

Algoritma Genetik: Alternatif Metode Penentuan Strata Optimum dalam Perancangan Survei

¹Yusma Yanti, ²Septian Rahardiantoro

¹Program Studi Ilmu Komputer, Universitas Pakuan

²Departemen Statistika, Institut Pertanian Bogor
 yusmayanti.fn@gmail.com

Abstrak

Tujuan dari pembuatan strata ketika pengambilan contoh dalam survei adalah untuk menghasilkan penduga parameter dengan varians kecil, sehingga penentuan alokasi strata perlu diperoleh. Menentukan banyaknya strata dan alokasi elemen strata dari suatu himpunan nilai respon akan menjadi fokus dari penelitian ini. Algoritma Genetik (AG) diaplikasikan untuk kasus ini dengan meminimalkan varians dalam strata pada himpunan yang tersedia, dari jumlah strata 2 sampai 6 strata. Studi empiris melalui simulasi dikembangkan dalam skema populasi yang telah diketahui banyaknya strata sebenarnya, kemudian dengan beberapa jenis banyaknya strata, AG diterapkan dalam data. Berdasarkan hasil simulasi, dapat disimpulkan bahwa AG dapat memberikan banyaknya strata yang sesuai dengan banyaknya strata sebenarnya, sehingga dapat menjadi metode alternatif yang baik untuk memilih banyaknya strata optimal dalam pengambilan contoh survei.

Kata Kunci: algoritma genetik, stratified sample design, perancangan survei

1. Pendahuluan

Perancangan pengambilan contoh dalam survei memiliki tujuan utama untuk memaksimalkan jumlah informasi untuk biaya tertentu [1]. Ketika terdapat keterbatasan biaya sedangkan populasi memiliki karakteristik yang sangat homogen, penarikan contoh acak sederhana (*simple random sampling*) sering memberikan pendugaan yang baik dari karakteristik populasi. Namun, ketika keheterogenan populasi meningkat, dengan biaya tertentu, penarikan contoh acak berstrata (*stratified random sampling*) akan menjadi pilihan terbaik untuk mengambil contoh. Karakteristik heterogenitas populasi ditangani dengan memisahkan populasi menjadi beberapa strata, kelompok *non-overlapping* dari elemen populasi, yang memiliki varians minimum dalam strata.

Berdasarkan alasan tersebut, pendefinisian banyaknya strata dan elemen di setiap strata akan menjadi topik yang diperlukan sebelum kita mengambil contoh menggunakan pengambilan contoh acak berstrata. Hal ini tentunya akan lebih sulit ketika data yang dihadapi dalam skala kualitatif. Secara aplikasi, pilihan umum untuk menentukan banyaknya strata menggunakan pengetahuan subjektif dari peneliti. Namun, dalam populasi tak terhingga, hal ini akan menghasilkan interpretasi yang mungkin keliru.

Metode umum yang sering digunakan adalah dengan menggunakan "akar kuadrat kumulatif dari metode frekuensi" (*cumulative square root of the frequency method*) [1]. Metode ini mengasumsikan bahwa terdapat beberapa data frekuensi yang memiliki korelasi dari variabel yang diamati. Hal ini memiliki beberapa masalah, seperti jika kita memiliki data frekuensi yang berbeda, kita akan memiliki strata yang berbeda, dan juga metode ini dibatasi hanya sampai lima atau enam strata. Oleh karena itu, pada penelitian ini diusulkan sebuah metode alternatif dengan menggunakan metode optimasi metaheuristik, yaitu Algoritma Genetik (AG).

AG merupakan teknik optimasi dan pencarian solusi berdasarkan prinsip-prinsip genetika dan seleksi alam. AG memungkinkan populasi terdiri dari banyak individu yang berkembang di bawah aturan seleksi tertentu [2]. Beberapa individu diambil dari populasi dengan dipilih secara acak untuk memberikan peluang yang sama. Konsep acak juga digunakan dalam kawin silang (*crossover*) dan proses mutasi untuk pembentukan generasi baru. Kemudian, individu terbaik pada generasi terakhir akan menjadi solusi dari algoritma ini [3].

Aplikasi GA dalam penentuan alokasi elemen optimum pada strata sudah banyak dilakukan. Ballin dan Barcarolli pada tahun 2013, menelaah AG sebagai solusi untuk penentuan stratifikasi optimum beserta pengalokasian jumlah contoh optimal untuk data multivariate [4]. Al-Kasab dan Dolmay (2015) memanfaatkan AG dalam penentuan batas stratifikasi dengan asumsi banyaknya strata dan banyaknya total contoh diketahui [5]. Selain itu, Er (2012) melakukan kajian efisiensi metode pengoptimuman (termasuk AG) untuk berbagai nilai banyaknya strata dalam berbagai populasi [6]. Berbeda dengan penelitian-penelitian tersebut, penelitian ini dilakukan dengan mengaplikasikan AG untuk menentukan banyaknya strata yang terdapat pada suatu kerangka populasi yang bersifat kontinu dengan meminimumkan keragaman elemen di dalam strata. Setelah itu, kajian penentuan kisaran selang nilai strata optimum dilakukan berdasarkan banyaknya strata yang terbentuk.

Pada praktiknya, tentunya dengan diasumsikan memiliki barisan data populasi, AG memilih nilai data pada strata yang ditentukan dengan meminimumkan varians dalam strata. Individu dalam AG dapat disajikan oleh bilangan biner juga bilangan real [2]. Penelitian ini akan menjelaskan representasi individu dalam bentuk strata pada AG, dan menggunakan varians dalam strata untuk kriteria pengoptimuman. Selanjutnya, aplikasi menggunakan perangkat lunak R dengan beberapa ilustrasi data simulasi, digunakan untuk menunjukkan bahwa AG bekerja dengan baik untuk tujuan di atas.

2. Metode Penelitian

2.1. Algoritma Genetik sebagai Alternatif Solusi

Misalkan, diketahui data univariat berukuran N yang ingin kita alokasikan pada L strata dengan ($N > L$). Pada algoritma ini, individu dinyatakan sebagai vektor berordo $N \times 1$ yang berisi bilangan 1, 2, ..., L , dengan setiap bilangan merepresentasikan posisi strata. Praktiknya, gen diwakili oleh setiap elemen dari vektor individu, yang juga berarti nilai tunggal dari nomor strata. Selanjutnya, populasi ialah matriks yang mengandung K vektor individu sebagai vektor kolom, sehingga dimensi matriks populasi adalah $N \times K$. Matriks ini disebut generasi pertama (populasi) dalam algoritma genetik.

Pada matriks populasi dilakukan evaluasi setiap vektor individu untuk mencari nilai *fitness* (*fitness value*). Pada penelitian ini, nilai *fitness* didefinisikan sebagai rata-rata dari varians dalam strata, dengan formula

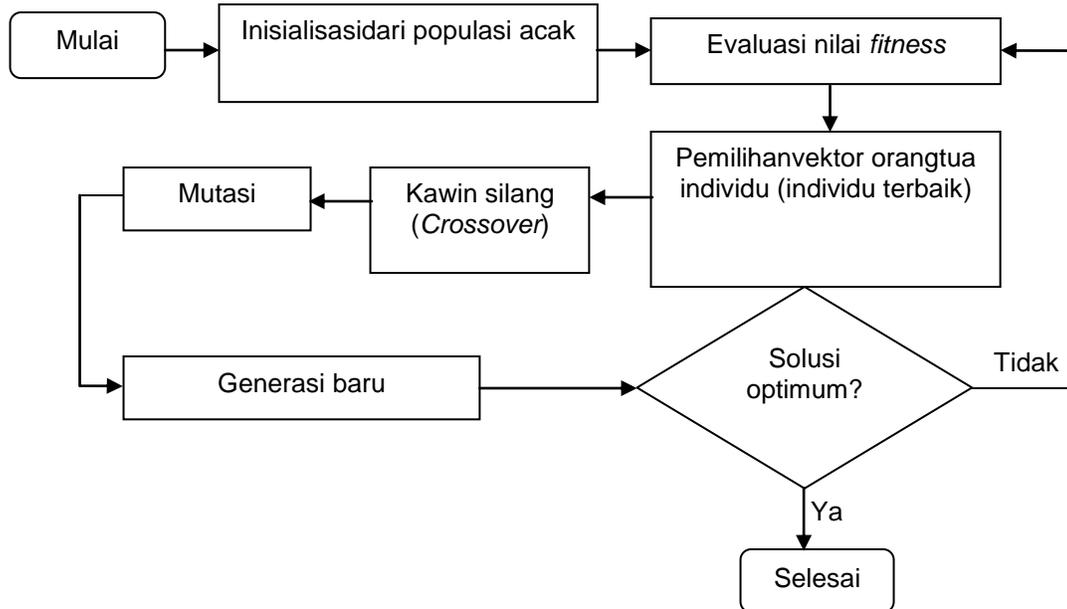
$$Fitness\ value = \frac{1}{L} \sum_{i=1}^L \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{n_i} \quad (1)$$

dengan: L = banyaknya strata, n_i = banyaknya elemen pada strata ke- i , x_{ij} = elemen ke- j dari strata ke- i , dan \bar{x}_i = rata-rata elemen strata ke- i . Proses pengoptimuman terletak pada penentuan strata yang memiliki rata-rata varians terendah dalam strata. Kemudian kita memiliki k vektor individu terbaik yang memiliki nilai *fitness* terbaik, disebut sebagai vektor orangtua individu (*individual parent vectors*).

Proses kawin silang (*crossover*) didefinisikan sebagai perubahan 50% bagian dari satu vektor orangtua individu ke vektor lainnya [7]. Jadi dari proses ini kita mendapatkan hasil kC_2 kawin silang, yang disebut vektor individumuda (*young individual vectors*). Vektor individu muda yang terpilih juga mencakup vektor orangtua, yang selanjutnya akan menjadi generasi baru setelah mengikuti proses mutasi. Mutasi terjadi pada peluang yang rendah (p) untuk proses perubahan gen secara acak.

Proses ini akan diulangi secara terus menerus untuk membentuk generasi baru, dan proses tersebut selesai sampai perbedaan minimum nilai *fitness* antara dua generasi kurang dari suatu nilai α . Akibatnya, vektor individu dengan nilai *fitness* minimum dalam generasi terakhir akan menjadi solusi dari algoritma ini. Gambar 1, menjelaskan secara detail algoritmanya.

Pada penelitian ini, proses algoritma tersebut diterapkan dalam data populasi yang telah diurutkan berdasarkan nilai terendah sampai nilai tertinggi, begitu juga untuk semua generasi yang terbentuk. Hal ini termasuk proses penting untuk memotong jumlah proses iterasi. Akibatnya, solusi dari algoritma ini akan tercapai lebih cepat.



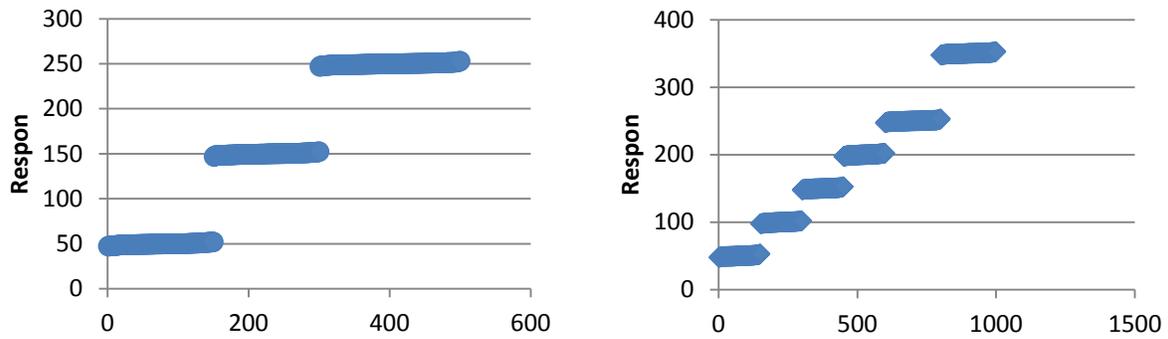
Gambar 1. Proses AG

2.2. Simulasi

Pada penelitian ini, metode penelitian yang digunakan yaitu dengan proses simulasi. Semua proses simulasi dalam penelitian ini menggunakan software R [8]. Data populasi didesain dengan kondisi data univariat dengan beberapa jumlah strata. Simulasi dilakukan dalam dua skema. Pertama, populasi dirancang dengan $N = 500$ dan $L = 3$ strata di dalamnya. Populasi ini mengikuti distribusi normal, yang berisi $S1 \sim N(50,1)$, $S2 \sim N(150,1)$, dan $S3 \sim N(250,1)$, dengan S_i menunjukkan strata ke- i . Pada populasi ini, AG diterapkan pada beberapa strata berbeda, yaitu $L = \{2, 3, 4, 5, 6\}$. Skema ini dirancang untuk menggambarkan populasi yang memiliki jumlah strata yang kecil.

Pada skema kedua, populasi dirancang dengan banyaknya strata yang tinggi, dalam hal ini $L = 6$ dengan $N = 1000$. Populasi ini juga berdistribusi normal, dengan masing-masing strata, $S1 \sim N(50,1)$, $S2 \sim N(100,1)$, $S3 \sim N(150,1)$, $S4 \sim N(200,1)$, $S5 \sim N(250,1)$, dan $S6 \sim N(300,1)$. Dalam populasi ini, AG juga diterapkan dengan ukuran yang berbeda dari strata, $L = \{2, 3, 4, 5, 6\}$.

Gambar 2 menyajikan diagram pencar data populasi dari skema 1 (a) dan skema 2 (b) yang sudah dibangkitkan tanpa melakukan proses pengacakan. Selanjutnya, sebelum AG diaplikasikan, populasi dari skema 1 dan 2 dilakukan proses pengacakan untuk mengkondisikan data sesuai dengan yang di lapangan.



(a) Data Populasi Skema Pertama ($L = 3$)

(b) Data Populasi Skema kedua ($L = 6$)

Gambar 2. Diagram pencar data populasi dua skema simulasi

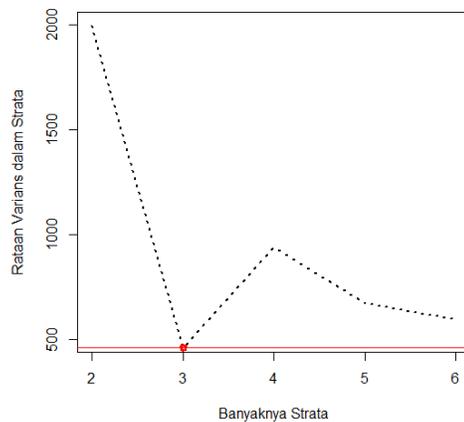
Spesifikasi AG yang diterapkan pada kedua populasi sebagai berikut:

- Jumlah vektor individu dalam generasi pertama : $K = 100$
- Jumlah vektor orangtua : $k = 10$
- Peluang mutasi : $p = 0,3$
- Batas kriteria penyelesaian GA : $\alpha = 0,1$

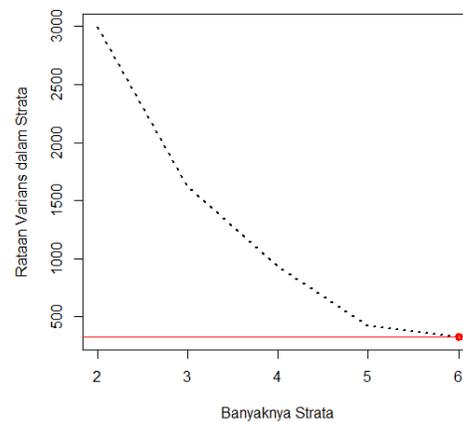
3. Hasil dan Pembahasan

Hasil simulasi dalam dua skema di atas disajikan pada Gambar 3. Proses evaluasi menggunakan rata-rata varians strata dalam solusi AG. Dari gambar 3 (a), aplikasi AG dalam skema pertama yang memberikan nilai terendah dari rata-rata varians dalam strata adalah ketika jumlah strata $L = 3$. Hal ini sesuai dengan jumlah strata awal dalam skema pertama simulasi yang menggambarkan populasi mengandung jumlah strata yang rendah. Meskipun jika dicoba dalam jumlah strata yang banyak, meningkatnya jumlah strata dapat menurunkan rata-rata varians dalam strata. Namun pada hal ini, ketika $L = 3$, AG memberikan nilai sebenarnya dari strata, dan solusinya bisa menjadi alternatif dari preferensi elemen dalam setiap strata.

AG juga bekerja dengan baik dalam skema simulasi keduanya yang menggambarkan populasi dengan strata tinggi. Nilai terendah dari rata-rata varians dalam strata terjadi ketika jumlah strata $L = 6$. Hal ini juga sesuai dengan jumlah strata awal dalam skema kedua. Akhirnya, dari dua skema simulasi ini, AG bisa menjadi metode alternatif pengoptimasian untuk menentukan jumlah strata dan unsur-unsur di setiap strata.



(a) Skema Pertama ($L = 3$)



(b) Skema kedua ($L = 6$)

Gambar3. Grafik evaluasi dua skema simulasi

Selanjutnya, rata-rata, simpangan baku, dan kisaran selang nilai pada data populasi berdasarkan banyaknya strata yang terbentuk untuk kedua skema disajikan pada Tabel 1. Berdasarkan hasil tersebut, dapat dijadikan suatu rekomendasi dalam penentuan banyaknya strata dan kisaran selang nilai untuk setiap strata karena hasilnya sudah sangat dekat kondisi data populasi sebenarnya.

Tabel 1. Rataan dan simpangan baku elemen data populasi pada setiap strata

	Keterangan Strata	Rataan	Simpangan baku	Kisaran selang
Skema 1	Strata 1	49.85	1.02	48.83 s.d. 50.87
	Strata 2	166.80	37.20	129.60 s.d. 204.00
	Strata 3	250.30	0.87	249.43 s.d. 251.17
Skema 2	Strata 1	52.81	11.14	41.67 s.d. 63.95
	Strata 2	104.45	13.92	90.53 s.d. 118.37
	Strata 3	158.66	18.44	140.22 s.d. 177.10
	Strata 4	211.11	20.21	190.90 s.d. 231.32
	Strata 5	260.39	29.89	230.50 s.d. 290.28
	Strata 6	350.20	0.88	349.32 s.d. 351.08

Pada hasil ini, terdapat beberapa hal yang menjadi perhatian. AG bekerja sangat baik ketika populasi yang sesungguhnya diketahui, sehingga dapat ditentukan jumlah strata dan alokasi unsur-unsurnya. Pada penelitian ini, diasumsikan bahwa populasi yang sebenarnya diketahui dan memiliki jumlah strata tertentu melalui simulasi. Namun, pada kenyataannya, data populasi seringkali tidak diketahui. Jadi, terdapat beberapa catatan di sini, bahwa metode ini akan berlaku ketika diketahui data yang menggambarkan populasi dengan baik. Data dapat dikumpulkan sebelum survei dengan melakukan survei pendahuluan.

Oleh karena itu, ketika diketahui data survei pendahuluan, hasil AG dapat diaplikasikan untuk menentukan jumlah strata dan alokasi unsur-unsurnya. Akhirnya, dari hasil ini, survei yang sebenarnya bisa diambil pada populasi menurut nilai kisaran dari masing-masing strata.

4. Kesimpulan

Berdasarkan studi simulasi dapat disimpulkan bahwa AG bisa menjadi metode alternatif yang sangat baik untuk menentukan jumlah optimal dari strata dan alokasi dari unsur-unsur di dalamnya. Hal ini membutuhkan persyaratan bahwa sebelum melaksanakan survei, peneliti harus melakukan survei pendahuluan. Jadi, dengan menggunakan data survei pendahuluan ini, AG akan bekerja dengan sangat baik untuk tujuan tersebut.

5. Daftar Pustaka

- [1] Scheaffer RL, Mendenhall W, Ott RL, Gerow KG. 2012. *Survey Sampling*. Canada: Brooks/Cole Cengage Learning
- [2] Haupt RL, Haupt SE. 2004. *Practical Genetic Algorithms Second Edition*. New Jersey: John Wiley & Sons, Inc

- [3] Rahardiantoro S, Martono T, Sartono B. 2013. *Genetic Algorithms Application for Case Study of Multi-Criteria Decision Analysis (MCDA) on the Data Contained Missing Value*. Proceeding of International Seminar on Sciences (ISS).2013; 1:259-264
- [4] Ballin M, Barcaroli G. 2013. *Joint Determination of Optimal Stratification and Sample Allocation using Genetic Algorithm*. Survey Methodology. Vol 39, 2:369-393
- [5] Al-Kassab MM, Dolmay SS. 2015. *A Genetic Algorithm to Determine the Stratum Boundaries using Proportional Allocation*. BEST: International Journal of Management, InformationTechnology and Engineering (BEST: IJMITE). Vol 3, 7:1:11
- [6] Er S. 2012. *Comparison of the Efficiency of the Various Algorithms in Stratified Sampling when the Initial Solutions are Determined with Geometric Method*. International Journal of Statistics and Applications. Vol 2, 1: 1-10
- [7] Sivanandam SN, Deepa SN. 2008. *Introductions to Genetic Algorithms*. New York: Springer
- [8] Lumley T. 2010. *Complex Surveys: A Guide to Analysis Using R*. New Jersey: John Wiley & Sons, Inc