PERBANDINGAN METODE *HOT-DECK IMPUTATION* DAN METODE KNNI DALAM MENGATASI *MISSING VALUES*

Penerapan Pada Data Susenas Maret Tahun 2017

(*Comparison of Hot-deck Imputation Method and KNNI Method in Overcoming Missing Values*)

Iman Jihad Fadillah[1], Siti Muchlisoh[2]

Politeknik Statistika STIS[1]
Politeknik Statistika STIS[2]
Jakarta Timur, DKI Jakarta, Indonesia
E-mail: 15.8662@stis.ac.id

ABSTRAK

Salah satu ciri data statistik yang berkualitas adalah *completeness*. Namun, pada penyelenggaraan sensus atau survei, sering kali ditemukan masalah data hilang atau tidak lengkap (*missing values*), tidak terkecuali pada data Survei Sosial Ekonomi Indonesia (Susenas). Berbagai masalah dapat ditimbulkan oleh *missing values*. Oleh karena itu, masalah missing values harus ditangani. Imputasi adalah cara yang sering digunakan untuk menangani masalah ini. Terdapat beberapa metode imputasi yang telah dikembangkan untuk menangani *missing values*. *Hot-deck Imputation* dan *K-Nearest Neighbor Imputation* (KNNI) merupakan metode yang dapat digunakan untuk menangani masalah *missing values*. Metode *Hot-deck Imputation* dan KNNI memanfaatkan variabel prediktor untuk melakukan proses imputasi dan tidak memerlukan asumsi yang rumit dalam penggunaannya. Algoritma dan cara penanganan *missing values* yang berbeda pada kedua metode tentunya dapat menghasilkan hasil estimasi yang berbeda pula. Penelitian ini membandingkan metode *Hot-deck Imputation* dan KNNI dalam mengatasi *missing values*. Analisis perbandingan dilakukan dengan melihat ketepatan estimator melalui nilai RMSE dan MAPE. Selain itu, diukur juga performa komputasi melalui penghitungan *running time* pada proses imputasi. Implementasi kedua metode pada data Susenas Maret Tahun 2017 menunjukkan bahwa, metode KNNI menghasilkan ketepatan estimator yang lebih baik dibandingkan *Hot-deck Imputation*. Namun, performa komputasi yang dihasilkan pada *Hot-deck Imputation* lebih baik dibandingkan KNNI.

Kata kunci: *missing values*, imputasi, *Hot-deck Imputation*, KNNI

ABSTRACT

One characteristic of quality statistical data is completeness. However, in carrying out censuses or surveys, problems are often found in missing or incomplete data (missing values), including the Indonesian Socio-Economic Survey (Susenas). Various problems can be caused by missing values. Therefore, missing values issues must be addressed. Imputation is a method commonly used to overcome this problem. Some values are missing. Hot-deck Imputation and KNNI are methods that can be used to overcome missing values problems. The Hot-deck Imputation and KNNI method uses predictor variables to carry out the imputation process and does not require complicated approvals in its use. Algorithms and how to manage missing values in different methods can produce different results. This study compared the Hot-deck Imputation and KNNI methods in overcoming missing values. Assessment analysis is done by looking at the accuracy of the estimator through the values of RMSE and MAPE. In addition, computational performance is also calculated by calculating the running time in the imputation process. The implementation of the two methods in the Susenas data in March 2017 shows that the KNNI method results in better estimator accuracy than Hot-deck Imputation. However, the computational performance generated in Hot-deck Imputation is better than KNNI.

***Keywords**: missing values, imputation, Hot-deck Imputation, KNNI*

PENDAHULUAN

Dalam usaha pengembangan *Official Statistics dalam m*endukung implementasi *Sustainable Development Goals* (SDGs), diperlukan data yang berkualitas. Dengan adanya data yang berkualitas, pencapaian tujuan-tujuan pembangunan berkelanjutan dapat didukung dengan baik. Han (2012) menjelaskan bahwa, salah satu ciri data statistik yang berkualitas adalah *completeness*. Namun, umumnya data yang dikumpulkan pada penyelenggaraan survei atau sensus masih dalam keadaan mentah, dimana sering dijumpai masalah data yang *incompleteness* (terdapat *missing values/missing data*). *Missing values* didefinisikan sebagai nilai data yang tidak disimpan untuk suatu variabel dalam suatu pengamatan atau observasi (Kang, 2013).

Banyak hal yang dapat menyebabkan terjadinya *missing values*. Mulai dari responden yang tidak mau diwawancara atau tidak dapat ditemui, data tidak terekam karena kesalahan petugas, serta kegagalan peralatan dan aplikasi (Batista, 2002). Selain itu, menurut Pearson (2005), *missing values* juga dapat muncul dalam bentuk *outlier* atau nilai yang tidak konsisten dengan nilai sebelumnya, ataupun isian yang tidak wajar pada data (Luengo, J, 2009). Luengo, J. (2011) menjelaskan bahwa beberapa masalah terkait dengan adanya *missing values*, mulai dari hilangnya efisiensi, komplikasi dalam menangani dan menganalisis data, ataupun masalah adanya bias yang dihasilkan antara data yang mengandung *missing values* dengan data lengkap (Kaiser, 2014). Oleh karena itu, perlu dilakukan penanganan lebih lanjut untuk mengatasi permasalahan *missing values*.

Missing values dapat terjadi baik pada unit observasi maupun pada beberapa item pertanyaan saja (Handayani, 2011). Permasalahan *missing values* pada unit observasi dapat ditangani dengan menghapus *cases* yang mengandung *missing value* kemudian melakukan modifikasi bobot dalam upaya penyesuaian untuk *nonresponse*. Menurut Little dan Rubin (2002), prosedur seperti ini dikenal dengan istilah *weighting procedures*. Namun, untuk kasus *missing values* yang terjadi pada item pertanyaan, penanganan *weighting procedures* menjadi kurang efisien, hal ini dikarenakan *missing values* hanya terjadi pada beberapa item pertanyaan saja. Menghapus item secara keseluruhan pada unit observasi, tentunya akan membuat hilangnya informasi yang telah dikumpulkan dan membuat pendugaan parameter menjadi tidak efisien. Menurut Little dan Rubin (2002), ketika *missing values* terjadi pada item pertanyaan, metode imputasi adalah prosedur yang dapat digunakan untuk menangani permasalah ini.

Menurut Biemer (2003), *missing values* ditemukan hampir disemua upaya pengumpulan data berskala besar dan menjadi masalah dalam pengerjaan survei. Survei Sosial Ekonomi Nasional (Susenas) adalah salah satu survei berskala besar yang rutin dilakukan oleh Badan Pusat Statistik setiap tahunnya. Pada dasarnya data awal yang dikumpulkan Susenas masih dalam keadaan mentah dan tidak dapat langsung dianalisis. Data tersebut mesti melewati tahap pengolahan terlebih dahulu. Salah satu tahapan pengolahan data Susenas adalah imputasi data. Imputasi data dilakukan pada isian yang hilang atau bermasalah berdasarkan hasil proses pengecekan kembali kewajaran data atau kekonsistenan isian antar variabel. Selain itu, secara umum, imputasi juga dilakukan sebagai perlakuan terhadap *outlier* dalam salah satu upaya peningkatan kualitas data (BPS, 2017). Oleh karena itu sangat penting untuk menentukan metode imputasi terbaik yang akan digunakan.

Menurut Jerez, et al (2010), metode imputasi dibagi kedalam dua jenis, yaitu metode imputasi berbasis statistik dan metode imputasi berbasis *machine learning*. *Mean imputation*, *hot-deck imputation*, metode *regression*, dan *multiple imputation* adalah beberapa contoh metode imputasi berbasis statistik. Metode *Hot-deck Imputation* merupakan salah satu metode imputasi yang sering digunakan. Metode ini adalah penyempuranaan dari metode sebelumnya yaitu *mean imputation*, khususnya pada pendugaan varians yang *underestimate* (Hendrawati, 2015). Metode ini juga lebih cocok digunakan pada banyak jenis data dibandingkan metode *regression*, dan metode *multiple*

imputation, dikarenakan penggunaan metode ini relatif sederhana dan tidak diperlukannya asumsi yang rumit dibanding keduanya. Selain itu metode *Hot-deck Imputation* dapat dilakukan untuk imputasi pada berbagai tipe data numerik, kategorik, ataupun *mixeddata*.

Selanjutnya adalah metode imputasi berbasis *machine learning*. metode imputasi berbasis *machine learning* adalah proses imputasi yang memanfaatkan pembelajaran (learning/training) pada data untuk melakukan prediksi nilai yang akan diimputasi. Adapun metode C4.5, CN2, dan *K-Nearest Neighbor Imputation* (KNNI) adalah beberapa contoh metode imputasi berbasis *machine learning*. Menurut Batista (2002), dari ketiga metode tersebut, KNNI memberikan hasil yang paling baik. Selain itu, Troyanskaya (2001) membandingkan metode *Singular Value Decomposition* (SVD) dan KNNI. Hasilnya menunjukkan bahwa KNNI tampaknya menyediakan metode yang lebih kuat dan sensitif untuk estimasi *missing values* dibandingkan SVD. Keuntungan dalam penggunaan metode KNNI ini adalah tidak diperlukannya asumsi apapun, tidak diperlukan pembentukan model prediksi dan dapat mengatasi *missing values* baik pada data numerik maupun kategorik (Azizah, 2016).

Algoritma dan cara penanganan *missing values* yang berbeda pada kedua metode tentunya dapat menghasilkan hasil estimasi yang berbeda pula. Walaupun demikian, kedua metode tersebut sama-sama memanfaatkan variabel prediktor untuk melakukan proses imputasi serta tidak memerlukan asumsi yang rumit dalam penggunaannya. Oleh karena itu dalam penelitian ini, peneliti tertarik untuk menentukan metode imputasi terbaik antara metode imputasi berbasis statisitik yaitu *Hot-deck Imputation* dengan metode imputasi berbasis *machine learning* yaitu KNNI dalam menangani *missing values* dengan mengimplementasikan data Susenas untuk mengestimasi nilai *missing values*. Secara khusus, tujuan dari penelitian ini adalah untuk membandingkan ketepatan estimator dan performa komputasi yang dihasilakan oleh metode *Hot-deck Imputation* dan metode KNNI dalam mengatasi *missing values* yang diimplementasikan pada data Susenas Maret Tahun 2017, serta mengidentifikasi kelebihan dan kekurangan dalam melakukan imputasi data.

METODE

Landasan teori

Hot-deck Imputation melibatkan penggantian *missing values* menggunakan nilai-nilai lain berdasarkan konsep *similarity*. *Hot-deck Imputation* salah satu metode imputasi yang popular digunakan. Meskipun popular dalam praktiknya, literatur tentang sifat-sifat teoretis dari berbagai metode sangat jarang. Menurut Kowarik (2016), *Hot-deck Imputation* umumnya mengacu pada *Sequential Hot-deck Imputation*, yang berarti bahwa set data diurutkan dan nilai-nilai yang hilang diimputasikan secara berurutan berjalan melalui observasi demi observasi. Pengurutan data menggunakan variabel prediktor dipilih berdasarkan hubungannya dengan variabel yang akan diimputasi (Grau, 2004). Saat ini, metode dengan sifat teoritis yang lebih baik telah tersedia, tetapi metode *Hot-deck Imputation* masih cukup popular karena kesederhanaan dan kecepatannya.

Algoritma KNN pertama kali dijelaskan pada awal tahun 1950-an (Han, 2012). Mirip dengan metode *Hot-deck Imputation*, metode KNNI juga didasarkan pada pengamatan donor. Sebelum melakukan imputasi, metode KNNI perlu menentukan jumlah tetangga terdekat (k) yang akan digunakan. Suyundikov (2015) pada penelitiannya, menggunakan nilai RMSE untuk mencari nilai k optimum dari metode KNNI. Troyanskaya (2001) berpendapat bahwa metode imputasi kurang sensitif terhadap pilihan k dalam kisaran 10-20. Pilihan nilai k yang terlalu kecil mungkin dapat mengurangi keakuratan nilai imputasi, namun sebaliknya semakin besar nilai k yang digunakan akan berakibat pada penurunan performa komputasi (Suyundikov, 2015). KNNI memanfaatkan jarak terdekat dengan objek terkait sebagai donor untuk melakukan imputasi. Perhitungan jarak terdekat dilakukan menggunakan variabel prediktor. Jarak ini kemudian digunakan sebagai dasar penentuan tetangga terdekat. Setelah itu, nilai *missing values* diimputasi menggunakan rata-rata tertimbang dari k tetangga terdekatnya berdasarkan hasil perhitungan sebelumnya dengan jarak sebagai bobotnya. Secara matematis dapat dituliskan sebagai berikut:

$$y_j = \frac{\sum_{k=1}^{K} w_k y_{ik}}{\sum_{k=1}^{K} w_k} \quad \text{..... (1)}$$

$$w_k = \frac{1}{d(i,j)_k} \quad \text{..... (2)}$$

dimana:

K = banyaknya tetangga terdekat
y_j = nilai variabel *missing values* pada observasi ke-j
y_{ik} = nilai variabel donor pada observasi ke-i pada tetangga ke-k
w_k = bobot variabel tetangga ke-k
$d(i,j)_k$ = jarak observasi i ke j tetangga ke-k

Variabel prediktor adalah variabel yang dinilai terkait dengan variabel imputasi (Hendrawati, 2015. Dalam penelitian ini, variabel prediktor yang akan digunakan untuk mengestimasi *missing values* diperoleh dari penelitian oleh Lhing (2013) dan Sekhampu (2013). Analisis perbandingan dilakukan dengan melihat ketepatan estimator yang dihasilkan melalui nilai RMSE (*Root Mean Square Error*) dan MAPE (*Mean Absolute Percentage Error*) dari data hasil imputasi masing-masing metode. RMSE dan MAPE adalah salah satu *error measurement* yang digunakan untuk evaluasi kinerja pada suatu pemodelan. Perhitungan RMSE dan MAPE Secara matematis dapat dituliskan sebagai berikut:

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (\hat{y}_i - y_i)^2} \quad \text{..... (3)}$$

$$MAPE = \frac{100}{M} \sum_{i=1}^{M} \frac{|\widehat{y_i - y_i}|}{y_i} \quad \text{..... (4)}$$

dimana:

$\hat{y}_i$ = nilai prediksi observasi ke-i
y_i = nilai aktual observasi ke-i
M = jumlah peramalan

Metode Pengumpulan Data

Data yang digunakan dalam penelitian adalah data bangkitan berdistirbusi normal *univariate* dan data sekunder yang berasal dari data sampel Susenas Kor dan Konsumsi/Pengeluaran Maret 2017. Data bangkitan berdistribusi normal *univariate* dibangkitkan dengan jumlah observasi sebesar 100 dan 1000. Data sebesar 100 dan 1000 dimaksudkan agar dapat dilakukannya analisis pada data berjumlah kecil dan besar. Data akan dibangkitkan menggunakan bantuan *software* R. Variablel prediktor dibangkitkan secara *random* menggunakan *Package MASS* di *software* R. Kemudian, nilai *slope* dan *error* pada fungsi linier juga diperoleh secara *random* menggunakan bantuan *software R*. Fungsi linier pembentukan data bangkitan normal *univariate* adalah sebagai berikut:

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon \quad \text{..... (5)}$$

dimana:

Y = variabel bangkitan normal *univariate*
$\alpha_1, \alpha_2, \alpha_3$ = *slope*
X_1, X_2, X_3 = variabel predikor
ε = error $\sim N(0, \sigma^2)$

Data selanjutnya adalah data sampel Susenas Kor dan Konsumsi/Pengeluaran Maret 2017. Variabel yang digunakan adalah Lapangan usaha atau bidang pekerjaan utama kepala rumah tangga, proporsi art yang bekerja, tingkat pendidikan krt, usia krt, dan pengeluaran perkapita perbulan. Variabel pengeluaran perkapita perbulan digunakan sebagai variabel imputasi, sedangkan sisanya merupakan variabel prediktor. Adapun prosedur penarikan sampel dilakukan menggunakan *systematic random sampling* dengan *implicit stratified* yang digunakan adalah pengeluaran perkapita perbulan. Data sampel yang digunakan akan dibuat menjadi dua ukuran data, yaitu kecil dan besar. Ukuran data dalam jumlah kecil diambil berdasarkan rata-rata sampel rumah tangga susenas 2017 perkabupaten yakni sebesar 578 observasi, sedangkan ukuran data dalam jumlah besar diambil berdasarkan rata-rata sampel rumah tangga susenas 2017 perprovinsi yakni sebesar 8.743 observasi (untuk kemudahan analisis disesuaikan menjadi 600 observasi dan 9.000 observasi). Data dengan ukuran tersebut dimaksudkan agar dapat dilakukannya analisis pada data berjumlah kecil dan besar. Kedua jenis data tersebut kemudian di-treatment sebagai populasi.

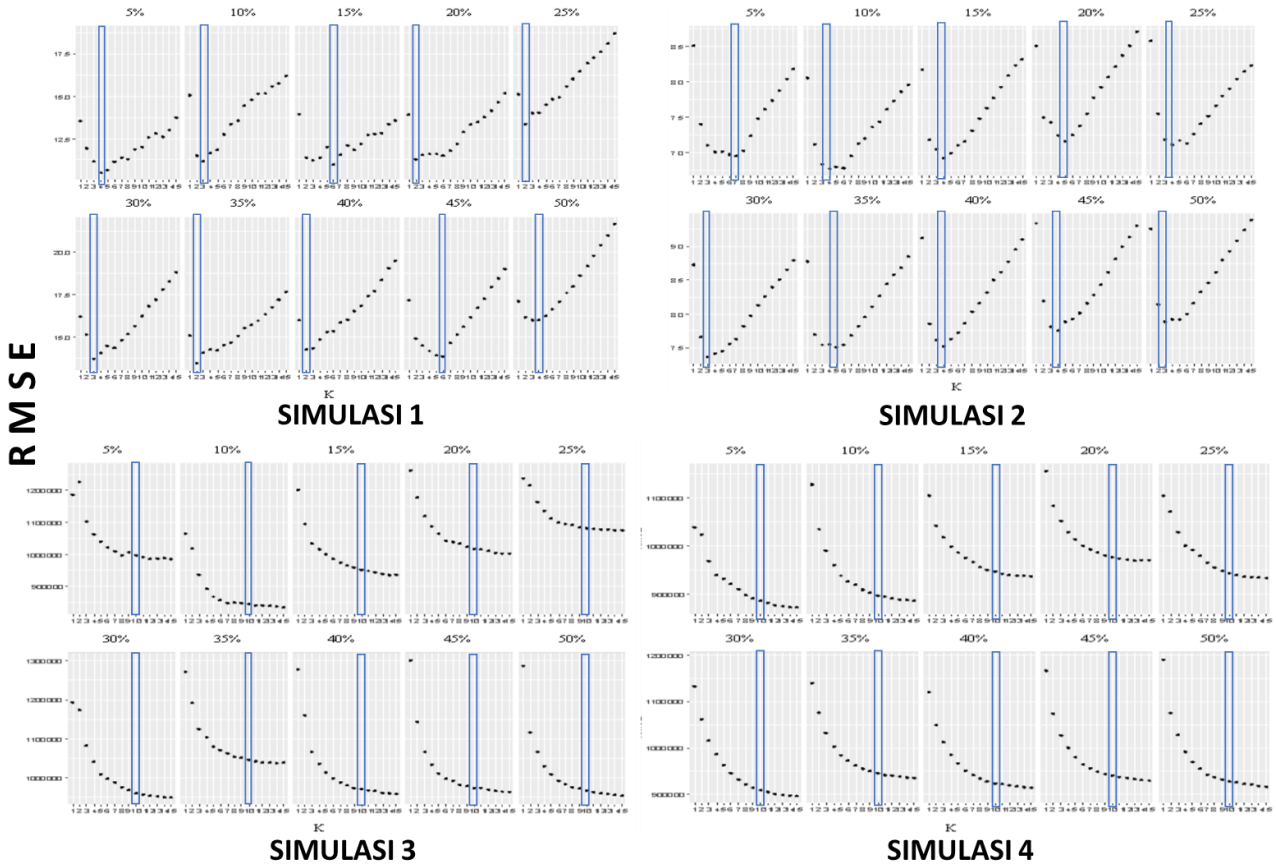
Metode Analisis

Pada penelitian ini, analisis dilakukan dalam bentuk simulasi. Simulasi akan dilakukan dalam tiga tahap. Tahap pertama adalah pembentukan dataset yang mengandung *missing values*. Pembentukan dataset ini dilakukan menggunakan mekanisme MCAR untuk sepuluh tingkatan *missing values* yakni 5 – 50 persen dengan interval 5 persen. Tahap kedua adalah proses imputasi data. Pada tahap ini, tiap dataset yang mengandung *missing values* tersebut akan diimputasi menggunakan metode *Hot-deck Imputation* dan metode KNNI. Metode *Hot-deck Imputation* yang digunakan pada analisis ini adalah *Sequential Hot-deck Imputation*. *Package VIM* dalam R digunakan untuk melakukan imputasi pada kedua metode tersebut.

Tahap ketiga adalah analisis. Sebelum dianalisis, terlebih dahulu dilakukan pemilihan nilai optimum k pada KNNI menggunakan perbandingan RMSE dari imputasi pada tiap k. Hasil imputasi dengan nilai k yang optimum akan digunakan untuk membandingkan hasil imputasi dengan metode *Hot-deck Imputation*. Ukuran RMSE dan MAPE digunakan untuk mengukur ketepatan estimator dari kedua metode imputasi. Sementara itu, penghitungan *running time* atau waktu yang diperlukan untuk melakukan proses imputasi digunakan untuk mengukur performa komputasi. Proses ini dilakukan dalam 10 kali pengulangan. Kemudian nilai rata-rata dari RMSE, MAPE, dan *running time* dari kedua metode dibandingkan dan dianalisis.

HASIL DAN PEMBAHASAN

Pemilihan K Optimum

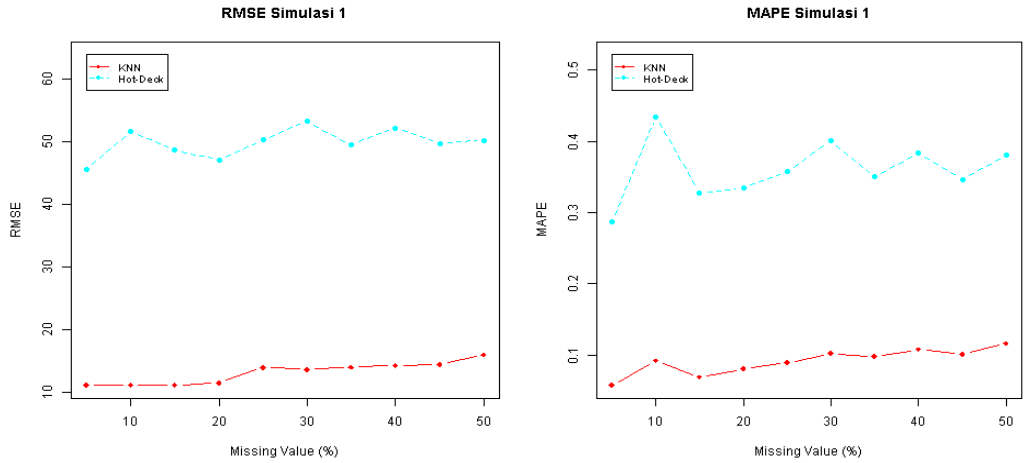


Gambar 1. Pemilihan k optimum

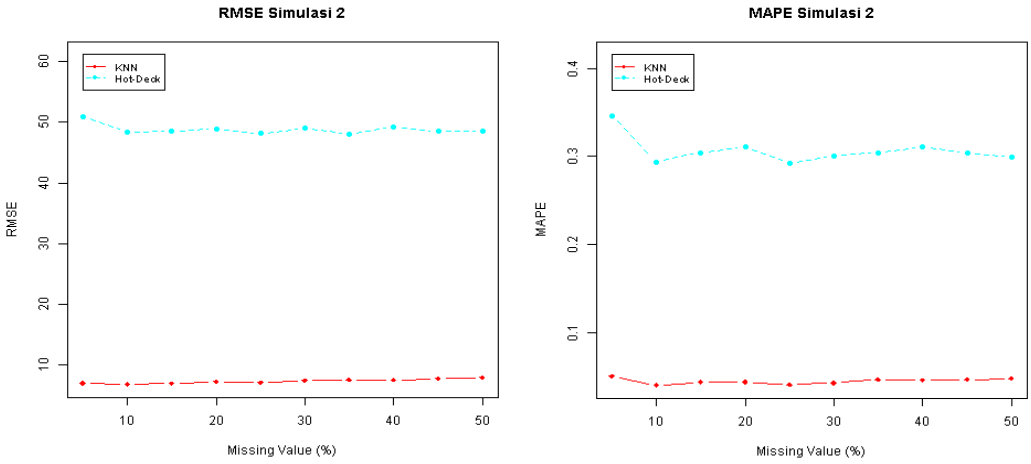
Sebelum melakukan analisis, dilakukan pemilihan k optimum untuk masing-masing simulasi. Gambar 1 menunjukkan efek dari jumlah tetangga terdekat, k, yang digunakan dalam metode KNNI. Hasil tersebut menunjukkan nilai RMSE untuk set data simulasi dengan masing-masing persentasi

missing values. Pada simulasi 1, nilai RMSE menurun pada nilai k dengan kisaran 1-6, sedangkan pada simulasi 2, nilai RMSE menurun pada nilai k dengan kisaran 1-7. Rata-rata nilai RMSE yang meningkatkan kinerja KNNI terjadi pada nilai k kisaran 3 (simulasi 1) dan k kisaran 4 (simulasi 2). Kemudian nilai RMSE meningkat seiring bertambahnya nilai k. Akibatnya kinerja imputasi menjadi kurang sensitif terhadap sisa nilai k yang ada. Sedangkan pada simulasi 3 dan 4 nilai RMSE menurun seiring peningkatan nilai k, dan menjadi hampir sama untuk nilai k dalam kisaran 10 keatas. Hal ini dikarenakan kinerja imputasi mengalami peningkatan yang kurang sensitif terhadap nilai k kisaran 10-15. Dengan demikian, nilai k sebesar 3 (3-NNI) pada simulasi 1, nilai k sebesar 4 (4-NNI) pada simulasi 2, dan nilai k sebesar 10 (10-NNI) pada simulasi 3 dan 4 akan digunakan sebagai pembanding dengan metode *Hot-deck Imputation*.

Analisis Data Bangkitan Normal *Univariate*



Gambar 2. Perbandingan RMSE dan MAPE simulasi 1

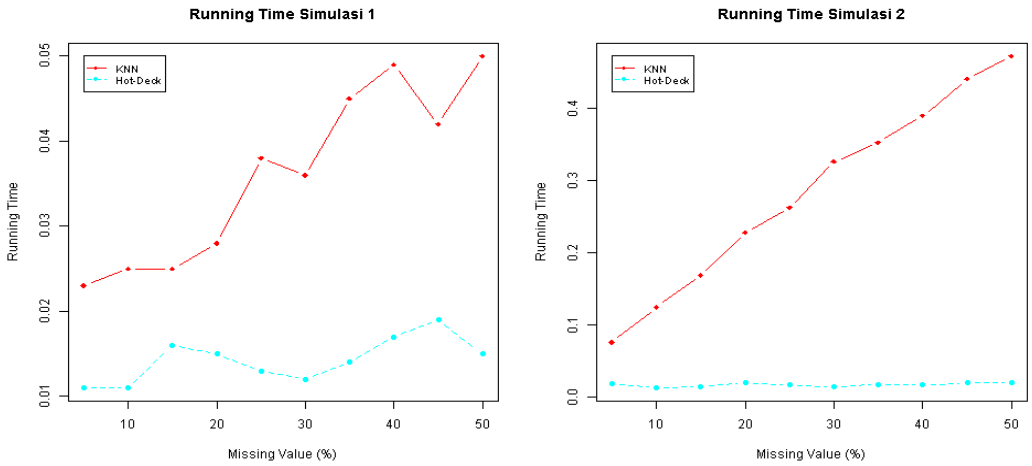


Gambar 3. Perbandingan RMSE dan MAPE simulasi 2

Analisis Data Bangkitan Normal *Univariate* akan dilakukan melalui dua simulasi. Simulasi pertama pada data dengan jumlah observasi 100 (Simulasi 1), dan simulasi kedua pada data dengan jumlah observasi 1000 (Simulasi 2). Simulasi 1 dan 2 merupakan analisis yang dilakukan pada tipe data berdistribusi normal.

Gambar 2 dan 3 menunjukkan hasil kinerja imputasi yang dihasilkan pada simulasi 1 dan 2. Berdasarkan hasil kedua simulasi diatas terlihat bahwa peningkatan jumlah data menyebabkan akurasi dari KNNI dan *Hot-deck Imputation* semakin baik. Hal ini dapat dilihat dari nilai RMSE dan MAPE pada simulasi 2 yang memiliki nilai lebih kecil dibandingkan simulasi 1. Berdasarkan hasil kedua simulasi, pada data bangkitan berdistribusi Normal *Univariate*, menunjukkan bahwa metode

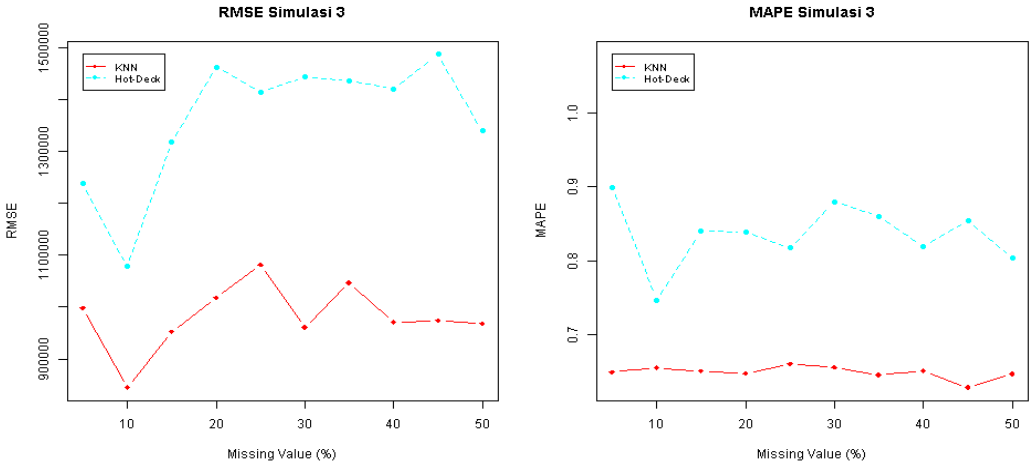
KNNI menghasilkan ketepatan estimator yang secara konsisten lebih baik daripada metode *Hot-deck Imputation*. Hal ini ditunjukkan dari nilai RMSE dan MAPE yang lebih rendah pada tiap persentase *missing values* yang ada.



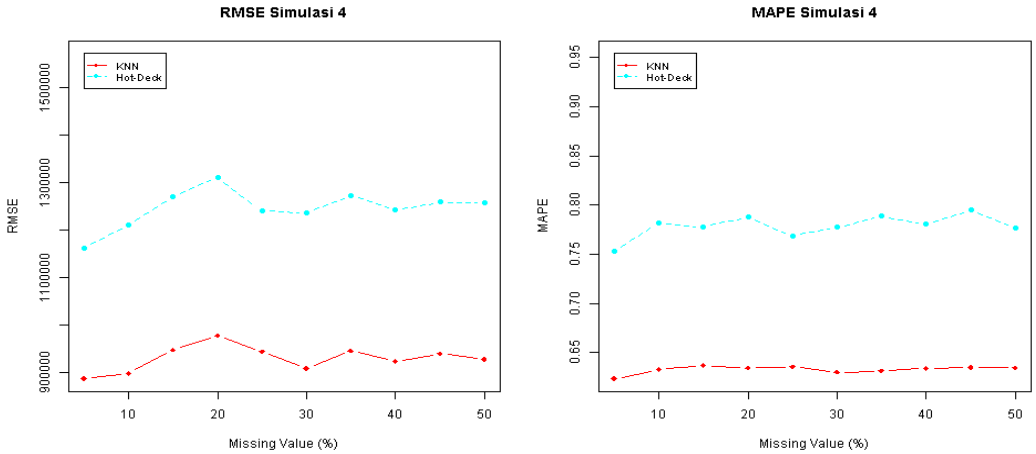
Gambar 4. Perbandingan *Running time* simulasi 1 dan simulasi 2

Gambar 4 menunjukkan *running time* yang diperlukan untuk melakukan imputasi pada simulasi 1 dan simulasi 2. Berdasarkan hasil kedua simulasi diatas terlihat bahwa peningkatan jumlah data pada dataset dan seiring meningkatnya persentase *missing values* menyebabkan *running time* untuk KNNI meningkat. Sedangkan pada metode *Hot-deck Imputation running time* yang diperlukan relatif sama seiring peningkatan jumlah data dan pada tiap persentase *missing value* yang ada. Berdasarkan hasil kedua simulasi, pada data bangkitan berdistribusi Normal *Univariate*, menunjukkan bahwa metode *Hot-deck Imputation* menghasilkan performa komputasi yang secara konsisten lebih baik daripada metode KNNI.

Analisis Data Sampel Susenas Kor dan Konsumsi/Pengeluaran Maret 2017



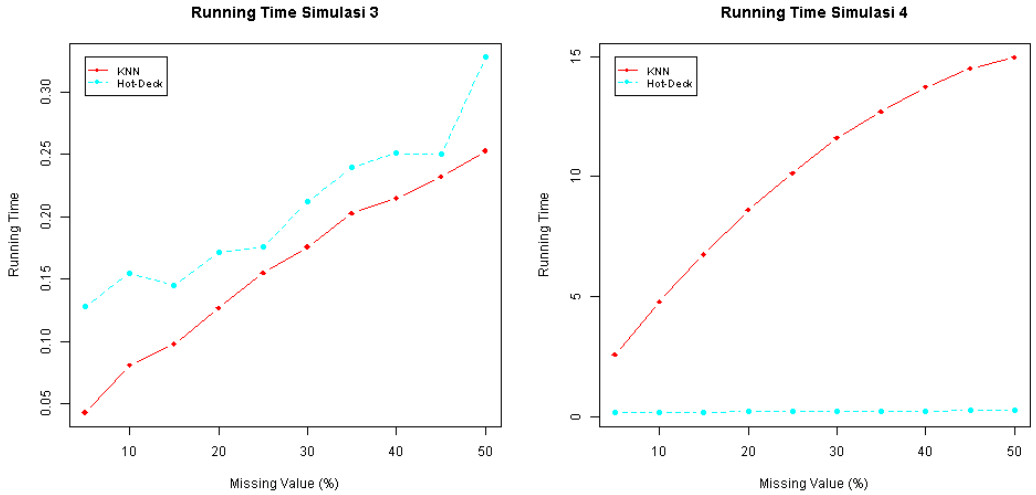
Gambar 5. Perbandingan RMSE dan MAPE simulasi 3



Gambar 6. Perbandingan RMSE dan MAPE simulasi 4

Analisis Data Sampel Susenas Kor dan Konsumsi/Pengeluaran Maret 2017 akan dilakukan melalui dua simulasi. Simulasi pertama akan dilakukan pada data dengan jumlah observasi 600 (Simulasi 3). Sedangkan, simulasi kedua akan dilakukan pada data dengan jumlah observasi 9000 (Simulasi 4). Simulasi 3 dan simulasi 4 merupakan analisis yang dilakukan pada tipe data berdistribusi tidak normal.

Gambar 5 dan 6 menunjukkan hasil kinerja imputasi pada simulasi 3 dan simulasi 4. Berdasarkan hasil diatas terlihat bahwa peningkatan jumlah data menyebabkan akurasi dari KNNI dan *Hot-deck Imputation* semakin baik. Hal ini dapat dilihat dari nilai RMSE dan MAPE pada simulasi 4 yang memiliki nilai lebih kecil dibandingkan simulasi 3. Berdasarkan hasil kedua simulasi, pada Data Sampel Susenas Kor dan Konsumsi/Pengeluaran Maret 2017, menunjukkan bahwa metode KNNI menghasilkan ketepatan estimator yang secara konsisten lebih baik daripada metode *Hot-deck Imputation*.



Gambar 7. Perbandingan *Running time* simulasi 3 dan simulasi 4

Gambar 7 menunjukkan *running time* pada simulasi 3 dan simulasi 4. Berdasarkan hasil kedua simulasi diatas terlihat bahwa peningkatan jumlah data pada dataset dan bertambah seiring meningkatnya persentase *missing values* menyebabkan *running time* yang diperlukan dalam proses imputasi KNNI meningkat. Pada metode *Hot-deck Imputation*, *running time* yang diperlukan relatif sama seiring peningkatan jumlah data. Pada simulasi 3, *running time* pada metode *Hot-deck Imputation* bertambah seiring meningkatnya persentase *missing values.* Namun, pada simulasi 4, *running time* pada proses *Hot-deck Imputation* cenderung sama untuk setiap persentase *missing*

values. Berdasarkan hasil kedua simulasi tersebut, Performa komputasi yang dihasilkan metode *Hot-deck Imputation* secara konsisten lebih baik dibanding metode KNNI seiring peningkatan jumlah dataset yang ada.

Analisis Perbandingan Metode *Hot-deck Imputation* dan KNNI

Berdasarkan hasil analisis sebelumnya pada keempat simulasi, dapat dilakukan analisis lebih lanjut mengenai perbandingan hasil imputasi dari metode *Hot-deck Imputation* dan KNNI. Keempat simulasi tersebut menunjukkan bahwa metode KNNI menghasilkan ketepatan estimator yang secara konsisten lebih baik daripada metode *Hot-deck Imputation*. Hal ini ditunjukkan dari nilai RMSE dan MAPE yang lebih rendah. Selain itu, ditunjukkan bahwa hasil imputasi semakin baik seiring dengan bertambahnya jumlah dataset yang digunakan. Hal ini disebabkan karena semakin banyaknya observasi-observasi yang mendekati atau mirip dengan karakteristik data yang mengandung *missing values*, sehingga donor yang digunakan semakin mendekati karakteristik data tersebut. Kinerja imputasi pada data bangkitan Normal *Univariate* menunjukkan hasil yang lebih baik dibandingkan pada data sampel Susenas Kor dan Konsumsi/Pengeluaran Maret 2017. Hal ini ditunjukkan dari nilai MAPE pada masing masing metode untuk setiap simulasi yang ada.

Walaupun demikian, performa komputasi menunjukkan hasil yang berlawanan dengan ketepatan estimator. Tiga dari empat simulasi yang dilakukan menunjukkan bahwa metode *Hot-deck Imputation* menghasilkan performa komputasi yang secara konsisten lebih baik daripada metode KNNI. Selain itu, melalui hasil simulasi tersebut, ditunjukkan performa komputasi metode KNNI memerlukan *running time* yang semakin lama seiring dengan bertambahnya jumlah dataset yang digunakan. Hal ini ditunjukkan pada peningkanan *running time* untuk simulasi data bangkitan Normal *Univariate* serta simulasi pada data sampel Susenas Kor dan Konsumsi/Pengeluaran Maret 2017. Sedangkan pada metode *Hot-deck Imputation*, performa komputasi yang dihasilkan menunjukkan bahwa *running time* pada observasi kecil dan besar tidak terlalu jauh berbeda. Berikut adalah ringkasan hasil imputasi untuk kedua metode pada setiap simulasi yang dilakukan.

Tabel 1. Ringkasan rata-rata hasil imputasi

Metode	Simulasi	MAPE (%)	RMSE	*Running Time* (detik)
Hot-deck Imputation	1	36,029	49,802	0,014
	2	30,665	48,833	0,016
	3	83,626	1.363.741,172	0,206
	4	77,856	1.245.706,136	0,226
KNNI	1	9,106	13,166	0,036
	2	4,500	7,321	0,284
	3	64,964	981.714,456	0,158
	4	63,270	929.524,460	10,042

Hasil ini menandakan bahwa terdapat konflik kualitas imputasi antara akurasi (ketepatan estimator) dengan waktu (performa komputasi). Berdasarkan penjelasan sebelumnya, imputasi pada kedua metode memberikan hasil yang lebih baik jika dilakukan pada data berdistribusi normal dengan jumlah observasi besar. Pada kenyataannya, hampir semua data mentah yang masih dalam proses pengolahan biasanya tidak memenuhi asumsi ini. Oleh karena itu, alternatif yang dapat dilakukan jika ingin meningkatkan ketepatan estimator adalah dengan melakukan tranformasi data menjadi normal terlebih dahulu, dengan resiko adanya tahapan tambahan yang harus dilakukan sebelumnya. Selain itu, alternatif lainnya yang dapat dilakukan adalah dengan melakukan imputasi

pada dataset dengan jumlah observasi besar. Namun, alternatif-alternatif tersebut menyebabkan terjadi penurunan performa komputasi. Hal ini ditandai dengan proses imputasi akan memakan waktu yang lebih lama. Oleh karena itu, jika proses imputasi dilakukan pada beberapa variabel saja maka metode KNNI akan lebih baik digunakan dibandingkan metode *Hot-deck Imputation*. Sebaliknya, jika imputasi dilakukan pada banyak variabel, terutama yang sifatnya mikro, maka metode *Hot-deck Imputation* akan lebih baik digunakan dibandingkan metode KNNI.

KESIMPULAN

Berdasarkan hasil simuasi dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa, implementasi metode *Hot-deck Imputation* dan metode KNNI pada data Susenas Maret Tahun 2017 menunjukkan bahwa, dari sisi kualitas akurasi, metode KNNI menghasilkan ketepatan estimator yang secara konsisten lebih baik dibandingkan metode *Hot-deck Imputation*, baik dari nilai RMSE dan MAPE. Sedangkan, jika dilihat dari sisi kualitas waktu, metode *Hot-deck Imputation* menghasilkan performa komputasi yang secara konsisten lebih baik dibandingkan metode KNNI.

Metode *Hot-deck Imputation* dan metode KNNI Menghasilkan ketepatan estimator yang lebih baik pada data berdistribusi normal dan meningkat seiring dengan bertambahnya jumlah data yang digunakan. Adapun pada metode KNNI Performa komputasi sangat dipengaruhi oleh jumlah dataset yang digunakan dan tingkat persentase *missing values*. Sedangkan metode *Hot-deck Imputation* tidak terlalu dipengaruhi oleh jumlah dataset yang digunakan dan tingkat *missing values*.

DAFTAR PUSTAKA

Azizah Nur. (2016). *Analisis Perbandingan Metode Multiple Imputation dan K-nearest neighbor imputation dalam Mengatasi Missing data.* [Skripsi]. Sekolah Tinggi Ilmu Statistik Jakarta.

Badan Pusat Statistik. *Survei Sosial Ekonomi Nasional (SUSENAS) Kor 2017*. (2017). Jakarta: BPS.

Batista, Gustavo E. A. P. A. Maria Carolina Monard. (2002). *A Study of K-Nearest Neighbour as an Imputation Method.* Second International Conference on Hybrid Intelligence, 8.

Biemer, P. P. & Lyberg, L. E. (2003). *Introduction to survey quality*. New Jersey: John Wiley & Sons, Inc.

Han, Jiawei., Kamber, Micheline., dan Pei, Jian. (2012). *Data Mining: Concepts and Techniques Third Edition*. Elsevier.

Handayani. (2011). *Perbandingan Metode Predictive Mean Matching dan Propensiti Scoring dalam Mengestimasi Data Hilang.* [Skripsi]. Sekolah Tinggi Ilmu Statistik Jakarta.

Hendrawati, Triyani. (2015). Kajian Metode Imputasi Dalam Menangani *Missing data*. *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS.* Surakarta: Universitas Muhammadiyah Surakarta.

Grau EA, Frechtel PA, Odom DM, Painter D. (2004). *A simple evaluation of the imputation procedures used in nsduh*. Toronto: American Statistical Association.

Jerez, J.M., dan Molina, I., (2010). *Missing data imputation using statistical and machine learning methods in a real breast cancer problem*. Artificial intelligence in medicine, 105-115.

Kaiser, J. (2014). Dealing with Missing valuess in Data. Journal of Systems Integration, 42.

Kang, Hyun. (2013). *The prevention and handling of the missing data*. Korea: Chung-Ang Universtiy College of Medicine.

Kowarik, Alexander., dan Templ, Matthias (2016), *Imputation with the R Package VIM*. Journal of Statistical. Foundation for Open Access Statistics, vol. 74(i07).

Little, R.J.A., dan Rubin, DA (2002), *Statistical Analysis with Missing data* (2 ed). New York: John Willey and Sons.

Lhing, Nem Nei. et. al. (2013). An Analysis of Factor Influencing Household Income: A Case Study of PACT Microfinance in Kayukpadaung Township of Myanmar. *American Journal of Human Ecology*.

Luengo, J. (2009). *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*. Elsevier Ltd.

Suyundikov, Anvar et. al. (2015). *Accounting for Dependence Induced by Weighted KNN Imputation in Paired Samples, Motivated by a Colorectal Cancer Study.* Taiwan: National Taiwan University.

Troyaska, Olga, et. Al. (2001). *Missing values estimation methods for DNA microarrays.* Bioinformatics, 17:520-525.

Sekhampu, T.J. (2013). *Anaysis Of the Factor Influencing Household Expenditure in A South African Township*. South Afica: Nonth-West University.