Methodische Aktualisierung des LEARN4SDGis-Projekts

(Stand 11.3.2021)

Johannes Gussenbauer

Matthias Till

Einleitung

Das Projekt “Machine Learning for Sample Data Geographic information systems” (LEARN4SDGis) war eine durch EUROSTAT finanzierte Machbarkeitsstudie (Merging Statistics and Geospatial Information Grant 2018-2020). Ziel war die kleinräumige Darstellung von Indikatoren die auf Stichprobendaten basieren. Untersucht wurden insbesondere Indikatoren für die Sustainable Development Goals (SDGs) der Agenda 2030 für nachhaltige Entwicklung. Der methodische Fokus lag auf der Verwendung von Maschinlernmethoden sowie die Integration unterschiedlicher Datenquellen. Maschinlernalgorithmen wurden darauf trainiert, Befragungsmerkmale vorherzusagen. Die Vorhersage stützte sich dabei auf Geoinformationen und Registerdaten, die für die Grundgesamtheit der österreichischen Wohnbevölkerung in Privathaushalten verfügbar sind. Die benötigten Individualmerkmale wurden anschließend in unterschiedlich detaillierte räumliche Darstellungen aggregiert. Ein Vergleich unterschiedlicher Algorithmen und Modellspezifikationen sowie Ergebnisse des Projektes sind sowohl in einem englischsprachigen Projektbericht (Till et al. (2020)) als auch in einer in Statistische Nachrichten publizierten deutschsprachigen Zusammenfassung beschrieben (Gussenbauer et al. (2020)). Das Projekt wurde im Frühjahr 2020 abgeschlossen. Die im Projektbericht beschriebene Methodik wurde seitdem weiterentwickelt. Im Frühjahr 2021 wurden die Ergebnisse schließlich in Form eines interaktiven Atlas als experimentelle Statistiken veröffentlicht[1]. Mit dem vorliegenden Kurzbericht werden die seit Projektende vorgenommenen Erweiterungen bzw. Änderungen der Methodik dokumentiert. Modellierungen und Analysen wurden mit der Programmiersprache R durchgeführt, siehe R Core Team (2019).

Datenquellen

Die verwendeten Datenquellen und Variablen wurden bereits detailliert im Anhang zum Projektbericht an Eurostat beschrieben (Till et al. (2020)). Seit Projektende wurden die zur Verfügung stehenden Trainingsdaten um die Stichprobendaten für EU-SILC 2019 erweitert. Die Modelle für die SDGs Armutsgefährdung, Armuts- oder Ausgrenzungsgefährdung,

[1] https://www.statistik.at/atlas/sdg/

erhebliche materielle Deprivation sowie subjektiver Gesundheitszustand werden somit anhand von EU-SILC 2014-2019 trainiert.

Modellwahl

Für die Modellierung der SDG-Indikatoren kommt nun ein verbessertes Neuronales Netz zum Einsatz. Diese Entscheidung ist insbesondere dadurch motiviert, dass die Schätzungen der Kreuzvalidierung besser mit den Punktschätzern aus den Erhebungsdaten, zu ausgewählten soziodemographischen Gruppen, übereinstimmen. Bei dem hier verwendeten Neuronal Netz handelt es sich um ein sog. *feedforward*-Netz bei dem *convolutional* sowie *fully-connected* Layer verwendet wurden. Da ein Großteil der hier betrachteten SDGs von der Haushaltsstruktur abhängt wurde der Input für personenspezifische Variablen für jede Person als Matrix dargestellt.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{8,1} & x_{8,2} & \cdots & x_{8,p} \end{bmatrix}$$

Die erste Zeile der Matrix $(x_{1,1}, x_{1,2}, \dots, x_{1,p})$ beinhaltet die Variablen der jeweiligen Person und alle weiteren Zeilen enthalten die Information von bis zu 7 weiteren Personen die im gleichen Haushalt gemeldet sind. Bei Haushalten mit weniger als 8 Personen wurde die Matrix mit 0 aufgefüllt. Dieser Input wird mit Hilfe von convolutional Layers im Neuronalen Netz verarbeitet um somit Variable abzuleiten die die Haushaltsstruktur berücksichtigen. Der Aufbau des Neuronalen Netzes ist in Abbildung 1 dargestellt. Die Hyperparameter wurden mittels Rastersuche über eine endliche Menge an Parametern für jeden SDG Indikator separat bestimmt. Für den Output-Layer wurde Aktivierungsfunktion *Softmax* und für alle anderen Layer eine sog. *Rectifier* verwendet.

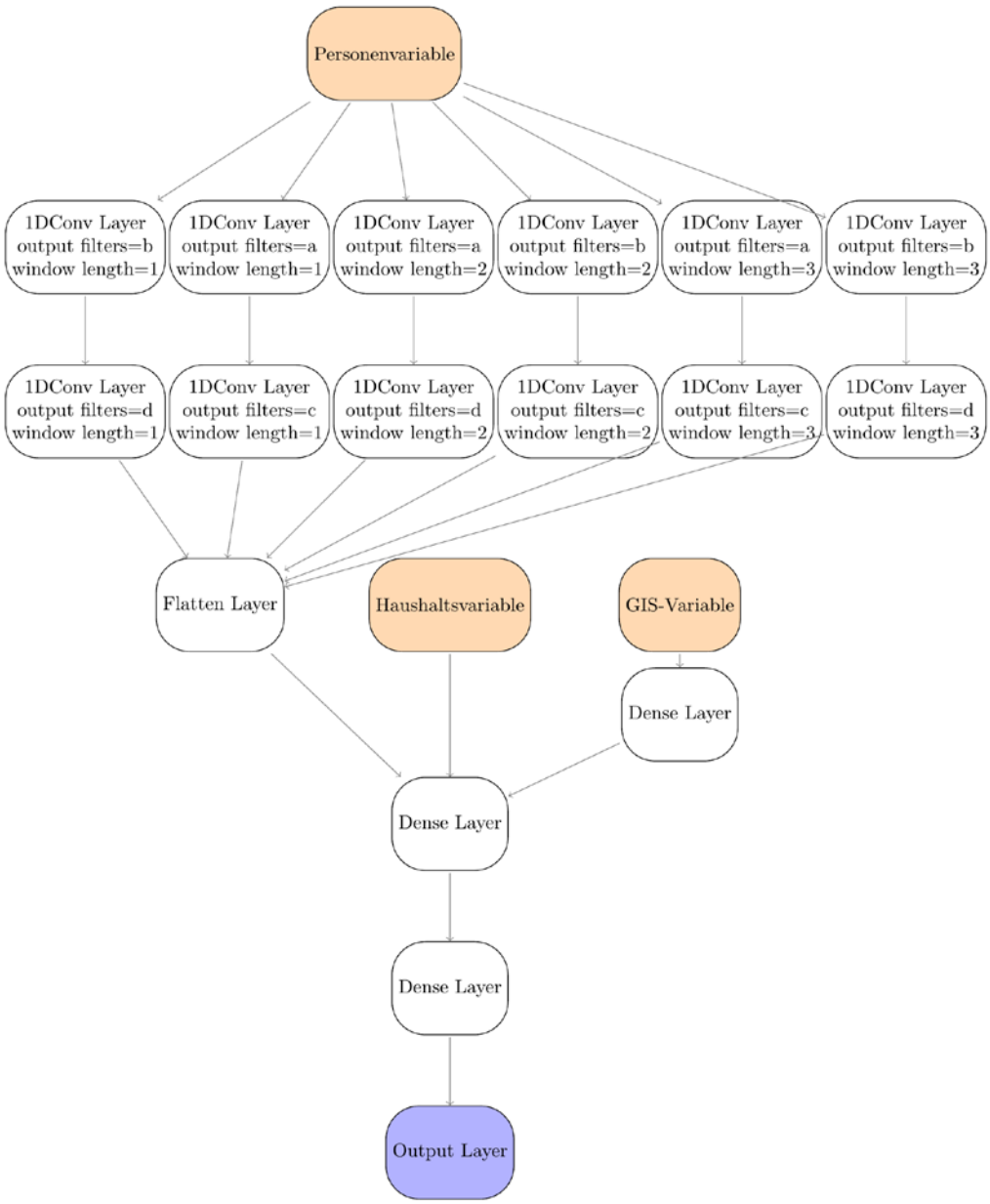

Abbildung 1: Architektur des Neuronalen Netzes

Kreuzvalidierung auf Stichprobendaten

Die Vorhersagegüte des Modells wurde analog zur Vorgangsweise im Endbericht an Eurostat (Till et al. (2020)) mittels Kreuzvalidierung evaluiert. Im Folgenden wird das Modell mit dem im ursprünglichen Projektbericht empfohlenen Random Forest Modell verglichen. Für beide Modelle wurden die Stichprobengewichte für das Modelltraining berücksichtigt. Für die Implementierungen der Modelle wird auf Wright und Ziegler (2017), Allaire und Chollet (2019) sowie Abadi et al. (2015) verwiesen.

Tabelle 1: Qualitätsindikatoren für die Modellvorhersage von Neural Network und Random Forest auf individueller Ebene

		povmd60	arose	deprived4	health3	si_lll
Genauigkeit (%)	Neural Network	90.67	89.02	95.09	89.67	81.77
	Random Forest	91.01	89.56	95.02	90.19	80.83
Sensitivität (%)	Neural Network	68.10	70.88	28.68	40.59	41.39
	Random Forest	68.88	71.94	30.80	41.41	38.43
AUC	Neural Network	0.94	0.92	0.88	0.83	0.75
	Random Forest	0.95	0.93	0.86	0.83	0.72
APRC	Neural Network	0.67	0.73	0.21	0.36	0.40
	Random Forest	0.70	0.75	0.22	0.36	0.37

Tabelle 1 zeigt Indikatoren zur Modellvorhersage auf individueller Ebene für die Merkmale Armutsgefährdung (povmd60), Armuts- oder Ausgrenzungsgefährdung (arose), erhebliche materielle Deprivation (deprived4), subjektiver Gesundheitszustand (health3) und lebenslanges Lernen (si_lll) (zur Definition dieser Merkmale bzw. Indikatoren siehe Till et al. (2020)). Für die Berechnung dieser Werte wurden ebenso die Stichprobengewichte berücksichtigt. Dargestellt sind Durchschnittswerte aus wiederholt durchgeführten Modellierungen. In den meisten Fällen ermöglichte das Random Forest Modell auf Personenebene scheinbar geringfügig genauere Schätzungen. Aufgrund eines möglichen 'overfitting' kann die Auswahl der geeigneten Methode nicht ausschließlich auf die Vorhersage auf individueller Ebene beruhen. Vergleicht man aggregierte Ergebnisse schneidet das neuronale Netz wesentlich besser ab. Die Kreuzvalidierung mit Punktschätzern aus den Erhebungsdaten zu ausgewählten soziodemographischen Gruppenzeigt, dass Neuronale Netz die aus den Erhebungsdaten bereits veröffentlichten Ergebnisse besser reproduzieren können als die auf dem Random Forest Algorithmus aufbauenden Schätzer (2). Die auf Neuralen Netzen basierenden Schätzer liegen meistens innerhalb der 95% Konfidenzintervalle der Punktschätzer aus den Stichprobendaten. Bei diesem Vergleich wurden die Ergebnisse für die Indikatoren die sich auf Variablen aus EU-SILC beziehen auf die Jahre 2016 bis 2018 und dem Indikator lebenslanges Lernen auf 2017 eingeschränkt. In Abbildung 3 sind die zugehörigen Streudiagramme abgebildet.

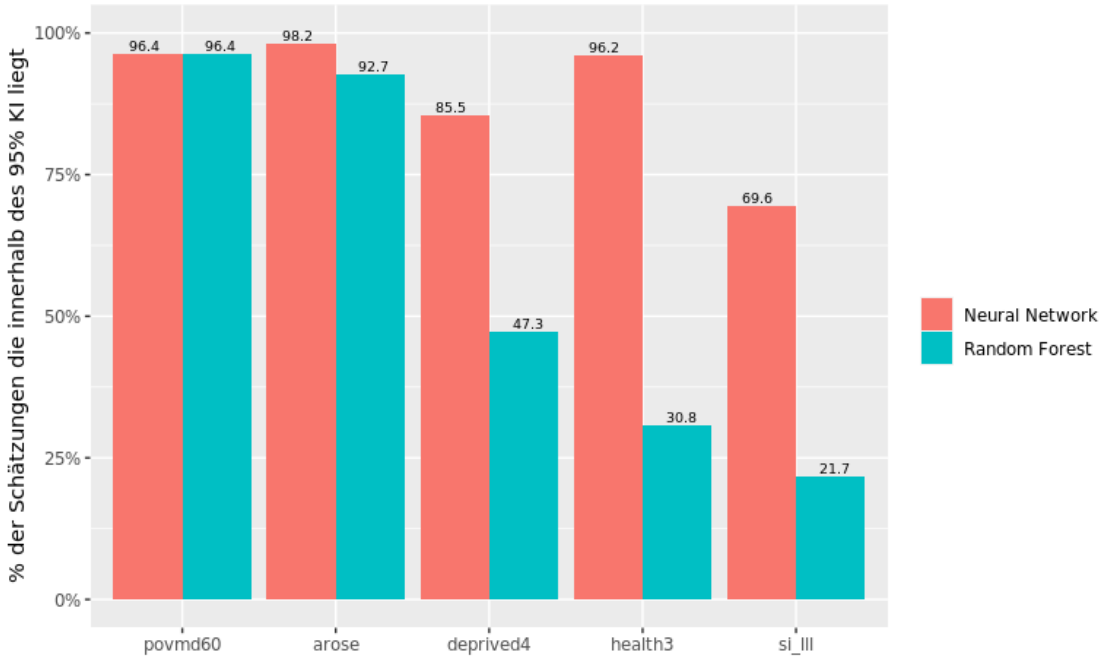

Abbildung 2: Anteil der aus den Modellvorhersagen auf Personenebene für ausgewählte soziodemographische Gruppen abgeleiteten Schätzer, die innerhalb des 95% KI der Punktschätzer aus den Stichprobendaten liegen.

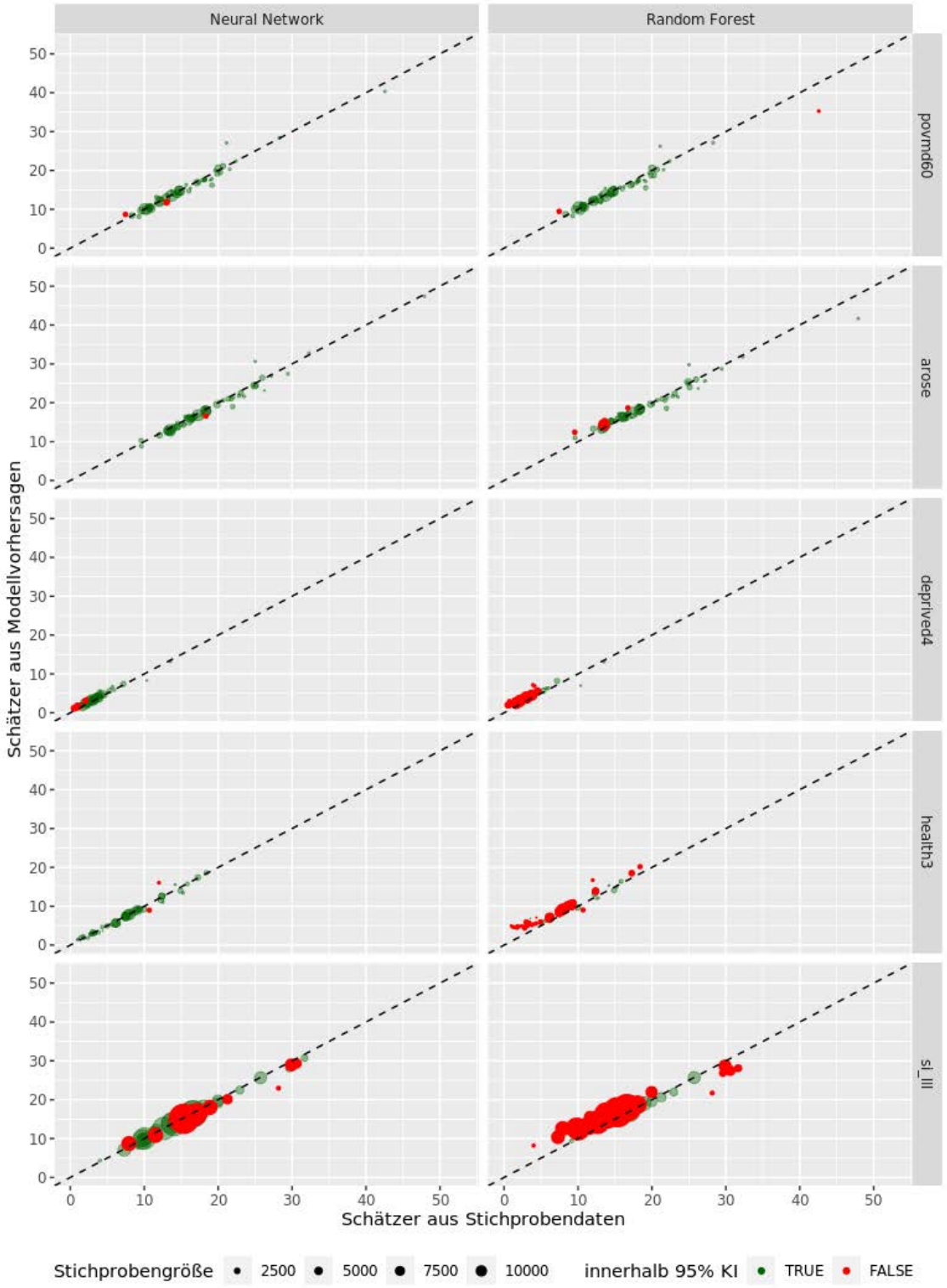

Abbildung 3: Streudiagramm für Punktschätzer abgeleitet aus den Modellvorhersagen (y-Achse) sowie den Stichprobendaten (x-Achse). Die Größe der Punkte spiegelt die dahinter liegende Stichprobengröße. Für Punkte in grün liegt der Schätzer aus den Modellvorhersagen innerhalb des 95% Konfidenzintervall des Punktschätzers aus den Stichprobendaten.

Der für das Neuronale Netz etwas schlechteren Trefferquote auf individueller Ebene steht also ein deutlich kohärenteres Ergebnis mit publizierten Stichprobenergebnissen gegenüber. Letzteres hat einen besonderen Stellenwert für die Glaubwürdigkeit dieser neuartigen experimentellen Schätzmethoden. Somit ist letztlich das Neuronale Netze für diese Schätzungen zu bevorzugen.

Anwendung auf Stichprobenrahmen

Die Grundgesamtheit wurde in diesem Projekt durch den bei Statistik Austria angelegten Stichprobenrahmen für Haushalts- und Personenbefragungen abgebildet. Um das Modell auf die Grundgesamtheit, den Stichprobenrahmen anzuwenden wurde das Neuronale Netz an den Stichprobendaten unter Berücksichtigung der Stichprobengewichte trainiert. Dabei dienten 10% zufällig ausgewählte Beobachtungen als Validierungsdatensatz. Anschließend wurde mit dem Modell für jedes Individuum durch das Modell eine prognostizierte Wahrscheinlichkeit berechnet. Dies wurde 10 mal wiederholt sodass für jede Person 10 prognostizierte Wahrscheinlichkeiten vorliegen. Die endgültige prognostizierte Wahrscheinlichkeit $\tilde{\theta}_i$ für jedes Individuum i und jeden Indikator θ ergibt sich aus dem Mittelwert dieser 10 Schätzungen. Zusätzlich dazu wurde das Modell für jeden Indikator noch weiter 100 mal trainiert, wobei dazu 100 verschieden Bootstrapgewichte verwendet wurden. Die Bootstrapgewichte wurden nach der Methodik aus Preston (2009) welche im R-Paket surveysd (Gussenbauer, Kowarik, und de Cillia (2020)), implementiert ist, erzeugt. Mit Hilfe der 100 Schätzungen $(\hat{\theta}_{i,1}, \dots, \hat{\theta}_{i,100})$ wurde für jedes Individuum i und jeden Indikator θ ein Standardfehler $\sigma_{i,\theta}$ für die Modellschätzungen bestimmt

$$\sigma_{i,\theta} = \sqrt{\frac{1}{100-1}\sum_{j=1}^{100}(\overline{\theta}_i - \hat{\theta}_{i,j})^2}$$

$$\overline{\theta}_i = \frac{1}{100}\sum_{j=1}^{100}\hat{\theta}_{i,j} \quad .$$

Der Standardfehler $\sigma_{i,\theta}$ wird für die Simulation von Konfidenzintervallen verwendet.

Vor Berechnung der Schätzer für die regionalen Einheiten, werden die Wahrscheinlichkeiten $\tilde{\theta}_i$ regional geglättet und auf die Ergebnisse der Stichprobendaten auf NUTS0 sowie NUTS2 kalibriert. Die Glättung verfolgt den Zweck mögliche extreme regionale Werte zu unterdrücken. Die Kalibrierung dient dazu Konsistenz mit den Stichprobenergebnissen herzustellen (für eine detaillierte Beschreibung siehe Gussenbauer et al. (2020), Till et al. (2020)).

Schätzer für regionale Einheiten bestimmen

Um den Schätzer $\tilde{\theta}_i^r$ für jede regionale Einheit r zu berechnen wird der Mittelwert über die prognostizierten Wahrscheinlichkeiten $\tilde{\theta}_i$ aller Individuen in der regionalen Einheiten r ermittelt

$$\tilde{\theta}^r = \frac{1}{N_r} \sum_{i \in r} \tilde{\theta}_i \quad ,$$

mit N_r als der Anzahl aller Individuen in r für die eine prognostizierte Wahrscheinlichkeit berechnet werden konnte.

Ein Konfidenzintervall für diesen Schätzer wird durch Simulation geschätzt. Dafür wird für jedes Individuum i innerhalb der regionalen Einheit r eine Zufallszahl $\hat{\tilde{\theta}}_i$ aus einer gestutzten Normalverteilung im Intervall (0,1) mit den Parametern $\tilde{\theta}_i$ und $\sigma_{i,\theta}$ erzeugt $\tilde{\theta}_i$. Mit diesen simulierten Werten kann wiederum ein Schätzer für regionalen Einheit berechnet werden

$$\hat{\tilde{\theta}}^r = \frac{1}{N_r} \sum_{i=1}^{n} I_{[\hat{\tilde{\theta}}_i > u_i]} \quad ,$$

mit $I_{[.]}$ als der Indikatorfunktion und u_i einer Zufallszahl generiert aus einer Gleichverteilung im Intervall [0,1]. Obige Simulation wird 1000 mal wiederholt um somit die Schätzer $\hat{\tilde{\theta}}_1^r, \dots, \hat{\tilde{\theta}}_{1000}^r$ zu erhalten. Das 95% Konfidenzintervall des Schätzers $\overline{\theta}_s$ wird dann über

$$(\hat{\tilde{\theta}}_{(\alpha/2)}^r, \hat{\tilde{\theta}}_{(1-\alpha/2)}^r)$$

geschätzt mit $\hat{\tilde{\theta}}_{(\alpha/2)}^r$ als das $\alpha/2$ Perzentil der Schätzer $\hat{\tilde{\theta}}_1^r, \dots, \hat{\tilde{\theta}}_{1000}^r$ und $\alpha = 0.05$.

Bei der Veröffentlichung werden Ergebnisse für alle regionalen Einheiten, für die ein Indikator für weniger als 50 Personen zur Verfügung steht oder das Konfidenzintervall des Schätzers der regionalen Einheiten mehr als 10 Prozentpunkte (bzw. 5 Prozentpunkte) umspannt, unterdrückt (Gussenbauer et al. (2020), Till et al. (2020)).

Vergleich mit Punktschätzern

Analog zur Vorgehensweise bei der Kreuzvalidierung können Schätzungen für ausgewählte soziodemographischen Gruppen mit den dazugehörigen Punktschätzern verglichen werden. Abbildung 4 zeigt pro Modell und Indikator den Anteil an Schätzern die innerhalb der 95% Konfidenzintervalle der Punktschätzer aus den Stichprobendaten liegen. In Abbildung 5 sind die zugehörigen Streudiagramme abgebildet. Es zeigt sich, dass die Schätzer schlechter reproduziert werden können als bei der Kreuzvalidierung. Das Neuronale Netz ermöglicht jedoch eine leichte Verbesserung der Schätzungen in Bezug auf die Punktschätzer.

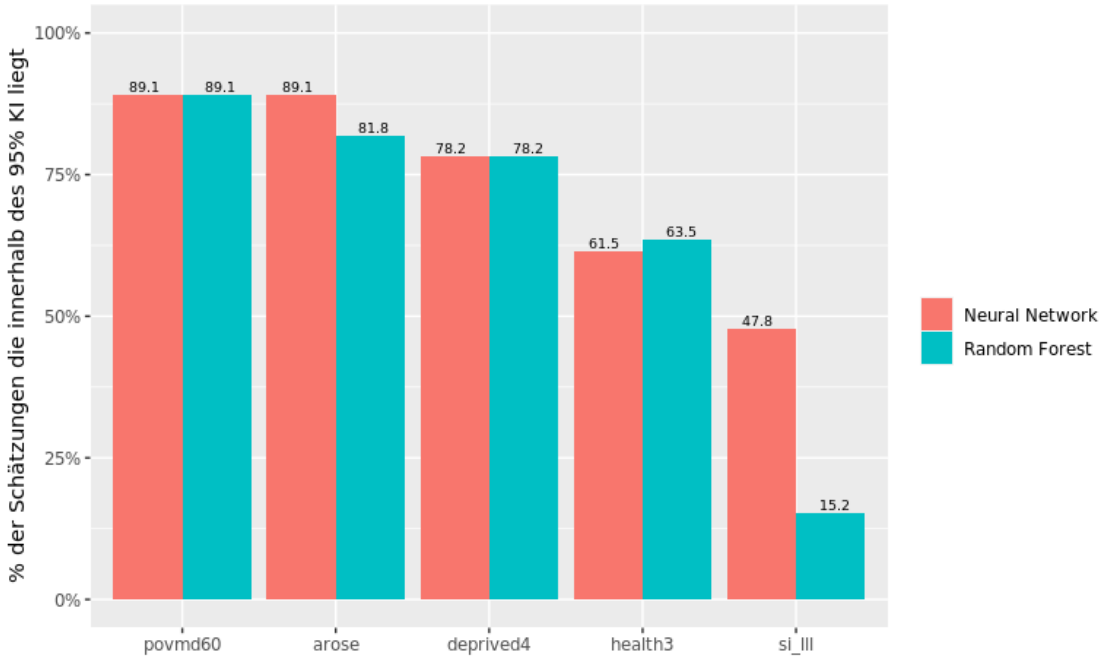

Abbildung 4: Anteil der aus den Modellvorhersagen auf Personenebene (Grundgesamtheit) für ausgewählte soziodemographische Gruppen abgeleiteten Schätzer, die innerhalb des 95% KI der Punktschätzer aus den Stichprobendaten liegen.

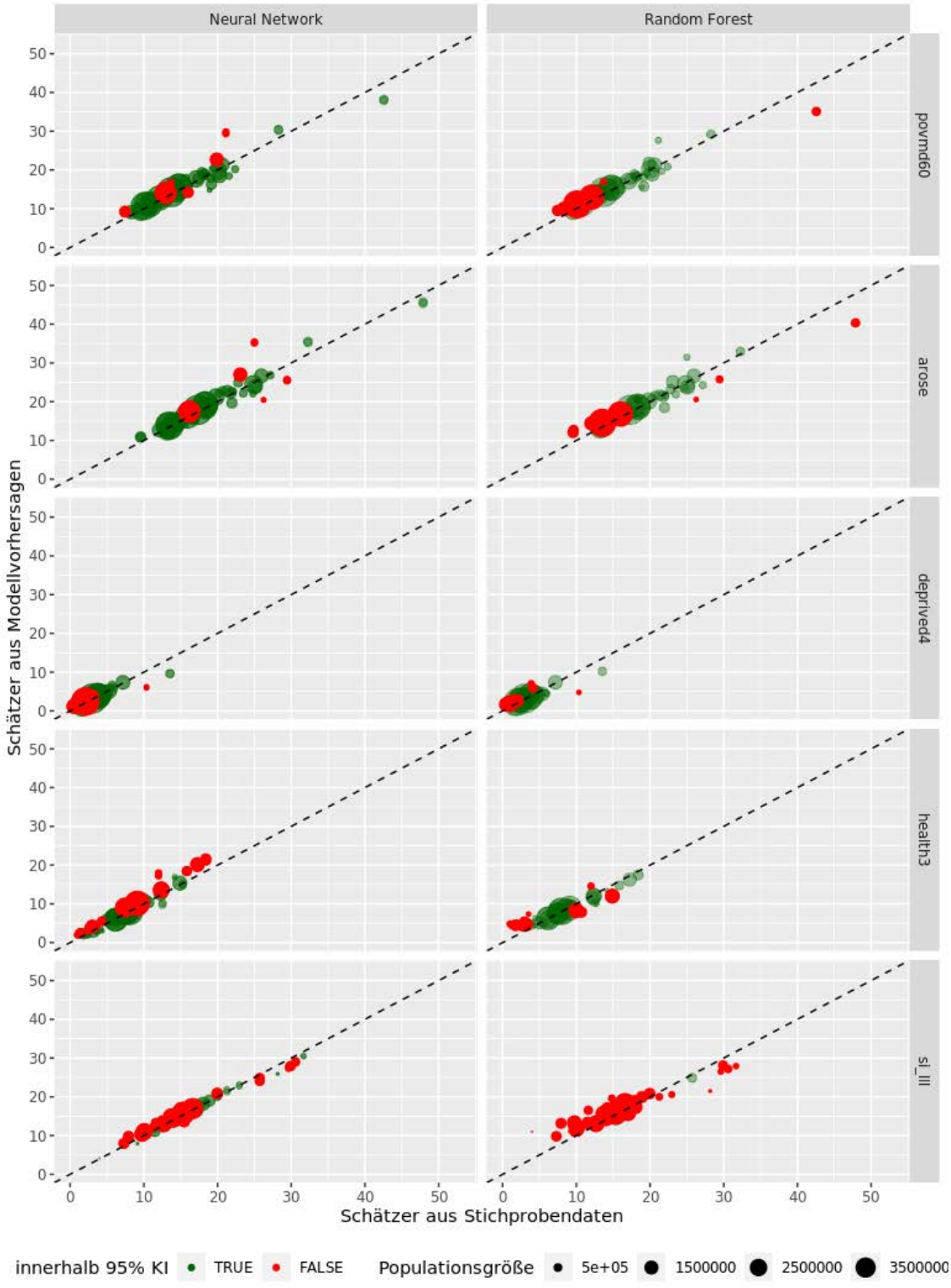

Abbildung 5: Streudiagramm für Punktschätzer abgeleitet aus den Modellvorhersagen für die Grundgesamtheit (y-Achse) sowie den Stichprobendaten (x-Achse). Die Größe der Punkte spiegelt die dahinter liegende Stichprobengröße. Für Punkte in grün liegt der Schätzer aus den Modellvorhersagen innerhalb des 95% Konfidenzintervall des Punktschätzers aus den Stichprobendaten.

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, und François Chollet. 2019. *Keras: R Interface to 'Keras'*. https://CRAN.R-project.org/package=keras.

Gussenbauer, Johannes, Ingrid Kaminger, Matthias Till und Alexandra Wegscheider-Pichler. 2020. "Kleinräumige Darstellung Durch Experimentelle Methoden - Machine Learning for Sample Data and Geographic Information Systems." *Statistische Nachrichten*, September, 857–71.

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

Preston, J. 2009. "Rescaled Bootstrap for Stratified Multistage Sampling." *Survey Methodology* 35 (December): 227–34. https://www.researchgate.net/publication/281735659_Rescaled_bootstrap_for_stratified_multistage_sampling.

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

Till, Matthias, Franz Bilek, Christine Bienzle, Thomas Glaser, Johannes Gussenbauer, Nina Hofer, Kaminger Ingrid, Alexander Kowarik, Sibylle Saul und Alexandra Wegscheider-Pichler. 2020. "LEARN4SDGis - Machine Learning for Sample Data Geographic Information Systems." Eurostat Grant Agreement Number: 08143.2017.001-2017.403. Eurostat / Statistics Austria.

Wright, Marvin, und Andreas Ziegler. 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.