VOL. 11, NO. 11, JUNE 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

RESTORATION OF HAZY DATA BASED ON SPECTRAL AND STATISTICAL METHODS

Nurul Iman Saiful Bahari[1], Asmala Ahmad[1], Burhanuddin Mohd Aboobaider[1], Muhammad Fahmi Razali[1], Hamzah Sakidin[2] and Mohd Saari Mohamad Isa[3]
[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, Malaysia
[2]Faculty of Science and Information Technology, Universiti Teknologi Petronas, Seri Iskandar, Perak, Malaysia
[3]Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, Malaysia
E-Mail: asmala@utem.edu.my

ABSTRACT

Remote sensing data recorded from passive satellite system tend to be degraded by attenuation of solar radiation due to haze. Haze is capable of modifying the spectral and statistical properties of remote sensing data and consequently causes problem in data analysis and interpretation. Haze needs to be removed or reduced in order to restore the quality of the data. In this study, initially, haze radiances due to radiation attenuation are removed by making use of pseudo invariant features (PIFs) selected among reflective objects within the study area. Spatial filters are subsequently used to remove the remaining noise causes by haze variability. The performance of hazy data restoration technique was evaluated by means of Support Vector Machine (SVM) classification accuracy. It is revealed that, the technique is able to improve the classification accuracy to the acceptable levels for data with moderate visibilities. Nevertheless, the technique is unable to do so for data with very low visibilities.

Keywords: haze, landsat, support vector machine, spectral, statistical.

INTRODUCTION

Haze is a partially opaque condition of the atmosphere caused by tiny suspended solid or liquid particles in the atmosphere, [13], [14], [20]. Malaysia experiences haze occurrences almost every year and are mainly caused by smoke originated from open fire in Indonesia due to plantation clean-up activities for the upcoming planting season[10].Haze degrades the quality of data recorded from remote sensing satellite by modifying the spectral and statistical properties of the data [4], [5]. A number of hazy data restoration studies can be found in literatures. The main objective of hazy data restoration is to remove the path radiance and retrieve the surface reflectance. Basically, there are two main approaches of hazy data restoration methods. The first one is an absolute correction method which is based on radiance transformation model. The second one is relative methods which mostly based on the image itself or several images from different acquisition date. Image based hazy data restoration method is an atmospheric correction method which is done by normalising an image to a reference image in order to remove atmospheric variation due to haze. This method is preferable when there is no in situ or ground truth data. The methods for image based include mean reflectance matching, dark pixel subtraction, histogram adjustment, band rationing and multi-date normalization using a regression approach.

Liang *et al.* [17] introduced mean reflectance matching methods to remove haze. This is done by subtracting mean reflectance from the clear region of Landsat TM bands 1, 2 and 3 (visible bands). In doing this, they assumed that bands 4, 5 and 7 (infrared bands) are not affected by the haze. The method seems most suitable for data with low concentration of haze. This contradicts the fact that that under thick haze, infrared bands are also affected by haze. The study claims that the method is able to remove haze visually but no in-depth analysis on the statistical properties of the restored data was carried out. Zhang and Guindon [21] developed haze optimized transform (HOT) to detect haze and further removed the haze layer by incorporating dark object subtraction method [16]. Moro and Halounova [8] further improved the method by adding HOT masking not only for dense vegetation but also for water and man-made features. The proposed method was applied on high resolution satellite data (IKONOS) and evaluated by means of vegetation index (VI) of both hazy and dehazed data. Hu *et al.* [11] developed haze detection, perfection and removal module coded in IDL language. Users can pick any method contained in each step or develop and use their own methods. Among the methods used for haze detection is HOT and DOS for removal of haze0. The methods are tested on a number of Landsat TM and QuickBirdsatellite data in which successfully reduced the effects of haze. Hu and Tang [7] carried out relative radiometric normalization (RRN) for atmospheric correction of remote sensing data. They normalised a hazy image based on a reference image that was free from haze. RRN used assumption that the relationship between the radiances recorded at two different times are homogenous and can be approximated by linear functions. For this purpose, pseudo invariance features (PIF) consisting of manmade objects, such as road lane, rooftop and parking lots, were determined from the scene within the data. The normalisation made use of a regression equation in which is developed by establishing relationship between the values of the PIF from both data. The relationship can be described as DN=a+b*DN. Where b is the multiplicative component which can correct for the difference in sun angle, atmosphere and else, the intercept a is an additive component, able to correct the difference in atmospheric

path radiance between the data. The important step in establishing the regression relationship is to obtain the a and b parameters. The result showed that the method was able to remove most haze from the data. This method will be applied in this study by choosing suitable PIF and generating a linear relationship between both hazy and non-hazy data and then normalising the data to remove haze effects.

DATA AND METHODOLOGY

In this study, a set of simulated hazy data was used in order to develop hazy data restoration technique [1]. The simulated hazy data sets were generated based on real Landsat-5 TM satellite data dated 11 February 1999. The data were obtained from the Malaysian Remote Sensing Agency (MRSA) covering the area of Klang, Selangor located in Peninsular Malaysia. For the purpose of this study, Landsat-5 TM data were spectrally subsetted by selecting only six bands, i.e. band 1, 2,3,4,5 and 7. Band 6 is a thermal band and was excluded because this study focuses on visible and near infrared bands only. The vector-based structure of hazy data $L_i(V)$ can be written as equation [1]:

$$L_i(V) = \left(1 \quad {}^{(1)}_{i}(V)\right) T_i + L_O + {}^{(2)}_{i}(V) H_i \qquad (1)$$

where $L_i(V)$, T_i, H_i, L_o, $\beta_i^{(1)}$ (V) and $\beta_i^{(2)}(V)$are the hazy data, the signal component, the pure haze component, the radiance scattered by the atmosphere, the signal attenuation factor and the haze weighting in satellite band i, respectively. Simulated datasets with different visibilities are used 0, 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 km [6].0 km visibility represents the atmosphere condition that is severely affected by haze. 20 km visibility represents clear atmosphere or in other words haze-free condition [1]. In this study the 20km visibility data was used as the reference data for accuracy assessment purpose.

Figure-1 shows the flowchart for hazy data restoration technique and its evaluation. Initially, suitable pseudo invariant features (PIF) were identified among bright objects. The pixel values from the selected PIF for both hazy and non-hazy data for each band were analysed using linear regression analysis. By doing so, we can then estimate the haze mean for each band by making use of the gain and offset values from the linear regression analysis. Haze mean subtraction was next carried out by using two data for image-to-image radiometric normalisation in order to remove the haze effects. The first data should be a clear or haze-free data while the second data is the data to be corrected for haze, i.e. the hazy data. Initially, band 1 from the 20 km visibility dataset was used as the first data and band 1 from the 18 km-visibility dataset was used as the second data. The relationship between the PIF values from both data was established using linear regression analysis. This produced a linear trend or best-fit line and its equation consisting the gain (slope) and offset value (y intersection) of the regression trend line. Eventually, the second data was then transformed using this equation. This was repeated for bands 2, 3, 4, 5 and 7 of the 18 km-visibility dataset. The overall procedure was repeated for the remaining hazy datasets, i.e. the 16, 14, 12, 10, 8, 6, 4, 2, and 0 km visibility datasets. Next, the haze mean was subtracted from the hazy data.

Spatial filtering was subsequently implemented to remove the remaining noise. Gaussian, average and median filtering with kernel size 3, 5, 7 and 9 was implemented [3]. The performance of the hazy data restoration technique was then evaluated by means of classification accuracy [2], [19]. In doing so, Support Vector Machine (SVM) classification was performed on the hazy datasets [15]. SVM is characterised by an efficient hyperplane searching technique that uses minimal training data and therefore consumes less processing time. The method is able to avoid over-fitting problem and requires no assumption on data type. Although non-parametric, the method is capable of developing efficient decision boundaries and therefore can minimise misclassification.

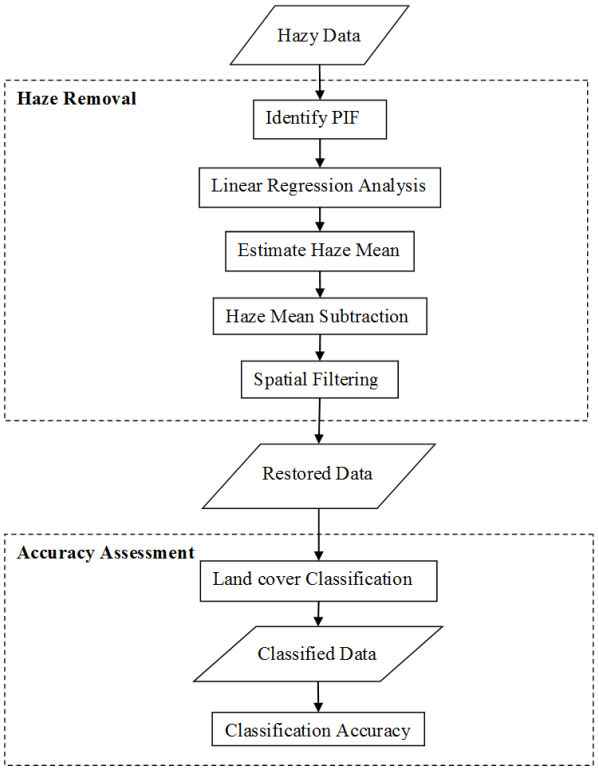
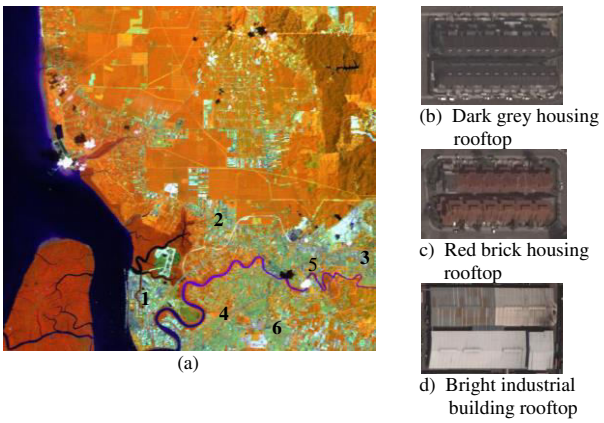


Figure-1. Flowchart for hazy data restoration technique and its evaluation.



(a)

(b) Dark grey housing rooftop

c) Red brick housing rooftop

d) Bright industrial building rooftop

Figure-2. 20 km visibility data, bands 4, 5 and 3 assigned to red, green and blue channels of Klang, Selangor, Malaysia. (b), (c) and (d) are an enlarged version of PIF location in (a) from Google Maps [9].

The SVM classification was next accuracy assessed by means of a confusion matrix, i.e. a square matrix with the number of rows and columns being equal to the number of classes. From this matrix two accuracy measures namely, producer accuracy and overall accuracy were computed. Producer accuracy is a measure of the accuracy of a particular classification scheme and shows the percentage of a particular ground class that has been correctly classified. The minimum acceptable accuracy for a class is 70% [12]. This is calculated by dividing each of the diagonal elements in the confusion matrix by the total of the column in which it occurs:

$$\text{Producer accuracy} = \frac{c_{aa}}{c_{\bullet a}} \quad (2)$$

where,

c_{aa} = element at position a^{th} row and a^{th} column

$c_{\bullet a}$ = column sum

A measure of behaviour of the ML classification can be determined by the overall accuracy, which is the total percentage of pixels correctly classified, i.e.:

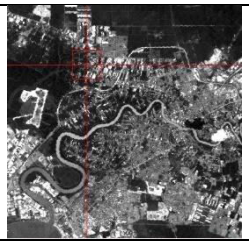
$$\text{Overall accuracy} = \frac{\sum_{a=1}^{U} c_{aa}}{Q} \quad (3)$$

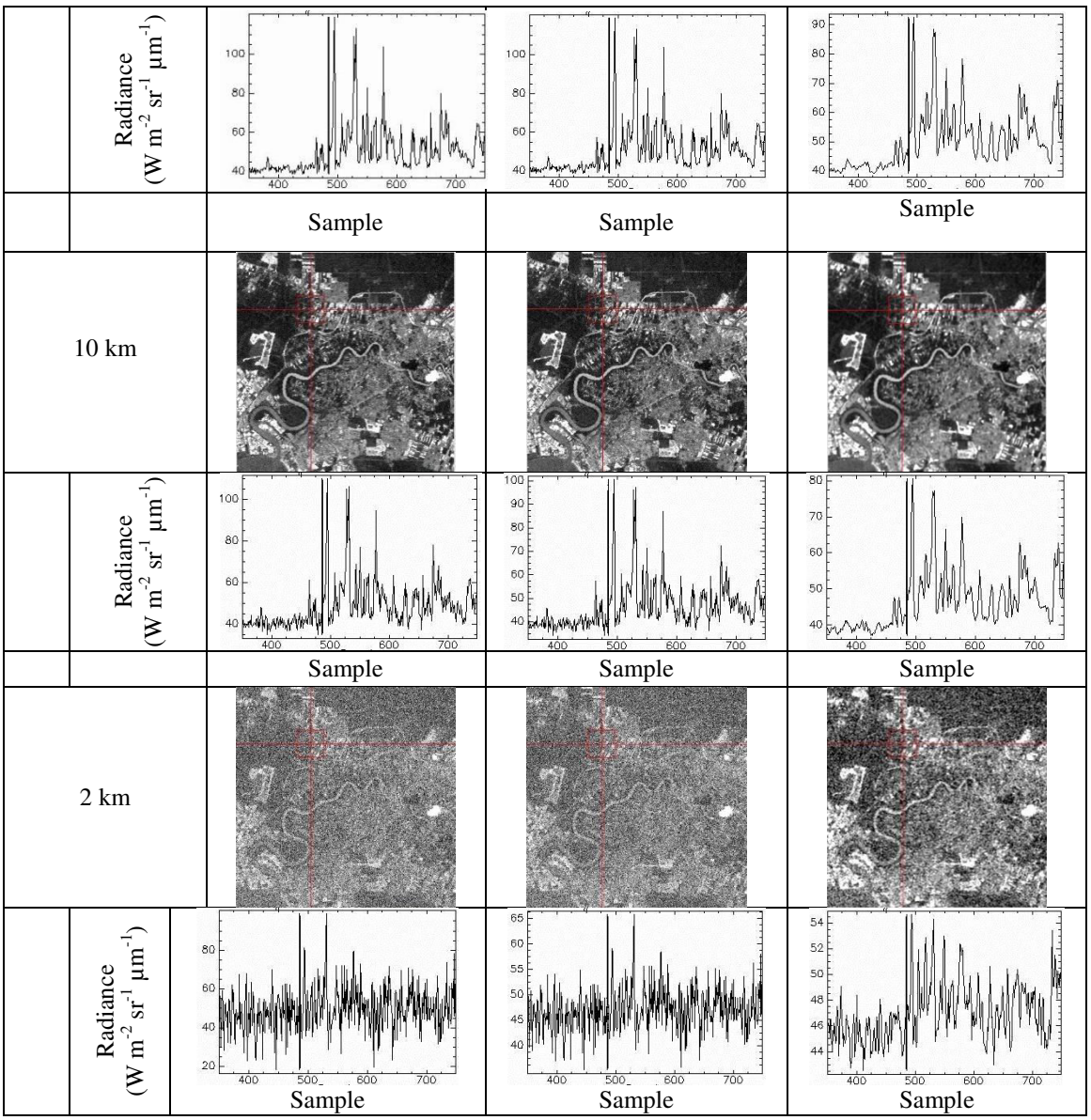
where Q and U represent the total number of pixels and classes respectively. The minimum acceptable overall accuracy is 80% [18].

RESULTS AND DISCUSSIONS

Table-1 shows band 1 of a hazy data (a) in original form (b) after haze mean subtraction and (c) after haze mean subtraction and filtering for 18, 10 and 2 km visibility. The radiance plots across x profile are shown at the bottom of the corresponding data where horizontal axis represents line while vertical axis represents radiance value. It can be seen that no apparent changes after haze mean subtraction can be seen from the horizontal profile and the radiance plot and value. As visibility decreases there is an apparent change in radiance value but not for radiance plot. This is because as the haze mean becomes larger as visibility decreases and subtraction of it cause a change in the radiance value. It can be therefore understood that mean subtraction only change the mean value for corresponding band but not the radiance signature. In addition, it can be seen that filtering has smoothen the data and this effect can be seen in the radiance plot. The smoothing has change the radiance mean value and radiance signature for corresponding band. Further analysis was performed by determining the overall classification accuracy of land classification applied on the hazy datasets before and after haze removal.

Table-1. Band 1 and radiance scatter plot of hazy data in (a) original form (b) after haze mean subtraction and (c) after haze mean subtraction and average filtering (3x3) for 18 km, 10 km and 2 km visibility.

Visibility	(a) Hazy data	(b) Hazy data after haze mean subtraction	(c) Hazy data after haze mean subtraction and average filtering (3x3)
18 km			



The accuracy of SVM classification for hazy data after restoration was assessed using confusion matrices where the corresponding overall classification accuracy (in percentage) and Kappa coefficient were computed [12]. Visually, it can be seen that average and median filtering has smoothen the pixel values (Figure-4). The same goes with median and average filtering. However Gaussian filtering still preserves data features as before removal. This is due to the way Gaussian filter works where the kernel has higher weight towards the centre, resulting in preservation of most data features.

Gaussian filtering is able to produce classifications with higher accuracy compared to average and median filtering (Figure-5). It is noticeable that the hazy data restoration is not needed at 18 to 10 km visibility since the removal seems even causing a decrease in classification accuracy compared to before haze removal, except for Gaussian filtering with 3x3 kernel size which maintains the accuracy of the data after hazy data restoration. Gaussian filtering is found to be the optimal filtering type compared to other filtering methods (Figure-6). The hazy data restoration method is unneeded for 18 to 12 km visibility. For 10 km visibility, the hazy data restoration has increased the classification accuracy from 84.02% to 84.21% which is within the acceptable accuracy range (> 80%). The improvement in classification accuracy for 8 km visibility is more obvious as the accuracy increases from 78.97% to 82.21%. There is also an increase in accuracy for the 6, 4 and 2 km visibility, but the accuracy is still less than the acceptable accuracy range.

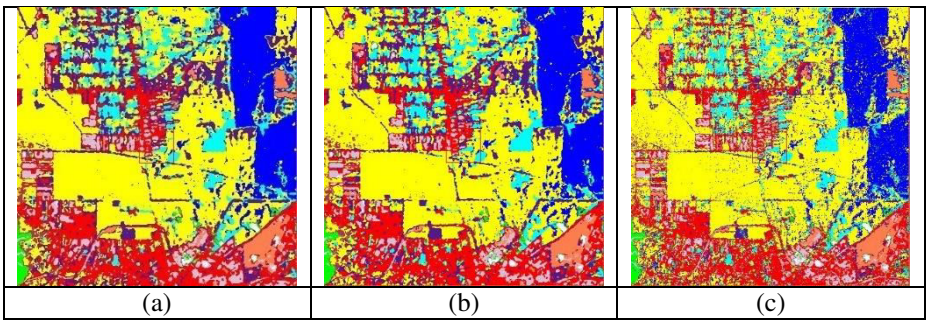


Figure-3. SVM classification for 12 km visibility data after haze mean subtraction and (a) average, (b) median and (c) Gaussian filtering.

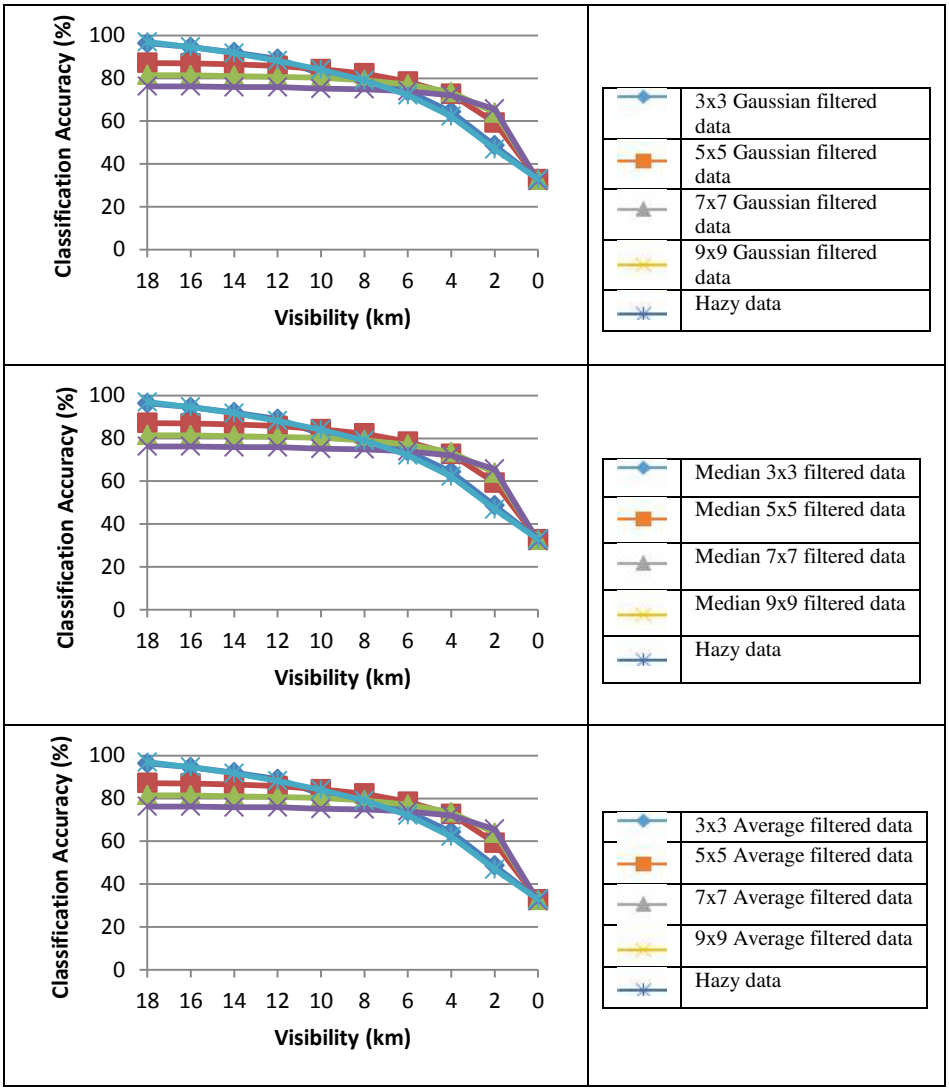


Figure-4. Classification accuracy versus visibility for Gaussian, median and average filtered data.

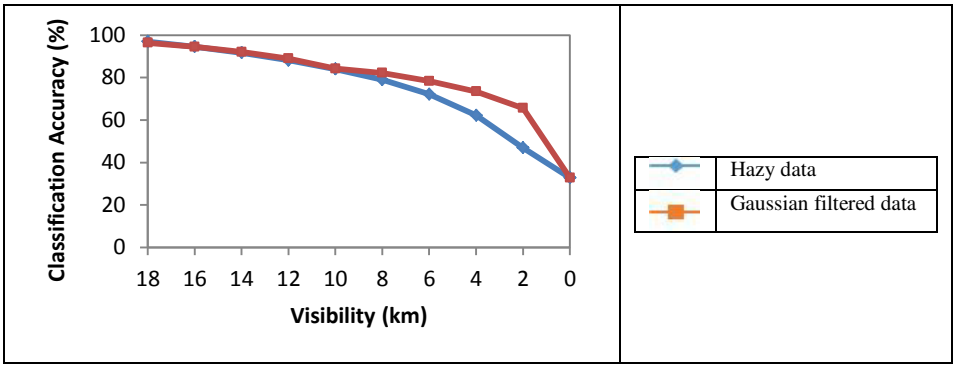


Figure-5. Classification accuracy versus visibility for hazy and Gaussian filtered data.

From the analysis, it can be deduced that at visibilities more than 12 km, the hazy data restoration is unnecessary as this will only further degrades the data. It is also revealed that hazy data restoration is applicable for visibility more than 6 km and less than 12 km since the restoration has improved the SVM classification accuracy. Nevertheless, hazy data restoration is not applicable anymore for visibility less than 6 km since the restoration unable to improve the classification accuracy.

Table-2. Visibility, Gaussian filter kernel size and the corresponding accuracies before and after restoration.

Visibility (km)	Gaussian kernel size	Accuracy of hazy data	Accuracy of after hazy data restoration
18	3x3	96.94	96.33
16	3x3	94.58	94.50
14	3x3	91.71	92.10
12	3x3	88.19	88.98
10	5x5	84.03	84.21
8	5x5	78.97	82.21
6	5x5	72.15	78.39
4	7x7	62.20	73.43
2	9x9	46.90	65.62
0	3x3	32.87	32.87

CONCLUSIONS

In this paper, we presented a restoration technique for hazy remote sensing data. The technique is able to improve hazy data with visibilities 6 to 12 km. Restoration is found not necessary for data more than 12 km visibility, while ineffective for data less than 6 km visibility. Nevertheless, the technique needs to undergo more testings particularly on data with different haze conditions and geographical locations in order to further validate its robustness.

ACKNOWLEDGEMENTS

We would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for funding this study under the Malaysian Ministry of Higher Education Grant (FRGS/2/2014/ICT02/FTMK/02/F00245).

REFERENCES

[1] A. Ahmad and S. Quegan. 2014. Haze modelling and simulation in remote sensing satellite data. Applied Mathematical Sciences.8 (159): 7909-7921.

[2] A. Ahmad and S. Quegan. 2012. Analysis of maximum likelihood classification technique on Landsat 5 TM satellite data of tropical land covers. Proceedings of 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE2012). pp. 1-6.

[3] A. Ahmad and S. Quegan. 2014. Haze reduction in remotely sensed data. Applied Mathematical Sciences. 8(36): 1755-1762.

[4] A. Ahmad and S. Quegan. 2014. The Effects of haze on the spectral and statistical properties of land cover classification. Applied Mathematical Sciences. 8(180): 9001-9013.

[5] A. Ahmad and S. Quegan. 2015. The Effects of haze on the accuracy of satellite land cover classification. Applied Mathematical Sciences. 9(49): 2433-2443.

[6] A. Asmala, M. Hashim, M. N. Hashim, M. N. Ayof and A. S. Budi. 2006. The use of remote sensing and GIS to estimate Air Quality Index (AQI) over Peninsular Malaysia. GIS development. p. 5.

[7] C.Hu and P. Tang. 2011. Converting DN value to reflectance directly by relative radiometric normalization. Proceedings - 4th International Congress on Image and Signal Processing, CISP 2011. 3: 1614-1618.

[8] G.D. Moro and L. Halounova. 2007. Haze removal for high-resolution satellite data: a case study. International Journal of Remote Sensing, 28(10): 2187-205.

[9] Google Maps. 2015. https://www.google.com.my/maps.

[10] H.A. Rahman. 2013. Haze phenomenon in Malaysia : domestic or transboudry factor, 3rd International Journal Conference on Chemical Engineering and its Applications (ICCEA'13), Sept. 28-29. pp. 597-599.

[11] J. Hu, W. Chen, X. Li and X. He. 2009. A haze removal module for mutlispectral satellite imagery. Urban Remote Sensing Event, 2009 Joint IEEE. pp. 1-4.

[12] J.R. Thomlinson, P.V. Bolstad and W.B. Cohen. 1999. Coordinating methodologies for scaling landcover classifications from site-specific to global: Steps toward validating global map products. Remote Sensing of Environment. 70(1): 16-28

[13] M. Hashim, K. D. Kanniah, A. Ahmad, A. W. Rasib. 2004. Remote sensing of tropospheric pollutants originating from 1997 forest fire in Southeast Asia. Asian Journal of Geoinformatics. 4: 57-68.

[14] M. F. Razali, A. Ahmad, O. Mohd and H. Sakidin. 2015. Quantifying haze from satellite using haze optimized transformation (HOT). Applied Mathematical Sciences. 9(29): 1407-1416.

[15] N.I.S. Bahari, A. Ahmad and B.M. Aboobaider. 2014. Application of support vector machine for classification of multispectral data. IOP Conference Series: Earth and Environmental Science. 20(1): 2038.

[16] P.S. Chavez. 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sensing of Environment. 24(3): 459-479.

[17] S. Liang, H. Fang, J.T. Morisette, M. Chen, C.J. Shuey, C.L. Walthall and C.S. Daughtry. 2002. Atmospheric correction of Landsat ETM+ land surface imagery. II. Validation and applications. IEEE Transactions on Geoscience and Remote Sensing. 40(12): 2736-2746.

[18] S.T. Knick, J.T. Rotenberry, T.J. Zarriello. 1997. Supervised classification of Landsat thematic mapper imagery in a semi-arid rangeland by nonparametric discriminant analysis. Photogrammetric Engineering and Remote Sensing. 63(1): 79-86.

[19] U. K. M. Hashim and A. Ahmad. 2014. The effects of training set size on the accuracy of maximum likelihood, neural network and support vector machine classification. Science International- Lahore. 26(4): 1477-1481.

[20] W. Morris. 1975. The Heritage Illustrated Dictionary of English Language of the English Language. American Heritage Publishing Co, New York, USA.

[21] Y. Zhang and B. Guindon. 2003. Quantitative assessment of a haze suppression methodology for satellite imagery: effect on land covers classification performance. IEEE Transactions on Geoscience and Remote Sensing. 41(5): 1082-1089.