# Review Jurnal
# The Use of Biplot Analysis and Euclidean Distance with Procrustes Measure for Outliers Detection

**Latar Belakang**

Pencilan sering ditemukan dalam analisis peubah ganda. Adanya pencilan menandakan objek yang berupa pencilan tersebut berbeda dibanding objek lainnya (Hair *et al.* 2014). Perbedaan pada objek ini menjadi suatu hal yang menarik untuk ditelusuri, misalnya penyebab objek tersebut berbeda dan baik atau burukkah perbedaan tersebut. Sebagai ilustrasi pada data jumlah produksi susu sapi terdapat pencilan salah satu sapi memproduksi susu sangat sedikit dibanding sapi lainnya atau salah satu sapi memproduksi susu sangat banyak dibanding sapi lainnya. Adanya pencilan ini menjadi suatu hal menarik yang dapat diuji oleh peneliti di bidang peternakan untuk mencari penyebab sapi yang terlalu sedikit atau terlalu banyak memproduksi susu. Alasan kedua perlu dilakukannya pendeteksian pencilan adalah pencilan membuat adanya kesalahan dalam menentukan rata-rata. Selain itu, adanya pencilan dapat menimbulkan kesalahan dalam menyimpulkan data misalnya karena nilai ragam yang terlalu besar sehingga mengakibatkan salahnya pengambilan keputusan. Agar hal seperti ini tidak terjadi, ada baiknya sebelum mengolah data, lakukan terlebih dahulu deteksi pencilan.

Metode yang dapat dilakukan untuk mendeteksi pencilan diantaranya ialah Determinan Koragam Minimum (DKM). DKM adalah metode dugaan yang kekar dalam menentukan lokasi dan plot data. Tujuan dari DKM adalah menentukan $h$ objek dari $n$ objek yang determinan matriks koragamnya terkecil lalu mencari jarak Mahalanobis dari rataan matriks yang terdiri dari $h$ objek dengan data asal (Rousseeuw 1984). Suatu objek dikatakan pencilan apabila berada di luar batas elips. Seiring dengan berkembangnya ilmu pengetahuan, DKM berkembang dengan metode bernama Determinan Koragam Minimum Cepat (DKMC). DKMC dapat merumuskan penduga fungsi elips yang lebih besar dimensi matriksnya, sehingga dapat lebih cepat mendapatkan hasil dalam iterasi (Rousseeuw dan Driessen 1999).

Pendeteksian pencilan dengan DKM dan DKMC sulit untuk menentukan matriks homogen jika data yang dicari berukuran besar, oleh karena itu dilakukan deteksi pencilan lain dengan analisis biplot. Analisis biplot adalah analisis berupa pendekatan berbentuk grafik dari matriks data yang berukuran besar untuk mengumpulkan sebanyak mungkin informasi yang terdapat dalam data ke dalam tampilan yang dapat dicerna dengan mudah (Greenacre 2010). Pendeteksian pencilan pada analisis biplot dilakukan dengan dua cara, yaitu secara langsung dan tidak langsung. Analisis langsung dilakukan dengan cara menghitung jarak Euclid setiap objek dengan sentroidnya dari matriks kordinat objek dari biplot. Objek yang dikategorikan calon pencilan adalah yang paling jauh dengan sentroid dari kordinat data objek. Analisis secara tidak langsung dilakukan dengan mencari dahulu sejumlah objek homogen yang dekat dengan sentroid, kemudian ditentukan jarak Euclid dari matriks baru yang terdiri dari matriks homogen dengan masing-masing objek. Pencilan pada metode ini adalah yang melebihi batas nilai Khi-kuadrat.

Analisis lain yang dapat digunakan adalah jarak Euclid. Terdapat dua metode yang dapat dilakukan dengan analisis ini, yaitu jarak Euclid langsung dan tidak langsung. Pada metode langsung, dilakukan dengan cara menghitung jarak setiap objek dengan sentroidnya, lalu pilih objek yang paling jauh dengan sentroid sebagai calon pencilan. Pada analisis langsung, lakukan terlebih dahulu pencarian sebagian objek sebagai kumpulan anak matriks data yang homogen, lalu deteksi pencilan dengan menghitung jarak Mahalanobis setiap objek dengan rataan matriks homogen. Pencilannya adalah yang melebihi batas Khi-kuadrat.

Pemilihan metode yang baik untuk mendeteksi pencilan dilakukan dengan ukuran kesesuaian Procrustes. Analisis Procrustes adalah teknik yang mengacu pada perbandingan dua matriks dengan berbagai kondisi dan menghasilkan ukuran kesesuaian. Perhitungan dengan jarak Euclid dan transformasi dengan cara translasi-rotasi dan dilasi adalah langkah yang efektif dalam analisis Procrustes (Bakhtiar dan Siswadi 2015). Matriks tanpa pencilan yang memunyai ukuran kesesuaian Procrustes paling kecil dengan matriks asal di mana banyaknya pengurangan objek yang disebut pencilan sama adalah metode yang terbaik.

Penelitian ini bertujuan untuk untuk mengenalkan metode pencilan dengan analisis biplot serta jarak Euclid dan membandingkannya dengan pendeteksian pencilan lain, yaitu DKM dan DKMC menggunakan ukuran kesesuaian Procrustes.

**Metode Penelitian**

Data dengan ukuran yang relatif kecil yang digunakan pada penelitian ini adalah data laporan keuangan bank umum syariah pada Maret 2016 yang memunyai enam peubah yaitu CAR (*Capital Adequency Ratio*), ROA (*Return On Asset*), ROE (*Return on Equity*), NI (*Net Income*), BOPO (Biaya Operasional Pendapatan Operasional) dan FDR (*Financing to Defisit Ratio*). Objek pada data ini merupakan 12 bank syariah, yaitu Bank Syariah Mandiri, Bank Muamalat, BNI Syariah, BRI Syariah, Bank Panin Dubai Syariah, Bank Jabar Banten Syariah, BTPN Syariah, Bank Syariah Bukopin, Bank Mega Syariah, BCA Syariah, Bank Victoria Syariah, Maybank Syariah.
Data dengan ukuran yang relatif sedang pada penelitian ini adalah data 34 provinsi di Indonesia berdasarkan indikator kesejahteraan rakyat menurut Badan Pusat Statistik yang diunduh tahun 2015 dan terdiri dari tujuh indikator. Tiap indikator terdapat beberapa peubah. Indikator dan peubahnya yaitu kependudukan (laju pertumbuhan penduduk ($X_1$), kepadatan penduduk ($X_2$), jumlah penduduk miskin ($X_3$)), jumlah pendapatan (PDRB perkapita ($X_4$), jumlah perusahaan industri ($X_5$), jumlah koperasi aktif ($X_6$)), pengeluaran atau konsumsi per kapita (konsumsi makanan ($X_7$), konsumsi bukan makanan ($X_8$)), pendidikan(angka melek huruf ($X_9$), angka partisipan sekolah usia 7-12 tahun ($X_{10}$), usia 13-15 ($X_{11}$), usia 16-18 ($X_{12}$), usia 19-24 ($X_{13}$), angka partisipasi kasar SD ($X_{14}$), SMP ($X_{15}$), SMA ($X_{16}$), angka partisipasi murni SD ($X_{17}$), SMP($X_{18}$), SMA ($X_{19}$)), kesehatan (jumlah sarana kesehatan ($X_{20}$), penduduk yang memiliki keluhan kesehatan ($X_{21}$)), ketenagakerjaan (tingkat partisipasi angkatan kerja ($X_{22}$), tingkat pengangguran terbuka ($X_{23}$)) serta lingkungan dan perumahan (status kepemilikan rumah sendiri ($X_{24}$), jenis atap rumah bukan ijuk ($X_{25}$), jenis dinding rumah bukan bamboo ($X_{26}$), jenis lantai bukan tanah ($X_{27}$), sumber listrik PLN ($X_{28}$), sumber air minum layak ($X_{29}$), sanitasi layak ($X_{30}$)).

Prosedur analisis yang digunakan dalam penelitian ini adalah

a. Eksplorasi Data

Eksplorasi data dilakukan untuk mengetahui sebaran nilai peubah seperti rata-rata, simpangan baku, median, nilai minimum, nilai maksimum, diagram kotak garis dan nilai korelasi antar peubah. Peubah yang memunyai rata-rata dan median yang berbeda jauh, maka berpotensi ada pencilan berdasarkan peubah tersebut. Peubah yang memunyai simpangan baku yang sangat besar dibandingkan peubah lainnya maka peubah tersebut mendominasi dibandingkan peubah lainnya. Nilai maksimum dan nilai minimum yang sangat berbeda jauh menunjukkan rentang data pada peubah tersebut cukup besar.
Eksplorasi data lain yang dilakukan adalah melihat sebaran data. Penyebaran data dapat dilihat dengan gambar data dalam bidang berdimensi lebih rendah, dapat dilakukan di antaranya dengan analisis biplot atau analisis komponen utama. Selain itu dapat ditentukan pula korelasi antar peubah untuk mengetahui peubah yang saling berkorelasi positif dan peubah yang berkorelasi negatif.

b. Pendeteksian pencilan

Ada empat metode yang akan digunakan untuk mendeteksi pencilan, yaitu:

1. Analisis Biplot

Analisis biplot yang akan dilakukan adalah dengan menggunakan $\alpha = 0$. Pemilihan nilai $\alpha$ ini karena pada saat $\alpha = 0$, analisis biplot dapat ditentukan karakteristik data berdasarkan koragam, simpangan baku, korelasi dan kuadrat jarak Euclid biplot sebanding dengan Mahalanobis data (Siswadi dan Suharjo 1999). Untuk memastikan data yang digambarkan dengan biplot memvisualkan data dengan baik, maka tentukan ukuran kesesuaian biplot dengan ukuran Procrustes.
Terdapat dua cara dalam mendeteksi pencilan dengan analisis biplot, yaitu secara langsung dan tidak langsung. Secara langsung dilakukan dengan cara mencari biplotnya, lalu deteksi objek yang berupa pencilan dengan menghitung jarak setiap data dengan sentroidnya. Kemudian urutkan data dari yang terdekat hingga terjauh. Semakin jauh objek, maka objek tersebut semakin terpencil. Kemudian pilih beberapa objek yang merupakan objek terjauh berdasarkan ranking jarak dengan sentroid. Pemilihan banyaknya objek yang dikategorikan sebagai pencilan yaitu sama dengan banyaknya pencilan yang telah didapatkan dengan metode lain.

Pendeteksian secara tidak langsung dengan biplot yaitu dengan cara mencari terlebih dahulu matriks homogen dari titik-titik biplot yang merupakan himpunan bagian dari data biplot. Matriks homogen ini dilakukan dengan pemilihan $h$ objek yang paling dekat dengan sentroid, dengan $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ (Lopuhä dan Rousseeuw 1991). Setelah terpilih $h$ objek, maka objek tersebut menjadi matriks baru $\mathbf{Z}_{h\times p}$, dengan $p$ adalah banyaknya peubah dan $n$ merupakan banyaknya objek. Tentukan pula rata-rata matriks $\mathbf{Z}$ ($\bar{\boldsymbol{z}}_{\boldsymbol{H}}$) dan korelasi dari matriks $\mathbf{Z}$ ( $\boldsymbol{\Sigma}_{\boldsymbol{zH}}{}^{-1}$ ). Kemudian cari jarak Euclid masing-masing objek dengan matriks $\mathbf{Z}$. Objek yang disebut pencilan merupakan objek yang memiliki jarak Euclid melebihi batas Khi-kuadrat.

2. Jarak Euclid

Terdapat dua metode yang dapat digunakan untuk mendeteksi pencilan dengan jarak Euclid, yaitu secara langsung dan tidak langsung. Analisis secara langsung dilakukan dengan mencari jarak Euclid antara setiap objek dengan sentroidnya, kemudian diurutkan dari yang terkecil ke yang terbesar. Objek yang memunyai jarak terjauh merupakan pencilan. Analisis secara tidak langsung dilakukan dengan cara mencari terlebih dahulu matriks homogen dengan memilih $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ objek yang memunyai jarak terkecil dengan sentroidnya, kemudian dicari rata-rata dan matriks koragamnya, setelah itu tentukan jarak Mahalanobis setiap objek dengan rata-rata matriks homogen. Objek yang jaraknya lebih dari batasan Khi-kuadrat adalah pencilan.

3. DKM

Dalam DKM, prosedur yang dilakukan adalah tentukan $h$ objek yang memunyai determinan dari matriks koragamnya minimum, lalu tentukan rataan ($\bar{\boldsymbol{\mu}}_{\boldsymbol{H}}$) dan invers matriks koragamnya ($\boldsymbol{\Sigma}_{\boldsymbol{H}}^{-1}$) (Hubert dan Debruyne 2010). Kemudian tentukan toleransi elips

$$(\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_{\boldsymbol{H}})^T \boldsymbol{\Sigma}_{\boldsymbol{H}}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_{\boldsymbol{H}}) \leq \chi^2_{p,0.975}$$

di mana $\boldsymbol{x}_i$ adalah objek ke-$i$, $\bar{\boldsymbol{\mu}}_{\boldsymbol{H}}$ adalah rataan $h$ objek dan$\boldsymbol{\Sigma}_{\boldsymbol{H}}^{-1}$ adalah invers matriks koragam $h$ objek. Suatu objek ke-$i$ ($\boldsymbol{x}_i$) disebut pencilan jika

$(\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_{\boldsymbol{H}})^T \boldsymbol{\Sigma}_{\boldsymbol{H}}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_{\boldsymbol{H}}) > \chi^2_{p,0.975}$ .

4. DKMC

Awal dilakukannya perhitungan pada DKMC sama seperti DKM, setelah itu ubah nilai $\boldsymbol{\Sigma}_{\boldsymbol{H}}^{-1}$dan $\bar{\boldsymbol{\mu}}_{\boldsymbol{H}}$ dengan

$$\boldsymbol{\mu}_{fmcd} = \frac{\sum_{i=1}^{n} w_i \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i}$$

$$\boldsymbol{\Sigma}_{fmcd} = \left( \sum_{i=1}^{n} w_i (\mathbf{x}_i - \boldsymbol{\mu}_{fmcd}) (\boldsymbol{x}_i - \boldsymbol{\mu}_{fmcd})^T \right) \left( \sum_{i=1}^{n} w_i \right)^{-1}$$

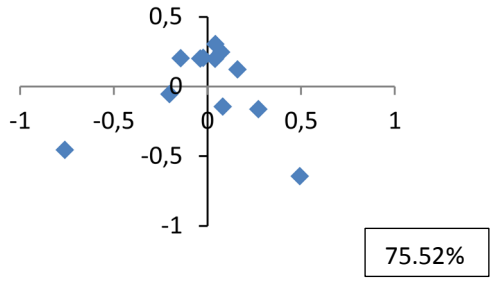
dengan bobot $w_i = 0$ jika $d_i > \sqrt{\chi^2_{p,0.975}}$ dan $w_i = 1$ jika lainnya.

Kemudian hitung kembali jarak Mahalanobis dengan rataan $\boldsymbol{\mu}_{fmcd}$ dan koragam $\boldsymbol{\Sigma}_{fmcd.}$ Suatu objek ke-$i$ ($\boldsymbol{x}_i$) disebut pencilan jika $(\boldsymbol{x}_i - \boldsymbol{\mu}_{fmcd})^T \boldsymbol{\Sigma}_{fmcd}{}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_{fmcd}) \geq \chi^2_{p,0.975}$.

c. Pemilihan metode terbaik

Pengukuran untuk memilih metode yang baik dapat dilakukan dengan ukuran kesesuaian Procrustes. Pada setiap metode hilangkan objek yang dikategorikan sebagai pencilan sama banyak, lalu tentukan matriks koragamnya. Metode yang terbaik adalah yang paling berbeda dengan data asal, yaitu yang memunyai ukuran kesesuaian Procrustes paling kecil.

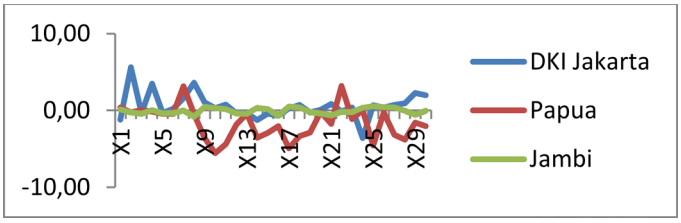
**Hasil Analisis**




Gambar 1 Biplot Data Laporan Keuangan Bank Umum Syariah Maret 2016

Ukuran kesesuaian dari koragam matriks data asal dan matriks koragam data yang pencilannya dihilangkan menunjukkan seberapa dekat jarak matriks data asal dengan matriks data tanpa pencilan, yang menginterpretasikan bahwa yang memiliki ukuran kesesuaian terkecil adalah metode terbaik. Alasan pernyataan ini adalah semakin kecil jarak antara dua matriks, maka semakin berbeda matriks tersebut. Bila matriks data asal dengan matriks data tanpa pencilan berbeda, maka objek yang dihilangkan adalah objek yang berbeda. Maka paling tepat objek tersebut disebut pencilan.

Sebelum dilakukan pendeteksian pencilan, tentukan terlebih dahulu analisis biplot untuk mengetahui seberapa besar analisis biplot dapat merepresentasikan data dalam dimensi berukuran kecil. Pada data laporan keuangan bank bulan Maret 2016, nilai ukuran kesesuaian titik biplot dengan dua peubah adalah 0.7552. Analisis biplot ini cukup merepresentasikan data. Setelah ditentukan ukuran kesesuaian biplot, lakukan kembali analisis bentuk sebaran data. Bentuk sebaran data laporan keuangan syariah ada pada Gambar 3. Gambaran data pada gambar tersebut tidak membentuk suatu kelompok yang memunyai banyak objek yang sama, sehingga sentroid data dapat mewakilkan objek-objek data yang homogen.




Gambar 2 Grafik Data Pencilan Data Kesejahteraan Rakyat Tiap Provinsi

DKI Jakarta dan Papua memiliki gambaran yang cukup curam pada beberapa peubah. Sedangkan Jambi memunyai nilai yang dekat dengan sumbu x sebagai acuan provinsi yang paling dekat dengan sentroid. Peubah yang mendominasi pada DKI Jakarta adalah $X_2$ (kepadatan penduduk), $X_4$ (pendapatan domestik regional bruto), $X_8$ (konsumsi bukan makanan), $X_{24}$ (status kepemilikan rumah sendiri) dan $X_{29}$ (sumber air minum layak). Untuk meningkatkan kesejahteraan diperlukan penurunan jumlah penduduk dan peningkatan persentase kepemilikan rumah sendiri atau menyarankan transmigrasi bagi rakyat yang belum memiliki rumah sendiri. Di sisi lain, pendapatan domestik regional bruto, konsumsi bukan makanan dan ketersediaan air bersih dapat ditingkatkan bagi provinsi lainnya untuk meningkatkan kesejahteraan. Peubah yang mendominasi pada Papua adalah rendahnya tingkat pendidikan yang digambarkan berdasarkan peubah $X_{10}$ (angka partisipasi sekolah usia 7-12 tahun), $X_{11}$(angka partisipasi sekolah usia 13-15 tahun), $X_{12}$ (angka partisipasi sekolah usia 16-18 tahun), $X_{14}$ (angka partisipasi kasar SD), $X_{15}$ (angka partisipasi kasar SMP), $X_{16}$ (angka partisipasi kasar SMA ), $X_{17}$ (angka partisipasi murni SD), $X_{18}$ (angka partisipasi murni SMP) dan $X_{19}$ (angka partisipasi murni SMA). Peubah lainnya yang perlu diperbaiki untuk Papua adalah mengenai perumahan dan lingkungan, yaitu peubah $X_{27}$ (jenis lantai rumah terluas bukan tanah), $X_{28}$ (sumber penerangan listrik PLN), $X_{29}$ (sumber air minum layak).

**Kesimpulan**

Deteksi pencilan dapat dilakukan dengan analisis biplot secara langsung dan tidak langsung serta jarak Euclid langsung dan tidak langsung. Analisis yang paling sederhana dan dengan langkah paling cepat adalah dengan jarak Euclid. Langkah tambahan yang dilakukan analisis biplot adalah dengan mencari terlebih dahulu titik kordinat objek yang berukuran lebih kecil dari data asal agar dapat digambarkan dalam bidang berdimensi rendah. Walau terdapat tambahan langkah, kelebihan analisis biplot adalah dapat memberi gambaran letak pencilan dalam ruang berdimensi dua. Analisis dengan biplot tak langsung dan jarak Euclid tak langsung dinilai lebih baik dibandingkan dengan analisis biplot langsung dan jarak Euclid langsung.

Sebagai acuan untuk metode pendeteksian pencilan dengan biplot dan jarak Euclid, dilakukan pula pendeteksian pencilan dengan DKM dan DKMC.Deteksi dengan biplot tak langsung, jarak Euclid tak langsung, DKM dan DKMC memerlukan matriks homogen sebagai acuan untuk menentukan jarak objek dengan suatu matriks homogen yang mewakilkan data. Penentuan matriks homogen dengan biplot dan jarak Euclid dinilai lebih baik dibandingkan dengan DKM dan DKMC dari segi kecepatan perhitungan data dengan komputer. Pemilihan pencilan oleh biplot dan jarak Euclid pun tak kalah tepat dengan metode DKM dan DKMC.