TPR and Findings from Paper 1

ESR 2.3: Model-based storage for time series

July 7, 2023

Abduvoris Abduvakhobov

Content

1. Thesis Proposal Report (TPR)
2. Findings from Paper 1

TPR

TPR, what has changed from DPP?

- State of the Art
- Project objectives
- Significance and outcome (will be covered in *Findings from Paper 1*)
- Work plan
- Publication plan
- PhD courses

State of the Art

Added *time series compression* techniques:

- **Dictionary-based**: TRISTAN, CORAD …
- **Functional approximation**: Piecewise Polynomial Approximation, Chebyshev Polynomial Transform …
- **Autoencoders**: Recurrent Convolutional Autoencoder, Recurrent Neural Network Autoencoder …
- **Sequential algorithms**: Huffman Encoding, Run-Length Encoding, Delta Encoding …
- **Others**: Continuous Hidden Markov Chain …

Project Objectives

RQ1: How is the effectiveness of ModelarDB's error-bounded model-based compression to varying error bounds and datasets?

RQ2: Depending on the outcomes of RQ1, what other model types can be implemented to improve the compression and query performance of ModelarDB on real-life RES datasets?

RQ3: How to integrate user requirements for compression effectiveness and data quality to error-bounded model-based time series compression?

RQ4: How to organize on disk storage of model-based time series compression?

Work Plan

- Shift in time due to a **parental leave**
- Main body of Paper 1 is ready, but **submission is delayed**.
- The rest is developing **according to the work plan**

Time	Plan
Spring' 22 (Home)	Literature study Problem formulation Preparation of the Doctoral Project Plan Begin work on Paper 1 Begin development of ModelarDB performance evaluation tool Data collection Start general and project-related courses Trying out ModelarDB on real-life datasets and analyzing the results
Milestones	**Submission of 2-month Doctoral Project Plan** **Establish collaboration with secondment partner and get access to data**
	Parental leave: June 1 - October 1, 2022
Spring' 23 (Home)	Continue project-related courses Develop and test performance evaluation tool for ModelarDB
Milestones	**Submission of Paper 1** **Submission of 11-month Doctoral Project Plan**
Fall' 23 (Host)	Develop new model types for ModelarDB Begin work on Paper 2 Refine and test new model types with real-life datasets
Milestones	**Submission of Paper 2**

Publication Plan

Tentative Title of Paper 3: Integrating user requirements for error-bounded model-based time series compression.

Type: Conference paper.

Description: This paper will focus on investigating the possibility of adjusting the sampling interval and model error bound depending on the user requirements including desired compression ratio, average actual error and distribution of actual errors. This will allow for optimized collection and analysis of high-frequency wind turbine datasets. An advisor tool will be implemented in addition to ModelarDB that helps to configure ModelarDB according to user requirements by ingesting a sample of an input dataset.

Datasets: Siemens Gamesa Renewable Energy and ENGIE data.

Authors: A. Abduvakhobov, E.Hedevang, S.K. Jensen, T. B. Pedersen, C. Thomsen, E. Zimányi.

Length: 12 pages. **Time of submission**: July, 2024. **Outlet**: SIGMOD.

Publication Plan

Tentative Title of Paper 4: Optimizing storage component for model-based time series compression systems.

Type: Conference paper.

Description: This paper will serve as a logical continuation of implementing new model types that improve compression in terms of functional approximation. As a next step, we answer the following main research question: How to organize storage component of an error-bounded model-based time series compression system? This will allow us to exploit error-bounded model-based compression for optimizing the use of memory and disk space. In addition to effective time series compression, it will also enable pruning and executing analytical queries efficiently.

Datasets: Siemens Gamesa Renewable Energy and ENGIE data.

Authors: A. Abduvakhobov, E.Hedevang, S.K. Jensen, T. B. Pedersen, C. Thomsen, E. Zimányi.

Length: 12 pages. **Time of submission:** March, 2025. **Outlet**: EDBT.

PhD Courses

Course Name	At	Type	ECTS	Time	Status
Applying the Danish Code of Conduct for Research Integrity to your Research	AAU	General	1	Spring'22	Completed
Introduction to the PhD Study	AAU	General	0.5	Summer'22	Completed
Incremental Machine Learning Algorithms for Time-series Data	AAU	Project	2.0	Spring'22	Completed
Writing and Reviewing Scientific Papers	AAU	General	3.75	Spring'23	Completed
Data quality management	AAU	Project	2.0	Spring'23	Ongoing
Academic Writing in English	AAU	General	2.5	Summer'23	Planned
Recent Advances within Specialized Data Management Systems	AAU	Project	2.0	Fall'23	Planned
TBD	AAU	Project	2.0	TBD	Planned
Winter School (ARC)	ARC	General	3	Spring'22	Completed
Summer School (ULB)	Cesena (Italy)	Project	3	Summer'22	Completed
Winter School (AAU)	AAU	General	4	Spring'23	Completed
Summer School (UPC)	UPC	Project	3	Summer'23	Mandatory
Conference Attendance	TBD	Project	3	TBD	Planned
Danish Lessons	TBD	General	TBD	TBD	Mandatory
Total: 30.75 Gen.: 13.75 Proj.: 17					

Paper 1

Title: *"Evaluation and Experiment on Losless and Lossy Error-Bounded Compression of High Frequency Wind Turbine Datasets"*

Authors: A. Abduvakhobov, S.K. Jensen, T. B. Pedersen, C. Thomsen**.**

Status: 11 pages written, to be submitted in the 3rd round of EDBT 2023.

Problem Statement:

- Lack of evaluation of ***Lossless and Lossy Error-Bounded Time Series Compression (EBLC)'s* compression effectiveness** on large high-frequency wind turbine datasets.
- Lack of **studies that compare EBLC with widely practised compression** methods in the industry.
- Need for understanding **what factors define EBLC's effectiveness**.
- EBLC hits **the sweet spot between compression effectiveness and data quality**.

Experimental Setup

Four Aspects of the experiment:

1. Dataset 2. Compression method 3. Sampling interval 4. Error bound

Dataset Aspect

Name	Description	Rows	Columns	Sampling Interval
PCD	~36 months of wind park power controller measurements	~480M	10	150ms
MTD	~11 months of multiple turbine measurements	~258M	6	2s
WTM	10 days of turbine measurements	432K	10	2s

Compression Method Aspect

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

**These baselines are chosen because they are widely applied in the industry*

Sampling Interval Aspect (SI)

Dataset	SI	Number of datapoints aggregated into 1
PCD	1.05s, 2.1s, 4.95s, 10.05s, 1m, 10m	7, 14, 33, 67, 400, 4000
MTD and WTM	6s, 10s, 30s, 1m, 10m	3, 5, 15, 30, 300

Error Bound Aspect (ε)

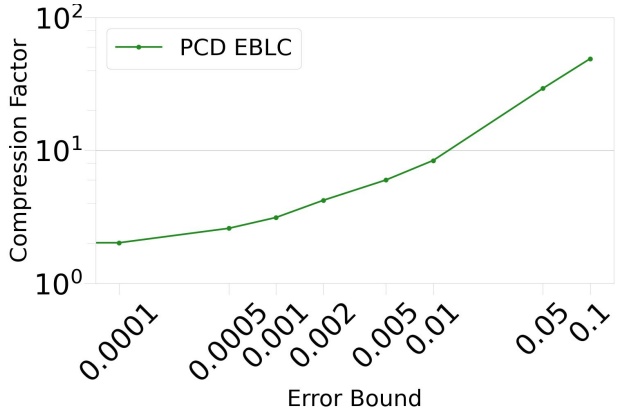
ε chosen for EBLC: 0.01%, 0.05%, 0.1%, 0.2%, 0.5%, 1%, 5%, 10%.

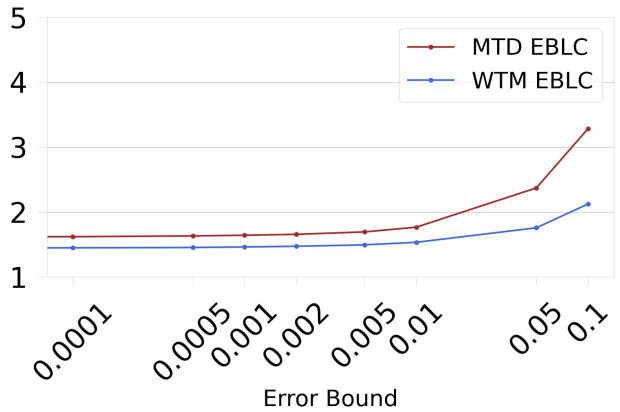
**error bounds were chosen based on discussions with secondment partner*

selected Results

***RQ1**: How well does a high-frequency wind turbine dataset compress with EBLC?*

***RQ1.1**: How does EBLC compare against the baseline lossless compression?*

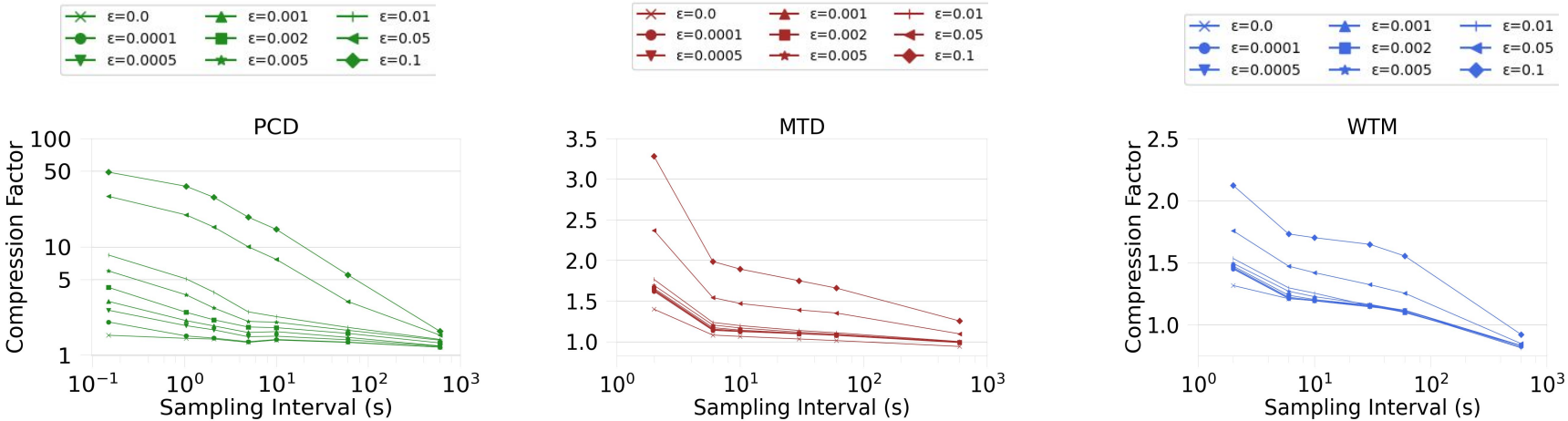




***Answer**:*

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

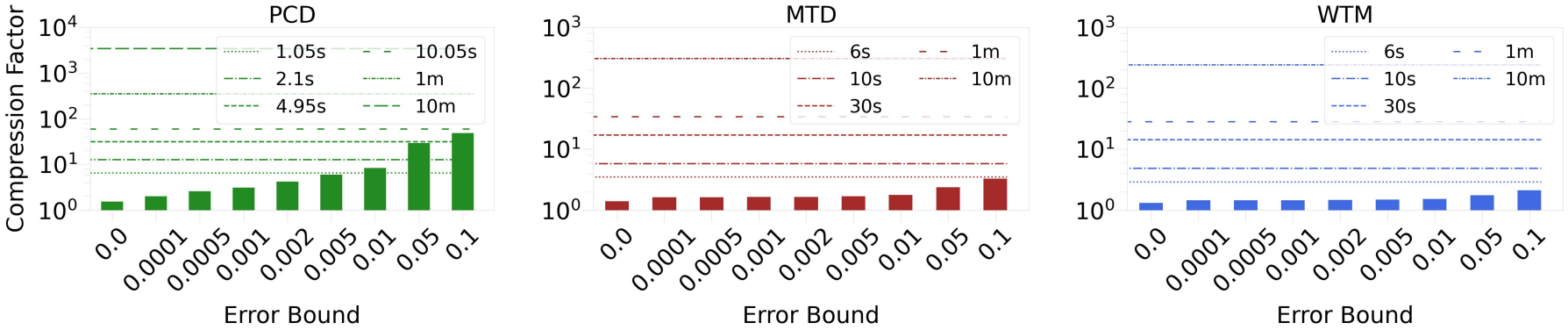
***RQ1.2**: How does the SI of a high-frequency wind turbine dataset affect EBLC?*



***Answer**:*

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

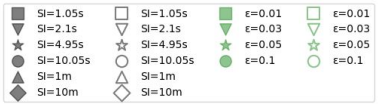
***RQ1.3**: How does EBLC compare against the baseline lossy compression (aggregation)?*



Answer:

- *For PCD, 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.*

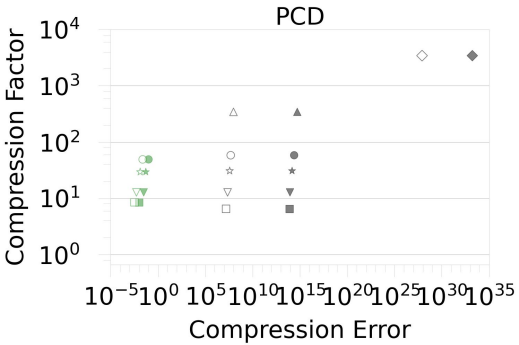
RQ1.3: How does EBLC compare against the baseline lossy compression (aggregation)?

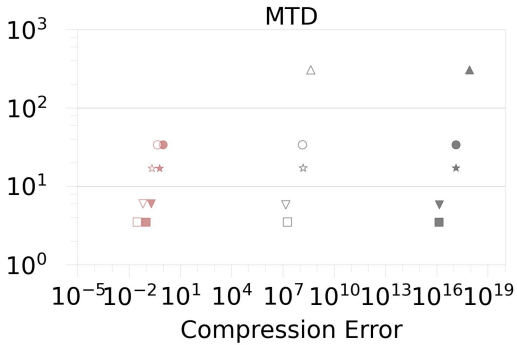


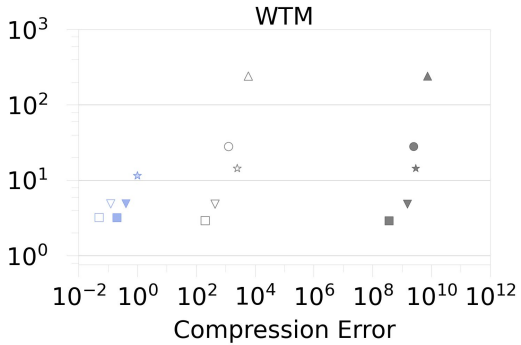
Filled: maximum error; Hollow: mean error; Green: EBLC; Black: Baseline Lossy Compression (Aggregation).

Filled: maximum error; Hollow: mean error; Brown: EBLC; Black: Baseline Lossy Compression (Aggregation).

Filled: maximum error; Hollow: mean error; Blue: EBLC; Black: Baseline Lossy Compression (Aggregation).







Answer:

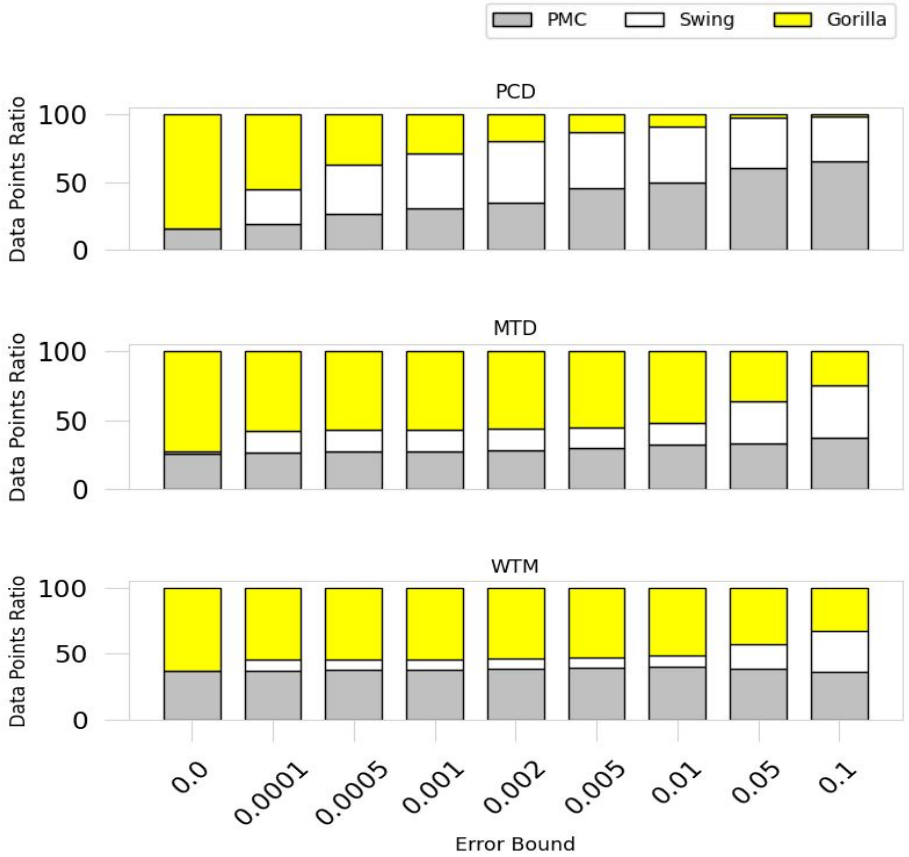
- *EBLC adds many orders of magnitude less error than aggregation.*

***RQ2**: How are model types used for the different aspects?*

***RQ2.1**: What is the distribution of model types?*

***Answer**:*

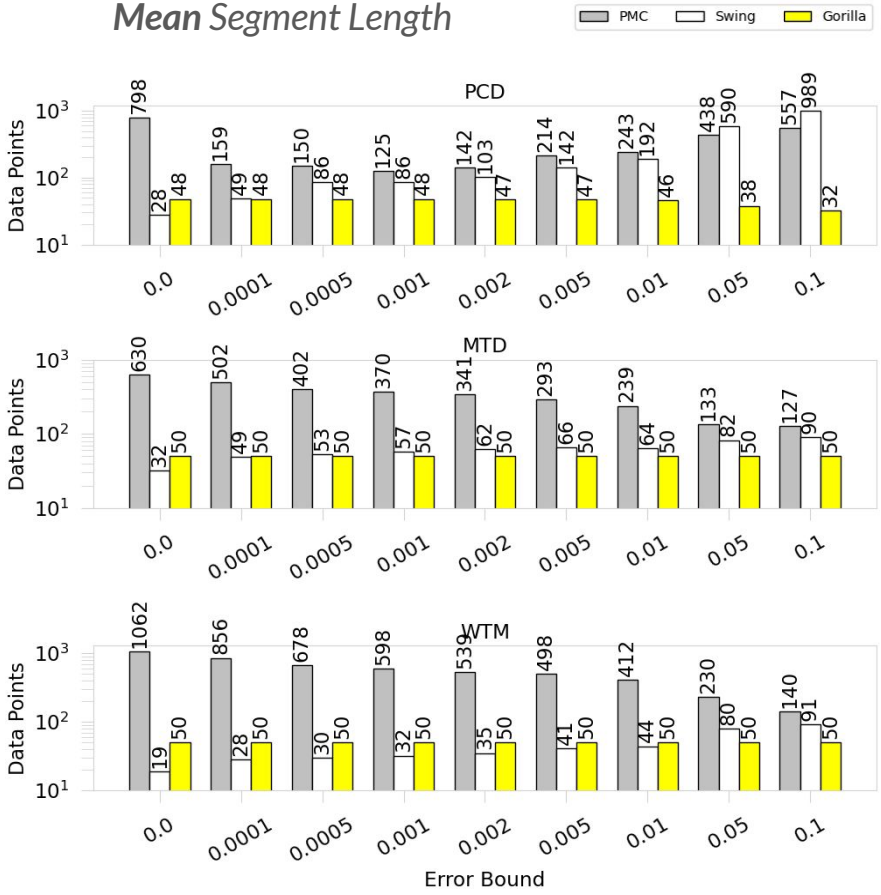
- *At ε=0%, Gorilla is the most used model type for all datasets.*
- *At ε>0% us of PMC and Swing increase for all datasets.*



***RQ2.2**: What is the length of segments for each model type?*

***Answer**:*

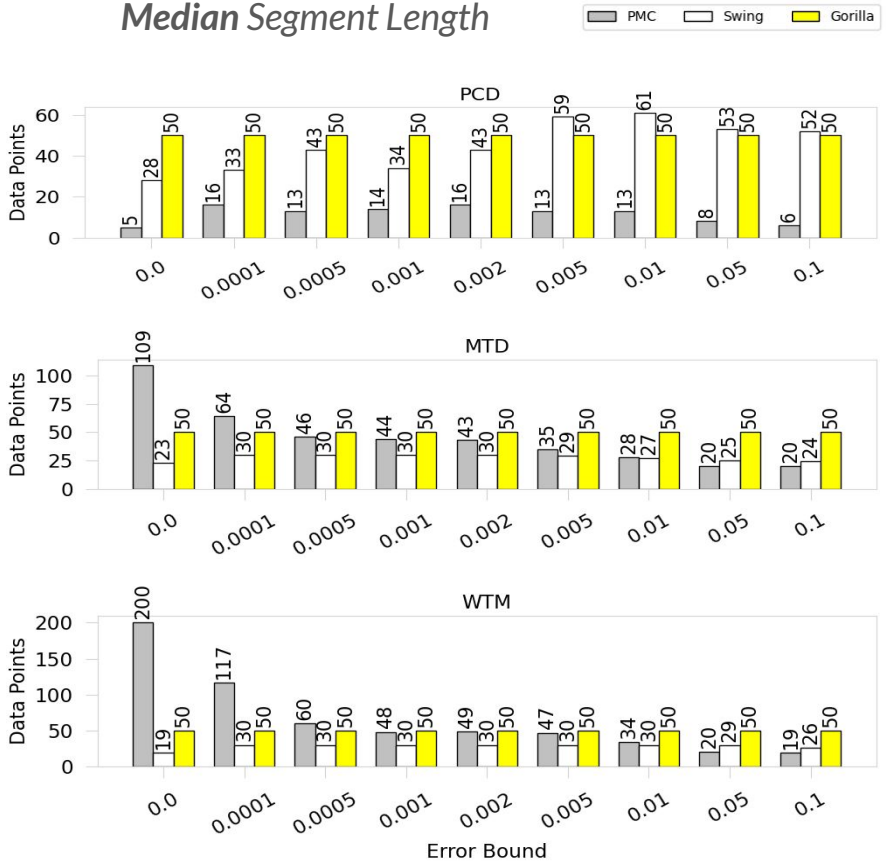
- *The segment lengths of model types depend on SI and **ε**.*
- *PMC has higher mean length than other model types.*
- *Swing creates segments with shorter mean length than PMC.*



***RQ2.2**: What is the length of segments for each model type?*

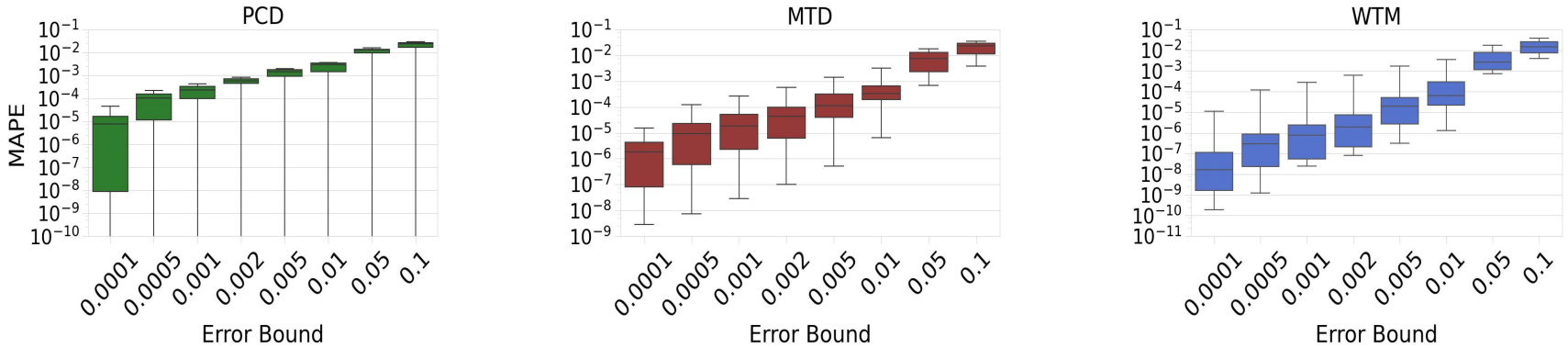
***Answer**:*

- ***Increasing mean** and **decreasing median** segment length mean better compression performance*
- *Both the mean and median length of Gorilla segments are always similar*



***RQ3**: How does EBLC affect the data quality of a decompressed high-frequency wind turbine dataset?*

***RQ3.1**: What is the compression error of a high-frequency wind turbine dataset when compressed using EBLC?*



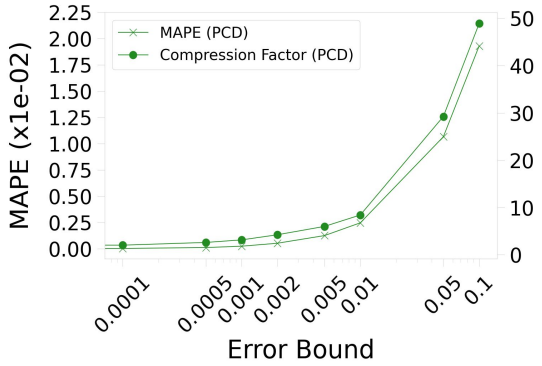
***Answer**:*

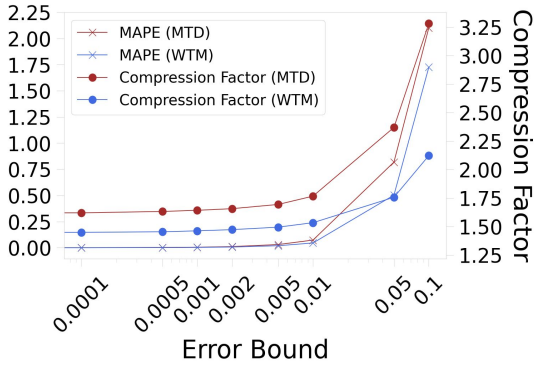
- *For PCD, MTD and WTM, the highest MAPE for an individual column reaches 0.47x, 0.36x and 0.38x of **ε**.*

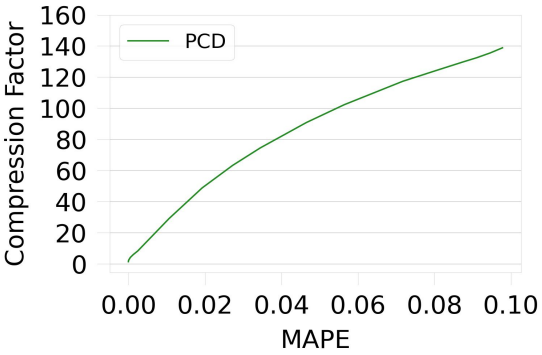
RQ4: *What is the relationship between the data quality and compression effectiveness for a high-frequency wind turbine dataset compressed using EBLC?*

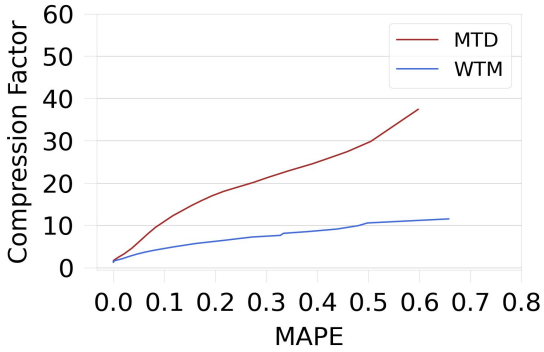
Answer:

- *MAPE and CF have a positive sublinear relationship*
- *PCD reaches around 24 increase in CF for every 1% MAPE*
- *MTD and WTD achieve ~0.8 and ~0.4 CF per 1% MAPE*









Conclusion

The paper gives insight into:

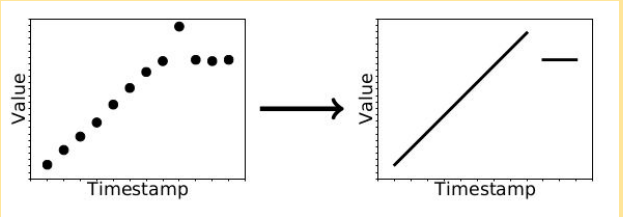
- EBLC's compression effectiveness on high-frequency wind turbine datasets.
- Detailed comparison of EBLC against other compression methods practiced in the industry.
- Aspects influencing the compression effectiveness and data quality of EBLC.
- Trade-off between EBLC's compression effectiveness and data quality.

Secondment

- Visited SGRE office in Brande, Denmark
- Presented research findings and made further plans with secondment supervisor
- Secondment continues until the end of PhD
- Blend of physical and virtual meetings

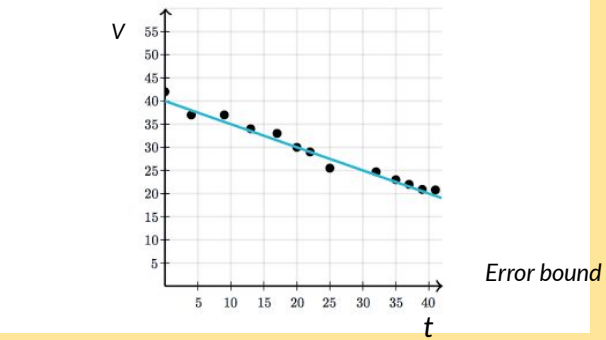
Model based storage of time series data

- Individual time series can be described with models:
- E.g., $v = a * t + b$ can represent a sub-sequence using only a and b.
- Correlated time series could also be grouped and stored as a single model



Goals:

- Approximate the time series values using mathematical functions (models) and store only model coefficients
- Group correlated time series and store as a single model



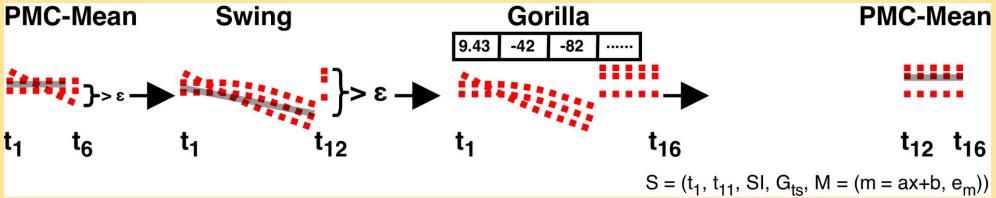
Error bound = 10 %

ModelarDB

- Open source time series database developed at AUU
- Two implementations: **ModelarDB (JVM version)** and **ModelarDB-RS (Rust-based)**

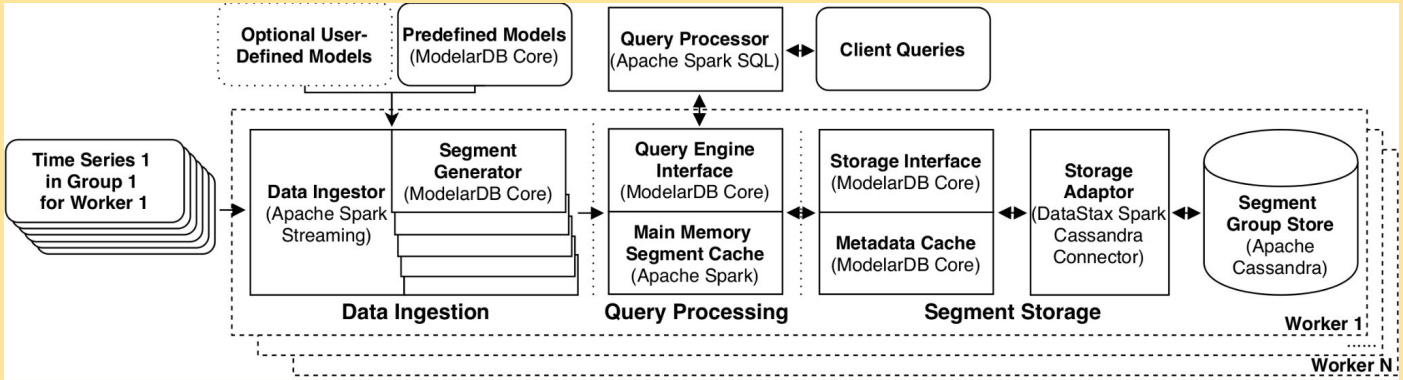
ModelarDB JVM:

- Uses Apache Cassandra (other storage options are also available) for storage and Apache Spark/H2 for query processing.
- Currently includes three different model types:

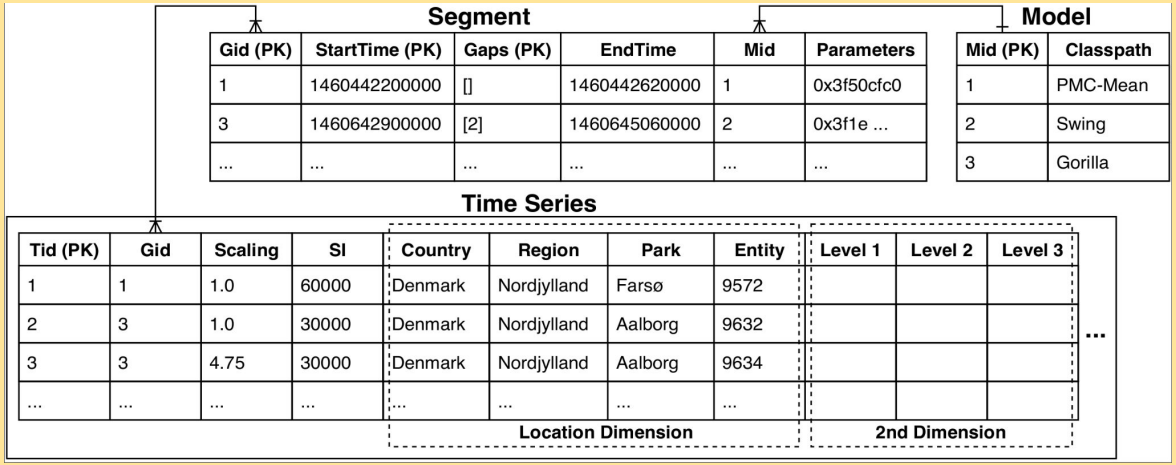


ModelarDB Architecture

- ModelarDB is a portable Java library (ModelarDB Core) interfaced with a query engine (Apache Spark/H2) and storage (Apache Cassandra/ORC, Parquet file formats/RDBMS).
- The architecture of a worker consists of three sets of components: Data Ingestion, Query Processing, and Segment Storage.



ModelarDB segment structure



Segment

Gid (PK)	StartTime (PK)	Gaps (PK)	EndTime	Mid	Parameters
1	1460442200000	[]	1460442620000	1	0x3f50cfc0
3	1460642900000	[2]	1460645060000	2	0x3f1e ...
...	...	...	...	...	...

Model

Mid (PK)	Classpath
1	PMC-Mean
2	Swing
3	Gorilla

Time Series

Tid (PK)	Gid	Scaling	SI	Country	Region	Park	Entity	Level 1	Level 2	Level 3
1	1	1.0	60000	Denmark	Nordjylland	Farsø	9572			
2	3	1.0	30000	Denmark	Nordjylland	Aalborg	9632			
3	3	4.75	30000	Denmark	Nordjylland	Aalborg	9634			
...	...	...	...	...	...	...	...			

Location Dimension

2nd Dimension

- Time Series and Model store metadata for time series and model types.
- Segment stores sub-sequences of time series as segments with a model.

Intended PhD papers

Paper 1: A tool for analysis of the efficiency of model-based compression in ModelarDB

- Analysis of the eciency of current model types deployed by ModelarDB.
- Integrated tool that explains the system performance and its usage of model types.
- Performance indicators and visualization.

Paper 2: New model types to achieve better compression rate and lower error bound for ModelarDB.

Paper 3: Automatic grouping of time series by deploying correlation statistics in ModelarDB

Paper 4: Adding dynamic sampling intervals and error bounds for time series ingestion of ModelarDB.

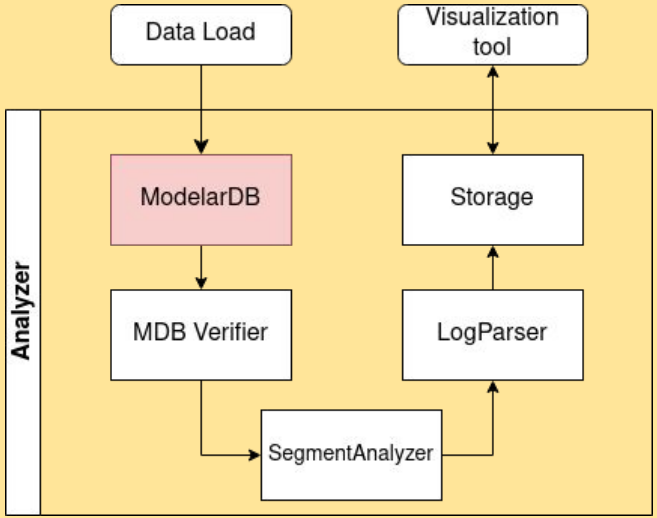
Current status

Tentative Title of Paper 1: A tool for analysis of the efficiency of model-based compression in ModelarDB

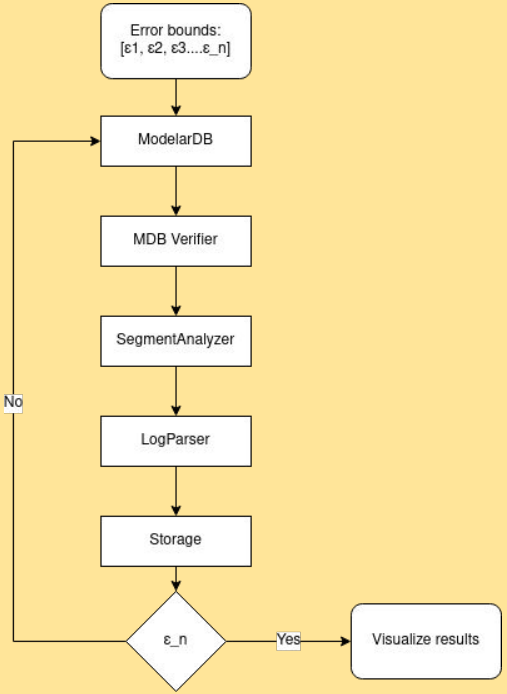
Tasks performed:

- Active collaboration with SGRE
- Python “Analyzer” tool
- Technical report with SGRE
- Scheduled visit to SGRE next week

Analyzer



Components of Analyzer

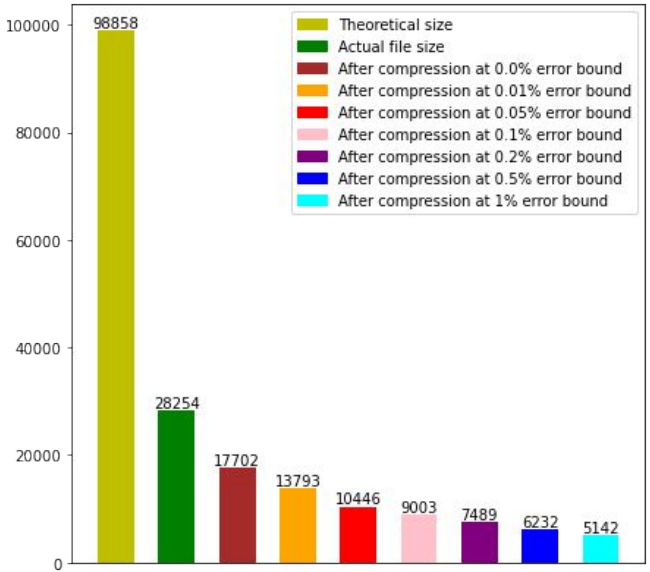


Analyzer running process

SOME RESULTS

Ingestion of the Full Dataset

- **2.5 years** of data from **Power Controller**
- **~480mln** rows, **46** columns (multivariate time series)
- **ORC file format**
- **150 ms** sampling interval (although not perfectly regular)
- **Error bounds used**: 0%, 0.01%, 0.05%, 0.1%, 0.2%, 0.5% and 1%
- ***Theoretical size (in bytes)*** *= Rows x (Columns x 4B + 8B)*

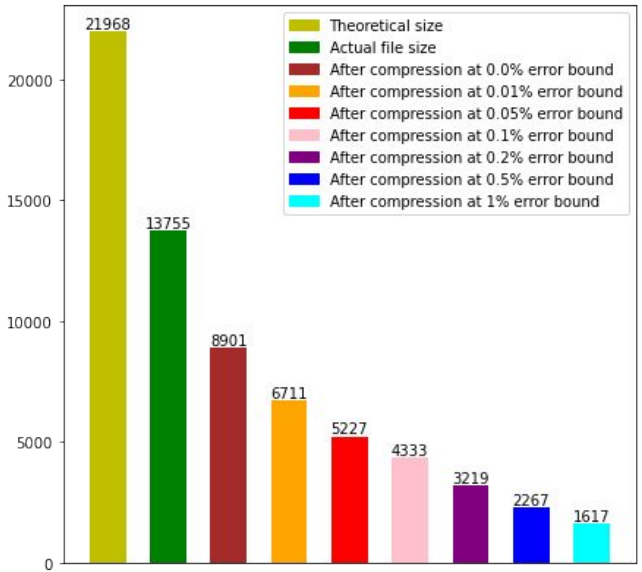


In MB

Ingestion of the Analytics Dataset

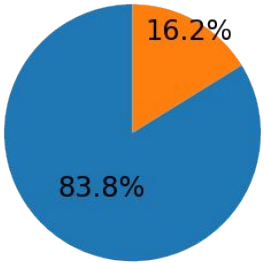
Used columns:

1. TimeStamp
2. ActivePower
3. ActivePower60
4. ActivePower600
5. AvailablePower
6. Frequency
7. PowerError
8. PowerLowerLimit
9. PowerUpperLimit
10. RawPower
11. RawReactivePower

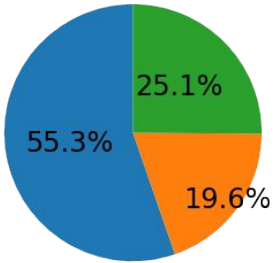


In MB

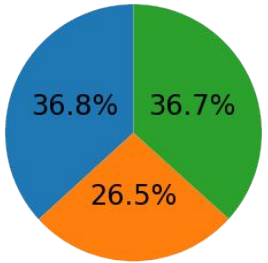
General model performance by data points (Analytics Dataset)



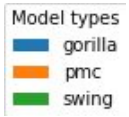
Error bound 0.0%

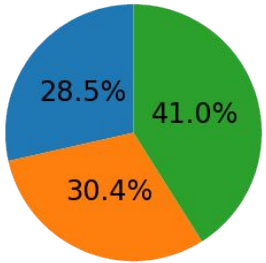


Error bound 0.01%

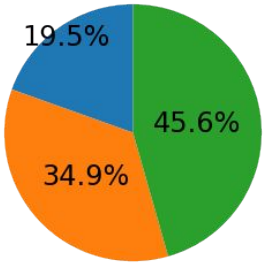


Error bound 0.05%

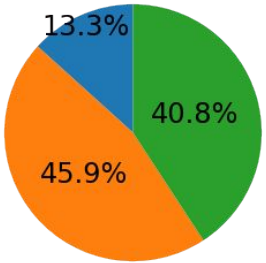




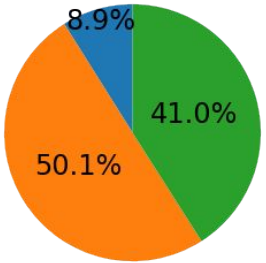
Error bound 0.1%



Error bound 0.2%

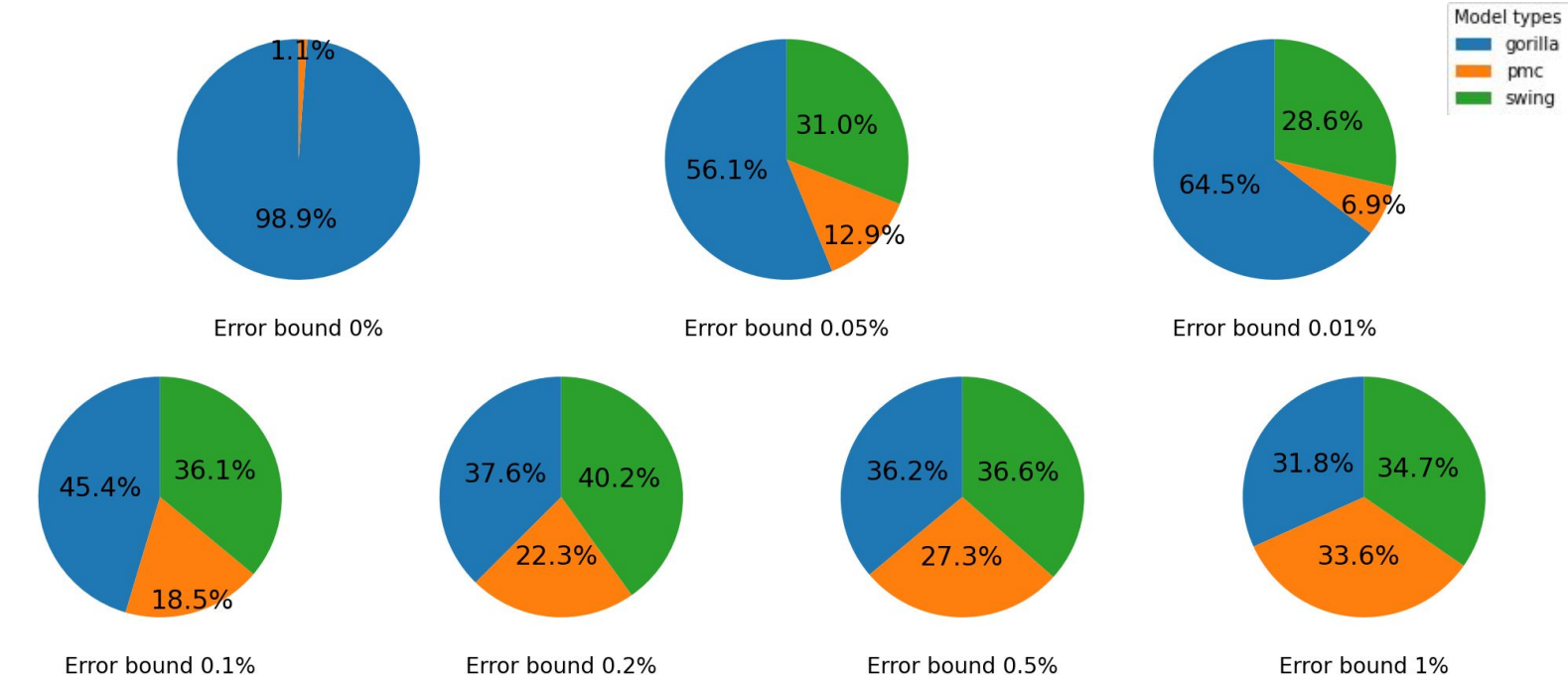


Error bound 0.5%

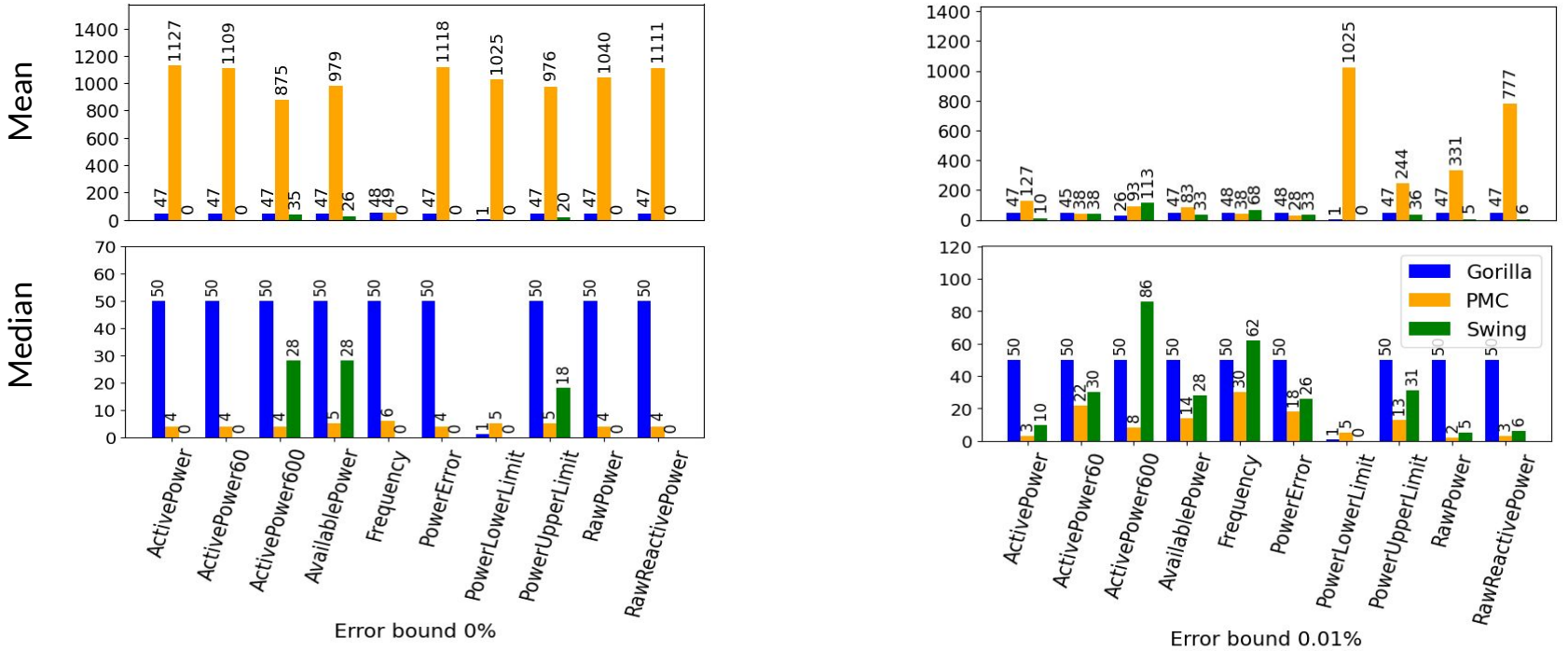


Error bound 1.0%

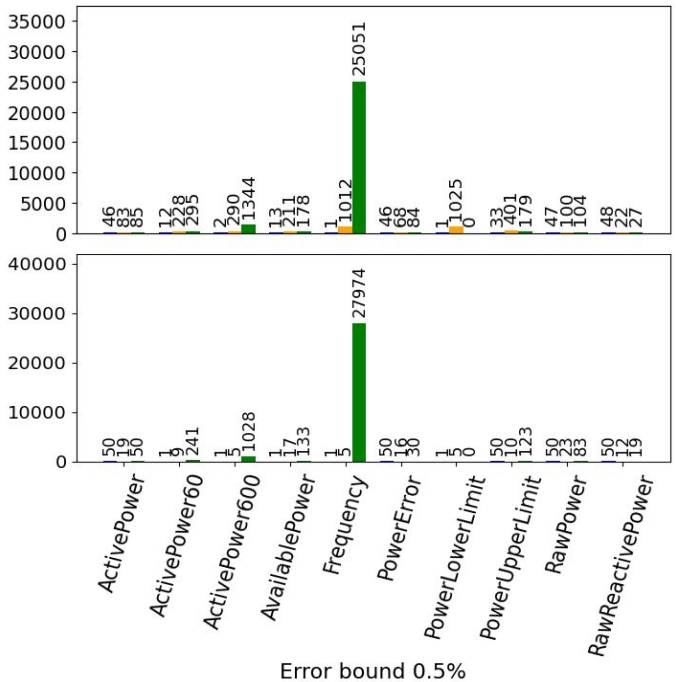
General model performance by segments (Analytics dataset)



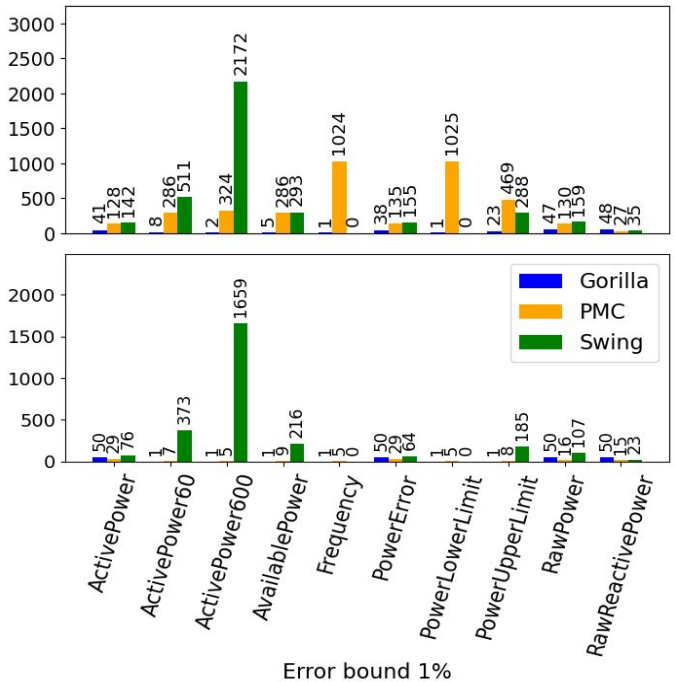
Length of segments



Length of segments

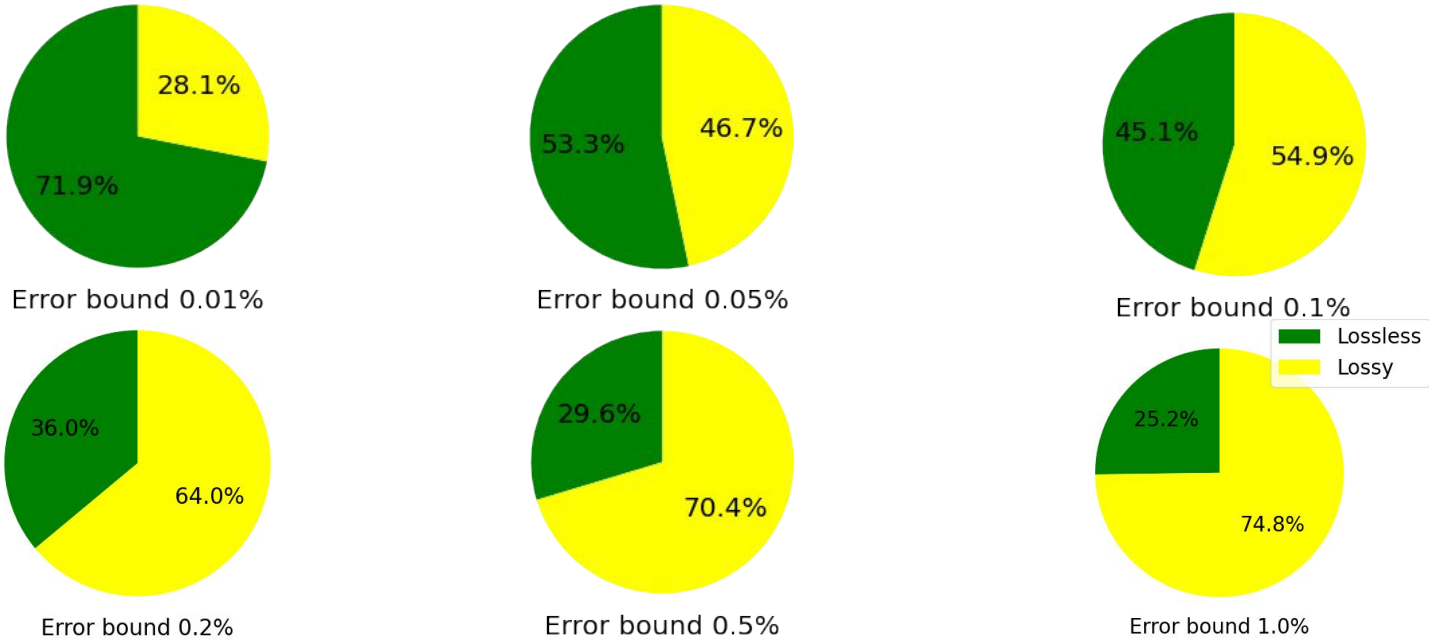


Error bound 0.5%

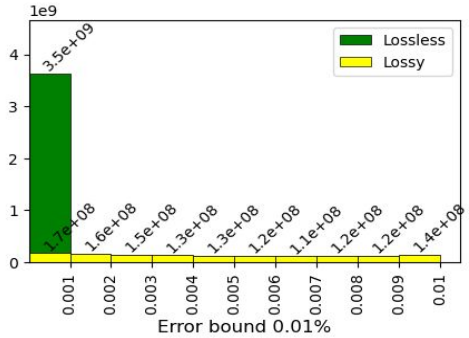


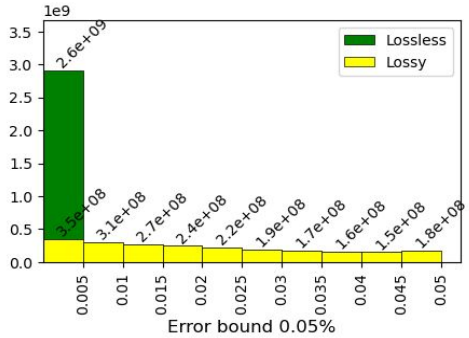
Error bound 1%

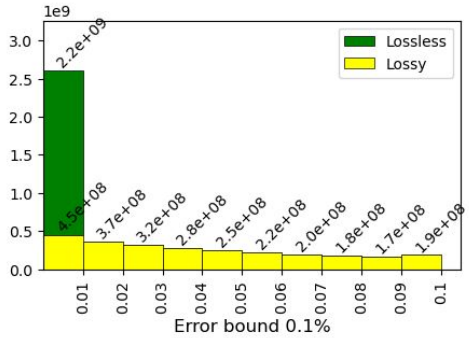
Losslessly compressed data points (Analytics Dataset)

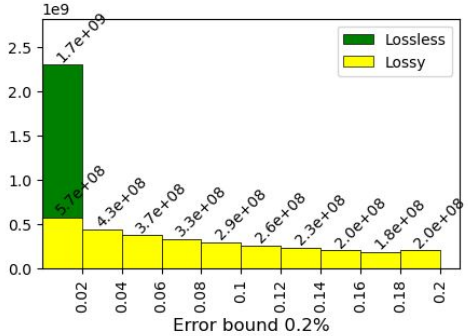


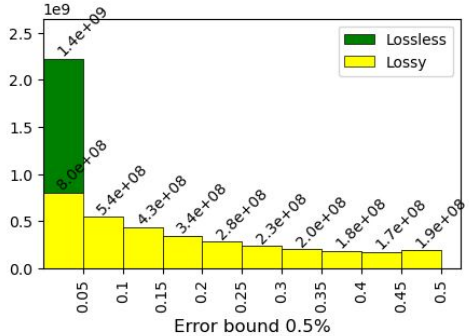
Distribution of actual errors (Analytics dataset)

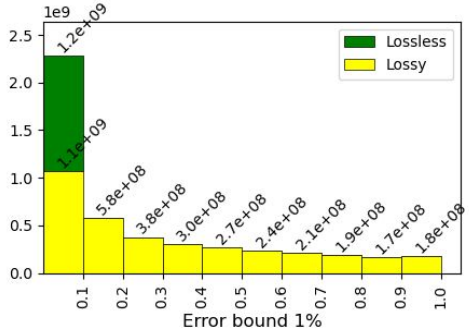






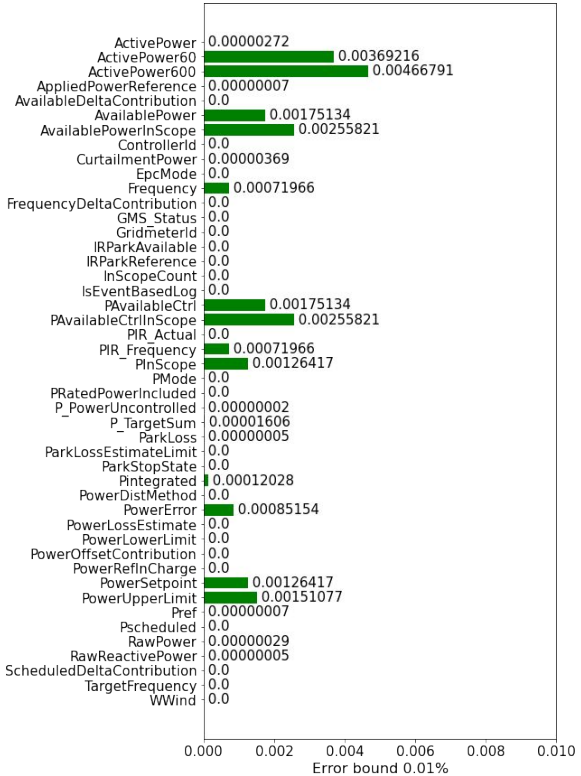




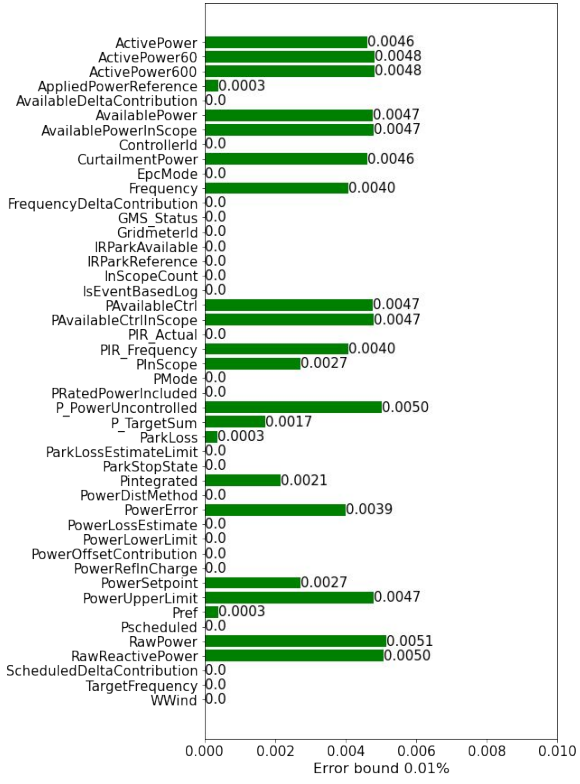


Average actual error (0.01% error bound)

MAPE (Mean Average Percentage Error)

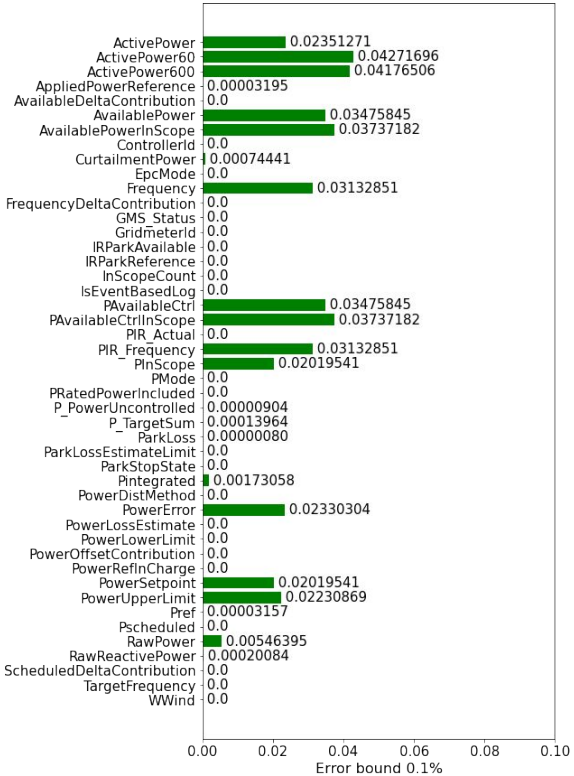


MAPE, |r| > 0

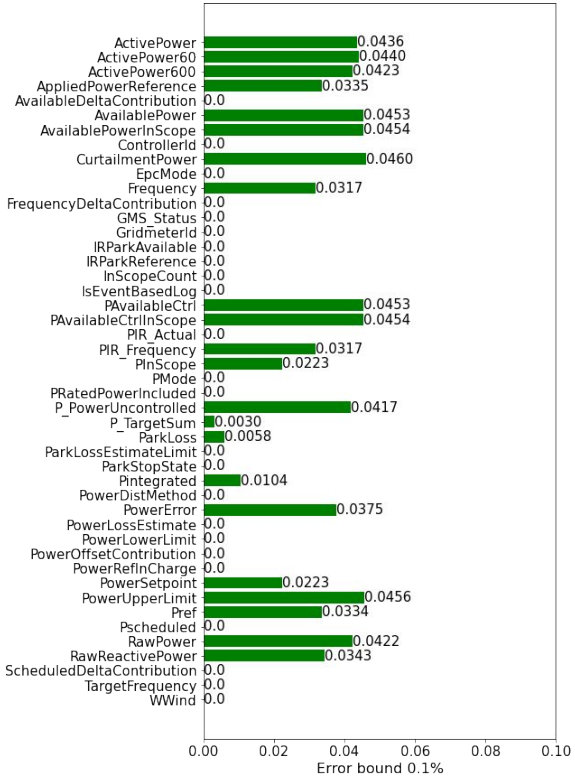


Average actual error (0.1% error bound)

MAPE

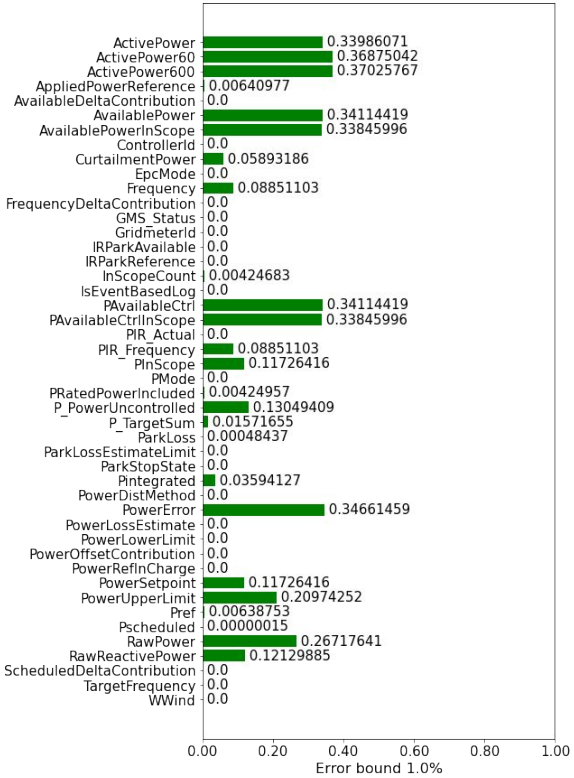


MAPE, |r| > 0

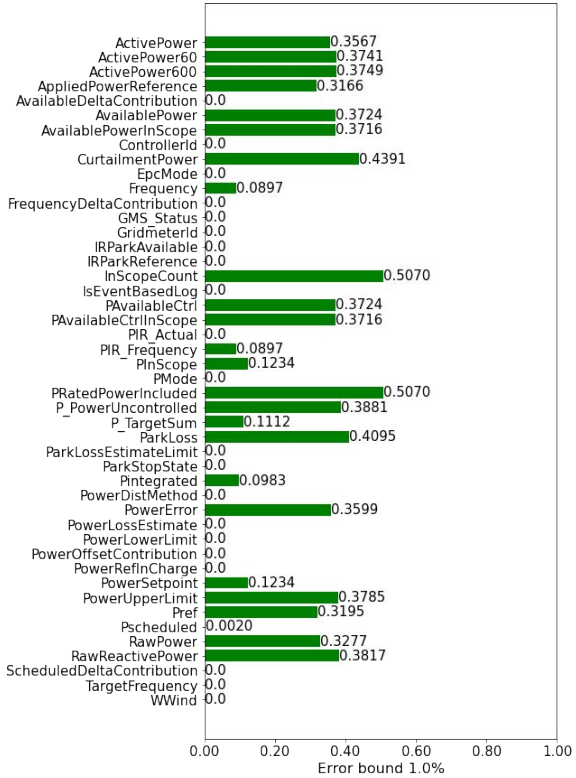


Average actual error (1% error bound)

MAPE



MAPE, |r| > 0



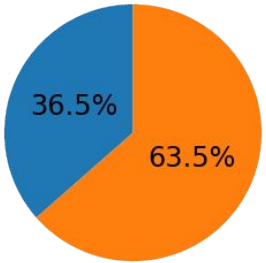
Next steps

- Technical report with SGRE
- Convert it into a paper
- New model types
- Optimization of existing ones
- Constant performance evaluation using Analyzer

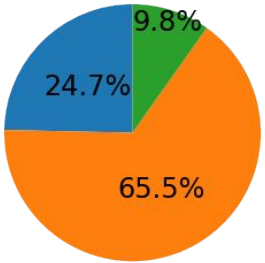
Feedbacks from audience

1. Mention a vision of this paper. What is the goal, targets you set, contribution you make
2. What could be improved in models you use?
3. For actual average error calculation, think about including percentiles to convey a more extensive message

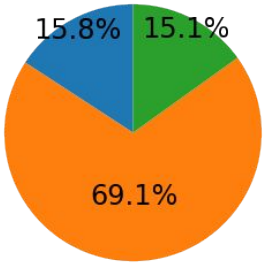
General model performance by data points (Full Dataset)



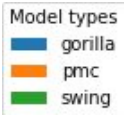
Error bound 0.0%

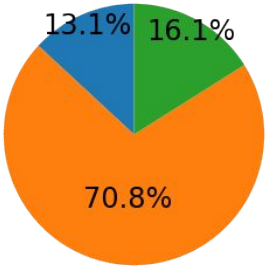


Error bound 0.01%

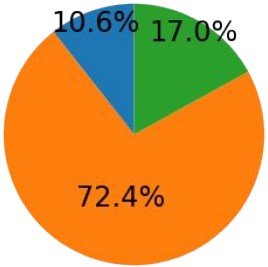


Error bound 0.05%

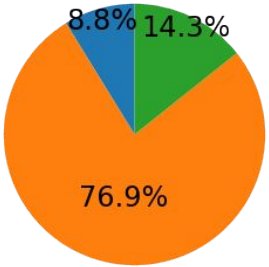




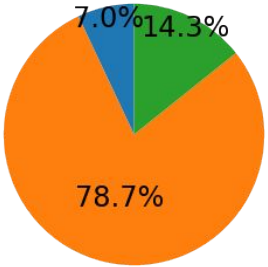
Error bound 0.1%



Error bound 0.2%

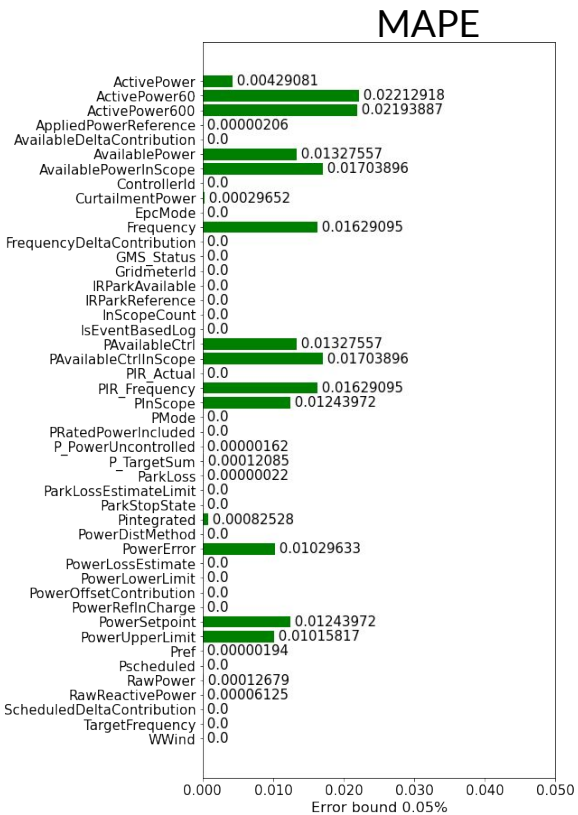


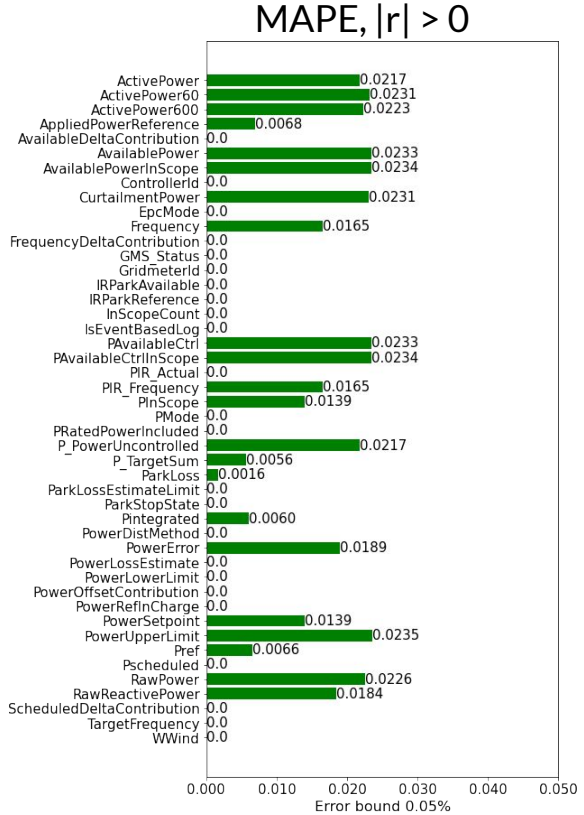
Error bound 0.5%



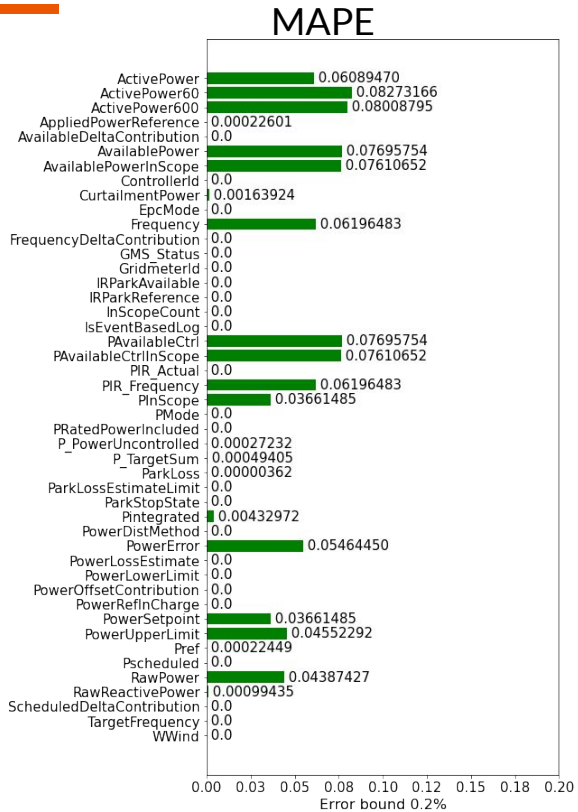
Error bound 1.0%

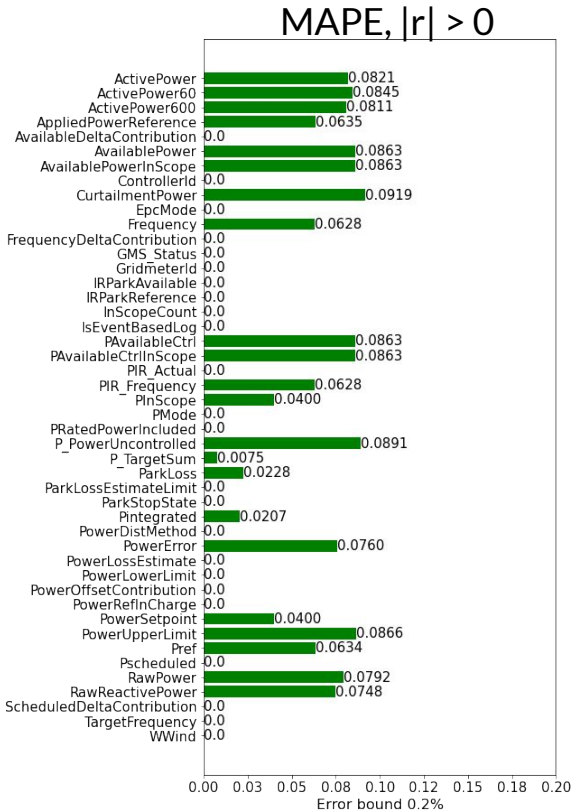
Average actual error (0.05% error bound)





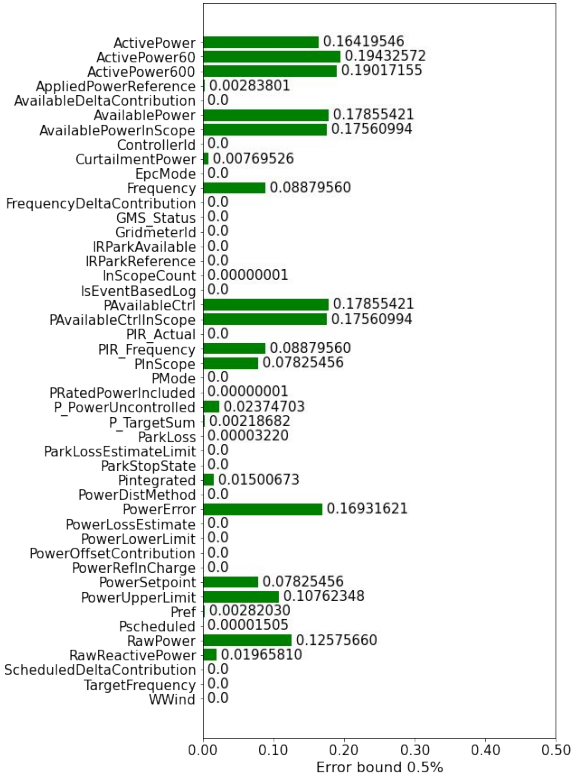
Average actual error (0.2% error bound)



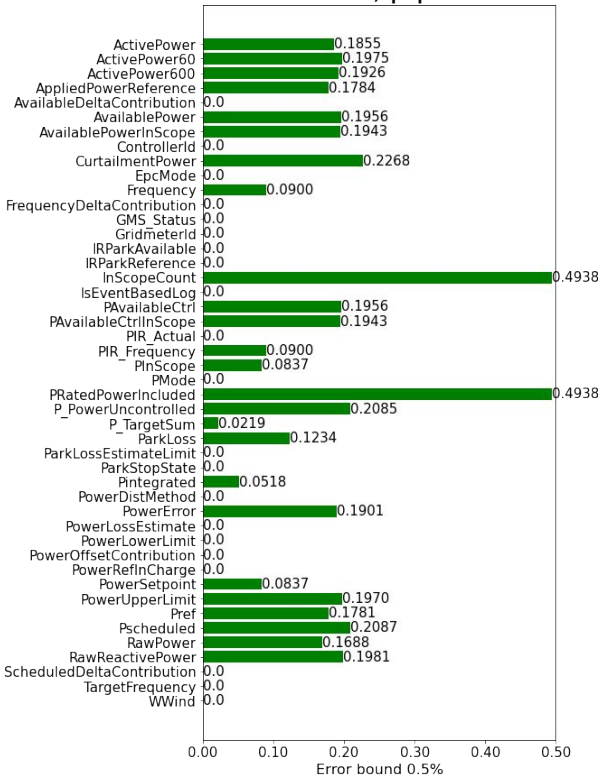


Average actual error (0.5% error bound)

MAPE



MAPE, |r| > 0



Length of segments