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

AKTUALIZOVANÝ POPIS PRÍSTUPNÝCH DÁT O SPOTREBE A ICH VZŤAH K SOCIO-DEMOGRAFICKÝM A HYDROMETEOROLOGICKÝM DÁTAM.

POŽADAVKY NA SOCIODEMOGRAFICKÉ A HYDROMETEOROLOGICKÉ DÁTA, PO SPOJENÍ DÁTOVEJ VZORKY I., II. A III., V KONTEXTE PROJEKTOM ZADEFINOVANÝCH VÝSKUMNÝCH OTÁZOK.

Predmetom tejto analýzy v rámci projektu bolo zhrnúť celkovú množinu vzoriek dát, ktorú máme k dispozícii do dátumu 20. 12. 2014, nad ktorou chceme v budúcnosti testovať modely real-time manažmentu dát, ich deskriptívnej a extrapolačnej analýzy a navrhnúť systém pre automatickú tvorbu agregátov a ich bezpečnú distribúciou na základe bezpečnostného request-access formátu. Táto správa je len formálnym výstupom analýzy, procesu spájania vytvorenia špecifickej databázy a modelovania troch samostatných dátových vzoriek, ktoré sme v rámci projektu získali. Budeme ich nazývať dátová vzorka I., II. a III.

1. Definovanie cieľov analýzy.

Ciele tejto analýzy sú:

1. spracovať novo-sprostredkované dáta do použiteľného formátu,

2. ohodnotiť potenciálnu pridanú hodnotu vzorky II. a III. samostatne,

3. vytvoriť databázu a automatizovaný systém na pred-spracovanie (tzv. automated data format pre-processing), ktorý nám umožní pridávať, spájať a dopĺňať nové dáta do databázy automaticky,

4. spojiť všetky vzorky (tzv. data merging),

5. skontrolovať kvalitu dát v spojenej databáze deskriptívnou a frekvenčnou analýzou,

6. aplikovať funkciu, ktorá automaticky hľadá duplikáty v dátovom sete,

7. popísať aký typ a akú granularitu socio-demografických a hydrometeorologických dát budeme potrebovať na zodpovedanie týchto otázok:

- S akou efektivitou vedia jednotlivé modely kombinovaných dát **predpovedať spotrebu** elektrickej energie na ďalšiu hodinu, deň, mesiac a rok?

- Aká je **závislosť spotreby** elektrickej energie v domácnostiach **na priebehu počasia**? Je hydrometeorologická predpoveď relevantným zdrojom dát, schopným priniesť pridanú hodnotu do modelu?

- Aká je **závislosť spotreby** elektrickej energie v domácnostiach **na socio-demografických parametroch obyvateľstva**? Je kvalitná socio-demografia relevantným zdrojom dát, schopným priniesť pridanú hodnotu do modelu? Príklady pod-otázok:

    - Aká je závislosť spotreby elektrickej energie domácností od veľkosti sídla odberateľa?

    - Aká je závislosť spotreby elektrickej energie domácností na veľkosti príjmov domácnosti?

    - Aká je závislosť spotreby elektrickej energie domácností na počte a veku členov domácností?

    - Aká je závislosť spotreby elektrickej energie domácností na stave zamestnanosti členov domácnosti?

- Aká je **závislosť výroby na** konkrétnom **mieste**, konkrétnej nadmorskej **výške** a konkrétnom **type solárneho zdroja**? Aká bude efektívnosť tejto investície?

2. Popis dátových vzoriek II. a III.

Popis dátovej vzorky I. a jej miery potenciálu/aplikácie na tvorbu modelov estimácie odchýlky výroby a spotreby elektrickej energie sme podrobne uviedli na strane č. 9, v reporte **č. 2**: **„Prehľad použiteľných close-data zdrojov (neverejné dáta), analýza ich kvality a možnosti ich využitia v modeloch predikcie a optimalizácie výroby a spotreby elektrickej energie na Slovensku.“**

Druhý dátový set (dátová vzorka II.) obsahoval dáta z jednej regionálnej distribučnej sústavy. Väčšina dát v ňom je z Nitrianskeho, Žilinského a Banskobystrického kraja. Skladal sa z 299 999 odberných miest. Dáta v tejto vzorke boli zbierané od 1. 1. 2013 do 31. 12. 2013 (jeden cely rok), s 15 minútovou frekvenciou.

Tretí dodaný dátový set bol anonymizovaný, o odberných miestach vieme iba ich PSČ. K faktu, že takáto anonymizácia je relatívne nedostatočná, sa vyjadríme v ďalšom reporte, kde si ukážeme na príkladoch, aké informácie je možné získať na základe takýchto dát, pomocou techník kros-referencie s verejne prístupnými databázami. Distribúcia sledovaných odberných miest je rovnomerná cez všetkých 8 krajov. Dáta sledujú 21 502 odberných miest po dobu od 1. 1. 2014 do 15. 12. 2014.

Dátové vzorky I., II. a III. sme spojili do jednej funkčnej databázy. K databáze je vyvinutý systém, ktorý nám pomáha k aktuálnej vzorke pridať dáta rôznej štruktúry, hľadá duplikáty a kalibruje kompatibilné dátové rady/dvojice (funkcie semi-automatického predspracovania dát).

3. Kompatibilné hydrometeorologické dáta, potrebné pre zodpovedanie projektových hypotéz

O štruktúre dát, ktoré zbiera Slovenský hydrometeorologický ústav (a historické dáta uchováva v klimatologickom a meteorologickom informačnom systéme) a Slovenská elektrická prenosová sústava sme informovali **v reporte č. 2.5**: **„Štruktúra hydrometeorologických dát používaná v energetike.“**

V kontexte našich hypotéz a spojenej vzorky dát o spotrebe elektrickej energie domácností, by sme mali pre projekt získať tieto premenné (zoradené podľa dôležitosti):

- Teplota 2 m nad zemou
- Globálne žiarenie (W/m2)

- Pokrytie oblačnosťou (v osminách)
- Rýchlosť vetra v 10 m nad zemou

V ideálnom prípade by sa nám malo podariť získať tieto dáta v časovom rozsahu od 1. 12. 2012 do 31. 12. 2014. Čo sa dátovej granularity týka, na modelovanie sú optimálne hodinové alebo minútové dáta z čo najväčšieho počtu staníc, ktoré obsahujú:

- aktuálne namerané údaje,
- predpovede na dnes,
- predpoveď D+1,
- predpoveď D+2,

Takéto dáta nám umožnia korelovať kvalitu predikcie elektrickej spotreby s použitím reálnych meteorologických údajov a predpovedí. Minimálnym zdroj meteorologických dát do modelov novej generácie, ktorý stále vie pridať hodnotu do výsledku sú hodinové dáta z aktuálne nameraných údajov.

4. Kompatibilné socio-demografické dáta, potrebné pre zodpovedanie projektových hypotéz

O štruktúre sociodemografických dát, ktoré zbiera Štatistický úrad Slovenskej republiky a Databáza regionálnej štatistiky sme informovali **v reporte č. 2**: **„Prehľad použiteľných close-data zdrojov (neverejné dáta), analýza ich kvality a možnosti ich využitia v modeloch predikcie a optimalizácie výroby a spotreby elektrickej energie na Slovensku.“**

Za účelom porovnania a verifikácie modelov, k týmto dátam priložíme aj dáta z otvorených a prístupných zdrojov zahraničných krajín. V kontexte našich hypotéz a spojenej vzorky dát o spotrebe elektrickej energie domácností, ktorú máme k dispozícii, by sme mali pre projekt získať prístup do týchto databáz:

Štatistický úrad Slovenskej republiky má schopnosť extrahovať pre nás za poplatok surové dáta (nie informácie – agregovanú štatistiku ku ktorej je prístup zdarma) o mestskej a obecnej štatistike, pre každú obec. Príkladom takýchto dát je napríklad:

- prítomnosť verejného vodovodu, kanalizácie, rozvodu plynu,
- prítomnosť rozvodov káblovej TV.

Samostatná Databáza DATAcube, ktorú spravuje Štatistický úrad Slovenskej republiky, obsahuje multidimenzionálne tabuľky za ukazovatele hospodárskeho a sociálno-ekonomického vývoja. Zaujímajú nás tieto premenné:

- Obyvateľstvo:
    - Veková štruktúra, ukazovatele veku (pre každú obec, každý rok chceme rozloženie obyvateľstva podľa veku).
    - Základné demografické údaje (počet obyvateľov, pôrodnosť, mortalita) podľa obcí, ročné.
- Práca:
    - Miera evidovanej nezamestnanosti podľa okresoch a rokov.
    - Bilancia ekonomickej aktivity obyvateľstva (všetky stĺpce, podľa krajov, štvrťročné dáta).
- Náklady práce:
    - Mzdy zamestnancov podľa SK NACE (podľa okresov, ročné).
- Príjmy a spotreba:
    - Príjmy a životné podmienky domácnosti (podľa krajov, ročné).
- Makroekonomické štatistiky - národné účty:
    - Regionálny hrubý domáci produkt (podľa krajov, ročné).

Kľúčové sú dáta v týchto premenných za obdobie od 1. 12. 2012 do 31. 12. 2014. Čím viac ich bude, tým lepšie, čím väčšia granularita pre jednotlivé premenné, tým lepšie. Pokiaľ budú takéto dáta k dispozícii, budeme schopní pre každú socio-demografickú premennú skúmať závislosť.

5. Záver

Našim cieľom v tomto reporte bolo skompletizovať finálnu vzorku dát o spotrebe elektrickej energie, dostať ju do požadovaného formátu, ohodnotiť jej potenciál, navrhnúť o aké socio-demografické a hydrometeorologické dáta ju máme v projekte doplniť a pripraviť systém automatickej integrácie a formátovania dát. Sprostredkované dáta sú dobrej kvality (variabilita) a je ich dostatočný počet. Naše možnosti skúmania závislostí medzi spotrebným profilom odberného miesta, meteorologickými podmienkami a socio-demografickým

charakterom spotrebiteľom budú závisieť, v akej kvalite dostaneme hydrometeorologické a socio-demografické dáta, ktoré sme špecifikovali v tomto reporte.

Zároveň chceme upozorniť, že sme popisovali vzorku dát, ktorá je optimálna pre naše ciele a pokiaľ bude možné získať len dáta menšieho časového rozsahu, poprípade menšej granularity, utrpí tým síce presnosť testov, ale nie základný vývoj predikčných modelov.

V ďalšom reporte sa budeme venovať prehľadu systémov inteligentnej analýzy dát a ich efektivity, použiteľných na vývoj modelov predikcie a optimalizácie výroby a spotreby elektrickej energie na Slovensku, postavených na v aktuálne prístupných dátach, ďalej príkladom de-anonymizácie dátových vzoriek. Na strane vývoja začneme testovať predikčný potenciál modelov na dátach, ktoré máme k dispozícii.