Continuous Imputation of Missing Values in Streams of Pattern-Determining Time Series

Kevin Wellenzohn[†] · Michael H. Böhlen[†]
Anton Dignös[‡] · Johann Gamper[‡] · Hannes Mitterer[‡]

† University of Zurich[UZH]
‡ unibz Freie Universität Bozen, Libera Università di Bolzano, Università Liedia de Bulsan

Problem

Problem. Streaming time series often have **missing values**, e.g. due to sensor failures or transmission delays!

Goal. Accurately **impute** (i.e. recover) the latest measurement by exploiting the **correlation** among streams.

Challenge. Streaming time series are often **non-linearly correlated**, e.g. due to **phase shifts**.

Approach (TKCM)

Intuition. Impute a missing value in time series s with past values from s when a set of correlated **reference time series** exhibited similar **patterns**.

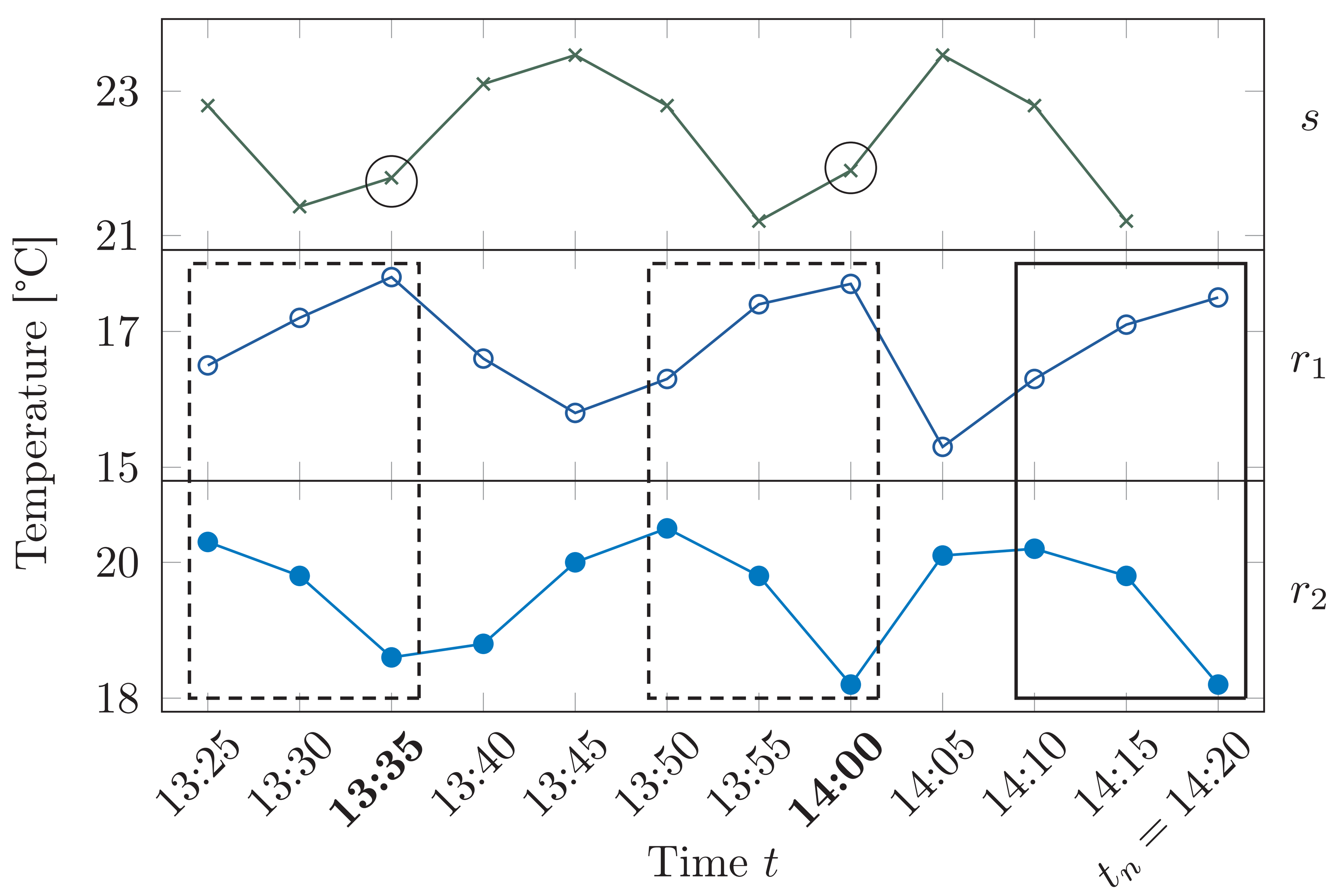
Top-k Case Matching (TKCM). To impute a missing value in a time series s at time t_n:

1. Define query pattern $P(t_n)$, spanning the values of d reference time series over a time frame of l time points anchored at time t_n
2. Look for the k most similar non-overlapping patterns in a sliding window over the time series
3. Impute the missing value $s(t_n)$ as the average of the values of s at the anchor time points of the k previously found patterns

Parameter l (called the "**pattern length**") enables TKCM to deal with non-linearly correlated time series, e.g. phase-shifted time series.

Application Scenario

Consider the set $\{s, r_1, r_2\}$ of streaming time series obtained from a sensor network. Time series s has a **missing value** at current time $t_n = 14{:}20$ that is **imputed** using the $d = 2$ **reference time series** r_1 and r_2.

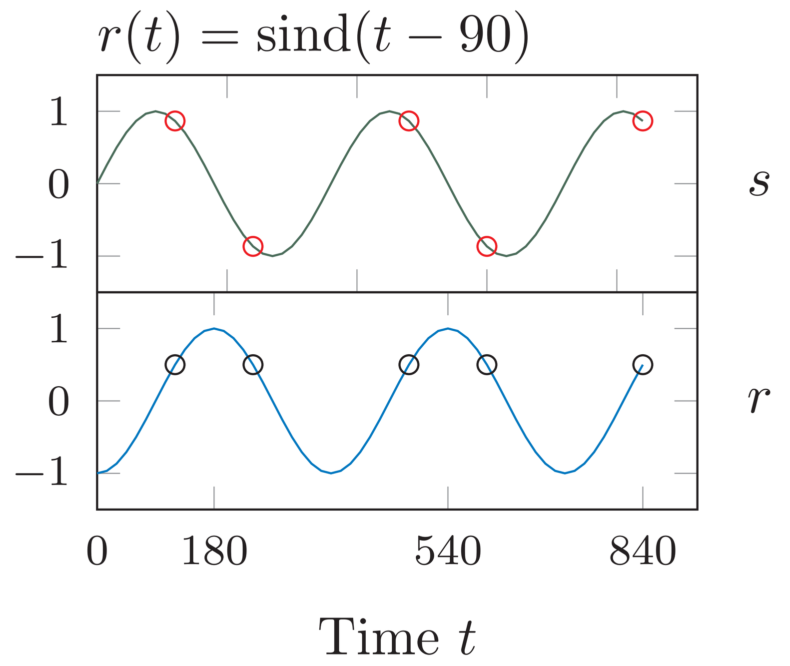


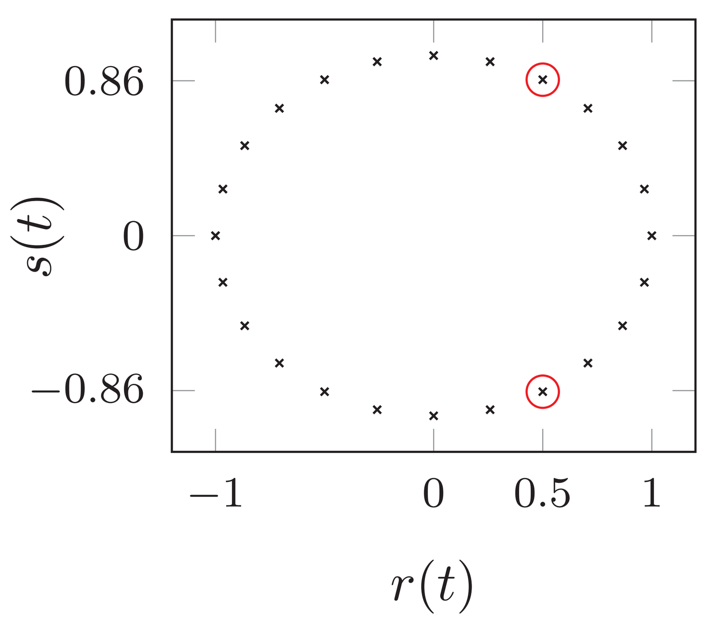
Applying TKCM.

1. Define query pattern $P(t_n) = P(14{:}20)$ over reference time series $\{r_1, r_2\}$ in a time frame of $l = 10$ minutes
2. The $k = 2$ most similar non-overlapping patterns are $P(14{:}00)$ and $P(13{:}35)$
3. Missing value is imputed as $\hat{s}(14{:}20) = \frac{1}{2}(s(14{:}00) + s(13{:}35)) = 21.85$°C

Phase Shifts

$s(t) = \text{sind}(t)$
$r(t) = \text{sind}(t - 90)$





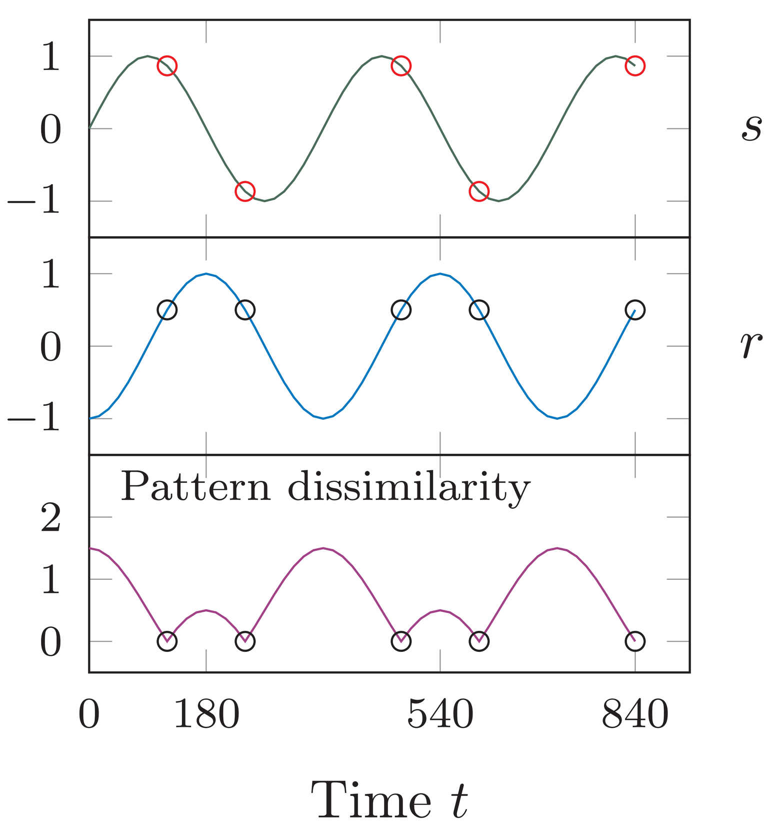
- Time series s and r are **phase-shifted** by 90 degrees
- The scatterplot shows that s and r are **non-linearly correlated**. Their Pearson Correlation Coefficient is 0!
- For example, whenever $r(t) = 0.5$, time series s has two different values, either $s(t) = 0.86$ or $s(t) = -0.86$

TKCM & Non-Linear Correlations

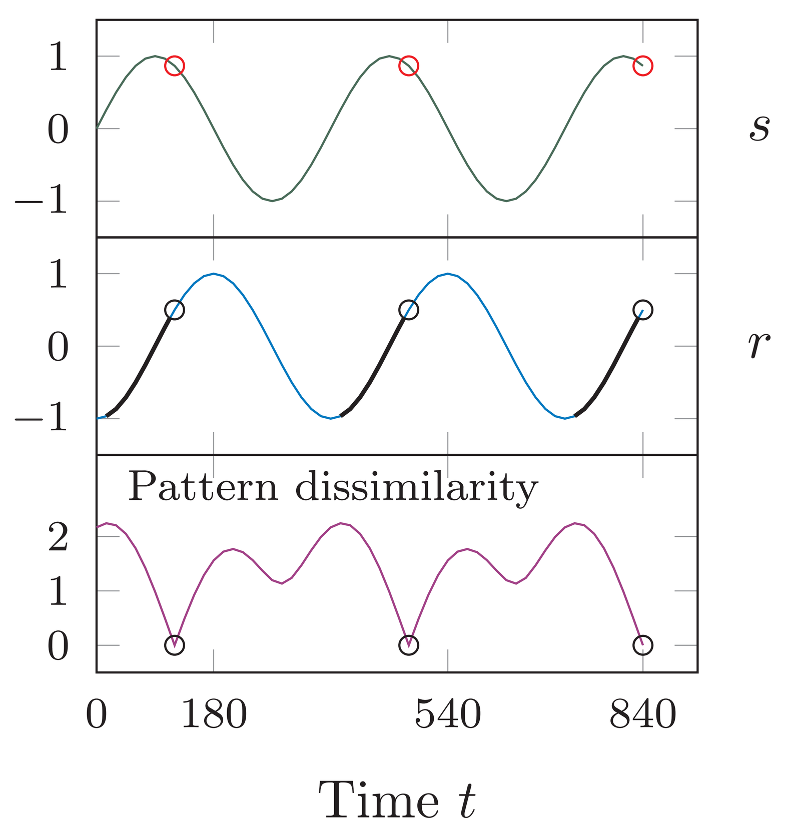
With pattern length $l > 1$,

- TKCM takes the temporal context into account and captures how time series change over time
- there are less patterns with pattern dissimilarity 0

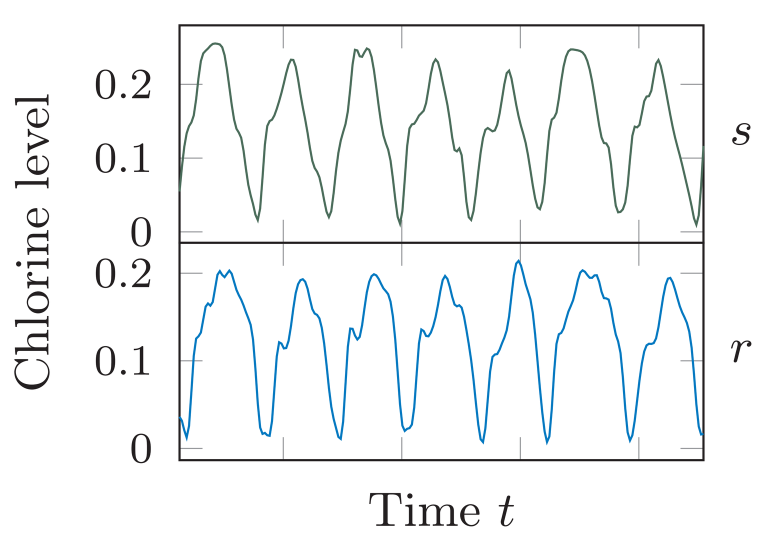
Pattern length $l = 1$

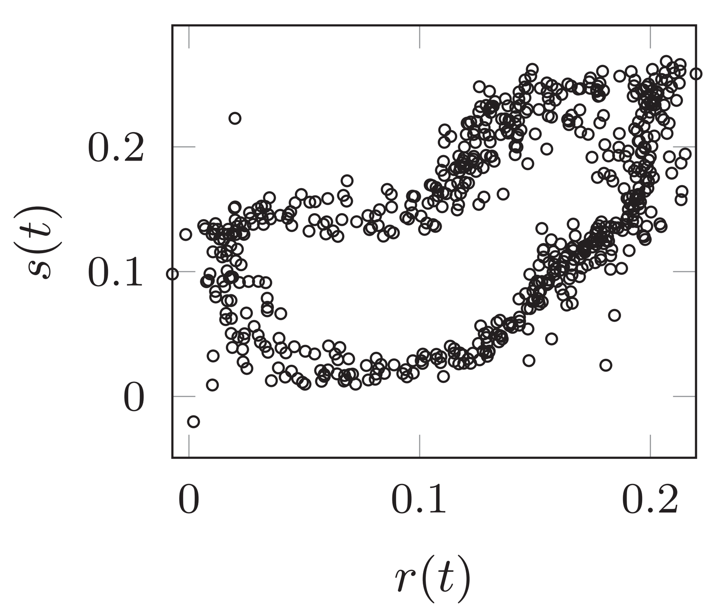


Pattern length $l = 100$

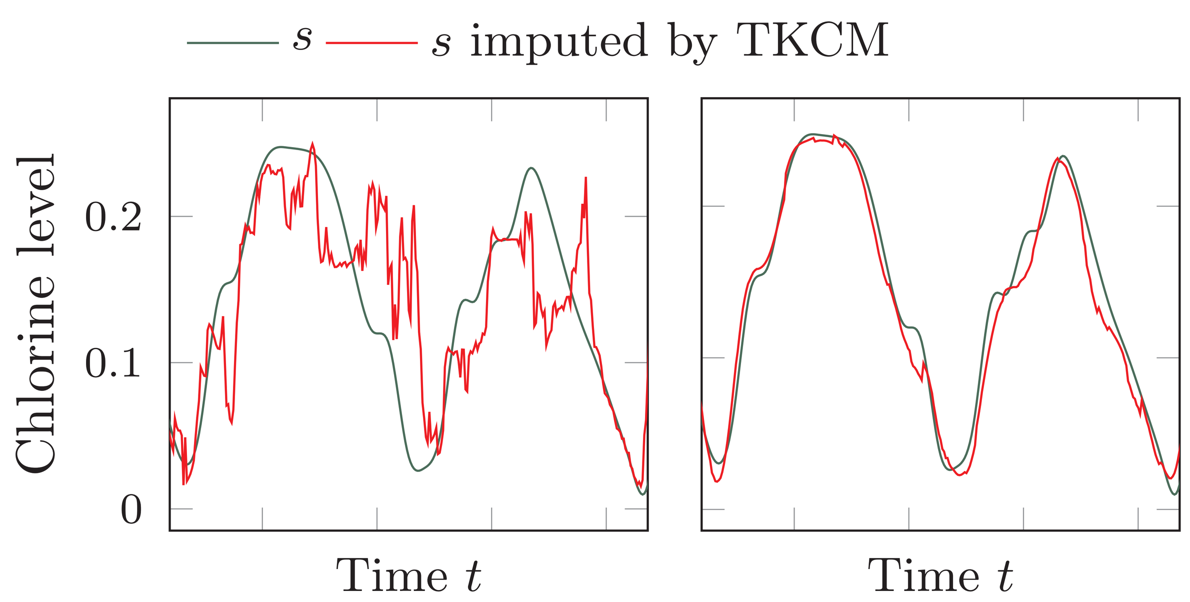


Experiments



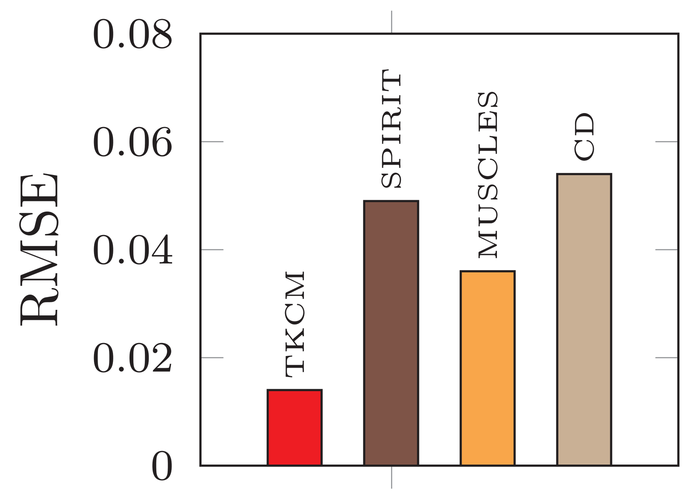


▷ The Chlorine dataset contains phase-shifted and hence non-linearly correlated time series



Pattern length $l = 1$ Pattern length $l = 72$

▷ A larger pattern length decreases the oscillation in the imputed time series



▷ TKCM is more accurate than its competitors for non-linearly correlated time series

Published in: *Proceedings of the 20th International Conference on Extending Database Technology*, Venice, Italy, March 2017.