MATEC Web of Conferences **52**, 02001 (2016)
DOI: 10.1051/matecconf/20165202001

Comparative Calibration of Corrosion Measurements Using K-Nearest Neighbour Based Techniques

Yaman Hamed[1,a], A'fza Shafie[1], Zahiraniza Bt Mustaffa[2] and Naila Rusma Binti Idris[3]

[1]*Fundamental and Applied Sciences Department, Universiti Teknologi PETRONAS, Tronoh, Malaysia*
[2]*Civil Engineering Department, Universiti Teknologi PETRONAS, Tronoh, Malaysia*
[3]*Upstream Technical Services (UTS), Technlogy & Engineering Division, Kuala Lumpur, Malaysia*

Abstract. Every measuring equipment or inspection tool is known to have its own accuracy, which may affect the reliability of its measurements. This includes oil and gas pipeline corrosion defects measurements. The inspection tolerance occurred in the measurements should be treated carefully for each equipment to prevent misinterpretation of the data which could lead to incorrect assessment. This paper presents a comparison between two K-Nearest Neighbour (KNN) interpolation techniques used to calibrate corrosion measurements collected by Magnetic Flux Leakage Intelligent Pig (MFL-IP) with the readings of Ultrasonic Testing (UT) scan device. The comparison has relied on the position of the interpolators, the weight sequence, and the error in the final enhanced metrics compared to the original measurements. Both techniques have the potential to calibrate and enhance IP measurements, with relative advantage for one technique in reducing over fitting problem. This enhancement will be used to improve the integrity assessment report that depends on the disturbed corrosion metrics of oil and gas pipelines, to decide whether the pipeline is fit for service or needs certain maintenance.

1 Introduction

The early detection of an operational oil pipeline defects can prevent failure which can lead to huge catastrophes. Corrosion is one of the major factors that can lead to a pipeline system failure; hence it is important to understand the actual condition of the operational pipeline via proper inspection and using the correct tools. Non-destructive corrosion detection techniques vary according to the purpose of the inspection, accessibility to the pipeline system, level of accuracy and the reliability of data required[1].

Smart/intelligent pigs are used to provide information about the conditions of a pipeline and can be used to locate the problem areas. Metal-loss inspection pigs are used to detect defects that have resulted in wall thinning in the pipeline. There are two main types of metal-loss pigs i.e. magnetic-flux leakage (MFL) and ultrasonic testing (UT). (MFL-IP)when compared to other scan techniques, is widely used as an in-line Inspection (ILI) in the corrosion detection field considering it's applicability in both offshore and onshore pipelines and its capability to discriminate defects to some extent, given its comparative cost and time consumption. UT scan technique is normally applied externally and is used to inspect localized section. UT is known to have better sizing accuracy since absolute

[a] Corresponding author : yaman_85@hotmail.com

measurement can be obtained when compared to MFL which provides relative measurements[2].Table 1 shows the sizing accuracy for different corrosion detection techniques [3].

Table 1. Sizing accuracy of MFL and UT tools

Inspection tool	Sizing accuracy	Confidence level
HR-MFL IP	± 10% T	80 %
XHR-MFL IP	± 5 % T	80 %
HR-UT Scan	± 6 % T	80 %
XHR-UT Scan	± 3 % T	80 %

HR: high resolution, XHR: extra high resolution, T: wall thickness.

As shown in Table 1, both MFL and UT technologies have certain limitations regarding the accurate measurement of the defects within the system since both devices suffer from a certain *error margin* depending on the device's sizing accuracy. Since UT measurements have a smaller sizing accuracy than IP measurements, a consideration of the UT device being more accurate than MFL-IP can be done. This measuring error affects the integrity assessment and may lead to an over or an under-estimation of the actual condition of the scanned pipeline. In this paper, by using two different techniques, K-nearest neighbor interpolation method (KNN) was used to enhance the inaccurate readings of the corrosion metrics collected by the MFL-IP device, with the metrics collected by a UT device, bearing the consideration that MFL-IP devices suffer from a wider error margin than the one affecting UT devices. A comparison between the two methods was applied to show which interpolator achieves better representation of the original measurements towards the goal ones.

2 Background on interpolation

Interpolation is the process of predicting a missing or an unknown value of a function or a sample point using the known points around it (neighbors)[4]. Different techniques can be applied as interpolators; Polynomial interpolation, Multivariate interpolation, Bilinear interpolation, Bi-cubic spline interpolation, K-Nearest-neighbor interpolation (KNN), Inverse distance weighting (IDW) Interpolation, quadratic interpolation, B-spline interpolation, Lagrange interpolation, Gaussian interpolation, among other techniques[5-7]. Interpolation techniques are well expressed in the application of image processing, data mining, artificial neural networks, as well as other variant applications. This paper, presents a new usage of interpolation techniques as calibration of the measurements of two non-destructive corrosion inspection tools. More precisely, K-Nearest Neighbor interpolation (KNN) was used to enhance the accuracy of corrosion readings collected by MFL-IP tool using the measurements of a UT scan device.

Nearest neighbor method is a statistical test that is used to determine the significance of a point's nearest neighbor in order to calculate the deviation from the general trend. Nearest neighbor algorithm selects the nearest point value and does not take into account the values of other neighboring points, hence producing a constant interpolation[5, 8]. Nearest neighbor technique is based on a comparison between the distribution of distances between a studied point and its nearest neighboring points in a set of randomly distributed data. The Distance function used in the comparison is a function that defines a Euclidean distance between each pair of elements of a set. The Euclidean distance between two points, $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ can be seen in (1):

$$d=\|x-y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

where*d* is the distance function, *n* is the sample size [9, 10].

The predicted point using the nearest neighbor method is simply the value of the nearest point among the n points of a sample to the unknown one, which is:

$$\hat{y}=y_i \;;\quad d_i=\{min \left\|\hat{y}-y_i\right\| \;; i=1,2,\dots n\} \tag{2}$$

The use of only the nearest point to predict the missing one, will exclude the effect of the other neighbor points which may lead to a biased estimator, herein to include the effect of multi neighbors to the predicted point and in order to reduce the effect of outliers of a sample point, K-nearest neighbor method is suitable for this purpose.

K-Nearest neighbor method: this method differs from the aforementioned nearest neighbor technique by considering the contiguity of the surrounding points to the required point. The contiguity can be estimated using what is called the weight function, which is defined as a function that measures the effect of each one of the neighbor points on the required one. In other words the estimated value of the missing or required point is the weighted average of its neighbors[11]. Loftsgaarden and Quesenberry introduced the KNN weight function in the relative field of density estimation[12], followed by Cover and Heart for the purpose of classification. The simplest weight function can be described as the ratio of the distance between each point of the neighborhood to the total sum of distances[11]:

$$w_i = \frac{d_i}{\sum d_i} \quad ; \quad d_i = \left\| \hat{y} - y_i \right\| \tag{3}$$

Franke & Nielson [13], used the classical form of weight function as seen in (4):

$$w_i = \frac{h_i^{-p}}{\sum h_i^{-p}} \tag{4}$$

where p is a positive real number chosen randomly and is called the power parameter (usually equals to 2), and h is the distance from point i to the interpolated point from the original data set.

Shepard [5], had used a modified weight function for superior results compared to the original one, as shown in(5):

$$w_i = \frac{\left[\frac{R-h_i}{Rh_i}\right]^2}{\sum \left[\frac{R-h_i}{Rh_i}\right]^2} \tag{5}$$

where h_i is the distance from the interpolated point to point i of a set of random data. R can be defined as the distance from the interpolated point to the farthest point of the set of data, and n is the number of the neighbors.

Hinton & Roweis [14], introduced a Stochastic Neighborhood embedding algorithm, an algorithm that uses a weight for the nearest neighboring points as seen in(6):

$$w_i = \frac{e^{-d_i}}{\sum e^{-d_i}} \tag{6}$$

where d_i is the distance between the interpolated point and its i'th neighbor.

Jianping Gou et al. [15], used a combination of the ratio between the interpolated point and its furthest neighbor, to the distance of its nearest neighbor as in (7):

$$w_i = \frac{d_i - d_{max}}{d_i - d_{min}} \times \frac{d_i + d_{max}}{d_i + d_{min}} \tag{7}$$

where d_i is the distance between the interpolated point and the point i, d_{max} is the distance between the interpolated point and furthest point in the set, d_{min} is the distance between the interpolated point and the nearest point in the set.

The interpolated point using the K-Nearest Neighbor algorithm can be defined by (8):

$$\hat{y}_i = \sum w_i y_i \tag{8}$$

where w_i is the proper weight for each one of the neighbor points y_i to the interpolated point $\hat{y}_i$ [11].

The methods described in this paper were built on Hinton's & Roweis's weight whilst using KNN. The exponential function defined in (6) will maximize the effect of the nearest neighbor to the point

that requires calibration, and by that, the interpolator will get the properties of the closest neighbor point, compared to the other surrounding points.

3 Proposed method

A 20 year operational oil pipeline used in Malaysia with a diameter of 25.4 cm, length of 3.9 kilometers, and internal wall thickness of 12.7 mm, was chosen to collect its internal corrosion measurements. Data on the corrosion geometry parameters (depth, length, width) were collected using two In-Line Inspection (ILI) tools; MFL-IP with a sizing accuracy of ±20% of the corrosion inspected measurements, and UT scan device with a sizing accuracy of ±5%.

Corrosion measurements collected from two different devices have the same properties however differ in formation. The measurements collected by a UT scan device is represented in a grid of the remaining wall thickness along the pipeline length, (the grid is set at constant dimensions horizontally and vertically) as shown in Table 2.

Table 2. Tabulation of UT corrosion measurements

			10:43	10:39	10:35	10:31	10:27
	Log Distance	X (mm)	0	5	10	15	20
1	108.705	-290	11.36	11.36	11.45	11.26	11.24
2	108.710	-285	11.36	11.36	11.31	11.24	11.22
3	108.715	-280	11.31	11.31	11.24	11.17	11.17
4	108.720	-275	11.19	11.19	11.24	11.29	10.94
5	108.725	-270	11.26	11.26	11.24	11.45	11.08
6	108.730	-265	11.4	11.4	11.24	11.36	11.19
7	108.735	-260	11.26	11.26	11.29	11.33	11.17

Table 3, shows the measurements collected by MFL-IP, which are represented as single points with relative metal loss, position and orientation.

Table 3. Tabulation of MFL-IP corrosion measurements

s-log distance [m]	s-o'clock position [h:m]	type	dimension	depth [%]	depth [mm] 11.1	length [mm]	width [mm]	ERF	at int. pipewall	Remaining wall thickness
108.527	0.320	MELO	PITT	11.000	1.221	15.000	15.000	0.819	YES	11.48
108.577	0.242	MELO	PITT	28.000	3.108	15.000	15.000	0.822	YES	9.59
108.582	0.195	MELO	PITT	13.000	1.443	20.000	20.000	0.821	YES	11.26
108.607	0.210	MELO	PINH	14.000	1.554	10.000	10.000	0.818	YES	11.15
108.637	0.359	MELO	PINH	10.000	1.110	10.000	10.000	0.818	YES	11.59
108.662	0.320	MELO	PITT	10.000	1.110	15.000	15.000	0.819	YES	11.59

To start the calibration process, each point was joined with its adjacent to form a pair, this is done by applying a mapping procedure to both of the corrosion measurement sets. The mapping procedure started from the same pipeline log-point for each device,which was recorded in the measurements report for each defect. The distance between the defect and the device's Log-point is illustrated in Fig.1.

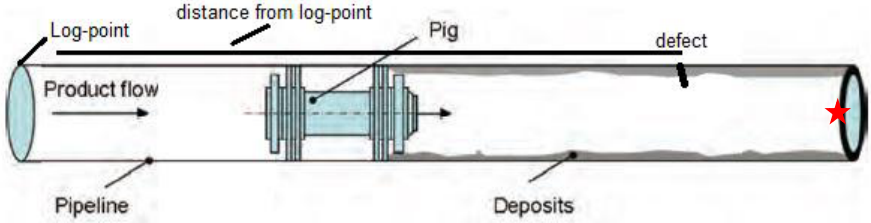

Figure 1. Distance from log-point to defect

This is followed by detecting the exact position of each pair by using a clockwise orientation method within the pipeline as shown in Fig. 2.

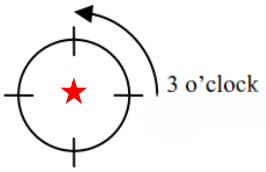

Figure 2. Clockwise orientation

Since corrosion measurements that are collected byMFL-IP are given as relative measurements compared to the UT device which are represented as a grid of actual remaining pipeline wall thickness, and in the light of the large sizing accuracy that UT device has compared to MFL-IP, the measurements of the UT device were assumed to be the goal measurements used to calibrate the measurements of the MFL-IP. In this paper bothproposed methods took in consideration a grid of the closest 8 measurements to the UT point to represent the neighbors of the *actual*/*goal* corrosion measurement. Fig.3 shows the distribution of points with respective neighbors where point ut_5 (the 5th element) is equivalent to is the goal for the point that requires calibration collected by MFL-IP.

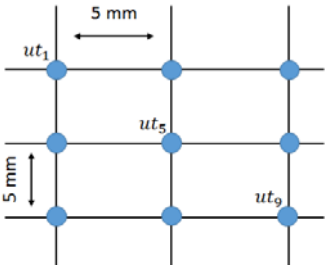

Figure 3. UT neighbor's grid

The first KNN method used in this paper to calibrate corrosion metrics of MFL-IP considered the center of UT measurements neighborhood as the position of the point that required the calibration. Using the Exponential weight function described by Hinton & Roweis, with the Euclidean function to describe the distance between the neighbors, (EwEf) method was applied as follows:

A. The neighbors grid were expanded by squeezing MFL-IP measurements in each grid as the 6'th element in the matrix, the point is placed such that it may seem to replace the original, however in actuality it is at a distance very near to the 5th point hence the vector now consists of 10 elements instead of 9, and the neighbors became 9 instead of 8 as shown in Fig. 4 and Fig. 5,hence will be used as $k$the number of neighbors in KNN interpolation. If $\{ y_i ; i: 1 \dots 10 \}$ is the variable representing the expanded remaining wall thickness matrix, then y_5 and y_6 represent the UT goal point and the IP point that requires calibration respectively.

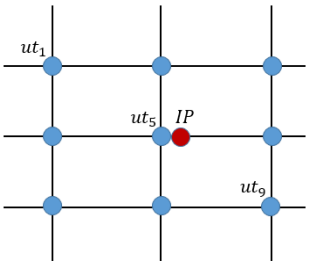

Figure 4. Expanded grid with IP measurements

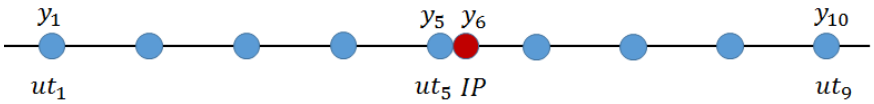

Figure 5. The vector representing remaining wall thickness expanded matrix

B. The distance between the center point (ut_5) and its neighbors were calculated using the Euclidean distance function as in (1).
C. A weight sequence was calculated for the expanded vector by using Hinton & Roweis weight function (6). The exponential function included within (6) will concentrate on the center point as it has the biggest effect on the predicted point. Adding to that, since the center point is the goal point, the trend well be stronger towards the actual required measurements.
D. By applying KNN interpolation technique (with k=9) on the 5th element in the vector, equation (8), will replace the IP value with the weighted average calculated by KNN technique which will result in an output that is the new calibrated corrosion measurements closer to the actual measurements collected by the UT scan device.
E. A comparison between the original UT measurements, original MFL-IP measurements, and the new interpolated corrosion measurements was conducted to determine the amount of enhancement that was done with the suggested method.

As mentioned earlier, EwEf calibration method is a method that applies KNN interpolation technique with an assumption that the position of the center point in the neighborhood (ut_5), is the position of the point that requires calibration, since (ut_5) is the goal point. However this assumption may lead to a biased estimator of the calibrated point towards the center of the grid most of the time, which could result in data over fitting. In light of this concern, the second method used in this paper to calibrate MFL-IP measurements used different position for the point that requires calibration in the UT neighborhood grid. The method was applied by assuming that the position of the point that requires calibration (IP) should consider the relationship between the original IP and UT measurements to avoid the over fitting possibility. To achieve that, a Maximum-Minimum based calibration method with Exponential weight and Euclidean distance functions (MMEwEf), was applied as follows:

A. if the MFL-IP measurement was larger than UT measurement, then the position of the point that require calibration will be assumed to be the position of the Maximum point in the neighborhood. On the contrary, if MFL-IP measurement was smaller than UT measurement, then the position of the point that require calibration will be assumed to be the position of the Minimum point in the neighborhood. The movement of the calibrated point's position will assure that the enhanced point after calibration will be relatively closer to the original MFL-IP measurement rather than the goal measurement (ut_5), which will reduce the possibility of a biased estimator. Fig 6, and Fig 7, show the distribution of the point that require calibration with UT measurements within the neighborhood grid in both cases respectively.

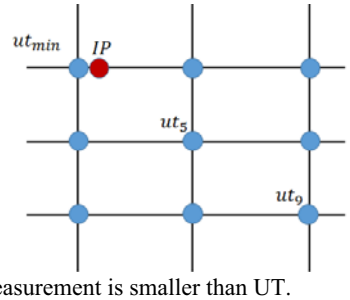

Figure 6. Neighbors' grid when IP measurement is smaller than UT.

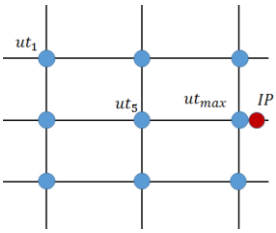

Figure 7. Neighbors' grid when IP measurement is larger than UT.

B. The expanded grid was used similarly (as with EwEf method) to generate a weight sequence with Hinton & Roweis weight function described in (6), and Euclidean distance function as in (1).
C. The weight sequence then was used to calibrate (IP) measurements by (8), the data set resulted by applying (8) is considered the calibrated corrosion measurements.

A comparison between the original UT measurements, original MFL-IP measurements, and the new interpolated corrosion measurements was conducted to determine the amount of enhancement that was done with the suggested method.

4 Results and discussion

The suggested methodology was applied on one segment of the studied pipeline, which contains 31 defects as reported in the reliability assessment report given by the MFL-IP operator. The calibration techniques used in this paper showed a remarkable enhancement in corrosion measurements. Fig. 8, and Fig.9 show comparisons between the original measurements, and the enhanced measurements using both proposed methods, EwEf, and MMEwEf calibration respectively. The enhanced IP measurements showed closer behavior to the goal UT points.

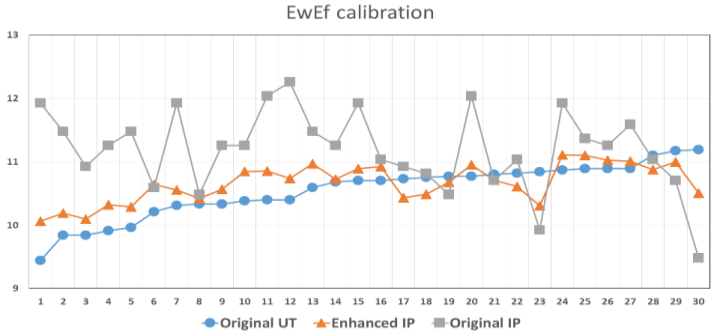

Figure 8. Corrosion measurements calibration by EwEf method

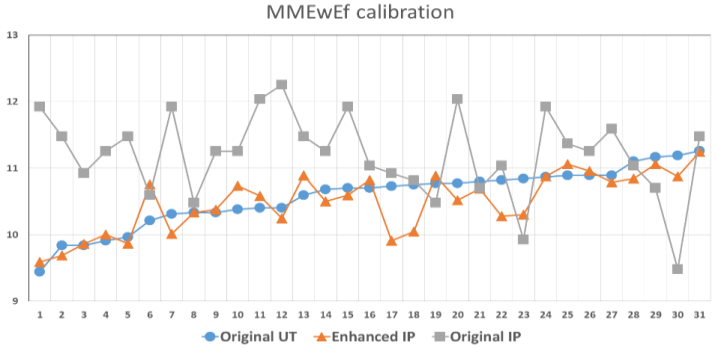

Figure 9. Corrosion measurements calibration by MMEwEf method

Table 4 shows the minimum remaining wall thickness in the studied segment of the pipeline collected by the UT device (actual size), compared to the thickness recorded by theMFL-IP device at the same point, and the thickness after the enhancement with both suggested methods.

Table 4. Comparison between minimum remaining wall thickness as recorded by different devices

	Minimum wall thickness		
Method	MFL-IP	Actual UT measurement	Calibrated measurement
EwEf	11.923 mm	9.44 mm	10.06 mm
MMEwEf	11.923 mm	9.44 mm	9.58 mm

The part of the pipeline that has the thinnest wall thickness is considered as the weakest point in the whole system. The point that has a minimum wall thickness of an operational pipeline is usually used in the extreme value analysis, which is applied to determine the probability of system failure. Table 4 shows the minimum wall thickness recorded using the Ultrasonic scan device as 9.44 mm. while the measurement reported by MFL-IP for the same point was 11.923. The calibrated measurement for the same point using the proposed EwEf, and MMEwEf calibration methods was found to be 10.06 mm, and 9.58 which is 14.66%, and 18.44% better than the original metrics respectively.

This difference between the calibrated measurement of the point that has the minimum wall thickness in the segment, and the original IP corrosion size using EwEf method is found to be 1.86 mm, which means that the service life of the studied segment is 4 years shorter than it was reported by the MFL-IP, considering that the average corrosion growth rate per year regarding to NACE (National Association of Corrosion Engineers) is reported to be 0.4 mm [16]. In other words, using the reported MFL-IP measurement without calibration will give a statement that the segment is fit-to-serve for an extra time approximated to be 4 years, while the actual size give much shorter service time. This huge difference may be lead to a possible system failure considering the lack of information related to the sizing accuracy of the MFL-IP device.

Likewise, the difference between the calibrated, and the original IP measurement using MMEwEf method was found to be 2.34 mm, which means an even shorter service life of the studied segment, which could be estimated to be 6 years.

Table 5 shows the error in the calibrated measurements compared to the original MFL-IP points using the proposed techniques.

Table 5. Error in calibrated measurements.

	Calibration method	
Source of error	EwEf	MMEwEf
Calibrated IP to original IP	0.98 mm	0.84 mm

Since the error in the calibrated measurements using MMEwEf method is smaller than the error committed using EwEf method,this means that MMEwEf calibration showed less biased estimation towards the goal measurements compared to the EwEf method. The 0.84 mm is equivalent to 6.61% of the pipeline wall thickness, while 0.98 mm is equivalent to 7.71%. The difference in the error between the two techniques refers to the superiority of the MMEwEf method on EwEf method when it comes to the biased estimation. This reduction of the error in calibrated measurements is due to the position movement of the point that requires calibration in the neighbors grid. The interpolator got the properties of the closest neighbor point to its own properties which reduce the error caused by KNN interpolation. This error reduction could be described as reducing the over fitting possibility in the calibrated measurements.

5 Conclusions

KNN interpolation technique can be used as proposed in this paper to calibrate corrosion measurements collected by different devices. The methods proposed in this paper showed remarkable results in reducing the error in corrosion measurements collected by MFL-IP compared to the actual corrosion metrics collected by UT scan device. Both EwEf, and MMEwEf calibration methods presented in this paper showed that using KNN will enhance the original IP corrosion measurements by 14.66%, and 18.44%, with an overall error of 6.61% and 7.71% of the actual pipeline wall thickness respectively. While EwEf calibration estimate closer measurements to the goal data, MMEwEf method reduces the bias in the estimators by considering the relationship between the original measurements. Further sample points testing should provide wider understanding to the differences between the proposed methods, since both had showed close results in the field of measurement calibration.

References

1. A. Barbian and M. Beller,*18th World Conference on NondestructiveTesting*, Durban, South Africa, (2012)
2. A. Pople and P. Wharf, *Pipeline Rehabilitation and Maintenance*, (2003)
3. P. Ivanov, Z. Zhang, C. Yeoh, L. Udpa, Y. Sun, S. Udpa, et al,*IEEE Transactions on Magnetics*,**34**(1998)
4. P. Miklos, *2nd Siberian-Hungarian Joint Symposium On Intelligent Systems*, (2004)
5. P. D. Dumitru, M. Plopeanu, and D. Badea, *1st European Conference of Geodesy & Geomatics Engineering GENG'13*, Antalya, Turkey,(2013)
6. R. Olivier and C. Hanqiang, "Nearest neighbor value interpolation," arXiv preprint arXiv:1211.1768, (2012)
7. T. M. Lehmann, C. Gönner, and K. Spitzer, IEEE Trans. Med. Imag, **18**,(1999)
8. R. Franke, Math. Comp, **38**, (1982)
9. K. Q. Weinberger, J. Blitzer, and L. K. Saul, *Advances in neural information processing systems*, (2005)
10. J. Walters-Williams and Y. Li, *Advanced Techniques in Computing Sciences and Software Engineering*, (2010)
11. W. Härdle and O. Linton, *Handbook of econometrics*, **4**, (1994)
12. D. O. Loftsgaarden and C. P. Quesenberry, Ann. Math. Stat, **36**,(1965)
13. R. Franke and G. Nielson, INT. J. NUMER. METH. ENG, **15**, (1980)
14. G. E. Hinton and S. T. Roweis, *Advances in neural information processing systems*, (2002)
15. J. Gou, L. Du, Y. Zhang, and T. Xiong, J. Inf. Comput. Sci,**9**, (2012)
16. F. Caleyo, A. Valor, V. Venegas, J. H. Espina Hernandez, J. C. Velazquez, and J. M. HALLEN, OIL GAS J, **110**, (2012)