# Modified Agglomerative Clustering by Klassen Disparities for Identification Hierarchical Cluster of Regional Developing Countries

Tb. Ai Munandar
Computer Sc. and Electronic Dept., Math. And Natural Science Faculty, Universitas Gadjah Mada, Yogyakarta, Indonesia,

Azhari Azhari
Computer Sc. and Electronic Dept., Math. And Natural Science Faculty, Universitas Gadjah Mada, Yogyakarta, Indonesia,

Aina Musdholifah
Computer Sc. and Electronic Dept., Math. And Natural Science Faculty, Universitas Gadjah Mada, Yogyakarta, Indonesia,

Lincolin Arsyad
Economics Department, Faculty of Economics and Business Universitas Gadjah Mada Yogyakarta, INDONESIA

## ABSTRACT
Inequality of regional development is a global problem and faced by many countries, including Indonesia. Various attempts were made to reduce inequality in the region, one of them is by analyzing the imbalance with appropriate methods that can be used as a basis for policy making prioritization of future development. Klassen methods typically used to analyze the inequality of the region according to the indicators Gross Regional Domestic Product (GRDP). However, the division of the region inequality using Klassen deemed too rigid, given the existence of a possible relationship between the regions and in each of the groups formed by Klassen. This research aims to develop a new approach that can be used to analyze the inequality of development of the region. Aggromerative cluster hierarchical cluster technique modified with Klassen named Modified Agglomerative Hierarchical Clustering with Klassen (MHACK). The results shows that the use of algorithms MHACK, besides being able to classify the area into four main clusters, are also capable of forming the new group hierarchy for each region in each of the main cluster. Cophenet distance coefficient showed that MHACK algorithm has 0.9950 for Quadrant I, and 0.9154 for Quadrant II. In addition, the city of Magelang is indicated as an advanced and rapidly growing region with a poor value of GRDP, while Cilacap, Kudus, Boyolali, Brebes and Wonogiri indicated as a potential and growing region but has the worst value of GRDP.

## General Terms
Data Mining, Decision Support Systems.

## Keywords
inequality of development, GDP, Klassen, agglomerative hierarchical clustering, MHACK.

## 1. INTRODUCTION
Inequality of regional development is a process of national development experienced by each country [1], including Indonesia. Metwally and Jensen in [2] states that inequality is closely related to regional development regional income inequality observed region against region income peers (national). In this case, the regional gross domestic income (GDP) is often used as an indicator in determining regional development imbalances. There are many imbalances analytical techniques used, one of them is the typology Klassen. This technique is used to see patterns and structure of economic development of a region and then divide it into four quadrants. Quadrant I is the area developed and grew rapidly; Quadrant II is advanced but depressed; Quadrant III is a potential area or they may develop; and Quadrant IV is relatively underdeveloped regions [3], [4].

Klassen is usually combined with other techniques such as Location Quotient [4] and Williamson Index [5]. Results of these combinations to form a hierarchical regional development imbalances. Topmost hierarchy typically shows groups of regions based on certain inequality under the provisions Klassen, then each group from the Klassen formation, forming hierarchical other form of information inequality index when using Williamson Index; the potential of the sector and the group hierarchical area if using Location Quotient. Analysis of development gaps using Williamson index and Location Quotient more focused on differences in the achievement levels of the economy of a region against region comparison. Grouping is done very firmly based on data from the economic achievements of the region. Notwithstanding the nature of information that is owned by a data.

Naturally the data has information that could be used for grouping data efficiently based on similarity and disimilarity [6]. This study rests on the assumption that the data GDP also have a natural information that can be used to support performance analysis of inequality of regional development by grouping them into specific groups based on similarity or disimilarity. The study also simultaneously aims to modify one cluster technique, namely hieararchical agglomerative clustering (HAC) for the identification of the regional development imbalance into a hierarchical model, where the sector GDP is used as the data being analyzed.

The use of HAC has been done by some researchers to extract new knowledge from the data into the model dendrogram tiered (hierarchical). Some of these studies, among others, the determination of educational curriculum relevant to industry [7], the grouping of medical documents to find information about the patient and his medical condition [8], the data analysis bio-medical [9], identification of user session on a web application [10], extract knowledge from text-based documents after reducing the dimensions of the data [11] and the behavior of the stock market trend analysis [12]. In

addition, for the needs analysis development gaps, algorithms HAC has been used in several EU countries [13], Germany [14], Romania [15], [16], Ukraine [17], as well as the evaluation inequality living standards of the region in the Czech Republic [18].

The discussion paper is divided into seven sections. The first section describes the background of the problems of the research conducted. The second and third part discuss the theoretical basis used in the study. The fourth section discusses the proposed methods HAC modified method Klassen typology. The fifth section contains the stages of research undertaken. The sixth section discusses the study and discussion of the results of the seventh part is the overall conclusion of the study.

## 2. HIERARCHICAL AGGLOMERATIVE CLUSTERING (HAC)

HAC is an algorithm for grouping of data which form the cluster results in the form of a dendrogram graphics visualization. Dendrogram represents the nested groups that form either at the same level at the time of the grouping, or even at different levels [19]. HAC cluster technique is largely a variant of the single-link and complete-link. Single-link group the two clusters into a single cluster based on its minimum distance while the complete-link is the opposite of single-link. HAC algorithm is shown in Figure 1.

**Hierarchical Agglomertaive Clustering (HAC)**

1. **Input :**
   $E = \{e_1,e_2,...,e_n\};$ (set of data objects)
2. **Output :**
   $C = \{c_1,c_2,...,c_n\};$ (set of cluster)
3. **For** $\{e_1,e_2\} \mid e_i \in E,\ 1 \le i \le n$ **do**
   (calculate the distances between data objects, eg using Euclidean distance)
   $D(e_1,e_2) \leftarrow sqrt(sqr(sum(e_1 - e_2)));$
4. **End for;**
5. *Determine the proximity matrix based on the distance D for all of the set E;*
6. *Determine the set of clusters based on singleton clusters, where each cluster represents a set of input E;*
7. **Repeat**
   (combine two nearest cluster, eg using single linkage)
   $d(e_1,e_2) \leftarrow min(D(e_1,e_2));$
   *Update the proximity matrix with the new D distance between the new cluster is formed with the original cluster E;*
8. **Until** $distanceD(e_1,e_2) = 1;$

**Fig 1: Algorithm of HAC**

Typology of Klassen divide a region into four quadrant based on regional economic growth and per capita income. Quadrant division Klassen development gaps as shown in Table 1.

**Table 1. Klassen Quadrant**

| | |
|---|---|
| **Quadrant I (K1)** (developed region) $r_i >= r$ dan $y_i >= y$ | **Quadrant II (K2)** (stagnant region) $r_i < r$ dan $y_i >= y$ |
| **Quadrant III (K3)** (developing region) $r_i >= r$ dan $y_i < y$ | **Quadrant IV (K4)** (underdeveloped region) $r_i < r$ dan $y_i < y$ |

Where, $r_i$ is the rate of economic growth in the Regency, r is the rate of economic growth in the Province, $y_i$ is the contributions district development, and y is the contribution of Provincedevelopment.

Economic growth rate can be calculated by the formulation as shown in equation (1), while the contribution of the development of a sector is calculated by equation (2).

$$r = \frac{P_t - P_{t-1}}{P_{t-1}} x100\% \quad (1)$$

$$y = \frac{P_t + P_{t-1}}{T_t + T_{t-1}} x100\% \quad (2)$$

Where, $P_t$ is the total value of GRDP of all indicators in the current year, $T_{t-1}$ is the total value of GDP throughout the previous year indicator, $P_t$ is the current year GRDP sector and $P_{t-1}$ is sector GRDP the previous year.

## 3. RESEARCH METHODOLOGY

Research conducted an effort to develop a new approach to the needs of regional development imbalances identification by using cluster technique and Klassen. The study begins with a study of literature related to the concept of regional development imbalances identified by some researchers either by using cluster technique and Klassen. Results of the study of literature is then used as a reference to modify the agglomerative hierarchical clustering techniques with Klassen, later called the Modified Hierarchical Clustering Agglomerative with Klassen (MHACK). MHACK then tested a new approach to the region GRDP data to identify development gaps that occur. GDP data used are the data of GRDP of 33 districts in Central Java province in 2012 and 2013.

## 4. DETAILS OF THE PROPOSE METHOD

This research is a complement several previous studies which is related to the identification of regional development imbalances. The proposed method is based on the consideration of the possibility of grouping with Klassen results can be grouped by similarity or disimilarity re-attached to the data development. Thus the grouping can be made deeper to see the closeness and the relationship between one region to another within the group. Proposed method in this study is a modified hierarchical clustering techniques with methods Klassen then called modified agglomerative hierarchical clustering with Klassen (MHACK). There are two stages that apply to MHACK before generating the output cluster as a whole. The first stage is the process of grouping data using Klassen regional development. This phase will classify the region into four main groups (major cluster). The second stage is the grouping of regional data on each major cluster with HAC. MHACK algorithm as shown in Figure 2.

The following is a description that applies to both phases of MHACK:

**Phase I:** grouping data Klassen regional development by forming four main clusters. At this stage, the method Klassen inserted into the HAC algorithm. The data used is the gross regional domestic product (GRDP) of each region will be grouped. The results from this stage in the form of four groups inequality of development of the region as the main cluster. In practice, the four groups do not have to always be formed, depending on the rules applicable Klassen in GRDP sector

data is entered. Input and output in the first phase is as follows:

Input :
E = {P,Q,R,S};
P = {$p_1$,$p_2$,...,$p_n$};
Q = {$q_1$,$q_2$,...,$q_n$};
R = {$r_1$,$r_2$,...,$r_n$};
S = {$s_1$,$s_2$,...,$s_n$};

Where E = {P, Q, R, S} is a set of data sector GRDP, P = {$p_1$, $p_2$, ..., $p_n$} is the set of sector data GDRP the current year, Q = {$q_1$, $q_2$, .. ., $q_n$} is the set of sector data GDRP previous year, R = {$r_1$,$r_2$,...,$r_n$} is a set of sector data in the current provincial GRDP, and S = {$s_1$,$s_2$,...,$s_n$} is the set of sector data Provincial GRDP a year earlier.

Output :
DK = {$e_1$,$e_2$,...,$e_n$};
DP = {$f_1$,$f_2$,...,$f_n$};
CK = {$g_1$,$g_2$,...,$g_n$};
CP = {$h_1$,$h_2$,...,$h_n$};
K = {E(i)) | i = 1,2,...,n};
Klabel = {Quadrant I, Quadrant II, Quadrant III, Quadrant IV};

Where, DK = {$e_1$,$e_2$,...,$e_n$} is the growthrate of the construction of a district, DP = {$f_1$,$f_2$,...,$f_n$} is a district development contributions, CK = {$g_1$,$g_2$,...,$g_n$} is the rate of growth of the construction of a province, CP = {$h_1$,$h_2$,...,$h_n$} is a Provincial development Contributions, K = {E (i)) | i = 1,2, ..., n} is quadrants development area has been grouped with Klassen, Klabel = {Quadrant I, Quadrant II, Quadrant III Quadrant IV} is the main cluster label. A key step in the first phase is described as follows:

***Step 1:*** calculate the value of the growth rate and development contribute to the region being analyzed and the reference region. Calculation of the value of the growth rate using equation (1), while the contribution of the construction is obtained by equation (2).

***Step 2:*** classify the region by comparing the rate of growth and the contribution derived from step 1. The grouping of data is done using rules Klassen region in Table 1.

**Phase II:** grouping the data region on each main cluster. Phase II consists of the following steps:

***Step 1:*** The input to this stage is the output of Phase I which sets K containing members of the set region.

***Step 2:*** Calculate the distance data at each of the major cluster on the set K. The calculation of distances for example performed using Eucledian Distance.

***Step 3:*** For each data in on each set of K, determine the set of singleton cluster $p_i \in$ P.

***Step 4:*** For each singleton cluster $p_i \in$ P in each set K, combine two singleton cluster.

***Step 5:*** update new distance between the clusters formed by the original cluster.

***Step 6:*** for every singleton cluster $p_i \in$ P in each set K, delete $p_1$ and $p_2$ of K, then add {$p_1$, $p_2$} into K.

***Step 7:*** Repeat steps 4-6 until no clusters can be grouped again in each set K.

**Modified HAC-Klassen (MHACK)**

**Input** :
*E = {P,Q,R,S};*
*P = {$p_1$,$p_2$,...,$p_n$};*
*Q = {$q_1$,$q_2$,...,$q_n$};*
*R = {$r_1$,$r_2$,...,$r_n$};*
*S = {$s_1$,$s_2$,...,$s_n$};*

**Output** :
*DK = {$e_1$,$e_2$,...,$e_n$};*
*DP = {$f_1$,$f_2$,...,$f_n$};*
*CK = {$g_1$,$g_2$,...,$g_n$};*
*CP = {$h_1$,$h_2$,...,$h_n$};*
*K = {C($E_{(i)}$)| i = 1,2,...,n};*
*Klabel = {Kuadran I, Kuadran II, Kuadran III, Kuadran IV};*

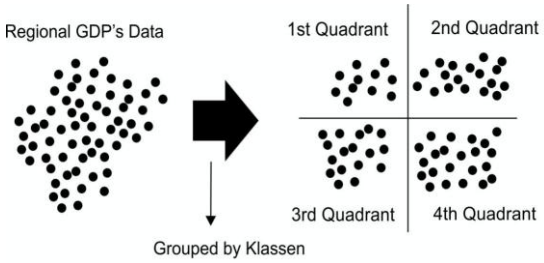
// Calculating the rate of growth and development contribution (Klassen typology)

**Foreach***input $p_i \in P$ ,$q_i \in Q$ ,$r_i \in R$,$s_i \in S$; $1 \le i \le n$* **do**
*DK($p_i$,$q_i$) ←sqrt(sqr(sum($p_i - q_i$)));*
*DP($r_i$,$s_i$) ←sqrt(sqr(sum($r_i - s_i$)));*
*CK($p_i$,$q_i$) ←sqrt(sqr(sum($p_i - q_i$)));*
*CP($r_i$,$s_i$) ←sqrt(sqr(sum($r_i - s_i$)));*

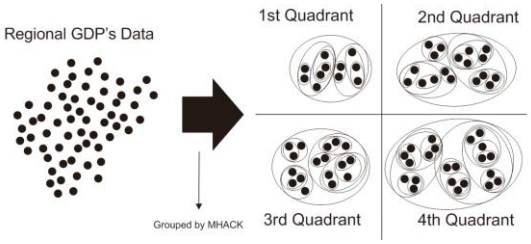
//Form a cluster of four main uses Klassen
**Foreach***$e_i \in DK$, $f_i \in DP$, $g_i \in CK$, $h_i \in CP$; $1 \le i \le n$* **do**
**If** *($e_i \ge f_i$) and ($g_i \ge h_i$)* **then**
*K ← {Ei};*
*Klabel ←Quadrant I;*
**Elseif***($e_i < f_i$) and ($g_i \ge h_i$)* **then**
*K ←{Ei};*
*Klabel ←Quadrant II;*
**Elseif***($e_i \ge f_i$) and ($g_i < h_i$)* **then**
*K ←{Ei};*
*Klabel ←Quadrant III;*
**Else**
*K ←{Ei};*
*Klabel ←Quadrant IV;*
**Endif**
**End**
*Calculate the distance matrix D for each object that already grouped by Klassen, for all $p_i \in$P on K;*
*Determine the set of cluster by cluster singleton, where each set of cluster represents every $p_i \in$P on K;*
// Merger singleton cluster on each of the main cluster
*t ← 0;*
**Repeat**
*t ← t+1;*
*Combine two single cluster $p_i \in$Pfrom the set of cluster K,*
*Update the proximity matrix with the new distance between the new cluster is*
*formed with the original cluster.*
*delete $p_1$ and $p_2$ of K*
*add {$p_1$, $p_2$} in K*
**Until** *cluster K = 1;*
**End**

**Fig 2: Algorithm of MHACK**

High-level visualization of HACK as shown in Figure 3. Figure 3a shows the process of grouping data sector GDP of a region using Klassen. Grouping the results are unequivocal, it means that the data meet the rules as in Table 1 will be grouped into four groups that have been determined. In Figure 3b, the result of a grouping which has been carried out using the method Klassen, deeper grouped to form a new group that shows the relationship between the region with other regions in each group formed by Klassen. Grouping in Figure 3b using a new approach that is MHACK.




**a. Regional disparities by Klassen**




**b. Regional disparities clustering by MHACK**

**Fig 3: Visualization of Klassen (a) and MHACK (b) regional disparities grouping**

To find out the needs of the execution time, the time complexity analysis performed using Big-O notation. Results of the analysis showed that MHACK algorithm has asymptotic time complexity of O ($n^2$).

## 5. RESULTS AND DISCUSSION

Thirty-one districts in Central Java province tested using algorithms MHACK and form two main clusters that describe the grouping information into the territory of a particular group inequality. Two main clusters formed is Quadrants I and II. Figure 4 shows the results of Quadrants Iclassification. Figure 5shows the results of Quadrants II classification. As mentioned before, the first stage MHACK grouping algorithm is forming a major cluster using Klassen. The next step is grouping the area carried out on each of the main cluster to form a group in hierarchical.

Results of grouping uses MHACK shows that Demak, Jepara and Kota Magelang grouped in Quadrant I. This means that a third of this region is a developed and grew rapidly area. In addition, the group formed a hierarchy in Figure 4 for Quadrant I shows that, although the three regions that are advanced areas, but all three still can be classified based on the value of its GDP. At Quadrant I, it appears that the District 1 (Demak) and 2 (Jepara) have a proximity that combined into one new group. The incorporation of the District 1 and 2 are then recombined with the District 3 (Kota Magelang) to form a new group. Analysis of the data of GRDP owned District 1 and 2 shows that these two regions have the highest GDP adjacent values. While District 3 has a value of GRDP is lower than both. In other words, the hierarchy bottom of the dendrogram that were formed then provide information about the area strata group advanced to the status of the best GRDP, while the topmost hierarchy indicates that the region is an advanced region to the status of the worst GRDP value.

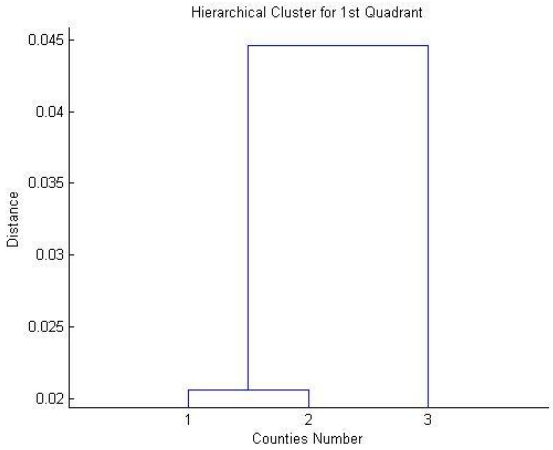



**Fig 4. Hierarchical cluster for 1st Quadrant**

For the thirty other regions, GRDP owned relationship can be seen in Figure 5 which shows a group hierarchical in Quadrant II. Results of the analysis shows that the Banyumas; Blora; Magelang; Pekalongan; Pemalang; Purbalingga; Rembang; Semarang; Temanggung; Wonosobo; Klaten; Grobogan; Tegal; Kota Tegal; Banjarnegara; Kota Surakarta; Boyolali and Pati; hierarchical strata are at the best value of GDP. Meanwhile, Cilacap district; Kudus; Boyolali; Brebesand Wonogiri stratum value of GDP is at worst. This can be seen from the fifth region located at the top level of hierarchical compared to other regions.

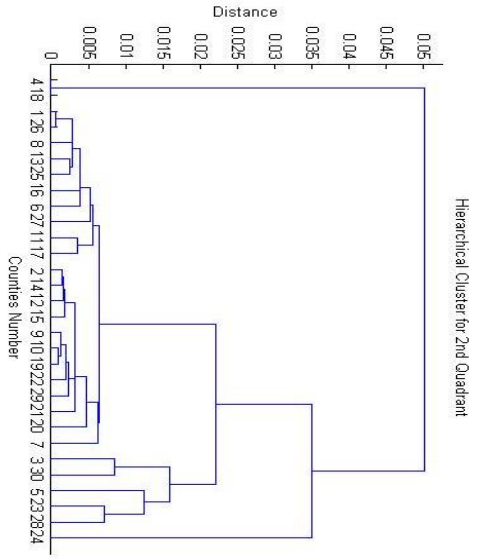



**Fig 5: Hierarchical cluster for 2nd Quadrant**

Hierarchical cluster testing is done by calculating the correlation coefficient cophenetic, for each cluster formed in each of the main cluster. Cophenet calculation results show that hierarchical cluster in Quadrant I has a value of 0.9950, while for the third quadrant of 0.9154. Cophenet value is obtained based on the distance cophenet of the tree that is formed and the object distance data used to form the tree itself. Cophenet value range should be close to 1 to indicate the quality of the results of the cluster.

GRDP data grouping districts with HACK algorithm can determine the hierarchy of groups from each district. This hierarchy shows the close relationship the district development results according to the GRDP indicator owned.

Relationships that are formed can be used as a reference for decision makers to determine define future development priorities.

## 6. CONCLUSION

The results showed that the grouping of regional development imbalances using an algorithm MHACK not only able to divide the region into a particular quadrant, but also can show the relationship between the region with other regions in each of the main cluster. The relationship shown hierarchical group based on the shape of the value of GRDP which is owned by each region. Two districts, Demak and Jepara is an advanced and rapidly growing region with the best value of GRDP, while Kota Magelang is an advanced and rapidly growing region with the worst value of GRDP. This was shown by the level of the hierarchy that is formed through MHACK. Region to establish the position of the top hierarchy (hierarchy outermost) is the region with the worst value of GRDP. Other cities fit into groups and developing potential areas, where Cilacap, Kudus, Boyolali, Brebes and Wonogiri are the strata worst value of GRDP compared to other regions in the same quadrant. Furthermore, the results of grouping the region inequality MHACK algorithm can be used by policy makers to determine the priority areas of development by looking at the hierarchy level of detail in each quadrant inequality of development.

## 7. REFERENCES

[1] Williamson, J.G., 1965, "Regional Inequality and the Process of National Development: A Description of the Patterns, Source: Economic Development and Cultural Change", Vol. 13, No. 4, Part 2 (Jul., 1965), pp. 184

[2] Akita, T., 2003, "Decomposing regional income inequality in China and Indonesia using two-stage nested Theil decomposition method", The Annals of Regional Science, Vol. 37, pp. 55–77

[3] Kuncoro, M., and Idris, A.N., 2010, "Mengapa Terjadi Growth Without Development Di Provinsi Kalimantan Timur", Jurnal Ekonomi Pembangunan, Volume 11, No. 2, Hal. 172 - 190, Desember

[4] Badrudin, R., 2012, "Pengembangan Ekonomi Lokal Kabupaten/Kota Provinsi Daerah Istimewa Yogyakarta Menggunakan Tipologi Klassen dan Location Quotient", Jurnal JRMB, Vol. 7, No. 1, Juni

[5] Barika, 2012, "Analisis Ketimpangan Pembangunan Wilayah Kabupaten/Kota Di Provinsi Bengkulu Tahun 2005 – 2009", Jurnal Ekonomi Dan Perencanaan Pembangunan, Volume :04 No. 03, Januari-Juni 2012

[6] Witten H, Ian dan Frank, Eibe, 2005, "Data Mining : Practical Machine Learning Tools and Techniques", United Kingdom (UK) : ELSEVIER.

[7] Sembiring, R.W., Zain, J.M., and Embong, A., 2010, "Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course", Journal Of Computing, Volume 2, Issue 12.

[8] Saad, F.H., Mohamed, O.I.E., and Al-Qutaish, R.E., 2012, "Comparison Of Hierarchical Agglomerative Algorithms For Clustering Medical Documents", International Journal of Software Engineering and Applications (IJSEA), Vol.3, No.3.

[9] Krishnaiah, V.V. J.R., Sekar, D.V.C., and Rao, K.R.H., 2012, "Data Analysis of Bio-Medical Data Mining using Enhanced Hierarchical Agglomerative Clustering", International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 3.

[10] Murray, G.C., Lin, J., and Chowdhury, A., 2006, "Identification of User Sessions with Hierarchical Agglomerative Clustering", Proceedings of the 2006 Annual Meeting of the American Society for Information Science and Technology, Texas.

[11] Patidar, G., Singh, A., and Singh, D., 2013, "An Approach for Document Clustering using Agglomerative Clustering and Hebbian-type Neural Network", International Journal of Computer Applications, Volume 75– No.9

[12] Marinova-Boncheva, V., 2008, "Using The Agglomerative Method of Hierarchical Clustering As A Data Mining Tool In Capital Market", International Journal "Information Theories & Applications" Vol.15.

[13] Poledníková, E., 2014, "Comparing Regions´ Ranking by MCDM methods: the Case of Visegrad Countries". WSEAS TRANSACTIONS on BUSINESS and ECONOMICS, Volume 11, 2014, pp. 496 – 507

[14] Kronthaler, F., 2003, A Study of the Competitiveness of Regions based on a Cluster Analysis: The Example of East Germany. Laporan Penelitian Institute for Economic Research Halle (IWH)

[15] Jaba, E., Ionescu, A.M., Iatu, Corneliu and Balan, C.B. 2009. "The Evaluation Of The Regional Profile Of The Economic Development In Romania". Analele Ştiintifice Ale Universităţii „Alexandru Ioan Cuza" Din Iaşi. Tomul LVI ŞtiinŃe Economice 2009

[16] Vincze, M., and Mezei, E., 2011, "The increase of rural development measures efficiency at the micro-regions level by cluster analysis: A Romanian case study". Eastern Journal Of European Studies, Volume 2, Issue 1, June 2011

[17] Nosova, O., 2013. "The Innovation Development in Ukraine: Problems and Development Perspectives". International Journal Of Innovation And Business Strategy. Vol. 02/August 2013

[18] Vydrová, H. V., and Novotná, Z., 2012. "Evaluation Of Disparities In Living Standards Of Regions Of The Czech Republic". Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis. Volume LX 42, Number 4.