

Dati sintetici e Modelli GAN

Teoria e Fondamenti

Francesca Liguori, Manuel Caccone

11 luglio 2025

Section 1

Introduzione ai dati sintetici

- I dati sintetici sono **dataset artificiali** progettati per replicare le **proprietà statistiche**, le **correlazioni** e le **distribuzioni** dei dati reali
- Preservano la struttura e le relazioni tra variabili senza contenere riferimenti a individui o a eventi specifici
- Costituiscono una soluzione efficace al trade-off tra utilità analitica e protezione della privacy
- Permettono lo sviluppo e la validazione di modelli attuariali in contesti sicuri

- **Completamente sintetici:** generati interamente attraverso modelli statistici o algoritmi
- **Parzialmente sintetici:** solo le variabili sensibili vengono sostituite con valori artificiali
- **Ibridi:** combinazione strategica di elementi reali e sintetici per obiettivi specifici

Section 2

Vantaggi e sfide nell'adozione dei dati sintetici

I dati sintetici eliminano i colli di bottiglia legati alla disponibilità e alla qualità dei dati reali, offrendo i seguenti vantaggi:

- **Scalabilità:** possibilità di generare volumi illimitati per addestrare modelli complessi
- **Compliance normativa:** assenza di informazioni sensibili, in piena conformità con GDPR e normative sulla protezione dei dati
- **Disponibilità di scenari:** generazione di casistiche rare ma rilevanti per la modellazione
- **Riduzione dei bias:** capacità di correggere distorsioni storiche presenti nei dataset reali, migliorando l'equità dei modelli predittivi
- **Costi ridotti:** risparmio notevole rispetto alla raccolta e annotazione manuale

- **Degradazione della qualità:** modelli generativi imperfetti possono introdurre artefatti o relazioni spurie, compromettendo l'affidabilità
- **Ambiguità della compliance normativa:** assenza di standard uniformi per la validazione legale, particolarmente rilevante in ambiti regolamentati come le assicurazioni
- **Complessità tecniche:** necessità di competenze specializzate in AI e statistica, spesso scarse nelle organizzazioni tradizionali

- **Pricing assicurativo:** combinazione di dati sintetici con informazioni tradizionali (età, parametri geografici, ...) per migliorare la personalizzazione del premio
- **Analisi di scenari catastrofici:** simulazione di eventi climatici estremi per migliorare la valutazione dei rischi
- **Rilevamento frodi:** creazione di pattern comportamentali anomali per training di modelli, senza esporre dati sensibili

Section 3

Ratemaking con dati sintetici

Fondamenti del ratemaking sintetico

- **Ratemaking tradizionale:** determinazione dei premi assicurativi basata su analisi statistica dei dati storici di sinistri, esposizioni e variabili di rischio
- **Innovazione sintetica:** utilizzo di dati artificiali per superare i limiti dei dataset storici e migliorare la precisione attuariale

Vantaggi specifici per il ratemaking

- **Ampliamento del dataset:** generazione di osservazioni per segmenti con pochi dati storici
- **Bias correction:** correzione di distorsioni storiche nei portafogli esistenti
- **Validazione robusta:** test di stress dei modelli su casistiche diversificate

Implementazione pratica nel ratemaking

Pipeline metodologica:

- ➊ **Analisi dei dati storici:** identificazione di pattern di sinistralità, frequenza e costo medio per classe di rischio
- ➋ **Generazione sintetica segmentata:** creazione di dati sintetici specifici per linee di business
- ➌ **Calibrazione attuariale:** validazione dei modelli sintetici contro benchmark di mercato e test di coerenza
- ➍ **Integrazione nel pricing:** utilizzo combinato di dati reali e sintetici per determinazione dei premi

Applicazioni emergenti

- **Micro-segmentazione:** pricing personalizzato attraverso variabili comportamentali sintetiche
- **Prodotti innovativi:** sviluppo di coperture per rischi emergenti (cyber, climate change)

Section 4

Generazione dei dati sintetici

- ➊ **Preparazione e analisi dei dati reali:** pulizia e codifica dei dati, seguite dalle analisi statistiche necessarie a comprenderne le caratteristiche fondamentali
- ➋ **Scelta e addestramento del modello generativo:** selezione del modello e addestramento sui dati reali, per apprendere le distribuzioni e le relazioni tra le variabili
- ➌ **Generazione e post-processing:** generazione dei dati sintetici, poi trasformati per renderli coerenti con il formato e la struttura dei dati reali di partenza
- ➍ **Validazione e utilizzo:** verifica della coerenza tra i dati sintetici e i dati reali rispetto alle distribuzioni e alle relazioni tra le variabili e verifica dell'anonimizzazione del dataset sintetico prima dell'utilizzo effettivo

- **Modelli parametrici:** utilizzo di distribuzioni note (Pareto, Gamma, Weibull) per la simulazione
- **Tecniche di ricampionamento:** bootstrap e varianti per la generazione di nuovi dataset
- **Simulazione Monte Carlo:** generazione di scenari basata su modelli probabilistici definiti

- **Generative Adversarial Networks (GAN):** competizione tra reti neurali per la generazione ottimale
- **Variational Autoencoders (VAE):** codifica e decodifica attraverso spazi latenti
- **Transformer e Large Language Models:** generazione di dati strutturati tramite modelli linguistici

Section 5

Modelli *Generative Adversarial Networks* (GAN)

- **Introdotte da Ian Goodfellow nel 2014**
- **Framework basato sulla teoria dei giochi a somma zero**
- Due reti neurali che competono in un processo antagonistico:
 - **Discriminator (D)**: impara a distinguere dati reali dai dati sintetici generati
 - **Generator (G)**: impara a creare dati sintetici sempre più realistici per ingannare il discriminatore
- **Minmax Game**: G cerca di minimizzare la capacità di D di distinguere i dati generati da quelli reali

- **Architettura competitiva:** due reti neurali (Generator e Discriminator) in competizione
- **Processo iterativo:** miglioramento progressivo della qualità dei dati generati
- **Equilibrio di Nash:** convergenza verso dati sintetici indistinguibili da quelli reali
- **Preservazione delle correlazioni:** mantenimento di relazioni complesse tra variabili

- **Funzione obiettivo:**

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{\xi \sim p_{\xi}(\xi)} [\log(1 - D(G(\xi)))]$$

Dove

- $p_{data}(x)$: distribuzione dei dati reali
- $p_{\xi}(\xi)$: distribuzione del rumore ξ in input al generatore (spesso normale multivariata)
- $D(x)$: probabilità che x provenga dai dati reali secondo il discriminatore
- $G(\xi)$: output del generatore dato un input casuale ξ

Funzionamento del training GAN (1/3)

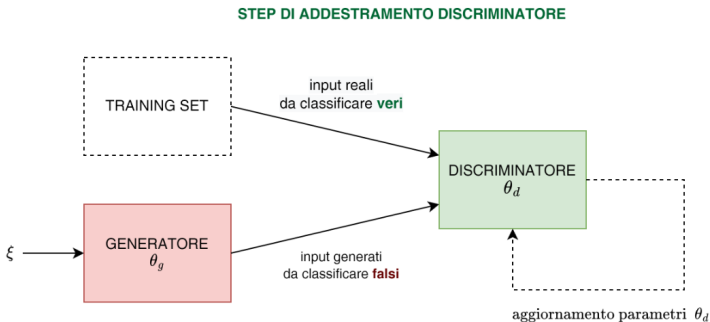


Figure 1: Step 1

Funzionamento del training GAN (2/3)

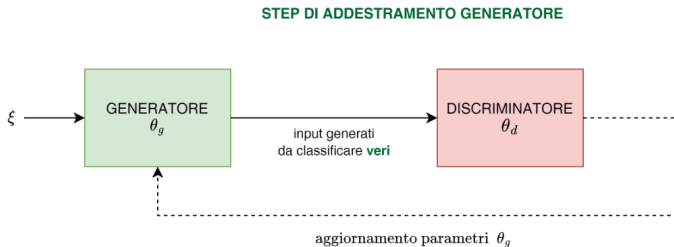


Figure 2: Step 2

Funzionamento del training GAN (3/3)

Il training è un processo iterativo che si interrompe quando si raggiunge l'**equilibrio di Nash**, ossia quando $D(x) = \frac{1}{2}$ per tutti gli x

- **Alta qualità del dato sintetico:**
 - Dati sintetici molto simili a quelli reali grazie al lavoro collaborativo tra le due reti neurali
- **Preservazione delle correlazioni complesse:**
 - Catturano relazioni non lineari tra variabili
 - Mantengono la struttura dei dati multivariati
- **Flessibilità:**
 - Capacità di generare scenari rari ma plausibili, utile per analisi di scenario
- **Scalabilità:**
 - Applicabilità a dataset di grandi dimensioni

Sfide nell'implementazione GAN

- **Mode collapse:** il generatore produce solo pochi esempi, con una ridotta capacità di generalizzazione
 - Soluzione: utilizzare le varianti GAN (e.g. WGAN)
- **Training instabile:**
 - Difficoltà nel bilanciare l'apprendimento di G e D
 - Vanishing gradients
 - Soluzioni: spectral normalization o altre tecniche per stabilizzare il training
- **Valutazione:** difficoltà nel misurare oggettivamente la qualità dei dati generati
 - Metriche: Inception Score, Fréchet Inception Distance, Maximum Mean Discrepancy che misurano la *fidelity* (qualità) e la *diversity* (varietà) dei dati generati

Section 6

Validazione e Controllo Qualità

- **Congruenza univariata:** test di Kolmogorov-Smirnov per la distribuzione delle singole variabili
- **Preservazione delle correlazioni:** analisi della matrice di correlazione tra dati reali e sintetici
- **Stabilità statistica:** confronto di momenti, quantili e statistiche descrittive

Section 7

Aspetti Normativi e Privacy

- **Status di anonimato:** i dati sintetici correttamente generati sono considerati anonimi
- **Esenzione dal consenso:** non richiedono consenso esplicito per il trattamento
- **Verifica dell'anonimato:** necessità di test per escludere rischi di re-identificazione
- **Privacy by design:** integrazione della protezione privacy nel processo di generazione

- **k-anonymity**: verifica che ogni combinazione di attributi quasi-identificativi sia condivisa da almeno k record diversi
- **l-diversity**: controllo della diversità dei valori sensibili all'interno dei gruppi
- **Differential privacy**: aggiunta di rumore calibrato per garantire la privacy matematica

Section 8

Implicazioni per la professione attuariale

- **Espansione delle possibilità analitiche:** accesso a scenari prima inaccessibili
- **Accelerazione dell'innovazione:** sviluppo più rapido di prodotti e servizi
- **Standardizzazione:** evoluzione verso best practice condivise nel settore
- **Competenze emergenti:** necessità di aggiornamento delle competenze professionali

Section 9

Conclusioni

I dati sintetici rappresentano un'evoluzione significativa per la pratica attuariale moderna

- **Privacy e utilità:** equilibrio ottimale tra protezione dei dati e capacità analitica
- **Metodologie mature:** disponibilità di strumenti consolidati e approcci emergenti
- **Integrazione strategica:** complemento essenziale al toolkit dell'attuario contemporaneo

- **Formazione continua:** investimento in competenze specifiche per massimizzare i benefici
- **Sperimentazione controllata:** approccio incrementale per acquisire esperienza pratica
- **Collaborazione professionale:** condivisione di esperienze e best practice
- **Monitoraggio normativo:** attenzione all'evoluzione del quadro regolamentare

Section 10

Bibliografia

- Goodfellow, I., et al. (2014). “Generative Adversarial Nets”. *Advances in Neural Information Processing Systems*, 27.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). “Wasserstein Generative Adversarial Networks”. *International Conference on Machine Learning*.
- Jordon, J., Yoon, J., & van der Schaar, M. (2019). “PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees”. *International Conference on Learning Representations*.

- Kuo, K., & Lupton, D. (2020). "Towards Explainability of Machine Learning Models in Insurance Pricing". *ASTIN Bulletin*, 50(1), 267-296.
- Lindholm, M., et al. (2022). "Machine Learning in Life Insurance: A Review of Methods and Applications". *European Actuarial Journal*, 12(1), 333-394.
- Denuit, M., & Hainaut, D. (2021). "Machine Learning Models for Motor Insurance Ratemaking". *European Actuarial Journal*, 11(1), 199-217.

- El Emam, K., et al. (2020). "A systematic review of re-identification attacks on health data". *PLoS ONE*, 15(12), e0243633.
- Rubin, D. B. (1993). "Statistical Disclosure Limitation". *Journal of Official Statistics*, 9(2), 461-468.
- European Commission (2016). "General Data Protection Regulation (GDPR)". *Official Journal of the European Union*.

Section 11

Grazie per l'attenzione!

- **Demo interattiva:** Webapp Shiny con dati sintetici
- **Vibe coding for actuaries**, pacchetti R essenziali