UNIVERSITY OF THESSALY
creative years
1988 — 2018

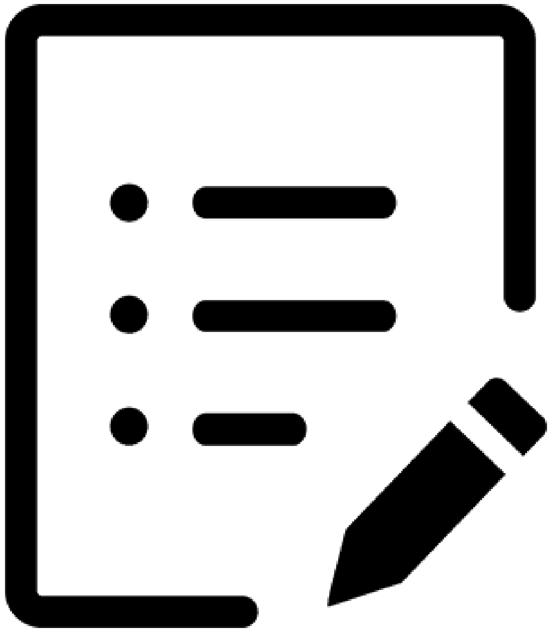
iPRISM
Intelligent Pervasive Systems Research Group
Dept. of Informatics and Telecommunications

Ensemble based Data Imputation at the Edge.

Panagiotis Fountas, Kostas Kolomvatsos
Email: pfountas@uth.gr, kostasks@uth.gr

ICTAI 2020
November 09-11,2020
32th International Conference on Tools with Artificial Intelligence.

Outline



- ➢ Introduction
- ➢ Problem Description
- ➢ Data Imputation Mechanism of Distance Based Model
- ➢ Data Imputation Mechanism of Prediction Based Model
- ➢ 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

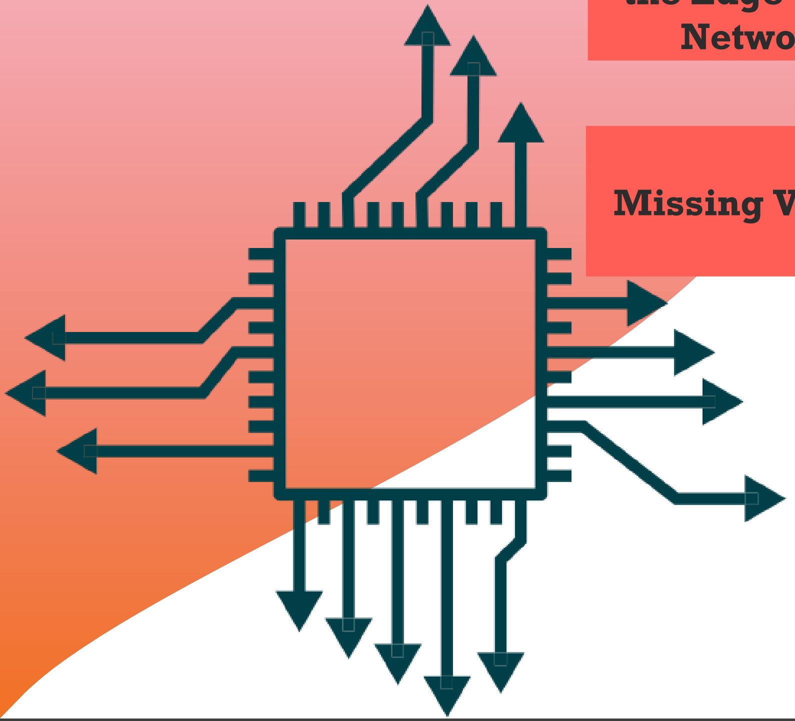
IoT devices and applications produce or collect data.
The data should have appropriate form to draw conclusions

Interaction with the Edge of the Network

The Edge Computing (EC) nodes are placed close to the data sources.
Edge minimize the latency in the provision of responses.
Edge Computing perform analytics over distributed data streams.

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 algorithms



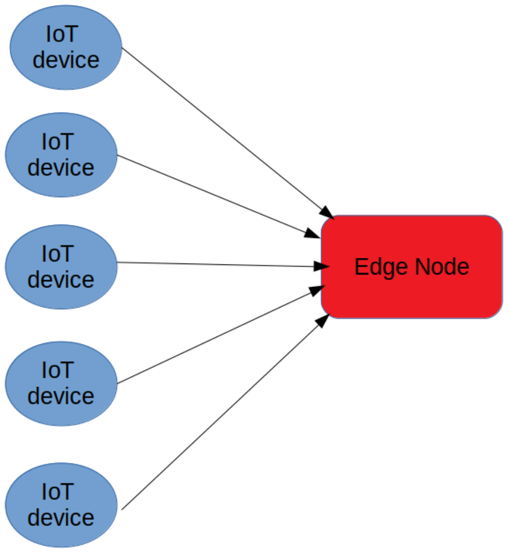
PROBLEM DESCRIPTION

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 of Distance Based Model (DBM)

01

The Data Imputation Mechanism use metrics i.e. the Cosine Similarity (CS) and the Mahalanobis Distance (MD).

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

02

The CS is applied over the latest reports of IoT devices and the Mahalanobis distance is applied over the W latest reports

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

03

Our ensemble scheme use the previous metrics for the calculations of the final correlation between reports of IoT devices.

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

04

Final correlation (F_C) pays attention on the CS result and uses a weighted model to reward devices with increased historical correlation with the device detecting the missing value.

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

05

The replacement value is calculated based on top-k F_C values.

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

Data Imputation Mechanism of Prediction Based Model (PBM)

01

The use of the CS and MD does not differ in the PBM compared to our previous effort, i.e., the Distance Based Model (DBM)

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

02

The replacement value is calculated taking into consideration the "group" view based on top-k F_C values and the "local" view.

03

We assume that we want to estimate the $x[W+1]$ of IoT device at the time instance $W+1$. The "local" view is based on Linear Regression Model to detect the linear relationship between $x[1], x[2], \ldots, x[W]$ and $x[W+1]$.

$$x[W+1] = f(X, B) + \epsilon = b_0 + b_1 x[1] + b_2 x[2] + \cdots, b_W x[W] + \epsilon$$

Data Imputation Mechanism of Prediction Based Model (PBM)

04

The "group" view based on the top-k IoT devices according to F_C and calculated using the Weighted Geometrical Mean where the weights of each report is the MD between IoT device with missing value and corresponding top-k IoT device.

$$WGM = \left(\prod_{i=1}^{k} x_i^{MD_i}\right)^{\frac{1}{\Sigma_{i=1}^{k} MD_i}}$$

05

We rely on a sigmoid function to calculate the weight of "local" view.

$$w_{local} = \frac{1}{1 + e^{\alpha\sigma - \beta}}$$

06

The replacement value resulting as a weighted scheme based on "local" and "group" view.

$$PD = w_{local} \cdot x[W+1] + (1 - w_{local}) \cdot WGM$$

EXPERIMENTAL EVALUATION

❖ In every dataset, we randomly annotate $V\%$ reports as missing values.

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 $W = 10$
Parameters adopted by our smoothing function $\alpha = 20, \beta = 2$

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

1st Metric(s)

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

2nd Metric

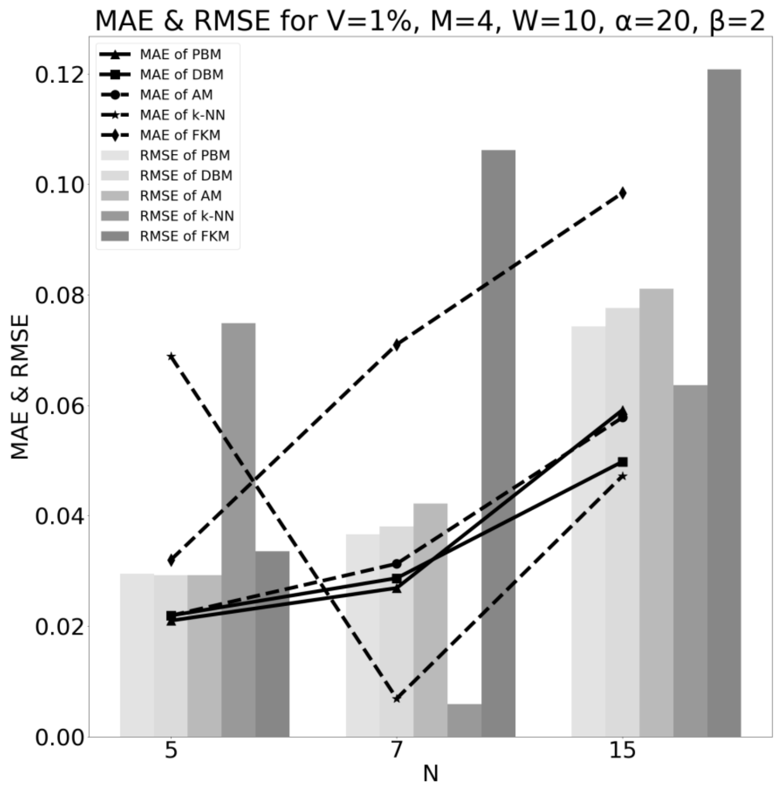
Time Requirements: Average Time per Replacement

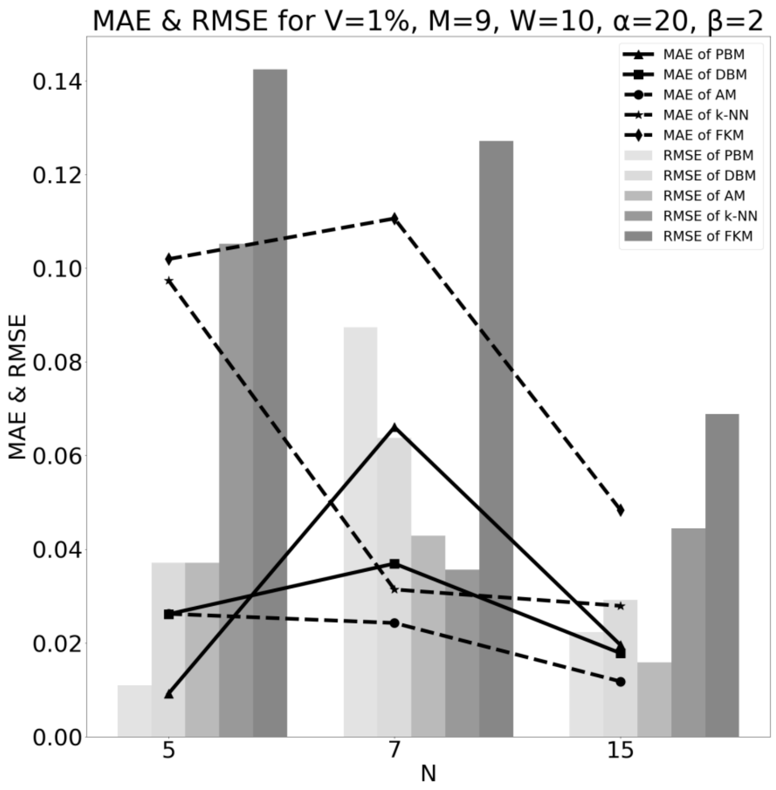
Comparison

Comparison with baseline Models:

- Averaging Model (AM)
- Fuzzy K-Means (FKM)
- k- Nearest Neighbors (k-NN)

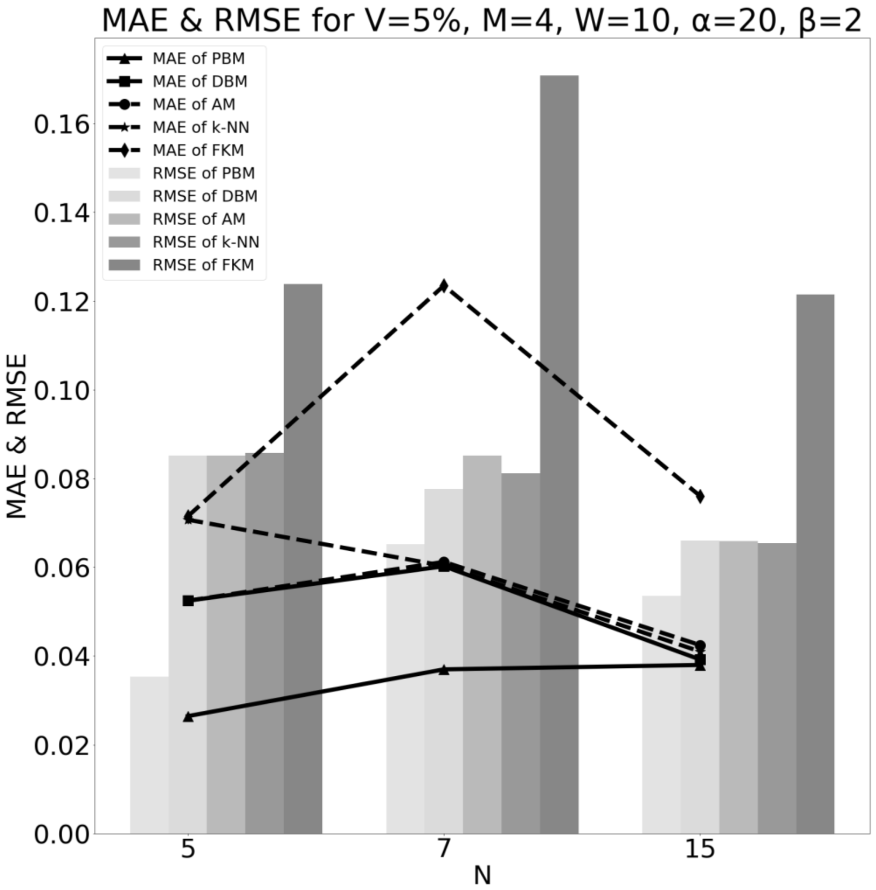
EXPERIMENTAL EVALUATION

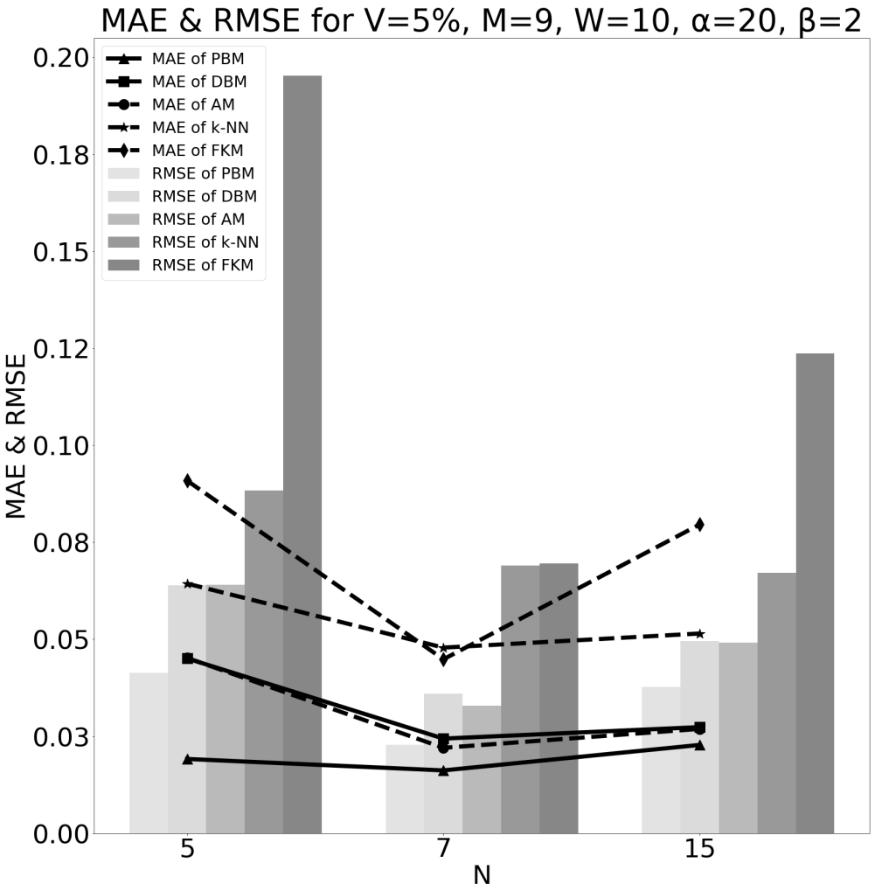




- In general, the increase of dimensions does not negatively affect our proposed models that are capable of efficiently replacing missing values.
- The DBM outperforms the PBM for $M = 9$ and $N = \{7,15\}$ and $M = 4$ and $N = 15$. In the other hand the PBM has better performance for $M = 4$ and $N = \{5,7\}$ and $M = 9$ and $N = 5$.

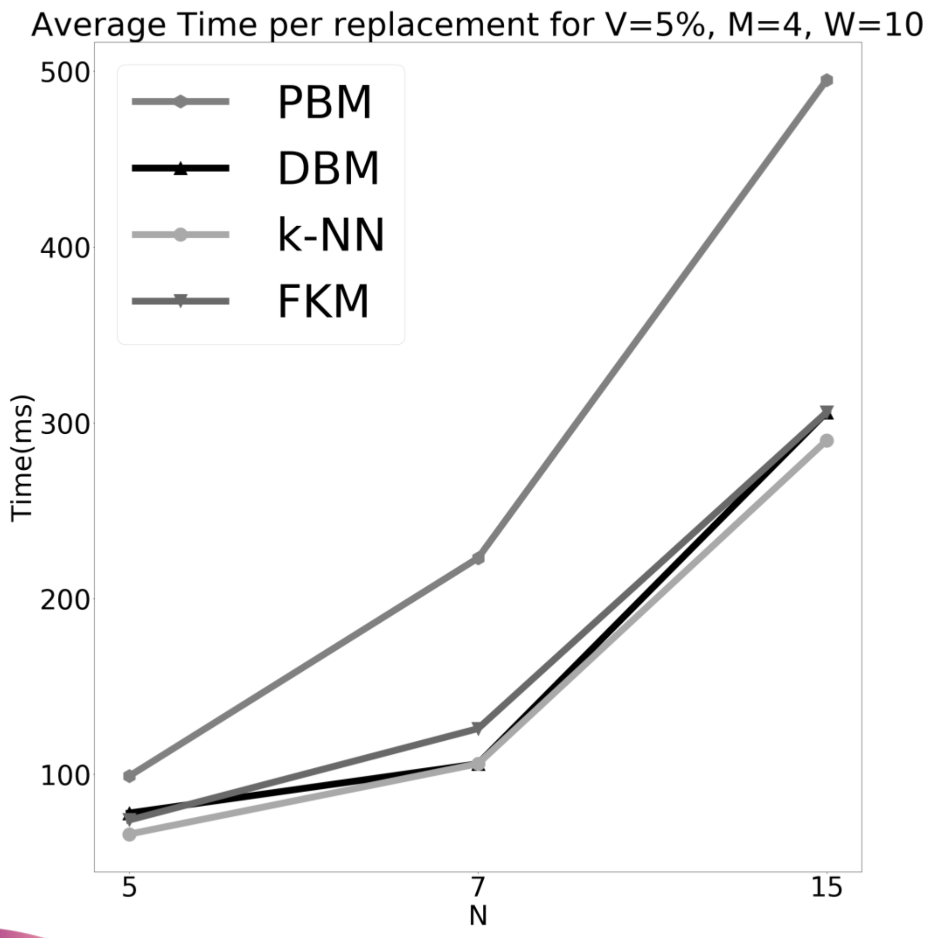
EXPERIMENTAL EVALUATION

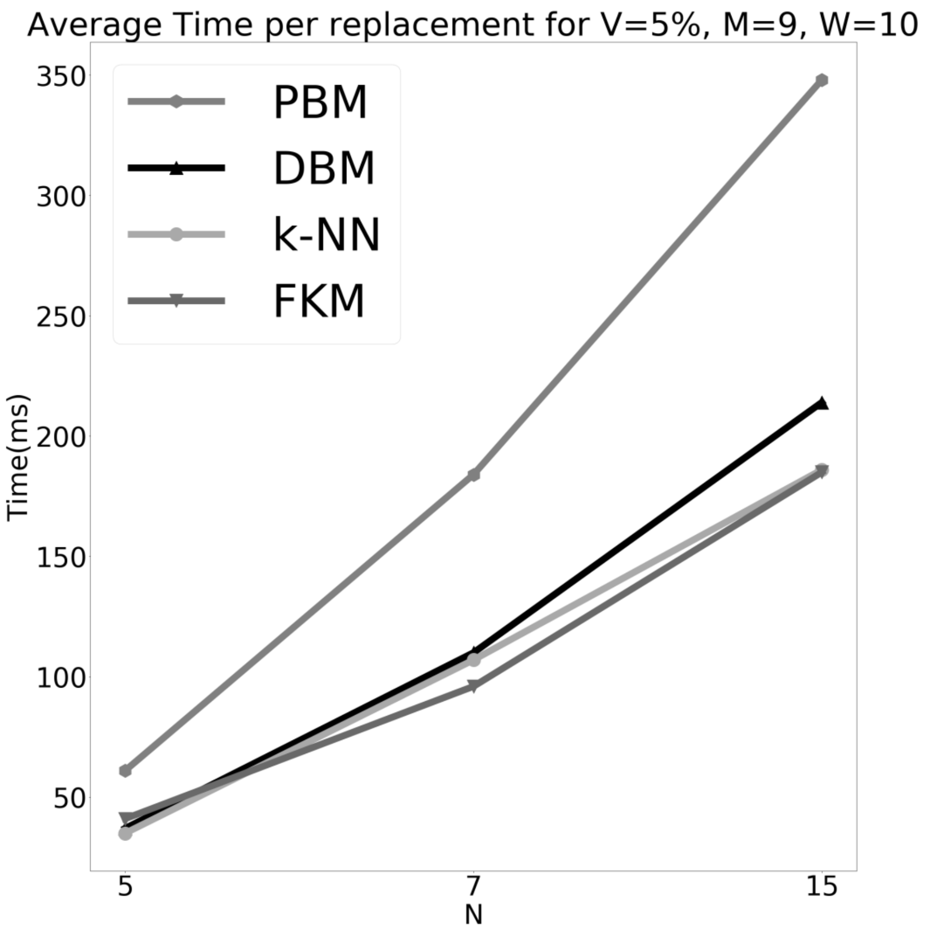




- ❖ The increase of number of missing values has as result the clearly dominance of PBM in both scenarios against all the other models.

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.

iPRISM

THANK YOU

Panagiotis Fountas
Email:pfountas@uth.gr

http://www.iprism.eu