PROSIDING SEMINAR NASIONAL STATISTIKA
UNIVERSITAS DIPONEGORO 2011
ISBN: 978-979-097-142-4

PEMULUSAN SEBARAN DATA MENGGUNAKAN PENAKSIR KERNEL NADARAYA-WATSON DAN LINIER LOKAL UNTUK KERNEL NORMAL

Sudarno[1]

[1)]Program Studi Statistika FMIPA Undip

dsghani@gmail.com

Abstrak

Sebaran data yang bebas distribusi, penaksiran fungsi regresinya menggunakan regresi nonparametrik dengan pendekatan fungsi mulus. Bentuk dari fungsi mulus tergantung dari parameternya. Untuk menaksir parameter dan fungsi mulusnya dapat menggunakan penaksir Nadaraya-Watson dan Kernel Linier Lokal. Akan dibahas simulasi metode kernel untuk menaksir nilai kurva mulus pada suatu nilai variable bebas, dengan *bandwidth* bervariasi. Adapun kernel yang dipakai untuk pendekatan yaitu kernel normal. Ingin diketahui sifat-sifat yang dihasilkan pada fungsi sinus atau cosinus dengan penambahan gangguan berdistribusi normal acak.

Kata kunci: Bandwidth, Kernel, Kernel Nadaraya-Watson, Kernel Linier lokal

1. Pendahuluan

Dalam statistika sebaran data sangat berarti. Sehingga ingin diketahui karakteristiknya. Tetapi terdapat sebaran data yang tidak diketahui fungsi sebarannya atau bebas distribusi. Untuk menentukan fungsi regresinya menggunakan regresi nonparametrik. Dalam regresi nonparametrik diperlukan suatu metode untuk mendapatkan fungsi mulus tetapi tidak seperti yang berparameter pada regresi parametrik. Dalam Green (1994) dikatakan bahwa model regresi nonparametrik berupa:

$$Y = g(\mathbf{x}) + \text{error}$$

dengan *g* merupakan fungsi mulus dan **x** adalah vector penjelas. Dalam hal ini diperlukan cara untuk mendapatkan fungsi mulus *g* tersebut.

Untuk mendapatkan fungsi mulus *g*, dapat menggunakan metode kernel, yaitu suatu fungsi pemulus hubungan antar titik dalam daerah pembicaraan dengan konsep seperti fungsi densitas peluang (Hardle, 1990) dan juga oleh Wand dan Jones [1995]. Sedangkan menurut Hardle (1991) dikatakan bahwa untuk mendapatkan fungsi mulus dapat menggunakan teknik pemulus dengan perangkat lunak **S**. Ingin diketahui dalam menentukan fungsi mulus atau taksiran regresi menggunakan perangkat lunak MATLAB. Akan dibahas kernel Nadaraya-Watson dan linier lokal untuk beberapa parameter

ISBN: 978-979-097-142-4

pemulusnya, agar didapat karakteristik dan nilai taksiran fungsi pada variabel bebas tertentu.

2. Metode Kernel

Kernel merupakan sembarang fungsi pemulusan K yang mempunyai sifat:

- $K(x) \geq 0$ (1)
- $\int K(x)\,dx = 1$ (2)
- $\int x\,K(x)\,dx = 0$ (3)
- $\sigma_K^2 = \int K(x)\,dx > 0$ (4)

Beberapa kernel yang biasa dipakai dalam pembahasan adalah:

Tabel 1. Macam-macam Kernel

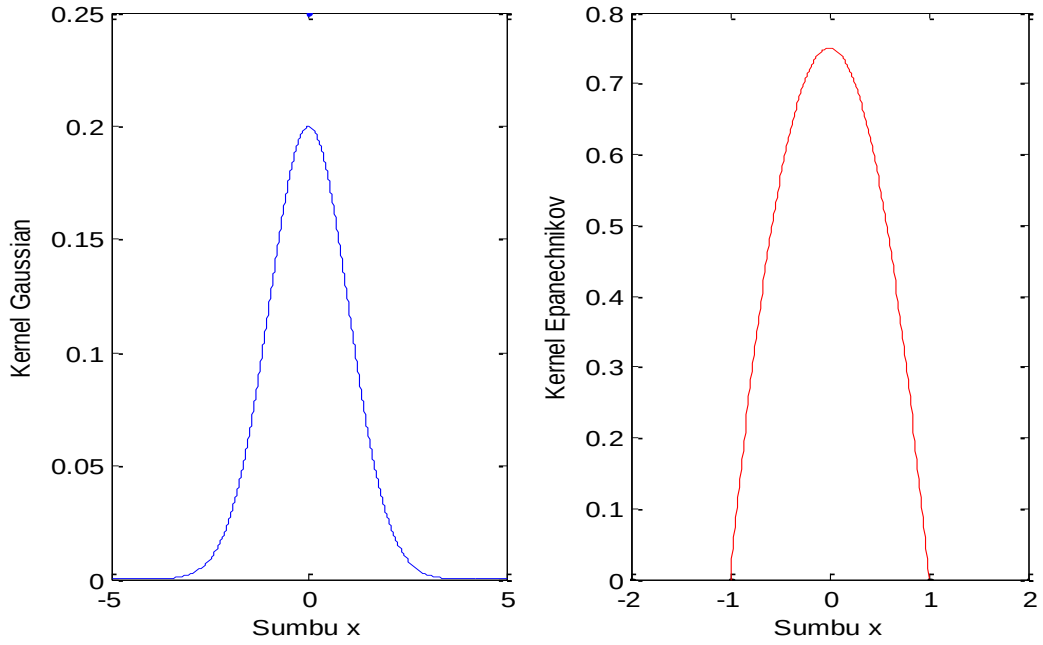
Kernel	Persamaan
Boxcar	$K(x) = \frac{1}{2} I(x)$
Gaussian	$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Epanechikov	$K(x) = \frac{3}{4}(1-x^2) I(x)$
Tricube	$K(x) = \frac{70}{81}(1-\lvert x\rvert^3)^3 I(x)$

dengan

$$I(x) = \begin{cases} 1 \text{ jika } |x| \leq 1 \\ 0 \text{ jika } |x| > 1 \end{cases}$$

Berikut ini diberikan contoh gambar kernel Gaussian dan Epanechikov:

ISBN: 978-979-097-142-4



Gambar 1. Kernel Gaussian dan Epanechikov

Kernel dipergunakan untuk pengambilan rataan lokal. Misal terdapat pasangan data $(X_1, Y_1),\ldots,(X_n, Y_n)$ dan diinginkan mendapatkan rataan dari semua Y_i yang sesuai dengan X_i dalam suatu jarak *h* dari suatu titik *x*. Rataan lokal ini adalah sama dengan

$$\sum_{i=1}^{n} Y_i \, l_i(x) \tag{5}$$

dimana

$$l_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \tag{6}$$

dengan K merupakan kernel. Jadi kernel merupakan rataan terboboti secara lokal. Kernel akan dipergunakan untuk penaksiran dari sebaran data yang bebas distribusi.

Metode pemulusan menggunakan kernel diberikan oleh Wand dan Jones [1995]. Akan ditampilkan klas metode pemulusan berdasar pada penaksir kernel, yang menentukan data dalam cara lokal. Cara ini disebut penaksir kernel polinomial lokal. Pertama didefinisikan penaksir secara umum dan selanjutnya menampilkan dua kasus khusus: **penaksir kernel Nadaraya-Watson** dan **penaksir kernel linier lokal**.

ISBN: 978-979-097-142-4

Dengan penaksir kernel polinomial lokal, dapat diperoleh suatu taksiran $\hat{y}_0$ pada titik x_0 dengan menentukan derajat polinomial *d* menggunakan kuadrat terkecil terboboti. Diinginkan memboboti titik-titik berdasarkan pada jaraknya dari x_0. Titik-titik yang lebih dekat terhadap pusat data, sebaiknya mempunyai bobot yang lebih besar. Sedangkan titik-titik yang lebih jauh jaraknya diberi bobot yang lebih kecil. Untuk mewujudkan ini, dipergunakan bobot yang dipergunakan untuk tinggi fungsi kernel yang berpusat pada x_0.

Kernel mempunyai suatu *bandwidth* atau parameter pemulus yang dinyatakan dengan *h*. Parameter pemulus ini akan mengontrol derajat pengaruh titik-titik pada penentuan lokasi. Jika *h* kecil, maka kurva akan bergelombang. Karena taksiran akan bergantung pada bobot dari titik-titik terhadap kedekatan dengan x_0. Dalam kasus ini, model akan dicoba menentukan nilai lokal dengan memilih nilai *h* dari kecil ke besar dengan melihat perubahan grafiknya. Dengan *h* cukup besar, akan dapat menentukan garis pada seluruh kumpulan data.

Sekarang diberikan gambaran untuk penaksir kernel polinomial lokal. Misal *d* menyatakan derajat polinomial yang ditentukan pada suatu titik *x*. Diperoleh penaksir $\hat{y} = \hat{f}(x)$ dengan menentukan polinomial

$$\beta_0 + \beta_1(X_i - x) + \cdots + \beta_d(X_i - x)^d \tag{7}$$

dengan menggunakan titik-titik (X_i, Y_i) dan menerapkan prosedur kuadrat terkecil terboboti. Bobotnya diberikan oleh fungsi kernel

$$K_h(X_i - x) = \frac{1}{h} K\left(\frac{X_i - x}{h}\right). \tag{8}$$

Nilai taksiran pada titik *x* adalah $\hat{\beta}_0$, dimana $\hat{\beta}_i$ diperoleh dengan meminimalkan

$$\sum_{i=1}^{n} K_h(X_i - x)(Y_i - \beta_0 - \beta_1(X_i - x) - \cdots - \beta_d(X_i - x)^d)^2 . \tag{9}$$

Karena titik-titik yang dipergunakan untuk menaksir model tersebut adalah semuanya berpusat di *x*, maka taksiran pada *x* diperoleh dengan mengambil argumen dalam model yaitu sama dengan nol. Dengan demikian, parameter yang tersisa adalah hanya suku konstan $\hat{\beta}_0$.

Jika kernel dipusatkan pada variable acak X_i. Sesuai dengan notasi yang dibuat oleh Wand dan Jones [1995] yang menunjukkan secara eksplisit bahwa pusat kernel pada titik *x*, ingin didapat nilai taksiran dari fungsi. Prosedur kuadrat terkecil terboboti dengan

ISBN: 978-979-097-142-4

menggunakan notasi matriks, sesuai dengan teori kuadrat terkecil terboboti baku (Draper dan Smith, 1981], didapat:

$$\hat{\beta} = \left(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{Y} \tag{10}$$

dengan **Y** merupakan vektor respon berukuran n x 1,

$$\mathbf{X}_x = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^d \\ \vdots & \vdots & & \cdots \ \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^d \end{bmatrix},$$

dan $\mathbf{W}_x$ adalah matriks berukuran n x n dengan bobotnya sepanjang diagonal. Bobot ini diberikan dengan

$$w_{ii}(x) = K_h(X_i - x) \tag{11}$$

Bobot di atas dapat bernilai nol, tergantung dari kernel yang dipergunakan. Penaksir $\hat{y} = \hat{f}(x)$ merupakan koefisien intersep β_0 dari taksiran lokal. Nilainya diperoleh dari

$$\hat{f}(x) = \mathbf{e}_1^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{Y} \tag{12}$$

dimana $\mathbf{e}_1^T$ adalah suatu vektor berdimensi (d + 1) x 1 dengan satu dalam tempat pertama dan nol pada tempat yang lainnya.

3. Penaksir Kernel Nadaraya-Watson dan Linier Lokal

3.1 Penaksir Kernel Nadaraya-Watson

Suatu pernyataan secara eksplisit ada bilamana d = 0 dan d = 1. Bila d bernilai nol, dicari fungsi konstan secara lokal yang diberikan titik x. Penaksir ini dikembangkan secara terpisah oleh Nadaraya [1964] dan Watson [1964]. Jika digabung mendapatkan Penaksir Nadaraya-Watson seperti yang diberikan berikut ini:

$$\hat{f}_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)\, Y_i}{\sum_{i=1}^{n} K_h(X_i - x)} \tag{13}$$

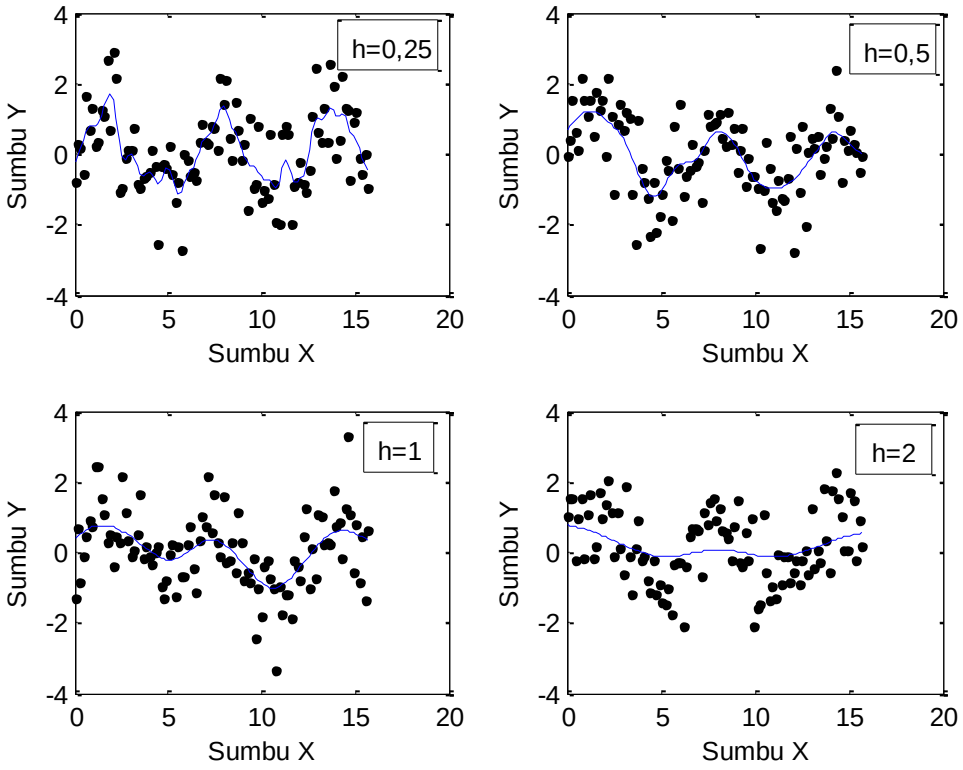
Penaksir ini untuk kasus rancangan acak. Bilamana titik rancangan adalah tetap, maka X_i, diganti dengan x_i, tetapi jika tidak demikian pernyatannya adalah sama [Wand dan Jones, 1995].

Berikut akan dibahas simulasi dari sebaran data fungsi binometri (sinus dan cosines) dengan penambahan gangguan sebaran acak berdistribusi normal. Taksiran fungsi

ISBN: 978-979-097-142-4

regresinya didekati menggunakan penaksir Nadaraya-Watson dengan kernel normal. Parameter pemulusnya mengambil nilai h = 0,25; 0,50; 1; 2.

- Untuk sebaran fungsi $y = \sin(x) + 0.9\,\text{randn}(\text{size}(x))$ dan taksirannya ditampilkan pada grafik berikut ini:



Gambar 2. Grafik Fungsi Sinus dengan Penaksir Nadaraya-Watson.

Terlihat bahwa makin kecil nilai *h*, fungsi taksirannya makin bergelombang. Sebaliknya makin besar nilai *h*, fungsi taksirannya makin menuju linier.

Adapun hasil taksirannya pada suatu nilai *x* dan *h* tertentu ditabelkan di bawah ini:

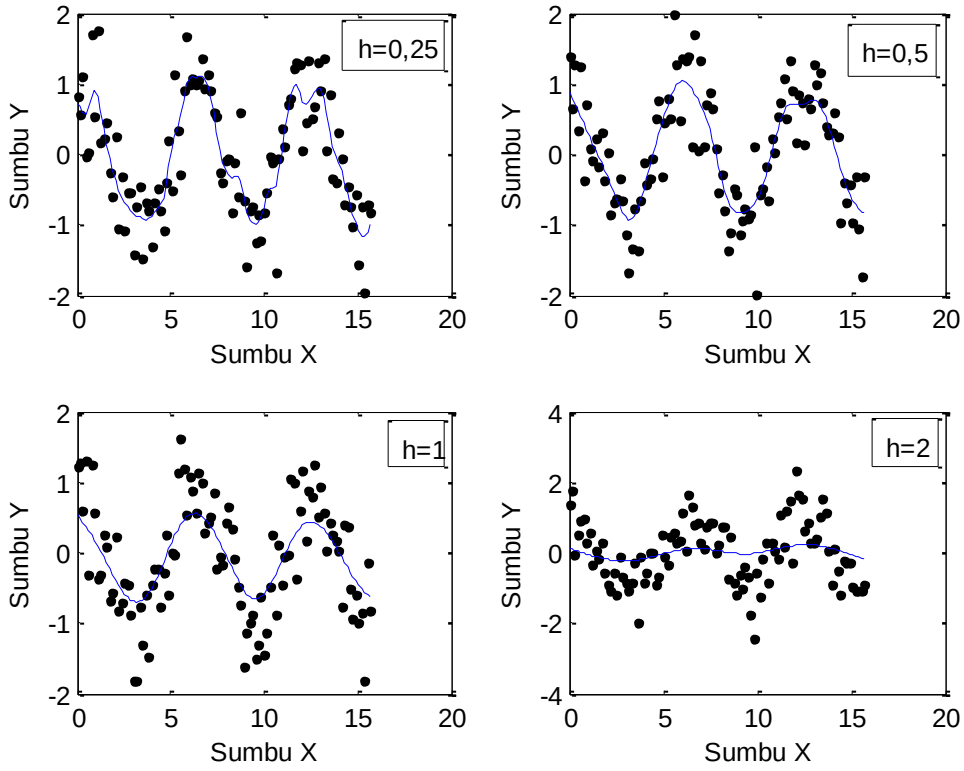
Tabel 2. Hasil taksiran nilai Y

X	**Y**			
	0,25	0,5	1	2
1	1,0246	0,0787	0,3922	0,7713
5	1,0810	0,5387	0,5114	0,6884

ISBN: 978-979-097-142-4

10	-0,2783	0,7843	0,5963	0,5448
15	0,4650	0,7497	0,5399	0,3594

- Untuk sebaran fungsi $y = \cos(x) + 0.5\,\text{randn}(\text{size}(x))$ dan taksirannya ditampilkan pada grafik berikut ini:



Gambar 2. Grafik fungsi cosinus dengan penaksir Nadaraya-Watson.

Terlihat bahwa makin kecil nilai *h*, fungsi taksirannya makin bergelombang. Sebaliknya makin besar nilai *h*, fungsi taksirannya makin menuju linier.

Adapun hasil taksirannya pada suatu nilai *x* dan *h* tertentu ditabelkan di bawah ini:

Tabel 3. Hasil Taksiran Nilai Y

X	**Y**			
	0,25	0,5	1	2
1	0,2640	1,5897	0,2278	0,6654

ISBN: 978-979-097-142-4

5	1,1128	1,1784	0,1025	0,4918
10	0,3807	0,2324	-0,1911	0,2771
15	0,0604	-0,5707	-0,5755	0,0921

3.2 Penaksir Kernel Linier Lokal

Jika ingin menentukan garis lurus pada suatu titik *x*, maka dapat menggunakan penaksir linier lokal. Hal ini sesuai dengan kasus dimana d = 1. Sehingga taksiran diperoleh sebagai penyelesaian $\hat{\beta}_0$ dan β_1 dengan meminimalkan:

$$\sum_{i=1}^{n} K_h(X_i - x)(Y_i - \beta_0 - \beta_1(X_i - x))^2 \qquad (14)$$

Akan diberikan rumus secara eksplisit untuk penaksir kernel linier lokal di bawah ini:

$$\hat{f}_{LL}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\{\hat{s}_2(x) - \hat{s}_1(x)(X_i - x)\}K_h(X_i - x)Y_i}{\hat{s}_2(x)\hat{s}_0(x) - \hat{s}_1(x)^2}, \qquad (15)$$

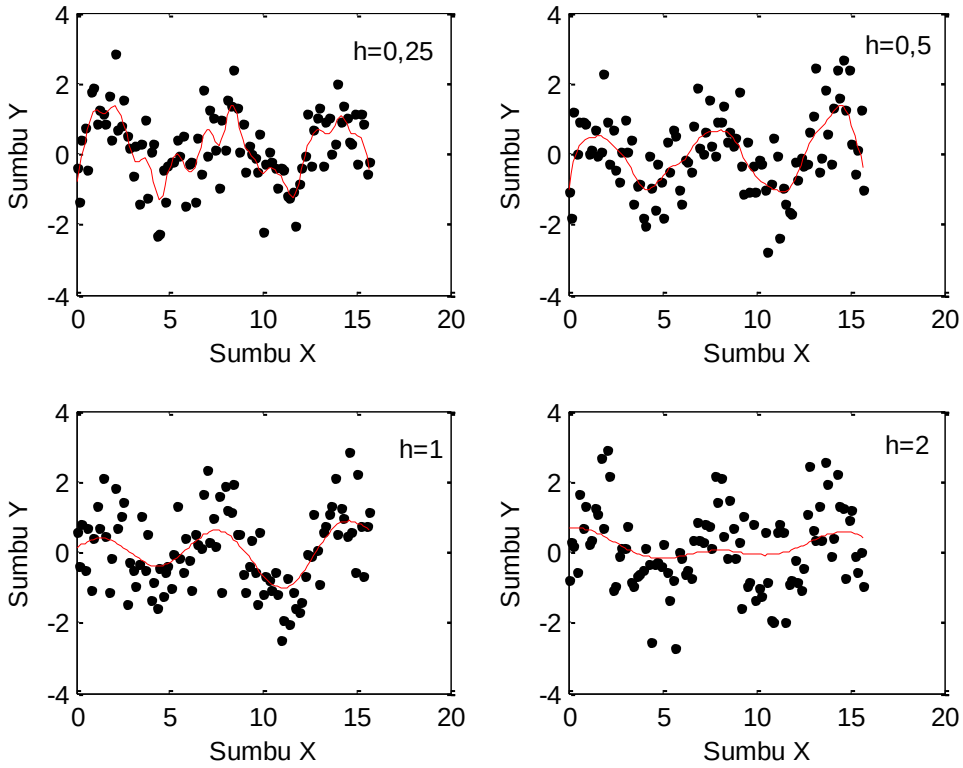
Dengan

$$\hat{s}_r(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i - x)^r K_h(X_i - x) \qquad (16)$$

Selanjutnya kasus rancangan tetap diperoleh dengan mengganti variable acak X_i dengan titik tetap x_i.

- Untuk sebaran fungsi $y = \sin(x) + 0.9\,\text{randn}(\text{size}(x))$ dan taksirannya ditampilkan pada grafik berikut ini:

ISBN: 978-979-097-142-4



Gambar 3. Grafik Fungsi Sinus dengan Penaksir Kernel Linier Lokal.

Terlihat bahwa makin kecil nilai *h*, fungsi taksirannya makin bergelombang. Sebaliknya makin besar nilai *h*, fungsi taksirannya makin mulus.

Adapun hasil taksirannya pada suatu nilai *x* dan *h* tertentu ditabelkan di bawah ini:

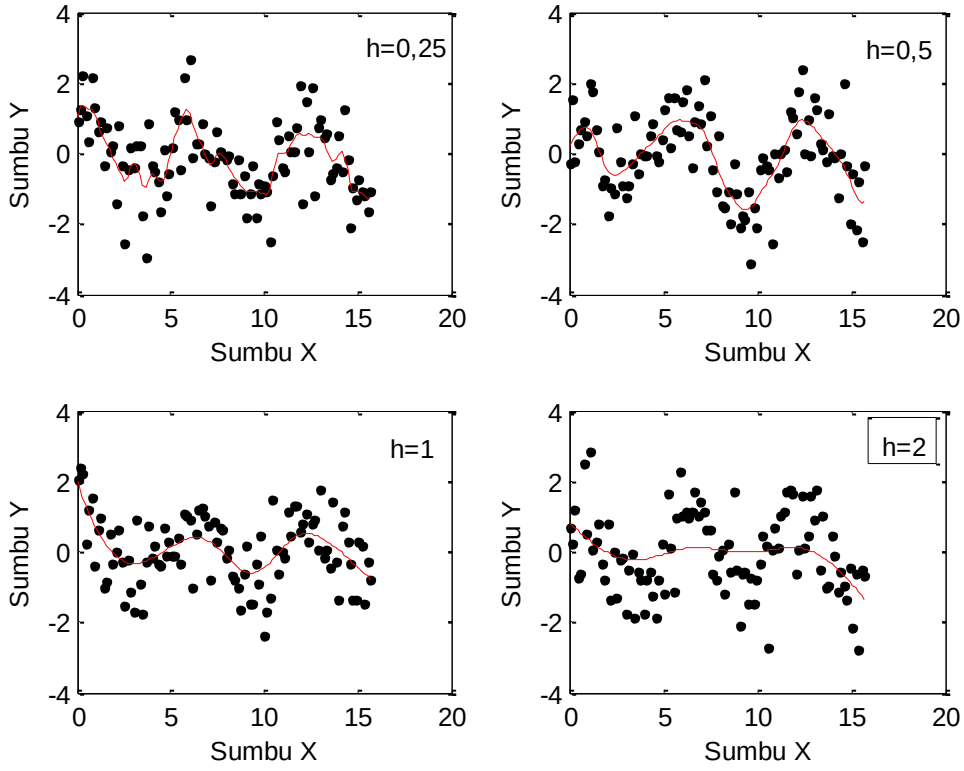
Tabel 4. Hasil Taksiran Nilai Y

X	Y			
	0,25	0,5	1	2
1	-0,7700	-1,0313	0,1112	0,6645
5	0,6567	0,2299	0,2845	0,6831
10	1,146	0,4727	0,3840	0,5668

ISBN: 978-979-097-142-4

	6			
15	1,242 6	0,3434	0,2415	0,3562

- Untuk sebaran fungsi $y = \cos(x) + 0.5\,\text{randn}(\text{size}(x))$ dan taksirannya ditampilkan pada grafik berikut ini:



Gambar 4. Grafik Fungsi Cosinus dengan Penaksir Kernel Linier Lokal.

Terlihat bahwa makin kecil nilai *h*, fungsi taksirannya makin bergelombang. Sebaliknya makin besar nilai *h*, fungsi taksirannya makin mulus.

Adapun hasil taksirannya pada suatu nilai *x* dan *h* tertentu ditabelkan di bawah ini:

Tabel 5. Hasil Taksiran Nilai Y

X	**Y**			
	0,25	0,5	1	2
1	1,0443	0,2142	2,0128	0,8143

ISBN: 978-979-097-142-4

5	1,2433	0,6536	1,1202	0,5276
10	0,4185	0,3432	0,3114	0,1899
15	-0,4354	-0,5781	-0,1750	-0,0708

4. Kesimpulan

Sebaran data yang bebas distribusi, penaksiran parameternya dapat menggunakan penaksir Nadaraya-Watson dan Kernel Linier Lokal. Untuk sebaran data yang berupa fungsi sinus atau cosines dengan penambahan gangguan berdistribusi normal acak didapat hasil bahwa:

- Makin kecil nilai *bandwidth*, makin bergelombang grafiknya.
- Makin besar nilai *bandwidth*, grafiknya makin linier atau mulus.
- Untuk menentukan besarnya nilai taksiran pada suatu titik dapat dengan memilih besarnya nilai *bandwidth* yang dikehendaki.

Daftar Pustaka

Draper, N.R. and Smith, H., (1981) *Applied Regression Analysis*. 2nd Edition, John Wiley & Sons, New York.

Green, P.J. and Silverman, B.W., (1994) *Nonparamertic Regression and Generalized Linear Models*, Chapman & Hall, London.

Hardle, W., (1990) *Applied nonparametrc regression*, Cambride University Press, New York.

Hardle, W., (1991) *Smoothing Techniques With Implementation in **S***, Springer-Verlag New York Inc., New York.

Martinez, W.L. and Martinez, A.R., (2002) *Computational Statistics Handbook with MATLAB*, Chapman & Hall, Florida.

Moore, H., (2007) *MATLAB for Engineers*, Pearson Prentice Hall, Inc., New Jersey.

Nadaraya, E.A., (1964) *Theory of Probability and Its Application*, **10**: pp. 186–190.

Scott, D.W., (1992) *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons, New York.

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

Wasserman, L., (2006) *All of Nonparametric Statistics*, Springer Science+Business Media, Inc., New York.

Watson, G.S., (1964) *Smooth Regression Analysis*, **26**: pp. 101 – 116.