Prosiding SPMIPA. pp. 185-191. 2006 *ISBN : 979.704.427.0*

PENENTUAN FAKTOR UTAMA PENYEBAB GANGGUAN LISTRIK DENGAN METODE VALIDASI-SILANG (STUDI KASUS DI KOTA SEMARANG)

Tarno
Program Studi Statistika FMIPA UNDIP Semarang
Jl. Prof. Soedarto, Kampus UNDIP Tembalang

Abstrak: Dalam tulisan ini dibahas tentang penentuan faktor utama yang berpengaruh secara signfikan terhadap pemadaman listrik di Semarang dengan menggunakan metode validasi silang. Pada awalnya faktor-faktor yang diduga berpengaruh terhadap pemadaman listrik di kota Semarang adalah: kerusakan jaringan transmisi, kerusakan trafo dan kerusakan fuse (sekering). Dengan melibatkan 3 faktor tersebut dibentuk model regresi yang menyatakan hubungan antara kerusakan jaringan, trafo dan fuse(sekering) terhadap pemadaman listrik di kota Semarang. Prosedur pemilihan model regresi terbaiknya dilakukan dengan menentukan estimasi sesatan prediksi atas semua model yang mungkin yaitu ada sebanyak 2^p-1 = 7 model, dengan p: banyaknya prediktor. Model yang terpilih adalah model yang memiliki rata-rata sesatan prediksi terkecil dan melibatkan variabel prediktor sesedikit mungkin. Prosedur pemilihan model terbaiknya dilakukan dengan menggunakan metode validasi-silang lepas-d (1<d<n). Prosedur pemilihan modelnya konsisten untuk $\frac{d}{n} \to 1$ dan $(n-d) \to \infty$ dengan n: ukuran sampel. Berdasarkan simulasi yang dilakukan dengan software R, diperoleh model regresi terbaik dengan melibatkan 2 variabel yaitu kerusakan jaringan dan kerusakan sekering. Dengan demikian faktor utama penyebab pemadaman listrik di kota Semarang adalah kerusakan jaringan dan kerusakan sekering.

Kata Kunci: pemilihan variabel, sesatan prediksi, validasi-silang

PENDAHULUAN

PT. PLN (Persero) merupakan Badan Usaha yang memberikan jasa pelayanan listrik kepada masyarakat. Keberhasilan PT. PLN dalam menyediakan jasa pelayanan listrik sangat tergantung pada alat-alat yang digunakan sebagai sarana penyampaian jasa listrik tersebut. Gangguan-gangguan pada peralatan sangat memungkinkan terjadinya pemadaman listrik di suatu wilayah tertentu. Dengan adanya pemadaman listrik tersebut berarti PT PLN dapat mengakibatkan kerugian pada masyarakat pengguna listrik dan juga bagi PT PLN sendiri. Kerusakan peralatan yang sering dapat menimbulkan pemadaman listrik antara lain: kerusakan trafo, kerusakan fuse/sekering, kerusakan jaringan transmisi. Bila sering terjadi pemadaman listrik, maka jumlah pemakaian listrik oleh konsumen menjadi sedikit, sehingga PT. PLN akan mengalami kerugian. Jika pemadaman listrik yang disebabkan oleh gangguan alat sering terjadi, maka PT. PLN perlu mengambil langkah yang tepat untuk melakukan perbaikan terhadap faktor-faktor penyebab pemadaman tersebut.

Kerusakan peralatan yang dapat menyebabkan gangguan atau pemadaman listrik seringkali terjadi di kota Semarang. Secara geografis kota Semarang terletak di daerah perbukitan, dimana wilayahnya dapat dibedakan menjadi dua bagian yaitu Semarang atas dan Semarang bawah. Terkait dengan kondisi geografis tersebut Semarang atas sering terjadi gangguan cuaca seperti: angin kencang dan petir, sedangkan di Semarang bawah sering terjadi banjir. Faktor-faktor alam tersebut dapat menyebabkan kerusakan pada jaringan transmisi PLN, sedangkan kerusakan trafo, fuse/sekering seringkali disebabkan oleh pemakaian listrik yang berlebihan.

Berdasarkan argumen-argumen diatas, maka dalam tulisan ini dilakukan pengkajian terhadap data gangguan listrik di kota Semarang yang diduga disebabkan oleh kerusakan/gangguan peralatan antara lain: kerusakan trafo, sekering dan jaringan transmisi. Diduga jumlah kerusakan peralatan tersebut berpengaruh secara signifikan terhadap jumlah pemadaman listrik dan mempunyai hubungan linier, sehingga hubungan fungsional antara jumlah kerusakan peralatan dengan jumlah gangguan listrik selama periode tertentu dapat dinyatakan dalam suatu model matematika.

Adapun model matematika yang sesuai dengan kenyataan tersebut adalah model regresi linier:

$$\mathbf{y}_i = \mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, i = 1,2,...,n \quad (1)$$

dengan $\mathbf{y}_i$ adalah respon ke-i menyatakan jumlah pemadaman pada kurun waktu ke-i , $\mathbf{x}_i$: 3-vektor variabel prediktor (jumlah kerusakan/gangguan trafo, fuse/sekering dan jaringan transmisi) yang berkaitan dengan $\mathbf{y}_i$, β : 3-vektor parameter yang tidak diketahui dan ε_i : sesatan random berdistribusi normal dengan mean 0 dan variansi konstan..

Untuk mengestimasi parameter dalam model regresi tersebut biasanya digunakan metode kuadrat terkecil. Jika estimasi parameter telah diperoleh berarti telah diperoleh estimasi model untuk respon **y** yang tergantung pada prediktor **x**, yang dapat digunakan untuk melakukan prediksi untuk nilai **y** yang akan datang berdasarkan prediktor **x**. Beberapa komponen dari **x** mungkin tidak menghasilkan prediksi yang akurat karena tidak berpengaruh secara signifikan terhadap respon **y**, oleh karena itu perlu dilakukan pemilihan model terbaik (dalam hal ini sama dengan pemilihan variabel prediktor), yang memiliki kemampuan prediksi yang paling akurat. Menurut Shao (1993), pemilihan variabel dalam model regresi linier ini dapat dilakukan dengan beberapa metode antara lain : Akaike Information Criterion (AIC) , Cp (Mallows), Bayesian Information Criterion (BIC) dan metode Validasi-Silang. Menurut [1], kriteria AIC secara eksak atau pendekatan merupakan estimator tak bias untuk model dengan semua parameternya tak nol, tetapi jika digunakan untuk memilih model dengan komponen parameternya ada yang sama dengan nol) kriteria ini kadang-kadang tidak konsisten (bias). Sedangkan untuk kriteria BIC secara asimptotis tidak konsisten untuk data pengamatan berukuran besar, dan lebih baik apabila diterapkan pada model runtun waktu. Dan menurut [2], [3], [4] pemilihan model linier dengan metode validasi-silang memiliki sifat konsisten untuk ukuran sampel besar.

Berdasarkan latar belakang tersebut, perlu dikaji lebih dalam tentang faktor utama penyebab gangguan listrik sekaligus menentukan model terbaik yang menyatakan hubungan antara jumlah kerusakan peralatan (trafo, fuse dan jaringan) terhadap pemadaman listrik di kota Semarang. Dalam tulisan ini digunakan metode pemilihan model regresi terbaik berdasarkan resampling data pengamatan yaitu metode Validasi–Silang (CV). Metode Validasi-silang merupakan metode pembangkitan data pengamatan berbasis komputer untuk mendapatkan data sampel berukuran besar, sehingga asumsi-asumsi yang disyaratkan dalam persamaan regresi akan terpenuhi terutama asumsi normalitas. Disamping itu sampel yang dikumpulkan di lapangan tidak perlu berukuran besar, sehingga peneliti dapat melakukan efisiensi waktu dan biaya untuk pengumpulan data di lapangan. Untuk mendapatkan sampel berukuran besar, cukup dilakukan pembangkitan data dengan simulasi komputer di laboratorium.

PEMILIHAN VARIABEL DAN SESATAN PREDIKSI

Prediksi nilai respon untuk masa yang akan datang menggunakan variabel prediktor **x**, secara aktual mungkin tidak tergantung pada semua komponen **x**, artinya penggunaan semua komponen dari **x** belum tentu menghasilkan prediksi yang akurat. Dibawah model (1),

$$y_i = x_i^{'}\beta + \varepsilon_i \, , i = 1,2,\ldots,n \qquad (1)$$

dengan **y** : variabel respon, **x** : p-vektor prediktor , β : p-vektor parameter yang tak diketahui dan ε : sesatan random dengan mean 0 dan variansi σ^2. Karena beberapa komponen dari β mungkin sama dengan 0 maka model yang menghasilkan prediksi yang lebih akurat (lebih kompak) adalah model yang berbentuk :

$$y_i = x_{i,\alpha}^{'}\beta_\alpha + \varepsilon_i \, , i = 1,2,\ldots,n \qquad (2)$$

dengan $\alpha \subset \{1,2,\ldots,p\}$.

Jika β_α dan $\mathrm{x}_{\mathrm{i},\alpha}$ sebagai subvektor yang memuat komponen-komponen dari β dan x_i berada dalam α, maka terdapat (2^p-1) model berbeda yang mungkin yang berbentuk (2), masing-masing terkait dengan suatu himpunan bagian α dan dinotasikan dengan $\hat{\alpha}$. Dimensi (ukuran) dari $\hat{\alpha}$ adalah banyaknya prediktor dalam $\hat{\alpha}$. Misalkan ***A*** menyatakan semua himpunan bagian dari $\{1,2,\ldots,\mathrm{p}\}$, jika diketahui masing-masing komponen dari β adalah 0 atau tidak, maka model-model $\hat{\alpha}$ dapat diklasifikasikan menjadi dua kategori :

- Kategori I (*incorrect model*) : Minimal satu komponen dari β yang tidak nol tidak berada dalam β_α.
- Kategori II (*correct model*) : β_α memuat semua komponen dari β yang tidak nol.

Memilih model dari kategori I berarti menghilangkan minimal satu prediktor yang penting, sedangkan memilih model dari kategori II berarti mengeliminasi semua variabel yang tak terkait dengan variabel respon. Dengan demikian model optimalnya adalah model (2) dengan α_0 sedemikian hingga β_{α_0} memuat semua komponen dari β yang semuanya tidak nol, yaitu model dalam kategori II dengan dimensi terkecil.

Model optimal tersebut tidak diketahui karena β tidak diketahui, sehingga perlu dilakukan pemilihan model. Yaitu memilih model dari model (2) berdasarkan data (y_1,x_1), (y_2,x_2), …, (y_n,x_n) yang memenuhi (1). Jika diasumsikan bahwa $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ independen dan berdistribusi identik dengan mean 0 dan variansi σ^2, maka dibawah model α, dengan Metode Kuadrat Terkecil diperoleh :

$$\hat{\beta}_\alpha = (X'_\alpha X_\alpha)^{-1} X'_\alpha y$$

dengan $\mathbf{y}=(y_1,y_2,\ldots,y_n)'$ dan $X_\alpha = (x_{1\alpha}, x_{2\alpha}, ..., x_{n\alpha})$.

Anggap bahwa $\mathbf{y_f}$ adalah nilai variabel respon untuk yang akan datang untuk suatu nilai variabel prediktor $\mathbf{x_f}$, maka :

$$\hat{y}_{f\alpha} = x'_{f\alpha}\hat{\beta}_\alpha .$$

Hal ini berakibat bahwa mean dari sesatan prediksi kuadrat $mse(x_f, \alpha)$ adalah :

$$mse(x_f,\alpha) = E(y_f - \hat{y}_{f\alpha})^2 = \sigma^2 + \sigma^2 x'_{f\alpha}(X'_\alpha X_\alpha)^{-1} x_{f\alpha} + \Delta(x_f,\alpha),$$

$$dengan\ \Delta(x_f,\alpha) = [x'_f\beta - x'_{f\alpha}(X'_\alpha X_\alpha)^{-1} X\beta]^2$$

Jika α dalam kategori II maka $X\beta = X_\alpha\beta_\alpha$, $x'_f\beta = x'_{f\alpha}\beta_\alpha$ dan $\Delta(x_f,\alpha) = 0$. Sehingga model optimalnya adalah model α dengan ukuran terkecil. Dengan demikian jika $mse(x_f,\alpha)$ diketahui, maka model optimal dapat dipilih dengan meminimalkan $mse(x_f,\alpha)$ atas semua $\alpha \in \boldsymbol{A}$. Model optimal dapat juga ditentukan dengan meminimalkan rata-rata dari sesatan prediksi kuadrat $mse(x_f,\alpha)$ atas $X=\{x_1,x_2,\ldots,x_n\}$:

$$\overline{mse}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} mse(x_i,\alpha) = \sigma^2 + \frac{\sigma^2 p}{n} + \Delta(\alpha),$$

dengan $\Delta(\alpha) = \frac{1}{n}\beta' X'(I - H_\alpha)X\beta$,

$H = X_\alpha(X'_\alpha X_\alpha)^{-1}X'_\alpha$ dan I matriks identitas pxp.

Namun, $mse(x_f,\alpha)$ dan $\overline{mse}(\alpha)$ kedua-duanya tidak diketahui. Sehingga mengestimasi $\overline{mse}(\alpha)$ lebih mudah dari pada mengestimasi $mse(x_f,\alpha)$ dengan menggunakan $\hat{\overline{mse}}(\alpha)$, kemudian memilih model dengan meminimalkan $\hat{\overline{mse}}(\alpha)$ atas $\alpha \in \boldsymbol{A}$.

Untuk mendapatkan model terbaik, lebih lanjut dilakukan pemilihan model dengan metode validasi silang.

PEMILIHAN VARIABEL DENGAN VALIDASI-SILANG

Berdasarkan ide jackknife-1, Shao,J. (1995) mengusulkan metode pemilihan variabel yang dikenal dengan metode validasi silang (*cross-validation*). Jika $\hat{\beta}_{\alpha,i}$ adalah estimator kuadrat terkecil dari β dibawah model α setelah menghapus elemen (y_i,x_i'), maka

$$\hat{\beta}_{\alpha,i} = \left(\sum_{j\neq i} x_{j\alpha}x'_{j\alpha}\right)^{-1}\sum_{j\neq i} x_{j\alpha}y_j,\ i = 1,2,...,n$$

Karena y_i dan $\hat{\beta}_{\alpha,i}$ independen, $\overline{mse}(\alpha)$ dapat diestimasi dengan:

$$\hat{\overline{mse}}_{CV}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x'_i\hat{\beta}_{\alpha,i})^2. \qquad (3)$$

Metode validasi silang lepas-1 (CV-1) memilih model dengan meminimalkan $\hat{\overline{mse}}_{CV}(\alpha)$ atas $\alpha \in \boldsymbol{A}$.

Lebih lanjut jika terdapat ketakkonsistenan dari validasi silang lepas-1, maka metode validasi silang lepas-d diharapkan dapat memperbaiki kelemahan dari validasi silang lepas-1 tersebut. Dalam metode validasi silang lepas-d , matriks (**y,X**) yang berordo n x (1+p) dikelompokkan ke dalam dua kelompok submatriks yaitu d x

(1+p) matriks ($\mathbf{y_s,X_s}$) yang memuat baris-baris dari ($\mathbf{y,X}$) dengan $\mathbf{s} \subset \{1,2,...,n\}$ berukuran d, dan (n-d) x (1+p) matriks (y_{s^c}, X_{s^c}) yang disebut data konstruksi. Sesatan ekses prediksi dinyatakan sebagai:

$$\left\| y_s - X_{\alpha,s}\hat{\beta}_{\alpha,s^c} \right\|$$

dengan $X_{\alpha,s}$ adalah matriks $(d \times p_\alpha)$ memuat kolom-kolom dari X_s yang diindekkan sama dengan bilangan bulat α. Sehingga ($\mathbf{y_s,X_s}$) disebut data validasi. Jika ***S*** adalah suatu koleksi himpunan bagian dari {1,2,…,n} berukuran d < n, maka metode validasi silang lepas-d (CV-d) memilih model dengan meminimalkan :

$$\hat{\overline{mse}}_{CV-d}(\alpha) = \frac{1}{B}\sum_{s \in S} \left\| y_s - X_{\alpha,s}\hat{\beta}_{\alpha,s^c} \right\|^2 \qquad (4)$$

dengan $\alpha \in \boldsymbol{A}$, B banyaknya himpunan bagian dalam ***S***. Himpunan ***S*** dapat diperoleh dengan mengambil sebuah sampel random sederhana dari koleksi semua himpunan bagian yang mungkin dari {1,2,…,n} berukuran d .

KONSISTENSI METODE VALIDASI-SILANG

Validasi Silang Lepas-1

Suatu syarat yang harus dipenuhi untuk suatu prosedur pemilihan variabel yang diberikan adalah tentang konsistensinya, yaitu:

$$\lim_{n\to\infty} P\{\hat{\alpha} = \alpha_0\} = 1$$

dengan $\hat{\alpha}$ adalah model terpilih dengan menggunakan prosedur pemilihan yang diberikan.

Teorema 1 [2], [3]. Diasumsikan bahwa ε_i independen dan berdistribusi identik(i.i.d.) dan $\max_{i\le n} h_{i\alpha} \to 0$ untuk semua $\alpha \in \boldsymbol{A}$, dengan $h_{i\alpha} = x_{i\alpha}'(X_\alpha' X_\alpha)^{-1} x_{i\alpha}$.

(i) Pandang suatu validasi silang-1. Apabila α dalam kategori I (*an incorrect model*),

$$\hat{\overline{mse}}_{CV}(\alpha) = \overline{mse}(\alpha) + o_p(n^{-1});$$

sedangkan apabila α dalam kategori II (*an correct model*),

$$\hat{\overline{mse}}_{CV}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{2\sigma^2 p_\alpha}{n} - \frac{\varepsilon' H_\alpha \varepsilon}{n} + o_p(n^{-1}).$$

(ii) Pandang validasi silang-d dengan ***S*** dibentuk dengan suatu rancangan blok tak lengkap berimbang. Lebih lanjut diasumsikan d dipilih sedemikian hingga

$$\frac{d}{n} \to 1, \quad \frac{n}{n-d}\max_{i\le n} h_{i\alpha} \to 0 \text{ untuk semua } \alpha \in \boldsymbol{A}, \text{dan}$$

$$\lim_{n\to\infty} \max_{s\in S} \left\| \frac{1}{d} X_{\alpha,s}' X_{\alpha,s} - \frac{1}{n-d} X_{\alpha,s^c}' X_{\alpha,s^c} \right\| = 0$$

Maka, apabila α dalam kategori I (*an incorrect model*),

$$\hat{\overline{mse}}_{CV}(\alpha) = \overline{mse}(\alpha) + o_p(n^{-1});$$ berlaku dengan $\hat{\overline{mse}}_{CV}(\alpha)$ diganti dengan $\hat{\overline{mse}}_{CV-d}(\alpha)$

sedangkan apabila α dalam kategori II (*a correct model*),

$$\hat{\overline{mse}}_{CV-d}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{\sigma^2 p_\alpha}{n-d} - \frac{\varepsilon' H_\alpha \varepsilon}{n-d} + o_p\left(\frac{1}{n-d}\right).$$

(iii) Pandang validasi silang-d, dengan ***S*** dibentuk dengan mengambil suatu sampel random sederhana berukuran B dari koleksi semua himpunan bagian dari {1,2,…,n}. Diasumsikan semua syarat dalam (ii) dan $n^2/[B(n-d)^2] \to 0$. Maka hasil dalam bagian (ii) berlaku dengan $\|\varepsilon\|^2/n$ diubah menjadi $\sum_{s\in S}\|\varepsilon_s\|^2/[B(n-d)]$.

(iv) Lebih lanjut diasumsikan bahwa $\liminf_{n} \inf_{\alpha\, dlm kategori II} \Delta(\alpha) > 0$.

Maka $\lim_{n\to\infty} P\{\hat{\alpha}_{CV} = \text{model } I\} = 0$ dan $\lim_{n\to\infty} P\{\hat{\alpha}_{CV} = \alpha_0\} < 1$ berlaku ; dan

$\hat{\alpha}_{CV-d}$ berlaku, yaitu $\lim_{n\to\infty} P\{\hat{\alpha}_{CV-d} = \alpha_0\} = 1$.

Meskipun $\hat{\overline{mse}}_{CV}(\alpha)$ merupakan estimator yang hampir tak bias dari $\overline{mse}(\alpha)$, Teorema 1 diatas telah menunjukkan bahwa, jika $\hat{\alpha}_{\text{CV}}$ adalah model terpilih dengan menggunakan CV-1, maka

$$\lim_{n\to\infty} P\{\hat{\alpha}_{CV} = \textit{Model } I\} = 0 \quad \text{dan} \tag{5}$$

$$\lim_{n\to\infty} P\{\hat{\alpha}_{CV} = \alpha_0\} < 1 \tag{6}$$

kecuali hanya model II, yaitu $\alpha = \{1,2,\ldots,p\}$. Hal ini berarti bahwa CV-1 tak konsisten (kecuali hanya jika semua komponen dari β tak nol) dan ini terlalu konservative yaitu cenderung memilih model dengan ukuran besar. Ketakkonsistenan dari CV-1 dapat dijelaskan sebagai berikut.

Pertama, konsistensi dari sebarang metode pemilihan model berdasarkan peminimalan $\hat{\overline{mse}}(\alpha)$ atas $\alpha \in \boldsymbol{A}$ ekuivalen dengan konsistensi dari $\hat{\overline{mse}}(\alpha) - \hat{\overline{mse}}(\gamma)$ sebagai suatu estimator dari selisih

$$\overline{mse}(\alpha) - \overline{mse}(\gamma) = \sigma^2(p_\alpha - p_\gamma)/n + \Delta(\alpha) - \Delta(\gamma)\ , \alpha, \gamma \in \boldsymbol{A} \tag{7}$$

dengan p_α ukuran dari α dan $\Delta(\alpha) = \frac{1}{n}\beta' X'(I - H_\alpha)X\beta$.

Kedua, apabila α dan γ kedua-duanya model II $(\Delta(\alpha) = \Delta(\gamma) = 0)$,

$$\hat{\overline{mse}}_{CV}(\alpha) - \hat{\overline{mse}}_{CV}(\gamma) = 2\sigma^2(p_\alpha - p_\gamma)/n - [\varepsilon'(H_\alpha - H_\gamma)\varepsilon]/n + o(n^{-1})$$

merupakan estimator hampir tak bias tetapi tak konsisten dari $\overline{mse}(\alpha) - \overline{mse}(\gamma)$.

Validasi Silang Lepas-d

Seperti ditunjukkan dalam teorema 1 diatas bahwa dibawah beberapa syarat yang lemah $\hat{\alpha}_{\text{CV-d}}$, model terpilih dengan menggunakan validasi silang lepas-d, adalah konsisten yaitu :

$\lim_{n\to\infty} P\{\hat{\alpha}_{CV} = \alpha_0\} = 1$ jika dan hanya jika $d/n \to 1$ *dan* $n - d \to \infty$.

Tentu saja hal ini sangat mengherankan karena ukuran data validasi d harus jauh lebih besar dari ukuran data konstruksi (n-d) yang secara total berlawanan dengan validasi silang lepas-1. Secara teknis syarat $d/n \to 1$ diperlukan karena hal ini merupakan syarat perlu dan cukup untuk konsistensi dari $\hat{\overline{mse}}_{CV-d}(\alpha) - \hat{\overline{mse}}_{CV-d}(\gamma)$ sebagai estimator dari $\overline{mse}(\alpha) - \overline{mse}(\gamma)$.

Dari persamaan (4) dapat dilihat bahwa $\hat{\overline{mse}}_{CV-d}(\alpha)$ merupakan estimator $\overline{mse}_{n-d}(\alpha)$ bukan $\overline{mse}_{n}(\alpha)$. Apabila α dalam kategori II, $\Delta(\alpha) = 0$,

$$\overline{mse}_{n-d}(\alpha) = \sigma^2 + \frac{\sigma^2 p_\alpha}{n-d}\ .$$

Perlu diketahui bahwa α_0 meminimalkan $\overline{mse}_{m}(\alpha)$ untuk suatu m tertentu. Sebagai suatu fungsi α, jika d kecil maka $\overline{mse}_{n-d}(\alpha)$ juga kecil. Hal ini berakibat dengan suatu d yang kecil, akan sangat sulit menentukan minimum dari $\overline{mse}_{n-d}(\alpha)$ untuk semua $\alpha \in \boldsymbol{A}$.

SIMULASI

Untuk memberikan gambaran yang jelas tentang prosedur penentuan factor utama penyebab pemadaman listrik di kota Semarang, pada bagian ini diberikan hasil simulasi terhadap data yang telah dicatat

pada kantor PLN Semarang selama 4 tahun belakangan. Adapun variabel-variabel yang terlibat dalam pemodelan ini adalah: Jumlah pemadaman listrik sebagai variabel respon y, dan variabel prediktornya adalah: jumlah kerusakan jaringan (x1), jumlah kerusakan trafo (x2) serta jumlah kerusakan sekering (x3). Hasil simulasi untuk menentukan estimasi rata-rata sesatan prediksi dengan menggunakan **software 'R'** disajikan dalam Tabel 1.

Tabel.1: estimasi mse untuk 7 model yang mungkin

No.	Variabel-variabel dalam Model			Estimasi mse (CV-1)	Estimasi mse (CV-7)	Estimasi mse (CV-8)	Estimasi mse (CV-9)	Estimasi mse (CV-10)	Estimasi mse (CV-11)
	x1	x2	x3						
1	x1			0.290	0.436	0.462	0.415	0.495	0.486
2		x2		**0.070**	**0.088**	**0.098**	**0.096**	**0.119**	**0.144**
3			x3	0.114	0.158	0.164	0.156	0.189	0.256
4	x1	x2		0.025	0.038	0.056	0.049	0.079	0.099
5	x1		x3	**0.015**	**0.024**	**0.032**	**0.036**	**0.038**	**0.045**
6		x2	x3	0.068	0.129	0.160	0.181	0.200	0.193
7	x1	x2	x3	**0.010**	**0.024**	**0.037**	**0.038**	**0.051**	**0.052**

Berdasarkan hasil perhitungan estimasi mse pada Tabel 1, diperoleh estimasi model regresi terbaik: ln(y) = 1,765 + 0,394 ln(x1) + 0,555 ln(x3).

KESIMPULAN

Prosedur pemilihan model dengan menggunakan metode validasi-silang lepas-d (1<d<n) konsisten untuk $\frac{d}{n} \to 1$ dan $(n-d) \to \infty$ dengan n: ukuran sampel. Berdasarkan simulasi yang dilakukan, diperoleh model regresi terbaik dengan melibatkan 2 variabel predictor: yaitu:

ln(y) = 1,765 + 0,394 ln(x1) + 0,555 ln(x3) dengan x1: kerusakan jaringan dan x3: kerusakan sekering. Dengan demikian faktor utama penyebab gangguan listrik di kota Semarang adalah kerusakan jaringan dan kerusakan sekering.

DAFTAR PUSTAKA :

[1]. Hjorth, J.S.U, *Computer Intensive Statistical Methods, Validation Model Selection and Bootstrap*, Chapman and Hall, New York, 1994.

[2]. Shao, J, Linier Model Selection by Cross-Validation, *Journal American Statistics Assosiation*, Vol. 88, pp. 486-494, 1993.

[3]. Shao, J. and Tu., *The Jackknife and Bootstrap*, Springer-Verlag, New York, 1995.

[4]. Shao, J, An Asymptotic Theory for Linear Model Selection, *Statistica Sinica*, Vol. 7, pp. 221-264, 1997.

[5]. Tarno, Pemilihan Model Regresi Linier Terbaik dengan Validasi-Silang lepas-d, *Jurnal Sains dan Matematika*, FMIPA UNDIP, 2004.

LAMPIRAN

Lampiran 1: Listing Program dengan Software 'R'

```
VSde<-function(d, M)
{
    x1 <- c(4.63,4.61,3.99,4.49,4.11,3.76,3.26,4.29,4.62,5.61,5.93,6.38,4.44,5.51,5.74,5.95)
    x2 <- c(3.81,3.69,3.33,4.14,3.81,2.48,2.77,2.83,3.64,3.14,4.16,3.30,4.97,6.05,6.45,6.64)
    x3 <- c(4.84,4.88,4.48,5.37,4.65,4.13,4.33,4.61,3.69,4.48,4.76,4.04,5.67,6.54,6.98,7.22)
    y <- c(6.20,6.28,5.86,6.53,6.26,5.36,5.33,6.01,5.79,6.29,6.58,6.65,6.74,7.55,7.92,8.12)
    b1 <- matrix(0, 2 * 16, nrow = 2)
    b2 <- matrix(0, 2 * 16, nrow = 2)
    b3 <- matrix(0, 2 * 16, nrow = 2)
    b12 <- matrix(0, 3 * 16, nrow = 3)
    b13 <- matrix(0, 3 * 16, nrow = 3)
    b23 <- matrix(0, 3 * 16, nrow = 3)
    b123 <- matrix(0, 4 * 16, nrow = 4)
    s1 <- rep(0, M)
    s2 <- rep(0, M)
    s3 <- rep(0, M)
    s12 <- rep(0, M)
    s13 <- rep(0, M)
    s23 <- rep(0, M)
    s123 <- rep(0, M)
    I <- c(1:16)
    Is <- matrix(0, d * M, nrow = d)
    Ic <- matrix(0, (16 - d) * M, nrow = 16 - d)
    for(j in 1:M) {
        Is[, j] <- sample(I, replace = F, size = d)
        Ic[, j] <- I[ - c(Is[, j])]
        yy <- y[Ic[, j]]
        xx1 <- x1[Ic[, j]]
        xx2 <- x2[Ic[, j]]
        xx3 <- x3[Ic[, j]]
        xxx1 <- x1[Is[, j]]
        xxx2 <- x2[Is[, j]]
        xxx3 <- x3[Is[, j]]
        C <- rep(1, d)
        y1 <- y[Is[, j]]
        b1[, j] <- glm(yy ~ xx1)$coef
        X1 <- cbind(C, xxx1)
        s1[j] <- 1/(d) * sum((y1 - (X1 %*% b1[, j]))^2)
        b2[, j] <- glm(yy ~ xx2)$coef
        X2 <- cbind(C, xxx2)
        s2[j] <- 1/(d) * sum((y1 - (X2 %*% b2[, j]))^2)
        b3[, j] <- glm(yy ~ xx3)$coef
        X3 <- cbind(C, xxx3)
        s3[j] <- 1/(d) * sum((y1 - (X3 %*% b3[, j]))^2)
        b12[, j] <- glm(yy ~ xx1 + xx2)$coef
        X12 <- cbind(C, xxx1, xxx2)
        s12[j] <- 1/(d) * sum((y1 - (X12 %*% b12[, j]))^2)
        b13[, j] <- glm(yy ~ xx1 + xx3)$coef
        X13 <- cbind(C, xxx1, xxx3)
        s13[j] <- 1/(d) * sum((y1 - (X13 %*% b13[, j]))^2)
        b23[, j] <- glm(yy ~ xx2 + xx3)$coef
        X23 <- cbind(C, xxx2, xxx3)
        s23[j] <- 1/(d) * sum((y1 - (X23 %*% b23[, j]))^2)
        b123[, j] <- glm(yy ~ xx1 + xx2 + xx3)$coef
        X123 <- cbind(C, xxx1, xxx2, xxx3)
        s123[j] <- 1/(d) * sum((y1 - (X123 %*% b123[, j]))^2)

    }
    cat(   "MSE.1     =", 1/M * sum(s1), "MSE.2     =", 1/M * sum(s2), "MSE.3     =", 1/M *
    sum(s3), "\n",
            "MSE.12   =", 1/M * sum(s12), "MSE.13   =", 1/M *sum(s13), "\n", "MSE.23   =", 1/M
    * sum(s23),  "\n",
            "MSE.123  =", 1/M * sum(s123), "\n")
    }
```