

Abstract English

The Research Data Center of the Federal Statistical Office of Germany offers data access to German official microdata for independent research institutions. Depending on the selected way of data access, the strict data privacy regulations for German official microdata mostly require sensitive attributes, such as precise geospatial information, to be anonymized, which makes the data products unusable for precise geospatial analysis. However, many research institutes might heavily benefit from research data with precise geo-referencing. Targeting this conflict, in this thesis paper we present and empirically examine five different data synthesis methods on the German Census Data from 2011, with the goal of finding a way for creating and providing useful synthetic versions of German official microdata with precise geospatial attributes. Regarding the synthesis models, we apply two parametric synthesizers, which are a Copula Synthesizer with Frequency Encoding and a Copula Synthesizer with One-Hot Encoding, as well as two synthesizers without parametric assumptions, which are a Random Forest Synthesizer and a CART (Classification and Regression Tree) Synthesizer, and finally Geomasking, a commonly used anonymization method for georeferenced data. We assess the potential the different synthesizers provide us using various metrics from the field of data synthesis for quantifying data utility and data privacy. The result of this thesis is that out of all examined synthesizers, the CART Synthesizer performs the best in terms of data utility, while offering a negligible risk in regards to data privacy. It would be conceivable to make such a synthetic data set with precise geospatial information accessible for independent research institutions via the remote access system of the Research Data Center.

Abstract Deutsch

Das Forschungsdatenzentrum des Statistischen Bundesamtes kann unabhängigen Forschungseinrichtungen Zugang zu amtlichen Einzeldaten gewähren. Je nach gewähltem Datenzugangsweg verlangen die strengen Datenschutzbestimmungen für amtliche Datenprodukte in der Regel, dass präzise räumliche Attribute, wie andere sensible Attribute auch, anonymisiert werden, was die Daten für Zwecke räumlicher Analysen unbrauchbar macht. Einige Forschungseinrichtungen könnten jedoch stark von Forschungsdaten mit präziser Georeferenzierung profitieren. Um diesem Konflikt Abhilfe zu verschaffen, präsentieren und analysieren wir in dieser Thesis fünf verschiedene Methoden der Datensynthese anhand der deutschen Zensusdaten von 2011, mit dem Ziel der Erstellung und Bereitstellung brauchbarer synthetischer Versionen amtlicher deutscher Datenprodukte mit präzisen räumlichen Attributen. Die angewendeten Synthesemodelle sind ein Copula Synthesizer mit Frequency Encoding und ein Copula Synthesizer mit One-Hot Encoding, welche beide parametrische Modelle sind, sowie ein Random Forest Synthesizer und ein CART (Classification and Regression Tree) Synthesizer, welche beide Synthesizer ohne parametrische Annahmen sind, und schließlich Geomasking, eine häufig verwendete Anonymisierungsmethode für georeferenzierte Daten. Wir bewerten das Potenzial der verschiedenen Synthesizer anhand verschiedener Metriken aus dem Bereich der Datensynthese zur Quantifizierung der Data Utility und Data Privacy. Das Ergebnis dieser Thesis ist, dass von allen untersuchten Synthesizern der CART Synthesizer in Bezug auf Data Utility am besten performt und gleichzeitig ein vernachlässigbares Risiko bezüglich Data Privacy aufweist. Es wäre denkbar, einen solchen synthetischen Datensatz über das Remote Access System des Forschungsdatenzentrums für unabhängige Forschungseinrichtungen zugänglich zu machen.