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

KONTROLA KVALITY DÁTOVÝCH ZDROJOV A AUTOMATIZÁCIA ANONYMIZAČNÝCH PROCESOV.

Predmetom piatej analýzy v rámci projektu je jeho upresnenie v kontexte požiadaviek na potrebnú kvalitu dátových zdrojov. Táto správa je len 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.

1. Predpoklady a podmienky kvalitného zdroju 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 kompletné, 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. Tejto téme sa budeme venovať v budúcom reporte.

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 náročný. 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.

2. Kritéria

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

1. **Použiteľnosť:**
    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

3. Naše získané energetické dáta

1. **Použiteľnosť**

    1. Dáta sú za obdobie 93 dní (od 1.1. 2014 do 3.4.2014) pri SMART metroch a rok (od 1.1.2013 do 1.1.2014) pri ostatných.
    2. Meranie sa uskutočňujú každých 15 minút pri SMART metroch, pri ostatných len mesačné alebo ročné.
    3. Dátová vzorka pozostáva s 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 kompletné, 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.

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.

5. Slovenský hydro-meteorologický ústav

1. **Použiteľnosť**
    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. Iné open dáta zdroje

Meteo dáta z FR

1. **Použiteľnosť**
    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. **Použiteľnosť**
    1. Dĺžka histórie uchovávania 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 Francúzka, 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. **Použiteľnosť**
    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.

7. Automatizácia procesu anonymizácie dát v SmartGrid

V minulom reporte sme uviedli, ako by mal vyzerať 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 spomenuli v minulom reporte, 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 dostavajú 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 administračného 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 inštitúcii zostali nemenné. Automatizácia autentifikačného systému a výmazu dát ďalej vyžaduje, aby zostalo nemenné rozloženie inštitúcii 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.

8. Záver

Našim 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).

Nasledujúci report sa bude venovať možnostiam extrapolácie nedostatočných dátových vzoriek. Predpokladom a podmienkam, ktoré musia byť naplnené. Metódam a nástrojom extrapolácie.