JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

VOLUME 10, NOMOR 1, JUNI 2018 ISSN 2086 – 4132
AKREDITASI NOMOR: 747/Akred/P2MI-LIPI/04/2016

PUSAT PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
POLITEKNIK STATISTIKA STIS

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

Jurnal "Aplikasi Statistika dan Komputasi Statistik" memuat karya ilmiah hasil penelitian dan kajian teori statistik 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.

Alamat Redaksi:
Politeknik Statistika STIS
Jl. Otto Iskandardinata 64C
Jakarta Timur 13330
Telp. 021-8191437

Redaksi menerima karya ilmiah atau artikel penelitian mengenai kajian teori statistik 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

Puji syukur kehadirat Allah, Tuhan Yang Maha Esa, "Jurnal Aplikasi Statistika dan Komputasi Statistik" Volume 10, Nomor 1, Juni 2018 dapat diterbitkan. Jurnal ilmiah ini dapat terwujud atas partisipasi semua pihak, penulis internal dilingkungan Politeknik Statistika STIS maupun penulis eksternal, serta mitra bestari.

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 agar publikasi yang dihasilkan menjadi salah satu sarana untuk memberikan sosialisasi statistika bagi masyarakat.

Jakarta, Juni 2018

Ketua Dewan Redaksi,

Setia Pramana, Ph.D

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

VOLUME 10, NOMOR 1, JUNI 2018
AKREDITASI NOMOR: 747/Akred/P2MI-LIPI/04/2016

DAFTAR ISI

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

(*Journal of Statistical Application & Statistical Computing*)

ISSN 2086 – 4132 **Volume 10, Nomor 1, Juni 2018**

DDC: 315.98

Sri Wahyuni dan Yogo Aryo Jatmiko

Pengelompokan Kabupaten/Kota di Pulau Jawa Berdasarkan Faktor-Faktor Kemiskinan dengan Pendekatan *Average Linkage Hierarchical Clustering*

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 1 – 8

Abstrak

Pulau Jawa masih merupakan pulau dengan persentase penduduk miskin terbesar di Indonesia. Dalam menentukan kebijakan penanggulangan kemiskinan, perlu diperhatikan faktor-faktor yang mempengaruhi kemiskinan.Selain itu, kemiskinan di setiap wilayah memiliki karakteristik yang berbeda, sehingga perlu adanya pengelompokan wilayah agar kebijakan yang akan dilaksanakan tepat sesuai dengan karakteristik wilayah. Tujuan dari penelitian ini adalah mengelompokkan kabupaten/kota di Pulau Jawa berdasarkan faktor-faktor kemiskinan tahun 2017 dengan pendekatan *average linkage* hierarchical clustering. Faktor-faktor kemiskinan yang digunakan sebagai dasar pengelompokan adalah tingkat pengangguran terbuka, persentase rumah tangga yang bekerja di pertanian, pengeluaran rumah tangga per kapita, dan rata-rata lama sekolah. Hasil penelitian menunjukkan ada dua kelompok wilayah kabupaten/kota di Pulau Jawa. Kelompok pertama, terdiri dari Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Timur, Kota Surabaya, Kota Jakarta Pusat, Kota Malang, Kota Bandung, Kota Yogyakarta, Kota Jakarta Utara, Kota Depok, Kabupaten Bantul, Kota Salatiga, Kota Tangerang Selatan, Kota Madiun, Kabupaten Sleman, Kota Bekasi, Kabupaten Sidoarjo, Kota Semarang, Kota Tangerang, Kota Surakarta. Sedangkan sebanyak 99 kabupaten/kota lainnya masuk dalam kelompok kedua. Kelompok pertama merupakan kota-kota besar di Indonesia yang tingkat kemiskinannya rendah, sedangkan kelompok kedua sebagian besar terdiri dari kabupaten/kota yang dicirikan dengan wilayah perdesaan yang tingkat kemiskinannya tinggi.

Kata kunci: Pulau Jawa, faktor kemiskinan, *average linkage hierarchical clustering*

DDC: 315.98

Takdir

Analisis Kinerja, Kualitas Data, dan *Usability* pada Penggunaan *CAPI* untuk Kegiatan Sensus/Survey

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 9 – 26

Abstrak

Pengumpulan data merupakan suatu tahapan pada Sensus/Survey yang sangat menentukan keberhasilan Sensus/Survey. Prosesnya yang memakan waktu lama akan mengakibatkan data yang disajikan tidak relevan dengan kondisi pada saat pelaksanaan. Dengan *Computer-Assisted Personal Interview* (*CAPI*), proses entri data dapat dilakukan pada saat proses interview berlangsung. Hal ini mempersingkat tahapan pengumpulan data hingga data tersedia pada sistem komputer dan siap untuk dianalisis. Pada penelitian ini, indikator-indikator penting penentu keberhasilan penerapan *CAPI*, yakni kinerja, kualitas data, dan usability diukur untuk melihat sejauh mana *CAPI*

memberikan penyempurnaan pada pengumpulan data. Penelitian ini memberikan rekomendasi, baik dari segi konsep, maupun teknis, mengenai desain *CAPI* untuk kegiatan sensus/survey.

Kata kunci: *CAPI*, sensus, survey, pengumpulan data

DDC: 315.98

Andri Yudhi S dan Aris Rusyiana

Beras atau Rokok?: Beban Ekonomis Rumah Tangga Miskin di Indonesia 2014

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 27 – 38

Abstrak
Fakta bahwa di beberapa negara berkembang, konsumsi rokok menimbulkan beban ekonomis yang signifikan (Toukan, 2016; Block dan Webb, 2009). Juga, untuk konteks Indonesia kontemporer, Kepala BPS mengatakan bahwa belanja rokok merupakan pengeluaran kedua terbesar dan memberikan kontribusi nyata terhadap angka kemiskinan nasional. Namun, kajian kontemporer yang secara komprehensif membahas beras dan rokok terhadap kemiskinan belum banyak dibahas. Celah penelitian tersebut menjadi dasar bagi kami untuk melakukan kajian mengenai hubungan konsumsi beras dan pengeluaran potensial rokok di antara rumah tangga miskin di Indonesia 2014. Untuk keperluan telaah kajian penelitian ini, kami membagi kategori rumah tangga berdasarkan tempat tinggal (perdesaan/perkotaan), rumah tangga dengan banyak anggota rumah tangga usia dewasa (di atas 15 tahun), dsb. Tujuan dari kajian ini adalah untuk menganalisa apakah rumah tangga miskin lebih memilih mengurangi konsumsi beras dibanding mengurangi konsumsi rokok .Untuk kajian ini, kami menggunakan Survei Sosial Ekonomi Nasional tahun 2014. Dengan menggunakan Model Regresi Linier Berganda, kami menggunakan sampel rumah tangga yang memiliki anggota rumah tangga dewasa yang merokok (NIndonesia = 285.371). Hasil penelitian kami menunjukkan bahwa rumah tangga miskin yang memiliki anggota rumah tangga perokok secara rata-rata mengkonsumsi beras relatif lebih sedikit dibandingkan rumah tangga yang tidak memiliki anggota rumah tangga perokok, baik yang termasuk kategori miskin maupun tidak. Hal ini mengindikasikan bahwa rumah tangga miskin lebih memprioritaskan konsumsi rokok dibandingkan konsumsi beras.

Kata kunci: Susenas, rumah tangga miskin, konsumsi rokok, regresi linier berganda

DDC: 315.98

Toza Sathia Utiayarsih, Jadi Suprijadi dan Bernik Maskun

Pengelompokan Pengguna Situs Web BPS Melalui Teknik *Bibliometric* dan Analisis Korespondensi

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 39 – 52

Abstrak
Salah satu upaya pemenuhan program percepatan (*quick wins*) terhadap produk BPS yang benar-benar dapat menyentuh kebutuhan para pengguna data adalah dengan melakukan segmentasi terhadap pengguna data. Segmentasi terhadap pengguna situs web BPS sebagai salah satu bentuk segmentasi terhadap pengguna data, sesuai program percepatan. Ukuran data pengguna web sangat besar dan berupa data teks sehingga tidak dapat langsung dianalisis melalui aplikasi statistik yang tersedia, maka perlu dilakukan suatu teknik untuk data pengguna web dengan menggunakan teknik *bibliometric*. Teknik tersebut mengubah data teks menjadi format numerik, selanjutnya dibuat menjadi matriks distribusi frekuensi. Matriks digunakan pada analisis korespondensi untuk pengelompokkan pengguna situs

web. Hasil dari analisis pengguna situs web BPS yang diwakili oleh alamat IP dapat dikelompokkan dengan halaman yang diakses berdasarkan asal negara, sehingga didapatkan segmentasi pengguna data situs web BPS antara negara dan halaman yang diakses.

Kata kunci: *Data mining*, *text mining*, *bibliometric*, *web mining*, analisis korespondensi

DDC: 315.98

Farid Ridho dan Arya Aji Kusuma

Deteksi Intrusi Jaringan dengan *K-Means Clustering* pada Akses *Log* dengan Teknik Pengolahan *Big Data*

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 53 – 66

Abstrak
Keamanan jaringan, adalah salah satu aspek penting dalam terciptanya proses komunikasi data yang baik dan aman. Namun, masih adanya serangan yang efektif membuktikan bahwa sistem keamanan yang berlaku belum cukup efektif untuk mencegah dan mendeteksi serangan. Salah satu metode yang dapat digunakan untuk mendeteksi serangan ini adalah dengan dengan *Intrusion Detection System* (IDS). Besarnya data (*volume*), cepatnya perubahan data (*velocity*), serta variasi data (*variety*) merupakan ciri-ciri dari *Big Data*. Akses log, secara teori termasuk dalam kategori ini sehingga dapat dilakukan pemrosesan menggunakan teknologi *Big Data* dengan *Hadoop*. Hal ini mendorong penulis untuk dapat menerapkan metode pengolahan baru yang dapat mengatasi perkembangan data tersebut, yaitu *Big Data*. Penelitian ini dilakukan dengan menganalisis akses log dengan *K-Means Clustering* menggunakan metode pengolahan *Big Data*. Penelitian menghasilkan satu model yang dapat digunakan untuk mendeteksi sebuah serangan dengan probabilitas deteksi sebesar 99.68%. Serta dari hasil perbandingan kedua metode pengolahan *Big Data* menggunakan *pyspark* dan metode tradisional menggunakan *python* standar, metode *Big Data* memiliki perbedaan yang signifikan dalam waktu yang dibutuhkan dalam eksekusi program.

Kata kunci: *IDS*, *big data*, akses log, k-means, *clustering*

DDC: 315.98

Sukim dan Rudi Salam

Pola Fertilitas Wanita Usia Subur di Indonesia: Perbandingan Tiga Survei Demografi dan Kesehatan Indonesia (2002, 2007 dan 2012)

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 67 – 78

Abstrak
Tingkat fertilitas merupakan salah satu faktor demografi yang paling menentukan dalam penurunan tingkat pertumbuhan penduduk di Indonesia. Salah satu ukuran fertilitas adalah *Total Fertility Rate* (TFR). Selama 20 tahun terakhir diketahui laju pertumbuhan penduduk di Indonesia stagnan pada angka 1,49 persen. Oleh karenanya, penelitian ini bertujuan untuk mengkaji pola TFR selama periode 20 tahun terakhir berdasarkan tiga Survei Demografi dan Kesehatan Indonesia (SDKI) tahun 2002, 2007 dan 2012. Metode yang digunakan adalah Regresi data count. Hasil penelitian menunjukkan bahwa dari ketiga SDKI tersebut, tanda koefisiennya adalah sama untuk semua variabel penjelas kecuali pada SDKI 2007 yaitu pada variabel tempat tinggal yang berbeda dengan SDKI 2002 dan 2012. Sejalan dengan temuan ini perlu studi lebih lanjut untuk mencari teori yang dapat menjelaskan temuan empirik tersebut.

Kata kunci: Fertilitas, TFR, SDKI, regresi data *count*

JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

(*Journal of Statistical Application & Statistical Computing*)

ISSN 2086 – 4132 **Volume 10, Nomor 1, Juni 2018**

DDC: 315.98

Sri Wahyuni dan Yogo Aryo Jatmiko

Pengelompokan Kabupaten/Kota di Pulau Jawa Berdasarkan Faktor-Faktor Kemiskinan dengan Pendekatan *Average Linkage Hierarchical Clustering*

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 1 – 8

Abstract
Java is still an island with the largest percentage of poor people in Indonesia. In determining poverty reduction policies, it is necessary to consider the factors that influence poverty. Moreover, poverty in each region has different characteristics, so there needs to be regional grouping so that the policies that will be implemented are in accordance with the characteristics of the region. The purpose of this study is to classify regencies in Java based on poverty factors in 2017 with the average linkage hierarchical clustering approach. The poverty factors that are used as a basis for grouping are level of open unemployment, percentage of agricultural households, household expenditure per CAPIta, and mean years schooling. The results showed that there were two groups of regencies in Java. The first group, consisti of West Jakarta City, South Jakarta City, East Jakarta City, Surabaya City, Central Jakarta City, Malang City, Bandung City, Yogyakarta City, North Jakarta City, Depok City, Bantul Regency, Salatiga City, South Tangerang City, Madiun City, Sleman Regency, Bekasi City, Sidoarjo Regency, Semarang City, Tangerang City, Surakarta City. Whereas 99 other regencies were included in the second group. The first group is large cities in Indonesia with a low poverty rate, while the second group consists mostly of districts / cities characterized by rural areas with high poverty levels.

Keywords: Java island, poverty factor, average linkage hierarchical clustering

DDC: 315.98

Takdir

Analisis Kinerja, Kualitas Data, dan *Usability* pada Penggunaan *CAPI* untuk Kegiatan Sensus/Survey

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 9 – 26

Abstract
Data collection is a phase in census/survey phases which highly affect the success of cencus or survey. Using Computer-Assisted Personal Interviewing (CAPI), data entry could be carried out during interview. It could shorten the data collection stage until data were available on a computer system and ready for analysis. In this study, the essential indicators which determine the success of CAPI implementation, i.e. performance, data quality, and usability are mesured to undestand the signifacancy of CAPI in improving data collection. This study proposed recommendation, either in the aspect of concept, or technical regarding CAPI design for census/survey.

Keywords: CAPI, Census, Survey, Data Collection

DDC: 315.98

Andri Yudhi S dan Aris Rusyiana

Beras atau Rokok?: Beban Ekonomis Rumah Tangga Miskin di Indonesia 2014

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 27 – 38

Abstract
Facts that in many developing countries, cigarettes consumption affects significantly toward economic burden (for instances see Toukan, 2016; Block and Webb 2009). Also, for Indonesian recently context, Chief of Statistics Indonesia says that cigarettes expenditure pose the second highest shared towards the national poverty rate. However, the recently comprehensive Indonesia researches on rice and cigarettes expenditure are still rare. Regarding those research gaps, we examine the linkage of rice consumption expenditure and the potential cost of cigarettes expenditure among poor households in Indonesia (includes the households characteristics: residency, social safety net receiver, adults smokers among households, etc). The objectives of this study is to examine whether poor households prefer to consume fewer rice rather than consuming fewer cigarettes. For this study, we use the National Social Economic Survey of the 2014 year dataset. By applying the multiple linear regression analysis, we use sample of adult smokers (N=285,371). Our results show that poor smoking-households relatively consume rice less than the non-smoking-households categories on average. This may indicate that poor households prioritize to consume more cigarettes rather than consuming rice.

Keywords: Susenas, poverty rate, cigarettes consumption, multiple linier regression

DDC: 315.98

Toza Sathia Utiayarsih, Jadi Suprijadi dan Bernik Maskun

Pengelompokan Pengguna Situs Web BPS Melalui Teknik *Bibliometric* dan Analisis Korespondensi

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 39 – 52

Abstract
The effort to fulfill one of quick wins program for BPS products that really can fulfill the needs of data users is by segmenting data users. Segmentation of BPS website users as a form of segmentation of data users, according to quick wins program. The size of web user data is very large and in the form of text data so that it cannot be directly analyzed through available statistical applications, it is necessary to do a technique for web user data using bibliometric techniques. This technique converts text data into numeric format, then it is made into a frequency distribution matrix. The matrix is used in correspondence analysis for grouping website users. The results of the analysis of BPS website users represented by IP addresses can be grouped with pages accessed based on national origin, so that segmentation users of BPS website data between the country and the page are accessed can be obtained.

Keywords: Data mining, text mining, bibliometric, web mining, correspondence analysis

DDC: 315.98

Farid Ridho dan Arya Aji Kusuma

Deteksi Intrusi Jaringan dengan *K-Means Clustering* pada Akses *Log* dengan Teknik Pengolahan *Big Data*

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 53 – 66

Abstract

Good network security planning ensures the safety and comfort of user data. However, the existence of effective attacks proves that the current security system is not effective to prevent and detect attacks. One of methods that can be used to detect this attack is by using Intrusion Detection System (IDS). The amount of data (volume), speed of which data change (velocity), and variations in data (variety) are characteristics of big data. Log access, theoretically is also a form of big data so a new approach in statistical data processing is needed to overcome big data. This research was conducted by analyzing log access with K-Means Clustering using the big data processing technique. The study produced a model that can be used to detect an attack with a detection probability of 99.68%. As well as a comparison between big data using Pyspark and traditional processing technique using standard python, which big data technique has a significant difference in time needed to execute the program.

Keywords: IDS, big data, log access, k-means, clustering

DDC: 315.98

Sukim dan Rudi Salam

Pola Fertilitas Wanita Usia Subur di Indonesia: Perbandingan Tiga Survei Demografi dan Kesehatan Indonesia (2002, 2007 dan 2012)

Jurnal Aplikasi Statistika & Komputasi Statistik, Volume 10, Nomor 1, Juni 2018, hal 67 – 78

Abstract
Fertility rate is one of the most decisive demographic factors in the decline in the rate of population growth in Indonesia. One measure of fertility is Total Fertility Rate (TFR). During the last 20 years, the population growth rate in Indonesia is stagnant at 1.49 percent. Therefore, this study aims to examine TFR patterns over the last 20 years based on the three Indonesia Demographic and Health Survey (SDKI) in 2002, 2007 and 2012. This study used Regression data count method. The results showed that of the three SDKIs, the coefficient values are the same for all explanatory variables except in SDKI 2007 i.e. in residential variables that are different from the 2002, 2012 SDKI. In line with this finding, further studies are needed to find a theory that can explain this empirical finding.

Keywords: Fertility, TFR, IDHS, regression data count

PENGELOMPOKAN KABUPATEN/KOTA DI PULAU JAWA BERDASARKAN FAKTOR-FAKTOR KEMISKINAN DENGAN PENDEKATAN AVERAGE LINKAGE HIERARCHICAL CLUSTERING

Sri Wahyuni[1], Yogo Aryo Jatmiko[2]

Badan Pusat Statistik
e-mail: [1]swahyuni@bps.go.id

Abstrak

Pulau Jawa masih merupakan pulau dengan persentase penduduk miskin terbesar di Indonesia. Dalam menentukan kebijakan penanggulangan kemiskinan, perlu diperhatikan faktor-faktor yang mempengaruhi kemiskinan.Selain itu, kemiskinan di setiap wilayah memiliki karakteristik yang berbeda, sehingga perlu adanya pengelompokan wilayah agar kebijakan yang akan dilaksanakan tepat sesuai dengan karakteristik wilayah. Tujuan dari penelitian ini adalah mengelompokkan kabupaten/kota di Pulau Jawa berdasarkan faktor-faktor kemiskinan tahun 2017 dengan pendekatan *average linkage hierarchical clustering.* Faktor-faktor kemiskinan yang digunakan sebagai dasar pengelompokan adalah tingkat pengangguran terbuka, persentase rumah tangga yang bekerja di pertanian, pengeluaran rumah tangga per kapita, dan rata-rata lama sekolah. Hasil penelitian menunjukkan ada dua kelompok wilayah kabupaten/kota di Pulau Jawa. Kelompok pertama, terdiri dari Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Timur, Kota Surabaya, Kota Jakarta Pusat, Kota Malang, Kota Bandung, Kota Yogyakarta, Kota Jakarta Utara, Kota Depok, Kabupaten Bantul, Kota Salatiga, Kota Tangerang Selatan, Kota Madiun, Kabupaten Sleman, Kota Bekasi, Kabupaten Sidoarjo, Kota Semarang, Kota Tangerang, Kota Surakarta. Sedangkan sebanyak 99 kabupaten/kota lainnya masuk dalam kelompok kedua. Kelompok pertama merupakan kota-kota besar di Indonesia yang tingkat kemiskinannya rendah, sedangkan kelompok kedua sebagian besar terdiri dari kabupaten/kota yang dicirikan dengan wilayah perdesaan yang tingkat kemiskinannya tinggi.

Kata kunci: Pulau Jawa, faktor kemiskinan, *average linkage hierarchical clustering*

Abstract

Java is still an island with the largest percentage of poor people in Indonesia. In determining poverty reduction policies, it is necessary to consider the factors that influence poverty. Moreover, poverty in each region has different characteristics, so there needs to be regional grouping so that the policies that will be implemented are in accordance with the characteristics of the region. The purpose of this study is to classify regencies in Java based on poverty factors in 2017 with the average linkage hierarchical clustering approach. The poverty factors that are used as a basis for grouping are level of open unemployment, percentage of agricultural households, household expenditure per CAPIta, and mean years schooling. The results showed that there were two groups of regencies in Java. The first group, consisti of West Jakarta City, South Jakarta City, East Jakarta City, Surabaya City, Central Jakarta City, Malang City, Bandung City, Yogyakarta City, North Jakarta City, Depok City, Bantul Regency, Salatiga City, South Tangerang City, Madiun City, Sleman Regency, Bekasi City, Sidoarjo Regency, Semarang City, Tangerang City, Surakarta City. Whereas 99 other regencies were included in the second group. The first group is large cities in Indonesia with a low poverty rate, while the second group consists mostly of districts / cities characterized by rural areas with high poverty levels.

.

***Keywords**: Java Island, poverty factor, average linkage hierarchical clustering*

PENDAHULUAN

1. Latar Belakang

Kemiskinan masih menjadi isu dunia karena jumlahnya yang besar dan dampak yang ditimbulkannya sangat buruk bagi kehidupan masyarakat. Sejak 25 September 2015, seluruh masyarakat dunia secara resmi berkomitmen untuk melaksanakan Agenda 2030 yang tersaji dalam Tujuan Pembangunan Berkelanjutan atau Sustainable Development Goals (SDGs) yang terdiri dari 17 tujuan dan 169 target. Dalam SDGs, penanggulangan kemiskinan menjadi tujuan pertama target pembangunan. Target yang ingin dicapai adalah mengakhiri kemiskinan dalam segala bentuk dimanapun.

Keseriusan pemerintah dalam upaya mencapai target penurunan kemiskinan tercantum dalam Rencana Pembangunan Jangka Menengah (RPJM) 2015-2019 yang menunjukkan bahwa salah satu visi pembangunan nasional adalah mempercepat pemerataan dan keadilan (Bappenas, 2014). Menurut Badan Pusat Statistik (2018), jumlah penduduk miskin di Indonesia diperkirakan sebesar 26,58 juta orang atau sekitar 10,12 persen dari total penduduk pada tahun 2017. Dari total ini sekitar 52 persen penduduk miskin berada di Pulau Jawa. Fenomena semacam ini mengindikasikan bahwa strategi pengentasan kemiskinan yang telah diterapkan belum mampu menciptakan pemerataan pendapatan (redistribution of income), mengatasi ketimpangan-ketimpangan serta mengurangi kemiskinan, terutama di Jawa. Problematika kemiskinan yang dialami masyarakat Jawa merupakan penghambat bagi upaya peningkatan kesejahteraan penduduk. Hal ini disebabkan oleh pelbagai sisi sosial budaya dan ekonomi yang melekat pada kondisi kemiskinan itu sendiri yang disebut sebagai lingkaran setan kemiskinan (vicious circle of poverty).

Untuk merancang penanggulangan kemiskinan harus memperhatikan beberapa aspek di setiap wilayah. Aspek-aspek tersebut mencakup aspek sosial, ekonomi, budaya, politik serta aspek waktu dan ruang. Faktor-faktor penyebab kemiskinan perlu terlebih dahulu diketahui agar strategi penanggulangan kemiskinan sesuai dengan kondisi masyarakat di setiap wilayah. Tujuan dari penelitian ini adalah melakukan pengelompokan kabupaten/kota di Pulau Jawa yang didasari adanya faktor-faktor kemiskinan agar program pengentasan kemiskinan menjadi lebih terarah, efektif dan tepat sasaran. Bararakbah dan Arai (2004) menyebutkan metode pengelompokan yang baik adalah metode yang mempunyai nilai simpangan baku dalam kelompok yang minimum dan nilai simpangan baku antar kelompok yang maksimum. Laraswati (2014) menemukan bahwa metode *average linkage* dan *complete linkage* merupakan metode yang lebih baik diantara metode pengelompokan K-Means. Laeli (2014) menemukan bahwa metode *average linkage* mempunyai kinerja yang lebih baik daripada metode Ward. Berdasarkan hasil penelitian tersebut dan untuk mencapai tujuan penelitian, maka pengelompokan kabupaten/kota dilakukan dengan metode analisis cluster *average linkage*.

2. Tinjauan Pustaka

Definisi kemiskinan sesungguhnya luas maknanya, karena faktor penyebab yang kompleks, indikator maupun permasalahan lain yang ada didalamnya. Kemiskinan tidak hanya dipandang dari dimensi ekonomi, namun juga pada dimensi sosial, kesehatan, pendidikan dan berbagai dimensi lainnya. Menurut Badan Pusat Statistik (BPS), penduduk miskin yaitu penduduk hidup di bawah garis kemiskinan, atau dengan kata lain penduduk yang tidak mampu memenuhi kebutuhan dasar minimum makanan dan non makanan. Garis kemiskinan adalah besarnya nilai pengeluaran (dalam rupiah) untuk memenuhi kebutuhan dasar minimum makanan dan non makanan. Nilai garis kemiskinan yang digunakan mengacu pada kebutuhan minimum 2.100 kilo kalori per kapita per hari ditambah dengan kebutuhan minimum non makanan yang merupakan kebutuhan dasar seseorang. Kebutuhan dasar tersebut meliputi papan, sandang,

sekolah, transportasi, serta kebutuhan rumahtangga dan individu yang mendasar lainnya. Pada penelitian ini, konsep kemiskinan mengacu pada konsep yang telah dibuat oleh BPS.

Penelitan yang terkait dengan kemiskinan sudah banyak dilakukan. Leasiwal (2013) menyebutkan bahwa kemiskinan di Maluku didominasi oleh penduduk yang tinggal di perdesaan. Adapun variabel yang secara signifikan mempengaruhi kemiskinan yakni daya beli masyarakat, inflasi, rata-rata lama sekolah, angka melek huruf, angka partisipasi kasar, angka harapan hidup, dan jumlah sekolah menengah atas. Chandra dan Nafisah (2017) dalam penelitiannya membagi wilayah Provinsi Jawa Timur ke dalam 3 kelompok kabupaten/kota. Kelompok tersebut dikategorikan ke dalam kelompok tingkat rendah, kelompok tingkat sedang dan kelompok tingkat tinggi berdasarkan faktor-faktor kemiskinan. Faktor-faktor yang digunakan yakni persentase angka melek huruf, persentase tingkat pengangguran terbuka, persentase angka partisipasi sekolah usia 16-18 tahun, dan persentase pendidikan.

Sementara Bachtiar, dkk (2016) melakukan pengkajian faktor-faktor yang mempengaruhi kemiskinan anak Balita di Provinsi Sumatera Barat. Hasil kajiannya menyebutkan beberapa faktor yang akan memberikan peluang anak balita jatuh pada kemiskinan, yaitu pendidikan yang rendah, pekerjaan ibu dan kepala rumah tangga, tinggal di wilayah perdesaan, orang tua memiliki balita lebih dari satu orang. Kemudian Kurniawan (2017) dalam penelitiannya menyebutkan bahwa faktor-faktor penyebab kemiskinan adalah pendidikan dan pendapatan.

Zuhdiyati dan Kaluge (2017) meneliti faktor-faktor yang mempengaruhi kemiskinan di Indonesia selama lima tahun terakhir yaitu 2011-2015. Penelitian ini menggunakan pendekatan kuantitatif dan regresi data panel, dengan sumber data dari Badan Pusat Statistik. Variabel yang dimasukkan dalam model meliputi Indeks Pembangunan Manusia (IPM), pertumbuhan ekonomi dan tingkat pengangguran terbuka (TPT). Hasil penelitian mereka menunjukkan bahwa yang berpengaruh terhadap kemiskinan di Indonesia adalah IPM, sedangkan pertumbuhan ekonomi dan TPT tidak berpengaruh.

Terkait dengan penelitian yang menggunakan analisis *cluster*, Ningsih dkk (2016) melakukan pengelompokan kabupaten/kota di Provinsi Kalimantan Timur berdasarkan data produksi palawija. Pendekatan metode yang dipakai adalah *complete linkage* dan *average linkage*. Hasilnya, terdapat 4 kelompok kabupaten/kota, yaitu kelompok pertama kabupaten/kota penghasil palawija sangat sedikit, kelompok kedua penghasil palawija cukup banyak, kelompok ketiga penghasil palawija terbanyak, dan kelompok keempat penghasil palawija sangat sedikit.

METODE PENELITIAN

1. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang bersumber dari Badan Pusat Statistik tahun 2017 yaitu:

X_1: Persentase rumah tangga yang bekerja di pertanian

X_2: Rata-rata lama sekolah

X_3: Pengeluaran rumah tangga per kapita

X_4: Tingkat pengangguran terbuka

Software yang digunakan untuk melakukan pengolahan adalah SPSS 22 dan STATA 13. Software SPSS digunakan untuk melakukan analisis faktor dan analisis *cluster*, sedangkan STATA 13 digunakan untuk membuat peta.

2. Langkah-Langkah Analisis Data

Dalam penelitian ini, langkah yang dilakukan dalam menghasilkan pengelompokan wilayah adalah sebagai berikut:

1. Melakukan analisis deskriptif faktor-faktor kemiskinan
2. Mengelompokkan faktor-faktor kemiskinan. Dalam analisis faktor, terlebih dahulu dilakukan pengujian adanya korelasi antar variabel dengan uji Barlett dan Kaiser-

Meyer- Olkin (KMO) untuk kelayakan suatu data.

3. Hasil analisis faktor kemudian digunakan sebagai input untuk melakukan pengelompokan kabupaten/kota dengan metode analisis *cluster average linkage*.

3. Analisis *Cluster* Metode *Average Linkage*

Analisis merupakan teknik analisis multivariate yang digunakan untuk mengelompokkan data observasi atau variabel-variabel ke dalam cluster berdasarkan faktor-faktor yang telah ditentukan. Tujuan analisis cluster adalah mengelompokkan obyek yang mirip ke dalam satu cluster yang sama.

Metode pengelompokan (clustering) dalam analisis cluster ada 2, yaitu metode hierarki dan metode nonhirarki. Analisis hierarki, pengklusteran datanya dilakukan dengan cara mengukur jarak kedekatan pada setiap objek yang kemudian disajikan dalam bentuk dendogram. Ada beberapa macam analisis cluster dengan metode hierarki, antara lain *single linkage*, *complete linkage*, dan *average linkage*. Single linkage, pembentukan cluster didasarkan pada jarak terkecil. Jika dua obyek terpisah oleh jarak yang pendek maka kedua obyek tersebut akan digabung menjadi satu cluster. Complete linkage, berlawanan dengan *single linkage*, pengelompokkannya berdasarkan jarak terjauh. Metode *average linkage* menghitung jarak dua cluster yang disebut sebagai jarak rata-rata. Keuntungan metode hierarki antara lain mempercepat proses pengolahan dan menghemat waktu karena data input akan membentuk hierarki atau tingkatan sehingga mempermudah dalam penafsiran.

Dalam metode *average linkage*, jarak dihitung pada masing-masing cluster dengan persamaan sebagai berikut:

$$d_{(uv)w} = \frac{\sum_i \sum_k d_{ik}}{V_{(uv)} N_w} \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Keterangan:

d_{ik}: jarak objek i dalam cluster (uv) dan objek k dalam cluster w.

$V_{(uv)}$: jumlah objek dalam cluster uv

N_w: jumlah objek dalam cluster w

HASIL DAN PEMBAHASAN

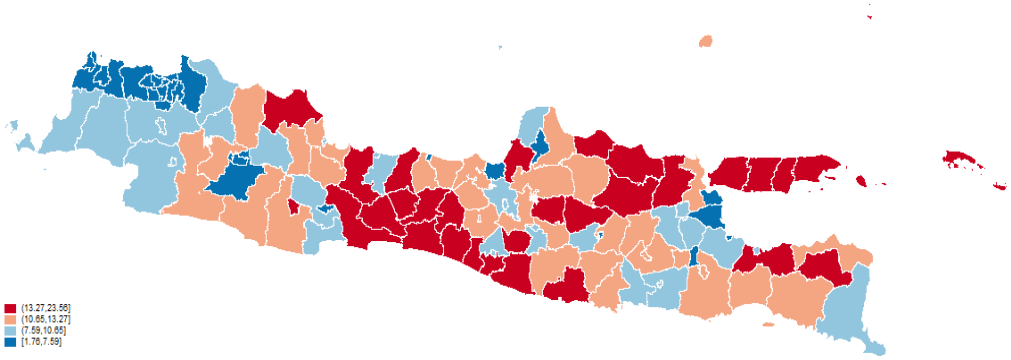
1. Analisis deskriptif

Analisis menggunakan 119 kabupaten/kota yang terletak di Pulau Jawa sebagai objek observasi. Rata-rata persentase rumah tangga miskin di kabupaten/kota sebesar 10,7 persen dengan nilai minimum 1,76 persen dan maksimum 23,56 persen. Pertanian masih menjadi mata pencaharian bagi sebagian rumah tangga di Jawa dengan rata-rata 22,88 persen, minimum 0,19 persen dan maksimum 65,58 persen. Tingkat pendidikan kepala rumah tangga di Pulau Jawa masih tergolong rendah. Tabel 1 menunjukkan rata-rata lama sekolah kepala rumah tangga sekitar 8,03 tahun atau setara dengan kelas 2 sekolah menengah pertama. Rata-rata lama sekolah minimum 4,12 tahun atau setara dengan kelas 4 SD dan maksimum 11,77 tahun atau setara dengan SMA kelas 2 atau 3. Dilihat dari pengeluaran per kapita, rata-ratanya 11.053,95 dengan nilai minimum 7.250 dan maksimum 23.098. Kecilnya pengeluaran per kapita ini menunjukkan bahwa di Pulau Jawa masih banyak rumah tangga yang belum mampu memenuhi kebutuhan hidupnya secara layak. Pengangguran di Pulau Jawa memiliki rata-rata 5,45 persen dengan nilai minimum 0,85 persen dan maksimum 13 persen. Semakin tingginya pengangguran akan berdampak pada berkurangnya pendapatan, sehingga rumah tangga akan sulit untuk hidup secara layak (lihat Tabel 1)

Tabel 1. Deskripsi Faktor-Faktor Kemiskinan di Pulau Jawa, 2017

Variabel	Observasi	Rata-Rata	Standar Deviasi	Minimum	Maksimum
Persentase rumah tangga miskin	119	10,70	4,64	1,76	23,56
Persentase rumah tangga yang bekerja di pertanian	119	22,88	17,03	0,19	65,58
Rata-rata lama sekolah	119	8,03	1,65	4,12	11,77
Pengeluaran rumah tangga per kapita (Rp)	119	11.053,95	2.780,70	7.250	23.098
Tingkat Pengangguran Terbuka (TPT)	119	5,45	2,53	0,85	13,00

Sumber: Badan Pusat Statistik, 2017



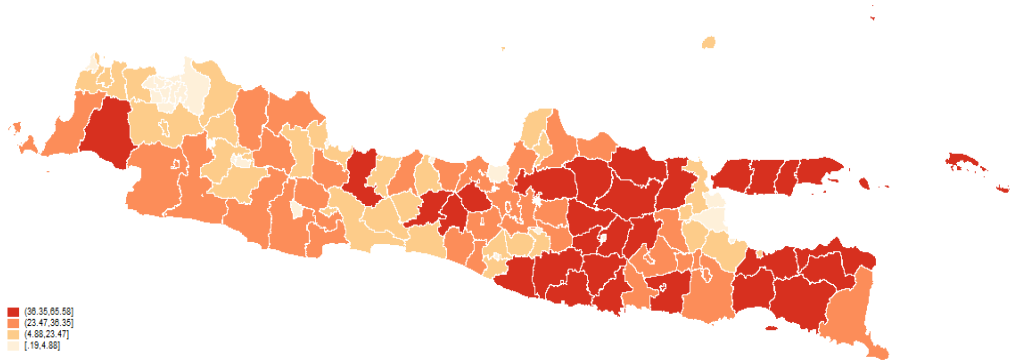
Gambar 1. Peta Wilayah Kabupaten/Kota di Pulau Jawa Berdasarkan Persentase Penduduk Miskin, 2017

Persentase Penduduk Miskin di Pulau Jawa

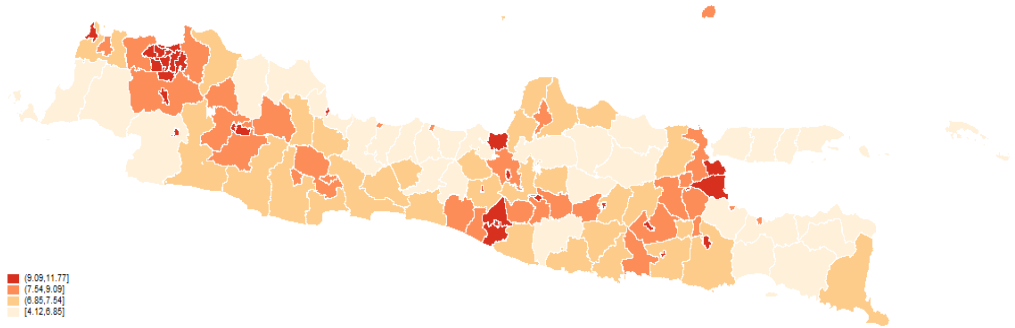
Gambar 1 menunjukkan peta wilayah berdasarkan persentase penduduk miskin di Pulau Jawa. Warna biru tua tua menunjukkan kabupaten/kota dengan persentase penduduk miskin paling rendah. Warna biru muda menunjukkan kabupaten/kota dengan persentase penduduk miskin cukup rendah. Warna merah muda menunjukkan kabupaten/kota dengan persentase penduduk miskin cukup tinggi dan warna merah tua menunjukkan kabupaten/kota dengan persentase penduduk miskin paling tinggi. Kabupaten/kota yang memiliki persentase penduduk miskin tertinggi (dalam peta ditunjukkan dengan warna merah tua) adalah sebagai berikut:

1. Provinsi Jawa Barat: Kabupaten Kuningan, Kabupaten Indramayu, dan Kota Tasikmalaya.
2. Provinsi Jawa Tengah: Kabupaten Grobogan, Kabupaten Demak, Kabupaten Purworejo, Kabupaten Cilacap, Kabupaten Sragen, Kabupaten Klaten, Kabupaten Banyumas, Kabupaten Banjarnegara, Kabupaten Pemalang, Kabupaten Rembang, Kabupaten Brebes, Kabupaten Kebumen, Kabupaten Purbalingga, dan Kabupaten Wonosobo.
3. Provinsi Yogyakarta: Kabupaten Bantul, Kabupaten Gunung Kidul, dan Kabupaten Kulon Progo.
4. Provinsi Jawa Timur: Kabupaten Bojonegoro, Kabupaten Lamongan, Kabupaten Bondowoso, Kabupaten Ngawi, Kabupaten Pacitan, Kabupaten Pamekasan, Kabupaten Tuban, Kabupaten Sumenep, Kabupaten Probolinggo, Kabupaten Bangkalan, dan Kabupaten Sampang.

Persentase Penduduk yang Bekerja di Pertanian



Gambar 2. Peta Wilayah Kabupaten/Kota di Pulau Jawa Berdasarkan Persentase Penduduk yang Bekerja di Pertanian, 2017



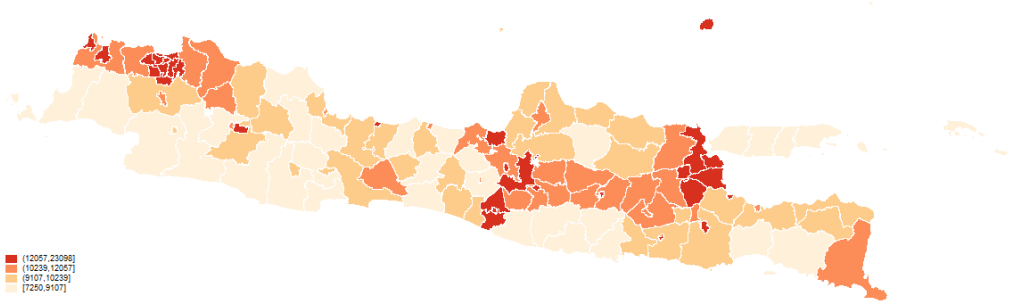
Gambar 3. Peta Wilayah Kabupaten/Kota di Pulau Jawa Berdasarkan Rata-Rata Lama Sekolah, 2017

Penduduk yang bekerja di sektor pertanian di Pulau Jawa ditunjukkan dalam peta pada Gambar 2. Semakin tua warnanya semakin tinggi persentase penduduk yang bekerja di sektor pertanian. Kabupaten dengan persentase penduduk yang bekerja di sektor pertanian pada kelompok tertinggi adalah sebagai berikut:

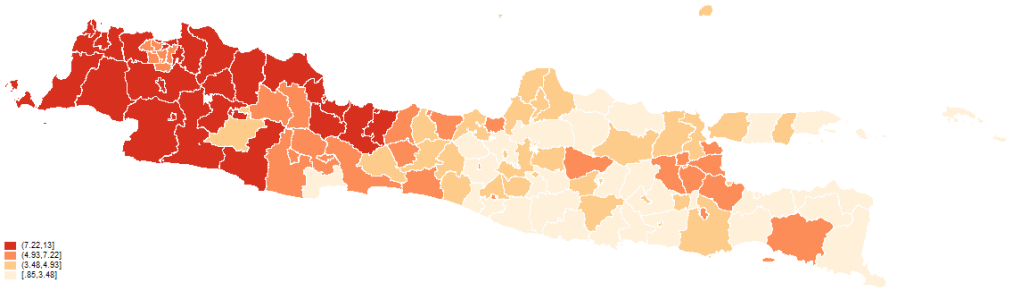
1. Provinsi Jawa Banten: Kabupaten Lebak.
2. Provinsi Jawa Tengah: Kabupaten Wonosobo, Kabupaten Brebes, Kabupaten Wonogiri, Kabupaten Grobogan, Kabupaten Temanggung, dan Kabupaten Blora
3. Provinsi Yogyakarta: Kabupaten Gunung Kidul.
4. Provinsi Jawa Timur: Kabupaten Jember, Kabupaten Tuban, Kabupaten Malang, Kabupaten Magetan, Kabupaten Lamongan, Kabupaten Nganjuk, Kabupaten Lumajang, Kabupaten Probolinggo, Kabupaten Ponorogo, Kabupaten Madiun, Kabupaten Situbondo, Kabupaten Blitar, Kabupaten Bondowoso, Kabupaten Bojonegoro, Kabupaten Trenggalek, Kabupaten Ngawi, Kabupaten Bangkalan, Kabupaten Pamekasan, Kabupaten Pacitan, Kabupaten Sumenep dan Kabupaten Sampang. Bahkan di Kabupaten Bangkalan, Kabupaten Pamekasan, Kabupaten Pacitan, Kabupaten Sumenep dan Kabupaten Sampang lebih dari separuh penduduknya bekerja di sektor pertanian (lihat Gambar 2)

Rata-Rata Lama Sekolah Penduduk di Pulau Jawa

Rata-rata lama sekolah menggambarkan rata-rata tingkat pendidikan yang dicapai oleh penduduk di suatu wilayah. Gradasi warna yang disajikan dalam peta pada Gambar 3 menggambarkan rata-rata lama sekolah di kabupaten/kota, dengan penjelasan semakin gelap warnanya semakin banyak jumlah tahun rata-rata lama sekolah. Jika dicermati dari peta tersebut, rata-rata lama sekolah di Pulau Jawa masih rendah. Kabupaten Sampang memiliki rata-rata lama sekolah terendah (4,12 tahun) dan Kota Tangerang Selatan memiliki rata-rata lama sekolah



Gambar 4. Peta Wilayah Kabupaten/Kota di Pulau Jawa Berdasarkan Pengeluaran per Kapita, 2017



Gambar 5. Peta Wilayah Kabupaten/Kota di Pulau Jawa Berdasarkan Tingkat Pengangguran Terbuka, 2017

tertinggi (11,77 tahun). Kabupaten/Kota dengan warna tua yang artinya memiliki rata-rata lama sekolah lebih tinggi dibanding kabupaten/kota lainnya, adalah sebagai berikut:

1. Provinsi Banten: Kota Tangerang, Kota Cilegon, dan Kota Tangerang Selatan.
2. Provinsi DKI Jakarta: Kota Jakarta Barat, Kota Jakarta Utara, Kota Jakarta Selatan, dan Kota Jakarta Timur.
3. Provinsi Jawa Barat: Kota Sukabumi, Kota Cirebon, Kota Bogor, Kota Bandung, Kota Depok, Kota Bekasi, dan Kota Cimahi.
4. Provinsi Jawa Tengah: Kota Salatiga, Kota Magelang, Kota Surakarta, Kota Semarang,
5. Provinsi Yogyakarta: Kabupaten Bantul dan Kabupaten Sleman
6. Provinsi Jawa Timur: Kota Pasuruan, Kota Blitar, Kota Kediri, Kota Mojokerto, Kota Malang, Kabupaten Sidoarjo, dan Kota Surabaya.

Pengeluaran rumah tangga per kapita

Pengeluaran rumah tangga per kapita menggambarkan rata-rata pengeluaran setiap penduduk di suatu wilayah. Semakin besar pengeluaran per kapita diartikan semakin tinggi tingkat kesejahteraan penduduk. Di Pulau Jawa, rata-rata pengeluran per kapita penduduknya sebesar 11.054 rupiah, artinya setiap penduduk memenuhi kebutuhannya baik makanan maupun non makanan sebesar 11.054 rupiah per bulan. Pengeluaran per kapita terndah di Kabupaten Tasikmalaya (7.250 rupiah), dan yang tertinggi di Kota Jakarta Selatan (23.098 rupiah). Kabupaten/kota dengan pengeluaran per kapita lebih tinggi dari kabupaten/kota lainnya ditunjukkan dalam gambar peta yang berwarna paling tua. Kabupaten/kota tersebut adalah sebagai berikut:

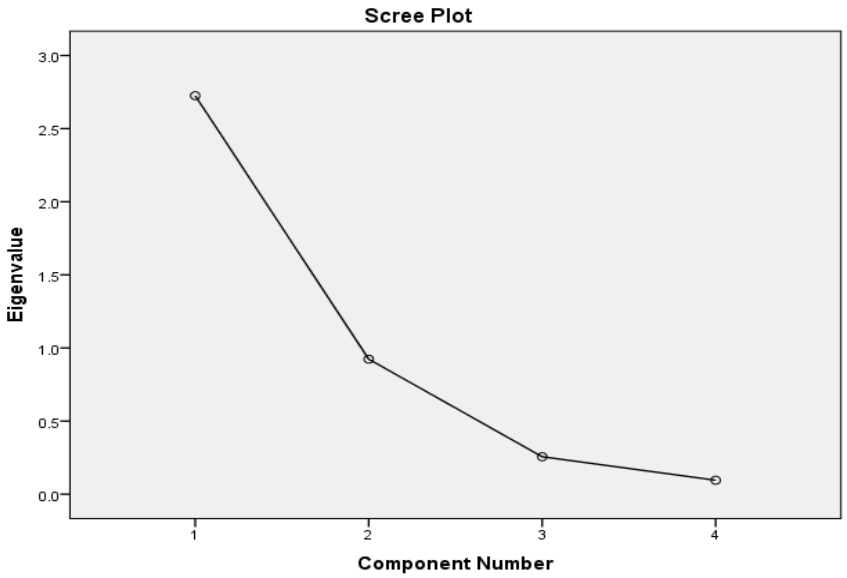
1. Provinsi Banten: Kota Cilegon, Kota Serang, Kota Tangerang, dan Kota Tangerang Selatan.
2. Provinsi DKI Jakarta: Kota Jakarta Pusat, Kota Jakarta Timur, Kota Jakarta Utara, Kota Jakarta Barat, dan Kota Jakarta Selatan.
3. Provinsi Jawa Barat: Kota Depok, Kota Bekasi, dan Kota Bandung.
4. Provinsi Jawa Tengah: Kabupaten Boyolali, Kota Tegal, Kota Surakarta, Kota Semarang, dan Kota Salatiga.
5. Provinsi Yogyakarta: Kabupaten Bantul, Kabupaten Sleman, dan Kota Yogyakarta.

Tabel 2. Hasil Uji KMO dan Bartlett

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.645
Bartlett's Test of Sphericity	Approx. Chi-Square	323.367
	df	6
	Sig.	.000

Tabel 3. *Total Variance Explained*

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.725	68.135	68.135	2.725	68.135	68.135
2	.923	23.086	91.221			
3	.256	6.399	97.620			
4	.095	2.380	100.000			



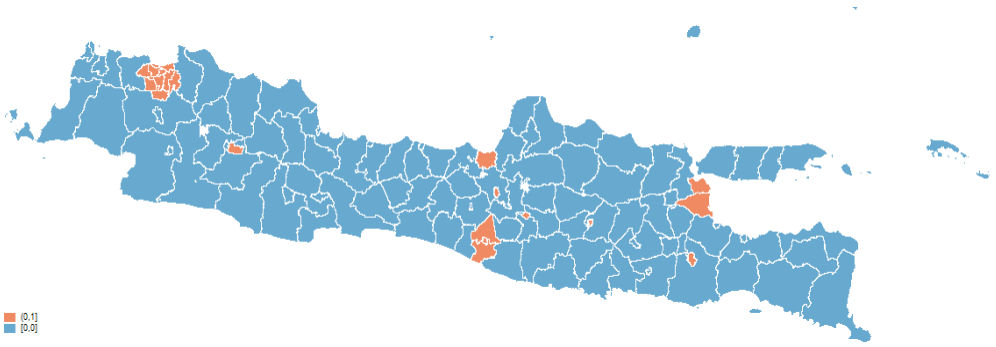
Gambar 6. *Scree Plot*

6. Provinsi Jawa Timur: Kota Batu, Kabupaten Mojokerto, Kabupaten Gresik, Kota Pasuruan, Kota Mojokerto, Kota Blitar, Kabupaten Sidoarjo, Kota Madiun, Kota Malang, dan Kota Surabaya.

Tingkat Pengangguran Terbuka

Menurut BPS, pengangguran adalah persentase jumlah pengangguran terhadap jumlah angkatan kerja. Gambar 5 menggambarkan peta wilayah berdasarkan tingkat pengangguran terbuka. Gradasi warna menjelaskan bahwa semakin tua warna suatu wilayah maka semakin tinggi tingkat pengangguran terbuka di wilayah tersebut. Kabupaten dengan tingkat pengangguran terendah di Pacitan, sedangkan yang tertinggi di Kabupaten Serang. Beberapa kabupaten/kota yang memilki tingkat pengangguran terbuka lebih tinggi dibanding kabupaten/kota lainnya digambarkan dalam peta dengan warna yang paling tua. Kabupaten/kota tersebut adalah:

1. Provinsi Banten: Kabupaten Pandeglang, Kota Serang, Kabupaten Lebak, Kabupaten Tangerang, Kota Cilegon, dan Kabupaten Serang.
2. Provinsi DKI Jakarta: Kota Jakarta Pusat dan Kota Jakarta Utara.
3. Provinsi Jawa Barat: Kabupaten Sukabumi, Kabupaten Garut, Kabupaten Kuningan, Kota Sukabumi, Kota Cimahi, Kota Bandung, Kabupaten Indramayu, Kabupaten Subang, Kabupaten Purwakarta, Kota



Gambar 7. Peta Pengelompokan Wilayah Berdasarkan Faktor-Faktor Kemiskinan

Cirebon, Kota Bekasi, Kabupaten Bandung Barat, Kabupaten Bogor, Kabupaten Karawang, Kota Bogor, Kabupaten Cirebon, Kabupaten Cianjur, Kabupaten Bekasi,

4. Provinsi Jawa Tengah: Kabupaten Tegal, Kabupaten Brebes, dan Kota Tegal.
5. Provinsi Jawa Timur: Kota Malang.

2. Analisis Faktor

Sebelum melakukan pengelompokan terlebih dahulu dilakukan uji KMO dan Bartlett untuk menguji apakah terdapat korelasi yang signifikan antar variabel. Variabel yang memiliki korelasi yang tinggi akan di reduksi.

Dari hasil pengujian KMO dan Bartlett, diperoleh p-value sebesar 0,000 sehingga tolak H0, artinya terdapat korelasi antar variabel. Selanjutnya, dilakukan uji KMO untuk mengetahui kecukupan data untuk analisis faktor. Hasil uji KMO menunjukkan besarnya KMO sebesar 0,645 sehingga nilai KMO > 0,5 yang artinya analisis faktor cukup baik untuk dilakukan (Sharma, 1996). Tujuan dari analisis faktor ini untuk menyederhanakan kumpulan 4 variabel yang digunakan sebagai faktor-faktor kemiskinan (Tabel 2).

Jumlah faktor yang akan dibentuk ditentukan dengan beberapa kriteria agar diperoleh faktor-faktor yang sesuai. Kriteria pertama yang digunakan sebagai penentu jumlah faktor adalah eigen value. Faktor yang memiliki eigen value lebih dari 1 adalah faktor 1, tetapi karena persentase cumulative eigen value yang mencapai lebih dari 90 persen di faktor 2, maka jumlah faktor yang digunakan adalah dua (Tabel 3).

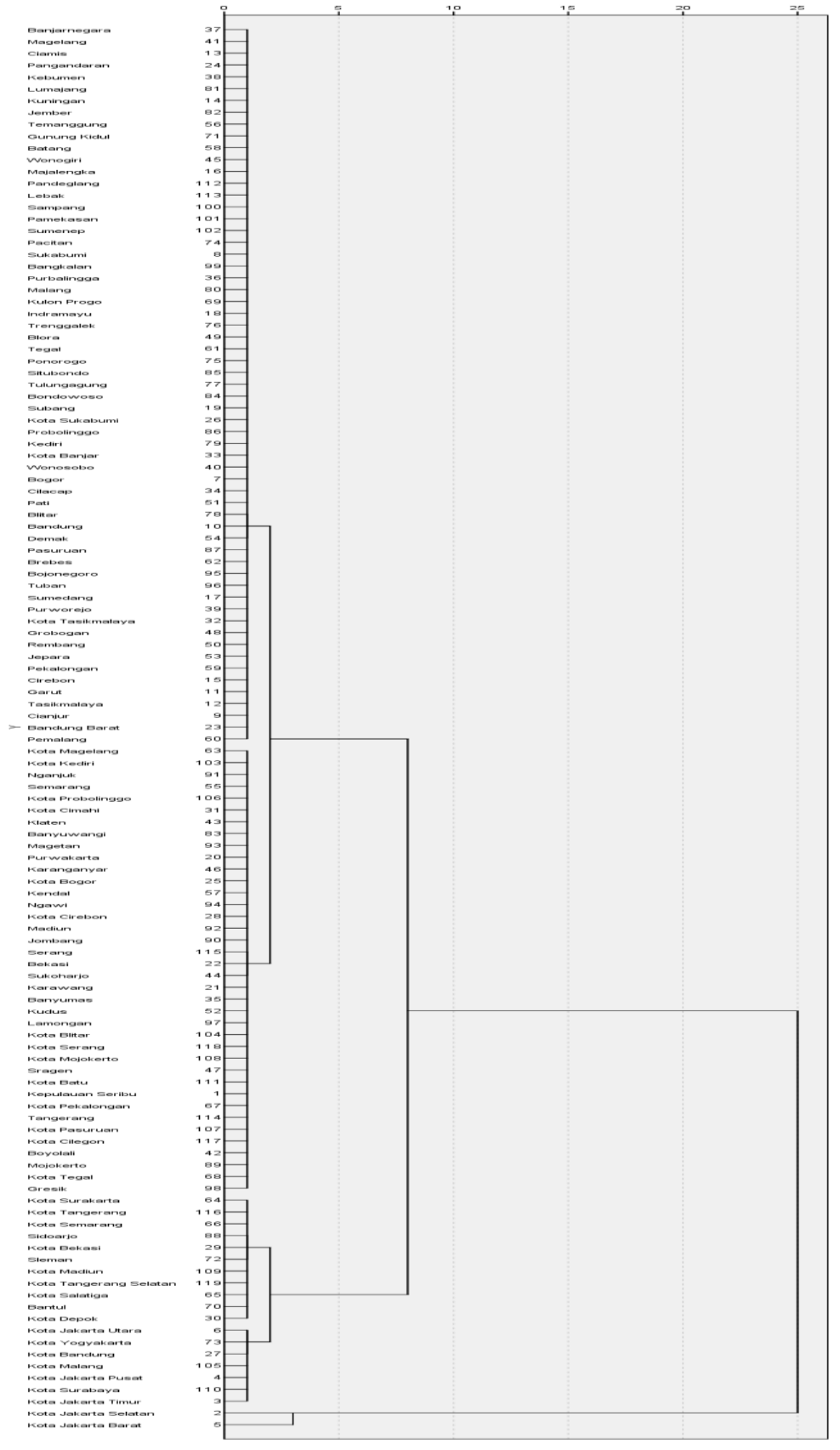
Penentuan jumlah faktor juga bisa menggunakan scree plot. Scree plot merupakan nilai plot eigen value terhadap jumlah faktor yang diekstraksi. Titik pada tempat dimana scree mulai terjadi menunjukkan banyaknya faktor yang sesuai, dimana scree terlihat mulai mendatar. Berdasarkan scree plot pada Gambar 6 dapat disimpulkan bahwa garis mulai mendatar di titik 3 sehingga hanya 2 faktor yang akan digunakan untuk membentuk cluster. Hasil ini sejalan dengan pembentukan faktor menggunakan PCA pada Tabel 3.

3.* Analisis *Cluster

Hasil dari analisis faktor digunakan sebagai input dalam melakukan analisis cluster. Hasil analisis cluster dengan metode *average linkage* membagi kabupaten/kota di Pulau Jawa menjadi 2 kelompok. Penentuan kelompok didasarkan pada Gambar dendogram seperti yang tersaji dalam Gambar 8.

Kelompok I merupakan kelompok kabupaten/kota yang kemiskinannya rendah, terdiri dari:

1. Provinsi Banten: Kota Tangerang dan Kota Tangerang Selatan.
2. Provinsi DKI Jakarta: Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Timur, Kota Jakarta Pusat, dan Kota Jakarta Utara.
3. Provinsi Jawa Barat: Kota Bandung, Kota Depok, dan Kota Bekasi.
4. Provinsi Jawa Tengah: Kota Semarang, Kota Salatiga, dan Kota Surakarta.



Gambar 8. *Dendogram using Average Linkage (Between Group)*

5. Provinsi Yogyakarta: Kota Yogyakarta, Kabupaten Bantul, dan Kabupaten Sleman.
6. Provinsi Jawa Timur: Kota Surabaya, Kota Malang, Kota Madiun, dan Kabupaten Sidoarjo.

Sedangkan kelompok 2 terdiri dari kabupaten/kota yang tingkat kemiskinannya lebih tinggi, yaitu sebanyak 99 kabupaten/kota, terdiri dari:

1. Provinsi Banten: Kabupaten Cilegon, Kabupaten Serang, Kabupaten Tangerang, Kota Serang, Kabupaten Lebak, dan Kabupaten Pandeglang.
2. Provinsi DKI Jakarta: Kabupaten Kepulauan Seribu.
3. Provinsi Jawa Barat: Kabupaten Bekasi, Kota Bogor, Kabupaten Bandung, Kota Cimahi, Kabupaten Sukabumi, Kabupaten Ciamis, Kota Sukabumi, Kabupaten Bogor, Kabupaten Purwakarta, Kota Cirebon, Kabupaten Pangandaran, Kabupaten Karawang, Kabupaten Sumedang, Kabupaten Subang, Kabupaten Tasikmalaya, Kabupaten Garut, Kabupaten Cianjur, Kabupaten Bandung Barat, Kabupaten Majalengka, Kabupaten Cirebon, Kabupaten Kuningan, Kabupaten Indramayu, dan Kota Tasikmalaya.
4. Provinsi Jawa Tengah: Kota Banjar, Kota Pekalongan, Kabupaten Kudus, Kabupaten Semarang, Kota Tegal, Kabupaten Jepara, Kabupaten Sukoharjo, Kota Magelang, Kabupaten Tegal, Kabupaten Batang, Kabupaten Kendal, Kabupaten Pati, Kabupaten Temanggung, Kabupaten Boyolali, Kabupaten Karanganyar, Kabupaten Magelang, Kabupaten Pekalongan, Kabupaten Wonogiri, Kabupaten Blora, Kabupaten Grobogan, Kabupaten Demak, Kabupaten Purworejo, Kabupaten Cilacap, Kabupaten Sragen, Kabupaten Klaten, Kabupaten Banyumas, Kabupaten Banjarnegara, Kabupaten Pemalang, Kabupaten Rembang, Kabupaten Purbalingga, Kabupaten Brebes, Kabupaten Kebumen, dan Kabupaten Wonosobo.
5. Provinsi Yogyakarta: Kabupaten Gunung Kidul, dan Kabupaten Kulon Progo.
6. Provinsi Jawa Timur: Kota Batu, Kota Mojokerto, Kota Pasuruan, Kota Probolinggo, Kota Blitar, Kabupaten Tulungagung, Kota Kediri, Kabupaten Banyuwangi, Kabupaten Blitar, Kabupaten Mojokerto, Kabupaten Pasuruan, Kabupaten Jombang, Kabupaten Magetan, Kabupaten Lumajang, Kabupaten Jember, Kabupaten Malang, Kabupaten Ponorogo, Kabupaten Nganjuk, Kabupaten Kediri, Kabupaten Madiun, Kabupaten Gresik, Kabupaten Trenggalek, Kabupaten Situbondo, Kabupaten Bojonegoro, Kabupaten Lamongan, Kabupaten Bondowoso, Kabupaten Ngawi, Kabupaten Pacitan, Kabupaten Pamekasan, Kabupaten Tuban, Kabupaten Sumenep, Kabupaten Probolinggo, Kabupaten Bangkalan, dan Kabupaten Sampang.

Penyajian kelompok kabupaten/kota hasil clustering tersaji dalam bentuk peta yang ditunjukkan pada Gambar 7.

KESIMPULAN DAN SARAN

Dari hasil pembahasan diperoleh kesimpulan bahwa dengan berdasarkan faktor-faktor kemiskinan, kabupaten/kota di Pulau Jawa dapat dibagi ke dalam dua kelompok. Sebanyak 16,8 persen kabupaten/kota di Pulau Jawa masuk dalam kelompok pertama, dan sisanya sebanyak 83,2 persen masuk ke dalam kelompok kedua. Kelompok pertama terdiri dari 20 kabupaten/kota, yaitu:

1. Provinsi Banten: Kota Tangerang dan Kota Tangerang Selatan.
2. Provinsi DKI Jakarta: Kota Jakarta Barat, Kota Jakarta Selatan, Kota Jakarta Timur, Kota Jakarta Pusat, dan Kota Jakarta Utara.
3. Provinsi Jawa Barat: Kota Bandung, Kota Depok, dan Kota Bekasi.
4. Provinsi Jawa Tengah: Kota Semarang, Kota Salatiga, dan Kota Surakarta.
5. Provinsi Yogyakarta: Kota Yogyakarta, Kabupaten Bantul, dan Kabupaten Sleman.
6. Provinsi Jawa Timur: Kota Surabaya, Kota Malang, Kota Madiun, dan Kabupaten Sidoarjo.

Sedangkan 99 kabupaten/kota lainnya termasuk dalam kelompok kedua.

Kelompok I merupakan kabupaten/kota yang tingkat kemiskinannya rendah, sedangkan kelompok II merupakan kabupaten/kota yang tingkat kemiskinannya tinggi. Saran, penelitian berikutnya dapat meneliti lebih lanjut pengaruh dari masing-masing faktor kemiskinan terhadap tingkat kemiskinan di setiap kelompok.

DAFTAR PUSTAKA

Bachtiar, N., Rasbi, M.J., dan Fahmi, R. (2016). *Analisis Kemiskinan Anak Balita pada Rumah Tangga di Provinsi Sumatera Barat.* Jurnal Kependudukan Indonesia Vol. 11 No. 1 Juni 2016, hal. 29-38.

Barakbah, A.R., dan Arai, K. (2004). Determining Constrains of Moving Variance to Find Global Optimum and Make Automatic *Clustering*. Proc. Industrial Electronics Seminar (IES) 2004, pp.409-413, October 12, 2004, Surabaya, Indonesia

Bappenas. (2014). Penyusunan Rencana Pembangunan Jangka Menengah Nasional 2015–2019. Kementerian Perencanaan Pembangunan Nasional/Badan Perencanaan Pembangunan Nasional. Jakarta.

Badan Pusat Statistik. (2017). *Data dan Informasi Kemiskinan Kabupaten/Kota 2017*. Badan Pusat Statistik. Jakarta. Diakses pada https://www.bps.go.id/

Chandra, N.E., dan Nafisah, Q. (2017). *Analisis Cluster Average Linkage Berdasarkan Faktor-Faktor Kemiskinan di Provinsi Jawa Timur*. Zeta-Math Journal Volume 3 No. 2, November 2017.

Kurniawan M.DP. (2017). *Analisis Faktor-Faktor Penyebab Kemiskinan di Kabupaten Musi Banyuasin (Studi Kasus di Kecamatan Sungai Lilin)*. Jurnal Ilmu Ekonomi Global Masa Kini Volume 8 No. 01 Juli 2017.

Laraswati, T.F., (2014). *Perbandingan Kinerja Metode Complete Linkage, Metode Average Linkage, dan Metode K-Means dalam Menentukan Hasil Analisis Cluster*. Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Negeri Yogayakarta. 2014

Leasiwal, T.C. (2013). *Determinan Kemiskinan dan Karakteristik Kemiskinan di Provinsi Maluku*. Jurnal Ekonomi, Cita Ekonomika, Vol. VII, No. 2, Desember 2013, ISSN: 1978-3612.

Ningsih, S., Wahyuningsih, S., dan Nasution, Y.N. (2016). *Perbandingan Kinerja Metode Complete Linkage dan Average Linkage dalam Menentukan Hasil Analisis Cluster*. Prosiding Seminar Sains dan Teknologi FMIPA Unmul, Vol. 1 No. 1 Juli 2016, Samarinda, Indonesia.

Sharma, S. (1996). Applied Multivariate Techniques, New York: John Wiley & Sons, Inc.

Zuhdiyati, N. dan Kaluge, D. (2017). *Analisis Faktor-Faktor yang Mempengaruhi Kemiskinan di Indonesia Selama Lima Tahun Terakhir (Studi Kasus pada 33 Provinsi)*. JIBEKA, Volume 11 No. 2 Februari 2017: 27-31

ANALISIS KINERJA, KUALITAS DATA, DAN *USABILITY* PADA PENGGUNAAN *CAPI* UNTUK KEGIATAN SENSUS/SURVEY

Takdir

Politeknik Statistika STIS
e-mail: takdir@stis.ac.id

Abstrak

Pengumpulan data merupakan suatu tahapan pada Sensus/Survey yang sangat menentukan keberhasilan Sensus/Survey. Prosesnya yang memakan waktu lama akan mengakibatkan data yang disajikan tidak relevan dengan kondisi pada saat pelaksanaan. Dengan *Computer-Assisted Personal Interview* (*CAPI*), proses entri data dapat dilakukan pada saat proses interview berlangsung. Hal ini mempersingkat tahapan pengumpulan data hingga data tersedia pada sistem komputer dan siap untuk dianalisis. Pada penelitian ini, indikator-indikator penting penentu keberhasilan penerapan *CAPI*, yakni kinerja, kualitas data, dan usability diukur untuk melihat sejauh mana *CAPI* memberikan penyempurnaan pada pengumpulan data. Penelitian ini memberikan rekomendasi, baik dari segi konsep, maupun teknis, mengenai desain *CAPI* untuk kegiatan sensus/survey.

Kata kunci: *CAPI*, sensus, survey, pengumpulan data

Abstract

Data collection is a phase in census/survey phases which highly affect the success of cencus or survey. Using Computer-Assisted Personal Interviewing (CAPI), data entry could be carried out during interview. It could shorten the data collection stage until data were available on a computer system and ready for analysis. In this study, the essential indicators which determine the success of CAPI implementation, i.e. performance, data quality, and usability are mesured to undestand the signifacancy of CAPI in improving data collection. This study proposed recommendation, either in the aspect of concept, or technical regarding CAPI design for census/survey.

Keywords*: CAPI, census, survey, data collection*

PENDAHULUAN

Data yang berkualitas sangat menentukan kebijakan pembangunan Negara dari berbagai arah, baik melalui kebijakan atau keputusan pemerintah secara langsung, maupun rekomendasi dari kegiatan penelitian. Badan Pusat Statistik (BPS) merupakan lembaga Negara yang ditugaskan khusus untuk menyediakan data statistik dasar yang dijadikan acuan oleh berbagai kalangan. Oleh karena itu, BPS dituntut untuk menjamin kualitas data yang dihasilkan.

Kegiatan Sensus dan Survey merupakan kegiatan pokok yang dilakukan oleh BPS. Tahapan pengumpulan data (data collection) merupakan salah satu tahapan pada kegiatan Sensus dan Survey yang harus dilaksanakan dan sangat menentukan keberhasilan pelaksanaan Sensus dan Survey. Tahapan pengumpulan data bertujuan untuk memperoleh data dan informasi dari responden, misalnya dengan melakukan wawancara secara langsung kepada responden. Tahapan ini sangat mempengaruhi kualitas data yang dihasilkan. Sebagai contoh, kesalahan perekaman data (*data entry*) akan mengakibatkan analisis data menghasilkan output yang tidak objektif. Selain itu, proses pengumpulan data yang memakan waktu yang lama akan mengakibatkan data yang nantinya disajikan tidak relevan dengan kondisi pada saat pengumpulan data dilakukan.

Computer-Assisted Personal Interview (*CAPI*) merupakan sebuah terobosan pada tahapan pengumpulan data. Dengan *CAPI*, proses interview dengan responden dan entri data dilakukan secara bersamaan. Hal ini akan mempersingkat tahapan pengumpulan data hingga data tersedia pada sistem komputer. Dengan demikian, dengan penerapan *CAPI* yang tepat, dapat dilakukan efsiensi, baik dari segi biaya, maupun waktu yang dibutuhkan pada tahapan pengumpulan data.

Saat ini teknologi pendukung *CAPI* telah berkembang pesat dan telah banyak diterapkan di berbagai Negara maju, khususnya Amerika, Inggris, Australia, dan Selandia Baru. Indonesia sebagai Negara dengan peringkat 4 jumlah penduduk terbesar di dunia (*CIA World Factbook* 2013) memerlukan solusi untuk memudahkan pengumpulan data agar kegiatan sensus/survey dapat berjalan lebih optimal.

Sekolah Tinggi Ilmu Statistik (STIS) merupakan perguruan tinggi kedinasan yang didirikan oleh BPS untuk memenuhi kebutuhan sumber daya manusia dalam menjalakan kegiatan perstatistikan di BPS. Setiap tahunnya STIS mengadakan kegiatan Praktil Kerja Lapangan (PKL) bagi mahasiswa semester ke-5 sebagai miniatur kegiatan perstatistikan yang dilakukan BPS. Penelitian ini bertujuan untuk mengukur kinerja, kualitas data yang dihasilkan, serta usability (kemudahan penggunaan) pengumpulan dan perekaman data dengan menggunakan *CAPI*. PKL Angkatan 54 STIS yang menggunakan 2 jenis metode/alat pengumpulan data, yakni *PAPI* (*Paper-and-pencil Personal Interview* dan *CAPI*, merupakan objek studi kasus yang akan diteliti. Untuk melengkapi hasil analisis, dilakukan perbandingan antara *PAPI* dan *CAPI* pada variabel-variabel yang dapat diperbandingkan, yakni kinerja dan kualitas data. Hasil penelitian menunjukkan bahwa *CAPI* memiliki potensi untuk diterapkan sebagai alat penumpulan dan perekaman data pada sensus/survey karena memiliki sejumlah kelebihan dari beberapa aspek. Aspek-aspek yang perlu menjadi perhatian utama dalam penerapan *CAPI* juga disajikan pada hasil penelitian ini. Selain itu, penelitian ini memberikan rekomendasi desain *CAPI* yang tepat, baik dari segi hardware maupun software, untuk diterapkan untuk di BPS pada survey yang memiliki kesamaan karakteristik dengan objek studi kasus pada penelitian ini, serta bentuk dukungan yang sesuai untuk diberikan kepada pengguna *CAPI* oleh organisasi.

TINJAUAN REFERENSI

1. Sejarah *CAPI*

Pada tahun Oktober 1988, *Bureau of Census* Amerika membentuk sub komite

yang membidangi *Computer Assisted Survey Information Collection* (*CASIC*) untuk meneliti potensi kemajuan di bidang teknologi untuk keperluan pengumpulan data statistik, transmisi data ke pusat data, dan masalah (*issue*) pada proses implementasinya (Bishop et al. 1990). Komite tersebut melakuan sejumlah studi mengenai teknologi-teknologi pengumpulan data yang memungkinkan untuk digunakan, khususnya *CATI* (*Computer-Assisted Telephone Interview*) dan *CAPI*. *CAPI* merupakan pengembangan dari *CATI* yang sebelumnya telah menjadi standard alat pengumpulan data dalam bidang penelitian (Bishop et al. 1990). Kemunculan metode *CAPI* dikuti dengan berbagai produk teknologi sebagai implementasi dari *CAPI*, seperti *Prepared Data Entry (PDE), Touchtone Data Entry (TDE),* dan *Voice Recognition Entry (VRE)* (Bishop et al. 1990).

Tahun 1989, *Bureau of Census* Amerika menggunakan *CAPI* pada *Current Population Survey (CPS)* (Couper and Geraldine Burt 1989). *UK Labour Force Survey* tahun 1990 merupakan survey berskala besar yang dilakukan *OPCS* (*Office of Population Censuses and Surveys*), yakni kantor statstik pemerintah Inggris, yang pertama kali menggunakan *laptop* untuk wawancara tatap muka (Matheson 1991). Pada sektor komersil, *British Telecom's* juga telah menggunakan *CAPI* untuk survey kepuasan pelanggan pada tahun 1990 (Sainsbury, Ditch, and Hutton 1993). Namun, survey di bidang sosial masih sedikit yang menggunakan *CAPI*. Hal ini disebabkan karena *CAPI* masih tergolong baru dan dianggap belum matang (*mature*), serta membutuhkan biaya awal yang tergolong besar (Sainsbury et al. 1993).

Beberapa *report papers* dan penelitian terbaru, misalnya (Shaw, Nguyen, and Nischan 2011) dan (Caviglia-harris et al. 2012), telah menunjukkan penggunaan dan pengembangan *CAPI* secara intensif. Di STIS, sistem *CAPI* telah digunakan pada kegiatan Praktik Kerja Lapangan (PKL) mahasiswa STIS sejak tahun 2011. Dimulai dengan aplikasi berbasis web yang memiliki kemampuan *offline storage*, hingga dalam bentuk aplikasi *smartphone native* seperti sekarang ini. *CAPI* yang dikembangkan di STIS terus mengalami pengembangan dari tahun ke tahun dan diuji melalui kegiatan PKL. Namun, sayangnya, *CAPI* belum dimanfaatkan secara optimal di Indonesia, khususnya di BPS. Penelitian mengenai *CAPI* di Indonesia juga sangat sedikit sehingga belum ada rujukan yang meyakinkan pihak yang berkepentingan untuk digunakan sebegai pengganti *PAPI*. Hal tersebut terlihat dari minimnya literatur ilmiah maupun laporan yang dapat diakses yang membahas penggunaan *CAPI* dalam melakukan survey. Penelusuran dengan kata kunci terkait *CAPI* dan "Indonesia" pada *search engine* dan repository karya ilmiah online tidak dapat memberikan hasil yang relevan dan pembahasan khusus terkait *CAPI*, begitu pula dengan daftar pustaka serta daftar tulisan ilmiah yang melakukan sitasi terhadap artikel-artikel populer yang membahas *CAPI*, yang juga terdapat pada daftar pustaka tulisan ini.

2. Kelebihan dan Kekurangan *CAPI*

Penerapan *CAPI* dengan tepat akan memberikan dampak positif berupa kualitas data yang lebih baik (*better quality*), durasi yang lebih cepat (*improved speed*), dan biaya operasional yang lebih rendah (*lower cost*) dibandingkan dengan metode *PAPI* (Manners 1990).

Better Quality

1. Adanya fitur *automatic routing* pada kuesioner yang didukung oleh *CAPI* menyebabkan kejadian missing value hanya akan terjadi apabila responden tidak ingin memberikan jawaban, bukan karena kesalahan *interviewer* yang melewatkan pertanyaan (Manners 1990).
2. Pada *CAPI* pengecekan konsistensi dan validitas isian dilakukan secara otomatis, sedangkan pada *PAPI*, hal tersebut dilakukan secara manual yang rentan terhadap kesalahan (Manners 1990).

3. Kalkulasi matematis diikutkan pada saat pencacahan sehingga penghitungan dapat dilakukan dengan komputer yang memberikan hasil akurat (Sainsbury et al. 1993).
4. Kesalahan (*error*) pada saat perekaman data yang diakibatkan oleh program data entri yang terpisah dengan kuesioner pada *PAPI* dapat dihindari (Sainsbury et al. 1993).

Improved Speed

Proses *editing* dokumen dan *data entry* yang membutuhkan alokasi waktu tersendiri pada metode *PAPI* tidak ditemui pada penerapan *CAPI*. Penerapan *CAPI* juga memungkinkan untuk mengirimkan data ke pusat data secara langsung pada saat pencacahan dilakukan sehingga pemrosesan data untuk tahapan selanjutnya dapat segera dilakukan (Manners 1990).

Lower Cost

Penghematan biaya pada *CAPI* dapat dicapai dengan 3 hal (Manners 1990). Pertama, tidak membutuhkan server dan mainframe dalam jumlah yang banyak untuk mendukung infrastruktur pengentrian data. Kedua, biaya yang diperlukan untuk proses *editing* dokumen dan pengentrian data dapat dihindari. Ketiga, kuesioer yang dikonversi ke dalam sistem komputer dapat diakses dan digunakan langsung dengan mudah oleh *interviewer* sehingga mengurangi jumlah tenaga spesialis komputer dan programmer (Manners 1990).

Disamping kelebihan tersebut, *CAPI* juga memiliki kelemahan-kelemahan yang secara umum dapat dijelaskan sebagai berikut (Matheson 1991).

Biaya *Setup*

Diperlukan biaya yang besar untuk investasi awal pada *CAPI*, khususnya untuk pengadaan infrastruktur.

Keterbatasan Device dan Kompleksitas

Keterbatasan device, misalnya dari segi ukuran, yang digunakan pada *CAPI* secara langsung juga memberikan dampak keterbatasan pada metode *CAPI* itu sendiri.

Pertanyaan Terbuka

CAPI memiliki kesulitan untuk menangani pertanyaan terbuka karena membutuhkan *coding* tertentu.

Kualitas Data

Selain memiliki kelebihan dari sisi kualitas data, *CAPI* juga memiliki kelemahan yang dapat mempengaruhi kualitas data. Apabila tedapat pertanyaan yang memiliki validasi yang *strict* (harus diisi) pada *CAPI* namun jawabannya tidak diketahui oleh responden, hal tersebut akan membuat *interviewer* mengisikan jawaban yang tidak sesuai agar dapat melanjutkan ke pertanyaan selanjutnya.

Kesalahan Perekaman Data

Apabila terjadi kesalahan pencacah dalam menginputkan data, sulit untuk menelusuri nilai yang benar untuk memperbaikinya karena dokumen (kuesioner kertas) tidak tersedia.

3. Issue pada Penerapan *CAPI*

Dalam perkembangannya, dengan model dan kebutuhan survey yang beragam, terdapat berbagai *issue* pada penerapan *CAPI* untuk melakukan pengumpulan data (Matheson, 1991). *Issue* tersebut merupakan hal yang perlu dipertimbangkan ketika akan mengimplementasikan *CAPI*.

Concurrent Interviewing

Pada kasus jumlah anggota rumah tangga yang akan dicacah cukup banyak, pencacah memiliki alternatif dengan membacakan pertanyaan cukup sekali dan dijawab bergantian oleh para responden. Perlu alternatif untuk melakukan hal yang sama pada *CAPI*.

Flexibility

Pencacah terkadang harus kembali ke pertanyaan atau blok (kelompok pertanyaan) sebelumnya untuk mengisi menanyakan kembali pertanyaan yang terlewatkan. Desain *CAPI* yang menampilkan pertanyaan satu per satu secara sequensial dapat menyulitkan melakukan hal ini. Oleh karena itu, desain yang baik perlu mengantisipasi hal ini.

Data Quality

Automatic routing (mengarahkan pertanyaan secara otomatis) merupakan salah satu fitur *CAPI* untuk meningkatkan kualitas data. Namun, fitur ini juga dapat berdampak negatif. Misalnya ketika pencacah salah melakukan input, maka akan diarahkan ke pertanyaan yang salah pula. Penggunaan fitur ini perlu memperhatikan kasus tersebut yang mungkin terjadi.

Diary Processing

Perlu dipertimbangkan untuk disediakan catatan tersendiri pada saat pencacahan dengan *CAPI* yang terpisah dengan kuesioner untuk mencatat hal-hal yang tidak dapat ditangani dengan mudah oleh *CAPI*.

Respondent/Interviewer Acceptability

Perlu diteliti lebih lanjut apakah responden bersedia datanya, termasuk data pribadi dan sensitif, dientrikan langsung ke sistem komputer. Kesediaan dan kemampuan pencacah untuk menggunakan *device* pendukung *CAPI* juga harus menjadi pertimbangan.

Timetable

Susunan jadwal kegiatan sensus/survey juga perlu didesain sedemikian rupa menyesuaikan dengan *CAPI*. Sistem komputer mengharuskan jadwal yang pasti dan setiap tahapan harus dijabarkan dengan detail.

4. Indikator Kinerja *Interviewer* pada *CAPI*

Pada *PAPI*, pengukuran kinerja *interviewer* dapat berupa variabel response rates, accuracy rates, dan production rates (Couper and Geraldine Burt 1989). Namun, *CAPI* membutuhkan indikator yang berda untuk mengukur kinerja *interviewer* karena beberapa indikator yang dipengaruhi oleh keterbatasan *interviewer* dapat ditangani oleh sistem komputer. Couper dan Geraldine merupakan peneliti yang pertamakali mengusulkan 3 indikator yang dapat digunakan untuk mengukur kinerja *interviewer* pada *CAPI* sebagai berikut (Couper and Geraldine Burt 1989).

1. Drop-out Rates, yaitu mencatat jumlah kasus di mana *interviewer* secara sepihak memutuskan berhenti untuk melakukan pencacahan. Indikator ini bertujuan untuk melihat sikap *interviewer* dalam menghadapi teknologi terkomputerisasi.
2. Data Quality Indicators, yaitu jumlah non-response dan penolakan oleh responden terhadap *interviewer*.
3. Self-reports of difficulties with *CAPI*, yaitu berdasarkan laporan kesulitan yang dihadapi oleh *interviewer* dalam menggunakan *CAPI*. Kesulitan dapat berupa aspek hardware, software, penanganan kasus khusus, dan jaringan komunikasi.

5. Durasi *Interview* pada *CAPI* dan *PAPI*

Durasi interview merupakan salah satu pertimbangan penting untuk menerapkan *CAPI*. Beberapa penelitian telah dilakukan untuk membandingkan durasi interview pada metode *PAPI* dan *CAPI*. Penelitan-penelitian tersebut memberikan hasil yang berbeda-beda. Beberapa diantara memberikan hasil bahwa *PAPI* memiliki durasi yang lebih lama (Baker, 1992; Baker et al, 1994; Lynn and Purdon, 1994), dan ada pula yang memberikan hasil yang sebaliknya (Martin and collegues, 1993; Muller and Kesselmann, 1996). Hasil yag komprehensif ditunjukkan pada penelitian Fuch (Fuchs, Couper, and Hansen 2000) dengan mengidentifikasi faktor-faktor yang mempengaruhi durasi interview pada metode *PAPI* dan *CAPI*. Terdapat 4 poin penting yang menyebabkan perbedaan durasi interview antara *PAPI* dan *CAPI* (Fuchs et al. 2000), yaitu:

Loop Design

Loop design pada umumnya diterapkan pada *CAPI* di mana responden diinterview satu per satu. Tiap responden harus menyelesaikan sebuah kuesioner sebelum menanyakan pertanyaan ke responden lainnya. Hal yang berbeda bisa dilakukan pada *PAPI* untuk rumah tangga

yang memiliki banyak jumlah anggota rumah tangga di mana setiap anggota rumah tangga diinterview secara bersamaan. Loop design berkaitan degan concurrent interviewing pada pembahasan sebelumnya.

Character Input and Banked Screens

Proses input data pada *CAPI* mengharuskan *interviewer* mengentrikan data sesuai dengan logic kuesioner *CAPI*. Hal ini berbeda dengan *PAPI* yang memungkinkan *interviewer* lebih bebas menuliskan data. Misalnya dalam kasus *interviewer* diharuskan menginputkan nama depan (first name) dan nama belakang (last name) pada *CAPI* yang membutuhkan waktu bagi *interviewer* untuk menentukan kedua fields tersebut.

Automated Calculations and Fills

Salah satu kelebihan *CAPI* adalah, *interviewer* dapat melakukan perhitungan yang rumit dengan memanfaatkan device yang dibawa secara otomatis, misalnya menghitung umur berdasarkan tanggal lahir yang diperoleh.

"Real" Comparison

Interviewer terkadang membacakan list/daftar anggota rumah tangga untuk meakukan konfirmasi dan memastikan tidak ada anggota rumah tangga yang tidak tercatat. Hal ini juga mempengaruhi perbedaan durasi waktu interview antara *PAPI* dan *CAPI*.

6. Kualitas Data

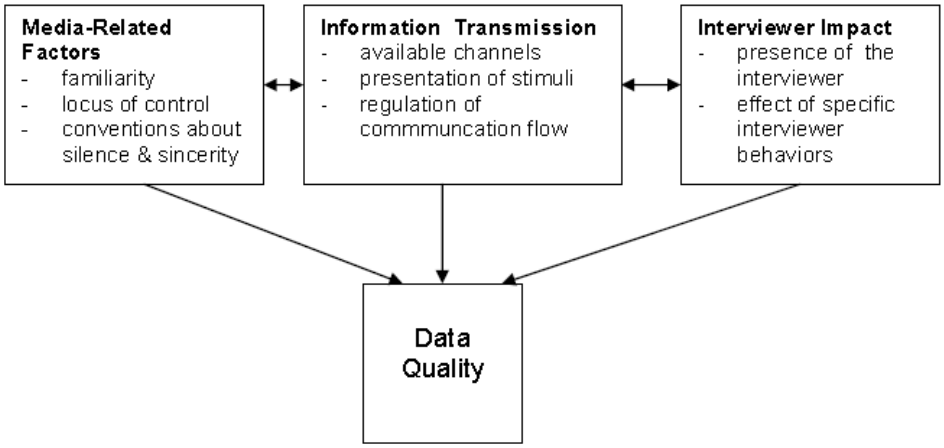
Diperlukan dasar yang kuat untuk megukur kualitas data yang dihasilkan pada *CAPI*. Ukuran yang digunakan sebisa mungkin tidak dipengaruhi oleh faktor diluar pengaruh penggunaan *CAPI* itu sendiri. Model yang dikembangkan De Leeuw (De Leeuw, 1992) mengenai efek pengumpulan data terhadap kualitas data merupakan model yang banyak drujuk untuk mengukur kualitas data yang dihasilkan dengan menerapkan *CAPI*.

Dalam peneitian yang lain (Sainsbury, Ditch, and Hutton 1995) yang membahasa model De Leeuw, dinyatakan pula bahwa terdapat 3 faktor pada *CAPI*, yaitu faktor teknologi/program, kehadiran (*presence*) perangkat komputer, dan efek penggunaan *CAPI* terhadap situasi pada saat interview seperti pada **Error! Reference source not found.**.

7. Spesifikasi Teknis *Software* dan *Hardware CAPI*

CAPI merupakan penerapaan teknologi komputer untuk memudahkan proses pengumpulan data pada Survei/Sensus. Oleh karena itu, spesifikasi teknis, seperti ukuran dan berat perangkat, jenis dan mekanisme pengentrian data, ketahanan baterai, jenis dna resolusi monitor, serta pemilihan software yang digunakan perlu ditentukan dengan tepat. Berikut adalah beberapa penelitian terkait yang mengusulkan spesifikasi teknis untuk penerapan *CAPI* (Caviglia-harris et al. 2012).

1. Couper and Groves (1992) menyimpulkan bahwa berat *hardware* yang digunakan merupakan faktor terpenting bagi *interviewer*. Dari pengujian menggunakan beberapa jenis komputer, mereka menemukan bahwa ukuran berat yang nyaman untuk dibawa adalah 7-8 *pounds* (kurang lebih 3-4 kilogram), sedangkan untuk pencacahan dengan keadaan berdiri hanya seberat 3 *pounds* (kurang lebih 1,4 kilogram).
2. Studi lain (Baker et al, 1995) menyebutkan bahwa kesulitan menginput data pada desain *CAPI* yang hanya menyertakan satu atau sedikit pertanyaan dalam satu kali tampilan di monitor, dan kesulitan membaca monitor di perangkat pada kondisi pencahayaan yang tidak baik merupakan 2 faktor yang menyebabkan durasi interview dengan *CAPI* lebih lama daripada *PAPI*.
3. Penelitian lain meghasilkan *CAPI* memberikan durasi interview yang lebih cepat dibandingkan dengan *PAPI* ketika interface dan desain survey ditetapkan dengan baik. Hal tersebut meliputi automatic skip, perhitungan aritmatika, dan desain survey yang kompleks (Couper 2000).



Gambar 1. *De Leeuw's Conceptual Model of Data Collection Effects on Data Quality* (Randolph et al. 2006)

4. *PDA (Personal Digital Assistance)* merupakan *device* yang paling sering dipilih untuk *CAPI* karena peprtimbangan berat, ukuran, dan biaya (Bernabe-Ortiz et al. 2008).
5. Untuk survey dengan desain yang kompleks, ukuran layar yang lebih besar (*laptop*) memberikan keuntungan yang signifikan dibadingkan dengan *PDA* (Childs and Landreth 2006).

Berdasarkan hasil penelusuran penulis terhadap sejumlah aplikasi *CAPI* yang tersedia baik secara gratis maupun komersial, diperoleh sejumlah produk yang telah populer dan banyak digunakan oleh berbagai kalangan, baik organisasi swasta, pemerintah, maupun peneliti. Diantaranya adalah *BLAISE* yang di-*develop* oleh *Statistics Netherland*, *CSPro* yang di-*develop* oleh *United States Census Bureau*, Survey Solutions yang di-*develop* oleh *World Bank*, *OpenDataKit* oleh *University of Washington's Department of Computer Science and Engineering*, dan *KoBoToolbox* oleh *Harvard Humanitarian Initiative*. *OpenDataKit* dan *KoBoToolbox* bersifat *opensource*, *CSPro* bersifat *freeware*, sedangkan software lainnya memiliki model lisensi komersil yang beragam.

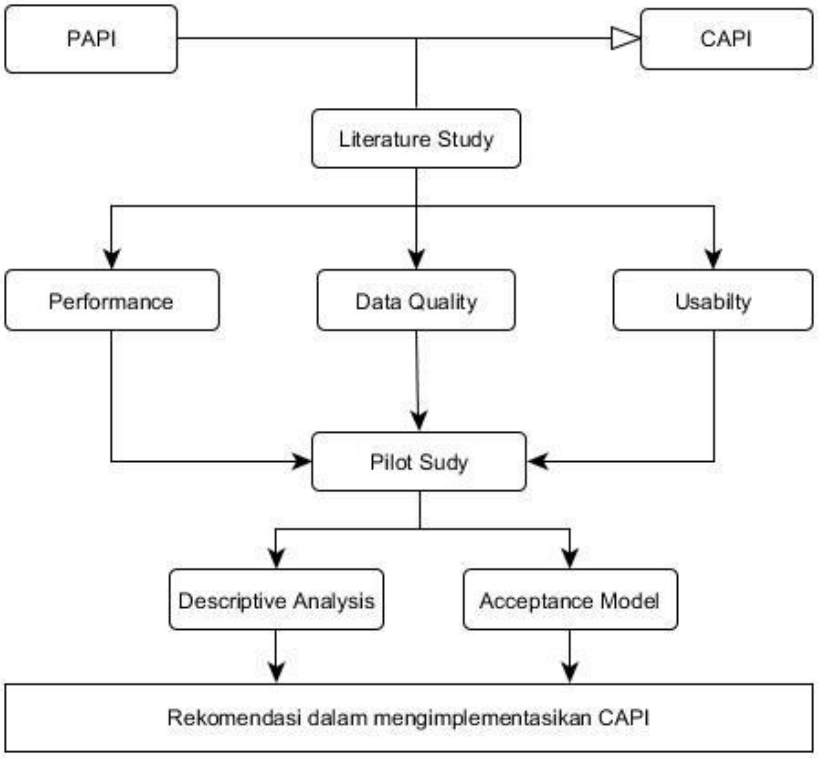
8. Usablity

Untuk melihat tingkat kegunaan dan kenyamanan pengguna (*interviewer*) dalam melakukan pengumpulan data *dengan CAPI*, diperlukan pengukuran kepuasan pengguna terhadap desain *CAPI* yang dibuat. Terdapat berbagai metode yang dapat digunakan untuk melakukan pengukuran tersebut. Salah satu *tools* yang sudah *mature* dan banyak digunakan adalah *QUIS (Questionnaire for User Interaction Satisfaction)* (Slaughter, Harper, and Norman 1994). *QUIS* digunakan untuk assessment kepuasan pengguna secara subjektif dengan aspek yang spesifik, yakni dari segi antar-muka (*interface*).

9. Kerangka Pikir

Dalam mengkaji penerapan *CAPI* pada survey, sejumlah aspek yang berpengaruh menjadi perhatian dalam penelitian ini. Kerangka pikir yang menjadi acuan dalam penelitian ini dapat dilihat pada Gambar 2 berikut.

Untuk melihat aspek yang perlu diperhatikan dalam melakukan transisi dari survey berbasis *PAPI* ke *CAPI*, dilakukan studi literatur dengan tema sumber literatur berupa kajian dan hasil penerapan *CAPI* dari berbagai negara, perusahaan, dan *NSO (National Statistics Office)* dalam kurun waktu 15 tahun terakhir. Dari studi literatur diperoleh sejumlah variabel yang dapat dikategorikan menjadi 3 jenis, yakni *performance*, data *quality*, dan *usability*. Ketiga variabel tersebut kemudian menjadi acuan untuk melakukan evaluasi pada *pilot study* yang dilakukan melalui kegiatan PKL



Gambar 2. Kerangka Pikir Penelitian

54 dan 55 STIS. Data yang dihasilkan dari evaluasi diolah dengan analisis deskriptif dan *Technology Acceptance Model (TAM)* untuk menghasilkan rekomendasi, khususnya untuk BPS, dalam mengimplementasikan *CAPI*. Analisis dengan *TAM* diluar pembahasan *paper* ini.

METODOLOGI

1. Objek dan Metode Penelitian

Penelitian ini merupakan pilot study dengan objek studi kasus Praktik Kerja Lapangan (PKL) Angkatan 54 dan 55 Sekolah Tinggi Ilmu Statistik. PKL 54 dan 55 STIS menggunakan 2 metode pencacahan yaitu pencacahan menggunakan kuesioner kertas (*PAPI*) dan menggunakan kuesioner elektronik (*CAPI*). Pada PKL 54 jumlah *interviewer* yang menggunakan *CAPI* sebanyak 108 orang atau 26,21 persen dari jumlah *interviewer* pada PKL 54, sedangkan pada PKL 55 sebanyak 228 orang atau 49 persen dari jumlah *interviewer* pada PKL 55. Adapun jumlah sampel yang dicacah dengan *CAPI* pada PKL 54 adalah sebanyak 1.755 responden atau 21,49 persen dari jumlah sampel, sedangkan pada PKL 55 sebanyak 3.406 responden atau 60 persen dari jumlah sampel.

Perangkat pendukung pencacahan *CAPI* berupa tablet/smartphone berbasis Android® yang disesuaikan dengan kebutuhan aplikasi untuk pencacahan. Setiap tim pencacah menerima empat buah tablet, tiga tablet digunakan oleh Petugas Cacah Lapangan (PCL) dan satu tablet untuk koordinator tim yang berfungsi sebagai perangkat cadangan dan perangkat pendukung monitoring.

Pada PKL 54, aplikasi *CAPI* yang digunakan masih bersifat statis, yaitu aplikasi didesain untuk tujuan survey yang spesifik pada PKL tersebut saja, sehingga untuk diterapkan pada survey lain atau PKL selanjutnya harus dilakukan perubahan kode program secara menyeluruh (*hardcode*). *CAPI* yang digunakan pada PKL 55 telah mengadopsi sistem kuesioner dinamis, daimana aplikasi dikembangkan berbasis software *opensorce OpenDataKit*. Dengan demikian, perubahan kuesioner dapat dilakukan dengan cepat tanpa harus mengubah kode sumber dari aplikasi *CAPI*. Selain itu, inovasi berupa proses *listing*

berbasis *CAPI* juga diterapkan pada PKL 55 dengan mengembangkan modul listing pada aplikasi *CAPI*. *Frame* hasil *listing* kemudian akan menghasilkan sampel responden terpilih secara otomatis dengan menambahkan fitur penarikan sampel otomatis pada sisi server. Sampel yang terpilih akan didistribusikan ke *device* pencacah berupa kuesioner elektronik yang siap digunakan untuk mencacah responden.

2. Variabel yang Diteliti

Terdapat sejumlah variabel yang mempengaruhi kualitas *CAPI* untuk diterapkan sebagai tools pengumpulan data. Pada penelitian ini, penulis mengategorikan variabel-variabel yang diteliti menjadi 3 kategori, yakni *performance*, *data quality*, dan *usability*.

Performance

Dalam dunia teknologi informasi, *performance* memiliki cakupan yang luas. Namun, untuk memudahkan analisis dan pendalaman masalah, penulis menetapkan beberapa variabel yang termasuk dalam kategori *performance* yang diperoleh dari studi literatur dengan penyesuaian terhadap kondisi studi kasus, yaitu durasi pencacahan hingga raw data siap untuk dianalisis, keluhan/laporan kerusakan, serta kinerja sistem yang meliputi *software* dan *hardware*, baik pada sisi *client*, maupun *server*.

Data Quality

Ukuran kualitas data mengacu pada hal-hal yang menyebabkan data yang dikumpulkan tidak valid atau memiliki anomali sehingga tidak merepresentasikan kenyataan sebenarnya. Untuk mengukur kualitas data yang bersifat laten/abstrak secara lengkap perlu memperhatikan berbagai aspek, sehingga tidak mudah untuk menarik kesimpulan absolut mengenai kualitas data. Untuk itu, perlu dibatasi variabel yang akan dipantau yang dapat dijadikan representasi terbaik untuk mewakili kualitas data. Pada penelitian ini, kualitas data diwakili oleh beberapa variable yang telah diterapkan pada *CAPI* oleh peneliti sebelumnya, yaitu inkonsistensi/kesalahan konsep dan definisi, nilai tidak valid, kesalahan entri, serta missing value, dengan penyesuaian terhadap kondisi studi kasus. Penilaian terhadap variabel-variabel tersebut mengasumsikan bahwa faktor penyebab selain akibat implementasi *CAPI* diabaikan. Oleh karena itu, dalam merepresentasikan hasil penelitian, perlu memahami asumsi tersebut.

Usability

Variabel yang digunakan dalam pengkuran ini berkaitan dengan kemudahan pengguna dalam menggunakan sistem *CAPI* untuk pengumpulan data. Aspek *user interface (UI)* dan *user experience (UX)* sangat menentukan pada *usability*. Komponen-komponen visual yang disajikan oleh aplikasi *CAPI*, seperti kesesuaian dan konsistensi tombol, tulisan, dan warna, merupakan hal yang dinilai pada aspek UI, sedangkan aspek UX berkaitan dengan kesan atau hal yang dirasakan oleh pengguna secara emosional dalam berinteraksi dengan aplikasi *CAPI*, seperti reaksi, antusiasme, serta ketertarikan dalam menggunakan aplikasi *CAPI*.

3. Pengumpulan Data

Data mengenai *CAPI* yang dibutuhkan untuk analisis dikumpulkan dengan tiga jenis pendekatan. Pertama, data empiris mengenai durasi pengisian kuesioner diperoleh dari *log* (catatan) khusus yang di-generate oleh aplikasi *CAPI*. Kedua, data mengenai jumlah non-response, total responden yang dicacah, serta kesalahan pemasukan/entri data diperoleh dari raw data yang dihaslkan oleh aplikasi *CAPI*. Data mengenai persepsi pengguna (*interviewer*) dikumpulkan dengan cara melakukan pencacahan lengkap (sensus) kepada pengguna *CAPI* (*self enumeration*), baik mengenai laporan kerusakan dan komplain, maupun kepuasan terhadap *user interface* dan *user experience* (*QUIS*). Sedangkan data yang berkaitan dengan *PAPI* diperoleh dengan pencatatan manual, baik berupa durasi pencacahan, durasi batching document, serta durasi pengentrian data.

Waktu pengumpulan data bervariasi sesuai dengan data yang dikumpulkan. Data mengenai kerusakan dan komplain petugas pencacahan dilaporkan setiap hari setelah melakukan kegiatan pencacahan di lapangan. Pada tahapan batching, *editing*, dan coding kuesioner, serta pengentrian data juga dilakukan pencatatan. Statistik dari raw data yang dihasilkan pada saat tabulasi juga dikumpulkan untuk melihat kualitas data. Pengukuran *QUIS* dilakukan dengan menyebarkan kuesioner online setelah seluruh kegiatan pencacaan di lapangan selesai dilaksanakan.

Pertanyaan yang harus dijawab oleh *interviewer* pada *QUIS* terdiri dari 7 kategori, yaitu:

1. Tanggapan umum terhadap kinerja sistem *CAPI*
2. Tampilan Layar Monitor
3. Penggunaan Istilah dan Informasi pada Aplikasi
4. Kemudahan Mempelajari Aplikasi
5. Kinerja Sistem
6. Panduan Penggunaan
7. Saran Terkait *Hardware* dan *Software*

Petanyaan untuk kategori 1 sampai dengan 6 berbentuk skala likert dengan nilai berupa 1 (respon negatif) hingga 9 (respon positif). Skala *likert* merupakan skala yang digunakan untuk mengukur sikap, pendapat, dan persepsi seseorang.

4. Metode Analisis

Data yang dikumpulkan dari berbagai sumber diolah, divalidasi, dan disajikan secara deskriptif untuk menggambarkan kondisi variabel yang diteliti. Penyajian dititik beratkan pada niai-nilai yang membutuhkan perhatian atau berbeda dari nilai rata-rata, misalnya hal-hal yang mengurangi performa dan kualitas data pada penerapan *CAPI*.

HASIL DAN PEMBAHASAN

1. *Performance*

Dari proses pencatatan durasi interview, baik pada *PAPI* maupun *CAPI* diperoleh hasil rata-rata durasi interview setiap responden dengan menggunakan *PAPI* pada PKL 54 adalah 1819,749 detik, sedangkan dengan menggunakan *CAPI* adalah 1531,229 detik. Adapun waktu yang diperlukan mulai dari pencacahan lapangan hingga menghasilkan *raw data* yang siap disajikan/ditabulasikan untuk keperluan analisis ditunjukkan pada **Error! Reference source not found.** berikut.

Statistik diatas menunjukkan bahwa *CAPI* memberikan dampak yang efek yang signifikan terhadap durasi survey, khususnya pada pencacahan dan pengolahan data. Efek terbesar terdapat pada proses *Batching, Editing*, dan *Coding* (*BEC*), di mana *CAPI* dapat menghemat waktu selama 16 hari. Hal ini memberikan dampak positif dari segi durasi pelaksanaan survey, namun dapat memberikan dampak negative terhadap kualitas isian kuesioner karena proses *BEC* tidak dilakukan pada *CAPI*.

Koordinator tim (kortim) memimpin 2 hingga 3 orang *interviewer* dalam kegiatan pengumpulan data survey. Konfirmasi ke kortim bertujuan untuk memeriksa kepastian isian kuesioner yang tidak valid, anomali, atau terdapat kuesioner yang belum terkirim ke *server*. Proses tersebut ditindaklanjuti dengan *database cleaning* untuk memperbaiki data yang terkoreksi. Total waktu 5 hari untuk kedua proses tersebut dapat mengganggu kualitas *CAPI*, di mana data sedapat mungkin dikoreksi pada saat pencacahan berlangsung atau dalam rentang waktu yang seminimal mungkin dengan pencacahan. Pada PKL 55 diterapkan mekanisme notifikasi di mana setiap pencacah akan menerima pesan untuk memeriksa data yang anomali. Pesan tersebut dibuat oleh Kortim dan dikirimkan ke pencacah yang bersangkutan secara *real time* (*near real time*) dengan meggunakan fasilitas yang disediakan oleh sistem *CAPI*. Dengan demikian, masalah terkait konfirmasi anomali data dapat terselesaikan dalam rentang waktu yang kecil dari proses interview.

Laporan kerusakan/keluhan dari *interviewer* terkait performa sistem *CAPI* yang tercatat pada PKL 54 untuk permasalahan software diantaranya adalah masih seringnya terjadi *error* dan isian Blok

Tabel 1. Waktu yang diperlukan dari proses pencacahan hingga siap ditabulasikan pada PKL 54

Rincian	*CAPI*	PAPI
Waktu pencacahan	7 hari	7 hari
Batching, Editing, Coding	-	16 hari
Entri Data	-	3 hari
Konfirmasi ke koordinator tim	3 hari	-
Database cleaning	2 hari	-
Total Waktu	**12 hari**	**26 hari**

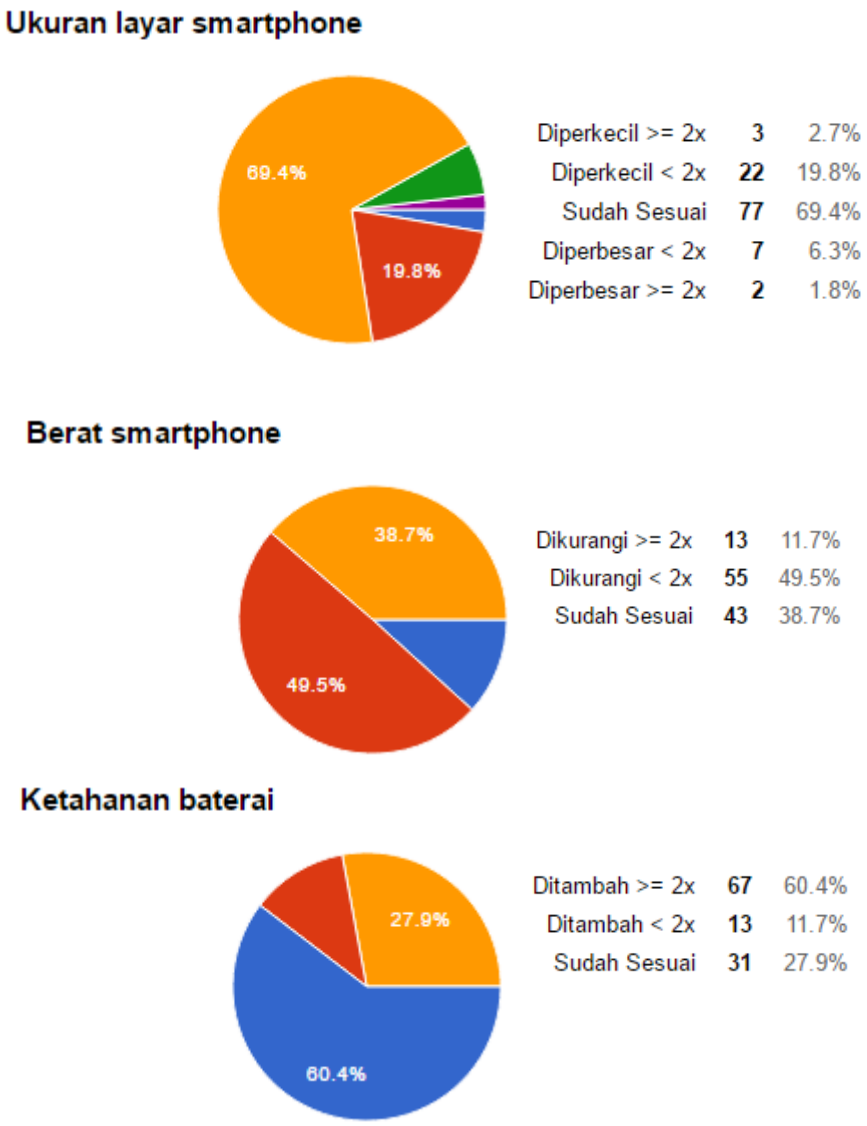
I, yang merupakan kelompok isian identitas responden, yang tidak dapat diedit. Error yang terjadi berupa *infinite loop* dan *force close*. Sedangkan untuk permasalahan hardware, terdapat sejumlah laporan terkait gangguan tablet PC yang tidak terdeteksi penyebabnya yang menyebabkan tablet PC tersebut tidak dapat bekerja secara normal. Gangguan seperti ini kemungkinan diakibatkan oleh sistem operasi, atau perangkat keras yang mengalami kegagalan fungsi. Permasalahan lain yang dilaporkan adalah kurang sensitifnya *touch screen* hardware yang digunakan yang menyulitkan *interviewer* mengentrikan jawaban responden.

Pada PKL 55, terdapat 25 kasus pengiriman kuesioner yang mengalami software *crash* (*force close*) yang menyebabkan kuesioner gagal terkirim, atau sekitar 0,51% dari total kuesioner yang dicacah dengan *CAPI*. Kegagalan pengiriman kuesioner yang diakibatkan oleh jaringan internet sebanyak 42 kasus atau sekitar 0,86%, sedangkan kegagalan akibat *error* pada sisi server sebanyak 24 kasus, atau sekitar 0,5%. Kasus tidak terdapatnya jaringan internet dapat diatasi dengan penyimpanan *offline* yang disediakan oleh aplikasi *CAPI* untuk kemudian dikirimkan ke server apabila sudah terkoneksi ke internet/server. Sedangkan untuk kasus kegagalan akibat *error* pada sisi server pada PKL 55 merupakan masalah yang diakibatkan oleh adanya perbaikan infrastruktur *server* yang digunakan pada saat pencacahan sedang berlangsung. *Server CAPI* yang digunakan di-*hosting* pada kampus STIS. Dari hasil tersebut, terlihat bahwa masalah yang timbul tidak memberikan dampak yang masif secara kuantitas terhadap pelaksanaan *CAPI*.

Selain mencatat permasalahan yang timbul, saran terkait hardware dari *interviewer* juga dicatat yang disajikan pada **Error! Reference source not found.** berikut ini.

Dari hasil tersebut, sebanyak 77 *interviewer*, atau 69,4%, menyatakan bahwa ukuran layar tablet yang digunakan, yakni 10.1 *inch*, sudah sesuai, dan 19.8% menyatakan ukuran layar perlu diperkecil namun tidak sampai diperkecil dua kali lipat. Berdasarkan hasil tersebut, untuk pelaksanaan survey sejenis, di mana rata-rata *interviewer* melakukan pencacahan dalam keadaan duduk, ukuran layar tersebut masih memadai atau dapat diperkecil lagi menjadi sekitar 8 *inch* (ukuran layar standard *smartphone* yang tersedia di pasaran). Pada aspek berat smartphone, sebagian besar *interviewer* merasa *smartphone*, yakni 560 *gram*, terlalu berat untuk dibawa sehingga perlu dikurangi. Dari segi ketahanan baterai diperlukan penambahan kapasitas sebesar lebih dari dua kali lipat dari kapasitas *tablet PC* yang digunakan pada PKL 54 dan 55, yaitu 3170 *mAh* dengan daya tahan 11 hingga 12 jam untuk pemakaian normal.

Saran terkait software pada umumnya terkait dengan tampilan yang akan dibahas pada sub bab *usability*. Namun, untuk menghindari kegagalan software yang diakibatkan oleh sistem operasi, penulis menyarankan agar sistem operasi diupgrade ke versi tebaru yang didukung oleh hardware yang akan digunakan, dan perlu dilakukan instalasi ulang untuk seluruh *device* agar *environment* sistem operasi yang digunakan seragam.



Gambar 3. Hasil survey persepsi interviewer PKL 54 terkait hardware

Dari hasil tersebut, sebanyak 77 *interviewer*, atau 69,4%, menyatakan bahwa ukuran layar tablet yang digunakan, yakni 10.1 *inch*, sudah sesuai, dan 19.8% menyatakan ukuran layar perlu diperkecil namun tidak sampai diperkecil dua kali lipat. Berdasarkan hasil tersebut, untuk pelaksanaan survey sejenis, di mana rata-rata *interviewer* melakukan pencacahan dalam keadaan duduk, ukuran layar tersebut masih memadai atau dapat diperkecil lagi menjadi sekitar 8 *inch* (ukuran layar standard *smartphone* yang tersedia di pasaran). Pada aspek berat smartphone, sebagian besar *interviewer* merasa *smartphone*, yakni 560 *gram*, terlalu berat untuk dibawa sehingga perlu dikurangi. Dari segi ketahanan baterai diperlukan penambahan kapasitas sebesar lebih dari dua kali lipat dari kapasitas *tablet PC* yang digunakan pada PKL 54 dan 55, yaitu 3170 *mAh* dengan daya tahan 11 hingga 12 jam untuk pemakaian normal.

Saran terkait software pada umumnya terkait dengan tampilan yang akan dibahas pada sub bab *usability*. Namun, untuk menghindari kegagalan software yang diakibatkan oleh sistem operasi, penulis menyarankan agar sistem operasi diupgrade ke versi tebaru yang didukung oleh hardware yang akan digunakan, dan perlu dilakukan instalasi ulang untuk seluruh *device* agar *environment* sistem operasi yang digunakan seragam.

2. *Data Quality*

Hasil pemantauan variabel kualitas data pada PKL 54 menunjukkan pada *CAPI* terdapat 13 isian yang tidak konsisten atau kesalahan konsep dan defininsi, sedangkan pada *PAPI* terdapat 123 isian. Pada PKL 55 kasus yang serupa terjadi sebanyak 87 kasus atau sekitar 5,1% dari total kuesioner yang dicacah dengan *CAPI*, sedangkan statistik kuesioner yang dicacah dengan *PAPI* masih dalam proses pengolahan ketika laporan penelitian ini dibuat. Kesalahan tersebut dapat diakibatkan oleh banyak hal, seperti kesalahan pemahaman konsep oleh pencacah, kesalahan entri, ataupun kesalahan validasi data (*routing*) pada saat mengisi kuesioner. Kesalahan yang dapat diminimalisir oleh *CAPI* adalah kesalahan entri, yang divalidasi langsung pada saat pencacahan di lapangan, dan kesalahan *routing* di mana *routing* dilakukan secara otomatis oleh aplikasi *CAPI*.

Untuk lebih menekan jumlah kesalahan pada kasus inkonsistensi atau kesalahan konsep dan definisi, penulis merekomendasikan untuk menambahkan fitur e-learning yang memungkinkan pencacah untuk mempelajari konsep dan definisi dengan mudah melalui aplikasi *CAPI*. Fitur dapat bersifat pasif, di mana trigger dilakukan oleh *interviewer*, atau bersifat pasif di mana aplikasi akan memantau dan mempelajari isian yang dientrikan oleh *interviewer*.

Kasus kesalahan pengentrian (*wrong key*) data yang tercatat adalah sebanyak 34 kasus pada kuesioner *CAPI* PKL 54. Kesalahan pengentrian tersebut berupa kesalahan menginputkan Nomor Kode Sampel (NKS). Pada PKL 55, kesalahan tersbut dapat dihilangkan dengan menerapkan mekanisme yang berbeda, di mana pencacah tidak perlu menginputkan NKS, tetapi kuesioner yang sudah dilengkapi dengan NKS dan biodata responden akan secara otomatis di-set oleh sistem *CAPI* kepada perangkat masing-masing *interviewer* sesuai dengan sampel yang akan dicacah.

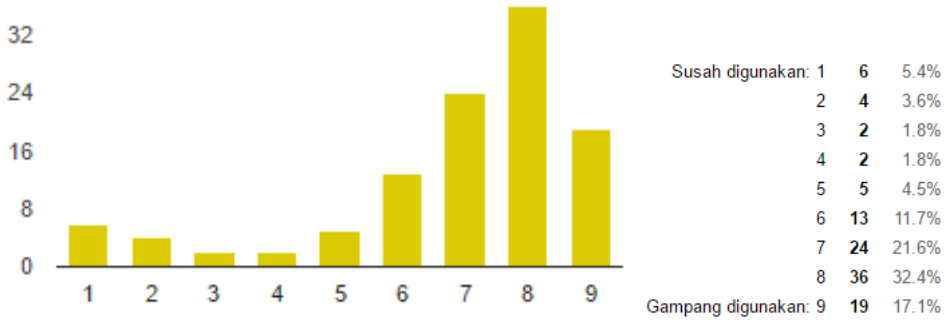
Pada PKL 55, terdapat 39 kuesioner atau sekitar 2,23% yang mengalami kasus *missing value* setelah interview dilakukan selama masa pengumpulan data di lapangan. Hal ini disebabkan oleh validasi rentang nilai yang telah ditanamkan pada kuesioner yang mengakibatkan nilai diluar rentang tersebut tidak dapat diterima oleh aplikai *CAPI*. Hal ini diakibatkan oleh adanya kasus yang tidak terpantau pada saat survey pendahuluan sehingga tidak dihandle oleh desain dan validasi kuesioner. Pada pencacahan dengan *PAPI*, hal tersebut dapat teratasi dengan adanya proses *editing*, namun pada *CAPI* hal tersebut mengakibatkan data tidak dapat diinputkan ke kuesioner digital. Untuk mengatasi hal tersebut, perlu dibuat mekanisme untuk menangkap nilai-nilai diluar rentang yang telah ditetapkan, misalnya memungkinkan untuk tetap mengisikan nilai diluar rentang dengan memunculkan pesan/notifikasi. Pendekatan ideal yang direkomendasikan penulis adalah dengan menerapkan sistem pelaporan berjenjang nilai yang anomali, mulai dari pencacah, Kortim, Intruktur Daerah, Instruktur Nasional, hingga ke subject matter yang diintegrasikan dengan updating validasi kuesioner secara broadcast. Namun, penerapan hal ini perlu dirancang dengan baik, karena melibatkan komunikasi data yang intens antara semua pihak yang terlibat pada pelaksanaan survey.

3. *Usability*

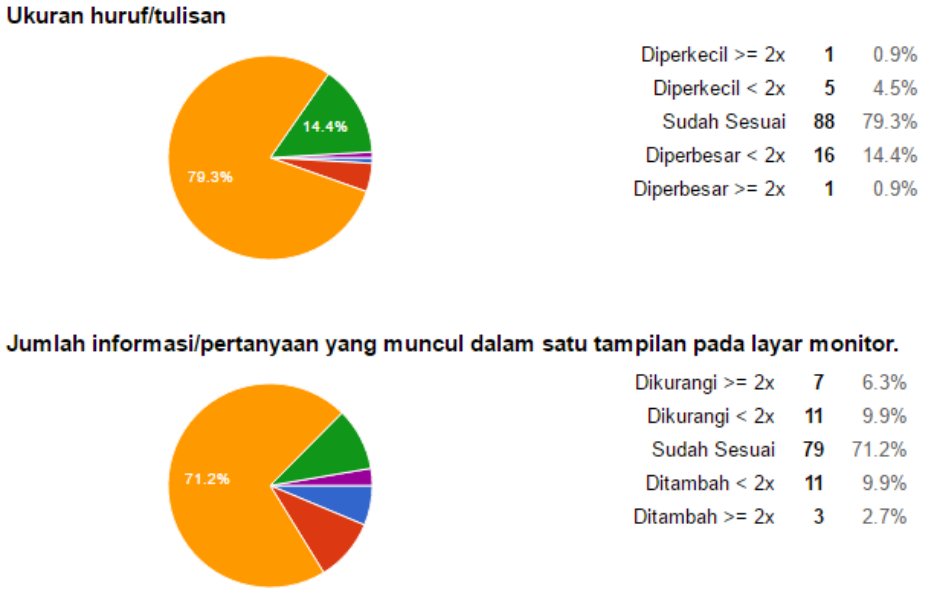
Usability diukur dengan *Questionnaire of User Interface Satisfaction*. (QUIS). Tanggapan Umum yang diberikan oleh *interviewer* pada PKL 54 berkisar antara 7 hingga 9 dari skala likert 1 (negatif) hingga 9 (positif). Hal ini menunjukkan bahwa desain kuesioner dengan *CAPI* secara keseluruhan (*overall*) sudah sesuai dengan yang diinginkan *interviewer*. **Error! Reference source not found.** menunjukkan contoh salah satu variable yang dinilai pada Tanggapan Umum. Hasil selengkapnya dapat dilihat pada Lampiran 2.

Pada bagian Tampilan Layar Monitor hasil *QIUS*, terdapat 3 hal yang membutuhkan perbaikan tampilan, yaitu:

1. Penggunaan *highlighting*, yakni Penggunaan warna, ukuran, ketebalan



Gambar 4. Hasil survey kemudahan penggunaan *CAPI* PKL 54



Gambar 5. Penilaian interviewer PKL 54 terhadap informasi yang ditampilkan aplikasi *CAPI* pada layar monitor

(*bold*), dan semacamnya yang bersifat khusus untuk menandai informasi penting,

2. Perlunya kemudahan navigasi kembali ke tampilan sebelumnya, dan
3. Tampilan progress penyelesaian pekerjaan, yakni Tampilan berapa bagian (persen) isian yang sudah dan belum diselesaikan pada aplikasi.

Proses yang membutuhkan waktu tunggu yang lama, seperti upload/download kuesioner perlu dilengkapi dengan *progress bar*. Selain itu, *interviewer* juga merasa khawatir untuk mengeskplorasi sendiri fitur-fitur yang ada padi *CAPI* dengan mekanisme *trial and error*. Hal ini dikarenakan tidak terdapatnya halaman simulasi dan fitur *undo/redo*. Untuk itu, pada pengembangan *CAPI* perlu ditambahkan fitur tersebut untuk menghasilkan *user experience* yang lebih baik terhadap aplikasi *CAPI*. Karena keterbatasan jumlah halaman, statistik pendukung pembahasan di atas tidak disertakan dalam tulisan ini.

Keluhan yang juga diutarakan oleh *interviewer* adalah fitur *auto correct* dan *auto complete* yang mengakibatkan tulisan yang diinput pencacah diubah secara otomatis oleh sistem sehingga tidak sesuai dengan yang diharapkan. Fitur tersebut perlu di-non-aktifkan atau menggunakan field input yang tidak terpangaruh oleh *auto complete*. Informasi yang tercakup pada panduan pengguna juga perlu dibuat lebih informatif agar mudah dipahami secara cepat oleh *interviewer*. Informasi yang ditampilkan pada aplikasi *CAPI* perlu dibedakan dengan buku panduan yang digunakan saat pelatihan dengan

mempertimbangkan efisiensi dan keterbatasan ukuran layar hardware.

Keterbatasan dimensi layar monitor pada *CAPI* membuat desain kuesioner perlu disesuaikan sehingga informasi yang tampil pada layar monitor dapat terbaca dengan jelas oleh *interviewer*. **Error! Reference source not found.** menunjukkan hasil *QUIS* untuk penilaian *interviewer* terhadap informasi yang ditampilkan aplikasi *CAPI* pada layar monitor.

Berdasarkan hasil tersebut, ukuran huruf dan jumlah informasi yang ditampilkan pada satu tampilan layar monitor sudah sesuai sehingga bisa terbaca dengan jealas oleh *interviewer*. Adapun ukuran huruf yang digunakan berkisar antara 12 hingga 14 *point (pt)* dan jumlah pertanyaan pada satu kali tampilan berkisar antara 4 hingga 7 pertanyaan.

KESIMPULAN DAN SARAN

1. Kesimpulan

Berdasarkan studi literatur yang dilakukan pada penelitian ini, yakni mengenai implementasi *CAPI* di berbagai negara selama 15 tahun terakhir, terdapat 3 kelompok variabel yang dapat digunakan sebagai ukuran keberhasilan penerapan *CAPI* pada kegiatan pengumpulan data, yakni *system performance*, *data quality*, dan *usability*. Hasil dari *pilot study* yang dilakukan, yakni pada PKL 54 dan 55, menunjukkan bahwa *CAPI* dapat memberikan performa yang lebih baik dibandingkan dengan *PAPI*, khususnya dalam hal durasi proses pencacahan hingga pengolahan data. Sistem notifikasi berjenjang antara Kortim dan pencacah dapat membantu Kortim memonitor kesalahan isian oleh pencacah sehingga mengoptimalkan peran Kortim.

Dari segi kualitas data, penggunaan *CAPI* dapat memberikan validasi data yang lebih baik dengan pengetrian dan validasi data melalui aplikasi di lapangan dan adanya fitur *automatic routing* yang meminimalisir inkonsistensi isian kuesioner. Namun, kendala kesulitan mengentri data pada *device* yang memiliki alat input yang kurang ergonomis merupakan hambatan yang mengganggu performa dan dapat menurunkan kualitas data yang dihasilkan dengan *CAPI*.

Desain antarmuka pada *CAPI* yang dikembangkan oleh STIS secara umum mendapat tanggapan positif dari *interviewer* yang menggunakan. Adapun beberapa hal yang menjadi masukan untuk meningkatkan *usability* dari *CAPI* STIS, diantara adalah penambahan fasilitas *undo/redo* serta simulasi aksi yang akan dilakukan pada aplikasi, dan menghindari pengaruh fitur *auto complete* dan *auto correct* pada saat pengentrian isian kuesioner *CAPI*.

Secara garis besar, desain *CAPI* yang diterapkan dan telah diujicoba pada *pilot study*, dengan jumlah sampel yang cukup representatif untuk menguji sistem *CAPI*, telah siap untuk digunakan dalam survey skala besar. Pengaruh negative *CAPI* terhadap variabel-variabel yang diteliti tidak signifikan dibandingkan dengan pengaruh positif yang diberikan dibandingkan dengan *PAPI* yang selama ini diterapkan.

2. Saran

Penelitian ini telah menghasilkan sejumlah instrument untuk mengukur kinerja *CAPI* dari berbagai aspek. Namun, pada penelitian ini masih terdapat beberapa keterbatasan, baik dari segi persiapan pencatatan data yang dibutuhkan, maupun jumlah variabel yang diamati. Oleh karena itu, selain mempelajari hasil pengolahan data dan analisis dari penelitian ini, hasil studi literatur pada penelitian ini sebaiknya dimanfaatkan pembaca sebagai referensi hal-hal yang perlu diperhatikan pada penerapan *CAPI*. Demikian pula untuk penelitian selanjutnya agar dapat meneliti variabel yang lebih lengkap seperti yang dipaparkan penulis pada bab studi literatur.

Penelitian dan beberapa penelitian sebelumnya masih menghasilkan beberapa *issue* terkait penerapan *CAPI* serta masih terdapatnya kelemahan pada *PAPI* yang belum teratasi, misalnya untuk kasus kesalahan konsep oleh pencacah yang menyebabkan kesalahan isian. Hal tersebut memiliki dampak negatif pada *CAPI* karena akan menghasilkan *routing* yang salah dan

tidak adanya bukti tertulis/analog sebagai backup. Tentu hal tersebut perlu diteliti lebih lanjut untuk menemukan solusinya. Kasus terdapatnya nilai diluar rentang validasi juga merupakan contoh lain yang perlu diperhatikan, misalnya dengan menerapkan sistem pelaporan dan *updating range* validasi berjenjang.

Meskipun penerapan sistem notifikasi dari dan ke Kortim dapat menghasilkan validasi yang berlapis pada *CAPI*, namun pada studi kasus PKL di STIS, beban Kortim menjadi lebih berat karena memiliki tanggung jawab membackup pekerjaan pencacah. Sebaiknya pada PKL selanjutnya, atau kegiatan survey yang menerapkan sistem ini, tanggungjawab Kortim sebagai backup dari pencacah perlu dihilangkan, dan digantikan dengan pencacah cadangan selain Kortim.

DAFTAR PUSTAKA

Baker, R.P. (1992): New Technology in Survey Research: Computer-Assisted Personal Interviewing (CAPI). Social Science Computer Review, 10, 145-157.

Baker, R.P., Bradburn, N., and Johnson, R. (1994): CAPI: An Experimental Evaluation. In American Statistical Association (Ed.), Proceedings of the Sec- tion on Survey Research Methods, 851-855.

Baker, R. P., Bradburn, N. M., & Johnson, R. A. (1995). Computer-assisted personal interviewing: an experimental evaluation of data quality and cost. Journal of Official Statistics, 11(4), 413-431.

Bernabe-Ortiz, A., Curioso, W. H., Gonzales, M. A., Evangelista, W., Castagnetto, J. M., Carcamo, C. P., ... & Holmes, K. K. (2008). Handheld computers for self-administered sensitive data collection: a comparative study in Peru. BMC medical informatics and decision making, 8(1), 11.

Bishop, Yvonne M., Warren L. Buckler, Robert P. Parker, and Charles E. Caudill. 1990. "Computer Assisted Survey Information Collection." (April).

Caviglia-harris, Jill et al. 2012. "Improving Household Surveys Through Computer-Assisted Data Collection: Use of Touch-Screen Laptops in Challenging Environments."

Chalmers, Neil, and Joachim De Weerdt. 2010. "A Comparison of CAPI and PAPI through a Randomized Field Experiment." (November):1–56.

Childs, J. H., & Landreth, A. (2006). Analyzing interviewer/respondent interactions while using a mobile computer-assisted personal interview device. Field methods, 18(3), 335-351.

Couper, Mick P., and Geraldine Burt. 1989. "THE IMPACT OF COMPUTER-ASSISTED PERSONAL INTERVIEWING (CAPI) ON INTERVIEWER PERFORMANCE: THE CPS EXPERIENCE." 189–93.

Couper, M.P. and Groves, R.M. (1992): Interviewer reactions to alternative hardware for computer-assisted personal interviewing. Journal of Official Sta- tistics, 8, 201-210.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. Social Science Computer Review, 18(4), 384-396.

De Leeuw, E. D. 1993. "Data Quality in Mail, Telephone, and Face to Face surveys".

Fuchs, Marek, Mick P. Couper, and Sue Ellen Hansen. 2000. "Technology Effects : Interview Duration in CAPI and Paper and Pencil Surveys."

Lynn, P., & Purdon, S. (1994). Time-series and lap-tops: the change to computer-assisted interviewing. INTERNATIONAL SOCIAL ATTITUDES, 141-141.

Manners, Tony. 1990. "THE DEVELOPMENT OF COMPUTER ASSISTED INTERVIEWING (CAI) FOR HOUSEHOLD SURVEYS : THE CASE OF THE BRITISH LABOUR FORCE SURVEY."

Martin, J. (1993, October). PAPI to CAPI: the OPCS experience. In Essays on

Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London (pp. 96-117).

Matheson, Jil. 1991. "APPLICATION OF COMPUTER ASSISTED INTERVIEWING TO THE FAMILY EXPENDITURE SURVEY." (February):1–48.

Müller, S., & Kesselmann, P. (1996). Akzeptanz von computergestützten Erhebungsverfahren. Ein empirischer Vergleich mit der traditionellen Fragebogentechnik. Marketing ZFP, 18(3), 191-202.

Randolph, Justus J., Marjo Virnes, Ilkka Jormanainen, and Pasi J. Eronen. 2006. "The Effects of a Computer-Assisted Interview Tool on Data Quality." 9:195–205.

Sainsbury, Roy, John Ditch, and Sandra Hutton. 1993. "Computer Assisted Personal Interviewing." (3).

Sainsbury, Roy, John Ditch, and Sandra Hutton. 1995. "The Effect of Computer-Assisted Interviewing on Data Quality: A Review."

Shaw, Arthur, Lena Nguyen, and Ulrike Nischan. 2011. "Comparative Assessment of Software Programs for the Development of Computer-Assisted Personal Interview (CAPI) Applications." (July).

Slaughter, Laura, Ben Harper, and Kent Norman. 1994. "Assessing the Equivalence of the Paper and On-Line Formats of the QUIS 5 . 5."

Wensing, Fred, Jane Barresi, David Finlay, and Australian Bureau. 2003. "Developing an Optimal Screen Layout for CAI." 63–76.

BERAS ATAU ROKOK?:
Beban Ekonomis Rumah Tangga Miskin di Indonesia 2014

Andri Yudhi Supriadi[1], Aris Rusyiana[2]

Badan Pusat Statistik
e-mail: [1]andri@bps.go.id

Abstrak

Fakta bahwa di beberapa negara berkembang, konsumsi rokok menimbulkan beban ekonomis yang signifikan (Toukan, 2016; Block dan Webb, 2009). Juga, untuk konteks Indonesia kontemporer, Kepala BPS mengatakan bahwa belanja rokok merupakan pengeluaran kedua terbesar dan memberikan kontribusi nyata terhadap angka kemiskinan nasional. Namun, kajian kontemporer yang secara komprehensif membahas beras dan rokok terhadap kemiskinan belum banyak dibahas. Celah penelitian tersebut menjadi dasar bagi kami untuk melakukan kajian mengenai hubungan konsumsi beras dan pengeluaran potensial rokok di antara rumah tangga miskin di Indonesia 2014. Untuk keperluan telaah kajian penelitian ini, kami membagi kategori rumah tangga berdasarkan tempat tinggal (perdesaan/perkotaan), rumah tangga dengan banyak anggota rumah tangga usia dewasa (di atas 15 tahun), dsb. Tujuan dari kajian ini adalah untuk menganalisa apakah rumah tangga miskin lebih memilih mengurangi konsumsi beras dibanding mengurangi konsumsi rokok .Untuk kajian ini, kami menggunakan Survei Sosial Ekonomi Nasional tahun 2014. Dengan menggunakan Model Regresi Linier Berganda, kami menggunakan sampel rumah tangga yang memiliki anggota rumah tangga dewasa yang merokok ($N_{Indonesia}$ = 285.371). Hasil penelitian kami menunjukkan bahwa rumah tangga miskin yang memiliki anggota rumah tangga perokok secara rata-rata mengkonsumsi beras relatif lebih sedikit dibandingkan rumah tangga yang tidak memiliki anggota rumah tangga perokok, baik yang termasuk kategori miskin maupun tidak. Hal ini mengindikasikan bahwa rumah tangga miskin lebih memprioritaskan konsumsi rokok dibandingkan konsumsi beras.

Kata kunci: Susenas, rumah tangga miskin, konsumsi rokok, regresi linier berganda

Abstract

Facts that in many developing countries, cigarettes consumption affects significantly toward economic burden (for instances see Toukan, 2016; Block and Webb 2009). Also, for Indonesian recently context, Chief of Statistics Indonesia says that cigarettes expenditure pose the second highest shared towards the national poverty rate. However, the recently comprehensive Indonesia researches on rice and cigarettes expenditure are still rare. Regarding those research gaps, we examine the linkage of rice consumption expenditure and the potential cost of cigarettes expenditure among poor households in Indonesia (includes the households characteristics: residency, social safety net receiver, adults smokers among households, etc). The objectives of this study is to examine whether poor households prefer to consume fewer rice rather than consuming fewer cigarettes. For this study, we use the National Social Economic Survey of the 2014 year dataset. By applying the multiple linear regression analysis, we use sample of adult smokers (N=285,371). Our results show that poor smoking-households relatively consume rice less than the non-smoking-households categories on average. This may indicate that poor households prioritize to consume more cigarettes rather than consuming rice.

***Keywords**: Susenas, poverty rate, cigarettes consumption, multiple linier regression*

PENDAHULUAN

Beban ekonomis rokok Indonesia tidak terlepas dari 5 (lima) fakta penting, sebagai berikut. Pertama, *Survey Global Adult Tobacco* (GAT) (*World Health Organization*, 2012) menggarisbawahi bahwa Indonesia merupakan negara produsen tembakau kelima terbesar di dunia. Posisi ini menempatkan Indonesia dalam 5 negara teratas produsen dan eksportir tembakau di dunia. Selain itu, Indonesia merupakan negara konsumen rokok terbesar keempat di dunia. Jumlah laki-laki dewasa yang merokok menempati urutan ketiga teratas, dan perempuan yang merokok termasuk ranking 17 besar dunia. Sebagai contoh, di tahun 2008, konsumsi rokok di Indonesia mencapai 255 milyar batang rokok per tahun, dan tahun 2017 telah mencapai lebih dari 400 milyar batang rokok per tahun. Kedua, Jumlah populasi perokok dewasa di Indonesia relatif banyak. Contohnya, GATS (*World Health Organization*, 2012) mencatat terdapat 59,8 juta orang dewasa (34% populasi penduduk), yang terdiri dari 67 % perokok laki laki and 2,7% perokok perempuan. Ketiga, rokok membebani anggaran pemerintah di fungsi kesehatan. Contoh, hasil Riset Kesehatan Dasar tahun 2010 yang dimuat di dalam laporan survey GAT tahun 2011 mencatat total biaya medis untuk mengobati kasus kesehatan akibat merokok tahun 2010 adalah Rp 1,85 triliun. Anggaran ini digunakan untuk membiayai 624.000 kasus rawat inap terkait penyakit yang diakibatkan merokok. Masih menurut Riskesdas, di tahun 2010 terjadi 191.000 total kasus kematian yang diakibatkan rokok, yang merupakan: 100.680 laki-laki dan 89.560 wanita meninggal dengan sebab penyakit yang berhubungan dengan rokok (*tobacco-related deseases*). Jumlah kejadian kematian ini merupakan 12,7 persen dari total kejadian kematian di Indonesia di tahun 2010.

Beberapa pustaka membahas tingginya prevalensi merokok di antara orang miskin di negara-negara berkembang (Kusumawardani, dkk, 2018; dan Toukan, 2016). Kusumawardani, dkk (2018) menunjukkan bahwa orang dewasa yang merokok di dalam kuantil orang-orang termiskin berpeluang merokok dua kali lipat dibanding orang-orang dewasa dalam kuantil orang-orang terkaya. Kusumawardani, dkk (2018) juga menemukan dari hasil olah data Riskesdas bahwa prevalensi orang dewasa yang merokok di Indonesia adalah 7,2 %, di mana tingkat prevalensi merokok lebih tinggi terdapat pada pria dewasa dibandingkan dengan wanita dewasa perokok Toukan (2016) juga menemukan bahwa di Yordania, prevalensi merokok itu tertinggi terdapat pada orang-orang Yordania yang termiskin.

Dengan mempertimbangkan kekurangan pustaka yang ada di dalam meneliti beban ekonomis rokok di dalam kajian pengentasan kemiskinan, kami memandang perlu untuk menyajikan sudut pandang baru di dalam memahami hubungan antara konsumsi beras dengan konsumsi rokok di Indonesia. Berdasarkan data survei sosial ekonomi nasional, kami menguji hipotesis nol mengenai konsumsi rokok yang tinggi di antara rumah tangga miskin tidak berhubungan erat dengan pengurangan konsumsi beras di Indonesia tahun 2014. Kami menduga bahwa prevalensi konsumsi rokok yang tinggi tidak berhubungan dengan pengurangan konsumsi beras di Indonesia.

Kajian ini mencoba menjawab 2 (dua) pertanyaan penelitian di dalam mengevaluasi hubungan konsumsi beras dan prevalensi rokok yang tinggi di Indonesia, yaitu:

1. Apakah prevalensi konsumsi rokok yang tinggi berhubungan dengan pengurangan konsumsi beras di Indonesia?;
2. Variabel-variabel ekonomis lain dan dan karakteristik demografis apa yang mempengaruhi pengeluaran belanja konsumsi rokok rumah tangga di Indonesia?

Penelitian ini mempunyai sekurang-kurangnya 3 (tiga) keterbatasan, antara lain. **Pertama**, kajian ini menggunakan dataset 1 (satu) tahun saja, koefisien estimasi di dalam model bisa jadi belum merupakan

temuan empiris yang robust untuk konteks Indonesia. **Kedua**, model penelitian ini dapat mengandung masalah endogenitas antara konsumsi beras dan konsumsi rokok. Kita harus berhati-hati mengenai kemungkinan hubungan kausalitas yang terjadi. Koefisien estimasi model seharusnya dapat dilihat sebagai ukuran hubungan, bukan ukuran pengaruh. Kausalitas dapat menyebabkan bias estimasi antara konsumsi beras rendah mempengaruhi konsumsi rokok yang tinggi atau konsumsi rokok yang tinggi dapat menyebabkan konsumsi beras yang rendah. **Ketiga**, penelitian kami belum secara tegas mengukur pengeluaran belanja beras dan belanja konsumsi rokok. Secara berturutan, kami menggunakan pendekatan pengeluaran rumah tangga untuk konsumsi makanan pokok per bulan dan pengeluaran untuk rokok, tembakau dan sirih, masing-masing untuk mengukur belanja beras dan beban pengeluaran konsumsi rokok. Penelitian mendatang diharapkan dapat lebih menghadirkan ukuran yang lebih spesifik untuk pengeluaran rumah tangga untuk belanja beras dan rokok.

Meskipun masih memiliki kelemahan, penelitian kami ini mempunyai beberapa peranan penting terhadap pustaka hubungan rokok dan beras serta konsekuensi negatif rokok yang masih terbatas (Lee dan Yi, 2016; Papadopoulu, dkk, 2017; Toukan, 2016; Block dan Webb, 2009; Semba, dkk, 2007; Hu, 2008; Prasad dan Dhar, 2017; Bergström, 2004). **Pertama**, kajian penelitian ini menggarisbawahi konsumsi rokok berasosiasi negatif dengan kebiasaan makan (Lee dan Yi, 2016; Papadopoulu, dkk, 2017). **Kedua**, kajian ini menunjukkan bahwa pengeluaran belanja rokok menimbulkan beban ekonomis terhadap Produk Domestik Bruto (PDB) di negara-negara berkembang dan juga negara-negara maju (Toukan, 2016; Sung, dkk, 2006; Kang, dkk, 2003). **Ketiga**, kajian ini menekankan konsekuensi negatif rokok di antara rumah tangga orang miskin sementara dalam pustaka-pustaka yang ada menunjukkan konsekuensi negatif konsumsi rokok pada resiko kanker dan penyakit-penyakit yang diakibatkan rokok, dan kehidupan di usia tua tak sehat dan tak produktif di negara berkembang dan negara maju (Prasad dan Dhar, 2017; Bergström, 2004). Contohnya, Korhonen, dkk (2015) menunjukkan kepada kita bahwa di Finlandia, konsumsi rokok secara signifikan dapat memprediksi usia pensiun yang tak berkemampuan (*disability retirement*) atau usia senja yang tak sehat (*unhealthly elderly*). Di dalam konteks Indonesia kontemporer, kajian kami fokus pada konsekuensi negatif dari konsumsi rokok pada rumah tangga miskin. Selain itu, kajian kami juga mengkaji dampak program keluarga harapan (PKH) dengan pengeluaran konsumsi makanan pokok rumah tangga penerima PKH.

Telaah penelitian ini kami sajikan dalam susunan pembahasan sebagai berikut. Pertama, di subbab A pendahuluan membahas lima fakta penting darurat merokok di Indonesia, dilanjutkan dengan signifikansi penelitian dan celah penelitian (*research gap*), serta pertanyaan penelitian. Berikutnya, di subbab B (Metodologi), tulisan kami menyajikan ulasan tentang data set, tinjauan referensi serta analisis yang digunakan di dalam penelitian ini. Sedangkan, di subbab C, kami menyajikan hasil dan pembahasan. Hasil penelitian kami sajikan dalam analisis statistik deskriptif menggunakan grafik garis (*line plot*) antara variabel angka kemiskinan (*headcount poverty rate*) antar provinsi, pengeluaran konsumsi beras dan makanan pokok, serta belanja konsumsi rokok rumah tangga. Selain itu, hasil dan pembahasan kami lengkapi dengan hasil analisis inferensi statistik menggunakan Analisis Regresi Linier Berganda. Adapun, di subbab D, kami sajikan kesimpulan dan saran dari hasil mengkaji hubungan antara konsumsi beras dan makanan pokok dengan karakteristik rumah tangga (Ruta), termasuk Ruta perokok dan miskin, serta karakteristik Ruta lainnya, berserta pengeluaran bulanan Ruta.

DATA DAN METODE

1. Data dan Tinjauan Referensi

Kami menggunakan data survei nasional yang representatif di dalam memahami hubungan antara konsumsi beras dengan pengeluaran rumah tangga untuk rokok. Kajian kami berdasarkan Survei Sosial Ekonomi Nasional (Susenas) tahun 2014. Pemilihan Susenas tahun 2014 sebagai batasan penelitian karena di tahun tersebut Susenas terakhir kali dilaksanakan secara triwulan (tahun berikutnya dilaksanakan semesteran). Dengan periode pendataan yang lebih panjang, informasi yang diperoleh lebih lengkap.

Susenas merupakan salah satu survei unggulan yang dilaksanakan Badan Pusat Statistik (BPS). Survei ini telah dilaksanakan sejak tahun 1993 untuk mengumpulkan informasi yang meliputi status sosial ekonomi dan akses individu dan rumah tangga terhadap layanan publik di seluruh kabupaten/kota di Indonesia. Saat ini, Susenas meliputi 300.000 rumah tangga sampel di 497 kabupaten/kota atau setara mendekati sejumlah 1,2 juta individu (BPS, 2015). Susenas juga berisi informasi mengenai karakteristik sosiodemografis kepala rumah tangga dan anggota rumah tangga, termasuk pendidikan, pekerjaan, dan pengeluaran mereka. Selain itu, Susenas berisi informasi mengenai pola konsumsi dan pengeluaran rumah tangga yang meliputi konsumsi makanan dan nonmakanan.

Kami menggunakan data Susenas Kor dan Modul Konsumsi sebagai sumber data utama penelitian ini karena beberapa alasan, sebagai berikut. Pertama, Susenas merupakan survei BPS yang menggunakan metode pengambilan sampel probabilitas yang telah teruji sejak lama. Teknik penarikan sampel probabilitas merupakan standar terbaik di dunia saat ini di dalam menciptakan sampel representatif karena secara matematis dapat memprediksi *sample error* (Neuman, 2014, halaman 247). Kedua, Susenas Kor menyajikan informasi yang lengkap, yang menjadi dasar penerbitan statistik kesejahteraan rakyat, bidang pendidikan, kesehatan, dan perumahan. Sementara Susenas modul konsumsi menyajikan statistik terkait rata-rata pengeluaran rumah tangga yang diperinci untuk jenis makanan dan non makanan, yang menjadi dasar perhitungan angka kemiskinan dan gini rasio (ketimpangan).

Untuk analisis inferensi di dalam kajian kami ini, kami menggunakan analisis Regresi Linier Berganda (RLB) karena sekurang-kurangnya 2 (dua) pertimbangan. Pertama, analisis RLB merupakan salah satu metode analisis yang paling populer di berbagai ranah penelitian (Darlington dan Hayes, 2017). Darlington dan Hayes (2017) berpendapat bahwa sejauh ini analisis RLB merupakan metode analisis paling populer di ranah ilmu sosial, ilmu analisa perilaku, kesehatan masyarakat, kedokteran, dan lain lain. Kedua, kajian kami memeriksa hubungan antara beberapa variabel *independent* dan variabel *dependent*. Darlington dan Hayes (2017) juga mengatakan bahwa RLB merupakan salah satu metode statistik untuk memodelkan hubungan antara variabel-variabel *independent* dengan variabel *dependent*.

2. Metode Analisis

Untuk analisis inferensi di dalam kajian kami ini, kami menyajikan suatu model kajian dalam notasi sederhana bila menggunakan RLB. Olive (2017) menyajikan dalam buku terbarunya berjudul "*Linier Regression*" suatu rumus sederhana RLB. Menurut Olive (2017), model RLB memiliki sekurang-kurangnya 2 *variable independent*. Olive (2017) menyajikan notasi RLB dengan menganggap variabel Y merupakan variabel skala interval/rasio dan sekurang-kurangnya ada 2 *variable independent* yang merupakan variabel kuantitatif. Berdasarkan notasi Olive untuk model RLB, kami menyajikan notasi penelitian kami dengan rumus sebagai berikut:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k + e \quad (1)$$

dengan Y merupakan variabel respon, $x_{1,2,\ldots,k}$ merupakan suatu vektor $k\ x\ 1$ untuk prediktor nontrivial. Adapun bo merupakan suatu konstanta tidak diketahui,

dan $b_1, b_2, .. b_k$ merupakan suatu vektor $k \times 1$ dari koefisien-koefisien tidak diketahui, dan e merupakan suatu variabel acak yang dinamakan error (galat).

Berdasarkan notasi rumus Olive, model penelitian kami dapat dirumuskan secara empiris dalam notasi RLB dengan mengakomodir interaksi antar variabel independen, sebagai berikut:

$$Y = b_0 + b_{1ij}X_{1i}.X_{2j} + b_2X_3 + b_{3i}X_{1i}.X_3 + b_4X_4 + b_5X_5 + b_6X_6 + e \quad (2)$$

di mana :

Y merupakan pengeluaran bulanan rumah tangga untuk konsumsi beras

X_{1i} merupakan *dummy variable* rumah tangga merokok (i=0 jika tidak merokok, i=1 jika merokok)

X_{2j} merupakan *dummy variable* rumah tangga miskin (j=0 jika tidak miskin, j=1 jika miskin)

X_3 merupakan jumlah anggota rumah tangga dewasa

X_4 merupakan *dummy variable* wilayah kota/desa

X_5 merupakan rumah tangga menerima Program Keluarga Harapan

X_6 merupakan total pengeluaran rumah tangga.

X_1X_2 merupakan *variable* interaksi untuk melihat perbedaan konsumsi beras antara rumah tangga merokok/tidak merokok dan miskin/tidak miskin

X_1X_3 merupakan *variable* interaksi untuk melihat perbedaan konsumsi beras antara rumah tangga merokok dengan jumlah anggota rumah tangga dewasa banyak/sedikit

HASIL DAN PEMBAHASAN

1. Karakteristik Sampel Penelitian

Tabel 1 menyajikan ringkasan statistik untuk ukuran-ukuran di dalam model asosiasi karakteristik rumah tangga dengan pengeluaran belanja beras dan rokok. Secara detail, masing-masing ukuran ringkasan statistik dapat dijabarkan, sebagai berikut. Pertama, Beras dihitung sebagai total belanja rumah tangga untuk beras dan makanan pokok selama satu bulan.

Tabel 1 menunjukkan kepada kita bahwa N=285.371 rumah tangga di Indonesia membayar belanja beras dan konsumsi makanan pokok rata-rata sekitar Rp 1,5 juta per bulan (dalam rentang data terkecil Rp 60 rb sampai dengan terbesar Rp 20 juta). Berikutnya, rumah tangga perokok dihitung dengan cara mengkategorikan rumah tangga yang ada anggota rumah tangga yang perokok sebagai "1". Sedangkan untuk kategori rumah tangga nonperokok dikategorikan "0" sebagai selain rumah tangga perokok.

Adapun, rumah tangga miskin diukur dengan cara mengkategorikan rumah tangga yang memiliki pendapatan perkapita kurang dari 1,2 kali garis kemiskinan wilayah dibedakan antara desa dan kota. Garis kemiskinan yang digunakan mengacu kepada perhitungan BPS, yakni pengeluaran makanan dan non makanan minimal setara 2.300 kilokalori per hari. Sementara itu, variabel rumah tangga dewasa merupakan jumlah anggota rumah tangga yang berusia di atas 15 tahun. Tabel 1 juga menunjukkan kepada kita bahwa secara rata-rata, rumah tangga di Indonesia memiliki jumlah anggota rumah tangga dewasa berkisar 1 sampai dengan 17 orang (rata-rata terdapat 3 anggota rumah tangga dewasa). Adapun angka ringkasan statistik lainnya, pengeluaran, perkotaan dan PKH secara berturutan menyajikan variabel dummy untuk pengeluaran bulanan rumah tangga, rumah tangga yang tinggal di perkotaan dan rumah tangga penerima Program Keluarga Harapan (PKH). Susenas 2014 menyajikan rata-rata pengeluaran rumah tangga Rp 2,9 Juta (minimal Rp 115 rb dan maksimal Rp 212 juta). Selain itu rumah tangga yang bertempat tinggal di wilayah perkotaan lebih sedikit dibandingkan yang tinggal di wilayah perdesaan terlihat dari rata-rata variabel perkotaan yang bernilai 42,76.

2. Hasil Analisis Inferensia

Tabel 2 menunjukkan bahwa terdapat N= 285.371 yang dilibatkan di dalam analisa Regresi Linier Berganda pada kajian mengenai hubungan belanja rokok terhadap pengeluaran belanja beras rumah tangga

Tabel 1. Ringkasan Statistik Karakteristik Rumah Tangga dan Pengeluarannya

Variabel	Obs	Mean	Std. Dev.	Min	Max
Beras	285.371	1.479.532	922.726,2	60.857,1	19.600.000
Rumah Tangga Perokok	285.371	0,60954	0,487853	0	1
Rumah Tangga Miskin	285.371	0,09467	0,292764	0	1
Dewasa	285.371	3	1	1	17
Perkotaan	285.371	0,42764	0,494736	0	1
Program Keluarga Harapan	285.371	0,02429	0,153963	0	1
Pengeluaran	285.371	2.885.034	3123113	115.352	212.000.000

Sumber: Perhitungan penulis menggunakan data Susenas 2014

miskin. Hasil menunjukkan model regresi linier berganda dengan estimasi Ordinary Least Square (OLS) dengan perhitungan STATA versi 14 *autorobust*. Dengan estimasi autorobust, asumsi heteroskedastisitas pada asumsi OLS menjadi lebih longgar[1]. Perhitungan menggunakan data N=285.371 Susenas 2014 menunjukkan bahwa model RLB dapat diterima dengan menggunakan *goodness of fit* R kuadrat 55,77% variabel dependen dapat diterangkan oleh variabel-variabel independennya.

Hasil model regresi berganda menunjukkan bahwa pola konsumsi beras serta makanan pokok lainnya bervariasi di antara karakteristik ruta yang berhubungan dengan prevalensi konsumsi rokok dan kategori kemiskinan ruta di Indonesia (Tabel 2).

Besaran serta arah koefisien parsial variabel pada model regresi linier berganda (Tabel 2) yang mengandung interaksi antar variabel, dengan salah satunya adalah variabel dummy mempunyai interpretasi berbeda dengan interpretasi model dengan variabel non dummy. Untuk yang mengandung variabel dummy, kita dapat melihat keterbandingan paling rendah atau paling tinggi dibandingkan dengan konstata regresinya (nilai konstanta mencerminkan rata-rata pengeluaran untuk variabel *dummy* yang bernilai ’0”). Dari hasil regresi dapat terlihat ada 4 (empat) kategori ruta terendah yang mengkonsumsi beras serta makanan pokok lainnya dibandingkan dengan rata-rata konsumsi beras untuk ruta nonperokok dan tidak miskin, yaitu: (1) Ruta miskin nonperokok; (2) Ruta miskin perokok, (3) Ruta penerima Program Keluarga Harapan, dan (4) Ruta Perokok Dewasa. Konsumsi beras dan makanan pokok lainnya pada ruta miskin nonperokok dan ruta miskin perokok secara signifikan berturut-turut terendah pertama (-313.393) dan terendah kedua (-167.756,20) dibandingkan dengan konsumsi beras dan makanan pokok lainnya pada rumah tangga nonperokok dan tidak miskin (konstanta=303.747,30). Sedangkan Ruta penerima Program Keluarga Harapan (PKH) (-46.678) serta Ruta Perokok Dewasa (-16.983,93) merupakan karakteristik terendah keempat dan kelima ruta pengkonsumsi beras dan makanan pokok lainnya bila dibandingkan dengan rata-rata pengeluaran ruta nonperokok dan tidak miskin. Temuan ini dapat dibaca dengan asumsi variabel penelitian lainnya *ceteris peribus*.Untuk pengeluaran tertinggi ruta dibandingkan dengan ruta nonperokok dan tidak miskin berturut-turut dilakukan oleh karakteristik ruta, sbb: (1) Ruta perokok dan tidak miskin mengkonsumsi beras dan makanan pokok tertinggi di bandingkan ruta non perokok dan tidak miskin. (2) Ruta yang tinggal di wilayah perkotaan mengkonsumsi beras dan makanan pokok kedua terbanyak dibandingkan dengan ruta non perokok dan tidak miskin.

Sedangkan hubungan pengeluaran beras dan makanan pokok lainnya dengan

[1] http://www3.grips.ac.jp/~yamanota/Lecture_Note_9_Heteroskedasticity

Tabel 2. Hasil Analisa Regresi Linier Berganda

Variabel	Koef	Std Error
Ruta nonperokok#miskin	-313.393,30*	10.581,31
Ruta perokok#tidakmiskin	341.154,40*	7.952,50
Ruta perokok#miskin	-167.756,20*	9.915,61
Dewasa	187.845,50*	5.050,97
Ruta perokok#dewasa	-16.983,93*	3.188,24
Perkotaan	73.597,75*	7.147,84
Program Keluarga Harapan (PKH)	-46.678*	6.032,92
Pengeluaran	0,17*	0,01
Konstanta	303.747,30*	5.439.552
N observasi	**285.371**	
R squared	**0,5577**	
F hitung	**26.169,44**	

penambahan anggota rumah tangga dewasa di atas usia 15 tahun (variabel Dewasa) serta hubungan pengeluaran beras dan makanan pokok lainnya dengan pengeluaran rumah tangga (variabel Pengeluaran) dapat dibaca sebagai arah korelasi parsial, sebagai berikut: (1) Setiap penambahan 1 (satu) orang dewasa akan menambah pengeluaran beras dan makanan pokok lainnya setara Rp 187.845,50 per bulan dengan asumsi variabel lain ceteris peribus. (2) Setiap penambahan Rp 1 pengeluaran rumah tangga berkorelasi dengan penambahan belanja beras dan makanan pokok lainnya sebesar Rp 0,17 per bulan.

Penelitian kami menemukan 3 (tiga) fakta menarik, yaitu: (1) Temuan bahwa ruta miskin bukan perokok mengeluarkan belanja konsumsi beras dan makanan pokok lainnya terendah di banding ruta tidak miskin dan tidak merokok, menunjukkan bahwa kategori rumah tangga ini sebagai paling rentan untuk masuk atau keluar dari garis kemiskinan absolut (*absolute poverty line*). (2) Rumah Tangga (ruta) miskin perokok menempati urutan kedua terendah di dalam mengkonsumsi beras dan makanan pokok lainnya di bandingkan ruta tidak miskin dan nonperokok. Temuan menarik ini dapat merupakan pendugaan awal bahwa di ruta miskin perokok, ada bias preferensi dalam hal mendahulukan belanja beras dan makanan pokok lainnya dengan pengeluaran untuk konsumsi rokok.(3) Ruta penerima Program Keluarga Harapan (PKH) secara signifikan juga menempati urutan terendah ketiga di dalam belanja pengeluaran beras dan makanan pokok lainnya.

Temuan ketiga ini bila dikaitkan dengan temuan kajian kedua cukup menarik, karena bisa jadi ada kekurangtepatan sasaran di dalam penggunaan bantuan sosial pemerintah (*social safety net* / *social protection program*). Alih-alih PKH menjadi pendukung program pengurangan kemiskinan (*poverty reduction program*), program bantuan sosial pemerintah ini malah disalahgunakan menjadi belanja pengeluaran konsumsi rokok. Oleh sebab itu, ke depan pemerintah perlu melakukan kajian evalausi dampak kebijakan komprehensif (*impact evaluation*) dan perbaikan database ruta sasaran penerima program PKH, atau program bantuan pengamanan sosial lainnya. Pemerintah hendaknya menerbitkan payung hukum untuk monitoring dan evaluasi dampak PKH, misalnya dengan menindaklanjuti diskursus penambahan syarat larangan merokok bagi kepala rumah tangga dan anggota rumah tangga selama menerima program, dan ada diskualifikasi bagi ruta yang melanggar aturan ini. Sejalan dengan evaluasi dampak PKH juga, hendaknya pemerintah memprioritaskan ruta miskin yang jelas jelas tidak ada anggota rutanya yang perokok untuk diberikan PKH, atau program sejenisnya. Daripada memasukkan

ruta miskin yang jelas-jelas anggota ruta nya ada yang perokok. Hal ini sejalan dengan temuan pertama dikaitkan dengan temuan kajian ketiga.

3. Pembahasan

Pertanyaan mengenai apa hubungan antara konsumsi beras, makanan pokok dengan belanja rokok serta pola pengeluaran lainnya telah menjadi perhatian peneliti ilmu-ilmu sosial sejak lama. Namun, dalam konteks Indonesia kontemporer, ranah kajian ini belum banyak disentuh apalagi menggunakan data dari survei nasional dengan cakupan pengamatan yang luas. Berdasarkan data Susenas 2014, kami menganalisa hubungan antara beras dan makanan pokok dengan pola konsumsi rokok di antara beberapa karakteristik rumah tangga miskin di Indonesia.

Berdasarkan temuan-temuan dari hasil penelitian, kajian kami menemukan beberapa fakta menarik bahwa rumah tangga miskin yang anggota rutanya merokok, tipe ruta ini mengkonsumsi beras dan makanan pokok lainnya lebih rendah dibandingkan rumah tangga tidak miskin di kategori tidak merokok. Hal ini mengindikasikan bahwa rumah tangga miskin di Indonesia cenderung memprioritaskan untuk membeli sebungkus rokok dibandingkan mengkonsumsi beras dan makanan pokok lainnya. Sedangkan pada karakteristik ruta non miskin perokok, konsumsi beras dan makanan pokok tetap relatif banyak. Meskipun bukti temuan kami masih lemah dalam kajian 1 (satu) tahun ini, temuan ini selaras dengan penelitian terdahulu seperti yang dilakukan oleh Kusumawardani, dkk (2013), Block dan Webb (2009), dan Semba, dkk (2007) yang menemukan bahwa rumah tangga miskin cenderung membelanjakan lebih banyak uang untuk merokok dibandingkan rumah tangga tidak miskin. Bila Kusumawardani, dkk (2013), dan Semba, dkk (2007) menunjukkan bukti berdasarkan data survey nasional lain (Riskesdan dan Survey Pengawasan Gizi Indonesia), serta Block dan Webb (2009) berdasarkan Susenas terdahulu, maka temuan kami menunjukkan kebaruan, baik dari data sets dan juga beberapa temuan menarik lainnya.

Temuan kami menunjukkan bahwa perokok cenderung membayar beberapa batang rokok dibandingkan konsumsi beras dan makanan pokok lainnya. Temuan awal sederhana kami mendukung temuan dalam kajian Lee dan Yi (2016) , juga Papadopoulu, dkk (2017) bahwa konsumsi rokok menunjukkan korelasi negatif terhadap pola makan. Lee dan Yi (2016) mengatakan bahwa perokok dewasa secara signifikan sedikit konsumsi buah buahan, sayur-mayur, dan susu/produk susu lainnya, dan mereka secara nyata lebih menyukai lebih banyak makanan cepat saji (*fast-food*) dibandingkan bukan perokok. Juga, Papadopoulu, dkk (2017) berpendapat bahwa perokok dewasa sedikit memilih makanan sehat dan disajikan secara higienis, di mana mereka cenderung memilih makanan-makanan dengan kandungan lemak yang relatif tinggi. Bila Lee dan Yi (2016) serta Papadopoulu, dkk (2017) menunjukkan bukti di dalam kajian studi kasus lingkup kecil, kami menunjukkan bukti yang sama menggunakan survey sosial ekonomi nasional yang terbukti bereputasi tinggi. Di dalam konteks Indonesia kontemporer, penelitian kami mengindikasikan bahwa perokok dewasa di rumah tangga miskin lebih cenderung tetap merokok dan mengirit untuk konsumsi beras dan makanan pokok lainnya.

KESIMPULAN DAN SARAN

Secara ringkas, kami dapat menyimpulkan 3 (tiga) poin utama dalam kajian ini. Pertama, Hasil penelitian kami menunjukkan bahwa rumah tangga miskin yang memiliki anggota rumah tangga perokok secara rata-rata mengkonsumsi beras relatif lebih sedikit dibandingkan rumah tangga yang tidak memiliki anggota rumah tangga perokok, baik yang termasuk kategori miskin maupun tidak. Hal ini mengindikasikan bahwa rumah tangga miskin lebih memprioritaskan konsumsi rokok dengan konsekuensi mengurangi konsumsi beras dan makanan pokok lainnya. Ini sedikit mencerminkan bahwa

semboyan orang Jawa "*mangan ora mangan kumpul (makan tidak makan (yang penting) kumpul*" ternyata kurang efektif untuk mencerminkan perilaku rumah tangga miskin perokok. Bagi rumah tangga miskin, semboyan tersebut di atas bila ditinjau dalam konteks konsumsi rokok dapat saja diplesetkan perokok menjadi " *ngudud ora ngudud kumpul* (merokok tidak merokok (yang penting) kumpul ".

Dengan mempertimbangkan cakupan penelitian kami, sepertinya fenomena "merokok atau tidak merokok yang penting kumpul " berlaku di hampir di seluruh daerah di Tanah Air, terutama di wilayah perdesaan. Seperti sudah menjadi kebiasaan turun temurun di dalam kultur masyarakat Indonesia, laki-laki dewasa senang duduk bercengkrama di kedai-kedai, baik di pagi, siang, dan sore hari. Acara kongkow-kongkow ini ini mendorong untuk mengobrol ngalor ngidul dengan ditemani kebiasaan merokok bareng-bareng. Terkadang, pengaruh pergaulan seperti ini menjadi pemicu seseorang menjadi perokok dengan alasan ketidakenakan menolak rokok yang ditawarkan teman-teman sepergaulannya (Nitcher, dkk, 2009).

Kedua, kajian kami menunjukkan korelasi negatif yang cukup signifikan antara program jaringan pengaman sosial (*social safety net*) Program Keluarga Harapan (PKH) dengan pola konsumsi beras dan konsumsi makanan pokok lainnya. Kecenderungan pengurangan belanja konsumsi ruta penerima PKH dapat saja kita artikan sebagai sebuah *alarm* kehati-hatian, terhadap salah peruntukan bantuan , yang salah satunya bisa saja diprioritaskan sebagai belanja konsumsi rokok dibandingkan dengan pengeluaran konsumsi beras dan makanan pokok lainnya. Meskipun ini tentu saja hanya dugaan awal yang perlu pembuktian kajian lanjut di penelitian mendatang, tetap saja kita harus ikut kritis mengevaluasi dampak kebijakan pemerintah, dalam hal pengentasan kemiskinan, yang kontra produktif dengan pengeluaran belanja ruta yang bukan peruntukan program.

Fakta bahwa Indonesia merupakan negara kelima penghasil tembakau terbesar, pemerintah pusat dan daerah diharapkan dapat menetapkan kebijakan-kebijakan yang tepat sehingga dapat menekan tingkat konsumsi rokok masyarakat. Selain itu, pemerintah pusat dan daerah diharapkan dapat bersinergi menciptakan kebijakan-kebijakan pro pengurangan konsumsi rokok masyarakat sekaligus mengurangi angka kemiskinan, seperti: pengawasan program jaring pengaman sosial (*social safety net*), yakni Program Keluarga Harapan, salah satunya. Agar program ini memiliki dampak positif dan tepat sasaran, misalnya dengan membuat payung hukum dan memberikan infrastruktur manusia dan sistem di dalam memperketat persyaratan penerima program ini tidak merokok selama jangka waktu tertentu.

Adapun penelitian mendatang dapat mempertimbangkan variabel-variabel penelitian baru dan metode mengukur evaluasi dampak (*impact evaluation*) di dalam mengukur dampak konsumsi rokok terhadap kemiskinan, misal dengan mempertimbangkan variabel untuk mengukur dampak evaluasi perluasan kawasan tanpa rokok (*non-smoking area*), perluasan kampanye anti rokok dengan gambar-gambar, video, poster mengenai bahaya rokok, dan juga ratifikasi regulasi tembakau dan cukai rokok dengan lebih realistis. Metode pengukuran dapat dipertimbangkan untuk penelitian lanjutan, misal dengan Prospensity Score Matching (PSM) di dalam mengukur evaluasi dampak (*impact evaluation*) perluasan kawasan tanpa rokok diperlukan untuk menekan pengaruh lingkungan di dalam peningkatan kebiasaan merokok masyarakat, dan juga menjadikan kualitas udara lingkungan lebih sehat. Juga diharapkan penelitian lanjutan dapat dilakukan untuk mengukur variabel kebijakan kebijakan lainnya di dalam ranah pengurangan konsumsi tembakau khususnya, dan pengurangan kemiskinan secara komprehensif, contohnya dampak kebijakan ratifikasi regulasi tembakau dan cukai rokok di dalam pengendalian penuh terhadap pengurangan tingkat konsumsi roko

Ketiga, meskipun belum cukup kuat bukti penelitian kami dalam menunjukkan

konsumsi rokok sebagai salah satu penyebab kemiskinan, pengurangan tingkat konsumsi rokok ke depan nampaknya bukan hanya urusan strategi kesehatan masyarakat, tetapi juga dapat dipertimbangkan sebagai strategi pengurangan kemiskinan. Meskipun, temuan awal di dalam penelitian ini masih menunjukkan kurang kuatnya peran konsumsi rokok dalam memperparah kemiskinan di Indonesia, tetapi ada indikasi perbaikan pengurangan konsumsi rokok untuk Ruta penerima Program Keluarga Harapan. Program PKH ini sempat digembar-gemborkan sebagai program keluarga sehat, yang melarang kepala ruta dan anggota ruta merokok. Namun, sampai saat ini, program anti rokok untuk penghargaan (*reward*)dan hukum (*punishment*) ruta PKH belum dikawal dengan ketat. Dan perlu bukti empiris lebih jauh dalam rentang penelitian lebih lama, untuk melihat apakah betul-betul efektif PKH di dalam mengedukasi ruta penerima dalam pengentasan kemiskinan, dan kampanye anti rokok. Sehingga, pengawasan tersebut bisa jadi cara pengawasan terhadap perilaku buruk anggota rumah tangga yang suka rokok cenderung menularkan kebiasaan tersebut kepada anggota ruta lainnya. Tentu perlu penelitian lanjutan di dalam melihat secara detail, peran signifikan konsumsi rokok sendirian dan konsumsi beras di dalam sumbangsihnya di dalam memperparah kemiskinan di tanah air ini. Penelitian lanjutan yang menggunakan metode penelitian lebih *robust* dan data lebih komprehensif, tidak hanya satu tahun, tetapi melibatkan panel tahun panjang, 5 sampai 10 tahun, agar hasil kajiannya menghasilkan hasil dan temuan yang robust, sehingga mampu menyumbangkan sumbangsih bagi kampanye anti rokok dan sekaligus strategi pemerintah di dalam pengentasan kemiskinan.

DAFTAR PUSTAKA

Bergström, J. (2004). Tobacco smoking and chronic destructive periodontal disease. *Odontology*, *92*(1), 1-8.

Block, S., & Webb, P. (2009). Up in smoke: Tobacco use, expenditure on food, and child malnutrition in developing countries. *Economic Development and Cultural Change*, *58*(1), 1-23.

BPS (2013). Pola Pengeluaran dan Konsumsi 2012. Badan Pusat Statistik.

BPS. (2015). Indonesia - Survei Sosial Ekonomi Nasional 2014 [Indonesia's national Socio-Economic Survey]. Indonesia's Central Bureau of Statistics: Jakarta.

Hayes, A. F., & Darlington, R. B. (2017). Regression analysis and linear models. Concepts, applications, and implementation. New York, London: Guilford Press (Methodology and the social sciences)..

Hu, T. W. (2008). Tobacco control policy analysis in China: economics and health (Vol. 12). World Scientific.

K Papadopoulou, S., N Hassapidou, M., Katsiki, N., Fachantidis, P., I Fachantidou, A., Daskalou, E., & P Deligiannis, A. (2017). Relationships Between Alcohol Consumption, Smoking Status and Food Habits in Greek Adolescents. Vascular Implications for the Future. *Current vascular pharmacology*, *15*(2), 167-173.

Kang, H. Y., Kim, H. J., Park, T. K., Jee, S. H., Nam, C. M., & Park, H. W. (2003). Economic burden of smoking in Korea. *Tobacco Control*, *12*(1), 37-44.

Korhonen, T., Smeds, E., Silventoinen, K., Heikkilä, K., & Kaprio, J. (2015). Cigarette smoking and alcohol use as predictors of disability retirement: a population-based cohort study. *Drug and alcohol dependence*, *155*, 260-266.

Kusumawardani, N., Tarigan, I., Suparmi, E. A., & Schlotheuber, A. (2018). Socio-economic, demographic and geographic correlates of cigarette smoking among Indonesian adolescents: results from the 2013 Indonesian Basic Health Research

(RISKESDAS) survey. *Global health action*, *11*(sup1), 54-62.

Lee, B., & Yi, Y. (2016). Smoking, physical activity, and eating habits among adolescents. *Western journal of nursing research*, *38*(1), 27-42.

Neuman, W. L. (2014). Social research methods: Qualitative and quantitative approaches: Pearson Education.

Nichter, M., Padmawati, S., Danardono, M., Ng, N., Prabandari, Y., & Nichter, M. (2009). Reading culture from tobacco advertisements in Indonesia. *Tobacco Control*, *18*(2), 98-107.

Olive, D. J. (2017). *Linear regression*. Springer.

Prasad, J. B., & Dhar, M. (2017). Tobacco use in India and its states: Burden of smoking and smokeless forms of tobacco (2015-25) and its predictors. *Journal of Cancer Policy*, *14*, 21-26.

Semba, R. D., Kalm, L. M., De Pee, S., Ricks, M. O., Sari, M., & Bloem, M. W. (2007). Paternal smoking is associated with increased risk of child malnutrition among poor urban families in Indonesia. *Public Health Nutrition*, *10*(1), 7-15.

Sung, H. Y., Wang, L., Jin, S., Hu, T. W., & Jiang, Y. (2008). Economic burden of smoking in China, 2000. In *Tobacco control policy analysis in China: Economics and health* (pp. 105-125).

Toukan, A. M. (2016). The Economic Impact of Cigarette Smoking on the Poor in Jordan. *Value in health regional issues*, *10*, 61-66.

World Health Organization. (2012). *Global adult tobacco survey: Indonesia report 2011*. WHO Regional Office for South-East Asia.

World Health Organization. (2015). WHO global report on trends in prevalence of tobacco smoking 2015. World Health Organization.

PENGELOMPOKAN PENGGUNA SITUS WEB BPS MELALUI TEKNIK *BIBLIOMETRIC* DAN ANALISIS KORESPONDENSI

Toza Sathia Utiayarsih[1], Jadi Suprijadi[2], Bernik Maskun[3]

[1]Politeknik Statistika STIS
[2,3]Universitas Padjajaran
e-mail: [1]toza@stis.ac.id

Abstrak

Salah satu upaya pemenuhan program percepatan (*quick wins*) terhadap produk BPS yang benar-benar dapat menyentuh kebutuhan para pengguna data adalah dengan melakukan segmentasi terhadap pengguna data. Segmentasi terhadap pengguna situs web BPS sebagai salah satu bentuk segmentasi terhadap pengguna data, sesuai program percepatan. Ukuran data pengguna web sangat besar dan berupa data teks sehingga tidak dapat langsung dianalisis melalui aplikasi statistik yang tersedia, maka perlu dilakukan suatu teknik untuk data pengguna web dengan menggunakan teknik *bibliometric*. Teknik tersebut mengubah data teks menjadi format numerik, selanjutnya dibuat menjadi matriks distribusi frekuensi. Matriks digunakan pada analisis korespondensi untuk pengelompokkan pengguna situs web. Hasil dari analisis pengguna situs web BPS yang diwakili oleh alamat IP dapat dikelompokkan dengan halaman yang diakses berdasarkan asal negara, sehingga didapatkan segmentasi pengguna data situs web BPS antara negara dan halaman yang diakses.

Kata kunci: *Data Mining, text mining, bibliometric, web mining*, analisis korespondensi

Abstract

The effort to fulfill one of quick wins program for BPS products that really can fulfill the needs of data users is by segmenting data users. Segmentation of BPS website users as a form of segmentation of data users, according to quick wins program. The size of web user data is very large and in the form of text data so that it cannot be directly analyzed through available statistical applications, it is necessary to do a technique for web user data using bibliometric techniques. This technique converts text data into numeric format, then it is made into a frequency distribution matrix. The matrix is used in correspondence analysis for grouping website users. The results of the analysis of BPS website users represented by IP addresses can be grouped with pages accessed based on national origin, so that segmentation users of BPS website data between the country and the page are accessed can be obtained.

***Keywords**: Data Mining, text mining, bibliometric, web mining, correspondence analysis*

PENDAHULUAN

Badan Pusat Statistik (BPS) selalu berupaya untuk melakukan perubahan dan reformasi yang mendasar terhadap sistem penyelenggaraan kegiatan statistik, melalui pembangunan profil dan perilaku aparatur BPS yang profesional, berintegritas, bertanggung jawab, serta mampu memberikan pelayanan prima kepada publik. BPS sebagai lembaga pemerintah non-kementerian mempunyai tugas untuk menyediakan data dan informasi statistik yang berkualitas, serta dituntut untuk melayani berbagai kepentingan pengguna data. Sejalan dengan keinginan reformasi birokrasi, ke depan BPS harus mampu menghasilkan data yang berkualitas, yang didukung oleh SDM profesional dan infrastruktur yang lebih modern.

Untuk membangun kepercayaan masyarakat perlu diupayakan suatu program percepatan (*quick wins*) terhadap produk BPS yang benar-benar dapat menyentuh kebutuhan para pengguna data. Program *quick wins* ini dipilih dengan memperhatikan produk statistik yang memiliki daya ungkit tinggi, inovatif, dan merupakan terobosan yang terkait dengan produk utama BPS. Program *quick wins* yang memenuhi kriteria tersebut di atas antara lain: (i). Peningkatan Kepuasan Pelanggan, (ii). Penyempurnaan Pelayanan Statistik yang terdiri pelayanan Elektronik (*e-Services*) dan pelayanan statistik terpadu yang menggabungkan pelayanan perpustakaan (digital dan non-digital), konsultasi statistik, toko buku (*e-Shop*) dan pelayanan lainnya, dan (iii). Membangunan *Advanced Release Calendar*.

Dalam upaya memenuhi kriteria tersebut muncul salah satu tujuannya yaitu segmentasi pengguna data baik melalui pelayanan langsung maupun pelayanan elektronik (*e-Service*) seperti situs web BPS (Laporan Reformasi Birokrasi Badan Pusat Statistik, 2011). Sejalan dengan hal tersebut, perlu diketahui tentang pola pengguna situs web itu sendiri dalam rangka mendapatkan segmentasi pengguna yang tepat. Untuk menganalisis pola pengguna situs web dibutuhkan suatu instrumen yang dapat menjembatani antara pengguna dengan pengelola situs web, yaitu melalui *web usage session*, yang merupakan interaksi antara pengguna dan *web server* dalam satu periode waktu tertentu yang berisi halaman web yang dikunjungi.

Data mining merupakan pendekatan yang sangat berguna pada aspek pengolahan data dan penelaahan penemuan. Pada dasarnya, data mining mengacu pada ekstraksi informasi data dalam jumlah besar, yang memiliki berbagai macam bentuk atau jenis data, seperti data transaksi pada aplikasi web (pembelian online, layanan konsumen, dll). Dalam sepuluh tahun terakhir, menurut Xu (2010), *data mining* berhasil masuk ke dalam dunia penelitian manajemen data web, seperti dokumen web, struktur tautan web, transaksi pengguna web, dan web semantics menjadi target penelaahan. Jelas bahwa informasi yang dapat digali dari berbagai jenis data web dapat membantu dalam menemukan hubungan antara berbagai obyek dalam web sehingga dapat meningkatkan manajemen data web.

Menurut wikipedia, *web mining* merupakan suatu aplikasi bagian dari *data mining* yang menggali pola-pola yang tersedia di dalam web itu sendiri. Jadi antara *data mining* dan *web mining* hanya berbeda dalam hal target data yang dianalisis. Data mining umumnya menganalisis data yang berasal dari OLTP (Online Transactional Process) dan data transaksi lainnya. Sedangkan *web mining* target analisisnya adalah data dari web, seperti data akses pengunjung, struktur halaman web, format halaman web dan sebagainya. Berdasarkan target analisisnya, *web mining* dibagi menjadi 3 (tiga) bagian, yaitu: (i). *web content mining*, (ii). *web structure mining*, dan (iii). *web usage mining*.

Menurut Srivastava (2000), *web usage mining* merupakan teknik *data mining* yang menggambarkan pola penggunaan dari halaman web, dalam rangka memahami dan meningkatkan pelayanan kebutuhan dari aplikasi berbasis web. Sumber data utama dari *web usage mining* adalah *server logs* dan *browser logs*.

Tabel 1. Kategori Halaman yang Diberi Label Kode Angka

Nama Halaman Web	Kode	Nama Halaman Web	Kode
Beranda	1	Publikasi BPS	9
Tentang BPS	2	Berita Resmi BPS	10
Rencana Strategis BPS	3	Unduh	11
Pusat Layanan	4	Berita	12
Istilah Statistik	5	Info Lelang	13
Jabatan Fungsional	6	Subyek Statistik	14
Sistem Rujukan Statistik	7	*Website* BPS Provinsi	15
Sekolah Tinggi Ilmu Statistik	8		

Teknik *server log analysis* digunakan jika memiliki akses penuh terhadap suatu situs web dan *server* web yang digunakan. Karena data tersimpan di dalam file, maka data log relatif mudah dikelola. Data yang tercatat pada *log server* memiliki format teks dalam jumlah yamg sangat besar. Data tersebut merupakan data tidak terstruktur, tetapi memungkinkan untuk diubah menjadi bentuk bibliografi sehingga bisa diterapkan metode untuk mengolahnya melalui teknik *bibliometric*. *Software bibliometric* sebagai alat untuk analisis informasi dalam jumlah yang besar berkembang dengan *output* format yang bervariasi, misalnya, distribusi frekuensi, matriks, peta, dan network (Supriyadi, 2011).

Ukuran data web sangat besar dan berupa data teks sehingga tidak dapat langsung dianalisis melalui *software* statistik biasa, maka perlu dilakukan suatu teknik untuk data pengguna web sehingga dapat berubah menjadi format numerik, seperti matriks distribusi frekuensi yaitu melalui teknik *bibliometric*. Selanjutnya hasil yang didapatkan melalui *bibliometric* dapat digunakan pada analisis statistik untuk pengelompokkan pengguna situs web BPS sehingga dapat dilihat pola segmentasi dari pengguna.

METODE

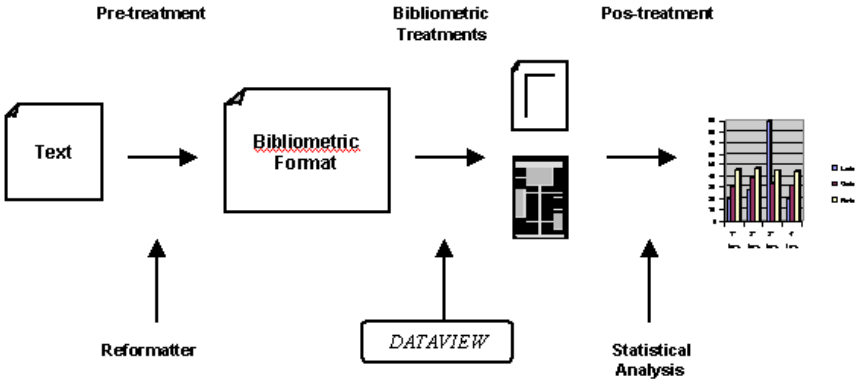
Berdasarkan uraian permasalahan yang disampaikan pada pendahuluan, dapat dirumuskan dalam penelitian ini adalah bagaimana mengolah data pengguna web pada halaman situs web BPS dengan format teks menjadi numerik, sehingga dapat dilakukan analisis terhadap data tersebut, dan diketahui pola segmentasi pengguna melalui salah satu analisis statistik. Dengan tahapan mengubah data pengguna situs web BPS yang berbentuk teks menjadi format numerik melalui teknik *bibliometric* yang dapat menghasilkan matriks kontingensi. Kemudian matriks tersebut bisa dilanjutkan dengan analisis statistik menggunakan analisis korespondensi. Sehingga didapatkan pola segmentasi pengguna situs web BPS melalui pengelompokkan berdasarkan asal negara.

1. Data *Web Usage* Situs Web BPS

Kategori halaman yang digunakan pada penelitian ini berasal dari peta situs web BPS yang merupakan kerangka dasar dalam sebuah situs web yang berisi informasi mengenai halaman-halaman yang ada dalam situs. Halaman pada situs web BPS terdapat 15 kategori, yaitu: “Beranda”, “Tentang BPS”, “Rencana Strategis BPS”, “Pusat Layanan”, “Istilah Statistik”, “Jabatan Fungsional”, “Sistem Rujukan Statistik”, “Sekolah Tinggi Ilmu Statistik”, “Publikasi BPS”, “Bertita Resmi BPS”, “Unduh”, “Berita”, “Info Lelang”, “Subyek Statistik”, “Website BPS Provinsi”.

Setiap kategori halaman direpresentasikan dengan label *integer*. Contohnya, “Beranda” diberi kode 1, “Tentang BPS” diberi kode 2, “Rencana Strategis BPS” diberi kode 3, dan seterusnya, seperti terlihat dalam Tabel 1.

Sumber data sebagian besar *web usage mining* adalah *web server log*, yang menyediakan data mentah untuk



Gambar 1. Posisi Dataview dalam Rantai Pengolahan *Bibliometric*

(Sumber: Rostaing, 2000, dalam Tarapanoff et al, 2001)

mengidentifikasi kumpulan data web atau web usage session. *Web server log* berisi catatan akses dari pengguna. Setiap *record* mewakili sebuah halaman yang diakses oleh pengguna dan umumnya berisi alamat IP (*Internet Protocol*) pengguna, tanggal dan waktu akses diterima, alamat URL yang diakses, kode balasan dari *server* yang menunjukkan status akses, dan ukuran file (byte) dari halaman yang diakses sempurna.

2. Teknik *Bibliometric*

Data text yang didapatkan dianalisis dengan menggunakan proses *bibliometric*. Tahapan Teknik *bibliometric* seperti yang dapat dilihat pada Gambar 1 adalah sebagai berikut:

1. Data web berupa text file yang tidak terstruktur diubah menjadi database terstruktur. Data yang diambil dari web log pada server perlu disiapkan sebelum memasuki proses pengolahan atau biasa disebut sebagai *preprocessing*. Proses ini terdiri dari 2 (dua) tahapan, yaitu: pemilihan data dan transformasi data menjadi data yang terstruktur. Hasil dari proses ini adalah database web server log.
2. *Database web server log* terdiri dari 5 field, yaitu: *Internet Protocol* (IP), waktu, halaman, status, dan ukuran.
3. Data Internet Protocol (IP) dan halaman ditransformasi menjadi format *bibliometric*. Data format *bibliometric* terdiri dari field nomor record (NO), alamat IP pengguna/*Internet Protocol* (IP), dan halaman yang diakses (HAL). Data ini berupa text file sehingga lebih mudah dikelola dalam proses bibilometric.
4. Data format *bibliometric* diolah dengan proses bibilometric sehingga menghasilkan *output* tabel kontingensi dengan baris adalah field Internet Protocol (IP) dan kolom adalah field halaman.
5. Tabel kontingensi disederhanakan dengan mengklasifikasikan alamat IP pengguna/ *Internet Protocol* (IP) berdasarkan negara.
6. Tabel kontingensi yang telah disederhanakan kemudian dianalisis dengan menggunakan analisis statistik.

3. Analisis Korespondensi

Tabel kontingensi yang dihasilkan melalui teknik *bibliometric* kemudian dianalisis dengan menggunakan analisis statistik, dalam penelitian ini digunakan analisis korespondensi sederhana. Menurut Izenman (2008), proses dari analisis sebagai berikut:

Tabel Kontingensi Dua Arah

Data kategorik adalah data yang dikumpulkan dari hasil hitungan yang disusun dalam tabel kontingensi. Sebuah tabel kontingensi dua arah ($r \times s$) dengan r baris (diberi label $A_1, A_2, \ldots, A_r$) dan s kolom (diberi label $B_1, B_2, \ldots, B_s$) terdiri dari rs sel. Sel ke-ij, n_{ij}, mewakili frekuensi yang diamati untuk baris kategori A_i dan kolom kategori B_j, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, s$. Total marjinal baris ke-i adalah $n_{i+} = \sum_{j=1}^{s} n_{ij}$, $i = 1, 2, \ldots, r$, dan total marjinal

Tabel 2. Tabel Kontingensi Dua Arah yang Menjelaskan Frekuensi Sel Pengamatan, Total Marjinal Baris & Kolom, dan Jumlah Sampel

Variabel Baris	Variabel Kolom						Total Baris
	B_1	B_2	…	B_j	…	B_s	
A_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	n_{1+}
A_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	n_{2+}
…	…	…		…		…	…
A_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	n_{i+}
…	…	…		…		…	…
A_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	n_{r+}
Total Kolom	n_{+1}	n_{+2}	…	n_{+j}	…	n_{+s}	n_{++}

kolom ke-j adalah $n_{+j} = \sum_{i=1}^{r} n_{ij}$, *j = 1, 2, ... , s*. Jika $n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$ individu diklasifikasikan oleh kategori baris dan kolom, kemudian Tabel 3, yang juga disebut tabel korespondensi, menunjukkan frekuensi sel, total marjinal, dan total ukuran sampel.

Notasi π_{ij} merupakan peluang bahwa seorang individu memiliki karakteristik A_i dan B_j, *i = 1, 2, ... , r, j = 1, 2, ... , s*. Dengan asumsi bahwa baris variabel *A* dan kolom variabel *B* adalah independen, sehingga $\pi_{ij} = \pi_{i+}\pi_{+j}$, dengan $\pi_{i+} = \sum_j \pi_{ij}$ dan $\pi_{+j} = \sum_i \pi_{ij}$, untuk semua *i = 1 2, ... , r* dan *j = 1, 2, ... , s*. Secara umum yang ingin dilihat adalah apakah *A* dan *B* memang variabel independen. Sebuah pertanyaan dapat diajukan sebagai alternatif dalam hal homogenitas dari distribusi peluang baris atau kolom, yaitu, apakah semua baris memiliki distribusi peluang yang sama di setiap kolom, atau sebaliknya, semua kolom memiliki distribusi peluang yang sama di setiap baris.

Variabel Dummy Baris dan Kolom

Pada tabel kontingensi dua arah, dapat melihat hubungan antara kategori baris dan kategori kolom seperti pada Tabel 2.

Merubah tabel kontingensi N menjadi "matriks korespondensi" sebagaimana Tabel 3

Jarak Chi-Square

Pada analisis korespondensi, penting untuk menggambarkan jarak diantara profil baris (yaitu baris pada matriks Pr) atau diantara profil kolom (yaitu kolom pada matriks Pc). Untuk mengukur jarak ini digunakan ukuran chi-squared.

1. Jarak Baris

Jika profil baris ke-i dan ke-i' adalah **ai** dan **ai'**, maka **ai - ai'** adalah s-vektor dengan elemen ke-j $n_{ij}/n_{i+} - n_{i'j}/n_{i'+}$. Kuadrat dari jarak *chi-squared* diantara $\mathbf{a_i}$ dan $\mathbf{a_i'}$ sebagai berikut:

$$d^2(a_i, a_i') = (a_i - a_i')^T D_c^{-1}(a_i - a_i')$$

$$= \sum_{j=1}^{s} \frac{n}{n_{+j}}\left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i'+}}\right)^2 \quad (1)$$

Perhatikan Persamaan (1), massa kolom ke-j (n_{+j}/n) masuk ke dalam persamaan tersebut berbanding terbalik dengan kuadrat jarak dari profil baris. Sehingga jumlah observasi (*n*) berpengaruh terhadap jarak antar profil baris.

Perhatikan bahwa **c** adalah sentroid baris. Matriks berukuran (*r x s*) dari titik pusat profil baris $\mathbf{P_r} - \mathbf{1_r c^T}$ dengan $\mathbf{P_r} = \mathbf{D_r^{-1}P}$, memiliki baris ke-I $(\mathbf{a_i} - \mathbf{c})^\mathbf{T}$ dengan elemen ke-j $n_{i+}^{-1}(n_{ij} - n_{i+}n_{+j}/n)$, *i = 1, 2, ... , r, j = 1, 2, ... , s*. Sehingga kuadrat dari jarak *chi-squared* antara $\mathbf{a_i}$ dan **c** adalah:

$$d^2(a_i, c) = (a_i - c)^T D_c^{-1}(a_i - c)$$

$$= \frac{1}{n_{i+}}\sum_{j=1}^{s} \frac{n}{n_{i+}n_{+j}}\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2$$

(2)

Tabel 3. Matriks Korespondensi Menjelaskan Frekuensi Relatif dari Sel Pengamatan, Total Marjinal Baris, dan Total Marjinal Kolom terhadap n

Variabel Baris	**Variabel Kolom**						**Total Baris**
	B_1	B_2	…	B_j	…	B_s	
A_1	p_{11}	p_{12}	…	p_{1j}	…	p_{1s}	p_{1+}
A_2	p_{21}	p_{22}	…	p_{2j}	…	p_{2s}	p_{2+}
…	…	…		…		…	…
A_i	p_{i1}	p_{i2}	…	p_{ij}	…	p_{is}	p_{i+}
…	…	…		…		…	…
A_r	p_{r1}	p_{r2}	…	p_{rj}	…	p_{rs}	p_{r+}
Total Kolom	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+s}	1

Penjumlahan dari semua profil baris pada Persamaan (2) menjadi:

$$n\sum_{i=1}^{r} p_{i+}d^2(a_i,c) = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(n_{ij}-\frac{n_{i+}n_{+j}}{n}\right)^2 \Big/ \left(\frac{n_{i+}n_{+j}}{n}\right) \quad (3)$$

dengan statistik uji *Pearson Chi-Squared* sebagai berikut:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (4)$$

Frekuensi sel hasil observasi O_{ij} dan frekuensi sel harapan E_{ij} (dengan asumsi baris dan kolom independen) sebagai berikut:

$$O_{ij} = n_{ij}, \; E_{ij} = \frac{n_{i+}n_{+j}}{n} \quad (5)$$

dengan *i = 1, 2, ... , r, j = 1, 2, ... , s*. Di bawah asumsi sampel acak, χ^2 pada Persamaan (4) mendekati *distribusi chi-squared* (χ^2) pada sampel besar dengan derajat bebas (*r*-1)(*s*-1).

2. Jarak Kolom

Sama seperti jarak baris, untuk jarak kolom, jika profil kolom ke-j dan ke-j' adalah $\mathbf{b_j}$ dan $\mathbf{b_j'}$, maka $\mathbf{b_j} - \mathbf{b_j'}$ adalah *r*-vektor dengan elemen ke-j $n_{ij}/n_{+j} - n_{ij'}/n_{+j'}$. Kuadrat dari jarak *chi-squared* diantara $\mathbf{b_j}$ dan $\mathbf{b_j'}$ sebagai berikut:

$$d^2(b_j,b_j') = (b_j-b_j')^T D_r^{-1}(b_j-b_j')$$

$$= \sum_{i=1}^{r} \frac{n}{n_{i+}}\left(\frac{n_{ij}}{n_{+j}}-\frac{n_{ij'}}{n_{+j'}}\right)^2 \quad (6)$$

Kuadrat dari jarak *chi-squared* antara $\mathbf{b_j}$ dan **r** adalah:

$$d^2(b_j,r) = (b_j-r)^T D_r^{-1}(b_j-r)$$

$$= \frac{1}{n_{+j}}\sum_{i=1}^{r} \frac{n}{n_{i+}n_{+j}}\left(n_{ij}-\frac{n_{i+}n_{+j}}{n}\right)^2 \quad (7)$$

Penjumlahan dari semua profil kolom pada Persamaan (7) menjadi:

$$n\sum_{j=1}^{s} p_{+j}d^2(b_j,r) = \chi^2 \quad (8)$$

dengan χ^2 seperti pada Persamaan (4).

Sehingga rata-rata tertimbang dari kuadrat jarak *chi-squared* pada semua profil baris terhadap sentroid baris atau pada semua profil kolom terhadap sentroid kolom (dengan penimbang massa baris/massa kolom) adalah χ^2/n. Jika baris dan kolom independen, maka χ^2/n akan kecil, sejalan dengan $p_{i+}d^2(a_i,c)$ dan $p_{+j}d^2(b_j,r)$.

Di sisi lain, jika χ^2/n besar, berarti minimal ada satu dari $p_{i+}d^2(a_i,c)$ atau $p_{+j}d^2(b_j,r)$ akan besar. Informasi ini penting untuk menentukan apakah independensi dalam tabel terpenuhi atau tidak. Bandingkan matriks tersebut dengan matriks **N** = (O_{ij}).

Total Inersia

Dengan menggunakan dummy variable untuk mewakili tabel kontingensi dua arah, memungkinkan untuk melihat suatu masalah sebagai suatu kasus khusus dari analisis kanonik. Bagaimanapun situasinya berbeda, bahwa apabila menggali struktur korelasi antara dua set dari vektor data statistik, akan berhadapan dengan struktur korelasi dari dua set dummy variable.

Tabel 4. Struktur Data *Web Server Log* Hasil Pemilihan Data

Nama Field	Deskripsi	Tipe Data
IP	Alamat IP Pengguna	*Text*
WAKTU	Tanggal dan Jam Akses	*Date*
URL	URL yang Diakses	*Text*
STATUS	Status Akses	*Numeric*
UKURAN	Ukuran Halaman yang Diakses	*Numeric*

Jika nilai dari χ^2 sangat besar, asumsi independensi dari variansi baris dan kolom pada tabel kontingensi tidak terpenuhi (ditolak). Selanjutnya menentukan dimana deviasi dari keindependenan terjadi. Nilai dari χ^2/n mengacu pada nilai total inersia pada tabel kontingensi. Nilai inersia utama merupakan persentase dari total variansi yang dijelaskan oleh beberapa komponen utama, yang biasanya terdiri dari 2 (dua) atau 3 (tiga) komponen utama.

Tampilan Grafis

Pada analisis korespondensi, dapat dipilih hanya dengan menganalisis profil baris atau profil kolom, atau menganalisis keduanya. Tampilan grafis dibentuk dengan membuat plot dari koordinat baris dan koordinat kolom yang merupakan *scatterplot*. Tampilan grafis terdiri dari 2 (dua) jenis, yaitu:

1. *Symetric map*: Baik koordinat baris dan koordinat kolom, keduanya dianggap sebagai koordinat utama.
2. *Asymetric map*: Koordinat baris (atau kolom) dianggap sebagai koordinat utama, sedangkan yang lainnya dianggap sebagai koordinat biasa.

Secara garis besar, titik yang terlihat dekat diantara satu sama lain menunjukkan hubungan antar kategori. Lebih jelasnya sebagai berikut:

1. Jika titik pada baris dekat, maka baris tersebut memiliki distibusi bersyarat yang sama pada setiap kolom.
2. Jika titik pada kolom dekat, maka kolom tersebut memiliki distibusi bersyarat yang sama pada setiap baris.
3. Jika titik pada baris dan kolom dekat, maka hal tersebut menyatakan bahwa deviasi tertentu dari independensi atau baris dan kolom menyimpang dari independensi.

HASIL DAN PEMBAHASAN

1. *Preprocessing* Data

Preprocessing data terdiri dari 2 (dua) tahapan, yaitu: pemilihan data dan transformasi data.

Pemilihan Data

Data web log pada server situs web BPS memiliki ukuran yang sangat besar dan berupa text file. Sehingga perlu ditentukan batasan dari segi waktu untuk analisis data pada penelitian ini. Pada penelitian ini, data yang dianalisis adalah data web server log bulan November 2011. Karena keterbatas *software*, data yang diproses adalah data 3 (tiga) hari pada bulan tersebut, yaitu tanggal 1 (satu), 2 (dua) dan 3 (tiga). Data text file kemudian dimasukkan ke dalam *database web server log* agar menjadi file yang terstruktur. Pada tahapan ini dilakukan juga proses *cleaning data* untuk menghilangkan data yang berulang (redundant) dan pemilihan data yang berstatus berhasil melakukan akses. Ukuran *database* untuk 3 hari sebanyak 61.759 *record*. Struktur data *web server log* hasil pemilihan data dapat dilihat pada Tabel 4.

Proses pemilihan data menggunakan program yang dirancang dengan menggunakan bahasa pemrograman Microsoft Visual Basic.NET yaitu melalui fasilitas tombol "Cleaning".

Tranformasi Data

Data *web server log* yang sudah dipilih masih belum sesuai dengan struktur data untuk analisis pada penelitian ini. Struktur data yang dimaksud adalah

Tabel 5. Struktur Data Hasil Transformasi

Nama Field	Deskripsi	Tipe Data
IP	Alamat IP Pengguna	*Text*
HALAMAN	Halaman yang Diakses	*Text*

Tabel 6. Matriks Kontingensi Hasil dari *Bibliometric*

IP	11	14	1	10	2	9	12	5	13	4	3	6	15	8
193.130.130.153	87	2342	916	280	1	0	0	0	0	0	0	0	0	0
223.255.225.75	2398	0	0	0	0	0	0	0	0	0	0	0	0	0
50.115.185.87	1524	108	499	0	0	2	0	0	0	0	0	0	0	0
66.249.69.24	68	422	10	616	174	40	168	17	6	1	10	1	1	1
69.191.249.202	0	548	132	563	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
103.10.169.235	1	0	0	0	0	0	0	0	0	0	0	0	0	0
101.255.16.202	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10.5.3.21	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1.113.17.82	0	0	1	0	0	0	0	0	0	0	0	0	0	0

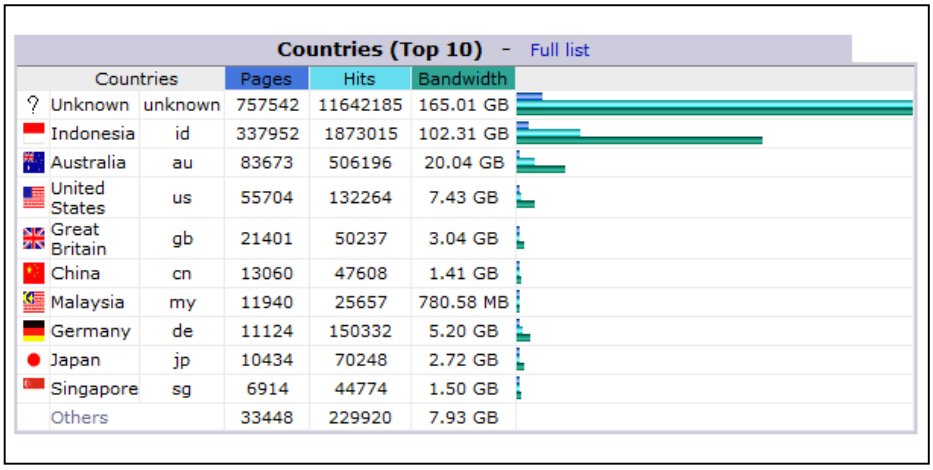
struktur data yang menggambarkan pola akses data berdasarkan halaman yang diakses. Sehingga perlu dilakukan proses transformasi data, yaitu dengan mentransformasi alamat URL yang diakses menjadi halaman-halaman yang ada dalam situs. Halaman pada situs web BPS terdapat 15 kategori, yaitu: "Beranda", "Tentang BPS", "Rencana Strategis BPS", "Pusat Layanan", "Istilah Statistik", "Jabatan Fungsional", "Sistem Rujukan Statistik", "Sekolah Tinggi Ilmu Statistik", "Publikasi BPS", "Berita Resmi BPS", "Unduh", "Berita", "Info Lelang", "Subyek Statistik", "Website BPS Provinsi". Setiap kategori halaman direpresentasikan dengan label integer. Contohnya, "Beranda" diberi kode 1, "Tentang BPS" diberi kode 2, "Rencana Strategis BPS" diberi kode 3, dan seterusnya.

Struktur data hasil transformasi dapat dilihat pada Tabel 5.

Data ditransformasi menjadi bentuk IP berdasarkan halaman web yang diakses. Setelah data terbentuk, diperhatikan bahwa terdapat beberapa pengguna memiliki karakteristik khusus, yaitu pengguna yang mengakses langsung pada halaman tertentu dan mengaksesnya berulangkali. Data web server log hasil transformasi data kemudian dirubah lagi menjadi format bibliografi yang kemudian akan digunakan dalam teknik biblometrik. Data diubah ke dalam field-field yang berisi form dalam format text.

2. Penerapan Teknik *Bibliometric*

Berdasarkan data yang telah dirubah bentuknya menjadi format bibliografi, maka selanjutnya diterapkan teknik *bibliometric* untuk mendapatkan bentuk yang dapat dianalisis lebih lanjut, dari format aslinya yang berupa teks. Pada tahap ini, data diolah dengan *software* khusus untuk format bibliografi. Selanjutnya data dalam dalam format bibliografi diubah bentuknya oleh *software* menjadi field "IP" (alamat IP) dan "HAL" (Halaman yang Diakses), sedangkan isi dari field tersebut menjadi form yang kemudian akan diekstrak dan dipasangkan (pair) antar field melalui suatu proses hingga menghasilkan matriks kontingensi dua arah berukuran 2.618 baris yang merepresentasikan pengguna, dalam hal ini IP, dan 14 kolom yang merepresentasikan halaman yang diakses, dalam hal ini kategori halaman seperti pada Tabel 1. Matriks tersebut dapat dilihat pada Tabel 6.



Countries (Top 10) - Full list

Countries		Pages	Hits	Bandwidth
Unknown	unknown	757542	11642185	165.01 GB
Indonesia	id	337952	1873015	102.31 GB
Australia	au	83673	506196	20.04 GB
United States	us	55704	132264	7.43 GB
Great Britain	gb	21401	50237	3.04 GB
China	cn	13060	47608	1.41 GB
Malaysia	my	11940	25657	780.58 MB
Germany	de	11124	150332	5.20 GB
Japan	jp	10434	70248	2.72 GB
Singapore	sg	6914	44774	1.50 GB
Others		33448	229920	7.93 GB

Gambar 2. 10 Negara Tertinggi yang Mengakses Situs Web BPS

Tabel 7. Kode Negara Klasifikasi Alamat IP

Negara	Kode	Negara	Kode
Indonesia	1	Malaysia	6
Australia	2	Jerman	7
USA	3	Jepang	8
Inggris	4	Singapura	9
Cina	5	Lainnya	10

Kategori halaman yang muncul pada matriks ini hanya 14 dari keseluruhan 15 kategori, hal ini disebabkan salah satu kategori tersebut, yaitu kode "7" tidak ada yang mengakses dalam 3 hari data yang dimasukkan ke dalam pengolahan. Pada *software* apabila isian kosong, maka otomatis akan hilang.

3. Pengklasifikasian Pengguna Data

Matriks yang dihasilkan tersebut memiliki ukuran yang cukup besar. Sehingga untuk menyesuaikan dengan tujuan segmentasi yang hendak dicapai, maka dilakukan pengklasifikasian pada pengguna data, dalam hal ini IP, berdasarkan negara asal pemilik IP. Negara yang dimunculkan pada klasifikasi ini diambil berdasarkan 9 (sembilan) negara yang memiliki frekuensi tertinggi mengakses situs web BPS di Bulan November 2011. Negara-negara yang memiliki frekuensi kecil masuk ke dalam klasifikasi lainnya. Seperti yang terlihat pada Gambar 2.

Gambar 2 di atas didapat dari statistik web pada situs web BPS pada bulan November 2011. Negara dengan kategori unknown adalah alamat IP yang tidak dapat ditelusuri asal negaranya, ada beberapa IP berbayar yang dirahasiakan kepemilikannya, atau yang dikenal dengan private IP number, dan dimasukkan ke dalam klasifikasi lainnya. Sehingga alamat IP berdasarkan asal negara terbagi menjadi 10 (sepuluh) klasifikasi yang dapat dilihat pada Tabel 7.

Berdasarkan klasifikasi tersebut maka seluruh alamat IP yang ada pada tabel kontingensi yang sudah didapat pada Tabel 6 ditransformasi berdasarkan asal negaranya. Untuk mengklasifikasikan alamat IP digunakan program yang dapat dilihat pada Gambar 4 melalui fasilitas tombol "Country Class". Database asal negara diperoleh dari situs web tentang lokasi alamat IP, yaitu http://www.ipaddresslocation.org. Dari transformasi tersebut didapat rekapitulasi akses web BPS berdasarkan negara pada Tabel 8.

Tabel 8 menunjukkan banyaknya akses setiap negara ke web BPS, terlihat bahwa banyaknya akses dari dalam negeri (kode negara 1) sebesar 40,42%. Sedangkan banyaknya akses dari luar negeri (kode

Tabel 8. Rekap Negara yang Mengakses Web BPS

Kode Negara	Frekuensi Akses	Persentase
1	24965	40,42
2	947	1,53
3	11571	18,74
4	4533	7,34
5	3346	5,42
6	626	1,01
7	991	1,60
8	1746	2,83
9	2019	3,27
10	11015	17,84
Total	61759	100

Tabel 9. Tabel Kontingensi Setelah Diklasifikasikan Berdasarkan Negara

Kode	H11	H14	H1	H10	H2	H9	H12	H5	H13	H4	H3	H6	H15	H8
1	21307	1049	1383	771	197	57	12	68	93	17	5	6	0	0
2	624	140	126	9	26	12	0	3	0	2	4	0	1	0
3	931	4228	2486	2760	461	175	333	134	24	16	17	3	2	1
4	374	2579	1183	357	26	9	0	3	0	1	1	0	0	0
5	259	1804	961	109	88	64	7	19	3	5	9	17	1	0
6	115	336	116	28	17	6	0	4	1	2	1	0	0	0
7	687	117	99	15	51	8	0	3	0	6	3	2	0	0
8	574	298	563	204	49	43	0	13	0	2	0	0	0	0
9	439	471	928	84	44	36	0	10	0	2	5	0	0	0
10	4571	2427	3085	249	261	145	143	104	4	12	14	0	0	0

negara 2 s/d 10) sebesar 59,58%. Akses luar negeri paling banyak berasal dari negara Amerika Serikat (kode 3) sebesar 18,74%. Hasil transformasi dari tabel frekuensi yang diklasifikasikan menurut negara dapat dilihat pada Tabel 9.

Perhatikan Tabel 9, kolom Kode menunjukkan kode negara. Sedangkan kolom H11 s/d H8 menunjukkan halaman yang diakses.

4. Penerapan Analisis Korespondensi

Analisis korespondensi dapat digunakan untuk mengetahui kedekatan hubungan antar kategori dari 2 (dua) variabel. Berdasarkan Tabel 9 terdapat 2 (dua) variabel yang dianalisis yaitu negara yang mengakses (Kode) dan halaman yang diakses (H1 s/d H15). Format data pada Tabel 9 diubah terlebih dahulu ke dalam format data yang sesuai dengan *software* statistik seperti pada Tabel 10.

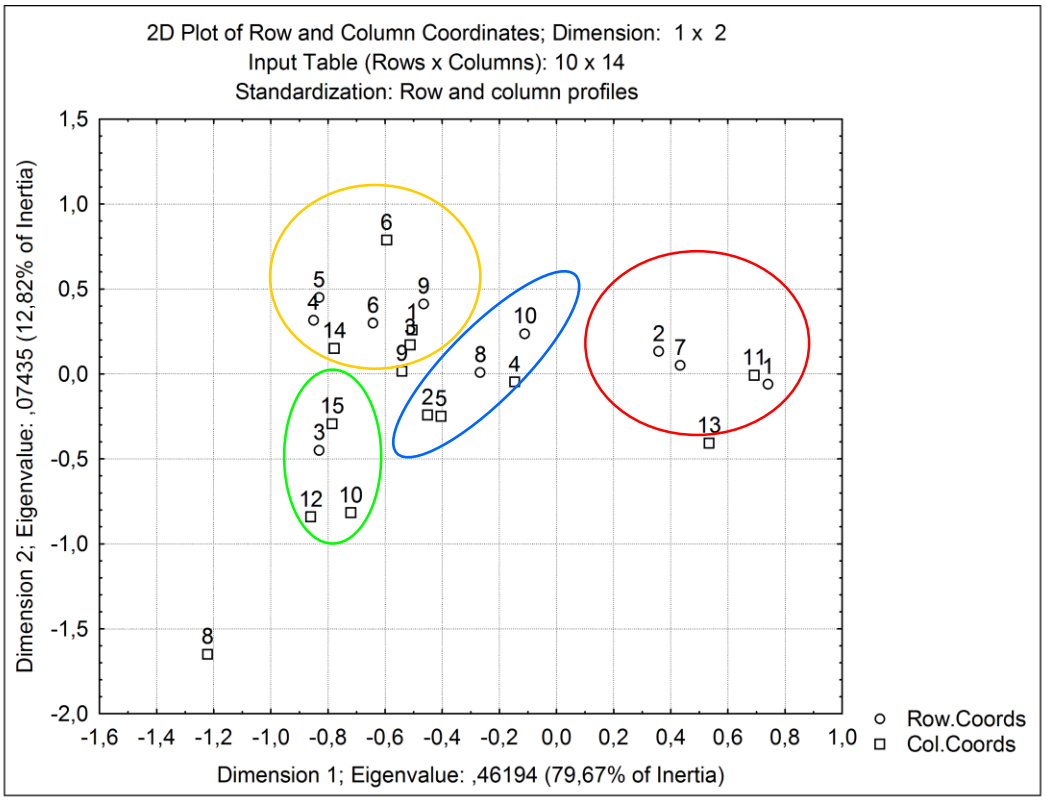
Data di atas kemudian diolah dengan *software* statistik menggunakan analisis korespondensi. *Output* pengolahan dari *software* statistik ditunjukkan pada Gambar 3.

Pada Gambar 3 menunjukkan grafik *symetric map* 2 (dua) dimensi dengan koordinat baris (row coordinates) adalah kode negara dan koordinat kolom (colomn coordinates) adalah kode halaman yang diakses. Sesuai dengan salah satu tujuan analisis korespondensi, terlihat pengelompokkan yang dapat diambil pada grafik. Pengambilan kelompok diambil dengan melihat jarak yang terdekat diantara koordinat tersebut secara subyektif.

Pertama, dapat dilihat pengelompokkan dengan batas merah, negara dengan kode 1, 2, dan 7 (Indonesia, Australia dan Jerman) berkelompok dengan halaman 11 dan 13 (Unduh dan Info Lelang). Sehingga dapat digambarkan

Tabel 10. Format Data Menurut Frekuensi

Klasifikasi	Halaman yang Diakses	Frekuensi
1	11	21307
1	14	1049
…	…	…
2	8	0
3	11	931
3	1	2486
3	10	2760



Gambar 3. *Output Software* Statistik untuk Analisis Koresponden

negara-negara tersebut mengakses halaman-halaman tersebut yang paling banyak. Bukan berarti negara lain tidak mengakses halaman tersebut ataupun bukan berarti negara tersebut tidak mengakses halaman lainnya. Jika dilihat pada pengelompokkan ini, korespondensi antara negara kode 1 (Indonesia) dengan halaman kode 11 (Unduh) sangat dekat.

Kedua, pengelompokkan terjadi pada negara dengan kode 8 dan 10 (Jepang dan Lainnya) dengan halaman berkode 2, 5, dan 4 (Tentang BPS, Pusat Layanan dan Istilah Statistik). Sehingga dapat digambarkan bahwa negara-negara ini berkorespondensi dengan halaman-halaman tersebut. Dapat juga dikatakan negara-negara ini tertarik untuk mengakses data pada BPS melalui bentuk selain web. Misalnya melalui perpustakaan maupun konsultasi statistik yang ada dalam Pusat Layanan.

Ketiga, pengelompokkan terjadi pada negara dengan kode 4, 5, 6 dan 9 (Inggris, Cina, Malaysia dan Singapura) dengan halaman berkode 1, 3, 9, dan 14 (Beranda, Rencana Strategis BPS, Publikasi BPS, dan Subyek Satistik) yang menggambarkan bahwa negara-negara ini mengakses dengan hampir merata terhadap halaman-halaman yang ada pada situs web BPS, tetapi paling berkorespondensi dengan halaman-hlaman tersebut. Jika dilihat pada pengelompokkan ini, negara dengan kode 4 (Inggris) sangat dekat korepondensinya halaman berkode 14

(Subyek Statistik) yang menggambarkan tingginya akses negara tersebut dalam membuka halaman Subyek Statistik yang berisi tabel-tabel statistik berdasarkan subyek.

Selanjutnya adalah pengelompokkan antara negara dengan kode 3 (USA) dan halaman dengan kode 10, 12 dan 15 (Berita Resmi BPS, Berita dan Website BPS Provinsi). Hal ini menggambarkan negara tersebut paling banyak mengakses halaman-halaman tersebut. Sedangkan pada halaman berkode 8 (Sekolah Tinggi Ilmu Statistik) berada pada posisi yang jauh dari kelompok manapun. Hal ini menggambarkan halaman ini yang paling jarang diakses oleh negara-negara tersebut.

KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan, dapat diambil beberapa kesimpulan bahwa pengguna situs web BPS yang diwakili oleh alamat IP dapat dikelompokkan dengan halaman yang diakses berdasarkan asal negara, sehingga didapat segmentasi pengguna data situs web BPS. Secara garis besar menjadi 3 (tiga) kelompok:

1. Berdasarkan Gambar 4 terjadi pengelompokkan pada negara Indonesia, Australia dan Jerman terhadap halaman "Unduh" dan "Info Lelang", bahkan korespondensi antara Indonesia dan halaman "Unduh" sangat dekat. Ini bisa diartikan bahwa yang mengunduh halaman web BPS paling banyak berasal dari Indonesia. Sedangkan untuk Australia dan Jerman juga banyak mengakses unduh dengan jarak yang hampir sama dengan Indonesia mengakses Info Lelang.
2. Kedua, pengelompokkan terjadi pada negara Jepang dan Lainnya dengan halaman "Tentang BPS", "Pusat Layanan" dan "Istilah Statistik". Sehingga dapat digambarkan bahwa negara-negara ini berkorespondensi dengan halaman-halaman tersebut. Dapat juga dikatakan negara-negara ini tertarik untuk mengakses data pada BPS melalui bentuk selain web. Misalnya melalui perpustakaan maupun konsultasi statistik yang ada dalam Pusat Layanan.
3. Kelompok ketiga terjadi pada negara Inggris, Cina, Malaysia dan Singapura dengan halaman "Beranda", "Rencana Strategis BPS", "Publikasi BPS", dan "Subyek Satistik". Negara Inggris sangat dekat korepondensinya halaman "Subyek Statistik" yang menggambarkan tingginya akses negara tersebut dalam membuka halaman Subyek Statistik yang berisi tabel-tabel statistik berdasarkan subyek.

Kesimpulan yang diwakili oleh ketiga kelompok ini, bukan berarti negara-negara tersebut tidak mengakses halaman-halaman lainnya. Secara korespondensi bisa dilihat kedekatan yang paling sering diakses. Yang menarik adalah halaman berkode 8 (Sekolah Tinggi Ilmu Statistik) yang berada jauh dari kelompok manapun, hal ini bisa dipelajari lebih lanjut.

Berdasarkan kesimpulan tersebut, maka penulis menyarankan beberapa hal, sebagai berikut:

1. Hasil pengelompokkan dapat digunakan sebagai bahan pertimbangan dalam mengembangkan situs web BPS, berhubungan dengan tampilan dan kemudahan akses dalam membuka halaman-halaman yang sering diakses.
2. Penyempurnaan untuk halaman web berbahasa asing, berhubungan dengan eratnya korespondensi negara luar dalam mengakses halaman-halaman yang ada pada situs web BPS.

DAFTAR PUSTAKA

Almind and Ingwersen. 1997. Informetric analyses on the World Wide Web: Methodological Approaches to Webometrics. E-Journal on-line Melalui http://www.cindoc.csic.es/cybermetrics/

Bjorneborn and Ingwersen. 2004. Toward a Basic Framework for Webometrics. E-Journal on-line Melalui http://www.interscience.wiley.com/cgi-bin/abstract/109594194/ABSTRACT

BPS. 2011. Laporan Reformasi Birokrasi Badan Pusat Statistik. Jakarta: BPS.

Cox, et al. 2001. *Multidimensional Scalling* (Second Ed.). New York: CRC Press LCC. E-book.

Greenacre, J. 1984. *Theory and Application of Correspondence Analysis*. London: Academic Press. E-book.

Izenman, A.J. 2008. *Modern Multivariate Statistical Techniques*. New York: Springer. E-book.

Khodra, M.L. 2003. Text Mining Kategori Teks Naive Bayes. E-Journal on-line Melalui http://kur2003.if.itb.ac.id/file/TextMiningKlasifikasiNB.pdf

Nicholson, S. 2006. The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services. E-Journal on-line Melalui http://arizona.openrepository.com/arizona/bitstream/10150/106175/1/nicholson2.pdf

Santoso, B. 2007. *Data Mining Teknis Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu.

Srivastava, et al. 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. E-Journal on-line Melalui http://nlp.uned.es/WebMining/Tema5.Uso/srivastava2000.pdf

Supriyadi, Y. 2011. *Aplikasi Teknik Bibliometric pada Analisis Data Paten*. Seminar Statistik Nasional 2011.

Tarapanoff, K, et al. 2001. *Intellegence Obtained by Applying Data Mining to a Database of French Theses on The Subject of Brazil* . Information Research, Vol. 7 No. 1, October 2001.

Thelwall, M. 2007. Bibliometrics to Webometrics. E-Journal on-line Melalui http:// www.scit.wlv.ac.uk/ ~cm1993/ papers/ JIS-0642-v4-Bibliometrics-to- Webometrics.pdf

Thelwall, M. 2009. Introduction to Webometrics: Quantitative Web Research for the Social Sciences. E-Journal on-line Melalui http://www.morganclaypool.com/doi/abs/10.2200/S001-76ED1V01Y200903ICR004

Web Mining. Melalui http://en.wikipedia.org/wiki/Web_mining

Xu, et al. 2011. *Web Mining and Social Networking*. New York: Springer. E-book.

DETEKSI INTRUSI JARINGAN DENGAN *K-MEANS CLUSTERING* PADA AKSES LOG DENGAN TEKNIK PENGOLAHAN *BIG DATA*

Farid Ridho[1], Arya Aji Kusuma[2]

Program Studi Komputasi Statistika
Politeknik Statistika STIS
e-mail:, [1]faridr@stis.ac.id, [2]aryaaku999@gmail.com

Abstrak

Keamanan jaringan, adalah salah satu aspek penting dalam terciptanya proses komunikasi data yang baik dan aman. Namun, masih adanya serangan yang efektif membuktikan bahwa sistem keamanan yang berlaku belum cukup efektif untuk mencegah dan mendeteksi serangan. Salah satu metode yang dapat digunakan untuk mendeteksi serangan ini adalah dengan dengan *Intrusion Detection System* (IDS). Besarnya data (*volume*), cepatnya perubahan data (*velocity*), serta variasi data (*variety*) merupakan ciri-ciri dari *Big data*. Akses log, secara teori termasuk dalam kategori ini sehingga dapat dilakukan pemrosesan menggunakan teknologi bigdata dengan *Hadoop*. Hal ini mendorong penulis untuk dapat menerapkan metode pengolahan baru yang dapat mengatasi perkembangan data tersebut, yaitu *Big data*. Penelitian ini dilakukan dengan menganalisis akses log dengan *K-Means Clustering* menggunakan metode pengolahan bigdata. Penelitian menghasilkan satu model yang dapat digunakan untuk mendeteksi sebuah serangan dengan probabilitas deteksi sebesar 99.68%. Serta dari hasil perbandingan kedua metode pengolahan bigdata menggunakan pyspark dan metode tradisional menggunakan *python* standar, metode bigdata memiliki perbedaan yang signifikan dalam waktu yang dibutuhkan dalam eksekusi program.

Kata kunci*:* IDS, *big data*, akses log, *k-means*, *clustering*

Abstract

Good network security planning ensures the safety and comfort of user data. However, the existence of effective attacks proves that the current security system is not effective to prevent and detect attacks. One of methods that can be used to detect this attack is by using Intrusion Detection System (IDS). The amount of data (volume), speed of which data change (velocity), and variations in data (variety) are characteristics of big data. Log access, theoretically is also a form of big data so a new approach in statistical data processing is needed to overcome big data. This research was conducted by analyzing log access with K-Means Clustering using the big data processing technique. The study produced a model that can be used to detect an attack with a detection probability of 99.68%. As well as a comparison between big data using Pyspark and traditional processing technique using standard python, which big data technique has a significant difference in time needed to execute the program.

Keywords: *IDS, big data, log access, k-means, clustering*

PENDAHULUAN

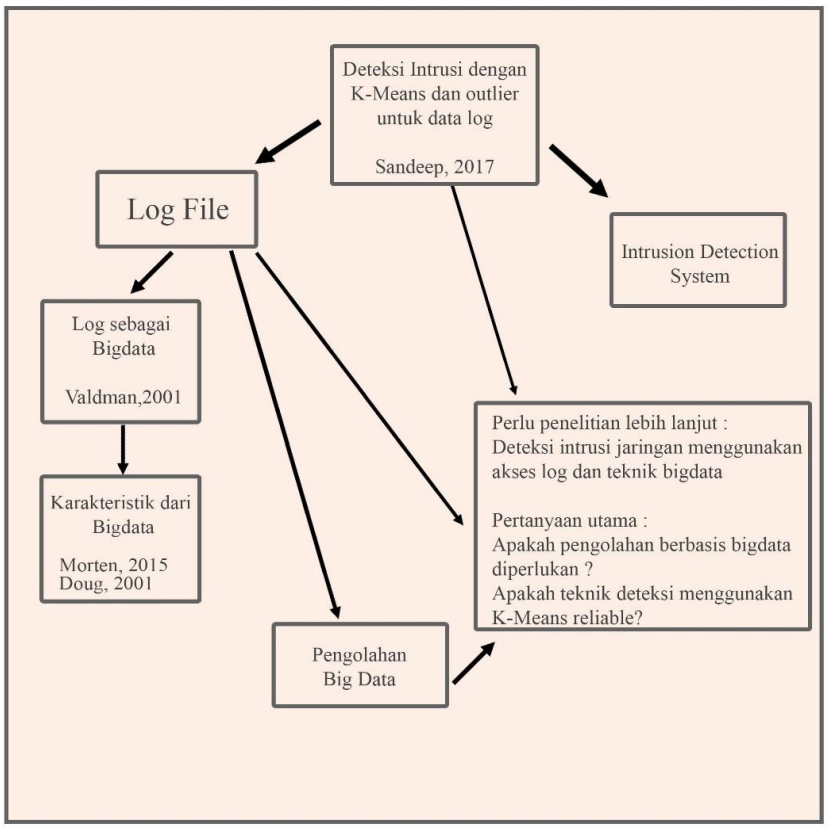
Komunikasi antar komputer merupakan hal sehari – hari yang sering ditemui saat ini. Komunikasi ini dapat terjadi karena adanya sebuah hubungan dalam jaringan komputer. Peranan jaringan komputer tidak hanya terletak pada komunikasi antar komputer, akan tetapi juga terletak pada pertukaran data yang terjadi, baik untuk masyarakat, industri maupun pemerintah. Pada implementasinya, perlu diperhatikan tiga aspek utama, yaitu *performance*, *reliability*, dan *security* agar kegiatan komunikasi dan pertukaran data dapat berjalan secara cepat, aman dan nyaman baik bagi seluruh kalangan.

Keamanan merupakan aspek utama dalam adanya jaringan komputer yang baik. Dalam penerapan keamanan jaringan, dapat dibagi menjadi empat tahapan yaitu prediksi, cegah, deteksi, dan tanggapan. Empat tahapan ini memberikan proteksi untuk menjaga data baik untuk server, maupun pengguna agar terhindar dari ancaman luar yang dapat merusak maupun merugikan pihak terkait. Namun, masih adanya kasus serangan pada jaringan membuktikan bahwa proteksi yang disediakan oleh sistem keamanan jaringan belum sempurna (Govscirt, 2018).

Sistem Deteksi Intrusi Jaringan merupakan salah satu pendekatan dalam deteksi serangan, metode ini mengklasifikasikan aktivitas jaringan yang sedang terjadi ke dalam model yang telah dibangun sistem ke dalam *library* sehingga dapat dikategorikan sebagai aktifitas normal atau serangan (Chakraborty, 2017). Pada praktiknya, sistem intrusi jaringan dibagi menjadi dua yaitu berbasis anomali dan berbasis aturan. Pada Sistem Deteksi Intrusi jaringan berbasis aturan, aktivitas atau lalu lintas akan dicek dengan membandingkan data tersebut dengan aturan yang telah dicatat sebelumnya menggunakan pola yang sering ditemukan pada aktivitas serangan. Sistem Akan tetapi, metode ini memiliki kekurangan yaitu metode ini tidak dapat mendeteksi serangan tipe baru yang belum pernah tercatat sebelumnya, yaitu *false-negative*. Ditambah metode ini sering mengklasifikasikan aktifitas atau lalu lintas baru sebagai serangan, dapat disebut sebagai *false-positive* yaitu aktivitas normal yang tercatat sebagai serangan.

Sedangkan Sistem deteksi intrusi jaringan berbasis anomali mengklasifikasikan aktivitas atau trafik ke dalam klasifikasi data normal maupun data anomali. Hal ini dapat dihasilkan dengan cara statistik. Dengan melihat data secara statistik, dapat dihasilkan model yang tepat dapat mendeteksi serangan dan dapat selalu terbaharui. Aktivitas dan lalu lintas dalam komunikasi antar server dan pengguna tercatat dalam data log (Iversen, 2015). Akses log, merupakan catatan data transaksi pengguna dan server yang meliputi URL, HTML, gambar, file, browser yang digunakan dalam kejadian transaksi tersebut. Secara umum, akses log dapat dianalisis secara statistik untuk informasi perihal monitoring jaringan, seperti banyaknya pengunjung, banyaknya jumlah akses data, maupun melihat popularitas halaman setiap harinya (Valdman, 2001). Karena data tercatat secara keseluruhan transaksi, maka dapat dilakukan analisis yang lebih lanjut. Akan tetapi, perlu diketahui bahwa dengan penggunaan akses log maka dapat menimbulkan permasalahan pada saat prosesing data karena file akses log memiliki ukuran yang besar, memiliki perubahan data yang sangat cepat, dan memuat berbagai macam informasi (Grace, 2011).

Besarnya ukuran data, cepatnya perubahan data, serta data yang bervariasi merupakan ciri – ciri dari *Big Data*. Sehingga, secara teori akses log merupakan *Big Data* maka dapat dilakukan metode pengolahan dengan *Big Data* untuk mengatasi permasalahan tersebut. Pengolahan dengan teknik *Big Data* akan melibatkan *Hadoop*, yaitu aplikasi berbasis open-source yang dapat mengatur dan memproses data secara terdistribusi. Dengan arsitektur *Hadoop*, juga akan digunakan *Spark* yaitu aplikasi pengolahan data secara terdistribusi (Scott, 2015).



Gambar 1. Peta Literatur

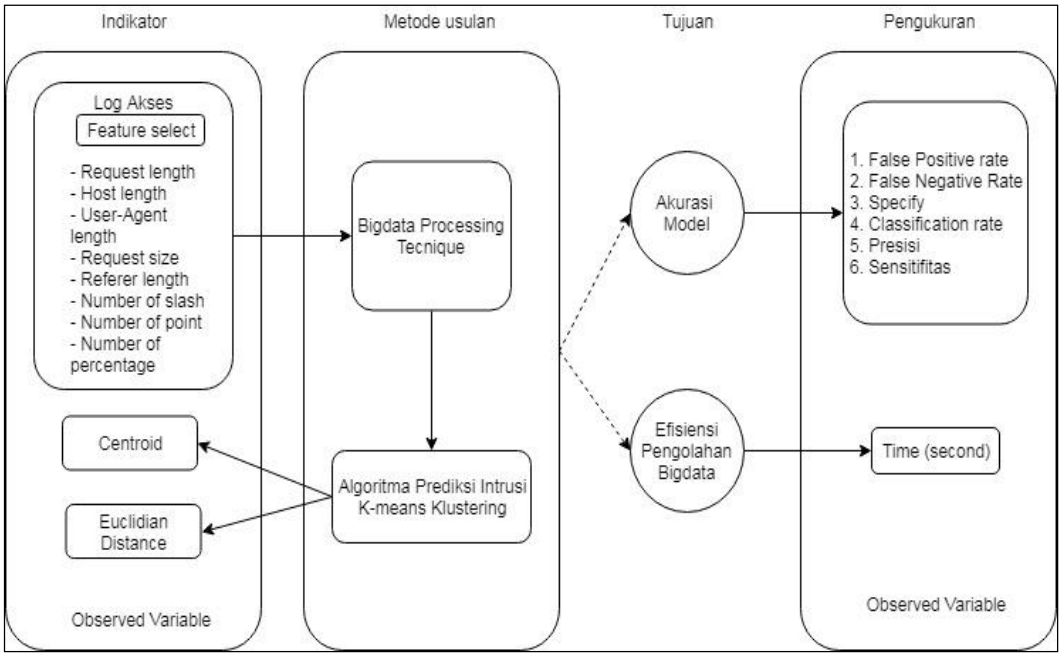
Algoritma analisis yang digunakan adalah *K-Means Clustering*. Dimana data akan di klaster kan menjadi 2 dan lalu akan diklasifikasikan sebagai cluster normal dan cluster anomali. Dari cluster bentukan itu akan dibangun sebagai model yang nantinya dapat digunakan untuk menganalisis aktivitas atau trafik baru yang masuk.

METODOLOGI

Logfile adalah catatan atas seluruh kejadian atau transaksi yang terjadi pada suatu sistem. Seiring perkembangan jaman, transaksi antar pengguna terjadi dalam hitungan menit bahkan detik. Sehingga untuk memenuhi kebutuhan atas pengolahan dan prosesing data, diperlukan teknik yang tepat yaitu teknik *Big Data*. Dalam penelitian ini akan digunakan akses log dari website www.stis.ac.id dalam periode waktu 27 Februari 2017 hingga 31 Oktober 2017. Akses log akan diproses menggunakan *Big Data* dengan *K-Means Clustering* sebagai metode analisisnya (Chandel, 2017).

1. Peta literatur bahasan

Peta literatur diatas menjelaskan adanya kesinambungan atas beberapa topik yang telah dijelaskan sebelumnya. Untuk mengatasi permasalahan deteksi intrusi jaringan dapat menggunakan analisis akses log, dimana akses log adalah catatan seluruh transaksi. Karena besarnya ukuran, cepatnya perubahan serta data yang memiliki variasi, akses log dapat dikategorikan sebagai bigdata. Maka dalam proses pengolahannya dapat digunakan teknik pengolahan *Big Data*. Pengolahan *Big Data* dilakukan dengan tujuan memeriksa keseluruhan data untuk menemukan pola didalamnya (Mukherjee dkk, 2016). Untuk melakukan proses komputasi secara cepat untuk data besar tersebut, tidak dapat menggunakan sistem konvensional seperti RDBMS maupun sistem penyimpanan lainnya, karena dapat memberikan beban yang terlalu besar pada komputer pengolah. Untuk menanggulangi hal ini dapat digunakan *Hadoop* sebagai arsitektur pengolahan *Big Data* (Parthiban, 2016).



Gambar 2. Kerangka Pikir Penelitian

2. Kerangka Pikir Penelitian

Dari penjelasan diatas, setiap tahapan dan langkah penelitian dapat digambarkan pada kerangka pikir (Gambar 2).

Dari kerangka pikir di atas, dapat dilihat bahwa dari log akses dilakukan clustering dengan fitur yang terpilih untuk mendapatkan pusat cluster untuk dibentuk model yang nantinya akan di evaluasi untuk melihat ukuran dari model tersebut (Fink dkk, 2012). Serta dalam proses pengolahan menggunakan bigdata, akan diuji perbedaannya dengan metode tradisional.

3. Metode Analisis

Dalam penelitian ini dilakukan serangkaian tahapan guna mencapai hasil analisis yang representatif. Analisis yang digunakan adalah dengan metode K-Means *Clustering*. K-Means *Clustering* dilakukan pada variabel akses log yang dipilih. Berikut tahapan yang dilewati dalam pengerjaan penelitian ini :

Data *preprocessing*

Data akses log berekstensi log akan di parsing terlebih dahulu menggunakan bahasa pemrograman *python*, hal ini dilakukan agar memudahkan sistem untuk mengolah data tersebut. Parsing data dilakukan dengan bantuan regex, yaitu regular expression memanfaatkan pola yang terdapat pada data akses log sendiri. Proses data *preprocessing* yang dilakukan adalah sebagai berikut:

1. Data cleaning*:* Pembersihan data dengan mengisi missing value, dan mengatasi inkonsistensi data yang ada pada log.
2. Data integrasi*:* Menghilangkan konflik dan menggabungkan data dengan tipe berbeda
3. Data selection*:* Memilih data sesuai dengan yang dibutuhkan, dalam istilah lain sering disebut dengan istilah feature selection
4. Data transformation*:* Data di normalisasi, agregasi, dan generalisasi

Ekstraksi dan Pemilihan Fitur

Analisis K-Means dilakukan pada beberapa variabel pilihan. Pemilihan variabel tersebut dilakukan dengan tahapan Ekstraksi dan Pemilihan Fitur. Ektraksi fitur untuk melakukan transformasi data menjadi fitur yang sesuai dengan model. Dalam penelitian terkait tentang ekstraksi fitur pada data log server web, didapatkan 30 fitur yang dapat diambil dari sebuah file log untuk deteksi serangan (Nguyen, dkk, 2011).

Feature Name	Feature Name
Length of the request ⋆ ◇	Length of the path ⋆
Length of the arguments ⋆ ◇	Length of the header “Accept” †
Length of the header “Accept-Encoding” †	Length of the header “Accept-Charset” †
Length of the header “Accept-Language” †	Length of the header “Cookie” †
Length of the header “Content-Length” †	Length of the header “Content-Type”
Length of the Host †	Length of the header “Referer” †
Length of the header “User-Agent” †	Method identifier
Number of arguments ⋆	Number of letters in the arguments ⋆
Number of digits in the arguments ⋆	Number of ’special’ char in the arguments ⋆ † • ◇
Number of other char in the arguments • ◇	Number of letters char in the path ⋆
Number of digits in the path ⋆ †	Number of ’special’ char in the path ⋆
Number of other char in path †	Number of cookies †
Minimum byte value in the request ◇	Maximum byte value in the request ⋆ †
Number of distinct bytes	Entropy ◇
Number of keywords in the path	Number of keywords in the arguments

Gambar 3. Fitur yang Dianggap Relevan pada Deteksi Serangan

Tabel 1. Fitur Spesial

No	Serangan	Contoh	Karakter Spesial
(1)	(2)	(3)	(4)
1	*Directory Transversal*	/admin/./index.html	*Slash* (/), *dot* (.)
2	*Hex-Encode HTTP Evasion*	/%69%6E%64%65	Persentase (%)
3	*Regex Attack*	^(([a-z])+.)+$	Semua simbol

Keterangan simbol untuk Gambar 3 adalah sebagai berikut: ⋆ fitur yang dipilih oleh CFS dari dataset CSIC-2010, † fitur yang dipilih oleh mRMR dari dataset CSIC 2010; • fitur yang dipilih oleh CFS dari dataset ECML/PKDD 2007 ; ◊ fitur yang dipilih oleh mRMR dari dataset ECML/PKDD 2007. Untuk fitur dalam akses log serta melihat faktor yang sering muncul pada serangan pada penelitian terkait, dipilih 6 variabel, yaitu: (a)Panjang karakter pada request, (b)Panjang karakter pada host, (c)Panjang karakter pada User-agent, (d)Besaran ukuran byte dalam transaksi, (e) panjang karakter pada Referer, serta (f) Banyaknya karakter spesial pada request. Beberapa karakter spesial merupakan ciri – ciri adanya serangan tertentu. Dengan detail pada karakter spesial mengikuti banyaknya okurensi yang mewakili serangan tersebut seperti yang ditampilkan pada Tabel 1.

Dari beberapa karakter spesial, dipilih karakter garis miring (/), titik(.), serta symbol persen (%). Karakter spesial tersebut dipilih karena dianggap dapat mendeteksi anomali yang ada pada transaksi pengguna dan *server*. Karakter ini di ekstraksi dari variabel request yang merupakan halaman yang dituju oleh pengguna terkait.

Setelah melakukan ekstraksi fitur kemudian akan dilakukan proses pemilihan fitur yang dilakukan untuk memilih fitur yang berpengaruh terhadap data dengan tujuan memberikan hasil analisis yang akurat sesuai masalah yang ada. Sehingga, setelah proses pemilihan fitur ini didapatkan 8 fitur utama yang akan digunakan untuk keperluan pengolahan dan analisis tujuan penelitian (Tabel 2).

Dari Tabel 2 terlihat beberapa fitur yang digunakan untuk keperluan analisis. Terdapat 8 variabel yang digunakan untuk keperluan analisis, yaitu besaran request pengguna, panjang karakter dari host, panjang karakter dari request line, panjang karakter dari User-Agent, panjang karakter dari Referer, banyaknya okurensi garis miring, dot, serta persentase. Variabel tersebut diasumsikan telah dapat mewakili beberapa karakteristik dari aktivitas jaringan, baik aktivitas normal maupun anomali.

Tabel 2. Fitur Terpilih sebagai Variabel Penelitian

No	Feature Name
(1)	(2)
1	Panjang dari request line
2	Panjang dari IP Host
3	Panjang header dari User-Agent
4	Besarnya ukuran byte setiap request
5	Panjang header dari Referer
6	Jumlah okurensi garis miring
7	Jumlah okurensi titik
8	Jumlah okurensi persentase

Tabel 3. *Confusion Matrix*

		Predicted value	
True value		P	N
	P	TP	FN
	N	FP	TN

Pengolahan dan Analisis Data

Setelah dilakukannya seleksi fitur untuk K-Means, maka hasil olahan akan siap diolah menggunakan *Hadoop* dan *Spark*. Hal ini dilakukan dengan menjalankan dua aplikasi yang bersangkutan, yaitu *Hadoop* Distributed File System (HDFS) serta *Spark*. HDFS adalah media penyimpanan secara terdistribusi antar komputer sehingga memungkinkan *Hadoop* dan *Spark* melakukan tugas pengolahan data secara terdistribusi. Pada percobaan ini digunakan 2 komputer, yaitu satu sebagai master node dan satunya sebagai slave node. Master node merupakan komputer pengatur jalannya pengolahan terdistribusi, sedangkan slave merupakan komputer pekerja. Dalam pengolahan datanya, digunakan *PySpark*, yaitu *Python* in *Spark*. Merupakan program yang memungkinkan untuk menggunakan bahasa pemrograman *python* dalam *Spark* melalui API yang telah disediakan oleh *Spark*.

Dari serangkaian proses diatas akan menghasilkan hasil akhir berupa sebuah model K-Means dan Bisecting K-Means sebagai pembandingnya pada *PySpark*. Sesuai dengan tujuan penelitian ini, evaluasi model ini dilakukan menggunakan metrik yang sama seperti penelitian – penelitian sebelumnya, yaitu dengan menghitung false-positive rate, false-negative rate, sensitifitas atas model, spesifik dari model, classification rate serta presisi atas hasil dari model. Keenam metrik ini dapat didapatkan dari perhitungan 4 aspek, yaitu *True positive*, *true negative*, *false positive* dan *false negative* seperti pada *confusion matrix* (Tabel 3).

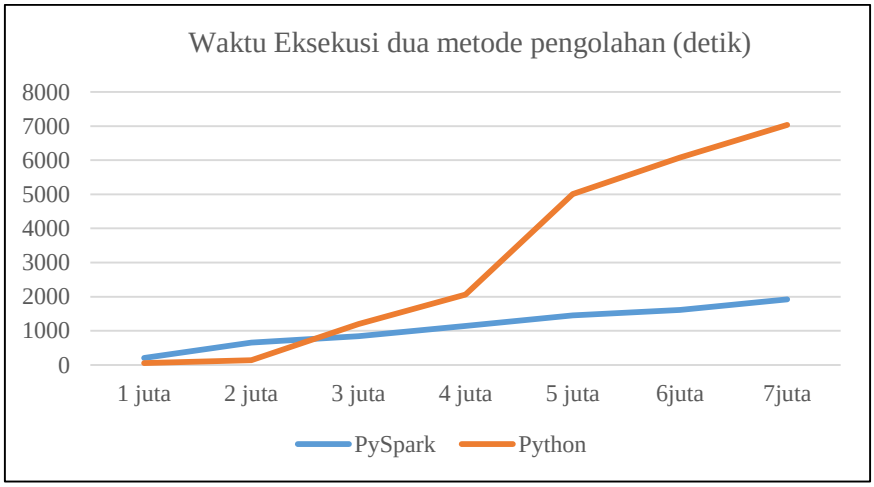
Beberapa nilai di atas dapat menggambarkan kondisi yang sesuai untuk bahasan keamanan jaringan komputer, baik dalam hal ketepatan dan akurasinya.

Spesifikasi yang digunakan

Dalam tahapan pengolahan diatas, digunakan dua perangkat komputer yang terhubung dalam virtualBox. Dimana satu komputer berperan sebagai master node, dan yang lainnya adalah slave node. Master node bertugas untuk mengendalikan dan membagi tugas antar komputer yang terhubung. Sedangkan slave node bertugas hanya sebagai penerima perintah dari master *node*. Kedua komputer tersebut memiliki spesifikasi yang sama, yaitu dengan 2GB RAM dan dengan sistem operasi Ubuntu 16.04. Perbedaannya terletak pada alokasi memori yang diberikan ke ekosistem *Hadoop*, dimana master node memberikan 512MB baik ke executor serta driver memory, sedangkan pada slave node hanya 512MB pada executor memory saja.

Tabel 4. Tabel Waktu Eksekusi Dua Metode Berbeda dalam Detik

Size	1 juta	2 juta	3 juta	4 juta	5 juta	6 juta	7 Juta
Pyspark	201.1	655.7307	841.1226	1140.61	1449.246	1607.348	1920.184
Python	51.78	140.7595	1197.417	2062.988	5011.45	6082.78	7032.3



Gambar 4. Grafik Perbandingan Waktu Eksekusi Dua Metode Pengolahan

Sedangkan piranti lunak yang digunakan dalam pengolahan dan uji coba bahan penelitian adalah sebagai berikut :

1. Bahasa pemrograman*: Python* 2.72
2. *Hadoop* 3.1.0 – *Hadoop Distributed File System*
3. *Spark* 2.3.0, dengan menggunakan *PySpark* (*Spark Python* API)

HASIL DAN PEMBAHASAN

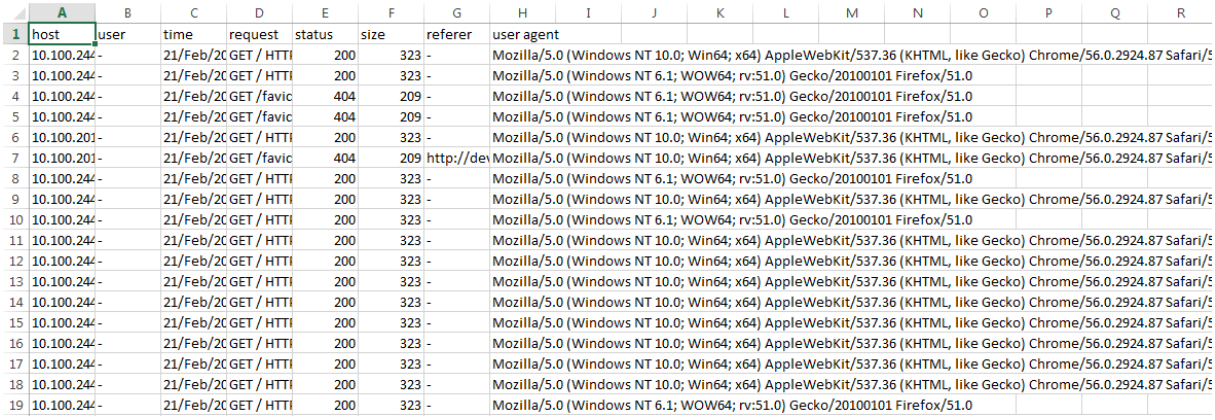
1. Perbandingan antar 2 Metode

Pada bagian sebelumnya telah dijelaskan bahwa dengan penggunaan teknik pengolahan *Big Data* dapat mempersingkat waktu eksekusi yang dibutuhkan untuk tiap data. Pernyataan ini diuji terlebih dahulu menggunakan beberapa data dummy dengan perbedaan pada banyaknya *record* yang ada pada data tersebut. Data yang digunakan mencakup data dengan 1 juta *record*, 2 juta hingga 7 juta *record* dengan masing – masing memiliki 10 variabel. Data tersebut diperlakukan sama seperti data olahan yang nantinya dipakai, yaitu dengan mengolahnya pada *PySpark* dengan menggunakan algoritma K-Means. Hal ini ditujukan untuk mencatat waktu yang dibutuhkan metode tradisional dan metode *Big Data* dalam memproses sebuah data dengan ukuran yang berbeda.

Pengukuran waktu eksekusi dalam percobaan menggunakan bantuan package time yang dimiliki oleh *python*. Time() memberikan nilai waktu terkini dalam detik yang dihitung dari awal masa. Perintah ini dimasukkan saat script dijalankan dan saat script selesai berjalan. Dalam pengukurannya, waktu eksekusi didapat dari selisih antara waktu selesai script berjalan dengan waktu script mulai dijalankan.

Dari Tabel 4 tersebut dapat dilihat bahwa setiap ukuran data akan menghasilkan ukuran waktu eksekusi yang berbeda. Pada metode tradisional, dapat dilihat bahwa pada ukuran data yang kecil proses eksekusi yang dibutuhkan lebih sedikit. Namun untuk kelipatan berikutnya, membutuhkan waktu yang secara drastis naik. Berbeda pada metode bigdata, di mana waktu eksekusi yang dibutuhkan naik secara stabil.

Dari hasil tersebut juga telah dilakukan uji T untuk membuktikan kedua metode adalah signifikan berbeda. Dengan tingkat signifikansi 5% dan df yaitu 6, didapatkan nilai t tabel sebesar 1,943. Dan berdasarkan perhitungan, didapat nilai t-



	A	B	C	D	E	F	G	H
1	host	user	time	request	status	size	referer	user agent
2	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
3	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:51.0) Gecko/20100101 Firefox/51.0
4	10.100.24	-	21/Feb/20	GET /favic	404	209	-	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:51.0) Gecko/20100101 Firefox/51.0
5	10.100.24	-	21/Feb/20	GET /favic	404	209	-	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:51.0) Gecko/20100101 Firefox/51.0
6	10.100.201	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
7	10.100.201	-	21/Feb/20	GET /favic	404	209	http://dev	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
8	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:51.0) Gecko/20100101 Firefox/51.0
9	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
10	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:51.0) Gecko/20100101 Firefox/51.0
11	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
12	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
13	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
14	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
15	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
16	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
17	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
18	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/5
19	10.100.24	-	21/Feb/20	GET / HTTI	200	323	-	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:51.0) Gecko/20100101 Firefox/51.0

Gambar 5. Hasil Parsing *Raw Data* Akses Log Website Politeknik Statistika STIS

18819	48	14	72	42	3	2	0
225	52	13	72	88	5	2	0
16762	46	15	165	22	3	2	0
18819	48	15	65	46	3	2	0
225	52	15	65	88	5	2	0
20396	46	14	137	48	3	2	0
24212	33	14	65	46	3	2	0
17492	80	11	72	1	11	1	0
24032	40	11	110	18	3	2	0
0	15	10	160	1	2	1	0
17298	14	12	160	1	2	1	0
225	25	15	100	28	2	2	0
20396	46	15	165	22	3	2	0
225	52	14	65	88	5	2	0
0	15	12	160	1	2	1	0
21067	78	14	65	41	3	2	0
17298	14	12	108	1	2	1	0
225	52	12	108	1	5	2	0
24276	40	14	90	80	3	2	0
225	52	14	90	48	5	2	0
18819	48	14	1	48	3	2	0

Gambar 6. Hasil Ekstraksi Fitur

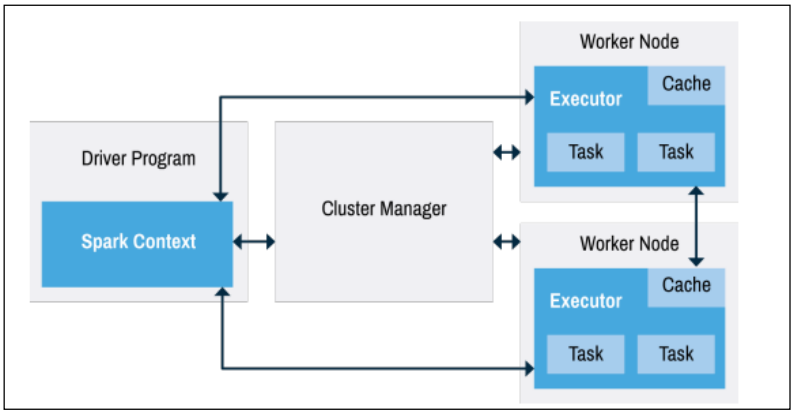
hitung sebesar 2.2165, dari kedua nilai ini dapat dilihat perbedaan kedua metode. Karena t hitung > t tabel, maka H0 ditolak, sehingga dapat diambil kesimpulan bahwa terdapat perbedaan signifikan antara waktu yang dibutuhkan untuk eksekusi script atau program pada metode tradisional menggunakan *python* dengan metode pengolahan bigdata.

2. *Clustering* pada *Big Data*

Preprocessing

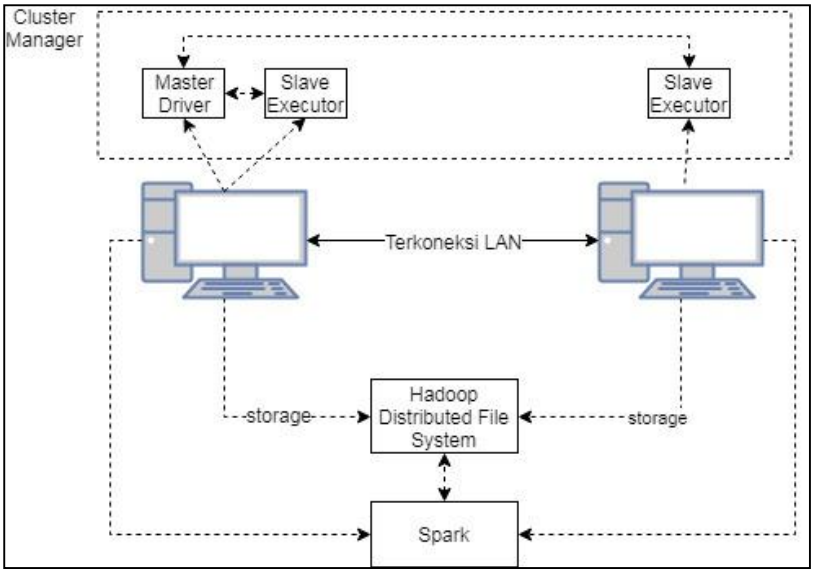
Sesuai yang telah dijelaskan pada sub-bab sebelumnya, tahapan ini terbagi menjadi 4 bagian yaitu data pre-prosesing, seleksi fitur dan ekstraksi fitur, pengolahan data, serta analisis hasil olahan. Pre-prosesing data dilakukan dengan parsing data file berekstensi .log menjadi bentuk yang lebih dapat dibaca yaitu .csv. Hal ini dilakukan dengan menggunakan regex yang telah disediakan pada salah satu script sebelumnya. Script ini akan menghasilkan file .csv baru yang diambil dari parsing data log.

Data hasil pre-prosesing tidak mengalami penambahan atau pengurangan record, di mana hanya terdapat catatan transaksi pada rentang waktu Februari 2017 hingga Oktober 2017. Dari data tersebut dilakukan seleksi dan ekstraksi fitur. Dari fitur yang telah dipilih, dilakukan ekstraksi dengan menghitung panjang dari variabel tertentu, serta menghitung banyak okurensi dari karakter tertentu. Penghapusan beberapa variabel yang tidak digunakan juga dilakukan.



Gambar 7. Alur Kerja Logical *Spark*

Sumber: Getting Started with Apache Spark, hal. 38



Gambar 8. Alur Kerja Fisik *Spark*

Pengolahan data

Setelah data siap olah, data tersebut dimasukkan kedalam ekosistem pengolahan *Hadoop* dengan melalui HDFS. Hal ini memungkinkan data tersebut dapat diolah secara terdistribusi, agar memenuhi tujuan dari penelitian ini. Pengolahan data dilakukan dengan *Spark*, yang dapat dijelaskan dalam Gambar 7 dan Gambar 8.

Pengolahan pada *Spark*, tidak melihat seluruhnya pada jumlah komputer yang digunakan. Namun melihat pada banyaknya node yang digunakan. Komposisi node pada *Spark* meliputi 1 master node dan beberapa slave node. Pada penelitian ini, menggunakan 2 slave node. Pengalokasian tugas pengolahan dilakukan pada master node dengan driver. Alokasi tugas dari driver lalu akan diberikan ke cluster manager, pada penelitian ini *Spark* menjadi cluster manager. Cluster manager lalu akan membagikan tugas dari driver ke slave node untuk dilakukan pengolahan secara terdistribusi. Pembagian tugas ini melewati HDFS, karena data yang akan diolah harus masuk kedalam HDFS untuk dapat dilakukan komputasi terdistribusi. Cluster manager akan memonitoring tugas yang telah diberikan ke slave node, dengan memonitoring secara terus menerus konsistensi dari pengolahan dapat terjaga. Setelah data selesai diolah, maka akan diberikan kembali ke cluster manager untuk disatukan kembali dan hasil tersebut akan diberikan ke driver untuk dapat ditampilkan pada layar atau file *output*.

Hasil pengolahan data

1. K-Means *Clustering*

Hasil dari pengolahan data di atas berupa pusat cluster dan model yang dapat

```
In [2]: from pyspark.mllib.clustering import KMeans, KMeansModel
        from numpy import array
        from math import sqrt

        data = sc.textFile("///user/hadoop/revisi_4.csv")
        parsedData = data.map(lambda line: array([float(x) for x in line.split(',')]))

        clusters = KMeans.train(parsedData, 2)

        for x in clusters.clusterCenters: print(x)

[1.14941476e+04 5.80004820e+01 1.32577988e+01 1.01088818e+02
 2.53420899e+01 4.68543727e+00 1.99703187e+00 1.47047355e-01]
[1.72912425e+06 5.76777824e+01 1.33801584e+01 1.09877345e+02
 3.97445811e+01 5.12390579e+00 2.06638183e+00 1.46540225e+00]
```

Gambar 9. Hasil Running Algoritma *K-Means* dengan *Pyspark*

```
In [1]: from pyspark.mllib.clustering import BisectingKMeans, BisectingKMeansModel
        from numpy import array
        from math import sqrt

        data = sc.textFile("///user/hadoop/revisi_4.csv")
        parsedData = data.map(lambda line: array([float(x) for x in line.split(',')]))

        model = BisectingKMeans.train(parsedData, 2, maxIterations=5)

        for x in model.clusterCenters: print(x)

[7.51238183e+02 6.79834723e+01 1.30996041e+01 8.99985370e+01
 2.14752879e+01 5.60538117e+00 1.88649086e+00 1.38719273e-01]
[3.25604176e+04 4.46917369e+01 1.34691553e+01 1.15906847e+02
 3.05531600e+01 3.46086011e+00 2.14465707e+00 1.63326513e-01]
```

Gambar 10. Hasil Running Algoritma *Bisecting K-Means* dengan *Pyspark*

dipanggil di dalam *PySpark*. Dari 5.700.375 record akses log dengan 8 variabel menghasilkan dua pusat cluster yaitu: [11.494,147 ; 58,0004 ; 13,257 ; 101,08 ; 25,342 ; 4,685 ; 1,997 ; 0,147] serta [1.729.124,25 ; 57,677 ; 13,38 ; 109,8 ; 39,744 ; 5,123 ; 2,066 ; 1,465] (Gambar 9).

2. Bisecting K-Means *Clustering*

Dari hasil running script terlihat bahwa dari 5.700.375 record akses log dengan 8 variabel menghasilkan dua pusat cluster yaitu: [751,238 ; 67,983 ; 13,099 ; 89,99 ; 21,475 ; 5,605 ; 1,8864 ; 0,1387] serta [3,256 ; 4,469 ; 13,469 ; 115,906 ; 30,55 ; 3,46 ; 2,144 ; 0,1633] (Gambar 10).

Evaluasi hasil Cluster terhadap IDS

Dari hasil clustering didapatkan sebuah model yang dapat digunakan untuk mendeteksi suatu data tergolong ke cluster normal atau anomali. Untuk menguji coba keefektifan model tersebut, maka dihitung serangkaian metrik yaitu *false positive rate* (FPR), *false negative rate* (FNR), sensitifitas, presisi, classification rate, dan spesifik (Tabel 5, Tabel 6 dan Tabel 7).

1. Interpretasi FPR

FPR adalah rasio IDS mengklasifikasikan aktivitas normal sebagai aktivitas serangan. Pada algoritma *K-Means Clustering*, didapatkan nilai FPR sebesar 0,0191% sedangkan pada Bisecting *K-Means Clustering* mendapat 53,58%. Semakin besarnya nilai FPR, maka model tersebut dapat dikatakan lebih sensitif dalam mendeteksi aktivitas jaringan. Namun, semakin tinggi nilai FPR juga menggambarkan bahwa akan terlalu banyak aktivitas normal yang terdeteksi sebagai anomali. Maka, dari kedua nilai berikut dapat menggambarkan bahwa algoritma K-Means lebih baik dalam aktivitas.

2. Interpretasi FNR

FNR adalah rasio yang mengklasifikasikan aktifitas serangan sebagai aktifitas normal. Semakin tinggi nilai FNR menunjukkan bahwa model akan lebih rentan terhadap serangan. Pada kedua algoritma, mendapatkan nilai FNR yang sama yaitu sebesar 79,51%. Namun, mengingat nilai FPR *K-Means Clustering* lebih rendah daripada Bisecting *K-Means Clustering* maka algoritma *K-Means*

Tabel 5. *Confusion Matrix* dari *K-Means*

K-Means	***Predicted value***		
True value		P	N
	P	76	295
	N	19	99.609

Tabel 6. *Confusion Matrix* dari *Bisect K-Means*

Bisect	***Predicted value***		
True value		P	N
	P	76	295
	N	53.390	46.238

Tabel 7. Ukuran Evaluasi Kedua Metode

Algoritma	**FPR**	**FNR**	**Sensitifitas**	**Spesifikasi**	**R. Klasifikasi**	**Presisi**
K-Means	0,019%	79,515%	0,205	0,999	99,68%	80%
Bisecting	53,58%	79,515%	0,205	0,464	46,31%	0,14%

Clustering tetap lebih unggul daripada Bisecting K-means.

3. Interpretasi dari Sensitifitas

Yaitu rasio model mengkategorikan aktivitas serangan sebagai serangan atau juga dapat disebut sebagai *True Positive Rate*. Semakin tinggi nilai sensitifitas, maka semakin tinggi FPR yang ada pada model tersebut. Kedua nilai ini memiliki keterkaitan satu sama lain. Terlihat bahwa dari kedua algoritma, bahwa keduanya memiliki nilai sensitifitas yang sama. Namun, K-Means memiliki FPR yang lebih rendah dari pada Bisect K-Means hal ini menggambarkan bahwa algoritma Bisect K-means lebih sensitif dan akan memberikan lebih banyak notifikasi kepada administrator.

4. Interpretasi dari Spesifikasi

Seperti halnya sensitifitas, spesifikasi menggambarkan nilai *True Negative Rate* (TNR) yaitu mengkategorikan aktivitas normal sebagai normal. Dapat dilihat pada table diatas bahwa nilai spesifikasi dari K-Means memiliki perbedaan yang sangat jauh dibanding Bisect *K-Means Clustering*. Hal ini menggambarkan bahwa algoritma *K-Means Clustering* lebih sanggup untuk mengidentifikasi aktifitas normal sebagai normal.

5. Interpretasi dari Rasio Klasifikasi

Rasio klasifikasi menggambarkan seberapa besar akurasi yang dihasilkan dari permodelan data. Dalam kasus deteksi intrusi jaringan, maka nilai rasio klasifikasi yang tinggi akan meningkatkan kualitas dari model tersebut, didorong dengan kecilnya nilai FPR dan FNR dari model tersebut. Sehingga dapat disimpulkan bahwa algoritma *K-Means Clustering* lebih baik dalam memberi gambaran besar atas aktivitas serangan ataupun normal.

6. Interpretasi dari Presisi

Presisi yang dimaksud adalah kemampuan model untuk mendeteksi sebuah aktivitas serangan. Algoritma *K-Means Clustering* memiliki nilai yang lebih besar daripada algoritma Bisect *K-Means Clustering*, sehingga dapat diambil kesimpulan bahwa *K-Means Clustering* lebih mampu atas mendeteksi aktivitas serangan.

Dari hasil evaluasi diatas, dapat terlihat bahwa algoritma *K-Means Clustering* memiliki keunggulan sebagai model pendeteksian intrusi jaringan.

KESIMPULAN DAN SARAN

Dari hasil pengolahan, analisis, hingga evaluasi data di atas maka dapat diambil kesimpulan berdasarkan tujuan dari penelitian ini, yaitu :

1. Implementasi pengolahan bigdata pada Deteksi Intrusi Jaringan dengan memanfaatkan akses log telah dilakukan. Hal ini dapat dilakukan karena akses log memiliki karakteristik dari bigdata, sehingga cocok untuk dilakukannya teknik pengolahan bigdata pada data tersebut, untuk tujuan Deteksi Intrusi Jaringan. Serta diperkuat dengan adanya hasil uji perbedaan antara metode tradisional dengan metode pengolahan bigdata yang menghasilkan bahwa metode pengolahan bigdata lebih baik dalam aspek efisiensi waktu eksekusi.
2. Implementasi *K-Means Clustering* serta Bisect *K-Means Clustering* dalam deteksi intrusi jaringan telah dilakukan dan menghasilkan model yang masing – masing mewakili algoritmanya. Karakteristik yang muncul dari model tersebut berbeda untuk kedua algoritma tersebut, pada model hasil algoritma *K-Means Clustering*, aktivitas anomali memiliki besaran request dari pengguna yang besar, berasal dari referrer dengan jumlah karakter yang lebih banyak, serta memiliki okurensi karakter spesial yang lebih banyak dibanding aktivitas normal. Aktivitas normal pada model *K-Means Clustering* memiliki kecenderungan bahwa pada request line, sedikit ditemukan karakter spesial yaitu persentase, berasal dari referrer yang memiliki jumlah karakter lebih sedikit, serta diakses melalui User-Agent yang hamper sama dengan aktivitas lain. Sedangkan pada algoritma Bisecting *K-Means Clustering*, aktivitas anomali memiliki besaran request dari pengguna yang besar, request pengguna yang lebih panjang, di akses dari User-Agent yang memiliki karakter lebih pendek daripada aktivitas normal, serta memiliki okurensi garing miring lebih banyak dibanding aktivitas normal. Pada aktivitas normal, memiliki perbedaan dibanding dari *K-Means Clustering*, yaitu pada Bisecting *K-Means Clustering* aktivitas normal memiliki okurensi persentase yang lebih banyak dari model *K-Means Clustering*, memiliki besaran request yang sangat kecil, serta panjang karakter request juga yang sedikit.
3. Dari hasil evaluasi yang telah di jabarkan pada Bab Hasil dan Pembahasan, telah terlihat bahwa algoritma *K-Means Clustering* memiliki keunggulan dibandingkan algoritma Bisect *K-Means Clustering*. Keunggulan yang dimaksud adalah dalam perihal kekuatan dan akurasi dari model bentukan dalam mendeteksi aktivitas normal maupun anomali. Hal ini dilihat dari interpretasi nilai FPR *K-Means Clustering* yang lebih kecil dengan nilai sebesar 0,019%, serta Spesifikasi, Rasio Klasifikasi dan Presisi yang lebih baik dibanding dari hasil Bisect *K-Means Clustering* dengan masing – masing nilainya adalah 0,99 untuk spesifikasi , rasio klasifikasi sebesar 99,68% serta 80% presisi model bentukan dari *K-Means Clustering*.

Dari proses dan hasil penelitian yang telah dilakukan pada data akses log dengan menggunakan pengolahan bigdata untuk tujuan deteksi intrusi jaringan, penulis memberikan beberapa poin saran yang mungkin dapat dikembangkan lebih lanjut :

1. Penggunaan metode statistik yang lebih kompleks untuk mendapatkan model yang lebih akurat.
2. Penambahan node dalam ekosistem *Hadoop* dan *Spark*, untuk memaksimalkan kinerja dan efisiensi dari pengolahan *Big Data*.
3. Menggunakan dataset yang lebih besar.
4. Membangun sistem yang dapat memanfaatkan hasil penelitian ini untuk memberi notifikasi ke administrator secara streaming.

DAFTAR PUSTAKA

Chakraborty, N. (2013). *Intrusion Detection System* and Intrusion Prevention System: A Comparative Study. International Journal of Computing and Business Research.

Chandel, S. K. (2017). *Intrusion Detection System* using K-Means Data Mining and Outlier Detection Approach. Bangalore: Faculty of Informatics, Masaryk University.

Fink, G., Chappell, B., Turner, T., & O'Donoghue, K. (2002). A Metrics-Based Approach to *Intrusion Detection System* Evaluation for Distributed Real-Time Systems. Florida: WPDRTS.

Grace, L. J., Maheswari, V., & Nagamalai, D. (2011). Analysis of Web Logs and Web User in Web Mining. International Journal of Network Security and Its Application, 99-110.

GovScirt. (2018). Statistik Insiden Respon Domain .Go.Id. govcsirt.kominfo.go.id. 22 Februari 2018. https://govcsirt.kominfo.go.id/statistik-insiden-respon-domain-go-id/

Iversen, M. A. (2015). When Logs Become *Big Data*. Oslo: Department of Informatics, University of Oslo.

Meyer, R. (2008, January 26). Detecting Attacks on Web Applications from Log Files. SANS Institute Infosec Reading Room, pp. 1-42.

Mukherjee, S., & Shaw, R. (2016). Big Data - Concept, Applications, Challenges, and Future Scope. International Journal of Advanced Research in Computer and Communication Engginering, 66-74.

Nguyen, H. T., Torrano-Gimenez, C., Alvarez, G., Petrovic, S., & Franke, K. (2011). Application of the Generic Feature Selection Measure in Detection of Web Attack. Computatuional Intelligence in Security for Information Systems, 25-32.

Parthiban, P., & Selvakumar, S. (2016). Big Data Architecture for Capturing, Storing, Analyzing and Visualizing of Web Server Logs. Indian Journal of Science and Technology, 1-9.

Scott, J. A. (2015). Getting Started With Apache *Spark*. San Jose: MapR Technologies, Inc.

Seyyar, M. B. (2017). Detection of Attack-Targeted Scans from Apache HTTP Server Access Log. Istanbul: Istanbul SEHIR University.

Suneetha, K., & Krishnamoorthi, R. (2009). Identifying User Behavior by Analyzing Web Server Access Log File. International Journal of Computer Science and Network Security, 327-332.

Troesch, M., & Walsh, I. (2014). Machine Learning for Network Intrusion Detection. Stanford.

Ularu, E. G., Puican, F. C., Apostu, A., & Velicanu, M. (2012). Perspective on Big Data and Big Data Analytics. Database Systems Journal, 3-14.

Valdman, J. (2001). Log File Analysis. Pilsen*:* Department of Computer Science and Engineering, University of West Bohemia.

Vijayalakshmi, S., Mohan, V., & Raja, S. (2010). Mining of Users Access Behaviour for Frequent Sequential Pattern from Web Logs. International Journal of Database Management System (IJDMS), 31-45.

Wei, L. (2007, Oktober 23). Evaluation of *Intrusion Detection System*s. pp. 1-10.

Zhong, S., Khoshgoftaar, T., & Seliya, N. (2007). *Clustering*-based Network Intrusion Detection. International Journal of Reliability, Quality, and Safety Engineering.

POLA FERTILITAS WANITA USIA SUBUR DI INDONESIA: PERBANDINGAN TIGA SURVEI DEMOGRAFI DAN KESEHATAN INDONESIA (2002, 2007 DAN 2012)

Sukim[1], Rudi Salam[2]

Politeknik Statistika STIS
e-mail: [1]sukim@stis.ac.id, [2]rudisalam@stis.ac.id

Abstrak

Tingkat fertilitas merupakan salah satu faktor demografi yang paling menentukan dalam penurunan tingkat pertumbuhan penduduk di Indonesia. Salah satu ukuran fertilitas adalah *Total Fertility Rate* (TFR). Selama 20 tahun terakhir diketahui laju pertumbuhan penduduk di Indonesia stagnan pada angka 1,49 persen. Oleh karenanya, penelitian ini bertujuan untuk mengkaji pola TFR selama periode 20 tahun terakhir berdasarkan tiga Survei Demografi dan Kesehatan Indonesia (SDKI) tahun 2002, 2007 dan 2012. Metode yang digunakan adalah Regresi data count. Hasil penelitian menunjukkan bahwa dari ketiga SDKI tersebut, tanda koefisiennya adalah sama untuk semua variabel penjelas kecuali pada SDKI 2007 yaitu pada variabel tempat tinggal yang berbeda dengan SDKI 2002 dan 2012. Sejalan dengan temuan ini perlu studi lebih lanjut untuk mencari teori yang dapat menjelaskan temuan empirik tersebut.

Kata kunci: Fertilitas, TFR, SDKI, regresi data *count*

Abstract

Fertility rate is one of the most decisive demographic factors in the decline in the rate of population growth in Indonesia. One measure of fertility is Total Fertility Rate (TFR). During the last 20 years, the population growth rate in Indonesia is stagnant at 1.49 percent. Therefore, this study aims to examine TFR patterns over the last 20 years based on the three Indonesia Demographic and Health Survey (SDKI) in 2002, 2007 and 2012. This study used Regression data count method. The results showed that of the three SDKIs, the coefficient values are the same for all explanatory variables except in SDKI 2007 i.e. in residential variables that are different from the 2002, 2012 SDKI. In line with this finding, further studies are needed to find a theory that can explain this empirical finding.

***Keywords**: Fertility, TFR, IDHS, regression data count*

PENDAHULUAN

Fertilitas dalam istilah demografi adalah kemampuan riil seorang wanita untuk melahirkan, yang dicerminkan dalam jumlah bayi yang dilahirkan (Yasin, 1981). Fertilitas merupakan salah satu faktor demografi yang paling menentukan di dalam penurunan tingkat pertumbuhan penduduk di Indonesia yang selama 20 tahun terakhir laju pertumbuhan penduduk di Indonesia stagnan pada angka 1,49 persen. Salah satu ukuran fertilitas adalah *total fertility rate* (TFR) dan salah satu sumber data TFR adalah survei demografi dan kesehatan Indonesia (SDKI). Survei terakhir dilaksanakan tahun 2017 tetapi yang sudah di-*release* datanya adalah hasil survey tahun 2012. Berdasarkan data SDKI tahun 2012, secara nasional, tingkat fertilitas di Indonesia relatif masih cukup tinggi dan variasi antar provinsi juga cukup besar.

Gambar 1.1. menunjukkan TFR yang dihitung dari enam SDKI yang dilakukan selama periode lebih dari 20 tahun antara tahun 1991 dan 2012. Hasil SDKI menunjukkan bahwa fertilitas hanya menurun relatif sedang selama dua dekade terakhir di Indonesia, dengan perubahan yang besar terjadi antara tahun 1991 dan 2002. TFR cenderung konstan di angka 2,6 kelahiran per wanita sejak SDKI 2002 sampai tahun 2012. Untuk tahun 2012, TFR terendah adalah 2,1 anak per wanita di Provinsi DI Yogyakarta dan tertinggi adalah 3,7 anak per wanita di Provinsi Papua.

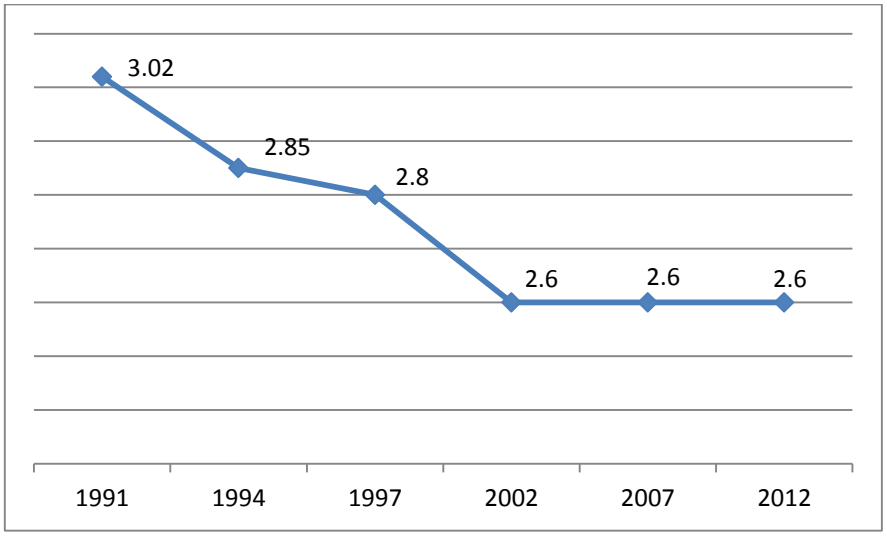
Berkaitan dengan fertilitas, dalam RPJMN 2015-2019, target pemerintah adalah Indonesia mempunyai tingkat fertilitas sebesar 2,3 anak per wanita pada tahun 2019. Dengan kondisi yang sekarang ada, sepertinya masih berat bagi pemerintah untuk memenuhi target tersebut. Oleh karena itu, berbagai upaya harus dilakukan untuk menurunkan tingkat fertilitas. Lebih jauh, dengan fakta bahwa TFR Indonesia yang stagnan pada angka 2,6 kelahiran per wanita sejak SDKI 2002 menunjukkan masih ada permasalahan serius di bidang fertilitas yang harus mendapatkan perhatian. Salah satu cara adalah dengan mengetahui faktor-faktor yang menjadi penyebab tingginya fertilitas. Keterkaitan faktor-faktor tersebut dengan fertilitas dapat didekati dengan analisis statistika yang tepat. Dengan diketahuinya faktor yang mempunyai pengaruh terhadap tingkat fertilitas diharapkan dapat dibuat kebijakan yang tepat sasaran dalam upaya menurunkan tingkat fertilitas.

Berbagai kerangka teoretis tentang perilaku dan penyebab fertilitas telah dikembangkan oleh beberapa ahli diantaranya: Davis dan Blake (1956), Freedman (1962), Hawthorne (1970), Leibenstein (1958), dan Becker (1960). Becker (1960) melihat bahwa variabel sosial ekonomi mempengaruhi fertilitas karena pengaruh mereka pada jumlah anak yang diinginkan (*demand for children*). Kemajuan dalam pembangunan menyebabkan kenaikan dalam pendapatan, dan hal ini akan meningkatkan jumlah anak yang diinginkan, karena mereka kini makin mampu membiayai jumlah anak yang lebih banyak. Easterlin (1975) menambahkan fertilitas alamiah dalam kerangka berpikir ekonom yang dipelopori oleh Becker (1960).

Adanya stagnansi fertilitas di Indonesia selama tiga SDKI terakhir menunjukkan masih ada permasalahan dalam penurunan fertilitas. Fertilitas didekati dengan jumlah anak pada setiap rumah tangga. Variabel-variabel yang diduga mempengaruhi fertilitas diantaranya adalah status bekerja istri, pendidikan istri, pendapatan rumah tangga, dan daerah tempat tinggal.

Dengan jumlah anak sebagai pendekatan untuk fertilitas, maka metode statistika yang bisa digunakan untuk analisis adalah metode regresi poisson. Namun penelitian-penelitan yang sudah dilakukan menunjukkan bahwa data fertilitas adalah *under dispersion*. Oleh karena itu, metode yang lebih tepat untuk digunakan adalah metode *generalized poisson regression*.

Adapun tujuan khusus dalam penelitian ini adalah:



Gambar 1. Tren TFR di Indonesia

1. Mendapatkan gambaran tingkat fertilitas dilihat dari beberapa karakteristik rumah tangga di Indonesia selama 2002, 2007, dan 2012.
2. Mendapatkan faktor yang berpengaruh terhadap fertilitas dan kecenderungannya di Indonesia selama 2002, 2007, dan 2012.

METODOLOGI

Metode analisis yang digunakan dalam penelitian ini adalah regresi poisson, salah satu metode Generalized Linear Model (GLM). Data yang digunakan adalah data jumlah anak lahir hidup dalam rumah tangga (Y) sebagai variabel terikat yang bersumber dari hasil Survei Demografi dan Kesehatan Indonesia (SDKI) tahun 2002, 2007, dan 2012. Untuk variabel bebas digunakan sepuluh variabel yaitu pendidikan isteri (X_1), status bekerja isteri (X_2), penggunaan kontrasepsi (X_3), umur kawin pertama (X_4), pendidikan suami (X_5), status bekerja suami (X_6), keinginan suami terhadap jumlah anak (X_7), tempat tinggal (X_8), status ekonomi (X_9), dan jumlah anak (X_{10}). Beberapa konsep dan definisi dari variabel yang berkaitan dengan *total fertility rate* (TFR), antara lain:

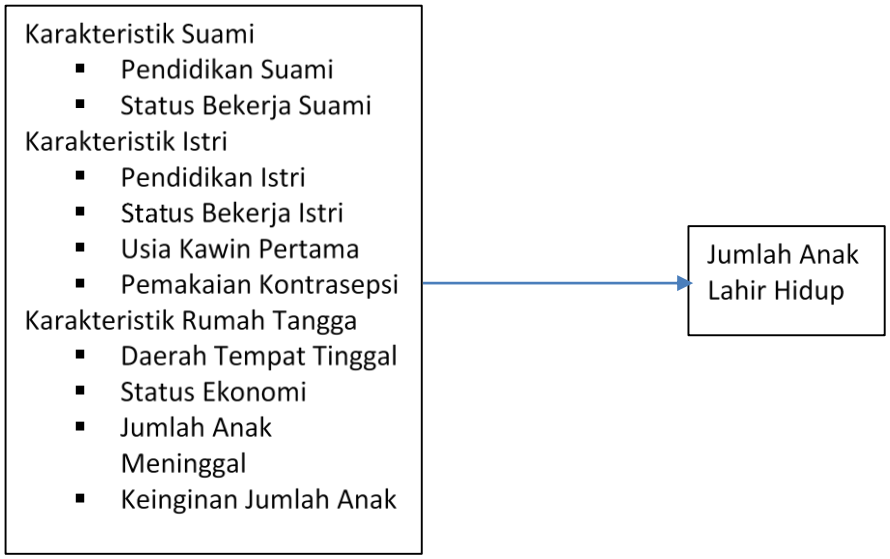
1. Fertilitas

Fertilitas sebagai istilah demografi diartikan sebagai hasil reproduksi yang nyata dari seorang wanita atau kelompok wanita. Dengan kata lain fertilitas ini menyangkut banyaknya bayi yang lahir hidup. Fertilitas mencakup peranan kelahiran pada perubahan penduduk. Istilah fertilitas adalah sama dengan kelahiran hidup (*live birth*), yaitu terlepasnya bayi dari rahim seorang perempuan dengan ada tanda-tanda kehidupan; misalnya berteriak, bernafas, jantung berdenyut, dan sebagainya (Mantra, 2003).

Seorang perempuan yang secara biologis subur (*fecund*) tidak selalu melahirkan anak-anak yang banyak, misalnya dia mengatur fertilitas dengan abstinensi atau menggunakan alat-alat kontrasepsi. Kemampuan biologis seorang perempuan untuk melahirkan sangat sulit untuk diukur. Ahli demografi hanya menggunakan pengukuran terhadap kelahiran hidup (*live birth*).

Pengukuran fertilitas lebih kompleks dibandingkan dengan pengukuran mortalitas, karena seorang perempuan hanya meninggal satu kali, tetapi ia dapat melahirkan lebih dari seorang bayi. Disamping itu seorang yang meninggal pada hari dan waktu tertentu, berarti mulai saat itu orang tersebut tidak mempunyai resiko kematian lagi. Sebaliknya seorang perempuan yang telah melahirkan seorang anak tidak berarti resiko melahirkan dari perempuan tersebut menurun.

Memperhatikan kompleksnya pengukuran terhadap fertilitas tersebut, maka memungkinkan pengukuran terhadap fertilitas ini dilakukan dengan dua macam pendekatan: *pertama*, Pengukuran Fertilitas



Gambar 2. Diagram Kerangka Pikir Penelitian

Tahunan (*Yearly Performance*) dan *kedua*, Pengukuran Fertilitas Kumulatif *(Reproductive History*).

Yearly Performance (current fertility) mencerminkan fertilitas dari suatu kelompok penduduk/berbagai kelompok penduduk untuk jangka waktu satu tahun. *Yearly Performance* terdiri dari :

1. Angka Kelahiran Kasar atau *Crude Birth Ratio* (CBR)
2. Angka Kelahiran Umum atau *General Fertility Rate* (GFR)
3. Angka Kelahiran menurut Kelompok Umur atau *Age Specific Fertility Rate* (ASFR)
4. Angka Kelahiran Total atau *Total Fertility Rate* (TFR)

Yang termasuk Reproductive History (cummulative fertility), diantaranya adalah

1. *Children Ever Born* (CEB) atau jumlah anak yang pernah dilahirkan.
2. Child Woman Ratio (CWR).

Dari beberapa penelitian sebelumnya, faktor-faktor yang berpengaruh terhadap TFR antara lain:

Status Bekerja Istri

Status bekerja istri diharapkan berhubungan negatif fertilitas. Menurut teori neoklasikal (Becker, 1960), istri yang bekerja mempunyai *opportunity cost* waktu yang lebih tinggi dibandingkan istri yang tidak bekerja. Oleh karena itu, rumah tangga dengan istri yang bekerja diharapkan memakan mempunyai anak yang lebih sedikit dibandingkan mereka yang tidak bekerja.

Pendidikan Istri

Pendidikan tertinggi istri diprediksi secara langsung berhubungan dengan *opportunity cost* dari waktu dia dan berhubungan secara berlawanan dengan keputusan fertilitas. Hubungan yang berlawanan diprediksi lebih kuat untuk istri dengan tingkat pendidikan yang lebih tinggi (Wang dan Famoye, 1997).

Pendapatan Rumah Tangga

Pengaruh dari pendapatan rumah tangga terhadap fertilitas agak ambigu. Jika anak-anak diperlakukan sebagai barang tahan lama, maka peningkatan pendapatan rumah tangga akan mempunyai pengaruh positif terhadap fertilitas, akan tetapi pendapatan bisa juga mempunyai pengaruh substitusi yang negatif. Pendekatan kuantitas-kualitas dari Becker dan Lewis (1973) memprediksi bahwa terdapat kemungkinan pengaruh substitusi dari kuantitas ke kualitas anak-anak dengan meningkatnya pendapatan. Peningkatan kualitas per anak akan berimplikasi pada peningkatan biaya membesarkan anak di mana hal ini akan menurunkan fertilitas. Pengaruh bersih dari pendapatan terhadap fertilitas tergantung pada kekuatan relatif pengaruh pendapatan terhadap pengaruh substitusi. Becker (1960) beralasan bahwa pengaruh substitusi akan besar

dibandingkan dengan pengaruh pendapatan. Dengan alasan keterbatasan data, variabel pendapatan rumah tangga akan diproksi menggunakan variabel pengeluaran rumah tangga.

Daerah Tempat Tinggal

Variabel daerah tempat tinggal dibedakan menjadi daerah perkotaan (kode 1) dan perdesaan (kode 0). Rumah tangga yang tinggal di kota akan mempunyai anak yang lebih sedikit dibandingkan rumah tangga yang tinggal di daerah perdesaan. Hal ini karena biaya membesarkan anak lebih murah di perdesaan. Selain itu, informasi mengenai kontrasepsi juga ada perbedaan antara desa dan kota.

Berdasarkan penjelasan di atas, maka disusun kerangka pikir apakah status bekerja isteri, tingkat pendidikan isteri, tingkat pendapatan rumah tangga, tipe daerah tempat tinggal berhubungan dengan jumlah anak dalam rumah tangga seperti pada Gambar 2.

2. Ruang Lingkup Penelitian

Penelitian ini menggunakan data SDKI tahun 2002, 2007, dan 2012 yang dilakukan oleh BKKBN dan BPS. Data pada penelitian ini adalah data individu wanita yang dikonversikan menjadi data rumah tangga sehingga observasi yang digunakan adalah rumah tangga.

3. Metode Analisis

Regresi Poisson

Regresi poisson merupakan analisis regresi nonlinier dari distribusi poisson, dimana analisis ini sangat cocok digunakan dalam menganalisis data diskrit (*count*). Model regresi poisson merupakan *Generalized Linier Model* (GLM) yang data respon diasumsikan berdistribusi poisson. Model regresi poisson diberikan sebagai berikut.

y_i = Poisson (μ_i), di mana $\mu_i = \exp(xT_i\,\beta)$ (1)

maka

$\ln(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ik}$ (2)

Estimasi parameter model regresi poisson menggunakan metode *Maximum Likelihood Estimator*. Fungsi log-likelihood poisson sebagai berikut.

$$\ln L(\beta) = -\sum_{i=1}^{n} \exp(x_i^T \beta) + \sum_{i=1}^{n} y_i x_i^T \beta - \sum_{i=1}^{n} \ln(y_i!) \quad (3)$$

Untuk memperoleh nilai taksiran β maka persamaan (3) diturunkan terhadap β dan disamadengankan nol menggunakan metode newton raphson.

Salah satu metode yang digunakan untuk menentukan statistik uji dalam pengujian parameter model regresi poisson adalah dengan menggunakan metode *Maximum Likelihood Ratio Test* (MLRT) dengan hipotesis:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1 : paling sedikit ada satu $\beta_i \neq 0$; $i = 1, 2, \cdots, k$

Statistik uji untuk kelayakan model regresi poisson adalah sebagai berikut.

$$D\left(\hat{\beta}\right) = -2\ln\left[\frac{L\left(\hat{\omega}\right)}{L\left(\hat{\Omega}\right)}\right] = 2\left[\ln L\left(\hat{\Omega}\right) - \ln L\left(\hat{\omega}\right)\right]$$

Keputusan yang akan diambil adalah tolak H_0 jika $D\left(\hat{\beta}\right) > \chi^2_{v,\alpha}$ dengan v adalah banyaknya parameter model dibawah populasi dikurangi dengan banyaknya parameter dibawah H_0. Parameter model regresi poisson yang telah dihasilkan dari estimasi parameter belum tentu mempunyai pengaruh yang signifikan terhadap model. Untuk itu perlu dilakukan pengujian terhadap parameter model regresi poisson secara individu.

Dengan menggunakan hipotesis sebagai berikut:

$H_0 : \beta_i = 0$

(pengaruh variabel ke-i tidak signifikan)

$H_0 : \beta_i \neq 0$

(pengaruh variable ke-i signifikan)

Statistik uji yang digunakan adalah:

$$z = \frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)}$$

Dengan $se\left(\hat{\beta}_i\right)$ adalah nilai *standar error* atau tingkat kesalahan dari parameter β_i. Keputusan yang akan diambil adalah tolak $H0$ jika $\left|z_{hitung}\right| > z_{\alpha/2}$ dimana α adalah tingkat signifikansi.

Regresi poisson dikatakan mengandung overdispersi apabila nilai variansnya lebih besar dari nilai meannya. Overdispersi memiliki dampak yang sama dengan pelang-garan asumsi jika pada data diskrit terjadi overdispersi namun tetap digunakan regresi poisson, anak dugaan dari parameter koefisien regresinya tetap konsisten namun tidak efisien. Hal ini berdampak pada nilai *standar error* yang menjadi *under estimate*, sehingga kesimpulannya menjadi tidak valid. Fenomena overdispersi (McCullagh dan Nelder [11]) dapat dituliskan *var*(Y) > *E*(Y).

Generalized Poisson Regression (GPR)

Penanganan pelanggaran asumsi equidispersi pada regresi poisson dilakukan pengembangan model menggunakan GPR. Pada model GPR selain terdapat parameter juga terdapat θ sebagai parameter dispersi. Model GPR mirip dengan regresi poisson yaitu pada persamaan (4) akan tetapi model GPR mengasumsikan bahwa komponen randomnya berdistribusi *general poisson*. Dalam analisis GPR, jika θ sama dengan 0 maka model GPR akan menjadi model poisson. Jika θ lebih dari 0 maka model GPR merepresentasikan data count yang mengandung kasus overdispersi dan jika θ kurang dari 0 merepresentasikan data *count* yang mengandung underdispersi.

Penaksiran parameter model GPR menggunakan metode *Maximum Likelihood Estimator* (MLE). Fungsi log-likelihood untuk model GPR adalah.

$$\ln L(\beta,\theta) = \sum_{i=1}^{n} y_i(x_i^T\beta) - y_i\ln(1+\theta\exp(x_i^T)\beta + (y_i-1)\ln(1+\theta y_i)) - \ln(y_{i!}) - \exp(x_i^T\beta)(1+\theta y_i)(1+\theta\exp(x_i^T\beta))^{-1} \quad (4)$$

Untuk mendapatkan taksiran parameter β dan θ maka persamaan (7) diturunkan terhadap β dan θ menggunakan metode numerik, iterasi Newton-Raphson. Pengujian parameter model GPR dilakukan sama seperti regresi poisson dengan menggunakan metode MLRT dan uji parsial menggunakan statistik uji *z*.

HASIL DAN PEMBAHASAN

Bab ini merupakan hasil pengolahan data SDKI 2002, 2007, dan 2012 menggunakan beberapa paket program. Untuk analisis deskriptif digunakan program Microsoft Excel, dan untuk inferensia digunakan STATA. Pada bab ini pula akan disajikan karakteristik umum fertilitas dari wanita usia subur untuk ketiga SDKI dan variabel-variabel yang memengaruhinya.

1. Gambaran Tingkat Fertilitas Dilihat dari Beberapa Karakteristik Rumah Tangga di Indonesia Selama 2002, 2007, Dan 2012

Fertilitas merupakan komponen pertumbuhan penduduk yang bersifat menambah jumlah penduduk. Pertumbuhan penduduk yang terlalu besar akan mengakibatkan berbagai masalah kependudukan seperti pengangguran, kemiskinan, dan masalah lainnya. Untuk itu diperlukan suatu pengendalian kelahiran supaya pertumbuhan penduduk tidak menjadi permasalahan yang besar.

Hasil dari perbandingan tiga SDKI dalam penelitian ini diharapkan dapat dengan lebih komprehensif melihat faktor-faktor apa saja yang memengaruhi fertilitas di Indonesia dan dapat menentukan faktor prioritas jika ada beberapa keterbatasan.

2. Faktor yang Berpengaruh Terhadap Fertilitas di Indonesia Selama 2002, 2007, dan 2012

Uji Goodness of Fit untuk Regresi Poisson

Pada penggunaan regresi Poisson, variabel respon pada data memiliki distribusi Poisson dan memiliki nilai *mean* yang sama dengan nilai *varians* ($\mu = \sigma^2$) atau dikenal dengan *equidispersion*. Untuk melihat apakah variabel respons berdistribusi Poisson atau tidak, dilakukan uji Kolmogorov-Smirnov untuk distribusi Poisson. Selain uji ini, dapat juga digunakan uji Anderson Darling.

Hasil kedua uji menunjukkan bahwa variabel respon fertilitas untuk ketiga SDKI mempunyai distribusi Poisson. Hasil kedua uji dapat dilihat pada Tabel 1

Tabel 1. Hasil Pengolahan Uji *Kolmogorov-Smirnov* dan Uji *Anderson Darling*

SDKI	Statistic Uji Kolmogorov-Smirnov	Statistik Uji Anderson-Darling	Keterangan
SDKI 2002	0.1918	1384.8	Berdistribusi Poisson
SDKI 2007	0.21476	1703.8	Berdistribusi Poisson
SDKI 2012	0.23726	3460.8	Berdistribusi Poisson

Tabel 2. Hasil Uji *Equidispersion*

SDKI	Nilai *Pearson Chi-Square* dibagi dengan derajat bebas	Keterangan
SDKI 2002	0.9398691	*Underdispersion*
SDKI 2007	0.8948355	*Underdispersion*
SDKI 2012	0.8678478	*Underdispersion*

Pendeteksian Equidispersi

Regresi Poisson bisa digunakan jika bisa memenuhi kondisi equidispersi atau mempunyai nilai *mean* dan *varians* yang sama. Jika kondisi tersebut tidak terpenuhi, maka telah terjadi *overdispersion* atau *underdispersion*. Suatu nilai yang bisa digunakan untuk menguji kondisi tersebut adalah nilai *Pearson Chi-Square*. Hasil uji *equidispersion* dapat dilihat pada Tabel 2.

Berdasarkan hasil pengujian equidispersi regresi Poisson, dapat disimpulkan bahwa terjadi underdispersi pada model yang digunakan, yang ditunjukkan dari hasil Pearson's Chi-Square dibagi dengan derajat bebas yang kurang dari satu. Keadaan ini mengakibatkan penggunaan regresi Poisson kurang sesuai untuk memodelkan variabel-variabel yang memengaruhi jumlah anak lahir hidup dari wanita usia subur di Indonesia. Untuk menangani penggunaan regresi Poisson yang tidak memenuhi asumsi equidispersi, dapat diterapkan metode regresi Poisson yang tergeneralisir, yaitu *Generalized Poisson Regression* yang dapat menanangi kondisi underdispersi maupun overdispersi pada regresi Poisson.

3. Model GPR pada Fertilitas WUS di Indonesia

Berdasarkan hasil estimasi model GPR, Tabel 3 menampilkan nilai dari *likelihood ratio* χ^2 dan *p-value* untuk tiap-tiap SDKI. *Likelihood ratio* adalah suatu *statistic* uji dari uji simultan apakah variabel bebas secara bersama-sama berpengaruh terhadap variabel terikat atau minimal satu variabel bebas yang berpengaruh terhadap terikat. Dari nilai *likelihood ratio* yang diperoleh dapat disimpulkan bahwa uji adalah tolak hipotesis nol atau minimal ada satu variabel bebas yang berpengaruh terhadap variabel terikat untuk semua tiga SDKI yang ada.

Setelah uji simultan menghasilkan keputusan menolak hipotesis nol, maka untuk mengetahui variabel mana saja yang berpengaruh terhadap fertilitas perlu dilanjutkan dengan pengujian secara parsial. Untuk SDKI 2002, hasil pengujian secara parsial terlihat pada Tabel 4

Pada SDKI 2002, berdasarkan hasil p-value dari tiap-tiap variabel dapat disimpulkan bahwa dengan tingkat signifikansi 5 persen estimasi parameter untuk semua variabel adalah signifikan memengaruhi jumlah anak lahir hidup wanita usia subur kecuali variabel pendidikan suami. Persamaan GPR yang terbentuk adalah

$$\begin{aligned}\ln(\hat{\mu}) = 1.6767 &- 0.0851x_1 + 0.1214x_2 \\ &+ 0.0834x_3 - 0.0405x_4 \\ &+ 0.0229x_3 - 0.1001x_6 \\ &- 0.0939x_7 + 0.0317x_8 \\ &- 0.0417x_9 + 0.3087x_{10}\end{aligned}$$

Pada SDKI 2007, dengan menggunakan tingkat signifikansi sebesar 5

Tabel 3. Nilai *Likelihood Ratio* χ^2 dan *p-value*

SDKI	*likelihood ratio* χ^2	*p-value*	Keterangan
SDKI 2002	6755.35	0.0000	Signifikan
SDKI 2007	9807.89	0.0000	Signifikan
SDKI 2012	11910.79	0.0000	Signifikan

Tabel 4. Hasil Pengujian Secara Parsial data SDKI 2002

Variabel	Coef.	Std. Err.	*p-value*	Selang Kepercayaan		Keputusan
X_1	-0.0851	0.0239	0.0000	-0.1318	-0.0383	Tolak H0
X_2	0.1214	0.0086	0.0000	0.1045	0.1383	Tolak H0
X_3	0.0834	0.0088	0.0000	0.0662	0.1006	Tolak H0
X_4	-0.0405	0.0012	0.0000	-0.0429	-0.0382	Tolak H0
X_5	**0.0229**	**0.0198**	**0.2490**	**-0.0160**	**0.0617**	**Tidak Tolak H0**
X_6	-0.1001	0.0266	0.0000	-0.1522	-0.0480	Tolak H0
X_7	-0.0939	0.0104	0.0000	-0.1143	-0.0734	Tolak H0
X_8	0.0317	0.0102	0.0020	0.0116	0.0518	Tolak H0
X_9	-0.0417	0.0102	0.0000	-0.0616	-0.0217	Tolak H0
X_{10}	0.3087	0.0052	0.0000	0.2985	0.3189	Tolak H0
_cons	1.6767	0.0358	0.0000	1.6066	1.7469	Tolak H0

persen, dapat disimpulkan estimasi parameter untuk semua variabel adalah signifikan memengaruhi jumlah anak lahir hidup wanita usia subur kecuali variabel status bekerja suami dan variabel tempat tinggal.

Persamaan GPR yang terbentuk adalah

$$\begin{aligned}\ln(\hat{\mu}) = 1.5420 &- 0.0988x_i + 0.0908x_2 \\ &+ 0.108x_3 + 0.078x_4 \\ &+ 0.0496x_5 - 0.0361x_6 \\ &- 0.0737x_7 - 0.0031x_8 \\ &- 0.500x_9 + 0.3281x_{10}\end{aligned}$$

Pada SDKI 2012, berdasarkan hasil p-value dari tiap-tiap variabel dapat disimpulkan bahwa dengan tingkat signifikansi 5 persen estimasi parameter untuk semua variabel adalah signifikan memengaruhi jumlah anak lahir hidup wanita usia subur kecuali variabel tempat tinggal Persamaan GPR yang terbentuk adalah

$$\begin{aligned}\ln(\hat{\mu}) = 1.4847 &- 0.1316x_1 + 0.0883x_2 \\ &+ 0.1579x_3 - 0.0364x_4 \\ &+ 0.0818x_5 - 0.0628x_6 \\ &- 0.1109x_7 + 0.0072x_8 \\ &- 0.0347x_9 + 0.3490x_{10}\end{aligned}$$

4. Perbandingan Model Fertilitas Tiga SDKI

Pada subbab berikut akan dijelaskan interpretasi untuk tiap-tiap variabel pada tiap-tiap SDKI dan membandingkan hasilnya pada tiga SDKI terakhir.

Hasil pengolahan dari tiga SDKI menunjukkan bahwa untuk tanda dari koefisien adalah sama untuk semua variabel kecuali pada SDKI 2007 di mana pada SDKI 2007 variabel tempat tinggal (X_8) mempunyai tanda negatif sedangkan pada SDKI 2002 dan 2012 mempunyai tanda positif.

Tabel 5. Hasil Pengujian Secara Parsial Data SDKI 2007

Variabel	Coef.	Std. Err.	p-value	Selang Kepercayaan		Keputusan
X_1	-0.0988	0.0203	0.0000	-0.1387	-0.0589	Tolak H0
X_2	0.0908	0.0082	0.0000	0.0747	0.1069	Tolak H0
X_3	0.1083	0.0083	0.0000	0.0920	0.1245	Tolak H0
X_4	-0.0378	0.0011	0.0000	-0.0399	-0.0357	Tolak H0
X_5	0.0496	0.0175	0.0040	0.0154	0.0838	Tolak H0
X_6	**-0.0361**	**0.0241**	**0.1340**	**-0.0834**	**0.0112**	**Tidak Tolak H0**
X_7	-0.0737	0.0097	0.0000	-0.0927	-0.0548	Tolak H0
X_8	**-0.0031**	**0.0095**	**0.7460**	**-0.0218**	**0.0156**	**Tidak Tolak H0**
X_9	-0.0500	0.0093	0.0000	-0.0683	-0.0317	Tolak H0
X_{10}	0.3281	0.0043	0.0000	0.3198	0.3365	Tolak H0
_cons	1.5420	0.0330	0.0000	1.4773	1.6068	Tolak H0

Tabel 6. Hasil Pengujian Secara Parsial Data SDKI 2012

Variabel	Coef.	Std. Err.	p-value	Selang Kepercayaan		Keputusan
X_1	-0.1316	0.0159	0.0000	-0.1628	-0.1004	Tolak H0
X_2	0.0883	0.0076	0.0000	0.0734	0.1032	Tolak H0
X_3	0.1579	0.0076	0.0000	0.1430	0.1728	Tolak H0
X_4	-0.0364	0.0009	0.0000	-0.0382	-0.0345	Tolak H0
X_5	0.0818	0.0146	0.0000	0.0533	0.1104	Tolak H0
X_6	-0.0628	0.0232	0.0070	-0.1083	-0.0173	Tolak H0
X_7	-0.1109	0.0076	0.0000	-0.1257	-0.0961	Tolak H0
X_8	**0.0072**	**0.0082**	**0.3790**	**-0.0089**	**0.0234**	**Tidak Tolak H0**
X_9	-0.0347	0.0083	0.0000	-0.0509	-0.0185	Tolak H0
X_{10}	0.3490	0.0043	0.0000	0.3405	0.3574	Tolak H0
_cons	1.4847	0.0296	0.0000	1.4267	1.5426	Tolak H0

Tabel 7. IRR (*Incidence Rate Ratio*) pada Tiga SDKI

Variabel	2002		2007		2012	
	Coef.	IRR	Coef.	IRR	Coef.	IRR
X_1	-0.0851	0.9185	-0.0988	0.9059	-0.1316	0.8767
X_2	0.1214	1.1291	0.0908	1.0950	0.0883	1.0923
X_3	0.0834	1.0869	0.1083	1.1143	0.1579	1.1710
X_4	-0.0405	0.9603	-0.0378	0.9629	-0.0364	0.9643
X_5	**0.0229**	**1.0231**	0.0496	1.0509	0.0818	1.0853
X_6	-0.1001	0.9048	**-0.0361**	**0.9645**	-0.0628	0.9391
X_7	-0.0939	0.9104	-0.0737	0.9289	-0.1109	0.8950
X_8	0.0317	1.0322	**-0.0031**	**0.9969**	**0.0072**	**1.0073**
X_9	-0.0417	0.9592	-0.0500	0.9512	-0.0347	0.9659
X_{10}	0.3087	1.3617	0.3281	1.3883	0.3490	1.4176

Dari penghitungan IRR (*Incidence Rate Ratio*), misalkan Variabel X_1 mempunyai IRR = Exp(β_1) = 0.92 artinya wanita dengan pendidikan lebih dari sltp akan memiliki jumlah anak lahir hidup sebesar 0.92 kali dibandingkan dengan yang kurang dari atau sama dengan sltp. Demikian juga untuk variabel bebas yang lain.

KESIMPULAN DAN SARAN

1. Kesimpulan

Berdasarkan hasil pembahasan yang diperoleh dari bab-bab sebelumnya, dapat ditarik beberapa kesimpulan penelitian sebagai berikut:

1. Tanda dari koefisien adalah sama untuk semua variabel kecuali pada SDKI 2007 di mana pada SDKI 2007 variabel tempat tinggal (X_8) mempunyai tanda negatif sedangkan pada SDKI 2002 dan 2012 mempunyai tanda positif
2. Pada SDKI 2002 hanya variabel pendidikan suami (X_5) yang tidak signifikan.
3. Pada SDKI 2007, variabel yang tidak signifikan adalah variabel status bekerja suami (X_6) dan variable tempat tinggal (X_8).
4. Pada SDKI 2012, variable yang tidak signifikan hanyalah variabel tempat tinggal (X_8).

2. Saran

Berdasarkan hasil dan kesimpulan yang telah diperoleh, maka peneliti dapat memberikan saran sebagai berikut:

1. Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) sebaiknya terus mensosialisasikan program KB khususnya penggunaan kontrasepsi modern yang efektif.
2. Wanita usia subur perlu lebih meningkatkan keterlibatan suami dalam penentuan jumlah anak yang diharapkan untuk lebih meningkatkan kesehatan anak, sehingga peluang kematian anak menjadi kecil dan dapat mengurangi jumlah anak yang dilahirkan
3. Untuk penelitian selanjutnya, dapat memasukkan variabel kontekstual atau spasial karena keberagaman wilayah di Indonesia.

DAFTAR PUSTAKA

Badan Kependudukan dan Keluarga Berencana Nasional, Badan Pusat Statistik, dan Kementerian Kesehatan Republik Indonesia. 2012. *Pedoman Survei Demografi dan Kesehatan Indonesia.* Agustus 2013. http://kesga.kemkes.go.id/images/pedoman/SDKI%202012-Indonesia.pdf

Becker Gary S. *An Economic Analysis of Fertility. In: Roberts George B, Chairman, Universities-National Bureau Committee for Economic Research , editor. Demographic and Economic Change in Developed Countries*. Columbia University Press. National Bureau of Economic Research; 1960. pp. 209–240. http://www.nber.org/chapters/c2387.

Davis, K. and J. Blake. 1956. *Social structure and fertility: an analytic framework. Economic and Cultural Change 4*(2):211-235.

Famoye F (1993) *Restricted generalized poisson regression model. Communications in Statis¬tics — Theory and Methods* 22:1335-1354

Friedman, Debra, Michael Hetcher, and Sathoshi Kanazawa. 1994. *A Theory of the Value of Children*. *Demography* 31: 375-401.

Freedman, Ronald. 1962. *The Sociology of Human Fertility: a Trend Report and Bibliography* 11 (2): 35-68

Gustavo Angeles, David K. Guilkey, and Thomas A. Mroz. *The Effects of Female Education and Health and Family Planning Programs on Child Mortality and Fertility in Indonesia. MEASURE Evaluation Working Papers No. wp-03-73-en. Carolina Population Center, 2003*

Michael Grimm, Robert Sparrow, Luca Tasciotti. *Does Electrification Spur the Fertility Transition? Evidence From Indonesia. Demography* 52(5): 1773–1796, 2015.

Wang S-X, Chen Y-D, Chen CHC, Rochat R, Chow LP, Rider R. *Proximate determinants of fertility and policy implications in Beijing*. Studies in Family Planning 18(4):222-228, 198

Petunjuk Penulisan
JURNAL APLIKASI STATISTIKA & KOMPUTASI STATISTIK

Naskah dikirim dalam bentuk *softcopy* ke alamat email pppm@stis.ac.id disertai dengan daftar riwayat hidup ringkas penulis. Format naskah mengacu pada Petunjuk Penulisan Naskah berikut:

Naskah dibuat menggunakan *Microsot Office Word* 2010. Seluruh bagian dalam naskah diketik dengan huruf *Times New Roman*, ukuran 12, spasi 1,5, ukuran kertas A4 dan marjin 2 cmuntuk semua sisi, serta jumlah halaman 15-20. Untuk kepentingan penyuntingan naskah, seluruh bagian naskah (termasuk tabel, gambar dan persamaan matematika) dibuat dalam format yang dapat disunting oleh editor.

Gaya penulisan naskah untuk Jurnal Aplikasi Statistika dan Komputasi Statistik ditulis dalam Bahasa Indonesia dengan gaya naratif. Pembabakan dibuat sederhana dan sedapat mungkin menghindari pembabakan bertingkat. Tabel dan gambar harus mencantumkan sumber jika dari data sekunder. Tabel, gambar dan persamaan matematika diberi nomor secara berurut sesuai dengan kemunculannya. Semua kutipan dan referensi dalam naskah harus tercantum dalam daftar pustaka, dan sebaliknya sumber bacaan yang tercantum dalam daftar pustaka harus ada dalam naskah. Format sumber: Nama Penulis dan Tahun. Nomor dan judul table diletakkan di bagian atas table dan dicetak tebal, sedangkan nomor dan judul gambar diletakkan di bagian bawah gambar dan dicetak tebal.

Bagian naskah berisi:

Judul. Judul tidak melebihi 12 kata dalam Bahasa Indonesia.

Data Penulis. Berisi nama lengkap semua penulis tanpa gelar, asal institusi, dan alamat email.

Abstrak. Ditulis dalam Bahasa Inggris dan Bahasa Indonesia, maksimum 100 kata untuk masing-masing abstrak dan berisikan tiga hal yaitu topik yang dibahas, metodologi yang dipergunakan dan hasil yang didapatkan.

Kata Kunci. Berisi kata atau frasa (maksimum 5 subjek) yang sering dipergunakan dalam naskah dan dianggap mewakili dan atau terkait dengan topik yang dibahas.

Pendahuluan. Memuat latar belakang, studi sebelumnya yang relevan, permasalahan ataupun hipotesis yang akan diuji dalam penelitian, ruang lingkup penelitian, serta tujuan dari penelitian.

Metodologi terdiri atas:

a. **Tinjauan Referensi.** Bagian ini menguraikan landasan konseptual dari tulisan dan berisi alasan teoritis mengapa pertanyaan penelitian dalam artikel diajukan. Di samping itu penulis dapat mengutip studi yang relevan sebelumnya untuk melengkapi justifikasi mengenai kerangka pikir penelitian.
b. **Metode Analisis.** Bagian ini berisi informasi teoritis dan teknis yang cukup memadai untuk pembaca dapat mereproduksi penelitian dengan baik termasuk di dalamnya uraian mengenai jenis dan sumber data serta variabel yang digunakan. Dalam hal keperluan verifikasi hasil, editor dan mitra bestari (*reviewer*) berhak meminta data mentah (*raw data*) yang digunakan penulis.

Hasil dan Pembahasan. Tuliskan hasil yang didapat berdasarkan metode yang digunakan disertai analisis terhadap variabel-variabelnya . Dapat disajikan berupa tabel, gambar, hasil pengujian hipotesis dengan disertai uraian analitis yang mengangkat poin-poin penting berdasarkan konsepsi teoritisnya.

Kesimpulan dan Saran. Bagian ini memuat kesimpulan dari hasil dan implikasinya secara akademis, dan saran yang dapat diberikan berdasarkan temuan dari pembahasan. Bagian ini juga memuat keterbatasan penelitian dan kemungkinan penelitian lanjutan yang dapat dilakukan dengan penggunaan/pengembangan variabel, metode analisis ataupun cakupan wilayah penelitian lainnya.

Daftar Pustaka. Daftar pustaka disusun berdasarkan urutan abjad dengan ketentuan sebagai berikut:

Publikasi Buku

1. Penulis satu orang
   Enders, Walter. 2010. *Applied Econometric Time Series, Third Edition*. New Jersey: Wiley.
2. Penulis dua orang
   Pyndick, Robert. S. dan Rubinfeld, Daniel L. 2009. *Microeconomics, Seventh Edition*. New Jersey: Pearson Education.
3. Penulis tiga orang
   Fotheringham, A. S., Brunsdon, C, dan Charlton, M. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. West Sussex: John Wiley & Sons.

Artikel dalam jurnal

Romer, P. 1993. Idea Gaps and Object Gaps in Economic Development. *Journal of Monetary Economics*, Vol. 32 (3), 543–573.

Artikel online

Woodward, Douglas P. 1992. Locational Determinants of Japanese Manufacturing Start-Ups in the United States. *Southern Economic Journal*, Vol. 58 (3), 690-708. http://www.jstor.org/discover/10.2307/1059836 (Diakses 1 Sepetember, 2014).

Buku yang ditulis oleh lembaga atau organisasi

BPS. 2009. *Analisis dan Penghitungan Tingkat Kemiskinan 2008*. Jakarta: BPS.

Kertas kerja (*working papers*)

Edwards, S. 1990. Capital Flows, Foreign Direct Investment, and Debt-Equity Swaps in Developing Countries. *NBER Working Paper*, 3497.

Makalah yang direpresentasikan

Zhang, Kevin H. 2006. Foreign Direct Investment and Economic Growth in China: A Panel Data Study for 1992-2004. *Conference of WTO, China, and Asian Economies*. Beijing.

Karya yang tidak dipublikasikan

Hartono, Djoni. 2002. Analisis Dampak Kebijakan Harga Energi terhadap Perekonomian dan Distribusi Pendapatan di DKI Jakarta: Aplikasi Model Komputasi Keseimabangan Umum (Computable General Equilibrium Model. *Tesis*. Jakarta.

Artikel di koran, majalah, dan periodik sejenis

Reuters. (2014, September 17). Where is Inflation?. *Newsweek*.