LEARN4SDGis – Machine Learning for Sample Data and Geographic information systems

Kleinräumige Darstellung durch experimentelle Methoden

JOHANNES GUSSENBAUER
INGRID KAMINGER
MATTHIAS TILL
ALEXANDRA WEGSCHEIDER-PICHLER

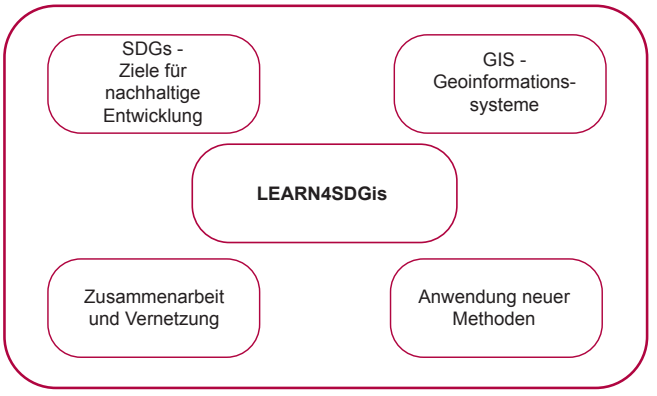
Das Innovationsprojekt "Machine Learning for Sample Data Geographic information systems" (LEARN4SDGis) zielte darauf ab, sozialstatistische Stichprobendaten kleinräumig darzustellen. Insbesondere wurde die kartographische Aufbereitung von Indikatoren erarbeitet, die im Kontext der Sustainable Development Goals (SDGs) der Agenda 2030 verwendet werden können. Diese Zielsetzung wurde durch Anwendung von Maschinenlernmethoden und Integration unterschiedlicher Datenquellen verfolgt. Kartographische Darstellungen zu Armut, Gesundheit und Bildung konnten als erste Ergebnisse gewonnen werden. Da diese Daten hinsichtlich Methodik oder europäischer Harmonisierung noch nicht vollständig ausgereift sind, werden sie als „Experimentelle Statistiken" gekennzeichnet.

Einleitung

Das Projekt „LEARN4SDGis" vereint unterschiedliche Zielsetzungen, die sich bereits im Projektnamen widerspiegeln. Die bereichsübergreifende Zusammenarbeit sowie der Einsatz moderner Methoden (LEARN4), die Sichtbarmachung und Disaggregation von SDG-Indikatoren (Indikatoren der Sustainable Development Goals) sowie die Verbindung von Stichprobendaten und räumlich tief gegliederten Daten auf Raster- und Zählsprengelebene aus dem Bereich der Geoinformationen (GIS) konnten durch das Projekt gefördert werden.

Es unterstützt die Verbreitung von Indikatoren für die Ziele der nachhaltigen Entwicklung der Agenda 2030 (SDGs) durch Kartogramme. Diese Kartogramme haben eine tiefere geographische Darstellung im Vergleich zu regionalen Schätzungen aus Stichprobendaten im Bereich der Sozialstatistik.

Projektüberblick



Ziele des Projektes sind demgemäß:

1) Verbesserung der bereichs- und direktionsübergreifenden Zusammenarbeit innerhalb Statistik Austria für SDG-Indikatoren, Big-Data-relevante Technologien und Datenquellen;
2) Integration auf Stichproben basierender Indikatoren zu Armut, Gesundheit und Bildung sowie räumlich tief gegliederter Datenpakete aus registerbasierten Vollerhebungen (Geoinformationen);
3) Potentialabschätzung für weitere Indikatoren, die auf Stichprobendaten basieren;
4) Test moderner Maschinenlernalgorithmen, die eine große Zahl an Variablen gleichzeitig verarbeiten können;
5) Kartographische Aufbereitung verbesserter regionaler Schätzungen für die SDG-Ziele Armut, Gesundheit und Bildung;
6) Erste Schritte im Bereich der Experimentellen Statistiken.

Das Projekt untersuchte die Verwendung neuer Datenquellen zu räumlichen Verteilungen, Registern und deren Integration mit Stichprobendaten durch Algorithmen des „maschinellen Lernens" (Machine Learning). Die ausgewählten Indikatoren zu Armut, Gesundheit und Bildung sind Teil des nationalen SDG-Indikatorensets.

Machine Learning wird eingesetzt, um synthetische Schätzungen aus Stichprobendaten und relevanter Zusatzinformationen aus registerbasierten Basisdaten abzuleiten. Die Machine-Learning-Resultate sollen die Ergebnisse etablierter Stichprobenerhebungen simulieren, als ob sie aus einer Vollerhebung stammen würden. Folglich zielen die synthetischen Schätzungen auf eine Übereinstimmung mit der offiziellen Statistik auf höherer Ebene ab. Insbesondere stehen die Armutsschätzungen im Einklang mit den Definitionen, die in der Statistik der Europäischen Gemeinschaft über Einkommen und Lebensbedingungen (EU-SILC) verwendet werden. Ebenso ist der Indikator zum Lebenslangen Lernen definitionsgemäß mit demselben aus der Mikrozensus-Arbeitskräfteerhebung (AKE) abgeleiteten Indikator kohärent.

Die Zusammenarbeit und Vernetzung unterschiedlicher Organisationseinheiten von Statistik Austria wurde wesentlich gefördert.

Die Projektleitung war in der Direktion Bevölkerung, im Bereich Analyse und Prognose, angesiedelt. Maßgeblich beteiligt waren die Stabstelle Qualitätsmanagement, Methodik und Klassifikationen, die frühere Stabstelle Analyse, der Bereich Geoinformation in der Direktion Raumwirtschaft sowie die Projektleitungen von Erhebungen in der Direktion Bevölkerung. Hinsichtlich der Veröffentlichungsstrategie wurde auch der Bereich Rechtsangelegenheiten und die Medieninformation aktiv einbezogen.

Indikatoren zu Armut, Gesundheit und Bildung konnten schlussendlich im Rahmen des Projekts auf der Ebene von Zählsprengeln sowie 500-Meter-Rastern in Österreich disaggregiert werden.

Der vorliegende Text wurde im April in ausführlicher Form als Projektbericht laut EU-Grant in englischer Sprache an Eurostat geliefert *(Till et al. 2020)*.

Experimentelle Statistiken

Im Projekt LEARN4SDGis kommen „Experimentelle Statistiken" zur Anwendung. Experimentelle Statistiken nutzen neue Quellen und wenden innovative Methoden der Datenerstellung und -darstellung an, die jedoch noch in Entwicklung sind. Der Vorteil Experimenteller Statistiken ist, dass neue Methoden getestet werden können, die sich nicht am fest vorgegebenen statistischen Rahmen der amtlichen Statistik orientieren müssen und die sich im Lauf der konkreten Projektrealisierung auch (z.B. auf Grundlage von Rückmeldungen durch die Nutzer und Nutzerinnen) noch stark weiterentwickeln können.

Zu beachten ist, dass Experimentelle Statistiken nicht den Status einer offiziell veröffentlichten Statistik haben. Über deren Einschränkung bezüglich der Interpretierbarkeit der Ergebnisse und der Daten muss daher bei jeder einzelnen Experimentellen Statistik genau informiert werden. Entsprechend sind Experimentelle Statistiken auch klar zu kennzeichnen.

Eurostat[1]) und einige andere nationale statistischen Ämter, wie jene der Schweiz, Deutschlands und der Niederlande, sind bereits im Bereich der Experimentellen Statistik tätig. Eurostat verwendet diese beispielsweise, um erstmals Preisentwicklungen in den verschiedenen Abschnitten der Lebensmittelproduktionsketten zu schätzen.

Statistik Austria plant derzeit die Darstellung Experimenteller Statistiken in einem eigenen Bereich auf der Homepage sowie die Kennzeichnung durch ein Logo.

Die Arbeit mit Experimentellen Statistiken ermöglicht langfristig die Erstellung neuer Analysen und die Schaffung neuer Indikatoren. Davon profitiert natürlich generell die Qualität der Statistikerstellung.

[1]) https://ec.europa.eu/eurostat/de/web/experimental-statistics.

Sustainable Development Goals und kleinräumig dargestellte Indikatoren

Die Sustainable Development Goals (SDGs) stehen inhaltlich im Mittelpunkt des Projekts. Diese wurden im September 2015 als „Agenda 2030 für nachhaltige Entwicklung" von der UN-Generalversammlung verabschiedet. Wesentlich für die Umsetzung der Agenda 2030 ist das Monitoring der Zielerreichung; dies soll durch die den 17 Zielen (Goals) bzw. 169 Unterzielen (Targets) zugeordneten Indikatoren geschehen.

Auf nationaler Ebene nehmen unabhängige nationale Statistikinstitute wie Statistik Austria dabei eine zentrale Rolle ein. Dementsprechend hat Statistik Austria 2017 ein erstes österreichspezifisches Indikatorenset auf Basis der UN-Indikatorenvorschläge erarbeitet. Dies geschah unter Berücksichtigung der Vorgaben des Europäischen Statistischen Systems (ESS) und in enger Konsultation mit weiteren nationalen Dateneigentümern wie Umweltbundesamt, Austrian Development Agency sowie Bundesministerien und Bundeskanzleramt. Die nationalen Indikatoren werden jährlich im Dezember aktualisiert und sind auf der Website von Statistik Austria unter dem Themenschwerpunkt „Agenda 2030 - Sustainable Development Goals" abrufbar.

Der Fortschritt in den Zielen soll durch quantitative Indikatoren messbar werden. Dadurch sollen möglichst alle globalen, regionalen, nationalen und auch lokalen Akteure mobilisiert werden. Eine Devise der Agenda 2030 lautet „leave no one behind", was die disaggregierte Darstellung von Daten unter anderem auf kleinräumiger Ebene erfordert. Aus den für Österreich verwendeten Indikatoren wurden in einem weitgehenden Evaluationsprozess fünf (basierend auf EU-SILC bzw. der Mikrozensus-Arbeitskräfteerhebung) für dieses Projekt ausgewählt:

Die ausgewählten Indikatoren mussten die folgenden Kriterien erfüllen:

- Die Indikatoren stammen aus dem nationalen SDG-Set.
- Die Indikatoren beruhen auf Stichprobendaten.
- Datenquelle der Indikatoren muss Statistik Austria sein.
- Die Indikatoren sollen auf nationaler Ebene relevant sein.
- Die Indikatoren sollten durch eine Verbesserung der regionalen Auflösung neue Erkenntnisse liefern.

Bei der Auswahl der Indikatoren wurde die Stichprobengröße berücksichtigt, da eine vernünftige regionale Auflösung mit einer ausreichenden Anzahl von Fällen einhergehen muss.

Folgende fünf Indikatoren wurden für die kleinräumige Schätzung ausgewählt:

Ziel 1.2: Armutsgefährdung (weniger als 60% des Medians des äquivalisierten Haushaltseinkommens); betroffen sind rund 14% der österreichischen Bevölkerung.

Ziel 1.2: Armuts- oder Ausgrenzungsgefährdung (Europa-2020-Zielgruppe: Armutsgefährdung oder erhebliche materielle Deprivation des Haushalts bei einer Liste von neun Merkmalen oder sehr geringe Erwerbsbeteiligung im Haushalt); betroffen sind rund 18% der Bevölkerung.

Ziel 1.4: Erhebliche materielle Deprivation (mehrfache Armuts- oder Ausgrenzungsgefährdung, zwei oder drei Merkmale treffen zu); dies betrifft rund 4% der Bevölkerung.

Ziel 3.4: Subjektiver Gesundheitszustand (Antworten auf eine fünfstufige Skala im EU-SILC Fragebogen); rund 8% der Bevölkerung ab 16 Jahren melden einen schlechten oder sehr schlechten Gesundheitszustand.

Ziel 4.3: Lebenslanges Lernen (Bildungsaktivität in den letzten zwölf Monaten, formale und nicht formale Aktivitäten bzw. Kurse, Personen ab 25 Jahren); rund 35% der Bevölkerung ab 25 Jahren sind bildungsaktiv.

GIS – Geoinformationen

Geoinformationen sind raumbezogene Daten, welche die Gegebenheiten eines Landes beschreiben – sei es in Form von Koordinaten, Zählsprengeln, Postadressen oder anderen Kriterien. Im Rahmen des Projekts LEARN4SDGis werden die Geokoordinaten sowie die daraus abgeleiteten kleinräumigen Statistik-Produkte als Geoinformationen zusammengefasst.

Das Angebot an **kleinräumigen statistischen Daten** hat sich in den vergangenen zwei Jahrzehnten durch die Verknüpfung mit den Gebäudekoordinaten ständig weiterentwickelt, und so zählen kleinräumige Datenprodukte, insbesondere auf Rasterebene, mittlerweile zu den Standardprodukten. Gleichzeitig ist es aber auch wichtig, als statistisches Grundprinzip den Datenschutz stets zu wahren, was bei kleinräumigen Datenprodukten besondere Aufmerksamkeit erfordert.

Für die räumliche Disaggregation der oben erwähnten Indikatoren kamen kleinräumige Basisdaten als Hilfsvariablen zum Einsatz. Eine wesentliche Basis für die kleinräumige Statistik bietet das Gebäude- und Wohnungsregister, das über Gebäudekoordinaten verfügt, über die alle damit verknüpften Daten mit Hilfe von Geoinformationssystemen räumlich flexibel aggregiert werden können.

Die verwendeten Merkmale stammen aus diversen Registern bzw. administrativen Quellen, die Statistik Austria österreichweit auf Personenebene zur Verfügung stehen, und sowohl aggregiert auf Haushaltsebene als auch auf kleinräumiger Raster- und Zählsprengelebene als Hilfsinformationen für die Schätzung herangezogen werden konnten.

Verwendete Datenquellen

Bei Statistik Austria werden alle Stichproben der Sozialerhebungen aus einem sogenannten "Rich Frame" gezogen. Der Rahmen wird aus dem Bevölkerungsregister abgeleitet und regelmäßig aktualisiert. Dieses Register wird regelmäßig mit dem Gebäude- und Wohnungsregister abgeglichen, das detaillierte Geoinformationen (die regionalen Zuordnungen einschließlich geographischer Koordinaten) enthält. Somit können prinzipiell alle Stichproben zum Zeitpunkt der Stichprobenziehung mit den verfügbaren Geoinformationen potentiell angereichert werden. Während Geoinformationen für die Datenerhebung unerlässlich sind, werden solche Informationen in der weiteren Datenverarbeitungsphase aus Datenschutzgründen auch für die statistikinterne Verwendung nicht automatisch bereitgestellt. Daher ist es notwendig, einen allgemeinen Rahmen zu definieren, der die Zusammenführung von Geoinformationen und Stichprobendaten für EU-SILC und eigentlich auch alle anderen von Statistik Austria durchgeführten Sozialerhebungen ermöglicht.

Um die kleinräumige Verteilung der genannten Indikatoren schätzen zu können, wurden EU-SILC-Datensätze aus den Jahren 2014 bis 2018 zusammengeführt und mit **Hilfsvariablen** angereichert. Diese lassen sich grob in drei Gruppen einteilen:

- Variablen (24) aus dem Samplingframe (Grunddaten zu Personen und Haushalten, Beschäftigung und Bildungsinformationen);
- Variablen (15) aus speziellen Einkommensdatensätzen die bereits für EU-SILC verwendet werden: Lohnsteuerdaten, Arbeitnehmerveranlagungsdaten, Pensionsjahresdaten, Qualifikationen laut Dachverband der Sozialversicherungsträger, Transferdaten, Familienbeihilfen- und Kinderbetreuungsgelddaten, Unfallrentendaten, Studienbeihilfendaten und Schülerbeihilfendaten;
- Regionale Datenpakete auf Zählsprengel-Ebene mit 281 Variablen zu den Themen Demographie, Bildung, Erwerbsstatus, Gebäude und Wohnungen (in weiterer Folge auch als Geoinformationen oder GIS-Daten bezeichnet), Daten zur Erreichbarkeit für jede Adresse sowie Schätzungen zu Wohnungs- und Häuserpreisen. Die Daten zur Erreichbarkeit sowie die Wohnungs- und Häuserpreise stehen nur für ein Jahr zur Verfügung und wurden für alle anderen Jahre als *Proxy-Variablen* verwendet.

Die oben angeführten Daten wurden über die sogenannte bereichsspezifische Personenkennzahl (bPk) oder Objektnummern miteinander verknüpft.

Datentransformationen

Bevor Modelle an den Daten zur Anwendung kommen, werden einige Variablen im Datensatz transformiert. Diese Transformationen und die daraus entstehenden *Features* sollen eine verbesserte Darstellung des Modellierungsproblems ermöglichen, womit ein trainiertes Modell die vorliegende Problemstellung besser generalisieren könnte und somit die Vorhersagen für neue, noch nicht im Training verwendete Daten akkurater sind.

Samplingframe-Variablen

Da alle in dieser Arbeit verwendeten Modelle auf Personenebene angewendet wurden, Variablen wie Armut (povmd60) jedoch auf Haushaltsebene definiert sind, ist es für die Problemstellung relevant, Informationen von Personen aus dem gleichen Haushalt in das Modell einfließen zu lassen. So könnte man zum Beispiel ein Feature aus Bildungsebene der Person mit dem höchsten Einkommen aus Verwaltungsdaten ableiten. Nachteil solcher Features ist, dass diese die Information aller weiteren Haushaltsmitglieder vernachlässigen und das je nach Problemstellung unterschiedlich sinnvoll ist.

In dieser Arbeit wurde versucht, eine pragmatische Transformation zu wählen, welche darauf abzielt, möglichst viel Information eines Haushalts zu erhalten. Dabei wird pro Haushalt die Anzahl an Personen pro Ausprägungsmerkmal aller personenbezogenen Variablen gezählt. Die personenbezogenen Variablen umfassen unter anderem Geschlecht, Altersgruppe, höchste abgeschlossenen Bildung sowie das Beschäftigungsverhältnis. Diese Features wurden zusätzlich zu den anderen bestehenden Variablen in die Modelle aufgenommen.

Einkommensvariablen

Die Einkommensvariablen wurden zusätzlich in aggregierter Form als *Features* für die Modellierung verwendet. Diese aggregierte Form ist definiert als die Summe der jeweiligen Einkommensvariablen pro Haushalt dividiert durch die äquivalisierte Haushaltsgröße.

Regionale Datenpakete

Die regionalen Datenpakete auf Zählsprengel-Ebene, welche den größten Teil der verwendeten Geoinformationen ausmachen, werden vor der Modellierung mittels Hauptkomponentenanalyse *(principal component analysis)* transformiert. Damit wird die Anzahl der Variablen von ursprünglich 281 auf 92 reduziert, wobei immer noch zumindest 95% der Varianz in den Daten enthalten bleibt. Diese Transformation dient dazu, die Laufzeit der Modelle zu reduzieren und ein mögliches *overfitting* auf den Trainingsdaten einzudämmen.

Modelle

Folgende Modelle wurden für die Schätzung der Indikatoren getestet:

- Random Forest
- Boosting
- Support Vector Machines
- Neuronale Netze

Bei **Random Forest**, vorgestellt von *Breiman* (2001), wird auf Basis vieler Entscheidungsbäume über die Vorhersagen aller Entscheidungsbäume gemittelt, um eine zuverlässigere Vorhersage zu erhalten. In dieser Arbeit wurde die Implementierung von Random-Forest-Modellen im R-Paket ranger *(Wright und Ziegler 2017)* verwendet. Für das Modell wurde die Anzahl an Bäumen auf 1.000 gesetzt.

Ähnlich wie bei Random Forest verfolgt auch **Boosting** einen sog. *Bagging-Ansatz*. Dabei wird ein Klassifizierungsverfahren viele Male auf einen Datensatz angewendet und über alle vorhergesagten Werte gemittelt. Anders als bei Random Forest werden nach jeder Anwendung die Beobachtungsgewichte vor der nächsten Anwendung, gegebenenfalls einer *Verlustfunktion*, angepasst. In dieser Arbeit wurde für den Boosting-Ansatz das R-Paket xgboost *(Chen et al 2019)* verwendet. Als Hyperparameter wurde die Anzahl an Bäumen auf 1.000, die Lernrate auf 0,01 und die maximale Tiefe jedes Baumes auf 5 gesetzt.

Die Idee von **Support Vector Machines** ist es, eine Hyperebene zu konstruieren, um zwei Klassen *optimal* zu separieren. In dieser Arbeit wurde das R-Pakte liquidSVM *(Steinwart und Thomann 2017)* verwendet. Bei dieser Implementierung wurden der Regularisierungsparameter und die Bandbreite des Kernels mittels 5-fold cross-validation bestimmt.

Neuronale Netze umfassen eine mächtige Familie an Modellen, welche bereits in vielen Anwendungen, wie Bilderkennung oder Textklassifikation, sehr gute Ergebnisse liefern können. Im Zusammenhang mit tabellarischen Daten, wie sie in dieser Arbeit vorliegen, ist nicht gesichert, dass diese Modelle ähnlich erfolgreich eingesetzt werden können. Für die Verwendung Neuronaler Netze wurde das R-Paket keras *(Allaire und Chollet 2019)*, welches ein Frontend zu Tensorflow *(Abadi et al. 2015)* darstellt, verwendet. Die Architektur des verwendeten Neuronalen Netzes findet sich im ausführlichen Projektbericht laut EU-Grant *(Till et al. 2020);* zusätzlich sollte erwähnt werden, dass die Zählsprengel-Datenpakete für das Neuronale Netz nicht mittels principal component analysis transformiert wurden, da die Dimensionsreduktion explizit durch die Architektur eines Neuronalen Netzes abgebildet werden kann.

Für alle Modelle wurden die Daten, sofern notwendig, im Vorfeld skaliert sowie explizite *Features* generiert *(siehe auch den Abschnitt zu Datenquellen)*. Zusätzliche wurden die implementierten Modelle derart verwendet, dass die prognostizierten Werte nicht binäre Variablen, sondern Wahrscheinlichkeiten zwischen 0 und 1 sind. Genauere Erläuterungen zu den einzelnen Modellen finden sich unter anderem in *Hastie, Tibshirani und Friedman* (2001).

Modellevaluierung

Die im letzten Abschnitt erwähnten Modelle wurden mittels Kreuzvalidierungsverfahren miteinander verglichen. Dabei werden die zur Verfügung stehenden Daten in Test- und Trainingsdaten aufgeteilt, das Modell auf den Trainingsdaten trainiert und die Vorhersage auf den Testdaten evaluiert. Für die Kreuzvalidierung wurden die Haushalte im vorliegenden Datensatz in fünf etwa gleich große Gruppen geteilt. Danach wurden aus diesen fünf Gruppen ein Teil als Testdaten und die anderen vier als Trainingsdaten verwendet. Dieses Pro-

zedere wurde viermal wiederholt, womit insgesamt 20 Testresultate konstruiert werden konnten. Für die Evaluation der Testresultate können verschiedene Metriken herangezogen werden, um die Modelle untereinander zu vergleichen.

Vorhersagegüte

Um die Modelle in Bezug auf deren Vorhersagegüte auf Einzelebenen zu vergleichen, wurden folgende Metriken in Betracht gezogen:

- Genauigkeit *(Accuracy)*
- Sensitivität *(Sensitivity);* auch True-Positive-Rate (TPR) oder Recall
- Spezifität *(Specificity);* auch True-Negative-Rate (TNR)
- Mittlerer absoluter Fehler zwischen den vorhergesagten Wahrscheinlichkeiten und der 0-1 Variablen (MAE)
- ROC-Kurve (Receiver Operating Characteristic Curve) und AUC oder AUROC (Fläche unter der ROC-Kurve)
- AUPR; Fläche unter der Precision-Recall Kurve

Die ersten drei – Genauigkeit, Sensitivität und Spezifität – können aus der sogenannten Konfusionsmatrix abgeleitet werden. Um diese Werte jedoch bestimmen zu können, müssen die vorhergesagten Wahrscheinlichkeiten auf 0 oder 1 abgebildet werden.

Für jeden Kreuzvalidierungsschritt wurde ein *cut-off* bestimmt, für den alle Wahrscheinlichkeiten größer als der cut-off auf 1 aufgerundet und sonst auf 0 abgerundet werden. Der cut-off ist in dieser Arbeit definiert durch den Anteil an positiven Fällen in den Trainingsdaten.

Die AUC oder AUPRC hängen nicht von der Wahl eines cut-off-Wertes ab und versuchen zu messen, wie trennscharf die vorhergesagten Wahrscheinlichkeiten sind. Die AUC bezieht sich dabei auf die Fläche unter der ROC-Kurve, welche das Verhältnis zwischen der True-Positive- und False-Positive-Rate darstellt. Wie in *Davis und Goadrich* (2006) dargelegt, verliert die AUC jedoch an Aussagekraft, wenn das beobachtete binäre Merkmal sehr unbalanciert ist. In solchen Fällen kann die AUPR herangezogen werden bei der die Präzision *(Precision)* und der Recall gegenübergestellt werden.

Vergleich mit Schätzern aus den Stichprobendaten

Neben herkömmlichen Gütemaßen werden die Modelle auch danach verglichen, wie gut die vorhergesagten Wahrscheinlichkeiten die Punktschätzer, abgeleitet aus den Stichprobendaten, für ausgewählte soziodemographische Gruppen reproduzieren können. Dieser Vergleich ist insbesondere relevant, da die Ergebnisse für Bevölkerungsgruppen im Vordergrund stehen und die Vorhersagegüte für das Individuum eher zweitrangig ist.

Aus den EU-SILC-Daten von 2016 und 2018 wurden Punktschätzer sowie 95%-Konfidenzintervalle mit dem R-Paket surveysd *(Gußenbauer, Kowarik und de Cillia 2020)* für folgende soziodemographische Gruppen bestimmt:

- Altersgruppe mal Geschlecht
- Gemeindegröße
- Haushaltstyp
- Staatsbürgerschaft und Geburtsland
- Geschlecht des Hauptverdieners bzw. der Hauptverdienerin pro Haushalt
- Bildung

Insgesamt wurden Indikatoren für 42 soziodemographische Subgruppen verwendet.

Die prognostizierten Wahrscheinlichkeiten aus der Kreuzvalidierung wurden pro Individuum gemittelt und danach für jede Subgruppe aggregiert. Dabei wurde das gewichtete Mittel unter Verwendung der Stichprobengewichte verwendet. Danach wurde über alle Subgruppen gezählt, wie oft der Schätzer, abgeleitet aus den prognostizierten Wahrscheinlichkeiten, innerhalb des 95%-Konfidenzintervalls des Punktschätzers aus den Stichprobendaten liegt.

Ergebnisse der Kreuzvalidierung

Für alle Kreuzvalidierungsschritte wurden die Modelle unter Verwendung unterschiedlicher Variablensets trainiert, einmal nur unter Verwendung der Variablen aus dem Stichprobenframe (~Frame) und dann noch zusätzlich mit den Zählsprengel-Daten (Geoinformationen/~GIS) bzw. Einkommensvariablen (~Income).

Nebenstehende *Grafik 1* zeigt die Ergebnisse der Kreuzvalidierung für die Variable Armutsgefährdung (povmd60). Für jedes Gütemaß sowie pro Modell und Variablenkombination ist die Verteilung der Werte der Kreuzvalidierung als Boxplot dargestellt. Es ist deutlich zu erkennen, dass sich die Güte der Schätzungen viel weniger nach dem jeweils angewendeten Algorithmus als nach den jeweils verwendeten Inputdaten unterscheidet. Insbesondere die Verwendung von Einkommensvariablen kann die Güte der Schätzungen massiv steigern.

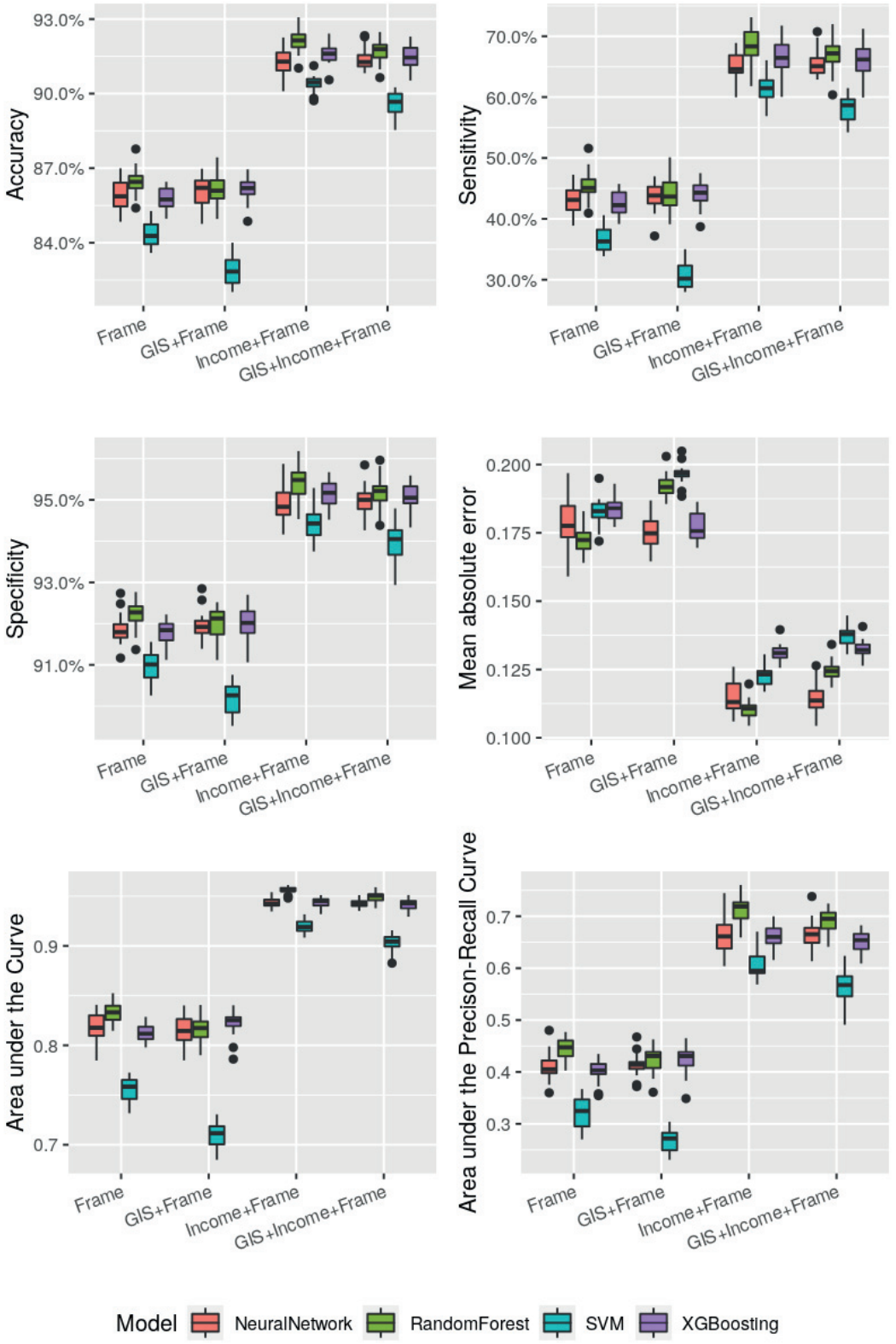
Zum Beispiel zeigen die Ergebnisse für Genauigkeit *(Accuracy)* im oberen linken Panel unabhängig von der Algorithmuswahl, dass etwa 91% bis 92% der Individuen aus den Testdaten korrekt klassifiziert werden, falls Einkommensvariablen in die Schätzung mit aufgenommen werden. Als Vergleich dazu: Ein uninformatives Modell, welches rein zufällig den Status povmd60=1 für etwa 12,19% der Testdaten verteilt würde erwartungsgemäß eine Genauigkeit von 78,59% erzielen.

Ähnliche Interpretationen wie oben lassen sich auch auf die Ergebnisse der anderen Gütemaße übertragen. Zu erwähnen sei noch, dass die Hinzunahme der Informationen auf Zählsprengel-Ebene (Geoinformation/~GIS) anscheinend keinen bzw. in manchen Fällen einen – eher sehr geringen – negativen Effekt auf die Vorhersagegüte hat. Es könnte sein, dass dies ein Zeichen für eine sogenannte Überanpassung *(overfitting)* der Modelle ist.

Ergebnis der Kreuzvalidierung für die Variable Armutsgefährdung (povmd60)

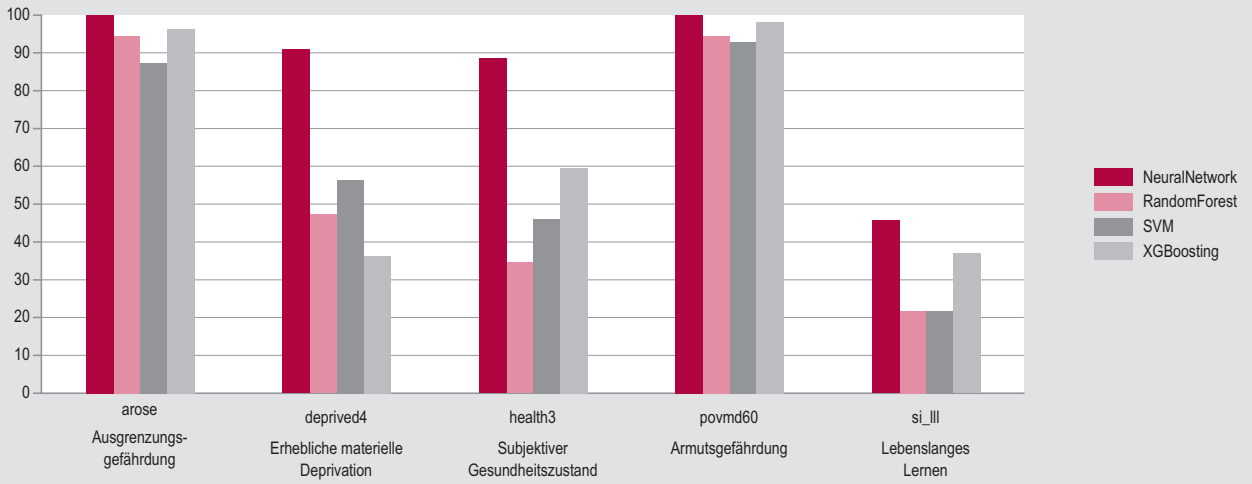
Grafik 1

Jedes Panel zeigt für jedes Modell sowie jede Inputvariablenkombination die Verteilung der Werte des jeweiligen Gütemaßes als Boxplot.



Q: STATISTIK AUSTRIA, Experimentelle Statistiken.

Anteil der modellbasierten Schätzungen, innerhalb des 95%-Konfidenzintervalls der Stichprobenschätzer Grafik 2



Q: STATISTIK AUSTRIA, Experimentelle Statistiken.

Für die anderen Variablen wie Armuts- oder Ausgrenzungsgefährdung, subjektiver Gesundheitszustand oder Lebenslanges Lernen zeigt sich ein sehr ähnliches Bild, wobei der Einfluss der Einkommensvariablen nicht in allen Fällen so ausgeprägt ist.

Neben konventionellen Gütemaßen wird in diesem Projekt auch eine Gegenüberstellung der prognostizierten Wahrscheinlichkeiten mit Punktschätzern abgeleitet aus den Stichprobendaten betrachtet.

Grafik 2 zeigt für alle Indikatoren den Anteil der Fälle bei denen die Schätzer abgeleitet von den prognostizierten Wahrscheinlichkeiten innerhalb des 95% Konfidenzintervall liegt. Die Modelle bei diesem Vergleich wurden stets auf allen Inputvariablen trainiert.

Aus der Grafik ist klar zu erkennen, dass die Modelle für jene Indikatoren, die direkt von den Einkommensdaten abhängen, wie Armutsgefährdung (povmd60) sowie Armuts- und Ausgrenzungsgefährdung (arose), sehr gute Ergebnisse erzielen. Für povmd60 erzielt jedes der Modelle einen Wert über 90% und Neuronale Netze sogar 100%. Für andere Indikatoren sinkt diese Zahl drastisch, was damit zusammenhängen könnte, dass die Punktschätzer im einstelligen Prozentbereich liegen und die dazugehörigen Konfidenzintervalle sehr schmal sind. Auch ist die Datengrundlage für manche der soziodemographischen Gruppen nicht sehr groß.

Für den Indikator Lebenslanges Lernen (si_lll) zeigt sich zudem, dass die Ergebnisse der Modelle hier deutlich von den Indikatoren aus EU-SILC abweichen. Eine mögliche Ursache dafür liegt vermutlich in der Wahl der Hyperparameter der Modelle sowie der Features, die für das Modelltraining verwendet wurden. Die hier gezeigten Ergebnisse lassen vermuten, dass die Hyperparameter und Features nicht allgemein auf Indikatoren aus verschiedenen Stichproben übertragbar sind.

Für den Vergleich mit den Schätzern aus den Stichprobendaten scheint das Neuronale Netz am besten abzuschneiden. Eine noch ausführlichere Diskussion der Ergebnisse findet sich im Projektbericht laut EU-Grant *(Till et al. 2020)*.

Obgleich es leichte Performanceunterschiede zwischen den Modellen gibt, scheint die Wahl der Inputvariablen von viel größerer Relevanz zu sein. Die Wahl des Modells für die Anwendung auf den gesamten Stichprobenrahmen (~österreichische Population) sollte daher auf jene Methode fallen, die technisch einfach umzusetzen ist.

Der Random-Forest-Algorithmus in *Wright und Ziegler* (2017) scheint hier eine plausible Wahl zu sein, da dieser schnell und effizient implementiert ist.

Übertragung der Schätzung auf Samplingframe

Mit dem Random-Forest-Algorithmus aus *Wright und Ziegler* (2017) wurden für alle Individuen aus dem Samplingframe sowie für jeden Indikator Wahrscheinlichkeiten berechnet, die angeben, wie wahrscheinlich eine Person von einem Merkmal betroffen ist. Das Modell wurde für jeden Indikator mit allen Prädiktoren, also inklusive Einkommensdaten und Zählsprengel-Daten, trainiert.

Insbesondere wurde beim Training der zur Verfügung stehende Datensatz in zehn etwa gleich große Teile geteilt und das Modell zehnmal trainiert, wobei jedes Mal einer der zehn Teildatensätze für das Training ausgeschlossen wurde. Damit ergaben sich pro Individuum zehn prognostizierte Wahrscheinlichkeiten, welche abschließend gemittelt wurden, um eine endgültige prognostizierte Wahrscheinlichkeit zu erhalten.

Für die kleinräumigen Ergebnisse der Indikatoren können diese Wahrscheinlichkeiten innerhalb räumlicher Einheiten aggregiert und durch die Anzahl an Personen innerhalb der räumlichen Einheit dividiert werden. Davor werden diese Wahrscheinlichkeiten jedoch noch geglättet und kalibriert.

Glättung und Kalibrierung von Wahrscheinlichkeiten

Es ist nicht auszuschließen, dass benachbarte kleinräumige Regionen sehr unterschiedliche Werte für die einzelnen Indikatoren haben. Um diesen Effekt auszugleichen, wurden die Indikatorwerte auf Ebene der Zählsprengel berechnet und diese anschließend geglättet. Jeder Zählsprengel hat dabei einen neuen Wert erhalten, welcher ein Mittel aus dem eigenen Wert und jenen der drei *nächsten* Zählsprengel ist. Als Distanzvariablen wurde die Anzahl der Personen im Zählsprengel sowie die X- und Y-Koordinate des Zählsprengelmittelpunktes verwendet.

Diese *neuen* Werte wurden anschließend auf die einzelnen Individuen pro Zählsprengel aufgeteilt, um somit geglättete prognostizierte Wahrscheinlichkeiten auf Personenebene zu erhalten.

Eine weitere Korrektur der Wahrscheinlichkeiten wurde vorgenommen, damit die regionalen Schätzungen konsistent mit den Ergebnissen aus der jeweiligen Stichprobe sind. Der nationale Indikatorwert für den Anteil an Armutsgefährdung unter Verwendung der prognostizierten Wahrscheinlichkeiten, noch vor der Glättung, beträgt 15,17%, wohingegen das gepoolte Ergebnis aus der Jahren 2016 bis 2018 nur 14,30% beträgt.

Um konsistent mit den Stichprobenergebnissen zu sein, wurden die prognostizierten Wahrscheinlichkeiten so angepasst, dass diese die Stichprobenergebnisse auf nationaler sowie auf Bundesländerebene reproduzieren können.

Gruppierung von Ergebnissen

Für die Ergebnisdarstellung werden die einzelnen Raten pro regionale Einheit in diskrete Gruppen eingeteilt. Dies hilft regionale Unterschiede hervorzuheben, unterdrückt jedoch auch sehr hohe oder niedrige Werte, welche unplausibel sein dürften.

Die Ergebnisse dieser Arbeit werden für alle Indikatoren in **fünf Kategorien** dargestellt. Dabei ist die mittlere Kategorie symmetrisch um das jeweilige nationale Ergebnis θ mit Länge $\pm l\%$

$$[\theta - \frac{l}{2}, \theta + \frac{l}{2}]$$

Die Kategorien darüber und darunter sind ebenso l Prozentpunkte lang usw. Für erhebliche materielle Deprivation wurde $l = 2$ und für alle anderen $l = 5$ gewählt.

Für den Indikator Armutsgefährdung, bei einem nationalen Ergebnis von 14,30%, ergeben sich somit folgende fünf Kategorien

[0,6.8) [6.8,11.8) [11.8,16.8] (16.8,21.8] (21.8,100]

Jedes regionale Ergebnis kann somit einer dieser fünf Kategorien zugeordnet werden.

Unterdrückung von Ergebnissen

Bei der Ergebnisdarstellung sind insbesondere die statistische Geheimhaltung als auch das Unterdrücken von möglicherweise *unplausiblen* Werten zu berücksichtigen. Bei den Ergebniskarten wird der Wert aller regionalen Einheiten mit weniger als 50 gemeldeten Personen unterdrückt. Zusätzlich wird ein 95%-Pseudo-Konfidenzintervall aus den prognostizierten Wahrscheinlichkeiten pro regionale Einheit bestimmt und alle regionalen Ergebnisse mit einem Konfidenzintervall, das breiter als $2 \cdot l$ ist, unterdrückt. Damit werden also jene Regionen unterdrückt, bei denen das Konfidenzintervall über mehr als zwei Kategorien abdeckt. Dieses Pseudo-Konfidenzintervall berücksichtigt nur einen kleinen Teil der Fehlerquellen, weswegen noch zusätzliche Forschung notwendig ist, um dies zu erweitern.

Mehr Details zur Konstruktion des Konfidenzintervalls finden sich im ausführlichen Projektbericht laut EU-Grant *(Till et al. 2020)*. Weitere Literatur zum methodischen Vorgehen findet sich am Ende des Beitrags.

Experimentelle Statistiken – Veröffentlichung der Ergebnisse im Internet

Der Projektbericht des als Grant geförderten Projekts wurde Ende März 2020 an Eurostat übermittelt. Zur Datenvalidierung der regionalisierten Ergebnisse wurden die Landesstatistiker und -statistikerinnen im Rahmen des Fachbeirats für Sozialstatistik informiert, zudem wurde eine Vorabversion von Ergebnissen mit dem Ersuchen um Information über Auffälligkeiten übermittelt.

Das endgültige Format der Veröffentlichung wird auf der Grundlage der Rückmeldung der Landesstatistiker und -statistikerinnen und möglicher methodischer Überarbeitungen, einschließlich z.B. Unterdrückungsregeln und Klassifizierungsintervalle, entschieden.

Die Veröffentlichung bei Statistik Austria soll in Form eines eigenen SDG-Atlas (nach Vorbild des STATatlas) erfolgen. Dabei sind die folgenden Karten lediglich als Illustrationen der möglichen Visualisierung zu verstehen, während die tatsächlichen Ergebnisse aus dem zukünftig veröffentlichten Atlas abgeleitet werden müssen. Dies entspricht auch dem Wesen der Experimentellen Statistiken, die einer ständigen Weiterentwicklung unterliegen sollten, und zwar im besten Fall, bis die Methodik den Anforderungen der offiziellen Statistikprodukte entspricht.

Eine spezielle Kennzeichnung als Experimentelle Statistik sowie begleitende Dokumentationen sind bei der Veröffentlichung unbedingt notwendig. Aktuell ist ein eigener Bereich auf der Homepage von Statistik Austria in Vorbereitung. Ein entsprechendes Label wurde soeben erarbeitet.

EXPERIMENTELLE STATISTIK

Zusammenfassend sollen bei der Veröffentlichung **folgende Prinzipien** verfolgt werden:

- Transparente und reproduzierbare Methodik
- Systematische Bewertung der Genauigkeit
- Raster/Zählsprengel als kleinste Disaggregationsebene
- Kennzeichnung der Ergebnisse als experimentell zur Unterscheidung von offiziellen Ergebnissen
- Maximale Kohärenz mit veröffentlichten amtlichen Statistiken
- Ergebnisunterdrückung bei unzuverlässigen Verdachtsfällen

Erste Ergebnisse im Rahmen des EU-Grants

Exemplarisch werden nachfolgend einzelne Kartogramme aus dem Projektbericht laut EU-Grant *(Till et al. 2020)* gezeigt, die den Indikator der Armutsgefährdung (zu SDG Unterziel 1.2) kleinräumig darstellen.

Zu beachten ist, wie erwähnt, dass sich Änderungen im oben angeführten methodischen Vorgehen auch in der Ergebnisdarstellung auswirken können. Derzeit umfassend diskutiert werden beispielsweise die Klassifizierungsintervalle oder das beste Vorgehen der Unterdrückung. Im Vergleich zum EU-Grant können daraus auch noch Änderungen in den Ergebnissen resultieren.

Im Projektbericht wurden z.B. verschiedene Gruppierungen getestet (z.B. fünf, sechs oder sieben Gruppen) oder die Farbgestaltung. Zudem sind verschiedene regionale Gliederungen dargestellt.

Nachfolgend ist die Armutsgefährdung nach dem Verstädterungsgrad (degree of urbanisation;[2]) *Grafik 3)* sowie auf Gemeindeebene dargestellt *(Grafik 4)*. Zudem wird sie in Wien und im Umland von Wien auch auf Zählsprengelebene gezeigt *(Grafik 5)*.

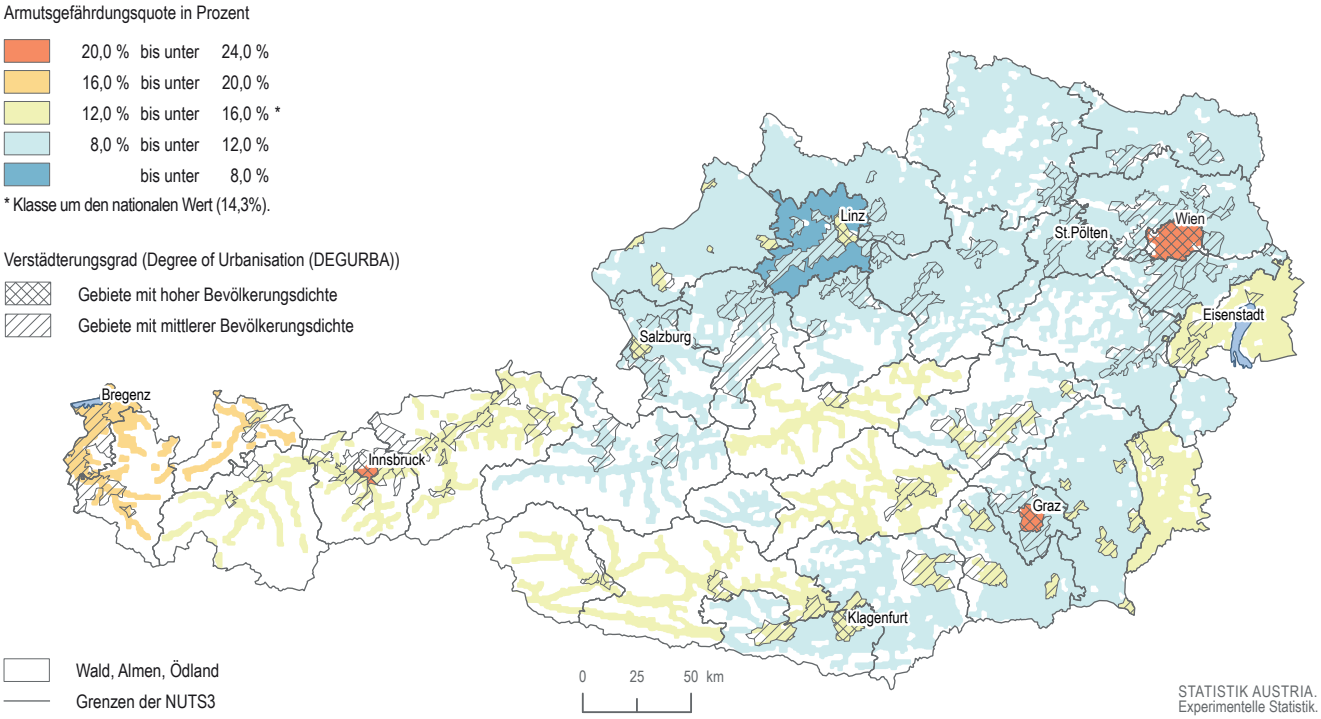
Die Ergebnisse des Projekts ermöglichen einen tiefen Einblick in räumliche Muster. Die nachfolgende Karte zeigt den Indikator zur Armutsgefährdung in den nach dem Grad der Verstädterung (NUTS 3 x DEGURBA).

Es zeigt sich, dass auch nach der kleinräumigen Darstellung die städtischen Ballungsräume Wien, Graz und Innsbruck am stärksten von Einkommensarmut betroffen sind. Gleichzeitig zeichnen sich die Gebiete rund um die oberösterreichische Landeshauptstadt Linz durch besonders niedrige Armutsraten aus *(Grafik 3)*.

Eine noch weitergehende Disaggregation auf Gemeindeebene zeigt, dass die Armutsgefährdung in vier Wiener Bezirken in der höchsten Kategorie liegt, während für viele Gemeinden in Nieder- und Oberösterreich die Armutsrate nur auf 4% bis 8% geschätzt wird *(Grafik 4)*. Aber auch einige Gemeinden in Grenzregionen zu Deutschland scheinen mit besonders hohen Schätzungen auf.

[2]) Der Verstädterungsgrad (DEGURBA) ist ein Kriterium zur Charakterisierung eines Gebietes. Auf der Grundlage des Anteils der lokalen Bevölkerung in städtischen Ballungsgebieten und städtischen Zentren werden sog. Lokale Verwaltungseinheiten (LAU) in drei Gebietstypen eingeteilt: **Städte** (dicht besiedelte Gebiete); **Kleinere Städte und Vororte** (Gebiete mit mittlerer Bevölkerungsdichte); **Ländliche Gebiete** (dünn besiedelte Gebiete).

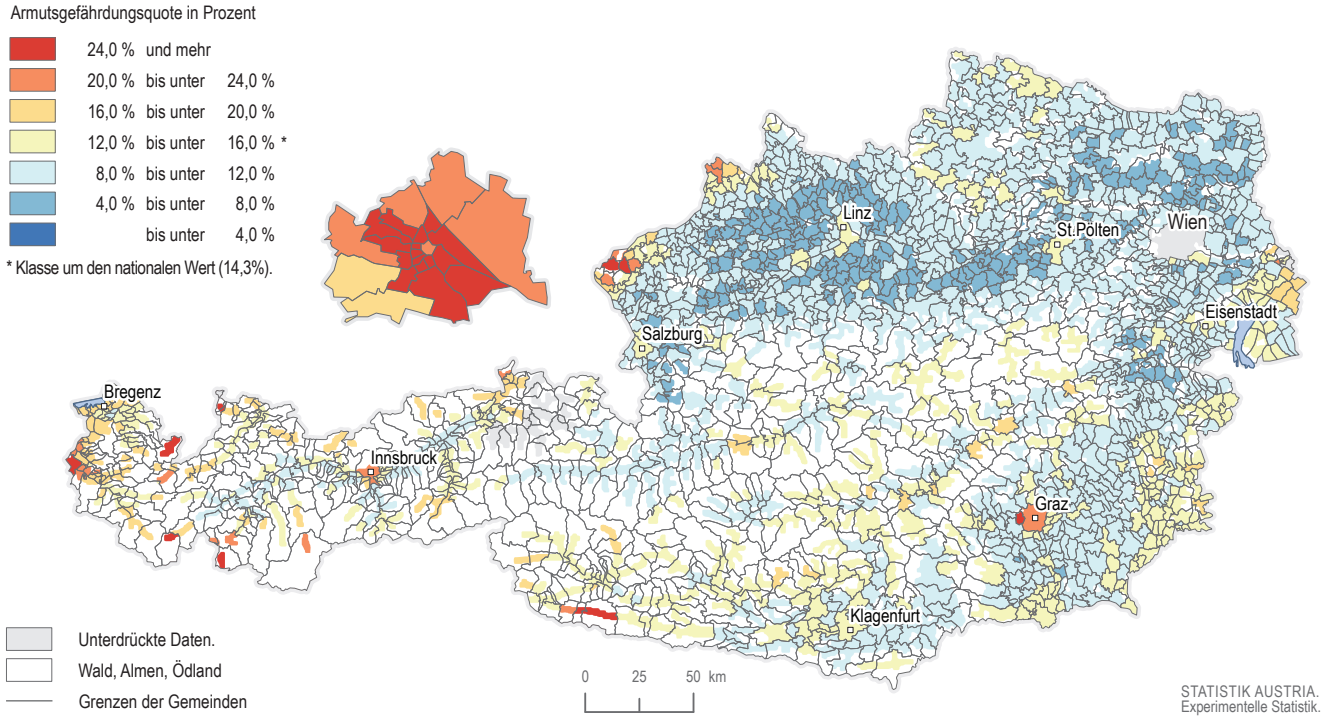
Armutsgefährdung in Österreich nach dem Grad der Verstädterung Grafik 3



Q: STATISTIK AUSTRIA, Experimentelle Statistiken nach EU-Grant 08143.2017.001-2017.403.

Armutsgefährdung in Österreich auf Gemeindeebene

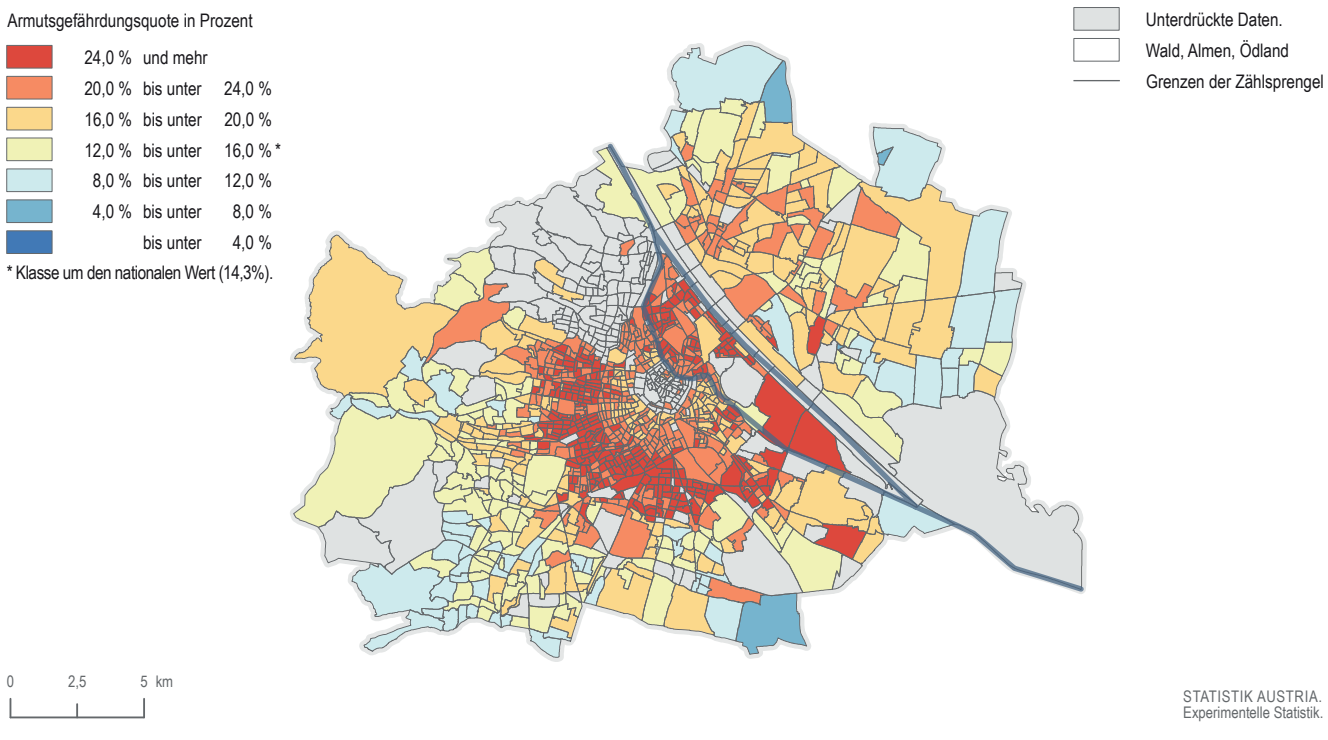
Grafik 4



Q: STATISTIK AUSTRIA, Experimentelle Statistiken nach EU-Grant 08143.2017.001-2017.403.

Armutsgefährdung in Wien und Umgebung auf Zählsprengelebene

Grafik 5



Q: STATISTIK AUSTRIA, Experimentelle Statistiken nach EU-Grant 08143.2017.001-2017.403.

In diesen Regionen könnten die Schätzungen aufgrund des Fehlens von Registerinformationen und der Knappheit an Grenzgängern in der EU-SILC-Ausbildungsstichprobe verzerrt sein.

Angesichts der Topographie Österreichs erwies sich diese Darstellung als nicht optimal. Erstens sind große Teile Österreichs bergig, und tatsächlich bewohnt sind dort Täler, die sich nur über ein relativ kleines Gebiet erstrecken. Infolgedessen dominieren in bestimmten Regionen einige Gemeinden, die nur von einer relativ kleinen Bevölkerung bewohnt sind, aber ein großes Gebiet abdecken, das Bild. Auch die Darstellung der Schätzungen nach Gemeinden ist politisch sensibel, insbesondere wenn diese für benachbarte Gemeinden deutlich voneinander abweichen. Zwei Situationen erscheinen besonders problematisch. Das erste Problem besteht darin, dass der Schwellenwert für die Klassifizierung im Wesentlichen willkürlich ist und eine andere Farbe in der Karte möglicherweise keinen statistisch robusten Unterschied bedeutet. Zweitens können Disparitäten innerhalb von Gemeinden verdeckt sein. Beispielsweise kann es Konzentrationen innerhalb des Gebiets einer Gemeinde geben, die übersehen werden.

Dies zeigt den Charakter und vor allem den Mehrwert von Experimentellen Statistiken, wo erste Ergebnisse Schwierigkeiten in der angewendeten Methodik oder der Datengrundlage aufzeigen, welche dazu benutzt werden können, entsprechende Weiterentwicklungen und Verfeinerungen durchzuführen.

Die *Grafik 5* zeigt den Indikator der Armutsgefährdung auf Zählsprengelebene in und um Wien. Hier wird deutlich, dass sich die Armutsgefährdung in Stadtteilen außerhalb und in der Nähe des äußeren Gürtels konzentriert. Diese Darstellung unterdrückt für eine Reihe von Zählsprengeln die Schätzungen: Hier leben entweder weniger als 50 Einwohner und Einwohnerinnen, mehr als ein Drittel der Einwohner und Einwohnerinnen sind ohne jegliche Registerangaben, oder es wird eine ungewöhnlich hohe Armutsquote (>20%) erreicht, während die Immobilienpreise mindestens doppelt so hoch sind wie im Durchschnitt der Stadt.

Obwohl die Zählsprengel eine zentrale Rolle für die Modellspezifikation spielten, da die meisten geographischen Merkmale auf dieser Ebene aggregiert wurden, wird voraussichtlich davon abgesehen, Ergebnisse auf dieser Ebene zu verbreiten. Zählsprengel haben in der Regel eine ähnlich große Einwohnerzahl, können aber in ihrer Fläche stark variieren, was die Ergebnisinterpretation erschweren könnte.

Weitere Ergebnisse zu Armutsindikatoren, dem Subjektiven Gesundheitszustand und dem Lebenslangen Lernen werden zukünftig über den SDG-Atlas zur Verfügung stehen. Aktuell geplant ist, die Indikatoren nach NUTS-3-Regionen, Politischen Bezirken und nach verschiedenen Rastergrößen bis Ende 2020 zu veröffentlichen.

Literatur

Abadi, Martín / Ashish Agarwal / Paul Barham / Eugene Brevdo / Zhifeng Chen / Craig Citro / Greg S. Corrado et al. (2015): "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems". http://tensorflow.org/.

Allaire, JJ. and Chollet, François (2019): "*Keras: R Interface to 'Keras'* ". https://CRAN.R-project.org/package=keras.

Breiman, Leo (2001): "Random Forests." Machine Learning 45 (1): 5-32. https://doi.org/10.1023/A:1010933404324.

Chen, Tianqi / Tong He / Michael Benesty / Vadim Khotilovich / Yuan Tang / Hyunsu Cho / Kailong Chen et al. (2019): "Xgboost: Extreme Gradient Boosting". https://CRAN.R-project.org/package=xgboost.

Cheng, Heng-Tze / Levent Koc / Jeremiah Harmsen / Tal Shaked / Tushar Chandra / Hrishi Aradhye / Glen Anderson et al. (2016): "Wide & Deep Learning for Recommender Systems" CoRR abs/1606.07792. http://arxiv.org/abs/1606.07792.

Davis, Jesse and Goadrich, Mark (2006): "The Relationship Between Precision-Recall and ROC Curves". In ICML '06: Proceedings of the 23rd International Conference on Machine Learning, 233-40. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1143844.1143874

Gower, J.C. (1971): "A General Coefficient of Similarity and Some of Its Properties" Biometrics 27 (4): 857-71. http://www.jstor.org/stable/2528823.

Greenwell, Brandon / Bradley Boehmke / Jay Cunningham / GBM Developers (2019). Gbm: Generalized Boosted Regression Models. https://CRAN.R-project.org/package=gbm.

Gussenbauer, Johannes / Kowarik, Alexander / de Cillia, Gregor (2020): "Surveysd: Survey Standard Error Estimation for Cumulated Estimates and Their Differences in Complex Panel Designs". https://github.com/statistikat/surveysd.

Hastie, Trevor and Pregibon, Daryl (1992): "Generalized Linear Models" In Statistical Models in S. Vol. 6.

Hastie, Trevor / Tibshirani, Robert / Friedman, Jerome (2001): "The Elements of Statistical Learning". Springer Series in Statistics. New York, USA: Springer New York Inc.

Kowarik, Alexander and Templ, Matthias (2016): "Imputation with the R Package Vim." Journal of Statistical Software, Articles 74 (7): 1-16. https://doi.org/10.18637/jss.v074.i07.

Meyer, David / Evgenia Dimitriadou / Kurt Hornik / Andreas Weingessel / Friedrich Leisch (2019): "E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)", TU Wien. https://CRAN.R-project.org/package=e1071.

Mikolov, Tomas / Ilya Sutskever / Kai Chen / Greg Corrado / Jeffrey Dean (2013): "Distributed Representations of Words and Phrases and Their Compositionality". In Advances in Neural Information Processing Systems 26, edited by C.J.C. Burges, L. / Bottou, M. / Welling, Z. Ghahramani / K.Q. Weinberger, 3111-9. Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

R Core Team (2019): "R: A Language and Environment for Statistical Computing". Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Steinwart, Ingo and Thomann, Philipp (2017): "liquidSVM: A Fast and Versatile Svm Package." ArXiv e-prints 1702.06899, February. http://www.isa.uni-stuttgart.de/software.

Till, Matthias / Bilek, Franz / Bienzle, Christine / Glaser, Thomas / Gußenbauer, Johannes / Hofer, Nina / Kaminger, Ingrid / Kowarik, Alexander / Saul, Sibylle / Wegscheider-Pichler, Alexandra (2020): „LEARN4SDGis – Machine Learning for Sample Data Geographic information systems". Eurostat Grant Agreement Number: 08143.2017.001-2017.403, Final Report, Eurostat / Statistics Austria.

Wright, Marvin and Ziegler, Andreas (2017): "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". Journal of Statistical Software, Articles 77 (1): 1-17. https://doi.org/10.18637/jss.v077.i01.

Summary

The inventive project "Machine Learning for Sample Data Geographic information systems" (LEARN4SDGis) aimed at presenting sample data from social statistics on a small scale. In particular, a cartographic presentation of indicators was developed. This was done in the context of the Sustainable Development Goals (SDGs) of the UN 2030 Agenda. For this machine learning methods were applied while integrating different data sources. Cartographic presentations of poverty, health and education were obtained as first results. As these data are not yet fully developed in terms of methodology or harmonization in European context, they are labeled "Experimental Statistics".