ANALYSIS OF EMISSION COMPLIANCE VIOLATIONS USING SELF ORGANIZING MAP

[1]JENNIFER T. CARPIO, [2]SHANETH C. AMBAT

Arellano University, AMA University
E-mail: jhennieffer@yahoo.com, shaneth_ambat@yahoo.com

Abstract- This study aims to explore the utility of Self Organizing Map (SOM) algorithm as an exploration instrument in analyzing violations in emission test results in the Philippines. A case of real-life data mining was performed using a dataset consisting of show cause orders issued by the Land Transportation Office (LTO) to erring Private Emission Testing Centers (PETCs) conducting non-appearance test violations from calendar year 2010 to the second quarter of 2016. The possibilities offered by the visualizations generated by SOM was explored to determine how it can be used to produce information that can support policy enhancement of the concerned government agency with the goal of minimizing, if not to totally eradicating, the occurrence of non-appearance testing violations that can lead to detrimental and more adverse air pollution in the country.

Index Terms- emission test, nonappearance, self organizing map.

I. INTRODUCTION

This paper describes a proposed analysis methodology based on Self Organizing Map (SOM) algorithm [1], [2] for discovering new knowledge from semi-structured data. Self-Organizing Map (SOM) performs a non-linear mapping of a high dimensional data space to a low-dimensional one (usually 2-dimensional). This produces a map which plots data similarities by grouping similar items together thus facilitating the visualization and exploration of the data structure. SOM reduce dimensions and display similarities. It has been used by several academic research and study in various fields like Oceanography, Geophysics, Economics, and Music, other fields as well can use SOM in effectively analyzing big data.

In this paper, SOM is used for automating the detection of patterns that can describe the nature of violations committed by agencies tasked to issue Certificate of Emission Compliance (CEC) to motorized vehicles. CEC compliance is critical because it is linked to certain environmental issues, the most prominent of which is that of climate change [3]. In the 2030 agenda for Sustainable Development formulated by the United Nation, the commitment to "take urgent action to combat climate change and its impacts" (Goal 13) is very well defined [4]. The Philippines, as a UN member, had pledged to cut carbon emissions by 70% in 2030 in support of this goal [5]. Senate committee inquiries have shown that the transport sector is one of the top sectors with the highest Green House Gas (GHG) contribution.

In response to this situation, implementation of the provisions of Republic Act 8749, otherwise known as the Clean Air Act was revitalized [6]. This act requires owners of motor vehicles to have their vehicles tested and certified for emissions compliance before the vehicle registration can be renewed. The mandatory of an emission testing prior to registration is intended to curve the rise of GHG emission. The Department of Transportation (DOTr), in cooperation with the Land Transportation Office (LTO) has accredited and authorized private emissions testing centers (PETC) to conduct these tests and issue the prerequisite certificates.

Unfortunately, the corrupt practice of no show or non-appearance of motor vehicle for emission testing has become rampant and pervasive among the various PETCs. Left unresolved, it will perpetrate adverse effects on the environmental, social and economical well-being of the Philippine Society. More efficient policies and procedures are needed in the assigned Agency to address this problem.

The objective of this study is to explore the utility of Data Mining particularly that of the self-organizing map (SOM) algorithm, to cluster and extract hidden patterns from the sanctioned CEC results of PETCs involved in non-appearance testing from 2010 to early 2016. The initial goal is to identify valid, novel, useful, and understandable patterns and formulate knowledge discovery information [7] that can be used to support improvement of policies and procedures in terms of effectively eradicating non-appearance testing violations that leads to uncontrollable air pollution. A corollary goal is to determine whether the clusters formed by SOM can be used to predict future transgressions in non-appearance testing and how the performance of the SOM model for this purpose can be improved.

II. PROBLEM SETTING

A. Motivation

LTO Department order no. 2012-10 provided revised rules and regulations on the authorization and monitoring of motor vehicle private emission testing

centers (PETCs) and IT providers. This order enabled the agency to monitor the conduct of Carbon Emission Testing in the Philippines. Any violations committed are quickly dealt with through issuance of show-cause orders to give violators the chance to defend the alleged violations and give appropriate sanctions.

Thus far, the collection of issued show-cause orders already provides LTO with a considerable amount of data regarding the committed violations. Processing these data could possibly provide some perspective on the nature of violations committed and how they are being perpetrated such as:

(a) Type of violations committed, (b) Geographic area with the highest rate of incidence of the violation, (c) Percentage of sanctioned PETCs handled by IT Providers, and (d) Motor Vehicle (MV) details often committing non-appearance testing. Unfortunately,

appropriate tools that facilitate an in-depth exploration of this data have yet to be developed.

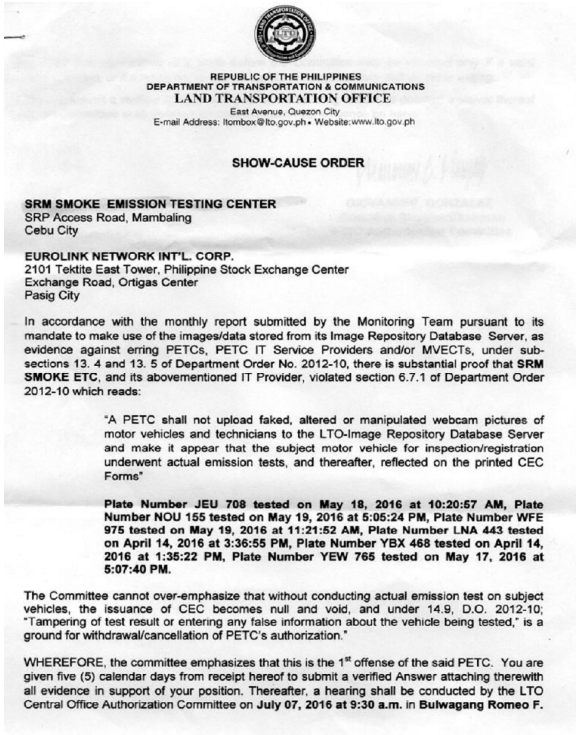
The Self Organizing Map Clustering Technique introduced by Teuvo Kohonen, seemed an interesting approach for processing this large amount of information. SOM has been defined as "a highly effective and sophisticated visualization tool for visualizing high dimensional, complex data with inherent relationships between the various features comprising the data". The SOM's output emphasizes the salient features of the data and subsequently leads to the automatic formation of clusters of similar data items.

The cluster models extracted from the dataset can then be used to identify causes of non-appearance testing in order to formulate recommendation plans. Given that the non-appearance violations could lead to the further deterioration of air quality in the country, these recommendation plans could go a long way towards repealing this serious threat to the environment.

B. Dataset

This study will use the nationwide data collected from sanctioned PETC that violated Department Order 2012-10 from 2010 to second quarter of 2016. A total of 2834 motor vehicles used as evidences in 734 show-cause orders describing violation incidents that occurred within that period are available for this purpose.

Each of these orders contains statements against the operator informing him/her of the Department Order 2012-10 violation and requiring him to file a verified explanation why the corresponding fine or penalty shall not be imposed. An example of such an order is displayed in Fig. 1.



REPUBLIC OF THE PHILIPPINES
DEPARTMENT OF TRANSPORTATION & COMMUNICATIONS
LAND TRANSPORTATION OFFICE
East Avenue, Quezon City
E-mail Address: ltombox@lto.gov.ph • Website:www.lto.gov.ph

SHOW-CAUSE ORDER

SRM SMOKE EMISSION TESTING CENTER
SRP Access Road, Mambaling
Cebu City

EUROLINK NETWORK INT'L. CORP.
2101 Tektite East Tower, Philippine Stock Exchange Center
Exchange Road, Ortigas Center
Pasig City

In accordance with the monthly report submitted by the Monitoring Team pursuant to its mandate to make use of the images/data stored from its Image Repository Database Server, as evidence against erring PETCs, PETC IT Service Providers and/or MVECTs, under sub-sections 13. 4 and 13. 5 of Department Order No. 2012-10, there is substantial proof that **SRM SMOKE ETC**, and its abovementioned IT Provider, violated section 6.7.1 of Department Order 2012-10 which reads:

"A PETC shall not upload faked, altered or manipulated webcam pictures of motor vehicles and technicians to the LTO-Image Repository Database Server and make it appear that the subject motor vehicle for inspection/registration underwent actual emission tests, and thereafter, reflected on the printed CEC Forms"

Plate Number JEU 708 tested on May 18, 2016 at 10:20:57 AM, Plate Number NOU 155 tested on May 19, 2016 at 5:05:24 PM, Plate Number WFE 975 tested on May 19, 2016 at 11:21:52 AM, Plate Number LNA 443 tested on April 14, 2016 at 3:36:55 PM, Plate Number YBX 468 tested on April 14, 2016 at 1:35:22 PM, Plate Number YEW 765 tested on May 17, 2016 at 5:07:40 PM.

The Committee cannot over-emphasize that without conducting actual emission test on subject vehicles, the issuance of CEC becomes null and void, and under 14.9, D.O. 2012-10; "Tampering of test result or entering any false information about the vehicle being tested," is a ground for withdrawal/cancellation of PETC's authorization."

WHEREFORE, the committee emphasizes that this is the 1st offense of the said PETC. You are given five (5) calendar days from receipt hereof to submit a verified Answer attaching therewith all evidence in support of your position. Thereafter, a hearing shall be conducted by the LTO Central Office Authorization Committee on **July 07, 2016 at 9:30 a.m.** in **Bulwagang Romeo F.**

Figure 1. A sample Show-Cause order

The profile of the data that can be explicitly extracted from the show-cause orders are shown in Table 1.

Table 1. Profile of Data that can be extracted from SCOs

Attributes	*Description/Possible Values*
Region	Regions 1,2,3,5,6,7,8,9,10,11,12,13, 4a,4b,car,ncr,armm
Province	Various provinces in the Phils
ITProv	LTO IT providers coded #1 … #9
ViolationType	Direct or Procedural
SCO Year	Year SCO was issued : 2010 to 2016
VehicleType	truck,uv,car,mc,jeep,tractor,suv,car go, dump,bus,mixer,auv,canter
Fuel	Gas or Diesel
Year Model	Vehicle model from 1961 to 2015

III. DATA MINING AND SELF ORGANZING MAP

The core of knowledge discovery process is data mining. According to Zaki and Meira [8], "Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data." It is a process of extracting information from large databases which produces interesting patterns and relationships. Data mining is an essential part of KDD process and considered as important subfield in knowledge management. It depends heavily on suitable machine learning tasks for pattern recognition and knowledge acquisition. There are two main tasks in datamining: (1)

Clustering and (2) Classification [9]. Clustering is the most prominent of these two core tasks. The main purpose of this task is to generate an optimal partitioning of the data set to be analyzed. Classification, on the other hand, is learning a function that maps (classifies) a data item into one of several predefined classes.

SOM is one of the most popular algorithms used in data mining [10]. It is a neural network tool based on unsupervised competitive learning and generates as output a topology which can preserve mapping from the high dimensional space onto a low-dimensional (typically two-dimensional) space in a way that similar inputs are positioned in locations closer to each other within the space (either a grid or lattice). This output is made possible through the use of a neighborhood function. A sample SOM output mapping is illustrated in Figure 2.

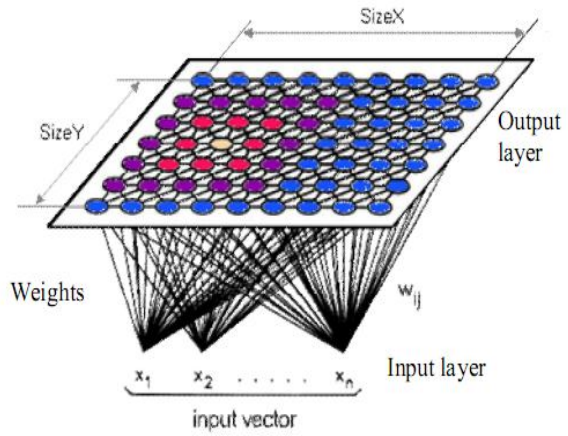


Figure 2. Sample SOM mapping

This output mapping is primarily used for the visualisation of high-dimensional data, but it can also be used for clustering and classification. Like other ANNs, the SOM has training parameters and is trained iteratively. The first step is to initialise the weights from N inputs to the M output nodes to small random numbers, and define an initial neighborhood radius for each node. The neighborhood d is the set of nodes which are similar to a given node. As the training progresses, the number of nodes in the neighborhood shrinks until only nodes representing similar inputs are left in the neighborhood. Thus, M > d(t1) > d(t2) . . . , where time t1 < t2 ....

As illustrated in Figure 2, each input vector x_i in the input layer is connected to every output node j in the output layer through a weight w_{ij}. At each step, an input vector x is presented to the SOM. The distance between x and the SOM weight vectors is calculated using the equation:

$||x - w_c|| = \min_i ||x - w_i||$ for i = 1, . . . , d (1)

where $||\cdot||$ is the Euclidian distance represented by the equation:

$$\epsilon = ||\mathbf{u} - \mathbf{o}|| = \sqrt{\sum_{i=1}^{M} [o_i - u_i]^2} \quad (2)$$

The neuron whose weight vector wc is closest to the input vector x is called the best-matching unit (BMU).

Whenever an example x is presented to the SOM, the closest representative weight vector wj is found. The weight update at each iteration for all i in the spatial neighbourhood of j is as follows:

$$wi = wi + [x - wi] \quad (3)$$

In this study the SOM implementation is trained by initially giving the training data their approximate position on the grid and initialising the weights to small random numbers. Input data is then presented to the SOM to determine the corresponding output map. The output for each input data instance is a point representing it in the SOM. As such, similar instances in the higher dimensional input space should have similar representation in the output space.

The SOM output can then be clustered by creating one cluster for every point in the output layer that has at least one data instance associated with it. While the SOM algorithm is primarily a clustering algorithm and not a classification algorithm, it can however, be used also for classification tasks. The latter simply requires applying labels to the map so that it can be used for classification. The SOM algorithm has an added advantage of requiring very little, if not no, preliminary data cleansing [11].

IV. EXPERIMENT SETUP

Figure 2 provides a diagram depicting the experiment setup. In line with the goals of the study, the experiment conducted encompasses two main steps. The first step involves training the self-organized map. Afterwards, in the second step, the resulting cluster is interpreted in order to identify patterns of relationship between the attributes that can be useful for policy improvement.

In this final step, the resulting map, is also treated as a classifier model and subsequently tested in order to explore and determine how well it would be able to predict categorical class labels from the dataset.

Pre-processing was conducted to convert the raw data extracted from the show-cause orders into a form suitable for processing. Codes were used to replace identifying values in order to anonymize the dataset. The data was also cleansed to correct/remove erroneous values and ensure consistency within the dataset.

Microsoft Excel was used during the initial preprocessing step to inspect and clean the data and generate summary statistics between the various attributes.

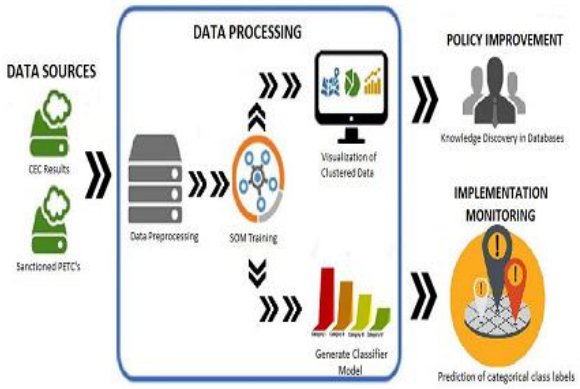


Figure 3. Conceptual Framework of the Study.

For actual SOM mapping, the Weka version 3.8 software installed with the SOM package tool was used. Weka have already been proven as an efficient tool for analyzing and predicting various types of data [12]. For this initial study, the default values of the parameters of SOM in Weka was utilized. Figure 4 shows the parameters available in Weka and their default values.

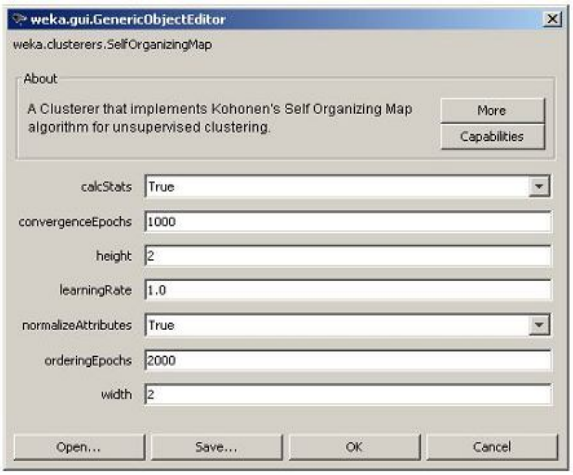


Figure 4. SOM parameters in WEKA

V. RESULTS

Figures and To achieve a better understanding of the nature of emission test violations and its underlying implications we have explored the dataset extracted from the Show-Cause orders issued by LTO using the SOM algorithm. Initial analysis was conducted by categorizing data instances based on the IT Provider that serviced the sanctioned transaction. This is followed by testing the level of accuracy that can be achieved if the SOM map was used as a classifier model. The results are discussed in succeeding sections.

C. Cluster Analysis

Figure 5 provides visualizations of the clusters produced using the default settings of SOM package in Weka. The interpretation of clusters obtained, on the other hand, is shown in Table V.

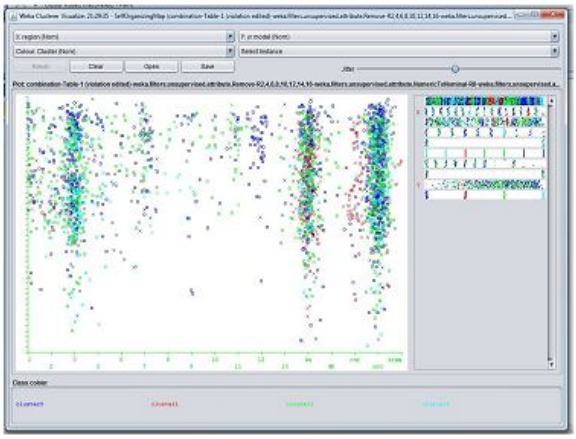


Figure 5. SOM clusters visualization in Weka

Even at this early stage, several remarkable observations can already be made. Foremost is the fact that in all clusters (except for cluster 2 which introduced CAR), the topmost regions where violations are often incurred is NCR, Region 4a, and Region 3. These are regions where some of the most highly urbanized cities are located such as Bulacan, Rizal, and Quezon City. The top IT providers in turn that have the higher propensity to commit violations are #1, #5, #2, and #7. It is worthy to note that these providers, which most often established themselves in highly urbanized locations, are run by small companies that are mainly dependent on the income that they generate in line with their service to LTO. On the other hand, bigger companies like #3, which have a wider network of business interests have less violations incurred.

The clusters generated also showed that owners of cars, trucks, motor cycles, and utility vehicles are the most often clients of emission violations where gasoline-based engines are the most dominant types. Looking at the models of these vehicles, it would not be surprising to see older vehicles (e.g. model 1993 and 1997) to be highlighted especially when the fuel type is diesel as these types of vehicles has the tendency to age quickly without proper maintenance. What is surprising however, is that for cluster 1 where the most recent SCOs are issued (i.e., 2015-2016), most of the vehicles involved are more recent models (i.e., 2010-2012) with gasoline-based engines. This is surprising because these vehicles are not yet that old to fail an emission test and yet their owners still chose to opt for a non-appearance test.

Such counter-intuitive examples of information already pinpoints the need for deeper and more thorough analysis of the dataset generated by SCOs in order to uncover the factors that induce the occurrence

of such violations and understand further the environment where violations abound. At the moment, the patterns seem to indicate that the motivations of motorists who opted to take non-appearance tests goes beyond the need to avoid repair or upgrade expenses as many of the vehicles involved are new models. Further information that can be generated from a more thorough analysis could provide stronger evidence to support more effective and adapted implementing guidelines and policies.

Table 2. Evaluation of SOM generated clusters

	Features	High end values	Low end values
Cluster1	Region	NCR, 4a, 3	NCR, 4a, 3
	Province	Bulacan, rizal, qc	Bulacan, rizal, qc
	IT Prov	#3, #8, #4	#9, #7, #2
	Type	1085:65	1085:65
	SCo Yr	2015(672);2016(478)	2015(672);
	Vehicle	Car, truck, UV, MC	2016(478)
	Fuel type	G(622): D(528)	Car, truck, UV, MC
	Model	2010 - 2012	G(622): D(528)
			2010 - 2012
Cluster2	Region	4a,CAR,NCR	4a,CAR,NCR
	Province	cavite, benguet, qc	cavite, benguet, qc
	IT Prov	#9, #3, #8	#7, #2, #5
	Type	304:6	304:6
	SCo Yr	2012	2012
	Vehicle	Car,uv,mc,truck	Car,uv,mc,truck
	Fuel type	G(149): D(161)	G(149): D(161)
	Model	1995,	1995,
		(1993,1997,2003),2005	(1993,1997,2003),2005
Cluster3	Region	Ncr,4a,3	Ncr,4a,3
	Province	Qc,rizal,Pampanga	Qc,rizal,Pampanga
	IT Prov	#3,#4,#8	#1,#2,#6,#7
	Type	951:46	951:46
	SCo Yr	2013-2014	2013-2014
	Vehicle	Car,truck,mc,uv	Car,truck,mc,uv
	Fuel type	G(493): D(604)	G(493): D(604)
	Model	2007,(2008,2009,20	2007,(2008,2009,20
		10),2005	10),2005
Cluster4	Region	Ncr, 3, 4a	Ncr, 3, 4a
	Province	Qc, bulacan, cavite	Qc, bulacan, cavite
	IT Prov	#9, #4, #8	#7,#1,#2,#6
	Type	255:122	255:122
	SCo Yr	2010	2010
	Vehicle	Car,truck,mc,uv	Car,truck,mc,uv
	Fuel type	G(214): D(163)	G(214): D(163)
	Model	1996,2007,1998	1996,2007,1998

D. Classification Analysis

On the other hand, to assess the performance of the generated map as a classification model, classification accuracy was used as the main metric. This computes the number of samples classified correctly over the number of instances in the dataset using the formula: Accuracy = # correctly classified/Total # instances (4)

The aim is to determine how effective the map is in predicting the occurrence of non-appearance violations given the various other attributes that can be extracted from the show-cause order form. The ability to predict whether or not violations will occur and which type (e.g., direct or procedural) can greatly help in formulating preemptive actions that minimize the occurrence of violations.

In this case, the performance of the SOM algorithm was tested by varying the percentage of data used to generate the SOM map. As with the clustering experiment, the SOM map used in the experiment also has a 2x2 dimension. Given this, it was found that as the size of the dataset increases, the ability of SOM networks to predict the presence of non-appearance violation also increases. The best accuracy achieved is 91.89% when the SOM was trained with around 2,500 data instances or 90% of the total dataset (see figure 6). Table 3 shows the results of SOM network with varying amounts of training data along with the amount of time it took to generate the classification results. It is clear from the table that although the SOM network's accuracy increases with the amount of training data, the time it took to generate the results, unfortunately, also increases proportionately.

Table 3. Classification accuracy of SOM

Percentage of Data	Acccuracy	Time
65%	88.82	32.8 sec
70%	89.57	35.3 sec
75%	90.26	37.6 sec
80%	90.87	40.3 sec
85%	91.41	42.6 sec
90%	91.89	45.3 sec

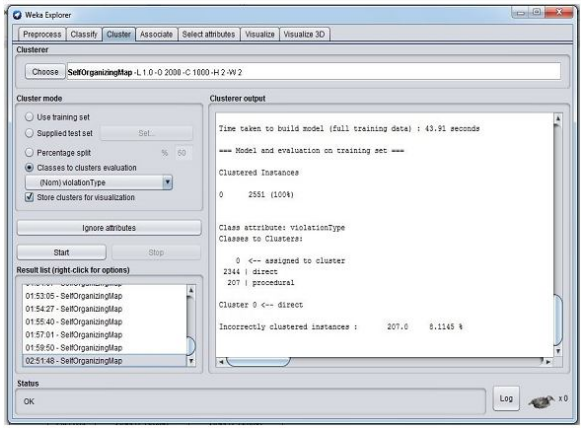


Figure 6. Highest accuracy achieved by SOM classifier

The results however, suggests a positive outlook with regards the proposed strategy. Although the results are only applicable to the dataset used, it does suggest that, with further study and development, there is a possibility that such a strategy can be put in place in the future.

CONCLUSIONS

This paper described a proposed strategy for analyzing emission compliance violations for vehicles applying for registration in LTO. The approach implemented consists of data preprocessing, then afterwards using the SOM algorithm to organize the data into clusters in order to identify patterns in the data that can be useful for policy makers. We have also presented the performance achieved when the SOM map generated was used as a classifier model that takes into consideration the various attributes of the dataset in

order to predict the occurrence of non-appearance violations. Although computationally becoming more expensive as the dataset grows, the experiments conducted shows that the SOM map can correctly predict the occurrence of violations ~91.89% of the time. The results also indicate that as the number of training instances increases, the ability of the classifier model to correctly predict violations also increases. It is possible that the approach described here can also be generalized for predicting other attributes within the dataset.

These initial findings only serve to encourage further attempts at applying SOM to the problem of analyzing emission compliance violations in LTO. Future work will involve exploring the effects of increasing the lattice size of the SOM and determining how it will affect the clusters formed. Procedures that can further improve the accuracy of the classifier model can also be investigated further. Needless to say, additional data to process can provide new samples that can only improve both the clustering and classification performance of SOM.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of the Department of Transportation (DoTr), Republic of the Philippines to the conduct of this study and for allowing access to the SCO dataset. Acknowledgment is also given to the University of Waikato in New Zealand, for the free software used in this study.

REFERENCES

[1] T. Kohonen, "Self-Organizing Maps. Springer, Berlin, Heidelberg," 1995.

[2] J. Vesanto, and E. Alhoniemi, "Clustering of the self-organizing map. IEEE Transactions on neural networks," vol. 11, no.3, pp. 586-600, 2000.

[3] A. Krupnick, R. Morgenstern, C. Fischer, K. Rolfe, J. Logarta, and B. Rufo, "Air pollution control policy options for Metro Manila," Resources for the Future Discussion Paper, Washington, DC., 2003.

[4] United Nation, "Transforming our world: the 2030 Agenda for Sustainable Development," 2015. Retrieved December 05, 2016, from https://sustainabledevelopment.un.org/post2015/transformingourworld/publication

[5] L. Legarda, "Keynote Speech of Senator Loren Legarda," September 2016. Retrieved January 1, 2017, from http://www.senate.gov.ph/press_release/2016/0926_legarda1.asp

[6] Eleventh Congress, "An act providing for a comprehensive air pollution control policy and for other purposes," June 1999. Retrieved November 11, 2016, from http://www.lawphil.net/statutes/repacts/ra1999/ra_8749_1999.html

[7] D. Srinivasan, and A.G. Tettamanzi, "An evolutionary algorithm for evaluation of emission compliance options in view of the clean air act amendments," IEEE Transactions on Power Systems, vol. 12 no.1, pp. 336-34, 1997.

[8] M.J. Zaki, W. Meira Jr, and W. Meira, "Data mining and analysis: fundamental concepts and algorithms." Cambridge University Press, 2014.

[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 1996. From "Data mining to knowledge discovery in databases," AI magazine, 17(3), 37.

[10] S.L. Shieh, and I.E. Liao, "A new approach for data clustering and visualization using self-organizing maps. Expert Systems with Applications", Vol. 39 no.15, pp. 11924-11933, 2012.

[11] R. Duda, P. Hart, and D. Stork, "Pattern Classification," Wiley, 1999.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, Volume 11, Issue 1, 2009. Reproduced with permission of the copyright owner. Further reproduction prohibited without permission, 2013.

★★★