

Prediction of Productivity of A Yangtze River Port Based on Data Analysis Using Combination of Visualization Method and Numeric Method - A Case Study

Abstract

There are huge data in industry, business, education, life science, meteorology, environmental and climate change studies (or to be called as Big Data) need to be analyzed. How to find effective and high-performance analytic algorithms that can fast and accurate predict the future performance of the selected data area based on the results of data analysis? This is still a challenge to Data Scientists, especially to Computer Scientists and Mathematicians. In this paper, the authors combine both data visualization method and numeric analysis method and bind them together to analyze large data(Big Data). We also included a case study selected from water transportation industry – A port of the Yangtze River and its annual import & export productivity (10^4 tons /year) as the original data, step by step, to show how our methods work. Theoretical and practical efforts in this paper suggest that the proposed methods (or algorithms) are efficient in data analysis and data prediction.

Key words *Big data analysis, Visualization Methods Least Square Methods, Case Study, Prediction of A Yangtze River Port Productivity*

1. Introduction

Data Science studies data collection, data storage, data transport / communication, data usage, data analysis, data visualization and knowledge discovery, etc. There are huge amount of data (massive data or Big Data) that are generated everyday in industry, business, education, sciences and engineering 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

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 the information graphics, such as 2-D or 3-D charts, curves, and graphs as well as their animation effects. An effective visualization helps users in analyzing and reasoning about data 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].

The numeric data analysis methods, such as the Least Square Methods, was long been used for data analysis. It started to be used in data analysis as early as in 1805 by Legendre [5] and in 1809 by Gauss [6]. There are many research articles and monographs in modern days and in history advanced the numeric data analysis methods [7 & 8]. This type of methods is 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 the selected industrial data. We also include a case study to show, step by step, how our methods of data analysis work well with the industrial data and how the methods can successfully make prediction for the future performance in the selected industry (A water port of the Yangtze River and its annual import & export productivity, in 10^4 tons/year, is selected in this paper).

2. Data Visualization Methods and their Software Tools

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 makes the users be able to communicate data or information by encoding it as visual objects .

There are many techniques and software knowledge to challenge Data Scientists to use the data visualization methods for analysis of large data.

There are many good data visualization software tools available in data analysis. The Google introduced 36 easy to use data visualization tools which are available on web. Some powerful data visualization software, such as MatLab, Maple, Mathematica, SAS Visualization tool, etc., can be selected depending what types of data, the size of the data and the purposes of the data analysis. If the data set is not too large, the 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 are huge number of experimental data, y_i , $i = 1, 2, 3, \dots, n$ (n can be a very large integer number) which contains or hides the law of the nature or function pattern that describes the data, $Y(X, A)$, need to be discovered, where $X = (x_1, x_2, x_3, \dots, x_m)$, and $A = (a, b, c, \dots, k)$. Then how to 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.

$$\begin{aligned} Q(X, A) &= Q(X, a, b, c, \dots, k) \\ &= \sum_1^n (y_i - Y(X, A))^2 = \text{Min} \end{aligned} \quad (1)$$

Therefore, we have

$$\begin{aligned} \frac{\partial Q}{\partial a} &= 2 \sum_1^n [y_i - Y(X, a, b, \dots, k)] \frac{\partial Y(X, A)}{\partial a} = 0 \\ \frac{\partial Q}{\partial b} &= 2 \sum_1^n [y_i - Y(X, a, b, \dots, k)] \frac{\partial Y(X, A)}{\partial b} = 0 \\ &\vdots \\ \frac{\partial Q}{\partial k} &= 2 \sum_1^n [y_i - Y(X, a, b, \dots, k)] \frac{\partial Y(X, A)}{\partial k} = 0 \end{aligned}$$

nonlinear function pattern into Equations (2), then 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

(2)

The above Equations (2) contain k of equations. The format of $Y(X, A)$ can be determined by the pattern of graphics of experimental data showing by the visualization tools. For example, if the given data showed by the data visualization tool is a linear relationship with variables, we can assume $Y(X, A)$ as

$$Y(x; a, b) = ax + b \quad (3)$$

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

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

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

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} M_{13} \\ M_{23} \end{pmatrix} \quad (5)$$

Where

$$\begin{aligned} M_{11} &= \sum_1^n x_i^2 & M_{12} &= \sum_1^n x_i \\ M_{13} &= \sum_1^n y_i x_i & M_{21} &= \sum_1^n x_i \\ M_{22} &= \sum_1^n 1 & M_{23} &= \sum_1^n y_i \end{aligned} \quad (6)$$

We can now use the regular linear algebraic method to find the solution of 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 discussed how to solve the 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 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, since they are fast change in data value, we don't use the Least Square Numeric Method,

instead, we use a very efficient numeric analytic methods, called Sinc numeric Methods [11 & 12].

4. Case Study

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

Data

Data: a port of Yangtze River annual import & export productivity of Material transportation (in 10^4 tons/year). The given data in Table 1 are real world 10 year's data of the port (1999 – 2008) in the unit of 10^4 tons /year.

Table 1. A 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

Task

Using the data analysis results to predict the future three year's annual productivity performance.

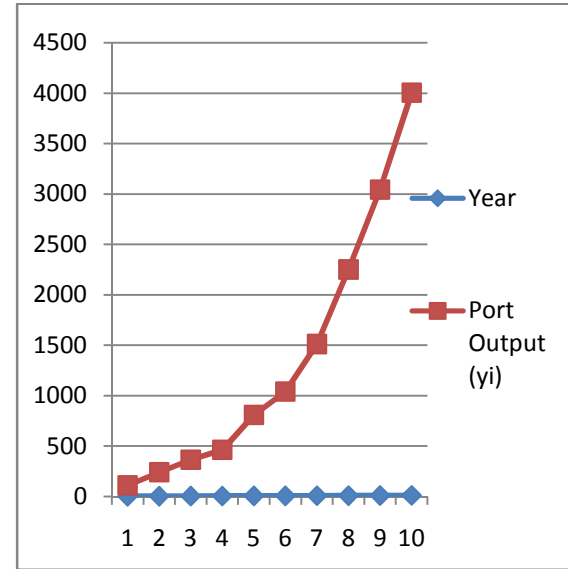
Data Analysis Procedure

Step 1

Since the given data is not too large, we selected the Microsoft Excel as our visualization tool.

Writing a computer program (completed by students at University of Guam, USA) to perform our numeric data analysis algorithms and 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 in the following).

Figure 1. Port Output (10^4 tons/year) during 1999 – 2008



Step 2:

It shows the function pattern of the data is a nonlinear exponential function. We try to use the following function pattern to describe the given data

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

Step 3

If we substitute the 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 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

$$G(x; b, c) = bx + c \quad (8)$$

Where

$$G(x; b, c) = \ln Y(x; a, b) \quad c = \ln a \quad (9)$$

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

Step 4:

Using a computer program to convert $g(x_i) = \ln(y_i)$ and list all the new converted data in the following Table 2.

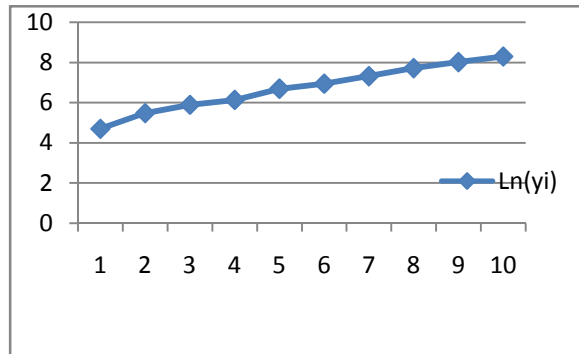
Table 2, Converted data from original data y_i

Year	x_i	$\text{Ln}(y_i)$
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

Step 5:

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

Figure 2. Converted Data , $\text{Ln}(y_i)$, in 1999 – 2008



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.

$$\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} M_{13} \\ M_{23} \end{pmatrix} \quad (10)$$

where

$$M_{11} = \sum_{i=1}^{10} x_i^2 = 385$$

$$M_{12} = M_{21} = \sum_{i=1}^{10} x_i = 55$$

$$M_{22} = \sum_{i=1}^{10} 1 = 10$$

$$M_{13} = \sum_{i=1}^{10} g_i x_i = 401.1704$$

$$M_{23} = \sum_{i=1}^{10} g_i = 67.2066$$

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

$$b = 0.3722$$

$$c = 4.6184$$

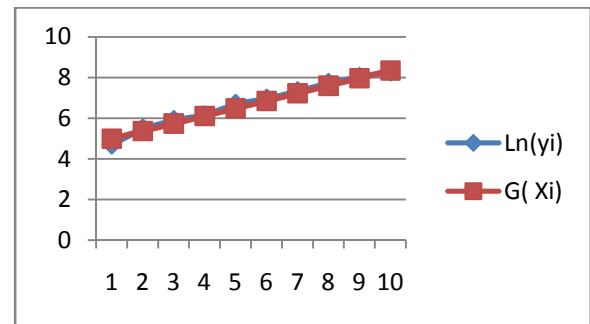
Therefore, we have solved new linear Equation (8):

$$G(x; b, c) = 0.3722 x + 4.6184 \quad (11)$$

Step 7:

Using the selected data visualization tool (Microsoft Excel in this case) and the data analysis resulted function pattern, Equation (11), to draw a predictive linear line, and compare with the linear line based on the data in Table 2 to see the two lines match closely or not (Fig. 3). Since the Least Square Methods are the optimal and the best fitting the data given, we can see from Fig. 3 that the data line and the theoretical line are matched very well.

Figure 3. Comparison of Converted Original Data $\text{Ln}(y_i)$ and Predictive Data $G(x_i)$



Step 8:

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

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

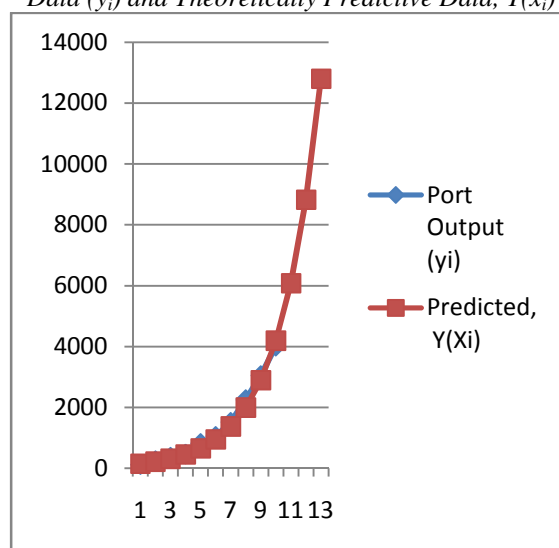
So the original non-linear exponential function pattern is

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

Step 9:

Using the predictive *Equation (12)* and the Microsoft Excel to draw the exponential curve and compare with the curve generated by original data given in Table 1 (Please see Fig. 4). From Fig. 4, we can see the that theoretical function pattern (Red line) very closely matches the data curve in (blue line). In the Fig. 4, we also employ the Formula (12) above to make the prediction of the Port future performance in 2009, 2010, and 2011. Is this prediction accurate enough or not. We can use the actual data or performance of the Yangtze River Port to verify.

Figure 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). In next Step, we will start to prepare the error analysis

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

Step 10:

Error analysis of the Least Square numeric algorithm [7].

Table 4 and Figure 5 in the following show our statistical results and the errors between original given data and the Least Square Method generate theoretically predicted data. We can see from the Figure 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 but error distributed along each data point is huge difference. In order to make the Least Squared total error become smaller, we can use a weighted factor to the Least Squared error value at each data point [13].

Figure 5. Comparison of Difference and Least Square Errors between Original Data and Predicted Data at Each Data Point

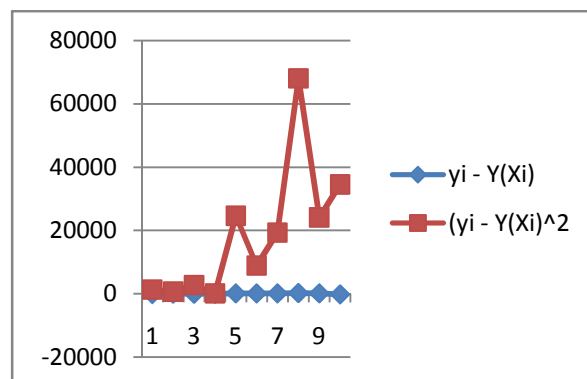


Table 4.. Statistical Analysis of Errors between Original Data y_i and Theoretically Predicted Data $Y(x_i)$

Error Type	Error Formula	Total Error Value
Total Difference of given and predicted data	$\sum_{i=1}^{10} (y_i - Y(x_i))$	667.3
Total Least Square Error	$Q(X) = \sum_{i=1}^{10} (y_i - Y(x_i))^2$	184658.4
Mean Value of Original Data	$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$	1383.2
Total		

Difference of given data and mean value	$\sum_1^{10}(y_i - \bar{y})$	0.0
Variance	$\delta_n^2 = \frac{1}{n} \sum_1^n (y_i - \bar{y})^2$	1565911.1
Standard Deviation	δ_n	1251.6

5. Conclusion

In this paper, we have summarized the physical and mathematical principles of Visualization and the Least Square Numerical data analysis methods. As an application and demonstration, we used a port industrial data as a case study, step by step, how to bind together to analyze massive data in industries. We used a mathematical transformation turned the nonlinear data pattern problem into a linear data analysis problem. We used both visualization tool and numeric tool based on graphics showing the data properties and then used Least Square numeric methods to find a theoretical formula to describe and predict the future performance of the selected industry. Finally, we used both graphics and numeric tools to analyze the errors in the future data performance and possible way to reduce the error between original industrial data and the data analysis-generated numeric predictive data formulas. The readers will be able to understand our methods through the case study in the paper, and to use this effective methods to analyze more real-world industrial large data with self-made computer programs or data analysis software.

References

- [1] C. Yadav, S. Wang, M. and Kumar, Algorithm and Approaches to Handle Large Data – A Survey, International Journal of Computer Science and Network, Vol. 2, Issue 3, 2013, pp. 2277 – 5420.
- [2] O. J. Reichman, M. B. Jones, and M. P. Schildhaver, (2011): Challenges and Opportunities of Open Data in Ecology, Science, 331 (6018), 2011, pp. 703 –705. DOI: 10.1126/Science 1197962.
- [3] U. Fayyad, G. G. Grinstein, A. Wierse edited, Information Visualization in Data Mining and Knowledge Discovery, Copyright 2002 by Academic Press, ISBN 1-555860- 689-0, pp 391.
- [4] C. R. Ehlschlaeger, A. M. Shortridge, and M. F. Goodchild, isualizing spatial data uncertainty using animation, Computers & Geosciences, Volume 23, Issue 4, 1997, pp. 387–395.
- [5] A. M. Legendre, (1805): Nouvelles Méthodes Pour La Détermination des Orbites des Comètes, Firmin Didot, Paris, 1805.
- [6] C. F. Gauss, Théorie Motus Corporum Coelestium in Sectionibus Coni cis Solem Ambientum, 1809.
- [7] W. Lichten, Data and Error Analysis, Allyn and Bacon, Inc. 1988, ISBN 0-205-11193-9, pp 171.
- [8] I. Petras, I. and D. Bednarova, D. , Total Least Square Approach to Modeling: A Mat Lab Toolbox, Acta Montanistica Slovaca, Vol. 15, Issue 2, 2010, pp.158 – 170.
- [9] A. Bjorck, Numerical Methods for Least Square Problems, SIAM, 1996, ISBN 978-0-89871-360-2.
- [10] J. Wolberg, Data Analysis Using the Method of Least Squares; Extracting the Most Information from Experiments, Springer, 2005, ISBN 3-540-25674-1.
- [11] L. P. Yaroslavsky, DFT and DCT based Discrete Reconstruction, Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, Vol. 1, 2003, pp.405 – 410.
- [12] F. Stenger, F, (2011): Handbook of Sinc Numeric Methods, CRC Press, 2010, ISBN 978-1-4398-2158-9, pp463.
- [13] T. Strutz,(2010): Data Fitting and Uncertainty – A Practical Introduction to Weighted Least Squares and Beyond, Vieweg and Teubner, 2010, ISBN 978-3-8348-1022-9.