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, 9.2.8 Umelá inteligencia
Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU v Bratislave
Vedúci práce: Ing. Marek Lóderer

apríl 2019

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Inteligentné softvérové systémy
Autor: Bc. Matúš Cuper
Bakalárska práca: Identifikácia neštandardného správania odberateľov v energetickej sieti
Vedúci práce: Ing. Marek Lóderer
máj 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 identifikácií anomálií v doméne energetiky, a ktorým musí čeliť aj naša metóda. Bližšie sme sa zamerali na zhlukovanie časových radov, ktoré prináša nové prístupy do zhlukovania vysokodimenzionálnych dát, medzi ktoré patrí aj vyhladzovanie, redukcia dimenzií alebo selekcia atribútov. Navrhovaná metóda zlúči diskretizované vyhladené časové rady a následne sú identifikované anomálie na základe vytvorených zhlukov a rozloženia profilu používateľa v zhlukoch.

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Intelligent software systems
Author: Bc. Matúš Cuper
Bachelor thesis: Identification of abnormal behavior of customers in the power grid
Supervisor: Ing. Marek Lóderer
May 2019

In the thesis we focused on anomaly identification in energy time series. Anomalies can be caused by abnormal behavior of customers or 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 company. Also it is necessary to identify only cases, when the behavioral change is temporal, whether it is result of different number of residents, weather or an exceptional occasion. Nowadays, electricity load profile of customers is changing as the 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 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, which includes also smoothing, dimension reduction and attribute selection. Proposed method clusters discretized smoothed time series and then, based on created clusters and layout of customers profile in cluster, identifies anomalies.

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.

.....
Matúš Cuper

POĎAKOVANIE

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

Obsah

1 Úvod

Jedným z problémov, ktorým v súčasnosti čelia distribučné spoločnosti, je detekcia neštandardného 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. Oba prípady sú pre distribučnú spoločnosť nežiaduce a je v záujme zníženia strát, ich čo najskôr identifikovať. Obvykle sú za týmto účelom vykonávané náhodné kontroly, ktoré pokrývajú iba nízky počet zákazníkov s anomálnym správaním. Na základe množstva dát získavaných z inteligentných meračov je možné modelovať správanie zákazníkov. Distribučné spoločnosti tak môžu znižovať svoje straty a preverovať iba odberateľov, ktorí svojím profilom nezapadajú medzi odberateľov so štandardným správaním.

2 Analýza problému

Tak ako je spomenuté v článku [24], straty v distribučných sieťach v niektorých krajinách tvoria až 30% z celkového objemu distribuovanej energie. Väčšinu strát vytvára svojimi vlastnosťami samotná sieť, no nezanedbateľnú 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 dáta 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

Merania časových radov 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 meraní jednej veličiny sa nazývajú jednorozmerné, pri meraní viacerých veličín sa jedná o viacrozmerné časové rady. Tiež ich môžeme rozdeliť na spojité a diskrétne. 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. Diskrétne časové rady sú pozorované spravidla v rovnakých časových intervaloch, napr. rokoch, dňoch či minútach. Stretnúť sa s nimi môžeme pri kurzoch mien, produkcii štátov či spotrebe elektrickej energie [1].

2.1.1 Analýza časových radov

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 pokiaľ model nedosahuje dostatočne uspokojivé výsledky [1].

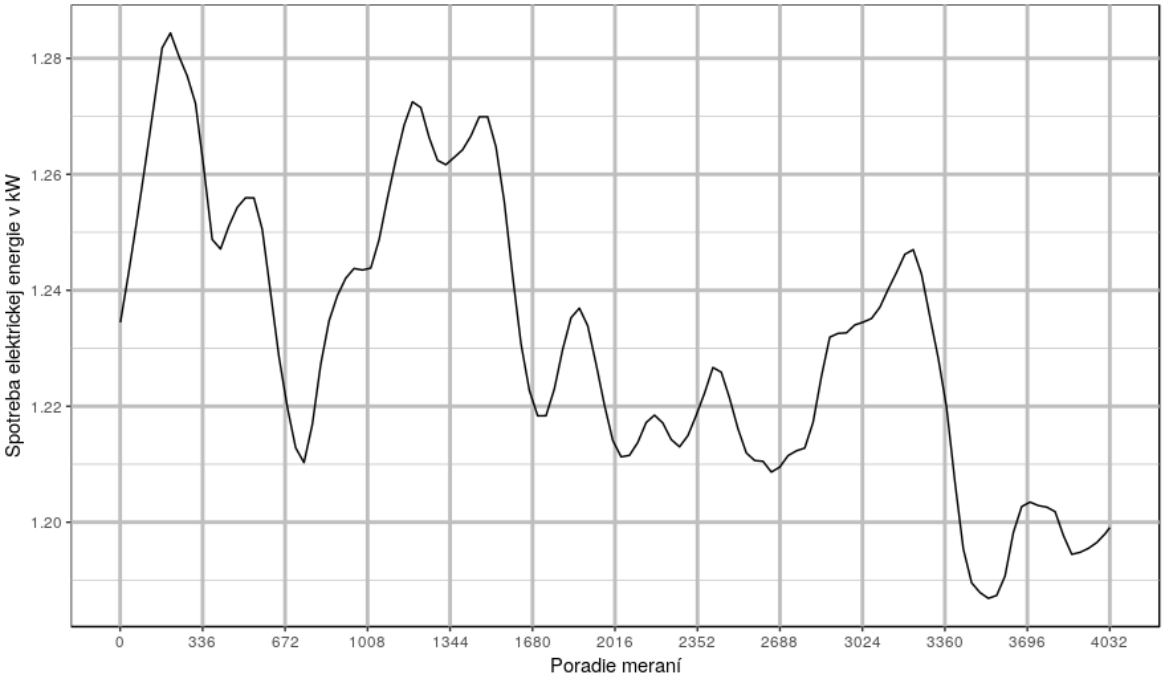
Premenná $\hat{x}$ vo vzorci 1 predstavuje predikovanú hodnotu časového radu x. Cieľom je nájsť funkciu $f(x)$, ktorá predikuje budúce hodnoty časového radu x tak, aby boli čo najpresnejšie, konzistentné a objektívne [33].

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

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, ekologickými, 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 periódy tohto cyklu môže byť rôzna, a to niekoľko 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 [15].

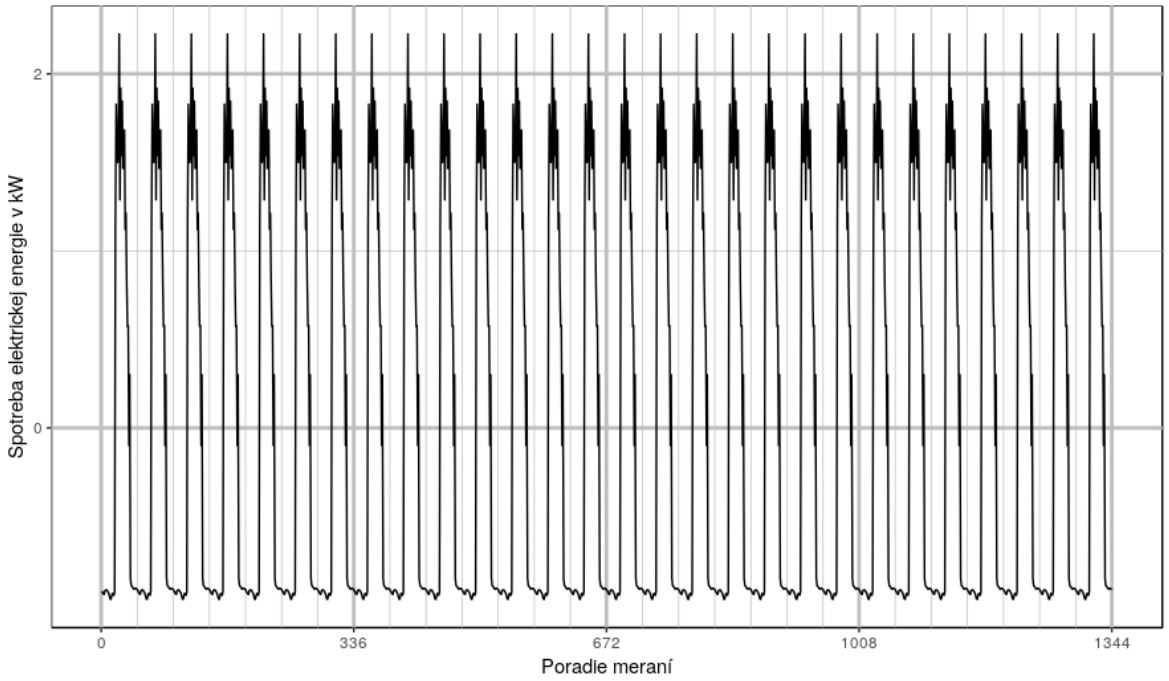
Trendová zložka zastupuje dlhodobé správanie časového radu. Ide o dlhodobé klesanie, rast alebo stagnáciu časového radu. Príkladom môže byť neustále predlžovanie priemernej doby dožitia alebo aj rast svetovej populácie. Priebeh dekompovanej trendovej zložky môžeme vidieť na obrázku 1 [1].



Obr. 1: Príklad trendovej zložky časového radu.

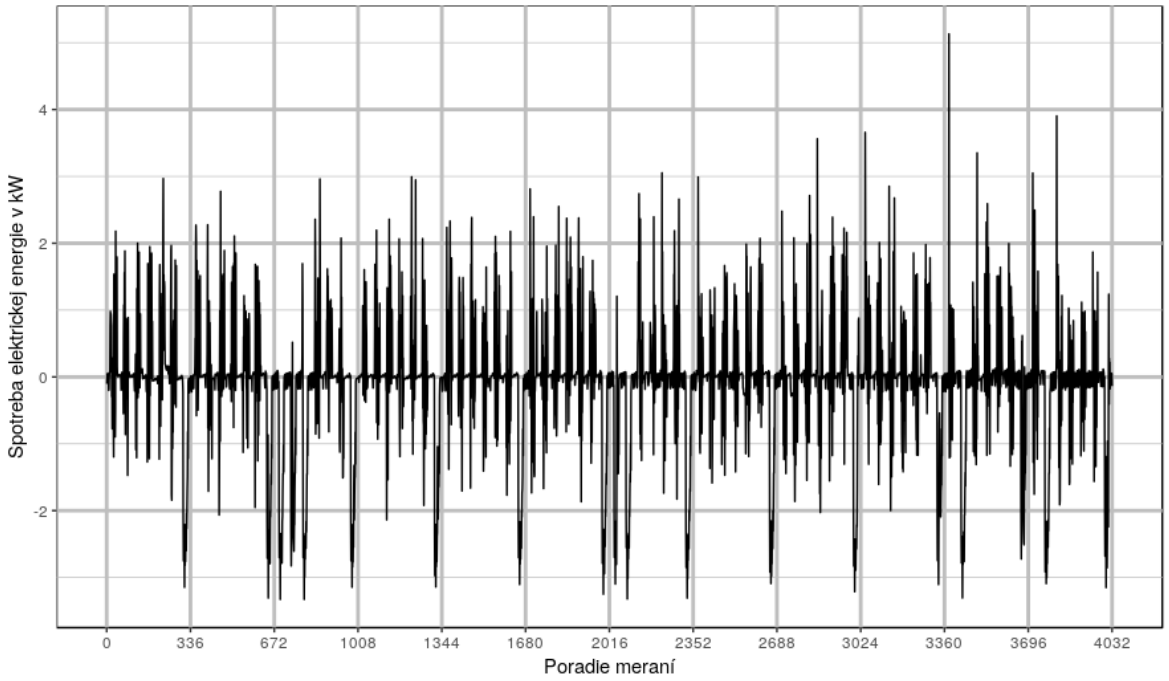
Cyklická zložka predstavuje strednodobú opakujúcu sa zmenu. Najčastejšie sa pri tom jedná o obdobie 2 a viac rokov. Táto zložka býva výrazne zastúpená pri ekonomických a finančných časových radoch. Príkladom môže byť aj podnikateľský cyklus, ktorý pozostáva zo 4 opakujúcich sa fáz [1].

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].



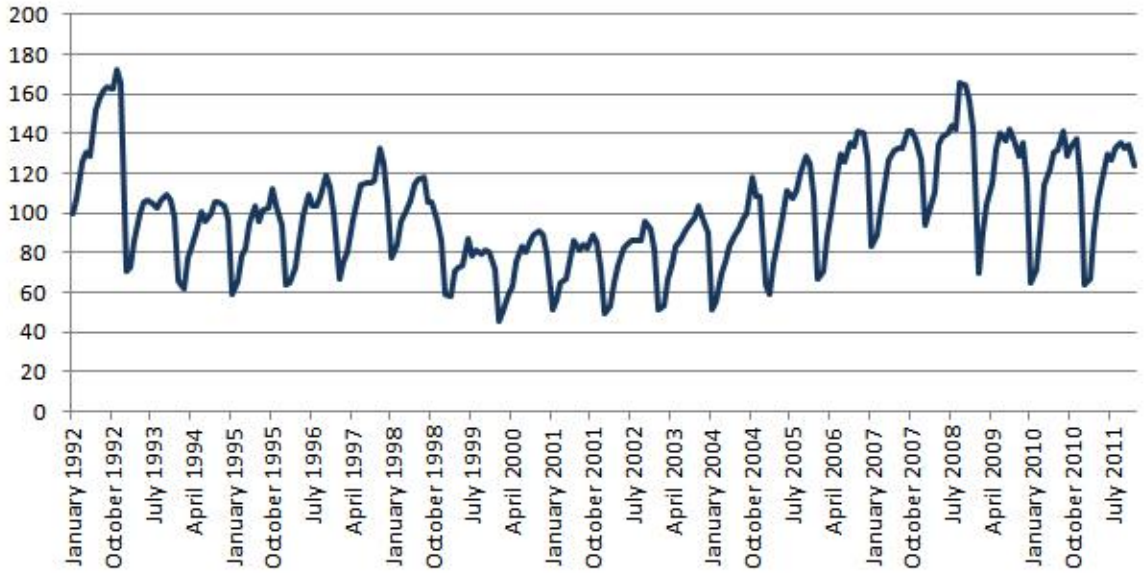
Obr. 3: Príklad reziduálnej zložky časového radu.

2.1.3 Typy modelov časových radov

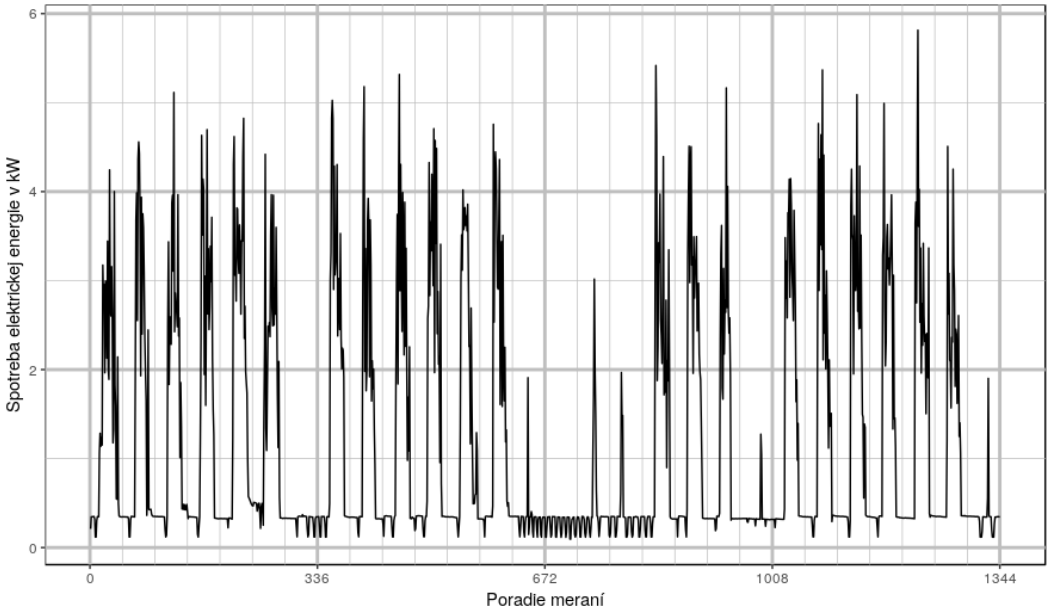
Kombináciou komponentov časových radov identifikovaných v predchádzajúcej kapitole vznikajú 2 typy modelov, aditívny a multiplikatívny.

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

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 modely predpokladáme ich nezávislosť. Multiplikatívny model je znázornený na obrázku 4 a aditívny na obrázku 5 [1].



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



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

2.1.4 Delenie časových radov

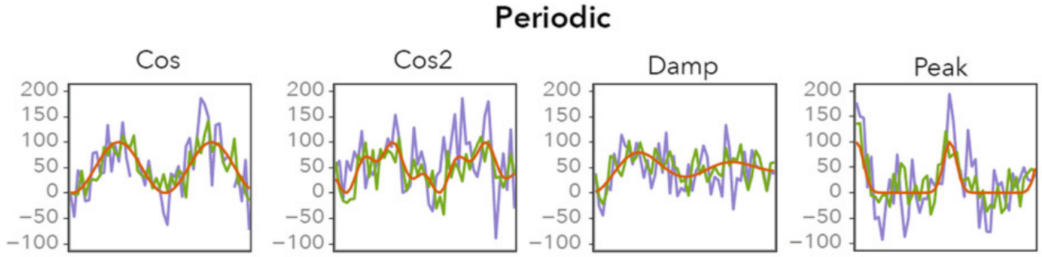
Výraznými vlastnosťami časových radov sú aj synchrónnosť a periodicita, znázornená na obrázkoch 6 a 7. Vznikajú tak 4 nasledujúce kategórie [39]:

- **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.

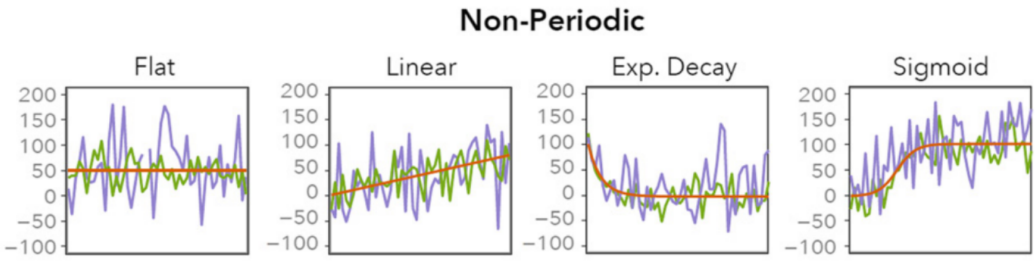
- **Neperiodické a synchrónne časové rady** nemajú žiadnu periodicitu, ale opäť sú časovo zarovnané.

- **Periodické a asynchrónne časové rady** nie sú časovo zarovnané, ale obsahujú periodicitu, čiže začiatok periódy v každom časovom rade je iný.

- **Neperiodické a asynchrónne časové rady** predstavujú skupinu, do ktorej spadajú ostatné časové rady, ktoré neobsahujú periodicitu a ani synchrónnosť.



Obr. 6: Príklad periodických časových radov [28].



Obr. 7: Príklad neperiodických časových radov [28].

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 [7]:

- **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 sťažuje 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 dáta neexistujú alebo ich je potrebné označiť, čo spravidla býva drahé.

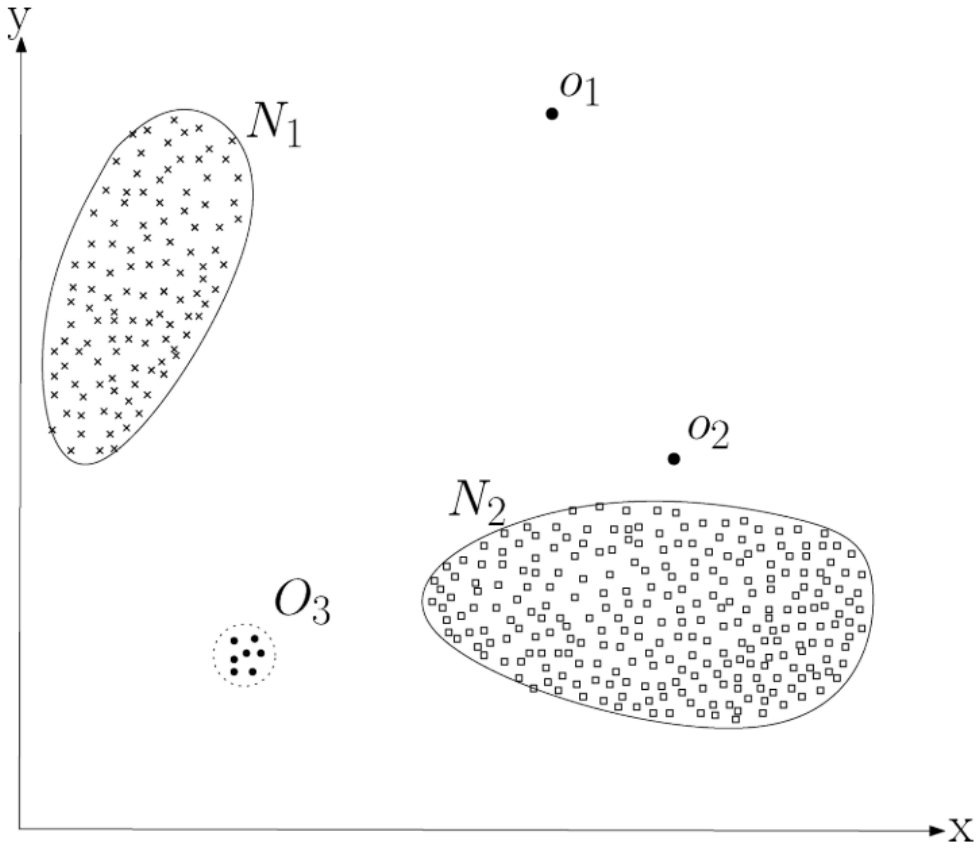
- **Biely šum** vyskytujúci sa v dátach má tendenciu skresľovať normálne dáta, ktorých identifikácia je následne zložitá.

Na detekciu anomálií sú používané aj 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 detekcie 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 dáta, ktoré nie sú anomáliou [7].



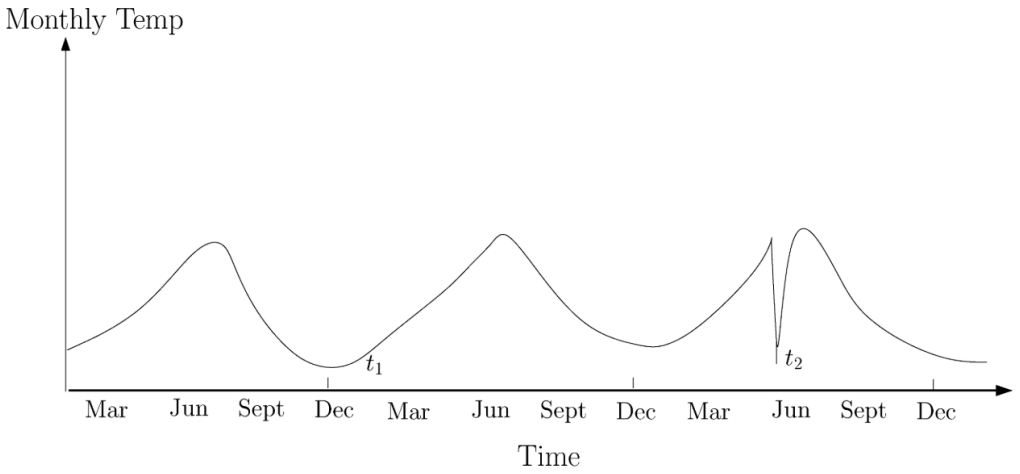
Obr. 8: Príklad bodových anomálií [7].

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 [7].

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 [7].

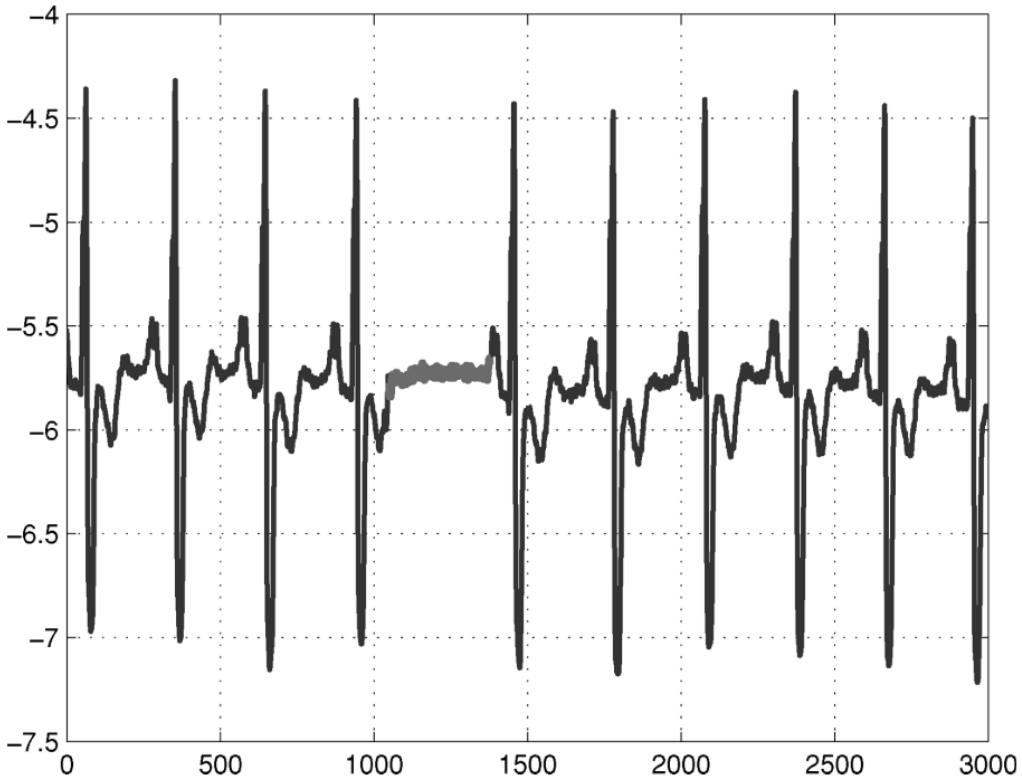
Zatiaľ čo v niektorých prípadoch je definovanie kontextu priamočiare, existujú domény, kde to jednoduché nie je. Dôležité je aby kontextové atribúty boli zmysluplne určené v cieľovej

doméne ich aplikácie [7].



Obr. 9: Príklad kontextových anomálií [7].

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 [7].



Obr. 10: Príklad skupinových anomálií [7].

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 detekcie bodových a skupinových anomálií je možné transformovať na problém detekcie 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 [7].

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 dispozícií obsahujú konečný počet meraní a teda nie sú spojité, anomália môže byť reprezentovaná aj jediným meraním. Anomálie môžme taktiež rozdeliť na pozitívne a negatívne, v závislosti od toho, aké sú očakávané a reálne hodnoty. Ak je ich rozdiel kladný budeme hovoriť o pozitívnych anomáliách, inak o negatívnych [4].

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 intervaly, ktorých sa môže vyskytovať rádovo viac. Robustné riešenie to musí zohľadniť a identifikovať iba signifikantné anomálie [4].

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 spätnej väzby môže byť zdĺhavý a nepraktický. Z toho dôvodu je dôležité zvoliť správny prístup pri identifikácií 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*) [7].

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 [7].

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 [7].

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 [7].

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. Popísané 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 [7, 38].

2.3.2 Analýza najbližšieho suseda

Metóda určí na základe vzdialenosti alebo podobnosti medzi dátovými inštanciami, či sa jedná o normálnu inštanciu alebo anomáliu. To je vypočítané pomocou vzdialeností medzi testovanou inštanciou a všetkými bodmi, alebo iba k najbližšími bodmi. Pri viacrozmerných 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 usporiadané, anomálie sú vzdialenejšie, prípadne na okraji vzniknutých oblastí. Aplikácia je možná pomocou techník založených na relatívnej hustote alebo vzdialenosti najbližších k susedných inštancií [7, 38].

2.3.3 Zhlukovanie

Jedná sa o učenie bez učiteľa, keďže zhluky inštancií sú vytvorené na základe ich vzdialenosti či podobnosti. Techniky ďalej delíme do kategórií na základe predpokladu o dátových inštanciách [7, 38].

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 detekcií anomálií, keďže sú primárne určené na riešenie zhlukovacích problémov [7].

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 vzdialenosti od centroidu daného zhluku. Používanými algoritmami sú neurónové siete (konkrétne 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 [7].

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-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 [32].

2.3.4 Štatistické metódy

Jedná sa o súbor metód založených na štatistike. 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 anomálie, 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 segmentácií je metóda úspešná iba v prípade, kedy sa odchýlka nepretržite mení [17].

Metóda kĺzavého priemeru má niekoľko modifikácií, na základe ovahánia jednotlivých meraní vzniká napr. metóda kĺzavého priemeru s exponencionálnym váhovaním (angl. *exponentially weighted moving average*), skrátene EWMA. Autori v práci [17] porovnávali okrem EWMA aj PEWMA (pravdepodobnostné exponencionálne váhovanie), kde v kombinácií s metódou ESD nedosiahli postačujúce výsledky. Kĺzavý priemer zlyhával pri identifikácií sezónnych anomálií.

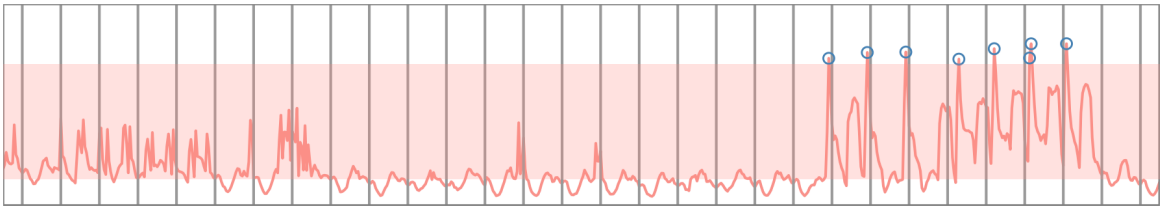
2.3.5 Extrémna Studentova odchýlka

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 zapísať jednoduchou rovnicou 3 definovanú pre i odobraných inštancií, ktorá je v štatistike často označovaná aj ako Grubbov test [21, 30].

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

Do rovnakej kategórie môžeme zaradiť aj sezónnu ESD (angl. *Seasonal Extreme Studentized Deviation*), ktorá rovnako využíva ESD na identifikáciu anomálií. Kľúčovým rozdielom je aplikovanie ESD až na dáta, 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 anomálií v reziduálnej zložke časového radu. Hlavnými obmedzením S-ESD sú datasety obsahujúce väčšií podiel anomálií. 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 [17].

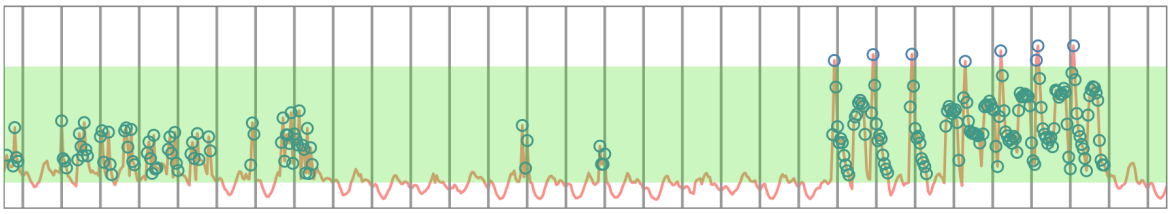


Obr. 11: Anomálie detegované pomocou S-ESD algoritmu [17].

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% anomálií v časovom rade. Jedná sa o absolútnu odchýlku mediánu MAD (angl. *Median Absolute Deviation*), ktorú vypočítame pomocou vzorca 4 [17].

$$MAD = median_i(|X_i - median_j(X_j)|) \tag{4}$$

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 [17].



Obr. 12: Anomálie detegované pomocou S-H-ESD algoritmu [17].

Ďalšou 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*), kedže klasická dekompozícia môže byť ovplyvnená výskytom anomálií vo vstupných dátach. Cieľom modifikovanej dekompozície je pomocou série vnorených 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 [41].

2.4 Metódy zhlukovania časových radov

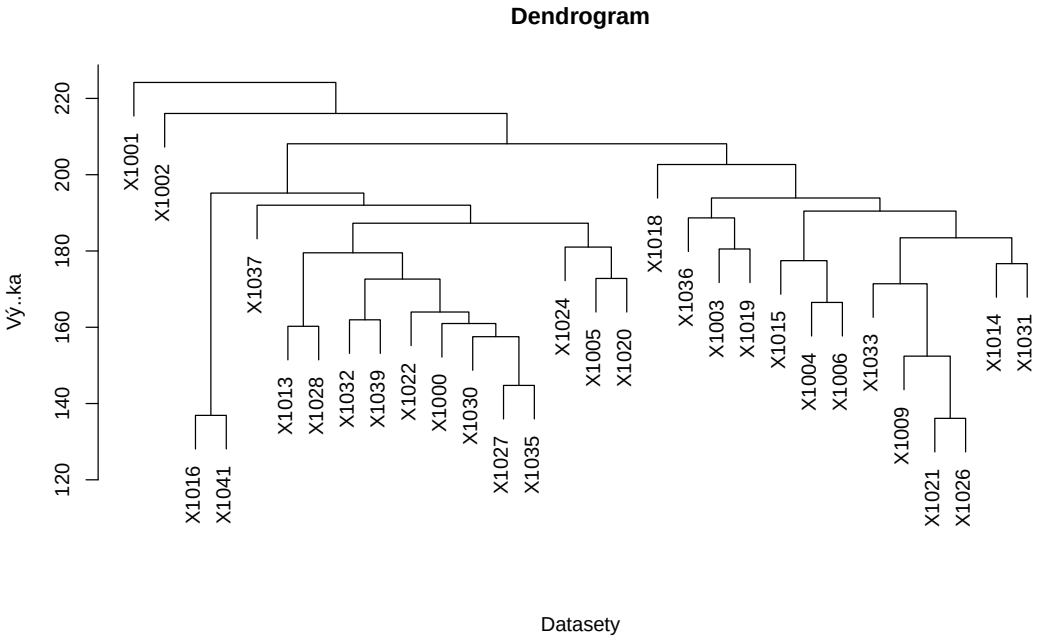
Cieľom zhlukovania je rozdeliť dátové inštancie do k zhlukov na základe spoločných čŕt. V prípade, že inštancie sú reprezentované nízkodimenzionálnym vektorom v Euklidovom priestore, môžu byť na zhlukovanie použité klasické techniky spomenuté v 2.3. Ak inštancie reprezentujú časový rad, nasadenie takýchto štandardných prístupov je zriedkavé [16].

Metódy používané na zhlukovanie časových radov môžeme rozdeliť do 3 skupín, na základe reprezentácie dát, s ktorými pracujú. Prvá skupina predpokladá surové dáta, 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 [29].

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 vzdialenostných alebo podobnostných metrík pre použité časové rady [29].

Hierarchické zhlukovanie produkuje vnorenú hierarchiu skupín podobných časových radov na základe vzdialenostných 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, aglomeratívne a deliace. Aglomeratívne 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 [42].

Aglomeratívne zhlukovanie na základe vzdialenosti 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 spätne zmeniť rozhodnutie zlúčiť zhluky [42].

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 vzdialenosti môžu byť použité metriky opísané v 2.4.5 [42].

2.4.2 Zhlukovanie na základe reprezentácie

Keďže manipulácia so surovými dátami je často náročná a dáta navyše obsahujú nadbytočné informácie, táto metóda (angl. *Representation based clustering approach*) najskôr transformuje dáta 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*) [29].

Samoorganizované mapy predstavujú triedu neurónových sietí, kde sú neuróny usporiadané v nízkodimenzionálnej štruktúre. Trénovanie prebieha iiteratívne 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 Euklidova 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 vzdialenosti od neuróna t. Nevýhodou je opäť náročné spracovanie časových radov rôznych dĺžok, keďže dĺžka časového radu definuje aj dĺžku váhového vektora w [20, 42].

2.4.3 Zhlukovanie na základe modelu

Metóda (angl. *Model based clustering approach*) predpokladá, že každý časový rad je generovaný nejakým modelom alebo pravdepodobnostnou distribúciou. Časové rady sú považované za podobné ak aj modely charakterizujúce jednotlivé časové rady sú si podobné [29].

ARIMA model navrhnutý v práci [44] zhlukuje jednorozmerné časové rady. Predpokladali, že časové rady sú vygenerované k rôznymi ARIMA modelmi. Vylepšili algoritmus na maximalizáciu očakávaní (angl. *expectation maximalization 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 nebol 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 [39].

Prístup 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 [39].

2.4.5 Metriky vzdialenosti

Kľúčovou záležitosťou pri zhlukovaní časových radov na základe ich podobnosti, je meranie vzdialenosti medzi nimi. Rovnako ako pri zhlukovaní bodových inštancií je potrebné definovať si metódy merania vzdialenosti. Najčastejšími metrikami sú Euklidova a Manhattanova 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, 42].

Metódy používané na meranie vzdialenosti 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. Euklidova 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 [16].

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

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

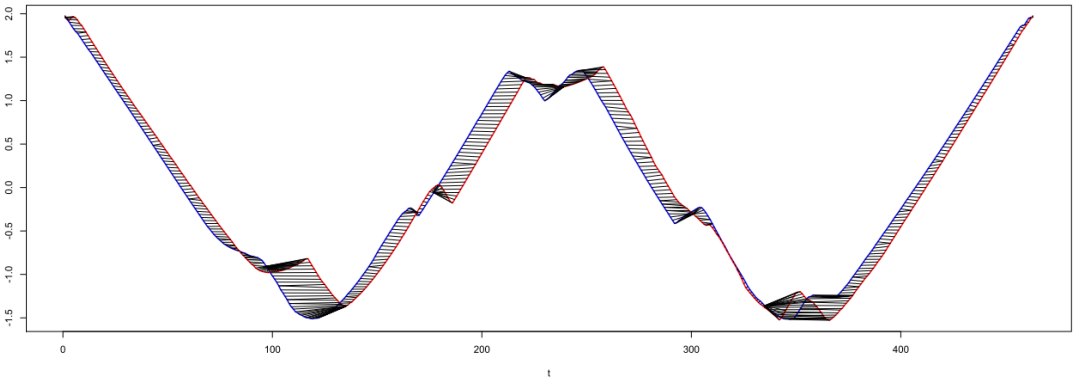
Korelácia blízka -1 znamená, že nárast kriviek časových radov je zrkadlový. Pri korelácií rovnej 0 hovoríme o rozdielnych časových radoch a pri hodnote 1 o podobných. Na základe hodnoty korelácie, potom môžeme vyjadriť vzdialenosť vzorcom 6. Nevýhodou je, ak máme k dispozícií iba malú, prípadne krátku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

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

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, 18].

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



Obr. 14: Príklad porovnávania časových radov pomocou dynamickej deformácií času [34].

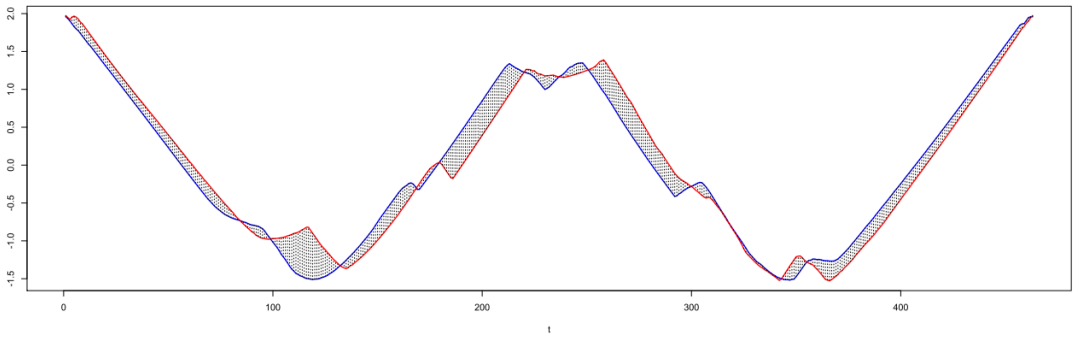
Do rovnakej rodiny vzdialenostných metrík patrí aj rýchle globálne zarovnávanie kernelov (angl. *Fast global alignment kernels*) skrátene GAK. Cieľom metódy je znížiť veľkú časovú náročnosť DTW s dosiahnutím porovnateľných výsledkov. Podobne aj metrika založená na tvare časových radov (angl. *Shape-based distance*) skrátene SBD, znižuje časovú náročnosť výpočtu vzdialenosti medzi časovými radmi. Narozdiel od GAK nepoužíva kernely, ale štatistické metódy založené na krížovej korelácií (angl. *cross-correlation*). Bližšie sa týmito metrikami zaoberali autori v prácach [11] a [27].

Kvalitatívna vzdialenosť je metóda založená na kvalitatívnom porovnávaní tvaru dvoch časových radov. 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 vyjadríme vzorcom 8, pomocou ktorého spočítame počet zhôd v raste časových radov. 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 vzdialenosti. Podobnosť tvarov kriviek je detegovaná aj v prípade, kedy neexistuje medzi časovými radmi lineárna alebo nelineárna závislosť [13].

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

Euklidova 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 [42].

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



Obr. 15: Príklad porovnávania časových radov pomocou Euklidovej vzdialenosti [34].

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

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

Pearsonov korelačný koeficient je používaný pri výpočte vzdialenosti, ktorá je založená na vzájomnej korelácií. Vo vzorci 11 reprezentuje $\widetilde{X}$ aritmetický priemer časového radu X. Vzdialenosť vyjadríme vzorcom 11 [42].

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

$$D_P\left(X, Y\right) = 2\cdot\left(1 - r\left(X, Y\right)\right) \tag{12}$$

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 nežiadúcich efektov škály sa používa z štandardizácia. Matematicky vzdialenosť vyjadríme vzorcom 13. Zložka t_k predstavuje čas [42].

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

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 [37].

2.5.1 Filtrovanie odberateľov

Dáta z inteligentných meračov bývajú často nekompletné a s chýbajúcimi hodnotami. Väčšina algoritmov nedokáže spracovať takéto dáta a všetky časové rady musia byť rovnakej dĺžky. Rovnako sú nepoužiteľné dáta, ktoré boli poškodené pri samotnom zbere dát, nie však pri meraní. Zatiaľ čo chybné meracie zariadenia môžu spadať do detekcie anomálií a zaujímajú nás, dáta ktoré boli zduplikované 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 [25].

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 vyčítať, 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 [8].

2.5.3 Extrakcia čŕt

Ďalšou technikou používanou pri príprave dát je tvorba nových atribútov založených na pôvodných, surových dátach. V súvisiacom článku [25] ide napr. o vytvorenie hodinového priemeru 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.

$$x_h = \frac{P_{h+1}}{D_{h+1} - D_h} \tag{14}$$

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 reťazci, počet prechodov medzi rôznymi behmi a počty prvých a posledných núl a jednotiek. Výhodami reprezentácie sú najmä redukcia dimenzií, zvýraznenie charakteru dát, paralelizmus a jednoduchá interpretácia dát [23].

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

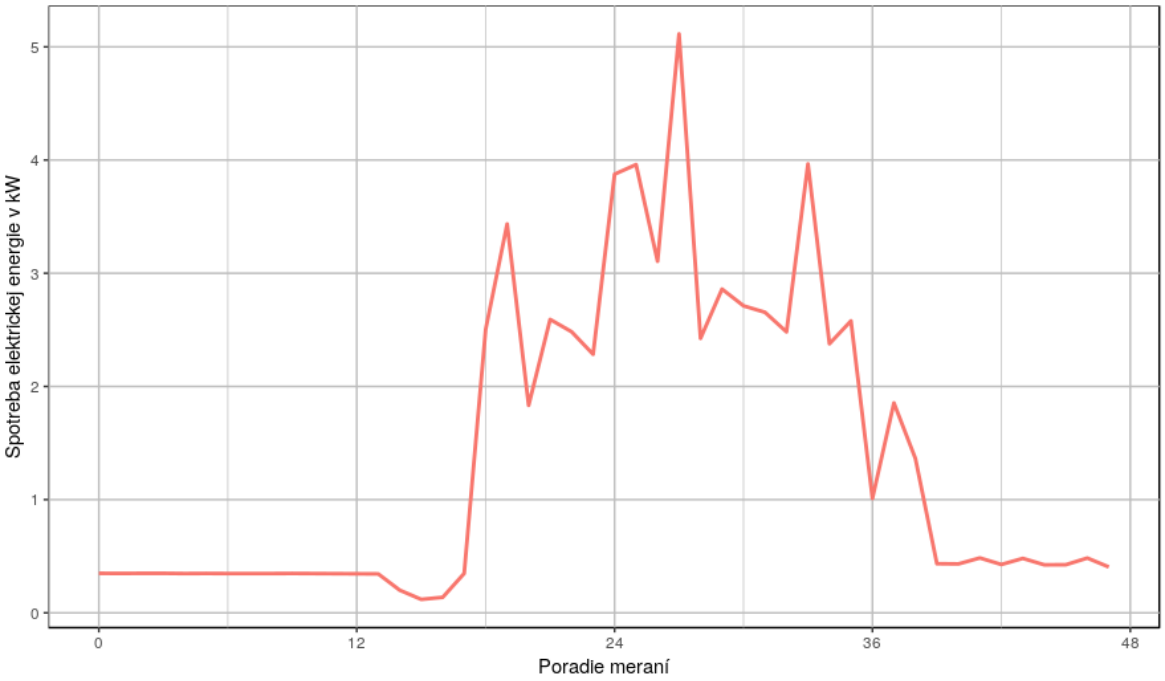
2.5.5 Agregácia dát

Dáta z meračov sú dostupné v pravidelných intervaloch. Pre jednoduchšiu manipuláciu s časovými radmi a redukciu dimenzií, môžu byť dáta 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 vyhladzuje malé odchýlky v časových radoch, čo môže sťažiť identifikáciu náhlej zmeny správania odberateľov. To môže viesť k nesprávnemu označeniu správania odberateľa za neštandardné [8].

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].



Obr. 16: Časový rad bez úprav

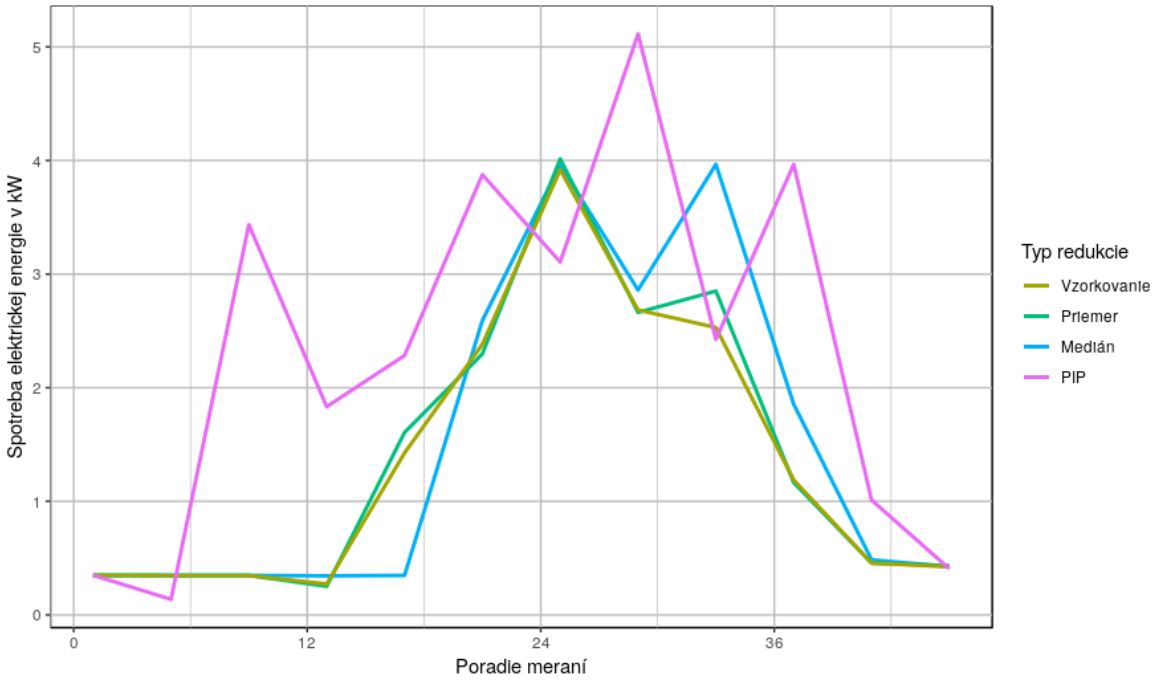
Lepšie výsledky dostaneme ak pri vzorkovaní budeme priemerovať hodnoty vo vzniknutých intervaloch. Táto metóda sa zvykne nazývať aj po častiach agregovaná aproximácia (angl. *piecewise aggregate approximation*), skrátene PAA. Vylepšenou verziou je metóda APCA, kde vzniknuté intervaly majú rôznu dĺžku, v závislosti od tvaru časového radu. Taktiež môžeme okrem priemeru použiť medián zvoleného intervalu. Obe metódy môžeme vidieť na obrázku 17 [6].

Ďalšou metódou je aproximácia pomocou rovných čiar, kde hlavnými kategóriami sú lineárna interpolácia a lineárna regresia. Bežnou metódou pri interpolácií je použiť po častiach lineárnu aproximáciu (angl. *piecewise linear approximation*). Algoritmus začína vytvorením odhadu časového radu, ktorý používa polovicu vytvorených intervalov. Tie sú následne zlučované, pokiaľ 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 vzdialenosti medzi dvoma susednými bodmi PIP. Proces pokračuje pokiaľ nie sú zoradené podľa dôležitosti všetky pôvodné body. Na obrázkoch 16 a 17 môžeme vidieť, že tvar kriviek pôvodného časového radu a redukovaného je mierne odlišný, čo je spôsobené roztiahnutím 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,



Obr. 17: Redukované časové rady.

čím sa znižuje dimenzionalita dát. Metóda má uplatnenie aj pri analýze či vizualizácií vysokodimenzionálnych dát. Ńajskôr 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 dáta 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 dvojdimenzionálnych dát [14, 35].

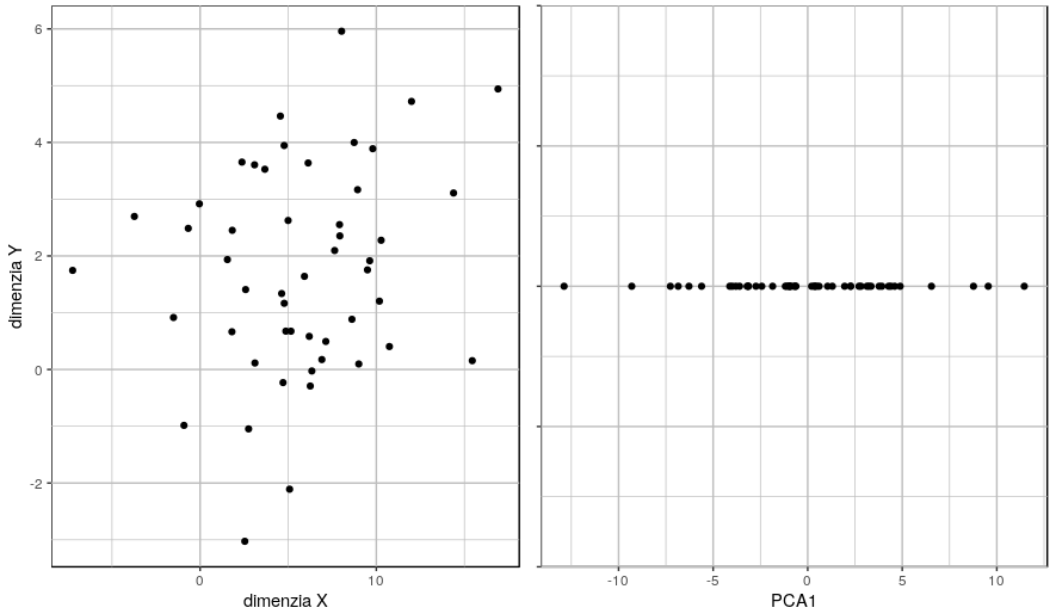
Na podobnom princípe ako DTW je založené aj hľadanie najdlhšej spoločnej podpostupnosti (angl. *longest common subsequence*), skrátene LCSS. Ide o variáciu editačnej vzdialenosti a spájania dvoch sekvencií, ktoré sa môžu natiahnuť a vynechať tak niektoré elementy bez toho, aby sa menilo ich poradie v rámci postupnosti. Narozdiel 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 čeliť viacerým problémom. Najjednoduchším prístupom je rozdeliť časový rad pomocou okna fixnej dĺžky do segmentov, z ktorých vznikajú jednoduché vzory. Jedinou úlohou je 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 líšiť, pri jednotlivých odberateľoch dokonca aj rádovo. Pri zhlukovaní takýchto časových radov je preto potrebná najskôr



Obr. 18: Príklad redukcie dimenzií pomocou PCA, pôvodný (vľavo) a redukovaný dataset (vpravo).

ich normalizácia, v prípade zhlukovania 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 vyjadríme nasledujúcim vzorcom 16 [2]

$$z = \frac{x - \mu}{\sigma} \tag{16}$$

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, 31].

Pri identifikácií anomálií 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 dispozícií označené iba anomálne dáta, je najskôr definovaná oblasť anomálnych dát a až následne normálna oblasť. Pri identifikácií anomálií v časových radoch v doméne energetiky je takýto prístup len ťažko aplikovateľný nakoľko podobné časové rady pri rôznych domácnostiach môžu, ale nemusia predstavovať normálne správanie [36].

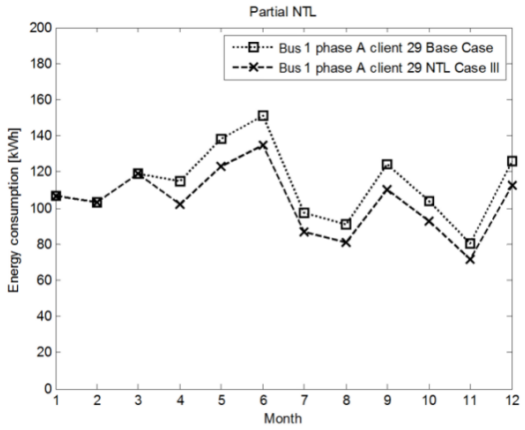
Najčastejšími metódami používanými pri nelegálnom odbere je obídenie meračov spotreby energie či samotná manipulácia s nimi. Merače tak poskytujú nesprávne informácie o spotrebovanej energie odberateľmi, čo je možné detegovať až po identifikácii celkových netechnických strát v sieti. Ďalšou populárnou metódou používanou na detekciu nelegálnych

odberov je analýza spotrebiteľského profilu zákazníka, kedy je našou snahou identifikovať nepravidelné vzory v nameraných spotrebiteľských dátach [31]. 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.

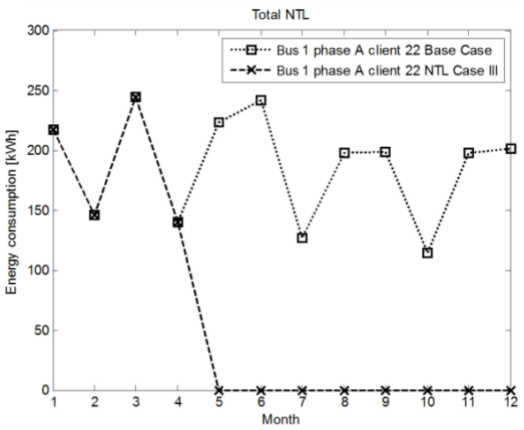
Vďaka inteligentným meračom je možné detegovať nelegálne odbery omnoho rýchlejšie, najmä kvôli vysokej frekvencii zberania údajov. Takto sú identifikované aj také odbery, ktoré by sa pri klasických meraniach stratili v týždenných alebo mesačných agregáciách. Úspešnosť detekcie nelegálnych odberov je 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 dispozícií väčšie množstvo dát a zároveň použiť zložitejšie algoritmy detekcie anomálií, ktoré sú popísané v súvisiacej práci [26].

V súvisiacich prácach sa autori zaoberali určením netechnických strát v elektrických distribučných sieťach s použitím rôznych štatistických metód alebo strojového učenia. Dostupné dáta od distribútorov pochádzali najmä z jedného zdroja, lokality a zameriavali sa na jeden zdroj energie. Dáta, ktoré budeme mať k dispozícií 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, miestností prípadne ich funkcie alebo zvýšením energetickej sebestačnosti. Následkom je nižšia nameraná spotrebe energie v dlhšom horizonte. Zníženie spotreby môže byť čiastočné alebo úplné, ako môžeme vidieť na obrázkoch 19 a 20 [36, 40].



Obr. 19: Čiastočné zníženie spotreby elektrickej energie [40].



Obr. 20: Úplné zníženie spotreby elektrickej energie [40].

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 [36].

2.7 Vyhodnocovacie metriky

Za predpokladu, že získané dáta 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 normálni odberatelia (angl. *true negative*) a správne označení normálni odberatelia (angl. *false negative*). Kvalitu klasifikácie potom môžeme zmerať pomocou presnosti a pokrytia. Presnosť vypočítame vzorcom 17, kedy ide o pomer správne označených anomálií a celkový počet označených anomálií. Tým vypočítame percento odberateľov, ktorých sme správne klasifikovali ako podozrivých.

$$\text{Presnosť} = \frac{TP}{TP + FP} \tag{17}$$

Pokrytie označuje pomer správne označených anomálií a celkový počet skutočných anomálií. Vyjadríme ju pomocou vzorca 18.

$$\text{Pokrytie} = \frac{TP}{TP + FN} \tag{18}$$

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 [40, 43].

Tabuľka 1: Matica zámen

		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 19, kde P predstavuje presnosť a C predstavuje pokrytie [40].

$$F = 2 \cdot \left(P^{-1} + C^{-1}\right)^{-1} \tag{19}$$

2.7.1 Zhlukovacie validačné indexy

Zhlukovanie je metóda, ktorej cieľom je určiť skupinu, do ktorej spadá daná inštancia. Triedenia 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* [5].

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 20, Jaccardov index vzorcom 21 a Fowlkes-Mallowsov index vzorcom 22. Indexy sú bližšie popísané v práci [5].

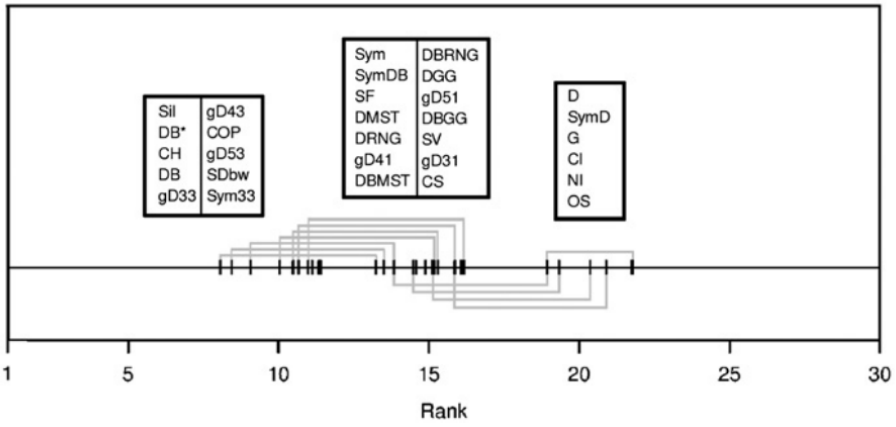
$$\text{RI} = \frac{TP}{FP + FN + TP} \tag{20}$$

$$\text{J} = \frac{TP + TN}{FP + FN + TP + TN} \tag{21}$$

$$\text{FM} = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} \tag{22}$$

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, Silhouetteov index a mnoho ďalších. V práci [3] autori analyzovali a porovnali 30 rôznych validačných indexov na rôznych datasetoch. Na syntetických datasetoch sa najviac osvedčili Silhouetteov 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.

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 nachádzajúcich 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. Silhouetteov index, modifikovaný Davies-Bouldinov index, Calinski-Harabasov index, Davies-Bouldinov index, generalizovaný Dunnov index a COP index [3].



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

2.8 Súvisiace práce v doméne energetiky a indentifikácií anomálií

V [16] bola pri zhlukovaní použitá aj kombinácia viacerých metód, konkrétne k-means, metóda náhodnej výmeny a aglomeratívne zhlukovanie. Ako už bolo spomenuté v 2.3, úlohou algoritmu k-means namapovať existujúce inštancie do k zhlukov. Aj keď metóda náhodnej výmeny je obmedzená na zhlukovacie problémy v Euklidovom priestore, bola použitá aj pri zhlukovaní časových radov a zabraňuje zaseknutiu zhluku v lokálnom minime. V princípe je náhodne vybraný zhluk, ktorý bude vymazaný a za centroid bude vybraný jeden časový rad z neho. Ak takéto riešenie je lepšie ako bez rozpustenia zhluku je nahradené pôvodným. Ako bolo spomenuté v 2.4.1, cieľom aglomeratívneho zhlukovania je všetky časové rady označiť ako zhluky a následne ich iteratívne zhlukovať. V momente, keď je vytvorených k zhlukov, je vypočítaný centroid zhluku a určená hierarchia zhlukov.

V práci [8] 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 dáta 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 [19] bolo navrhnúť novú vlnovú techniku na reprezentovanie viacerých vlastností meraných dát. Tiež vytvorili nový model, ktorý v sebe zahŕňa 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 nie je citlivý na 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 intervaly, ktoré nezodpovedajú štandardnému správaniu. Tie označujeme ako intervaly s výskytom anomálií. Cieľom našej práce ich bude nájsť a zmenšiť dĺžku nájdeného intervalu tak, aby bol čo najmenší, no zároveň v sebe zahŕňal 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 unikátne. Z dostupných dát však vieme po normalizácií extrahovať vzory, ktoré po následnom zhlukovaní predstavujú rádovo menej skupín, s ktorými ďalej pracujeme ako s definíciou normálneho správania. Väčšina článkov zaoberajúca 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 intervaly obsahujúce anomálie. Pri zhlukovaní časových radov experimentálne overíme vhodnosť voľby hyperparametrov ako je napr. počet zhlukov, vzdialenostná metrika alebo veľkosť použitého posuvného okna. Riedke zhluky budeme považovať za anomálne a budú podrobené ďalšej analýze, kedy budú identifikované zlomy, lokálne a globálne anomálie.

Vzhľadom na to, že dostupné dáta neobsahujú informáciu o anomáliách, budeme pri evaluácií 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 vzdialeností 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 dispozícií 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 intervaly 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.

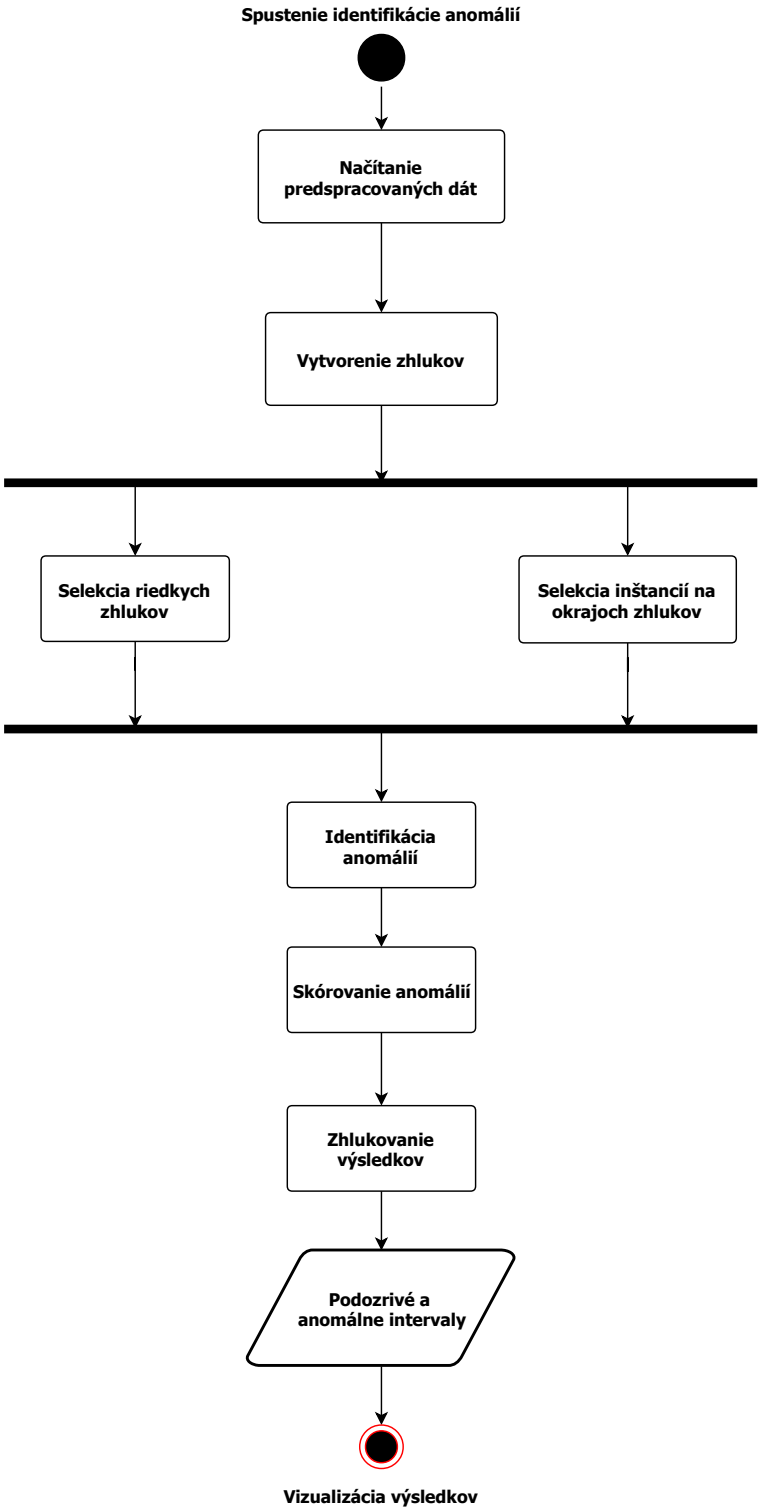
Výstupom opísaného procesu sú podozrivé a anomálne časové rady a jednotlivé merania v nich, ktoré sú taktiež považované za anomálie. Na výstupe sa môže podieľať viacero algoritmov, čo je potrebné zohľadniť pri vytváraní výsledného skóre. Na záver je potrebné zlúčiť jednotlivé merania do intervalov, ktoré svojim skóre opisujú mieru istoty, že označený interval obsahuje anomáliu. Výhodou takéhoto 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 neprekrývalo. 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 Euklidovej. Zhluky vo vytvorenom zhlukovaní sú rozdelené na základe početnosti jednotlivých skupín na majoritné a minoritné. Z majoritnej skupiny sú vybrané časové rady nachádzajúce sa na okraji zhluku. Časové rady z oboch skupín sú následne analyzované pomocou SHESD metódy, čím vznikájú jednotlivé merania v časových radoch označené ako anomálie. Vzniknutým bodom je pridelené skóre, ktoré opisuje mieru istoty, že dané meranie je anomálne. Body je následné nutné zlúčiť do intervalov, ktoré sú roztriedené do skupín.

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 [22] sme sa rozhodli pre metódu k-medoids, ktorá ako stred zhluku používa inštanciu, ktorej súčet vzdialeností od ostatných inštancií v zhluku je čo najnižšia. Takýto



Obr. 22: Stavový diagram procesu identifikácií anomálií.

vzťah môžeme zapísať rovnicou 23. 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 vzdialeností, následné je proces opakovaný so zamenením medoidov a inštanciami. Posledným krokom je výber riešenia, ktoré poskytuje najlepšie zhlukovanie.

$$\hat{\gamma} = min \sum_{j=1}^{k} \sum_{x \in K_j(\lambda)} d(x, m_j) \tag{23}$$

Pri práci so zhlukovacími metódami je nutné určiť viacero hyperparametrov, ako je napr. výsledný počet zhlukov, metrika vzdialenosti, 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 vzdialenosti sú popísané 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 vyhladiť existujúce anomálie do takej miery, že by neboli identifikované. Na druhej strane agregácia zabezpečuje elimináciu menších anomálií. 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 zapísať vzorcami 24 a 25, pre každé okno z intervalu $< 1, pocet\_tyzdnov - dlzka\_okna >$. 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. Výbrané okno časových radov je porovnávané na základe tvaru krivky, preto je nutné dáta 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 a jej priemerom a štandardnej odchýlky, čo môžeme zapísať vzorcom 26.

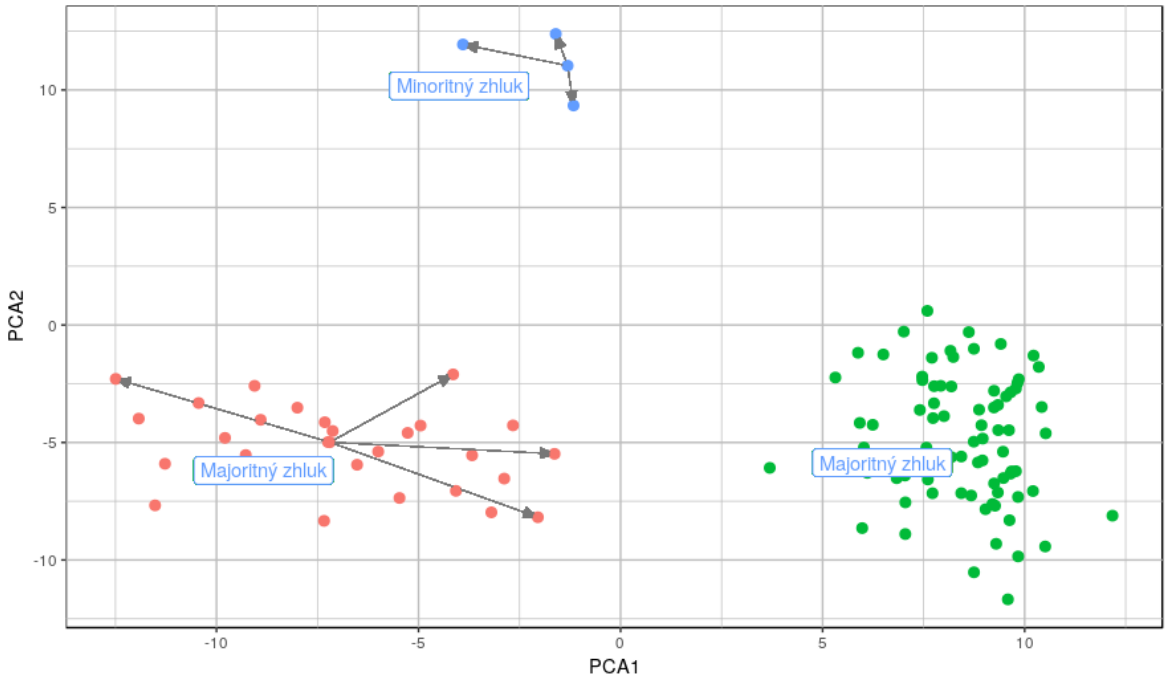
$$index_{zaciatok} = (poradie\_tyzdna - 1) * pocet\_merani\_tyzdenne \tag{24}$$

$$index_{koniec} = (poradie\_tyzdna + dlzka\_okna - 1) * pocet\_merani\_tyzdenne \tag{25}$$

$$z = \frac{x - \mu}{\sigma} \tag{26}$$

3.2 Ohodnotenie podozrivých zhlukov a intervalov

Pre výber riedkych zhlukov a inštancií na okraji zhlukov je potrebné vypočítať skóre, na základe ktorého bude daný časový rad považný 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



Obr. 23: Skórovanie podozrivých inštancií a zhlukov.

inštancie. Ich násobením je vypočítané anomálne skóre časového radu, zapísané rovnicou 27. Predpokladom pre výpočet skóre je zhlukovanie, ktoré okrem rozdelenia inštancií do zhlukov obsahuje aj informáciu o vzdialenosti jednotlivých inštancií od centroidu zhluku, do ktorého patrí.

$$skore_i = skore_{instancia_i} * skore_{zhluk_j}, \text{ pre } instancia_i \in zhluk_j \tag{27}$$

Vzorec pre výpočet skóre zhluku môžeme zapísať vzorcom 28 a skóre inštancie vzorcom 29. 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 vzdialeností inštancií od centroidu v rovnakom zhluku.

$$skore_{zhluk_i} = \sum_{j=1}^{n} \begin{cases} 1 \text{ ak } P(pocetnost_i \leq Q_j) \\ 0 \text{ inak} \end{cases} \text{, pre j } \in (0.05, 0.1, ..., 1) \tag{28}$$

$$skore_{instancia_i} = \frac{vzdialenost_i}{priemerna\_vzdialenost\_zhluku_j}, \text{ pre } instancia_i \in zhluk_j \tag{29}$$

Navrhnuté skórovanie 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 podozrivých inštancií

Vypočítané skóre anomálnosti je potrebné vyhodnotiť a porovnávať navzájom voči ostatným navrhovaným skórovaniam. 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 skórovanie je založené na vhodnej reprezentácií časového radu v dvojdimenzionálnom priestore 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 vysokodimenzionálnych dát a následná jednoduchá vizualizácia. Keďže metóda FeaClip je založená na transformácií dát podľa vzorca 30 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 [23]. Pre vizualizáciu vybraného intervalu časového radu je nutná najskôr FeaClip transformácia po oknách, spriemerovanie výsledkov a následný výber vhodných atribútov alebo aplikovanie PCA.

$$\hat{x_i} = \begin{cases} 1 \text{ ak } x_i > \mu \\ 0 \text{ inak} \end{cases} \quad \text{, pre i } \in (1, 2, ..., n) \tag{30}$$

Z extrahovaných vlastností nás zaujímajú najmä počet jednotiek v reťazci a počet prechodov medzi rôznymi behmi. Beh predstavuje súvislú postupnosť jedného znaku. Operácie môžeme zapísať vzorcami 31 a 32. Za anomálne sú považované intervaly časových radov, ktorých vlastnosti nespadajú do invervalu medzikvartilového pravidla $< Q1 - 1.5 * IQR, Q3 + 1.5 * IQR >$. Pri vizualizácií pomocou PCA, sú anomáliami inštancie nachádzajúce sa mimo oblasti väčšiny dát.

$$sum_1 = \sum_{i=1}^{n} casovy\_rad_i \tag{31}$$

$$RLE = dlzka_{RLE}(casovy\_rad_i) - 1 \tag{32}$$

Rovnako ako pri FeaClip reprezentácií, tak aj pri skórovaní inštancií 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.

3.4 Vyhladenie časových radov

TODO Pridat nieco o vyhladzovani + graf

3.5 Analýza metódou S-H-ESD

Cieľom je vylepšiť skóre vypočítané v predchádzajúcich krokoch a spresniť tak interval, v ktorom je výskyt anomálií výrazne väčší. Výstupom je pôvodný časový rad s novým atribútom, ktorý pre každé meranie obsahuje skóre opisujúce podozrenie na anomálnosť. Atribút budeme nazývať skóre odberateľa, nakoľko je vypočítané pre celé sledované obdobie spotreby odberateľa. Môžeme ho vyjadriť vzorcom 33. 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 dvojitou sezónnosťou.

(33)

Dosiahnuté výsledky je nutné vhodne vizualizovať a vyhodnotiť. Príkladom môže byť vizualizácia výsledkov na obrázku **??** TODO pridat obrazok. Vzhľadom na fakt, že metóda je založená na učení bez učiteľa, vyhodnocovanie prebieha na základe už spomenutých riešení, konkrétne FeaClip transformácie a metóde S-H-ESD.

4 Experimentálne overenie

Pri experimentoch sme pracovali v jazyku R. Použité knižnice s verziami sú zobrazené pomocou tabuľky 3.

Tabuľka 3: Použité knižnice jazyka R.

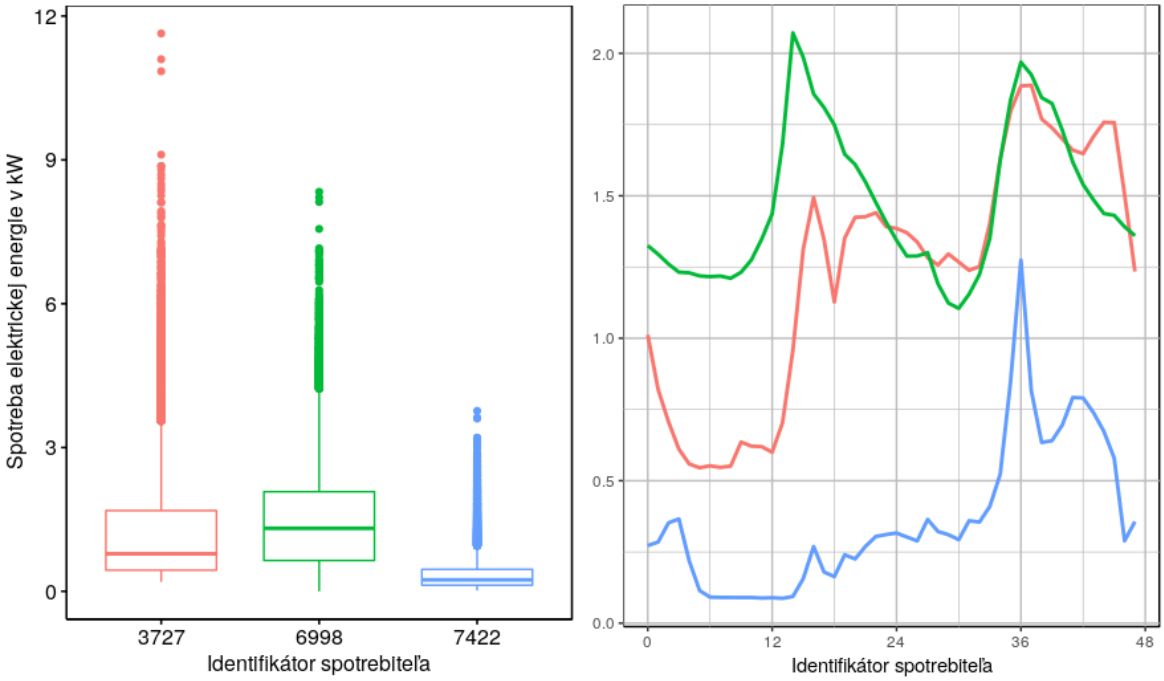
Názov	**Použitá verzia**
AnomalyDetection	1.0
BreakoutDetection	1.0.1
cluster	2.0.6
clusterCrit	1.2.8
data.table	1.12.0
devtools	1.13.6
dplyr	0.7.8
dtw	1.20-1
dtwclust	5.5.1
ggplot2	3.1.0
lubridate	1.7.4
pkgmaker	0.27
plotly	4.8.0
proxy	0.4-22
registry	0.5
rngtools	1.3.1
stringr	1.3.1
TSrepr	1.0.1
zoo	1.8-4

Pri experimentoch sme 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, spotrebu 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 24 na vizualizáciu náhodne vybraných odberateľov.

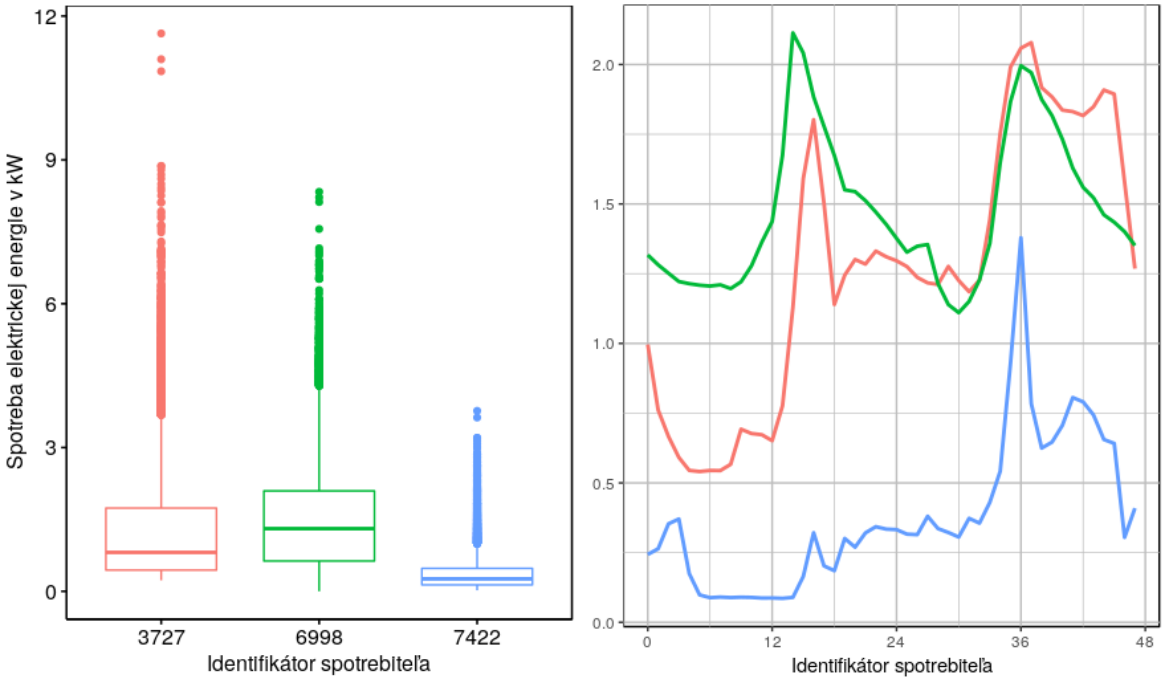
Tabuľka 4: Charakteristiky polohy použitého datasetu.

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 zhlukovania 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 25 a 26.



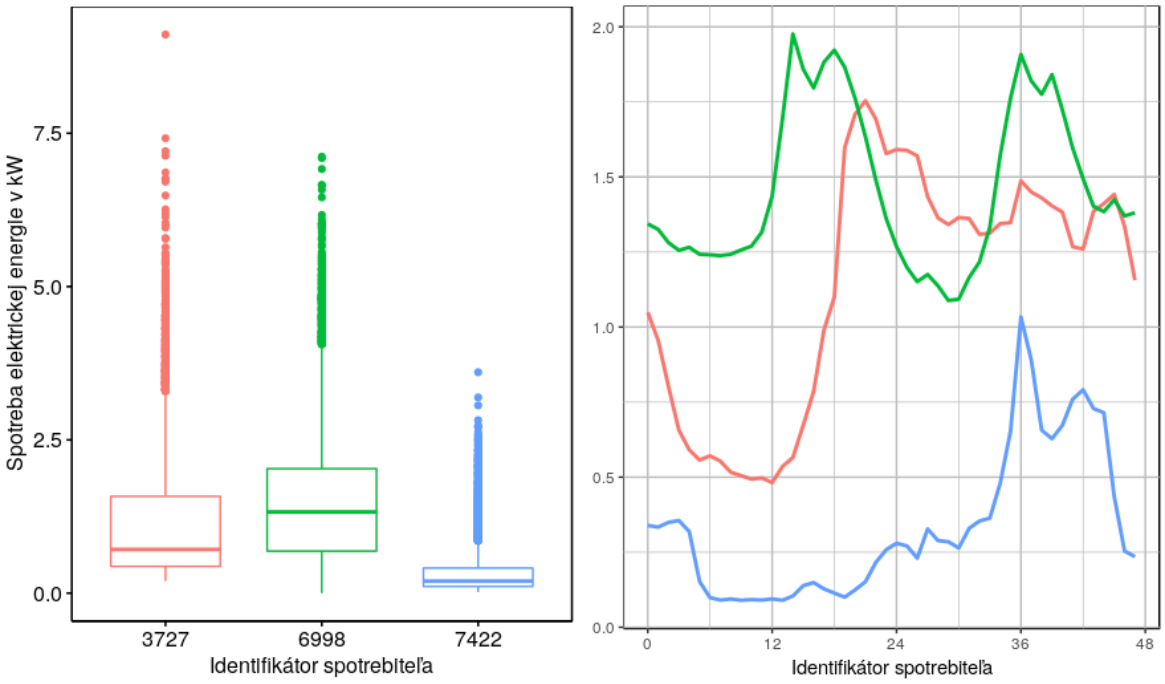
Obr. 24: Krabicový graf znázorňujúci spotrebu vybraných odberateľov.



Obr. 25: Krabicový graf znázorňujúci spotrebu vybraných odberateľov počas pracovných dní.

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

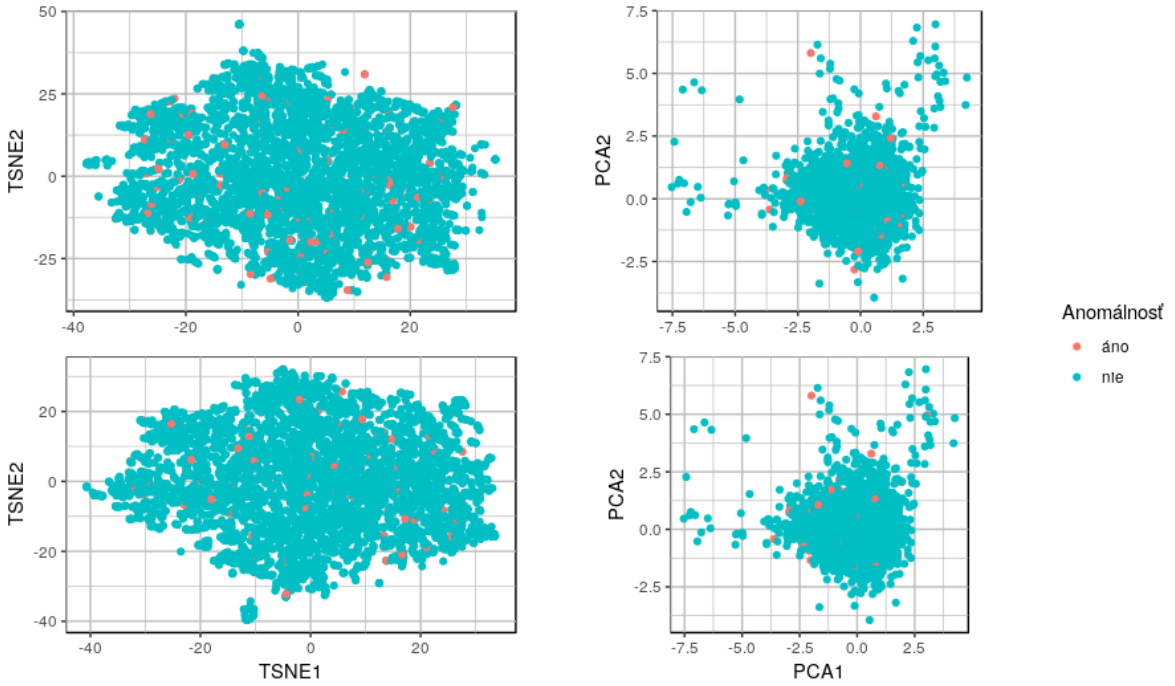


Obr. 26: Krabicový graf znázorňujúci spotrebu vybraných odberateľov počas dní voľna.

4.1 Existujúce riešenia

Experimenty, ktoré sme vykonali sme porovnávali s existujúcou implementáciou metódy S-H-ESD, ktorá sa nachádza v balíčku *AnomalyDetection* [1]. 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 27. V prvom riadku

[1] https://github.com/twitter/AnomalyDetection



Obr. 27: Grafy zobrazujúce identifikovaných anomálnych odberateľov pomocou metódy S-H-ESD.

sú výsledky funkcie *AnomalyDetectionTs* s 3.05% anomálnych odberateľov, v druhom riadku sú výsledky funkcie *AnomalyDetectionVec* s 2.27% anomálnych odberateľov.

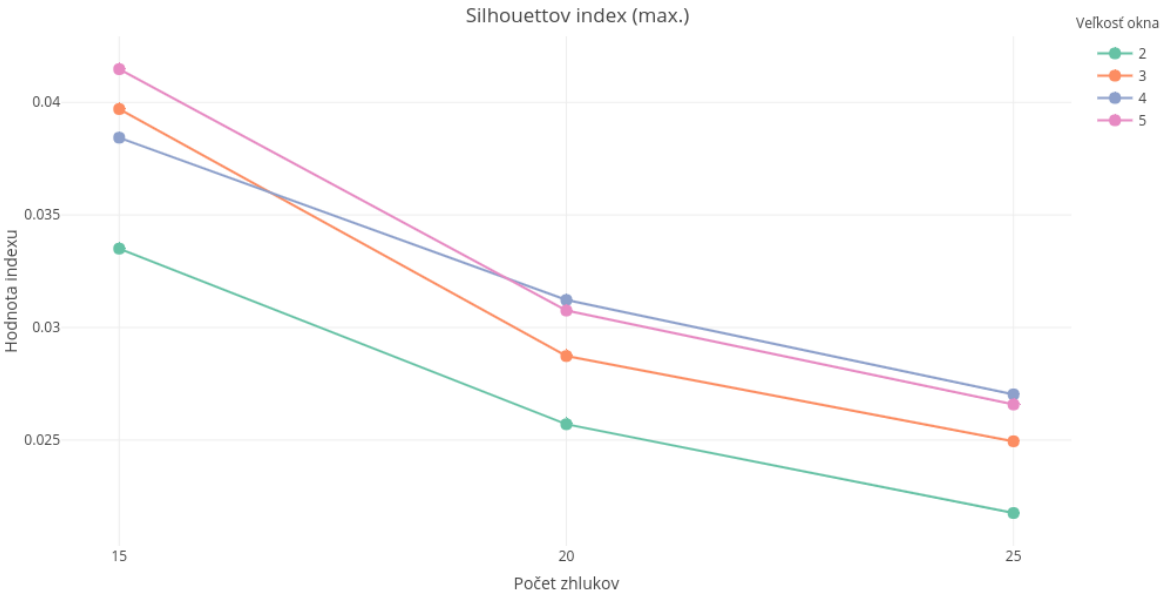
Funkcia *AnomalyDetectionTs* očakáva na vstupe časový rad, ktorý obsahuje časovú známku a hodnotu sledovanej veličiny v danom čase. Nutnou podmienkou je dostatočne dlhý časový rad, nad ktorým je spustená identifikácia anomálií. Pri experimentoch, kedy bola analyzovaná iba časť časového radu je kvôli 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 rýchlosti identifikácie anomálií.

4.2 Výber hyperparametrov zhlukovania

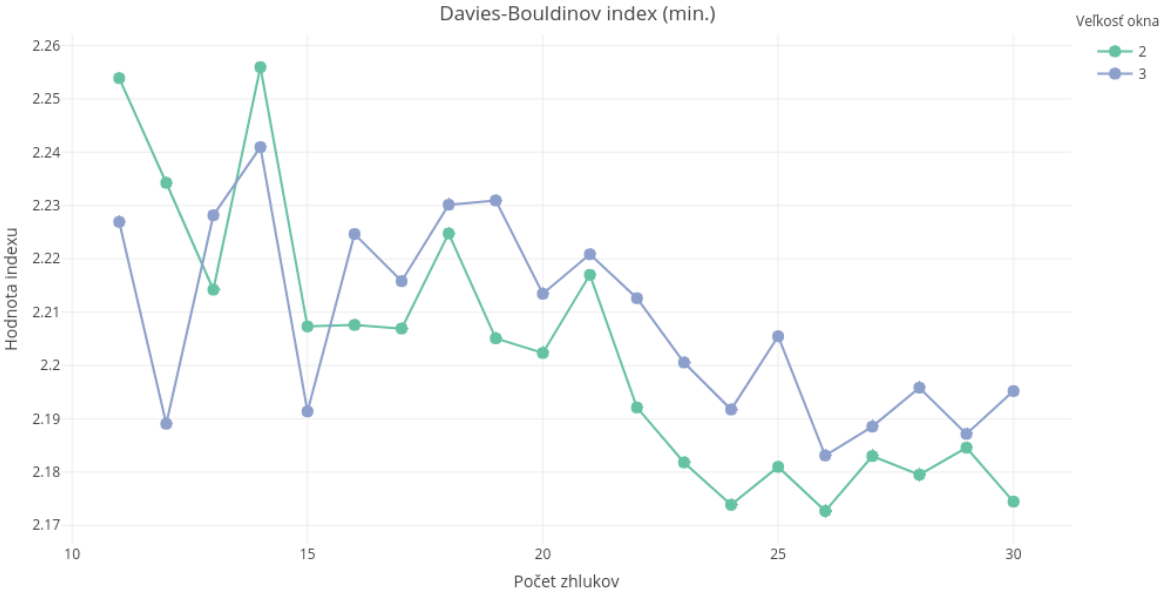
Zhlukovacie metódy poskytujú viacero parametrov, ktoré ovplyvňujú výsledné zhlukovanie, jeho kvalitu alebo časovú náročnosť. Pri práci sme sa zamerali najmä na dosahovanú presnosť, ktorú sme merali pomocou zhlukovacích validačných indexov, bližšie opísaných v kapitole 2.7.1. Niektoré hyperparametre sme testovali iba na požadovanom rozmedzí. Veľkosť posuvného okna by nemala presahovať 4-5 týždňov, aby okno neobsahovalo sezónnosť jednotlivých ročných období. Všetky výsledky experimentov sa nachádzajú v prílohe v kapitole B. Z vybraných grafov 28 a 29 je zrejmé, že najlepším nastavením hyperparametrov je práve nízky počet okien, ktoré budú agregované. Výsledný počet zhlukov by mal byť približne 25. Ostatné grafy podporujú naše tvrdenie, prípadne neposkytujú dostatočnú výpovednú hodnotu, keďže rozdiel medzi jednotlivými pokusmi je minimálny.

Ďalším testovaným hyperparametrom sú vzdialenostné metriky, ktoré sú použité implementované v knižnici *dtwclust* [2]. Metriky sú bližšie popísané v kapitole 2.4.5. Z kapitoly 2.7.1

[2]https://CRAN.R-project.org/package=dtwclust

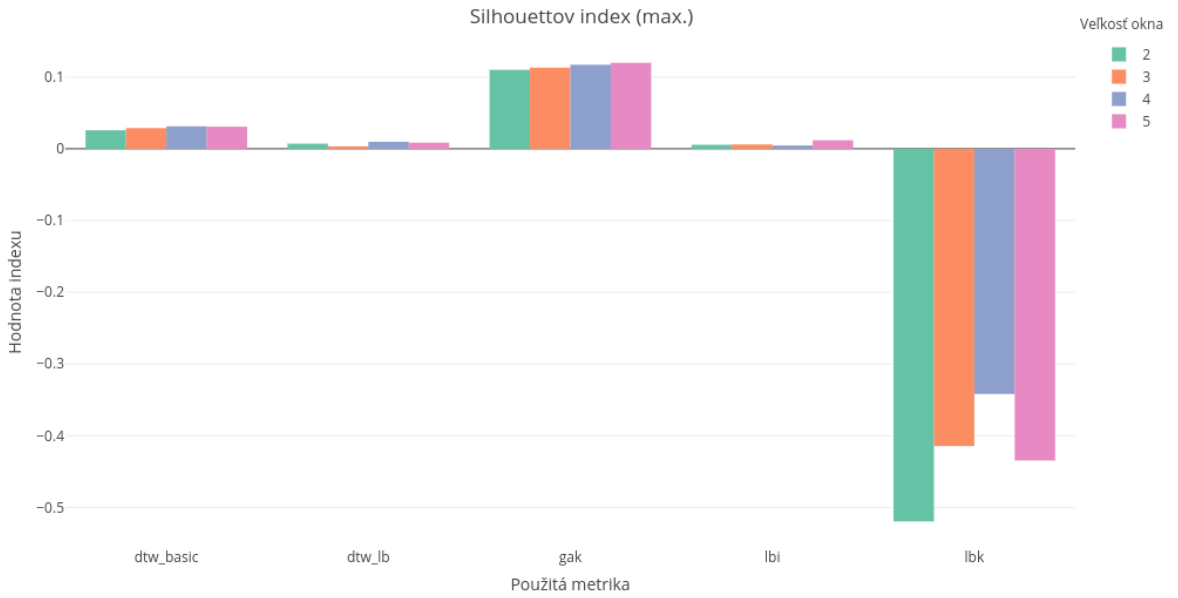


Obr. 28: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.

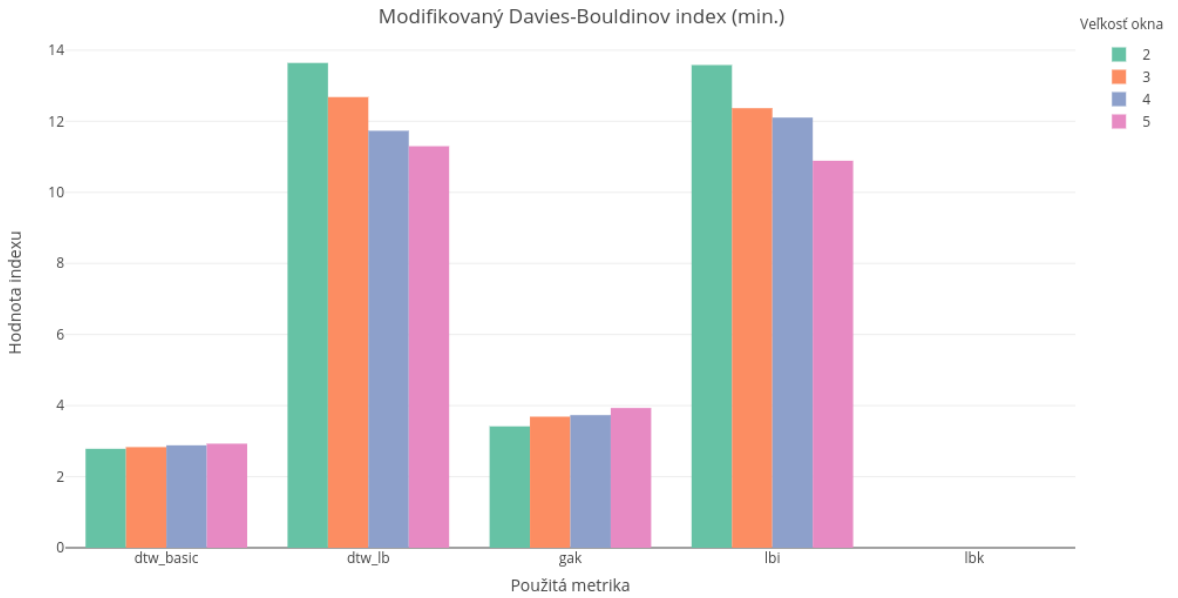


Obr. 29: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.

je zrejmé, že najlepšiu informáciu o kvalite zhlukovania poskytujú práve Silhouetteov index a modifikovaný Davies-Bouldinov index, vizualizované na grafoch 30 a 31. Najvhodnejšími vzdialenostnými metrikami sú potom GAK a DTW, pri ďalších experimentoch preto budeme používať GAK 2.4.5. Je dôležité poznamenať, že pri rovnakom nastavení funkcie, sú výsledky medzi jednotlivými behmi nezávislé a rôzne. Exprimentmi sme však overili, že rozdiely sú štatisticky nevýznamné.



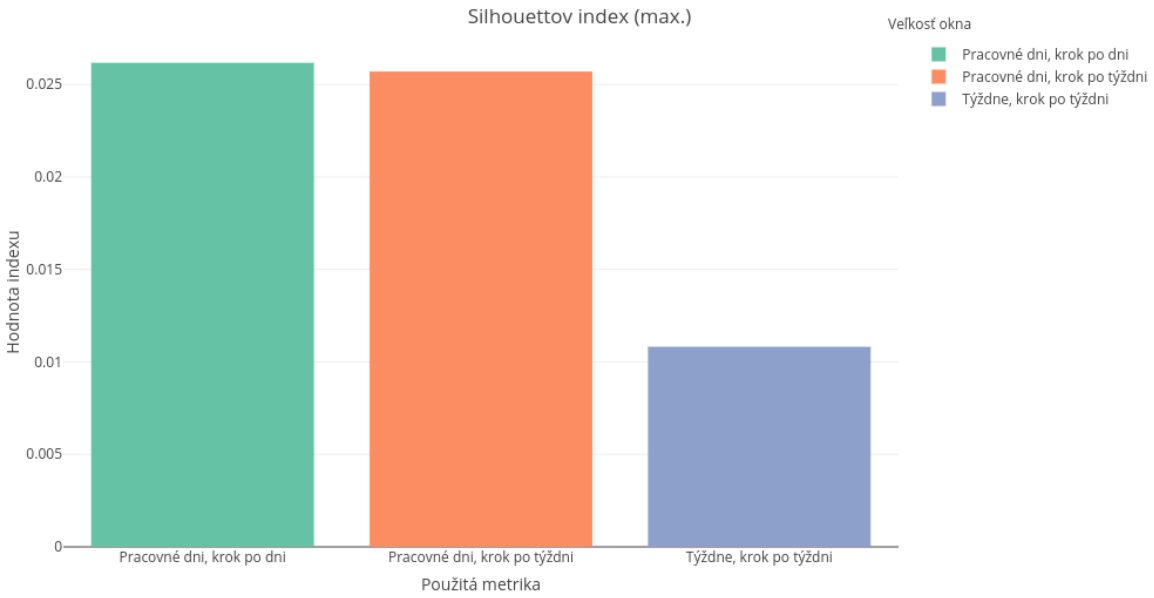
Obr. 30: Graf zhlukovania, porovnanie vzdialenostných metrík.



Obr. 31: Graf zhlukovania, porovnanie vzdialenostných metrík.

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 32 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.

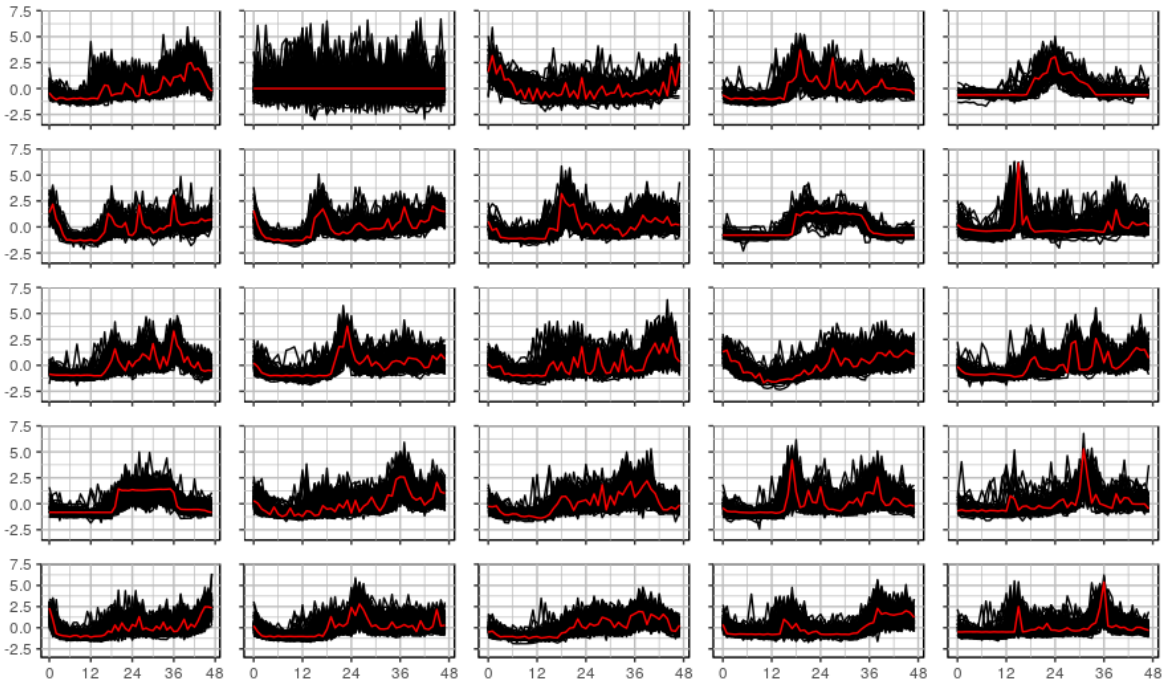


Obr. 32: Graf zhlukovania, porovnanie veľkostí a typov posuvných okien.

Predspracovanie dataset pozostáva aj z normalizácie dát pomocou z-skóre, ktoré je bližšie opísané v kapitole 2.5.8. Použitá knižnica *dtwclust* [3] 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 oboch súčasne alebo iba externej normalizácie sú dosahované výsledky porovnateľné.

Výsledkom experimentov je nové zhlukovanie pre každé posuvné okno. Náhodne vybrané zhlukovanie je zobrazené aj na obrázku 33. Je zrejmé, že jednotlivé medoidy sa navzájom dostatočne líšia a potvrdzujú správnu voľbu hyperparametrov zhlukovacieho algoritmu.

[3] https://CRAN.R-project.org/package=dtwclust



Obr. 33: Vytvorené zhluky so zvýraznenými medoidmi.

Literatúra

[1] Adhikari, R.: *An Introductory Study on Time Series Modeling and Forecasting*. Saarbrucken: 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.
URL http://doi.acm.org/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.
URL http://dx.doi.org/10.1016/j.patcog.2012.07.021

[4] Arun Kejariwal, S. W., James Tsiamis: 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.

[5] Bilgic, E.; Cakir, O.: Comparing clusterings: a store segmentation application, 10 2018, (čaká na publikovanie).

[6] 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.
URL http://doi.acm.org/10.1145/568518.568520

[7] 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.

[8] 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.

[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.
URL https://doi.org/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.
URL http://dl.acm.org/citation.cfm?id=3104482.3104599

[12] Depuru, S. S. S. R.: *Modeling, detection, and prevention of electricty theft for enhanced performance and security of power grid*. Dizertačná práca, The University of Toledo, 2012.

[13] Dzeroski, S.; Gjorgjioski, V.; Slavkov, 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.
URL http://dx.doi.org/10.1016/j.engappai.2010.09.007

[15] Grmanová, G.; Laurinec, P.; Rozinajová, V.; aj.: Incremental Ensemble Learning for Electricity Load Forecasting. *Acta Polytechnica Hungarica*, ročník 13, č. 2, 2016.

[16] 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.

[17] Hochenbaum, J.; Vallis, O. S.; Kejariwal, A.: Automatic Anomaly Detection in the Cloud Via Statistical Learning. *CoRR*, ročník abs/1704.07706, 2017, 1704.07706.
URL http://arxiv.org/abs/1704.07706

[18] 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.

[19] 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.

[20] 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.

[21] Kuppusamy, M.; Kaliyaperumal, S.: Comparison of Methods for detecting Outliers. *International Journal of Scientific & Engineering Research*, ročník 4, 01 2013: s. 709–714.

[22] Laurinec, P.; Lucka, M.: *Improving Forecasting Accuracy through the Influnce of Time Series Representations and Clustering*. Dizertačná práca, Slovak University of Technology in Bratislava, Ilkovičova 2, 842 16 Bratislava, Slovakia, 5 2018.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.
URL http://doi.acm.org/10.1145/2949741.2949758

[28] 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.
URL https://doi.org/10.1186/s12859-015-0645-6

[29] 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.
URL http://research.ijcaonline.org/volume52/number15/pxc3881278.pdf

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Simon Malinowski, L. R. T.: Recent advances in Time Series Classification. URL: http://www.antoniomucherino.it/events/CDs/CD03/TimeSeriesClassification.pdf, 6 2017.

[35] Smith, L. I.: A tutorial on principal components analysis. Technická správa, Cornell University, USA, February 26 2002.
URL http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[36] 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.

[37] 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.
URL http://www.sciencedirect.com/science/article/pii/S0142061514002750

[38] Tan, P.-N.; Steinbach, M.; Kumar, V.: *Introduction to Data Mining*. Addison Wesley, us ed vydanie, May 2005, ISBN 0321321367.

[39] 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.

[40] 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.

[41] 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

[42] 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.

[43] 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.
URL http://portal.acm.org/citation.cfm?doid=1150402.1150498

[44] 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 Obsah elektronického média

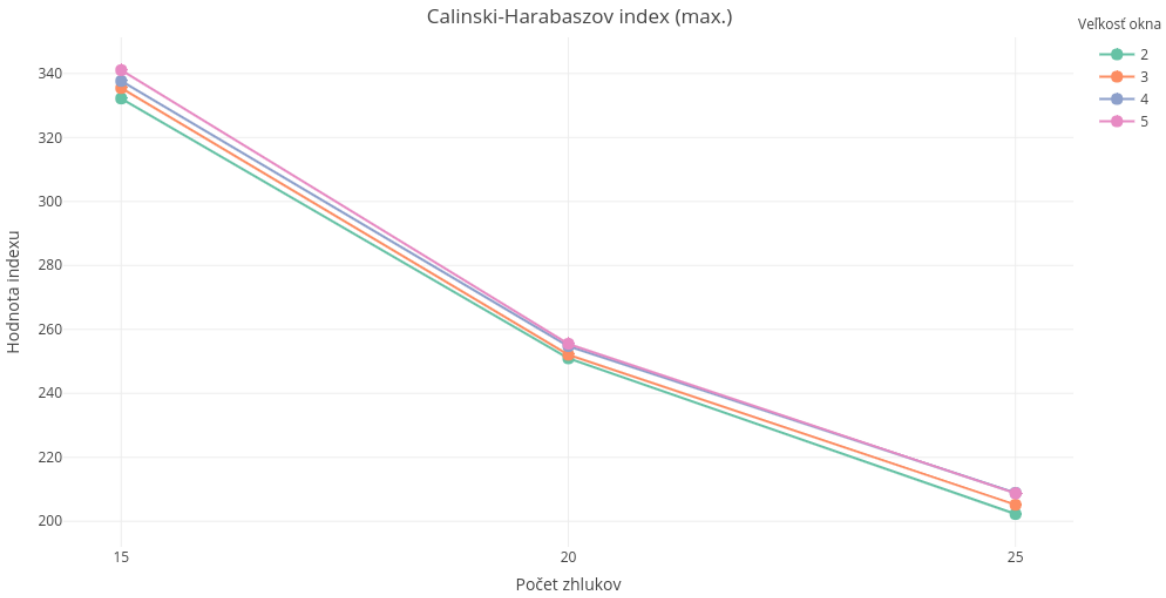
```
CD nosič
├── doc
│   └── DP_MATUS_CUPER.pdf
└── src
    ├── aggregators.R
    ├── analyzators.R
    ├── anomalyDetectors.R
    ├── boilerplate.R
    ├── filters.R
    ├── loaders.R
    ├── oneliners.sh
    ├── presentation.R
    ├── ts-sample-decomposition.R
    ├── utilities.R
    └── visualizators.R
```

- **doc** dokumentácia, obrázky použité v nej a zdrojové súbory pre LaTeX
- **src** skripty, prototypy a časti zdrojových kódov, ktoré boli použité pri experimentoch

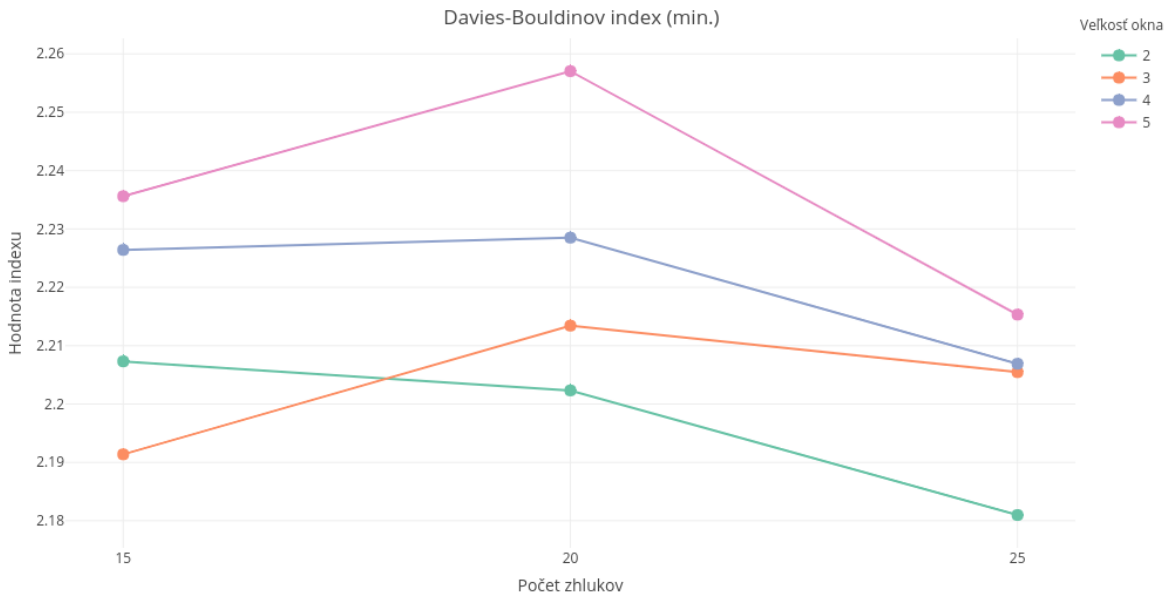
B Vizualizácie experimentov pre výber hyperparametrov

Každý index obsahuje aj informáciu o jeho optimálnych hodnotách. Pri indexoch, ktoré obsahujú $(max.)$ znamenajú väčšie hodnoty lepšie výsledné zhlukovanie. Pri indexoch s $(min.)$ nižšie hodnoty znamenajú lepšie výsledky zhlukovania.

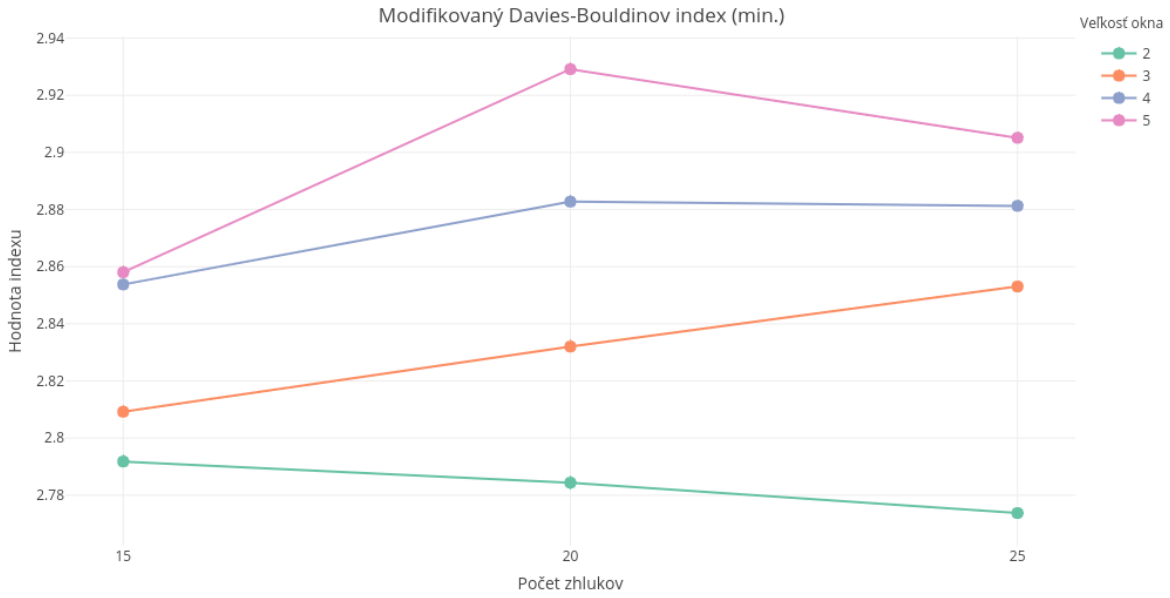
Obr. 34: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



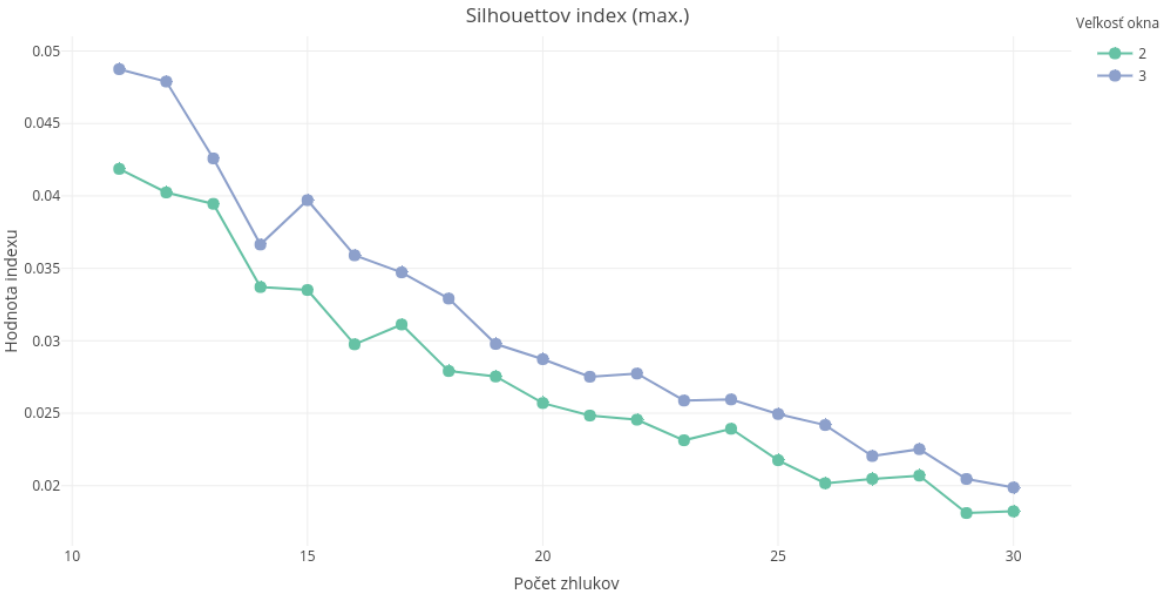
Obr. 35: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



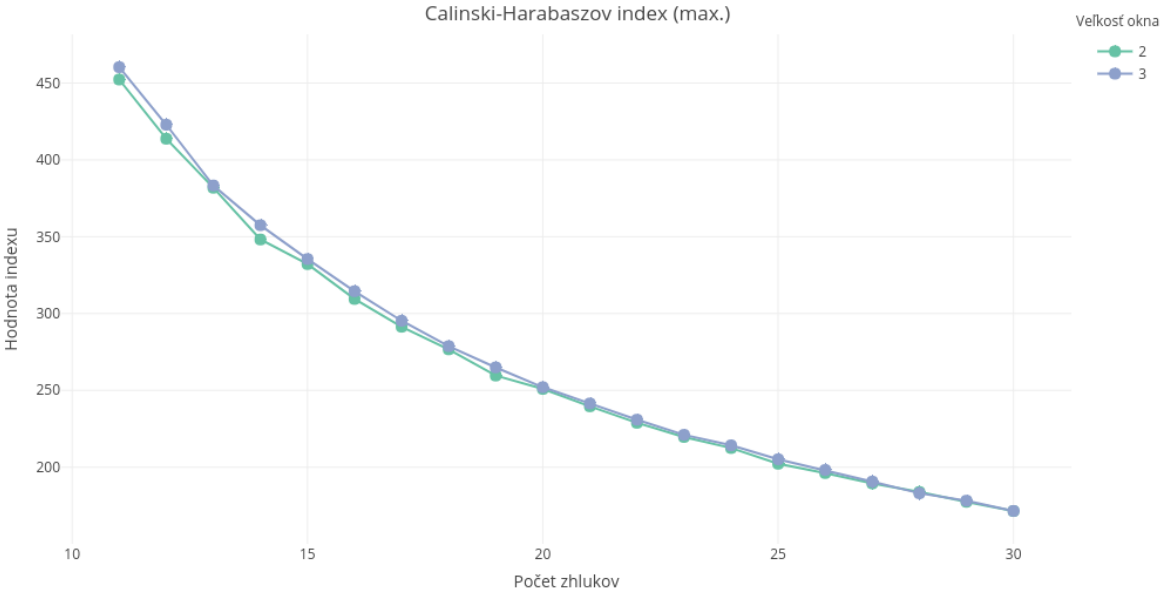
Obr. 36: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



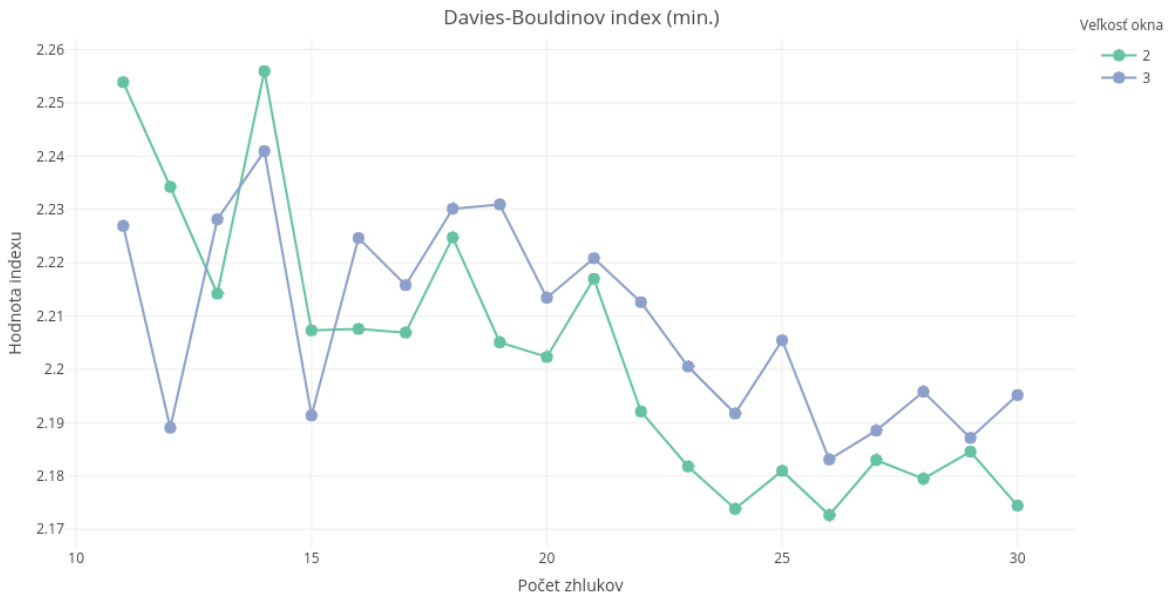
Obr. 37: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



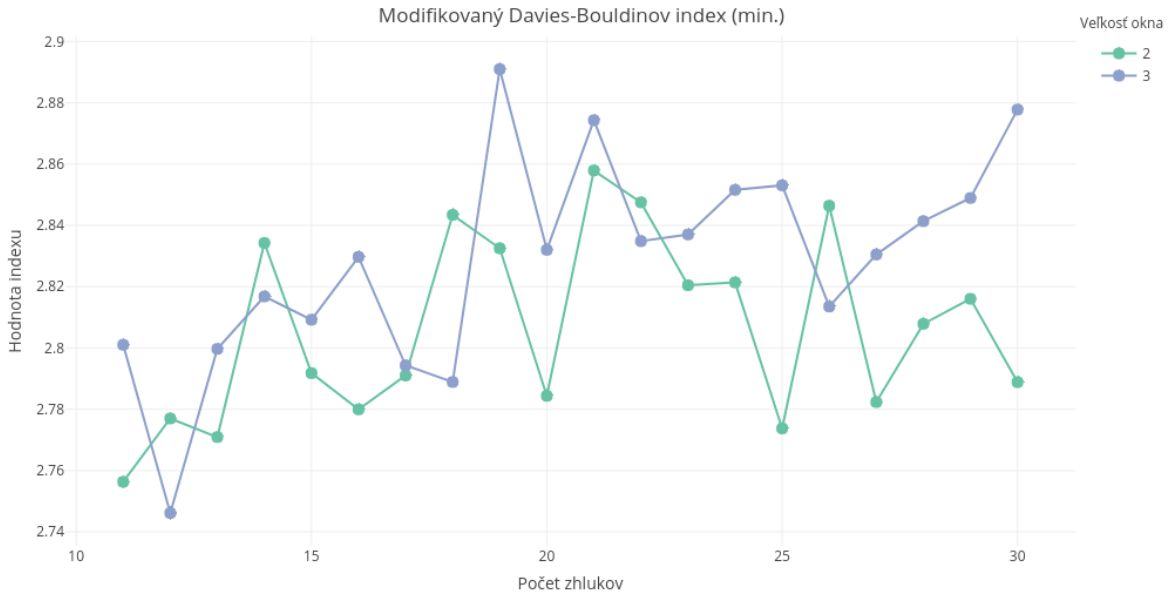
Obr. 38: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



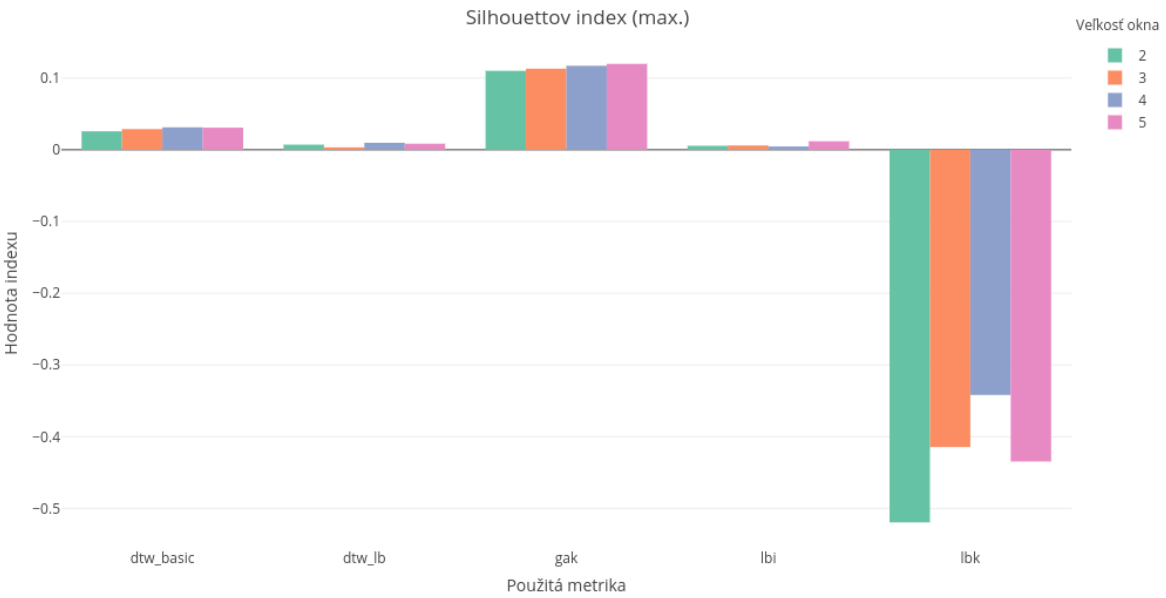
Obr. 39: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



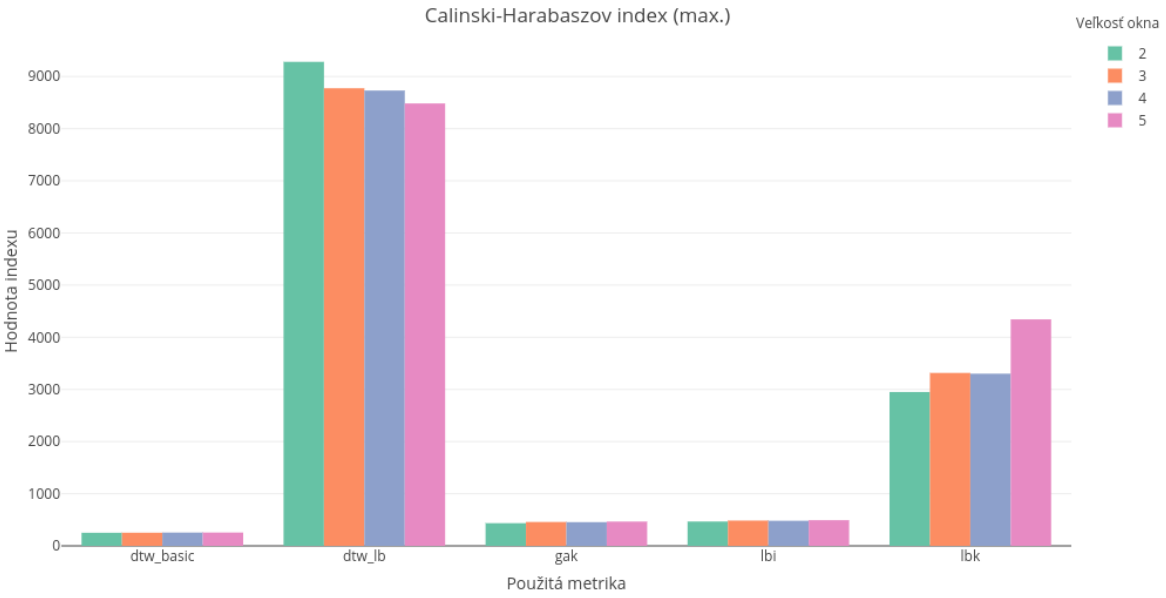
Obr. 40: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



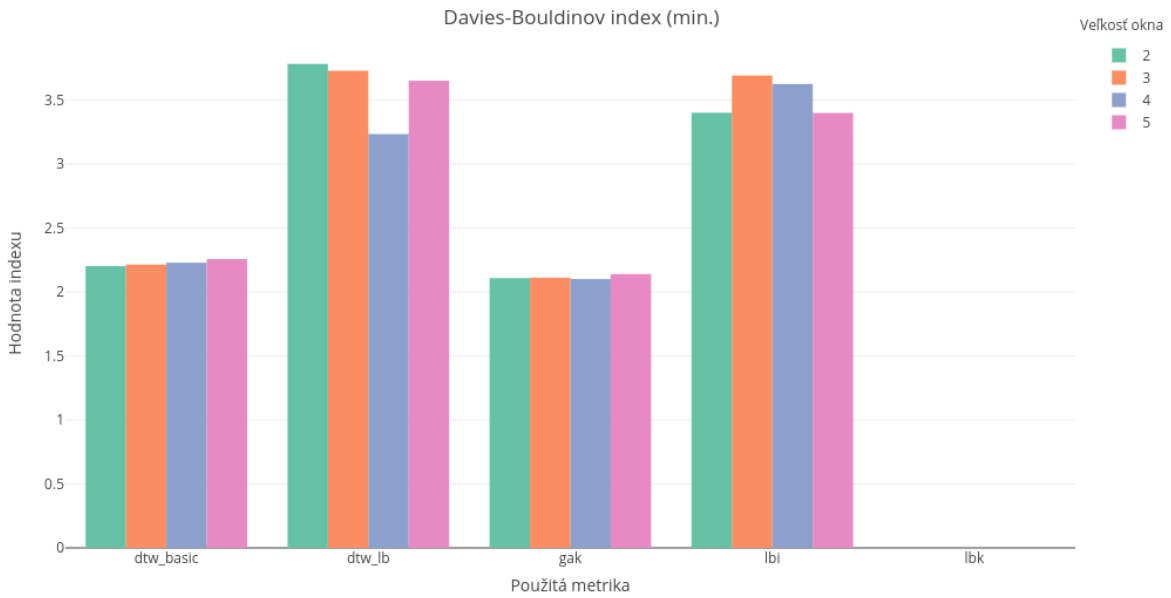
Obr. 41: Graf zhlukovania, porovnanie veľkosti posuvného okna a počtu zhlukov.



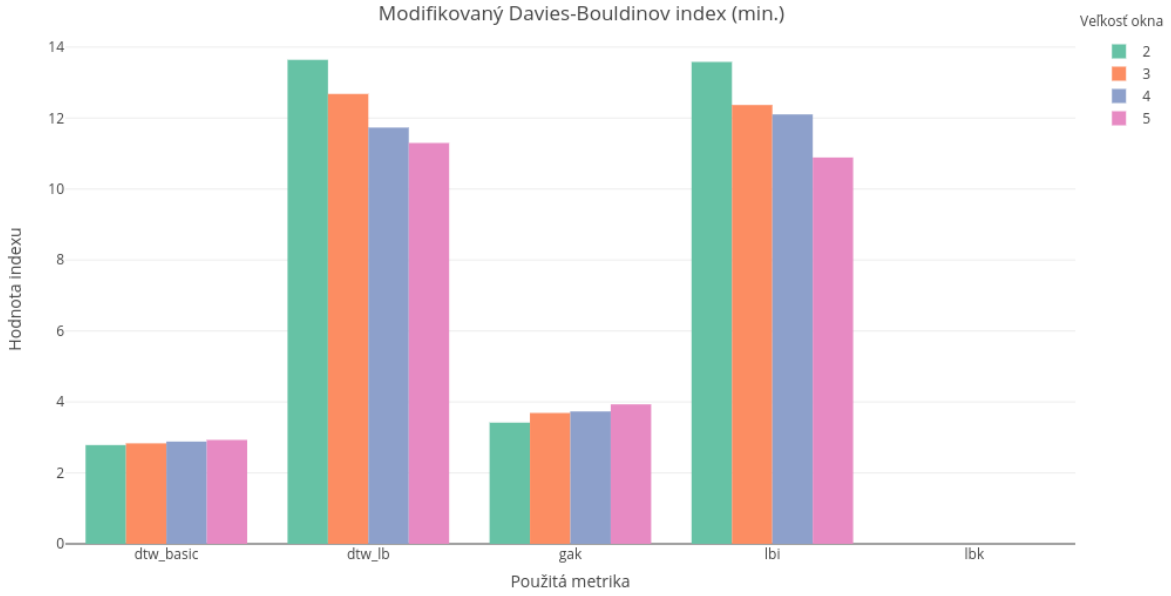
Obr. 42: Graf zhlukovania, porovnanie vzdialenostných metrík.



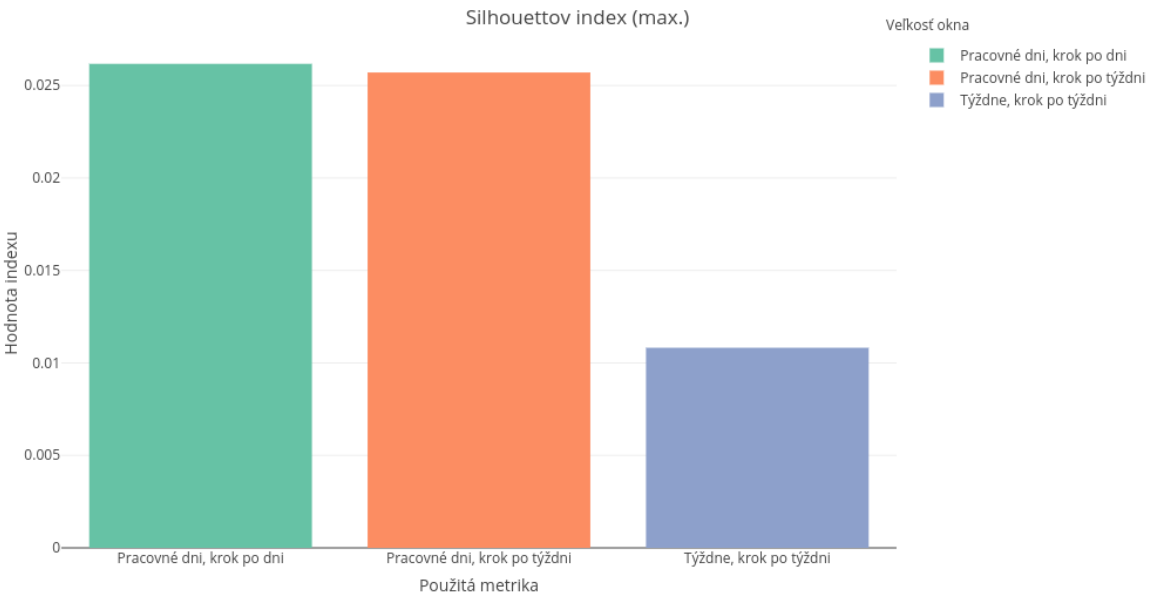
Obr. 43: Graf zhlukovania, porovnanie vzdialenostných metrík.



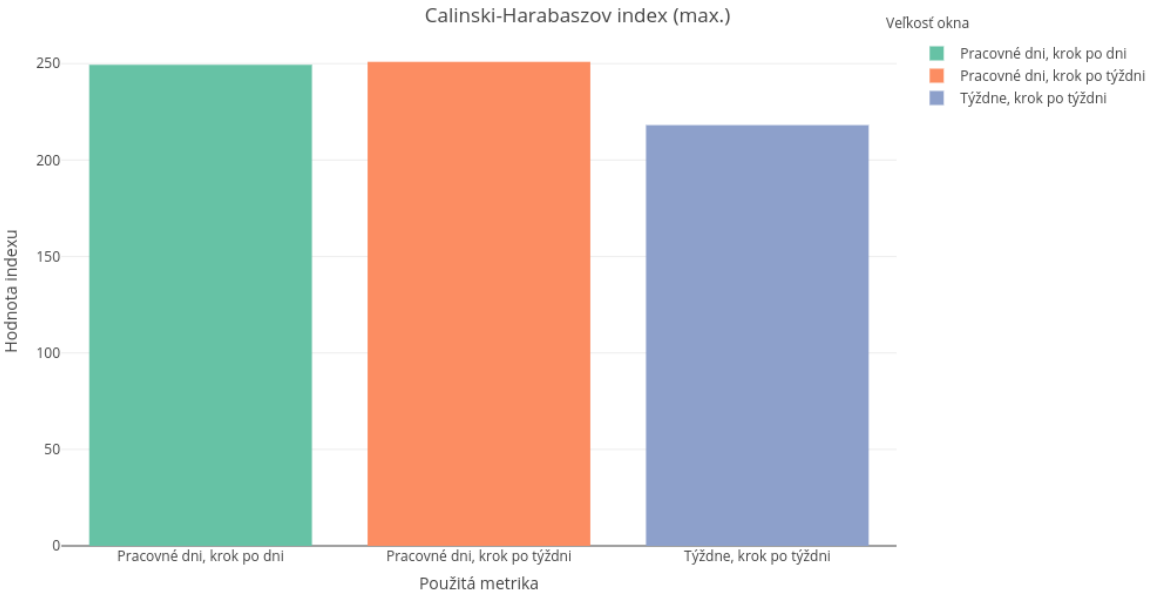
Obr. 44: Graf zhlukovania, porovnanie vzdialenostných metrík.



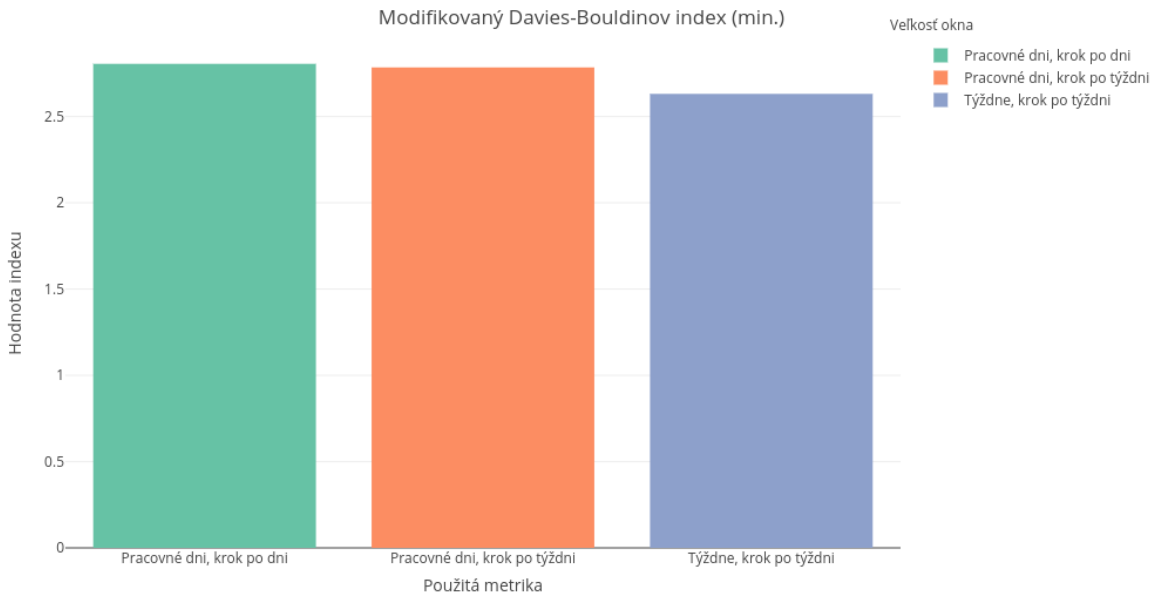
Obr. 45: Graf zhlukovania, porovnanie vzdialenostných metrík.



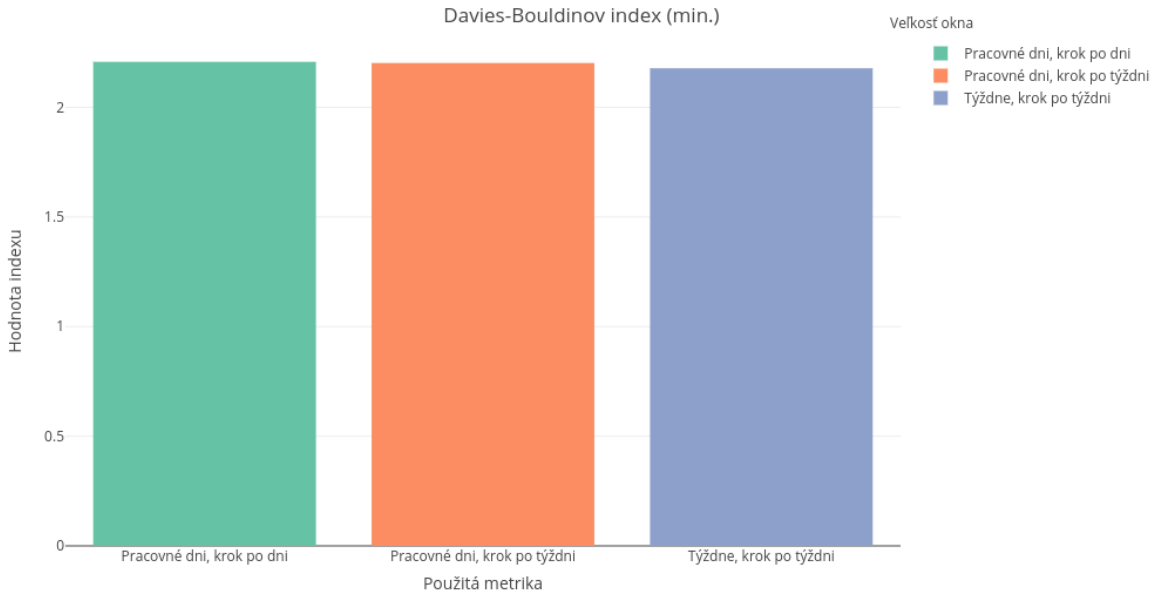
Obr. 46: Graf zhlukovania, porovnanie veľkostí a typov posuvných okien.



Obr. 47: Graf zhlukovania, porovnanie veľkostí a typov posuvných okien.



Obr. 48: Graf zhlukovania, porovnanie veľkostí a typov posuvných okien.



Obr. 49: Graf zhlukovania, porovnanie veľkostí a typov posuvných okien.