FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Bakalářská práce

Analýza dat pro správu chytrých budov

Studijní program: B0613A140005 Informační technologie
Studijní obor: Aplikovaná informatika

Autor práce: **David Jansa**
Vedoucí práce: Ing. Jan Kraus, Ph.D.
Ústav mechatroniky a technické informatiky

Liberec 2023

FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Zadání bakalářské práce

Analýza dat pro správu chytrých budov

Jméno a příjmení:	**David Jansa**
Osobní číslo:	M20000063
Studijní program:	B0613A140005 Informační technologie
Specializace:	Aplikovaná informatika
Zadávající katedra:	Ústav mechatroniky a technické informatiky
Akademický rok:	2022/2023

Zásady pro vypracování:

1. Seznamte se s metodami pro zpracování archivních dat z měření veličin v chytrých budovách – shluková analýza (klastrování), hledání podobných intervalů v různých časech, detekce odlehlých hodnot a anomálií, doplnění chybějících záznamů, predikce apod.
2. S využitím existujících nástrojů pro prostředí .NET navrhněte a realizujte vlastní výpočetní moduly pro vybrané úlohy z bodu 1 zadání.
3. Vytvořený software otestujte na rozsáhlejších množinách měřených posloupností dat a ověřte jejich správnou a přiměřeně efektivní funkci.
4. V textové části práce shrňte předepsanou formou dosažené výsledky, uveďte přehledně a srozumitelně hlavní přednosti a nedostatky zvolených alternativ, dále uveďte možnosti praktického využití vytvořených modulů a diskutujte příležitosti dalšího rozvoje tématu.

Rozsah grafických prací:	dle potřeby dokumentace
Rozsah pracovní zprávy:	30-40 stran
Forma zpracování práce:	tištěná/elektronická
Jazyk práce:	Čeština

Seznam odborné literatury:

[1] RUIZ, Luis G. Baca, et al. A time-series clustering methodology for knowledge extraction in energy consumption data. *Expert Systems with Applications*, 2020, 160: 113731.

[2] SEEM, John E. Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy and buildings*, 2007, 39.1: 52-58.

[3] KHAN, Imran, et al. Fault detection analysis of building energy consumption using data mining techniques. *Energy Procedia*, 2013, 42: 557-566.

[4] MOLINA-SOLANA, Miguel, et al. Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews*, 2017, 70: 598-609.

Vedoucí práce:	Ing. Jan Kraus, Ph.D. Ústav mechatroniky a technické informatiky
Datum zadání práce:	12. října 2022
Předpokládaný termín odevzdání:	15. května 2023

L.S.

prof. Ing. Zdeněk Plíva, Ph.D.
děkan

doc. Ing. Josef Černohorský, Ph.D.
vedoucí ústavu

V Liberci dne 12. října 2022

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědom toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

22. května 2023 David Jansa

Analýza dat pro správu chytrých budov

Abstrakt

Cílem této bakalářské práce je otestovat vybrané metody pro analýzu archivních dat chytrých budov. Základní přístupy analýzy, tj. klasifikace, regrese, detekce anomálií, shluková analýza a analýza časových řad jsou v této práci popsány s příklady různých algoritmů a jejich možnými úskalími při aplikaci. Větší pozornost je věnována primárně metodám pro vyhledávání vzorců chování, specificky metodám symbolic aggregate approximation a shlukování metodou nejbližších středů. Hlavně zmíněné metodě shlukování je věnována patřičná pozornost zvláště při popisu popisných a vzdálenostních metrik. V případě vzdálenostních metrik je část rešerše také věnována metodě dynamic time warping, hojně využívané v oblasti časových řad.

V praktické části jsou tyto vybrané metody aplikovány na archivní data měření spotřeby energie rodinného domu. Před otestováním metod jsou data podrobena explorační datové analýze a v závislosti na zjištěných vztazích je následně provedena analýza za pomocí zmíněných metod. Práce se zabývá primárně analýzou dat a porovnáním výsledků jednotlivých metod za použití různých parametrů.

Klíčová slova: chytré budovy, analýza časových řad, shlukování metodou nejbližších středů, dynamic time warping, symbolic aggregate approximation

Data analysis for building energy management

Abstract

The goal of this bachelor thesis is to test chosen methods for analyzing archived data from smart buildings. The pivotal analytic aproach that are classification, regression, anomaly detection, cluster analysis and time series analysis are described in this work with examples of various algorithms and their advantages in their aplication. More attention was paid primarily to methods that look for pattern recognition, specifically methods symbolic aggregate approximation and k-means clustering. Primarily the previously mentioned clustering method is being paid due attention, particularly in describing descriptive and distance metrics. In the case of distance metrics, segment of the research is designated to dynamic time warping, that is abundantly used in the field of time series analysis.

In the practical part, these methods are aplied on archival data of energy consumption measurements in a single-family house. Before testing these methods, the data is thoroughly subjected to exploratory data analysis and dependent on the discovered relationships, an analysis with aforementioned methods is made. This thesis mainly looks into data analysis and comparing the results of each individual method with the use of distinct paramaters.

Keywords: smart buildings, time series analysis, k-means, dynamic time warping, symbolic aggregate approximation

Poděkování

Děkuji vedoucímu bakalářské práce Ing. Janu Krausovi, Ph.D. za cenné rady v oblasti energy management systémů.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

IB	Inteligentní Budovy
OZ	Obnovitelné Zdroje
HVAC	Heating, Ventilation, and Air Conditioning
IS	Inteligentní Sítě
IOT	Internet Of Things
SAX	Symbolic Aggregate approXimation
DTW	Dynamic Time Warping

Jednotky

ms	Milisekunda
W	Watt

1 Úvod

Cílem této bakalářské práce je seznámit se s metodami zpracování archivních dat z měření veličin v chytrých budovách, implementovat vybrané metody a aplikovat je na reálná data.

První část práce je věnována problematice chytrých budov. Inteligentní budovy jsou zde popsány z hlediska energy managementu v závislosti na stanovisku Evropské unie, mj. z hlediska úspor, efektivního využívání elektrické energie a rostoucího zastoupení ekologických elektráren. Kromě toho je zde popsána definice samotných chytrých budov za pomocí základních vlastností, které by tyto budovy měly splňovat.

Druhá část práce se zaměřuje na studium existujících metod pro analýzu dat v kontextu chytrých budov a energy managementu. Přesněji jsou zde popsány úlohy klasifikace, regrese, detekce anomálií, shlukové analýzy a analýzy časových řad. Pozornost je věnována nejpoužívanějším metodám v těchto oblastech. Tyto metody jsou stručně popsány společně s jejich hlavními přednostmi a úskalími při použití. Podrobněji se práce věnuje metodám shlukování metodou nejbližších středů a symbolic aggregate approximation, které jsou implementovány v prostředí .NET společně s různými metrikami a přístupy pro vizualizaci dat.

V praktické části jsou výše zmíněné metody aplikovány na roční data měření spotřeby energie rodinného domu. Před použitím obou metod je provedena explorační analýza dat. Metoda shlukování je otestována na datech za použití různých hodnot nastavených parametrů. Část každého cyklu testování je věnována analýze nastavení jednotlivých parametrů. Parametry s nejlepšími výsledky jsou poté použity pro finální zpracování dat. Metoda symbolic aggregate approximation je otestována převážně z hlediska použití při explorační analýze dat, kde jsou konstatovány její možné způsoby použití v kombinaci s vizualizacemi dat. Výsledky jednotlivých cyklů testování pro obě metody jsou mezi sebou porovnány a jsou diskutovány dosažené výsledky a případné přednosti a nedostatky použitých metod. V průběhu praktické části je kladen důraz na nastavení parametrů obou metod s přihlédnutím k úrovni získaných informací při použití.

V závěru práce jsou shrnuty dosažené výsledky a hlavní přednosti a nedostatky otestovaných metod. Jsou zde uvedeny možnosti praktického použití modulu zahrnující vybrané metody, metriky a vizualizace v rámci energy management systémů.

2 Problematika chytrých budov

Inteligentní systémy se stále více stávají neoddělitelnou součástí nejrůznějších typů odvětví, od elektrotechnického průmyslu přes automobilový průmysl až po energetiku. Zvláště poslední zmíněná energetika je do hloubky spjatá s ostatními typy průmyslu a to v kontextu ať už výroby, distribuce či úspory samotné energie. Výjimkou není ani stavebnictví, které prošlo v souvislosti se spotřebou elektrické energie a stále větší integrací obnovitelných zdrojů (OZ) energie dlouhou řadou změn, jež postupně definovaly koncepty, jako nízkoenergetické, pasivní nebo nulové domy.

V podobném kontextu je definována koncepce chytrých či inteligentních budov (IB), jichž metodologie v sobě propojuje teze OZ, úspory energie, chytrých systémů a v neposlední řadě uživatelského komfortu.

2.1 Proč se věnovat chytrým budovám

Budovy byly a stále jsou jedním z hlavních témat diskuzí o úspoře, efektivního využívání elektrické energie, změně klimatu i inovacích. V rámci Evropské unie se jedná o jeden z deseti pilířů v oblasti pro strategické investice hlavně v souvislosti s dekarbonizací výroby elektřiny. V Evropské unii spotřeba budov v roce 2019 představovala celých 40 % celkové spotřeby energie. Evropská unie ve smyslu snížení spotřeby energie, která úzce souvisí s ekologií, již dříve zveřejnila několikrát přepracované směrnice či doporučení. Například v roce 2010 definovala nový pojem – budovy s téměř nulovou spotřebou energie (nearly zero energy buildings). Jedná se o budovy, jejichž energetická náročnost je velmi nízká. Tato minimální spotřeba by měla být z většiny rozsahu pokryta energií z OZ s důrazem na energii, která byla vyrobena v místě spotřeby či v jeho blízkosti. Zároveň se tyto směrnice věnují například strategiím snížení emisí skleníkových plynů cestou renovace starších budov, do roku 2020 bylo za cíl snížit emise o 20 % oproti roku 1990 a dlouhodobější cíl vztahující se k roku 2050 nastavuje tuto hodnotu na 80–95 % či hovoří o úplné dekarbonizaci.[1, 2]

V souvislosti s čistou energií je vhodné uvést, že v Evropské unii z celkové spotřebované energie v roce 2021 pocházelo 21,8 % z OZ, což je více než dvojnásobek v porovnání s hodnotami z roku 2004. Z celkové čisté energie bylo spotřebováno 22,9 % systémy vytápění, ventilace a klimatizace (HVAC) v budo-

vách. Rostoucí zastoupení ekologických elektráren, a to zvláště z hlediska budov, způsobilo určité potíže v oblasti energy managementu z důvodu nekonzistence a decentralizovanosti OZ. Změnilo se uvažování nad samotným managementem, kdy je z hlediska budov vhodnější, aby spotřeba a generování energie bylo řízeno na úrovni budov, což implikuje zapojení chytrých měřáků a inteligentních sítí (IS).[1, 3]

Paralelně s těmito okolnostmi se objevuje zvýšená poptávka po budovách s interaktivním řízením či chováním za cílem zvýšit uživatelský komfort a snížit spotřebu energie. Také se zvyšují nároky na tyto budovy v souvislosti s distribuční sítí, kdy je požadována určitá adaptace na aktuální stav sítě, tj. aby systém reagoval na cenu energie či výpadky v síti. Za této situace budovy prochází přeměnou od necitelných k vysoce efektivním v oblastech spotřeby, výroby, ukládání a dodávky energie.[1]

Toto společensko-ekonomické pozadí vytváří příhodné prostředí pro koncept, který označujeme termínem chytré či inteligentní budovy. Přestože dodnes neexistuje přesná definice, většina odborných rešerší se shoduje ve specifických vlastnostech.

2.2 Definice chytrých budov

IB lze popsat za pomoci 4 základních schopností, které by tyto technologie měly splňovat. Mezi tyto schopnosti patří:

- Schopnost reagovat na klima. Schopnost budovy aktivně odpovídat na lokální externí klimatické podmínky, tj. identifikovat operační profil, a to v čase aktuálním i budoucím. Data o klimatu lze získat a použít za pomoci propojení internetu věcí (IOT) a řídícího systému. Tato schopnost dokáže významně ovlivnit, například: HVAC, osvětlení či stínící systém.[1]

- Schopnost reagovat na síť. Schopnost budovy reagovat na informace přijaté ze sítě. Hlavním cílem je maximalizovat efektivní využívání energie, což následně zapříčiní snížení výše záloh za energie či omezení zátěže distribuční sítě. Hlavním prvkem této funkce jsou IS, které zajišťují obousměrnou komunikaci mezi zákazníky a dodavateli energie. Neméně důležitou roli také hrají chytré měřáky a specializované procesory.[1]

- Schopnost reagovat na uživatele. Umožnit budově interagovat s uživatelem a chytrými technologiemi v reálném čase. Důraz je zde kladen na snahu vytvořit komfortní prostředí pro uživatele tím, že splní jeho požadavky.

- Schopnost monitoringu a kontroly. Schopnost monitorovat real-time i historické operace budovy nebo jejích přidružených systémů a chování uživatele. Cílem této funkce je ulehčit fungování prvních tří zmíněných schopností za pomoci aplikace výpočetních metod, například: predikce, shlukové analýzy či detekce anomálií, na nashromážděná data.[1]

S těmito základními schopnostmi souvisí i další pokročilé požadavky. Ze strany Evropské unie to může být například cíl, přiblížit IB směrem k formátu téměř nulových budov, integraci pasivních strategií ve stavebnictví, aplikací energeticky efektivních technologií v oblasti HVAC a osvětlení, kde tyto segmenty dohromady tvoří až 70 % celkové spotřeby v budovách, nebo přes rozsáhlejší integraci OZ. Schopnost reagovat na síť souvisí s požadavkem, aby IB byly flexibilní v rámci energy managementu, tj. aby zvládaly řídit vlastní vyrobenou energie a poptávku po energii.[1, 2]

3 Metody pro zpracování dat

Za účelem složitější a hlubší analýzy dat v IB byly implementovány a testovány různé metody pro zpracování naměřených dat. Statistickými modely, data mining metodami a data science je poté možné vytěžit netriviální informace z naměřených vzorků a tyto získané poznatky následně použít při rozhodování v dílčích aspektech energy managementu.

V oblasti energy managementu budov se, mimo jiné aktuálně objevují tři největší výzvy pro analýzu:

- Predikce poptávky po energii.
- Analýza chování přístrojů a uživatelů budov.
- Detekce opakujících se vzorců v měření spotřeby energie.

Následující podkapitoly shrnují základní i pokročilé metody, které se využívají ke zpracování dat a ke splnění nejen výše uvedených výzev.[4]

3.1 Klasifikace

Hlavním cílem klasifikace je označit a zařadit vzorky do určitého počtu skupin, v závislosti na jejich atributech. Často využívanými modely v oblasti energy managementu budov jsou rozhodovací stromy (Decision Trees), jejichž výhodou je jednoduchost a možnost přehledné vizualizace výsledků. Nejznámější algoritmy patřící do této skupiny jsou, například: CLS (Concept Learning System), ID3, C4.5, C5.0 a CART (Classification And Regression Trees). Dalším důležitým zástupcem je náhodný les (Random Forest), který v sobě kombinuje několik rozhodovacích stromů. Hlavní nevýhodou rozhodovacích stromů je jejich tendence ke snadnému přeučení, proto je nutné dbát na vhodné nastavení parametrů. Mezi další často používané metody patří například metoda podpůrných vektorů (Support Vector Machines). Dále se často aplikují metody využívající Bayesův teorém při zařazování vzorků, přičemž teorém vyžaduje silnou nezávislost mezi atributy vzorků. Mezi ty nejznámější metody také patří algoritmus k-nejbližších sousedů (K-Nearest Neighbors) a různé typy neuronových sítí. Klasifikační algoritmy jsou jedny z nejčastěji využívaných v oblasti energy managementu budov. Uplatňují se v detekci opakujících se vzorců chování, predikci hodnot, klasifikaci chování uživatelů a budov, analýze závislostí a detekci podvodů.[4, 5, 6]

3.2 Regrese

Regresní analýza se využívá především v predikci výsledků a vyhledávání závislostí na úrovni vzorků nebo jejich atributů. Zvláště předpověď spotřeby energie je jedním z hlavních cílů energy managementu. Mezi lineární modely patří, například: metoda nejmenších čtverců, Bayesovská lineární regrese či zobecněné lineární modely (Generalized Linear Models). Nicméně lineární modely často nedokážou nabývat vyšší úspěšnosti při zpracování reálných dat. Kromě lineárních modelů se využívají také nelineární modely, jejichž účinnost je při zpracování reálných dat významně vyšší. Mezi nelineární modely patří, například: polynomiální regrese, logistická regrese atd. Kromě těchto modelů lze využít i modely, které jsou využívány konvenčně pro klasifikaci. Z této oblasti lze využít, například rozhodovací stromy či metodu k-nejbližších sousedů. Závislost mezi vzorky či jejich atributy lze také analyzovat pomocí Pearsonovi nebo Spearmanovi korelace a korelačních koeficientů. V případě regresní analýzy lze také aplikovat různé druhy neuronových sítí.[4, 7]

3.3 Detekce anomálií

Jedním z nejdůležitějších nástrojů pro analýzu dat energy managementu jsou algoritmy pro detekci anomálií. Tyto metody dokáží s různou efektivitou odhalovat anomálie či odlehlé hodnoty v měření. Detekované anomálie lze poté odstranit a nahradit validními vzorky za pomoci různých metod pro nahrazení dat nebo je dále analyzovat. Mezi základní a často používané algoritmy patří GESD (Generalized Extreme Studentized Deviate), modifikované z-skóre nebo použití mezikvartilového rozpětí. Stejně jako v případě regresní analýzy lze využít pro detekci odlehlých hodnot i metody z jiných oblastí, příkladem může být algoritmus DBSCAN, který se primárně využívá ve shlukové analýze. Častěji jsou poté využívány kombinace různých metod za účelem zvýšení účinnosti detekce. Výjimkou není ani využití klasifikačních algoritmů. Detekce anomálií se hojně využívá v oblasti vyšetřování podvodů či identifikace energetických špiček.[4, 8]

3.4 Shluková analýza

Shluková analýza nebo tzv. klastrování je dalším možným přístupem k analýze dat energy managementu. Princip této analýzy spočívá v rozdělení dat do skupin (klastrů) v závislosti na jejich vzájemné podobnosti. Vzorky v jednom klastru by si měly být podobnější než se vzorky v jiných klastrech. Jedná se o metody strojového učení typu učení bez učitele, tedy shlukovací algoritmy se snaží sami nalézt podobnosti mezi jednotlivými vzorky a rozřadit je do skupin, čímž vykazují samoorganizaci bez potřeby validních výsledků na vstupu.[4, 9]

Mezi nejčastěji používané algoritmy patří shlukování metodou nejbližších středů (k-means), kde metoda rozdělí jednotlivé vzorky do k různých shluků na základě

vzdálenosti. Každý vzorek je přidělen do právě jednoho klastru. Na podobném principu funguje metoda fuzzy c-means, která umožňuje přidělit každý vzorek do více než jednoho klastru, s různou mírou příslušnosti. Obě tyto metody je vhodné používat na detekci sférických shluků. Další často používanou metodou je algoritmus DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN vytváří shluky na základě oblastí s vysokou koncentrací bodů. Navíc je tato metoda schopna detekovat anomálie, tj. vzorky, které se nachází mimo koncentrované oblasti, a zároveň je robustní vůči tvaru výsledného shluku. Kromě těchto algoritmů se navíc využívají, například metody hierarchické, metody založené na mřížce či metody založené na modelu. Shluková analýza se využívá v oblasti detekce opakujících se vzorců chování, klasifikaci chování budovy a uživatelů, odhalování skrytých spojitostí mezi vzorky nebo detekci anomálií.[4, 9]

V případě shlukovacích metod nelze určit účinnost aplikované metody za pomocí metrik jako je přesnost či preciznost. Speciálně pro shlukovou analýzu jsou definovány metriky, například:

- Silhouette score (silueta) jednotlivých vzorků definuje jejich podobnost s ostatními vzorky ve stejném klastru (koheze) v porovnání s podobností se vzorky z nejbližšího cizího klastru (separace).

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}, \tag{3.1}$$

  kde $b(i)$ je průměrná vzdálenost i-tého vzorku od všech bodů z nejbližšího cizího klastru. $a(i)$ představuje průměrnou vzdálenost i-tého vzorku od všech bodů ve stejném klastru. Hodnota siluety se pohybuje v intervalu od -1 do 1, kde -1 představuje chybnou klasifikaci vzorku a naopak 1 značí správnou klasifikaci vzorku do shluku.[10]

- Distortion score je metrika, která popisuje vzdálenost jednotlivých vzorků v klastru od jeho centroidu, čímž podává informaci o ucelenosti shluku.

3.4.1 Shlukování metodou nejbližších středů

Shlukování metodou nejbližších středů známější jako metoda k-means je shlukovací algoritmus, který rozděluje jednotlivé vzorky datasetu do k různých klastrů na základě vzdálenosti od jejich centroidů. V počáteční fázi je definován celkový počet shluků. V závislosti na tomto parametru se inicializují centroidy. Tato inicializace může probíhat různými způsoby, například náhodnou inicializací, což je nejjednodušší ale nespolehlivý přístup, protože se mohou centroidy vytvořit v těsné blízkosti a tím snížit efektivitu shlukování. Namísto náhodné inicializace je vhodnější použít metodu k-means++, která se snaží rozprostřít centroidy co nejdále od sebe a tím zvýšit efektivitu klastrování. Následně se každý vzorek přiřadí k nejbližšímu centroidu. Souřadnice centroidu se poté přepočítají tak, aby vzdálenost k jeho přiřazeným bodům byla co nejnižší. Poslední dva zmíněné kroky

se opakují, dokud není splněn počet iterací. K měření vzdálenosti se nejčastěji využívá euklidovská metrika, v případě časových řad je efektivnější využít metodu dynamic time warping, která je popsána v podkapitole 3.5.2.[9]

Efektivita této metody silně závisí na zadaném počtu klastrů, proto je definováno několik přístupů, jak tento parametr nastavit. Příkladem je možné uvést:

- Pravidlo rule of thumb, kde je doporučeno nastavit počet centroidů k na hodnotu odmocniny z poloviny počtu dat, kde n je počet dat.[10]

$$k \approx \sqrt{n/2} \tag{3.2}$$

- Nejstarší metoda lokte využívá vizualizaci, kde osu y tvoří postupně vypočítáná tzv. nákladová funkce (cost function) pro počet klastrů od 2 do n, kde tento počet tvoří osu x. Ve výsledném grafu se následně zvolí počet klastrů, kde nastal významný pokles či zlom. Problém může nastat, kdy graf nevykazuje žádné významné poklesy nebo naopak vykazuje více takových změn. Jako nákladová funkce se může využít, například: celkový čas shlukování, distortion score nebo siluetu.[10]

3.5 Analýza časových řad

Časové řady tvoří přesně danou sekvenci dat, tedy jednotlivé vzorky jsou uspořádané za sebou tak, jak byly naměřeny v souvislosti s časem. Tento typ dat je nedílnou součástí energy managementu, jelikož valná většina měření dat se provádí průběžně, příkladem může být měření spotřeby energie. Pomocí časové značky lze poté získat zajímavé poznatky o naměřených datech, například je možné analyzovat odděleně pracovní dny a víkendy či svátky, jejichž naměřené vzorky se od sebe mohou významně lišit.

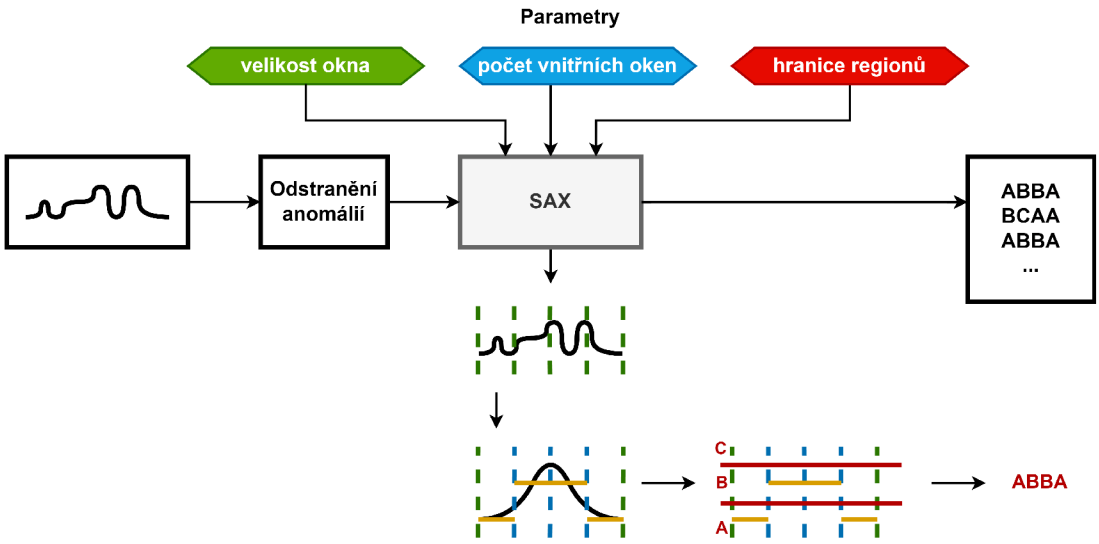
Jedním z důležitých přístupů k analýze časových řad je jejich dekompozice na trendovou, sezónní a náhodnou složku, popřípadě ještě složku cyklickou, kdy jednotlivé složky mohou být podrobeny dalšímu zpracování. Mimo jiné se zde využívají metody Fourierovy analýzy, autokorelace nebo různé typy klouzavých průměrů. Dále se často využívají různé třídy modelů časových řad, jejichž hlavním cílem je pochopení chování časových řad a následná předpověď jejich chování do budoucna. Příkladem může být statistický model ARIMA, který se skládá ze 3 částí – autoregresní (AR), která se snaží nalézt vzájemnou závislost mezi vzorky a vzorky zpožděnými. Autoregrese vychází z myšlenky, že hodnoty minulých vzorků ovlivňují vzorky budoucí, tedy mohou pomoci při predikci. Integrační (I) část diferencuje časovou řadu před použitím AR či MA, čímž se snaží zpracovávanou sekvenci transformovat do stacionárního stavu, tj. aby rozdělení pravděpodobnosti řady bylo v čase neměnné. Stacionarita časové řady je jedním z nejdůležitějších předpokladů pro další analýzy. Poslední částí jsou klouzavé průměry (MA), které vyjadřují

závislost mezi vzorky a náhodnou složkou (reziduem) za pomoci posuvného okna.[4]

Je nutné podotknout, že metody, které jsou popsány v předchozích podkapitolách, je možné aplikovat také na data typu časových řad. Nicméně před použitím těchto metod je nutné vstupní data transformovat, například rozdělit vzorky do samostatných celků – intervalů (dnů, měsíců apod.) za pomoci diskretizace. Takovým víceúčelovým příkladem může být metoda Symbolic aggregate approximation (SAX), která je popsána v následující podkapitole.

3.5.1 Symbolic aggregate approximation

SAX metoda se využívá především ke snížení velikosti dat typu časových řad. Oproti ostatním metodám, zabývající se podobnou problematikou, například analýza hlavních komponent (PCA), má výhodu, že zpracovávaná data nepřevádí do jiného vektorového podprostoru, čímž vzorky neztrácí svůj původní reálný význam. Metoda funguje na principu rozdělení sekvence dat na menší části, které se poté transformují do symbolických textových řetězců. Tímto procesem metoda provádí diskretizaci dat za pomoci posuvných oken stejné velikosti bez překrývání. Tato posuvná okna se poté rozdělí na další vnitřní okna. Vnitřní okna jsou aproximovány průměrnou hodnotou vzorků vnitřního okna. Tyto aproximace jsou následně označeny kódováním (znakem) podle toho, v jakém regionu se nachází. Diagram metody je zobrazen na obrázku 3.1.[6]



Obrázek 3.1: Diagram metody SAX

Definice hranic jednotlivých regionů je signifikantní z hlediska účinnosti metody. V případě dat pocházejících z normálního rozdělení lze využít vzdálenosti násobků směrodatné odchylky od střední hodnoty či použít jiné statistické charakteristiky. Před použitím SAX metody je každopádně nutné provést explorační datovou

analýzu zpracovávaných dat a zjistit základní informace o zkoumaném jevu. Počet regionů, přesněji jejich hranice, se nastaví v závislosti na volbě velikosti abecedy – znaky, které jsou použity ke kódování jednotlivých aproximací. Větší velikost abecedy umožňuje získat přesnější aproximace pro jednotlivá okna, nicméně snižuje míru výsledné generalizace. Uvádí se, že nejvhodnější velikost vnitřních oken je 6 hodin (v případě denní velikosti posuvného okna) a velikost abecedy 4.[6]

Výsledné kódy jednotlivých oken vytváří symbolické textové řetězce – tzv. slova SAX. Tyto slova se mohou následně analyzovat, například skrze výpočet vzdálenosti mezi jednotlivými slovy (jejich znaky) je možné aplikovat shlukovou analýzu či vyhledávat anomálie.[6]

SAX je, vzhledem ke své flexibilitě, rychlosti výpočtu a multifunkčnosti, široce používanou metodou v oblasti energy managementu. Využívá se jako metoda pro snížení dimenze dat, detektor vzorců chování nebo předzpracování vstupních dat shlukové analýzy.[6]

3.5.2 Dynamic time warping

V rámci analýzy časových řad se algoritmus dynamic time warping (DTW) využívá jako výpočet podobnosti (vzdálenosti) mezi dvěma časovými řadami. DTW nahrazuje klasickou euklidovskou metriku, která nesplňuje flexibilitu vůči posunům v čase, tj. nebere v úvahu časovou dimenzi. V případě silné korelace mezi dvěma časovými řadami, přičemž jedna z nich je posunuta o určité zpoždění, euklidovská vzdálenost často chybně definuje nízkou podobnost mezi těmito řadami. Namísto euklidovské metriky je vhodnější využít právě metodu DTW, která dokáže odhalit podobnost mezi dvěma časovými řadami i za předpokladu, že se liší v čase, rychlosti nebo délce.[11]

$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2}, \tag{3.3}$$

kde x a y jsou časové řady a $\pi = [\pi_0, ..., \pi_K]$ je cesta, která splňuje následující body:

- vektor indexů dvojic $\pi_k = (i_k, j_k)$ splňující vztah: $0 \leq i_k < n \wedge 0 \leq j_k < m$
- $\pi_0 = (0, 0)$ a $\pi_K = (n - 1, m - 1)$
- $\forall k > 0$, $\pi_k = (i_k, j_k)$ odpovídají $\pi_{k-1} = (i_{k-1}, j_{k-1})$, kde:
  - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
  - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

Nejdříve se vytvoří matice vzdáleností představující vzdálenosti mezi každou dvojicí vzorků časových řad. Následně se inicializuje matice nákladů (cost matrix), která udržuje minimální akumulovanou vzdálenost mezi odpovídajícími prvky obou

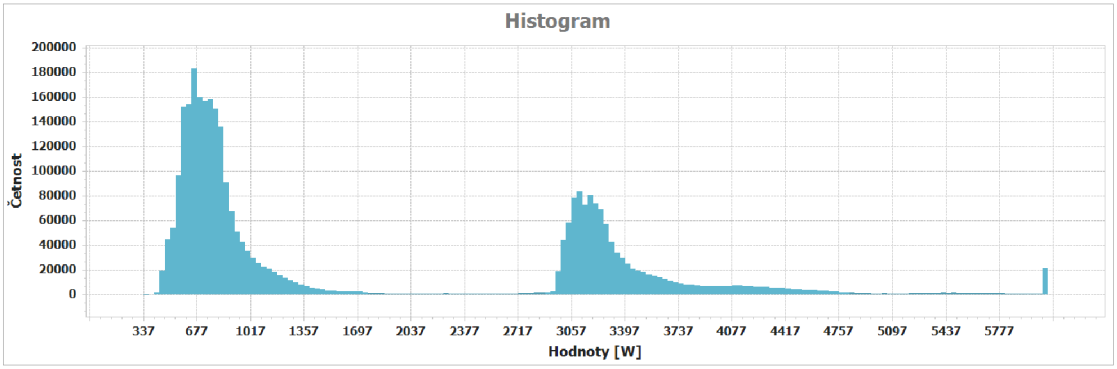
časových řad při všech možných zarovnáních. Na začátku obou časových řad je matice nákladů inicializována na nulu. Matice nákladů je definována pomocí iterací přes každou buňku matice vzdáleností a kalkulací minimální akumulované vzdálenosti mezi odpovídajícími prvky obou časových řad s přihlédnutím k vzdálenostem mezi sousedními prvky v obou časových řadách. Cílem je najít optimální zarovnání mezi časovými řadami, které minimalizuje celkovou vzdálenost. Nakonec se vypočítá celková vzdálenost mezi řadami nalezením minimální cesty prostřednictvím matice nákladů. Tato cesta reprezentuje optimální zarovnání.

4 Analýza dat a testování vybraných metod

V rámci této kapitoly jsou otestovány vybrané metody z předchozí kapitoly 3 na reálných datech z rodinného domu. Analýza dat společně s implementací a otestováním metod je provedena v prostředí .NET verze 7 za použití knihoven ML.NET verze 2.0.1, DevExpress zkušební verze 22.2.3 a MathNet.Numerics verze 5.0.0. Za použití zmíněných technologií je naimplementováno testovací prostředí, které disponuje nástroji pro práci s časovými řadami. Příkladem těchto nástrojů jsou základní popisné statistiky, různé typy vizualizace dat, metody pro detekci anomálií, metody pro nahrazení anomálií, různé druhy metrik a vybrané metody podrobněji popsané v předchozí kapitole. Před otestováním jednotlivých metod je v následující podkapitole provedena explorační analýza zpracovávaných dat.

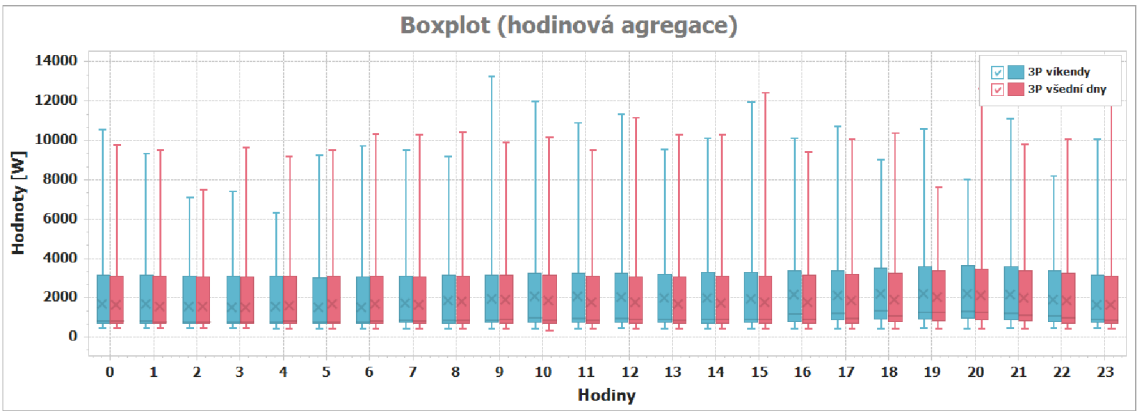
4.1 Explorační analýza dat

Analyzovaný dataset obsahuje sekvenci měření činné energie pocházející z rodinného domu. Časové rozmezí sběru dat je od 1. ledna do 31. prosince roku 2020. Jedná se o časovou řadu kde každý vzorek představuje 10sekundový agregační interval měření spotřebované energie. Chybějící hodnoty tvoří 0,07 % celého datasetu, k doplnění těchto vzorků je použita dopředná metoda. V naměřených agregačních intervalech často dochází k odchylkám v periodě měření v rámci jednotek až desítek milisekund. Největší naměřená odchylka je 35 ms. Tyto nepřesnosti následně vytváří posunutí vzorků až přetečení mimo agregační interval. Jelikož metody pro zpracování časových řad předpokládají pravidelné intervaly mezi vzorky, je potřeba data znovu agregovat pro zvýšení efektivity aplikovaných metod.



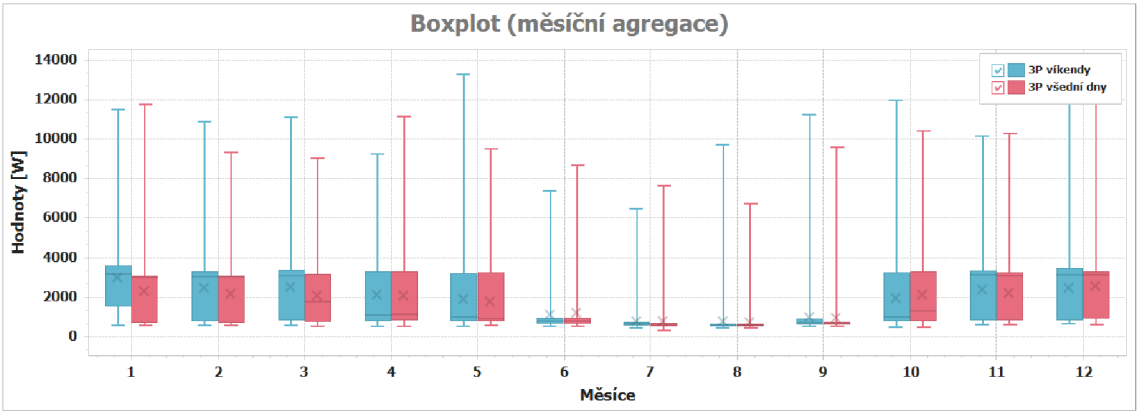
Obrázek 4.1: Histogram celého datasetu

Na obrázku 4.1 je vyobrazen histogram celého datasetu. Z grafu je patrné, že data nesplňují normální rozdělení, což následně ovlivňuje výběr metod, nastavení parametrů metod a statistik pro práci s daty. Nicméně histogram vykazuje bimodální rozdělení, což implikuje možný výskyt dvou typů měření v datasetu.



Obrázek 4.2: Boxplot celého datasetu, hodinová agregace rozdělená na všední dny (modrá) a víkendy (červená)

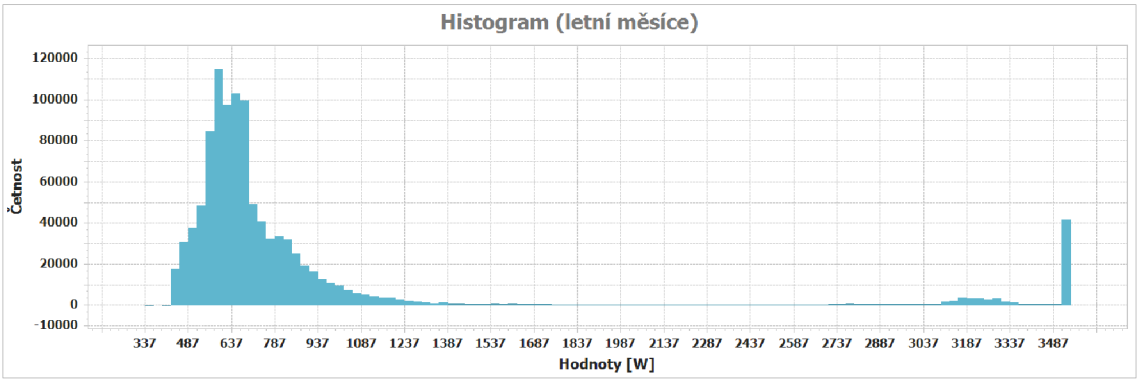
Krabicový graf na obrázku 4.2 zohledňuje hodinovou agregaci v rámci dne rozdělenou na všední dny a víkendy. Z grafu je patrné, že data nevykazují významné rozdíly jak mezi hodinami měření, tak mezi víkendy a všedními dny.



Obrázek 4.3: Boxplot celého datasetu, měsíční agregace rozdělená na všední dny (modrá) a víkendy (červená)

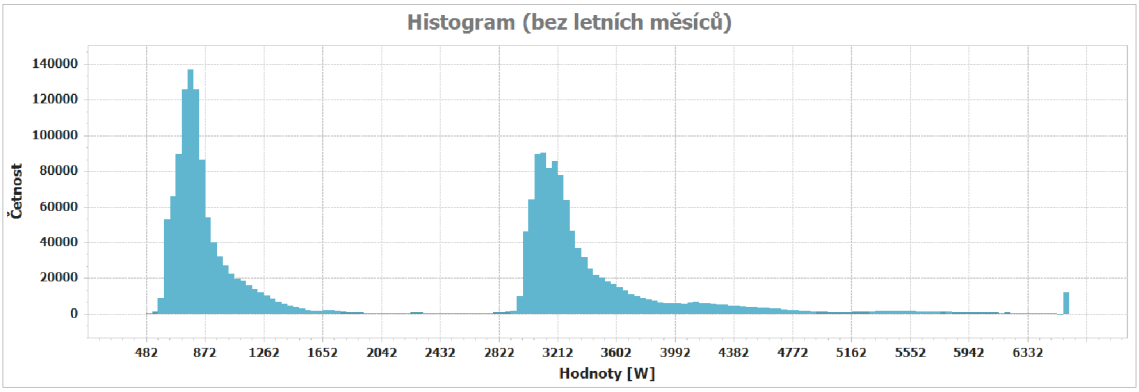
Další krabicový graf na obrázku 4.3 zobrazuje měsíční agregaci dat rozdělenou na všední dny a víkendy. Z grafu je patrný náhlý výkyv hodnot v letních měsících, tj. červen, červenec, srpen a září, kde se průměrné hodnoty činné energie pohybují okolo 1000 wattů. Tento jev může být způsoben náhlou změnou chování uživatele,

například odjezdem na dovolenou nebo defektem přístroje apod. Porovnání víkendů a pracovních dnů nevykazuje významné rozdíly v chování, pouze u vybraných měsíců, přesněji v lednu, únoru a březnu, se hodnoty činné energie o víkendech pohybují nepatrně nad hodnotami všedních dnů. Při bližším průzkumu letních měsíců lze vyčíst z histogram na obrázku 4.4, že většina hodnot se pohybuje přibližně od 400 do 1100 wattů. V porovnání s histogramem celého datasetu letní měsíce tvoří hodnoty činného výkonu v rozmezí od 3000 do 3400 wattů pouze nepatrné množství, což může být způsobeno již zmíněnou změnou chování či jiným faktorem.



Obrázek 4.4: Histogram dat letních měsíců

Histogram vzorků naměřených mimo letní měsíce, jenž je znázorněn na obrázku 4.5, vykazuje stále bimodální rozdělení dat.



Obrázek 4.5: Histogram dat bez letních měsíců

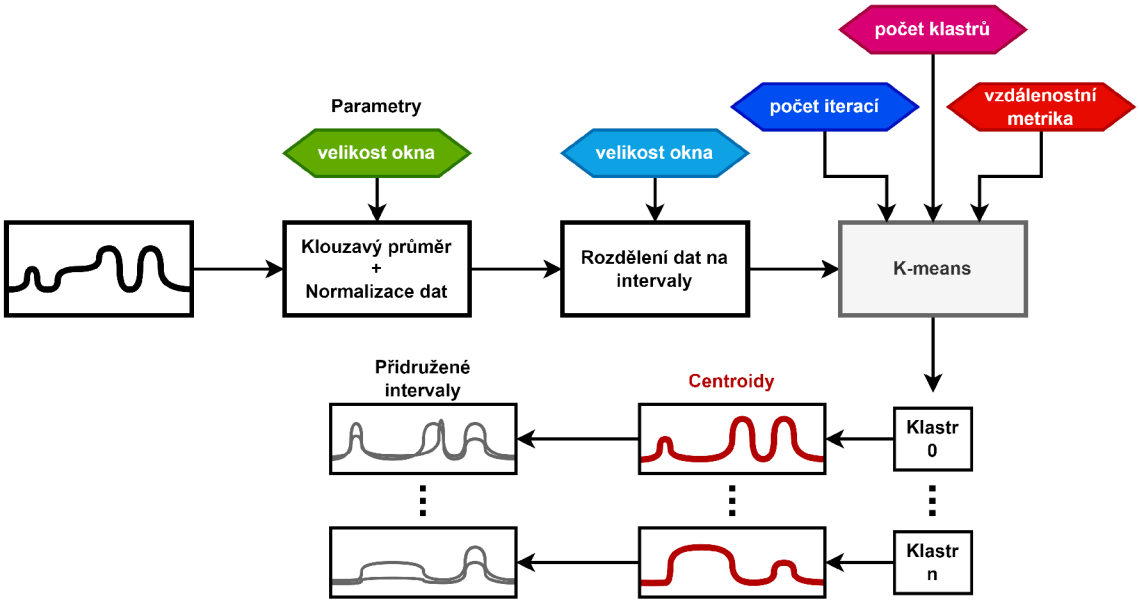
Hlubší analýzou datasetu pomocí různých časových filtrů nebo metod by bylo možné odkrýt další poznatky, podle kterých by se následně data mohla rozdělit a analyzovat odděleně. V rámci této práce jsou vybrané metody aplikovány na celý dataset, jenž vykazuje nenormální, bimodální rozdělení, ale také na různé časové výběry z datasetu. Za využití této informace jsou nastaveny specifické parametry

jednotlivých metod. Nenormální rozdělení ztěžuje výběr parametrů určitých algoritmů a často volba těchto parametrů závisí na matematickém modelu, například strojového učení, nebo samotném analytikovi. V případě známého rozdělení lze využít jeho vlastnosti k přesnějšímu nastavení parametrů metod.

4.2 Aplikace metody k-means

V rámci této kapitoly je popsána aplikace shlukovací metody k-means na analyzovaná data, včetně nastavení parametrů a porovnání jednotlivých výsledků. Jelikož je použití této metody na časové řady podmíněno určitými kroky, je následující podkapitola 4.2.1 věnována metodologii použití této metody.

4.2.1 Metodologie použití metody k-means



Obrázek 4.6: Diagram metodologie použití k-means

Na obrázku 4.6 je zobrazen diagram metodologie aplikace metody k-means. Na časové řady očištěné od chybějících hodnot je aplikován klouzavý průměr se zvolenou velikostí okna, výsledek je poté normalizován za pomocí min-max normalizace, kde jsou vzorky standardizovány v rozsahu 0 (minimum) až 1 (maximum). Následně je normalizovaná časová řada rozdělena na intervaly ve velikosti zvoleného okna například denního. Tyto intervaly vstupují do metody k-means, přesněji k-means++, jako samostatné vzorky. Současně se vzorky vstupují do metody parametry – počet iterací, počet klastrů a vzdálenostní metrika. Algoritmus roztřídí jednotlivé vzorky (intervaly) do různých shluků v závislosti na jejich chování. Výsledkem jsou klastry

definované centroidy a jejich přidruženými vzorky. Klastry a jednotlivé vzorky jsou ohodnoceny za pomoci metrik siluety a distortion score.

4.2.2 První cyklus testování metody k-means

V prvním cyklu testování je dataset rozdělen na 2 skupiny v závislosti na explorační analýze. První skupinu tvoří měření z letních měsíců (červen, červenec, srpen, září) a druhou skupinu tvoří měření zbylých měsíců. Pro obě skupiny je zpracována analýza nastavení parametru výsledného počtu klastrů pro různé agregační intervaly. Data jsou po předzpracování rozdělena na vzorky po dnech. Jako vzdálenostní metrika je zvolena dynamic time warping.

Tabulka 4.1: První cyklus - analýza nastavení parametru počtu klastrů letních měsíců

agregace	velikost okna klouzavého průměru	průměrný počet klastrů	průměrné distortion score
60 min	3	8,3	0,133
30 min	6	8,4	0,213
15 min	13	7,6	0,498

Letní měsíce tvoří přesně 122 dní. Podle pravidla rule of thumb je vhodné nastavit parametr počtu klastrů přibližně na 8. Pro spolehlivější nastavení parametru je pro každý agregační interval aplikována metoda k-means od 2 do 10 klastrů při 50 iteracích, kde je vždy vybrán počet klastrů s nejnižším distortion score. Takto je provedena analýza pro každý agregační interval celkem 10krát a výsledky jsou následně zprůměrovány a zasazeny do tabulky 4.1. Nejlepší výsledky z hlediska distortion score vykazují případy, kdy se nastavený počet klastrů pohybuje okolo 8, což odpovídá již zmíněnému rule of thumb.

Tabulka 4.2: První cyklus - analýza nastavení parametru počtu klastrů ne-letních měsíců

agregace	velikost okna klouzavého průměru	průměrný počet klastrů	průměrné distortion score
60 min	3	8,9	0,212
30 min	6	9,6	0,371
15 min	13	10,2	0,748

Zbytek roku tvoří 243 dní. Podle pravidla rule of thumb je vhodné nastavit parametr počtu klastrů na 11. V tabulce 4.2 jsou uvedeny výsledky analýzy, kde jsou testovány počty klastrů od 2 do 12. Analýza je provedena stejným způsobem jako u letních měsíců. Zde se nejlepší výsledky pohybují okolo 9 až 10 klastrů.

Následně je aplikována metoda k-means na jednotlivé agregace se zvoleným počtem klastrů. Pro letní měsíce je nastaven počet klastrů na 8 a pro druhou skupinu, tj. zbytek roku, je nastaven parametr na 10. Pro každý agregační interval jsou otestovány různé velikosti oken klouzavého průměru, přičemž se výpočet shlukování provádí celkem 10x pro každou velikost okna a výsledky, tj. silueta a distortion score, jsou následně zprůměrovány a zasazeny do tabulek níže.

V případě výsledků letních měsíců, které jsou zaznamenány v tabulce 4.3, se hodnota průměrné siluety, se zvětšující se velikostí okna klouzavého průměru, zvyšuje. Tento jev je způsobem právě vyhlazením jednotlivých nedokonalostí denních vzorků a tím spíše si navzájem odpovídají svým průběhem. Na druhou stranu pokud úroveň vyhlazení překročí určitou úroveň, jednotlivé vzorky si mohou být významně podobné i mezi shluky a tím snižovat hodnotu získané informace. Letní měsíce tento trend vykazují v případě 60minutového a 30minutového agregačního intervalu. Naopak výsledky 15minutové agregace tento trend nesplňují. Nejlepší výsledky z hlediska průměrné siluety představuje 30minutový agregační interval s velikostí okna klouzavého průměru nastavenou na 47 hodnot.

Tabulka 4.3: První cyklus - analýza nastavení parametru klouzavého průměru letních měsíců

agregace	velikost okna klouzavého průměru	průměrná silueta	průmerné distortion score
60 min	8	0,344	0,73
	16	0,438	0,103
	23	0,542	0,100
30 min	16	0,378	0,172
	32	0,518	0,246
	47	**0,592**	**0,206**
15 min	32	0,224	0,638
	64	0,182	0,612
	95	0,377	0,354

Stejným způsobem je aplikována analýza na druhou skupinu dat s tím rozdílem, že parametr počtu klastrů je nastaven na 10. Výsledky této skupiny, které jsou k dispozici v tabulce 4.4, nedosahují takových hodnot průměrné siluety jako v případě letních měsíců, to může být způsobeno vyšším počtem vzorků, což může mít za následek větší množství různých vzorců v měřeních. Trend zvyšující se siluety a zvyšující se velikostí okna klouzavého průměru platí i zde, se stejnou podmínkou jako v případě letních měsíců, že 15minutový agregační interval tento trend nesplňuje.

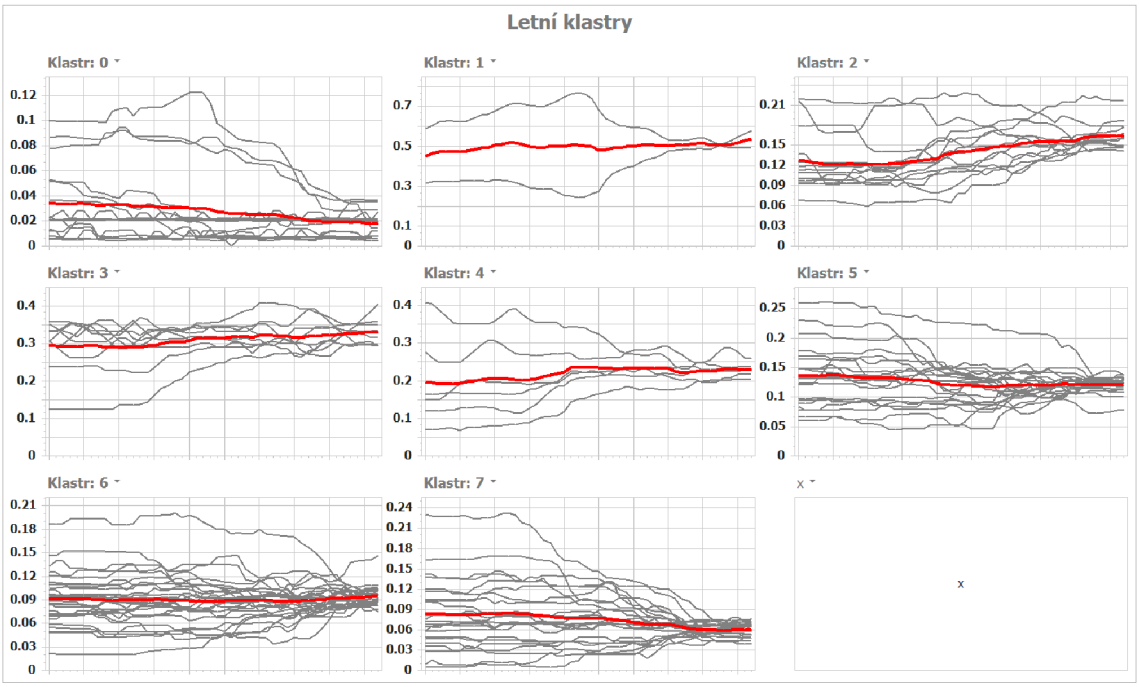
Průměrné hodnoty siluety v případě druhé skupiny dat dosahují v některých měřeních záporných hodnot, což naznačuje začlenění určitých vzorků do nesprávných shluků. Nejlepší výsledky z hlediska průměrné siluety představuje 30minutový

agregační interval s velikostí okna klouzavého průměru nastavenou na 47 hodnot.

Tabulka 4.4: První cyklus - analýza nastavení parametru klouzavého průměru ne-letních měsíců

agregace	velikost okna klouzavého průměru	průměrná silueta	průmerné distortion score
60 min	8	-0,010	0,130
	16	0,228	0,134
	23	0,290	0,140
30 min	16	0,128	0,238
	32	0,344	0,254
	47	**0,406**	**0,254**
15 min	32	-0,006	0,622
	64	0,216	0,562
	95	0,164	0,598

Letní měsíce jsou zpracovány metodou k-means, kde jsou měření agregovány po 30 minutách, počet klastrů je nastaven na 8, velikost okna klouzavého průměru je 47 vzorků. Maximální počet iterací je nastaven na 50. Výsledek ve formě grafů nalezených klastrů je na obrázku 4.7, kde černé křivky představují jednotlivé dny a červené křivky reprezentují centroid klastru. Grafy představují závislost normalizovaných hodnot spotřeby energie na čase, přesněji hodinách v průběhu dne.



Obrázek 4.7: První cyklus - graf nalezených shluků v letních měsících

V tabulce 4.5 jsou zaznamenány hodnoty siluety a distortion score pro jednotlivé letní klastry, které jsou seřazeny podle siluet sestupně. Nejvyšší hodnoty siluety dosahuje první klastr, zároveň má ale nejvyšší distortion score. Tento klastr obsahuje pouze dva dny, a to 29. a 30. září. Spotřeba energie v těchto dnech odpovídá spotřebě ne-letních měsíců, což souhlasí s faktem, že se jedná o poslední zaznamenané vzorky v letní skupině dat. První klastr tím vykazuje schopnost detekovat anomálie v letních měsících, v tomto případě v datech s minimální spotřebou energie. Nultý klastr má druhou nejvyšší siluetu a zároveň nejnižší distortion score. Obsahuje data od 22. července do 2. srpna a data od 12. srpna do 18. srpna současně s 27. srpnem. Jen podle časové značky lze poznat, že se klastr snaží najít vzorce chování v červenci a srpnu. Pro bližší informace by bylo potřeba zahrnout další data o budově, například počet lidí v budově, počasí atd. Zajímavé výsledky zde představuje třetí klastr, který zahrnuje dny od 2. do 9. června a 28. září. Tím klastr vykazuje snahu rozpoznat přechod mezi různými typy tarifu spotřeby energie. Ve zpracovávaném datasetu na počátku června přechází spotřeba z klasické spotřeby na spotřebu minimální a tím, že algoritmus zahrnuje do tohoto klastru i 28. září, tak značí, že je schopný detekovat i přechod zpět do klasické spotřeby zbytku roku.

Tabulka 4.5: První cyklus - vyhodnocení nalezených shluků v letních měsících

klastr	silueta	distortion score	počet dnů
1	0,779	1,094	2
0	0,700	0,029	17
3	0,654	0,115	8
7	0,509	0,066	24
4	0,509	0,220	7
6	0,490	0,038	30
5	0,450	0,068	22
2	0,389	0,057	11

Sedmý klastr obsahuje různé dny z července, srpna a září. Oproti ostatním klastrům obsahuje vyšší zastoupení pracovních dnů oproti víkendům. Součástí sedmého klastru jsou 2 víkendy a 22 všedních dnů. Poslední tři klastry, tj. šestý, pátý a druhý, obsahují různorodé dny, které při ruční kontrole průběhu spotřeby nevykazují společné zřetelné znaky, ale i tak mohou ukrývat cenné informace, které dokázal algoritmus k-means odhalit.

Na druhou skupinu dat bez letních měsíců je aplikována metoda k-means, kde jsou měření agregovány po 30 minutách, počet klastrů je nastaven na 10, velikost okna klouzavého průměru je 47 vzorků. Maximální počet iterací je nastaven na 50. Na obrázku A.1 je zobrazen graf objevených klastrů. Informace o jednotlivých klastrech jsou uvedeny v tabulce 4.6, kde jsou klastry setříděny sestupně podle hodnoty siluety.

Tabulka 4.6: První cyklus - vyhodnocení nalezených shluků v datech bez letních měsíců

klastr	silueta	distortion score	počet dnů
7	0,594	0,207	33
5	0,591	0,168	16
3	0,484	0,212	15
9	0,408	0,154	28
2	0,391	0,093	20
4	0,361	0,937	12
6	0,360	0,303	29
0	0,326	0,144	24
8	0,314	0,145	35
1	0,276	0,210	30

Většina klastrů obsahuje dny z různorodých měsíců bez očividných specifických znaků, které by je spojovaly. Výjimku zde tvoří několik klastrů. Třetí klastr obsahuje převážně dny z měsíců května, října a prosince. Jeho hlavním rysem je, že obsahuje jedny z posledních dnů těchto měsíců, přesněji v každém uvedeném měsíci obsahuje 24. až 27. den. Čtvrtý klastr obsahuje převážně dny měsíců ledna a prosince. Oproti ostatním dnům zde zřetelně vyčnívají svými hodnotami 28. a 29. prosinec, které jak v rámci tohoto klastru, tak i celého datasetu jsou anomáliemi. Z tohoto důvodu má tento klastr nejvyšší distortion score. Nultý klastr tvoří převážně dny listopadové a prosincové, zároveň obsahuje pouze 4 víkendy oproti 20 dnům všedním. Osmý klastr obsahuje převážně dny od února do května.

4.2.3 Druhý cyklus testování metody k-means

Druhý cyklus testování je proveden stejným způsobem jako první s tím rozdílem, že jako vzdálenostní metrika je zvolena euklidovská vzdálenost. Specificky jsou data rozdělena do 2 skupin – letní měsíce a zbytek roku. Data jsou po předzpracování rozdělena na vzorky po dnech.

Tabulka 4.7: Druhý cyklus - analýza nastavení parametru počtu klastrů letních měsíců

agregace	velikost okna klouzavého průměru	průměrný počet klastrů	průměrné distortion score
60 min	3	8,6	0,063
30 min	6	8,1	0,122
15 min	13	8,5	0,248

Analýza nastavení parametru počtu klastrů pro letní měsíce, kde výsledky jsou uvedeny v tabulce 4.7, vykazuje nejlepší výsledky pro 8 shluků. což odpovídá pravidlu rule of thumb a shoduje se s výsledky z prvního cyklu testování. Výsledky

analýzy pro druhou skupinu dat jsou zaznamenány v tabulce 4.8, kde algoritmus k-means dosahuje nejlepších výsledků při 9 klastrech. Hodnoty průměrné distortion score metriky dosahují poloviční hodnot oproti hodnotám metriky z prvního cyklu testování, což implikuje vyšší celistvost jednotlivých klastrů.

Tabulka 4.8: Druhý cyklus - analýza nastavení parametru počtu klastrů ne-letních měsíců

agregace	velikost okna klouzavého průměru	průměrný počet klastrů	průměrné distortion score
60 min	3	9	0,110
30 min	6	9,7	0,216
15 min	13	9	0,382

Analýza nastavení parametru klouzavého průměru letních měsíců, kde výsledky analýzy jsou zaznamenány v tabulce 4.9, vykazuje nejlepší výsledky v případě 60minutového agregačního intervalu s velikostí okna klouzavého průměru nastavenou na 23 vzorků. Použití euklidovské metriky výrazně snižuje průměrné distortion score oproti DTW v prvním cyklu testování. Na druhou stranu DTW dosahuje obecně lepších výsledků průměrné siluety, což souhlasí s faktem, že se DTW primárně volí jako vzdálenostní metrika v případě časových řad.

Tabulka 4.9: Druhý cyklus - analýza nastavení parametru klouzavého průměru letních měsíců

agregace	velikost okna klouzavého průměru	průměrná silueta	průmerné distortion score
60 min	8	0,204	0,076
	16	0,264	0,067
	23	**0,320**	**0,022**
30 min	16	0,220	0,143
	32	0,272	0,156
	47	0,297	0,036
15 min	32	0,218	0,274
	64	0,265	0,148
	95	0,279	0,070

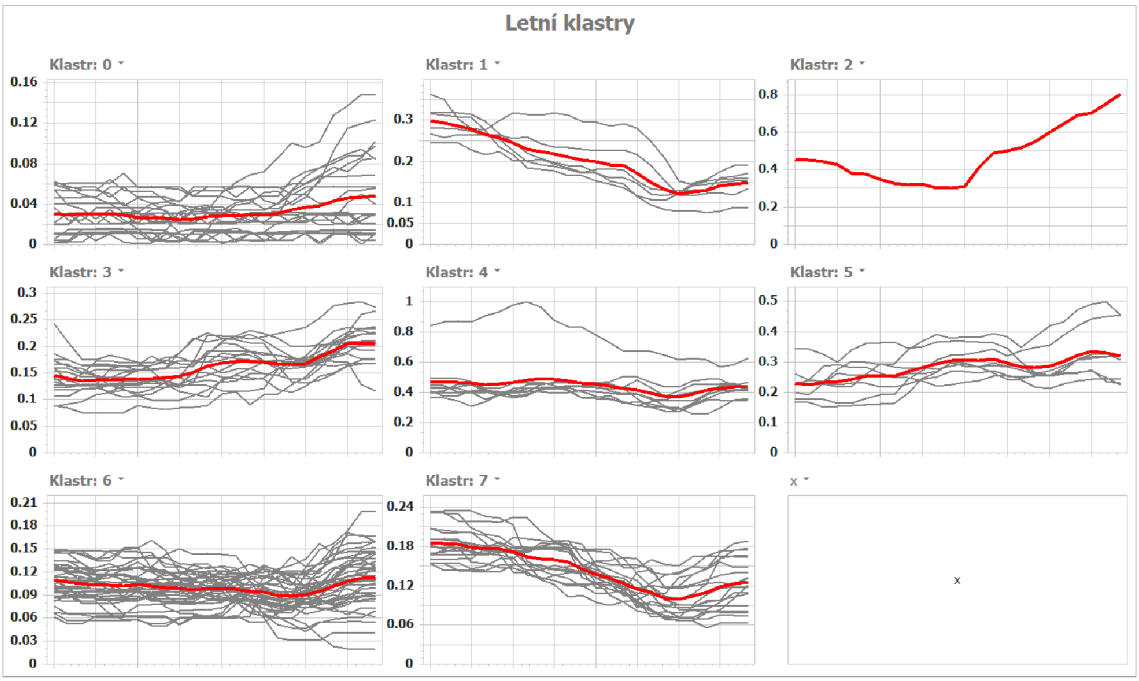
V případě analýzy nastavení parametru klouzavého průměru ve skupině dat bez letních měsíců, kde výsledky jsou uvedeny v tabulce 4.10, dosahuje nejlepších výsledků 30minutový agregační interval s velikostí okna klouzavého průměru nastavenou na 47 vzorků. Stejně jako v případě letních měsíců hodnoty průměrné distortion score metriky dosahují nižších hodnot než za použití DTW v prvním cyklu testování a opět v případě průměrné siluety dosahuje DTW vyšších hodnot. Algoritmus

za použití euklidovské metriky nedosahuje oproti DTW žádných záporných hodnot průměrné siluety.

Tabulka 4.10: Druhý cyklus - analýza nastavení parametru klouzavého průměru ne-letních měsíců

agregace	velikost okna klouzavého průměru	průměrná silueta	průmerné distortion score
60 min	8	0,149	0,084
	16	0,209	0,060
	23	0,248	0,053
30 min	16	0,162	0,153
	32	0,200	0,137
	47	**0,258**	**0,096**
15 min	32	0,141	0,326
	64	0,208	0,265
	95	0,242	0,194

Metodou k-means jsou zpracovány letní měsíce, přičemž měření jsou agregována po hodině. Pro tento proces je použito 8 klastrů a velikost okna klouzavého průměru je 23 vzorků. Maximální počet iterací je nastaven na 50. Výsledné grafy nalezených klastrů jsou prezentovány na obrázku 4.8.



Obrázek 4.8: Druhý cyklus - graf nalezených shluků v letních měsících

V tabulce 4.11 jsou zaznamenány výsledné metriky letních klastrů seřazené sestupně dle hodnoty siluety. Nejvyšší hodnoty siluety zde dosahuje nultý klastr, jehož součástí jsou dny od 22. do 31. července společně s 12. až 28. srpnem. V pozdních hodinách jsou v nultém klastru patrné anomálie 19. a 23. srpen, které přesahují normalizovanou hodnotu 0,12. Podobným způsobem se chová i pátý klastr, který představuje vybrané dny z června a září. Zde 28. září a 5. červen dosahují vyšších hodnot v pozdních hodinách. Stejně jako v prvním cyklu testování i zde je ve čtvrtém klastru snaha rozpoznat přechod mezi letním tarifem spotřeby a spotřebou zbytku roku. Oproti prvnímu cyklu je součástí tohoto klastru 30. září, které dosahuje maximálních hodnot letní spotřeby kolem sedmé hodiny ranní. Jelikož se 30. září významně odchyluje od zbytku dat v klastru, nabývá distortion score shluku vyšších hodnot. Zajímavý je zde druhý klastr, který obsahuje pouze jeden den a to 29. září, což naznačuje jeho abnormalitu v rámci letních dat.

Tabulka 4.11: Druhý cyklus - vyhodnocení nalezených shluků v letních měsících

klastr	silueta	distortion score	počet dnů
0	0,574	0,012	23
6	0,351	0,015	39
3	0,283	0,022	17
7	0,278	0,016	18
5	0,229	0,085	8
4	0,215	0,410	9
1	0,206	0,037	6
2	0	0	1

Druhá skupina dat bez letních měsíců je zpracována pomocí metody k-means s agregačním intervalem 30 minut. Počet klastrů je nastaven na 9, velikost okna klouzavého průměru je 47 vzorků a maximální počet iterací je nastaven na 50. Výsledné grafy klastrů jsou zobrazeny v příloze B.1. V tabulce 4.12 jsou uvedeny výsledné metriky pro každý klastr.

Tabulka 4.12: Druhý cyklus - vyhodnocení nalezených shluků v datech bez letních měsíců

klastr	silueta	distortion score	počet dnů
5	0,436	0,081	15
6	0,320	0,047	38
7	0,274	0,241	12
0	0,267	0,137	34
3	0,246	0,060	50
1	0,231	0,074	36
8	0,158	0,097	39
4	0,120	0,091	17
2	0	0	1

Pátý klastr dosahuje nejlepších výsledků z hlediska siluety, jeho součástí jsou vybrané dny z května a října společně s výjimkami z dubna a prosince. Tyto data nevykazují z hlediska časové značky zřetelné kontinuity. Pro lepší pochopení tohoto klastru, ale i klastrů ostatních, by bylo potřeba vzít v úvahu i jiná data, například: počet lidí v budově, lokální počasí či teplotu. Sedmý shluk v sobě spojuje převážně data z ledna a prosince. Zřetelně zde vystupuje z řady 29. prosinec, který dosahuje okolo 12 hodiny maximálních hodnot spotřeby energie. Třetí klastr zde představuje vybrané dny všech neletních měsíců kromě prosince. S přihlédnutím k hodnotě distortion score, která dosahuje druhé nejnižší hodnoty, a počtu dnů zahrnutých v tomto klastru, může tento shluk představovat například standardní denní spotřebu. První klastr spojuje vybrané denní spotřeby energie měsíců od ledna do března a od října do prosince. V prvním klastru částečně vyčnívá 11. leden, který přesahuje po 15. hodině hodnoty 0,6. Stejně jako v případě letních měsíců i zde je nalezen klastr, který je definován pouze jedním dnem – 28. prosincem, který představuje anomálii z hlediska spotřeby energie.

4.2.4 Diskuze výsledků shlukové analýzy

Algoritmus k-means obecně dosahuje lepších výsledků za použití DTW jako vzdálenostní metriky. Euklidovská vzdálenost oproti DTW definuje v obou skupinách dat vždy jeden klastr v závislosti pouze na jednom vzorku. Důležitým rozdílem mezi DTW a euklidovskou metrikou je složitost výpočtu, přičemž DTW má výrazně vyšší komplexnost výpočtu. Při agregaci ročních dat 15minutovým agregačním intervalem a zpracováním těchto dat algoritmem k-means, za použití 10 klastrů a maximální počtem 50 iterací, jsou získány následující výsledky časové náročnosti výpočtu v závislosti na použité vzdálenostní metrice. Za využití euklidovské metriky trvá výpočet přibližně 360 ms. Oproti tomu použití DTW prodlouží čas výpočtu až na 47000 ms. Je vhodné zmínit, že v této práci není využita jedna z hlavních výhod metody DTW, což je možnost výpočtu vzdálenosti i mezi časovými řadami, které mají různý počet vzorků. Touto vlastností euklidovská metrika nedisponuje.

Jelikož je v této práci zpracováván dataset obsahující pouze časovou složku a měření spotřeby energie, je poměrně obtížné ze strany analytického pracovníka odhalit spojitosti mezi vzorky v jednotlivých klastrech. Většina klastrů v sobě spojuje vzorky z vybraných měsíců. Objevují se ale i shluky, které nevykazují ani toto rozdělení. V případě klastrů s nižším počtem vzorků je možné předpokládat jejich unikátnost se specifickými vlastnostmi. Jelikož dataset nevykazuje významné změny chování v porovnání víkendů a všedních dnů, není možné ani předpokládat rozdělení vzorků do shluků v závislosti na této informaci. Pro pochopení nalezených shluků v této práci z pohledu člověka by bylo vhodné znát i doplňující informace o jednotlivých měřeních. Z hlediska chytrých budov je možné do analýzy zahrnout například počasí, interní a externí teplotu či vlhkost, počet lidí v budově, aktivní spotřebiče, chování HVAC systému, aktivní energetický tarif, strukturu samotné budovy a další.

4.3 Aplikace metody SAX

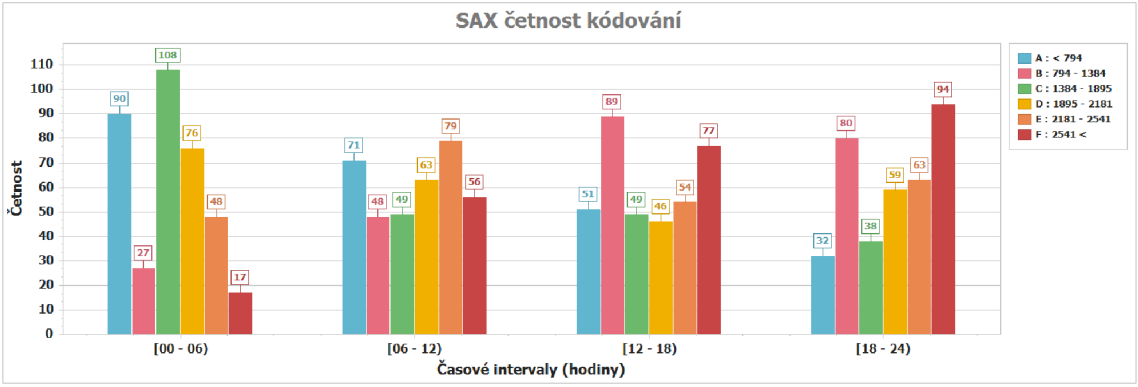
V rámci této kapitoly je popsána aplikace metody symbolic aggregate approximation na analyzovaná data, včetně nastavení parametrů a porovnání jednotlivých výsledků. Základní metodologie této metody je popsána v kapitole 3.5.1. Pro první cyklus testování metody jsou vybrány na základě explorační analýzy dat parametry, které jsou uvedeny v tabulce 4.13.

4.3.1 První cyklus testování metody SAX

Tabulka 4.13: První cyklus - nastavení parametrů a kódování metody SAX

velikost okna	počet vnitřních oken	hranice regionů	kódování regionů
8640	4	min - 794	A
		794 - 1384	B
		1384 - 1895	C
		1895 - 2181	D
		2181 - 2541	E
		2541 - max	F

Velikost okna 8640 vzorků představuje v 10sekundovém agregačním intervalu jeden den. Počet vnitřních oken je nastaven na 4, což vytváří rozdělení hlavního okna na 4 díly, tedy díly po 6 hodinách, tj. 0–6, 6–12, 12–18 a 18–24 hodin. Jednotlivé hranice regionů jsou nastaveny na sextily průměrných hodnot aproximovaných vnitřních oken.



Obrázek 4.9: První cyklus - četnost kódovacích znaků vnitřních oken

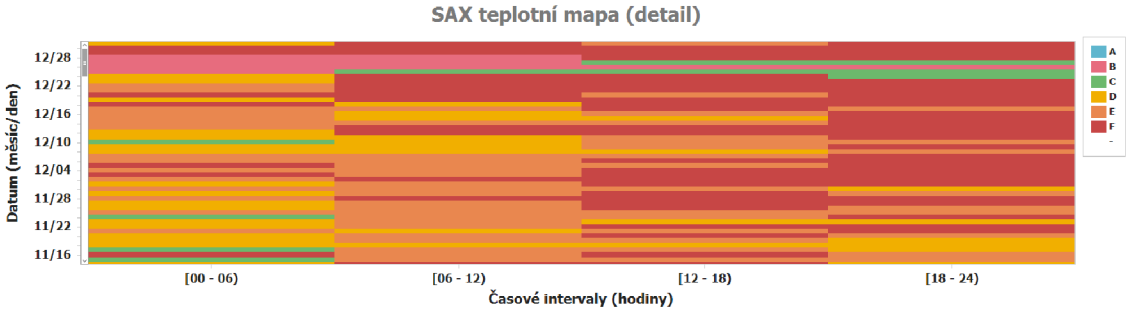
Na obrázku 4.9 je zobrazen sloupcový graf, který popisuje jednotlivé zastoupení kódovacích znaků (regionů) v rámci vnitřních oken. Graf v okně 0–6 hodin vykazuje významný počet zástupců regionů A, C a D. Naopak jsou aproximace tohoto

okna součástí regionů B a F pouze v 27 a 17 případech. V případě následujícího okna 6–12 hodin, jsou kódování zastoupena rovnoměrně s převažujícím regionem D čítající 79 pozorování. Vnitřní okno 12–18 hodin má rovnoměrně zastoupené regiony A, C, D a E. Regiony B a F zde převládají s počty 89 a 77 aproximací. V posledním vnitřním okně 18–24 hodin převažují, stejně jako v okně předchozím, regiony B a F.

Tabulka 4.14: První cyklus - nejčetnější slova SAX

slovo SAX	celkový počet	všední dny	víkendy	měsíce
AAAA	25	19	6	7 - 9
AABB	24	17	7	6 - 9
ABBB	16	7	9	7 - 9
EFFF	15	6	9	1 - 4, 11, 12

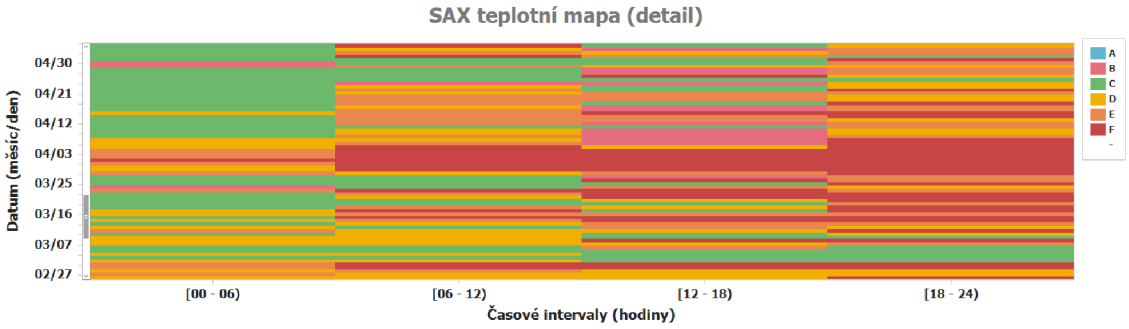
V tabulce 4.14 jsou zaznamenány nejčetnější slova SAX, kde převážně letní dny tvoří většinu těchto slov, které si udržují stálý trend nízkých průměrných hodnot aproximací. Čtvrté nejčastější slovo se naopak objevuje převážně u dnů mimo letní období.



Obrázek 4.10: První cyklus - detail teplotní mapy zaměřený na měsíce listopad a prosinec

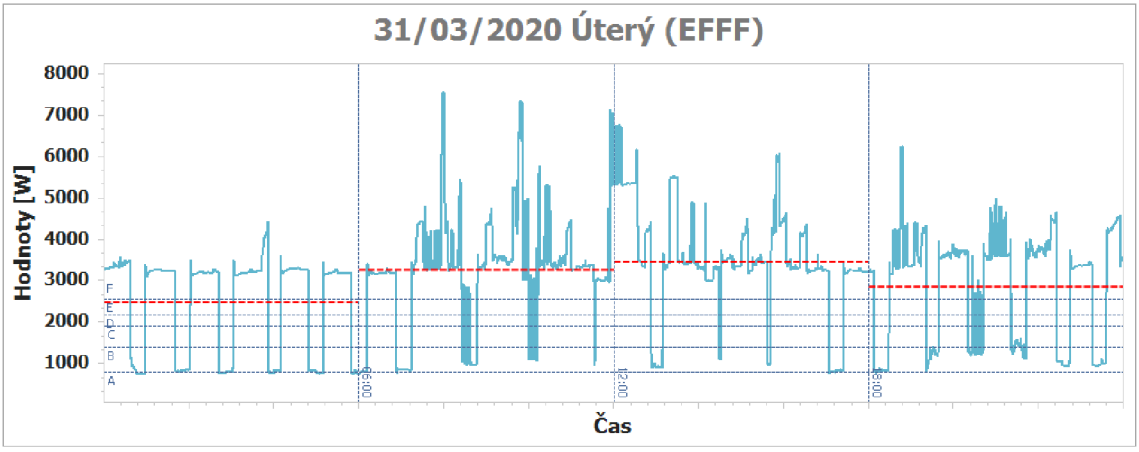
Na obrázku 4.10 je zobrazena teplotní mapa, kde jednotlivé buňky představují výsledné kódování časových oken. Specificky tato teplotní mapa zobrazuje detail kódování měsíců listopadu a prosince. Je patrné, že převládají regiony D, E a F. Výjimku zde představuje období Vánoc, přesněji dny od 24. do 28. prosince kde tyto dny jsou označeny kódováním B a C, tedy nižší průměrnou spotřebou oproti ostatním dnům v tomto období. Další zajímavé informace vykazují dny 29. a 30. prosinec, jelikož jsou všechna jejich vnitřní okna označena regionem F, tedy nejvyšší průměrnou spotřebou. Totožné chování s těmito dny sdílí už pouze jen 3. až 5. leden a 1. duben.

Na obrázku 4.11 je prezentován detail teplotní mapy zaměřený na měsíce březen a duben. Patrné jsou zde vyčnívající regiony B v časovém intervalu 12–18 hodin, mimo jiné se jedná o dny od 6. dubna do 10. dubna. Kromě toho je zde zřejmá vyšší průměrná spotřeba v období od 29. března do 4. dubna, přesněji v intervalu od 6. hodiny ranní po půlnoc.



Obrázek 4.11: První cyklus - detail teplotní mapy zaměřený na měsíce březen a duben

Na obrázku 4.12 je zobrazen graf měření spotřeby energie dne 31. března, kde červené linie značí průměrnou spotřebu v rámci jednotlivých vnitřních oken. Modré vodorovné linie značí jednotlivé hranice regionů. Vnitřní okna jsou zde ohraničena svislými přerušovanými čárami. Červené linie zde definují výsledné slovo SAX, tj. EFFF. Z grafu je patrné jak významně metoda SAX generalizuje jednotlivé vzorky dnů. Původní den čítající celkem 8640 vzorků měření spotřeby energie je zobecněn kódováním 4 znaků, což představuje významnou ztrátu informace za cenu snížení celkové velikosti dat. Z toho důvodu je před použitím této metody nutné zvážit mnoho faktorů, tj. dostupnou výpočetní kapacitu, požadovanou přesnost následné analýzy, očekávanou úroveň získaných informací z dat a další.



Obrázek 4.12: První cyklus - měření spotřeby energie dne 31. března s vyznačenými průměrnými spotřebami energie (červená)

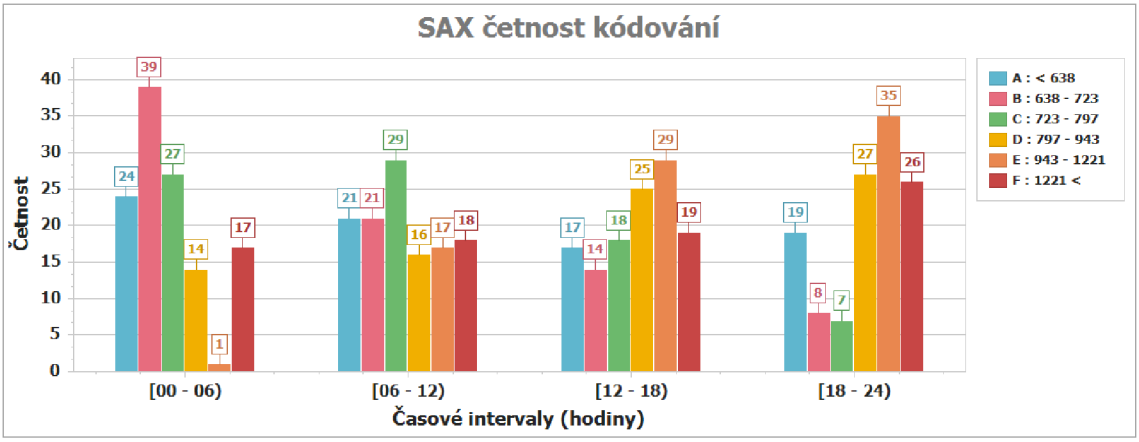
4.3.2 Druhý cyklus testování metody SAX

V případě druhého cyklu testování jsou analyzovány pouze letní měsíce, tj. červen, červenec, srpen a září.

Tabulka 4.15: Druhý cyklus - nastavení parametrů a kódování metody SAX

velikost okna	počet vnitřních oken	hranice regionů	kódování regionů
8640	4	min - 638	A
		638 - 723	B
		723 - 797	C
		797 - 943	D
		943 - 1221	E
		1221 - max	F

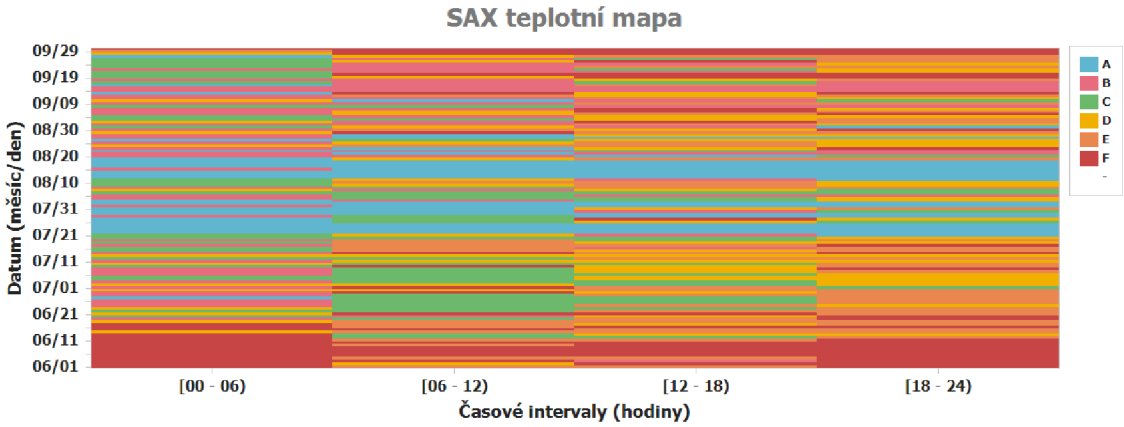
Parametry jsou použité stejné jako v předchozím cyklu, tj. denní velikost okna rozdělena na 4 části po šesti hodinách. Hranice regionů jsou opět nastaveny na sextily průměrných hodnot aproximovaných vnitřních oken. Jelikož je spotřeba energie v rámci letních měsíců nižší, jsou hranice regionů definovány nižšími hodnotami oproti prvnímu cyklu testování.



Obrázek 4.13: Druhý cyklus - četnost kódovacích znaků vnitřních oken

Na obrázku 4.13 je zobrazen sloupcový graf četnosti kódování. V časovém intervalu 0–6 hodin se nejčastěji vyskytují regiony A, B a C, přičemž region B v tomto čase označuje téměř jednu třetinu všech dnů. Naopak region E se vyskytuje v tomto okně pouze v jednom případě. Vnitřní okno 6–12 hodin má rovnoměrně zastoupené všechny regiony kromě kódování C, které tvoří přes čtvrtinu aproximací v tomto okně. V časovém intervalu převládají regiony D a E s počty 25 a 29 aproximací. Ostatní kódování jsou v tomto okně zastoupeny rovnoměrně. Poslední vnitřní okno 18–24 hodin je převážně označeno regiony D, E a F, přičemž převládá region E

s téměř jednou třetinou průměrných aproximací. Regiony B a C jsou součástí tohoto okna pouze v 8 a 7 případech.



Obrázek 4.14: Druhý cyklus - teplotní mapa letních měsíců

Na obrázku 4.14 je zobrazena teplotní mapa letních měsíců. Nejčastěji se v těchto měsících vyskytuje slovo AAAA, přesněji celkem dvanáctkrát, kde se jedná pouze o dny měsíců července a srpna. Druhé nejčastější slovo je FFFF, které označuje dny na začátku června a na konci září. Na teplotní mapě je vidět výrazný červený pruh regionu F na začátku června, kde přechází klasický tarif spotřeby energie v letní spotřebu.

4.3.3 Diskuze výsledků metody SAX

Metoda symbolic aggregate approximation se převážně využívá v oblasti předzpracování dat za účelem snížení dimenze. V této podkapitole je věnována pozornost hlavně vizualizaci výsledků této metody, kdy je následně možné identifikovat různé vzorce chování či anomálie ve spotřebě energie. Při aplikaci této metody je nutné stanovit požadovanou detailnost výsledné informace, a podle toho nastavit parametry. V případě této práce se osvědčila denní okna rozdělena na 4 díly po 6 hodinách a nastavení regionů na sextily průměrných hodnot aproximovaných vnitřních oken. Při větším počtu vnitřních oken a větší velikosti abecedy se zvyšuje úroveň podrobnosti získaných informací z dat, v takovém případě může nastat i stav, kdy není možné rozeznat jakékoliv vztahy mezi vzorky. Na druhou stranu při nastavení příliš nízkých hodnot stejných parametrů může dojít k nadměrné generalizaci získaných výsledků.

5 Závěr

Hlavním cílem této bakalářské práce bylo vytvořit výpočetní modul v prostředí .NET pro analýzu dat pocházejících z chytrých budov.

Teoretická část práce byla věnována definici a popisu chytrých budov především z hlediska energy managementu. Byla provedena studie problematiky zpracování archivních dat z měření veličin v chytrých budovách. V práci byly popsány různé přístupy pro analýzu dat, výčtem: klasifikace, regrese, detekce anomálií, shluková analýza a analýza časových řad. Každé z těchto oblastí byly věnovány podkapitoly pro popis nejpoužívanějších metod z hlediska jejich přínosu a možných komplikací při jejich aplikaci. Podrobnější popis byl věnován metodám shlukování metodou nejbližších středů a symbolic aggregate approximation.

V praktické části byly otestovány vybrané metody na datasetu měření spotřeby energie rodinného domu v průběhu jednoho roku. Před otestováním metod byla provedena explorační analýza dat, při které byly objeveny významné rozdíly v hodnotách spotřeby energie v měsících červen, červenec, srpen a září oproti zbytku roku. Tento jev následně způsoboval bimodální rozdělení dat. V závislosti na této informaci byly následně otestovány metody shlukování metodou nejbližších středu a symbolic aggregate approximation.

V prvním cyklu testování metody k-means byla použita vzdáleností metrika dynamic time warping. Data byla rozdělena do dvou skupin – letní měsíce a zbytek roku. Pro každou skupinu dat byla provedena analýza nastavení parametrů agregačního intervalu, velikosti okna klouzavého průměru a počtu výsledných klastrů. Výsledky analýzy byly zprostředkovány za pomocí metrik siluety a distortion score. V případě letních měsíců dosahovala metoda nejlepších výsledků pro 8 klastrů při nastavení 30minutového agregačního intervalu a velikosti okna klouzavého průměru nastavené na 47 vzorků. S totožnými parametry pouze s počtem klastrů nastaveným na hodnotu 10 dosahovala metoda nejlepších výsledků v případě druhé skupiny dat. S přihlédnutím k výsledným hodnotám siluety dosahovala metoda poměrně dobrých výsledků. V případě letních dat byla nejlepší hodnota siluety 0,799 a nejnižší 0,389. Druhá skupina dat dosahovala nejvyšší siluety 0,594 a naopak nejnižší 0,276. V druhém cyklu testování byla pro porovnání použita euklidovská vzdálenostní metrika při stejném rozdělení dat na dvě skupiny. Výsledné hodnoty siluety dosahovaly nižších hodnot než v případě prvního cyklu. Při aplikaci metody na letní měsíce dosáhla silueta nejvyšší hodnoty 0,574 a nejnižší

0,206 s pominutím klastru obsahující pouze jeden den. Zbytek roku dosahoval nejlepší hodnoty siluety 0,436 a nejhorší 0,120 opět s pominutím klastru obsahující jeden den. Jednotlivé klastry vykazovaly převážně snahu rozdělit vzorky dnů podle měsíců. Jeden klastr byl schopen identifikovat převážně všední dny. Nejzajímavější výsledek představoval klastr, který detekoval přechod mezi tarifem spotřeby energie letních měsíců a klasickou spotřebou zbytku roku. Jelikož zpracovávaný dataset obsahoval pouze časovou složku a měření spotřeby energie, bylo poměrně obtížné odhalit spojitosti mezi vzorky v jednotlivých klastrech, tj. pochopit výsledek shlukování. Pro hlubší pochopení výsledků klastrování by bylo potřeba znát i jiné informace, například: počasí, interní a externí teploty či vlhkost, počet lidí v budově, aktivní spotřebiče, aktivní energetický tarif, strukturu budovy a další.

Metoda symbolic aggregate approximation byla použita stylem explorační analýzy. Kde za pomocí rozličných způsobů vizualizací výsledků této metody bylo možné identifikovat v datech různé vzorce chování. V případě prvního cyklu kdy byla metoda použita na celý dataset, bylo možné za pomocí grafu četnosti kódování a teplotní mapy identifikovat anomálie z hlediska spotřeby energie. Příkladem může být průměrná spotřeba energie v časovém intervalu 12–18 hodin v období od 6. dubna do 10. dubna, kde spotřeba energie v uvedeném čase dosahovala nižších hodnot než v předchozích či následujících dnech. Primárně je tato metoda používána pro snížení dimenze zpracovávaných dat, ale v této práci byla prezentována z hlediska možného použití při explorační analýze dat za využití různých přístupů vizualizace.

Metody pro zpracování a vizualizaci dat byly implementovány a otestovány v prostředí .NET. Výsledný modul lze primárně využít pro detekci vzorců chování v datech. Z hlediska praktického použití je možné tento modul aplikovat v energy management systémech jak na straně distributora energie, tak i na straně spotřebitele. Z hlediska rozvoje tématu je možné otestovat a porovnat další shlukovací algoritmy za účelem zvýšení efektivity detekce vzorců chování v datech.

Použitá literatura

[1] AL DAKHEEL, Joud et al. Smart buildings features and key performance indicators: A review. *Sustainable Cities and Society*. 2020, roč. 61, s. 102328. ISSN 2210-6707. Dostupné z DOI: https://doi.org/10.1016/j.scs.2020.102328.

[2] SOVACOOL, Benjamin K. a Dylan D. FURSZYFER DEL RIO. Smart home technologies in Europe: A critical review of concepts, benefits, risks and policies. *Renewable and Sustainable Energy Reviews*. 2020, roč. 120, s. 109663. ISSN 1364-0321. Dostupné z DOI: https://doi.org/10.1016/j.rser.2019.109663.

[3] EUROSTAT. *Renewable energy statistics* [Eurostat Statistics Explained]. 2021. Dostupné také z: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics.

[4] MOLINA-SOLANA, Miguel et al. Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews*. 2017, roč. 70, s. 598–609. ISSN 1364-0321. Dostupné z DOI: https://doi.org/10.1016/j.rser.2016.11.132.

[5] KHALIL, Mohamad et al. Transfer Learning Approach for Occupancy Prediction in Smart Buildings. In: *2021 12th International Renewable Engineering Conference (IREC)*. 2021, s. 1–6. Dostupné z DOI: 10.1109/IREC51415.2021.9427869.

[6] CAPOZZOLI, Alfonso et al. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. *Energy*. 2018, roč. 157, s. 336–352. ISSN 0360-5442. Dostupné z DOI: https://doi.org/10.1016/j.energy.2018.05.127.

[7] REVATI, G. et al. Smart Building Energy Management: Load Profile Prediction using Machine Learning. In: *2021 29th Mediterranean Conference on Control and Automation (MED)*. 2021, s. 380–385. Dostupné z DOI: 10.1109/MED51440.2021.9480170.

[8] SEEM, John E. Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy and Buildings*. 2007, roč. 39, č. 1, s. 52–58. ISSN 0378-7788. Dostupné z DOI: https://doi.org/10.1016/j.enbuild.2006.03.033.

[9] WARREN LIAO, T. Clustering of time series data—a survey. *Pattern Recognition*. 2005, roč. 38, č. 11, s. 1857–1874. ISSN 0031-3203. Dostupné z DOI: https://doi.org/10.1016/j.patcog.2005.01.025.

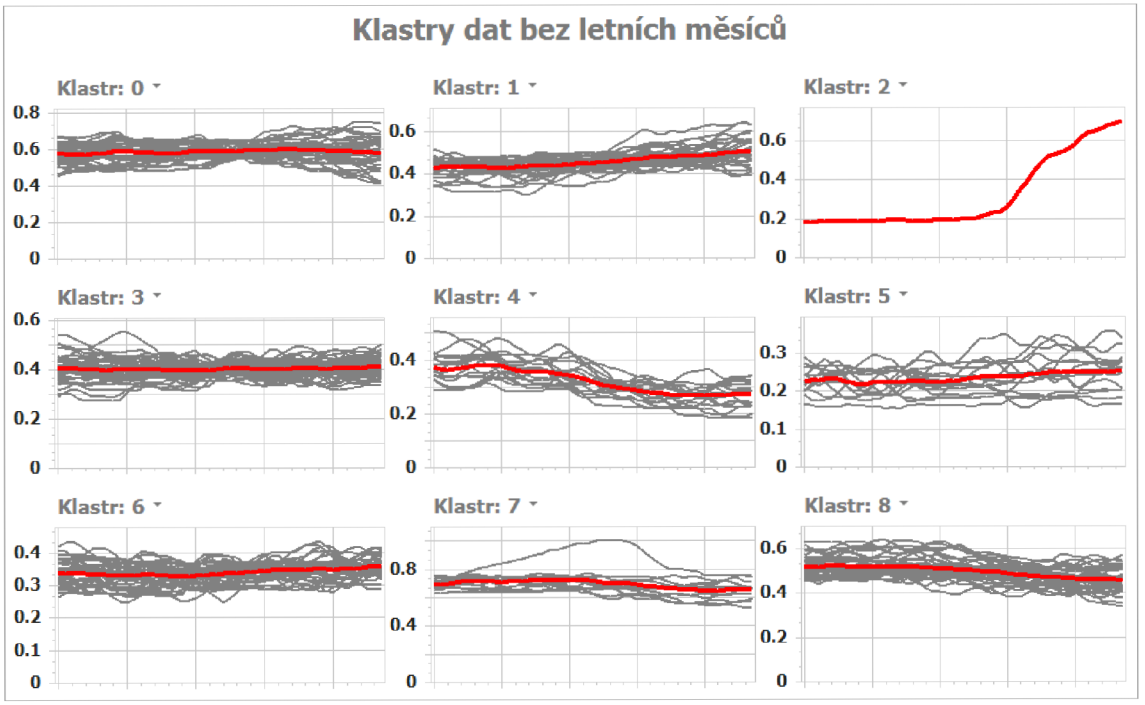
[10] KODINARIYA, Trupti a Prashant MAKWANA. Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*. 2013, roč. 1, s. 90–95.

[11] AMIDON, Alexandra. How to Apply K-Means Clustering to Time Series Data. *Towards Data Science* [online]. 2020 [cit. 2023-04-02]. Dostupné z: https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3.

A Příloha - graf klastrů dat bez letních měsíců prvního cyklu

Obrázek A.1: První cyklus - graf nalezených shluků dat bez letních měsíců

B Příloha - graf klastrů dat bez letních měsíců druhého cyklu



Obrázek B.1: Druhý cyklus - graf nalezených shluků dat bez letních měsíců

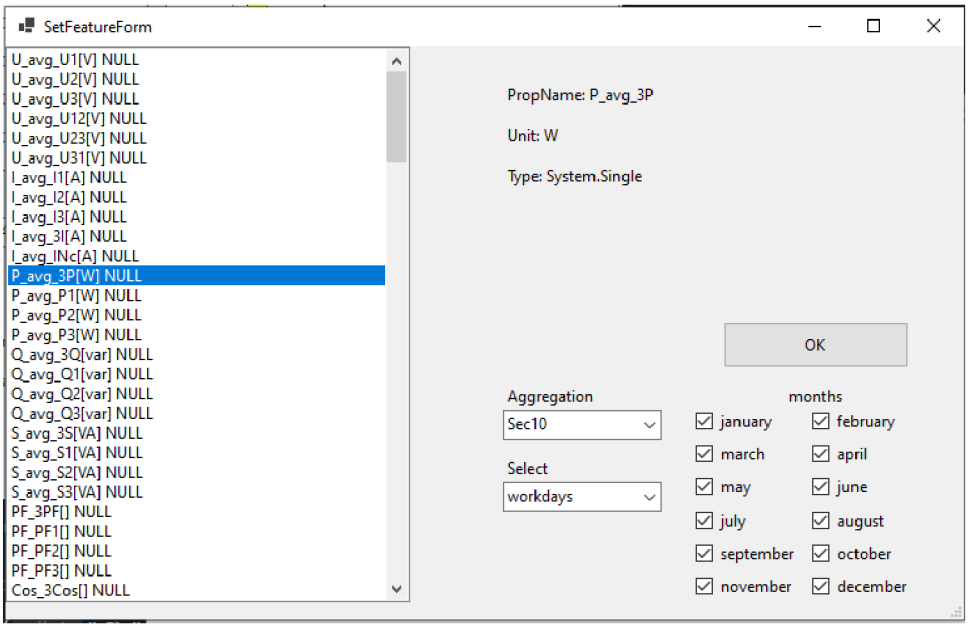
C Příloha - Návod k použití modulu

Navržený modul je součástí interního systému, který disponuje nástroji pro načtení dat. Tento návod neobsahuje postup pro načtení dat, ale zaměřuje se výhradně na práci s již načtenými daty v rámci modulu.

Po nahrání dat do systému se zobrazí úvodní okno s horním panelem tlačítek.

1. Výběr a filtrace atributu datasetu

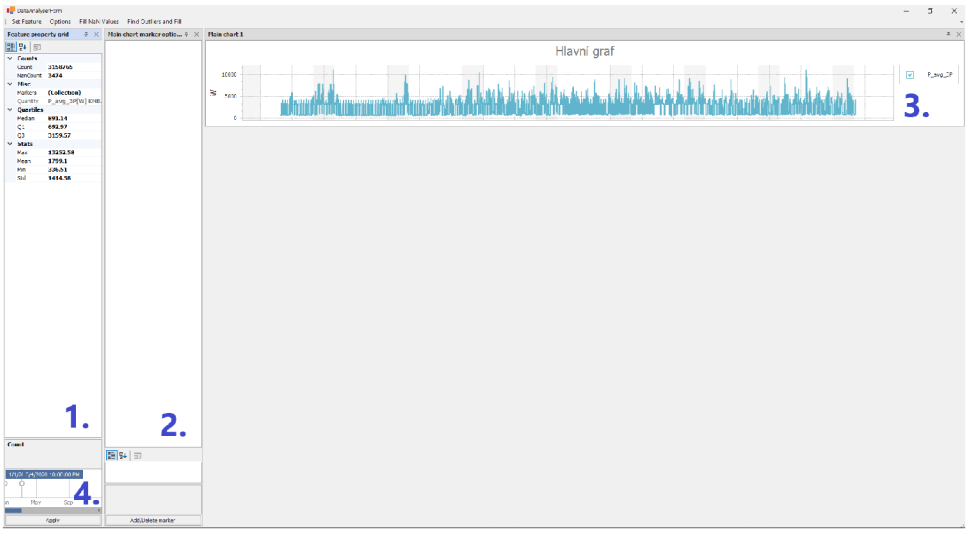
Stisknutím tlačítka “Set Feature” v horním panelu úvodního okna, se zobrazí okno pro výběr atributu.



- V levé části okna lze vybrat právě jeden atribut ze seznamu atributů nahraného datasetu.
- V poli “Aggregation” lze vybrat agregační interval. K dispozici jsou přesně definované intervaly. Je možné zvolit i možnost “None”, tedy bez agregačního intervalu.
- V poli “Select” lze filtrovat měření podle toho, zda se jedná o všední dny (workdays) nebo víkendy (weekends).
- Pomocí zaškrtávacích polí lze filtrovat měření podle měsíců.

Stisknutím tlačítka “OK” se potvrdí výběr a provede se selekce dat.

Po selekci dat se v úvodním okně zobrazí panely.

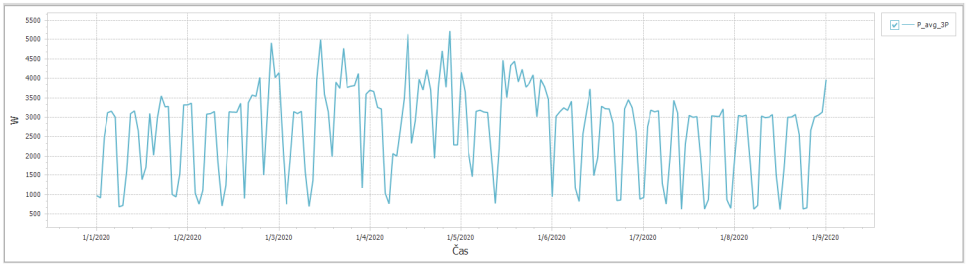


1. Panel obsahující tabulku popisných dat vybraného atributu.
    - “Count” (počet validních vzorků),
    - “NanCount” (počet chybějících vzorků),
    - “Markers” (seznam značek),
    - “Quantity” (název vybraného atributu),
    - “Median” (medián),
    - “Q1” (první kvartil),
    - “Q3” (třetí kvartil),
    - “Max” (maximální hodnota),
    - “Mean” (průměrná hodnota),
    - “Min” (minimální hodnota),
    - “Std” (směrodatná odchylka)
2. Panel obsahující seznam značek - viz 4. Značky.
3. Panel obsahující hlavní graf.
4. Panel obsahující časovou osu a tlačítko pro výběr časového intervalu.
    - Podržením levého tlačítka myši v poli časové osy lze vybrat časový interval.
    - Stisknutím pravého tlačítka myši v poli časové osy se nastaví týdenní časový interval.
    - Stisknutím tlačítka “Apply” se vybraný časový interval aplikuje a případně se projeví v aktivních grafech.

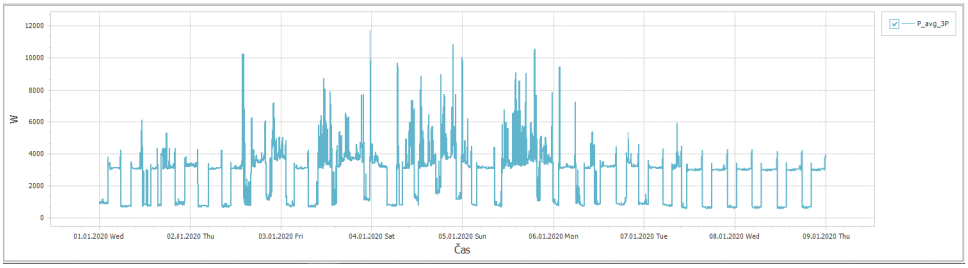
2. Vizualizace dat

Stisknutím tlačítka “Options” v horním panelu úvodního okna se zobrazí seznam tlačítek.

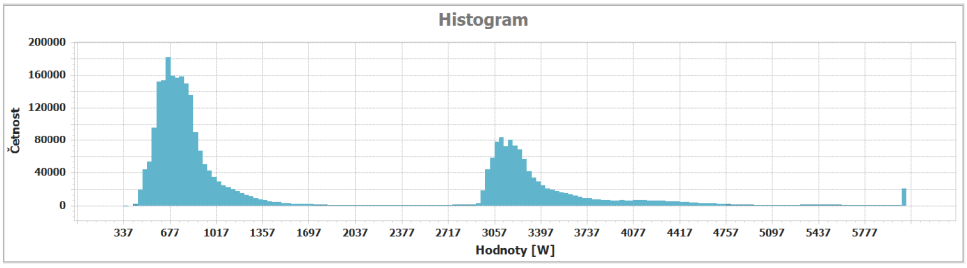
- “Line Graph”
  - Stisknutím tlačítka se zobrazí spojnicový graf s automatickou agregací dat, tj. v závislosti na úrovní přiblížení se mění agregační interval.



- “Line Graph (seconds)”
  - Stisknutím tlačítka se zobrazí spojnicový graf se statickým sekundovým agregačním intervalem.

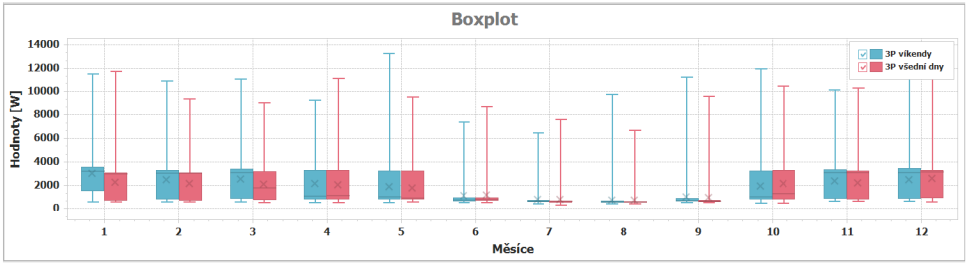


- “Histogram”
  - Stisknutím tlačítka se zobrazí histogram.

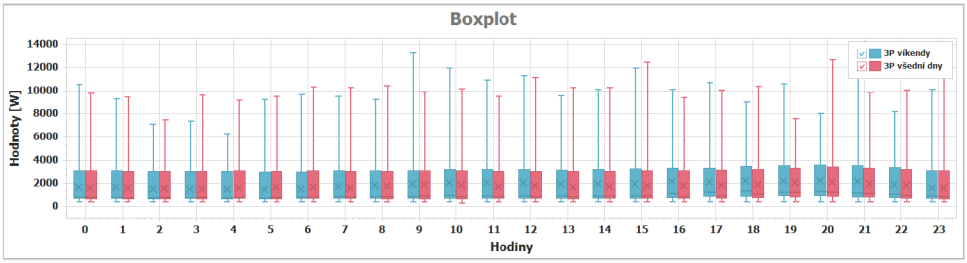


- “Markers” - viz 3. Metody pro analýzu dat.

- “Boxplot (month)”
    - Před použitím je nutné data očistit od chybějících hodnot. (viz 5. Chybějící a odlehlé hodnoty)
    - Stisknutím tlačítka se zobrazí krabicový graf s rozdělením dat podle měsíce měření.



- “Boxplot (hour)”
    - Před použitím je nutné data očistit od chybějících hodnot. (viz 5. Chybějící a odlehlé hodnoty)
    - Stisknutím tlačítka se zobrazí krabicový graf s rozdělením dat podle hodiny měření.

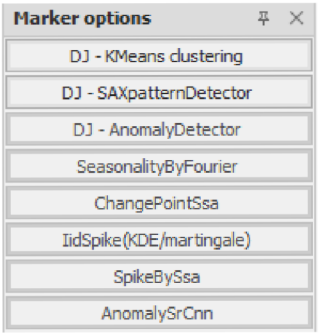


2.1. Ovládání grafů

- **Přiblížení**
    - Podržením klávesy “shift” a levého tlačítka myši v poli grafu lze přiblížit vybraný interval.
    - Podržením klávesy “shift” a stisknutím levého tlačítka myši v poli grafu lze přiblížit oblast kurzoru.
- **Oddálení**
    - Podržením klávesy “alt” a stisknutím levého tlačítka myši v poli grafu lze oddálit křivky.
- **Legenda**
    - Za pomocí zaškrtávacích tlačítek v legendě lze skrýt a zobrazit vybrané křivky.

3. Metody pro analýzu dat

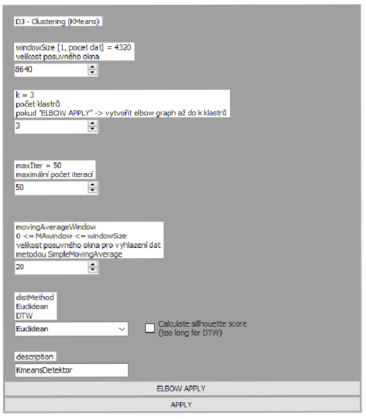
Stisknutím tlačítka “Options” v horním panelu úvodního okna se zobrazí seznam tlačítek. Stisknutím tlačítka “Markers” v seznamu se v pravém dolním rohu úvodního okna zobrazí panel se seznamem metod.



Stisknutím některého z tlačítek se zobrazí pod seznamem panel s nastavením parametrů pro vybranou metodu. Jednotlivé parametry jsou v panelu stručně popsány.

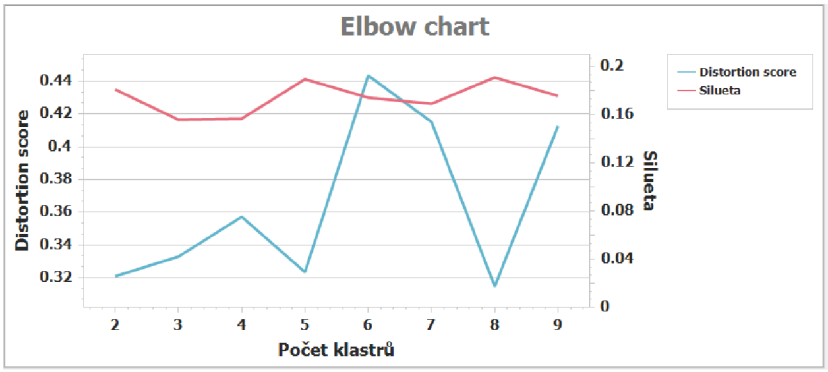
3.1. DJ - KMeans clustering

- Metoda shlukování metodou nejbližších středů

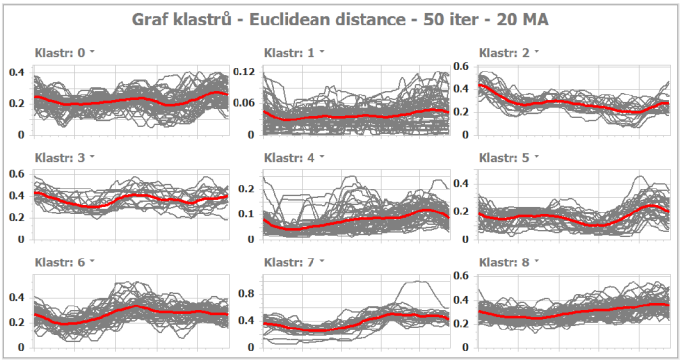


- Parametry:
  - “windowSize” (velikost posuvného okna, tj. kolik vzorků tvoří jeden vzorek v procesu klastrování),
  - “k” (výsledný počet klastrů),
  - “maxIter” (maximální počet iterací),
  - “movingAverageWindow” (velikost posuvného okna),
  - “distMethod” (vzdálenostní metrika, 2 možnosti - euklidovská metrika (Euclidean) a dynamic time warping (DTW)),
  - “Calculate silhouette score” (zda během zpracování metody počítat siluetu),
  - “description” (popis výsledné značky)

- Stisknutím tlačítka “ELBOW APPLY” se vytvoří Elbow chart a do panelu seznamu značek se přidá element popisující výsledek metody.

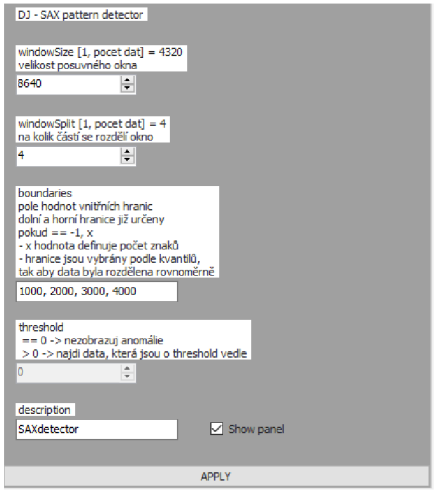


- Stisknutím tlačítka “APPLY” se vytvoří graf s nalezenými klastry a do panelu seznamu značek se přidá element popisující výsledek metody.

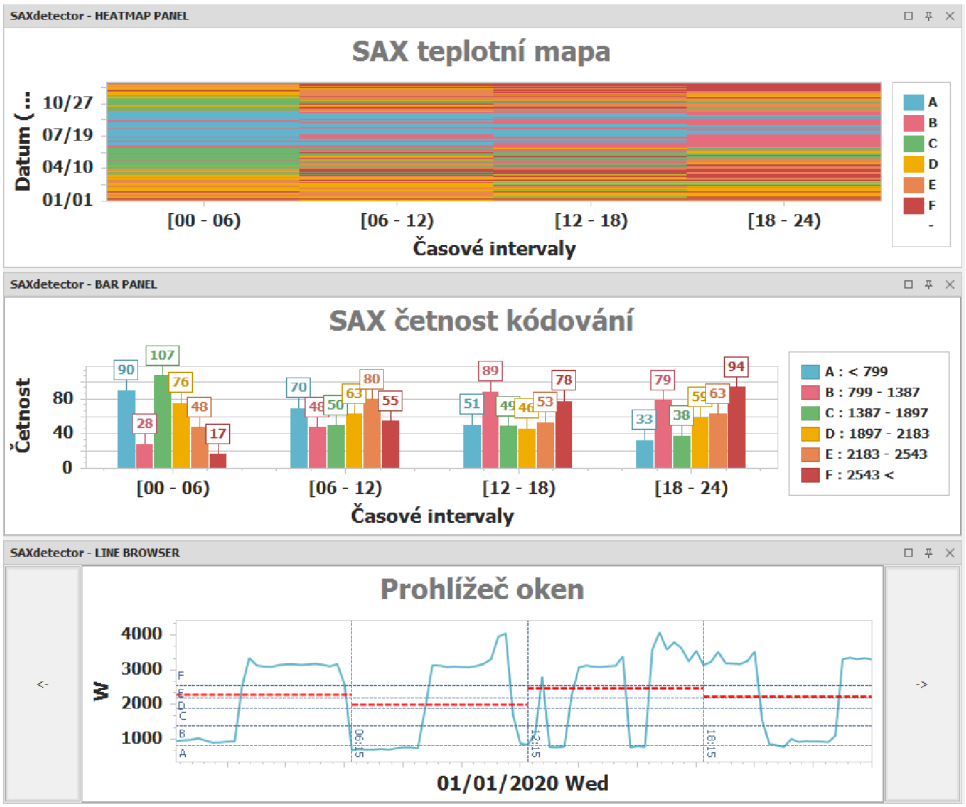


3.2. DJ - SAXpatternDetector

- Metoda symbolic aggregate approximation

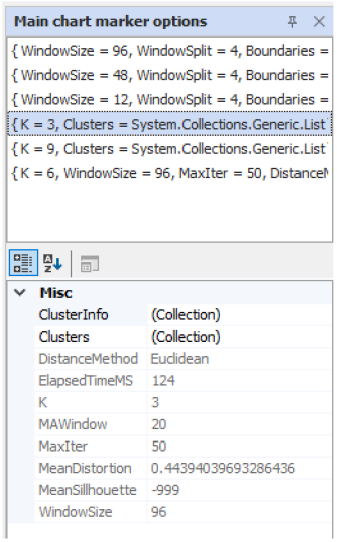


- Parametry:
    - “windowSize” (velikost posuvného okna, tj. kolik vzorků tvoří jeden vzorek v metodě symbolic aggregate approximation),
    - “windowSplit” (na kolik vnitřních oken se rozdělí hlavní okno),
    - “boundaries” (pole hodnot vnitřních hranic),
    - “description” (popis výsledné značky),
    - “Show panel” (zda nechat zobrazit výsledné grafy)
- Stisknutím tlačítka “APPLY” se vytvoří 3 grafy (teplotní mapa, četnost kódování a prohlížeč oken) a do panelu seznamu značek se přidá element popisující výsledek metody.



4. Značky

Při aplikaci libovolné metody se její výsledky a statistiky uloží jako element v seznamu značek, který je k dispozici v panelu úvodního okna.



Výběrem libovolného elementu ze seznamu pomocí levého tlačítka myši se jeho položky zobrazí v tabulce pod seznamem.

5. Chybějící a odlehlé hodnoty

- Stisknutím tlačítka “Fill NaN Values” v horním panelu úvodního okna lze doplnit chybějící hodnoty.
- Stisknutím tlačítka “Find Outliers and Fill” v horním panelu úvodního okna lze detekovat odlehlé hodnoty a nahradit je validními hodnotami.