Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Evidenčné číslo: FIIT-XXXX-73688

Bc. Matúš Cuper

Identifikácia neštandardného správania odberateľov v energetickej sieti

Diplomová práca

Vedúci práce: Ing. Marek Lóderer

máj 2019

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Evidenčné číslo: FIIT-XXXX-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

máj 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 **1**

2 Analýza problému **2**
2.1 Časové rady . . . 2
2.1.1 Analýza časových radov . . . 2
2.1.2 Zložky časových radov . . . 3
2.1.3 Typy modelov časových radov . . . 5
2.1.4 Delenie časových radov . . . 6
2.2 Detekcia anomálií . . . 7
2.2.1 Typy anomálií . . . 8
2.2.2 Rozsah výskytu anomálií . . . 10
2.2.3 Prístupy k identifikácii anomálií . . . 10
2.2.4 Techniky detekcie anomálií . . . 11
2.3 Metódy zhlukovania časových radov . . . 12
2.3.1 Zhlukovanie na základe dočasnej susednosti . . . 12
2.3.2 Zhlukovanie na základe reprezentácie . . . 13
2.3.3 Zhlukovanie na základe modelu . . . 13
2.3.4 Ďalšie prístupy k zhlukovaniu . . . 14
2.3.5 Metriky vzdialenosti . . . 14
2.4 Predspracovanie dát . . . 17
2.4.1 Filtrovanie odberateľov . . . 17
2.4.2 Výber atribútov . . . 18
2.4.3 Extrakcia čŕt . . . 18
2.4.4 Agregácia dát . . . 18
2.4.5 Redukcia dát . . . 18
2.4.6 Segmentácia časových radov . . . 19
2.4.7 Normalizácia číselných vektorov . . . 19
2.5 Anomálie v energetických časových radoch . . . 19
2.6 Vyhodnocovacie metriky . . . 21
2.7 Súvisiace práce v doméne energetiky . . . 22
2.8 Existujúce riešenia . . . 23
2.8.1 Zhlukovanie časových radov . . . 23
2.8.2 Identifikácia anomálií . . . 23
2.8.3 Detekcia zlomov . . . 23

3 Návrh riešenia **24**

Dodatok A Plán do letného semestra **28**

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.

% TODO dopíš niečo o anomáliách a predpovediach časových radov
% TODO výsledky práce
% TODO členenie práce

2 Analýza problému

Tak ako je spomenuté v článku [12], 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 veličiny 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].

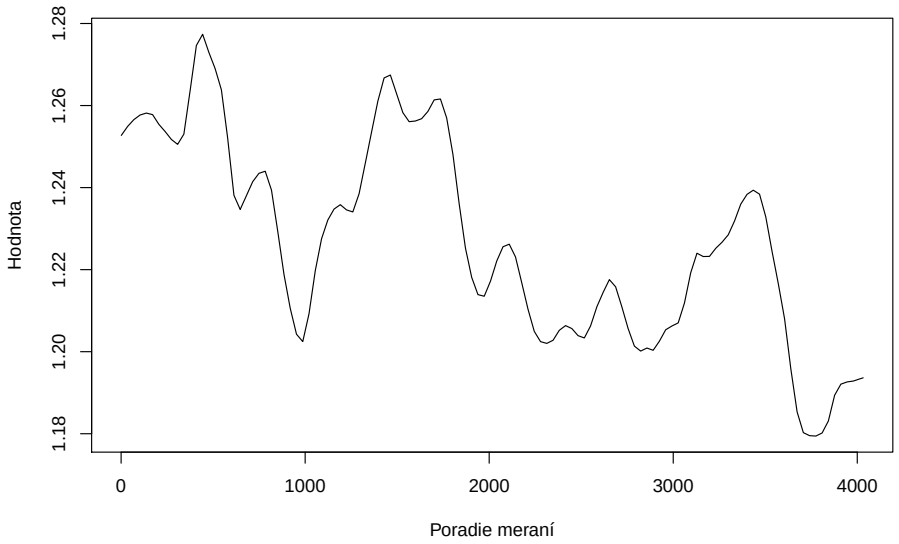
Premenná $\hat{x}$ vo vzorci 1 predstavuje predikovanú hodnotu časového radu x. Cieľom je nájsť funkciu $f(x)$, ktorá predikuje budúce hodnoty časového radu x tak, aby boli čo najpresnejšie, konzistentné a objektívne [19].

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

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 do 4 komponentov, a to trendová, cyklická, sezónna a reziduálna zložka [9].

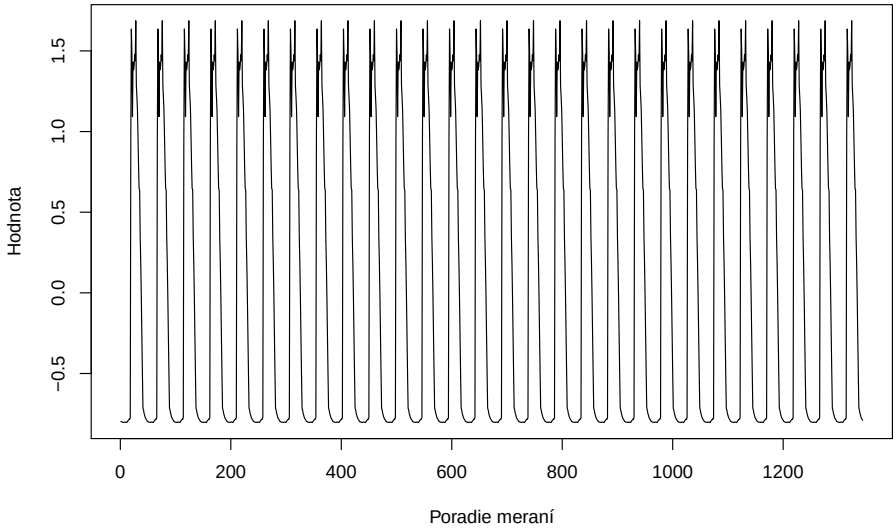
Trendová zložka zastupuje dlhodobé správanie časového radu. Ide o dlhodobé klesanie, rast alebo stagnáciu časového radu. Príkladom môže byť neustále predlžovanie priemernej doby dožitia alebo aj rast svetovej populácie. Priebeh dekompovanej trendovej zložky môžeme vidieť na obrázku 1 [1].



Obr. 1: Príklad trendovej zložky časového radu.

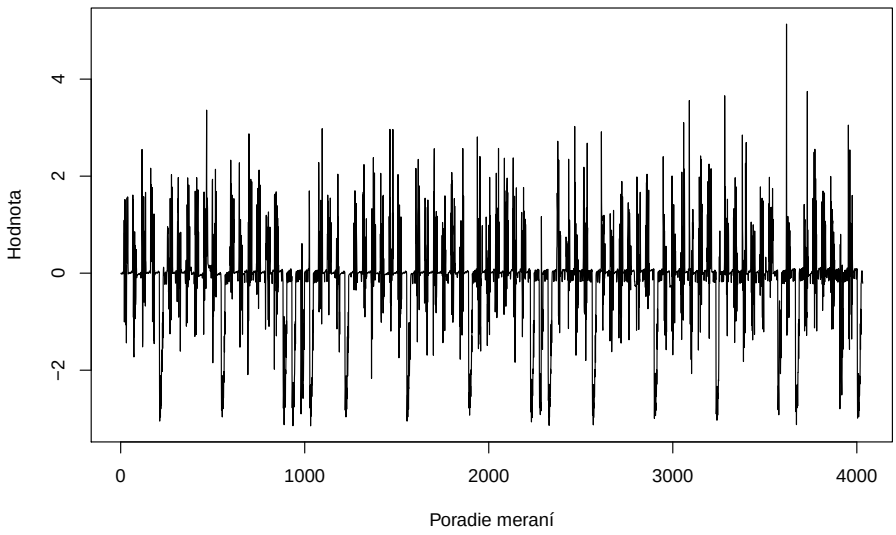
Cyklická zložka predstavuje strednodobú opakujúcu sa zmenu. Najčastejšie sa pri tom jedná o obdobie 2 a viac rokov. Táto zložka býva výrazne zastúpená pri ekonomických a finančných časových radoch. Príkladom môže byť aj podnikateľský cyklus, ktorý pozostáva zo 4 opakujúcich sa fáz [1].

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].



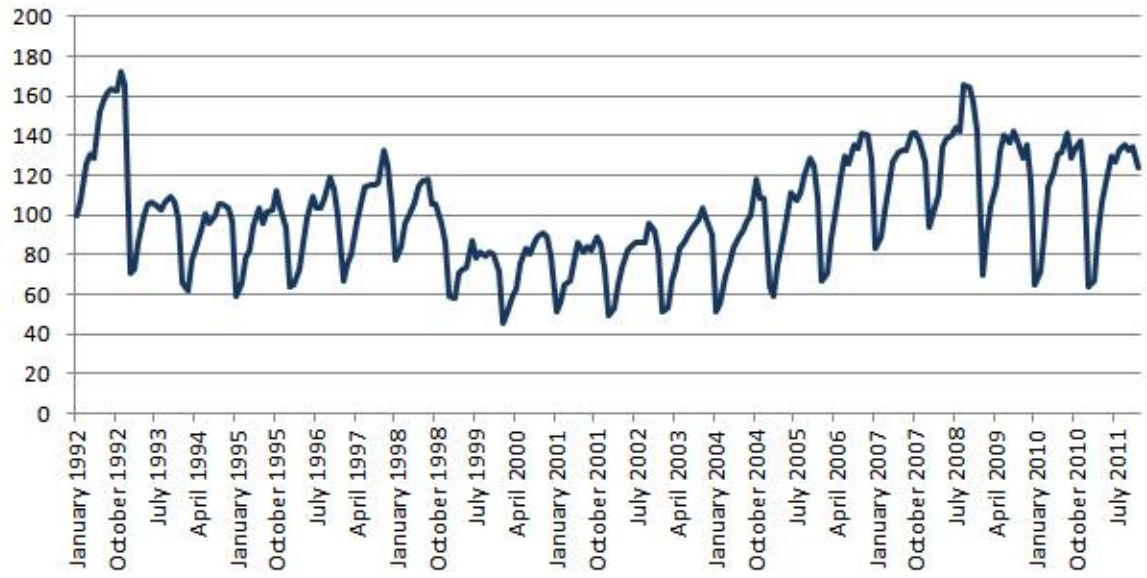
Obr. 3: Príklad reziduálnej zložky časového radu.

2.1.3 Typy modelov časových radov

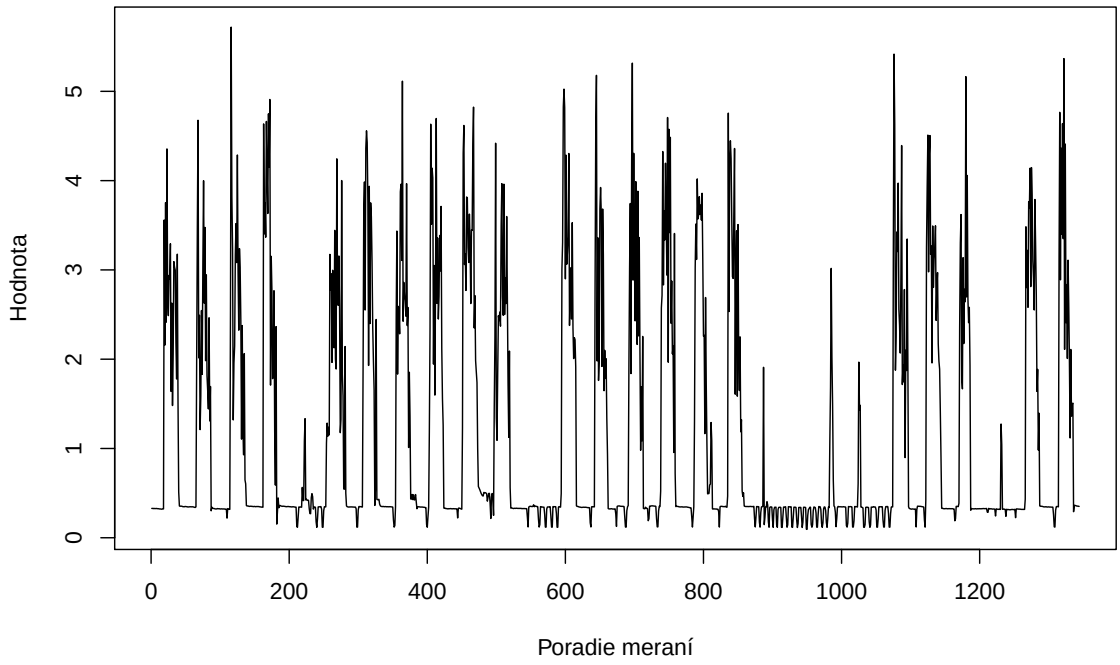
Kombináciou komponentov časových radov identifikovaných v predchádzajúcej kapitole vznikajú 2 typy modelov, aditívny a multiplikatívny.

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

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.

% TODO add some figures

2.1.4 Delenie časových radov

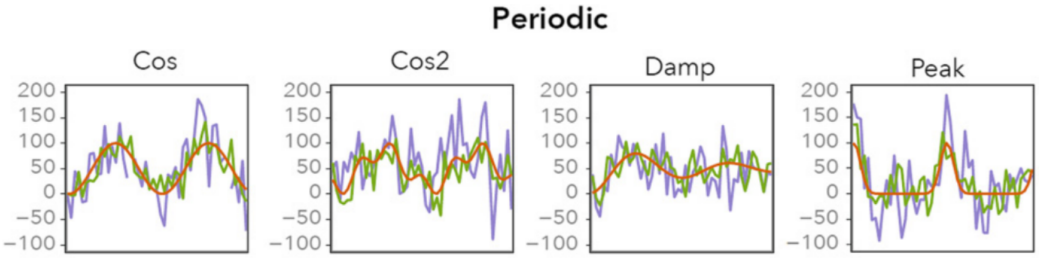
Výraznými vlastnosťami časových radov sú aj synchrónnosť a periodicita, znázornená na obrázkoch 6 a 7. Vznikajú tak 4 nasledujúce kategórie [25]:

- **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.

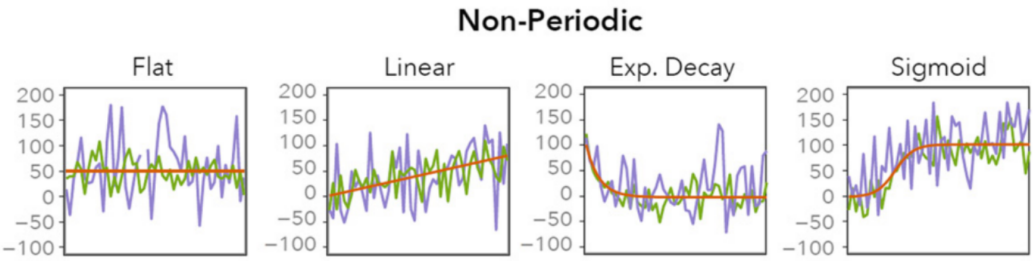
- **Neperiodické a synchrónne časové rady** nemajú žiadnu periodicitu, ale opäť sú časovo zarovnané.

- **Periodické a asynchrónne časové rady** nie sú časovo zarovnané, ale obsahujú periodicitu, čiže začiatok periódy v každom časovo 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 [15].



Obr. 7: Príklad neperiodických časových radov [15].

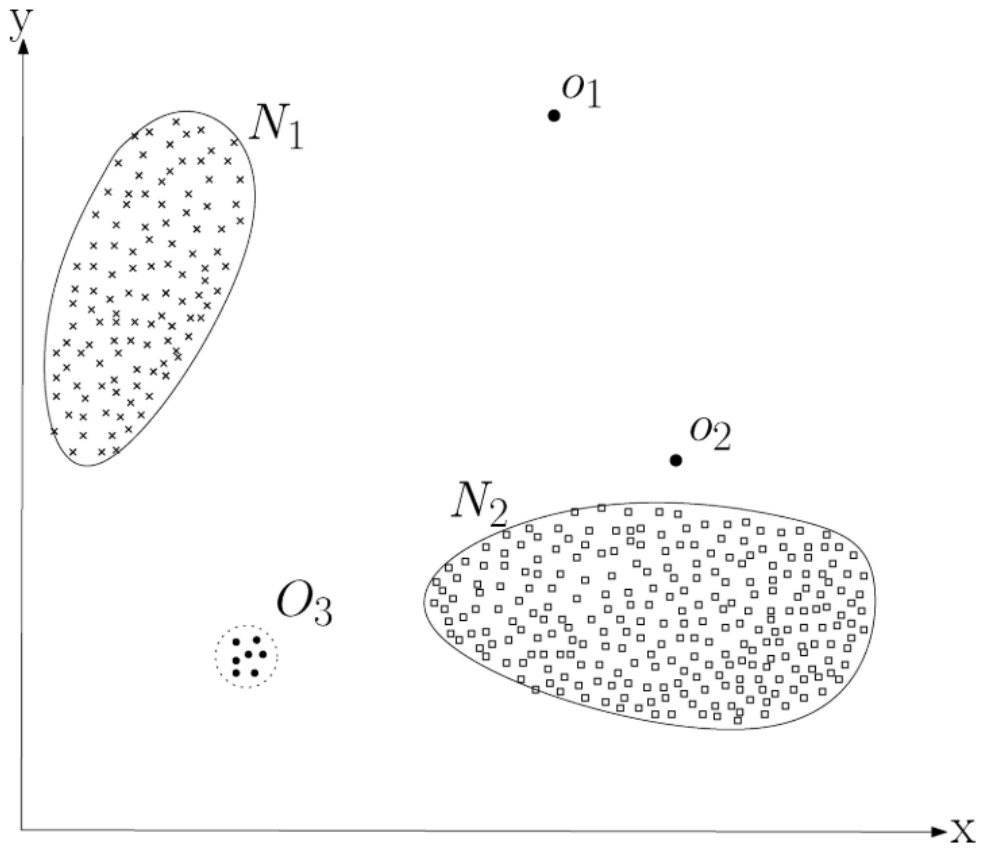
2.2 Detekcia anomálií

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 [3]:

- **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.
- **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*) [5].



Obr. 8: Príklad bodových anomálií [3].

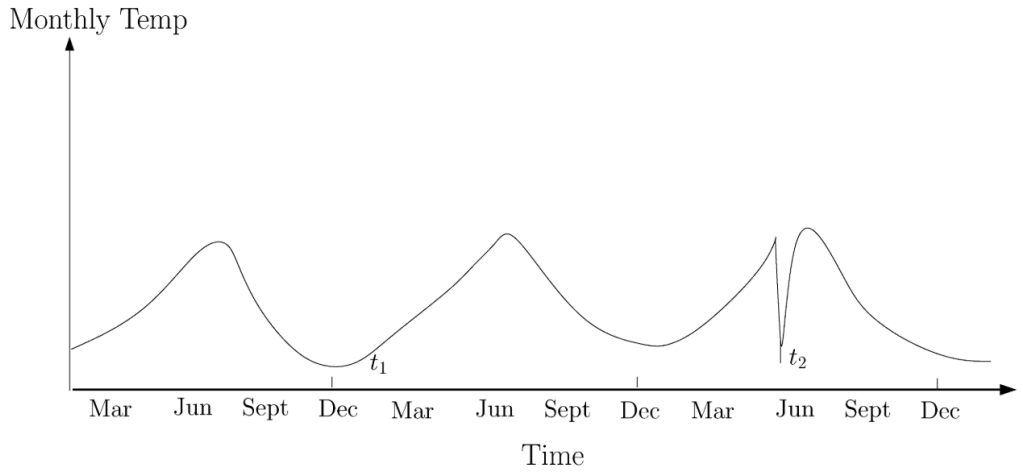
2.2.1 Typy anomálií

Dôležitým aspektom pri uplatnení detekcie anomálií je charakter anomálie. Z toho 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 [3].

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 [3].

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

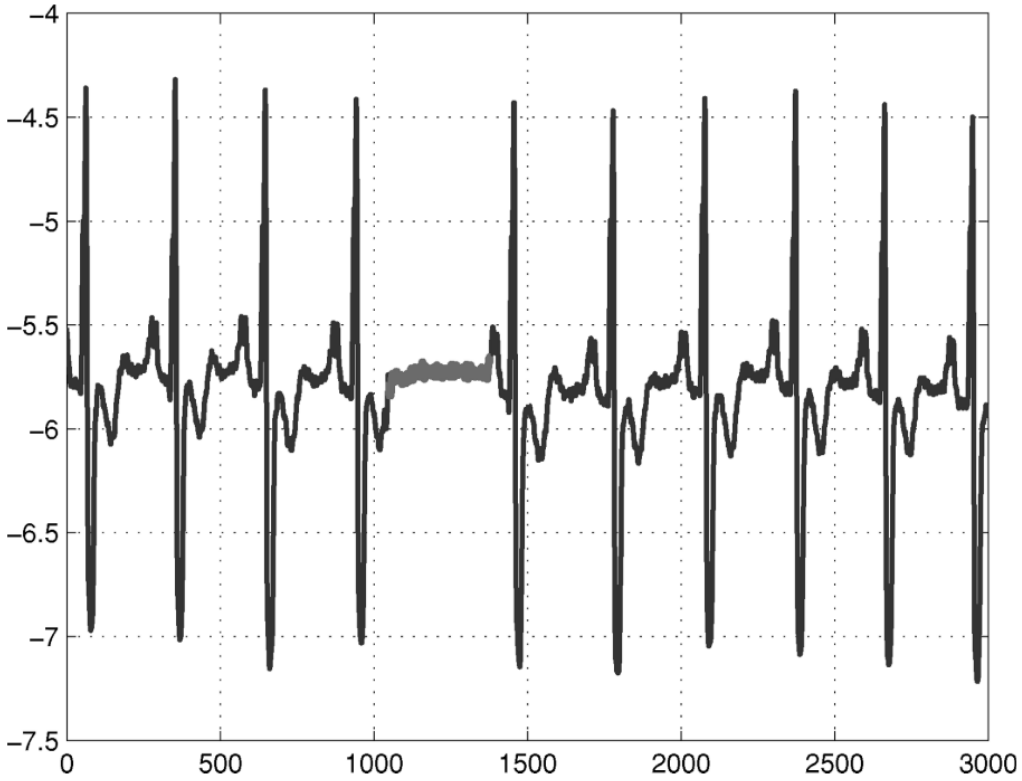


Obr. 9: Príklad kontextových anomálií [3].

môžu byť opäť transakcie väčšieho objemu peňazí, ktoré sú bežné v období pred Vianocami, ale neštandardné v inom ročnom období [3].

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 [3].

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 [3].



Obr. 10: Príklad skupinových anomálií [3].

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 [3].

2.2.2 Rozsah výskytu anomálií

% TODO add local anomalies
% TODO add global anomalies

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*) [3].

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, dochádza tak často k falošnému alarmu [3].

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, že anomálnych inštancií v porovnaní s normálnymi je omnoho menej a označenie dát ľudským expertom môže byť pri anomálnej inštancií náročné [3].

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 [3].

2.2.4 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.

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 [3, 24].

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 [3, 24].

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 [3, 24].

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 [3].

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 učiť aj pomocou kombinovaného učenia [3].

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 [18].

% TODO

K -means
% TODO

S -H-ESD

2.3 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 Euklidovskom priestore, môžu byť na zhlukovanie použité klasické techniky spomenuté v 2.2.4. Ak inštancie reprezentujú časový rad, nasadenie takýchto štandardných prístupov je zriedkavé [10].

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. Rozdelenie môžeme vizualizovať pomocou obrázka **??**. Prístupy sú opísané v nasledujúcich podkapitolách [16].

2.3.1 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 vystriedanie viacerých vzdialenostných alebo podobnostných metrík pre použité časové rady [16].

% TODO add figure from real dataset

Hierarchické zhlukovanie Táto metóda produkuje vnorenú hierarchiu skupín podobných časových radov na základe vzdialenostných matíc jednotlivých inštancií. 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 [7].

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, až potom ich postupne spájajú

do väčších zhlukov až 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 [27].

Aglomeratívne zhlukovanie Vzdialenosť medzi dvoma zhlukmi je meraná pomocou dvojice najbližších časových radov umiestnených v rôznych zhlukoch, ktoré sú potenciálnymi kandidátmi 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, až potom 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 [27].

Deliace zhlukovanie Algoritmus nie je obmedzený 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 2.3.5 [27].

% TODO Add k-means / dbscan
% TODO Add random swap

2.3.2 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 až následne sú aplikované zhlukovacie algoritmy. V literatúre sa zvykne označovať aj ako zhlukovanie na základe vlastností (angl. *Feature based clustering approach*) [16].

% TODO add equation ???

Samoorganizované mapy Trieda neurónových sietí, kde neuróny sú usporiadané v nízkodimenzionálnej štruktúre a trénované iteratívne a bez učiteľa. Trénovací proces začína pridelením náhodných hodnôt váhovým vektorom w. Každá iterácia trénovania pozostáva z 3 krokov a to náhodného výberu vstupného vektoru z trénovacej množiny, evaluácie siete a aktualizovaní váhových vektorov. Po natrénovaní je vypočítaná Euklidovská vzdialenosť medzi vstupným vzorom a váhovým vektorom. Následne je neurón s najmenšou vzdialenosťou označený ako t a ostatné váhy ostatných neurónov sú aktualizované v závislosti od vzdialenosti od neuróna t. Nevýhodou je opäť náročné spracovanie časových radov rôznych dĺžok, keďže dĺžka časového radu definuje aj dĺžku váhového vektora w [11, 27].

2.3.3 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 [16].

ARIMA V práci [29] autori navrhli metódu zhlukujúcu 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.

2.3.4 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 [25].

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 [25].

2.3.5 Metriky vzdialenosti

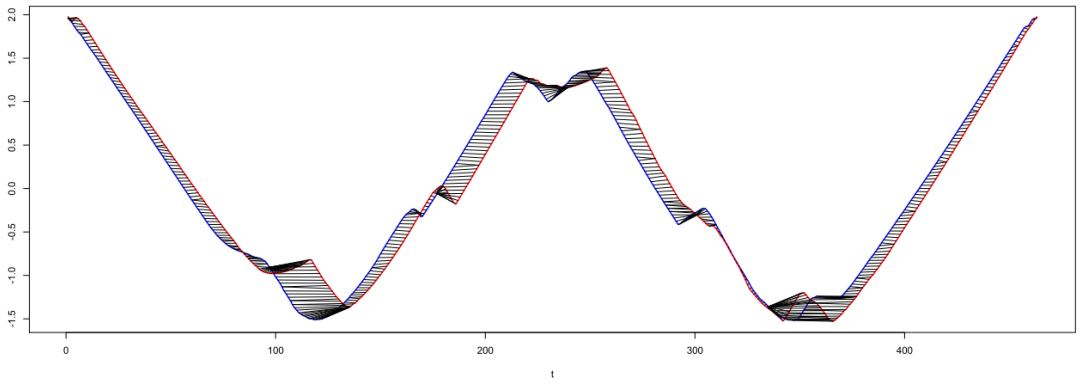
Kľúčovou záležitosťou pri zhlukovaní časových radov na základe ich podobnosti, je meranie vzdialenosti medzi nimi. Rovnako ako pri zhlukovaní bodových inštancií je potrebné definovať si metódy merania vzdialenosti. Najčastejšími metrikami sú Euklidovská a Manhattanovská vzdialenosť. Vhodnosť aplikovania 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 [7, 27].

% TODO explain different approaches add figure from real dataset
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. Euklidovská vzdialenosť medzi jednotlivými inštanciami. Modelové techniky používajú parametrický model, do ktorého vstupujú časové rady. Vzdialenosť je potom definovaná ako vzdialenosť medzi jednotlivými modelmi. Metódy porovnávajúce tvary kriviek sa snažia prispôsobiť výsledný tvar časového radu nelineárnym rozťahovaním a kontrakciou časových osí [10].

Korelácia Korelačný koeficient $r(X, Y)$ meria stupeň lineárnej závislosti medzi dvoma časovými radmi X a Y. Vyjadríme ho vzorcom 3.

$$r\left(X,Y\right) = \frac{E\left[\left(X - E\left[X\right]\right) \cdot \left(Y - E\left[Y\right]\right)\right]}{E\left[\left(X - E\left[X\right]\right)^2\right] \cdot E\left[\left(Y - E\left[Y\right]\right)^2\right]} \tag{3}$$

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 4. Nevýhodou je, ak máme k dispozícií iba malú, prípadne krátku časť datasetu, podobnosť touto metrikou sa určuje



Obr. 11: Príklad porovnávania časových radov pomocou dynamickej deformácií času [20].

len ťažko. Keďže korelácia zachytáva iba lineárnu podobnosť časových radov, pri aplikovaní metriky na dva nelineárne podobné časové rady, sú vyhodnotené ako vzdialené [7].

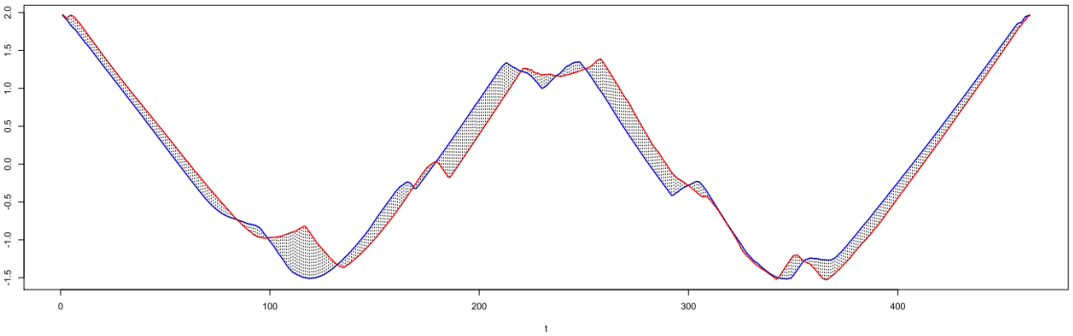
$$D_r\left(X,Y\right) = \sqrt{0.5 \cdot \left(1 - r\left(X,Y\right)\right)} \tag{4}$$

Dynamické deformovanie času Ide o 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 11 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 5 [7, 8]. % TODO add ref from Flexible Dynamic Time Warping for Time Series Classification

$$D_{DTW}\left(i,j\right) = \begin{cases} d\left(x_i, y_j\right) + min \begin{cases} D_{DTW}\left(i-1,j\right) \\ D_{DTW}\left(i,j-1\right) \text{ ak } i \neq 0 \text{ a } j \neq 0 \\ D_{DTW}\left(i-1,j-1\right) \end{cases} \\ 0 \text{ ak } i = 0 \text{ a } j = 0 \\ \infty \text{ inak} \end{cases} \tag{5}$$

% TODO Add GAK
% TODO Add SBD

Kvalitatívna vzdialenosť Metóda je 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 6, 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



Obr. 12: Príklad porovnávania časových radov pomocou Euklidovskej vzdialenosti [20].

v prípade, kedy neexistuje medzi časovými radmi lineárna alebo nelineárna závislosť [7].

$$D_q(X,Y) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2 \cdot Diff\left(q(X_i, X_j), q(Y_i, Y_j)\right)}{N \cdot (N-1)} \tag{6}$$

% TODO compare euclidian and manhattan distance metrics

Euklidovská vzdialenosť Je používaná najmä pri klasických zhlukovacích problémoch. Ak zvolený časový rad má dĺžku n, vzdialenosť vypočítame vzorcom 7. Na obrázku 12 sú vždy porovnávané hodnoty vyskytujúce sa v rovnakom časte t [27].

$$D_E(X,Y) = \sqrt{\sum_{k=1}^{n} (X_{ik} - Y_{jk})^2} \tag{7}$$

Manhattanovská vzdialenosť Je rovnako ako Euklidovská vzdialenosť používaná najmä pri klasických zhlukovacích problémoch. Výpočet je tiež veľmi podobný, môžeme ho vyjadriť vzorcom 8 % TODO add som refference

$$D_M(X,Y) = \sum_{k=1}^{n} |X_{ik} - Y_{jk}| \tag{8}$$

Pearsonov korelačný koeficient Vo vzorci 9 reprezentuje $\widetilde{X}$ aritmetický priemer časového radu X. Koeficient je používaný pri výpočte vzdialenosti, ktorá je založená na vzájomnej korelácií. Vzdialenosť vyjadríme vzorcom 9 [27].

$$r(X,Y) = \frac{\sum_{i=1}^{n} \left(X_i - \widetilde{X}\right) \cdot \left(Y_i - \widetilde{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \widetilde{X}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} \left(Y_i - \widetilde{Y}\right)^2}} \tag{9}$$

$$D_P(X,Y) = 2 \cdot (1 - r(X,Y)) \tag{10}$$

Vzdialenosť medzi krátkymi časovými radmi Metóda (angl. *Short time series*) meria vzdialenosť ako sumu štvorcových rozdielov medzi krivkami jednotlivých časových radov. Na odstránenie nežiaducich efektov škály sa používa z štandardizácia. Matematicky vzdialenosť vyjadrime vzorcom 11. Zložka t_k predstavuje čas [27].

$$D_{STS}\left(X,Y\right) = \sqrt{\sum_{k=1}^{n}\left(\frac{Y_{j(k+1)} - Y_{jk}}{t_{(k+1)} - t_k} - \frac{X_{i(k+1)} - X_{ik}}{t_{(k+1)} - t_k}\right)^2} \tag{11}$$

2.4 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 [22].

2.4.1 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 [13].

2.4.2 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 [4].

2.4.3 Extrakcia čŕt

Ďalšou technikou používanou pri príprave dát je tvorba nových atribútov založených na pôvodných, surových dátach. V súvisiacom článku [13] 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 12

$$x_h = \frac{P_{h+1}}{D_{h+1} - D_h} \tag{12}$$

2.4.4 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 [4].

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 anomálne aktivite odberateľa, čo je nutné zvážiť pri konkrétnej implementácii.

2.4.5 Redukcia dát

% TODO add figure
Jednou z najjednoduchších metód používaných pri redukcií dát je práve vzorkovanie (angl. *sampling*). Parametrami sú m a n, ktoré predstavujú počet dimenzií pred a po procese vzorkovania. Vzdialenosť medzi jednotlivými sa zväčšuje, no medzi všetkými inštanciami je rovnaká. Nevýhodou je, že tvar výsledného časového radu je oproti pôvodnému skreslený [8].

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, kedy vzniknuté intervaly majú rôznu dĺžku, v závislosti od tvaru časového radu [2].

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 [8].

% TODO add figure

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 [8].

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 [8].

Ďalšou metódou je analýza hlavných komponentov (angl. *principal component analysis*), skrátene PCA. Obvykle sa PCA používa na elimináciu menej významných komponentov, čím sa znižuje dimenzionalita dát. Metóda má uplatnenie aj pri analýze či vizualizácií vysokodimenzionálnych dát [8].

Podobným problémom ako DTW je 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 [8].

2.4.6 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 [8].

2.4.7 Normalizácia číselných vektorov

%TODO add znorm

2.5 Anomálie v energetických časových radoch

V distribučných sieťach vznikajú straty, ktoré vo všeobecnosti môžeme rozdeliť na technické a netechnické straty. 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 [5, 17].

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 [21].

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 identifikácií 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 [17]. Tak ako je spomenuté v práci [6], 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 sa 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 [14].

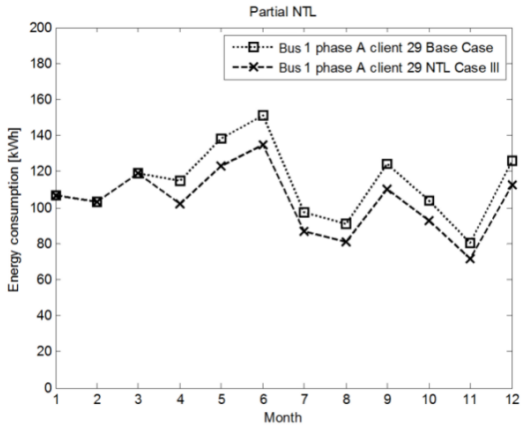
V súvisiacich prácach sa autori zaoberali určením netechnických strát v elektrických distribučných sieťach s použitím rôznych štatistických metód alebo strojového učenia. Dostupné dáta od distribútorov pochádzali najmä z jedného zdroja, lokality a zameriavali sa na jeden zdroj energie. Dáta, ktoré budeme mať k dispozícií disponujú podobnými vlastnosťami. V súvisiacej práci [5] 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 počtom ľudí, miestností prípadne ich funkcie alebo zvýšením energetickej sebestačnosti. Následkom je nižšia nameraná spotrebe energie v dlhšom horizonte. Zníženie spotreby môže byť čiastočné alebo úplne, ako môžeme vidieť na obrázkoch 13 a 14 [21, 26].

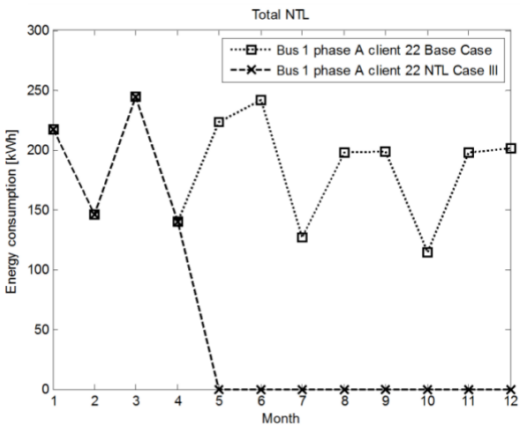
Z pohľadu výskytu anomálie môžu nastať nasledovné scenáre:

- 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.



Obr. 13: Čiastočné zníženie spotreby elektrickej energie [26].



Obr. 14: Úplné zníženie spotreby elektrickej energie [26].

- Anomália vznikne v priebehu sledovaného intervalu a nie je odhalená. Táto skupina ja predmetom celej 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 [21].

% TODO add confusion matrix table

2.6 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 čierneho 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 označený normálny odberatelia (angl. *TRUE NEGATIVE*) a správne označený normálny odberatelia (angl. *FALSE NEGATIVE*). Kvalitu klasifikácie potom môžeme zmerať pomocou presnosti a pokrytia. Presnosť vypočítame vzorcom 13, kedy ide o pomer správne označených anomálií a celkový počet označených anomálií. Tým vypočítame percento odberateľov, ktorých sme správne klasifikovali ako podozrivých.

$$\text{Presnosť} = \frac{TP}{TP + FP} \tag{13}$$

Pokrytie označuje pomer správne označených anomálií a celkový počet skutočných anomálií. Vyjadríme ju pomocou vzorca 14.

$$\text{Pokrytie} = \frac{TP}{TP + FN} \tag{14}$$

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 [26, 28].

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 15, kde P predstavuje presnosť a C predstavuje pokrytie [26].

$$F = 2 \cdot \left(P^{-1} + C^{-1}\right)^{-1} \tag{15}$$

% TODO add evaluation with syntetic dataset
% TODO add evaluation with clustering TS

2.7 Súvisiace práce v doméne energetiky

V [10] bola pri zhlukovaní použitá aj kombinácia viacerých metód, konkrétne k-means, metóda náhodnej výmeny a aglomeratívne zhlukovanie. Ako už bolo spomenuté v 2.2.4, úlohou algoritmu k-means namapovať existujúce inštancie do *k* zhlukov. Aj keď metóda náhodnej výmeny je obmedzená na zhlukovacie problémy v Euklidovskom 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 2.3.1 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 [4] 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 [23] 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 nevýrazne zmeny a trendy v dátach.

2.8 Existujúce riešenia

fdsafsd

2.8.1 Zhlukovanie časových radov

fdfd

2.8.2 Identifikácia anomálií

fsd

2.8.3 Detekcia zlomov

fds

https://github.com/robjhyndman/anomalous-acm https://sites.google.com/site/andreaventurini65/hon
detection https://cran.r-project.org/web/packages/tsoutliers/tsoutliers.pdf https://cran.r-project.org/web/pa
https://cran.r-project.org/web/packages/rainbow/rainbow.pdf https://cran.r-project.org/web/packages/kmo

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, ktoré sú sviatky.

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.

Zvýšiť presnosť odhaľovania anomálií je možné aj osobitným prístupom k jednotlivým odberateľom. Práve vďaka identifikácií zlomov v časových radoch je možné presnejšie určiť správanie odberateľa. Zároveň algoritmus poskytuje ďalší mechanizmus na určovanie intervalu, v ktorom sa mení správanie odberateľa, ktoré môžeme považovať za anomáliu.

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 a môže byť predmetom záujmu distribútorov elektrickej energie.

Literatúra

[1] Adhikari, R.: *An Introductory Study on Time Series Modeling and Forecasting*. Saarbrucken: LAP LAMBERT Academic Publishing, 2013, ISBN 9783659335082.

[2] 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`

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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`

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

[10] Hautamaki, V.; Nykanen, P.; Franti, P.: Time-series clustering by approximate prototypes. *2008 19th International Conference on Pattern Recognition*, 2008: s. 1–4, ISSN 1051-4651, doi:10.1109/ICPR.2008.4761105.
URL `http://ieeexplore.ieee.org/document/4761105/`

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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

[16] 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

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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

[23] Tagaris, H.; Lachsz, A.; Jeffrey, M.: Wavelet based feature extraction and multiple classifiers for electricity fraud detection. *IEEE/PES Transmission and Distribution Conference and Exhibition*, ročník 3, č. November 2002, 2002: s. 2251–2256, doi:10.1109/TDC.2002.1177814.

URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=1177814'escapeXml='false'/{%}3E

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

[25] Teng, M.: Anomaly detection on time series. *2010 IEEE International Conference on Progress in Informatics and Computing*, ročník 1, 2010: s. 603–608, doi:10.1109/PIC.2010.5687485, 1708.02975.
URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=5687485

[26] 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.

[27] 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.

[28] 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

[29] 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 Plán do letného semestra

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

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