A CONTINUOUS DATA IMPUTATION MECHANISM BASED ON STREAMS CORRELATION

Panagiotis Fountas, **Kostas Kolomvatsos**

Email: pfountas@uth.gr

Department of Informatics and Telecommunications,
University of Thessaly, Lamia, Greece

The 10th Workshop on Management of Cloud and Smart City Systems

In conjunction with IEEE ISCC 2020

OUTLINE

- Introduction
- Problem Description
- Data Imputation Mechanism
- Experimental Evaluation
- Conclusions & Future Work

INTRODUCTION

Numerous Devices	☑	The increased adoption of Internet of Things (IoT). The development of IoT application and usage IoT devices.
Huge Volumes of Data	☑	IoT devices and applications produce or collect data. The data should have appropriate form to draw conclusions
Interaction with the Edge of the Network	☑	The Edge Computing (EC) nodes are placed close to the data sources. Edge Computing perform analytics over distributed data streams. Edge minimize the latency in the provision of responses.
Missing Values	☑	Data streams can be characterized by missing values. The goal is to eliminate the error between the replacement and actual value. The replacement value calculated by the proposed algorithm

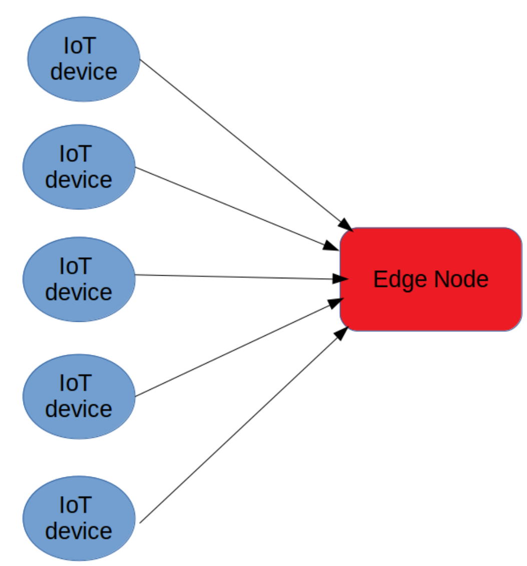
PROBLEM DESCRIPTION

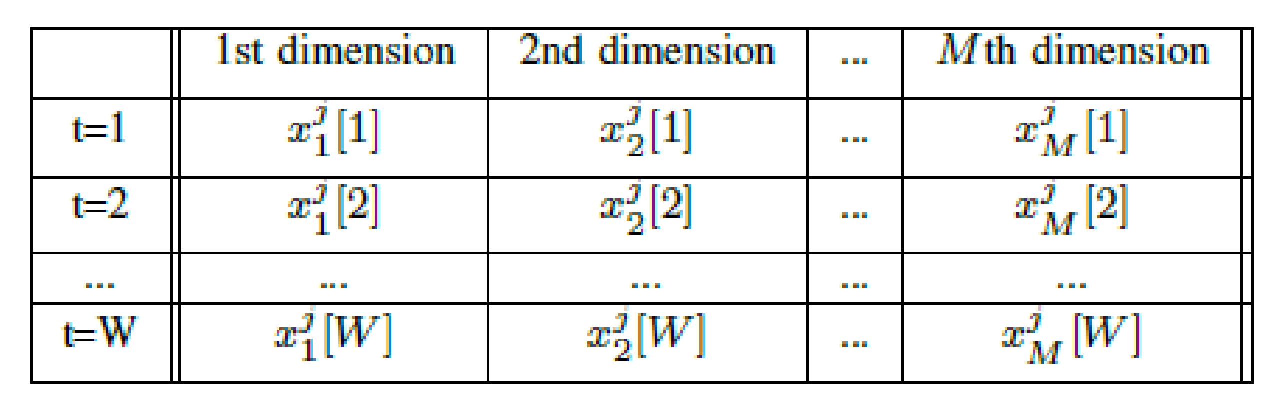
IoT devices /applications collect multivariate data from their environment.

Edge Nodes store the information from IoT devices in the appropriate format.

Edge Nodes use the proposed monitoring mechanism to detect if missing values exist.

Edge Nodes use the proposed imputation mechanism to calculate the replacement value





	1st dimension	2nd dimension	...	Mth dimension
t=1	$x_1^j[1]$	$x_2^j[1]$	...	$x_M^j[1]$
t=2	$x_1^j[2]$	$x_2^j[2]$	...	$x_M^j[2]$
...	...	...	...	...
t=W	$x_1^j[W]$	$x_2^j[W]$	...	$x_M^j[W]$

DATA IMPUTATION MECHANISM

- The Data Imputation Mechanism use metrics i.e. the Cosine Similarity(CS) and the Mahalanobis Distance(MD).
- The CS is applied over the latest reports of IoT devices.
- The Mahalanobis distance is applied over the W latest reports

$$CS((x)^i[t],(x)^j[t]) = \frac{(x)^i[t]\cdot(x)^j[t]}{\|(x)^i[t]\|\cdot\|(x)^j[t]\|} = \frac{\sum_{l=1}^{M}(x)_l^i[t](x)_l^j[t]}{\sqrt{\sum_{l=1}^{M}((x)_l^i[t])^2}\cdot\sqrt{\sum_{l=1}^{M}((x)_l^i[t])^2}}$$

$$MD(\vec{x}-\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}$$

DATA IMPUTATION MECHANISM

- Our ensemble scheme use the previous metrics for the calculations of the final correlation between reports of IoT devices.
- Final correlation (F_C) pays attention on the CS result and uses a weighted model to reward devices with increased correlation with the device detecting the missing value.
- The replacement value is calculated based on top-k F_C values.

$$F_C = w \cdot CS((x)^i[t], (x)^j[t]), \forall i, j, i \neq j$$

$$w = \frac{1}{MD}$$

$$PD = \frac{\sum_{i=1}^{k} MN_i \cdot x_d}{\sum_{l=1}^{k} CS_l}$$

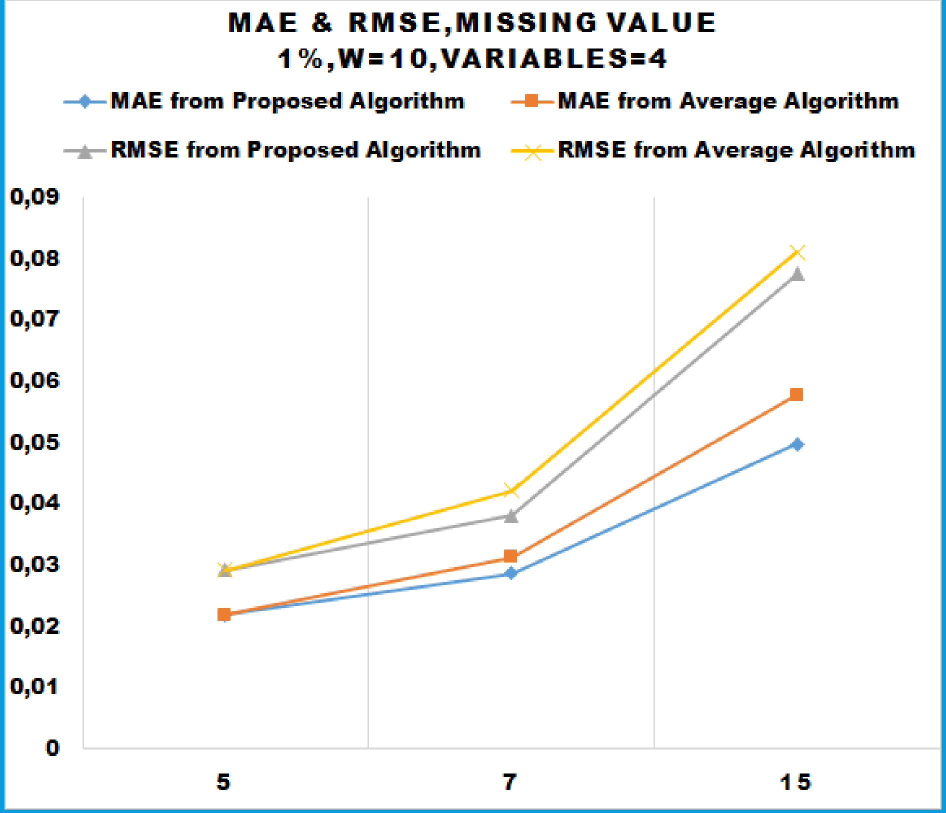
EXPERIMENTAL EVALUATION

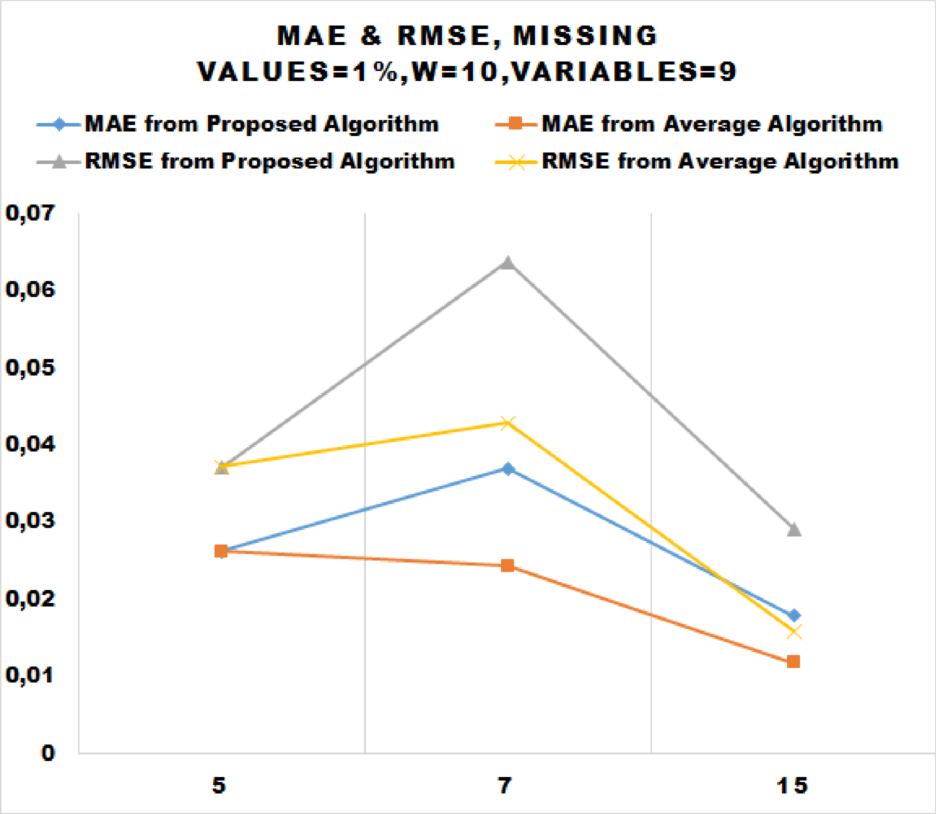
- Metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- In every dataset, we randomly annotate $V\%$ reports as missing values.
- We compare our scheme with a baseline model i.e. Averaging Model (AM) and provide experiments for time requirements.

Parameters
Percentage of missing values in dataset $V \in \{1,5,10\}$
Number of top correlated nodes $k = 4$
Number of IoT devices $\in \{ 5,7,15\}$
Number of Variables $\mathrm{M} \in \{4,9\}$
Number of W latest reports $W = 10$

Dataset	Source
GNFUV Unmanned Surface Vehicles Sensor Data Set	https://archive.ics.uci.edu/ml/datasets/GNFUV+Unmanned+Surface+Vehicles+Sensor+Data+Set+2
Intel Berkeley Research Lab dataset	http://db.csail.mit.edu/labdata/labdata.html

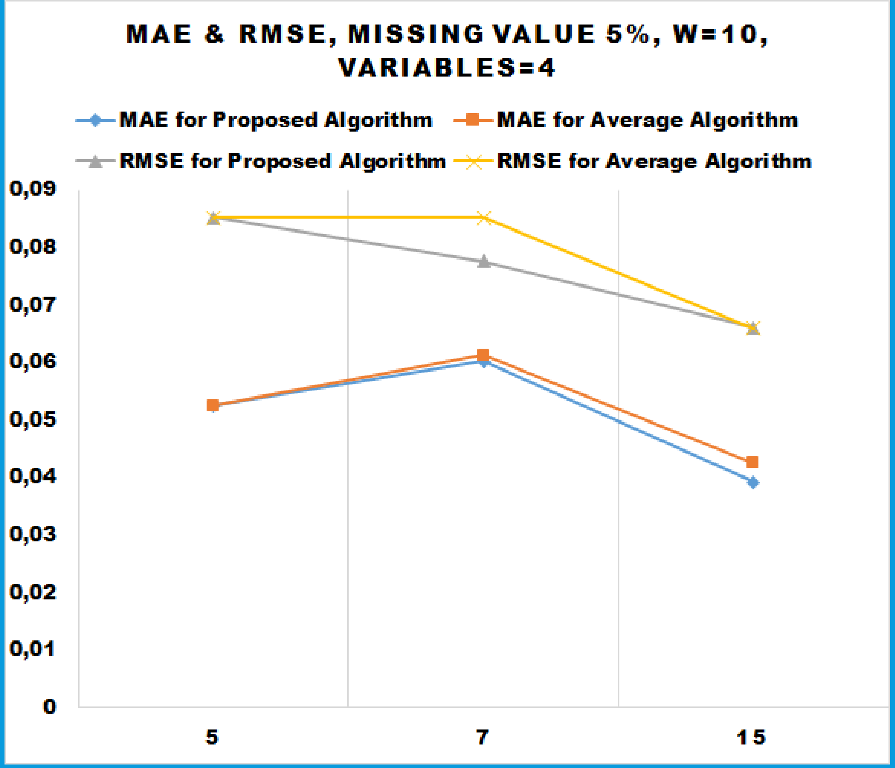
EXPERIMENTAL EVALUATION

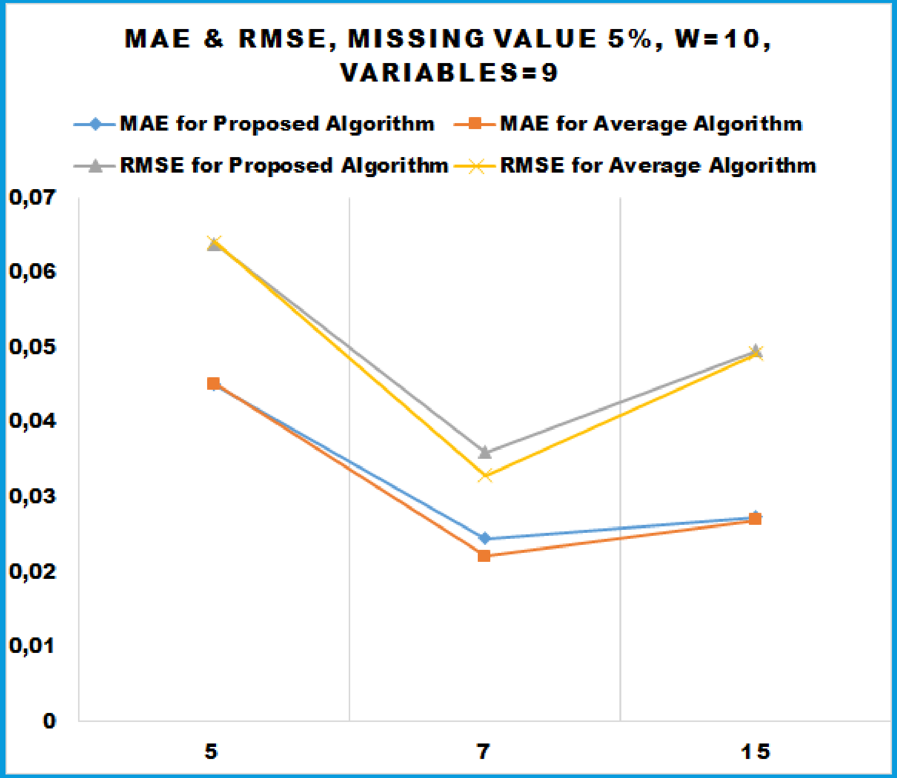




A high number of dimensions does not negatively affect our model that is capable of efficiently replacing missing values.

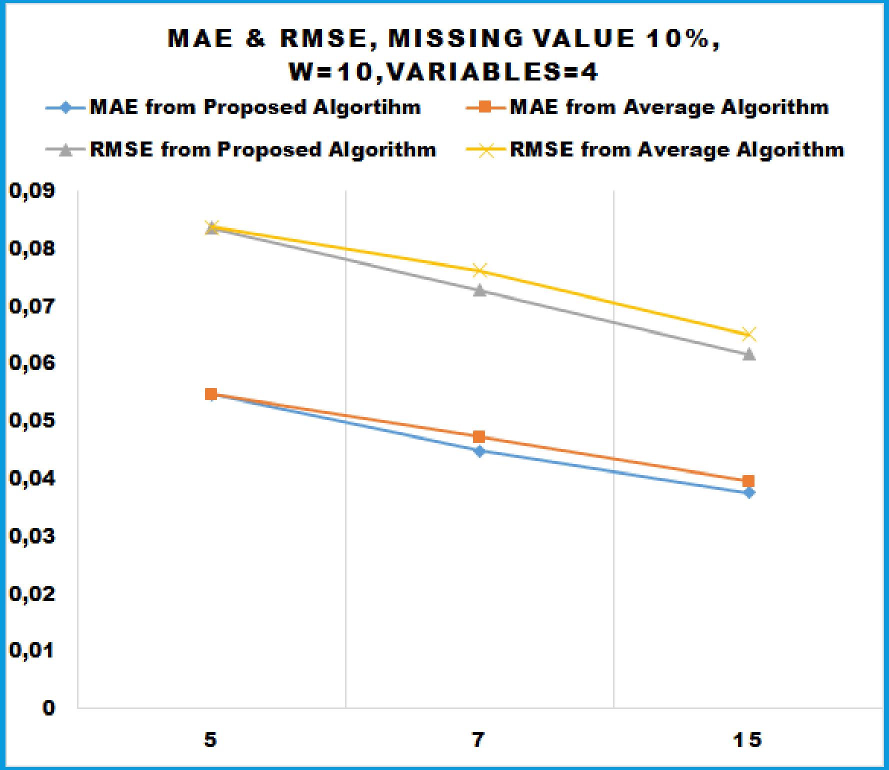
EXPERIMENTAL EVALUATION

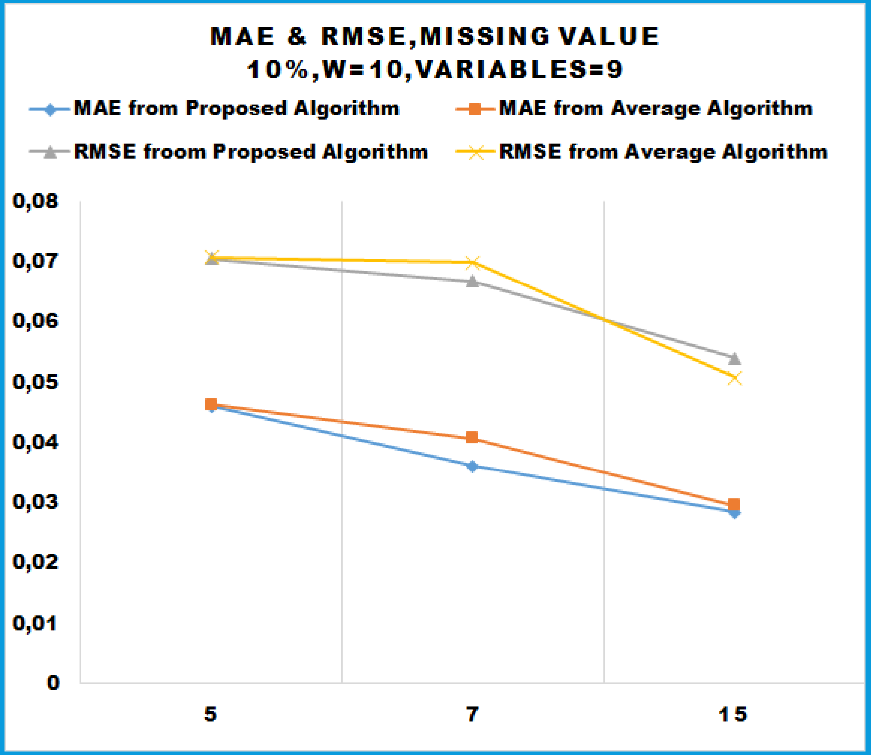




The number of IoT devices affects both scenarios, however, in the ‘opposite’ direction

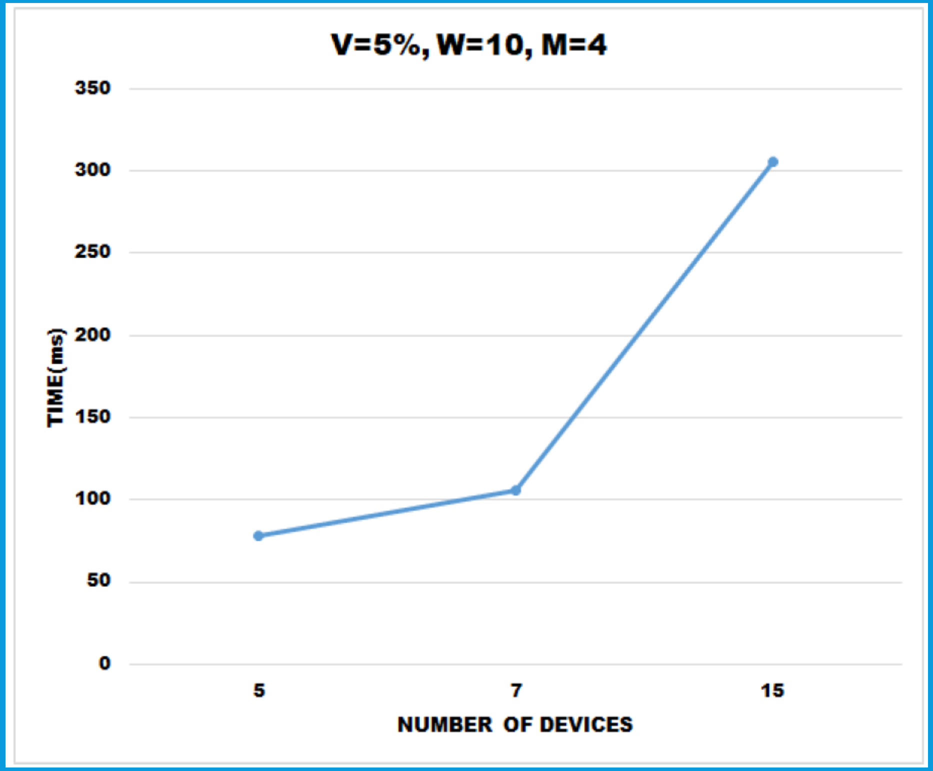
EXPERIMENTAL EVALUATION

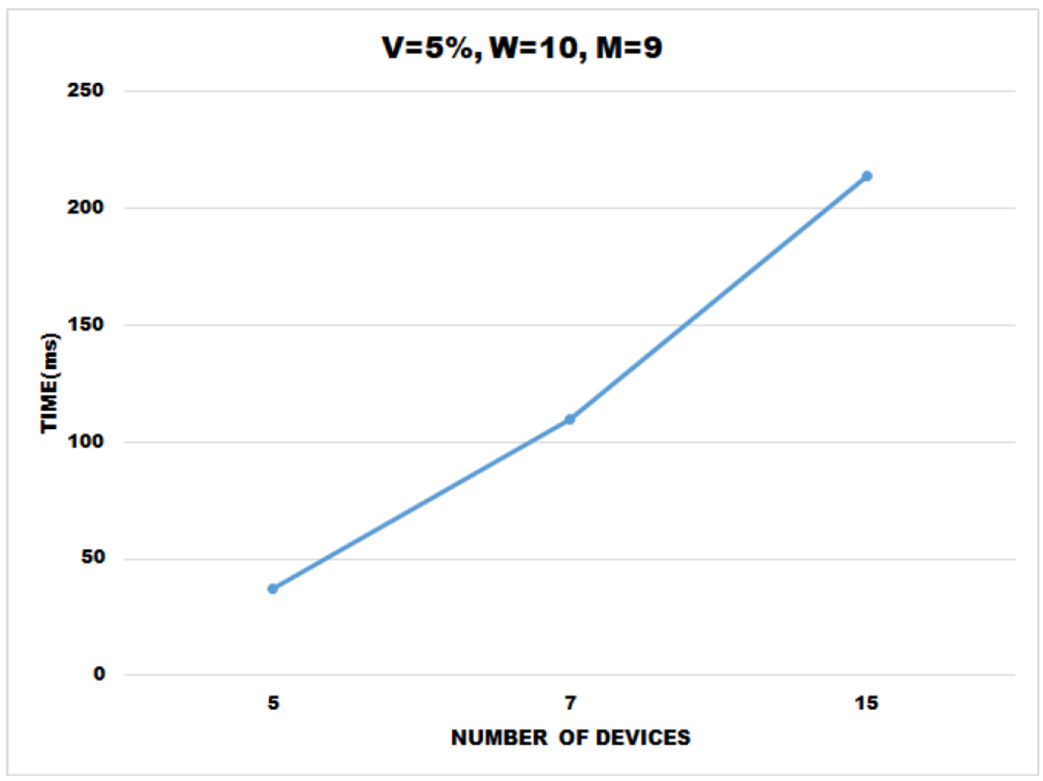




In this set of experiments, a high number of nodes leads to a decreased MAE & RMSE

EXPERIMENTAL EVALUATION





We observe that the number of the devices affects the final outcome and leads to an increased computation time.

CONCLUSIONS & FUTURE WORK

- Missing values imputation is a significant research subject for supporting efficient data analysis.
- We have to adopt data imputation techniques that are capable of providing the final result in the minimum time.
- Our future research plans involve the definition and adoption of a more complex methodology to deal with uncertainty related to the replacement of missing values.

Questions?

Thank you!