DAV/DGVFM
Herbsttagung
2025

Dr. Guido Grützner, QuantAkt Consulting
Dr. Maik Urban, Deloitte

Erzeugung synthetischer Daten: Use Cases, Software und Anleitung

Fachgruppe ADS/AI 18. November 2025

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Agenda

1. Vorstellung der AG
2. Motivation und Definition synthetischer Daten
3. Softwaretools zur Erzeugung synthetische Daten
4. Synthetische Daten mit Bäumen
5. Anwendungsfälle
6. Bisherige Ergebnisse

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Die AG Synthetisierung von Daten

Ziele

Untersuchung und Bewertung von Methoden zur Datensynthetisierung für aktuarielle Anwendungen.

Unterarbeitsgruppen

1. Synthetisierung mit Tools und echten Daten
2. Neue Datensätze erzeugen

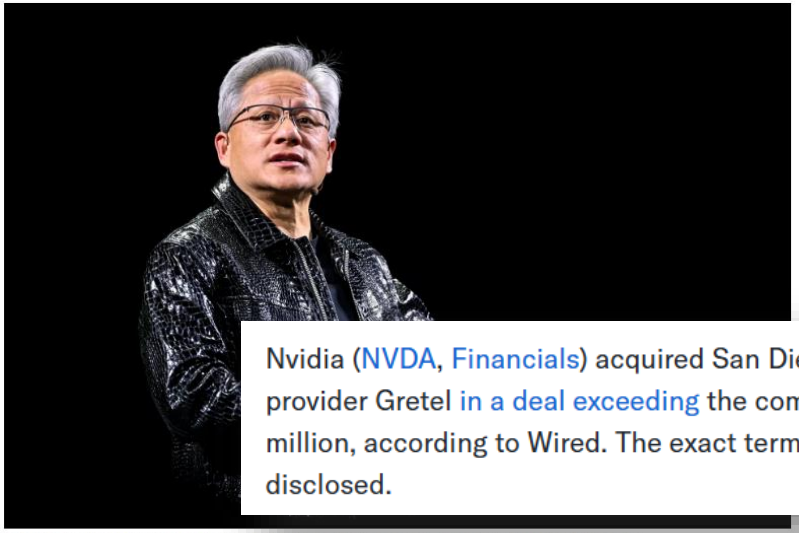
Mitglieder

- Dariush Sadeghi-Yam (Leitung)
- Friedrich Loser (Leitung)
- Vanessa Dietze
- Dr. Guido Grützner
- André Jambon
- Kristina Reisenauer
- Dr. Maik Urban
- Prof. Dr. Christian Weiß
- Dr. Sven Wiesinger
- Leonid Zeldin

Synthetische Daten sind kein Nischenthema

Nvidia Bets Big on Synthetic Data

Nvidia has acquired synthetic data startup Gretel to bolster the AI training data used by the chip maker's customers and developers.



Nvidia CEO Jensen Huang addresses participants at the keynote of CES 2025 in Las Vegas, Nevada. PHOTOGRAPH: ARTUR WIDAK/GETTY IMAGES

Nvidia (NVDA, Financials) acquired San Diego-based synthetic data provider Gretel in a deal exceeding the company's last valuation of $320 million, according to Wired. The exact terms of the transaction were not disclosed.

Quelle: https://www.wired.com/story/nvidia-gretel-acquisition-synthetic-training-data/

TECHNOLOGY

Nvidia, Google, OpenAI Turn To 'Synthetic Data' Factories To Train AI Models

Quelle: https://www.investors.com/news/technology/nvidia-stock-tech-giants-use-synthetic-data-train-ai-models/

How Apple Will Analyze Your Data to Train Its AI -- While Protecting Your Privacy

In a blog post, the company outlined a new approach designed to expand its AI capabilities while safeguarding user privacy, especially as competitors like OpenAI and Google advance more quickly with fewer restrictions. Apple said it will train its AI models using synthetic data, known as information that mimics the format and characteristics of real-world messages without including any actual user-generated content.

Quelle: How Apple Will Analyze Your Data to Train Its AI -- While Protecting Your Privacy

Die Kunst, Realität nachzubilden, ohne sie zu kopieren

Begriffsdefinition[1]

Synthetische Daten sind **künstlich erzeugte Daten**, die reale Daten in ihrer **Struktur** und **statistischen Eigenschaften** nachbilden, aber **keine echten Personen oder Vorgänge** repräsentieren.

Abzugrenzen von

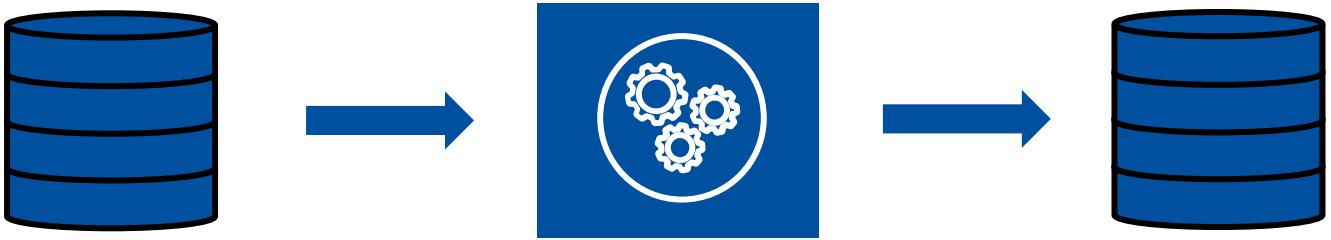
- **Anonymisierten Daten**: Reale Daten, bei denen alle identifizierbaren Merkmale entfernt wurden.
- **Pseudonymisierten Daten**: Reale Daten, bei denen identifizierbare Merkmale durch Schlüssel ersetzt wurden, die rückführbar sind.

[1] Angelehnt an die Definition des World Economic Forums: https://reports.weforum.org/docs/WEF_Synthetic_Data_2025.pdf

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Struktur synthetischer Daten und deren Erzeugung



Reale Daten

- Schadendaten
- Gesundheitsdaten
- …

Synthetisierungs-Software

Synthetische Daten

- Identische Struktur
- Ähnliche statistische Eigenschaften

Der Fokus liegt auf einer **datengetriebenen Synthetisierung** von **tabellarischen** Daten.

Softwaretools zur datengetriebenen Synthetisierung

In der Arbeitsgruppe haben wir folgende Pakete/Bibliotheken untersucht

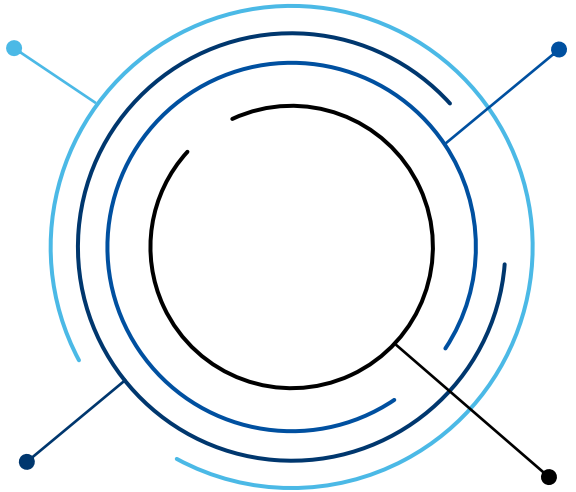
Python	R
• SDV (Copula, CTGAN, TVAE) • DataSynthesizer • Synthomat[1] • SynDiffix	• synthpop

Wir haben **keine kommerziellen Anbieter** geprüft.

[1] Paket im Rahmen der Arbeitsgruppe von Dr. Guido Grützner entwickelt

Vergleichskriterien für synthetische Daten

Laufzeit
... meint die **Gesamtdauer** für die Synthetisierung eines gegebenen Datensatzes



Fidelity
... ist ein Maß für die **Ähnlichkeit** der synthetischen Daten zum Originaldatensatz

Privacy
... misst wie gut synthetische Daten die **(Re-)Identifikation** sensibler Informationen im **Originaldatensatz** schützen

Utility
... bezieht sich darauf, wie nützlich oder **brauchbar** die synthetischen Daten für einen gegebenen **Verwendungszweck** sind

Gütemaße für Fidelity: Wie viel Ähnlichkeit ist gut genug?

Datentypen

Visualisierung

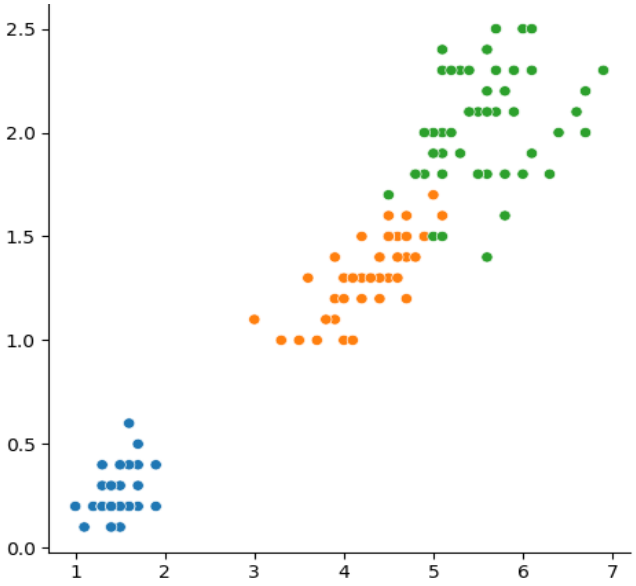
Statistische Tests

Einschätzungen der AG zu den Softwaretools

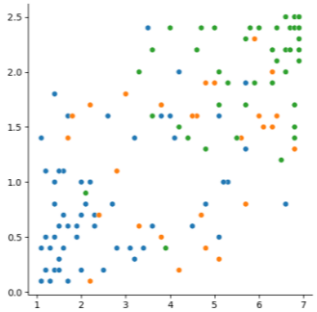
Softwarepaket	Fidelity	Laufzeit
SDV	Nicht überzeugend	Auch für große Datenmengen geeignet
DataSynthesizer	Überzeugend	Für kleine Datenmengen geeignet
Synthomat	Sehr überzeugend	Auch für große Datenmengen geeignet
synthpop	Sehr überzeugend	Auch für große Datenmengen geeignet

Als AG fanden wir Synthomat und synthpop für die betrachteten Datenmengen überzeugend.

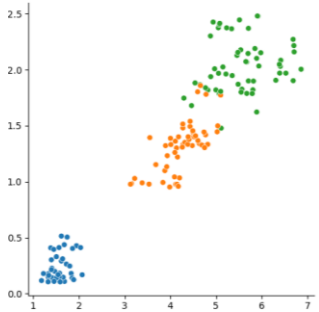
Visueller Vergleich von Clustern



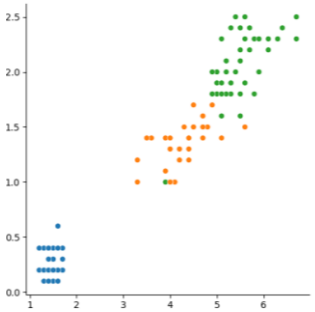
Originaldatensatz



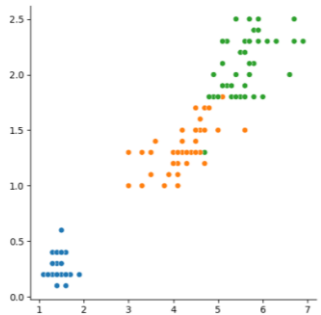
SDV



DataSynthesizer



synthpop



Synthomat

French Motor Datensatz

- ca. 680.000 Einträge à 11 numerische & kategoriale Attribute

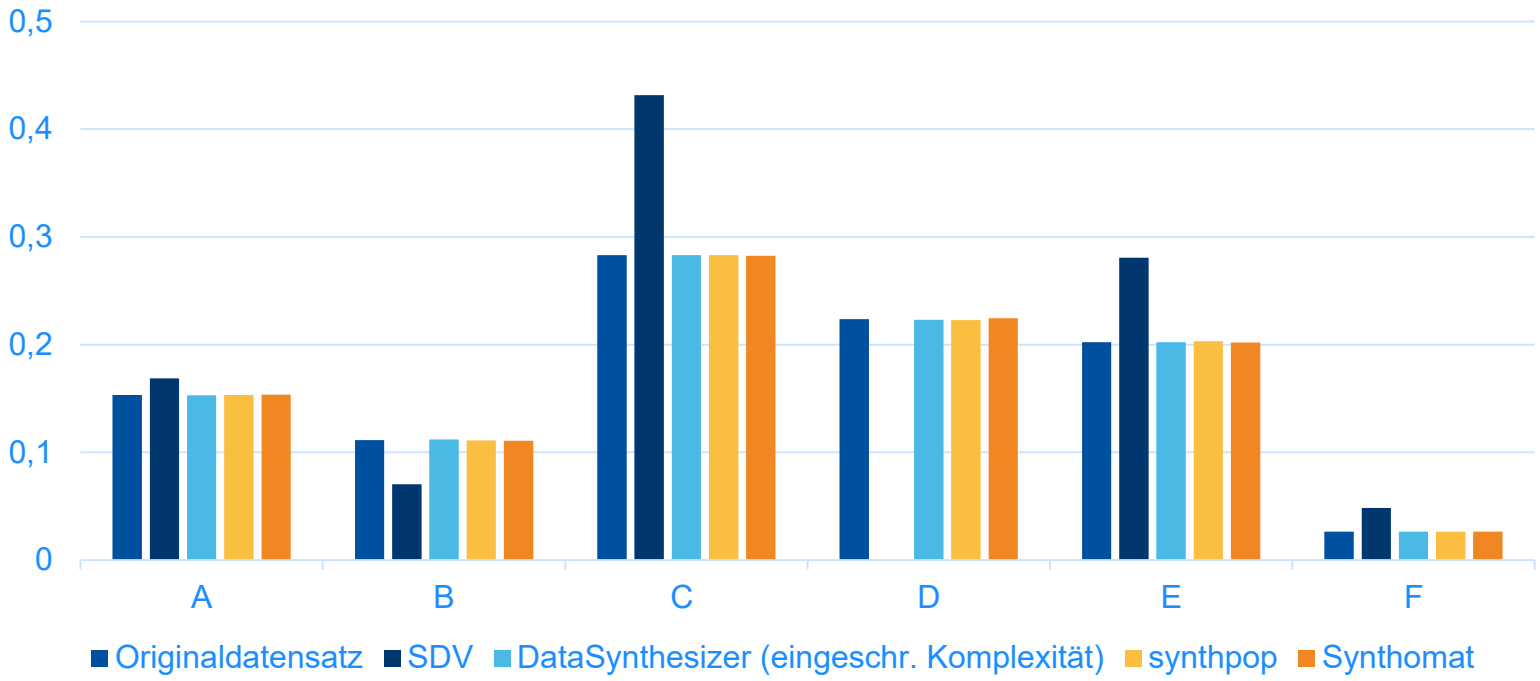
Softwarepaket	Laufzeit
SDV	6 Minuten
DataSynthesizer	> 12 Stunden
DataSynthesizer (eingeschränkte Komplexität)	16 Minuten
synthpop	49 Minuten
Synthomat	12 Minuten

Datenquelle: OpenML French-Motor-Claims-Datasets-freMTPL2freq

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Histogramme für das Attribut *Area*



Auffäliges Segment für die Modellierung im French Motor Datensatz

Segment mit
- Schadenhäufigkeit: 66%
- Segmentgröße: 3%

Datenmenge	Schadenhäufigkeit	Segmentgröße
SDV	0%	6%
DataSynthesizer	105%	1%
synthpop	60%	3%
Synthomat	59%	3%

Baumbasierte Generatoren - Grundlagen

- Die Verteilung, die dem Sample zu Grunde liegt, schreibt man als

$$\mathbb{P}(Z, Y, X) = \mathbb{P}(Z \mid X, Y) \cdot \mathbb{P}(Y \mid X) \cdot \mathbb{P}(X)$$

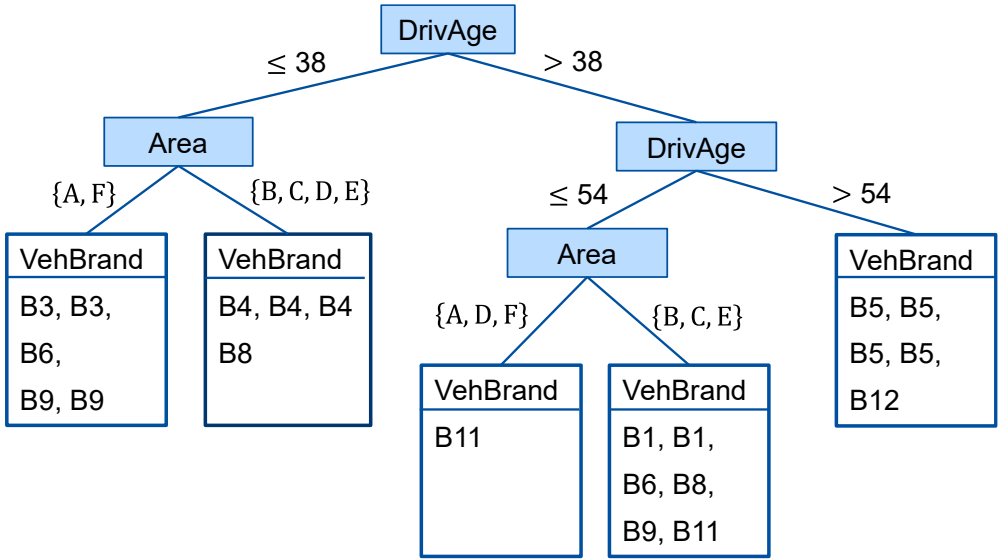
1. Trainiere erst je einen Baum für jeden bedingten Faktor
   a. Ein Baum für $\mathbb{P}(Y \mid X)$
   b. Ein Baum für $\mathbb{P}(Z \mid X, Y)$
   c. Und so weiter … falls es noch mehr Faktoren geben sollte

2. Dann sequentielles Sampling
   a. Ziehe ein X gemäss der Verteilung $\mathbb{P}(X)$
   b. Ziehe dann ein Y gemäss der Verteilung $\mathbb{P}(Y \mid X)$
   c. Ziehe dann ein Z gemäss der Verteilung $\mathbb{P}(Z \mid Y, X)$
   d. Und so weiter … falls es noch mehr Faktoren geben sollte

Baumbasierte Generatoren im Beispiel

- Der Baum sortiert alle Tupel $(\text{DrivAge}, \text{Area}, \text{VehBrand})$ des Samples
- In den Blättern ist jeweils das zu DrivAge und Area passende VehBrand
- Beim Sampling bedingt auf ein Paar $(\text{DrivAge}, \text{Area})$ wird einfach ein Vertreter aus dem passenden Blatt gezogen.

Der Baum für
$\mathbb{P}(\text{VehBrand} \mid \text{DrivAge}, \text{Area})$



Beispiele für Anwendungen

- Drei Anwendungsfälle die keine oder wenig Privacy beinhalten.
- Als Beispiele/Ideen was man mit synthetischen Daten machen kann
- Datensatz: French Motor
  - ~ 680‘000 Policen, 11 Attribute
  - Attribute sind «typisch aktuariell» also nominal, ordinal und numerisch
- Synthesizer: Synthomat
- Vier Machine Learning Modelle
  - GLM: Main Effects Poisson GLM (Package «GLUM»)
  - GBM: Gradient Boosted Tree (Scikit-Learn) in drei Varianten
    - Rein additiv («no_interactions»)
    - Paarweise Interaktionen («pairwise»)
    - Interaktionen jeder Ordnung («None»)

Anwendungsfall: Kauf einer aktuariellen Software

- Der kleine Kfz-Versicherer KarmaKasko möchte seine bisherige Software für Tarifierung und Reservierung durch ein besser integriertes und modernes Softwarepaket ersetzen.
- Dazu sollen verschiedene Anbieter verglichen werden. Ein guter Vergleichsansatz wäre es, den Anbietern ein Musterportfolio zur Verfügung zu stellen und die Darstellung und Qualität der Vorhersage von Schadenfrequenzen zu vergleichen.
- Allerdings gibt es einen alten Aufsichtsratsbeschluss, der jede Herausgabe von Originaldaten – egal zu welchen Zwecken – verbietet. Eine Herausgabe eines synthetischen Portfolios ist aber gemäss Vorstand OK «solange die Daten nicht zu ähnlich sind».

Die drei Dimensionen

Utility:

1. Die synthetischen Daten sollen genauso umfangreich und komplex sein wie der Originalbestand.
2. Für typische Verfahren (z.B. GLMs oder Boosted Trees) sollen die synthetischen Daten zu analogen Ergebnissen wie die Originaldaten führen.

Privacy:

- Das Aktuariat möchte «nicht zu ähnlich» mit einem Ähnlichkeitsmass objektivieren.

Fidelity:

- Keine zusätzlichen Anforderungen, das folgt einfach aus dem Kompromiss zwischen «Utility» und «Privacy».

Utility: Funktioniert!

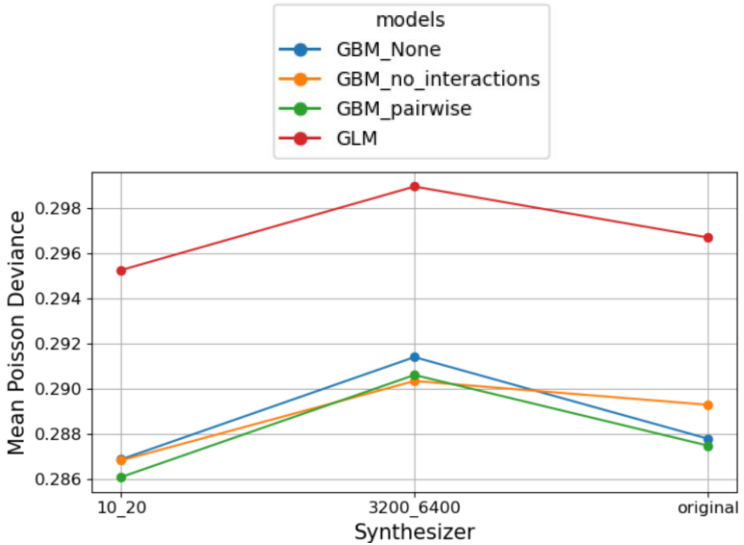
Vorgehen: «Train on Synthetic, Test on Original»

1. Split der Originaldaten in «Train» und «Test»
2. Trainiere Modelle auf Train (wie immer)
3. Generiere synthetische Daten «Synth» aus Train
4. Trainiere einen zweiten Satz von Modellen auf Synth
5. Evaluiere Modelle aus 2. und 4. auf Test

Fazit Utility:

- Loss in ähnlicher Grössenordnung
- Reihenfolge der Modelle vergleichbar

Loss (kleiner ist besser) der vier Modelle



Privacy light: «Nicht zu ähnlich»

Problem:

- Synthesiser sollen neue Datensätze «erfinden», gleichzeitig sollen die Daten aber «genau wie die Originaldaten sein»?
- Lösung: Generiere Daten, die so divers sind wie die Originaldaten!
- Algorithmus:
  - Teile Originaldaten zufällig in zwei gleich grosse Blöcke und zähle die gemeinsamen Datensätze
  - Wiederhole dies mehrfach
  - Ergebnis: Ähnlichkeit der Originaldaten
  - Ersetze nun im Algorithmus oben jeweils einen der beiden Blöcke durch synthetische Daten
  - Ergebnis: Ähnlichkeit synthetischer zu Originaldaten

Ergebnis:

Mittlere Anzahl gemeinsamer Datensätze in zwei Blöcken

	Count	
original	15872	Original zu Original
synth	8674	Synthetisch zu Synthetisch
mixed	8648	Synthetisch zu Original

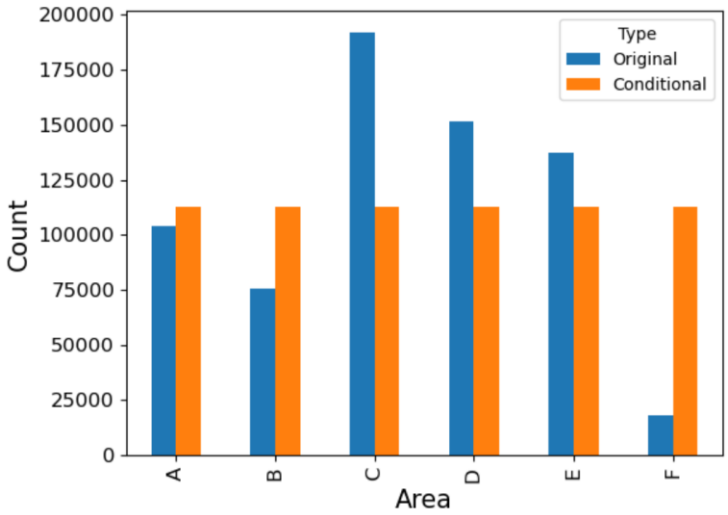
Fazit:

- Die Daten sind nicht zu ähnlich
- Im Gegenteil: Der Synthesiser ist eher zu divers

Anwendungsfall: Gezielte Generierung

- Es gibt nun doch neue Bedenken beim Vorstand. Aus den synthetischen Daten lässt sich die regionale Verteilung des Portfolios sehr genau ablesen. Aus Wettbewerbsgründen möchte man dies vermeiden.
- Die Lösung: Bedingte Synthese
  - Man generiert bedingt auf vorgegebene Attribute
  - Hier ist das «Area».
  - Die Samples folgen der Originalverteilung aber bedingt auf die vorgegebenen(!) Attribute
- Die Kalibrierung des Generators erfolgt dabei auf allen Daten. Nur das Sampling wird eingeschränkt. Dadurch bleibt die Qualität auch bei kleinen Beständen (z.B. Area F) gut.

Im bedingten Sample kommt jede Area genau 113’000-mal vor



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

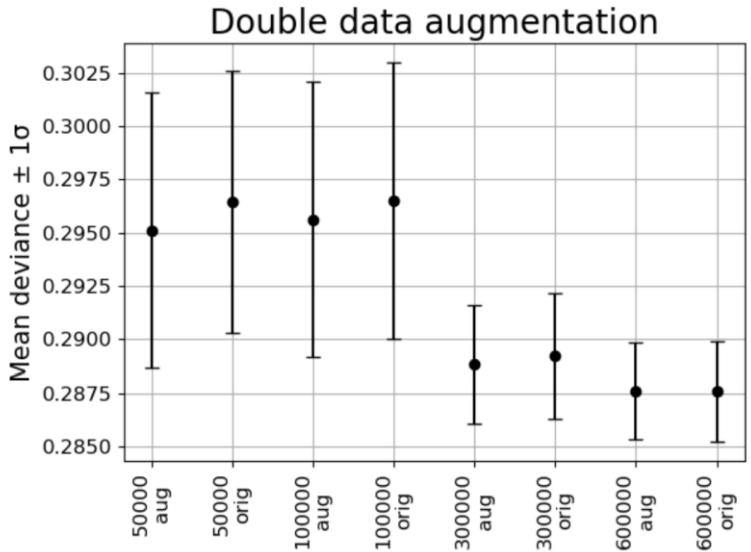
Anwendungsfall: Augmentation

- Augmentation (Anreicherung) ist die gezielte Vermehrung der Trainingsdaten durch zusätzliche, hier synthetische, Daten zur Verbesserung der Generalisierung
- Dies ist einer der wichtigsten Anwendungsfälle für synthetische Daten. Im speziellen Fall aktuarieller Anwendungen ist die Nützlichkeit aber nicht klar.
- Aktuarielle Anwendungen unterscheiden sich vom typischen Fall der Augmentation

Unterschied in	Aktuariell	Typisch
Modellen	GLMs, Gradient Boosting	Sehr(!) tiefe neuronale Netze
Daten	Tabellarisch, ohne Symmetrien	Symmetrien wie Rotation, Translation
Vorhersagen	Wahrscheinlichkeiten, Frequenzen	Klassifikation

Test von Augmentation

- Vorgehen im Experiment
  1. Split in Train und Test, generiere Synth aus Train
  2. Kalibriere ein Modell auf Train
  3. Kalibriere ein Modell auf Train und Synth
  4. Vergleiche beide Modelle auf Test
- Verschieden Grössen für das Sample aus 1.
- Mehrfache bootstrap Wiederholungen, um Variabilität der Ergebnisse abzuschätzen
- **Fazit**: Für diesen Datensatz und diesen Generator bringt diese Art der Augmentation nichts.



Modell: HistGradientBoostingRegressor aus scikit-learn

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Bisherige Ergebnisse

Alle Ergebnisse werden auf dem DAV-Github veröffentlicht und sind frei verfügbar!

Handbuch:

- Einführung in die Synthese mit synthpop

Synthomat:

- Sourcen des Generators

Tool comparison:

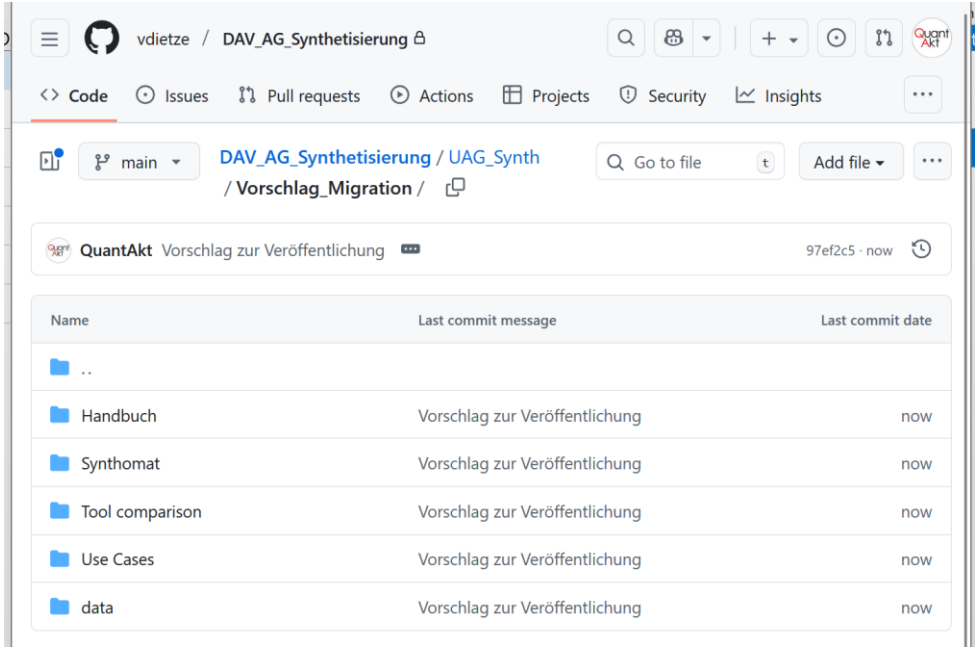
- Die Notebooks zum Toolvergleich

Use Cases:

- Die Notebooks zu den Anwendungsfällen

Data:

- Die verwendeten Datensätze



Fragen, Vorschläge oder Bemerkungen?

Auch nach dem Vortrag gerne per E-Mail an:

- maiurban@deloitte.de
- guido.gruetzner@quantakt.com

oder via LinkedIn

- www.linkedin.com/in/dr-maik-urban
- www.linkedin.com/in/guido-gruetzner