UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA

Tesis

untuk memenuhi sebagian persyaratan
mencapai gelar Magister

Program Studi Magister Teknologi Informasi
Konsentrasi *e-Goverment*
Departemen Teknik Elektro dan Teknologi Informasi

diajukan oleh
FAHMI DHIMAS IRNAWAN
18/434927/PTK/12490

Kepada

PROGRAM PASCASARJANA
FAKULTAS TEKNIK
UNIVERSITAS GADJAH MADA
YOGYAKARTA
2022

UNIVERSITAS
GADJAH MADA

METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

TESIS

METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK, PROVINSI DI YOGYAKARTA

Dipersiapkan dan disusun oleh

Fahmi Dhimas Irnawan
18/434927/PTK/12490

Telah dipertahankan di depan dewan penguji
Pada tanggal : **31 Desember 2021**
Susunan Dewan Penguji

Ketua	Anggota
Dr. Ir. Rudy Hartanto, M.T., IPM.	**Ir. Lukito Edi Nugroho, M.Sc., Ph.D.**
Anggota	Anggota
Dr. Indriana Hidayah, S.T., M.T.	**Adhistya Erna Permanasari, S.T., M.T., Ph.D.**

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister

Tanggal: **31 Januari 2022**
Ketua Program Studi Magister Teknologi Informasi

Dr. Ir. Rudy Hartanto, M.T., IPM.
NIP. 196403151990031003

Mengetahui,
Ketua Departemen
Teknik Elektro dan Teknologi Informasi

Ir. Hanung Adi Nugroho, S.T., M.E., Ph.D., IPM.
NIP. 197802242002121001

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini:

Nama	: Fahmi Dhimas Irnawan
NIM	: 18/434927/PTK/12490
TahunTerdaftar	: 2018
Program Studi	: Magister Teknologi Informasi
Fakultas/Sekolah	: Fakultas Teknik

Menyatakan bahwa dalam dokumen ilmiah Tesis ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan sumbernya secara lengkap dalam daftar pustaka.

Dengan demikian saya menyatakan bahwa dokumen karya ilmiah ini bebas dari unsur-unsur plagiasi dan apabila dokumen ilmiah Tesis ini di kemudian hari terbukti merupakan plagiasi dari hasil karya penulis lain dan/atau dengan sengaja mengajukan karya atau pendapat yang merupakan hasil karya penulis lain, maka pennulis bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Yogyakarta, 31 Januari 2022

Fahmi Dhimas Irnawan

18/434927PTK/12490

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

PRAKATA

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan barokah-Nya sehingga penulis dapat menyelesaikan tesis dengan judul " Metode Imputasi Data Hilang pada Daerah Aliran Sungai Opak, Provinsi DI Yogyakarta". Laporan tesis ini disusun untuk memenuhi salah satu syarat dalam memperoleh gelar *Master of Engineering (M.Eng.)* pada Program Studi Magister Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada Yogyakarta.

Dalam melakukan penelitian dan penyusunan laporan tesis ini penulis telah mendapatkan banyak dukungan dan bantuan dari berbagai pihak. Penulis mengucapkan terima kasih yang tak terhingga kepada:

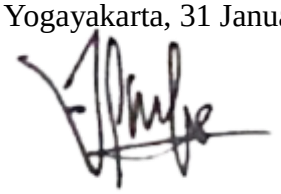
1. Ibu Indriana Hidayah, Dr., S.T., M.T. selaku dosen pembimbing utama, dan Bapak Ir. Lukito Edi Nugroho, M.Sc., Ph.D selaku dosen pembimbing pendamping yang telah dengan penuh kesabaran dan ketulusan memberikan ilmu dan bimbingan terbaik kepada penulis.
2. Bapak. Ir. Hanung Adi Nugroho, S.T., M.E., Ph.D., IPM. selaku Ketua Departemen Teknik Elektro dan Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada dan Bpk. Rudy Hartanto, Dr. Ir., M.T., IPM.selaku Ketua Program Studi Magister Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada atas izin belajar yang diberikan.
3. Bapak/Ibu Dosen Program Studi Magister Teknologi Informasi Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada yang telah memberikan bekal ilmu kepada penulis.
4. Para Karyawan/wati Program Studi Magister Teknologi Informasi Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada yang telah membantu penulis dalam proses belajar.
5. Seluruh keluarga Bapak Purnawan, S.T., M.Eng., Ibu Tetty Murniati, dan Adik Farchan Aldi Irnawan serta Stefanus Okky Setya Mahendra yang senantiasa memberikan dukungan, do'a, semangat dan segala apapun yang menjadikan dapat terselesaikannya tesis ini.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

6. Teman-teman seperjuangan mahasiswa Magister Teknologi Informasi 2018, dan khususnya teman – teman dari konsentrasi e-Gov yang telah memberikan motivasi, semangat, dan bantuan atas terselesainya tesis ini.
7. Kakak senior selaku pembimbing ke-tiga, Mbak Eko Mulyani, Mas Zaenuri Putro Utomo, Alm. Mas Beny Rustam dan Mas Syaifulloh Amien Pandega serta rekan - rekan bimbingan Ibu Indriyana Hidayah yang telah membimbing dalam penyelesaian Thesis.
8. Berbagai pihak yang telah memberikan bantuan dan dorongan serta berbagi pengalaman pada proses penyusunan tesis ini.

Penulis menyadari sepenuhnya bahwa laporan tesis ini masih jauh dari sempurna, untuk itu semua jenis saran, kritik dan masukan yang bersifat membangun sangat penulis harapkan. Semoga tulisan ini dapat memberikan manfaat dan wawasan tambahan bagi para pembaca dan khususnya bagi penulis sendiri.

Yogayakarta, 31 Januari 2022



Fahmi Dhimas Irnawan

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

ARTI LAMBANG DAN SINGKATAN

AWLR	=	*AutomaticWater Level Recorder*
DAS	=	Daerah Aliran Sungai
k-NNi	=	*k-Nearest Neighbor Imputation*
k-NN	=	*K-Nearest Neighbor*
MICE	=	*Multivariate Imputation and Chained Equation*
RMSE	=	*Root Mean Squared Error*
R^2	=	*R Squared*
MAE	=	*Mean Absolute Error*
SISDA	=	Sistem Informasi Sumber Daya Air
BBWSDA	=	Balai Besar Wilayah Sumber Daya Air
NA	=	*Not Available*
IDE	=	*Integrated Development Environment*
Rstudio	=	*R Studio*
MCAR	=	*Missing Completely at Random*

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

The availability of water resources data in Indonesia has several complex problems related to the completeness of the data. Problems that occur during data collection in several agencies at Indonesia are the accuracy of the data and the completeness of the data. The concept of water in the form of objects can be said to be very dynamic starting from the shape, color, discharge, and smell. The impact of climate change and natural disasters is also a contributing factor to the dynamic changes of water. The characteristics of the watershed in DI Yogyakarta Province which is fan-shaped with many downstream branches of the river make the complexity of water resource data more widespread.

One of the problems that arise is the missing value of water resources data which can affect the processing of water resources data. There are several methods that can be used for Missing Value Imputation, one of which is k-Nearest Neighbors Imputation (k-NNi) and Multivariate Imputation by Chained Equation (MICE). The two methods proposed are used to compare and find the most appropriate method using the Opak watershed in Yogyakarta D.I Province.

The results of the statistical validation comparison show that the most consistent average value of RMSE and MAE is the k-NNi method with a value of k=8. As for the comparison of R2 values, the k-NNi method with a value of k=8 gets the best average value of 80%, followed by the k-NNi method of k=20 as the default k value with a percentage of 78%. The MICE comparison method gets the lowest average percentage value from other methods by only getting a value of 63%.

Keywords: Watershed, Waterflow, k-NNi, MICE, imputation, *missing value, statistical validation*

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

Ketersediaan data sumber daya air di Indonesia memiliki beberapa permasalahan yang kompleks terkait dengan kesempurnaan dan kelengkapan data. Permasalahan yang terjadi pada saat pendataan di beberapa instansi di Indonesia adalah keakuratan data dan kelengkapan data. Konsep air dalam wujud benda dapat dikatakan sangat dinamis mulai dari bentuk, warna, debit, dan bau. Dampak perubahan iklim dan bencana alam juga menjadi faktor pendukung perubahan dinamis dari air. Karakteristik DAS yang ada di Provinsi DI Yogyakarta yang berbentuk kipas dengan hilir cabang sungai yang banyak menjadikan kompleksitas dari data sumber daya air menajdi semakin meluas.

Salah satu permaslahan yang timbul adalah nilai hilang dari data sumber daya air yang dapat berpengaruh terhadap pengolahan data sumber daya air. Ada beberapa metode yang dapat digunakan untuk *imputasi missing value*, salah satunya adalah *k-Nearest Neighbors Imputation* (k-NNi) dan *Multivariate Imputation by Chained Equation* (MICE). Kedua metode akan yang diusulkan digunakan untuk membandingkan dan menemukan metode yang paling tepat menggunakan DAS Opak di Provinsi D.I Yogyakarta.

Hasil perbandingan validasi statistik, nilai rata-rata RMSE dan MAE yang paling konsisten adalah metode k-NNi dengan nilai k=8. Sedangkan untuk perbandingan nilai R2, metode k-NNi dengan nilai k=8 mendapatkan nilai rata-rata terbaik sebesar 80%, disusul dengan metode k-NNi sebesar k=20 sebagai nilai k default dengan persentase 78 %. Metode perbandingan MICE mendapatkan nilai persentase rata-rata terendah dari metode lainnya dengan hanya mendapatkan nilai sebesar 63%.

Kata kunci --DAS, Debit, k-NNi, MICE, imputasi, *data hilang*, *statistical validation*

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR ISI

HALAMAN JUDUL ..... **i**
HALAMAN PENGESAHAN ..... **ii**
PERNYATAAN BEBAS PLAGIASI ..... **iii**
PRAKATA ..... **iv**
ARTI LAMBANG DAN SINGKATAN ..... **vi**
ABSTRACT ..... **vii**
INTISARI ..... **viii**
DAFTAR ISI ..... **ix**
DAFTAR GAMBAR ..... **xi**
DAFTAR TABEL ..... **xii**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Perumusan Masalah 5
1.3 Batasan Masalah 5
1.4 Keaslian Penelitian 6
1.4 Tujuan Penelitian 10
1.5 Manfaat Penelitian 10
BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI **11**
2.1 Tinjauan Pustaka 11
2.1.1 Imputasi Data Hilang 11
2.1.2 Penelitian Saat ini 12
2.2 Landasan Teori 14
2.2.1 K- Nearest Neighbours Imputation (k-NNi) 14
2.2.2 Multivariate Imputation by Chained Equation (MICE) 15
2.2.3 Statistical-Validation 17

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

BAB III METODOLOGI 20
3.1 Alat dan Bahan 20
3.1.1 Alat Penelitian 20
3.1.2 Pemrograman R 20
3.1.2 Bahan Penelitian 21
3.1.3 Menentukan Parameter Metode Imputasi Nilai Hilang 27
BAB IV HASIL DAN PEMBAHASAN 33
4.1 Implementasi Imputasi Data Hilang 33
4.1.1 Imputasi Menggunakan Metode MICE 33
4.1.2 Imputasi Menggunakan Metode k-NNi 36
4.1.3 Perbandingan Hasil Imputasi 37
4.1.4 Perbandingan Validasi Statistik k-NNi dan MICE 38
4.2 Kelebihan dan Keterbatasan Penelitian 41
4.2.1 Kelebihan Penelitian 41
4.2.2 Keterbatasan Penelitian 42
BAB V KESIMPULAN DAN SARAN 43
5.1 Kesimpulan 43
5.2 Saran 43
DAFTAR PUSTAKA 45
LAMPIRAN 53
L.1 Data lengkap AWLR Stasiun Bunder 53
L.2 Pengisian Data NA Stasiun Bunder 56

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR GAMBAR

Gambar 2. 1 Langkah dalam konsep MICE [18] 16
Gambar 3. 1 Karakteristik DAS [49] 21
Gambar 3. 2 Peta DAS Opak [48] 22
Gambar 3. 3 *Curva Hidrograf* karakteristik DAS Opak 23
Gambar 3. 4 Diagram Alir Penelitian 25
Gambar 3. 5 Experiment Scenario 27
Gambar 3. 6 Plot kluster dataset Bunder 29
Gambar 3. 7 Hasil diagram 10 kali iterasi 30
Gambar 3. 8 Nilai titik pusat cluster 30
Gambar 4. 1 MICE Margin Plot 34
Gambar 4. 2 Xyplot sebagai inspeksi distribusi original dan imputed data 35
Gambar 4. 3 Pooling modelFit1 36
Gambar 4. 4 Pooling modelFit2 36
Gambar 4. 5 Perbandingan dataset original dengan imputasi dataset 37
Gambar 4. 6 R Squared 38
Gambar 4. 7 RMSE 39
Gambar 4. 8 MAE 40

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR TABEL

Tabel 1. 1 Penelitian saat ini dan sebelumnya 8
Tabel 3. 1 AWLR Description..... 23
Tabel 3. 2 Dataset Attribut 24
Tabel 3. 3 Normalisasi Dataset 26
Tabel 3. 4 Perhitungan jarak *ecluidean* antara cluster *k-Means* dengan data stasiun Bunder..... 31
Tabel 3. 5 Deskripsi Penentuan Parameter k..... 32
Tabel 4. 1 NA Input..... 33
Tabel 4. 2 R2 (R Squared)..... 39
Tabel 4. 3.RMSE (Root Mean Squared Error)..... 40
Tabel 4. 4 MAE (Mean Absolute Error) 41

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

BAB I
PENDAHULUAN

1.1 Latar Belakang

Air adalah sumber daya alam yang hadir dalam berbagai bentuk seperti sungai, sumur, danau dan waduk. Pengembangan sumber daya untuk berbagai keperluan termasuk konsep hidrologi, sumber daya air, dan penanggulangan banjir merupakan dasar bagi pembangunan sosial ekonomi masyarakat [1]. Dampak perubahan iklim juga menjadi faktor pendukung perubahan dinamis dari airdalam wujud benda mulai dari bentuk, warna, debit, dan bau. DAS (Daerah Aliran Sungai) berperan penting dalam perubahan tersebut seperti kegiatan sosial yang meningkat, dan perkembangan tutupan lahan menyebabkan limpasan air yang mengalir dari hulu ke hilir. Hal ini menyebabkan kenaikan muka air di DAS yang sangat signifikan di atas normal. Sehingga mengakibatkan meluapnya air sungai yang dikenal dengan banjir [2].

Kenaikan muka air sungai sangat sulit diprediksi terlebih di Indonesia dengan ribuan sungai yang mengalir pada setiap provinsi. Perkembangan keilmuan multi disiplin memiliki pengaruh positif dalam menjaga sumber daya air di Indonesia. Sumber daya air dan kecerdasan buatan sangat penting karena permintaan air dan pasokan listrik yang besar, kebutuhan irigasi, mitigasi kekeringan dan pengendalian banjir. Untuk merencanakan dan merancang kecerdasan buatan, diperlukan kumpulan data hidrologi yang lengkap dan handal [1]. Masalah paling umum dalam dataset nyata dan analisis statistik adalah data yang hilang dengan persentase nilai yang hilang bervariasi dari satu kumpulan data ke kumpulan data lainnya. Umumnya, kumpulan data berisi persentase berbeda dari nilai yang hilang di setiap kolom [3][4]. Apabila rate data yang hilang kurang dari 1% disebut trivial dan rasio data yang hilang dengan kisaran 1-5% bersifat fleksibel. Metode lanjutan yang diterapkan untuk menangani data yang hilang dengan kisaran 5-15%, dan <15% berdampak sangat besar pada analisis [5].

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Sangat penting untuk diperhatikan bahwa terdapat perbedaan antara nilai kosong dan nilai yang hilang. Nilai kosong berarti tidak ada nilai yang dapat diberikan sedangkan nilai yang hilang berarti nilai aktual untuk variabel itu ada tetapi tidak tersedia atau ditangkap dalam kumpulan data karena beberapa alasan [6]. Pentingnya pendekatan yang tepat terhadap penanganan nilai hilang merupakan syarat mutlak terhadap kondisi sebuah data. Sehingga dibutuhkan beberapa teknik untuk dapat mengakomodasi data hilang dan meminimalkan efek negative terhadap sebuah data [7]. Tiga dari pendekatan yang paling banyak digunakan yang diidentifikasi oleh Little [8] adalah memeriksa kasus yang tidak lengkap, mengganti nilai untuk data yang hilang, dan menyediakan statistik bobot untuk menyelesaikan kasus.

Sebelum menggunakan metode apa pun untuk menangani nilai yang hilang, penting untuk memahami mengapa data hilang [6]. Little dan Rubin [8] dan Rubin [9] memformulasikan tiga kemungkinan mekanisme data yang hilang: Missing Completely at Random (MCAR), Missing at Random (MAR), dan Missing Not at Random (MNAR). MNAR sering dianggap sebagai tipe hilang yang paling buruk, hal ini dikarenakan dapat menyebabkan hasil yang bias sedangkan MCAR dan MAR dapat menyebabkan hilangnya kekuatan statistik [10]. Berbagai jenis data yang hilang ini penting karena menentukan perlakuan statistik mana dari data yang hilang dapat digunakan secara efektif [6].

Istilah imputasi muncul dalam Materi metodologi *"Glossary of Terms on Statistical Data Editing"* yang disusun berdasarkan permintaan negara-negara peserta kegiatan *Statistical Data Editing* yang diselenggarakan oleh *UN/ECE Statistical Division* dalam rangka program kerja *Conference of European Statisticians* [11]. Proses untuk dapat memperbaiki data dengan melakukan imputation perlu dilakukan dengan estimasi untuk mengisi kekosongan data agar kesimpulan statistik menjadi lengkap dan efisien [12]. Hal ini menyangkut topik yang relatif baru, tentang istilah penyuntingan dan imputasi sebagai bagian dari program kerja Konferensi Ahli Statistik Eropa [11].

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Penggunaan imputasi dalam statistic telah digunakan oleh *Statistics Canada Quality Guidelines* yang menyebutkan bahwa imputasi adalah proses yang digunakan untuk menentukan dan menetapkan nilai pengganti untuk data yang hilang, tidak valid, atau tidak konsisten [13]. Dalam Penelitian Donald Bruce Rubin, Imputasi Data merupakan proses memperkirakan data yang hilang dari suatu pengamatan berdasarkan nilai-nilai valid dari variabel lain [9]. Sehingga yang perlu diperhatikan dalam imputasi adalah memastikan metode yang digunakan untuk Imputasi data hilang. Secara spesifik Imputasi Data diklasifikasikan menjadi 2 jenis yaitu Single Imputation dan Multiple Imputation [6].

Dalam penelitian Anil Jadhav dengan data *UCI machine learning repository***,** Single Imputation dengan metode k-NNi (*k-Nearest Neighbors Imputation)* memiliki performa yang cukup signifikan dibanding dengan *Multiple Imputation* [6]. Beberapa mekanisme *single imputation* lain seperti *mean imputation, imputation with distribution,* dan *regression imputation* masih memiliki performa di bawah k-NNi. Sedangkan pada penelitian analisis perbandingan imputasi Cali Curley dengan memanfaatkan data ICSD (*Integrated City Sustainability Database*) *Multiple Imputation* menggunakan metode MICE (*Multivariate Imputation by Chained Equation*) dalam mekanisme *Multiple imputation* yang memungkinkan analisis dengan ukuran sampel yang lebih besar, bias yang lebih sedikit, dan kemampuan untuk menginterpretasikan data seolah-olah data tersebut tidak hilang [7].

Pada penelitian yang dilakukan Peter Schmitt [14], nilai RMSE (Root *Mean Squared Error*) dari MICE didasarkan pada algoritme yang jauh lebih kompleks dan perilakunya tampaknya terkait dengan ukuran kumpulan data yaitu cepat dan efisien pada kumpulan data kecil. Akan tetapi, kinerja sedikit menurun dan time estimated meningkat ketika diterapkan pada kumpulan data besar. Sedangkan metode k-NNi menghasilkan nilai yang stabil meningkat berdasarkan ukuran data yang terkecil hingga ukuran data besar.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Penelitian ini akan fokus menggunakan beberapa *package* pada pemrograman R. *Package* MICE sangat compatible dengan menggunakan pemrograman R [15], sedangkan k-NNi dapat dilakukan dengan menggunakan paket VIM di pemrograman R. Detail tentang paket VIM dijelaskan oleh Kowarik dan Templ yang menyebutkan bahwa Paket VIM dikembangkan untuk mengeksplorasi dan menganalisis struktur nilai yang hilang dalam data menggunakan metode visualisasi. Hal ini diuperlukan untuk menghubungkan nilai yang hilang dengan metode imputasi bawaan dan memverifikasi proses imputasi menggunakan alat visualisasi, serta untuk menghasilkan grafik berkualitas tinggi sebagai bahan publikasi [16].

Manfaat yang diharapkan dari penelitian ini adalah mampu memberikan pengetahuan tentang mengisi data yang hilang sebagai bahan untuk menyusun prosedur Preprocessing yang menangani anomali beberapa data dalam suatu dataset. Hal ini dilakukan karena pendekatan konvensional memperlakukan setiap anomali data sebagai kasus yang terisolasi, tetapi anomali data juga terjadi pada data. Metode MICE dan k-NNi menjadi fokus penelitian yang akan dikembangkan dengan data DAS Sungai Opak dikarenakan memiliki kategori data numeric dan timeseries dengan ukuran data sedang.

Di samping pengaruh pengolahan data terhadap hasil analisis statistik belum banyak diteliti, maka perlu dilakukan penelitian di bidang ini untuk menghasilkan hasil yang akurat terutama dalam pengisian data yang hilang. Akibatnya cakupan yang lebih luas dari penggabungan imputation gap dan data nilai yang hilang berdampak pada tingkat akurasi dalam penerapan prosedur *preprocessing*. Untuk itu diperlukan penerapan prosedur Imputation Missing data agar data anomali, bias dan *noise* dapat diminimalisir dan dapat memberikan perbandingan untuk menemukan metode yang paling tepat pada kasus DAS Opak di Provinsi D.I Yogyakarta. Hasil perbandingan tersebut dapat menjadi acuan untuk melengkapi data yang hilang pada beberapa stasiun AWLR (*Automatic Water Level Recorder*).

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

1.2 Perumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan pada bagian 1.1 Latar Belakang, perumusan masalah pada penelitian ini adalah :

1. Mengidentifikasi karakteristik data sumber daya air khususnya Data Daerah Aliran Sungai Opak berdasarkan karakteristik DAS.
2. Melakukan identifikasi ketidaklengkapan data yang dikumpulkan oleh beberapa stasiun pemantauan aliran debit air di Daerah Aliran Sungai Opak. Dikarenakan beberapa metode konvensional pra-pemrosesan nilai hilang hanya dengan menghilangkan data, maka dilakukan identifikasi penerapan metode algoritme imputasi data hilang dalam melakukan pengisian data hilang dengan mempertimbangkan akurasi imputasi pada tahap pra-pemrosesan sebelum menggunakan data untuk pengelompokan pada langkah analisis dari stasiun yang dipilih dari dataset.
3. Membandingkan hasil setiap metode dengan perbandingan validasi statistik

1.3 Batasan Masalah

Pemberian batasan masalah pada penelitian ini bertujuan untuk memudahkan jalannya penelitian. Batasan masalah pada penelitian ini yaitu sebagai berikut:

A. Penelitian ini berfokus kepada imputasi data hilang dengan pada data debit DAS Opak dari 5 stasiun Automatic Water Level Recorder (AWLR) pada Provinsi DI. Yogyakarta menggunakan metode *K-Nearest Neighbors Imputation* (k-NNi) dan *Multivariate Imputation by Chained Equation* (MICE).

B. Data Stasiun Bunder DAS Opak merupakan data yang digunakan untuk menganalisis metode imputasi dikarenakan memiliki data yang lengkap dengan presentase nilai hilang sebesar 0%.

C. Pada tahap percobaan pengisian data yang hilang pada penelitian ini menggunakan bahasa pemrogaman R dengan aplikasi *integrated development*

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

environment (IDE) *Rstudio*.

D. Untuk mengetahui apakah model yang dirancang berfungsi dengan baik proses perbandingan pada penelitian ini menggunakan *validation* agar dapat menilai hasil statistik analisis imputation.

1.4 Keaslian Penelitian

Keaslian penelitian bertujuan menunjukkan perbedaan antara penelitian ini dan penelitian sebelumnya pada area penelitian yang sama.Terdapat beberapa penelitian yang sudah melakukan upaya imputation data hilang. Berbagai metode baik *machine learning* maupun *deep learning* telah diaplikasikan oleh beberapa peneliti di dunia terkait data hilang. Namun, yang menggunakan pilihan kursus yang sama kami menemukan beberapa penelitian yang fokus pada imputasi nilai data hilang. atau perhitungan multi disiplin ilmu untuk lebih mengurangi bias dan inefisiensi. Pada Tabel 1.1 menunjukkan studi tentang imputasi data hilang dan studi kasus metode terhadap dataset yang digunakan.

Pada penelitian Doreswamy, Ada tiga kategori metode dalam menangani data yang hilang yaitu penghapusan kasus, estimasi parameter, dan teknik imputasi [17]. Penghapusan kasus nilai hilang merupakan metode yang paling umum digunakan dengan menghapus setiap baris yang berisi nilai hilang. Sedangkan untuk metode estimasi parameter lebih efisian dibanding dengan metode penghapusan karena dapat menggunakan semua data yang tersedia dalam dataset. Kemudian untuk metode ketiga yaitu teknik imputasi yang merupakan sebuah prosedur untuk mengisi data yang hilang dengan yang diprediksi berdasarkan nilai yang tersedia dalam dataset [18][19].

Beberapa metode yang diusulkan oleh Kamwanga pada data limpasan di sungai Little Ruaha Tanzania, menunjukkan penelitian tersebut dapat memberikan estimasi nilai yang baik. Namun, hasil ini hanya dapat dicapai jika ada korelasi yang cukup tinggi antara stasiun donor dan stasiun yang membutuhkan pengisian [1]. Selain itu, studi harus diperluas pada metode lain untuk secara spesifik dapat memilih metode yang paling cocok untuk mengisi celah yang terjadi selama

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

periode tertentu dalam satu tahun hidrologi.

Metode pengisian data hilang merupakan hal krusial sebelum proses analisis data. Penelitian yang dilakukan oleh Murti D.M.P, mengusulkan penelitian menggunakan metode k-NNi dengan pertimbangan klasifikasi kualitas dari data training [20]. Berdasarkan hasil dan pembahasan pada penelitian ini, dapat disimpulkan bahwa metode imputasi k-NN mampu menjadi salah satu pendekatan yang dapat digunakan ketika dataset hilang. Hal ini terlihat dari performa metode k-NN yang diukur berdasarkan nilai akurasi dimana akurasi hasil imputasi mendekati akurasi data lengkap dengan model missing data yang berbeda [20].

Penelitian yang dilakukan oleh Cali Curley dalam studi perkotaan dengan membandingkan dan mengevaluasi tiga pendekatan yang umum digunakan untuk menangani data yang hilang yaitu penghapusan data hilang, *single imputation*, dan *multiple imputation*. Hasilnya menunjukkan manfaat menggunakan pendekatan terhadap data yang hilang berdasarkan *multiple imputation* dapat dilakukan dengan baik [7].

Salah satu metode *multiple imputation* yang cukup sering digunakan adalah MICE. Geeta Chabra mengemukakan penggunaan metode MICE merupakan Analisis kami menegaskan bahwa kekuatan MICE terletak pada *standart error* yang lebih kecil dan interval waktu eksekusi yang lebih baik [21]. Geeta menambahkan bahwa Metode MICE selanjutnya dapat dikombinasikan dengan *machine learning* dan Algoritme Genetika pada kumpulan data.

Dalam penelitian terdahulu proses imputasi nilai hilang dapat dilakukan dengan baik. Imputasi nilai hilang dengan metode *multiple imputation* dan k-NNi memiliki keberhasilan yang cukup baik. Namun pada semua penelitian tidak membandingkan antara kedua metode antara multiple imputation dan k-NNi. Pada penelitian ini nantinya akan memberikan kontribusi dengan menunjukkan keberhasilan imputasi nilai hilang dengan membandingkan terlebih dahulu pada karakteristik dataset dan kemudian membandingkan kedua metode yang telah dilakukan pada penelitian terdahulu.

METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Tabel 1. 1 Penelitian saat ini dan sebelumnya

Judul	Tahun	Peneliti	Deskripsi	Data dan Instrumen	Missing Value	MICE	k-NNi	Validasi dan Perbandingan
Performance Evaluation of Predictive Models for Missing Data Imputation in Weather Data	2017	Doresmwamy [17]	Melakukan imputasi data hilang menggunakan dataset Cuaca dengan beberapa metode dan teknik baru pada *machine learning.*	NCDC (National Climatic Data Center) Dataset	√	×	√	*5-Fold Cross Validation*
Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha catchment Tanzania	2018	Kamwanga [1]	Melakukan pengisian data menggunakan metode empiris dan regresi pada data curah hujan berdasar limpasan sungai	Data stasiun penangkar curah hujan little Ruaha Tanzania	√	×	×	*Nash-Sutcliffe efficiency coefficient (NSE), Coefficient of determination (R2)* dan *standard error of estimate (SE)*
K-Nearest Neighbor (K-NN) based Missing Data Imputation	2019	Murti [20]	Metode K-NN sebagai imputasi dilakukan pada beberapa kasus yang memiliki mekanisme yang berbeda dan model data yang hilang.	Data sekunder dari salah satu situs Journal Ranking, SCImago Journal Rank (SJR) menggunakan	√	×	√	*confusion matrix*
A Review on Missing Data Value Estimation Using Imputation Algorithm	2019	Geeta Chhabra [15]	Untuk menganalisis metode imputasi yang mudah digunakan, lebih akurat, efisien dan menghasilkan estimasi yang tidak bias	*Literature review*	√	×	√	*Advantage* dan *limitation*

METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database	2019	Cali Curley [7]	Penelitian tentang perawatan data yang hilang dalam penelitian studi perkotaan dengan metode perbandingan dan evaluasi	*2nd gen Integrated City Sustainability Database* (ICSD)	√	√	×	Variable tak bebas menggunakan Policy Index dan variabel independen menggunakan Standart Error of Estimate (SE)
Comparison of Performance of Data Imputation Methods for Numeric Dataset	2019	Anil Jadhav [6]	membandingkan secara komprehensif metode imputasi data.	*UCI Machine Learning Repository*	√	×	√	*Normalized Root Mean Square Error* (NRMSE)
A Comparison of Multiple Imputation Methods for Data with Missing Values	2017	Geeta Chhabra [21]	mengeksplorasi Multiple Imputasi menggunakan MICE melalui pemeriksaan kumpulan data sampel.	dataset iris dari *UCI Machine Learning Repository*	√	√	×	*Mean Standard Error* dan *Mean Confidence Interval Length*
A Comparison of Six Methods for Missing Data Imputation	2015	Peter Schmitt [14]	Memperkenalkan elemen ambiguitas ke dalam analisis data dan mempertimbangkan dengan tepat untuk memberikan analisis yang efisien dan valid.	*Iris* dataset, *e. coli UCI machine learning repository,* dan *Breast cancer dataset*	√	√	×	*Root mean square error* (RMSE), *Unsupervised classification erro*r (UCE), *supervised classification erro*r (SCE) dan *time execution*

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

1.4 Tujuan Penelitian

Tujuan dari penelitian Metode Imputasi Data Hilang pada Debit Daerah Aliran Sungai dengan Studi Kasus Daerah Aliran Sungai Opak, Provinsi DI Yogyakarta ini adalah :

1. Memberikan kontribusi dengan melakukan metode imputasi data yang hilang pada dataset DAS Opak.
2. Memberikan perbandingan metode terbaik dengan menggunakan metode imputasi pada data hilang.

1.5 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitan ini adalah sebagai berikut:

1. Memberikan solusi pengisian data hilang sebagai sarana dalam pengembangan preprocessing dan normalisasi dataset
2. Dapat menghilangkan Bias dan anomaly dalam sebuah dataset.
3. Memberikan sarana bantuan terhadap penerapan penelitian dibidang keilmuan multi disiplin seperti Ilmu geografi, Ilmu Sipil, dan Ilmu Lingkungan Hidrologi Air.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

2.1.1 Imputasi Data Hilang

Data yang hilang dapat sangat mengganggu kualitas dan utilitas data. Masalah kesenjangan dalam seri data dapat diselesaikan secara teoritis dengan melengkapi catatan aliran harian dari data yang ada di stasiun pengukuran terdekat, baik hulu atau hilir aliran air yang sama (misalnya, teknik interpolasi), meskipun pemilihan stasiun donor mungkin sangat penting. faktor yang mempengaruhi hasil[22].

Dalam penelitian di Tanzania oleh Kamwanga, menunjukkan bahwa terdapat pendekatan umum yang sering diadaptasi dalam menangani kesenjangan data yang hilang [1]. Cara pertama adalah hanya menggunakan catatan data secara kontinyu dan mengabaikan peristiwa sebelumnya. Hal ini berdasarkan kepada sebuah peristiwa atau kejadian sebelum data set tersedia dengan melakukan asumsi bahwa data adalah satu seri catatan yang berkelanjutan. Cara selanjutnya adalah menghapus waktu hilangnya data dan menganggap data yang tersisa sebagai kumpulan data berkelanjutan.

Engels et all [23] memberikan beberapa jenis teknik mean imputation techniques yang diperkenalkan pada metode *missing value imputation*. Nilai yang hilang dilakukan perhitungan menggunakan setiap metode dan dibandingkan dengan nilai yang diamati.

Pada *literature review* yang dilakukan oleh Anil Jadhav [6], terdapat strategi untuk mengatasi nilai hilang. Kedua strategi tersebut adalah mengabaikan atau menghapus bagian yang hilang, dan strategi kedua adalah menggunakan *missing value imputation*. Pada teknik strategi mengabaikan nilai hilang banyak digunakan dan cenderung menjadi metode populer untuk menangani data yang hilang. Masalah serius dengan metode ini adalah mengurangi ukuran dataset. Hal ini akan mempengaruhi ketika jumlah dataset memiliki sejumlah kecil nilai yang

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

hilang. Ada dua pendekatan umum untuk mengabaikan data yang hilang: penghapusan listwise (penghapusan kasus atau analisis kasus lengkap) dan penghapusan berpasangan (analisis kasus yang tersedia) [6]. Sedangkan untuk strategi kedua adalah imputasi nilai hilang. Imputasi nilai hilang bertujuan untuk memberikan estimasi parameter populasi yang akurat sehingga kekuatan data mining dan teknik analisis atau jumlah data tidak berkurang [6]. Pada penelitian ini hasil analisis menunjukkan bahwa metode imputasi k-NN mengungguli metode lainnya.

Salah satu metode *Machine Learning* yang dikembangkan untuk menangani data yang hilang adalah k-NN [20]. Metode ini menggunakan jarak antar data training sebagai klasifikasi untuk melakukan testing [24]. k-NN merupakan metode yang fleksibel baik dalam data kontinyu maupun data diskrit [25]. Metode ini dapat digunakan karena pada beberapa data filler yang hilang[26], tidak memerlukan model prediksi untuk setiap atribut[27]. Sedangkan salah satu kelemahan dari metode k-NN adalah memakan waktu yang sangat tinggi dalam menganalisis dataset besar karena metode ini mencari data yang serupa di semua dataset, selain itu akurasi k-NN dapat sangat terdegradasi dengan data berdimensi tinggi karena ada sedikit perbedaan antara tetangga terdekat dan terjauh[3].

2.1.2 Penelitian Saat ini

Mempertimbangkan keterbatasan dan masalah-masalah yang dihadapi dalam imputasi nilai hilang dalam suatu dataset, pendekatan untuk imputasi dengan metode yang tepat menjadi tantangan yang harus diselesaikan. Penelitian ini, mengusulkan dengan menggunakan 2 metode utama yaitu MICE dan k-NNi. Metode MICE dalam penelitian Cali Curley memiliki pertimbangan dalam penggunaan dataset *2nd Generation Integrated City Sustainable Database* yang memiliki sedikit kemiripan dengan dataset DAS OPAK [7]. Kemiripan tersebut berdasarkan nilai hilang cukup besar dengan komposisi diatas 15% pada klasifikasi MAR dan MCAR.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Kemudian pada penelitian yang dilakukan oleh Murti [20] dan Doreswamy [17], metode k-NN memiliki hasil performa yang baik dalam pendekatan nilai hilang. Penggunaan dataset *Journal Ranking Scimago Dataset* dan *NCDC (National Climatic Data Center) Dataset* imputasi nilai hilang dapat dilakukan menggunakan kedua dataset tersebut dengan baik.

Berdasarkan beberapa penelitian terdahulu, penggunaan metode MICE dan k-NNi menjadi pokok acuan dalam metode yang digunakan. Hal ini dikarenakan metode MICE dan k-NNi dapat melakukan imputasi dengan baik pada persentase nilai hilang diatas 15% ataupun lebih kecil dari 15%. Selain itu, metode k-NNi merupakan metode yang dapat dikombinasikan dengan gabungan algoritma lain untuk menentukan nilai parameter k.

Berdasarkan kajian yang dilakukan terhadap beberapa penelitian di atas, dapat ditarik kesimpulan sebagai berikut:

1. Pemilihan metode-metode yang diterapkan oleh peneliti-peneliti pada paper yang menjadi acuan untuk penelitian ini dilakukan tanpa penanganan pada penentuan parameter nilai k. Penentuan nilai k dianggap perlu karena dapat memberikan akurasi nilai imputasi pada suatu data dengan nilai hilang dengan persentase diatas 15%.
2. Penggunaaan dataset dengan nilai hilang tidak bervariasi pada penelitian paper yang menjadi acuan menjadi dasar penelitian ini untuk membandingkan beberapa jenis variasi nilai hilang pada dataset DAS Opak.
3. Melakukan perbandingan dari hasil imputasi pada tiap metode yang digunakan yaitu MICE dan k-NNi menggunakan validasi statistik. Validasi statistik akan mampu n memeriksa efektifitas model machine learning.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

2.2 Landasan Teori

2.2.1 K- Nearest Neighbours Imputation (k-NNi)

k-NN pertama kali diusulkan oleh Fix [28] pada tahun 1951 dan dikembangkan oleh Cover dan Hart [29]. Algoritme k-NN adalah sebuah algoritme dalam melakukan klasifikasi terhadap objek berdasarkan pada data pembelajaran yang jaraknya paling dekat dengan mencari kelompok objek tetangga yang terdekat [30].

Dalam k-NN, untuk melakukan prediksi data yang hilang menggunakan rata-rata dari contoh data yang memiliki kemiripan[24]. Kemiripan yang terjadi diambil dengan rata – rata fungsi jarak antara stasiun AWLR hulu dan hilir. Dengan pertimbangan tersebut data yang hilang di lakukan pembanding dengan stasiun sebelumnya.

k-NN melakukan pendekatan yang berbeda dengan melakukan estimasi dan asumsi harus memiliki hasil keluaran yang serupa dengan data dengan fungsi jarak sebelumnya. Dengan demikian, tingginya resistensi terhadap mekanisme dan model data yang hilang menjadikan metode k-NNi sebagai salah satu pendekatan penanganan data yang hilang [31]. Hambatan ini juga dilihat dengan imputasi menggunakan parameter k yang berbeda. *Neighbours* dari titik data yang ditanyakan harus digunakan untuk menentukan nilainya dan bukan titik yang jauh. Dalam kasus regresi, nilai *k-Neighbours* dipilih dan rata-ratanya dianggap mengatur nilai titik yang ditanyakan[17]. Secara default p = 2 dengan demikian menjadi *Euclidean Distance*.

$$\boldsymbol{d_{(x,y)}} = \sqrt{\sum_{\boldsymbol{j=1}}^{\boldsymbol{s}} (\boldsymbol{x_j} - \boldsymbol{y_j})^{\boldsymbol{2}}} \qquad (1)[20]$$

$d_{(x,y)}$: *Euclidian distance*,

j : atribut datadengan *j*=1,2,3,.....*s*,

s : dimensi data,

x_{aj} : nilai dari j- attribute berisi data yang hilang,

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

y_{bj} : nilai selain j- attribute berisi data yang lengkap,

Asumsi inti yang dibuat dengan metode k-NN adalah bahwa contoh dengan vektor fitur serupa harus memiliki keluaran yang serupa. Tetangga dari titik data yang ditanyakan harus digunakan untuk menentukan nilainya dan bukan titik yang jauh. Dalam kasus regresi, nilai k tetangga dipilih dan meannya dianggap menetapkan nilai titik yang ditanyakan[17].

k-NN termasuk dalam algoritme *non-parametric* yang merupakan tidak ada parameter atau jumlah parameter tetap terlepas dari besaran ukuran data. Sebuah sampel pada sebuah dataset diklasifikasikan berdasarkan suara terbanyak (*majority voting)* dari tetangganya (*neighbor*) [32]. Sampel tersebut akan diklasifikasikan pada kelas mayoritas di antara tetangga terdekat k yang diukur dengan fungsi jarak. Jika k = 1, maka secara sederhana, sampel tersebut diklasifikasikan ke dalam kelas tetangga terdekatnya.

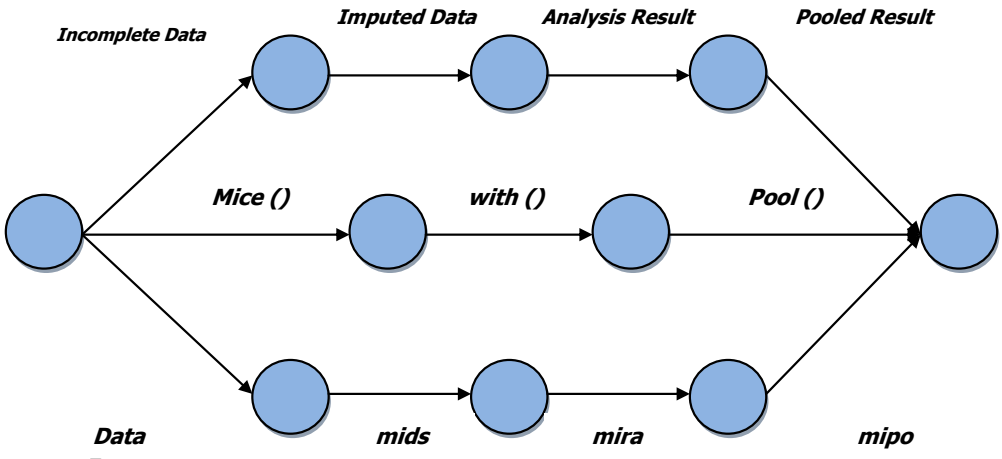
Rumus jarak yang digunakan pada algoritme k-NN ada banyak yakni: *Euclidean Distance, Standardized Euclidean Distance, Mahalanobis Distance, City Block Distance, Minkowski Distance, Cosine Distance, Correlation Distance, Hamming Distance, Jaccard Distance dan Spearman Distance.* Dari kesemua rumus jarak (*distance metric)* yang digunakan, *Euclidean Distance* merupakan rumus jarak yang paling banyak digunakan. Ditambah lagi, *default setting* untuk parameter *metric* yang ada pada *library* scikit-learn untuk algoritme k-NN adalah *Euclidean Distance* (*Minkowski Distance* dengan nilai p=2).

2.2.2 Multivariate Imputation by Chained Equation (MICE)

MICE dikenal juga dengan “Fully Conditional Spesification” atau “Sequential Regression Multiple Imputation” yang digunakan dalam acuan statistika sebagai salah satu metode untuk menangani *Missing data* [33]. Pada Gambar 1 mekanisme Multiple Imputation dengan metode MICE terbagi menjadi 3 tahap yaitu data imputasi, data analisis, dan pooling [34].

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Pada tahap pertama yaitu data imputasi merupakan tahap dimana dataset yang dilakukan imputasi dari distribusi yang menghasilkan dataset lengkap. Kemudian pada tahap data analisis adalah tahap dimana data hilang telah terisi dengan nilai yang mendekati nilai asli. tahap selanjutnya adalah pooling dimana output yang diperoleh setelah analisis data dikumpulkan untuk mendapatkan hasil akhir menggunakan aturan sederhana terhadap mekanisme Multiple Imputasi [21].



Gambar 2. 1 Langkah dalam konsep MICE [18]

Metode Multiple Imputation digunakan untuk mengganti nilai yang hilang dengan probabilitas nilai yang tepat. Kumpulan data yang tidak lengkap kemudian diubah menjadi kumpulan data yang lengkap dengan menggunakan metode imputasi yang kemudian dapat dianalisis dengan metode analisis standar apa pun. Oleh karena itu, Multiple Imputasi dengan metode MICE dapt dilakukan dengan baik untuk menangani data yang hilang[21].

MICE dapat digunakan untuk berbagai model data, seperti data kontinyu, data biner (regresi logistik), data kontinu 2-level, regresi logistik polikotomus, dan odds proporsional [33]. Prosedur MICE mengikuti serangkaian model regresi, dimana masing-masing variabel dari data yang hilang dimodelkan bersyarat pada variabel lain dalam data tersebut [35]. Hal ini menunjukkan bahwa setiap variabel dapat dimodelkan menurut distribusinya.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Pada tahap pertama yaitu data imputasi merupakan tahap dimana dataset yang dilakukan imputasi dari distribusi yang menghasilkan dataset lengkap. Kemudian pada tahap data analisis adalah tahap dimana data hilang telah terisi dengan nilai yang mendekati nilai asli. tahap selanjutnya adalah pooling dimana output yang diperoleh setelah analisis data dikumpulkan untuk mendapatkan hasil akhir menggunakan aturan sederhana terhadap mekanisme Multiple Imputasi [21].

Metode Multiple Imputation digunakan untuk mengganti nilai yang hilang dengan probabilitas nilai yang tepat. Kumpulan data yang tidak lengkap kemudian diubah menjadi kumpulan data yang lengkap dengan menggunakan metode imputasi yang kemudian dapat dianalisis dengan metode analisis standar apa pun. Oleh karena itu, Multiple Imputasi dengan metode MICE dapt dilakukan dengan baik untuk menangani data yang hilang[21].

MICE dapat digunakan untuk berbagai model data, seperti data kontinyu, data biner (regresi logistik), data kontinu 2-level, regresi logistik polikotomus, dan odds proporsional [33]. Prosedur MICE mengikuti serangkaian model regresi, dimana masing-masing variabel dari data yang hilang dimodelkan bersyarat pada variabel lain dalam data tersebut [35]. Hal ini menunjukkan bahwa setiap variabel dapat dimodelkan menurut distribusinya.

2.2.3 Statistical-Validation

Tantangan terbesar pada pada proses *Machine Learning* adalah membuat membuatnya berfungsi secara akurat pada data yang tidak terlihat. Untuk mengetahui apakah model yang dirancang berfungsi dengan baik atau tidak, kita harus mengujinya terhadap titik-titik data yang tidak ada selama pelatihan model[35]. Salah satu teknik terbaik untuk memeriksa efektifitas model pembelajaran mesin adalah teknik Validasi Statistik yang dapat dengan mudah diterapkan dengan menggunakan bahasa pemrograman R. Pada statistical-validation nantinya akan memberikan hasil validasi dengan metode akurasi *Root Mean Squared Error* (RMSE), *R-squared* (R^2), dan *Mean Absolute Error* (MAE).

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Alasan penggunaan R^2, RMSE, dan MAE adalah mempertimbangkan bahwa penelitian ini bersifat prediksi. Ketepatan prediksi adalah suatu hal yang penting untuk peramalan atau *forecasting*, yaitu bagaimana mengukur kesesuaian antara data yang yang sudah ada dengan data peramalan [36]. Berikut penjelasan mengenai R^2, RMSE, dan MAE.

1) RSquared (R^2)

RSquared (R^2) adalah ukuran persentase variasi total dalam variabel dependen yang diperhitungkan oleh variabel independen [37]. R^2 sebesar 1,0 menunjukkan bahwa data sangat cocok dengan model linier. Apabila nilai R^2 yang kurang dari 1,0 menunjukkan bahwa setidaknya beberapa variabilitas dalam data tidak dapat diperhitungkan oleh model. Sebagai contoh, R2 0,5 menunjukkan bahwa 50% variabilitas dalam data hasil tidak dapat dijelaskan oleh model atau metode tertentu. Berikut nilai R^2 dapat diturunkan secara matematis dengan rumus [37].

$$\boldsymbol{R^2} = 1 - \frac{SSE}{SS_{yy}} \qquad (3)[37]$$

R^2 : RSquared coefficient of determination

SSE : Sum of squares of residuals

SS_{yy} : Total sum of squares

2) Root mean square error (RMSE)

Root mean square error (RMSE) merupakan mengukur perbedaan antara nilai yang diperhitungkan dan nilai yang sebenarnya. Pada dasarnya, ini mewakili standar deviasi sampel dari perbedaan antara data sebelum dan sesudah dilakukan imputasi [14]. Di bawah ini merupakan rumus dari RMSE.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\boldsymbol{X}_i^{obs} - \boldsymbol{X}_i^{imputed}\right)^2}{n}} \qquad (2)[14]$$

RMSE : Root mean square error

$\boldsymbol{X}_i^{obs}$: Corresponding prediction

$\boldsymbol{X}_i^{imputed}$: i[th] measurement

n : The number of data points

3) Mean Absolute Error (MAE)

Mean Absolute Error (MAE) adalah nilai mutlak dari selisih antara nilai prakiraan dengan nilai sebenarnya [38]. Untuk evaluasi model peramalan, MAE lebih intuitif dalam memberikan rata – rata error dari keseluruhan data [39]. Berikut untuk rumus matematis dari MAE.

$$\boldsymbol{MAE} = \frac{\sum_{n=1}^{N} |\check{r}_n - r_n|}{N} \qquad (4)[40]$$

MAE : Mean Absolute Error

$\check{r}_n$: Prediction rating

r_n : The true rating in testing data set

Statistical-validation digunakan sebagai validasi model untuk menilai hasil statistik analisis imputation dengan menggeneralisasi kumpulan data independent [41][42]. Menurut Shao dan Rao Wu Statistical-validation dapat memberikan wawasan teoritis dan teori asimtotik pemilihan model dengan sejumlah variabel tetap sebagai model linier [43][44].

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

BAB III
METODOLOGI

3.1 Alat dan Bahan

Alat dan bahan yang digunakan selama melakukan penelitian ini dan penulisan laporan sebagai berikut:

3.1.1 Alat Penelitian

Pengembangan algoritme *Missing value imputation* menggunakan bahasa pemrograman *r* dan *RStudio* sebagai *interface*. Alat penunjang lainnya yang digunakan dalam penelitian ini terbagi menjadi dua macam, yaitu perangkat keras *(hardware)* dan perangkat lunak *(software).*

- Perangkat keras *Notebook,* dengan spesifikasi minimum:
  - *Intel core i3*
  - *RAM 4 GB.*
- Perangkat Lunak
  - *R x64 4.0.4*
  - *RStudio*
  - *Google Chrome*
  - *Windows 10*

3.1.2 Pemrograman R

Perangkat lunak pemrograman "R" adalah OSS *(open source software)* untuk pengolahan data dan analisis statistik berbasis bahasa program dan merupakan perangkat lunak yang menggunakan GUI *(grafik user interface).* Pemrograman "R" muncul setelah "S" dan "S plus" software berkembang [45]. R merupakan perangkat lunak yang digunakan untuk manipulasi data, perhitungan, simulasi, penayangan grafik, dan sekaligus sebagai bahasa pemograman yang bersifat *interpreter*. R diturunkan dari bahasa S, suatu bahasa pemograman yang dikembangkan di Laboratorium Bell [46]. Software ini sudah digunakan dan dikembangkan oleh banyak volunter di berbagai penjuru dunia. Comprehensive R Archive Network (CRAN) adalah sumber informasi internet utama, mulai dari file

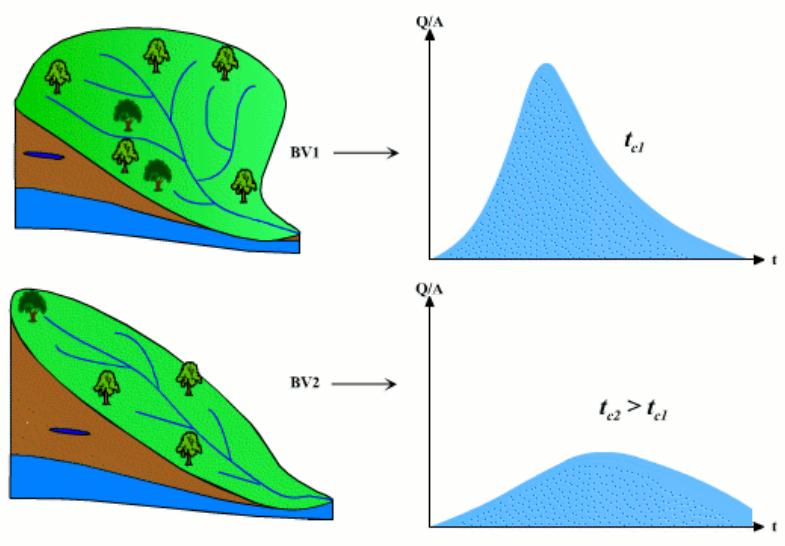
UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

untuk instalasi, sumber kode, dokumentasi, *Newsletter, mailing list*, hingga paket "R" [46].

R menjadikan bahasa pemograman komputer yang memungkinkan pengguna untuk melakukan pemrograman algoritme dan menggunakan alat yang telah dikembangkan melalui R oleh pengguna lain. R merupakan bahasa pemograman tingkat tinggi dan juga merupakan lingkungan untuk analisis data dan grafik [47].

3.1.2 Bahan Penelitian

Proses alur penelitian diawali dengan tinjauan pustaka dan pengambilan data dari pihak terkait dalam hal ini dari SISDA Balai Besar Sumber Daya Air Progo-Serayu-Opak Kementrian Pekerjaan Umum dan Perumahan Rakyat. Penelitian ini fokus pada *missing value imputation* dari data debit DAS Opak di Provinsi Daerah Istimewa Yogyakarta dengan area cakupan seluas 2,9 km2 yang melintasi 2 Kabupaten dan 1 Kota di DI Yogyakarta dengan panjang aliran sejauh 62,83 km [48].

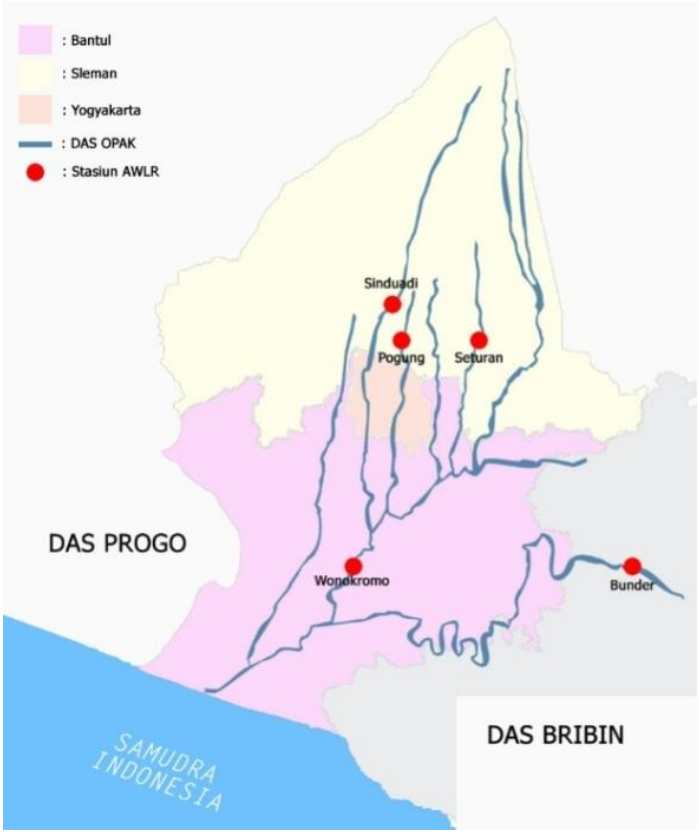


Gambar 3. 1 Karakteristik DAS [49]

Bentuk DAS akan mempengaruhi bentuk hidrograf karakteristiknya. Pada Gambar 3.1, DAS bentuk memanjang menjadikan DAS dengan curah hujan yang sama, namun memiliki aliran yang keluar lebih rendah, karena waktu konsentrasi

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

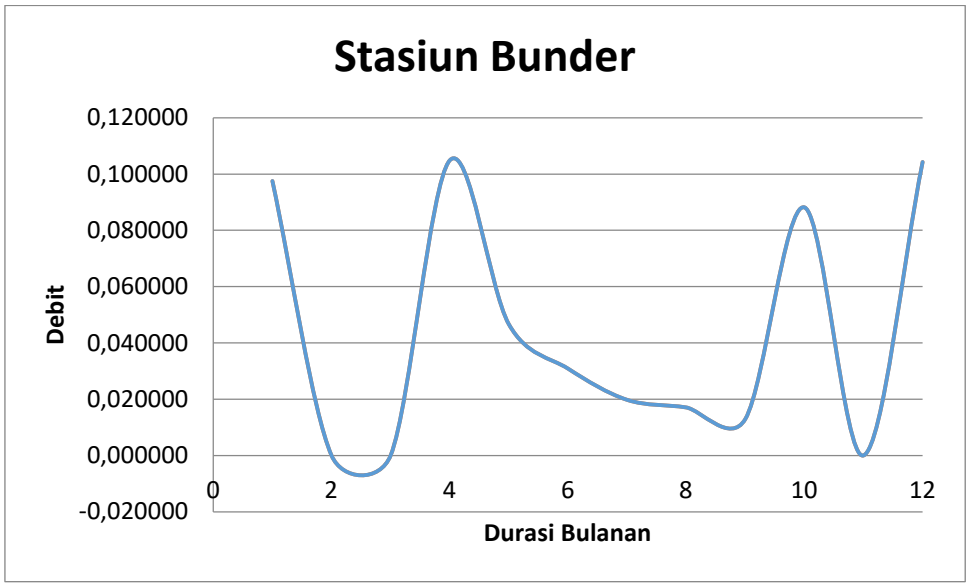
lebih tinggi. Sedangkan DAS OPAK yang berbentuk kipas memberikan waktu konsentrasi yang lebih rendah, dan menghasilkan aliran yang lebih tinggi [49]. DAS Opak merupakan DAS dengan karakteristik berbentuk kipas seperti yang ditunjukkan pada Gambar. 3.2.



Gambar 3. 2 Peta DAS Opak [48]

Hidrograf aliran terdiri dari tiga komponen, yaitu: (1) sisi naik (*rising limb*), (2) bagian puncak (*crest*), (3) sisi resesi (*recession limb*). Pada Gambar 3.3 merupakan curva hidrograf karakteristik DAS Opak pada tahun 2008 dan 2010. Sisi naik menandakan masih adanya kontribusi hujan terhadap debit aliran. Puncak hidrograf adalah debit maksimum yang terjadi dalam suatu aliran dengan waktu naik yang merupakan selang waktu antara mulai bertambahnya aliran sampai tercapainya debit puncak. Sisi turun merupakan proses penurunan intensitas daerah tangkapan. Waktu dasar yaitu waktu mulai bertambahnya debit aliran sampai kembali ke debit aliran dasar [49].

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/



Gambar 3. 3 *Curva Hidrograf* karakteristik DAS Opak

Pengambilan data rerata bulanan berdasar kepada ketersediaan pengaturan AWLR pada stasiun pengambilan sampel. Data yang diambil pada dari SISDA Balai Besar Sumber Daya Air Progo-Serayu-Opak merupakan data *private* yang digunakan oleh Kementrian Pekerjaan Umum dan Perumahan Rakyat untuk menghitung probabilitas debit andalan. Pada Tabel 3.1 menunjukkan aliran dan letak koordinat dari DAS Opak beserta SubDAS sungai yang terdapat di DAS Opak.

Tabel 3. 1 AWLR Description

Stasiun AWLR	Sungai	Daerah Adm.	Koordinat	Hulu Sungai	Data Hilang
Stasiun Bunder	Sungai Oyo	Gunung Kidul	-7.896006, 110.513925	Gajah Mungkur	0
Stasiun Pogung	Sungai Code	Sleman	-7.759429, 110.370078	Boyong	4
Stasiun Sinduadi	Sungai Winongo	Sleman	-7.748396, 110.357493	Denggung, Doso	5
Stasiun Seturan	Sungai Tambakbayan	Sleman	-7.747839, 110.357912	Opak	17
Stasiun Wonokromo	Sungai Gajahwong	Bantul	-7.866613, 110.394481	Lereng Merapi	25

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Dataset tersebut merupakan data yang diambil langsung dari Sistem Informasi Daerah (SISDA) Unit Balai Besar Wilayah Sumber Daya Air (BBWSDA) Provinsi DI Yogyakarta. Data debit DAS Opak di Provinsi Daerah Istimewa Yogyakarta merupakan dataset *timeseries* yang diambil dari 5 stasiun AWLR DAS Opak dengan durasi waktu dataset dimulai dari Januari 2007 hingga Desember 2017 Dari kelima stasiun terdapat 1 stasiun AWLR yang memiliki ratio 0% Missing Data dan keempat stasiun lainya berada pada rasio 3-19% Missing Data. Pada Tabel 3.2 menunjukkan atribut dari dataset 5 stasiun AWLR.

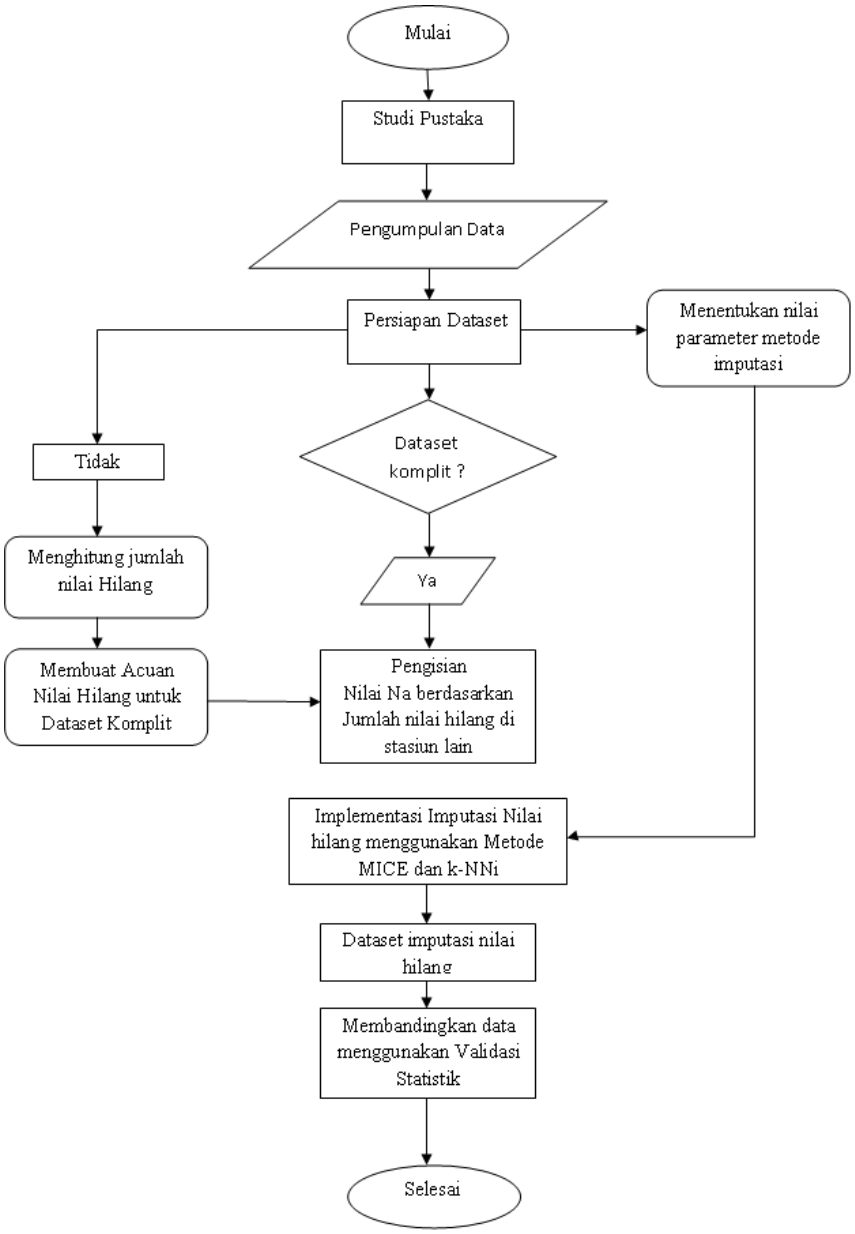
Tabel 3. 2 Dataset Attribut

Attribute	Data Type	Range of Value
Monthyear	Date	(Jan-07 – Des-17)
StationName_waterflow	Real	(0.07-131.42)
Rain	Real	(0,07 - 19.95)
Temp	Real	(0,01 - 4,16)

Pada kolom parameter Rain merupakan Curah hujan dari stasiun penangkar hujan otomatis atau *Automatic Rain Recorder* (ARR). Dataset DAS Opak menggunakan konsep Metode Rasional USSCS 1973 yang terbatas untuk DAS dengan ukuran kecil yaitu kurang dari 300 ha [50]. Hal ini dikarenakan luas DAS Opak adalah 2,9 km2 berkisar 290 ha atau lebih kecil dari 300 ha.

Metode rasional berhubungan dengan Koefisien aliran permukaan, intensitas hujan, dan waktu konsentrasi . Sehingga parameter dataset DAS Opak yaitu Debit Puncak Bulanan, curah hujan, dan temperatur udara. Untuk Curah hujan dan temperature sebagai variable yang mempengaruhi Debit bulanan.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/



Gambar 3. 4 Diagram Alir Penelitian

Pada Gambar 3.4 merupakan diagram alir sebagai tahapan yang dilakukan dalam penelitian ini dan terbagi menjadi beberapa tahapan sebagai berikut:

1. Studi pustaka melakukan kajian pustaka terhadap beberapa pustaka acuan untuk dijadikan landasan penelitian sehingga memperoleh permasalahan (research gap) yang akan dicarikan solusi dengan metode yang diterapkan dalam penelitian ini.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

2. Pengumpulan Dataset diperlukan karena bentuk data masih berupa data kasar dan perlu dinormalisasi agar dapat diproses pada sistem Machine Learning. Proses normalisasi dilakukan pada kolom Bulan dan Tahun menjadi kolom terpisah yang menyatakan Bulan dan tahun dalam kolom tersendiri. Tahap awal proses penelitian ini adalah persiapan dataset untuk dinormalisasi agar dapat dilakukan proses Machine Learning pada pemrograman R. Data yang diproses menggunakan data *numeric* sehingga poin data berupa bulan dan tahun di rubah menjadi angka dengan menambahkan kolom Tahun dan Bulan ke dalam dataset kemudian menghapus kolom *Monthyear.* Pada Tabel 3. 3 hasil dataset DAS Opak yang telah melalui proses normalisasi.

Tabel 3. 3 Normalisasi Dataset

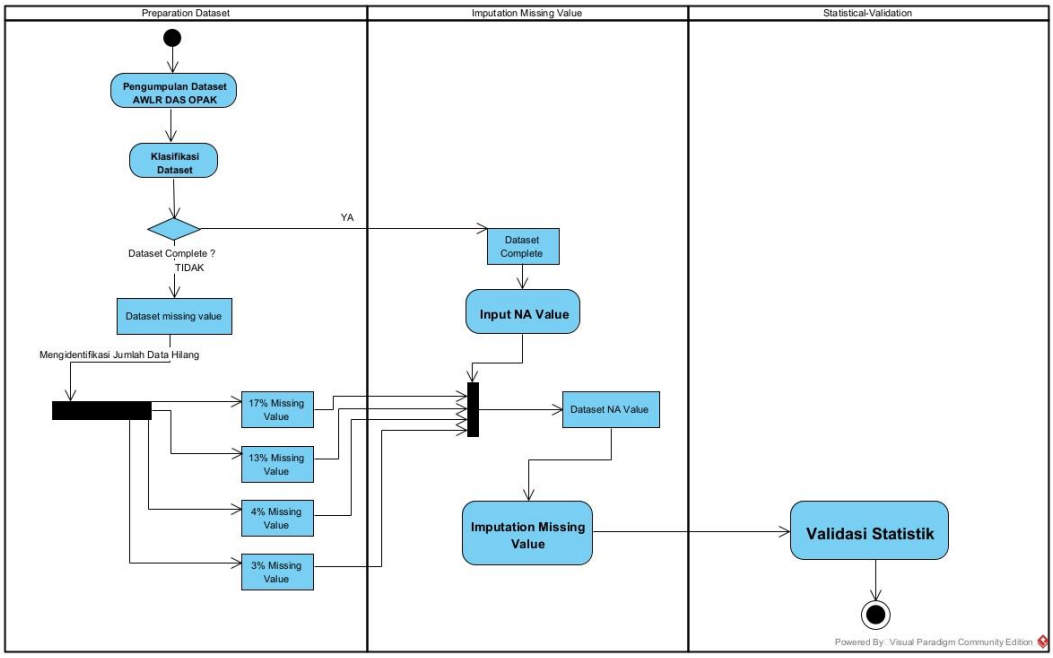
Attribute	*Data Type*
X1	*Numeric*
Bunder Station	*Numeric*
Rain	*Numeric*
Temp	*Numeric*
Year	*Numeric*
Month	*Numeric*

3. Pada data stasiun AWLR yang memiliki 0% rasio *Missing Value* yaitu stasiun Bunder akan menjadi dataset yang dilakukan *Missing value imputation* sebagai bahan validasi metode yang dilakukan. Sedangkan 4 stasiun AWLR lain dihitung jumlah nilai hilang dan digunakan acuan dengan melakukan imputasi nilai hilang ke stasiun Bunder.

4. Tahap selanjutnya adalah memberikan data Na dari dataset dengan 0% *missing value* yaitu Bunder_waterflow. Pemberian Na Value dibagi menjadi 4 dataset dengan masing – masing presentase berbeda dengan 4 dataset stasiun AWLR lain sesuai missing data pada Table 3.1. Pada tahap ini nilai Na akan diberikan ke dalam dataset Stasiun Bunder dengan masing masing nilai hilang yaitu sebesar 3% Na, 4%Na, 13%Na dan 19%Na.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

5. Proses imputasi nilai hilang dilakukan dengan 2 metode yaitu MICE dan k-NNi. Metode MICE hanya menggunakan 1 parameter yaitu m=5 dan metode k-NNi menggunakan 5 parameter yaitu k=3, k=5, k=7, k=8, dan k=20. Dengan bantuan *software R Studio* proses imputasi akan secara langsung melakukan pengisian nilai hilang dengan berdasarkan satuan nilai parameter yang telah ditentukan.
6. Setelah proses *missing value imputation* dilakukan kemudian dilanjutkan proses validasi menggunakan *statistical-validation* untuk mengetahui nilai r^2, MAE, dan RMSE. Hasil perhitungan validasi akan membandingkan metode imputasi menggunakan MICE, k-NNi k=3, k-NNi k=5, k-NNi k=7, k-NNi k=8, dan kNNi k=20. Berikut Gambar 3.4 *activity diagram experiment scenario* proses penelitian yang dilakukan pada penelitian ini.



Gambar 3. 5 Experiment Scenario

3.1.3 Menentukan Parameter Metode Imputasi Nilai Hilang

Proses persiapan dataset dilanjutkan untuk mencari Nilai k pada merode k-NNi. Terdapat beberapa metode khusus, dalam menentukan nilai k pada k-NNi

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

[32]. Akan tetapi, pada metode MICE tidak terdapat metode atau prosedur khusus untuk penentuan nilai m [35]. Menurut Stef Van Burren [35], secara kesimpulan substantif tidak terdapat perubahan dari peningkatan m melebihi atau kurang dari m=5.

Pada nilai parameter k-NNi perlu mencari tahu dengan berbagai nilai parameter dengan percobaan dan dengan asumsi bahwa data pelatihan tidak diketahui. Dengan demikian, tingginya resistensi terhadap mekanisme dan model data yang hilang menjadikan metode k-NNi sebagai salah satu pendekatan penanganan data yang hilang [32].

Hambatan ini juga dilihat dengan imputasi menggunakan parameter k yang berbeda. Jika nilai k terlalu kecil maka akan terjadi banyak *noise* dan bias yang mengurangi tingkat akurasi dalam imputasi, dan apabila terlalu besar akan dapat menyebabkan kesalahan dalam membatasi nilai yang diambil dan secara tidak langsung mempengaruhi keakuratan [23]. Selain itu, Syarat nilai k adalah tidak boleh lebih besar dari jumlah data, parameter k sebaiknya ganjil dan harus lebih dari satu [51][52]. Terkait nilai parameter k dengan nominal ganjil, nilai k pada k-NN harus menggunakan nilai ganjil jika digunakan untuk proses klasifikasi beda halnya jika digunakan untuk melakukan prediksi nila k pada k-NN dapat berupa bilangan ganjil ataupun genap [53].

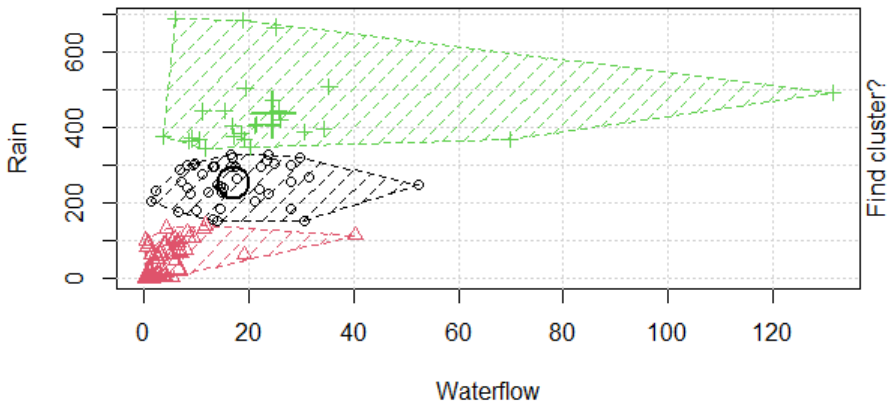
Penentuan parameter nilai k menggunakan rumus *Euclidean Distance* atau jarak *Euclidean*. Jarak *Euclidean* pada penelitian ini menggunakan penentuan jarak 1 dimensi. Hal ini dikarenakan hanya stasiun AWLR Bunder yang memiliki data lengkap atau tidak memiliki nilai hilang. Hasil dari perhitungan jarak *Euclidean* nilai rata – rata dari stasiun bunder menunjukkan angka 7,0754 dan untuk nilai terkecil adalah 0,01. Dengan syarat nilai k yang terlalu kecil akan menimbulkan *noise* dan bias maka nilai 0,01 ditetapkan menjadi nilai k=3. Sedangkan untuk hasil rata – rata nilai jarak *Euclidean* 7,0754 dibulatkan menjadi nilai k=7.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Penggunaan jarak *Euclidean* merupakan konsep umum dalam metode k-NNi. Penambahan algoritma lain diperlukan untuk dapat mencari nilai k yang tepat agar hasil dari imputasi nilai hilang tidak mengalami bias dan anomali. Imputasi nilai hilang menggunakan metode kluster merupakan salah satu metode yang dikenal dalam pengelolaan data mining [54].

Tujuan dari metode cluster adalah memisahkan objek data yang sama atau mirip dengan data yang berbeda tidak terkait antar kelompok. Perhitungan untuk menentukan nilai k dapat diambil dari perhitungan jarak *ecluidean* yang diambil dari nilai centroid *k-Means* untuk dihitung menggunakan rumus jarak *ecluidean*. Proses *clustering* menggunakan *machine learning* dengan pemrograman R dan dijalankan pada perangkat lunak Rstudio. Skema proses *cluster k-Means* diawali dengan penentuan jumlah cluster, kemudian memproses menggunakan algoritma *k-Means.* Selanjutnya hasil *clustering* dari algoritma *k-Means* akan menunjukkan nilai pengelompokan data ke pusat cluster terdekat. Kemudian nilai data cluster pusat dihitung jarak pusat cluster dengan *ecluidean Distance.* Hasil dari rumus jarak *ecluidean* merupakan nilai k yang akan digunakan untuk imputasi missing value pada dataset NA stasiun Bunder.

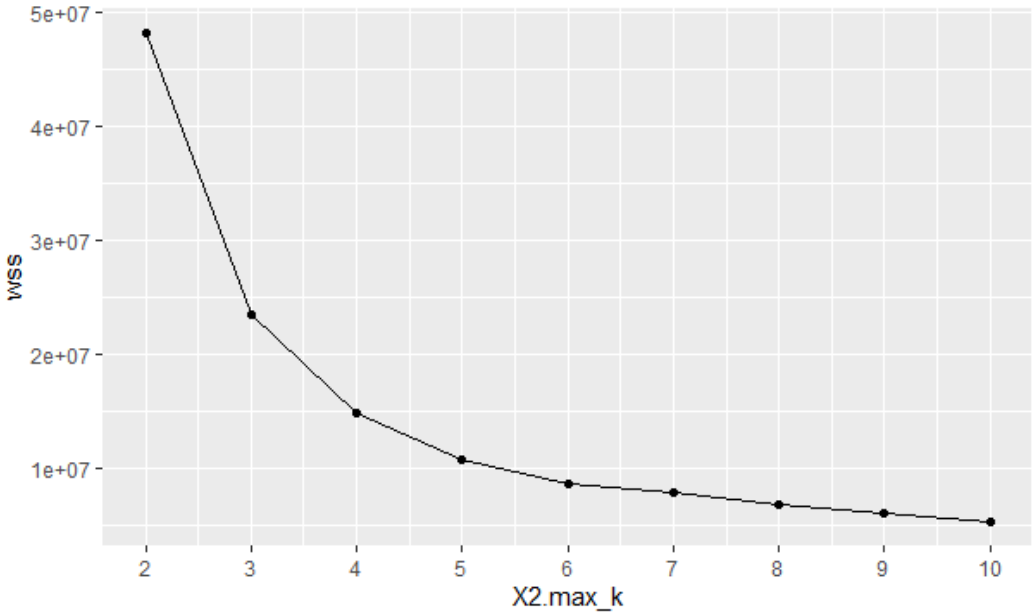
Hasil dari proses *clustering* pada dataset Bunder menunjukkan hasil terbagi menjadi 3 kluster seperti yang ditunjukkan pada Gambar 3.5 tentang hasil training penentuan cluster pada pemrograman R.



Gambar 3. 6 Plot kluster dataset Bunder

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Hasil pembagian cluster tersebut selanjutnya dilakukan optimalisasi untuk menentukan nilai k *clustering* dengan 10 kali iterasi. Jumlah iterasi sebanyak 10 kali dimaksudkan untuk mengurangi bias nilai dikarenakan semakin tinggi nilai k akan semakin mengurangi akurasi. Hasil dari optimalisasi dalam proses iterasi tersebut akan menghasilkan sebuah bagan grafik hasil iterasi sebanyak 10 kali.



Gambar 3. 7 Hasil diagram 10 kali iterasi

Pada gambar 3.6 menunjukkan hasil dari proses iterasi menunjukkan nilai 8 mengalami penuruman hingga iterasi 10. Maka nilai iterasi 8 akan dioptimalisasi untuk mencari nilai titik pusat cluster. Hasil titik pusat cluster ditunjukkan pada Gambar 3.7 dibawah ini.

```
> center_reshape <- gather(center_k_meansExmp, features, values,Waterflow: Rain)
> head(center_reshape)
  cluster    Month  features    values
1       1 42491.62 Waterflow 16.896923
2       2 42020.46 Waterflow  3.400769
3       3 42016.40 Waterflow 16.716000
4       4 41472.00 Waterflow 28.460000
5       5 39843.58 Waterflow  6.597895
6       6 42887.31 Waterflow 10.143077
```

Gambar 3. 8 Nilai titik pusat cluster

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Selanjutnya nilai titik pusat cluster akan dihitung menggunakan jarak *ecluidean* dengan data yang sesuai pada variable bulan titik tengah stasiun bunder. Pada table 3.4 merupakan hasil antara titik tengah cluster dan data stasiun Bunder yang menghasilkan nilai *eclidean distance* sebesar 19,85. Dari hasil berikut nilai k diambil dari pembulatan 19,85 menjadi k=20. Berikut untuk table perhitungan jarak *ecluidean* antara cluster *k-Means* dengan data stasiun Bunder.

Tabel 3. 4 Perhitungan jarak *ecluidean* antara cluster *k-Means* dengan data stasiun Bunder

Cluster	Month	Values Cluster k-Means	Data	ED
1	May-16	16,90	11,75	19,85
2	Jan-15	3,40	17,08	
3	Jan-15	16,72	17,08	
4	Jul-13	28,46	30,47	
5	Jan-09	6,60	17,58	
6	Jun-17	10,14	2,68	

Pada penentuan parameter k selanjutnya menggunakan pendekatan *rule-of-thumb* dimana proses ini menginisialisasi dalam menentukan nilai k [55]. Pendekatan *rule-of-thumb* dapat dijelaskan dengan persamaan sebagai berikut :

$$\boldsymbol{k} = \sqrt{\frac{n}{2}} \qquad (5)[55]$$

k = jumlah nilai k

n = jumlah kolom data

Perhitungan pendekatan *rule-of-thumb* dimulai dengan jumlah kolom data dari stasiun Bunder yang berjumlah 132 kolom data dibagi 2. Hasil dari pendekatan *rule-of-thumb* adalah 8,12 dibulatkan menjadi k=8. Berikut Tabel 3.5 deskripsi penentuan parameter nilai k.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Tabel 3. 5 Deskripsi Penentuan Parameter k

Nilai k	Metode Penentuan Parameter
k=3	Menggunakan nilai terkecil dari jarak Euclidean 1 Dimensi
k=5	Nilai parameter k *default*
k=7	Menggunakan nilai rata - rata dari jarak *Euclidean* 1 Dimensi
k=8	Menggunakan pendekatan *rule-of-thumb*
k=20	Menggunakan *Ecluidean Distance* pada Algoritma *k-Means*

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

BAB IV
HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan mengenai hasil pengembangan algoritme usulan, pengujian algoritme, dan membandingkan kinerja dari k-NNi dan MICE. Kemudian kedua metode tersebut dilakukan perbandingan menggunakan *validation* untuk dapat menentukan dilai RMSE, R^2, dan MAE.

A. Preparation Dataset

Tabel 4. 1 NA Input

Bunder_waterflow (Ori_Dataset)	**Bunder_waterflow (Na_input)**			
	3%	***4%***	***13%***	***25%***
5.83	5.83	5.83	5.83	5.83
18.54	18.54	18.54	18.54	18.54
9.55	9.55	9.55	9.55	9.55
26.07	26.07	26.07	26.07	26.07
3.00	3	3	3	3.00
2.77	2.77	2.77	2.77	2.77
1.64	1.64	1.64	Na	***Na***
0.34	0.34	0.34	Na	***Na***
0.17	0.17	0.17	0.17	0.17
.....	.....	.....	.....	.....
5.93	5.93	5.93	5.93	5.93
9.01	9.01	9.01	Na	Na

Proses persiapan dataset pada tahap awal akan memasukkan nilai Na pada stasiun bunder dengan prosentase 3%, 4%, 13%, 19%. Sehingga akan menjadi 4 dataset Stasiun bunder seperti yang ditunjukkan pada Table 4.1.

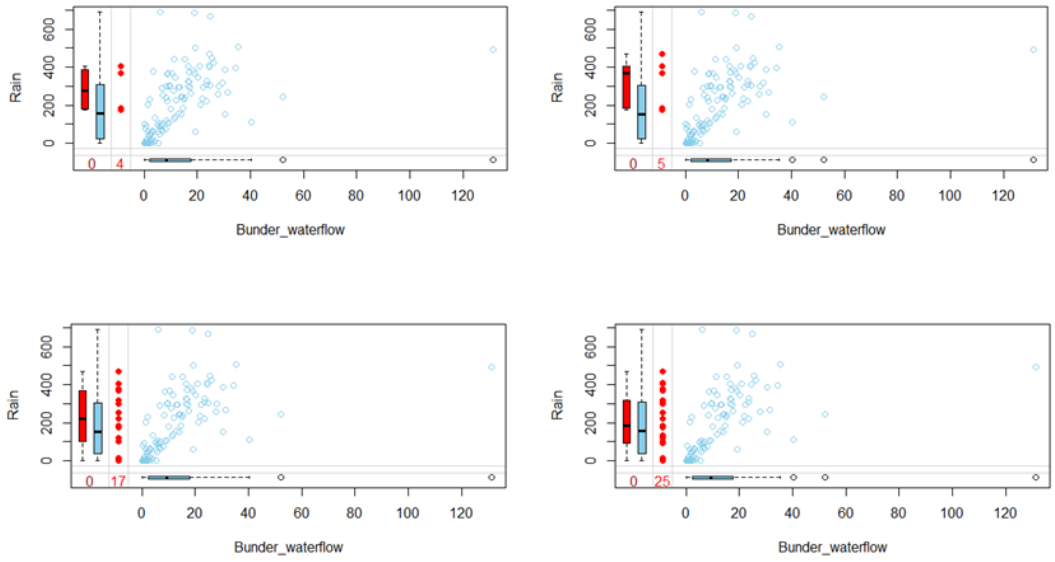
4.1 Implementasi Imputasi Data Hilang

4.1.1 Imputasi Menggunakan Metode MICE

Proses pengisian missing value menggunakan Metode MICE dilakukan dengan menggunakan *Software Rstudio* dengan *package library* MICE. Pada implementasi pengisian missing value menggunakan MICE terdapat 4 langkah yaitu preparation dataset, input missing value, pooling, dan validasi.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Proses selanjutnya adalah melihat *pattern* dari Missing value dengan menggunakan *library(mice)* dengan *package margin plot*. Pada Gambar 4.2 menunjukkan sebaran nilai yang hilang dari 2 variable Bunder_waterflow dan variabel Rain dengan asumsi nilai pada plot kotak merah di sebelah kiri menunjukkan distribusi variabel *Rain* dengan variabel *Bunder_waterflow* yang hilang. Sedangkan pada plot kotak biru, menunjukkan distribusi titik data yang tersisa antara 2 variabel. Asumsi plot ini berdasar dengan model MCAR (*Missing Completely at Random*).



Gambar 4. 1 MICE Margin Plot

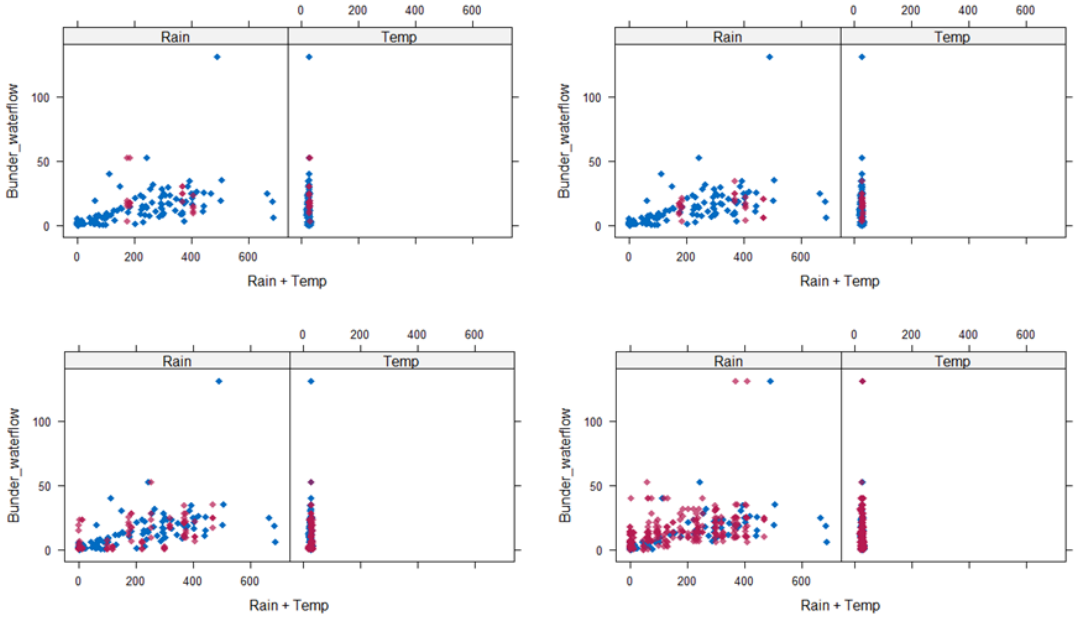
Pengisian *missing value* dalam metode MICE menggunakan nilai m dengan nilai *default* m=5 yang mengacu pada jumlah dataset yang diperhitungkan. Pertimbangan nilai m=5 terdapat pada variabel dan jumlah poin data set yang tidak terlalu banyak yang hanya berisi 132 objek dari 5 variabel. Metode imputasi yang dilakukan mengacu pada *Predictive Mean Matching* (pmm) dari bagian library MICE yang tersedia.

Selanjutnya melakukan pemeriksaan distribusi data asli dengan data imputation dengan menggunakan *xyplot* pada Gambar 4.3. Perbandingan ini dapat memberikan gambaran bahwa titik magenta atau yang diperhitungkan memiliki

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

kecocokan dengan titik biru atau yang teramati bahwa nilai yang diperhitungkan adalah nilai yang masuk akal *(plausible value)*.

```
tempData <mice(data,m=5,maxit=50,
meth='pmm',seed=500)
summary(tempData)
```



Gambar 4. 2 Xyplot sebagai inspeksi distribusi original dan imputed data

Proses pooling merupakan proses terakhir sebelum validasi. Proses ini berfokus pada penyesuaian model linier dengan data yang berisi hasil penyesuaian pada data set yang diperhitungkan dengan fungsi *pool()* untuk mengumpulkan semua penyesuaian data. Pada Gambar 4.4 dan Gambar 4.5 adalah model fungsi *pool()* untuk penyesuaian dataset.

METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

```
modelFit1     <-     with(tempData,lm(Bunder_waterflow~
Rain+Temp))

summary(pool(modelFit1))
```

```
         term    estimate    std.error  statistic       df      p.value
1 (Intercept) -4.39502140 11.559206333 -0.3802183 40.61615 7.057646e-01
2        Rain  0.04906175  0.006715251  7.3060180 86.14412 1.300338e-10
3        Temp  0.26897847  0.414053442  0.6496226 37.37584 5.199088e-01
```

Gambar 4. 3 Pooling modelFit1

```
tempData2 <- mice(data,m=50,seed=245435)

modelFit2<-with(tempData2,lm(Bunder_waterflow~
Rain+Temp))

summary(pool(modelFit2))
```

```
         term    estimate    std.error  statistic       df      p.value
1 (Intercept) -4.48141857 12.388788965 -0.3617318 91.82561 7.183835e-01
2        Rain  0.05045911  0.008242111  6.1221103 77.53909 3.542561e-08
3        Temp  0.27054906  0.438840950  0.6165082 92.06479 5.390806e-01
> |
```

Gambar 4. 4 Pooling modelFit2

4.1.2 Imputasi Menggunakan Metode k-NNi

Pada experimen k-NN Imputation proses pengisian missing value menggunakan dataset yang telah dilakukan pengisian NA Value pada Bunder_waterflow pada proses preparation dataset Penelitian ini menggunakan parameter k = 3,5,7,8,20.

Selanjutnya melakukan import dataset Bunder_waterflow yang telah dibagi menjadi 4 dataset yang telah diisi dengan Na Value pada proses preparation dataset. tersebut kemudian dilakukan pengisian data hilang menggunakan nilai k=3, k=5, k=7, k=8, dan k=20. Berikut ini merupakan kode pemrograman R dengan metode k-NN imputation.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

```
Bunder3<-k-
NN(s_bunder,variable=c("Bunder_waterflow"),k=3)
```

```
Bunder5<-k-
NN(s_bunder,variable=c("Bunder_waterflow"),k=5)
```

```
Bunder7<-k-
NN(s_bunder,variable=c("Bunder_waterflow"),k=7)
```

```
Bunder8<-k-
NN(s_bunder,variable=c("Bunder_waterflow"),k=8)
```

```
Bunder20<-k-
NN(s_bunder,variable=c("Bunder_waterflow"),k=20)
```

4.1.3 Perbandingan Hasil Imputasi

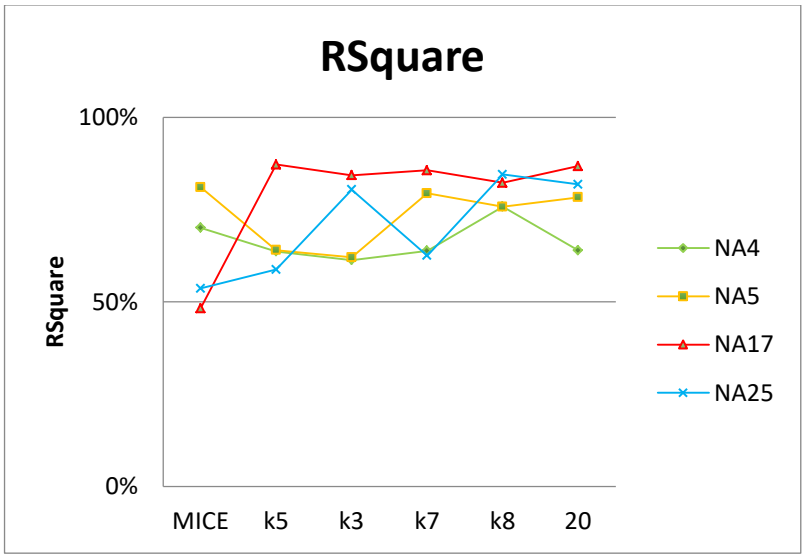
Gambar 4. 5 Perbandingan dataset original dengan imputasi dataset

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Pada proses imputasi nilai hilang menggunakan metode MICE dan k-NNi hamper semua nilai hilang dapat terisi dengan baik. Semua metode dapat bekerja dan mengisi pada setiap jumlah nilai hilang. Pada Gambar 4.5 menunukkan jumlah nilai hilang sebanyak 17 Na dapat diisi menggunakan metode MICE ataupun menggunakan metode k-NNi dengan masing – masing parameter nilai k.

4.1.4 Perbandingan Validasi Statistik k-NNi dan MICE

Dalam proses validasi dari 2 metode disajikan dalam bentuk grafik sebagai bahan perbandingan. Hasil R^2 pada Gambar 4.6 dan Tabel 4.25 menunjukkan bahwa metode k-NNi dengan nilai k=8 memiliki nilai yang lebih konsisten pada rentang nilai yang berdekatan antara na4 hingga na25 yaitu pada rentang 76%-85% dibanding keelima metode lainya. Sedangkan metode imputation MICE hasil persentase variasi total mendapatkan nilai cukup rendah pada rentang nilai hilang 17-25. Sedangkan hasil persentase RSquare pada nilai k pada algoritma *clustering k-Means* dengan *ecluidean distance* mendapatkan persentase tertinggi pada rentang nilai hilang 17 dengan nilai 86,76% atau sedikit lebih rendah dibanding k=5 dengan 87,21% apabilai nilai tersebut tidak dibulatkan.



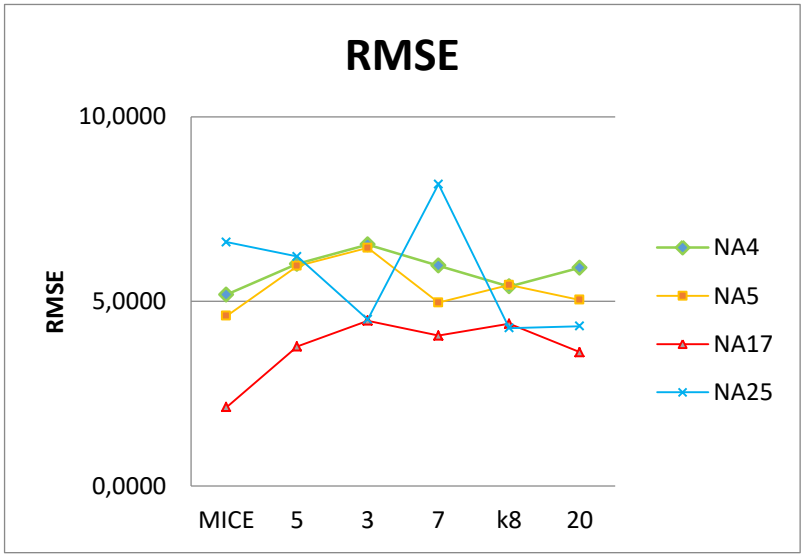
Gambar 4. 6 R Squared

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Tabel 4. 2 R2 (R Squared)

Rsquare						
	MICE	**k=5**	**k=3**	**K=7**	**k=8**	**k=20**
na4	70%	64%	61%	64%	76%	64%
na5	81%	64%	62%	79%	76%	78%
na17	48%	87%	84%	86%	82%	87%
na25	54%	59%	80%	63%	85%	82%
AVG	*63%*	*68%*	*72%*	*73%*	*80%*	*78%*
MIN	*48%*	*59%*	*61%*	*63%*	*76%*	*64%*
MAX	*81%*	*87%*	*84%*	*86%*	*85%*	*87%*

Sedangkan nilai RMSE dengan konsep regresi linier, menunjukkan bahwa variasi nilai yang dihasilkan mendekati variasi nilai observasinya. Nilai RMSE MICE berada pada 6,605 dengan nilai tertinggi kedua dan dianggap masih belum memiliki keakuratan dengan variasi nilai observasi. Pada metode k-NNi k=8 memiliki nilai akurasi RMSE yang paling baik dibanding metode lainya. Hanya k-NNi k=20 dalam k-NNi saja yang memiliki error lebih rendah. Pada Gambar 4.7 dan Tabel 4.3 menggambarkan grafik perbandingan antara metode yang digunakan dalam penelitian ini.



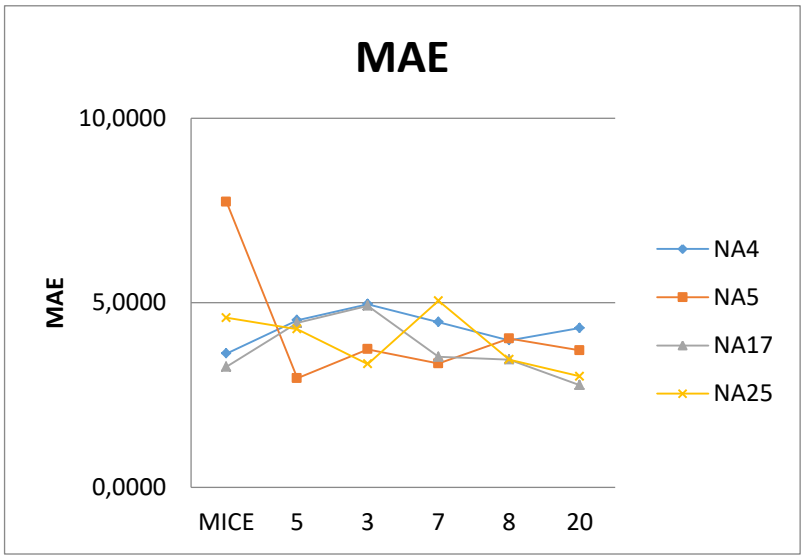
Gambar 4. 7 RMSE

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Tabel 4. 3.RMSE (Root Mean Squared Error)

RMSE						
	MICE	**k=5**	**k=3**	**K=7**	**k=8**	**k=20**
na4	5,1821	6,0157	6,5407	5,9661	5,4011	5,9093
na5	4,6029	5,9575	6,4457	4,9650	5,4417	5,0404
na17	2,1337	3,7764	4,4743	4,0721	4,3934	3,6213
na25	6,6053	6,2122	4,5113	8,1720	4,2736	4,3281
AVG	*4,6310*	*5,4904*	*5,4930*	*5,7938*	*4,8774*	*4,7248*
MIN	*2,1337*	*3,7764*	*4,4743*	*4,0721*	*4,2736*	*3,6213*
MAX	*6,6053*	*6,2122*	*6,5407*	*8,1720*	*5,4417*	*5,9093*

Nilai *Mean Absolute Error* (MAE) dari kelima metode yang dilakukan uji validasi statistik rata – rata nilai error terendah diperoleh dengan metode MICE. Akan tetapi MICE memiliki kekurangan pada nilai hilang pada rasio tinggi yaitu nilai hilang lebih dari 17 nilai hilang. Terlihat pada pada Gambar 4.8 dan Tabel 4.4 nilai MICE mengalami peningkatan signifikan na17 sebesar 2,1337 meningkat drastis hingga 6,60353.



Gambar 4. 8 MAE

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Tabel 4. 4 MAE (Mean Absolute Error)

MAE						
	MICE	**k=5**	**k=3**	**K=7**	**k=8**	**k=20**
na4	3,6233	4,5277	4,9585	4,4766	3,9815	4,3141
na5	7,7253	2,9549	3,7377	3,3535	4,0264	3,7065
na17	3,2636	4,4551	4,9173	3,5324	3,4615	2,7668
na25	4,5924	4,2854	3,3388	5,0502	3,4604	3,0027
AVG	*4,8011*	*4,0558*	*4,2381*	*4,1032*	*3,7325*	*3,4475*
MIN	*3,2636*	*2,9549*	*3,3388*	*3,3535*	*3,4604*	*2,7668*
MAX	*7,7253*	*4,5277*	*4,9585*	*5,0502*	*4,0264*	*4,3141*

4.2 Kelebihan dan Keterbatasan Penelitian

Dari pembahasan di atas terdapat beberapa kelebihan dan keterbatasan penelitian ini dari penelitian sebelumnya serta keterbatasan penelitian yang dapat menjadi masukan untuk penelitian lanjutan yang lebih baik.

4.2.1 Kelebihan Penelitian

1. Penelitian ini menguji kehandalan model dengan variasi persentase nilai hilang yang berbeda pada dataset stasiun Bunder. Pada penelitian sebelumnya tidak dilakukan dengan variasi nilai hilang yang berbeda pada dataset. Jenis dataset pada penelitian sebelumnya [1] [17] [20] [7] jenis dataset hanya pada menggunakan satu dataset dengan satu persentaase nilai hilang.. Variasi nilai hilang berdasarkan dataset acuan pada stasiun lain DAS Opak dapat menguji kehandalan dalam melakukan imputasi nilai hilang.
2. Pada bagian penentuan parameter metode imputasi nilai hilang menjelaskan bagaimana proses penentuan nilai k dilakukan berdasarkan beberapa algoritma lain. Metode tersebut menggunakan *Euclidean distance, rule-of-thumbs* dan *k-means*. Dengan penggabungan beberapa

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

algoritma lain dalam penentuan nilai k diperlukan untuk dapat mencari nilai k yang tepat agar hasil dari imputasi nilai hilang tidak mengalami bias dan anomali.

4.2.2 Keterbatasan Penelitian

1. Penelitian ini menggunakan dua metode yang terdapat penelitian sebelumnya [7] [6] [20] [21]. Metode MICE dan k-NNi yang digunakan dalam penelitian ini hanya terbatas pada hasil dari metode yang dijelaskan pada penelitian sebelumnya. Penggunaan dataset DAS Opak sebagai dataset privat memiliki beberapa kekurangan salah satunya terbatas pada data debit, curah hujan dan suhu. Data debit masih berupa data *monthly seasonal data* dengan durasi bulanan pada setiap dataset DAS Opak pada masing masing stasiun AWLR. Durasi bulanan diketahui akan berpengaruh terhadap tingkat performa pada model metode imputasi. Oleh karena itu, dataset dengan durasi waktu baik itu dalam satuan jam, menit maupun detik sangatlah penting untuk dilakukan pada penelitian selanjutnya.
2. Penelitian ini masih terbatas pada beberapa kendala penggabungan metode yang berdasarkan pada multi disiplin keilmuan lain. Dimulai dari dataset yang masih dapat ditambahkan variable lain sesuai dengan keilmuan Hidrologi Lingkungan akan dapat mempengaruhi hasil performa imputasi yang lebih baik. Sehingga hasil dengan penggabungan metode keilmuan *expert* sangat diperlukan untuk penelitian selanjutnya pada tahap yang lebih tinggi nilai performanya.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari metode imputasi dengan MICE dan k-NNi pengisian *missing value* dapat dilakukan dengan baik pada setiap metode.

1. Setelah dilakukan validasi, nilai rata – rata RMSE dan MAE yang paling konsisten adalah metode k-NNi dengan nilai k=8 menggunakan pendekatan *rule-of-thumb*.
2. Sementara untuk perbandian nilai R^2 metode k-NNi dengan nilai k=8 medapatkan nilai rata –rata presentase terbaik yaitu pada nilai 80% disusul dengan metode k-NNi k=20 sebagai nilai k yang diambil dari pendekatan algoritma *k-Means* dan jarak *euclidean* dengan persentase 78%.
3. Sedangkan metode pembanding MICE mendapatkan rata –rata nilai persentase paling rendah dari metode lain dengan hanya mendapatkan nilai 63%.
4. Maka metode MICE kurang sesuai untuk melakukan proses imputasi dengan menggunakan dataset DAS Sungai Opak yang berada di Provinsi DI Yogyakarta.

5.2 Saran

Penelitian ini masih memiliki beberapa keterbatasan yang perlu ditingkatkan dalam penelitian di masa depan. Berikut beberapa saran yang dapat dilakukan untuk penelitian selanjutnya:

1. Pada metode k-NNi pencarian nilai parameter k menggunakan pertimbangan variable tetangga dengan mekanisme pendekatan *k-Means* dan jarak *Euclidean* menggunakan pemrograman R menjadi tahap yang penting dalam proses imputes data hilang. Pengambilan parameter k yang lebih tepat dapat meminimalisir resiko tingginya resistensi

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

terhadap mekanisme dan model data yang hilang. Sehingga masih dibutuhkan algoritma baru untuk dapat memberikan niai k yang lebih akurat dan tepat.

2. Dari penelitian yang telah dilakukan, model imputasi data yang hilang akan sangat membantu untuk penelitian lanjutan. Hal ini dikarenakan metode k-NNi dapat memberikan akurasi yang cukup baik. Beberapa keilmuan multi displin seperti Mekanika Fluida dalam keilmuan Fisika, Limpasan air sungai dalam keilmuan Hidrologi Geaografi, dan Dimensional Sungai dalam keilmuan Teknik Sipil dapat dikombinasikan dengan metode *Machine Learning* seperti k-NNi. Hasil kombinasi tersebut nantinya dapat dilakukan pada penelitian selanjutnya salah satunya prediksi Banjir dan Klasifikasi Tata Kelola Ruang dengan gabungan algoritme multi disiplin ilmu.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR PUSTAKA

[1] S. Kamwaga, D. M. M. Mulungu, and P. Valimba, "Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha catchment Tanzania," *Phys. Chem. Earth*, vol. 106, no. May 2016, pp. 17–28, 2018, doi: 10.1016/j.pce.2018.05.008.

[2] R. J. Abrahart *et al.*, "Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting," *Prog. Phys. Geogr.*, vol. 36, no. 4, pp. 480–513, 2012, doi: 10.1177/0309133312444943.

[3] E. Acu, "Classification, Clustering, and Data Mining Applications," *Classif. Clust. Data Min. Appl.*, no. January 2004, 2004, doi: 10.1007/978-3-642-17103-1.

[4] J. Luengo, S. García, and F. Herrera, "A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method," *Neural Networks*, vol. 23, no. 3, pp. 406–418, Apr. 2010, doi: 10.1016/J.NEUNET.2009.11.014.

[5] L. Sunitha, M. Balraju, and J. Sasikiran, "Data Mining: Estimation of Missing Values Using Lagrange Interpolation Technique," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 4, pp. 1579–1582, 2013, [Online]. Available: http://ijarcet.org/wp-content/uploads/IJARCET-VOL-2-ISSUE-4-1579-1582.pdf.

[6] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

of Data Imputation Methods for Numeric Dataset," *Appl. Artif. Intell.*, vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.

[7] C. Curley, R. M. Krause, R. Feiock, and C. V. Hawkins, "Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database," *Urban Aff. Rev.*, vol. 55, no. 2, pp. 591–615, 2019, doi: 10.1177/1078087417726394.

[8] R. J. A. Little, "Missing-data adjustments in large surveys," *J. Bus. Econ. Stat.*, vol. 6, no. 3, pp. 287–296, 1988, doi: 10.1080/07350015.1988.10509663.

[9] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.

[10] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002, doi: 10.1037/1082-989X.7.2.147.

[11] D. Ferguson and W. Winkler, "Glossary of Terms on Statistical Data Editing," *Conf. Eur. Stat. Methodol. Mater. Gloss.*, 2000, [Online]. Available: https://ec.europa.eu/eurostat/ramon/statmanuals/files/UN_editing_glossary_2000.pdf.

[12] W. M. Campion and D. B. Rubin, "Multiple Imputation for Nonresponse in Surveys," *J. Mark. Res.*, vol. 26, no. 4, p. 485, 1989, doi: 10.2307/3172772.

[13] Statistics Canada, "Catalogue no. 12-539," *Stat. Canada Qual. Guidel.*, no.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

12, pp. 21–24, 2003.

[14] S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," *J. Biom. Biostat.*, vol. 06, no. 01, pp. 1–6, 2015, doi: 10.4172/2155-6180.1000224.

[15] G. Chhabra, V. Vashisht, and J. Ranjan, "A review on missing data value estimation using imputation algorithm," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 7 Special Issue, pp. 312–318, 2019.

[16] A. Kowarik and M. Templ, "Imputation with the R package VIM," *J. Stat. Softw.*, vol. 74, no. 7, 2016, doi: 10.18637/jss.v074.i07.

[17] Doreswamy, I. Gad, and B. R. Manjunatha, "Performance evaluation of predictive models for missing data imputation in weather data," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 1327–1334, 2017, doi: 10.1109/ICACCI.2017.8126025.

[18] E. C. Blessie, E. Karthikeyan, and V. Thavavel, "An Extended RELIEF-DISC for Handling of Incomplete Data to Improve the Classifier Performance," 2012.

[19] S. Jain and K. Jain, "Estimation of Missing Attribute Value in Time Series Database in Data Mining," *Glob. J. Comput. Sci. Technol.*, 2017.

[20] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," *Proceeding - 2019 5th Int. Conf. Sci. Inf. Technol. Embrac. Ind. 4.0 Towar. Innov. Cyber Phys. Syst. ICSITech 2019*, pp. 83–88, 2019, doi:

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

10.1109/ICSITech46713.2019.8987530.

[21] G. Chhabra, V. Vashisht, and J. Ranjan, "A Comparison of Multiple Imputation Methods for Data with Missing Values," *Indian J. Sci. Technol.*, vol. 10, no. 19, pp. 1–7, 2017, doi: 10.17485/ijst/2017/v10i19/110646.

[22] L. D. Jackel, "Limits in Learning Machine Accuracy Imposed by Data Quality," *KDD-95 Proc.*, vol. 1, no. KDD-95, pp. 57–62, 1994.

[23] J. M. Engels and P. Diehr, "Imputation of missing longitudinal data: A comparison of methods," *J. Clin. Epidemiol.*, vol. 56, no. 10, pp. 968–976, 2003, doi: 10.1016/S0895-4356(03)00170-7.

[24] J. Maillo, S. Ramírez, I. Triguero, and F. Herrera, "kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data," *Knowledge-Based Syst.*, vol. 117, pp. 3–15, 2017, doi: 10.1016/j.knosys.2016.06.012.

[25] G. E. A. P. A. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," *Front. Artif. Intell. Appl.*, vol. 87, no. January, pp. 251–260, 2002.

[26] B. Suthar, H. Patel, and A. Goswami, "A Survey: Classification of Imputation Methods in Data Mining," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 1, pp. 309–312, 2012, [Online]. Available: www.ijetae.com.

[27] D. Priya, R. Sivaraj, R. Assistant, and S. Gr, "a Review of Missing Data Handling Methods," *Int. J. Eng. Technol. Sci. – IJETS^{TM} ISSN*, vol. 2, no. 2, pp. 2349–3968, 2015.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

[28] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *Int. Stat. Rev. / Rev. Int. Stat.*, vol. 57, no. 3, pp. 238–247, Oct. 1989, doi: 10.2307/1403797.

[29] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[30] G. L. Pritalia, "Klasifikasi Gaya Belajar Global Dan Sekuensial Pada Multimedia Learning Menggunakan Pendekatan Eye-Tracking Dan Machine Learning Pada Multimedia Learning Menggunakan Pendekatan Eye-Tracking Dan Machine Learning," Yogyakarta, 2020.

[31] Q. Wang and J. N. K. Rao, "Empirical likelihood-based inference under imputation for missing response data," *Ann. Stat.*, vol. 30, no. 3, pp. 896–924, 2002, doi: 10.1214/aos/1028674845.

[32] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009, doi: 10.1142/S0218001409007326.

[33] A. MELISSA J, "Multiple imputation by chained equations: what is it and how does it work?," *Int. J. Methods Psychiatr. Res.*, vol. 17 Suppl 1, no. 1, pp. S78–S82, 2008, doi: 10.1002/mpr.

[34] P. Hayati Rezvan, K. J. Lee, and J. A. Simpson, "The rise of multiple imputation: A review of the reporting and implementation of the method in medical research Data collection, quality, and reporting," *BMC Med. Res. Methodol.*, vol. 15, no. 1, pp. 1–14, 2015, doi: 10.1186/s12874-015-0022-1.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

[35] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.

[36] U. Azmi, Z. N. Hadi, and S. Soraya, "ARDL METHOD: Forecasting Data Curah Hujan Harian NTB," *J. Varian*, vol. 3, no. 2, pp. 73–82, 2020, doi: 10.30812/varian.v3i2.627.

[37] D. F. Hamilton, M. Ghert, and A. H. R. W. Simpson, "Interpreting regression models in clinical outcome studies," *Bone Jt. Res.*, vol. 4, no. 9, pp. 152–153, 2015, doi: 10.1302/2046-3758.49.2000571.

[38] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[39] A. A. Suryanto, "Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi," *Saintekbu*, vol. 11, no. 1, pp. 78–83, 2019, doi: 10.32764/saintekbu.v11i1.298.

[40] W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 324, no. 1, 2018, doi: 10.1088/1757-899X/324/1/012049.

[41] M. J. Hartmann and G. Carleo, "Neural-Network Approach to Dissipative Quantum Many-Body Dynamics," *Phys. Rev. Lett.*, vol. 122, no. 25, p. 250502, 2019, doi: 10.1103/PhysRevLett.122.250502.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

[42] K. Crammer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res. - JMLR*, vol. 2, no. 2, pp. 265–292, 2002.

[43] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[44] J. Shao, "Linear model selection by cross-validation," *J. Am. Stat. Assoc.*, vol. 88, no. 422, pp. 486–494, 1993, doi: 10.1080/01621459.1993.10476299.

[45] Y. Sarvina, "Pemanfaatan Software Open Source 'R' Untuk Penelitian Agroklimat," *Inform. Pertan.*, vol. 26, no. 1, p. 23, 2017, doi: 10.21082/ip.v26n1.2017.p23-30.

[46] R-project, "What is R?," *CRAN*. https://www.r-project.org/about.html.

[47] A. F. Zuur, E. N. Ieno, and E. Meesters, "Introduction," in *A Beginner's Guide to R*, New York, NY: Springer New York, 2009, pp. 1–27.

[48] P. L. S. D. Air, "Katalog Basis Data 2014 Sumber Daya Air," *Badan Penelit. dan Pengemb. Kementrian Pekerj. Umum*, vol. 1, no. 1, pp. 341–353, 2014.

[49] S. L. Dingman, *Physical Hydrology: Third Edition*, Third Edit. Illinois: Waveland Press, 2015.

[50] S. J. Goldman, T. A. Bursztynsky, and K. Jackson, *Erosion and sediment control handbook*, 1st ed. New York: McGraw-Hill, 1986.

[51] M. Rivki and A. M. Bachtiar, "IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN FOLLOWER TWITTER YANG MENGGUNAKAN BAHASA INDONESIA," *J. Sist.*

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

Inf., vol. 13, no. 1, pp. 31–37, May 2017, doi: 10.21609/jsi.v13i1.500.

[52] J. Han, M. Kamber, and J. Pei, "3 - Data Preprocessing," in *Data Mining (Third Edition)*, Third Edit., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 83–124.

[53] A. Khairi, "IMPLEMENTASI K-NEAREST NEIGHBOR (KNN) UNTUK KLASIFIKASI MASYARAKAT PRA SEJAHTERA DESA SAPIKEREP KECAMATAN SUKAPURA," *TRILOGI J. Ilmu Teknol. Kesehatan, dan Hum.*, vol. 2, no. 3, pp. 319–323, 2021.

[54] T. Hartono, "IMPUTATION USING FUZZY K-MEANS FOR MISSING DATA CASE STUDY LARGE MANUFACTURING ESTABLISHMENT DATA OF EAST JAVA PROVINCE 2008," Institut Teknologi Sepuluh Nopember Surabaya, 2011.

[55] M. Nishom and M. Y. Fathoni, "Implementasi Pendekatan Rule-Of-Thumb untuk Optimasi Algoritma K-Means Clustering," *J. Inform. J. Pengemb. IT*, vol. 3, no. 2, pp. 237–241, 2018, doi: 10.30591/jpit.v3i2.909.

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

LAMPIRAN

L.1 Data lengkap AWLR Stasiun Bunder

Bunder_waterflow	Rain	Temp	Year	Month
5,83	104,7	27,18	2007	1
18,54	384,0	27,26	2007	2
9,55	105,9	27,97	2007	3
26,07	421,4	28,08	2007	4
3,00	59,6	34,5	2007	5
2,77	52,2	34,15	2007	6
1,64	4,5	34,98	2007	7
0,34	0,4	35,08	2007	8
0,17	1,8	33,82	2007	9
0,54	80,9	35,16	2007	10
2,23	231,2	28,3	2007	11
18,88	685,9	29,16	2007	12
7,20	254	29,24	2008	1
27,95	257,3	29,13	2008	2
24,50	470,7	29,23	2008	3
8,65	362,6	29,21	2008	4
3,81	15,5	29,3	2008	5
2,70	18	29,31	2008	6
1,70	0	26,56	2008	7
1,40	0	26,35	2008	8
0,87	3	25,85	2008	9
6,41	71,7	26,96	2008	10
25,00	667,2	30,05	2008	11
8,25	301,2	20,03	2008	12
17,58	263,1	20,66	2009	1
16,51	326,3	26,69	2009	2
11,32	130,7	30,89	2009	3
14,95	228,6	30,12	2009	4
4,17	132,4	30,55	2009	5
1,77	49,1	31,1	2009	6
0,98	21,3	30,85	2009	7

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

0,57	0,8	29,73	2009	8
0,29	3,2	30,55	2009	9
0,30	100	30,29	2009	10
6,82	101,6	30,32	2009	11
12,19	226,7	20,02	2009	12
11,23	442,7	21,09	2010	1
9,71	304,6	22,67	2010	2
13,42	296,2	21,87	2010	3
10,04	180,2	21,67	2010	4
6,81	289,0	24,76	2010	5
8,27	120,6	23,5	2010	6
7,10	67,6	23,67	2010	7
5,91	88,3	28,42	2010	8
10,53	369,1	28,45	2010	9
13,04	297,3	28,78	2010	10
11,69	346,1	29,78	2010	11
15,27	443,6	30,12	2010	12
34,28	395,7	24,66	2011	1
16,71	404,5	24,67	2011	2
21,86	233,9	27,72	2011	3
11,14	275	26,87	2011	4
14,65	184,2	28,93	2011	5
3,79	4,5	30,31	2011	6
1,76	0	26,37	2011	7
0,70	0	25,9	2011	8
0,38	0	26,7	2011	9
1,15	25,7	28,65	2011	10
15,17	241,1	27,75	2011	11
23,43	310,3	21,66	2011	12
22,31	294,9	26,52	2012	1
30,62	388,2	26,67	2012	2
16,94	320,5	27,83	2012	3
13,86	246,5	23,66	2012	4
5,51	63,1	28,04	2012	5
3,90	4,2	27,13	2012	6
2,42	0,3	25,99	2012	7
1,26	0,1	26,1	2012	8
1,00	0	27,62	2012	9
3,28	66,8	28,62	2012	10
8,77	222,3	29,02	2012	11
21,46	406,7	28,25	2012	12

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

131,42	492,9	28,21	2013	1
69,72	369	28,54	2013	2
52,30	246,1	28,96	2013	3
40,27	112,8	29,22	2013	4
23,71	222,4	28,93	2013	5
30,47	151,6	27,94	2013	6
19,19	62,2	27,37	2013	7
5,35	1,9	27,02	2013	8
1,46	5	27,88	2013	9
0,83	91,6	28,68	2013	10
9,36	300,7	27,86	2013	11
25,70	445,8	27,6	2013	12
24,79	304,7	27,54	2014	1
28,12	298,1	28,27	2014	2
12,97	157,4	29,32	2014	3
6,41	176,6	32,29	2014	4
5,23	96,8	29,57	2014	5
3,18	65,9	28,95	2014	6
2,47	51,4	30,61	2014	7
0,28	0	27,69	2014	8
0,14	0	27,77	2014	9
0,07	3,2	29,4	2014	10
3,62	376,4	29,19	2014	11
19,26	503,5	27,95	2014	12
17,08	375	25,21	2015	1
21,24	203	24,38	2015	2
24,55	402,6	24,16	2015	3
19,10	368,4	24,88	2015	4
8,03	76,2	23,68	2015	5
2,53	14,2	22,6	2015	6
0,42	0	21,52	2015	7
0,26	0	23,01	2015	8
0,22	0	23,45	2015	9
0,16	0	25,56	2015	10
1,43	205	26,68	2015	11
10,52	369,6	26,65	2015	12
14,03	152,8	29,4	2016	1
29,65	320,2	28,65	2016	2
24,59	409,9	29,02	2016	3
27,94	184,7	29,53	2016	4
11,75	140,4	29,48	2016	5

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

17,42	296,5	28,55	2016	6
8,58	105,7	28,77	2016	7
6,57	94,5	28,35	2016	8
8,27	240	28,53	2016	9
23,62	327,2	27,33	2016	10
35,29	508,2	28,28	2016	11
31,43	267,1	27,52	2016	12
16,52	297,6	23,42	2017	1
20,40	347,7	25,71	2017	2
21,09	403,4	25,42	2017	3
14,41	243,4	27,98	2017	4
5,87	45,7	22,71	2017	5
2,68	9,2	23,42	2017	6
1,19	12,7	25,9	2017	7
0,47	0	24,71	2017	8
1,46	62,8	24,67	2017	9
1,40	60,3	25,89	2017	10
5,93	689	25,12	2017	11
9,01	370,3	29,79	2017	12

L.2 Pengisian Data NA Stasiun Bunder

s_Bunder_Original	s_Bunder_NA4	s_Bunder_NA5	s_Bunder_NA17	s_Bunder_NA25
5,83	5,83	5,83	5,83	5,83
18,54	18,54	18,54	18,54	18,54
9,55	9,55	9,55	9,55	9,55
26,07	26,07	26,07	26,07	26,07
3	3	3	3	3
2,77	2,77	2,77	2,77	2,77
1,64	1,64	1,64	NA	NA
0,34	0,34	0,34	NA	NA
0,17	0,17	0,17	0,17	0,17
0,53	0,53	0,53	0,53	0,53
2,23	2,23	2,23	2,23	2,23
18,88	18,88	18,88	18,88	18,88
7,2	7,2	7,2	NA	NA
27,95	27,95	27,95	27,95	27,95
24,5	24,5	NA	NA	NA
8,65	8,65	8,65	8,65	8,65
3,8	3,8	3,8	3,8	3,8

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

2,7	2,7	2,7	2,7	2,7
1,7	1,7	1,7	1,7	1,7
1,4	1,4	1,4	1,4	1,4
0,86	0,86	0,86	0,86	0,86
6,41	6,41	6,41	6,41	6,41
25	25	25	25	25
8,25	8,25	8,25	NA	NA
17,58	17,58	17,58	17,58	17,58
16,51	16,51	16,51	16,51	16,51
11,32	11,32	11,32	11,32	NA
14,95	14,95	14,95	14,95	14,95
4,17	4,17	4,17	4,17	NA
1,77	1,77	1,77	1,77	1,77
0,97	0,97	0,97	0,97	0,97
0,56	0,56	0,56	0,56	0,56
0,29	0,29	0,29	0,29	0,29
0,3	0,3	0,3	NA	NA
6,82	6,82	6,82	6,82	6,82
12,18	12,18	12,18	12,18	12,18
11,23	11,23	11,23	11,23	11,23
9,71	9,71	9,71	9,71	9,71
13,41	13,41	13,41	13,41	13,41
10,04	10,04	10,04	10,04	10,04
6,8	6,8	6,8	6,8	6,8
8,27	8,27	8,27	NA	NA
7,09	7,09	7,09	7,09	7,09
5,91	5,91	5,91	5,91	5,91
10,52	10,52	10,52	10,52	10,52
13,04	13,04	13,04	13,04	13,04
11,69	11,69	11,69	11,69	11,69
15,27	15,27	15,27	15,27	15,27
34,28	34,28	34,28	34,28	34,28
16,71	16,71	16,71	16,71	16,71
21,86	21,86	21,86	21,86	21,86
11,13	11,13	11,13	11,13	11,13
14,65	14,65	14,65	14,65	14,65
3,79	3,79	3,79	3,79	3,79
1,76	1,76	1,76	1,76	1,76
0,7	0,7	0,7	0,7	0,7
0,38	0,38	0,38	0,38	0,38
1,15	1,15	1,15	1,15	1,15

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

15,17	15,17	15,17	15,17	15,17
23,43	23,43	23,43	23,43	23,43
22,3	22,3	22,3	22,3	22,3
30,62	30,62	30,62	30,62	30,62
16,94	16,94	16,94	16,94	16,94
13,86	13,86	13,86	13,86	13,86
5,51	5,51	5,51	5,51	5,51
3,9	3,9	3,9	3,9	3,9
2,42	2,42	2,42	2,42	2,42
1,25	1,25	1,25	1,25	1,25
0,99	0,99	0,99	0,99	0,99
3,28	3,28	3,28	3,28	3,28
8,77	8,77	8,77	NA	NA
21,46	NA	NA	NA	NA
131,42	131,42	131,42	131,42	131,42
69,72	NA	NA	NA	NA
52,3	52,3	52,3	52,3	52,3
40,27	40,27	40,27	40,27	40,27
23,71	23,71	23,71	23,71	23,71
30,47	30,47	30,47	30,47	30,47
19,19	19,19	19,19	19,19	19,19
5,34	5,34	5,34	5,34	5,34
1,46	1,46	1,46	1,46	1,46
0,83	0,83	0,83	0,83	NA
9,36	9,36	9,36	9,36	9,36
25,7	25,7	25,7	25,7	25,7
24,79	24,79	24,79	24,79	NA
28,12	28,12	28,12	28,12	NA
12,97	12,97	12,97	12,97	12,97
6,4	NA	NA	NA	NA
5,22	5,22	5,22	5,22	5,22
3,18	3,18	3,18	3,18	3,18
2,47	2,47	2,47	2,47	2,47
0,28	0,28	0,28	0,28	NA
0,14	0,14	0,14	0,14	0,14
0,07	0,07	0,07	0,07	0,07
3,61	3,61	3,61	NA	NA
19,26	19,26	19,26	19,26	19,26
17,07	17,07	17,07	17,07	17,07
21,24	21,24	21,24	21,24	21,24
24,55	24,55	24,55	24,55	24,55

UNIVERSITAS
GADJAH MADA
METODE IMPUTASI DATA HILANG PADA DAERAH ALIRAN SUNGAI OPAK PROVINSI DI YOGYAKARTA
FAHMI DHIMAS IRNAWAN, Indriana Hidayah, Dr., S.T., M.T., Ir. Lukito Edi Nugroho, M.Sc., Ph.D
Universitas Gadjah Mada, 2022 | Diunduh dari http://etd.repository.ugm.ac.id/

19,09	19,09	19,09	19,09	19,09
8,03	8,03	8,03	8,03	8,03
2,53	2,53	2,53	NA	NA
0,42	0,42	0,42	0,42	NA
0,26	0,26	0,26	0,26	0,26
0,21	0,21	0,21	0,21	0,21
0,16	0,16	0,16	0,16	0,16
1,43	1,43	1,43	1,43	1,43
10,52	10,52	10,52	10,52	10,52
14,03	14,03	14,03	14,03	14,03
29,65	29,65	29,65	NA	NA
24,59	24,59	24,59	24,59	NA
27,94	NA	NA	NA	NA
11,75	11,75	11,75	11,75	11,75
17,42	17,42	17,42	17,42	17,42
8,58	8,58	8,58	8,58	8,58
6,57	6,57	6,57	6,57	6,57
8,27	8,27	8,27	8,27	8,27
23,62	23,62	23,62	23,62	23,62
35,29	35,29	35,29	35,29	35,29
31,43	31,43	31,43	31,43	31,43
16,52	16,52	16,52	16,52	16,52
20,4	20,4	20,4	20,4	20,4
21,09	21,09	21,09	21,09	21,09
14,41	14,41	14,41	14,41	14,41
5,87	5,87	5,87	5,87	5,87
2,68	2,68	2,68	2,68	2,68
1,19	1,19	1,19	1,19	1,19
0,47	0,47	0,47	NA	NA
1,46	1,46	1,46	1,46	1,46
1,4	1,4	1,4	1,4	1,4
5,93	5,93	5,93	5,93	5,93
9,01	9,01	9,01	NA	NA