TESIS SS14 2501

IMPUTASI DATA HILANG PADA SURVEI INDUSTRI BESAR SEDANG SUMATERA UTARA MENGGUNAKAN *FUZZY C-MEANS* DIOPTIMALKAN DENGAN ALGORITMA GENETIKA

ERVIN NODERIUS MEI BUNAWOLO

NRP. 1315201710

DOSEN PEMBIMBING :
Dr. Irhamah, S.Si., M.Si

Dr. Brodjol Sutijo Suprih Ulama, M.Si

PROGRAM MAGISTER

JURUSAN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2017

THESIS SS14 2501

MISSING DATA IMPUTATION ON LARGE AND MEDIUM MANUFACTURING SURVEY OF SUMATERA UTARA USING *FUZZY C-MEANS* OPTIMIZED WITH GENETIC ALGORITHM

ERVIN NODERIUS MEI BUNAWOLO

NRP. 1315201710

SUPERVISOR :
Dr. Irhamah, S.Si., M.Si

Dr. Brodjol Sutijo Suprih Ulama, M.Si

PROGRAM OF MAGISTER

DEPARTMENT OF STATISTICS

FACULTY OF MATHEMATICS AND NATURAL SCIENCES

SEPULUH NOPEMBER INSTITUTE OF TECHNOLOGY

SURABAYA

2017

IMPUTASI DATA HILANG PADA SURVEI INDUSTRI BESAR SEDANG SUMATERA UTARA MENGGUNAKAN *FUZZY C-MEANS* DIOPTIMALKAN DENGAN ALGORITMA GENETIKA

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Sains (M.Si)
di
Institut Teknologi Sepuluh Nopember
Oleh :

ERVIN NODERIUS MEI BUNAWOLO
NRP. 1315201710

Tanggal Ujian : 4 Januari 2017
Periode Wisuda : Maret 2017

Disetujui oleh :

1. Dr. Irhamah, S.Si., M.Si. (Pembimbing I)
   NIP. 19780406 200112 2 002
2. Dr. Brodjol Sutijo Suprih Ulama, M.Si. (Pembimbing II)
   NIP. 19660125 199002 1 001
3. Santi Puteri Rahayu, M.Si., Ph.D. (Penguji)
   NIP. 19750115 199903 2 003
4. R. Mohamad Atok, S.Si., M.Si., Ph.D. (Penguji)
   NIP. 19710915 199702 1 001
5. Dr. Erni Tri Astuti, M.Math. (Penguji)
   NIP. 19671022 199003 2 002

an. Direktur Program Pascasarjana
Asisten Direktur

Prof. Dr. Ir. Tri Widjaja, M.Eng.
NIP. 19611021 198603 1 001

Direktur Program Pascasarjana,

Prof. Ir. Djauhar Manfaat, M.Sc., Ph.D.
NIP. 19601202 198701 1 001

IMPUTASI DATA HILANG PADA SURVEI INDUSTRI BESAR SEDANG SUMATERA UTARA MENGGUNAKAN *FUZZY C-MEANS* DIOPTIMALKAN DENGAN ALGORITMA GENETIKA

Nama Mahasiswa	: Ervin Noderius Mei Bunawolo
NRP	: 1315201710
Pembimbing I	: Dr. Irhamah, S.Si, M.Si
Pembimbing II	: Dr. Brodjol Sutijo Suprih Ulama, M.Si

ABSTRAK

Nonrespon pada suatu survei mengakibatkan data tidak lengkap sehingga menyebabkan potensi bias dalam estimasi parameter dan memperlemah generalisasi penelitian. Salah satu cara yang dapat dilakukan untuk mengatasi hal ini adalah dengan mengganti data yang *missing* dengan suatu nilai yang dikenal sebagai metode imputasi. Penelitian ini menggabungkan *Fuzzy C-Means* yang merupakan metode imputasi menggunakan metode klaster dalam menangani *missing data*, dengan Algoritma Genetika yang diharapkan akan meningkatkan akurasi *Fuzzy C-Means*. *Root Mean Square Error* (RMSE) dipakai sebagai kriteria untuk mengevaluasi kinerja kedua metode. Hasil penelitian menunjukkan bahwa kinerja imputasi metode hibrida FCM-GA secara umum lebih baik dibandingkan metode imputasi FCM konvensional khususnya pada persentase *missing* yang tidak terlalu besar. Pilihan nilai *weighting exponent m* dan jumlah klaster *c* yang sesuai sangat berpengaruh terhadap nilai RMSE yang dihasilkan. Dengan kata lain kinerja imputasi FCM bergantung dari ketepatan pemilihan parameter *c* dan *m*. Pemilihan ukuran jarak Euclidean dan Manhattan dibawah kombinasi c dan *m* tidak menunjukkan perbedaan dan kinerja dari dua ukuran jarak pada tingkat *missing* yang berbeda mengindikasikan tidak ada yang lebih unggul diantara keduanya. Hibrida Fuzzy C-Means-GA memberikan parameter *c* dan *m* yang sesuai untuk menghasilkan imputasi dengan nilai RMSE yang lebih kecil dibandingkan FCM dan mengatasi data tidak lengkap industri sedang pada hasil survei industri besar dan sedang di Sumatera Utara.

Kata kunci : Algoritma genetika, *Fuzzy C-Means*, Imputasi, Klaster.

MISSING DATA IMPUTATION ON LARGE AND MEDIUM MANUFACTURING SURVEY OF SUMATERA UTARA USING FUZZY C-MEANS OPTIMIZED WITH GENETIC ALGORITHM

By	: Ervin Noderius Mei Bunawolo
Student Identity Number	: 1315201710
Supervisor I	: Dr. Irhamah, S.Si, M.Si
Supervisor II	: Dr. Brodjol Sutijo Suprih Ulama, M.Si

ABSTRACT

Nonrespon on a survey result in incomplete data and might cause potential bias in the estimation of parameters and weaken the generalizability of the study. One way that could be done to overcome this problem is to replace missing data with a value known as the imputation method. This research combines the Fuzzy C-Means, which is the imputation method using cluster method in dealing with missing data, with the Genetic Algorithm which is expected to improve the accuracy of Fuzzy C-Means. Root Mean Square Error (RMSE) is used as the performance evaluation criteria on both methods. The results showed that the performance of hybrid FCM-GA is generally out performs the conventional FCM imputation method particularly on relatively small percentage of missing values. Appropriate chosen of weighting exponent m and the number of clusters c affected on RMSE values. In other words, the performance of FCM imputation depends on the right choice of parameters c and m. The Euclidean Distance and the Manhattan Distance under a combination of c and m did not show differences in the performance of the two distances measurement at several missing levels, thus indicates there is none more or less superior distance measurement of the two. The Hybrid Fuzzy C-Means-GA provides the parameters c and m which are best to produce imputation with smaller value of RMSE than the FCM and to handle incomplete data problem on medium scale industry from large and medium manufacturing survey in Sumatera Utara.

Key words : Genetic Algorithm, Fuzzy C-Means, Imputation, Cluster.

KATA PENGANTAR

Perjalanan yang dimulai kurang dari dua tahun yang lalu mengantarkan pada sebuah pengalaman luar biasa untuk mendapatkan guru, teman, dan pengetahuan yang menambah ilmu, wawasan, dan khazanah batin. Setiap langkah diwarnai oleh banyak kesulitan, kegundahan, dan keletihan sebanyak kemudahan, kesenangan dan kegembiraan yang didapatkan. Tesis ini adalah manifestasi dari proses itu.

Terima kasih untuk Pimpinan BPS di pusat dan unit kerja daerah yang telah memberi kesempatan dan dukungan selama menempuh pendidikan, juga buat mba Yuli dkk. di Pusdiklat yang sabar mengurus keperluan kita semua.

Terima kasih untuk jajaran pimpinan ITS, seluruh karyawan, dosen, dosen wali, dosen penguji dan dosen pembimbing yang sudah memberi dan berbagi banyak hal melampaui ilmu pengetahuan yang diberikan.

Untuk teman-teman Angkatan 9, terima kasih untuk uluran tangan pertemanan dan kebersamaan selama di Surabaya. Terima kasih untuk '*someone*' yang selalu bermurah hati untuk berbagi. Buat mas Syahrul terima kasih untuk *sharing* ilmu dan bantuan yang tak ternilai. Tanpa beliau tesis ini tidak akan selesai.

Terima kasih buat banyak pihak yang layak disebutkan satu persatu, yang kelihatan dan yang tidak kelihatan, untuk doa, dukungan dan kebaikan yang diberikan. Semoga yang terbaik dicurahkan berlimpah bagi kehidupan kalian.

Terima kasih buat keluarga, secara khusus untuk kedua orang tua yang selalu mendoakan yang terbaik buat anak-anak mereka. Puji syukur dan hormat bagi Tuhan Yesus Kristus sumber kekuatan hati, tempat penghiburan, yang selalu memberi kelegaan dan harapan. *You are the Greatest*.

Akhir kata, terima kasih bagi pembaca budiman yang berkenan meluangkan waktu membaca tesis ini yang notabene mengandung banyak kelemahan. Terima kasih.

Surabaya, Januari 2017

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Sektor industri adalah sektor penting dalam perekonomian Indonesia. Hal ini ditandai dengan kontribusi sektor industri pengolahan terhadap Produk Domestik Bruto (PDB) yang konsisten cukup besar dari tahun ke tahun. Pada tahun 2013 kontribusi dari sektor ini sebesar 21,03 persen, terbesar dibandingkan dengan sektor-sektor lainnya. Untuk memperoleh data di sektor industri pengolahan , BPS setiap tahun melakukan pengumpulan data industri besar dan sedang melalui Survei Tahunan Perusahaan Industri Manufaktur. Survei ini dilaksanakan dengan mengirimkan kuesioner kepada seluruh perusahaan skala besar dan sedang di seluruh wilayah Indonesia. Perusahaan-perusahaan tersebut diminta untuk melaporkan banyaknya pekerja/karyawan rata-rata per hari kerja dalam satu tahun, pengeluaran untuk pekerja dibayar termasuk upah/gaji, pemakaian bahan bakar dan pelumas, tenaga listrik yang dibeli, pengeluaran lain seperti sewa gedung, pajak, telepon, air, pos dan lain-lain, bahan baku dan bahan penolong yang digunakan, jenis barang yang dihasilkan atau diproduksi, nilai produksi dan nilai taksiran barang modal tetap perusahaan.

Di Sumatera Utara, peranan sektor ini berada pada peringkat kedua setelah sektor pertanian. Pada tahun 2013 sektor industri pengolahan menyumbang 19,86 persen terhadap pembentukan Produk Domestik Regional Bruto (PDRB) Sumatera Utara. Rata-rata kontribusi selama 5 tahun terakhir sebesar 20,58 persen (BPS Sumatera Utara, 2015). Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara pada tahun 2013 dilakukan pada 1.006 perusahaan/industri pengolahan berskala besar dan sedang yang tercatat. Terdiri dari 352 perusahaan (35 persen) yang tergolong industri besar dengan tenaga kerja diatas 100 orang, dan 654 perusahaan (65 persen) tergolong industri sedang dengan jumlah tenaga kerja 20 sampai 99 orang. Bila dilihat dari jenis produksi barang atau jasa industri yang dihasilkan, tiga industri terbesar adalah industri makanan dengan 435 perusahaan (43 persen). Diikuti oleh industri karet, barang dari karet dan plastik dengan 146 perusahaan (15 persen). Kemudian industri

kayu, barang dari kayu dan gabus (selain furnitur) dan barang anyaman dari bambu, rotan dan sejenisnya 76 perusahaan (8 persen).

Dalam pelaksanaan survei ini kuesioner yang telah disampaikan kepada perusahaan seringkali tidak dikirimkan kembali ke BPS pada waktu yang telah ditetapkan. Kuesioner yang telah dijawab oleh perusahaan juga masing banyak yang tidak diisi dengan lengkap. Sehingga menyebabkan terjadinya nonrespon unit pengamatan dan nonrespon item pertanyaan. Pada survei ini tahun 2013 di Sumatera Utara, jumlah perusahaan yang respon adalah sebanyak 844 perusahaan (84 persen), dan dari perusahaan yang respon tersebut masih ditemukan sebagian perusahaan yang tidak memberikan jawaban pada beberapa item pertanyaan.

Missing data dapat terjadi pada dua tingkatan atau level yaitu unit atau item. Nonrespon pada tingkat unit terjadi ketika tidak ada informasi yang dikumpulkan dari responden. Contohnya jika responden menolak mengikuti survei atau tidak mengisi kuesioner. Nonrespon pada tingkat item merujuk pada ketidaklengkapan informasi yang dikumpulkan dari responden. Misalnya jika responden tidak mengisi satu atau dua pertanyaan dalam sebuah survei tetapi menjawab pertanyaan yang lain.

Missing data pada umumnya bukan menjadi hal yang utama pada studi yang bersifat substantif, namun jika tidak ditangani secara tepat dapat mengakibatkan masalah yang serius (Dong dan Peng, 2013). Pertama, *missing data* dapat menyebabkan potensi bias dalam estimasi parameter dan memperlemah generalisasi dari hasil penelitian. Kedua, mengabaikan *missing data* mengakibatkan hilangnya informasi yang berakibat berkurangnya *statistical power* dan meningkatnya *standard error*. Ketiga, hampir semua prosedur statistik dirancang untuk menangani data lengkap.

Metode untuk mengatasi *missing data* dapat dibagi ke dalam tiga kategori (Li dkk., 2004). Pertama dengan mengabaikan dan membuang observasi yang mengandung *missing data*. *Listwise deletion* dan *pairwise deletion* adalah dua metode yang dipakai secara luas dalam kategori ini. Kedua adalah estimasi parameter, yang menggunakan variasi dari algoritma *Expectation-Maximization* untuk melakukan estimasi parameter dimana terdapat *missing data*. Metode ini mengasumsikan bahwa distribusi bersama dari data adalah multivariat normal dan

mekanisme *missing data* adalah *Missing At Random*. (Piggot, 2001). Kategori ketiga adalah metode imputasi dengan mengisi atau mengganti nilai yang *missing* dengan nilai-nilai yang mungkin berdasarkan informasi yang tersedia pada data set tersebut.

Penelitian untuk mengatasi *missing data* telah lama dilakukan oleh banyak peneliti dari waktu ke waktu. Dempster, Laird dan Rubin (1976) menerapkan suatu pendekatan umum untuk perhitungan secara iterasi dari estimasi *Maximum Likelihood* ketika observasi diketahui berupa data tak lengkap. Efron (1992) memperkenalkan metode *bootstrap* dalam melakukan imputasi pada data hilang. King dkk. (2001) menawarkan sebuah algoritma alternatif untuk imputasi berganda dalam melakukan analisis data yang mengandung *missing value* di bidang politik.

Kelebihan melakukan imputasi untuk mengatasi data hilang adalah seluruh informasi yang ada pada data tetap terjaga. Setelah seluruh *missing value* diimputasi, dataset kemudian dapat dianalisis dengan menggunakan teknik standar untuk data yang lengkap. Diantara semua pendekatan imputasi, terdapat banyak pilihan yang bervariasi dari metode yang sederhana sampai beberapa metode yang lebih rumit berdasarkan analisis hubungan antara atribut. Metode imputasi utama dalam praktek meliputi (a) *Mean Imputation*; (b) *Regression Imputation*; (c) *Hot Deck Imputation*; dan (d) *Multiple Imputation* (Li dkk., 2004). Secara khusus dalam *Hot Deck Imputation*, algoritma pengklasteran telah digunakan secara luas. Salah satu yang paling terkenal adalah metode *K-Means*, yang mengambil sejumlah klaster yang diinginkan, *K*, sebagai parameter input, dan output berupa partisi *K* klaster dari sekumpulan objek. Metode *K-Means* juga disebut sebagai *C-Means* (Hathaway dan Bezdek, 2001).

Algoritma klaster konvensional memiliki batasan yang jelas. Akan tetapi pada beberapa kasus, suatu objek dapat masuk pada lebih dari satu klaster. Oleh karena itu, fungsi keanggotaan *fuzzy* (*fuzzy membership function*) dapat diterapkan pada klaster *C-Means* yang modelnya merupakan derajat keanggotaan dimana suatu objek lebih dekat kepada sebuah klaster. Salah satu kriteria yang digunakan dalam klaster adalah ukuran kemiripan yang dilakukan dengan menggunakan

fungsi jarak. Fungsi jarak yang biasa dikenal antara lain Euclidean, Manhattan, Minkowski dan sebagainya.

Masalah dalam metode imputasi berbasis *Fuzzy C-Means* adalah berkurangnya keefektifan dalam pemilihan nilai-nilai awal dan seringkali memberi hasil solusi yang optimum lokal. Hal ini karena nilai fungsi keanggotaannya tidak selalu terpetakan dengan baik. Dalam perkembangannya, sebuah metode sering dikombinasikan dengan teknik optimasi untuk memperoleh hasil yang lebih optimal. Teknik optimasi yang cukup populer adalah algoritma genetika yang memiliki daya tarik pada kesederhanaan dan kemampuan untuk mencari solusi yang baik dan cepat untuk masalah yang kompleks (Haupt dan Haupt, 2004). Algoritma genetika merupakan salah satu teknik optimasi yang mendasarkan konsepnya seperti pada proses genetika yang ada pada makhluk hidup dan dikenal luas sebagai perangkat teknik yang efektif untuk memetakan solusi yang optimal (Tang dkk., 2015).

Penelitian untuk mengatasi *missing data* pada survei industri besar dan sedang telah banyak dilakukan. Wardani (2011) membandingkan metode *Artificial Neural Network* (ANN) dan hibrida ANN-Algoritma Genetika (ANN-GA) untuk imputasi pada variabel nilai produksi. Metode hibrida ANN-GA menghasilkan nilai *Mean Square Error* (MSE) yang lebih kecil dan waktu *running* program yang lebih cepat daripada metode ANN. Metode ANN-GA juga lebih sederhana karena arsitektur terbaik langsung didapatkan. Sementara dengan metode ANN lebih rumit karena dilakukan dengan proses uji coba (*trial and error*).

Hartono (2011) membandingkan metode *Fuzzy K-Means Imputation* (FKMI) dengan metode *K-Nearest Neighbors Imputation* (KNNI) pada perusahaan industri besar di Jawa Timur tahun 2008 dengan variabel jumlah tenaga kerja, nilai bahan bakar, total nilai bahan baku, dan total nilai produksi. Kesimpulannya menunjukkan nilai rata-rata *Root Mean Square Error* (RMSE) dengan menggunakan metode FKMI memiliki nilai yang lebih besar dibandingkan dengan metode KNNI. Namun metode FKMI masih dapat dikembangkan lebih lanjut mengingat parameter FKMI yang lebih variatif jika dibandingkan dengan metode KNNI. Kesimpulan lainnya adalah bahwa fungsi jarak Euclidean secara

rata-rata menghasilkan nilai *error* yang lebih kecil jika dibandingkan dengan fungsi jarak Manhattan baik itu nilai RMSE, *Centroid Error* maupun *U Error*.

Penelitian Mawarsari (2012) dengan K-Nearest Neighbor dan Algoritma Genetika mengkaji metode KNNI dan metode hibrida KNNI-Algoritma Genetika (KKNI-GA). Hasil penelitiannya menyimpulkan bahwa imputasi *missing data* dengan metode hibrida KNNI-GA untuk pembobotan variabel memberikan hasil yang lebih baik daripada metode KNNI dan hibrida KNNI-GA untuk seleksi variabel karena menghasilkan RMSE yang lebih kecil. Metode hibrida KNNI-GA untuk pembobotan variabel dapat digunakan sebagai alternatif metode hibrida ANN-GA karena secara umum dapat menghasilkan nilai imputasi dengan RMSE yang kecil dan waktu *running* program yang lebih cepat.

Tang dkk. (2015) menerapkan pendekatan hibrida metode imputasi berbasis *Fuzzy C-Means* (FCM) dengan Algoritma Genetika untuk menangani data hilang pada volume lalu-lintas. Dengan memanfaatkan kesamaan mingguan diantara data, struktur data berbasis vektor ditransformasi ke dalam pola data berbasis matrik. Kemudian Algoritma Genetika diterapkan untuk mengoptimalkan fungsi keanggotaan dan *centroid* dalam model FCM. Dengan menggunakan RMSE, koefisien korelasi dan akurasi relatif (RA) sebagai indikator, hasil imputasi dibandingkan dengan beberapa metode konvensional. Hasilnya menunjukkan bahwa imputasi berbasis FCM-GA mengungguli metode konvensional dibawah kondisi lalu lintas yang berlaku.

Distribusi *missing data* pada data industri besar dan sedang di Sumatera Utara 2013 diasumsikan memiliki mekanisme *Missing At Random* (MAR). Untuk mekanisme ini dapat dilakukan pendefinisian untuk data-data yang hilang dengan metode *Hot Deck Imputation* (Hair dkk., 2010). Dalam metode *Hot Deck*, nilai yang digunakan untuk estimasi berasal dari observasi yang lain di dalam sampel yang dianggap serupa. Setiap observasi dengan *missing value* dipasangkan dengan observasi lain yang serupa pada variabel yang ditentukan. Kemudian *missing data* diganti dengan nilai yang valid yang berasal dari observasi yang serupa. Dalam pengklasteran *C-Means* dilakukan pengelompokkan data yang memiliki karakteristik yang sama ke dalam satu klaster yang sama. Dengan memanfaatkan

fuzzy dan algoritma genetika diharapkan dapat menghasilkan pengelompokkan dan imputasi yang lebih baik.

Berangkat dari hal tersebut, maka penelitian ini akan mengkaji metode imputasi *Fuzzy C-Means* dengan Algoritma Genetika, yang akan diterapkan untuk mengatasi *missing data* yang terjadi pada survei industri besar dan sedang di Sumatera Utara tahun 2013.

1.2 Perumusan Masalah

Pengumpulan data industri besar dan sedang hasil survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara selalu mengalami unit nonrespon dan item nonrespon. Item nonrespon tersebar pada sejumlah variabel tertentu dan terjadi setiap kali pelaksanaan survei. Nonrespon pada unit juga terjadi dengan masih banyak kuesioner yang tidak terisi atau dikembalikan oleh perusahaan. Nonrespon pada survei ini terjadi bukan karena ketidaktersediaan data, namun lebih disebabkan keengganan sebagian responden untuk mengikuti survei dan menjawab beberapa item pertanyaan. Akibat dari hal tersebut adalah munculnya *missing data* yang menyebabkan data menjadi tidak lengkap. Data yang tidak lengkap akan menyebabkan potensi bias dalam estimasi parameter dan memperlemah generalisasi penelitian karena ukuran data yang berkurang.

1.3 Tujuan Penelitian

Untuk mengatasi permasalahan di atas ditetapkan tujuan dari penelitian ini sebagai berikut :

1. Mengetahui perbandingan metode *Fuzzy C-Means* dan metode hibrida *Fuzzy C-Means* dan Algoritma Genetika dalam mengatasi data yang tidak lengkap pada data survei industri besar dan sedang di Sumatera Utara.
2. Mendapatkan metode imputasi dengan kinerja terbaik untuk mengatasi data yang tidak lengkap pada data hilang survei industri besar dan sedang di Sumatera Utara.

1.4 Manfaat Penelitian

Manfaat yang ingin dicapai dari penelitian ini adalah memberikan metode alternatif dalam mengatasi data tidak lengkap pada data hasil survei Badan Pusat Statistik. Secara khusus pada data hasil survei industri besar dan sedang di Sumatera Utara. Sehingga tersedia data yang lengkap dan dapat dianalisis untuk menghasilkan estimasi yang valid terkait dengan parameter populasi.

1.5 Batasan Masalah Penelitian

Penelitian ini dibatasi pada data Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara tahun 2013 pada kelompok usaha industri sedang. Pembatasan pada kelompok industri sedang mempertimbangkan jumlah industri yang masuk pada kategori ini adalah yang terbesar di Sumatera Utara dan *missing data* paling banyak terjadi pada industri yang masuk dalam kategori industri sedang. *Missing data* yang diteliti adalah *missing* pada item karena tidak lengkapnya isian pada beberapa variabel. Asumsi yang digunakan adalah mekanisme *missing data* mengikuti *Missing At Random* (MAR) dengan pola *General Pattern*.

BAB 2
TINJAUAN PUSTAKA

Bab ini memuat definisi dan tinjauan teoritis tentang m*issing data* dan tahapan-tahapan dalam imputasi *Fuzzy C-Means* dan Algoritma Genetika, serta deskripsi ringkas tentang Survei Tahunan Industri Manufaktur di Indonesia yang datanya digunakan dalam penelitian ini.

2.1 Missing Data

Tujuan dari analisis data adalah membuat estimasi yang valid terkait dengan parameter populasi. Adanya *missing data* mengancam keberhasilan tujuan ini jika data yang hilang atau tidak lengkap mengakibatkan sampel berbeda dengan populasinya (sampel yang bias). *Missing data* terjadi ketika nilai yang valid pada satu atau lebih variabel tidak tersedia untuk analisis. Sebagian besar *missing data* terjadi karena nonrespon pada survei baik itu disengaja (penolakan dari responden untuk disurvei atau tidak menjawab beberapa pertanyaan) atau tidak disengaja karena lupa. *Missing data* juga dapat terjadi karena kesalahan teknis seperti peralatan atau komputer yang tidak berfungsi dengan baik. Graham (2012) membaginya ke dalam dua jenis yaitu *item nonresponse* dan *wave nonresponse*. *Item nonresponse* terjadi jika responden tidak mengisi beberapa bagian dari kuesioner yang seharusnya terisi, sementara *wave nonresponse* terjadi jika responden tidak mengikuti sebagian atau seluruh periode pendataan.

Hair dkk. (2010) menyatakan bahwa sebelum melakukan perbaikan terhadap *missing data* perlu menentukan terlebih dulu tipenya apakah *ignorable* atau *not ignorable*, lalu sejauh mana data yang hilang cukup rendah untuk tidak mempengaruhi hasil. *Ignorable* jika *missing data* yang terjadi memang sudah diduga dan merupakan bagian dari desain penelitian. *Not ignorable* jika proses *missing data* disebabkan hal yang dapat diidentifikasi atau sukar diidentifikasi. Dapat diidentifikasi seperti kesalahan input data dan isian kuesioner yang tidak lengkap. Sukar diidentifikasi seperti penolakan responden menjawab pertanyaan tertentu atau responden tidak memahami pertanyaan. Kemudian jika *missing data*

tidak cukup rendah, maka perlu terlebih dahulu menentukan *randomness* (mekanisme) dari proses *missing data* sebelum melakukan perbaikan atau imputasi.

Untuk menangani *missing data* pada item memerlukan perhatian pada tiga aspek, yaitu : proporsi *missing data*, mekanisme *missing data*, dan pola *missing data* (Dong dan Peng, 2013). Proporsi *missing data* berkaitan langsung dengan kualitas inferensi statistik. Tetapi belum ada batasan yang tegas persentase *missing data* yang dapat diterima dalam sebuah kumpulan data untuk menghasilkan inferensi statistik yang valid. Schafer (1999) menegaskan bahwa tingkat *missing* kurang dari 5 persen tidak memberikan pengaruh. Hair (2010) mengatakan bahwa *missing data* dibawah 10 persen secara umum dapat diabaikan kecuali jika terjadi secara non random dengan pola tertentu.

2.1.1 Mekanisme *Missing Data*

Schafer dan Graham (2002) mengatakan dalam prosedur *missing data* modern, *missingness* dianggap sebagai fenomena probabilistik sebagai satu set variabel acak yang memiliki distribusi probabilitas gabungan. Dalam literatur statistik, distribusi ini kadang-kadang disebut mekanisme respon atau mekanisme *missingness*.

Menurut Little dan Rubin (2002) terdapat tiga mekanisme *missing data* :

1. *Missing Completely at Random* (MCAR) yang berarti bahwa terjadinya *missing data* tidak berkaitan dengan nilai semua variabel, apakah itu variabel dengan *missing values* atau dengan variabel pengamatan.
2. *Missing At Random* (MAR), berarti terjadinya *missing data* hanya berkaitan dengan variabel pengamatan.
3. *Missing Not At Random* (MNAR) bahwa terjadinya *missing data* pada suatu variabel berkaitan dengan variabel itu sendiri, sehingga tidak bisa diprediksi dari variabel lain pada suatu set data.

Misalkan untuk setiap set data, didefinisikan variabel indikator *R* yang mengidentifikasi apa yang diketahui dan apa yang hilang. *R* sebagai set variabel acak yang memiliki distribusi probabilitas gabungan. Karena *missingness* mungkin terkait dengan data, distribusi untuk *R* diklasifikasikan sesuai dengan

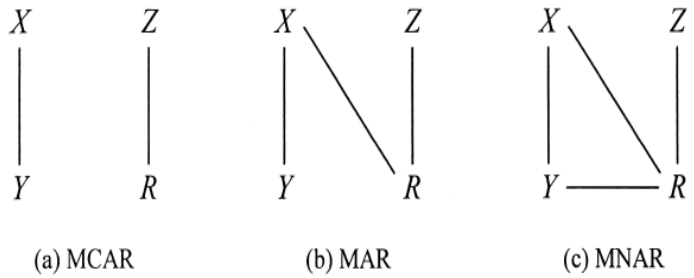
sifat hubungan itu. Misalkan Y_{com} menyatakan data lengkap dan mempartisinya sebagai $Y_{com} = (\mathrm{Y_{obs}}, \mathrm{Y_{mis}})$, dimana Y_{obs} dan Y_{mis} masing-masing adalah bagian yang diamati dan bagian hilang. Little dan Rubin (1976) mendefinisikan data yang hilang menjadi *MAR* jika distribusi *missingness* tidak tergantung pada Y_{mis},

$$P(R \mid Y_{com}) = P(R \mid Y_{obs}) \qquad (2.1)$$

Dengan kata lain, *MAR* memungkinkan probabilitas *missingness* bergantung pada data yang diamati tetapi tidak pada data yang hilang. Sebuah kasus khusus penting dari *MAR*, yang disebut *Missing Completely At Random (MCAR)*, terjadi ketika distribusi juga tidak tergantung pada Y_{obs},

$$P(R \mid Y_{com}) = P(R)$$

Ketika persamaan (2.1) dilanggar dan distribusi tergantung pada Y_{mis}, *missing data* dikatakan *Missing Not At Random (MNAR).*



Gambar 2.1 Representasi grafis dari (a) *missing completely at random* (MCAR), (b) *missing at random(MAR)*, dan (c) *missing not at random* (MNAR) dalam pola univariat. *X* menyatakan variabel yang lengkap, *Y* menyatakan sebuah variabel yang sebagian hilang, *Z* menyatakan komponen penyebab *missing*ness yang tidak terkait terhadap *X* dan *Y*, dan *R* menyatakan *missingness*. (Schafer dan Graham, 2002).

Definisi data hilang oleh Little dan Rubin (1976) menjelaskan hubungan statistik antara data dengan *missingness* bukan hubungan sebab-akibat. Penyebab *missingness* tidak ada dalam set data, tetapi beberapa penyebab tersebut mungkin terkait dengan X dan Y dan oleh karenanya dapat menyebabkan hubungan antara X atau Y dan R. Penyebab lain mungkin sama sekali tidak terkait dengan X dan Y dan dapat dianggap sebagai gangguan eksternal. Jika Z menunjukkan komponen penyebab yang tidak terkait dengan X dan Y, maka MCAR, MAR, dan MNAR

dapat direpresentasikan oleh hubungan grafis seperti pada Gambar 2.1. (a) MCAR mensyaratkan bahwa penyebab *missingness* sepenuhnya terkandung dalam bagian yang tidak terkait Z, (b) MAR mengizinkan beberapa kemungkinan terkait dengan X, dan (c) MNAR mensyaratkan beberapa penyebab terkait dengan Y setelah hubungan antara X dan R diperhitungkan.

Jika set data kecil maka secara visual dapat ditentukan mekanisme *missing data* dengan melihat pola *missing data*. Untuk data yang besar dibutuhkan uji diagnostik empiris untuk menentukan tingkat *randomness* dari *missing data*. Pendekatan pertama dilakukan terhadap proses *missing data* pada variabel tunggal dengan membentuk dua kelompok. Kelompok yang pertama adalah observasi dengan *missing data*, kelompok yang kedua adalah observasi tanpa *missing data*. Kemudian dilakukan uji *t* untuk melihat signifikansi statistik dari perbedaan dalam rata-rata variabel antara dua kelompok. Jika kedua kelompok berbeda maka mengindikasikan kemungkinan proses *missing data* tidak random.

Pendekatan yang kedua adalah uji keseluruhan untuk menentukan apakah *missing data* dapat dikategorikan sebagai MCAR, salah satunya dengan Uji *Little*. Uji ini menganalisis pola *missing data* pada semua variabel dan membandingkannya dengan suatu pola *missing data* yang random. Jika tidak ditemukan perbedaan yang signifikan, *missing data* dapat dikategorikan sebagai MCAR (Hair dkk., 2010).

2.1.2 Pola *Missing Data*

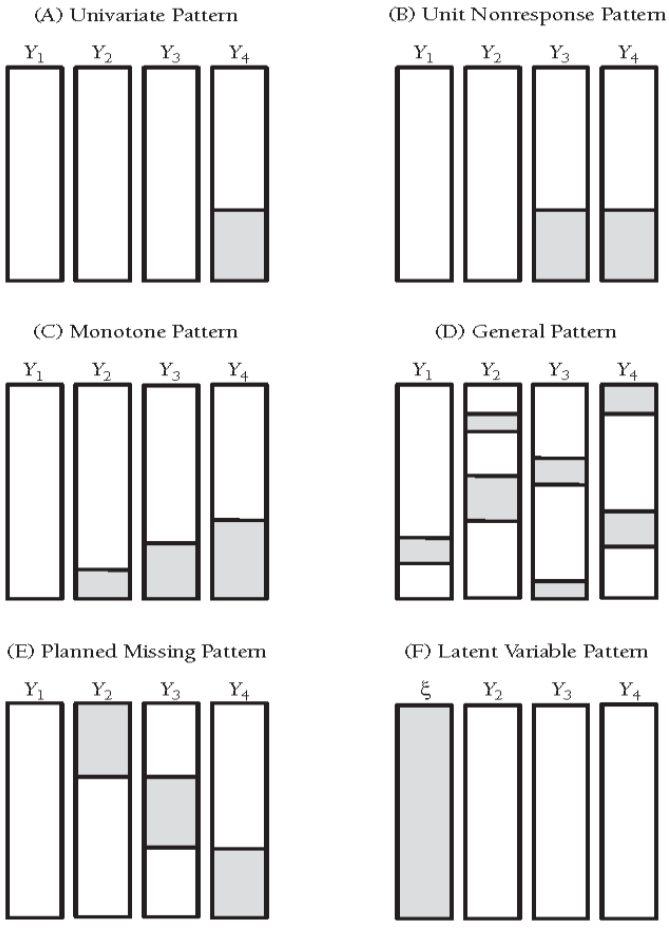
Menurut Enders (2010), ada beberapa pola *missing data* atau data hilang. Pola ini hanya menggambarkan lokasi dari *missing values* dan bukan sebab dari terjadinya *missing values*.

Pola univariat (A) adalah pola data hilang yang data hilangnya terdapat pada satu variabel. Pola ini relatif jarang namun dapat muncul dalam studi yang bersifat percobaan. Pola nonrespon unit (B) umum terjadi dalam survei penelitian, terdapat variabel yang lengkap dan beberapa variabel lain terdapat data hilang seperti responden menolak untuk menjawab. Sebuah pola data hilang dikatakan monoton (C), jika data yang hilang pada pengukuran tertentu selalu hilang pada

pengukuran berikutnya. Pola ini biasanya terkait dengan studi longitudinal dimana responden atau partisipan keluar dan tidak kembali lagi.

Pola general (D) adalah bentuk yang paling umum dari missing values yaitu memiliki pola yang menyebar pada data secara acak. Pola data hilang terencana (E) yaitu pola data yang secara sengaja direncanakan hilang yang digunakan untuk mengumpulkan sejumlah besar item kuesioner sekaligus mengurangi beban responden. Yang terakhir adalah pola variabel laten (F), yaitu pola unik untuk analisis variabel laten seperti model persamaan struktural.

Gambar 2.2 memperlihatkan enam *prototipe* pola *missing data* yang dapat terjadi dimana bagian yang berwarna abu-abu menyatakan lokasi *missing values* dalam set data.



Gambar 2.2 Enam pola *missing data*. Bagian yang berwarna abu-abu menyatakan lokasi *missing values* dalam set data dengan empat variabel (Enders, 2010).

2.2 Fuzzy C-Means (FCM) Clustering

Metode pengelompokkan dibedakan menjadi dua, yaitu metode hirarki dan metode non hirarki. *C-Means* (CM) adalah salah satu metode non hirarki yang bertujuan mempartisi data ke dalam satu atau lebih klaster/kelompok berdasarkan kemiripan objek, dan meminimalkan ketidakmiripan di dalam klaster. Algoritma dasar untuk CM (Tan dkk., 2005) dimulai dengan menentukan *c* inisial *centroid*, dimana *c* adalah parameter jumlah klaster awal yang diinginkan. Sebuah *centroid* klaster menyatakan nilai rata-rata dari objek di dalam klaster. Setiap data kemudian ditentukan pada *centroid* terdekat dan setiap kumpulan data yang ditentukan pada sebuah *centroid* adalah sebuah klaster. *Centroid* dari setiap klaster kemudian dihitung kembali berdasarkan kumpulan data yang ditentukan terhadap klaster tersebut. Proses tersebut diulang sampai tidak ada data yang berpindah klaster atau *centroid-centroid* tidak berubah.

Dalam pengklasteran CM, ketidakmiripan di dalam klaster diukur oleh penjumlahan jarak antara objek-objek dan *centroid* dari klaster dimana objek-objek tersebut ditentukan. Untuk menghitung jarak antara data dan *centroid*, *distance space* yang sering digunakan adalah Euclidean dan Manhattan. Euclidean sering digunakan karena penghitungan jarak dalam *distance space* ini merupakan jarak terpendek yang bisa didapatkan antara dua titik yang dihitung. Keunggulan dari Manhattan adalah kemampuannya dalam mendeteksi keadaan khusus seperti keberadaan *outliers* dengan lebih baik.

Setelah c*entroid* dari setiap klaster diperbaharui, data kemudian dialokasikan kembali ke dalam masing-masing klaster. Proses iterasi ini dapat dilakukan dengan dua cara. Pertama adalah pengalokasian secara tegas (*hard*) yaitu data dialokasikan ulang secara tegas ke klaster yang mempunyai *centroid* terdekat dengan data tersebut dan tidak menjadi anggota klaster lainnya. Kedua adalah dengan cara halus/samar (*fuzzy*) dimana masing-masing data diberikan nilai kemungkinan untuk bisa bergabung ke setiap klaster yang ada.

Dalam metode *Fuzzy C-Means* (FCM) digunakan variabel fungsi keanggotaan (*membership function*) u_{ik}, yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota suatu klaster. FCM mengalokasikan kembali data ke dalam masing-masing klaster dengan mengubah

domain u_{ik}, yaitu dari u_{ik} yang bernilai 0 atau 1 ($u_{ik} \in \{0,1\}$) menjadi u_{ik} yang bernilai 0 sampai dengan 1 ($u_{ik} \in [0,1]$). Data yang mempunyai tingkat kemungkinan yang lebih tinggi ke suatu kelompok akan mempunyai nilai fungsi keanggotaan u_{ik} ke kelompok tersebut yang mendekati angka 1 dan ke kelompok yang lain mendekati angka 0. Pal dan Bezdek (1995) memperkenalkan suatu variabel *m* yaitu bobot eksponen (*weighting exponent*) dari fungsi keanggotaan. *Weighting exponent* memiliki peran dalam model FCM karena mempengaruhi kualitas kesimpulan tentang validitas FCM (*U,V*) yang dihasilkan oleh algoritma yang mencoba untuk mengoptimalkan J_m, fungsi objektif *Fuzzy C-Means*. Hasil perhitungan mereka menyarankan pilihan nilai *m* yang disarankan pada interval 1,5< *m* < 2,5. Nilai *m* yang umum digunakan adalah 2.

Tujuan dari pengelompokkan data adalah untuk meminimalisasi fungsi objektif. Fungsi objektif yang digunakan ditentukan berdasarkan pendekatan yang digunakan dalam (i) *distance space* untuk menghitung jarak antara data dan *centroid* (ii) metode pengalokasian ulang data ke dalam masing-masing klaster.

$$\min_{(U,v)} \left\{ J_m(U,v) = \sum_{i=1}^{n} \sum_{k=1}^{c} U_{ik}^{m} \left\| x_i - v_k \right\|_A^2 \right\} \qquad (2.2)$$

dimana $m > 1$ adalah *weighting exponent*, $\|x\|_A^2 = x^T A x$ adalah vektor norm-*A*, *n* = banyak data, *c* = banyak klaster, v_k = nilai *centroid* klaster ke-*k*, u_{ik} = derajat keanggotaan data ke-*i* klaster ke-*k*, $0 \le u_{ik} \le 1$.

Kegagalan untuk mencapai konvergen dapat terjadi karena perpindahan suatu data ke suatu klaster tertentu dapat mengubah karakteristik model klaster, yang berakibat data yang telah dipindahkan lebih sesuai berada di klaster semula sebelum dipindahkan. Kejadian seperti ini mengakibatkan proses tidak akan berhenti dan kegagalan untuk konvergen terjadi. Pada FCM kemungkinan hal ini untuk terjadi sangat kecil karena setiap data dilengkapi dengan fungsi keanggotaan u_{ik} untuk mejadi anggota klaster yang ditemukan. Setiap objek data x_i memiliki sebuah fungsi keanggotaan yang menggambarkan derajat objek data

ini menjadi milik klaster tertentu v_k. Algoritma FCM (Xie dan Beni, 1991) adalah sebagai berikut :

(1) Inisialisasi keanggotaan u_{ik} dari data x_i pada klaster ke-*k*, dimana $\sum_{k=1}^{c} u_{ik} = 1$

(2) Menghitung *centroid* masing-masing klaster :

$$v_k = \frac{\sum_{i=1}^{n} u_{ik}{}^{m} x_i}{\sum_{i=1}^{n} u_{ik}{}^{m}}, \text{untuk } 1 \le k \le c$$

(3) Perbaharui nilai fungsi keanggotaan masing-masing data ke masing-masing klaster :

$$u_{ik} = \frac{d\left(x_i, v_k\right)^{-2/(m-1)}}{\sum_{j=1}^{c} d\left(x_i, v_j\right)^{-2/(m-1)}}$$

untuk $1 \le i \le n$, $1 \le k \le c$,dimana $\sum_{j=1}^{c} u_{ij} = 1$

(4) Ulangi langkah (2) dan langkah (3) sampai nilai J_m pada persamaan (2.2) tidak lagi menurun.

Algoritma FCM selalu konvergen pada minimum lokal yang ketat dari J_m yang dimulai dari sebuah perkiraan awal dari u_{ik}. Tetapi pilihan yang berbeda terhadap nilai inisial u_{ik} bisa mengarah kepada minimum lokal yang berbeda.

Salah satu kriteria yang digunakan dalam klaster adalah ukuran kemiripan yang diukur dengan menggunakan fungsi jarak. Beberapa fungsi jarak yang dikenal adalah Euclidean, *Squared* Euclidean, Manhattan, dan Minkowski. Jarak antara *centroid* v_k dan objek x_{ij} dalam *Fuzzy C-Means* dinotasikan dengan $d\left(x_i, v_k\right)$ dengan *s* adalah jumlah atribut. *Generalized* L_p norm digunakan untuk mengukur jarak antara sebuah *centroid* dan objek data dalam klaster.

$$d\left(x_i, v_k\right) = \left(\sum_{j=1}^{s} \left| x_{ij} - v_k \right|^{p} \right)^{1/p}$$

Jarak Euclidean adalah jarak dengan $p = 2$ $\left(L_2\right)$ dan jarak Manhattan adalah jarak dengan $p = 1$ $\left(L_1\right)$.

2.3 Algoritma Genetika

Algoritma genetika merupakan algoritma pencarian berdasarkan mekanisme seleksi alam yang dikembangkan oleh John Holland untuk meneliti proses adaptasi dari sistem alam serta mendesain perangkat lunak yang memiliki kecerdasan buatan. Algoritma ini meniru proses genetika pada makhluk hidup dimana perkembangan generasi dalam suatu populasi mengikuti prinsip seleksi alam yaitu yang kuat yang akan bertahan.

Algoritma genetika bekerja dengan mempertahankan sebuah populasi yang terdiri dari individu-individu, dimana tiap individu merepresentasikan sebuah solusi pada permasalahan yang dihadapi. Individu dikodekan dalam bentuk kromosom yang terdiri dari komponen genetik terkecil yaitu gen. Dari individu yang ada, dihitung nilai *fitness* yang digunakan sebagai kriteria solusi terbaik. Individu-individu yang lolos proses seleksi kemudian melakukan reproduksi dengan perkawinan silang sehingga menghasilkan keturunan. Keturunan-keturunan ini kemudian dievaluasi untuk membuat populasi baru yang memiliki kriteria yang lebih baik. Setelah beberapa generasi terbentuk, algoritma akan konvergen pada individu terbaik yang diharapkan merepresentasikan solusi optimal dari permasalahan yang dihadapi (Gen dan Cheng, 2000).

Individu dalam algoritma genetika direpresentasikan oleh kromosom yang menggambarkan sebuah solusi dari masalah yang akan diselesaikan. Kromosom terdiri dari sekumpulan gen. Representasi kromosom dapat berupa bilangan biner, integer, ataupun bilangan riil. Algoritma genetika diawali dengan menginisialisasi himpunan solusi yang dibangkitkan secara acak. Himpunan solusi ini disebut populasi. Ukuran populasi bergantung pada kerumitan masalah yang dihadapi. Semakin besar populasi akan memiliki konsekuensi terhadap

kinerja akurasi dan waktu yang dibutuhkan untuk konvergen (Engelbrecht, 2002). Kromosom-kromosom dapat berubah terus menerus yang disebut dengan proses regenerasi. Pada setiap generasi, kromosom dievaluasi dengan mengunakan alat ukur yang disebut fungsi *fitness*. Untuk membuat generasi berikutnya, kromosom-kromosom baru yang disebut *offspring* (keturunan) terbentuk dengan cara menggabung dua kromosom dari generasi sekarang dengan menggunakan operator *crossover*/persilangan atau mengubah sebuah kromosom dengan menggunakan operator mutasi. Generasi baru dibentuk dengan cara seleksi yang dilakukan terhadap *parent* dan *offspring* berdasarkan nilai *fitness*nya dan menghilangkan yang lainnya. Kromosom-kromosom yang lebih sesuai memiliki probabilitas untuk dipilih. Setelah beberapa generasi, algoritma ini akan konvergen ke arah bentuk kromosom yang terbaik, dengan harapan dapat menyatakan solusi optimal dari permasalahan yang diselesaikan.

2.3.1 Fitness

Setiap individu dievaluasi oleh sebuah fungsi *fitness* dalam algoritma genetika. *Fitness* mengukur kualitas dari kromosom (individu), yaitu seberapa dekat kepada solusi optimal. Tujuannya adalah mencari individu dengan nilai *fitness* yang paling tinggi (Haupt dan Haupt, 2004). Fungsi *fitness* yang digunakan adalah :

$$f = \frac{1}{\left(RMSE + h\right)}$$

dimana h adalah nilai yang kecil ($0 \leq h \leq 1$), dan RMSE adalah nilai *Root Mean Square Error*.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

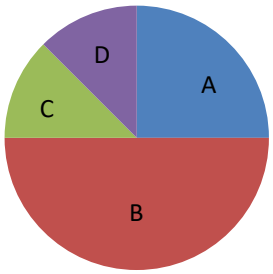
dengan n adalah jumlah observasi, y_i adalah output yang diinginkan pada observasi ke-i dan $\hat{y}_i$ adalah nilai output dugaan pada observasi ke-i.

2.3.2 Seleksi

Seleksi merupakan proses pemilihan individu-individu untuk dilakukan rekombinasi/perkawinan silang dan mutasi yang bertujuan untuk memperoleh calon induk yang baik. Langkah pertama yang dilakukan dalam seleksi adalah pencarian nilai *fitness*. Semakin tinggi nilai *fitness* suatu individu kemungkinan terpilih juga semakin besar. Ada beberapa metode seleksi :

a. Seleksi Roda Roulette

Metode ini seperti permainan roda roulette dimana masing-masing kromosom menempati bagian lingkaran pada roda roulette secara proporsional sesuai dengan nilai *fitness*nya. Kromosom dengan nilai *fitness* yang lebih tinggi menempati bagian potongan lingkaran yang lebih besar dibandingkan dengan kromosom bernilai *fitness* rendah sehingga mempunyai kemungkinan lebih besar untuk terpilih seperti dapat dilihat pada Gambar 2.3 berikut :



Gambar 2.3 Ilustrasi penggunaan metode roda roulette

Karena terpilihnya suatu kromosom dalam populasi untuk dapat berkembang biak adalah sebanding dengan nilai *fitness*nya maka cenderung kromosom yang baik akan terpelihara terus sehingga dapat membawa ke hasil optimum lokal (konvergensi dini). Jika semua kromosom dalam populasi memiliki *fitness* yang hampir sama, maka seleksi akan bersifat acak.

b. Seleksi Ranking

Kelemahan dari seleksi roda roulette adalah jika terdapat kromosom dengan nilai *fitness* yang perbedaannya jauh, maka akan ada kromosom yang mempunyai kesempatan terpilih sangat kecil. Dengan seleksi ranking, maka setiap kromosom akan diberi nilai *fitness* baru dengan mengurutkan kromosom dari nilai *fitness* terkecil sampai terbesar. Kromosom dengan nilai

fitness terkecil akan mendapat nilai *fitness* 1, yang kedua mendapat nilai *fitness* 2 dan seterusnya sampai yang terbaik akan mendapat nilai *fitness* N, yaitu jumlah kromosom dalam populasi.

c. Seleksi Turnamen

Seleksi ini merupakan variasi dari seleksi roulette dan seleksi ranking. Sejumlah k kromosom tertentu dari populasi dengan n kromosom ($k \leq n$) dipilih secara acak dengan probabilitas yang sama. Dari k kromosom yang diperoleh kemudian dipilih kromosom dengan *fitness* terbaik yang didapatkan dengan mengurutkan ranking *fitness* kromosom yang dipilih tersebut. Pemilihan kromosom yang akan digunakan untuk berkembang biak tidak berdasarkan *fitness* dari populasi. Untuk $k = 1$, seleksi turnamen akan sama dengan seleksi secara acak karena hanya akan melibatkan satu kromosom. Untuk $k = 2$, maka dua kromosom akan dipilih secara acak dari populasi, lalu dari kedua kromosom tersebut akan dipilih salah satu dengan kriteria *fitness* terbaik. Banyaknya k kromosom yang dipilih dari populasi disebut *tournament size*. *Tournament size* yang sering digunakan adalah *binary tournament* yaitu *tournament size* dengan $k = 2$.

Memilih metode seleksi yang akan digunakan memperhatikan nilai *fitness* dari individu. Seleksi dengan Roda Roulette merupakan metode yang paling sederhana namun paling besar variansinya sehingga munculnya individu superior sering terjadi pada model ini. Salah satu cara untuk mengurangi kemungkinan terjadinya individu superior yang dapat menyebabkan optimum lokal adalah menggunakan rank yaitu metode seleksi dengan Turnamen.

Untuk menentukan individu mana yang akan tetap bertahan selama proses seleksi ada dua model populasi yang dikenal :

1. *Generation Model,* dimana setiap generasi dimulai dengan sebuah populasi berukuran *n* yang dari populasi tersebut dipilih kelompok pasangan *n parents.* Kemudian *m offspring* dengan jumlah yang sama *(m = n)* dihasilkan. Setelah setiap siklus, seluruh populasi digantikan oleh *offspring,* membentuk generasi berikutnya.

2. *Steady-state Model,* berbeda dengan sebelumnya, seluruh generasi tidak diganti pada saat yang bersamaan, tapi sejumlah *m* individu lama (*m* < *n*) digantikan oleh *m* individu baru yang dinamakan *offspring*

2.3.3 Persilangan (*Crossover*)

Pada persilangan (*crossover*) dilakukan operasi pertukaran gen-gen yang bersesuaian dari dua induk untuk membentuk individu baru. *Crossover* terjadi pada probabilitas tertentu yaitu $P_c \in [0,1]$. *Crossover* dilakukan dengan mengambil bilangan random $\xi \sim Uniform(0,1)$, jika $\xi \le P_c$ maka terjadi *crossover,* dan sebaliknya. Artinya penyilangan bisa dilakukan hanya jika suatu bilangan random $[0,1]$ yang dibangkitkan tidak melebihi dari P_c yang ditentukan. Beberapa cara untuk melakukan *Crossover* yaitu :

a. Persilangan (*Crossover*) Satu Titik

   Proses persilangan satu titik adalah memilih satu titik persilangan. Posisi *crossover p* (*p* = 1,2,…,*q*-1), dengan *q* = panjang kromosom dipilih secara random. Kemudian gen-gen antar kromosom induk ditukar pada titik tersebut.

b. Persilangan Dua Titik atau lebih

   Pada *crossover* lebih dari satu titik penentuan *m* posisi *crossover* yaitu p_i (*p* = 1,2,…,*q*-1; i = 1,2,…,*m*) dengan *q* = panjang kromosom dipilih secara acak dan tidak boleh ada posisi yang sama serta diurutkan naik. Lalu gen-gen antar kromosom induk ditukar pada titik-titik tersebut.

c. Persilangan *Uniform*

   Persilangan ini memperlakukan setiap gen secara individu, kemudian memilih secara random dari *parent* yang mana gen tersebut akan diturunkan. Langkah pertama adalah membangkitkan sejumlah *l* bilangan random berurutan (*l* adalah panjang kromosom) dari distribusi *Uniform* [0,1]. Berikutnya. untuk *offspring* pertama untuk setiap posisi urutan gen, nilainya dibandingkan dengan parameter *p* (biasanya 0.5). Jika nilainya dibawah *p* gen diambil dari *parent* pertama, dan jika sebaliknya dari *parent* kedua. *Offspring* kedua secara natural akan dibentuk dari proses kebalikannya.

d. Persilangan Aritmatika

Crossover aritmatika digunakan untuk representasi kromosom berupa bilangan pecahan (*float*). *Crossover* ini dilakukan dengan menentukan bilangan *r* sebagai bilangan random bernilai 0 sampai 1. Selain itu juga ditentukan posisi *crossover p* (*p* = 1,2,…,*q*-1), dengan *q* = panjang kromosom yang dilakukan *crossover* secara random. Nilai gen baru pada keturunan (*offspring*) mengikuti rumus sebagai berikut :

$$x_1^{'}(p) = rx_1(p) + (1-r)x_2(p)$$

$$x_2^{'}(p) = rx_2(p) + (1-r)x_1(p)$$

2.3.4 Mutasi

Setelah proses seleksi selesai kemudian dilakukan proses mutasi pada keturunannya. Mutasi adalah proses mengubah nilai dari satu atau beberapa gen yang ada dalam satu kromosom. Jika dalam proses pemilihan kromosom cenderung terus pada kromosom yang baik saja maka konvergensi dini dapat dengan mudah terjadi, yaitu mencapai solusi optimum lokal. Untuk menghindari konvergensi dini dan tetap menjaga perbedaan-perbedaan kromosom dalam populasi operasi mutasi dapat dilakukan. Proses mutasi bersifat acak sehingga tidak ada jaminan bahwa akan diperoleh kromosom dengan *fitness* yang lebih baik, namun dengan mutasi diharapkan dapat menghasilkan kromosom dengan nilai *fitness* yang lebih baik dibandingkan sebelum operasi mutasi dilakukan.

Mutasi juga terjadi pada probabilitas tertentu, yaitu $P_m \in [0,1]$ dengan nilai yang digunakan biasanya kecil untuk memastikan bahwa solusi yang baik tidak terdistorsi terlalu banyak (Engelbrecht, 2002). Mutasi pada individu dilakukan dengan mengambil bilangan random $\xi \sim Uniform(0,1)$, jika $\xi \le P_m$ maka terjadi mutasi. Posisi gen *p* (*p* = 1,2,…,*q* - 1), dengan *q* = panjang kromosom yang akan dilakukan mutasi dipilih secara random.

2.3.5 Replacement dan Elitism

Replacement merupakan tahap akhir dari satu siklus algoritma genetika. Ada dua teknik yang dapat digunakan, yaitu mengganti semua populasi dengan keturunan yang dihasilkan dari proses *crossover* dan mutasi sebagai generasi baru

atau dengan masih mempertahankan beberapa populasi awal yang memiliki nilai *fitness* tertinggi (*elitism*) dan menambahkan beberapa keturunan sebagai generasi yang baru.

2.4 Integrasi Imputasi *Fuzzy C-Means* dengan Algoritma Genetika

2.4.1 Imputasi *Fuzzy C-Means*

Masalah mendasar dalam imputasi *missing data* adalah untuk mengisi informasi yang hilang tentang sebuah objek berdasarkan pengetahuan dari informasi lain tentang objek tersebut. Misalkan diberikan sekumpulan objek, tujuan dari pengklasteran adalah untuk membagi kumpulan data ke dalam kelompok-kelompok berdasarkan kemiripan objek, dan untuk meminimalkan ketidakmiripan dalam klaster. Dalam *C-Means* ketidakmiripan dalam klaster ini diukur dengan penjumlahan jarak antara dua objek dan *centroid* klaster dari objek tersebut. Sebuah *centroid* klaster menyatakan nilai rata-rata dari objek-objek dalam klaster. Metode FCM menyempurnakan metode C-Means orisinil sebagai alat pengklasteran statistik yang lebih baik ketika klaster-klaster tidak terpisah secara baik. Metode ini mengatasi keterbatasan dengan masalah optimum lokal pada metode pengklasteran orisinil (Li dkk., 2004).

Misalkan sekumpulan dari *n* objek $X = \{x_1, x_2, ..., x_n\}$ dimana setiap objek memiliki *s* atribut ($1 \le i \le n$ dan $1 \le j \le s$), x_{ij} menyatakan nilai atribut *j* dalam objek x_i. Objek x_i disebut sebuah objek yang lengkap, jika $\{x_{ij} \ne \varnothing | \forall 1 \le j \le s\}$ dan sebuah objek tidak lengkap, jika $\{x_{ij} = \varnothing | \exists 1 \le j \le s\}$, dan dikatakan bahwa objek x_i memiliki *missing value* pada atribut *j*. Untuk setiap objek x_i tidak lengkap, digunakan $R = \{ j | x_{ij} \ne \varnothing, 1 \le j \le s\}$ untuk menyatakan kumpulan atribut yang nilainya tersedia, dan atribut ini disebut atribut *reference.* Tujuannya adalah untuk mendapatkan nilai dari atribut *non-reference* untuk objek-objek yang tidak lengkap.

Tang dkk. (2015) menuliskan enam langkah metode imputasi berbasis FCM konvensional :

Langkah-1 Tetapkan nilai-nilai untuk parameter, *c* dan *m*, inisialisasi nilai keanggotaan fungsi *U*, dan hitung *centroid* klaster $V = \{v_1, v_2, ..., v_c\}$ dengan persamaan berikut :

$$v_k = \frac{\sum_{i=1}^{n} U(x_i, v_k)^m . x_i}{\sum_{i=1}^{n} U(x_i, v_k)^m} \qquad (2.3)$$

Langkah-2 Definisikan jarak *generalized norm* L_p, antara *centroid* v_k dan data spesifik x_i.

$$d(x_i, v_k) = \left(\sum_{j=1}^{s} \left| x_{ij} - v_k \right|^p \right)^{1/p}$$

Ketika *p* = 2, L_2 adalah jarak Euclidean. Ketika *p* = 1, L_1 menyatakan jarak Manhattan.

Langkah-3 Definisikan fungsi objektif seperti berikut :

$$J(U,V) = \sum_{i=1}^{n} \sum_{k=1}^{c} U(x_i, v_k)^m . d(x_i, v_k)$$

Langkah-4 Putuskan apakah batasan kondisi dipenuhi, jika nilai dari fungsi objektif lebih besar dari ambang batas (*threshold*) yang ditentukan sebelumnya, kemudian perbaharui nilai *U* dalam persamaan (2.4) hitung kembali nilai *V* dalam persamaan (2.3), dan kembali ke Langkah-2. Jika tidak, lanjutkan ke tahap berikutnya.

$$U(x_i, v_k) = \frac{d(x_i, v_k)^{-2/(m-1)}}{\sum_{j=1}^{c} d(x_i, v_j)^{-2/(m-1)}} \qquad (2.4)$$

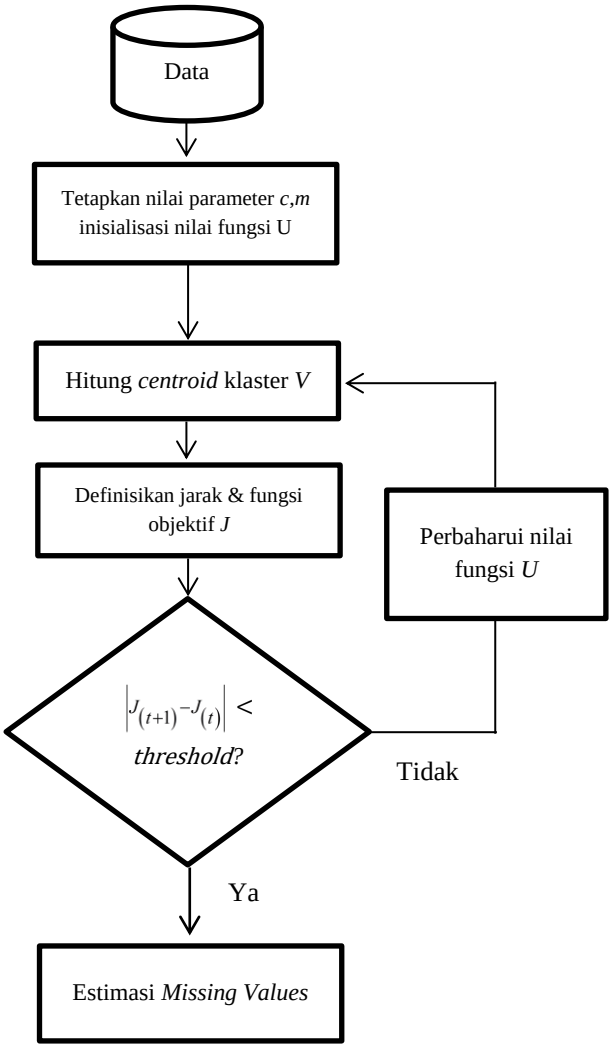
dimana fungsi keanggotaan, *U*, menyatakan derajat bahwa x_i adalah milik v_k. Parameter *m*, adalah faktor pembobot pengukur derajat ketidakjelasan (*fuzziness*) dalam proses pengklasteran. *m* > 1 dan $\sum_{j=1}^{c} U(x_i, v_j) = 1$ untuk setiap objek data x_i $(1 \le i \le n)$.

Langkah-5 Perbaharui parameter-parameter baru, *c* dan *m*, dan ulangi langkah-1 sampai langkah-4.

Langkah-6 Dapatkan fungsi objektif minimal dan nilai optimal untuk *U* dan *C*, dan estimasi nilai atribut yang hilang untuk elemen x_i yang tidak lengkap.

$$\hat{x}_{ij} = \sum_{k=1}^{c} U\left(x_i, v_k\right).v_{k,j} \qquad (2.5)$$

dimana $\hat{x}_{ij}$ mengindikasikan *missing data* yang didefinisikan sebagai atribut *non-reference*.



Gambar 2.4 Diagram alir imputasi berbasis *Fuzzy C-Means*

Hathaway dan Bezdek (2001) menyebutkan bahwa cukup beralasan untuk memodifikasi FCM berdasarkan jarak pada persamaan (2.6) dan persamaan (2.7) untuk mengganti optimisasi alternatif FCM dengan menambahkan faktor skala dalam perhitungan jarak.

$$d\left(x_i, v_k\right) = \frac{s}{I_i}\left(\sum_{j=1}^{s}\left|x_{ij} - v_k\right|^p\right)^{1/p} I_{ij} \qquad (2.6)$$

dimana,

$$I_{ij} = \begin{cases} 0, & \text{jika } x_{ij} \in X_M \\ 1, & \text{jika } x_{ij} \in X_P \end{cases} \qquad 1 \le j \le s \text{ dan } 1 \le i \le n$$

$$X_M = \left\{ x_{ij} \text{ untuk } 1 \le j \le s, 1 \le i \le n \,\middle|\, \text{nilai } x_{ij} \text{ ada dalam } X \right\}$$

$$X_P = \left\{ x_{ij} =? \text{ untuk } 1 \le j \le s, 1 \le i \le n \,\middle|\, \text{nilai } x_{ij} \text{ hilang dari } X \right\}$$

dan

$$I_i = \sum_{j=1}^{s} I_{ij} \qquad (2.7)$$

Faktor skala akan memiliki efek dalam memberikan pembobot tambahan kepada komponen dari *missing data,* dibandingkan dengan komponen dari $x_i \in X_P$ dalam perhitungan rata-rata klaster.

2.4.2 Imputasi Hibrida FCM-GA

Efektifitas dari metode FCM konvensional berkurang disebabkan oleh seleksi nilai awal dan sering menghasilkan solusi yang optimal lokal. Secara parsial dapat mengoptimalkan fungsi keanggotaan *U* dan *centroid V* pada nilai awal pendahuluan terpilih dari parameter, *c* dan *m*. Untuk mengatasi keterbatasan ini, optimisasi metode berbasis FCM dapat ditingkatkan dengan menggunakan teknik Algoritma Genetika/*Genetic Algorithm* (GA). Sebagai sebuah alat yang efektif untuk mendapatkan lokasi solusi optimal, GA cukup populer dalam banyak wilayah penelitian. GA menyertakan operasi reproduksi, *crossover,* dan mutasi. Secara khusus, prosedur perhitungan GA untuk mencari solusi optimal berbasis FCM sebagai berikut (Tang dkk., 2015) :

Langkah-1 Atur jumlah klaster, *c*, dan faktor pembobot, *m*, inisialisasi fungsi keanggotaan *U*, dan hitung nilai awal *centroid* klaster *V* menggunakan persamaan **(2.3)**. Tentukan ukuran populasi *n*, jumlah generasi evolusi T, probabilitas *crossover* P_c, probabilitas mutasi P_m;

Langkah-2 Estimasi *missing data* dalam persamaan **(2.5)**, kemudian definisikan RMSE antara imputasi dan nilai aktual seperti berikut :

$$Errors(U,V) = \sqrt{\frac{1}{n.s}\sum_{i=1}^{n}\sum_{j=1}^{s}\left(x_{ij} - \hat{x}_{ij}\right)^2} \qquad (2.8)$$

Langkah-3 Definisikan ukuran *goodness of fit* seperti berikut :

$$f(U,V) = \frac{1}{Errors(U,V) + \xi} \qquad (2.9)$$

dimana ξ adalah sebuah konstanta.

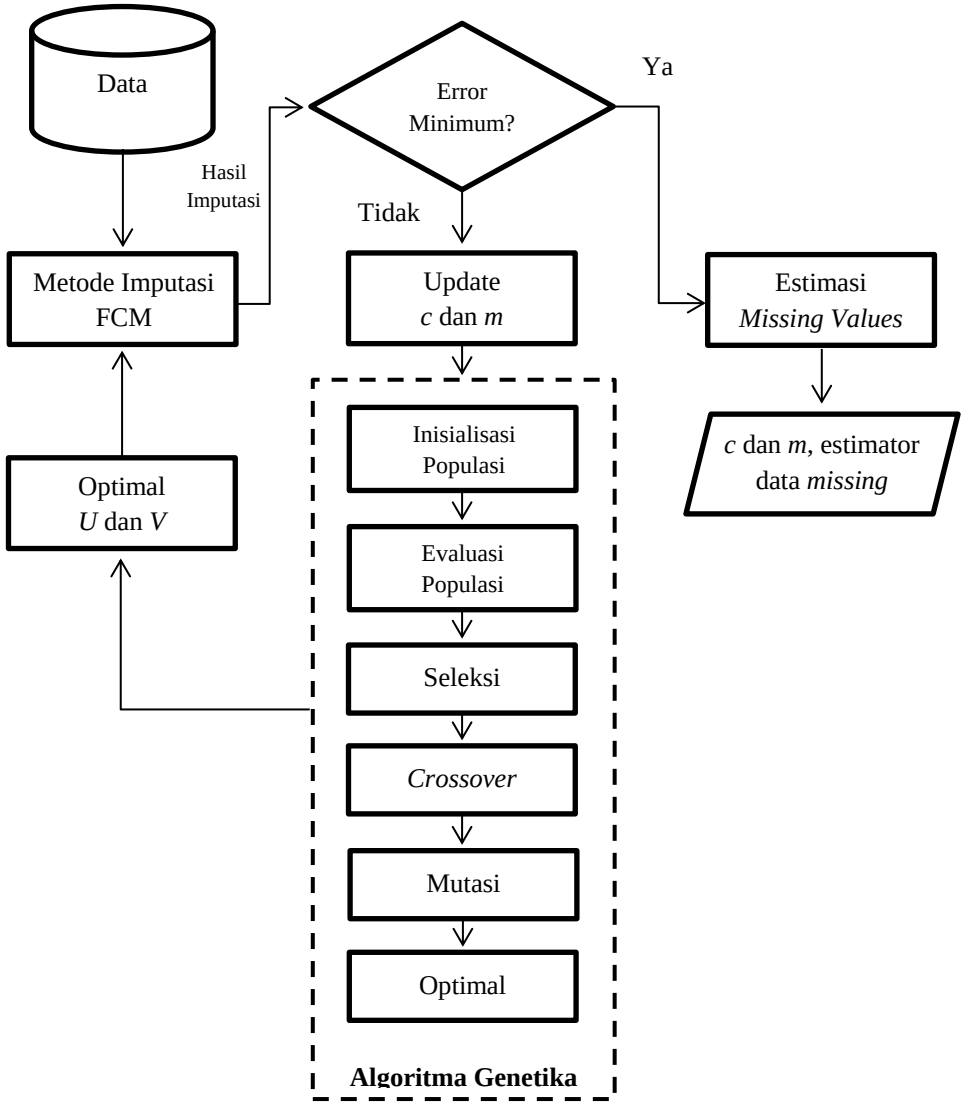
Langkah-4 Hasilkan populasi awal;

Langkah-5 Gunakan ukuran *goodness of fit* pada persamaan (2.9) untuk meng evaluasi populasi dan selesaikan operasi seleksi, *crossover* dan mutasi;

Langkah-6 Identifikasi solusi-solusi optimal dan evaluasi kondisi konvergen evolusi. Jika hal ini terpenuhi lanjutkan ke Langkah-7, jika sebaliknya, kembali ke Langkah-5. Kondisi konvergen diukur oleh jumlah generasi evolusi dalam penelitian.

Langkah-7 Dapatkan nilai optimal untuk *U* dan *V*, dan hitung nilai imputasi $\hat{x}_{ij}$ dalam persamaan (2.5).

Langkah-8 Hitung *error* imputasi dalam persamaan (2.8) dan evaluasi apakah kriteria konvergen dipenuhi. Jika *error* mencapai *thresholds* minimal, *missing data* diestimasi dengan model yang dioptimasi. Sebaliknya, perbaharui parameter, *c* dan *m*, dan kembali ke Langkah-5.



Gambar 2.5 Diagram alir pendekatan integrasi metode imputasi FCM dan optimisasi GA

2.5 Survei Tahunan Perusahaan Industri Manufaktur

Perusahaan yang dijadikan responden dalam survei Tahunan Perusahaan Industri Manufaktur adalah seluruh perusahaan industri manufaktur skala besar dan sedang (*complete enumeration*) yang ada di wilayah Indonesia. Yang dimaksud dengan perusahaan industri manufaktur besar adalah perusahaan industri manufaktur dengan jumlah tenaga kerja 100 orang atau lebih. Perusahaan

industri manufaktur menengah adalah perusahaan industri manufaktur yang mempunyai jumlah tenaga kerja 20 sampai dengan 99 orang. Dikutip dari situs resmi BPS perusahaan industri pengolahan dibagi dalam 4 golongan. Industri Besar (banyaknya tenaga kerja 100 orang atau lebih), Industri Sedang (banyaknya tenaga kerja 20-99 orang), Industri Kecil (banyaknya tenaga kerja 5-19 orang) dan Industri Rumah Tangga (banyaknya tenaga kerja 1-4 orang). Penggolongan perusahaan industri pengolahan ini semata-mata hanya didasarkan kepada banyaknya tenaga kerja yang bekerja. Tanpa memperhatikan apakah perusahaan itu menggunakan tenaga mesin atau tidak, serta tanpa memperhatikan besarnya modal perusahaan itu.

Metode yang digunakan untuk mencatat informasi dalam pengumpulan data ini adalah kombinasi antara wawancara langsung dan tidak langsung (*self–enumeration*). Wawancara langsung biasanya untuk pertanyaan-pertanyaan yang dapat dijawab langsung oleh penanggung jawab perusahaan. Sedangkan wawancara tidak langsung adalah memberikan kuesioner disertai dengan penjelasan teknis tata cara pengisiannya, kemudian kuesioner ditinggal untuk diisi perusahaan. Jika kuesioner sudah diisi lengkap oleh perusahaan pengembaliannya bisa melalui petugas survei, dikirim melalui pos, dikirim melalui *email*, dan dikirim melalui faksimili ke BPS.

Data yang dihasilkan dari survei ini digunakan untuk menyusun publikasi yang menyajikan data industri pengolahan skala besar dan sedang agar dapat membantu para pengguna data dalam menganalisa secara langsung perkembangan sektor industri tanpa harus melakukan pengolahan. Selain itu, publikasi yang dihasilkan juga menyediakan data untuk pemerintah dalam hal pembuatan kebijakan terkait sektor industri pengolahan.

Selain penggolongan industri berdasarkan banyaknya tenaga kerja yang bekerja, industri juga dikelompokkan berdasarkan jenis produksi barang atau jasa industri yang dihasilkan. Menurut jenis produksi barang atau jasa industri yang dihasilkan, industri pengolahan dapat dikelompokkan menjadi 24 kelompok Klasifikasi Baku Lapangan Usaha Indonesia (KBLI). Jenis produksi barang dan jasanya disebut dengan istilah Klasifikasi Komoditi Indonesia (KKI) 2 digit dan 5 digit.

BAB 3
METODOLOGI PENELITIAN

3.1 Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini bersumber dari hasil Survei Tahunan Industri Manufaktur tahun 2013 yang diselenggarakan oleh Badan Pusat Statistik (BPS). Data tersebut meliputi data perusahaan industri manufaktur yang tercatat di Sumatera Utara yang mempunyai tenaga kerja 20 orang atau lebih pada tahun 2013. Terdapat sebanyak 1.006 perusahaan berskala besar dan sedang yang terdiri dari 352 perusahaan besar dan 654 perusahaan sedang.

Kategori industri dengan skala sedang dipilih untuk diteliti karena merupakan industri yang paling dominan di Sumatera Utara. Sampai dengan tahun 2013 jumlah perusahaan yang termasuk dalam kategori ini berjumlah 654 perusahaan atau sekitar 65 persen dari keseluruhan industri besar dan sedang yang terdaftar di Sumatera Utara. Dalam kategori industri ini juga termasuk industri unggulan di Sumatera Utara seperti industri pengolahan kelapa sawit dan pengolahan kopra.

Data yang digunakan sebanyak 534 perusahaan yang diperoleh dengan mengeluarkan perusahaan yang datanya mengandung *missing value*. Pemilihan variabel penelitian dilakukan melalui evaluasi terhadap proporsi *missing value* pada variabel yang terdapat *missing* karena nonrespon. Batas kriteria yang dipakai adalah dengan mengambil batas 5 persen (Schafer, 1999). Berdasarkan hal tersebut maka variabel-variabel yang digunakan dalam penelitian ini adalah sebagai berikut :

X_1 = Bahan bakar dan pelumas, yaitu bahan bakar dan pelumas yang digunakan sebagai bahan pembakar untuk menjalankan mesin, memasak dan lainnya yang dipakai untuk usaha.

X_2 = Tenaga listrik yang dibeli, adalah tenaga listrik yang dibeli dari PLN dan perusahaan non PLN.

X_3 = Pengeluaran lain, yaitu pengeluaran untuk sewa atau kontrak meliputi sewa gedung, tanah, mesin, serta peralatan, pajak tak langsung, biaya yang

dikeluarkan untuk ongkos produksi/pengolahan yang dilakukan oleh pihak lain, pembayaran royalti, kemasan, suku cadang, alat tulis dan komputer suplai, biaya promosi, perjalanan dinas, rekening air, biaya telepon, fax, internet dan surat menyurat dan lain-lain.

3.2 Metode Penelitian

Penelitian diawali dengan melakukan eksplorasi data untuk identifikasi proporsi dan mekanisme *missing data.* Kemudian mengidentifikasi pola *missing data* dan korelasi untuk melihat kekuatan dan arah hubungan antara variabel penelitian.

Percobaan dilakukan menggunakan data industri sedang yang tidak mengandung *missing value* sebanyak 534 perusahaan. Metode *Fuzzy C-Means* dan optimasi dengan Algoritma Genetika kemudian diterapkan untuk melihat kinerja masing-masing metode dan mendapatkan hasil imputasi terbaik.

Percobaan menggunakan data lengkap industri sedang hasil Survei Tahunan Perusahaan Industri Manufaktur 2013 di Sumatera Utara bertujuan untuk melihat kinerja imputasi *Fuzzy C-Means* (FCM) dan hibrida FCM dengan Algoritma Genetika (GA), yaitu melihat performa dari metode apabila diterapkan pada data yang mengandung persentase *missing* yang berbeda-beda. Data untuk penelitian ini terdiri atas tiga variabel $X_1, X_2, \text{dan } X_3$ dengan jumlah 534 observasi.

Pada set data industri sedang dilakukan penghilangan beberapa nilai pada variabel $X_1, X_2, \text{dan } X_3$ dan memperlakukannya sebagai *missing value* untuk mendapatkan lima set data tidak lengkap secara random. Untuk setiap set data persentase *missing value* masing-masing adalah : 5 persen, 10 persen, 15 persen, 20 persen dan 25 persen dari keseluruhan data. Penghilangan secara random dilakukan dengan kondisi bahwa setiap observasi berisi paling tidak satu nilai variabel yang diketahui dan setiap variabel yang mengandung *missing value* paling sedikit terdapat satu komponen yang diketahui. Penghilangan nilai pada variabel dilakukan dengan menggunakan teknik sistematik random sampling dengan langkah-langkah sebagai berikut :

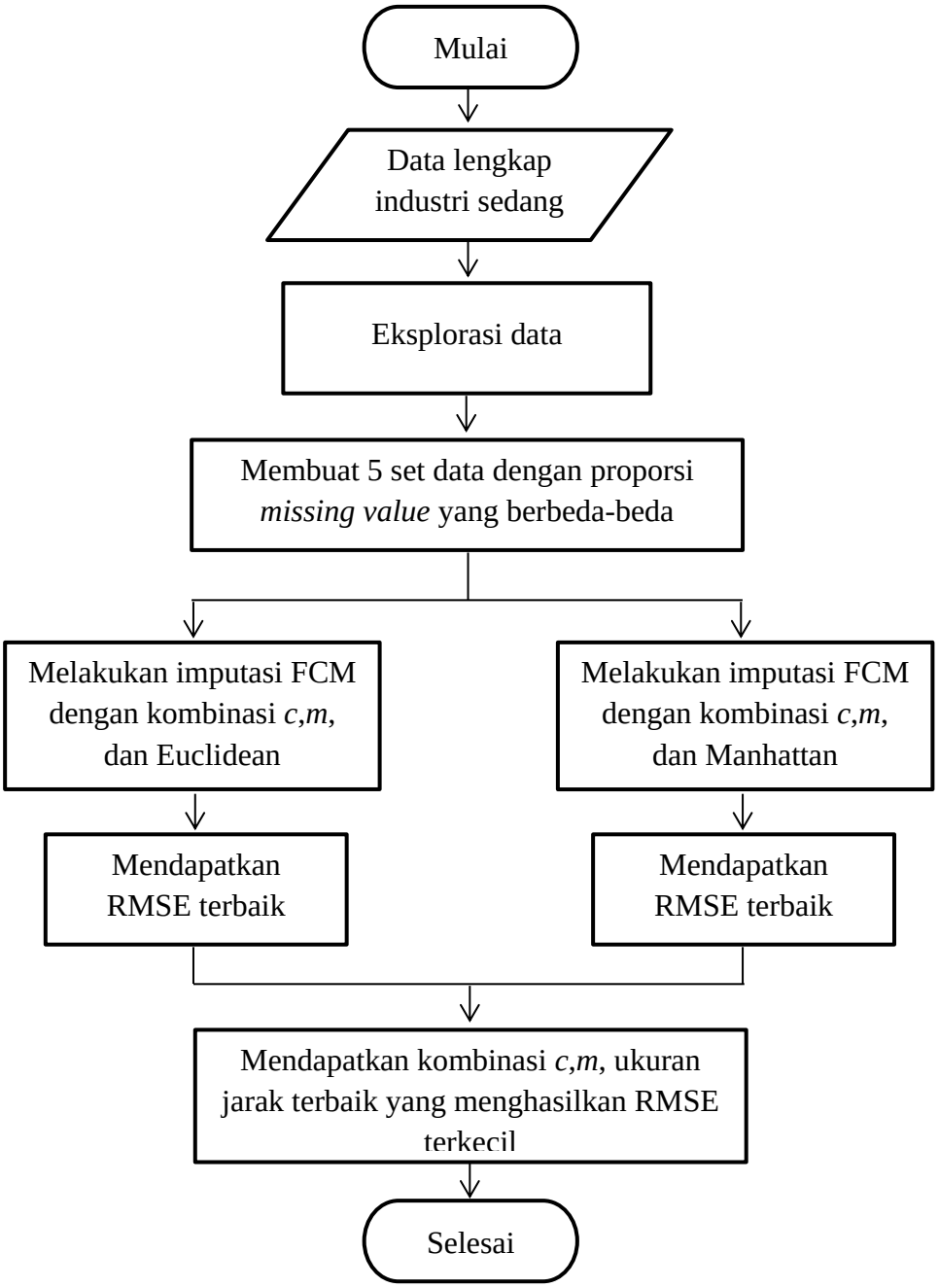
1. Mengurutkan data masing-masing variabel dari yang terkecil sampai yang terbesar.
2. Menentukan interval pengambilan sampel $I = N/n$, dimana *I* adalah interval pengambilan sampel, *N* adalah jumlah observasi, dan *n* adalah jumlah sampel.
3. Menentukan nomor urut observasi yang terpilih sebagai sampel pertama (R_1) secara random, dimana $R_1 < I$. Nomor urut observasi selanjutnya yang terpilih sebagai sampel diperoleh dengan cara menghitung $R_2 = R_1 + I$, $R_3 = R_2 + I$, dan seterusnya.

Selanjutnya dilakukan proses imputasi terhadap setiap set data menggunakan metode FCM dan FCM-GA. Percobaan dilakukan dengan menggunakan ukuran jarak Euclidean dan Manhattan, dan nilai *weighting exponent m* yang berbeda-beda. Pengukuran kinerja hasil imputasi dilakukan dengan membandingkan *Root Mean Square Error (RMSE)* yang dihasilkan dari masing-masing metode.

Adapun langkah-langkah yang dilakukan dalam imputasi dengan metode FCM adalah sebagai berikut :

1. Melakukan eksplorasi data untuk mengidentifikasi pola *missing data* dan korelasi untuk melihat kekuatan dan arah hubungan antara variabel penelitian.
2. Menghilangkan secara random beberapa nilai pada data dengan persentase 5 persen, 10 persen, 15 persen, 20 persen, dan 25 persen untuk menghasilkan 5 set data tidak lengkap dengan tahapan yang telah diuraikan sebelumnya.
3. Melakukan imputasi dengan metode FCM pada setiap set data tidak lengkap dengan mencobakan berbagai kombinasi jumlah klaster *c* dan *weighting exponent* yang sudah ditetapkan dengan menggunakan ukuran jarak Euclidean dan Manhattan
4. Mendapatkan RMSE terkecil untuk masing-masing set data tidak lengkap yang dihasilkan oleh kombinasi parameter jumlah klaster *c, weighting exponent* dan ukuran jarak yang dicobakan.

Untuk lebih jelasnya dapat dilihat melalui diagram alir pada Gambar 3.1 berikut :

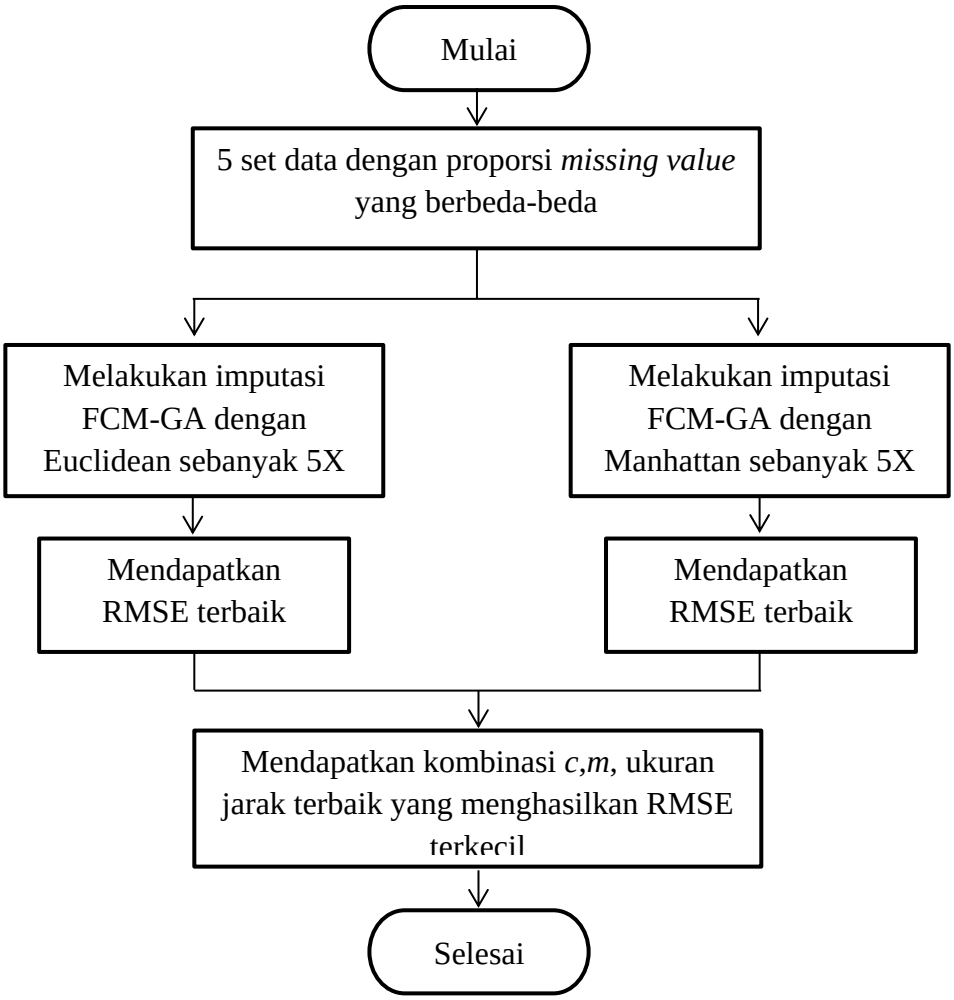


Gambar 3.1 Diagram alir percobaan dengan FCM

Berikutnya adalah melakukan percobaan dengan metode hibrida FCM-GA pada set data yang sama dengan yang digunakan pada percobaan dengan metode FCM, dengan tahapan-tahapan sebagai berikut :

1. Melakukan imputasi dengan metode FCM-GA pada setiap set data tidak lengkap dengan menggunakan ukuran jarak Euclidean dan Manhattan masing-masing sebanyak 5 kali ulangan percobaan.
2. Mendapatkan RMSE terkecil untuk masing-masing set data tidak lengkap yang dihasilkan oleh jumlah klaster *c* dan *weighting exponent* tertentu menggunakan ukuran jarak Euclidean dan Manhattan.

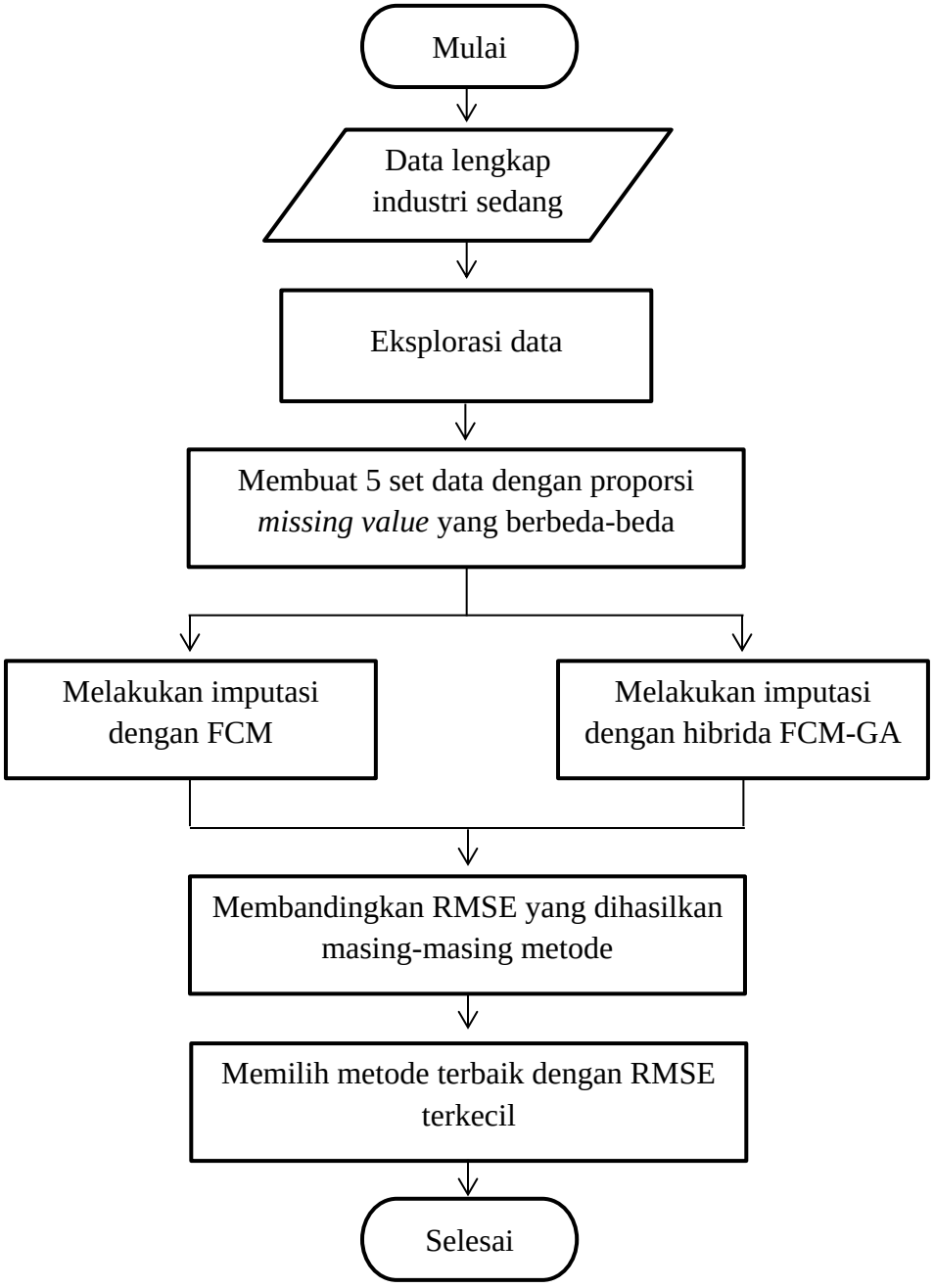
Gambar 3.2 berikut menampilkan diagram alir penelitian dengan FCM-GA :



Gambar 3.2 Diagram alir percobaan dengan FCM-GA

Setelah hasil percobaan diperoleh kemudian akan dibandingkan nilai RMSE terkecil yang didapatkan dari masing-masing metode untuk menentukan

metode yang terbaik. Kriteria metode terbaik adalah yang menghasilkan nilai RMSE terkecil. Secara umum tahapan-tahapan yang dilakukan untuk membandingkan hasil kedua metode untuk mendapatkan hasil terbaik dapat dilihat melalui Gambar 3.3 berikut :



Gambar 3.3 Diagram alir penelitian

BAB 4
HASIL DAN PEMBAHASAN

Percobaan dilakukan untuk menguji performa dari metode Fuzzy C-Means dibandingkan dengan hibrida Fuzzy C-Means dan Algoritma Genetika menggunakan data lengkap industri sedang yang bersumber dari data survei industri besar dan sedang Sumatera Utara tahun 2013.

4.1 Analisis Deskriptif Industri Sedang Sumatera Utara Tahun 2013

Data dalam penelitian ini bersumber dari dari hasil Survei Tahunan Perusahaan Industri Manufaktur 2013 di Sumatera Utara. Data yang digunakan adalah data industri berskala sedang. Variabel yang digunakan adalah bahan bakar dan pelumas (X_1), tenaga listrik yang dibeli (X_2), dan pengeluaran lain (X_3). Eksplorasi dilakukan untuk mengetahui karakteristik data yang digunakan dalam penelitian. Statistik deskriptif industri sedang Sumatera Utara tahun 2013 dapat dilihat pada Tabel 4.1 berikut :

Tabel 4.1 Statistik Deskriptif Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013 (juta rupiah)

Var	N	Rata-rata	St. Dev.	Varian	Min	Max
X_1	604	354,08	1.039,88	1.081.359,40	0,23	16.774,94
X_2	606	410,50	1.755,07	3.080.273,97	0,03	28.922,03
X_3	617	498,94	1.656,45	2.743.815,24	0,15	21.598,23

Dari tabel di atas dapat dilihat bahwa jumlah observasi dengan data lengkap pada masing-masing variabel berbeda-beda. Jumlah observasi lengkap pada variabel X_1 adalah 604 perusahaan. Jumlah observasi dengan data lengkap pada variabel X_2 adalah 606 perusahaan. Jumlah observasi lengkap pada variabel X_3 adalah 617 perusahaan.

Tabel 4.2 menampilkan *missing* data yang tersebar pada data perusahaan industri sedang di Sumatera Utara tahun 2013. Dari 654 perusahaan terdapat 534

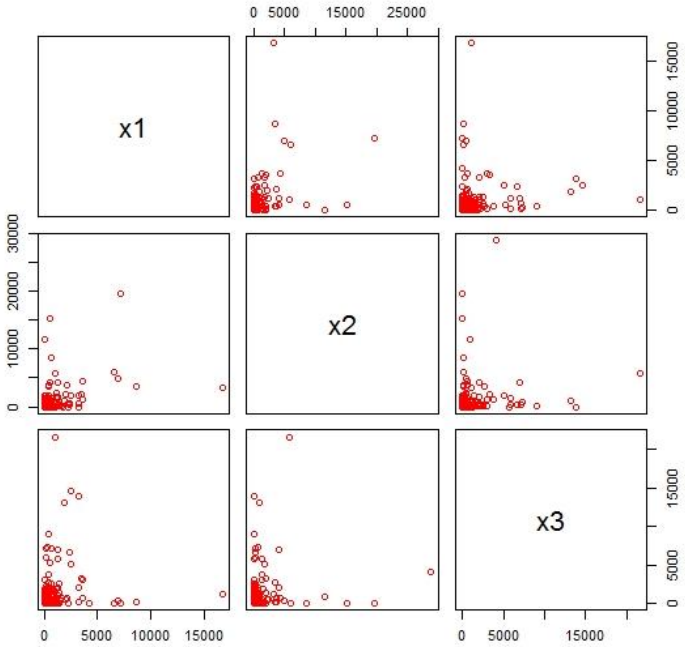
perusahaan dengan data yang lengkap. Sementara pada 120 perusahaan terdapat data *missing* yang tersebar pada item variabel.

Tabel 4.2 *Missing Data* Pada Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013 (juta rupiah)

Perusahaan	X_1	X_2	X_3
1	10,70	?	0,92
2	27,40	?	165,50
3	35,31	?	223,65
4	9,28	?	0,55
…	…	…	…
145	53,07	173,24	?
146	82,97	253,03	42,48
147	?	345,00	25,80
…	…	…	…
248	347,42	?	4,50
249	2,50	?	?
250	?	0,30	14,50
251	26,52	0,60	25,00
…	…	…	…
Jumlah *Missing* (Persen)	7,65	7,34	5,66

Keterangan : Simbol '?' menyatakan data yang *missing*

Jumlah pengeluaran paling sedikit yang dikeluarkan oleh perusahaan untuk bahan bakar dan pelumas selama setahun pada tahun 2013 adalah 0,23 juta rupiah atau sekitar 230 ribu rupiah. Sementara pengeluaran paling besar untuk bahan bakar dan pelumas dalam setahun pada tahun 2013 adalah 16.774,94 juta rupiah atau sekitar 16,77 milyar rupiah. Rentang nilai minimum dan maksimum pada variabel ini sangat lebar jika dibandingkan dengan rata-rata pengeluaran perusahaan untuk bahan bakar dan pelumas sebesar 354,08 juta rupiah setahun. Deskripsi data yang sama juga dapat dilihat pada variabel tenaga listrik yang dibeli dan variabel pengeluaran lain yang memiliki perbedaan nilai minimum dan maksimum yang sangat besar.



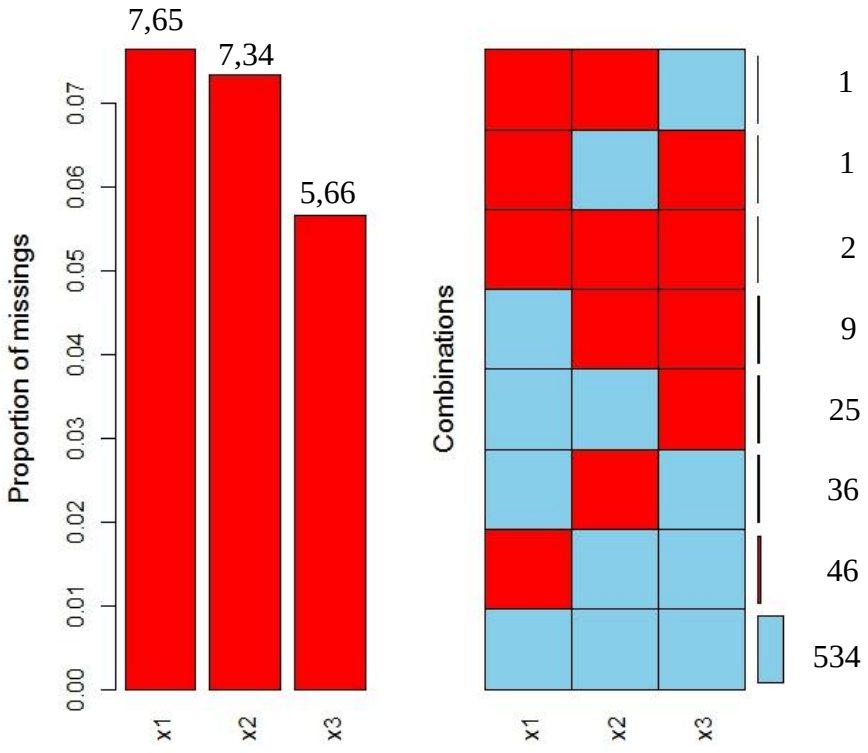
Gambar 4.1 *Scatterplot* Hubungan Antara Variabel Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013.

Pola hubungan antara variabel-variabel dapat dilihat dengan memperhatikan sebaran data pada *scatterplot* pada Gambar 4.1. Dari plot data diperoleh gambaran bahwa hubungan antar variabel tidak menunjukkan hubungan yang linier. Baik hubungan antara variabel bahan bakar dan pelumas dengan variabel tenaga listrik yang dibeli, variabel bahan bakar dan pelumas dengan variabel pengeluaran lain, dan variabel listrik dengan variabel pengeluaran lain, tidak menunjukkan adanya suatu pola hubungan yang linier. Sebagian besar data mengelompok pada rentang nilai tertentu. Dari gambar sebaran data dapat dilihat bahwa plot data bahan bakar dan pelumas dengan listrik tidak menampakkan suatu pola tertentu. Demikian pula plot data bahan bakar dan pelumas dengan data pengeluaran lain tidak menunjukkan adanya suatu pola tertentu. Gambaran yang sama ditunjukkan oleh plot data tenaga listrik yang dibeli dengan data pengeluaran lain menampakkan sebaran data yang tidak beraturan. Hal ini menunjukkan bahwa tidak ada suatu pola hubungan yang dapat disimpulkan dari ketiga variabel tersebut.

Kekuatan dan arah hubungan antara ketiga variabel dapat dilihat melalui matriks korelasi berikut :

$$\hat{\rho} = \begin{bmatrix} 1 & 0,430 & 0,230 \\ 0,430 & 1 & 0,192 \\ 0,230 & 0,192 & 1 \end{bmatrix}$$

Dari matrik korelasi dapat dilihat bahwa semua variabel memiliki hubungan yang positif. Meski demikian, tidak terdapat korelasi yang kuat di antara ketiga variabel yang diteliti. Korelasi antara variabel bahan bakar dan pelumas dengan variabel listrik masuk dalam kategori korelasi lemah. Korelasi antara variabel bahan bakar dan pelumas dengan variabel pengeluaran lain juga korelasi lemah. Serupa dengan korelasi antara variabel tenaga listrik yang dibeli dengan variabel pengeluaran lain juga korelasi lemah. Korelasi lemah diantara ketiga variabel sejalan dengan pola sebaran data yang tidak beraturan dan tidak menunjukkan pola hubungan yang linier.



(a) (b)

Gambar 4.2 Jumlah Data Hilang Pada Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013 a) *Bar Plot* , (b) *Aggregation Plot*.

Seperti yang ditunjukkan pada Gambar 4.2 *Bar Plot* (a) merupakan grafik batang untuk setiap variabel, dan tinggi grafik batang menyatakan jumlah *missing value* pada variabel tersebut dalam persentase. *Aggregation Plot* (b) menunjukkan semua kombinasi *missing* dan *non-missing* dari setiap variabel yang ada pada observasi. Warna merah mengindikasikan *missingness*, warna biru menyatakan data yang tersedia. Bagian kanan plot menunjukkan frekuensi observasi dari kombinasi *missing* dan *non-missing* pada variabel terkait. Dari plot *missing* data pada item variabel tersebut dapat dilihat bahwa kombinasi data hilang pada data industri sedang tersebar pada semua variabel secara acak atau memiliki pola yang random.

4.2 Percobaan Dengan Data Riil

4.2.1 Analisis Deskriptif Data Lengkap Industri Sedang Sumatera Utara Tahun 2013

Percobaan dengan menggunakan data riil menggunakan data lengkap industri sedang yang bersumber dari hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013. Variabel yang digunakan adalah variabel bahan bakar dan pelumas (X_1), variabel tenaga listrik yang dibeli (X_2), dan variabel pengeluaran lain (X_3).

Pemilihan ketiga variabel yang digunakan dalam penelitian ini berdasarkan evaluasi terhadap proporsi *missing value* pada variabel yang terdapat *missing* karena nonrespon dengan batas kriteria 5 persen. Persentase *missing value* pada masing-masing variabel adalah 7,65 persen pada variabel X_1, variabel X_2 sebanyak 7,34 persen, dan pada variabel X_3 sebanyak 5,66 persen. Statistik deskriptif data lengkap industri sedang Sumatera Utara tahun 2013 dapat dilihat pada Tabel 4.3.

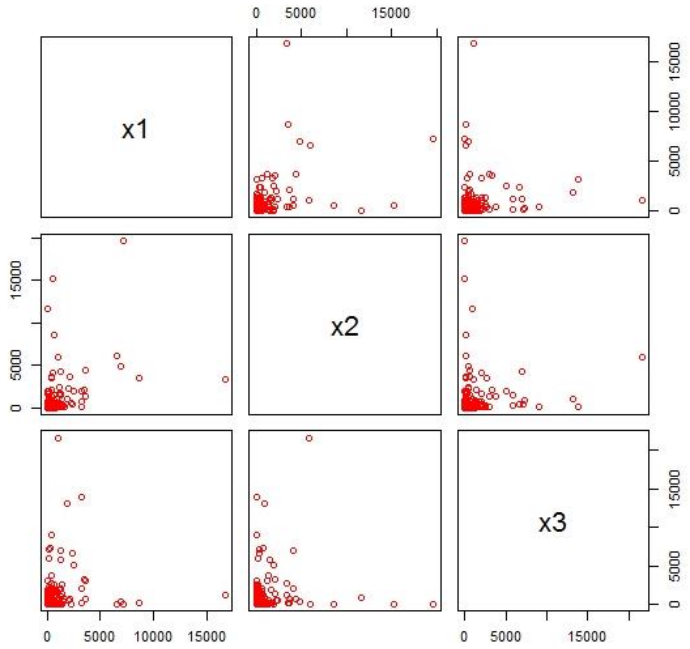
Dari tabel diketahui bahwa jumlah observasi dengan data lengkap adalah sama untuk ketiga variabel yaitu sebanyak 534 perusahaan. Jumlah pengeluaran paling sedikit dan paling besar yang dikeluarkan oleh perusahaan untuk bahan bakar dan pelumas selama setahun pada tahun 2013 sama dengan jumlah

pengeluaran minimum dan maksimum pada data awal. Sementara rata-rata pengeluaran perusahaan untuk bahan bakar dan pelumas sebesar 343,88 juta rupiah setahun. Terdapat selisih yang sangat besar antara nilai minimum dan maksimum pada variabel ini dibandingkan dengan nilai rata-ratanya. Rentang nilai yang lebar ini juga ditunjukkan oleh nilai varians yang besar. Sama halnya dengan pengeluaran perusahaan untuk membeli listrik dan pengeluaran lain juga memiliki rentang nilai minimum dan maksimum yang sangat besar dan nilai varians yang besar.

Tabel 4.3 Statistik Deskriptif Data Lengkap Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013 (juta rupiah)

Var	N	Rata-rata	St.Dev.	Varian	Min	Max
X_1	534	343,88	1.073,04	1.151.405,32	0,23	16.774,94
X_2	534	389,15	1.398,41	1.955.553,02	0,18	19.669,20
X_3	534	470,93	1.577,54	2.488.619,00	0,15	21.598,23

Pola hubungan antar variabel-variabel pada data lengkap industri sedang hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013 dapat dilihat dengan memperhatikan sebaran data pada Gambar 4.9. Sebaran data lengkap industri sedang terlihat tidak berbeda jika dibandingkan dengan sebaran pada data awal sebelum dilakukan pemisahan objek yang ada *missing value* yaitu data industri besar dan sedang. Antara variabel X_1 dengan variabel X_2, variabel X_1 dengan variabel X_3 dan variabel X_2 dengan variabel X_3, tidak menunjukkan adanya suatu pola hubungan yang linier. Sebagaimana pada data industri besar dan sedang, sebagian besar data pada data lengkap industri sedang juga mengelompok pada rentang nilai tertentu.



Gambar 4.3 *Scatterplot* Hubungan Antara Variabel Data Lengkap Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013.

Plot data X_1 dengan data X_2 tidak memperlihatkan suatu pola hubungan yang linier. Hal yang sama juga terlihat pada plot data X_1 dengan data X_3 tidak menunjukkan adanya suatu pola tertentu. Plot sebaran data X_2 dengan data X_3 juga memperlihatkan sebaran data yang tidak beraturan. Jadi serupa dengan data awal, sebaran data pada data lengkap industri sedang menunjukkan bahwa tidak ada suatu pola hubungan yang linier dari ketiga variabel tersebut.

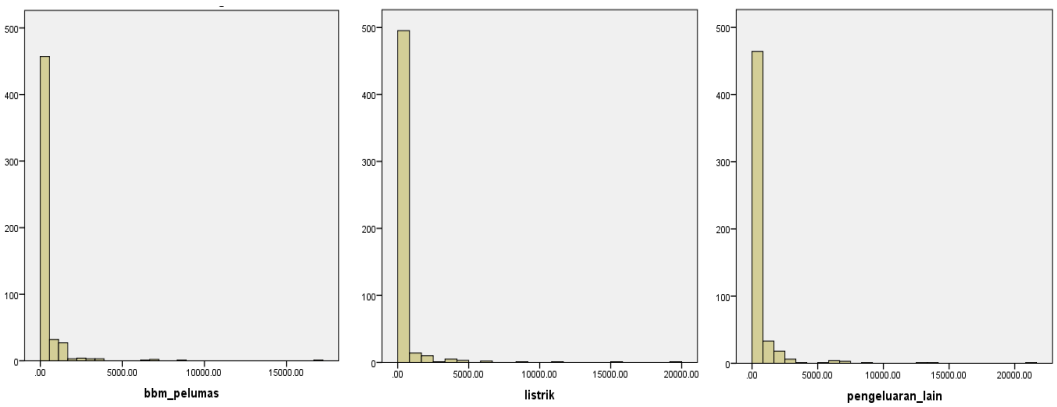
Kekuatan dan arah hubungan antara variabel pada data lengkap industri sedang juga menunjukkan kesamaan dengan data awal industri sedang yang dapat dilihat pada matriks korelasi berikut :

$$\hat{\rho} = \begin{bmatrix} 1 & 0,435 & 0,221 \\ 0,435 & 1 & 0,173 \\ 0,221 & 0,173 & 1 \end{bmatrix}$$

Dari matrik korelasi dapat dilihat bahwa antar variabel memiliki hubungan yang positif. Namun tidak terdapat korelasi yang kuat di antara ketiga

variabel yang diteliti. Korelasi antara ketiga variabel tergolong korelasi lemah. Baik itu korelasi antara variabel X_1 dengan variabel X_2, korelasi antara variabel X_1 dengan variabel X_3, dan korelasi antara variabel X_2 dengan X_3. Korelasi yang lemah diantara ketiga variabel sejalan dengan pola sebaran data yang tidak beraturan dan tidak menunjukkan pola hubungan yang linier.

Selanjutnya dapat dilihat distribusi data masing-masing variabel melalui histrogram pada Gambar 4.4 berikut. Pada histrogram dapat dilihat bahwa distribusi data menceng kanan karena terdapat beberapa observasi dengan nilai ekstrim yang terdapat pada semua variabel. Nilai-nilai ekstrim tersebut adalah pencilan.



(a) (b) (c)

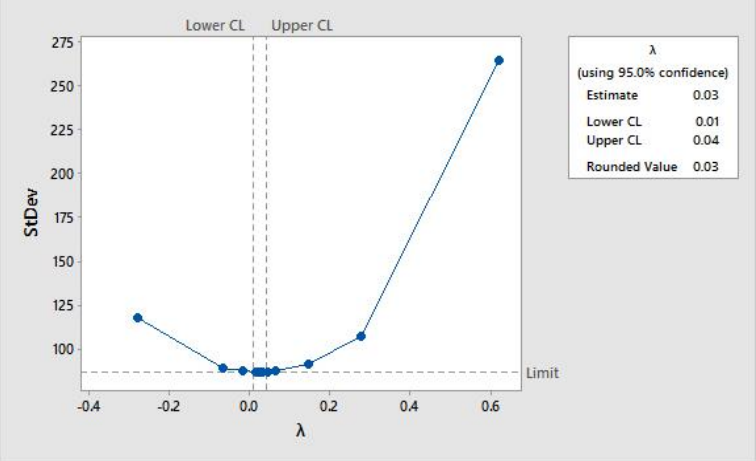
Gambar 4.4 Histogram Data Lengkap Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013 a) Variabel bahan bakar dan pelumas, (b) Variabel listrik yang dibeli, dan (c) Variabel pengeluaran lain.

Untuk mengatasi rentang nilai yang sangat lebar dan distribusi data menceng pada variabel penelitian, maka dilakukan tranformasi data. Hasil tranformasi Box-Cox pada Gambar 4.5 menunjukkan bahwa nilai λ = 0,03 atau mendekati nol. Sehingga diputuskan untuk menggunakan tranformasi algoritma natural. Statistik deskriptif data lengkap industri sedang Sumatera Utara tahun 2013 setelah data ditransformasi dapat dilihat pada Tabel 4.4 berikut :

Tabel 4.4 Statistik Deskriptif Data Transformasi Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara 2013

Var	N	Rata-rata	St.Dev.	Varian	Min	Max
X_1	534	4,1731	1,8784	3,5285	-1,4697	9,7276
X_2	534	3,9656	2,2325	4,9842	-1,7148	9,8868
X_3	534	4,1831	1,9463	3,7880	-1,8971	9,9804

Dengan jumlah observasi yang tetap sama sebanyak 534 perusahaan, rentang nilai minimum dan maksimum pada masing-masing variabel menjadi lebih pendek. Besaran nilai varians juga tidak berbeda jauh dibandingkan dengan nilai rata-rata.



Gambar 4.5 Box-Cox plot data transformasi.

4.2.2 Pengolahan Data Industri Sedang Sumatera Utara Tahun 2013

Data yang digunakan dalam penelitian ini adalah data industri sedang hasil Survei Tahunan Perusahaan Industri Manufaktur di Sumatera Utara tahun 2013 yang tidak terdapat *missing value* di dalamnya sesuai dengan batasan masalah yang disampaikan pada bab sebelumnya. Setelah data ditranformasi kemudian dilakukan penghilangan nilai secara random pada data yang akan dijadikan sebagai observasi.

Dalam proses penghilangan nilai secara random kondisi yang ditetapkan adalah setiap observasi harus memiliki setidaknya satu nilai variabel yang

diketahui dan setiap variabel harus memiliki setidaknya satu nilai observasi yang diketahui. Nilai yang dihilangkan diperlakukan sebagai *missing data* dan observasi dengan nilai variabel yang *missing* tersebut akan dijadikan sebagai observasi yang terdapat data hilang, dengan mengikuti asumsi mekanisme *Missing At Random* (MAR).

Jumlah observasi dengan nilai yang hilang ini disimulasikan masing-masing sebanyak 5 persen, 10 persen, 15 persen, 20 persen, dan 25 persen dari keseluruhan data. Penentuan nilai proporsi *missing* untuk masing-masing kelompok data tersebut didasarkan atas pertimbangan proporsi *missing* pada data riil yang berada pada kisaran 5 persen sampai 25 persen. Sementara interval proporsi *missing* ditetapkan 5 persen mengacu pada interval yang digunakan oleh penelitian-penelitian terdahulu yang menggunakan data hasil survei industri besar dan sedang. Statistik deskriptif dari masing-masing kelompok data yang dibentuk disajikan pada Tabel 4.5 berikut ini :

Tabel 4.5 Statistik Deskriptif Data Transformasi Industri Sedang Berdasarkan Persentase *Missing*

Kelompok Data	Var	N	Rata-rata	St.Dev.	Varian	Min	Max
Missing 5 persen	X_1	502	4,1645	1,8434	3,3980	-1,4697	9,7276
	X_2	515	4,0050	2,2284	4,9659	-1,7148	9,8868
	X_3	505	4,2125	1,9240	3,7019	-1,8971	9,9804
Missing 10 persen	X_1	471	4,1866	1,8498	3,4216	-1,4697	9,7276
	X_2	488	3,9951	2,2018	4,8477	-1,7148	9,8868
	X_3	483	4,1811	1,9206	3,6888	-1,8971	9,9804
Missing 15 persen	X_1	459	4,1316	1,8595	3,4577	-1.4697	9,7276
	X_2	452	3,9856	2,2434	5,0328	-1,7148	9,8868
	X_3	451	4,2379	1,9684	3,8745	-1,8971	9,9804
Missing 20 persen	X_1	428	4,2727	1,8581	3,4524	-1,2040	9,7276
	X_2	420	4,0743	2,2185	4,9216	-1,6607	9,8868
	X_3	434	4,1454	1,9806	3,9229	-1,8971	9,9804

Tabel 4.5 (Lanjutan)

Kelompok Data	Var	N	Rata-rata	St.Dev.	Varian	Min	Max
Missing 25 persen	X_1	400	4,2225	1,8601	3,4600	-1,4697	9,7276
	X_2	395	3,9237	2,3284	5,4216	-1,7148	9,8868
	X_3	407	4,1888	1,9842	3,9370	-1,8971	9,9804

4.2.2.1 Imputasi Menggunakan Algoritma *Fuzzy C-Means* (FCM)

Berikutnya adalah melakukan percobaan menggunakan algoritma FCM untuk menduga nilai yang hilang kemudian melakukan imputasi pada masing-masing kelompok data yang *missing*. Kinerja dari algoritma imputasi FCM dilihat dengan menghitung nilai *Root Mean Square Error* (RMSE) yang dihasilkan.

Percobaan dilakukan berdasarkan dua input parameter. Pertama adalah memilih ukuran jarak Euclidean. Lalu menentukan nilai *weighting exponent* dalam rentang nilai *m* = 1,2 sampai dengan *m* = 3 yang nanti akan digunakan. Kemudian metode FCM diterapkan menggunakan jumlah klaster yang berbeda-beda. Klaster yang dicobakan mulai dari klaster = 2 sampai dengan klaster = 9. Percobaan dengan metode yang sama juga dilakukan dengan menggunakan ukuran jarak Manhattan. Menggunakan beberapa nilai *weighting exponent* dalam rentang *m* = 1,2 sampai dengan *m* = 3. Kemudian menerapkan metode FCM menggunakan klaster =2 sampai dengan klaster = 9.

Tabel 4.6 RMSE Algoritma FCM Dengan Persentase *Missing Value* dan *Weighting Exponent* Yang Berbeda-beda Menggunakan Ukuran Jarak Euclidean

Persentase *Missing*	Euclidean, Klaster = 2					
	m = 1,2	*m* =1,5	*m* =1,7	*m* =2	*m* = 2,5	*m* = 3
5	265,28	265,23	266,01	267,37	268,98	269,85
10	296.53	296.62	297.98	300.34	303.11	304.59
15	319,23	320,49	322,40	325,13	328,20	329,89
20	464.63	464.78	465.76	467.43	469.63	471.02
25	348,62	349,36	351,06	353,70	356,95	358,93

Tabel 4.6 menampilkan nilai RMSE hasil imputasi dengan metode FCM menggunakan jarak Euclidean pada klaster = 2. Percobaan dilakukan pada lima kelompok data dengan persentase *missing* yang berbeda-beda. Dari tabel tersebut dapat dilihat untuk setiap *weighting exponent* yang diberikan nilai RMSE yang dihasilkan cenderung mengalami peningkatan. Pada kelompok *missing* 20 persen terjadi peningkatan nilai yang cukup tinggi, namun nilainya kembali turun pada kelompok *missing* berikutnya yaitu 25 persen. Meski turun tetapi nilai RMSE pada kelompok *missing* 25 persen masih lebih besar dibandingkan pada kelompok *missing* 15 persen. Pola fluktuasi yang terjadi diikuti dengan kecenderungan peningkatan RMSE seiring dengan peningkatan persentase *missing* data. Peningkatan nilai RMSE dengan pola seperti ini terjadi pada semua nilai *m* yang dicobakan.

Peningkatan nilai RMSE juga terlihat dipengaruhi oleh besaran nilai *weighting exponent* yang diberikan. Pada masing-masing kelompok *missing data*, semakin besar nilai *m* cenderung menghasilkan nilai RMSE yang semakin tinggi. Pola peningkatan nilai RMSE ini terlihat pada semua kelompok *missing data* yang ada. Secara umum dapat dikatakan bahwa untuk nilai-nilai *m* yang dicobakan pada kelompok *missing data* berbeda-beda, semakin tinggi persentase *missing* akan menghasilkan nilai RMSE yang juga semakin besar.

Jika diperhatikan w*eighting exponent* yang memberikan nilai RMSE terkecil untuk masing-masing kelompok *missing* selalu dihasilkan oleh nilai *m* yang sama, kecuali pada kelompok *missing* 5 persen. Untuk kelompok data dengan *missing* 10 persen,15 persen, 20 persen dan 25 persen dihasilkan oleh m = 1,2. Pada kelompok data dengan *missing* 5 persen dihasilkan oleh nilai m = 1,5.

Percobaan dengan jumlah klaster yang sama dengan menggunakan jarak Manhattan ditampilkan pada Tabel 4.7. Hasil yang diperoleh secara umum juga mencerminkan pengaruh persentase *missing* terhadap RMSE dengan pola yang mirip. Pola fluktuasi yang terjadi diikuti dengan kecenderungan peningkatan RMSE seiring dengan peningkatan persentase *missing* data. Semakin tinggi persentase *missing* akan meningkatkan nilai RMSE yang dihasilkan. Pengaruh nilai *weighting exponent* pada tiap kelompok data juga terlihat. Seperti hasil percobaan menggunakan jarak Euclidean, semakin besar nilai *m* yang digunakan

cenderung menghasilkan nilai RMSE yang semakin tinggi pada masing-masing kelompok *missing data*. Hal ini mengindikasikan bahwa pemilihan *weighting exponent* yang sesuai adalah penting untuk mendapatkan hasil terbaik.

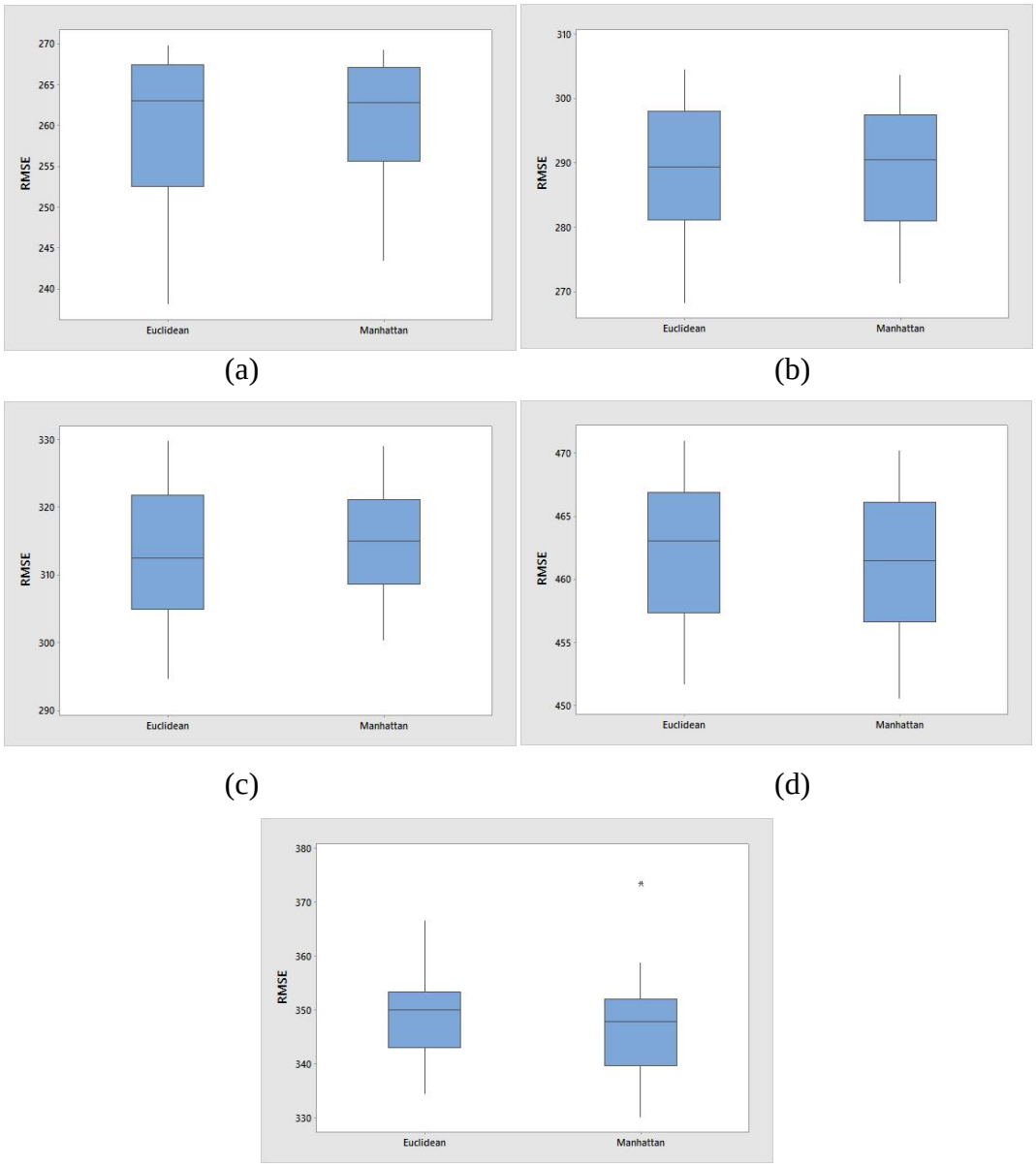
Tabel 4.7 RMSE Algoritma FCM Dengan Persentase *Missing Value* dan *Weighting Exponent* Yang Berbeda-beda Menggunakan Ukuran Jarak Manhattan

Persentase *Missing*	Manhattan, Klaster =2					
	m = 1,2	m =1,5	m =1,7	m =2	m = 2,5	m = 3
5	264,47	264,41	265,14	266,54	268,31	269,33
10	295.9	295.86	297.04	299.28	302.15	303.80
15	318,59	319,49	321,25	323,97	327,22	329,10
20	463.04	463.60	464.58	466.29	468.66	470.22
25	347,03	347,84	349,46	352,10	355,50	357,70

Serupa dengan percobaan menggunakan ukuran jarak Euclidean, *weighting exponent* yang menghasilkan nilai RMSE terkecil untuk masing-masing kelompok *missing* data sama kecuali untuk kelompok *missing* 5 persen. Untuk kelompok data dengan *missing* 5 persen dihasilkan oleh m = 1,5. Sementara pada kelompok data yang lain dihasilkan oleh nilai m = 1,2.

Berikutnya, pada Tabel 4.8 dapat dilihat bahwa penggunaan ukuran jarak Euclidean dan Manhattan tidak menunjukkan ada yang lebih superior diantara keduanya, baik pada persentase *missing data* tertentu atau pada jumlah klaster tertentu. Pada jumlah klaster dan persentase *missing data* yang berbeda, penggunaan Euclidean dan Manhattan secara bergantian menjadi ukuran jarak terbaik yang memberikan nilai RMSE terkecil.

Gambar 4.6 adalah *boxplot* perbandingan RMSE yang dihasilkan dengan menggunakan ukuran jarak Euclidean dan Manhattan. Dalam masing-masing *boxplot*, garis tengah menyatakan median dan garis batas bawah dan batas atas menyatakan persentil ke-25 dan persentil ke-75 dari nilai RMSE. Nilai median yang lebih kecil menunjukkan akurasi yang lebih baik. Untuk masing-masing kelompok *missing* dapat dilihat perbandingan akurasi antara dua ukuran jarak tersebut.



Gambar 4.6 *Boxplot* perbandingan RMSE ukuran jarak Euclidean dan Manhattan (a) *missing* 5 persen, (b) *missing* 10 persen, (c) *missing* 15 persen, (d) *missing* 20 persen, dan (e) *missing* 25 persen.

Untuk kelompok *missing* 5 persen, median Manhattan lebih kecil daripada median Euclidean. Hal ini mengindikasikan bahwa pemakaian ukuran jarak Manhattan pada berbagai kombinasi jumlah klaster dan *weighting exponent* untuk kelompok data tersebut secara umum memberikan hasil yang lebih baik dibandingkan Euclidean. Sementara pada kelompok *missing* 10 persen dan 15 persen, giliran ukuran jarak Euclidean yang lebih baik. Seterusnya untuk

kelompok *missing* 20 persen, dan 25 persen, median Manhattan lebih kecil dibandingkan dengan median Euclidean yang berarti secara umum menghasilkan akurasi yang lebih baik. Tetapi perbedaan nilai median yang tidak terlalu besar mengindikasikan bahwa perbedaan akurasi yang disebabkan pemakaian kedua ukuran jarak tersebut tidak terlalu besar.

Tabel 4.8 RMSE Algoritma FCM Dengan Ukuran Jarak dan Jumlah Klaster Yang Berbeda-beda (m = 2)

Persentase *Missing*	c = 3		c = 6		c = 9	
	Euclidean	Manhattan	Euclidean	Manhattan	Euclidean	Manhattan
5	263,15	262,38	264,46	263,78	260,22	261,24
10	289.97	290.59	292.94	290.49	284.66	291.67
15	314,58	315,20	310,65	310,24	315,62	315,57
20	459.66	460.41	462.34	460.61	463.82	462.83
25	342,74	341,98	343,62	345,60	359,79	354,21

Bila performa kedua ukuran jarak dikombinasikan dengan penggunaan *weighting exponent* yang berbeda pada jumlah klaster tertentu, dapat diamati adanya indikasi ukuran jarak Manhattan cenderung memberikan hasil yang lebih baik dibandingkan ukuran jarak Euclidean pada kelompok data dengan persentase *missing* yang lebih tinggi dan nilai m yang lebih besar.

Tabel 4.9 RMSE Algoritma FCM Dengan Ukuran Jarak dan *Weighting Exponent* Yang Berbeda-beda (klaster = 6)

Persentase *Missing*	m = 1.2		m = 2		m = 3	
	Euclidean	Manhattan	Euclidean	Manhattan	Euclidean	Manhattan
5	238,14	249,30	264,46	263,78	268,65	268,41
10	268.11	271.24	292.94	290.49	301.19	300.10
15	294,57	318,30	310,65	310,24	322,12	320,54
20	454.16	456.61	462.34	460.61	469.01	467.26
25	342,58	347,77	344,52	345,60	352,32	353,37

Tabel 4.9 adalah nilai RMSE hasil imputasi metode FCM yang dicobakan pada klaster = 6 dengan menggunakan nilai m = 1,2, m = 2 dan m = 3. Dapat diamati bahwa pemakaian ukuran jarak Euclidean memberikan nilai RMSE

yang lebih kecil dibandingkan jarak Manhattan menggunakan *weighting exponent* m = 1,2. Namun ukuran jarak Manhattan didapati lebih sering mengungguli jarak Euclidean ketika *weighting exponent* yang digunakan m = 2 dan m = 3.

Setelah rangkaian percobaan selesai dilakukan pada setiap kelompok *missing data*, dari perhitungan hasil imputasi dengan metode FCM didapatkan parameter terbaik untuk setiap kelompok data. Kriteria terbaik adalah kombinasi parameter yang memberikan nilai RMSE terkecil.

Tabel 4.10 Parameter Terbaik Algoritma FCM Berdasarkan Nilai RMSE Terkecil

Persentase *Missing*	Parameter			
	Euclidean		Manhattan	
	c	***m***	***c***	***m***
5	6	1,20	9	1,20
10	6	1,50	6	1,20
15	6	1,20	5	1,50
20	7	1,20	9	1,70
25	3	1,50	3	1,20

Tabel 4.10 menampilkan kombinasi parameter terbaik untuk dua ukuran jarak yang digunakan. Dari hasil percobaan dapat dilihat bahwa persentase *missing* yang berbeda menghasilkan kombinasi parameter yang bervariasi untuk mendapatkan nilai RMSE terbaik. Tidak ada kombinasi parameter yang sama pada setiap kelompok *missing data* yang menghasilkan RMSE terbaik. Jika diperhatikan ada indikasi bahwa nilai RMSE terbaik dihasilkan oleh *weighting exponent* yang kecil. Hal ini dapat dilihat dari nilai *weighting exponent* yang menghasilkan nilai RMSE terbaik, baik menggunakan ukuran jarak Eclidean dan Manhattan selalu bernilai < 2.

Setelah percobaan selesai dilakukan menggunakan berbagai kombinasi parameter yang dikemukakan sebelumnya, kemudian dihitung nilai RMSE dari setiap kombinasi parameter tersebut. Pada Tabel 4.11 disajikan nilai RMSE terbaik dari kombinasi parameter hasil imputasi pada masing-masing kelompok *missing data* menggunakan algoritma FCM sebagai berikut :

Tabel 4.11 Nilai RMSE Terbaik Menggunakan Algoritma FCM

Persentase *Missing*	***c***	***m***	RMSE	Ukuran Jarak
5	6	1,20	238,14	Euclidean
10	6	1,50	268,11	Euclidean
15	6	1,20	294,57	Euclidean
20	9	1,70	450,54	Manhattan
25	3	1,20	329,97	Manhattan

Nilai RMSE terbaik pada kelompok *missing* 5 persen, 10 persen, dan 15 persen dihasilkan dari kombinasi parameter menggunakan jarak Euclidean. Sementara pada kelompok dengan persentase *missing* 20 persen dan 25 persen dihasilkan oleh ukuran jarak Manhattan. Hal ini mendukung indikasi sebelumnya bahwa pada persentase *missing* yang lebih tinggi, kinerja ukuran jarak Manhattan cenderung lebih baik dibandingkan Euclidean.

4.2.2.2 Imputasi Menggunakan Hibrida *Fuzzy C-Means* dan Algoritma Genetika (FCM-GA)

Setelah selesai melakukan percobaan dengan algoritma FCM dan hasilnya telah diperoleh, kemudian percobaan dilanjutkan dengan menerapkan metode hibrida FCM dan Algoritma Genetika (FCM-GA). Sebelum melakukan optimasi dengan Algoritma Genetika (GA) didefinisikan terlebih dahulu beberapa ukuran dan batasan yang digunakan. Ukuran populasi ditetapkan $N = 50$ jumlah generasi $T = 50$, dan probabilitas persilangan $P_c = 0,8$. Solusi terbaik diperoleh setelah kriteria yang ditentukan terpenuhi, yaitu ketika mencapai generasi maksimum 50 generasi atau selisih nilai *fitness* terbaik dalam 5 generasi terakhir tidak lebih dari 1×10^{-8}. Parameter dipilih berdasarkan hasil terbaik dari lima kali ulangan percobaan.

Tabel 4.12 menampilkan kombinasi parameter hasil optimasi dengan GA pada masing-masing kelompok *missing data*. Kombinasi parameter tersebut menghasilkan nilai RMSE terkecil untuk masing-masing kelompok *missing data* yang dicobakan. Seperti yang dapat dilihat pada tabel tersebut, hasil optimasi GA menghasilkan kombinasi parameter yang berbeda-beda. Jika dibandingkan antara

ukuran jarak Euclidean dan Manhattan kombinasi parameter yang dihasilkan tidak ada yang sama, baik jumlah klaster dan nilai *weighting exponent.*

Tabel 4.12 Optimisasi Parameter Dengan Algoritma FCM-GA

Persentase *Missing*	Parameter			
	Euclidean		Manhattan	
	c	***m***	***c***	***m***
5	6	1,04	9	1,09
10	6	1,03	8	1,05
15	8	1,34	6	1,38
20	8	1,06	3	1,03
25	3	1,40	3	1,11

Kemudian dari hasil percobaan yang dilakukan didapatkan nilai RMSE terbaik hasil imputasi menggunakan algoritma hibrida FCM-GA yang ditampilkan pada Tabel 4.13 berikut :

Tabel 4.13 Nilai RMSE Terbaik Menggunakan Hibrida FCM-GA

Persentase *Missing*	***c***	***m***	RMSE	Ukuran Jarak
5	6	1,04	227,14	Euclidean
10	6	1,03	257,20	Euclidean
15	8	1,34	289,49	Euclidean
20	3	1,03	445,79	Manhattan
25	3	1,11	329,21	Manhattan

Kombinasi parameter menggunakan jarak Euclidean memberikan hasil terbaik pada persentase *missing* 5 persen, 10 persen, dan 20 persen. Sementara kombinasi parameter menggunakan ukuran jarak Manhattan memberikan hasil terbaik pada persentase *missin*g 20 persen dan 25 persen.

4.2.2.3 Perbandingan Hasil Imputasi Algoritma FCM dan Hibrida FCM-GA

Tabel 4.14 menampilkan perbandingan nilai RMSE dan kombinasi parameter yang menghasilkan nilai RMSE terbaik untuk masing-masing metode menggunakan ukuran jarak Euclidean.

Tabel 4.14 RMSE Algoritma FCM dan FCM-GA Dengan Ukuran Jarak Euclidean

Persentase *Missing*	FCM			FCM-GA		
	c	***m***	RMSE	***c***	***m***	RMSE
5	6	1,20	238,14	6	1,04	227,14
10	6	1,50	268,11	6	1,03	257,20
15	6	1,20	294,57	8	1,34	289,49
20	7	1,20	451,68	8	1,06	449,74
25	3	1,50	334,28	3	1,40	333,90

Dari tabel di atas dapati dilihat bahwa nilai RMSE yang dihasilkan oleh metode hibrida FCM-GA pada masing-masing kelompok persentase *missing* seluruhnya lebih baik dibandingkan nilai RMSE yang diperoleh dengan FCM. Hal ini berarti proses optimasi yang dilakukan dengan metode GA berhasil meningkatkan kinerja FCM yang dibuktikan dengan nilai RMSE yang lebih kecil jika dibandingkan dengan nilai RMSE yang dihasilkan tanpa optimasi GA. Nilai RMSE terbaik menggunakan FCM dihasilkan oleh *weighting exponent* yang kecil. Demikian juga dengan hasil terbaik dengan menggunakan FCM-GA dihasilkan dari nilai *weighting exponent* yang kecil.

Tabel 4.15 RMSE Algoritma FCM dan FCM-GA Dengan Ukuran Jarak Manhattan

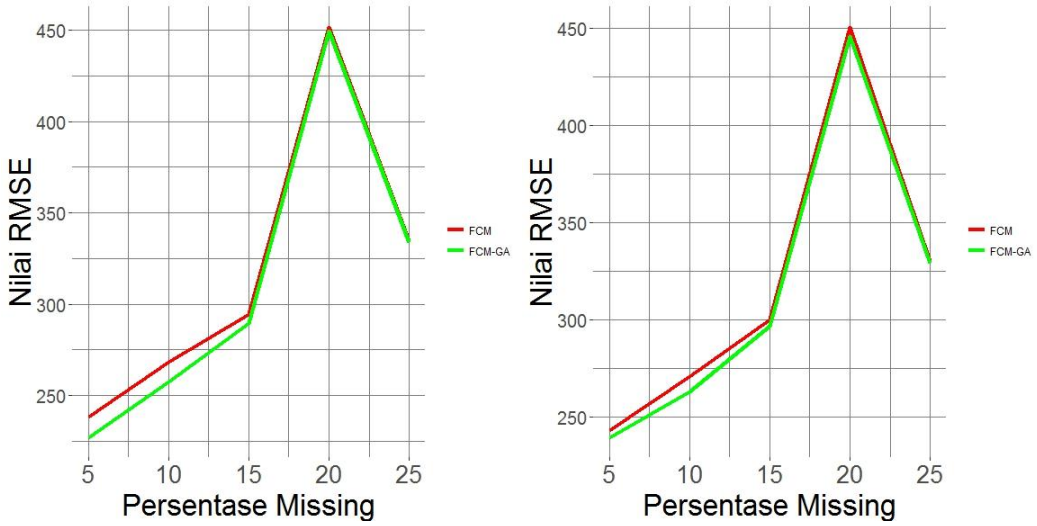
Persentase *Missing*	FCM			FCM-GA		
	c	***m***	RMSE	***c***	***m***	RMSE
5	9	1,20	243,39	9	1,09	239,82
10	6	1,20	271,24	8	1,05	263,36
15	5	1,50	300,33	6	1,38	296,95
20	9	1,70	450,54	3	1,03	445,79
25	3	1,20	329,97	3	1,11	329,21

Perbandingan nilai RMSE yang dihasilkan kedua metode menggunakan ukuran jarak Manhattan dapat dilihat pada tabel 4.15. Nilai RMSE yang dihasilkan dengan metode hibrida FCM-GA pada masing-masing kelompok *missing* seluruhnya juga lebih baik dibandingkan nilai RMSE yang dihasilkan dengan metode FCM. Dengan algoritma FCM hasil RMSE terbaik dihasilkan

dengan *weighting exponent* yang kecil. Sama halnya dengan FCM-GA dimana nilai RMSE terbaik diperoleh dengan nilai *weighting exponent* yang kecil.

Dari perbandingan kombinasi parameter yang memberikan hasil terbaik untuk kedua metode dapat dilihat ada pola fluktuasi peningkatan nilai RMSE yang sama dan nilai RMSE terbaik dihasilkan oleh nilai *weighting exponent* yang kecil. Dari Tabel 4.14 dapat dilihat untuk setiap kelompok persentase *missing* dari 5 persen sampai 25 persen, nilai m yang digunakan kedua metode untuk menghasilkan nilai RMSE terbaik $m < 2$. Pada Tabel 4.15 juga diamati terdapat pola yang serupa. Untuk setiap kelompok persentase *missing* 5 persen sampai 25 persen nilai m yang memberikan hasil terbaik adalah $m < 2$.

Untuk lebih jelas perbandingan kinerja FCM dan FCM-GA dapat dilihat melalui tampilan Gambar 4.7. Perbandingan RMSE yang dihasilkan oleh metode FCM (garis merah) dengan hasil optimasi (garis hijau) disajikan pada gambar tersebut. Untuk persentase *missing* yang berbeda-beda, hasil optimasi GA seluruhnya berhasil memberikan nilai yang lebih baik dibandingkan percobaan dengan metode FCM.

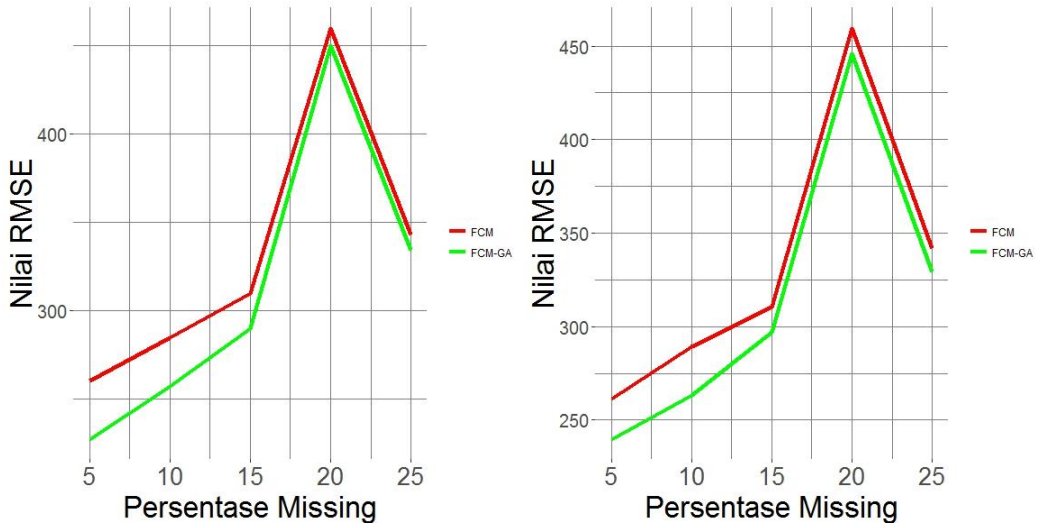


(i) Euclidean (ii) Manhattan

Gambar 4.7 Perbandingan RMSE FCM dan FCM-GA Dengan Persentase *Missing Value* Yang Berbeda-Beda

Jika dibandingkan hasil optimasi dibandingkan dengan nilai RMSE yang dihasilkan dari *weighting exponent* $m = 2$ yang biasa digunakan dalam FCM

konvensional, perbandingan kinerja optimasi dengan GA lebih jelas terlihat karena jarak menjadi lebih lebar. Seperti yang dapat diamati pada Gambar 4.8, RMSE hasil optimasi GA (garis hijau) seluruhnya berada di bawah hasil metode FCM (garis merah). Ini menunjukkan bahwa nilai RMSE yang dihasilkan dengan FCM-GA lebih kecil. Dapat disimpulkan bahwa optimasi GA berhasil meningkatkan kinerja algoritma FCM.



(i) Euclidean (ii) Manhattan

Gambar 4.8 Perbandingan RMSE FCM dan FCM-GA Dengan Persentase *Missing Value* Yang Berbeda-Beda (m = 2)

4.3 Hasil Imputasi Data Industri Sedang

Dari hasil percobaan pada data industri sedang, metode hibrida FCM-GA memiliki kinerja yang baik. Hal ini ditunjukkan dengan kemampuannya untuk meminimalkan RMSE. Jika nilai imputasi untuk setiap observasi diperhatikan, hasil imputasi untuk setiap variabel pada observasi yang *missing* relatif berbeda dari nilai aktualnya. Ada yang selisihnya cukup besar dan ada juga yang mendekati nilai aktualnya.

Tabel 4.16 menyajikan beberapa hasil imputasi data industri sedang pada kelompok persentase *missing* 25 persen. Dapat dilihat dari tabel tersebut bahwa nilai imputasi pada beberapa observasi berbeda dengan nilai aktual pada data

lengkap (bagian yang berwarna). Misalnya pada variabel X_2 pada observasi yang ke-2. Nilai imputasi yang dihasilkan adalah 104,06 sementara nilai aktual variabel tersebut sebesar 2,60. Namun ada hasil imputasi yang relatif mendekati nilai data aktualnya seperti pada nilai variabel X_3 pada observasi ke-521. Nilai imputasi yang dihasilkan adalah 61,73 tidak jauh berbeda dari nilai aktualnya yang sebesar 74,33. Hasil imputasi selengkapnya dapat dilihat pada bagian Lampiran.

Tabel 4.16 Hasil Imputasi Data Industri Sedang Pada Kelompok *Missing* 25 Persen

No. Obs	Data Aktual			Hasil Imputasi		
	X_1	X_2	X_3	X_1	X_2	X_3
2	62.91	2.60	3.50	62.91	104.06	61.73
3	61.31	19.79	64.25	66.55	104.06	64.25
7	114.00	85.20	17.15	66.55	85.20	17.15
10	127.90	189.81	179.59	127.90	104.06	179.59
11	127.90	189.81	179.59	66.59	189.81	61.78
12	127.90	189.81	179.59	66.64	104.16	179.59
14	213.16	316.36	299.31	566.94	316.36	918.28
15	321.40	300.96	386.95	321.40	300.96	918.28
16	321.40	300.96	386.95	566.94	300.96	386.95
19	896.22	222.82	856.02	125.75	222.82	137.65
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
521	41.25	191.41	74.33	41.25	104.06	61.73
523	717.94	138.20	37.54	717.94	138.20	918.15
525	132.80	48.72	1172.79	132.80	48.72	61.73
526	2.40	25.00	2.40	14.96	4.13	2.40
528	975.00	570.00	80.00	975.00	570.00	918.28
529	48.99	81.00	86.00	66.55	104.06	86.00
530	77.09	174.31	96.90	66.55	104.06	96.90
531	2.63	2.40	26.14	2.63	2.40	12.76
532	1.26	7.20	15.70	1.26	4.13	15.70
533	2.91	18.17	31.25	2.91	18.17	12.77

Selanjutnya, Tabel 4.17 menyajikan kesalahan yang terjadi dari hasil imputasi pada 1 variabel *missing* yang dihasilkan dari percobaan pada kelompok dengan persentase *missing* 25 persen.

Tabel 4.17 Hasil Imputasi Pada 1 Variabel *Missing* Data Industri Sedang Pada Kelompok *Missing* 25 Persen

No. Obs	Variabel *Missing*	Data Aktual	Hasil Imputasi	$\frac{\left\|\hat{x}_{ij} - x_{ij}\right\|}{x_{ij}}$
7	X_1	114,00	66,55	0,42
10	X_2	189,81	104,06	0,45
15	X_3	386,95	918,28	1,37
16	X_1	321,40	566,94	0,76
24	X_2	840,00	13,25	0,98
25	X_1	36,61	66,55	0,82
26	X_1	13,50	25,77	0,91
28	X_2	60,00	104,06	0,73
29	X_2	3,60	4,13	0,15
33	X_2	180,00	103,98	0,42
…	...	…	…	…
…	...	…	…	…
…	...	…	…	…
514	X_2	33,60	4,60	0,86
516	X_3	57,25	12,76	0,78
517	X_1	63,29	66,55	0,05
519	X_2	90,00	104,06	0,16
523	X_3	37,54	918,15	23,46
525	X_3	1.172,79	61,73	0,95
528	X_3	80,00	918,28	10,48
531	X_3	26,14	12,76	0,51
532	X_2	7,20	4,13	0,43
533	X_3	31,25	12,77	0,59

Kesalahan pada tabel di atas didapatkan dari nilai absolut selisih hasil imputasi dengan nilai aktual dibandingkan nilai aktual. Kesalahan yang beragam ini ditengarai terjadi karena tidak adanya pola korelasi yang kuat antar variabel. Sehingga nilai variabel pada tiap observasi tidak mengikuti suatu pola tertentu terkait dengan hubungannya terhadap variabel lain. Sebagai gambaran adalah observasi data aktual pada observasi ke-525 dan ke-528. Kedua observasi sama-sama mengalami *missing* pada 1 variabel yaitu variabel X_3. Nilai variabel X_1

dan X_2 pada kedua observasi tersebut berturut-turut adalah 132,80 dan 48,72 untuk observasi ke-525, kemudian 975,00 dan 570,00 untuk observasi ke-528. Namun nilai aktual pada variabel X_3 justru berbanding terbalik dimana pada observasi ke-525 adalah 1.172,79 sementara pada observasi ke-528 adalah 80,00.

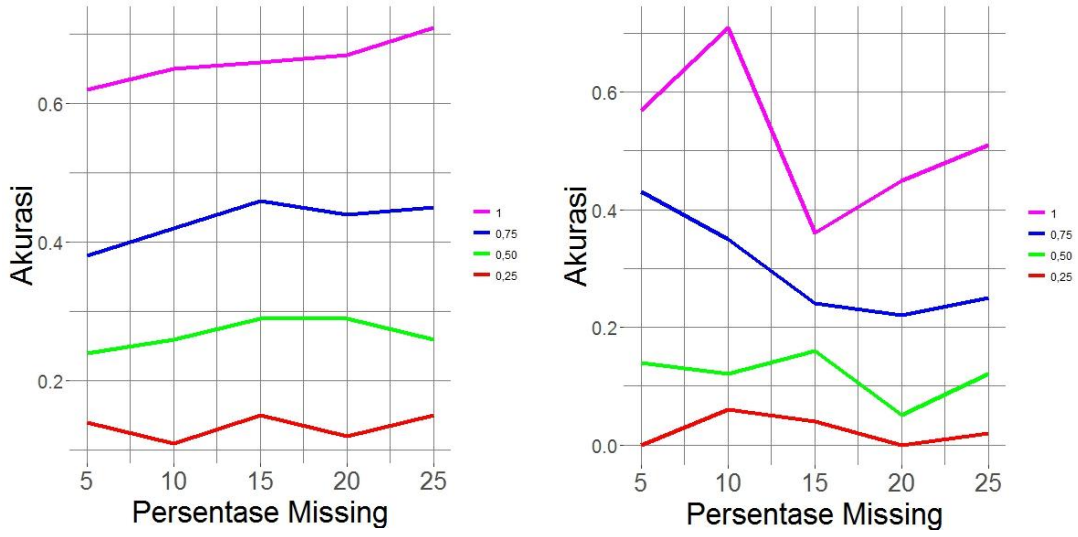
Berikutnya Tabel 4.18 menyajikan kesalahan dari hasil imputasi pada 2 variabel *missing* pada kelompok dengan persentase *missing* 25 persen.

Tabel 4.18 Hasil Imputasi Pada 2 Variabel *Missing* Data Industri Sedang Pada Kelompok *Missing* 25 Persen

No. Obs	Variabel *Missing*	Data Aktual		Hasil Imputasi		$\frac{\lvert\hat{x}_{ij} - x_{ij}\rvert}{x_{ij}}$	
		1	2	1	2	1	2
2	X_2, X_3	2,60	3,50	104,06	61,73	39,02	16,64
3	X_1, X_2	61,31	19,79	66,55	104,06	0,09	4,26
11	X_1, X_3	127,90	179,59	66,59	61,78	0,48	0,66
12	X_1, X_2	127,90	189,81	66,64	104,16	0,48	0,45
14	X_1, X_3	213,16	299,31	566,94	918,28	1,66	2,07
19	X_1, X_3	896,22	856,02	125,75	137,65	0,86	0,84
21	X_1, X_3	80,50	235,00	66,55	61,73	0,17	0,74
22	X_2, X_3	55,08	355,45	104,06	61,73	0,89	0,83
32	X_2, X_3	160,00	10,00	104,06	61,73	0,35	5,17
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
468	X_2, X_3	22,33	45,12	4,13	12,76	0,82	0,72
485	X_2, X_3	26,83	1143,70	170,35	144,75	5,35	0,87
488	X_2, X_3	402,64	53,63	495,37	916,54	0,23	16,09
494	X_2, X_3	14,00	37,00	4,13	12,76	0,70	0,66
498	X_1, X_3	163,60	39,50	66,55	61,73	0,59	0,56
518	X_1, X_3	52,08	196,00	66,55	61,73	0,28	0,69
521	X_2, X_3	191,41	74,33	104,06	61,73	0,46	0,17
526	X_1, X_3	2,40	25,00	14,96	4,13	5,23	0,83
529	X_1, X_3	48,99	81,00	66,55	104,06	0,36	0,28
530	X_1, X_3	77,09	174,31	66,55	104,06	0,14	0,40

Dapat dilihat pada tabel di atas kesalahan yang beragam terjadi seperti hasil imputasi pada 1 variabel *missing*. Tidak adanya pola korelasi yang kuat antar variabel menjadikan nilai variabel pada tiap observasi tidak mengikuti suatu pola tertentu terkait dengan hubungannya terhadap variabel lain. Pola data yang tidak menentu ini menjadi salah satu andil terhadap tingkat kesalahan yang terjadi. Dengan 2 variabel yang *missing*, informasi yang didapatkan menjadi sangat terbatas. Sebagai contoh dapat diamati pada observasi ke-2 dan ke-3. Nilai yang *missing* pada observasi ke-2 adalah variabel X_2 dan X_3. Sementara pada observasi ke-3 adalah variabel X_1 dan X_2. Informasi yang didapatkan dari observasi lain yang memiliki informasi yang bersesuaian mengindikasikan suatu nilai untuk variabel yang *missing* tersebut. Namun data aktual tidak mengikuti pola tersebut sehingga potensi kesalahan yang terjadi menjadi bervariasi.

Gambar 4.9 menampilkan perbandingan akurasi hasil imputasi untuk melihat sejauh mana jumlah observasi dengan kesalahan dalam level toleransi yang ditetapkan bervariasi terhadap persentase *missing* dan informasi yang tersedia.

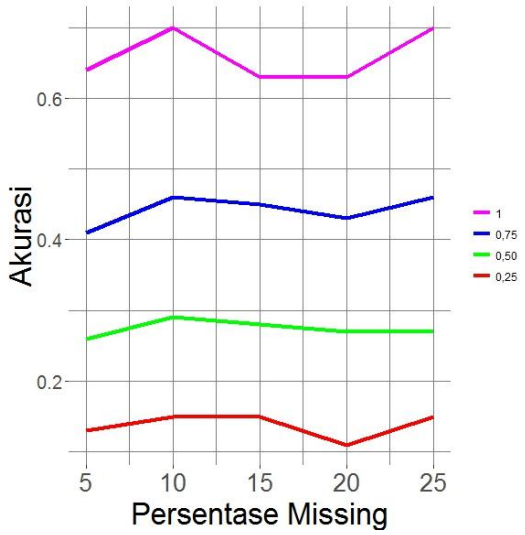


(a) (b)

Gambar 4.9 Perbandingan akurasi imputasi berdasarkan tingkat toleransi pada persentase *missing* data yang berbeda-beda (a) 1 variabel *missing*, (b) 2 variabel *missing*.

Dari gambar dapat dilihat bahwa semakin tinggi level toleransi yang ditetapkan maka tingkat akurasi akan semakin rendah. Ada indikasi tingkat akurasi pada 1 variabel *missing* memiliki kecenderungan meningkat seiring peningkatan persentase *missing*. Hal ini dipahami karena tingkat probabilitas juga semakin meningkat seiring semakin tinggi jumlah data yang akan diimputasi. Sebaliknya pada kelompok data dengan 2 variabel *missing* tidak ditemukan pola demikian. Walaupun ada kecenderungan untuk turun pada level toleransi yang lebih rendah tetapi akurasi berfluktuasi terhadap peningkatan persentase *missing* yang terjadi. Hal ini dipahami sebagai akibat berkurangnya informasi yang tersedia dan korelasi antar variabel yang lemah memiliki pengaruh yang kuat terhadap kualitas hasil imputasi.

Gambaran secara umum dapat dilihat pada Gambar 4.10, dimana peningkatan level toleransi akan menurunkan tingkat akurasi. Pengaruh persentase *missing* mengakibatkan tingkat akurasi yang beragam namun cenderung stabil pada tingkatan yang sama.



Gambar 4.10 Perbandingan akurasi imputasi berdasarkan tingkat toleransi pada persentase *missing* data yang berbeda-beda.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil penelitian yang telah dilakukan dapat diambil beberapa kesimpulan :

1a. Pilihan nilai *weighting exponent m* dan jumlah klaster *c* yang sesuai sangat berpengaruh terhadap nilai RMSE yang dihasilkan. Dengan kata lain kinerja imputasi FCM bergantung dari ketepatan pemilihan parameter *c* dan *m*.

b. Pemilihan ukuran jarak Euclidean dan Manhattan dibawah kombinasi c dan *m* tidak menunjukkan perbedaan besar. Kinerja dari dua ukuran jarak pada tingkat *missing* yang berbeda mengindikasikan tidak ada yang lebih unggul diantara keduanya.

c. Hasil penelitian menunjukkan bahwa GA lebih unggul pada persentase *missing* yang kecil. Pada tingkat *missing* data yang tinggi, kinerja GA tetap lebih baik namun perbedaan yang ada tidak terlalu signifikan.

2. Melalui optimasi GA mendapatkan informasi parameter *c* dan *m* yang mampu meminimumkan RMSE ketika diterapkan pada metode FCM untuk mendapatkan hasil imputasi yang lebih baik.

5.2 Saran

Berdasarkan penelitian yang dilakukan, terdapat beberapa saran untuk penelitian selanjutnya :

1. Mengkaji penerapan FCM-GA pada struktur data yang berbeda melalui simulasi atau data survei yang lain.
2. Mengkombinasikan dengan metode statistik lain yang mampu meningkatkan akurasi imputasi dengan metode FCM.
3. Menggunakan ukuran jarak selain Euclidean dan Manhattan.
4. Menerapkan metode optimalisasi selain GA.
5. Membuat perbandingan dengan metode imputasi atau optimalisasi yang lain.

DAFTAR PUSTAKA

BPS (2011), *Pedoman Pencacahan Survei Perusahaan Industri Manufaktur Skala Menengah Besar*, BPS, Jakarta.

BPS (2015), *Analisis Kondisi Sektor Industri Pengolahan (Manufacture) Sumatera Utara Tahun 2014*, BPS Provinsi Sumatera Utara, Medan.

Choir, A.S. (2011), *Imputasi Berganda K-Medoid General Regression Neural Network Untuk Menangani Missing Data*, Institut Teknologi Sepuluh Nopember, Surabaya.

Dempster, A.P., Laird, N.M. dan Rubin, D.B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm", *Journal of the Royal Statistical Society*, hal. 1-38.

Efron, B. (1992), *Missing Data, Imputation, and The Bootstrap*, Division of Biostatistics Stanford University, California.

Enders, C.K. (2010), *Applied Missing Data Analysis*, The Guilford Press, New York.

Engelbrecht, A.P. (2002), *Computational Inteligence : An Introduction*, John Wiley and Sons.

Gen, M. dan Cheng, R. (2000), *Genetic Algorithm and Optimization Engineering*, John Wiley and Sons.

Graham, J.W. (2012), *Statistics For Social and Behavioural Sciences*, Springer, New York.

Grittner, U., Gmel, G., Ripatti, S., Bloomfield, K. dan Wicki, M. (2011), "Missing Value Imputation in Longitudinal Measures of Alcohol Consumption", *International Journal of Methods in Psychiatric Research*, hal. 50-61.

Hair, J.F., Black, W.C., Babin, B.J. dan Anderson, R.E. (2010), *Multivariate Data Analysis 7th edition*, Prentice Hall.

Hartono, T. (2011), *Imputasi Menggunakan Fuzzy K-Means Dalam Penanganan Missing Data (Studi Kasus Data Perusahaan Industri Besar Provinsi Jawa Timur Tahun 2008)*, Institut Teknologi Sepuluh Nopember, Surabaya.

Hathaway, .J. dan Bezdek, J.C. (2001), "Fuzzy c-Means Clustering of Incomplete Data", *IEEE Transactions on Systems, Man dan Cybernetics-Part B : Cybernetics Vol. 31*, hal. 735-744.

Haupt, R.L. dan Haupt, S.E. (2004), *Practical Genetic Algorithms*, John Wiley and Sons.

King, G., Honaker, J., Joseph, A. dan Scheve, K. (2001), "Analyzing Incomplete Political Science Data : An Alternative Algorithm for Multiple Imputation", *American Political Science Review Vol. 95 No. 1*, hal. 49-69.

Li, D., Deogun, J., Spaulding, W. dan Shuart, B. (2004), "Towards Missing Data Imputation : A Study of Fuzzy K-Means Clustering Method", *International Conference on Rough Sets and Current Trends in Computing*, Springer Berlin Heidelberg, hal. 573-579.

Little, R.J.A. dan Rubin, D.B. (2002), *Statistical Analysis With Missing Data 2nd Edition*, Wiley and Sons, Inc., New Jersey.

Mawarsari, U. (2012), *Imputasi Missing Data Dengan K-Neares Neighbor dan Algoritma Genetika (Studi Kasus Data Survei Industri Besar dan Sedang 2008)*, Institut Teknologi Sepuluh Nopember, Surabaya.

Pal, N.R. dan Bezdek, J.C. (1995), "On Cluster Validity for the Fuzzy C-Means Model", *IEEE Transaction on Fuzzy System Vol.3, No.3*, hal. 370-379.

Piggot, T.D. (2001), "A Review of Methods of Missing Data", *Educational Research and Evaluation Vol. 7 No. 4*, hal. 353-383.

Schafer, J.L. dan Graham, J.W. (2002), "Missing Data ; Our View of the State of the Art", *Psychological Methods*, hal. 147-177.

Sukim (2011), *Studi Tentang Metode C-Means Cluster dan Fuzzy C-Means Cluster Serta Aplikasinya Pada Kasus Pengelompokkan Desa/Kelurahan Berdasarkan Status Ketertinggalan*, Institut Teknologi Sepuluh Nopember, Surabaya.

Tang, J., Zhang, G., Wang, Y., Wang, H. dan Liu, F. (2015), "A Hybrid Approach to Integrate Fuzzy C-Means Based Imputation Method With Genetic Algorithm for Missing Traffic Volume Data Estimation", *Transportation Research Part C Emerging Technologies*.

Tan, P.-N., Steinbach, M. dan Kumar, V. (2005), *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co.,Inc, Boston, MA.

Wardani, S. (2011), *Metode Jaringan Saraf Tiruan-Algoritma Genetika Untuk Imputasi (Studi Kasus : Survei Tahunan Industri Besar dan Sedang Provinsi Jawa Timur)*, Institut Teknologi Sepuluh Nopember, Surabaya.

Xie, X.L. dan Beni, G. (1991), "A Validity Measure for Fuzzy Clustering", *IEEE Trans. Pattern Anal. Machine Intell.*.

LAMPIRAN

Lampiran 1. Sintaks M-File Algoritma Imputasi Fuzzy C-Means

```
function [output] = fkmites(cdata,mdata,option)
% fungsi utama FKMI
% mdata : data yang akan diimputasi
% cdata : data lengkap
% cluster_n : jumlah kluster yg diinginkan
% expo : fuzifier (m)
% max_iter : maksimum iterasi
% min_impro : minimum improvisasi/batas konvergen
% distmet : pilihan fungsi jarak 1. euclidean 2. manhattan
cdata=cdata;
data=mdata;
cluster_n=option.cluster_n;
expo=option.expo;
max_iter=option.max_iter;
min_impro=option.min_impro;
distmet=option.distmet;
%*******************************************************
data_n = size(data, 1);
U = initfkmi(cluster_n, data_n); % Initial fuzzy partition
obj_fcn = zeros(max_iter, 1); % Array for objective function
y=[];
% Main loop
for i = 1:max_iter,
    [U, center,obj_fcn(i)] = stepfkmi(data, U, cluster_n,
expo,distmet);
    % check termination condition
    if i > 1
        if abs(obj_fcn(i) - obj_fcn(i-1)) < min_impro
           flag=1;
            break;
        else
            flag=2;
        end
    end

    x=U'*center;
end

% mengganti nilai yang missing
[m,n]=size(data);
inan= find(isnan(data));
nmis=sum(sum(isnan(data)));
[i,j] = ind2sub([m,n], inan);
xd=data;
for ii=1:numel(i);
    for jj=1:numel(j);
        ik=i(ii);
        jk=j(ii);
        xd(ik,jk)=x(ik,jk);
    end
end
```

Lampiran 1. (Lanjutan)

```
%menghitung rmse
idata=exp(xd);
err = (cdata-xd);
ss_err = sum(sum(err.^2));
mse = ss_err/(m*n);
rmse = sqrt(mse);


iter_n = i; % Actual number of iterations
obj_fcn(iter_n+1:max_iter) = [];

%output
output.center=center;
output.U=U;
output.obj_fcn=obj_fcn;
output.flag=flag;
output.xd=xd;
output.idata=idata;
output.rmse=rmse;


function U = initfkmi(cluster_n, data_n)
%fungsi untuk inisialisasi awal matriks partisi FKM secara acak
U = rand(cluster_n, data_n);
col_sum = sum(U);
U = U./col_sum(ones(cluster_n, 1), :);

function [U_new, center, obj_fcn] =
stepfkmi(data,U,cluster_n,expo,distmet)
%StepFCM One step in fuzzy c-mean clustering.
[m1 n]=size(data);
data0=noldata(data);
mf = U.^expo; % MF matrix after exponential modification
Ikj=inisnan(data,m1,n);
center = mf*(Ikj.*data0)./(mf*Ikj); % new center
dist = distfkmi(center,data,distmet,Ikj);
obj_fcn=objectfunction(data0,center,Ikj,mf,distmet);
tmp = dist.^(-2/(expo-1)); % calculate new U, suppose expo != 1
U_new = tmp./(ones(cluster_n, 1)*sum(tmp));

function data=noldata(data)
% fungsi untuk mengidentifikasi missing data
inan = find(isnan(data));
data(inan) = zeros(size(inan));
function I=inisnan(data,m,n);
% fungsi faktor skala pada missing data
I=ones(m,n);
inan = find(isnan(data));
I(inan) = zeros(size(inan));
function out = distfkmi(center,data,distmet,Ikj)
% fungsi jarak 1. euclidean 2. manhattan
% center : centroid kluster
% data : data yang akan diimputasi
% distmet : pilihan fungsi jarak 1. euclidean 2. manhattan
```

Lampiran 1. (Lanjutan)

```
% Ikj : faktor skala untuk missing data
data0=noldata(data);
[m n]=size(data);
[jj p]=size(center);
out = zeros(jj,m);
if distmet==1,
    for i = 1:jj,
        out(i, :) =sqrt(n./sum(Ikj').*(sum((((data0-
ones(m,1)*center(i,:)).^2).*Ikj)')));
    end
elseif distmet==2
    for i = 1:jj,
        out(i, :) = (n./sum(Ikj').*(sum(((abs(data0-
ones(m,1)*center(i,:))).*Ikj)')));
    end
end

function obj=objectfunction(data0,center,Ikj,mf,distmet);
% fungsi untuk menghitung objective function
[m n]=size(data0);
[jj p]=size(center);
Ik=sum(Ikj')';
kn=n./Ik;
if distmet==1,
    r=sum((((data0-ones(m,jj)*center).^2).*Ikj)');
    obj =sum(sum(mf*kn*r));
elseif distmet==2
    r=sum((((abs(data0-ones(m,jj)*center)).*Ikj))');
    obj =sum(sum(mf*kn*r));
  end
```

Contoh Sintaks *Command* yang digunakan

```
%ketik perintah berikut di jendela command atau sorot dan
tekan F9
%membuat data lengkap.mat
cdata=load('ibs.txt');
save('cdata.mat','cdata');
%data missing
mdata=load('ibs_m5.txt');
save('mdata.mat','mdata');
%cdata=cdata; %mendefinisikan data lengkap
%mdata=mdata;  %mendefinisikan data yg akan dimputasi
%mendefinisikan parameter FCM
option.cluster_n=2;
option.expo=2;
option.max_iter=10000;
option.min_impro=0.000001;
option.distmet=1;
%perintah untuk mengeksekusi fungsi dan mengeluarkan output
output= fkmites(cdata,mdata,option);
```

Lampiran 2. Sintaks M-File Fungsi *Fitness* dan Algoritma Genetika

```
function fitness=fga(cdata,mdata,par,option)
option.cluster_n=par(1);
option.expo=par(2);
[output] = fkmites(cdata,mdata,option);
fitness= -(1/(output.rmse+0.0000001));


function output=gafclust(cdata,mdata,option)
f=@(par) fga(cdata,mdata,par,option);
OptionsGA = gaoptimset('Display','iter',...
     'PopulationSize',50,...
     'Generations',50,...
     'StallGenLimit',5,...
     'TolFun', 1e-8,...
     'SelectionFcn',@selectionroulette,...
     'PlotFcns',@gaplotbestf);
%GA Description using: x =
ga(fitnessfcn,nvars,A,b,[],[],LB,UB,nonlcon,IntCon,options)

 par= ga(f,2,[],[],[],[],[1 1],[9 3],[],1,OptionsGA);
 option.cluster_n=par(1);
 option.expo=par(2);
 output=fkmites(cdata,mdata,option);
%output for GA optimization
 output.cluster_n = par(1);
 output.expo = par(2);

%The algorithm stops if max generation (Generations)reached, or
the average relative change in the best fitness function value
over Stall generations (StallGenLimit) is less than or equal to
Function tolerance (TolFun)
```

Contoh Sintaks *Command* yang digunakan

```
%ketik perintah berikut di jendela command atau sorot dan
tekan F9
%membuat data lengkap.mat
cdata=load('ibs.txt');
save('cdata.mat','cdata');
%data missing
mdata=load('ibs_m5.txt');
save('mdata.mat','mdata');
%cdata=cdata; %mendefinisikan data lengkap
%mdata=mdata;  %mendefinisikan data yg akan dimputasi
%mendefinisikan parameter FCM
option.max_iter=10000;
option.min_impro=0.000001;
option.distmet=1;
%mengeksekusi fungsi dan mengeluarkan output GA
output=gafclust(cdata,mdata,option);
```

Lampiran 3. Nilai RMSE Yang Dihasilkan Oleh Algoritma Imputasi FCM

m	*missing*	Euclidean							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	5%	265,28	246,62	249,52	246,71	238,14	238,96	246,43	238,54
1,5	5%	265,23	253,32	252,71	247,57	252,46	252,12	247,66	248,64
1,7	5%	266,01	258,12	257,57	254,90	259,62	259,73	253,36	254,24
2	5%	267,37	263,15	263,03	262,57	264,46	265,18	264,03	260,22
2,5	5%	268,98	267,09	267,20	267,05	267,06	267,90	267,41	267,74
3	5%	269,85	268,73	269,11	268,69	268,65	268,67	268,74	268,53

m	*missing*	Manhattan							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	5%	264,47	250,89	254,61	251,87	249,30	252,33	257,91	243,39
1,5	5%	264,41	255,13	255,22	253,81	252,09	257,71	256,93	249,91
1,7	5%	265,14	258,38	258,54	257,35	257,59	262,05	262,58	253,67
2	5%	266,54	262,38	262,74	262,79	263,78	265,58	264,42	261,24
2,5	5%	268,31	266,36	266,35	266,73	267,05	267,27	267,21	267,66
3	5%	269,33	268,36	268,06	268,20	268,41	268,48	268,44	268,60

m	*missing*	Euclidean							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	10%	296,53	269,21	273,16	269,33	268,11	272,42	280,85	282,19
1,5	10%	296,62	276,34	274,88	269,56	271,61	282,19	283,87	284,71
1,7	10%	297,98	282,46	280,81	277,86	285,82	286,68	284,22	284,58
2	10%	300,34	289,97	288,82	291,10	292,94	294,00	288,05	284,66
2,5	10%	303,11	297,38	296,53	297,18	297,86	298,60	297,61	298,03
3	10%	304,59	301,60	299,52	301,06	301,19	301,15	301,24	300,56

m	*missing*	Manhattan							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	10%	295,97	275,09	277,33	274,32	271,24	274,89	283,48	280,20
1,5	10%	295,86	280,39	278,63	275,95	274,05	283,11	281,52	282,49
1,7	10%	297,04	284,84	283,03	280,89	280,65	288,03	282,04	281,86
2	10%	299,28	290,59	289,24	289,26	290,49	293,97	291,44	291,67
2,5	10%	302,15	297,42	295,44	296,57	297,15	297,40	297,41	298,13
3	10%	303,80	301,37	298,90	299,84	300,10	300,55	300,12	300,39

Lampiran 3.(Lanjutan)

m	*missing*	Euclidean							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	15%	319,23	300,30	303,56	309,75	294,57	296,12	305,78	310,91
1,5	15%	320,49	303,42	303,14	306,18	304,68	303,57	303,91	301,32
1,7	15%	322,40	307,80	306,15	304,74	302,02	305,61	305,88	308,63
2	15%	325,13	314,58	311,60	309,68	310,65	313,30	315,48	315,62
2,5	15%	328,20	322,98	319,39	317,56	318,40	319,88	320,28	320,70
3	15%	329,89	327,35	324,85	322,56	322,12	322,47	322,93	323,36

m	*missing*	Manhattan							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	15%	318,59	305,55	307,47	316,35	318,30	309,27	313,32	311,97
1,5	15%	319,49	306,73	304,89	300,33	308,26	305,51	304,14	305,81
1,7	15%	321,25	309,66	307,34	310,24	303,14	308,48	312,00	311,95
2	15%	323,97	315,20	312,08	310,26	310,24	314,19	314,82	315,57
2,5	15%	327,22	322,56	319,46	318,27	317,06	317,81	319,39	322,50
3	15%	329,10	326,61	324,53	322,55	320,54	321,27	322,59	323,01

m	*missing*	Euclidean							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	20%	464,63	454,34	459,33	457,34	454,16	451,68	456,62	457,16
1,5	20%	464,78	453,10	454,63	457,88	455,15	453,66	458,55	465,14
1,7	20%	465,76	455,00	455,48	459,42	458,05	457,39	463,12	464,42
2	20%	467,43	459,66	459,78	463,05	462,34	462,10	462,31	463,82
2,5	20%	469,63	465,47	465,50	467,21	465,83	464,12	465,67	468,34
3	20%	471,02	468,77	468,56	468,72	469,01	467,58	467,27	468,04

m	*missing*	Manhattan							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	20%	463,04	454,39	457,24	458,57	456,61	456,54	462,83	464,24
1,5	20%	463,60	454,92	454,88	457,71	455,15	455,64	453,58	454,01
1,7	20%	464,58	456,63	456,05	458,49	457,22	458,10	456,58	450,54
2	20%	466,29	460,41	459,63	461,63	460,61	460,88	461,37	462,83
2,5	20%	468,66	465,51	464,99	465,36	466,28	463,27	464,88	468,11
3	20%	470,22	468,47	468,17	468,01	467,26	466,80	467,37	467,61

Lampiran 3. (Lanjutan)

m	*missing*	Euclidean							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	25%	348,62	334,76	342,05	348,79	342,58	343,94	366,15	366,65
1,5	25%	349,36	334,28	340,89	341,30	340,02	352,26	352,66	362,17
1,7	25%	351,06	336,96	341,23	342,15	339,66	349,78	350,14	360,66
2	25%	353,70	342,74	344,92	346,28	343,62	344,52	350,65	359,79
2,5	25%	356,95	349,80	350,75	352,07	350,89	349,32	348,19	350,81
3	25%	358,93	354,00	354,50	355,55	354,36	352,32	351,50	352,41

m	*missing*	Manhattan							
		c=2	*c*=3	*c*=4	*c*=5	*c*=6	*c*=7	*c*=8	*c*=9
1,2	25%	347,03	329,97	334,75	340,04	347,77	373,58	337,60	349,57
1,5	25%	347,84	333,01	338,84	341,11	339,48	337,48	337,81	355,98
1,7	25%	349,46	336,13	339,57	341,39	338,95	342,09	334,29	358,81
2	25%	352,10	341,98	342,51	345,03	345,60	345,23	346,69	354,21
2,5	25%	355,50	349,14	349,38	350,87	349,78	348,81	352,66	351,93
3	25%	357,70	353,43	353,61	353,93	353,37	351,74	351,54	351,87

Lampiran 4. Nilai RMSE Yang Dihasilkan Oleh Optimalisasi GA (5 kali ulangan untuk setiap kelompok *missing data*)

missing	Euclidean		
	c	*m*	rmse
5%	6	1,08	240,90
	6	1,04	234,42
	6	1,05	251,18
	6	1,04	236,22
	6	1,04	227,14
10%	6	1,03	258,37
	6	1,03	271,54
	6	1,03	257,20
	6	1,04	257,84
	6	1,03	260,08
15%	3	1,21	300,29
	8	1,42	293,76
	6	1,22	308,03
	8	1,34	289,49
	3	1,21	300,29
20%	4	1,07	460,84
	8	1,42	463,75
	6	1,03	459,47
	4	1,19	457,01
	8	1,06	449,74
25%	3	1,40	333,90
	3	1,40	333,90
	3	1,40	333,90
	3	1,40	333,90
	3	1,40	333,90

missing	Manhattan		
	c	*m*	rmse
5%	9	1,09	249,09
	8	1,07	260,93
	9	1,08	255,82
	9	1,09	239,82
	9	1,09	247,29
10%	6	1,05	274,69
	8	1,03	278,49
	6	1,03	274,85
	6	1,03	274,40
	8	1,05	263,36
15%	6	1,34	310,94
	6	1,30	314,32
	6	1,38	309,54
	6	1,40	305,32
	6	1,38	296,95
20%	3	1,03	445,79
	3	1,05	447,15
	3	1,04	446,44
	3	1,03	446,16
	8	1,36	452,63
25%	3	1,12	329,21
	3	1,12	329,21
	3	1,11	329,21
	3	1,11	329,21
	3	1,11	329,21

Lampiran 5. Data Lengkap Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur Sumatera Utara 2013 (juta rupiah)

No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain
1	20,64	2,38	20,98
2	62,91	2,60	3,50
3	61,31	19,79	64,25
4	86,61	260,00	95,00
5	15,75	156,00	6,00
6	650,00	900,00	17,00
7	114,00	85,20	17,15
8	193,99	142,07	90,99
9	21,47	196,77	101,51
10	127,90	189,81	179,59
11	127,90	189,81	179,59
12	127,90	189,81	179,59
13	63,00	137,00	184,48
14	213,16	316,36	299,31
15	321,40	300,96	386,95
16	321,40	300,96	386,95
17	341,06	506,17	478,90
18	59,71	636,20	1367,30
19	896,22	222,82	856,02
20	195,50	157,59	1365,91
21	80,50	108,00	235,00
22	90,52	55,08	355,45
23	508,80	14,40	15,00
24	116,00	840,00	8,00
25	36,61	295,50	28,46
26	13,50	14,00	40,00
27	169,11	158,36	203,60
28	70,00	60,00	255,00
29	21,54	3,60	6,15
30	82,97	253,03	42,48
31	20,41	376,42	18,76
32	129,11	160,00	10,00
33	297,70	180,00	27,00
34	1269,86	1707,30	230,00
35	16,19	1,56	2,00

No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain
…	….	….	….
…	….	….	….
…	….	….	….
508	33,12	33,63	11,53
509	3,70	4,20	14,00
510	104,00	360,00	23,50
511	54,85	25,00	23,60
512	4,85	12,00	26,50
513	72,37	63,73	42,73
514	2,76	33,60	43,40
515	71,96	331,94	56,99
516	21,13	6,85	57,25
517	63,29	24,00	142,30
518	52,08	96,00	196,00
519	83,09	90,00	203,70
520	428,38	62,49	314,62
521	41,25	191,41	74,33
522	243,08	197,79	155,83
523	717,94	138,20	37,54
524	16,50	271,37	73,60
525	132,80	48,72	1172,79
526	2,40	25,00	2,40
527	23,75	2,65	4,00
528	975,00	570,00	80,00
529	48,99	81,00	86,00
530	77,09	174,31	96,90
531	2,63	2,40	26,14
532	1,26	7,20	15,70
533	2,91	18,17	31,25
534	340,07	70,00	89,19

Sumber : Data diolah dari hasil Survei Tahunan Perusahaan Industri Manufaktur 2013 Sumatera Utara

Lampiran 6. Hasil Imputasi Simulasi *Missing* 5 Persen Data Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur Sumatera Utara 2013 (juta rupiah)

Aktual				Imputasi			
No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain	No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain
25	36,61	295,50	28,46	25	83,47	295,50	43,27
29	21,54	3,60	6,15	29	21,54	3,60	25,79
37	8,90	1,50	2,62	37	8,90	1,36	2,62
41	30,50	1,45	4,26	41	5,94	1,45	4,26
44	3,40	1,52	4,94	44	5,94	1,52	4,94
53	3,82	0,36	13,00	53	3,82	0,36	7,68
55	15,90	2,60	13,66	55	15,90	35,07	13,66
56	10,03	1,20	14,90	56	10,03	35,07	14,90
61	10,20	1,45	17,50	61	5,94	1,45	7,68
84	8.691,12	3.533,63	195,70	84	978,37	3.533,63	195,70
88	2.099,95	3.667,00	660,00	88	1.027,45	3.667,00	660,00
102	1.285,53	118,12	2.032,76	102	1.021,87	118,12	2.032,76
104	3.510,98	2.101,84	3.305,61	104	3.510,98	2.101,84	1.395,51
109	1,36	8,40	1,00	109	1,36	8,40	7,93
133	74,50	75,00	37,50	133	10,56	75,00	37,50
137	21,43	15,22	46,71	137	83,47	327,48	46,71
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
458	492,65	70,00	92,00	458	492,65	70,00	313,66
459	318,50	138,00	8.969,56	459	1.027,45	138,00	8.969,54
471	83,86	81,82	33,64	471	83,86	81,82	313,66
472	693,16	108,00	21,00	472	693,16	108,00	697,27
481	142,65	111,00	7,00	481	10,56	111,00	7,00
495	80,08	1.158,10	194,00	495	80,08	1.158,10	43,27
497	30,70	250,00	400,00	497	30,70	250,00	43,27
505	6,15	1,75	3,64	505	5,94	1,75	3,64
506	7,50	4,01	8,96	506	7,50	4,01	7,68
509	3,70	4,20	14,00	509	3,70	4,20	7,68
515	71,96	331,94	56,99	515	71,96	331,94	43,27
519	83,09	90,00	203,70	519	83,09	82,15	203,70
520	428,38	62,49	314,62	520	116,87	62,49	314,62
524	16,50	271,37	73,60	524	16,50	271,37	43,27
531	2,63	2,40	26,14	531	83,11	2,40	26,14
533	2,91	18,17	31,25	533	2,91	18,17	22,49

Lampiran 7. Hasil Imputasi Simulasi *Missing* 10 Persen Data Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur Sumatera Utara 2013 (juta rupiah).

Aktual				Imputasi			
No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain	No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain
2	62,91	2,60	3,50	2	7,87	2,60	3,50
18	59,71	636,20	1.367,30	18	790,21	636,20	1.367,30
24	116,00	840,00	8,00	24	116,00	174,55	8,00
25	36,61	295,50	28,46	25	36,61	295,50	139,50
29	21,54	3,60	6,15	29	21,54	3,60	7,13
37	8,90	1,50	2,62	37	7,87	1,22	2,62
41	30,50	1,45	4,26	41	7,87	1,45	4,26
43	3,33	1,35	4,78	43	7,87	1,35	4,78
44	3,40	1,52	4,94	44	7,87	1,52	4,94
46	25,49	3,00	6,42	46	25,49	1,89	6,42
50	7,49	1,86	11,58	50	7,49	1,23	11,58
53	3,82	0,36	13,00	53	3,82	1,22	7,12
55	15,90	2,60	13,66	55	15,90	24,78	13,66
56	10,03	1,20	14,90	56	10,03	24,78	14,90
58	5,52	1,08	16,22	58	5,52	1,08	7,12
61	10,20	1,45	17,50	61	7,87	1,45	7,12
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
497	30,70	250,00	400,00	497	30,70	250,00	139,50
501	37,50	388,74	97,85	501	37,50	200,03	97,85
502	15,10	33,04	9,87	502	15,10	33,04	16,05
504	23,25	4,15	19,60	504	104,03	4,15	19,60
505	6,15	1,75	3,64	505	7,87	1,22	3,64
506	7,50	4,01	8,96	506	7,50	4,01	7,12
509	3,70	4,20	14,00	509	3,70	4,20	7,12
515	71,96	331,94	56,99	515	71,96	331,94	139,50
519	83,09	90,00	203,70	519	83,09	200,03	203,70
520	428,38	62,49	314,62	520	36,61	62,49	314,62
522	243,08	197,79	155,83	522	243,08	6,94	155,83
524	16,50	271,37	73,60	524	16,50	271,37	139,50
527	23,75	2,65	4,00	527	23,75	2,65	7,12
530	77,09	174,31	96,90	530	36,61	174,31	96,90
531	2,63	2,40	26,14	531	101,88	2,40	26,14
533	2,91	18,17	31,25	533	2,91	18,17	16,05

Lampiran 8. Hasil Imputasi Simulasi *Missing* 15 Persen Data Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur Sumatera Utara 2013 (juta rupiah).

Aktual				Imputasi			
No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain	No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain
1	20,64	2,38	20,98	1	20,64	24,06	20,98
7	114,00	85,20	17,15	7	80,35	85,20	17,15
8	193,99	142,07	90,99	8	26,92	142,07	90,99
11	127,90	189,81	179,59	11	127,90	46,89	179,59
15	321,40	300,96	386,95	15	747,11	905,54	386,95
16	321,40	300,96	386,95	16	321,40	154,84	386,95
17	341,06	506,17	478,90	17	341,06	506,17	29,45
22	90,52	55,08	355,45	22	90,06	55,08	355,45
24	116,00	840,00	8,00	24	243,84	840,00	8,00
25	36,61	295,50	28,46	25	164,21	295,50	32,23
26	13,50	14,00	40,00	26	11,56	14,00	40,00
28	70,00	60,00	255,00	28	90,00	60,00	255,00
30	82,97	253,03	42,48	30	32,30	64,47	42,48
32	129,11	160,00	10,00	32	268,73	160,00	10,00
35	16,19	1,56	2,00	35	7,46	1,56	2,00
44	3,40	1,52	4,94	44	3,40	1,28	4,94
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
496	40,71	123,56	483,87	496	40,71	123,56	109,91
498	163,60	174,86	39,50	498	163,60	174,86	29,81
501	37,50	388,74	97,85	501	22,85	388,74	97,85
503	19,07	138,00	16,00	503	19,07	326,61	83,31
504	23,25	4,15	19,60	504	23,25	327,41	83,40
505	6,15	1,75	3,64	505	6,15	1,75	7,96
511	54,85	25,00	23,60	511	54,85	25,00	181,85
513	72,37	63,73	42,73	513	72,37	20,01	69,38
520	428,38	62,49	314,62	520	90,02	62,49	314,62
521	41,25	191,41	74,33	521	23,62	191,41	74,33
522	243,08	197,79	155,83	522	243,08	51,14	155,83
526	2,40	25,00	2,40	526	2,40	25,00	25,95
528	975,00	570,00	80,00	528	975,00	304,85	80,00
530	77,09	174,31	96,90	530	77,09	174,31	92,10
533	2,91	18,17	31,25	533	2,91	18,17	25,72
534	340,07	70,00	89,19	534	340,07	157,01	89,19

Lampiran 9. Hasil Imputasi Simulasi *Missing* 20 Persen Data Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur Sumatera Utara 2013 (juta rupiah).

Aktual				Imputasi			
No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain	No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain
1	20,64	2,38	20,98	1	20,64	4,38	20,98
4	86,61	260,00	95,00	4	626,14	260,00	757,16
14	213,16	316,36	299,31	14	213,16	354,75	757,16
17	341,06	506,17	478,90	17	626,14	506,17	757,16
18	59,71	636,20	1.367,30	18	626,14	354,75	1.367,30
20	195,50	157,59	1.365,91	20	195,50	351,98	742,37
21	80,50	108,00	235,00	21	80,50	124,94	235,00
23	508,80	14,40	15,00	23	508,80	77,92	15,00
26	13,50	14,00	40,00	26	13,50	14,00	11,83
27	169,11	158,36	203,60	27	169,11	125,10	203,60
32	129,11	160,00	10,00	32	56,73	160,00	10,00
34	1.269,86	1.707,30	230,00	34	1.269,86	1.707,30	757,16
37	8,90	1,50	2,62	37	8,90	1,50	11,83
38	21,75	2,24	2,72	38	18,29	2,24	2,72
39	10,50	0,61	2,90	39	10,50	0,61	11,83
41	30,50	1,45	4,26	41	30,50	4,38	4,26
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
498	163,60	174,86	39,50	498	163,60	174,86	54,46
499	57,04	85,26	142,80	499	57,04	124,94	142,80
501	37,50	388,74	97,85	501	37,50	388,74	54,46
504	23,25	4,15	19,60	504	23,25	4,15	11,83
510	104,00	360,00	23,50	510	626,14	360,00	757,16
511	54,85	25,00	23,60	511	54,85	25,00	54,46
512	4,85	12,00	26,50	512	18,29	12,00	26,50
519	83,09	90,00	203,70	519	83,09	124,94	203,70
523	717,94	138,20	37,54	523	56,73	138,20	37,54
526	2,40	25,00	2,40	526	18,29	25,00	2,40
527	23,75	2,65	4,00	527	23,75	4,38	4,00
528	975,00	570,00	80,00	528	975,00	354,75	80,00
531	2,63	2,40	26,14	531	56,59	124,06	26,14
532	1,26	7,20	15,70	532	1,26	4,38	11,83
533	2,91	18,17	31,25	533	2,91	4,38	11,83
534	340,07	70,00	89,19	534	340,07	70,00	740,42

Lampiran 10. Hasil Imputasi Simulasi *Missing* 25 Persen Data Industri Sedang Hasil Survei Tahunan Perusahaan Industri Manufaktur Sumatera Utara 2013 (juta rupiah).

Aktual				Imputasi			
No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain	No.	Bahan Bakar & Pelumas	Listrik Yang Dibeli	Pengeluaran Lain
2	62,91	2,60	3,50	2	62,91	104,06	61,73
3	61,31	19,79	64,25	3	66,55	104,06	64,25
7	114,00	85,20	17,15	7	66,55	85,20	17,15
10	127,90	189,81	179,59	10	127,90	104,06	179,59
11	127,90	189,81	179,59	11	66,59	189,81	61,78
12	127,90	189,81	179,59	12	66,64	104,16	179,59
14	213,16	316,36	299,31	14	566,94	316,36	918,28
15	321,40	300,96	386,95	15	321,40	300,96	918,28
16	321,40	300,96	386,95	16	566,94	300,96	386,95
19	896,22	222,82	856,02	19	125,75	222,82	137,65
21	80,50	108,00	235,00	21	66,55	108,00	61,73
22	90,52	55,08	355,45	22	90,52	104,06	61,73
24	116,00	840,00	8,00	24	116,00	13,25	8,00
25	36,61	295,50	28,46	25	66,55	295,50	28,46
26	13,50	14,00	40,00	26	25,77	14,00	40,00
28	70,00	60,00	255,00	28	70,00	104,06	255,00
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
…	….	….	….	…	….	….	….
511	54,85	25,00	23,60	511	54,85	104,01	23,60
514	2,76	33,60	43,40	514	2,76	4,60	43,40
516	21,13	6,85	57,25	516	21,13	6,85	12,76
517	63,29	24,00	142,30	517	66,55	24,00	142,30
518	52,08	96,00	196,00	518	66,55	96,00	61,73
519	83,09	90,00	203,70	519	83,09	104,06	203,70
521	41,25	191,41	74,33	521	41,25	104,06	61,73
523	717,94	138,20	37,54	523	717,94	138,20	918,15
525	132,80	48,72	1.172,79	525	132,80	48,72	61,73
526	2,40	25,00	2,40	526	14,96	4,13	2,40
528	975,00	570,00	80,00	528	975,00	570,00	918,28
529	48,99	81,00	86,00	529	66,55	104,06	86,00
530	77,09	174,31	96,90	530	66,55	104,06	96,90
531	2,63	2,40	26,14	531	2,63	2,40	12,76
532	1,26	7,20	15,70	532	1,26	4,13	15,70
533	2,91	18,17	31,25	533	2,91	18,17	12,77

BIOGRAFI PENULIS

Ervin Noderius Mei Bunawolo dilahirkan di Kabupaten Nias Selatan, Sumatera Utara pada 27 Mei 1980. Anak kedua dari tiga bersaudara dari pasangan orang tua Sanehaoni Bunawolo dan Syamsiar Bidaya. Menyelesaikan pendidikan formal dari SD hingga SMU di Kota Gunungsitoli, Nias. Dari tahun 1999-2003 menempuh pendidikan ikatan dinas di Sekolah Tinggi Ilmu Statistik (STIS) di Jakarta.

Terhitung mulai tahun 2003 diangkat sebagai Calon Pegawai Negeri Sipil di Badan Pusat Statistik (BPS). Penempatan awal ditugaskan sebagai staf di BPS Kabupaten Nias pada tahun 2004. Kemudian tahun 2006 dimutasi ke BPS Kabupaten Nias Selatan sebagai pelaksana tugas Seksi Neraca Wilayah dan Analisis Statistik. Kemudian bertugas sebagai Kepala Seksi Integrasi Pengolahan dan Diseminasi Statistik (IPDS) di BPS Kota Binjai tahun 2010. Tahun 2015 mendapatkan kesempatan untuk melanjutkan pendidikan pasca sarjana di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember (ITS) Surabaya.