



„Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ“

Názov projektu :

Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov

ITMS : 26240120039

Názov výstupu :

Pracovný balík 6: Kvalita dát - kvalita vstupných dát, kvalita zdrojov ako základ pre ďalšie procesy - získavanie veľkých objemov dát s prihliadnutím na kvalitu dát

Obsah

Obsah.....	2
Zoznam obrázkov	3
Zoznam skratiek	4
1. Úvod.....	5
2. Prehľad použiteľných close-data zdrojov. Neverejné dáta.....	6
2.1 Štruktúra hydrometeorologických dát používaná v energetike	6
2.2 Prehľad použiteľných zdrojov dát. Verejné dáta.....	19
3. Základné aspekty správy kvality dát.....	27
4. Metodologický prehľad štatistických nástrojov použiteľných pri tvorbe a testovaní modelov predikcie a optimalizácie výroby a spotreby elektrickej energie na Slovensku.....	31
5. Skúmanie kvality spracovaných energetických dát.....	41
5.1.1 Predspracovanie dát pre experimenty.....	44
6. Ochrana citlivých údajov v kontexte dát zo siete smartgrid.....	45
6.1 Kontrola kvality dátových zdrojov a automatizácia anonymizačných procesov	59
7. Štatistické metódy.....	67
7.1 Jednorozmerné dáta.....	67
7.2 Viacrozmerné dáta	69
7.3 Údaje časových radov.....	71
7.4 Vzorkovanie dát.....	72
8. Logické metódy.....	73
8.1 Funkčné závislosti dát.....	73
8.2 Inkluzívne závislosti dát	75
8.3 Sekvenčné závislosti dát	76
8.4 Ďalšie typy závislosti dát.....	77
9. Hybridné a iné metódy	78
10. Extrapolácia nedostatočných dátových vzoriek	79
11. Záver	92
12. Literatúra	93

Zoznam obrázkov

Obr. 1. Celková spotreba 65 slovenských odberno-odovzdávacích miest za obdobie troch mesiacov (93 dní).	41
Obr. 2. Priemerný týždenný priebeh spotreby elektrickej energie v jednom OOM.	42
Obr. 3. Rozdelenie OOM podľa prvých dvoch čísel PSČ.	43
Obr. 4 Graf normálneho a Zipfovho rozdelenia	67
Obr. 5 Príklad súmerného a nesúmerného dvojrozmerného normálneho rozdelenia.....	69
Obr. 6 Základná schéma (štruktúra) analyzovanej bázy dát a použitá terminológia.....	72
Obr. 7 Príklad relácie Zákazníci (prevzatý z [15])	73
Obr. 8 Príklad jednoduchých schém pre internetový obchod	75
Obr. 9 Príklad reálnych „usporiadaných“ dát	76

Zoznam skratiek

CFD	Conditional Functional Dependency
CID	Conditional Inclusion Dependency
CSD	Conditional Sequential Dependency
FD	Functional Dependency
MAD	Median Absolute Deviation
MCD	Minimum Covariance Determinant
MVE	Minimum Volume Ellipsoid
OCR	Optical Character Recognition
OD	Order Dependency
OGK	Orthogonalized Gnanadesikan-Kettering
PNRS	Personal Name Recognizing Strategy

1 Úvod

V súčasnosti sa v rámci mnohých spoločností ako aj v iných oblastiach generujú veľké objemy dát, pričom aj rýchlosť ich prísunu neustále narastá. Vzhľadom na to, že tieto dáta taktiež reprezentujú rôzne heterogénne dáta, kvalita výsledných zhromaždených dát je výrazne ovplyvnená týmito faktormi a nedosahuje požadovanú úroveň. Niektoré zdroje uvádzajú, že tieto chybné dáta stoja americký obchod 600 miliárd dolárov ročne. Miera chybných dát sa u spoločností pohybuje na úrovni 1 – 5% (u niektorých dokonca až 30%) [1]. Táto situácia vytvára tlak na potrebu správy kvality dát, aby sa zabezpečilo, že dáta v databázach budú konzistentné, presné, úplné, načas a jedinečne reprezentovali entity reálneho sveta, ktorých sa týkajú. To znamená, že narastá v priemysle potreba účinných manažovacích systémov pre udržanie požadovanej kvality dát, čo znamená, účinne detekovať a opravovať chyby v dátach. S príchodom veľkých objemov dát, ktoré sú charakterizované pojmami ako objem, rýchlosť a rôznorodosť [2], táto potreba získala ešte väčšiu váhu ako doteraz. Ku uvedeným pojmom sa pričleňuje ešte ďalší dôležitý a to vierohodnosť alebo správnosť dát. Pri veľkých objemoch dát a ich rýchlosti hromadenia je nutné nájsť aj rýchle metódy na pochopenie ako aj odstránenie chybných dát. Mnohé štruktúrované dáta, často z mnohých zdrojov, sú integrované navzájom, a preto vzniká dôležitá úloha určiť sémantiku dát a pochopiť vzájomné vzťahy medzi ich atribútmi. Je nutné nájsť alebo najlepšie naučiť sa priamo z dát pravidlá, ktoré nám umožnia odstrániť chybné dáta a zvýšiť tak ich kvalitu [3].

Metódy, ktoré sa pokúšajú hľadať a prípadne opravovať chyby v dátach, sa označujú ako metódy na čistenie dát (data cleaning/cleansing) [4]. V prvej fáze je nutné vykonať analýzu samotných dát a zistiť typ chýb a nekonzistentností dát. Potom je možné aplikovať príslušné metódy, ktoré budú dáta opisovať vhodným modelom, pravidlami alebo metrikami, a ktoré poukážu na chyby a nezrovnalosti v dátach. Následne je nutné aplikovať opravy a vytvoriť novú „čistú“ bázu dát, ktorá sa nahrá do konečného skladu dát (data warehouse). V súčasnosti sa stretávame najmä s dvomi prístupmi spojenými s kvalitou dát. Prvý, ktorý sa snaží nájsť v dátach záznamy (record matching), ktoré sú identické (duplikáty, tuples), ktoré identifikujú ten istý fyzický objekt. Druhý sa týka priamo opravy dát (data repairing), ktorý sa snaží nájsť druhú databázu údajov tým, že hľadá a opravuje nezrovnalosti v pôvodnej databáze. Väčšina systémov na čistenie dát poskytovala prvý prístup a len niektoré ten druhý, pričom boli vykonávané ako oddelené a nezávislé procesy. Ukázala sa ale potreba ich vzájomnej interakcie.

Aplikácia metód čistenia dát sa odvíja aj od typu dát, či sú dané dáta neštruktúrované, štruktúrované alebo pološtruktúrované (tzv. schemaless).

V nasledovných kapitolách budú okrem iného rozobrané základné typy metód použiteľných na čistenie dát. U mnohých týchto metód existujú rôzne variácie navrhnuté rôznymi autormi.

2. Prehľad použiteľných close-data zdrojov. Neverejné dáta.

Táto časť výstupnej správy je formálnym výstupom analýzy vhodnosti rôznych zdrojov dát, vybraných na základe ich prítomnosti v dátových štruktúrach systému súčasného sledovania pohybu výroby a spotreby elektrickej energie na Slovensku. Cieľom analýzy bolo nájsť a popísať neverejné zdroje dát vhodné k tvorbe inteligentných systémov, podporujúcich alebo nahradzujúcich dátové zdroje súčasného bilančného modelu estimácie odchýlky výroby a spotreby. Pokúsime sa podať návrhy na vylepšenie zberu, extrakcie a implementácie týchto dát v kontexte modelov novej generácie.

Pri tomto zadaní sme prihliadali:

- na hodnotu metadát získaných z potenciálnych dátových zdrojov,
- na relevantnú legislatívu, najmä: 1.) Zákon č. 211/2000 Z. z. o slobodnom prístupe k informáciám a o zmene a doplnení niektorých zákonov, 2.) Zákon č. 275/2006 Z. z. o informačných systémoch verejnej správy, 3.) Štandardy pre informačné systémy úradov verejnej správy, a 4.) Zákon č. 540/2001 Z. z. o štátnej štatistike,
- na bezpečnosť jednotlivých neverejných zdrojov dát z pohľadu anonymity dát,
- na kvalitu potenciálnych dátových zdrojov,
- na náročnosť extrakcie a implementácie dát z neverejných zdrojov do aktuálne bežiacich systémov,
- na relevanciu potenciálnych dátových zdrojov na pri cieľoch krátkodobej predikcie (hodinový, denný interval) a cieľoch dlhodobej predikcie (týždenný a mesačný interval) spotreby a výroby elektrickej energie,
- na charakteristiky súčasného systému poskytovania, evidencie a využitia údajov o výrobe a spotrebe elektrickej energie.

Legislatíva prístupu k dátam, dátovým produktom a informáciám

Pred popísaním dátových zdrojov s analytickým potenciálom, považujeme za dôležité uviesť legislatívne predpoklady, ktoré ohraničujú ich extrakciu a použitie dát z neverejných zdrojov. Popíšeme tu legislatívny priestor, ktorého sa dotýkame našou analýzou.

Okrem zákona č. 211/2000 Z. z. o slobodnom prístupe k informáciám, ktorý nám určuje všeobecné mantinely našej interakcie s dátami vo vlastníctve orgánov štátnej správy, sa nás bytostne dotýka Zákon NR SR č. 540/2001 o štátnej štatistike. Tento zákon upravuje

podmienky získavania štatistických informácií potrebných na posudzovanie sociálno-ekonomického vývoja, postavenie a pôsobnosť orgánov vykonávajúcich štátnu štatistiku, úlohy orgánov verejnej moci v oblasti štátnej štatistiky, práva a povinnosti spravodajských jednotiek, ochranu dôverných štatistických údajov pred zneužitím, poskytovanie a zverejňovanie štatistických údajov, zabezpečovanie porovnateľnosti štatistických informácií a plnenie záväzkov vyplývajúcich z medzinárodných zmlúv v oblasti štátnej štatistiky, ktorými je SR viazaná.

Zákon o štátnej štatistike okrem iného definuje informačné registre, ktorými disponuje Štatistický úrad SR. Podľa § 20, odseku (1), úrad vedie najmä register organizácií, poľnohospodársky register, register ubytovacích zariadení, register sčítacích obvodov, register priestorových jednotiek a register subjektov zahraničného obchodu.

Register organizácií

O údajoch a obsahu registra organizácií sa píše v Zbierke zákonov č. 298/2010, Čiastka 113, § 21. Do registra organizácií sa zapisujú právnické osoby a fyzické osoby - podnikatelia na základe údajov, na ktoré má Štatistický úrad SR právo podľa zákona o štátnej štatistike a podľa ďalších osobitných zákonov. Register sa využíva na štatistické účely a údaje v ňom evidované majú informatívny charakter. Informácie o počte subjektov evidovaných v registri organizácií sa mesačne zverejňujú. Na jednoznačnú identifikáciu subjektov evidovaných v registri organizácií sa používa identifikačné číslo (IČO).

Používateľom, ktorí potrebujú informácie z registra podľa vlastných špecifických požiadaviek, ponúkajú možnosť objednať si kompletnú databázu, resp. výber z databázy, vrátane programového vybavenia. Údaje uvedené v registri organizácií, sú verejné údaje a možno ich poskytnúť každému, kto o ne požiada. Súkromným a právnickým žiadateľom sú poskytnuté za odplatu. Údaje o rodnom čísle, adrese trvalého pobytu a štátnej príslušnosti podnikateľa a o počte zamestnancov sú však identifikované ako neverejné údaje.

Poľnohospodársky register

O údajoch a obsahu registra je popísaný v Zbierke zákonov č. 298/2010, Čiastka 113, § 22. Poľnohospodársky register, poskytuje údaje o právnických osobách a fyzických osobách, ktoré vykonávajú poľnohospodársku výrobu. Tieto údaje sú neverejné.

Register ubytovacích zariadení

O údajoch a obsahu registra ubytovacích zariadení sa píše v Zbierke zákonov č. 298/2010, Čiastka 113, § 23. Register ubytovacích zariadení poskytuje dáta o ubytovacích zariadeniach, ktoré poskytujú prechodné ubytovanie osôb. Poskytuje tiež údaje o ekonomických subjektoch, ktoré ich prevádzkujú. Register ubytovacích zariadení je verejným registrom.

Register sčítacích obvodov

Údaje a obsah registra sčítacích obvodov nájdeme v Zbierke zákonov č. 298/2010, Čiastka 113, § 24. Register sčítacích obvodov, poskytuje dáta o sčítacích obvodoch z celého územia Slovenskej republiky, ktoré sú vytvorené podľa osobitného zákona. Tento register je neverejný.

Register priestorových jednotiek

O údajoch a obsahu registra sa píše v Zbierke zákonov č. 298/2010, Čiastka 113, § 25. Na účely priestorovej identifikácie štatistických údajov evidujú priestorové jednotky vo väzbe na územné a správne usporiadanie Slovenskej republiky. Údaje evidované v registri priestorových jednotiek sú verejné.

Register subjektov zahraničného obchodu

Charakteristiky údajov a obsahu registra subjektov zahraničného obchodu nájdeme v Zbierke zákonov č. 298/2010, Čiastka 113, § 25a. Do registra subjektov zahraničného obchodu sa zapisujú všetky spravodajské jednotky uskutočňujúce obchod s tovarom medzi Slovenskou republikou a inými členskými štátmi Európskej únie a ďalšie subjekty uskutočňujúce obchod s tovarom medzi Slovenskou republikou a nečlenskými krajinami Európskej únie. Údaje z tohto registra sú neverejné.

Manipulácia a extrakcia dôverných a neverejných štatistických údajov

Zákon o štátnej štatistike takisto definuje, ktoré z údajov v týchto registroch sú údajmi dôvernými, ktoré sú údajmi verejnými a ktoré neverejnými. Dôverným štatistickým údajom je údaj týkajúci sa jednotlivcej spravodajskej jednotky, ktorý umožňuje jej priamu alebo nepriamu identifikáciu a bol získaný na štatistické účely podľa tohto zákona. Za priamu identifikáciu sa považuje identifikácia spravodajskej jednotky podľa obchodného mena, sídla a identifikačného čísla organizácie (ďalej len „identifikačné číslo“) alebo podľa mena a priezviska, adresy trvalého bydliska alebo rodného čísla. Za nepriamu identifikáciu sa považuje identifikácia spravodajskej jednotky využitím iných údajov získaných štatistickým zisťovaním (Z. z. č. 298/2010, Čiastka 113, § 20, odsek f)). Podľa § 30, odseku (1) toho istého zákona, sa dôverné štatistické údaje nesmú bez súhlasu právnických osôb alebo fyzických osôb, ktoré ich poskytli a ktorých sa týkajú, zverejňovať ani komukoľvek oznamovať, ani použiť na iné než štatistické účely.

Positívne sú však dve informácie. Prvá je, že podľa § 30, odseku (6) toho istého zákona, sa za dôverné štatistické údaje nepovažujú údaje, ktoré sa nachádzajú vo všeobecne dostupných

zdrojoch vytvorených v súlade s právnymi predpismi (napríklad adresáre, zoznamy, číselníky, registre), údaje, ktoré o sebe zverejnili spravodajské jednotky alebo ktoré sa dajú jednoduchým spôsobom zistiť, a údaje, ktoré vznikli sumarizáciou dostatočného počtu (najmenej troch) dôverných štatistických údajov tak, že nemožno určiť, ktorých spravodajských jednotiek sa týkajú. Druhá je, že podľa § 30, odseku (4) toho istého zákona, orgány vykonávajúce štátnu štatistiku môžu poskytnúť dôverné štatistické údaje na vedecké účely, právnickým osobám alebo výskumným pracovníkom, ktorých základným poslaním je vedecký výskum, ak právnická osoba, ktorej sa tieto údaje poskytnú, zabezpečí podmienky na ich ochranu. Údaje sa poskytnú vo forme, ktorá neumožňuje priame určenie právnickej osoby alebo fyzickej osoby, ktorých sa poskytnuté údaje týkajú.

Podľa § 31 toho istého zákona, je tak úrad oprávnený poskytovať štatistické informácie získané štatistickým zisťovaním za cenu dohodnutú v súlade s cenovými predpismi, a to každému, kto o ne požiada. Napriek tomu, že s pojmami “neverejný údaj” a “verejný údaj” sa v tomto zákone intenzívne narába, ich presná definícia v ňom chýba. Poskytovaním týchto (nedostatočne definovaných) neverejných údajov Štatistický úrad SR monetizuje tieto dáta. Pekným príkladom takejto zákonom podporovanej monetizácie dát je napríklad Slovenský hydrometeorologický ústav, ktorý je ochotný poskytnúť konkrétnu informáciu, alebo informáciu o ich dátových balíkoch a produktoch, no nie čisté dáta. Toto im dovoľuje zákon č. 201/2009 Z. z. o štátnej hydrologickej službe a štátnej meteorologickej službe, ktorý v § 13 v odseku (3) hovorí, že hydrologické produkty a meteorologické produkty, hydrologické informácie a meteorologické informácie sa poskytujú bezodplatne orgánom štátnej a verejnej správy, súdom a Národnej rade Slovenskej republiky pre výkon ich právomocí. V ostatných prípadoch sa hydrologické produkty a meteorologické produkty a hydrologické informácie a meteorologické informácie poskytujú odplatne.

Štatistický úrad Slovenskej republiky

Štatistický úrad SR je ústredným orgánom štátnej správy Slovenskej republiky pre oblasť štátnej štatistiky. Úrad plní úlohy najmä v kontexte zákona č. 540/2001 Z. z. o štátnej štatistike v znení neskorších predpisov. V kontexte našej analýzy je dôležité, že Štatistický úrad SR:

- zostavuje v súčinnosti s ministerstvami a štátnymi organizáciami program zisťovaní,
- určuje metodiku štatistických zisťovaní, zhromažďuje a spracováva štatistické údaje,
- vytvára, zverejňuje a spravuje v súčinnosti s ministerstvami a štátnymi organizáciami štatistické klasifikácie, číselníky a registre,
- určuje spôsob tvorby registrov, prideluje a oznamuje identifikačné čísla,

- vypracováva analýzy vybraných charakteristík sociálno-ekonomického a ekologického vývoja Slovenskej republiky,
- spolupracuje s medzinárodnými orgánmi a organizáciami pri zavádzaní štandardov a klasifikácií v oblasti štatistiky,
- zverejňuje výsledky štatistických zisťovaní za SR a pravidelne poskytuje štatistické informácie.

V súčasnosti zbierané dáta

Z pohľadu relevantnosti dát, Štatistický úrad SR reprezentuje prehľadný zdroj dát, kaskádovito organizovaný do dátových celkov na úrovni oblastí, krajov okresov a obcí. Okrem základných informácií ako celková výmera obce a hustota obyvateľstva na km², sú zaujímavé najmä dáta o ich technickej vybavenosti, ktoré korelovaním s dátami o užívaní elektrickej energie odbernými miestami môžu byť použité pri identifikovaní spotrebných profilov obyvateľov. Ide najmä o informácie:

- existencia pobočky pošty (ideálne poloha a počet pobočiek).
- prítomnosť káblového distribučného systému pre televízne, zvukové a interaktívne multimedialne signály,
- verejný vodovod,
- verejná kanalizácia,
- rozvod plynu,
- najbližšia zastávka vlakov osobnej dopravy (názov zastávky a vzdialenosť).

K technickým dátam pri profilovaní spotrebného správania značne prispievajú demografické premenné. Štatistický úrad SR zbiera tieto, nám relevantné údaje:

- počet obyvateľov na danom územnom celku,
- rozdelenie podľa veku (predproduktívny, produktívny, poproduktívny),
- počty sobášov, rozvodov na danom územnom celku,

- štatistika mortality a pôrodnosti na danom územnom celku,
- rozloženie a pomer národností a vierovyznaní na danom územnom celku,
- zamestnanosť a nezamestnanosť na danom územnom celku.

V neposlednom rade, môžu byť nápomocné dáta o existujúcej infraštruktúre na danom území, nakoľko táto výrazne ovplyvňuje životný štýl populácie a teda aj ich spotrebné správanie. Štatistický úrad SR zbiera tieto, nám relevantné údaje:

- športoviská: kúpaliská, telocvične, ihriská,
- kultúra: knižnica, videopožičovňa, kino,
- zdravotnícka infraštruktúra: lekárne, ambulancie (pre dospelých, pre deti a dorast, stomatologická, gynekologická),
- životné prostredie: spracovanie komunálneho odpadu,
- služby: predajňa potravinárskeho tovaru, pohostinské odbytové stredisko, predajňa nepotravinárskeho tovaru, predajňa pohonných látok, zariadenie pre údržbu a opravu motorových vozidiel, predajňa súčiastok a príslušenstva pre motorové vozidlá, hotel (motel, botel), penzión *** až *, turistická ubytovňa ** až *, chatová osada *** až *, kemping **** až *, ostatné hromadné ubytovacie zariadenia, komerčná poisťovňa, komerčná banka, bankomat.

Kvalita dát

Proces a frekvencia zberu a analýzy štatistických dát Štatistickým úradom SR sa riadi vyhláškou: “Program štátnych štatistických zisťovaní na trojročné obdobia”, ktorú zostavuje úrad v súčinnosti s ministerstvami a štátnymi organizáciami. Program obsahuje všetky dôležité a potrebné zisťovania vrátane zisťovaní vykonávaných inými ústrednými orgánmi a ministerstvami, ktoré majú prebehnúť v nasledujúcom trojročnom období. Tento program je k dispozícii na vyžiadanie, hneď po jeho schválení. Okrem zisťovaní zahrnutých v programe, vykonáva úrad i osobitné zisťovania u fyzických osôb (napríklad sčítanie obyvateľov domov a bytov, poľnohospodárske súpisy a.i.).

Hodnota metadát získaných zo zdrojov Štatistického úradu SR je značná. Najmä pri vytváraní spotrebných profilov užívateľov ako jednej dimenzie bilančného modelu estimácie odchýlky výroby a spotreby elektrickej energie na Slovensku. Ako sme písali v časti Legislatíva, získateľ

prístup k základným dátam a nie len agregovaným dátam si bude vyžadovať bližšiu spoluprácu s úradom.

Náročnosť extrakcie a implementácie dát

Základné dáta poskytované Štatistickým úradom SR je možné získať cez webové rozhranie. Masívne sťahovanie následne zabezpečujú nástroje, špecifické pre konkrétne databázy (napr. PX-Web a PX-iMap pre databázu regionálnej štatistiky). Jednotlivé databázy sa od seba líšia v podmienkach používania ich nástrojov a prístupu k surovým dátam.

Databáza regionálnej štatistiky

Databáza regionálnej štatistiky obsahuje časové rady ukazovateľov hospodárskeho a sociálno-ekonomického vývoja za regióny Slovenskej republiky (okresy, kraje a oblasti). Regionálne údaje z rôznych štatistických okruhov sú prezentované vo forme tabuliek v ročných časových radoch. Napriek tomu, že administratívne a organizačnou štruktúrou patrí pod Štatistický úrad SR, uvádzame je tu ako samostatný zdroj, z dôvodu, že poskytuje inú štruktúru dát a má samostatné podmienky ich poskytovania.

V súčasnosti zbierané dáta

Z pohľadu relevantnosti dát, databáza regionálnej štatistiky reprezentuje prehľadný zdroj, zameraný na dátové celky na úrovni regiónov a okresov. Okrem základných informácií o rozlohe regiónov a hustote obyvateľstva, sú obsahom databázy dátové rady v témach:

- **Demografia:** počet obyvateľov a migrácia obyvateľstva podľa demografických premenných.
- **Trh práce:** Výberové zisťovanie pracovných síl, Štruktúra miezd - výberové zisťovanie, Úplné náklady práce - výberové zisťovanie, Evidovaná nezamestnanosť, Voľné pracovné miesta, Zamestnanosť a mzdy (hospodárstvo SR - štvrťročné, podniky a historické údaje).
- **Sociálne štatistiky:** Príjmy, výdavky a spotreba domácností, Príjmy a životné podmienky domácnosti.
- **Sociálna ochrana:** Výdavky na dôchodky, Vyplatené dôchodky, Príjmy zdravotných poisťovní z poisťného, Úhrady zdravotných poisťovní za zdravotnú starostlivosť, Príjmy Sociálnej poisťovne z poisťného, Výdavky Sociálnej poisťovne na sociálne dávky, Sociálne zariadenia.

- **Zdravie:** Pracovná neschopnosť - prípady, Odškodnené pracovné úrazy a choroby z povolania, Zdravotnícke zariadenia, Nemocenský poistení a zamestnanci na rizikových pracoviskách.
- **Kriminalita a požiare:** Trestné činy, Násilné trestné činy, Požiare.
- **Makroekonomické štatistiky:** Regionálne účty (Regionálny hrubý domáci produkt, Regionálny hrubý domáci produkt na obyvateľa, Hrubá pridaná hodnota v bež. cenách, Tvorba hrubého fixného kapitálu v bež. cenách, Regionálna zamestnanosť a zamestnanci, Účet rozdelenia prvotných dôchodkov domácností, Účet rozdelenia druhotných dôchodkov domácností), Dlhodobý majetok (Byty dokončené podľa počtu izieb, Dokončené byty v rodinných domoch za obce, Dokončené byty v obciach).
- **Financie:** Rozmiestnenie organizačných jednotiek komerčných bánk v SR, Vklady a úvery, Skupina pre bankovníctvo.
- **Organizačná štatistika:** Fyzické osoby - podnikatelia podľa právnych foriem za obce SR, Slobodné povolania, Právnické osoby za obce SR, Živnostníci podľa vybraných ekonomických činností, Podniky podľa ekonomických činností, Podniky podľa počtu zamestnancov, Podniky podľa vybraných druhov vlastníctva.
- **Priemysel:** Štatistika priemyslu, Priemyselné závody - vybrané ukazovatele
- **Energetika:** Spotreba palív, elektriny, tepla podľa územia, typu ukazovateľa a roku.
- **Stavebníctvo:** Stavebníctvo - celé odvetvie, Stavebníctvo - podniky s 20 a viac zamestnancami, Tržby za vlastné výkony a tovar v stavebníctve.
- **Obchod, pohostinstvo a ubytovanie:** Tržby za vlastné výkony a tovar, Produkcia trhových tovarov a služieb, Tržby za vlastné výkony a tovar za ubytovacie a stravovacie služby, Tržby za vlastné výkony a tovar - hotely a reštaurácie, Tržby na obyvateľa, Tržby za vlastné výkony a tovar v maloobchode na 1 obyvateľa, Sieť pohostinstva veľkých podnikov s 20 a viac zamestnancami, Sieť maloobchodu veľkých podnikov s 20 a viac zamestnancami v SR.
- **Doprava:** Tržby za vlastné výkony a tovar v doprave (bez cestovných kancelárií) z mesačného štatistického zisťovania, Dĺžka ciest (v km), Miestne komunikácie, objekty na miestnych komunikáciách, Prehľad objektov na cestách I., II., III. triedy, Počet motorových vozidiel, Nehodovosť, Vybrané ukazovatele cestnej dopravy.

- **Poštové služby, telekomunikácie:** Vybrané ukazovatele za pošty a telekomunikácie.
- **Cestovný ruch:** Kapacita a výkony ubytovacích zariadení, Štatistika cestovného ruchu podľa okresov.
- **Informácie a komunikácia:** Počet a typ osobných počítačov v podnikoch s počtom zamest. 20 a viac, Počet zamestnancov pracujúcich s PC, Informačno-komunikačné technológie v domácnostiach.
- **Vybrané trhové služby:** Nehnuteľnosti, prenájom, obchodné činnosti a ostatné služby, Vybrané trhové služby.
- **Poľnohospodárstvo:** Poľnohospodárska produkcia, Tržby za predaj poľnohospodárskych výrobkov, Intenzita chovu hospodárskych zvierat, Stav hospodárskych zvierat, Produkcia vybraných poľnohospodárskych plodín, Hektárová úroda.
- **Vodné hospodárstvo:** Rozvod a čistenie vody, Verejné vodovody a kanalizácie, Dĺžka siete a odpadová voda.
- **Veda a technika:** Výdavky na výskum a vývoj, vedná oblasť, Vedecká činnosť, Zamestnanci výskumu a vývoja vo fyzických osobách.
- **Školstvo:** Materské školy, Základné školy, Gymnáziá, Stredné odborné školy, Stredné odborné učilištia, Združené stredné školy, Konzervatória a Vysoké školy.
- **Kultúra:** Galérie, Múzeá, Kultúrno-osvetové zariadenia, Divadlá podľa druhu vlastníctva, Knížnice - knižničné jednotky a výpožičky, Ostatné kultúrne zariadenia.

Kvalita dát

Proces a frekvencia zberu dát databázy regionálnej štatistiky sa riadi vyhláškou Štatistického úradu SR: "Program štátnych štatistických zisťovaní na trojročné obdobia", ktorú zostavuje úrad v súčinnosti s ministerstvami a štátnymi organizáciami. Okrem zisťovaní zahrnutých v programe, vykonáva databáza regionálnej štatistiky aj osobitné zbery u fyzických osôb (poľnohospodárske súpisy). Databáza regionálnej štatistiky sa aktívne rozvíja a postupne rozširuje. Je taktiež otvorená návrhom na dodatočné zbery dát. Aktualizácia údajov prebieha podľa harmonogramu, v čo najkratšom čase po oficiálnom zverejnení údajov.

Hodnota metadát získaných zo zdrojov databáz regionálnej štatistiky je veľmi cenná. Vďaka svojej komplexnosti, má veľký potenciál pri vytváraní spotrebných profilov užívateľov ako

jedného z výstupov nového bilančného modelu estimácie odchýlky výroby a spotreby elektrickej energie na Slovensku. Pri plánovaní modelu ďalšej generácie odporúčame bližšiu spoluprácu s úradom.

Náročnosť extrakcie a implementácie dát

Databáza regionálnej štatistiky je uložená v prostredí IIS servera a prezentovaná prostredníctvom nástroja PX-Web a pred pripravených dátových blokov vo forme PX súborov. PX-Web je web aplikácia určená na dynamickú prezentáciu štatistických údajov. Jej súčasťou je ďalšia web aplikácia PX-iMap určená na mapové prezentácie štatistiky. Web stránky sa vytvárajú dynamicky na základe výberu ukazovateľov používateľa.

Napriek tomu, že pre prístup k údajom nie je podmienený registráciou, údaje sprístupnené používateľom v rámci systému databáz regionálnej štatistiky sú duševným majetkom Štatistického úradu SR a sú chránené autorským zákonom. Ich rozširovanie tretím osobám je podmienené súhlasom Štatistického úradu SR.

Prvá vzorka dát

V poslednej časti správy si popíšeme pilotné dáta, ktoré predstavovali dátové jadro našej analytickej, modelovacej, optimalizačnej a predikčnej činnosti.

Pilotné dáta

Pilotná databáza popisuje 255 sústav. Z nich je 250 miestnych, 1 prenosová a 4 regionálne, distribučné. Z pilotných dát nevieme presne, o ktorú sústavu sa jedná. Dáta o odberných miestach máme len pre 11 sústav. V prvej vzorke máme dáta o 4581 OOM. Z nich je 4565 odberov a 16 dodavok. V dátach má 797 z nich deskripciu "vyrobna_id". Máme 68 meraní typu A, 7 meraní typu B, a 4506 meraní typu C. Čo sa týka meraní, pre 3 odberné miesta typu A nemáme merania (8901997, 8892074, 8902072). Pre ostatné máme dáta za 93 dní (od 1.1.2014 do 3.4.2014). Nepriebehové merania prirodzene variujú na mesačné, ročné a dátumové, s rozsahom od 1.1.2013 do 1.4.2014. Databáza poskytuje niekoľko TDO diagramov.

Čo by nám vylepšilo vzorku

O hodnote týchto a dát a ich metadát nie je pochyb. Vo vzťahu k ich objemu, 93 dní je relatívne krátky úsek, nedovoľujúci nám všetky vybrané typy analýz, poprípade predlžujúci

tvorbu modelu analýzy o náročnú extrapoláciu dát. Kvalite dátových zdrojov a možnostiam extrapolácie dátových radov budeme venovať samostatný report. Na tomto mieste by sme radi operačne informovali, čo k vzorke pridať:

- **Veľkosť vzorky:** všeobecne platí - čím viac dát budeme mať k dispozícii, tým lepšie. Súčasná štruktúra ukazuje, že na komplexnejšie analýzy by sme (ideálne) mali disponovať aspoň dvojročnými dátami (730 dní). Takáto vzorka by nám umožnila oveľa väčšie množstvo korelácií a dala by nám voľnosť pri spájaní dát o spotrebe a výrobe elektrickej energie so socio-demografickými, spoločensko-ekonomickými a hydrometeorologickými trendmi na Slovensku.
- **Stratifikácia vzorky:** Radi by sme disponovali dátami z väčšieho množstva odberných miest a to hlavne typu A (máme 68 meraní) a typu B (máme 7 meraní).
- **Chýbajúce dáta:**
 - V tabuľke "BD_SUSTAVY" máme len typ sústavy a bez informácie o ktorú sústavu sa jedná.
 - Nemáme príklady dát o externých sústavách a priamych vedeniach.
 - Chýbajú dáta pre triedy TDO 1224, 1227, 1178.
 - Nemáme merania pre 3 odberné miesta typu A (8901997, 8892074, 8902072) popísané v databáze,
 - Chýbajú nám dáta pre triedy TDO 1224, 1227, 1178.
- **vyrobna_id:** V tabuľke BD_OOM nevieme presne čo znamená premenná "vyrobna_id", nevieme si ju s ničím spojiť a nevieme na čo sa používa.

Záver

Naším cieľom v tejto časti správy bolo urobiť prehľad použiteľných close-data zdrojov (neverejné dáta), analyzovať ich legislatívne ukotvenie a ich kvalitu. Následne ohodnotiť možnosti ich využitia v systémoch, podporujúcich alebo nahradzujúcich dátové zdroje súčasného bilančného modelu estimácie odchýlky výroby a spotreby. Vymenované neverejné zdroje dát a predovšetkým ich kombinácie majú potenciál posunúť súčasné spracovanie dát v Slovenskej energetike na novú úroveň a zároveň reagovať na očakávané zmeny v kontexte dátovej štruktúry, ktorá bude analytickým subjektom k dispozícii v blízkej

budúcnosti (smartmetering). Nasledujúca časť správy sa bude venovať potenciálu hydrometeorologických dátových zdrojov a prehľadu použiteľných verejných dát, analýze ich kvality a možnosti ich využitia v modeloch predikcie a optimalizácie výroby a spotreby elektrickej energie na Slovensku.

Návrhy na vylepšenie zberu verejných a neverejných dátových vzoriek, ich kvalita, popis metód a nástrojov ich anonymizácie, extrapolácie a vizualizácie budú predmetom samostatných kapitol.

V poslednej časti, poukazujeme na potrebu doplnenia a rozšírenia vzorky dát.

2.1 Štruktúra hydrometeorologických dát používaná v energetike.

1. Všeobecne k produktom Slovenského hydrometeorologického ústavu

Energetici vo svete využívajú meteorologické produkty ako vstupné údaje do numerických modelov odhadu výroby a spotreby elektriny v najbližších hodinách a dňoch. V podstate im ide o pokles alebo nárast:

- spotreby elektriny - spôsobený ochladením (kúrenie), oteplením (klimatizácie),
- výroby elektriny - solárne, veterné a vodné elektrárne (dážď, topenie snehu).

Na základe odhadu si denne elektrinu rezervujú u dodávateľov, nakupujú alebo predávajú.

2. Odoberané údaje (v SR)

Distribučné spoločnosti na Slovensku odoberajú hydrometeorologické dáta nezáväzne na sebe, paradoxne, každá iné produkty SHMU. To, že všetci slovenskí klienti SHMU energetike neberú všetky produkty je spôsobené dvomi faktormi:

- menším počtom vstupov do ich modelov (jednotlivé distribučné spoločnosti majú svoje modely). Tieto modely dostávajú od zahraničných (aj materských) firiem a viaceré z nich sú už staršie.
- finančným limitom na nákup vstupných meteorologických údajov (= šetrenie prevádzkových nákladov).

Menší počet vstupov samozrejme vedie k „hrubším“ odhadom a väčšej neistote pri rozhodovaní.

3. Štruktúra hydrometeorologických dát podľa odberateľov.

Celkový zoznam premenných, ktoré sa využívajú v aktuálne existujúcich modelov výroby a spotreby elektriny v najbližších hodinách a dňoch:

- Teplota v 2m nad zemou
- Rýchlosť vetra v 10m nad zemou
- Globálne žiarenie (W/m²)
- Pokrytie oblačnosťou (v osminách)

Z dátových radov týchto premenných sa počítajú priemery v rôznej frekvencii:

- 4x denne (krok = 1 h)
- 2x denne (krok = 6 h)
- 1x denne (krok = 1 h)
- 1x týždenne (krok = 1 deň)

Časové rozsahy v ktorých si distribučné spoločnosti dáta od SHMU žiadajú:

- Namerané hodnoty a hodinová predpoveď D (dnes)
- Predpoveď D+1 (+ 24 hod)
- Predpoveď D+2 (+ 48 hod)
- Namerané hodnoty D-1 (včera)

Približne dve tretiny distribučných spoločností využívajú namerané hodnoty (D-1) v premenných teplota v 2m nad zemou, globálne žiarenie (W/m²) a pokrytie oblačnosťou (v osminách). Tieto premenné sa zbierajú na 4+3 staniciach (dosť málo) a počítajú sa z nich 2 priemery (veľmi málo). Tieto spoločnosti nakupujú predpoveď zo systému ALADIN na predikciu (D+1,+2,+3) na 72 hodín (pracuje s teplotou, globálnym žiarením a oblačnosťou). Hodnoty sa upravujú 4x denne. Ďalej nakupujú predpoveď z ECMWF na 4.-10. deň (pracuje s teplotou, globálnym žiarením a oblačnosťou). Hodnoty sa upravujú 2x denne. Posledný typ

predpovede ktoré distribučné spoločnosti nakupujú je predpoveď na 10.- 30. deň (pracuje s teplotou, globálnym žiarením a oblačnosťou). Hodnoty sa upravujú 1x týždenne.

Približne jedna tretina distribučných spoločností využíva okrem nameraných hodnôt D-1 aj hodnoty z aktuálneho dňa (D). Okrem teploty v 2m nad zemou, globálneho žiarenia a pokrytia oblačnosťou si berú aj rýchlosť vetra v desiatich metroch nad zemou. Tieto dáta prichádzajú v podobe štyroch priemerov a merané sú na 23. staniciach, so špecifikáciou 2-3 staníc na západnom a východnom Slovensku, z ktorých si vyžadujú hodinové, maximálne minútové dáta. Slovenská elektrizačná prenosová sústava, a. s. (SEPS) má vlastný zdroj hydrometeorologických dát. Ich sieť staníc je asymetrická, nastavená podľa relevantných spotrebiteľov a výrobcov.

Zbierajú sa merania:

- Vzduch teplota ©, vlhkosť (%), tlak (hPa),
- Vietor: – rýchlosť (m/s), smer (degree), 10 min nárazy (m/s),
- Vnútoraná teplota na čidle (c)
- Globálne žiarenie (W/m²)
- Zrážky: 10 min (mm), 60 min (mm), 24 hod (mm)
- Trvanie osvitu: 10min (s), 60min (s), 24hod (s)

4. Navrhované (potrebné) údaje (v SR)

Predikčný systém môže pracovať efektívnejšie, v prípade, ak bude používať kompletný balík údajov a predpovedí, ktoré SHMU zbiera a vyhodnocuje.

Ako príklad sa dá uviesť skutočnosť, že zahraniční zákazníci SHMU programujú komplexnejšie modely a uzatvárajú zmluvy so SHMU aj na také ich údajové produkty, ako sú odber údajov o úhrnoch zrážok a výške snehovej pokrývky. Pomocou týchto produktov je potom možné optimalizovať nákupy, výrobu, predaj a reguláciu elektriny v krátkodobej (1 - 3 dni) aj v strednodobej (1 -10 dní) perspektíve.

Výber staníc SHMU, z ktorých distribučné spoločnosti zbierajú dáta je neoptimálny, ich počet je v niektorých prípadoch nedostačujúci.

Naproti tomu -existujúca meteorologická sieť SEPS-u je potenciálne veľmi dobre využiteľný zdroj dát, najmä kvôli relevantnosti rozloženia meracích staníc.

2.2 Prehľad použiteľných zdrojov dát. Verejné dáta.

Predmetom tohto pokračovania druhej analýzy v rámci projektu bolo jeho doplnenie v oblasti dátového usmernenia. Táto správa je len formálnym výstupom analýzy vhodnosti rôznych verejných zdrojov dát, vybraných na základe ich prítomnosti v dátových štruktúrach systému súčasného sledovania pohybu výroby a spotreby elektrickej energie na Slovensku. Cieľom analýzy bolo nájsť a popísať ďalšie verejné zdroje dát vhodné k tvorbe inteligentných systémov, podporujúcich alebo nahradzujúcich dátové zdroje súčasného bilančného modelu estimácie odchýlky výroby a spotreby.

Pri tom sme prihliadali:

- na hodnotu metadát získaných z potenciálnych dátových zdrojov,
- na relevantnú legislatívu, najmä: 1.) Zákon č. 211/2000 Z. z. o slobodnom prístupe k informáciám a o zmene a doplnení niektorých zákonov, 2.) Zákon č. 275/2006 Z. z. o informačných systémoch verejnej správy, 3.) Štandardy pre informačné systémy úradov verejnej správy, 4.) Zákon č. 540/2001 Z. z. o štátnej štatistike a 5. ustanovenia na úrovni Európskej únie a medzinárodného práva,
- na bezpečnosť jednotlivých verejných zdrojov dát z pohľadu anonymity dát,
- na kvalitu potenciálnych dátových zdrojov,
- na náročnosť extrakcie a implementácie dát z verejných zdrojov do aktuálne bežiacich systémov,
- na relevanciu potenciálnych dátových zdrojov pri cieľoch krátkodobej predikcie (hodinový, denný interval) a cieľoch dlhodobej predikcie (týždenný a mesačný interval) spotreby a výroby elektrickej energie,
- na charakteristiky súčasného systému poskytovania, evidencie a využitia údajov o výrobe a spotrebe elektrickej energie.

Verejné dáta predstavujú zaujímavý zdroj poznania pre projekty oblasti výskumu a vývoja. Množstvo verejných zdrojov aj kvalita a objem dát v nich v posledných rokoch zaznamenávajú prudký nárast. Viaceré štáty, na čele s Veľkou Britániou uvoľňujú štátne dáta a to nielen v agregovanej, ale aj surovej podobe. Ich motivácia na takúto aktivitu je dvojaká:

- a) občianske združenia a spolky sa dovoľávajú svojich práv na prístup k informáciám, ktoré boli zbierané za ich dane a pokiaľ sa netýkajú bezpečnosti štátu, mali by byť zverejnené,
- b) zverejnenie dát umožňuje inováciu cez crowdsourcing mechanizmus.

Na nasledujúcich stranách popíšeme vybrané zdroje týchto dát vo svete, ktoré sa po našej analýze ukázali ako kvalitné a relevantné pre našu výskumnú činnosť.

Hydrometeorologické dáta

Francúzsko

Zdroj: <http://www.data.gouv.fr/en/dataset/donnees-d-observation-des-principales-stations-meteorologiques>

Popis zdroja: Na stránke sú k dispozícii hydrometeorologické dáta z územia Francúzska (nie len kontinentálneho), rozmiestnených v najväčších metropolách a geograficky relevantných bodoch.

Zbierané dáta: Zbierajú sa atmosférické parametre ako teplota, vlhkosť, smer a sila vetra, atmosférický tlak a hladina zrážok. Merané sú aj deskriptívne parametre ako časová senzitivita, vizibilita a meranie oblačnosti. Poslednou skupinou dát sú dáta špecifické k polohe stanice, napríklad hĺbka snehu a povrchový stav zeminy.

Kvalita dát: Na 63 staniciach (budú pribúdať) sa základné dáta zapisujú raz za 4 hodiny. Na serveri je v každom momente ponechaná dvojtýždňová história dát.

Náročnosť extrakcie a implementácie dát: Dáta sú voľne šíriteľné a stiahnuteľné v štandardnom formáte, ktorý umožňuje ich ľahkú manipuláciu. Dáta bude nutné zbierať pravidelne (raz za dva týždne) aby sa vytvorila relevantná vzorka. Túto aktivitu je možné zautomatizovať.

Hodnota získaných metadát: Prostredníctvom tohto zdroja máme k dispozícii dáta použiteľné na tréning algoritmov zapracovávajúcich hydrometeorologické trendy do celkovej predikcie.

USA

Energetické dáta

Veľká Británia

Zdroj: <http://www2.nationalgrid.com/uk/Industry-information/Electricity-transmission-operational-data/Data-Explorer/>

Popis zdroja: Prehľadové dáta o celkovom stave prenosovej sústavy v UK ponúkajú komplexný pohľad na sieť.

Zbierané dáta: Dátové rady o celkovom stave siete, prenosových aktivitách a výrobe solárnej a veternej energie.

Kvalita dát: Kvalitatívne primeraná frekvencia zapisovania dát. Databáza nám dáva k dispozícii dáta s 9 ročnou históriou (od 2005). Okrem toho publikuje aj predpoveď stavu siete na týždeň dopredu, fázovanú po dňoch.

Náročnosť extrakcie a implementácie dát: Dáta sú voľne šíriteľné a stiahnuteľné v štandardnom formáte, ktorý umožňuje ich ľahkú manipuláciu. Zber bude možné zautomatizovať.

Hodnota získaných metadát: Prostredníctvom tohto zdroja máme k dispozícii dáta ukazujúce vývoj a zaťaženie prenosovej sústavy na systémovej úrovni. Dáta sú preto použiteľné na tréning algoritmov v sledovaní trendov vo výrobe elektrickej energie. Chybovosť predpovedí, ktoré sa z týchto dát generujú na týždennej báze, nám môže slúžiť ako referencia výkonu nášho systému.

Zdroj: http://data.gov.uk/dataset/energy_consumption_in_the_uk

Popis zdroja: Zdroj predstavuje databázu údajov o spotrebe elektrickej energie a iných energií v UK.

Zbierané dáta: Štatistiky o spotrebe elektrickej energie sú tu rozdelené do piatich sekcií: celková spotreba energií, spotreba v doprave, domácnostiach, priemysle a v sektore služieb. Štatistiky obsahujú klasické premenné umiestnené v čase.

Kvalita dát: Zdroj drží dáta od roku 2007 do roku 2013. Ide o ročné súhrny spotrieb elektrickej energie podľa odvetví.

Náročnosť extrakcie a implementácie dát: Dáta sú voľne šíriteľné a stiahnuteľné v štandardnom formáte, ktorý umožňuje ich ľahkú manipuláciu. Zber bude možné zautomatizovať.

Hodnota získaných metadát: Tieto dáta nám umožnia pozrieť sa na vzťahy medzi spotrebou elektrickej energie a ostatných energií v rôznych ekonomických kontextoch a rôznom čase.

Zdroj: <http://data.gov.uk/dataset/energy-generation-from-solar-pv-arrays-for-selected-bristol-buildings>

Popis zdroja: Solárny smartmetering dataset z UK.

Zbierané dáta: Dáta sú z bežiaceho projektu Energy UK z mesta Bristol. Zbierajú sa dáta o vygenerovanej elektrickej energii skupinami solárnych panelov (PV arrays), umiestnených na vybraných budovách v meste. Projekt má za cieľ testovať efektivitu zberu slnečnej energie v mestskej infraštruktúre.

Kvalita dát: Dáta sa zbierajú na 32 staniciach na dennej báze, hodnoty sú merané smartmetrami v krátkych intervaloch, generujú sa aj týždenné agregované dátové agregáty. História zberu závisí od stanice, najstaršie sú z roku 2012.

Náročnosť extrakcie a implementácie dát: Dáta sú voľne šíriteľné a stiahnuteľné v štandardnom formáte, ktorý umožňuje ich ľahkú manipuláciu. Zber bude možné zautomatizovať.

Hodnota získaných metadát: vzorka je zaujímavá, nakoľko je zbieraná na malom území (je tam vidno minimálne vzťahy) a smartmetreringové dáta sú postupne agregované do použiteľnej dátovej štruktúry. Môžeme z nich odsledovať úroveň, do akej sa oplatí predprocesovať surové dáta zo smartmetrov ak by sme ich chceli použiť na analýzy rôznych úrovní, či ľahko interpretovateľnú vizualizáciu.

Nemecko

Zdroj:

<http://www.transparency.eex.com/en/Statutory%20Publication%20Requirements%20of%20the%20Transmission%20System%20Operators/Power%20generation/Actual%20wind%20power%20generation>

Popis zdroja: Stránka predstavuje zdroj dát o veterných a solárnych elektrárnach, primárne na území Nemecka.

Zbierané dáta: Zo zdroja sa vieme dozvedieť, koľko veternej a solárnej energie kto v Nemecku vyrobil. Sú tam aj dáta o výrobe solárnej a veternej energie na Slovensku, no nie je ich veľa.

Kvalita dát: Dáta zo staníc sa zbierajú a zapisujú v reálnom čase, výstupom sú denné agregované dáta o tom, koľko vyrobili jednotlivé bilančné skupiny, napríklad 50Hertz (DE), Amprion (DE), APG (AT), TenneT (DE), TransnetBW (DE). Historické dáta tu sú minimálne od začiatku roku 2010.

Náročnosť extrakcie a implementácie dát: Samotné dáta sú len ako ukážka z flashovej aplikácie. Predpokladáme, že so sťahovaním dát budú problémy, aj keď zdroj sa javí, že je verejným dátovým zdrojom.

Hodnota získaných metadát: Nemecká výroba solárnej a veternej energie má sprostredkovane relatívne veľký vplyv na výrobu na predikciu výroby elektrickej energie na Slovensku. Zdroj by bolo dobré korelovať s nemeckými hydrometeorologickými dátami.

USA

Zdroj: <http://www.nrel.gov/midc/>

Popis zdroja: Vybrané dáta z „Clean energy“ produkcie na území Spojených štátov amerických.

Zbierané dáta: Na stránke sú k dispozícii dáta o zelenej produkcii, vybrané relevantne k USA hydrometeorologickým dátam z databázy popísaným v prvej časti reportu.

Kvalita dát: Dáta sú obmedzené na producentov, s geografickou blízkosťou ku staniciam, ktoré zbierajú hydrometeorologické dáta.

Náročnosť extrakcie a implementácie dát: Dáta sú voľne šíriteľné a stiahnuteľné v štandardnom formáte, ktorý umožňuje ich ľahkú manipuláciu. Zber bude možné zautomatizovať.

Hodnota získaných metadát: Zdroj umožňuje skúmať vzťahy medzi hydrometeorologickými trendami a výrobou.

Uvedené dáta je možné doplniť o databázu produkcie veterných elektrární v USA, ktoré zapisujú dáta v desať minútových intervaloch (variácia podľa jednotlivých premenných) zbieraných od roku 2004:

http://www.nrel.gov/electricity/transmission/western_wind_methodology.html

Nápomocné budú aj dáta zbierané z fotovoltaiických článkov v USA:

<http://wiki.dbpedia.org/Downloads39>

<https://openpv.nrel.gov/search>

Zaujímavá je aj databáza veterných turbín, popisujúca statické dáta o lokáciách a vlastnostiach (napríklad výkon) veterných jednotiek:

http://www.thewindpower.net/services_en.php#windfarms

Dáta o infraštruktúre a demografii

Zdroj: <http://data.worldbank.org/country>

<http://data.worldbank.org/country/slovakrepublic>

Popis zdroja: Dáta národnej banky o hospodárskych ukazovateľoch krajín (aj SR).

Zbierané dáta: Sú tam všetky relevantné dáta k analýze ekonomického rastu krajiny (aj SR).

Kvalita dát: Obsahuje 1300 ukazovateľov v číselných dátach za jednotlivé roky ku konkrétnej krajine, obsahujú aj dáta o elektrickej energii.

Náročnosť extrakcie a implementácie dát: Dáta sú voľne šíriteľné a stiahnuteľné v štandardnom formáte, ktorý umožňuje ich ľahkú manipuláciu. Zber bude možné zautomatizovať.

Hodnota získaných metadát: Cez tieto dáta vidíme trendy v spotrebe a smerovaní krajiny (aj SR). Je možné testovať vzťahy medzi ukazovateľmi ako export (mesačný, ročný) a spotrebou elektrickej energie. Vzorku možno doplniť o dáta o exporte produktov v čase:

http://www.thewindpower.net/country_zones_en_33_slovakia.php

Verejné zdroje demografických dát a dát o infraštruktúre:

<http://www.pmt.sk/Login.aspx?ReturnUrl=%2fstatistika.aspx>

<http://datahub.io/dataset/slovakian-company-registry>

Záver

Naším cieľom v tejto kapitole bolo urobiť prehľad použiteľných open-data zdrojov (verejné dáta), analyzovať ich legislatívne ukotvenie a ich kvalitu. Následne ohodnotiť možnosti ich využitia v systémoch, podporujúcich alebo nahradzujúcich dátové zdroje súčasného bilančného modelu estimácie odchýlky výroby a spotreby. Vymenované príklady verejných zdrojov dát a predovšetkým ich kombinácie majú potenciál posunúť vývoj algoritmov v slovenskej energetike na novú úroveň a zároveň reagovať na očakávané zmeny v kontexte dátovej štruktúry, ktorá bude analytickým subjektom k dispozícii v blízkej budúcnosti (smartmetering). Verejné dáta nie sú ešte dostatočne stabilné a bezpečné na to, aby mohli byť stálym zdrojom pre bežiaci finálny bilančný systém, no majú potenciál doplniť medzery vo vzorkách, použitých pri výskume a vývoji (tréningu) novej generácie algoritmov.

3 Základné aspekty správy kvality dát

Aby bolo možné navrhnúť efektívne metódy na čistenie dát, je nutné si ujasniť, aké chyby a nezrovnalosti v dátach je možné očakávať a nájsť, a aké sú ich zdroje. Potom je možné navrhnúť fázy v „živote“ dát, v ktorých je vhodné zakročiť a spôsoby, akými vykonať príslušné akcie.

3.1 Zdroje chýb v dátach

Skôr než údaje skončia v databázach, prechádzajú obyčajne niekoľkými fázami, ktoré v sebe obsahujú interakciu s ľuďmi a výpočtové spracovanie. Chyby v údajoch sa môžu zakomponovať v každom kroku tohto procesu počínajúc zberom údajov a končiac ich archivovaním [4]. Pochopenie týchto procesov vzniku chýb môže napomôcť pri návrhu vhodných metód zberu a liečenia dát ako aj pri vývoji vhodných metód na čistenie dát, ktoré dokážu odhaliť a následne odstrániť chyby. Zdroje chýb v databázach možno väčšinou rozdeliť do nasledovných kategórií:

- chyby v položkách údajov – keďže položky dát sú zadávané ľuďmi do príslušných formulárov, pri prevode z rečou podaných informácií alebo prepisovaním z iných písomných zdrojov, zadávatelia často vnášajú typografické chyby do dát alebo nesprávne pochopené informácie od zdroja. Patria tu aj prípady, keď niektoré položky údajov, ktoré majú byť zadané, nie sú poskytnuté zdrojom a je zvolená implicitná hodnota, ktorá je síce správna, ale nevystihuje pravú skutočnosť.
- chyby meraní – údaje sa získavajú na základe správania sa neakého fyzikálneho procesu (pohyb vozidla, stav meradla spotreby, množstvo populácie, a pod.). Často sú hodnoty charakterizujúce proces snímané ľuďmi, ktorí nesprávnymi metódami alebo nástrojmi odčítajú nesprávne hodnoty. Hoci údaje môžu byť odčítavané priamo senzormi, stále je možnosť zlého nastavenia (kalibráciou, interferenciou) senzorov alebo napríklad ich samotným návrhom a výrobou.
- chyby spôsobené predspracovaním dát - v niektorých prípadoch sú surové údaje predspracované a zosumarizované, kým sa vložia do databázy (napr. odstránenie šumu a zložitosti údajov, vykonávanie špecifických štatistických analýz, zdôraznenie spoločných vlastností údajov, alebo len zmenšenie objemu dát).
- chyby spôsobené integráciou dát – objemné a dlhodobo budované databázy často obsahujú údaje z rôznych zdrojov zbierané postupne rôznymi metódami. Často veľké databázy vznikajú zlúčením viacerých menších databáz alebo pričlenením inej databázy do hlavnej databázy. Toto zlučovanie je často spojené s riešením vzájomných nekompatibilit ako rôzne jednotky, reprezentácie dát, doby meraní a pod. Akákoľvek metóda, ktorá integruje údaje z rôznych zdrojov, môže vnášať chyby do dát.

3.2 Spôsoby zlepšenia kvality dát

Život údajov je proces, ktorý obsahuje zber/získavanie údajov, ich transformáciu, uchovanie, kontrolu, čistenie a analýzu (predikciu) [4]. Obyčajne tento proces zahŕňa ľudí a zariadenia viacerých organizácií a má dlhodobý a priestorovo náročný charakter. Každý krok tohto procesu môže byť navrhnutý spôsobom, ktorý berie ohľad na výslednú kvalitu dát. Prístupy, ktoré možno použiť, sú:

- inteligentné rozhranie zberu dát – inteligentné používateľské rozhranie môže zabezpečiť dodržanie základných pravidiel integrity dát dané databázou už pri ich zadávaní obsluhou (napr. kontrola typu dát, intervalu dovolených hodnôt, výberu z dovolených položiek). Môže ale súčasne spôsobiť, že zadávatelia dát často vkladajú nesprávne hodnoty. Štúdie ukázali, že aplikácia pravidiel integrity nezabráni zlým údajom (napr. obmedzenie, aby dané pole neostalo prázdne, neznamená, že tam zadávateľ vloží zmysluplný údaj, ak ho napríklad nemá k dispozícii) a vynucovanie týchto pravidiel vedie k určitej frustrácii zadávateľov (ktorí vykonávajú neustále sa opakujúcu a nudnú prácu, pri ktorej strácajú motiváciu udržiavať potrebnú kvalitu vkladných dát alebo prípadne nemajú dostatočné znalosti, aby prípadné neúplne údaje vedeli sami opraviť). Pri návrhu takéhoto inteligentného rozhrania by sa malo vziať do úvahy niekoľko pravidiel. Úroveň kvality dát by sa mala udržiavať nie vynucovaním ale prostredníctvom spätnej väzby, využitím intuitívnych vysvetlení a rád uľahčujúcich zadávateľovi dodatočnú prácu pri zadávaní údajov.
- organizovaný manažment kvality dát – zakladá na využití technologických riešení ale aj organizačných štruktúr a stimulov na modernizovanie celého procesu (t.j. získavanie, archivovanie a analýza dát) automatizáciou zberu dát, získavaním metadát na zlepšenie interpretácie dát a motiváciou zúčastnených udržiavať kvalitu získaných dát.
- automatizovaná kontrola a čistenie dát – existuje množstvo výpočtových metód, ktoré sa venujú identifikovaniu chýb v dátach a následným pokusom o ich opravu. Tieto metódy budú rozkategorizované a potom detailnejšie charakterizované v nasledujúcich častiach správy.
- objasňujúca analýza a čistenie dát – v mnohých prípadoch je možné opraviť chyby v dátach iba s pomocou ľudského zásahu. Obyčajne existuje vzájomná súčinnosť medzi nástrojmi na čistenie dát a systémami pre vizualizáciu dát. Účasť človeka sa vyžaduje pri pochopení vlastností danej vzorky dát (data profiling) [5] a následnom odhalení a prípadnej oprave chýb. Pri vizualizácii dát analytici často sledujú štatistické vlastnosti dát.

3.3 Typy a metódy na čistenie dát

Ako už bolo uvedené vyššie, veľké objemy dát uložené v databázach spoločností obsahujú tzv. špinavé (dirty) dáta, ktoré znižujú celkovú kvalitu dát. Patria sem nekompletné

a nesprávne údaje a duplikáty. Preto v rámci výskumov boli hľadané metódy, ktoré by dokázali identifikovať tieto dáta (chyby) a prípadne sa pokúsili ich aj opraviť [6]. Vo všeobecnosti možno vyčleniť nasledovné kategórie metód na čistenie dát:

1. štatistické (kvantitatívne) metódy – ako názov napovedá, tieto metódy sú založené na využití štatistických vlastností dát a ich rozdelení. Základom metód je nájsť tzv. outsajderov (outliers) – budeme ich volať odľahlé hodnoty, ktoré sa vymykajú skutočnosti alebo dovoleným intervalom.
2. logické metódy – pri týchto metódach čistenia dát je základom nájsť logický model dát a potom sa hľadajú údaje, ktoré porušujú tento model. Model vychádza zo stanovených pravidiel a obmedzení, na základe ktorých sa potom realizuje detekcia týchto údajov (chýb).
3. hybridné metódy – kombináciou štatistických, logických a iných metód možno zvýšiť účinnosť čistenia dát (hľadania a vymazania alebo opravy chýb v dátach).

3.4 Typy dát v databázach

Metódy na čistenie dát úzko súvisia aj s typom dát uložených v databázach. Jedna bežná databáza dát (tabuľka) môže obsahovať kombináciu viacerých typov dát. Preto sa aj na čistenie dát v databáze pre rôzne typy dát využívajú rôzne metódy. Je možné sa stretnúť najčastejšie s nasledovnými typmi dát [7]:

- kvantitatívne údaje – celé čísla alebo reálne (s pohyblivou radovou čiarkou) čísla, ktoré udávajú množstvo zvolenej veličiny. Môže to byť súbor jednoduchých čísiel alebo viacrozmerných (komplexných) čísiel často v úzkom vzťahu s časom. Sú často spojené s jednotkou, ktorá by mala byť jednotná pre všetky dáta.
- kategorizované údaje - sú to mená (slová) alebo kódy, ktoré sa používajú na pridelenie dát do kategórií alebo skupín. Nemajú prirodzený spôsob usporiadania ani vzdialenosť medzi dvojicami. Nemusia podporovať jednotný priestor mien (rovnaká položka môže mať v rôznych súboroch odlišný názov). Ťažko sa identifikujú údaje so zlým priradením ku kategóriám alebo typografické chyby (potrebné slovníky).
- poštové adresy – patria medzi kategorizované údaje, ale majú podstatný význam, aby sa s nimi zaobchádzalo špeciálne. Obyčajne je to voľný text, ktorý má svoju štruktúru a vnútornú nadbytočnosť. Problémy môže spôsobovať napr. prítomnosť či neprítomnosť slova „ulica“, s čím súvisí aj tzv. deduplikácia, t.j. identifikovanie duplicitných adries v zozname.
- identifikátory – špeciálny prípad kategorizovaných údajov, ktoré jedinečným spôsobom pomenúvajú objekty alebo vlastnosti. Vo väčšine prípadov sa vyžaduje, aby dodržiavali dohodnutú štruktúru, ktorá podáva určitú informáciu (telefónne čísla, kódy produktov, IP

adresy a pod.). Identifikátor v rámci jednej množiny dát by mal identifikovať jeden objekt (vyskytnúť sa iba raz), ale v rámci inej množiny sa už môže vyskytovať v mnohých položkách. Táto vlastnosť definuje referenčnú integritu dát a jej chyby sa označujú aj ako nekonzistentnosti integrity dát. Pri ich hľadaní sa využívajú funkčné závislosti.

Z pohľadu dát používaných v energetike možno vyčleniť dve skupiny. Na jednej strane spoločnosti musia udržiavať údaje o svojich zákazníkoch v štruktúrovanej podobe. Na druhej strane sú to dáta, ktoré reprezentujú informácie o odberno-odovzdávacích miestach a o meraniach z týchto miest.

4. Metodologický prehľad štatistických nástrojov použiteľných pri tvorbe a testovaní modelov predikcie a optimalizácie výroby a spotreby elektrickej energie na Slovensku.

Predmetom prvej analýzy v rámci projektu bolo jeho metodologické usmernenie. Táto správa je formálnym výstupom analýzy vhodnosti rôznych prognostických metód a štatistických nástrojov na predikovanie vývoja viacerých časových radov, vybraných na základe ich prítomnosti v dátových štruktúrach systému súčasného sledovania pohybu výroby a spotreby elektrickej energie na Slovensku. Cieľom analýzy bolo nájsť a popísať metodologicky vhodné prístupy k tvorbe inteligentných systémov, podporujúcich alebo nahradzujúcich analytickú stranu súčasného bilančného modelu estimácie odchýlky výroby a spotreby.

Pri tomto zadaní sme prihliadali:

- na charakteristiky súčasného systému poskytovania, evidencie a využitia údajov o výrobe a spotrebe elektrickej energie,
- na hodnotu meta-dát získaných analýzy,
- na bezpečnosť jednotlivých prístupov z pohľadu anonymity dát,
- na možnosť interpretácie výsledkov
- na jednoduchosť implementácie metód do aktuálne bežiacich systémov.
- na efektivitu modelov pri cieľoch krátkodobej predikcie (hodinový, denný interval) a cieľoch dlhodobej predikcie (týždenný mesačný interval) spotreby a výroby elektrickej energie.

Pri analýze potenciálu metodologických prístupov boli brané do úvahy dáta o evidencii sústav, výrobcov elektriny a odberných a odovzdávacích miest pripojených do sústav (v ráťane TDO hodnôt a nameraných údajov odberu a dodávky elektriny). Predpokladali sme tiež prístup k ďalším dátam o odberných a odovzdávacích miestach (PSČ, Druh, Napäťová úroveň, Typ merania, Trieda TDO, Predpokladané ročné množstvo, Zaradenie do sústavy), o ich odbere a dodávke elektriny zo a do sústavy.

Predspracovanie dát

Dáta, získané zo systému v surovej podobe, je pred analýzou potrebné skontrolovať a upraviť do formátu, ktorý uľahčí ich manipuláciu a interpretáciu. Pre väčšinu testovaných metód predikcie je najvhodnejším vstupným (učiacim) profilom taký profil, ktorý má dlhodobý charakter stabilného toku dát v štandardne sa vyskytujúcich hodnotách, bez signifikantných posunov. Trénovacie profily s nepravidelným (resp. neprirodzeným) charakterom hodnôt a skokovými zmenami spomaľujú tréning systému a v horšom prípade, znemožňujú štatisticky robustnú predikciu.

Tvorba algoritmickej modelov sa nezaobíde bez primárnej kontroly kvality získaných dát. Je dôležité, aby dáta neobsahovali evidentné chyby (e.g. teplota $+100^{\circ}\text{C}$, podozrivo vysoká spotreba). Šetrenie takýchto chýb môže prebiehať dvoma spôsobmi. Ak stavíme model na citlivých dátach, v ktorých je minimálna chybovosť esenciálna alebo na malých vzorkách dát, snažíme sa takéto infekty vyhľadať a po jednom ošetriť. V prípade, že model stavíme na dostatočne veľkých vzorkách dát, môžeme takéto extrémne hodnoty ošetriť nastavením hodnotovej hranice, ktorú získame pomocou indikátorov deskriptívnej štatistiky - priemeru, mediánu, maxima minima a štandardnej odchýlky. Neštandardné zmeny tvaru dát, ktoré je potrebné odfiltrovať, je možné detekovať prostredníctvom validačnej funkcie.

Pre potreby kvalitnej analýzy je taktiež dôležité, aby väčšina vstupov mala podobný rozsah (i.e. nie je dobré, keď teploty sú z rozsahu desiatok a spotreby z rozsahu tisícok, desať tisícok). Štandardný postup je každý vstup preškálovať do rozsahu 0 až 1, prípadne tak, aby mal priemer 0 a štandardnú odchýlku 1. Medzi ďalšie štandardné úpravy dát patrí definovanie rozdielu vo významoch nulovej hodnoty a prázdnej hodnoty, kde prázdna hodnota zväčša označuje nevalidnú hodnotu (e.g. výpadok merania) a nulová hodnota označuje meranie hodnoty 0.0, aj napriek tomu, že priemer viacerých meraní v hodnote 0.0 nie je frekventovaný.

Kritickým predpokladom validity a reliability záverov je dostatočná veľkosť vzorky (počet dátových profilov, dĺžka ich trvania a frekvencia zberu bodových meraní). V niektorých prípadoch je možné získané dáta matematicky extrapolovať, pokiaľ je vzorka nedostatočná na získanie štatisticky signifikantných prípadov. Takto rozšírená dátová vzorka však musí spĺňať veľmi striktné požiadavky robustnosti, internej konzistentnosti a dynamiky v čase.

Tréning takýchto systémov môže byť realizovaný off-line, na tzv. mŕtvej vzorke dát alebo on-line, na tréningových dátach z bežiaceho dátového toku. Napriek tomu, že pri off-line tréningu je šanca, že takto trénovaný systém nezoberie do úvahy všetky signály

manipulujúce predikciu v reálnom dátovom toku (lebo off-line vzorka tréningových profilov ich neobsahovala), prístup dovoľuje intenzívnejšie testovanie modelu a prípadné chyby nemajú priamy a dopad na aktuálne bežiace systémy. On-line tréning je intenzívnejší, predstavuje istotu započítania všetkých existujúcich signálov v aktuálne bežiacich dátových profiloch do tvorby modelu. Hrozbou je, dopad prvotnej chybovosti modelu na bežiaci systém a faktu, že množstvo irelevantných informácií v dátovom toku a profily s dlhším časom vyhodnotenia by značne spomaľovali výpočet predikcie. V kontexte energetiky je akceptovateľným kompromisom prístup dvoj krokovej analýzy, pri ktorej sa na tréning predikcie on-line dátových profilov používajú ich uložené verzie a bodové merania, ktoré sa v potrebnom časovom horizonte (napr. denne) pred výpočtom predikcií aktualizujú podľa hodnôt počítaného dátového profilu. V nasledujúcej časti správy si popíšeme vybrané predikčné modely, ktoré budeme v rámci projektu estimácie výroby a spotreby elektrickej energie testovať.

Validácia modelu

Pri tréningu modelu netrénujeme tak, aby sa dobre správal na tréningových dátach, pretože nemáme záruku, že sa bude dobre správať na nových dátach. Pri niektorých modeloch je častý jav tzv. pretrénovanie, t.j. stav kedy sa model veľmi dobre správa na tréningových dátach, ale má vysokú chybovosť na nových dátach. Na odhalenie tohto fenoménu a odhadnutie skutočnej chybovosti modelu na nových dátach sa používajú obvykle nasledujúce postupy:

- z tréningových dát vyčleníme tzv. testovaciu vzorku na ktorej model netrénujeme, ale po tréningu na nej overíme chybovosť modelu
- K-fold cross-validácia – tréningovú vzorku rozdelíme na K častí, v každom kroku jednu z častí vyčleníme z tréningovej vzorky a použijeme ako testovaciu, následne natrénujeme model a overíme jeho výkonnosť; výsledný odhad výkonnosti modelu je priemer z K overení.

Metódy strojového učenia a ich efektivita v kontexte modelov predikcie výroby a spotreby elektrickej energie.

Pri strojovom učení vytvárame model, ktorý sa snaží predikovať cieľovú premennú na základe iných vstupných premenných (napríklad spotrebu energie pomocou teploty vzduchu, ročného obdobia, toho, či je víkend, atď.). Základom tohto analytického prístupu je databáza takzvaných vstupných a výstupných profilov, definovaných bodovými meraniami (stavmi) v časovom slede ohraničujúcim analyzovaný jav. Bodové merania v týchto profiloch sú

organizované do jednotlivých premenných alebo faktorov (skupín premenných), ktoré pre systém fungujú ako trénovacie dáta. Pomocou trénovacích dát automatizovane a dynamicky nastavujeme parametre modelu tak, aby zodpovedali týmto profilom. Potom už môžeme pomocou získaného modelu vykonávať predikciu.

Lineárna regresia

Pri modeloch založených na lineárnej regresii sa spotreba predikuje ako lineárna kombinácia viacerých vstupov (predchádzajúca spotreba, teplota, binárne indikátory o dni v týždni, voľných dňoch, známe časové rady). Pomocou týchto vstupných dát sú pri tejto metóde identifikované modely neznámych parametrov pomocou lineárnych funkcií. V kontexte lineárnej regresie sa tieto parametre nazývajú vstupné atribúty (features) a snažia sa predikovať sériu cieľových premenných (target variables) v priestore všetkých možných funkcií (hypotheses). Cieľom modelu je popísať takú hypotézu, ktorá najlepšie predikuje pohyb cieľovej premennej.

Efektivitu vybranej hypotézy môžeme merať pomocou jednoduchých chybových funkcií, napríklad funkciu ktorá sleduje štvorec rozdielu reálnej hodnoty z trénovacích dát a predikovanej hodnoty. Cieľom je teda o čo najnižšiu sumu funkčných hodnôt chybovej funkcie na vzorke všetkých vstupných atribútov. Druhou alternatívou sú chybové funkcie na základe lineárneho programovania. Tieto sú síce komutačne náročnejšie, no majú výhodu, ak distribúcia dát nie je normálna.

Výhodou lineárnej regresie je relatívne nízka časová náročnosť trénovania a možnosť distribuovaného trénovania v prípade veľkého množstva dát. Dost' často sa ale stáva, že závislosti medzi vstupmi a spotrebou nie sú lineárne (napr. teplota & spotreba; najvyššia spotreba je práve pri veľmi nízkych a veľmi vysokých teplotách). Na odstránenie tohto problému sa obvykle používa niekoľko rôznych postupov. Jednou z alternatív riešenia, ktorú budeme testovať je, že sa teplota na vstupe uvedie vo viacerých vyšších mocninách (aby sme mohli modelovať aj nelineárnu závislosť). Druhým validným prístupom k zadaniu je rozloženie dát o spotrebe na spotrebu závislú a nezávislú od teploty a tieto časti modelovať separátne (vzťah premenných teplota a celková spotreba tu používame ako príklad). O nameranej efektívite týchto a alternatív a ostatných prístupov spomenutých v tejto správe budeme informovať v správach v ďalšej fáze projektu. Samostatnou otázkou, ktorú sa budeme snažiť zodpovedať je, či je vhodné mať separátne modely pre jednotlivé časové horizonty (napr. hodiny, dni, mesiace), alebo jeden model, ktorý dostane konkrétny časový horizont ako premennú. Lineárna regresia predstavuje jeden z klasických prístupov s relatívne slabými, no stabilnými modelmi. Jej neschopnosť zachytiť nelineárne závislosti sa dá parciálne obísť pridaním binárií atribútov.

Pri podpore tejto metódy nám môže byť nápomocná metóda generalizovanej lineárnej regresie, ktorá sa líši tým, že množinu atribútov najprv predspracujeme bázovými funkciami, čo umožní cieľovým premenným mať inú ako normálnu distribúciu chybovosti. Ďalšou alternatívou s potenciálom je proces lokálnej váhovanej aproximácie. Táto metóda definuje kľúčový stav v tréningových dátach a aj keď akceptuje všetky ostatné stavy vo vzorke, váži ich vplyv podľa funkcie ich vzdialenosti od kľúčového stavu. Príkladom takto využiteľnej funkcie je napr. známa bázová funkcia Kernel. V kontexte takejto grafovej interpretácie sa klasická lineárna regresia snaží optimálne vykresliť priamku cez všetky tréningové stavy, bez váhovania tréningových stavov. Nedostatkom lokálnej váhovej aproximácie je fakt, že pre každé nové tréningové dáta musíme definovať špecifickú sústavu kľúčových stavov.

Otázku, ktorú testovaním metód zodpovieme, je akým spôsobom optimálne zdefinovať kľúčové stavy v tréningových dátach v predikcii spotreby elektrickej energie. Ak by boli dáta rozvrhnuté rovnomerne, kľúčové stavy by bolo možné rozmiestniť v štruktúre pravidelného gridu. Pokiaľ máme dostatočný veľký počet dátových profilov, kľúčové stavy je možné vybrať aj náhodne. Posledným prístupom, je dynamicky vyberať optimálnu polohu kľúčových stavov pomocou metód reinforcement learning a analýzy zhlukov, čo je náročnejšie na výpočet ale pokiaľ v predikcii je dostatočná časová rezerva, metóda je veľmi efektívna aj pri menších, nerovnomerne rozvrhnutých dátových vzorkách.

Umelé neurónové siete

Neurónové siete predstavujú komplexný výpočtový systém, tvorený sériou jednoduchších efektívne prepojených elementov, ktoré spracovávajú informácie prostredníctvom ich dynamickej odpovede na externé vstupy, na základe ich aktuálneho stavu. Štandardne pozostávajú z väčšieho počtu vzájomne prepojených rozhodovacích uzlov, ktoré sa riadia aktivačnou funkciou. Uzly sú organizované do vrstiev. Dáta sú vložené do vstupnej vrstvy, ktorá ich komunikuje viacerým skrytým vrstvám, v ktorých prebieha ich spracovávanie. Neurónové siete majú viaceré špecifiká. Nie sú klasicky sekvenčné a nevyhnutne deterministické. V klasických modeloch v tomto prístupe absentuje centrálny procesor, ktorý je nahradený paralelnými odpoveďami uzlov na prezentované vstupy. Neexistujú tu ani oddelené pamäťové adresy na ukladanie dát počas analýzy, finálna informácia je obsiahnutá v celkovom aktivačnom nastavení všetkých uzlov v sieti. Novo získané poznatky sú teda reprezentované stavom siete samotnej, ktorý má väčšiu interpretačnú silu, ako súčet stavov všetkých uzlov v nej. Konexie medzi uzlami sa váhujú a výsledok je prezentovaný na výstupnej vrstve uzlov. Váhy uzlov sa nastavujú podľa dopredu zadaného pravidla učenia. Týchto pravidiel je niekoľko, najfrekvencovanejšie v klasických modeloch je tzv. delta pravidlo, ktoré administruje spätnú propagáciu chyby z jednej vrstvy, do uzlov vrstvy predchádzajúcej. Na základe pravidla učenia sa umelá neurónová sieť posúva na učiacej

krivke od náhodného typovania, k expertnému výberu. Efektivitu posilňuje aplikácia sigmoidálnych aktivačných funkcií, ktoré polarizujú aktivitu siete a pomáhajú ju stabilizovať.

Umelá neurónová sieť je v princípe metóda univerzálnej aproximácie, ktorá modeluje najlepšie také systémy, ktoré majú dostatočnú toleranciu k prvotnej chybovosti. Je ideálna na zachytávanie asociácií a detekovanie pravidielností v setoch dátových profilov, na identifikovanie vzťahov medzi premennými, ktoré sú neúplne popísané, na analýzu dátových štruktúr, ktoré disponujú veľkým počtom premenných, poprípade veľkými objemami. Po tom, čo sa nám podarí natrénovať sieť na požadovanú úroveň, môže byť použitá ako analytický nástroj na iných vzorkách dát. V tomto stave už nešpecifikujeme tréning a pustíme na neurónovej sieti iba propagáciu vpred. Výstupom takto nastavenej siete (forward propagation run) je predikčný model pre dáta, ktorý môže byť následne použitý v komplexnejšej analýze, či interpretácii.

Okrem všeobecných metodologických limitácií spomenutých pri popise metód lineárnej regresie, neurónové siete založené na princípe spätnej propagácie (ale aj viaceré iné typy) predstavujú v princípe čiernu skrinku. Okrem definovania primárnej architektúry siete a prvých testov s náhodnými číslami, nemá administrátor inú kompetenciu, ako kŕmiť sieť vstupnými dátami a zapracovávať výstupy. Niektoré nástroje (napr. NevProp, poprípade, Mactivation) dovoľujú získavať vzorky postupu siete v definovaných intervaloch, ale učiaca krivka siete sa rozvíja samostatne. Finálnym produktom tohto procesu je sieť sama, a teda nie rovnica alebo séria koeficientov (ako pri lineárnej regresii) ktoré by popisovali vzťahy premenných mimo architektúry siete. Potenciálnym limitom je, že neurónové siete je možné pretrénovať, t.j. sieť bude dávať dobré výsledky na tréningových dátach, ale jej výsledky na nových dátach budú slabé. Na odstránenie tohto fenoménu sa používajú nasledovné postupy: predčasné zastavenie tréningu, regulácia (ku chybe pripočítame aj vhodne prenasobenú hodnotu parametrov siete a túto hodnotu sa snažíme minimalizovať), dropout (počas tréningu náhodne vypneme niektoré neuróny), dropconnect (počas tréningu náhodne vypneme niektoré spojenia). Neutrálne siete a ich modely sú citlivé na voľbu parametrov a ich výber sa vo vedeckej komunite neriadi striktnými pravidlami. Preto je jedným z dôležitých cieľov nášho projektu testovať viaceré parametrické nastavenia neurónových sietí, ktoré budú tvoriť model predikcie spotreby elektrickej energie na Slovensku. Výzvou v projekte teda zostáva voľba vstupných parametrov a ich počet, ako aj veľkosť a typ skrytých vrstiev v sieti tak, aby mal systém dostatočnú predikčnú silu, stabilitu a boli zároveň časovo efektívne.

Hlavnou výhodou neurónových sietí je, že sú schopné modelovať nelineárne a dynamické závislosti. Sú tiež presnejšie ako metódy lineárnej regresie a rýchlo sa trénujú. Keďže sú

schopné zachytiť rôzne druhy vzťahov medzi premennými môžeme pomocou nich modelovať komplexné fenomény, ktoré by bolo náročné (niekedy až nemožné) modelovať klasickými štatistickými nástrojmi. Od roku 2006 sa do popredia dostávajú takzvané *deep networks*. Sú to neurónové siete s veľkým počtom vrstiev. Napriek tomu, že nie je presne popísané akým spôsobom ich skryté vrstvy fungujú, majú potenciál byť presnejšie ako obyčajné neurónové siete. Predbežné výsledky ukazujú, že pri dostatku dát majú vysokú efektivitu skoro v ľubovoľnej oblasti. Ich nedostatkom je, že sa ich niekedy skrátka nepodari natrénovať.

Support vector machines

Support vector machines fungujú tak, že v trénovacích dátach nájdú množinu dôležitých vzoriek (support vectors) a následne sledujú podobnosti medzi vstupom a dôležitými vzorkami, na základe čoho predikujú výsledok. Formálnejšie, táto metóda konštruuje nadroviny v priestore s veľkým, alebo nekonečným počtom dimenzií. Tieto nadroviny môžu byť následne použité na riešenie regresie alebo klasifikačných problémov. Support vector machines detekujú funkčné krajnice, dátové body jednej s klás, ktoré majú najväčšiu vzdialenosť k okrajovým dátovým bodom ostatných klás (lem). čím väčšie sú tieto vzdialenosti, tým menšia je generalizačná chyba klasifikátora. S využitím kernelu v hľadaní nadrovin s maximálnym lemom, je touto metódou možné vytvoriť aj nelineárne klasifikátory. Využívajú sa aj multi-klasové support vector machines, ktoré pripisujú inštanciám znaky, ktoré sú vyberané s konečného počtu dopredu identifikovaných elementov. Atraktívne sú najmä návrhy modelov, ktoré sa pozerajú na multi-klasový problém klasifikácie ako na jeden optimalizačný problém a nerozbíjajú ho na sériu problémov klasifikácie binárnych stavov (napr. Crammerova a Singerova metóda). Jedinečnou vlastnosťou support vector machines je, že maximalizovanie geometrického lemu medzi klasami a minimalizovanie klasifikačnej chyby prebieha zároveň. Nevýhodou je ich vysoká časová náročnosť pri tréningu (väčšinou kvadratická od počtu vzoriek), vďaka čomu sa nedajú použiť na veľké vzorky dát. V praxi sa úspešne používajú napríklad ako klasifikátory hypertextov, na kategorizovanie obrázkov, rozpoznávanie písaného textu, či klasifikovanie proteínov v medicíne.

Smartmetering a metodologický prístup k predikcii spotreby v domácnostiach

Pri téme smartmeteringu sa dostávame do polohy, ktorá má svoje metodologické špecifiká, na ktoré je nutné reagovať. Fakt, že zberné systémy budú kumulovať masívne množstvá dát znamená, že nebude možné vložiť všetky merania do jedného algoritmu ako jednorazový vstup. Je preto vhodné zberné jednotky (napr.. domácnosti) najprv zmysluplne klastrovať a na tvorbu modelov používať agregované dáta zhlukov dátových profilov. Pridanou hodnotou tohto prístupu je aj to, že sa v rámci klastrov dajú deskriptívne sledovať rôzne spotrebiteľské trendy (ich stabilita, intenzita rastu/klesania), ktoré môžu ovplyvniť vývoj v ďalšom období. Príkladom je sledovanie vývoja geografických oblastí a sledovať a predikovať vývoj spotreby v nich. V projekte prinesieme odpoveď na otázku, akým kľúčom je možné vytvárať dátové klastre najoptimálnejšie. Na efektívne klastrovanie dát existuje viac štatistických nástrojov,

tie akceptovateľné v kontexte profilu dátových setov, s ktorými počítame po nástupe smartmeteringu, si v krátkosti predstavíme.

K-means

K-means je metóda pôvodom z teórie venujúcej sa spracovávaníu signálov v šume. Jej princípom je rozdeľovanie dátových bodov do k-klastrov podľa ich podobnosti. Podobnosť tu definujeme na základe body s najbližším priemerom, ktorý v každom klastri funguje ako prototyp. K-means je rýchla a dobre paralelizovateľná metóda. Jej nevýhodou je citlivosť na škálovanie a korelované vstupné dáta.

Sparse coding

Táto metóda hľadá sadu vzorov v dátových setoch, ktorých naskladaním sa dáta vysvetlia. Optimálne, chceme každý dátový bod vysvetliť pomocou iba niekoľkých vzorov (preto sparse coding). Proces učenia prebieha autonómne, bez supervízora. Metóda nám umožní ľahšie interpretovať niektoré pohyby v dátach. Sparse coding sa dá použiť na získanie hlbšieho vhľadu do dát o správaní spotrebiteľov a výrobcov. Je ním možné napríklad tvorba profilov spotrebiteľských jednotiek, či detekovať aktivitu zapnutých elektrických spotrebičov (napr. za akú časť odberu zodpovedajú ktoré spotrebiče a za akých podmienok ich ľudia zapínajú). Tieto zistenia nielenže posilnia celkový predikčný model spotreby, ale majú aj samostatnú interpretačnú silu smerom k sledovaným vzťahom. Tu je kritickou premennou frekvencia dát (e.g. čím častejšie meriame, tým viac spotrebičov vieme odlíšiť). Pri malej frekvencii vieme sledovať len zmeny celkovej spotreby. Tradične sa v dátovom toku odlišujú len zariadenia s veľkou spotrebou, resp. sa zariadenia delia na spotrebné kategórie. Existuje niekoľko pomerne komplikovaných modelov, napríklad faktorizované skryté markovove modely, ktoré môžu finálnemu modelu zdvihnúť efektivitu, no ich popis v tejto fáze projektu nie je potrebný.

Optimalizačné nástroje

Predikcia ja síce dôležitou zložkou projektu, no nie je jedinou. Výstupné dáta z aplikovaných predikčných modelov spotreby nám môžu slúžiť ako interpretačné médium pri poradenstve (optimalizácii) smerom k výrobcovi a spotrebiteľom. Štatistické nástroje používané na optimalizáciu procesov a systémov nám umožnia túto úlohu do veľkej miery automatizovať. Ponúkame tu tri prístupy, ktoré budeme počas projektu testovať.

Integer linear programming

Integer linear programming predstavuje štandardnú techniku, ktorá sa vie adaptovať na riešenie problémov rôznych typov a štruktúr. Primárne sa zadajú premenné ako cieľ, ktorý treba optimalizovať (lineárna funkcia premenných), obmedzenia pre hodnoty premenných (tiež lineárne) a konkretizujúca inštrukcia niektorým premenným (napr. že môžu byť iba

celočíselné). Metóde nevádi linearita a v závislosti na charaktere dátovej vzorky je schopná namodelovať aj nelineárne závislosti s použitím vhodných transformácií.

Heuristické metódy

Heuristické metódy prehľadávajú stavový priestor riešení a robia malé lokálne optimalizácie. Heuristické riešenie predstavuje sériu nenáročných procedúr, ktorých cieľom je poskytnúť v závislosti na zadaní akceptovateľné, no nie nevyhnutne najlepšie riešenie z optimálnych riešení problému s vysokou komplexitou. Jedným z najpoužívanějších skupín optimalizačných heuristik sú genetické algoritmy. Tie sú založené na prepozíciách ako možnosť iterácie testovania, miera akceptovanej mutácie, pravidla prirodzenej selekcie (preferencia vhodnejších vlastností pred menej vhodnými) a pravidla ohraničeného počtu vstupov, agentov, či parametrov. Genetické parametre sú schopné optimalizovať veľké počty parametrov jedného agenta. Heuristické riešenia sú väčšinou šité na konkrétne zadaný problém.

Reinforcement Learning

Metódy Reinforcement Learning predstavujú komplexnú zbierku algoritmov a pomocných modulov, ktoré sa snažia optimalizovať cestu agenta cez problémový priestor, tvorený množinou akcií (transferov) a problémových stavov. Hlavnou motiváciou agenta je vyhýbanie sa trestom a/alebo získavanie odmeny. Týmto štýlom vie systém identifikovať optimálne riešenie poznaného problému, rozdeliť komplexný problém na jednoduchšie problémy tak, aby každé riešenie menšieho problému významné zvýšilo pravdepodobnosť dosiahnutia celkového riešenia alebo popísať problémový priestor, priradiť hodnoty jednotlivým akciám a stavom. Výhodou tohto prístupu je jeho pomerne jednoduchá skladba, kontrola efektivity riešenia, absencia pred-trénovania a fakt, že (na rozdiel od neurónových sietí) je celý proces viditeľný, popísaný vo výsledku a supervízor do procesu môže hocikedy priamo zasiahnuť. Systémy budované touto metódou sú schopné sa dynamicky prispôbiť zmene problémového priestoru alebo parametrov. Výsledkom nie je len finálna stratégia, ale skôr zoznam alternatív riešenia daného problému s hodnotami o ich efektivite a popis problémového priestoru a všetkých jeho zložiek, s ich impaktom na stratégiu riešenia. Nevýhodou je potreba iterácií (čas), ktorá narastá s komplexitou problémového stavu.

Záver

Cieľom analýzy bolo nájsť a popísať metodologicky vhodné prístupy k tvorbe inteligentných systémov, podporujúcich alebo nahradzujúcich analytickú stranu súčasného bilančného modelu estimácie odchýlky výroby a spotreby. Vymenované predikčné aj optimalizačné expertné systémy a predovšetkým ich kombinácie majú potenciál posunúť súčasné spracovanie dát v slovenskej energetike na novú úroveň a zároveň reagovať na očakávané zmeny v kontexte dátovej štruktúry.

5 Skúmanie kvality spracovaných energetických dát

Dostupné atribúty a ich možné hodnoty o odberno-odovzdávacích miestach a o meraniach z týchto miest sú uvedené v tabuľke 1.

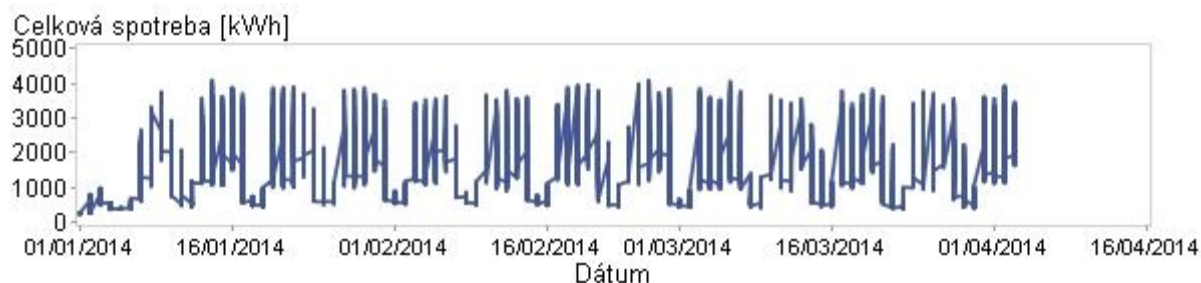
Tab. 1. Informácie dostupné o meraniach elektrickej energie na Slovensku.

Atribút	možné hodnoty	význam
odberno-odovzdávacie miesta (OOM)		
identifikátor	číslo	identifikačné číslo OOM
typ	Odber Dodávka Dodávka a odber	identifikácia, či dané OOM dodáva alebo odoberá el. energiu do / zo siete
napäťová hladina	NN VN VVN	napätie, na ktorom operuje OOM (nízke, vysoké, veľmi vysoké)
typ sústavy	Prenosová sústava Externá prenosová sústava Regionálna distribučná sústava Miestna distribučná sústava Priame vedenie (výrobca – odberateľ)	typ sústavy, v ktorej sa nachádza OOM
typ merania	A – priebehové meranie s možnosťou diaľkového odpočtu B – priebehové meranie bez možnosti diaľkového odpočtu C – meranie bez priebehového merania a bez možnosti diaľkového odpočtu	typ merania deklarovaný vo vyhláske Úradu pre reguláciu sieťových odvetví č. 24/2013 Z. z., ktorá ustanovuje pravidlá vnútorného trhu s energiami (elektrina, plyn)
lokalita	PSC	poštové smerovacie číslo, na ktorom s nachádza merač
merania		
dátum	dátum	deň merania
čas	poradové číslo	poradové číslo štvrťhodiny v dni
hodnota	reálne číslo	nameraná hodnota spotreby
jednotka merania	kW	celková spotreba je meraná v kWh, údaje za 15 minút sú v $\text{kW} \frac{1}{4} \text{h}$

Prvá vzorka dát, ktorú sme mali k dispozícii na experimentovanie (pozri obrázok 1.) pochádzala zo 65 odberno-odovzdávacích miest na Slovensku. Obsahovala merania za dobu troch mesiacov. Všetky miesta boli napojené na miestnu distribučnú sústavu a boli typu A (t. j. mali zapojené inteligentné merače elektrickej energie). V súčasnosti sa tento typ merania využíva len pri väčších odoberateľoch, akými sú rôzne výrobné, továrne, firmy a väčšie organizácie. Táto dátová množina neobsahovala údaje za dostatočne dlhé obdobie,

a preto bolo ťažké pomocou nej vytvoriť spoľahlivé a reálne použiteľné modely predpovede. Opis prvej použitej dátovej vzorky sa nachádza v tabuľke 2.

Merané údaje reprezentujú údaje o spotrebe energie odčítavané pre každú napr. štvrthodinu zvlášť. Priebeh celkovej spotreby odberateľov z prvej vzorky dát je vizualizovaný na **Error! Reference source not found...**



Obr. 1. Celková spotreba 65 slovenských odberno-odovzdávacích miest za obdobie troch mesiacov (93 dní).

Lokalita odberno-odovzdávacích miest podľa poštového smerovacieho čísla slúži pre doplnenie meteorologických alebo demografických údajov, ktoré by mohli obsahovať faktory ovplyvňujúce spotrebu elektrickej energie.

Tab. 2. Opis prvej dátovej vzorky.

Odberno-odovzdávacie miesta (OOM)	
počet	65
PSČ	60 (Martin – 030601) + 3 (Košice – 04454) + 2 (Bratislava – 83101)
typ	63 odberov a 2 dodávky
napätie	45 NN a 20 VN
typ sústavy	Miestna distribučná sústava
typ merania	A
merania	
obdobie	01.01.2014 – 03.04.2014 (93 dní)
počet záznamov/deň	96 (92 na 30.03.2014 – posun času)
spolu záznamov	579 868

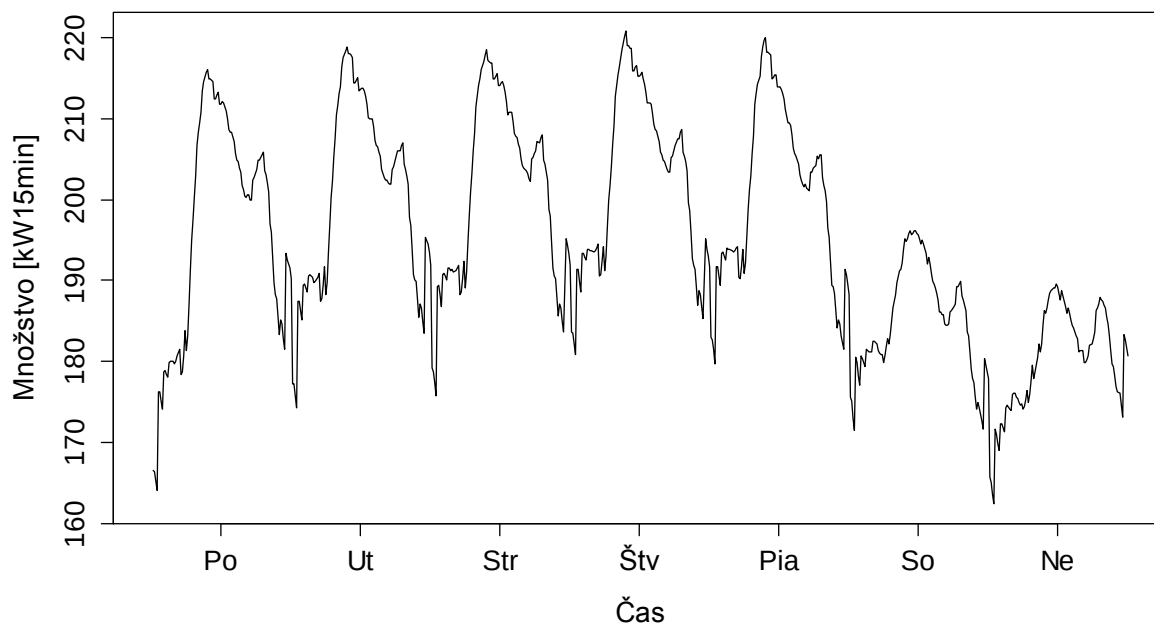
Táto vzorka bola v priebehu projektu doplnená o historické aj o nové získané merania. Konečná dátová množina je zhrnutá v tabuľke 3.

Tab. 3. Opis finálnej dátovej množiny.

Odberno-odovzdávacie miesta (OOM)	
počet	21 502
PSČ	1377 rôznych PSČ, najviac OOM (562) v okolí Košíc, na Východnom Slovensku
typ	19 446 odberov, 2 056 dodávok

	a 663 odberov aj dodávok
napätie	- neznáme -
typ sústavy	Miestna distribučná sústava
typ merania	A
merania	
obdobie	01.07.2011 – 14.05.2015
počet záznamov/deň	96 posuny časov – údajov za deň 30.10.2011 100 25.03.2012 92 28.10.2012 100 31.03.2013 92 27.10.2013 100 30.03.2014 92 26.10.2014 100
spolu záznamov	1 354 779 164
minimum	0
maximum	980 856,96 kW $\frac{1}{4}$ h
priemer	195,56 kW $\frac{1}{4}$ h
smerodajná odchýlka	8 785,14 kW $\frac{1}{4}$ h

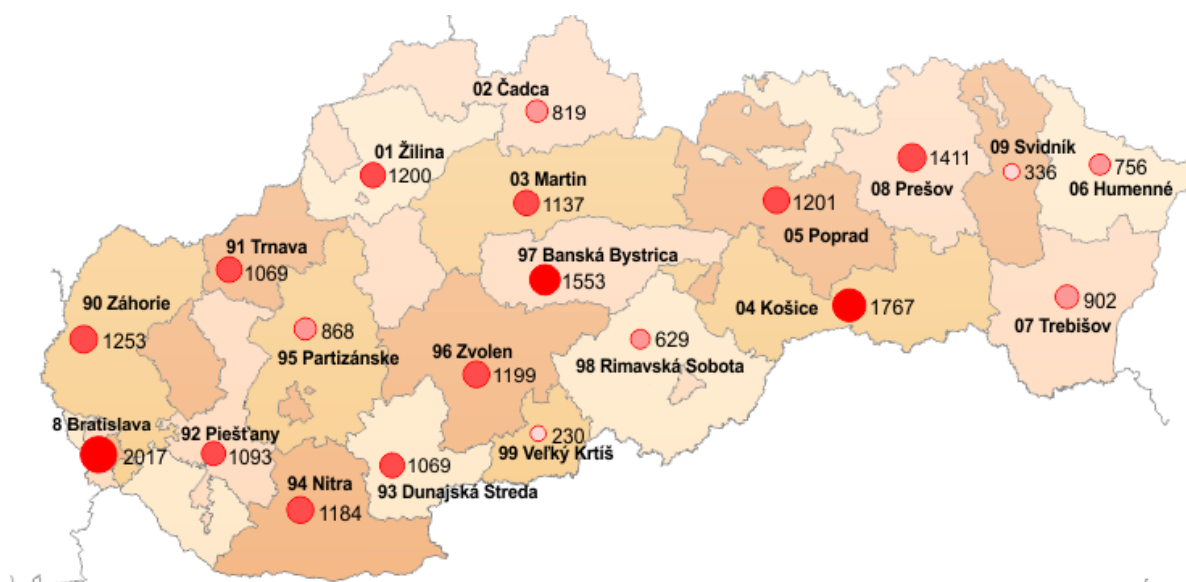
Priemerný týždenný priebeh jedného OOM je zobrazený na obrázku 2.



Obr. 2. Priemerný týždenný priebeh spotreby elektrickej energie v jednom OOM.

5.1.1 Predspracovanie dát pre experimenty

Pre experimentovanie sme vybrali z týchto dát vzorku, ktorá obsahovala iba OOM, ktoré odoberali energiu a mali úplné záznamy za čo najdlhšie obdobie. Vybrali sme úsek 01.07.2013 – 15.02.2015. Dáta sme rozdelili na základe prvých dvoch čísel PSČ do skupín a ich spotrebu sme zosumovali, aby bolo možné predpovedať spotrebu celých regiónov Slovenska. Rozdelenie OOM podľa PSČ je zobrazené na obrázku 3. Charakteristiky vybraných dát sú uvedené v tabuľke 4.



Obr. 3. Rozdelenie OOM podľa prvých dvoch čísel PSČ.

Tab. 4. Charakteristiky použitých dát z vybraných regiónov.

PSČ	región	počet OOM	priemerná spotreba	priemer na jedno OOM
04	Košice	722	35 501,854	49,172
05	Poprad	471	17 135,133	36,380
07	Trebišov	382	11 571,184	30,291
08	Prešov	580	18 671,795	32,193
8	Bratislava	1314	119 691,911	91,090
90	Záhorie	773	41 402,715	53,561
92	Piešťany	706	74 340,781	105,296
93	Dunajská Streda	594	28 196,959	47,470
95	Partizánske	584	34 298,912	58,731
99	Veľký Krtíš	114	2 124,887	18,639

6 Ochrana citlivých údajov v kontexte dát zo siete smartgrid

PREDPOKLADY A PODMIENKY, KTORÉ JE NUTNÉ NAPLNENIŤ, METÓDY, NÁSTROJE A AUTOMATIZÁCIA ANONYMIZAČNÝCH PROCESOV.

Predmetom tejto kapitoly je bezpečnostné usmernenie v oblasti ochrany citlivých údajov. Táto časť správy je formálnym výstupom analýzy vhodnosti rôznych prístupov k procesu anonymizovania a manažmentu osobných údajov v dátových štruktúrach potencionálneho systému sledovania pohybu výroby a spotreby elektrickej energie na Slovensku. Cieľom analýzy je popísať predpoklady a podmienky, ktoré treba splniť pri zbere, analýze, správe, uskladnení a distribúcii z inteligentných systémov estimácie odchýlky výroby a spotreby.

1. Všeobecne k ochrane osobných údajov v IS

Ochrana osobných údajov predstavuje jednu z hlavných priorít každého systému, ktorý zbiera, spracováva a uchováva informácie o svojich zložkách v takej forme, v ktorej sú tieto jeho zložky identifikovateľné externými entitami, ktoré k týmto informáciám nemajú oprávnený prístup.

Moderné systémy podporujúce súčasné bilančné modely estimácie odchýlky výroby a spotreby elektrickej energie v sieti nie sú výnimkou. Získanie dát a meta-dát o koncových užívateľoch, ktorými tieto systémy disponujú, predstavujú potenciálne veľký zdroj peňazí pre jednotlivcov a skupiny organizovaného kybernetického zločinu. Bohužiaľ, ochrana týchto údajov je vo svete často kriticky podhodnotená tak na úrovni legislatívnej (legislatívna definícia pojmu osobný údaj, aplikácia legislatívy) ako aj na úrovni operatívnej (práca s dátami, ich anonymizovanie, ich uskladnenie, fyzický prístup k dátam). Kybernetické útoky na vládne databázy európskych štátov zaznamenávajú v posledných rokoch prudký nárast. Dôvody tohto trendu sú tri: 1.) zvyšujúca sa počítačová gramotnosť populácie, 2.) znižujúca sa cena výpočtovej techniky a rast jej výkonu, 3.) uvoľňovanie štátnych dát v agregovanej a surovej podobe ako súčasť iniciatívy voľného prístupu k informáciám (OpenData), s nezvládnutým systémom anonymizovania týchto databáz. Viaceré štáty, na čele s Veľkou Britániou, Ruskom, Nórskom a Nemeckom vyhlásili otvorenú diskusiu o ochrane osobných údajov a financujú internú agendu, ktorá má systém ochrany štátnych databáz modernizovať.

Na nasledujúcich stranách popíšeme túto problematiku najprv v jej teoretických základoch, v kontexte limitov anonymizovania, rôznych prístupov k riešeniu tohto problému a nakoniec v kontexte systému bilančného modelu estimácie odchýlky výroby a spotreby elektrickej energie po spustení inteligentnej siete (SmartGrid).

2. Osobné údaje

Pojem osobný údaj je slovenskou legislatívou identifikovaný prostredníctvom zákona 122/2013 Z. z. o ochrane osobných údajov. Jeho aktuálne znenie je kompatibilné so Smernicou 95/46/ES Európskeho parlamentu a rady EU o ochrane jednotlivcov, spracovania osobných údajov a o voľnom pohybe osobných údajov.

Definícia pojmu v plnom znení: Osobnými údajmi sú údaje týkajúce sa určenej alebo určiteľnej fyzickej osoby, pričom takou osobou je osoba, ktorú možno určiť priamo alebo nepriamo, najmä na základe všeobecne použiteľného identifikátora alebo na základe jednej či viacerých charakteristík alebo znakov, ktoré tvoria jej fyzickú, fyziologickú, psychickú, mentálnu, ekonomickú, kultúrnu alebo sociálnu identitu.

Definíciou údaje je jednotlivý fakt a osobný údaj je charakterizovaný ako súbor údajov, teda súbor jednotlivých faktov, na základe ktorých je osoba určiteľná alebo určená. Preto náhodný súbor údajov musíme charakterizovať ako osobné údaje vtedy keď je možné jednotlivca na základe týchto údajov identifikovať, či už priamo alebo nepriamo.

3. Náročná aplikácia legislatívnych definícií v praxi

Táto definícia osobných údajov je síce evidentná, no v aplikácii často nedostatočná. To, či je údaj osobný alebo nie do veľkej miery závisí od kontextu, v ktorom je uvedený. Pokiaľ je jednotlivec popísaný ako veriaci vo vzorke, v ktorej je väčšina veriacich, tento údaj môžeme len ťažko použiť na jeho identifikáciu. Pokiaľ by bol ale jeden z mála veriacich vo vzorke, stáva sa tento údaj dobrým nástrojom na jeho identifikáciu, a teda podľa legislatívnej definície jeho osobným údajom. Takýmto spôsobom sa môžu stať osobnými údajmi informácie, ktoré by sme za takéto pôvodne vôbec nepovažovali (napr. doba odchodu do práce, preferencia jedla, ľubovoľná aktivita, či pasivita v ľubovoľnom čase na ľubovoľnom mieste).

Treba si preto uvedomiť, že osobné údaje nie sú len štandardne používané údaje ako meno, priezvisko, rodné číslo, adresa, číslo poistenca či informácie o zdravotnom stave. Tieto informácie sú frekventovane identifikované ako osobné údaje len preto, lebo sa v súčasnom systéme často používajú ako identifikátory. V správnom kontexte môže byť identifikátorom ktorýkoľvek dostupný údaj o skupine ľudí. Problém je, že údaje ako číslo občianskeho preukazu alebo rodné číslo sami o sebe poskytujú aj pridané informácie o človeku, ktoré v danom kontexte nie sú potrebné, no uľahčujú identifikovanie konkrétnej osoby treťou stranou.

Riešením tohto problému by mohlo byť zavedenie nového štandardu v systémoch štátnej správy prostredníctvom MV SR, ktoré by do budúcnosti zabránilo takýmto kros-validáciám dát. Z pohľadu štruktúry by malo ísť o alfanumerickú sadu znakov priradenú všetkým osobám evidovaným napríklad v registri obyvateľov. Takýto identifikátor nemôže obsahovať žiaden údaj popisujúci ľubovoľnú charakteristiku osoby, čím sa odlišuje od teraz používaných systémových identifikátorov ako rodné číslo, z ktorého možno odvodiť pohlavie, vek alebo rasu osoby. Takýto identifikátor by bol prirodzene menej náchylný na zneužitie kros-validáciu a vhodnejší ako kľúč pri procese anonymizovania dát.

4. Proces anonymizovania dát

Ak má ľubovoľný komunikačný systém vytvorený kvalitný proces anonymizovania dát, umožňuje mu to prenos dôležitých informácií cez miesta s nižšou informačnou bezpečnosťou, ako napríklad prenos medzi organizáciami. Anonymizovaný prenos bude disponovať relatívne nízkym rizikom zverejnenia citlivých informácií a stále umožňuje následnú analýzu dát a hodnotenie prenesených informácií.

Vo všeobecnosti, je anonymizovanie dát proces konvertovania čistých (surových) dát do sofistikovanej podoby, z ktorej surové dáta nemožno extrahovať reverzným procesom ich primárnej manipulácie. Podané inak, je to proces, v ktorom zoberieme dáta, z ktorých sme schopní identifikovať konkrétnu entitu (napr. osobu alebo skupinu osôb) a pretransformujeme ich na dáta, z ktorých sa tieto entity nedajú identifikovať.

K základným definíciám však treba doplniť niekoľko podmienok. Prvá, dáta anonymizované (za konkrétnym cieľom) podľa hore uvedených definícií môžu stále poslúžiť ako pomocný nástroj na identifikáciu entít, na ktorých ochranu sme sa pri anonymizovaní explicitne nezamerali. Druhá, na anonymizovanie dát sa používa súbor rozmanitých techník, ktorých cieľom je zachovať rozloženie dátového poľa (charakteristik dátového setu) ako jeho veľkosť, pozíciu dát, typ dát a ich frekvenciu. Dáta musia aj po anonymizovaní vyzerať realisticky, aby mohli byť použité v testovacom prostredí.

Z uvedeného vyplýva jeden veľmi dôležitý fakt. Z absolútne anonymizovaných dát (vynulovanie hodnôt) bude mať analytik len malý úžitok, nakoľko sa anonymizovaním znižuje ich hĺbka. Zmenšujú sa možnosti ich analýzy. Pri procese anonymizovania preto ide vždy o nájdenie kompromisu medzi bezpečnosťou a utilitou dát.

5. Základné metódy anonymizovania dát

Prvou, relatívne intuitívnou metódou anonymizovania je **vytvorenie náhradných symbolov**. Nahradenie číslíc dátumu narodenia náhodnými znakmi je pre útočníka neprekonateľná prekážka. Je ale jasné, že takýto proces anonymizovania výrazne znižuje informačnú hodnotu

dát. Anonymizovanie tvorbou náhradných dát podľa kľúča (nie náhodným priradovaním symbolov) predstavuje skôr pseudo-anonymizovanie dát, nakoľko po obdržaní kľúča vieme dáta dostať do pôvodnej podoby. Existencia kľúča zvyšuje pravdepodobnosť úspešného útoku ale zároveň zvyšuje informačný potenciál dát.

Druhou metódou anonymizovania je **nulovanie dát**. Je to radikálny prístup k anonymizovaniu, pri ktorom sa citlivým dátam pripíše hodnota N/A alebo NULL. Týmto ťahom sa však informačná hodnota dát znižuje. Na druhú stranu, dáta takto anonymizované sú maximálne odolné voči nechcenej interpretácii. Tento prístup neodporúčame použiť na dáta, ktoré majú prejsť komplexnejšou post-anonymizačnou analýzou. Z takýchto dát vieme totiž vyčítať maximálne ich frekvenciu výskytu v čase a ich objem. Podobne drastickým prístupom k anonymizovaniu, ktorý je dobré používať najmä pri extrahovaní za účelom odovzdania dát tretej strane nad ktorou nemáme kontrolu (nevieme na čo dáta potrebuje ani ako ich chce analyzovať) je odovzdanie dát v agregovanej forme. Zo surových dát sa spravia dáta s menšou granularitou, napríklad priemerovaním.

Treťou metódou anonymizovania je tvorba virtuálnych dátových vzoriek. **Virtuálne dátové vzorky** sú také vzorky, ktoré sú požadovanými, dopredu definovanými charakteristikami identické s pôvodnou vzorkou dát (napríklad poradím, typom, frekvenciou, rozložením dát), no v ostatných charakteristikách sú ich stavy nahradené inými hodnotami (napríklad vymyslené tel. čísla, dátumy narodenia). Zo štandardných metód anonymizovania dát je tvorba virtuálnych vzoriek asi tá najefektívnejšia. Na aplikáciu tohto typu anonymizovania potrebujeme vytvoriť knižnicu, z ktorej sa budú virtuálne údaje vyberať a pripisovať reálnym dátam a logické pravidlá, podľa ktorých ich budeme pripisovať jednotlivým stĺpcom. Anonymizovanie pomocou virtuálnych vzoriek poskytuje dobrý kompromis medzi zachovaním informačnej hodnoty dát po anonymizovaní a ich bezpečnosťou.

Klasickou chybou pri aplikácii metódy virtuálnych vzoriek je takzvaný data overflow. Stane sa to vtedy, keď pri tvorbe virtuálnej vzorky nedodržíme charakteristiky pôvodnej vzorky (zväčšíme-zmenšíme počet riadkov alebo stĺpcov alebo zmeníme distribúciu dát). Ak chceme virtualizovať naše dáta naozaj kvalitne, treba ich v prvom rade popísať so všetkými ich charakteristikami, ktoré potrebujeme zachovať. Napríklad rodné čísla sú všetky deliteľné číslom 11. Ak by sme ich chceli virtualizovať ale stále použiť ako identifikátor v systéme, ktorý kontroluje rodné čísla deliteľnosťou jedenástimi, musíme túto charakteristiku zachovať aj vo virtuálnych dátach. Z bezpečnostného hľadiska sa oplatí virtualizovať dáta tak, že sú na prvý pohľad na nerozoznanie od pôvodných dát. Číselné hodnoty nahradzujeme hodnotami v podobnej sústave, geografické názvy geografickými názvami a nie len náhodným zhukom písmen.

Štvrtou metódou anonymizovania je **shuffling**. Je to prístup, pri ktorom prehádzame originálnu vzorku dát po dopredu definovaných osách, podľa konkrétnych premenných (dátových reťazcov). Metóda vyžaduje prípravu a zamyslenie sa nad jej vhodnosťou. Aby bola

efektívna, musí byť prehádzanie dát dostatočne komplikované. Podobné premenné by sa nemali prehadzovať ak je premenných takého typu v dátach málo (umožňuje to ľahké de-anonymizovanie). Metódu neodporúčame použiť pri vzorkách s malým počtom premenných. Pri dostatočne veľkých vzorkách, kde dáta predstavujú názvy, mená alebo iné nominálne premenné má anonymizovanie prostredníctvom shufflingu lepší výstup z pohľadu informačnej hodnoty dát ako anonymizovanie prostredníctvom virtuálnych vzoriek. Pri anonymizovaní numerických intervalových hodnôt a niektorých ordinálnych premenných vieme použiť **metódu posunutia**, pri ktorej posunieme hodnoty o vopred definovaný interval alebo o náhodný interval. Manipuláciou veľkosti intervalu vieme manipulovať zachovanie štatistickej hodnoty dát.

Vo všeobecnosti, predovšetkým ale pri shufflingu, by sa pri anonymizovaní dát mala venovať pozornosť hodnotám deviuúcim od priemeru, unikátne, nezodpovedajúce štandardnému rozloženiu dát vo vzorke. Platí tu pravidlo, čím sú dáta vo vzorke monotónnejšie, tým ľahšie je vzorka anonymizovateľná, čím viac štatistických výnimiek je vo vzorke dát, tým jednoduchšie je pre útočníka extrahovať zo vzorky parciálne dáta.

Pri implementácii základných anonymizačných techník je vhodné jednotlivé prístupy kombinovať. Správnym prístupom vieme za pomoci jednoduchých nástrojov anonymizovania dosiahnuť značnú úroveň bezpečnosti a zachovať pri tom interpretatívnu hodnotu dát v post-anonymizačnej analýze. K návrhu a aplikácii systému procesov anonymizovania dát je treba pristupovať systematicky a s cieľom nájsť kompromis medzi ochranou osobných informácií a zachovaním analytického potenciálu týchto informácií.

Neriadené alebo prehnané anonymizovanie dát je často kontraproduktívne a nepotrebné lebo:

- Útočníci vo väčšine prípadov nemajú i tak všetky potrebné informácie na systematickú extrakciu dát.
- Nie každá informácia z databázy je citlivá v kontexte cieľu zberu týchto dát.
- Nie všetky dáta sú rovnako senzitívne v kontexte analýz a interpretácie, za ktorých účelom ich zbierame. Nie všetky vzťahy medzi premennými sú predmetom anonymizačného procesu. Napríklad vzťahy medzi používateľmi v on-line komunikačných systémoch sú často verejné, no vzťahy užívateľa k službe (napr. prihlásenie do služby, frekvencia používania) sú skoro vždy považované za senzitívne.
- Systém anonymizovania je vhodné od začiatku stavať na poznatkoch o tom, ktoré informácie v dátach sú a ktoré nie sú senzitívne. V projekte treba definovať čo presne je pokladané za osobný údaj a čo nie. Následne je nutné dopredu nadizajnovať plán pomeru anonymizovania dát proti strate ich informačnej hodnoty a riadiť sa ním počas procesu vývoja anonymizačného systému.

Na anonymizovanie dát sa používajú aj pokročilé prístupy a nástroje ako hashing, **syntaktické bezpečnostné modely** (K-anonymity, l-diversity, t-closeness, e-differential privacy) zameriavajúce sa na publikáciu dát a špecifikáciu ako by dáta mali vyzerieť, ak majú byť zverejnené spolu s inými typmi dát, či systém **crowd blending privacy**. Každý z prístupov má samozrejme svoje výhody (väčšinou komplexitu) a nevýhody. Crowd blending privacy je napríklad ľahko dosiahnuteľný a dynamický, no neponúka silnú záruku bezpečnosti. Algoritmy typu **differential privacy** je náročné dizajnovať, ak majú po anonymizovaní na dátach bežať komplikovanejšie analytické úlohy, nakoľko potrebujú vlastnú optimalizáciu pre každú úlohu. Tento problém nedostatočnej utility majú aj syntaktické bezpečnostné modely (optimalizujú správne k pre k-anonymity, správne l pre l-diversity, správne e pre e-differential privacy pre každú úlohu). Problém s často používanou metódou **hashing** si popíšeme nižšie ako dobrý príklad procesu de-anonymizovania dát.

6. Pseudo-anonymizovanie dát

Ľudia si tradične zamieňajú anonymizované dáta za pseudo-anonymizované dáta. **Pseudo-anonymizácia** je proces v ktorom zoberieme surové dáta, v ktorých môžeme identifikovať osobné údaje, a konvertujeme ich do takého formátu, v ktorom tieto údaje nie je možné identifikovať. Zároveň disponujeme metódou, ktorou z týchto pseudo-anonymizovaných dát vieme konvertovať vzorku naspäť do stavu, v ktorom vieme opäť identifikovať osobné údaje. Pseudo-anonymizovanie je teda reverzibilný proces, ktorý nám umožňuje uzamykať a odomykať hlbšiu informačnú a štatistickú hodnotu dát.

Tento prístup je dostatočný a dokonca často požadovaný, lebo v zamknutom stave disponujú dáta relatívne dobrou bezpečnosťou (riziko, ktoré je vždy prítomné je získanie kódovacieho kľúča útočníkom) a po odomknutí majú dáta svoj plný štatistický potenciál. Nebezpečenstvo predstavuje stav, keď si ľudia myslia, že ich dáta sú plne anonymizované a také ich ponúknu tretej strane, pri čom sú len pseudo-anonymizované. Naopak, plne anonymizovať dáta v kontexte, v ktorom je ideálne či dostačujúce pseudo-anonymizovanie je zbytočný krok, ktorý okliešťa ich analytický potenciál.

7. Proces de-anonymizovania a nedostatočné anonymizovanie dát

Napriek tomu, že podľa vyššie uvedených definícií, je proces pravého anonymizovania nereverzibilný, v reálnych aplikáciách existujú cesty, ako anonymizované dáta de-anonymizovať. Tieto metódy predstavujú hrozby pre systémy, v ktorých je anonymizovanie dát podmienkou. De-anonymizovanie je proces, ktorým z viacerých anonymizovaných databáz vieme pomocou **kros-referencie** a **kros-validizácie** extrahovať aspoň časť senzitívnych dát o užívateľoch systému. Problémom, vďaka ktorému je de-anonymizovanie možné, je vzrastajúci počet dát charakterizovaných ako osobné údaje (citlivé dáta) a zvyšujúci sa počet verejných zdrojov dát, ktoré môžeme podrobiť kros-referencii. Ako príklad, len na základe kros-referencie dát z verejných databázach USA, kde každá z nich je

anonymizovaná pre svoj kontext, no ponecháva zostatkové (viditeľné) údaje, ktoré v tomto kontexte nie sú považované za citlivé údaje, je možné do hĺbky identifikovať 87% obyvateľov USA len na základe ich PSČ, veku a pohlavia. Legislatívne rozširovanie pojmu „osobný údaj“ preto nie je cestou k zvýšeniu bezpečnosti systému.

De-anonymizovaniu dát je možné sa brániť dodržaním niekoľkých podmienok:

- i) uvedomme si a prijmime fakt, že riziko de-anonymizovania je reálne,
- ii) tvorbu bezpečnostného systému vnímajme ako snahu o dynamické a kontinuálne minimalizovanie rizika,
- iii) komunikovať tento prístup zúčastneným inštitúciám (často požadujúcim absolútne riešenia aj tam, kde nie sú možné alebo zmysluplné) je náročné, no čím viac zúčastnených to pochopí, tým ľahší bude vývoj a implementácia,
- iv) racionálne limitujme detail dát,
- v) počas vývoja, by sme sa mali opakovane pozrieť na dáta, ktoré plánujeme zo systému posielať tretím stranám a predstaviť si, ako by jednotlivci mohli byť týmito dátami identifikovaní.

Sú v dátach segmenty ktoré sú viac senzitívne ako sú potrebné pre spracovanie a interpretáciu dát? Ak áno, dáta sú nedostatočne chránené. Ak ste schopní agregovať jednotlivcov v dátach do stabilných kohort a malých skupín, ktoré reprezentujú priemerné správanie v dátach, dáta sú náchylnejšie k de-anonymizovaniu.

Ukážku nedostatočného anonymizovania dát, či neefektívne aplikovanej metódy popíšeme na frekventovane používanej metóde anonymizovania - metóde **hashing**. Pre pochopenie popis zjednodušíme. Predstavme si hashing ako funkciu, v ktorej na jednej strane dáme údaj (napr. rodné číslo) a na výstupe nám funkcia dá jedinečný identifikátor ako: k9089jthg876fe6534. Funkcia má tri dobré vlastnosti: a) pre ten istý vstup dodá stále rovnaký výstup, b) pre iný vstup dodá iný výstup, bez viditeľného vzťahu k vstupu prvému, c) hashing nie je reverzibilná funkcia, takže útočník ktorý by chcel získať rodné číslo z výstupu k9089jthg876fe6534 otočením procesu funkcie sa nikde nedostane. Tieto vlastnosti sú ušľachtilé, no dostatočne skúsený útočník vie stále hashing obísť. Stačí že má k dispozícii hashovaciu funkciu a vie, že boli hashované rodné čísla. Ďalej vie, že rodné čísla majú konkrétnu štruktúru. Stačí mu už len vygenerovať všetky existujúce kombinácie rodných čísel (nie je ich až tak veľa), vložiť ich postupne do hashovacej funkcie a porovnávať výstupy s naším až kým neidentifikuje naše rodné číslo. Ešte pripravenější útočník si vygeneruje kombinácie dopredu a vytvorí ich index, aby naše rodné číslo získal v priebehu niekoľkých sekúnd.

Príklad nedostatočného anonymizovania dát z energetiky predstavuje link <http://open.enernoc.com/data/> kde sú k dispozícii 5-minútové dáta o spotrebe elektrickej energie od 100 anonymizovaných komerčných a industriálnych klientov. Samotné mená organizácií sú síce zakódované, no aspoň niektoré z nich bude možné identifikovať na základe metadát (sú tiež k dispozícii) o randomizovanej geografickej polohe, typu priemyselného odvetvia a rozlohy v stopách štvorcových. To je pochopiteľne neprijateľné.

8. Dizajn bezpečného systému

Pri dizajne a vývoji systému ktorý bezpečne pracuje s dátovým tokom sa musíme riadiť niekoľkými princípmi. Prvým je interoperabilita existujúceho a vyvíjaného softvéru. Aplikácie musia odpovedať na viacero typov dotazov, vzniká požiadavka, aby existujúci softvér a softvér vo vývoji bežal na anonymizovaných a pseudo-anonymizovaných dátach. Druhou požiadavkou bude pochopiteľne extenzibilita existujúceho softvéru, predpokladáme, že tento bude musieť generovať agregované štatistiky. Škálovateľnosť systémov je podmienkou, nakoľko anonymizované dáta budú mať aspoň takú veľkosť ako originálna vzorka dát, pričom počet agregátov a metadát požadovaných jednotlivými zložkami systému môže dramaticky narásť. Dôležitým princípom je kontinuálne vylepšovanie utility dát z dátového toku, nakoľko dáta nebudú používané v izolovanom systéme a dátové agregáty vyťahované pre potreby jednotlivých inštitúcií budú v závislosti na ich rozmanitosti potenciálnymi zdrojmi dát pre proces de-anonymizovania kros-validáciou. Služby poskytovania agregovaných dát zo systému by preto mali byť koordinované.

Pri tvorbe anonymizačného systému je potrebné otvoriť aktívnu diskusiu s vlastníkom dát, ktorý bude systém spracovávať a budúcim administrátorom tohto systému. Cieľom diskusie bude, aby všetky strany intuitívne chápali nepriamu úmeru vzťahu medzi mierou anonymizovania dát (bezpečnosťou) a použiteľnosťou dát v analýze. Skupina by mala spoločne definovať mieru tohto pomeru, ktorú sú ochotní prijať v kontexte modelov (napr. modelov estimácie odchýlky spotreby elektrickej energie na Slovensku) a na úrovni konkrétnych dátových parametrov. Výstupom diskusie by mala byť aj zadefinovaná metrika bezpečnosti dát a metrika ich utility.

Na druhej strane, systém má byť pripravený na jeho užívateľov a ich požiadavky. Užívateľ má byť informovaný, že dáta, ktoré dostane z databázy sú anonymizované alebo sú to agregované dáta s menšou granularitou. Obmedzenia a štandardný formát, v ktorom sa budú dáta poskytovať tretím stranám, musia byť kvalitne popísané, aby sa predišlo nedorozumeniam pri tvorbe nástrojov, ktoré budú tieto upravené dátové vzorky automaticky generovať a zasielať. V ideálnom prípade by mal byť vytvorený systém na mapovanie query a requestov na databázu, ktorý by pomohol v ich manažmente, ukazoval by zaťaženie databázy požiadavkami a umožnil by detekovať podozrivé požiadavky (napr. časté, neautorizované). S takýmto systémom by bolo možné vytvoriť interaktívny prístup k databázam, ktoré by poskytovali agregované dáta podľa požiadaviek užívateľov (ak by ich požiadavka spĺňala bezpečnostné podmienky), namiesto poskytovania objemných anonymizovaných dátových vzoriek, nad ktorými by systém po odovzdaní nemal kontrolu.

9. Podmienky ochrany citlivých údajov v inteligentnej elektrickej sieti

Jedným z cieľov Smart Grid iniciatívy je umožniť správcovským inštitúciám, užívateľom, dodávateľom a tretím stranám monitorovať a kontrolovať spotrebu elektrickej energie. Predpokladá sa, že vyššia frekvencia zbieraných dát spolu so silnejšími analytickými nástrojmi, interpretáciou a integráciou vnútro štruktúrnej komunikácie záverov analýz zefektívni systém pre vyššie menované strany. Umožní im robiť lepšie rozhodnutia, lepšie pochopiť potreby a požiadavky klientov a distribučných partnerov, čo bude mať pozitívny dopad na efektivitu distribúcie elektrickej energie.

Prirodzene, Smart Grid iniciatíva so sebou prináša sériu vážnych otázok spojených s ochranou osobných údajov koncových užívateľov, nakoľko dáta o spotrebe domácností budú zbierané so značnou frekvenciou a granularitou sekundárnych premenných (napr. geografická poloha smartmetru). Monitorovanie spotreby elektrickej energie je v princípe monitorovanie správania sa ľudí na dennej (minútovej) báze. Tieto dáta majú extrémnu hodnotu napríklad pre reklamné spoločnosti, štátne bezpečnostné zložky a kriminálne aktívne skupiny. Spreneverenie či de-anonymizovanie takéhoto typu informácie by zasiahlo do základných ľudských práv užívateľa a preto by mala byť tvorba kvalitného bezpečnostného systému na zber, manažment, analýzu a distribúciu dát o elektrickej výrobe a spotrebe absolútnym imperatívom iniciatívy Smart Grid.

10. Podmienky na vývoj bezpečnostného systému v kontexte Smart grid

V kontexte tvorby tohto projektu musíme zachovať stav, v ktorom bude možné pripísať dáta ku špecifickému smartmetru alebo OOM a zároveň ochrániť osobné údaje fyzických a právnických osôb, ktoré tieto dáta generujú svojim správaním. Tento cieľ musí byť splnený bez negatívneho dopadu na kontinuálny tok odmerných dát vo vysokej frekvencii a bez negatívneho dopadu na prebiehajúce operácie v sieti.

Z kontextu ďalej môžeme definovať tieto špecifikácie budúceho systému:

1. Odmerné dáta potrebné pre účtovanie spotreby a fakturácie musia byť pripísateľné OOM, bezpečne priradené ku konkrétnemu klientovi alebo majiteľovi účtu.
2. Pre účely účtovania je dostatočné generovať agregované štatistiky pripísateľných dát k OOM na dennej mesačnej a štvrťročnej báze.
3. Odmerné dáta pre účely estimácie odchýlky výroby a spotreby (pre bilančné systémy) a pre účely manažmentu siete nemusia byť pripísateľné konkrétnym domácnostiam, osobám či právnickým osobám. Anonymizované dáta sú dostatočné, pokiaľ: i) sú overiteľné ii) je ich možné priradiť ku konkrétnej správcovskej distribučnej stanici, spravujúcej sub-set smartmetrov (distribučné spoločnosti a sub-sety na úrovni okresov, mestských častí, či obytných blokov).

4. Anonymizované dáta budú zbierané s vysokou frekvenciou, aby bolo možné v reálnom čase reagovať na fluktuáciu dostupnej elektrickej energie v sieti a na potenciálne problémy s distribúciou v sieti.

5. Najmenšia (akceptovateľná) jednotka anonymizovaných dát o spotrebiteľoch pre účely estimácie odchýlky výroby a spotreby, je dátový agregát na úrovni distribučných staníc, z ktorých sa energia dodáva priamo jednotlivým koncovým konzumentom (napr. prvá úroveň nad smartmetrami domácností). Systém nepotrebuje vedieť aké dáta sa generujú ktorým smartmetrom, zaujímajú ho vlastnosti týchto dát a ich agregát v čase. Systém by však mal byť schopný spravovať jednotlivé smartmetre a ich funkcie (napr. vypnúť dodávku) aj keď užívateľ bude anonymný, nakoľko je to požiadavka zo strany manažmentu siete.

6. Podmienkou pre stavbu bezpečného systému je dôverný vzťah medzi podnikmi verejných služieb zodpovedných za správu siete, úložiskom dát a dodávateľmi smartmetrov. V súčasnosti nie je úplne jasné, ako bude na Slovensku postavené vlastníctvo dát generovaných smartmetrami v systéme Smart Grid a kto všetko bude mať oprávnenie prístup k týmto dátam v čistej forme, či ich uchovávať. Kým podľa niektorých nariadení by mal klient (užívateľ) vlastniť dáta o vlastnej spotrebe a výrobe, nakoľko ich generuje vlastným správaním, iné už zabehnuté systémy dávajú formálne vlastníctvo dát na stranu inštitúcií spravujúcich sieť, či prekvapivo, poskytovateľom služieb. Prístup k dátam pre jednotlivé strany je väčšinou definovaný legislatívou na úrovni štátu, ktorá sa snaží prihliadať napríklad aj na prístup k dátam počas mimoriadnych situácií. Tieto vzťahy by mali byť jasné pred začatím vývoja bezpečnostného systému, ktorý, ako sme už uviedli, do značnej miery závisí od kontextu, do ktorého je zasadený. Na základe týchto špecifikácií je následne možné odvodiť pravidlá manažmentu a manipulácie dát v takomto systéme:

Zber dát by mal byť v každej časti systému prísne obmedzený len na dáta potrebné k takým operáciám, za ktoré každá inštitúcia v systéme Smart Grid zodpovedá (napr. plánovanie a manažment, testovanie efektivity distribúcie, výpočet bilancie, manažment účtov a fakturácia).

Uchovávanie dát, z ktorých je možné extrahovať osobné údaje či informácie o správaní užívateľov, by malo byť v každej časti systému prísne obmedzené na dobu potrebnú na splnenie aktivít s ktorými bol koncový užívateľ oboznámený a ktoré akceptoval (napr. aktivity potrebné na tvorbu predikcií budú zrejme predstavovať horný strop, lebo je na ne potrebné značné množstvo dát). Po tom, ako boli všetky schválené aktivity na dátach vykonané, inštitúcie sú povinné nenávratne zničiť pôvodné dáta, všetky ich kópie a metadáta z nich vytvorené.

Metodologickou podmienkou distribúcie dát je **zadefinovanie a zachovanie bezpečných dátových intervalov**. Aby sme mohli dobre implementovať princíp minimalizácie zbieraných dát, potrebujeme vedieť, čo vieme odvodiť zo zbieraných dát ako funkciu rôznych intervalov. Zistiť, ako funguje vzťah medzi veľkosťou intervalov zberu, množstvom informácií, ktoré z takýchto zberov vieme odvodiť a bezpečnosťou v kontexte Smart Grid siete, si vyžaduje ďalší výskum.

Okrem intuitívnej **agregácie dát v čase**, je vhodné vytvárať aj **agregáciu dát na základe užívateľov**. Bezpečnostné riziká spojené s meraním energetického spotrebného správania

užívateľov vieme značne obmedziť, ak bude systém vytvárať agregáty na základe blízkosti alebo podobnosti užívateľov (napr. podľa obytných blokov, podľa podobného profilu spotreby etc.) a individuálne dáta anonymizovať. Aj tu je potrebné definovať do akej miery treba dáta agregovať a aký je v konkrétnych prípadoch správny pomer medzi bezpečnosťou a utilitou agregátov. Pritom treba mať na pamäti, že agregácia neuchráni jednotlivca, ak majú všetci v agregovanej vzorke podobné hodnoty vo viacerých premenných. Vtedy je metódu potrebné kombinovať s anonymizovaním iného typu. Minimálne charakteristiky vzorky, potrebné na to, aby bola agregácia efektívna sú: i) vzorka o veľkosti minimálne 15-ich užívateľov alebo ordinálne/nominálne rozlíšiteľných skupín užívateľov, ii) v každej rozlíšiteľnej skupine užívateľov nesmie byť jedna premenná, podľa ktorej je možné identifikovať len jedného užívateľa a zároveň ani jeden užívateľ nesmie generovať viac ako 15% všetkých dát, ktoré budú spoločne agregované. Pokiaľ pravidlo 15/15 nie je v agregáte dodržané, takéto dáta neodporúčame posúvať tretím stranám, nad ktorými nemáme kontrolu.

Latencia dát. Na všetky dáta, ktoré budú odovzdávané priamo užívateľom, fyzickým osobám, či komerčným a nekomerčným právnickým osobám mimo systému, navrhujeme aplikovať princíp prístupu k dátam s latenciou. Jednou z obáv zneužitia dát o spotrebe elektrickej energie je, že budú použité na odhad neprítomnosti užívateľov v ich domácnostiach. Pokiaľ by systém tretím stranám ponúkal dáta s časovým odstupom a zároveň s upravenými časovými radami, tento problém by odpadol.

11. Analýza anonymizovania Smart Grid dát v zahraničných systémoch

Tvorba systému, ktorý zaručuje bezpečnú manipuláciu s dátami získanými v inteligentnej energetickej sieti je relatívne neprebádaná téma. Pri analýze už nasadených systémov sme zistili, že systémové regulácie a bezpečnostné požiadavky na dátový manažment sú často špecifikované príliš všeobecne, ak vôbec. Výsledkom je, že nastavenie implementačnej praxe pre systémy manipulácie dát je ponechané na verejnú správu nižšej úrovne, na komerčné a konzultačné firmy. Tieto inštitúcie si preto sami regulujú bezpečnosť dát s ktorými pracujú. Bohužiaľ, v niektorých prípadoch nezávisle na sebe a s nedostatočnou odbornosťou.

Vlády a organizácie, vydávajúce štandardy, vypracovávajú aj bezpečnostné štandardy a nariadenia, usmerňujúce vývoj systémov na ochranu osobných údajov v kontexte iniciatívy Smart Grid. Komplexné a profesionálne odporúčania v severoamerickom prostredí poskytli napríklad Výbor pre severoamerickú energetiku (North American Energy Standards Board), Národný inštitút technologických štandardov (National Institute of Standards and Technology) a Oddelenie energetiky spojených štátov amerických (U.S. Department of Energy) a The Smart Grid Policy Framework, vydaný Prezidentskou kanceláriou v roku 2011. Konštruktívnym je aj Zborník č. 2 expertnej skupiny Smart Grid pri Európskej komisii, popisujúci regulačné odporúčania pre bezpečnosť dát, manipuláciu dát a ochranu osobných údajov pre štáty Európskej únie.

Praktické inšpirácie pri plánovaní a vývoji bezpečnostných systémov pre Smart Grid odporúčame získavať z podobne orientovaných systémov v odvetviach manažmentu

medicínskych dát a anonymizačných systémoch telekomunikačných spoločností. V prvom z menovaných odvetví je kvalitne vyriešený problém pomeru medzi bezpečnosťou dát a zachovaním ich štatistického potenciálu, v druhom odvetví existujú vhodné riešenia na bezpečnostný manažment konštantného toku dát zbieraných vo vysokej frekvencii. Telekomunikačné spoločnosti majú tiež kvalitné štandardizované postupy pre komunikáciu s tretími stranami. Konkrétne pre limitovanie informácií, ktoré zdroje osobných údajov o klientoch môžu sprostredkovať tretím stranám bez súhlasu spotrebiteľa (lebo pýtať sa spotrebiteľa o povolenie zakaždým, keď agregát s jeho dátami posúvame v systéme by bolo nerealizovateľné), pre rozhodnutia o tom, aké informácie o zákazníkoch môžu dať tretím stranám centrá zákazníckeho servisu a pravidlá oznamovacích povinností pre časti systému, ktoré disponujú dátami v takej forme, z ktorej ešte možno extrahovať osobné údaje užívateľov.

Bohužiaľ, aj v týchto oblastiach sa kontext, v ktorom bezpečnostné systémy pracujú, občas podcení. Dobrým príkladom je, ako telekomunikačné inštitúcie vnímajú telefónne číslo. Telefónne číslo samo o sebe nie je osobným údajom kým nie je verejne spojené s ďalšími údajmi o jeho majiteľovi. Na Slovensku neexistovala služba, ktorá by pre telefónne číslo vrátila meno vlastníka linky, kým ju nezačal poskytovať Slovak Telekom. Týmto krokom spoločnosť premenila telefónne čísla, pôvodne bežne dostupný údaj o svojich klientoch, na osobný údaj, ktorý je však stále bežne dostupný.

12. Zhodnotenie

Naším cieľom v tejto časti analýzy bolo urobiť prehľad problematiky bezpečnosti zberu, manažmentu, analýzy uchovávanía a distribúcie dát získaných v pripravovanom systéme Smart Grid. Následne testovať rôzne prístupy k anonymizovaniu týchto dát na sprostredkovanej vzorke. Technický popis tohto procesu je príliš komplexný pre potreby tohoto reportu, ktorý zaraďujeme ako výstup projektu, preto report obsahuje hlavné závery, odporúčania a špecifikácie, ktoré sme extrahovali počas našej práce.

Anonymizačné metódy sú veľmi užitočné nástroje, ale ako je to so všetkým v paradigme bezpečnosti a správe osobných údajov, kľúčom k tvorbe odolného systému je pochopenie kontextu, v ktorom bezpečnostný systém bude aplikovaný. Zložky aktívne integrované do systému musia plne porozumieť svojmu rozhodnutiu anonymizovať svoje dáta a aký má a nemá daný proces anonymizovania vzťah k ochrane osobných údajov užívateľov.

V procese prípravy realizácie projektu bude nutné identifikovať, ktoré údaje zo Smart Gridu potrebujú jednotlivé inštitúcie, za akým cieľom a komu tieto údaje budú sprostredkovať. Po tejto fáze bude možné identifikovať pomer medzi mierou anonymizovania a mierou ponechania štatistického potenciálu dát pre každú z inštitúcií a pre ich vzájomnú dátovú interakciu. V záujme vytvorenia kvalitného bezpečnostného systému je nutné si vyjasniť, aké sú potreby jednotlivých inštitúcií v systéme.

Na strane vývoja IS je nutné prepojiť teoretické poznatky o anonymizovaní dát s praxou v energetike. Posledný cieľ je síce náročný a dlhodobý, no nemôžeme ho nespomenúť. Z analýzy podobných bezpečnostných systémov sa ukazuje, že jedno z najslabších miest

systému je samotný užívateľ, ktorý si vypýta dáta o svojej spotrebe (na čo má nárok) a následne ich sprostredkuje tretej strane, ktorá má potenciál ich zneužiť. Vzdelávanie koncových užívateľov v oblasti bezpečného narábania s dátami o svojej spotrebe preto má potenciál eliminovať značné množstvo útokov.

13. Záverečné poznámky

V rámci zhodnotenia vyššie napísaného uvádzame ešte nasledovné poznámky:

- Zložky systému by mali mať prístup k údajom o užívateľoch v takých agregátoch, ktoré im umožňujú realizovať procesy, ktorými sú poverené (napr. optimalizácia a manažment siete, koordinácia s distribučným a prenosovým systémom, účtovné a fakturačné služby, tvorba predikcií do bilančného systému).
- Klienti by mali byť schopní sa rozhodnúť, či poskytnú svoje dáta o spotrebe tretím stranám za účelom iným ako sprostredkovanie elektrickej energie.
- Všetky zložky systému sú schopné pracovať s kvalitne extrahovanými agregátmi dát (vytvorenými špecificky pre ich účel) a neexistuje dôvod, prečo by mali mať k dispozícii surové dáta. Keď bude ktorákoľvek zložka pracovať so surovými dátami, bezpečnosť celého systému sa značne zníži.
- Zložky systému by nemali dávať agregované dáta k dispozícii tretím stranám bez súhlasu užívateľov.
- Autorizačný systém by mal byť postavený na princípe autorizačného procesu, v ktorom jednotlivé zložky systému a tretie strany dostávajú autorizáciu na prístup k dátam na limitovaný čas, a iba na základe ich požiadavky, v ktorej musí byť jasne uvedený účel, pre ktorý o dáta žiadajú.
- Pokiaľ je tretia strana autorizovaná na prístup k dátam, jej povinnosťou je tieto dáta chrániť, nedistribúovať a iniciovať na nich len operácie, ktoré boli popísané v žiadosti o autorizáciu. Nedodržanie týchto limitov by malo byť penalizované.

14. Otvorené otázky

Z uvedeného vyplývajú nasledovné otázky, ktoré v súčasnosti, ešte nie sú zodpovedané:

- Kto bude vlastníkom dát generovaných smartmetrami v sieti Smart Grid na Slovensku?
- Kto bude zodpovedný za správu systému ochrany osobných údajov a bezpečnostného manažmentu dát (napr. tvorby agregátov pre jednotlivé zložky systému, užívateľov a tretie strany)?

- Budú mať klienti možnosť odmietnuť inštaláciu smartmetru alebo moderovať, aké dáta v akom intervale sú o nich zbierané (v Kanade a niektorých štátoch U.S.A. takú možnosť majú)?
- Aké možnosti budú mať užívatelia, ak chcú ohlásiť problém so smartmetrom alebo únik informácií o ich osobe?
- Akú rolu budú v systéme hrať štátne bezpečnostné zložky Ministerstva vnútra SR a Ministerstva obrany SR (budú priamou súčasťou systému alebo budú figurovať ako tretia strana ktorá pre prístup k dátam potrebuje autorizáciu)?
- Aké dáta budú systémové zložky schopné odovzdať autorizovaným tretím stranám? Akú kontrolu bude mať koncový užívateľ nad odovzdávaním dát o jeho správaní zložkami systému tretím stranám?
- Za akých podmienok môže legislatíva zakázať poskytovať dáta tretím stranám či zložkám systému?
- Aké informácie bude nutné predložiť spolu so žiadosťou o autorizáciu prístupu k dátam?
- Akými nástrojmi budú užívatelia vzdelávaní o bezpečnej manipulácii s ich dátami?

6.1 Kontrola kvality dátových zdrojov a automatizácia anonymizačných procesov.

Predmetom tejto časti analýzy v rámci projektu je jeho upresnenie v kontexte požiadaviek na potrebnú kvalitu dátových zdrojov. Táto časť správy je formálnym výstupom analýzy kvality rôznych zdrojov dát, vybraných na základe ich prítomnosti v dátových štruktúrach systému súčasného sledovania pohybu výroby a prítomnosti spotreby elektrickej energie na Slovensku. Cieľom analýzy bolo kvalitatívne popísať verejné a uzavreté dátové zdroje identifikované v minulých reportoch ako zdroje vhodné k tvorbe inteligentných systémov, podporujúcich alebo nahradzujúcich dátové zdroje súčasného bilančného modelu estimácie odchýlky výroby a spotreby. Vybrané zdroje sme preto sledovali dlhšie obdobie, aby sme mohli definovať ich kvalitu.

6.1.1 Predpoklady a podmienky kvalitného zdroja dát

Dátové zdroje vhodné na dlhodobú predikciu pomocou metód strojového učenia, musia splniť niekoľko podmienok.

Aby mali dáta štatistickú relevantnosť, musí ich byť dostatočné množstvo. Podmienku objemu dát môžeme naplniť buď dostatočnou frekvenciou dát, alebo prítomnosťou podobných dát z iného zdroja (napr. na tréning inteligentného systému je možné použiť aj dáta o spotrebe z krajín, ak majú podobný profil). Dáta, ktoré sa dajú získať iba raz za rok stále môžu byť štatisticky významné, pokiaľ sa dajú získať pre viac štátov. Ďalšími dôležitými parametrami je presnosť dát a ich spoľahlivosť. Presnosť a spoľahlivosť dát, z ktorých vychádzame priamo ovplyvňuje presnosť a spoľahlivosť samotného modelu. Spoľahlivosť zdroja zahŕňa aj to, či sú dáta kompletne, či existujú časy, z ktorých sú neúplné, alebo ich v nich chýbajú celé zložky. Pokiaľ nie je zdroj úplne spoľahlivý, je možné použiť metódy na odhadnutie chýbajúcich informácií (extrapolácia dát) alebo model, ktorý si nevyžaduje vždy kompletnú dátovú vzorku.

Nakoľko by bola ideálna predikcia v reálnom čase, sú kladené vyššie podmienky aj na získavanie dát. Pokiaľ nie sú tieto podmienky splnené, predikcia stráca presnosť. Jedným z hlavných kritérií je pravidelná obnova dát, teda prístup k novým, aktuálnym dátam. Pokiaľ sú dáta obnovované s malou frekvenciou, nie sú vhodné na krátkodobú predikciu. Čím častejšie aktualizujeme dáta, tým presnejší výsledok dosiahneme v krátkodobých predikciách a tým lepšie máme vyhliadky aj na predpovedanie trendu dlhodobého vývoja. S predikciou v reálnom čase súvisí aj zber samotných dát (v reálnom čase). Čo sa týka zberu, môže byť rôzne náročné. Táto náročnosť je ale väčšinou jednorazová. Zahŕňa napísanie vhodného

programu, ktorý potom pracuje autonómne a vyžaduje minimum pozornosti. Zber dát z externých zdrojov preto odporúčame automatizovať. Ďalším kritériom je oneskorenie dát. Pokiaľ sú dáta obnovované každý deň, ale s mesačným oneskorením, tak je na ich základe nemožné realizovať krátkodobú predikciu v reálnom čase. Do analýz sledovania dlhodobého trendu vývoja sa však stále hodia.

Pri kontrole kvality jednotlivých zdrojov sme sa riadili nami zostaveným a štandardizovaným zoznamom kritérií, ktorého architektúru uvádzame nižšie. Následne si v skratke prejdeme výsledky sledovania kvality dátových zdrojov.

6.1.2 Kritéria

Nami zostavený a štandardizovaný zoznam kritérií, ktorý budeme ďalej používať na hodnotenie informačných a dátových zdrojov, je nasledovný:

1. Použitelnosť:

1. množstvo historických dát (pre nájdenie signifikantných závislostí musí byť dát dostatok)
2. frekvencia zberu dát
3. granularita dát podľa iných kritérií (zvyčajne poloha)
4. relevancia v energetike
5. množstvo chýb / chýbajúcich dátových bodov

2. Postupnosť:

1. frekvencia obnovy dát
2. náročnosť získavania dát

6.1.3 Naše získané energetické dáta

1. Použitelnosť

1. Dáta sú za obdobie 01.07.2011 – 14.05.2015
2. Meranie sa uskutočňujú každých 15 minút pri SMART metroch, pri ostatných len mesačné alebo ročné.
3. Dátová vzorka pozostáva z miest so 75 SMART metrov a z 4506 ostatných miest.

4. Tieto dáta sú pre predikciu veľmi relevantné, nakoľko zaznamenávajú pravidelnosti a mikro ale aj makro zmeny v historických odberoch počas oka.
5. Dáta nie sú úplne kompletne, nakoľko pre niektoré odberné miesta chýbajú merania. Okrem toho sú dáta kvalitné.

2. Dostupnosť

1. Pri SMART metroch je frekvencia obnovy každých 15 minút, ostatné dáta sú aktualizované mesačne, alebo ročne.
2. Náročnosť získavania závisí od dohody s OKTE.

Vzorka nami získaných energetických dát je kvalitná, no aby sme mohli testovať všetky predikčné modely identifikované v reporte č. 1, potrebovali by sme ju doplniť o tieto parametre:

- **Veľkosť vzorky:** všeobecne platí - čím viac dát budeme mať k dispozícii, tým lepšie. Súčasná štruktúra ukazuje, že na komplexnejšie analýzy by sme (ideálne) mali disponovať aspoň dvojročnými dátami (730 dní). Takáto vzorka by nám umožnila oveľa väčšie množstvo korelácií a dala by nám voľnosť pri spájaní dát o spotrebe a výrobe elektrickej energie so socio-demografickými, spoločensko-ekonomickými a hydrometeorologickými trendmi na Slovensku.
- *Stratifikácia vzorky: Radi by sme disponovali dátami z väčšieho množstva odberných miest a to hlavne typu A (máme 68 meraní) a typu B (máme 7 meraní).*
- *Chýbajúce dáta:*
 - *V tabuľke "BD_SUSTAVY" máme len typ sústavy a bez informácie, o ktorú sústavu sa jedná.*
 - *Nemáme príklady dát o externých sústavách a priamych vedeniach.*
 - *Chýbajú dáta pre triedy TDO 1224, 1227, 1178.*
 - *Nemáme merania pre 3 odberné miesta typu A (8901997, 8892074, 8902072) popísané v databáze.*
 - *Chýbajú nám dáta pre triedy TDO 1224, 1227, 1178.*
- *vyrobna_id: V tabuľke BD_OOM nevieme presne čo znamená premenná "vyrobna_id", nevieme si ju s ničím spojiť a nevieme na čo sa používa.*

6.1.4 Štatistický úrad

Dáta zo štátnych registrov (napr. register organizácií, poľnohospodársky register, register sčítacích obvodov, register priestorových jednotiek, register subjektov zahraničného obchodu a register ubytovacích zariadení), z databázy regionálnej štatistiky a zo Štatistického úradu Slovenskej republiky majú veľkú hodnotu. Keďže ide o socio-demografické dáta, ktoré sa v priebehu času veľmi nemenia, tak otázky o množstve historických dát a frekvencii sú irelevantné. Granularita dát ide až na úroveň obce a konkrétne typy dát sú popísané v reporte č. 2. Dáta sa dajú získať cez webové rozhranie, ale zber treba automatizovať. Dáta budú mať cenu najmä pri vytváraní spotrebných profilov užívateľov. Do profilov dodávajú socio-demografické premenné a charakteristiky infraštruktúry, ktoré poslúžia na segmentáciu užívateľov. Identifikácia vzťahu medzi týmito premennými a konzumným správaním užívateľov nám môže do značnej miery pomôcť optimalizovať výrobu a dodávku elektrickej energie. Pri vývoji a nasadzovaní predikčného modelu v kontexte SmartGrid odporúčame bližšiu spoluprácu s týmito inštitúciami.

6.1.5 Slovenský hydro-meteorologický ústav

1. Použitelnosť

1. Množstvo historických údajov pri slovenskom hydrometeorologickom ústave je dostatočné, SHMÚ disponuje značnou historickou databázou.
2. Dáta sú zbierané každú hodinu, alebo každý deň.
3. Granularita sa pre rôzne stanice v SR líši, v prípade predpovedí rôzne hustá mriežka (medzi 1 km - 5 km).
4. Nakoľko počasie ovplyvňuje produkciu alternatívnych zdrojov energie, prenosovú sústavu, ako aj každodenný život ľudí, považujeme tie dáta za vysoko relevantné.
5. Očakávame vysokú kvalitu a málo chýbajúcich údajov, nakoľko sú tieto zbery aj komerčne ponúkané.

2. Dostupnosť

1. Predpoveď sa prepočítava každé 3 hodiny.
2. Náročnosť získavania závisí od dohody so SHMÚ.

6.1.6 Iné open dáta zdroje

1.1.1 Meteo dáta z FR

1. Použitelnosť

3. Táto dátová vzorka má dvojtýždňový cyklus.
4. Dáta sú zbierané každé 4 hodiny.
5. Celkovo ide o 63 staníc na území FR.
6. Aj keď sú dáta z oblasti mimo Slovenska, je možné že pomôžu odhaliť spoločné závislosti. Hodnotíme ich preto ako stredne relevantné.
7. Dáta veľmi kvalitné a takmer bez chýb.

2. Dostupnosť

1. Dáta sa obnovujú každé 4 hodiny.
2. Nakoľko sú na stránke len dáta za posledné 2 týždne, treba ich sťahovať dlhšie časové obdobie, aby sme získali väčšiu vzorku.

Meteo dáta z USA

1. Použitelnosť

1. Dĺžka histórie uchovávaní dát závisí od konkrétnej stanice, ale väčšinou obsahuje aspoň 3 roky záznamov.
2. Dáta sú zbierané každých 10 minút, alebo každý deň (10 min - deň).
3. Celkovo ide o 33 staníc v USA.
4. Podobne ako dáta z Francúzska, aj tieto nie sú priamo spojené s energetikou na Slovensku, môžu ale pomôcť a preto ich hodnotíme ako stredne relevantné.
5. Zdroj je veľmi spoľahlivý, bez chýbajúcich dátových bodov.

2. Dostupnosť

6. Obnovovanie databáz závisí od stanice, niektoré sa neobnovujú pravidelne, niektoré každý deň.
7. Je možné priamo stiahnuť aktualizovaný súbor s dátami.

Veľká Británia energetika

(http://data.gov.uk/dataset/energy_consumption_in_the_uk)

1. Použitelnosť

1. Dáta z obdobia od roku 2007 do roku 2013
2. Sú zbierané ročne.

3. Granularita závisí od typu dát.
4. Považujeme ich za stredne relevantné, nakoľko sa nejedná o dáta zo Slovenska.
5. Nevyskytujú sa v nich chyby alebo absentujúce úseky.

2. Dostupnosť

6. Nové dáta pribúdajú raz za rok.
7. Dáta je možné priamo stiahnuť v súbore.

EEX dáta - Nemecko

<http://www.transparency.eex.com/en/Statutory%20Publication%20Requirements%20of%20the%20Transmission%20System%20Operators/Power%20generation/Actual%20wind%20power%20generation>

1. Použiteľnosť

1. Tieto dáta sú zbierané od roku 2010.
2. Merania sú uskutočňované každú hodinu.
3. Podľa bilančnej skupiny (5 skupín).
4. Vysoko relevantné (pohyby v Nemecku sa dotýkajú územia SR).
5. Nevyskytujú sa v nich chyby alebo absentujúce úseky.

2. Dostupnosť

6. Dáta tečú do zdroja v reálnom čase a je k nim okamžitý prístup.
7. Stránka dáta zobrazuje v podobe flashovej aplikácie, je nutné napísať a používať špeciálny program na sťahovanie.

6.1.7 Automatizácia procesu anonymizácie dát v SmartGrid

V predchádzajúcich častiach sme uviedli, ako by mal vyzeráť bezpečnostný systém, postavený na princípoch komplexnej anonymizácie dát, po zavedení SmartGrid. Takýto systém má za cieľ minimalizovať nebezpečenstvo zneužitia osobných údajov koncových užívateľov SmartGrid a stále zachovať analytickú hodnotu dát pre špecifické inštitúcie s rôznymi cieľmi. V tejto časti by sme chceli krátko popísať požiadavky a možnosti automatizácie procesov anonymizácie.

V systéme anonymizácie bude dôležitá automatizácia troch typov procesov.

Prvým bude automatizácia agregovania surových dát zo SmartGrid na metadáta vo forme dostatočnej pre potreby konkrétnej inštitúcie, ktorá si ich od systému vyžiadala. Ako sme predtým spomenuli, musíme zaručiť aby inštitúcie mali prístup k údajom o užívateľoch

SmartGrid v takých agregátoch, ktoré im umožňujú realizovať procesy, ktorými sú poverené (napr. optimalizácia a manažment siete, koordinácia s distribučným a prenosovým systémom, účtovné a fakturačné služby, tvorba predikcií do bilančného systému).

Druhým typom procesu, ktorý bude musieť byť automatizovaný je štandardizovaný autorizačný systém. Z pohľadu minimalizovania rizika informačného zneužitia, by prístup k agregátom mal byť postavený na princípe autorizačného procesu, v ktorom jednotlivé zložky systému a tretie strany dostávajú autorizáciu. Podanie a vybavenie autorizácie by malo byť automatizované.

Tretím a zároveň najdiskutabilnejším procesom, z ktorého automatizácie by bezpečnosť celého systému značne profitovala, je automatizácia výmazu dát. Všetky inštitúcie, ktoré získajú autorizáciu, dostanú prístup k dátam na limitovaný čas, na základe ich požiadavky, v ktorej musí byť jasne uvedený účel, pre ktorý o dáta žiadajú. Po dosiahnutí cieľa, za ktorým boli dáta vyžiadané (napr. dátový agregát už pre danú inštitúciu ďalej nemá relevantnú hodnotu) musia byť tieto dátové vzorky aj všetky ich kópie vymazané. Existencia nevymazaných irelevantných dát držaných jednotlivými inštitúciami by značne zvýšila riziko úniku osobných informácií, či iného zneužitia týchto dát. Preto automatizácia procesu likvidácie použitých agregátov je zmysluplnou požiadavkou na systém. Diskutabilný je proces realizácie. Je tu viacero variant od sledovania vzoriek priradením identifikátora, ktorého platnosť vyprší (napr. všetky agregáty s daným identifikátorom budú vymazané alebo uzamknuté) až po špecifickú vlastnosť rozhrania, v ktorom sa ku anonymizovaným agregátom bude môcť pristupovať.

Pri špecifikácii jej automatizácie anonymizačných procesov do veľkej miery záleží na nastavení operatívy a administratívneho aparátu, v ktorom sa systém nasadí. Konkrétne na kompetenciách a právach inštitúcií, ktoré budú spravovať systém, na motiváciách a požiadavkách inštitúcií, ktoré budú systém využívať (viď. Otázky na konci reportu č. 4). Tak ako výber samotných nástrojov anonymizácie dát, aj ich automatizácia má isté bezpečnostné a realizačné kritériá. Automatizovať popísané procesy anonymizácie je možné, ak bude charakter vstupných dát nemenný. Je nevyhnutné aby od koncových užívateľov SmartGrid prichádzali dáta o spotrebe a výrobe elektrickej energie v štandardizovanom sete premenných. Pre automatizáciu sprostredkovania agregovaných výstupov je taktiež nevyhnutné, aby ciele analýz konkrétnych inštitúcií zostali nemenné. Automatizácia autentifikačného systému a výmazu dát ďalej vyžaduje, aby zostalo nemenné rozloženie inštitúcií v systéme, v zmysle ich práv a povinností (napr. kto má právo kontrolovať systém anonymizácie, kto ho udržiava po softvérovej a hardvérovej stránke). Pokiaľ prebehne zmena v niektorom z troch menovaných stavov, nasadený systém bude musieť prejsť úpravou jeho nastavení. Inštitúcia zodpovedná za manažment systému by mala kontinuálne sledovať trendy v dátovej bezpečnosti a pravidelne revidovať systém anonymizácie aj automatizácie manažmentu dát. Posledným kritériom bezpečnosti je teda aktualizácia systému.

6.1.8 Záver

Naším cieľom bolo urobiť si prehľad o kvalite relevantných zdrojov dát, ktoré sú potencionálne k dispozícii buď verejne alebo v uzavretej forme. Zdroje, ich aktualizáciu, chybovosť, presnosť a spoľahlivosť v čase sme sledovali niekoľko mesiacov. Naša analýza ukázala nedostatky niektorých zdrojov a pripravila nám otázky, ktoré si musíme zodpovedať (ideálne v diskusii s partnermi projektu). Vo všeobecnosti je však väčšina dátových zdrojov dostatočne kvalitná a obsiahla, teda vhodná na ďalšie použitie. Čo sa týka dlhodobého získavania dát z externých jednotlivých zdrojov, nebudú v tom výrazné technické problémy. Bude potrebné komunikovať s inštitúciami ako SHMU a Štatistický úrad SR.

Druhým cieľom bolo informovať o možnostiach automatizácie systému anonymizácie a zberu dát z externých zdrojov. Popísali sme zložky systému manažmentu dát, ktoré z bezpečnostného a praktického hľadiska môžu byť a mali by byť automatizované (napr. dátový zber, agregácia dát a ich anonymizácia, systém udeľovania autorizácií a systém výmazu nadbytočných dát).

7 Štatistické metódy

Najjednoduchším prípadom dát sú jednorozmerné dáta, ktoré často reprezentujú jeden stĺpec tabuľky databázy. Kontrolou takýchto dát je možno odhaliť množstvo chybných dát, ktoré sú založené na typografickej chybe alebo dáta obsahujú extrémne (v podstate nereálne) alebo nedovolené hodnoty. Na zvýšenie kvality takýchto dáta sa často aplikujú štatistické metódy. V prípade dát, ktoré sú uložené v tabuľkách s rôznymi atribútmi, t.j. majú svoju štruktúru, hovoríme o viacrozmerných dátach. Preto aj pri návrhu metód čistenia týchto dát je nutné aplikovať komplexnejšie prístupy.

7.1 Jednorozmerné dáta

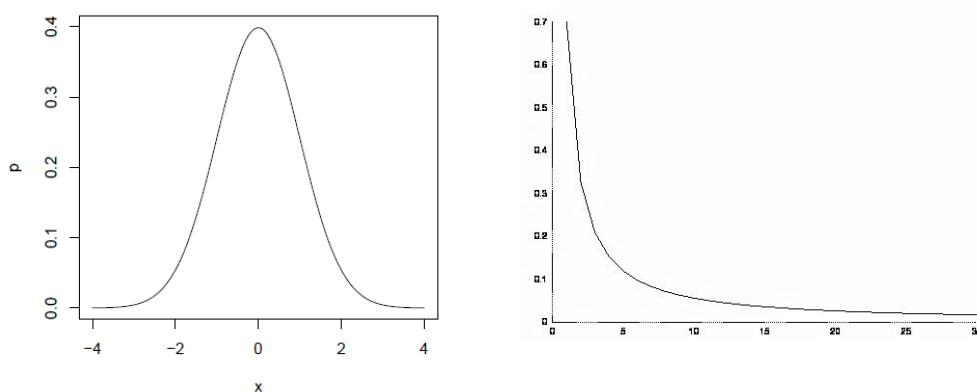
Štatistické metódy skúmajú vzorku jednorozmerných dát z pohľadu procesu, ktorý tieto dáta vygeneroval. Základnou úlohou je nájsť jednoduchý štatistický opis alebo model, ktorý bude reprezentovať tento proces a teda aj vzorku dát. Tvorba algoritmických modelov sa nezaobíde bez primárnej kontroly kvality získaných dát. Ako už bolo spomenuté, je dôležité, aby dáta neobsahovali evidentné chyby (napr. teplota +100°C, podozrivo vysoká spotreba). Šetrenie takýchto chýb môže prebiehať dvoma spôsobmi. Ak stavíme model na citlivých dátach, v ktorých je minimálna chybovosť esenciálna alebo na malých vzorkách dát, snažíme sa takéto infekty vyhľadať a po jednom ošetriť. V prípade, že model stavíme na dostatočne veľkých vzorkách dát, môžeme takéto extrémne hodnoty ošetriť nastavením hodnotovej hranice, ktorú získame pomocou indikátorov deskriptívnej štatistiky – priemeru, mediánu, maxima, minima a štandardnej odchýlky. Neštandardné zmeny tvaru dát, ktoré je potrebné odfiltrovať, je možné detekovať prostredníctvom validačnej funkcie. Ako štandardný štatistický model sa používa často lineárna regresia, ktorá využíva pravdepodobnostný prístup k dátam (viacrozmerným dátam). Tento model sa dá taktiež využiť na extrapoláciu dát, ktoré chýbajú vo vzorke v dôsledku chyby pri zbere dát alebo dáta je potrebné predpovedať do budúcnosti v rámci potrebných prognóz.

Medzi základné štatistické metriky, ktoré je možné aplikovať pri hľadaní odľahlých hodnôt (extrémnych hodnôt, ktoré by sa nemali vyskytovať vo vzorke dát), patrí stredná (priemerná) hodnota a disperzia. Stredná hodnota, ktorá sa nemusí vyskytovať v samotných dátach, reprezentuje určitý matematický stred hodnôt dát. Disperzia udáva relatívnu vzdialenosť od tohto stredu. Pri štatistickej analýze sú tieto veličiny previazané s nájdením vhodného rozdelenia (hustotou rozdelenia pravdepodobností), ktoré bude reprezentovať dáta a udáva, aká je pravdepodobnosť výskytu každej hodnoty dát. Najznámejším je normálne rozdelenie, ktoré je definované strednou hodnotou a smerodajnou odchýlkou. Smerodajná (štandardná) odchýlka (ako odmocnina rozptylu) hovorí, ako úzko alebo široko sú rozmiestnené hodnoty dát.

Ak sa sústredíme na hľadanie odľahlých hodnôt (outliers), potom potrebujeme nájsť stred a disperziu hodnôt všetkých dát a potom hodnoty, ktoré sú príliš ďaleko od stredu (napr. dve smerodajné odchýlky), aby ich bolo možné vyhlásiť za odľahlé hodnoty. Problém ale vzniká, keď tieto odľahlé hodnoty sa zakomponujú do výpočtu týchto štatistických veličín a výrazne posunú stred dát niektorým smerom. Tento efekt spôsobujú najmä tie hodnoty, ktoré sa odlišujú rádovo od ostatných (chyba pri vkladaní, chybná jednotka hodnoty a pod.). Preto je vhodné nájsť spôsob eliminácie tohto typu hodnôt pred výpočtom spomínaných štatistických veličín. Jedným z použiteľných prístupov, ak uvažujeme veľké objemy dát (t.j. manuálna oprava je nemožná) sú robustné štatistiky založené na výpočte mediánu [8]. Medián na rozdiel od stredu je výrazne odolnejší voči vplyvu príliš extrémnych odľahlých hodnôt. Medián sa snaží rozdeliť dáta na dve takmer rovnaké skupiny. Ak vylúčime odľahlé hodnoty podľa mediánu a aplikujeme výpočet týchto štatistických veličín na „opravnú“ vzorku dát, tak zistíme, že nový stred sa výrazne posunie, zatiaľ čo medián iba minimálne. Miera odľahlých hodnôt (dát) musí byť príliš veľká v dátach, aby to výraznejšie ovplyvnilo hodnotu mediánu.

Inou robustnou štatistikou, ktorú je možné použiť na riešenie tohto problému, je tzv. $k\%$ trimmed mean. Tento prístup vynecháva pri výpočte výsledného stredu dát k percent hodnôt dát s najvyššími a najnižšími hodnotami. Týmto spôsobom je možné zamedziť nežiaduci vplyv čudných (extrémnych) hodnôt. Parameter k je voliteľný, často volený v intervale 5 až 25 percent. Existuje aj variant, ktorý extrémne hodnoty nezahadzuje, ale nahrádza najväčšou (alebo najmenšou) zahrnutou hodnotou.

Aritmetický priemer nie je vhodný pre všetky typy dát. V niektorých prípadoch je vhodné využiť geometrický alebo harmonický priemer (napr. pre priemerné rýchlosti). Oba tieto priemery sú ale citlivé na hodnoty blízke nule a nie sú teda príliš robustné. Na zvýšenie ich robustnosti je opäť možné použiť metódu krátenia dát. V prípade, že nemáme informáciu o charakteristike dát, odporúča sa vypočítať všetky tri typy priemerov, prípadne ich robustné verzie. Aby sme dosiahli robustnú disperziu, je vhodné pri jej výpočte využiť robustný priemer, t.j. medián. Tým sa získa robustná metrika disperzie označovaná ako MAD (Median Absolute Deviation) [8].



Obr. 4 Graf normálneho a Zipfovho rozdelenia

Ako sme už spomínali, základom štatistického prístupu je pre danú bázu dát (hodnôt) nájsť vhodné rozdelenie. Normálneho (gaussovské) rozdelenie (Obr. 4 vľavo) je častou reprezentáciou dát, ale môžu sa vyskytnúť aj iné rozdelenia. Multimodálne rozdelenia sú charakterizované viacerými špičkami. Takéto rozdelenie môže ale byť reprezentované ako zlúčenie viacerých gaussovských rozdelení do jedného. Iným príkladom je Zipfovo (Zipfian, Yule, Zeta) rozdelenie [9], ktoré je špecifické pre dáta, ktoré sú ovplyvnené obľúbenosťou. Príkladom je početnosť výskytov slov anglického jazyka. Je charakterizované údajmi, ktoré sa vyskytujú veľmi zriedka, a ktoré tvoria dlhý chvost krivky (Obr. 4 vpravo).

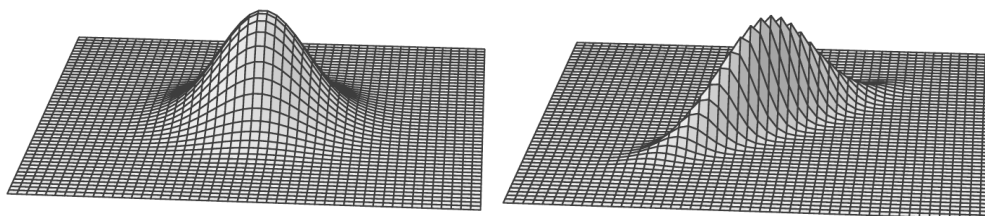
Pri aplikácii štatistických metód čistenia dát je potrebné na začiatku zistiť, ktoré rozdelenie najlepšie modeluje bázu dát. Často sa to realizuje tak, že sa nakreslí histogram dát a ten sa preloží krivkou najbližšieho rozdelenia. V prípade, že dáta nezodpovedajú normálnemu rozdeleniu, je možné aplikovať niekoľko prístupov. Jedna možnosť je tá, že sa napriek tomu aplikuje normálne rozdelenie a ráta sa s určitými nedokonalosťami odhadu odľahlých dát. Vtedy je potrebný zásah analytika. Druhá možnosť je aplikovať rozdelenie, ktoré sa očakáva. Rozdiel medzi modelom a dátami potom reprezentujú dáta volané ako reziduá, ktoré zvyčajne majú charakter normálneho rozdelenia a aplikuje sa na ne vyššie opísaný postup. Bázu dát je možné aj rozdeliť na menšie celky tak, aby sledovali normálne rozdelenie. Alebo sa použije niektorý z navrhnutých neparametrických prístupov. Celkovo sa dá ale zosumarizovať, že pre efektívny prístup k tomuto problému, je potrebný zásah analytika, ktorý v niektorých fázach rozhodne o ďalšom postupe (výber rozdelenia, parametrov a pod.).

7. 2 Viacrozmerné dáta

Viacrozmerné dáta sú uchovávané v tabuľkách databáz v rámci viacerých stĺpcov a môžu napr. reprezentovať priebeh spotreby pre množstvo domácností v danom regióne alebo ekonomické ukazovatele pre rôzne krajiny a pod. Zvyčajne existuje určitá korelácia alebo podobné správanie sa hodnôt medzi dátami v rôznych stĺpcoch. Metódy používané pre spracovanie viacrozmerných dát vychádzajú z tých pre jednorozmerné, ale hľadanie vzájomných vzťahov medzi dátami vo viacerých rozmeroch je komplikovanejšie a časovo náročnejšie.

Aj pri viacrozmerných dátach sa hľadá stred (priemer) v dátach ako aj rozdelenie, ktoré by reprezentovalo tieto dáta. Napríklad v prípade dvojrozmerných dát po aplikovaní normálneho rozdelenia dostávame trojrozmerný graf (Obr. 5), ktorý udáva pravdepodobnosť výskytu jednotlivých dvojíc hodnôt (z oboch stĺpcov). V prípade súmernosti tohto grafu dostávame jednu hodnotu stredu a smerodajnej odchýlky dát a dvojice s rovnakou pravdepodobnosťou výskytu by sa nachádzali na kružnici. Analogicky by to platilo aj pre viacrozmerné dáta. V prípade, keď ale toto normálne rozdelenie nebude spĺňať vlastnosť súmernosti (Obr. 5 vpravo), dostaneme viacrozmerný stred dát a kovariančnú maticu, ktorá bude reprezentovať rozptyly (druhé mocniny smerodajných odchýlok) pre jednotlivé premenné (dáta v stĺpcoch) a korelácie medzi jednotlivými dvojicami. To znamená, že

z pohľadu čistenia dát je pre nás podstatný stred dát (reprezentovaný viacrozmerným priemerom dát) a disperziou, ktorá je reprezentovaná kovariančnou maticou. V tomto prípade teda hľadáme, ako ďaleko sa nachádza daný bod od stredu vo všetkých smeroch s ohľadom na disperziu danou kovariančnou maticou. Na tento účel je možné požiť metriku volanú Mahalanobisova hĺbka, ktorá sa ráta z inverznej kovariančnej matice.



Obr. 5 Príklad súmerného a nesúmerného dvojrozmerného normálneho rozdelenia

Rovnako ako pri jednorozmerných dátach, tak aj v tomto prípade platí, že jedna extrémna hodnota v dátach výrazne ovplyvní výslednú hodnotu vypočítaného stredu ako aj kovariančnú maticu (maskovanie). Preto aj pri viacrozmerných dátach je potrebné prejsť na robustné metriky. Ako použiteľné techniky možno spomenúť:

- Iteratívne mazanie Mahalanobisových bodov – bod s najväčšou vzdialenosťou od stredu sa vymaže z množiny dát a opätovne sa vyráta stred a disperzia. Tento proces sa postupne opakuje. Potrebne je ale rozhodnúť, kedy sa tento proces ukončí. Jednou z možností je proces ukončiť, keď sa stred a disperzia mení už málo.
- Robustné estimátory pre strednú hodnotu a kovarianciu dát – je niekoľko prístupov, ktoré poskytujú optimálne estimátory pre stred a disperziu dát. Mahalanobisové vzdialenosti [10] od robustných stredov dát môžu byť vypočítané pre všetky body (hodnoty) s použitím robustnej disperzie. Často používaná metóda využíva MCD (Minimum Covariance Determinant), ktorá vypočíta podskupinu hodnôt z dát, ktorá minimalizuje determinant kovariančnej matice. Potom stredná hodnota a kovarianca týchto hodnôt sa použije ako stred a disperzia pre všetky dáta. Ďalšie metódy využívajú napr. MVE (Minimum Volume Ellipsoid) alebo OGK (Orthogonalized Gnanadesikan-Kettering).
- Techniky založené na výpočte hĺbok – pri jednorozmerných dátach je možné usporiadať hodnoty dát do častí (úrovní) s využitím štatistických prístupov založených na výpočte mediánu a tzv. kvantilov. Techniky založené na výpočte hĺbok sa snažia aplikovať tento prístup do viacrozmerných dát. Potom hodnoty, ktoré by sa mali vylúčiť z bázy dát, by mali nulovú hĺbku. Systém hĺbok sa môže taktiež aplikovať aj pri krátení viacrozmerných dát. Hodnoty s hĺbkou menšou ako stanovená hranica sú vylúčené zo súboru dát. Z dát, ktoré ostali, sa vypočíta „orezaný“ stred a kovariančná matica dát a z nich Mahalanobisové metriky na hľadanie odľahlých dát. Detailnejšie informácie o niektorých metódach možno nájsť v [11].

- Úprava dát robustnými estimátormi – stred dát sa vypočíta pomocou robustného prístupu (napr. mediánu) a potom sa vypočíta vzdialenosť hodnôt od tohto stredu v štandardizovaných jednotkách disperzie. Každá premenná (reprezentujúca jeden súbor dát; t.j. stĺpec/rozmer) je vylepšená (tzv. robustifikácia) jednorozmerným estimátorom. Čiže na výpočet stredu dát sa použije medián, ktorý má svoju reprezentáciu pre každú premennú (súbor dát).

V niektorých prípadoch je ťažké nájsť štatistické rozdelenie, ktoré by vystihovalo súbor dát či už jedno alebo viacrozmerného charakteru. To znamená, že je komplikované nájsť parametrický model, ktorý by následne mohol napomôcť pri hľadaní čudných hodnôt. Niektorí autori navrhli metódy, ktoré nevyužívajú tento typ štatistického prístupu a v podstate nevyužívajú hľadanie modelu pre dáta, ale vychádzajú priamo zo vzdialeností. Princíp je v určení vzdialeností medzi jednotlivými párami hodnôt a potom odľahlé hodnoty budú tie, ktoré majú najväčšiu vzdialenosť od svojich susedov. Ako príklad možno uviesť spôsob, kde bod (hodnotu dát) definuje ako $DB(k,D)$ odľahlý, ak najviac k bodov leží bližšie ako vzdialenosť D . Tieto metódy ale často sú neefektívne, keď hustota bodov v rôznych častiach dát (klastroch), sa dosť odlišuje. Jedným z problémov, ktorým tieto prístupy trpia tým, že nevyužívajú štatistickú analýzu, je aj to, že vždy nájdú neaké odľahlé hodnoty v dátach, dokonca aj keď tam už nie sú. Problém môže nastať aj pri zmene jednotky dát.

7.3 Údaje časových radov

Údaje časových radov (time series) reprezentujú hodnoty premennej, ktoré sú odoberané v pravidelných časových intervaloch (roky, mesiace, dni, hodiny a pod.). Je to teda chronologická postupnosť čísel, ktorá sa odlišuje od bežných numerických údajov v korelácii medzi susednými odoberanými hodnotami. Takéto údaje reprezentujú rôzne ukazovatele ako aj dáta merané v energetike. Údaje časových radov občas vykazujú sezónnu závislosť. Napr. spotreba energie v domácnostiach je väčšia večer. Odľahlé hodnoty môžu byť najčastejšie aditívne alebo inovačné. Aditívne sa vyskytujú izolovane a po nich sa hodnoty okamžite vracajú do normálneho správania. Obyčajne sú to chyby v dátach. Inovačné sú nesprávne hodnoty, ktoré kontaminujú hodnoty, ktoré po nich nasledujú a ich účinok postupne zoslabuje.

Typickým príkladom metódy, ktorá sa používa proti odľahlým hodnotám v časových radoch je vyhladzovanie prostredníctvom okien. Na zvolenom okne dát sa vypočíta tzv. kĺzavý priemer. Výsledný časový rad sa použije na porovnanie s originálnymi hodnotami, hodnota, ktorá je príliš vzdialená od kĺzavého priemeru, sa vyhlási za čudnú hodnotu. Ako príliš musí byť vzdialená, sa odvinie od smerodajnej odchýlky. Rozptyl časových radov sa považuje za konštantný a môže byť rátaný na oknách dát relevantnej dĺžky. Je možné vypočítať smerodajnú odchýlku na jednotlivých oknách a sledovať tak trend zmien rozptylu. Zovšeobecnením tohto prístupu je však aplikovať flexibilné váhy na vplyv minulých dát, pričom sa tieto váhy postupne zmenšujú s rastúcou vzdialenosťou v čase. Typickým príkladom je schéma EWMA (Exponentially Weighted Moving Average), ktorá zmenšuje

exponenciálnym spôsobom váhy v čase [12]. Podobným spôsobom je možné rátať aj kľzavý rozptyl a sledovať jeho zmeny v čase. Je určite zaujímavé sledovať vývoj „oknovaných“ a exponenciálne váhovaných kľzavých priemerov pri hľadaní čudných hodnôt.

Ako už bolo spomínané, pri používaní priemerov (stredných hodnôt) a rozptylov, už jedna extrémna hodnota môže výrazne negatívne ovplyvniť účinnosť týchto prístupov. Preto sa odporúča aplikovať okná na časové rady hodnôt a používať „oknované“ mediány a MAD hodnoty. Je však požadovaný ešte ďalší výskum v tejto oblasti na nájdenie účinných prístupov. Populárny model pre časové rady je ARIMA (AutoRegressive Integrated Moving Average), ktorý sa ale skôr používa na predikciu dát.

7.4 Vzorkovanie dát

Medzi ďalšie alternatívne metódy hľadania odľahlých dát možno zaradiť metódu vzorkovania dát. Táto metóda využíva princíp riadeného opakujúceho sa spracovania vzoriek dát za účelom určiť sumárnu štatistiku vzorky a následnou kombináciou vzoriek získať finálnu vlastnosť celej bázy dát. Zriedkavé extrémne hodnoty sa objavujú iba v málo vzorkách a preto neovplyvnia výrazne finálny estimátor.

Pre menšie súbory dát je možné použiť prístup, ktorý vyberie vzorku z pôvodného súboru dát a vypočíta sledovanú štatistickú vlastnosť. Tento proces sa opakuje veľmi veľký počet krát (tisíc, milión krát) a z jednotlivých výsledkov sa potom vytvorí sumárna štatistika reprezentovaná tzv. výberovým rozdelením. Iný prístup opakuje výpočet pre daný súbor dát, pričom sa zakaždým vynechá niektorá hodnota. Zo získaných výsledkov sa opäť vypočíta výberové rozdelenie. Z takéhoto rozdelenia je potom možné odhadnúť priemer a rozptyl na hľadanie odľahlých hodnôt. Tento prístup opakovaného výpočtu nad vzorkami dát môže byť použitý aj pre tvorbu zložitejších sumárnych štatistík, ktoré môžu byť previazané s lineárnou regresiou na hľadanie odľahlých hodnôt [13].

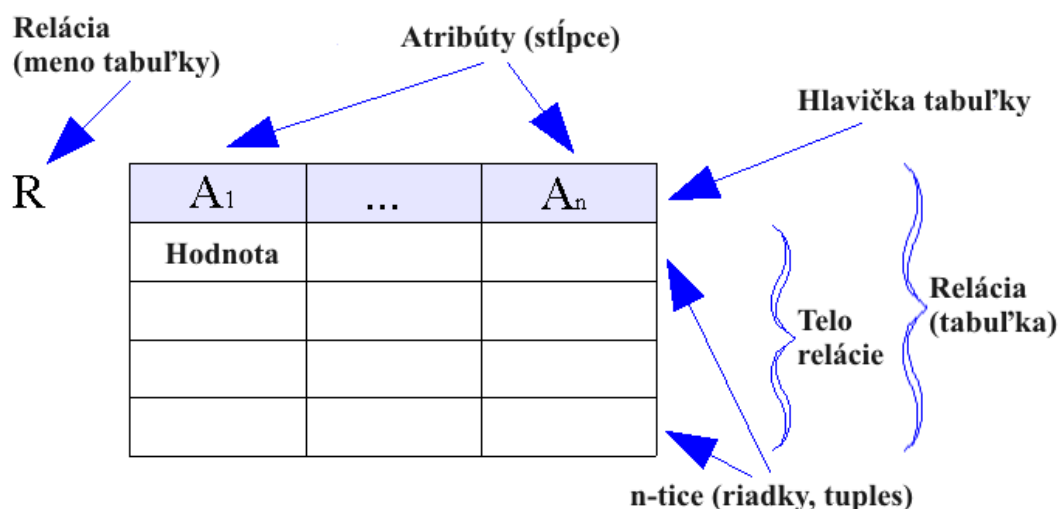
8 Logické metódy

Logické metódy alebo metódy, ktoré sú založené na obmedzeniach (obmedzujúcich podmienkach), sa pokúšajú vyvinúť logický model, ktorý bude reprezentovať dáta a akékoľvek porušenie tohto modelu bude signalizovať nekonzistentnosť (chyby) dát. Základom je teda nájsť obmedzujúce podmienky integrity dát (data integrity constraints), aby bolo následne možné zlepšiť kvalitu najčastejšie štruktúrovaných dát v databázach. Na tento účel sa používajú rôzne typy kľúčových a kontrolných obmedzení, funkčné závislosti a ich rôzne podmienené verzie.

Vzhľadom na to, že údaje sú získané a integrované z rozličných zdrojov, pre ich veľké objemy a meniacu sa povahu je v podstate nemožné manuálne špecifikovať pravidlá pre zabezpečenie požadovanej kvality dát. Veľké objemy dát (big data) na druhej strane prinášajú také množstvo dát, že by bolo možné aplikovať pravidlo: nechať hovoriť dáta samých za seba. To znamená, naučiť sa automatizovane príslušné pravidlá priamo z týchto dát. Jednotlivé pravidlá ale musia byť naučené efektívne.

8.1 Funkčné závislosti dát

Závislosti medzi dátami v databázach hrajú významnú úlohu nielen pri zlepšení kvality dát ale aj pri návrhu samotnej databázy a reprezentácii známych a nových znalostí. Základom je teda proces extrakcie (objavovanie) všetkých závislostí (vzťahov), ktoré budú zodpovedať existujúcim dátam. Hoci oblasť objavovania nových znalostí z dát prináša nové možnosti vo výskume, my sa budeme zaujímať najmä hľadaním závislostí, ktoré prispievajú k hľadaniu nesprávnych dát a ich následnej úprave či oprave.



Obr. 6 Základná schéma (štruktúra) analyzovanej bázy dát a použitá terminológia

Pri definícii funkčných závislostí (FD, functional dependency) budeme vychádzať z Obr. 6, kde je znázornená základná štruktúra bázy dát, nad ktorou budeme hľadať jednotlivé

pravidlá. Obrázok poskytuje aj základnú terminológiu, ktorú budeme ďalej používať. Relačný model databázy združuje dáta do tabuliek (relácií), ktoré sú tvorené jednotlivými (neusporiadanými) riadkami – n-ticami (tuples). Každá tabuľka pozostáva z neusporiadaných stĺpcov (atribútov), ktoré obsahujú hodnoty pre jednotlivé riadky (n-tice), a ktoré sú dané svojim názvom, typom a rozsahom (čo predstavuje ich doménu). Inak povedané, tabuľka je daná svojimi záznamami, ktoré reprezentujú jednotlivé n-tice. Relačný model nám potom umožňuje zaviesť a využívať pojmy ako relačná integrita, kľúče (cudzí, primárny) atď.

Nech máme reláciu R , súbor atribútov X z R funkčne určuje iný súbor atribútov Y z R , ak a iba ak každá hodnota z X je úzko spojená (zviazaná) s presne jednou hodnotou z Y . Potom R vyhovuje funkčnej závislosti $X \rightarrow Y$. Inak povedané, ak poznáme hodnoty atribútu X (označme ich x), potom hodnoty atribútu Y zodpovedajúce hodnotám x môžeme určiť na základe akejkoľvek n-tice z R obsahujúcej x . Dve n-tice obsahujúce rovnakú hodnotu v X budú mať zodpovedajúcu rovnakú hodnotu v Y [14].

	CC	AC	phn	name	street	city	zip
t_1 :	44	131	1234567	Mike	Mayfield	NYC	EH4 8LE
t_2 :	44	131	3456789	Rick	Crichton	NYC	EH4 8LE
t_3 :	01	908	3456789	Joe	Mtn Ave	NYC	07974

Obr. 7 Príklad relácie Zákazníci (prevzatý z [15])

Uvažujme relačnú schému z Obr. 7 obsahujúcu údaje o zákazníkoch. Pre každého zákazníka bude záznam obsahujúci nasledovné atribúty: tri pre telefónne číslo (CC – kód krajiny, AC – kód oblasti, phn – číslo), meno (name) a tri pre adresu (street – ulica, city – mesto, zip – PSČ). Potom možno stanoviť nasledovné funkčné závislosti:

f_1 : [CC, AC, phn] $\rightarrow$ [street, city, zip],

f_2 : [CC, AC] $\rightarrow$ [city]

Funkčná závislosť f_1 hovorí, že telefónne číslo zákazníka jednoznačne určuje jeho adresu a závislosť f_2 udáva, že kód krajiny a oblasti určujú mesto zákazníka. Pomocou týchto závislostí možno skúmať konzistentnosť v zákazníckych údajoch, t.j. hľadať tie chyby (hodnoty), ktoré ich porušujú. Aplikáciou týchto závislostí na tabuľku sa nenašli také chyby a preto je možné vyhlásiť túto reláciu ako „čistú“ (bez chýb). Pri detailnejšom skúmaní relácie Zákazníci z Obr. 7 je však možné objaviť, že žiaden zo záznamov nie je bez chyby. Nasledovná závislosť $cfd1$: ([CC = 44, zip] $\rightarrow$ [street]) udáva, že pre zákazníkov z Veľkej Británie (CC = 44), PSČ (zip) jednoznačne určuje ulicu. Čiže $cfd1$ je FD, ktorá platí pre podskupinu n-tíc, pre ktoré platí CC = 44 (napr. t_1 a t_2). $cfd1$ nie je tradičná FD, pretože je definovaná s konštantami a nemusí platiť na celej relácii (napr. v USA PSČ neidentifikuje ulicu). Napríklad nasledovné závislosti $cfd2$: ([CC = 44, AC = 131, phn] $\rightarrow$ [street, city = 'edi', zip]) a $cfd3$: ([CC = 01, AC = 908, phn] $\rightarrow$ [street, city = 'mh', zip]) vylepšujú FD f_1 . Hovorí, že

pre akúkoľvek dvojicu zákazníkov z Veľkej Británie, ktorí majú kód oblasti 131 a rovnaké tel. číslo, musia mať rovnakú ulicu a PSČ a naviac mesto musí byť „EDI“. Záznamy t_1 a t_2 porušujú cfd2, pretože majú odlišnú ulicu. Hoci táto celá relácia spĺňa f_1 , obe n -tice t_1 a t_2 porušujú cfd2. Tieto obmedzenia rozširujú funkčné závislosti a označujú sa ako **podmienené funkčné závislosti** (CFD, Conditional Functional Dependency). Tento typ FD je účinnejší pri hľadaní chýb a nekonzistentností v dátach, pretože dokáže nájsť nezrovnalosti súvisiace s reálnym životom (skutočnosťou) [15].

8.2 Inkluzívne závislosti dát

Jednou z vlastností dát je referenčná integrita, ktorá, keď je splnená, hovorí, že všetky hodnoty jedného atribútu relácie musia existovať ako hodnoty iného atribútu inej (alebo aj rovnakej) relácie. Potom sa môžeme stretnúť s tzv. inkluzívnou závislosťou (ID, Inclusion Dependency), ktorú pre dve relácie V a W môžeme zapísať ako $V[A_1, \dots, A_n] \subseteq W[B_1, \dots, B_n]$, kde A_i, B_i sú rozdielne atribúty V a W . To znamená, že n -tice hodnôt v stĺpcoch $A_1, \dots, A_n$ z relácie V sa musia tiež objaviť ako n -tice hodnôt v stĺpcoch $B_1, \dots, B_n$ pre niektoré skutočnosti v relácii W .

Uvažujme príklad z Obr. 8, kde sú znázornené dve schémy. Prvá obsahuje iba jednu reláciu pre všetky jednotovarové objednávky daného obchodu a druhá obsahuje dve relácie pre dostupné produkty (tovary): pre knihy a CD média. Potom by bolo možné napísať nasledovné inkluzívne závislosti: $\text{Objednávky}[\text{titul}, \text{cena}] \subseteq \text{knihy}[\text{titul}, \text{cena}]$ a $\text{Objednávky}[\text{titul}, \text{cena}] \subseteq \text{CD}[\text{album}, \text{cena}]$ a hľadať chyby v týchto databázach. Je ale jasné, že nemožno nájsť danú knihu v objednávkach v relácii CD médií, pokiaľ by daná kniha nebola vydaná ako CD. Z toho vyplýva, že v takýchto prípadoch platia tieto inkluzívne závislosti iba za určitých podmienok. Tým sme zistili, že potrebujeme podmienené inkluzívne závislosti (CID, Conditional ID) [16], ktoré môžu vyzeráť napr. takto:

cind1: $\text{Objednávky}[\text{titul}, \text{cena}, \text{typ} = \text{„kniha“}] \subseteq \text{knihy}[\text{titul}, \text{cena}]$

cind2: $\text{Objednávky}[\text{titul}, \text{cena}, \text{typ} = \text{„CD“}] \subseteq \text{CD}[\text{album}, \text{cena}]$

cind3: $\text{CD}[\text{album}, \text{cena}, \text{žáner} = \text{„rozprávka“}] \subseteq \text{knihy}[\text{titul}, \text{cena}, \text{formát} = \text{„audio“}]$

Cind1 hovorí, že pre každú n -ticiu t z relácie *Objednávky*, ak jej typ je *kniha*, existuje n -tica t' z relácie *knihy* taká, že platí t aj t' majú rovnaké atribúty *titul* a *cena*. Analogicky to platí pre cind2. Cind3 stanovuje, že pre každú n -ticiu (záznam) t z relácie *CD*, ak jej žáner je *rozprávka*, musí byť záznam t' v relácii *knihy* taký, že titul a cena záznamu t' je zhodný s albumom a cenou záznamu t a naviac atribút formát t musí byť *audio*. Ako je vidieť, cind1 a cind2 platia, ale cind3 nie je splnená. Záznam t_7 sa nezhoduje so záznamom t_9 . Týmto sme identifikovali nezrovnalosť v databázach. Podobne ako CFD aj od CID sa očakáva, že platí iba na podskupine záznamov. CID nám umožňuje objaviť chyby naprieč rôznymi reláciami, čo tradičné ID nedokážu.

	kód	titul	typ	cena
t ₄ :	X2154	Harry Potter	kniha	21
t ₅ :	X3425	Snehulienka	CD	8,5

a.) Príklad dát pre reláciu Objednávky

	ISBN	titul	cena	formát
t ₆ :	2345-3567	Harry Potter	21	viazaná kniha
t ₇ :	6790-6528	Snehulienka	8,5	papierový obal

b.) Príklad dát pre reláciu Knihy

	ID	album	cena	žáner
t ₈ :	i457	Elán	15,5	rock
t ₉ :	i528	Snehulienka	8,5	rozprávka

c.) Príklad dát pre reláciu CD

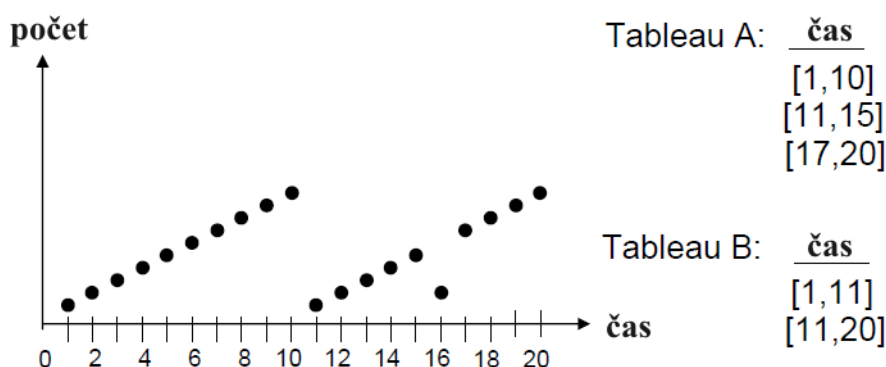
Obr. 8 Príklad jednoduchých schém pre internetový obchod

CFD aj CID je možné doplniť o operácie disjunkcia a nerovnosť, čím je možné rozšíriť ich schopnosti.

8.3 Sekvenčné závislosti dát

Funkčné ani inkluzívne závislosti dát sa nevyjadrujú k usporiadaniu hodnôt dát v atribútoch relácií. Niektoré atribúty dát (napr. časové značky, poradové čísla, namerané hodnoty predaja alebo teploty a pod.) sa vyznačujú vlastnosťou usporiadania. Pochopenie sémantiky takýchto dát môže napomôcť aj pri zlepšení ich kvality. Pre popis integrity týchto dát je možné použiť tzv. sekvenčné závislosti [17], ktoré vychádzajú z radových závislostí (OD, Order Dependency).

Nech X a Y sú atribúty relácie R a g je interval, potom zápis $X \rightarrow_g Y$ bude reprezentovať sekvenčnú závislosť týchto atribútov a hovorí, že vzdialenosť medzi hodnotami Y akýchkoľvek dvoch po sebe idúcich záznamov, keď sú usporiadané podľa X , bude v rámci intervalu g . Na to, ako dobre vystihujú sekvenčné závislosti dáta, autori používajú mieru konfidencie. Keďže niektoré dáta, ak napr. počet paketov, ktoré boli poslané linkou, majú narastajúci charakter, predsa len dané počítadlo má hornú hranicu (alebo nastane reštart počítania z iných príčin) a v niektorom okamihu sa začne počítať „od znovu“ a hodnoty potom môžu mať charakter, ako na Obr. 9. Aby bolo možné sa vysporiadať s takýmto charakterom dát, ktoré už nie je možné popísať jednoduchým zápisom ($\text{čas} \rightarrow_{(0,\infty)} \text{počet}$), nasadili sa opäť podmienené sekvenčné závislosti (CSD, Conditional Sequential Dependency), ktorých tabuľka (tableau) obsahuje podsúbory hodnôt, ktoré vyhovujú stanoveným SD. Ako je vidieť na Obr. 9, vhodnejšia tabuľka by bola B, pretože je možné z nej jednoduchšie identifikovať výkyvy, ktoré reprezentujú nesprávne hodnoty.



Obr. 9 Príklad reálnych „usporiadaných“ dát

8.4 Ďalšie typy závislosti dát

Autori v [18] navrhli **diferenčné závislosti** dát, ktoré sú založené na určení rozdielu medzi záznamami, to znamená, že kladú obmedzujúce podmienky na ich vzájomnú diferenciu. Diferenčná závislosť stanovuje, že ak dve n-tice (záznamy) majú vzdialenosť na atribúte X zodpovedajúcu určitej diferenčnej funkcii, potom ich vzdialenosť na atribúte Y by mala tiež zodpovedať tejto funkcii. Napríklad, zápis $[date(\leq 7)] \rightarrow [price(< 100)]$ udáva, že rozdiel v cene medzi dvomi dňami v rámci týždňa by nemal byť väčší ako 100 dolárov.

Pri niektorých procesoch je možné na príslušné dáta aplikovať zákon zachovania meranej veličiny. Týka sa to dát, ktoré charakterizujú vlastnosti vstupov a výstupov uzavretých systémov (napr. prichádzajúce a odchádzajúce pakety zo smerovača). Na základe tohto zákona zachovania určitej metriky je možné postaviť pravidlá a závislosti (**conservation dependency**), ktoré vyjadrujú sémantiku a charakterizujú kvalitu takýchto dát [19].

Nevýhodou funkčných a inkluzívnych závislostí je napríklad to, že sa nevedia jednoznačne vysporiadať s tým, keď rovnaké hodnoty neakého atribútu je možné zapísať rôznym spôsobom (napr. ulice v adresách). Aby bolo možné sa vysporiadať s takýmito prípadmi z reálneho života (byť tolerantný k rôznym formátom reprezentácie hodnôt), W. Fan v [20] navrhol koncept **porovnávacích závislostí** (matching dependences), ktoré využívajú niektorý porovnávací operátor, ako napr. Euklidova vzdialenosť alebo kosínusová podobnosť. Ak máme reláciu Kontakty, potom napríklad takúto závislosť možno napísať ako md1: $([ulica] \rightarrow [mesto], <0,8, 0,7>)$, čo stanovuje, že pre akékoľvek n-tice z tejto relácie, ak majú rovnaký atribút ulica (zhodujúca sa podobnosť daná použitým operátorom je tu väčšia ako hranica 0,8), potom zodpovedajúci atribút mesto by sa mal zhodovať tiež (podobnosť musí byť väčšia ako 0,7).

9 Hybridné a iné metódy

Hybridné metódy vznikajú vzájomnou integráciou štatistických a logických metód, alebo niektorou z týchto metód s iným prístupom. Spoločnou integráciou metód je možné zvýšiť efektivitu výslednej metódy pre čistenie dát. Na druhej strane sa to ale môže prejavíť väčšími nárokmi na výpočtový výkon ako aj čas, čo pri veľkých objemoch dát môže hrať dosť podstatnú úlohu.

Sú však aj metódy, ktoré na hľadanie určitého typu chýb v dátach používajú iné prístupy, ako boli tie spomínané v predošlých kapitolách. Príkladom je prístup, ktorý je založený na slovníkoch. Slovníky obsahujú veľkú databázu slov získanú napr. automatizovaným zozbieraním a môže mať statický, dynamický alebo zmiešaný charakter. Táto metóda je účinná, keď sa pracuje s údajmi získanými OCR (Optical Character Recognition) metódou pri prevode tlačenej textu do elektronickej podoby.

Varol a kol. [21] navrhli PNRS (Personal Name Recognizing Strategy) metódu, ktorú je možné použiť na čistenie fonetických a typografických chýb. Skladá sa z dvoch algoritmov: tzv. „Near Miss Strategy“ a fonetický algoritmus. Prvý zisťuje, aké blízke sú si dva porovnávané slová, pričom druhý vychádza z toho, ako sa dané slovo vysloví.

Niektorí autori vychádzajú z algoritmu Tranzitívny uzáver (uzavretie) [22]. Tento algoritmus predspracováva dáta, aby rozkategorizoval jednotlivé záznamy do skupín príbuzných dát. Týmto prístupom je možné identifikovať a odstrániť redundancie, doplniť prázdne polia, ako aj nájsť skupinové vzájomné vzťahy medzi odlišnými záznamami. Rozkategorizovanie do skupín sa deje pomocou kľúčov, čo sú vybrané atribúty dát. Keď sú záznamy v skupinách, je možné pristúpiť k ich ďalšej analýze a oprave (v rámci skupín).

Kombináciu algoritmov PNRS a Tranzitívneho uzáveru prinášajú autori v [23] pod názvom HADCLEAN, ktorou sa snažia zužitkovať výhody oboch algoritmov.

10 Extrapolácia nedostatočných dátových vzoriek.

PREDPOKLADY A PODMIENKY, KTORÉ MUSIA BYŤ NAPLNENÉ, METÓDY, NÁSTROJE EXTRAPOLÁCIE.

Predmetom tejto kapitoly je zhrnúť použitie extrapolačných štatistických techník v kontexte estimácie odchýlky výroby a spotreby elektrickej energie. Táto časť správy je formálnym výstupom analýzy vhodnosti rôznych prístupov k procesu extrapolácie dostupných dát. Cieľom analýzy bolo popísať predpoklady a podmienky, ktoré musia byť naplnené pri extrapolácii dát v energetike, prínosy a riziká použitia extrapolačných techník, popísať typy dát použiteľných na extrapoláciu, uviesť príklady extrapolačných techník, popísať postup extrapolácie v časových krokoch a uviesť možnosti umelého rozšírenia vzorky v prípade nedostatočných vzoriek dát.

1. Extrapolácia

Extrapolácia je štatistická technika inferencie nepoznaného zo známych stavov - okolností. Pomocou nástrojov extrapolácie sa snažíme predvídať budúce dátové body, pri čom sa spoliehame na historické dáta, ktoré máme k dispozícii v určitej granularite (napr. odhad veľkosti populácie za niekoľko rokov na základe medziročného tempa rastu populácie od začiatku 19. storočia). Inak povedané, dátová extrapolácia je predikcia dát mimo hranice tabuľky. Tento proces sa dá jednoducho popísať ako vzorku dátových bodov s hodnotami $x_1, \dots, x_n$, ku ktorým sa snažíme aproximovať bod s hodnotou mimo rozsah daných bodov vo vzorke. Ak tento novo identifikovaný bod započítame do vzorky, môžeme iteráciami popísať ďalšie body a vytvoriť tak prediktívny trend. Každý ďalší bod, bude mať menšiu mieru pravdepodobnosti odhadu jeho hodnoty.

2. Podmienky a limity sledovania trendov pomocou extrapolácie

Ako každá štatistická technika, extrapolácia má podmienky, ktoré musia byť splnené za účelom jej efektívnej aplikácie a metodologické limity, v kontexte ktorých sa pohybuje. Predpovede na princípe extrapolácie môžu byť veľmi efektívne, pokiaľ kontext, z ktorého analyzované dáta pochádzajú, nenaznačuje žiadne prerušenie, alebo **nepravidelné výkyvy v dlhodobom vyskytujúcom sa trende**, či minulých dátovo paralelných trendoch. Príkladom takéhoto výkyvu sú hodnoty premenných na začiatku ekonomickej krízy v dátach finančnej burzy alebo zníženie globálnej teploty po výbuchu sopky v meteorologických dátach.

Samozrejme, výkyvy v trende nám môžu pripadať ako nepravidelné alebo neočakávané z dôvodu nedostatočnej vzorky (napr. krátkych časových radov).

Avšak, takzvaná **extrapolácia z priamky** (pri ktorej predikujeme ďaleko do budúcnosti na základe krátkodobého trendu) so sebou vždy nesie riziko, pretože niektoré nepredvídateľné faktory doň takmer vždy zasiahnu. Extrapolácie môžeme použiť ako orientačnú evidenciu, napr. pre zistenie smeru vzťahu trendu a konštantnej premennej (napr. v porovnaní s terajším priemerom zrážok bude v blízkej budúcnosti pršať viac či menej?), určite však nie na predikciu presných hodnôt. Ak chceme aplikáciu extrapoláčnych metód správne použiť, je treba zvážiť nasledujúce okolnosti:

- **Chyba predpovede sa zvyšuje so zmenšovaním vzorky dát**, z ktorej chceme extrapolovať a naopak, pravdepodobnosť kvalitnej predpovede sa zvyšuje s objemom historických dát.
- Extrapolácia funguje obzvlášť dobre, pokiaľ sú **testovacie sety dát zbierané z konštantne bežiaceho signálu v pravidelných intervaloch** a tieto dáta sú použité na predpovedanie budúcich dátových bodov pri zachovanom intervale (napr. v takom istom intervale, aký bol interval zberu).
- Pravdepodobnosť chyby predpovede sa zvyšuje s požiadavkou predpovedať dlhšie časové horizonty a naopak, **čím kratší horizont, čím bližší k súčasnosti, tým máme väčšiu pravdepodobnosť generovať presnú predpoveď**.
- Technikám extrapolácie, úspechom priamo naviazaným na veľkosť vzorky, **štandardne pomáha aj širšie pole zberu dát**. Napríklad v energetike a meteorológii pôjde o rozdiel medzi v extrapolovaním dát pre celé Slovensko verzus pre jeho lokálnu oblasť.
- Pri aplikácii extrapoláčnych metód musíme mať na pamäti, že sme vzali za svoj predpoklad, že aktuálny a budúci pohyb trendu sledovaného javu má priamy vzťah k minulým stavom, sa odvíja od jeho minulých stavov. Esenciálne informácie o jeho pohybe v budúcnosti sú teda uložené v dátach o jeho minulosti a sú extrahovateľné. **Predpokladáme, že minulé trendy budú pokračovať do budúcnosti**. Aby sme zabezpečili, že je tento predpoklad dlhodobo správny, je treba vytvoriť systematický prístup k identifikácii elementov v dátach, ktoré predpokladáme, že zostanú stabilné a dlhodobo sledovať rozdiel medzi stabilnými a dynamickými zložkami trendu.

3. Výhody a prínosy vs. nevýhody a riziká extrapoláčnych techník

Výhody extrapoláčnych techník im pomohli udržať sa pri zavádzaní do praxe v oblastiach, kde je dôležité sledovať a spresňovať trendy v krátkych a častých intervaloch. Uvádžame najväčšie výhody aplikácií týchto metód:

- Extrapoláčne techniky vynikajú svojou **komputačnou jednoduchosťou**. To znamená, že pri dobrom nastavení mechanizmov kumulujúcich surové dáta do použiteľných agregátov, je systém, postavený na extrapoláčnych technikách, schopný bežať v reálnom čase a sledovať trendy intenzívnych dátových tokov.
- Druhou výhodou je **metodologická transparentnosť** extrapoláčnych systémov. Na rozdiel od predikcií generovaných pomocou neurónových sietí, pri extrapoláčnych technikách neexistuje „black box“, máme pri nich **úplnú kontrolu nad vývojom predikcie**. Na rozdiel od metód algoritmov posilňovaného učenia, má extrapoláčná štatistika **menšie nároky na iterácie**.
- Extrapolácie majú dobrú schopnosť **pracovať s agregovanými dátami** v kontexte času (s nižšou granularitou dát), bez umelého zníženia ich relevancie.
- Vďaka jednoduchosti, je jej **nasadenie bezproblémové**.

Extrapolácia je vhodná pre spracovávanie dát z veľkých oblastí, vykresľovanie trendov v krátkych horizontoch a na sledovanie premenných, v ktorých predpokladáme pomalé rasty a poklesy (teda nie náhle a extrémne výkyvy v hodnotách). Samozrejme, extrapolácia má aj nevýhody a nedostatky, ktoré paradoxne ako jej výhody, plynú z jej jednoduchosti:

- Štatistická extrapolácia neprináša odpoveď na otázku, čo je príčinou sledovaného trendu. Ide o čisto deskriptívnu techniku. Aj ako deskripcia, má nedostatky, nakoľko pracuje s agregátmi a teda **nesleduje špecifické zmeny stavu štruktúry** dátových vzoriek.
- To znamená, že napríklad tvorba nových kohort v dátach (napr. nový sektor populácie, nová skupina rovesníkov) je pre interpretátora vykresleného trendu neviditeľný. Na sledovanie dát o štruktúre a systematickom kontexte **musíme použiť pomocné frekvenčné a cluster analýzy**, ktoré nám pomôžu zvážiť zmyslupnosť aplikácie extrapoláčnych štatistických techník. Takúto analýzu štruktúry je potrebné robiť opakovane.
- Aplikácia extrapoláčnych techník ako primárneho nástroja tvorby predikcií je možná len vtedy, pokiaľ môžeme **vylúčiť prípadné externé vplyvy**, mimo systému, s ktorým

počítame. Ide o neznáme premenné, ktoré by mohli posunúť budúce hodnoty mimo škálu historických hodnôt (napríklad energetika je jedným z takýchto systémov).

- Automatizácia rozhodovacieho systému, založeného na dátach z extrapolačných predikcií, sa v praxi často nepoužíva. Výsledky extrapolácie **je nutné vizuálne skontrolovať a posúdiť**, či zodpovedajú našim poznatkom o správaní sa hodnôt v danej doméne. Tento typ analýzy má teda skôr verifikačný ako exploratívny charakter.
- Pri extrahovaní testovacích vzoriek dát za účelom extrapolácie je potrebné zozbierať dva typy vzoriek, nakoľko musíme **porovnávať výsledky extrapolácie málo a viac roztrúsenej vzoriek** dát (citlivosť na dátovú štruktúru).
- Podobný problém ako vyššie popísaný limit **extrapolácie z priamky** nastáva ak má systém počítať s veľkým množstvom premenných. Tento problém je však oveľa ťažšie detekovať, ako problém krátkych časových sledov dát. Riešením je často agregácia dát cez premenné, na základe identifikácie vzťahov medzi nimi a ich následnej korelácie.
- Pri sledovaní trendu je dobré dopredu vedieť aký typ trendu (model) máme očakávať. Ak máme informácie o očakávanom tvare trendu pre daný časový úsek (trend môže mať podobu lineárnej, kvadratickej, exponenciálnej funkcie atď.), môžeme použiť **metódu least-squares** (metóda najmenších štvorcov) za účelom nájdania optimálne vykreslenej krivky trendu. Chyba extrapolácie krivky generovanej pomocou **least-squares** je často signifikantne menšia ako chyba extrapolácie z priamky pomocou interpolovaného polynomiálu (matematickej funkcie odpovedajúcej priebehu setu dát, modelu alebo trendu). Treba si však uvedomiť, že fakt, že sme pomocou least-squares identifikovali krivku veľmi dobre fitujúcu (najlepšie kopírujúcu) na originál, nie je dôkazom toho, že extrapolácia je metóda predikcie na základe konkrétnej vzorky dát v jej kontexte a kvalite – t.j. predikcia dosiahnutá pomocou takejto extrapolácie nemusí dosiahnuť kontext a kvalitu konkrétnej zdrojovej vzorky.
- Vo väčšine prípadov, **systémy v reálnom svete** nemajú tendenciu nasledovať aktuálne trendy dlhodobo.

4. Dáta použiteľné na extrapoláciu

Kvalita extrapolácie závisí priamo od kvality vzorky dát, z ktorej extrapolujeme, preto by sme radi zbežne popísali typy dát, vhodných priamo na extrapoláciu alebo aspoň jej podporu. Vieme využiť štyri zdroje dát: historické dáta, dáta z analogických kontextov, simulácia dát -

dáta z laboratórnych alebo experimentálnych situácií a simulácia dát - dáta z priameho (poľného) zberu.

Historické dáta sú často priamou a prvou voľbou pri extrapolovaní. Pokiaľ máme dlhšie časové obdobie priamu skúsenosť (evidenciu) s javom, ktorý sa snažíme predikovať, je to jednoznačný zdroj dát, z ktorého chceme pri predikcii vychádzať. Presnosť extrapolácie na základe historických dát závisí na dvoch premenných: a) presnosť merania historických dát a ich všeobecná kvalita, b) premenlivosť trendu.

Dáta z analogických kontextov sa používajú ako podporné dáta alebo v prípade, keď historické dáta nie sú k dispozícii (sú neúplné). Steinov paradox naznačuje, že pridaním analogických dát vieme značne spresniť výsledok extrapolácie. Podmienkou je evidencia o tom, že jav, ktorý vieme sledovať, je naozaj analogický k javu, ktorý sa snažíme predikovať. Tieto dva javy by teda mali byť v stabilnom vzťahu, či už korelačnom alebo kauzálnom. Na sledovanie - nakoľko sú dva javy analogické a dáta z nich použiteľné pre jednu extrapoláciu - existujú rôzne indexy.

Simulácie môžu byť použité na generovanie dát v prípade, že nemáme k dispozícii iné zdroje dát, alebo keď sa chceme uistiť o tom, že jav v historických dátach sa správa štandardne. Simulácie môžu prebiehať experimentálne v laboratórnych podmienkach alebo v poľnom výskume. Laboratórne podmienky nám dávajú lepšiu kontrolu nad externými vplyvmi a poľné zbory majú lepšiu ekologickú validitu (lepšie simulujú podmienky reálneho sveta). V tabuľke ponúkame hodnotenie jednotlivých zdrojov dát v kontexte cieľa predikcie:

HODNOTENIE JEDNOTLIVÝCH ZDROJOV DÁT V KONTEXTE CIEĽA PREDIKCIE

(1 = najpriaznivejšia alternatíva)

Zdroj dát	Očakávam že jav je reprezentovaný				Cena zberu dát
	Odhad súčasného stavu trendu	Malými zmenami	Veľkými zmenami	Efekt výskumníkovej zaujatosti (bias)	
Historické dáta	1	1	4	1	1
Dáta z analogických kontextov	2	4	3	2	2
Laboratórna simulácia	4	3	2	4	3
Poľný výskum	3	2	1	3	4

5. Dva príklady extrapoláčnych techník

Popíšeme si dva z mnohých nástrojov extrapolácie. Vybrali sme základné techniky, ktoré sa ale často používajú v praxi.

a. Technika Markov Chains

Pri technike Markov Chains používame vzorec nedávneho správania sa javu ako základ pre tvorbu prognózy. Pri tomto prístupe sa predpokladá sa, že správanie javu v budúcnosti sa dá odvodiť zo znalosti súčasného stavu javu a z analýzy toho, ako sa jav presunie z jedného stavu do druhého. Ide teda o progresívnu a nie regresívnu metódu, lebo sa primárne zameriava na analýzu krokov vpred ($n(t) - n(t+1)$) a nie na sledovanie historickej stability trendu. Dôležitým krokom je tu kvalitne zadefinovať stavy a akcie, ktorými sa jav presúva do ďalšieho stavu (to môžu byť napríklad jednotlivé zbery, alebo zmena ročného obdobia).

b. Technika Exponential Smoothing

Exponencial Smoothing je pomerne jednoduchá, intuitívna, lacná, a dobre známa metóda. Nadväzuje na filozofiu dekompozície. Predpokladáme pri nej, že časové rady sa skladajú z niekoľkých základných prvkov: priemer, trend, sezónnosť a chyba (odchýlka). Prvé tri zložky predstavujú historické dáta a sú tak aj vpisované do grafov. V extrapoláčnom grafe sa tak dá znázorniť odhad aktualizovaného priemeru, predpoveď po tom, ako sme započítali trend do odhadu aktuálneho stavu a celkovú prognózu, ktorá bude kombináciou súčasného stavu, trendu a sezónnosti. K tejto kombinácii vieme vypočítať absolútnu a relatívnu chybu odhadu.

Exponencial Smoothing je podobný metóde kľzavého priemeru, ale novším údajom priradzuje väčšiu váhu. Priradené váhy klesajú exponenciálne od najaktuálnejších dát smerom k historickým, takže čím staršie údaje, tým menší je ich vplyv. To, že váhovanie historických dát dáva zmysel, je podporené viacerými empirickými štúdiami. Exponencial Smoothing môže byť použitý spolu s dekompozičnými technikami na analýzu historických dát.

K popisu základných metód prikladáme všeobecný popis hlavných krokov a princípov, ktoré musia byť dodržané pri extrapoláciách rôznych druhov. Kroky sú radené podľa časovej priority:

Výber a príprava dát:

- Získaj dáta reprezentujúce jav, ktorého správanie sa snažíme extrapolovať;
- Použi všetky dostupné dáta, najmä pri dlhodobých predpovediach;
- Popíš štruktúru javu a kontext, v ktorom sa jav nachádza (predikcia odchýlky výroby a spotreby elektrickej energie na Slovensku), aby si mohol na spresnenie extrapolácie využiť doménové znalosti;
- Vyčisti dáta a zníž chybu merania;
- Prispôsob dáta časovým sériám, v akých boli zbierané;
- Prispôsob dáta relevantným historickým zmenám.

Sezónne úpravy na dátach (napasovanie dát na intervaly, ktoré poznáme):

- Prispôsob dáta sezónnym zmenám v správaní javu, ak tvoje doménové poznatky poukazujú na existenciu fluktuácií javu v ich kontexte a pokiaľ je vzorka na takúto úpravu dostatočná;
- Použi multiplikatívne faktory pre stabilné časové úseky a javy, ktoré majú presné pomerovo škálovateľné dáta;
- Vypusti sezónne úpravy dát, ak čelíš vyššej miere štatistickej neistoty.

Extrapolovanie:

- Kombinuj odhady rôznych úrovní predikcií;
- Použi jednoduchú reprezentáciu trendu, až na stavy kde je silná evidencia v prospech komplexnejších kriviek;
- Naváhuj aktuálnejšie dáta viac ako staršie dáta v prípade, že je chyba merania malá, horizont predpovede je krátky a opakované merania na rôznych vzorkách sú stabilné;
- Buď konzervatívny v príprave dát, analýze a interpretácii, ak je prítomná veľká miera neistoty (osobnej alebo/aj štatistickej);
- Použi doménové poznatky na vopred špecifikované úpravy extrapolácie.

- Použi deskriptívne štatistické metódy ako pomôcku pri výbere metódy extrapolácie;
- Aktualizuj odhady parametrov predikčného modelu často;
- Používaj cykly iba vtedy, ak sú dôkazy o budúcom načasovaní a amplitúde trendu veľmi presné.

Zohľadni štatistickú neistotu (pravdepodobnosť a chybovosť):

- Použi empirické odhady vychádzajúce z out-of-sample testov,
- Pre pomerovo škálovateľné dáta, odhadni predikčné intervaly pomocou log transformácie reálnych a predikovaných hodnôt;
- Použi bezpečnostné faktory pre ostatné typy dátových sérií.

Použi extrapoláčne techniky:

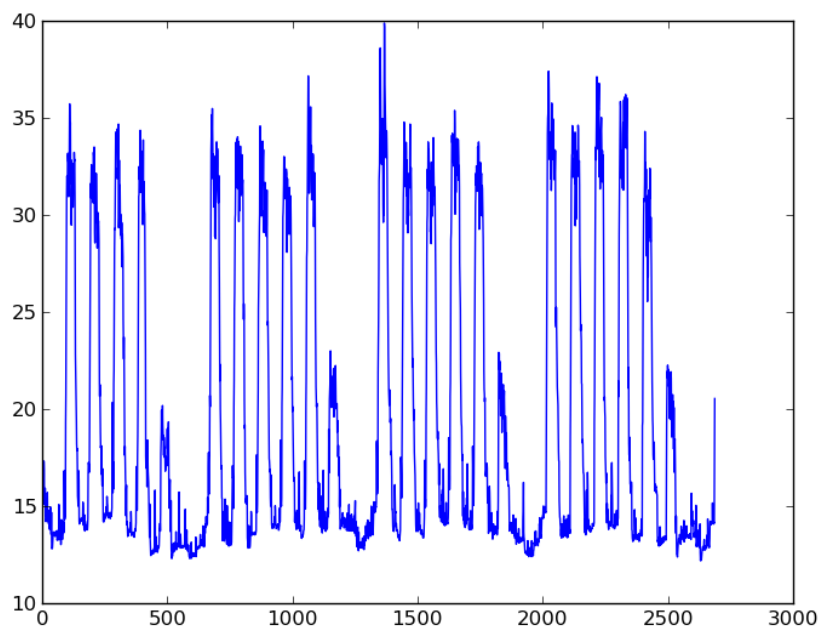
- Keď je potrebné robiť veľa predpovedí alebo opakované predpovede;
- Keď si môžeš dovoliť ignorovať väčšinu kontextuálnych premenných;
- Keď je situácia stabilná v čase;
- Keď použitie iných komplexnejších metód je nemožné;
- Keď potrebuješ merania nižšej miery efektivity predikcií (napr. identifikovanie predikčného potenciálu dát alebo stability správania sa javu).

6. Extrapolácia časových radov v kontexte odhadu odchýlky elektrickej spotreby

Ako sme videli v časti 3. Výhody a prínosy vs. nevýhody a riziká extrapoláčnych techník, štatistické techniky extrapolácie sú veľmi špecifické vo svojej aplikácii. Pre potreby predikcie v kontexte odhadu výroby a spotreby elektrickej energie na Slovensku je však extrapolácia veľmi použiteľným nástrojom. Striktne definované pravidlá výroby a spotreby, stabilné a očakávateľné vývoje trendov vo vzťahu ku klimatickým zmenám (napr. teplota), možnosť agregovania dát, pravidelnosť ich zberu, množstvo zbieraných dát, ich kvalitné namapovanie na rozlohu a relatívne malý počet premenných umožňuje využiť extrapoláciu v plnej sile najmä pri krátkych, frekventovane opakovaných predikciách na minút, hodín.

Ak by sme sa rozhodli extrapolovať dáta zo smartmetrov, t.j. 65 odberných miest zbieraných v časovom úseku od 1.1.2014 do 3.4.2014, na úrovni dní a týždňov, presnosť predikcie

značne klesne. Na druhej strane, vďaka štatistickej kontrole extrapoláčného procesu môžeme vidieť, že sa nám stále darí nájsť v dátach trendy. Ako je jasne vidieť na nasledujúcom grafe, časové rady vykazujú aj na tejto úrovni pomerne periodické správanie:



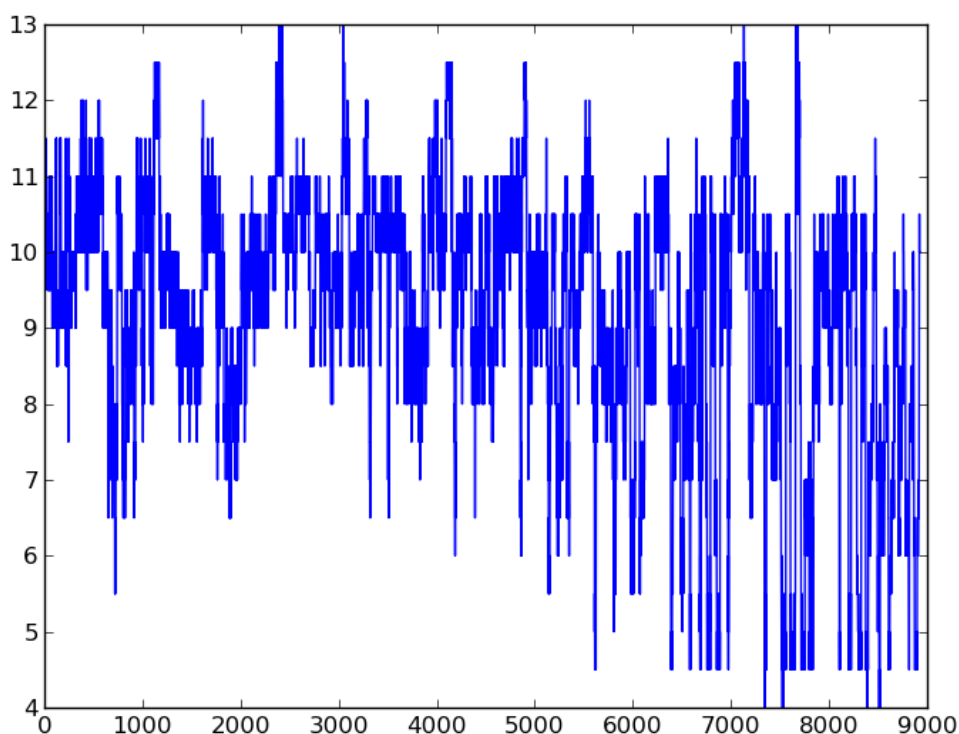
Periódy sú hlavne týždenné a denné. Pokiaľ by sme chceli extrapolovať dáta automaticky na istú dobu dopredu pomocou automatu, ktorý by používal triviálnu metódu: “predikuj to isté, čo bolo x dní dozadu”, tak by sme dostali výsledky, ktoré nám určia hranicu, chybovosti, pod ktorú sa nikdy nedostaneme.

Do dostaneme nasledovné relatívne chyby predikcie:

Interval na predikciu dopredu [dní]	Relatívna chyba (cez všetky rady)	Relatívna chyba (na najlepšom rade)	Relatívna chyba (na najhoršom rade)
1	222%	5.83%	700%
7	75%	5.80%	189%
14	86%	5.78%	227%
21	90%	5.70%	258%
28	95%	5.80%	270%

Tieto chyby môžu pôsobiť ako veľké čísla, ale treba si uvedomiť, že automat predikuje spotrebu jedného odmerného miesta v krátkom časovom úseku. To znamená, že pracuje s pomerne veľkým výskytom rôznych náhodných faktorov, na ktoré, ako sme si povedali, sú extrapoláčnne techniky veľmi citlivé. Ďalej si je treba uvedomiť, že väčšina predikcií pomocou extrapolácie sa doteraz zameriavala na hodinové úseky a väčšinou predikovala len lokálnu časť siete. My sa týmto testom snažíme určiť, kde má využitie extrapolácie limit. Preto, sa z tohto základného výsledku odrážame a posúvame výkonnosť systému smerom k väčšej presnosti tak, aby mohol byť jeho beh stále manažovaný automatom.

Pokiaľ zlepšíme kvalitu vzorky, t.j. spotrebu predikujeme z viacerých dátových bodov (smartmetering) alebo ak chceme predikovať spotrebu na týždeň dopredu, tak pridáme okrem dát o súčasnej spotrebe aj spotrebu pred týždňa a dvoch týždňov, dostaneme lepšie výsledky. Ak použijeme nielen súčasný 15 minútový interval, ale aj intervaly okolo neho a tieto hodnoty vhodne skombinujeme, výsledok sa zase zlepší. Ešte lepšia presnosť by sa dala dosiahnuť, keby sme odberné miesta mohli segmentovať do viacerých skupín (napr. domácnosti, veľké fabriky), keďže sa dá predpokladať, že u týchto segmentov existuje iné správanie v odbere energie. Predpokladáme, že signifikantný posun k presnosti predikcie pridá zakomponovanie sezónnosti (sezónneho správania spotrebiteľov a výrobcov). Takto sa dá chyba pri týždňovej predikcii dopredu znížiť na 55%. To je veľmi dobrý výsledok na metódu všeobecne považovanú za použiteľnú najmä na odhad minútových a hodinových intervalov. Ešte lepší spôsob je natrénovať lineárnu regresiu, ktorá na vstupe dostane posledný týždeň časového radu a predikuje spotrebu na týždeň dopredu. Ukazuje sa, že je možné dosiahnuť chybu okolo 50%. Zaujímavým faktom je, že u niekoľkých radov ale periódu nevidno (viď nasledujúci graf). Dôvodov môže byť viac, od neštandardného správania sa odberateľa, po nedostatky vo vzorke dát, s ktorou pracujeme.



7. Umelé rozširovanie vzorky

Pri nedostatku dát sme v niektorých situáciach schopní použiť metódy na umelé rozšírenie testovacej vzorky. Tieto metódy však treba používať s veľkou opatrnosťou, nakoľko s rošírením trendu často v nepriamom pomere rozširujeme aj chybu predikcie.

Jednou z metód je pridávanie nových časových radov. Rozširovanie vzorky je štandardná technika pri tvorbe automatických expertných systémov rozpoznávaní obrazu, keď ľahko vyrábame nové vzorky malou modifikáciou známych vzoriek (posunutie, otočenie, natiahnutie). Technika sa používa hlavne na zlepšenie robustnosti algoritmov, teda na vylepšenie schopnosti systému predikovať - identifikovať trendy, nie na vylepšenie deskriptívnych štatistík. Tento prístup sa dá presunúť aj do energetiky, kde vyrobíme nové časové rady malým pozmenením predchádzajúcich (pomocou manipulácie šumu, prenasobením konštantou, atď.) a potom môžeme zlepšiť robustnosť algoritmov na predikciu a segmentáciu.

Druhou možnosťou je už spomenutá simulácia, pridávanie umelých budúcich meraní. Presnosť týchto meraní nebude veľká, no dala by sa zlepšiť použitím analogických (napr. teplotných) dát. Na týchto dátach nemá zmysel predikovať na viac ako týždeň dopredu a majú slúžiť na orientačnú analýzu.

Poslednou užitočnou technikou rozšírenia vzorky je pridávanie nových atribútov. Nové atribúty môžu predstavovať štatistické konštrukty alebo agregáty veličín. Dajú sa dopočítať dodatočné štatistiky pre každý časový rad ako: plávajúce priemery, hodinové/denné dáta, štandardné odchýlky, časové priemery.

8. Záver

Naším cieľom v tejto kapitole bolo urobiť prehľad problematiky použitia extrapoláčnych metód na vzorkách dát, ktoré máme k dispozícii a vyjadriť sa k ich použiteľnosti pri tvorbe nového predikčného systému v kontexte smartmeteringu.

Technický popis tohto procesu je príliš komplexný pre potreby výstupného reportu, ktorý odovzdávame v rámci projektu, preto report obsahuje hlavné závery, odporúčania a špecifikácie, ktoré sme extrahovali počas našej práce.

Extrapoláčne techniky sú veľmi užitočné a jednoduché nástroje na popisovanie vývoja trendov v minútových, hodinových a denných časových úsekoch. Majú potenciál byť relevantným zdrojom verifikačných dát pri predikciách na dlhšie časové úseky – týždňov a mesiacov. Najlepší výkon dosahujú pri jemne rastúcich a klesajúcich hodnotách trendov a pri optimalizácii identifikovania trendov z periodicky zbieraných dát.

Doménovo-špecifické poznatky o energetike, dátovej štruktúre a sezónnych fluktuáciách signifikantne zlepšujú výkon extrapoláčnych nástrojov. Automatizované identifikovanie trendov pomocou extrapolácie je veľmi efektívne kontexte zvládania dátového objemu vtedy, ak sú správne nastavené systémy tvorby dátových agregátov a periodického vyťahovania testovacích vzoriek z toku surových dát, bežiaceho v reálnom čase.

11 Záver

Nespracované (špinavé) dáta uložené v skladoch dát obsahujú často množstvo chýb a nezrovnalostí, a aby nespôsobovali problémy a s nimi spojené dodatočné náklady spoločnostiam, je nutné ich odstrániť alebo opraviť. Tento aspekt sa týka mnohých oblastí zamerania vrátane energetického priemyslu. To má za následok, že spoločnosti investujú do systémov, ktoré zabezpečia požadovanú kvalitu dát v databázach. Metódy, ktoré tieto systémy integrujú, obsahujú algoritmy na čistenie dát.

V rámci tejto správy sme ukázali, že existuje niekoľko prístupov, ako sa postaviť k riešeniu tohto problému. Na jednej strane sú prístupy, ktorých základom je štatistická analýza dát, a ktoré sa štatistickými nástrojmi pokúšajú odhaliť tzv. odľahlé dáta, t.j. dáta, ktoré majú príliš extrémnu hodnotu alebo nespĺňajú stanovené kritéria. Na druhej strane je možné pristupovať k riešeniu tohto problému tak, že sa pre jednotlivé dáta v relačných databázach stanovujú základné pravidlá a obmedzujúce podmienky integrity dát. Tie dáta, ktoré vyplynú ako porušenia týchto pravidiel a obmedzení, reprezentujú chyby a nezrovnalosti v dátach. Okrem týchto dvoch prístupov možno nájsť aj ďalšie, ktoré fungujú na iných princípoch. Napríklad využívajú slovníky, referenčné dáta, šablóny a pod.

Dáta spoločností so zameraním na energetický priemysel možno rozčleniť na dve skupiny. Prvá, ktorá je spoločná pre väčšinu spoločností, obsahuje údaje o jednotlivých zákazníkoch. Druhá skupina obsahuje údaje, ktoré vznikli ako výsledok merania veličín o produktoch, ktoré ponúkajú svojim zákazníkom. V tomto prípade sa to týka najmä spotreby energie zákazníkmi.

Ak sa sústredíme na prvú skupinu dát o zákazníkoch (mená, adresy, identifikátory, typ sústavy, napäťové hladiny, atď.), tak pracujeme s kategorizovanými dátami, ktoré majú svoju štruktúru a reprezentáciu v relačných databázach. Na čistenie takýchto dát sa javia, ako účinné logické metódy, ktoré pri aplikácii podmienených funkčných a inkluzívnych závislostí môžu poskytnúť prijateľnú úroveň zabezpečenia ich kvality.

V prípade druhej skupiny dát – meraných dát, ktoré môžu byť charakteru jednorozmerných ale aj viacrozmerných numerických dát, sa javia ako vhodné štatistické metódy. Pri využití robustných štatistík založených na mediáne a rozdelení, ktoré najlepšie reprezentuje tieto dáta, je možné nájsť odľahlé hodnoty v dátach. Ak tieto dáta budeme charakterizovať ako dáta časových radov, potom metódy využívajúce okná a kľzavé priemery môžu byť aplikované tiež.

Ako už bolo spomínané, vhodnou kombináciou spomenutých metód je možné zvýšiť účinnosť čistenia dát, ale je nutné nájsť efektívne algoritmy, ktoré dokážu v prijateľnom čase spracovať veľké objemy dát a aplikovať príslušné čistiace úkony.

Literatúra

- [1] T. Redman: *The impact of poor data quality on the typical enterprise*. Communications of the ACM, 1998, pp. 79-82
- [2] M. R. Trifu, M. L. Ivan: *Big Data: present and future*. Database Systems Journal, vol. V, no. 1/2014, pp. 32-41
- [3] F. Chiang, R. J. Miller: *Discovering data quality rules*. Proceedings of the VLDB 2008, pp. 1166-1177.
- [4] E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches*. Data Engineering, Vol. 23 No. 4, 2000, pp. 3-13
- [5] L. Sebastian-Coleman: *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Elsevier, 2013, 376 pages
- [6] H. Müller, J.-Ch. Freytag: *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. HUB-IB-164, 2003
- [7] J. M. Hellerstein: *Quantitative Data Cleaning for Large Databases*. White Paper, 2008
- [8] C. Leys, et al.: *Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median*. Journal of Experimental Social Psychology 49, 2013, pp. 764–766
- [9] D. M. W. Powers: *Applications and explanations of Zipf's law*. Association for Computational Linguistics, 1998, pp. 151–160
- [10] M. Hazewinkel, ed.: *Mahalanobis distance*, Encyclopedia of Mathematics, Springer, 2001
- [11] T. Dasu and T. Johnson: *Exploratory Data Mining and Data Cleaning*. Wiley, 2003
- [12] R. Sparks, C. OkuGami: *Data Quality: Algorithms for Automatic Detection of Unusual Measurements*. Frontiers in Statistical Quality Control 10, 2012
- [13] M. Martin and S. Roberts: *An evaluation of bootstrap methods for outlier detection in least squares regression*. Journal of Applied Statistics, 33(7), 2006
- [14] J. Liu, J. Li, C. Liu, Y. Chen: *Discover Dependencies from Data - A Review*. IEEE Transactions on Knowledge and Data Engineering, vol.24, no.2, 2012, pp.251-264
- [15] W. Fan, F. Geerts, X. Jia: *Conditional Dependencies: A Principled Approach to Improving Data Quality*. BNCOD 2009: pp. 8-20

- [16] J. Bauckmann, Z. Abedjan, U. Leser, H. Müller, F. Naumann: *Discovering conditional inclusion dependencies*, Proc. of the 21st ACM international conference on Information and knowledge management, 2012, Maui, Hawaii, USA
- [17] L. Golab, H. J. Karloff, F. Korn, A. Saha, D. Srivastava: *Sequential Dependencies*. PVLDB 2009, pp. 574-585
- [18] S. Song, L. Chen: *Differential dependencies: Reasoning and discovery*. ACM Trans. Database Syst. 36(3): 16, 2011
- [19] L. Golab, H. J. Karloff, F. Korn, B. Saha, D. Srivastava: *Discovering Conservation Rules*. ICDE, 2012, pp. 738-749
- [20] W. Fan: *Dependencies revisited for improving data quality*. In PODS, 2008, pp. 159–170
- [21] C. Varol, C. Bayrak, R. Wagner and D. Goff: *Application of the Near Miss Strategy and Edit Distance to Handle Dirty Data*. Data Engineering - International Series in Operations Research & Management Science, vol. 132, 2010, pp. 91 -101
- [22] W. N. Li, R. Bheemavaram, X. Zhang: *Transitive Closure of Data Records: Application and Computation*. Data Engineering - International Series in Operations Research & Management Science, Springer US, vol. 132, 2010, pp. 39-75
- [23] A. Paul, V. Ganesan, J. S. Challa and Y. Sharma: *HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses*. Department of Computer Science & Information Systems. Birla Institute of Technology & Science, Pilani, Rajasthan, India – 333031.