

PEMBANGKITAN POPULASI SINTETIS KEPENDUDUKAN KOTA CIREBON MENGGUNAKAN HIBRIDA CTGAN DAN ITERATIVE PROPORTIONAL FITTING

Achmad Syaiful Hidayat^{1*}, Marsani Asfi², Petrus Sokibi³

^{1,2,3}Universitas Catur Insan Cendekia, Indonesia

achmad.hidayat.id@gmail.com^{1*},

marsani.asfi@cic.ac.id²,

petrus.sokibi@cic.ac.id³

***corresponding author**

Info Artikel

Pengajuan : 29-07-2026

Peninjauan : 07-08-2026

Revisi : 18-08-2026

Diterima : 25-08-2026

Diterbitkan : 31-08-2026

Kata Kunci:

CTGAN, Iterative Proportional Fitting, Populasi Sintetis, Privasi Data, Data Kependudukan

Abstrak

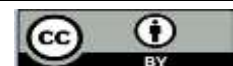
Keterbatasan akses terhadap data mikro kependudukan di Indonesia akibat regulasi perlindungan data pribadi menyulitkan analisis level individu, sementara data agregat publik berisiko kehilangan variasi antarindividu dan menimbulkan ecological fallacy. Penelitian ini mengusulkan pipeline pembangkitan populasi sintetis dengan studi kasus Disdukcapil Kota Cirebon yang mengintegrasikan Conditional Tabular Generative Adversarial Network (CTGAN) dengan Iterative Proportional Fitting (IPF). Indonesia Family Life Survey 5 (IFLS5) digunakan sebagai data mikro referensi berskala nasional untuk mempelajari pola hubungan antarvariabel demografis, dengan keterbatasan bahwa distribusinya tidak secara langsung merepresentasikan karakteristik lokal Kota Cirebon. Secara teoritis, struktur hubungan antarvariabel demografis pada data mikro cenderung lebih stabil lintas wilayah dibandingkan distribusi marginalnya, sehingga IFLS5 tetap dapat menjadi benih awal yang valid sebelum disesuaikan melalui constraint agregat Kota Cirebon pada tahap kalibrasi IPF. Dataset IFLS5 sebanyak 89.382 baris hasil penggabungan disaring menjadi 21.871 baris untuk melatih CTGAN, kemudian benih populasi sintetis dikalibrasi menggunakan data agregat Disdukcapil dan diintegerkan melalui metode Truncate-Replicate-Sample (TRS). Model CTGAN terbaik dipilih pada epoch 200 dari enam variasi epoch yang diuji, berdasarkan keseimbangan kemiripan distribusi dan preservasi hubungan antarvariabel. Setelah kalibrasi IPF, populasi sintetis akhir memperoleh rata-rata Jensen-Shannon Similarity (JSS) 0,9856, Mean Absolute Error (MAE) 0,0062, dan MAE matriks Cramer's V 0,0279, sementara seluruh constraint menghasilkan Weighted Absolute Percentage Error (WAPE) di bawah 10%. Hasil tersebut menunjukkan CTGAN dan IPF saling melengkapi dalam membentuk populasi sintetis yang mempertahankan struktur mikro data referensi sekaligus menyesuaikan distribusi agregat wilayah target, dengan tetap mempertimbangkan keterbatasan representasi lokal dan risiko privasi pada kombinasi atribut bergranularitas tinggi.

How to Cite : Hidayat, A. S., Asfi, M., & Sokibi, P. (2026). Pembangkitan populasi sintetis kependudukan Kota Cirebon menggunakan hibrida CTGAN dan iterative proportional fitting. *Jurnal Rekayasa Sistem Informasi dan Teknologi (JRSIT)*, 4(1), 79–94.

DOI : 10.70248/jrsit.v4i1.4471

This is an open access article under <https://creativecommons.org/licenses/by/4.0/>

Copyright© 2026 by Author. Published by Yayasan Nuraini Ibrahim Mandiri



PENDAHULUAN

Di era digital, data menjadi aset penting dalam pengambilan keputusan dan perencanaan kebijakan, sekaligus meningkatkan perhatian terhadap risiko kebocoran dan penyalahgunaan data pribadi. Beberapa laporan menunjukkan tingginya insiden kebocoran data di Indonesia (Fadliansyah et al., 2025; Putra et al., 2025; Rendi et al., 2026), termasuk pada layanan publik (Soleh & Tjenreng, 2025), sehingga pemerintah

menerapkan Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi sebagai landasan hukum perlindungan data pribadi (Indonesia, 2022).

Implementasi UU PDP turut memengaruhi pengelolaan dan publikasi data kependudukan oleh instansi pemerintah, termasuk Dinas Kependudukan dan Pencatatan Sipil Kota Cirebon. Data yang dipublikasikan umumnya berbentuk agregat, sedangkan akses terhadap data mikro individu dibatasi untuk melindungi data pribadi. Kondisi ini membatasi analisis pada tingkat individu. Data agregat dapat menghilangkan variasi antarindividu dan menimbulkan risiko *ecological fallacy* ketika hubungan pada tingkat kelompok digeneralisasikan ke tingkat individu (Gulo, 2025; Juanuari et al., 2026). Selain itu, agregasi dapat mengubah estimasi dispersi dan hubungan antarvariabel dibandingkan data mikro (Errington et al., 2022).

Pembangkitan data sintesis dapat menjembatani kebutuhan analisis tingkat individu dengan keterbatasan akses terhadap data mikro. Sejumlah penelitian telah mengembangkan populasi sintesis menggunakan model generatif dan kalibrasi statistik pada berbagai konteks (Darsel et al., 2026; Kim & Bansal, 2023; Mwigereri et al., 2025), termasuk analisis kerentanan dan perencanaan wilayah (Gelb et al., 2024). Namun, penggunaan data mikro referensi untuk mempelajari struktur hubungan antarvariabel yang kemudian dikalibrasi menggunakan data agregat resmi pemerintah daerah masih relatif terbatas dalam konteks kependudukan Indonesia pada tingkat kota. Penelitian ini mengisi celah tersebut melalui integrasi CTGAN sebagai pembangkit *seed population* dan IPF sebagai metode kalibrasi terhadap distribusi agregat Kota Cirebon.

CTGAN bukan satu-satunya pendekatan untuk pembangkitan data tabular sintesis. TVAE, CTAB-GAN+, dan Bayesian Network (BN) juga dapat memodelkan distribusi serta ketergantungan antarvariabel (Pathare, 2023; Sun & Erath, 2015; Zhao, 2023), dengan tantangan masing-masing dalam merepresentasikan data tabular (Miletic & Sariyar, 2024). Penelitian ini tidak membandingkan metode tersebut secara empiris sehingga tidak dimaksudkan untuk menyimpulkan keunggulan CTGAN secara umum.

CTGAN dipilih karena dirancang untuk data tabular dengan kombinasi variabel kontinu dan kategorikal serta distribusi kategori yang tidak seimbang. Mekanisme *mode-specific normalization*, *conditional vector*, dan *training-by-sampling* memungkinkan model memperhatikan kategori yang relatif jarang (Xu et al., 2019). Karakteristik tersebut sesuai dengan data penelitian yang mengombinasikan umur dengan variabel jenis kelamin, status perkawinan, tingkat pendidikan, dan aktivitas utama.

Secara praktis, populasi sintesis yang dikalibrasi terhadap data agregat dapat digunakan untuk simulasi perencanaan pemerintah daerah yang membutuhkan karakteristik penduduk pada tingkat individu tanpa menggunakan data individu asli, sebagaimana ditunjukkan pada analisis kerentanan dan perencanaan wilayah (Gelb et al., 2024). Pendekatan ini juga relevan dengan modernisasi layanan administrasi kependudukan di Indonesia, termasuk SIAK (Andika & Agustiono, 2024), layanan elektronik (Ramadhan & Bintari, 2025), dan identitas kependudukan digital (Yulanda & Frinaldi, 2023). Populasi sintesis dapat menjadi lapisan analitis untuk eksplorasi kebijakan, misalnya simulasi kebutuhan fasilitas berdasarkan umur dan kecamatan atau estimasi kebutuhan transportasi, tanpa mengakses maupun mempublikasikan data individu penduduk sesungguhnya.

Penelitian ini mengintegrasikan CTGAN, model GAN untuk data tabular (Sharma et al., 2024; Xu et al., 2019), dengan Iterative Proportional Fitting (IPF), yaitu metode

kalibrasi distribusi berdasarkan *marginal constraints* (Deming & Stephan, 1940) yang telah diterapkan dalam pembangkitan populasi sintetis (Beckman et al., 1996). IFLS5 digunakan sebagai sumber struktur mikro dan pembentukan *seed population*, kemudian disesuaikan melalui *constraint* agregat Disdukcapil Kota Cirebon pada tahap IPF. Dengan demikian, IFLS5 tidak diposisikan sebagai representasi langsung penduduk Kota Cirebon dan penelitian tidak memerlukan data mikro penduduk Kota Cirebon.

Penelitian ini bertujuan untuk: (1) menghasilkan populasi sintetis kependudukan yang sesuai dengan distribusi agregat populasi Kota Cirebon tanpa menggunakan data mikro asli; (2) mengevaluasi kualitas populasi sintetis berdasarkan kemiripan distribusi, preservasi hubungan antarvariabel, dan kesesuaian terhadap *constraint* agregat; serta (3) mengembangkan *pipeline* yang dapat direplikasi pada wilayah lain dengan ketersediaan data mikro referensi dan *constraint* agregat yang sebanding.

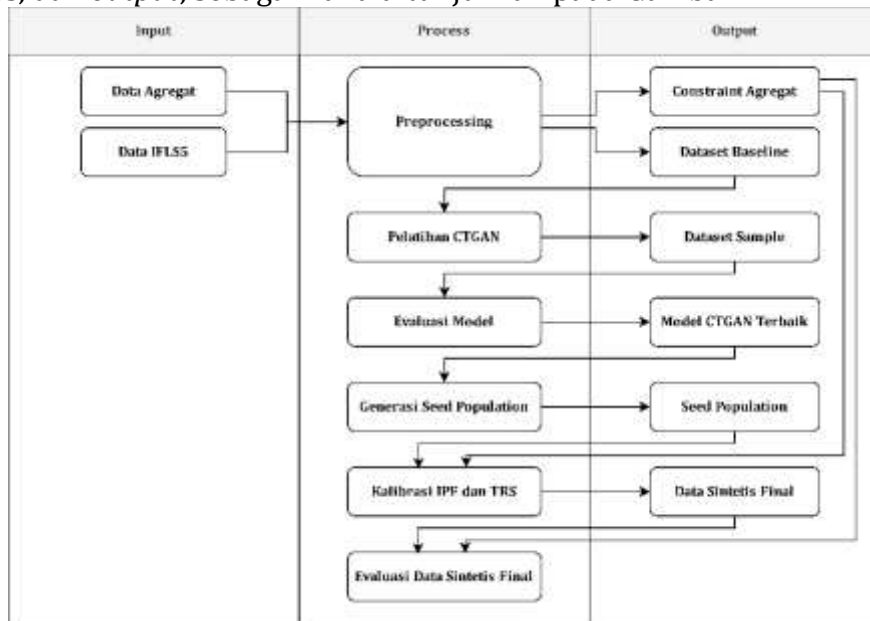
METODE PENELITIAN

Data dan Sumber Data

Penelitian ini menggunakan dua jenis data sekunder. Pertama, *Indonesia Family Life Survey 5*, survei longitudinal berskala nasional yang tersedia untuk penelitian (Corporation, 2016; Strauss et al., 2016), digunakan sebagai data mikro referensi untuk mempelajari pola hubungan antarvariabel demografis dalam pelatihan CTGAN, bukan sebagai representasi langsung populasi Kota Cirebon. Kedua, data agregat Data Konsolidasi Bersih Kota Cirebon Semester II Tahun 2025 dari Disdukcapil Kota Cirebon digunakan sebagai *constraint* kalibrasi, mencakup 274.753 penduduk berusia 15 tahun ke atas pada lima kecamatan. Variabel penelitian meliputi kecamatan, kelompok umur, jenis kelamin, status perkawinan, tingkat pendidikan, dan aktivitas utama.

Kerangka dan Tahapan Penelitian

Penelitian dirancang sebagai *pipeline* yang terdiri atas tiga komponen, yaitu *input*, *proses*, dan *output*, sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Kerangka Pipeline Pembangkitan Populasi Sintetis

Input terdiri atas data mikro IFLS5 dan data agregat Disdukcapil Kota Cirebon. Proses mencakup *preprocessing*, pelatihan dan evaluasi CTGAN, pembentukan synthetic

seed population, kalibrasi IPF, integerisasi TRS, serta evaluasi populasi sintetis final. *Output* berupa *dataset baseline*, model CTGAN terpilih, *seed population*, dan populasi sintetis final yang disesuaikan dengan distribusi agregat Kota Cirebon.

Preprocessing Data Referensi

Tahap *preprocessing* dilakukan untuk menyiapkan *dataset* IFLS5 sebelum digunakan dalam pelatihan model. Tahap ini terdiri atas empat langkah utama: (1) penggabungan (*merge*) beberapa tabel IFLS5 menggunakan identitas responden *hhid14* dan *pid14* sebagai *primary key*; (2) *filtering* untuk memilih responden berusia 15 tahun ke atas serta variabel penelitian yang relevan; (3) harmonisasi kategori variabel jenis kelamin, status perkawinan, tingkat pendidikan, dan aktivitas utama agar seragam dengan skema penelitian; dan (4) *stochastic imputation* untuk mengisi nilai hilang pada variabel tingkat pendidikan dan aktivitas utama. *Dataset* hasil *preprocessing* selanjutnya divalidasi menggunakan aturan logis (*rule-based validation*) sebelum digunakan sebagai *dataset baseline* untuk pelatihan model.

Secara formal, *stochastic imputation* dilakukan dengan mengambil sampel acak dari distribusi probabilitas kategori berdasarkan karakteristik demografis yang tersedia dan tidak memiliki *missing value*, yaitu umur, jenis kelamin, dan status perkawinan. Untuk variabel tingkat pendidikan E , probabilitas kategori ke- k untuk observasi ke- i dihitung berdasarkan strata umur dan jenis kelamin:

$$p_{ik} = P(E = k | U_i, S_i)$$

sedangkan untuk variabel aktivitas utama A , probabilitas kategori dihitung berdasarkan umur, jenis kelamin, dan status perkawinan:

$$p_{ik} = P(A = k | U_i, S_i, M_i)$$

dengan U menyatakan umur, S jenis kelamin, dan M status perkawinan. Probabilitas empiris tersebut dihitung dari observasi *complete case* pada strata yang sesuai dan diberikan *Laplace smoothing* dengan $\alpha = 0,5$ untuk menghindari probabilitas nol pada kategori yang belum muncul dalam strata. Secara umum, probabilitas kategori ke- k dihitung sebagai:

$$p_{ik} = \frac{[n_k(x_i) + \alpha]}{\sum_j [n_j(x_i) + \alpha]}$$

dengan $n_k(x_i)$ jumlah observasi lengkap pada kategori k dalam strata x_i , K jumlah kategori yang diperbolehkan, dan α parameter *Laplace smoothing*. Nilai hilang diisi melalui pengambilan sampel dari distribusi kategorikal tersebut menggunakan probabilitas p_{ik} , bukan penetapan kategori deterministik, mengikuti prinsip umum *stochastic/multiple imputation* yang mempertahankan variabilitas data (R. J. A. Little & Rubin, 2019; van Buuren, 2018), meski tetap mengandung asumsi bahwa hubungan antara variabel target dan ketiga variabel acuan cukup merepresentasikan pola nilai hilang, sehingga hasil imputasi tidak dapat diklaim sepenuhnya bebas bias.

Apabila tidak terdapat observasi lengkap pada strata yang sesuai, distribusi marginal seluruh observasi lengkap digunakan sebagai *fallback*. Pendekatan ini mempertahankan unsur acak dalam imputasi sekaligus mempertimbangkan struktur demografis dan batasan logis antarvariabel.

Pelatihan dan Evaluasi Model CTGAN

CTGAN merupakan model generatif berbasis *deep learning* yang terdiri atas jaringan *generator* dan *discriminator* yang dilatih secara *adversarial* (Xu et al., 2019). Model ini menggunakan *mode-specific normalization*, *conditional vector*, dan *training-by-*

sampling untuk menangani data tabular dengan kombinasi variabel numerik dan kategorikal serta distribusi kategori yang tidak seimbang (Xu et al., 2019). Pada penelitian ini, CTGAN dilatih menggunakan pustaka Synthetic Data Vault (SDV) dengan akselerasi GPU. *Hyperparameter* selain *epoch* ditetapkan, yaitu *embedding dimension* 128, *generator dimension* dan *discriminator dimension* (256,256), *learning rate* masing-masing 0,0002, *batch size* 500, dan PAC 10. Variasi pelatihan terdiri atas 50, 100, 150, 200, 250, dan 300 *epoch* untuk mengamati pengaruh durasi pelatihan terhadap kualitas data sintetis.

PAC 10 dipertahankan karena mekanisme *packing* membantu *discriminator* membedakan data nyata dan sintetis berdasarkan beberapa sampel secara bersamaan (Lin et al., 2018). Eksplorasi kombinasi *hyperparameter* lain melalui *grid search* atau *random search* tidak dilakukan karena akan meningkatkan jumlah eksperimen. Penelitian ini berfokus pada variasi *epoch*, sedangkan *hyperparameter tuning* yang lebih luas menjadi keterbatasan dan peluang penelitian lanjutan.

Kualitas model dievaluasi menggunakan empat metrik, yaitu Jensen-Shannon Similarity (JSS) untuk mengukur kemiripan distribusi marginal (Endres & Schindelin, 2003); Mean Absolute Error (MAE) untuk mengukur penyimpangan proporsi antarkategori (Willmott & Matsuura, 2005); MAE matriks Cramér's V untuk mengukur preservasi asosiasi antarvariabel kategorikal (Bergsma, 2013); dan Distance to Closest Record (DCR) untuk mengukur kedekatan setiap rekam sintetis terhadap rekam referensi terdekat (Yao et al., 2025). DCR digunakan sebagai indikator tambahan kedekatan data sintetis terhadap data referensi, bukan sebagai ukuran tunggal risiko privasi.

Standar Interpretasi Metrik

Penelitian ini menggunakan standar interpretasi untuk membantu pembacaan hasil evaluasi. Nilai JSS mendekati 1 menunjukkan kemiripan distribusi yang semakin tinggi (Pathare, 2023), sedangkan WAPE <10% digunakan sebagai indikasi akurasi sangat baik berdasarkan kriteria ketepatan peramalan berbasis persentase (Hyndman & Koehler, 2006; Lewis, 1982). MAE diinterpretasikan berdasarkan sifat metrik, yaitu semakin mendekati 0 menunjukkan penyimpangan yang semakin kecil. Karena belum terdapat ambang baku yang disepakati luas untuk MAE pada evaluasi data mikro sintetis (Voas & Williamson, 2001), interpretasinya tidak ditetapkan sebagai standar universal.

Tabel 1. Standar Interpretasi Metrik Evaluasi

No	Metrik	Rentang Nilai	Kategori	Sumber
1	JSS	Mendekati 1	Kemiripan distribusi semakin tinggi	Berdasarkan sifat metrik
2	MAE	Mendekati 0	Penyimpangan distribusi semakin kecil	Berdasarkan sifat metrik
3	WAPE	<10%	Sangat baik (Highly Accurate)	Adaptasi Lewis
4	WAPE	10-20%	Baik (Good)	Adaptasi Lewis
5	WAPE	20-50%	Cukup (Reasonable)	Adaptasi Lewis
6	WAPE	>50%	Kurang baik (Inaccurate)	Adaptasi Lewis

Kalibrasi Iterative Proportional Fitting dan Integerisasi

Benih populasi sintetis dibangkitkan menggunakan model CTGAN terbaik dengan oversampling lima kali populasi target. Karena Kecamatan tidak tersedia pada IFLS5, atribut tersebut diinjeksi secara stokastik berdasarkan distribusi agregat Kecamatan menurut kelompok umur dan jenis kelamin dari data Disdukcapil. Untuk kombinasi

kelompok umur a dan jenis kelamin s , probabilitas pemilihan Kecamatan c dihitung sebagai:

$$P(K = c|A = a, S = s) = \frac{N_{c,a,s}}{\sum_{c'} N_{c',a,s}}$$

dengan $N_{c,a,s}$ merupakan jumlah penduduk pada Kecamatan c , kelompok umur a , dan jenis kelamin s . Kecamatan kemudian dipilih secara acak berdasarkan probabilitas tersebut dan setiap individu diberi *initial weight* sebesar satu.

Pendekatan ini membatasi informasi lokasi pada hubungan kondisional Kecamatan–umur–jenis kelamin yang tersedia dalam data agregat dan tidak mengasumsikan independensi Kecamatan terhadap seluruh karakteristik demografis. Hubungan Kecamatan dengan pendidikan dan aktivitas utama tidak dimodelkan secara langsung karena data agregat tidak menyediakan *cross-constraint* untuk kedua variabel tersebut. Kondisi ini merupakan simplifikasi yang berpotensi menimbulkan bias apabila karakteristik tersebut berbeda secara sistematis antarwilayah. Untuk mengurangi keterbatasan tersebut, hubungan Umur $\times$ Jenis Kelamin $\times$ Kecamatan digunakan sebagai *constraint* IPF. Dengan demikian, Kecamatan merupakan atribut yang diinjeksi dari informasi agregat, bukan informasi mikro dari IFLS5.

IPF menyesuaikan bobot individu sintetis secara iteratif agar distribusinya mendekati target (Deming & Stephan, 1940). Faktor penyesuaian setiap kelompok dihitung menggunakan $F = \frac{T}{E}$, dengan T sebagai target dan E sebagai jumlah bobot saat ini, kemudian bobot diperbarui menggunakan:

$$w_i^{(t+1)} = w_i^{(t)} F$$

Tiga *constraint* diterapkan secara berurutan pada setiap iterasi, yaitu Umur $\times$ Aktivitas Utama $\times$ Jenis Kelamin, Umur $\times$ Jenis Kelamin $\times$ Kecamatan, dan Status Perkawinan. Iterasi dihentikan ketika rata-rata perubahan faktor penyesuaian antarkelompok $< 0,001$ (0,1%) atau mencapai maksimum 1.000 iterasi. Toleransi digunakan untuk menyeimbangkan kesesuaian terhadap target dan biaya komputasi, sedangkan batas iterasi mencegah proses berjalan tanpa konvergensi.

Setelah konvergen, bobot diintegerisasi menggunakan metode Truncate-Replicate-Sample (TRS) (Lovelace & Ballas, 2013). Bagian bulat setiap bobot digunakan sebagai jumlah replikasi awal, sedangkan bagian pecahan menjadi *residual*. Sisa populasi kemudian dialokasikan melalui *sampling* tanpa pengembalian dari rekam dengan *residual* positif dengan probabilitas sebanding dengan *residual*. Pemilihan metode *rounding* dapat memengaruhi akurasi populasi sintetis dibandingkan metode lain seperti *simulated annealing* (Duran-Heras et al., 2018). Hasil akhirnya berupa populasi sintetis dengan bobot setiap rekam bernilai satu dan jumlah observasi sama dengan populasi target.

Evaluasi Populasi Sintetis Final

Populasi sintetis akhir dievaluasi berdasarkan empat aspek: (1) kemiripan distribusi terhadap populasi benih menggunakan JSS dan MAE; (2) preservasi hubungan antarvariabel menggunakan MAE matriks Cramér's V; (3) *constraint accuracy*, yaitu kesesuaian terhadap data agregat menggunakan MAE, RMSE, Weighted Absolute Percentage Error (WAPE), dan Maximum Error; serta (4) *small-cell disclosure*, yaitu proporsi sel pada tabel *constraint* dengan jumlah populasi di bawah 5 dan 10 individu. Pemeriksaan *small-cell* digunakan sebagai indikator awal sel berfrekuensi rendah yang berpotensi meningkatkan risiko pengungkapan ketika dikombinasikan dengan

karakteristik lain, termasuk Kecamatan, dengan mengacu pada prinsip *k-anonymity* (Sweeney, 2002) dan *statistical disclosure control* (Willenborg & Waal, 2001).

Evaluasi privasi juga mempertimbangkan *membership inference attack* (MIA), yaitu skenario ketika penyerang mencoba menentukan apakah suatu rekam individu digunakan dalam data pelatihan model (Shokri et al., 2017), termasuk pada model pembangkit data tabular (Hyeong et al., 2022; Zhang et al., 2022). Analisis MIA pada penelitian ini dilakukan secara kualitatif berdasarkan struktur ruang kombinasi kategorikal data referensi, tanpa simulasi serangan berbasis *shadow model* (Miura et al., 2024; van Breugel et al., 2023). Oleh karena itu, hasil evaluasi privasi bersifat indikatif dan bukan jaminan anonimitas formal. Pengujian MIA empiris dan evaluasi *disclosure risk* yang lebih komprehensif menjadi peluang penelitian lanjutan.

HASIL PENELITIAN DAN PEMBAHASAN

Hasil Preprocessing Data Referensi

Proses merge beberapa tabel IFLS5 menghasilkan 89.382 baris data terintegrasi. Setelah *filtering* berdasarkan usia dan variabel penelitian, diperoleh 21.871 observasi sebagai *dataset* referensi. Stochastic imputation mengisi 2.121 nilai hilang pada tingkat pendidikan dan 1.474 pada aktivitas utama tanpa mengurangi jumlah observasi, sebagaimana ditunjukkan pada Tabel 2.

Tabel 2. Hasil Stochastic Imputation pada Dataset Referensi

No	Variabel	Sebelum Imputasi	Sesudah Imputasi
1	Tingkat Pendidikan	2.121	0
2	Aktivitas Utama	1.474	0
	Jumlah	3.595	0

Validasi rule-based menunjukkan tidak terdapat *missing value* maupun pelanggaran aturan logis terkait status perkawinan, status pekerjaan, tingkat pendidikan, dan aktivitas utama. Evaluasi distribusi menggunakan Total Variation Distance (TVD) menghasilkan 0,0489 untuk tingkat pendidikan dan 0,0072 untuk aktivitas utama, menunjukkan perubahan distribusi yang relatif kecil setelah imputasi. Dataset *baseline* sebanyak 21.871 observasi selanjutnya digunakan sebagai data pelatihan CTGAN.

Pelatihan dan Evaluasi Model CTGAN

Model CTGAN dilatih pada enam konfigurasi jumlah *epoch*, yaitu 50, 100, 150, 200, 250, dan 300 *epoch*. Setiap konfigurasi digunakan untuk membangkitkan data sintetis yang kemudian dievaluasi menggunakan Jensen-Shannon Similarity (JSS), Mean Absolute Error (MAE), MAE matriks Cramer's V, dan Distance to Closest Record (DCR). Hasil evaluasi disajikan pada Tabel 3.

Tabel 3. Hasil Pelatihan Model CTGAN pada Berbagai Variasi Epoch

No	Epoch	Mean JSS	Mean MAE	Mean Cramer's V MAE	Mean DCR
1	50	0,8819	0,0346	0,1019	0,0253
2	100	0,8740	0,0351	0,0937	0,0254
3	150	0,8927	0,0265	0,1262	0,0186
4	200	0,9070	0,0236	0,0542	0,0179
5	250	0,9057	0,0182	0,0933	0,0176
6	300	0,9097	0,0207	0,0551	0,0185

Penambahan jumlah *epoch* menunjukkan perubahan pada seluruh metrik, tetapi tidak konsisten pada semua aspek evaluasi. JSS meningkat dari 0,8819 pada *epoch* 50 menjadi 0,9097 pada *epoch* 300, sedangkan MAE terendah diperoleh pada *epoch* 250

sebesar 0,0182. Nilai Cramer's V MAE terendah diperoleh pada *epoch* 200 sebesar 0,0542, sementara DCR relatif berdekatan pada konfigurasi 200–300 *epoch*.

Epoch 300 menghasilkan JSS tertinggi, sedangkan *epoch* 250 menghasilkan MAE terendah. Namun, *epoch* 200 memberikan Cramer's V MAE terendah dan DCR yang tetap kompetitif. Dengan mempertimbangkan keseimbangan antara kemiripan distribusi, preservasi hubungan antarvariabel, kedekatan terhadap data referensi, dan efisiensi komputasi, *epoch* 200 dipilih sebagai model CTGAN untuk pembangkitan *seed population*. Model terpilih menghasilkan JSS sebesar 0,9070, MAE sebesar 0,0236, Cramer's V MAE sebesar 0,0542, dan DCR sebesar 0,0179. Hasil tersebut menunjukkan bahwa model menghasilkan data sintetis dengan kemiripan distribusi dan hubungan antarvariabel yang relatif baik terhadap data referensi, sedangkan DCR digunakan sebagai indikator tambahan untuk menilai kedekatan data sintetis terhadap data referensi.

Tabel 4. Nilai Jensen-Shannon Similarity (JSS) per Variabel

No	Variabel	Nilai JSS
1	Jenis Kelamin	0.9805
2	Aktivitas Utama	0.9041
3	Status Perkawinan	0.9005
4	Tingkat Pendidikan	0.8430

Berdasarkan Tabel 4, variabel jenis kelamin memperoleh JSS tertinggi sebesar 0,9805, sedangkan tingkat pendidikan memperoleh nilai terendah sebesar 0,8430. Hasil ini menunjukkan bahwa distribusi marginal variabel penelitian memiliki tingkat kemiripan yang baik, meskipun tingkat pendidikan memiliki kemiripan relatif lebih rendah dibandingkan variabel dengan kategori yang lebih sederhana seperti jenis kelamin.

Tabel 5. Nilai Mean Absolute Error (MAE) per Variabel

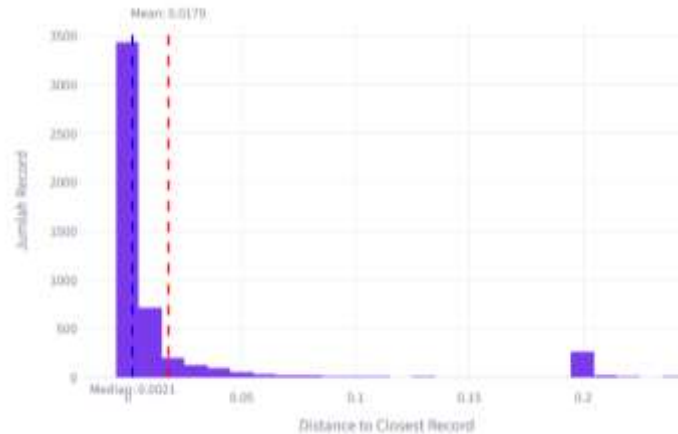
No	Variabel	Nilai MAE
1	Umur	0.0029
2	Aktivitas Utama	0.0127
3	Jenis Kelamin	0.0229
4	Status Perkawinan	0.0375
5	Tingkat Pendidikan	0.0418

Tabel 5 menunjukkan variabel umur memiliki MAE terkecil sebesar 0,0029, sedangkan tingkat pendidikan memiliki MAE terbesar sebesar 0,0418. Nilai rata-rata MAE sebesar 0,0236 menunjukkan bahwa penyimpangan proporsi antarkategori secara keseluruhan relatif rendah.

Evaluasi hubungan antarvariabel menggunakan matriks Cramer's V menunjukkan MAE sebesar 0,0542 (Tabel 6). Selisih terbesar terdapat pada pasangan jenis kelamin-tingkat pendidikan sebesar 0,0767 dan status perkawinan-tingkat pendidikan sebesar 0,0703. Hasil tersebut menunjukkan bahwa perbedaan asosiasi antarvariabel antara data sintetis dan data referensi relatif kecil, dengan MAE sebesar 0,0542.

Tabel 6. Pasangan Variabel dengan Selisih Korelasi Terbesar

No	Pasangan Variabel	Selisih Cramer's V
1	Jenis Kelamin - Tingkat Pendidikan	0,0767
2	Status Perkawinan - Tingkat Pendidikan	0,0703
3	Jenis Kelamin - Status Perkawinan	0,0670
4	Tingkat Pendidikan - Aktivitas Utama	0,0501
5	Status Perkawinan - Aktivitas Utama	0,0477



Gambar 2. Distribusi Distance to Closest Record (DCR)

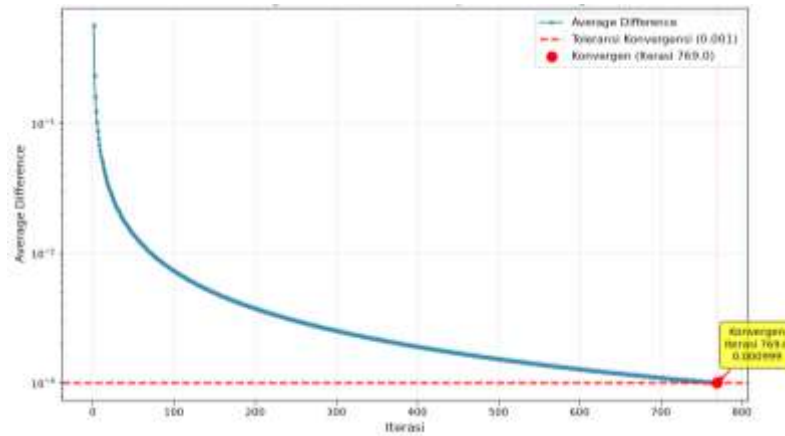
Histogram DCR pada Gambar 2 menunjukkan sebagian besar rekam sintesis memiliki jarak relatif kecil terhadap rekam referensi terdekat, dengan median 0,0021, rata-rata 0,0179, dan maksimum 0,2442. Distribusi yang condong ke kanan menunjukkan bahwa sebagian besar rekam memiliki kedekatan relatif tinggi terhadap data referensi, sementara sebagian kecil memiliki jarak lebih besar.

Evaluasi tambahan menghasilkan *exact match rate* sebesar 64,75% dan *novelty* sebesar 57,31%, dengan 8.749 kombinasi atribut unik pada data sintesis dibandingkan 6.380 kombinasi pada data *baseline*. Tingginya *exact match rate* dapat dipengaruhi oleh terbatasnya ruang kombinasi kategorikal pada data referensi. Kemunculan kombinasi atribut yang identik tidak dengan sendirinya menunjukkan bahwa model menghafal rekam individu, tetapi juga bukan bukti bahwa data sepenuhnya aman dari risiko privasi.

DCR, *exact match rate*, dan *novelty* belum cukup menggambarkan risiko privasi secara menyeluruh. Risiko tersebut dapat mencakup *re-identification*, *membership inference*, *attribute disclosure* (Reiter, 2005), dan *small-cell disclosure* (Sweeney, 2002). Pada penelitian ini, MIA belum diuji secara empiris melalui *shadow model* (Miura et al., 2024; van Breugel et al., 2023), dan *differential privacy* formal juga belum diterapkan. Oleh karena itu, evaluasi privasi bersifat indikatif dan tidak dimaksudkan sebagai jaminan keamanan atau anonimitas formal. Pengujian MIA empiris dan *differential privacy* diusulkan sebagai penelitian lanjutan sebelum populasi sintesis dipublikasikan secara terbuka. Penelitian ini hanya menghasilkan populasi sintesis dan tidak mempublikasikan rekam asli IFLS5.

Kalibrasi IPF dan Integerisasi TRS

Benih populasi sintesis dibangkitkan sebanyak 1.373.765 *record*, yaitu lima kali populasi target sebesar 274.753 jiwa, untuk menyediakan jumlah rekam yang memadai sebelum proses kalibrasi. Proses IPF mencapai konvergensi pada iterasi ke-769 dengan nilai *average difference* sebesar 0,000999, di bawah batas toleransi 0,001 (Gambar 3). Hasil tersebut menunjukkan bahwa proses pembobotan telah mencapai kriteria konvergensi yang ditetapkan terhadap *constraint* kalibrasi.



Gambar 3. Kurva Konvergensi Kalibrasi IPF

Setelah kalibrasi, dilakukan integerisasi menggunakan metode Truncate-Replicate-Sample (TRS). Tahap *truncate* menghasilkan 44.963 individu, sedangkan *residual* sebesar 229.790 dialokasikan melalui tahap *sample* tanpa pengembalian. Proses ini menghasilkan populasi sintetis akhir sebanyak 274.753 individu, sesuai dengan populasi target, dengan seluruh individu memiliki bobot satu, sebagaimana dirangkum pada Tabel 7.

Tabel 7. Hasil Implementasi Integerisasi TRS

No	Parameter	Nilai
1	Total weight awal	274.753
2	Individu hasil Truncate	44.963
3	Residual dialokasikan (Sample)	229.790
4	Jumlah populasi akhir	274.753
5	Selisih terhadap target	0

Evaluasi Populasi Sintetis Final

Populasi sintetis akhir dibandingkan terhadap benih populasi (*seed population*) untuk menilai kesamaan distribusi setelah kalibrasi. Tabel 8 menunjukkan rata-rata JSS meningkat menjadi 0,9856 dan rata-rata MAE menurun menjadi 0,0062, dengan status perkawinan memperoleh JSS tertinggi (0,9934).

Tabel 8. Kesamaan Distribusi Populasi Sintetis Final terhadap Seed Population

No	Variabel	JSS	MAE
1	Jenis Kelamin	0,9890	0,0128
2	Status Perkawinan	0,9934	0,0035
3	Aktivitas Utama	0,9742	0,0022
	Rata-rata	0,9856	0,0062

JENIS_KELAMIN	0.000	0.039	0.002	0.153	0.003
STATUS_PERKAWINAN	0.039	0.000	0.008	0.023	0.006
TINGKAT_PENDIDIKAN	0.002	0.008	0.000	0.039	0.002
AKTIVITAS_UTAMA	0.153	0.023	0.039	0.000	0.005
KECAMATAN	0.003	0.006	0.002	0.005	0.000

JENIS_KELAMIN
STATUS_PERKAWINAN
TINGKAT_PENDIDIKAN
AKTIVITAS_UTAMA
KECAMATAN

Gambar 4 Preservasi Hubungan Antarvariabel pada Populasi Sintetis Final

Evaluasi hubungan antarvariabel menggunakan MAE matriks Cramér's V menghasilkan nilai sebesar 0,0279. Pada Gambar 4, selisih Cramér's V terbesar ditemukan pada hubungan Jenis Kelamin × Aktivitas Utama sebesar 0,15, sejalan dengan penerapan constraint Umur × Aktivitas Utama × Jenis Kelamin pada IPF. Evaluasi *constraint accuracy* pada Tabel 9 menunjukkan seluruh *constraint* memiliki WAPE di bawah 10% atau kategori sangat baik berdasarkan kriteria yang digunakan (Lewis, 1982). WAPE terendah terdapat pada Status Perkawinan (1,43%), sedangkan tertinggi pada Umur × Aktivitas Utama × Jenis Kelamin (8,24%). Dengan demikian, seluruh *constraint* yang digunakan berada pada kategori sangat baik berdasarkan standar interpretasi tersebut.

Tabel 9. Constraint Accuracy Populasi Sintetis Final

No	Constraint	MAE	RMSE	WAPE	Max Error
1	Umur × Aktivitas × Jenis Kelamin	58,08	156,80	8,24%	1.202
2	Umur × Jenis Kelamin × Kecamatan	80,94	109,76	3,83%	429
3	Status Perkawinan	983,50	1.324,07	1,43%	1.871

Analisis *small-cell* menunjukkan bahwa *constraint* Umur × Jenis Kelamin × Kecamatan (130 sel) dan Status Perkawinan (4 sel) tidak memiliki sel di bawah 5 maupun 10 individu. Sebaliknya, *constraint* Umur × Aktivitas Utama × Jenis Kelamin memiliki 116 dari 390 sel (29,74%) dengan jumlah di bawah 5 individu. Kondisi tersebut berasal dari kombinasi kategori yang memang jarang pada data agregat dan tidak disebabkan oleh penambahan atribut baru pada tahap IPF maupun TRS.

Dengan ambang $k = 5$ sebagai indikator frekuensi minimum (Sweeney, 2002), sel-sel tersebut berpotensi meningkatkan risiko *disclosure* apabila dikombinasikan dengan atribut eksternal. Karena keempat variabel belum tersedia dalam satu *constraint* silang, risiko *disclosure* pada irisan tersebut belum sepenuhnya terukur. Oleh karena itu, *cell suppression* atau *generalization* tetap diperlukan sebelum publikasi pada granularitas kecamatan (Willenborg & Waal, 2001).

Tabel 10. Tiga Kombinasi Kategori dengan Selisih Terbesar terhadap Target

No	Kombinasi (Umur Aktivitas Jenis Kelamin)	Target	Aktual	Selisih	(%)
1	15-19 Pelajar/Mahasiswa Perempuan	9.121	10.323	1.202	13,18%
2	15-19 Belum/Tidak Bekerja Perempuan	5.276	4.189	1.087	20,60%
3	55-59 Mengurus Rumah Tangga Perempuan	7.476	6.672	804	10,75%

Tabel 10 menunjukkan deviasi absolut terbesar terdapat pada kelompok perempuan usia 15–19 tahun dengan aktivitas pelajar/mahasiswa, yaitu 1.202 individu (13,18%). Meskipun populasi akhir telah disesuaikan dengan total target, deviasi masih dapat terjadi pada tingkat sel.

Konsentrasi deviasi pada kelompok perempuan usia 15–19 tahun dengan aktivitas pelajar/mahasiswa dapat mengindikasikan variasi transisi status aktivitas pada kelompok usia tersebut (misalnya peralihan antara status pelajar dan belum/tidak bekerja) yang secara alami lebih tinggi dibandingkan kelompok usia dewasa yang aktivitasnya relatif stabil. Namun, penelitian ini tidak menguji hipotesis tersebut secara empiris sehingga interpretasi ini bersifat tentatif dan dapat dikonfirmasi pada penelitian lanjutan.

Temuan tersebut menunjukkan *trade-off* antara privasi dan utilitas, sejalan dengan penelitian sebelumnya (C. Little et al., 2025; Ponte et al., 2024). Kalibrasi IPF meningkatkan utilitas statistik, ditunjukkan oleh peningkatan JSS dari 0,9070 menjadi 0,9856 dan penurunan MAE Cramer's V dari 0,0542 menjadi 0,0279. Namun, penyesuaian terhadap *constraint* yang *granular* dapat mengurangi ruang variasi pada

kombinasi kategori langka, tercermin dari *exact match rate* 64,75% dan 116 sel dengan $k < 5$.

Evaluasi utilitas penelitian ini berfokus pada kemiripan distribusi dan *constraint accuracy* (Karr et al., 2006). Oleh karena itu, meskipun utilitas statistik meningkat, mitigasi SDC tambahan pada sel berisiko tetap direkomendasikan sebelum publikasi terbuka pada granularitas kecamatan.

Secara keseluruhan, integrasi CTGAN dan IPF menghasilkan populasi sintetis yang memenuhi *constraint* agregat resmi Kota Cirebon dengan tingkat kesalahan yang rendah, dengan tetap mempertimbangkan keterbatasan privasi pada sel-sel *granular*.

KESIMPULAN

Penelitian ini menghasilkan *pipeline* pembangkitan populasi sintetis kependudukan yang mengintegrasikan CTGAN dan Iterative Proportional Fitting (IPF). CTGAN digunakan untuk mempelajari struktur hubungan antarvariabel demografis dari data referensi IFLS5, sedangkan IPF mengkalibrasi *seed population* terhadap *constraint* agregat resmi Kota Cirebon. Integrasi kedua metode memungkinkan pembentukan populasi sintetis berdasarkan struktur mikro referensi tanpa menggunakan data mikro penduduk Kota Cirebon.

Model CTGAN terpilih menghasilkan kemiripan distribusi dan preservasi hubungan antarvariabel yang baik terhadap data referensi, dan kualitas tersebut meningkat secara konsisten setelah kalibrasi IPF, dengan seluruh *constraint* agregat mencapai kategori akurasi sangat baik terhadap target Kota Cirebon. Namun, sebagian kecil sel pada *constraint* Umur $\times$ Aktivitas $\times$ Jenis Kelamin masih berfrekuensi rendah, injeksi Kecamatan tetap bergantung pada hubungan kondisional yang terbatas, dan risiko privasi belum diuji secara empiris melalui *membership inference attack* maupun *differential privacy* formal. Oleh karena itu, populasi sintetis yang dihasilkan perlu dipandang sebagai representasi statistik yang mendekati data agregat target, bukan sebagai data yang telah terbukti aman secara privasi.

Bagi Disdukcapil dan Pemerintah Kota Cirebon, *pipeline* ini dapat dipertimbangkan sebagai acuan penyediaan data sintetis untuk kebutuhan riset dan simulasi kebijakan yang memerlukan karakteristik individu. Penerapannya perlu disertai pembaruan *constraint* berdasarkan DKB terbaru serta penilaian *disclosure risk* dan mitigasi *statistical disclosure control* pada sel berisiko sebelum publikasi.

Penelitian lanjutan disarankan untuk: (1) menguji risiko privasi secara empiris melalui simulasi *membership inference attack* berbasis *shadow model* dan mengevaluasi *differential privacy* formal; (2) membandingkan CTGAN dengan TVAE atau Bayesian Network menggunakan protokol evaluasi yang setara; serta (3) melakukan *hyperparameter tuning* yang lebih luas dan memperluas *constraint* agregat, khususnya hubungan Kecamatan dengan tingkat pendidikan dan aktivitas utama, untuk mengurangi keterbatasan asumsi pada *pipeline* saat ini.

Pernyataan mengenai Penggunaan Kecerdasan Buatan (AI)

AI digunakan dalam penyusunan naskah ini. Penulis menggunakan perangkat kecerdasan buatan semata-mata untuk membantu perbaikan bahasa, pemeriksaan tata bahasa, serta penyuntingan demi kejelasan dan keterbacaan. Penulis tetap bertanggung jawab penuh atas akurasi, orisinalitas, integritas, dan konten akademis naskah tersebut. Seluruh konten yang dibantu AI telah ditinjau dan diverifikasi oleh penulis sebelum naskah diserahkan.

Konflik Kepentingan

Penulis menyatakan tidak ada konflik kepentingan.

Kontribusi Penulis

Konseptualisasi, Achmad Syaiful Hidayat, Marsani Asfi, dan Petrus Sokibi; metodologi, Achmad Syaiful Hidayat; perangkat lunak, Achmad Syaiful Hidayat; validasi, Marsani Asfi dan Petrus Sokibi; analisis formal, Achmad Syaiful Hidayat; investigasi, Achmad Syaiful Hidayat; sumber daya, Achmad Syaiful Hidayat; kurasi data, Achmad Syaiful Hidayat; penulisan—draf awal, Achmad Syaiful Hidayat; penulisan—peninjauan dan penyuntingan, Marsani Asfi dan Petrus Sokibi; visualisasi, Achmad Syaiful Hidayat; supervisi, Marsani Asfi dan Petrus Sokibi; administrasi proyek, Achmad Syaiful Hidayat. Seluruh penulis telah membaca dan menyetujui versi akhir naskah ini.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Universitas Catur Insan Cendekia atas dukungan dan fasilitas yang diberikan selama pelaksanaan penelitian ini. Penulis juga menyampaikan apresiasi kepada RAND Corporation dan seluruh pihak yang terlibat dalam penyelenggaraan Indonesia Family Life Survey (IFLS) atas penyediaan data IFLS5 yang digunakan sebagai data mikro referensi dalam penelitian ini. Ucapan terima kasih juga disampaikan kepada Dinas Kependudukan dan Pencatatan Sipil Kota Cirebon atas penyediaan data agregat yang digunakan sebagai constraint dalam proses kalibrasi populasi sintetis.

DAFTAR PUSTAKA

- Andika, S. D., & Agustiono, W. (2024). Analisis Penerapan Sistem Informasi Administrasi Kependudukan (SIK) di Dinas Kependudukan dan Catatan Sipil Kabupaten Tuban dengan Metode End-User Computing Satisfaction. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(4). <https://doi.org/10.36040/jati.v8i4.10264>
- Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6 PART A), 415–429. [https://doi.org/10.1016/0965-8564\(96\)00004-3](https://doi.org/10.1016/0965-8564(96)00004-3)
- Bergsma, W. (2013). A bias correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*, 42(3), 323–328. <https://doi.org/10.1016/j.jkss.2012.10.002>
- Corporation, R. (2016). *The Indonesia Family Life Survey (IFLS5): Wave 5*. <https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS.html>
- Darsel, V., Côme, E., & Oukhellou, L. (2026). Robust and reproducible evaluation framework for population synthesis models—Application to probabilistic and deep generative models. *Transportation Research Part C: Emerging Technologies*, 186. <https://doi.org/10.1016/j.trc.2026.105627>
- Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4), 427–444. <https://doi.org/10.1214/aoms/1177731829>
- Duran-Heras, A., Garcia-Gutierrez, I., & Castilla-Alcala, G. (2018). Comparison of Iterative Proportional Fitting and Simulated Annealing as synthetic population generation

- techniques: Importance of the rounding method. *Computers, Environment and Urban Systems*, 68, 78–88. <https://doi.org/10.1016/j.compenvurbsys.2017.11.001>
- Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7), 1858–1860. <https://doi.org/10.1109/TIT.2003.813506>
- Errington, A., Einbeck, J., Cumming, J., Rössler, U., & Endesfelder, D. (2022). The effect of data aggregation on dispersion estimates in count data models. *The International Journal of Biostatistics*, 18(1), 183–202. <https://doi.org/10.1515/ijb-2020-0079>
- Fadliansyah, F., Mubarak, M. K., Aziz, M. D., & Augustia, A. E. (2025). Evaluasi Pembelajaran dari Kasus Kebocoran Data di Indonesia sebagai Negara dengan Data Breach Terbesar Ke-8 Dunia. *TEKNOBIS: Jurnal Teknologi, Bisnis Dan Pendidikan*, 3(2), 272–277. <https://jurnalmahasiswa.com/index.php/teknobis/article/view/3025>
- Gelb, J., Apparicio, P., & Alizadeh, H. (2024). A synthetic vulnerable population dataset for fine scale geographical equity analysis and urban planning. *Scientific Data*, 11, 954. <https://doi.org/10.1038/s41597-024-03771-6>
- Gulo, J. (2025). Identifikasi Faktor Penentu Kemiskinan di Indonesia Melalui Pendekatan Regresi Linier Berganda. *J. Publ. Sist. Inf. Dan Manaj. Bisnis*, 4(2), 299–310. <https://doi.org/10.55606/jupsim.v4i2.4157>
- Hyeong, J., Kim, J., Park, N., & Jajodia, S. (2022). An empirical study on the membership inference attack against tabular data synthesis models. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, 4064–4068. <https://doi.org/10.1145/3511808.3557546>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Indonesia, R. (2022). *Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi*.
- Juanuari, J., Manzis, I., Musyaffa, A., Ngara, S. M., Syahla, H., & Putra, J. L. (2026). Implementasi K-Nearest Neighbor untuk Klasifikasi Tingkat Kemiskinan Sebagai Pendukung Pengambilan Keputusan Sosial di Indonesia. *Infomatek*, 28(1), 1–10. <https://doi.org/10.23969/infomatek.v28i1.26866>
- Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3), 224–232. <https://doi.org/10.1198/000313006X124640>
- Kim, E.-J., & Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C: Emerging Technologies*, 148. <https://doi.org/10.1016/j.trc.2023.104053>
- Lewis, C. D. . (1982). *Industrial and business forecasting methods : a practical guide to exponential smoothing and curve fitting*. Butterworth Scientific.
- Lin, Z., Khetan, A., Fanti, G., & Oh, S. (2018). *PacGAN: The power of two samples in generative adversarial networks*.
- Little, C., Allmendinger, R., & Elliot, M. (2025). Synthetic Census Microdata Generation: A Comparative Study of Synthesis Methods Examining the Trade-Off Between Disclosure Risk and Utility. *Journal of Official Statistics*, 41(1). <https://doi.org/10.1177/0282423X241266523>
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley. <https://doi.org/10.1002/9781119482260>

- Lovelace, R., & Ballas, D. (2013). "Truncate, replicate, sample": A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41, 1–11. <https://doi.org/10.1016/j.compenvurbsys.2013.03.004>
- Miletic, M., & Sariyar, M. (2024). Challenges of Using Synthetic Data Generation Methods for Tabular Microdata. *Applied Sciences*, 14(14), 5975. <https://doi.org/10.3390/app14145975>
- Miura, T., Kii, M., Shibahara, T., Iwahana, K., Okuda, T., Ichikawa, A., & Yanai, N. (2024). SETSUBUN: Revisiting Membership Inference Game for Evaluating Synthetic Data Generation. *Journal of Information Processing*, 32, 757–766. <https://doi.org/10.2197/ipsjip.32.757>
- Mwigereri, D. G., Kamotho, N. T., Waljee, A. K., Rego, R. T., Weinheimer-Haus, E. M., Alarakhiya, F., Ngugi, A. K., Price, W. N., Zhu, J., Wong, S. P., & Siwo, G. H. (2025). Synthetic data generation of health and demographic surveillance systems data: a case study in a low- and middle-income country. *JAMIA Open*, 8(6). <https://doi.org/10.1093/jamiaopen/ooaf137>
- Pathare, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *International Journal of Information Management Data Insights*, 3(2), 100177. <https://doi.org/10.1016/j.ijime.2023.100177>
- Ponte, G. R., Wieringa, J. E., Boot, T., & Verhoef, P. C. (2024). Where's Waldo? A framework for quantifying the privacy-utility trade-off in marketing applications. *International Journal of Research in Marketing*, 41(3), 529–546. <https://doi.org/10.1016/j.ijresmar.2024.05.003>
- Putra, R. K., Agustin, Y., Ihsan, L. N., & Dafiqi, Z. A. (2025). Institutional Dysfunction in Personal Data Protection: A Legal-Political Analysis Based on New Institutional Theory. *PERKARA: Jurnal Ilmu Hukum Dan Politik*, 3(2), 938–949. <https://doi.org/10.51903/bkktey52>
- Ramadhan, R., & Bintari, A. (2025). Kualitas E-Service dalam Pelayanan Administrasi Kependudukan melalui Aplikasi E-OPEN di Kota Bekasi Tahun 2023. *Jurnal Governansi*, 11(1), 1–14. <https://doi.org/10.30997/jgs.v11i1.14204>
- Reiter, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1), 185–205. <https://doi.org/10.1111/j.1467-985X.2004.00343.x>
- Rendi, M., Tamza, F. B., & Fardiansyah, A. I. (2026). Perlindungan Data Pribadi dalam Era Digitalisasi dan Tantangan Keamanan Siber di Indonesia. *Jurnal Hukum Lex Generalis*, 6(12). <https://doi.org/10.56370/jhlg.v6i12.2452>
- Sharma, P., Kumar, M., Sharma, H. K., & Biju, S. M. (2024). Generative adversarial networks (GANs): Introduction, Taxonomy, Variants, Limitations, and Applications. *Multimedia Tools and Applications*, 83, 88811–88858. <https://doi.org/10.1007/s11042-024-18767-y>
- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy* (pp. 3–18). <https://doi.org/10.1109/SP.2017.41>
- Soleh, M., & Tjenreng, Z. (2025). Strategi Pencegahan Kebocoran Data Pelayanan Publik di Era Digital. *Jurnal Kajian Pemerintah: Journal of Government, Social and Politics*, 11(1). [https://doi.org/10.25299/jkp.2025.vol11\(1\).20524](https://doi.org/10.25299/jkp.2025.vol11(1).20524)

- Strauss, J., Witoelar, F., & Sikoki, B. (2016). *The Fifth Wave of the Indonesia Family Life Survey: Overview and Field Report: Volume 1*. RAND Corporation. <https://doi.org/10.7249/WR1143.1>
- Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61, 49–62. <https://doi.org/10.1016/j.trc.2015.10.010>
- Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570. <https://doi.org/10.1142/S0218488502001648>
- van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M. (2023). Membership Inference Attacks against Synthetic Data through Overfitting Detection. In *Proceedings of AISTATS* (Vol. 206, pp. 3493–3514).
- van Buuren, S. (2018). Flexible Imputation of Missing Data (2nd ed.). In *Chapman and Hall/CRC*. <https://doi.org/10.1201/9780429492259>
- Voas, D., & Williamson, P. (2001). Evaluating Goodness-of-Fit Measures for Synthetic Microdata. *Geographical and Environmental Modelling*, 5(2), 177–200. <https://doi.org/10.1080/13615930120086078>
- Willenborg, L. C. R. J., & Waal, T. de. (2001). *Elements of statistical disclosure control*. Springer.
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79–82. <https://doi.org/10.3354/cr030079>
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 32, pp. 7333–7343).
- Yao, Z., Krčo, N., Ganev, G., & de Montjoye, Y.-A. (2025). *The DCR Delusion: Measuring the Privacy Risk of Synthetic Data*.
- Yulanda, A., & Frinaldi, A. (2023). Inovasi Program Identitas Kependudukan Digital dalam Upaya Meningkatkan Kualitas Layanan Kependudukan di Indonesia. *Titian: Jurnal Ilmu Humaniora*, 7(2), 415–426. <https://doi.org/10.22437/titian.v7i2.30348>
- Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. *Journal of Biomedical Informatics*, 125. <https://doi.org/10.1016/j.jbi.2021.103977>
- Zhao, Z. (2023). CTAB-GAN+: Enhancing Tabular Data Synthesis. *Frontiers in Big Data*. <https://doi.org/10.3389/fdata.2023.1296508>