Proceedings World Geothermal Congress 2015
Melbourne, Australia, 19-25 April 2015

Integration of Stream Sediment Geochemical and Airborne Gamma-Ray Data for Surficial Lithologic Mapping Using Clustering Methods

Husin Setia Nugraha[*], Emmanuel J. M. Carranza[**], and Mark Van Der Meijde[***]

[*]Directorate for Geothermal, Pegangsaan Timur 1A, Jakarta, Indonesia

[**]James Cook University, Townsville, Qld 4811, Australia

[***]University of Twente, Hengelosestraat 99 Enschede, The Netherlands

husin_setia_n@yahoo.com, john.carranza@jcu.edu.au, m.vandermeijde@utwente.nl

Keywords: closure problems, compositional data (CoDa), partition around medoids (PAM), model-based clustering (Mclust), British Columbia Canada.

ABSTRACT

In surficial lithologic mapping, geologists use remotely sensed data prior to fieldwork, however, the utility of these datasets are limited due to vegetation cover. Thus, the use of other sources of information about chemical and physical properties of rocks such as geochemical data (e.g., from stream sediment samples) and airborne geophysical data (e.g., radiometric data) becomes important. In this study, two clustering algorithms, partition around medoids (PAM) and model-based clustering (Mclust) were performed in stream sediment geochemical (SSG) and airborne-gamma-ray (AGR) data as well as in SSG and AGR together to map surficial lithologies in vegetation-covered areas in Central Part of British Columbia Province-Canada. Prior to clustering two approaches, conventional and compositional (CoDa), were applied to SSG and AGR data in order to study the influences of closure problems within the data. In SSG data analysis, clustering was applied using all 13 elements and selected nine elements. In addition, two types of data integration was done SSG all element and AGR (Reference Data I); and SSG selected element and AGR (Reference Data II). Overall accuracy and kappa coefficient was computed for the results and, two references were used to assess accuracy of the classification which is simplified existing lithological map (Reference Data I) and the lithological map based on the interpretation of airborne magnetic data (Reference Data II).

The integration of SSG and AGR data produces better results than those using both SSG and AGR data separately. The percentage accuracies of integration data compare to their separated data increase quite significant up to 17% and 0.15 for overall accuracy and kappa coefficient, respectively. In addition, Mclust produces better classifications for lithilogical mapping relatively to PAM clustering base on both qualitative and quantitative assessments. Qualitatively, from visual evaluation, the patterns of Mclust results are more similar to lithological patterns in the existing lithological map than PAM clustering. Quantitatively, assessments results in each separated data (SSG or AGR) show up to 5% and 0.7 differences for overall accuracy and kappa coefficient, respectively, whereas for the integrated data (SSG and AGR) produces non-significant difference results (1% and 0% differences for overall accuracy and kappa coefficient, respectively). Therefore, Mclust could be applied to integrate and classify SSG and AGR data for lithological mapping in regional scale.

1. INTRODUCTION

Surficial lithologic mapping in vegetation-covered areas is not simple task for geologists. The situation is worse when, in those areas, only limited outcrops of rocks exist. For lithological mapping in those areas, geologists usually have to derive optimum prior information from available remote sensing data before going on fieldwork. Nevertheless, the use of satellite spectral images will be limited because of vegetation cover. Therefore, in addition to field data, the use of surficial geochemical data (e.g., from stream sediment samples) and airborne geophysical data (e.g., radiometric data) becomes important sources of information about the chemical and physical properties in those areas.

Stream sediment and airborne gamma-ray data contain geochemical properties; thus, it will be advantageous to integrate information from these data sets. Stream sediment data are point data with irregular pattern of sample locations and non-uniform sampling density because the samples are taken by following rivers. Stream sediment data usually contain concentrations of many elements. In contrast, airborne gamma-ray data contain concentrations of only three elements but these data have regular sampling pattern and uniform sampling density. Consequently, when these two types of data are integrated, the strength of one data type compensates the weakness of the other. For example, the multiple elements in stream sediment data compensate for the only three elements in airborne gamma-ray data. In addition, the high sampling density of airborne gamma-ray would result in integrated data with higher spatial resolution than the stream sediment data.

However, a problem that arises when integrating airborne gamma-ray and stream sediment data is in representing point data of stream sediment into continuous data because stream sediment samples represent only materials within catchment basins of every sampling site. Some authors tried to find appropriate technique for representing stream sediment geochemical data. Bonham-Carter et al. (1987); Carranza and Hale (1997) and Spadoni et al., (2004) applied catchment basin approach to represent stream sediment data. This approach considers that stream sediment samples represent several sources and processes within catchment basins of every sampling site. The sources and processes include minerals of bedrock, minerals formed during weathering, minerals typical of mineralization, and anthropogenic substances (Howarth, 1984; Naseem et al., 2002). Robinson et al. (2004) demonstrated the use of inverse distance weighting (IDW) and kriging interpolation in order to observe regional-scale spatial variation of stream sediment and water geochemical data in New England (USA). Recently, Carranza (2010) explained that representing stream sediment geochemical data as discrete or continuous landscapes depend on mapping scale. For regional scale (e.g., 1:100.000 or smaller), representing stream sediment geochemical data as continuous landscapes by interpolation technique is plausible because its purpose to delineate anomalous areas for further investigations at higher scales could be achieved, whereas representing the data as discrete landscapes such as sample catchment basins could be both tedious and impractical.

Other problems that might rise in using geochemical data such as from stream sediments or airborne gamma ray data are related to “closure” that is inherent in compositional data such concentration of elements. Compositional data are characterized by its relative contained information because the data are ratio values (e.g., expressed as ppm, %, etc.) but not absolute values. Other characteristics of compositional data are that they always have positive values and the sums of the element data per sample are constrained to a constant value (*k*) such as 100 wt% or 1,000,000 ppm. Therefore, compositional data always have limited range between 0 and k (Pawlowsky-Glahn and Egozcue, 2006). One of the problems, which might is caused by this closure property of geochemical data, is the skewed data distribution which means not following a normal distribution. Direct application of statistical techniques to the non-normally distributed data could produce improper results because many statistics techniques rely on the assumption of normal data distribution. Moreover, data transformation such as logarithmic transformation is a common technique in order to solve this problem. However, according to Filzmoser et al. (2009a), conventional data transformations such as logarithmic transformation do not solve problems associated with closure property of compositional data. Furthermore, Pawlowsky-Glahn and Egozcue (2006) explained that closure-related problems also produce less or no significant information in geologic sense when multivariate techniques such as principal components analysis are applied. Other problem associated with closure is untrue correlation among compositional variables, which is caused by the ratio values that are contained to a constant sum in compositional data.

The main objective of the research is to map surficial lithologies in vegetation-covered areas to assist field work preparation for lithological mapping by integrating stream sediment geochemical data and airborne gamma-ray data in regional scale. The following sub-objectives are composed in order to achieve the main objective to quantify the significance of compositional data approach application in stream sediment geochemical and gamma ray data for surficial lithologic mapping; to perform clustering methods in stream sediment geochemical and airborne gamma-ray data for mapping the lithologies; to perform clustering methods for integrating stream sediment geochemical and airborne gamma-ray data as applied to surficial lithologic mapping.

The present research used clustering methods in order to integrate stream sediment geochemical data and airborne gamma-ray data. These methods were chosen because they are unsupervised and, thus, are appropriate in areas where no or little a-priori information about the objects to be mapped is available. Moreover, clustering methods are independent of grid size and, thus, are more robust to the influence of significant spatial resolution differences such as between stream sediment and airborne gamma-ray data.

The two clustering algorithms used in this research are Model-based clustering (Mclust) and Partition Around Medoids (PAM), representing respectively model-based and distance-based clustering. In distance-based clustering, cluster members are determined by calculating the distances between the samples. In model-based clustering, clusters are determined by selecting an appropriate model for the data. Furthermore, both of those clustering techniques were chosen because their algorithms are robust to existence of outliers in data (Gan et al., 2007; Kaufman and Rousseeuw, 2005). These two methods were also applied to stream sediment geochemical data and airborne gamma-ray data in order to investigate the difference in the performance with respect to these two types of data.

2. STUDY AREA AND DATASETS

2.1. Study Area

2.1.1 Location

The study area is situated at regional district of Bulkley-Nechako in Northern-Central of British Columbia province (figure 1). The area was chosen due to its characteristics and data availability that appropriate with the objectives of the research such as vegetation-covered areas (Delong, 1996). Regarding to data availability, besides input data such as stream sediment geochemical and airborne gamma-ray data, reliable geologic map for validation is also available. In addition, the dominant landform of this regional district is the Nechako plateau. The areas consist of Bulkley Valley, the northern part of the Nechako District, and the Omineca District, including portions of the Hazelton Mountains and Omineca Mountains in the west and north of the regional district, respectively (http://en.wikipedia.org/wiki/Regional_District_of_BulkleyNechako). The study area bounded by geographic coordinates (372750mW, 6095500mN) and (423500mW, 6134500 mN) and covers an area of ~2,000 km^2.

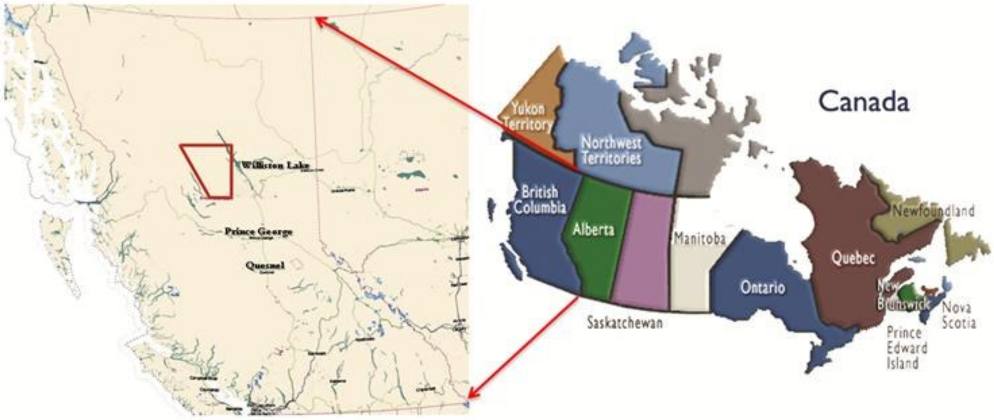


Figure 1. Location of study area (the polgyon) in the northern central part of the British Colombia Province.

2.1.2 Geology

The area is dominantly underlain by the Quesnel Terrane or Quesnelia. Two groups of rocks form this terrane, the Takla Group at the northern part and the Nicola Group at the southern part. The terrane is intruded by the northwest-elongate Hogem batholiths. The Takla Group consists of sedimentary units of Late Triassic in age. This group is overlain by volcanic, pyroclastic, and piclastic rocks; and intruded by early a Jurassic pluton. Augite phyric rocks are dominant with plagioclase and hornblende (Nelson et al.,1992; Nelson, 1991). Takla Group volcanics are unusually K-rich and alkalic (Delong, 1996).

Nelson (1991) divided the Takla Group into four interfingering formations, the Rainbow Creek, Inzana Lake, Witch Lake and Chuchi Lake Formations. In stratigraphy, Rainbow Creek is the lowest unit overlain by the Inzana Lake, Witch Lake and Chuci Lake Formation, in upward sequence. The Rainbow Creek Formation is comprised of dark grey to black slates or phyllites with interbeded quartz-rich siltstone and sandstone. The Inzana Lake Formation consists of epiclastic and sedimentary rocks with minor pyroclastic rocks. The Witch Lake Formation is dominated by an augite porphyry suite, which was produced from explosive intermediate volcanism. The Chuchi Lake Formation is made up of volcanic rocks with andesitic to latite-andesite composition. The phenocryst assemblage of these volcanic rocks is dominated by plagioclase with variable amounts of augite and hornblende (Delong, 1996; Nelson, 1991).

Figure 2 is simplified lithologic map, which was used for validation of results of this study. The geologic map from Massey et al. (2005a, b, c,d) were simplified base on regional geologic map from Nelson (1991). The lithological units were divided according to age group. For sedimentary rocks, the lithological units were divided into four lithological units based on age (from Proterozoic to Quaternary). Small lithological units such as ultramafic rocks and metamorphic rocks were merged with the larger lithological unit wherein they lie. Intrusive rocks comprise two formations, which are the Chuchi Syenite and Klawli Pluton Formations. Two formations, the Chuchi Lake Succession Formation and Witch Lake Formation, comprise volcanic rocks. Therefore, there are six lithological units for validation which are Intrusive Rocks, Volcanic Rocks and four sedimentary rocks units. The four sedimentary rocks units are Sedimentary Rocks 1 which comprises of sedimentary rocks from Jurassic to Quaternary, Sedimentary Rocks 2 that is contained sedimentary rocks from Triassic to Jurassic and small parts of metamorphic rocks, Sedimentary Rocks 3 which are constituted by sedimentary rock from Ordovician to Jurassic and Sedimentary Rocks 4 that comprises sedimentary rock from Proterozoic to Ordovician and small parts metamorphic rocks. The subset maps from the map shown in figure 2 were used for validation of the results of the study.

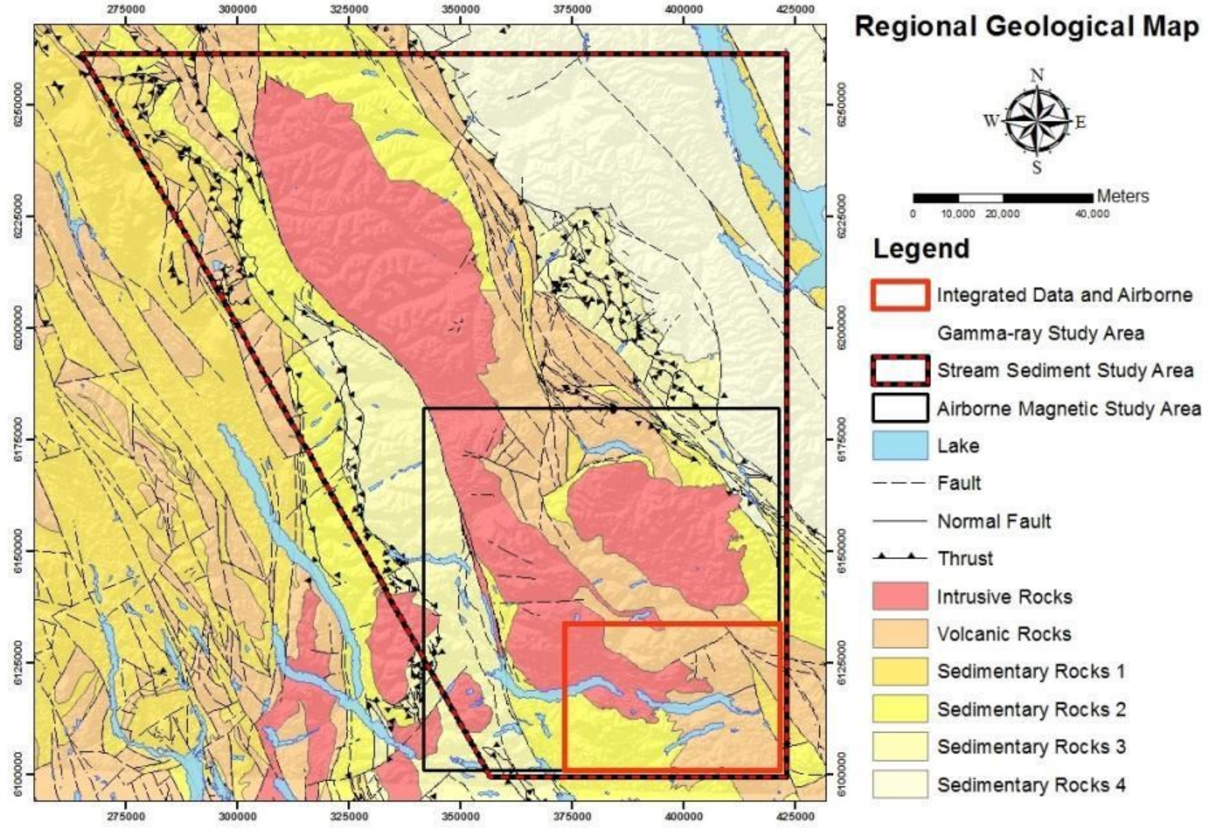


Figure 2. Simplified geological map used for validation of results (modified from Massey et al., 2005a, b, c, d).

2.2. Description of stream sediment geochemical datasets

2.2.1 The stream sediment geochemical (SSG) datasets

Stream sediment geochemical data used here were collected by Geological Survey of British Columbia during a National Geochemical Reconnaissance Program of Canada (NGR) that began in 1975. The data were sampled from the first and/or second order streams, producing average density about a sample per 13 km^2. The samples were taken from active part of stream channel with two-thirds of the sample paper bag was filled with silt or fine sand. In the laboratory, the samples were air dried at temperature below 40°C and sieved using a minus 80-mesh (177 µm) screen. The samples were analyzed for base and precious metals, pathfinder elements and rare earth elements by instrumental neutron activation analysis (INAA) and inductively coupled plasma mass spectrometry (ICP-MS). For quality control, control reference and blind duplicate samples were inserted into each block of twenty stream sediment samples (Jackaman and Balfour, 2008).

The stream sediment geochemical data were downloaded from the Geoscience Data Repository of Natural Resources Canada website (http://gdrdap.agg.nrcan.gc.ca/geodap/home/Default.aspx?lang=e), then subset to the research area. The data consist of sixteen elements (Zn, Cu, Pb, Ni, Co, Ag, Mn, Fe, Mo, Hg, Sb, As, Ba, Ce, Cr and Rb) with 2,478 sampling points including 284 duplicate samples. The concentrations of elements were measured in ppm, except for Fe in percentage whereas Ag and Hg were measured in ppb. Duplicate samples were used to analyze data quality and were excluded from statistical analysis.

2.2.2 The airborne gamma-ray (AGR) datasets

The airborne radiometric data used in this research were downloaded from the National Gamma-Ray Spectrometry Program (NATGAM) Data Base from the Geological Survey of Canada website (Natural Resources Canada, 2010b). Data on concentrations of radioelements (K, eTh, and eU) are available in 250 m grid size. K concentration is measured in percentage (%), whereas eTh and U are in parts per million (ppm). The data are the result of several hundreds of airborne surveys of the Geological Survey of Canada for over 30 years since 1970. The aircraft flew along a pattern of parallel flight lines at 120 m terrain clearance with line spacing 200-500 m. The aircrafts flew at speed around 190 km/h. The data were acquired by sampling every 1 second interval, which equivalent around 60 m on the ground (Natural Resources Canada, 2010b).

3. METHODOLOGY

The integration of SSG and AGR data consist of four stages, namely: data preparation; data integration using clustering method; post clustering; and assessment. Data preparation comprises of data transformation and standardization of SSG data, and then their interpolation in order to convert the point SSG data into a continuous data layer like the AGR data. The interpolation of the SSG data was performed using *ArcGIS 10*, whereas variogram analysis before interpolation and clustering analyses were executed using *R* statistical software. Two clustering algorithms – Mclust and PAM – were performed to integrate the SSG and AGR datasets. Post clustering processes such as reclassification and filtering were employed before the assessments. From the error matrix, overall and producer's accuracy and kappa coefficient were calculated to assess the classification results. The assessments were conducted using the two reference maps shown in *figure 3*.

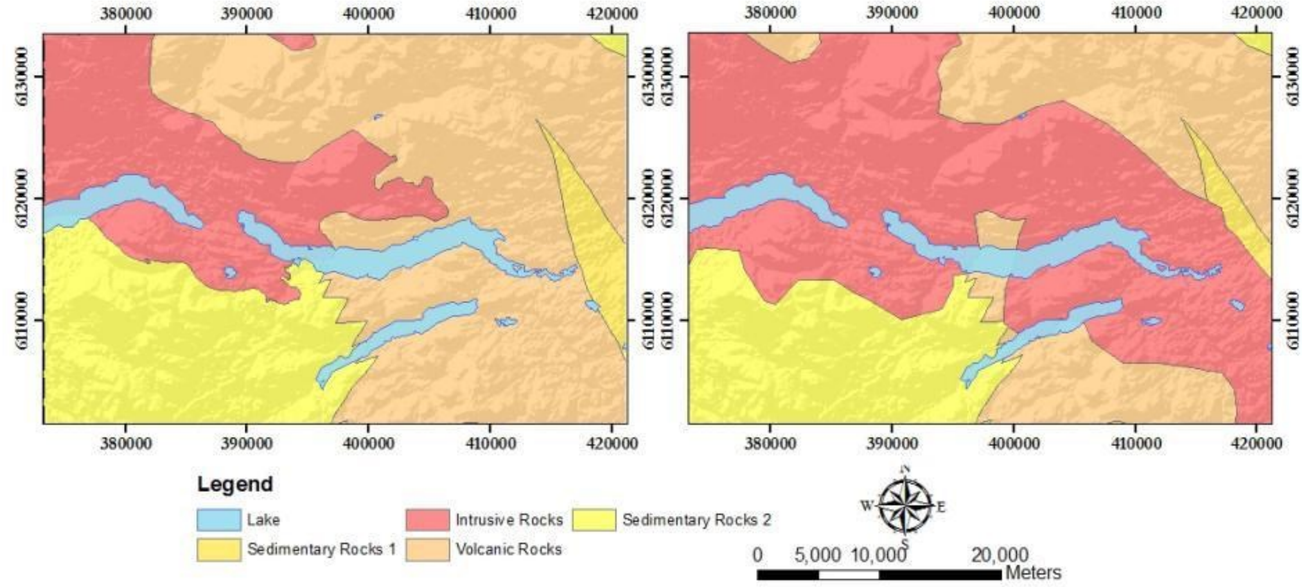


Figure 3. The study area and maps used for validation of results: (a) existing lithological map; (b) lithological map based on the interpretation of airborne magnetic data (modified from Massey et al. (2005a,b,c,d)).

3.1. Data preparation

To investigate the spatial data structure of the geochemical data for each element, variogram analyses were performed. The variogram models and the point geochemical data were used to interpolate unsampled areas using universal kriging. As explained by Carranza (2010), interpolation is a plausible technique to make continuous geochemical landscapes from stream sediment geochemical data in regional scale. Furthermore, Robinson et al. (2004) stated that universal kriging is an appropriate technique to interpolate stream sediment geochemical data in regional scale.

Data transformation and standardization were then applied to the SSG data before clustering techniques were applied. Base-10 logarithmic and square root transformation were applied to the SSG and AGR data, respectively, to get symmetric shape in data distribution. Transformation is chosen based on skewness that is nearest to zero. The zero skewness value usually shows symmetric shape of data distribution even it is not necessary. Standardization using Median Absolute Deviation (MAD) and median, which was developed by Yusta et al. (1998), equation 1 and equation 2, was employed in order to make comparable data range. This type of standardization is preferred to be employed than standardization using mean and standard deviation because its robustness to existence of outlier data (Carranza, 2008; Reimann et al., 2008).

$$Z_{ij} = \frac{X_{ij} - median_j}{MAD_j} \tag{1}$$

where

$$MAD = median[|X_i - median(X_i)|] \tag{2}$$

X = measurements of element concentrations; i = sample number; and j = element number

3.2. Clustering

Mclust and PAM clustering methods were applied to stacked raster images of SSG and AGR data. As experiments besides using all 13 elements in the SSG data, clustering was also applied to the nine selected elements in the SSG data. Thus, there are two data integration results: Integrated Data I is from combination of SSG all elements with AGR elements and Integrated Data II is from combination of SSG selected elements with AGR elements.

The main purpose of clustering is to find patterns such as groupings in characteristics or behaviours within observation datasets. Observation data are measured as element concentrations of stream sediment geochemistry. The data, based on their characteristics, are classified into groups/clusters/classes based on particular similarity/dissimilarity criteria. The aim of clustering algorithms is to minimize the dissimilarity objects within a group. Consequently, objects with a high degree of similarity are classified into the same cluster.

Furthermore, according to similarity/dissimilarity criteria, clustering algorithms could be divided into two approaches, distance-based and model-based approaches (Gan et al., 2007; Reimann et al., 2008). The distance-based approach clustering algorithm determines cluster members by calculating the distances between the samples; whereas the model-based by selecting appropriate model for the data.

PAM clustering and Mclust, representing distance-based and model-based approaches, respectively, were applied to the data. Both of these clustering techniques were chosen because of their robustness technique to outlier data (Gan et al., 2007; Kaufman and Rousseeuw, 2005). In addition, Templ et al., (2008) stated that partitioning method such as PAM performs better than hierarchy method for large data and the results from model-based clustering are more reliable and interpretable. Therefore, these two techniques were chosen for the purpose of comparing the performance of distance- and model-based clustering algorithms in classifying the multivariate geochemical to assist lithological mapping.

3.2.1 Model-based Clustering (Mclust)

The Mclust algorithm optimizes the fit of the shape between the data and the models. The algorithm chooses cluster shape models and assigns memberships of individual samples into particular clusters. A cluster is describes by density of multivariate normal distribution with a particular mean and covariance. For this purpose, the Expectation Maximization (EM) algorithm is used. This algorithm is applied to several clusters and with several sets covariance matrices of the clusters. The best model with certain cluster number was determined by the highest BIC value (Fraley and Raftery, 2002; Fraley and Raftery, 2006; Reimann et al., 2008; Templ et al., 2008).

Gan et al., (2007) divided model-based clustering algorithm into three main steps. First is initializing the EM algorithm using the partitions from model-based agglomerative hierarchical clustering. Then, the parameters are estimated using the EM algorithm. The last step is choosing the model and the number of clusters according to the BIC (Fraley and Raftery, 2002; Gan et al., 2007). R package from Fraley and Raftery (2002; 2006), mclust package, was employed to transformed data both for conventional and CoDa approach.

3.2.2 Partition Around Medoids (PAM) clustering

The aim of PAM algorithm is to minimize sum average distances to the cluster medians. These medians are representative objects which represent the structure of the data. These medians are so-called medoids of the cluster. The first step in the algorithm is to set number of medoids (k) then k clusters are constructed by assigning each object of the dataset to the nearest medoids. The nearest criterion is determined based on Euclidean distance or Manhattan distance. In this research, Euclidean distance was employed.

According to Kaufman and Rousseeuw (2005), the algorithm of PAM clustering is as follow. Let set of objects is denoted as X = $\{x_1, x_2, ..., x_n\}$ and the dissimilarity between objects x_i and x_j denoted by d(i,j).The algorithm consist two steps. First is selecting of objects as medoids in cluster: y_i is defined as binary variable (1 or 0). The value of y_i will equal to 1 if the object xi (i=1,2,..., n) is selected as a medoids. Second step is to assign each object x to one of the selected medoid. The value of z_{ij} is also binary value (0 or 1).The z_{ij} has value of 1 if only if the object x_j is assigned to cluster of which x_i is the medoid. The PAM clustering was performed using cluster package in R statistic software (Maechler, 2005).

3.3. Post clustering

For the post clustering stage, reclassification and filtering were applied to images of clustering results. Interpolation and rasterization are the next steps after clustering in order to assess the quality of classification. Thiessen polygon was performed to interpolate unsample areas. Then polygons were converted to raster image with particular grid size. The grid size was determined by using Equation 2-18 proposed by Hengl (2006).

$$p = 0.25\sqrt{\frac{A}{N}} \tag{3}$$

where A is study area in m2 and N is the total number of observations/samples. The formula is suitable for random or clustered distribution sample points such as stream sediment sample points.

Reclassification using majority rules, the same that developed by Lang et al. (2008) for labelling the classification results, was applied to clustering results in order to make number of cluster the same as number of reference classes (the existing lithological map). The images were reclassified into five classes representing five features as in the reference maps (figure 3). First, all clusters were assigned into category which the highest pixel number of the cluster lies on. If there are two or clusters having the same categories, the cluster with the highest pixel numbers among them was selected as 'key' cluster. If the same cluster was selected as key cluster, the second highest was taken as key cluster and so on. In the case where there is no key cluster in category, a cluster with the highest pixel number within that category was assigned as key cluster. The final step is joining non-key cluster into key cluster with the same category.

Majority filtering was applied to the images of clustering results as a suggested filtering technique by Lillesand and Kiefer (2000) to "smooth" classified data. The purpose of clustering is to eliminate a single or small cluster; thus, the final results only show the dominant cluster that presumably is the correct classification. The filtering employs a moving window of 3x3 pixels. The window is moved over the image like a moving kernel. At every position, the filtering will change the identity of centre pixel in a moving window to majority class when its class is not the majority one. In the case where there is no majority class; the identity of the centre pixel still remains.

3.4. Assessments

Two lithological maps, figure 3(a) and figure 3(b), were used to validate the images resulting from the clustering analyses. The map in figure 3(a), the existing lithological map, was designated as Reference Data I, whereas the map in figure 3(b), derived from the interpretation of airborne magnetic data as explained by Nugraha et al. (2013), was designated as Reference Data II. This latter reference data were used to test if interpreted lithological boundaries from analysis of airborne magnetic are consistent with results of clustering the SSG and AGR data together. If assessment results using Reference Data II are better than those using Reference Data I, it means that the lithological boundaries interpreted from the airborne magnetic data might be better than those portrayed in the existing lithological map. Furthermore, overall and producer's accuracy from error matrix and kappa coefficients were used to assess the quality of the classification results.

3.4.1 Error matrix

Error matrix is one of the most common ways to express accuracy of the classification. Error matrix shows relationship between known reference data and the corresponding cluster/class of clustering results. An existing lithological map was used as the reference data for the research. The matrices have the same rows and columns as the numbers of categories as lithological features in existing lithological map (Lillesand and Kiefer, 2000).

In an error matrix, two type of accuracy, overall and producer's accuracy, were taken to assess the accuracy of the clustering. Overall accuracy is a ratio of total pixels number correctly classified to total pixels number, whereas producer's accuracy is a ratio between the numbers of correctly classified pixels by the total pixel number in each category. The values describe how well the pixels are classified using clustering method.

3.4.2 Kappa coefficient

Kappa coefficient could be used to indicate the level of the percentage correct values in an error matrix caused by true agreement or only chance agreement. Equation 4 is used to calculated kappa coefficient:

$$k = \frac{n\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+} - x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+} - x_{+i})} \tag{4}$$

where

r = number of rows in the error matrix

x_{ii} = the number of observation in row i and column i (on the major diagonal)

x_{i+} = total of observations in row i (shown as marginal total to right of matrix)

x_{+i} = total of observations in row i (shown as marginal total at bottom of matrix)

N = total number of observations included in matrix.

The kappa coefficient ranges from 0 to 1. One indicates true agreements and zero indicates chance agreements.

4. RESULTS AND DISCUSSION

4.1. Interpolated images

Figure 4 and table 1, respectively, show the variogram model for Zn, as an example, and summary of variogram components for 13 elements in the SSG dataset. A model was chosen based on the smallest root mean square values from three variogram models which had been tried, exponential, gaussian and spherical. The variogram model show spatial data structure of the element at the study area. All the elements data fit with the exponential model that means the data have high variance change within small distance. The range values, the last column of table 1, show maximum distance at which the samples still have spatial correlation. Therefore, the models are reliable to be used due to the fact the smallest distance between the samples at the study area are less than 500 m. In addition, figure 5 shows the spatial distribution of Zn as a result of universal kriging interpolation.

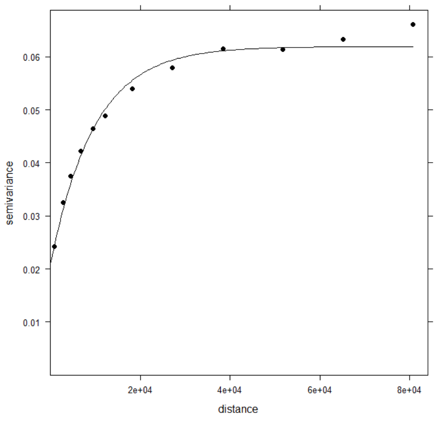


Figure 4. Exponential variogram model for logarithmic base-10 transformed Zn data.

Table 1. Summary of variogram components of individual elements in the SSG dataset.

Element	Model	Nugget	Partial Sill	Range
Zn	Exponential	0.02	0.04	9,791
Cu	Exponential	0.02	0.10	13,795
Pb	Exponential	0.05	0.09	15,692
Ni	Exponential	0.03	0.16	15,181
Co	Exponential	0.02	0.05	8,623
Ba	Exponential	0.03	0.04	8,988
Mn	Exponential	0.04	0.05	5,706
Fe	Exponential	0.02	0.03	9,188
Ce	Exponential	0.03	0.07	23,845
Cr	Exponential	0.05	0.15	14,717
Hg	Exponential	0.04	0.09	18,299
Sb	Exponential	0.03	0.09	14,135
As	Exponential	0.06	0.15	12,860

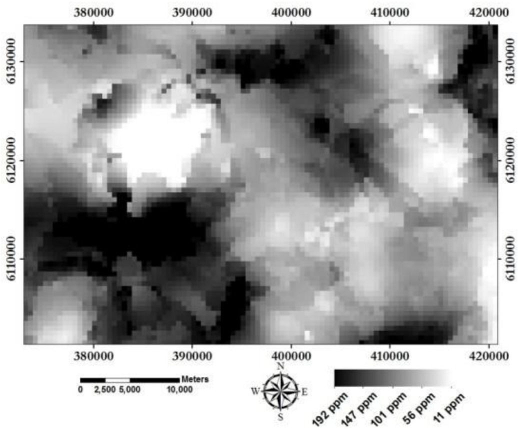


Figure 5. Image of spatial distribution of Zn as a result of universal kriging.

4.2. Clustered images

Figure 6(a) is an image of Mclust results using SSG and AGR data together (Integrated Data I). The best model for the data base on BIC is an EEV-model with nine clusters. This model means that the nine clusters have equal volumes and shapes but vary in orientations of ellipsoidal distributions in feature space. In addition, the image depicts several homogenous clusters with clear separation between them.

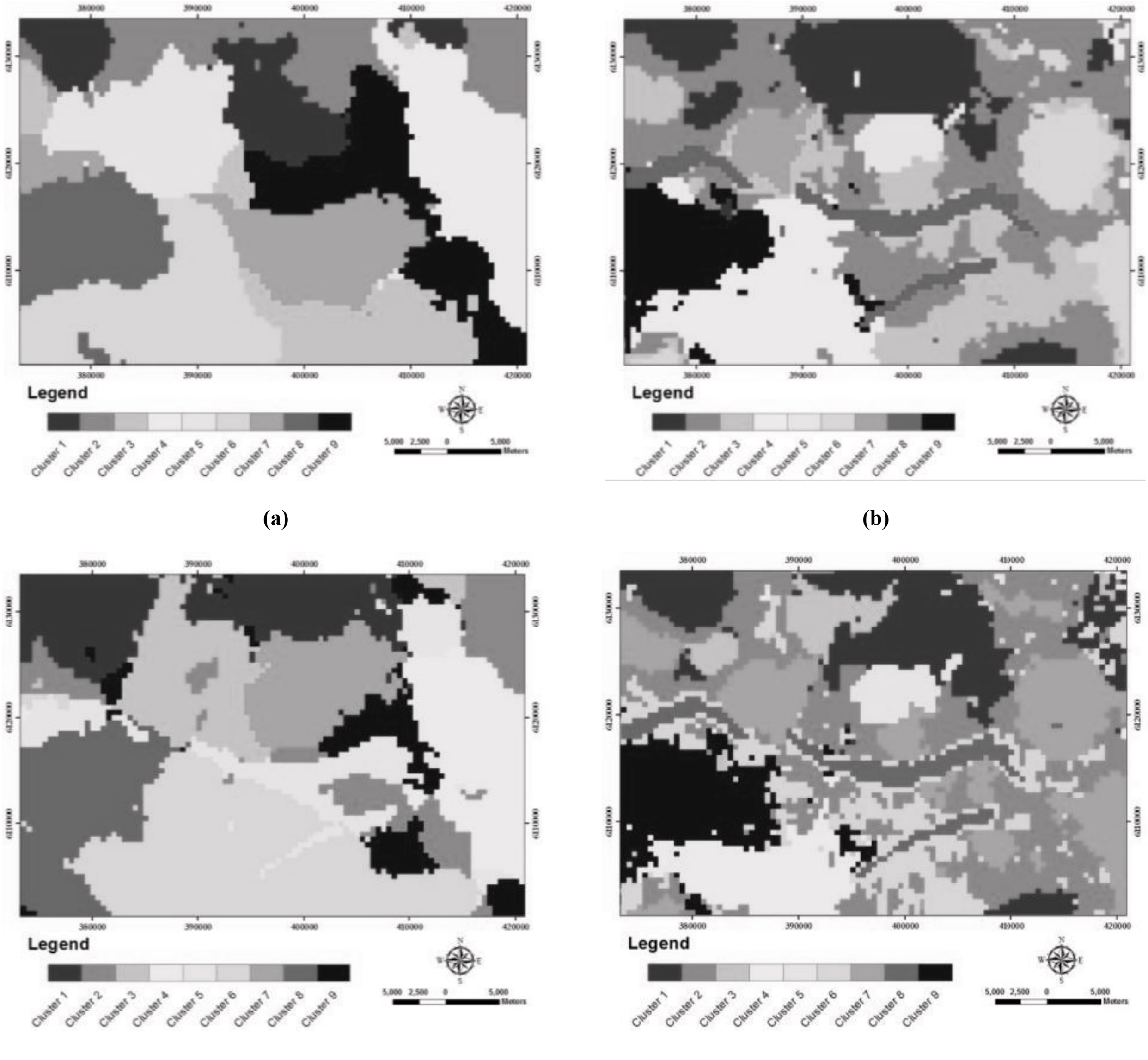
For PAM clustering, the optimum cluster number for Mclust for the same data was taken as an input. It is because values of SC (Silhouette Coefficient) in PAM clustering for cluster number (k) from 3 to 15 did not show an optimum cluster number but rather fluctuate at a constant range between 1.3 to 1.5. According to Kaufman and Rousseeuw (2005) when the SC value is below 0.25, it means that there is no substantial structure in the data. When there is no substantial in the data, the data using this technique become less reliable to be used to explain the processes/phenomena. Furthermore, figure 6(b), the result of PAM clustering using Integrated Data, shows several clusters with clear boundaries among them and several individual clusters within them. However, the PAM clustering image shows that lake boundaries could be detected quite well as can be seen in figure 6(b).

Figure 6(c) and figure 6(d) are the images resulting from clustering using the nine selected elements in the SSG data and the AGR data (Integrated Data II). These images show similar patterns as their corresponding resulting from Integrated Data I. For Mclust, the same model as for Integrated Data I, an EEV-model with nine clusters, was also obtained. In addition, PAM clustering used the same cluster number from the Mclust result as an input in the process. The number was taken due to the difficulties in determining an optimum cluster number based on SC.

Comparing images resulting from Mclust, figure 6(a) and figure 5-5(c), with those resulting from PAM clustering, figure 6(b) and figure 6(d), reveals that the clusters of Mclust are larger and more homogeneous than those of PAM clusters. Furthermore, it is apparent that Mclust is a better technique than PAM clustering in distinguishing large features such as both sedimentary rocks. However, PAM clustering could detect small features, such as lakes, better.

Figure 7 and figure 8, which are results of reclassification and filtering, demonstrate that patterns of particular clusters are similar to patterns of particular feature in reference data. Both figures have similar patterns, except for Intrusive Rocks. It is due to the facts that figure 8 using Reference Data II as a basis for reclassification. As explained in previous section, the Reference data II data is a lithological map base on airborne magnetic data analysis which updating new boundary for Intrusive Rocks. Furthermore, for example, in figure 7(a), pattern of Cluster D is similar to that of the Sedimentary Rocks 2 in southwestern parts of the areas. This pattern is found also in figure 7(c), and figure 7(d), whereas for figure 7(b) the pattern is somewhat elongated to the east. Besides the pattern of Cluster D, in figure 7(a), the pattern of Cluster A is also similar to that of Intrusive Rocks in the existing lithological map. Moreover, among all images, figure 6 (a) has the clearest boundaries of lithological units, except for the lake features.

Like in figure 7, in figure 8, the sedimentary rocks unit could also be recognized. The same as figure 7(a), Mclust results for Integrated Data I, figure 8(a) could identify the Instrusive Rocks and give the most obvious boundaries among lithological units.



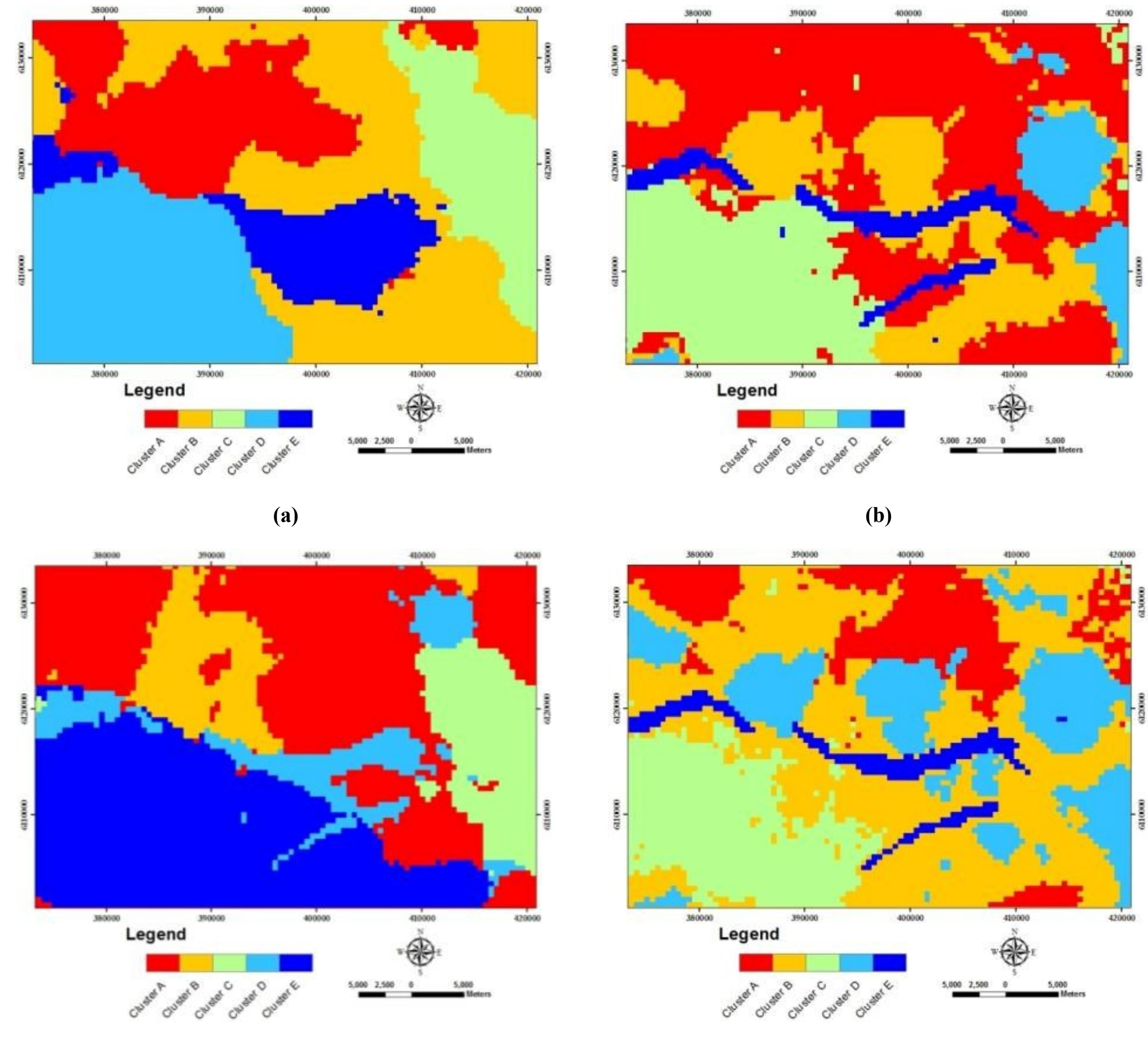
(a)

(b)

(c)

(d)

Figure 3. Clustering results image, (a) Mclust for all elements of stream sediment geochemical data and gamma-ray data integration, (b) PAM clustering for all element of stream sediment geochemical data and gamma-ray data integration, (c) Mclust for nine selected element of stream sediment geochemical and gamma-ray data integration, (d) PAM clustering for nine selected element of stream sediment geochemical and gamma-ray data integration.



(a)

(b)

(c)

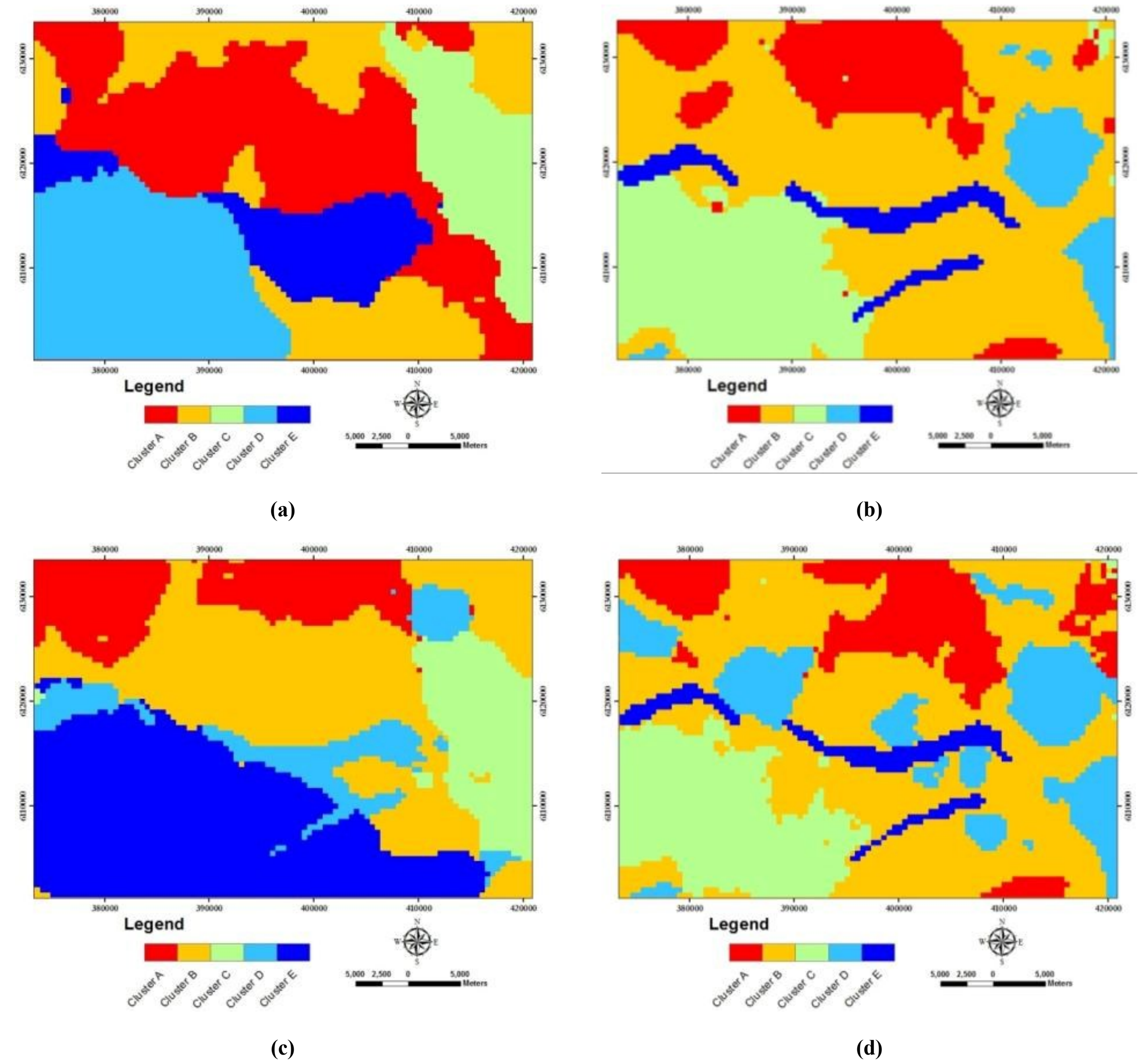
(d)

Figure 4. Images after reclassification and filtering using existing lithological map (Reference Data I), (a) Mclust for all element of stream sediment geochemical and gamma-ray data integration, (b) PAM clustering for all element of stream sediment geochemical and gamma-ray data integration, (c) Mclust for selected nine elements of stream sediment geochemical and gamma-ray data integration, (d) PAM clustering for nine selected element of stream sediment geochemical and gamma-ray data integration.

4.3. Quality of classification

Regarding producer’s accuracy, consistent results, in terms of differences among lithological units, are obtained as displayed in figure 9. For assessment using Reference I, the consistency could be observed such as for Features A, D, and E. The differences are roughly 10% between the highest and the lowest percentage. Producer’s accuracy for Feature A is roughly the same at 50-60% whereas producer’s accuracy for Features units D and E are higher at 70-80%. In addition, producer’s accuracy for Features B and C is variable depend on clustering type and data preparation.

Table 2 shows that assessment results vary from 51% to 62% and from 0.3 to 0.45 for overall accuracy and kappa coefficient, respectively. The lowest accuracy of 51% is obtained in PAM clustering of Integrated Data II with respect to both Reference Data I and II. The lowest accuracy, even is still in moderate level, might be due to its low SC value which describe that the data are less representative to depict the process in the area. The highest accuracy is 62% when Mclust clustering was applied into Integrated Data I with respect to Reference Data I as shown in table 2 (a). Similar to overall accuracy, kappa coefficient is worse when Integrated Data II was used with respect to both Reference Data I and Reference Data II. In addition, table 2 also shows that the assessments using Reference Data I are better than using Reference Data II.



(a)

(b)

(c)

(d)

Figure 8. Images after reclassification and filtering using lithological map based on the interpretation of airborne magnetic data (Reference Data I), (a) Mclust for all element of stream sediment geochemical and gamma-ray data integration, (b) PAM clustering for all element of stream sediment geochemical and gamma-ray data integration, (c) Mclust for selected nine elements of stream sediment geochemical and gamma-ray data integration, (d) PAM clustering for selected nine elements of stream sediment geochemical and gamma-ray data integration.

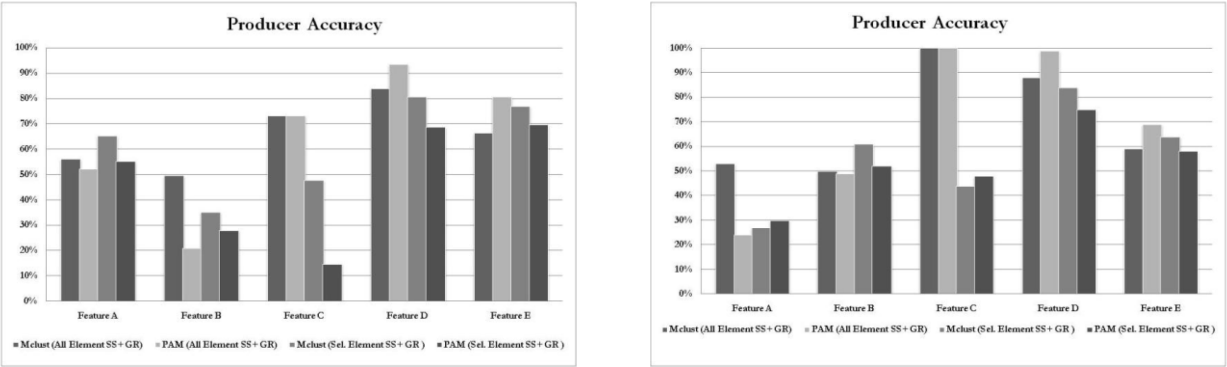
According to assessment results, Mclust and PAM clustering have similar accuracy values. However visually, Mclust images are better than PAM clustering images. It is because the cluster patterns from Mclust have similar general patterns as in the existing lithological map except for the lakes features. Thus, for the next research, it is suggested to integrate satellite imagery such as Landsat or Aster imagery in order to identify water bodies such as lakes features. In addition, Mclust result images have more homogeneous cluster areas than PAM clustering images. It might be because in PAM clustering an optimum cluster number was not achieved.

4.4. Comparison of individual data

Producer's accuracies of integrated data with respect to Reference Data I tend to be similar to the results of using AGR data. It might be due to the influence of sample density and regularity pattern of AGR data. AGR data have denser sampling density than SSG data, the former have density about 1 sample per 0.25 km^2 whereas the latter have density of 1 sample per 13 km^2. In addition, the pattern of the sample points may also influence accuracy in interpolation results. Furthermore, in interpolating SSG data, it was difficult to find appropriate variogram models when only using SSG data within the integrated/AGR study area; thus, data within the whole stream sediment study area were employed.

Overall accuracies of clustering the SSG and AGR data together are higher than these of SSG and AGR data. However, their better results are inconsistent with respect to producer's accuracy except for Mclust using combination of SSG all 13 elements with AGR data (Integrated Data I). In addition, when the accuracies of SSG data is compared to these of AGR data, Intrusive Rock could be well detected in SSG data whereas Volcanic Rocks could be well distinguished in AGR data. For other features, accuracy of both data shows almost the same results. Furthermore, producer's accuracies of Mclust using Integrated Data I have relative a higher or the same as these of SSG and AGR data. The selection elements of SSG data are not produce better results when they are combined with AGR data. It might be because structures in SSG data only are different from these in the SSG and AGR data together.

Accuracy of clustering results using the integrated data is dependent on spatial data structure. It is because values in the integrated datasets that were used inputs for clustering are interpolated data. The spatial data structure will influence the selection of appropriate interpolation technique. In the case when the data have no spatial structure, interpolation technique such as kriging is not an appropriate technique.



(a) **(b)**

Figure 5. Producer's accuracy diagram from assessment of clustering results, (a) an assessment using existing lithological map for the reference (Reference Data I), (b) an assessment using map of the interpretation of airbone magnetic data (Reference Data II); Feature A, Volcanic Rocks; Feature B, Intrusive Rocks; Feature C, Sedimentary Rocks 1; Feature D, Sedimentary Rocks; Feature E, Lake (for explanations about the lithology at the study area see section 1.6.2).

Table 1. Comparison of clustering results assessment based on stream sediment geochemical data only, airborne gamma-ray only, and integration of both datasets, (a)accuracy using combination of stream sediment geochemical (SSG) all 13 elements with airborne gamma-ray (AGR) elements (Integrated Data I), (b) accuracy using combination of SSG selected nine elements with AGR elements; *) an assessment using existing lithological map for the reference (Reference Data I), **) an assessment using map of the interpretation of airborne magnetic data (Reference Data II).

(a)

Assesement	Mclust				PAM			
	SSG	AGR	(SSG$^{1)}$+AGR)$^{*)}$	(SSG+AGR)$^{**)}$	SSG	AGR	SSG+AGR$^{*)}$	SSG+AGR$^{**)}$
Overall Accuracy	50%	50%	62%	59%	44%	45%	61%	57%
Kappa Coefficient	0.39	0.36	0.45	0.37	0.32	0.30	0.45	0.35

(b)

Assesement	Mclust				PAM			
	SSG	AGR	(SSG+AGR)$^{*)}$	(SSG$^{2)}$+AGR)$^{**)}$	SSG	AGR	(SSG+AGR)$^{*)}$	(SSG+AGR)$^{**)}$
Overall Accuracy	51%	50%	56%	54%	43%	45%	51%	51%
Kappa Coefficient	0.41	0.36	0.37	0.30	0.31	0.30	0.30	0.26

5. CONCLUSIONS

Mclust and PAM clustering could be alternatives techniques for classifying stream sediment geochemical data to help lithological mapping in area with limited information. The images of their results depict pattern similarities to the existing litholigical map. In addition, the assessments of the results show moderate accuracy up to 51% and 0.41 for overall accuracy and kappa coefficient, respectively. In addition, for a large homogeneous lithology, the producer's accuracy is quite high up to 80%.

In general, the application of CoDa approach in data preparation to both SSG and AGR data do not produce better accuracy than conventional approach. The assessments show the differences between the application of CoDa and conventional approach reach up to 10% and 0.1 for overall accuracy and kappa coefficient, respectively.

The integrated data of stream sediment geochemistry and airborne gamma-ray produce better results than those using stream sediment geochemical or airborne gamma-ray data separately. The percentage accuracy increments of integration data compare to their separated data are quite significant up to 17% and 0.15 for overall accuracy and kappa coefficient, respectively.

Mclust produces better classifications for lithological mapping relatively to PAM clustering base on both qualitative and quantitative assessments. Qualitatively, from visual evaluation, the patterns of Mclust results are more similar to lithological patterns in the existing lithological map than PAM clustering. Quantitatively, assessments results show up to 5% and 0.7 difference for overall accuracy and kappa coefficient, respectively, in each separated data (SSG or AGR) whereas for integrated data (SSG and AGR) produces non-significant difference in the assessments (1% and 0% differences for overall accuracy and kappa coefficient, respectively). Therefore, Mclust could be applied to integrate and classify SSG and AGR data for lithological mapping in regional scale.

REFERENCES

Bonham-Carter, G.F., Rogers, P.J., and Ellwood, D.J.: Catchment basin analysis applied to surficial geochemical data, Cobequid Highlands, Nova Scotia, *Journal of Geochemical Exploration*, **29**, (1987), 259-278.

Carranza, E.J.M.: Geochemical anomaly and mineral prospectivity mapping in GIS, Amsterdam, Elsevier, (2008), 366 p.

Carranza, E.J.M.: Mapping of anomalies in continuous and discrete fields of stream sediment geochemical landscapes, *Geochemistry-Exploration Environment Analysis*, **10**, (2010), 171-187.

Carranza, E.J.M., and Hale, M.,: A catchment basin approach to the analysis of reconnaissance geochemical-geological data from Albay Province, Philippines, *Journal of Geochemical Exploration*, **60**, (1997), 157-171.

Delong, R.C.: Geology, alteration, mineralization and metal zonation of the Mt. Milligan Porphyry Copper-gold Deposits, Vancouver, The University of British Columbia, (1996)

Filzmoser, P., Hron, K., and Reimann, C.: Principal component analysis for compositional data with outliers, *Environmetrics*, **20**, (2009a), 621-632.

Fraley, C., and Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, (2002), 611-631.

Fraley, C., and Raftery, A.E.: Mclust version 3 for r: normal mixture modeling and model-based clustering, *Technical Report No. 504*, Department of Statistics, University of Washington (http://cran.r-project.org/web/packages/mclust/index.html), (2006).

Gan, G., Ma, C., and Wu, J.: Data clustering: theory, algorithm, and applications, SIAM, Philadelphia, ASA, Alexandria, VA, (2007), 466 p.

Hengl, T.: Finding the right pixel size, *Computers & Geosciences*, **32**, (2006), 1283-1298.

Howarth, R.J.: Statistical applications in geochemical prospecting: a survey of recent developments, *Journal of Geochemical Exploration*, **21**,(1984), 41-61.

Jackaman, W., and Balfour, J.S.: QUEST project geochemistry: field surveys and data reanalysis, Central British Columbia (parts of NTS 093A, B, G, H, J, K, N, O), Geoscience BC Summary of Activities 2007; Geoscience BC, *Report 2008-1*, (2008), 150.

Kaufman, L., and Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis: New Jersey, John Wiley & Sons, Inc, (2005).

Lang, R., Shao, G., Pijanowski, B.C., and Farnsworth, R.L.: Optimizing unsupervised classifications of remotely sensed imagery with a data-assisted labeling approach, *Computers & Geosciences*, **34**, (2008), 1877-1885.

Lillesand, T.M., and Kiefer, R.W.: Remote sensing and image interpretation, New York, etc., John Wiley and Sons, (2000), 724 p.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M.: Cluster analysis basics and extensions, unpublished (http://cran.r-project.org/web/packages/cluster/index.html), (2005).

Massey, N.W.D., MacIntyre, D.G., Desjardins, P.J., and Cooney, R.T.,: Digital Geology Map of British Columbia - Tile NN10 Central B.C., GeoFile 2005-6, B.C. Ministry of Energy and Mines, (2005b).

Massey, N.W.D., MacIntyre, D.G., Desjardins, P.J., and Cooney, R.T.,: Digital Geology Map of British Columbia: Tile NO10 Northeast B.C, GeoFile 2005-10, B.C. Ministry of Energy and Mines, (2005b).

Massey, N.W.D., MacIntyre, D.G., Desjardins, P.J., and Cooney, R.T.,: Digital Geology Map of British Columbia:Tile NO9 North Central B.C., GeoFile 2005-9, B.C. Ministry of Energy and Mines, (2005c).

Massey, N.W.D., MacIntyre, D.G., Haggart, J.W., Desjardins, P.J., Wagner, C.L., and Cooney, R.T.: Digital Geology Map of British Columbia: Tile NN8-9 North Coast and Queen Charlotte Islands/Haida Gwaii, GeoFile 2005-5, B.C. Ministry of Energy and Mines, (2005d).

Naseem, S., Sheikh, S.A., Qadeeruddin, M., and Shirin, K.: Geochemical stream sediment survey in Winder Valley, Balochistan, Pakistan, *Journal of Geochemical Exploration*, **76**, (2002), 1-12.

Natural Resources Canada: Geoscience data repository, aeromagnetic and electromagnetic data, Natural Resources Canada (http://gdr.nrcan.gc.ca/aeromag/index_e.php), (2010a).

Natural Resources Canada: Geoscience data repository, radioactivity data, Natural Resources Canada (http://gdr.nrcan.gc.ca/gamma/index_e.php), (2010b).

Nelson, J., Bellefontaine, K., Rees, C., and MacLean, M.: Regional geological mapping in the Nation Lakes Area (93N/2E,7E), Geological Field Work 1991, *Paper 1992-1*, British Columbia Ministry of Energy, Mines and Petroleum Resources, (1992), 118.

Nelson, J., Bellefontaine, K.,Green, K., and MacLean, M.: Regional geological mapping near the Mount Milligan Deposit (93N/l,93k/16), Geological Fieldwork 1990, *Paper 1991-1*, British Columbia Ministry of Energy, Mines and Petroleum Resources (1991).

Nugraha, H. S., Carranza, E. J. M., and Van Der Meijde, M.: Airborne Magnetic Data for Lithologic Mapping using Edge Detection Method, *Proceedings*, 13th Indonesia International Geothermal Convention & Exhibition 2013, Jakarta, (2013).

Pawlowsky-Glahn, V., and Egozcue, J.J.: Compositional data and their analysis: an introduction, in Buccianti, A., Mateu-Figueras, G., and Pawlowsky-Glahn, V., eds., Compositional data analysis in the geosciences: from theory to practice: London, The Geological Society, (2006), 264.

Reimann, C., Filzmoser, P., Garrett, R., and Dutter, R.: Statistical data analysis explained: applied environmental statistics with R, West Sussex, John Wiley and Sons, (2008), 341 p.

Robinson, G.R., Kapo, K.E., and Grossman, J.N.: Chemistry of stream sediments and surface waters in New England, Volume 2010, *p. open-file report 2004-1026*, (2004).

Spadoni, M., Cavarretta, G., and Patera, A.: Cartographic techniques for mapping the geochemical data of stream sediments: the “sample catchment basin” approach, *Environmental Geology*, **45**, (2004).

Templ, M., Filzmoser, P., and Reimann, C.: Cluster analysis applied to regional geochemical data: problems and possibilities, *Applied Geochemistry*, **23**, (2008), 2198-2213.

Yusta, I., Velasco, F., and Herrero, J.-M.: Anomaly threshold estimation and data normalization using EDA statistics: application to lithogeochemical exploration in Lower Cretaceous Zn-Pb carbonate-hosted deposits, Northern Spain, *Applied Geochemistry*, **13**, (1998), 421-439.