METODE POHON REGRESI DAN PROSEDUR REGRESI BERTATAR UNTUK SEGMENTASI DATA

BUDI SUHARJO

Departemen Matematika,
Fakultas Matematika dan Imu Pengetahuan Alam,
Institut Pertanian Bogor
Jln. Meranti, Kampus IPB Dramaga, Bogor 16680, Indonesia

Abstrak : Metode regresi berstruktur pohon merupakan suatu suatu metode alternatif untuk segmentasi yang semakin luas penggunaannya. Dalam prosesnya, segmentasi dilakukan melalui proses penyeleksian peubah dan pemilahan terhadap data berdasarkan peubah terpilih. Sementara dalam regresi bertatar, walaupun proses penyeleksian peubah juga dilakukan secara bertahap, namun tidak dilakukan pemilahan terhadap data, sehinggga tidak dapat diperoleh informasi yang berarti untuk melakukan segmentasi data dengan analisis ini. Namun demikian, analisis ini memililki suatu kelebihan; yaitu setiap tahap dari proses penyeleksian peubah dilengkapi dengan suatu persamaan regresi, dimana hal ini penting artinya untuk mengetahui sejauh mana keterkaitan struktural antara peubah-peubah penjelas terhadap peubah respon. Dalam tulisan ini, kelebihan yang dimiliki oleh metode pohon regresi dan analisis regresi bertatar akhirnya dapat digabungkan dalam suatu teknik segmentasi dengan menggunakan prosedur regresi bertatar, dimana metode ini akan memilah data berdasarkan peubah yang muncul pada tahap pertama proses penyeleksian peubah, sehingga metode ini bisa dimanfaatkan untuk segmentasi data sekaligus menemukan hubungan antar peubah. Penerapan metode baru ini untuk melakukan segmentasi terhadap 1002 nasabah bank berdasarkan frekuensi transaksi memberikan hasil yang cukup memuaskan, dimana keakuratan yang dihasilkan pada taraf nyata α=0.05 lebih baik dari segmentasi yang dilakukan dengan metode pohon regresi dengan N_{min} =50. Hal ini merupakan suatu indikasi bahwa metode ini perlu dipertimbangkan sebagai metode alternatif untuk kasus segmentasi.

1. PENDAHULUAN

1.1. Latar Belakang. Segmentasi merupakan suatu upaya untuk mengelompokkan objek-objek pengamatan ke dalam beberapa kelompok, dimana pengamatan yang berada pada satu kelompok umumnya bersifat lebih homogen dibandingkan dengan kelompok yang lain. Teknik segmentasi ini banyak diterapkan pada berbagai bidang seperti dalam bidang riset pemasaran, sehingga muncul istilah yang dikenal dengan segmentasi pasar (market segmentation); yaitu proses pemilahan konsumen, produk, jasa atau industri

yang tujuannya adalah untuk membantu mempelajari prilaku pasar sehingga memudahkan dalam merancang strategi pemasaran seperti positioning dan targeting.

Metode regresi berstruktur pohon (*regression tree*) merupakan suatu teknik yang termasuk dalam kelompok *dependence methods* yang penggunaannya untuk segmentasi akhir-akhir ini semakin luas. Cara kerja dari metode pohon regresi ini hampir serupa dengan prosedur regresi bertatar, dimana proses penyeleksian peubah - peubah penjelas yang berpengaruh terhadap peubah respon dilakukan secara bertahap. Perbedaannya adalah bahwa setiap tahap dari proses penyeleksian peubah pada metode pohon regresi diikuti dengan pemilahan terhadap amatan respon berdasarkan peubah yang muncul pada tahap tersebut. Sedangkan pada analisis regresi bertatar setiap tahap dari proses penyeleksian peubah melibatkan keseluruhan data; tanpa melakukan pemilahan terhadap data berdasarkan peubah yang muncul pada setiap tahapnya, sehinggga tidak dapat diperoleh informasi yang berarti untuk melakukan segmentasi data dengan analisis ini . Namun demikian, analisis ini memililki suatu kelebihan; yaitu setiap tahap dari proses penyeleksian peubah dilengkapi dengan suatu persamaan regresi, dimana hal ini penting artinya untuk mengetahui sejauh mana tingkat keterkaitan struktural antara peubah-peubah penjelas terhadap peubah respon, dan hal ini tidak bisa didapatkan melalui metode pohon regresi. Berdasarkan hal tersebut, perlu dikaji suatu metode yang dapat mengakomodasi kelebihan dari kedua metode ini; segmentasi sekaligus mengetahui struktur keterkaitan antara peubah.

1.2. Tujuan Berdasarkan uraian pendahuluan di atas, maka penelitian ini akan mencoba mengkaji suatu metode dimana nantinya prosedur regresi bertatar dapat diterapkan untuk melakukan segmentasi terhadap data. Selanjutnya akan dilihat sejauh mana keakuratan segmentasi yang dilakukan oleh metode pohon regresi dan prosedur regresi bertatar ini, yang salah satunya dapat diketahui melalui jumlah kuadrat sisaan dari kelompok / segmen yang dihasilkan.

2. METODE POHON REGRESI

Metode pohon regresi adalah salah satu metode yang menggunakan kaidah pohon keputusan (*decision tree*) yang dibentuk melalui suatu algoritma penyekatan secara rekursif. Metode ini diilhami oleh program AID yang dikembangkan oleh Morgan dan Sonquist pada tahun 1960-an. Metode ini menganalisa suatu gugus data dengan cara menyekatnya menjadi beberapa anak gugus (simpul) secara bertahap.

2.1. Algoritma Pohon Regresi. Notasi-notasi yang digunakan dalam pembentukan pohon regresi sama dengan notasi yang digunakan pada regresi biasa dengan p peubah penjelas X_1, X_2, X_3, …, Xp dan satu peubah respon Y. Pembentukan pohon regresi memerlukan empat komponen yaitu:

1. Segugus pertanyaan dikotomous **Q** dengan bentuk "Apakah $\mathbf{x}_i \in \mathbf{A}$?" dengan $\mathbf{x}_i$ merupakan suatu amatan contoh dan $\mathbf{A} \subset X$ (ruang peubah penjelas). Jawaban dari pertanyaan tersebut menentukan sekatan (*partition*) bagi ruang peubah penjelas.
2. Kriteria kebaikan-sekatan $\phi(s,g)$.
3. Aturan penghentian dari proses penyekatan.
4. Statistik ringkasan dari setiap simpul akhir.

2.2. Aturan Penyekatan. Pohon regresi dibentuk melalui penyekatan data pada tiap simpul ke dalam dua simpul anak. Aturannya adalah sebagai berikut:

1. Tiap pemilahan tergantung pada nilai yang hanya berasal dari suatu peubah penjelas.
2. Untuk peubah numerik X_j, pemilahan hanya berasal dari pertanyaan "Apakah $X_j \le c$? " untuk $c \in R$. Jadi, jika ruang contohnya berukuran n dan terdapat sebanyak-banyaknya n nilai amatan yang berbeda pada peubah X_j, maka akan terdapat sebanyak-banyaknya n-1 *split* yang berbeda yang dibentuk oleh gugus pertanyaan {"Apakah $X_j \le c_i$? "}, dengan i=1,2,3,…,n-1 dan c_i adalah nilai tengah antara dua nilai amatan peubah X_j yang berbeda dan berurutan.
3. Untuk peubah penjelas kategori, pemilahan yang terjadi berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua anak gugus yang saling lepas (*disjoint*).

2.3. Tahap Penyekatan. Proses yang dilakukan Breiman et al. (1993) untuk menyekat suatu simpul adalah sebagai berikut:

1. Tentukan semua penyekat yang mungkin untuk setiap peubah penjelas.
2. Pilih sekat yang terbaik dari kumpulan sekat tersebut dan sekat simpul tersebut menjadi dua anak simpul yaitu simpul kiri dan simpul kanan.

Jumlah Kuadrat Sisaan (JKS) akan digunakan sebagai kriteria Kehomogenan di dalam masing-masing simpul. Nilai respon dalam suatu simpul g diduga oleh rataan respon pada simpul g tersebut, yang dihitung sebagai berikut:

$$\bar{y}(g) = \frac{1}{n(g)} \sum_{x_n \in g} y_n$$

maka Jumlah Kuadrat Sisaan di dalam simpul g adalah:

$$JKS(g) = \sum_{x_n \in g} [y_n - \bar{y}(g)]^2$$

Misalkan s menyekat simpul g menjadi simpul kiri g_L dan simpul kanan g_R, maka fungsi penyekatan yang digunakan adalah:

$$\phi(s,g) = R(g) - R(g_L) - R(g_R);$$

dengan R(g) adalah Jumlah Kuadrat Sisaan pada simpul g (JKS(g)). Sekat terbaik s^* adalah sekat yang memenuhi kriteria :

$$\phi(s^*,g) = \max_{s^* \in \Omega} \phi(s,g) ;$$

dimana Ω adalah himpunan semua sekat s yang mungkin pada simpul g. Hal ini berarti bahwa sekat yang dipilih adalah sekat yang mampu menghasilkan penurunan jumlah kuadrat sisaan terbesar.

Algoritma pembentukan struktur pohon ini akan terus dilakukan sampai dipenuhi suatu aturan penghentian tertentu. Kriteria yang sering dijadikan aturan penghentian adalah N_{min}; banyaknya objek pengamatan pada setiap simpul akhir.

2.4. Penentuan Ukuran Pohon. Suatu aspek yang penting pada metode pohon regresi adalah penentuan ukuran pohon yang layak. Suatu upaya yang dapat dilakukan untuk mengatasi hal ini adalah dengan melakukan pemangkasan *(prunning)* terhadap pohon yang terbentuk. Inti dari pemangkasan minimum adalah pemotongan jalur terlemah (*weakest–link*). Untuk sembarang G_g yang merupakan anak cabang dari G_1, didefinisikan:

$$R(G_g) = \sum_{g' \in \tilde{G}_g} R(g') ;$$

dimana $\tilde{G}_g$ adalah gugus simpul akhir dari G_g. Untuk sembarang simpul dalam g dari pohon G_1 berlaku sifat $R(g) > R(G_g)$, dan ukuran biaya kompleksitas dari g didefinisikan sebagai:

$$R_\alpha(\{g\}) = R(g) + \alpha$$

Ukuran biaya kompleksitas dari subpohon G_g adalah:

$$R_\alpha(G_g) = R(G_g) + \alpha \left| \tilde{G}_g \right|$$

Ukuran biaya kompleksitas suatu simpul g akan bernilai sama dengan ukurun kompleksitas pada subpohon G_g bila :

$$\alpha = \frac{R(g) - R(G_g)}{|\tilde{G}_g| - 1}$$

Untuk setiap $g \in G_1$, didefinisikan suatu fungsi $h_1(g)$ sebagai berikut:

$$h_1(g) = \begin{cases} \dfrac{R(g) - R(G_g)}{|\tilde{G}_g| - 1}, & g \notin \tilde{G} \\ +\infty, & g \in \tilde{G} \end{cases}$$

Jalur terlemah dalam G_1 dinotasikan dengan $\bar{g}_1$, adalah simpul yang memenuhi kriteria :

$$h_1(\bar{g}_1) = \min_{g \in G1} h_1(g)$$

Sedangkan nilai parameter kompleksitas α_2 dihitung sebagai berikut:

$$\alpha_2 = h_1(\bar{g}_1).$$

Selanjutnya dibentuk pohon baru dengan cara memangkas cabang dari simpul $\bar{g}_1$, dan pohon baru ini dinamakan G_2. Jadi pohon G_2 diperoleh dengan cara:

$$G_2 = G_1 - G_{\bar{g}_1} ;$$

dimana $G_{\bar{g}_1}$ adalah cabang dari subpohon yang simpul utamanya adalah $\bar{g}_1$. Dengan demikian G_2 adalah pohon yang memenuhi kriteria biaya kompleksitas minimum dengan parameter kompleksitas α_2. Selanjutnya dilakukan lagi proses pemangkasan pada pohon G_2 dengan prosedur seperti di atas, dan seterusnya, sehingga diperoleh sekuens (deretan) pohon yang tersarang dan makin kecil, yaitu $\{G_1, G_2, \ldots \{g_1\}\}$, dimana $G_1 > G_2 > \ldots > \{g_1\}$ dan sekuens

α dalam urutan meningkat, yakni $\{\alpha_1, \alpha_2, \ldots\}$, dimana $\alpha_1 = 0$, $\alpha_2 > \alpha_1$, dan seterusnya.

Langkah terakhir adalah pemilihan pohon terbaik. Dalam pemilihan pohon terbaik ini, digunakan suatu penduga yang dinamakan penduga jujur bagi R(G) (Breiman *et al*. 1993). Penduga jujur yang akan digunakan adalah penduga contoh uji $R^{ts}(G)$ yang didefinisikan sebagai:

$$R^{ts}(G_k) = \frac{1}{n_2} \sum_{(x_i, y_i) \in \ell_2} [y_i - \hat{y}_k(x_i)]^2 ;$$

dimana n_2 adalah ukuran dari *test sample* dimana n_2 adalah ukuran dari *test sample* L_2 dan $\hat{y}_k(x_i)$ adalah dugaan respon dari amatan ke-i pada pohon ke-k. Pohon terbaik adalah G_{k0} yang memenuhi:

$$R^{ts}(G_{k0}) = \min_k R^{ts}(G_k)$$

3. ANALISIS REGRESI BERTATAR

Pembentukan model dengan analisis regresi bertatar dilakukan dengan menyisipkan peubah penjelas satu demi satu. Urutan penyisipan ditentukan dengan menggunakan koefisien korelasi parsial sebagai ukuran kepentingan peubah yang masih berada di luar persamaan.

3.1. Prosedur dasar analisis regresi bertatar

1. Hitung korelasi semua peubah penjelas (X) dengan peubah respon (Y). Peubah yang dipilih pertama kali adalah yang memiliki korelasi tertinggi dengan peubah respon, misalkan X_1.
2. Hitunglah regresi linier ordo pertama $\hat{Y} = f(X_1)$.
3. Lakukan pengujian terhadap koefisien regresi yang terbentuk untuk mengetahui apakah peubah X_1 nyata atau tidak. Jika hasil pengujian ini tidak nyata, proses berhenti dan model terbaik adalah $Y = \bar{Y}$.
4. Jika hasil pengujian nyata, cari peubah penjelas kedua untuk dimasukkan ke dalam model dengan memeriksa koefisien korelasi parsial Y dengan semua peubah penjelas yang berada di luar persamaan regresi, yaitu X_j; $j \neq 1$. Dengan kata lain, Y dan X_j keduanya dikoreksi melalui hubungan garis lurusnya dengan X_1. Misalkan peubah yang terpilih adalah X_2, dan selanjutnya cari persamaan regresi ke dua $\hat{Y} = f(X_1, X_2)$.
5. Setelah itu lakukan lagi pemeriksaan terhadap koefisien regresi, dan nilai-F parsial kedua peubah yang ada di dalam persamaan diuji. Nilai-F parsial terendah dibandingkan dengan nilai-F tabel. Hasil pengujian ini akan menentukan apakah peubah-peubah ini akan dipertahankan atau dikeluarkan dari model.
6. Proses terus berlanjut sampai akhirnya tidak ada lagi peubah yang akan dimasukkan atau dikeluarkan dari model.

4. METODE SEGMENTASI DENGAN PROSEDUR REGRESI BERTATAR

Segmentasi dengan prosedur regresi bertatar dilakukan dengan cara memilahan amatan respon berdasarkan peubah yang muncul pada tahap pertama proses penyeleksian peubah. Ini berarti bahwa amatan Y akan dipilah berdasarkan peubah bebas X yang paling berkorelasi dengan peubah respon Y. Lebih jelasnya, segmentasi data dengan prosedur regresi bertatar ini akan dilakukan dengan metode sebagai berikut:

1. Lakukan analisis regresi bertatar untuk keseluruhan data.
2. Lakukan pemilahan terhadap data berdasarkan peubah yang muncul pada tahap pertama proses penyeleksian peubah.
3. Lakukan kembali analisis regresi bertatar untuk masing-masing kelompok data yang terbentuk pada tahap 2.
4. Lakukan lagi pemilahan berdasarkan peubah yang muncul pada tahap pertama regresi bertatar.
5. Ulangi proses di atas sampai tidak ada lagi peubah yang muncul pada proses regresi bertatar pada taraf nyata α untuk masing - masing kelompok pengamatan yang ada.

Pemilahan suatu gugus data (simpul) berdasarkan peubah yang muncul pada tahap pertama analisis ditujukan untuk menghasilkan dua sub gugus (dua simpul anakan). Namun demikian tidak menutup kemungkinan untuk pemilahan yang menghasilkan lebih dari 2 simpul anakan. Dengan demikian, penentuan titik pemilah dari suatu peubah penyekat terpilih (yang muncul pada tahap pertama analisis regresi bertatar ini) selanjutnya sama dengan yang dilakukan oleh metode pohon regresi, yaitu yang mampu menghasilkan 2 simpul anakan dengan jumlah kuadrat sisaan terkecil.

Segmentasi dengan metode di atas selanjutnya akan dinamakan dengan segmentasi dengan prosedur regresi bertatar. Hasil analisis dengan prosedur regresi bertatar ini dapat pula digambarkan dalam bentuk struktur pohon seperti halnya pada metode *regression tree*, dimana hal ini dapat memberikan gambaran yang jelas tentang pengelompokan amatan respon.

5. PEMBANDINGAN SEGMENTASI ANTARA METODE POHON REGRESI DAN PROSEDUR REGRESI BERTATAR

Terdapat perbedaan antara kedua metode; metode regresi berstruktur pohon dengan metode regresi bertatar, dalam memilih suatu peubah yang akan menyekat sebuah simpul menjadi simpul 2 anakan, dimana metode regresi berstruktur pohon akan memilih peubah yang menghasilkan penurunan JKS terbesar dari simpul awal terhadap 2 simpul anakan yang dihasilkan. Sedangkan prosedur regresi bertatar mendasari pemilihan peubah penyekat pada tingkat korelasi peubah tersebut dengan peubah respon Y, dimana peubah yang terpilih adalah peubah yang tingkat korelasinya dengan Y paling tinggi.

Perbedaan kedua metode dalam menentukan peubah penyekat ini dan perbedaan alat kontrol yang digunakan untuk menghentikan pemilahan sebuah simpul (N_{min} dan taraf nyata α), secara bersama-sama akan menyebabkan terjadinya perbedaan struktur pohon klasifikasi yang terbentuk, dan ini akan mempengaruhi keakuratan dari segmentasi yang dilakukan oleh kedua metode.

6. SEGMENTASI NASABAH BANK BERDASARKAN FREKUENSI TRANSAKSI

Untuk melihat penerapan kedua metode; metode pohon regresi dan prosedur regresi bertatar, dalam melakukan segmentasi terhadap data, akan dilakukan segmentasi terhadap 1002 nasabah bank berdasarkan frekuensi transaksi yang dilakukan dalam sebelas bulan pada tahun 2000. Data nasabah ini berasal dari sebuah bank swasta di Jakarta yang dikumpulkan pada tahun 2000. Peubah - peubah penjelas yang diperhatikan adalah:

1. Jenis Kelamin:
   1. wanita 2. laki-laki

2. Agama:
   1. islam 2. katolik
   3. kristen 4. hindu 5. budha

3. Status pernikahan:
   1. lajang 2. menikah 3. janda/ duda

4. Usia:
   1. $\geq$ 61 tahun 2. 56-60 tahun
   3. 51-55 tahun 4. 46-50 tahun
   5. 41-45 tahun 6. 36-40 tahun
   7. 31-35 tahun 8. 26-30 tahun
   9. 21-25 tahun 10. $\leq$ 20 tahun

5. Pendidikan terakhir:
   1. SD 2. SLTP 3. SLTA
   4. Akademi 5. Universitas

6. Pekerjaan:
   1. Pelajar / mahasiswa
   2. Ibu rumah tangga
   3. Karyawan
   4. Wiraswasta
   5. Profesional
   6. Pensiunan
   7. Lain-lain

7. Saldo rata-rata / bulan

Seperti terlihat di atas bahwa peubah jenis kelamin, agama, status pernikahan, usia, pendidikan terakhir, dan peubah jenis pekerjaan adalah peubah kategori. Untuk keperluan analisis regresi bertatar, maka masing - masing kategori dari peubah ini akan dijadikan suatu peubah boneka (*dummy variable)*.

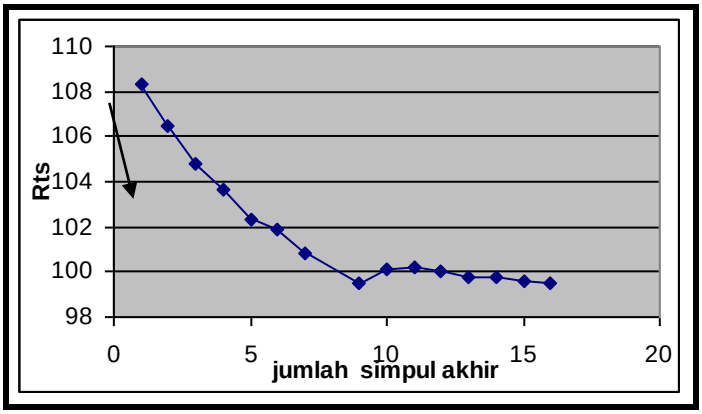
6.1 Segmentasi Nasabah Bank Berdasarkan Frekuensi Transaksi dengan Analisis Pohon Regresi. Analisis pohon regresi (*regression tree*) untuk mengelompokkan 1002 nasabah bank berdasarkan frekuensi transaksi dilakukan

dengan menetapkan banyaknya amatan minimum pada suatu simpul sebanyak 50 amatan. Hal ini berlandaskan pada apa yang dilakukan oleh Schmoor et. *al*. 1993 dimana untuk ukuran contoh 447 ia menetapkan banyaknya amatan pada daun sebanyak 25 (Kudus, 1999).

Pohon awal yang terbentuk terdiri dari 31 simpul, 16 diantaranya adalah simpul akhir. Terbentuknya 16 simpul akhir ini berarti bahwa para nasabah dalam melakukan transaksi dapat dikelompokkan menjadi 16 segmen. Karakteristik dari masing-masing segmen yang terbentuk terlihat dari faktor (peubah) yang muncul sebagai pemisah. Ketujuh peubah penjelas yang dimasukkan dalam analisis muncul sebagai pemisah pada pohon regresi ini.

Pada Gambar 1 terlihat bahwa peubah status pernikahan merupakan peubah pertama yang membedakan nasabah bank dalam hal frekuensi transaksi. Berdasarkan peubah status pernikahan ini terbentuk 2 segmen nasabah bank, yaitu kelompok nasabah yang masih lajang (belum menikah) akan berada pada simpul sebelah kiri (simpul 2), dan kelompok nasabah dengan status pernikahan kawin atau janda / duda akan berada pada simpul sebelah kanan (simpul 3). Secara rata-rata nasabah dengan status pernikahan kawin atau janda / duda lebih sering melakukan transaksi dibandingkan dengan nasabah yang belum menikah. Kelompok nasabah yang belum menikah selanjutnya akan terpilah berdasarkan pendidikan terakhir yang ditamatkan, jenis kelamin, agama, saldo rata-rata per bulan, pekerjaan, dan usia masing-masing nasabah. Sedangkan untuk kelompok nasabah yang sudah pernah menikah (kawin atau janda/ duda) akan terpisah lebih lanjut berdasarkan usia, agama, dan saldo rata-rata per bulan.

6.2. Pemangkasan dan Pemilihan Pohon '*Regression Tree*' Terbaik untuk Segmentasi Nasabah Bank. Pemangkasan yang dilakukan pada pohon utama menghasilkan 15 sub pohon (G_1, …, G_{15}). Pohon terbaik yang dihasilkan dengan penduga contoh uji adalah pohon dengan 9 simpul akhir.



Gambar 2. Plot R^{ts} terhadap jumlah simpul akhir

Simpul akhir yang dimiliki oleh pohon terbaik G_8 adalah simpul 6, 8, 10, 11, 15, 18, 19, 20, dan simpul 21. Struktur pohon terbaik *'regression*

tree' ini dapat dilihat pada Gambar 3. Dari 9 segmen yang terbentuk pada pohon terbaik ini, segmen yang dibentuk oleh simpul 15 memiliki rata-rata frekuensi transaksi paling tinggi. Segmen ini merupakan kelompok nasabah yang sudah menikah dan memiliki saldo rata-rata / bulan lebih dari Rp 65.988,90; serta menganut agama katolik, kristen, atau hindu. Sedangkan simpul 8 merupakan segmen dengan frekuensi transaksi paling rendah.

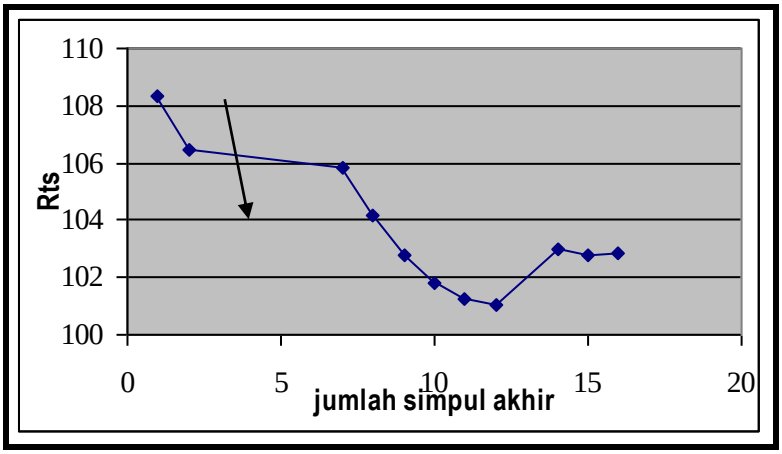
6.3. Segmentasi Nasabah Bank Berdasarkan Frekuensi Transaksi dengan Prosedur Regresi Bertatar. Segmentasi nasabah bank dengan menggunakan prosedur regresi bertatar dilakukan dengan menetapkan taraf nyata α=0.05. Hal ini berarti bahwa pemilahan amatan akan terus berlanjut sampai tidak ada lagi peubah yang muncul untuk suatu regresi bertatar pada taraf nyata α=0.05. Peubah status pernikahan lajang / belum menikah (SP1) muncul pada tahap pertama analisis regresi bertatar yang dilakukan terhadap keseluruhan data. Model lengkap dari regresi bertatar pada keseluruhan nasabah adalah sebagai berikut:

Frek = 9.12 - 3.53 SP1 + 2.9 JK + 3.3 EDU5 -2.42 Ag1

Dari model di atas dapat diketahui bahwa terdapat 4 peubah bebas yang mempengaruhi frekuensi transaksi nasabah secara keseluruhan pada taraf nyata α=0.05, yaitu peubah status pernikahan, jenis kelamin, pendidikan terakhir yang ditamatkan, serta agama yang dianut oleh nasabah. Selanjutnya berdasarkan peubah SP1 ini, yaitu peubah pertama yang muncul pada analisis regresi bertatar untuk keseluruhan data, nasabah bank akan terpilah menjadi dua kelompok, yaitu kelompok nasabah dengan SP1=0 (sudah menikah atau janda / duda) dan kelompok nasabah dengan SP1=1 (nasabah yang masih lajang / belum menikah). Selanjutnya analisis regresi bertatar pada kelompok nasabah yang sudah menikah atau janda/ duda, peubah yang muncul pada tahap pertamanya adalah peubah Ag1, sehingga kelompok nasabah yang sudah menikah atau janda/ duda akan terpilah lagi menjadi dua segmen yaitu kelompok nasabah yang beragama selain islam (Ag1=0) dan kelompok nasabah beragama islam (Ag1=1). Kelompok nasabah yang sudah menikah atau janda/ duda dan menganut agama selain islam (katolik, kristen, hindu, atau budha) masih terpilah lebih lanjut berdasarkan jenis kelamin, pekerjaan, pendidikan terakhir yang ditamatkan, saldo rata-rata/ bulan, dan usia. Sedangkan kelompok nasabah yang sudah menikah atau janda/ duda dan beragama islam tidak terpilah lebih lanjut. Hal ini berarti bahwa tidak ada peubah yang muncul pada analisis regresi bertatar yang dilakukan pada kelompok ini dengan taraf nyata α=0.05. Dengan kata lain bahwa untuk kelompok ini tidak ada satu peubah penjelas pun yang berkorelasi nyata dengan peubah responnya (frekuensi transaksi). Kelompok nasabah yang belum menikah (lajang) selanjutnya terpilah berdasarkan peubah P5, yaitu kelompok nasabah dengan pendidikan akhir bukan di universitas (SD, SLTP, SLTA, atau akademi) dan kelompok nasabah dengan pendidikan akhirnya universitas. Pemilahan kelompok nasabah selanjutnya dapat diperhatikan pada Gambar 4.

6.4. Pemangkasan dan Pemilihan Pohon 'Regresi Bertatar' Terbaik untuk Segmentasi Nasabah Bank

Pemangkasan pada pohon klasifikasi utama "regresi bertatar" menghasilkan 11 sub pohon ($G_1 > G_2 \ldots . > G_{11}$). Pemilihan pohon terbaik yang dilakukan dengan penduga contoh uji menghasilkan pohon terbaik dengan 12 simpul akhir. Simpul akhir yang masuk ke dalam kelompok pohon terbaik ini adalah simpul 7, 8, 10, 11, 16, 17, 18, 21, 26, 27, 28, dan simpul 29. Struktur pohon terbaik untuk segmentasi nasabah bank dengan prosedur regresi bertatar dapat dilihat pada Gambar 6. Dari 12 segmen nasabah yang terbentuk pada pohon terbaik ini, segmen nasabah dengan frekuensi transaksi tertinggi



Gambar 5. Plot R^{ts} terhadap jumlah simpul akhir regresi bertatar

dimiliki oleh segmen yang dibentuk oleh simpul 27 dengan karakteristik nasabah sudah menikah, janda / duda; beragama selain islam (kristen, katolik, hindu, atau budha); jenis kelamin laki-laki; dan bekerja sebagai profesional dengan saldo rata-rata / bulan lebih dari Rp 233.162,00. Nasabah dengan frekuensi transaksi paling rendah dicirikan oleh nasabah yang belum menikah; pendidikan terakhir lebih rendah dari universitas; berjenis kelamin wanita; dengan usia kurang 20 tahun (simpul 17).

6.5. Pembandingan Hasil Segmentasi Nasabah Bank antara Metode *Regression Tree* dengan Metode Regresi Bertatar. Seperti terlihat pada Gambar 1 dan Gambar 4 bahwa pemilahan simpul akar oleh kedua metode sama-sama dilakukan oleh peubah status pernikahan, sehingga menghasilkan 2 simpul anakan, yaitu kelompok yang belum menikah (simpul 2) dan kelompok yang sudah pernah menikah menikah (simpul 3). Berdasarkan metode *regression tree,* hal ini berarti bahwa peubah status pernikahan adalah peubah yang mampu menghasilkan penurunan JKS terbesar dari simpul akar terhadap JKS kedua simpul anakan yang dihasilkan (simpul 2 dan simpul 3), dan memenuhi N_{min} yang dikehendaki, dimana simpul 2 dan simpul 3 masing-masing terdiri dari 530 dan 472 amatan. Sedangkan menurut prosedur regresi bertatar ini berarti bahwa peubah status pernikahan adalah peubah bebas yang paling berkorelasi dengan peubah respon (frekuensi transaksi) bila dibandingkan dengan peubah bebas lainnya, dan korelasi ini nyata pada taraf nyata α=0.05. Pemilahan yang sama juga terjadi pada pemilahan simpul 2, 4, 5, dan simpul 8. Sedangkan untuk pemilahan data selanjutnya terdapat

perbedaan antara kedua metode. Seperti halnya pemilahan simpul 8 yang sama-sama terpilah berdasarkan jenis agama, selanjutnya untuk kelompok beragama selain kristen (islam atau budha) dengan metode *regression tree* terpilah berdasarkan jenis pekerjaan (simpul 16, Gambar 1), sedangkan dengan prosedur regresi bertatar kelompok ini (simpul 14, Gambar 4) tidak terpilah lebih lanjut. Hal ini berarti bahwa untuk kelompok ini tidak ada peubah penjelas yang berkorelasi nyata dengan peubah respon pada taraf nyata α=0.05. Sebaliknya untuk kelompok yang beragama kristen, pada prosedur regresi bertatar (Gambar 4) masih terpilah lebih lanjut berdasarkan jenis pekerjaan, sedangkan dengan metode *regression tree* kelompok ini (simpul 17, Gambar 1) sudah merupakan simpul akhir, yang berarti bahwa tidak ada lagi pemilahan yang mungkin dengan jumlah amatan minimum pada tiap simpul sebanyak 50 amatan.

Perbedaan dalam pemilihan peubah penyekat oleh kedua metode (penurunan jumlah kuadrat sisaan terbesar pada metode pohon regresi dan tingkat korelasi linier pada prosedur regresi bertatar) secara keseluruhan akan menyebabkan perbedaan pada hasil segmentasi yang diperoleh. Pada segmentasi nasabah bank ini, walaupun kedua metode menghasilkan pohon awal yang terdiri dari 16 simpul akhir, namun karakteristik dari simpul-simpul akhir yang dihasilkan berbeda. Hal ini dapat terlihat pada segmen dengan frekuensi transaksi tertinggi, dimana karakteristik dari segmen ini untuk kedua metode tidaklah sama. Segmen nasabah dengan frekuensi transaksi tertinggi untuk pohon awal yang dihasilkan dengan metode pohon regresi dicirikan oleh kelompok nasabah yang sudah menikah atau janda/ duda; beragama katolik, kristen, atau hindu; dengan saldo rata-rata / bulan lebih dari Rp 163.959,00. Anggota dari segmen ini adalah sebanyak 52 amatan dengan frekuensi transaksi rata-rata adalah sebesar 15.1 kali. Sedangkan segmen nasabah dengan frekuensi transaksi tertinggi untuk pohon awal yang dihasilkan dengan prosedur regresi bertatar dicirikan oleh nasabah yang sudah menikah atau janda/ duda; beragama selain islam (katolik, kristen, hindu, atau budha); jenis kelamin laki-laki; bekerja sebagai profesional dengan saldo rata-rata / bulan lebih dari Rp 233.162,00. Anggota dari segmen ini hanya 2 amatan, namun dengan rata-rata frekuensi transaksi yang jauh lebih besar dari yang dihasilkan oleh metode pohon regresi, yaitu sebesar 45 kali. Dari sini terlihat adanya suatu indikasi bahwa penetapan jumlah amatan minimum pada suatu simpul sebagai suatu kriteria penghentian pertumbuhan pohon akan berakibat tidak terdeteksinya amatan berpengaruh yang jumlahnya sedikitl (amatan pencilan), dimana kalau ditetapkan amatan minimum yang lebih besar dari jumlah amatan pencilan ini, baik pencilan di kanan (nilai respon jauh lebih besar dari amatan lain) atau pun pencilan di kiri (nilai respon jauh lebih kecil dari amatan lainnya), akan mengakibatkan amatan-amatan ini tidak bisa membentuk kelompok/ segmen tersendiri sehingga kita tidak bisa mengetahui karakteristik dari amatan-amatan yang mempunyai nilai-nilai ekstrem ini. Selanjutnya hal ini akan berpengaruh pada tingkat kehomogenan dari segmen-segmen yang dihasilkan pada proses segmentasi, karena segmen / kelompok yang terbentuk mengandung nilai-nilai ekstrem, yang apabila dikeluarkan (membentuk segmen tersendiri) sebenarnya dapat meningkatkan kehomogenan dari kelompok tersebut (keragaman dalam kelompok / segmen akan jauh berkurang).

Tabel 1. JKS untuk masing - masing segmen nasabah bank yang terbentuk pada pohon awal

$JKS(i) = \sum_{x_n \in i} [y_n - \bar{y}(i)]^2$				
Simp. akhir	**resi pohon**		**Simp. akhir**	**Regresi bertatar**
10	4480		5	44074
11	10831		12	4480
12	1281		13	10831
13	4615		14	2651
17	2114		17	5109
19	5249		18	2273
22	10385		20	10148
23	14815		22	23
24	713		23	288
25	1260		24	7707
26	1471		25	4100
27	3278		27	365
28	4440		28	21
29	9607		29	8
30	11057		30	782
31	12216		31	30
Total	**97812**		**Total**	**92890**
E(JKS)	**97.62**		**E(JKS)**	**92.71**

Keakuratan pengelompokan yang dilakukan oleh kedua metode lebih jelasnya dapat diketahui dari besarnya nilai harapan kuadrat sisaan, yang dapat diduga melalui total jumlah kuadrat sisaan untuk semua simpul akhir dibagi dengan total amatan keseluruhan. Dari Tabel 1 dapat diketahui bahwa segmentasi nasabah bank dengan prosedur regresi bertatar dengan menggunakan taraf nyata α=0.05 lebih akurat dari segmentasi yang dilakukan dengan metode pohon regresi dengan N_{min}=50, karena nilai harapan jumlah kuadrat sisaan untuk prosedur regresi bertatar lebih kecil dari nilai harapan jumlah kuadrat sisaan yang dihasilkan oleh segmentasi dengan metode pohon regresi dengan N_{min}=50 (92.71 < 97.62). Hal serupa juga terjadi pada struktur pohon terbaik yang dihasilkan, dimana nilai harapan kuadrat sisaan yang dihasilkan prosedur regresi bertatar pada taraf nyata α=0.05 lebih kecil dari nilai harapan jumlah kuadrat sisaan yang dihasilkan metode pohon regresi dengan N_{min}=50 (93.58 < 98.9). Hal ini adalah merupakan suatu indikasi bahwa prosedur regresi bertatar layak dipertimbangkan sebagai suatu metode alternatif untuk segmentasi.

7. KESIMPULAN DAN SARAN

7.1. Kesimpulan. Adalah suatu hal yang sangat riskan dan terlalu awal untuk menyimpulkan bahwa salah satu metode (metode pohon regresi dan prosedur regresi bertatar) lebih baik dari yang lainnya, karena masing-masing memiliki

kelebihan dan kekurangan tersendiri. Namun beberapa hal berikut dapat menjadi pertimbangan bagi penggunaan kedua analisis ini untuk melakukan segmentasi terhadap suatu gugus data.

1. **Tinjauan teoritis**; dimana dalam melakukan segmentasi terhadap segugus data, kedua metode memiliki prosedur yang berbeda, diantaranya dapat dilihat pada:
   - ⇨ **Penentuan peubah yang berpengaruh.** Pada metode pohon regresi, peubah yang paling berpengaruh adalah peubah yang mampu menghasilkan penurunan JKS terbesar dari suatu simpul induk terhadap 2 simpul anakan yang dihasilkan. Sedangkan prosedur regresi bertatar menggunakan koefisien korelasi parsial (linier) sebagai ukuran kepentingan peubah yang masih berada di luar persamaan. Perlu diketahui, bahwa keterkaitan peubah-peubah penjelas dengan peubah respon dalam suatu gugus data pada umumnya tidak saja berbentuk linier, sehinga penggunaan tingkat korelasi linier untuk mendeteksi peubah yang berpengaruh pada prosedur regresi bertatar akan berakibat tidak terdeteksinya peubah yang memiliki hubungan yang tidak linier dengan peubah respon.
   - ⇨ **Aturan yang digunakan untuk menghentikan pemilahan suatu simpul.** Hal ini akan berpengaruh pada jumlah simpul akhir yang akan terbentuk, yang tidak lain adalah jumlah kelompok data yang terbentuk dari proses segmentasi. Jumlah amatan minimum (N_{min}) pada suatu simpul digunakan sebagai kriteria untuk menghentikan pemilahan pada metode pohon regresi. Salah satu kelemahan dari kriteria N_{min} ini adalah kurang bisa mengidentifikasi amatan pencilan (amatan berpengaruh yang jumlahnya lebih kecil dari N_{min} yang ditetapkan), dimana amatan ini tidak bisa membentuk segmen tersendiri sehingga kita tidak bisa mengetahui karakteristik dari amatan-amatan yang mempunyai nilai-nilai ekstrem ini. Selanjutnya hal ini akan berpengaruh pada tingkat kehomogenan dari segmen-segmen yang dihasilkan pada proses segmentasi, karena segmen / kelompok yang terbentuk mengandung nilai-nilai ekstrem, yang apabila dikeluarkan (membentuk segmen tersendiri) dapat meningkatkan kehomogenan dari kelompok tersebut (keragaman dalam kelompok / segmen akan jauh berkurang), yang selanjutnya hal ini jelas akan mempengaruhi ketepatan / keakuratan segmentasi yang diinginkan. Sedangkan prosedur regresi bertatar lebih mendasarkan kriteria pemberhentian pada dimana amatan ini tidak bisa membentuk segmen tersendiri sehingga kita tidak bisa mengetahui karakteristik dari amatan-amatan yang mempunyai nilai-nilai ekstrem ini. Selanjutnya hal ini akan menurunkan tingkat kehomogenan pada segmen-segmen yang dihasilkan oleh proses segmentasi, karena segmen / kelompok yang terbentuk mengandung nilai-nilai ekstrem, yang apabila dikeluarkan (membentuk segmen tersendiri) dapat meningkatkan kehomogenan dari kelompok tersebut (keragaman dalam kelompok / segmen akan jauh berkurang), dan hal ini jelas akan mempengaruhi ketepatan / keakuratan dari segmentasi yang diperoleh. Sedangkan prosedur regresi bertatar lebih mendasarkan kriteria pemberhentian pada besarnya taraf nyata α yang digunakan dalam analisis.

2. Bentuk *outcome* yang dihasilkan

Kedua metode (metode pohon regresi dan prosedur segmentasi dengan regresi bertatar) dapat memberikan informasi mengenai segmentasi terhadap data (amatan respon) berdasarkan peubah-peubah yang mempengaruhinya. Sedangkan informasi tentang sejauh mana tingkat keterkaitan struktural antara peubah penjelas dengan peubah respon hanya bisa diberikan oleh prosedur regresi bertatar, dengan kata lain informasi ini tidak bisa didapatkan melalui metode pohon regresi. Namun Suatu hal yang tidak bisa diabaikan adalah adanya berbagai asumsi yang mendasari pemodelan dengan analisis regresi konvensional termasuk analisis regresi bertatar, sehingga penggunaan persamaan yang dihasilkan pada tiap-tiap tahap untuk pemodelan tetap harus hati - hati.

7.2. Saran. Perlu dilakukan penelitian lebih lanjut untuk menemukan suatu algoritma yang cepat dan fleksibel agar prosedur segmentasi dengan regresi bertatar ini dapat diterapkan pada berbagai data.

DAFTAR PUSTAKA

[1]. Breiman, L., J. H. Friedman, R. A. Olshen dan C. J. Stone. 1993. Classification and Regression Tree. Chapman and Hall, New York.

[2]. Draper, N. dan H. Smith. 1992. *Analisi Regresi Terapan.* Ed. ke-2. Terjemahan Bambang Sumantri. Gramedia, Jakarta.

[3]. Kinnear, T.C. & J.R. Tailor. 1996. *Marketing Research an Applied Approach*. Ed.

[4]. Kudus, A. 1999. *Penerapan Metode Regresi Berstruktur Pohon Pada Pendugaan Masa Rawat Kelahiran Bayi (Studi Kasus di Rumah Sakit Hasan Sadikin Bandung).* Karya Ilmiah S2. Pasca Sarjana IPB. Tidak dipublikasikan.

[5]. Ryan, Thomas. 1997. *Modern Regression Methods.* Jhon Wiley & Sons, Inc., New York.

[6]. Yozza, H. 2000. Analisis Data Longitudina dengan Metode Regresi Berstrukrur Pohon. (Kasus Penyakit Kencing Manis). Karya Ilmiah S2. Pasca Sarjana IPB. Tidak dipublikasikan.