Vorhersage von zukünftigem Slum-Wachstum durch Data-Mining

Bachelor-Thesis von Bálint Klement

1. Gutachten: Prof. Dr. Johannes Fürnkranz
2. Gutachten: Lea Rausch

Erklärung zur Bachelor-Thesis

Hiermit versichere ich, die vorliegende Bachelor-Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

_____

(Balint Klement)

Darmstadt, den 30. September 2017

Inhaltsverzeichnis

1 Einführung

Laut einem Bericht der Vereinten Nationen lebt jeder siebte Mensch der Erde in einem Slum[1]. Slums sind informelle Siedlungen, die hohe Armuts- und Arbeitslosenquoten aufweisen und ein Hort für soziale Probleme sind – Menschen in Slums leben oft unter menschenunwürdigen Bedingungen. Präzise Vorhersagen darüber, welche Länder in Zukunft von starkem Slum-Wachstum betroffen sein werden, ermöglichen es den örtlichen Regierungen, rechtzeitig Gegenmaßnahmen zu ergreifen.

Gegenwärtig gängige Methoden der Slum-Modellierung befolgen einen bottom-up Ansatz und sind meist agentenbasiert oder arbeiten mit Geographischen Informationssystemen [SimMod]. Diese Arbeit versucht hingegen, die Slum-Entwicklung anhand Informationen auf Länder-Ebene zu modellieren. Diese Analyse auf globalen Daten bietet die interessante Möglichkeit, Gemeinsamkeiten in der Slum-Entwicklung auf der ganzen Erde festzustellen.

Die verwendeten Daten stammen aus den „World Development Indicators“ [2] der Weltbank, welche seit 1960 jährlich eine große Anzahl Indikatoren aus verschiedenen offiziellen, internationalen Quellen sammelt. Insgesamt werden aus 217 Ländern und ökonomischen Zonen 1453 Indikatoren aus den Bereichen Landwirtschaft, Ökonomie, Bildungswesen, Gesundheit, Infrastruktur, städtische Entwicklung, und vieles mehr für Forschungszwecke zur Verfügung gestellt. Einer dieser Indikatoren, „Population living in Slums (% of Urban)“, beschreibt die Slum-Bevölkerung und ermöglicht damit die Verwendung des Datensatzes in dieser Arbeit.

[1] https://sustainabledevelopment.un.org/content/documents/745habitat.pdf
[2] https://data.worldbank.org/data-catalog/world-development-indicators

1.1 Aufbereitung des Datensatzes

Das Vorkommen des benötigten Indikators „Slum Populaton (% of Urban)“ bestimmt maßgeblich die betrachteten Daten:

- Die Messungen erfolgten in 7 Jahren: 1990, 1995, 2000, 2005, 2007, 2009, 2014
- Für 106 Länder stehen Messdaten zur Verfügung. Es gibt Angaben zu Ländern aus jedem Gebiet der Erde außer Europa, wobei der Slum-Bevölkerungsanteil zwischen 3% und 97% liegt und durchschnittlich 50,63% beträgt.
- Somit haben wir insgesamt 436 einzelne Angaben über die Slumbevölkerung eines Landes zu einer gegebenen Zeit.

Diese 436 Angaben zur Slum-Bevölkerung mit den jeweiligen Land-Jahr-Kombinationen und dort gemessenen Indikatoren bilden die Datensätze, die in den Kapiteln 3 und 4 für die Untersuchungen verwendet werden. Die Anzahl betrachteter Indikatoren wurde auf 500 reduziert, wobei redundante und irrelevante Indikatoren entfernt wurden.

Die Datensätze sind Tabellen, in denen ein Eintrag durch einen Primärschlüssel aus Land und Jahr eindeutig bestimmt ist. Ein Eintrag, also eine Reihe, wird Instanz, und eine Spalte Attribut oder Indikator genannt.

Country Name	Year	Access to electricity (% of population)	Acces to non-solid fuel (% of population)	…	Slum Population (% of urban population)	…
Algeria	1990	94,03	85,80		11,8	
Bangladesh	1990	21,62	6,78		87,3	
Bangladesh	1995	-	-		84,7	
Bangladesh	2000	32,00	11,30		77,8	

Abb. 1: Ausschnitt aus dem Datensatz (dieser ist teilweise unvollständig)

Diese Datensätze werden mithilfe eines Skripts aus dem Ursprünglichen zusammengeschnitten, was spätere Modifikationen erleichtert. So können Attribute angepasst, oder neue hinzugefügt werden, wie zum Beispiel die Veränderung der Slum-Einwohnerzahl zwischen zwei gemessenen Werten.

Es gibt sehr viele Missing Values, nur die Hälfte aller Attribute hat Angaben für mindestens 50% aller Instanzen. Das muss bei der Auswahl der verwendeten Algorithmen in Betracht gezogen werden. Einfache statistische Berechnungen wurden auf den Tabellen mit Excel durchgeführt, für spezielle Algorithmen wurde das WEKA[3] Framework verwendet.

[3] http://www.cs.waikato.ac.nz/ml/weka/

2 Grundlagen des maschinellen Lernens

Methoden und Algorithmen aus dem Bereich des maschinellen Lernens haben das Ziel, aus Wissen Erfahrung zu generieren. Erfahrung steht dabei für strukturierte Daten in der Form eines Datensatzes. Darin sollen Gesetzmäßigkeiten erkannt werden, die auch auf ungesehene Instanzen zutreffen.

2.1 Klassifizierung

Hauptsächlich werden Klassifizierungsalgorithmen verwendet. Diese Algorithmen beobachten, wie sich ein bestimmtes Attribut, genannt Zielattribut, in Abhängigkeit von den restlichen Attributen verhält. Auf dieser Grundlage wird ein Modell gebaut, mit dessen Hilfe man auch bisher ungesehenen Instanzen einen Wert für das Zielattribut zuweisen kann. Diesen Vorgang nennt man Klassifikation.

Diese Arbeit benutzt die Standardimplementierungen verbreiteter Algorithmen im WEKA[4] Framework. Diese Algorithmen sind darauf ausgelegt, für den Klassifizierungsvorgang optimale Modelle zu finden. Diese Modelle werden anschließend bewertet und interpretiert.

Die Algorithmen unterscheiden sich einerseits in der Herangehensweise, wie die Zusammenhänge zwischen den einzelnen Attributen erkannt werden, andererseits in der Art der Modelle, die sie als Ergebnis liefern. In diesem Kapitel stelle ich drei Algorithmen vor, die in dieser Arbeit verwendet werden. (Hinweis: die verwendeten Beispiele sind zu Demonstrationszwecken abgeändert.)

2.1.1 JRip[5]

JRip erzeugt ein Modell in Form einer Regelmenge. Für die Klassifizierung einer Instanz muss man von oben nach unten überprüfen, ob die Bedingungen des IF-Zweiges einer Regel erfüllt sind. Wenn das der Fall ist, wird das Zielattribut auf den Wert gesetzt, wie in THEN beschrieben. Ansonsten betrachtet man den IF-Zweig der nächsten Regel. Hat man bis zum Schluss keinen IF-Zweig betreten, tritt der ELSE-Fall ein.

IF (Lifetime risk of maternal death ≥ 3.65%)
    THEN Slum Development = higher than average (34/5)

IF (Crop production index ≥ 88.95) & (Access to electricity ≤ 21.84%)
    THEN Slum Development = higher than average (7/1)

ELSE Slum Development = lower than average (47/3)

Abb. 2: Beispiel einer Regelmenge

Country, Year	Lifetime risk of maternal death	Crop Production Index	Access to electricity	…	Slum Development	Regel
Afghanistan, 2005	9,53%	70,12%	33,01%	…	higher than avg.	1. IF
India, 1990	3,12%	92,15%	15,08%	…	higher than avg.	2. IF
Benin, 2000	1,97%	91,08%	24,91%	…	lower than avg.	ELSE

Abb. 2: Beispielhafte Klassifizierung dreier Beispiele anhand der Regelmenge aus 2.1.1

[4] http://www.cs.waikato.ac.nz/ml/weka/
[5] Die Weka-Implementierung ist angelehnt an [FERI]

In Abb. 2 ist das Zielattribut „Slum Development“, welches entweder „higher than average“ oder „lower than average“ ist. „Lifetime risk of maternal death “, „Crop production index“ und „Access to electricity“ sind weitere Attribute des Datensatzes, die für die Klassifizierung benötigt werden.

Die Zahlenwerte der Form (XX/YY) nach der Zuweisung zum Zielattribut bewerten die jeweilige Regel und werden im Kapitel 2.1.4 unter Heuristiken genauer erklärt.

2.1.2 J48[6]

Der zweite Algorithmus erzeugt als Modell einen Entscheidungsbaum. Das Zielattribut ist erneut Slum Development und kann dieselben Werte annehmen, wie im vorherigen Beispiel. Zur Klassifizierung einer Instanz mit diesem Modell geht man den Baum von oben nach unten durch, wobei immer der Attributwert der zu klassifizierenden Instanz entscheidet, welchen Pfad man wählt. Dies wird so lange wiederholt, bis man an einem Blatt-Knoten angekommen ist.

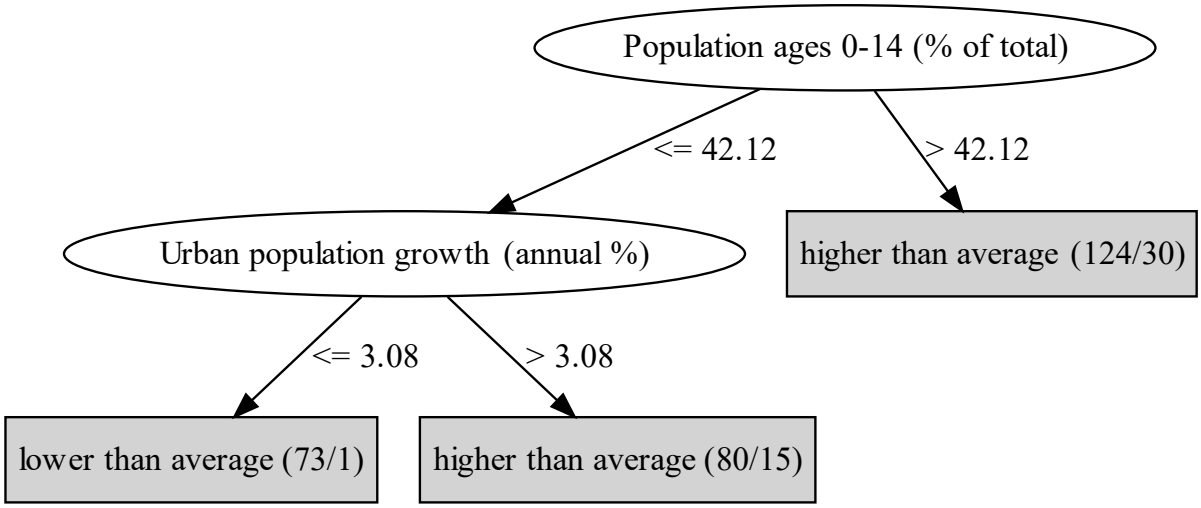

Abb. 3: Beispielhafter Entscheidungsbaum

Die Faustregel für die Interpretation dieses Modells lautet: je weiter oben ein Attribut, desto größer sein Einfluss auf das Zielattribut. Dies ist der Art und Weise geschuldet, wie der Algorithmus den Baum aufbaut. Beginnend bei der Wurzel wird für jeden Knoten das Attribut ausgewählt (i.d.F. Population ages 0-14), welche die Menge aller Instanzen an einem beliebigen Scheidepunkt (i.d.F. 42,12) am saubersten in zwei Teilmengen aufteilt, sodass die Instanzen der Teilmengen möglichst dasselbe Zielattribut haben.

2.1.3 Random Forest[7]

Der dritte Algorithmus erstellt mehrere Entscheidungsbäume und lässt diese über das Zielattribut abstimmen. Damit sich die Bäume unterscheiden, verwenden sie jeweils unterschiedliche Teilmengen der Attribute und trainieren auf einer Teilmenge aller Instanzen, haben aber stets dasselbe Zielattribut. Der zu klassifizierenden Instanz wird schließlich die Klasse zugewiesen, die die Mehrheit der Entscheidungsbäume vorhergesagt hat.

Da in der Arbeit Random Forests mit 100 Entscheidungsbäumen benutzt werden, werden die einzelnen Bäume nicht abgebildet oder interpretiert. Allerdings eignen sich Random Forests gut für die Bewertung der Attribute, da man beobachten kann, wie oft einzelne Attribute für den

[6] Ausführliche Beschreibung in [PML]
[7] Ausführliche Beschreibung in [RanF

Aufbau der Bäume verwendet wurden, und wie effektiv diese im jeweiligen Baum zur Klassifizierung beigetragen haben. Dazu mehr im Kapitel „2.2 Bewertung der Indikatoren“.

Dieser Algorithmus liefert von den drei Vorgestellten in den meisten Fällen die Modelle mit der höchsten Vorhersagegenauigkeit.

2.1.4 Bewertung

Das Ziel des maschinellen Lernens ist es, übertragbares Wissen zu generieren, welches auch auf ungesehene Beispiele gut Anwendbar ist. Ein wichtiges Kriterium für die Bewertung eines Modells ist es deshalb, inwiefern es auch neue, bis dahin ungesehene Instanzen korrekt klassifiziert (die Annahme ist natürlich, dass die ungesehenen Instanzen dieselbe Datenstruktur haben und denselben Sachverhalt abbilden). Um das zu bewerten teilt man die verfügbare Menge an Instanzen in eine Lern- und eine Testmenge auf: Die Lernmenge wird benutzt, um ein Modell zu trainieren, und anschließend wird dieses Modell auf der Testmenge bewertet.

Unter Heuristiken wird erklärt, wie diese Bewertung erfolgt und welche Form ihre Ergebnisse haben. Anschließend werden zwei Auswertungsmethoden vorgestellt, um die Instanzmenge effizient in Lern- und Testmenge aufzuteilen und dabei möglichst genaue Angaben über die voraussichtliche Anwendbarkeit des Modells auf neue Instanzen machen.

Heuristiken

Um ein Modell zu bewerten, überprüft man für jede Instanz der Testmenge, ob das Modell die Klasse korrekt vorhergesagt hat - die tatsächlichen Klassen sind uns bereits bekannt. Der prozentuale Anteil der korrekten Vorhersagen beschreibt die Vorhersagegenauigkeit eines Modells und ist der wichtigste Vergleichswert. Nach der Vorhersagegenauigkeit wird in Klammern die Anzahl korrekt klassifizierter Instanzen angegeben.

Ferner ist für jede Regel (im Falle von JRip) bzw. für jeden Blattknoten (im Falle von J48) ein Zahlenwert der Form (XX/YY) angegeben. Dabei steht XX für die Anzahl Instanzen der Trainingsmenge, bei denen diese Regel angewendet wurde, und YY für die Anzahl an falscher Klassifizierung durch diese Regel. Dadurch kann man die Abdeckung und Vorhersagegenauigkeit einzelner Regeln errechnen.

	Vorhersage: Klasse A	Vorhersage: Klasse B
Ist: Klasse A	97	17
Ist: Klasse B	37	83

Abb. 4: Beispiel Konfusionsmatrix

Die Konfusionsmatrix beschreibt, wie oft Verwechslungen zwischen den einzelnen Klassen vorkamen, und welche Verwechslungen das waren. Abb. 4 hat 234 (=97+37+17+83) Instanzen, wovon 114 (=97+17) die Klasse A besitzen. Davon wurde jedoch nur 97 korrekt als Klasse A klassifiziert, die restlichen 17 als Klasse B. An Abb. 4 kann man z.B. erkennen, dass Instanzen der Klasse B verhältnismäßig öfter als Klasse A klassifiziert wurden, als andersrum. Aus einer Konfusionsmatrix kann man außerdem die Vorhersagegenauigkeit errechnen:

$$Vorhersagegenauigkeit = \frac{Anzahl\ korrekter\ Vorhersagen}{Anzahl\ Vorhersagen} = \frac{(97 + 83)}{234} = 76{,}92\%$$

Es gibt noch weitere Heuristiken, die in den umfassenden Ergebnissen vorkommen, für die Interpretation reichen uns diese beide Maße allerdings aus.

Auswertungsmethoden

Damit die angegebene Genauigkeit des Modells möglichst wenig von der Aufteilung der Instanzmenge in Lern- und Testmenge abhängt und die bereits klassifizierten Instanzen dennoch effizient genutzt werden, verwenden wir besondere Auswertungsmethoden.

Am weitesten verbreitet ist „n-fold cross-validation“[8]. Bei dieser Methode werden die Instanzen in n – meistens, auch in dieser Arbeit, 10 - gleichgroße Sets geteilt. Alle n Sets werden einmal als Testmenge verwendet, um das Modell zu testen, welches auf den restlichen n-1 Sets trainiert wurde. Das heißt, bei einer 10-cross fold-validation wird jede Instanz 9x verwendet, um ein Modell zu bauen. Alle 10 Modelle werden unabhängig voneinander ausgewertet. Da jede Instanz genau einmal als Teil einer Testmenge klassifiziert wurde, kann man die Ergebnisse der einzelnen Auswertungen addieren. Je größer der Parameter n gewählt wird, desto unabhängiger ist das Gesamtergebnis von den einzelnen Aufteilungen in Lern- und Testset.

Eine Version von cross-validation, bei der das n maximal groß gewält wird (d.h. n = Anzahl Instanzen) heißt „leave-one-out“[9]. Dabei wird, analog zur Funktionsweise von n-fold cross-validation, für jede Instanz aus allen anderen Instanzen als Trainingsmenge ein Modell gebaut, um mit dieser Instanz als Testmenge das Modell zu bewerten. In der Arbeit wird eine leicht abgewandelte Methode verwendet, genannt Leave-One-Country-Out-Validation, bei der für jedes Land ein Set erstellt wird. Der Hintergedanke dieser Methode ist, dass Instanzen aus demselben Land zu unterschiedlichen Zeitpunkten sich teilweise nur minimal unterscheiden. Wenn auf einige dieser Instanzen trainiert wird, kann man bei einer Instanz desselben Landes in der Testmenge nicht von einem ungesehenen Beispiel sprechen. Um die Algorithmen mit dieser Auswertungsmethode bewerten zu können, wurde eine Funktion des WEKA-APIs entsprechend angepasst.

Die Bewertung eines Modells beschreibt also die Lernbarkeit eines Datensatzes von dem gegebenen Algorithmus, und ist ein Maßstab dafür, wie das letztendliche Modell neue, ungesehene Instanzen klassifizieren kann. Nach der Bewertung wird ein Modell auf der gesamten Instanzmenge gelernt; dieses wird zum Schluss ausgegeben.

2.1.5 Pruning

Bei der Erstellung des Modells dürfen die Regeln nicht zu spezifisch werden, da sie sich sonst sehr an den Trainings-Datensatz anpassen. Als Pruning bezeichnet man verschiedene Methoden, die solch eine Anpassung vermeiden sollen. Die 3 vorgestellten Algorithmen verwenden deshalb Pruning, um das Modell möglichst generell zu halten. Man unterscheidet zwischen Pre- und Post-Pruning.

Beim Pre-Pruning werden während des Aufbaus der Regelmenge oder des Baums Kriterien festgelegt, die sehr spezifische Regeln bzw. Knoten verhindern sollen. So kann man sagen, eine Regel oder ein Blattknoten muss mindestens x Instanzen abdecken. Beim Post-Pruning wird im Nachhinein überprüft, ob Regeln oder Knoten relevante Verbesserungen der Vorhersagegenauigkeit zur Folge haben. Wenn nicht, werden sie weggekürzt.

[8] Ausführliche Beschreibung in [DatM]
[9] Ausführliche Beschreibung in [DatM]

2.2 Attribut-Evaluation

Außer der stichpunktartigen Interpretation der Klassifizierungsmodelle wird eine weitere Art von Machine Learning-Algorithmen verwendet, um eine Korrelation zwischen dem Zielattribut und anderen Attributen festzustellen. Dabei werden alle Attribute nach bestimmten Kriterien bewertet. In diesem Abschnitt werden drei solcher Kriterien vorgestellt, die in der Arbeit auch Verwendung finden.

2.2.1 Gain Ratio

Bewertet ein Attribut anhand der maximalen Verringerung der Entropie, der mit diesem Attribut möglich ist. Entropie beschreibt in der Informationstheorie die Ungeordnetheit einer Instanzmenge bezüglich eines Attributs. In den folgenden Beispielen haben wir jeweils 7 Instanzen mit den beiden verschiedenen Klassen A oder B.

- 7 Instanzen, 4x Klasse A, 3x Klasse B => hohe Entropie
- 7 Instanzen, 6x Klasse A, 1x Klasse B => niedrigere Entropie

$$GainRatio(Klasse, Attribut) = \frac{E(Klasse) - E(Klasse|Attribut)}{I(Attribut)}$$

Die Formel zeigt, wie die GainRatio eines Attributs bezüglich einer Klasse (Zielattribut) berechnet wird. $E(Klasse)$ ist dabei die Entropie der ursprünglichen Instanzmenge bzgl. der Klasse. $E(Klasse|Attribut)$ ist die Summe der Entropiewerte der Teilmengen, in denen das Attribut die Instanzmenge teilt. $I(Attribut)$ gibt die Anzahl dieser Teilmengen an, und soll vermeiden, dass sehr spezifische oder einzigartige Attribute gut bewertet werden. Diese teilen die Instanzmenge zwar in reine Teilmengen auf, sind aber für die Klassifizierung neuer Instanzen nicht geeignet.

Schauen wir uns beispielhaft zwei Aufteilungen einer Instanzmenge mit 7 Instanzen, davon 4x A und 3x B, durch die Attribute X und Y an.

1. 1. Teilmenge: 4 Instanzen mit $x_i \leq X$: 4x A, 0x B
   2. Teilmenge: 3 Instanzen mit $x_i > X$: 0x A, 3x B
   => Entropie wurde stark verringert. Das Attribut hat eine hohe Gain Ratio.
2. 1. Teilmenge: 5 Instanzen mit $y_i \leq Y$: 3x A, 2x B
   2, Teilmenge: 2 Instanzen mit $y_i > Y$: 1x A, 1x B
   => Entropie wurde nicht verringert. Das Attribut hat eine niedrige Gain Ratio.

2.2.2 ReliefF

ReliefF bewertet Attribute gut, die folgende Eigenschaften haben:
- Ähnliche Instanzen mit derselben Klasse haben ähnliche Attributwerte
- Ähnliche Instanzen mit unterschiedlichen Klassen haben (stark) unterschiedliche Attributwerte

Dabei wird die Entfernung zwischen zwei Instanzen durch die Summe der normalisierten Differenzen zwischen den einzelnen Attributwerten berechnet. Das heißt ähnliche (=nahe) Instanzen haben viele Attribute mit ähnlichen Attributwerten.

1. Setze alle Attributwerte auf $w_A = 0$
2. Für $i = 0$ bis n (Anzahl Instanzen)
   a. wähle Instanz $\#n$
   b. finde
      i. h: die nächste Instanz mit derselben Klasse (near hit)
      ii. m: die nächste Instanz mit einer anderen Klasse (near miss)

c. für jedes Attribut A

$$w_A = w_A + \frac{1}{n} \times \left(d_A(m,x) - d_A(h,x)\right)$$

$d_A(x,y)$ ist der Abstand im Attribut A zwischen x und y (normalisiert auf [0,1])[10]

2.2.3 Random Forest Evaluation

Wie schon in Kapitel 2.1.3 beschrieben, erstellt der Random Forest Algorithmus mehrere Entscheidungsbäume, die jeweils eine Teilmenge aller Attribute zur Verfügung haben. Die Aufteilung erfolgt so, dass alle Attribute von gleich vielen Bäumen verwendet werden können. Die Bäume wählen ausgehend von der Wurzel für jeden Knoten jeweils das Attribut, welches die größte Gain Ratio (siehe 2.2.1) für die in dem Teilbaum vorkommenden Instanzen hat. Die Auflistung, wie oft ein Attribut von allen Entscheidungsbäumen insgesamt verwendet wurde, ist deswegen ein guter Indikator dafür, ob ein Attribut auf verschiedenen Teilmengen aller Instanzen mit dem Zielattribut korreliert.

Eine Liste der Random Forest Evaluation beschreibt mit der ersten Zahl, wie oft das Attribut in den einzelnen Bäumen des Waldes verwendet wurde, und mit der zweiten Zahl, wie stark die durchschnittliche Verringerung der Entropie bei der Verwendung des Attributes war.

[10] Algorithmus entnommen aus: http://www.ke.tu-darmstadt.de/lehre/ws-16-17/mldm/dt.pdf

3. Prognose der Slum-Entwicklung

Das Hauptziel dieser Arbeit ist es, zukünftige Slum-Entwicklungen vorhersagen zu können. Wenn wir davon ausgehen, dass die Bedingungen für die Slum-Entwicklung konstant sind, also sich mit der Zeit nicht ändern, gelten unsere Modelle, die die Slum-Entwicklung der letzten Jahre korrekt vorhersagen, auch für die Gegenwart. Diese Annahme reduziert das Problem auf die Erstellung solcher Modelle, die die vergangene Slum-Entwicklung beschreiben.

Zuvor gilt es in einem ersten Schritt festzustellen, inwiefern die Angaben zur gegenwärtigen Slum-Bevölkerung in einem kausalen Zusammenhang mit den restlichen Attributen stehen bzw. inwiefern die von uns verwendeten Algorithmen in der Lage sind, diese Zusammenhänge zu erkennen. Wenn das möglich ist, können wir auch anspruchsvollere Untersuchungen durchführen.

3.1 Klassifizierung der aktuellen Slum-Bevölkerung

Das Attribut „Slum Population (% of Urban Population)“ wurde in 4 Klassen aufgeteilt um das Zielattribut zu bilden:

- 1: 0% - 10% der Stadtbevölkerung lebt in Slums (189 Instanzen)
- 2: 10% - 40% der Stadtbevölkerung lebt in Slums (193 Instanzen)
- 3: 40% - 80% der Stadtbevölkerung lebt in Slums (182 Instanzen)
- 4: 80% - 100% der Stadtbevölkerung lebt in Slums (49 Instanzen)

Um eine gleichmäßige Verteilung der Klassen zu erreichen, wurden 178 Instanzen der Klasse 1 hinzugefügt, bei denen davon ausgegangen werden kann, dass die Slum-Bevölkerung 10% der Stadt-Bevölkerung nicht übersteigt. Dazu zählen diverse europäische und andere Industriestaaten. So besteht dieser Datensatz aus insgesamt 613 Instanzen.

Klassifikationsalgorithmus	Cross-Validation	Leave-One-Country-Out
JRip	81,40% (499)	76,02% (466)
J48	82,87% (508)	72,43% (444)
Random Forest	87,11% (534)	80,26% (492)

Abb. 5: Vorhersagegenauigkeit der verschiedenen Modelle aus Untersuchung 3.1.

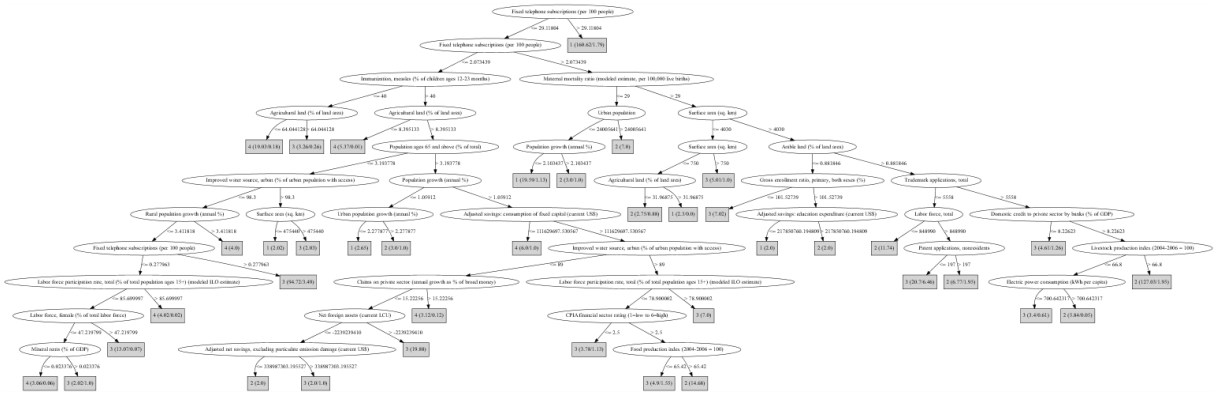

Abb. 6: Entscheidungsbaum (von J48 erstellt) zur Untersuchung 3.1
Diese Abbildung dient nur der Darstellung der Komplexität

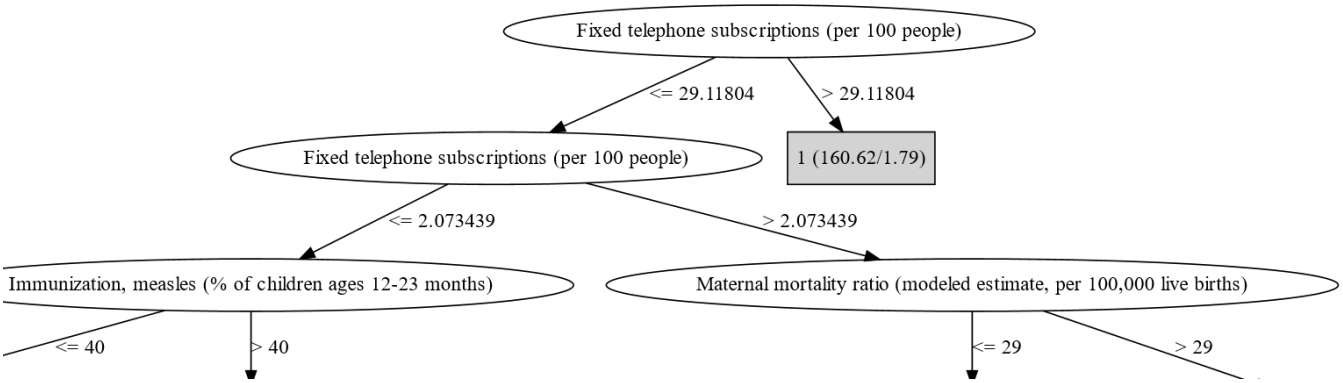

Abb. 7: Oberer Ausschnitt des Entscheidungsbaumes aus Abb. 6

Auswertung

Eine fundierte Vorhersage der Klassen anhand der Attribute ist mithilfe der verwendeten Algorithmen möglich. Eine Genauigkeit von 80% (siehe Abb.5) heißt, dass die Slum-Bevölkerungsklasse in 4 von 5 Fällen korrekt vorhergesagt wurde. Das dafür verwendete Modell ist allerdings ziemlich komplex (siehe Abb. 6).

Die für die Klassifizierung verwendeten Attribute stehen in verschiedenen Zusammenhängen mit dem Zielattribut. Schauen wir uns den oberen Ausschnitt des Baums (Abb. 7) an:

- „Fixed Telephone Subscriptions“: Slums sind per Definition Orte mit unzureichender Infrastruktur, sodass die starke Korrelation nicht überraschend ist. Direkt nach dem Wurzelknoten wird 162 Instanzen mit jeweils mehr als 29 Festnetzanbindungen die Klasse 1 zugewiesen – damit wurden 160 von 189 Instanzen der Klasse 1, also 85%, korrekt klassifiziert. Dennoch wäre es unsinnig zu sagen, dass die Anzahl Festnetzleitungen der Grund für eine bestimmte Slum-Größe ist.
- „Maternal Mortality Ratio“: Dieses Attribut steht nicht per Definition mit Slums im Zusammenhang, ist aber vermutlich eine Folge der schlechten gesundheitlichen Versorgung in Slums. Bei Knoten, die weiter unten im Baum sind, darf muss man auch den zurückgelegten Weg beachten. Falls die mütterliche Sterblichkeitsrate kleiner gleich 29 ist, werden die Instanzen in die Kategorien 1 oder 2 eingeteilt – aber nur, wenn es $2{,}07 < x < 29{,}11$ Festnetzleitungen pro 100 Einwohner gibt. Der rechte Teilbaum hat Blattknoten der Kategorien 2-4.

In Abb. 8 kann man erkennen, dass ein großer Anteil an Klasse 4 Instanzen der Klasse 3 zugeordnet wird. Dies kann entweder daran liegen, dass diese Klasse weniger Instanzen hat, und somit Regeln, die diese Klasse benachteiligen, trotzdem gut bewertet werden können. Oder es ist tatsächlich schwieriger anhand der Indikatoren die Entscheidung zwischen den Klassen 3 und 4 zu treffen.

183	3	3	0
5	168	20	0
2	18	161	1
0	0	27	22

Abb. 8: Konfusionsmatrix der 10-fold Cross-Validation des Random Forests in Untersuchung 3.1 (Werte repräsentativ für alle 6 Konfusionsmatrizen)

3.2 Prognose der zukünftigen Slum-Entwicklung

Nun kommen wir zu dem wichtigsten Teil dieser Arbeit, der Vorhersage der zukünftigen Slum-Bevölkerung aufgrund statischer Daten. Dafür müssen wir, anstatt die Slum-Bevölkerung als prozentualen Anteil der Stadtbevölkerung anzugeben, die absolute Einwohnerzahl dieser Slums berechnen.

$$Total\ Slum\ Population = Slum\ Population\ (\%\ of\ Urban\ Population) \times Urban\ Population$$

Wenn wir diesen Wert für dasselbe Land für zwei Zeitpunkte „Now“ und „Later“ haben, können wir die Entwicklung der Einwohnerzahl zwischen den beiden Zeitpunkten angeben.

$$Slum\ Developtment = \frac{(TotalSlumPopLater - TotalSlumPopNow)}{TotalSlumPopNow}$$

Die Anzahl verwendbarer Instanzen wird dadurch stark verringert, da wir für eine Instanz zwei Angaben zur Slum-Bevölkerung aus demselben Land mit einem festen Abstand brauchen. Haben wir alle Slum-Entwicklungen für einen gegebenen Zeitschritt berechnet, können wir für jede Instanz entscheiden, ob dessen Entwicklung über oder unter dem Durchschnitt liegt. Diese binäre Angabe wird unser Zielattribut für die nächsten Untersuchungen. Die Klassen so zu definieren hat einige Vorteile:

1. Die Klassen sind relativ ausgewogen, d.h. beide haben circa gleich viele Instanzen.
2. Man hat keine eigenhändig festgelegten Grenzwerte.
3. Man erhält Attribute, die im Wesentlichen das Slum-Wachstum beeinflussen.

Angenommen wir haben als Beispiel die Instanz Afghanistan im Jahr 2000. Dann nehmen wir für die Vorhersage alle Indikatoren die im Jahr 2000 gemessen wurden, und berechnen zusätzlich die Slum-Entwicklung von 2000 bis 2010 (wenn der Zeitschritt 10 Jahre beträgt).

Ergebnisse

Zeitschritt	5 Jahre		10 Jahre		20 Jahre	
Anzahl Instanzen	244		210		88	
Ø Wachstum	12,90%		29,98%		77,68%	
Instanzen mit Wachstum geringer als Ø	134 (54,92%)		119 (56,67%)		50 (56,82%)	
Auswertungsmethode	Cross-Validation	L.O.C.O.	Cross-Validation	L.O.C.O.	Cross-Validation	L.O.C.O.
JRip	73,77% (180)	68,03% (166)	77,62% (163)	69,52% (146)	70,45% (62)	76,14% (67)
J48	73,77% (180)	70,49% (172)	75,24% (158)	69,52% (146)	80,68% (71)	72,73% (64)
Random Forest	83,61% (204)	78,28% (191)	81,90% (172)	79,05% (166)	81,82% (72)	79,55% (70)

Abb. 9: Auswertung der Modelle der Untersuchung 3.2

Im Vergleich zu der Untersuchung 3.1 erkennen wir, dass die Vorhersagegenauigkeit bei allen drei Zeitschritten geringer ist, obwohl hier nur zwei statt vier Klassen voneinander unterschieden werden müssen. Dies weist auf eine komplexere Aufgabenstellung hin. Dennoch sind die Modelle eindeutig besser als zufälliges Raten – das hätte bloß ca. 55% Vorhersagegenauigkeit, da die größte Klasse 55% aller Instanzen hat und man immer diese Klasse wählen würde.

IF (Maternal mortality ratio (modeled estimate, per 100,000 live births) ≥ 544) AND
(Surface area (sq. km) ≤ 587000) AND
(Women's share of population ages 15+ living with HIV (%) ≤ 59.26)
THEN Relative Slum Development = higher than average (48 /2)

IF (Population growth (annual %) ≥ 2.26) AND
(Rural population (% of total population) ≥ 63.18) AND
(Import value index (2000 = 100) ≤ 109.91)
THEN Relative Slum Development = higher than average (29/1)

IF (Lifetime risk of maternal death (%) ≥ 1.34) AND
(Final consumption expenditure, etc. (% of GDP) ≥ 94.49) AND
(Agriculture, value added (annual % growth) ≥ -0.27)
THEN Relative Slum Development = higher than average (12/1)

IF (Population ages 65 and above (% of total) ≤ 3.52) AND
(Adjusted net enrolment rate, primary, both sexes (%) ≥ 76.36)
THEN Relative Slum Development = higher than average (17/5)

IF (Fertility rate, total (births per woman) ≥ 5.9) AND
(Refugee population by country or territory of asylum ≤ 11000)
THEN Relative Slum Development = higher than average (8/1)

ELSE Relative Slum Development = lower than average (130/6)

Abb. 10: Regelmenge für Zeitschritt 5 Jahre

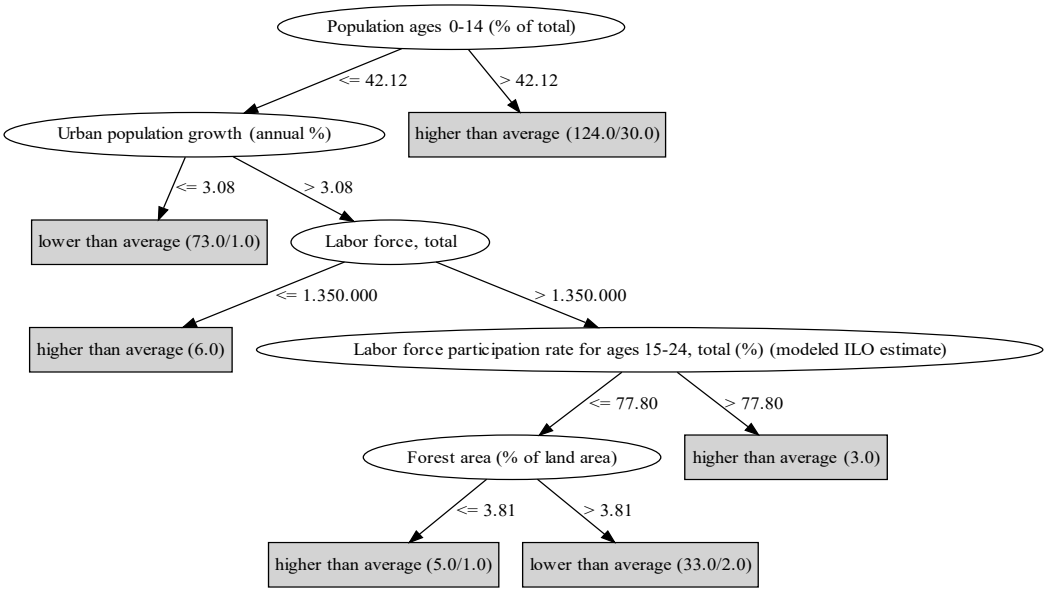

Abb. 11: Entscheidungsbaum für Zeitschritt 5 Jahre

	JRip		J48	
	Anzahl Regeln	Ø Bedingung. / Regel	Anzahl Knoten	Max. Tiefe
5 Jahres-Modell	5	2,6	5	5
10 Jahres-Modell	1	1	4	4
20 Jahres-Modell	2	1	2	2

Abb. 12: Kennwerte der verschiedenen Modelle

Die erzeugten Modelle (Abb. 10, 11, 13 - 16) sind wesentlich kleiner als die der vorherigen Untersuchung, was einerseits an der verringerten Klassenanzahl, andererseits am Akzeptieren relativ ungenauer Regeln liegt. Zum Beispiel deckt das rechte Blatt des ersten Knotens in Abb. 11 mehr als die Hälfte aller Instanzen, klassifiziert jedoch fast 25% falsch.

IF (Fertility rate, total (births per woman) ≥ 5.15)
THEN Relative Slum Development = higher than average (97/24)

ELSE Relative Slum Development = lower than average (113/18)

Abb. 13: Regelmenge für Zeitschritt 10 Jahre

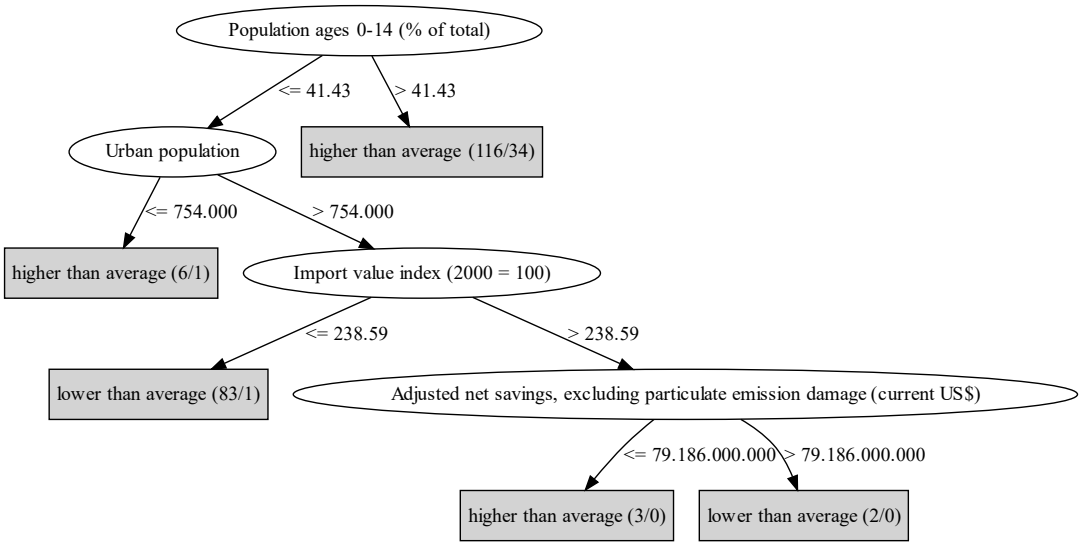

Abb. 14: Entscheidungsbaum für Zeitschritt 10 Jahre

IF (Lifetime risk of maternal death (%) ≥ 3.65)
THEN Relative Slum Development = higher than average (34/5)

IF (Crop production index (2004-2006 = 100) ≥ 88.95)
THEN Relative Slum Development = higher than average (7/1)

ELSE Relative Slum Development = lower than average (47/3)

Abb. 15: Regelmenge für Zeitschritt 20 Jahre

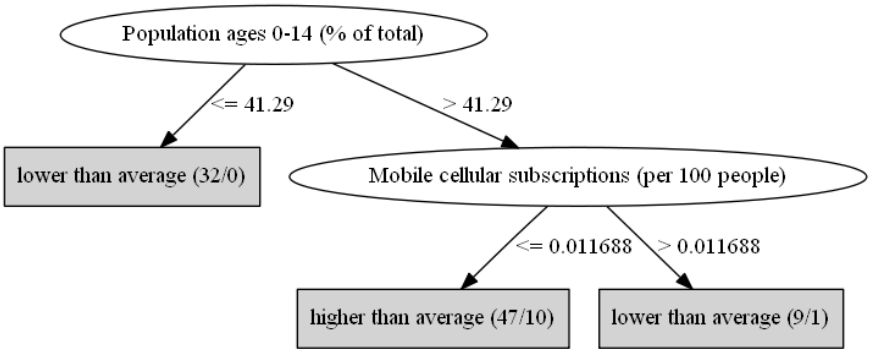

Abb. 16: Entscheidungsbaum für Zeitschritt 20 Jahre

Auswertung

Vergleichen wir nun die sechs Modelle (Abb. 10, 11, 13 - 16) und deren Auswertungen (Abb. 9) miteinander, erkennen wir, dass die Vorhersagegenauigkeit sich nur minimal unterscheidet. Es ist genauso gut möglich, zu bestimmen, ob ein Slum in 5 Jahren überdurchschnittlich wächst, oder ob es in 20 Jahren überdurchschnittlich wächst. Jedoch werden die Modelle zunehmend einfacher (Abb. 12); während man bei der Vorhersage für die nächsten 5 Jahre also viele Fälle unterscheiden muss (mehrere lange Regeln, bzw. größere Anzahl Knoten), so bestimmen auf lange Sicht wenige Faktoren entscheidend über das Slum-Wachstum. Dies bedeutet wenige, kurze Regeln bzw. eine geringe Anzahl Knoten in den Bäumen (siehe Abb. 12).

Untersuchen wir die bestimmenden Attribute der Modelle, stellen wir starke Ähnlichkeiten fest. Die größte Gruppe bilden demographische Attribute, wie „Population ages 0-14“, „Urban population growth“, „Fertility Rate“, usw. Alle drei Bäume haben als Wurzelknoten das Attribut „Population ages 0-14“, und weisen anhand dieser ersten Abfrage direkt der einen Teilmenge eine Klasse zu. Die Regelmenge für 10 Jahre besteht sogar nur aus einer einzigen Regel mit dem Attribut „Fertility Rate“. Außerdem gibt es volks- und landwirtschaftliche Faktoren, wie „Crop Production Index“, „Import Value Index“ und „Final Consumption Expanditure“, die oft mit demographischen Attributen eine Regel bilden oder in durch sie bedingte Teilbäume auftreten. Außerdem treten gesundheitliche (z.B. „Lifetime risk of maternal Death“) auf.

Auffällig ist, dass das Attribut „Population living in Slums (% of Urban)“, welches in dieser Untersuchung normaler Teil der Datensätze ist, nicht in den Modellen auftaucht. Auch bei späteren Untersuchungen bezüglich der Korrelation mit dem Zielattribut stellen wir fest, dass die Entwicklung eines Slums relativ unabhängig von seiner Ausbreitung ist.

3.3 Prognose der zukünftigen Slum-Entwicklung mit zusätzlichen dynamischen Daten

Bisher haben wir die Attributwerte nur zu einem Zeitpunkt betrachtet, und versucht, mit diesen Werten eine Vorhersage zu treffen. In diesem Schritt wollen wir nicht nur den aktuellen Wert der Attribute, sondern auch ihre Entwicklung über die letzten Jahre in die Untersuchung einbeziehen. Diese Entwicklung können wir durch die Differenz aus dem aktuellen Wert und dem Wert aus einem vergangenen Jahr berechnen. Wir betrachten die Entwicklungen der Attribute in den vergangenen 5, 10 und 20 Jahren, und nennen sie Entwicklungsattribute. Die Zielvariable bleibt die selbe, wie in Untersuchung 3.2, allerdings betrachten wir nur die Slum-Entwicklung in den nächsten 5 Jahren, um das Ergebnis überschaubar zu halten. Zu dem bestehenden Datensatz aus 3.2 fügen wir nacheinander die Entwicklung der Attribute über jeweils 5, 10 und 20 Jahre hinzu.

Auswertungs-methoden	Klassifikations-methoden	Nur statische Indikatoren	+ Entwicklung 5 Jahre	+ Entwicklung 5 & 10 Jahre	+ Entwicklung 5,10,20 Jahre
Cross-Validation	JRip	73,77% (180)	76,23% (186)	74,18% (181)	72,95% (178)
	J48	73,77% (180)	72,54% (177)	72,13% (176)	70,08% (171)
	RandomForest	83,61% (204)	79,51% (194)	79,92% (195)	79,51% (194)
Leave-One-Country-Out	JRip	68,03% (166)	74,59% (182)	70,49% (172)	69,26% (169)
	J48	70,49% (172)	69,67% (170)	69,67% (170)	69,67% (170)
	RandomForest	78,28% (191)	78,28% (191)	75,82% (185)	76,64% (187)

Abb. 17: Auswertung der Untersuchung 3.3

Das Ergebnis ist überraschend: die Klassifikationsmodelle sind ungenauer geworden, obwohl wir zusätzliche Informationen zur Verfügung hatten. Das mag auf den ersten Blick irrational sein, aber ist durch den Aufbau der verwendeten Algorithmen zu erklären. Diese benutzen Pruning, wie in Kapitel 2.1.4 beschrieben, um möglichst allgemeingültige Aussagen treffen zu können.

Die endgültigen Modelle sind klein und enthalten sehr wenige (in den meisten Fällen keine) Entwicklungsattribute. Vermutlich wurden im ersten Schritt Modelle mit Entwicklungsattributen gebaut, die jedoch wegen schlechter Generalisierung beim anschließenden Pruning entfernt wurden. Der Verdacht, dass die verwendeten Entwicklungsattribute nicht so gut für diese Art der Vorhersage geeignet sind, wird durch die Liste der Attribute bestätigt, die bei der GainRatio- und Random-Forest-Evaluation die höchsten Werte erreichen:

In den Top10 GainRatio Attributen ist „Population ages 65 and above (% of total)“ jeweils mit der Entwicklung der letzten 5 bzw. 20 Jahre auf Platz 8 und 9. Bei Random Forest ist „Cereal Yield (kg per hectare) – 20 years development“ auf Platz 2 und „Age dependency ratio – 10 years development“ auf Platz 4. Die Entwicklungsattribute sind relativ schlecht repräsentiert, wenn man bedenkt, dass auf jedes Attribut drei Entwicklungsattribute kommen.

4. Korrelationen mit der Slum-Entwicklung

Während uns die Modelle aus Kapitel 3 plausible Erklärungen für das Slum-Wachstum liefern und beispielhaft die Möglichkeit geben, das Auftauchen bestimmter Attribute zu interpretieren, so liefern sie uns dennoch keine vollständige Übersicht über das Korrelationsverhältnis zwischen jedem einzelnen Attribut und dem Zielattribut. Aus diesem Grund verwenden wir die in Kapitel 2.2 vorgestellten Evaluations-Algorithmen, um die Attribute nach ihrem Einfluss auf bestimmte Fragestellungen zu bewerten.

4.1 Steigender Einfluss demographischer Attribute bei größeren Zeitschritten

Schon bei der Analyse der Ergebnisse der Untersuchung 3.2 ist der immense Einfluss, den demographische Attribute auf die Slum-Entwicklung haben, aufgefallen. Wir wollen nun die Attribute für die drei Zeitschritte miteinander vergleichen:

GainRatio der Top 5 Attribute für 5 Jahre:

1.	0.3095	Lifetime risk of maternal death (%)
2.	0.3088	Population ages 0-14 (% of total)
3.	0.3020	Population ages 15-64 (% of total)
4.	0.3020	Age dependency ratio (% of working-age population)
5.	0.2961	Birth rate, crude (per 1,000 people)

GainRatio der Top 5 Attribute für 10 Jahre:

1.	0.3037	Population ages 0-14 (% of total)
2.	0.2964	Birth rate, crude (per 1,000 people)
3.	0.2920	Life expectancy at birth, total (years)
4.	0.2899	Survival to age 65, female (% of cohort)
5.	0.2755	Lifetime risk of maternal death (%)

GainRatio der Top 5 Attribute für 20 Jahre:

1.	0.4336	Population ages 0-14 (% of total)
2.	0.4185	Birth rate, crude (per 1,000 people)
3.	0.4040	Age dependency ratio (% of working-age population)
4.	0.4040	Population ages 15-64 (% of total)
5.	0.3688	Maternal mortality ratio (modeled estimate, per 100,000 live births)

In allen drei Fällen sind die Top 5 Attribute fast ausschließlich Demographie-bezogen. Man erkennt, dass die Zusammensetzung der Bevölkerung aus verschiedenen Altersgruppen einen sehr großen Einfluss darauf hat, ob die Slum-Entwicklung über- oder unterdurchschnittlich ist. Auch, wenn uns die GainRatio-Werte keine Formel für die Art der Beeinflussung angeben, können wir anhand der Modelle aus Kapitel 3.2 erkennen, wie die Attribute die Slum-Entwicklung beeinflussen. So sehen wir, dass „Population ages 0-14“ über ca. 41% in allen drei Entscheidungsbäumen zu einer Klassifikation als „higher than average“ führen.

Zusätzlich steigt der GainRatio-Wert beim Sprung von 10 auf 20 Jahre um bis zu 50%, aber mindestens um 30% an. Dies erklärt, wieso unsere Modelle aus der Untersuchung 3.2 mit immer weniger Regeln/Knoten die Vorhersagegenauigkeit bei 80% halten konnten. Sichtbar wird diese Steigerung auch bei den Vorhersagegenauigkeiten der Wurzelknoten (Abb. 11, Abb. 14 & Abb. 16). Für die Zeitschritte 5 und 10 Jahre war die Klassifikation aufgrund von „Population ages 0-14“ in 75,81% bzw. 70,69% der Fälle erfolgreich. Bei 20 Jahren waren alle Klassifikationen anhand dieses Attributs korrekt – wobei der Grenzwert jedes Mal sehr ähnlich war.

4.2 Vergleich zwischen Slum-Entwicklung und urbaner Entwicklung

Bei der Menge an demographischen Attributen, die mit überdurchschnittlichem Slum-Wachstum korrelieren, kommt die Frage auf, inwiefern Slum-Wachstum mit dem allgemeinen Städte-Wachstum korreliert. Sogar die Modelle in Kapitel 3.1 benutzen demographische Attribute, um den momentanen Anteil der in Slums lebenden Bevölkerung (sehr erfolgreich) vorherzusagen!

GainRatio der Top 5 Attribute aus Untersuchung 3.1

1. 0.4700 Fixed telephone subscriptions (per 100 people)
2. 0.4257 Lifetime risk of maternal death (%)
3. 0.3994 Birth rate, crude (per 1,000 people)
4. 0.3985 Maternal mortality ratio (modeled estimate, per 100,000 live births)
5. 0.3937 Population ages 65 and above (% of total)

Zu Beginn stellen wir für die Instanzen, für denen wir Slum-Wachstumswerte berechnet haben, auch das relative Stadt-Wachstum dar (siehe Kapitel 3.2 Einleitung, die Stadt-Entwicklung wurde mit derselben Formel errechnet). Nun können wir beurteilen, ob die Stadtbevölkerung für einzelne Länder schneller oder langsamer wächst als der Durchschnitt.

Der Zeitschritt beträgt 5 Jahre. Das durchschnittliche Städte-Wachstum in den beobachteten Ländern liegt bei 18,63% (bei Slums: 12,90%), die durchschnittliche Abweichung 6,45% (bei Slums: 13,72%). Das heißt die Stadtbevölkerung entwickelt sich im Durchschnitt stärker und konstanter als die Slumbevölkerung. (Die 2 Fälle Irak und Zimbabwe 2000 -> 2005 abgezogen beträgt die Standardabweichung bei Slums 10,94% - Erklärung folgt im nächsten Kapitel).

	Slum-Dev. above average	Urban-Dev. below average
Urban-Dev. above average	34,26% (84)	11,07% (27)
Urban-Dev. below average	10,66% (26)	43,85% (107)

Abb. 18: Korrelationen zwischen Städte- und Slum-Wachstum

In Abb. 18 erkennen wir, dass Städte- und Slum-Wachstum sich bei fast 80% der betrachteten Fälle gleich entwickeln. Das wirft ein interessantes Licht auf unser Verständnis von Slum-Wachstum: Die Entwicklung von Slums isoliert zu betrachten von der Entwicklung der Städte und Länder, in denen sie sich befinden, ist nur bis zu einem gewissen Grad real. Mit den meisten Fällen von überdurchschnittlichem Slum-Wachstum geht auch ein überdurchschnittliches Wachstum der Stadtbevölkerung einher und umgekehrt.

4.3 Bewertung der Korrelationen nach anderen Kriterien

Außer GainRatio stehen uns auch andere Kriterien zur Bewertung von Attributen zur Verfügung. In diesem Kapitel untersuchen wir die Korrelation mit dem Zielattribut „Slum-Development above or below average within 5 years“ (siehe Kapitel 3.2) mit den Evaluations-Funktionen ReliefF und Random Forest-Evaluation (siehe Kapitel 2.2).

ReliefF-Evaluation

0.1603	International tourism, number of departures
0.1138	Improved sanitation facilities, urban (% of urban population with access)
0.0982	Fertility rate, total (births per woman)
0.0946	Trademark applications, total
0.0939	Women's share of population ages 15+ living with HIV (%)
0.0934	Birth rate, crude (per 1,000 people)
0.0924	Fossil fuel energy consumption (% of total)
0.0820	Contributing family workers, total (% of total employment)
0.0814	Patent applications, residents
0.0812	Patent applications, nonresidents

Bei der Analyse dieser Liste muss man sich nochmal bewusstmachen, welche Attribute ReliefF gut bewertet: Attribute, deren Attributwerte sich bei ähnlichen Instanzen mit derselben Klasse ähneln und bei ähnlichen Instanzen mit unterschiedlichen Klassen klar unterscheiden. Diese Attribute kommen in den Modellen aus Kapitel 3 nur selten vor, da eine gute Bewertung durch ReliefF nicht heißt, dass sie für diese Art des Modellaufbaus nützlich sind. Umso interessanter und vielfältiger ist die Liste. Sie sind nicht nur in die zwei Bereiche Demographie und Gesundheit einzuordnen, und bieten einige neue Ansätze für die Kausalität der Slum-Entwicklung.

Random-Forest-Evaluation

60	0.28	Age dependency ratio (% of working-age population)
58	0.33	Urban population (% of total)
57	0.34	Forest area (% of land area) – Platz 369 bei GainRatio
57	0.28	Birth rate, crude (per 1,000 people)
56	0.36	Arable land (% of land area)
54	0.29	Maternal mortality ratio (modeled estimate, per 100,000 live births)
53	0.33	Fertility rate, total (births per woman)
52	0.27	Labor force, total
51	0.27	Lifetime risk of maternal death (%)
49	0.28	Adolescent fertility rate (births per 1,000 women ages 15-19)

Die Random-Forest-Evaluation ist nicht so stark von dem Datensatz abhängig wie die anderen beiden Evaluationen, da die Attribute auf unterschiedliche Teilmengen der Instanzen bewertet werden. Die hier aufgeführten Instanzen bringen also bei diversen Teilmengen eine hohe GainRatio und können damit generell als gut mit dem Zielattribut korrelierend angesehen werden. Auch hier sehen wir wieder einiges aus den Bereichen Demographie und Gesundheit, aber auch die Beschaffenheit des Landes (Forest area und Arable Land) sowie absolute Werte wie Labor force, total. Forest area ist interessanter weise bei der GainRatio Tabelle nur auf Platz 369 mit einem Informationsgewinn von 0. Das bedeutet, auf den gesamten Datensatz kann man das Attribut nicht verwenden, um reine Teilmengen zu kriegen, jedoch auffällig oft auf Teilmengen aller Instanzen, die evtl. nach anderen Kriterien gesplittet wurden.

4.4 Einordnung der erstellten Modelle

In diesem Kapitel werden die Modelle dieser Bachelor-Arbeit anhand einer anderen Forschungsarbeit [SimMod] bewertet. Dort werden unter anderem sieben Faktoren genannt, die die Slum-Entwicklung maßgeblich beeinflussen. Zuerst folgt eine Auflistung dieser sieben Faktoren inklusive kurzer Beschreibung, danach die Analyse, ob bzw. inwiefern die besagten Faktoren modelliert werden.

1. Population Dynamics: Demographische Veränderungen in der Bevölkerung, wie z.B. ein starkes Bevölkerungswachstum, können urbane Probleme wesentlich verstärken.
2. Economic Growth: Wirtschaftliches Wachstum in Entwicklungsländern konzentriert sich oft auf einige wenige Großstädte. Der Unterschied zwischen der Wirtschaftsleistung der Großstädte und der des Landes beeinflusst die urbane Zuwanderung, welches wiederum potenzielles Slum-Wachstum verstärkt.
3. Housing Market Dynamics: Die Preisentwicklungen auf dem Immobilienmarkt, sowohl auf dem offiziellen wie auch auf dem inoffiziellen, beeinflussen die Erschwinglichkeit von Immobilien.
4. Informal Economy: Ein großer Teil der jungen Leute, die in urbanen Gebieten informeller Arbeit nachgehen, leben laut UN-Analysen in Slums. So kann eine stabile informelle Wirtschaft das Slum-Wachstum verstärken.
5. Local Topography: Auch die lokalen Gegebenheiten in und um den einzelnen Slums, wie z.B. Wasserversorgung oder unbesetztes Gebiet in der Nähe können das Slum-Wachstum fördern.
6. Street Pattern: Da es in Slums kaum offizielle Straßen und wenige Eigentumsrechte gibt, haben Laufwege der Bevölkerung und der lokale Aufbau wie z.B. Positionen von Märkten einen großen Einfluss auf Form und Richtung der Slum-Ausbreitung.
7. The Politics of Slums: Politische Entscheidungen in den Ländern und Städten, in denen sich Slums befinden, haben massiven Einfluss auf die Entwicklung dieser Slums.

Wichtig ist zu beachten, dass dieses Paper die Slum-Entwicklung aus einer anderen Perspektive betrachtet. Es geht darum, Slum-Simulationen anhand ihres Umfangs zu bewerten. Dafür wird unter anderem untersucht, wie viele der oben genannten sieben Punkte bei der Simulation berücksichtigt werden. Eine wesentliche Einschränkung dieser Arbeit ist, dass Indikatoren ausschließlich auf Länder-Ebene bewertet werden, und lokale Aspekte nicht berücksichtigt werden können. Deshalb kommen die Punkte 3, 5 und 6 kommen in dieser Arbeit nicht vor.

Auch Punkt 7 wird nicht explizit betrachtet, allerdings weist eine statistische Auffälligkeit im Datensatz auf die wichtige Rolle hin, die die politische Situation bei der Entwicklung von Slums spielt. In der Untersuchung 3.2 wird die Slum-Entwicklung innerhalb von 5 Jahren berechnet. Dabei weichen zwei Instanzen mit ihren Werten massiv vom Durchschnitt ab: Irak und Simbabwe hatten zwischen 2000 und 2005 jeweils einen Slum-Bevölkerungszuwachs von 260,66% bzw. 364,29%. Das sind die zwei einzigen Fälle, in denen innerhalb von 5 Jahren eine Entwicklung von mehr als 100% stattfand. Diese Entwicklung ist auf die politische Situation im Land zurückzuführen: Irak wurde 2003 von den Vereinigten Staaten und der „Koalition der Willigen“ angegriffen und anschließend belagert. In Simbabwe startete die Armee 2005 auf Befehl des Präsidenten Mugabe „Operation Murambatsvina“[11] mit dem Ziel, illegale Behausungen zu zerstören. Laut einem Bericht der UN[12] wurden dabei 700.000 Leute aus ihrem Zuhause vertrieben, das entspricht ungefähr dem absoluten Anstieg der Slum-Bevölkerung (Stand 2000: 140.000; Stand 2005: 790.000).

[11] https://en.wikipedia.org/wiki/Operation_Murambatsvina
[12] http://www.un.org/News/dh/infocus/zimbabwe/zimbabwe_rpt.pdf

Der erste Punkt hingegen spielt in den erstellten Modellen die entscheidende Rolle. Population dynamics ist in dem Paper beschrieben als Zusammenspiel von Geburtenrate, Sterblichkeitsrate und Migrationsquote. Gerade die Geburtenrate ist Teil vieler Modelle und hat eine sehr hohe GainRatio – in Untersuchung 4.1 findet sich die Geburtenrate in allen drei Fällen unter den Top 5. Es gibt auch weitere Attribute bezüglich der Bevölkerungsentwicklung mit einer hohen Gain-Ratio. Dazu gehören „Death Rate“, „Life Expectency at Birth“, „Fertility Rate“ und „Population Growth”.

Es gibt zwar keine Angaben über das Verhältnis der Wirtschaftsleistung in den Städten zu dem im gesamten Land (Punkt 2), dennoch finden wir einzelne wirtschaftliche Indikatoren, hauptsächlich aus der Landwirtschaft oder Angaben über die Import-/Export-Verhältnisse, die einen hohen Einfluss auf die Slum-Entwicklung haben. Dazu gehören „Cereal Yield” (Plätze 19, 13, 9)[13] und „Agriculture, value added” (Plätze 18, 38, 50) bzw. „Imports of goods and services“ (Plätze 20, 15, 32) und „Merchandise imports” (Plätze 15, 12, 14).

Informationen über die informelle Ökonomie (Punkt 4) erhalten wir in diesem Datensatz hauptsächlich durch „Labor Force Participation Rate“ (Plätze 39, 41, 7). Es gibt zwar das Attribut „Informal Employment“, aber dieses kommt nicht in den Modellen vor und wird von allen drei Evaluationsalgorithmen schlecht bewertet.

[13] Diese Angabe ist wie folgt zu lesen: Das Attribut hat bei den GainRatio-Listen bzgl. der über-/unterdurchschnittlichen Slum-Entwicklung diese Plätze für die 3 Zeitschritte (5, 10, 20 Jahre)

5. Ausblick

Die Arbeit hat einen interessanten Einblick in die Art und Weise gegeben, wie die Slum-Entwicklung eines Landes anhand bestimmter Attribute vorhergesagt werden kann. Zudem basieren die analysierten Modelle auf tatsächliche Angaben über die Slum-Entwicklung auf Landes-Ebene, und haben damit eine mathematisch fundierte, reale Grundlage.

Dass die Demographie starken Einfluss auf die Slum-Entwicklung hat, war schon bekannt. Aber dass mehr noch als Geburten- oder Sterberate der Anteil von Kindern in der Gesamtbevölkerung mit der Slum-Entwicklung korreliert, ist eine überraschende Erkenntnis. Die Zusammensetzung der Bevölkerung könnte ein wichtigerer Faktor sein, als bisher angenommen. Zudem wird diese Korrelation in allen Untersuchungen dieser Arbeit bestätigt.

Des Weiteren sind Slum- und Städtewachstum stark gekoppelt und können nur bedingt unabhängig voneinander betrachtet werden. Dieser Faktor wird bei Forschungsarbeiten, die versuchen, Slum-Wachstum „von innen“ zu simulieren, häufig vernachlässigt. In [SimMod] wird dieses Verhältnis zwischen Stadt und Slum innerhalb der Population Dynamics lediglich als Differenz zwischen „Immigration Rate“ und „Emigration Rate“ dargestellt.

Letztendlich gibt es noch einige Punkte, an denen zukünftige Untersuchungen anknüpfen können:

- Man könnte versuchen, anhand der Ergebnisse in Abb. 14 die Instanzen bezüglich ihrer Einteilung in die 4 Gruppen zu klassifizieren. Generell sind die Beispiele sehr interessant, in denen die Entwicklung der Slums von dem der Städte abweicht, da in diesen Fällen Slums eine eigene Entwicklungs-Dynamik haben. Zudem kann auch diese Untersuchung auf 10 bzw. 20 Jahre ausgedehnt werden.
- Das schlechte Abschneiden der Entwicklungsattribute aus Kapitel 3.3 könnte an der Berechnung oder der Darstellung liegen. Alternative Formeln zur Berechnung (wie das relative Wachstum) oder eine andere Darstellung, z.B. als Reihe von Werten, würden vielleicht neue Querverbindungen aufdecken.
- Die UN hat zusätzlich einen Datensatz explizit für Großstädte[14]. Das ermöglicht ähnliche Untersuchungen auf Städte-Ebene oder einen direkten Vergleich zwischen der Entwicklung einzelner Slums und ihren Ländern.

[14] https://unhabitat.org/urban-indicators-guidelines-monitoring-the-habitat-agenda-and-the-millennium-development-goals/

5.1 Prognosen zur Slum-Entwicklung

Ein ansehnliches Ergebnis dieser Arbeit ist die Möglichkeit, die in 3.2 analysierten Modelle auf aktuelle Daten anzuwenden. In Abb. 19 wurden die Länderdaten der 106 Länder auf dem Stand von 2015 klassifiziert, für die Angaben zur Slum-Bevölkerung gemacht worden sind – für 2016 und 2017 lagen leider noch nicht alle Daten vor. Dabei ist zu beachten, was die Aufgabenstellung der Klassifikation war: Es sollte entschieden werden, ob die Slum-Entwicklung über oder unter einem gegebenen Grenzwert ist.

Man erkennt, dass einige Modelle sehr einseitig klassifizieren. So prognostiziert das 10-Jahres-Modell von J48 sehr oft Slum-Wachstum von über 30% (in 84% aller Fälle), während das 20-Jahre-Modell von J48 ausschließlich unterdurchschnittliche Slum-Entwicklung (weniger als +77% in 20 Jahren) prognostiziert.

Das 20-Jährige Slum-Wachstum konnte nur auf Instanzen aus den Jahren 1990 und 1995 berechnet werden. Der Baum klassifiziert unter anderem Aufgrund des Attributs „Mobile Cellular Subscriptions (per 100 people)“. Diese Anzahl ist in 2015 wesentlich höher als in 1990/1995, als der Grenzwert festgelegt wurde. Deshalb wird für alle Instanzen eine Slum-Entwicklung von unter 77,68% prognostiziert. In diesem Fall ist die grundlegende Annahme falsch, die zu Beginn des 3. Kapitels getroffen wurde. Die Bedingungen für die Slum-Entwicklung sind nicht konstant, zumindest dieser Wert unterliegt einer starken Veränderung.

Dennoch, der Trend dieser Klassifikationen gibt Hoffnung. Es wird in sehr vielen Fällen eine Slum-Entwicklung prognostiziert, die unterhalb des jeweiligen Grenzwertes liegt. Dies bedeutet einen Rückgang der relativen Geschwindigkeit der Slum-Entwicklung – für alle drei betrachteten Zeitschritte. Auf den bisherigen Daten ließ sich so ein Trend bisher nicht erkennen. Da die Random Forest Modelle die höchste Vorhersagegenauigkeit hatten, ist eine Orientierung an ihrer Vorhersage empfehlenswert.

Land	Rip	J48	RF	Rip	J48	RF	Rip	J48	RF
Afghanistan	-	+	+	-	+	+	-	-	-
Algeria	-	-	-	-	+	-	-	-	-
Angola	-	+	-	+	+	-	-	-	-
Antigua and Barbuda	-	-	-	-	+	-	-	-	-
Argentina	-	-	-	-	-	-	-	-	-
Armenia	-	-	-	-	+	-	-	-	-
Bangladesh	-	-	-	-	+	-	-	-	-
Belize	-	-	-	-	+	-	-	-	-
Benin	+	+	-	-	+	-	-	-	-
Bhutan	-	+	-	-	+	-	-	-	-
Bolivia	-	-	-	-	+	-	-	-	-
Botswana	-	-	-	-	+	-	-	-	-
Brazil	-	-	-	-	-	-	-	-	-
Burkina Faso	-	+	+	+	+	+	-	-	+
Burundi	+	+	+	+	+	+	+	-	+
Cabo Verde	-	-	-	-	+	-	-	-	-
Cambodia	-	-	-	-	+	-	-	-	-
Cameroon	+	+	-	-	+	-	-	-	-
Central African Republic	+	-	-	-	-	-	+	-	-
Chad	-	+	+	+	+	+	+	-	+
Chile	-	-	-	-	+	-	-	-	-

China	-	-	-	-	-	-	-	-	-
Colombia	-	-	-	-	+	-	-	-	-
Comoros	+	-	+	-	+	-	-	-	-
Congo, Dem. Rep.	-	+	+	+	+	+	+	-	-
Congo, Rep.	-	+	-	-	+	-	-	-	-
Costa Rica	-	-	-	-	+	-	-	-	-
Cote d'Ivoire	+	+	+	-	+	+	-	-	-
Djibouti	-	-	-	-	+	-	-	-	-
Dominica	-	-	-	-	+	-	-	-	-
Dominican Republic	-	-	-	-	-	-	-	-	-
Ecuador	-	-	-	-	+	-	-	-	-
Egypt, Arab Rep.	-	-	-	-	+	-	-	-	-
El Salvador	-	-	-	-	-	-	-	-	-
Equatorial Guinea	-	+	-	-	+	-	-	-	-
Eritrea	-	-	-	-	-	-	-	-	-
Ethiopia	+	-	-	-	+	-	-	-	-
Gabon	-	-	-	-	+	-	-	-	-
Gambia, The	-	+	+	+	+	-	+	-	-
Ghana	+	-	-	-	+	-	-	-	-
Grenada	-	-	-	-	+	-	-	-	-
Guatemala	-	-	-	-	+	-	-	-	-
Guinea	+	+	+	-	+	+	-	-	-
Guinea-Bissau	-	+	+	-	+	-	-	-	-
Guyana	-	-	-	-	+	-	-	-	-
Haiti	-	+	-	-	+	-	-	-	-
Honduras	-	-	-	-	+	-	-	-	-
India	-	-	-	-	-	-	-	-	-
Indonesia	-	-	-	-	-	-	-	-	-
Iran, Islamic Rep.	-	-	-	-	+	-	-	-	-
Iraq	-	+	-	-	+	-	-	-	-
Jamaica	-	-	-	-	-	-	-	-	-
Jordan	-	-	-	-	+	-	-	-	-
Kenya	-	-	+	-	+	-	-	-	-
Lao PDR	-	-	-	-	+	-	-	-	-
Lebanon	-	-	-	-	+	-	-	-	-
Lesotho	-	+	-	-	+	+	-	-	-
Liberia	-	+	+	-	+	-	-	-	-
Libya	-	-	-	-	+	-	-	-	-
Madagascar	-	+	+	-	+	-	-	-	-
Malawi	-	+	+	-	+	+	-	-	+
Mali	-	+	+	+	+	+	+	-	-
Mauritania	-	+	-	-	+	-	-	-	-
Mexico	-	-	-	-	-	-	-	-	-
Mongolia	-	-	-	-	+	-	-	-	-
Morocco	-	-	-	-	+	-	-	-	-
Mozambique	+	+	+	+	+	+	-	-	+
Myanmar	-	-	-	-	+	-	-	-	-

Namibia	-	+	-	-	+	-	-	-	-
Nepal	-	-	-	-	+	-	-	-	-
Nicaragua	-	-	-	-	+	-	-	-	-
Niger	-	+	+	+	+	+	+	-	+
Nigeria	-	+	-	+	+	-	+	-	-
Oman	+	+	-	-	+	-	-	-	-
Pakistan	-	+	-	-	+	-	-	-	-
Panama	-	-	-	-	+	-	-	-	-
Paraguay	-	-	-	-	+	-	-	-	-
Peru	-	-	-	-	+	-	-	-	-
Philippines	-	-	-	-	-	-	-	-	-
Rwanda	-	-	-	-	+	-	-	-	-
Sao Tome and Principe	+	+	-	-	+	-	-	-	-
Saudi Arabia	+	-	-	-	+	-	-	-	-
Senegal	-	+	-	-	+	-	-	-	-
Sierra Leone	+	+	+	-	+	-	+	-	-
Somalia	+	+	+	+	+	+	+	-	-
South Africa	-	-	-	-	+	-	-	-	-
South Sudan	-	-	+	-	+	+	+	-	+
Sri Lanka	-	-	-	-	+	-	-	-	-
St. Lucia	-	-	-	-	+	-	-	-	-
Sudan	-	-	-	-	-	-	-	-	-
Suriname	-	-	-	-	+	-	-	-	-
Swaziland	-	-	-	-	+	+	-	-	-
Syrian Arab Republic	-	-	-	-	-	-	-	-	-
Tanzania	-	+	+	-	+	-	-	-	-
Thailand	-	-	-	-	+	-	-	-	-
Togo	+	+	-	-	+	-	-	-	-
Trinidad and Tobago	-	-	-	-	+	-	-	-	-
Tunisia	-	-	-	-	-	-	-	-	-
Turkey	-	-	-	-	-	-	-	-	-
Uganda	-	+	+	+	+	+	-	-	+
Venezuela, RB	-	-	-	-	-	-	-	-	-
Vietnam	-	-	-	-	+	-	-	-	-
Yemen, Rep.	-	+	-	-	+	-	-	-	-
Zambia	-	+	-	+	+	-	-	-	-
Zimbabwe	-	-	-	-	+	-	-	-	-

Abb. 19: Anwendung der Modelle aus 3.2 auf die Länderdaten von 2015

Literaturverzeichnis

[FERI] William W. Cohen: Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, 115-123, 1995.

[PML] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA

[RanF] Leo Breiman (2001). Random Forests. Machine Learning. 45(1):5-32.

[SimMod] Debraj Roy, Michael Harold Lees, Bharath Palavalli, Karin Pfeffer, M.A. Peter Sloot et al. „The emergence of slums: A contemporary view on simulation models", Environmental Modelling & Software Volume 59, 15 pages, 2014

[DatM] Ian H. Witten, Eibe Frank. „Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann; 4. Auflage, 2016