ISBN No. 978-602-98559-3-7

Analisis *Cluster* Data Longitudinal pada Pengelompokan Daerah Berdasarkan Indikator IPM di Jawa Barat

A.S Awalluddin1, I. Taufik2
UIN SunanGunungDjati Bandung
Email : aasolih@uinsgd.ac.id

ABSTRAK

Anaisis cluster dapat digunakan untuk melakukan pengelompokan objek berdasarkan kesamaan karakteristik data yang ada pada setiap objek tersebut. Umumnya analisis cluster terbatas pada struktur data cross section (data silang), dengan asumsi pada satu waktu pengamatan. Penelitian ini bertujuan untuk menentukan pendekatan baru dalam analisis cluster untuk data longitudinal dengan struktur data yang bukan hanya cross section tapi juga time series (rumpun waktu) untuk data multivariabel. Perluasan metode analisis cluster hirarki Ward dengan kombinasi analisis data cross section dan time series dijadikan sebagai dasar analisis matematik dalam penelitian ini. Implementasi metode dilakukan pada studi kasus dataIndeks Pembangunan Manusia (IPM) Jawa Barat berupa empat variabel indikator komponen IPM yaitu : Angka Harapan Hidup (AHH), Harapan Lama Sekolah (HLS), Rata-rata Lama Sekolah (RLS), dan Pengeluaran Perkapita (PP), untuk mengetahui pengelompokkan kabupaten/kota, sehingga dapat diketahui daerah mana saja yang perlu mendapatkan prioritas peningkatan nilai variabel indikator IPM dalam kebijakan pembangunan Provinsi.

***Kata kunci** : Analisis cluster, Analisis Multivariat,Data Longitudinal, Metode Ward*

1. PENDAHULUAN

Analisis *cluster* adalah teknik multivariat yang mempunyai tujuan utama untuk mengelompokkan objek penelitian berdasarkan karakteristik yang dimilikinya. Analisis *cluster* mengklasifikasi objek sehingga setiap objek yang paling dekat karakteristiknya dengan objek lain berada dalam *cluster* yang sama. *Cluster* yang baik terbentuk memiliki homogenitas internal dan heterogenitas eksternal yang tinggi. Solusi analisis *cluster* bergantung pada variabel-variabel yang digunakan sebagai dasar untuk menilai kesamaan. Metode analisis *cluster* yang ada umumnya untuk analisis data *cross section*. Kajian analisis *cluster* untuk data longitudinal masih terbatas.

Beberapa literatur telah menjelaskan analisis *cluster* secara umum dan dapat dijadikan sebagai kerangka teoretis yang dapat dijadikan dasar dalam penelitian ini diantaranya ditulis oleh [1] yang lebih mengkaji secara metematis, sedangkan [2] lebih mengkaji secara aplikatif. Landasan teoritis analisis data longitudinal dapat diperoleh dalam tulisan [3] yang membahas secara mendasar konsep data longitudinal, [4] membahas beberapa analisis dengan beragam variasi kondisi dan asumsi metode analisis data longitudinal, dan pembahasan beberapa penelitian mengenai analsis data longitudinal dikumpulkan oleh [5].[6] menguraikan penggunaan data panel untuk analsis ekonomi.

Metode dasar analisis baik yang berkaitan dengan analisis *cluster* maupun data longitudinal, menjadi dasar pengembangan metode analisis *cluster* untuk data longitudinal yang dijelaskan dalam makalah ini. Beberapa penelitian yang berkaitan dengan analisis *cluster*dan data longitudinal diantaranya telah ditulis oleh [7] yang mengkaji analisis *cluster* untuk data panel variabel tunggal (*single*) dengan menggunakan metode regresi bertahap. [8] melakukan pengkajian pengelompokan (*clustering*) untuk data longitudinal multivariabel kontinyu dan diskrit dengan menggunakan inferensi bayesian pada model dan menggunakan simulasi MCMC (*Markov*

Chain Monte Carlo). Bagaimana dampak pengelompokan pada nilai varians dalam analisis data longitudinal dibahas oleh [9].

Makalah ini menjelaskan bagaimana analisis *cluster* untuk data longitudinal khususnya data multivariabel. Analisis *cluster* data longitudinal lebih kompleks dibandingkan dengan analisis *cluster* data *cross section*, karena bukan hanya dari dimensi objek yang diperhatikan tetapi juga dari dimensi waktu (*time series*). [10] telah mengkaji analisis *cluster*menggunakan perluasan metoda aglomerasihirarki Ward dengan *sum of squared deviation (SSD)* untuk data longitudinal multivariabel. Makalah ini menawarkan alternatif lain pendekatan perluasan Ward dengan *sum of absolute deviation* (SAD) untuk data longitudinal multivariabel.

Penerapan metoda dilakukan untuk menentukan pengelompokkan Kabupaten dan Kota di Provinsi JawaBarat berdasarkan pada variabel-variabel indikator Indeks Pembangunan Manusia (IPM) untuk tahun 2010 – 2014. Data merupakan data longitudinal multivariabel dengan pengamatan waktu selama empat tahun dan banyaknya objek pengamatan 26 Kabupaten dan Kota dan variabel yang diukur adalah Angka Harapan Hidup (AHH), Harapan Lama Sekolah (HLS), Rata-rata Lama Sekolah (RLS), dan Pengeluaran Perkapita (PP).

Pembahasan dalam makalah ini terbagi dalam beberapa bagian.Bagian 2. Membahasstruktur data longitudinal multivariabel untuk memberikan gambaran kompleksitas struktur data jika dibandingkan dengan dengan data *cross section* maupun data longitudinal dengan variabel tunggal (*single*). Bagian 3. Membahas analsis *cluster* secara umum yang menjadi dasar pengembangan metode *cluster* data longitudinal dengan dua pendekatan yaitu perluasan Ward SSD dan alternatif lain perluasan metode Ward SAD untuk data longitudinal multivariabel. Bagian 4. Berisi penerapan metoda untuk kasus pengelompokkan daerah Kabupaten dan Kota dengan data longitudinal multivariabel berdasarkan variabel indikator IPM tahun 2010-2014 dengan menganalisis perbandingan hasil pengelompokkan yang diperoleh dengan kedua metode.

2. DATA LONGITUDINAL MULTIVARIABEL

Struktur data longitudinal dapat dibedakan menjadi struktur data longitudinal dengan variabel tunggal dan data longitudinal multivariabel. Struktur data longitudinal tunggal dapat dianggap sebagai struktur data dengan tabel dua dimensi, di mana baris menyatakan objek/sampel sedangkan kolom menyatakan variabel tunggal dari waktu ke waktu. Misalkan $X_i(t)$ adalah variabel pengamatan objek ke-I pada saat pengamatan ke-t. Himpunan data longitudinal terdiri dari pengamatan pada objek penelitian ke-I selamat pengamatan untuk setiap i = 1,2…,N dan t = 1,2…,T. Struktur data seperti ini sama dengan struktur data *cross section*, dengan kolom menyatakan *p* buah variabel, sedangkan untuk struktur longitudinal menyatakan *t* waktu pengamatan. Desain data longitudinal variabel tunggal dapat dilihat dalam Tabel 1. Analisis cluster untuk data longitudinal tunggal dapat menggunakan analisis cluster data *cross section*. Pengelompokkan untuk data longitudinal variabel tunggal mudah dan dapat menggunakan *software* yang sudah tersedia seperti SPPS, MINITAB, dan lain sebagainya.

Tabel 1. *Desain Data Longitudinal Variabel Tunggal*

waktu(t) **objek(i)**	1	2	. . .	T
1	$x_{1(1)}$	$x_{1(2)}$	. . .	$x_{1(T)}$
2	$x_{2(1)}$			$x_{2(T)}$
.	.	.	.	.
.				
N	$x_{N(1)}$	.	.	$x_{N(T)}$

Struktur data longitudinal multivariabel lebih kompleks sehingga tidak dapat dibuat dalam bentuk tabel dua dimensi. Misalkan $X_{ij}(t)$ adalah variabel pengamatan untuk objek ke-i, variabel ke-j, dan saat pengamatan ke-t. Himpunan data longitudinal terdiri dari pengamatan pada objek penelitian ke-i, variabel ke-j selamat pengamatan untuk setiap i = 1,2…,N, j = 1,2…,p dan t = 1,2,...,T. Desain struktur data longitudinal multivariabel dapat dilihat dalam Tabel 2.

Tabel 2*. Desain Data Longitudinal Multivariabel*

waktu (t)	1	...	t	. . .	T
variabel (j) objek(i)	1. . . j . . . p		1 . . . j . . . p		1. . .j . . . p
1	$x_{11(1)}$. . . $x_{12(1)}$. . . $x_{1p(1)}$	...	...$x_{1j(t)}$...	. . .	... $x_{1p(T)}$
2	$x_{21(1)}$. . . $x_{22(1)}$. . . $x_{2p(1)}$	...			... $x_{2p(T)}$
.					
.	.		.	.	.
.					
N	$x_{N1(1)}$. . . $x_{N2(1)}$. . .$x_{Np(1)}$	...	...$x_{Nj(t)}$...	.	... $x_{Np(T)}$

Untuk menentukan statistik rata-rata dan varians dari data longtudinal multivariabel, dapat ditentukan berturut turut sebagai berikut :

Rata-rata dari variabel ke-j pada waktu ke-t

$$\bar{X}_j(t) = \frac{1}{N}\sum_{i=1}^{N} X_{ij}(t) \tag{1}$$

Rata-rata dari variabel ke-j

$$\bar{X}_j(t) = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} X_{ij}(t) \tag{2}$$

Varians dari variabel ke-j pada waktu ke-t

ISBN No. 978-602-98559-3-7

$$Var_{X_j(t)} = \frac{1}{N-1}\sum_{i=1}^{N}\left[X_{ij}(t) - \bar{X}_j(t)\right]^2 \quad (3)$$

Varians dari variabel ke-j

$$Var_{X_j} = \frac{1}{T(N-1)}\sum_{i=1}^{N}\left[X_{ij}(t) - \bar{X}_j(t)\right]^2 \quad (4)$$

Statistik ini akan digunakan dalam analisis cluster.

3. ANALISIS *CLUSTER* DATA LONGITUDINAL

Dalam menentukan analisis *cluster* longitudinal, dasar pemahaman dibangun berdasarkan pada analisis *cluster* secara umum untuk data *cross section*. Dua hal utama yang perlu ditetapkan dalam analisis cluster. *Pertama*, teknik apa yang akan digunakan, apakahteknik non-hirarki atau hirarki.*Kedua*, metode pengelompokkan apa yang akan dipilih, jika teknik non-hirarki yang dipilih maka metode yang dapat dipilih adalah metode partisi (*partitioning*) seperti *K-mean*, metode campuran distribusi (*mixture of distribution*), dan *density estimation*. Selain itu jika teknik hirarki yang dipilih maka metode yang dapat digunakan diantaranya : metode jarak terdekat (*single linkage*), metode jarak terjauh (*complete linkage*), metode rata-rata, metode centroid, metode jumlah kuadrat error / *sum of squares deviation* (Ward), dan lain sebagainya.

[10] menggunakan teknik aglomerasi hirarki dengan metode perluasan Ward dalam analisis *cluster* data longitudinal multivariabel yang juga digunakan sebagai salah satu metode analisis dalam makalah ini. Pengelompokkan yang optimal dapat terbentuk dengan diperoleh nilai observasi yang homogen antar anggota di dalam*cluster*, tetapi berbeda secara nyata antar *cluster*. Teknik pengelompokkan ini dimulai dengan *n* cluster dengan menganggap bahwa setiap objek/sampel adalah sebuah *cluster*. Selanjutnya pengukuran jarak antar objek dilakukan untuk mengelompokkan sesuai dengan kedekatan antar objek atau kelompok objek sebagai *cluster* baru. Proses ini terus dilakukan sampai terbentuk *cluster* yang lebih kecil. Penentuan jarak antara *cluster* baru yang terbentuk dengan sisa *cluster*ditentukan dengan metodepautan pengelompokkan (*linkage* method).

Makalah ini memilih metode pautan pengelompokkan Ward, metode ini tidak hanya mempertimbangkan jarak antar cluster, tetapi juga di dalam *cluster*. Berbeda dengan metode lain yang hanya mempertimbangkan jarak antar *cluster* [1]. Metode Ward juga dikenal seabagai metode jumlah kuadrat deviasi (*sum of squares deviation*). Fungsi jumlah kuadrat deviasi (SSD) untuk data longitudinal didefinisikan dalam rumus (5) berikut :

$$S_h = \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i\in i^h}\left[X_{ij}(t) - \bar{X}_j^h\right]^2 \quad (5)$$

dengan S_h menyatakan jumlah kuadrat deviasi dari *clusterh*, dan $\bar{X}_j^h$ adalah rata-rata variabel ke-j pada waktu ke-t untuk *clusterh* yang dapat diperoleh dengan menggunakan rumus (1) , dan i^h menyatakan seluruh objek yang termasuk *clusterh*. Perluasan metode Ward dengan menggunakan *least absolute deviation* telah diteliti oleh [11]

ISBN No. 978-602-98559-3-7

untuk jenis data *cross section*. Perluasan Ward dengan *sum of absolute deviation* (SAD) untuk data longitudinal yang digunakan dalam makalah ini didefinisikan dalam rumus (6) berikut :

$$M_h = \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i\in i^h}\left|X_{ij}(t) - \bar{X}_j^h\right| \tag{6}$$

denganM_h menyatakan jumlah absolut deviasi dari *clusterh*.

Pemilihan *cluster* baru dilakukan dengan teknik aglomerasi hirarki Ward dapat dilakukan dengan formula Lance-Williams [1] sebagai berikut :

$$D_{(C,AB)} = \frac{n_A + n_c}{n_A + n_B + n_c}S_{AC} + \frac{n_B + n_c}{n_A + n_B + n_c}S_{BC} - \frac{n_c}{n_A + n_B + n_c}S_{AB} \tag{7}$$

dengan $n_A,\ n_B,\ n_c$ berturut-turut adalah banyaknya objek/sampel pada *cluster* A, B, dan C.

Menentukan nilai S_{AC}, S_{BC}, dan S_{AB}diperoleh dengan rumus (5). Formula yang sama juga diterapkan ketika menggunakan M_h dengan rumus (6). Pembentukan *cluster*data longitudinal multivariabel dapat dilakukan dengan langkah berikut :

1. Tentukan data $X_{ij}(t)$; i=1,2,...N, j=1,2,...p, t=1,2,..T
2. Mulai dengan N *cluster*
3. Hitung persamaan (5) dan (6)
4. Pilih min $\{S_h\}$ dan min $\{M_h\}$ sebagai *cluster* N-1
5. Hitung $D_{(C, AB)}$pada persamaan (7)
6. Ulangi langkah 3 s.d 5 untuk cluster N-2, N-3, dan seterusnya
7. Sampai terbentuk 1 *cluster* untuk setiap metode

4. ANALISIS *CLUSTER* DATA IPM JAWA BARAT

Human Development Index (HDI) atau sering dikenal Indeks Pembangunan Manusia (IPM) merupakan indikator yang dapat menjelaskan perkembangan pembangunan manusia secara representatif. Tiga dimensi dasar yang menjadi tolak ukur besar kecilnya nilai IPM di suatu wilayah adalah harapan hidup yang tinggi, pendidikan yang cukup dan standar hidup yang layak. Dalam perkembangannya indikator dan perhitungan IPM mengalami perubahan sejak tahun 2010, perbedaan yang cukup signifikan ini menjadikan perhatian pemerintah daerah untuk memperhatikan variabel indikator dalam rangka meningkatkan nilai IPM. Tidak hanya itu perhatian pengelompokkan Kabupaten dan Kota berdasarkan capaian indikator IPM masing-masing menjadi perhatian untuk menentukan prioitas pembangunan daerah.

ISBN No. 978-602-98559-3-7

Penerapan metode analsis *cluster* data longitudinal yang telah dijalaskan dalam pembahasan sebelumnya dilakukan pada data IPM Kabupaten dan Kota di Provinsi Jawa Barat dengan metode perhitungan baru. Data diambil dari 26 Kabupaten dan Kota di Jawa Barat (tidak termasuk Kab. Pangandaran) selama periode 2010-2014 bersumber dari BPS Jawa Barat [12]. Empat variabel indikator IPM digunakan yaitu : Angka Harapan Hidup (AHH), Harapan Lama Sekolah (HLS), Rata-rata Lama Sekolah (RLS), dan Pengeluaran Perkapita (PP). Perhitugan metode baru digunakan dalam menentukan nilai dari setiap variabel. Secara desriptif statistik data yang digunakan dapat dilihat alam Tabel 3 dan rata-rata IPM Kab/Kotadapat dilihat dalam Tabel 4.

***Tabel 3**. Statistik Data Variabel Indikator IPM*

Variabel	Mean	Var	Max	Min
AHH	71,29	1,67	74,18	67,54
HLS	11,69	0,95	13,71	9,62
RLS	7,72	1,52	10,78	4,93
PP	9344,96	2178,87	15048,47	6149,57

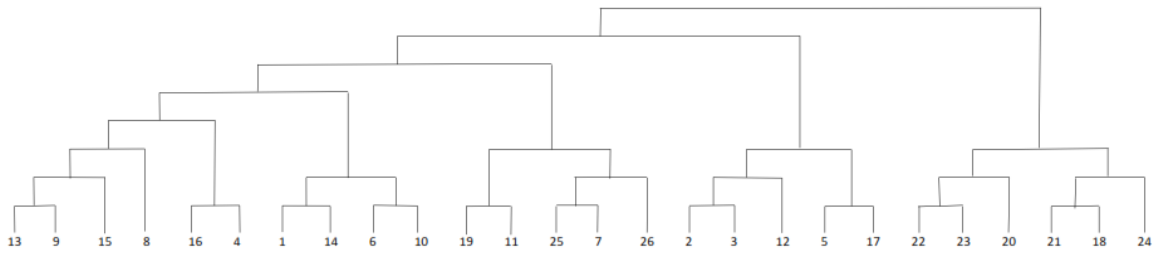
***Tabel 4.** Rata-rata IPM Kab/Kota Jawa Barat*

No.	Kabupaten	Rata-rata	No.	Kabupaten/Kota	Rata-rata
1	Bogor	65,78	14	Purwakarta	66,23
2	Sukabumi	62,36	15	Karawang	65,89
3	Cianjur	60,40	16	Bekasi	69,24
4	Bandung	68,17	17	Bandung Barat	63,02
5	Garut	61,14	18	Kota Bogor	72,24
6	Tasikmalaya	61,63	19	Kota Sukabumi	69,67
7	Ciamis	66,25	20	Kota Bandung	78,29
8	Kuningan	65,57	21	Kota Cirebon	71,88
9	Cirebon	64,57	22	Kota Bekasi	77,89
10	Majalengka	63,18	23	Kota Depok	77,55
11	Sumedang	67,36	24	Kota Cimahi	75,02
12	Indramayu	62,19	25	Kota Tasikmalaya	67,86
13	Subang	64,78	26	Kota Banjar	67,57

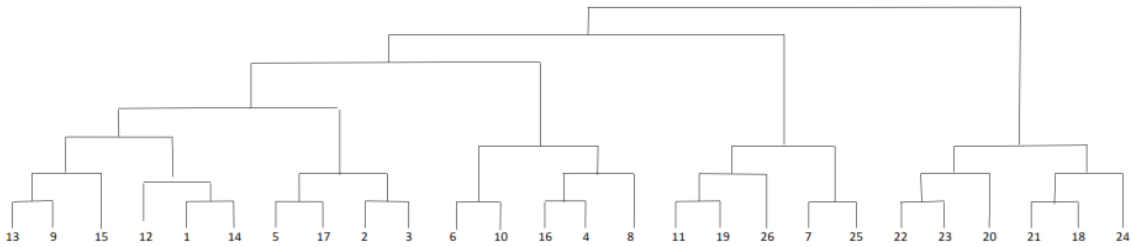
Penentuan *cluster* dengan metode yang telah dijelaskan dilakukan dengan dua pendekatan yaitupenerapan analisis *cluster* data longitudinal dengan perluasan Ward SSD rumus (7) dan (9) dan perluasan Ward SAD rumus (8) dan (9). Hasil pengelompokkan ditunjukkan berturut-turut dalam dendogram **Gambar 1** dan **Gambar 2**.

Secara umum kedua metode menghasilkan lima pengelompokkan (*cluster*) besar dengan dua *cluster* memiliki objek yang sama. *Cluster* pertama yaitu : Kota Bekasi, Kota Depok, Kota Bandung, Kota Cirebon, Kota Bogor, dan Kota Cimahi. *Cluster* kedua yaitu : Kota Sukabumi, Kab. Sumedang, Kota Tasikmalaya, Kota Cimahi, dan Kota Banjar. Kedua *cluster* secara berturut-turut menunjukkan bahwa dari keempat indikator IPM kedua cluster di atas merupakan daerah yang yang paling baik dan baik jika dilihat dari keempat variabel indikator IPM. Tiga

cluster lainnya dari kedua metode memberikan hasil yang sedikit berbeda. Khusus untuk *cluster* dengan objek daerah Kab. Sukabumi, Kab. Cianjur, Kab. Garut, dan Kab. Bandung Barat, hasil pengelompokkan metode perluasan Ward SAD, ditambah Kab. Indramayu yang diperoleh metode perluasan Ward SSD menunjukkan *cluster* yang paling rendah , sehingga kelima daerah tersebut dapat dikategorikan seabagai daerah prioritas dalam pembangunan di Provinsi Jawa Barat untuk meningkatkan nilai keempat variabel indikator IPM.



Gambar 1. Dendogram Perluasan Ward SSD



Gambar 2. Dendogram Perluasan Ward SAD

5. SIMPULAN

Analisis *cluster* dengan data longitudinal dengan memperhatikan dimensi waktu dan objek berbeda pendekatannya dengan analisis umum yang digunakan karena mengabaikan pengaruh perubahan waktu terhadap perubahan nilai variabel. Alternatif analisis *cluster* untuk data longitudinal telah diuraikan dalam makalah ini dengan dua pendekatan alternatif yaitu jumlah kuadrat deviasi (SSD) dan jumlah absolut deviasi (SAD) untuk data longitudinal dengan perluasan Ward sebagai *linkage method*.

Pengelompokkan studi kasus pada data empat variabel indikator IPM untuk 26 Kabupaten dan Kota di Jawa Barat dari5*cluster* daerah yang terbentuk, memberikan hasil yang sama untuk tiga *cluster*. Daerah dengan *cluster* sangat baik berdasarkan empat variabel indikator IPM yaitu Kota Bekasi, Kota Depok, Kota Bandung, Kota Cirebon, Kota Bogor, dan Kota Cimahi. Sedangkan cluster rendah yang perlu mendapat perhatian pemerintah baik Provinsi maupun Pusat untuk menjadikan prioritas dalam meningkatkan variabel indikator IPM yaitu : Kab. Sukabumi, Kab. Cianjur, Kab. Garut, Kab. Bandung Barat, dan Kab. Indramayu

6. UCAPAN TERIMA KASIH

Penelitian ini dibiayai oleh dana BOPTN UIN Sunan Gunung Djati. Kami mengucapkan terima kasih atas dukungan yang diberikan.

KEPUSTAKAAN

[1] Rencher, A.C., 2002. *Methods of Multivariate Analysis*, 2nd Edition, New York; John Wiley & Sons.

[2] Hair, J.E., dkk., 1998. *Multivariate Data Analysis fifth Ed*., Prentice Hall International.

[3] Toon W. Taris,2000. *A Primer in Longitudinal Data Analysis*., Sage Publications Ltd.

[4] Frees, Edward. W, 2004. *Longitudinal and Longitudinal Data : Analysis and Applications in Social Science*., Cambridge University Press.

[5] Hsio, Cheng,dkk., 2004. *Analysis of Longitudinals and Limited Dependent Variable Models*.,Cambridge Un.iversity Press.

[6] Baltagi, Badi H.,2005. *Econometric Analysis of Panel Data third edition*, John Wiley and Sons.

[7] Mouchart,Michel, Jeroen V.K., *2005. Clustered panel data Clustered Panel Data Models: An Efficient Approach for Nowcasting from Poor Data*. *I*nternational Journal of Forecasting **5**:577-594.

[8] Komarek, A., Kamarkova, L., 2013. *Clustering for multivariate continuous and discrete longitudinal data.*The Annals of Applied Statistics, **7**(1), 177–200.

[9] C.Skinner., M.T.Vieire., 2007. *Variance Estimation in the Analysis of Clustered Longitudinal Survey Data*., Survey Methodology Vol.33, No.1, pp.3-12.

[10] Zheng, B., Li, S., 2014. *Multivariable Panel Data Cluster Analysis and Its Application., Computer Modelling & New Technologies***18**., 553-557.

[11] Strauss, T., J Von Maltitz, M., 2014. *Statistical Classification of Languages : Generalising Method for Use with Manhattan Distance*., Technical Report., University of The Free State.

[12] Badan Pusat Statistika Provinsi Jawa Barat. *Indeks Pembangunan Manusia Metode Baru Provinsi Jawa Barat dan Kabupaten/Kota Tahun 2010-2014*. (Online). (http://jabar.bps.go.id/linkTabelStatis/view/id/95)