JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

VOLUME 7, NOMOR 2, DESEMBER 2015 ISSN 2086 - 4132

UNIT PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
SEKOLAH TINGGI ILMU STATISTIK
(UPPM-STIS)

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

Jurnal "Aplikasi Statistika dan Komputasi Statistik" memuat karya ilmiah hasil penelitian dan kajian teori statistika dan komputasi statistik yang diterapkan khususnya pada bidang ekonomi dan sosial kependudukan, serta teknologi informasi yang terbit dua kali dalam setahun setiap bulan Juni dan Desember

Penanggung Jawab: Ketua Sekolah Tinggi Ilmu Statistik

Alamat Redaksi:
Sekolah Tinggi Ilmu Statistik
Jl. Otto Iskandardinata 64C
Jakarta Timur 13330
Telp. 021-8191437

Redaksi menerima karya ilmiah atau artikel penelitian mengenai kajian teori statistika dan komputasi statistik pada bidang ekonomi dan sosial kependudukan, serta teknologi informasi. Redaksi berhak menyunting tulisan tanpa mengubah makna substansi tulisan. Isi Jurnal Aplikasi Statistika dan Komputasi Statistik dapat dikutip dengan menyebutkan sumbernya.

PENGANTAR REDAKSI

Syukur *Alhamdulillah*, di akhir tahun 2015 "Jurnal Aplikasi Statistika dan Komputasi Statistik" Volume 7, Nomor 2, Desember 2015 dapat diterbitkan. Jurnal ini terwujud atas partisipasi Bapak/Ibu dosen di STIS dan luar STIS yang telah mengirimkan artikel kepada redaksi melalui koreksi konstruktif dari mitra bestari serta ketelitian dari para editor jurnal. Untuk atensi dan kerjasama yang baik guna keberlangsungan terbitnya jurnal ini redaksi mengucapkan terimakasih.

Artikel yang dimuat dalam edisi kali ini menyajikan berbagai variasi penggunaan metode statistika yang diterapkan dalam membahas daya saing industri, ketertinggalan kabupaten, pemberian imunisasi, di samping juga mengenai permasalahan dalam pemanfaatan teknologi informasi.

Semoga artikel dalam jurnal ini dapat menambah pengetahuan para pembaca tentang penggunaan metode statistika serta komputasi statistik pada berbagai jenis data. Redaksi terus menunggu artikel-artikel ilmiah selanjutnya dari Bapak/Ibu guna dapat menghasilkan publikasi yang menjadi salah satu sarana untuk memberikan sosialisasi statistika dan komputasi bagi masyarakat.

Jakarta, Desember 2015

Salam,

E k a r i a

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

VOLUME 7, NOMOR 2, DESEMBER 2015

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK
(Journal of Statistical Application & Statistical Computing)
ISSN 2046 – 4132 Volume 7, Nomor 2, Desember 2015

DDC : 315.98

Metty Nurul Romadhona

Model Probit Biner Bivariat Pada Pemberian Imunisasi Dasar dan Air Susu Ibu
(Studi Kasus di Provinsi Kalimantan Selatan Tahun 2013)

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Nomor 2, Desember 2015, hal. 67-80

Abstrak
Tujuan ke empat *Millenium Development Goals* (MDG's) adalah menurunkan angka kematian anak. Salah satu upaya untuk mengurangi angka kematian anak adalah meningkatkan kekebalan tubuh pada anak. Kekebalan tubuh pada anak diperoleh dengan pemberian imunisasi dasar yang lengkap dan ASI eksklusif. Penelitian ini bertujuan mengaplikasikan model probit biner bivariat untuk mengetahui faktor-faktor yang mempengaruhi pemberian imunisasi dasar dan ASI eksklusif. Sumber data yang digunakan dalam penelitian ini adalah data Survei Sosial Ekonomi Nasional (SUSENAS) Provinsi Kalimantan Selatan Tahun 2013. Pemilihan model terbaik berdasarkan kriteria AIC *(Akaike Information Criterion)* menghasilkan informasi bahwa umur perkawinan pertama ibu, pendidikan ibu, pekerjaan bapak, penolong kelahiran terakhir dan status daerah berpengaruh signifikan terhadap pemberian imunisasi dasar dan ASI eksklusif.
Kata kunci : Imunisasi, ASI Eksklusif, Model Probit Biner Bivariat, AIC

DDC : 315.98

Rani Nooraeni

Metode Cluster Menggunakan Kombinasi Algoritma Cluster K-Prototype dan Algoritma Genetika untuk Data bertipe Campuran

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Nomor 2, Desember 2015, hal. 81-98

Abstrak
Clustering adalah salah metode utama pada data mining yang berguna untuk mengeksplorasi data. Membagi suatu data set berukuran besar ke dalam cluster yang sehomogen mungkin adalah tujuan dalam metode data mining. Salah satu metode clustering konvensional yaitu algoritma K-Means efisien untuk dataset berukuran besar dan tipe data numerik tapi tidak untuk data kategorikal. Algoritma K-Prototype menghilangkan keterbatasan pada data numerik tapi dapat juga digunakan pada data kategorikal. Namun solusi yang dihasilkan oleh kedua algoritma tersebut merupakan solusi lokal optimal dimana salah satu penyebabnya adalah penentuan pusat cluster awal. Untuk menghadapi masalah tersebut maka algoritma genetika menjadi salah satu usulan yang dapat digunakan untuk mengoptimalkan hasil pengclusteran dengan K-Prototype. Hasil dari penelitian menunjukkan optimasi pusat cluster dengan algoritma genetika berhasil meningkatkan akurasi hasil cluster dengan K-Prototype.
Kata kunci: Data Mining, Analisis Cluster, Data Campuran, Algoritma K-Prototype, Algoritma Genetika

DDC : 315.98

Retno Indrawati dan Ernawati Pasaribu

Bagaimana Daya Saing Industri *Life Sciences* di Indonesia: Sebuah Pembandingan dengan Negara-Negara Lain

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Nomor 2, Desember 2015, hal. 99-114

Abstrak
Indonesia adalah negara terbesar di Asia Tenggara dengan lebih dari 20 juta penduduknya adalah kelas menengah yang dewasa ini memiliki pengaruh penting dan semakin menginspirasi. Indonesia telah menjadi pasar yang menarik karena perkembangan pesat jumlah konsumen, khususnya dari kelompok penduduk berpendapatan menengah tersebut. Tingginya jumlah populasi (lebih dari 250 juta penduduk) juga mengindikasikan besarnya potensi sumber tenaga kerja. Industri Life Sciences (LS) secara luas mulai dikenal sebagai

aliran baru ekonomi berbasis ilmu pengetahuan. Studi ini mengidentifikasi posisi relatif Indonesia dikaji dari investasi langsung luar negeri (foreign direct investment-FDI) pada industri LS, sekaligus dari sisi daya saing (competitiveness) dengan negara-negara lain di dunia Berdasarkan sektor LS, pesaing utama Indonesia adalah Portugal, Turki, Saudi Arabia, dan Nigeria, sedangkan berdasarkan aktivitas LS, Argentina dan Bulgaria adalah saingan utama. Studi ini juga mengungkapkan bahwa FDI yang masuk ke Indonesia dipengaruhi terutama oleh tingkat inflasi dan return on investment.

Kata Kunci : Indonesia, *life sciences*, daya saing, investasi langsung luar negeri

DDC : 315.98

Siskarossa Ika Oktora

Analisis *Multivariate Adaptive Regression Splines* (MARS) Pada Prediksi Ketertinggalan Kabupaten Tahun 2014

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Nomor 2, Desember 2015, hal. 115-128

Abstrak

Kabupaten tertinggal merupakan kabupaten yang masyarakat serta wilayahnya relatif kurang berkembang dibandingkan daerah lain dalam skala nasional berdasarkan kategori perekonomian masyarakat, Sumber Daya Manusia (SDM), infrastruktur, kemampuan keuangan daerah, aksesbilitas, dan karakteristik daerah. Pengklasifikasian kabupaten tertinggal tidaklah mudah karena melibatkan variabel dan observasi dalam jumlah yang cukup banyak. Selain itu diantara variabel yang digunakan memiliki keterkaitan antara satu dengan yang lain. MARS adalah salah satu metode pengklasifikasian yang mampu menangani data berdimensi tinggi dengan pola data yang tidak diketahui sebelumnya.

Dari model MARS yang dibangun, terdapat lima variabel utama yang berpengaruh terhadap ketertinggalan kabupaten diantaranya adalah pengeluaran konsumsi per kapita, angka harapan hidup, persentase rumah tangga pengguna listrik, rata-rata jarak dari kantor desa/kelurahan ke kantor kabupaten yang membawahi, serta jumlah desa yang memiliki pasar tanpa bangunan permanen. Akurasi dari model MARS yang terbentuk sangat tinggi, yakni mencapai 97,83 persen dan dapat dipergunakan untuk melakukan prediksi ketertinggalan kabupaten.

Kata kunci : Multivariate Adaptive Regression Splines (MARS), Kabupaten tertinggal, Prediksi ketertinggalan

DDC : 315.98

Hafshoh Mahmudah dan Ricky Yordani

Visualisasi Penggerombolan Wilayah Berdasarkan Teori Pertumbuhan Ekonomi Menggunakan Aplikasi Integrasi Self Organizing Map (SOM) dan Sistem Informasi Geografis

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Nomor 2, Desember 2015, hal. 129-142

Abstrak

Pertumbuhan ekonomi merupakan salah satu faktor penting untuk menentukan kesejahteraan suatu wilayah. Akan tetapi, perbedaan kondisi geografis dan potensi wilayah menyebabkan perbedaan kondisi ekonomi yang berbeda antarwilayah. Studi kasus dilakukan terhadap Provinsi Jawa Tengah karena merupakan salah satu kontributor PDRB terbesar di Indonesia, yang ternyata masih memiliki ketimpangan perekonomian antar kota dan antar kabupaten. Untuk memudahkan visualisasi pertumbuhan ekonomi maka dibuatlah suatu aplikasi yang mampu melihat secara mudah efek pertumbuhan dan penggerombolan dalam wilayah Provinsi Jawa Tengah tersebut. Metode yang bisa digunakan untuk analisis gerombol sangat beragam. Salah satu metode alternatif adalah menggunakan metode *Self Organizing Map* (SOM) yang mampu menggerombolkan data multidimensi disertai dengan visualisasinya dengan teknik *Unsupervised Artificial Neural Network*. Aplikasi ini memudahkan visualisasi dan analisisnya karena diintegrasikan dengan Sistem Informasi Geografis (SIG). Aplikasi yang dibuat selanjutnya digunakan untuk melakukan analisis gerombol dengan data studi kasus Provinsi Jawa Tengah. Visualisasi yang dihasilkan mampu menunjukkan pola pertumbuhan ekonomi di Provinsi Jawa Tengah namun belum terlihat adanya pemusatan kutub pertumbuhan ekonomi di Provinsi Jawa Tengah karena pola penggerombolan berdasarkan indikator pertumbuhan ekonomi masih menyebar.

Kata kunci : Kutub Pertumbuhan Ekonomi, Self

Organizing Map, Analisis Gerombol

DDC : 315.98

Takdir dan Choerul Afifanto

Analisis Preferensi Mahasiswa STIS Berdasarkan Akun Facebook yang Dimiliki
Studi Kasus: Mahasiswa STIS Angkatan 54 sampai 57

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Nomor 2, Desember 2015, hal. 143-154

Abstrak
Penggunaan sosial media saat ini sangat masif di berbagai kalangan. Facebook merupakan salah satu sosial media yang memiliki jumlah dan frekuensi penggunaan yang besar serta memuat banyak data, khususnya data yang berupa relasi antarentitas. Penelitian ini mengidentifikasi preferensi, yakni kecenderungan topik yang digemari, mahasiswa STIS aktif berdasarkan akun Facebook yang dimiliki. Akun Facebook tersebut diperoleh dari grup-grup angkatan. Preferensi diperoleh dengan melakukan *crawling* terhadap halaman (*page*) yang di-*like* serta *group* yang diikuti oleh mahasiswa. Hasil dari penelitian ini adalah gambaran karakteristik preferensi mahasiswa berupa statistik mengenai jenis-jenis topik yang diminati oleh mahasiswa STIS serta visualisasi terbentuknya *cluster*/komunitas mahasiswa untuk topik tertentu. Pendekatan yang digunakan pada penelitian ini untuk mengekstraksi dan menganalisis data pada sosial media diharapkan dapat menjadi referensi bagi berbagai bidang penelitian yang memanfaatkan data social media.
Kata kunci : Facebook, Analisis Sosial Media, *Social Graph*

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK
(Journal of Statistical Application & Statistical Computing)
ISSN 2046 – 4132 Volume 7, Nomor 2, Desember 2015
Kata kunci bersumber dari artikel. Lembar abstrak ini boleh diperbanyak tanpa izin dan biaya

DDC : 315.98

Metty Nurul Romadhona

Binary Bivariate Probit Model on Giving Basic Immunization and Breast Milk

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Number 2, December 2015, pg. 67-80

Abstract
The fourth goal of the Millennium Development Goals (MDGs) is to reduce child mortality. One of the efforts to reduce child mortality is increasing immunity for children. Immunity for children can be obtained by providing complete basic immunization and exclusive breastfeeding This study aimed to apply the bivariate binary probit model in determining factors that affect provision of basic immunization and exclusive breastfeeding. The data source used in this research is data of the 2013 National Socio Economic Survey (SUSENAS) in South Kalimantan Province. The best model selection criterion based on the AIC (Akaike Information Criterion) values provided information that the age of first marriage, mother's education, father's job, the birth attendants and status of the living area have significant effects on the provision of basic immunization and exclusive breastfeeding.
Keywords : Immunization, Exclusive Breastfeeding, Bivariate Binary Probit Model, AIC.

DDC : 315.98

Rani Nooraeni

Cluster Method Using A Combination of Cluster K-Prototype Algorithm and Genetic Algorithm for Mixed Data

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Number 2, December 2015, pg. 81-98

Abstract
The government in setting policies require complete and accurate data that targeted programs .As more data is collected, the more complex types of data held. Data mining is one of the methods used for this data type . Clustering is one of the main methods in data mining that useful to explore the data. One conventional clustering methods namely the K - Means algorithm efficient for large dataset and numeric data types but not for categorical data type. K-prototype algorithm eliminates the limitations of the numerical data but can also be used on categorical data . But the solutions generated by the algorithm is a local optimal solution in which one of the causes is the determination of the initial cluster's center. Deal with these problems, the genetic algorithm was proposed for solving this global optimitation problem. The results of the study indicate that the cluster's center optimization with genetic algorithm success to improve the accuracy of the results of the cluster with K– Prototype algorithm.
Keywords : Data Mining, Cluster Analysis, Mixed Data, K-Prototype Algorithm, Genetic Algorithm

DDC : 315.98

Retno Indrawati dan Ernawati Pasaribu

How Competitive is Life Sciences Industry in Indonesia: Compared to Other World Countries

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Number 2, December 2015, pg. 99-114

Abstract
Indonesia is the South East Asia's largest economy and has a substantial and increasingly inspirational middle class of over 20 million. Indonesia has become an attractive market due to her strongly growing consumer market, especially those of the middle income segment. The high number of population (more than 250 million people) also indicates the existing potential pool of labour. Life Sciences (LS)

industry is widely recognised as the new wave of knowledge-based economy. This study identifies relative position of Indonesia in terms of foreign direct investment (FDI) in LS industry and competitiveness of the LS industry in Indonesia compared with other countries. Based on LS sector, Indonesia has to compete mainly with Portugal, Turkey, Saudi Arabia, and Nigeria, while based on LS activities, Argentina and Bulgaria are the main competitors. This study also reveals that FDI inflow to LS industry in Indonesia is influenced mainly by inflation and return on investment.

Keywords : Indonesia, life sciences, competitiveness, foreign direct investment

DDC : 315.98

Siskarossa Ika Oktora

Multivariate Analysis Adaptive Regression Splines (MARS) on Prediction The Underdeveloped District in 2014

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Number 2, December 2015, pg. 115-128

Abstract

Underdeveloped districts are districts where community and the region is relatively less developed than other regions on a national scale by economic categories, Human Resources (HR), infrastructure, fiscal capacity, accessibility, and regional characteristics. The classification of underdeveloped districts is not easy because it involves many variables and observations. The variables also have a relationship each other. MARS is one of classification method which able to handle high dimensional data with unknown patterns previously.

From the MARS model, there are five main variables that affect the underdeveloped districts, which is consumption expenditure per capita, life expectancy, the percentage of household electricity users, the average distance from the village office to the district office, and the number of villages which has a market without a permanent building. The accuracy of the MARS model is very high, 97.83 percent and can be used to predict the underdeveloped district.

Keywords : Multivariate Adaptive Regression Splines (MARS), Underdeveloped districts, predict the underdeveloped.

DDC : 315.98

Hafshoh Mahmudah dan Ricky Yordani

Visualitation of Clustering Region by Economic Growth Theory Using The Integration Of Self Organizing Map (SOM) and Geographic Information System

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Number 2, December 2015, pg. 129-142

Abstract

Economic growth is one of factor that is critical to determining the welfare of a region. However, differences in geographical conditions and the potential of the area led to differences in economic conditions differ between regions. The case studies conducted on Central Java Province because it is one of the largest contributors to GDP in Indonesia, which still has economic inequality between cities and between districts. To make more easy for visualize the economic growth, researcher then made an application that is able to easily see the effect of growth and clustering in the province of Central Java. There are many methods that can be used for cluster analysis. One of the most common methods used are the K-Means. However, K-Means has some drawbacks. One alternative method is using the Self Organizing Map (SOM) which is capable clustering accompanied by visualization of multidimensional data with techniques Unsupervised Artificial Neural Network. This application allows visualization and analysis because it is integrated with Geographic Information Systems (GIS). Applications are made subsequently used to analyze clustering with case study data of Central Java province. The resulting visualization capable of showing a pattern of economic growth in Central Java Province but has not seen the concentration of economic growth pole in Central Java because clustering pattern based on indicators of economic growth spread.

Keywords : Economic Growth Pole, Self Organizing Map, Cluster Analysis

DDC : 315.98

Takdir dan Choerul Afifanto

Students Preference Analysis Based on Facebook Account Held of STIS

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 7, Number 2, December 2015, pg. 143-154

Abstract
Currently, social media is used massively in various societies. Facebook is one of the greatest social media in terms of total and frequency of uses, as well as the number of collected information, especially the information about relationships between entities. This study aims for analyzing preference of STIS's students based on their Facebook account. Their Facebook accounts are collected from their Facebook group communities. The preference data are collected by crawling the liked pages and joined groups. The results of this study are the characteristics view of students' preferences in form of statistics of interesting topic types and visualization of students' clusters for certain topics.
Keywords **:** *Facebook, Social Media Analysis, Social Graph*

MODEL PROBIT BINER BIVARIAT PADA PEMBERIAN IMUNISASI DASAR DAN AIR SUSU IBU (STUDI KASUS DI PROVINSI KALIMANTAN SELATAN TAHUN 2013)

BINARY BIVARIATE PROBIT MODEL ON GIVING BASIC IMMUNIZATION AND BREAST MILK

Metty Nurul Romadhon
Sekolah Tinggi Ilmu Statistik

Masuk tanggal: 04-12-2015, revisi tanggal: 17-01-2016, diterima untuk diterbitkan tanggal: 19-01-2016

Abstrak

Tujuan ke empat *Millenium Development Goals* (MDG's) adalah menurunkan angka kematian anak. Salah satu upaya untuk mengurangi angka kematian anak adalah meningkatkan kekebalan tubuh pada anak. Kekebalan tubuh pada anak diperoleh dengan pemberian imunisasi dasar yang lengkap dan ASI eksklusif. Penelitian ini bertujuan mengaplikasikan model probit biner bivariat untuk mengetahui faktor-faktor yang mempengaruhi pemberian imunisasi dasar dan ASI eksklusif. Sumber data yang digunakan dalam penelitian ini adalah data Survei Sosial Ekonomi Nasional (SUSENAS) Provinsi Kalimantan Selatan Tahun 2013. Pemilihan model terbaik berdasarkan kriteria AIC *(Akaike Information Criterion)* menghasilkan informasi bahwa umur perkawinan pertama ibu, pendidikan ibu, pekerjaan bapak, penolong kelahiran terakhir dan status daerah berpengaruh signifikan terhadap pemberian imunisasi dasar dan ASI eksklusif.

Kata kunci : Imunisasi, ASI Eksklusif, Model Probit Biner Bivariat, AIC

Abstract

The fourth goal of the Millennium Development Goals (MDGs) is to reduce child mortality. One of the efforts to reduce child mortality is increasing immunity for children. Immunity for children can be obtained by providing complete basic immunization and exclusive breastfeeding This study aimed to apply the bivariate binary probit model in determining factors that affect provision of basic immunization and exclusive breastfeeding. The data source used in this research is data of the 2013 National Socio Economic Survey (SUSENAS) in South Kalimantan Province. The best model selection criterion based on the AIC (Akaike Information Criterion) values provided information that the age of first marriage, mother's education, father's job, the birth attendants and status of the living area have significant effects on the provision of basic immunization and exclusive breastfeeding.

Keywords : *Immunization, Exclusive Breastfeeding, Bivariate Binary Probit Model, AIC*

PENDAHULUAN

Angka kematian bayi merupakan indikator yang penting untuk mencerminkan keadaan derajat kesehatan di suatu masyarakat, karena bayi yang baru lahir sangat sensitif terhadap keadaan lingkungan tempat orang tua bayi tinggal dan status sosial orang tua bayi. Dengan demikian angka kematian bayi merupakan tolok ukur yang sensitif dari semua upaya intervensi yang dilakukan pemerintah khususnya di bidang kesehatan. Angka kematian anak dan angka kematian balita dapat berguna untuk mengembangkan program imunisasi, serta program-program pencegahan penyakit menular terutama pada anak-anak, program tentang gizi dan pemberian makanan sehat untuk anak dibawah usia 5 tahun.

Angka kematian bayi dan anak di Indonesia berdasarkan hasil SDKI 2012 lebih rendah dari hasil SDKI 2007. Untuk periode lima tahun sebelum survei, angka kematian bayi hasil SDKI 2012 adalah 34 kematian per 1000 kelahiran hidup dan untuk angka kematian anak adalah 9 kematian per 1000 anak dengan umur yang sama pada pertengahan tahun tersebut. Angka tersebut mengalami penurunan

dibandingkan dengan hasil SDKI 2007 yaitu sebesar 35 kematian per 1000 kelahiran hidup untuk angka kematian bayi dan sebesar 10 kematian per 1000 anak dengan umur yang sama pada pertengahan tahun tersebut untuk angka kematian anak (BPS, BKKBN, Kementrian Kesehatan dan Measure DHS ,2012).

Penurunan angka kematian bayi dan anak tersebut memang sesuai dengan tujuan MDG's keempat yaitu menurunkan angka kematian anak. Penurunan angka kematian anak telah menunjukkan kemajuan yang signifikan dan diharapkan dapat tercapai pada tahun 2015. Namun penurunan angka kematian bayi maupun anak tersebut cenderung stagnan. Penyebab utama kematian balita adalah masalah neonatal (asfiksia, berat badan lahir rendah dan infeksi neonatal), penyakit infeksi (utamanya diare dan pneumonia) serta terkait erat dengan masalah gizi (gizi buruk dan gizi kurang). Kondisi ini disebabkan oleh masalah akses dan kualitas pelayanan kesehatan, masalah sosial ekonomi dan budaya, pertumbuhan infrastruktur serta keterbukaan wilayah tersebut akan pembangunan ekonomi dan pendidikan.

Salah satu upaya untuk mengurangi angka kematian balita adalah dengan meningkatkan kekebalan tubuh balita tersebut. Balita sangat mudah terserang penyakit, hal ini disebabkan masih belum kuatnya sistem kekebalan tubuh yang terdapat pada balita. Untuk menjaga sistem kekebalan tubuh terhadap balita diantaranya dengan memberikan imunisasi dan ASI eksklusif. Imunisasi adalah proses menginduksi imunitas secara buatan baik dengan vaksinasi (imunisasi aktif) maupun dengan pemberian antibody (imunisasi pasif). Sedangkan pemberian ASI bermanfaat sebagai nutrisi, untuk meningkatkan daya tahan tubuh dan meningkatkan kecerdasan. Sehingga pemberian imunisasi dan ASI dapat menjaga kesehatan tubuh pada balita.

Pemberian imunisasi lengkap pada balita di Indonesia berdasarkan SUSENAS 2013 sebanyak 71,70 persen (BPS, 2014). Cakupan imunisasi lengkap ini meningkat dari 67,67 persen SUSENAS 2012 (BPS, 2013). Sedangkan pemberian ASI Eksklusif kepada anak berusia 2-4 tahun di Indonesia dalam SUSENAS 2013 sebesar 44,50 persen (BPS, 2014) lebih tinggi dibandingkan dengan hasil SUSENAS 2012 sebesar 43,03 persen (BPS, 2013). Meskipun pemberian imunisasi dan ASI Eksklusif mengalami peningkatan namun masih dibawah yang diharapkan.

Kalimantan Selatan adalah salah satu provinsi di Indonesia yang terletak di pulau Kalimantan. Angka kematian anak menjadi salah satu masalah yang dihadapi di Provinsi Kalimantan Selatan. Hasil SDKI 2012 menunjukkan bahwa angka kematian anak di Provinsi Kalimantan Selatan sebesar 13 kematian per 1000 anak dengan umur yang sama di pertengahan tahun tersebut. Angka ini di atas angka nasional dan merupakan angka tertinggi di antara provinsi di pulau Kalimantan (BPS, BKKBN, Kementrian Kesehatan dan Measure DHS ,2012). Persentase balita yang mendapat imunisasi dasar lengkap di Provinsi Kalimantan Selatan pada tahun 2013 sebesar 76,61 persen (BPS, 2014). Hal ini mengalami penurunan dari tahun 2012 sebesar 76,99 persen (BPS, 2013). Persentase anak usia 2-4 tahun yang mendapatkan ASI eksklusif di Provinsi Kalimantan Selatan sebesar 37,39 persen pada tahun 2013 (BPS, 2014) yang turun sebesar 1,28 persen dari tahun 2012 sebesar 38,67 persen (BPS, 2013). Persentase anak usia 2-4 tahun yang mendapatkan ASI eksklusif di Provinsi Kalimantan Selatan tahun 2013 lebih rendah dari angka nasional sebesar 44,50 persen (BPS, 2014) dan terendah diantara provinsi di Pulau Kalimantan.

Sehubungan dengan latar belakang masalah, dalam penelitian ini melihat kekebalan tubuh yang dapat diperoleh dari pemberian imunisasi dasar dan pemberian ASI eksklusif maka diperlukan pengembangan model multivariat. Metode yang digunakan adalah model probit biner bivariat. Menggunakan model probit biner bivariat karena variabel respon berbentuk kategorik/kualitatif.

Adapun tujuan khusus dari penelitian ini adalah: (1) mendapatkan model terbaik dari penelitian; (2) mengetahui faktor-faktor yang signifikan berpengaruh terhadap pemberian imunisasi dasar dan ASI eksklusif di Provinsi Kalimantan Selatan berdasarkan model probit biner bivariat. Manfaat dari penelitian ini yaitu: (1) Sebagai bahan evaluasi pemerintah dalam menentukan variabel yang signifikan berpengaruh terhadap pemberian imunisasi dasar dan ASI eksklusif dalam upaya mengurangi angka kematian anak; (2) Mengembangkan keilmuan dan memberikan informasi mengenai model probit biner bivariat dalam melihat faktor-faktor yang berpengaruh terhadap pemberian imunisasi dasar dan ASI eksklusif.

METODOLOGI

Imunisasi Dasar dan ASI Eksklusif

Pada balita kekebalan tubuh dari suatu penyakit sangat diperlukan karena dapat mencegah dari kematian. Pada usia bayi hingga balita merupakan usia yang sangat rentan terhadap penyakit terutama yang diakibatkan oleh bakteri dan virus. Sehingga daya tahan tubuh yang kebal akan membuat balita terjaga dan terlindungi dari penyakit. Kekebalan tubuh pada balita dapat diperoleh dari pemberian imunisasi dasar atau vaksinasi dan pemberian ASI eksklusif.

Tabel 1. Jadwal Imunisasi Dasar

No	Umur	Imunisasi
1	0-7 hari	Hepatitis B1
2	< 1 bulan	BCG
3	2 bulan	Hepatitis B2, DPT 1, Polio 1
4	3 bulan	Hepatitis B3, DPT 2, Polio 2
5	4 bulan	DPT 3, Polio 3
6	9 bulan	Campak, Polio 4

Pemberian imunisasi dasar dan ASI eksklusif termasuk dalam determinan perilaku kesehatan. Dalam teori yang dikembangkan oleh Lawrence Green sejak 1980 yang dikenal dengan teori preced-proceed kesehatan seseorang atau masyarakat dipengaruhi oleh dua faktor pokok, yakni faktor perilaku (behavior causes) dan faktor diluar perilaku (non-behaviour causes). Selanjutnya faktor perilaku kesehatan seseorang dipengaruhi oleh tiga faktor utama yaitu (Notoatmodjo, 2014): (1) Faktor predisposisi (predisposing factors) yang komponennya antara lain faktor demografi, faktor struktur sosial, dan faktor keyakinan terhadap kesehatan; (2) Faktor pemungkin (enabling factors) yang komponennya antara lain sumber daya keluarga dan sumber daya masyarakat; (3) Faktor pendorong (renforcing factors), yang terwujud dalam sikap dan perilaku petugas kesehatan atau petugas lainnya yang merupakan kelompok-kelompok panutan dari perilaku masyarakat.

Model Probit Biner Bivariat

Model probit biner bivariat adalah model yang menggambarkan hubungan antara dua variabel respon yang berbentuk data kategorik biner dengan satu atau lebih variabel prediktor yang berbentuk data kategorik, data kontinu maupun gabungan data kategorik dan data kontinu. Asumsi yang digunakan dalam model probit biner bivariat adalah antar variabel respon memiliki hubungan. Dalam penelitian ini digunakan uji *chi-square* untuk melihat hubungan antar variabel respon. Untuk melihat hubungan antar variabel respon digunakan Uji Chi-Square (Ramachandran dan Tsokos, 2009).

Misal diberikan variabel respon Y_1 dan Y_2 dimana kedua variabel tersebut terbentuk dari variabel yang tidak teramati Y_1^* dan Y_2^*. Persamaan kedua variabel tersebut adalah sebagai berikut:

$$y_1^* = \boldsymbol{\beta}_1^T \mathbf{x} + \varepsilon_1 \quad \text{.....(1)}$$

dan

$$y_2^* = \boldsymbol{\beta}_2^T \mathbf{x} + \varepsilon_2 \quad \text{.....(2)}$$

dengan:

$$\mathbf{x} = \begin{bmatrix} 1 & x_1 & \cdots & x_p \end{bmatrix}$$

$$\boldsymbol{\beta}_1 = \begin{bmatrix} \beta_{10} & \beta_{11} & \cdots & \beta_{1p} \end{bmatrix}^T$$

$$\boldsymbol{\beta}_2 = \begin{bmatrix} \beta_{20} & \beta_{21} & \cdots & \beta_{2p} \end{bmatrix}^T$$

$\mathbf{x}, \boldsymbol{\beta}_1$ dan $\boldsymbol{\beta}_2$ adalah vektor berukuran $(p+1)\times 1$ dimana p adalah banyaknya variabel prediktor. Dalam model probit biner bivariat terdapat beberapa asumsi, antara lain:

1. $E(\varepsilon_1) = E(\varepsilon_2) = 0$
2. $Var(\varepsilon_1) = Var(\varepsilon_2) = 1$
3. $Cov(\varepsilon_1, \varepsilon_2) = \rho$

Dari asumsi pada ε_1 dan ε_2 sehingga kedua variabel respon mengikuti distribusi normal yang dapat dinotasikan menjadi $Y_1^* \sim N(\boldsymbol{\beta}_1^T \mathbf{x}, 1)$ dan $Y_2^* \sim N(\boldsymbol{\beta}_2^T \mathbf{x}, 1)$. Seperti halnya dengan model probit biner univariat, pembentukan kategori pada variabel respon model probit biner bivariat dengan menentukan *threshold* pada variabel respon yang tidak teramati. Misalnya pengkategorian tersebut adalah sebagai berikut:

a. Model $y_1^* = \boldsymbol{\beta}_1^T \mathbf{x} + \varepsilon_1$ dengan memisalkan *threshold* adalah γ sehingga pengkategoriannya adalah:
   $Y_1 = 0$ jika $y_1^* \le \gamma$ dan
   $Y_1 = 1$ jika $y_1^* > \gamma$
b. Model $y_2^* = \boldsymbol{\beta}_2^T \mathbf{x} + \varepsilon_2$ dengan memisalkan *threshold* adalah δ sehingga pengkategoriannya adalah:
   $Y_2 = 0$ jika $y_2^* \le \delta$ dan
   $Y_2 = 1$ jika $y_2^* > \delta$

Tabel 2. Tabel Frekuensi Dua Arah untuk Variabel Y_1 dan Y_2

Variabel Respon Y_1	Variabel Respon Y_2	
	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	Y_{00}	Y_{01}
$Y_1 = 1$	Y_{10}	Y_{11}

Metode *Maximum Likelihood Estimation* (MLE) digunakan dalam estimasi parameter model probit biner bivariat. Karena persamaan yang dihasilkan dari proses penurunan estimasi dengan MLE tidak *close formed* maka penyelesaian untuk mendapatkan estimasi parameter model dengan cara *Newton Raphson* (Ratnasari, 2012).

Terdapat dua pengujian signifikansi parameter model, yaitu pengujian signifikansi secara simultan dan parsial.

Hipotesis pengujian parameter secara simultan adalah sebagai berikut:

$H_0 : \beta_{11} = \beta_{12} = \cdots = \beta_{1p} = 0$ dan
$\beta_{21} = \beta_{22} = \cdots = \beta_{2p} = 0$

H_1 : paling sedikit ada satu $\beta_{uv} \neq 0$ dengan u=1,2 dan v=1,2,…,p

Statistik uji untuk pengujian parameter secara simultan (Ratnasari, 2012) terlihat pada persamaan 3 (Lampiran 1). Keputusan tolak H_0 pada tingkat signifikansi sebesar α jika $G^2 > \chi^2_{\alpha,df}$ dengan derajat bebas (db) adalah banyaknya parameter dibawah populasi dikurangi banyaknya parameter dibawah H_0 atau tolak H_0 jika *p-value* < α .

Adapun hipotesis dalam pengujian parameter secara parsial adalah sebagai berikut:

$H_0 : \beta_{uv} = 0$

$H_1 : \beta_{uv} \neq 0$ dengan u=1,2 dan v=1,2,…,p

Statistik uji untuk pengujian parameter secara parsial (Ratnasari, 2012) terlihat pada persamaan 4 (Lampiran 1): Keputusan menolak H_0 pada tingkat signifikansi sebesar α jika $G^2 > \chi^2_{\alpha,1}$, hal ini dikarenakan apabila $n \to \infty$ maka *likelihood ratio* G^2 asymtotik berdistribusi χ^2_1.

Multikolinieritas adalah kejadian adanya korelasi yang tinggi antar variabel bebas. Artinya ada korelasi yang tinggi antara $X_1, X_2, \ldots, X_p$ (Suharjo, 2013). Menurut Hocking (1996) salah satu cara pendeteksian multikolinieritas adalah dengan melihat besarnya nilai korelasi antar variabel prediktor. Jika nilai korelasi antar variabel prediktor kurang dari 0,95 maka dapat disimpulkan bahwa tidak terjadi multikolinieritas pada variabel prediktor.

AIC merupakan suatu kriteria kebaikan model dari parameter yang diestimasi berdasarkan metode maksimum *likelihood*

(Konishi dan Kitagawa, 2008). Semakin kecil nilai AIC maka model tersebut semakin baik. Nilai AIC diperoleh dari formula sebagai berikut:

$$AIC = -2\ln L\left(\hat{\theta}\right) + 2p \quad \cdots \quad (5)$$

dengan:

$L\left(\hat{\theta}\right)$ = nilai maksimum fungsi *likelihood*

p = banyaknya parameter

Sumber Data dan Variabel Penelitian

Penelitian ini menggunakan data sekunder yang berasal dari hasil Survei Sosial Ekonomi Nasional (SUSENAS) provinsi Kalimantan Selatan tahun 2013. Unit analisis yang akan diteliti adalah rumah tangga yang memiliki balita usia 2-4 tahun atau 24-59 bulan.

Variabel penelitian yang digunakan pada penelitian adalah sebagai berikut:

Tabel 3. Variabel Penelitian

Variabel	Nama Variabel	Kategori
Y_1	Pemberian Imunisasi Dasar	Imunisasi dasar tidak lengkap
		Imunisasi dasar lengkap
Y_2	Pemberian ASI Eksklusif	Tidak Eksklusif
		Eksklusif
X_1	Umur Ibu	-
X_2	Umur Perkawinan Pertama	-
X_3	Pendidikan Ibu	Tidak punya ijazah SD *(reference)*
		SD dan SMP sederajat ($D_{3.1}$)
		SMA dan PT sederajat ($D_{3.2}$)
X_4	Status Kerja Ibu	Ya *(reference)*
		Tidak (D_4)
X_5	Pekerjaan Bapak	Pertanian *(reference)*
		Non Pertanian (D_5)
X_6	Pendidikan Bapak	Tidak punya ijazah SD *(reference)*
		SD dan SMP sederajat ($D_{6.1}$)
		SMA dan PT sederajat ($D_{6.2}$)
X_7	Jumlah Anak	-
	Lahir Hidup	
X_8	Penolong Kelahiran Terakhir	Medis *(reference)*
		Non Medis (D_8)
X_9	Status Daerah	Perkotaan *(reference)*
		Perdesaan (D_9)

HASIL DAN PEMBAHASAN

Gambaran Pemberian Imunisasi Dasar dan ASI Eksklusif

Untuk mengurangi angka kematian anak dapat dilakukan dengan meningkatkan kekebalan tubuh pada anak. Kekebalan tubuh pada anak bisa diperoleh dengan pemberian imunisasi dasar yang lengkap dan ASI eksklusif. Pada tahun 2013 di Provinsi Kalimantan Selatan pemberian imunisasi dasar yang lengkap kepada anak berusia 2-4 tahun sebesar 71,89 persen. Belum optimalnya pemberian imunisasi kepada anak dapat menyebabkan daya tahan tubuh anak kurang sehingga anak mudah terserang penyakit. Tabel 4 merupakan penyebaran anak usia 2-4 tahun yang mendapatkan imunisasi dasar lengkap di kabupaten/kota di Provinsi Kalimantan Selatan.

Persebaran anak usia 2-4 tahun yang mendapat imunisasi dasar lengkap pada tahun 2013 menempatkan Kabupaten Tapin dengan persentase anak usia 2-4 tahun yang mendapatkan imunisasi dasar lengkap tertinggi yaitu sebesar 85,35 persen dan Kabupaten Tabalong dengan persentase anak usia 2-4 tahun yang mendapatkan imunisasi dasar lengkap terendah yaitu sebesar 41,38 persen. Persentase anak usia 2-4 tahun yang mendapat imunisasi dasar lengkap di setiap kabupaten/kota diatas 60 persen kecuali di Kabupaten Tabalong. Namun hal ini belum cukup untuk menurunkan angka kematian anak mengingat pentingnya imunisasi bagi tubuh anak.

Tabel 4. Persentase Anak Usia 2-4 tahun yang Mendapat Imunisasi Dasar Lengkap Tahun 2013

No.	Kabupaten/Kota	Persentase Imunisasi Dasar Lengkap
1.	Tanah Laut	84,83
2.	Kotabaru	76,32
3.	Banjar	71,09
4.	Barito Kuala	75,57
5.	Tapin	85,35
6.	Hulu Sungai Selatan	64,91
7.	Hulu Sungai Tengah	69,45
8.	Hulu Sungai Utara	75,07
9.	Tabalong	41,38
10.	Tanah Bumbu	67,83
11.	Balangan	67,60
12.	Kota Banjarmasin	74,86
13.	Kota Banjarbaru	66,85

Sumber data: SUSENAS 2013 (*data diolah*)

Selain pemberian imunisasi, untuk meningkatkan kekebalan tubuh anak juga bisa diperoleh dari pemberian ASI eksklusif. Karena dalam ASI mengandung zat gizi yang tidak terdapat dalam susu formula. Komposisi zat dalam ASI antara lain 88,1 persen air; 3,8 persen lemak; 0,9 persen protein; 7 persen laktosa serta 0,2 persen zat lainnya yang berupa DHA, DAA, shpynogelin dan zat gizi lainnya (Prasetyono, 2009). Karena banyaknya kandungan yang terdapat dalam ASI, maka ASI sangat dibutuhkan oleh anak untuk menjaga daya tahan tubuh dari serangan penyakit.

Tabel 5. Persentase Anak Usia 2-4 tahun yang Mendapat ASI Eksklusif Tahun 2013

No.	Kabupaten/Kota	Persentase ASI Eksklusif
1.	Tanah Laut	31,07
2.	Kotabaru	38,48
3.	Banjar	39,58
4.	Barito Kuala	36,00
5.	Tapin	28,18
6.	Hulu Sungai Selatan	38,03
7.	Hulu Sungai Tengah	32,39
8.	Hulu Sungai Utara	47,81
9.	Tabalong	32,64
10.	Tanah Bumbu	49,42
11.	Balangan	14,14
12.	Kota Banjarmasin	40,48
13.	Kota Banjarbaru	28,79

Pada Tabel 5 dapat dilihat bahwa persentase anak usia 2-4 tahun yang mendapatkan ASI eksklusif relatif rendah. Di setiap kabupaten/kota di Provinsi Kalimantan Selatan persentase anak usia 2-4 tahun yang mendapatkan ASI eksklusif dibawah 50 persen. Dibandingkan dengan pemberian imunisasi dasar, pemberian ASI eksklusif cenderung lebih rendah. Persentase terendah terdapat di Kabupaten Balangan dan persentase tertinggi terdapat di Kabupaten Tanah Bumbu. Kurangnya anak usia 2-4 tahun yang mendapatkan ASI eksklusif disebabkan banyak faktor. Oleh karena itu, program ASI eksklusif sebaiknya lebih digalakkan di seluruh penjuru daerah. Hal ini disebabkan manfaat yang terkandung dalam ASI sangat besar dalam menjaga daya tahan tubuh anak terhadap serangan penyakit.

Pemodelan Probit Biner Bivariat

Pemberian imunisasi dasar dan ASI eksklusif secara bersama-sama diduga dipengaruhi oleh umur ibu, umur perkawinan pertama ibu, pendidikan ibu, status kerja ibu, pekerjaan bapak, pendidikan bapak, jumlah anak lahir hidup, penolong kelahiran terakhir dan status daerah. Pemberian imunisasi dasar dibedakan menjadi dua kategori yaitu lengkap dan tidak lengkap sedangkan ASI eksklusif dibedakan menjadi dua kategori yaitu eksklusif dan tidak eksklusif. Untuk mengetahui adanya independensi antara pemberian imunisasi dasar dan ASI eksklusif dilakukan uji independensi dengan menggunakan uji Pearson Chi-Square (Agresti, 2002). Menurut Gani dan Amalian (2015), dalam penelitian bidang sosial tingkat signifikansi (α) sampai dengan 20% atau 0,20. Sehingga dalam penelitian ini menggunakan tingkat signifikansi (α) 10% atau 0,10.

Dengan uji Chi-Square didapatkan kesimpulan bahwa antara pemberian imunisasi dasar dan ASI eksklusif saling dependen. Karena nilai Chi-Square yang didapatkan sebesar 3,436 lebih besar dibandingkan dengan $\chi^2_{0,10;1} = 2,706$ atau dengan *p-value* sebesar 0,064 (nilai *p-value* kurang dari $\alpha = 10\%$).

Untuk mengidentifikasi adanya multikolinieritas antar variabel prediktor maka terlebih dahulu melihat korelasi antar variabel prediktor. Untuk melihat korelasi dengan melihat nilai koefisien korelasi *momen pearson, rank's spearman dan kendall's tau.* Berdasarkan hasil pengolahan terlihat bahwa nilai koefisien korelasi antar variabel prediktor tidak ada yang memiliki koefisisen korelasi yang sangat kuat. Menurut Hocking (1996) jika nilai korelasi antar variabel prediktor kurang dari 0,95 maka dapat disimpulkan bahwa tidak terjadi multikolinieritas pada variabel prediktor. Sehingga dapat disimpulkan bahwa tidak terjadi multikolinieritas antar variabel prediktor.

Dengan menggunakan metode *backward elimination* dan berdasarkan kriteria AIC menghasilkan model terbaik yang sama. AIC pada model terbaik adalah sebesar 2688,643 (Tabel 6 (Lampiran 2).

Persamaan model probit biner bivariat terbaik terlihat pada Lampiran 3.

Pengujian parameter secara simultan pada model terbaik berdasarkan nilai *wald chi-square* (G^2) sebesar 57,60 $(G^2 > \chi^2_{0,10;12} = 18,549)$ atau *p-value* sebesar 0,000 lebih kecil dari 0,10 yang dapat ditarik kesimpulan bahwa paling sedikit ada satu variabel prediktor yang signifikan terhadap variabel respon. Pengujian parameter secara parsial didapatkan variabel prediktor yang berpengaruh signifikan adalah variabel prediktor umur perkawinan pertama ibu (X_2) dan pekerjaan bapak (X_5) berpengaruh signifikan terhadap pemberian ASI eksklusif sedangkan variabel pendidikan ibu (X_3), penolong kelahiran terakhir (X_8) dan status daerah (X_9) berpengaruh signifikan terhadap pemberian imunisasi dasar.

Interpretasi Model Probit Biner Bivariat Terbaik

Untuk menginterpretasikan model probit biner bivariat, dimisalkan jika dalam sebuah rumah tangga umur perkawinan pertama ibu adalah 30 tahun $(X_2 = 30)$, pendidikan terakhir adalah perguruan tinggi $(D_{3.1} = 0 \text{ dan } D_{3.2} = 1)$, pekerjaan bapak di sektor non pertanian $(D_5 = 1)$, penolong kelahiran terakhir adalah bidan $(D_8 = 0)$ dan tinggal di perkotaan $(D_9 = 0)$ maka nilai $\hat{y}_1^*$ dan $\hat{y}_2^*$ terlihat pada Lampiran 4.

Dari persamaan tersebut, maka diperoleh nilai probabilitas sebagai berikut:

Tabel 7. Tabel Kontingensi Probabilitas (2×2) untuk Variabel Y_1 dan Y_2

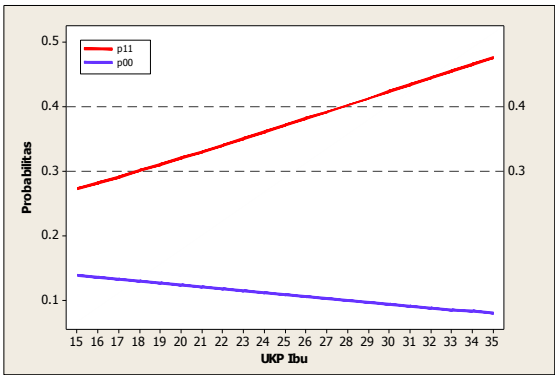
Variabel Y_1	Variabel Y_2		Total
	$Y_2 = 0$	$Y_2 = 1$	
$Y_1 = 0$	0,0939	0,0816	0,1755
$Y_1 = 1$	0,3993	0,4252	0,8245
Total	0,4932	0,5068	1

Setelah diperoleh nilai probabilitasnya maka dapat disimpulkan bahwa rumah tangga tersebut mempunyai probabilitas sebesar 0,4252 untuk masuk dalam kategori rumah tangga yang memberikan imunisasi dasar lengkap dan memberikan ASI eksklusif.

Efek marginal pada model probit biner bivariat digunakan untuk mengetahui besarnya pengaruh perubahan suatu variabel prediktor terhadap variabel respon dengan asumsi variabel lainnya konstan. Pada model probit biner bivariat terbaik, dimisalkan jika dalam sebuah rumah tangga umur perkawinan pertama ibu adalah 30 tahun $(X_2 = 30)$, pendidikan terakhir adalah perguruan tinggi $(D_{3.1} = 0 \text{ dan } D_{3.2} = 1)$, pekerjaan bapak di sektor non pertanian $(D_5 = 1)$, penolong

kelahiran terakhir adalah bidan $(D_8 = 0)$ dan tinggal di perkotaan $(D_9 = 0)$.

Pada variabel umur perkawinan pertama (X_2) efek marginalnya terhadap $\hat{p}_{11}$ sebesar 0,0106 yang berarti bahwa perubahan umur perkawinan pertama (X_2) sebesar satu satuan akan meningkatkan 0,0106 terhadap probabilitas rumah tangga yang memberikan imunisasi dasar lengkap dan memberikan ASI eksklusif. Namun secara statistik, umur perkawinan ibu hanya signifikan pada variabel respon Y_2 yaitu pemberian ASI eksklusif. Hal ini sesuai dengan hasil penelitian Santosa (2009) menunjukkan bahwa umur perkawinan pertama ibu berpengaruh signifikan terhadap pemberian ASI eksklusif pada rumah tangga miskin di Provinsi Sulawesi Tengah.



Gambar 1. Grafik Hubungan antara Probabilitas $\hat{p}_{11}$, $\hat{p}_{00}$ dan UKP Ibu

Berdasarkan Gambar 1 terlihat bahwa semakin tinggi umur perkawinan pertama ibu maka semakin tinggi probabilitas rumah tangga tersebut untuk memberikan imunisasi dasar lengkap dan ASI eksklusif. Jika umur perkawinan pertama ibu adalah 18 tahun maka probabilitas rumah tangga tersebut memberikan imunisasi dasar lengkap dan ASI eksklusif sebesar 30 persen. Jika umur perkawinan pertama ibu adalah 28 tahun maka probabilitas rumah tangga tersebut memberikan imunisasi dasar lengkap dan ASI eksklusif sebesar 40 persen. Sehingga dapat disimpulkan bahwa umur perkawinan ibu yang rendah dapat mengurangi perilaku pemberian imunisasi dasar lengkap dan ASI eksklusif. Oleh karena itu perlu dilakukan penyuluhan tentang usia perkawinan pertama seorang ibu. Mudanya usia perkawinan pertama ibu diduga mengakibatkan kurangnya pengetahuan seorang ibu dalam pentingnya memberikan imunisasi dasar yang lengkap dan ASI eksklusif.

Pada Tabel 8 (Lampiran 5) efek marginal variabel pendidikan ibu $(D_{3.1})$ terhadap $\hat{p}_{11}$ adalah 0,0946. Hal ini berarti bahwa rumah tangga yang pendidikan ibunya adalah SD/sederajat atau SMP/sederajat, probabilitas untuk memberikan imunisasi dasar lengkap dan ASI eksklusif lebih besar 0,0946 dibandingkan dengan rumah tangga yang pendidikan ibunya tidak punya ijazah. Sedangkan efek marginal variabel pendidikan ibu $(D_{3.2})$ terhadap $\hat{p}_{11}$ adalah 0,0500. Hal ini berarti bahwa rumah tangga yang pendidikan ibunya adalah SMA/sederajat atau PT/sederajat, probabilitas untuk memberikan imunisasi dasar lengkap dan ASI eksklusif lebih besar 0,0500 dibandingkan dengan rumah tangga yang pendidikan ibunya tidak punya ijazah. Secara statistik, variabel pendidikan ibu hanya signifikan pada variabel respon Y_1 yaitu pemberian imunisasi dasar. Hal ini sesuai dengan hasil penelitian Wardhana (2001) bahwa ibu berpendidikan rendah status imunisasinya cenderung tidak lengkap. Probabilitas memberikan imunisasi dasar lengkap dan ASI eksklusif tertinggi pada rumah tangga yang pendidikan terakhir ibu adalah SD/sederajat atau SMP/sederajat $(D_{3.1} = 1 \text{ dan } D_{3.2} = 0)$, dimana variabel yang lain konstan yaitu sebesar 46,92 persen. Namun probabilitas ini tidak berbeda jauh jika dibandingkan dengan ibu yang pendidikan terakhirnya SMA dan PT.

Efek marginal variabel pekerjaan bapak (D_5) terhadap $\hat{p}_{11}$ adalah 0,0393. Hal ini berarti bahwa rumah tangga yang sektor pekerjaan bapaknya adalah non pertanian,

probabilitas rumah tangga tersebut untuk memberikan imunisasi dasar lengkap dan ASI eksklusif lebih besar 0,0393 dari rumah tangga yang pekerjaan bapaknya di sektor non pertanian. Secara statistik, variabel pekerjaan bapak hanya signifikan pada variabel respon Y_2 yaitu pemberian ASI eksklusif. Probabilitas rumah tangga dengan pekerjaan bapak di sektor pertanian dimana variabel prediktor yang lain adalah konstan, probabilitas untuk memberikan imunisasi dasar lengkap dan ASI eksklusif adalah 38,37 persen. Sedangkan untuk pekerjaan bapak di sektor non pertanian adalah 42,52 persen. Menurut Litman dan Weiss (1994), wanita-wanita yang menyusui bayinya adalah wanita yang disusui ketika masih bayi, mempunyai teman yang menyusui bayinya, dan menerima dukungan dari tenaga kesehatan dan suaminya.

Efek marginal variabel penolong kelahiran terakhir (D_8) terhadap $\hat{p}_{11}$ adalah -0,0862. Hal ini berarti bahwa rumah tangga yang penolong kelahiran terakhir dengan non medis, probabilitas rumah tangga tersebut untuk memberikan imunisasi dasar lengkap dan ASI eksklusif lebih kecil 0,0862 dibandingkan rumah tangga yang penolong kelahirannya dengan medis. Secara statistik, variabel penolong kelahiran hanya signifikan pada variabel Y_1 yaitu pemberian imunisasi dasar. Hal ini diperkuat oleh penelitian Sandra (2010) yang menunjukkan bahwa penolong kelahiran berpengaruh signifikan terhadap status imunisasi dasar pada anak dan hasil penelitian Maryati (2009) menyatakan bahwa penolong kelahiran berpengaruh signifikan terhadap pemberian ASI eksklusif. Probabilitas rumah tangga dengan penolong kelahiran terakhir adalah medis akan memberikan imunisasi dasar lengkap dan ASI eksklusif adalah 42,52 persen dimana variabel prediktor yang lain konstan. Dari hasil tersebut maka perlu dilakukan pemerataan tenaga kesehatan di semua wilayah. Sehingga dengan meratanya tenaga kesehatan dapat memberikan pengetahuan dan kesadaran masyarakat terhadap perilaku kesehatan.

Efek marginal variabel status daerah (D_9) terhadap $\hat{p}_{11}$ sebesar -0,0638. Hal ini berarti bahwa rumah tangga yang tinggal di daerah pedesaan, probabilitas kategori rumah tangga yang memberikan imunisasi dasar lengkap dan ASI eksklusif lebih kecil 0,0638 dibandingkan dengan rumah tangga yang tinggal di daerah perkotaan. Secara statistik, variabel status daerah hanya signifikan pada variabel Y_1 yaitu pemberian imunisasi dasar lengkap. Probabilitas rumah tangga yang tinggal di daerah perkotaan untuk memberikan imunisasi dasar dan ASI eksklusif adalah sebesar 42,52 persen dimana variabel prediktor yang lain konstan. Dari hasil tersebut maka pemerintah perlu memperhatikan aspek sarana prasarana kesehatan, sehingga tidak ada kesenjangan fasilitas kesehatan antara daerah perkotaan dan pedesaan. Hal ini sesuai dengan penelitian Idwar (2000) yang menyatakan bahwa ada hubungan antara status imunisasi dengan jarak ke fasilitas kesehatan. Seorang ibu akan mencari pelayanan kesehatan yang terdekat dengan rumahnya karena pertimbangan aktivitas lain yang harus diselesaikan. Hasil penelitian Purnamawati (2003) juga menyatakan bahwa status daerah tempat tinggal berpengaruh signifikan terhadap pola pemberian ASI.

Ketepatan Klasifikasi Model Probit Biner Bivariat Terbaik

Ketepatan klasifikasi adalah ketepatan antara data aktual dengan hasil prediksinya. Berdasarkan model probit biner bivariat terbaik, ketepatan klasifikasi sebesar 40,89 persen. Ketepatan klasifikasi yang kecil diduga karena dalam penelitian ini, tidak ada variabel prediktor yang berpengaruh ke semua variabel respon. Sehingga efek marginal yang dihasilkan dari variabel prediktor yang signifikan cenderung berpengaruh ke salah satu variabel respon.

KESIMPULAN DAN SARAN

Kesimpulan

Uji Chi-Square untuk tabel kontingensi (2×2) menunjukkan bahwa ada hubungan yang signifikan antara pemberian dasar dan ASI eksklusif. Dengan menggunakan model probit biner bivariat menghasilkan model terbaik dengan nilai AIC sebesar 2688,643 dengan variabel yang signifikan dalam model berdasarkan pemberian imunisasi dasar dan ASI eksklusif adalah variabel umur perkawinan pertama ibu (X_2), pendidikan ibu (X_3), pekerjaan bapak (X_5), penolong kelahiran terakhir (X_8) dan status daerah (X_9). Namun secara statistik, pengujian parameter secara parsial didapatkan variabel prediktor yang berpengaruh signifikan adalah variabel prediktor umur perkawinan pertama ibu (X_2) dan pekerjaan bapak (X_5) berpengaruh signifikan terhadap pemberian ASI eksklusif sedangkan variabel pendidikan ibu (X_3), penolong kelahiran terakhir (X_8) dan status daerah (X_9) berpengaruh signifikan terhadap pemberian imunisasi dasar. Dan untuk ketepatan klasifikasi berdasarkan model probit biner bivariat terbaik, ketepatan klasifikasi sebesar 40,89 persen.

Saran

Dengan mempertimbangkan hasil penelitian ini sebaiknya pemerintah Provinsi Kalimantan Selatan khususnya lebih memperhatikan aspek sarana kesehatan dan prasarana kesehatan yang mampu menjangkau daerah terpencil serta memperhatikan aspek pendidikan untuk meningkatkan kualitas sumber daya manusia. Misalnya dengan membangun puskesmas dan sekolah-sekolah di daerah yang sulit dijangkau. Sehingga meskipun tinggal di daerah yang sulit dijangkau, kualitas sumber daya manusia tetap terjaga. Dalam penelitian ini, nilai koefisien korelasi antar variabel respon sangat rendah meskipun asumsi dependensi antar variabel respon terpenuhi. Sehingga untuk penelitian selanjutnya dalam model probit bivariat, selain memenuhi asumsi dependensi juga diperlukan nilai koefisien korelasi yang tinggi.

DAFTAR PUSTAKA

Agresti, A. 2002. *Categorical Data Analysis, Second Edition*. Wiley-Inter-Science A John Wiley & Sons, Inc.

BPS. 2013. *Statistik Kesejahteraan Rakyat 2012*. BPS, Jakarta.

BPS. 2014. *Statistik Kesejahteraan Rakyat 2013.* BPS, Jakarta.

BPS, BKKBN, Kementrian Kesehatan, dan Measure DHS. 2012. *Laporan Pendahuluan Survei Demografi dan Kesehatan 2012.* Jakarta.

Bokosi, F. K. 2007. Household Poverty Dynamics in Malawi: A Bivariate Probit Analysis, *Journal of Applied Sciences*: Asian Network for Scientific Information, Vol. 7, No. 2, pp. 573-578.

Chen, G., dan Hamori, S. 2010. Bivariate Probit Analysis of Differences of Between Male and Female Formal Employment in Urban Cina, *Journal of Asian Economics*: Vol. 21, pp. 494-501.

Dudewics, E. J. dan Mishra, S. N. 1988. *Modern Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons.

Gujarati, D. N. dan Porter, D. C. 2013. *Dasar-dasar Ekonometrika*. Edisi Kelima Buku 2, Jakarta: Salemba Empat.

Hidayat, A. Aziz Alimul. 2008. *Pengantar Ilmu Kesehatan Anak untuk Pendidikan Kebidanan*. Jakarta: Salemba Medika.

Hocking, R. 1996. *Methods and Application of Linier Models.* John Wiley and Sons, Inc., New York.

Hosmer, D. W. dan Lemeshow, S. 2000. *Applied Logistic Regression, Second Edition*. Wiley-Interscience A John Wiley & Sons, Inc.

Konishi, S. dan Kitagawa, G. 2008. *Information Criteria and Statistical Modeling*. Springer Science + Business Media, LCC, New York.

Kutner, M.H., Nachtsheim, C.J., dan Nether, J. 2008. *Applied Linear Regression Model.* McGraw-Hill Companies. New York.

Mahayu, P. 2014. *Imunisasi dan Nutrisi (Panduan Pemberian Imunisasi dan Nutrisi pada Bayi, Batita, Balita, dan Manfaatnya).* Yogyakarta: Bukubiru.

Nugraha, J. 2010. *Pemodelan Pilihan Diskrit Menggunakan Model Probit dan Model Fixed Logit pada Respon Multivariat*. Disertasi, Universitas Gajah Mada, Yogyakarta.

Prasetyono, D. S. 2009. *Buku Pintar ASI Eksklusif*. Yogyakarta: Diva Press.

Ramachandran, K.M. dan Tsokos, C. P. 2009. *Mathematical Statistics with Appications.* Elseiver Inc, USA.

Ratnasari, V. 2012. *Estimasi Parameter dan Uji Signifikansi Model Probit Bivariat.* Disertasi, Institut Teknologi Sepuluh Nopember, Surabaya.

Roesli, U. 2000. *Mengenal ASI Eksklusif.* Jakarta: Niaga Swadaya.

Suharjo, B. 2013. *Statistika Terapan (Disertai Contoh Aplikasi dengan SPSS).* Yogyakarta: Graha Ilmu.

Wahyudi, C. D. 2014. *Model Kemiskinan Perdesaan dan Perkotaan dengan Pendekatan Garis Kemiskinan Menggunakan Regresi Probit Biner Bivariat di Provinsi Bengkulu.* Tesis, Institut Teknologi Sepuluh Nopember, Surabaya.

LAMPIRAN

Lampiran 1

$$G^2 = 2\sum_{i=1}^{n}\left[y_{11i} \ln\left(\frac{\hat{p}_{2i} - \hat{p}_{01i}}{\hat{p}^{*}_{2i} - \hat{p}^{*}_{01i}} \right) + y_{10i} \ln\left(\frac{\hat{p}_{1i} - \hat{p}_{2i} + \hat{p}_{01i}}{\hat{p}^{*}_{1i} - \hat{p}^{*}_{2i} + \hat{p}^{*}_{01i}} \right) + y_{10i} \ln\left(\frac{\hat{p}_{01i}}{\hat{p}^{*}_{01i}} \right) + y_{00i} \ln\left(\frac{1 - \hat{p}_{1i} - \hat{p}_{01i}}{1 - \hat{p}^{*}_{1i} - \hat{p}^{*}_{01i}} \right) \right] \quad (3)$$

$$G^2 = 2\sum_{i=1}^{n}\left[y_{11i} \ln\left(\frac{\hat{p}_{11i}}{\hat{p}^{**}_{11i}} \right) + y_{10i} \ln\left(\frac{\hat{p}_{10i}}{\hat{p}^{**}_{10i}} \right) + y_{10i} \ln\left(\frac{\hat{p}_{01i}}{\hat{p}^{**}_{01i}} \right) + y_{00i} \ln\left(\frac{\hat{p}_{00i}}{\hat{p}^{**}_{00i}} \right) \right] \quad (4)$$

Lampiran 2

Tabel 6. Nilai Koefisien, *Standar Error* dan *p-value* pada Masing-masing Parameter Model Probit Biner Bivariat Terbaik

Variabel Prediktor		Imunisasi Dasar			ASI Eksklusif		
		Coeff	*Std. Err*	*p-value*	*Coeff*	*Std. Err*	*p-value*
Umur Perkawinan Pertama Ibu (X_2)	-	0,0055	0,0124	0,659	0,0301	0,0116	0,009
Pendidikan Ibu (X_3)	SD atau SMP sederajat ($D_{3.1}$)	0,3841	0,1224	0,002	0,1426	0,1260	0,258
	SMA atau PT sederajat ($D_{3.2}$)	0,4038	0,1539	0,009	-0,0003	0,1527	0,999
Pekerjaan Bapak (X_5)	Sektor Non Pertanian (D_5)	-0,1041	0,0947	0,272	0,1588	0,0906	0,080
Penolong Kelahiran Terakhir (X_8)	Non Medis (D_8)	-0,3576	0,1106	0,001	-0,1272	0,1127	0,259
Status Daerah (X_9)	Pedesaan (D_9)	-0,2310	0,0963	0,016	-0,1069	0,0889	0,229
Konstanta	-	0,4679	0,2711	0,084	-1,0443	0,2579	0,000

Lampiran 3

$$\hat{y}^{*}_1 = 0,4679 + 0,0055X_2 + 0,3841D_{3.1} + 0,4038D_{3.2} - 0,1041D_5 - 0,3576D_8 - 0,2310D_9$$

dan

$$\hat{y}^{*}_2 = -1,0443 + 0,0301X_2 + 0,1426D_{3.1} - 0,0003D_{3.2} + 0,1588D_5 - 0,1272D_8 - 0,1069D_9$$

Lampiran 4

$$\hat{y}^{*}_1 = 0,4679 + 0,0055(30) + 0,3841(0) + 0,4038(1) - 0,1041(1) - 0,3576(0) - 0,2310(0) = 0,9326$$

$$\hat{y}^{*}_2 = -1,0443 + 0,0301(30) + 0,1426(0) - 0,0003(1) + 0,1588(1) - 0,1272(0) - 0,1069(0) = 0,0172$$

Lampiran 5

Tabel 8. Efek Marginal dan Probabilitas Variabel Bebas Kategorik Terhadap p_{11}

Nama Variabel	**Kategori**	**Efek Marginal Terhadap p_{11}**	**Probabilitas Terhadap p_{11}**
Pendidikan Ibu	Tidak punya ijazah SD *(reference)*	-	0,3655
	SD dan SMP sederajat $(D_{3.1})$	0,0946	0,4692
	SMA dan PT sederajat $(D_{3.2})$	0,0500	0,4252
Pekerjaan Bapak	Pertanian *(reference)*	-	0,3837
	Non Pertanian (D_5)	0,0393	0,4252
Penolong Kelahiran Terakhir	Medis *(reference)*	-	0,4252
	Non Medis (D_8)	-0,0862	0,3368
Status Daerah	Perkotaan *(reference)*	-	0,4252
	Perdesaan (D_9)	-0,0638	0,3609

METODE CLUSTER MENGGUNAKAN KOMBINASI ALGORITMA *CLUSTER K-PROTOTYPE* DAN ALGORITMA GENETIKA UNTUK DATA BERTIPE CAMPURAN

CLUSTER METHOD USING A COMBINATION OF CLUSTER K-PROTOTYPE ALGORITHM AND GENETIC ALGORITHM FOR MIXED DATA

Rani Nooraeni
Sekolah Tinggi Ilmu Statistik

Masuk tanggal: 05-12-2015, revisi tanggal: 15-01-2016, diterima untuk diterbitkan tanggal: 19-01-2016

Abstrak

Clustering adalah salah metode utama pada *data mining* yang berguna untuk mengeksplorasi data. Membagi suatu data set berukuran besar ke dalam *cluster* yang sehomogen mungkin adalah tujuan dalam metode *data mining*. Salah satu metode clustering konvensional yaitu algoritma *K-Means* efisien untuk dataset berukuran besar dan tipe data numerik tapi tidak untuk data kategorikal. Algoritma *K-Prototype* menghilangkan keterbatasan pada data numerik tapi dapat juga digunakan pada data kategorikal. Namun solusi yang dihasilkan oleh kedua algoritma tersebut merupakan solusi lokal optimal dimana salah satu penyebabnya adalah penentuan pusat *cluster* awal. Untuk menghadapi masalah tersebut maka algoritma genetika menjadi salah satu usulan yang dapat digunakan untuk mengoptimalkan hasil pengclusteran dengan *K-Prototype*. Hasil dari penelitian menunjukkan optimasi pusat *cluster* dengan algoritma genetika berhasil meningkatkan akurasi hasil cluster dengan *K-Prototype*.

Kata kunci : *Data Mining*, Analisis Cluster, Data Campuran, Algoritma *K-Prototype*, Algoritma Genetika

Abstract

Clustering is one of the main methods in data mining that useful to explore the data. One conventional clustering methods namely the K -Means algorithm efficient for large dataset and numeric data types but not for categorical data type. K-prototype algorithm eliminates the limitations of the numerical data but can also be used on categorical data. But the solutions generated by the algorithm is a local optimal solution in which one of the causes is the determination of the initial cluster's center. Deal with these problems, the genetic algorithm was proposed for solving this global optimasitation problem. The results of the study indicate that the cluster's center optimization with genetic algorithm success to improve the accuracy of the results of the cluster with K–Prototype algorithm.

Keywords : *Data Mining, Cluster Analysis, Mixed Data, K-Prototype Algorithm, Genetic Algorithm*

PENDAHULUAN

Clustering merupakan salah satu metode dalam *data mining*. Metode cluster dalam *data mining* berbeda dengan metode konvensional yang biasa digunakan untuk pengelompokkan. Perbedaannya adalah *data mining* memiliki dimensi data yang tinggi yaitu bisa terdiri dari puluhan ribu atau jutaan record dengan puluhan ataupun ratusan atribut. Selain itu pada *data mining* data bisa terdiri dari tipe data campuran seperti data numerik dan kategorikal.

Metode cluster standar hierarki dapat menangani data bertipe campuran numerik dan kategorikal tetapi ketika data berukuran besar maka timbul masalah dalam hal efisiensi waktu penghitungan. K-Means dapat diterapkan pada data berukuran besar tetapi efisien untuk data bertipe numerik. Hal ini disebabkan pada K-means optimasi *cost function* menggunakan jarak euclidean yang mengukur jarak antara data poin dengan rata-rata cluster. Meminimalkan fungsi *cos*t dengan menghitung rata-rata cocok digunakan untuk data numerik.

Salah satu dataset yang terdiri dari ratusan atribut adalah data hasil Pendataan Potensi Desa (PODES). Data PODES

menyediakan data potensi/keadaan pembangunan hingga level terendah yaitu desa/kelurahan. Data PODES meliputi informasi mengenai keadaan sosial, ekonomi, sarana dan prasarana, serta potensi yang dimiliki suatu desa/kelurahan. Unit observasi dalam PODES adalah desa yang merupakan level wilayah terendah, sedangkan atribut yang terkandung dalam PODES jika dirinci jumlahnya maka jumlahnya mencapai 593 atribut dengan jumlah record mencapai 77.961 unit desa/kelurahan (Tabel 1 (Lampiran 1)).

Berdasarkan karakteristik tersebut maka struktur data PODES merupakan struktur data kompleks. Untuk mengeksplorasi data kompleks diperlukan suatu metode yang tepat sehingga dapat diperoleh hasil yang lebih akurat dengan informasi yang mendalam dan berharga sekaligus dapat menjadi suatu pengetahuan baru dan bermanfaat. Metode analisis yang dapat diterapkan pada kasus tersebut adalah metode analisis *data mining*.

Struktur dataset PODES sejalan dengan realita data yang tersedia dalam kehidupan sehari-hari, dimana data yang tersedia tidak hanya terdiri dari data numerik saja atau data kategorikal saja namun terdapat juga data bertipe campuran. Tabel 1 memperlihatkan struktur dataset PODES secara rinci. Begitu juga dengan data hasil sensus/survey biasanya memiliki tipe data campuran. Hal ini dikarenakan tidak semua persoalan atau pertanyaan bisa dijawab dengan suatu nilai berskala ukur. Oleh karena itu, diperlukan suatu metode analisis yang dapat digunakan untuk menganalisis data bertipe campuran.

Teknik analisis yang dapat menggambarkan karakteristik sekelompok wilayah berdasarkan satu atau lebih variabel salah satunya adalah teknik *clustering*. Dengan teknik clustering akan diperoleh kelompok-kelompok desa, dimana setiap desa yang berada dalam satu kelompok memiliki karakteristik yang mirip dan dengan desa pada kelompok lain sangat berbeda karakteristiknya. Dengan teknik tersebut dapat mempermudah dalam melihat profil suatu desa berdasarkan variabel ciri yang mendominasinya. Dan dengan teknik ini juga mempermudah pengguna data untuk melihat perbandingan karakteristik suatu desa terhadap desa lainnya.

Salah satu metode *clustering konvensional* yang biasa digunakan dalam teknik pengclusteran dan efisien digunakan pada data berukuran besar adalah algoritma *K-Means.* Akan tetapi metode ini cocok untuk data yang bertipe numerik dan tidak efektif jika digunakan untuk tipe data kategorikal. Hal ini dikarenakan cost function yang dihitung menggunakan jarak euclidean hanya cocok untuk data bertipe numerik (Jayaraj, 2014).

Menghadapi kendala tersebut Huang mengusulkan sebuah algoritma yang disebut dengan algoritma *K-Prototype*, untuk menangani masalah *clustering* dengan data bertipe campuran numerik dan kategorikal. *K-Prototype* adalah salah satu metode *clustering* yang berbasis *partitioning*. Algoritma ini merupakan hasil pengembangan dari algoritma cluster *K-Means* untuk menangani *clustering dengan* atribut data bertipe campuran numerik dan kategorikal. *K-Protoype* memiliki keunggulan karena algoritmanya yang tidak terlalu kompleks dan mampu menangani data yang besar serta lebih baik dibandingkan dengan algoritma yang berbasis hierarki (Huang, 1997). Algoritma K-prototype ini telah mendasari banyak penelitian yang menghadapi data besar bertipe campuran seperti penelitian yang dilakukan oleh D.T. Pham (2011), Jengyou He (2011).

Namun demikian baik metode K-Means maupun K-Prototype menghasilkan solusi yang lokal optimum. Kedua metode tersebut sensitif terhadap penentuan inisialisasi posisi pusat cluster dan cenderung mengalami *konvergensi prematur* sehingga menghasilkan solusi optimum lokal akibatnya hasil peng*cluster*an bisa berbeda jika menggunakan inisial pusat cluster random yang berbeda (Dash & Dash, 2012) dan (Feng & Wang, 2011). Begitu juga Duc truong Pham, 2011 dalam penelitiannya mengatakan bahwa proses algoritma K-Means dan K-Prototype seringkali

konvergen pada lokal minimum dan bukan pada global minimum.

Dalam penelitiannya, Huang mengusulkan untuk menerapkan teknik pengoptimasi yang dapat mengatasi masalah optimum lokal, salah satunya dengan menerapkan algoritma genetika (Huang, 1997). Algoritma genetika (GA) merupakan suatu alat optimasi yang dapat digunakan untuk mengotimalkan hasil dari suatu metode. GA dapat digunakan untuk mengoptimalkan inisial center cluster sehingga dapat diperoleh hasil pengclusteran yang global optimum.

Penelitian lainnya yang menunjukkan efektifitas GA dalam pengclusteran misalnya penelitian yang dilakukan Li Jie (2003) menerapkan algoritma genetika dalam pengclusterannya, dan menghasilkan kesimpulan bahwa algoritma genetika efektif dalam menangani data yang kompleks baik dari sisi jumlah record maupun dari jumlah cluster. Begitupula dalam penelitian Rajashree Dash (2012), pengclusteran dengan algoritma genetika menghasilkan cluster yang lebih optimal dibandingkan algoritma K-Means. Sedangkan Dianhu Cheng (2014) mengkombinasikan antara algoritma K-Means dengan algoritma genetika untuk menggabungkan kelebihan dari kedua metode tersebut untuk memperoleh jumlah cluster yang optimal.

Berdasarkan berbagai macam manfaat tersebut maka selanjutnya algoritma genetika akan digunakan dalam penelitian ini untuk mengoptimasi algoritma *K-prototype* yang dapat melakukan pengclusteran pada data bertipe campuran.

Maksud penelitian ini adalah menerapkan metode cluster algoritma *k-prototype* yang dioptimalkan dengan algoritma genetika pada data beratribut campuran dengan tujuan memperoleh solusi optimum global sehingga hasil peng-*cluster*-an menjadi lebih baik dan akurat.

Manfaat yang diharapkan dari penelitian ini adalah kontribusi dalam bidang keilmuan berkaitan dengan pengelompokan wilayah dengan atribut bertipe campuran sehingga dapat menghasilkan kelompok-kelompok desa yang lebih baik, lebih erat kesamaan karakteristiknya, dan lebih akurat dengan kombinasi algoritma *cluster K-Prototype* dan algoritma genetika. Sehingga akan bermanfaat untuk berbagai kalangan baik pemerintah, akademisi dan masyarakat luas dalam memperoleh gambaran/karakteristik dan kondisi suatu wilayah hingga level desa yang lengkap dan akurat dan mendukung dalam menentukan suatu kebijakan agar lebih tepat sasaran.

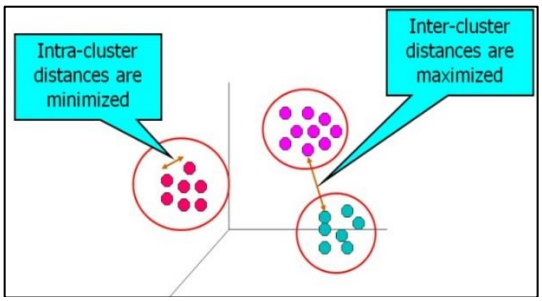
METODOLOGI

Tinjauan Referensi

Clustering merupakan salah satu metode utama pada *data mining*. Tipe data yang dapat dikerjakan dengan metode *data mining* adalah data kompleks, yaitu data yang terdiri dari puluhan ribu atau ratusan ribu *record* dan puluhan atau ratusan atribut. Terdapat berbagai algoritma pengclusteran konvensional yang umum digunakan namun salah satu algoritma yang efisien untuk data berukuan besar adalah algoritma *K-Means*. Keterbatasa *K-means* terhadap data numerik dikembangkan menjadi algoritma *K-prototype*. Masalah dari algoritma *K-Means* dan Algoritma *K-Prototype* adalah hasil ukuran similaritas yang optimum lokal (Huang, 1997). Untuk mengatasi masalah tersebut telah dilakukan pengembangan metode oleh para peneliti. Gambar 1 menunjukkan peta penelitian yang telah dilakukan oleh peneliti lain terkait dengan algoritma *clustering* konvensional, algoritma *clustering* untuk data campuran, algoritma genetika, dan kombinasi antar algoritma yang menjadi rujukan penelitian ini (Gambar 1 (Lampiran 2)).

Clustering adalah pengelompokkan sekumpulan objek yang mirip dengan properti yang sama dalam satu kelompok dan tidak mirip terhadap objek di kelompok lainnya. *Clustering* dan *Classification* adalah dua teknik utama dalam *data mining* yang diikuti kemudian

oleh *association rules*, prediksi, estimasi, dan regresi. (Han & Kamber, 2006), *Clustering* dikenal sebagai *unsupervised learning* karena tidak terdapat informasi label kelas sehingga *clustering* merupakan *learning by obseravtion* daripada *learning by examples.*



Gambar 2. Ilustrasi *Clustering*

Metode analisis cluster membutuhkan suatu ukuran ketakmiripan (jarak) yang didefinisikan untuk setiap pasang objek yang akan dikelompokan. Jarak yang biasa digunakan dalam analisis penggerombolan diantaranya:

1) Ukuran Data numerik, ukuran yang umum digunakan untuk data bertipe numerik adalah ukuran jarak euclidean, sedangkan ukuran lainnya adalah mahalanobis, Manhatan, minkowski, chebyshev dan lain-lain.
2) Ukuran untuk data kategorikal, terdapat ukuran rasio ketidakcocokan, Goodal3 similarity, gambaryan similaruty, dan lain-lain.

Ukuran Kesamaan (*Similarity Measure*)

Bentuk umum ukuran kesamaan dinyatakan sebagai berikut

$$d(X_i, Z_l) = \sum_{j=1}^{m} \delta(x_{ij}, z_{lj}) \quad (1)$$

$z_l = [z_{i1}, z_{i2}, \dots, z_{im}]^T$ adalah prototype untuk cluster l. Ukuran kesamaan untuk atribut numerik dikenal dengan jarak euclidean ditunjukkan dalam persamaan (2) berikut ini

$$d(X_i, Z_l) = \left(\sum_{j=1}^{m_r} \left(x_{ij}^r - z_{lj}^r\right)^2\right)^{1/2} \quad (2)$$

${x_{ij}}^r$ adalah nilai pada atribut numeric j, ${z_{lj}}^r$ adalah rata-rata atau prototype atribut numerik ke j cluster l. m_r adalah jumlah atribut numerik.

Sedangkan ukuran kesamaan untuk data kategorikal adalah

$$d(X_i, Z_l) = \gamma_l \sum_{j=l+1}^{m_c} \delta\left(x_{ij}^c, z_{lj}^c\right) \quad (3)$$

Dimana *simple matching similarity measure* untuk data kategorikal adalah

$$\delta\left(x_{ij}^c, z_{lj}^c\right) = \begin{cases} 0 & \left(x_{ij}^c = z_{lj}^c\right) \\ 1 & \left(x_{ij}^c \neq z_{lj}^c\right) \end{cases} \quad (4)$$

γ_l adalah bobot untuk atribut kategori pada cluster l yang nilainya merupakan nilai standar deviasi untuk atribut numerik pada masing-masing cluster. ketika ${x_{ij}}^c$ adalah nilai atribut kategorikal , z_{lj}^c adalah modus atribut ke j cluster l. m_c adalah jumlah atribut kategorikal.

He, memodifikasi *simple matching similarity measure* menjadi persamaan (5) untuk meningkatkan kemiripan objek dalam cluster dengan atribut kategorikal sehingga hasil pengclusteran menjadi lebih baik. Jika

$$\delta\left(x_{ij}^c, z_{lj}^c\right) = \begin{cases} 1 - \omega(x_{ij}^c, l) & \left(x_{ij}^c = z_{lj}^c\right) \\ 1 & \left(x_{ij}^c \neq z_{lj}^c\right) \end{cases} \quad (5)$$

$\omega(x_{ij}^c, l)$ adalah nilai penimbang untuk x_{ij}^c dimana nilai $\omega(x_{ij}^c, l)$ adalah

$$\omega\left(x_{ij}^c, l\right) = \frac{f\left(x_{ij}^c | C_l\right)}{|C_l| f\left(x_{ij}^c | D\right)} \quad (6)$$

$f\left(x_{ij}^c | C_l\right)$ adalah frekuensi nilai x_{ij}^c dalam kluster l, dan $|C_l|$ adalah jumlah objek dalam kluster l, dan $f\left(x_{ij}^c | D\right)$ adalah frekuensi nilai x_{ij}^c pada keseluruhan dataset. Pada paper ini *matching similarity measure* yang digunakan untuk data kategorikal menggunakan formula He.

Berdasarkan persamaan (1)-(5), maka ukuran kesamaan untuk data yang memiliki atribut numerik dan atribut kategorikal adalah [*2*]

$$d(X_i, Z_l) = \left(\sum_{l=1}^{m_r} \left(x_{ij}^r - z_{lj}^r\right)^2 + \gamma_l \sum_{j=l+1}^{m_c} \delta\left(x_{ij}^c, z_{lj}^c\right)\right)^{1/2} \quad (7)$$

Huang Cost Function

Huang menyatakan persamaan *cost function* untuk data campuran numerik dan kategorikal adalah

$$Cost_l = \sum_{l=1}^{k} u_{il} \sum_{j=1}^{m_r} \left(x_{ij}^r - z_{lj}^r\right)^2 + \gamma_l \sum_{j=1}^{m_c} u_{il} \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, z_{lj}^c\right) \quad (8)$$
$$Cost_l = Cost_l^r + Cost_l^c$$

dimana $Cost_l^r$ adalah biaya total untuk semua atribut numerik dari *object* dalam *cluster l*. $Cost_l^r$ diminimalkan jika z_{jl} dihitung dengan persamaan (9) berikut ini.

$$z_{lj} = \frac{1}{n_l} \sum_{i=1}^{n} u_{il} x_{ij} \quad (9)$$

untuk $j = 1, \ldots, m$

dimana $n_l = \sum_{i=1}^{n} u_{il} x_{ij}$ adalah jumlah *object* di dalam *cluster l.*

Pada atribut kategorikal misalkan C_j adalah sekumpulan nilai unik yang terdapat dalam atribut kategorikal *j* dan $p(c_j \in C_j|l)$ adalah probabilitas dari kemunculan nilai c_j di dalam *cluster l.* maka $Cost_l^c$ dalam persamaan (5) bisa ditulis ulang menjadi

$$Cost_l^c = \gamma_l \sum_{j=1}^{m_c} n_l \left(1 - p\left(q_{jl}^c \in C_j|l\right)\right) \quad (10)$$

dimana n_l adalah jumlah *object* di dalam *cluster l.* Solusi untuk meminimalisasi $Cost_l^c$ dijelaskan dengan Lemma 1 berikut.

Lemma 1: untuk sebuah *cluster* khusus *l*, $Cost_l^c$ diminimalisasi jika dan hanya jika $p\left(z_{lj}^c \in C_j|l\right) \geq p\left(c_j \in C_j|l\right)$ untuk $z_{lj}^c \neq c_j$ untuk semua atribut kategorikal. Akhirnya *Cost* bisa dituliskan ulang dengan

$$Cost = \sum_{l=1}^{k} (Cost_l^r + Cost_l^c) = \sum_{l=1}^{k} Cost_l^r + \sum_{l}^{k} Cost_l^c = Cost^r + Cost^c \quad (11)$$

Persamaan (10) adalah *cost function* untuk *Clustering dataset* dengan atribut bernilai numerik dan kategorikal. Karena $Cost^r$ dan $Cost^c$ adalah non-negatif, minimalisasi *Cost* bisa dilakukan dengan meminimalkan $Cost^r$ dan $Cost^c$, *total cost* dari atribut numerik dan kategorikal untuk semua *cluster*.

Algoritma K-Prototype

Algoritma *K-Prototype* adalah salah satu metode *Clustering* yang berbasis *partitioning*. Algoritma ini adalah hasil pengembangan dari algoritma *K-Means* (Huang,1998) untuk menangani c*lustering* pada data dengan atribut bertipe campuran numerik dan kategorikal. Pengembangan yang dilakukan oleh Huang mempertahankan efisiensi algoritma K-Means dalam menghadapi data berukuran besar dan dapat diterapkan pada data numerik dan kategorikal. Pengembangan yang mendasar pada algoritma K-Prototype terdapat pada pengukuran kesamaan (*similarity measure*) antara *object* dengan *centroid (prototype)*-nya.

Secara umum algoritma K-Prototype terbagi kedalam tiga tahapan utama (Huang 1997), yaitu:

1. Inisialiasi awal *prototype*. Pada proses ini akan dilakukan pemilihan sejumlah *k prototype* secara acak dari *dataset X* sesuai dengan jumlah *cluster* yang ditentukan.
2. Alokasi objek di dalam *X* ke *Cluster* dengan *prototype* terdekat. Ukur Jarak Objek ke semua *prototype* dan tempatkan objek pada *cluster* terdekat. Tahap ini algoritma *K-Prototype* mengalokasikan semua *object* didalam *dataset* ke *cluster* dimana *prototype* dari *cluster* tersebut memiliki jarak yang paling dekat ke *object* data. Pengalokasian semua *object* di dalam daset *X* ke *cluster* yang memiliki jarak *prototype* terdekat dengan *object* yang diukur. Untuk setiap kali *object X* selesai dialokasikan, maka selanjutnya akan dilakukan penghitungan *(update)* terhadap *prototype cluster* yang berkaitan.
3) Realokasi *object* Jika terjadi perubahan *prototype*. Setelah semua *object* dalam X selesai dialokasikan, selanjutnya akan dilakukan pengukuran ulang jarak antara semua *object* di dalam X terhadap semua *prototype* yang ada. Jika ditemukan adanya *object* yang ternyata lebih dekat ke *prototype* yang lain, maka akan dilakukan pemindahan

keanggotaan dan kemudian akan dilakukan update terhadap *prototype cluster* lama dan *prototype cluster* baru. Proses ini akan terus dilakukan sampai tidak ada lagi perubahan *prototype* atau sampai kriteria *stopping* terpenuhi.

Evaluasi Hasil Cluster

Metode yang umum digunakan untuk mengukur hasil pengclusteran dengan data campuran adalah *Total Cost* dan *Categorical Variance Criterion* (CVC) (Hsu & Huang, 2008). CVC menggabungkan antara metode *Category Utility* dan pengukuran variance untuk data numerik. Semakin besar nilai CVC maka semakin bagus juga hasil *clustering.* Persamaan CVC sebagai berikut:

$$CVC = CU/(1+\sigma^2) \qquad (12)$$

Fungsi *Categorical Utility* (CU) bertujuan untuk memaksimalkan kemungkinan atau probabilitas bahwa dua buah object di dalam cluster yang sama memiliki nilai atribut yang sama dan probabilitas bahwa dua object pada cluster yang berbeda memiliki atribut yang berbeda. *Categorical utility* untuk sebuah dataset dapat dihitung sebagai berikut:

$$CU = \sum_l \left(\frac{|C_l|}{|D|} \sum_j \sum_i \left[P(A_j = V_{ij} \mid C_l)^2 - P(A_j = V_{ij})^2 \right] \right) \qquad (13)$$

$P(A_j = Vij \mid C_l)$ adalah probabilitas kondisional dimana atribut *j* memiliki nilai V_{ij} di dalam cluster C_l, dan $P(A_j = V_{ij})$ probabilitas keseluruhan bahwa atribut *j* memiliki nilai V_{ij} di seluruh dataset.

Variance (σ^2), bisa digunakan untuk mengevaluasi kualitas *clustering* untuk data numerik. Total *variance* dapat diperoleh dengan melakukan penambahan semua *variance* di setiap cluster, dimana pada setiap *cluster* akan dilakukan penambahan *variance* dari setiap data numerik. Persamaannya sebagai berikut:

$$\sigma^2 = \sum_l \frac{1}{|C_l|} \sum_j \sum_i (V_{ij}^l - V_{j,avg}^l)^2 \qquad (14)$$

Dalam hal ini, V_{ij}^l dan $V_{j,avg}^l$ adalah record ke-*i* dan nilai rata-rata atribut numerik ke-*j* pada cluster ke C_l.

Algoritma Hibrida K-Prototype-GA

Kim dkk (2008), menggunakan kombinasi algoritma *K-Means* dengan algoritma Genetika pengelompokan pelanggan dalam membuat *recomender system* pada *online shoping market.* Dari hasil penelitian tersebut, bisa disimpulkan bahwa *GA-K-Means* mampu menghasilkan pengelompokan *(clustering)* yang lebih baik dibandingkan dengan *Self Organising Map* (SOM) yang berbasis *Neural Network.*

Min Feng melakukan penelitian untuk mengoptimalkan Algoritma *K-Means* dalam menentukan pusat awal *cluster,* dimana hasil penelitian menunjukkan bahwa algoritma *K-Means* memiliki kelemahan tidak hanya memiliki ketergantungan pada data awal (*inisial center cluster*), tetapi juga konvergensi yang cepat (konvergensi prematur) dan hasil *clustering* yang kurang akurat (Feng & Wang, 2011). Untuk memperoleh *cluster* yang efektif dan akurat maka Min Feng dan Zhenyan-wang mengoptimalkan Algoritma *K-Means* (PKM) dengan Algoritma Genetika menjadi sebuah Algoritma Hibrid (PGKM). Percobaan menunjukkan bahwa algoritma ini dapat mengatasi masalah pada penentuan *inisial center cluster,* konvergensi premature dan waktu pengolahan yang lebih efisien.

Hasil penelitian Liu dkk, tahun 2008 menggunakan algoritma genetik yang dikombinasikan dengan algoritma *K-Means* untuk menemukan variabel yang valid dan jumlah *cluster* optimal secara simultan. Hasil penelitian menunjukkan, metode hibrid tersebut behasil menghilangkan variabel yang tidak relevan dan menghasilkan jumlah *cluster* secara otomatis, dan berhasil meningkatkan hasil pengelompokan pelanggan secara signifikan (Liu & Ong, 2008).

Berdasarkan pada fakta-fakta yang didapatkan dari beberapa penelitian sebelumnya maka penulis melalui penelitian ini mengusulkan untuk mengkombinasikan algoritma genetika dengan metode clustering yang diusulkan dalam penelitian ini yaitu algoritma *K-*

Prototype untuk memperoleh tingkat akurasi yang lebih baik.

Dalam penelitian ini metode algoritma genetika digunakan untuk memperoleh inisial *center cluster* yang optimal. Tahapan algoritma metode hibrida *K-prototype*-GA dalam penelitian ini adalah sebagai berikut:

1) Menentukan inisial populasi, meliputi jumlah gen dalam kromosom dan jumlah kromosom dalam individu
2) Melakukan proses algoritma K-Prototype untuk setiap kromosom dalam populasi.
3) Menghitung nilai *fitness* dari tiap kromosom dalam populasi berdasarkan nilai *cost function*.
4) Memilih kromosom berdasarkan nilai *fitness*.
5) Melakukan perkawinan silang (*crossover*) dan mutasi untuk mendapatkan keturunan (*offspring*).
6) Melakukan *elitism* dan *replacement* sehingga diperoleh populasi baru.
7) Kembali ke langkah 2 hingga kriteria yang ditentukan terpenuhi
8) Setelah diperoleh hasil dari proses algoritma genetika kemudian digunakan untuk proses *clustering* dengan algoritma *K-Prototype*.

Metode Analisis

Algoritma Genetika (GA) banyak digunakan dalam masalah pencarian parameter optimal, dengan demikian algoritma genetika akan digunakan untuk mengoptimalkan hasil pengclusteran dengan *K-prototype*. Penerapan GA dalam penelitian ini memiliki fungsi untuk menghasilkan inisial *center cluster* yang optimal sehingga pada tahap awal inisialisasi populasi dengan *K-Prototype* menggunakan *center cluster* dari hasil pencarian Algoritma Genetika (Feng & Wang, 2011).

Preprocessing

Sebelum melakukan pemodelan akan dilakukan preprocessing, yang pertama pemeriksaan data missing, yang kedua tranformasi data numerik. Pada variabel penelitian ini terdapat dua variabel kategorikal yang memiliki missing value yaitu variabel permukaan jalan dan variabel kondisi jalan yang dapat dilalui kendaraan beroda empat atau lebih. Missing terjadi karena terdapat desa yang tidak memiliki sarana transportasi darat oleh sebab itu penulis mengganti nilai missing kategori tersebut dengan kategori tidak memiliki transportasi darat.

Transformasi yang dilakukan adalah melakukan standardisasi data numerik menjadi Z-Score. Hal ini dilakukan karena data memiliki satuan yang berbeda. Proses standarisasi menjadikan dua data dengan perbedaan satuan yang lebar akan otomatis menjadi menyempit (Santoso,2010). Dengan demikian analisis perbandingan antar variabel pun dapat dibandingkan.

Selain standarisasi dan transformasi, akan dilakukan juga pembuatan *look up table* untuk mengefisienkan waktu penghitungan jarak pada saat menjalankan algoritma genetika. *Look up table* merupakan tempat yang menyimpan semua jarak antar objek, strategi ini digunakan untuk mengefisienkan waktu computing seperti yang dilakukan oleh Lin dan Yang (2005).

Kemudian hal lain yang perlu dipersiapkan pada saat akan melakukan proses clustering adalah menyiapkan beberapa inputan kategori sebagai berikut:

1) Ukuran Populasi. Ukuran populasi diperlukan pada saat akan mengeksekusi algoritma genetika. Tidak ada ketentuan dalam menentukan ukuran populasi. Jika jumlah kromosom yang digunakan terlalu sedikit, maka individu yang dapat digunakan untuk proses *crossover* dan mutasi akan sangat terbatas, sehingga menyia-nyiakan proses yang ada. Sedangkan jika jumlah kromosom yang digunakan terlalu banyak, akan memperlambat proses algoritma genetika yang dilakukan. Semakin besar ukuran populasi dalam satu generasi, maka akan menghasilkan solusi yang lebih baik. Dalam penelitian ini ukuran populasi adalah 1000 kromosom untuk

fungsi inisial center dan 50 untuk fungsi penentuan variabel relevan.

2) Generasi Maksimal. Generasi maksimal atau iterasi maksimal mempengaruhi jumlah komputasi yang dilakukan pada saat pengolahan data, dimana 1 generasi dapat mewakili sebesar populasi yang telah ditentukan. jika terdapat 10 generasi dan ukuran populasi 1000 maka komputasi yang dilakukan akan sebanyak 10000 kali sehingga menghasilkan kromosom yang lebih variatif dalam proses *fitness*. Kromosom menjadi variatif dikarenakan pada saat pencarian nilai *fitness* terbaik pada generasi pertama telah selesai, maka selanjutnya dilakukan proses pindah silang dari populasi yang ada, sehingga pada generasi selanjutnya akan menghasilkan populasi yang baru. Pada penelitian ini generasi maksimal yang ditetapkan adalah 100.
3) Jumlah *Cluster*. Jumlah *Cluster* (*K*) ditentukan di awal untuk menggelompokkan data yang diolah sesuai dengan jumlah *cluster* yang diinginkan. Jumlah cluster minimal (*Kmin*) adalah 2 dan maksimal (*Kmaks*) adalah *n/2* atau $\sqrt{n}$ (Lin, Yang, & Kao, 2005) dimana *n* adalah jumlah data. Kemudian jumlah cluster bisa ditentukan secara random dengan rentang [*Kmin, Kmaks*]. Dalam penelitian ini jumlah *cluster* awal ditetapkan berdasarkan hasil pengamatan perubahan nilai cost function pada saat nilai *K=2* sampai dengan *K=20*.
4) Parameter algoritma genetika yang disarankan De Jong (A.A., 2001) dalam (Zukhri, 2013) adalah (1) Probabilitas penyilangan cukup besar lebih dari 50%, (2) Probabilitas mutasi cukup kecil (sebuah gen untuk sebuah kromosom), (3) Ukuran populasi berkisar antara 50 sampai 500 kromosom. Walaupun demikian tidak ada batasan yang pasti mengenai besaran nilai probabilitas crossover, mutasi dan ukuran kromosom tergantung dari tujuan penelitian. Dengan mengacu ketetapan De Jong maka batas probabilitas penyilangan yang digunakan pada penelitian ini adalah 0,5, probabilitas mutasi adalah 0,1 dan ukuran populasi untuk proses penentuan center cluster optimal dan pemilihan variabel relevan berturut-turut adalah 1000 dan 50.

Pemodelan

K-Prototype

Pada proses *Clustering* dengan *K-Prototype* dilakukan beberapa proses utama yang terbagi kedalam 3 tahapan utama sebagai berikut (Huang, 1997):

1) Inisialiasi awal *prototype*. Pada proses ini akan dilakukan pemilihan sejumlah k *prototype* secara acak dari *dataset* X sesuai dengan jumlah *cluster* yang ditentukan.
2) Alokasi objek di dalam X ke *Cluster* dengan *prototype* terdekat. Berikut *pseudocode* dari algoritma alokasi objek kedalam *cluster* pada *K-Prototype* (Gambar 3 (Lampiran 3)):
3) Realokasi *object* Jika terjadi perubahan *prototype*. Proses ini akan terus dilakukan sampai tidak ada lagi perubahan *prototype* atau sampai kriteria *stopping* terpenuhi. Berikut *pseudocode* algoritmanya (Gambar 4 (Lampiran 4)):

K-Prototype-GA* untuk Optimasi Inisial *Center Cluster

Desain dari optimasi yang dilakukan pada *K-Prototype* terletak pada inisial pusat *cluster*. Jadi ketika pusat *cluster* di optimasi dengan Algoritma Genetik diharapkan *K-Prototype* mempunyai awalan *prototype* yang bagus, sehingga untuk proses selanjutnya dapat memperoleh *cluster* yang lebih akurat. Alur proses penggabungan metode K-Prototype dan Genetika untuk memperoleh center cluster optimal dalam penelitian ini dapat dilihat pada Gambar 5 (Lampiran 5).

1) Inisialisasi Populasi Awal

Penelitian ini melibatkan 37 variabel dan 77.961 record data dalam penelitian ini. Unit record adalah unit desa yang diberikan indeks mulai dari nomor 1 sd 77.961. Fase inisialisasi populasi ini digunakan untuk menentukan sejumlah kromosom awal yang akan digunakan untuk komputasi selanjutnya. Berdasarkan pada penelitian Jie dan Li-Cheng (2003) maka dalam membentuk kromosom individu, panjang gen ini adalah sama dengan jumlah *K* (jumlah *cluster*) dari proses *Clustering*. Dimana masing-masing nilai yang ada pada gen akan mewakili no record data pada proses *Clustering* (Jie, Xinbo, & Li-Cheng, 2003). Jadi nilai yang ada pada gen adalah no ID desa yang terpilih secara acak dari nomor 1-77.961 sebanyak *k* cluster. Jika jumlah kromosom yang akan dibentuk adalah 1000 maka akan ada 1000 x *k* yaitu 1000*k* desa yang akan terpilih pada populasi awal ini.

Dalam penelitian ini jumlah cluster ditetapkan setelah menganalisa grafik perubahan nilai cost function dari $k \in [2,20]$. Jumlah cluster yang paling signifikan penurunan nilai cost functionnya akan dijadikan nilai *k* pada metode pengclusteran dengan *k-prototype*.

2) Evaluasi Fitness

Proses evaluasi dengan alat ukurnya adalah fungsi *fitness* merupakan proses untuk mengevaluasi setiap populasi dengan menghitung nilai *fitness* setiap kromosom dan mengevaluasinya sampai terpenuhi kriteria berhenti. Nilai *fitness* menyatakan nilai dari fungsi tujuan. Tujuan dari algoritma genetika adalah memaksimalkan nilai *fitness*. Berikut formulanya:

$$f = \frac{1}{(h+a)} \quad (15)$$

h adalah suatu nilai yang sangat kecil untuk menghindari pembagian dengan nilai 0. Sedangkan *a* adalah fungsi cost pada *K-Prototype*.

Kemudian untuk menghindari optimum lokal, dibuatlah suatu mekanisme yang disebut dengan *Linier Fitness Rangking* (LFR). Tujuan dari mekanisme ini sebenarnya adalah untuk melakukan penskalaan nilai-nilai *fitness* dengan menggunakan persamaan berikut:

$$LFR(i) = f_{max} - (f_{max} - f_{min}) \left(\frac{R(i)-1}{N-1}\right) \quad (16)$$

Keterangan:

- LFR(i) = nilai LFR individu ke-i
- N = jumlah individu dalam populasi
- R(i) = rangking individu ke-i setelah diurutkan dari nilai fitness terbesar hingga terkecil
- fmax = nilai *fitness* tertinggi
- fmin = nilai *fitness* terendah

3) *Elitisme* dan *Replacement*

Proses elitisme diperlukan untuk mencegah kehilangan solusi terbaik, tahap ini dapat meningkatkan performansi algoritma genetika secara cepat. Pada saat membuat populasi baru dengan kawin silang dan mutasi, kromosom terbaik dapat hilang. Elitism adalah metode untuk mengganti kromosom terjelek dengan kromosom terbaik.

Individu terbaik ini ditentukan berdasarkan nilai fitnessnya, individu/kromosom dirangking berdasarkan besaran nilai fitnessnya. Semakin besar nilai fitnessnya senakin baik kromosom/individu tersebut. penentuan individu terbaik dilakukan untuk kebutuhan proses crossover dan mutasi. Jika ukuran populasi adalah ganjil maka kromosom terbaik di copy sebanyak satu kromosom, sedangkan jika ukuran populasi genap maka kromosom yang dicopy sebanyak dua kromosom. Kromosom terbaik ini akan dibandingkan dengan kromosom hasil penyilangan jika nilainya lebih besar dari hasil penyilangan maka copy kromosom elit kedalam iterasi berikutnya. Jika lebih kecil maka replace kromosom elit dengan kromosom terbaik hasil penyilangan.

4) Seleksi

Seleksi dilakukan dalam rangka untuk mendapatkan calon induk yang baik

yang akan menjalani proses *crossover* dan *mutasi*. Metode yang banyak digunakan dalam proses seleksi adalah teknik *roullete wheel*. Pendekatan ini dilakukan dengan menghitung nilai probabilitas seleksi (p) tiap individu/kromosom berdasarkan nilai fitnessnya dengan persamaan sebagai berikut:

$P_i = \frac{f_i}{f_{total}}$ i=1,2, ..., *pop size*

f_i menyatakan nilai *fitness* dari individu ke-i dan f_{total} adalah total nilai *fitness* dari semua individu.

Setelah diperoleh nilai *p* kemudian dihitung *probabilitas kumulatif* yang akan digunakan pada proses seleksi tiap individu. Kemudian untuk memilih tiap individu bangkitkan nilai peluang *r* secara random. Pilih individu yang nilai probabilitas kumulatifnya $p_{kum} \geq r$.

5) Proses Penyilangan (*Crossover*)

Crossover adalah operator dalam algoritma genetika untuk melakukan operasi pertukaran gen-gen yang bersesuaian dari dua induk untuk membentuk individu baru. Proses perkawinan silang dilakukan berdasarkan probabilitas kawin silang yaitu $P_c \in [0,1]$. Dibangkitkan suatu bilangan random p untuk menentukan terjadi kawin silang atau tidak. Apabila $p \geq P_c$ maka tidak terjadi kawin silang. Menurut De Jong nilai Pc disarankan untuk ditetapkan cukup besar berkisar 50% sampai 70% (A.A., 2001) dalam (Zukhri, 2013). Kemudian untuk menentukan titik potong maka dilakukan juga dengan membangkitkan suatu bilangan acak [1,panjang gen-1].

6) Mutasi

Mutasi dilakukan untuk mencegah algoritma berada pada optimum lokal. Mutasi merupakan proses menggantikan gen yang hilang dari populasi akibat proses seleksi yang memungkinkan munculnya kembali gen yang tidak muncul pada inisialisasi populasi. Mutasi juga terjadi pada probabilitas tertentu yaitu $P_m \in [0,1]$. Pada tahap ini pada setiap gen dibangkitkan suatu bilangan *p*, jika *p* lebih kecil dari p_m yang ditetapkan maka gen tersebut akan dikenai proses mutasi. Proses menggantikan nilai dalam gen yang terkena mutasi terdapat beberapa cara. Pertama dengan membangkitkan bilangan acak dari separuh jumlah record. Kedua menukar nilai pada gen tersebut dengan gen lain yang juga terkena mutasi.

Setelah proses crossover dan mutasi selesai maka akan dilakukan kembali proses evaluasi dengan menghitung fitness dan membandingkan dengan kromosom elite. Kemudian dilakukan replacement jika kromosom baru lebih baik dibandingkan kromosom elite. Begitu seterusnya hingga kriteria berhenti terpenuhi.

Data Penelitian

Dalam penelitian ini, penulis akan melakukan studi kasus menggunakan dataset PODES 2011 se-Indonesia. Dataset yang digunakan dalam penelitian ini terdiri dari 77.961 *record* yang menunjukkan 77.961 desa dan 71 atribut yang dikelompokkan menjadi 37 variabel.

Variabel yang digunakan dalam penelitian ini berdasarkan pada kajian Identifikasi Desa Tertinggal tahun 2002 yang diselenggarakan oleh BPS tahun 2003 menggunakan PODES 2002. Variabel penelitian yang digunakan akan disesuaikan dengan kondisi kuesioner PODES 2011.

HASIL DAN PEMBAHASAN

Hasil *K-Prototype* Tanpa GA

Eksekusi program utama K-Prototype bertujuan untuk mendapatkan pengelompokan dengan nilai total *cost function* terkecil. Total cost menunjukkan total jarak setiap objek terhadap prototype cluster. Semakin kecil nilai total cost maka semakin dekat jarak antara objek dengan prototype clusternya.

Pada Gambar 6 (Lampiran 6) terlihat perubahan nilai total cost dengan beberapa kali percobaan, mulai dari K = 2 dan sampai dengan K = 20. Pada Gambar 6 jumlah cluster dimulai dari dua dengan total nilai cost adalah 1,490438 x 10^6. Semakin besar jumlah cluster yang ditentukan maka nilai total cost semakin mengecil. Penurunan yang paling signifikan adalah pada saat *k* bernilai 4, 6, dan 13, yang mengalami penurunan sebesar 7,44657 x 10^5 , 5,10553 x 10^5 dan 6,63805 x 10^5. Berdasarkan efisiensi waktu pengolahan dan pertimbangan kecukupan jumlah cluster maka nilai *k* yang ditetapkan adalah *k* = 6.

Hasil dari proses clustering dengan K-Prototype, dimana K = 6, maka diperoleh jumlah anggota tiap cluster seperti yang tertera pada Tabel 2.

Pada Tabel 2 diperoleh informasi bahwa anggota *cluster* terbanyak terdapat pada cluster 2 dengan persentase sebesar 41,44 persen. Sedangkan *cluster* 6 memiliki persentase terkecil yaitu 0,08 persen. Jika dilihat dari aspek pemerataan kondisi sosial ekonomi dan prasarana berdasarkan indikator ketertinggalan desa maka desa-desa pada *cluster* 6 merupakan kelompok desa yang sangat berbeda karakteristik sosial ekonominya dibandingkan dengan kelompok besar desa lainnya.

Tabel 2. Jumlah anggota Per Cluster

Cluster i	Jumlah Anggota (Desa)	%
Cluster 1	12929	16,58
Cluster 2	32308	41,44
Cluster 3	28502	36,56
Cluster 4	2364	3,03
Cluster 5	1797	2,30
Cluster 6	61	0,08

Hasil Hibrid *K-Prototype* dengan Algoritma Genetika

Metode hibrid yang berbasis *K-Prototype* merupakan metode untuk membangkitkan inisial pusat *cluster* yang sudah dioptimasi dengan metode algoritma Genetika, kemudian inisial pusat cluster tersebut digunakan dalam melakukan pengclusteran dengan algoritma *K-Prototype*. Sehingga algoritma genetika dalam metode ini hanya digunakan untuk memperoleh calon inisial pusat *cluster* yang baik.

Tahapan algoritma genetika dalam memperoleh inisal pusat *cluster* terbaik adalah sebagai berikut:

1) Menentukan populasi kromosom awal secara acak. Tahapan menentukan kromosom awal dilakukan pada saat input kategori yang dipilih secara acak sebanyak 1000 kromosom dengan jumlah cluster 6.
2) Evaluasi nilai *fitness*. Evaluasi nilai fitness seluruh kromosom untuk mencari nilai terbaik dari clustering yang dilakukan. Kemudian evaluasi nilai *fitness* terbaik per iterasi dalam setiap jumlah cluster. Hasilnya dapat dilihat pada Gambar 7 (Lampiran 7). Gambar 7 memperlihatkan pergerakan perubahan nilai fitness dari best kromosom per iterasi. Perubahan nilai fitness masih bergerak hingga iterasi ke 14 setelah itu mencapai nilai konvergen pada iterasi ke 15 dan seterusnya dengan nilai fitness adalah 9,4423 x 10^{-8}. Dengan demikian nilai fitness terbaik pada saat jumlah cluster 6 adalah 9,4423 x 10^{-8}. Maka *the best chromosom* adalah sebagai berikut:

180	13158	26151	39048	52138	65201

Hasil dari *the best chromosom* merupakan inisial center cluster pada proses clustering dengan algoritma *k-prototype.* Nilai pada setiap gen adalah no id desa. Panjang gen pada Gambar 8 (Lampiran 8) menunjukkan jumlah cluster.

Nomor objek pada Gambar 7 dan 8 akan menjadi inisial center pada proses pengclusteran dengan K-Prototype. Persentase jumlah anggota per cluster yang dihasilkan dari algoritma K-Prototype setelah mengoptimasi inisal centernya dapat dilihat pada Tabel 3.

Tabel 3. Jumlah Anggota per Cluster

Cluster i	Jumlah Anggota (Desa)	%
Cluster 1	2179	2,79
Cluster 2	16341	20,96
Cluster 3	24630	31,59
Cluster 4	61	0,078
Cluster 5	1789	2,29
Cluster 6	32961	42,28

Evaluasi Perbandingan Hasil Clustering

Baik atau tidaknya *cluster* yang dihasilkan dari kedua model tersebut akan dilihat dari beberapa alat ukur yang dapat digunakan untuk data campuran yaitu *Total Cost* dan *Categorical Variance Criterion*. *Total cost* adalah total jarak dari setiap objek ke cluster tempat dia berada. Semakin kecil nilai *total cost* maka cluster yang terbentuk semakin *compact.*. Model yang akan dibandingkan tersebut dapat dilihat pada Tabel 4.

Tabel 4. Perbandingan hasil *Clustering*

No	Model	jumlah cluster	Total Cost	CVC
1	Model K Prototype tanpa Genetika	6	$1204{,}9 \times 10^3$	0,0031
2	Model K Prototype –Genetika untuk center cluster optimal	6	$1202{,}9 \times 10^3$	0,0054

Berdasarkan Tabel 4 maka model K Prototype-Genetika lebih baik dibandingkan dengan model K-Prototype saja tanpa dioptimasi dengan genetika. Ditunjukkan oleh nilai total cost pada K-Prototype yaitu $1204{,}9 \times 10^3$ lebih besar dibandingkan dengan model K-Prototype-Genetika yang menggunakan optimasi dengan genetika yaitu $1202{,}9 \times 10^3$. Nilai CVC pada model K-Prototype-Genetika untuk center cluster optimal merupakan nilai yang terbesar yaitu 0,0054. Semakin besar nilai CVC maka semakin bagus clustering yang dihasilkan. Maka dalam penelitian ini model K-Prototype-Genetika untuk center cluster optimal menghasilkan akurasi cluster yang lebih baik diantara metode lainnya. Artinya dengan kondisi data yang sama model K-Prototype-Genetika untuk center cluster optimal lebih mampu menghasilkan cluster yang lebih homogen dibandingkan model lainnya. Hal ini menunjukan jika model ini lebih baik dibandingkan dengan model lainnya.

Tabel 5. Nilai CU dan Varians pada setiap model penelitian

No	Model	CU	Varians
1	Model K Prototype tanpa Genetika	1,1406	366,7799
2	Model K Prototype –Genetika untuk center cluster optimal	1,1351	209,073

Jika dilihat dari hasil evaluasi dengan melihat nilai *total cost* dan CVC kedua model tersebut, maka perbedaan kedua model terlihat tidak terlalu signifikan. Maka penulis menyarankan pada penelitian selanjutnya dilakukan pembuangan outlier dan pemilihan variable yang relevant terlebih dahulu.

KESIMPULAN DAN SARAN

Metode gabungan *K-Prototype* dengan Algortima Genetika yang diusulkan dalam penelitian ini menghasilkan inisial pusat *cluster yang* optimal. Hal ini terlihat dari hasil percobaan yang telah dilakukan, dimana pada saat dilakukan pengujian menggunakan total cost, model K-Prototype- GA menghasilkan nilai total cost sebesar $1202{,}9 \times 10^3$. Model K-Prototype tanpa GA menghasilkan nilai total cost sebesar $1204{,}9 \times 10^3$. Dengan demikian model K-Prototype-GA untuk kasus penelitian pengelompokan desa berdasarkan indikator ketertinggalan desa menggunakan dataset PODES 2011 memiliki tingkat akurasi yang lebih baik dibandingkan model cluster dengan K-Prototype tanpa Genetika.

Berdasarkan nilai index clustering criterion maka pada kasus penelitian ini model K-Prototype-Genetika untuk center cluster optimal merupakan model terbaik

karena nilainya lebih tinggi dari yang lainnya yaitu 0.0054 dibandingkan 0,0031 dimana hal ini menunjukkan bahwa tingkat kesamaan ciri atau karakteristik dari setiap kelompok yang terbentuk pada model K-Prototype-Genetika untuk optimasi inisial center cluster lebih mirip.

Penelitian ini memanfaatkan metode k-prototype dengan GA sebagai metode utama dalam proses *clustering*. Perlu dilakukan penelitian lebih lanjut untuk dapat menghasilkan clustering yang lebih baik mengingat begitu kompleksnya struktur data dalam penelitian ini serta tipe atribut berupa campuran, numerik dan kategorikal, menyebabkan proses pengolahan semakin komplek dan waktu pengolahan yang panjang.

Untuk mengevaluasi hasil pengelompokan, penulis menyarankan untuk mencari dan menggunakan alat ukur lainnya yang cocok digunakan untuk mengevaluasi hasil pengclusteran dengan data yang bertipe campuran.

DAFTAR PUSTAKA

Amir Ahmaddan Lipika Dey, "A k-mean clustering algoritm for mixed numeric and categorical data," *DAta & Knowledge Engineering*, vol. 63, pp. 503-527, 2007.

BPS, *Metodologi dan Profil Kemiskinan Tahun Tahun 2002*., 2003.

Ch. D. V. Subba Rao, C. Kishore and Shreyash Raju Srinivasulu Asadi, "Clustering the Mixed Numerical and Categorical Datasets Using Similarity Weight and Filter Method," *VSRD International Journal of Computer Science & Information Technology*, vol. 2 (5), pp. 373-385, 2012.

Dharmendra K Roy, Lokesh K Sharma, "Genetic k-Means Clustering Algorithm for Mixed Numeric and Categorical Data Sets," *International Journal of Artificial Intelligece & Applications (IJAIA)*, vol. 1, April 2010.

Gil David, Amir Averbuch, "SpectralCAT:Categorical spectral clustering of numerical and nominal data," *Pattern Recognition*, vol. 45, pp. 416-433, 2012.

J.Han Kamber, *Data Mining Concepts and Techniques*, 2nd ed. San Fransisco, United States of America: Dianne Cerra, 2006.

J. Suguna, M.Arul Selvi, "Ensemble Fuzzy Clustering for Mixed Numeric and Categorical Data," *International Journal of Computer Applications (0975-8887)*, vol. 42 - No 43, Maret 2012.

M. Ramakrishnan, D. Tennyson Jayaraj, "Modified K-Means Algorithm for effective Clustering of Categorical Data Sets," *International Journal of Computer Applications (0975-8887)*, vol. 89 - No 7, Maret 2014.

Ramesh Valaboju, N. Raghava Rao V.N. Prasad Pinisetty, "Hybrid Algorithm for Clustering Mixed Data Sets," *IOSR Jornal of Computer Engineeering (IOSRJCE)*, vol. 6, no. 2, pp. 09-13, Sep-Okt 2012.

Zhexue Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values".

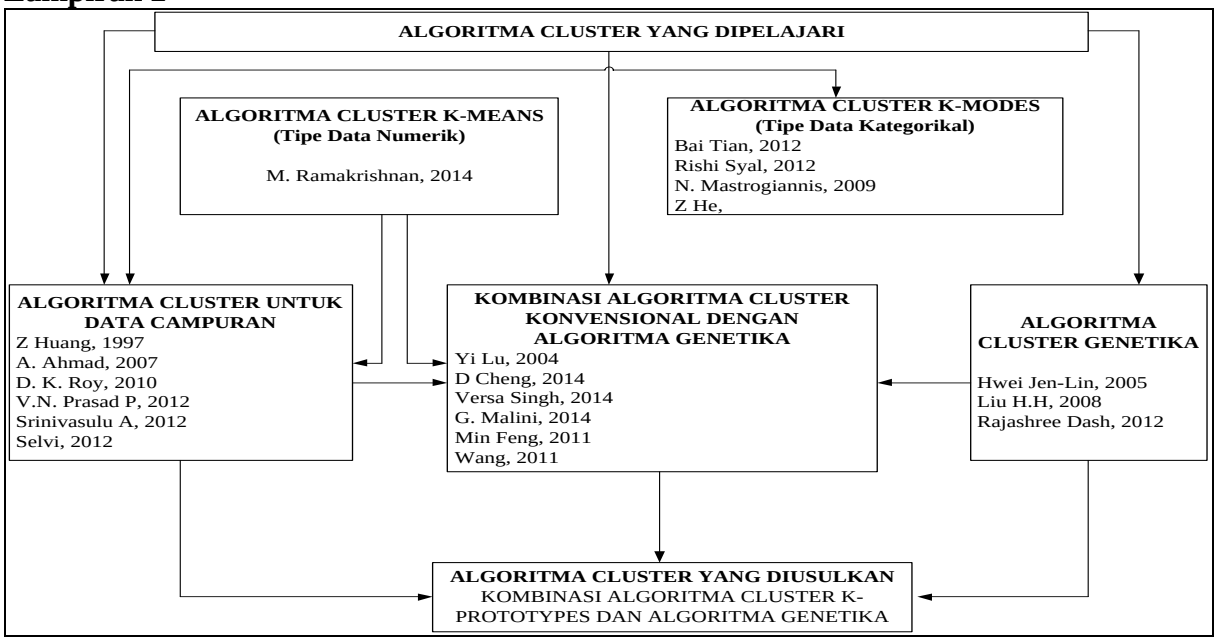
LAMPIRAN

Lampiran 1

Tabel 1. Jumlah atribut kuesioner PODES 2011 menurut bagian-bagian pertanyaan

Bagian kuesioner	Uraian	Jumlah Atribut		
		Numeric	Categorical	Total
BLOK I	Pengenalan tempat	-	15	15
BLOK II	Keterangan petugas	-	4	4
BLOK III	Keterangan umum desa/kelurahan	2	19	21
BLOK IV	Kependudukan dan ketenagakerjaan	7	2	9
BLOK V	Perumahan dan lingkungan hidup	11	38	49
BLOK VI	Bencana alam dan penanganan bencana alam	30	56	86
BLOK VII	Pendidikan dan kesehatan	87	40	127
BLOK VIII	Sosial budaya	24	20	44
BLOK IX	Hiburan dan olahraga	2	19	21
BLOK X	Angkutan, komunikasi dan Informasi	6	31	37
BLOK XI	Penggunaan lahan	6	6	12
BLOK XII	Ekonomi	25	13	38
BLOK XIII	Keamanan	11	56	67
BLOK XIV	Otonomi desa dan program pemberdayaan masyarakat	7	48	55
BLOK XV	Keterangan aparatur desa	2	6	8
Total		**220**	**373**	**593**

Lampiran 2



Gambar 1. Peta Penelitian yang Terkait *Clustering*

Lampiran 3

```
FOR i = 1 TO NumberOfObjects
  Mindistance=       Distance(X[i],O_prototypes[1])+       gamma*
  Sigma(X[i],C_prototypes[1])
   FOR j = 1 TO NumberOfClusters
      distance=     Distance(X[i],O_prototypes[j])+     gamma     *
      Sigma(X[i],C_prototypes[j])
        IF (distance < Mindistance)
          Mindistance=distance
          cluster=j
        ENDIF
   ENDFOR
  Clustership[i]=cluster
  ClusterCount[cluster] + 1
   FOR j=1 TO NumberOfNumericAttributes
      SumInCluster[cluster,j] + X[i,j]
      O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCoun
      t[cluster]
   ENDFOR
   FOR j=1 TO NumberOfCategoricAttributes
      FrequencyInCluster[cluster,j,X[i,j]] + 1
      C_prototypes[cluster,j]=HighestFreq(FrequencyInCluster,clus
      ter,j)
   ENDFOR
ENDFOR
```

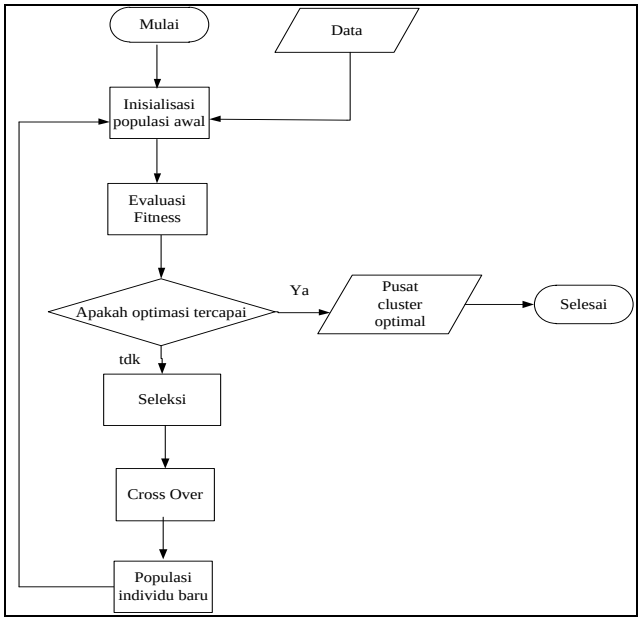
Gambar 3. *Pseudocode* K-Prototype pada tahap pengalokasian objek kedalam cluster (Huang 1998)

Lampiran 4

```
moves=0
FOR i = 1 TO NumberOfObjects
  …
  (To find the cluster whose prototype is the nearest to object
  i.)
  …
   IF (Clustership[i]<>cluster)
    moves+1
    oldcluster=Clustership[i]
    ClusterCount[cluster] + 1
    ClusterCount[oldcluster] - 1
      FOR j=1 TO NumberOfNumericAttributes
       SumInCluster[cluster,j] + X[i,j]
       SumInCluster[oldcluster,j] - X[i,j]
       O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCo
       unt[cluster]
       O_prototypes[oldcluster,j]=
       SumInCluster[oldcluster,j]/ClusterCount[oldcluster]
      ENDFOR
      FOR j=1 TO NumberOfCategoricAttributes
       FrequencyInCluster[cluster,j,X[i,j]] + 1
       FrequencyInCluster[oldcluster,j,X[i,j]] - 1
       C_prototypes[cluster,j]=HighestFreq(cluster,j)
       C_prototypes[oldcluster,j]=HighestFreq(oldcluster,j)
      ENDFOR
  ENDIF
ENDFOR
```

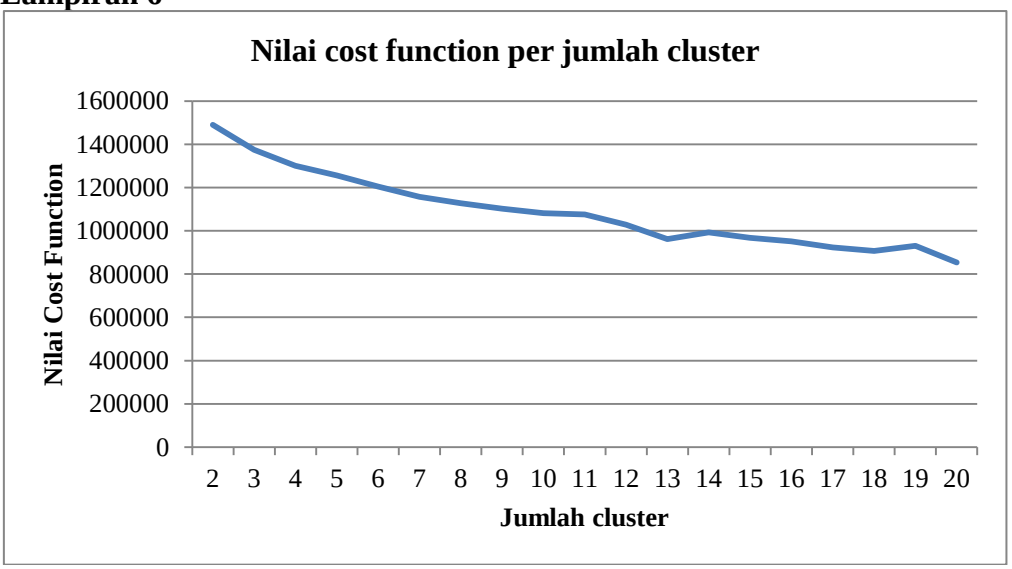
Gambar 4. *Pseudocode* Realokasi Objek

Lampiran 5



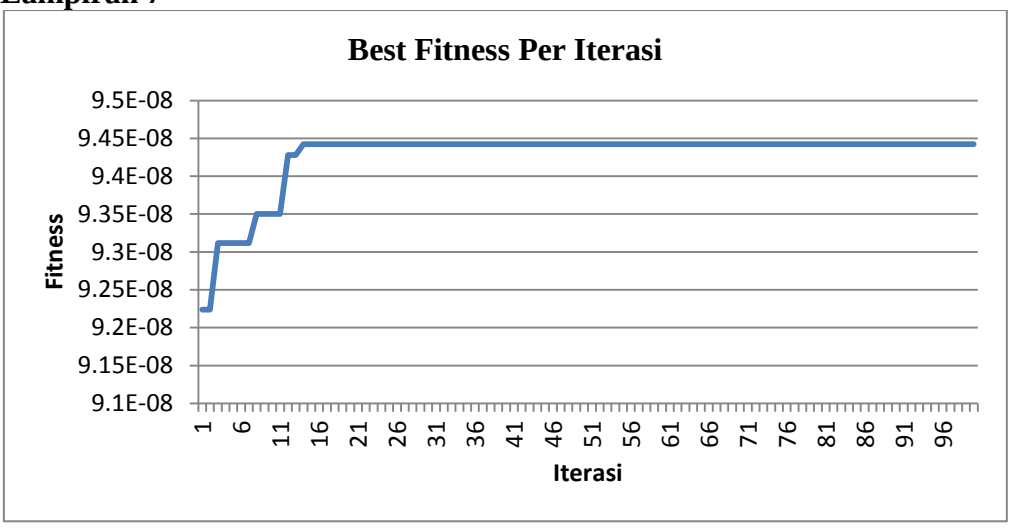
Gambar 5. Alur Proses KPrototype-Genetika untuk center cluster optimal

Lampiran 6



Gambar 6. Perubahan nilai *cost* menurut jumlah *cluster*

Lampiran 7



Gambar 7. Nilai fitness terbaik per iterasi

Lampiran 8

Variabel kategorikal

no objek	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
180	3	1	1	5	5	2	2	2	2	2	2	2	2	2	2	1	4	2	1	2
13158	3	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1	5	2	1	7
26151	3	1	1	1	1	2	2	2	2	2	2	2	2	1	2	1	2	2	1	4
39048	2	1	1	1	1	2	2	2	2	2	2	2	2	2	1	1	4	2	2	2
52138	3	1	2	2	1	2	2	2	2	2	2	2	2	2	2	1	4	2	1	5
65201	3	1	1	2	1	2	2	2	2	2	2	2	2	1	2	1	4	5	1	3

variabel numerik

no objek	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
180	-0.4872	0.2892	-0.5935	-0.1693	0.1555	-0.6512	-0.3237	-0.203	0.6001	-0.2835	-0.0913	-0.125	-0.1352	0.161	-0.0197	-0.3418	-0.1246
13158	0.24763	-0.063	0.13121	-0.1693	0.1463	-0.6512	0.28498	0.4323	-1.116	0.2129	-0.0913	-0.125	-0.1352	-0.047	0.0987	-0.3418	-0.1246
26151	-0.5239	0.2411	-1.0819	-0.1693	0.6412	-0.126	0.45527	0.2744	0.588	-0.2835	-0.0913	-0.125	-0.1352	0.176	-0.5141	-0.3418	-0.1246
39048	-0.2667	-0.64	0.63537	-0.1693	-0.213	-0.6512	-0.0546	0.2465	0.4948	-0.2835	-0.0913	-0.125	-0.1352	-0.047	-0.5141	-0.3418	-0.1246
52138	-0.4504	-0.095	-0.373	-0.1693	-0.379	-0.6512	-0.5677	1.0667	0.6001	-0.2835	-0.0913	-0.125	-0.1352	-0.428	-0.5141	-0.3418	-0.1246
65201	-0.4504	0.6896	-0.5148	-0.1693	-0.234	-0.6512	-0.4333	0.7323	-0.008	-0.2835	-0.0913	-0.125	-0.1352	-0.268	0.1207	0.3074	-0.1246

Gambar 8. Atribut pada objek desa terpilih sebagai inisial center cluster

BAGAIMANA DAYA SAING INDUSTRI *LIFE SCIENCES* DI INDONESIA: SEBUAH PEMBANDINGAN DENGAN NEGARA-NEGARA LAIN

HOW COMPETITIVE IS LIFE SCIENCES INDUSTRY IN INDONESIA: COMPARED TO OTHER WORLD COUNTRIES

Retno Indrawati
Direktorat Statistik Harga - Badan Pusat Statistik

Ernawati Pasaribu
Sekolah Tinggi Ilmu Statistik

Masuk tanggal: 15-12-2015, revisi tanggal: 18-01-2016, diterima untuk diterbitkan tanggal: 19-01-2016

Abstrak

Indonesia adalah negara terbesar di Asia Tenggara dengan lebih dari 20 juta penduduknya adalah kelas menengah yang dewasa ini memiliki pengaruh penting dan semakin menginspirasi. Indonesia telah menjadi pasar yang menarik karena perkembangan pesat jumlah konsumen, khususnya dari kelompok penduduk berpendapatan menengah. Tingginya jumlah populasi juga mengindikasikan besarnya potensi sumber tenaga kerja. Industri Life Sciences (LS), secara luas mulai dikenal sebagai aliran baru ekonomi berbasis ilmu pengetahuan. Studi ini mengidentifikasi posisi relatif Indonesia dikaji dari investasi langsung luar negeri (foreign direct investment-FDI) pada industri LS, sekaligus dari sisi daya saing *(competitiveness)* dengan negara-negara lain di dunia Berdasarkan sektor LS, pesaing utama Indonesia adalah Portugal, Turki, Saudi Arabia, dan Nigeria, sedangkan berdasarkan aktivitas LS,Argentina dan Bulgaria adalah saingan utama.Studi ini juga mengungkapkan bahwa FDI yang masuk ke Indonesia dipengaruhi terutama oleh tingkat inflasi dan return on investment.

Kata kunci : Indonesia, *life sciences*, daya saing, investasi langsung luar negeri

Abstract

Indonesia is the South East Asia's largest economy and has a substantial and increasingly inspirational middle class of over 20 million. Indonesia has become an attractive market due to her strongly growing consumer market, especially the middle income segment. The high number of population also indicates the existing potential pool of labour. Life Sciences (LS) industry is widely recognised as the new wave of knowledge-based economy. This study identifies relative position of Indonesia in terms of foreign direct investment (FDI) in LS industry and competitiveness of the LS industry in Indonesia compared with other countries. Based on LS sector, Indonesia has to compete mainly with Portugal, Turkey, Saudi Arabia and Nigeria, while based on LS activities, Argentina and Bulgaria are the main competitors. It also reveals that FDI inflow to LS industry in Indonesia is influenced mainly by inflation and return on investment..

Keywords : *Indonesia, life sciences, competitiveness, foreign direct investment*

INTRODUCTION

Some countries in Asia and the Pacific, namely China, India, Indonesia, Thailand, and Vietnam have a large population with a growing middle class. These countries attract more market especially in terms of research location and manufacturing base of Life Sciences industry. Also, as a consequence of rising disposable incomes and shift in lifestyle, demands for medicine increase.

Indonesia, as South East Asia's largest economy, has a population of more than 250 million. It has affluent and increasing middle class households of over 20 million. As a reflection of rising disposable incomes, Indonesia's pharmaceutical market has registered double-digit growth since 2009 and by 2016 it is anticipated to rank as the sixth largest pharmaceutical market in the region (Jones Lang, 2012).

In 2011, four countries out of 10 members of the Association of Southeast Asian Nations (ASEAN) namely Brunei Darussalam, Indonesia, Malaysia, and Singapore saw a considerable rise in Foreign Direct Investment (FDI) inflows. As revealed by the United Nations Conference on Trade and Development (UNCTAD) (2012), Indonesia and Thailand are among the top priority host economies chosen by transnational corporations (TNCs). In addition, the possibility of further increase in FDI inflows to the two countries is growing.

The top five prospective host economies 2012-2014 are China, United States, India, Indonesia, and Brazil. This fact proves the importance of developing regions to transnational corporations (TNCs) as locations for international production (UNCTAD, 2012).

Indonesia has shown tremendous economic recovery after the 1997/1998 Asian financial crisis. The GDP (Gross Domestic Product) growth of Indonesia in 1998 was -13.33%. It settled above 4.5% since 2002. In the 2008 Global Financial Crisis, which was began in the US sub-prime mortgage markets, Indonesia had also affected. Depreciation of the rupiah (Indonesian currency) exchange rate by the end of 2008 was 30 percent. Still, Indonesia together with China and India are the only countries experienced with positive growth of GDP. In 2012, the growth of GDP stood at 6.2%. Even USA just acknowledge having better economy after almost 8 years (December 2015) by increasing their Federal rate. According to the World Economic Forum (2012) in Tan and Amri (2013), Indonesia has the 16th largest GDP in the world amounting to US$846.8 billion in 2011. The stable growth of Indonesia's economy over the last decade, along with her progress in transition to democracy, has led to Indonesia as a prosperous and enabling environment for investment.

Despite all of those excellent records about Indonesia, statistics has shown that inward FDI to Indonesia is still relatively modest. Sjoholm dan Lipsey (2010) measure the role of inward FDI- in different East Asian countries by the ratio of the inward stock to GDP, as can seen in Table 1 (Appendix 1).

In 2009 the ratio of inward FDI to GDP for South East Asia was 46.34%, while for Indonesia the ratio was only 13.48%. As seen in table 1, there are only two countries which have ratio of inward FDi below Indonesia namely Taiwan and Korea. In addition, Indonesia also shows poor performance in competitiveness compared to other ASEAN economies. Based on the Global Competitiveness Index ranking, Indonesia ranked 50th (out of 144 countries) in 2012-2013, while Singapore is second and Malaysia is ranked 25th (table 2).

Table 2. Global Competitiveness Index (GCI) ranking

Country	2011-2012	2012-2013
Singapore	2	2
Malaysia	21	25
Brunei	28	28
Thailand	39	38
Indonesia	46	50
Philippines	75	65
Vietnam	65	75
Cambodia	97	85
Lao PDR*)	-	-
Myanmar*)	-	-

Source: The Global Competitiveness Report 2012-2013, *) data not available

The problem statement for this study basically to compare the weaknesess of Indonesia in term of competitiveness and some advantages which can be the factor to increase the competitiveness. Indonesia is an attractive market for FDI in LS industry due to its strongly growing consumer market, especially the middle income segment. The high population also indicates the potential pool of labor. Since FDI has a significant role in accelerating economic growth due its many benefits to receiving country, economies have been competing for attracting FDI.

This study identifies several location factors as the main determinants of LS industry. It also measuring the competitiveness of Indonesia's LS industry and investigating which country is the

main competitor, which are important to design a proper strategy to attract FDI.

LITERATURE REVIEW

Competitiveness

There are many concepts to measure competitiveness viewed from different perspective.

Wignaraja (2002) distinguished the competitiveness by macroeconomic, business strategy, as well as technology and innovation perspectives. He argued that macroeconomic perspective, which has been widely used to measure competitiveness in developed and developing countries, gives an incomplete framework for structuring public policies.

While Storper (1997) defines competitiveness as "*the ability of an (urban) economy to attract and maintain firms with stable or rising market shares while maintaining standards of living for those participating in it"* (Storper 1997, p.20). He also mentioned indicators of city ability to attract investments, such as investment climate, infrastructure availability, capacity of innovation and learning, the business environment, productivity, standard of living/quality of life and top down/sector and macro influences.

Competitiveness in term of national scale is explained by Onsel et al. (2008. It defines competitiveness as productivity of a country which produces goods and services under free and fair market. Those production are meet the international market standards and could increase the real income of its citizen.This concept also includes the set of institutions, policies, and factors that determine the level of productivity of a country.

Some scholars defend that competition among cities are exist in terms of investment. According to Alderson and Beckfield (2004), the level of cities is determined by the ability to attract investments and how they take control of the world economy.

Likewise, Gordon (1999) proposed that product markets, FDI, and hosting of high profile events are among various fields which cities could compete. Phillips and Ryan (2007) argued that the global life-science research has been significantly transformed. The main reasons behind this transformation are the complexity and specialization of this field which makes it difficult to isolate. The second reason is the extension of intellectual property (IP) rights into new subject areas and new jurisdictions.

Foreign Direct Investment (FDI)

Foreign private investment can be distinguished by FDI and portfolio investment. This research only discusses the FDI, which categorized into outward FDI and inward FDI. Outward FDI is direct investment abroad, whereas inward FDI is direct investment coming from abroad in to this country. UNCTAD (2007) defined FDI as " *an ivestment involving a long-term relationship and reflecting a lasting interest and control by a resident entity in one economy (foreign direct investor or parent enterprise) in an enterprise resident in an economy other than that of the foreign direct investor (FDI enterprise or affiliate enterprise or foreign affiliate)".*

Why is it Important to Find Out The Determinants of FDI?

For many developing countries which do not have access to international capital market, attracting FDI is important. As mentioned by Chakrabarti (2001) and Asiedu (2002), most of the developing countries rely on two forms of foreign financing: FDI and official loans. The latter has been a problem for heavily indebted countries due to the 'debt overhang' in 1982-1983 (break down of normal financial relations). This led to the decline in official lending, foreign aid, investment, and growth rates in those countries. This backdrop revealed the

importance of FDI as provider of capital for investment.

Khondoker and Mottaleb (2007) argued that FDI has a significant role in rapid economic growth by bridging the gap between domestic savings and investment. It is also bringing the latest technology and management know-how from developed countries to developing and even to under-developed countries. Foreign investment offers many benefits to host country, such as enhancing its efficiency since the existence of foreign firms increase competition. Also, from the workers' side, it may support the increasing income by providing higher wage and salary in the host countries.

Crespo and Fontura (2007) mentioned other benefits of FDI to host country: providing capital foreign exchange, technology, competition, and raising access to foreign market. Azam and Lukman (2010) revealed FDI as an important factor for national economic development by transferring innovative technology, up to date management, and marketing techniques to the host countries.

Studies on Determinants of FDI

As explained earlier, the importance thing to analyze the determinants of FDI in terms of economic growth has led scholars to do a lot of empirical studies. They concentrate more on location factors rather than the capital propriety advantages (Nonnemberg and Mendonca, 2004). This is because the capital propriety advantages as *push factor* are more difficult to analyze as it heavily involves firms in its survey. Several empirical studies on different determinants and observed effect on FDI are presented in table 3 (Appendix 2).

Life Sciences Industry

Stremersch and Van Dyck (2009) defined LS industry as an industry that comprises companies in pharmaceuticals, biotechnology, and therapeutic medical devices, and it forms the innovative producer side of the health care industry. Two basic dimensions that underlie the LS industry are science-based knowledge ("know-why") and quality of life.

Gertler and Vinodrai (2009) noted that activity related with life science is expected to produce employment and to raise income for regions and nations, contributing to their economic competitiveness and prosperity, and to generate highly skilled and well-paying jobs. Therefore, academics and policymakers have paid more concerns to understanding the enabling conditions, institutional forces and policy mechanisms that have nurtured and developed the innovative capacity and economic success of LS industry activities in particular regions and nations, as argued by Gertler and Vinodrai (2009).

FDI and Competitiveness in Indonesia

In Indonesia, study and analyses related to competitiveness have also been conducted. The ranking of Investment Climate for 33 Indonesian provinces was provided by Regional Autonomy Watch (*Komite Pemantauan Pelaksanaan Otonomi Daerah*-KPPOD) and Indonesia Investment Coordinating Board (BKPM) in 2008 (KPPOD, 2008a). This ranking was based on six indicators: investment services, investment promotion, commitment of provincial government to the private sector, infrastructure, labor, and accessibility to land. KPPOD (2008b, 2011) also measured rankings in the city and district level based on surveys to business operators in more than 240 cities and districts through the Local Economic Governance.

METHODOLOGY

The Purpose of The Study

The purpose of this research is to identify the relative position of LS industry in Indonesia in terms of FDI and competitiveness, by using descriptive and explanatory analysis. Both of those

analyses are quantitative approach by processing the existing raw data using software: UCINET, SPSS version 20, and EViews version 5.1.

The World Data Set

This study comprises all the firms, cities, and countries across the world in the database of 'FDI Markets', particularly in the cluster of LS industry. Combined with the location factors data set from Global Competitiveness Report, 117 countries from seven regions (Africa, Asia and Pacific, Middle East, West Europe, Rest of Europe, North America, and Latin America) are being analysed.

FDI in this study is based on green field data, because it is a kind of investments where parent companies start an entirely new venture in a foreign country by constructing new operational facilities from the ground up. Therefore it indicates traceable developments between firms and are beneficial in studying their impact on regional development (Wall and Burger, 2012). Another constraint with the purchased data is that roughly 60% of the investment values are not known and have therefore been estimated by FDI Markets. By using a high degree of estimated data in the analyses, the results could possibly be misleading. The solution is using the number of investments instead of the value of investment as a proxy.

Indonesian Data Set

Previous study about determinants of inward FDI in India, Indonesia and Pakistan was written by Azam and Lukman (2010). They found out that determinants of inward FDI in Indonesia do not match with those of Pakistan and India. Almost all of the results of determinants of inward FDI in Indonesia are statistically insignificant. This study assessed the similar variables used by Azam and Lukman to analyze the determinants of inward FDI in LS industry in Indonesia. All variables are compiled for each Indonesian province, including variables that were not processed in the previous study (government consumption, infrastructure, tax, and return on investment). Variable of external debt is the only one excluded due to the difficulty of finding external debt data in the provincial level.

The panel data methodology was used, which combines information on the variation of the Indonesian provinces. It comprises 29 provinces out of 33 provinces and covers period from 2003 to 2011. Four provinces namely Riau Archipelago, West Sulawesi, North Maluku, and West Papua were dropped because they do not have data on FDI in LS industry.

The selected explanatory variables were trade openness (TO), market size (MS), domestic investment (DI), infrastructure expenditure (IE), government consumption (GC), tax (TAX), inflation (INF), and return on investment (RI). The dependent variable (Y variable) was data of inward FDI value which were collected from the Indonesia Investment Coordinating Board (BKPM).

This study utilised panel (longitudinal) data which is defined as data set that follows a given sample of individuals over time, and thus provide multiple observations on each individual in the sample (Hsiao, 2003, p.2). Panel data were distinguished between balanced and unbalanced data. In panel data, variables of the same cross-sectional subject are observed over time. Let $i = 1, 2, \ldots, N$ be an index of the cross-sectional subject and $t = 1, 2, \ldots, T_i$ be an index of time for subject i.

A panel is called balanced if each cross-section subject has the same number of observations. That is, if $T_i = T$ for $i = 1, 2, \ldots, N$ and the total number of observations is

$n = NT$

If each individual subject has a different number of observations over time, that is $T_i \neq Tj$, then we have an unbalanced panel. The total number of observations for unbalanced panel is

$n = \sum_{i=1}^{N} T_i$.

Also if $N > T$, it is called a short panel and if $N < T$, then it is called a long panel.
Generally panel data regression model is written as

$$Y_{it} = \alpha_{it} + \beta' X_{it} + \mu_{it} ,$$
$$i = 1, 2, \ldots, N , \quad t = 1, 2, \ldots, T$$

Where Y_{it} is the dependent variable of the individual i at time t, intercept α_{it} is an effect of individual i at time t, variable β' is constant vector $K \times 1$, X_{it} is a $K \times 1$ vector of explanatory variables, and μ denotes error regression of individual i at time t. Panel data analysis has three approach methods namely Pooled Least Square (PLS), Fixed Effect (FE) and Random Effect (RE).

How to Measure Relative Position and Competitiveness

The ranking of countries is developed by processing number of FDI in LS industry using excel software. It is also classified by types of investments (outward and inward), each of them has been analysed based on region and country.

The main competitors of Indonesia in LS industry are answered with the results of Manhattan Distance analysis. This analysis measuring the distance between two points which is calculated by summing the absolute differences of their coordinates, using UCINET software by processing matrix of number of FDI in LS industry and name of countries.

RESEARCH FINDINGS

In terms of three main sectors in the LS industry, Indonesia had only two outward FDI which were only in pharmaceutical sectorscomparing with other ASEAN member, after Malaysia, Singapore, and Thailand. Table 4 shows that Indonesia has 56th position among 66 countries worldwide, classified by number of outward FDI in three main sectors. Table 5 shows that as destination country, Indonesia came with better result, rank 35 out of 117 countries with inward FDI in three main sectors. (Appendix 3 and 4).

As seen in Table 6 and Table 7 (Appendix 5 and 6), by three main activities, Indonesia ranked 54 out of 66 as source country (outward FDI). As a destination country by activities, Indonesia stood at rank 36 out of 117.

The Main Competitors of Indonesia

Table 8 below illustrates the 20 competitor countries of Indonesia (by sectors). As can be seen, viewed from outward FDI, Indonesia has two competitors from Asia and the Pacific region: Australia and Philippines. From Africa region the competitors are Algeria, Nigeria, Egypt, and South Africa. From the region of Latin America there are two competitors of Indonesia, namely: Argentina and Colombia. Israel and Saudi Arabia are the competitors from Middle East region. From West Europe, Indonesia has Portugal, Finland, and Denmark as competitors. Several countries from the Rest of Europe also become the competitors of outward FDI, with the top namely Romania, Turkey, and Serbia.

Table 8. Competitors of Indonesia by Sectors in LS industry

No.	Sectors	
	Outward FDI	Inward FDI
1	Argentina	Bulgaria
2	Romania	Portugal
3	Portugal	Nigeria
4	Turkey	Turkey
5	Serbia	Finland
6	Slovakia	Israel
7	Australia	Chile
8	Algeria	Saudi Arabia
9	Bulgaria	Serbia
10	Saudi Arabia	Algeria
11	Nigeria	Egypt
12	Slovenia	Ukraine
13	Colombia	Philippines
14	Ukraine	Qatar
15	Egypt	Malaysia
16	Finland	Taiwan
17	Philippines	Slovenia
18	Denmark	Croatia
19	South Africa	Tunisia
20	Israel	Malta

As the destination countries, Indonesia has several competitors from various regions. Philippines, Malaysia, and Taiwan become competitors from Asia and the Pacific. From Africa, Nigeria and Egypt also become competitors in terms of outward FDI. The main competitors of Indonesia are Portugal, Turkey, Saudi Arabia, and Nigeria.

Table 9 describes the competitors of Indonesia by activities. Argentina as can be seen clearly is the main competitor, both as source and destination country of FDI. In terms of outward FDI, only Vietnam and Sri Lanka are the competitors of Indonesia from Asia and the Pacific region. While in terms of inward FDI in the same region; Philippines, Australia, and Taiwan are the main competitors of Indonesia. It can be concluded that the main competitors of Indonesia by activities in LS industry are Argentina and Bulgaria.

Table 9. Competitors of Indonesia by Activities in LS Industry

No.	Sectors	
	Outward FDI	Inward FDI
1	Argentina	Argentina
2	Bulgaria	Portugal
3	Vietnam	Nigeria
4	Jordan	Philippines
5	Lithuania	Saudi Arabia
6	Chile	Serbia
7	Ghana	Algeria
8	Malta	Slovakia
9	Puerto Rico	Australia
10	Sri Lanka	Bulgaria
11	Macedonia FYR	Egypt
12	Mexico	Slovenia
13	Norway	Ukraine
14	Portugal	South Africa
15	Serbia	Turkey
16	Ukraine	Colombia
17	Colombia	Croatia
18	Egypt	Dominican Rep.
19	Kenya	Finland
20	Liechtenstein	Taiwan

Panel Regression Analysis

This section presents the econometric results of the determinants for inward FDI in LS industry in Indonesia. The data used are unbalanced panel data comprising time series data from 2003-2011 (trade openness, market size, domestic investment, inflation, return of investment) and three variables which are only available from 2003-2008 (infrastructure, government consumption, indirect tax). The cross-section data only used 29 provinces in Indonesia which had FDI in LS industry

Regression model for inward FDI by province in Indonesia year 2003-2011 is

$y_{it} = \alpha_{0i} + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \beta_4 X_{4it} + \beta_5 X_{5it} + \beta_6 X_{6it} + \beta_7 X_{7it} + \beta_8 X_{8it}$

i = 1, 2, …, 29 (number of province as individual sample unit)

t = 1, 2, …, 9 (number of year observation)

with

y_{it} = value of inward FDI in LS industry

X_{1it} = trade openness (TO)

X_{2it} = market size (MS)

X_{3it} = domestic investment (DI)

X_{4it} = infrastructure expenditure (IE)

X_{5it} = government consumption (GC)

X_{6it} = taxes (TAX)

X_{7it} = inflation (INF)

X_{8it} = return on investment (RI)

Determine the Estimation Method

1). Pooled Least Squares (PLS) method will be used to develop the regression model for FDI inward

FDI = -2143.37 – 16.98 TO – 35.93 MS + 0.53 DI + 3.30 IE – 2.01 GC + 4.80 TAX + 55.58 INF + 4248.56 RI

(0.88) (0.01) (0.10) (0.00) (0.05) (0.00) (0.44) (0.24)

*) The value between brackets represent the t sig

As can be seen from individual test (t-test probability) there are four variables found significant. MS (market size), IE (infrastructure expenditure), and TAX are significant at 5% level of significance, while GC (government consumption) is significant at 10% level of significance. Two variables are insignificant, which are TO (trade openness) and INF (inflation) The empirical results obtained are acceptable and significant on the basis of R-squared (R^2) 0.67. The Durbin-Watson

statistics is 1.84 (close to 2), shows no autocorrelation problem.

2). Fixed Effects Method (FEM) has been assessed to calculate for possible unobserved heterogeneity across provinces. The regression model is:

FDI = -346.63 + 11.07 TO – 24.33 MS + 0.04 DI + 0.44 IE – 0.33 GC + 0.84 TAX + 5.76 INF + 650.86 RI

(0.64) (0.25) (0.62) (0.23) (0.15) (0.18) (0.00) (0.00)

*) The value between bracket represent the t sig

INF (inflation) and RI (return of investment) are significant at 5% level of significance. Here the value of R-squared (R^2) is 0.76 and it is higher than result of R^2 from PLS method. Similar with PLS method, FEM also highly significance which shown by F-stat value (0.00000). The Durbin-Watson statistics is 2.45.

3). Since the results from those two approaches are somewhat significant, restricted F-test should be implemented to determine which method will be better to use. The F-test hypothesis is as follows

H_0: Pooled Least Squares Model (restricted)

H_1 : Fixed Effect Model (unrestricted)

Table 10. Redundant Fixed Effects Test

Pool: FEM
Test cross-section fixed effects

Effects Test	Statistic	d.f.	Prob.
Cross-section F	5.202851	(28,125)	0.0000

The p-values associated to the F-statistic is 0.0000, which provides strong evidence against the null hypothesis meaning FEM should be used to estimate panel regression model.

4). Random Effect Method (REM) takes the residual error into account using least square method.

FDI = -2129.77 + 9.76 TO – 36.61 MS + 0.13 DI + 3.95 IE – 2.62 GC + 5.37 TAX + 52.89 INF + 3264.12 RI

(0.94) (0.23) (0.74) (0.00) (0.16) (0.03) (0.50) (0.48)

*) The value between bracket represent the t sig

The result obtained is only two variables are significant at 5% level of significance, IE (infrastructure expenditure) and TAX. The value of R-squared is 0.44, REM is also highly significance which shown by F-stat value (0.00000). The Durbin-Watson statistics is 1.99 which indicated no autocorrelation.

5). The REM assumes that random effects are uncorrelated with the explanatory variables. Hausman test should be used to determine whether FEM or REM more suitable to estimate the model. The hypothesis of Hausman Test is:

H_0: Random effect (RE)

H_1: Fixed effect (FE)

Table 11. Hausman Test

Correlated Random Effects - Hausman Test
Pool: FEM
Test cross-section fixed effects

Summary	Chi-Sq. Statistic	Chi-Sq. d.f	Prob.
Cross-section F	4.468903	8	0.8125

As can be seen on the Table 11 above, the test fails to reject the null hypothesis at 5% level of significance. Meaning that the assumption that the random effects should be uncorrelated to the explanatory variables is true for this dataset. Therefore the panel regression model should be estimated by using the REM method.

6) At last, this research should compare the statistical results between FEM and REM to determine which one is the most suitable model.

Table 12. Comparison of Statistical Result between FEM and REM

Model	Fixed Effects Method (FEM)	Random Effects Method (REM)
R- Squared	0.76	0.44
Adjusted R-Squared	0.68	0.41
Prob (F-Statistic)	0.00	0.00

Based on Table 12, Statistical result for FEM shows that this model is the best to be used as estimator tool for panel regression. It also beneficial using this model since the different characteristic of each individual sample and time series are taken into account.

Estimation Model Panel Regression for FDI

Based on the several test in the previous section, FEM has been found more efficient than REM. Then the estimated model panel regression for FDI is:

Inward FDI = -346.63 + 11.07 TO – 24.33 MS + 0.04 DI + 0.44 IE – 0.33 GC + 0.84 TAX + 5.76 INF + 650.86 RI

FEM allows us to explore the relationship between predictor variable (X variables) and outcome variables within province. When using FEM we assume that something within the province may impact or bias the predictor or outcome variables and we need to control it. FEM removes the effect of those time-invariant characteristics from the predictor variables so that we can assess the predictors' net effect. The fixed-effects model controls for all time-invariant differences between the individuals, so the estimated coefficients of the fixed-effects models cannot be biased because of omitted time-invariant characteristics (i.e culture, environment)

Another significant assumption of the FEM is that those time-invariant characteristics are unique to the province and should not be correlated with other province characteristics. Each province is different therefore the entity's error term and the intercept represent as a constant (which captures provincial characteristics) should not be correlated with the others.

From eight variables, only two variables are found significant,they are Inflation (INF) and Return on investment (RI) The other variables are not significant when to be tested partially, meaning that if they stand alone as a determinant of FDI, they are insignificant. Those variables had been statistically significant as the determinant of LS industry when they are assessed simultaneously, and represents by the R-square of FEM (0.67) and F stat-value (0.00).

Inflation is found as significant with expected positive sign. Azam & Lukman (2010) also found a positive relationship between inflation rates and inward FDI. Return on investment, with proxy 1/GDRP, is found significant with positive expected sign. Tsai (1994) and Azam & Lukman (2010) also found positive significant relationship.

Trade openness is find insignificant with expected positive sign. Schmitz & Bieri (1972), Wheeler & Moody (1992) also found insignificant relationship between trade openness and FDI.

Market size had been found insignificant with expected positive sign. The previous study that also found positive relationship between market size and FDI are from Chakrabarti (2001, 003), Ioannatos (2003), Banga (2003), and Eli et al., (2006). Domestic investment had been found insignificant with expected positive sign. The similar findings are from Razin (2003) and Yasmin et al.,(2003).

Infrastructure Expenditure and Government consumption had been found insignificant with expected positive sign. Tax had been found insignificant with unexpected positive sign. Only Wheeler & Mody (1992), Jackson & Markowski (1995), Yulin & Reed (1995) and Porcano & Price (1996) which had similar result. Meanwhile as can be seen in table 3 above, many researches found out taxes has negative effect on FDI.

Table 13. Intercept estimation ($\hat{\alpha}_{0i}$) of each province for FEM with cross section weight

No	Province	$\hat{\alpha}_{0i}$
1	NAD	-1178.32
2	Sumatera Utara	-2686.48
3	Sumatera Barat	-741.17
4	Riau	1486.37
5	Jambi	-244.46
6	Sumatera Selatan	-1166.80

7	Bengkulu	-44.72
8	Lampung	-439.64
9	Bangka Belitung	-57.62
10	DKI Jakarta	15756.52
11	Jawa Barat	2281.28
12	Jawa Tengah	-3990.92
13	DI Yogyakarta	-253.60
14	Jawa Timur	-3891.21
15	Banten	1657.82
16	Bali	-683.47
27	Nusa Tenggara Barat	-363.54
18	NusaTenggara Timur	-115.74
19	Kalimantan Barat	-491.15
20	Kalimantan Tengah	-197.80
21	Kalimantan Selatan	-692.93
22	Kalimantan Timur	-1846.50
23	Sulawesi Utara	-86.25
24	Sulawesi Tengah	-126.87
25	Sulawesi Selatan	-852.92
26	Sulawesi Tenggara	-114.06
27	Gorontalo	-8.05
28	Maluku	-21.59
29	Papua	-509.24

The fixed-effects parameters, α_i, capture the net effects of all variables, both observable and unobservable, that vary across provinces but are constant over time. Constant intercept for this model is -346,6247, therefore we have to sum up this intercept with the province's intercept as presented in table 13 to develops model for each province.

Indonesia has 33 provinces; unfortunately only 29 provinces could be examined. Kepulauan Riau, Papua Barat, Sulawesi Barat, and Maluku Utara are excluded. This panel regression analysis requires time series data of the dependent variable (Y). Since those four provinces in certain year within period 2003 to 2012 did not receive FDI, so they were excluded from analysis.

The constant value ($\hat{\alpha}_{0i}$) of intercept for each of Indonesian province ranged from 15409,9 (DKI Jakarta) and -4337.5 (Jawa Tengah). Only four out of 29 provinces that have positive intercept, namely: DKI Jakarta (15409.9), Jawa Barat (1934.66), Banten (1311.2), and Riau (1139.75). DKI Jakarta, Jawa Barat, and Banten are also having highly competitiveness ranking compared to other provinces.

Three of those provinces are located in Java, and only Riau is located in Sumatra Island. As reported in Life Sciences Cluster Report 2012, Indonesia has 55 industrial park firms but unfortunately none of them are dedicated fully to the LS industry. Java, which is seen as a destination option for industry, has about 75 percent of Indonesian's industrial estate. 50 percent of them are located in Jawa Barat province. The biggest pharmaceutical company also located in Greater Jakarta Industrial estate, covers Tangerang, Bogor, Bekasi, and Karawang.

Riau has a relatively higher intercept because of their positions in Sumatra which is near Singapore as one of the biggest receiver of FDI, and also becomes province in Indonesia which has the free tax policy. This policy has been a factor that enhances inward FDI to Riau Province.

Surprisingly, Jawa Timur and Jawa Tengah -ranked 2nd and 3rd in competitiveness- had the lowest intercept compared to other provinces. This results need further research to find out whether this condition only exist in LS industry or else.

CONCLUSIONS AND RECOMMENDATION

Conclusions

This research aimed to find the most competitors of Indonesia in the LS industry. As the source country of FDI, Indonesia has two competitors from Asia and the Pacific region, namely: Australia and Philippines. From Africa region the competitors are Algeria, Nigeria, Egypt, and South Africa. Latin America also becomes the competitors of Indonesia, with Argentina and Colombia as the countries. Israel and Saudi Arabia are the competitors from Middle East region. From West Europe, Indonesia has Portugal, Finland, and Denmark as

competitors. Several countries from Rest of Europe also become the competitors of outward FDI, namely Romania, Turkey, and Serbia.

As the destination country, Indonesia has Philippines, Malaysia and Taiwan as the competitors. Overall, based on LS sectors Indonesia has Portugal, Turkey, Saudi Arabia and Nigeria as the main competitors. Based on activities, Indonesia has Argentina and Bulgaria as the main competitors. Whilst in terms of outward FDI, the competitors of Indonesia are Argentina, Bulgaria, Vietnam, Jordan, and Lithuania. By inward FDI, there are Argentina, Portugal, Nigeria, Philippines, and Saudi Arabia as the competitors. This study revealed that FDI inward in LS industry in Indonesia influenced mainly by inflation (INF) and return on investment (RI). It clearly shows that for Indonesia macroeconomic variables (inflation) and return on investment have significant relationship with inward FDI.

The dynamics of price of primary goods in Indonesia -which is reflected on inflation rates-tends to be viewed as an opportunity for inward FDI. Return on investment (RI) also significantly positive affected inward FDI. The different result is: GDP is found positive affected inward FDI in the world model but as likely negative determinants for Indonesia, since it is used as a proxy for return on investment variable (1/GDRP). In Indonesia, the increase of inward FDI will be gained coherent with increasing price of primary goods.

This study had also found that provinces with higher ranking of competitiveness, such as Jawa Timur and Jawa Tengah had the lowest intercept of estimates model compared to other provinces. This results need further research to find out whether this condition only exist in LS industry or anything else.

Recommendation

The disparity in economic and social sector between province in Java and other islands is the main issue for attracting FDI. The empirical study resulted that only four provinces has a positive intercept as the host of FDI, three of them are located in Java, since Java provides better infrastructure, higher skilled labor, better facilities of science, etc.

As argued by Sethi et al. (2003) FDI brings several benefits for the host country, such as the inflow of capital, the creation of job opportunities, transfer of technological knowledge —which is translated into the development of skilled workers—, higher productivity, and higher value-added activities. These advantages will enhanced the income distribution among Indonesian provinces will be diatributed evenly.

The economic structure of Indonesia is now primarily focused on agriculture and industries which extract and utilize natural resources. Industries which is focused on products with significant added value are still limited. There is no other way to attract more FDI in LS industry in Indonesia, but improvement on infrastructure and human capital resources. The Masterplan for Acceleration and Expansion of Indonesia's Economic Development (abbreviated MP3EI)[1] is expected to be the solution to accelerate and expand economic development among regions.

[1] The Masterplan for Acceleration and Expansion of Indonesia's Economic Development (abbreviated MP3EI) is an ambitious plan by the Indonesian government to accelerate the realization of becoming a developed country. It aims to established Indonesia as one of the world's developed countries by 2025.

REFERENCES

Asiedu, E.2002. On the Determinants of Foreign Direct Investment to Developing Countries: Is Africa Different?. *World development* Vol. 30, No. 1, pp. 107-119.

Azam, M and Lukman, L.2010. Determinants of Foreign Direct Investment in India, Indonesia and Pakistan: A Quantitative Approach. *Journal of Managerial Sciences.* Vol. IV, No. 1, pp. 31-44.

Baltagi, B.H. Econometric Analysis of Panel Data. John Wiley and Sons.1995.

Chakrabarti, A.2001. The Determinants of Foreign Direct Investment: Sensitivity Analyses of Cross-Country Regressions. KYKLOS. Vol. 54, pp. 89-114.

Crespo, Nuno and Fontoura, Maria P.2007. "Determinant Factors of FDI Spillovers- What do We Really Know?". *World Development*. Vol.35, No. 3, pp 410-425.

Gertler, M.S., and Vinodrai, T.2009. LS industry and Regional Innovation: One Path or many?. *European Planning Studies.* Vol.7, No.2.

Gordon. I.1999. Internationalization and Urban Competition. *Urban Studies.* Vol. 36. No. 5-6 pp. 1001-1016.

Jones Lang LaSalle.2012. LS Cluster Report-Global.

Khondoker and Mottaleb, A.2007. Determinants of Foreign Direct Investment and Its Impact on Economic Growth in Developing Countries. MPRA Paper No. 9457.

Komite Pemantauan Pelaksanaan Otonomi Daerah.2008b. *Local economic governance in Indonesia*: *A survey of businesses in 243 regencies/cities in Indonesia, 2007.* Jakarta, Indonesia: Komite Pemantauan Pelaksanaan Otonomi Daerah. Retrieved from http://www.kppod.org/datapdf/laporan/rating2007_eng.pdf

Nonnemberg, M.B., and de Mendonca, M.J.C.2004. The Determinants of Foreign Direct Investment in Developing Countries.

Phillips, P.W.B., and Ryan, C.D.2007. The Role of Clusters in Driving Innovation. Handbook of Best Practices. Chapter 3.11, pp. 281-294.

Sethi, D., Guisinger, S.E., Phelan, S.E. and Berg, D.M.2003. 'Trends in foreign direct investment flows: a theoretical and empirical analysis'. *Journal of International Business Studies*. Vol. 34(4), pp. 315–326.

Sjoholm, F, and Lipsey, R.E.2010.FDI and Growth in East Asia:Lessons for Indonesia.

Storper, M.1997. *The Regional World - Territorial Development in a Global Economy*. New York: Guilford Press.

Stremersch, S., and Dyck, W.V.2009. Marketing of the LS: A New Framework and Research Agenda for a Nascent Field. *Journal of Marketing* Vol. 73 (July 2009), pp. 4–30

Sum, N.L., and Jessop, B.2013. Competitiveness, The knowledge-based economy and Higher Education.

Tan, K.G, and Amri, M.2013. Subnational Competitiveness and National Performance: Analysis and Simulation for Indonesia. *Journal of Centrum Cathedra.* Vol. 6. Issue 2, pp 173-192

UNCTAD.2007. World Investment Report 2007: Transnational Corporations, Extractive Industries and Development.

Wall, R.S. and Burger, M.,2012. Research report: De Strijd om Kapitaal. Den Haag: Province of South-Holland.

Wignaraja, G.2002. Creating value: From comparative to competitive advantage. Competitiveness Strategy in Developing Countries. Executive Forum on National Export Strategies.

APPENDIX

Appendix 1

Table 1.The Stock of inward FDI as percent of GDP, year 1980-2009

	1980	1985	1990	1995	2000	2005	2009
China/HK	53.37	55.01	46.33	36.47	47.44	32.95	27.06
Taiwan	5.69	4.62	5.91	5.75	6.08	12.13	12.75
Indonesia	5.73	5.98	6.95	9.32	15.2	14.41	13.48
Korea	1.78	1.87	1.97	1.84	7.45	13.25	13.31
Malaysia	20.33	22.8	22.57	31.15	56.24	32.23	39.01
Philippines	2.82	5.98	10.22	13.69	23.92	15.17	14.63
Singapore	45.66	60.03	82.57	78.21	119.26	162.44	193.98
Thailand	3.03	5.14	9.66	10.53	24.38	34.24	37.52
Vietnam	59.1	30.25	25.49	34.48	66.07	58.93	51.93
Northeast Asia	41.85	38.91	25.9	20.96	32.11	26.01	25.35
Southeast Asia	9.44	12.54	18.09	22.46	44.47	44.8	46.34

Source: Lipsey and Sjoholm, 2010

Appendix 2

Table 2. Determinants of FDI and observed effect on FDI

Determinants of FDI	Positive effect	Negative effect	Insignificant
1. Market size	Bandera & White (1968) Schmitz & Bieri (1975) Swedenborg (1979) Lunn (1980) Dunning (1980) Root & Ahmed (1979) Kravis & Lipsey (1982) Nigh (1985) Schneider & Frey (1985) Culem (1988) Papanastassiou & Pearce (1990) Wheeler & Mody (1992) Sader (1993) Tsai (1994) Shamsuddin (1994) Billington (1999) Pistoresi (2000)		
2. Inflation rate		Garibaldi et al (2001) Naeem, Ijaz & Azam (2005) Azam & Lukman (2010)	
3. Domestic investment	Razin (2003) Yasmin *et al*. (2003) Naeem, Ijaz & Azam (2005) Azam & Lukman (2010)		
4. Trade openness	Kravis & Lipsey (1982) Culem (1988) Edwards (1990) Gastanaga *et al.* (1998) Pistoresi (2000) Hausmann & Fernandez-Arias (2000) Aseidu (2002) Ioannatos (2003) Azam & Lukman (2010)		Schmitz & Bieri (1972) Wheeler & Mody (1992)
5. Government consumption			Azam & Lukman (2010)
6. Infrastructure	Wheeler & Mody (1992) Kumar (1994) Loree and Guisinger (1995) Aseidu (2002) Ioannatos (2003) Azam & Lukman (2010)		

7. Taxes and tariffs	Swenson (1994)	Hartman (1984) Grubert & Mutti (1991) Hines & Rice (1994) Loree & Guisinger (1995) Guisinger (1995) Cassou (1997) Kemsley (1998) Barrel & Pain (1998) Gastanaga *et al.* (1998) Billington (1999) Wei (2000)	Wheeler & Mody (1992) Jackson & Markowski(1995) Yulin & Reed (1995) Porcano & Price (1996)
8. Return on investment	Tsai (1994) Azam & Lukman (2010)		

Source: Compiled from Chakrabarti (2001), Asiedu (2002), Azam & Lukman (2010)

Appendix 3

Table 4. Rank of ASEAN Countries by Outward FDI in Three Main Sectors period 2003-2012

World Ranking	Source Countries	Number of FDI			Total
		Pharmaceuticals	Medical Devices	Healthcare	
18	Malaysia	2	10	26	38
20	Singapore	2	11	16	29
47	Thailand	0	0	5	5
56	Indonesia	2	0	0	2
60	Philippines	0	0	1	1
62	Vietnam	1	0	0	1
*)	Cambodia	0	0	0	0
*)	Myanmar	0	0	0	0
*)	Laos	0	0	0	0
*)	Brunei	0	0	0	0

*) not having outward FDI

Appendix 4

Table 5. Rank of ASEAN Countries by Inward FDI in Three Main Sectors period 2003-2012

World Ranking	Destination Countries	Number of FDI			Total
		Pharmaceuticals	Medical Devices	Healthcare	
8	Singapore	71	27	11	109
26	Vietnam	16	12	5	33
30	Thailand	9	16	3	28
31	Malaysia	10	10	5	25
35	Indonesia	9	3	8	20
50	Philippines	4	3	2	9
69	Cambodia	1	1	2	4
106	Laos	0	0	1	1
108	Myanmar	1	0	0	1
*)	Brunei Darussalam	0	0	0	0

*) not having inward FDI

Appendix 5

Table 6. Rank of ASEAN Countries by Outward FDI in Three Main Activities period 2003-2012

World Ranking	Source Countries	Number of FDI			Total
		Manufacturing	Sales, Marketing & Support	Research & Development	
21	Singapore	9	8	0	17
22	Malaysia	9	6	1	16
54	Indonesia	1	0	1	2
55	Thailand	0	1	0	1
56	Philippines	0	1	0	1
62	Vietnam	1	0	0	1
*)	Cambodia	0	0	0	0
*)	Myanmar	0	0	0	0
*)	Laos	0	0	0	0
*)	Brunei Darussalam	0	0	0	0

*) not having outward FDI

Appendix 6

Table 7. Rank of ASEAN Countries by Inward FDI in Three Main Activities period 2003-2012

World Ranking	Destination Countries	Number of FDI			Total
		Manufacturing	Sales, Marketing & Support	Research & Development	
8	Singapore	40	28	52	120
27	Thailand	20	5	4	29
28	Vietnam	15	14	0	29
32	Malaysia	15	4	6	25
36	Indonesia	10	7	1	18
54	Philippines	2	6	0	8
96	Myanmar	1	0	0	1
*)	Laos	0	0	0	0
*)	Cambodia	0	0	0	0
*)	Brunei Darussalam	0	0	0	0

*) not having inward FDI

ANALISIS *MULTIVARIATE ADAPTIVE REGRESSION SPLINES* (*MARS*) PADA PREDIKSI KETERTINGGALAN KABUPATEN TAHUN 2014

MULTIVARIATE ANALYSIS ADAPTIVE REGRESSION SPLINES (MARS) ON PREDICTION THE UNDERDEVELOPED DISTRICT IN 2014

Siskarossa Ika Oktora
Sekolah Tinggi Ilmu Statistik

Masuk tanggal: 04-12-2015, revisi tanggal: 15-01-2016, diterima untuk diterbitkan tanggal: 19-01-2016

Abstrak

Penelitian ini bertujuan untuk membentuk model kabupaten tertinggal dan melakukan prediksi ketertinggalan kabupaten pada tahun 2014 berdasarkan kriteria perekonomian masyarakat, SDM, infrastruktur, kemampuan keuangan daerah, aksesbilitas, dan karakteristik daerah dengan metode MARS. MARS adalah salah satu metode pengklasifikasian yang mampu menangani data berdimensi tinggi dengan pola data yang tidak diketahui sebelumnya, serta dapat diterapkan untuk melihat interaksi diantara variabel yang digunakan. MARS digunakan untuk mengatasi beberapa kelemahan dari metode yang selama ini digunakan serta sebagai metode alternatif ketika data yang digunakan tidak memenuhi asumsi yang dibutuhkan pada statistika parametrik. Dari model MARS yang dibangun, terdapat tiga variabel utama yang berpengaruh terhadap ketertinggalan kabupaten diantaranya adalah pengeluaran konsumsi per kapita, angka harapan hidup, dan persentase rumah tangga pengguna listrik. Akurasi dari model MARS yang terbentuk sangat tinggi, yakni mencapai 97,83 persen dan dapat dipergunakan untuk melakukan prediksi ketertinggalan kabupaten. Berdasarkan model MARS, maka di akhir periode RPJM Nasional 2010-2014 diprediksikan terjadi transisi yang signifikan dari kabupaten dengan kondisi tertinggal menjadi tidak tertinggal serta terdapat beberapa kabupaten yang diindikasikan salah klasifikasi (yang sebelumnya dinyatakan tidak tertinggal namun seharusnya terkategorikan sebagai kabupaten tertinggal). Model ini juga dapat digunakan untuk memprediksi kondisi ketertinggalan daerah otonom baru berdasarkan data empiris yang ada, karena sebelumnya pengklasifikasian DOB hanya mengikuti status ketertinggalan daerah induknya saja.

Kata kunci : *Multivariate Adaptive Regression Splines* (*MARS*), Kabupaten tertinggal, Klasifikasi

Abstract

The purposes of this research are to build underdeveloped regency model and make a prediction in 2014 based on economic categories, Human Resources (HR), infrastructures, fiscal capacity, accessibility, and regional characteristics with MARS method. MARS is a classification method which can handle high-dimensional data with unknown pattern in advance, and can be applied to see the interaction between variables. MARS is an alternative method when the data doesn't fulfil the parametric statistics assumptions. From MARS model, there are three variables that affect underdeveloped regency, they are consumption expenditure per capita, life expectancy, and percentage of household electricity users. The accuracy of MARS model is very high, 97.83 percent and can be used to make a prediction. Based on MARS model, at the end of the National Development Plan 2010-2014 is predicted a significant transitions in regency's status. This model can also be used to predict the condition of new regency based on empirical data, because in the earlier classification, the status of regency just follows the status of parent region.

Keywords : *Multivariate Adaptive Regression Splines (MARS), Underdeveloped regency, Classification*

PENDAHULUAN

Di era reformasi dan otonomi daerah saat ini, ketimpangan antar wilayah di berbagai daerah di Indonesia masih sangat tinggi. Hal tersebut tercermin dari masih tingginya disparitas antar wilayah dari segi pendidikan, perekonomian, infrastruktur, dan kualitas sumber daya manusia. Ketimpangan tersebut mengakibatkan beberapa daerah masuk ke dalam kategori kabupaten tertinggal.

Kabupaten tertinggal merupakan kabupaten yang masyarakat serta wilayahnya relatif kurang berkembang dibandingkan daerah lain dalam skala nasional berdasarkan kategori perekonomian masyarakat, Sumber Daya Manusia (SDM), infrastruktur, kemampuan keuangan daerah, aksesbilitas, dan karakteristik daerah (berdasarkan RPJM 2010-2014 yang ditetapkan dengan Perpres No. 5 Tahun 2010).

Ketertinggalan suatu wilayah dapat terjadi akibat kondisi geografis yang menyebabkan daerah tersebut terisolir dan terpencil seperti daerah perbatasan negara, daerah pulau-pulau kecil, daerah pedalaman, serta daerah rawan bencana. Konflik sosial dan politik pun tidak luput menjadi salah satu penyebab ketertinggalan suatu wilayah. Untuk mengurangi tingkat kesenjangan tersebut, setiap tahunnya negara menganggarkan Dana Alokasi Khusus untuk membantu kabupaten tertinggal dengan harapan agar pemerintah pusat dapat mengarahkan belanja daerah untuk percepatan pembangunan kabupaten tertinggal, dan sebagai implikasinya dapat meningkatkan kesejahteraan rakyat.

Kementerian Pembangunan Daerah Tertinggal (KPDT) bersama dengan Bappenas dan Kementerian Dalam Negeri melakukan evaluasi bersama pada tahun 2004 mengenai kabupaten tertinggal. Dari evaluasi tersebut ditetapkan 199 kabupaten yang tergolong kabupaten tertinggal, dimana 62 persen diantaranya berada di wilayah Indonesia Timur. Kabupaten-kabupaten tersebut menjadi target berbagai program percepatan pembangunan kabupaten tertinggal selama Rencana Pemerintah Jangka Menengah (RPJM) 2004-2009. Pada akhir periode tersebut, 50 kabupaten berhasil keluar dari daftar kabupaten tertinggal berdasarkan ukuran ketertinggalan.

Perubahan sistem pemerintahan Indonesia dari sentralisasi menjadi desentralisasi melahirkan cukup banyak provinsi dan kabupaten baru, atau yang biasa disebut Daerah Otonom Baru (DOB). Untuk pengkategorian ketertinggalan bagi DOB didasarkan pada kabupaten induknya, jika kabupaten induknya bukan merupakan kabupaten tertinggal, maka DOB tersebut otomatis tidak masuk dalam kategori kabupaten tertinggal, dan berlaku kondisi sebaliknya. Dengan semakin banyaknya daerah otonom baru, maka bertambah pula daerah-daerah yang masuk ke dalam kategori kabupaten tertinggal.

Pada RPJM Nasional 2010-2014 terdapat 183 kabupaten yang masuk dalam kategori kabupaten tertinggal dan menjadi fokus kinerja pemerintah dalam penanganan kabupaten tertinggal, yang terdiri dari 149 kabupaten lama dan 34 kabupaten baru hasil pemekaran. Berdasarkan Tabel 1 berikut dapat diketahui bahwa jika dilihat per provinsi, maka dari 33 provinsi yang ada di Indonesia, 26 provinsi diantaranya memiliki kabupaten tertinggal. Provinsi yang memiliki jumlah kabupaten tertinggal terbanyak adalah Provinsi Papua yaitu 27 kabupaten, diikuti Provinsi Nusa Tenggara Timur sebanyak 20 kabupaten, dan Provinsi Aceh sebanyak 12 kabupaten. Jika dibandingkan dengan jumlah kabupaten/kota yang ada pada provinsi bersangkutan, Provinsi Sulawesi Barat merupakan provinsi yang paling tertinggal karena seluruh kabupatennya (100%) termasuk dalam kategori kabupaten tertinggal.

Untuk melakukan evaluasi kabupaten tertinggal periode 2010-2014 Badan Pusat Statistik dilibatkan untuk melakukan penghitungan dengan menggunakan metode yang sudah dikembangkan. Jika sebelumnya hanya dilakukan perbandingan dengan rata-rata hitung pada masing-masing variabel, maka untuk evaluasi saat ini, ke-27 variabel tersebut diberikan bobot berdasarkan hasil analisis faktor dan dilakukan penghitungan *Z-score* untuk masing-masing variabel. Nilai tersebut kemudian dikelompokkan ke dalam lima kelas interval dengan kategori Potensi Maju, Agak Tertinggal, Tertinggal, Sangat Tertinggal, dan Sangat Parah. Dari kelima kategori tersebut, hanya kabupaten yang berada pada status Potensi Maju saja

yang dikeluarkan dari kategori ketertinggalan.

Berdasarkan kondisi tersebut di atas, penulis tertarik untuk melakukan kajian yang lebih mendalam tentang fenomena kabupaten tertinggal di Indonesia melalui variabel-variabel yang telah digunakan sebelumnya dan melihat ketepatan klasifikasi dengan sebuah alat statistik sehingga menghasilkan suatu keterbandingan dengan metode yang selama ini sudah digunakan. Penulis juga melihat beberapa kelemahan dari metode yang selama ini digunakan, diantaranya kementerian terkait belum pernah melakukan pemodelan dan melihat sejauh mana variabel-variabel yang digunakan tersebut memberikan kontribusi terhadap ketertinggalan suatu wilayah. Selain itu dengan mengklasifikasikan DOB berdasarkan kabupaten induknya tanpa dilakukan evaluasi ulang terhadap kabupaten tersebut akan menimbulkan misklasifikasi, dimana kondisi DOB bisa saja berbeda dengan daerah induknya seiring dengan berjalannya pembangunan di daerah tersebut. Dengan demikian diperlukan adanya penilaian ketepatan klasifikasi dari masing-masing wilayah. Selain itu, RPJM 2010-2014 sudah menginjak periode akhir, sehingga diperlukan suatu kajian analisis untuk mengevaluasi sejauh mana pencapaian yang terjadi dalam hal penanganan kabupaten tertinggal.

Penentuan kabupaten tertinggal dan tidak tertinggal pada dasarnya adalah bagaimana cara mengelompokkan atau mengklasifikasikan sejumlah observasi ke dalam kelompok tersebut dengan memperhatikan indikator yang ada. Friedman (1991) memperkenalkan metode klasifikasi yang relatif fleksibel untuk menyelidiki pola hubungan antara variabel respon dan variabel prediktor tanpa asumsi awal terhadap bentuk hubungan fungsionalnya yang dikenal dengan *Multivariate Adaptive Regression Splines* (MARS). Metode ini merupakan kombinasi yang kompleks dari spline dan *recursive partitioning* serta melibatkan dimensi data yang besar yakni dengan jumlah observasi dan jumlah variabel yang cukup banyak. Selain itu MARS dapat secara efektif mengeksplorasi hubungan non linier yang tersembunyi diantara variabel respon dan variabel prediktor serta efek interaksi pada struktur data yang kompleks (Li-Yen Chang, 2014).

Tujuan yang ingin dicapai dari penelitian ini diantaranya membentuk model kabupaten tertinggal berdasarkan kriteria perekonomian masyarakat, SDM, infrastruktur, kemampuan keuangan daerah, aksesbilitas, dan karakteristik daerah dengan metode MARS dimana pemodelan dilakukan secara simultan (*multivariate*) dengan kondisi data yang tidak diketahui polanya; menentukan ketepatan klasifikasi kabupaten tertinggal dengan menggunakan metode MARS, membuat prediksi ketertinggalan kabupaten pada akhir periode RPJM Nasional 2010-2014 berdasarkan model MARS yang terbentuk.

METODOLOGI

Tinjauan Referensi

Proses penghitungan dan penentuan daerah tertinggal mengalami perubahan dari waktu ke waktu guna penyempurnaan. Penentuan kabupaten tertinggal pada tahun 2004 menggunakan metode rata-rata hitung, dimana dari seluruh kabupaten yang ada pada saat itu diperoleh rata-rata hitung untuk 27 variabel dari 6 kriteria yang digunakan, yaitu perekonomian masyarakat, Sumber Daya Manusia (SDM), infrastruktur, kemampuan keuangan daerah, aksesbilitas, dan karakteristik daerah. Kabupaten-kabupaten yang memiliki nilai variabel di bawah rata-rata hitung akan dikategorikan sebagai daerah tertinggal. Berdasarkan kajian statistik, metode ini memiliki kelemahan, yaitu jika hanya digunakan rata-rata hitung, maka penentuan kabupaten tertinggal akan menjadi bias akibat adanya outlier. Selain itu, metode ini hanya bersifat multiindikator, yakni melibatkan banyak sekali indikator namun tidak melakukan penghitungan secara simultan

(*multivariate*), dan tidak melihat efek interaksi diantara variabel-variabel yang digunakan.

Selanjutnya berdasarkan panduan Penjelasan Penetapan Daerah Tertinggal dijelaskan bahwa teknis penghitungan daerah tertinggal menggunakan data hasil standardisasi karena masing-masing data memiliki variasi satuan. Dari hasil standardisasi tersebut selanjutnya dikalikan dengan bobot untuk masing-masing variabel dan dilakukan penjumlahan. Namun, sebelum dilakukan penjumlahan, hasil perkalian tersebut harus dengan arah yang sama. Indikator yang bersifat mengukur tingkat keburukan seperti jumlah penduduk miskin maka arahnya positif, dan sebaliknya. Hasil total indeks inilah yang dijadikan patokan penetapan kabupaten tertinggal, dimana kabupaten-kabupaten yang memiliki total indeks di atas 0 merupakan kabupaten tertinggal. Secara statistika kondisi penghitungan tersebut baik, karena melakukan proses standardisasi yang disebabkan satuan yang berbeda dari masing-masing variabel. Namun akibat proses standardisasi ini, nilai yang semula positif dan bisa dikalikan ± 1 guna membedakan mana variabel yang mengukur tingkat keburukan dan yang tidak, justru akan diperoleh hasil yang kurang representatif, karena standardisasi akan menghasilkan nilai yang tidak hanya positif, melainkan juga negatif. Sehingga jika kemudian dikalikan dengan ± 1 akan memberikan peluang untuk menghasilkan kesimpulan yang salah.

Selain itu asumsi bagi Daerah Otonom Baru (DOB) yang dimekarkan dari daerah induk dengan status daerah tertinggal dan kemudian langsung ditetapkan sebagai daerah tertinggal membutuhkan kajian yang lebih mendalam, karena bisa saja terjadi bahwa DOB tersebut justru merupakan daerah tidak tertinggal. Atau sebaliknya, DOB yang dimekarkan dari non daerah tertinggal justru kondisinya lebih tertinggal yang tentunya menjadi lebih berhak untuk mendapatkan perhatian.

Untuk melakukan pemodelan dengan respon biner (dalam hal ini kabupaten tertinggal dan kabupaten tidak tertinggal, biasanya dilakukan dengan analisis regresi. Analisis regresi digunakan untuk memperlihatkan hubungan dan pengaruh variabel prediktor terhadap variabel respon dengan terlebih dahulu melihat pola hubungan dari variabel tersebut. Hal ini dapat dilakukan dengan dua pendekatan. Pendekatan yang paling umum dan seringkali digunakan adalah pendekatan parametrik, yang mengasumsikan bentuk model sudah ditentukan sebelumnya. Namun apabila tidak ada informasi apapun tentang bentuk dari fungsi regresi, maka pendekatan yang digunakan adalah pendekatan nonparametrik. Karena pendekatan ini tidak tergantung pada asumsi bentuk kurva tertentu, maka akan memberikan fleksibilitas yang lebih besar (Budiantara, dkk., 2006).

Metode klasifikasi merupakan bagian dari analisis statistika. Metode yang paling sering digunakan untuk masalah klasifikasi adalah analisis diskriminan. Penggunaan analisis ini membutuhkan sejumlah asumsi diantaranya populasi berdistribusi normal dengan varians-kovarians sama. Tetapi pada penerapannya analisis diskriminan sering melibatkan variabel-variabel kategorik yang tidak mengikuti pola distribusi normal sehingga akibatnya diperoleh hasil yang tidak optimal. Metode lain yang juga sering digunakan untuk masalah klasifikasi adalah regresi logistik. Analisis regresi logistik digunakan untuk analisis data respon kategorik dengan variabel-variabel bebas dapat berupa kategorik maupun kontinu (Otok, 2003). Namun analisis ini mensyaratkan adanya asumsi tidak terjadinya multikolinieritas pada variabel prediktornya (Nash dan Bradford, 2001). Padahal dalam penelitian di bidang sosial, masalah multikolinieritas seringkali tidak bisa dihindari. Selain itu, jika asumsi independensi tidak terpenuhi maka akan memberikan pendugaan yang tidak tepat.

Metode Analisis

Multivariate Adaptive Regression Splines (MARS)

MARS adalah salah satu pendekatan regresi non parametrik serta merupakan metode yang relatif baru dan dikembangkan oleh Jerome H. Friedman pada tahun 1991 untuk mengatasi kelemahan *recursive partitioning*. MARS difokuskan untuk mengatasi permasalahan data berdimensi tinggi dengan jumlah variabel dan observasi yang cukup banyak dan menghasilkan model yang kontinu pada knots. Prinsip dasar MARS adalah memberikan fleksibilitas tinggi untuk mengeksplorasi hubungan non linier yang terjadi diantara variabel respon dan variabel prediktor melalui fungsi yang berbeda untuk setiap interval yang berbeda. Selain itu, melalui metode ini juga dapat diketahui interaksi yang terjadi diantara variabel prediktor.

Pemilihan model pada MARS dilakukan dengan menggunakan metode *stepwise* yang terdiri dari *forward* dan *backward*. *Forward stepwise* dilakukan untuk mendapatkan jumlah fungsi basis maksimum dengan kriteria pemilihan fungsi basis adalah dengan meminimumkan *Average Square Residual* (ASR). Sedangkan untuk memenuhi konsep parsemoni (model sederhana) dilakukan *backward stepwise* yaitu memilih fungsi basis yang dihasilkan dari *forward stepwise* dengan meminimumkan *Generalized Cross Validation* (GCV) (Friedman, 1991, Budiantara, dkk, 2006). Secara umum model MARS dapat ditulis sebagai berikut :

$$\hat{f}(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right] \quad (1)$$

dimana :

α_0 = koefisien konstan fungsi basis

α_m = koefisien dari fungsi basis ke-m

M = jumlah maksimum fungsi basis (*non constant* fungsi basis)

K_m = derajat interaksi

s_{km} = nilainya ± 1

$x_{v(k,m)}$ = variabel prediktor

t_{km} = nilai knots dari variabel prediktor $x_{v(k,m)}$

Menurut Friedman (1991), klasifikasi pada model MARS dapat didasarkan pada pendekatan analisis regresi. Regresi logistik linier sering digunakan ketika variabel respon diasumsikan memiliki dua nilai atau yang biasa disebut *binary response*.

Berdasarkan B. W. Otok (2009) jika $f(x) = y$ dan $y \sim Ber(1, \pi(x))$ dengan $y \in (0,1)$ dan $x \in \Re^p$ maka :

$P(Y_i = 1) = \pi(x)$ dan $P(Y_i = 0) = 1 - \pi(x)$

$x \in \Re^p$ adalah vektor dari p variabel prediktor dan $P(Y = 1|x) = \pi(x)$

Lemma 1 : Jika hubungan dengan model logistik, $\sigma_L : R \rightarrow (0,1)$, dimana

$$\sigma_L = \pi(x) = \left[\frac{e^z}{1 + e^z} \right]$$

maka invers dari σ_L dapat dikatakan sebagai transformasi logit, yakni :

$$logit\ \pi(x) = ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = z$$

$$z = \hat{f}(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right]$$

maka dapat ditulis ke dalam model :

$$logit\ \pi(x) = ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \cdot \left(x_{v(k,m)} - t_{km} \right) \right] \quad (2)$$

Dan dalam bentuk matriks dapat ditulis sebagai berikut :

$$logit\ \pi(x) = \boldsymbol{B\alpha} \quad (3)$$

Karena variabel respon memiliki 2 kategori (biner), maka digunakan titik potong (*cut off*) sebesar 0,5 dengan ketentuan apabila $\pi(x) \geq 0{,}5$ maka hasil prediksi adalah 1. Dan jika $\pi(x) < 0{,}5$ maka hasil prediksi adalah 0.

Data dan Variabel yang Digunakan

Data yang digunakan dalam penelitian ini diperoleh dari Kementerian Pembangunan Daerah Tertinggal yang bersumber dari Pendataan PODES 2008, PODES 2011, SUSENAS 2009, SUSENAS 2012-2013 (yang bersumber dari Badan Pusat Statistik), serta realisasi Kemampuan Keuangan Daerah (KKD) 2009 dan realisasi KKD 2012 (yang bersumber dari Kementerian Keuangan). Sementara variabel yang digunakan pada penelitian ini dapat dilihat pada Tabel 2 berikut :

Tabel 2. Variabel yang Digunakan dalam Pemodelan Kabupaten Tertinggal

Variabel	Keterangan
Y	Kabupaten tertinggal (1) Kabupaten tidak tertinggal (0)
X1	Persentase penduduk miskin
X2	Pengeluaran konsumsi per kapita
X3	Angka harapan hidup
X4	Rata-rata lama sekolah
X5	Angka melek huruf
X6	Jumlah desa dengan jenis permukaan jalan terluas aspal/beton
X7	Jumlah desa dengan jenis permukaan jalan terluas diperkeras
X8	Jumlah desa dengan jenis permukaan jalan terluas tanah
X9	Jumlah desa dengan jenis permukaan jalan terluas lainnya
X10	Persentase rumah tangga pengguna listrik
X11	Persentase rumah tangga pengguna telepon
X12	Persentase rumah tangga pengguna air bersih
X13	Jumlah desa yang memiliki pasar tanpa bangunan permanen
X14	Jumlah prasarana kesehatan per 1000 penduduk
X15	Jumlah dokter per 1000 penduduk
X16	Jumlah SD dan SMP per 1000 penduduk
X17	Kemampuan keuangan daerah
X18	Rata-rata jarak dari kantor desa/kelurahan ke kantor kabupaten yang membawahi
X19	Jumlah desa dengan akses ke pelayanan kesehatan >5 km
X20	Jarak desa ke pelayanan pendidikan dasar
X21	Persentase desa gempa bumi
X22	Persentase desa tanah longsor
X23	Persentase desa banjir
X24	Persentase desa bencana lainnya
X25	Persentase desa di kawasan hutan lindung
X26	Persentase desa berlahan kritis
X27	Persentase desa konflik satu tahun terakhir

HASIL DAN PEMBAHASAN

Pembentukan Model Kabupaten Tertinggal

Pembentukan model kabupaten tertinggal dengan metode MARS dilakukan dengan cara *trial* and *error* terhadap kombinasi antara jumlah maksimum *basis function* (BF) yang nilainya adalah 2 sampai dengan 4 kali jumlah variabel prediktor, dalam kasus ini BF yang digunakan adalah 54, 81, dan 108; jumlah maksimum interaksi (MI) yaitu 1, 2, dan 3, dengan asumsi model yang melibatkan lebih dari 3 interaksi akan menghasilkan model yang terlalu kompleks; serta nilai minimum observasi (MO) yaitu sebesar 12, yang didasarkan atas teori dari Friedman (1991) bahwa jumlah minimum observasi terbaik dihasilkan dari rumus berikut ini:

$$L(\alpha) = 3 - log_2(^{\alpha}/_{n})$$
$$= 3 - log_2\left(^{0,05}/_{27}\right)$$
$$= 12{,}077 \approx 12$$

dimana α = tingkat kesalahan, dan n = jumlah variabel prediktor. Penentuan model yang terbaik didasarkan pada nilai GCV minimum.

Setelah dilakukan proses pengolahan dengan kombinasi BF, MI dan MO, maka diperoleh hasil kombinasi seperti terlihat pada Tabel 3 (Lampiran 2).

Berdasarkan kombinasi tersebut dan kriteria dalam pemilihan model, maka model terbaik yang dihasilkan adalah model 9 (sembilan) yang merupakan

kombinasi antara BF = 108, MI = 3, dan MO = 12. Model tersebut merupakan model dengan nilai GCV minimum yaitu 0,0662. Dari kombinasi tersebut dihasilkan 20 variabel yang berpengaruh terhadap model dengan jumlah fungsi basis sebanyak 38.

Dari tabel tersebut juga dapat disimpulkan bahwa dari sisi jumlah interaksi antar variabel, model-model dengan maksimum interaksi sebanyak 3 memiliki nilai GCV yang paling minimum dibandingkan dengan model untuk 2 interaksi maupun tanpa interaksi. Kondisi tersebut mengindikasikan bahwa diantara variabel-variabel yang mempengaruhi ketertinggalan suatu wilayah saling berinteraksi satu dengan lainnya. Selain itu juga terihat bahwa untuk semua kombinasi, variabel X2 (pengeluaran konsumsi per kapita) adalah variabel yang paling berpengaruh terhadap kondisi ketertinggalan wilayah dibandingkan dengan variabel lainnya di dalam model.

Model Kabupaten Tertinggal

Berdasarkan hasil kombinasi BF, MI, dan MO, maka model yang dihasilkan adalah sebagai berikut :

$$\begin{aligned}\hat{Y} = {} & 1{,}168 - 0{,}0278 * BF1 \quad - 0{,}2188 \\ & * BF2 + 0{,}1162 * BF3 \\ & - 0{,}01775 * BF4 \\ & + 0{,}004347 * BF5 \\ & + 4{,}49x10^{-5} * BF6 \\ & + 0{,}000232 * BF7 \\ & - 0{,}00677 * BF8 \\ & + 0{,}01119 * BF9 \\ & - 0{,}00507 * BF10 \\ & - 0{,}02704 * BF11 \\ & + 0{,}000929 * BF12 \\ & - 9{,}1x10^{-5} * BF13 \\ & + 0{,}004137 * BF14 \\ & + 0{,}007318 * BF15 \\ & - 0{,}00272 * BF16 \\ & + 0{,}000212 * BF17 \\ & + 1{,}68x10^{-5} * BF18 \\ & + 4{,}39x10^{-5} * BF19 \\ & + 0{,}004126 * BF20 \\ & - 0{,}00068 * BF21 \\ & + 0{,}000224 * BF22 \\ & + 0{,}006894 * BF23 \\ & + 0{,}197 * BF24 \\ & - 2{,}4x10^{-5} * BF25 \\ & - 0{,}00506 * BF26 \\ & - 0{,}00248 * BF27 \\ & - 0{,}01275 * BF28 \\ & - 0{,}01113 * BF29 \\ & + 0{,}000859 * BF30 \\ & + 0{,}001872 * BF31 \\ & + 0{,}000261 * BF32 \\ & - 0{,}1033 * BF33 \\ & + 0{,}000207 * BF34 \\ & - 0{,}00614 * BF35 \\ & + 0{,}007584 * BF36 \\ & + 0{,}000239 * BF37\end{aligned}$$

dimana :

BF1 = h(X2-600.79)

BF2 = h(X3-69.62)

BF3 = h(69.62-X3)

BF4 = h(X10-73.58)

BF5 = h(X2-600.79)*h(X3-69.38)

BF6 = h(X2-600.79)*h(69.38-X3)*h(X18-25.57)

BF7 = h(X2-600.79)*h(69.38-X3)*h(25.57-X18)

BF8 = h(600.79-

BF20 = h(X1-24.96)*h(X10-73.58)

BF21 = h(24.96-X1)*h(X10-73.58)

BF22 = h(24.96-X1)*h(X3-67.82)*h(X10-73.58)

BF23 = h(X2-600.79)*h(X12-53.03)*h(X16-1.74)

BF24 = h(69.62-X3)*h(X5-98.88)*h(X26-38.87)

BF25 = h(X2-600.79)*h(69.38-X3)*h(X6-95)

BF26 = h(610.39-X2)*h(69.62-X3)

BF27 = h(X2-

X2)*h(X11-1.9)

BF9 = h(600.79-X2)*h(1.9-X11)

BF10 = h(40.68-X21)
BF11 = h(600.79-X2)*h(X10-90.98)*h(X11-1.9)

BF12 = h(X2-611.66)*h(X10-73.58)
BF13 = h(X2-600.79)*h(69.38-X3)*h(30.22-X22)
BF14 = h(611.66-X2)*h(X10-73.58)*h(X12-85.76)

BF15 = h(X2-600.79)*h(3.36-X11)
BF16 = h(X2-600.79)*h(3.36-X11)*h(3-X20)
BF17 = h(X2-600.79)*h(X3-69.38)*h(X8-58)
BF18 = h(X2-600.79)*h(69.38-X3)*h(X13-100)
BF19 = h(X2-600.79)*h(69.38-X3)*h(100-X13)

610.39)*h(69.62-X3)*h(X5-96.31)
BF28 = h(610.39-X2)*h(69.62-X3)*h(X26-45.7)
BF29 = h(69.62-X3)*h(4.69-X11)
BF30 = h(611.66-X2)*h(X10-73.58)*h(X20-1.76)
BF31 = h(611.66-X2)*h(X10-73.58)*h(1.76-X20)
BF32 = h(X1-14.73)*h(X2-600.79)*h(X24-48.61)

BF33 = h(1-X9)
BF34 = h(69.62-X3)*h(X11-4.69)*h(147-X13)
BF35 = h(X1-14.73)*h(X2-600.79)*h(X4-8.38)
BF36 = h(X6-111)*h(X8-106)*h(1-X9)

BF37 = h(X6-111)*h(1-X9)*h(X25-38.13)

Interpretasi pada model MARS tidak hanya melibatkan 1 variabel saja, melainkan juga terdapat interaksi antar variabel. Karena model yang digunakan adalah MARS untuk respon biner, maka interpretasi model didekati dengan model logistik, yakni menggunakan *odds ratio*.

Variabel-variabel yang Berpengaruh Signifikan Terhadap Model Kabupaten Tertinggal

Dari pemodelan dengan metode MARS, dapat diketahui variabel-variabel mana saja yang berpengaruh secara signifikan terhadap model MARS yang dibangun. Selain itu, variabel tersebut dapat diurutkan berdasarkan tingkat kepentingan variabel tersebut di dalam model (*variable importance*). Kriteria yang digunakan untuk mengestimasi tingkat kepentingan variabel pada model MARS adalah *nsubsets*, GCV, dan RSS (*Residual Sum of Squares*). Kriteria *nsubsets* digunakan dengan cara menghitung jumlah *subsets* model yang memasukkan suatu variabel di dalamnya. Variabel yang dilibatkan dalam *subset* yang lebih banyak dianggap sebagai variabel yang lebih penting. Pada kriteria RSS, yang dilakukan adalah dengan menghitung penurunan RSS untuk setiap *subsets*. Setiap penambahan variabel akan dihitung penurunan RSS-nya. Variabel yang menyebabkan penurunan RSS yang lebih besar dianggap sebagai variabel yang lebih penting. Hal yang serupa juga dilakukan jika menggunakan kriteria GCV. Penambahan variabel yang mengakibatkan peningkatan nilai GCV, dianggap memiliki pengaruh yang tidak baik pada model. Selanjutnya untuk kemudahan interpretasi, penurunan nilai RSS atau GCV dibuat skala sehingga penurunan terbesar memiliki skala 100.

Dari hasil pemodelan, dapat diketahui bahwa dari 27 variabel yang digunakan dalam penentuan kabupaten tertinggal, hanya 20 variabel yang signifikan dalam pembentukan model. Variabel-variabel tersebut telah diurutkan berdasarkan jumlah kemunculannya pada *subsets* dan pengaruhnya terhadap penurunan GCV dan RSS. Variabel X2 (pengeluaran konsumsi per kapita) adalah variabel yang paling berpengaruh terhadap status ketertinggalan suatu kabupaten. Hal tersebut dapat dijelaskan bahwa dengan mengeluarkan variabel X2 dari model maka akan sangat berpengaruh terhadap model secara keseluruhan karena secara otomatis akan mengakibatkan peningkatan nilai RSS dan GCV, sehingga model yang dihasilkan menjadi kurang baik. Selanjutnya untuk melihat 10 variabel utama yang berkontribusi dan seberapa besar peranannya terhadap pembentukan model, dapat dilihat pada Tabel 4 berikut ini.

Tabel 4. Sepuluh Variabel Utama yang Berpengaruh terhadap Pembentukan Kabupaten Tertinggal di Indonesia

Variabel	Jumlah Subset	GCV	RSS
(1)	*(2)*	*(3)*	*(4)*
X2 (Pengeluaran konsumsi per kapita)	37	100,0	100
X3 (Angka harapan hidup)	36	68,1	72,7
X10 (Persentase rumah tangga pengguna listrik)	35	54,8	61,8
X18 (Rata-rata jarak dari kantor desa/kelurahan ke kantor kabupaten yang membawahi)	32	45,6	53,3
X13 (Jumlah desa yang memiliki pasar tanpa bangunan permanen)	31	45,9	52,9
X11 (Persentase rumah tangga pengguna telepon)	30	44,0	51,0
X5 (Angka melek huruf)	30	43,5	50,9
X26 (Persentase desa berlahan kritis)	29	42,4	49,5
X1 (Persentase penduduk miskin)	27	38,7	45,7
X24 (Persentase desa bencana lainnya)	27	38,7	45,7

Evaluasi Model Kabupaten Tertinggal Tahun 2009

Berdasarkan hasil pemodelan kabupaten tertinggal tahun 2009, maka diperoleh tabel klasifikasi sebagai berikut :

Tabel 5. Klasifikasi Kabupaten Tertinggal di Indonesia Tahun 2009

Kelompok Aktual	Kelompok Prediksi		Total
	0	**1**	
(1)	*(2)*	*(3)*	*(4)*
0	202	5	**207**
1	3	159	**162**
Total	**205**	**164**	**369**

Dengan menggunakan metode MARS kombinasi BF = 108, MI = 3, dan MO = 12, maka ketepatan klasifikasi secara keseluruhan yang dihasilkan mencapai 97,83 persen. *Sensitivity* atau dengan kata lain kabupaten yang tepat diklasifikasikan sebagai kabupaten tidak tertinggal mencapai 0,976. Sementara *Specificity* atau kabupaten yang tepat diklasifikasikan sebagai kabupaten tertinggal mencapai 0,981. Nilai Press's Q yang dihasilkan untuk menilai ketepatan dalam pengelompokkan adalah sebesar 337,69. Nilai ini lebih besar dari nilai kritis yakni sebesar 3,841. Dengan demikian dapat dikatakan bahwa klasifikasi ini konsisten secara statistik. Selain itu model juga dapat dikatakan baik karena selain memiliki GCV minimum juga memiliki tingkat akurasi yang tinggi dan tingkat kesalahan yang sangat kecil.

Jika dikaji lebih jauh, maka akan dapat dilihat perbedaan antara kondisi yang sebenarnya dengan hasil prediksi yang digambarkan oleh Gambar 1 di atas. Berdasarkan gambar tersebut dapat disimpulkan bahwa ketepatan klasifikasi pada model MARS cukup baik karena dari 33 provinsi yang ada di Indonesia, hanya 4 provinsi yang memiliki perbedaan antara kondisi aktual dengan prediksinya, yaitu Provinsi Sumatera Utara, Provinsi Jambi, Provinsi Sumatera Selatan dan Provinsi Gorontalo. Rasio misklasifikasi terhadap total kabupaten hanya sebesar 0,02. Berdasarkan indikator-indikator tersebut, maka selanjutnya model dapat digunakan untuk memprediksi ketertinggalan suatu kabupaten di akhir periode RPJM Nasional 2010-2014.

Estimasi Pengklasifikasian Kabupaten Tertinggal pada Akhir Periode RPJM Nasional 2010-2014

Dengan menggunakan pemodelan dari data untuk penentuan kabupaten tertinggal tahun 2009, selanjutnya akan dilakukan evaluasi terhadap pencapaian yang terjadi dalam hal penanganan kabupaten tertinggal di Indonesia di akhir

periode RPJM Nasional 2010-2014. Tujuan yang ingin dicapai adalah memprediksi kabupaten-kabupaten mana yang masih berstatus kabupaten tertinggal dan kabupaten tidak tertinggal. Selain itu, melakukan prediksi kabupaten mana saja yang berpotensi lepas dari ketertinggalan dan kabupaten yang mengalami misklasifikasi, yakni kabupaten yang seharusnya berstatus kabupaten tertinggal namun diklasifikasikan sebagai kabupaten tidak tertinggal.

Tabel 6. Transisi Kabupaten Tertinggal di Indonesia pada akhir RPJM Nasional 2010- 2014

Kelompok Aktual	**Kelompok Prediksi**		**Total**
	0	**1**	
(1)	*(2)*	*(3)*	*(4)*
0	195	20	**215**
1	68	115	**183**
Total	**263**	**135**	**398**

Dari tabel tersebut dapat disimpulkan bahwa setelah kurun waktu 5 tahun proses pembangunan yang terfokus pada kabupaten tertinggal, diprediksikan bahwa di akhir periode (tahun 2014) terdapat 68 kabupaten yang berpotensi lepas dari ketertinggalan. Sehingga saat ini terdapat 263 kabupaten dengan status kabupaten tidak tertinggal. Dari 183 kabupaten tertinggal pada periode sebelumnya, 62,84 persen diantaranya tetap berada pada status yang sama. Selain itu, dari hasil pemodelan dengan menggunakan metode MARS, diperkirakan terdapat 20 kabupaten yang salah klasifikasi (misklasifikasi), salah satu penyebabnya adalah penentuan status ketertinggalan DOB yang hanya didasarkan oleh status ketertinggalan kabupaten induknya. Dengan demikian sekitar 33,92 persen kabupaten menjadi target pembangunan kabupaten tertinggal untuk periode selanjutnya.

KESIMPULAN DAN SARAN

Kesimpulan

Dari hasil penelitian tentang kabupaten tertinggal di Indonesia dengan model MARS maka dapat disimpulkan beberapa hal sebagai berikut :

1. Metode MARS merupakan pendekatan yang sesuai untuk penentuan kabupaten tertinggal jika dilihat dari kondisi data dan keterbatasan beberapa metode statistika yang ada dan metode yang digunakan oleh KPDT selama ini. Dari hasil pemodelan diperoleh tingkat akurasi yang sangat tinggi, yakni mencapai 97,83 persen dan dapat dipergunakan untuk melakukan prediksi ketertinggalan kabupaten.
2. Dari model MARS yang dibangun dapat diketahui bahwa terdapat keterkaitan/interaksi antara variabel prediktor yang digunakan dalam penentuan kabupaten tertinggal.
3. Lima prediktor utama yang berpengaruh terhadap ketertinggalan kabupaten diantaranya adalah pengeluaran konsumsi per kapita, angka harapan hidup, persentase rumah tangga pengguna listrik, rata-rata jarak dari kantor desa/kelurahan ke kantor kabupaten yang membawahi, serta jumlah desa yang memiliki pasar tanpa bangunan permanen.
4. Berdasarkan model MARS, maka di akhir periode RPJM Nasional 2010-2014 diprediksikan terjadi transisi yang signifikan dari kabupaten dengan kondisi tertinggal menjadi tidak tertinggal serta terdapat beberapa kabupaten yang diindikasikan salah klasifikasi (yang sebelumnya dinyatakan tidak tertinggal namun seharusnya terkategorikan sebagai kabupaten tertinggal).

Saran

Untuk penelitian dan pengembangan lebih lanjut, maka berdasarkan hasil penelitian ini dapat disarankan beberapa hal, diantaranya :

1. Pengklasifikasian yang dilakukan pada saat ini adalah dengan respon biner, sehingga untuk penyempurnaan selanjutnya diharapkan dapat dilakukan pengklasifikasian dengan

lebih dari dua kategori agar dapat dibedakan kabupaten mana saja yang berada pada kondisi maju, berkembang, tertinggal, dan sangat tertinggal.

2. Penelitian yang dilakukan saat ini adalah pemodelan kabupaten tertinggal dengan skala nasional agar hasilnya dapat diperbandingkan dengan metode sebelumnya. Namun demikian diperlukan pemodelan MARS dan prediksi ketertinggalan kabupaten untuk masing-masing wilayah Indonesia Barat, Indonesia Tengah, dan Indonesia Timur yang diperkirakan memiliki karakteristik yang berbeda.
3. Model MARS yang digunakan berbasis pada spline, untuk pengembangan metode selanjutnya dapat dilakukan modifikasi dengan fungsi basis yang setara dengan spline seperti *wavelet* untuk kemudian dapat dilihat perbedaan tingkat akurasi yang diperoleh.

DAFTAR PUSTAKA

Agresti, A. 2002. *Categorical Data Analysis*. New Jersey : John Wiley & Sons, Inc

Budiantara, I.N., Suryadi, F., Otok, B.W., Guritno, S. 2006. *Pemodelan B-Spline dan MARS pada Nilai Ujian Masuk terhadap IPK Mahasiswa Jurusan Disain Komunikasi Visual UK. Petra Surabaya*. Jurnal Teknik Industri Vol 8, No 1, hal 1-13.

Chang, Li-Yen. 2014. *Analysis of Bilateral Air Passenger Flows: A Non-Parametric Multivariate Adaptive Regression Spline Approach.*Journal of Air Transport Management 34 : 123-130

Direktoral Jenderal Perimbangan Keuangan Kementerian Keuangan. 2013. *Affirmative Policy Dalam Percepatan Pembangunan Daerah Untuk Peningkatan Kesejahteraan Rakyat*. Jakarta : Kementerian Keuangan.

Fernandez, J. R. A., Nieto, P. J. G., Muniz, C. D., Anton, J. C. A. 2014. *Modelling Eutrophication and Risk Prevention in a Reservoir in the Northwest of Spain by Using Multivariate Adaptive Regression Splines Analysis.* Ecological Engineering 68 : 80-89

Friedman, J. H. 1991. *Multivariate Adaptive Regression Splines*. The Annals of Statistics, Vol. 19, No. 1, hal. 1-141

Hair, J.F, Rolph E. Anderson, Ronald L. Tatham, William C. Black. 2006. *Multivariate Data Analysis*. *Sixth Edition,* Pearson Education Prentice Hall, Inc.

Kementerian Pembangunan Daerah Tertinggal. 2010. *Rencana Strategis Tahun 2010-2014.* Jakarta : KPDT

Nash, M. S. dan David F.B. 2001. *Parametric and Non Parametric Logistic Regression for Prediction of Precense/ Absence of an Amphibian.* Las Vegas, Nevada : US Environmental Protection Agency Office of Research and Development National Exposure Research Laboratory Environmental Sciences Division

Otok, B. W. 2003. *Perbandingan MARS dengan Regresi Logistik pada Respon Biner*. Prosiding Seminar Nasional Matematika dan Statistika VI. ITS, Surabaya.

Otok, B. W., Akbar, M. S., Guritno, S., Subanar. 2007. *Pendekatan Bootstrap pada Klasifikasi Pemodelan Respon Ordinal*. Jurnal Ilmu Dasar, Vol. 8 No. I, hal. 54-67.

Otok, B. W. 2009. *Konsistensi dan Asimtotik Normalitas Model Multivariate Adaptive Regression Splines (MARS) pada Respon Biner*. Jurnal Ilmu Dasar, Vol. 10 No. 2, hal. 133-140.

Quiros, E., Felicimo, A. M., Cuartero, A. 2009. *Testing Multivariate Adaptive Regression Splines (MARS) as a Method of Land Cover Classification of TERRA-ASTER Satellite Images.* Sensors 2009, 9.

LAMPIRAN

Lampiran 1

Tabel 1. Daerah Tertinggal per Provinsi di Indonesia Periode 2010-2014

No	Provinsi	Jumlah Kabupaten/Kota	Jumlah Daerah Tertinggal	% Daerah Tertinggal
(1)	*(2)*	*(3)*	*(4)*	*(5)*
1	Aceh	23	12	52,17
2	Sumatera Utara	33	6	18,18
3	Sumatera Barat	19	8	42,11
4	Kepulauan Riau	7	2	28,57
5	Sumatera Selatan	15	7	46,67
6	Bangka Belitung	7	1	14,29
7	Bengkulu	10	6	60,00
8	Lampung	14	4	28,57
9	Jawa Barat	26	2	7,69
10	Banten	8	2	25,00
11	Jawa Timur	38	5	13,16
12	Kalimantan Barat	14	10	71,43
13	Kalimantan Tengah	14	3	21,43
14	Kalimantan Timur	14	3	21,43
15	Sulawesi Utara	15	3	20,00
16	Gorontalo	6	3	50,00
17	Sulawesi Tengah	11	10	90,91
18	Sulawesi Selatan	24	4	16,67
19	Sulawesi Barat	5	5	100,00
20	Sulawesi Tenggara	12	9	75,00
21	NTB	10	8	80,00
22	NTT	21	20	95,24
23	Maluku	11	8	72,73
24	Maluku Utara	9	7	77,78
25	Papua	29	27	93,10
26	Papua Barat	11	8	72,73
Total		**406**	**183**	**45,07**

Sumber : Kementerian Keuangan

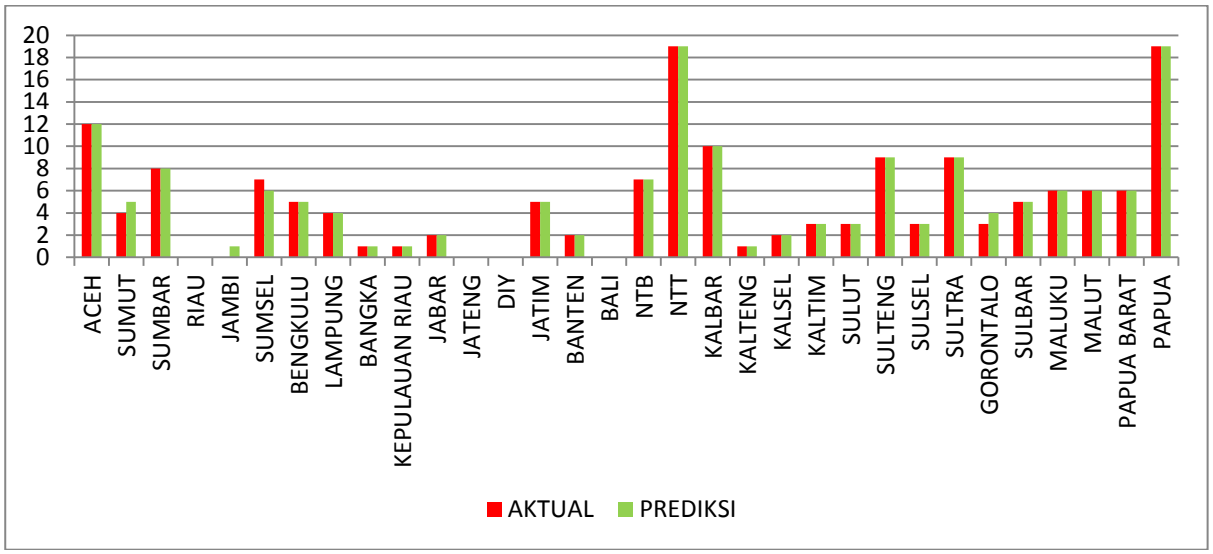
Lampiran 2

Tabel 3. Hasil Kombinasi dari BF, MI, dan MO untuk Model Kabupaten Tertinggal

Model MARS	BF	MI	MO	GCV	R^2	Jumlah Variabel Sig.	Jumlah BF Sig.	Var. X yang Berkontribusi
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*
1	54	1	12	0.0934	0.686	13	17	X2
2	54	2	12	0.0778	0.784	17	26	X2
3	54	3	12	0.0727	0.808	14	29	X2
4	81	1	12	0.0949	0.703	14	23	X2
5	81	2	12	0.0746	0.809	19	31	X2
6	81	3	12	0.0669	0.846	16	37	X2
7	108	1	12	0.0940	0.702	13	22	X2
8	108	2	12	0.0777	0.855	21	48	X2
9	**108**	**3**	**12**	**0.0662**	**0.850**	**20**	**38**	**X2**

Keterangan : BF = *Basis Function* MI = Maksimum Interaksi MO = Minimum Observasi

Lampiran 3



Gambar 1 Perbandingan Kabupaten Tertinggal Tahun 2009 antara Kondisi Aktual dan Prediksi per Provinsi di Indonesia

VISUALISASI PENGGEROMBOLAN WILAYAH BERDASARKAN TEORI PERTUMBUHAN EKONOMI MENGGUNAKAN APLIKASI INTEGRASI *SELF ORGANIZING MAP* (SOM) DAN SISTEM INFORMASI GEOGRAFIS

VISUALIZATION OF CLUSTERING REGION BY ECONOMIC GROWTH THEORY USING THE INTEGRASI SELF ORGANIZING MAP (SOM) AND GEOGRAPHIC INFORMATION SYSTEM

Hafshoh Mahmudah
Sekolah Tinggi Ilmu Statistik

Ricky Yordani
Sekolah Tinggi Ilmu Statistik

Masuk tanggal: 07-12-2015, revisi tanggal: 13-01-2016, diterima untuk diterbitkan tanggal: 19-01-2016

Abstrak

Pertumbuhan ekonomi merupakan salah satu faktor penting untuk menentukan kesejahteraan suatu wilayah. Akan tetapi, perbedaan kondisi geografis dan potensi wilayah menyebabkan perbedaan kondisi ekonomi yang berbeda antarwilayah. Studi kasus dilakukan terhadap Provinsi Jawa Tengah karena merupakan salah satu kontributor PDRB terbesar di Indonesia, yang ternyata masih memiliki ketimpangan perekonomian antar kota dan antar kabupaten. Untuk memudahkan visualisasi pertumbuhan ekonomi maka dibuatlah suatu aplikasi yang mampu melihat secara mudah efek pertumbuhan dan penggerombolan dalam wilayah Provinsi Jawa Tengah tersebut. Metode yang bisa digunakan untuk analisis gerombol sangat beragam. Salah satu metode alternatif adalah menggunakan metode *Self Organizing Map* (SOM) yang mampu menggerombolkan data multidimensi disertai dengan visualisasinya dengan teknik *Unsupervised Artificial Neural Network.* Aplikasi ini memudahkan visualisasi dan analisisnya karena diintegrasikan dengan Sistem Informasi Geografis (SIG). Aplikasi yang dibuat selanjutnya digunakan untuk melakukan analisis gerombol dengan data studi kasus Provinsi Jawa Tengah. Visualisasi yang dihasilkan mampu menunjukkan pola pertumbuhan ekonomi di Provinsi Jawa Tengah namun belum terlihat adanya pemusatan kutub pertumbuhan ekonomi di Provinsi Jawa Tengah karena pola penggerombolan berdasarkan indikator pertumbuhan ekonomi masih menyebar.

Kata kunci : Kutub Pertumbuhan Ekonomi, Self Organizing Map, Analisis Gerombol

Abstract

Economic growth is one of factor that is critical to determining the welfare of a region. However, differences in geographical conditions and the potential of the area led to differences in economic conditions differ between regions. The case studies conducted on Central Java Province because it is one of the largest contributors to GDP in Indonesia, which still has economic inequality between cities and between districts. To make more easy for visualize the economic growth, researcher then made an application that is able to easily see the effect of growth and clustering in the province of Central Java. There are many methods that can be used for cluster analysis. One of the most common methods used are the K-Means. However, K-Means has some drawbacks. One alternative method is using the Self Organizing Map (SOM) which is capable clustering accompanied by visualization of multidimensional data with techniques Unsupervised Artificial Neural Network. This application allows visualization and analysis because it is integrated with Geographic Information Systems (GIS). Applications are made subsequently used to analyze clustering with case study data of Central Java province. The resulting visualization capable of showing a pattern of economic growth in Central Java Province but has not seen the concentration of economic growth pole in Central Java because clustering pattern based on indicators of economic growth spread.

Keywords : *Economic Growth Pole, Self Organizing Map, Cluster Analysis*

PENDAHULUAN

Pertumbuhan ekonomi yang terus menunjukkan peningkatan menggambarkan bahwa perekonomian

negara atau wilayah tersebut berkembang dengan baik (Amri, 2007), hal ini dapat menjadi salah satu indikator keberhasilan pembangunan ekonomi suatu wilayah. Akan tetapi kondisi geografis dan potensi sumber daya yang berbeda-beda antar daerah menyebabkan perbedaan kondisi ekonomi pada wilayah tersebut. Hal tersebut membuat munculnya daerah yang memiliki potensi sebagai pusat pertumbuhan ekonomi dan ada yang tidak.

Daerah dengan potensi sebagai pusat pertumbuhan ekonomi merupakan salah satu alternatif untuk menggerakkan dan memacu pembangunan guna meningkatkan pendapatan masyarakat. Secara tidak langsung kemajuan daerah akan membuat masyarakat mencari kehidupan yang lebih layak di daerah tersebut. Selain itu, penciptaan pusat pertumbuhan ekonomi yang dimulai dari beberapa sektor yang dinamis dan mampu memberikan output rasio yang tinggi akan dapat memberikan dampak yang luas (*spread effect*) dan dampak ganda (*multiple effect*) pada sector lain dengan wilayah yang lebih luas. Atau dengan kata lain, wilayah yang menjadi pusat pertumbuhan ekonomi akan membuat wilayah di sekitarnya turut mengalami peningkatan pertumbuhan ekonomi.

Komite Percepatan dan Perluasan Pembangunan Ekonomi Indonesia (KP3EI) dalam strategi utamanya menjadikan pendekatan pusat-pusat pertumbuhan ekonomi sebagai dasar mencapai percepatan dan perluasan pembangunan ekonomi Indonesia. Pendekatan ini pada intinya merupakan integrasi dari pendekatan sektoral dan regional sehingga pengembangan wilayah pusat pertumbuhan ekonomi tersebut dapat memaksimalkan keuntungan aglomerasi, menggali potensi dan keunggulan daerah serta memperbaiki ketimpangan spasial pembangunan ekonomi Indonesia. Berkaitan dengan usaha pembangunan ekonomi yang berkonsentrasi pada wilayah pertumbuhan ekonomi, maka diperlukan suatu analisis tertentu yang dapat digunakan untuk mengetahui persebaran pusat pertumbuhan ekonomi pada suatu wilayah, salah satunya adalah menggunakan analisis gerombol.

Provinsi Jawa Tengah merupakan kontributor Produksi Domestik Regional Bruto (PDRB) terbesar keempat di Indonesia setelah DKI Jakarta, Jawa Timur dan Jawa Barat. Akan tetapi, ketimpangan PDRB perkapita antar kabupaten/kota di provinsi ini masih cukup besar dalam beberapa tahun terakhir. Hal itu terlihat dari besarnya kesenjangan antara kabupaten atau kota dengan PDRB perkapita tertinggi dan PDRB perkapita terendah. Jika dilihat perbandingan nilai PDRB Atas Dasar Harga Berlaku (ADHB) dengan migas ter- lihat adanya kesenjangan pendapatan yang cukup tinggi yaitu PDRB tertinggi mencapai 65.137 miliar rupiah (Kabupaten Cilacap) dan PDRB terendah sebesar 1.370 miliar rupiah (Kota Salatiga), (Badan Pusat Statistik Provinsi Jawa Tengah, 2008). Kondisi tersebut mendorong Pemerintah Provinsi Jawa Tengah membuat kebijakan dan kerjasama regional antar wilayah yang bertujuan untuk membantu percepatan pembangunan.

Metode yang sering digunakan dalam mengkaji tentang penggerombolan adalah K-Means. K- Means adalah suatu metode analisis data atau metode *data mining* yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. K-Means dapat mengolah data dalam jumlah yang sangat besar dengan lebih efektif dan tidak memerlukan waktu yang lama. Namun algoritma ini juga memiliki permasalahan dalam menentukan titik awal centroid (Wang Huai-bin, 2010).

Metode lain yang bisa digunakan adalah salah satu metode *Artificial Neural Network* (ANN) yaitu *Self Orginizing Map* (SOM). Analisis gerombol yang digunakan pada metode tersebut bersifat *unsupervised* karena

tidak ada satu atributpun yang digunakan untuk memandu proses pembelajaran dan seluruh variabel input diperlakukan sama. Metode SOM mampu mengatasi permasalahan berkaitan dengan data multidimensi seperti data yang memiliki banyak variabel yang menjadikannya sulit diintepretasi. Metode SOM memberikan kemudahan intepretasi data multidimensi dengan visualisasi serta memiliki keunggulan pada akurasi dan ketahanan *(accuracy and robustness)* (Yan Li Subana S, 2007).

SOM sering dianggap sebagai metode penggerombolan, visualisasi dan reduksi dimensi yang baik dan telah dikembangkan pada berbagai analisis data eksploratori di berbagai bidang seperti medis, segmentasi konsumen, pasar finansial dan teknik industri (Oja et al, 2002).

Meskipun demikian, aplikasi yang mengimple mentasikan metode SOM masih sedikit dan memiliki keterbatasan bagi pengguna karena pengguna harus mengetikkan baris kode sendiri pada *command line*. Oleh karena itu, dalam penelitian ini penulis ingin menggembangkan metode SOM yang akan diintegrasikan dengan Sistem Informasi Ge- ografis untuk dapat diterapkan sebagai alat penggerombolan dan visualisasi kota dan kabupaten di Provinsi Jawa Tengah berdasarkan variabel-variabel pertumbuhan ekonomi.

METODOLOGI

Teori Pertumbuhan Ekonomi (*Growth Pole Theory*)

Pertumbuhan ekonomi yang optimal akan membawa kepada kehidupan yang lebih baik. Salah satu yang menjadi teori dasar pertumbuhan ekonomi adalah teori *growth pole*. Teori kutub pertumbuhan pertama kali dikemukakan oleh Perroux pada Tahun 1955. Setelah itu, teori *growth pole* berkembang dengan pesat dan digunakan sebagai dasar pengambilan kebijakan baik pada negara berkembang maupun negara maju. Penerapan teori tersebut secara serius dimulai sejak tahun 1970 (Miyoshi, 1997).

Konsep *growth pole* didasarkan pada teori ekonomi makro. Oleh karenanya, dasar utama adalah konsentrasi pertumbuhan ekonomi pada ruang tertentu. Boundeville melengkapi penelitian dari Perroux tentang teori kutub pertumbuhan dengan menambah implikasi spasial terhadap teori tersebut. Boundeville mendefinisikan kutub pertumbuhan regional sebagai aglomerasi geografis sekelompok industri propulsif yang mengalami ekspansi yang berlokasi di suatu daerah perkotaan dan mendorong perkembangan kegiatan ekonomi lebih lanjut ke seluruh wilayah pengaruhnya. Dengan kata lain, dalam konteks pertumbuhan ekonomi, kegiatan-kegiatan industri yang akan menjadi medan magnet dan membentuk kutub pertumbuhan sehingga dapat menyebarkan pertumbuhan ekonomi melalui efek kumulatif.

Analisis Gerombol

Gerombol dapat diartikan sebagai ’kelompok’, dengan demikian pada dasarnya analisis gerombol akan menghasilkan sejumlah gerombol (kelompok). Analisis ini diawali dengan pemahaman bahwa sejumlah data tertentu sebenarnya mempunyai kemiripan di antara anggotanya; karena itu, dimungkinkan untuk mengelompokkan anggota- anggota yang mirip atau mempunyai karakteristik yang serupa tersebut dalam satu atau lebih dari satu gerombol (Santoso, 2010).

Terdapat dua kriteria yang digunakan untuk memilih skema penggerombolan yang optimal, antara lain: (Salazar et al, 2002)

1. *Compactness*, yaitu anggota dari masing-masing gerombol harus sedekat mungkin dengan yang lain.
2. *Separation*, yaitu gerombol harus terpisah secara luas dari gerombol lain.

Secara umum metode utama penggerombolan dapat diklasifikasikan menjadi:

1. Metode Hierarki
   Metode hierarki ialah metode yang memulai penggerombolannya dengan dua atau lebih obyek yang mempunyai kesamaan paling dekat, kemudian proses dilanjutkan ke objek lain yang mempunyai kedekatan kedua. Demikian seterusnya sehingga gerombol akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai dengan yang paling tidak mirip.

2. Metode Non Hierarki
   Metode non hierarki ialah metode yang dimulai dengan menentukan terlebih dahulu jumlah gerombol yang diinginkan (dua gerombol, tiga gerombol atau yang lain). Dan kemudiaan baru dilakukan proses gerombol tanpa mengikuti proses hierarki. Biasa disebut metode *K-Means Cluster*. Dua kelemahan dari prosedur non hierarki ialah bahwa banyaknya gerombol harus disebutkan atau ditentukan sebelumnya dan pemilihan pusat gerombol sembarang. Lebih lanjut, hasil gerombol mungkin tergantung pada bagaimana pusat dipilih. Banyak metode non hierarki dalam memilih gerombol tergantung pada urutan observasi dalam data, sehingga metode gerombol non hierarki lebih cepat daripada metode hierarki dan lebih menguntungkan kalau jumlah objek/kasus atau observasi besar sekali (sampel besar).

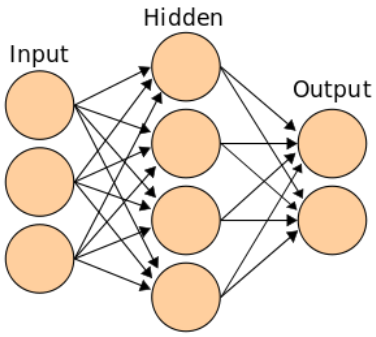
Artificial Neural Network (ANN)

Artificial Neural Network (ANN) merupakan model komputasi yang terinspirasi oleh jaringan syaraf. ANN terdiri dari gabungan sejumlah elemen yang memproses informasi dari input sehingga memberikan suatu informasi keluaran. Sekelompok objek yang dipelajari oleh sistem belajar dengan tujuan untuk mengenali bentuk pola. Proses ini dilakukan dengan melatih sistem belajar (*train neural network*) melalui pemberian bobot dan bias (pada kesalahan minimum yang dicapai) untuk semua pola yang dipelajari.

ANN mempunyai distribusi pararel arsitektur dengan sejumlah besar simpul mempunyai bobot dan bias tertentu. Kontruksi ANN terdiri dari penentuan perangkat jaringan, penentuan perangkat simpul, penentuan sistem dinamik. Selain itu, ANN terdiri dari sejumlah lapisan dan simpul yang berbeda untuk tiap *layer*. Jenis *layer*nya dibedakan menjadi:

1. *Input layer:* terdiri dari unit-unit simpul yang berperan sebagai input proses pengolahan data pada neural network.
2. *Hidden layer:* terdiri dari unit-unit simpul yang dianalogikan sebagai lapisan tersembunyi dan berperan sebagai lapisan yang meneruskan respon dari input; 3. *Output layer*: terdiri dari unit-unit simpul yang berperan memberikan solusi dari data input.

Secara umum, struktur ANN dapat dilihat dalam Gambar 1 di bawah ini.



Gambar 1: Model Struktur ANN

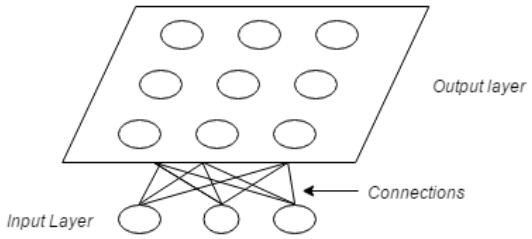
Self Organizing Map (SOM)

Self Orginizing Map (SOM) merupakan jenis dari *Artificial Neural Network* (ANN) yang dikembangkan oleh Teuvo Kohonen (Kohonen, 2001). SOM menjadi metode dengan pendekatan

ANN untuk melakukan penggerombolan (*clustering*) setelah melakukan *competitive learning* (Han dan Kamber, 2001). Jaringan SOM merupakan salah satu jaringan yang banyak dipakai, antara lain untuk membagi pola masukan ke dalam beberapa kelompok/gerombol/*cluster* (Siang, 2009).

Masukan dalam metode SOM adalah berupa vektor yang terdiri atas n komponen yang akan dikelompokkan dalam maksimum k buah kelompok (disebut vektor contoh). Keluaran jaringan adalah kelompok yang paling dekat/mirip dengan masukan yang diberikan. Ada beberapa ukuran kedekatan yang dapat dipakai. Ukuran yang sering dipakai adalah jarak Euclidean yang paling minimum (Siang, 2009).

SOM merupakan generalisasi dari jaringan kompetitif, dan merupakan jaringan tanpa supervisi (Siang, 2009). SOM disusun oleh sebuah lapisan unit input yang dihubungkan seluruhnya ke lapisan unit output, yang kemudian unit unit diatur di dalam topologi khusus seperti struktur jaringan. Secara umum arsitektur jaringan SOM dapat dilihat pada Gambar 2 berikut :



Gambar 2: *Self Organizing Map* (SOM)

K-Means

Pada analisis gerombol, metode yang paling umum digunakan adalah K-Means karena K-Means berdasarkan pada konsep yang sederhana dan menghasilkan hasil yang baik (Kumar dan Asger, 2015). Pendekatan K-Means digunakan untuk mendapatkan dua estimasi yaitu:

1. Pusat lokasi (*center*) dari masing-masing gerombol
2. Partisi dari data menurut gerombolnya.

Metode K-Means membagi x_n data (dengan x merupakan suatu variabel, dengan jumlah sebanyak n) ke dalam suatu set data sebanyak k gerombol, dengan perbedaan yang besar antar gerombol. Tujuan dari K-Means adalah untuk meminimalkan total varians dalam gerombol (*total intra-cluster variance*).

Within-Cluster Sum of Squares (WCSS)

*Within-cluster sum of square*s (WCSS) merupakan salah satu kriteria yang paling sering digunakan dalam analisis gerombol. WCSS digunakan untuk mengukur kualitas penggerombolan dengan cara mengukur varians-kovarian dalam sebuah gerombol.

Davies-Bouldin Index

Indeks Davies-Bouldin merupakan salah satu indeks validitas yang digunakan sebagai metode validasi gerombol untuk evaluasi kuantitatif dari hasil penggerombolan. (Salazar dkk, 2002).

Silhouette Index

Indeks Silhouette merupakan salah satu indeks validasi untuk analisis gerombol. Indeks Silhouette mengukur seberapa mirip suatu titik dengan titik yang lainnya dalam satu gerombol ketika dibandingkan dengan titik pada gerombol lainnya.

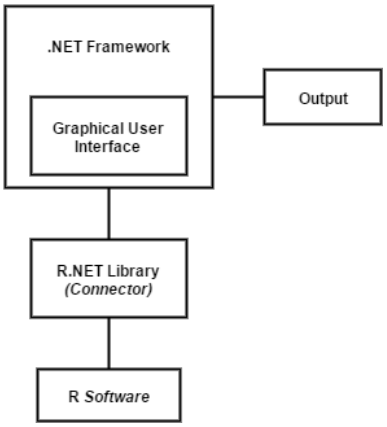
System Development Life Cycle (SDLC)

Metode pembangunan sistem menggunakan empat tahap dalam metodologi *System Development Life Cycle* (SDLC). Berdasarkan metodologi SDLC, pembangunan aplikasi meliputi beberapa fase sebagai berikut:

1. Identifikasi dan inisiasi kebutuhan aplikasi.
   Langkah awal dalam pembangunan sistem adalah menggali informasi yang dibutuhkan oleh sistem dengan cara mengidentifikasi masalah yang potensial untuk dikembangkan lebih lanjut. Identifikasi kebutuhan dapat dilakukan dengan mencari dan mempelajari referensi-referensi ilmiah penelitian yang terkait dengan masalah. Hal ini diperlukan untuk mendapatkan fakta yang mendukung pembangunan sistem aplikasi.
2. Analisis.
   Tahap analisis dilakukan dengan mengamati aplikasi yang terkait dengan domain penelitian. Tujuan utama pada tahapan ini adalah mendokumentasikan keadaan aplikasi yang ada dan menghasilkan prasyarat-prasyarat yang harus dipenuhi oleh sistem usulan berdasarkan hasil pengamatan yang dilakukan.
3. Perancangan aplikasi.
   Langkah selanjutnya adalah menindaklanjuti kebutuhan yang masih berupa konseptual menjadi spesifikasi sistem yang lebih nyata.
4. Implementasi.
   Tahap keempat akan mengimplementasikan rancangan-rancangan yang telah dibuat pada tahap sebelumnya. Tahapan ini dilakukan dengan menerjemahkan metode dan rancangan yang telah dibuat ke dalam kode program sehingga menjadi suatu aplikasi yang utuh.
5. Validasi.
   Selanjutnya, pengujian terhadap hasil implementasi rancangan sistem dilakukan guna mengevaluasi sistem yang telah dibangun

Rancangan Aplikasi

Aplikasi yang dikembangkan merupakan aplikasi berbasis dekstop yang menggunakan *software* R sebagai backend dalam pengolahan data. Untuk menghubungkan aplikasi utama dan aplikasi R dibutuhkan suatu *connector*. *Connector* yang akan digunakan dalam aplikasi ini adalah R.NET. Selanjutnya hasil pengolahan aplikasi utama akan ditampilkan pada antar muka agar dapat dilihat oleh pengguna. Secara visual, rancangan aplikasi dapat dilihat pada Gambar 3 di bawah ini.



Gambar 3. Aristektur Aplikasi

Data dan Variabel

Variabel yang digunakan dalam penelitian ini berasal dari penelitian Isnainy (2012) yang membahas tentang kinerja perekonomian di Jawa Tengah dengan sumber data dari berbagai publikasi BPS. Pada penelitian tersebut menghasilkan penggerombolan berdasarkan 14 variabel yang signifikan, yang antara lain sebagai berikut, yaitu Produk Domestik Regional Bruto (PDRB), tingkat kepadatan pen- duduk (*DENSITY*), tingkat produktivitas tenaga kerja (*PRODUCT*), presentase penduduk miskin (*POVERTY*), Jumlah Penduduk (POP), Angka Pastisipasi Sekolah (APS1) usia 7-12 tahun, Angka Partisipasi Sekolah (APS2) usia 13-15 tahun, Angka Harapan Hidup (AHH), jumlah pasar (*MARKET*), jumlah Sekolah Dasar (SD), jumlah Sekolah Menengah Pertama (SMP), jumlah Sekolah Menengah Atas (SMA), panjang jalan (*ROAD*), dan penerimaan daerah (PAD).

Pembangunan Sistem

Pembangunan sistem menggunakan empat tahap dalam metodologi *System*

Development Life Cycle (SDLC). Berdasarkan metodologi SDLC, pembangunan aplikasi meliputi beberapa fase sebagai berikut:

1. Identifikasi dan inisiasi kebutuhan aplikasi.
   Langkah awal dalam pembangunan sistem adalah menggali informasi yang dibutuhkan oleh system dengan cara mengidentifikasi masalah yang potensial untuk dikembangkan lebih lanjut. Identifikasi kebutuhan dapat dilakukan dengan mencari dan mempelajari referensi-referensi ilmiah penelitian yang terkait dengan masalah. Hal ini diperlukan untuk mendapatkan fakta yang mendukung pembangunan sistem aplikasi.
2. Analisis
   Tahap analisis dilakukan dengan mengamati aplikasi yang terkait dengan domain penelitian. Tujuan utama pada tahapan ini adalah mendokumentasikan keadaan aplikasi yang ada dan meng- hasilkan prasyarat-prasyarat yang harus dipenuhi oleh sistem usulan berdasarkan hasil pengamatan yang dilakukan. Dalam penelitian ini, dilakukan pengamatan terhadap *software* penggerombolan yang sudah ada seperti R *software* dan SOM *tool box* pada Matlab.
3. Perancangan aplikasi
   Langkah selanjutnya adalah menindaklanjuti kebutuhan yang masih berupa konseptual menjadi spesifikasi sistem yang lebih nyata. Perancangan yang dilakukan pada penelitian ini yaitu: (1) merancang spesifikasi input, output, dan proses pada aplikasi, (2) memilih teknologi baik perangkat keras dan perangkat lunak yang akan digunakan (3) merancang arsitektur aplikasi (4) merancang antar muka pengguna.
4. Implementasi
   Tahap keempat akan mengimplementasikan rancangan-rancangan yang telah dibuat pada tahap sebelumnya. Tahapan ini dilakukan dengan menerjemahkan metode dan rancangan yang telah dibuat ke dalam kode program sehingga menjadi suatu aplikasi yang utuh. Tahapan implementasi dilakukan dengan: a. Menentukan bahasa pemprogaman yang digunakan; b. Menentukan struktur data yang digunakan; c. Mengimplementasikan antar muka yang telah dirancang ke dalam kode program; d. Memberikan perintah ke pada komponen dalam antar muka dalam menghadapi kondisi-kondisi tertentu (*event*); e. Memberikan aturan validasi pada form berdasarkan input pengguna dan pada kotak dialog; f. Mengimplementasikan algoritma metode *Self Organizing Map* ke dalam kode program; 5. Validasi. Selanjutnya, pengujian terhadap hasil implementasi rancangan sistem dilakukan guna mengevaluasi sistem yang telah dibangun. Uji coba pada sistem menggunakan *blackbox test* dan *whitebox test*.

Penerapan Penggunaan Metode SOM

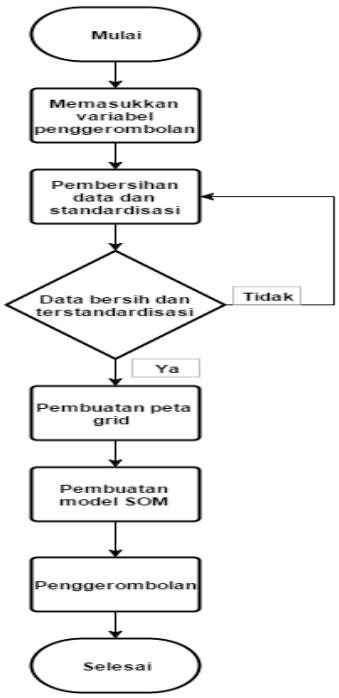
Penerapan penggunaan metode *Self Organizing Map* (SOM) untuk mengelompokkan kota dan kabupaten di Provinsi Jawa Tengah berdasarkan variabel potensi pertumbuhan ekonomi dengan mengintegrasikanya melalui Sistem Informasi Geografis sehingga didapat peta tematik yang menampilkan hasil penggerombolan.

Berikut ini adalah tahapan untuk mengkaji penggunaan metode *Self Organizing Map* (SOM) untuk mengelompokkan kota/kabupaten di Provinsi Jawa Tengah :

a. Menentukan variabel yang akan digunakan dalam penelitian
   Variabel yang digunakan dalam penelitian ini berasal dari penelitian Isnainy (2012) yang membahas tentang kinerja perekonomian di Jawa Tengah. Data-data tersebut bersumber dari berbagai publikasi BPS.

b. Melakukan pembersihan data dan standarisasi
   Pembersihan data dilakukan dengan maksud mendapatkan data yang bersih. Data yang bersih adalah data yang konsisten dan tidak mengandung nilai yang tidak lengkap dan *noise*. Secara umum data yang tidak bersih adalah nilai yang tidak lengkap, data yang mengandung *noise* dan data yang tidak konsisten (Han Kamber, 2001). Setelah itu, proses transformasi dilakukan dengan normal isasi sehingga data siap untuk dianalisis.
c. Pembuatan peta grid
   Pembuatan peta grid atau *layer* output sebagai dasar input dalam pemodelan dengan SOM.
d. Melakukan analisis gerombol dengan metode Self Organizing Map

Langkah-langkah pengkajian metode SOM di atas digambarkan dalam diagram alur pada Gambar 4 di bawah ini.



Gambar 4: *Flow Chart* Penggerombolan Menggunakan SOM

Implementasi

Perangkat lunak yang digunakan dalam pembangunan aplikasi adalah:

1. C# sebagai bahasa pemprogaman dalam pengembangan sistem, dengan bahasa inilah peneliti melakukan pembuatan aplikasi.
2. Microsoft .NET *framework* 4.0 sebagai *platform* pengembangan aplikasi. Melalui *platform* ini peneliti membangun, mengembangkan dan menyebarkan serta menjalankan aplikasi ini. NET *framework* merupakan komponen *OS Windows* yang terintegrasi yang dibuat dengan tujuan untuk mendukung pengembangan berbagai macam jenis aplikasi serta untuk dapat menjalankan berbabagai macam aplikasi generasi mendatang.
3. *MockupBuilder* dan *Microsoft Visio* sebagai alat bantu untuk memodelkan rancangan antar muka.
4. R software versi 3.1.2 sebagai *backend processing.* Aplikasi R sebagai aplikasi utama dalam melakukan kegiatan komputasi yang bekerja dibelakang aplikasi yang dibangun.
5. *Library* R.NET 1.5.22.0 sebagai penghubung antara R dan C# dalam NET *framework.*
6. *Library* kohonen sebagai pendukung dalam pembuatan model SOM. *Library* ini untuk melakukan pemodelan dan penghitungan dalam melakukan analisa dengan SOM
7. *Package* maptools sebagai dasar dalam pembuatan dan pembacaan peta. Alur setelah dilakukan analisa dengan menggunakan *library* SOM, kemudian peneliti melakukan penggambaran dan penerapan pada peta mengenai gerombol yang terbentuk menggunakan *package* maptools.
8. *Package* clusterSim sebagai dasar untuk melakukan validasi. *Package* ini digunakan dalam melakukan validasi dari gerombol yang terbentuk dengan menggunakan Indeks Silohouette.

Implementasi Kode Program

Algoritma metode SOM yang telah dijabarkan sebelumnya kemudian

diimplementasikan ke dalam *script* bahasa pemograman R. *Script* tersebut ditulis dalam bentuk *String* pada bahasa pemograman C#. Berikut ini cuplikan *script* R metode *Self Organizing Map* pada Gambar 5 (Lampiran 1). Pada Gambar 6 (Lampiran 2) ditampilkan implementasi jendela utama dari aplikasi yang dibangun.

HASIL DAN PEMBAHASAN

Dalam penelitian ini dilakukan penggerombolan kota dan kabupaten di Provinsi Jawa Tengah dengan tujuan agar bisa melihat gambaran penggerombolan kota dan kabupaten di Jawa Tengah berdasarkan variabel pertumbuhan ekonomi tentang teori *Growth Pole*. Pengelompokan dilakukan menggunakan 14 variabel signifikan pada penelitian sebelumnya (dijelaskan pada Bab II) yaitu Produk Domestik Regional Bruto (PDRB), tingkat kepadatan penduduk (DENSITY), tingkat produktivitas tenaga kerja (*PRODUCT*), presentase penduduk miskin (*POVERTY*), Jumlah Penduduk (POP), Angka Pastisipasi Sekolah (APS1) usia 7-12 tahun, Angka Partisipasi Sekolah (APS2) usia 13-15 tahun, Angka Harapan Hidup (AHH), jumlah pasar (*MARKET*), jumlah Sekolah Dasaar (SD), jumlah Seko- lah Menengah Pertama (SMP), jumlah Sekolah Menengah Atas (SMA), panjang jalan (*ROAD*), dan penerimaan daerah (PAD).

Analisis Gerombol Data Studi Kasus

Sebelum melakukan analisis penggerombolan menggunakan metode *Self Organizing Map*,, maka ditentukan terlebih dahulu berapa jumlah gerombol yang optimal. Pada Gambar 7 (Lampiran 3) di bawah ini menampilkan hasil saran aplikasi berdasarkan Indeks Silhouette.

Dapat diketahui bahwa gerombol optimal berdasarkan Indeks Silhouette pada Gambar 7 adalah tiga gerombol karena memiliki nilai Indeks Silhouette paling besar, karena Indeks Silhouette mengukur tingkat kemiripan dalam suatu gerombol dibandingkan dengan gerombol lainnya. Sehingga yang dicari adalah yang mempunyai tingkat kemiripan (indeks Silhouette) yang paling besar. Sedangkan pada gambar di bawah ini menunjukkan tampilan saran berdasarkan *Within-cluster sum of squares* (WCSS). Berdasarkan kriteria dari WCSS yang menyatakan bahwa jumlah gerombol terbaik adalah yang meminimalkan nilai WCSS atau dengan kata lain yang meminimalkan variasi dalam gerombol. Pemilihan gerombol tiga atau empat dianggap optimal karena nilai WCSS lebih signifikan turun pada jumlah tersebut dibandingkan gerombol 5,6 dan seterusnya.

Berdasarkan hasil Indeks Silhouette dan WCSS (Gambar 7 dan Gambar 8 (Lampiran 3 dan 4)), peneliti bisa memilih gerombol tiga atau empat. Dalam penelitian ini peneliti memilih tiga gerombol sebagai dasar penggerombolan. Oleh karena itu, analisis gerombol menggunakan data studi kasus akan membagi Jawa tengah menjadi tiga gerombol yang dapat dilihat pada Gambar 9 (Lampiran 5).

Peneliti melakukan analisis penggerombolan wilayah Provinsi Jawa Tengah untuk Tahun 2008 sampai dengan Tahun 2010. Hasil penggerombolan menggunakan metode SOM menghasilkan tiga gerombol tiap tahunnya. Anggota gerombol yang cenderung tetap adalah pada gerombol ketiga sementara pada gerombol satu dan dua cenderung mengalami perubahan anggota gerombol. Dari hasil pemetaan, terlihat belum ada arah pemusatan pertumbuhan ekonomi. Pola yang terlihat masih menyebar. Hal ini berarti, pusat pertumbuhan ekonomi di Jawa Tengah belum bisa mendorong pertumbuhan ekonomi wilayah di sekitarnya.

Sejak tahun 2008 sampai 2010, kota dan kabupaten yang masuk pada gerombol dengan nilai variabel tertinggi

yaitu kabupaten Cilacap dan Kota Semarang, terlihat belum mempengaruhi wilayah disekitarnya. Meskipun begitu, Kabupaten Cilacap dan Kota Semarang terbukti sebagai kota dan kabu- paten yang memiliki ekonomi yang relatif stabil berdasarkan variabel yang digunakan pada tahun 2008 sampai tahun 2010.

KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini pada akhirnya mendapatkan kesimpulan yaitu aplikasi SOMgis yang dibangun telah dapat digunakan untuk melakukan penggerombolan dengan metode SOM dan dapat terintegrasi dengan Sistem Informasi Geografis. Akan tetapi, melihat dari hasil penggerombolan di peta dan hasil penggerombolan di codes plot, terlihat belum ada arah pemusatan pertumbuhan ekonomi. Hal ini berarti, pusat pertumbuhan ekonomi di Jawa Tengah belum bisa mendorong pertumbuhan ekonomi wilayah di sekitarnya. Serta terdapat perubahan pola penggerombolan kota dan kabupaten di Jawa tengah pada tahun 2008 sampai 2010.

Saran

Dalam rangka pengembangan penelitian khususnya yang terkait dengan metode SOM di masa yang akan datang, maka saran yang ditawarkan yaitu perlu dilakukan pengembangan aplikasi ini dengan metode SOM lain, seperti SuperSOM dan KSOM. Selain itu, perlu dipertimbangkan pula analisis jarak dalam membangun aplikasi terkait dengan aspek teori kutub pertumbuhan.

DAFTAR PUSTAKA

Arribas- Bel, Daniel dan Schmidt, H.R. (2011). *Self Organizing Maps and the US Urban Spatial Structure*. Arizona: Arizona State University.

Amri, Amir. (2007). Pengaruh Inflasi dan Pertumbuhan Ekonomi Terhadap Pengangguran di Indonesia. Jambi*: Jurnal Inflasi dan Pengangguran*, Vol. I no. 1.

Badan Pusat Statistik. (2011). *Produk Domestik Regional Bruto Atas Dasar Harga Konstan 2000 Menurut Provinsi, 2006-2010 (Milyar Rupiah).* http://bps.go.id/. (Diakses 19 Mei, 2015.)

Badan Pusat Statistik Provinsi Jawa Tengah. (2008). *Jawa Tengah Dalam Angka 2008.* Semarang: BPS Provinsi Jawa Tengah.

Badan Pusat Statistik Provinsi Jawa Tengah. (2009). *Jawa Tengah Dalam Angka 2009.* Semarang: BPS Provinsi Jawa Tengah.

Badan Pusat Statistik Provinsi Jawa Tengah. (2010). *Jawa Tengah Dalam Angka 2010*. Semarang: BPS Provinsi Jawa Tengah.

Badan Pusat Statistik Provinsi Jawa Tengah. (2011). *Jawa Tengah Dalam Angka 2011*. Semarang: BPS Provinsi Jawa Tengah.

Chen, Hao. (2010). *Comparative Study of C, C++, C and Java Programming Languages*. Finland: University of Applied Sciences.

Demuth, H. dan Beale, M.H. (2003). *Neural Network Toolbox for Use with MATLAB*. USA: The MathWorks, Inc.

Glasson, John. (1974). *An Introduction to Regional Planning*. London: Hucthinson and Co Publisher Ltd.

Han, J. dan Kamber, M. (2001). *Data Mining: Concepts and Techniques*. USA: Academic Press.

Isnainy, Mira Ayu. (2012). Kinerja Perekonomian Kabupaten/Kota di Provinsi Jawa Tengah Periode 1983-2010. [Skripsi]. Jakarta: STIS.

Jain, A. K. dan Dubes, R. C. (1998). *Algorithm for Clustering Data*. New Jesrsey: Prentice Hall.

Karima, Kourtit (2012). *Benchmarking of World Cities through Self Organizing Maps*. Amsterdam: vrije Universiteit Amsterdam.

Komite Percepatan dan Perluasan Pembangunan Ekonomi Indonesia. (2015). *Kerangka Acuan Kerja (KAK)*. Jakarta: Kementerian Koordinator Bidang Perekonomian.

Kumar, Satish dan Asger, Mohammed (2015). Analysis Clustering Techniques in Biological Data with R. *International Journal of Computer Science and Information Technologies*. Vol 6(2), 1859-1864.

Kulkarani, Rajendra. (2002). A Kohonen Self Organizing Map Approach to Modeling Growth Pole Dynamics. *Journal Network and Spatial Economics 2*, page 175-189.

Larose, D.T. (2004). *Discovering Knowledge in Data: An Introduction to Data Mining*. USA: John WileySons Inc.

Li, Yan. (2007). *Social Area Analysis Using SOM and GIS*. Ritsumeikan Center for Asia Pacific Studies: Ritsumeikan Asia Pacific University.

Mangiameli, P., Chen, S. K. dan West, D. (1996). A Comparison of SOM Neural Network and Hierachical Clustering Method. *European Journal of Operational Reserach*, 93, 402-417.

Nuningsih, S. (2010). K-Means Clustering (Studi Kasus Pada Data Pengujian Kualitas Susu di Koperasi Peternakan Bandung Selatan. [Skripsi]. Bandung:UPI.

Oja, M, Kaski, S. dan Kohonen, T. (2002). Bibliography of Self Organizing Map (SOM) Papers. 1998-2001.http://www.cis.hut.fi/4D56A069-A106-4B00-B4D6-8BA39A0EC385/Final Download/DownloadId-823AF7FF0046E49553BA41F31870

8C79/4D56A069-A106-4B00-B4D6-8BA39A0EC385/research/refs/NCS_vol3_1.pdf.(Diakses 1 September, 2015)

Salazar, E.J dkk. (2002). A Cluster Validity Index for Comparing Non Hierarchical Clustering Methods. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.5206&rep=rep1& type=pdf (Diakses 10 Mei, 2015)

Santoso, S. (2010). *Statistik Multivariat*. Jakarta: PT Elex Media Komputindo.

Siang, J. J. (2009). *Jaringan Syaraf Tiruan dan Pemrogramannya Menggunakan MATLAB (Ed. II)*. Yogyakarta: Andi Offset.

Sukirno, Sadono. (2006). *Ekonomi Pembangunan: Proses, Masalah, dan Dasar Kebijakan*. Jakarta: Kencana.

Turban, Efraim dkk. (2005*). Introduction to Information Technology, 3rd Edition*. New York: John Wiley Sons, Inc.

Wang Huai-Bin, dkk.(2010). A Clustering Algorithm Use SOM and K-Means in Intrusion Detection. *International Conference E-Business and E-Government (ICEE)*: 1281-1284. Guangzhou.

Yin, Huyun. (2012). *The Self Organizing Maps: Background, Theories, Extensions and Applications*. Berlin: Springer.

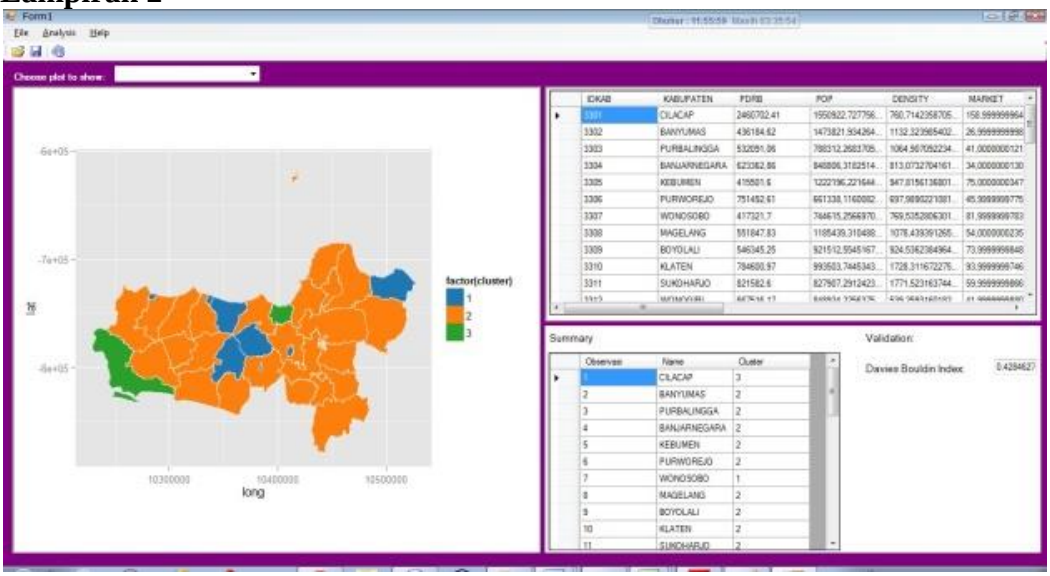
LAMPIRAN

Lampiran 1

```
//fungsi GRID
engine.Evaluate(@"somgriddo<-function(xd,yd,topo=c('rectangular','hexagonal')){
    topo<-match.arg(topo);
    x <- 1L:xd
    y <- 1L:yd
    pts <- as.matrix(expand.grid(x = x, y = y))
    if (topo == 'hexagonal') {
        pts[, 1L] <- pts[, 1L] + 0.5 * (pts[, 2L]%%2)
        pts[, 2L] <- sqrt(3)/2 * pts[, 2L]
    }
    res <- list(pts = pts, xdim = xd, ydim = yd, topo = topo)
    class(res) <- 'somgrid'
    res
    }");
engine.Evaluate("som_grid <- somgriddo(xd =xaxis , yd=yaxis, topo='hexagonal')");
```

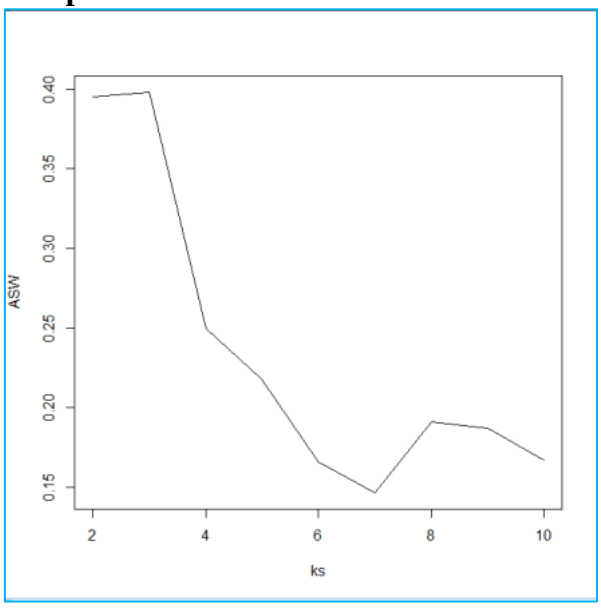
Gambar 5. Cuplikan *script* R di C#

Lampiran 2



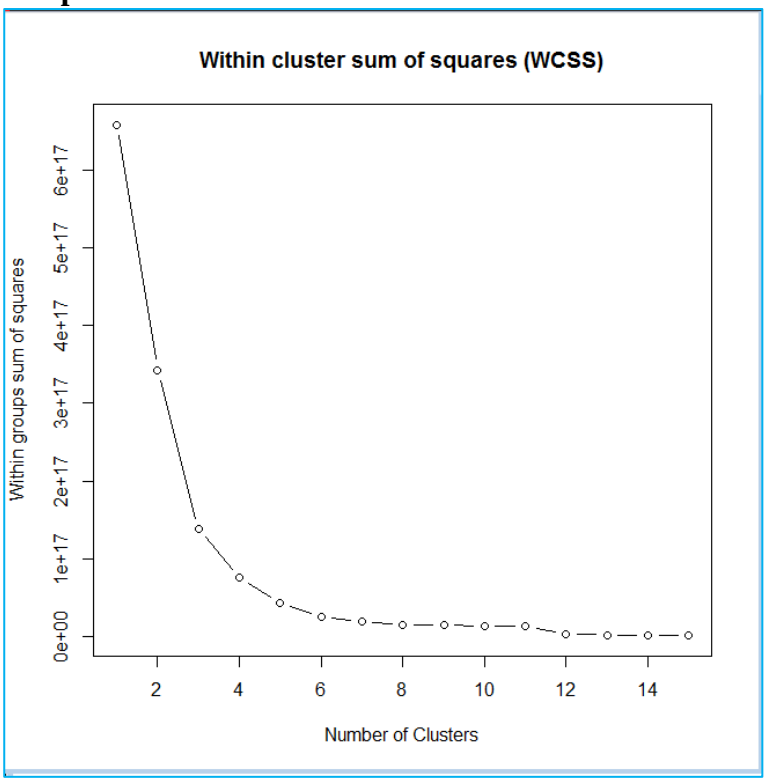
Gambar 6: Cuplikan Tampilan *Output* pada Jendela Utama

Lampiran 3



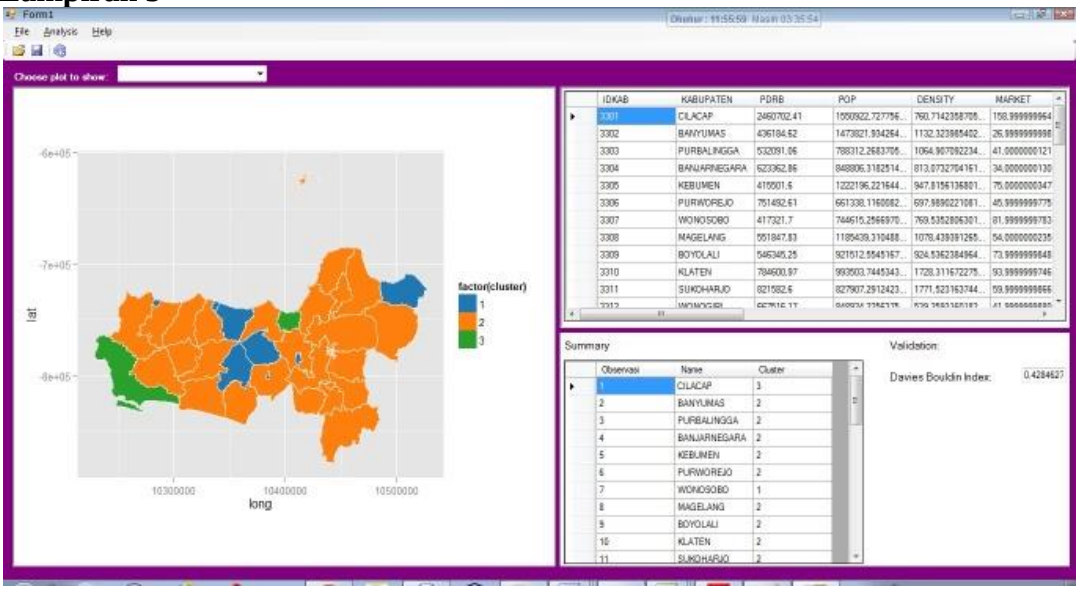
Gambar 7: Tampilan Saran Jumlah Gerombol Maksimal Berdasarkan Indeks Silhouette

Lampiran 4



Gambar 8: Tampilan Saran Jumlah Gerombol Maksimal Berdasarkan WCSS

Lampiran 5



Gambar 9: Hasil Analisis Penggerombolan Studi Kasus Tahun 2008 Menggunakan SOMgis

ANALISIS PREFERENSI MAHASISWA STIS BERDASARKAN AKUN FACEBOOK YANG DIMILIKI
STUDI KASUS: MAHASISWA STIS ANGKATAN 54 SAMPAI 57

STUDENTS PREFERENCE ANALYSIS BASED ON FACEBOOK ACCOUNT HELD IN STIS

Takdir
Sekolah Tinggi Ilmu Statistik

Choerul Afifanto
Sekolah Tinggi Ilmu Statistik

Masuk tanggal: 22-12-2015, revisi tanggal: 17-01-2016, diterima untuk diterbitkan tanggal: 19-01-2016

Abstrak

Penggunaan sosial media saat ini sangat masif di berbagai kalangan. Facebook merupakan salah satu sosial media yang memiliki jumlah dan frekuensi penggunaan yang besar serta memuat banyak data, khususnya data yang berupa relasi antarentitas. Penelitian ini mengidentifikasi preferensi, yakni kecenderungan topik yang digemari, mahasiswa STIS aktif berdasarkan akun Facebook yang dimiliki. Akun Facebook tersebut diperoleh dari grup-grup angkatan. Preferensi diperoleh dengan melakukan *crawling* terhadap halaman (*page*) yang di-*like* serta *group* yang diikuti oleh mahasiswa. Hasil dari penelitian ini adalah gambaran karakteristik preferensi mahasiswa berupa statistik mengenai jenis-jenis topik yang diminati oleh mahasiswa STIS serta visualisasi terbentuknya *cluster*/komunitas mahasiswa untuk topik tertentu. Pendekatan yang digunakan pada penelitian ini untuk mengekstraksi dan menganalisis data pada sosial media diharapkan dapat menjadi referensi bagi berbagai bidang penelitian yang memanfaatkan data social media.

Kata kunci: Facebook, Analisis Sosial Media, *Social Graph*

Abstract

Currently, social media is used massively in various societies. Facebook is one of the greatest social media in terms of total and frequency of uses, as well as the number of collected information, especially the information about relationships between entities. This study identifies preference of active STIS's students based on their Facebook account. Their Facebook accounts are collected from their Facebook group communities. The preference data are collected by crawling the liked pages and joined groups. The results of this study are the characteristics view of students' preferences in form of statistics of interesting topic types and visualization of students' clusters for certain topics. Approaches used in this research to extract and analyze data in social media could become a reference for another research fields which use social media data.

Keywords : *Facebook, Social Media Analysis, Social Graph*

PENDAHULUAN

Sosial media merupakan sebuah media yang populer untuk saling berinteraksi, berkomunikasi dan berkolaborasi antar pengguna secara *online* melalui internet (Wilson, Sala, Puttaswamy, & Zhao, Beyond Social Graphs: User Interactions in Online Social Networks and their Implications, 2012). Popularitas sosial media seperti Facebook, LinkedIn, dan Google+ semakin meningkat dalam beberapa tahun terakhir, sejak pertama kali dikenal di akhir tahun 1990an. Hal ini disebabkan oleh kemampuan sosial media yang dapat menghubungkan ratusan juta manusia di seluruh dunia tanpa adanya batasan geografis (Heidemann, Klier, & Probst, 2012). Selain itu, perbedaan antara sosial media dan halaman web tradisional juga menyebabkan popularitasnya meningkat di mata pengguna. Halaman web tradisional secara garis besar disusun dengan berorientasi pada konten, sedangkan sosial media disusun berdasarkan pengguna beserta

preferensinya (Mislove, Online social setworks: Measurement, analysis, and applications to distributed information system, 2009).

Seiring dengan popularitas sosial media yang semakin meningkat, skala penggunaannya juga semakin meningkat. Tercatat sebanyak 1.4 milyar pengguna internet mengakses sosial media di tahun 2012 dan semakin berkembang hingga hampir mencapai 2 milyar pengguna di tahun 2015 (Statista, 2015a). Sosial media yang paling populer yaitu Facebook dengan jumlah pengguna aktif sebesar 1,5 milyar pengguna di tahun 2015 (Statista, 2015b). Bahkan, Indonesia berada di posisi keempat dunia sebagai pengguna Facebook terbanyak dengan jumlah pengguna sebesar 60,3 juta pengguna, sedangkan posisi pertama ditempati oleh Amerika Serikat dengan jumlah pengguna sebesar 151.8 juta pengguna di tahun 2014 (Statista, 2014). Rata-rata waktu yang dihabiskan pengguna tiap harinya untuk mengakses Facebook selama 20+ menit (DMR Digital Statistics, 2015). Selain itu, setiap menit dalam sehari ada sebanyak 150.000 pesan terkirim, 10.000 permintaan pertemanan, 500.000 Facebook *likes*, serta 1.3 juta konten dibagikan oleh pengguna pada Facebook (Jeffbullas, 2015). Hal-hal tersebut menggambarkan seberapa berpengaruhnya sosial media khususnya Facebook dalam kehidupan manusia serta mengubah cara manusia untuk saling berkomunikasi dan berinteraksi.

Sosial media khususnya Facebook menangkap data-data yang berkaitan dengan individu melalui akun profil, interaksi antar pengguna secara langsung maupun melalui grup, dan konten yang disukai maupun dibagikan. *Dataset* tersebut dapat diperoleh dengan menggunakan aplikasi *crawling* data Facebook. Hanya saja *dataset* dalam jumlah besar hampir sulit didapatkan karena Facebook telah menerapkan beberapa pengaturan privasi pada data penggunanya, sehingga data yang didapat hanya sebatas data dari pengguna yang memiliki hubungan pertemanan dengan pelaku *crawling* (Rohman, Dewi, Riza, & Takdir, Sosial Graf untuk Visualisasi Data Facebook Menggunakan Visual Interaction System (Vis.js), 2014).

Popularitas, besarnya data yang dihasilkan dan tersedianya data tersebut secara publik memberikan peluang sekaligus tantangan bagi peneliti untuk melakukan penelitian terkait analisis sosial media, misalnya preferensi pengguna dalam skala besar melalui aktivitasnya di sosial media (Abbasi, Chai, Liu, & Sagoo, 2012). Tantangan bagi peneliti jaringan sosial adalah menemukan teknik terbaik untuk mengumpulkan dan memproses data dari jejaring sosial secara otomatis dan strategi untuk menyingkap ciri yang menggambarkan tipe-tipe jaringan yang kompleks. Selain itu, metode yang digunakan sebaiknya bisa bekerja dengan baik pada skenario skala besar (Ferrara, 2012).

Salah satu ciri yang ada pada jaringan sosial yang kompleks yaitu representasi struktur graf yang membentuk kumpulan simpul yang disebut komunitas, dengan kepadatan yang tinggi dari ikatan antar simpul dalam satu komunitas dan kepadatan yang rendah dari ikatan antar simpul pada komunitas yang berbeda (Lancichinetti, Kivela, Saramaki, & Fortunato, 2010).

Eksplorasi komunitas pada jaringan sosial sangat penting untuk beberapa alasan di antaranya: 1) mengungkap organisasi jaringan pada level yang tinggi sehingga bisa membantu untuk memformulasikan mekanisme untuk evolusinya; 2) memberikan pemahaman yang lebih baik mengenai proses dinamis pada jaringan sosial; 3) menemukan hubungan antar simpul yang semu dengan menginspeksi graf secara menyeluruh dan yang bisa menjadi atribut untuk fungsi dari sebuah sistem (Lancichinetti, Kivela, Saramaki, & Fortunato, 2010). Gambaran komunitas di dunia nyata sangat beragam. Salah satu di antaranya adalah berbentuk preferensi yaitu kecenderungan topik yang digemari.

Kemunculan dan *progress* social media yang pesat menciptakan sumber data baru untuk berbagai bidang penelitian.

Penelitian ini mengidentifikasi dan menganalisis preferensi mahasiswa STIS berdasarkan akun Facebook yang dimiliki dengan studi kasus: mahasiswa STIS angkatan 54 sampai angkatan 57. Sebuah studi yang mempelajari tentang kecenderungan topik yang digemari mahasiswa STIS angkatan 54-57 dengan cara melakukan visualisasi *social graph* untuk melihat keterkaitan sosial (*sociometric*) antarmahasiswa. Hasil penelitian dapat menunjukkan statistik preferensi mahasiswa, serta visualisasi terbentuknya *cluster* komunitas berdasarkan preferensi tersebut. Pendekatan yang digunakan pada penelitian ini untuk mengekstraksi dan menganalisis data sosial media diharapkan dapat memberikan referensi bagi berbagai bidang penelitian yang memanfaatkan sosial media.

Struktur penulisan paper ini dimulai dari studi literatur yang membahas tentang media sosial, analisis sosial media, dan *social graph*. Setalah itu, diikuti penjelasan mengenai metodologi yang digunakan dalam membuat visualisasi *social graph*, dimulai dari pemilihan sampel yang representatif, proses *scraping*/*crawling* data facebook, eksplorasi data, *tools* serta implementasi dan evaluasi.

METODOLOGI

Data yang digunakan dalam penelitian ini diperoleh dari jaringan sosial digital Facebook. Dalam bentuk sederhana, data meliputi simpul dan ikatan. Simpul merepresentasikan *fanpage* dan pengguna, sedangkan ikatan merepresentasikan hubungan antara *user* dan pengguna berupa "like" dari pengguna terhadap *fanpage* tersebut.

Data diperoleh dari Facebook menggunakan sebuah aplikasi bernama Netvizz[2] yang merupakan sebuah aplikasi yang memungkinkan peneliti untuk mengekstrak data yang dibutuhkan dari berbagai macam bagian Facebook dan menyimpan atau menampilkan file hasilnya dalam format yang standar (Rieder, 2013). Format standar yang digunakan adalah *matrix database* (GDF) yang hampir mirip dengan *comma separated file* (CSV). Selain itu, dilakukan pula crawling terhadap Facebook untuk memperoleh data publik.

Langkah-langkah ekstraksi data dengan menggunakan Netvizz dimulai dengan memberikan izin aplikasi untuk mengakses koneksi pertemanan Facebook. Kemudian membuka halaman aplikasi dan memilih parameter apa saja yang disertakan dalam ekstraksi data. Setelah proses ekstraksi selesai, akan didapatkan *file* GDF yang selanjutnya akan divisualisasikan dengan menggunakan *software* Gephi[3]. Daftar *users* yang akan diamati diperoleh dengan menggunakan aplikasi ini dengan *keywords* pencarian berupa "stis 54", "stis 55", "stis 56", dan "stis 57". Dari berbagai percobaan keyword yang dilakukan, jenis keyword tersebut yang memberikan hasil yang representatif terhadap target populasi yang ingin diamati. Penggunaan keyword dengan pola yang sama bertujuan untuk menyeragamkan metode yang digunakan dalam menelusuri populasi pada masing-masing angkatan sehingga mengurangi subjektifitas.

Pada penelitian ini, ada tiga tahapan proses ekstraksi data menggunakan Netvizz untuk mendapatkan data preferensi (kecenderungan topik yang digemari) mahasiswa STIS berdasarkan akun Facebook yang dimiliki dengan studi kasus: mahasiswa STIS angkatan 54-57. Tahapan tersebut dimulai dari ekstraksi data grup yang memiliki unsur STIS beserta angkatannya, *user* yang tergabung dalam grup tersebut, dan data preferensi *user*. Grup yang mewakili STIS beserta angkatannya ditetapkan dengan 2 kriteria berikut:

1. Grup angkatan
2. Grup kelas

Tahapan pengumpulan data ditunjukkan pada Gambar 1 (Lampiran 1). Setelah data melalui tahapan ekstraksi, data terlebih dahulu melalui tahapan *cleaning data*.

[2] https://apps.facebook.com/netvizz/

[3] https://gephi.org/

Hasil pencarian dengan Netvizz yang tidak mewakili target *user* yang diinginkan disaring untuk kemudian dilakukan *crawling* terhadap *member user* yang memenuhi kriteria yang telah ditentukan seperti contoh pada Tabel 1. Metode penyaringan dilakukan secara manual dengan memperhatikan dua kriteria yang telah ditetapkan, yakni harus berupa grup angkatan dan/atau grup kelas.

Tabel 1. Contoh Output Aplikasi Netvizz dan Pemilihan Grup Facebook

Terpilih	Nama Grup	Deskripsi
Ya	KS 54 STIS	
Ya	STIS 54 - 3SE5 2014/2015	Grup Kelas 3SE5 Angkatan 54. PKL54 lancar dan wisudanya bareng-bareng tahun 2016 :) Semangat 3SE5!
Tidak	PROBABILITA STIS'54	Kelompok 1 ------ PROBABILITA PK : Kak Nanda Adi Pradana (085273305460) …
Ya	2K STIS54	
Ya	STIS 54 C dan G	
Ya	2J STIS'54 2013/2014	
Ya	2KS2 STIS 54	Grup kelas 2KS2 angkatan 54 di Sekolah Tinggi Ilmu Statistik. Bervisi PKL bersama 2015, wisuda bersama-sama 2016!

Hal ini bertujuan untuk memastikan data tersebut valid dan dapat digunakan untuk analisis, sekaligus mereduksi ukuran data agar tidak terlalu besar ketika dianalisis menggunakan *software* Gephi.

Data yang diperoleh kemudian divisualisasi dan dianalisis dengan *software* Gephi untuk melihat karakteristik preferensi *user* secara jelas serta mendapatkan ukuran-ukuran statistik yang digunakan untuk analisis jaringan sosial. Untuk mendapatkan visualisasi yang jelas dan bermakna, digunakan *Force Atlas layout* dengan parameter-parameter yang disesuaikan dengan kebutuhan. *Layout* ini menghasilkan tampilan jaringan yang lebih jelas dari tiap-tiap komunitas atau *cluster*.

Algoritma *Force Atlas layout* merupakan algoritma *layout* spasial untuk jaringan web atau lebih dikenal dengan *small-world network*. Algoritma ini lebih berfokus pada kualitas tampilan daripada waktu. Algoritma ini bekerja dengan memastikan tampilan memiliki ikatan yang saling memotong seminimal mungkin. Oleh karena itu, algoritma ini mempermudah peneliti untuk interpretasi data yang sebenarnya walaupun waktu komputasinya cukup lama. Algoritma *Force Atlas layout* termasuk dalam kategori algoritma *force-directed* (Khokhar, 2015).

Beberapa parameter yang digunakan pada algoritma *Force Atlas layout* adalah:

- *Inertia* menunjukkan frekuensi simpul untuk mengubah posisinya pada ruang grafis untuk setiap iterasi pada algoritma. Nilai *default* yang digunakan adalah 0,1 yang berarti simpul tidak berubah posisinya secara signifikan pada ruang grafis.
- *Repulsion Strength* menunjukkan kekuatan setiap simpul untuk mendorong simpul lain. Semakin besar nilainya semakin terlihat renggang jaringan yang terbentuk. Nilai ini bisa diubah untuk memudahkan interpretasi tampilan yang dihasilkan. Nilai yang digunakan dalam penelitian ini adalah sebesar 500.
- *Attraction Strength* menunjukkan kekuatan setiap pasang simpul yang saling terhubung dalam menarik satu sama lain. Nilai *default* yang digunakan adalah sebesar 10.
- *Maximum Displacement* merupakan nilai yang digunakan untuk membatasi jumlah simpul yang bisa disingkirkan pada tampilan akhir jaringan. Nilai *default* yang digunakan adalah 10.
- *Auto Stabilize Function* membantu untuk menstabilkan jaringan pada saat algoritma dijalankan dengan membekukan simpul yang tidak stabil.

Namun, hal ini bisa mengurangi efisiensi dari algoritma.

- *Autostab Strength* menunjukkan kekuatan dari *Auto Stabilize Function* ketika opsi ini dipilih. Semakin tinggi nilainya, perpindahan simpul yang tidak stabil semakin jarang.
- *Autostab Sensibility* menunjukkan taraf dan kecepatan yang diadaptasi oleh *inertia* saat algoritma dieksekusi. Semakin besar nilainya, semakin tinggi taraf dan kecepatan yang diadaptasi oleh *inertia*. Nilai *default* yang digunakan adalah sebesar 0,2.
- *Gravity* menunjukkan kekuatan semua simpul terhadap pusat graf. Nilai ini mencegah penyebaran tampilan graf yang besar akibat simpul yang tidak saling terhubung. Nilai *default* yang digunakan adalah 30.
- *Attraction Distribution*, ketika opsi ini dipilih, algoritma akan memusatkan keanggotaan dari komunitas dan mendorong pusat komunitas mendekati tepi tampilan. Hal ini memudahkan peneliti dalam mendefinisikan komunitas.
- *Adjust by Sizes*, ketika opsi ini dipilih, algoritma akan berusaha untuk meminimalkan jumlah simpul yang saling tumpang tindih pada tampilan akhir.
- *Speed* menunjukkan kecepatan algoritma dalam melakukan penyebaran simpul. Semakin besar nilainya semakin cepat penyebaran yang dilakukan, namun mengurangi kualitas tampilan yang dihasilkan Nilai *default* yang digunakan adalah sebesar 1.

Untuk menganalisis preferensi pengguna, hal utama yang perlu diperhatikan adalah mengenai identifikasi komunitas. Komunitas dapat diidentifikasi dengan menjalankan statistik *Modularity* (kekuatan pembagian sebuah jaringan menjadi beberapa komunitas atau *cluster*) dengan membedakan tiap-tiap kemunitas dengan warna yang berbeda.. Algortima yang diimplementasikan pada statistik *Modularity* adalah algoritma *fast unfolding of communities in large networks* (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

Algoritma *fast unfolding of communities in large networks* mencari nilai *modularity* yang tinggi dari setiap partisi pada jaringan yang besar dalam waktu yang singkat dan membuka struktur hierarki komunitas secara lengkap dari jaringan. Algoritma ini terdiri dari dua fase yang dieksekusi pada setiap iterasi.

- Fase pertama mengenai identifikasi komunitas. Setiap simpul dalam jaringan dijadikan komunitas yang berbeda. Sehingga pada partisi awal, komunitas yang terbentuk adalah sebanyak simpul yang tersedia. Untuk setiap simpul *i* yang saling bertetangga dengan simpul *j* dihitung nilai *modularity* dengan cara mengambil *i* dari komunitasnya dan menempatkan simpul *i* ke komunitas simpul *j*. Simpul *i* diletakkan pada komunitas yang memiliki nilai *modularity* tertinggi dengan syarat hanya nilai *modularity* yang bernilai positif.
- Fase kedua mengenai pembangunan jaringan baru yang terdiri dari komunitas yang terbentuk pada fase pertama dengan cara memberi penimbang pada ikatan antar simpul baru dengan nilai jumlah dari penimbang pada ikatan antar simpul pada dua komunitas yang bersangkutan (Arenas, Duch, Fernandez, & Gomez, 2007).

Selain itu, pada panel "*Statistics*" ada banyak ukuran statistik seperti *average degree, graph density, modularity, dan average path length* yang bisa di-*run* untuk memudahkan analisis sekaligus membuat visualisasi data lebih bermakna (Keatinge, 2015). Penjelasan untuk statistik masing-masing adalah sebagai berikut (Wasserman, Stanley, & Faust, 1994):

- *Average Degree* menunjukkan rata-rata derajat (ikatan) yang ada di satu buah simpul dalam sebuah jaringan. Semakin besar nilai *average degree* semakin banyak ikatan yang ada pada sebuah simpul dalam jaringan.
- *Graph Density* menunjukkan perbandingan antara banyaknya ikatan yang tersedia dan ikatan yang mungkin tersedia dalam jaringan sosial. Semakin

besar nilai *graph density*, semakin banyak simpul yang saling terhubung.

- *Modularity* menunjukkan ukuran untuk mengukur kekuatan pembagian sebuah jaringan menjadi beberapa komunitas atau *cluster*. Sebuah jaringan yang memiliki nilai *modularity* tinggi, memiliki hubungan yang erat antar simpul dalam satu komunitas/*cluster*, namun memiliki hubungan yang lemah antar simpul di komunitas/*cluster* yang berbeda.
- *Average Path Length* adalah rata-rata dari semua jarak antar simpul dalam sebuah jaringan. *Average Path Length* juga menunjukkan ukuran untuk informasi yang mengalir dalam jaringan. Semakin besar nilainya, semakin cepat dan lancar informasi yang mengalir dalam jaringan tersebut.

Inti utama dari penelitian ini adalah untuk menganalisis preferensi mahasiswa STIS berdasarkan akun Facebook yang dimiliki. Oleh karena itu, identifikasi dari komunitas/*cluster* dalam jaringan adalah kunci utama dalam penelitian. Statistik yang digunakan adalah *modularity* (kekuatan pembagian sebuah jaringan menjadi beberapa komunitas/*cluster*) untuk mendeteksi tiap-tiap simpul masuk ke dalam komunitas/*cluster* tertentu sehingga bisa dibedakan warnanya tiap komunitas/*cluster*.

HASIL DAN PEMBAHASAN

Setelah melakukan visualisasi terhadap data yang diperoleh dari Facebook, diperoleh sejumlah grafik yang menujukkan dominasi suatu entitas. Dari keempat angkatan STIS yang diamati, 20 *fanpages* yang paling diminati ditunjukkan pada Tabel 3 (Lampiran 2).

Pada Tabel 3 dapat dilihat bahwa *fanpage* Senat Mahasiswa STIS dan Masa Pengenalan dan Pembentukan Karakter menempati peringkat teratas. Banyaknya jumlah *likes* pada kedua kategori tersebut menunjukkan bahwa mahasiswa di STIS, khususnya mahasiswa baru yang menjalani Masa Pengenalan dan Pembentukan Karakter sebelum memulai perkuliahan di STIS, telah mengenal dan bergelut dengan dunia sosial media, dalam hal ini Facebook. *Fanpage official* STIS menempati peringkat ke-4 setelah *fanpage* STIS yang dikelola oleh alumni dan mahasiswa STIS (*unofficial*). Daftar kategori top 20 *fanpages* tersebut menunjukkan hasil yang beragam. Hal ini menunjukkan pula keragaman secara makro preferensi yang dimiliki oleh mahasiswa STIS.

Setiap *fanpage* pada Facebook memiliki kategori yang ditentukan oleh pengelola *fanpage* tersebut. Statistik top 10 kategori *fanpages* yang paling banyak di-*like* ditunjukkan pada Tabel 2 berikut.

Tabel 2. Top 10 Kategori Fanpages yang Terbanyak Di-like pada Angkatan 54-57

Kategori Fan Page	Jumlah Likers
Musician/Band	3797
Organization	1929
Community	1732
App Page	1389
Education	1274
Movie	1154
Public Figure	1094
College & University	1066
News/Media Website	940
TV Show	934

Meskipun pada Tabel 3, kategori Musician/Band berada pada peringkat 20 untuk dengan *fanpage* Justin Bieber, namun pada Tabel 2, kategori tersebut merupakan kategori yang memiliki jumlah *likers* terbanyak secara signifikan. Justin Bieber, Taylor Swift, dan Avril Lavigne merupakan 3 musisi teratas berdasarkan jumlah *like* pada kategori tersebut. Hal ini menunjukkan bahwa mahasiswa STIS memiliki antusiasme terhadap musisi/band favorit yang cukup besar.

Nilai statistik untuk top 200 *fanpages* pada setiap angkatan yang diperoleh dari proses analisis sosial media ditunjukkan pada Tabel 4 (Lampiran 3).

Pada Tabel 4 terlihat bahwa angkatan 54 memiliki nilai *average degree* dan *graph density* terbesar dengan nilai 13,921 dan 0,071. Angka ini menunjukkan rata-

rata ikatan (derajat) untuk setiap simpul dalam jaringan serta rasio antara ikatan yang tersedia dan yang mungkin tersedia. Sedangkan nilai *modularity* terbesar diperoleh oleh angkatan 57 sebesar 0,258. Angka ini menunjukkan kekuatan pembagian jaringan menjadi beberapa komunitas pada angkatan 57 lebih besar dibandingkan angkatan lainnya. Banyaknya komunitas yang terbentuk yaitu sebanyak 5 komunitas untuk masing-masing angkatan selain angkatan 56 yang mencapai 195. Hal ini menunjukkan angkatan 56 memiliki data yang sangat heterogen. Selain itu, nilai *average path length* terbesar diperoleh oleh angkatan 57 sebesar 2,608 yang berarti informasi yang mengalir dalam jaringan sosial angkatan 57 lebih lancar daripada angkatan lainnya.

Angkatan 54

Angkatan 54 pada saat penelitian ini dilakukan adalah mahasiswa tingkat 4 STIS. Gambar 2 (Lampiran 4) menunjukkan visualisasi 200 *fanpages* yang paling banyak di-*like* oleh angkatan 54. *Dataset* mencakup 194 *users* dan 58 kategori *fanpage*. Semakin besar diameter *node* suatu *fanpages* pada visualisasi, maka semakin banyak jumlah *likes* yang diperoleh. Warna *node*s dan *edges* menunjukkan terbentuknya *cluster* sosial di mana sejumlah *fanpages* di-*like* oleh beberapa user tertentu yang membentuk suatu *cluster* secara tidak langsung. Secara kasat mata, terdapat 5 *cluster* sosial yang terbentuk. Selain itu terlihat pula bahwa *cluster* yang berwarna hijau dan merah memiliki perbedaan yang cukup signifikan, baik dari segi jarak/lokasi antar-*cluster* maupun keseragaman warna. Interpretasi dari hal tersebut adalah terdapat dua komunitas pengguna yang memiliki perbedaan preferensi pada *fanpages* yang terdapat pada Facebook. Sedangkan *cluster* yang berada di antara kedua kelompok tersebut merupakan kelompok pengguna yang menghubungkan keduanya. Pada cluster hijau, topik *fanpages* yang banyak dibahas adalah terkait kerohanian, seperti Rohis STIS, Wish Muharram 1435 H, Kartun Muslimah, dan Kajian Islam Statistik. Sedangkan pada cluster merah, topik-topik banyak berkaitan dengan entertainment, seperti Harry Potter, SpongeBob SquarePants, Cinema 21, dan Batik Indonesia.

Angkatan 55

Gambar 3 (Lampiran 5) merupakan visualisasi top 200 *fanpages* dengan jumlah *likes* terbanyak. Terdapat 132 *users* dan 60 kategori *fanpages* pada *dataset* yang divisualisasikan. Berbeda dengan angkatan 54 yang cenderung membentuk 2 buah *cluster* yang signifikan, visualisasi angkatan 55 menunjukkan karakteristik user yang lebih heterogen sehingga setiap *cluster* memiliki *node* yang tersebar diantara *cluster-cluster* lainnya. Dari sejumlah *fanpages* yang ada pada angkatan 55, belum terlihat dominasi *fanpages* yang diinisiasi oleh angkatan tersebut. *Likes* masih mendominasi *fanpages* komunitas umum yang ada di STIS, seperti STIS Bersih, Senat Mahasiswa STIS, dan Sekolah Tinggi Ilmu Statistik (official). *Cluster* merah pada angkatan 55 mengandung sejumlah *fanpages* yang beririsan dengan *cluster* hijau pada angkatan 54.

Angkatan 56

Dengan mekanisme pemilihan *dataset* dan teknik visualisasi yang sama, pada angkatan 56 diperoleh visualisasi yang memiliki lebih terstruktur. Dari visualisasi pada Gambar 4 (Lampiran 6) terlihat jelas bahwa terbentuk dua buah *cluster* besar, yakni hijau dan biru, di mana kedua *cluster* tersebut dihubungkan oleh *cluster* oranye. Masa Pengenalan dan Pembentukan Karakter merupakan *fanpage* yang dominan menyatukan mahasiswa angkatan 56. Topik yang banyak dibahas pada cluster hijau adalah topik mengenai hiburan, seperti Harry Potter, Justin Bieber, SpongeBob SquarePants, dan Dahsyat. Sedangkan pada cluster biru topik yang dibahas didominasi oleh *fanpages* seputar organisasi dan kegiatan di STIS,

seperti Senat Mahasiswa STIS, Dies Natalis STIS, Media Kampus STIS, dan Sekolah Tinggi Ilmu Statistik (Official). Namun demikian, terdapat pula beberapa *fanpages public figure* pada cluster biru yang memiliki jumlah like yang banyak, seperti Mario Teguh, Ustadz Felix Siauw, Susilo Bambang Yudhoyono, dan Yusuf Mansur (Official).

Angkatan 57

Angkatan 57 merupakan mahasiswa tahun pertama pada saat penelitian ini dilaksanakan, artinya mahasiswa pada angkatan ini baru mengikuti perkuliahan di STIS selama sekitar 3 bulan. Dengan mengambil 200 top *fanpages*, diperoleh sebanyak 273 *users*. Berdasarkan visualisasi yang diperoleh pada Gambar 5 (lampiran 7), *social graph* yang dihasilkan bersifat sangat heterogen. Keberagaman ini dapat dipengaruhi oleh karakteristik sosial media dari mahasiswa baru yang berasal dari lingkungan sekolah dan daerah yang berbeda sebelum menempuh pendidikan di STIS. Terlihat juga bahwa terdapat *gap*/*space* pada pusat visualisasi yang belum berisi *node* penghubung. Gap tersebut dapat dimanfaatkan untuk membuat *node* penghubung yang menjembatani keragaman karakteristik *user* pada angkatan 56 dengan memperkenalkan *fanpage* yang memuat topik dari semua *cluster* yang terbentuk.

KESIMPULAN DAN SARAN

Berdasarkan objek studi kasus yang diamati pada penelitian ini, dapat diperoleh kesimpulan bahwa preferensi mahasiswa STIS dapat tergambar melalui aktifitas yang dilakukan di sosial media. Hal ini dibuktikan dengan diperolehnya sejumlah *fanpages* yang beragam dari berbagai kategori dengan menelusuri aktifitas sosial media pada akun-akun Facebook yang mewakili mahasiswa STIS dari 4 angkatan yang aktif menjalani perkuliahan di STIS pada saat penelitian ini dilakukan.

Dari segi komunitas, penggunaan sosial media juga efektif untuk menyatukan berbagai komunitas yang berbeda namun memiliki tujuan yang sama. Hal ini ditunjukkan dengan jumlah *likes* yang diperoleh pada *fanpages* yang diciptakan terkait dengan organisasi dan kepanitiaan di STIS, seperti Senat Mahasiswa, Masa Pengenalan dan Pembentukan Karakter, Dies Natalis, dan STIS Bersih. Dengan mengetahui komunitas-komunitas yang ada pada STIS, khususnya tiap angkatan, dosen dapat melakukan inovasi pengajaran sesuai dengan komunitas yang diminati oleh mahasiswa yang diajarkan. Hal ini tentu akan menghindari kebosanan dan membuat mahasiswa antusias mengikuti pembelajaran.

Sosial media merupakan salah satu sumber data yang memiliki kelebihan tersendiri dibandingkan dengan pengumpulan data secara konvensional seperti survei. Data pada sosial media dapat diperoleh dengan *effort* yang lebih kecil, namun dengan *filtering* dan analisis data yang tepat dapat memberikan *insight* yang jauh lebih cepat dibandingkan dengan survei. Namun demikian, data yang terdapat pada sosial media sulit untuk dipertanggungjawabkan kebenarannya dan berpotensi terkena *spam* yang dapat mengaburkan data sesungguhnya. Oleh karena itu, penelitian lanjutan diperlukan, serta pengembangan teknik-teknik analisis agar dapat menghasilkan data yang representatif. Perbaikan regulasi bagi pengguna internet oleh pemerintah dan pengawasan pengguna yang dilakukan oleh vendor sosial media yang semakin baik juga dapat menjadi harapan untuk menjadikan sosial media sebagai sumber data yang valid.

DAFTAR PUSTAKA

Abbasi, M. A., S. K. Chai, H. Liu, and K. Sagoo. 2012. "Real-world behavior analysis through a social media lens." *Social Computing, Behavioral-Cultural Modeling and Prediction.* Springer. 18-26.

Aggarwal, Charu C. 2011. *Social Network Data Analytics.* Springer.

2015. *DMR Digital Statistics.* Accessed Desember 16, 2015. http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/.

Heidemann, J., M. Klier, and F. Probst. 2012. "Online social networks: A survey of a global phenomenon." *Computer Network, 56(18)* 3866-3878.

2015. *Jeffbullas.* Accessed Desember 16, 2015. http://www.jeffbullas.com/2015/04/17/21-awesome-facebook-facts-and-statistics-you-need-to-check-out/.

Keatinge, Fergus J.D. 2015. "Examining the effects of digital social networks on new physical human interactions and social networks: A validation of Dunbar's Numbers." *Social Networking* 72-79.

Mislove, A. E. 2009. "Online social setworks: Measurement, analysis, and applications to distributed information system." *ProQuest.*

Rieder, B. 2013. "Studying Facebook via data extraction: The Netvizz application." *WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference.* New York: ACM. 346-355.

Rohman, Abdul, Ardani Yustriana Dewi, Kemas M. Irsan Riza, and Takdir. 2014. "Sosial Graf untuk Visualisasi Data Facebook Menggunakan Visual Interaction System (Vis.js)."

2015b. *Statista.* Accessed Desember 16, 2015. http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

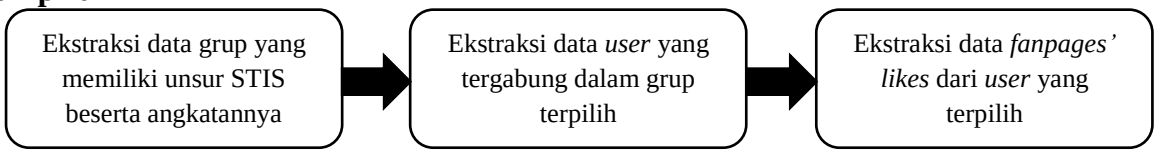
2015a. *Statista.* Accessed Desember 16, 2015. http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

2014. *Statista.* Accessed Desember 16, 2015. http://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/.

Wilson, Christo, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. 2012. "Beyond Social Graphs: User Interactions in Online Social Networks and their Implications." *ACM Transactions on Web.*

LAMPIRAN

Lampiran 1



Gambar 1. Tahapan Pengumpulan Data

Lampiran 2

Tabel 3. Top 20 Fanpages dengan Jumlah Likers Terbanyak pada Angkatan 54-57

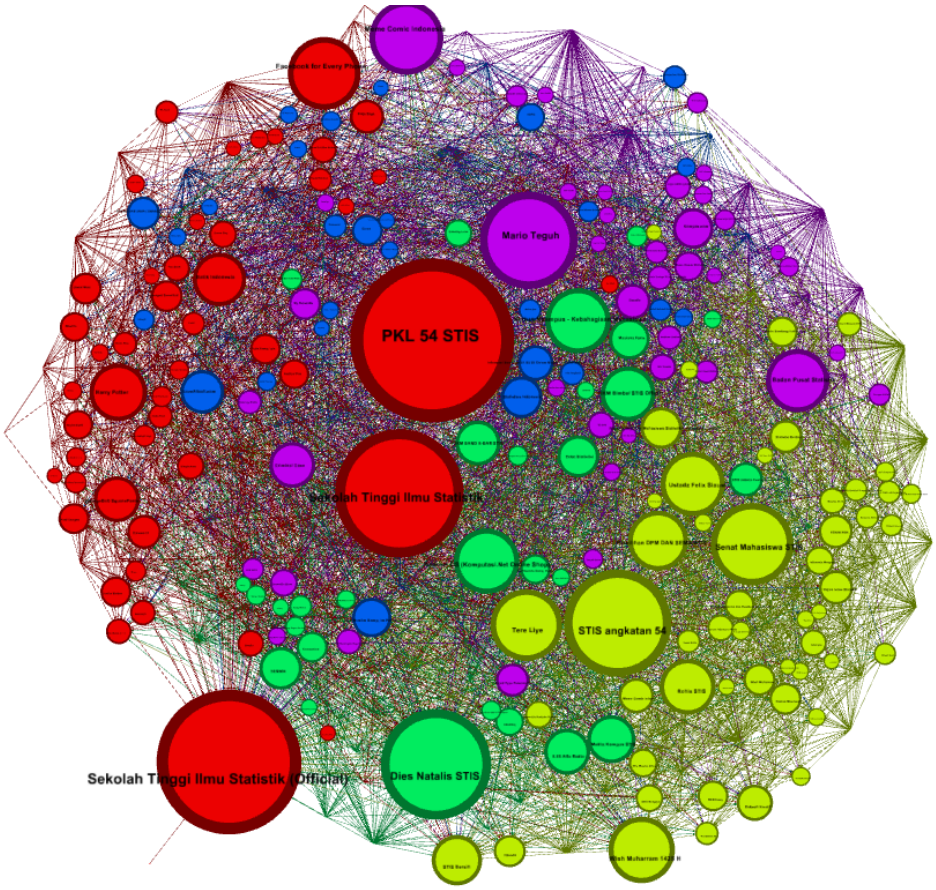
Fan Page	Kategori	Jumlah Likers
Senat Mahasiswa STIS	Organization	541
Masa Pengenalan dan Pembentukan Karakter	Organization	539
Sekolah Tinggi Ilmu Statistik	College & University	493
Sekolah Tinggi Ilmu Statistik (Official)	College & University	476
Facebook for Every Phone	App Page	475
Mario Teguh	Public Figure	440
Harry Potter	Movie	419
Dies Natalis STIS	Education	317
Meme Comic Indonesia	Entertainment Website	256
SpongeBob SquarePants	TV Show	249
UKM Bimbel STIS Official	Education	234
PKL 54 STIS	News/Media Website	230
Ninja Saga	App Page	227
Media Kampus STIS	Media/News/Publishing	224
Batik Indonesia	Clothing	215
Ustadz Felix Siauw	Author	211
Badan Pusat Statistik	Government Organization	204
Pemilihan DPM DAN SEMA STIS	Organization	199
Tere Liye	Writer	197
Justin Bieber	Musician/Band	185

Lampiran 3

Tabel 4. Nilai Statistik untuk Top 200 fanpages pada Setiap Angkatan

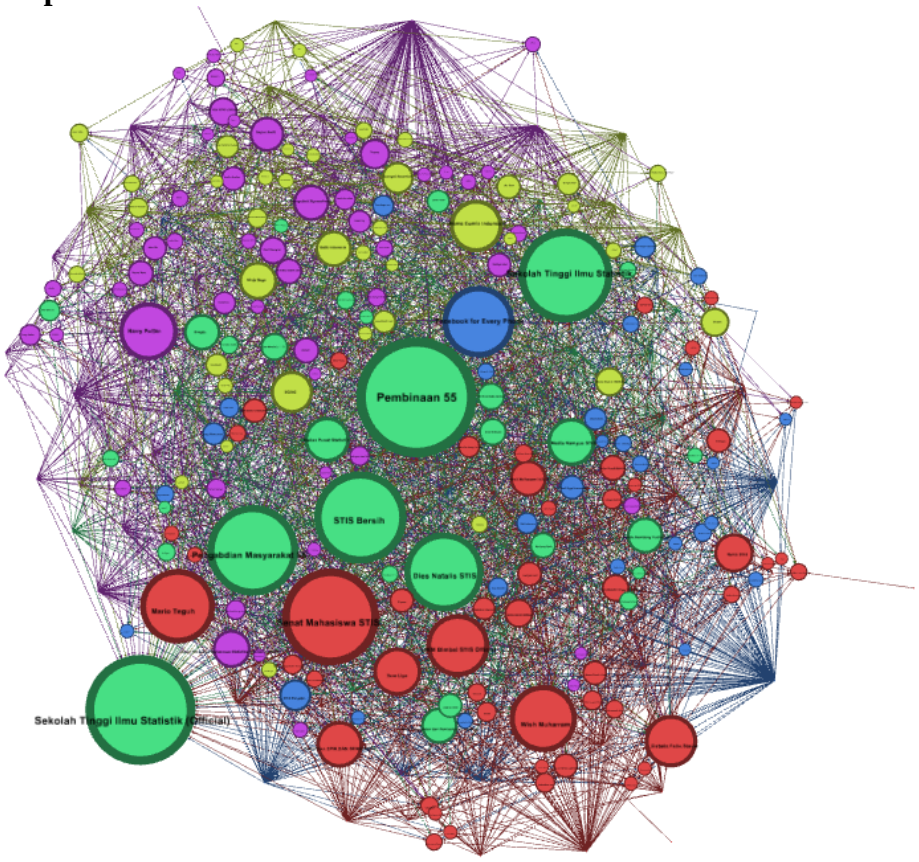
Ukuran Jaringan	Angkatan 54	Angkatan 55	Angkatan 56	Angkatan 57
Simpul	391	332	782	473
Ikatan	5443	3165	8636	5226
Avg. Degree	13,921	9,533	11,043	11,049
Graph Density	0,071	0,058	0,028	0,047
Modularity	0,196	0,216	0,24	0,258
Number of Communities	5	5	195	5
Avg. Path Length	2,407	2,513	2,46	2,608

Lampiran 4



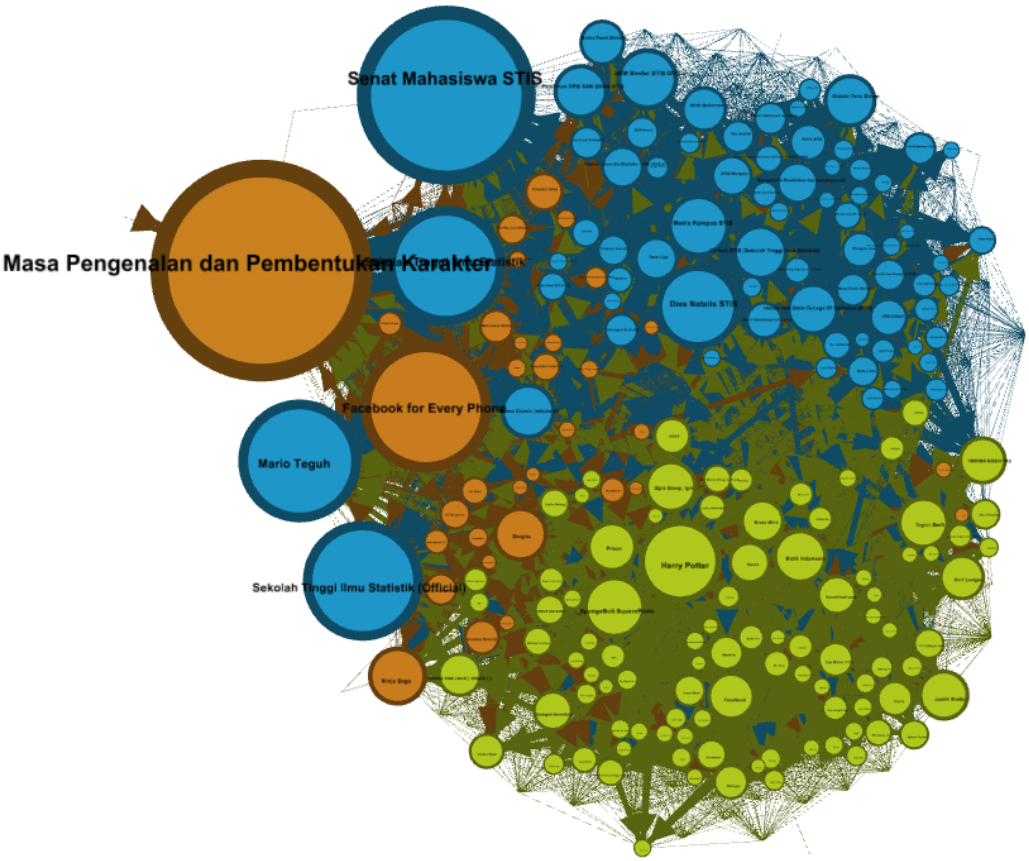
Gambar 2. Top 200 Likes pada Angkatan 54

Lampiran 5



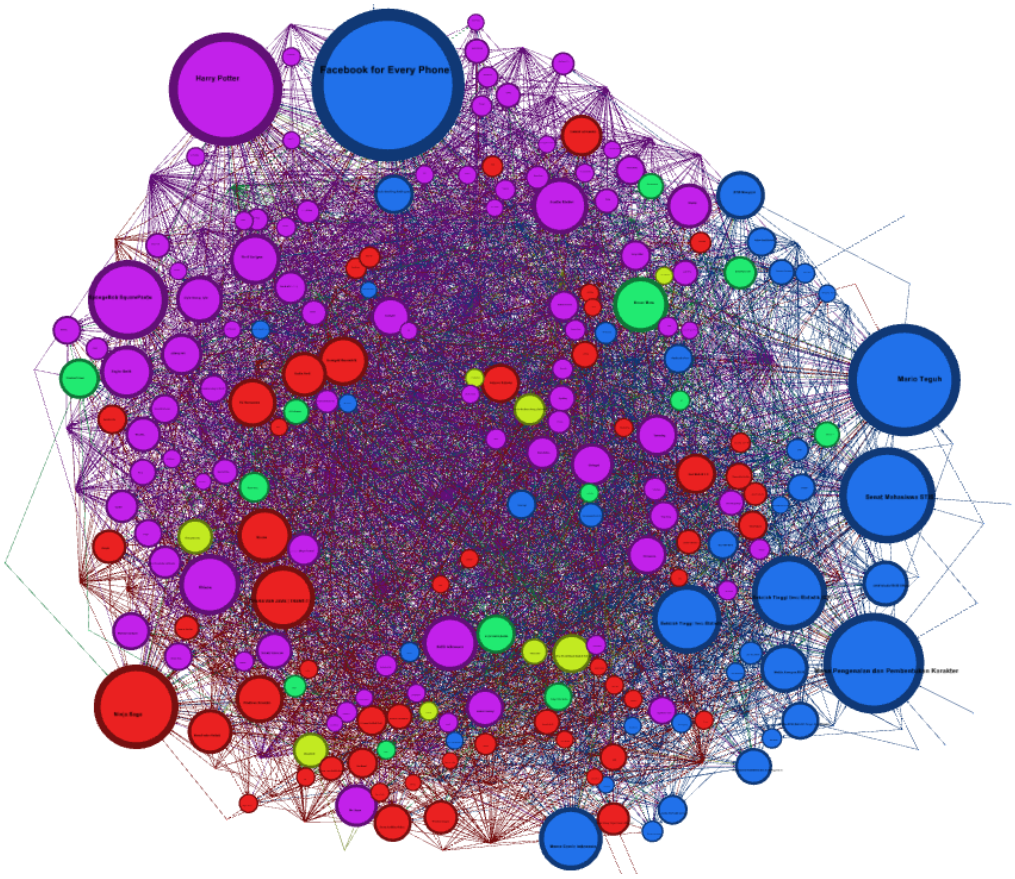
Gambar 3. Top 200 Likes pada Angkatan 55

Lampiran 6



Gambar 4. Top 200 Likes pada Angkatan 56

Lampiran 7



Gambar 5. Top 200 Likes pada Angkatan 57

Indeks

Indeks Penulis