Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

MAS Imputation 2021

Aktualisierung der Analyse für die Daten der MAS 2021

zuhanden der FMH und Ärztekasse

Olten, 30. Juli 2025

Impressum

Bibliographische Angaben

Titel: MAS Imputation 2021. Aktualisierung der Analyse für die Daten der MAS 2021
Autoren: Schoch, T., Hulliger, B. und Müller, R.
Auftraggeber: FMH und Ärztekasse
Ort: Olten
Datum: 30. Juli 2025

Projektteam

Tobias Schoch
Beat Hulliger
Roman Müller

Kontakt

Prof. Dr. Tobias Schoch
Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft, Institut ICC

Riggenbachstrasse 16
CH-4600 Olten, Schweiz

E-Mail: tobias.schoch@fhnw.ch
Tel. (direkt): +41 (0)62 957 21 02

Anmerkung

Der Bericht gibt die Auffassung des Projektteams wieder, die nicht notwendigerweise mit derjenigen des Auftraggebers bzw. der Auftraggeberin oder der Begleitorgane übereinstimmen muss. Für den Inhalt ist allein der Auftragnehmer / die Auftragnehmerin verantwortlich.

Bildrechte

Titelbild: Freepik.com (this cover has been designed using assets from Freepik.com)

Inhaltsverzeichnis

Abkürzungsverzeichnis

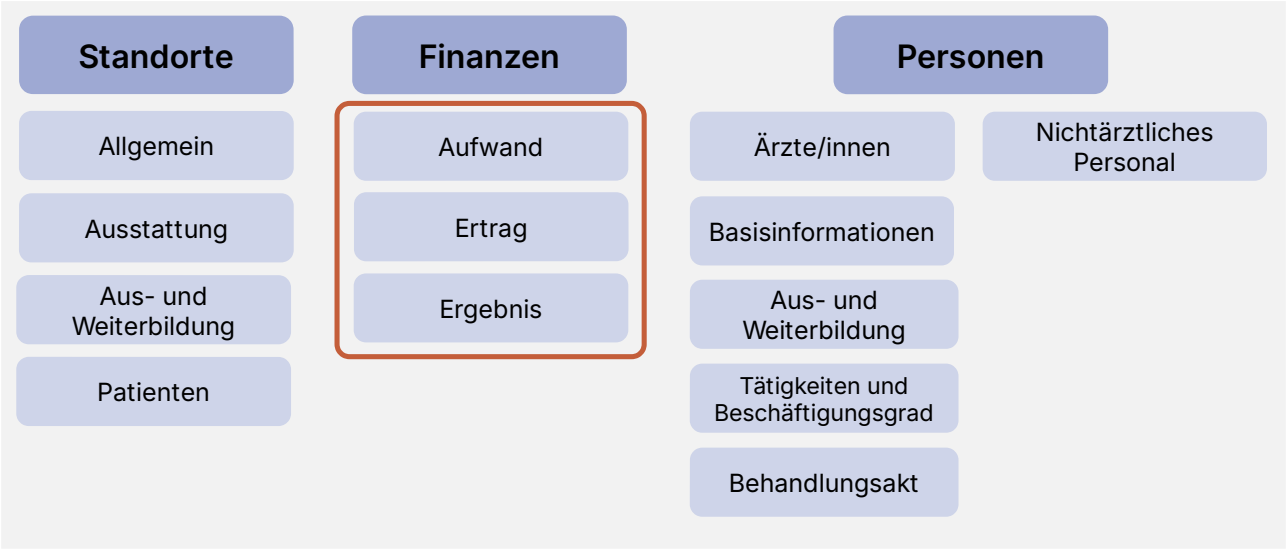
AG	Aktiengesellschaft
BFS	Bundesamt für Statistik
GmbH	Gesellschaft mit beschränkter Haftung
kNN	k-Nearest Neighbors Methode
KVG	Bundesgesetz über die Krankenversicherung
MAE	Mean absolute error (mittlerer absoluter Fehler)
MAS	Strukturdaten der Arztpraxen und ambulanten Zentren
OECD	Organisation for Economic Cooperation and Development
OKP	Obligatorische Krankenpflegeversicherung
QQ-Plot	Quantil-Quantil-Plot
RFPE	Robust final prediction error
UNCE	United Nations Economic Commission for Europe

1 Einleitung

Die Erhebung zu den Strukturdaten der Arztpraxen und ambulanten Zentren (MAS, Medical Ambulatory Structure) ist eine Querschnitterhebung des Bundesamts für Statistik (BFS), die seit 2017 jährlich durchgeführt wird (vgl. BFS, 2018a). Es handelt sich um eine Unternehmenserhebung, weshalb die Teilnahme der Erhebungssubjekte (Unternehmen:[1] Arztpraxen und ambulanten Zentren) obligatorisch ist.[2]

Die MAS bildet eine wesentliche Datengrundlage für die ambulante medizinische Versorgung in der Schweiz. Der Standardfragebogen gliedert sich in drei Teile: Standorte, Finanzen und Personen; siehe Abbildung 1.

Abbildung 1: Hauptaspekte der MAS-Erhebung



Quelle: Zusammenstellung aus: Bundesamt für Statistik, Strukturdaten Arztpraxen: Variablenliste MAS 2021 (BFS, 2023).

Für ein Einzelunternehmen[3], die nicht Teil einer Gruppenpraxis ist, werden die Finanzkennzahlen der Erfolgsrechnung entnommen. Tabelle 1 listet die Kennzahlen zur Erfolgsrechnung von Einzelunternehmen auf.

[1] Bei den Unternehmen werden folgende Arten unterschieden: 1) Einzelunternehmen (nicht Teil einer Gruppenpraxis), 2) Einzelunternehmen, das Teil einer Gruppenpraxis ist, 3) Unternehmen mit der Rechtspersönlichkeit einer Aktiengesellschaft (AG) oder Gesellschaft mit beschränkter Haftung (GmbH); siehe BFS (2018b, S. 2).

[2] Vgl. Verordnung über die Durchführung von statistischen Erhebungen des Bundes, SR 431.012.1.

[3] Das Einzelunternehmen ist die häufigste Rechtsform in der ambulanten ärztlichen Versorgung. Im Jahr 2021 hatte sie einen Anteil von über 80 %.

Tabelle 1: In der MAS erfasste Kennzahlen/ Merkmale zur Erfolgsrechnung der Einzelunternehmen

xet_tot*	**Aufwand**	**Ertrag**	xit_tot*
	Personalaufwand	Ertrag aus Leistungen	
xet_per_sdoc	Lohnaufwand angestellte Ärzte (ohne Praxisinhaber/in)	Ertrag aus medizinischen Leistungen von Ärzten	xit_pra_doc
xet_per_shpad	Lohnaufwand nichtärztliche Gesundheitsfachkräfte	Ertrag aus Leistungen von nichtärztl. Personal	xit_pra_ndoc
xet_per_soth	Lohnaufwand übriges Personal	Ertrag aus Laboranalysen	xit_pra_lab
xet_per_soc	Sozialleistungen und berufl. Vorsorge (nur Personal)	TOTAL Ertrag aus Leistungen	xit_pra_prest
xet_per_ext	Einkauf ärztlicher Leistungen	Ertrag aus Medikamenten	xit_pra_drug
xet_per_other	Übriger Personalaufwand	Ertrag aus Mittel und Gegenständen/ Material	xit_pra_migel
xet_per_total	TOTAL Personalaufwand	TOTAL Ertrag aus Medikamenten und Gegenständen	xit_pra_mtot
	Sachaufwand für medizinische Tätigkeiten	Übriger Ertrag aus Praxistätigkeit	xit_pra_other
xet_mat_drug	Medikamente	TOTAL Ertrag aus Praxistätigkeit	xit_pra_total
xet_mat_lab	Labormaterial	Ertrag durch Ärzte ausserhalb der Praxis	
xet_mat_migel	Mittel und Gegenstände	Ertrag aus Spitaltätigkeit	xit_npr_hosp
xet_mat_other	Übriger Materialaufwand	Ertrag aus übrigen ärztl. Dienstleistungen	xit_npr_other
xet_mat_total	TOTAL Sachaufwand für medizinische Tätigkeiten	TOTAL Ertrag durch Ärzte ausserhalb der Praxis	xit_npr_total
	Sozialleistungen und Vorsorge Praxisinhaber/in	Übriger betrieblicher Ertrag	
xet_ins_total	TOTAL Sozialleistungen und Vorsorge Praxisinhaber/in	Miet- und Kapitalerträge	xit_oth_occap
	Übriger betrieblicher Aufwand	Entschädigung für Verdienstausfall (ohne Praxisinhaber)	xit_oth_coloe
xet_oth_occup	Raum- und Mietaufwand	Sonstiger Ertrag	xit_oth_other
xet_oth_itadm	IT- und Verwaltungsaufwand	TOTAL übriger betrieblicher Ertrag	xit_oth_total
xet_oth_vehic	Fahrzeugaufwand		
xed_oth_capit	Zinsaufwand (Kapitalaufwand)		
xet_oth_depre	Abschreibungen		
xet_ins_pra	Praxisversicherungen		
xet_oth_other	Sonstiger Aufwand		
xet_oth_total	TOTAL übriger betrieblicher Aufwand		
xrt_tot*	Betriebsergebnis		

Quelle: Zusammenstellung aus: Bundesamt für Statistik, Strukturdaten Arztpraxen: Variablenliste MAS 2019/2021, Standardfragebogen, Abschnitt D.2.
*Legende: * berechnete Variable; obligatorische Variablen sind fett gedruckt.*

Tabelle 2: Betroffene Variablen

Variablennamen	Bezeichnung
xit_pra	Erträge
xit_pra_drug	Ertrag aus Medikamenten: Total Ertrag
xit_pra_lab	Ertrag aus Laboranalysen: Total Ertrag
xit_pra_migel	Ertrag aus Mittel und Gegenstände: Total Ertrag
xit_pra_other	Übriger Ertrag aus Praxistätigkeit: Total Ertrag
xit_pra_ndoc	Ertrag aus Leistungen von nichtärztlichem Personal: Total Ertrag
xit_pra_doc	Ertrag aus medizinischen Leistungen von Ärzten: Total Ertrag
xet_mat	Materialaufwände
xet_mat_drug	Aufwand Medikamente
xet_mat_lab	Aufwand Labormaterial
xet_mat_migel	Aufwand Mittel und Gegenstände
xet_mat_other	Aufwand übriger Materialaufwand
xet_per	Personalaufwände
xet_per_sdoc	Rechtsform 1 (Einzelunternehmen): Lohnaufwand angestellte Ärzte (ohne Inhaber der Arztpraxis) Rechtsform 2-34: Lohnaufwand Ärzte
xet_per_shpad	Lohnaufwand nichtärztliche Gesundheitsfachkräfte und Personal für Administration
xet_per_soth	Lohnaufwand übriges Personal
xet_per_soc	Rechtsform 1 (Einzelunternehmen): Sozialleistungen und berufliche Vorsorge (angestelltes Personal) Rechtsform 2-34: Sozialleistungen und berufliche Vorsorge
xet_per_ext	Einkauf ärztlicher Leistungen
xet_per_oth	Übriger Personalaufwand
xet_oth	Sonstige Aufwände
xet_oth_occup	Raum- und Mietaufwand
xet_oth_itadm	IT- und Verwaltungsaufwand (z.B. Telekommunikation, Software, Büromaterial)
xet_oth_vehic	Fahrzeugaufwand
xet_oth_capit	Kapitalaufwand
xet_oth_depre	Abschreibungen
xet_oth_other	Übriger Aufwand
xet_oth_inspra	Praxisversicherungen

Quelle: Eigene Darstellung.

Das Betriebsergebnis eines Einzelunternehmens ist in Tabelle 1 als Saldo zum Ausgleich der Erfolgsrechnung auf der Auftragsseite eingetragen — unter der Annahme, dass der Saldo positiv ist; siehe Leimgruber und Prochinig (2023, S. 28 ff.). Bei Einzelunternehmen ent-

spricht das Betriebsergebnis dem «Einkommen» der selbständig praktizierenden Ärztin bzw. des selbständig praktizierenden Arztes.

Bei Praxen mit einer anderen Rechtsform als der des Einzelunternehmens (z. B. Aktiengesellschaft) enthält die Erfolgsrechnung zusätzliche bzw. andere Positionen. Diese werden ebenfalls mit Fragebogen der MAS erfasst; auf eine Darstellung der Einzelheiten wird an dieser Stelle verzichtet; siehe BFS (2023). Überdies wird – für jede Position der Erfolgsrechnung – zwischen KVG-Erträgen und sonstigen Erträgen unterschieden.[4]

1.1 Gegenstand der Studie

Die Daten der MAS 2021 weisen fehlende Werte und inkonsistente **Finanzdaten** auf, welche die Aussagekraft der Daten einschränkt. Der vorliegende Bericht beschreibt die Methodik und die statistischen Verfahren zur Einsetzung von fehlenden Werten und zur Verbesserung der Datenqualität. Damit werden die folgenden **Ziele** verfolgt:

- fehlende Werte zu untersuchen,
- Widersprüche und Fehler in den Daten aufzudecken,
- fehlende oder widersprüchliche Daten durch plausible Werte zu ersetzen, um einen möglichst vollständigen und konsistenten Datensatz zu erhalten.

Die in diesem Bericht besprochenen Datenarbeiten beziehen sich nur die in Tabelle 2 aufgeführten Variablen. Dabei kann auf die Vorarbeiten von Hulliger und Bisang (2021) zurückgegriffen werden. Die in diesem Bericht vorgeschlagenen Methoden, Einsetzungen und Anpassungen unterscheiden sich jedoch (teilweise grundlegend) von denen in Hulliger und Bisang (2021).

Die Ausführungen in diesem Bericht sind bewusst **beispielhaft** gehalten. Auf die Darstellung aller formalen und technischen Details wird verzichtet. Gleichwohl sind die Ausführungen so detailliert beschrieben, dass sie von Dritten nachvollzogen und reproduziert werden können.

1.2 Prinzipien

Bei den Einsetzungen und Anpassungen sind die folgenden Grundsätze und Prinzipien handlungsleitend:

- **Anpassungen** der Originaldaten sind auf **ein Minimum** zu beschränken. Eine Anpassung der Originaldaten ist nur dann vorgesehen, wenn diese als unplausibel erachtet werden.

[4] Als KVG-Erträge werden die Erträge bezeichnet, die für Leistungen der obligatorischen Krankenpflegeversicherung (OKP) entrichtet werden.

- Die Imputation von fehlenden Werten soll sich auf die **vorhandenen Daten** stützen oder auf Modelle, die aus diesen Daten geschätzt wurden.
- Die Totale der Finanzdaten werden nach einer anfänglichen Plausibilitätsprüfung und Ausreisserentdeckung als unveränderlich angenommen. Diese Priorisierung erleichtert die Lokalisierung von Fehlern und Inkonsistenzen in den Daten.
- Der Schwerpunkt der Arbeiten liegt auf ausgewählten Finanzdaten, insbesondere auf den für Tarifierungsfragen relevanten Variablen. Es wird nicht der Anspruch erhoben, fehlende Werte für alle Variablen zu imputieren.

1.3 Herausforderungen

Die (meisten) Finanzdaten können als gemischte Zufallsvariablen aufgefasst werden, die reelle Zahlen als Werte annehmen. Sei X eine solche Variable. Die Verteilungsfunktion F von X besitzt eine Zerlegung der Gestalt $F = pF_d + (1-p)F_{as}$, $0 \leq p \leq 1$, wobei F_d eine diskrete und F_{as} eine absolut stetige Verteilungsfunktion ist.[5] Für $0 < p < 1$ ist F weder diskret noch absolut stetig, sondern gemischt. Das hat Implikationen z. B. für die:

- Berechnung von Distanz- oder Divergenzmassen (z. B. Hellinger-Distanz) zwischen zwei Verteilungen F und G;
- Transformation der Variablen und die Wahl der Methoden; z. B. Methoden zu compositional data analysis (siehe bspw. Templ, 2023, Kapitel 10) können nicht verwendet werden;
- Ausreisserentdeckung (Unterscheidung von Nullwerten vs. positive Werte bei den Finanzkennzahlen).

1.4 Weitere Anmerkungen

Alle Variablennamen der MAS werden mit Kleinbuchstaben geschrieben; z. B. `entid_aleat` anstelle von `ENTID_ALEAT`.

Die MAS ist eine Vollerhebung (Zensuserhebung), allerdings haben nicht alle Unternehmen an den jährlichen Erhebungen teilgenommen. Das BFS hat deshalb eine Gewichtung der Einzeldaten vorgenommen, um für allfällige Verzerrungen infolge von Teilnahmeausfällen zu korrigieren (BFS, 2018a).[6] Alle im vorliegenden Bericht ausgewiesenen Schätzungen berücksichtigen die Gewichtung des BFS.

[5] Nach dem Lebesgueschen Zerlegungssatz ist die Zerlegung eindeutig (und beinhaltet noch eine stetige singuläre Komponente, die jedoch für unsere Anwendung keine Rolle spielt); siehe bspw. Loeve (2017, S. 178).

[6] Siehe auch Särndal und Lundström (2005, Kapitel 2).

Alle Berechnungen wurden mit der R-Statistiksoftware (Language and Environment for Statistical Computing, siehe R Development Core Team, 2025) durchgeführt.

1.5 Aufbau des Berichts

Kapitel 2 gibt eine Übersicht zur Anzahl der fehlenden Werte. In Kapitel 3 wird die Grundgesamtheit der MAS definiert und erläutert, welche Praxen nicht zur Grundgesamtheit gehören und daher ausgeschlossen werden (z. B. Spitalamublatorien). Im Anschluss daran werden folgende Aspekte besprochen:

- Ausreisserentdeckung und -entfernung (Kapitel 4),
- deduktive Eingriffe (Kapitel 5),
- Dateneinsetzung/ Imputation (Kapitel 6) und
- Ausgleichseingriffe (Kapitel 7).

In den Kapiteln 8 und 9 werden die Einsetzungen evaluiert und Empfehlungen ausgesprochen.

2 Übersicht zu den fehlenden Werten

Die Finanzdaten der MAS 2021 sind in unterschiedlichem Ausmass von fehlenden Werten betroffen; siehe Tabelle 3. In dieser Darstellung wird zwischen bekannten und fehlenden Subtotalen unterschieden. Dabei wird zusätzlich zwischen vollständig und teilweise fehlenden Werten differenziert.

Lässt man die Position der Versicherungsaufwände ausser Acht, so ergibt sich folgendes Bild: Zwischen 41 % und 61 % der Subtotale sind bei den Erträgen und Aufwänden bekannt. Bemerkenswert ist, dass gut 30 % der Subtotale vollständig fehlen (Ausnahme: Versicherungsaufwände und andere Aufwände). Als vollständig fehlend werden Ertrags- oder Aufwandspositionen bezeichnet, bei denen alle Subtotale fehlen. Zum Beispiel: Das Total des Sachaufwands für medizinische Tätigkeiten setzt sich aus den Subtotalen Medikamente, Labormaterial, etc. (siehe Tabelle 4) zusammen. Vollständiges Fehlen bedeutet in diesem Beispiel, dass nur das Total `xet_mat_total` bekannt ist, nicht aber die Subtotale. Dies ist darauf zurückzuführen, dass die Angabe der Subtotale im Fragebogen fakultativ ist.

Tabelle 3: Anteil der bekannten und fehlenden Subtotale bei den relevanten Finanzdaten

		Erträge	Material-aufwände	Personal-aufwände	Versicherungs-aufwände	Andere Aufwände
Bekannt		43%	50%	41%	0%	61%
Fehlend	Vollst.	34%	32%	33%	83%	4%
	Teilw.	14%	16%	24%	3%	9%

Quelle: Eigene Darstellung, Daten: MAS 2021.

Tabelle 4: Beispiel: Position Sachaufwand für medizinische Tätigkeiten

Sachaufwand für medizinische Tätigkeiten	
`xet_mat_drug`	Medikamente
`xet_mat_lab`	Labormaterial
`xet_mat_migel`	Mittel und Gegenstände
`xet_mat_other`	Übriger Materialaufwand
`xet_mat_total`	**TOTAL Sachaufwand für medizinische Tätigkeiten**

Quelle: Eigene Darstellung, Auszug aus Tabelle 1.

3 Daten integrieren

3.1 Definition der Grundgesamtheit und Ausschlüsse

Die Grundgesamtheit der MAS 2021 umfasst alle Unternehmen, die als Leistungserbringer in der ambulanten Gesundheitsversorgung im Jahr 2021 in der Schweiz tätig waren. Das BFS unterscheidet dabei zwischen der Hauptpopulation und der peripheren Population (siehe bspw. BFS, 2018a).

- Die **Hauptpopulation** umfasst Unternehmen mit eigener Infrastruktur und einem Jahresumsatz von mehr als CHF 30 000. Diese Unternehmen füllen den Standardfragebogen aus.
- Zur peripheren Population gehören Unternehmen mit einem Jahresumsatz bis CHF 30 000 oder Unternehmen, die einen Jahresumsatz von mehr als CHF 30 000 ausweisen, jedoch über keine eigene Infrastruktur verfügen. Diese Unternehmen müssen nur einen Kurzfragebogen ausfüllen.

Im ersten Schritt erfolgt eine **Eingrenzung der Grundgesamtheit** auf die Hauptpopulation gemäss BFS. Unternehmen, die der peripheren Population zugerechnet werden, werden folglich aus den Daten entfernt. Im anschliessenden Schritt wird die Grundgesamtheit weiter eingegrenzt, sodass sie nur Praxen von selbständig praktizierenden Ärztinnen und Ärzten umfasst (d. h. Ausschluss der Spitalambulatorien).

Die Eingrenzung der Grundgesamtheit, wie sie geschildert wurde, erfolgt unter Zuhilfenahme von **Ausschlusskriterien**. Unternehmen/ Praxen werden von der Grundgesamtheit ausgeschlossen, sofern sie eines oder mehrere der folgenden Ausschlusskriterien erfüllen:

- Praxis mit einem Umsatz $\leq$ CHF 30 000 (periphere Population),
- ambulantes Zentrum/ Spitalambulatorium,
- Praxis ohne eigene Infrastruktur (periphere Population),
- Praxis, die keine ambulanten ärztlichen Leistungen erbringt.

Zusätzlich werden alle Praxen ausgeschlossen, die erst im Laufe des Erhebungsjahres ihre Tätigkeit aufgenommen haben. Ebenfalls ausgeschlossen werden Praxen mit einem unplausiblen Praxisertrag (Variable `xit_pra_total = 9999`). Die Entscheidung, ob ein Wert unplausibel ist oder nicht, wurde vom BFS getroffen.

Alle Ausschlusskriterien sind in Tabelle 5 aufgeführt, einschliesslich der technischen Umsetzung in Bezug auf die MAS-Variablennamen.

Tabelle 5: Ausschlusskriterien

Ausschlusskriterium	Technische Umsetzung
Unternehmen erbringt keine ambulanten Leistungen	`hs_med_amb = 0`
Unternehmen hat keine eigene Infrastruktur	`hs_own_infra = 0`
Unternehmen mit einem Umsatz kleiner als CHF 30 000	`30000 > xit_tot`
Unternehmen rechnet nach Spital-TPW ab	`tpw_hosp > 0`
Unternehmen rechnet nicht nach ambulantem TPW ab	`tpw_pra = 0`
Unternehmen wurde während des Erhebungsjahres neu aktiv	`is_new > 0`
Unternehmen mit irreführenden/ unplausiblen Werten	`xit_pra_total = 9999`

Quelle: Eigene Darstellung.

3.2 Aggregation

Die MAS-Daten werden vom BFS in mehreren Dateien/ Tabellen geliefert, deren Inhalte sich auf unterschiedliche Analyseebenen beziehen. Es können folgende drei Ebenen unterschieden werden:

- Unternehmen/ Praxis
- Niederlassung
- Personal (Ärztin/ Arzt)

Mithilfe von pseudonymisierten Identifikatoren (auf den Stufen Unternehmen, Niederlassungen und Personen) können die Dateien/ Tabellen miteinander verknüpft werden. Wir verwenden die folgenden Tabellen[7]: `f_local_unit`, `f_finance`, `f_degree_fach`, `f_doctor_local_unit_fach`, `f_typology` und `f_weight`.

Nach Ausschluss der Fälle, die nicht zur Grundgesamtheit zählen (siehe Kapitel 3.1), wurden die Dateien/ Tabellen zusammengeführt bzw. aggregiert.[8]

[7] Die Namen aller Tabellen sind mit dem Suffix _[Jahr] versehen. Z. B. `f_finance_2021` meint die Finanzdaten für das Jahr 2021. Um die Notation zu vereinfachen, haben wir auf die Angabe des Suffixes verzichtet.

[8] Die Tatsache, dass alle notwendigen Ausschlüsse – die sich auf unterschiedliche Analyseebenen beziehen – direkt zu Beginn durchgeführt werden, hat den Vorteil, dass sich bei der nachfolgenden Aggregation und Vermengung der unterschiedlichen Analyseebenen keine ungewollten Fälle einschleichen können. Im Anschluss werden die unterschiedlichen Dateien zusammengefügt und so aggregiert, dass die Daten auf Ebene der Unternehmen ausgewertet werden können. Das Aggregationsverfahren wird im Folgenden beschrieben.

Zum Umfang mit Unternehmen, die mehrere Niederlassungen haben

Besitzt ein Unternehmen mehrere Niederlassungen im gleichen Kanton, so werden diese zu einer Gesamtpraxis zusammengefasst.[9] Da die Daten der MAS als Grundlage für diese Verhandlungen dienen, ist es notwendig, dass die MAS-Daten nach Kantonen ausgewertet werden können. Aus diesem Grund wird der Kanton bzw. die Kantonsgruppenzugehörigkeit einzelner Niederlassungen eines Unternehmens bei der Aggregation berücksichtigt. Konkret bedeutet dies, dass ausschliesslich Niederlassungen eines Unternehmens zusammengefasst werden, deren Standorte sich in der gleichen Kantonsgruppe befinden. Zusammenfassen bedeutet in solchen Fällen, dass die Aufwands- und Ertragsposten der Niederlassungen aufsummiert werden.

Nachfolgend wird das Zusammenfassen anhand eines fiktiven Beispiels mit den beiden Unternehmen «A» und «B» und Variablen «Materialaufwand» illustriert.

Abbildung 2: Zusammenfassen von Niederlassungen (vgl. Beispiel 3.1)

Unternehmen	Kantonsgruppe	Materialaufwand
A	BE	210'500
A	BE	50'200
B	ZH	424'600
B	SZ, ZG	163'300
B	AR, AI, SG	145'900

Aggregation →

Unternehmen	Kantonsgruppe	Materialaufwand
A	BE	260'700
B	ZH	424'600
B	SZ, ZG	163'300
B	AR, AI, SG	145'900

Quelle: Eigene Darstellung.

Beispiel 3.1: Das Unternehmen «A» hat zwei Niederlassungen im Kanton Bern, deren Materialaufwände bei der Aggregation aufsummiert werden (210'500 + 50'200 = 260'700). Bei Unternehmen «B» verteilen sich die drei Niederlassungen hingegen auf unterschiedliche Kantonsgruppen. Aus diesem Grund werden die Materialaufwände in diesem Beispiel nicht aufsummiert.

3.3 Resultierender Datensatz

Nach Anwendung der Ausschlusskriterien und der Aggregation von Niederlassungen resultiert ein Datensatz im Umfang von 8'327 Praxen mit 148 Variablen.

[9] Für die nach KVG abgerechneten medizinischen Leistungen wird der Taxpunktwert pro Kanton individuell zwischen den Tarifpartnern ausgehandelt.

4 Ausreisserentdeckung und -entfernung

Für die Imputation haben die Totale der Finanzdaten pro Praxis, deren Beantwortung im Fragebogen als **obligatorisch** erklärt wurde (siehe Tabelle 1), eine zentrale Bedeutung. Es handelt sich um folgende Total/Variablen:

- `xet_mat_total` (Total Sachaufwand),
- `xet_per_total` (Total Personalaufwand),
- `xet_ins_total` (Total Sozialleistungen und Vorsorge der/des Inhaber/in),
- `xet_oth_total` (Total übriger betrieblicher Aufwand),
- `xit_pra_total` (Total Ertrag aus Praxistätigkeit),
- `xit_oth_total` (Total übriger betrieblicher Ertrag),

Angesichts ihrer grossen Bedeutung für das weitere Vorgehen ist es wichtig, **Artefakte und Ausreisser in diesen Variablen zu identifizieren und zu entfernen**.

Die Variable `xit_npr_total` (Total Ertrag durch Ärzt/innen ausserhalb Praxis) wird nicht berücksichtigt, weil sie für die Abrechnung nach OKP nicht relevant ist. Die Variablen `xet_tot` (Gesamtaufwand), `xrt_tot` (Betriebsergebnis) und `xit_tot` (Gesamtertrag) werden auch nicht berücksichtigt, da sie sich als Summen aus (einigen) der oben genannten Variablen ergeben.[10]

Jede der oben aufgeführten sechs Variablen kann entweder null oder grösser null sein (in abgekürzter Notation als «> 0» geschrieben) – also zwei Zustände annehmen (vgl. Kapitel 1.3, gemischte Verteilung). Insgesamt resultieren $2^6 = 64$ verschiedene Muster (Kombinationen von Zuständen).[11] Für einige der Muster existieren keine Daten bzw. die wenigen verfügbaren Beobachtungen sind statistische Artefakte. Ein Beispiel dazu ist Muster 16, bei dem nur die Variable `xit_pra_total` (Total Ertrag aus der Praxistätigkeit) > 0 ist. Praxen mit diesem Muster weisen zwar einen positiven Ertrag aus der Praxistätigkeit auf, jedoch keine

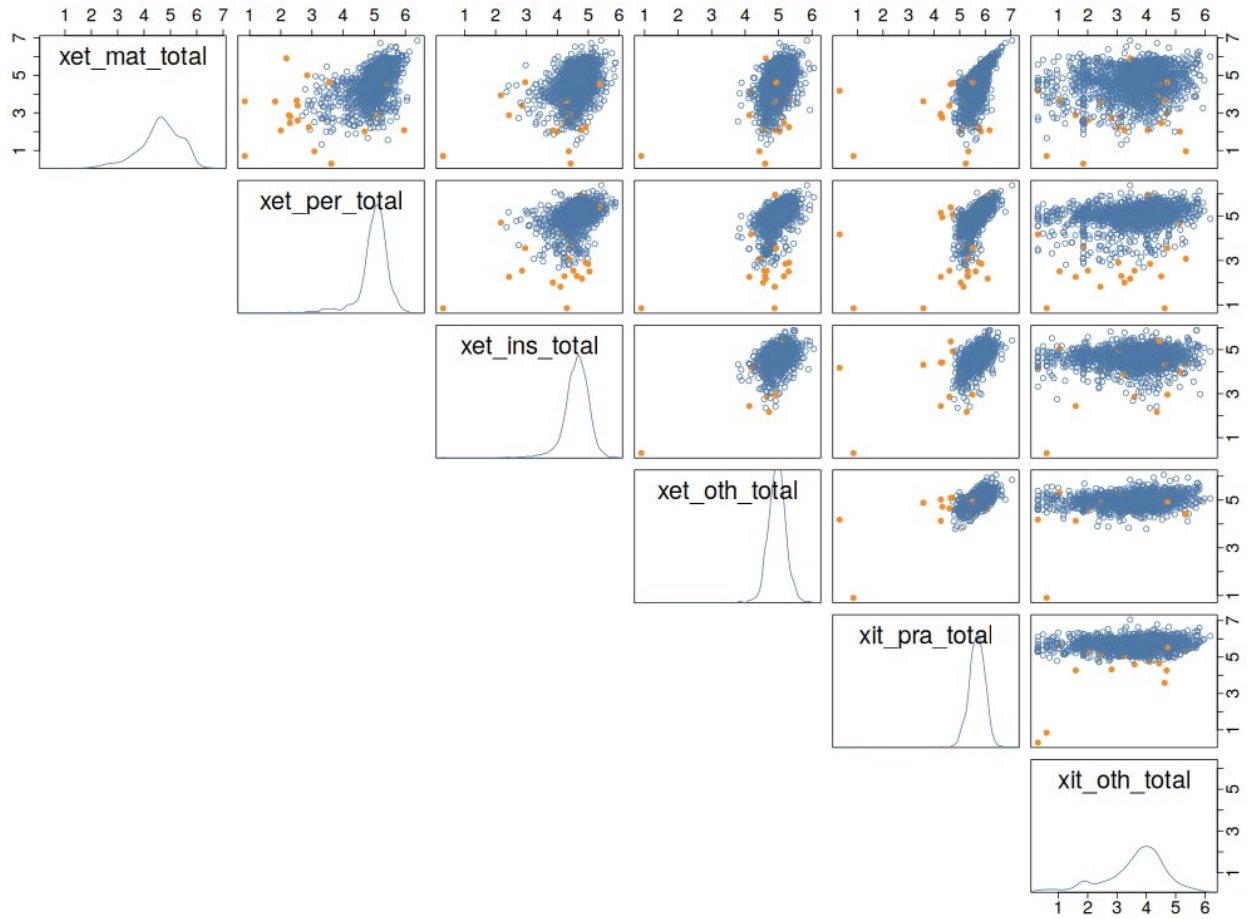
[10] Zum Beispiel: Der Gesamtertrag `xit_tot` setzt sich zusammen aus den Variablen `xit_pra_total`, `xit_npr_total`, `xit_oth_total` und `xrt_tot`. Diese Variablen sind bereits berücksichtigt (bzw. das Betriebsergebnis `xrt_tot` ist zwar nicht enthalten, ergibt sich jedoch als Differenz der Erträge abzüglich der Aufwände), so dass es nicht notwendig (bzw. aus Gründen der Überbestimmung nicht möglich) ist, den Gesamtertrag `xit_tot` separat einzubeziehen.

[11] Die Muster werden mit den Ziffern 1, ..., 63 bezeichnet (Muster 0 wird weggelassen). Diese Zahlen leiten sich aus einer binären Codierung ab: Code 0 bedeutet, dass die Variable null ist; 1 bedeutet, dass die Variable einen Wert > 0 besitzt. Die sechs Variablen, `xet_mat_total`, ..., `xit_oth_total`, sind für die Codierung in der Reihenfolge angeordnet, wie sie in der Aufzählung (siehe oben) aufgeführt sind. Beispiel: 010000 bedeutet, dass nur die fünfte Variable, d. h. `xet_pra_total`, einen Wert > 0 besitzt. Der Code 010000 entspricht dem Muster 16. Die Coderung erlaubt es dem/der Leser/in, die Variablen-Konfiguration der Muster zu identifizieren, z. B. um diese in ihrer/seiner Analyse zu reproduzieren.

Aufwände. Es handelt sich um zehn Praxen, für die das Muster 16 zutrifft. Diese und ähnliche Fälle werden als Artefakte (mit dysfunktionalen Geschäftsmodellen im Kontext der OKP) betrachtet und aus dem Datensatz ausgeschlossen. Insgesamt werden **59 Artefakte** (Praxen) ausgeschlossen.

Die übrigen Muster repräsentieren Praxen, die nicht per se zu Artefakten erklärt werden. Für diese Fälle wird eine Ausreisserentdeckung mit dem BACON-Algorithmus von Billor et al. (2000) für jedes Muster einzeln durchgeführt.[12] Alle Variablen werden zuvor einer Symmetrietransformation unterzogen (in den meisten Fällen logarithmiert), so dass eine (approx.) symmetrisch-konturierte multivariate Verteilung (siehe bspw. Fang et al., 1990, Kapitel 2) resultiert. Wir haben den BACON-Algorithmus konservativ parametrisiert, so dass nur extreme Beobachtungen als potenzielle Ausreisser deklariert werden.

Abbildung 3: Ausreisserentdeckung für das Muster 63



Quelle: Eigene Darstellung, Daten: MAS 2021, alle Beobachtungen sind auf der log-log-Skala dargestellt.
Anm.: Die potenziellen Ausreisser sind orange markiert; BACON mit den Parametern alpha = 0.01, collect = 2.

Beispiel 4.1: Abbildung 3 zeigt die Streudiagrammmatrix zur Ausreisserentdeckung mit dem BACON-Algorithmus für das Muster 63. Bei diesem Muster sind alle sechs Variablen > 0. Die

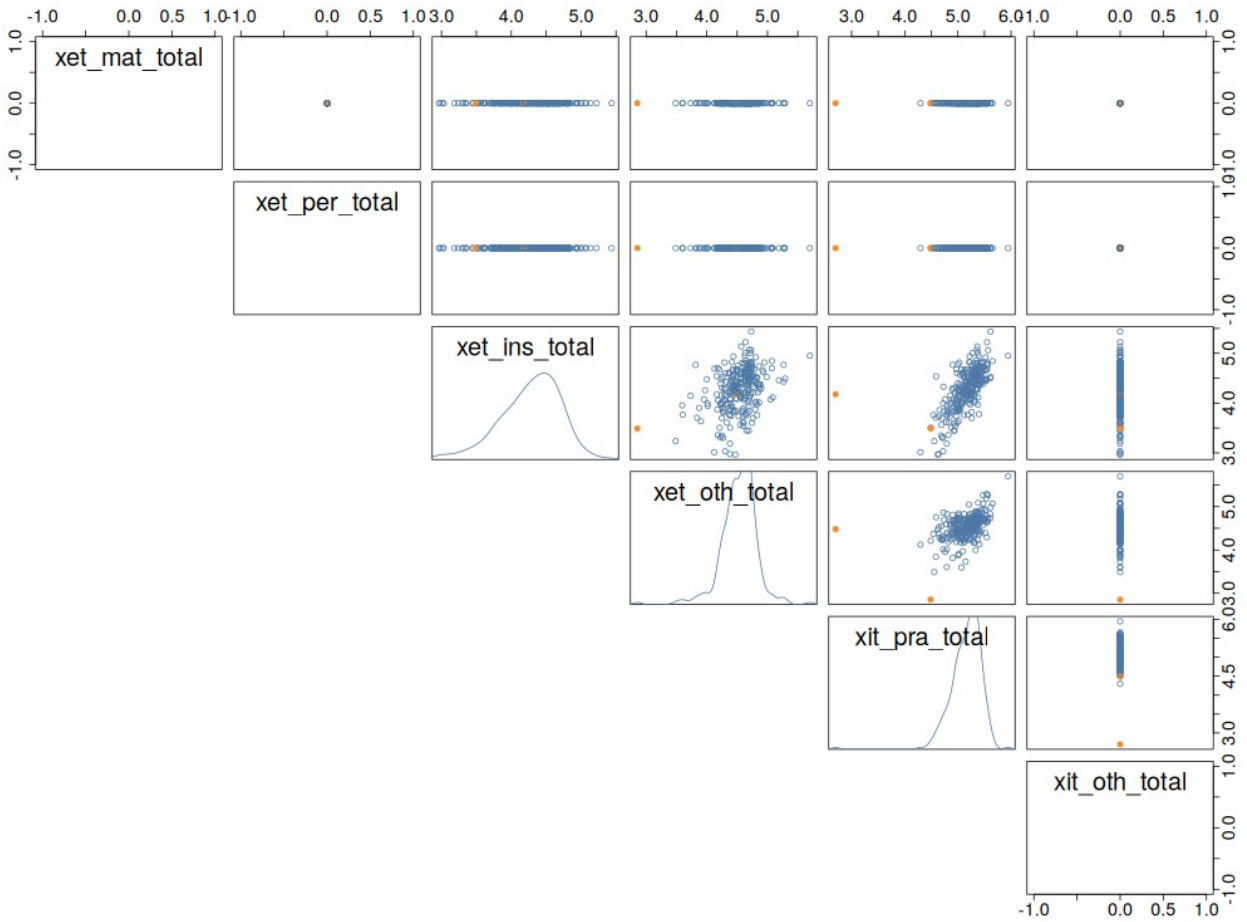
[12] Wir verwenden die Softwareimplementation des BACON-Algorithmus in Schoch (2021).

potenziellen Ausreisser sind orange markiert. Insgesamt wurden 29 Praxen als Ausreisser deklariert.

Beispiel 4.2: Ein weiteres Beispiel ist in Abbildung 4 dargestellt. Es handelt sich um die Streudiagrammmatrix für das Muster 28. Bei diesem Muster sind nur die Variaben xet_ins_total, xet_oth_total und xit_pra_total grösser als null; alle anderen Variablen sind gleich null (siehe Punkte, die sich auf einer vertikalen oder horizontalen Linie aufreihen). Es wurden nur 2 Ausreisser identifiziert.

Mit Hilfe der multivariaten Ausreissererkennung konnten **insgesamt 85 Ausreisser** (Praxen) identifiziert werden. Die Anzahl der Ausreisser pro Muster ist in Tabelle 6 dargestellt. Die identifizierten Ausreisser wurden aus dem Datensatz entfernt. Nach Ausschluss der Ausreisser und Artefakte umfasst der Datensatz 8'183 Praxen.

Abbildung 4: Ausreisserentdeckung für das Muster 28



Quelle: Eigene Darstellung, Daten: MAS 2021, alle Beobachtungen sind auf der log-log-Skala dargestellt.
Anm.: Die potenziellen Ausreisser sind orange markiert; BACON mit den Parametern alpha = 0.01, collect = 2.

Tabelle 6: Anzahl identifizierte (potenzielle) Ausreisser nach Mustern

Muster	Anzahl potenzielle Ausreisser	Muster	Anzahl potenzielle Ausreisser
25	1	31	37
27	1	60	3
28	2	61	2
29	4	62	2
30	4	63	29

Quelle: Eigene Darstellung, Daten: MAS 2021; BACON mit den Parametern alpha = 0.01, collect = 2.

5 Deduktive Eingriffe

In Kapitel 4 wurden Ausreisser und Artefakte unter den Praxen identifiziert und entfernt. Dort setzte die Untersuchung auf der **Ebene der Totale** der Finanzdaten an. In diesem Kapitel werden die **Subtotale** (Unterpositionen) betrachtet, die sich zu den Totalen addieren. Zum Beispiel: Das Total der Sachaufwände für medizinische Tätigkeiten (`xet_mat_total`) wird als Summe von vier Subtotalen `xet_mat_drug`, `xet_mat_lab`, `xet_mat_migel` und `xet_mat_other` gebildet(siehe Tabelle 1). Für dieses Beispiel muss also gelten:

$$\begin{aligned} \text{xet\_mat\_total} = \text{xet\_mat\_drug} + \text{xet\_mat\_lab} + \text{xet\_mat\_migel} + \\ \text{xet\_mat\_other} \end{aligned} \tag{S1}$$

und

$$\begin{aligned} \text{xet\_mat\_drug} &\geq 0, \\ \text{xet\_mat\_lab} &\geq 0, \\ \text{xet\_mat\_migel} &\geq 0, \\ \text{xet\_mat\_other} &\geq 0, \\ \text{xet\_mat\_total} &\geq 0. \end{aligned} \tag{S2}$$

Mit anderen Worten: Den Finanzdaten liegt eine hierarchische Struktur zu Grunde (hier am Beispiel der Sachaufwände für medizinische Tätigkeiten illustriert), die die Beziehungen zwischen den einzelnen Subtotalen bestimmten Bedingungen oder Regeln unterwirft. So müssen sich etwa die Subtotale zum jeweiligen Total addieren und es dürfen keine negativen (oder fehlenden) Aufwände und Erträge vorkommen; vgl. (S1) und (S2). Für die übrigen Totalwerte können analoge Regeln oder Bedingungen formuliert werden.

Es gibt eine Vielzahl von Möglichkeiten, wie Daten im Widerspruch zu den Regeln sein können. Wir gehen im folgenden Kapitel auf einige Beispiele ein und zeigen auf, welche **Eingriffe** (edits) nötig sind, um die Widersprüche oder Verstösse gegen die Regeln aufzulösen.

5.1 Beispiele

Im Folgenden arbeiten wir mit fiktiven Daten einer Praxis und betrachten nur Fälle, bei welchen ein Subtotal fehlt (NA, not available) oder nicht korrekt ist.

Beispiel 5.1: Ein fehlender Wert

	xet_mat_drug	xet_mat_lab	xet_mat_migel	xet_mat_other	xet_mat_total
Wert:	12	0	**NA**	29	42

Das Subtotal `xet_mat_migel` fehlt in Beispiel 5.1. Darum ist Bedingung (S1) verletzt; auch die Bedingung $\texttt{xet\_mat\_migel} \geq 0$ ist verletzt. Durch die Einsetzung von $\texttt{xet\_mat\_migel} = 1$ sind die Bedingungen (S1) und (S2) erfüllt. Dies ist ein trivialer Fall.

Beispiel 5.2: Zahlendreher und Schreibfehler

	xet_mat_drug	xet_mat_lab	xet_mat_migel	xet_mat_other	xet_mat_total
Wert:	**21**	0	1	29	42

Die Beobachtung mit dem Wert 21 in Variable `xet_mat_total` (siehe Beispiel 5.2) verstösst gegen Bedingung (S1), weil die Summe der Subtotale den Wert 51 hat, das Total (`xet_mat_total`) jedoch den Wert 42. Es handelt sich hierbei um einen Zahlendreher. Mit dem Eingriff $\texttt{xet\_mat\_drug} = 12$ (anstatt 21) kann das Problem behoben werden.

Beispiel 5.3: Masseinheitsfehler

	xet_mat_drug	xet_mat_lab	xet_mat_migel	xet_mat_other	xet_mat_total
Wert:	**120**	0	1	29	42

Beispiel 5.3: Die Beobachtung mit dem Wert 120 in Variable `xet_mat_total` verstösst gegen Bedingung (S1), weil die Summe der Subtotale den Wert 150 hat, das Total `xet_mat_total` jedoch den Wert 42. Es handelt sich hierbei um einen Masseinheitsfehler. Der Wert bei `xet_mat_drug` ist um den Faktor 10 zu gross. Mit dem Eingriff $\texttt{xet\_mat\_drug} = 12$ kann das Problem behoben werden.

5.2 Deduktive Eingriffe

In den Beispielen 5.1 – 5.3 wurden einzelne deduktive Eingriffe (deductive edits oder deductive corrections) vorgestellt. Um die **Möglichkeiten und Grenzen** dieser Eingriffe zu verstehen und sie im Kontext der MAS-Daten anzuwenden, ist es notwendig, eine Definition anzugeben. Dabei orientieren wir uns an De Waal (2009), siehe auch de Waal et al. (2011, Kapitel 2), und halten fest: Deduktive Eingriffe sind **nur dann möglich**, wenn ein Widerspruch oder eine Inkonsistenz in den Daten (z. B. eine Verletzung der Bedingungen S1 und S2 in den Beispielen 5.1–5.3) durch **logische Argumente eindeutig** aufgelöst werden kann. Treten, wie im nachfolgenden Beispiel 5.4 dargestellt, zwei fehlende Werte auf, so ist es (mit dem vorhandenen Wissen) nicht möglich, die Werte der Variablen `xet_mat_drug` und `xet_mat_lab` eindeutig zu bestimmen. Es gibt unendlich viele Möglichkeiten, Werte so einzusetzen, dass die Bedingungen (S1) und (S2) erfüllt sind.

Beispiel 5.4: Mehrere fehlende Werte

	xet_mat_drug	xet_mat_lab	xet_mat_migel	xet_mat_other	xet_mat_total
Wert:	**NA**	**NA**	1	29	42

Der Spielraum für deduktive Eingriffe ist weniger begrenzt, als es auf den ersten Blick scheint. Es können auch **Verkettungen** von Inkonsistenzen aufgelöst werden. So können z. B. ein Zahlendreher (z. B. 67 statt 76) in einem Eintrag und eine Ziffernauslassung (z. B. 427 statt 4 627) in einem anderen Eintrag aufgelöst werden. Weitere mögliche Fehler, die identifiziert werden können sind: fälschliches Hinzufügen von Ziffern (z. B. 1 201 anstatt 121) oder Vorzeichenfehler (negative Werte), siehe De Waal et al. (2011, Kapitel, 2.3.3).

Die deduktiven Eingriffe folgenden dem **Sparsamkeitsprinzip**, das heisst, der Zielzustand (der durch vorab definierte Bedingungen festgelegt ist) soll mit einer möglichst geringen Anzahl von Eingriffen erreicht werden.

Wir haben zu allen Total- und Subtotalwerten (siehe Tabelle 1) Bedingungen analog zu (S1) und (S2) formuliert; hierzu verwenden wir die Funktionalität des R-Pakets `validate` (Van der Loo und de Jonge, 2021; Van der Loo et al., 2024). Diese Bedingungen sind sparsam parametrisiert und fordern nur ein, dass alle Subtotale ≥ 0 sind und, dass das Total der Summe der Subtotale entspricht. Für die Überprüfung der Einhaltung der Regeln und die entsprechenden Eingriffe verwenden wir (mehrheitlich) die Funktionen in den R-Paketen `editrules` (De Jonge und van der Loo, 2024) und `deducorrect` (Van der Loo et al., 2015).

Tabelle 7: Anzahl deduktive Eingriffe (bei den Subtotalen zu `xit_pra` und `xet_mat`)

Subtotal	Anz. Eingriffe
xit_pra_doc	869
xit_pra_drug	806
xit_pra_lab	16
xit_pra_migel	654
xit_pra_ndoc	853
xit_pra_other	874

Subtotal	Anz. Eingriffe
xet_mat_drug	892
xet_mat_lab	509
xet_mat_migel	980
xet_mat_other	935
Total	**7 388**

Quelle: Eigene Darstellung, Daten: MAS 2021.

Die Anzahl der deduktiven Eingriffe bei den Subtotalen zu `xit_pra` und `xet_mat` sind in Tabelle 7 nach Subtotalen differenziert aufgeführt. Insgesamt werden die Werte für 7'388 Beobachtungen angepasst oder eingesetzt. Das scheint auf den ersten Blick viel zu sein. Wenn wir uns jedoch vergegenwärtigen, dass es sich um 10 Variablen mit Angaben zu jeweils 8'183 Praxen handelt, dann erscheint die Gesamtanzahl kleiner. Nur etwa 8.9 % der Werte wurden angepasst oder eingesetzt.

6 Dateneinsetzung (Imputation)

Fehlende Werte, die nicht bereits durch die deduktiven Eingriffe eingesetzt oder angepasst werden konnten, werden auf der Grundlage modellbasierter Verfahren eingesetzt (imputiert). Dabei wird ein zweistufiges Vorgehen verfolgt (vgl. De Waal et al., 2011, Kapitel 10):

1. Einsetzung fehlender Werte;
2. Anpassung/Korrektur der eingesetzten Werte mit Ausgleichseingriffen; siehe Kapitel 7.

Wahl der Einsetzungsmethoden

Die Wahl der Einsetzungsmethoden (und deren Parametrisierung) stützt sich auf theoretische Überlegungen und Erkenntnisse einer Simulationsstudie, deren wichtigsten Erkenntnisse (beispielhaft) in Anhang A aufgeführt sind.[13] Es wurden folgende Methoden ausgewählt:

- Einsetzung nach der Zero-to-Zero-Methode (wurde bereits im Bericht von Hulliger und Bisang, 2021, verwendet);
- Einsetzung auf der Grundlage robust geschätzter Regressionsmodellen (wurde bereits im Bericht von Hulliger und Bisang, 2021, verwendet);
- Einsetzung nach der Methode des Gradient Boostings.

Die gewählten Methoden werden in den folgenden Abschnitten besprochen.

6.1 Einsetzung nach der Zero-to-Zero-Methode

Die wichtigsten Eigenschaften der Methode Zero-to-Zero und deren Eignung im Kontext der MAS-Daten wurden in Hulliger und Bisang (2021) besprochen. Aus diesem Grund halten wir uns hier kurz.

Als Zero-to-Zero Einsetzungen werden Imputationen bezeichnet, bei denen fehlende Werte bei Ertragskonten mit einer Null eingesetzt werden, wenn das entsprechende Aufwandskonto auch einen Nullwert aufweist. Und umgekehrt (d. h. wenn das Aufwandskonto Null ist und beim korrespondierenden Ertragskonto kein Wert eingetragen wurde, wird beim Ertragskonto eine Null eingesetzt). Die Idee dabei ist, dass die Aufwände und Erträge einander möglichst entsprechen sollten; wenn z. B. beim Labormaterial keine Aufwände anfallen, ist es plausibel anzunehmen, dass auch beim Laborertrag keine Erträge durch Labor-

[13] Die Resultate der Einsetzung nach der k-Nearest-Neighbormethode (kNN, vgl. bspw. Templ, 2023, Kapitel 7.2 oder De Waal et al., 2011, Kapitel 7.6) waren den anderen Methoden in der Simulationsstudie unterlegen. Darum wird der Einsatz der kNN-Methode nicht weiterverfolgt.

analysen erwirtschaftet wurden. Die Einsetzungen nach der Zero-to-Zero-Methode werden auf die in Tabelle 8 aufgeführten Variablenkombinationen angewendet. Ebenda sind die Häufigkeiten der Einsetzungen pro Variable aufgeführt. Die Anzahl der Einsetzungen ist bei den Erträgen grösser als bei den Aufwänden. Dies liegt (neben anderem auch) daran, dass der Anteil der fehlenden Werte bei den Erträgen grösser ist.

Tabelle 8: Einsetzungen nach der Zero-to-zero-Methode

Aufwandsposition	Anzahl Einsetzungen
xet_mat_drug	0
xet_mat_lab	166
xet_mat_migel	252
xet_mat_other	223

Ertragsposition	Anzahl Einsetzungen
xit_pra_drug	494
xit_pra_lab	798
xit_pra_migel	680
xit_pra_ndoc	145
xit_pra_other	554

Quelle: Eigene Darstellung, Daten: MAS 2021.

6.2 Einsetzung auf Grundlage geschätzter Regressionsmodelle

Einsetzungen auf der Grundlage (robust) geschätzter Regressionsmodelle wurden bereits in Hulliger und Bisang (2021) besprochen. Die wichtigsten Eigenschaften und ihre Eignung im Kontext der MAS-Daten können diesem Bericht entnommen werden. Die (robuste) Regression wird überdies in der Literatur im Zusammenhang mit fehlenden Werten ausführlich diskutiert; siehe bspw. Templ (2023, Kapitel 8), Van der Loo und de Jonge (2018, Kapitel 10) und De Waal et al. (2011, Kapitel 7). Aus diesem Grund wird an dieser Stelle nur kurz darauf eingegangen.

Die Parameter der Modelle wurden mit dem MM-Regressionsschätzer geschätzt; siehe bspw. Maronna et al. (2019, Kapitel 5.5). Die Auswahl der Variablen basierte auf der schrittweisen MM-Regression unter Minimierung des robust final prediction error (RFPE). Dabei wird eine Variable nach der anderen entfernt (backward elimination) bis der Wert des RFPE minimal ist; siehe Maronna et al. (2019, Kapitel 5.6). Wir verwenden die Implementation des MM-Schätzers im R-Package `RobStatTM` (Yohai et al., 2023) und den folgenden Annahmen:

Annahme:
a) Die (robusten) Regressionen werden nur dann für die Einsetzung verwendet, wenn der (robuste) adjustierte Determinationskoeffizient (R^2) grösser als 0.3 ist.
b) Für die Modellschätzung werden nur Beobachtungen verwendet, deren Werte für die relevanten Variablen vollständig sind, d. h. nicht fehlen.

Beispiel 6.1: Die geschätzten Parameter sind nachfolgend *beispielhaft* für die Regression zu den Sachaufwänden für Medikamente (xet_mat_drug) in Tabelle 9 aufgeführt. Aufgeführt ist

das Modell, das das Kriterium des robust final prediction error (RFPE) minimiert. Tabelle 9 zeigt, dass einige der geschätzten Koeffizienten auf dem 5%-, 1%- bzw. 0.1%-Signifikanzniveaus nicht signifikant von Null verschieden sind (siehe Sternchen und entsprechende Signifikanzcodes). Die entsprechenden Variablen wurden im Modell belassen, weil die Modellselektion auf dem Kriterium des RFPE und nicht auf der Inferenzstatistik basiert.

Tabelle 9: Regressionstabelle zur Modellierung der Variablen `log1p(xet_mat_drug)`

Variable/ Effekt	Koeffizient	Std. Fehler	t-Wert	P-Wert
(Intercept)	1.03E-01 ***	3.48E-02	2.96	0.00314
log1p(xit_pra_drug)	9.77E-01 ***	3.40E-03	287.52	< 2.00E-16
pe_med_assist_nb	5.38E-04	3.69E-04	1.46	0.14518
contact_tot_nb	-4.66E-07	3.02E-07	-1.55	0.12214
main_activity_cd_ano_1_dicho	-4.87E-02 ***	1.46E-02	-3.33	0.00088
med_off_orient_cd_1_dicho	-2.97E-02 *	1.41E-02	-2.10	0.03608
drug_supplier_cd_1_dicho	-1.63E-02	1.20E-02	-1.36	0.17334

Anzahl Beobachtungen: 2 369, Daten: MAS 2021

Robust residual standard error: 0.2111

Multiple R-squared: 0.9534, Adjusted R-squared: 0.9533

Signifikanzcodes: *** < 0.001, ** < 0.01, * < 0.05

`log1p` steht für die Funktion $x \mapsto \log(x+1)$

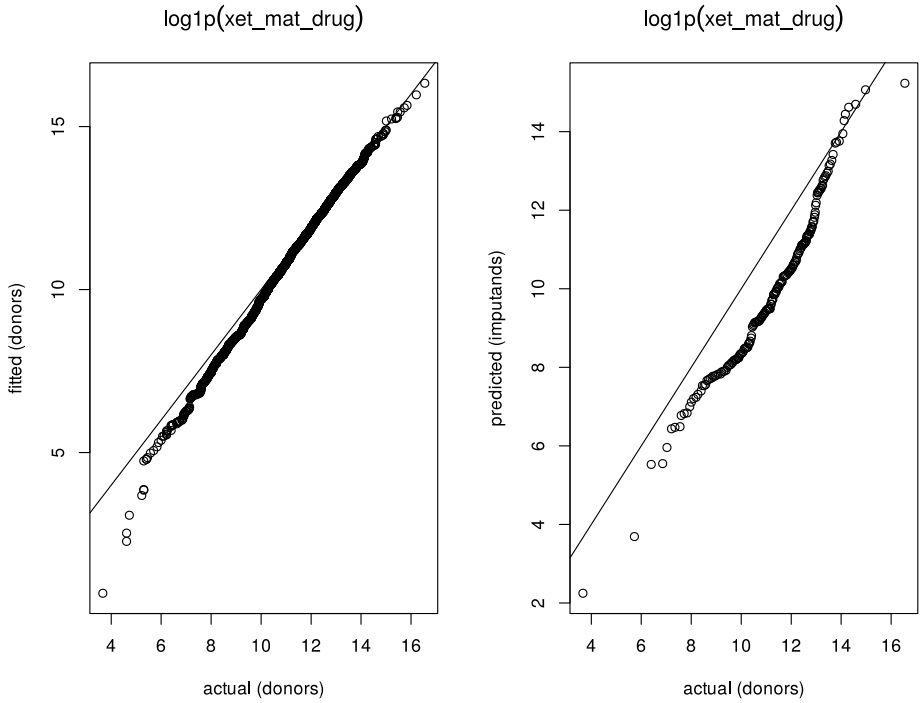
Eigene Darstellung, Daten: MAS 2021

In Abbildung 5 sind zwei Quantil-Quantil-Plots (QQ-Plot) zum Modell aus Tabelle 9 dargestellt. Auf der linken Seite befindet sich der QQ-Plot der tatsächlichen Werte (actual values) im Vergleich zu den angepassten Werten (fitted values). Aus diesem QQ-Plot geht hervor, dass die Verteilung der beiden Werte ein hohes Mass an Übereinstimmung aufweist.

Auf rechten Seite ist der QQ-Plot der tatsächlichen Werte (actual values) gegen die vorhergesagten Werte (predicted values) dargestellt. Bei den vorhergesagten Werten handelt es sich um Beobachtungen, für die die Sachausgaben für Arzneimittel (xet_mat_drug) fehlen, die aber mit dem Modell vorhergesagt werden können. Dieser QQ-Plot zeigt, dass die Verteilung der tatsächlichen Werte und die Verteilung der vorhergesagten Werte eine akzeptable Übereinstimmung aufweisen.

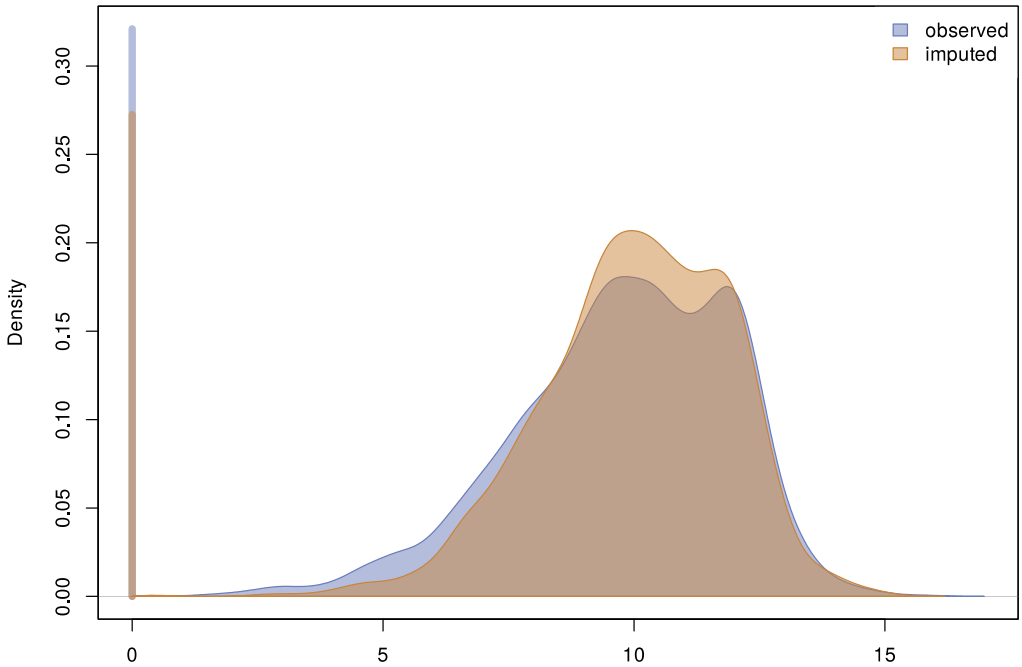
Abbildung 6 zeigt die Schätzung der Kerndichte für die beobachteten (observed) und eingesetzten (imputed) Beobachtungen für die Sachaufwände für Medikamente nach der Imputation (einschliesslich der Schätzung des Punktmasses bei Null). Aus dieser Darstellung geht hervor, dass der Anteil der Nullwerte bei den eingesetzten Beobachtungen etwas geringer ist als bei den beobachteten. Die Unterschiede fallen jedoch nicht ins Gewicht. Für Beobachtungen mit Werten > 0 unterscheiden sich die geschätzten Kerndichten nur geringfügig.

Abbildung 5: QQ-Plot actual vs. fitted (Panel links) und actual vs. predicted (Panel rechts) für die Variable `xet_mat_drug`



Quelle: Eigene Darstellung, Daten: MAS 2021; Anm.: `log1p` steht für die Funktion $x \mapsto \log(x+1)$

Abbildung 6: Separate Kerndichteschätzung (und Schätzung des Punktmasses bei 0) zur Variablen `xet_mat_drug` für die Originalbeobachtungen (observed) und die eingesetzten Beobachtungen (imputed)



Quelle: Eigene Darstellung, Daten: MAS 2021 .

Zwischenbillanz zur Einsetzung nach der Regressionsmethode

Die Resultate wurden nur für das Beispiel der Sachaufwände für Medikamente erläutert. Bei den übrigen Variablen ist das Vorgehen analog.

- Der Erklärungsgehalt der geschätzten Modelle ist generell hoch. Dies liegt auch daran, dass wir per Annahme nur dann Einsetzungen vornehmen, wenn der (robuste) adjustierte Determinationskoeffizient (R^2) grösser als 0.3 ist. Keine Einsetzungen wurden für die Subtotale `xet_mat_other` und `xit_pra_other` vorgenommen, weil der Erklärungsgehalt der Modelle zu gering war; siehe Tabelle 10.
- Die in Tabelle 10 dargestellte Anzahl der Einsetzungen zeigt, dass Einsetzung nach der Regressionsmethode bei den Sachaufwänden – im Vergleich zu den Erträgen – seltener angewendet werden konnte. Dies ist vor allem darauf zurückzuführen, dass die erklärenden Variablen in den Modellen für die Sachaufwände viele fehlende Werte enthalten.

Tabelle 10: Einsetzungen nach der Regressionsmethode

Aufwandsposition	Anzahl Einsetzungen
`xet_mat_drug`	194
`xet_mat_lab`	188
`xet_mat_migel`	107
`xet_mat_other`	[keine]

Ertragsposition	Anzahl Einsetzungen
`xit_pra_drug`	2 540
`xit_pra_lab`	2 359
`xit_pra_migel`	2 460
`xit_pra_ndoc`	648
`xit_pra_doc`	184
`xit_pra_other`	[keine]

Quelle: Eigene Darstellung, Daten: MAS 2021.

6.3 Einsetzung nach der Gradient-Boosting-Methode

Nach Anwendung der Zero-Zero-Methode und der Regressionsmethode verbleibt eine beträchtliche Anzahl fehlender Werte. Der nächste Schritt besteht darin, diese fehlenden Werte zu imputieren. Dabei kann nicht auf einfache, aber erklärungsstarke Regressionsmodelle zurückgegriffen werden.

Da keine einfachen Modelle zur Verfügung stehen, wird die Methode des Gradient Boosting verwendet. Dabei handelt es sich um eine Technik des maschinellen Lernens; siehe Hastie et al. (2009, Kapitel 10) und – im Kontext fehlender Daten – Templ (2023, Kapitel 9). Gradient Boosting liefert ein prädiktives Modell in Form eines Ensembles schwacher prädiktiver Modelle (weak learner), d. h. geschätzte Modelle, die nur sehr wenige Annahmen über die Daten machen und als einfache Entscheidungsbäume aufgebaut sind.

Wir verwenden die Implementation XGBoost (Chen et al., 2024) mit der Tweedie Loss-Funktion. Für die Schätzung der Parameter werden nur vollständige Beobachtungen verwendet, die nicht bereits vorgängig (d. h. mit den Methoden Zero-to-Zero oder der Regression) imputiert wurden. Die Menge der möglichen Erklärungsvariablen sind in Tabelle 11 aufgeführt. Die Tuning-Parameter von XGBoost wurden im Rahmen der Simulationsstudie (siehe Anhang A) und mittels Kreuzvalidierung (MAS-Datensätze 2017–2021) festgelegt.

Die Anzahl der Einsetzungen nach der Gradient-Boosting-Methode sind in Tabelle 12 aufgeführt. Erwartungsgemäss wurden viele fehlende Werte eingesetzt.

Tabelle 11: Mögliche Erklärungsvariablen

Variablen	Variablen (Fortsetzung)	Variablen (Fortsetzung)
xet_mat_total	xet_mat_lab	main_activity_cd_ano_4_dicho
xit_pra_total	xet_mat_migel	med_off_orient_cd_1_dicho
xet_per_total	xet_mat_other	med_off_orient_cd_2_dicho
xet_oth_total	xet_per_sdoc	drug_supplier_cd_1_dicho
xit_pra_drug	xet_per_shpad	drug_supplier_cd_2_dicho
xit_pra_lab	xet_oth_occup	drug_supplier_cd_3_dicho
xit_pra_migel	xet_oth_capit	pe_med_assist_nb
xit_pra_other	contact_tot_nb	pe_care_nb
xit_pra_doc	main_activity_cd_ano_1_dicho	pe_dir_admin_nb
xit_pra_ndoc	main_activity_cd_ano_2_dicho	pe_informatics_nb
xet_mat_drug	main_activity_cd_ano_3_dicho	

Quelle: Eigene Darstellung

Tabelle 12: Einsetzungen nach der Gradient-Boosting-Methode

Aufwandsposition	Anzahl Einsetzungen
xet_mat_drug	4 510
xet_mat_lab	4 976
xet_mat_migel	5 039
xet_mat_other	5 199

Ertragsposition	Anzahl Einsetzungen
xit_pra_drug	3 035
xit_pra_lab	3 158
xit_pra_migel	3 140
xit_pra_ndoc	5 539
xit_pra_doc	5 992
xit_pra_other	5 510

Quelle: Eigene Darstellung, Daten: MAS 2021.

7 Ausgleichseingriffe

Wir hatten in Kapitel 5 das Total `xet_mat_total` und dessen Subtotale untersucht. Für dieses Beispiel wurden die Bedingungen (S1) und (S2) formuliert, die wir an dieser Stelle (zur besseren Auffindbarkeit) wieder aufgreifen:

$$\begin{aligned} \text{xet\_mat\_total} = \text{xet\_mat\_drug} + \text{xet\_mat\_lab} + \text{xet\_mat\_migel} + \\ \text{xet\_mat\_other} \end{aligned} \tag{S1}$$

und

$$\begin{aligned} \text{xet\_mat\_drug} &\geq 0, \\ \text{xet\_mat\_lab} &\geq 0, \\ \text{xet\_mat\_migel} &\geq 0, \\ \text{xet\_mat\_other} &\geq 0, \\ \text{xet\_mat\_total} &\geq 0. \end{aligned} \tag{S2}$$

Die Einsetzung fehlender Werte (siehe Kapitel 6) führt in den meisten Fällen dazu, dass die Bedingungen (S1) und (S2) nicht erfüllt sind. Diesem Problem begegnen wir mit Ausgleichseingriffe (engl. balance edits). Damit sind Eingriffe gemeint, die dazu führen, dass sich die Subtotale auf das jeweilige Total addiert. Im Unterschied zu den deduktiven Eingriffen (vgl. Kapitel 5) werden bei den Ausgleichseingriffen aber nicht Vorzeichen-, Rundungs- und Tippfehler in den Originaldaten korrigiert. Stattdessen werden ausschliesslich zuvor eingesetzte Subtotale angepasst.

Es bezeichne $\mathbf{x}_0$ den p-Vektor *einer* eingesetzten und anzupassenden Beobachtungen aus Subtotalen. Wir wollen das Vorgehen anhand eines fiktiven Beispiels zu den Materialaufwänden illustrieren. In diesen Fall hat $\mathbf{x}^0$ nach Abschluss der Einsetzung die folgenden (fiktiven) Werte:

Beispiel 7.1

	`xet_mat_drug`	`xet_mat_lab`	`xet_mat_migel`	`xet_mat_other`	`xet_mat_total`
Wert:	12	0	3	29	42
Unverändert:	Ja	Nein	Nein	Nein	Ja

In der Zeile «Unverändert» ist vermerkt, ob der Wert der Variablen den (ursprünglich) beobachten Erhebungsdaten entspricht oder während des Datenaufbereitungsprozesses verändert (z. B. eingesetzt oder angepasst) wurde.

Wir erkennen leicht, dass die Summe der Subtotale `xet_mat_drug`, `xet_mat_lab`, `xet_mat_migel` und `xet_mat_other` in Beispiel 7.1 den Wert 44 hat und somit das ausgewiesene Total

(`xet_mat_total`) in der Höhe von 42 um 2 übersteigt. Ausgleichseingriffe werden nun unter der folgenden Annahme durchgeführt:

Annahme: Die Einsetzung erfolgte modellgestützt. Darum ist sie mit Unsicherheiten behaftet. Den ursprünglichen (unveränderten) Beobachtungen wird eine höhere Glaubwürdigkeit beigemessen. Aus diesen Gründen werden Ausgleichseingriffe nur auf Subtotale angewendet, die während des Datenaufbereitungsprozesses eingesetzt oder angepasst wurden.

7.1 Ausgleichseingriffe als Optimierungsproblem

Mit den zur Verfügung stehenden Informationen ist es nicht möglich zu eruieren, welches Subtotal bzw. welche Subtotale in Beispiel 7.1 angepasst werden müssen, damit die Bedingungen (S1 und S2) eingehalten sind. Es gibt – ausser in trivialen Fällen – eine *Vielzahl* von Möglichkeiten. Unter allen Möglichkeiten soll nun derjenige Ausgleichseingriff ausgewählt werden, bei welchem das Ausmass des Eingriffs möglichst *klein* ist (im Vergleich zur ursprünglichen Beobachtung $\mathbf{x}^0$). Dies kann als Optimierungsproblem unter Nebenbedingungen formuliert werden (De Waal et al., 2011, Kapitel 10.2.2)

$$\begin{aligned} &\min_{\mathbf{x}} (\mathbf{x}-\mathbf{x}^0)^\mathsf{T}\mathbf{W}(\mathbf{x}-\mathbf{x}^0) \\ &\text{s.t.c.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \end{aligned}$$

wobei $\mathbf{A}$ eine $(k \times p)$ Matrix und $\mathbf{b}$ ein k-Vektor ist, die für die Formulierung der k Nebenbedingungen (siehe oben) verwendet werden. Die Matrix $\mathbf{W}$ ist eine $(p \times p)$ Diagonalmatrix, $\mathbf{W} = \text{diag}(w_1, \ldots, w_p)$ mit sog. confidence weights $w_1, \ldots, w_p$. Diese Gewichte widerspiegeln das «Vertrauen», das wir in die Gültigkeit der Werte $\mathbf{x}^0 = (x_1^0, \ldots, x_p^0)^\mathsf{T}$ haben. Werte mit grossen Gewichten werden weniger angepasst, weil ihre Anpassung einen grösseren Einfluss auf das Kriterium der gewichteten kleinsten Quadrate hat als Werte mit kleineren Gewichten. In der Praxis werden die Gewichte häufig so gewählt, dass sie dem reziproken Wert von $\mathbf{x}^0$ entsprechen; vgl. De Waal et al. (2011, Kapitel 10.2.2).

Wichtiger als die Wahl der Gewichte ist, dass Ausgleichseingriffe nur bei denjenigen Beobachtungen erfolgen, die während des Datenaufbereitungsprozesses verändert oder eingesetzt wurden (in Beispiel 7.1 sind dies: `xet_mat_lab`, `xet_mat_migel` und `xet_mat_other`). Nach den Ausgleichseingriffen resultiert für Beispiel 7.1 die folgende Verteilung:

Beispiel 7.1 (nach den Ausgleichseingriffen)

	xet_mat_drug	xet_mat_lab	xet_mat_migel	xet_mat_other	xet_mat_total
Wert	12	0	**2.8125**	**27.1875**	42
Unverändert	Ja	Nein	Nein	Nein	Ja

Wir erkennen leicht, dass nur die Werte der Subtotalen `xet_mat_migel` und `xet_mat_other` angepasst wurden. Die Summe der Subtotale entspricht dem Total (`xet_mat_total`).

7.2 Umsetzung

Die Umsetzung der Ausgleichseingriffe erfolgte mit den Funktionen im R-Paket `rspa` (Van der Loo, 2022), wie sie in Kapitel 7.1 (kursorisch) beschrieben wurden. Bei den Subtotalen zu den Materialaufwänden wurden jeweils zwischen 2'200 und 2'500 Eingriffe vorgenommen, was einem Anteil von etwa 26 % – 31 % entspricht;[14] siehe Tabelle 13. Anteile in dieser Grössenordnung sind nicht erstaunlich, weil bei den Materialaufwänden viele Subtotale eingesetzt und angepasst werden müssen; vgl. auch Tabelle 3.

Tabelle 13: Anzahl Ausgleichseingriffe am Beispiel Subtotale der Materialaufwände

	xet_mat_drug	xet_mat_lab	xet_mat_migel	xet_mat_other
Anzahl Ausgleichseingriffe	2 233	2 280	2 329	2 482

Quelle: Eigene Darstellung, Daten: MAS 2021.

In der Simulationsstudie (siehe Anhang A) hatten wir festgestellt, dass die Ausgleichseingriffe bei einzelnen Variablen zu vergleichsweise grossen Abweichungen führt. Aus diesem Grund wurden die Ausgleichseingriffe nicht für alle Variablen durchgeführt. Bei welchen Variablen dies der Fall ist, wird in Kapitel 8 erläutert.

Werden die Ausgleichseingriffe weggelassen, dann addieren sich die Subtotale nicht zum Total. Es gibt Diskrepanzen. Dies wird jedoch als unproblematisch erachtet, solange die Diskrepanzen gering sind (d. h. weniger als 100 CHF betragen).

[14] Die Anteile sind in Bezug auf die Anzahl der Praxen berechnet. Der MAS-Datensatz umfasst (nach Auschluss der Ausreisser und Artefakte) 8'183 Praxen.

8 Ergebnisse evaluieren

In diesem Kapitel sollen die Verteilungen der Variablen und die Verhältnisse zwischen den Aufwänden und Erträgen vor und nach der Einsetzung evaluiert werden.

Die mittleren Verhältnisse zwischen den Erträgen und Aufwänden vor (als Original bezeichnet) und nach der Einsetzung sind in Tabelle 14 dargestellt. Dies gilt auch für die Verteilung der Subtotale in Prozent des Totals, wie sie am Beispiel der Erträge in Tabelle 15 dargestellt ist. Daraus wird ersichtlich, dass die Verhältnisse und die Verteilungen durch die Einsetzung nur geringfügig verändert werden.

Tabelle 14: (Mittlere) Verhältnisse zwischen Erträgen und Aufwänden: vor (Original) und nach der Einsetzung

	Original	Nach der Einsetzung
`xit_pra_drug / xet_mat_drug`	1.13	1.13
`xit_pra_migel / xet_mat_migel`	0.50	0.44
`xit_pra_lab / xet_mat_lab`	2.23	2.65
`xit_pra_other / xet_mat_other`	1.41	1.92
`xit_pra_total / xet_tot`	1.22	1.22

Quelle: Eigene Darstellung, Daten: MAS 2021.

Tabelle 15: Verteilung der Ertragspositionen (in Prozent des Ertragstotals) vor (Original) und nach den Einsetzungen

		xit_pra_drug	xit_pra_migel	xit_pra_lab	xit_pra_doc	xit_pra_ndoc	xit_pra_other
Grundversorgung	Original	27%	1%	6%	64%	1%	2%
	nach Einsetzung	27%	1%	7%	59%	3%	3%
OP_Spezialisierung	Original	7%	1%	1%	88%	1%	2%
	nach Einsetzung	8%	1%	2%	82%	3%	4%
Psychiatrie	Original	3%	0%	0%	79%	16%	2%
	nach Einsetzung	3%	0%	1%	76%	17%	3%
Spezialisierung	Original	23%	1%	3%	68%	4%	3%
	nach Einsetzung	23%	1%	5%	64%	4%	3%

Quelle: Eigene Darstellung, Daten: MAS 2021.

9 Übersicht und Empfehlungen

Tabelle 16 gibt einen Überblick darüber, ob und wie die Einsetzungen und Anpassungseingriffe vorgenommen wurden. Es handelt sich dabei um eine Zusammenfassung und Konkretisierung der Ausführungen aus den Kapiteln 6 und 7. In der Spalte Qualität ist angegeben, ob die Qualität als ungenügend einzustufen ist. Für diese Variablen empfehlen wir, keine Einsetzungen vorzunehmen.

Tabelle 16: Übersicht zu den Einsetzungen

Variablen	Qualität	Anmerkung
xit_pra_doc		ohne Zero-To-Zero
xit_pra_drug		ohne Ausgleichseingriffe (Balancing)
xit_pra_lab		ohne Ausgleichseingriffe (Balancing)
xit_pra_migel		ohne Ausgleichseingriffe (Balancing)
xit_pra_ndoc		keine Einsetzung nach Regressionsmethode
xit_pra_other	ungenügend	
xet_mat_drug		ohne Zero-To-Zero
xet_mat_lab		
xet_mat_migel		
xet_mat_other		keine Einsetzung nach Regressionsmethode
xet_oth_capit	ungenügend	
xet_oth_depre		keine Einsetzung nach Regressionsmethode
xet_oth_inspra	ungenügend	
xet_oth_itadm		keine Einsetzung nach Regressionsmethode
xet_oth_occup		keine Einsetzung nach Regressionsmethode
xet_oth_other		keine Einsetzung nach Regressionsmethode
xet_oth_vehic	ungenügend	
xet_per_ext	ungenügend	
xet_per_other		
xet_per_sdoc		
xet_per_shpad	ungenügend	
xet_per_soc		keine Einsetzung nach Regressionsmethode
xet_per_soth		keine Einsetzung nach Regressionsmethode

Quelle: Eigene Zusammenstellung.

In der Spalte Anmerkungen von Tabelle 16 ist auch angegeben, wenn keine Einsetzungen nach Regressionsmethode durchführt wurden. Dies ist dann der Fall, wenn der Erklärungsgehalt der geschätzten Modelle als zu gering eingeschätzt wurde; siehe Kapitel 6. Darüber hinaus ist vermerkt, wenn einzelne Schritte (z. B. «ohne Zero-To-Zero») weggelassen wurden, weil sie zu nennenswerten Verzerrung geführt haben (vgl. Simulation).

Für Analysen im Kontext der Tarifierung wird empfohlen, nur für diejenigen Variablen/Subtotale Einsetzungen vorzunehmen (unter Berücksichtigung der in Tabelle 16 aufgeführten Einschränkungen), deren (Einsetzungs-) Qualität nicht als ungenügend bewertet wurde.

Literaturverzeichnis

Billor, N., Hadi, A. S. und Vellemann, P. F. (2000). BACON: Blocked Adaptive Computationally-efficient Outlier Nominators. Computational Statistics and Data Analysis 34, S. 279–298, DOI: 10.1016/S0167-9473(99)00101-2.

Bundesamt für Statistik (2018a). Erste Erhebung «Strukturdaten der Arztpraxen und ambulanten Zentren» (MAS 2015): Analyse von Teilnahme und Grundgesamtheit, Neuchâtel. (BFS-Nummer: be-d-14.04.05-01)
URL: https://dam-api.bfs.admin.ch/hub/api/dam/assets/4842232/master

Bundesamt für Statistik (2018b). Die Erhebung MAS: Möglichkeiten und Grenzen der Interpretation der Resultate, Neuchâtel. (BFS-Nummer: be-d-14.04.05-02)
URL: https://dam-api.bfs.admin.ch/hub/api/dam/assets/4946471/master

Bundesamt für Statistik (2022a). Rapport sur la qualité des données financières. Relevé des données structurelles des cabinets médicaux et des centres ambulatoires MAS, Neuchâtel (Numéro OFS be-f-14.04.05-03).
URL: https://dam-api.bfs.admin.ch/hub/api/dam/assets/25345407/master

Bundesamt für Statistik (2022b). Strukturdaten Arztpraxen – Kapitel Finanzen: Hilfe zum Ausfüllen des Fragebogens, Neuchâtel (BFS-Nummer: do-d-14.04.05.06).
URL: https://dam-api.bfs.admin.ch/hub/api/dam/assets/23745152/master

Bundesamt für Statistik (2023). Strukturdaten Arztpraxen: Variablenliste MAS 2021, Neuchâtel (BFS-Nummer: do-d-14.04.05.18)
URL: https://dam-api.bfs.admin.ch/hub/api/dam/assets/24625130/master

Chen, T., He, T., Benesty, M. et al. (2024). xgboost: Extreme Gradient Boosting, R package version 1.7.8.1. DOI: 10.32614/CRAN.package.xgboost

De Jonge, E. und van der Loo, M. P. J. (2024). editrules: Parsing, Applying, and Manipulating Data Cleaning Rules, R package version 2.9.5, DOI: 10.32614/CRAN.package. editrules

De Waal, T. (2009). Statistical Data Editing. In: Sample Surveys: Design, Methods and Applications hrsg. v. Pfeffermann, D. und Rao, C. R., Band 29A: Handbook of Statistics, Amsterdam: Elsevier, Kap. 9, S. 187–214, DOI: 10.1016/S0169-7161(08)00009-6

De Waal, T., Pannekoek, J. und Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation, Hoboken (NJ): John Wiley & Sons, DOI: 10.1002/9780470904848

Fang, K.-T., Kotz, S. und Ng, K.-W. (1990). Symmetric Multivariate and Related Distributions, London: Chapman and Hall / CRC Press.

Hastie, T., Tibshirani, R. und Friedman, J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction, New York: Springer-Verlag, 2. Aufl.

Hulliger, B. und Bisang, L. (2021). Datenaufbereitung der Finanzdaten der MAS 2017, Hauptbericht, Olten. Bericht zuhanden der FMH/ Ärztekasse.

Leimgruber, J. und Prochinig, U. (2023). Das Rechnungswesen der Unternehmung, Zürich: Verlag SKV.

Loeve, M. (2017). Probability Theory: Mineola (New York): Dover Publications, 3. Aufl.

Little, R. J. A. und Rubin, D. B. (2002). Statistical Analysis with Missing Data, New York: John Wiley & Sons, 2. Aufl.

Maronna, R. A., Martin, D. R., Yohai, V. J. und Salibian-Barrera, M. (2019). Robust statistics: Theory and Methods (with R): John Wiley & Sons.

R Development Core Team (2025). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

Särndal, C.-E. und Lundström, S. (2005). Estimation in Surveys with Nonresponse. Hoboken (NJ): John Wiley & Sons.

Schoch, T. (2021). wbacon: Weighted BACON algorithms for multivariate outlier nomination (detection) and robust linear regression. Journal of Open Source Software 6, S. 3238, DOI: 10.21105/joss.03238

Templ, M. (2023). Visualization and Imputation of Missing Values: With Applications in R, Cham: Springer Nature.

Van der Loo, M. P. J, de Jonge, E. Und Hsieh, P. (2024)validate: Data Validation Infrastructure, R package version 1.1.5, DOI: 10.32614/CRAN.package.validate

Van der Loo, M. P. J (2022). rspa: Adapt Numerical Records to Fit (in)Equality Restrictions, R package version 0.2.8, DOI: 10.32614/CRAN.package.rspa

Van der Loo, M. P. J. und de Jonge, E. (2018). Statistical Data Cleaning with Applications in R, New York: John Wiley and Sons, DOI: 10.1002/ 9781118897126

Van der Loo, M. P. J. und de Jonge, E. (2021). Data Validation Infrastructure for R. Journal of Statistical Software 97, S. 1–31, DOI: 10.18637/jss.v097.i10

Van der Loo, M. P. J., de Jonge, E. und Scholtus, S. (2015). deducorrect: Deductive Correction, Deductive Imputation, and Deterministic Correction, R package version 1.3.7, DOI: 10.32614/CRAN.package.deducorrect

Yohai, V. J., Maronna, R. A., Martin, D. R., Brownson, G., Konis, K. und Salibian-Barrera, M. (2023). RobStatTM: Robust Statistics: Theory and Methods, R package version 1.0.8. DOI: 10.32614/CRAN.package.RobStatTM

Anhang A: Simulationen

In den drei Unterkapiteln A.1 – A.3 geben wir einen Überblick zu den **simulierten** Einsetzungen für die drei Subtotale `xet_mat_lab`, `xet_mat_drug` und `xit_pra_other`. Diese Variablenauswahl gibt das Spektrum gut wieder. Die besten Ergebnisse (betr. den mittleren absoluten Fehler, siehe unten) werden bei der Variablen `xet_mat_lab` erzielt, die schlechtesten bei der Variablen `xit_pra_other`.

In der Simulation wurden fehlende Werte mit dem Mechanismus *missing completely at random* (MCAR) nach Little und Rubin (2002, Kapitel 1) in den Daten der MAS 2021 erzeugt. Anschliessend wurden die fehlenden Werte eingesetzt und mit den Originaldaten verglichen. Die Ergebnisse der Einsetzungen und Eingriffe sind für die folgenden Schritte in separaten Grafiken dargestellt:

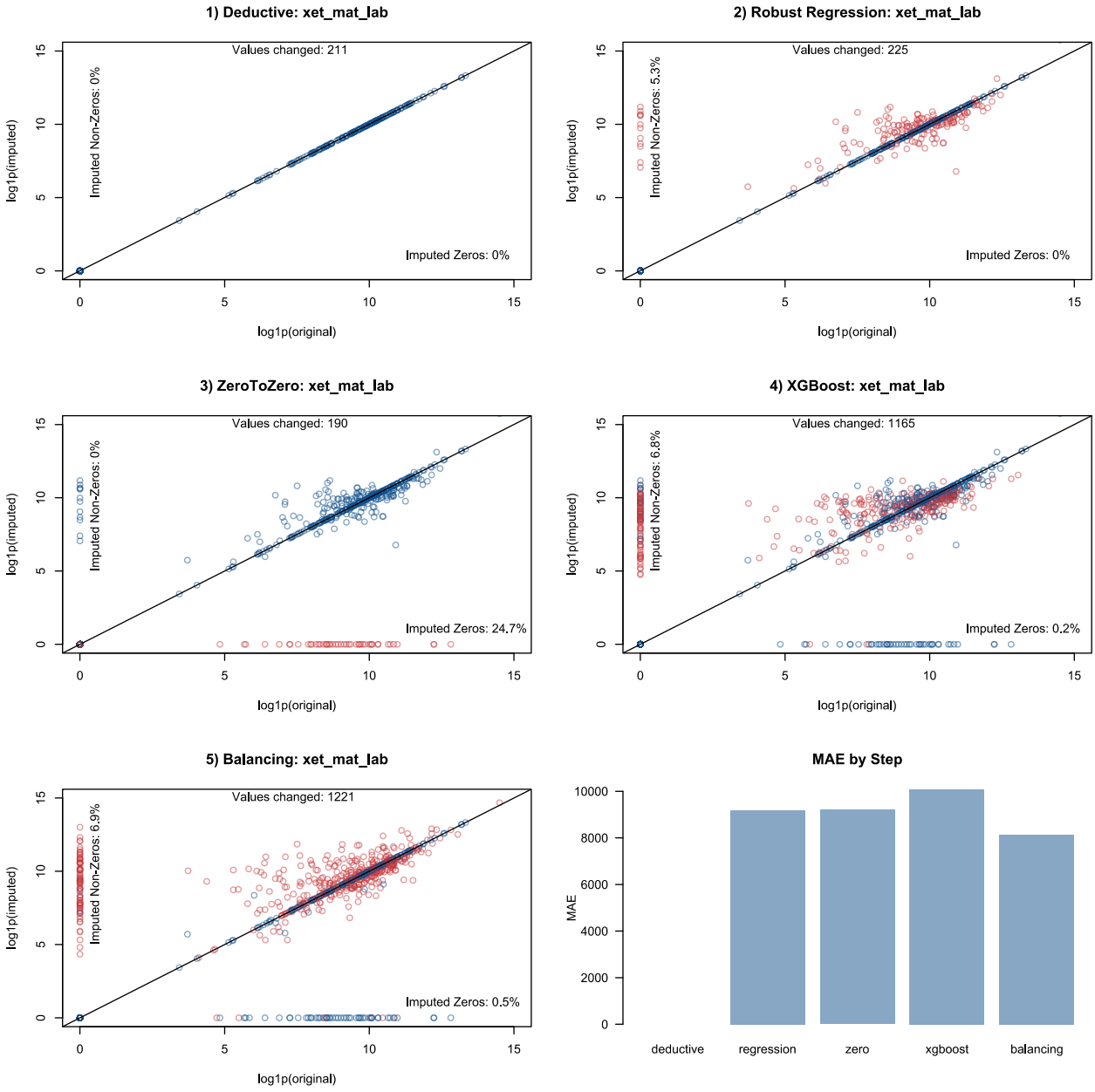
1. Deduktive Eingriffe (deductive edits)
2. Robuste Regression
3. Zero-to-Zero
4. Einsetzung mit Gradient Boosting (XGBoost)
5. Ausgleichseingriffe (balancing)

Weitere Anmerkungen:

- Bei der hier vorgestellten Simulation wurden nur fehlende Werte erzeugt, jedoch keine Fehler (z. B. Zahlendreher wie 67 anstatt 76) eingespielt. Aus diesem Grund werden bei der Simulation keine deduktiven Eingriffe vorgenommen.
- Die eingesetzten oder angepassten Beobachtungen sind rot dargestellt; die unveränderten Beobachtungen sind blau eingefärbt.
- Die sechste Grafik gibt einen Überblick zu den Schritten 1–5. Sie zeigt eine Auswertung zum mittleren absoluten Fehler (mean absolute error, MAE) für jeden Schritt (Einheit: Schweizer Franken).

Die Streudiagramme zeigen die (Original-) Beobachtungen auf der Abszisse vs. die angepassten bzw. imputierten Beobachtungen auf Ordinaten. Die Beobachtungen sind gegen die log-log-Skala aufgetragen; `log1p` steht für die Funktion $x \mapsto \log(x + 1)$

A.1 Überblick zur Simulation für `xet_mat_lab`

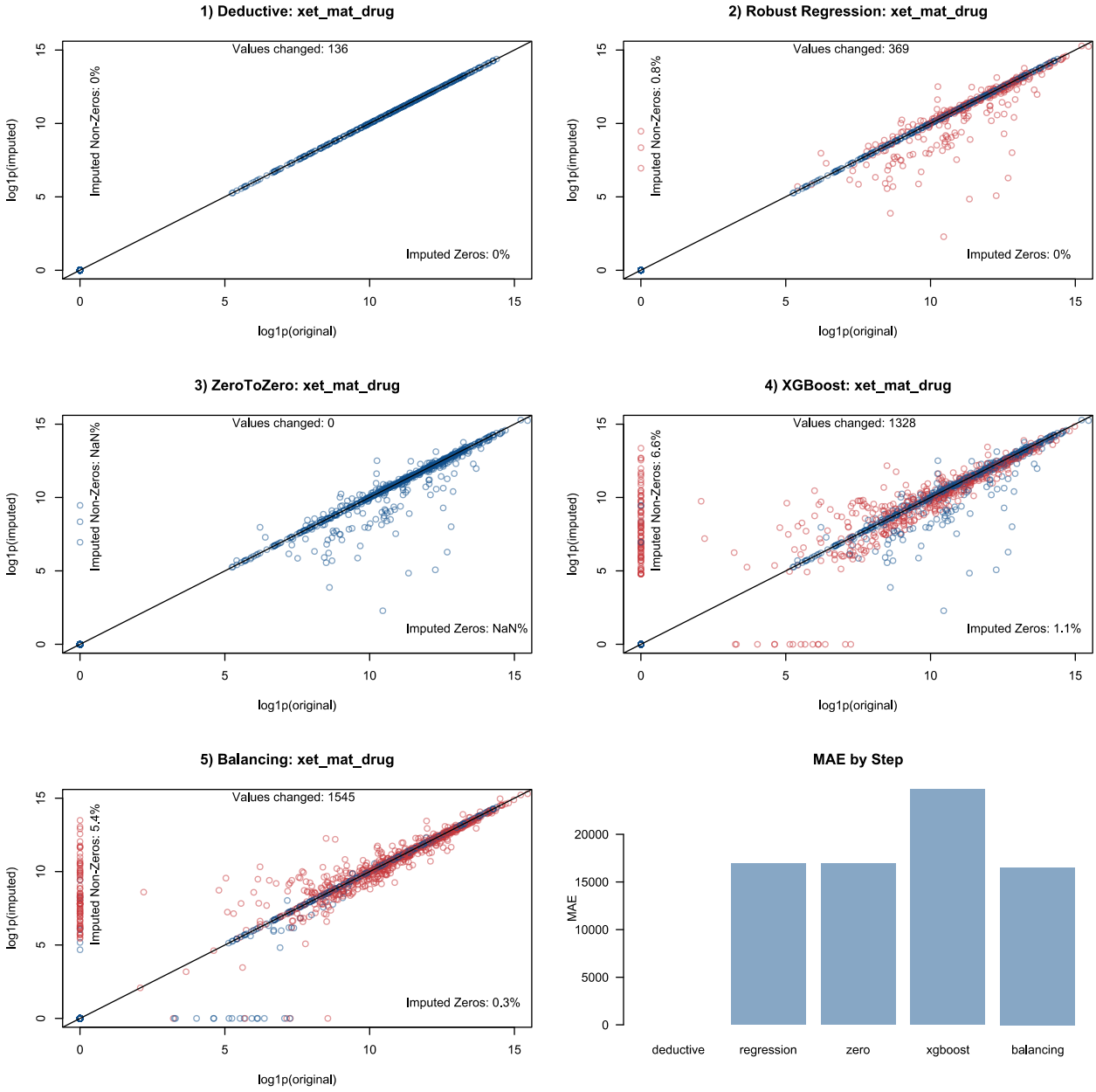


Quelle: Eigene Darstellung, Daten: MAS 2021.

Anmerkungen zur Grafik A.1:

- Der Anteil der eingesetzten Nullwerte, obwohl der Originalwert > 0 ist, ist bei allen Schritten gering.
- Der mittlere absolute Fehler (MAE) liegt (je nach Schritt) bei etwa CHF 8 000 bis 10 000.

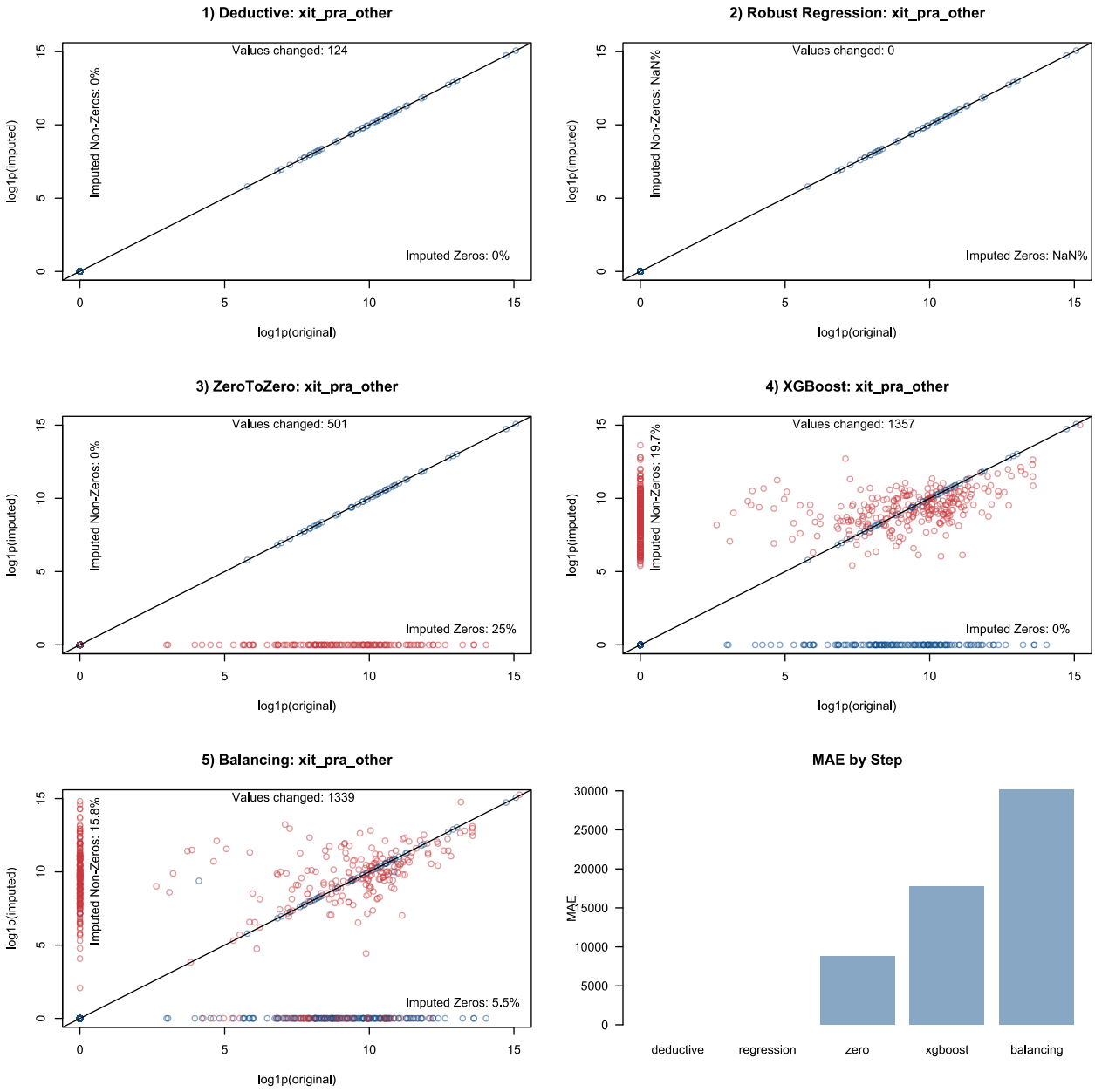
A.2 Überblick zur Simulation für `xet_mat_drug`



Quelle: Eigene Darstellung, Daten: MAS 2021.

Anmerkungen zur Grafik A.2: Die Einsetzung XGBoost (Schritt 4) führt zu einem MAE von fast CHF 25 000. Dieser Fehler ist im Vergleich zu A.2 relativ gross. Durch die darauffolgenden Ausgleichseingriffe (balancing) wird der Fehler wieder etwas reduziert, liegt jedoch immer noch bei etwa CHF 15 000.

A.3 Überblick zur Simulation für `xit_prat_other`



Quelle: Eigene Darstellung, Daten: MAS 2021.

Anmerkungen zur Grafik A.3:

- Die Einsetzung mit der Zero-to-Zero-Methode (Schritt 3) ist als problematisch anzusehen; sie erzeugt einen Anteil von 25 % Nullwerten, obschon die Originalwerte > 0 sind.
- Die Einsetzungen mit XGBoost (Schritt 4) weichen relativ stark von der Winkelhalbierenden ab; daraus resultiert ein vergleichsweise grosser MAE.
- Die Ausgleichseingriffe (balancing) führen zu einer Verschlechterung der eingesetzten Werte. Alle Einsetzungen und Eingriffe führen (kumulativ) zu einem vergleichsweise grossen MAE.