

Toelichting Open Dataset STORM onderstation data

Achtergrond

Deze dataset betreft metingen en "bottom-up"-inschattingen van de belasting over 180 verschillende installaties in onder-, regel en schakelstations. De meting is altijd een optelling van meerdere fysieke metingen van meerdere parallelle transformator- of kabelcircuits die een installatie in een station voeden. Samen vormen deze metingen de belasting van het station, die binnen bepaalde grenzen moet blijven om ervoor te zorgen dat kabels en transformatoren niet teveel opwarmen.

Verder is er voor elk tijdstip opgeslagen of het een verschakeling of andere outlier is (het label). Alle gemeten onderstations voeden een groot aantal klanten, zodat er geen individuele gebruikspatronen te herleiden zijn.

De dataset geeft inzicht in hoe inschattingen van de belasting van het net gefilterd moeten worden om een goede inschatting te krijgen van de daadwerkelijke netbelasting. Algoritmen om dit automatisch te doen kunnen op deze dataset geëvalueerd worden. Door de data te open-sourcen is er de verwachting dat meerdere onderzoeksgroepen vergelijkend en verbeterend onderzoek kunnen doen.

Zie bijvoorbeeld Bouman et al. 2024: <https://arxiv.org/abs/2405.16164>

Formaat dataset

De dataset met metingen bestaat uit 180 CSV bestanden, deze zitten in de mappen "X". Elk van deze bestanden correspondeert met een fysiek onderstation dat gemeten is over een periode van 1 jaar. Elk bestand is ongeveer 2MB in grootte, en omvat metingen om de 15 minuten van de meting en "bottom-up load" evenals de bijbehorende timestamp.

De dataset met de labels bestaat uit opnieuw 180 CSV bestanden, deze zitten in de mappen "y". De data is daarnaast gesplitst in Train, Test en Validatie, om goede onafhankelijke tests te kunnen doen over de kwaliteit van algoritmes. Omdat er geen informatie nodig is over b.v. de locatie van het station zijn de ids/namen vervangen door willekeurige nummers. De bestandsnaam correspondeert met dit willekeurige nummer.

Het tekstscheidingsteken van de csv bestanden is “,”.

Geldigheid

Het betreft een geheel jaar aan data per station, of over 2020, of over 2021. De metingen zelf komen realtime binnen, en historisch gezien jaarlijks verwerkt. De labeling is voor dit project eenmalig gedaan.

Betekenis van de kolommen

- *M_TIMESTAMP* – Datum en tijdstip in UTC
- *S_original* – Meting in MVA
- *BU_original* – Bottom-Up load in MW
- *label* – waarden 0 (geen verschakeling/outlier), 1 (verschakeling/outlier) en 5 (onduidelijk, gaat wel mee in algoritme maar telt niet mee bij de beoordeling)