia. Dáta podezrivých odberateľov sú vyhľadené a analyzované pomocou metódy S-H-ESD, čím vznikajú jednotlivé merania v časových radoch označené ako anomálie. Vzniknutým bodom je pridelené skóre, ktoré opisuje mieru istoty, že dané meranie je anomálne. Anomálne intervale sú pozlučované a výsledky vizualizované.

3.1 Vytvorenie zhlukov

Prvým krokom pri návrhu zhlukovania je výber vhodnej zhlukovacej metódy. Existujúce metódy sú bližšie popísané v kapitole 2.3.3. Aj na základe experimentov vykonaných autormi v práci [24] sme sa rozhodli pre metódu k-medoids, ktorá ako stred zhluku používa



Obr. 22: Stavový diagram procesu identifikácií anomálií.

inštanciu, ktoréj súčet vzdialenosí od ostatných inštancií v zhluku je čo najnižšia. Takýto vzťah môžeme zapísť rovnicou 24. Jej výhodou je najmä jednoduchosť a rýchlosť konvergencie k postačujúcim výsledkom. Rovnako ako pri k-means ide NP problém, kvôli čomu sú na vyriešenie problému použité heuristiky. Najpopulárnejšou z nich je metóda delenia okolo medoidov (angl. *Partitioning around medoids*), skrátene PAM. Najskôr je pre každú inštanciu vypočítaný najbližší medoid a súčet vzdialenosí, následne je proces opakovany so zamenením medoidov a inštanciami. Posledným krokom je výber riešenia, ktoré poskytuje najlepšie zhlukovanie.

$$\hat{\gamma} = \min \sum_{j=1}^k \sum_{x \in K_j(\lambda)} d(x, m_j) \quad (24)$$

Pri práci so zhlukovacími metódami je nutné určiť viaceru hyperparametrov, ako je napr. výsledný počet zhlukov, metrika vzdialnosti, ale aj špecifické parametre ako je veľkosť kroku a dĺžka posuvného okna. Pod dĺžkou posuvného okna rozumieme dĺžku vybraného intervalu, ktorý udávame v týždňoch. Veľkosť kroku posuvného okna je rovnako udávaná v týždňoch a predstavuje veľkosť posunu, o ktorý sa okno zmení. V prípade, že dĺžka posuvného okna a veľkosť kroku sú rovnaké, nedochádza k prekryvu okien. Pri dĺžke okna n a menšej veľkosti kroku napr. $n - 1$, sa okná prekrývajú práve v $n - 1$ týždňoch. Výhody a nevýhody metrík vzdialenosí sme bližšie analizovali už v kapitole 2.4.5. Kritériami na výber je presnosť a rýchlosť výpočtu, prípadne schopnosť spracovať aj časové rady s rôznymi dĺžkami. Veľkosť posuvného okna by nemala vyhľadávať existujúce anomálie do takej miery, že by neboli identifikované. Na druhej strane agregácia zabezpečuje elimináciu menších anomálí. Cieľom práce je identifikovať najmä rozsiahlejšie anomálie v správaní odberateľov. Veľkosť kroku posuvného okna je nutné zadefinovať tak, aby pri posune dochádzalo k prekryvu okien.

Výpočet intervalov posuvného okna môžeme zapísť vzorcami 25 a 26, pre každé okno z intervalu $<1, pocet_tyznov - dlzka_okna>$. Všetky posuvné okná sa prekrývajú minimálne v jednom týždni, práve toľko krát, koľko je veľkosť kroku posuvného okna v týždňoch. Vybraný interval dát je agregovaný na základe poradia merania v danom dni, čím vznikne denná reprezentácia odberateľa. Vybrané okno časových radov je porovnávané na základe tvaru krivky, preto je nutné dátá najskôr normalizovať a až potom analyzovať zhlukovacím algoritmom. Normalizácia pomocou z-skóre je bližšie opísaná v podkapitole 2.5.8. Jedná sa o výpočet podielu medzi rozdielom nameranej hodnoty x a jej priemerom μ a štandardnej odchýlky σ , čo môžeme zapísť vzorcom 27.

$$index_{zaciato} = (poradie_tyzdna - 1) * pocet_merani_tyzdenne \quad (25)$$

$$index_{koniec} = (poradie_tyzdna + dlzka_okna - 1) * pocet_merani_tyzdenne \quad (26)$$

$$z = \frac{x - \mu}{\sigma} \quad (27)$$

3.2 Skórovacie podozrivých zhlukov a inštancií

Pre výber riedkych zhlukov a inštancií na okraji zhlukov je potrebné vypočítať skóre, na základe ktorého bude daný časový rad považovaný za anomálny vo vybranom časovom rozmedzí. Skóre označujúce hustotu pozorovaného zhluku budeme ďalej označovať ako skóre zhluku

a skóre označujúce vzdialenosť konkrétneho časového radu od centroidu zhluku ako skóre inštancie. Ich násobením je vypočítané anomálne skóre časového radu, zapísané rovnicou 28. Predpokladom pre výpočet skóre je zhlukovanie, ktoré okrem rozdelenia inštancií do zhlukov obsahuje aj informáciu o vzdialenosťi jednotlivých inštancií od centroidu zhluku, do ktorého patrí.

$$skore_i = skore_{instancia_i} * skore_{zhluk_j}, \text{ pre } instancia_i \in zhluk_j \quad (28)$$

Vzorec pre výpočet skóre zhluku môžeme zapísať vzorcom 29 a skóre inštancie vzorcom 30. Skóre zhluku zabezpečuje penalizáciu malých zhlukov, čím je kvantilov viac a sú menšie, tým je penalizácia výraznejšia. Funkcia pre dané skóre je potom nerastúca a nadobúda hodnoty z intervalu $< 0, pocet_kvantilov >$. Skóre inštancie predstavuje pomer medzi vzdialenosťou inštancie od centroidu zhluku, do ktorého patrí a priemerom vzdialenosťí inštancií od centroidu v rovnakom zhluku.

$$skore_{zhluk_i} = \sum_{j=1}^n \begin{cases} 1 \text{ ak } P(pocetnost_i \leq Q_j) \\ 0 \text{ inak} \end{cases}, \text{ pre } j \in (0.05, 0.1, \dots, 1) \quad (29)$$

$$skore_{instancia_i} = \frac{vzdialenosť_i}{priemerna_vzdialenosť_zhluku_j}, \text{ pre } instancia_i \in zhluk_j \quad (30)$$

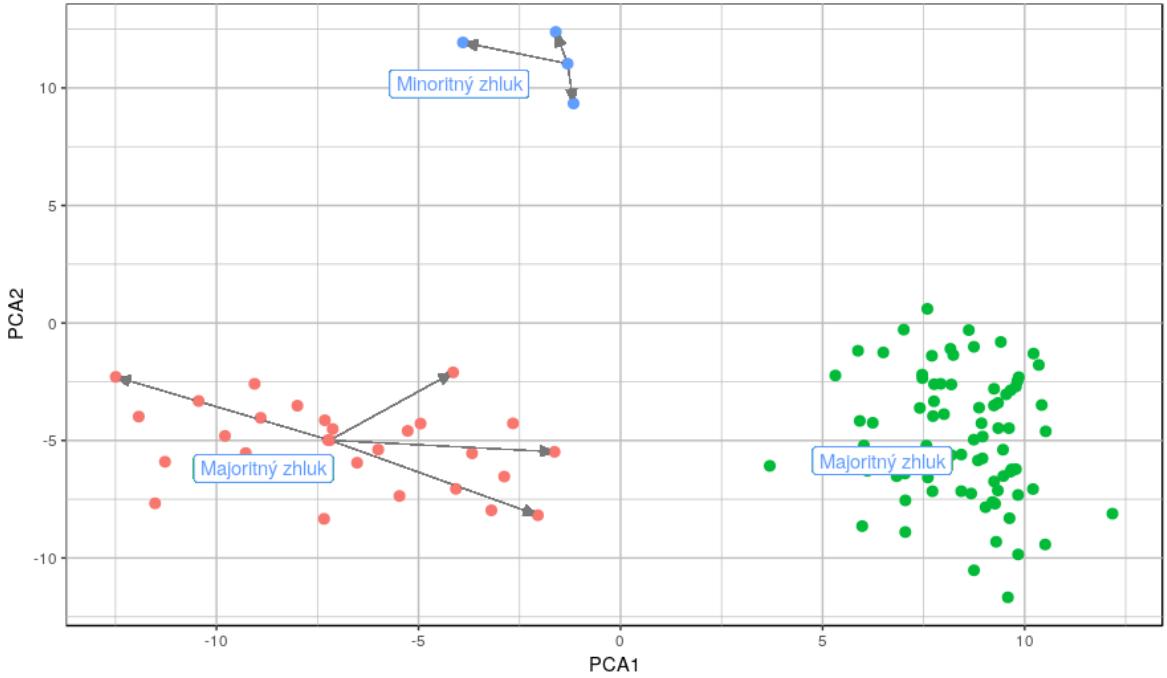
Navrhnuté skórovacie je vypočítané pre každé analyzované posuvné okno. Výpočet skórovania je zobrazený na obrázku 23. Skóre zhluku je založené na rozdelení zhlukov na majoritné a minoritné zhluky. Majoritné zhluky predstavujú zhluky, ktorých početnosť je väčšia ako kvantil Q_j pre aktuálny beh j , minoritné sú všetky ostatné. Ich početnosť nespĺňa dané kvantilové kritérium. Skóre inštancie je znázornené na obrázku 23 šedou šípkou, ktorá predstavuje vzdialenosť inštancie od medoidu daného zhluku.

3.3 Selekcia podezrivých odberateľov

Vypočítané skóre anomálnosti je potrebné vyhodnotiť a porovnávať navzájom voči ostatným navrhovaným skórovaniám. V prípade dostupnosti dát s označenými anomálnymi inštanciami je jednoduché pomocou vyhodnocovacích metrík analyzovaných v kapitole 2.7 určiť presnosť daného riešenia. Ako už bolo spomenuté, vytvorenie takéhoto datasetu je nesmierne časovo a finančne náročné.

Vyhodnocovanie vytvoreného skórovania je založené na vhodnej reprezentácii časového radu v dvojdimenziom priestoru pomocou FeaClip reprezentácie, opísanej v kapitole 2.5.4 a metódy PCA (prípadne TSNE), ktorá je bližšie opísaná v kapitole 2.5.6.

Ako už bolo spomenuté, metóda FeaClip extrahuje z dát spotreby elektrickej energie odberateľov ďalšie vlastnosti časových radov. Tým je zabezpečená redukcia vysokodimenziólnych dát a následná jednoduchá vizualizácia. Keďže metóda FeaClip je založená na transformácií dát podľa vzorca 31 a až následnej extrakcii 8 vlastností, je nutné vzniknuté časové rady opäť redukovať, napr. pomocou analýzy hlavných komponentov alebo vhodnou selekciou vzniknutých atribútov. Takými atribútmi môže byť práve počet jednotiek alebo počet prechodov medzi rôznymi behmi v pozorovanom okne časového radu, čo bližšie opísali autori v práci [25]. Pre vizualizáciu vybraného intervalu časového radu je nutná najskôr



Obr. 23: Skórovanie podozrivých inštancií a zhľukov.

FeaClip transformácia po oknách, spriemerovanie výsledkov a následný výber vhodných atribútov alebo aplikovanie PCA.

$$\hat{x}_i = \begin{cases} 1 \text{ ak } x_i > \mu \\ 0 \text{ inak} \end{cases}, \text{ pre } i \in (1, 2, \dots, n) \quad (31)$$

Z extrahovaných vlastností nás zaujímajú najmä počet jednotiek v reťazci a počet prechodov medzi rôznymi behmi. Beh predstavuje súvislú postupnosť jedného znaku. Operácie môžeme zapísť vzorcami 32 a 33. Za anomálne sú považované intervaly časových radov, ktorých vlastnosti nespadajú do intervalu medzikvartilového pravidla $< Q1 - 1.5 * IQR, Q3 + 1.5 * IQR >$. Pri vizualizácii pomocou PCA, sú anomáliami inštancie, nachádzajúce sa mimo oblasti väčšiny dát.

$$sum_1 = \sum_{i=1}^n casovy_rad_i \quad (32)$$

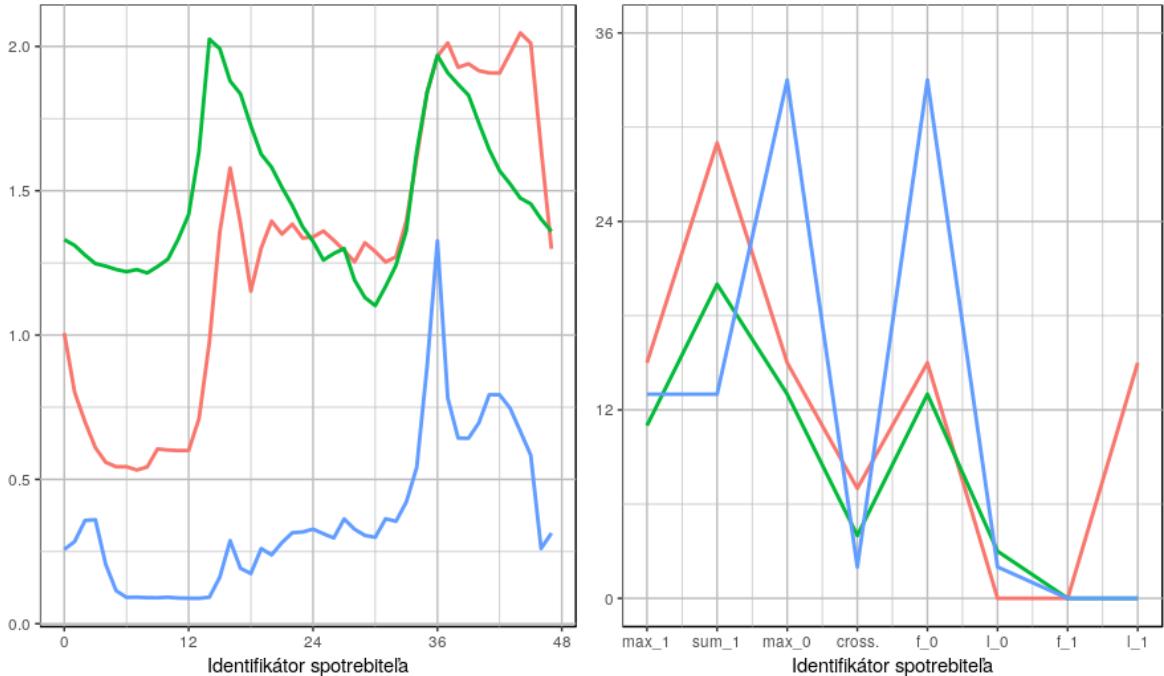
$$|prechody| = dlzka_{RLE}(casovy_rad_i) - 1 \quad (33)$$

Príkladom FeaClip transformácie môžu byť časové rady zobrazené na obrázku 24. Popis atribútov sa nachádza v tabuľke 3

Rovnako ako pri FeaClip reprezentácií, tak aj pri skórovanej inštancii môžeme na vypočítané skóre opäť uplatniť medzikvartilové pravidlo $< Q1 - 1.5 * IQR, Q3 + 1.5 * IQR >$. Vzhľadom na to, že nás zaujímajú najmä odberatelia, ktorých skóre je výrazne väčšie, budeme uvažovať iba prípady, ktoré prekračujú hornú hranicu pravidla. V prípade, že spotreba odberateľov, je identifikovaná ako anomália vo viacerých posuvných oknách, je časový rad dodatočne analyzovaný v ďalších krokoch.

Tabuľka 3: Atribúty metódy FeaClip a ich opis.

Názov atribútu	Popis atribútu
max_1	Maximálna veľkosť jednotkového behu
sum_1	Počet jednotiek v reťazci
max_0	Maximálna veľkosť nulového behu
cross.	Počet prechodov medzi rôznymi behmi
f_0	Počet nul v prvom behu
l_0	Počet nul v poslednom behu
f_1	Počet jednotiek v prvom behu
l_1	Počet jednotiek v poslednom behu

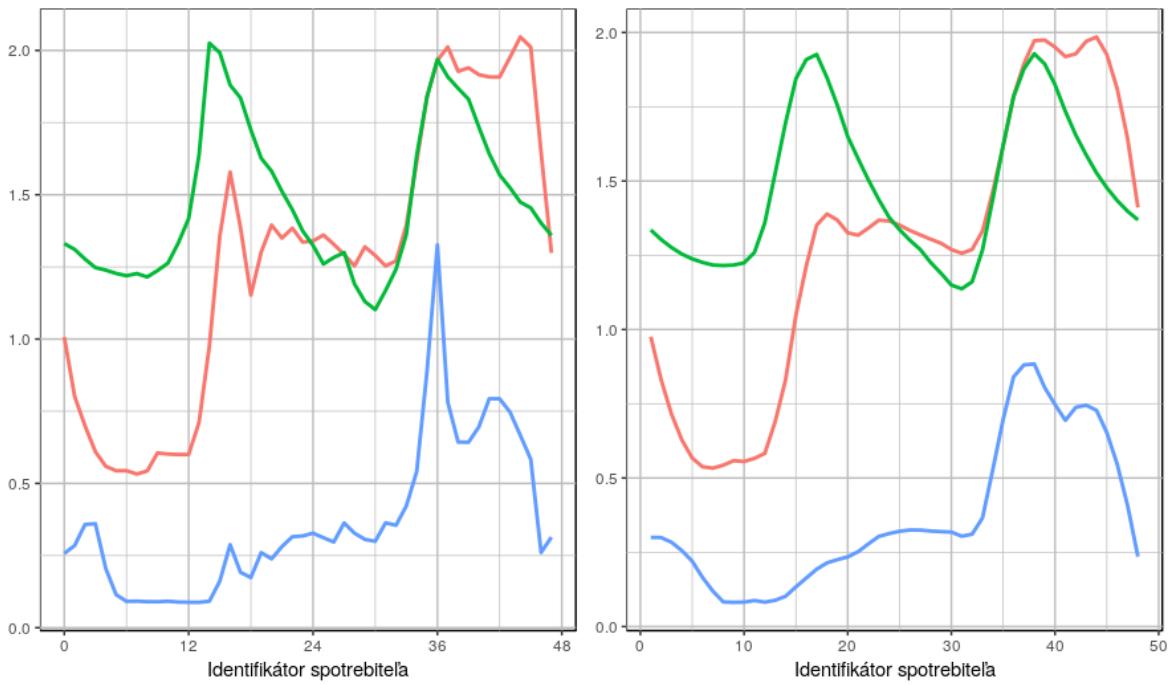


Obr. 24: Reprezentácia spotrebiteľov pomocou metódy FeaClip.

3.4 Vyhladzovanie časových radoch podozrivých odberateľov

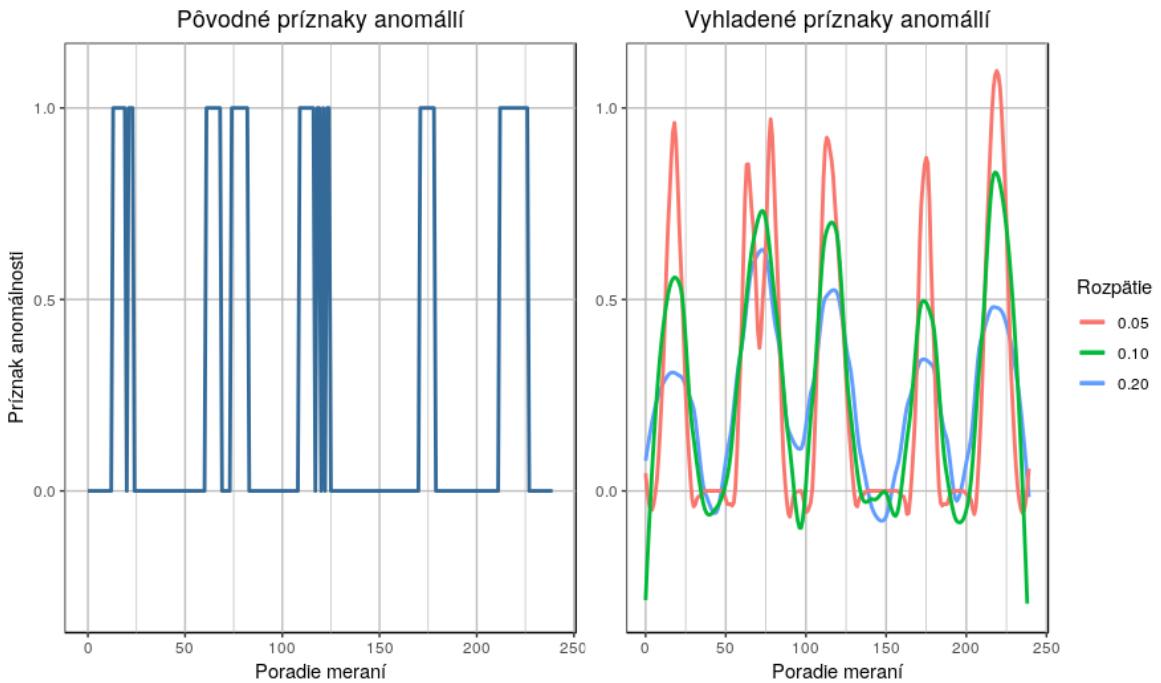
Odberatelia označení najvyšším anomálnym skóre sú ďalej analyzovaní pomocou metódy S-H-ESD. Najskôr je však potrebné spracovávané dátá vyhladiť, čím sú v podstatnej miere eliminované extrémne hodnoty a lokálne anomálie. V analýze v kapitole 2.5.9 sme opísali metódu LOESS, ktorej vstupným parametrom je rozpätie rádusu, ktoré ovplyvňuje mieru vyhľadenia spracovávaného časového radu. Je nutné správne určiť daný parameter, nakoľko je tým ovplyvnené aj ďalšie spracovanie časového radu. Vzhľadom na to, že cieľom je určiť anomálne intervale spotreby elektrickej energie, viac nás zaujímajú globálne anomálie. Vďaka ich identifikácií môže distribútor optimalizovať výrobu, distribúciu a spotrebu elektrickej energie. Príkladom aplikovania metódy LOESS, môžu byť pôvodné a vyhľadené časové rady zobrazené na obrázku 25.

Výstupom metódy S-H-ESD je príznak anomálnosti pre každé analyzované meranie. Vizualizáciou takýchto výsledkov dostaneme zúbkovitý graf, kde niektoré anomálne intervale sú jasne identifikovateľné, pri iných sa hodnota príznaku často mení a za anomálne sú považované iba samotné merania. Daný jav môžeme vidieť aj na obrázku 26, kde prechody



Obr. 25: Vyhladenie časových radov pomocou metódy LOESS s rozpäťím $\alpha = 0.25$.

medzi normálnymi a anomálnymi intervalmi v pôvodnom časovom rade príznakov sú ostré, pomocou vyhladzovania sú početné užšie intervale zlúčené a prechody zabrúsené.



Obr. 26: Vyhladenie príznakov anomálnosti pomocou metódy LOESS.

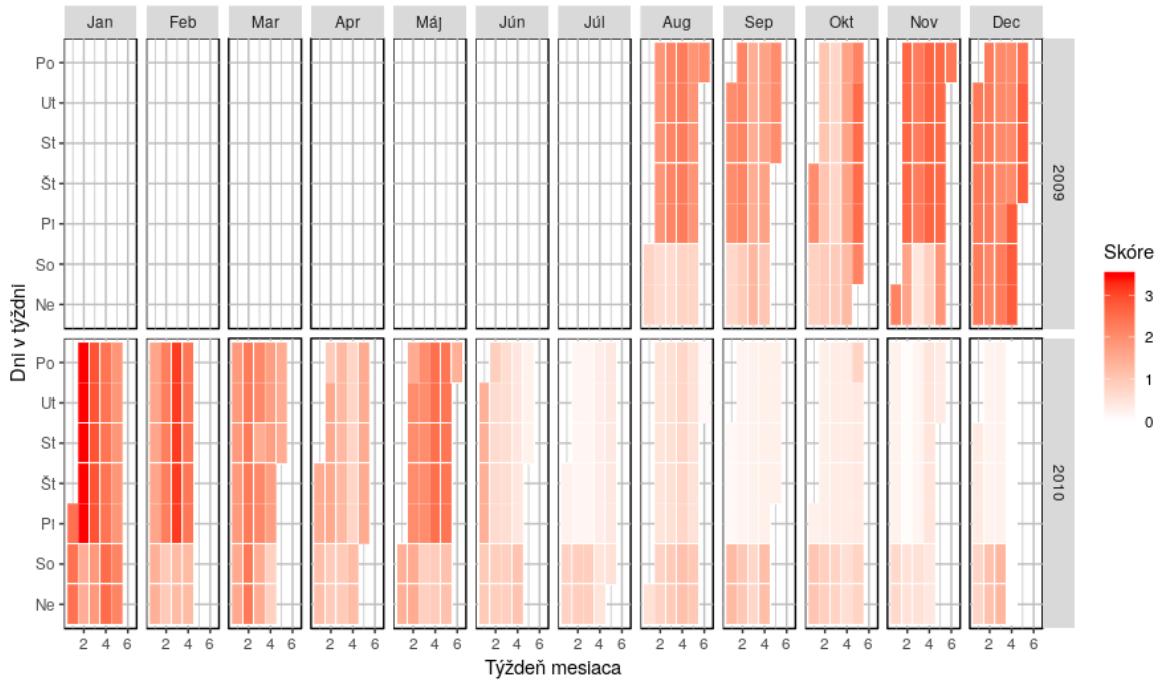
3.5 Skórovanie odberateľov metódou S-H-ESD

Cieľom je vylepšiť skóre vypočítané v predchádzajúcich krokoch a spresniť tak interval, v ktorom je výskyt anomálií výrazne väčší. Výstupom metódy S-H-ESD je pôvodný časový

rad s príznakom, ktorým sú označené identifikované anomálie. Nový atribút budeme nazývať príznak anomálnosti. Po jeho skombinovaní s navrhnutým skórováním inštancie a zhluku budeme hovoriť o skóre odberateľa, nakoľko je vypočítané pre celé sledované obdobie spotreby odberateľa. Môžeme ho vyjadriť vzorcom 34. Dôležité je uvedomiť si, že zatiaľ čo skóre inštancie a zhluku sú vypočítané pre posuvné okno, ktoré je rozdelené na pracovné dni a dni pokoja, skóre odberateľa je vypočítané pre celé sledované obdobie bez ohľadu na typ dňa. Predpokladom je, že metóda S-H-ESD dokáže spracovať časové rady s dvojitosou sezónnosťou.

$$skore_{ik} = skore_{instancia_i} * skore_{zhluk_j} + skore_{odberatel_k}, \text{ pre} \\ instancia_i \in zhluk_j \wedge \\ instancia_i \in odberatel_k \quad (34)$$

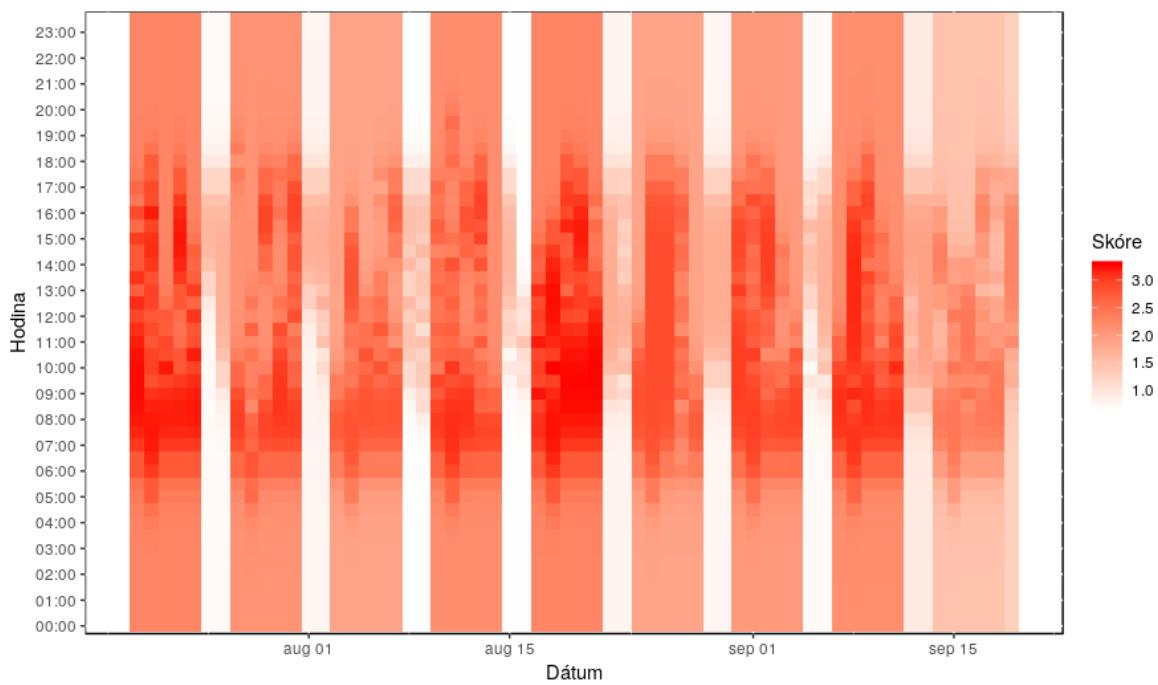
Dosiahnuté výsledky bez aplikovania S-H-ESD sú agregované v rámci jedného posuvného okna, čiže vytvorené skóre okna je rovnaké. Je možné ho spriemerovať s prekrývajúcimi sa oknami, no ak by bola veľkosť prekryvu príliš malá, nárast časovej a pamäťovej zložitosti by bol niekoľkonásobný. Pri zmenení veľkosti prekryvu na hodiny je výpočet dokonca rádovo náročnejší. Z toho dôvodu je zhlukovanie časových radov realizované pomocou posuvných okien, ktorých veľkosť je vždy niekoľko týždňov. Pri vizualizácii výsledkov pomocou tepelnej mapy, môžeme na obrázku 27 vidieť nedostatok tejto metódy, keďže rovnaký typ dňa v rámci jedného týždňa má vždy rovnaké skóre anomálnosti. Mapa zobrazuje skóre jedného odberateľa v celom pozorovanom období. Čím je skóre anomálnosti vyššie, tým je väčšia miera istoty výskytu anomálií.



Obr. 27: Vizualizácia skóre odberateľa pred pridaním S-H-ESD.

Ako už bolo spomenuté po aplikovaní metódy S-H-ESD na časový rad dostaneme pre každé meranie príznak, ktorý označuje anomálie. Časový rad príznakov je často zúbkovitý a označené intervale obsahujú jediné meranie. Po vyhľadení časového radu je sú vysledkom grafy zobrazené na obrázku 26. Dosiahnuté skóre je pripočítané k pôvodnému skóre na základe

vzorca 34. Výsledky je nutné vhodne vizualizovať a vyhodnotiť. Príkladom môže byť vizualizácia výsledkov na obrázku 28. Vzhľadom na fakt, že granularita výsledkom je o mnoho väčšia ako pred pridaním metódy, je použitý na vizualizáciu iný typ tepelnej mapy a zoobrazovaný interval je len niekoľko týždňov. Kedže navrhnutá metóda je založená na učení bez učiteľa, vyhodnocovanie prebieha na základe už spomenutých riešení, konkrétnie FeaClip transformácie a metóde S-H-ESD.



Obr. 28: Vizualizácia skóre odberateľa po pridaní S-H-ESD.

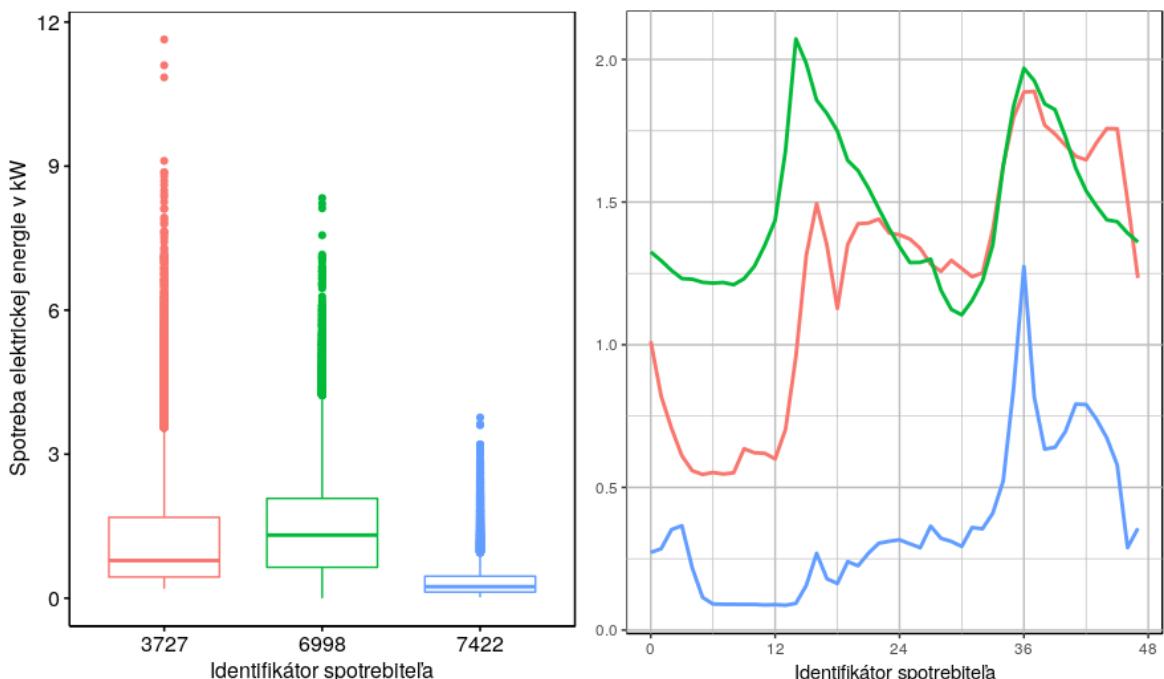
4 Experimentálne overenie

Pri experimentoch sme pracovali v jazyku R. Použili dataset 4621 írskych domácností, ktorých spotreba elektrickej energie bola počas 17 mesiacov sledovaná pomocou inteligentných meračov. Dáta boli zberané každých 15 minút v období medzi 15. júlom 2009 a 31. decembrom 2010. Dataset obsahuje iba časovú známku, spotrebú v kW a príznak sviatku. Spotreba elektrickej energie meraná v kW nadobúda hodnoty v intervale $< 0, 66.815 >$ a priemerná spotreba je 0.6727399 kW. Medián, dolný a horný kvantil je zobrazený v tabuľke 4. Štandardná odchýlka súboru je 1.372831. Je náročné prehľadne vizualizovať množstvo meraní od odberateľov, preto sme použili čiarové a krabicové grafy 29 (angl. *box plots*) na vizualizáciu náhodne vybraných odberateľov.

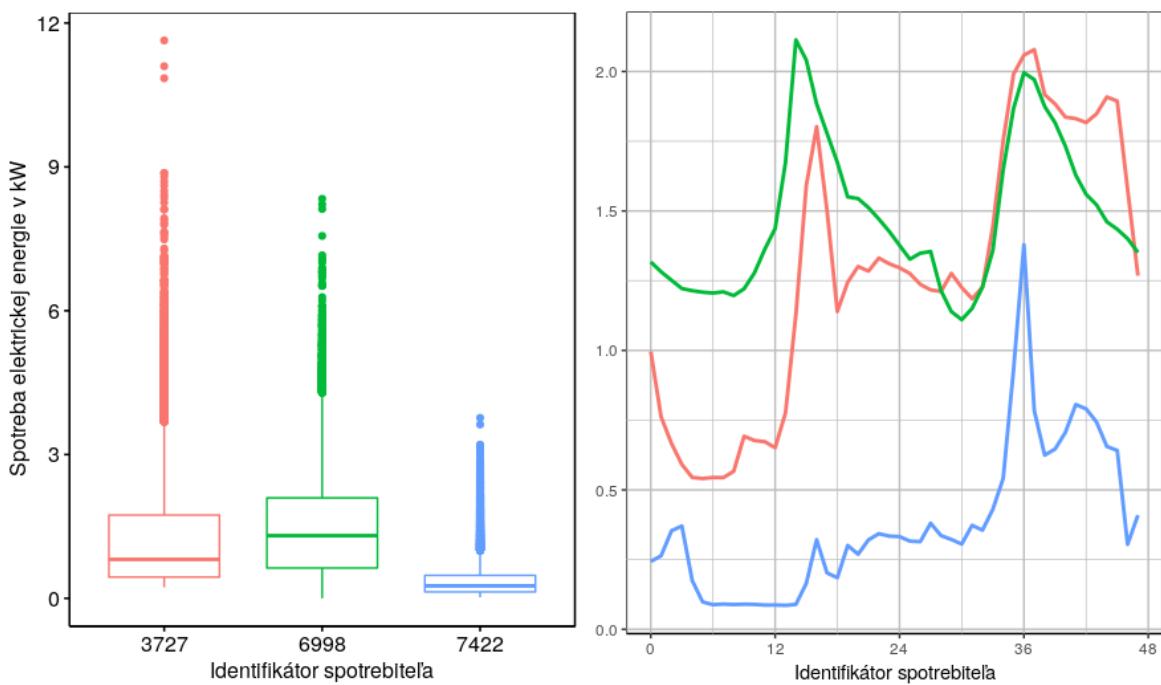
Tabuľka 4: Charakteristiky polohy použitého datasetu.

Dolný kvantil	Medián	Horný kvantil
0.121	0.269	0.666

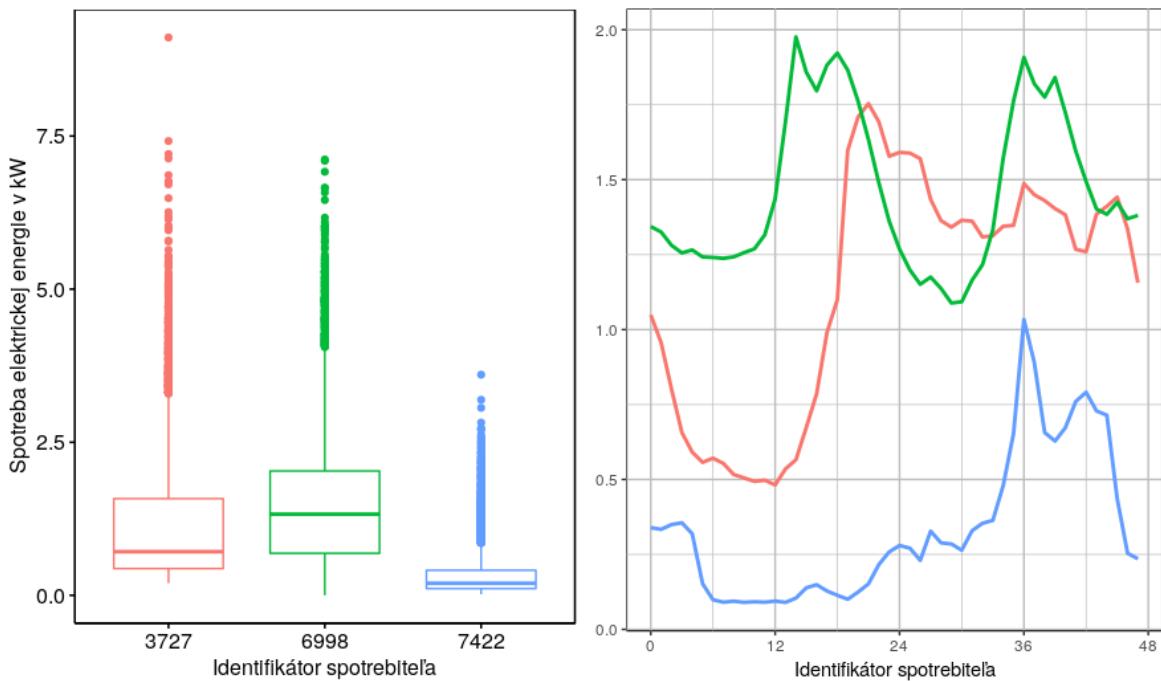
Profil spotrebiteľa sa výrazne líši počas pracovných dní a víkendov, preto je celý proces opísaný v kapitole 3 aplikovaný osobitne na pracovné dni a osobitne na dni voľna, čiže sviatky, soboty a nedele. Cieľom je zvýšiť presnosť zhľukovania a následne identifikácie anomálnych intervalov v pôvodnom datasete. Charakteristiky polohy sú prehľadne zobrazené v tabuľke 5. Štandardná odchýlka pracovných dní je 1.418979 a dní voľna 1.24807. Pre lepšiu vizualizáciu rozdielov medzi pracovnými dňami a dňami voľna sme vizualizovali spotrebu elektrickej energie rovnakých odberateľov grafmi 30 a 31.



Obr. 29: Krabicový graf spotreby odberateľov.



Obr. 30: Krabicový graf spotreby odberateľov počas pracovných dní.



Obr. 31: Krabicový graf spotreby odberateľov počas dní voľna.

Tabuľka 5: Charakteristiky polohy po rozdelení datasetu.

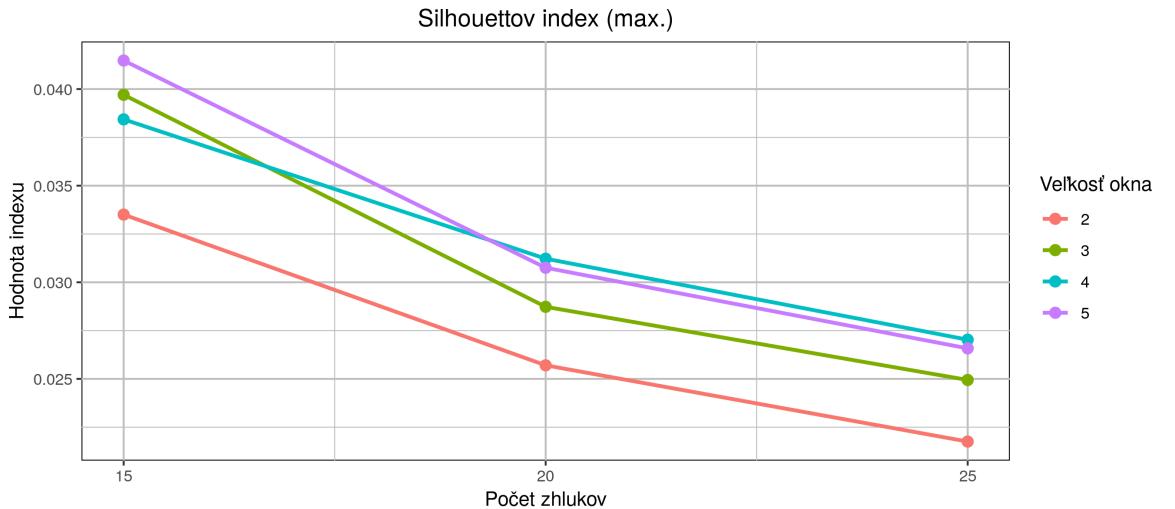
	Pracovné dni	Víkendy	Sviatky	Dni voľna
Priemer	0.6826072	0.6480718	0.7035064	0.6495951
Minimum	0	0	0	0
Dolný kvantil	0.121	0.124	0.127	0.124
Medián	0.266	0.275	0.303	0.276
Horný kvantil	0.644	0.670	0.778	0.673
Maximum	66.815	42.326	38.530	42.326

4.1 Výber hyperparametrov zhlukovania

Zhlukovacie metódy poskytujú viacero parametrov, ktoré ovplyvňujú výsledné zhlukovanie, jeho kvalitu alebo časovú náročnosť. Pri práci sme sa zamerali najmä na dosahovanú presnosť, ktorú sme merali pomocou zhlukovacích validačných indexov, bližšie opísaných v kapitole 2.7.1. Rozhodovali sme sa najmä na základe Silhouetteevho a Davies-Bouldinovho indexu. Silhouetteov index môžeme definovať pomocou rovnice 35 [3], kde a_i predstavuje priemernú vzdialenosť inštancie i od ostatných bodov nachádzajúcich sa v rovnakom zhluku, a b_i predstavuje najmenšiu priemernú vzdialenosť inštancie i od všetkých bodov nachádzajúcich sa v iných zhlukoch, čím je nájdený najbližší susedný zhluk. Davies-Bouldinov index je možné zapísť pomocou rovnice 36[3]. Hodnota S_i predstavuje mieru rozptylu meraní v rámci zhluku i a $M_{i,j}$ mieru separácie medzi zhlukmi i a j . Maximalizovaním ich podielu dostaneme priemernú hodnotu Davies-Bouldinovho indexu pre každý zhluk.

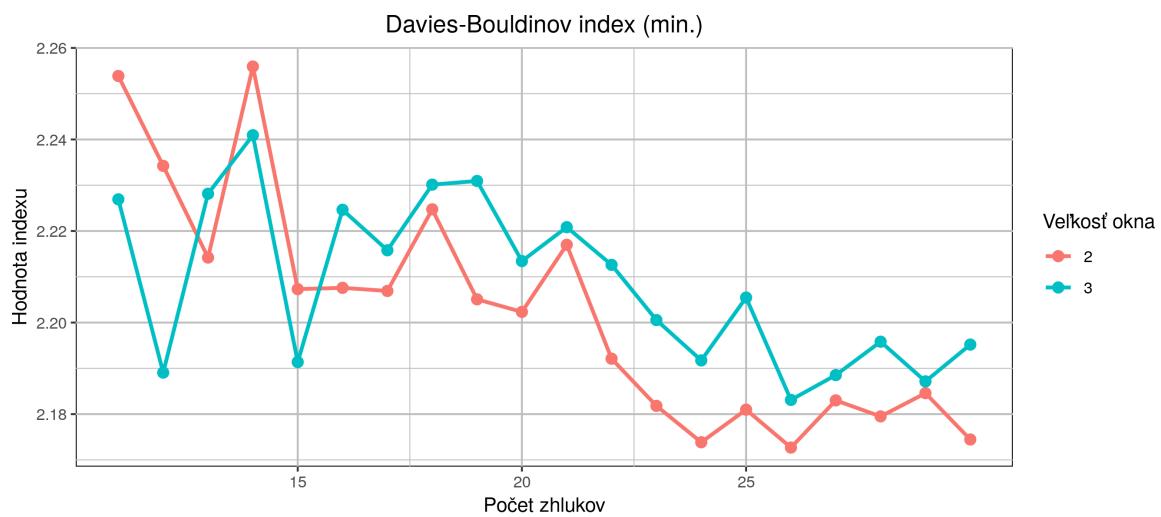
$$S_i = \begin{cases} 1 - \frac{a_i}{b_i} & \text{ak } a_i < b_i \\ 0 & \text{ak } a_i = b_i \\ \frac{b_i}{a_i} - 1 & \text{ak } a_i > b_i \end{cases} \quad (35)$$

$$DB = \frac{1}{N} \sum_{i=1}^N \max \frac{S_i + S_j}{M_{i,j}} \text{ pre } i \neq j \quad (36)$$



Obr. 32: Porovnanie veľkosti posuvného okna a počtu zhlukov pre Silhouetteov index.

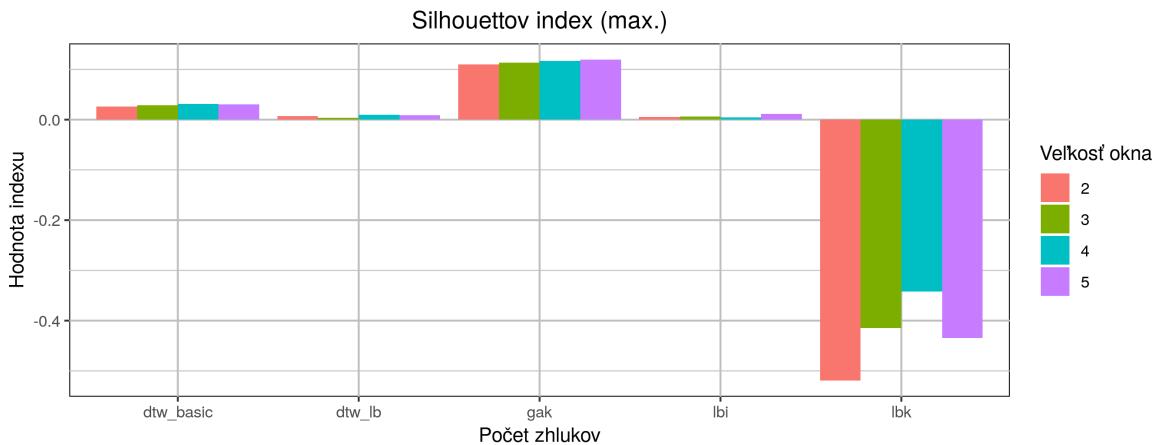
Hyperparametre sme testovali iba na požadovanom rozmedzí. Veľkosť posuvného okna by nemala presahovať 4-5 týždňov, aby okno neobsahovalo sezónnosť jednotlivých ročných období. Z vybraných grafov 32 a 33 je zrejmé, že najlepším nastavením hyper parametrov je práve nízky počet okien, ktoré budú agregované. Výsledný počet zhlukov by mal byť približne 25. Všetky výsledky experimentov sa nachádzajú v prílohe v kapitole B. Grafy podporujú naše tvrdenie, prípadne neposkytujú dostatočnú výpovednú hodnotu, keďže rozdiel medzi jednotlivými pokusmi je minimálny.



Obr. 33: Porovnanie veľkosti posuvného okna a počtu zhlukov pre Davies-Bouldinov index.

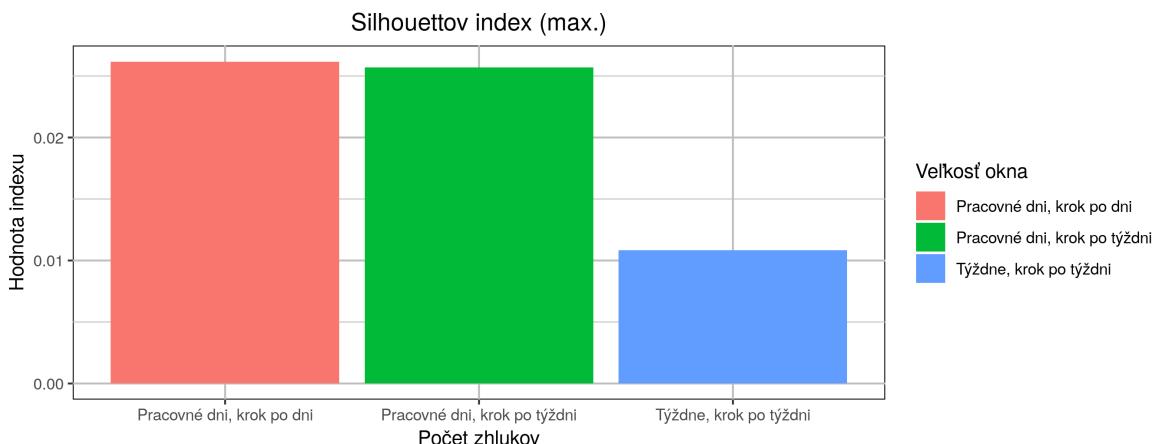
Ďalším testovaným hyperparametrom sú použité vzdialenosné metriky, ktoré sú implementované v knižnici *dtwclust*¹. Metriky sú bližšie popísané v kapitole 2.4.5. Z kapitoly 2.7.1 je zrejmé, že najlepšiu informáciu o kvalite zhlukovania poskytujú práve Silhouetteov index a modifikovaný Davies-Bouldinov index, ktoré autori v práci [3] označili za najkvalitnejšie na základe množstva experimentov na rôznych typoch datasetov. Na základe nich sme sa rozhodovali pre najlepšiu vzdialenosnú metriku. Pre Silhouetteov index dosahuje jednoznačne najlepšie výsledky GAK a za ním nasleduje DTW. Pri Davies-Bouldinovom indexe, je lepšie zhlukovanie dané nižším skóre, ktoré dosahuje DTW, tesne za ním nasleduje metrika GAK. Medzi najvhodnejšie vzdialenosné metriky pre náš dataset preto patrí GAK a DTW. Metriky sme vizualizovali pomocou grafu 34. Pri ďalších experimentoch preto budeme používať najmä GAK, ktoré dosahuje pri Silhouetteovom indexe výrazne lepšie výsledky. Vzdialenosná metrika je bližšie opísaná v kapitole 2.4.5. Je dôležité poznamenať, že pri rovnakom nastavení funkcie, sú výsledky medzi jednotlivými behmi nezávislé a rôzne. Experimentmi sme však overili, že rozdiely sú štatisticky nevýznamné.

¹<https://CRAN.R-project.org/package=dtwclust>



Obr. 34: Porovnanie vzdialenosných metrík pomocou Silhouetteovho indexu.

Dôležitým nastavením posuvného okna je jeho tvar a posun. Pri výbere tvaru sme sa zamerali najmä na pracovné dni, no na porovnanie sme vykonali experimenty aj s celými týždňami. Predpokladali sme, že zhlukovanie vytvorené iba z pracovných dní bude kvalitnejšie. Na grafe 35 si môžeme všimnúť približne rovnaké výsledky zhlukovania s posuvným oknom nad pracovnými dňami. Pri veľkosti posunu sme porovnávali iba experimenty vykonané nad pracovnými dňami. Výsledky experimentov nie sú signifikantne rozdielne, preto sme zvolili časovo menej náročný výpočet s posunom po týždňoch. Beh zhlukovania s dňovým posunom trval 5-krát dlhšie oproti týždňovému posunu.



Obr. 35: Porovnanie veľkostí a typov posuvných okien pomocou Silhouetteovho indexu.

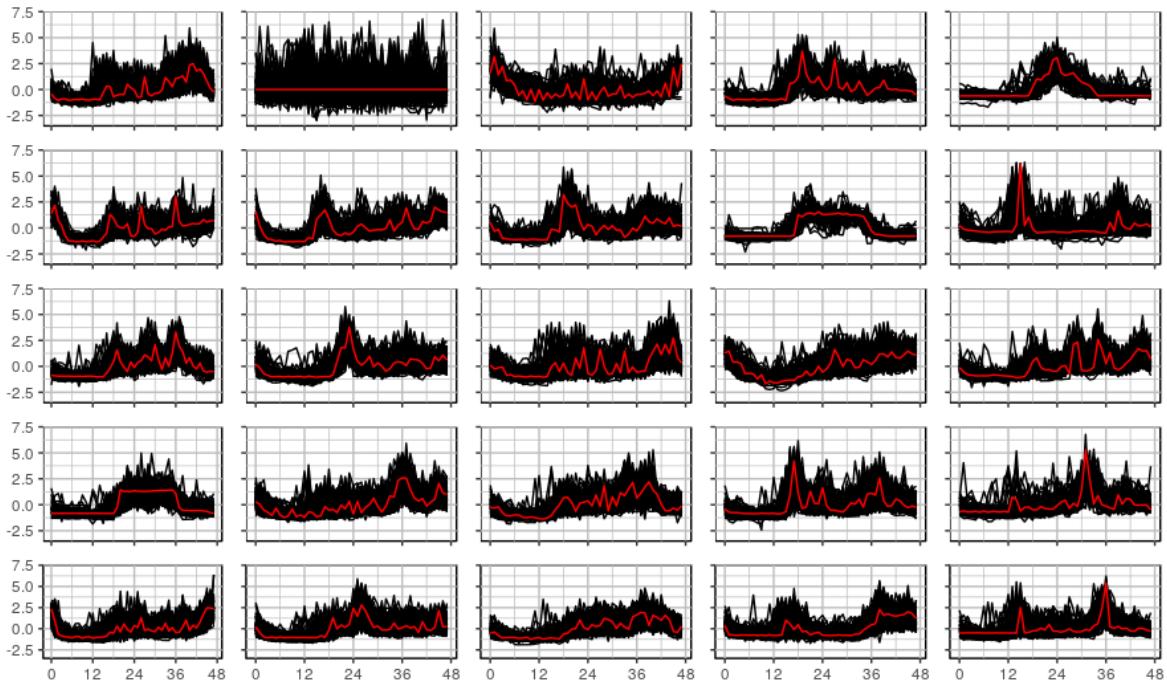
Predspracovanie datasetu pozostáva aj z normalizácie dát pomocou z-skóre, ktoré je bližšie opísané v kapitole 2.5.8. Môžeme ho zapísať vzorcom 37, kde rozdiel nameranej hodnoty a priemeru sledovanej veličiny je podelený štandardnou odchýlkou. Použitá knižnica *dtw-clust*² v jazyku R poskytuje taktiež predspracovanie vstupnej množiny dát pomocou rovnakej normalizácie. Preto sme vykonali niekoľko experimentov pre porovnanie časovej náročnosti a presnosti výsledného zhlukovania, pri použití vstavanej a externej normalizácie. Časová náročnosť pri použití oboch normalizácií súčasne alebo iba jednej z nich bola približne rovnaká. Rozdiel bol vo výsledkoch, ktoré nepoužívali externú normalizáciu. V prípade použitia

²<https://CRAN.R-project.org/package=dtwclust>

oboch súčasne alebo iba externej normalizácie sú dosahované výsledky porovnateľné.

$$z = \frac{x - \mu}{\sigma} \quad (37)$$

Výsledkom experimentov je nové zhlukovanie pre každé posuvné okno. Náhodne vybrané zhlukovanie je zobrazené aj na obrázku 36. Je zrejmé, že jednotlivé medoidy sa navzájom dostatočne líšia a potvrdzujú správnu voľbu hyperparametrov zhlukovacieho algoritmu.



Obr. 36: Vizualizácia vytvoreného zhlukovania so zvýraznenými medoidmi.

4.2 Vyhodnotenie navrhovanej metódy

Prvým krokom nami navrhutej metódy je predspracovanie a normalizácia dát, kedy sú dátá normalizované a následne rozdelené na dve skupiny. Prvá skupina obsahuje pracovné dni v týždni, druhá skupina dni voľna, čiže sviatky a víkendy. Rozdelenie je navrhnuté na základe porovnania charakteristických polohy datasetu v kapitole 4.

Obe skupiny dát sú následne postupne zhlukované pomocou k-medoidov. Každé zhlukovanie je vykonané na agregovaných časových radoch v rámci jedného posuvného okna. Dĺžku okna sme určili na základe experimentov popísaných v kapitole 4.1 na hodnotu 2. Snahou bolo zabezpečiť rýchlu odozvu na nové trendy v správaní používateľov a zároveň sčasti eliminovať množstvo lokálnych anomalií. Podobnými experimentmi sme identifikovali najvhodnejšiu vzdialenosťnú metriku pre PAM heuristiku a počet zhlukov, do ktorých bude množina odberateľov rozdelená. Používanou metrikou je GAK a počet zhlukov s najlepšími výsledkami je 25.

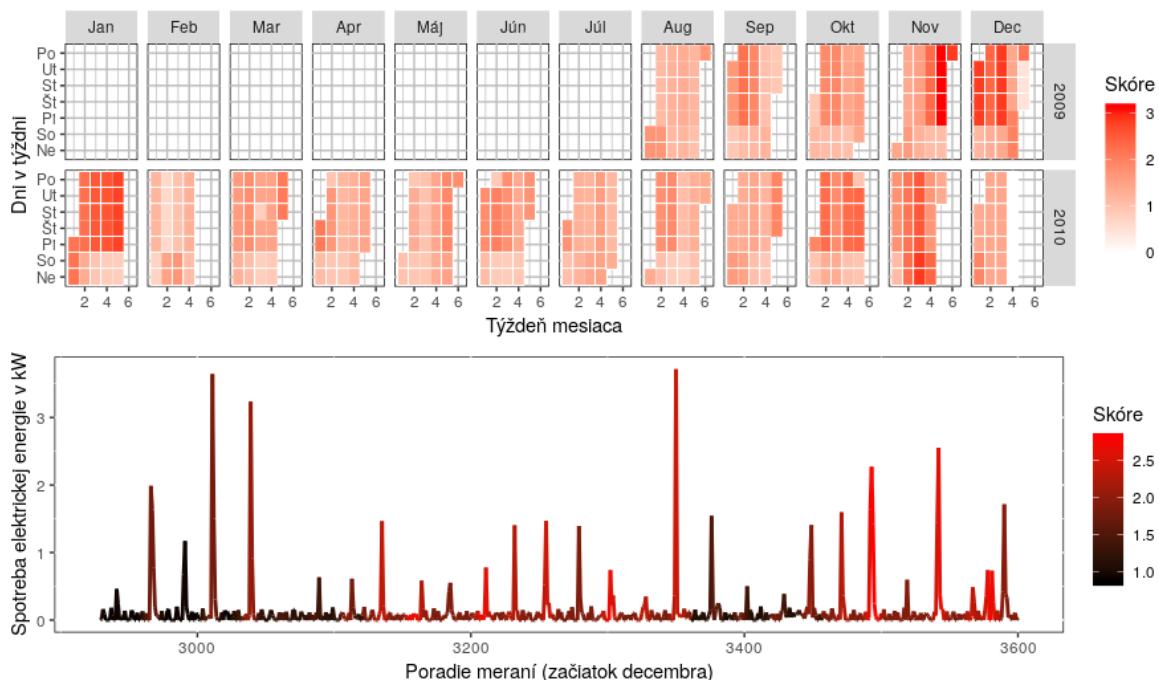
Výstupom zhlukovania je pre každého odberateľa vo všetkých posuvných oknách skóre, ktoré opisuje jeho polohu v rámci zhluku, do ktorého spadá ako aj veľkosť samotného zhluku oproti ostatným zhlukom. Dodatočne je pre každé posuvné okno taktiež vypočítané skóre, ktoré vychádza z FeaClip transformácie. Odberatelia, ktorých skóre sa v jednotlivých oknách

často vyskytuje mimo intervalu medzikvartilového pravidla sú označení a analyzovaní v ďalších krokoch procesu. Ich počet závisí od parametra, ktorý udáva používateľ. Výstupom sú identifikátory odberateľov, ktorých anomálne skóre bolo najvyššie.

Pôvodné časové rady odberateľov sú následne predspracované vyhľadzovacou metódou LOESS, ktorej cieľom je vyhľadiť vysokú volatilitu spotreby elektrickej energie. Vďaka tomu sú menej výrazné lokálne anomálie zanedbané a metóda S-H-ESD ich neidentifikuje.

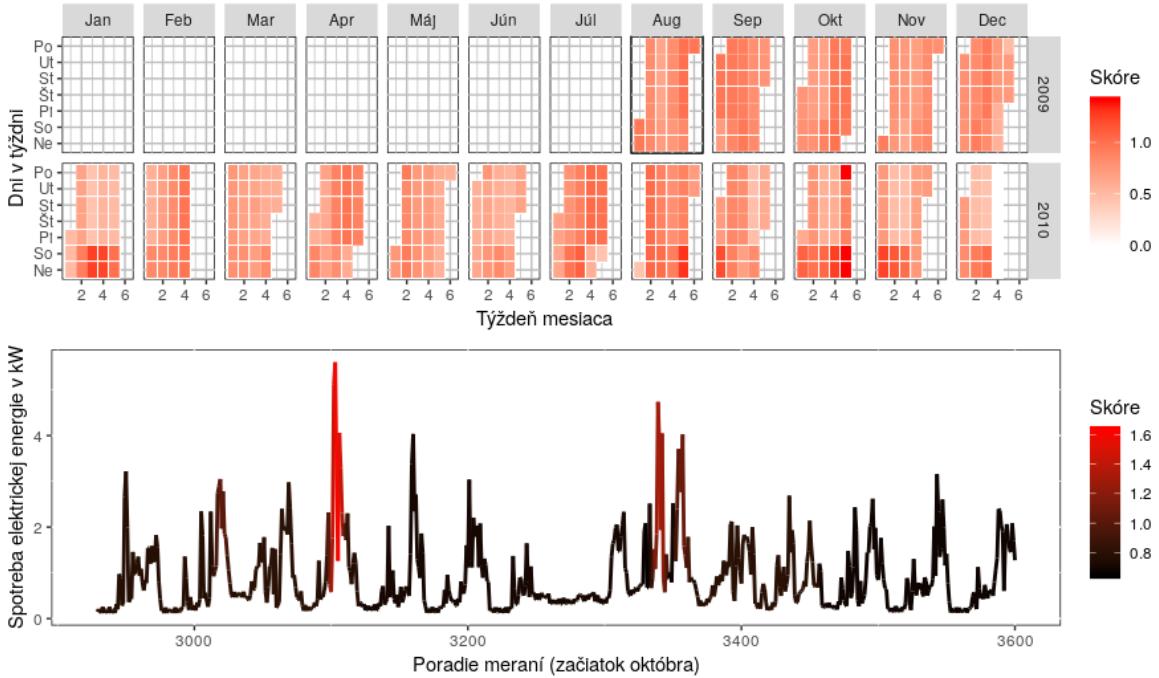
Posledným krokom je aplikovanie štatistickej metódy S-H-ESD na vyhľadené časové rady. Získame tým skóre odberateľa, ktoré na najnižšej granularite identifikuje podozrivé merania. Je dôležité si uvedomiť, že skóre vypočítané v predchádzajúcich krokoch určuje anomálie na maximálne dňovej granularite. Vzhľadom na fakt, že skóre odberateľa nadobúda iba hodnoty 0 a 1, je pred kombinovaním s pôvodným skóre vyhľadené. Použitím metódy LOESS je zabezpečené čiastočné zaniknutie ojedinelých anomálií a zároveň zlúčenie intervalov, kde je výskyt anomálií vysoký. Kombináciou skórovania získavame pre každé pôvodné meranie novú hodnotu, ktorým je určená miera podozrenia, že dané meranie je anomáliou. Čím je výsledné skóre väčšie, tým je väčšie aj podozrenie.

Dosiahnuté výsledky je možné vizualizovať rôznymi grafmi, napr. aj pomocou tepelnej mapy, ako je to v kapitole 3.5. Nevýhodou však je, že nemáme k dispozícii pôvodný časový rad odberateľa. Preto sme sa rozhodli použiť tepelnú mapu v kombinácii s bežným čiarovým grafom, ktorého farba sa mení v závislosti od vypočítaného skóre. Vybrané časové rady sú zobrazené na obrázkoch 37 a 38. Môžeme si všimnúť, že spotreba elektrickej energie na obrázku 37 je veľmi nepravidelná a náhodná. Anomáliami sú označené intervale, v ktorých sa nachádza extrémne vysoká spotreba. Naopak časový rad na obrázku 38 má pomerne pravidelný denný priebeh a anomálnym intervalom je označený deň, ktorého priebeh nie je štandardný.



Obr. 37: Vizualizácia vypočítaného skóre pre odberateľa 2172.

Množstvo identifikovaných odberateľov je udávané vstupným parametrom. Pri experimentoch sme použili 1% tých, ktorých skóre anomálnosti bolo najväčšie. Počet identifikovaných anomálií jednotlivých odberateľov sa pohybuje na úrovni 3-4% v závislosti od nastavenia



Obr. 38: Vizualizácia vypočítaného skóre pre odberateľa 6536.

vstupných premenných a samozrejme aj od zhlukovania, ktoré je stochastické. Doba spracovania odberateľa je o 19.42 sekundy dlhšia ako pri metóde S-H-ESD, ktorej to zaberie priemerne 63.24 sekúnd. Dôležité je poznamenať, že zatiaľ čo pri metóde S-H-ESD je potrebné analyzovať všetky časové rady, pri nami navrhovanej metóde je nutné aplikovať na celý dataset iba zhlukovanie. Priemerný čas výpočtu na odberateľa je pritom iba 14.79 sekundy. Z uvedených výsledkov je možné vyhodnotiť, že metóda je schopná identifikovať anomálne intervale, ktoré sa svojou charakteristikou líšia od predchádzajúcich períód. Môže sa pritom jednať o opakujúce sa anomálie, no zároveň predstavuje vhodné riešenie identifikácie skokových anomalií.

4.3 Porovnanie existujúcich riešení

Experimenty, ktoré sme vykonali sme porovnávali s existujúcou implementáciou metódy S-H-ESD, ktorá sa nachádza v balíčku *AnomalyDetection*³. Autori ju bližšie opísali v práci [44]. Rovnako sme výsledky porovnávali aj s metódou FeaClip, ktorú navrhli autori v práci [25]. Výsledky porovnania sa nachádzajú v tabuľke 6.

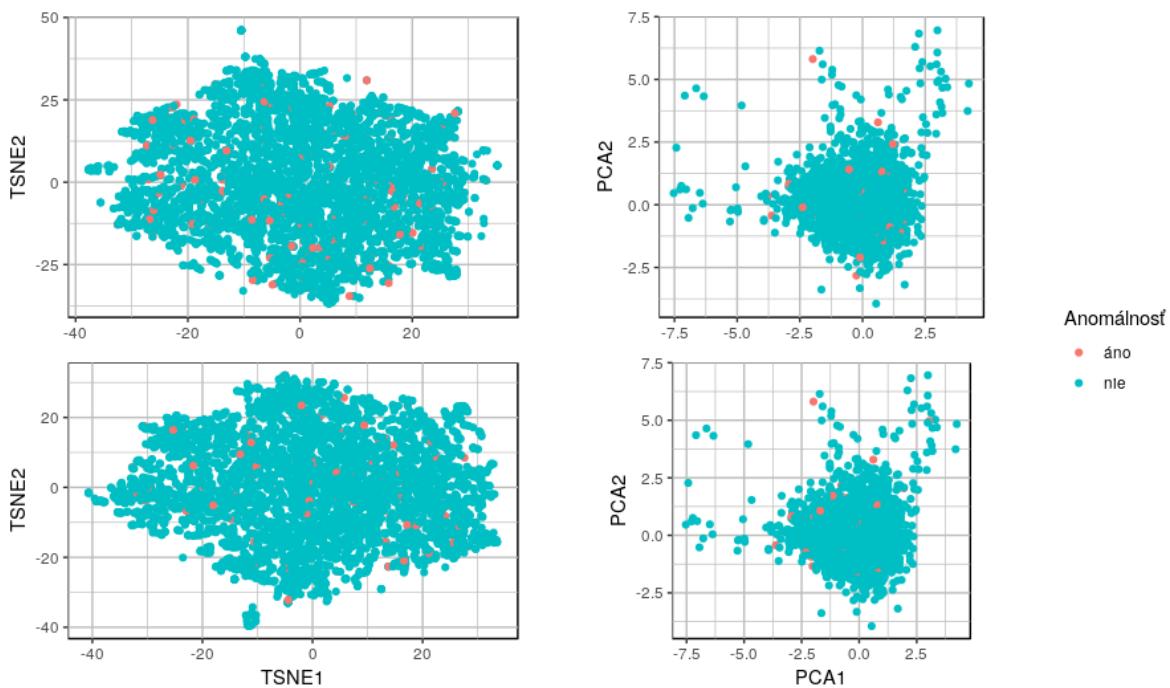
Tabuľka 6: Porovnanie výsledkov spracovania existujúcich riešení.

	Priemerná doba spracovania jedného odberateľa	Počet identifikovaných odberateľov
S-H-ESD <i>AnomalyDetectionTs</i>	63.24 sekúnd	3.05 %
S-H-ESD <i>AnomalyDetectionVec</i>	12.31 sekúnd	2.27 %
FeaClip	0.0197 sekúnd	34.83 %
FeaClip agregované	0.4605 sekúnd	2.66 %

³<https://github.com/twitter/AnomalyDetection>

4.3.1 Identifikácia anomálií metódou S-H-ESD

Vstupnými parametrami metódy je hladina významnosti, na ktorej je hypotéza o anomálnosti inštancie zamietaná, typ anomálie (kladná alebo záporná) a maximálny počet anomálií v danom datasete. Vďaka robustnosti metódy S-H-ESD, je možné identifikovať až 50% anomálií, preto sme sa rozhodli toto nastavenie ponechať a upraviť iba hladinu významnosti na $\alpha = 0.001$. Pri experimentoch bola priemerná doba spracovania jedného odberateľa 63.24 sekúnd so štandardnou odchýlkou 1.8 sekundy. Najrýchlejšie spracovanie trvalo presne minútu, najpomalšie 70 sekúnd. Priemerný počet identifikovaných anomálií je 14.11%. V prípade, že sme ako anomálnych označili tých odberateľov, ktorých počet identifikovaných anomálií nespĺňa medzikvartilové pravidlo, čiže nespadá do intervalu $< Q1 - 1.5 * IQR, Q3 + 1.5 * IQR >$, ich počet bol približne 2-3% z celkového počtu odberateľov. Opäť sme uvažovali iba odberateľov, ktorých skóre nespĺňa hornú hranicu pravidla, nakoľko nás nezaujímajú štandardný odberatelia. Výsledky sme vizualizovali pomocou transformácie FeaClip a sú zobrazené na obrázku 39. V prvom riadku sú výsledky funkcie *AnomalyDetectionTs* s 3.05% anomálnych odberateľov, v druhom riadku sú výsledky funkcie *AnomalyDetectionVec* s 2.27% anomálnych odberateľov. Anomálne inštancie sa obvykle nachádzajú na okraji zhlukov, no nie je to pravidlom. Pri vizualizácii vysokodimenzionálnych dát často nie je možné jednoznačne určiť dostatočne reprezentatívne atribúty a z toho dôvodu nám metóda PCA a TSNE neposkytuje žiadnu informáciu o anomálnosti odberateľa.

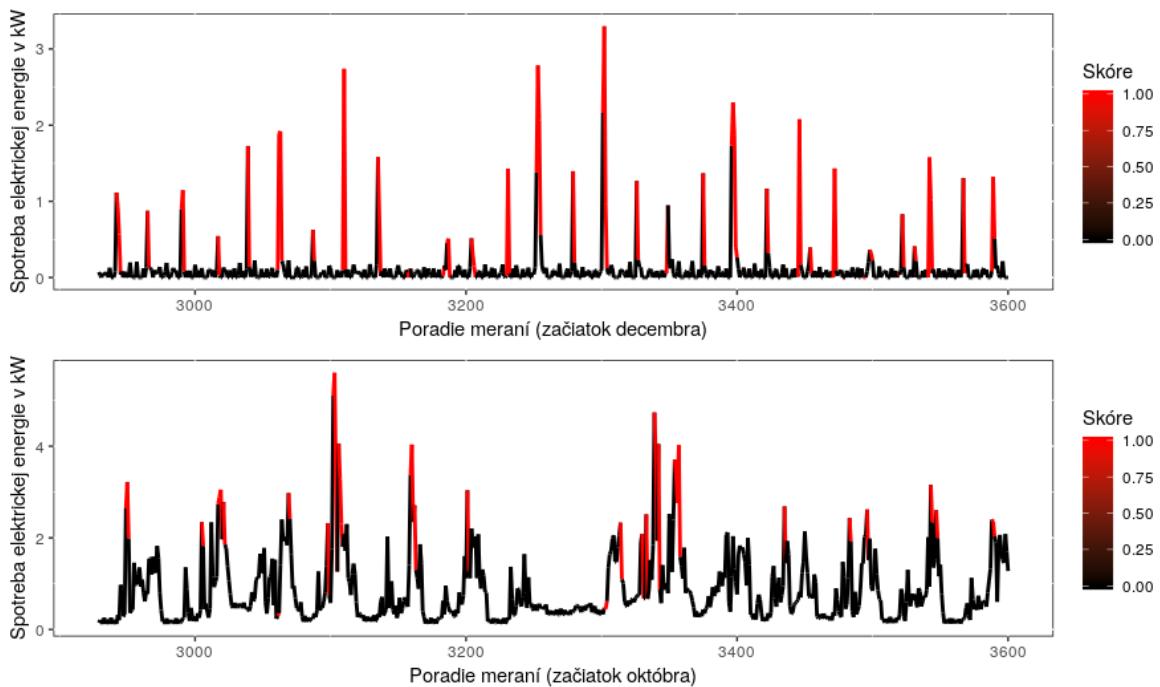


Obr. 39: Vizualizácia všetkých odberateľov, anomálie identifikované pomocou S-H-ESD, hore *AnomalyDetectionTs*, dole *AnomalyDetectionVec*

Funkcia *AnomalyDetectionTs* očakáva na vstupe časový rad, ktorý obsahuje časovú známku a hodnotu sledovanej veličiny v danom čase. Nutnou podmienkou je dostatočne dlhý časový rad, nad ktorým je spustená identifikácia anomálií. Pri experimentoch, kedy bola analyzovaná iba časť časového radu je kvôli tomu nutné zvoliť obdobie aspoň 9 týždňov. Pre funkciu *AnomalyDetectionVec* používateľ parametrami definuje krátkodobú a dlhodobú periódu dát a spomínané obmedzenie neexistuje. Navyše výsledky sú rovnaké ako pri analýze celého

časového radu, tak aj pri čiastkových analýzach, čo môže prispieť k rýchlosťi identifikácie anomalií.

Vizualizované výsledky môžeme vidieť na obrázku 40. Z oboch je zrejmé, že anomálnymi meraniami sú označené merania, ktorých hodnota je lokálnym extrémom. Pri dolnom obrázku odberateľa 6536 je tak označená časť takmer každého dňa pozorovaného intervalu, hoc môžeme vidieť, že existuje určitá pravidelnosť v spotrebe elektrickej energie.



Obr. 40: Vizualizácia S-H-ESD skóre odberateľov, hore 2172, dole 6536.

4.3.2 Identifikácia anomalií metódou FeaClip

Jediným vstupným parametrom metódy je postupnosť meraní, nad ktorou je vykonaný výpočet. Metóda je extrémne rýchla, odberateľa transformuje a aplikuje medzikvartilové pravidlo primerne za 0.0197 sekundy. Za anomálnych je však označených až tretina odberateľov na celom pozorovanom intervale. Kvôli tomu je nutné vykonať transformáciu iba nad posuvným oknom, ktorého výsledky sú agregované a je naň aplikované medzikvartilové pravidlo na identifikáciu anomalií. Výpočet sa tak predĺži na 0.4605 sekundy pre každého odberateľa. Počet identifikovaných anomálnych odberateľov sa snížil približne na 2.66%, pri výbere 10% najčastejšie označovaných odberateľov. Priemerne je v každom okne označených ako anomálnych 63.4 odberateľov. Oba časové rady identifikované metódou S-H-ESD zobrazené na obrázku 40 sú označené anomálnymi aj FeaClip metódou. Jej nevýhodou však je, že ako anomáliu označí celé analyzované posuvné okno.

5 Zhodnotenie

Cieľom našej práce bolo navrhnuť a overiť riešenie, ktoré v časových radoch spotreby elektrickej energie identifikuje podozrivé intervaly a merania. Nemusí pritom nutne ísť o anomálie spôsobené manipuláciou s meracím zariadením, ale môže sa jednať aj o jeho poškodenie prípadne môže byť správanie odberateľa nepredvídateľné. Identifikácia podozrivých odberateľov je kvôli množstvu dát a rýchlosťi s akou pribúdajú, nesmierne finančne a časovo náročná. Zároveň musí byť navrhnuté riešenie dostatočne robustné voči meniacim sa trendom spotreby elektrickej energie. Preto je požiadavkou na riešenie robustnosť, ale najmä schopnosť učenia bez učiteľa. Navrhovaná metóda je založená na zhlukovaní časových radov postupne po niekoľko týždňových intervaloch, čím je zabezpečená rýchla adaptácia na zmenu v prostredí. Zhlukovanie bolo validované pomocou zhlukovacích validačných indexov. Anomálnosť intervalu a merania je určená na základe skórovania, ktoré vychádza zo spomínaneho zhlukovania a štatistickej metódy S-H-ESD, ktorá je primárne určená na identifikáciu anomalií. Pri experimentoch však S-H-ESD nedosahuje postačujúce výsledky a často dochádza k falošne pozitívnym označeniam anomalií. Pridaním zhlukovania sme sa snažili zredukovať takéto prípady a docieliť presnejšiu identifikáciu anomalií.

V analýze sme sa zamerali na používané metódy v doméne identifikácie anomalií, spôsoby predspracovania vysokodimenziólnych dát, ale aj existujúce práce v danej doméne. Bližšie sme analyzovali oblasť zhlukovania časových radov a štatistickej metód používaných pri detekcii anomalií. Pri predspracovaní dát sme sa zamerali najmä na vhodnú redukciu dimenzií a normalizáciu dát, nakoľko sme pri zhlukovaní používali vzdialenosťné metriky založené na tvare kriviek časových radov. V návrhu riešenia sme opísali nami navrhovanú metódu a bližšie sme opísali jednotlivé kroky metódy. Dáta sú počas spracovania rozdelené, normalizované, agregované, zhlukané, vyhľadzované a nakoniec aj vizualizované pomocou vhodne aplikovaných grafov.

Experimentmi sme overili nami navrhnuté riešenie a výsledky sme vhodne vizualizovali a následne porovnali s existujúcimi metódami. Na základe výberu náhodnej vzorky časových radov môžeme zhodnotiť až štvornásobné zrýchlenie výpočtu a zároveň spresnenie výskytu anomálnych intervalov. Pri niektorých inštanciách sa jedná aj o dosiahnutie lepších výsledkov. Pod tým môžeme rozumieť identifikovanie anomálneho intervalu namiesto jednotlivých hodnôt, ale aj zredukovanie falošne pozitívnych hlásení.

Literatúra

- [1] Adhikari, R.: *An Introductory Study on Time Series Modeling and Forecasting*. Saarbrücken: LAP LAMBERT Academic Publishing, 2013, ISBN 9783659335082.
- [2] Arampatzis, A.; Kamps, J.: A Signal-to-noise Approach to Score Normalization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, New York, NY, USA: ACM, 2009, ISBN 978-1-60558-512-3, s. 797–806, doi:10.1145/1645953.1646055.
- [3] Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; aj.: An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recogn.*, ročník 46, č. 1, jan 2013: s. 243–256, ISSN 0031-3203, doi:10.1016/j.patcog.2012.07.021.
- [4] Bilgic, E.; Cakir, O.: Comparing clusterings: a store segmentation application, 10 2018, (čaká na publikovanie).
- [5] Chakrabarti, K.; Keogh, E.; Mehrotra, S.; aj.: Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. *ACM Trans. Database Syst.*, ročník 27, č. 2, Jún 2002: s. 188–228, ISSN 0362-5915, doi:10.1145/568518.568520.
- [6] Chandola, V.; Banerjee, A.; Kumar, V.: Anomaly Detection: A Survey. *ACM Comput. Surv.*, ročník 41, č. 3, jul 2009: s. 15:1–15:58, ISSN 0360-0300, doi:10.1145/1541880.1541882.
- [7] Cody, C.; Ford, V.; Siraj, A.: Decision Tree Learning for Fraud Detection in Consumer Energy Consumption. *the 14th IEEE International Conference on Machine Learning and Applications*, 2015, doi:10.1109/ICMLA.2015.80.
- [8] Cohen, R. A.: An introduction to PROC LOESS for local regression. 01 1999.
- [9] Coma-Puig, B.; Carmona, J.; Gavalda, R.; aj.: Fraud detection in energy consumption: A supervised approach. *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics*, DSAA 2016, 2016: s. 120–129, doi:10.1109/DSAA.2016.19.
- [10] Craw, S.: *Manhattan Distance*, kapitola Manhattan Distance. Boston, MA: Springer US, 2017, ISBN 978-1-4899-7687-1, s. 790–791, doi:10.1007/978-1-4899-7687-1_511.
- [11] Cuturi, M.: Fast Global Alignment Kernels. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, USA: Omnipress, 2011, ISBN 978-1-4503-0619-5, s. 929–936.
- [12] Depuru, S. S. S. R.: *Modeling, detection, and prevention of electricity theft for enhanced performance and security of power grid*. Dizertačná práca, The University of Toledo, 2012.
- [13] Dzeroski, S.; Gjorgjioski, V.; Slakov, I.; aj.: Analysis of time series data with predictive clustering trees. *Knowledge Discovery in Inductive Databases*, 2007: s. 47–58, ISSN 03029743, doi:10.1007/978-3-540-75549-4_5.
- [14] Fu, T. C.: A review on time series data mining. *Engineering Applications of Artificial Intelligence*, ročník 24, č. 1, 2011: s. 164–181, ISSN 09521976, doi:10.1016/j.engappai.2010.09.007.

- [15] Goldberger, A.; Amaral, L.; Glass, L.; aj.: PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, ročník 101, 07 2000: s. E215–20, doi:10.1161/01.CIR.101.23.e215.
- [16] Grmanová, G.; Laurinec, P.; Rozinajová, V.; aj.: Incremental Ensemble Learning for Electricity Load Forecasting. *Acta Polytechnica Hungarica*, ročník 13, č. 2, 2016.
- [17] Hautamaki, V.; Nykanen, P.; Franti, P.: Time-series clustering by approximate prototypes. In *2008 19th International Conference on Pattern Recognition*, Dec 2008, ISSN 1051-4651, s. 1–4, doi:10.1109/ICPR.2008.4761105.
- [18] Hochenbaum, J.; Vallis, O. S.; Kejariwal, A.: Automatic Anomaly Detection in the Cloud Via Statistical Learning. *CoRR*, ročník abs/1704.07706, 2017, 1704 . 07706.
- [19] Hsu, C.-J.; Huang, K.-S.; Yang, C.-B.; aj.: Flexible Dynamic Time Warping for Time Series Classification. *Procedia Computer Science*, ročník 51, 12 2015: s. 2838–2842, doi:10.1016/j.procs.2015.05.444.
- [20] Jiang, R.; Tagaris, H.; Lachsz, A.; aj.: Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In *IEEE/PES Transmission and Distribution Conference and Exhibition*, ročník 3, Oct 2002, s. 2251–2256 vol.3, doi:10.1109/TDC.2002.1177814.
- [21] Kejariwal, A.; Tsiamis, J.; Wong, S.: Introducing practical and robust anomaly detection in a time series. URL: https://blog.twitter.com/engineering/en_us/a/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series.html, 1 2015.
- [22] Kohonen, T.; Schroeder, M. R.; Huang, T. S. (editori): *Self-Organizing Maps*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., tretie vydanie, 2001, ISBN 3540679219.
- [23] Kuppusamy, M.; Kaliyaperumal, S.: Comparison of Methods for detecting Outliers. *International Journal of Scientific & Engineering Research*, ročník 4, 01 2013: s. 709–714.
- [24] Laurinec, P.; Lucka, M.: *Improving Forecasting Accuracy through the Influence of Time Series Representations and Clustering*. Dizertačná práca, Slovak University of Technology in Bratislava, Ilkovičova 2, 842 16 Bratislava, Slovakia, 5 2018.
- [25] Laurinec, P.; Lucka, M.: Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting. *Data Mining and Knowledge Discovery*, 11 2018, doi:10.1007/s10618-018-0598-2.
- [26] Malinowski, S.; Team, L. R.: Recent advances in Time Series Classification. URL: <http://www.antoniomucherino.it/events/CDs/CD03/TimeSeriesClassification.pdf>, 6 2017.
- [27] Meffe, A.; de Oliveira, C. C. B.: Technical loss calculation by distribution system segment with corrections from measurements. In *CIRED 2009 - 20th International Conference and Exhibition on Electricity Distribution - Part 1*, June 2009, ISSN 0537-9989, s. 1–4, doi:10.1049/cp.2009.0962.

- [28] Nagi, J.; Yap, K. S.; Tiong, S. K.; aj.: Detection of Abnormalities and Electricity Theft using Genetic Support Vector Machines. In *TENCON 2008 - 2008 IEEE Region 10 Conference*, 12 2008, ISSN 2159-3442, s. 1–6, doi:10.1109/TENCON.2008.4766403.
- [29] Nikovski, D. N.; Wang, Z.; Esenther, A.; aj.: Smart Meter Data Analysis for Power Theft Detection. *Machine Learning and Data Mining in Pattern Recognition*, 2013: s. 379–389, ISSN 03029743, doi:10.1007/978-3-642-39712-7·29.
- [30] Paparrizos, J.; Gravano, L.: k-Shape: Efficient and Accurate Clustering of Time Series. *SIGMOD Rec.*, ročník 45, č. 1, jun 2016: s. 69–76, ISSN 0163-5808, doi:10.1145/2949741.2949758.
- [31] Perea, J. A.; Deckard, A.; Haase, S. B.; aj.: SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC Bioinformatics*, ročník 16, č. 1, Aug 2015: str. 257, ISSN 1471-2105, doi:10.1186/s12859-015-0645-6.
- [32] Rani, S.; Sikka, G.: Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications*, ročník 52, č. 15, 2012: s. 1–9, ISSN 09758887, doi:10.5120/8282-1278.
- [33] Rosner, B.: Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, ročník 25, č. 2, 1983: s. 165–172, doi:10.1080/00401706.1983.10487848.
- [34] Sahoo, S.; Nikovski, D.; Muso, T.; aj.: Electricity theft detection using smart meter data. *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2015: s. 1–5, doi:10.1109/ISGT.2015.7131776.
- [35] Salvador, S.; Chan, P.: Learning states and rules for detecting anomalies in time series. *Applied Intelligence*, ročník 23, č. 3, 2005: s. 241–255, ISSN 0924669X, doi:10.1007/s10489-005-4610-3.
- [36] Sapankevych, N. I.; Sankar, R.: Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine*, ročník 4, č. 2, May 2009: s. 24–38, ISSN 1556-603X, doi:10.1109/MCI.2009.932254.
- [37] Smith, L. I.: A tutorial on principal components analysis. Technická správa, Cornell University, USA, February 26 2002.
- [38] Song, X.; Wu, M.; Jermaine, C. M.; aj.: Conditional Anomaly Detection. *IEEE Trans. Knowl. Data Eng.*, ročník 19, č. 5, 2007: s. 631–645.
- [39] Spirić, J. V.; Dočić, M. B.; Stanković, S. S.: Fraud detection in registered electricity time series. *International Journal of Electrical Power and Energy Systems*, ročník 71, 2015: s. 42–50, ISSN 01420615, doi:10.1016/j.ijepes.2015.02.037.
- [40] Stankovic, S. S.; Doc, M. B.; Popovic, T. D.; aj.: Using the rough set theory to detect fraud committed by electricity customers. *International Journal of Electrical Power & Energy Systems*, ročník 62, 2014: s. 727–734, ISSN 0142-0615, doi:10.1016/j.ijepes.2014.05.004.
- [41] Tan, P.-N.; Steinbach, M.; Kumar, V.: *Introduction to Data Mining*. Addison Wesley, us ed vydanie, May 2005, ISBN 0321321367.

- [42] Teng, M.: Anomaly detection on time series. In *2010 IEEE International Conference on Progress in Informatics and Computing*, ročník 1, Dec 2010, s. 603–608, doi: 10.1109/PIC.2010.5687485.
- [43] Trevizan, R. D.; Bretas, A. S.; Rossoni, A.: Nontechnical Losses detection: A Discrete Cosine Transform and Optimum-Path Forest based approach. *2015 North American Power Symposium, NAPS 2015*, October 2015, doi:10.1109/NAPS.2015.7335160.
- [44] Vallis, O.; Hochenbaum, J.; Kejariwal, A.: A Novel Technique for Long-term Anomaly Detection in the Cloud. In *Proceedings of the 6th USENIX Conference on Hot Topics in Cloud Computing*, HotCloud14, Berkeley, CA, USA: USENIX Association, 2014, s. 15–15.
- [45] Vieira, R. G.; Filho, M. A. L.; Semolini, R.: An Enhanced Seasonal-Hybrid ESD Technique for Robust Anomaly Detection on Time Series. *Proceedings of the Brazilian Symposium on Computer Networks and Distributed Systems (SBRC)*, 2018.
URL <http://portaldeconteudo.sbc.org.br/index.php/sbrc/article/view/2422>
- [46] Warren Liao, T.: Clustering of time series data - A survey. *Pattern Recognition*, ročník 38, č. 11, 2005: s. 1857–1874, ISSN 00313203, doi:10.1016/j.patcog.2005.01.025.
- [47] Wei, L.; Keogh, E.: Semi-supervised time series classification. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, 2006: str. 748, ISSN 01651684, doi:10.1145/1150402.1150498.
- [48] Xiong, Y.; Yeung, D.-Y.: Mixtures of ARMA models for model-based time series clustering. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 2002, s. 717–720, doi:10.1109/ICDM.2002.1184037.

A Technická dokumentácia

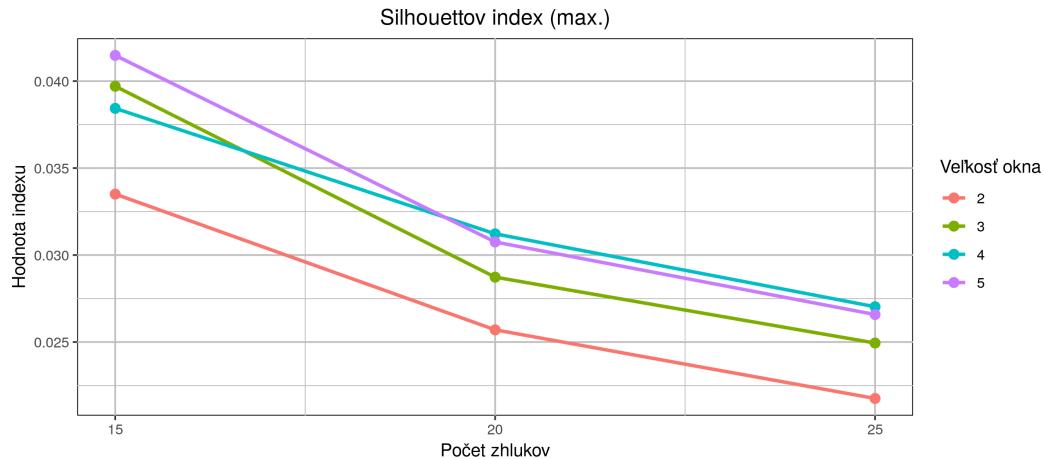
Použité knižnice s verziami sú zobrazené pomocou tabuľky 7.

Tabuľka 7: Použité knižnice jazyka R.

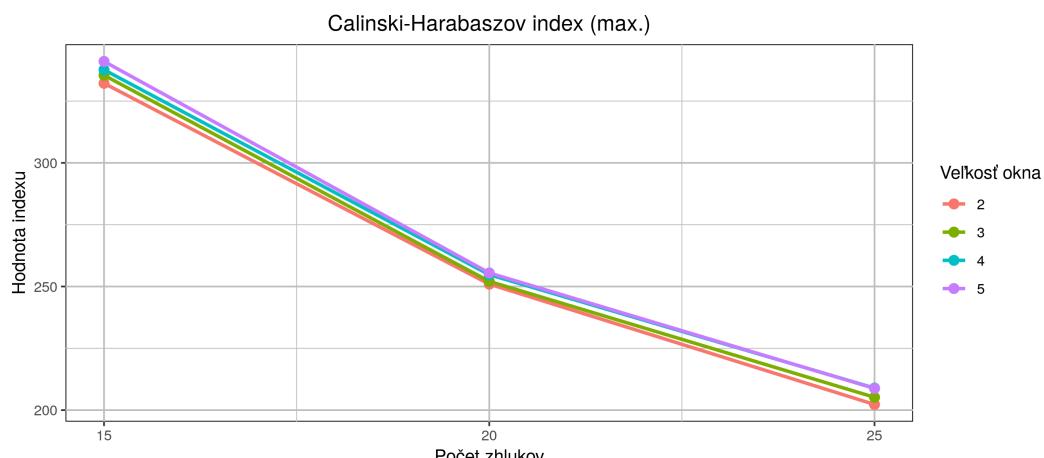
Názov	Použitá verzia
AnomalyDetection	1.0
BreakoutDetection	1.0.1
cluster	2.0.8
clusterCrit	1.2.8
data.table	1.12.2
devtools	2.0.2
dplyr	0.8.0.1
dtw	1.20-1
dtwclust	5.5.2
dygraphs	1.1.1.6
ggbiplot	0.55
ggplot2	3.1.1
ggpubr	0.2
gridExtra	2.3
gttable	0.3.0
lubridate	1.7.4
pkgmaker	0.27
plotly	4.8.0
proxy	0.4-22
registry	0.5
rngtools	1.3.1
Rtsne	0.15
scales	1.0.0
stringr	1.4.0
TSrepr	1.0.2
tidyquant	0.5.6
zoo	1.8-5

B Vizualizácie experimentov pre výber hyperparametrov

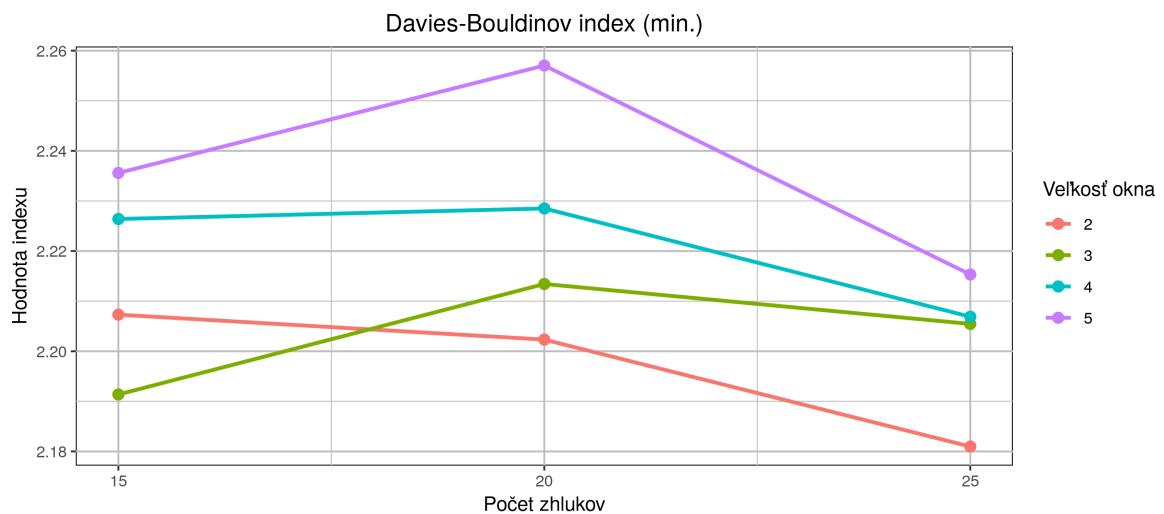
Každý index obsahuje aj informáciu o jeho optimálnych hodnotách. Pri indexoch, ktoré obsahujú (*max.*) znamenajú väčšie hodnoty lepšie výsledné zhľukovanie. Pri indexoch s (*min.*) sú za lepšie výsledky zhľukovania považované nižšie hodnoty. Indexy, ktorých výsledky boli veľmi podobné sme pri rozhodovaní vyniechali, používali sme najmä Silhouetteov a Davies-Bouldinov index, ktoré najlepšie opisujú kvalitu zhľukovania.



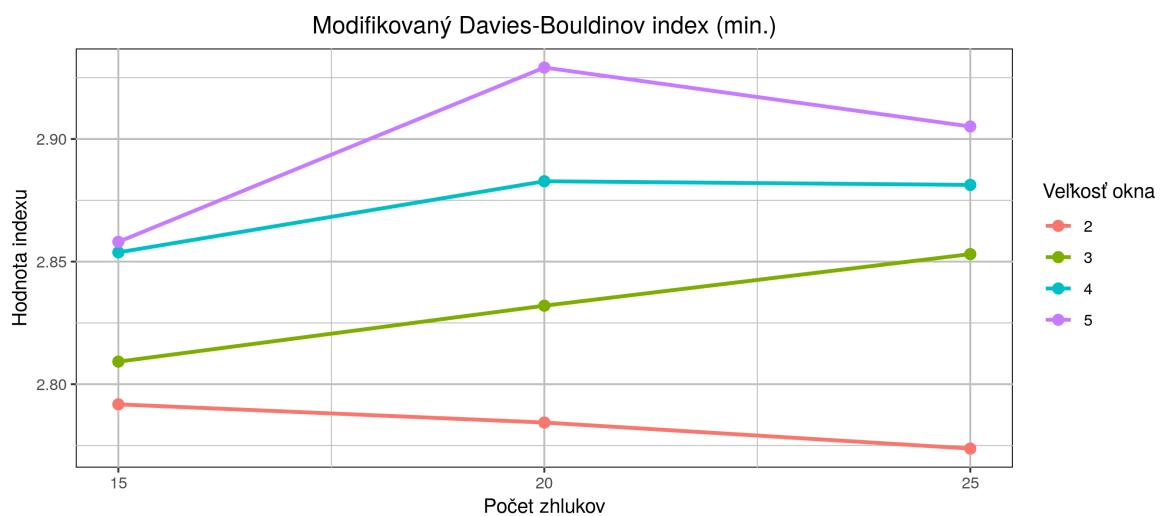
Obr. 41: Porovnanie veľkosti posuvného okna a počtu zhlukov (Silhouetteov index).



Obr. 42: Porovnanie veľkosti posuvného okna a počtu zhlukov (Calinski-Harabaszov index).



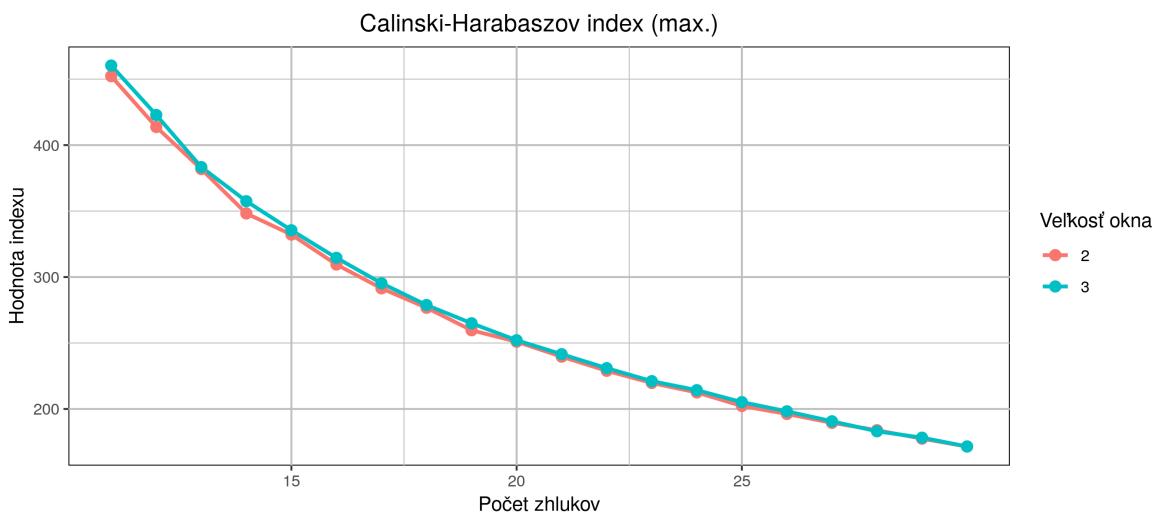
Obr. 43: Porovnanie veľkosti posuvného okna a počtu zhlukov (Davies-Bouldinov index).



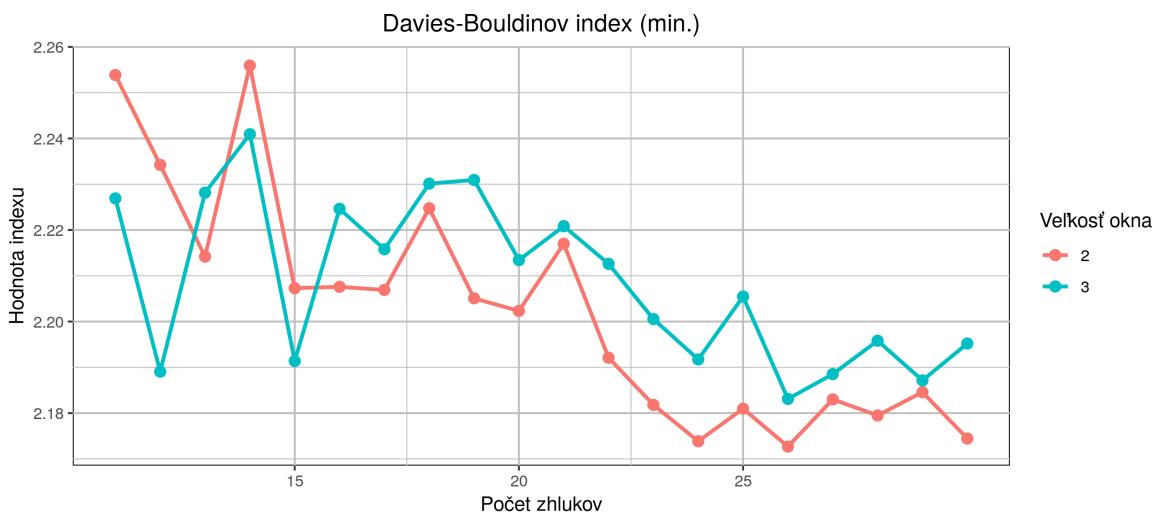
Obr. 44: Porovnanie veľkosti posuvného okna a počtu zhlukov (Modifikovaný Davies-Bouldinov index).



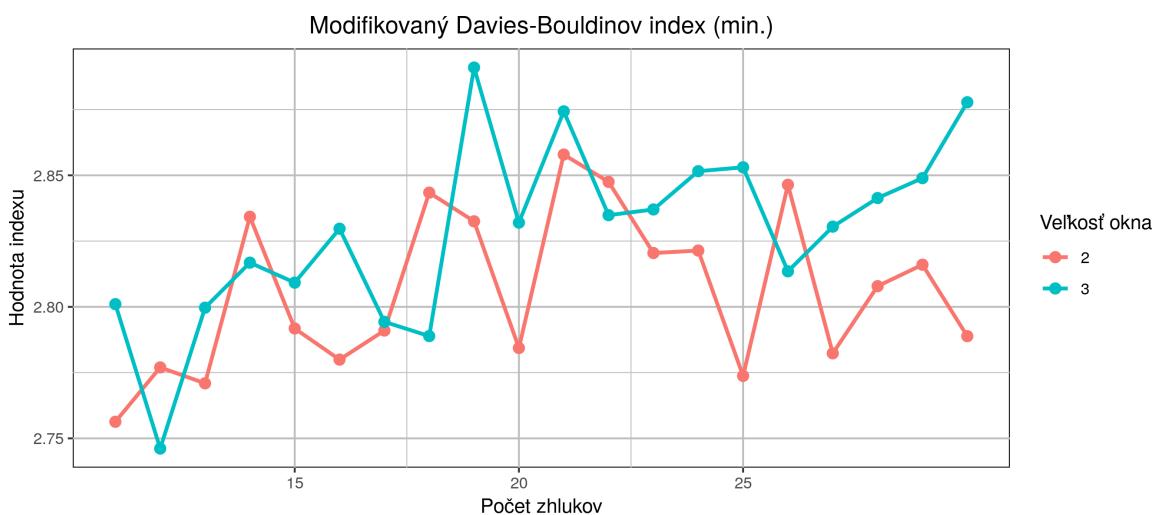
Obr. 45: Porovnanie veľkosti posuvného okna a počtu zhlukov (Silhouetteov index).



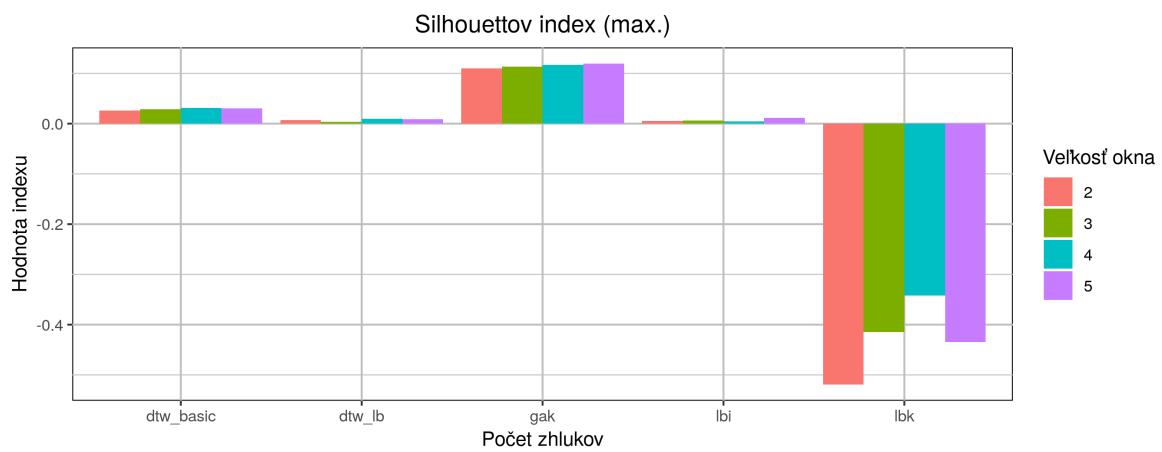
Obr. 46: Porovnanie veľkosti posuvného okna a počtu zhlukov (Calinski-Harabaszov index).



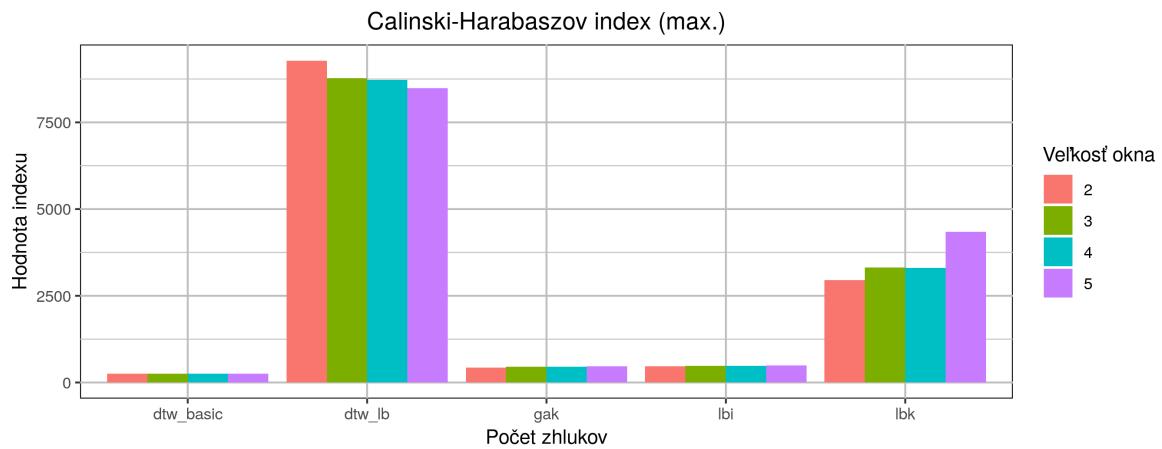
Obr. 47: Porovnanie veľkosti posuvného okna a počtu zhlukov (Davies-Bouldinov index).



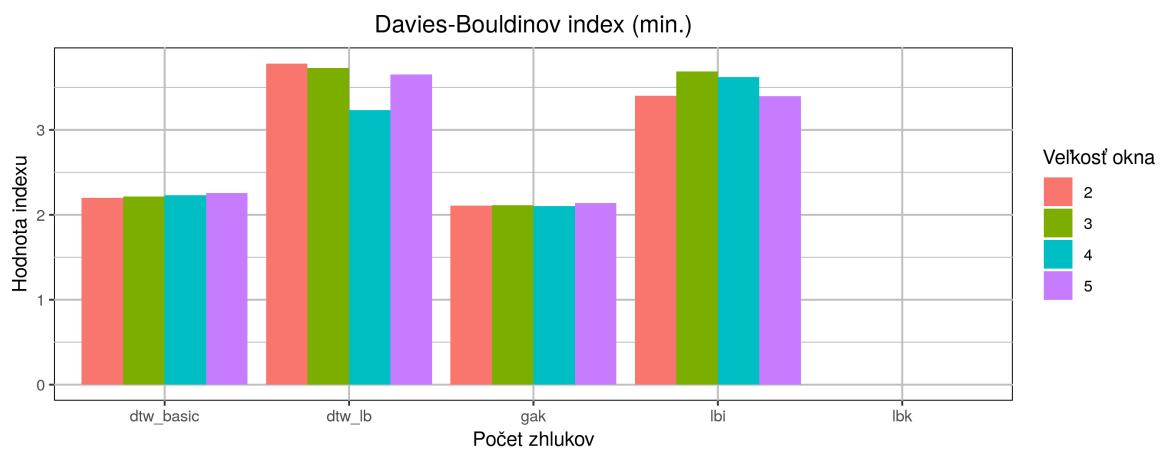
Obr. 48: Porovnanie veľkosti posuvného okna a počtu zhlukov (Modifikovaný Davies-Bouldinov index).



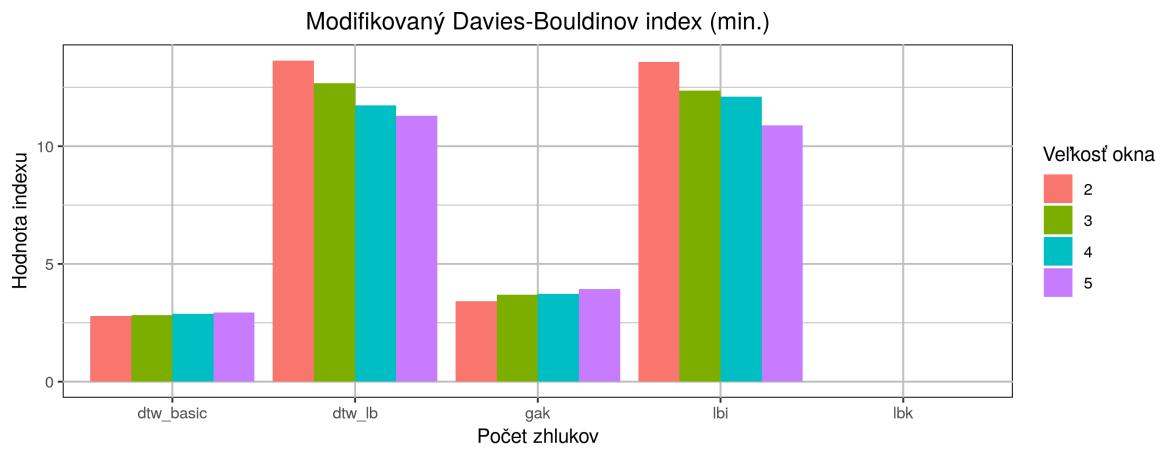
Obr. 49: Porovnanie vzdialenosných metrík (Silhouetteov index).



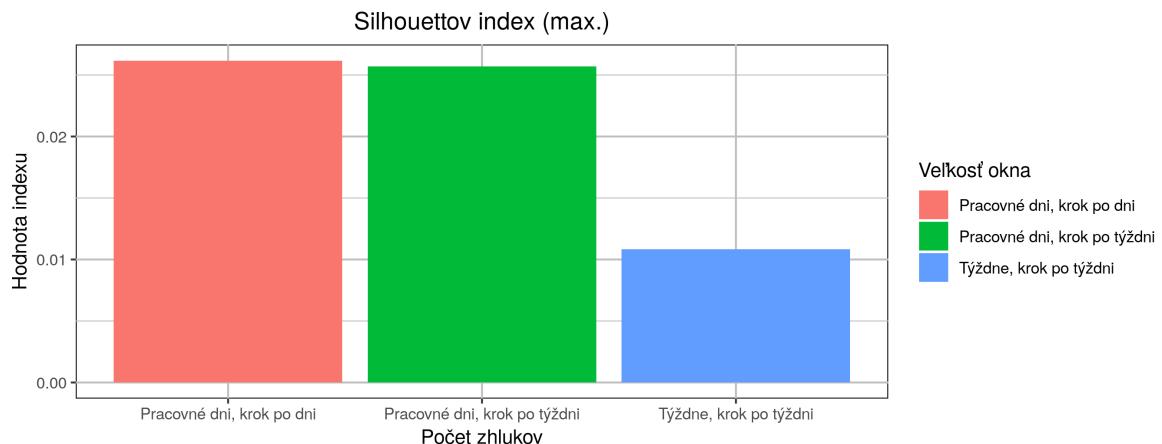
Obr. 50: Porovnanie vzdialenosných metrík (Calinski-Harabaszov index).



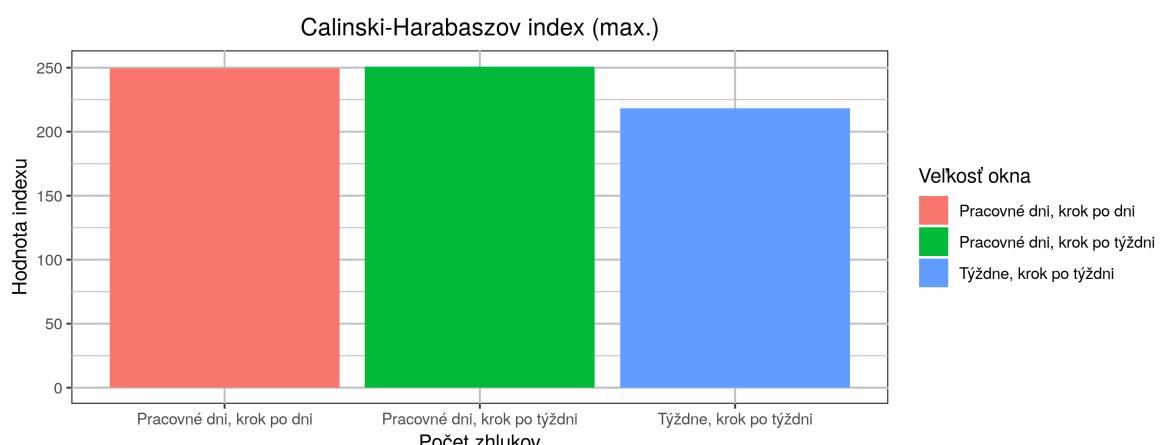
Obr. 51: Porovnanie vzdialenosných metrík (Davies-Bouldinov index).



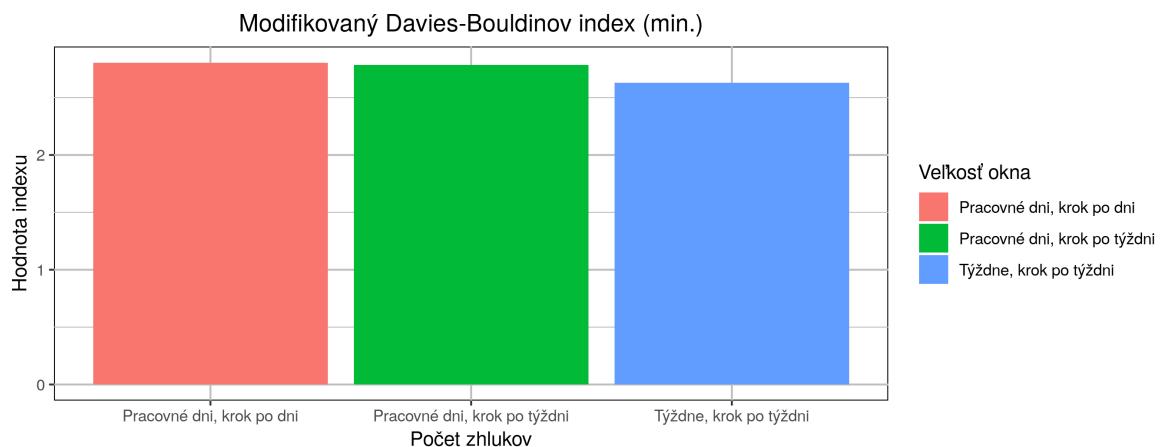
Obr. 52: Porovnanie vzdialenosných metrík (Modifikovaný Davies-Bouldinov index).



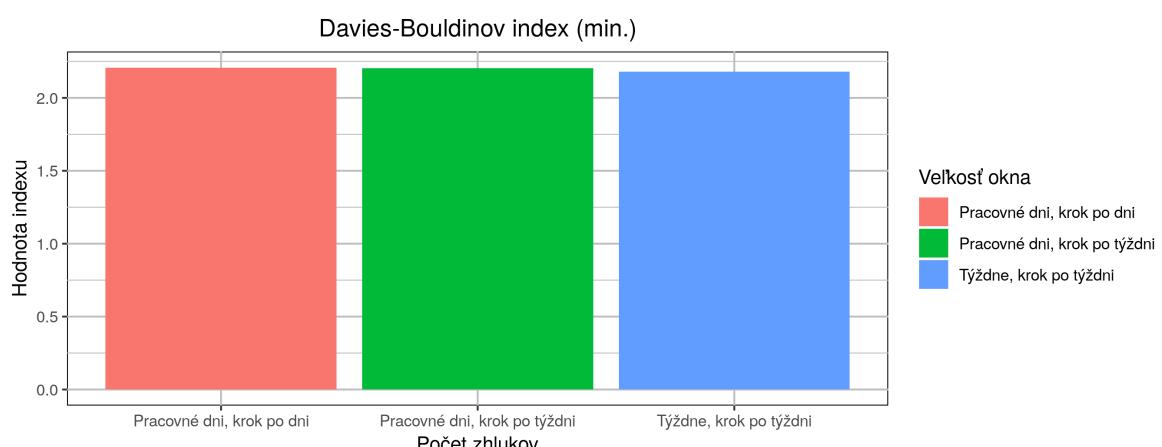
Obr. 53: Porovnanie veľkostí a typov posuvných okien (Silhouetteov index).



Obr. 54: Porovnanie veľkostí a typov posuvných okien (Calinski-Harabaszov index).



Obr. 55: Porovnanie veľkostí a typov posuvných okien (Davies-Bouldinov index).



Obr. 56: Porovnanie veľkostí a typov posuvných okien (Modifikovaný Davies-Bouldinov index).

C Plán práce na riešení projektu

C.1 Plán do zimného semestra

Poradie týždňa v zimnom semestri	Popis plánovanej činnosti
1. týždeň	dokončenie analýzy literatúry, špecifikácie, výber vhodných metód na experimenty
2. týždeň	nájdenie viacerých vhodných zdrojov dát, analýza datasetov
3. týždeň	analýza datasetov, doplnenie datasetov dátami z iných zdrojov
4. týždeň	výber atribútov, spojenie dát, príprava prostredia pre dátu a manipuláciu s nimi
5. týždeň	hlbšia analýza zvolených datasetov, agregácia časových radov, výber technológií
6. týždeň	vyhľadenie a diskretizácia časových radov
7. týždeň	tvorba prototypu modelu zhlukujúceho časové rady
8. týždeň	prvé zhlukovacie experimenty, skúmanie vytvorených zhlukov
9. týždeň	tvorba identifikátora pre anomálne časové rady nachádzajúce sa mimo početných zhlukov
10. týždeň	tvorba prototypu modelu identifikujúceho anomálie v rámci zhlukov
11. týždeň	prvé experimenty, skúmanie identifikovaných anomálií
12. týždeň	tvorba prezentácie a príprava na obhajoby

Počas zimného semestra budú analyzované dostupné datasety. Následne budú spracované a upravené do formátu potrebného pre model. Po vytvorení prototypu modelu budú čo najskôr evaluované výsledky. Na základe nich budú ďalej spracovávané dátá a menené datasety. Taktiež sa bližšie pozrieme na identifikované anomálie dát a určíme ich pôvod na základe dostupných vysvetľujúcich premenných. V závere semestra budú prebiehať prípravy na obhajoby tvorbou prezentácie a súvisiacich materiálov.

Použitý dataset bol analyzovaný a transformovaný podľa potrieb ďalších experimentov v súlade s časovým plánom. Úspešne sme vykonali a vyhodnotili experimenty súvisiace s použitím zhlukovaním. Nemali sme však k dispozícii žiadne ďalšie testované datasety a preto sme obmedzili rozsah experimentov iba na spomínaný dataset. Prototyp hodnotiaci anomálnosť časových radov jednotlivých odberateľov bol úspešne vytvorený. Príprava na obhajobu práce si vyžiadala väčšie úsilie vzhľadom na zoznamovanie sa s prostredím pre vizualizáciu výsledkov. V rámci výskumného semináru prebehla aj skúšobné prezentovanie výsledkov kolegom.

C.2 Plán do letného semestra

Poradie týždňa v letnom semestri	Popis plánovanej činnosti
1. týždeň	vytvorenie validačného modelu pre zhlukovanie časových radov
2. týždeň	výber hyperparametrov a metrík vzdialenosť pre použité zhlukovacie algoritmy
3. týždeň	použitie rôznych diskretizačných okien a validácia použitých metrík
4. týždeň	použitie identifikátora pre anomálne časové rady nachádzajúce sa mimo početných zhlukov
5. týždeň	porovnanie výsledkov z experimentov s naštudovanou literatúrou, overenie výsledkov
6. týždeň	optimalizácia modelu a výpočtov, vyhodnotenie experimentov
7. týždeň	vytvorenie článku na študentskú konferenciu, dokončovanie experimentov
8. týždeň	príprava na študentskú vedeckú konferenciu, vizualizácia dosiahnutých výsledkov
9. týždeň	zapracovanie prípomienok z konferencie, návrh ďalších experimentov
10. týždeň	písanie technickej dokumentácie k softvérovému dielu, posledné experimentovanie
11. týždeň	vyhodnotenie experimentov, evaluácia riešenia, zhodnotenie výsledkov
12. týždeň	tvorba prezentácie a príprava na obhajobu projektu

V úvode semestra bude vytvorený validačný mechanizmus pre model, zhlukujúci analyzované časové rady na základe podobnosti priebehov. Cieľom bude overiť vytvorené zhluky, ktoré vznikli na základe profilov spotreby elektrickej energie. Dôležitý je aj výber hyperparametrov pre algoritmi strojového učenia, ale aj vhodná dĺžka posuvného okna, prípadne metrika použitá na meranie podobnosti jednotlivých časových radov. Z analyzovaných riešení je potrebné vybrať najvhodnejšie pre náš problém. Ďalšie navrhované vylepšenia budú zapracované na základe vlastností dát a dosiahnutých výsledkov. V prípade veľkej časovej náročnosti budú optimalizované jednotlivé procesy pri manipulácii dát. Na konci semestra sa koná študentská konferencia, do ktorej by sme chceli prispieť článkom a získať tak dôležitú spätnú väzbu. Na záver by sme chceli zapracovať jednotlivé prípomienky a znova vyhodnotiť výsledky, ktoré budú použité pri obhajobe projektu.

Navrhnuté zhlukovanie bolo sériou experimentov vylepšené. Taktiež boli optimalizované procesy s vysokou časovou náročnosťou. Na základe vytvoreného zhlukovania bolo navrhnuté a odskúšané skóre, ktorým je určená miera anomálnosti daného odberateľa. Výsledky boli porovnávané s ďalšími existujúcimi riešeniami. Naše skóre sme skombinovali s existujúcimi metódami kvôli spresneniu intervalov. Výsledky sme overili ich vizualizovaním a porovnaním.

D Opis digitálnej časti práce

Evidenčné číslo práce v informačnom systéme: FIIT-182905-73688

Obsah digitálnej časti práce (archív ZIP):

```
archív ZIP
├── doc
│   ├── img
│   ├── iso-690
│   ├── DP_MATUS_CUPER.pdf
│   ├── dp.tex
│   └── bibliography.bib
└── src
    ├── graphs
    ├── server
    ├── agg-functions.R
    ├── dependencies.R
    ├── filter-functions.R
    ├── load-functions.R
    ├── score-and-visualize-ad.R
    ├── score-and-visualize-own.R
    ├── utilities.R
    ├── utilities.sh
    └── visualize-functions.R
```

- **doc** adresár obsahujúci dokumentáciu k diplomovej práci

- **img** adresár s obrázkami použitými v práci
- **iso-690** adresár s formátovaním pre BibTeX
- **DP_MATUS_CUPER.pdf** PDF dokument diplomovej práce
- **dp.tex** zdrojový súbor dokumentácie v LaTeXu
- **bibliography.bib** zdrojový súbor bibliografických odkazov v BibTeXu

- **src** skripty, vizualizácie a časti zdrojových kódov v jazyku R použité pri experimentoch

- **graphs** adresár obsahujúci skripty vykreslujúce grafy, ktoré sú použité v práci
- **server** adresár obsahuujci skripty, ktoré boli spúšťané na výpočtovom servery
- **agg-functions.R** funkcie používané na agregáciu dát
- **dependencies.R** skript obsahujúci všetky závislosti projektu
- **filter-functions.R** funkcie selektujúce dátá na základe kritérií
- **load-functions.R** skripty načítavajúce dátá
- **score-and-visualize-ad.R** skript vizualizujúci skórovanie metódy S-H-ESD
- **score-and-visualize-own.R** skript vizualizujúci skórovanie nami navrhutej metódy
- **utilities.R** nezaradené funkcie
- **utilities.sh** skripty použité pri transformácií vstupných dát
- **visualize-functions.R** skript s vizualizáciami použitými pri experimentoch

Názov odovzdaneho archívu: DP_prilohy_digital_cuper.zip