RecovDB: accurate and efficient missing blocks recovery for large time series

Ines Arous[1], Mourad Khayati[1], Philippe Cudré-Mauroux[1]
Ying Zhang[2], Martin Kersten[2], Svetlin Stalinlov[2]

[1]{firstname.lastname}@unifr.ch
[2]{firstname.lastname}@monetdbsolutions.com

Goal and contributions

Motivation: Real-world time series (sensor) data often contain missing values. Missing values are harmful to upper-level time series analytics.

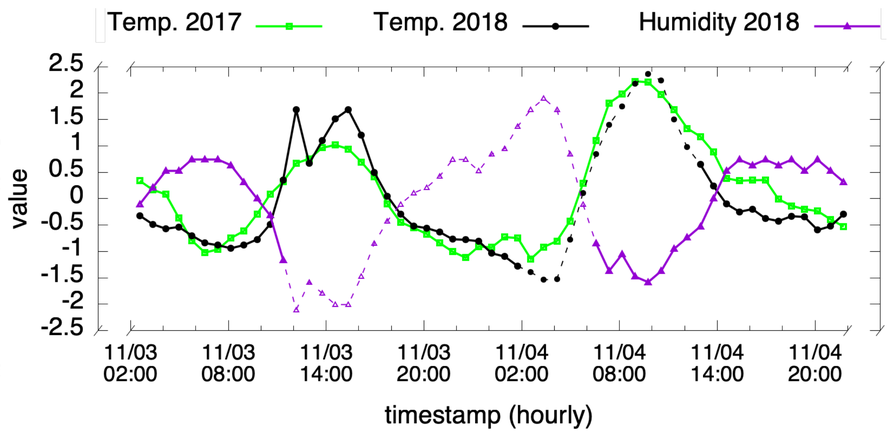
Goal: An efficient and accurate DB system to recover missing values in time series.

Contribution: A system called RecovDB with:

- Parameter-free recovery
- Correlation-aware recovery
- Full-fledged DBMS (MonetDB) support

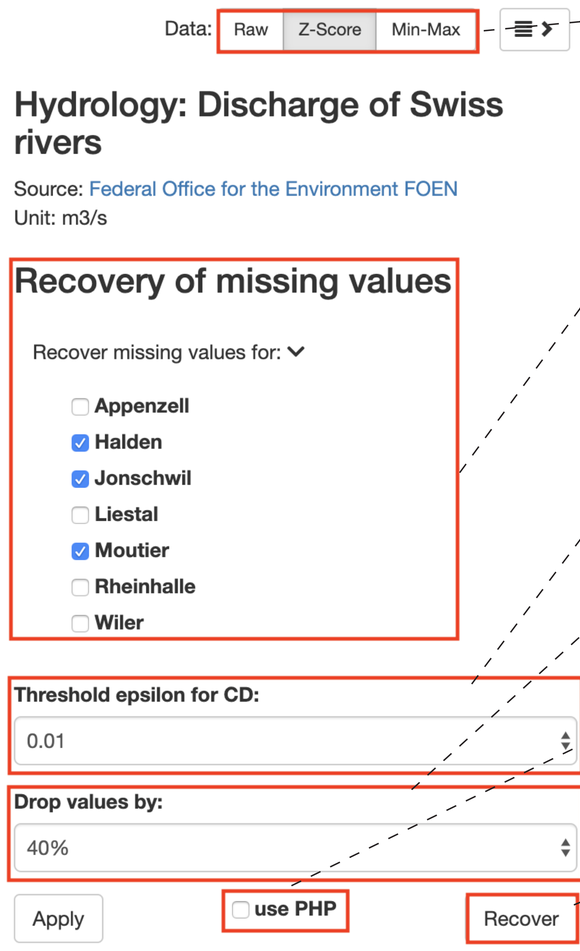
Recovery Challenges

- Long and (linearly) correlated time series.
- Large missing blocks (up to 80% of consecutive observations are missing per time series).
- Multiple incomplete time series.
- Integrate the recovery in the MonetDB system.



Recovery GUI

Menu

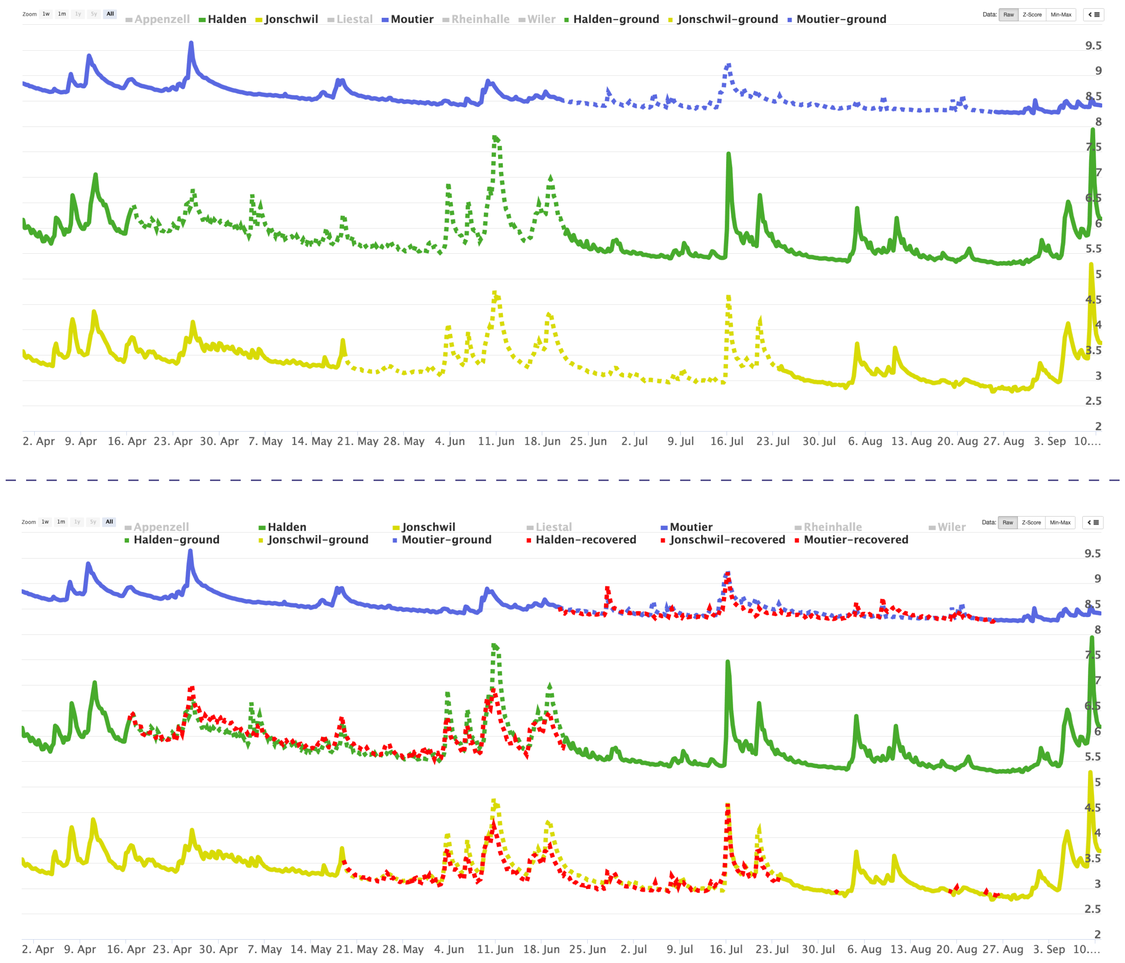


1. Data normalization options.
2. Select one or multiple TS from which to drop blocks of missing values.
3. Select a recovery termination threshold.
4. Select percentage of additional missing values.
5. Use PHP-based recovery instead of UDF.
6. Recover the missing blocks with the selected setup.

- The tool is accessible via `revival.exascale.info`
- The tool can be easily extended with new datasets

Recovery Example

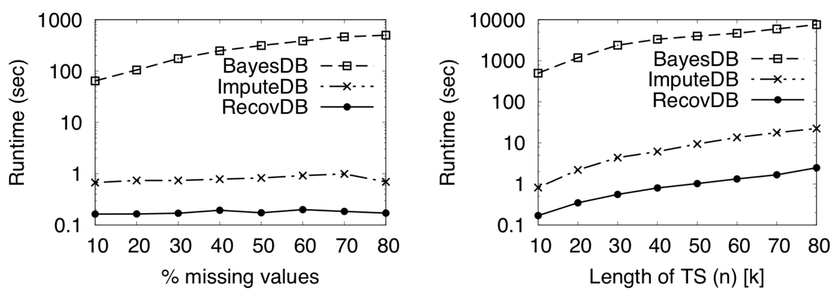
- Missing values are denoted by dashed lines and recovered values are denoted by red lines.
- RecovDB leverages the inter-time series correlation to accurately recover multiple TS in one go.



Empirical evaluation

Efficiency

- RecovDB is up to **10000x** and **10x** faster than BayesDB[a] and ImputeDB[b] respectively.
- To achieve this performance, RecovDB exploits the analytical power of MonetDB to handle the data management and pre-/post-processing.

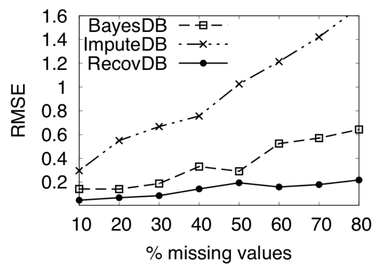


[a]Saad, F. and Mansinghka, V.K. *A probabilistic programming approach to probabilistic data analysis.* NIPS , 2016

[b]Cambronero, J., Feser, J.K., Smith, M.J. and Madden, S. *Query optimization for dynamic imputation* VLDB, 2017

Accuracy

- RecovDB outperforms the STOA using RMSE, MSE and MAE metrics.
- RecovDB is up to **66%** and **87%** more accurate than BayesDB and ImputeDB, respectively.
- The accuracy of RecovDB is steady with increasing % of missing values.



Demo Scenarios

Scenario 1 Recover multiple incomplete time series at one.

Scenario 2 Increase the size of the missing block and of the data.

Scenario 3 Compare RecovDB against STOA recovery DB systems (i.e., BayesDB and ImputeDB).

Conclusions

- We present RecovDB which recovers large missing blocks in multiple time series.
- RecovDB leverages the correlation across time series during the recovery.
- RecovDB outperforms STOA in both efficiency and accuracy when increasing i) the length/number of time series and ii) the size of the missing blocks.

Published in: *Proceedings of the 35th IEEE International Conference on Data Engineering (ICDE 2019)*, Macau, China, April 2019.