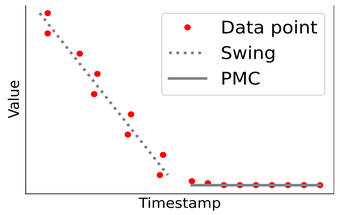
Evaluation of Lossless and Lossy Error-Bounded Compression on High-Frequency Wind Turbine Datasets

Abduvoris Abduvakhobov
abduvorisa@.cs.aau.dk
ESR 2.3: Model-based Storage for Time Series

Problem statement

- Conventional time series compression methods are either ineffective in use of storage or do not preserve desired data quality.
- Error-bounded model-based time series compression hits the sweet spot between compression effectiveness and data quality.
- Lack of relevant studies using very large high-frequency wind turbine datasets.



Example of a model.

Experimental Setup

Four aspects of the experiment:

Dataset Aspect

- **PCD:** ~36 months of wind park power controller measurements, cols=10, len=~480M, SI=150ms.
- **MTD:** ~11 months of multiple wind turbine measurements, cols=6, len=~258M, SI=2s.
- **WTM:** 10 days of turbine measurements, cols=10, len=~432K, SI=2s.

Compression Method Aspect

- **Baseline Lossless Compression**: Multivariate time series stored in a single Apache ORC file compressed with Snappy.
- **Baseline Lossy Compression:** Aggregation method by n period using a function of mean.
- **Lossless and Lossy Error-Bounded Compression (EBLC):** Combination of Gorilla, PMC-Mean (PMC) and Linear Swing (Swing).

Sampling Interval Aspect (SI)

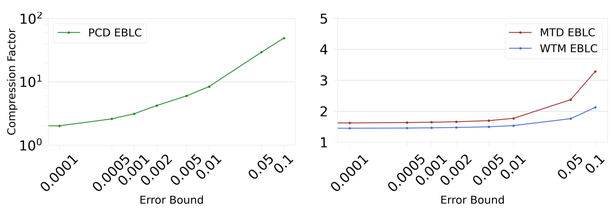
Downsampling of datasets using the SI:
PCD: *1.05s (7x), 2.1s (14x), 4.95s (33x), 10.05s (67x), 1m (400x), 10m (4000x)*.
MTD and **WTM**: *6s (3x), 10s (5x), 30s (15x), 1m (30x), 10m (300x)*.

Error Bound Aspect (ε)

ε chosen for EBLC**:** *0.01%, 0.05%, 0.1%, 0.2%, 0.5%, 1%, 5%, 1%.*

RQ1.1

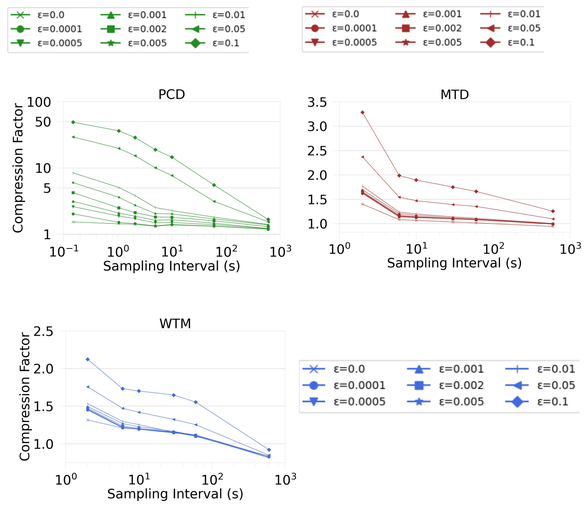
RQ1: How well does a high-frequency wind turbine dataset compress with EBLC?
RQ1.1: How does EBLC compare against the baseline lossless compression?



- Up to 1.5x and 49.5x compression than the baseline lossless method for ε=0% and 10%.

RQ1.2

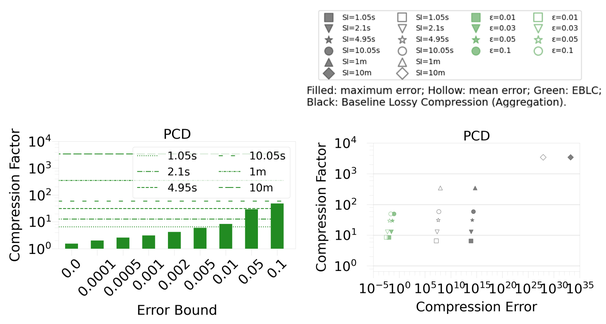
RQ1.2: How does the SI of a high-frequency wind turbine dataset affect EBLC?



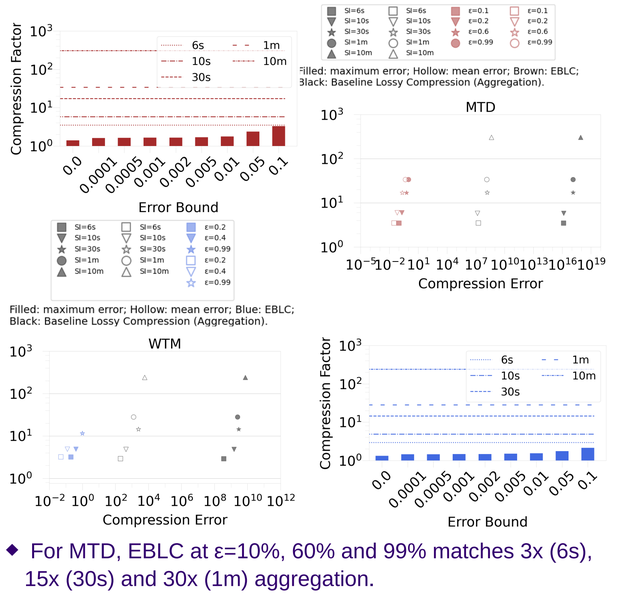
- *Negative correlation* between SI and CF (compression factor).
- Increase in the ε further increases correlation.

RQ1.3

RQ1.3: How does EBLC compare against the baseline lossy compression?



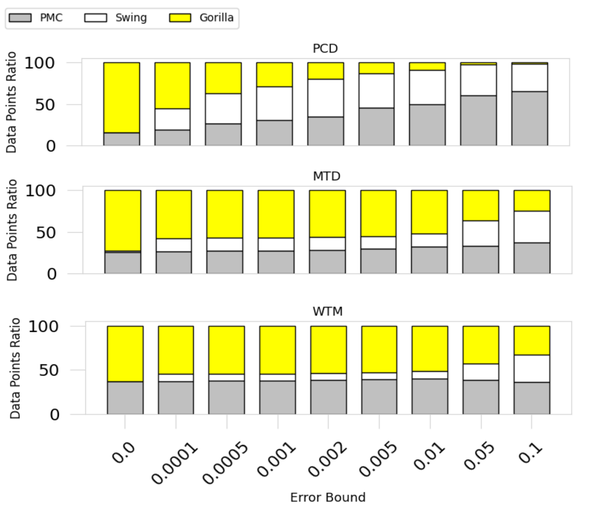
- EBLC with ε=0.5% matches 1.05s (7x) aggregation.
- ε=10% matches 10.05s (67x) aggregation.



- For MTD, EBLC at ε=10%, 60% and 99% matches 3x (6s), 15x (30s) and 30x (1m) aggregation.
- EBLC adds many orders of *magnitude less error* than aggregation.

RQ2.1

RQ2: How model types are used for the different aspects?
RQ2.1: What is the distribution of model types?



- At ε=0%, Gorilla is the most used model type for all datasets.
- At ε>0% us of PMC and Swing increase for all datasets.