PENTAPISAN DAN DETEKSI DATA OUTLIER DALAM PROSES SISTEM AKUSISI DATA PADA PROSES SINTERING

Adhi Mahendra

Program Studi Teknik Elektro, Fakultas Teknik Universitas Pancasila
Srengseng Sawah, Jagakarsa, Jakarta Selatan
e-mail: johanes70am@yahoo.com

Abstract– Industrial data is often in large databases, including both direct process measurements and calculated values. Unmeasured disturbances together with process and instrument failures contaminate the data called as outlier. Outliers are observations that deviate significantly from the majority of observations, and also effect on statistical properties of the data, if included in calculations. One main aim of data pre-processing is the detection of outliers, removing them and replacing them with better data. This paper will examine the outlier filtering techniques on time series data of temperature and flow rate of hydrogen gas as a result of the measurement sensors on the sintering process and outlier detection method based on LOF (Local Outlier Factor) of hydrogen gas as output in the sintering process. Data from the LOF method then will be compared with DBSCAN method. From the experimental results shown that the optimal parameter filtering technique will produce an adequate signal to noise ratio while still to maintain peak signal on the measurement results. And of the two methods used in detecting it turns out have the same data in the average detection is above 96% of the detected data as normal data. Both methods LOF and DBSCAN are highly sensitive to the threshold value. While the parameters have a high degree of dependence on the data input. Therefore we need seek a method to determine the parameters on both the data outlier detection method that can be adapted to the real data to be processed.
Key Words– outlier detection; filtering, time series data, sintering process, local outlier factor

Abstrak – Data pengukuran yang dihasilkan dalam sebuah proses industri sering dikategorikan sebagai sebuah basis data yang besar, termasuk hasil pengukuran proses langsung dan nilai-nilai yang dihitung. Gangguan yang terjadi dalam sebuah proses pengamatan yang diukur timbul secara bersamaan dengan kegagalan proses dan instrumen, yang dapat mengkontaminasi data yang sesungguhnya, hal ini dapat disebut sebagai outlier. Outliers adalah suatu deviasi observasi yang signifikan dari suatu observasi utama, yang juga berdampak pada sifat data statistik jika ikut dikalkulasi. Salah satu tujuan utama data preprocessing adalah mendeteksi outlier, memindahkannya dan menggantikannya dengan data yang lebih baik. Tulisan ini akan meneliti teknik penyaringan outlier pada data time series dari suhu dan laju aliran gas hidrogen sebagai hasil pengukuran sensor pada proses sintering dan deteksi outlier berbasiskan pada metode LOF (Local Outlier Factor) dari gas hidrogen sebagai hasil keluaran pada proses sintering. Data dari metode LOF kemudian akan dibandingkan dengan metode DBSCAN. Dari hasil percobaan terlihat bahwa parameter yang optimal pada teknik penyaringan akan menghasilkan sinyal yang memadai terhadap rasio kebisingan sementara yang tetap untuk mempertahankan sinyal puncak pada hasil pengukuran. Dan dari kedua metode yang digunakan dalam mendeteksi ternyata memiliki kesamaan data dalam rata rata pendeteksian yaitu diatas 96 % dari data yang terdeteksi sebagai data yang normal. Kedua metode baik LOF maupun DBSCAN sangatlah sensitif pada nilai ambang batas (threshold). Sementara parameternya memiliki tingkat ketergantungan yang tinggi pada data masukannya. Oleh karena itu dibutuhkan mencari suatu metode untuk menentukan parameter pada kedua metode pendeteksi outlier data yang dapat diadaptasikan pada data real untuk diproses.
Kata Kunci– outlier detection; filtering, time series data, sintering process, local outlier factor

I. PENDAHULUAN

Outlier adalah sebuah data observasi yang memiliki deviasi yang signifikan atau anomali dari kebanyakan observasi [1]. *Outlier* dapat disebabkan oleh sebuah sinyal menusuk, pengukuran sensor yang keliru, derau yang disebabkan oleh proses peralatan, degradasi peralatan atau beberapa kesalahan manusia. Dalam aplikasi sistem kontrol data outlier akan membawa kita pada analisi data yang menjadi tidak berguna karena outlier dapat membawa mspesifikasi model yang tidak sesuai, estimasi parameter menjadi bias bahkan akan menghasil-kan analisa dan kesimpulan yang tidak akurat atau benar.

Terdapat beberapa pendekatan yang berbeda untuk mendeteksi *outlier*. Kali ini kita fokus pada metode berbasis kerapatan (*density based*) untuk mendeteksi *outlier*. Metode ber-basis kerapatan menganggap bahwa *outlier* yang terjadi jauh dari sinyal sinyal yang khas yaitu seperti dalam kawasan kerapatan yang rendah dalam ruang masukan. Penggunaan luas komputer dalam instrumentasi proses kimia dan fisika serta fleksibilitas pemrograman perangkat lunak membuat penggunaannya untuk deteksi noise dan filtering [1], pembersihan data outlier [2-4] dan median filtering [5-7] banyak diimplementasikan dengan hasil yang cukup memuaskan.

Mengestimasi kerapatan dapat dilaksana-kan dengan menggunakan berbasis regresi, parametric, semi parametric dan metode non parametric. Di dalam metode berbasis regresi dimana sisa dari hasil sinyal tes dipergunakan sebagai skor atau nilai *outlier* [2]. Dalam metode parametric, sampel dari sinyal sinyal yang khas

dianggap untuk mematuhi distribusi yang diketahui dan test statistik dipergunakan untuk mengidentifikasi *outlier* [3,4]. Metode semi-parametric mengasumsikan sebuah campuran distribusi parametrik untuk dan atau sampel *outlier* yang khas dan memakai test statistik untuk memperoleh *outlier* [5,6]. Metode non parametric menggunakan historgram atau kernel sebagai basis cara untuk mengestimasi kerapatan untuk menghitung nilai atau skor *outlier* [7,8]. Kemudian, kerapatan estimasi di-transformasikan ke dalam *outlier* yang bisa diukur. Metode berbasis pengelompokan semi parametric membuat asumsi yang berbeda pada *outlier*.

Dalam [9], sebuah sinyal yang bukan milik sebarang kelompok dianggap *outlier*, sementara dalam [10] sebuah sinyal yang jauh dari pusat sebarang kelompok dan atau di dalam sebuah kelompok kecil dianggap sebagai *outlier*.

Metode *non-parametric* berbasis tetangga terdekat digunakan untuk pengukuran yang berbeda untuk menentukan *outlier* seperti jarak k ke tetangga [11], jumlah jarak k tetangga terdekat [12] atau jumlah jarak k tetangga dalam lingkungan dari ukuran yang pasti [13]. Terdapat juga beberapa cara yang memperhitungkan jumlah kerapatan relatif di sekitar sinyal untuk menghitung bskor atau nilai *outlier* seperti metode *Local Outlier Factor* (LOF) [14].

Dalam tulisan ini, metode *Local Outlier Factor* (LOF) dipergunakan untuk mendeteksi *outlier* dalam waktu data runtun dari laju aliran rata rata Hidrogen sebagai hasil atau keluaran sensor laju hidrogen. Nilai faktor *outlier* lokal mengkalkulasi untuk berbagai nilai k dan mengambil nilai maksimumnya untuk setiap sinyal. Hal ini dilakukan untuk mendapatkan parameter optimum untuk waktu data runtun dari laju rata rata hidrogen. Hal ini menjadi penting bagi nilai LOF secara individu dan tidak mengambil resiko kemungkinan kehilangan sebarang *outlier*.

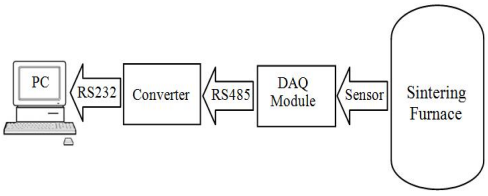
II. MATERIAL DAN METODE

A. Proses Sintering dan Akusisi Data

Sintering adalah sebuah perlakuan panas yang diterapkan pada sebuah bubuk kompak untuk memberi kekuatan dan integritas. Temperatur yang dipergunakan untuk proses sintering berada dibawah titik peleburan dari unsur material bubuk metalurgi. Data percobaan diperoleh dari industri perlakuan panas pada tungku pembakaran sintering dengan atmosfir hidrogen yang murni.

Analisa pengukuran instrumentasi antar-muka pada PC untuk data akusisi secara *on line* saat ini sudah menjadi strandar praktis dalam laboratorium yang modern. Untuk meng-eliminasi ketidakpastian waktu dari pengukuran temperatur dan laju aliran, sebuah program akusisi data yang baru sudah di implementasikan, khusus untuk tungku pembakaran. Seperti yang ditunjukkan pada Gambar 1 yang adalah diagram sistem akusisi data untuk pengukuran laju rata-rata hidrogen.

Pengumpulan data dilakukan selama satu putaran proses sintering sekitar 36 jam untuk mendapatkan waktu data runtun dari laju aliran hydrogen.



Gambar 1 Sistem Data Akusisi untuk pengukuran laju rata rata gas hidrogen

Pengumpulan data dan percobaan dilaku-kan dengan parameter proses sintering sebagai berikut: panas rata rata 250/jam, temperatur perendaman adalah 1700 C, kondisi pendinginan dilakukan secara alami dan menggunakan nitogen dan hidrogen sebagai gas atmosfir dalam proses sintering. Data percobaan yang diukur selama proses sintering mendekati 36 jam dengan total sampel data sebesar 73.324 data. Dari keseluruhan data sekitar 25000 yang diambil sebagai waktu runtun data dari laju aliran rata rata gas hidrogen dan dipergunakan sebagai data percobaan dalam studi ini seperti pada Gambar 2 dan Gambar 3.

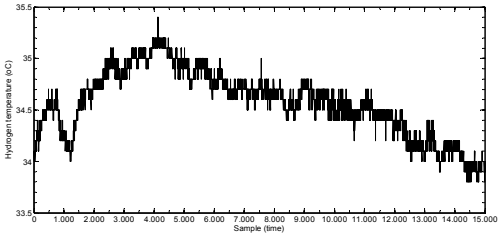
B. Median Filter

Dengan asumsi bahwa pembangkitan *outlier* dapat dijelaskan oleh model *outlier* aditif yang populer pada *robust-time series analysis* [2,8] sebagai berikut

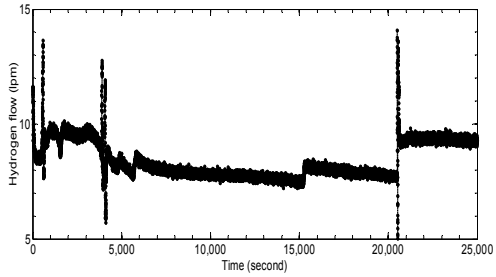
$$y_k = x_k + o_k \tag{1}$$

dimana y_k adalah urutan data hasil pengukuran, x_k adalah urutan data nominal yang kita inginkan dan o_k menggambarkan urutan data yang berisi *outlier*. Nilai-nilai urutan o_k diasumsikan menjadi nol kecuali untuk kasus waktu sesaat dengan nilai o_k yang jauh lebih besar dibandingkan dengan variasi nominal yang terlihat dalam data. Pencarian nilai perkiraan x_k didasarkan pada pengamatan data saat ini dan sebelumnya $y_k - j$ untuk $k \geq j \geq 0$. Secara khusus data, y_k ,N - 1 dan $y_k - j$ disimpan dalam jendela data W_k dengan lebar N.

$$W_k = \{y_k - N + 1, y_k - N + 2, ..., y_k\} \tag{2}$$



Gambar 2 Data Temperatur Gas Hidrogen pada Proses Sintering



Gambar 3. Laju rata-rata Gas Hidrogen pada proses sintering

Nilai data dalam jendela kemudian disusun berdasarkan peringkat untuk mendapatkan R_k sebagai berikut:

$$R_k = \left\{ y^k_{(1)} \le y^k_{(2)} \le \ldots \le y^k_{(N)} \right\}, \quad (3)$$

dan median y^m_k dari urutan R_k dihitung sebagai berikut:

$$y^m_k = \begin{cases} y^k_{((N+1)/2)} & \text{for } N \text{ odd} \\ (y^k_{(N/2)} + y^k_{(N/2+1)})/2 & \text{for } N \text{ even} \end{cases} \quad (4)$$

Nilai median y^m_k menyediakan referensi data nominal sebagai kompensasi data saat ini y_k, kemudian dievaluasi dengan menentukan jarak-jarak d_k antara y_k dan y^m_k yang ditentukan sebagai berikut:

$$d_k = \left| y^m_k - y_k \right| \quad (5)$$

Jika jarak tersebut melebihi ambang batas yang ditentukan yaitu Tk ≥ 0, maka y_k dinyatakan sebagai *outlier* dan menggantinya dengan nilai prediksi y^m_k untuk mendapatkan urutan data setelah di filter f_k, dengan ketentuan sebagai berikut:

$$f_k = \begin{cases} y_k & if \;\; d_k \le T_k \\ y^{pred}_k & if \;\; d_k \succ T_k \end{cases} \quad (6)$$

Proses untuk menghapus *outlier* pada data runtun waktu temperatur dan laju alir gas hidrogen dilakukan dengan memvariasikan ukuran jendela geser *N,* yang bertujuan untuk mencari ukuran jendela yang tepat yang menghasilkan rasio sinyal terhadap *noise* yang terbaik.

C. Faktor *Density Local Outlier*

Metode ini menggunakan distribusi kerapatan dari titik data dalam himpunan data. Untuk menentukan *outlier* menggunakan basis kerapatan sebagai data titik diambil dalam pertimbangan dimana kerapatannya dibanding-kan dengan kerapatan tetangga.

Perbandingan ini dipertimbangkan sebagai skor untuk mendetekasi *outlier*. Usulan atau ide *outlier* lokal berbasis kerapatan yang dipakai dibandingkan dengan kerapatan tetangga lokal diperkenalkan oleh Breuing et al. disebut sebagai *Local Outlier Factor* (LOF) yang sudah luas dipergunakan [14]. Metode ini menghitung derajat keberadaan *outlier* untuk setiap sinyal berdasarkan pada kerapatan lokal disekitarnya.

Umumnya metode mendeteksi *outlier* tetangga terdekat mengukur ketiadaan *outlier* dalam konteks jarak dalam sinyal yang lain dalam himpunan data. Itulah sebabnya, pendekatan ini beresiko kehilangan *outlier* dalam himpunan data dimana kerapatan lokal sangatlah bervariasi. Dua kelompok C_1 dan C_2 dapat dilihat pada Gambar 3 dimana hal ini diharapkan bahwa o_1 dan o_2 akan dianggap *outlier* oleh metode pendeteksi. Namun, kebanyakan metode berbasis tetangga terdekat yang mengandalkan jarak akan menamakan sinyal sinyal dari kelompok C_1 sebagai *outlier* jika disetel untuk mendteksi o_2.

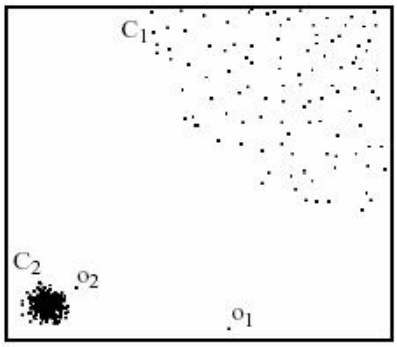
Metode LOF mengatasi kekurangan ini dengan mempertimbangakan perbedaan dalam kerapatan lokal disekitar sinyal sebagai *outlier* yang bisa diukur. Jika sinyal kerapatan di sekitar o lebih tinggi dari kerapatan disekitar tetangga, LOF dari o akan menjadi lebih tinggi.

Himpunan data masukan diberi notasi $X = \{(x^{(t)}, r^{(t)})\}_{t=1}^{N}$, $x, x^{(i)}$ and $x^{(j)}$ yang merepresentasikan sinyal dari himpunan data $d = (x, x^{(i)})$ dan menotasikan jarak antara x and $x^{(i)}$. Jarak k dari sinyal x dinotasikan oleh k jarak (x), diberikan jarak dari sinyal x ke k^{th} itusendiri sebagai tetangga terdekat.

Jarak -k $(x) \equiv d(x, x^{(i)})$ sehingga

$d(x, x^{(j)} \le d(x, x^{(i)})$ untuk k kejadian

$d(x, x^{(j)} \prec d(x, x^{(i)})$ untuk k-1 kejadian (7)



Gambar 4. Sinyal Himpunan Data [14]

Lingkungan jarak k dari suatu sinyal x, dinyatakan oleh $N_k(x)$, mengandung semua sinyal yang dekat kepada x daripada nilai jarak k:

$$N_k(x) = \{x^{(t)} \in X \setminus \{x\} \mid d(x,x^{(t)}) \prec k - \text{distance}(x)\} \quad (8)$$

Kemampuan mencapai jarak dari suatu sinyal x dengan sehubungan dengan sinyal $x^{(i)}$ lain adalah jarak diantara dua sinyal tetapi agar mencegah fluktuasi, ketercapaian kemampuan jarak dimuluskan dalam lingkungan jarak k oleh penandaan jarak ketercapaian kemampuan yang sama yang ada dalam lingkungan $x^{(i)}$:

$$reachdist_k(x,x^{(i)} = \max\{k - \text{distance}(x^{(i)}, d(x,x^{(i)})\} \quad (9)$$

Kerapatan ketercapaian kemampuan lokal mengukur bagaimana mudahnya hal ini untuk mencapai sebuah kejadian pasti dan dikalkulasi sebagai invers dari jarak rata-rata ketercapaian kemampuan sinyal dalam jarak lingkungan k

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{x^{(i)} \in N_k(x)} reachdist_k(x, x^{(i)})} \quad (10)$$

Setelah mendefinisikan kerapatan ketercapaian kemampuan lokal, kita dapat berpindah pada langkah puncak dimana kita menghitung faktor lokal *outlier* untuk setiap sinyal. Kita berharap sinyal-sinyal dengan variasi tinggi antara kerapatan ketercapaian kemampuan lokal dari longkungannya mengambil nilai LOF yang lebih tinggi:

$$LOF_k(x) = \frac{\sum_{x^{(i)} \in N_k(x)} \frac{lrd_k(x^{(i)})}{lrd_k(x)}}{|N_k(x)|} \quad (11)$$

Setiap permasalahan yang spesifik memerlukan sebuah jarak parameter yang melangkah mencari nilai k yang optimum.Selain itu, nilai LOF cukup sensitif pada nilai k dan bisa menunjukkan sifat ketidakstabilan sebagai perubahan k. Dalam hal meminimisasi ketidakstabilan nilai LOF, nilai LOF dihitung untuk perubahan nilai k dan mengambil maksimumnya dari setiap sinyal. Meskipun, hal ini memungkinkan untuk diterapkan pada perbedaan heuristik untuk mengkombinasikan nilai LOF berlapis, mengambil nilai maksimum memberikan hal yang penting bagi nilai LOF dan tidak beresiko kehilangan sebarang *outlier*.

D. Aplikasi *Density-Based Spatial Clustering* dengan Derau.

DBSCAN adalah algoritma kerapatan berbasis pengelompokkan.Algoritma ini mengembangkan kawasan kawasan dengan kerapatan tinggi ke dalam kelompok dan menemukan kelompok dari perubahan bentuk dalam basis data spasial dengan derau. DBSCAN mendefinisikan sebuah kelompok sebagai himpunan kerapat titik terhubung yang maksimal. Setiap titik dalam himpunan data dibagi kedalam kelompok titik dan noise. Ide kuncinya adalah bahwa setiap titik dalam kelompok lingkungan dalam radius yang diberikan mengandung setidaknya sejumlah titik minimum. Misalnya kerapatan dalam lingkungan memiliki nilai yang melampaui ambang batas.

Untuk menyediakan deskripsi yang dapat dimengerti dari algoritma ini definisi-definisi yang sama harus disajikan. Lingkungan dalam radius epsilon (ε) dari obyek yang diberikan disebut obyek lingkungan ε. Jika obyek lingkungan ε mengandung setidaknya sebuah bilangan minimum, *MinPts* dari obyek, maka obyeknya disebut inti obyek. Diberikan seperangkat obyek *X*, kita katakan bahwa obyek p adalah kerapatan yang mampu dicapai secara langsung dari obyek *q* dalam lingkungan ε dari *q*, dan *q* adalah obyek inti

Sebuah obyek p adalah kerapatan yang dapat dicapai dari obyek q terkait ε dan MinPts didalam himpunan obyek *X*, jika terdapat rantai obyek $p_1,\ldots, p_n$, dimana $p_1=p$dan$p_n = p$ sehingga p_{i+1} adalah kerapatan yang dapat dicapai langsung dari p_i berkenaan dengan ε dan *MinPts*, for $1 \leq i \leq$ n, $p_i \in X$.

Obyek p adalah kerapatan yang terhubung dengan obyek q sehubungan dengan e dan minpts dalam himpunan obyek, X, jika terdapat obyek $o \in X$ sehingga p dan q kerapatan yang dapat tercapai dari 0 berkenaan dengan ε and *MinPts*.

Ketercapaian kemampuan kerapatan adalah pelengkap penutup dari kemampuan ketercapaian kerapatan langsung, dan hubungan ini adalah asimetrik. Hanya obyek inti yang kemampuan ketercapaian kerapatannya satu

dengan yang lain. Hubungan kerapatannya adalah simetrik.

Kelompok berbasis kerapatan adalah seperangkat himpunan kerapatan yang terhubungan dengan obyek yang maksimal berkenaan dengan ketercapaian kemampuan kerapatannya. Setiap obyek tidak mengandung sebarang kelompok yang dipertimbangkan sebagai derau.

III. HASIL PERCOBAAN DAN SIMULASI

Pada penelitian ini bentuk sinyal asli terutama puncak sinyal sangat penting sehingga akan dicari ukuran lebar jendela filter yang optimal dengan menentukan kriteria rata-rata kuadrat error (MSE) terendah dan rasio sinyal terhadap *noise* (SNR) yang paling besar sebagai lebar jendela filter yang optimal untuk kasus ini. Untuk mencapai hal tersebut pada penelitian ini variasi ukuran jendela yang digunakan dibagi menjadi ukuran jendela kecil yaitu 115, 135, 175 dan ukuran jendela yang lebih besar yaitu 215, 235 dan 275. Kriteria MSE dan SNR dihitung menggunakan persamaan berikut:

$$MSE = \frac{1}{N}\sum_{n=0}^{N}(V(n)-V_R(n))^2 \quad (12)$$

$$SNR = \log_{10}\frac{\sum_{n=0}^{N} V_R^{\,2}(n)}{\sum_{n=0}^{N} S_R^{\,2}(n)} \quad (13)$$

dimana N adalah jumlah data runtun waktu, $V(n)$ adalah sinyal asal sebelum difilter, V_R(n) adalah sinyal hasil rekonstruksi setelah difilter, dan S_R(n) merupakan perbedaan antara sinyal asal dan sinyal hasil rekonstruksi setelah difilter.

Hasil perhitungan MSE dan SNR untuk data temperatur dan laju alir gas hidrogen setelah difilter untuk menghilangkan data outlier dengan ukuran jendela kecil dan ukuran jendela yang lebih besar diperlihatkan pada Tabel 1 dan Tabel 2.

Tabel 1. Hasil perhitungan MSE dan SNR setelah di filter dengan ukuran jendela kecil

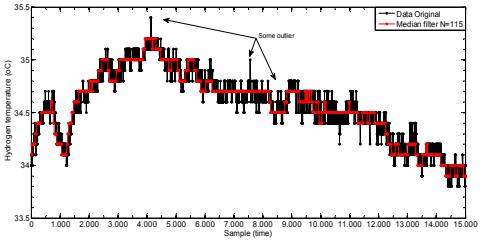
Time series data	*Criterion*	Window size (*N*)		
		115	***135***	***175***
Temperature	MSE	0,0045	0,0047	0,0052
	SNR	5,3908	5,3693	5,3288
Flowrate	MSE	0,0703	0,0699	0,0657
	SNR	3,0349	3,0373	3,0640

Tabel 2. Hasil perhitungan MSE dan SNR setelah di filter dengan ukuran jendela lebih besar

Time series data	*Criterion*	Window size (*N*)		
		215	***235***	***275***
Temperature	MSE	0,0057	0,0061	0,0067
	SNR	5,2846	5,2613	5,2168
Flowrate	MSE	0,0673	0,0689	0,0705
	SNR	3,0535	3,0436	3,0334

Pada Tabel 1 dan Tabel 2 terlihat bahwa variasi lebar jendela (*window size*) pada median filter semakin besar lebar jendela maka MSE naik yang diikuti oleh penurunan SNR dengan kata lain semakin besar ukuran jendela maka kinerja filter semakin progresif tetapi deformasi pada sinyal hasil filtering akan semakin besar terhadap sinyal aslinya.

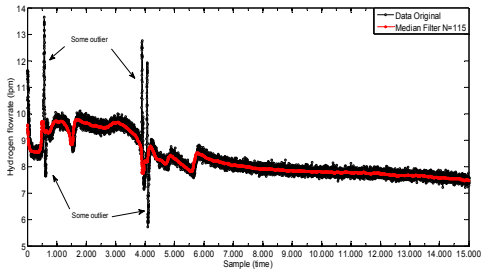
Gambar 5 dan 6 memperlihatkan beberapa potensi outlier pada data time series temperatur dan laju alir hydrogen dengan lebar jendela median filter 115.



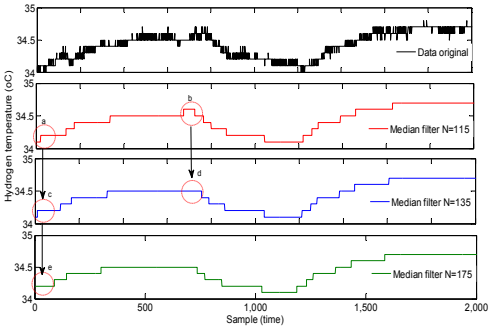
Gambar. 5 Beberapa potensi outlier pada data time series temperatur hydrogen dengan lebar jendela N=115.

Ukuran jendela geser filter (*moving size window*) pada median filter yang semakin besar mengakibatkan perubahan bentuk sinyal yang semakin progresif seperti yang diperlihatkan pada Gambar 6 untuk data *time series* temperatur hidogen dengan lebar jendela 115, 135 dan 175. Oleh karena itu pemilihan ukuran jendela median filter optimal selain dengan mempertimbangkan nilai MSE dan SNR juga pengamatan visual terhadap perubahan bentuk sinyal pada setiap ukuran jendela filter. Pengamatan visual ini merupakan *engineering judgment* dalam menentukan ukuran lebar jendela filtering sesuai kebutuhan untuk analisa selanjutnya.

Dalam repositori data dunia nyata, adalah sulit untuk menemukan himpunan data untuk mengevaluasi algoritma pendeteksi *outlier*, sebab hanya sedikit himpunan data di dunia nyata yang pasti diketahui yang mana obyeknya sangat nyata berkelaluan secara berbeda [15].



Gambar 6 Beberapa potensi outlier pada data time series flow rate hydrogen dengan lebar jendela *N*=115.



Gambar 7 Perbandingan Bentuk Sinyal Sebelum dan Sesudah di Filter dengan Ukuran Jendela Kecil.

Dalam percobaan ini kita menggunakan himpunan data industri dari laju rata rata hidrogen pada proses sintering. Pada bagian ini, kita sajikan beberapa hasil percobaan meng-gunakan aliran data waktu runtun yang real.

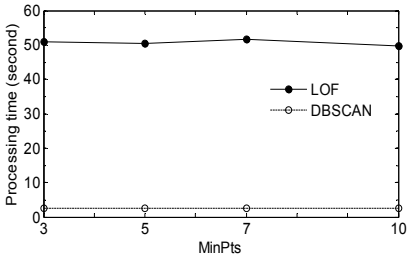
Kita garap pertama metode *local outlier factor* dengan 15.000 himpunan data untunk *training* dan 10.000 himpunan data untuk *testing*. Pada percobaan ini parameter *MinPts* diubah dari 3 ke 10. Dari beberapa referensi, dalam praktisnya nilai parameter *theta* bervariasi dari 0.1 to 0.9. Dalam studi ini kita gunakan parameter *theta* dari 0.1 ke 0.5.

Selanjutnya dengan menggunakan himpunan data tes yang sama diberikan kepada metode DBSCAN. Tabel 3 menunjukkan hasil dari pendeteksi outlier dengan menggunakan metode LOF dan DBSCAN untuk data runtun waktu yang *real* dari laju aliran hidrogen. Hasilnya menunjukkan rata-rata dari tiga percobaan. Seperti yang ditunjukkan pada tabel, peningkatan pada parameter minpts metode keduanya menumbuh-kan prosentase rendah dari pendeteksian data sebagai data normal juga sebaliknya, untuk data yang terdeteksi sebagai *outlier*.

Untuk biaya komputasi seperti yang ditunjukkan Gambar 4, metode LOF mengambil lebih banyak waktu untuk menunjukkan metode deteksi *outlier* dengan DBSCAN pada parameter *MinPts* yang sama

Tabel 3. Hasil deteksi outlier menggunakan metode lof dan dbscan untuk laju rerata hidrogen waktu runtun data

Methods	***MinPts***	**Data Points Detected as**	
		Normal (%)	***Outlier(%)***
LOF	3	98.9	1.1
	5	98.8	1.2
	7	98.3	1.7
	10	97.0	3.0
DBSCAN	3	98.5	1.5
	5	98.1	1.9
	7	97.6	2.4
	10	96.0	4.0



Gambar 8 MinPts Parameter vs Waktu Pemrosesan pada deteksi outlier dari laju rerata Hidrogen untuk waktu data runtun

IV. SIMPULAN

Dari hasil percobaan terlihat bahwa parameter yang optimal pada teknik penyaringan akan menghasilkan sinyal yang memadai terhadap rasio kebisingan sementara yang tetap untuk mempertahankan sinyal puncak pada hasil pengukuran.

Menemukan *outlier* adalah pekerjaan penting untuk semua kawasan aplikasi pada khusunya untuk memastikan kualitas data pengukuran sensor yang akan dipergunakan pada proses selanjutnya. Dalam tulisan ini, kita menunjukkan bahwa penerapan dari dua metode deteksi outlier LOF dan DBSCAN untuk data waktu runtun dari laju aliran gas hidrogen sudah diselidiki dengan perbedaan parameter minpts.Dari percobaan dihasilkan bahwa kedua metode memiliki kesamaan rata rata pendeteksian diatas 96 persen untuk mendeteksi data sebagai data normal.

Kedua metode (LOF dan DBSCAN) sangat peka pada nilai parameter ambang batas. Sementara parameter parameter ini sangat tergantung dari data masukan. Oleh karena itu diperlukan untuk mencari suatu metode untuk

menentukan parameter untuk metode pendeteksi outlier keduanya yang dapat di adaptasikan pada data real yang dapat diproses.

REFERENSI

[1] Liu, H.C., Shah, S., and Jiang, W., "On-line outlier detection and data cleaning," Computers and Chemical Engineering, Vol.28, pp. 1635-1647, 2004.

[2] Bhausaheb Shinde, A.R. Dani," Noise Detection and Removal Filtering Techniques in Medical Images", International Journal of Engineering Research and Applications, Vol. 2, Issue 4, p.311-316, July-August 2012

[3] P.H. Menold, R.K. Pearson, F. Allg¨ower," Online outlier detection and removal", Proceedings of the 7th Mediterranean Conference on Control and Automation (MED99) Haifa, Israel - June 28-30, p.1110-1133, 1999.

[4] R. Ganguli," Noise and outlier removal from jet engine healt signal using weighted FIR median hybrid filter", Mechanical Systems and Signal Processing, 16(6), p.967–978, 2002.

[5] Hancong Liu, Sirish Shah, Wei Jiang," On-line outlier detection and data cleaning", Elsevier, Computers and Chemical Engineering, 28, p.1635–1647, 2004.

[6] Lin Yin, Ruikang Yang," Weighted Median Filters: A Tutorial", IEEE Transactions on Circuit and System: Analog and Digital Signal Processing, Vol. 43, No. 3, p. 157-192, March 1996.

[7] [66] Ruikang Yang, dkk," Optimal Weighted Median Filterin Under Structural Constraints", IEEE Transactions on Signal Processing, Vol. 43, No. 3, p. 591-604, March 1995.

[8] H. Hwang and R. A. Haddad," Adaptive Median Filters: New Algorithms and Results", IEEE Transactions on Image Processing, Vol. 4, No. 4, p. 499-502, April 1995.

[9] Martin, R. and V. Yohai,"Influence Functionals for Time Series," Ann. Statist., 14, p. 781-818,1986

[10] Rousseeuw, P. J. and A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, Inc., New York, NY, USA, 1987.

[11] Laurikkala, J., M. Juhola and E. Kentala, "Informal Identification of Outliers in Medical Data", The Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, pp. 17-29, 2000.

[12] Ye, N. and Q. Chen, "An Anomaly Detection Technique Based on a Chi-square Statistic for Detecting Intrusions into Information Systems", Quality and Reliability Engineering International, Vol. 17, No. 2, pp. 105-112, 2001.

[13] Eskin, E., "Anomaly Detection over Noisy Data Using Learned Probability Distributions", Proceedings of the International Conference on Machine Learning, pp.255-262, 2000.

[14] Abraham, B. and G. E. P. Box, "Bayesian Analysis of Some Outlier Problems in Time Series", Biometrika, Vol. 66, No. 2, pp. 229-236, 1979.

[15] Eskin, E., "Modeling System Calls for Intrusion Detection with Dynamic Window Sizes", Proceedings of DARPA Information Survivabilty Conference and Exposition II (DISCEX), pp. 143-152, 2001.

[16] Desforges, M. J., P. J. Jacob and J. E. Cooper, "Applications of Probability Density Estimation to the Detection of Abnormal Conditions in Engineering", Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, Vol. 212, No. 8, pp. 687-703, 1998.

[17] Yu, J. X., W. Qian, H. Lu and A. Zhou, "Finding Centric Local Outliers in Categorical/Numerical Spaces", Knowledge Information Systems, Vol. 9, No. 3, pp.309-338, 2006.

[18] He, Z., X. Xu and S. Deng, "Discovering Cluster-based Local Outliers", Pattern Recognition Letters, Vol. 24, No. 9-10, pp. 1641-1650, 2003.

[19] Byers, S. and A. E. Raftery, "Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes", Journal of the American Statistical Association, Vol. 93, No. 442, pp. 577-584, 1998.

[20] Eskin, E., A. Arnold, M. Prerau, L. Portnoy and S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data", Applications of Data Mining in Computer Security, pp. 319-330, 2002.

[21] Knorr, E. M., R. T. Ng and V. Tucakov, "Distance-based Outliers: Algorithms and Applications", The International Journal on Very Large Data Bases, Vol. 8,No. 3-4, pp. 237-253, 2000.

[22] Breunig, M. M., H. P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying Densitybased Local Outliers", SIGMOD Record, Vol. 29, pp. 93-104, 2000