Verbundprojekt „Anonymität bei integrierten und georeferenzierten Daten (AnigeD)“

Technische Hochschule Köln

Förderkennzeichen: 16KISA095

Laufzeit des Vorhabens: 11/2022 - 11/2025

Teil II – Eingehende Darstellung

1. Projekthintergrund und Ziele

Das Ziel des Teilprojektes bestand zum einen in der Beurteilung und Weiterentwicklung bestehender Anonymisierungsmethoden für georeferenzierte Mikrodaten. Zum anderen waren Kriterien zur Bewertung der Datensicherheit und der Datenqualität für potenzielle (wissenschaftliche) Nutzer zu entwickeln. Zu Projektbeginn existierten bereits Veröffentlichungen über Methoden des sogenannten Geomaskings. Es gab jedoch keine umfassenden Vergleichsanalysen der Methoden untereinander. Außerdem fanden sich in der Fachliteratur keine Ergebnisse zur systematischen Bewertung der Datensicherheit, etwa im Hinblick auf realistische Datenangriffsszenarien. Auch hinsichtlich des beachtlichen Analysepotenzials von Geodaten mangelte es an einer strukturierten Klassifikation der Analysemöglichkeiten speziell für die empirisch arbeitende Wissenschaft. Das Projekt sollte damit einen methodischen Beitrag zur Verbesserung der Geodateninfrastruktur insgesamt leisten.

2. Projektablauf

Der Ablauf des Teilprojektes folgte untenstehendem Balkenplan, der sich über 12 Quartale erstreckt.

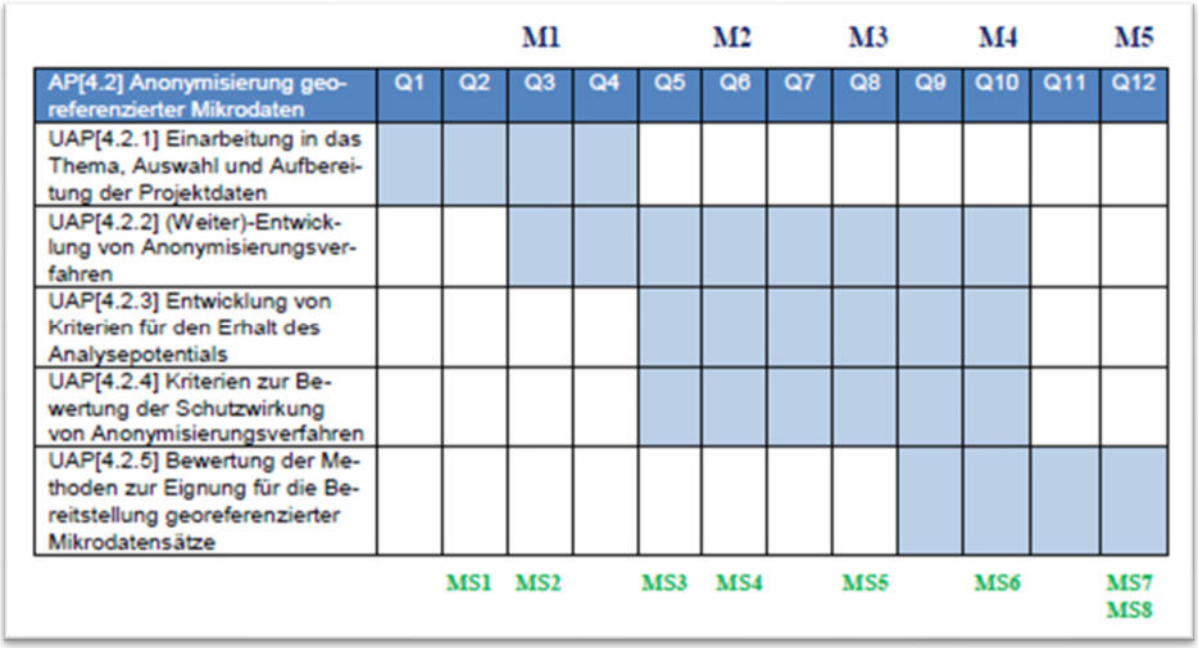

Abbildung: Balkenplan des Teilvorhabens

Auf die im Balkenplan dargestellten Unterarbeitspakete 4.2.1 bis 4.2.5 gehen wir unten in Kapitel 3 ein. Wir beginnen mit den fünf Meilensteinen (M1-M5) des Teilprojektes. Auf die Erreichung der acht Meilensteine des Gesamtvorhabens (MS1-MS8) wird an anderer Stelle im entsprechenden Berichtsdokument des Statistischen Bundesamtes eingegangen.

Meilenstein 1 (M1) – Einarbeitung in das Thema, Auswahl und Aufbereitung der Projektdaten:

> „Die (Test-)Datensätze für die weitere Projektarbeit sind verfügbar und der WMA hat eine detaillierte Übersicht über bestehende und für das Projekt relevante Geomasking-Methoden erstellt."

Dieser Meilenstein wurde mit einer Verzögerung gegenüber dem ursprünglich vorgesehenen Termin erreicht. Gründe waren der verzögerte Projektstart und der Mangel an Zugang zu amtlichen Mikrodaten (siehe Kurzbericht und Erfolgskontrollbericht) in der ersten Projektphase, die während der Projektlaufzeit kompensiert werden konnten.

Meilenstein 2 (M2) – Entwicklung und Implementierung von Anonymisierungsmethoden:

> „Es steht ein Werkzeugapparat in Form von Skriptprogrammen bereit, der transformatorisch auf die Projektdaten anwendbar ist."

Dieser Meilenstein wurde ohne Verzögerung zum vorgesehenen Termin erreicht.

Meilenstein 3 (M3) – Bewertung des Analysegehalts geomaskierter Daten:

> „Es liegt eine Typologie der Analyseverfahren samt jeweiliger Zielkriterien vor."

Dieser Meilenstein wurde ohne Verzögerung zum vorgesehenen Termin erreicht.

Meilenstein 4 (M4) – Kriterien zur Bewertung des Datenschutzes bei geomaskierten Daten:

> „Die zur Datenschutzbewertung erforderlichen Privacy-Metriken liegen vor und können für Datenangriffsszenarien eingesetzt werden."

Dieser Meilenstein wurde mit einer Verzögerung gegenüber dem ursprünglich vorgesehenen Termin erreicht.

Meilenstein 5 (M5) – Ableitung von Handlungsempfehlungen zur Erzeugung standardisierter Mikrodatensätze:

> „Die Handlungsempfehlungen für Datenanbieter und –halter liegen für den jeweiligen Datenkontext vor."

Dieser Meilenstein wurde ohne Verzögerung zum vorgesehenen Termin erreicht.

3. Durchgeführte Arbeiten und Ergebnisse

In Unterkapitel 3.1 werden die durchgeführten Arbeiten zusammengefasst. Anschließend folgen Erläuterungen über Abweichungen von der Projektplanung (3.2) und über die Zusammenarbeit unter den Projektpartnern (3.3).

3.1 Beschreibung der durchgeführten Arbeiten

In diesem Unterkapitel werden die Ergebnisse der Unterarbeitspakete 4.2.1 - 4.2.5 im Einzelnen besprochen.

Unterarbeitspaket 4.2.1: Einarbeitung in das Thema, Auswahl und Aufbereitung der Projektdaten

Die auf dem Markt existierenden und für das Projekt relevanten Mikrodatensätze wurden identifiziert und systematisiert. Diese stammen von Datenhaltern wie den Statistischen Ämtern

von Bund und Ländern, dem MeDIC der Universität zu Köln, dem Bundesamt für Kartographie und Geodäsie (Geodatensatz des Bundes) und auch von kommerziellen Datenanbietern wie beispielsweise dem Unternehmen Infas360. Auch innerhalb von professionellen Geoinformationssystemen wie der Umgebung von ArcGIS-Pro kann auf georeferenziertes Kartenmaterial zugegriffen werden.

Aufgrund einer bestehenden rechtlichen Problematik bei der Datenbereitstellung auf Seiten der am Cluster AnigeD beteiligten Datenhalter wurde wie im Teilvorhaben-Projektantrag (siehe dort unter „Risiko 1“) formuliert entsprechend reagiert und ein partiell synthetischer Mikrodatensatz generiert; und zwar im Wesentlichen basierend auf Publikationen der Stadt Köln und dem oben angesprochenen georeferenzierten Kartenmaterial aus ArcGIS-Pro. Der synthetische Datensatz wurde danach sukzessive um weitere Analysemerkmale angereichert, weist aber dennoch ein limitiertes Analysepotenzial auf. Er stellt damit keinen gleichwertigen Ersatz für amtliche georeferenzierte Mikrodaten dar.

Weiterhin wurde basierend auf einschlägigen Fachpublikationen eine Übersicht über bestehende Geomasking-Verfahren erstellt, strukturiert nach den beiden Kerncharakteristika “informationsreduzierend” und “datenverändernd”, sowie ein detailliertes Verständnis der Wirkungsweise sowohl der informationsreduzierenden als auch der datenverändernden Anonymisierungsverfahren entwickelt.

Unterarbeitspaket 4.2.2: (Weiter)-Entwicklung von Anonymisierungsverfahren

Ziel des Unterarbeitspaketes war zum einen die Entwicklung und Implementierung spezifischer datenverändernder und informationsreduzierender Verfahren zur Anonymisierung von Überschneidungsmerkmalen und/oder sensitiven Merkmalen in georeferenzierten Mikrodatensätzen. Zum anderen sollten effiziente, an die jeweilige Datenstruktur angepasste Programme zu den in 4.2.2.1 (informationsreduzierende Verfahren) und 4.2.2.2 (datenverändernde Verfahren) entwickelten Methoden transformatorisch implementiert werden.

UAP 4.2.2.1: Methodische Weiterentwicklung bestehender informationsreduzierender Verfahren

Die Hauptaufgabe bestand in einer adaptiven Clusterung der Daten in Gruppen möglichst ähnlicher Objekte, um bei der anschließenden Aggregation bzw. Datenveränderung deren Informationsverlust zu minimieren. Es wurde basierend auf der in UAP 4.2.1. erstellten Übersicht der relevanten Methoden eine Vorauswahl der vielversprechendsten Methoden für die weitere Projektarbeit getroffen. Mit diesen wurde danach weitergearbeitet.

Eine besondere Erkenntnis ist, dass der Umgebungsbegriff für einzelne geografische Datenpunkte in der Praxis variabel bzw. adaptiv in Abhängigkeit zur angewendeten Anonymisierungsmethode anzupassen ist. Speziell im Falle der informationsreduzierenden Methoden stellt eine solche Umgebung meist eine rechteckige Kachel oder, wie beim sogenannten Voronoi-Masking, ein konvexes Polygon dar. Für diese Methoden existieren nunmehr prototypische Implementierungen.

Ergänzend zu diesen Ansätzen wurde mit der Attribute-Based Adaptive Spatial Aggregation (ABASA) ein merkmalszusammenfassendes Anonymisierungsverfahren entwickelt, das eine merkmalsabhängige Anonymisierung räumlicher Informationen ermöglicht. Ausgangspunkt hierfür ist die Beobachtung, dass sich die Schutzwirkung georeferenzierter Mikrodaten häufig erst aus der Kombination mehrerer inhaltlicher Merkmale ergibt und eine rein globale räumliche Anonymisierung diesem Umstand nicht ausreichend Rechnung trägt.

ABASA zerlegt den Gesamtdatensatz zunächst in mehrere disjunkte Teildatensätze. Für jeden dieser Teildatensätze wird anschließend eine eigenständige Anonymisierung des räumlichen Merkmals durchgeführt. Reicht die Anzahl der Beobachtungen eines Teildatensatzes nicht zur Sicherstellung einer ausreichenden Schutzwirkung aus, werden benachbarte Teildatensätze sukzessive zusammengefasst, bis die benötigte Mindestfallzahl an Beobachtungen erreicht ist.

Durch dieses Vorgehen wird eine räumliche Aggregation erreicht, die an Merkmalskombinationen gekoppelt ist und genau dort erfolgt, wo sie aus Sicht des Datenschutzes erforderlich ist. Dies vermeidet eine unnötige Generalisierung räumlicher Informationen und ermöglicht zugleich eine flexible Anpassung an unterschiedliche Datensätze, Merkmalskombinationen und Mindestfallzahlen. Die Methode ist sowohl auf Datensätze mit bereits aggregierten räumlichen Einheiten, etwa Postleitzahlen oder andere administrative Gebiete, als auch auf geografische Punktdaten anwendbar. Im Falle von Punktdaten kann ein frei parametrisierbares Raster über die Daten gelegt werden, dessen Zellen als räumliche Grundeinheiten dienen und adaptiv zusammengefasst werden.

Die Anwendung der simultan in Unterarbeitspaket 4.2.4 entwickelten Anonymisierungs-metriken zeigte, dass ABASA auch bei moderater räumlicher Aggregation eine ausreichende Schutzwirkung erzielt und damit eine praktikable Ergänzung zu rein räumlichen Anonymisierungsverfahren darstellt. Um die ABASA-Maskierung an realen Daten testen zu können, wurden die Projektmitarbeiter im letzten Projektjahr kostenneutral nach Destatis abgeordnet, wo ein Zugriff auf amtliche Daten des Mikrozensus und der Krankenhausstatistik DRG bestand.

<u>UAP 4.2.2.2:</u> Methodische Weiterentwicklung bestehender datenverändernder Verfahren

Im Rahmen dieses Unterarbeitspakets hat sich die Verfahrensgruppe der sogenannten Donut-Maskierung aufgrund ihrer Flexibilität mit zwei Parametern (Radien) als besonders geeignet für die Entwicklung qualitativ hochwertiger anonymisierter Datenprodukte herausgestellt. Daher wurde dieser Ansatz methodisch weiterentwickelt, um dessen Schwächen in Bezug auf heterogene Punktedichten zu adressieren. Das konventionelle Donut-Masking arbeitet mit global festgelegten Mindest- und Maximalradien, die unabhängig von der lokalen Dichte der Datenpunkte angewendet werden. Dies führt dazu, dass Punkte in dicht besiedelten Gebieten häufig stärker als erforderlich modifiziert werden, während in dünn besiedelten Regionen die Verschiebung unter Umständen nicht ausreicht, um ein angemessenes Datenschutzniveau zu gewährleisten.

Zur Überwindung dieser Einschränkungen wurde das Verfahren des Adaptiven Nächste-Nachbarn-Donut-Masking (ANNDM) entwickelt, welches die Verschiebungsradien individuell für jeden Datenpunkt bestimmt. Die Radien werden dabei aus den Distanzen zu den jeweils nächsten Nachbarn abgeleitet. Konkret wird für jeden Punkt ein innerer Radius als durchschnittliche Distanz zu den nächsten k_1 Nachbarn sowie ein äußerer Radius als durchschnittliche Distanz zu den nächsten k_2 Nachbarn definiert, wobei die Wahl dieser Parameter k_1 und k_2 flexibel an den jeweiligen Datenkontext angepasst werden kann. Das ANNDM führt dazu, dass die Verschiebung eines Punktes unmittelbar an die lokale Punktdichte gekoppelt ist und somit eine angemessene Anonymisierung ermöglicht.

Wissenschaftlich-technisch stellt ANNDM eine Weiterentwicklung datenverändernder Geomasking-Verfahren dar, bei der die im konventionellen Donut-Masking verwendeten, für alle Datenpunkte identischen Verschiebungsradien durch individuell bestimmte Radien ersetzt werden, die eine punktindividuelle Steuerung der Verschiebungsdistanzen in Abhängigkeit von der lokalen Punktedichte erlauben.

Die Anwendung auf synthetische Projektdaten zeigt, dass mit ANNDM bei vergleichbarer Anonymisierungsstufe im Mittel geringere Verschiebungsdistanzen erforderlich sind als bei nicht-adaptiven Varianten des Donut-Maskings. Dies wirkt sich positiv auf den Erhalt räumlicher Strukturen und damit auf das Analysepotenzial anonymisierter Punktdaten aus. Das Ziel dieses Unterarbeitspakets, ein datenveränderndes Anonymisierungsverfahren mit verbesserter Anpassung an lokale Punktdichten zu entwickeln, wurde damit erreicht.

Zur Bewertung der Schutzwirkung und der aus der Anonymisierung resultierenden Verschiebungsdistanzen wurden die simultan in Unterarbeitspaket 4.2.4 entwickelten Anonymisierungsmetriken herangezogen. Dabei zeigte sich, dass ANNDM trotz moderat gewählter Verschiebungsdistanzen eine hohe Schutzwirkung entfaltet.

UAP 4.2.2.3: Implementierung der informationsreduzierenden und datenverändernden Verfahren

Zunächst wurde auf Basis der Übersichtsarbeiten in Unterarbeitspaket 4.2.1 die zu implementierenden Routinen identifiziert. Hierzu zählten Verfahren zur Aggregation von Mikrodaten, ausgewählte Verfahren der Datenveränderung inklusive Zufallsprojektionen und sogenannte Bloom-Filter-Verfahren.

Ein besonderes Augenmerk wurde auf die in 4.2.2.1 weiterentwickelten Varianten der Donut-Maskierung gerichtet, da sich diese Verfahrensgruppe in der Projektarbeit als besonders praktikabel herausgestellt hat.

Speziell für das ANNDM-Verfahren wurden transformatorische Programmcodes entwickelt, unter denen die Anzahl der zu berücksichtigenden Nachbarn eines Datenpunktes zur Berechnung der inneren und äußeren Radii frei parametrisierbar sind. Darüber hinaus wurden die entwickelten Codes bewusst flexibel gestaltet, sodass wahlweise entweder nur die im Datensatz enthaltenen nächstgelegenen Nachbarn oder alle real existierenden Adressen als Nachbarn in die Radiusberechnung einbezogen werden können. Auf diese Weise kann das Verfahren an unterschiedliche Datenverfügbarkeiten und Anwendungsszenarien angepasst werden, ohne dabei grundlegende Modifikationen an der Implementierung vornehmen zu müssen.

Die implementierten Programmcodes wurden im Laufe der Projektarbeiten kontinuierlich anhand der Projektdaten getestet und hinsichtlich ihrer zu Beginn der Arbeiten beachtlichen Laufzeiten optimiert.

Unterarbeitspaket 4.2.3: Entwicklung von Kriterien für den Erhalt des Analysepotenzials

Ziel des hier dargestellten Teilvorhabens war die Entwicklung von Kriterien, mit denen der Erhalt des analytischen Nutzens nach Anwendung von Geomasking-Verfahren systematisch beurteilt werden kann. Zu den wesentlichen Aufgaben des Arbeitspakets zählt die umfassende

Analyse des Stands der Forschung zu Geomasking-Verfahren sowie der in der Literatur verwendeten Bewertungsansätze. Dabei wurde festgestellt, dass bestehende Arbeiten ihren Schwerpunkt überwiegend auf datenschutzrechtliche Fragestellungen und die Reduktion von Re-Identifikationsrisiken legen, während der Erhalt des Analysepotenzials häufig nur fall- oder analysespezifisch betrachtet wird. Vor diesem Hintergrund wurde im Rahmen des Teilprojekts eine Systematisierung von Metriken, die für die Erfassung von Auswirkungen auf den Datennutzen herangezogen werden können, vorgenommen.

Ein zentrales Ergebnis des Arbeitspakets ist die Feststellung, dass sich keine universell anwendbare Kennzahl zur Bewertung des Analysepotenzials geomaskierter Daten ableiten lässt. Der zugrunde liegende Qualitätsbegriff ist mehrdimensional und abhängig von der jeweiligen Analyseform sowie vom Nutzungskontext der Daten. Einzelne Bewertungsmetriken erfassen jeweils nur spezifische Aspekte räumlicher Information. Eine Aggregation zu einer Gesamtkennzahl würde normative Gewichtungen erfordern und die Transparenz sowie Nachvollziehbarkeit der Bewertung einschränken. Dieses Ergebnis wurde im Projektverlauf als strukturelle Eigenschaft der Problemstellung identifiziert und entsprechend berücksichtigt, indem statt der Entwicklung einer übergreifenden Kennzahl eine systematische Klassifikation der in der Literatur verwendeten Bewertungsmetriken erarbeitet und weiterentwickelt wurde. Diese Klassifikation stellt ein zentrales Ergebnis des Arbeitspakets dar und ermöglicht eine strukturierte und technisch fundierte Bewertung des Analysepotenzials geomaskierter Daten.

Eine erste Kategorie bilden globale lagebezogene Metriken, insbesondere zentrografische Analysen zur Bewertung globaler Lage- und Streuungsmaße. Diese Metriken erfassen aggregierte Eigenschaften des gesamten Datensatzes, wie etwa Lagezentrum oder räumliche Ausdehnung, und erlauben Aussagen darüber, inwieweit sich die makroskopische räumliche Struktur infolge des Geomaskings verändert. Sie liefern damit eine globale Perspektive auf die räumliche Verteilung der Beobachtungen, ohne direkten Bezug auf lokale Strukturen oder Lagebeziehungen zwischen einzelnen Datenpunkten.

Eine zweite Kategorie umfasst lokale strukturbezogene Metriken, die sich auf die räumlichen Beziehungen der Punkte untereinander beziehen. Hierzu zählen Punktmuster-Analysen zur Untersuchung räumlicher Verteilungen und Abstände, Clusteranalysen zur Bewertung räumlicher Verdichtungen sowie Analysen räumlicher Autokorrelation zur Erfassung räumlicher Abhängigkeiten. Diese Metriken ermöglichen eine differenzierte Beurteilung lokaler Strukturen und sind insbesondere für Analysen relevant, die auf Nachbarschaftsbeziehungen, relativen Distanzen oder räumlichen Interaktionen basieren.

Darüber hinaus wurde eine dritte Kategorie identifiziert, die kontext- bzw. attributbezogene Metriken umfasst. Diese Analysen beziehen sich nicht primär auf die relative Lage der Punkte zueinander, sondern auf deren Einbettung in den räumlichen Kontext. Bewertet wird hierbei, inwieweit ortsbezogene Umwelt- und Kontextinformationen durch das Geomasking erhalten bleiben, etwa die Zuordnung zu Flächennutzungs- oder Landschaftstypen oder die Erhaltung von Entfernungen zu relevanten Infrastrukturobjekten wie Straßen oder Gewässern. Diese Dimension des Analysepotenzials ist insbesondere für Expositions- und Kontextanalysen von Bedeutung und ergänzt die lage- und strukturbezogene Bewertung um eine inhaltlich-analytische Perspektive.

Ein weiteres wesentliches Ergebnis des Arbeitspakets ist die klare Herausarbeitung der Abhängigkeit zwischen Analyseziel und Eignung eines Geomasking-Verfahrens. Die Untersuchungen zeigen, dass jene Verfahren nicht unabhängig vom vorgesehenen Nutzungsszenario bewertet werden können. Vielmehr ist die Auswahl des Verfahrens an der Frage auszurichten, welche räumliche Information für die geplante Analyse als wesentlich anzusehen sind. Entsprechend unterscheiden sich die Anforderungen zwischen Analysen, bei denen Nachbarschaftsbeziehungen im Vordergrund stehen, Analysen mit Fokus auf ortsbezogene Expositionen oder Analysen zur Identifikation räumlicher Cluster.

Neben diesen wissenschaftlich-technischen Ergebnissen ergeben sich aus dem Arbeitspaket auch praktische Implikationen. Insbesondere für Datenhalter bietet der entwickelte Kategorisierungsansatz eine belastbare Entscheidungsgrundlage für die anwendungsorientierte Auswahl von Geomasking-Verfahren. Dies ist vor dem Hintergrund relevant, dass in der bisherigen Forschung häufig der Anspruch verfolgt wird, ein einzelnes, allgemein überlegenes Geomasking-Verfahren zu identifizieren. Die Ergebnisse des Arbeitspakets zeigen jedoch, dass ein solcher Ansatz den unterschiedlichen Anforderungen wissenschaftlicher Analysen nicht gerecht wird. Stattdessen wird deutlich, dass insbesondere im Kontext der Bereitstellung von Daten für die empirische Forschung ein dynamischer und nutzungsorientierter Ansatz erforderlich ist, bei dem eine Auswahl datenverändernder und informationsreduzierender Geomasking-Verfahren als Werkzeugkasten verfügbar ist. Die Auswahl eines konkreten Verfahrens oder die Kombinationen mehrerer Verfahren sollte dabei in Abhängigkeit vom beabsichtigten Analysezweck und den relevanten Qualitätsdimensionen erfolgen. Durch eine solche zielgerichtete Ausrichtung der Anonymisierung wird der Nutzen bereitgestellter Mikrodaten optimiert und gleichzeitig das vorgegebene Datenschutzniveau erreicht. Langfristig trägt dieser Ansatz zu einer effizienten Nutzung vorhandener Datenbestände sowie zur Steigerung des analytischen Mehrwerts georeferenzierter Erhebungen, insbesondere aus der amtlichen Statistik, bei.

Das Ziel, Kriterien zur Bewertung des Erhalts des Analysepotenzials zu entwickeln, wurde damit erreicht, wenngleich nicht in Form einer einzelnen Kennzahl. Eine wesentliche Herausforderung bestand in der Heterogenität des Qualitätsbegriffs sowie in der Vielfalt räumlicher Analyseansätze. Diese Herausforderung wurde dadurch adressiert, dass die unterschiedlichen Qualitätsdimensionen nicht reduziert, sondern systematisch strukturiert und transparent dargestellt wurden. Risiken hinsichtlich einer zu starken Analysespezifik konnten durch die abstrakte Kategorisierung der Bewertungsmetriken reduziert werden. Zusammenfassend konnte im Arbeitspaket gezeigt werden, dass der Erhalt des Analysepotenzials unter Anwendung von Geomasking-Verfahren nur durch eine differenzierte und kontextbezogene Betrachtung angemessen bewertet werden kann. Die entwickelten Kriterien und die Systematisierung der Bewertungsmetriken stellen eine belastbare Grundlage für die weitere methodische Ausgestaltung und praktische Anwendung von Geomasking-Verfahren dar.

Unterarbeitspaket 4.2.4: Entwicklung von Kriterien zur Bewertung der Schutzwirkung von Anonymisierungsmaßnahmen

Das Ziel dieses Unterarbeitspakets bestand zum einen in der Entwicklung einer realitätsnahen und vollständigen Klassifikation potenzieller Angriffsszenarien für georeferenzierte Mikrodaten. Zum anderen sollte ein belastbarer Kriterienkatalog zur Bewertung der Schutzwirkung von Geomasking-Verfahren entwickelt werden. Beide Ziele wurden erreicht, wie nachfolgend dargestellt wird.

UAP 4.2.4.1: Systematisierung potenzieller Angriffsszenarien für georeferenzierte Mikrodaten

Im Rahmen dieses Unterarbeitspakets wurde eine systematische Typologie realistischer Angriffsszenarien auf anonymisierte georeferenzierte Mikrodaten entwickelt. Ziel war es, potenzielle Datenschutzrisiken nicht abstrakt, sondern aus der Perspektive realer, rational handelnder Datenangreifer zu strukturieren und damit eine belastbare Grundlage für die Bewertung der Schutzwirkung von Geomasking-Verfahren zu schaffen.

Zentrales Ergebnis ist die Unterscheidung zweier grundlegender Angriffsperspektiven, die sich danach differenzieren, ob die ursprüngliche Adresse als sogenanntes Zielattribut (Perspektive 1) oder als Quasi-Identifikator (Perspektive 2) fungiert. Für jede dieser Perspektiven wurden jeweils vier Angriffsszenarien formuliert, die sich systematisch nach dem Vorliegen von Zusatzwissen über

1. die Teilnahme einzelner Personen an der Erhebung (Participation Knowledge) und
2. Details der eingesetzten Geomasking-Methodik (Method Knowledge) zu unterscheiden.

Zu Perspektive 1 (Originaladresse als Zielattribut) wurden folgende vier Szenarien potenziellen Zusatzwissens von Datenangreifern identifiziert:

1. Kein Wissen über Teilnahme und Methodik,
2. Wissen über Teilnahme, kein Wissen über Methodik,
3. Wissen über Methodik, kein Wissen über Teilnahme und
4. Wissen über Teilnahme und Methodik.

Diese Szenarien beschreiben Situationen, in denen ein Datenangreifer versucht, ausgehend von anonymisierten Geokoordinaten auf die ursprünglichen Wohnadressen der Auskunftgebenden zu schließen und damit eine Identität offen zu legen.

Zu Perspektive 2 (Originaladresse als Quasi-Identifikator) wurden analog zu Perspektive 1 vier Szenarien potenziellen Zusatzwissens definiert. Hierbei steht ein sogenannter inverser Datenangriff im Fokus, bei dem bekannte reale Adressen gezielt mit anonymisierten Punkten verknüpft werden, um sensible Attribute offenzulegen.

Damit leistet dieses Arbeitspaket einen zentralen Beitrag zur Weiterentwicklung der Bewertung räumlicher Anonymität. Während bestehende Ansätze – darunter insbesondere die räumliche *k*-Anonymität – nur einzelne Angriffsszenarien implizit adressieren, ermöglichen die hier entwickelten Szenarien eine explizite und differenzierte Risikoanalyse.

Durch die klare Trennung der Angriffsperspektiven wird deutlich, dass der Schutzbedarf stark vom angenommenen Datenangreifer-Modell abhängt und letztlich auch von der Art des Datenzugangs (z.B. die Arbeit im heimischen Büro mit einem Mikrodatensatz, an einem Gastwissenschaftlerarbeitsplatz des Datenhalters, via kontrollierten Datenfernzugriff usw.). Dies erlaubt erstmals eine methodisch eindeutige Zuordnung spezifischer Anonymisierungsmetriken zu potenziellen Angriffsszenarien.

Aus Sicht der Datenhalter ist diese Systematisierung relevant, da sie eine fundierte Grundlage für risikoadäquate Datenschutzentscheidungen bietet. Diese können damit Geomasking-Verfahren gezielt auswählen und ihre Auswahl begründen, was langfristig die Bereitstellung anonymisierter georeferenzierter Forschungsdaten und deren Akzeptanz verbessert und rechtliche Unsicherheiten reduzieren kann, soweit die vorgestellten Metriken in aktuelle Rechtsgrundlagen aufgenommen werden.

Eine zentrale Herausforderung des Arbeitspakets bestand in der Formulierung von Datenangriffsszenarien, die weder rein theoretischer Natur noch auf sehr spezifische Einzelfälle beschränkt sind. Durch die klare Strukturierung entlang der beiden Dimensionen des Zusatzwissens (Teilnahme- und Methodenkenntnis) wurde eine konsistente und überschneidungsfreie Typologie entwickelt. Die resultierenden Szenarien erwiesen sich in der täglichen Projektarbeit als tragfähig und anschlussfähig mit Blick auf die nachfolgenden Arbeitspakete, speziell hinsichtlich der Entwicklung und Anwendung sogenannter Privacy-Metriken.

UAP 4.2.4.2: Zusammenstellung von Kriterien zur Bewertung der Schutzwirkung

Aufbauend auf der im Unterarbeitspaket 4.2.4.1 entwickelten Systematik realistischer Angriffsszenarien wurde im Rahmen dieses Unterarbeitspakets ein Kriterienkatalog zur Bewertung der Schutzwirkung von Geomasking-Verfahren erarbeitet. Ziel war es, die Auswirkungen unterschiedlicher Anonymisierungsstrategien auf das Datenschutzniveau georeferenzierter Mikrodaten messbar und vergleichbar zu machen.

Es wurden vier komplementäre Anonymisierungsmetriken entwickelt, die jeweils gezielt auf unterschiedliche Angriffsperspektiven und Wissensstände potenzieller Datenangreifer ausgerichtet sind. Diese Metriken erweitern bestehende Ansätze zur Bewertung räumlicher Anonymität und ermöglichen erstmals eine systematische, szenariobasierte Einschätzung der Schutzwirkung von Geomasking-Verfahren.

Die Kriterien adressieren sowohl Angriffe, bei denen von anonymisierten Koordinaten auf Originaladressen geschlossen werden soll, als auch inverse Angriffe, bei denen originale Adressen mit anonymisierten Daten verknüpft werden. Gleichzeitig wird - wie in den Ausführungen zu Unterarbeitspaket 4.2.4.1 erläutert - berücksichtigt, ob ein Angreifer über Zusatzwissen zur Teilnahme an der Erhebung oder zu den eingesetzten Anonymisierungsmethoden verfügt.

Aus wissenschaftlich-technischer Sicht stellen die entwickelten Metriken einen wesentlichen Fortschritt gegenüber der bislang vorherrschenden Bewertung räumlicher Anonymität dar. Während die räumliche *k*-Anonymität lediglich einen Ausschnitt möglicher Angriffsszenarien abdeckt, ermöglichen die hier entwickelten Kriterien eine differenzierte Bewertung entlang der in Unterarbeitspaket 4.2.4.1 identifizierten Angriffsperspektiven.

Konkret wurden vier Metriken definiert und empirisch untersucht:

- *k-original:* zur Bewertung der Datenschutzwirkung gegenüber Angriffen ohne Methodenwissen, ausgehend von anonymisierten Adressen,

- *k-original (methodenbezogen):* zur Bewertung bei dem Datenangreifer bekannter Geomasking-Methodik,

- *k-moved:* zur Bewertung der Datenschutzwirkung gegenüber Angriffen ausgehend von bekannten Originaldressen ohne Methodenwissen,

- *k-moved (methodenbezogen):* zur Bewertung entsprechender Angriffe bei dem Datenangreifer bekannter Geomasking-Methodik.

Diese vier Metriken erlauben es, die Schutzwirkung eines Geomasking-Verfahrens nicht nur global, sondern abhängig vom angenommenen Angreifermodell zu beurteilen.

Die exemplarische Überprüfung der Kriterien anhand probeweise anonymisierter Projektdaten zeigt, dass die Schutzwirkung einzelner Geomasking-Verfahren deutlich davon abhängt, welches Angriffsszenario zugrunde gelegt wird. Damit hat sich bestätigt, dass eine eindimensionale Bewertung – etwa allein auf Basis der in der Fachliteratur vorherrschenden räumlichen *k*-Anonymität – nicht ausreicht, um das tatsächliche Datenschutzniveau eines (anonymisierten) georeferenzierten Mikrodatensatzes angemessen darzustellen.

Das wesentliche Ziel des Arbeitspakets, in die Praxis umsetzbare und auf verschiedene Anonymisierungsstrategien übertragbare Kriterien zu entwickeln, wurde damit erreicht.

<u>Unterarbeitspaket 4.2.5:</u> Bewertung der Methoden zur Eignung für die Bereitstellung georeferenzierter Mikrodatensätze

Die im Projekt (weiter)-entwickelten Anonymisierungsmaßnahmen wurden hinsichtlich der Eignung zur Erzeugung standardisierter Mikrodatensätze für Datennutzer insbesondere aus der Forschung untersucht. Die in Arbeitspaket 4.2.2.2 entwickelte Methode der adaptiven Donut-Maskierung (ANNDM) ist sehr gut geeignet zur Erzeugung solcher Forschungsdaten. Aus wirtschaftlicher und institutioneller Perspektive ist das Verfahren insbesondere für Datenhalter geeignet, die georeferenzierte Mikrodaten über räumlich heterogene Gebiete (z.B. urbane und ländliche Regionen innerhalb eines Datensatzes) bereitstellen wollen. Durch die adaptive Festlegung der Verschiebungsradien und der Umgebungsparameter (Einbezug von

Punkten aus der direkten Nachbarschaft) kann der Zielkonflikt zwischen Datenschutz und analytischer Nutzbarkeit aufgelöst werden.

Ein weiteres Ziel des Unterarbeitspakets bestand in der Entwicklung von Handlungsempfehlungen für Datenanhalter zur Erzeugung von anonymisierten Datenprodukten aus georeferenzierten Mikrodaten. Die Erzeugung anonymisierter georeferenzierter (Forschungs-)Daten ist nach Einschätzung des Teilprojektteams bei einer Anwendung der zuvor erläuterten Kriterien samt Anonymisierungsmetriken möglich.

Der Einsatz des entwickelten Werkzeugkastens, bestehend aus Anonymisierungsmethoden und Kriterien zur Bewertung von Datensicherheit und Analysepotenzial, auf Seiten der Datenhalter hängt allerdings von rechtlichen Fragen und institutionellen Entscheidungen ab, die sich - in der Diskussion mit den Projektpartnern auch anderer Teilprojekte des Clusters AnigeD – als sehr komplex herausgestellt haben.

3.2 Änderungen/Abweichungen im Projektplan

Anders als geplant standen dem Projektteam aus rechtlichen Gründen keine amtlichen georeferenzierten Mikrodaten zur Verfügung. Dieser Mehraufwand, der aufgrund der Nicht-Verfügbarkeit amtlicher Mikrodaten nötig war, wurde bereits zu Projektbeginn als potenzielles Problem identifiziert und im Projektantrag als mögliche Reaktion (Risiko 1) die Erzeugung eines synthetischen Mikrodatensatzes vorgesehen. Die Erzeugung eines synthetischen georeferenzierten Mikrodatensatzes, und zwar zu Punktdaten der Stadt Köln, ist gelungen. Somit konnten die entwickelten Programmroutinen getestet werden. Im letzten Halbjahr der Projektlaufzeit konnten die entwickelten Geomasking-Verfahren zusätzlich auf einem durch den Unterauftragnehmer bereitgestellten Mikrodatensatz mit breiterem Merkmalsspektrum validiert werden, weshalb es gelungen ist, die wesentlichen Projektziele zu erreichen (i.e. Entwicklung der Methodik und Ableitung von Handlungsempfehlungen für die Praxis).

3.3 Zusammenarbeit und Rolle der Partner

Das Forschungsvorhaben wurde an der Technischen Hochschule Köln durchgeführt. Unteraufträge wurden an Prof Schnell (Universität Duisburg-Essen) vergeben.

Es gab einen regelmäßigen Austausch aller an den Arbeitspaketen beteiligten Institutionen, und zwar einerseits durch monatliche Treffen auf Projektleiterebene und andererseits durch bilaterale Treffen auf Ebene der Mitarbeiter und mit dem Unterauftragnehmer. So konnten neue Erkenntnisse zeitnah ausgetauscht und mit Fachkollegen diskutiert werden.

Der im Projekt entwickelte synthetische Datensatz zu Punktdaten der Stadt Köln wurde darüber hinaus auch von den Projektpartnern in anderen Arbeitspaketen verwendet. Hier gab es einen regen Austausch mit dem Statistischen Bundesamt (Mitarbeiter Herr Tang). Der Projektmitarbeiter Simon Cremer wirkt in einer Bund-Länder-Arbeitsgruppe zur Statistischen Geheimhaltung mit, die Handreichungen für eine praktische Umsetzung von Anonymisierungsmethoden entwickelt, in Zusammenarbeit mit dem von Sarah Gießing geleiteten Teilprojekt und ihrem Mitarbeiter Martin Möhler.

Da sich im Projektverlauf auch der Projektpartner FU Berlin (Prof. Rendtel und Mitarbeiterin Frau Grill) mit der Donut-Maskierung beschäftigte, entstand ein bilateraler Austausch zu dieser Methodik.

Hilfreich waren auch die Diskussionen mit Prof. Martini und seinen Mitarbeitern Frau Benner und Herrn Wagner bei Fragen rund um den Datenschutz und die Einhaltung von Restriktionen bei der Erzeugung anonymisierter Daten zu Forschungszwecken.

Die Projektergebnisse wurden zudem zweimal jährlich auf den Werkstatttreffen im Statistischen Bundesamt in Wiesbaden präsentiert.

Die wissenschaftliche Relevanz der im Projekt AnigeD bearbeiteten Fragestellungen zeigte sich auch in der erfolgreichen Aufnahme der beiden Projektmitarbeiter in das interdisziplinäre Promotionsprogramm IPHS der Universität zu Köln. Aufbauend auf den Arbeiten zur Bewertung des Analysepotenzials sowie zur Messung der Anonymität geomaskierter Datensätze werden die im Projekt entwickelten methodischen Ansätze damit im Rahmen von Promotionsvorhaben weiterverfolgt.

4. Notwendigkeit und Angemessenheit der Projektarbeiten

Die im Rahmen dieses Teilprojekts entwickelte Typologie der Angriffsszenarien stellte eine wesentliche konzeptionelle Grundlage für die Entwicklung neuer Metriken zur Bewertung der Schutzwirkung von Geomasking-Verfahren, die weit über die bis dato aus der Fachliteratur bekannte räumliche *k*-Anonymität hinausgehen. Es wurde damit nicht nur

Grundlagenforschung geleistet, sondern ein praktischer Mehrwert für die Halter georeferenzierter Makrodatensätze generiert, denen nun ein Werkzeugkasten für die alltägliche Arbeit an der Schnittstelle zwischen Datenanbietern und den (wissenschaftlichen) Datennutzern zur Verfügung steht.

Gleichzeitig kommt der systematischen Aufstellung von Angriffsszenarien auch bei der künftigen (Weiter-)Entwicklung von Anonymisierungsverfahren eine zentrale Rolle zu, da die Wirksamkeit solcher Methoden maßgeblich davon abhängt, wie robust sie gegenüber realen Angreifermodellen ausgestaltet sind. In diesem Sinne leistet die entwickelte Szenarien-Systematik einen Beitrag zum aktuellen wissenschaftlichen Diskurs im Bereich Geoprivacy und schließt eine große Lücke, nämlich die bislang fehlende explizite Einbindung realer Angriffssituationen in die Bewertung der Anonymität räumlicher Mikrodaten.

Die Entwicklung der Methode der adaptiven Donut-Maskierung (ANNDM) greift wissenschaftliche Bestrebungen auf, Verschiebungsdistanzen in Geomasking-Verfahren stärker daten- und dichteabhängig zu steuern, und erweitert damit den bestehenden Methodenkanon um einen praktikablen Ansatz zur Austarierung der beiden meist gegenläufigen Ziele des Datenschutzes und des Analysepotenzials.

Im Gleichklang dazu wurde im Projekt auch eben jener bislang unscharf verwendete Begriff des Analysepotenzials systematisch aufgearbeitet und in inhaltlich unterscheidbare Dimensionen räumlicher Information überführt. Anstelle der Entwicklung einer eindimensionalen Qualitätskennzahl wurde ein klassifikatorischer Ansatz gewählt, der unterschiedliche Analyseanforderungen explizit berücksichtigt. In der Zusammenschau mit der parallel entwickelten Typologie realistischer Angriffsszenarien liegen damit nun sowohl auf der Schutz- als auch auf der Nutzenseite die konzeptionellen Grundlagen vor, um Geomasking-Verfahren künftig strukturiert, transparent und anwendungsorientiert (weiter) zu entwickeln. Die im Projekt erarbeiteten Systematisierungen ermöglichen es, neue Anonymisierungsansätze gezielt entlang konkreter Anforderungen an Anonymität und Analysepotenzial auszurichten und deren Wirkungen nachvollziehbar zu bewerten

5. Wichtigste Positionen des zahlenmäßigen Nachweises

Die Zuwendungen wurden in erster Linie für folgende Positionen eingesetzt:

- Personalkosten für wissenschaftliche Mitarbeiter und studentische Hilfskräfte, um Recherchearbeiten durchzuführen sowie die Forschungsergebnisse insgesamt zu

entwickeln, transformatorisch zu implementieren, zu dokumentieren und zu veröffentlichen.

- Reisekosten für die interdisziplinäre Zusammenarbeit mit den Projektpartnern und die Diskussion der erzielten Ergebnisse mit anderen Forschern auf nationalen und internationalen Fachkonferenzen (inklusive Projektsitzungen des Clusters AnigeD und Workshops des Forschungsnetzwerks AnoSiDat).
- Sachmittel für Unteraufträge und Rechner, um die Arbeiten innerhalb des Projektzeitrahmens durchführen zu können und die erzielten empirischen Forschungsergebnisse einer breiten Öffentlichkeit zugänglich machen zu können.

Detaillierte Angaben finden sich im simultan eingereichten zahlenmäßigen Verwendungsnachweis. Die Mittelverwendung erwies sich als notwendig und angemessen, da die Entwicklung von Forschungsergebnissen dieses Umfangs ohne eine entsprechende personelle Unterstützung nicht zu leisten gewesen wäre und speziell die Zusammenarbeit mit den Projektpartnern und dem Unterauftragnehmer von zentraler Bedeutung für die Validierung der erzielten Ergebnisse war.

6. Voraussichtlicher Nutzen und Verwertbarkeit

Eine Nachnutzbarkeit der Projektergebnisse ist ohne Zweifel gegeben. Für einen Transfer der Ergebnisse und die zukünftige Anwendung der entwickelten Anonymisierungsmethodik in der alltäglichen Praxis der Datenhalter erscheint die Implementierung einer gemeinsamen Transferstelle zur Beratung von staatlicher und privater Datenhalter bei der Anwendung der Methoden und Publikation vertraulichen georeferenzierten Datenmaterials sinnvoll und folgerichtig. Eine solche Transferstelle könnte durch die Partner des Verbundprojektes, die in den vergangenen drei Jahren das notwendigen Grundlagen erarbeitet haben, betrieben werden. Durch den Betrieb einer solchen Transferstelle kämen die im Projekt entwickelten Methoden und damit gewonnenen empirischen Erfahrungen der an den Mikrodaten interessierten wissenschaftlichen Gemeinschaft zugute, besonders aus dem Bereich der Wirtschafts-, Sozial- und Gesundheitswissenschaften. Zu den Aufgaben einer Transferstelle zählen: die Bereitstellung von Leitfäden zur Erstellung von anonymisierten Datenprodukten, die Formulierung von Leitlinien zur Anwendung der entwickelten Privacy- und Utility-Metriken bei der Generierung von Datenprodukten, die Bereitstellung standardisierter Programmcodes samt Handbuch zu den Anonymisierungsverfahren. Hierzu zählt eine flexible Adaption der Anonymisierungsparameter, um verschiedenen (Forschungs)-Zielen der Nutzer über verschiedene Datenzugangswege - wie Scientific-Use-Files, Campus-Files für die Lehre,

Remote Access oder die Arbeit an einem Gastwissenschaftlerarbeitsplatz - gerecht zu werden. Auf diese Weise zahlen die Forschungsergebnisse des Clusters direkt in die Bestrebungen der europäischen Datenstrategie, der DSGVO und des in Bearbeitung befindlichen Forschungsdatengesetzes ein.

Da den Mitarbeitern während der Projektlaufzeit keine realen Testdatensätze kleinräumig geoferenzierter Mikrodatensätze zur Verfügung standen, sollte die entwickelte Methodik in der weiterführenden Forschung (neben den im Rahmen von AnigeD eigens erzeugten Mikrodatensätzen) auf reale georeferenzierte Mikrodaten angewendet werden, beispielsweise auf Mikrodaten der Luftrettung in Kombination mit Einzeldaten der amtlichen Krankenhausstatistik (DRG) aus der Inanspruchnahme akuter medizinischer Notfälle. Die Vorarbeiten des Projektes haben bereits datenverändernde und aggregierende Anonymisierungstechniken und Kenngrößen zur Messung von räumlicher Anonymität und Datennutzen hervorgebracht, deren Anwendung zur Erzeugung von Public-Use- und Scientific-Use-Files folgerichtig der nächste Schritt ist.

7. Veröffentlichungen / geplante Veröffentlichungen

Nachfolgend sind in Abschnitt 7.1 die Konferenzen und Workshops gelistet, an denen das Projektteam teilgenommen hat. In Abschnitt 7.2 finden sich die Publikationen während der Projektlaufzeit sowie die geplanten Publikationen im Rahmen eines durch das das Cluster AnigeD organisierten Sonderbandes im Wirtschafts- und Sozialstatistischen Archiv.

7.1 Konferenzen und Workshops

- Statistische Woche in Dortmund, September 2023
  Beitrag: *Anonymisierung georeferenzierter Mikrodaten für die Forschung – ein Werkstattbericht*
- Werkstatttreffen AnigeD im März 2023 und September 2023 in Wiesbaden mit eigenen Beiträgen
- Zwischenkonferenz AnigeD im Oktober 2023 im Spreespeicher Berlin
  Beitrag: *Anonymization of georeferenced microdata*
- Werkstatttreffen AnigeD in Wiesbaden im März 2024 mit eigenen Beiträgen
- Zwischenkonferenz AnoSiDat in Lübeck im April 2024
  Beitrag: *The dilemma of anonymizing georeferenced microdata*
- Konferenz Privacy in Statistical Databases (PSD), Antibes, September 2024

Beitrag: *Masking georeferenced health data - an analysis taking the example of partially synthetic data on sleep disorder*

- Werkstatttreffen AnigeD in Wiesbaden im April 2025 mit eigenen Beiträgen
- Konferenz New Techniques and Technologies for Statistics (NTTS), Brüssel, März 2025
  Beitrag: *A risk-utility framework for anonymized georeferenced health microdata*
- Konferenz Journées Statistique de la Societé Francaise (JdS), Marseille, Juni 2025
  Beitrag: *Anonymization of georeferenced health microdata – a glimpse of work in progress*
- Abschlusskonferenz AnoSiDat im AXICA-Kongresszentrum Berlin, Oktober 2025
  Beiträge:
  1) Extension of Spatial *k*-Anonymity - New Metrics for Assessing the Anonymity of Geomasked Data Considering Realistic Attack Scenarios
  2) Vorstellung der Ergebnisse des Clusters AnigeD
- 34. Wissenschaftlichen Kolloquium in Wiesbaden, November 2025
  Zwei Beiträge:
  1) *Neue Metriken zur Bewertung der Anonymität nach Geomaskingverfahren unter Berücksichtigung realer Angriffsszenarien*
  2) Zur Qualität geomaskierter Daten

7.2 Publikationen

- L. Jehmlich, S. Cremer (2024): *The dilemma of anonymizing georeferenced microdata*, Poster, Forschungskonferenz Anonymisierung für eine sichere Datennutzung (AnoSiDat), Lübeck 2024
- S. Cremer, L. Jehmlich, R. Lenz (2024): *Masking Georeferenced Health Data - An Analysis Taking the Example of Partially Synthetic Data on Sleep Disorder*, Privacy in Statistical Databases (PSD) 2024, Proceedings, Lecture Notes in Computer Science (LNCS) vol.14915, S. 297-309
- S. Cremer, R. Lenz, L. Jehmlich (2025): *Extension of Spatial k-Anonymity - New Metrics for Assessing the Anonymity of Geomasked Data Considering Realistic Attack Scenarios*, Poster, Forschungskonferenz ‘Anonymisierung für eine sichere Datennutzung’ (AnoSiDat), Berlin 2025
- L. Jehmlich, R. Lenz, S. Cremer (2025): *A risk-utility framework for anonymized georeferenced health microdata*, New Techniques and Technologies for Statistics (NTTS), Brüssel, S. 19-22

- R. Lenz, L. Jehmlich, S. Cremer (2025): *Anonymization of georeferenced health microdata - a glimpse of work in progress*, Journees de Statistique de la SFdS, Marseille 2025, S. 741-747
- S. Cremer, L. Jehmlich, R. Lenz (2025): *Anonymization of Georeferenced Public Health Microdata*, American Journal of Public Health Research, 2025, Vol. 13, No. 6, S. 257-262
- S. Cremer, L. Jehmlich, R. Lenz (2025): *Extension of Spatial k-Anonymity: New Metrics for Assessing the Anonymity of Geomasked Data Considering Realistic Attack Scenarios*, arXiv, https://doi.org/10.48550/arXiv.2509.07505
- Es wird in 2026 ein Sonderband *Anonymization and Confidentiality in the Use of Sensitive Data* im Wirtschafts- und Sozialstatistischen Archiv (AStA) der Deutschen Statistischen Gesellschaft veröffentlicht, bei dem der Teilprojektleiter der TH Köln (Rainer Lenz) als Mitherausgeber fungiert. Die Projektmitarbeiter werden darin ebenfalls Fachaufsätze platzieren.