

การพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ โดยใช้วิธีการเติมเต็มแบบจัดกลุ่มข้อมูลให้สมบูรณ์
และวิธีประสานผลของตัวแบบโครงข่ายประสาท



นางสิรภัทร เชี่ยวชาญวัฒนา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-17-6750-1

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

INCOMPLETE TIME-SERIES DATA FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE
AND ENSEMBLING NEURAL NETWORK MODEL

Mrs. Sirapat Chiewchanwattana

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Science

Department of Mathematics

Faculty of Science

Chulalongkorn University

Academic year 2005

ISBN 974-17-6750-1

481901

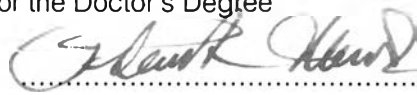
Thesis Title INCOMPLETE TIME-SERIES DATA FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE AND ENSEMBLING NEURAL NETWORK MODEL

By Mrs. Sirapat Chiewchanwattana

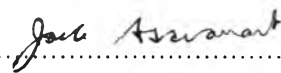
Field of Study Computer Science


Thesis Advisor Professor Chidchanok Lursinsap, Ph.D.


Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctor's Degree

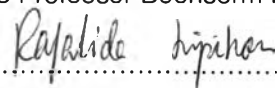
 Dean of the Faculty of Science
(Professor Piamsak Menasveta, Ph.D.)

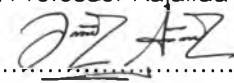
THESIS COMMITTEE

 Chairman
(Associate Professor Jack Asavanant, Ph.D.)

 Thesis Advisor
(Professor Chidchanok Lursinsap, Ph.D.)

 Member
(Associate Professor Boonserm Kijirikul, Ph.D.)

 Member
(Assistant Professor Rajalida Lipikorn, Ph.D.)

 Member
(Chularat Tanprasert, Ph.D.)

สิริภัทร เชี่ยวชาญวัฒนา : การพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ โดยใช้วิธีการเติมเต็มแบบจัดกลุ่มข้อมูลให้สมบูรณ์ และวิธีประสานผลของตัวแบบโครงข่ายประสาท :

(INCOMPLETE TIME-SERIES DATA FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE AND ENSEMBLING NEURAL NETWORK MODEL). อ. ที่ปรึกษา :

ศาสตราจารย์ ดร. ชิดชนก เหลือสินทรัพย์, 133 หน้า. ISBN 974-17-6750-1.

วิทยานิพนธ์นี้นำเสนอ การพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ โดยอาศัยการจำลองรูปแบบของโครงข่ายประสาทเทียม ซึ่งการจำลองนั้นสามารถแบ่งได้เป็นสองขั้นตอนดังนี้ ขั้นตอนหนึ่ง ทำการเติมเต็มข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์นั้นให้สมบูรณ์ ในขั้นตอนที่สองทำการพยากรณ์ข้อมูลอนุกรมเวลาที่ได้จากขั้นตอนที่หนึ่ง การแก้ปัญหาในงานนี้คือพัฒนาแบบจำลองโครงข่ายประสาทเทียมใหม่ สำหรับการพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ และยังสามารถให้ความถูกต้องในการพยากรณ์เพิ่มขึ้นด้วย โดยได้นำเสนอแบบจำลองโครงข่ายประสาทเทียม สองแบบ แบบแรก ใช้วิธีการเติมเต็มข้อมูลแบบ EM หลายลักษณะ และวิธีการเติมเต็มข้อมูลแบบ Spline ซึ่งข้อมูลหลายๆ ชุดที่ถูกเติมเต็มจากหลายๆ วิธีนั้น จะถูกนำมาสอนโดยใช้โครงข่ายประสาทเทียม MLP โดยใช้แบบขยาย Kalman Filtering จากนั้นทำการประสานผลลัพธ์ของโครงข่ายประสาทเทียมทุกโครงข่ายเข้าด้วยกัน แบบจำลองโครงข่ายนี้ให้ชื่อว่า โครงข่าย FI-GEM แบบที่สองปรับเปลี่ยนมาใช้โครงข่ายประสาทเทียม FIR เพื่อทำการพยากรณ์ จากนั้นผลลัพธ์ของโครงข่ายประสาทเทียมทุกโครงข่ายจะถูกประสานเข้าด้วยกันโดยใช้วิธีการเลือกโครงข่ายแบบ genetic algorithm ให้ชื่อแบบจำลองโครงข่ายนี้ว่า โครงข่าย RMD-FSE นอกจากนั้นยังได้นำเสนอวิธีการเติมเต็มข้อมูลแบบใหม่ เพื่อปรับปรุงการประมาณค่าข้อมูลที่หายไปนั้นให้ได้ค่าที่ถูกต้องมากยิ่งขึ้น โดยได้ใช้เทคนิคการจัดกลุ่ม โดยอาศัยคุณลักษณะของรูปแบบข้อมูลที่มีอยู่จริง แนวคิดหลักคือ ทำการตัดแบ่งข้อมูลอนุกรมเวลาออกเป็นหลายๆ ชิ้นที่มีขนาดต่างๆ กัน วิธีการคำนวณหาค่าข้อมูลที่หายไป จะคำนวณหาจากชิ้นข้อมูลที่มีความคล้ายกับชิ้นที่มีข้อมูลที่หายไปมากที่สุด แล้วทำการคำนวณหาค่าข้อมูลที่หายไปนั้น ให้ชื่อว่า ขั้นตอนวิธี WDC ซึ่งสามารถให้ผลที่เทียบเท่าหรือดีกว่าวิธีอื่น เช่น EM, MI, OCSFCM และ Spline ในกรณีของข้อมูลอนุกรมเวลาที่ไม่คงที่

ภาควิชา คณิตศาสตร์
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2548

ลายมือชื่อนิติ.....
ลายมือชื่ออาจารย์ที่ปรึกษา.....

ก.ก.พ.

C.L.P.

4373850523 : MAJOR COMPUTER SCIENCE

KEY WORD: INCOMPLETE TIME-SERIES PREDICTION /MISSING DATA /CLUSTERING
FILL-IN TECHNIQUE / ENSEMBLING NEURAL NETWORK.

SIRAPAT CHIEWCHANWATTANA: INCOMPLETE TIME-SERIES DATA
FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE AND ENSEMBLING
NEURAL NETWORK MODEL. THESIS ADVISOR: PROF. CHIDCHANOK
LURSINSAP, Ph.D., 133 pp. ISBN 974-17-6750-1.

This dissertation demonstrates the problem of incomplete time-series prediction by modelling the forecasting of several natural and social phenomena. The modeling consists of two main steps. The first step is to estimate the collected incomplete data, which are considered as missing data or missing values. The second step is to predict new data based on the nature of the data obtained from the first step. Our solution is to develop a new neural network model for forecasting incomplete time-series data and improving the accuracy of prediction. Two neural network models are proposed. First, various versions of EM-based algorithm and smoothing spline interpolation are used to preprocess the incomplete data sets. The individual networks are trained by supervised multilayer perceptron (MLP) with extended Kalman filtering. The ensemble construction is used for the combination of the individual networks. We name this type of network Fill-In - Generalized Ensemble Method (FI-GEM) networks. Second, each individual network uses a Finite Impulse Response model to perform the prediction. The outputs of all individual neural networks are combined by the genetic algorithm-based selective neural network ensemble method (GASEN). We denote this network as a reconstructed missing data-finite impulse response selective ensemble (RMD-FSE) network. Moreover, we proposed a new fill-in technique that is improved for estimating missing values based on clustering technique for characterizing the pattern of incomplete time-series data. The main idea is the time-series data are divided into separate subsequences of different sizes and, therefore, each subsequence can be viewed as a window. The imputation of missing samples is achieved by finding a complete subsequence similar to the missing sample subsequence and imputing the missing samples from this complete subsequence.

The imputation accuracy of the proposed algorithm, namely varied window clustering (WDC) algorithm is comparable or better than the others traditional methods such as: the spline interpolation, the multiple imputation (MI), and the optimal completion strategy fuzzy c-means algorithm (OCSFCM) in case of the non-stationary time-series data.

Department **Mathematics**

Field of study **Computer Science**

Academic year **2005**

Student's signature.....

Advisor's signature.....

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Chidchanok Lursinsap, for his contributions to all aspects of my work on this thesis and encouragement and for giving lots of his precious time to supervise this research work. There is one more unofficial supervisor for this thesis: Professor Chee Hung Henry Chu of the CACS, ULL. I thank him both for the close collaboration on this thesis. His sincere help and guidance have been extremely helpful to me.

The financial support of the ministry of university affairs of Thailand and in part by the National Electronic and Computer Technology Center (NECTEC), the National Science and Technology Development Agency (NSTDA) are also gratefully acknowledged. I would like to thank the thesis committee for their helpful comments and suggestions.

Special thanks are also expressed to Dr. Khamron Sunat, Professor Suchada Siripant and Sub.Lt. Supannika Lursinsap for giving me the encouragement and support. I am extremely grateful to my family, especially my parents, for their love, patience, and countless encouraging words.

Tables of Contents

	page
Thai Abstract.....	iv
English Abstract.....	v
Acknowledgments.....	vi
List of Tables.....	x
List of Figures.....	xii
List of Abbreviation.....	xix
List of Symbols.....	xx
Chapter	
1 Introduction.....	1
1.1 Motivation.....	2
1.2 Objectives.....	3
1.3 Area of interest.....	4
1.4 Performance Measure.....	5
1.4.1 Prediction Performance Index.....	6
1.4.2 Imputation Performance Index.....	7
1.5 Contributions of the Dissertation.....	8
1.6 Dissertation Organization.....	8
2 Incomplete Time-series Prediction.....	10
2.1 Managing Incomplete Data.....	10
2.1.1 Cubic Smoothing Spline Interpolation.....	11
2.1.2 Multiple Imputation by EM Algorithm: Random Selection....	11
2.1.3 Multiple Imputation by EM Algorithm: Average Selection...	12
2.1.4 Imputation by the regularized EM algorithm: Random Selection.....	12

Chapter	page
2.1.5 Imputation by the regularized EM algorithm: Average Selection.....	13
2.1.6 k-Segment Principal Curves.....	13
2.2 Time-Series Prediction.....	14
2.2.1 Multilayer Feedforward Neural Network.....	14
2.2.2 Ensemble Neural Network Design.....	15
3 Applications of Time-Series Prediction.....	16
3.1 FI-GEM Network.....	18
3.2 RMD-FSE Network.....	20
3.3 Extended RMD-FSE Network.....	22
3.4 Results.....	25
3.4.1 Results of FI-GEM Network [18].....	26
3.4.2 Results of RMD-FSE Network [19].....	28
3.4.3 Results of extended RMD-FSE Network [20].....	30
3.5 Summary.....	32
3.5.1 Cited Reference.....	32
3.5.2 Limitation.....	33
4 WDC algorithm.....	39
4.1 Classifying The Characteristic of Incomplete Time-Series Data.....	39
4.1.1 Pattern Characterization with Varied Window Sizes.....	42
4.1.2 The Imputation of Missing Value.....	43
4.2 Proposed Algorithm Based on Similarity Measure.....	45
4.3 Experimental Results and Discussion.....	48

Chapter	page
4.3.1 Mackey-Glass Chaotic Time-Series.....	49
4.3.2 The monthly sunspots.....	50
4.3.3 The daily gauge height at Ban Luang gauging station, Mae Tun stream, Ping river, Thailand.....	51
4.3.4 The daily air temperature at Nakhon Ratchasima province, Thailand.....	52
4.4 Discussion.....	52
4.4.1 Appropriate Partitioning Window Size.....	52
4.5 Summary.....	54
5 Conclusion.....	98
References.....	102
Appendix A.....	107
Appendix B.....	110
Vita.....	111

List of Tables

Table	page
3.1 Average performance index P'_d for Mackey-Glass chaotic time-series and sunspots data. There are six types of individual networks, which are used to compare with FI-GEM networks. We found that all values of P'_d are greater than zero.....	33
3.2 Average performance index P'_d for sunspots data and the daily gauge height data. We found that all values of P'_d are greater than zero.....	34
3.3 Average performance index P'_d for sunspots data and the daily gauge height data. We found that all values of P'_d are greater than zero.....	34
4.1 The mean square error (MSE) and the Pearson's correlation (CORR) of Mackey-Glass data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.....	56
4.2 The mean square error (MSE) and the Pearson's correlation (CORR) of the sunspots data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.....	57
4.3 The mean square error (MSE) and the Pearson's correlation (CORR) of the gauge height data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.....	58
4.4 The mean square error (MSE) and the Pearson's correlation (CORR) of the air temperature data for the spline interpolation method, the MI method, the OCSFCM algorithm, and the WDC algorithm.....	59
4.5 Average imputation performance index P_{Imp} for Mackey-Glass data, sunspots data, gauge height data and air temperature data. We found that all values of P_{Imp} are greater than zero.....	60

4.6 The appropriate range of partitioning window size of the four case studies:
Mackey-Glass chaotic time-series, the monthly sunspots data, the gauge
height data and the air temperature. 89

List of Figures

Figure	page
2.1 Multilayer Feedforward Neural Networks.....	14
3.1 FI-GEM Network Structure.....	18
3.2 Prediction performance index for (a) Mackey-Glass, and (b) sunspots. There are six groups of bars for 2.5%, 5%, 7.5%, 10%, 12.5% and 15% missing as shown on the x-axis. Each Group has five individual networks and FI-GEM networks. From left to right, the first bar is for MLP for cubic smoothing spline, the second bar is for MLP for EM with random selection, the third bar is for MLP for the regularized EM with random selection, the fourth bar is for MLP for EM with average selection, the fifth bar is for MLP for the regularized EM with average selection, and the last bar is for FI-GEM network. Prediction performance index Pd is shown on the y-axis. The prediction performance of our proposed are the lowest at all levels of missing.....	35
3.3 Weight in the selective ensemble network assigned to different FIR net works for the (a) sunspots and (b) gauge height data sets. They differ depending on the amount of data points missing.....	36
3.4 Prediction performance index for the (a) sunspots and (b) gauge height data sets.....	36
3.5 Weight in the selective ensemble network assigned to different FIR net works for the (a) sunspots and (b) gauge height data sets. They differ depending on the amount of data points missing.....	37
3.6 Prediction performance index for the (a) sunspots and (b) gauge height data sets.....	38

Figure	page
4.1 The figure shows the two groups of time-series data. The directions of subsequences of data in each circle of pattern a are similar. It is also true in each of pattern b	40
4.2 The scatter plot of the minimum β values at various fixed subsequence lengths of the air temperature data.....	43
4.3 The subsequences \mathbf{v}_τ and \mathbf{v}_q have a very similar shape. Since the value of $x_m^{(v_\tau)}$ is missing, both left and right differences are used to impute the value of $x_m^{(v_\tau)}$. The circle denotes the missing value.....	44
4.4 A comparison of imputing values between the cubic spline interpolation and the linear interpolation from the similar subsequence \mathbf{v}_q . The result of the proposed linear interpolation from characterization pattern is better than that of the cubic spline interpolation.....	46
4.5 An example of the similarity measure and different window partitions. The circle denotes the missing value. (a) Partitioning with a width of K . (b) Partitioning with a width of $K + c$, for some constant c	47
4.6 The MSE of (a) Mackey-Glass data and (b) the sunspots data for spline interpolation, the MI method, the OCSFCM algorithm, and the WDC algorithm.....	54
4.7 The MSE of (a) gauge height data and (b) the air temperature for spline interpolation, MI method, the OCSFCM algorithm and the WDC algorithm.....	55

Figure	page	
4.8	Scatter plots of reconstructed versus true values of the missing data for the 60% missing case using the Mackey-Glass data set. Missing values were initially set to random values (top left) and filled-in by the WDC algorithm (top right). The WDC output can be compared to the reconstructions by spline interpolation (lower left) and by the MI method (lower right).....	60
4.9	Scatter plots of reconstructed versus true values of the missing data for the 30% missing case using the sunspot data set. Missing values were initially set to random values (top left) and filled-in by the WDC algorithm (top right). The WDC output can be compared to the reconstructions by spline interpolation (lower left) and by the MI method (lower right).....	61
4.10	Scatter plots of reconstructed versus true values of the missing data for the 50% missing case using the gauge height data set. Missing values were initially set to random values (top left) and filled-in by the WDC algorithm (top right). The WDC output can be compared to the reconstructions by spline interpolation (lower left) and by the MI method (lower right).....	62
4.11	Scatter plots of reconstructed versus true values of the missing data for the 50% missing case using the air temperature data set. Missing values were initially set to random values (top left) and filled-in by the WDC algorithm (top right). The WDC output can be compared to the reconstructions by spline interpolation (lower left) and by the MI method	

Figure	page
(lower right).....	63
4.12 For each scatter plot, the number of most-likely subsequences of (a) Mackey-Glass (b) sunspots (c) gauge height data and (d) the air temperature are plotted.....	64
4.13 Scatter plots of the reference MSE values versus segment length for all levels of missing of the Mackey-glass data set.....	65
4.14 Scatter plots of the reference MSE values versus segment length at 10% missing of the Mackey-glass data set.....	66
4.15 Scatter plots of the reference MSE values versus segment length at 20% missing of the Mackey-glass data set.....	67
4.16 Scatter plots of the reference MSE values versus segment length at 30% missing of the Mackey-glass data set.....	68
4.17 Scatter plots of the reference MSE values versus segment length at 40% missing of the Mackey-glass data set.....	69
4.18 Scatter plots of the reference MSE values versus segment length at 50% missing of the Mackey-glass data set.....	70
4.19 Scatter plots of the reference MSE values versus segment length at 60% missing of the Mackey-glass data set.....	71
4.20 Scatter plots of the reference MSE values versus segment length at 70% missing of the Mackey-glass data set.....	72
4.21 Scatter plots of the reference MSE values versus segment length for all levels of missing of the sunspots data set.....	73
4.22 Scatter plots of the reference MSE values versus segment length at 10%	

Figure	page
missing of the sunspots data set.....	74
4.23 Scatter plots of the reference MSE values versus segment length at 20% missing of the sunspots data set.....	75
4.24 Scatter plots of the reference MSE values versus segment length at 30% missing of the sunspots data set.....	76
4.25 Scatter plots of the reference MSE values versus segment length at 40% missing of the sunspots data set.....	77
4.26 Scatter plots of the reference MSE values versus segment length at 50% missing of the sunspots data set.....	78
4.27 Scatter plots of the reference MSE values versus segment length at 60% missing of the sunspots data set.....	79
4.28 Scatter plots of the reference MSE values versus segment length at 70% missing of the sunspots data set.....	80
4.29 Scatter plots of the reference MSE values versus segment length for all levels of missing of the gauge height data set.....	81
4.30 Scatter plots of the reference MSE values versus segment length at 10% missing of the gauge height data set.....	82
4.31 Scatter plots of the reference MSE values versus segment length at 20% missing of the gauge height data set.....	83
4.32 Scatter plots of the reference MSE values versus segment length at 30% missing of the gauge height data set.....	84
4.33 Scatter plots of the reference MSE values versus segment length at 40% missing of the gauge height data set.....	85

Figure	page
4.34 Scatter plots of the reference MSE values versus segment length at 50% missing of the gauge height data set.....	86
4.35 Scatter plots of the reference MSE values versus segment length at 60% missing of the gauge height data set.....	87
4.36 Scatter plots of the reference MSE values versus segment length at 70% missing of the gauge height data set.....	88
4.37 Scatter plots of the reference MSE values versus segment length for all levels of missing of the air temperature data set.....	90
4.38 Scatter plots of the reference MSE values versus segment length at 10% missing of the air temperature data set.....	91
4.39 Scatter plots of the reference MSE values versus segment length at 20% missing of the air temperature data set.....	92
4.40 Scatter plots of the reference MSE values versus segment length at 30% missing of the air temperature data set.....	93
4.41 Scatter plots of the reference MSE values versus segment length at 40% missing of the air temperature data set.....	94
4.42 Scatter plots of the reference MSE values versus segment length at 50% missing of the air temperature data set.....	95
4.43 Scatter plots of the reference MSE values versus segment length at 60% missing of the air temperature data set.....	96
4.44 Scatter plots of the reference MSE values versus segment length at 70% missing of the air temperature data set.....	97
5.1 The time-series data of (a) Mackey-Glass (b) sunspots (c) gauge height	

Figure	page
data and (d) the air temperature.....	106

List of Abbreviations

FI-GEM	Fill-In - Generalized Ensemble Method
GEM	Generalized Ensemble Method
RMD-FSE	Reconstructed Missing Data-Finite Impulse Response Selective Ensemble
WDC	Varied Window Clustering
MLP	Multi-Layer Perceptron
ML	Maximum Likelihood
EM	Expectation and Maximization
MI	Multiple Imputation
REG-EM	Regularized Expectation and Maximization
MAR	Missing at Random
OCSFCM	Optimal Completion Strategy Fuzzy C-Mean
MSE	Mean Square Error
CORR	Pearson's Correlation
GASEN	Genetic Algorithm-based Selective Neural Network Ensemble

List of Symbols

x_t	Value of a System at Time t
P_d	Performance Index
P'_d	Average Performance Index
E_i^T	Mean Square Error of the Tested Network of Run i
E_j^{RN}	Mean Square Error of the Reference Network of Run j
E_j^{RF}	Mean Square Error of Proposed Network of Run j
R	Number of Runs per Fill-In Method
L	Number of Different Percentage of Missing Data
M	Number of Missing Value
MSE	Mean Squared Error
x	Actual Values
\hat{x}	Prediction Values
CORR	Pearson's correlation
N_{ij}^{Actual}	the Normalized Zero Mean of the Actual Values
$N_{ij}^{Predict}$	the Normalized Zero Mean of the Estimated Values
P_{Imp}	the Average Imputation Performance Index
f_{GEM}	Outputs of All Individual Neural Networks
$f_i(\mathbf{x})$	Output Value of Network i
α_i	Weighting Parameter for Network i
C_{ij}	Elements of the Covariance Matrix of the Errors from the Function Estimators f_i and f_j

k	Window Size of k
$o_j^{(q)}$	Output Signal of j^{th} Neuron in the q^{th} Layer
$w_{i,j}^{(q)}$	Connection Weight coming from the i^{th} Neuron in the $(q - 1)$ Layer to the j^{th} Neuron in the q^{th} layer
δ	Similarity Correlation in terms of Cosine
K	Length of the Subsequences
β	Distance between two Subsequences
\mathbf{v}_τ	Target Subsequence
\mathbf{v}_q	Reference Subsequence
$x_m^{(\mathbf{v}_\tau)}$	the Considered Missing Value at Time m of the Target Subsequence \mathbf{v}_τ
$x_m^{(\mathbf{v}_q)}$	Value at Time m of the Reference Subsequence \mathbf{v}_q
θ_l	Left Difference between $x_{m-1}^{(\mathbf{v}_q)}$ and $x_{m-1}^{(\mathbf{v}_\tau)}$
θ_r	Right Difference between $x_{m+1}^{(\mathbf{v}_q)}$ and $x_{m+1}^{(\mathbf{v}_\tau)}$
$\hat{x}_m^{(\mathbf{v}_\tau)}$	Estimated Value of $x_m^{(\mathbf{v}_\tau)}$
$G_{(d\%)}$	the Goodness of Partitioning Window Size at $d\%$ of Missing
κ_j	Window Size at Order j
q_{κ_j}	the number of points in window of size κ_j