

Pemodelan Hubungan Kepadatan Penduduk dan Indeks Kualitas Udara (IKU) di Indonesia Menggunakan Regresi Kuantil *Smoothing Splines*

You Ari Faeni

Badan Pusat Statistik Kabupaten OKU Timur

you@bps.go.id

Abstract. The level of pollution in Indonesia is the top 10 worst in the world. Judging from the Air Quality Index (IKU), there are 9 provinces that have KPI values below the expected target. This paper aims to perform IKU modeling using the population density variable as a predictor variable. Modeling using linear regression in the parametric method cannot be used because the model residuals are not normally distributed, so a nonparametric smoothing splines approach is carried out. However, the presence of outliers in the smoothing splines residual model causes the residuals of the model to be too large so that it affects the prediction accuracy, so the smoothing splines quantile regression is used in the IKU modeling. Apart from using the median (quantile $\tau = 0.5$), the quantiles of 0.2 were also used; 0.4; 0.6; and 0.8 to generate models at various quantiles. The results of the analysis using the R package quantreg Software prove that the smoothing splines (median) quantile regression model is more robust against the presence of outliers seen from the lower RMSE value than the smoothing splines regression model (mean). In addition, it is concluded that there are 5 provinces that are below the quantile 0.2, which means that the IKU level is very low or there is very high pollution based on the level of population density. Likewise, there are 3 provinces with KPI values above the quantile of 0.8, which means they have very high IKU levels or areas with low levels of pollution.

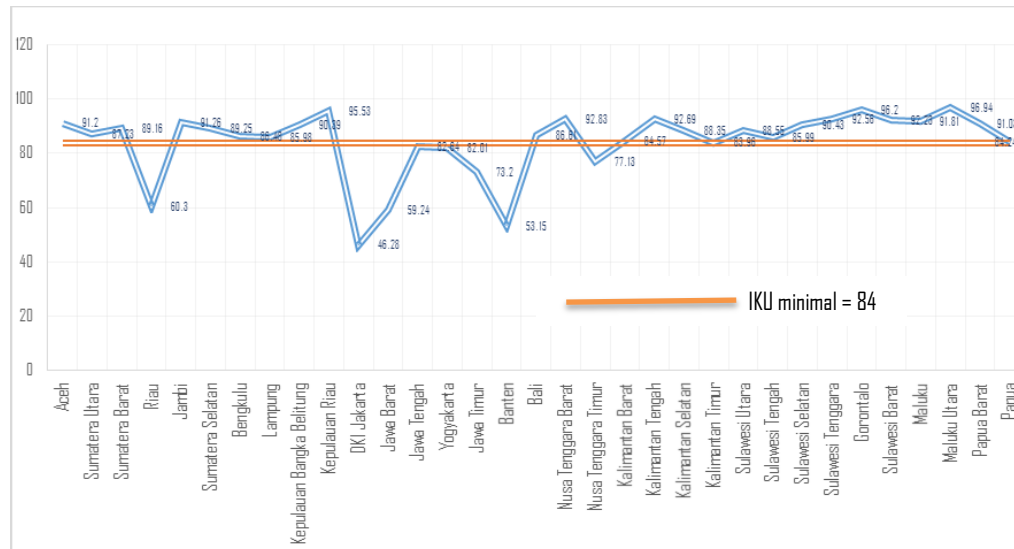
Keywords: *IKU, Population Density, Smoothing Splines, Quantile Regression*

I. PENDAHULUAN

Data Badan Kesehatan Dunia (WHO) menunjukkan bahwa 9 dari 10 orang di dunia menghirup udara dengan tingkat polutan yang tinggi. Polusi udara luar ruangan menjadi penyebab utama kematian dini di dunia ke-4, dan kerugian ini diperkirakan membebani ekonomi global dengan biaya tahunan yang tidak sedikit yaitu sebesar 225 miliar USD (Green peace[3]). Indonesia sendiri merupakan salah satu negara di dunia dengan tingkat polusi yang tinggi. Kota Jakarta masuk peringkat sepuluh besar ibu kota negara dengan kualitas udara terburuk di dunia. Konsentrasi rata-rata tahunan PM2.5 di Jakarta tahun 2018 sangat buruk, dimana pada Jakarta Selatan mencapai 42.2 $\mu\text{g}/\text{m}^3$ dan Jakarta Pusat mencapai 37.5 $\mu\text{g}/\text{m}^3$. Konsentrasi ini mencapai empat kali lipat di atas batas aman tahunan menurut standar WHO, yaitu 10 $\mu\text{g}/\text{m}^3$. Lebih lanjut lagi, Guru Besar Tetap Bidang Ilmu Kesehatan Lingkungan dari Fakultas Kesehatan Masyarakat UI (FKM UI), Budi Haryanto, menyatakan bahwa 50% dari angka kesakitan di Indonesia saat ini terkait dengan polusi udara. Jumlah penyakit yang terkait dengan kasus pencemaran udara telah diprediksi lebih tinggi dan lebih parah.

Kualitas udara di Indonesia dapat dilihat dari beberapa indikator antara lain Indeks Standar Pencemar Udara (ISPU), Konsentrasi Partikel Terlarut, dan Indeks Kualitas Udara (IKU). IKU bagian dari Indeks Kualitas Lingkungan Hidup (IKLH) yang sudah di hitung oleh Kementerian Lingkungan Hidup dan Kehutanan (KLHK) di lokasi-lokasi yang mewakili daerah permukiman, industri, dan padat lalu lintas kendaraan bermotor dan

parameter yang diukur adalah SO_2 dan NO_2 (BPS [1]). Nilai IKU berkisar antara 0 sampai dengan 100. Nilai ideal adalah 100, yang menggambarkan kualitas terbaik. Sementara nilai 0 menggambarkan kualitas terburuk. Seperti diketahui bahwa RPJMN 2015-2019 telah menargetkan IKLH pada tahun 2019 sebesar 66,5-68,5 (kategori cukup). Penetapan ini didasarkan pada asumsi pencapaian IKU minimal sebesar 84.



Gambar 1. Indeks Kualitas Udara Menurut Provinsi di Indonesia, Tahun 2014

Berdasarkan target tersebut, pada tahun 2014 terdapat 9 provinsi yang memiliki nilai IKU di bawah target yang diharapkan (**Gambar 1**).

Permasalahan Polusi Udara ini merupakan masalah yang penting, maka penulis tertarik untuk melakukan Pemodelan Indeks Kualitas Udara (IKU). Penelitian tentang polusi banyak dikaitkan dengan permasalahan kependudukan, diantaranya Kousa [6] yang memodelkan kepadatan penduduk terhadap polusi udara di daerah perkotaan menggunakan analisis *spatio temporal*. Penelitian oleh Cheng [2] menyatakan bahwa peningkatan populasi di China menyebabkan peningkatan polusi PM_{10} . Rodriguez [8] juga melakukan penelitian serupa pada kota-kota di Eropa. Oleh karena itu, dalam makalah ini digunakan variabel kepadatan penduduk sebagai variabel prediktor polusi udara (yang digambarkan melalui indikator IKU)

II. METODE PENELITIAN

Data Indeks Kualitas Udara dalam penelitian ini berasal dari data sekunder untuk 34 Provinsi di Indonesia. Namun karena keterbatasan data yang ada maka digunakan periode penelitian adalah Tahun 2014 dengan menghilangkan data Kalimantan Utara karena adanya *missing values*, sehingga total observasi sebanyak 33 provinsi di Indonesia. Variabel-variabel yang digunakan dirangkum dalam **Tabel 1**.

Tabel 1. Rangkuman Variabel Penelitian

Variabel	Nama Variabel	Definisi Variabel	Sumber Data
y (Variabel Respon)	Indeks Kualitas Udara (IKU)	Nilai yang menunjukkan mutu atau tingkat kebaikan udara menurut sifat-sifat unsur pembentuknya. IKU dihitung berdasarkan emisi dari dua polutan udara yaitu karbon monoksida (CO) dan nitrogen oksida (NO _x). Kedua jenis polutan ini dijadikan sebagai komponen IKU karena pengaruh keduanya yang sangat signifikan terhadap kehidupan manusia.	BPS 2017
x (Variabel Prediktor)	Kepadatan Penduduk (jiwa/km ²)	Banyaknya jumlah penduduk untuk setiap kilometer persegi luas wilayah.	BPS 2017

Data diolah menggunakan *Software R package quantreg* oleh Koenker [5].

Pada tahap persiapan awal data, dilakukan transformasi data Kepadatan Penduduk (x) menggunakan transformasi logaritma natural untuk memperoleh data dengan variansi yang tidak terlalu besar. Kemudian data di-plot untuk mengetahui bentuk hubungan antara variabel respon dan variabel y. Setelah itu data dianalisis dengan tahapan sebagai berikut

Regresi Linier

Analisis regresi menjelaskan hubungan variabel prediktor dan variabel respon dengan menggambarkan nilai variabel respon yang dipengaruhi oleh setiap nilai variabel prediktor. Model regresi yang fokus pada nilai rata-rata dari distribusi respon disebut regresi *mean*, atau *ordinary least square* (OLS). Secara umum model regresi linier adalah :

$$y = x\beta + \varepsilon$$

dengan β adalah parameter yang ditaksir dan ε adalah residual model.

Metode OLS dilakukan dengan meminimumkan jumlah kuadrat residual atau $L_2 - loss function$. Penaksir yang diperoleh dengan metode ini merupakan penaksir parameter yang BLUE (*Best Linear Unbiased Estimator*). Namun regresi linier memiliki beberapa kelemahan diantaranya jika digunakan pada data dengan sebaran yang menceng (*skewness*), maka regresi *mean* kurang tepat digunakan sebab tidak dapat mewakili bagian lain dari distribusi respon. Selain itu, regresi *mean* sensitif terhadap *outlier*. Keberadaan *outlier* dalam data menyebabkan *residual* tidak berdistribusi normal. Akibatnya, nilai *residual* yang akan menjadi besar dan diikuti varians *residual* yang besar dan tidak homogen dan estimator yang dihasilkan pada data yang mengandung *outlier* tidak akan bersifat *Best Linear Unbiased Estimator* (BLUE).

Regresi Smoothing Splines

Regresi *smoothing splines* merupakan pencocokan kurva dengan pemulusan untuk suatu set pengamatan dengan menggunakan fungsi *splines*. Estimasi $f(xi)$ dengan OLS akan menghasilkan fungsi yang kasar. Alasannya, $f(xi)$ dalam persamaan model nonparametrik merupakan fungsi yang tidak diketahui, sehingga diperlukan fungsi yang lebih halus agar jumlah kuadrat *residual* menjadi kecil. Salah satu cara untuk mendapatkan fungsi yang cukup halus adalah dengan memberikan pemulus (*penalty*). estimasi fungsi *splines* $f(xi)$ adalah meminimisasi :

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_0^1 [f''(x)]^2 dx \quad (1)$$

dimana $\sum (y_i - f(x_i))^2$ merupakan jumlah kuadrat *residual*, serta ukuran kemulusan $\lambda \int_0^1 [f''(x)]^2 dx$ yang disebut *roughness pinalty* dan λ sebagai parameter pemulus.

Peranan parameter pemulus (λ) sangat penting dalam mengontrol atau mengendalikan ketepatan model dan kemulusan kurva. Nilai λ adalah 0 sampai ∞ . Jika $\lambda \rightarrow 0$ maka ketepatan model akan tinggi namun kurva kurang mulus. Kurva akan cenderung mengikuti pola data, sehingga terlihat lebih kasar. Jarak antara data dan dugaan sangat dekat, sehingga (1) akan sama dengan jumlah kuadrat *residual*. Sebaliknya, jika $\lambda \rightarrow \infty$ maka ketepatan model akan rendah namun kurva sangat mulus. Kelengkungan kurva sangat kecil, bahkan mendekati fungsi linier, sehingga (1) akan sama dengan *roughness pinalty*. Kurva akan cenderung menjauhi pola data sehingga jarak antara data dan dugaan sangat jauh.

Dalam makalah ini parameter pemulus (λ) dipilih dengan kriteria *Cross Validation* (CV) dimana

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \frac{y_i - f(x_i)}{1 - h_{ii}}$$

Regresi Kuantil *Smoothing Splines*

Regresi kuantil adalah teknik regresi yang menjelaskan hubungan antara variabel respon dan variabel prediktor tidak hanya pada ukuran pemusatan (bersyarat median) variabel respon tetapi pada berbagai kuantil. Regresi kuantil sangat baik digunakan pada sebaran data yang berdistribusi asimetris, padat pada ujung sebaran data, maupun terdapat *outlier*, karena estimator yang dihasilkan akan lebih efisien (Andriyana dalam [7]).

Metode regresi kuantil *smoothing splines* merupakan pemodelan regresi yang mengestimasi kurva tidak hanya pada ukuran pemusatan (bersyarat median) variabel respon, namun pada berbagai kuantil dengan menggunakan teknik *smoothing splines*. Dalam Koenker et.al [5] disebutkan solusi regresi kuantil *smoothing splines* diperoleh dengan meminimumkan:

$$\sum_{i=1}^n \rho_{\tau}(y_i - f(x_i)) + \lambda \int_0^1 |f''(x)| dx \quad (2)$$

$$\text{Dengan } 0 = x_0 < x_1 < \dots x_n < x_{n+1} = 1.$$

Fungsi obyektif kuantil regresi kuantil *smoothing splines* atau L_1 mempunyai sifat tidak dapat diturunkan/diminimumkan (*not differentiable*) pada titik nol, sehingga tidak diperoleh solusi yang eksplisit. Oleh sebab itu, solusi kuantil dapat diperoleh dengan mengubah fungsi L_1 ke dalam bentuk sistem persamaan linier yang dapat dipecahkan dengan *LP-problem*.

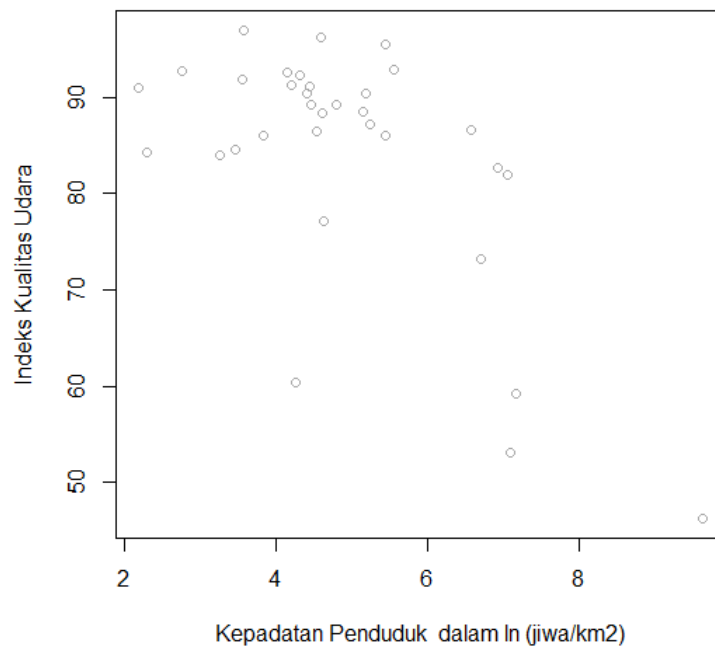
Pada tahap akhir makalah, dilakukan perbandingan metode Regresi *Smoothing Splines* menggunakan pendekatan *mean* dengan Regresi *Smoothing Splines* dengan pendekatan median menggunakan nilai *Root Mean Square Error* (RMSE).

III. HASIL DAN PEMBAHASAN

Spesifikasi Model

Hubungan Kepadatan Penduduk dan Indeks Kualitas Udara (IKU) dalam suatu pemodelan dapat diketahui dengan menggunakan teknik analisis regresi. Langkah awal

yang dilakukan adalah menelaah spesifikasi model secara deskriptif terhadap pola sebaran datanya. Apabila pola sebaran data tidak dapat ditentukan fungsi *slope* dan *intercept*-nya, maka dapat digunakan pendekatan nonparametrik. Bentuk pola sebaran data kedua variabel tersebut dapat dilihat dengan *scatter plot* sebagai berikut:



Gambar 2. Pola Hubungan Kepadatan Penduduk dan IKU menurut Provinsi di Indonesia, Tahun 2014

Berdasarkan **Gambar 2.** diatas, terlihat bahwa sebaran data cenderung mengumpul di ujung, sehingga pendekatan nonparametrik disarankan. Namun untuk memastikan, pendekatan parametrik dapat dilakukan dengan syarat *residual* harus memenuhi asumsi normalitas. Untuk mengetahui apakah *residual* berdistribusi normal atau tidak, dapat dilakukan dengan uji Shapiro Wilk dengan hipotesis:

H_0 : residual berdistribusi normal

H_1 : residual tidak berdistribusi normal

Tolak H_0 atau residual tidak berdistribusi normal jika $p - value$ kurang dari α (0,05). Hasil uji normalitas pada **Lampiran 2.** menggunakan Uji Shapiro Wilk mendapatkan $p - value = 0.01374$ sehingga dapat ditarik kesimpulan bahwa residual tidak berdistribusi normal.

Tidak terpenuhinya asumsi kenormalan, maka pembentukan model regresi linier dalam metode parametrik tidak dapat dilakukan. Selanjutnya pemodelan Kepadatan Penduduk dan IKU menggunakan pendekatan nonparametrik *Smoothing Splines*.

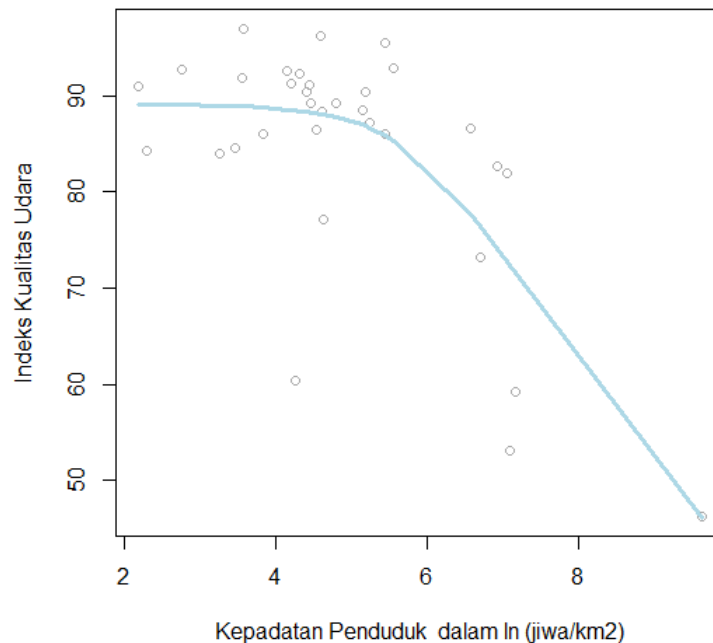
Pemodelan Regresi *Smoothing Splines*

Pemodelan regresi *Smoothing Splines* dilakukan dengan menggunakan fungsi `smooth.spline` dengan kriteria *Cross Validation* (CV). Nilai parameter λ optimum diperoleh secara default menggunakan kriteria CV. Dari pemodelan yang dilakukan diperoleh parameter-parameter sebagai berikut :

Tabel 2. Parameter dalam Regresi *Smoothing Spline*

Parameter	Nilai
Spar	0.9944586
Lambda	0.008100737
Degrees of Freedom	3.721798

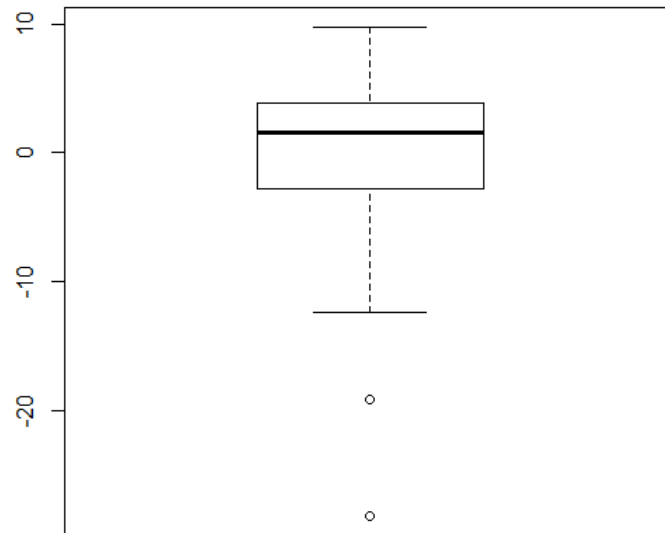
Parameter optimum tersebut digunakan untuk menentukan estimator sehingga didapatkan model regresi *Smoothing Splines* seperti gambar di bawah ini



Gambar 2. Kurva Regresi *Smoothing Splines* (Mean)

Pemodelan Regresi Kuantil *Smoothing Splines*

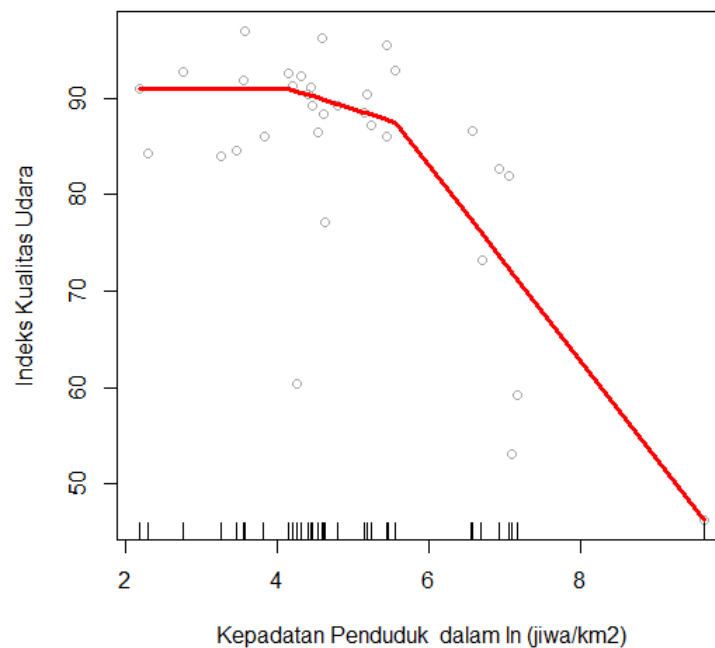
Jika sebaran data mengandung pencilan, pendekatan regresi *mean* tidak disarankan dan sebaiknya menggunakan pendekatan regresi kuantil. Dengan menggunakan *boxplot* dari *residual* dengan menggunakan regresi *smoothing splines*, diperoleh hasil sebagai berikut:



Gambar 3. Boxplot Residual Pemodelan Regresi *Smoothing Splines*

Berdasarkan **Gambar 3.** dapat diketahui bahwa hubungan kedua variabel menunjukkan adanya data yang mengandung *outlier*. Berdasarkan hal ini, maka penggunaan regresi kuantil sebagai alternatif *mean* lebih optimal digunakan karena dapat mengakomodir *outlier* dan menangkap informasi sebaran data melalui berbagai garis kuantilnya.

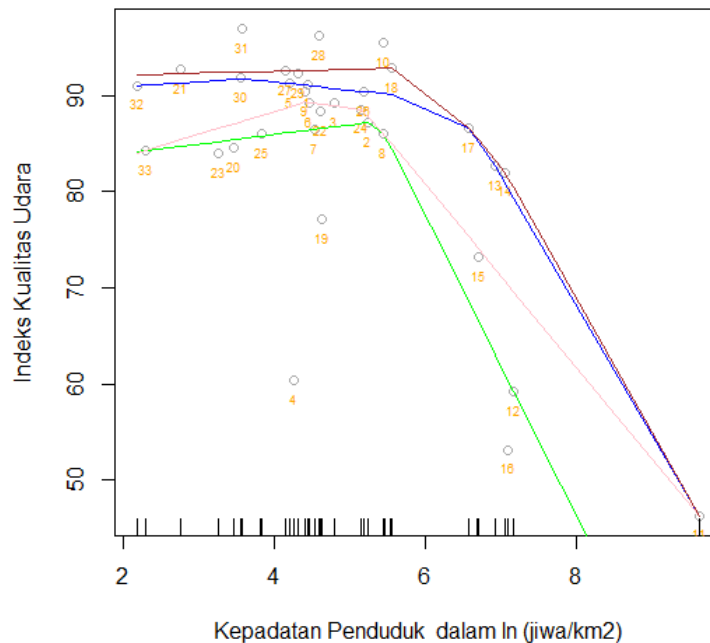
Model Regresi Kuantil *Smoothing Splines* (Median, $\tau = 0,5$) diperoleh sebagai berikut:



Gambar 4. Kurva Regresi Kuantil *Smoothing Splines* (Median, $\tau = 0,5$)

Selain regresi median ($\tau = 0,5$), pada regresi kuantil dapat diperoleh berbagai nilai kuantil sehingga diperoleh garis regresi sebanyak kuantil yang diinginkan (Koenker [4]).

Pada makalah ini, dalam memodelkan Kepadatan Penduduk dan Indeks Kualitas Udara (IKU) digunakan nilai kuantil 0.2, 0.4, 0.6, dan 0.8, dengan kategorisasi tingkat IKU sangat rendah, rendah, sedang, tinggi, dan sangat tinggi. Hasil pemodelan tersebut ditampilkan dalam **Gambar 5**.

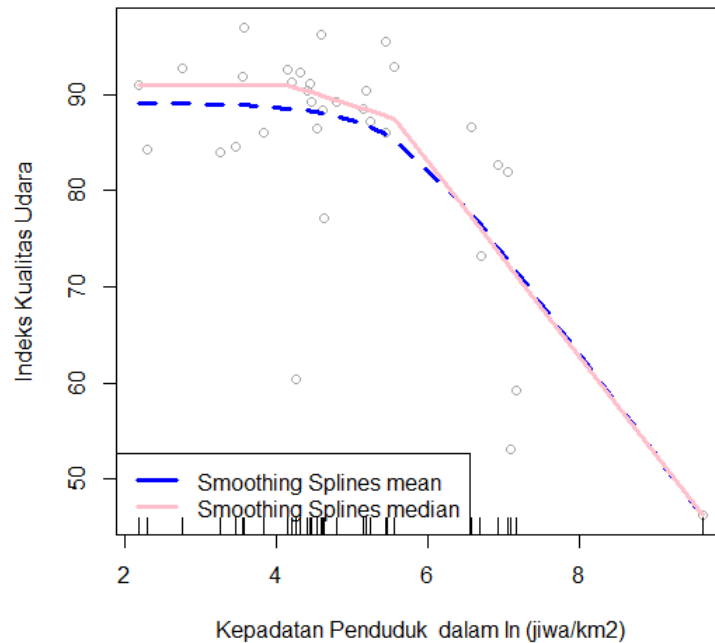


Gambar 5. Kurva Regresi Kuantil *Smoothing Splines*

Selanjutnya dapat dilakukan estimasi nilai IKU berdasarkan kepadatan penduduk. Dalam hal ini, penggunaan pendekatan nonparametrik akan membatasi estimasi hanya pada data *driven* di dalam penelitian. Berdasarkan **Gambar 5**, di atas, terlihat bahwa terdapat 5 provinsi yang berada di bawah kuantil 0,2, yang berarti bahwa tingkat IKU sangat rendah atau terjadi pencemaran yang sangat tinggi berdasarkan tingkat kepadatan penduduknya. Begitupun terdapat 3 provinsi dengan nilai IKU di atas kuantil 0,8 yang berarti memiliki tingkat IKU sangat tinggi atau daerah dengan tingkat pencemaran yang rendah.

Perbandingan Model

Pada sebaran data yang menyebar asimetris dan adanya data *outlier*, penggunaan pendekatan kuantil lebih efisien dibandingkan pendekatan *mean*. Untuk mengetahuinya, model regresi kuantil *smoothing splines* (median, $\tau=0.5$) akan dibandingkan dengan regresi *smoothing splines* (*mean*). Median dapat dibandingkan dengan *mean* sebab merupakan salah satu ukuran pemusatan selain *mean*. Perbandingan kedua model dapat dilihat dalam gambar berikut:



Gambar 6. Perbandingan Kurva Regresi *Smoothing Splines* Pendekatan *Mean* dan *Median*

Sedangkan dalam pengujian performa kedua model di dalam penelitian ini menggunakan penghitungan *Root Mean Square Error* (RMSE) dengan hasil sebagai berikut:

Tabel 4. Perbandingan Nilai RMSE

Model Regresi	RMSE
Regresi <i>Smoothing Splines</i> (<i>Mean</i>)	8.043714
Regresi <i>Smoothing Splines</i> (<i>Median</i>)	5.962126

Berdasarkan Tabel 4.8 dapat diketahui bahwa estimasi model regresi *smoothing splines* (*mean*) mempunyai RMSE sebesar 8,04 lebih tinggi dibandingkan estimasi model regresi kuantil *smoothing splines* (*median*, $\tau=0.5$) dengan RMSE sebesar 5,96. Nilai RMSE pada model regresi *mean* akan menghasilkan nilai yang lebih tinggi, sebab pada data yang menyebar asimetris dan mengandung *outlier*, penggunaan OLS sangat sensitif sehingga menghasilkan nilai *residual* yang lebih besar.

IV. KESIMPULAN

Pemodelan regresi kuantil *smoothing splines* memiliki beberapa kelebihan dibandingkan pendekatan *mean* dalam regresi. Salah satu kelebihannya adalah pada regresi kuantil dapat diperoleh berbagai nilai kuantil sehingga diperoleh garis regresi sebanyak kuantil yang diinginkan. Berbagai garis regresi yang dihasilkan dapat digunakan untuk analisis tingkat kewajaran suatu titik observasi berdasarkan nilai variabel prediktornya. Dalam makalah ini, diperoleh kesimpulan bahwa terdapat 5 provinsi yang berada di bawah kuantil 0,2, yang berarti bahwa tingkat IKU sangat rendah atau terjadi pencemaran yang sangat tinggi berdasarkan tingkat kepadatan penduduknya. Begitupun terdapat 3 provinsi

dengan nilai IKU di atas kuantil 0,8 yang berarti memiliki tingkat IKU sangat tinggi atau daerah dengan tingkat pencemaran yang rendah.

Kelebihan lain regresi kuantil *smoothing splines* adalah model ini lebih robust terhadap adanya outlier. Hal ini terlihat dari nilai RMSE model regresi *smoothing splines* (mean) lebih tinggi dibandingkan estimasi model regresi kuantil *smoothing splines* (median, $\tau = 0.5$) dengan RMSE yaitu sebesar 8,04 dan 5,96.

DAFTAR PUSTAKA

- [1] Badan Pusat Statistik. 2017. Statistik Lingkungan Hidup Indonesia 2017. Jakarta: Badan Pusat Statistik.
- [2] Cheng, Zhen, et.al. 2013. Characteristics and health impacts of particulate matter pollution in China (2001–2011). *Atmospheric Environment*, 65, Pages 186-194
- [3] Greenpeace. Data Terkini Kualitas Udara Kota-kota di Seluruh Dunia. <https://www.greenpeace.org/indonesia/publikasi/2217/data-terkini-kualitas-udara-kota-kota-di-seluruh-dunia/>
- [4] Koenker, Roger, et.al. 1994. Quantile Smoothing Splines. *Biometrika* 81,4, pages 673-680.
- [5] Koenker, Roger. Quantile Regression In R: A Vignette. CRAN: <http://cran.r-project.org>.
- [6] Kousa, Anu, et.al. 2002. A model for evaluating the population exposure to ambient air pollution in an urban area. *Atmospheric Environment*, 13, Pages 2109-2119.
- [7] Mulyani, Sri. 2017. Pemodelan Hubungan Indeks Pembangunan Manusia Dan Persentase Penduduk Miskin Menggunakan Regresi Kuantil *Smoothing Splines*. Tesis: Universitas Padjadjaran.
- [8] Rodriguez, Miguel Cardenas. 2016. Air pollution and urban structure linkages: Evidence from European cities. *Renewable and Sustainable Energy Reviews*, 53, Pages 1-9.

LAMPIRAN 1. Data yang Digunakan

No	Provinsi	Indeks Kualitas udara- IKU	Kepadatan Penduduk (Jiwa/km ²)
1	Aceh	91.20	85
2	Sumatera Utara	87.23	189
3	Sumatera Barat	89.16	122
4	Riau	60.30	71
5	Jambi	91.26	67
6	Sumatera Selatan	89.25	87
7	Bengkulu	86.48	93
8	Lampung	85.98	232
9	Bangka Belitung	90.39	82
10	Kepulauan Riau	95.53	234
11	DKI Jakarta	46.28	15.173
12	Jawa Barat	59.24	1.301
13	Jawa Tengah	82.64	1.022
14	DI Yogyakarta	82.01	1.161
15	Jawa Timur	73.20	808
16	Banten	53.15	1.211
17	Bali	86.61	710
18	Nusa Tenggara Barat	92.83	257
19	Nusa Tenggara Timur	77.13	103
20	Kalimantan Barat	84.57	32
21	Kalimantan Tengah	92.69	16
22	Kalimantan Selatan	88.35	101
23	Kalimantan Timur	83.96	26
24	Kalimantan Utara	-	8
25	Sulawesi Utara	88.55	172
26	Sulawesi Tengah	85.99	46
27	Sulawesi Selatan	90.43	180
28	Sulawesi Tenggara	92.56	64
29	Gorontalo	96.2	99
30	Sulawesi Barat	92.23	75
31	Maluku	91.81	35
32	Maluku Utara	96.94	36
33	Papua Barat	91.03	9
34	Papua	84.24	10