

Stand 2023-04-21

Demonstrator Synthetische Daten – Einblick in Anwendungen und Eigenschaften

Dorian Wachsmann, Jens Tiemann, Jan Dennis Gumz, Fabian Manzke

Gefördert durch:



Übersicht

Was sind synthetische Daten?

- Herausforderung & Lösungsansatz
- Algorithmen im Demonstrator

3 Zugänge zum Demonstrator

- schützenswerte Daten anonymisieren
- synthetische mit realen Daten vergleichen
- Illustration von Re-Identifizierbarkeit

Detail zum Demonstrator

Siehe auch Blogbeitrag:

Ort	Standort	Art	Jahr	Höhe	Umfang
Mitte	Washingtonplatz	Japanischer Schnurbaum	2019	24cm	5m
Treptow-Köpenick	-	Pyramiden-Hainbuche	2018	20cm	5m

5 Beispieldaten

Zufällige Auswahl

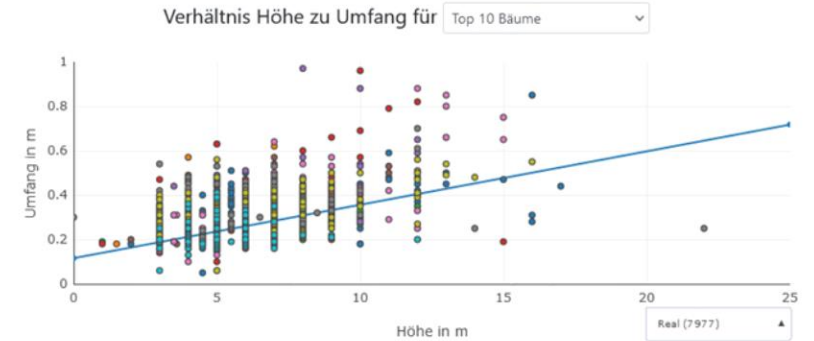
2. Wählen Sie eine Methode zum Generieren der synthetischen Daten:

Gaussian Copula

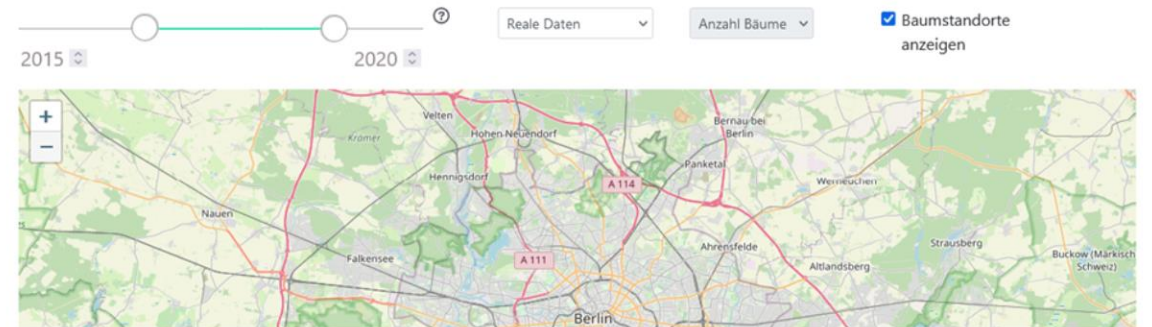
Methode	Score (KSCompl.)	MLP Classifier
Reale Daten	100%	0.337
Gaussian Copula	86.25%	0.093
GAN	77.5%	0.267
AutoEncoder	86.09%	0.218

Durch das Zählen des Auftretens der einzelnen Werte lässt sich über jede Reihe & Spalte eine sogenannte *Randhäufigkeit* berechnen. Typischerweise gibt es mehrere solcher Randverteilungen, eine Copula stellt einen funktionalen Zusammenhang zwischen diesen auf und kann auf diese Weise stochastische Abhängigkeiten modellieren. Anders gesagt lassen sich mit der Methode Rückschlüsse auf die Art der stochastischen Abhängigkeit zweier Zufallsvariablen machen.

3. Ausgewählte Szenarien zum qualitativen Vergleich der Daten



4. Datenpunkte visualisieren und aggregieren



Abschnitt 1

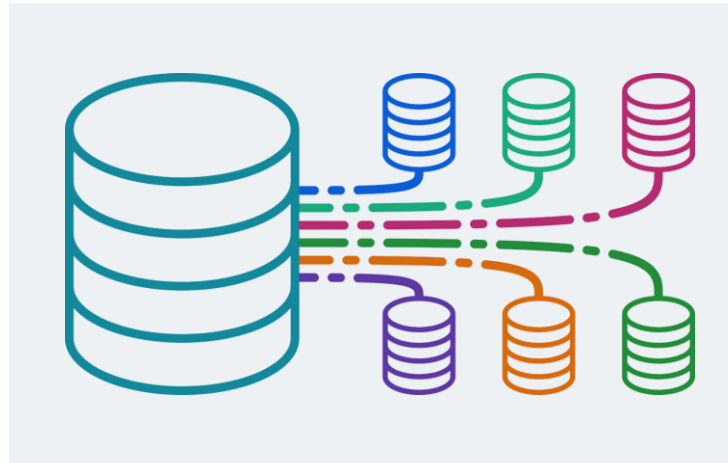
Theorie: Was sind synthetische Daten?

Was sind synthetische Daten?



Herausforderungen im Umgang mit Daten

- enge Grenzen bei Nutzung personenbezogener Daten
- sensible / schützenswerte Daten in allen Bereichen



Daten als (begrenzte) Ressource

- ggf. eingeschränkte Verfügbarkeit
- ggf. Unvollständigkeit, Stichwort Fairness
- multiple Quellen nötig?



Herausforderung des „Daten-Labeling“

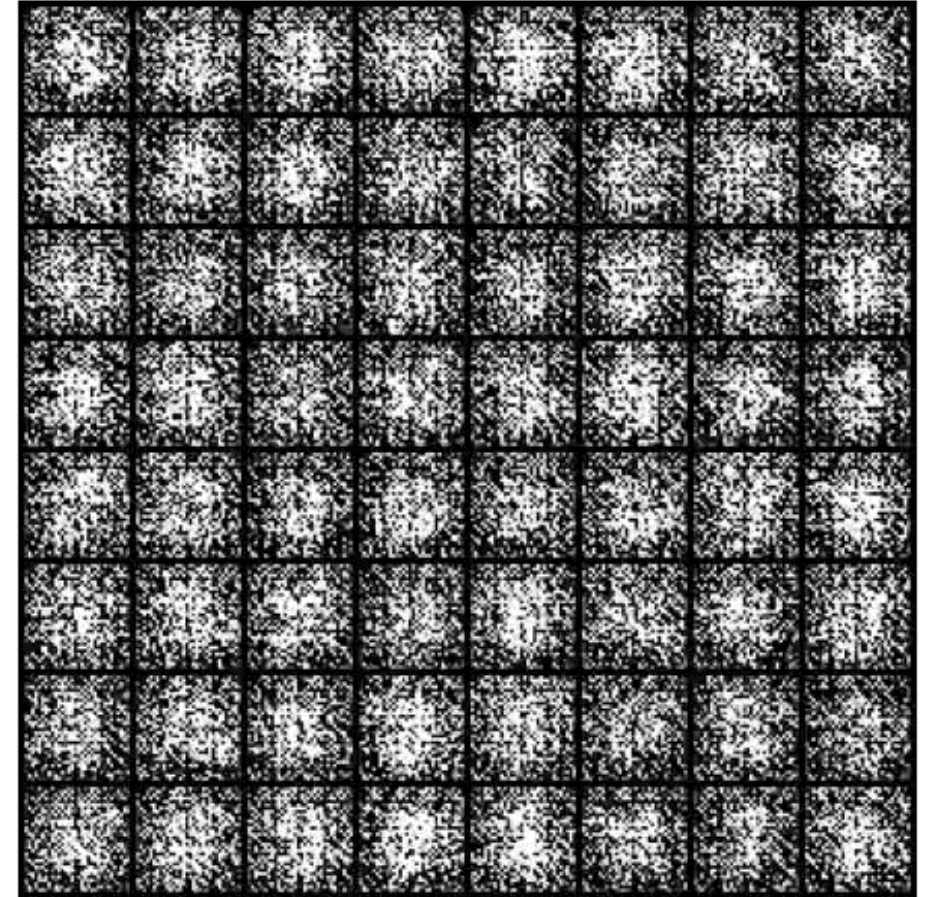
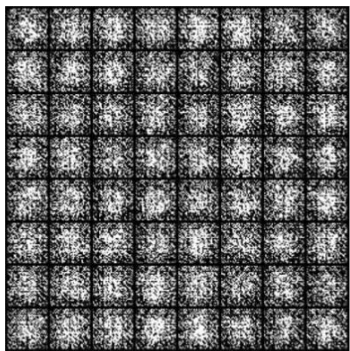
- Überwachtes Lernen von KI Modellen benötigt „gelabelte“ Daten: Woher kommen die Labels, »wer tut die Arbeit«?

Was sind synthetische Daten?

Lösungsansatz

Synthetische Daten

- sind „künstlich“ erzeugt, also algorithmisch
- sollten idealerweise Struktur & Eigenschaften realer Daten abbilden
 - abhängig von Ziel der Nutzung und verwendeter Methode
- Realdaten bleiben geschützt
 - Schutz von Individuen bzw. Datenpunkten

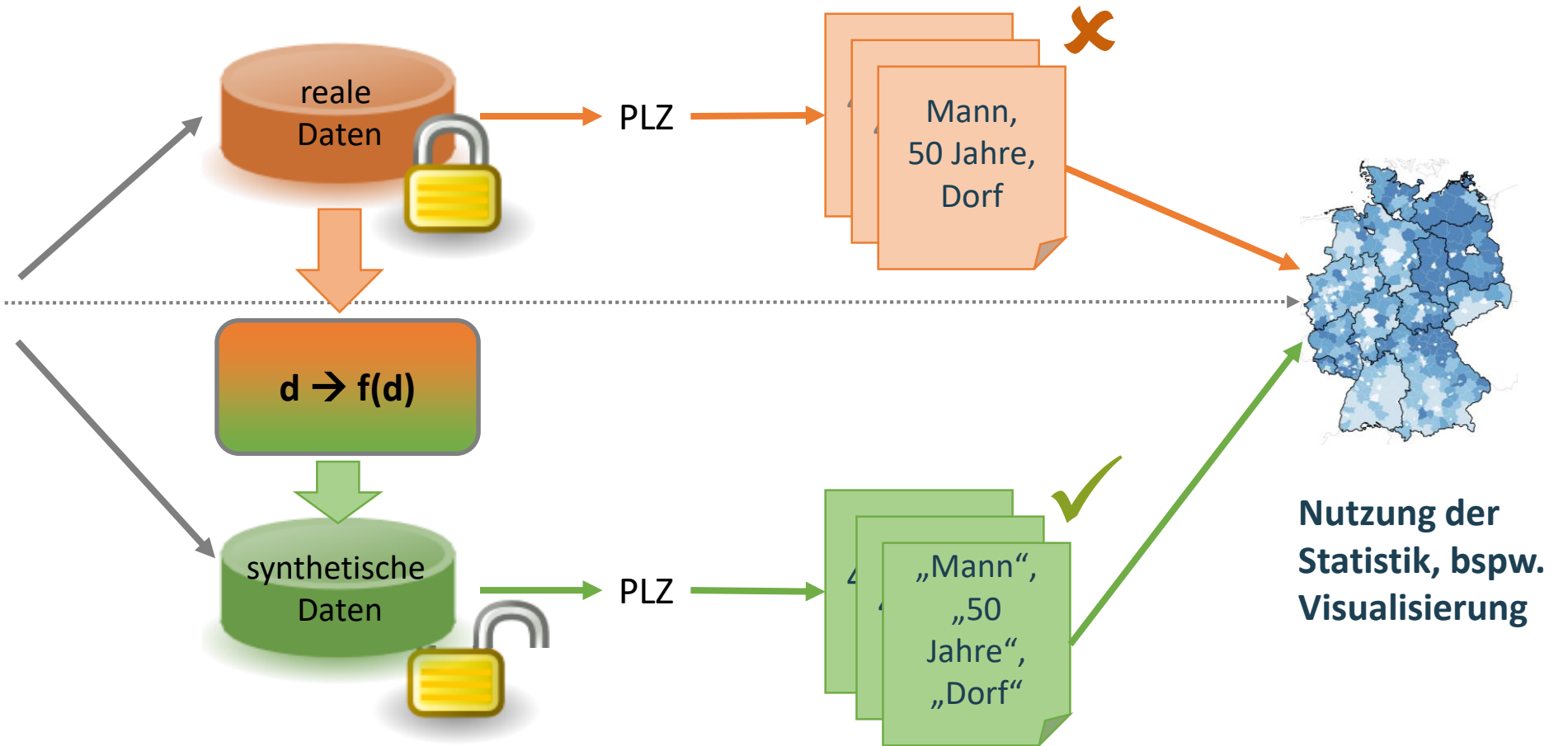


(Video)

Altersverteilung nutzen, ohne mit realen Datensätzen arbeiten zu müssen

Beispiel zur Verwendung synthetischer Daten

Fragestellung:
Wohnorte („männlich,
45-64 Jahre“)



Abschnitt 2

Der Demonstrator: Methoden und Zugänge

Demonstrator gibt ersten Einblick

Einleitung

reale Daten als Basis → Erzeugung synthetischer Daten

- statistische Eigenschaften der Originaldaten bestmöglich abbilden

Genutzte Datenbasis: Baumbestand Berlin

- Annahme: Eigenschaften von Bäumen enthalten sensible Informationen
- vergleichbar mit personenbezogenen Daten z. B. Pflanzjahr := Geburtsjahr

Anmerkung:

Die hier vorgestellten Verfahren dienen der Illustration und müssen für konkrete Anwendungen angepasst werden



Demonstrator gibt ersten Einblick

Was ist unser Ansatz?

reale Daten als Basis → Erzeugung synthetischer Daten

- „Anonymisierung“ durch Vergrößerung
- 3 generative Methoden:
 - Gaussian Copula
 - Generative Adversariale Netze
 - Variationeller Autoencoder

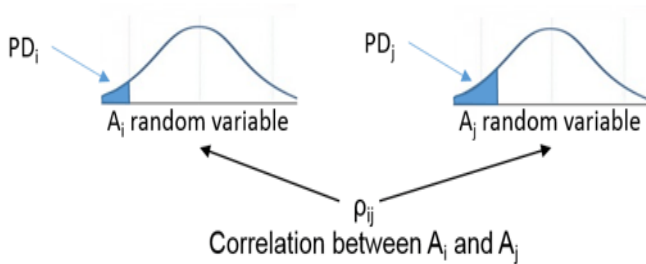
Jede Methode hat Vorteile und Nachteile,
der Demonstrator ermöglicht Vergleich
anhand einer praktischen Anwendung



Generierung synthetischer Daten

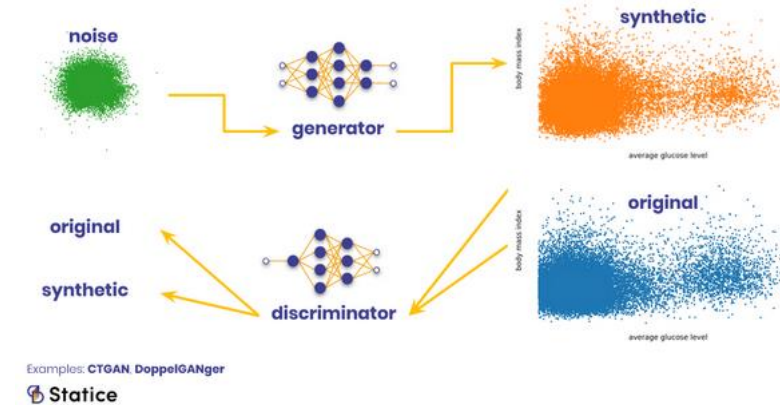
Allen Methoden gemeinsam: Rekonstruktion der unbekanntenen Wahrscheinlichkeitsverteilung, die den Originaldaten zugrunde liegt

Gaussian Copula



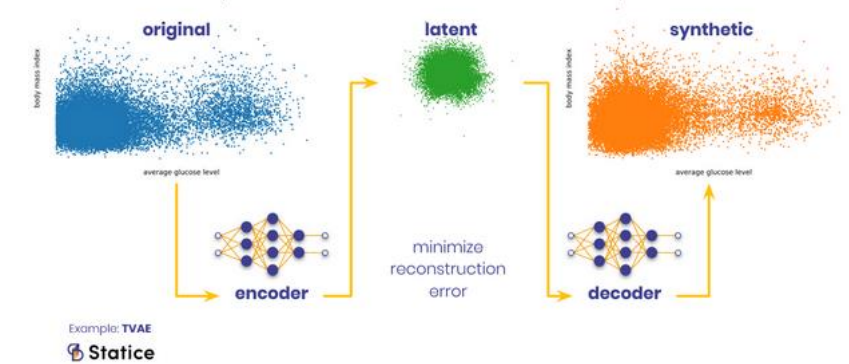
- Funktionale Zusammenhänge zwischen Randverteilungen
- Modellierung stochastischer Abhängigkeiten

Generative Adversariale Netze



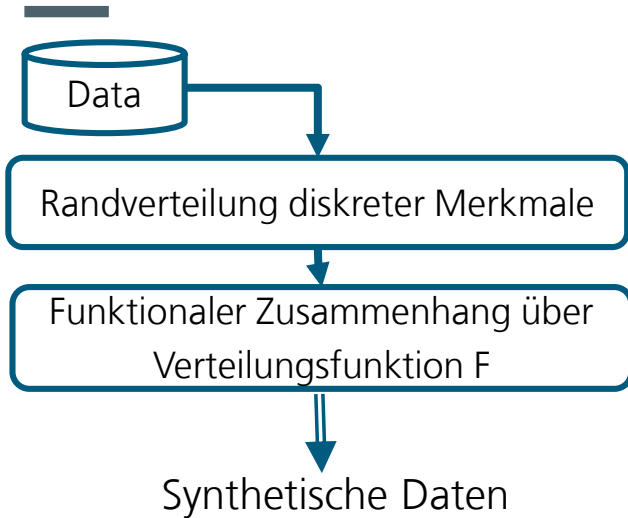
- Deep Learning Methode
- Spieltheoretischer Ansatz: Generator vs. Diskriminator
- Nachteil: Relative kompliziertes Training

Variationeller Autoencoder



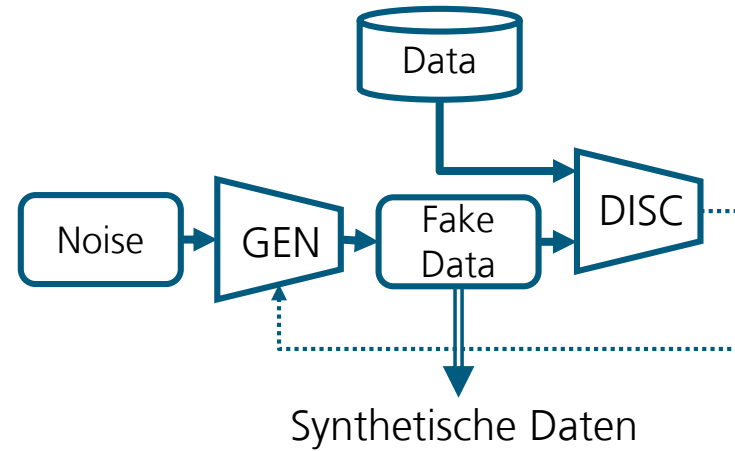
- Deep Learning Methode
- Encoder kodiert Basisdaten, Decoder decodiert Daten → lernt über Ähnlichkeit
- Daten Neu-Generierung durch Variieren des latent-space (Input des Decoders)

Generierung synthetischer Daten



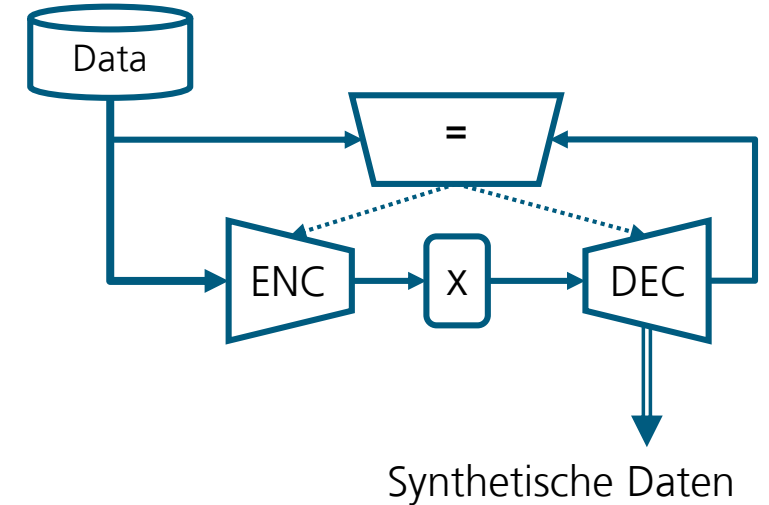
Gaussian Copula

- Sampling von der (multivariaten) Verteilungsfunktion F zur Datengenerierung



Generative Adversariale Netze

- Deep Learning Methode
- Spieltheoretischer Ansatz: Generator (GEN) vs. Diskriminator (DISC)



Variationeller Autoencoder

- Deep Learning Methode
- Idee Training: ENC kodiert Data d zu x , DEC decodiert x zu $\sim d$. Abstandsmaß d zu $\sim d$ definiert den Loss
- Idee Generierung: wenn x ausreichend strukturiert, sample random x und nutze DEC für Generierung

Erster Zugang: Klassische Anonymisierung

2. Abschnitt: Wählen Sie eine Methode zum Generieren der synthetischen Daten:

Anonymisiert

Methode	Score (KSCompl.) ?	MLP Classifier ?	Download
Reale Daten	100%	0.337	
Gaussian Copula	86.25%	0.093	
GAN	77.5%	0.267	
AutoEncoder	86.09%	0.218	

last_name	first_name
Smith	Bob
Doe	Jane
King	Stephen
Savage	Randal
Downer	Debbie

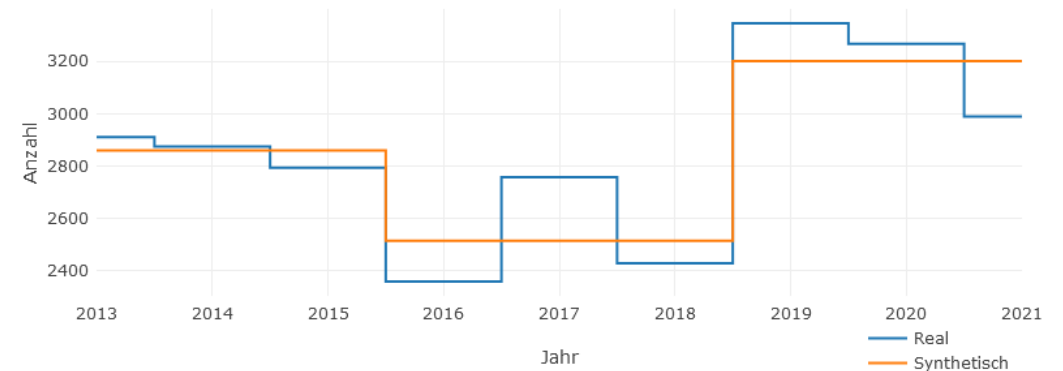
Anonymisieren

klassische Methode zum Schutz sensibler Informationen: Vergrößerung

- Effekt: Individuum „verstecken“ in einer Gruppe

☞ Auswahl der Methode »Anonymisiert« im 2. Abschnitt

Gepflanzte über Zeit ?



☞☞ Auswahl „Gepflanzte Bäume über Zeit“ im 3. Abschnitt

Demonstrator zeigt im 3. Abschnitt den Vergleich realer und – in diesem Fall – einfach anonymisierten Daten

- Wirkung: jahresgenaue Auflösung ist nicht mehr möglich
- Nachteil: Auswertung entsprechend ungenau unmöglich

Erster Zugang: Anonymisierung mit synthetischen Daten



2. Abschnitt: Wählen Sie eine Methode zum Generieren der synthetischen Daten:

Gaussian Copula

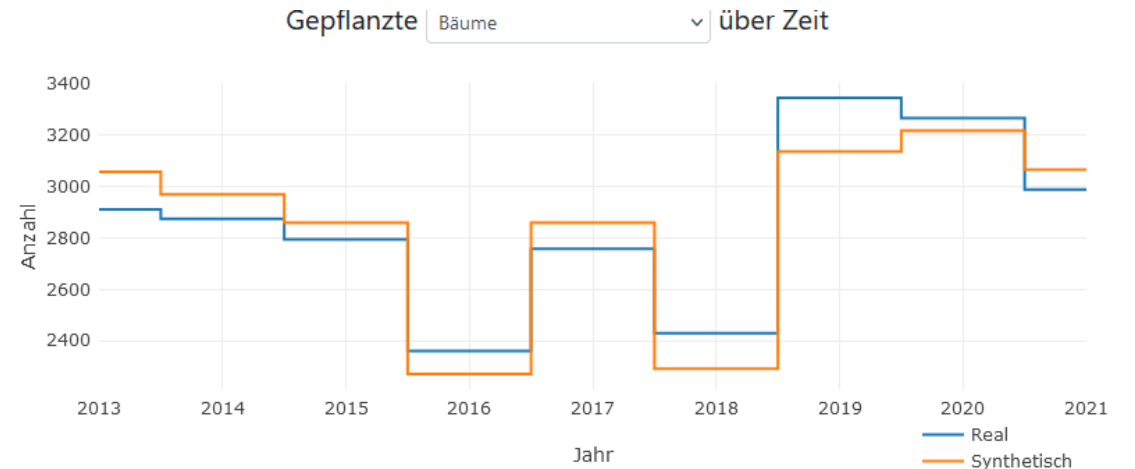
Methode	Score (KSCmpl.) ?	MLP Classifier ?	Download
Reale Daten	100%	0.337	
Gaussian Copula	86.25%	0.093	
GAN	77.5%	0.267	
AutoEncoder	86.09%	0.218	

Durch das Zählen einer sogenannten Randverteilung auf und kann anlassen sich mit zweier Zufallsva

Ansatz zum Schutz sensibler Daten:
Generierung künstlicher Daten mit ähnlichen statistischen Eigenschaften

- Effekt: idealerweise sind die Datenobjekte komplett künstlich und lassen keine Rückschlüsse auf reale Echtdaten zu

[Wechsel der Methode im 2. Abschnitt zu »Gaussian Copula«](#)



Hier: genaue Darstellung der Eigenschaft „Pflanzjahr“

Verschiedene Methoden zeigen eine mehr oder weniger gute Annäherung an die Eigenschaft „Jahr“

Zweiter Zugang: Vergleichbarkeit Synthetische Daten & Echtdaten

2. Abschnitt: Wählen Sie eine Methode zum Generieren der synthetischen Daten:



Gaussian Copula

Methode	Score (KSCmpl.) ?	MLP Classifier ?	Download
Reale Daten	100%	0.337	
Gaussian Copula	86.25%	0.093	
GAN	77.5%	0.267	
AutoEncoder	86.09%	0.218	

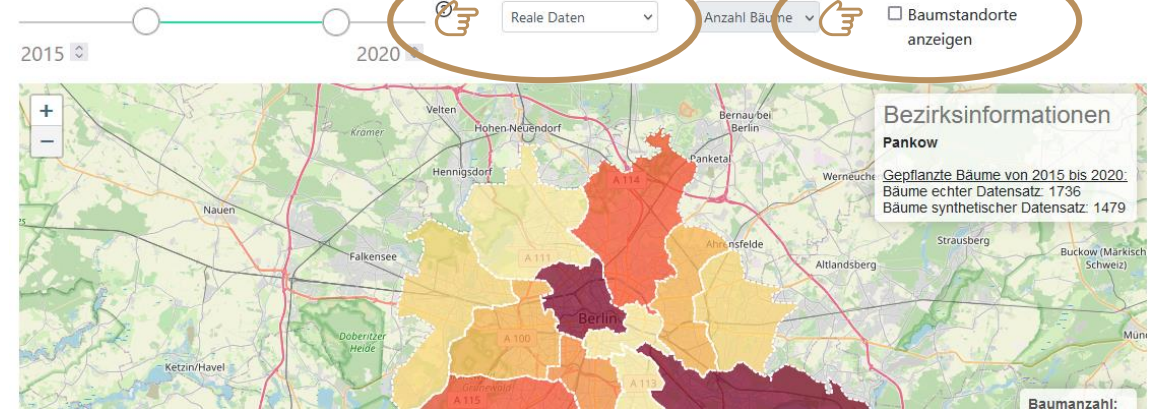
Durch das Zählen einer sogenannten Randverteilung auf und kann lassen sich mit zweier Zufalls

Nutzung einer Methode zur Generierung synthetischer Daten:

- Gaus
- GAN
- Auto

☞ im 2. Abschnitt »Gaussian Copula« wählen

4. Datenpunkte visualisieren und aggregieren

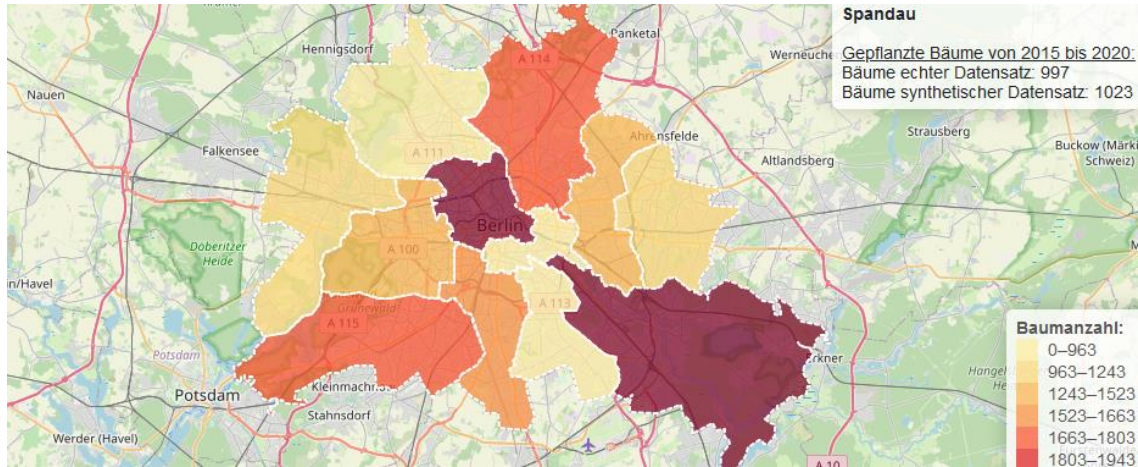


Karte zeigt im 4. Abschnitt reale Daten: Anzahl der Bäume pro Bezirk der Pflanzjahre 2015 bis 2020

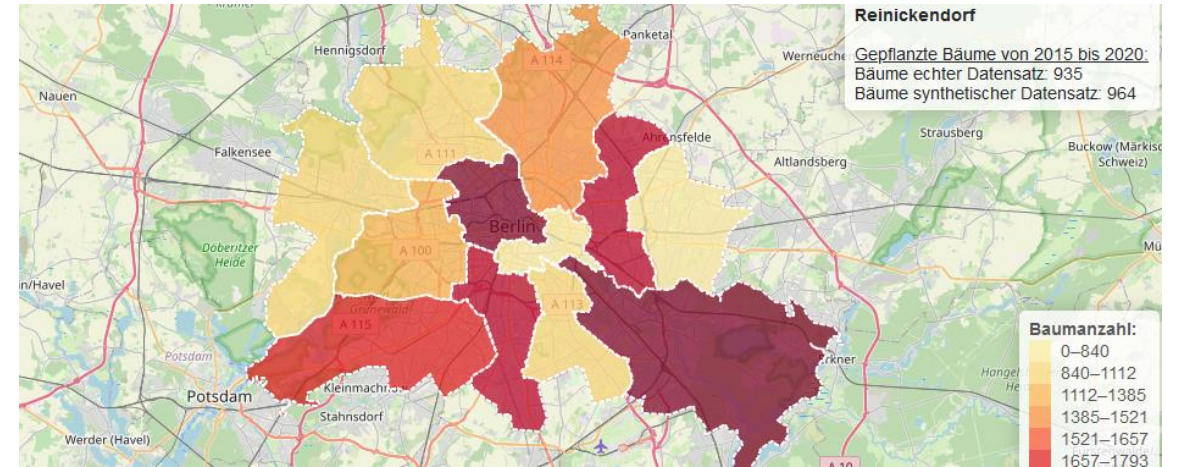
Bedienelemente:

☞ »reale / synthetische Daten« ☞ »Baumstandorte anzeigen«

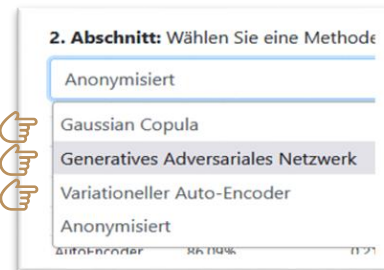
Zweiter Zugang: Vergleichbarkeit mit realen Daten – aggregierte Datensätze



☞ Einstellung: reale Daten

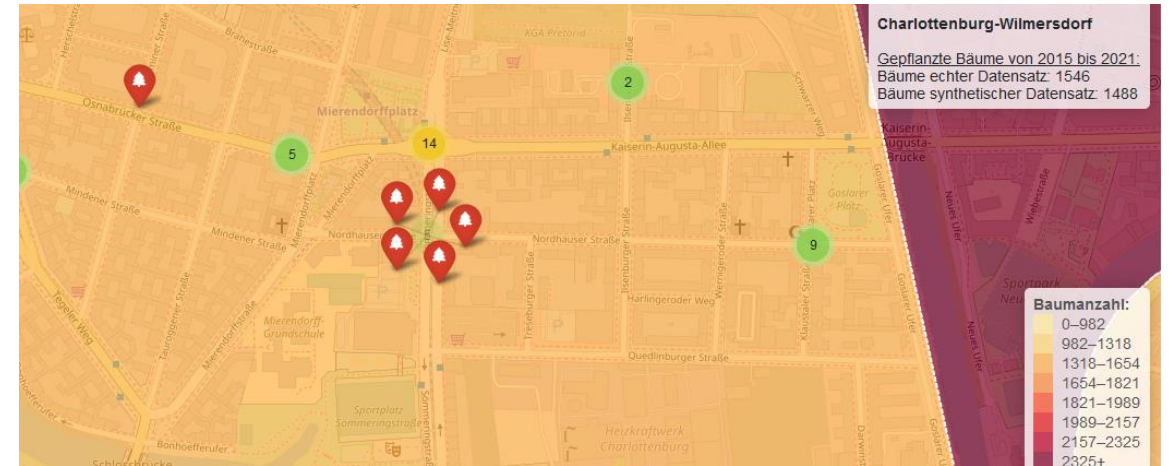
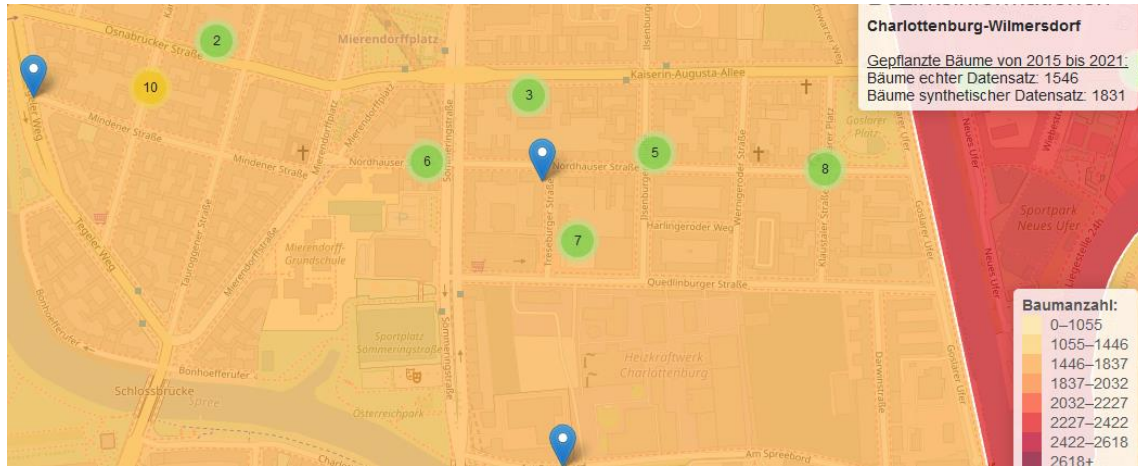


☞ Einstellung: synthetische Daten



Wechsel der Methode im 2. Abschnitt zeigt unterschiedlich ähnliche Verteilungen in Bezug auf reale Daten

Zweiter Zugang: Vergleichbarkeit mit realen Daten – einzelne Bäume



☞ Einstellung: Baumstandorte anzeigen

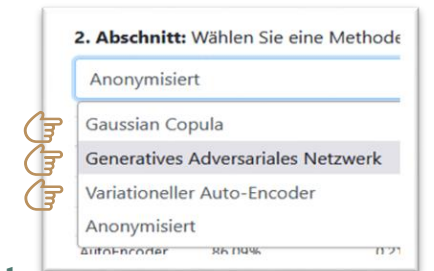
☞ Einstellung: reale Daten

Anmerkung: Die Datenbasis enthält nur Bäume, die in den letzten 10 Jahren gepflanzt wurden, geringe Anzahl von Standorten.

☞ Einstellung: synthetische Daten

Clusterbildung aufgrund von Reverse Geocoding

Wechsel der Methode im 2. Abschnitt zeigt unterschiedlich ähnliche Verteilungen in Bezug auf reale Daten



Dritter Zugang: Illustration der Re-Identifizierbarkeit

1. Abschnitt: Wählen Sie eine Datenmenge: (vorerst nur Baumdaten 2022)

Berliner Baumdaten 2022

Daten 25958 Berliner Bäume
Erhebung 2022
Infos Sachdaten zum Baumbestand - Straßenbäume - mit Angaben zur Baumart, Adresse, Pflanzjahr, Höhe etc. für alle in den letzten 10 Jahren gepflanzten Bäume: [Mehr Infos hier](#)

Bezirk	Str.Name	Baumart	Pflanzjahr	Umfang	Höhe
Pankow	Blankenburger Straße	Kaiser-Linde	2014	32cm	3m
Neukölln	Gerlachsheimer Weg	Birne	2017	23cm	Keine Angabe
Treptow-Köpenick	Puschkinallee	Erle	2016	26cm	5m
Treptow-Köpenick	Am Falkenberg	Schwarz-Ahorn 'Faassen's Black'	2021	20cm	6m
Mitte	Neue Hochstraße	Esche	2016	20cm	6m

5 Beispieldaten Zufällige Auswahl

Zur Datenbasis im 1. Abschnitt werden zur Illustration der Daten zufällig ausgewählte Datensätze angezeigt

- Datenbasis enthält Bezirk, Straße, Baumart, Pflanzjahr usw.

➔ Auswahl der Methode »Auto Encode« im 2. Abschnitt

2. Wählen Sie eine Methode zum Generieren der synthetischen Daten:

Variationeller Auto-Encoder

Methode	Score (KSCmpl.)	MLP Classifier
Reale Daten	100%	0.337
Gaussian Copula	86.25%	0.093
GAN	77.5%	0.267
AutoEncoder	86.09%	0.218

Variationelle Autoencoder bestehen aus zwei Komponenten, 1 Encoder und einem Decoder. Der Encoder transformiert c einfachere Verteilung, eine sogenannte *latent distribution*. Da versucht diese Verteilung in den Ursprungsraum zurück zu tran Daten sind die synthetischen Daten und können durch Ar Anwendungsszenario angepasst werden.

5. Datensätze vergleichen & filtern. Anhand weniger Parameter lassen sich schnell eindeutige Datenpunkte erfassen

Bezirke Baumart Pflanzjahr

Echtdatenpunkte: 25958 Fakedatenpunkte: 25958

ID	Baumname	Umfang	Höhe
0	Schwarz-Erle	37cm	8m
1	Erle	49cm	8m
2	Erle	48cm	9m
3	Erle	26cm	8m
4	Ahornblättrige Platane	38cm	10m
5	Berg-Ahorn, Weiss-Ahorn	19cm	4m
6	Spitz-Ahorn	36cm	3m
7	Amberbaum	33cm	Keine Angabe
8	Kaiser-Linde	24cm	7m
9	Kirsche	23cm	Keine Angabe

ID	Baumname	Umfang	Höhe
0	Winter-Linde 'Greenspire'	20cm	4m
1	Säulen-Weiss-Dorn	20cm	5m
2	Sumpf-Eiche	29cm	5m
3	Eisenholzbaum, Parrotie	0cm	0m
4	Spitz-Ahorn	25cm	0m
5	Kaiser-Linde	34cm	7m
6	Japanischer Schnurbaum	39cm	8.52m
7	Amberbaum	26cm	5m
8	Winter-Linde 'Greenspire'	20cm	5m
9	Winter-Linde 'Greenspire'	34cm	7m

Im 5. Abschnitt werden Einträge der realen und der synthetischen Datenbasen miteinander verglichen und können gefiltert werden

Dritter Zugang: Illustration der Re-Identifizierbarkeit

5. Datensätze vergleichen & filtern. Anhand weniger Parameter lassen sich schnell eindeutige Datenpunkte erfassen

Bezirke Baumart Pflanzjahr

Echtdatenpunkte: 25958 Fakedatenpunkte: 25958

ID	Baumname	Umfang	Höhe
0	Schwarz-Erle	37cm	8m
1	Erle	49cm	8m
2	Erle	48cm	9m
3	Erle	26cm	8m
4	Ahornblättrige Platane	38cm	10m
5	Berg-Ahorn. Weiss-Ahorn	19cm	4m
6	Spitz-Ahorn	36cm	3m
7	Amberbaum	33cm	Keine Angabe
8	Kaiser-Linde	24cm	7m
9	Kirsche	23cm	Keine Angabe

ID	Baumname	Umfang	Höhe
0	Winter-Linde 'Greenspire'	20cm	4m
1	Säulen-Weiss-Dorn	20cm	5m
2	Sumpf-Eiche	29cm	5m
3	Eisenholzbaum, Parrotie	0cm	0m
4	Spitz-Ahorn	25cm	0m
5	Kaiser-Linde	34cm	7m
6	Japanischer Schnurbaum	39cm	8.52m
7	Amberbaum	26cm	5m
8	Winter-Linde 'Greenspire'	20cm	5m
9	Winter-Linde 'Greenspire'	34cm	7m

5. Datensätze vergleichen & filtern. Anhand weniger Parameter lassen sich schnell eindeutige Datenpunkte erfassen

Spandau Kirsche Pflanzjahr

Echtdatenpunkte: 3 Fakedatenpunkte: 3

ID	Baumname	Umfang	Höhe
5383	Japanische Zierkirsche 'Kanzan'	47cm	4m
5761	Scharlach-Kirsche	25cm	5m
5769	Scharlach-Kirsche	25cm	8m

ID	Baumname	Umfang	Höhe
12551	Japanische Zierkirsche 'Kanzan'	22cm	3m
13635	Spiegelrinden-Kirsche	21cm	5m
16875	Kirsche	20cm	3m

Anhand von bekannten Merkmalen lässt sich ein Objekt „einkreisen“ und schließlich identifizieren

Beispiel: Seltener Nachname in einer kleinen Stadt – Person kann anhand dieser Informationen identifiziert werden

Synthetische Daten: keine reale Personen, sondern realistische statistische Eigenschaften nach festgelegten Kriterien

legt man einen Bezirk und eine (seltene) Baumart fest, reduziert sich die Anzahl infrage kommender Bäume schnell

☞ Auswahl Bezirk »Spandau« und Baumart »Kirsche«

Listen zeigen ähnliche Bäume, aber die synthetischen Bäume existieren nicht

Abschnitt 3

Weiterführendes

Nutzung entlang der Abschnitte

Übersicht Demonstrator

1. Abschnitt: Datenmenge

- vorerst nur Baumdaten 2022

2. Abschnitt: Methode zum Generieren der synthetischen Daten

- Gaussian Copula
- Generative Adversariale Netze (GAN)
- Variationelle Autoencoder
- Anonymisierung – zum Vergleich

3. Abschnitt: Szenarien zum qualitativen Vergleich der Daten

- Verschiedene statistische Auswertungen, Vergleich mit Realdaten

4. Abschnitt: Datenpunkte visualisieren und aggregieren

- Reale oder synthetische Datenpunkte auf Karten

5. Abschnitt: Datensätze vergleichen und filtern

- Filter nach Eigenschaften kann Datenpunkte effektiv eingrenzen

Demonstrator Synthetische Daten

Synthetische Datensätze bestehen aus künstlichen, meist durch spezielle Algorithmen generierte Daten. Sie finden Anwendung in vielen sensiblen Domänen, wo Originaldaten aufgrund von Datenschutz und Privatsphäre nicht veröffentlicht werden dürfen, synthetische Daten jedoch ohne negative Auswirkungen für Betroffenen verfügbar gemacht werden können. Der von uns entwickelte interaktive Demonstrator bietet die Möglichkeit, Potentiale und Schwächen synthetischer Daten auf spielerische Weise eigenständig zu entdecken.

Infoleuten:

Lerne den Demonstrator kennen Welche Methode performt am »besten«?

Dafür werden wir drei verschiedenen Methoden auf Echtdaten an. Da wir selbstverständlich nicht mit personenbezogenen Originaldaten arbeiten dürfen, nutzen wir stattdessen einen Datensatz von gepflanzten Bäumen der letzten 10 Jahre in Berlin. Bei der Auswertungssetzung mit den Daten wird klar, dass diese vergleichbare Eigenschaften besitzen, datenschutztechnisch jedoch unbedenklich sind. Beispielsweise lässt sich das Pflanzjahr des Baumes wie das Geburtsjahr einer Person lesen, oder die Höhe des Baumes wie das jährliche Einkommen. Wir wünschen Ihnen viel Spaß beim Ausprobieren!

1. Abschnitt: Wählen Sie eine Datenmenge (vorerst nur Baumdaten 2022)

Berliner Baumdaten 2022

Daten 23958 Berliner Bäume
Erhebung 2022
Infos Sachdaten zum Baumbestand - Straßenbäume - mit Angaben zur Baumart, Adresse, Pflanzjahr, Höhe etc. für alle in den letzten 10 Jahren gepflanzten Bäume. Mehr Infos hier

Bezirk	StrName	Baumart	Pflanzjahr	Umfang	Höhe
Treptow-Köpenick	Peter-Hille-Straße	Rosa-Ulm 'New Horizon'	2019	18cm	6m
Marzahn-Hellersdorf	Clara-Immensee-Straße	Silber-Linde	2018	28cm	4m
Pankow	Schulstraße	Winter-Linde 'Greengirl'	2014	44cm	8m
Steglitz-Zehlendorf	Neurospinger Straße	Baum-Hassel	2020	18cm	Keine Angabe
Treptow-Köpenick	Wilhelmshofstraße	Gewöhnliche Hageleibche	2014	20cm	7m

5 Beispieldaten:

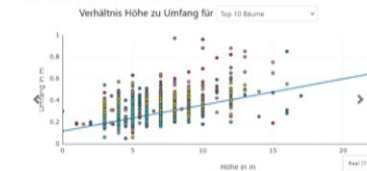
2. Abschnitt: Wählen Sie eine Methode zum Generieren der synthetischen Daten:

Gaussian Copula

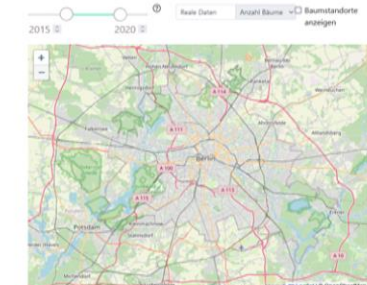
Methode	Score	MIP	Qualität	Datenart
Reale Daten	100%	0.337	B	Reale Daten
Gaussian Copula	98.23%	0.693	B	Synthetische Daten
GAN	77.3%	0.267	B	Synthetische Daten
Autoencoder	55.05%	0.215	B	Synthetische Daten

Durch das Zählen des Auftretens der einzelnen Werte lässt sich über jede Reihe & Spalte eine sogenannte Ähnlichkeitsmatrix berechnen. Typischerweise gibt es mehrere solcher Randverteilungen, eine Copula stellt einen funktionalen Zusammenhang zwischen diesen auf und kann auf diese Weise stochastische Abhängigkeiten modellieren. Anders gesagt lassen sich mit der Methode Rückschlüsse auf die Art der stochastischen Abhängigkeit zweier Zufallsvariablen machen.

3. Abschnitt: Ausgewählte Szenarien zum qualitativen Vergleich der Daten



4. Abschnitt: Datenpunkte visualisieren und aggregieren



5. Abschnitt: Datensätze vergleichen & filtern. Anhand weniger Parameter lassen sich schnell eindeutige Datenpunkte erfassen

Bezirk: Echtdatenpunkte: 23958 Baumart: Synthetische Datenpunkte: 23958 Pflanzjahr:

ID	Baumname	Umfang	Höhe	ID	Baumname	Umfang	Höhe
0	Schwarz-Erle	37cm	8m	0	Baum-Hassel	22cm	3,67m
1	Erle	49cm	8m	1	Chinesischer Waldahorn	1cm	2,15m

<https://www.oeffentliche-it.de/>

Kompetenzzentrum Öffentliche IT

Öffentliche IT

- IT im öffentlichen Raum
- IT der öffentlichen Verwaltung

Kompetenzzentrum Öffentliche IT (ÖFIT)

- angesiedelt am Fraunhofer-Institut für offene Kommunikationssysteme (FOKUS)
- gefördert vom Bundesministerium des Innern und für Heimat (BMI)

Demonstrator Synthetische Daten

<https://www.oeffentliche-it.de/werkstatt/synthetische-daten-demonstrator>



Navigation



Aktuell



Blog



Veranstaltungen



Trendschau



Publikationen



Umfragen



Werkstatt



Deutschland-Index

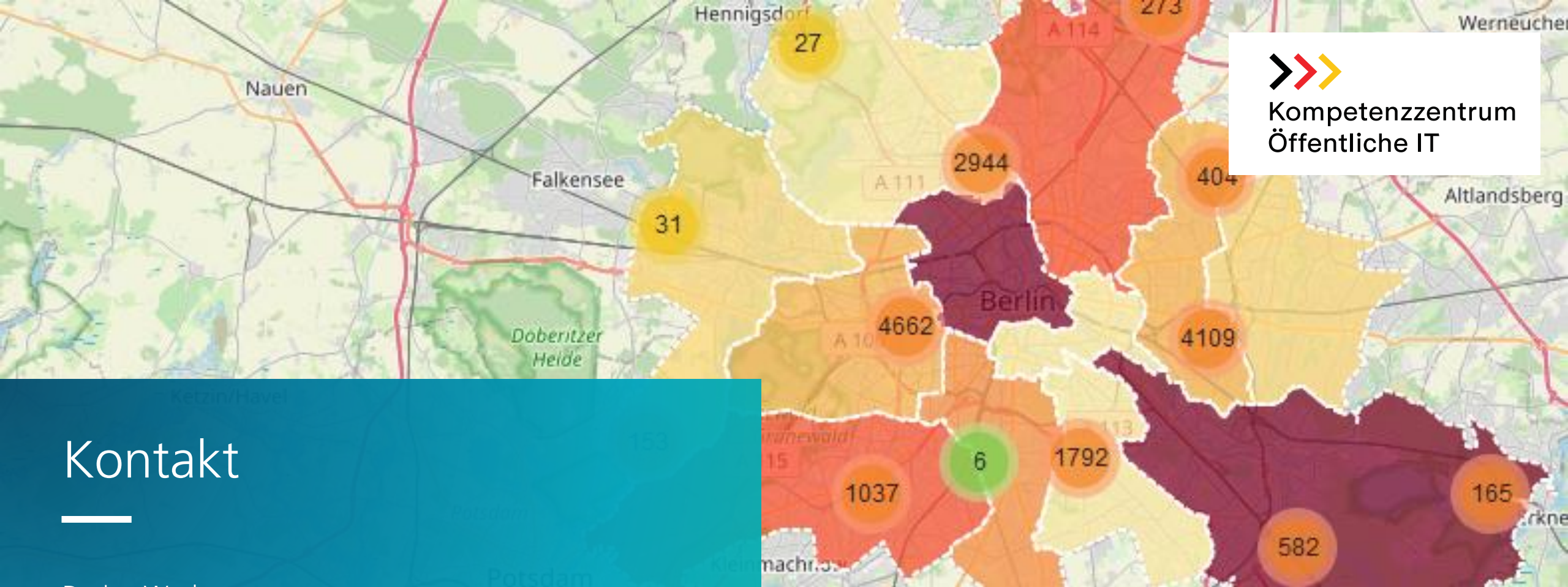
Was ist Öffentliche IT?

Unter öffentlicher IT versteht man Informationstechnologien, die in einem öffentlichen Raum durch die gesamtgesellschaftliche Relevanz unter besonderer Berücksichtigung der staatlichen Verantwortung stehen. Öffentliche IT verdeutlicht die Notwendigkeit IT als kritische Infrastruktur wahrzunehmen. Das Konzept der öffentlichen IT eröffnet die dringend erforderliche proaktive Diskussion über eine innovations- und kommunikationstreibende übergreifende öffentliche IT in Deutschland unter besonderer Beachtung der Verantwortung des Staates und des Datenschutzes.

[... mehr](#)

Aktuell






Kompetenzzentrum
Öffentliche IT

Kontakt

Dorian Wachsmann
Kompetenzzentrum Öffentliche IT (ÖFIT)
Tel. +49 30 3463-7338
dorian.wachsmann@fokus.fraunhofer.de

Gefördert durch:

