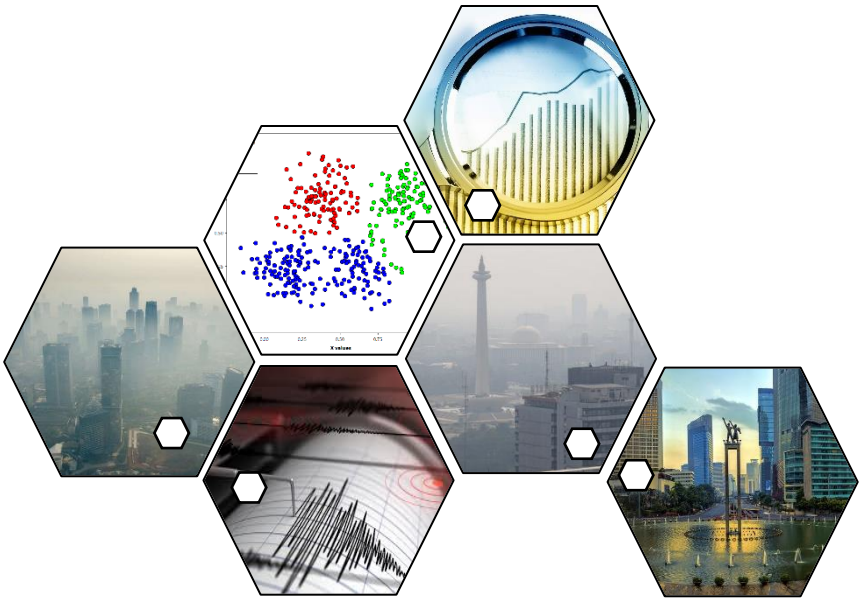
EVALUASI MODEL *HYBRID CLUSTERING LARGE APPLICATIONS* DAN *FUZZY TIMESERIES MARKOV CHAIN* PADA PERAMALAN POLUSI UDARA *PARTICULAR MATTER* DI KOTA JAKARTA

EVALUATION OF HYBRID CLUSTERING LARGE APPLICATIONS AND FUZZY TIMESERIES MARKOV CHAIN MODEL ON FORECASTING AIR POLLUTION PARTICULAR MATTER IN JAKARTA CITY



ANKAZ AS SIKIB

H062231015

PROGRAM STUDI MAGISTER STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR

2024

EVALUASI MODEL *HYBRID CLUSTERING LARGE APPLICATIONS* DAN *FUZZY TIMESERIES MARKOV CHAIN* PADA PERAMALAN POLUSI UDARA *PARTICULAR MATTER* DI KOTA JAKARTA

ANKAZ AS SIKIB

H062231015

PROGRAM STUDI MAGISTER STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2024

EVALUASI MODEL *HYBRID CLUSTERING LARGE APPLICATIONS* DAN *FUZZY TIMESERIES MARKOV CHAIN* PADA PERAMALAN POLUSI UDARA *PARTICULAR MATTER* DI KOTA JAKARTA

Tesis

sebagai salah satu syarat untuk mencapai gelar magister

Program Studi Magister Statistika

Disusun dan diajukan oleh

ANKAZ AS SIKIB
H062231015

kepada

PROGRAM STUDI MAGISTER STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024

TESIS

EVALUASI MODEL *HYBRID CLUSTERING LARGE APPLICATIONS* DAN *FUZZY TIMESERIES MARKOV CHAIN* PADA PERAMALAN POLUSI UDARA *PARTICULAR MATTER* DI KOTA JAKARTA

ANKAZ AS SIKIB

H062231015

Telah dipertahankan di hadapan Panitia Ujian Magister pada 20 November 2024 dan dinyatakan telah memenuhi syarat kelulusan

pada

Program Studi Magister Statistika
Departemen Statisika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Hasanuddin
Makassar

Mengesahkan:

Pembimbing Utama

Prof. Dr. Dr. Georgina Maria Tinungki, M.Si.
NIP. 19620926 198702 2 001

Pembimbing Pendamping

Prof. Dr. Nurtiti Sunusi, S.Si., M.Si.
NIP. 19720117 199703 2 002

Ketua Program Studi
Magister Statistika

Dr. Erna Tri Herdiani, S.Si., M.Si.
NIP. 19750429 200003 2 001

Dekan Fakultas Matematika dan
Ilmu Pengetahuan Alam
Universitas Hasanuddin

Dr. Eng. Amiruddin, M.Si.
NIP. 19720515 199702 1 002

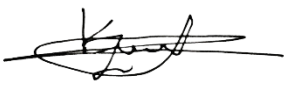
PERNYATAAN KEASLIAN TESIS
DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa, tesis berjudul "Evaluasi Model *Hybrid Clustering Large Applications* dan *Fuzzy Timeseries Markov Chain* pada Peramalan Polusi Udara *Particular Matter* di Kota Jakarta" adalah benar karya saya dengan arahan tim pembimbing Prof. Dr. Dr. Georgina Maria Tinungki, M.Si. sebagai Pembimbing Utama dan Prof. Dr. Nurtiti Sunusi, S.Si., M.Si. sebagai Pembimbing Pendamping. Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka tesis ini. Sebagian dari isi tesis ini telah dipublikasikan di AIMS Environmental Science dengan judul "*Evaluation of Hybrid Clustering Large Applications and Fuzzy Time Series Markov Chain Model on Particular Matter Air Pollution Forecasting in Jakarta City*".

Dengan ini saya melimpahkan hak cipta (hak ekonomis) dari karya tulis saya berupa tesis ini kepada Universitas Hasanuddin.

Makassar, 20 November 2024

Yang menyatakan,



Ankaz As Sikib
NIM.H062231015

UCAPAN TERIMA KASIH

Alhamdulillah, segala puji atas kehadirat Allah SWT, atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyusun dan menyelesaikan tesis ini. Penulis menyadari bahwa tesis ini masih jauh dari kesempurnaan. Namun, segala sesuatu yang tercantum dalam tesis ini merupakan usaha terbaik penulis dalam menysusun tesis ini.

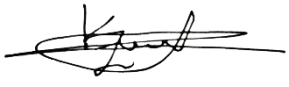
Penulis percaya, tesis ini dapat selesai bukan hanya dengan kekuatan pikiran penulis semata akan tetapi karena bantuan dari berbagai pihak, baik selama proses perkuliahan hingga proses pengerjaan tesis di Program Magister Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin. Terima kasih yang tak terhingga kepada kedua orang tua tercinta Bapak Khoiruddin dan Ibu Saltini Thohir, juga kepada adik saya Serly Al-Fajria dan Riyadh Ar-Rizqi atas doa yang tak pernah putus, dukungan serta kasih sayang yang tiada henti. Selanjutnya, saya ingin menyampaikan juga rasa hormat dan terima kasih kepada:

1. Prof. Dr. Ir. Jamaluddin Jompa, M.Sc. selaku Rektor Universitas Hasanuddin.
2. Dr. Eng. Amiruddin, M.Si. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh jajarannya.
3. Dr. Erna Tri Herdiani, S.Si., M.Si. Ketua Program Studi Magister Statistika Departemen Statistika Fakultas MIPA Universitas Hasanuddin yang menjadi salah satu tim penguji yang telah memberikan arahan dan dukungan dalam penyelesaian tesis ini.
4. Prof. Dr. Dr. Georgina Maria Tinungki, M.Si. selaku Pembimbing Utama yang dengan tulus ikhlas memberikan ilmu pengetahuan dan pengalaman yang dimilikinya serta meluangkan banyak waktunya dalam membimbing, memberikan motivasi, serta masukan sehingga memberikan banyak manfaat bagi penulis dalam menyelesaikan tesis ini maupun di masa mendatang.
5. Prof. Dr. Nurtiti Sunusi, S.Si., M.Si. selaku Pembimbing Pendamping yang dengan tulus ikhlas memberikan ilmu pengetahuan dan pengalaman yang dimilikinya serta meluangkan banyak waktunya dalam membimbing, memberikan motivasi, serta masukan sehingga memberikan banyak manfaat bagi penulis dalam menyelesaikan tesis ini maupun di masa mendatang.

6. Dr. Anna Islamiyati, S.Si., M.Si. selaku Ketua Program Studi Statistika Statistika Fakultas MIPA Universitas Hasanuddin yang menjadi salah satu tim penguji yang telah memberikan arahan dan dukungan dalam penyelesaian tesis ini.
7. Dr. Nirwan, M.Si. selaku salah satu tim penguji yang telah bersedia memberikan arahan dan dukungan dalam penyelesaian tesis ini.
8. Bapak dan Ibu Dosen serta Staf Departemen Statistika Fakultas MIPA Universitas Hasanuddin, yang dengan tulus ikhlas memberikan ilmu pengetahuan dan pengalaman yang dimilikinya sehingga memberikan banyak manfaat bagi penulis untuk saat ini maupun di masa mendatang.
9. Dr. Ruliana S.Pd., M.Si. selaku Ketua Program Studi Statistika, FMIPA Universitas Negeri Makassar. Sekaligus Penasehat Akademik penulis selama menjalani perkuliahan di Universitas Negeri Makassar.
10. Aswi, S.Pd., M.Si., Ph.D. Selaku penasehat akademik II penulis semasa menjalani perkuliahan di Universitas Negeri Makassar.
11. Seluruh teman-teman Mahasiswa Program Studi Magister Statistika, Delta-19, Santuystika, Secret Garden, HMPS Statistika FMIPA UNM dan LPM Penalaran UNM. Terima kasih telah menjadi bagian dari perjalanan penulis.
12. Franz Magnis Suseno, Edward De Bono, James Clear, Henry Manampiring, Gita Wirjawan dan Ferry Irwandi. Terima kasih atas karya hebatnya yang berpengaruh besar kepada penulis.
13. Semua pihak yang telah membantu, penulis tidak memiliki banyak teman. Oleh karena itu, terima kasih untuk siapapun yang mau meluangkan waktunya untuk sekadar cerita dan tertawa bersama penulis. Jazakumullah Khairan Katsiran.

Semoga Allah SWT memberikan pahala yang berlipat ganda atas segala kebaikan yang telah diberikan kepada penulis dan semoga penulisan tesis ini bermanfaat bagi perkembangan ilmu pengetahuan dan teknologi, khususnya dalam dunia statistika dan sains.

Makassar, 20 November 2024



Ankaz As Sikib

ABSTRAK

ANKAZ AS SIKIB. **Evaluasi Model *Hybrid Clustering Large Applications* dan *Fuzzy Timeseries Markov Chain* pada Peramalan Polusi Udara *Particular Matter* di Kota Jakarta.** (dibimbing oleh Prof. Dr. Dr. Georgina Maria Tinungki, M.Si. dan Prof. Dr. Nurtiti Sunusi, S.Si., M.Si.)

Analisis *timeseries* merupakan metode statistika dalam memperkirakan peristiwa yang akan terjadi di masa depan. *Fuzzy timeseries markov chain* merupakan pendekatan yang menggabungkan konsep logika *fuzzy* dengan model rangkaian (*chain*) untuk meramalkan nilai di masa depan berdasarkan partisi fuzzy pada data *timeseries*. Partisi memungkinkan berbagai keadaan yang terjadi dalam *timeseries* dan konsep fuzzy memungkinkan kita untuk mengukur derajat keanggotaan dalam setiap partisi. Namun, FTS memiliki beberapa masalah, seperti penggunaan jumlah partisi dan panjang interval yang berubah-ubah untuk semesta wacana. Tujuan penelitian ini adalah mengevaluasi model *hybrid clustering large applications* dan *fuzzy timeseries* dalam meramalkan polusi udara *particular matter* di Kota Jakarta. Metode yang digunakan adalah *hybrid clustering large applications* dan *fuzzy timeseries markov chain*. Analisis membentuk lima cluster, dengan pemilihan *medoid* berhenti pada iterasi pertama dan membentuk matriks probabilitas transisi berordo 5x5. Hasil ketepatan model untuk data PM10 dan PM2.5 berdasarkan MAPE dikategorikan baik.

Kata Kunci: CLARA, *Fuzzy Timeseries*, *Hybrid Fuzzy Timeseries Clustering*, Polusi Udara.

ABSTRACT

ANKAZ AS SIKIB. **Evaluation of Hybrid Clustering Large Applications and Fuzzy Timeseries Markov Chain Model on Forcasting Air Pollution Paticular Matter in Jakarta City.** (Supervised by Prof. Dr. Georgina Maria Tinungki, M.Si. and Prof. Dr. Nurtiti Sunusi, S.Si., M.Si.)

Timeseries analysis is a statistical method of forecasting events that will occur in the future. Fuzzy timeseries Markov chain is an approach that combines the concept of fuzzy logic with a chain model to forecast future values based on fuzzy partitions in the timeseries data. Partitions allow various states to occur in the timeseries and the fuzzy concept allows us to measure the degree of membership in each partition. However, FTS has some problems, such as the use of an arbitrary number of partitions and interval lengths for the universe of discourse. The aim of this research is to evaluate the hybrid clustering large applications and fuzzy timeseries model in forecasting particular matter air pollution in Jakarta City. The method utilized is hybrid clustering large applications and fuzzy timeseries Markov chain. The analysis forms five clusters, with medoid selection stopping at the first iteration and forming a 5x5 transition probability matrix. The results of model accuracy for PM10 and PM2.5 data based on MAPE are categorized as good.

Keywords: Air Pollution, CLARA, Fuzzy Timeseries, Hybrid Fuzzy Timeseries Clustering.

DAFTAR ISI

HALAMAN JUDUL i
PERNYATAAN PENGAJUAN ii
HALAMAN PENGESAHAN iii
PERNYATAAN KEASLIAAN TESIS iv
UCAPAN TERIMA KASIH v
ABSTRAK vii
ABSTRACT viii
DAFTAR ISI ix
DAFTAR TABLE xi
DAFTAR GAMBAR xii
DAFTAR LAMPIRAN xii
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Tujuan Penelitian 3
1.3 Manfaat Penelitian 3
1.4 Peramalan 3
1.5 *Fuzzy Logic* 4
1.6 *Fuzzy Timeseries* 4
1.7 *Fuzzy Timeseries Markov Chain* 5
1.8 *Gap Statistics* 6
1.9 *Euclidean Distance* 6
1.10 Analisis Cluster 7
1.11 Algoritma *Partition Around Medoid* (PAM) 7
1.12 *Clustering Large Applications* (CLARA) 7
1.13 Ketepatan Model Peramalan 8
1.14 Polusi Udara 8
BAB II METODE PENELITIAN 9
2.1 Jenis Penelitian 9
2.2 Sumber Data 9
2.3 Definisi Operational Peubah 9
2.4 Tahapan Analisis Data 9
BAB III HASIL DAN PEMBAHASAN 11
3.1 Analisis Deskriptif 11
3.2 Pemilihan Jumlah Cluster Yang Optimal 11
3.3 Analisis *Clustering Large Applications* (CLARA) 12
3.3.1 Pemilihan *Medoid* Awal 12
3.3.2 Menghitung Total Jarak Objek dengan *Medoid* Awal 13
3.3.3 Pemilihan *Medoid* Baru 14
3.3.4 Menghitung Total Jarak Objek dengan *Medoid* Baru 15
3.3.5 Membandingkan Jarak Total Medoid Awal Dan Medoid Baru 16
3.3.6 Iterasi Pertama: Pemilihan *Medoid* Baru Lagi 16
3.3.7 Menghitung Total Jarak Objek dengan *Medoid* Iterasi Pertama 16
3.3.8 Membandingkan Jarak Total Medoid Baru Dan Medoid Iterasi Pertama 17
3.3.9 Interval Medoid 17
3.4 Analisis *Fuzzy Timeseries Markov Chain* 18
3.4.1 Fuzzifikasi dan *Fuzzy Logic Relationships* 18

3.4.2 *Fuzzy Logic Relationships Grup* 18
3.4.3 Matriks *Probabilitas* Transisi Markov 18
3.4.4 Menghitung Nilai Peramalan Awal 19
3.4.5 Menghitung Penyesuaian Peramalan 19
3.1 Menghitung Ketepatan Model Peramalan 22
3.2 Pembahasan 23
BAB IV KESIMPULAN DAN SARAN 25
4.1 Kesimpulan 25
4.2 Saran 25
DAFTAR PUSTAKA 26
LAMPIRAN 29

DAFTAR TABEL

Nomor urut		Halaman
1.	Kriteria nilai ketepatan model peramalan…….	8
2.	Indeks standar pencemaran udara berdasarkan peraturan menteri lingkungan hidup dan kehutanan republik indonesia……..	8
3.	Data polusi udara PM10 dan PM2.5 di Kota Jakarta per 1 Januari 2021-31 Juli 2024……..	11
4.	Analisis deskriptif data polusi udara PM10 dan PM2.5 di Kota Jakarta per 1 Januari 2021- 31 Juli 2024……..	11
5.	Sampel dari algoritma CLARA dalam pemilihan medoid awal, dimana tiap sampel mewakili urutan data pada data aktual polusi udara PM10 dan PM2.5…….	12
6.	*Medoid* awal berdasarkan sampel data polusi udara PM10 dan PM2.5…..	13
7.	Hasil perhitungan jarak objek dengan *medoid* awal data polusi udara PM10 dan PM2.5…….	14
8.	Sampel dari algoritma CLARA dalam pemilihan *medoid* baru, dimana tiap sampel mewakili urutan data pada data aktual polusi udara PM10 dan PM2.5..…….	14
9.	*Medoid* baru berdasarkan sampel data polusi udara PM10 dan PM2.5…	15
10.	Hasil perhitungan jarak objek dengan *medoid* baru data polusi udara PM10 dan PM2.5……	15
11.	Sampel dari algoritma CLARA dalam pemilihan *medoid* iterasi pertama, dimana tiap sampel mewakili urutan data pada data aktual polusi udara PM10 dan PM2.5……..……..	16
12.	*Medoid* iterasi pertama berdasarkan sampel data polusi udara PM10 dan PM2.5…….……...	16
13.	Hasil perhitungan jarak objek dengan *medoid* iterasi pertama data polusi udara PM10 dan PM2.5……..	17
14.	Interval *medoid* berdasarkan *medoid* pada Table 9, data polusi udara PM10 dan PM2.5……..……..	17
15.	Hasil fuzzifikasi dan FLR data polusi udara PM10 dan PM2.5…….…….	18
16.	Hasil *fuzzy logic relactionship grup* data polusi udara PM10……	18
17.	Hasil fuzzy logic relactionship grup data polusi udara PM2.5……	18
18.	Hasil peramalan awal *hybrid clustering large applications* dan *fuzzy timeseries markov chain* data polusi udara PM10 dan PM2.5…….	19
19.	Hasil penyesuaian peramalan *hybrid clustering large applications* dan *fuzzy timeseries markov chain* data polusi udara PM10 di Kota Jakarta…	20
20.	Hasil penyesuaian peramalan *hybrid clustering large applications* dan *fuzzy timeseries markov chain* data polusi udara PM2.5 di Kota Jakarta..	21
21.	Hasil ketepatan model peramalan *hybrid clustering large applications* dan *fuzzy timeseries markov chain* data polusi udara PM10 dan PM2.5 di Kota Jakarta…….	23

DAFTAR GAMBAR

Nomor urut		Halaman
1.	Alur Penelitian…….	10
2.	Penentuan jumlah *cluster* yang optimal menggunakan *GAP Statistic*, dengan maksimal *cluster* (15) dan *Bootstrapping* (100)……..	12
3.	Grafik data aktual dan peramalan *particular matter* 10 *hybrid clustering large applications* dan *fuzzy timeseries markov chain*……..	20
4.	Grafik data aktual dan peramalan *particular matter* 2.5 *hybrid clustering large applications* dan *fuzzy timeseries markov chain*……..	22

DAFTAR LAMPIRAN

Nomor urut		Halaman
1.	Data polusi udara harian *particular matter* di Kota Jakarta……..	30
2.	Hasil perhitungan cost total medoid awal, baru dan iterasi pertama……	31
3.	Hasil peramalan *hybrid hybrid clustering large applications* dan *fuzzy timeseries markov chain* pada peramalan polusi udara particular matter di Kota Jakarta……	34
4.	*Curriculum Vitae*……	35

BAB I

PENDAHULUAN

1.1. Latar Belakang

Analisis *timeseries* merupakan metode statistika dalam memperkirakan peristiwa yang akan terjadi di masa depan. Analisis *timeseries* memiliki peran penting dalam perencanaan, pengambilan keputusan dan mengoptimalkan proses bisnis dengan mempertimbangan tren, pola, dan perilaku yang mendasari data empiris (Aditya Satrio et al., 2021). Fokus utama dari proses peramalan adalah mengurangi tingkat ketidakpastian dan menghasilkan prediksi yang lebih akurat mengenai kejadian di masa depan, selain melibatkan hal subsektif dan pengalaman (Petropoulos et al., 2022; Zaenurrohman et al., 2021).

Salah satu pendekatan yang menarik dalam analisis *timeseries* adalah penggunaan metode berbasis logika *fuzzy*. Logika *fuzzy* telah terbukti memberikan hasil yang lebih efektif dalam menyelesaikan berbagai masalah, termasuk dalam peramalan data *timeseries* (Zhang et al., 2017). Logika *fuzzy* memungkinkan kita untuk mengatasi ketidakpastian dalam data *timeseries*, yang seringkali disebabkan oleh variasi dan faktor eksternal yang sulit ditangani oleh metode analisis klasik (Cheng et al., 2016).

Fuzzy timeseries diperkenalkan oleh (Song & Chissom, 1993) dengan memprediksi pendaftar di Alabama University (Zhang, 2012). Sejak saat itu, berbagai metode *fuzzy timeseries* banyak dikembangkan seperti *weighted* (Yu, 2005), *chen* (Chen, 2006), *markov* (Sullivan & Woodall, 1994) dan *multiple atribut* (Cheng dan Wang, 2008). Salah satu metode baru yang mendapat perhatian dalam analisis *timeseries* berbasis *fuzzy* adalah *fuzzy timeseries markov chain*.

Berdasarkan penelitian yang dilakukan oleh (Alyousifi et al., 2020; Ramadani & Devianto, 2020) *fuzzy timeseries markov chain* merupakan metode yang paling handal berdaraskan MSE dan MAPE dibandingkan metode *fuzzy timeseries* lainnya. *Fuzzy timeseries markov chain* merupakan pendekatan yang menggabungkan konsep logika *fuzzy* dengan model rangkaian (*chain*) untuk meramalkan nilai di masa depan berdasarkan partisi *fuzzy* pada data *timeseries*. Partisi memungkinkan berbagai keadaan yang terjadi dalam *timeseries* dan konsep *fuzzy* memungkinkan kita untuk mengukur derajat keanggotaan setiap keadaan dalam setiap partisi.

Rantai markov didefinisikan oleh matriks peluang transisi yang berisi informasi untuk mengatur sistem dari satu kejadian ke kejadian lainnya. Matriks peluang transisi bergantung pada jumlah partisi yang ditentukan. Jumlah partisi juga mempengaruhi proses fuzzifikasi, yang menjadi acuan terhadap hubungan keanggotaan (FLR). Namun, FTS memiliki beberapa masalah, seperti penggunaan jumlah partisi dan panjang interval yang berubah-ubah untuk semesta wacana. Karenanya, pemilihan jumlah partisi yang optimal merupakan masalah menarik yang perlu diatasi.

Pada dasarnya penentuan jumlah partisi dan panjang interval analisis *fuzzy timeseries markov chain* menggunakan pendekatan hirarki, yaitu penentuan dengan melewati proses analisis terlebih dahulu (Tsaur, 2012). Namun, dalam metode *fuzzy timeseries*, penentuan jumlah partisi dan panjang interval tidak memiliki rumus pasti dalam perhitungannya (Zaenurrohman et al., 2021). Hubungan jumlah partisi terhadap panjang interval terbentuk berdasarkan penelitinya (Alyousifi et al., 2020; Mubarrok et al., 2022). Meskipun penentuan jumlah partisi dan panjang interval sangat berpengaruh terhadap terbentuknya hubungan keanggotaan (FLR) yang mengakibatkan perbedaan akurasi hasil peramalan.

Oleh karena itu, untuk mengoptimalkan *fuzzy timeseries markov chain*, diperlukan solusi pasti untuk pemilihan partisi dan panjang interval. Beberapa penelitian sebelumnya telah berupaya melakukan gabungan metode *clustering* dalam mengoptimalkan partisi metode *fuzzy timeseries* (Dewi et al., 2023; Surono, Goh, et al., 2022; Vovan & Lethithu, 2020; Vovan & Phamtoan, 2021). Namun, dalam menentukan partisi yang optimal, metode *k-means* dan *k-medoids* masih memiliki kekurangan, karena kurang efektif dalam menganalisis data berukuran besar dibandingkan dengan metode pengembangan seperti *clustering large applications* (CLARA).

CLARA merupakan pengembangan metode *cluster partition around medoid* (PAM) yang menerapkan *medoid* sebagai pusat *cluster*nya. CLARA memiliki sifat *robust* terhadap data berjumlah besar dan dapat mengatasi *outlier* (Gentle et al., 1991). CLARA menggunakan sampel dalam mengelompokan data berskala besar, dengan menerapkan algoritma PAM. *Medoid* diambil secara acak dan menjadi perwakilan *cluster* sampai semua objek terpilih. CLARA menghasilkan *cluster* yang tetap pada setiap iterasi (Kamber & Han, 2006).

Polusi udara merupakan perubahan komposisi udara akibat pencampuran dengan *particular matter* (PM), *sulfur dioksida* (SO2), *karbon monoksida* (CO) atau logam lainnya. Konsentrasi zat pencemar yang melebihi ambang batas toleransi dapat berdampak negatif pada lingkungan dan kesehatan manusia (Auliana et al., 2024; Choi, 2018). Menurut *World Health Organitation* (WHO), polusi udara merupakan faktor risiko yang signifikan terhadap berbagai masalah kesehatan, mulai dari iritasi kulit, hidung, tenggorokan, dan mata, hingga kondisi serius seperti penyakit jantung, pneumonia, bronkitis, dan kanker paru-paru. Selain itu, polusi udara dapat merusak lapisan ozon dan berkontribusi pada pemanasan global (WHO, 2021).

Berdasarkan *world air quality report* 2023, Indonesia menempati peringkat pertama negara paling berpolusi di Asia Tenggara (IQAir, 2023). Sementara itu, per 23 Juni 2024, Jakarta merupakan kota paling berpolusi kedua di dunia dengan *air quality index* sebesar 160 (IQAir, 2024). Hal tersebut menjadi pertanyaan besar, karena Indonesia memiliki hutan yang sangat luas sebagai penghasil oksigen terbesar di dunia. Berbagai kebijakan Pemerintah DKI Jakarta dalam menekan tingginya polusi udara seperti mobilisasi transportasi umum, membentuk satgas pengendali udara, emisi kendaraan bermotor dan kebijakan disinsentif tarif parkir. Namun, indeks kualitas udara Kota Jakarta masih terus melonjak.

Oleh karena itu, diperlukan upaya solutif dan acuan dalam pengambilan keputusan untuk menekan angka polusi udara di Kota Jakarta. Penelitian mengenai prediksi kualitas udara menggunakan *fuzzy timeseries markov chain* dan *hybrid clustering fuzzy timeseries* telah dilakukan oleh beberapa peneliti (Alyousifi et al., 2020; Dewi et al., 2023; Kingsy et al., 2017; Surono et al., 2022; Vovan & Lethithu, 2022). Namun, Berdasarkan penelusuran yang dilakukan, pemodelan *hybrid clustering large applications* dan *fuzzy timeseries* belum pernah dilakukan. Oleh karena itu, penelitian ini mempertimbangkan evaluasi model *hybrid clustering large applications* dan *fuzzy timeseries markov chain* dalam meramalkan polusi udara *particular matter* di Kota Jakarta.

1.2. Tujuan Penelitian

Berdasarkan uraian latar belakang diatas, tujuan utama dari penelitian ini, antara lain:

1. Memperoleh hasil evaluasi model *hybrid clustering large applications* dan *fuzzy timeseries markov chain* dalam meramalkan polusi udara *particular matter* di Kota Jakarta.
2. Memperoleh hasil peramalan polusi udara *particular matter* di Kota Jakarta dengan menggunakan metode *hybrid clustering large applications* dan *fuzzy timeseries markov chain.*

1.3. Manfaat Penelitian

Setelah tercapainya tujuan penelitian ini, maka diharapkan:

1. Hasil evaluasi model *hybrid clustering large applications* dan *fuzzy timeseries markov chain* dapat menjadi bahan kajian dalam pengembangan metode *timeseries*.
2. Hasil peramalan polusi udara *particular matter* di Kota Jakarta dapat menjadi acuan bagi instansi terkait dalam pengambilan keputusan guna menekan tingginya polusi udara di Kota Jakarta dan edukasi kepada Masyarakat.

1.4 Peramalan

Peramalan merupakan teknik analisis dalam memperkirakan nilai pada masa dapan dengan memperhatikan data masa lalu dan masa sekarang. Peramalan juga dapat diartikan sebagai cara untuk mendapatkan gambaran di masa depan dengan mempertimbangkan data historis dan penggunaan suatu model matematis. Hasil dari peramalan akan dijadikan dasar dan pertimbangangan dalam pengambilan keputusan. Peramalan dikategorikan menjadi tiga, berdasarkan jangka waktunya.

1) Jangka pendek ($< 3\ bulan$).
2) Jangka menengah ($3\ bulan\ \leq X \leq 3\ tahun$).
3) Jangka panjang ($> 3\ tahun$) (Heizer, Jay. Render, 2014).

1.5 *Fuzzy Logic*

Fuzzy logic merupakan pengembangan dari logika *boolean.* Konsep *fuzzy logic* dikenalkan oleh (Zadeh, 1968) dengan tidak membatasi keanggotaan himpunan menjadi absolut 0 atau 1, tetapi mentolerir berbagai tingkat keanggotaan. *Fuzzy logic* memiliki domain atribut interval [0, 1]. *Fuzzy logic* memiliki keunggulan dalam perhitungannya karena mempertimbangkan Kemungkinan tidak pasti dan tidak kaku (Sabahi & Akbarzadeh-T, 2016). Sistem *fuzzy logic* dapat menangani istilah nonlinear yang tidak diketahui dan metode kontrol yang diusulkan secara efektif (Li et al., 2021).

1.6 *Fuzzy Timeseries*

Metode *fuzzy timeseries* umumnya menggunakan data historis berupa data linguistik (Efendi et al., 2015). Tahapan FTS meliputi pendefinisian semesta wacana (U), pembagian U ke dalam beberapa interval, fuzzifikasi, pembentukan hubungan *fuzzy*, defuzzifikasi, dan penentuan nilai prediksi.

Definisi 1 Misalkan $U = \{u_1, u_2, u_3, \dots, u_n\}$ adalah semesta wacana, maka u_n $(i = 1, \dots, n)$ adalah nilai linguistik yang mungkin dalam U. Himpunaan *fuzzy* variable linguistik A; dari U didefinisikan seabagai berikut:

$$A_i = \frac{f_{A_i}(u_1)}{u_1} + \frac{f_{A_i}(u_2)}{u_2} + \cdots + \frac{f_{A_i}(u_n)}{u_n} \quad (1.1)$$

dimana f_{A_i} adalah fungsi keanggotaan dari himpunan *fuzzy* $f_{A_i}: U \rightarrow [0,1]$, $f_{A_i}(u_r) \in [0,1]$ dan $1 < r < n$.

Definsi 2 Misalkan $Y(t)$ $(t = 1, 2, \dots, n)$, merupakan bagian dari bilangan real dan semesta wacana himpunan *fuzzy* $f_i(t)$, Jika $F(t)$ adalah himpunan $f_i(t), i = 1,2,3, \dots, n$. Maka $F(t)$ didefinisian sebagai *fuzzy time series* dari $Y(t)$.

Definsi 3 Misalkan $Y(t) = A_j$ disebabkkan oleh $Y(t-1) = A_i$, maka *fuzzy logical relationship* (FLR) didefinisikan sebagai $A_i \rightarrow A_j$.

Definisi 4 Jika terdapat FLR yang diperoleh dari *state* A_2, maka transisi dibuat ke *state* yang lain yaitu A_j, $j = 1,2,3, \dots n$, seperti $A_2 \rightarrow A_3$, $A_2 \rightarrow A_2$, $A_2 \rightarrow A_1$. Maka, FLR dikelompokkan menjadi *fuzzy logical relationship group* (FLRG) seperti berikut:

$$A_2 \rightarrow A_1, A_2, A_3 \quad (1.2)$$

Fuzzifikasi adalah tahap di mana data diubah menjadi nilai linguistik untuk membentuk FLR. Proses ini memerlukan nilai batas atas dan batas bawah yang diperoleh dari persamaan:

$$ub_i = \frac{cluster\ center_i + cluster\ center_{i+1}}{2} \quad (1.3)$$

dimana $i = 1, 2, \dots, k$. ub_i adalah batas atas dari interval ke-i, dan batas bawah interval ke-$i + 1$. Karena tidak ada pusat *cluster* sebelum *cluster center* pertama dan

terakhir, maka nilai batas bawah pada lb_1 dan batas atas pada ub_k diperoleh dengan menggunakan aturan sebagai berikut:

$$ub_k = cluster\ center_k + |max_{data} - cluster\ center_k| \quad (1.4)$$

$$lb_1 = cluster\ center_1 - |cluster\ center_1 - min_{data}| \quad (1.5)$$

(Dewi et al., 2023; Surono, Goh, et al., 2022)

1.7 *Fuzzy Timeseries Markov Chain*

Penentukan matriks peluang transisi markov chain dibentuk berdimensi $p \times p$, dimana p merupakan banyaknya himpunan *fuzzy* (Li et al., 2021). Persamaan untuk menentukan probabilitas transisi *state* sebagai berikut:

$$P_{ij} = \frac{r_{ij}}{r_i} \quad (1.6)$$

dimana:

P_{ij} : probabilitas transisi *state* A_i ke A_j

r_{ij} : banyaknya transisi *state* A_i ke A_j

r_i : banyaknya data yang termasuk dalam *state* A_i

Matriks Peluang transisi **P** dapat ditulis sebagai berikut:

$$\boldsymbol{P} = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1p} \\ P_{21} & P_{22} & \dots & P_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ P_{p1} & P_{p2} & \dots & P_{pp} \end{bmatrix}$$

1) Menghitung nilai peramalan awal, dengan aturan sebagai berikut:

Aturan 1 Jika terdapat himpunan *fuzzy* yang tidak memiliki FLR ($A_i \rightarrow \emptyset$), dan kemudian $Y(t-1)$ pada waktu t-1 masuk dalam A_i, maka nilai peramalan F_t adalah m_i, dimana m_i adalah nilai tengah dari interval x_i.

Aturan 2. Jika FLRG A_i adalah relasi satu ke satu ($A_i \rightarrow A_q$), dimana $Y(t-1)$ pada waktu t-1 masuk dalam *state* A_i, maka nilai peramalan F_t adalah m_q dengan m_q merupakan nilai tengah dari x_q pada FLRG yang terbentuk pada data ke t-1.

Aturan 3. Jika FLRG A_i adalah relasi satu ke banyak *($A_j \rightarrow A_1, A_2, A_3, \dots, A_q$, j = 1, 2, 3, …, q)* dimana $Y(t-1)$ pada waktu t-1 masuk dalam *state* A_j, maka peramalan $F(t)$ adalah sebagai berikut:

$$F(t) = m_1P_{i1} + m_2P_{i2} + \cdots + m_{i-1}P_{i(i-1)} + Y(t)P_{ii} + m_{i+1}+1P_{i(i+1)} + \cdots + m_nP_{in} \quad (1.7)$$

dimana, $m_1, m_2, \dots, m_n$ adalah nilai tengah dari $u_1, u_2, \dots, u_n$ dan m_i diganti dengan $Y(t)$ pada state A_i untuk mendapatkan nilai akurasi yang lebih baik.

2) Mengatur penyesuaian nilai peramalan, dengan menambahkan selisih nilai aktual $Y(t)$, yang dapat menyesuaikan nilai peramalan untuk mengurangi kesalahan. persamaan berikut:

$$\hat{F}(t+1) = F(t+1) + \text{diff}\,(Y(t)) \tag{1.8}$$

$diff\ (Y(t))$ merupakan selisih antara nilai data aktual $(Y(t))$ pada waktu ke-i dengan nilai sebelumnya $(Y(t_{i-1})$.

$$diff\big(Y(t)\big) = \begin{cases} 0 & ,\text{jika } Y(t=1) \\ Y(t) - Y(t-1) & ,\text{jika } Y(t \geq 2) \end{cases} \tag{1.9}$$

dimana:

$Y(t)$: data aktual pada periode ke-t
$F(t)$: hasil peramalan awal pada periode ke-t
$\hat{F}(t)$: hasil peramalan setelah penyesuaian pada periode ke-t

1.8 *GAP Statistics*

Gap Statistics merupakan metode yang digunakan untuk menentukan jumlah *cluster* yang optimal. *Gap Statistics* menentukan jumlah *cluster* optimal dengan membandingkan jumlah variasi dalam *cluster* dari data aktual dengan ekspektasi dari data acak. *Gap Statistic* lebih optimal dalam algoritma CLARA yang bekerja dengan sampel dari dataset besar. *Gap Statistics* diperoleh berdasarkan persamaan berikut:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log(W_{kb}) - \log\,(W_k) \tag{1.10}$$

dengan:

$Gap(k)$: Gap untuk jumlah cluster k yang optimal
B : Jumlah bootstrap yang digunakan dalam metode Gap statistik.
W_{kb} : Dispersi intra-cluster untuk jumlah cluster k pada bootstrap ke-b.
W_k : Hasil within cluster variation untuk jumlah cluster k pada data asli.

1.9 *Euclidean Distance*

Euclidean distance merupakan metode menghitung jarak antar titik dalam ruang *euclidean* dan kemudian mengelompokan titik tersebut menjadi beberapa *cluster* berdasarkan jarak antar titik (Alguliyev et al., 2020). CLARA menggunakan *euclidean distance* untuk menghitung jarak antar titik yang kemudian mengelompokan titik kedalam *cluster.* Rumus *euclidean distance* sebagai berikut: (Thamrin & Wijayanto, 2021)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_k - y_i)^2} \quad ,k = 1,2,\dots,c \tag{1.11}$$

dimana:

$d(x,y)$: Jarak *euclidean distance* antara $x_k - y_i$
x_k : nilai cluster center ke-k
y_i : nilai data aktual ke-i $(i = 1,2,\dots,n)$

1.10 Analisis *Cluster*

Analisis *cluster* bertujuan mengelompokan sebuah objek kedalam suatu kelompok dengan mempertimbangkan tingkat kemiripan suatu objek. Analisis *cluster* dapat secara efektif mengatur kumpulan data ke dalam kelas terpisah, memaksimalkan kesamaan dalam kelompok, sekaligus meminimalkan redundasi data (Vehkalahti, K., & Everitt, 2018).

1.11 Algoritma *Partition Around Medoids* (PAM)

PAM diperkenalkan oleh (Gentle et al., 1991). PAM menggunakan *medoid* sebagai pusat *cluster* (Shamsuddin, N., & Mahat, 2019). Pemilihan *medoid* sebuah K *cluster* dari n objek ditentukan dari titik *random* yang diasumsikan representatif dari setiap *cluster* (Wu et al., 2022). Prinsip dasar algoritma PAM adalah meminimalisir jumlah ketidaksamaan antar objek pada *cluster*, dengan menukar objek *medoid* dan *non-medoid* hingga konvergen (Arora et al., 2016). Umumnya proses untuk mencari *medoid* baru dilakukan secara berulang sehingga terbentuk *medoid* terbaik dengan jumlah jarak terkecil yang mewakili *cluster*.

S = jarak total euclidean *medoid* baru − jarak total euclidean *medoid* lama (1.12)

dimana:

Jika (S) < 0 : Proses pemilihan *medoid* baru dengan mengganti objek non-*medoid* diulang kembali

jika (S) > 0 : Maka iterasi berhenti.

1.12 *Clustering Large Applications* (CLARA)

Clustering Large Applications (CLARA) merupakan metode pengembangan yang dikenalkan oleh (Gentle et al., 1991) dalam mengatasi kelemahan dari algoritma PAM yang hanya efektif pada data berukuran kecil. Selayaknya metode pengembangan dari PAM, CLARA juga menggunakan *medoid* sebagai pusat *cluster* untuk mengelompokan data berskala besar serta kokoh terhadap *outlier* (Gupta & Panda, 2019). CLARA membagi data berskala besar menjadi beberapa subset yang lebih kecil dengan mempertimbangkan pemilihan *medoid* yang optimal. Penentuan jumlah sampel untuk setiap subset berdasarkan persamaan berikut:

$$min\ (40 + 2\ x\ K) \quad (1.13)$$

dimana:

K : banyaknya *cluster*

Objek dipilih secara acak sebagai perwakilan *cluster* sampai semua objek terpilih. Perbedaan PAM dan CLARA terletak pada proses pencarian *medoid* terbaik, jika PAM mencari diantara keseluruhan data. Maka CLARA mencari *medoid* di antara sampel yang dipilih dari keseluruhan data. Oleh sebab itu, CLARA tidak dapat menemukan pengelompokan yang baik jika salah satu sampel jauh dari titik pusat *medoid* (Schubert & Rousseeuw, 2021).

1.13 Ketepatan Model Peramalan

Besarnya galat pada hasil peramalan diketahui menggunakan *Mean absolute percentage error* (MAPE). Semakin rendah nilai yang diperoleh maka akurasi peramalan semakin tinggi dan berlaku sebaliknya (Hodson, 2022; Sunusi, 2022). Persamaan untuk mengukur keakuratan hasil analisis *timeseries* sebagai berikut:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y(t)-\hat{F}(t)}{Y(t)}\right| \times 100\% \qquad (1.14)$$

Table 1. Kriteria nilai ketepatan model peramalan

Nilai MAPE	Kriteria
≤ 10	Sangat Baik
10 < Nilai ≤ 20	Baik
20 < Nilai ≤ 50	Cukup
> 50	Buruk

1.14 Polusi Udara

Polusi udara merupakan perubahan komposisi udara akibat pencampuran seperti *particular matter* (PM), sulfur dioksida (SO2), karbon monoksida (CO) atau logam berat lainnya (Kingsy et al., 2017). Badan Meteorologi, Klimatologi dan Geofisika (BMKG) mendefinisikan polusi udara sebagai kehadiran substansi fisik, kimia, atau biologi di atmosfer dalam jumlah yang dapat membahayakan kesehatan manusia, hewan, dan tumbuhan. Polusi udara dapat berasal dari sumber alami maupun aktivitas manusia, seperti emisi kendaraan, industri, dan pembakaran bahan bakar.

Proses terbentuknya PM pada dasarnya terbentuk secara langsung. Namun, reaksi polutan dari sumber tertentu di atmosfer juga dapat membentuk PM. PM10 dan PM2.5 memiliki ukuran yang sangat kecil, sehingga dapat masuk dan menyebabkan masalah kesehatan serius pada organ tubuh seperti paru-paru (Choi, 2018; IQAir, 2023).

Table 2. Indeks standar pencemaran udara berdasarkan peraturan menteri lingkungan hidup dan kehutanan republik Indonesia.

Kategori	ISPU	PM10 (µg/m³)	PM2.5 (µg/m³)	SO2 (µg/m³)	CO (µg/m³)	O3 (µg/m³)	NO2 (µg/m³)
Baik	0-50	50	15.5	52	4000	120	80
Sensitif	51-100	150	55.4	180	8000	235	200
Tidak Sehat	101-200	350	150.4	365	15000	400	1130
Sangat Tidak Sehat	201-300	420	250.4	800	30000	800	2260
Berbahaya	>300	500	500	1200	45000	1000	3000

source: *Peraturan menteri lingkungan hidup dan kehutanan nomor 14 tahun 2020 tentang indeks standar pencemar udara (ISPU)*, 1–16.

BAB II

METODE PENELITIAN

2.1 Jenis Penelitian

Penelitian ini, menggunakan pendekatan kuantitatif, dengan fokus utama pada data numerik yang diolah menggunakan metode statistika, *hybrid clustering large applications* dan *fuzzy timeseries markov chain.*

2.2 Sumber Data

Sumber data dalam penelitian ini, diperoleh melalui situs web https://lingkunganhidup.jakarta.go.id/. Dinas Lingkungan Hidup DKI Jakarta berupa data harian indeks standar pencemaran udara pada 1 Januari 2021- 31 Juli 2024.

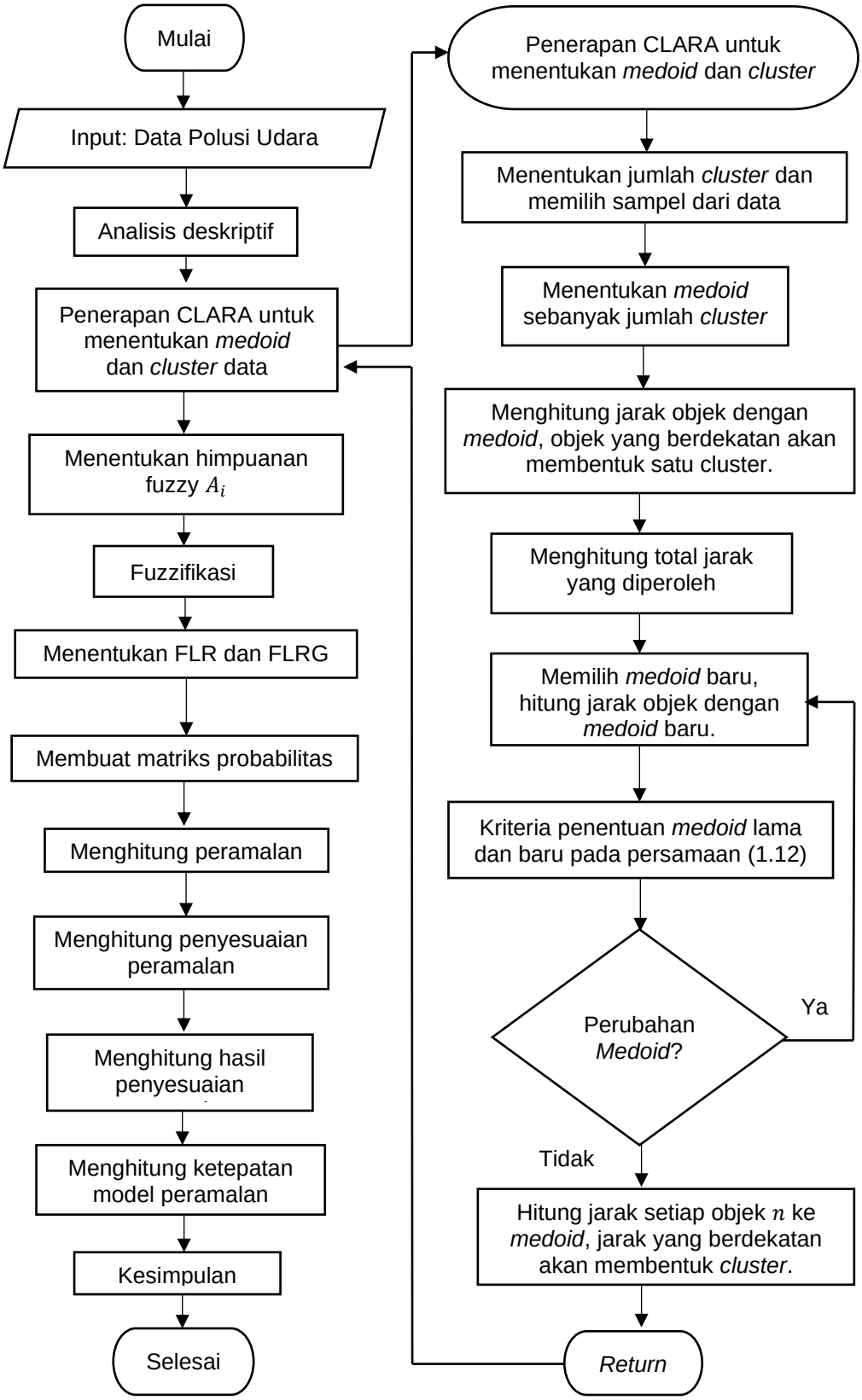
2.3 Definisi Operational Peubah

Y_1 : polusi udara PM10 di Kota Jakarta

Y_2 : polusi udara PM2.5 di Kota Jakarta

2.4 Tahapan Analisis Data

1. Menginput data dan melakukan analisis deskriptif.
2. Menentukan jumlah *cluster* yang optimal dari n objek.
3. Memilih sampel dengan ukuran pada persamaan (1.13).
4. Menetukan *medoid* awal sebanyak jumlah *cluster*.
5. Menghitung jarak objek dengan *medoid* awal setiap *cluster*, objek yang berdekatan akan membentuk satu *cluster.*
6. Menempatkan objek berdasarkan jarak terdekat dengan *medoid* awal.
7. Menghitung total jarak *euclidean distance* yang diperoleh.
8. Memilih secara acak objek non-*medoid* sebagai kandidat *medoid* baru. lalu hitung total jarak objek non-*medoid* dengan kandidiat *medoid* baru. Kriteria penentuan *medoid* awal dan baru sebagai *medoid* tetap pada persamaan (1.12).
9. Ulangi Tahap 8-9, sampai tidak ada perubahan.
10. Hasil *clustering*, akan dilanjutkan pada tahapan *fuzzy timeseries markov chain*.
11. Selanjutnya, dilakukan fuzzifikasi dengan mengelompokkan data aktual pada himpunan *fuzzy* Ai yang sesuai.
12. Menentukan *Fuzzy Logic Relationship* (FLR) dan membentuk *Fuzzy Logic Relationship Group* (FLRG).
13. Menentukan matriks *probabilitas* transisi markov
14. Menghitung nilai peramalan awal.
15. Mengatur penyesuaian kecenderungan nilai peramalan.
16. Menghitung nilai peramalan akhir.
17. Menghitung ketepatan model peramalan.


Mulai
Input: Data Polusi Udara
Analisis deskriptif
Penerapan CLARA untuk menentukan *medoid* dan *cluster* data
Menentukan himpuanan fuzzy A_i
Fuzzifikasi
Menentukan FLR dan FLRG
Membuat matriks probabilitas
Menghitung peramalan
Menghitung penyesuaian peramalan
Menghitung hasil penyesuaian
Menghitung ketepatan model peramalan
Kesimpulan
Selesai
Penerapan CLARA untuk menentukan *medoid* dan *cluster*
Menentukan jumlah *cluster* dan memilih sampel dari data
Menentukan *medoid* sebanyak jumlah *cluster*
Menghitung jarak objek dengan *medoid*, objek yang berdekatan akan membentuk satu cluster.
Menghitung total jarak yang diperoleh
Memilih *medoid* baru, hitung jarak objek dengan *medoid* baru.
Kriteria penentuan *medoid* lama dan baru pada persamaan (1.12)
Perubahan *Medoid*?
Ya
Tidak
Hitung jarak setiap objek n ke *medoid*, jarak yang berdekatan akan membentuk *cluster*.
Return