British Journal of Mathematics & Computer Science
15(4): 1-10, 2016, Article no.BJMCS.19773
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Binding Visualization Method and Numeric Method Together to Analyze Large Data - With a Case Study

Yousuo Zou[1*], Zhiqiang Chen[2] and Jinhe Xu[2]

[1]Computer Science Program, University of Guam, Mangilao, GU 96923, USA.
[2]Chien-Shiung Institute of Technology, Taicang, Jiangsu 215411, China.

Authors' contributions

This work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJMCS/2016/19773
Editor(s):
(1) Mohamed Rabea Eid Said, Department of Science and Mathematics, Assiut University, Egypt.
(2) Raducanu Razvan, Department of Applied Mathematics, Al. I. Cuza University, Romania.
(3) Metin Basarir, Department of Mathematics, Sakarya University, Turkey.
(4) Kai-Long Hsiao, Taiwan Shoufu University, Taiwan.
(5) Tian-Xiao He, Department of Mathematics and Computer Science, Illinois Wesleyan University, USA.
Reviewers:
(1) C. R. Kikawa, Tshwane University of Technology, South Africa.
(2) Shagufta Kanwal, International Islamic University, Pakistan.
(3) Carlos Polanco, Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran, Mexico.
(4) Anonymous, Kocaeli University, Turkey.
Complete Peer review History: http://sciencedomain.org/review-history/13895

Case Study

Received: 26th June 2015
Accepted: 18th February 2016
Published: 28th March 2016

Abstract

In this paper, the authors combine a data visualization method and a numeric analysis method, binding them together to analyze large data (Big Data). Also included is a case study selected from the water transportation industry – A port on the Yangtze River and its annual import & export productivity (10^4 tons /year) – used to show how our methods work step by step. Theoretical and practical efforts in this paper suggest that the proposed methods (or algorithms) are efficient for data analysis and data prediction.

Keywords: Big data analysis; visualization methods; the least square methods; case study; prediction of annual productivity; the Taicang port of Yangtze River.

**Corresponding author: E-mail: yjzou@uguam.uog.edu;*

1 Introduction

Data Science study data collection, data storage, data transport / communication, data usage, data analysis, data visualization, knowledge discovery, etc. There are huge amount of data (massive data or Big Data) generated every day in industry, business, education, sciences and engineering, etc. that need to be analyzed [1]. It is still a challenging task for Data Scientists, especially Computer Scientists and Mathematicians, to find fast and efficient data analysis methods (or algorithms) that can quickly and accurately produce data analysis results for knowledge discovery or for prediction of future performance in the selected data area [2].

Data Scientists often employ both data visualization methods and numeric analysis methods to analyze data. Data visualization methods and data numeric analysis methods are different analysis tools in Data Science. Data Scientists frequently use data visualization methods in data mining and knowledge discovery [3]. There are many techniques and software skills in data visualization for Data Scientists to learn and to use [4].

Numeric data analysis methods, such as the Least Square Methods, have long been used for data analysis. Their use in data analysis began as early as in 1805 by Legendre [5] and in 1809 by Gauss [6]. Many research articles and monographs in both modern day and in history have advanced these numeric data analysis methods [7,8]. These type of methods are called classical or traditional data analysis methods [9,10].

In this paper, we want to share our research methods and experience on binding the visualization method and numeric analysis method together to analyze selected industrial data. We also include a case study in order to show, step by step, how our methods of data analysis work well with the industrial data and how the methods can successfully make predictions for future performance in the selected industry (A water port of the Yangtze River and its annual import &export productivity, in 104 tons/year, is selected in this paper).

2 Data Visualization Methods and Their Software Tools

Data visualization involves the creation and study of the visual representation of data. A primary goal of data visualization is to communicate clearly and efficiently to users via information graphics such as 2-D or 3-D charts, curves, and graphs as well as their animation effects. An effective visualization helps users to analyze and reason through data and evidence. It makes complex data more accessible, understandable and usable.

Data visualization is a portion of Data Science frequently employed by Data Scientists. Data visualization methods are powerful tools for data analysis. It enables the users to communicate data or information by encoding the data into visual representations. There are many techniques and much software knowledge that challenge Data Scientists to use data visualization methods for the analysis of large data.

There are many good data visualization software tools available in data analysis. For example, Google has introduced 37 easy to use data visualization tools which are available on web [11]. A powerful data visualization software application (MatLab, Maple, Mathematica, SAS Visualization tool, etc.) can be selected depending on the type of data, the size of the data and the purposes of the data analysis. If the data set is not too large, then Microsoft Excel is a very convenient data visualization tool to be used.

3 Numeric Analytic Algorithms

In this research, we use the traditional data analysis methods (or algorithms): the Least Square Methods. The following tells how the methods work.

Assume there is a huge amount of experimental data y_i , where i = 1, 2, 3, …, n (n can be a very large integer) which contains or hides the law of the nature or function pattern that describes the data, Y (X, A), which needs to be discovered through data analysis, where $X = (x_1, x_2, x_3, \ldots, x_m)$, and A = (a, b, c, …, k). How do

we find the pattern, Y(X, A), which best fits the given data? We let Q(X, A) be the sum of square of the difference of (y_i – Y(X, A)), and Q(X, A) becomes minimum.

$$Q(X, A) = Q(X, a, b, c, \ldots, k)$$

$$= \sum_1^n (\boldsymbol{y_i} - \boldsymbol{Y(X, A)})^2 = Min \tag{1}$$

Therefore, we have

$$\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{a}} = \sum_1^n [\boldsymbol{y_i} - \boldsymbol{Y(X, a, b, \ldots, k)}] \frac{\partial \boldsymbol{Y(X, A)}}{\partial \boldsymbol{a}} = \boldsymbol{0}$$

$$\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{b}} = \sum_1^n [\boldsymbol{y_i} - \boldsymbol{Y(X, a, b, \ldots, k)}] \frac{\partial \boldsymbol{Y(X, A)}}{\partial \boldsymbol{b}} = \boldsymbol{0}$$

$$\vdots$$

$$\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{k}} = \sum_1^n [\boldsymbol{y_i} - \boldsymbol{Y(X, a, b, \ldots, k)}] \frac{\partial \boldsymbol{Y(X, A)}}{\partial \boldsymbol{k}} = \boldsymbol{0} \tag{2}$$

The above Equations (2) contains k equations. The format of *Y(X, A)* can be determined by the pattern of experimental data using visualization tools. For example, if the given data showed by a data visualization tool is a linear relationship with variables x, a, and b, we can assume *Y(X, A)* to be

$$Y(x; a, b) = ax + b \tag{3}$$

Now substitute Equation (3) into Equations (2), we then have two algebraic linear equations:

$$\begin{aligned} M_{11}\, a + M_{12}\, b &= M_{13} \\ M_{21}\, a + M_{22}\, b &= M_{23} \end{aligned} \tag{4}$$

M_{11}, M_{12}, M_{13}, M_{21}, M_{22}, M_{23} are all constants obtained by experimental data. Equations (4) can also be written in the format of matrix,

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{12} & M_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} M_{13} \\ M_{23} \end{pmatrix} \tag{5}$$

where

$$M_{11} = \sum_1^n \boldsymbol{x_i^2} \qquad M_{12} = \sum_1^n \boldsymbol{x_i}$$

$$M_{13} = \sum_1^n \boldsymbol{y_i}\, \boldsymbol{x_i} \qquad M_{21} = \sum_1^n \boldsymbol{x_i}$$

$$M_{22} = \sum_1^n \boldsymbol{1} \qquad M_{23} = \sum_1^n \boldsymbol{y_i} \tag{6}$$

We can now use the regular linear algebraic method to find the solution of the coefficients a and b by solving the above linear equations. When the numeric solution of a and b are found, the linear pattern or the natural laws hiding in the given data Y(x; a, b) = ax + b is found. We can use this quantitative formula to predict the future performance of the data area.

In the above, we only discuss how to solve for a linear data pattern. However, if the given original data is shown by visualization software tool to be a nonlinear function pattern, we substitute the nonlinear function pattern into Equations (2), and a group of nonlinear algebraic equations will be generated. We may use the Newton-Gaussian Iteration Method to solve the nonlinear equations (though a little complicated) [10], or we may use some type of Mathematical transformation to turn the nonlinear problem into a linear problem, and use the above linear algebraic method to find the solution [8].

When analyzing and processing the signal and image data pattern, we do not use the Least Square Numeric Method, as data values change very quickly. Instead, we use very efficient numeric analytic methods, called Sinc numeric Methods [12,13].

4 Data and Analysis Procedures: A Case Study

In order to show readers of this paper how our data analysis methods work, we include this case study, to show step by step how to use our analysis methods.

4.1 Data

A port of Yangtze River, the TaiCang Port, annual import & export productivity of material transportation are listed in Table 1, which are real-world 10 year's data of the port (1999 – 2008) in the unit of 104 tons /year.

Table 1. Taicang port of Yangtze River annual import & export materials (in 10^4 tons/year)

Year	x_i	y_i (10^4 tons/year)
1999	1	100.0
2000	2	240.40
2001	3	362.40
2002	4	462.00
2003	5	808.61
2004	6	1039.88
2005	7	1510.69
2006	8	2251.02
2007	9	3043.04
2008	10	4003.90

Our task is to use the data analysis results to predict the future three year's annual productivity performance of the Taicang Port of the Yangtze River.

4.2 Data analysis procedure

4.2.1 Step 1

Since the given data is not too large, we chose to use Microsoft Excel as our visualization tool. A computer program was written to execute our numeric data analysis algorithms and to import and export the data in Table 1 into the selected data visualization software tool to show the function pattern of the given data (Fig. 1).

4.2.2 Step 2

The graph shows that the function pattern of the data is a nonlinear exponential function, so we try to use the following function pattern to describe the given data:

$$Y(x; a, b) = a\, e^{bx} \tag{7}$$

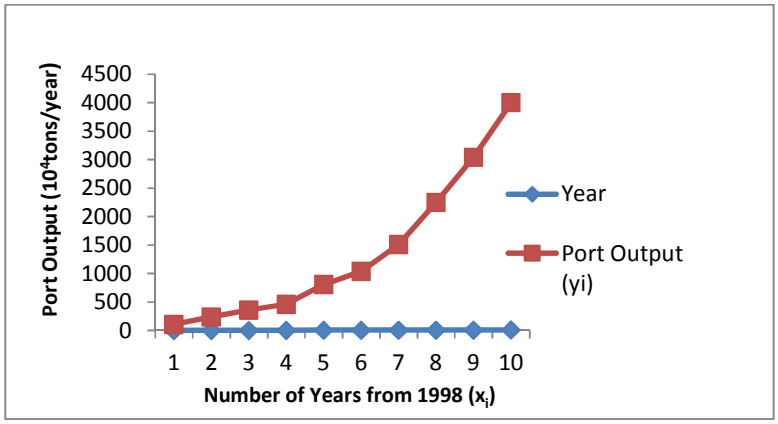

Fig. 1. Port output (10^4 tons/year) during 1999 – 2008

4.2.3 Step 3

If we substitute Equation (7) into the Least Square Equations (2), we will get two nonlinear algebraic equations with *a* and *b* as the variables. Of course, we can use the complicated Newton-Gaussian Iteration Method to solve the nonlinear equation to get *a* and *b*. However, we may also use a mathematical transformation to turn Equation (7) into a linear equation. We take a logarithm operation to both sides of Equation (7) and get

$$G(x; b, c) = bx + c \tag{8}$$

where

$$G(x; b, c) = \ln Y(x; a, b) \qquad c = \ln a \tag{9}$$

We have successfully turned the nonlinear Equation (7) into a linear Equation (8).

4.2.4 Step 4

We use a computer program to convert $g(x_i) = \ln(y_i)$. All the new converted data is listed in the following Table 2.

Table 2. Converted data from original data y_i

Year	x_i	$\ln(y_i)$
1999	1	4.7005
2000	2	5.4806
2001	3	5.8927
2002	4	6.1356
2003	5	6.6953
2004	6	6.9469
2005	7	7.3203
2006	8	7.7191
2007	9	8.0206
2008	10	8.2950

4.2.5 Step 5

Using the selected data visualization tool (Microsoft Excel in this case) and the data given in Table 2, Fig. 2 shows the function pattern of the new data set. Fig. 2 shows this is a very good linear function pattern described as in Equation (8).

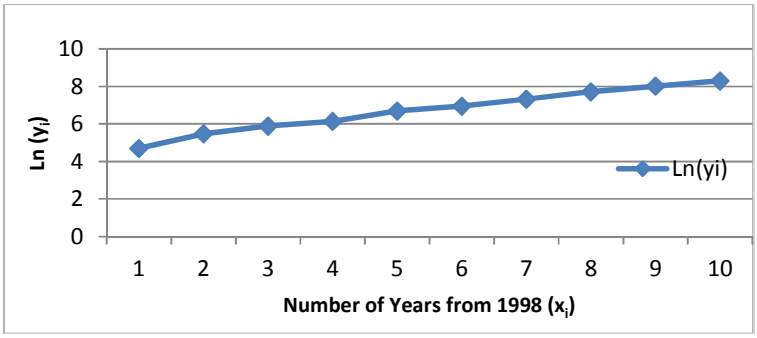

Fig. 2. Converted data, *ln* (y_i), in 1999 – 2008

4.2.6 Step 6

Substitute Equation (8) into the Least Square Equations (2), we get two linear algebraic equations similar to Equations (5) but with different values of M_{11}, M_{12}, M_{13}, M_{21}, M_{22}, M_{23} since the data in Table 2 are different from data in Table 1.

$$\begin{pmatrix} \boldsymbol{M_{11}} & \boldsymbol{M_{12}} \\ \boldsymbol{M_{21}} & \boldsymbol{M_{22}} \end{pmatrix} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M_{13}} \\ \boldsymbol{M_{23}} \end{pmatrix} \tag{10}$$

where

$M_{11} = \sum_1^{10} \boldsymbol{x_i^2} = 385$

$M_{12} = M_{21} = \sum_1^{10} \boldsymbol{x_i} = 55$

$M_{22} = \sum_1^{10} \boldsymbol{1} = 10$

$M_{13} = \sum_1^{10} \boldsymbol{g_i}\ \boldsymbol{x_i} = 401.1704$

$M_{23} = \sum_1^{10} \boldsymbol{g_i} = 67.2066$

Using the above constants and the regular linear algebraic routine to solve Equation (10) and get the numeric values of *b* and *c*.

$b = 0.3722$ $\quad$ $c = 4.6184$

Therefore, we have solved new linear Equation (8):

$$G(x; b, c) = 0.3722\ x + 4.6184 \tag{11}$$

4.2.7 Step 7

Using the selected data visualization tool (Microsoft Excel in this case) and the function pattern (Equation (11)) obtained from the data analysis, we draw a predictive linear line and compare it with the linear line

based on the data in Table 2 to see the two lines match closely or not (Fig. 3). Since the results obtained from the Least Square Methods are the most optimal and the best-fitting for the data given, we can see from Fig. 3 that the data line and the theoretical line match very well.

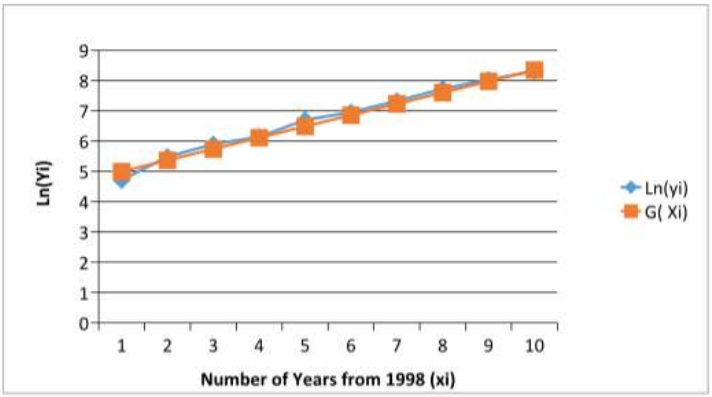

Fig. 3. Comparison of converted original data Ln (y_i) and predictive data G (x_i)

4.2.8 Step 8

Now turn the linear function pattern back into original nonlinear function pattern,

$$Y(x; a, b) = e^{G(x,b,c)} \qquad a = e^{c} = 101.332$$

So the original non-linear exponential function pattern is

$$Y(x; a, b) = 101.332\, e^{0.3722x} \tag{12}$$

5 Results and Discussion on Error Analysis

Using the predictive Equation (12) and Microsoft Excel, we draw the exponential curve and compare it with the curve generated by original data given in Table 1 (Please see Fig. 4). From Fig. 4, we can see that the theoretical function pattern (Red line) very closely matches the data curve (blue line). In Fig. 4, we also employ Formula (12) above to make predictions of the port's future performance in 2009, 2010, and 2011. Is this prediction accurate enough? We can use the actual data or performance of the Yangtze River Port to verify.

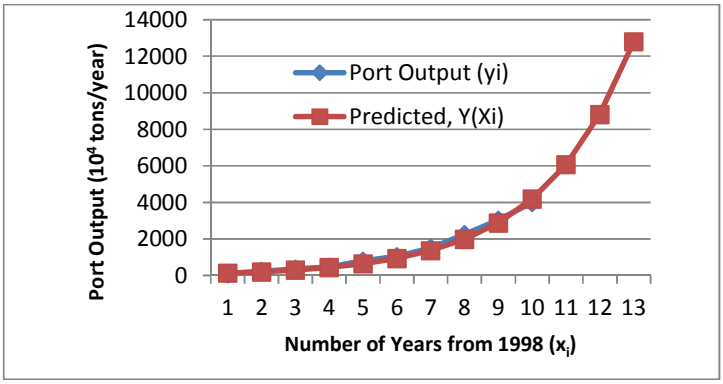

Fig. 4. Comparison of original port output data (y_i) and theoretically predictive data, Y(x_i)

The predictive numeric values are listed in the following Table 3 (in 10^4 tons/year). We can compare to the actual data when the data becomes available.

Table 3. Comparison of predicted and actual annul output of the Yangtze River port (10^4 tons/year)

Year	2009	2010	2011
x_i	11	12	13
Predicted, $Y(x_{i)}$	6079.041	8820.221	12797.463
Actually achieved, y_i	To be verified	To be verified	To be verified

Finally, let's discuss the errors of data analysis generated by the Least Square methods and possible ways to reduce such errors [7].

We first use computer programs to conduct error analysis. Table 4 and Fig. 5 in the following show our statistical results and the errors between original given data and the Least Square Method generated theoretically predicted data. We can see from the Fig. 5 that direct error between original data and predicted data is very small and almost equally distributed at each data point. However the total Least Squared error is huge: 184658.4, and the error distributed along each data point is hugely different. In order to make the Least Squared total error smaller, we can use a weighted factor to the Least Squared error value at each data point, as discussed in reference [14].

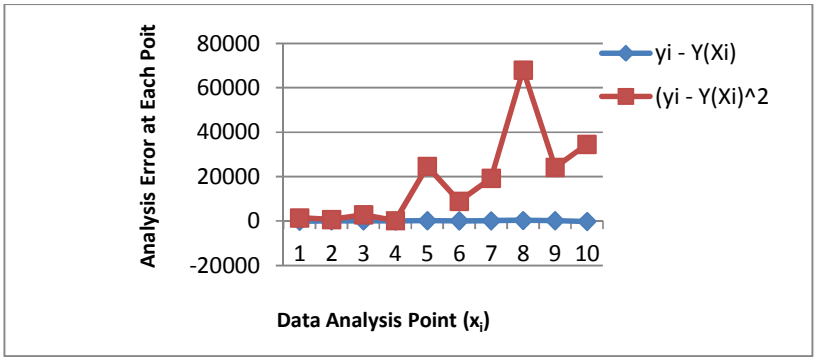

Fig. 5. Comparison of differences (errors) and squared differences (squared errors) between original data and predicted data at each data point

Table 4. Statistical analysis of errors between original data y_i and theoretically predicted data $Y(x_i)$

Error type	Error formula	Total error value
Total difference of given and predicted data	$\sum_1^{10}(y_i - Y(x_i))$	667.3
Total least square error	$Q(X) = \sum_1^{10}(y_i - Y(x_i))^2$	184658.4
Mean value of original data	$\bar{y} = \frac{1}{n}\sum_1^n y_i$	1383.2
Total difference of given data and mean value	$\sum_1^{10}(y_i - \bar{y})$	0.0
Total variance	$\delta_n^2 = \frac{1}{n}\sum_1^n(y_i - \bar{y})^2$	1565911.1
Standard total deviation	δ_n	1251.6

6 Summary and Discussion

In this paper, we have briefly discussed the physical and mathematical principles of visualization methods and the Least Square numerical methods for analysis of massive or Big Data. As an application and

demonstration, we used a port's industrial data as a case study to show, step by step, how to bind the two methods together to analyze massive data in industries. We employed a mathematical transformation and turned the nonlinear data pattern problem into a linear data pattern and easily solved the data analysis problem. Finally, we analyzed and discussed the errors of the data analysis methods. The research suggests that the combination of both graphic and numeric methods is an efficient way to analyze massive or Big Data.

However, there are still great challenges in the data analysis process that the authors of this paper should consider and make further efforts toward improving. First, the methods by which the group of equations, Equations (2) with different linear and nonlinear data patterns Y(X, A), was solved can be improved upon. It is a challenging job to solve a large group of equations, especially a large group of nonlinear equations. Second, it is a not easy task to determine the data pattern or function that is accurately modeling or describing the selected experimental data since Big Data is characterized by 4 huge Vs (Volume, Velocity, Variety, and Veracity). Any simple data pattern (or function) would be unable to exactly simulate the massive experimental data of the veracity that makes the Least Square methods produce errors in data analysis. Though the Weighted Least Square Method (W-LSQM) has been proposed to reduce the error of data analysis [14], there is still a long way to go in reducing the errors of the Least Square methods. Therefore the authors of this research are trying to apply the new numeric methods, the Sinc Methods, into the Big Data analysis [13].

Disclaimer

Some part of this manuscript was previously presented and published in the following conference.

Conference name: The International Conference On Green And Human Information Technology (ICGHIT).

Dates: Feb. 4-6, 2015

Location: Da Nang, Vietnam.

Web Link of the proceeding:
http://cdn1.spotidoc.com/store/data/000701930.pdf?key=ce05ce2bc06023a7020e8d27c38037ef&r=1&fn=701930.pdf

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Yadav C, Wang S, Kumar M. Algorithm and approaches to handle large data – a survey. International Journal of Computer Science and Network. 2014;2(3):2277–2420.

[2] Reichman OJ, Jones MB, Schildhaver MP. Challenges and opportunities of open data in ecology. Science. 2011;331(6018):703–705.
DOI: 10.1126/Science 1197962

[3] Fayyad U, Grinstein GG, Wierse A, editors. Information visualization in data mining and knowledge discovery. Academic Press. 2002;391.
ISBN 1-555860-689-0

[4] Ehlschlaeger CR, Shortridge AM, Goodchild MF. Visualizing spatial data uncertainty using animation. Computers & Geosciences. 1997;23(4):387–395.

[5] Legendre AM. Nouvelles méthodes pour la détermination des orbites des comètes. Paris: Firmin Didot; 1805. French.

[6] Gauss CF. Théorie motus corporum coelestium in sectionibus coni cis solem. Ambientum; 1809. French.

[7] Lichten W. Data and error analysis. Allyn and Bacon, Inc. 1988;171. ISBN 0-205-1193-9

[8] Petras I, Bednarova D. Total least square approach to modeling: A mat lab toolbox. Acta Montanistica Slovaca. 2010;15(2):158–170.

[9] Bjorck A. Numerical methods for least square problems. SIAM; 1996. ISBN 978-0-89871-360-2.

[10] Wolberg J. Data analysis using the method of least squares: Extracting the most information from experiments. Springer; 2005. ISBN 3-540-25674-1.

[11] Creative Bloq. The 37 best tools for data visualization; 2015.
Available: http://www.creativebloq.com/designtools/data-visualization-712402

[12] Yaroslavsky LP. DFT and DCT based discrete reconstruction. Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis. 2003;1:405–410.

[13] Stenger F. Handbook of Sinc numeric methods. CRC Press; 2011;463. ISBN 978-1-4398-2158-9

[14] Strutz T. Data fitting and uncertainty – a practical introduction to weighted least squares and beyond. Vieweg and Teubner; 2010. ISBN 978-3-8348-1022-9.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/13895