A Scheme for Sensor Data Reconstruction in Smart Home

Yegang Du

Dalian Maritime University

Outline

1. Introduction
   - Smart Home Environment, Research Problem
2. Background
   - Definition, Example and Observed Unavailable Data, Investigation of Unavailable Data, Related Works, Motivation and Objective
3. Proposed HADI Scheme
   - HADI Architecture, Generalized HADI Module, DI Model Data Restoration
4. Numerical Simulation
   - Solar Irradiance, Relative Humidity, Performance on Processing Time
5. Conclusion and Future Work
6. List of Publications

- Smart home is one of the most popular Internet of Things (IoT) application today
- Sensor data takes a significant role in the performance of automated systems in smart home environment

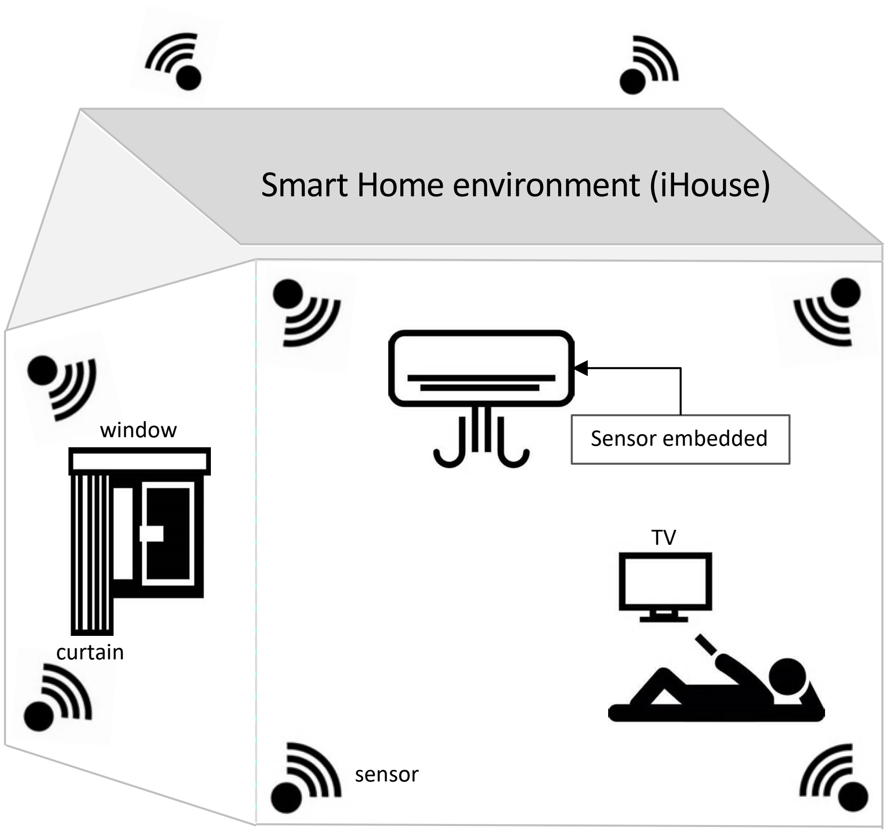


Fig. 1 Sensing Area Networks

Sensors are distributedly placed throughout the smart home environment

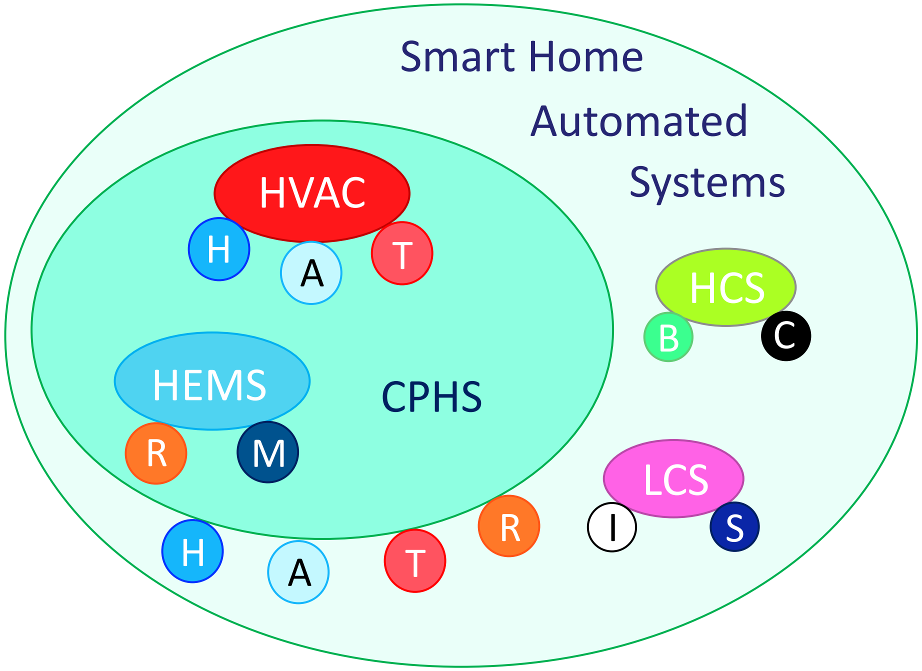


Fig. 2 Smart Home Automated Systems

HVAC : Heating, ventilation, and air conditioning
HEMS : Home energy management system
CPHS : Cyber-physical home system
HCS : Healthcare system
LCS : Lighting control system

H : Humidity sensor
A : Air flow sensor
T : Temperature sensor
R : Solar irradiance sensor
M : (electricity) Meter
B : Biosensor
C : Camera
I : Illuminance sensor
S: Switch

1.1 Smart Home Environment and CPHS

Fig. 3 iHouse in Nomi City

- Advanced experimental environment for future smart home
- Over 300 sensors and actuators deployed
- ECHONET Lite v1.1 protocol implemented
- Various smart home automated systems, e.g., Cyber-Physical Home System (CPHS)

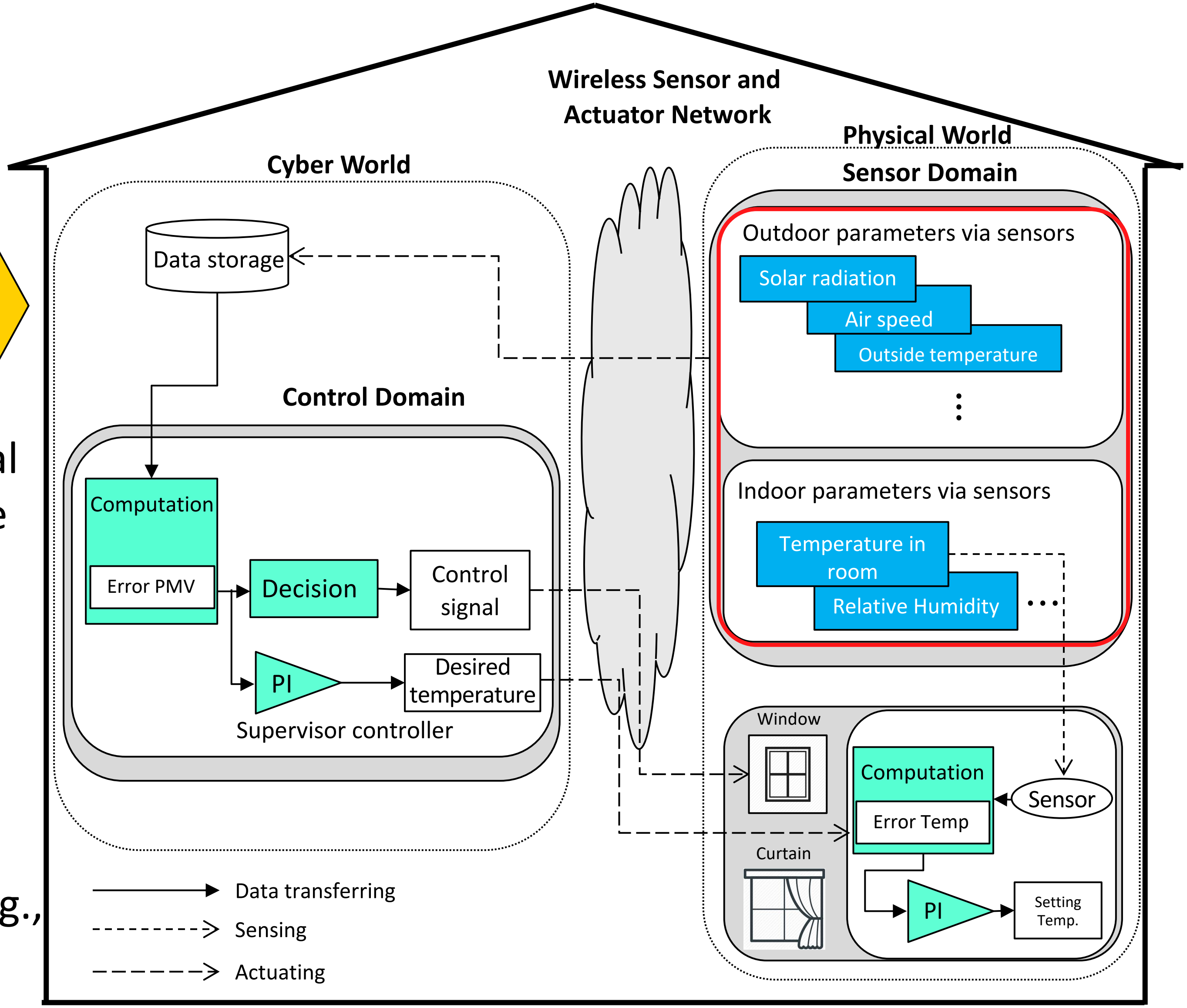


Fig. 4 Cyber-Physical Home System[1]

[1] Z. Cheng, W.W. Shein, Y. Tan, and A.O. Lim, "Energy efficient thermal comfort control for cyber-physical home system," *in Proc. of the IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Vancouver, Canada, pp. 797-802, 21-24 October 2013.

1.2 Research Problem

Date	Time	Temperature	Relative Humidity
2016/01/01	01:06:19	2.72	112.49
⋮	⋮	⋮	⋮
2016/01/01	06:17:34	1.82	100.5
2016/01/01	06:17:39	1.82	99.90
⋮	⋮	⋮	⋮
2016/01/05	15:46:19	6.99	99.96
2016/01/05	15:46:24	6.98	-9999
⋮	⋮	⋮	⋮
2016/01/06	09:00:49	6.56	-9999
2016/01/06	09:00:54	6.56	99.73
⋮	⋮	⋮	⋮

Table 1 Example of Observed Data

Sensors' raw data

↓

Unavailable data exist

↓

Automated Systems Failure

Consequences
- increase energy consumption
- decrease indoor thermal comfort
- and so on

Fig. 5 Consequences of Unavailable Data

Research Problem

Automated systems meet a potential hazard caused by unavailable data problem, especially for a unique sensor. To prevent automated system suffering from unavailable data, a data restoration scheme is considered in this research.

2.0 Background

In general, availability defined by failure[2]

$$\lim_{t \to \infty} A(t) = A = \frac{MTTF}{MTTF + MTTR}$$

MTTF : Mean Time to Failure
MTTR : Mean Time to Repair

In this research, data availability of a sensor x is:

$$\lim_{t \to T} A_x(t) = A_x = \frac{IAD}{(IAD + IUD)}$$

IAD: Interval of available data
IUD: Interval of unavailable data

In [3], data fault has been defined as

Unavailable Data Type	Definition
Outlier	Isolated data point or sensor unexpectedly distant from models
Stuck-at	Multiple data points with a much greater than expected rate of change
Calibration	Sensor reports values that are offset from the ground truth
Spike	Multiple data points with a much greater than expected rate of change

Table 2 Unavailable Data Type

Sensor Type	Outlier	Stuck-at	Calibration	Spike
Temperature	○	○	○	○
Humidity	○	X	X	X
Solar irradiance*	△	X	X	△
Wind speed	△	X	○	△

△ Unavailable data exists, but is acceptable ○ Unavailable data doesn't exist X Unavailable data exists

Table 3 Unavailable Data Type on Sensors

Duration Type	Definition
Intermittent	Data show an unavailable less than k samples. Most intermittent unavailable data are caused by outlier
Continual	Data show an unavailable more than and equal to k samples. Most continual unavailable data are related with Spike, Stuck-at and calibration

Table 4 Duration Type

[2] M. Rausand and H. Arnljot, *System reliability theory: models, statistical methods, and applications*. John Wiley & Sons, vol. 396, 2004.
[3] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava, "Sensor network data fault types," *ACM Transactions on Sensor Networks* (TOSN), vol. 5, no. 3, 2009.

2.1 Example and Observed of Unavailable Data

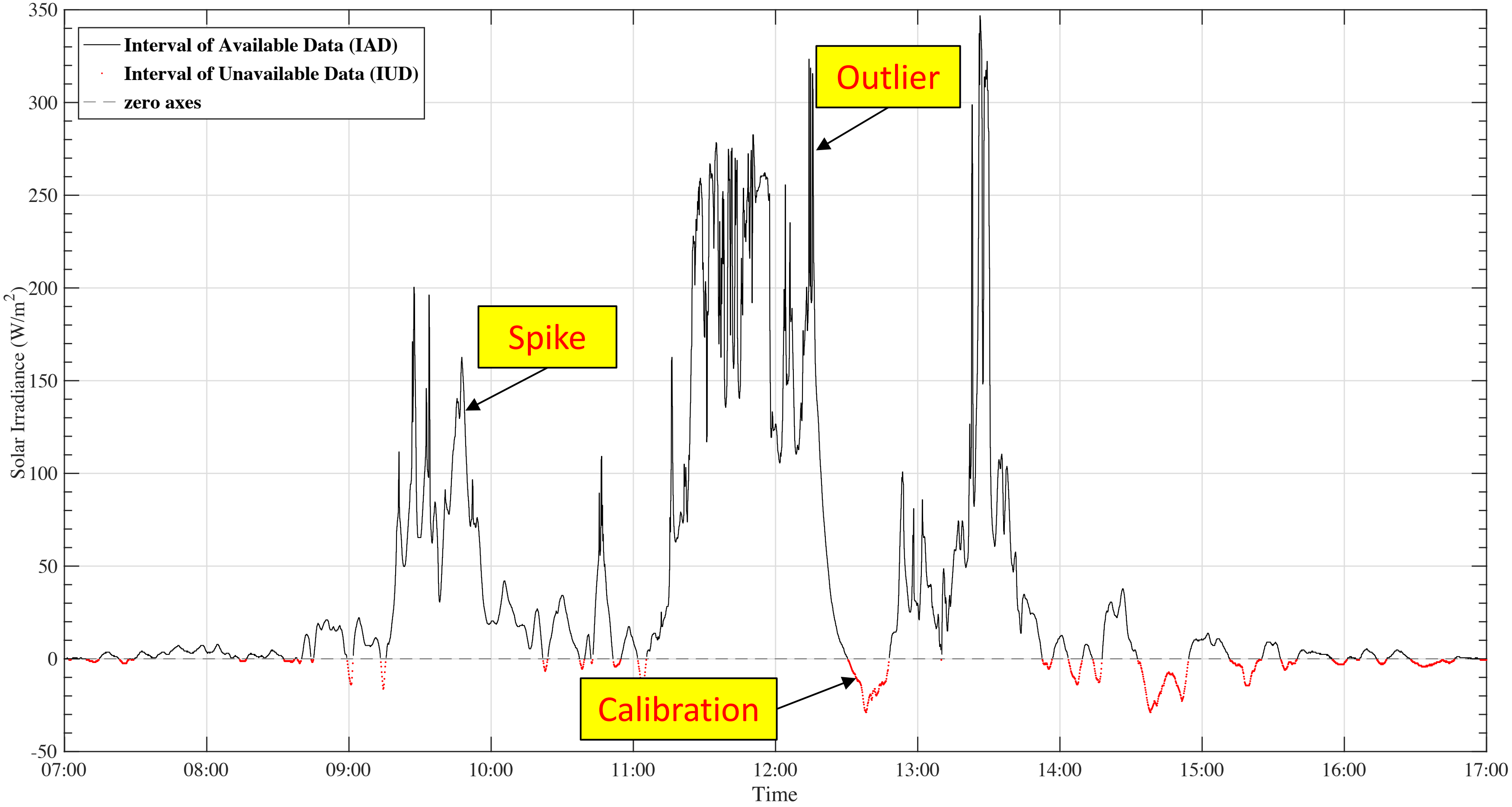
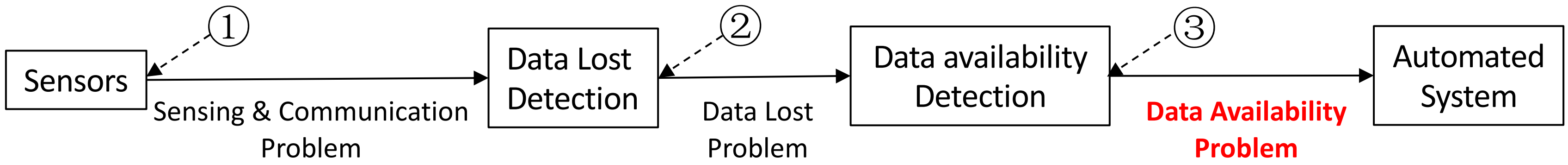


Fig. 6 Raw Data of Solar Irradiance

2.2 Investigation of Unavailable Data at iHouse



Humidity sensor	①	②	③
2nd floor	100% Delay:0.83%	98.16%	98.16%
Bedroom	100% Delay:1.45%	98.74%	98.74%
Entrance	100% Delay:0.82%	99.07%	99.07%
Japanese room	100% Delay:0.85%	98.83%	98.83%
Kitchen	100% Delay:0.85%	98.75%	98.75%
Living room	100% Delay:1.45%	98.15%	98.15%
Spare room	100% Delay:0.81%	98.52%	98.52%
Utility room	100% Delay:0.84%	98.83%	98.83%
Western room 1	100% Delay:1.59%	98.77%	98.77%
Western room 2	100% Delay:0.96%	98.82%	98.82%
Outdoor	100% Delay:0%	100%	**64.29%**

Table 5 Availability of Humidity Sensor

Temperature sensor	①	②	③
2nd floor	100% Delay:6.75%	98.82%	98.82%
Bedroom	100% Delay:6.67%	98.77%	98.77%
Entrance	100% Delay:6.74%	98.59%	98.59%
Japanese room	100% Delay:6.72%	98.76%	98.76%
Kitchen	100% Delay:6.72%	98.76%	98.76%
Living room	100% Delay:6.72%	98.83%	98.83%
Spare room	100% Delay:6.73%	98.83%	98.83%
Utility room	100% Delay:6.72%	98.83%	98.83%
Western room 1	100% Delay:6.67%	98.78%	98.78%
Western room 2	100% Delay:6.61%	98.82%	98.82%
Outdoor	100% Delay:6.68%	100%	100%

Table 6 Availability of Temperature Sensor

Solar irradiance	①	②	③
outdoor	100% Delay:0	100%	**93.4***

Table 7 Availability of Pyranometer

Wind speed	①	②	③
outdoor	100% Delay:0	100%	99.96%

Table 8 Availability of Anemometer

In the unavailable data, a **continual duration** is major type

Unavailable data of solar irradiance are defined as:

- Minus value in daytime

In 2016 whole year, data availability of solar radiation are calculated, and we found:

- Unavailable data in daytime last 574.2 hours
- Pyranometer is unavailable nearly 1.6 hours (during daytime) every single day

Methods are required to restore those unavailable data (solar irradiance/relative humidity) with homogeneous and/or heterogeneous data

2.3 Related Works, Motivation and Objective

Artificial Neural Network (ANN)

- Use enormous completely available temporal data to estimate future data
- Aiming to achieve a highly accurate restoration for HVAC system

Efficient Temporal and Spatial Data Recovery (ETSDR)

- Use spatiotemporal homogeneous data
- Apply the linear regression mechanism ARIMA
- Achieve extremely high accuracy intermittent data restoration

Deep Multimodal Encoder (DME)

- An optimized ANN network
- DME specialize the hidden layer for adapting to heterogeneous data computing

D
Resto

Objective : 'high availability' and 'comparatively accurate' data restoration scheme with rapid computation to maintain a regular operation of any automated system

Principle components analysis (PCA)

- Use spatiotemporal homogeneous data
- Taking the lead of applying PCA on data restoration in HVAC system

Recursive-PCA

- Remarkable efficiency on data fault detection, data aggregation and recovery accuracy
- Consider the spatiotemporal homogeneous and heterogeneous data

Intermittent Unavailable Problem

Continual Unavailable Problem

Current research on continual unavailable problem

- Require highly available data as training sample
- Too much time consumption on data process

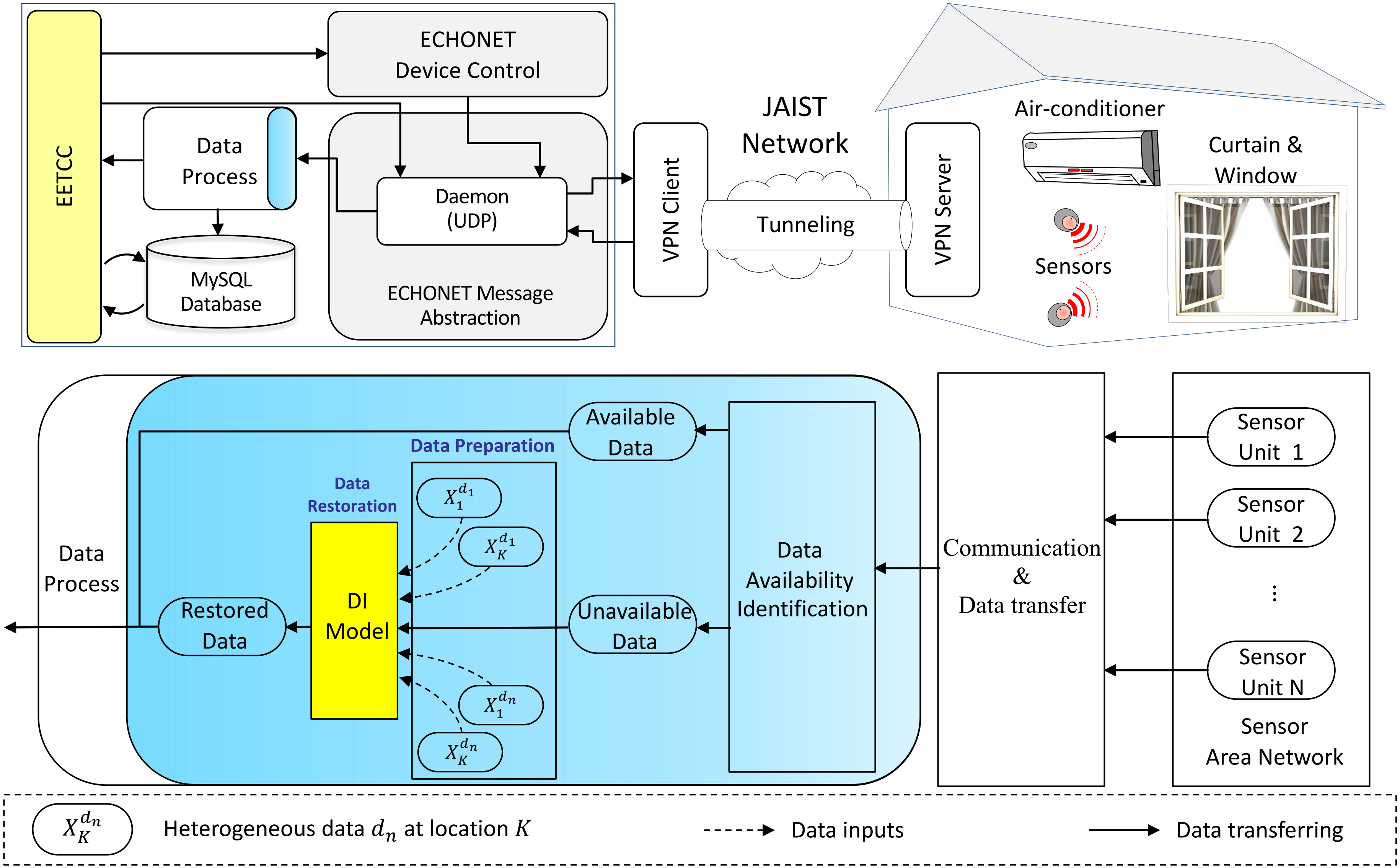
Current research on intermittent unavailable problem

- Emphasize accuracy too much, ignore requirement of data in real-world condition
- The higher accuracy, the longer processing time

Motivation

3.0 Proposed HADI Scheme

Highly Available Data Interpolation (HADI) Scheme and its Architecture



3.1 HADI Modules

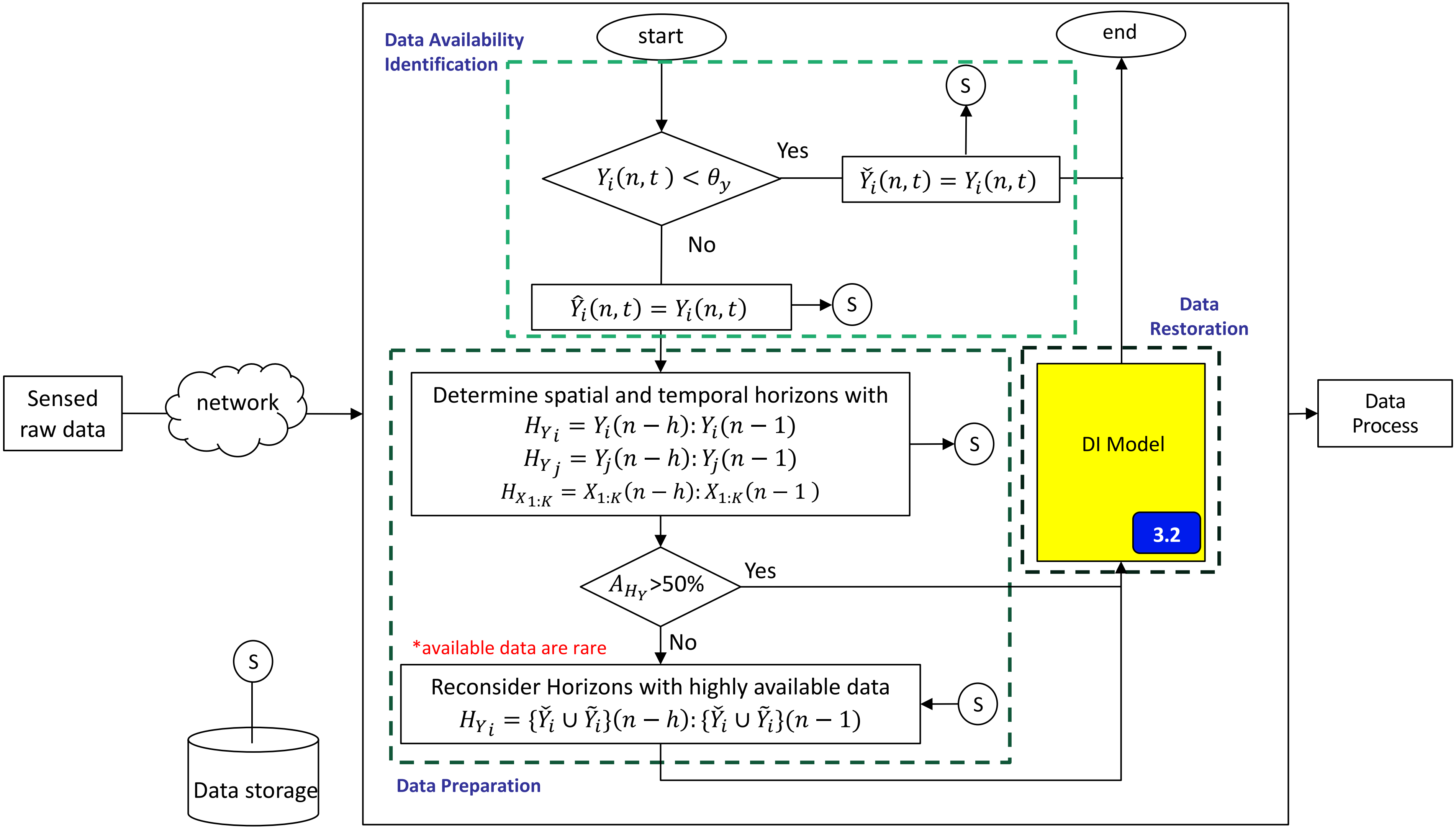
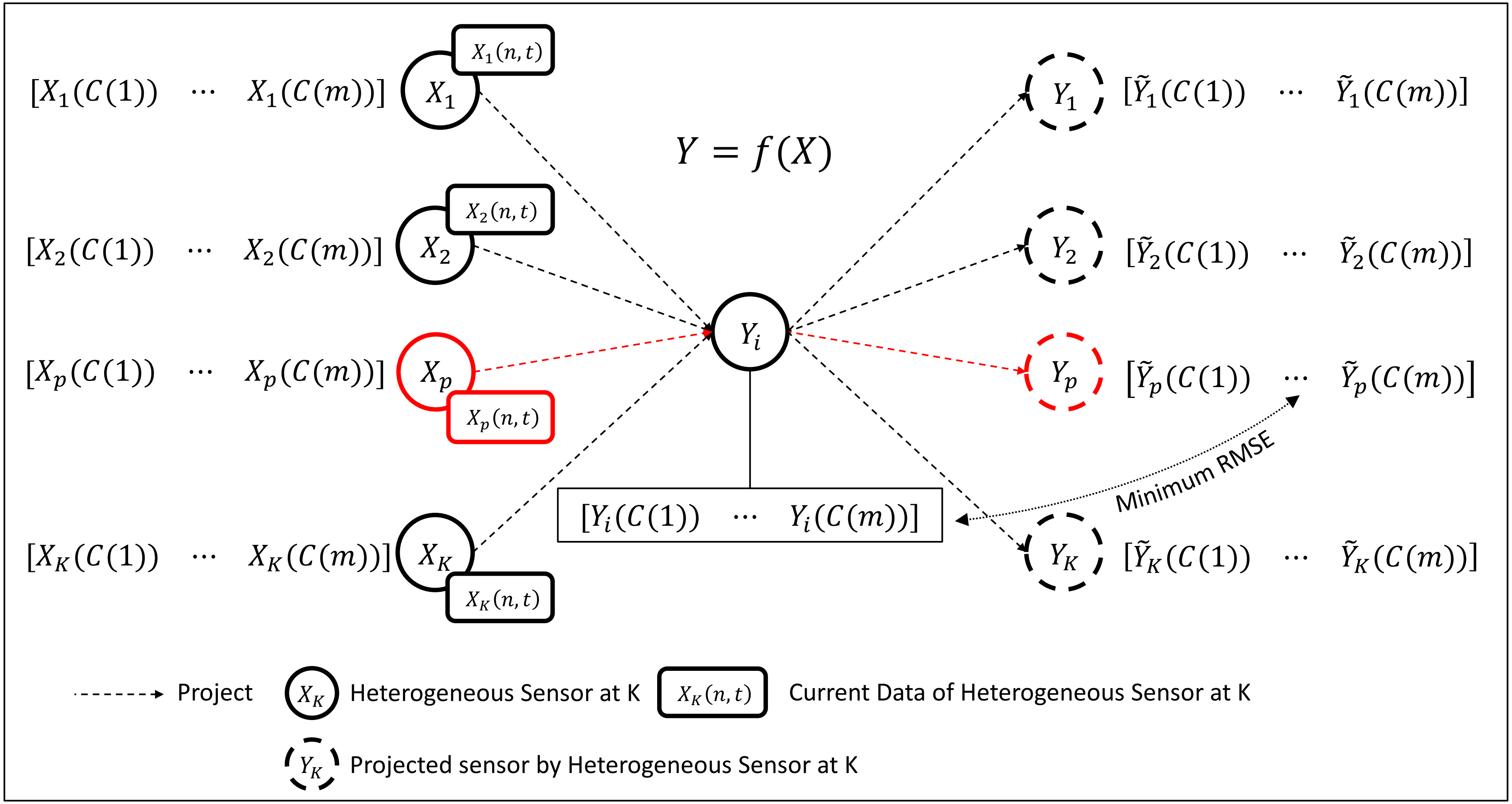


Fig. 7 Flow Chart and its HADI Modules

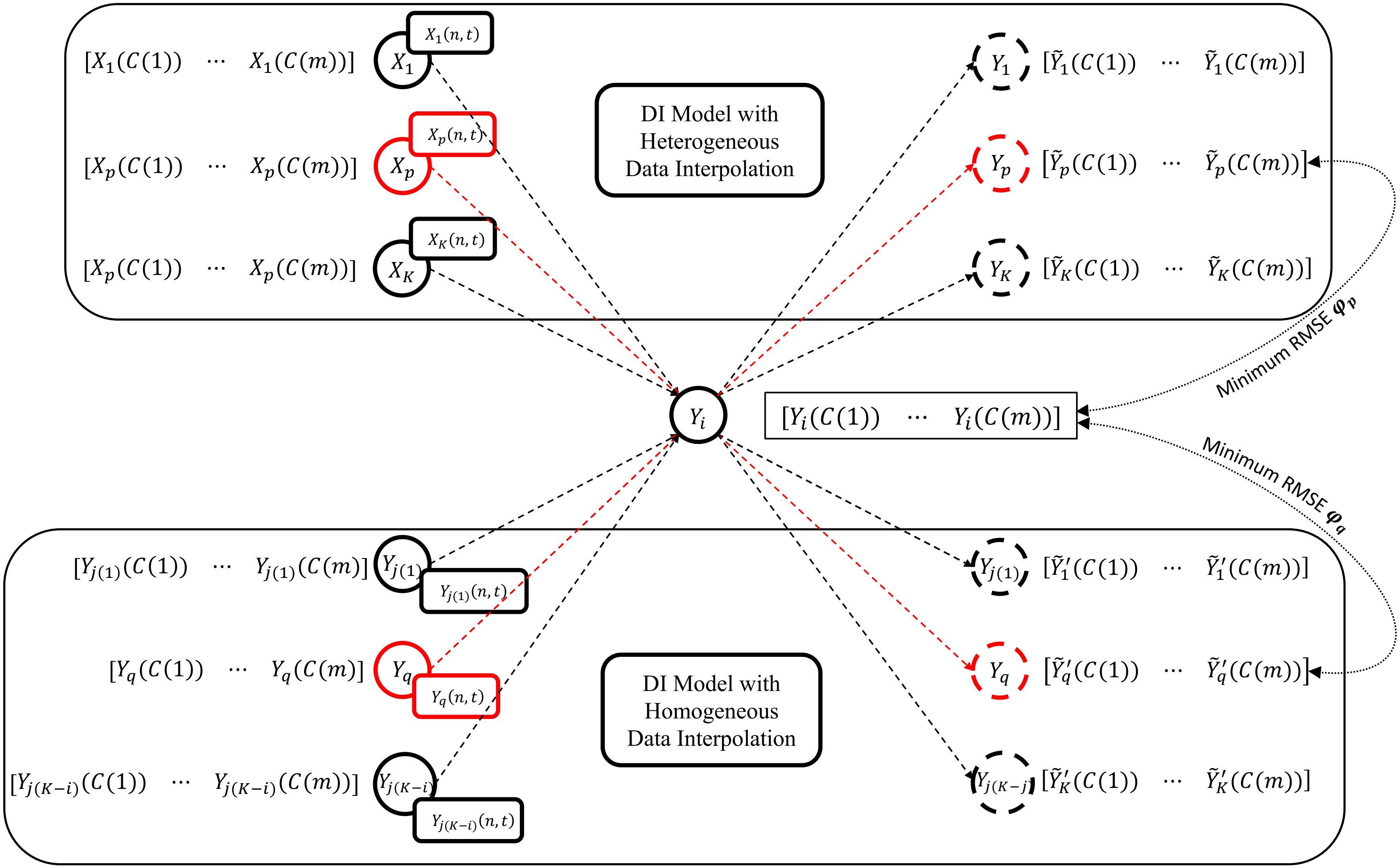
3.2 DI Model Data Restoration

3.2.1 With Heterogeneous Data Interpolation



DI Model Data Restoration

3.2.2 With Heterogeneous and/or Homogeneous Data Interpolation



DI Model Data Restoration

4.0 Simulation Verification

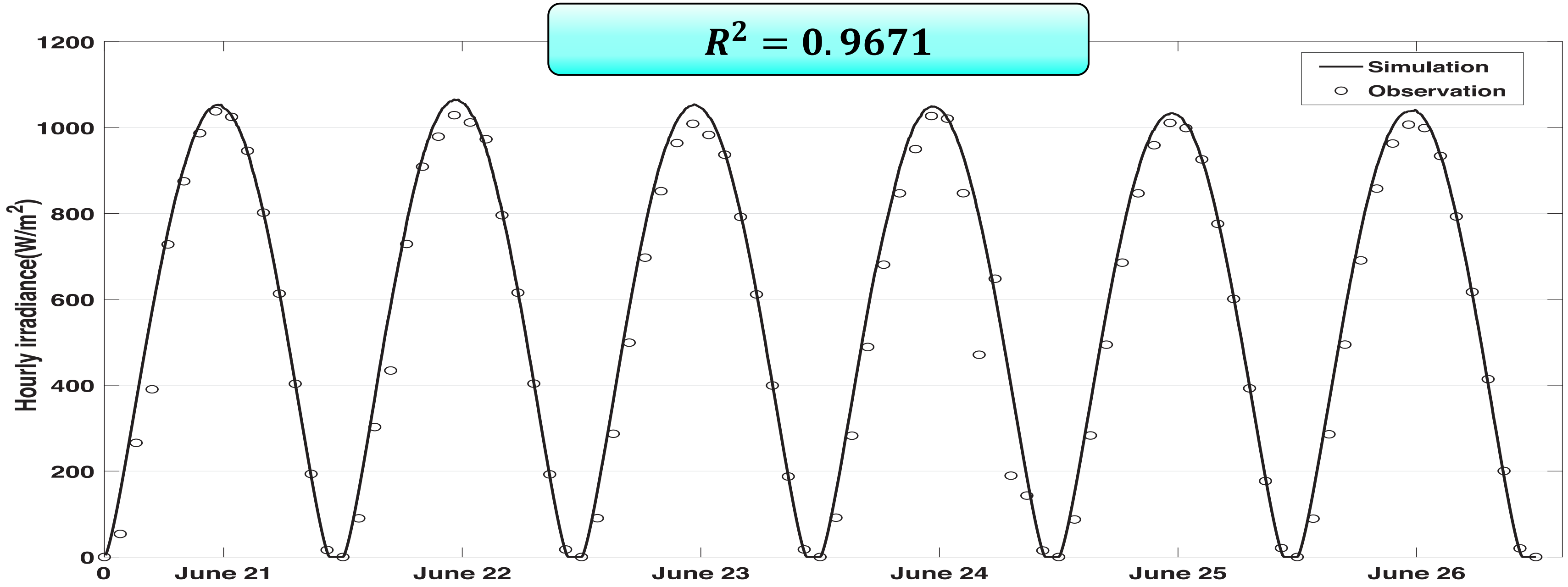


Fig. 8 Verification of Solar Irradiance

Parameter	Description
Surface elevation	1219 m
Latitude	31.80°
Location	USA EP (El Paso)

Table 9 Parameter and Description for Verification

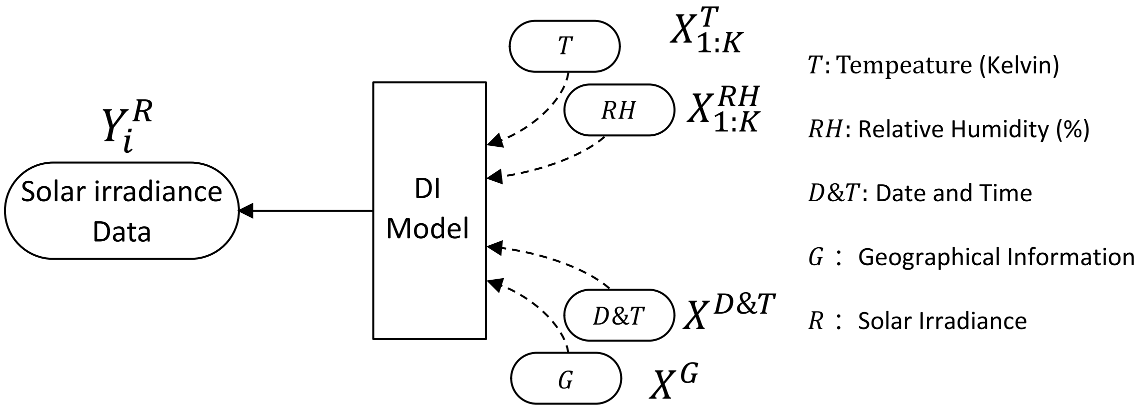


Fig. 9 Input and Output of DI Model

Parameter	Description
Surface elevation	**132 m**
Latitude	**36.40°**
Location	**JP Nomi**

Table 10 Parameter and Description for DI Model with Heterogeneous Data Interpolation

4.1 Solar Irradiance Data Restoration

4.1.1 Comparison of Intermittent Unavailable Data Restoration

RMSE ：Root Mean Square Error
MAE : Mean Absolute Error

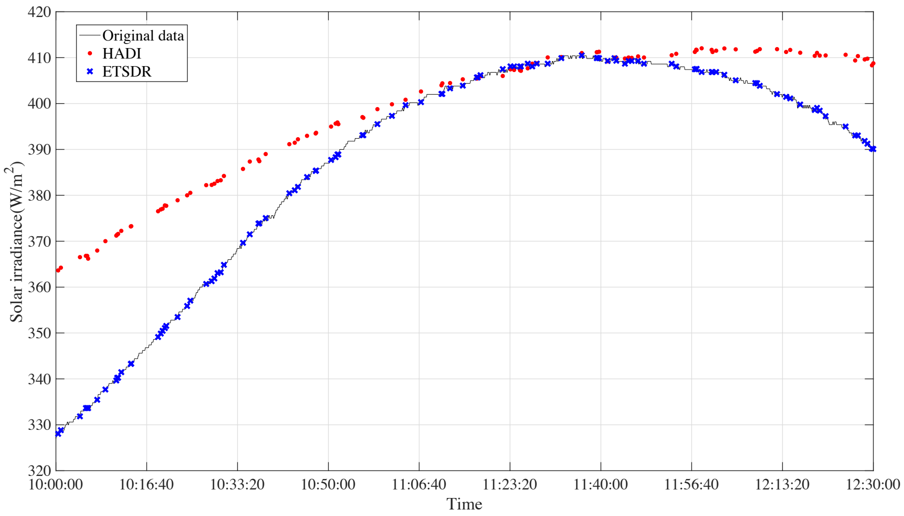


Fig. 10 Example of Intermittent Unavailable Data Restoration

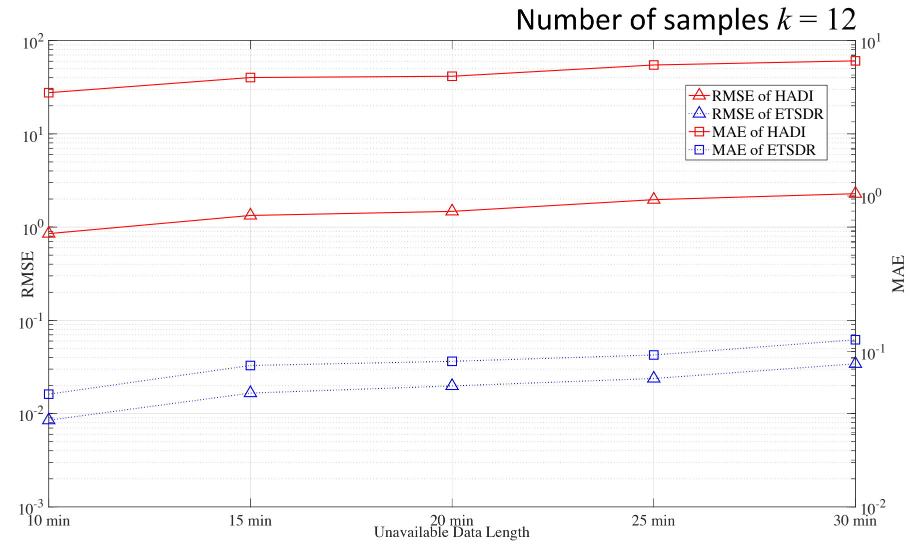


Fig. 11 Performance of Intermittent Unavailable Data Restoration (Number of simulations :10)

- Despite of HADI results in bigger error, both HADI and ETSDR show the feasibility of data restoration
- With the unavailable data length growing, accuracy of either HADI or ETSDR will decrease
- Average growth rate of RMSE and MAE of HADI are 12.89% and 29.14%, respectively. However, ETSDR holds higher accuracy, in which these values are 23.61% and 44.62%, respectively

4.1.2 Comparison of Continual Unavailable Data Restoration

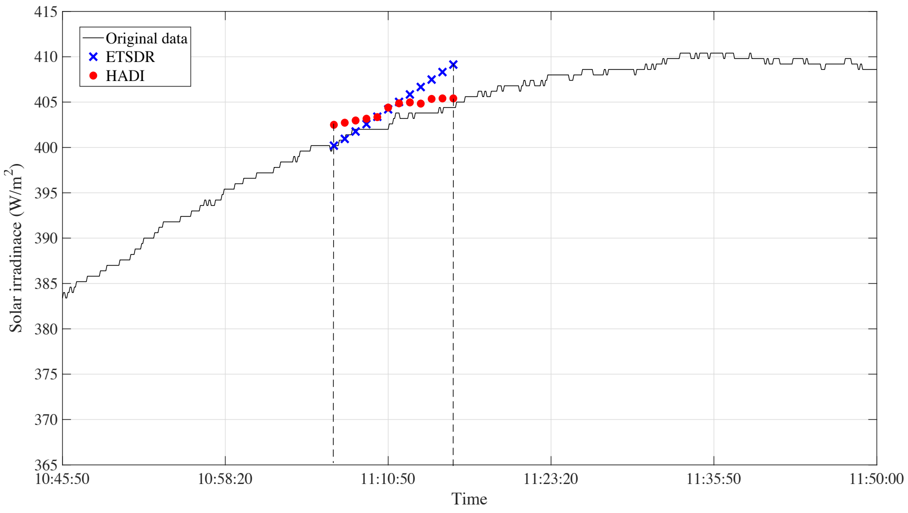


Fig. 12 Example of Continual Unavailable Data Restoration

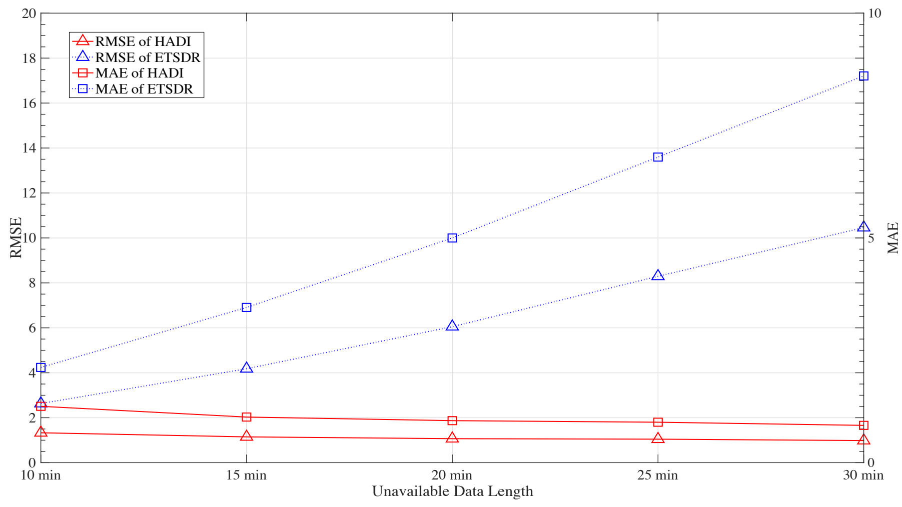
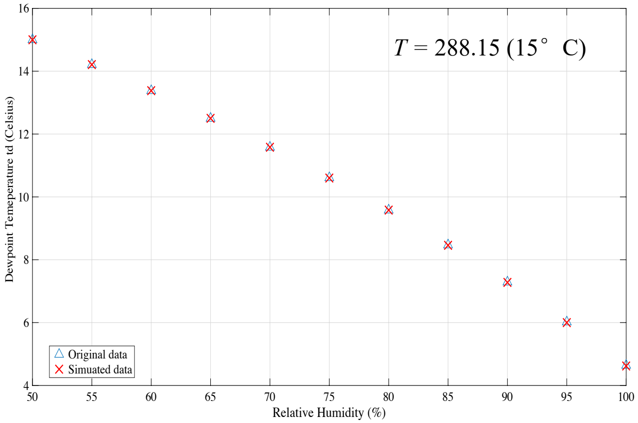


Fig. 13 Performance of Continual Unavailable Data Restoration

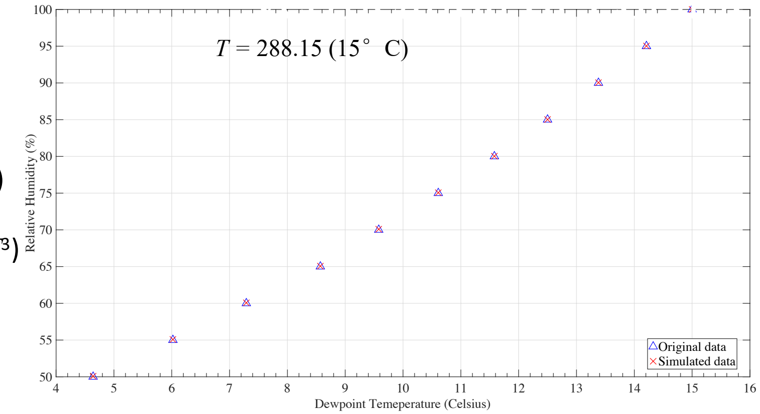
- ETSDR cannot update dynamic linear model with continual unavailable problem. Thus, the restored data will regress in straight line gradually
- Apparently, HADI shows stable errors with original data, accuracy of HADI is nearly same with intermittent unavailable data restoration
- On the contrary, it is astonishing that RMSE due to 30 min unavailable data interpolation increased by 400% compared with condition of 10 min

4.2 Relative Humidity Data Restoration

4.2.1 Verification for Relative Humidity Equations



R_w = 461.5J/(K*kg) (constant for water vapor)
T_d= dew point temperature (Kelvin)
t_d = dew point temperature (Celsius)
L (enthalpy of vaporization)
= (2500.8-2.36T+0.0016T^2-0.00006T^3) J/g



$$f\big(g(X)\big) = Y$$

$$T_d = \left[1 - \frac{T \times ln(\frac{RH}{100})}{L/R_w}\right]^{-1}$$

$$RH = 100exp\left[-\frac{L}{R_w T T_d}(T - T_d)\right]^{-1}$$

RH: Relative Humidity (%)

T: Tempeature (Kelvin)

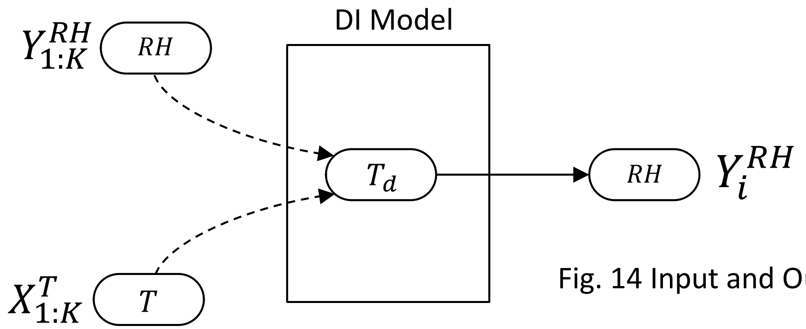


Fig. 14 Input and Output of DI Model

RH(%)	100.0	90.0	80.0	70.0	60.0	50.0
Original data	15.00	13.38	11.58	9.58	7.29	4.64
Simulated result	15.00	13.38	11.58	9.58	7.27	4.62

t_d	4.64	7.29	9.58	11.58	13.38	15.00
Original data	50.00	60.00	70.00	80.00	90.00	100.0
Simulated data	50.08	60.07	70.09	80.03	90.04	100.0

4.2.2 Comparison of Intermittent Unavailable Data Restoration

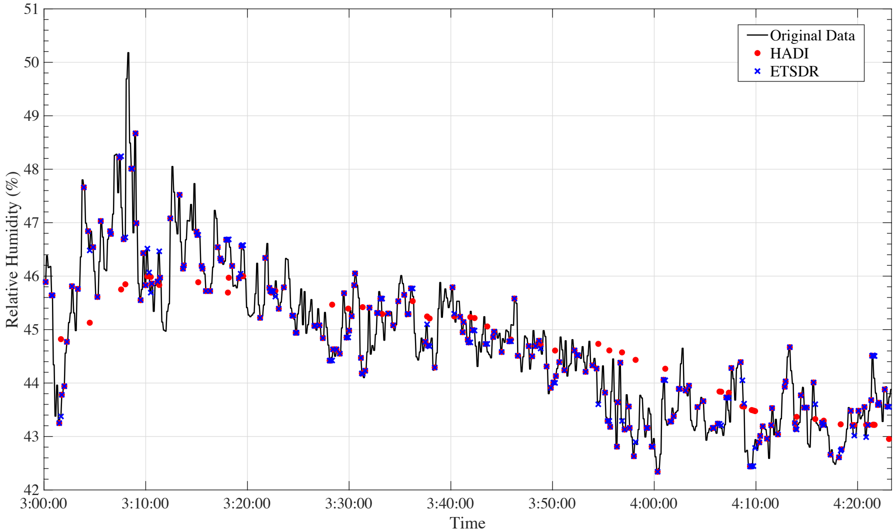


Fig. 15 Example of Intermittent Unavailable Data Restoration

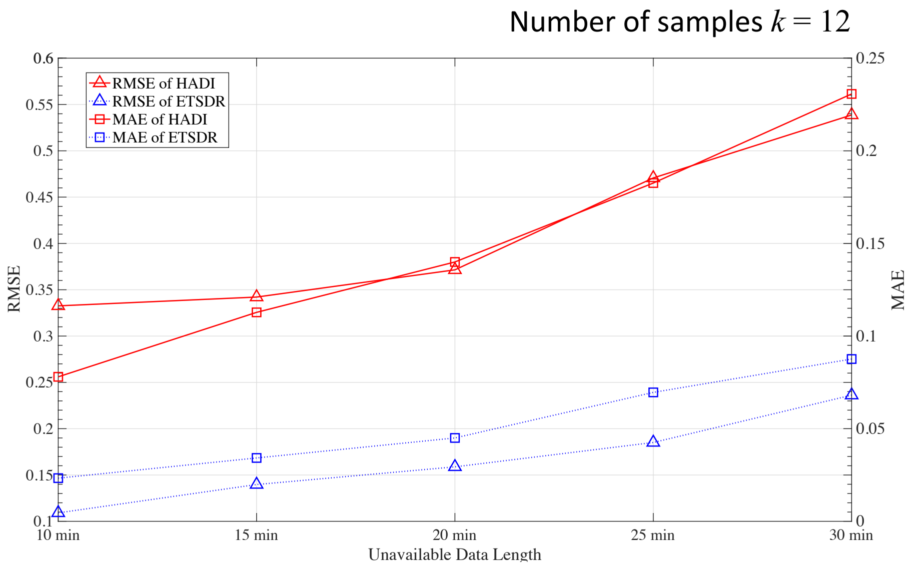


Fig. 16 Performance of Intermittent Unavailable Data Restoration (Number of simulations :10)

- Relative humidity has a lower standard deviation. This means that values are less spread out from their mean value
- It is noticeable that most of restored data by HADI and ETSDR are closely located at original data curve, most of them is even coincident
- Despite restored data by HADI reveal a certain error, however 89.1% restored data are within 0.1 error value

4.2 Relative Humidity Data Restoration

4.2.3 Comparison of Continual Unavailable Data Restoration

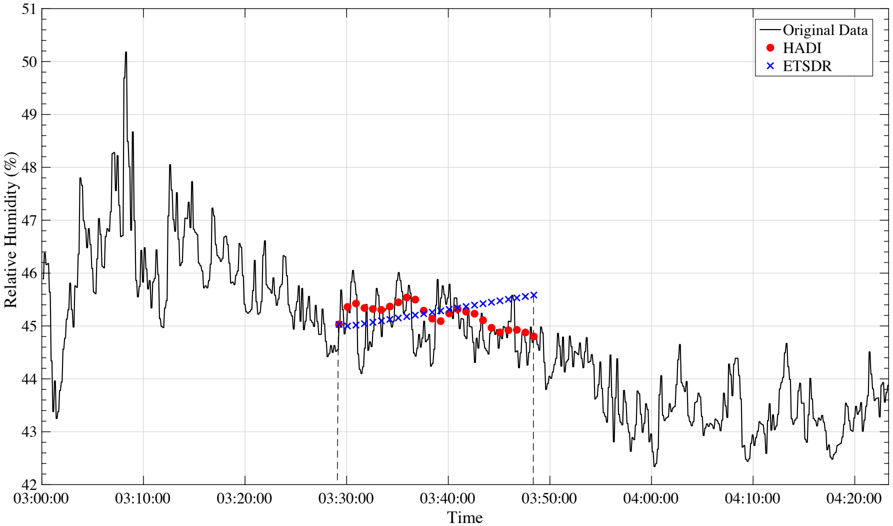


Fig. 17 Example of Continual Unavailable Data Restoration

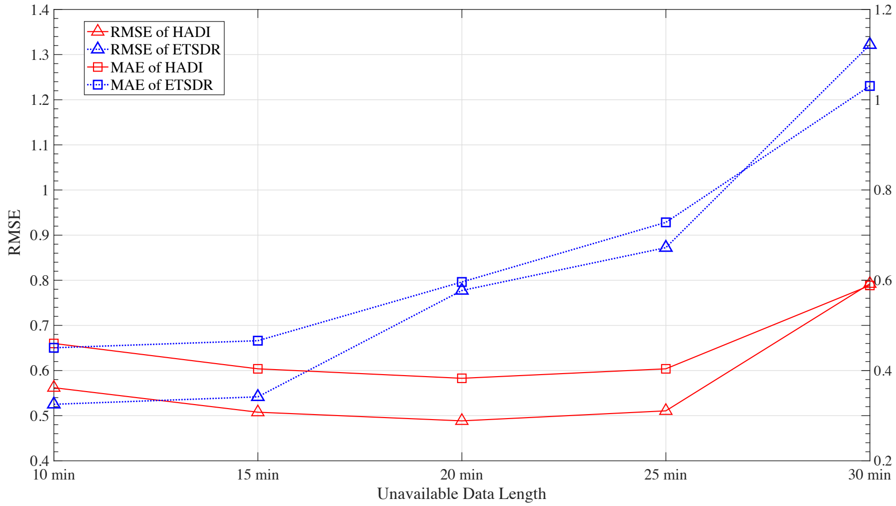


Fig. 18 Performance of Continual Unavailable Data Restoration

- HADI represents an excellent tracking character, although there are errors with original data
- After a length of tiny variation which is hardly to recognize, the data restoration by ETSDR regress in an obvious linearity
- HADI shows a stable variation on RMSE and MAE, the results of data restoration with 30 min reveals that the horizon is not so instructive as before when the raw data are sparse in target horizon

4.3 Processing Time Performance

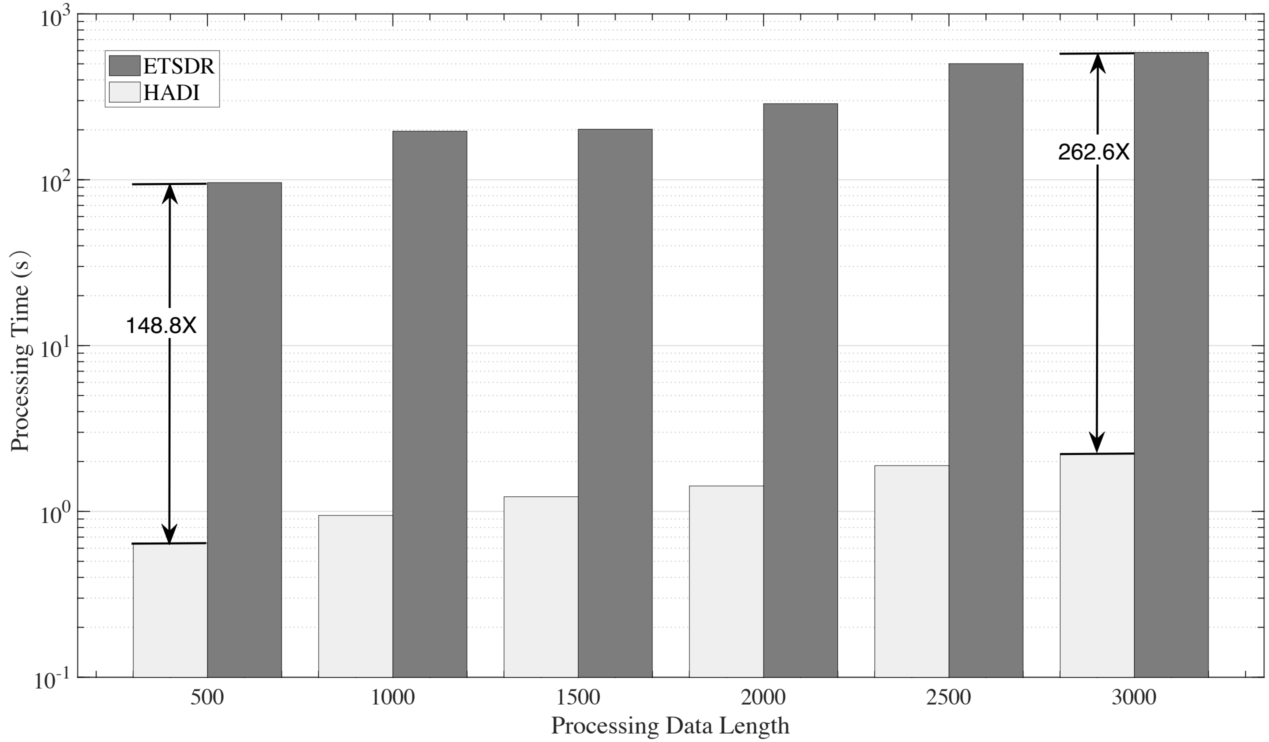


Fig. 19 Processing Time in Logarithmic Scale
*Probability of unavailable data = 10%

- It is obvious that the processing time of HADI is more than 100 times shorter than ETSDR
- Regardless of the processing data length grows, HADI keeps an extremely low processing time

Concluding Remarks

1. This research focused on sensors with low data availability, especially a unique sensor in the smart home environment
2. HADI succeeded in highly available data restoration and comparatively accurate
3. HADI shown a great performance on continual unavailable problem
4. HADI can enrich and enhance the correlation between heterogeneous sensors in smart home environment
5. Unlike previous works, HADI performs high efficiency on processing time, in which it can reduce the burden on the processor of smart automated system

Future Works

1. Focus on attempting to introduce more DI models and their correlations into HADI scheme, so that the automated systems are free from unavailable data
2. In this research, unavailable data of sensor type are observed in iHouse. It is necessary to figure out a detection method for identifying the unavailable data

Thanks for your attention!

DI Model Data Restoration #1

With Heterogeneous Data Interpolation

In

Available Target Data Aggregation

$$H_{Y_i}{}^T = \begin{bmatrix} \check{Y}_i(n-h) \\ \tilde{Y}_i(n-h+1) \\ \hat{Y}_i(n-h+2) \\ \check{Y}_i(n-h+3) \\ \vdots \\ \check{Y}_i(n-1) \end{bmatrix} = \begin{bmatrix} H_{Y_i}(\mathbf{1}) \\ H_{Y_i}(\mathbf{2}) \\ H_{Y_i}(3) \\ H_{Y_i}(\mathbf{4}) \\ \vdots \\ H_{Y_i}(\boldsymbol{h}) \end{bmatrix} \to \begin{bmatrix} H_{Y_i}(C(1)) \\ H_{Y_i}(C(2)) \\ H_{Y_i}(C(3)) \\ \vdots \\ H_{Y_i}(C(m)) \end{bmatrix} = \boldsymbol{H_{Y_i}}{}^{\boldsymbol{T}}(\boldsymbol{C})$$

And $\mathbb{Y} = H_{Y_i}(C)$ [Target Vector]

Find all available data(include restored data) in the horizon

Therefore, we compose a vector of available target data

Heterogeneous Data Aggregation

$$H_{X_{1:K}}{}^T = \begin{bmatrix} H_{X_1} \\ H_{X_2} \\ H_{X_3} \\ H_{X_4} \\ \vdots \\ H_{X_K} \end{bmatrix} = \begin{bmatrix} X_1(n-h) & \cdots & X_1(n-1) \\ \vdots & \ddots & \vdots \\ X_K(n-h) & \cdots & X_K(n-1) \end{bmatrix}$$

Integrate the heterogeneous into matrix

Available Heterogeneous Data Aggregation

$$(H_{X_{1:K}}(C))^T = \begin{bmatrix} H_{X_1}(C) \\ H_{X_2}(C) \\ H_{X_3}(C) \\ H_{X_4}(C) \\ \vdots \\ H_{X_K}(C) \end{bmatrix} = \begin{bmatrix} X_1(C(1)) & \cdots & X_1(C(m)) \\ \vdots & \ddots & \vdots \\ X_K(C(1)) & \cdots & X_K(C(m)) \end{bmatrix}$$

Then $\mathbb{X} = (H_{X_{1:K}}(C))^T$

Mark the simultaneous data corresponding to set C,

Compose a synchronous matrix

$$Y = f(X)$$

Data Transformation

$$\tilde{\mathbb{Y}} = (\tilde{Y}_{1:K}(C))^T = \begin{bmatrix} \tilde{Y}_1(C) \\ \tilde{Y}_2(C) \\ \tilde{Y}_3(C) \\ \tilde{Y}_4(C) \\ \vdots \\ \tilde{Y}_K(C) \end{bmatrix} = \begin{bmatrix} \tilde{Y}_1(C(1)) & \cdots & \tilde{Y}_1(C(m)) \\ \vdots & \ddots & \vdots \\ \tilde{Y}_K(C(1)) & \cdots & \tilde{Y}_K(C(m)) \end{bmatrix}$$

Hence, we have *K* groups simulated data

$\tilde{\mathbb{Y}}$ represents the matrix of restored data

$$RMSE = \sqrt{\frac{(\mathbb{Y}-\tilde{\mathbb{Y}})^2}{m}}$$

DI Model Data Process

Restored Data Election

$$\Phi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \vdots \\ \varphi_K \end{bmatrix} = \begin{bmatrix} \sqrt{\dfrac{(\tilde{Y}_1(C(1)) - H_{Y_i}(C(1)))^2 + \cdots + (\tilde{Y}_1(C(m)) - H_{Y_i}(C(m)))^2}{m}} \\ \vdots \\ \sqrt{\dfrac{(\tilde{Y}_K(C(1)) - H_{Y_i}(C(1)))^2 + \cdots + (\tilde{Y}_K(C(m)) - H_{Y_i}(C(m)))^2}{m}} \end{bmatrix}$$

Assume that $\varphi_p = \min[\Phi]$,

location set p is voted to compute the current data

Current Data Generation

$$\tilde{Y}_i(n,t) = f(X_p(n,t))$$

Eventually, data restoration at the current time t is accomplished

Out

DI Model Data Restoration #1

Proposed Algorithm (Heterogeneous)

Symbol	Description
n	Number of data
t	A certain time instance
K	Number of locations
C	Set of sequence number of available data in the horizon
m	Number of available data in target data horizon
i	A set of location of target data and $i \in [1, K]$
j	A set of other locations and $i \cup j = K$
$Y_i(n,t)$	Target data at time t
$\check{Y}_i(n,t)$	Available target data at time t
$\hat{Y}_i(n,t)$	Unavailable target data at time t
$\tilde{Y}_i(n,t)$	Restored data by heterogeneous data
$\check{Y}_i(C(1{:}m)$	Available target data in horizon
$X_{1:K}(n,t)$	Set of other heterogeneous data at time t
$X_{1:K}(C(1{:}m))$	Heterogeneous data in horizon at the same moment with target data
A_{H_Y}	Percentage of available data in horizon
$H_{X_{1:K}}$	Horizon of other heterogeneous data
H_{Y_i}	Horizon of target data
θ_y	Threshold of target data
h	Determined length of horizon (h = 120 data equal to 10 min)
p	Set of locations determine the minimum root mean square error (RMSE)
q	Set of locations determine the maximum absolute Pearson correlation coefficient

Algorithm 1 DI Model (Heterogeneous Data Interpolation)

if $Y_i(n,t) > \theta_y$ **then** // $Y_i(n,t)$ is unavailable.
  $\hat{Y}_i(n,t) \leftarrow Y_i(n,t)$

Availability Identification

  // Determine spatial and temporal horizon as:
  $H_{Y_i} = Y_i(n-h) : Y_i(n-1)$
  $H_{X_{1:K}} = X_{1:K}(n-h) : X_{1:K}(n-1)$
  if $A_{H_{Y_i}} < 50\%$ **then**
    // Available data are rare, reconsider horizon with restored data
    $H_{Y_i} = \left\{\check{Y}_i \bigcup \tilde{Y}_i\right\}(n-h) : \left\{\check{Y}_i \bigcup \tilde{Y}_i\right\}(n-1)$
  end if

Data Preparation

  // Aggregate the available data set in horizon $H_{Y_{(i)}}$
  $\check{Y}_i(C(1:m)) = \left[\check{Y}_i(C(1)), \cdots, \check{Y}_i(C(m))\right]$
  for each $u = 1 : K$ **do**
    $X_u(C(1:m)) = [X_u(C(1)), \cdots, X_u(C(m))]$
    // Substitute $X_u(C(1:m))$ for Data Interpolation(DI) model
    $\tilde{Y}_u(C(1:m)) = \left[\tilde{Y}_u(C(1)), \cdots, \tilde{Y}_u(C(m))\right]$
    // Calculate the Root Mean Square Error
    $\varphi(u) = \sqrt{\frac{\left(\tilde{Y}_u(C(1)) - \check{Y}_u(C(1))\right)^2 + \cdots + \left(\tilde{Y}_u(C(m)) - \check{Y}_u(C(m))\right)}{m}}$
  end for
  $\varphi(p) = min(\varphi(1:K))$
  // Calculate current time $\tilde{Y}_i(n,t)$
  $\tilde{Y}_i(n,t) = f(X_p(n,t))$

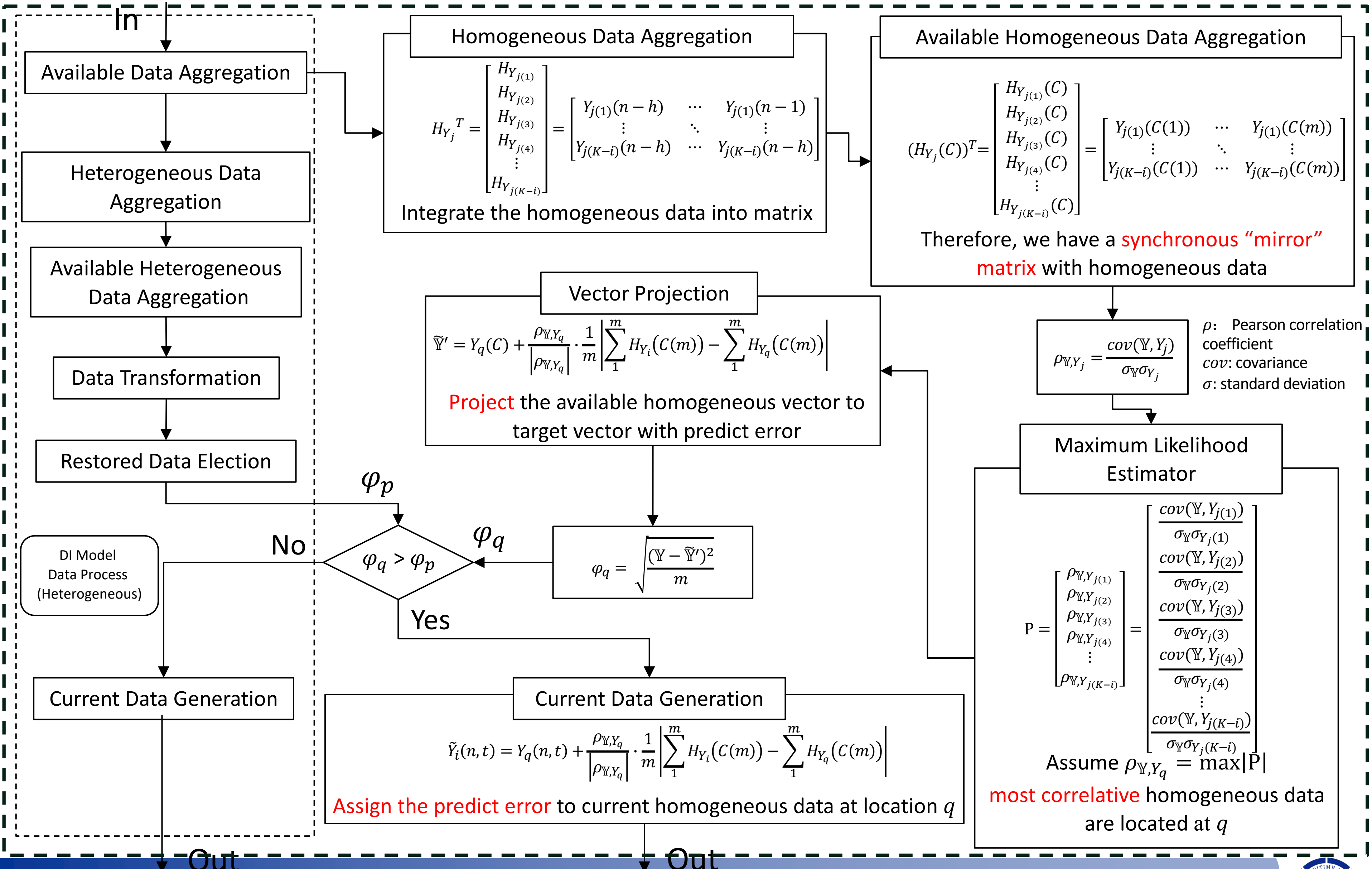
DI Model with Heterogeneous Data Interpolation

else
  // $Y_t(n, l_i)$ is available.
  $\check{Y}_t(n, l_i) \leftarrow Y_t(n, l_i)$
end if

Algorithm DI Model with Heterogeneous Data Interpolation

DI Model Data Restoration #2

With Heterogeneous & Homogeneous Data Interpolation



DI Model Data Restoration #2

Proposed Algorithm (Heterogeneous & Homogenous)

Symbol	Description
n	Number of data
t	A certain time instance
K	Number of locations
C	Set of sequence number of available data in the horizon
m	Number of available data in target data horizon
i	A set of location of target data and $i \in [1, K]$
j	A set of other locations and $i \cup j = K$
$Y_i(n,t)$	Target data at time t
$\check{Y}_i(n,t)$	Available target data at time t
$\hat{Y}_i(n,t)$	Unavailable target data at time t
$\tilde{Y}_i(n,t)$	Restored data by heterogeneous data
$\check{Y}_i(C(1{:}m)$	Available target data in horizon
$X_{1:K}(n,t)$	Set of other heterogeneous data at time t
$X_{1:K}(C(1{:}m))$	Heterogeneous data in horizon at the same moment with target data
A_{H_Y}	Percentage of available data in horizon
$H_{X_{1:K}}$	Horizon of other heterogeneous data
H_{Y_i}	Horizon of target data
θ_y	Threshold of target data
h	Determined length of horizon (h = 120 data equal to 10 min)
p	Set of locations determine the minimum root mean square error (RMS
q	Set of locations determine the maximum absolute Pearson correlation coefficient

Algorithm 2 DI Model (Homogeneous & Heterogeneous Data Interpolation)

Availability Identification

if $Y_i(n,t) > \theta_y$ **then** // $Y_i(n,t)$ is unavailable.

$\quad \hat{Y}_i(n,t) \leftarrow Y_i(n,t)$

Heterogeneous Data Preparation

$\quad H_{Y_i} = Y_i(n-h) : Y_i(n-1)$ // Target data horizon

$\quad H_{X_{1:K}} = X_{1:K}(n-h) : X_{1:K}(n-1)$ //Heterogeneous data horizon (spatial)

Homogeneous Data Preparation

$\quad H_{Y_j} = Y_j(n-h) : Y_j(n-1)$ // Homogeneous data horizon (spatial)

$\quad$ **if** $A_{H_{Y_i}} < 50\%$ **then**

$\quad\quad H_{Y_i} = \left\{\check{Y}_i \bigcup \tilde{Y}_i\right\}(n-h) : \left\{\check{Y}_i \bigcup \tilde{Y}_i\right\}(n-1)$

$\quad$ **end if**

DI Model with Heterogeneous Data Interpolation

$\quad \hat{Y}_i(C(1:m)) = \left[\hat{Y}_i(C(1)), \cdots, \hat{Y}_i(C(m))\right]$ // $\mathbb{Y} = \hat{Y}_i(C(1:m))$

$\quad$ **for each** $u = 1 : K$ **do**

$\quad\quad X_u(C(1:m)) = [X_u(C(1)), \cdots, X_u(C(m))]$

$\quad\quad \tilde{Y}_u(C(1:m)) = \left[\tilde{Y}_u(C(1)), \cdots, \tilde{Y}_u(C(m))\right]$

$\quad\quad \varphi(u) = \sqrt{\frac{\left(\tilde{Y}_u(C(1)) - \hat{Y}_u(C(1))\right)^2 + \cdots + \left(\tilde{Y}_u(C(m)) - \hat{Y}_u(C(m))\right)^2}{m}}$

$\quad$ **end for**

$\quad \varphi(p) = min(\varphi(1:K))$

DI Model with Homogeneous Data Interpolation

$\quad$ **for each** $v = j(1) : j(end)$ **do**

$\quad\quad Y_v(C(1:m)) = [Y_v(C(1)), \cdots, Y_v(C(m))]$

$\quad\quad \rho_{\mathbb{Y},Y_v} = \frac{cov(\mathbb{Y}, Y_v)}{\sigma_{\mathbb{Y}} \sigma_{Y_v}}$

$\quad$ **end for**

$\quad \rho_{\mathbb{Y},Y_{j(q)}} = max|\rho_{\mathbb{Y},Y_{j(1:end)}}|$

$\quad \tilde{\mathbb{Y}}' = Y_{j(q)}(C) + \frac{\rho_{\mathbb{Y},Y_{j(q)}}}{|\rho_{\mathbb{Y},Y_{j(q)}}|} \cdot \frac{1}{m} \left| \sum_1^m H_{Y_i}(C(m)) - \sum_1^m H_{Y_{j(q)}}(C(m)) \right|$

$\quad \varphi(q) = \sqrt{\frac{(\mathbb{Y} - \tilde{\mathbb{Y}}')^2}{m}}$

$\quad$ **if** $\varphi(p) > \varphi(q)$ **then** // Execute heterogeneous interpolation.

$\quad\quad \tilde{Y}_i(n,t) = f(X_p(n,t))$ // Heterogeneous interpolation end.

$\quad$ **else**

$\quad\quad \tilde{Y}_i(n,t) = Y_{j(q)}(n,t) + \frac{\rho_{\mathbb{Y},Y_{j(q)}}}{|\rho_{\mathbb{Y},Y_{j(q)}}|} \cdot \frac{1}{m} \left| \sum_1^m H_{Y_i}(C(m)) - \sum_1^m H_{Y_{j(q)}}(C(m)) \right|$

$\quad$ **end if**

else

$\quad$ // $Y_t(n, l_i)$ is available.

$\quad \check{Y}_t(n, l_i) \leftarrow Y_t(n, l_i)$

end if

Algorithm DI Model with Heterogeneous & Homogeneous Data Interpolation

Solar Irradiance Data Restoration

Equations for Solar Irradiance Restoration

X

- Temperature
- Relative Humidity
- Date & Time
- Geographical Information

$f(X)$

$$l = 0.44 - 0.16\sqrt{[(|\emptyset| - 80)/60]^2 + [(d - 120)/(263 - |\emptyset|)]^2}$$
$$w = 0.00493(RH)T^{-1}\exp[26.23 - 5416T^{-1}]$$
$$\beta = (0.025 + 0.1\cos^2\emptyset)\exp(-0.7z/1000)$$

$$m' = m \cdot p/p_0$$
$$p = p_0 \cdot \exp\left(-\frac{z}{H_T}\right) \quad H_T = 8340\text{m}$$
$$m = 1/[\sin(h) + 0.15(57.296h + 3.885)^{-1.253}]$$

$$\bar{\tau}_{oz} = \exp[-0.0365(ml)^{0.7136}]$$
$$\bar{\tau}_w = \min[1.0, 0.909 - 0.036\ln(mw)]$$
$$\bar{\tau}_g = \exp(-0.0117m'^{0.3139})$$
$$\bar{\tau}_a = \exp\{-m\beta[0.6777 + 0.1464(m\beta) - 0.00626(m\beta)^2]\}^{-1.3}$$
$$\bar{\tau}_r = \exp[-0.008735m'(0.547 + 0.014m')^2 - 0.00038m'^2 + 4.6 * 10^{-6}m'^3]^{-4.08}$$

$$\bar{\tau}_{d,clear} \approx 0.5\,[\,\bar{\tau}_{oz}\,\bar{\tau}_w\,\bar{\tau}_g\,(1 - \bar{\tau}_r\,\bar{\tau}_a) + 0.013]$$
$$\bar{\tau}_{b,clear} \approx \max(0, \bar{\tau}_{oz}\,\bar{\tau}_w\,\bar{\tau}_g\,\bar{\tau}_r\,\bar{\tau}_a - 0.013)$$

$$R_{b,clear} = R_0\bar{\tau}_{b,clear}$$
$$R_{d,clear} = R_0\bar{\tau}_{d,clear}$$

$$\tau_c = 0.4560 + \frac{0.3566n}{N} + 0.1874(n/N)^2$$

$$R_{clear} = R_{b,clear} + R_{d,clear}$$

Solar Irradiance $\boldsymbol{R}$ $\rightarrow$ $\boldsymbol{R = \tau_c R_{clear}}$

Y

Solar Irradiance

R_{clear} : global solar irradiance

$R_{b,clear}$: surface solar beam irradiance

$R_{d,clear}$: solar diffuse irradiance

R_0 : solar radiation at the top of the atmosphere

m: air mass

m': air mass pressure-corrected air mass

β: Angstrom turbidity coefficient

H_T: the scale height of isothermal atmosphere

τ_c : radiative transmittance due to cloud extinction

$\tau_{b,clear}$: solar beam radiative transmittance under clear skies

$\tau_{d,clear}$: solar diffuse radiative transmittance under clear skies

τ_{oz} : radiative transmittances due to ozone absorption

τ_w : radiative transmittances due to water vapour absorption

τ_g : radiative transmittances due to permanent gas absorption

τ_a: radiative transmittances due to aerosol extinction

τ_r: radiative transmittances due to Rayleigh scattering

RH: relative humidity

T: temperature (Kelvin)

h: solar elevation(radian)

$\emptyset$: latitude

z: surface elevation form the mean sea level