Improvement Design of Fuzzy Geo-demographic Clustering using Artificial Bee Colony Optimization

Arie Wahyu Wijayanto, Ayu Purwarianti
School of Electrical Engineering and Informatics
Insitut Teknologi Bandung
Bandung, Indonesia

Abstract—**Geo-demographic analysis (GDA) is the study of geo-demographic that refers to spatial or geographical area, utilizing some spatial based analysis explicitly. Fuzzy Geographically Weighted Clustering (FGWC), a variant of Fuzzy C-Means (FCM), has been serving as an effective algorithm in Geo-demographic Analysis. FGWC is sensitive because of its initialization by determining random cluster centers makes the greater probability of clustering result falling into the local optima that affect the clustering quality. Artificial Bee Colony (ABC), one of metaheuristic algorithms is usually used as a global optimization tools. This research aims to propose a integration design of ABC based optimization and FGWC for improving geo-demographic clustering accuracy.**

I. INTRODUCTION

Geo-demographic analysis (GDA) studies the attributes of population demographics based on location, utilizing some spatial based analysis explicitly [1]. GDA is ussually employed for data exploration and analysis, and is widely applied to support effective development policies [2].

Fuzzy c-means (FCM) is a popular data mining algorithm which has been widely used as a clustering method and also as a conventional method in geo-demographic clustering [3]. FCM is an improvement method from k-means algorithm which permits an item to have a place with two or more groups with a fuzzy membership value somewhere around 1 and 0 [3]. FCM is also widely used in the most studies concerning GDA and fuzzy logic [4]. Fuzzy Geographically Weighted Clustering (FGWC) is a variant of FCM, which has been serving as the state of the art in geo-demographic clustering [2].

Some research recognized that FCM is effective [5], and its simple centroid-based iterative procedure is very engaging when managing substantial volume of data [3]. However, FCM is sensitive because of its initialization by determining random cluster centers makes the greater probability of clustering result falling into the local optima [5]. This weakness of FCM does not ensure to give the global optimal solution [6]. All of those weaknesses of FCM also found in FGWC.

A metaheuristic is a set of procedures for algorithmic development that can be implemented to differing optimisation issues with non significant changes [7]. Metaheuristic algorithms is usually used as a global optimization tools.

Artificial bee colony, one of metaheuristic algorithm, is effective when used on optimization of fuzzy clustering [8]. This research aims to propose an integration design of ABC based optimization and FGWC for improving geo-demographic clustering accuracy.

II. THEORETICAL BACKGROUND

A. Geo-demographic Analysis (GDA)

GDA is the study of geo-demographic that refers to spatial or geographical area and is generally utilized as a part of people in government and private areas in the arranging and procurement of goods and administrations [2]. Geo-demographic Analysis often uses clustering methods that are used to perform the segmentation of geo-demographic data, making the data more easy to maintain for analysis purposes [2].

In geo-demographic studies, there are two basic priciples: Firstly, it must begin with assumption that people living in the same geographical area are inclining to have resemble demographic profile than people at randomly chosen area; Secondly, it also give assumption that geographical areas can be characterized regarding their number of population and people interaction among them. Due to these two assumptions, clustering process is then implemented to determine geo-demographic data segmentation into meaningful clusters that describe existing profile, or relevant geo-demographic characteristics, in order to perform an easy and high quality data management and analysis [2].

B. Fuzzy Geographically Weighted Clustering (FGWC) Algorithm

FGWC Algorithm, which firstly proposed by Mason and Jacobson in 2007, provides a geographically aware alternative to a classical FCM method by providing the capacity to perform number of population and geographical distance effects for analyzing a geo-demographic cluster [9]. The influence of one area upon another is considered by FGWC as the multiplication of the areas number of populations. The divisor implements a distance decay effect through the weighting factor [9]. The fuzzy membership matrix is then adjusted using the geographical weight in each clustering cycle as defined in equation below [9]:

$$\mu'_i = \alpha\mu_i + \beta \frac{1}{A}\sum_j^n w_{ij}\ \mu_j \quad (1)$$

Where μ'_i is the adjusted fuzzy membership of area i and μ_i is the old fuzzy membership of area i. The n is number of area. The w_{ij} is defined as interaction weighting measure of two geographical areas i and j. The weight is decided by distance between area centers and the number of population of those areas. The A parameter is determined to ensure that the average of weighted membership values is still in the range of 0 and 1 [9].

$$w_{ij} = \frac{(m_i\, m_j)^b}{d_{ij}^a} \qquad (2)$$

Where $m_i\, m_j$ are the population of areas i and j respectively, d_{ij} indicates the geographical distance among areas i and j, then a and b are constanta parameters that is determined by user. The α and β are weights to old fuzzy membership value and the mean of membership values of neighborhood areas respectively and are calculated as follow this equation [9]:

$$\alpha + \beta = 1 \qquad (3)$$

FGWC improves the previous research, a proposed method using neighbourhood effect (NE) by Feng and Flowerdew (1998), which gives *ex post facto* modification of the cluster memberships after standard fuzzy clustering [9]. FGWC incorporating geography into geo-demographic analysis, so the cluster being sensitive to neighbourhood effects and will modify the cluster center values to perform more "geographically aware" segmentation [9]. A comparison overview of this geographically weighted clustering and its modification to classical fuzzy clustering method is presented in Figure 1.

Feng and Flowerdew incorporated neighbourhood effects after the process of fuzzy clustering, which gives better result for fuzzy clustering [9]. Mason and Jacobson integrated those two process in FGWC clustering iteration steps [9].

C. Artificial Bee Colony (ABC) Optimization

ABC Algorithm, firstly developed by Dervis Karaboga in 2005, is one of the metaheuristic algorithm [10]. The principle thought is that the most imperative piece of optimization is the main methods used to discover ideal solutions for a given issue under different constraint [7], [11]. A set of techniques are called metaheuristic methods, and are frequently nature-enlivened, impersonating some effective qualities in nature, which are regularly productive in practice in tackling troublesome optimization issues [7].

There are three kind of bees that ABC utilized: employed bees, onlooker bees and scout bees. The employed bees bring heaps of nectar originally from the food asset to its colony hive and may impart the data about it to the others in the dancing territory. Employed bees bring data about selected food sources and transfer the data with a certain probability through dancing activity in its colony hive. The onlooker bees wait in the dancing territory to determine the selected food source depends on the probability calculated by employed bees [12]. The each search iteration of the ABC algorithm consists of three process. First, sending the employed bees into promising food sources and then calculate the probability value of selected nectar. Then onlooker bees determine the food source areas and evaluating the value of nectar in the food sources. The final step are selecting the scout bees for finding another new promising sources [12].

Figure 2 presents an artificial bee colony (ABC) method for optimizing the n-dimensional function $f(x)$, where x, is the i -th candidate solution.

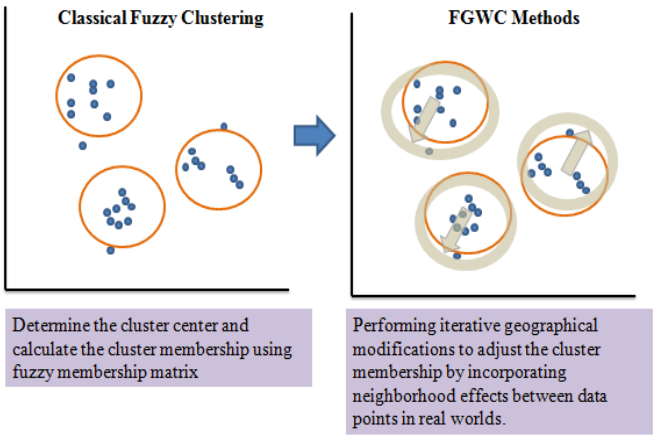

Figure 1. Comparison overview of FGWC Methods and Classical Fuzzy Clustering [9]

III. The Proposed Methods

A. The Improved FGWC Using ABC Optimization

FGWC algorithm have some limitations in initialization phase. First, the number of geo-demographic groups (clusters) must be define manually by user. Second, the center point of each cluster (centroids) are created randomly, that makes the greater probability of clustering result falling into the local optima.

The basic ideas is using ABC algorithm to select number of cluster and its centoids automatically in the initialization phase of FGWC clustering.

The basic objective function which will be minimized is:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m\, {d_{ij}}^2 \rightarrow \min \qquad (4)$$

where m is a weigting exponent that defines the fuzziness of the groups (clusters), u_{ij} is an element of partition matrix, d_{ij} is distance between i -th cluster center and j -th data point in Euclidean method.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \qquad (5)$$

Here is step by step of the improved FGWC Algorithm using ABC optimization:

Step 1: Determine the number of clusters c, threshold $\varepsilon > 0$ and other parameters such as weighted fuzziness exponent(m).

Step 2: Set initial value of cluster centers V_i at $t = 0$ using Artificial Bee Colony process in Figure 2 by minimizing the objective function in equation (4). Number of food sources, employed bees and onlooker bees is defined by number of cluster respectively. The ABC dimension is equal to number of clustering data variable. Best solution provided by ABC is chosen as the cluster center.

Set initial value of the positive integer L, which is the stagnation limit
Set initial value of the population size of employed bees $P_f < N$
Set initial value of the population size of onlooker bees $P_o = N - P_f$
Set initial value of a random population of employed bees $\{x_i\}$ *for* $i \in [1, P_f]$
Set initial value of the employed bee trial counters $T(x_i) = 0$ *for* $i \in [1, P_f]$
While not (termination criterion)
  Employed Bee:
  For each employed bee $x_i, i \in [1, P_f]$
    $k \leftarrow$ *random integer* $\in [1, N]$ *such that* $k \neq i$
    $s \leftarrow$ *random integer* $\in [1, n]$
    $r \leftarrow U[-1,1]$
    $v_i(s) \leftarrow x_i(s) + r(x_i(s) - x_k(s))$
    If $f(v_i)$ *is better than* $f(x_i)$ *then*
      $x_i \leftarrow v_i$
      $T(x_i) \leftarrow 0$
    Else
      $T(x_i) \leftarrow T(x_i) + 1$
    End if
  Next employed bee
  Onlooker Bees:
  For each onlooker $v_i, i \in [1, P_o]$
    Select an employed bee x_j, *where Pr(*x_j*)* $\propto$ *fitness (*x_j*) for* $j \in [1, P_f]$
    $k \leftarrow random\ integer \in [1, P_f]\ such\ that\ k \neq j$
    $s \leftarrow random\ integer \in [1, n]$
    $r \leftarrow U[-1, 1]$
    $v_i(s) \leftarrow x_j(s) + r(x_j(s) - x_k(s))$
    If $f(v_i)$ *is better than* $f(x_j)$ *then*
      $x_j \leftarrow v_i$
      $T(x_j) \leftarrow 0$
    Else
      $T(x_j) \leftarrow T(x_j) + 1$
    End if
  Next onlooker
  Scout Bees:
  For each employed bee $x_i, i \in [1, P_f]$
    If $T(x_i) > L$ *then*
      $x_i \leftarrow$ *randomly-generated individual*
      $T(x_i) \leftarrow 0$
    End if
  Next employed bee
Next generation

Figure 2. Artificial Bee Colony Algorithm Pseudo Code [13]

Step 3: Define geographic parameters $\alpha, \beta, a,$ and b which will be used to adjust the partition matrix following by geographical characteristics.

Step 4: The equation (5) is then utilized to determine the fuzzy membership values.

Step 5: Perform geographical cluster modifications using equations (1), (2) and (3) to involve the neighbourhood effect.

Step 6: Use ABC and formula (1) to compute the cluster centers at $t + 1$ by minimizing the objective function in equation (4).

Step 7: If the error of $\left\|V^{(t+1)} - V^{(t)}\right\| \leq \varepsilon$ then stop the iterative procedure. Otherwise, assign $V^{(t)} = V^{(t+1)}$ and return to Step 2.

The detailed flowchart of the proposed method are shown in Figure 3.

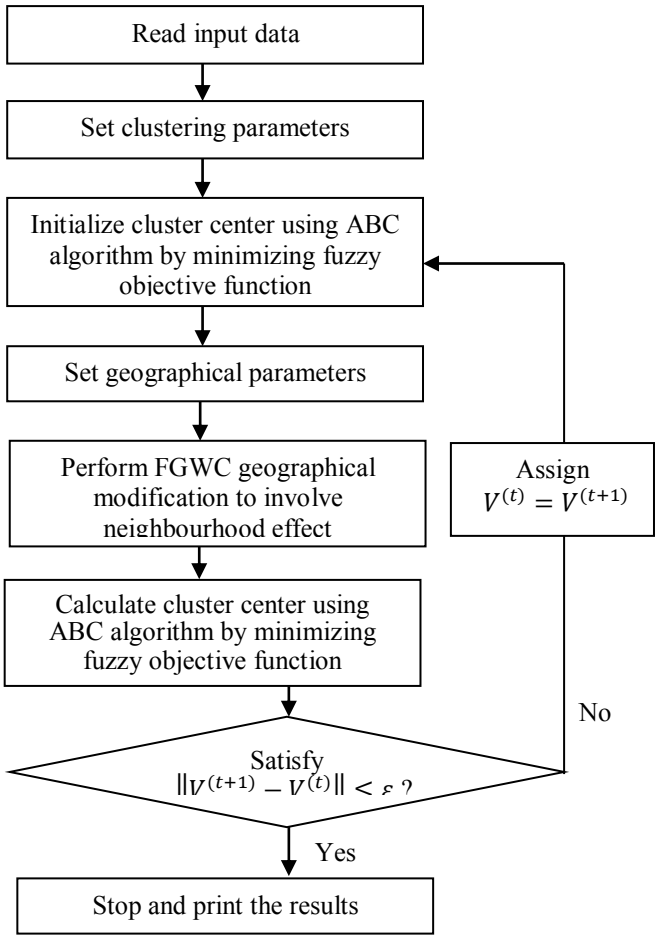

Figure 3. Flowchart of improved FGWC using ABC algorithm

B. Clustering Validity

The objective of this proposed method is model accuracy, which can be measured by Partition Coefficient (PC) validity

index, Classification Entropy (CE) index, Partition Index (SC), Separation Index (S), Xie and Beni's Index (XB). Those measurement are usually used to measure the performance of clustering algorithms [4] and are used as observed variables in this research. According to this research framework, the accuracy of fuzzy geo-demographic clustering problem is aimed to be improved. The Partition Coefficient index calculates the overlapped data points between clusters and for c number of cluster is defined as follows [4]:

$$PC = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^{2} \quad (6)$$

where μ_{ij} defines the fuzzy membership of item j in cluster i.

The CE index computes the fuzziness of the group (cluster) segmentation and was determined as follows [4]:

$$CE = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log(\mu_{ij}) \quad (7)$$

The greater value of PC index and the smaller value of CE index indicates the better clustering quality resulted. The SC index defines the ratio of the total compactness and separation of the clusters and is defined as follows [4]:

$$SC = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}(\mu_{ij})^{m}\|x_j - v_i\|^2}{N_i\sum_{k=1}^{c}\|v_k - v_i\|^2} \quad (8)$$

A better partition is indicated by a lower value of SC. The S index uses a minimum-distance separation for partition validity, while XB determines the ratio quantification of the total variation inside the clusters and the separation among clusters.

$$S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{2}\|x_j - v_i\|^2}{N\, min_{i,k}\|v_k - v_i\|^2} \quad (9)$$

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{m}\|x_j - v_i\|^2}{N\, min_{i,j}\|x_j - v_i\|^2} \quad (10)$$

The better number of cluster is indicated by smaller value of S index and XB index [4], [14].

IV. EXPERIMENTAL RESULTS

The preliminary result of experiment will reported in this section. We develop the implementation code of the proposed methods in Matlab R2013a environment. Then it has executed under the environment of Intel Core i5-3210M CPU @2.50GHz, 4GB RAM and Windows 7 64bit operating system. The public data set from Indonesian Population Census 2010 [15] which consist of 110 sociodemographic variables and 33 regions/provinces were utilized to evaluate the proposed method. The improved algorithm called FGWC-ABC was tested against existing original FGWC [9], and NE [9]. We set the parameters of those algoritms using threshold $\varepsilon = 10^{-4}$, a = 1, b = 1, α = 0.5 and β = 0.5 and m=2. All methods were run 100 times.

We use various number of cluster to evaluate the methods. The evaluation using Partition Coefficient validity index shown in table 1. The greater PC value resulted the better clustering process performed. The result of proposed FGWC-ABC is better than the original Fuzzy Geographically Weighted Clustering (FGWC). The result of FGWC and NE are not significantly different using three (3) decimal number rounding. Actually their PC value are not equal.

TABLE I. COMPARISON OF PC INDEX VALUE

Number of cluster	*NE*	*FGWC*	*FGWC-ABC*
2	0.500	0.500	**0.537**
3	0.333	0.333	**0.375**
4	0.250	0.250	**0.287**
5	0.200	0.200	**0.233**
6	0.167	0.167	**0.218**
7	0.125	0.125	**0.149**
8	0.125	0.125	**0.193**
9	0.111	0.111	**0.142**

The result of Classification Entropy validity index is shown in table 2. Better clustering process is measured by minimum value of CE validity index. The proposed FGWC-ABC gives better result than original FGWC and NE in various number of clusters. Similar with PC index evaluation, the result of FGWC and NE are not significantly different using three (3) decimal number rounding. Actually their CE validity index value are not equal.

TABLE II. COMPARISON OF CE INDEX VALUE

Number of cluster	*NE*	*FGWC*	*FGWC-ABC*
2	0.693	0.693	**0.655**
3	1.099	1.099	**1.039**
4	1.386	1.386	**1.313**
5	1.609	1.609	**1.537**
6	1.792	1.792	**1.666**
7	2.079	2.079	**2.010**
8	2.079	2.079	**1.888**
9	2.197	2.197	**2.094**

Table 3 present the comparison of SC value evaluation for those four methods. The minimum value of SC are representing the better clustering process. It is shown that the proposed method (FGWC-ABC) gives better result than original FGWC and NE invarious number of clusters.

TABLE III. COMPARISON OF SC INDEX VALUE

Number of cluster	*NE*	*FGWC*	*FGWC-ABC*
2	3.57E+02	7.01E+07	**1.94E-04**
3	1.96E+02	1.64E+07	**1.46E-04**
4	7.97E+01	1.50E+07	**3.16E-05**
5	3.53E+01	5.44E+06	**3.52E-05**
6	6.17E+01	5.49E+06	**6.87E-05**
7	1.96E+01	4.65E+06	**2.94E-05**
8	1.96E+01	4.65E+06	**6.87E-05**
9	1.13E+01	6.00E+05	**3.78E-05**

We evaluated the performance of those three methods in S validity index as shown in table 4. It is clear that FGWC-ABC result can reach better geo-demographic clustering quality measured by S index. FGWC-ABC gives minimum value of S index than other method.

TABLE IV. COMPARISON OF S INDEX VALUE

Number of cluster	*NE*	*FGWC*	*FGWC-ABC*
2	3.57E+02	7.01E+07	**1.94E-04**
3	2.63E+02	2.71E+07	**1.86E-04**
4	1.17E+02	2.25E+07	**4.01E-05**
5	5.46E+01	8.81E+06	**4.34E-05**
6	8.68E+01	7.98E+06	**9.95E-05**
7	2.92E+01	6.31E+06	**3.46E-05**
8	2.92E+01	6.31E+06	**9.74E-05**
9	1.68E+01	9.30E+05	**5.74E-05**

Comparison of XB validity index result is shown in table 5. The detailed simulation on various number of clusters shown that the proposed method can improve original FGWC to reach a better clustering quality measured by Xie and Beni's (XB) index value.

TABLE V. COMPARISON OF XB INDEX VALUE

Number of cluster	*NE*	*FGWC*	*FGWC-ABC*
2	7.97E-01	1.93E+00	**1.35E+03**
3	5.31E-01	1.29E+00	**6.30E+03**
4	3.98E-01	9.64E-01	**1.01E+03**
5	3.19E-01	7.71E-01	**8.14E+02**
6	2.66E-01	6.43E-01	**2.40E+04**
7	1.99E-01	4.82E-01	**5.29E+02**
8	1.99E-01	4.82E-01	**1.32E+08**
9	1.77E-01	4.29E-01	**4.10E+03**

Using those five different fuzzy clustering validity evaluation, we can conclude that the proposed design can improve the original Fuzzy Geographically Weighted Clustering (FGWC) to reach a better geo-demographic clustering quality. The FGWC-ABC is also better than NE method.

V. CONCLUSION

This paper aims to propose the design for improvement of the limitations in fuzzy geo-demographic clustering algorithm by proposing an integration of Artificial Bee Colony (ABC) based optimization and Fuzzy Geographically Weighted Clustering (FGWC) algorithm to reach a better geo-demographic clustering accuracy.

The design are using ABC algorithm to select the cluster centers (centroids) automatically in the initialization phase of FGWC clustering. Preliminary experimental simulation give promising result that the proposed method give better clustering quality than the original FGWC. The result also shown that this improvement gives better clustering quality than the popular Neighborhood Effect algorithm (NE) This proposed integration design can be implemented as a contribution to enrich the fuzzy geo-demographic clustering field.

REFERENCES

[1] A. Páez, M. Trépanier, and C. Morency, "Geodemographic analysis and the identification of potential business partnerships enabled by transit smart cards," *Transp. Res. Part A Policy Pract.*, vol. 45, no. 7, pp. 640–652, Aug. 2011.

[2] L. H. Son, B. C. Cuong, P. L. Lanzi, and N. T. Thong, "A novel intuitionistic fuzzy clustering method for geo-demographic analysis," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9848–9859, Aug. 2012.

[3] J. Wu, H. Xiong, C. Liu, and J. Chen, "A Generalization of Distance Functions for Fuzzy c-Means Clustering With Centroids of Arithmetic Means," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 3, pp. 557–571, Jun. 2012.

[4] G. Grekousis and H. Thomas, "Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods," *Appl. Geogr.*, vol. 34, pp. 125–136, May 2012.

[5] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1835–1838, Mar. 2011.

[6] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook : Second Edition*. Springer, 2010.

[7] X. Yang, "Nature-Inspired Mateheuristic Algorithms: Success and New Challenges," *J. Comput. Eng. Inf. Technol.*, vol. 01, no. 01, 2012.

[8] D. Karaboga and C. Ozturk, "Fuzzy clustering with artificial bee colony algorithm," *Sci. Res. Essays*, vol. 5, no. 14, pp. 1899–1902, 2010.

[9] G. A. Mason and R. D. Jacobson, "Fuzzy Geographically Weighted Clustering," in *Proceedings of the 9th International Conference on Geocomputation*, 2007, no. 1998, pp. 1–7.

[10] D. Karaboga and B. Basturk, “On the performance of artificial bee colony (ABC) algorithm,” *Appl. Soft Comput.*, vol. 8, no. 1, pp. 687–697, Jan. 2008.

[11] X. Yang and L. Press, *Nature-Inspired Metaheuristic Algorithms Second Edition*, 2nd ed. Luniver Press, 2010.

[12] M. Horng, “Multilevel thresholding selection based on the artificial bee colony algorithm for image segmentation,” *Expert Syst. Appl.*, vol. 38, no. 11, pp. 13785–13791, May 2011.

[13] D. Simon, *Evolutionary Optimization Algorithms : Biologically Inspired and Population-Based Approaches to Computer Intelligence*. Wiley, 2013.

[14] B. Balasko, J. Abonyi, and B. Feil, “Fuzzy Clustering and Data Analysis Toolbox: For Use with Matlab,” Veszprem, Hungary, 2005.

[15] BPS, “Statistics Indonesia - 2010 Population Census,” 2014. [Online]. Available: sp2010.bps.go.id. [Accessed: 01-Sep-2014].