ESTIMATOR NADARAYA-WATSON DENGAN KERNEL ORDE BERHINGGA DAN TAK HINGGA

Maria Suci Apriani

Dosen Program Studi Pendidikan Matematika, FKIP, Universitas Sanata Dharma
Alamat Korespondensi: Kampus III Paingan, Maguwoharjo, Depok, Sleman, Yogyakarta
Email: *maria.suci@usd.ac.id*

ABSTRACT

This article compared the performance between finite order kernel (normal) and infinite order kernel (sinus and cosinus) in Nadaraya Watson estimator by comparing *the MSE values. This research used literature study method based on the article* entitled *"Minimally Biased Nonparametric Regression and Autoregression" by Timothy and Dimitris. The estimation result* based on *MSE* values, *showed that three kernels have different strengths on estimation process.*

Keywords *: Nonparametric regression, Nadaraya Watson estimator, finite order kernel, infinite order kernel.*

1. PENDAHULUAN

Andaikan terdapat n pengamatan pasangan $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ sampel dengan X_i adalah variabel prediktor dan Y_i adalah variabel respon, maka hubungan linear antara variabel respon dengan variabel prediktor yang memenuhi model $Y_i = r(X_i) + \varepsilon_i$ dapat dicari. Mengestimasi fungsi regresi $r(X_i)$ dapat dilakukan dengan pendekatan parametrik maupun non parametrik. Pendekatan nonparametrik dilakukan jika tidak ada asumsi tentang fungsi $r(X_i)$ dan akan diestimasi dengan teknik *smoothing* menggunakan estimator kernel. Menurut Hardle (1994) ketepatan suatu pemulus kernel sebagai estimator dari r ditentukan oleh dua hal yaitu *bandwidth* dan fungsi kernel yang digunakan sebagai bobot. Ukuran tingkat kesalahan suatu estimator dapat dilihat dari MSE (*Mean Squared Error*) atau MISE (*Mean Integrated Squared Error*). Semakin kecil nilai MSE atau MISE, maka hasil estimasi akan semakin mendekati fungsi aslinya. Menurut Timothy dan Dimitris (2003) jika kernel K mempunyai orde v dan fungsi kepadatan r mempunyai turunan kontinu sebanyak k kali maka:

$$\text{Bias } (\hat{r}(x)) = C_{K,r}(x)\, h^n + o(h^n),$$

Ketika fungsi r cukup mulus atau dapat dideferensialkan sebanyak k kali dimana $v \geq k$, maka bias $\hat{r}(x)$ dapat direduksi menjadi $o(h^k)$ dengan secara tepat memilih kernel dengan orde yang lebih besar dari banyaknya diferensial. Namun kita kesulitan untuk menentukan orde kernel berapakah yang harus dipilih. Sehingga ditetapkan suatu kernel yang memiliki "*infinite orde*". Menggunakan data hasil pengamatan rata-rata volume air sungai di Indonesia yang pengalirannya lebih dari 1000 km^2, akan dibandingkan *performance* estimator Nadaraya Watson ketika menggunakan kernel orde tak hingga dan berhingga dengan melihat nilai MSE-nya.

2. KERNEL ORDE BERHINGGA

Estimator fungsi regresi yang diusulkan oleh Nadaraya-Watson untuk fungsi densitas f yang tidak diketahui adalah:

$$\hat{r}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)Y_i}{\frac{1}{n}\sum_{k=1}^{n} K_h(x - X_k)} = \frac{\hat{g}(x)}{\hat{f}(x)}$$

Suatu kernel dikatakan berorde 2 jika $K(x) \geq 0$, $\int K(x)\,dx = 1$, $\int xK(x)\,dx = 0$ dan $\int x^2K(x)\,dx < \infty$, untuk semua nilai $x \in \mathbb{R}$.

Definisi 2.1 (Hardle, 1991). Secara umum fungsi Kernel dengan bandwidth h didefinisikan sebagai berikut

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), -\infty<x<\infty \text{ dan } h>0$$

yang memenuhi sifat-sifat:

(1) $K(x)\geq 0$,

(2) $\int K(x)dx=1$,

(3) $\int xK(x)dx=0$,

(4) $\int x^2K(x)dx\neq 0$,

(5) $\int x^2K(x)dx<\infty$, untuk semua nilai $x\in\mathbb{R}$. .

Contoh fungsi kernel berorde berhingga, antara lain: Kernel Uniform, Kernel Triangle, Kernel Epanechnikov, Kernel Quartic, Kernel Triweight, Kernel Cosinus, Kernel Gausian atau Normal. Kernel Gausian memiliki fungsi sebagai berikut

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), -\infty<x<\infty$$

Teorema 2.1 (Wand dan Jones, 1995). Bila $\hat{f}_h$ estimator densitas kernel maka

(i) Bias $\left(\hat{f}_h(x)\right) = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2),\ h\to 0$

(ii) var $\left(\hat{f}_h(x)\right) = (nh)^{-1} f(x) \| K \|_2^2 + o\left((nh)^{-1}\right)$,

untuk $nh\to\infty$

Teorema 2.2 (Wand dan Jones, 1995). Bila $\left(\hat{f}_h(x)\right)$ estimator densitas kernel maka

$$MSE\left(\hat{f}_h(x)\right) = (nh)^{-1} f(x) \| K \|_2^2 \frac{h^4}{4} (f''(x)$$

$$\mu_2(K))^2 + o((nh))^{-1} + o(h^4), .\ h\to 0,\ nh\to\infty.$$

Menurut Hardle (1991) nilai-nilai statistik pembilang dari estimator Nadaraya-Watson dengan fungsi kernelnya mempunyai orde dua adalah sebagai berikut:

$$\text{Bias } \left(\hat{g}(x)\right) = \frac{h^2}{2} g''(x)\mu_2(K) + o(h^2),\ h\to 0$$

$$\text{var } \left(\hat{g}(x)\right) = (nh)^{-1} f(x)s^2(x) \| K \|_2^2 + o((nh))^{-1},$$

untuk $nh\to\infty$

$$MSE\left(\hat{g}(x)\right) = (nh)^{-1} f(x)s^2 \| K \|_2^2 \frac{h^4}{4} (g''(x)\mu_2$$

$$(K))^2 + o((nh))^{-1+} o(h^4),\ h\to 0,\ nh\to\infty.$$

dengan $s^2(x) = E[Y^2|X = x]$.

3. KERNEL ORDE TAK HINGGA

Definisi 3.1 (Berg, 2008). $K(x)$ dikatakan berorder tak hingga jika memenuhi

$$\int_{-\infty}^{\infty} x^i K(x)dx = 0,\ i = 1,2, \ldots$$

Definisi 3.2 (McMurry dan Politis, 2003). Sebuah flat-top Kernel K dengan order tak hingga secara umum dibentuk melalui Transformasi Fourier λ, yaitu untuk nilai tetap $c > 0$

$$\lambda(s) = \begin{cases} 1 & jika|s| \leq c \\ g(|s|) & jika|s| > c \end{cases},$$

dengan fungsi g dipilih sehingga membuat $\lambda(x)$, $\lambda^2(s)$ dan $s\lambda(s)$ dapat diintegralkan. Flat-top Kernel diberikan sebagai berikut:

$$K(x) = \frac{1}{2\pi} \int_{-\infty}^{-\infty} \lambda(s)e^{-isx}ds.$$

Berikut diberikan contoh yang memenuhi definisi di atas.

1) Diberikan fungsi $\lambda(x)$ sebagai berikut:

$$\lambda(s) = \begin{cases} 1 & jika|s| \leq 1 \\ 0 & jika|s| > 1 \end{cases}.$$

Menurut definisi: $K(x)$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \lambda(s)e^{isx}ds$$

$$= \frac{1}{2\pi} \left(\int_{-\infty}^{-1} 0.e^{-isx}ds + \int_{-1}^{1} 1.e^{-isx}ds + \int_{1}^{-\infty} 0.e^{-isx}ds \right)$$

$$= \frac{\sin(x)}{\pi x}$$

2. Diberikan fungsi $\lambda(s)$ sebagai berikut:

$$\lambda(s)=\begin{cases}1 & jika\ |s|\le 1/2\\ 2(1-|s|) & jika\ 1/2<|s|\le 1\\ 0 & jika\ |s|>1\end{cases}$$

Kernel yang bersesuaian adalah

$$K(x)=\frac{2\left(\cos\left(x/2\right)-\cos(x)\right)}{\pi x^2}.$$

Asumsi 3.1 Ketika $n\to\infty$, bandwidth $h\to 0$ dan $nh\to\infty$.

Asumsi 3.2 ε_i adalah random error dengan asumsi independen, $E(\varepsilon_i|X_i=x)=0$ dan $E(\varepsilon^2{}_i|X_i=x)=\sigma^2$.

Asumsi 3.3 (X_1,Y_1), (X_2,Y_2), …, (X_n,Y_{2n}) berdistribusi identik dan independen dengan densitas *f*.

Lemma 3.1 Jika x berada dalam interval terbuka dimana $f(x)$ mempunyai turunan kontinu terbatas p dan $r(x)$ mempunyai turunan kontinu terbatas q, maka berdasarkan asumsi 3.1 dan 3.2:

a. $E[\hat{f}(x)]-f(x)=o(h^p)$

b. $E[\hat{g}(x)]-g(x)=o(h^k)$

dengan $k=min\{p,q\}$.

Asumsi 3.4 Titik x merupakan titik kontinu dari $\sigma^2(x)$, $f(x)>C$ untuk $C>0$ dan fungsi r serta fungsi f masing-masing terdiferensial di sekitar x.

Lemma 3.2 Jika x berada dalam interval terbuka dimana $f(x)$ mempunyai turunan kontinu terbatas p dan $r(x)$ mempunyai turunan kontinu terbatas q, berdasarkan asumsi 3.1 –3.4 maka:

a. $\text{var}\left[\hat{f}(x)\right]=\frac{f(x)}{nh}\int_{-\infty}^{\infty}K^2(z)ds+o\left(\frac{1}{nh}\right)+O\left(\frac{1}{n}\right)$

b. $\text{var}\left[\hat{g}(x)\right]=\frac{\left(r^2(x)+\sigma^2(x)\right)f(x)}{nh}\int_{-\infty}^{\infty}K^2(z)ds+o\left(\frac{1}{nh}\right)+O\left(\frac{1}{n}\right)$

c. $\text{cov}\left[\hat{f}(x),\hat{g}(x)\right]=\frac{r(x)f(x)}{nh}\int_{-\infty}^{\infty}K^2(z)dz+o\left(\frac{1}{nh}\right)+O\left(\frac{1}{n}\right).$

Akibat 3.1 Berdasarkan asumsi 3.1 serta lemma 3.1 dan lemma 3.2 maka nilai *MSE* dari masing-masing $\hat{f}(x)$ dan $\hat{g}(x)$:

a. $MSE(\hat{f}(x))=O\left(\frac{1}{n}\right)$

b. $MSE(\hat{g}(x))=O\left(\frac{1}{n}\right).$

Bukti:

a. Berdasarkan definisi 2.4 dan lemma 3.1 dan lemma 3.2 maka:

$$MSE\left(\hat{f}(x)\right)=\frac{1}{nh}\int_{-\infty}^{\infty}K^2(s)f(x)ds+o\left(\frac{1}{nh}\right)+O\left(\frac{1}{n}\right)+\left[o(h^p)\right]^2$$

Ketika $n\to\infty$ maka nilai $MSE(\hat{f}(x))$ secara asimtotik adalah

$$MSE(\hat{f}(x))=O\left(\frac{1}{n}\right).$$

b. Berdasarkan definisi 2.4 dan lemma 3.1 dan lemma 3.2 maka:

$$MSE\left(\hat{g}(x)\right)=\frac{\left(r^2(x)+\sigma^2(x)\right)f(x)}{nh}\int_{-\infty}^{\infty}K^2(s)ds+o\left(\frac{1}{nh}\right)+O\left(\frac{1}{n}\right)+\left[o(h^k)\right]^2.$$

Ketika $n\to\infty$ maka nilai $MSE(\hat{g}(x))$ secara asimtotik adalah

$$MSE(\hat{g}(x))=O\left(\frac{1}{n}\right).$$

4. STUDI KASUS

Melalui regresi nonparametrik dengan menggunakan estimator Nadaraya Watson, akan dilihat *performance* antara kernel yang berorde tak hingga dan berhingga dengan membandingkan nilai MSE. Data yang digunakan adalah hasil pengamatan rata-rata volume air sungai di Indonesia yang pengalirannya lebih dari 1000 km^2. Data tersebut diambil dari Statistik Indonesia, *Statistical Yearbook of Indonesia 2013* yang dapat dilihat pada situs resmi Badan Pusat Statistik

(BPS). Fungsi kernel berorde tak hingga yang digunakan adalah

$$K(x) = \frac{\sin(x)}{\pi x} \text{ dan } K(x) = \frac{2\left(\cos\left(x/2\right) - \cos(x)\right)}{\pi x^2},$$

sedangkan untuk fungsi kernel orde berhingga yang digunakan adalah kernel Gausian atau Normal. Proses estimasi dengan pendekatan nonparametrik salah satu syarat yang harus dipenuhi adalah data kontinu. Pada studi kasus ini, penulis menggunakan data sungai di Indonesia yang daerah pengalirannya lebih dari 1000 km^2 tahun 2010 dimana variabel independen yaitu tinggi aliran air (juta m) dan volume air (juta dam^3) sebagai variabel dependen. Data aliran sungai dalam penelitian ini digunakan untuk membandingkan *performance* antara estimator dengan fungsi kernel yang berorde berhingga dan tak hingga.

4.1 Pengolahan Data dengan Program R

Proses yang dilakukan dalam melakukan pengolahan data dengan R untuk melakukan estimasi adalah sebagai berikut:

1. Masukkan data berpasangan (x_i, y_i)
2. Masukkan kernel yang digunakan sebagai pembanding. Kernel yang digunakan adalah sebagai berikut.

Kernel normal:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), -\infty < x < \infty$$

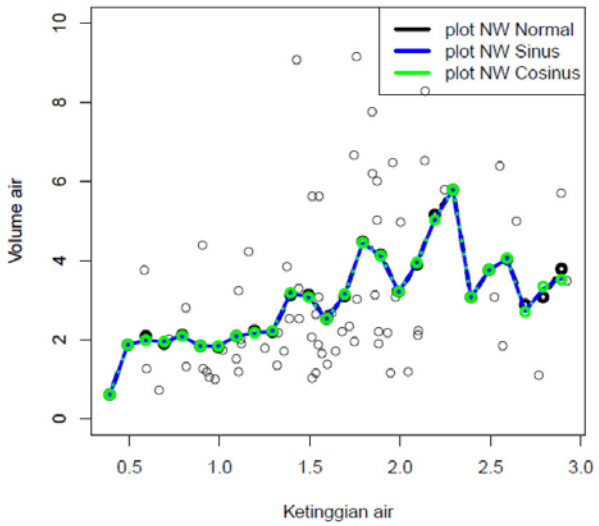
Kernel sinus $K(x) = \frac{\sin(x)}{\pi x}$ dan Kernel cosinus:

$$K(x) = \frac{2\left(\cos\left(x/2\right) - \cos(x)\right)}{\pi x^2}$$

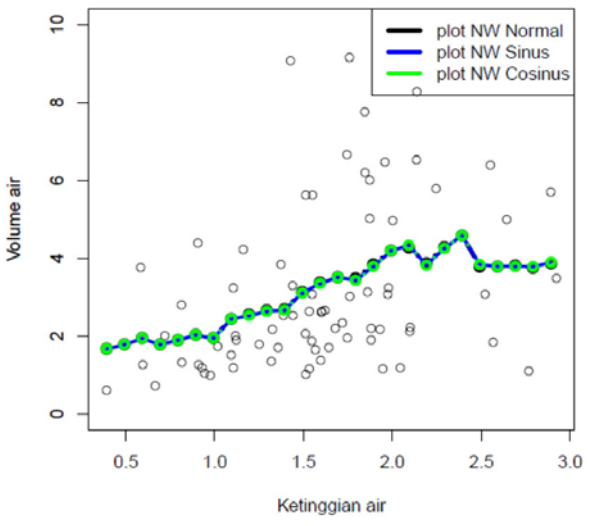
3. Masukkan nilai kelipatan untuk titik x yang akan diestimasi
4. Masukkan nilai *bandwidth*.
5. Plot pasangan data (x_i, y_i)
6. Plot hasil estimasi dengan kernel orde berhingga (normal)
7. Plot estimasi dengan kernel orde tak hingga (sinus dan cosinus)
8. Mendapatkan nilai MSE dari ketiga kernel
9. Membandingkan antara ketiga nilai MSE dari ketiga kernel

Kernel yang tersedia pada program R hanyalah kernel berorde berhingga. Sedangkan untuk melakukan proses dengan menggunakan kernel berorde tak hingga, penulis harus membuat program tersendiri.

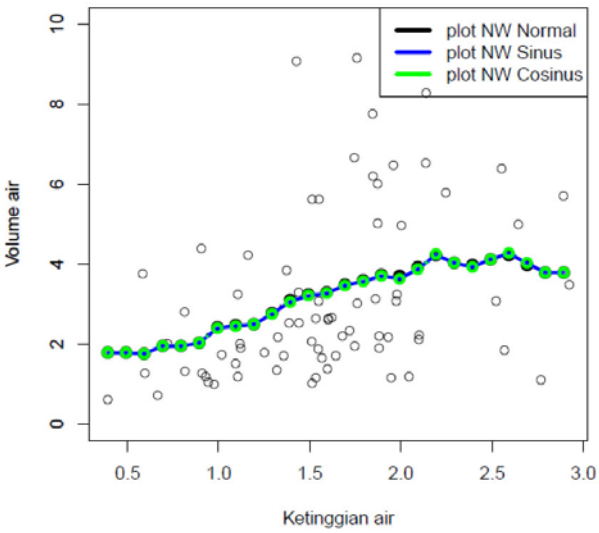
Berikut grafik hasil proses estimasi menggunakan data aliran sungai dengan nilai kelipatan titik-titik estimasi 0,1 dan 0,7



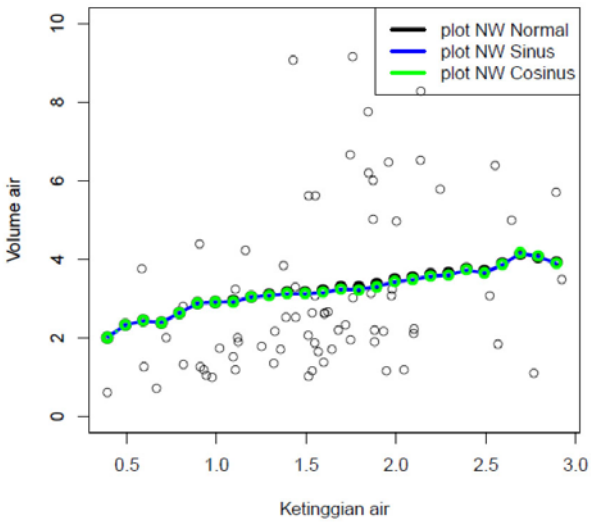
a. Grafik dengan *bandwidth* 0.13



b. Grafik dengan *bandwidth* 0,3445996

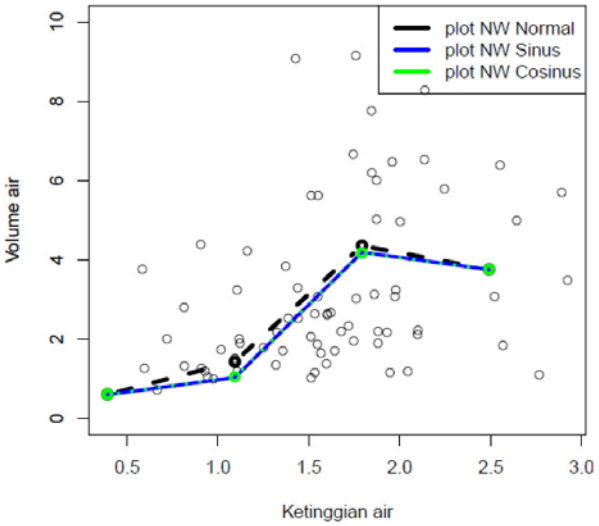


c. Grafik dengan *bandwidth* 0.5

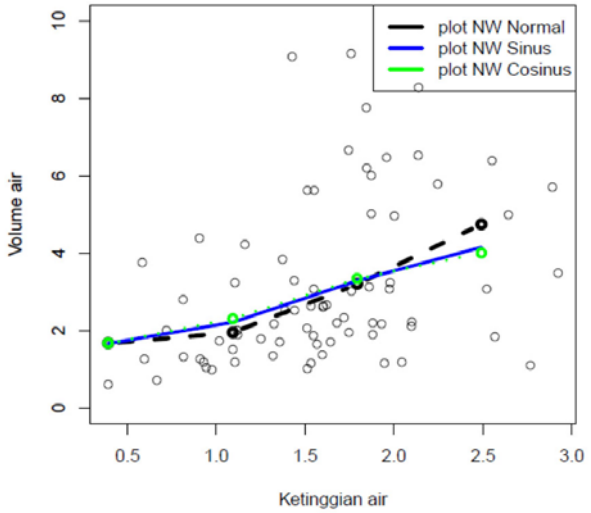


d. Grafik dengan *bandwidth* 1

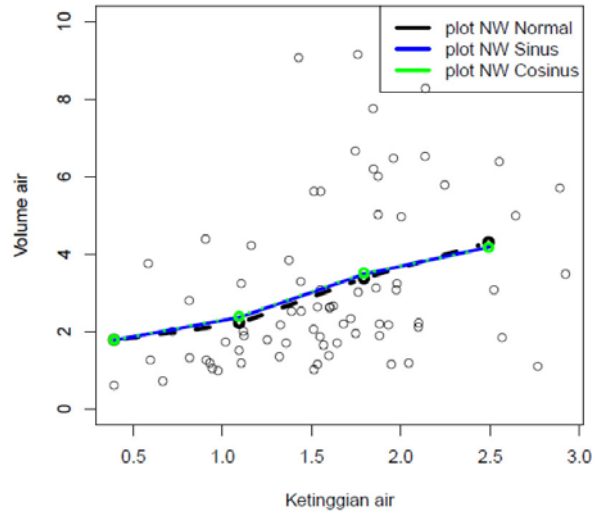
Gambar 4.1: Grafik estimasi dengan kelipatan nilai *x* sebesar 0,1



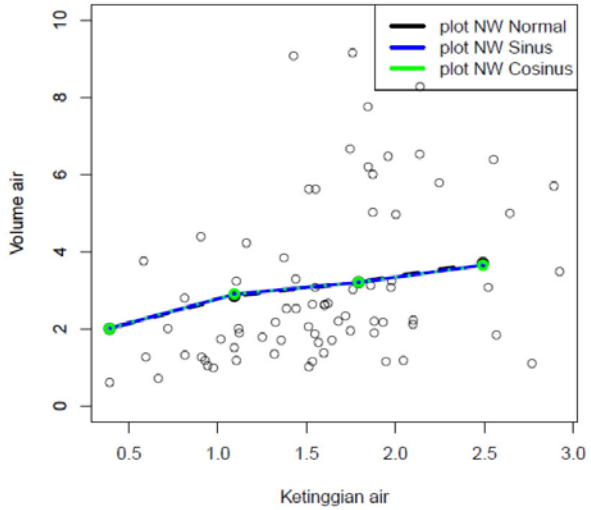
a. Grafik dengan *bandwidth* 0.13



b. Grafik dengan *bandwidth* 0,3445996



c. Grafik dengan *bandwidth* 0.5



d. Grafik dengan *bandwidth* 1

Gambar 4.2 Grafik estimasi dengan kelipatan nilai *x* sebesar 0,7

Berikut nilai-nilai MSE yang dihasilkan dengan menggunakan program R:

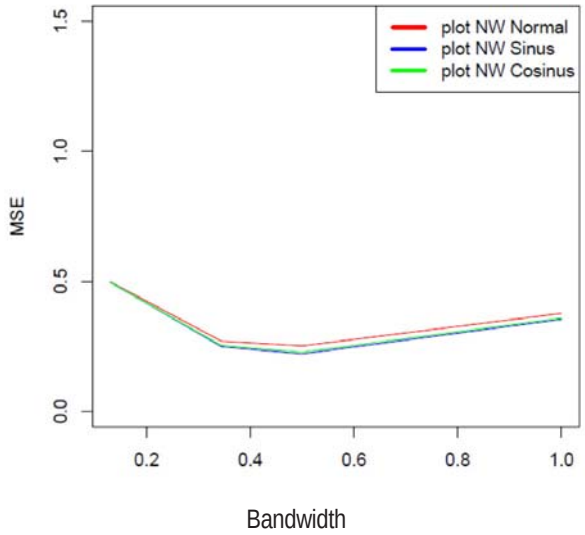
Tabel 4.1: Nilai MSE dari Masing-masing Kernel

Kelipatan Titik Estimasi	*Bandwidth*	Nilai MSE		
		Normal	Sinus	Cosinus
0,1	0,13	0,4982647	0,4963074	0,4957802
	0,3445996	0,269451	0,2487284	0,2530287
	0,5	0,2515239	0,2208997	0,2274729
	1	0,3768402	0,3539623	0,3585203
0,2	0,13	0,6237044	0,4963074	0,5001564
	0,3445996	0,2533656	0,2487284	0,2495126
	0,5	0,2422269	0,2208997	0,2255545
	1	0,3734509	0,3539623	0,3578464
0,3	0,13	1,030359	0,4963074	0,536074
	0,3445996	0,2291352	0,2487284	0,2436954
	0,5	0,227195	0,2208997	0,2223609
	1	0,3678283	0,3539623	0,3567232
0,4	0,13	1,303002	0,4963074	0,6681886
	0,3445996	0,2010336	0,2487284	0,2356519
	0,5	0,2072396	0,2208997	0,2178995
	1	0,3600139	0,3539623	0,3551501
0,5	0,13	1,352872	0,4963074	1,0337
	0,3445996	0,1751454	0,2487284	0,2255095
	0,5	0,1836912	0,2208997	0,2121826
	1	0,3500706	0,3539623	0,3531269
0,6	0,13	1,356637	0,4963074	1,965438
	0,3445996	0,1579634	0,2487284	0,2134736
	0,5	0,1584852	0,2208997	0,2052315
	1	0,3380866	0,3539623	0,3506531
0,7	0,13	1,356785	0,4963074	5,011454
	0,3445996	0,1536994	0,2487284	0,1998664
	0,5	0,1340825	0,2208997	0,1970795
	1	0,3241813	0,3539623	0,3477281

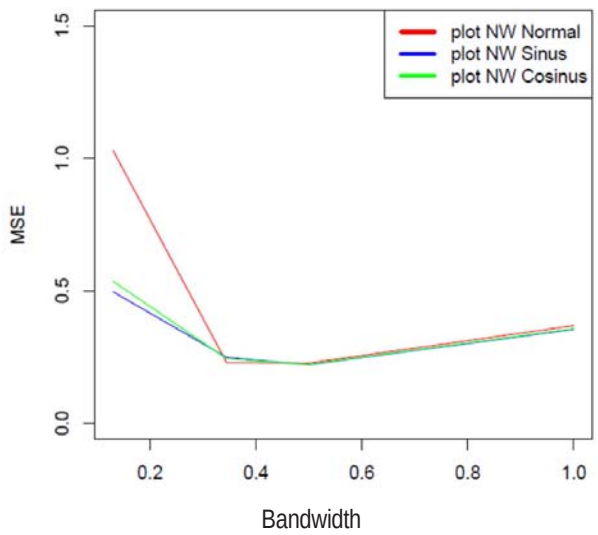
Hasil di atas menunjukkan bahwa masing-masing kernel mempunyai kekuatan yang berbeda-beda. Ketika kelipatan *x* dipilih yang kecil, dalam kasus ini kelipatan *x* kurang dari 0,4, maka estimator dengan menggunakan kernel *infinite* orde yaitu sinus akan menghasilkan nilai MSE kecil yang berarti bahwa kernel sinus akan memiliki *performance* lebih baik (memberikan hasil estimasi yang lebih baik), berapapun *bandwidth* yang dipilih, dibandingkan kernel yang lainnya. Sedangkan nilai MSE terkecil akan dihasilkan oleh estimator dengan menggunakan kernel normal ketika kelipatan *x* yang dipilih cukup besar dalam kasus ini untuk kelipatan *x* lebih dari 0,4.

Dari tabel di atas, secara keseluruhan didapatkan hasil sebagai berikut: 12 MSE bernilai kecil dihasilkan oleh estimator dengan menggunakan kernel normal, 15 MSE bernilai kecil dihasilkan oleh estimator dengan kernel sinus dan 1 MSE bernilai kecil dihasilkan oleh estimator dengan kernel cosinus. Sehingga dari hasil tabel di atas terlihat bahwa MSE terkecil paling banyak dihasilkan oleh estimator yang menggunakan kernel sinus. Berikut akan ditampilkan grafik dari MSE dari beberapa kelipatan titik *x*.

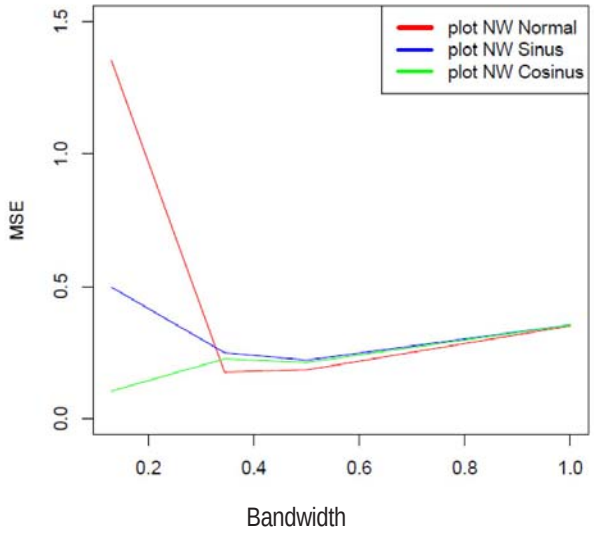
Berikut grafik MSE dari beberapa kelipatan nilai *x*:



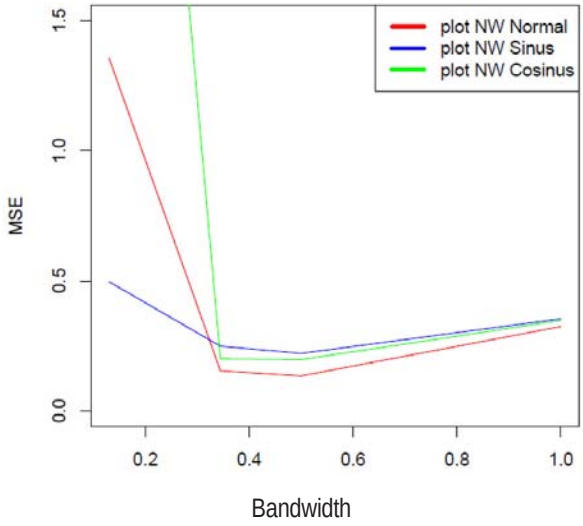
a. Grafik MSE dengan kelipatan titik x = 0.1



b. Grafik MSE dengan kelipatan titik x = 0.3



c. Grafik MSE dengan kelipatan titik x = 0.5



d. Grafik MSE dengan kelipatan titik x = 0.7

Gambar 4.3 Grafik MSE

Tinggi rendahnya grafik memperlihatkan besar dan kecilnya nilai MSE. Berikut program yang digunakan untuk menggambar grafik MSE:

```
maria=read.table("f://Data//Range_0.1.csv",header=TRUE,sep=";")
x<-maria[,4]
N<-maria[,1]
S<-maria[,2]
C<-maria[,3]
plot(x,N,type="l",col="red",xlab="x",ylab="MSE",ylim=c(0,1.5))
lines(x,S,type="l",col="blue")
lines(x,C,type="l",col="green")
legend("topright",legend=c("plot NW Normal","plot NW Sinus","plot NW Cosinus"),lwd=c(3,3,3),col=c("red", "blue", "green"))
title(sub="Grafik MSE dengan kelipatan titik x yang dipilih").
```

5. KESIMPULAN

Kesimpulan yang didapatkan mengenai *performance* estimator Nadaraya Watson menggunakan kernel berorde berhingga dan tak hingga pada kasus 74 sungai di Indonesia yang daerah pengalirannya lebih dari 1000 km^2 tahun 2010 diperoleh:

a. Masing-masing kernel mempunyai kekuatan yang berbeda-beda. Kernel dengan *infinite*

order memberikan hasil estimasi volume air yang optimal jika kelipatan titik estimasi *x* dipilih sekecil mungkin. Nilai estimasi yang dipilih adalah 0,1, 0,2 dan 0,3. Sedangkan kernel dengan *finite order* akan memberikan hasil estimasi volume air yang optimal jika kelipatan titik estimasi yang dipilih cukup besar. Nilai estimasi *x* yang digunakan 0,5, 0,6 dan 0,7.

b. Secara keseluruhan, berdasarkan nilai MSE yang dihasilkan dengan menggunakan titik estimasi *x* dan *bandwidth* tertentu, kernel sinus memiliki *performance* yang lebih optimal dibandingkan kernel cosinus dan normal.

DAFTAR PUSTAKA

Berg, Arthur. 2008. *Nonparametric Function Estimation with Infinite-Orde Kernels*. Department of Statistics, University of Florida.

Hardle, Wolfgang. 1991. *Smoothing Techniques With Implementation in S*. Springer-Verlag, New York.

Hardle, Wolfgang. 1994. *Applied Nonparametric Regression*. Berlin

McMurry, T.L and Politis, D.N. 2003. *Nonparametric Regression with Infinite Orde Flat-Top Kernel. www.math.ucsd.edu/~politis/PAPER/McMurryPolitis04.pdf*, diakses pada tanggal 13 Februari 2013.

Wand, M.P and Jones, M.C. 1995. *Kernel Smoothing*. Chapman and Hall. London.