Martin Christian Wiltsche BSc.

Ableitung des Bruttoinlandprodukts von Brasilien auf Basis von Nacht- Satellitenbildern und weiteren Geodaten durch Anwendung eines Machine-Learning Modells

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science

Masterstudium

Geospatial Technologies

eingereicht an der

Technischen Universität Graz

Betreuer

Ao.Univ.-Prof. Mag. Dr.rer.nat, Wolfgang Sulzer

Institut für Geographie und Raumforschung

Karl-Franzens-Universität Graz

Co-Betreuer

Dr., Michael Wurm

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)

Deutsches Fernerkundungsdatenzentrum

Graz, Mai 2023

Vorwort

Kurzfassung

Der Begriff „Machine Learning“ ist mittlerweile im Alltag der meisten Menschen präsent, aber auch in der Wissenschaft spielt die Methode eine immer größer werdende Rolle in der Datenverarbeitung. Diese Masterarbeit widmet sich der Ableitung des Bruttoinlandsprodukts von Brasilien auf Basis von Nacht- Satellitenbildern und weiteren Geodaten durch die Anwendung eines Machine Learning Modells. Das Ziel dieser Arbeit ist die Erstellung eines Datensatzes, mit dem ein Machine Learning Modell trainiert werden kann, um das Bruttoinlandsprodukt bestimmter Regionen in Brasilien vorherzusagen. Die Forschungsfragen, die im Zuge dieser Arbeit behandelt werden, beschäftigen sich damit, ob das Modell auf unterschiedliche Testgebiete innerhalb Brasiliens angewandt werden kann und ob ein Zusammenhang zwischen der nächtlichen Beleuchtung und der Wirtschaftskraft einer Region besteht.

Der Datensatz, der zur Beantwortung der Forschungsfragen erstellt werden muss, besteht aus Referenzdaten, die in einer Auflösung von 1x1km das BIP enthalten und den Inputdaten des Modells, die in der gleichen geometrischen Auflösung akquiriert werden müssen. Die Referenzdaten des Bruttoinlandsprodukts werden durch eine einkommensbasierte Disaggregation erstellt. Die Inputdaten hingegen bestehen aus Sentinel-2- und Black Marble Satellitenbildern, sowie aus den Sentinel-2 Bandkombinationen „NDVI“, „NDBI“ und „MNDWI“. Nach der Erstellung des Datensatzes werden unterschiedlichste Parameterkombinationen für das Modell getestet, um ein optimales Ergebnis zu erzielen. Beim Machine Learning Modell handelt es sich um ein Fusionsmodell aus einem Convolutional Neural Network (CNN) und einem Multilayer Perceptron (MLP), welches vom DLR speziell für diesen Einsatz entwickelt und zur Verfügung gestellt wurde.

Durch die Anwendung des Vorhersagemodells konnten Ergebnisse für 14 der bevölkerungsreichsten Städte Brasiliens berechnet werden. Darunter Sao Paulo, mit einem Bruttoinlandsprodukt von 687.035.890 brasilianischen Real (R$). Das BIP konnte mit einem Determinationskoeffizienten R^2 von 0,64 und einer Pearson Korrelation R von 0,8 nachmodelliert werden. Die Genauigkeit der Modellierung variiert jedoch stark zwischen den unterschiedlichen Testgebieten. Es stellt sich heraus, dass die Werte durch die Modellierung geglättet werden und somit Ausreißer in den BIP-Werten verloren gehen. Weiters werden Kacheln, die sich in ruralen Regionen befinden, zu hohe Werte zugewiesen. Der Einfluss der nächtlichen Beleuchtung auf das BIP zeigt sich in der Verbesserung der Performance des Modells durch die Einbindung der Black Marble Daten.

Weiterführende Forschung in diesem Bereich wäre eine Disaggregation der BIP-Daten nach Wirtschaftssektoren unter Einbindung von Landbedeckungs- und OpenStreetMap Daten, sowie eine weitere Optimierung des Machine Learning Modells hinsichtlich ruraler Räume.

Abstract

Machine learning nowadays is a present term in most people's everyday lives, but the method is also playing an increasingly important role in data science. This master thesis is dedicated to the derivation of the gross domestic product (GDP) of Brazil based on nighttime satellite imagery and other geodata by applying a machine learning model. The goal of this thesis is to create a dataset, that can be used to train a machine learning model to predict the GDP of specific regions in Brazil. The research questions deal with whether the model can be applied to different test areas within Brazil and whether there is a connection between the nighttime illumination and the economic strength of a region.

The dataset which must be acquired to answer the research questions consists of reference data and input data for the machine learning model, both in a spatial resolution of 1x1km. The GDP reference data is generated by income-based disaggregation. The input data, on the other hand, consists of Sentinel-2 and Black Marble satellite images, as well as the Sentinel-2 band combinations "NDVI", "NDBI" and "MNDWI". After the dataset has been created, a variety of parameter combinations are tested to achieve an optimal model outcome. The machine learning model is a fusion model of a convolutional neural network (CNN) and a multilayer perceptron (MLP), which was developed specially for this purpose and made available by DLR.

By applying the prediction model, results could be calculated for 14 of the most populous cities in Brazil. Among them Sao Paulo, with a gross domestic product of 687,035,890 Brazilian Reals (R$). The GDP could be modeled with a determination coefficient R^2 of 0.64 and a Pearson correlation R of 0.8. However, the accuracy of the modeling varies greatly between the different test areas. It turns out that the values are smoothed by the modeling and that outliers in the GDP values get lost by the modeling. Furthermore, tiles that are located in rural regions get assigned values that are way too high. The impact of nighttime illumination on GDP is proved by the improvement of model performance through the inclusion of Black Marble data.

Further research in this area would be a disaggregation of the GDP data by economic sectors, including landcover and OpenStreetMap data, as well as a further optimization of the machine learning model regarding rural areas.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Problemstellung

Sozioökonomische Indikatoren sind eine wichtige Entscheidungsgrundlage für Regierungen und politische Entscheidungsträger. Nur auf Basis vollständiger und richtiger Daten können politische Entscheidungen getroffen werden. Hoch entwickelte Länder verfügen über statistische Datenbestände, die ökonomische Indikatoren wie das Bruttoinlandsprodukt (BIP) beinhalten. Dieser Indikator beziffert den Gesamtwert aller Waren und Dienstleistungen, die innerhalb eines Jahres in einem Land hergestellt werden. „Es ist die am weitesten anerkannte Kennzahl für die Wirtschaftskraft und Leistung eines Landes“ (vgl. Conway 2011). In Entwicklungsländern besteht jedoch das Problem, dass gewisse Datenbestände, vor allem auf regionaler Ebene, fehlerhaft oder nicht vorhanden sind. Gründe dafür liegen in mangelnder Infrastruktur und Ressourcen, um fundierte statistische Datenbestände erheben und erhalten zu können. Ein mangelhafter Datenbestand hindert den Einblick in wirtschaftliches Wachstum, Armut, Gesundheit und Lebensqualität innerhalb eines Landes (vgl. Chen und Nordhaus 2011).

Die folgende Arbeit soll beschreiben, wie ein Machine Learning Modell eingesetzt werden kann, um aus Geodaten das Bruttoinlandsprodukt (BIP) von Brasilien abzuleiten. Es soll somit versucht werden, aus frei zugänglichen Daten, mit Hilfe von künstlicher Intelligenz (KI), die Wirtschaftsleistung von bestimmten Regionen in Brasilien zu prädizieren. Dieser Ansatz würde es ermöglichen, Datenbestände zu erneuern oder statistische Daten für Regionen zu akquirieren, die aus wirtschaftlichen oder politischen Gründen nicht erfasst sind. Eine Modellierung des Bruttoinlandsprodukts wäre daher neben dem wissenschaftlichen Mehrwert auch für die Wirtschaft und Verwaltung eines Landes von Interesse.

Das beschriebene Modell verarbeitet die Daten in einer geometrischen Auflösung von 1x1km, modelliert daher auch das BIP in dieser Auflösung. Als Inputdaten dienen sowohl Sentinel-2- als auch Blackmarble- Satellitenbilder und unterschiedliche Bandkombinationen (Indizes) aus Sentinel-2 Bildern. Im Zuge der Analysen soll getestet werden, welche Daten sich am besten für das Training eines Machine Learning Modells eigenen und ob es möglich ist, aus ihnen das BIP zu modellieren. Als Forschungsgebiet wurde Brasilien gewählt, da für dieses Land ein guter statistischer Datenbestand vorliegt und eine sehr große Fläche abgedeckt wird. Dadurch soll auch getestet werden, ob das Modell unabhängig von geographischer Lage und den einhergehenden klimatologischen Unterschieden angewandt werden kann.

1.2 Zielsetzung

Das Ziel der Arbeit ist es, einen flächendeckenden Datenbestand für Brasilien zu generieren und einen Machine Learning Datensatz bestehend aus Trainings-, Validierungs-, und Testdaten für das Vorhersagemodell zu erstellen. Der Datenbestand soll sowohl die Referenzdaten des Bruttoinlandsprodukts als auch die Inputdaten für das Machine Learning Modell enthalten. Da es sich bei den Daten um Informationen unterschiedlichster Datenformate und Auflösungen handelt, wird die Datenaufbereitung einen großen Bestandteil der Arbeit in Anspruch nehmen. Zusätzlich zur Datenaufbereitung soll das Modell angewandt werden, um Ergebnisse unterschiedlicher Testgebiete miteinander zu vergleichen. Dabei werden mehrere Parameter und Einstellungen im Modell getestet. Das Machine Learning Modell wird hierfür vom Deutschen Zentrum für Luft- und Raumfahrt (DLR), entwickelt von Thomas Stark, zur Verfügung gestellt. Das Ziel dieser Forschung im Allgemeinen ist es, frei zugängliche Daten zu nutzen, um ökonomische Indikatoren ableiten zu können. Somit könnte ein Modell in Gebieten mit vorhandenen Referenzdaten trainiert werden, um es auf Gebiete ohne Referenzdaten anzuwenden. Dies betrifft vor allem Entwicklungsländer mit geringem statistischen Datenbestand (vgl. Chen und Nordhaus 2011).

Im Zuge dieser Arbeit sollen somit folgende Forschungsfragen beantwortet werden:

- Ist es möglich, mit Hilfe von Machine Learning, die Wirtschaftskraft einer Region aus Geodaten abzuleiten?
- Kann das entwickelte Machine Learning Modell auf unterschiedliche Testgebiete in Brasilien angewandt werden?
- Gibt es einen Zusammenhang zwischen der nächtlichen Beleuchtung und der Wirtschaftskraft einer Region?

Die erste Forschungsfrage hinterfragt allgemein die Vorhersage von ökonomischen Indikatoren mit Hilfe von Machine Learning. Die Beantwortung dieser Forschungsfrage setzt eine präzise Datenvorverarbeitung sowie ein funktionierendes Vorhersagemodell voraus. Als Geodaten dienen, in Bezug auf diese Fragestellung, sowohl Nacht- als auch Tagsatellitenbilder, sowie Zusatzdaten, die die Spektralwerte unterschiedlichster Bandkombinationen aus Sentinel-2 Daten enthalten. In der zweiten Forschungsfrage soll herausgefunden werden, inwiefern die geographische Lage einen Einfluss auf die Ergebnisse des Machine Learning Modells hat. Da Brasilien eine große Nord-Süd Ausdehnung aufweist, werden unterschiedliche Klimazonen abgedeckt. Das hat vor allem einen Einfluss auf die Pixelwerte der Sentinel-2 Daten, die unter anderem als Input in das Modell eingehen. Auf diese Thematik wird schon in der Datenvorverarbeitung eingegangen, um die klimatischen Effekte

zu minimieren. Bei der Anwendung des Machine Learning Modells soll auf eine gleichmäßige Verteilung der Testgebiete geachtet werden. Ein weiterer Faktor, der die Beantwortung dieser Forschungsfrage beeinflusst sind die Stadtstrukturen, die je nach Lage oder Größe der Städte variieren können. Die dritte Forschungsfrage bezieht sich auf die Blackmarble Daten, deren Pixelwerte die nächtliche Beleuchtung darstellen. Hier soll folgende Hypothese überprüft werden: Je höher die nächtliche Beleuchtung, desto höher ist das Bruttoinlandsprodukt einer Region. Es besteht also die Annahme, dass in den Black Marble Daten die bevölkerten Regionen ersichtlich sind und die Bevölkerungsdichte mit der Höhe des BIP korreliert.

1.3 State of the art

Die Ableitung von ökonomischen Indikatoren aus Geodaten ist ein aktuelles Forschungsgebiet. Durch die steigende Rechenleistung von Computern und den daraus resultierenden Fortschritten in Bereich Machine Learning ergeben sich viele Möglichkeiten, in diesem Forschungsgebiet laufend neue Erkenntnisse zu erlangen. Bei Geodaten wird prinzipiell zwischen Raster- und Vektordaten unterschieden. Im Zusammenhang mit der Ableitung von sozioökonomischen Indikatoren können beide Datenformate eingesetzt werden. Als Vektordaten können OpenStreetMap-Daten herangezogen werden, die unter anderem die Verkehrsinfrastruktur beinhalten. Weitere Inhalte der OpenStreetMap, die im Zusammenhang mit sozialgeographischen Strukturen stehen, sind bestimmte Points of Interest. Dabei kann über die Verteilung von gewissen Objekten auf die wirtschaftliche und soziale Situation einer Region geschlossen werden. Nach Feldemeyer et al. (2020) gibt es beispielsweise einen Zusammenhang zwischen der räumlichen Verteilung von Points of Interest der Kategorien „Tourism“ und „Natural sights“ und der Verteilung von Arbeitslosigkeit in Baden Wüttemberg. Neben den Vektordaten spielen jedoch auch die Rasterdaten eine große Rolle in diesem Forschungsbereich. Durch die freie Zugänglichkeit und teilweise hohe Auflösung von Satellitenbildern, dienen diese als gute Informationsquelle über die Erdoberfläche. Der Einsatz von Satellitenbildern in Kombination mit Machine Learning wurde bereits von Jean et al. (2016) untersucht, die in ihrer Arbeit „Combining satellite imagery and machine learning to predict poverty“ wirtschaftliche Indikatoren aus Nacht- und Tagessatellitenbildern ableiteten. Dabei wurde ein Convolutional Neural Network (CNN) trainiert, um aus Tagessatellitenbildern bestimmte Muster zur Vorhersage von Armut zu extrahieren. Es wurde auch eine sogenannte „transfer learning“ Methode angewandt, um aus Tagessatellitenbildern die nächtliche Beleuchtung vorauszusagen. Zur Quantifizierung von Armut, die durch das Modell prädiziert werden sollte, wurde ein Indikator für Konsumausgaben und ein weiterer Indikator für den Haushaltsvermögenswert herangezogen. Die Konsumausgaben entsprechen dem Betrag, den ein Haushalt pro Jahr für Konsumgüter ausgibt. Dieser wird von einer Studie der Weltbank (LSMS) definiert. Der Haushaltsvermögenswert wird von der Institution „Demographic and

Health Surveys“ erhoben. Das Untersuchungsgebiet der Studie umfasst fünf afrikanische Länder, die größten darunter Nigeria und Tansania. Ein Ziel der Arbeit war es, zu quantifizieren, wie sich Modelle verhalten, die in einem bestimmten Land trainiert und auf ein anderes Land angewandt werden.

Eine Studie, die sich auf das gleiche Untersuchungsgebiet bezieht, wie diese Arbeit, ist jene von Charris et al.. In ihrer Publikation „Mapping the Human Development Index using Nighttime Lights inside Brazil” wird der Zusammenhang zwischen lokalem wirtschaftlichen Entwicklungsstand (MHDI) und nächtlicher Beleuchtung untersucht. Der MHDI (Munizipal Human Development Index) beschreibt den HDI (Human Development Index) auf einer kleineren Raumebene und gibt Auskunft über den Lebensstandard, Bildung und Einkommen in einer Region. Die Studie zeigt, dass Gemeinden mit intensiverer nächtlicher Beleuchtung einen höheren MHDI aufweisen. Dazu kommt eine niedrigere Geburtensterblichkeit und größere ökonomische Aktivität, gemessen an der Anzahl von Kraftwerken. Zusammenfassend liefern die Ergebnisse Hinweise darauf, dass die nächtliche Beleuchtung ein guter Indikator für sozioökonomische Indikatoren auf subnationaler Ebene in Entwicklungsländern sein kann (vgl. Charris et al.).

Ein weiterer Index, der durch Kombination von Bevölkerungsdichte und Satellitenbildern abgeleitet werden kann, ist der Night Light Development Index (NLDI). Elvidge et al. (2012) bezeichnet diesen als einen simplen, objektiven und global verfügbaren empirischen Messwert zur Bewertung des Entwicklungsstands einer Region. Er basiert auf Nachtsatellitenbildern und der Einwohnerverteilung in Form eines 1x1km Rasters. Der NLDI steht mit einer Korrelation R^2 von 0,71 in starken Zusammenhang zum HDI.

Nach Chen und Nordhaus (2011) besteht auch ein Zusammenhang zwischen dem Bruttoinlandsprodukt und der nächtlichen Beleuchtung einer Region. Unter Berücksichtigung dieser Korrelation soll in ihrer Arbeit „Using luminosity data as a proxy for economic statistics“ mit Hilfe eines statistischen Modells der Datenbestand des Bruttoinlandsprodukts auf regionaler Ebene verbessert werden. Bei den Inputdaten handelt es sich hierbei um 1x1Grad Gitterzellen. Es werden mehrere Länder unterschiedlicher Entwicklungsstufen, und damit einhergehend unterschiedlicher Qualität statistischer Datenbestände, in das Modell miteinbezogen. Es zeigt sich, dass mit Hilfe des Modells die BIP-Daten von Ländern mit mangelhaftem Datenbestand verbessert werden können. In hoch entwickelten Ländern ist jedoch der Modellierungsfehler aus den Satellitenbildern größer als jener in den statistischen Daten, was daher keine Verbesserung ermöglicht.

Die Nachtsatellitenbilder, die in allen genannten Arbeiten herangezogen werden, sind jene des United States Air Force Defense Meteorological Satellite Program (DMSP). Diese Daten wurden mit dem Operational Linescan System (OLS)- Sensor aufgenommen und liefern Bilder

im sichtbaren Bereich und im nahen Infrarot. Die Daten sind für den Zeitraum von 1992 bis 2013 in einer Auflösung von 30 Bodensekunden erhältlich.

Tabellarische Darstellung Literatur

Autoren	Daten	Methodik	Vorhersageparameter
Jean, Neal; Burke, Mashall; Xie, Michael; Davis, Matthew; Lobell, David B.; Ermon, Stefano	Tagsatellitenbilder (Google Static Maps) Nachtsatellitenbilder (DMSP-OLS)	Convolutional Neural Network (CNN), transfer learning	Konsumausgaben (LSMS), Haushaltsvermögenswert (DHS)
Charris, Carlos; Velilla, Raul; Chaves, Leonardo	Nachtsatellitenbilder (DMSP-OLS)	Statistisches Modell	Municipality Human Development Index (MHDI)
Elvidge, C. D.; Baugh, K. E.; Anderson, S. J.; Sutton, P. C.; Ghosh, T.	Nachtsatellitenbilder (DMSP-OLS)	Statistisches Modell	Night Light Development Index (NLDI)
Chen, Xi; Nordhaus, William D.	Nachtsatellitenbilder (DMSP-OLS)	Statistisches Modell	Gross Domestic Product (BIP)
Feldemeyer, Daniel; Meisch, Claude; Sauter, Holger; Birkmann, Joern	OpenStreetMap	Lineare Regression, Random Forest, Deep Neural Network (DNN)	Einwohner, Arbeitslosigkeit, Migration, Senioren

In Tabelle xy ist die beschriebene Literatur aufgelistet, um einen Überblick über die verwendeten Daten und Methoden zu gewinnen. Der Lösungsansatz dieser Arbeit, der in der Kombination von Machine Learning mit Satellitenbildern liegt, lehnt sich an jenen von Jean et al. (2016) an. Es wird ebenfalls ein Convolutional Neural Network (CNN) trainiert, um Muster aus den Satellitenbildern zu extrahieren. Der Unterschied liegt jedoch darin, dass in dieser Arbeit auch noch weitere Geodaten und ein zweites Machine Learning Modell zur Verarbeitung dieser miteinbezogen werden. Ein weiterer Unterschied zur genannten Literatur liegt in der Datengrundlage der Nachtsatellitenbilder. In diesem Fall werden nämlich nicht die DMSP-Daten verwendet, sondern Black Marble Daten der NASA. Der Grund dafür liegt in der temporalen und geometrischen Auflösung der Satellitenbilder. Die Daten der US-Air Force sind nur bis zum Jahr 2013 in einer Auflösung von 30 Bogensekunden erhältlich, während die Black Marble Bilder seit 2012 in einer Auflösung von 15 Bogensekunden zugänglich sind. Die Methode, das Modell in einem Land zu trainieren, um es auf andere Länder anzuwenden wird in dieser Arbeit auch abgeändert, da sich die Analysen ausschließlich auf Testgebiete innerhalb Brasiliens begrenzen.

Auch Charris et al. und Elvidge et al. (2012) nutzen Nachtsatellitenbilder, um sozioökonomische Indikatoren abzuleiten. Diese beschränken sich in den beiden

Publikationen jedoch auf den HDI und den NLDI. Da der NLDI jedoch sehr stark mit dem HDI korreliert und beide Indikatoren sozioökonomische Strukturen beschreiben, stehen sie in Zusammenhang mit den voraussichtlichen Ergebnissen dieser Arbeit. Gemäß der Studie „Relationship between Gross Domestic Product and Human Development Index" von Hudakova (2017) besteht auch eine signifikante Korrelation zwischen dem HDI und dem BIP. Sie beschreiben zwar unterschiedliche wirtschaftliche Eigenschaften eines Staates, beeinflussen sich jedoch gegenseitig. Beispielsweise wird das Bildungs- und Gesundheitssystem, welches in den HDI miteinfließt, stark vom Bruttoinlandsprodukt beeinflusst.

Chen und Nordhaus (2011) dagegen beschäftigten sich mit dem Zusammenhang zwischen dem BIP und der nächtlichen Beleuchtung, was von der Fragestellung sehr der dieser Arbeit ähnelt. Die Methodik unterscheidet sich jedoch, da Chen und Nordhaus ein statistisches Modell benutzten, welches 22 unterschiedliche Länder beinhaltet. Die Resultate der genannten Arbeit stützen jedoch die Hypothese, dass die Intensität der nächtlichen Beleuchtung mit dem Bruttoinlandsprodukt korreliert. Die Ableitung von sozioökonomischen Indikatoren aus Geodaten ist also bereits Thema der Wissenschaft und es wurden auch schon einige Zusammenhänge bewiesen. Die Methodik dieser Arbeit unterscheidet sich jedoch zu den Ansätzen in der Literatur, da als Informationsquelle nicht nur Tages- und Nachtsatellitenbilder, sondern auch Zusatzdaten in Form von Bandkombinationen aus Sentinel-2 Daten in das Modell mit einfließen.

2 Fallstudie: Ableitung des BIP von Brasilien

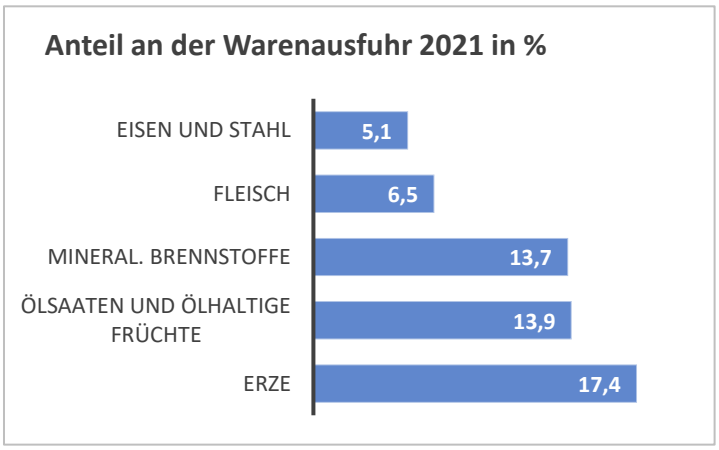
Beginnend mit einem Überblick über das Untersuchungsgebiet und einer Darlegung der wirtschaftlichen Situation in Brasilien wird in diesem Kapitel die konkrete Fallstudie, die im Zuge dieser Arbeit durchgeführt wurde, beschrieben. Das weiterführende Unterkapitel widmet sich der Datengrundlage. Dabei werden die unterschiedlichen Geodaten vorgestellt und auf Basis der räumlichen und temporalen Auflösung miteinander verglichen. Die Methodik dieser Arbeit umfasst einerseits die Datenvorverarbeitung und andererseits das Machine Learning Modell und die damit verbundenen Analysen. Ein großer Teil der Datenvorverarbeitung wird von der Disaggregation eingenommen, mit Hilfe derer der Referenzdatensatz für das Machine Learning Modell erstellt wird. Das Sentinel-2 Mosaik, welches für ganz Brasilien erstellt wird, ist ebenfalls Teil der Datenvorverarbeitung. Auf die einzelnen Arbeitsschritte wird in den entsprechenden Unterkapiteln eingegangen. Dieses Kapitel abschließend wird das verwendete Machine Learning Modell und die durchgeführten Experimente beschrieben.

2.1 Untersuchungsgebiet

Das Untersuchungsgebiet für diese Arbeit wird abgegrenzt durch die Festlandgrenze Brasiliens, wodurch eine Fläche von 8.515.664 Quadratkilometer abgedeckt wird. Der Staat wird in 5 statistische Großregionen unterteilt: den Norden, Nordosten, Südosten, Süden und Mittelwesten. Die räumlich nächstkleinere Unterteilung erfolgt in die 26 Bundesstaaten und einen Bundesdistrikt (Distritio Federal), in dem sich die Hauptstadt Brasilia befindet. Die

Bundesstaaten werden wiederum weiter untergliedert in die Gemeinden. In Abbildung xy sind die Großregionen, sowie die Bundesstaaten abgebildet.

Aus wirtschaftlicher Sicht gibt es in Brasilien große Disparitäten, insgesamt zählt es jedoch zu den größten Volkswirtschaften weltweit. Im Jahr 2021 betrug das Bruttoinlandsprodukt von Brasilien umgerechnet 1,61 Milliarden US$ (vgl. weltbank.org 2023). Ein brasilianischer Real (R$) entspricht 0,19 US$ (Stand März 2023). Der Anteil der Landwirtschaft an der Bruttowertschöpfung ist von 2005 bis 2021 von 4,7% auf 6,9% gestiegen, was im globalen Vergleich einen hohen Prozentsatz darstellt. Im Vergleich dazu verzeichnet die USA, die mit einem BIP von 23 Billionen US$ im Jahr 2021 als stärkste Volkswirtschaft angesehen werden kann, im Sektor Landwirtschaft lediglich 1% (vgl. wko.at). Im gleichen Zeitraum, in dem die Landwirtschaft eine Zunahme von 2,2% erfahren hat, ist der Anteil des Produktionsbereichs an der Bruttowertschöpfung um 5,3% auf einen Anteil von 18,9% gesunken. Der Anteil der Dienstleistungen betrug im Jahr 2021 59,4% (vgl. wko.at). Der Anteil des Exports von Waren und Dienstleistungen am BIP für das Jahr 2021 betrug 20,1%. Die beliebtesten Exportgüter waren dabei Erze, Ölsaaten und ölhaltige Früchte, mineralische Brennstoffe, Fleisch, Eisen und Stahl.



Den größten Anteil an der Warenausfuhr im Jahr 2021 nehmen die Erze ein, was die Bedeutung Brasiliens als Rohstofflieferant verdeutlicht. „Nicht nur aufgrund der enormen Größe, sondern auch wegen der geologischen Beschaffenheit ist das Land reich an unterschiedlichen mineralischen Rohstoffen“ (Colucci et al.). Rund 90% der geförderten mineralischen Rohstoffe werden exportiert. Den Größten Anteil (68%) nimmt dabei Eisenerz ein, gefolgt von den Edelmetallen Gold (9%) und Kupfer (9%) (vgl. ebd.). Der hohe Anteil der mineralischen Rohstoffe an der Bruttowertschöpfung von Brasilien stellt eine Herausforderung in der Ableitung des BIP durch Geodaten dar, da die Höhe des Bruttoinlandprodukts in den Referenzdaten nach dem Einkommen innerhalb der Gemeinden verteilt wird und die Rohstoffe meist nicht am gleichen Ort abgebaut und verkauft werden. Der Dienstleistungssektor erwirtschaftet jedoch den größten Teil des Bruttoinlandprodukts. Hierbei steht das Einkommen im direkten Zusammenhang mit dem BIP, was für die Modellierung von Vorteil ist.

2.2 Datengrundlage

Als Datengrundlage für diese Arbeit dienen mehrere Geodatensätze bestehend aus Sentinel-2, Black Marble, NDVI *(*Normalized Difference Vegetation Index*),* NDBI (Normalized Difference Built-up Index) und MNDWI (Modified Normalized Difference Water Index) - Daten. Alle genannten Datensätze sind Rasterdaten in unterschiedlich hoher geometrischer Auflösung. In Tabelle xy ist zu sehen, dass die Sentinel-2 Daten mit einer räumlichen Auflösung von 10 Metern am genauesten sind. Mit der hohen geometrischen Auflösung geht auch hohes Speichervolumen einher, was in weiterer Folge zu beachten sein wird. Die Black Marble Daten werden in einer Auflösung von 15 Bogensekunden veröffentlicht. Die Daten für diese Arbeit stammen jedoch vom DLR, die bereits auf eine geometrische Auflösung von 500m umgerechnet worden sind. Die Zusatzdaten NDVI, NDBI und MNDWI entsprechen derselben räumlichen Auflösung. In der Datenvorverarbeitung werden, mit Ausnahme der Sentinel-2 Daten, alle Rasterdaten in eine geometrische Auflösung von 1x1km gebracht. Die temporale Auflösung der Daten variiert teilweise um ein paar Jahre, was jedoch aufgrund der Verfügbarkeit vernachlässigt werden muss.

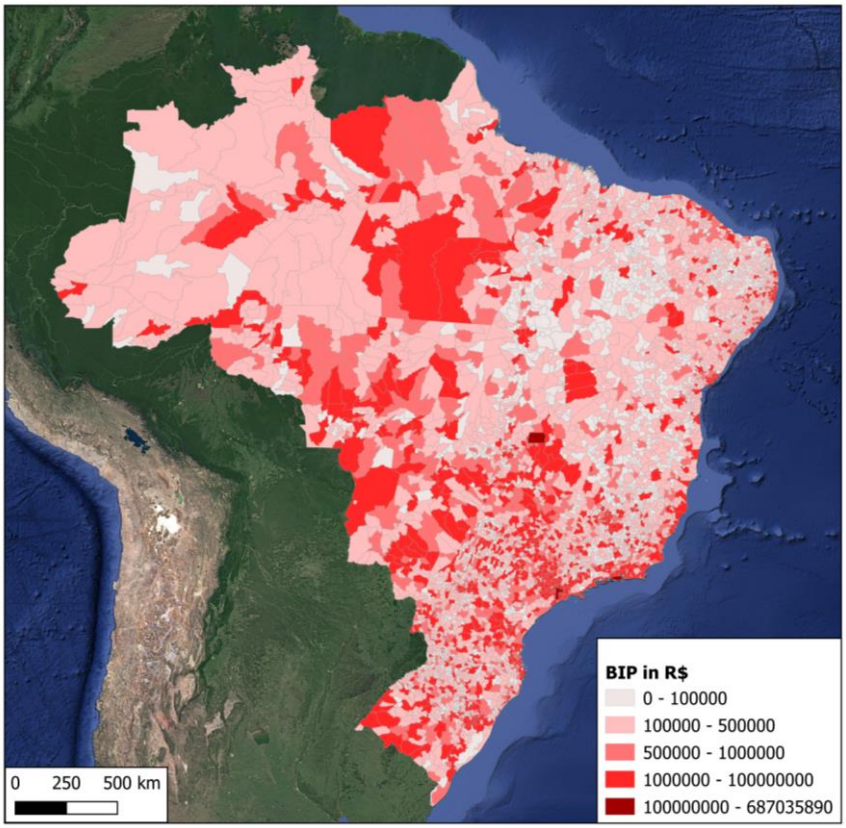
Tabellarische Darstellung der Datengrundlage

Datensätze		**räumliche Auflösung**	**Temporale Auflösung**	**Quelle**
1 Fernerkundungsdaten				
1.2	Black Marble	500m	2016	NASA
1.3	Sentinel 2	10m	2019-2021	Google Earth
1.4	NDVI	500m	2018-2020	Google Earth
1.5	NDBI	500m	2018-2020	Google Earth
1.6	MNDWI	500m	2018-2020	Google Earth
2 Statistische Daten				
2.1	BIP	Gemeindeebene	2016	IBGE
2.2	Zensus	Nachbarschaftsebene	2010	IBGE

Die statistischen Daten stehen als Vektordatensätze zur Verfügung. Der Vorteil dieses Datenformats ist, dass zu den statistischen Informationen auch die Geometrien mitgespeichert werden. Dadurch können die Datensätze räumlich miteinander verschnitten werden. Die BIP-Daten stehen auf Gemeindeebene zur Verfügung und werden mit Hilfe der Zensusdaten auf eine geometrische Auflösung von 1x1km disaggregiert. Die Zensusdaten dienen hierbei dazu, die räumliche Auflösung zu verbessern in dem das BIP nach dem Einkommen innerhalb der Gemeinden aufgeteilt wird. Eine nähere Beschreibung der Datensätze und der Datenvorverarbeitung folgt in den nachfolgenden Kapiteln.

2.2.1 BIP

Der BIP-Datensatz für Brasilien liegt in Form einer Shape-Datei vor und beinhaltet 5569 Gemeindepolygone. Die Größe der Gemeinden ist dabei sehr unterschiedlich. Gebiete mit hoher Bevölkerungsdichte werden kleinräumiger unterteilt als bevölkerungsärmere Gebiete. Auch die Höhe des Bruttoinlandprodukts weist eine große Schwankungsbreite auf. Bei den Werten handelt es sich um das aggregierte BIP innerhalb der Polygone, was in höheren Werten für größere Gebiete resultieren müsste. Da geringe Bevölkerungsdichte jedoch eher geringeres BIP indiziert, gleichen sich die Werte besser aus. Wie in Abbildung 1 zu erkennen, dominieren vor allem die Metropolregionen Sao Paulo, Rio de Janeiro und Brasilia mit BIP-Werten von über 100.000.000 Real.



2.2.2 Zensus

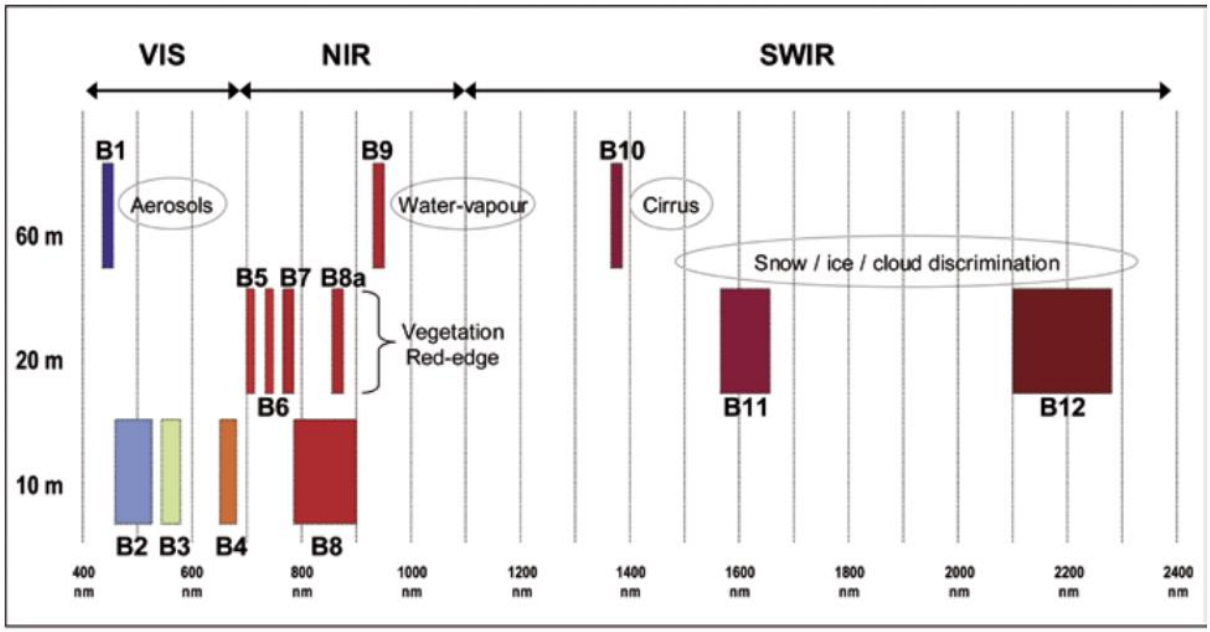
Die Zensusdaten stehen ebenfalls als Shape-Dateien zur Verfügung. Sie decken das gesamte Untersuchungsgebiet ab und sind aufgeteilt in die 26 Bundesstaaten und das Bundesdistrikt Districio Federal. Die räumliche Auflösung der Zensusdaten ist deutlich höher als die der BIP-Daten und unterteilt die Gemeinden in Nachbarschaftsregionen. Da sich die Grenzen mit den BIP-Daten decken, können die zwei Datensätze im Zuge der Datenvorverarbeitung miteinander verschnitten werden. Das Attribut, welches im Zuge des Zensus erhoben wurde und für diese Arbeit von Relevanz ist, ist das aggregierte Einkommen pro Region.

2.2.3 Sentinel-2

Die Satellitenmission „Sentinel-2“ besteht aus dem Satellitenpaar „Sentinel-2A“ und „Sentinel-2B“ und ist Teil des Copernicus Programms der Europäischen Union. Die Mission liefert öffentliche Daten zur Beobachtung von Veränderungen auf der Erdoberfläche. Seit 2017 befinden sich beide Satelliten in einem sonnensynchronen Orbit in einer Entfernung von 786km zur Erde. Die Wiederholrate eines Satelliten auf Höhe des Äquators beträgt 10 Tage, was in der zwei-Satellitenkonstellation eine Wiederholrate von 5 Tagen bedeutet. In mittleren Breiten verkürzt sich die Wiederholrate auf 2-3 Tage (vgl. esa.int 2023).

Die Sentinel-2 Satelliten liefern multispektrale Aufnahmen in einem Wellenlängenbereich von 443-2190nm, aufgeteilt auf 13 Kanäle. Für diese Arbeit wird die Produktreihe Level-2A verwendet, die bereits atmosphärisch korrigierte Bilder liefert. Die maximale geometrische Auflösung beträgt 10m im sichtbaren Bereich und im nahen Infrarot (vgl. copernicus.de 2023). Dies betrifft die Bänder B2, B3, B4 und B8, welche für diese Arbeit herangezogen werden.

Tabelle mit Wellenlängenbereichen (Quelle: ESA, Sentinel2.pdf):



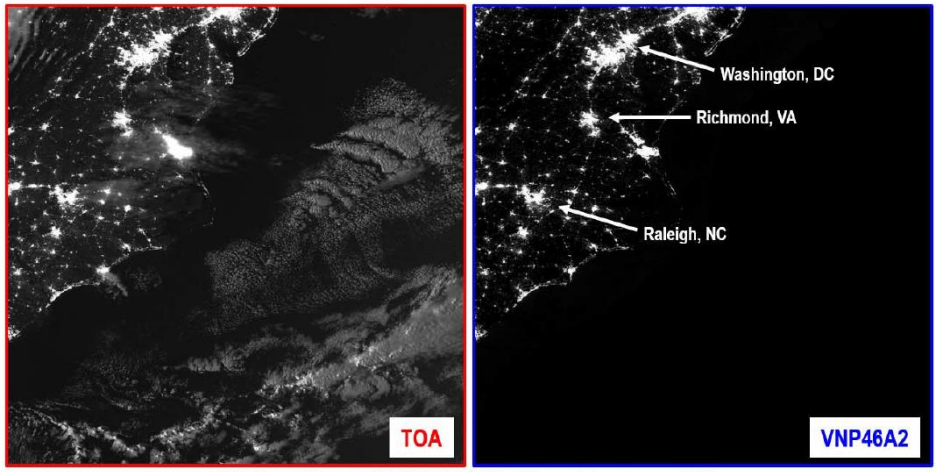
2.2.4 Black Marble

Die Nachtsatellitenbilder, die in dieser Arbeit zum Einsatz kommen, stammen von der Suomi-NPP (National Polar-orbiting Plattform) – Mission der NASA. Dabei handelt es sich um einen Wetter- und Umweltsatelliten, der sich seit 2011 in einer Entfernung von 830km zur Erde befindet. Der Satellit weist eine Wiederholrate von 16 Tagen auf und liefert mit Hilfe des VIIRS-Instruments (Visible Infrared Imaging Radiometer Suite) Aufnahmen in 22 Spektralbändern. Das DNB (day/night band) ist eines der Bänder im sichtbaren Wellenlängenbereich und nahem Infrarot, welches die Rohdaten für die Black Marble nighttime lights product suite (VNP46)

liefert. Dies ist eine Produktreihe, die seit 2012 Nachtsatellitenbilder in einer geometrischen Auflösung von 15 Bogensekunden öffentlich bereitstellt (vgl. nasa.gov 2023).

Die Wolkenbedeckung wird bei den Black Marble Bildern mit Hilfe einer Wolkenmaske bereinigt. Außerdem wird das direkte sowie gestreute Licht des Modes aus den Daten gefiltert, um ausschließlich anthropogen bedingte Lichtquellen darzustellen (vgl. Roman et al. 2018).

Quelle: Roman et al



US East Coast - DOY 2015-090 ; Moon Fraction = 83%

Der Datensatz für diese Arbeit wird vom DLR zur Verfügung gestellt und deckt das gesamte Untersuchungsgebiet in einer geometrischen Auflösung von 500m ab. Die Pixelwerte liegen zwischen 0 (schwarz) und 255 (weiß).

2.2.5 Zusatzdaten

Die im folgenden beschriebenen Zusatzdaten wurden vom DLR zur Verfügung gestellt und sind ein Produkt aus Sentinel-2 Bildern aus den Jahren 2018-2020. Durch mathematische Operationen von Spektralbändern lassen sich Indizes berechnen, die es ermöglichen, spezielle Beschaffenheiten der Erdoberfläche besser zu erkennen.

2.2.5.1 NDVI

Der NDVI (Normalized Difference Vegetation Index) ist eine Vegetationsvariable, die aus Fernerkundungsdaten berechnet werden kann. Dieser Wert gibt Auskunft über das Vorkommen und den Zustand von Vegetation auf der Erdoberfläche und ist eng verbunden mit der Vegetationsdichte und -produktivität (vgl. Tucker und Sellers 1986). Der Index wird nach folgender Formel berechnet:

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

Quelle: dlr.de

NIR und RED stehen hierbei für spektrale Reflektanzen im nahen Infrarot und im Rot. Bei Sentinel-2 Bildern wird der Index aus den Bändern „B8“ und „B4“ berechnet. Der NDVI kann Werte zwischen -1 und +1 einnehmen (vgl. dlr.de 2023).

2.2.5.2 NDBI

Der NDBI (Normalized Difference Buit-Up Index) nutzt Spektralwerte im kurzwelligen- und nahen Infrarot, um bebaute Flächen aus Fernerkundungsdaten zu extrahieren. Bei Sentinel-2 Daten wird er aus den Bändern „B11“ und „B8“ berechnet (vgl. Xu et al. 2018).

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR}$$

Quelle: Xu et al., 2018

Der NDBI kann Werte zwischen -1 und +1 einnehmen, wobei im Allgemeinen ein Pixelwert größer als 0 als bebaute Fläche angenommen werden kann. Da diese Klassifizierung einer gewissen Ungenauigkeit unterliegt, kann die Genauigkeit verbessert werden, indem für einen Datensatz ein gewisser Schwellwert berechnet wird (Otsu’s method), über dem ein Pixel der bebauten Fläche zugeordnet werden kann (vgl. Xu et al. 2018). In dieser Arbeit werden jedoch die originalen NDBI-Werte zwischen -1 und +1 verwendet, da das Machine Learning Modell die metrischen Zahlenwerte als Input erfordert.

2.2.5.3 MNDWI

Der MNDWI (Modified Normalized Difference Water Index) berechnet sich aus spektralen Reflektanzen im kurzwelligen Infrarot und im grünen Kanal des sichtbaren Lichts. Dies entspricht den Bändern „B3“ und „B11“ der Sentinel-2 Daten.

$$MNDWI = \frac{Green - SWIR}{Green + SWIR}$$

Ouelle: Xu et al.. 2018

Der Index dient dazu, Wasserflächen in Fernerkundungsdaten zu lokalisieren. Der Wertebereich liegt auch bei diesem Index zwischen -1 und +1, wobei Pixelwerte über 0 als

Wasserflächen klassifiziert werden können. Der Vorteil gegenüber dem NDWI ist die klare Trennung zwischen Wasserflächen und bebauten Flächen, die eine ähnliche spektrale Signatur aufweisen (vgl. Xu 2006).

2.3 Methodik

Eine wichtige Voraussetzung für ein Machine Learning Modell sind vollständige und akkurate Referenzdaten. Die Referenzdaten dienen als Richtwerte für das Modell und stellen jene Werte dar, an die sich das Modell annähern soll. Aus diesem Grund ist es wichtig, dass diese so genau wie möglich der Realität entsprechen. Als Referenzdatensatz dient das 1x1km Raster des Bruttoinlandprodukts, welches in der Datenvorverarbeitung generiert wird. Das BIP steht auf Gemeindeebene zur Verfügung und wird unter Einbeziehung des Einkommens auf eine geometrische Auflösung von 1x1km disaggregiert. Die Höhe des Einkommens ist in den Zensus-Daten gespeichert, die ebenfalls die Gemeindegrenzen enthalten, jedoch noch kleinräumiger untergliedert sind. Da sich die Grenzlinien der zwei Datensätze decken, können diese miteinander verschnitten werden. Das BIP wird dadurch in die Raumgliederung der Zensus-Daten gebracht. Durch diesen Verarbeitungsschritt wird die räumliche Auflösung des BIPs vergrößert, bevor die Werte in das 1x1km Raster übertragen werden.

Ein weiterer Datensatz, der in der Datenvorverarbeitung erstellt werden muss, ist ein Sentinel-2 Mosaik. Dieses wird in der Google Earth Engine berechnet und besteht aus den Medianwerten der Pixel aus mehreren Szenen eines definierten Zeitraums. Zusätzlich werden die Szenen mit einer Wolkenmaske bereinigt, um die Wolkenbedeckung zu minimieren. Die geometrische Auflösung der Sentinel-2 Daten beträgt 10m. Die Datensätze Black Marble, NDVI, MNDWI, NBI werden vom DLR zur Verfügung gestellt. Mit Ausnahme der Sentinel-2 Daten werden alle genannten Datensätze im Zuge der Datenvorbereitung in eine geometrische Auflösung von 1x1km gebracht.

Nach Abschluss der Datenvorverarbeitung wird der Datensatz für das Machine Learning Modell vorbereitet. In der Datenvorverarbeitung werden die Daten für ganz Brasilien akquiriert, was zu einem sehr großen Speichervolumen führt, welches zu groß wäre, um es in das Modell einzubinden. Aus diesem Grund werden Untersuchungsgebiete definiert, welche die 30 größten Städte Brasiliens beinhalten. Es wird ein Datensatz erstellt, der für jedes Untersuchungsgebiet alle Informationen in 1x1km Kacheln enthält.

Das Machine Learning Modell, welches zur Beantwortung der Forschungsfragen verwendet wird, basiert auf der Programmiersprache Python unter der Verwendung der Open-Source Bibliothek PyTorch. Diese Bibliothek bietet ein Framework für maschinelles Lernen und beinhaltet einige vorgefertigte Module, die es ermöglichen, neuronale Netzwerke zu entwickeln und zu trainieren. Für diese Arbeit wurde eine Kombination aus einem Convolutional Neural

Network (CNN) und einem Multilayer Perceptron (MLP) entwickelt und zur Verfügung gestellt. Die Sentinel 2 Daten werden im CNN verarbeitet, während die restlichen Daten in das MLP eingehen. Die zwei Netzwerke können entweder separat trainiert, oder in einem Fusion Layer miteinander kombiniert werden. In den Analysen werden unterschiedliche Parameter getestet, um den bestmöglichen Output des Modells zu erzielen.

2.3.1 Datenvorverarbeitung

Bei der Datenvorverarbeitung werden die räumlichen Daten in ein einheitliches Format gebracht. Als Koordinatenreferenzsystem dafür dient Sirgas 2000 (EPSG:5880), welches Zentral- und Südamerika vollständig abdeckt (vgl. epsg.io 2023). Das Ziel der Datenvorverarbeitung ist es, ein Raster zu generieren, welches das BIP in einer geometrischen Auflösung von 1x1km flächendeckend für Brasilien enthält. Dieses Raster dient in weiterer Folge als Referenzdatensatz für das Machine Learning Modell. Da die Ausgangsdaten als Polygondatensatz auf Gemeindeebene vorliegen, werden diese Anhand von höher aufgelösten Zensusdaten nach dem Einkommen disaggregiert, um die räumliche Auflösung zu verbessern.

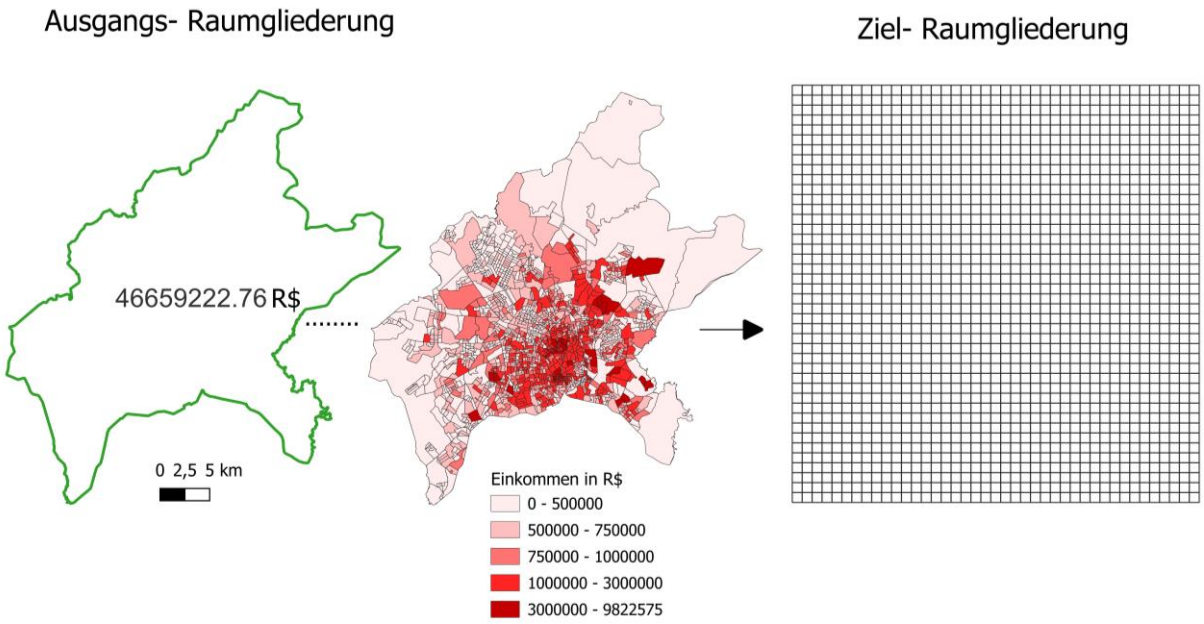
Ein weiterer Teil der Datenvorverarbeitung befasst sich mit der Erstellung eines Sentinel-2 Mosaiks für Brasilien. Die Datenakquisition dafür erfolgt über die Google Earth Engine, welche als Webanwendung den Zugriff auf Sentinel-2 Daten ermöglicht.

Die Zusatzdaten „NDVI“, „NDBI“ und „MNDWI“ werden im Zuge der Datenvorverarbeitung in das richtige Koordinatenreferenzsystem gebracht und geometrisch an die Auflösung des BIP-Rasters angepasst.

2.3.1.1 Disaggregation

Die Disaggregation wird in den Geowissenschaften angewandt, um räumliche Daten in einen gewünschten Raumbezug zu bringen. In dieser Arbeit geht es bei der Disaggregation darum, die räumliche Auflösung der BIP-Daten anhand von den Zensusdaten zu erhöhen und die Information in ein 1x1km Raster zu übertragen. Dabei wird die räumliche Auflösung verändert, indem die Daten in eine andere Raumgliederung übertragen werden. Eine einfache Methode ist dabei die flächengewichtete Interpolation nach Goodchild und Lam (1980). Diese beschreiben die Methodik anhand von Bevölkerungszahlen, die als Merkmal einer Ausgangs-Raumgliederung in eine Ziel-Raumgliederung übertragen werden. In den Schnittflächen wird die relative Fläche, in Bezug auf die Ausgangsfläche, mit dem Merkmal der Ausgangsfläche multipliziert und in die Zielfläche übertragen. Dadurch können Merkmale in höher aufgelöste Raumeinheiten übertragen werden. Voraussetzung dafür ist jedoch eine homogene Verteilung des Merkmals innerhalb der Ausgangsflächen (vgl. Goodchild und Lam 1980). In dieser Arbeit

wird die Gemeindeebene mit dem Merkmal des Bruttoinlandprodukts als Ausgangs-Raumgliederung angenommen, und in die Raumgliederung der Zensusdaten gebracht. Dadurch wird die geometrische Auflösung der BIP-Daten erhöht. Anschließend werden die Daten in die Ziel-Raumgliederung des 1x1km Rasters gebracht.



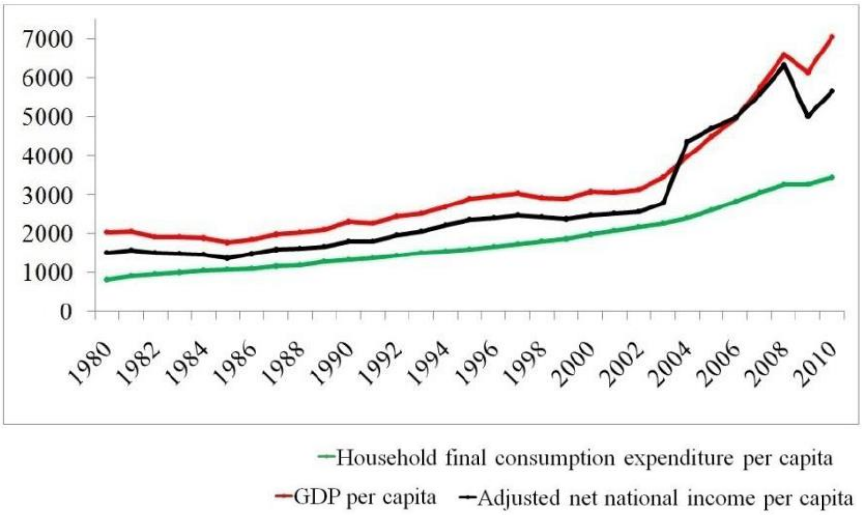
In Abbildung xy sind die Raumgliederungen anhand der Gemeinde Goias zu erkennen. Der Wert in der Ausgangs-Raumgliederung beziffert das BIP der Gemeinde in Real (R$). Anhand der Einkommensverteilung wird dieser Wert auf die Ziel-Raumgliederung aufgeteilt. Es wird also eine einkommensbasierte Disaggregation durchgeführt.

2.3.1.1.1 Einkommensbasierte Disaggregation

Um die BIP-Daten in kleinere räumliche Einheiten zu bringen, werden die BIP-Werte jeder Gemeinde aufgeteilt auf die Zensus-Regionen, die sich innerhalb dieser befinden. Die Gewichtung der Verteilung erfolgt dabei jedoch nicht nur nach Flächenanteilen, sondern auch nach der Höhe des Einkommens jeder Zensus-Region. Die Höhe des BIP-Wertes steht direkt proportional zur Höhe des Einkommens. Nach Diacon und Maha (2015) besteht eine Korrelation zwischen Konsum, Einkommen und BIP innerhalb einer Volkswirtschaft. Die Weltbank unterscheidet Länder nach Einkommen in vier Klassen: Low Income, Lower Middle Income, Upper Middle Income, High Income (vgl. weltbank.org 2023). Der Zusammenhang zwischen Konsum und Einkommen ist größer bei Ländern mit hohem Einkommen und Ländern mit niedrigem Einkommen. Der Grund dafür liegt in der höheren Bereitschaft der Bevölkerung mit mittlerem Einkommen Geld zu sparen, während in Ländern mit niedrigem Einkommen das Budget für Artikel des täglichen Lebens aufgebraucht wird. In Ländern mit hohem Einkommen

besteht die Annahme, dass mit einem Teil des Budgets Investitionen getätigt werden. Es besteht jedoch eine signifikante Korrelation zwischen Einkommen und BIP in allen 4 Einkommensklassen. Am höchsten ist die Korrelation jedoch in Ländern niedrigen- und mittleren Einkommens (vgl. Diacon und Maha 2015).

Nach Einstufung der Weltbank wird Brasilien der Klasse „Upper Middle Income“ zugeordnet und zählt daher zu Ländern mit mittlerem Einkommen. Diese Klasse wird definiert nach dem GNI (Bruttoinlandseinkommen) pro Kopf, welches zwischen 4.1US$ und 12.1US$ liegen muss (vgl. weltbank.org 2023).

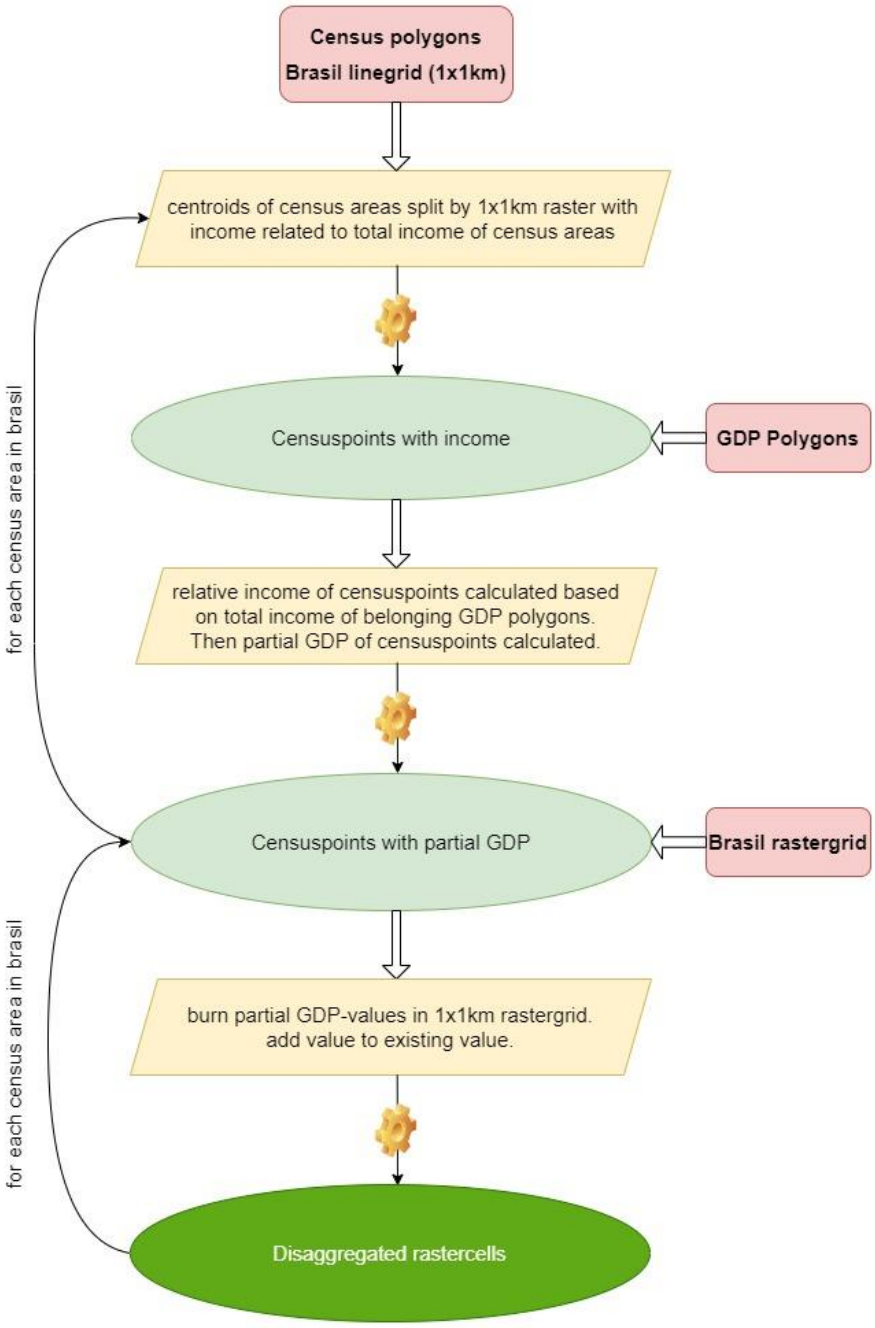


Quelle: Diacon und Maha 2015, S1537

Unter Einbeziehung dieser Erkenntnisse ist es also möglich, das Bruttoinlandprodukt einer Gemeinde nach der Höhe des Einkommens der darin vorkommenden Regionen aufzuteilen. Dabei handelt es sich um eine Annäherung, die gewisse Unsicherheiten birgt, die jedoch bei einer künstlichen Verbesserung der räumlichen Auflösung nicht zu verhindern sind.

2.3.1.1.2 Prozessierung der einkommensbasierten Disaggregation

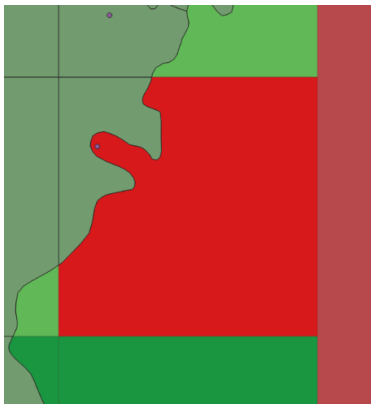
Die Durchführung der einkommensbasierten Disaggregation erfolgt in QGIS, wobei die meisten sequenziellen Prozessierungsschritte mit Hilfe von Python automatisiert wurden. Zu Beginn wird eine leere Rasterdatei mit einer geometrischen Auflösung von 1x1km erstellt. Dieses Raster deckt das gesamte Untersuchungsgebiet ab und definiert die räumliche Einheit des Ziel-Datensatzes. Die Verarbeitungsschritte erfolgen in einer Schleife, die über die einzelnen Bundesländer iteriert. Dadurch werden die Gebiete hintereinander prozessiert, was die Performance deutlich verbessert, da die meisten Analysewerkzeuge von QGIS nicht für große Datensätze ausgelegt sind.

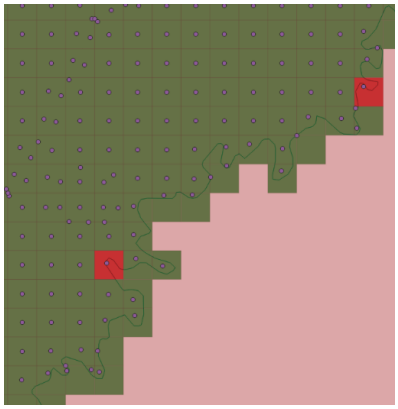


Im ersten Schritt werden die Flächen der Zensus-Polygone berechnet, danach werden sie am Raster zerschnitten. Daraus resultiert, dass die Teilgebiete der Zensus-Regionen als eigenständige Objekte abgespeichert sind. Das Attribut der Fläche des Ursprungspolygons bleibt dabei dennoch erhalten. Dadurch kann die relative Fläche der Teilgebiete in Bezug auf die Ursprungspolygone berechnet werden. Die relative Fläche wird als Attribut gespeichert, da diese für eine spätere Berechnung benötigt wird.

Im nächsten Schritt werden die geometrischen Schwerpunkte (Centroids) der Zensus-Teilregionen berechnet. Hierbei ist es bei der Auswahl des Analysewerkzeugs wichtig, dass

sich die resultierenden Punkt-Geometrien innerhalb der Ausgangspolygone befinden. Liegt der geometrische Schwerpunkt außerhalb des Polygons, soll der Punkt an der nächsten Stelle innerhalb des Polygons gesetzt werden, da es sonst zu einer Verfälschung des Ergebnisses kommen würde. In Abbildung xy ist zu sehen, dass bei der klassischen Schwerpunktberechnung, die Punkte Außerhalb des ursprünglichen Polygons liegen können. Handelt es sich dabei um Kacheln in der Randregion einer Gemeinde, kommt es dadurch zu Fehlern in der weiteren Prozessierung, wenn die Punkte nicht innerhalb des Gemeindepolygons gesetzt werden.

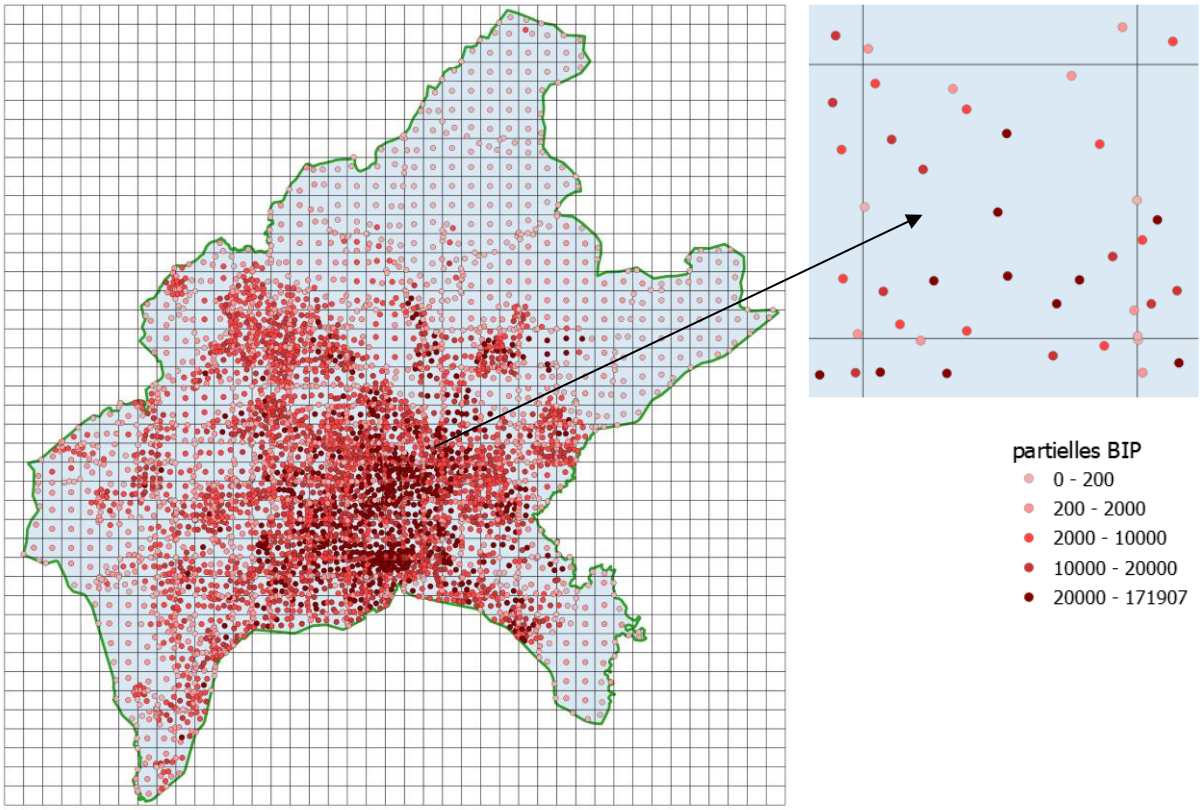




Bei diesem Prozessierungsschritt resultiert ein Punktlayer in dem die Punkte die gleichen Attribute aufweisen, wie in dem Polygonlayer im vorherigen Schritt, also die relative Fläche der Teilregionen. Da diese bekannt ist, kann sie nun mit dem Einkommen des Ausgangspolygons multipliziert werden. Daraus resultiert das absolute Einkommen der Zensus-Teilregionen. Die Zensus-Punkte mit dem Attributwert des absoluten Einkommens, sind das erste Zwischenprodukt in der Prozesskette (Abbildug x).

Im nächsten Prozessierungsschritt soll jedem Zensuspunkt ein partieller BIP-Wert zugewiesen werden. Dafür wird das Einkommen aller Punkte pro Gemeinde addiert, um das Gesamteinkommen jeder Gemeinde zu berechnen. Wie bereits zuvor beschrieben, ist es wichtig, dass sich die Punkte innerhalb der entsprechenden Polygone befinden. Dies gilt vor allem für die Grenzregionen der Gemeinden. Wenn sich die Punkte außerhalb der Gemeinde befinden würden, wäre das aggregierte Gesamteinkommen niedriger als der reale Wert.

Der Attributwert des Gesamteinkommens wird im BIP-Datensatz abgespeichert. Im nächsten Schritt wir ein „Spatial Join“ durchgeführt, der allen Punkten, die sich in einer Gemeinde befinden, das entsprechende Gesamteinkommen anhängt. So kann für jeden Punkt das relative Einkommen in Bezug auf das Gesamteinkommen der Gemeinde berechnet werden. Wird nun dieser Wert mit dem BIP-Wert der Gemeinde multipliziert, resultiert das partielle BIP pro Zensuspunkt. Aus diesem Prozessierungsschritt resultiert das zweite Zwischenergebnis in Abbildung xy (Censuspoints with partial GDP).



Alle bisherigen Schritte werden sequenziell für jedes Bundesland in Brasilien durchgeführt. Zuzüglich des Distritio Federal resultieren somit 27 Dateien, die die Punkte mit den partiellen BIP-Werten beinhalten. Nun kommt die leere Rasterdatei zum Einsatz, indem das partielle BIP auf das Raster übertragen wird. Dabei wird jeder Rasterzelle der Attributwert des enthaltenen Punktes zugewiesen. In urbanen Gebieten, in denen die Zensus-regionen räumlich sehr klein gegliedert sind, befinden sich mehrere Punkte innerhalb einer Rasterzelle. In diesem Fall wird das partielle BIP der Punkte aufsummiert und der Rasterzelle zugewiesen. In Abbildung xy ist die Ausgangssituation dieser Berechnung zu sehen. Innerhalb der grünen Linie, die die Gemeindegrenze der Stadt Goias darstellt, befinden sich die Punkte mit den partiellen BIP-Werten. In der vergrößerten Darstellung ist eine Rasterzelle abgebildet, die mehrere Punkte enthält, deren Werte in diesem Prozessierungsschritt aufsummiert werden.

Da jeder Punkt einen Teilwert der zugehörigen Gemeinde besitzt, verschwimmen somit auch die Gemeindegrenzen. Deckt eine Rasterzelle mehrere Gemeinden ab, so werden die Werte ebenfalls aufsummiert. Dadurch, dass immer die gleiche Rasterdatei bearbeitet wird, besteht auch kein Problem in den Grenzregionen zwischen Bundesländern. Die 27 Dateien, die jeweils einen Punktlayer enthalten, werden nacheinander abgearbeitet. Befindet sich daher ein Punkt in einer Rasterzelle eines schon prozessierten Bundeslandes, wird der Wert einfach zu dem bestehenden dazu addiert.

2.3.1.1.3 Ergebnis der einkommensbasierten Disaggregation

Durch die Ausführung der im vorherigen Kapitel beschriebenen Prozesskette konnte ein Raster-Datensatz generiert werden, der das Bruttoinlandsprodukt von Brasilien in einer Auflösung von 1x1km enthält. Durch die Verteilung des BIP innerhalb der Gemeinden nach der Höhe des Einkommens konnte die räumliche Auflösung der Daten verbessert werden. Im Zuge der Datenprozessierung stellte sich heraus, dass für die Bundesländer Para und Tocantins im Norden des Landes keine Einkommens-Information im Zensusdatensatz vorhanden ist. Die Zensusdaten wurden 2016 von der Webseite des IBGE (ibge.gov.br 2023) heruntergeladen und mit den Geometrien verschnitten. Da die Daten sowohl online als auch nach persönlicher Anfrage beim dortigen Amt nicht mehr verfügbar sind, werden die zwei Bundesländer vom Untersuchungsgebiet ausgeschlossen. Für die übrigen Bundesstaaten konnte ein flächendeckender Datensatz berechnet werden.

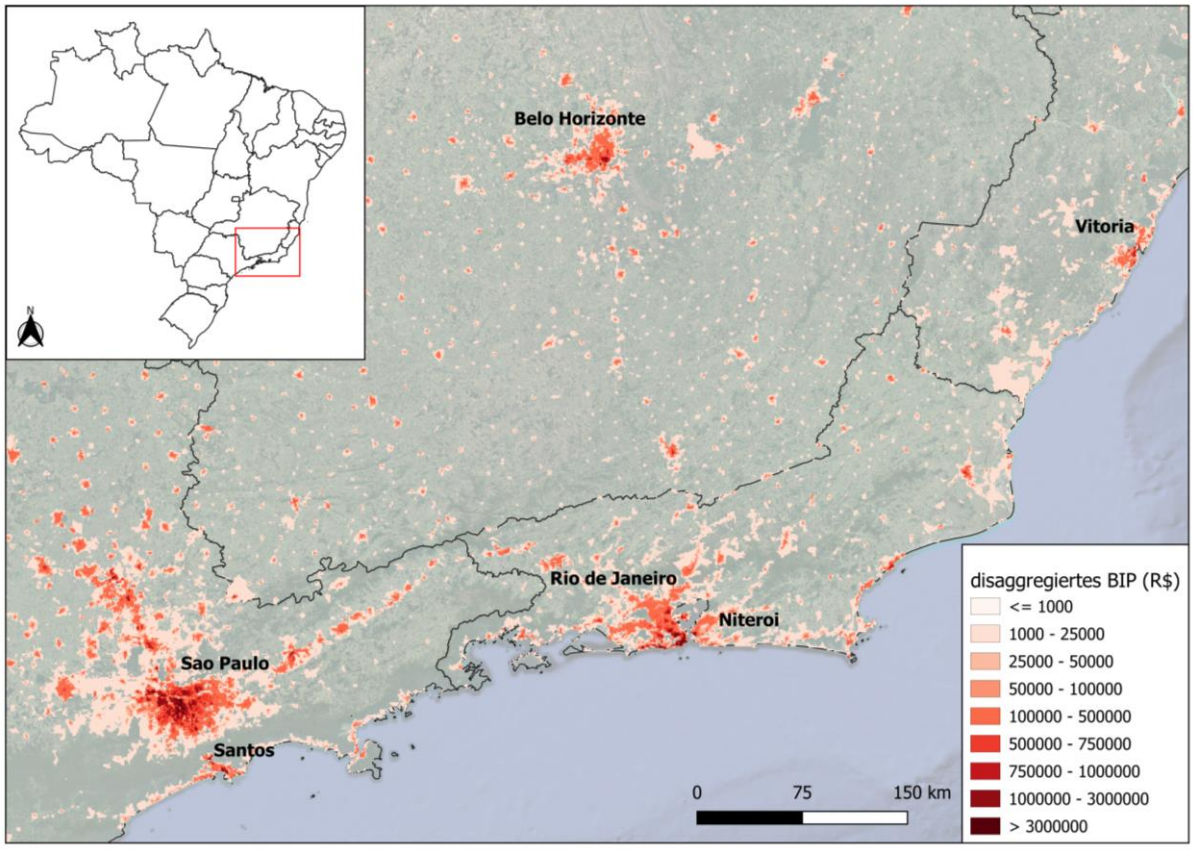


Abbildung xy zeigt einen Ausschnitt der Daten an der Südostküste Brasiliens. Mit knapp über 10 Millionen Einwohnern zählt Sao Paulo zur größten Stadt des Untersuchungsgebiets. Aufgrund der hohen Bevölkerungszahl ist auch das BIP in dieser Region besonders hoch. Im innerstädtischen Bereich sind BIP-Werte von über 3 Millionen Real (~571.000 USD) zu verzeichnen. Der Höchstwert in Sao Paulo und gleichzeitig im gesamten Datensatz liegt bei 6.465.621 Real, was umgerechnet 1.230.445 US-Dollar entspricht (Stand März 2023). Auch die umliegenden Städte zählen zu den bevölkerungsreichsten des Landes. In der Verteilung

des Bruttoinlandprodukts sind große Disparitäten zwischen ruralen und urbanen Räumen zu erkennen. Ein Großteil der Rasterzellen, vor allem im Landesinneren, repräsentiert BIP-Werte von unter 1000 Real, was einer wirtschaftlichen Wertschöpfung von weniger als 190 US-Dollar entspricht (Stand März 2023). Die großen Unterschiede sind der Bevölkerungsverteilung und dem einhergehenden Gesamteinkommen zu verschulden.

2.3.1.1.3 Verifizierung der einkommensbasierten Disaggregation

Um das Ergebnis der Disaggregation zu verifizieren, wird für jedes Bundesland das Bruttoinlandsprodukt wieder aggregiert. Dabei werden die partiellen BIP-Werte der Punkte, die sich innerhalb eines Bundeslands befinden, addiert. Dies betrifft die Punkt-Layer „Censuspoints with partial GDP“ aus Abbildung xy. Da das BIP innerhalb der Bundesstaaten nur in kleinere Raumeinheiten aufgeteilt wurde, dürfte sich der rückaggregierte Wert nicht vom originalen BIP-Wert unterscheiden.

Bundesland	BIP rückaggregiert	BIP original	Abweichung	Genauigkeit
Acre	13.751.125,72	13.751.125,72	0	100%
Alagoas	49.456.361,74	49.456.361,74	0	100%
Amapa	14.338.837,51	14.338.837,51	0	100%
Amazonas	89.017.164,71	89.017.164,71	3,01E-06	100%
Bahia	258.649.048,90	258.649.048,90	0	100%
Ceara	138.378.784,70	138.378.784,70	0	100%
Distritio Federal	235.497.106,60	235.497.106,60	1,01E-06	100%
Espiritio Santo	109.226.783,20	109.226.783,20	0	100%
Goias	181.692.435,10	181.692.435,10	0	100%
Maranhao	85.286.225,65	85.286.225,65	0	100%
Mato Grosso	123.834.253,10	123.834.253,10	1,99E-06	100%
Mato Grosso do sul	91.865.802,62	91.865.802,62	-1,98E-06	100%
Minas Gerais	544.633.968,20	544.633.968,20	0	100%
Paraiba	59.088.985,93	59.088.985,93	0	100%
Parana	401.661.680,70	401.661.680,70	0	100%
Pernambuco	167.289.930,20	167.289.930,20	0	100%
Piaui	41.405.815,50	41.405.815,50	0	100%
Rio de Janeiro	640.185.779,60	640.185.779,60	0	100%
Rio Grande do Norte	59.660.847,34	59.660.847,34	0	100%
Rio Grande do Sul	408.645.099,60	408.645.099,60	0	100%
Rondonia	39.450.586,66	39.450.586,66	-1,01E-06	100%
Roraima	11.011.454,33	11.011.454,33	0	100%
Santa Catarina	256.661.190,00	256.661.190,00	0	100%
Sao Paulo	2.038.004.931,00	2.038.004.931,00	3,10E-06	100%
Sergipe	38.866.963,9	38.866.963,9	0	100%

In Tabelle xy ist zu erkennen, dass bei der Disaggregation der BIP-Werte kein Fehler unterlaufen ist. Die Abweichungen, die in manchen Regionen zu sehen sind, betreffen die sechste Nachkommastelle. Da die Ausgangsdaten jedoch nur zwei Nachkommastellen aufweisen, können diese Fehler als Rundungsfehler ignoriert werden.

Um die Disaggregation auf Gemeindeebene zu überprüfen, werden die Gemeinden der größten 30 Städte Brasiliens als Stichprobendaten rückaggregiert. Gleich wie bei den Bundesstaaten werden die Werte der Punkte, die das partielle BIP enthalten, innerhalb der Grenzen aufsummiert. In dieser Berechnung dienen die Gemeindegrenzen als Abgrenzung der räumlichen Analyse.

BIP original R$	BIP rückaggregiert R$	Abweichung R$	Genauigkeit %
2.084.279.725,57	2.084.279.725,57	0	100%

Das Ergebnis der Rückaggreagation der Punktdaten zeigt eine Genauigkeit von 100%. Das bedeutet, dass auch innerhalb der Gemeindegrenzen die Disaggregation fehlerfrei durchgeführt wurde.

Die Rückaggregation geht aus geometrischen Gründen vom Punktdatensatz aus, da die Punktgeometrien keine Fläche einnehmen und dadurch ein sogenannter „Spatial Join“ durchgeführt werden kann, der die Werte aller Punkte aufsummiert, die sich innerhalb eines Polygons befinden. Würden die Rasterwerte innerhalb der Polygone aufsummiert werden, würden die Rasterzellen an den Polygongrenzen die Genauigkeit stark beeinflussen. Besonders groß ist dieser Effekt bei kleinen Gemeinden mit hohem Bruttoinlandsprodukt. Je nach Definition der Zugehörigkeit einer Rasterzelle zu einem Polygon (Flächenanteil, Flächenschwerpunkt, Boolean), wird dem Polygon ein unkorrekter Wert zugewiesen. Die einzige Methode die ein akkurates Ergebnis ausgehend von Rasterdaten erzielen würde, wäre eine flächengewichtete Zuweisung des BIP jener Rasterzellen, die die Gemeindegrenzen schneiden. Da der Punktlayer exakt diese Information enthält, wird dieser zur Validierung der Disaggregation verwendet (siehe Abbildung xy in Kapitel 2.3.1.1.2).

Um die Gesamtgenauigkeit des Rasterdatensatzes von ganz Brasilien zu überprüfen, werden die Rasterwerte jedes Pixels des Datensatzes aufsummiert. Da die Bundesstaaten Para und Tocantins nicht disaggregiert werden konnten, wird für die Verifizierung das BIP dieser Bundesstaaten zur Summe der Rasterwerte addiert. So kann der rückaggregierte Wert mit dem Originalwert verglichen werden.

BIP original R$	BIP rückaggregiert R$	Abweichung R$	Genauigkeit %
6.267.205.002,69	6.267.209.535,83	4533,14	99,993

Wie in Tabelle xy zu erkennen, weicht der rückaggregierte Wert um 4533.14 R$ vom originalen Bruttoinlandsprodukt ab. Dieser Fehler entsteht beim letzten Prozessschritt der Disaggregation, bei dem die Punktwerte in das 1x1km Raster übertragen werden. Da es sich dabei um eine Abweichung von nur 0.007% handelt, kann der Fehler vernachlässigt werden. Somit sind die Referenzdaten für das Machine Learning Modell validiert und können als Datensatz zur Implementierung des Modells verwendet werden.

2.3.1.2 Sentinel-2 Mosaik

Der Prozess der Mosaikbildung beginnt mit der Auswahl repräsentativer Testgebiete, um auf Basis deren Phänologie die jahreszeitlichen Zeiträume festzulegen, in denen die Aufnahmen überlagert werden können. Im nächsten Schritt wird eine Zeitreihenanalyse durchgeführt, um die Monate zu definieren, in denen einerseits der Bewölkungsgrad am geringsten ist und andererseits der Einfluss phänologischen Effekte am niedrigsten ist. Nach der Festlegung der Zeitspannen wird eine Wolkenmaske auf die ausgewählten Szenen angewandt, um die Restbewölkung so gut wie möglich zu entfernen.

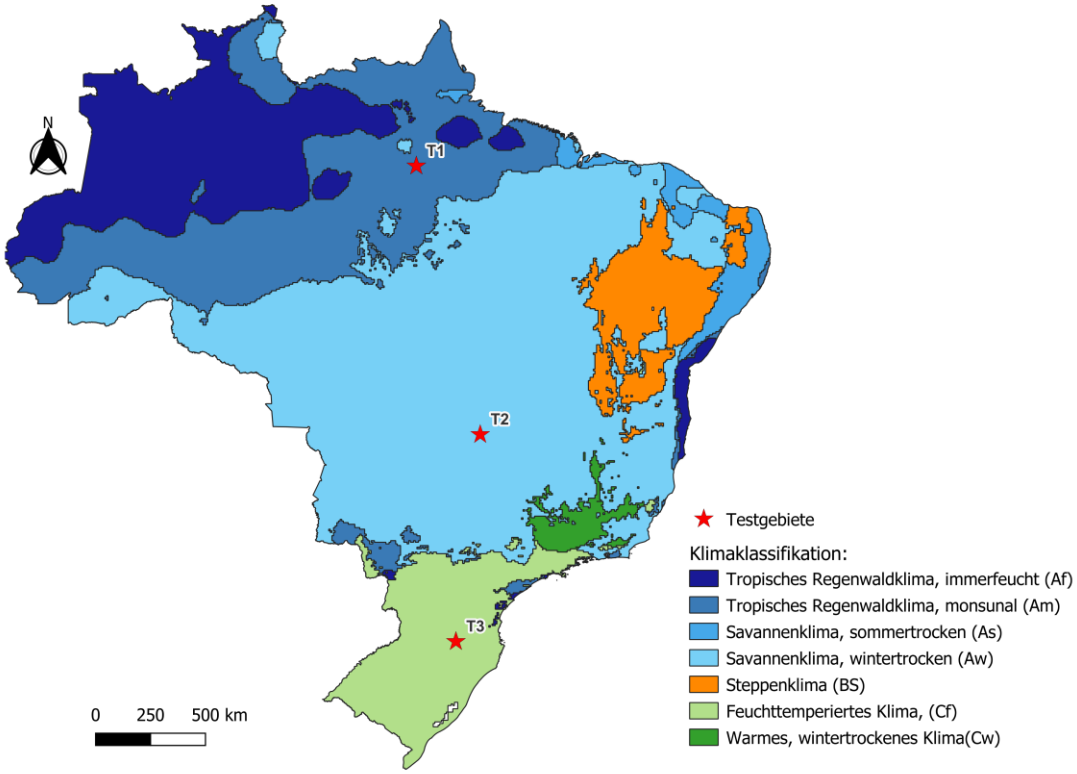
Der Sentinel-2 Datensatz für diese Arbeit wird von der Google Earth Engine bezogen, die als Webanwendung den Zugriff auf Satellitendaten ermöglicht. Dabei dient der Earth Engine Code Editor als webbasierte Entwicklungsumgebung. In dieser Entwicklungsumgebung ist es möglich, mit Hilfe von JavaScript auf Datenbanken zuzugreifen und Datensätze direkt im Editor zu bearbeiten. Die Daten können manipuliert, visualisiert und heruntergeladen werden. Durch vorgefertigte Module können Standardanalysen und Berechnungen an den Daten in wenigen Code-Zeilen durchgeführt werden. Der Code Editor dient daher als Schnittstelle zwischen User und Geodatenbank.

Bei der Datenabfrage wird gefiltert nach Sensor, Bewölkungsgrad und Aufnahmedatum. Die Bezeichnung des Datensatzes, welcher die Sentinel-2 Daten beinhaltet, lautet „Copernicus/S2_SR“. Dabei handelt es sich um Level-2A Daten, die bereits atmosphärisch korrigiert wurden. Dies bedeutet, dass Streuungen, die nicht den direkt reflektierten Lichtstrahlen entsprechen, bereits rechnerisch bereinigt sind. Streuungen werden beispielsweise verursacht durch Luftmoleküle, atmosphärische Gase (z.B. Ozon), Wasserdampf und Aerosole. Level-2A Daten sind auf globaler Ebene seit Dezember 2018 verfügbar. Die Kacheln, in denen die Satellitenbilder verfügbar sind, decken jeweils eine Fläche von 110x110km auf der Erdoberfläche ab (vgl. sentinels.copernicus.eu 2023a). Um die vorliegenden Kacheln in ein homogenes Satellitenbild zu verwandeln, muss ein Mosaik erstellt werden. Bei der Akquisition von Geodaten ist das Datenmanagement ein wichtiger Prozess, der schon vor dem Download der Daten beginnt. Im Fall der Setinel-2 Daten müssen Raumeinheiten definiert werden, in denen die Mosaikbildung die Hardware- und Softwareressourcen nicht übersteigt. Eine Sentinel-2 Szene, mit einer Ausdehnung von 100x100km, besitzt ein Datenvolumen von 600MB (vgl. ESA 2015). Wird das Datenvolumen auf die Fläche von Brasilien hochgerechnet, ergibt dies eine Dateigröße von rund 511GB. Da die Google Earth Engine nur limitierte Dateigrößen verarbeiten kann, und die Satellitenbilder von Softwareprogrammen lesbar und bearbeitbar sein sollen, wird das Untersuchungsgebiet für den Download aufgeteilt in die Bundesstaaten. Diese werden, je nach Größe, weiter unterteilt in insgesamt 42 Regionen. Die Mosaike der Teilgebiete können anschließend problemlos in einer Desktop-GIS Anwendung zusammengefügt werden.

2.3.1.2.1 Festlegung der Testgebiete - Klimaklassifikation

Im ersten Schritt der Datenerhebung für das Sentinel-2 Mosaik muss ein Zeitraum definiert werden, in dem die Satellitenbilder überlagert werden können. Aufgrund der Abschattung durch Bewölkung müssen meistens mehrere Szenen derselben Region kombiniert werden, um ein einheitliches Bild zu erhalten. Ein wichtiger Faktor ist dabei die Phänologie, die die jahreszeitlichen Veränderungen in der Vegetation beschreibt. Bei der Überlagerung von Satellitenbildern ist es wichtig, dass dasselbe Objekt in unterschiedlichen Aufnahmen einen ähnlichen Spektralwert aufweist. Um dies gewährleisten zu können, darf die Zeitspanne, in der die Bilder kombiniert werden, nicht zu groß sein. Um den Datenbestand zu vergrößern, können Aufnahmen aus unterschiedlichen Jahren verwendet werden.

Um die Phänologie im Untersuchungsgebiet zu untersuchen, wurden drei Testgebiete definiert. Diese Testgebiete sollen so gut wie möglich die klimatischen Zonen im Untersuchungsgebiet abdecken.



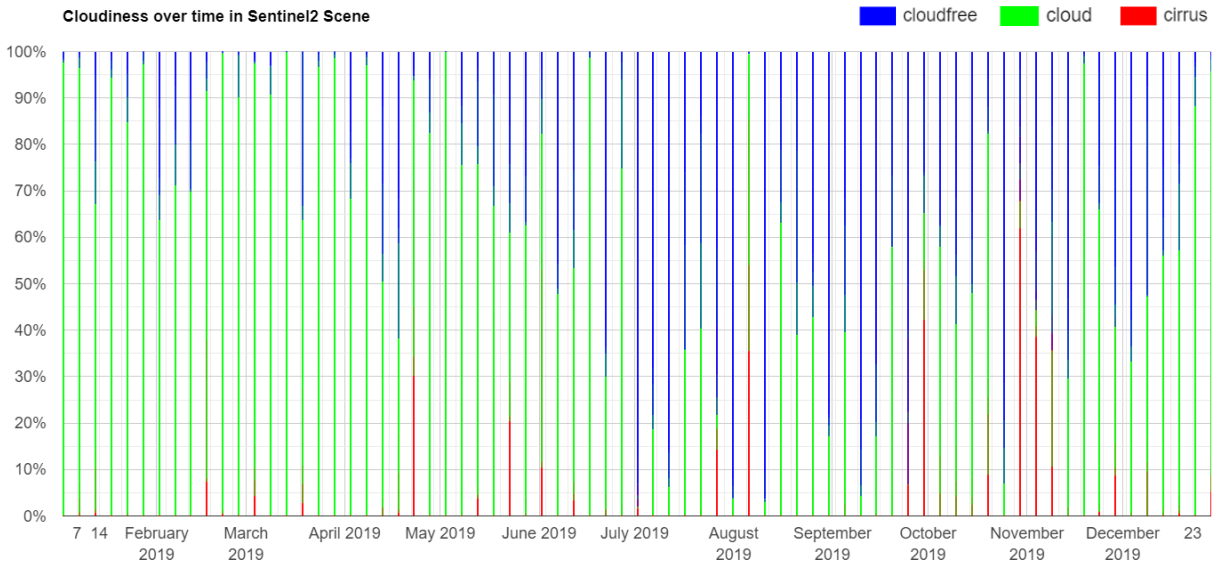
Da Brasilien eine Fläche von 8.515.664 Quadratkilometern abdeckt, beinhaltet das Untersuchungsgebiet mehrere Klimazonen. In Abbildung xy ist eine Klimaklassifikation nach Köppen zu sehen. Die Daten repräsentieren das Klima, berechnet aus dem Zeitraum von 1986-2010, in einer geometrischen Auflösung von 5 Bogenminuten. Als Testgebiete für das Sentinel-2 Mosaik wurden drei Regionen gewählt, die in Nord-Süd Richtung im Untersuchungsgebiet verteilt sind. Das nördlichste Testgebiet T1 befindet sich im monsunalen

tropischen Regenwaldklima (Am). Tropische Regenklimate werden dadurch definiert, dass das Temperatur-Monatsmittel des kältesten Monats mehr als 18 Grad-Celsius beträgt (vgl. Gebhart et al. 2016). Das Monsunklima zeichnet sich durch die jahreszeitlichen Schwankungen in der Niederschlagsmenge aus. In der Regel gibt es eine Trocken- und eine Regenzeit. Das Testgebiet T2 befindet sich ungefähr in der Mitte des Landes und deckt die Stadt Goias und die umliegende Landschaft ab. Das Testgebiet liegt im wintertrockenen Savannenklima (Aw). Diese Klimazone schließt an den tropischen Regenwald an und ist geprägt von Regen- und Trockenzeiten, flachen Relief und tief liegendem Grundwasser. Das bedeutet, dass der Boden in den Wintermonaten stark austrocknet, was zu karger Vegetation und Erosion führt (vgl. Geographie Infothek 2004) . Das dritte Testgebiet T3 liegt im Süden des Landes, was dem feuchttemperierten Klima (Cf) zugeordnet wird. Der Klimatyp zählt zu den warmgemäßigten Klimaten, in denen der kälteste Monat zwischen +18 und -3 Grad-Celsius liegt. Ein weiteres Kennzeichen ist die ganzjährig gleichbleibende Niederschlagsintensität.

2.3.1.2.2 Zeitreihenanalyse

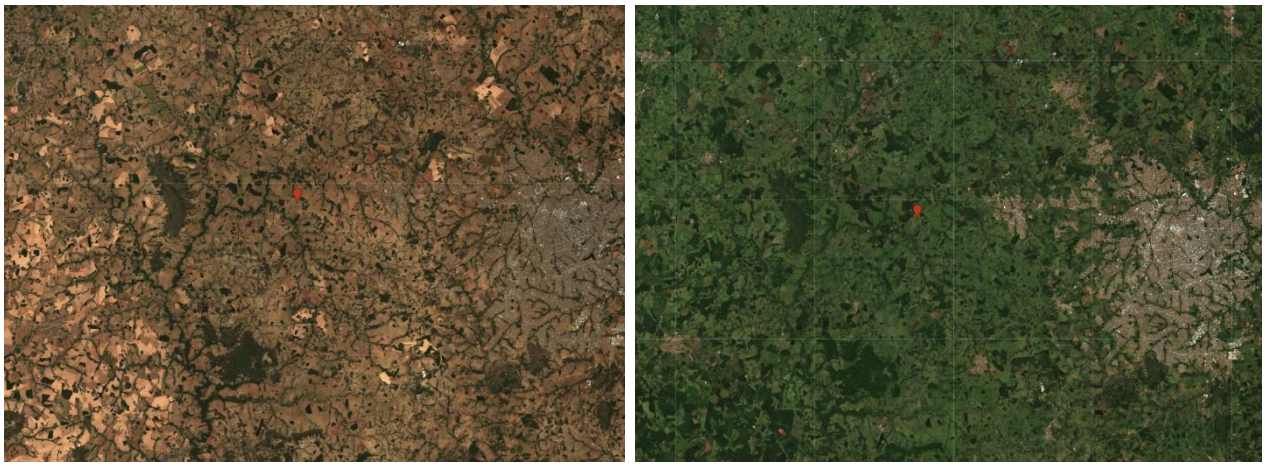
Die Klimazonen wirken sich auf die jahreszeitlichen Schwankungen in der Bewölkung, sowie die Änderungen in der Phänologie aus. Die Jahreszeiten in Brasilien unterscheiden sich grundlegend zu denen in Europa, da sie genau umgekehrt ausfallen. Aus diesem Grund wurde für jedes Testgebiet für jeden Monat des Jahres 2019 ein Mosaik berechnet, um eine Zeitreihenanalyse durchführen zu können. Im nächsten Schritt wurden die Bilder visuell miteinander verglichen, um herauszufinden, in welchen Monaten sich die Vegetation am wenigsten verändert. Um die Schwankungen in der Bewölkung zu analysieren, wurde mit Hilfe des „QA60“-Bands in der Google Earth Engine für jedes Testgebiet ein Diagramm erstellt, was den Bewölkungsgrad im Jahresverlauf visualisiert.

Beim Testgebiet 1, welches sich im monsunalen Regenwaldklima befindet, besteht eine große Herausforderung darin, einen flächendeckenden Datensatz zu erhalten. Vor allem in den Monaten von Dezember bis Juni kann kein flächendeckendes Mosaik erstellt werden. Der Grund dafür liegt in der Bewölkung, die sich speziell in dieser Region als besonders anhaltend aufweist. Aufgrund der Monsunzeit führt die hohe Luftfeuchtigkeit zu Wolkenbildung, was eine Abtastung der Erdoberfläche verhindert. Ein flächendeckendes Mosaik ist in dieser Region nur in den Monaten von Juli bis November möglich. In diesem Testgebiet wurde neben der visuellen Auswertung, auch eine quantitative Auswertung durchgeführt, um den Bewölkungsgrad im Testgebiet zu untersuchen (vgl. Gärtner 2020).



Die Abbildung xy zeigt alle Sentinel-2 Szenen, die das Testgebiet T1 im Jahresverlauf von 2019 vollständig oder teilweise abdecken. In den Monaten von Dezember bis Juni sind die Aufnahmen vor allem durch undurchsichtige Wolken verdeckt. Cirrus-Wolken kommen nur vereinzelt, am häufigsten im Oktober und November vor. Für die Mosaikbildung eignen sich nach der quantitativen Auswertung die Monate von Juli bis November am besten. In dieser Zeitspanne darf auch der Bewölkungsgrad bei der Datenfilterung nicht zu niedrig sein, da sonst zu wenig Szenen für das Mosaik vorhanden sind. Durch die Auswertung aller drei Testgebiete wurde ein maximaler Bewölkungsgrad von 20% festgelegt. Dies bedeutet, dass nur Szenen, in denen weniger als 20% der Pixel durch Bewölkung verdeckt werden, aus der Datenbank „Copernicus/S2_SR“ importiert werden. Die Wolken-Pixel, die sich dennoch in den Szenen befinden, werden mit Hilfe einer Wolkenmaske vor dem Überlagern ausgeschnitten.

Im Testgebiet T2 sind die Sentinel-2 Szenen von Juni bis September fast ausschließlich wolkenfrei. Der Grund dafür liegt in der Trockenzeit, die in den dortigen Wintermonaten vorherrscht. Aufgrund der unterschiedlichen Niederschlagsmengen im Jahresverlauf kommt es zu großen Unterschieden in der Phänologie.



Die Abbildung xy zeigt das Testgebiet T2 im August (links) und im Januar (rechts). Aufgrund des tief liegenden Grundwasserspiegels und der Dürre, trocknen nahezu alle Grünflächen über die Wintermonate (Juli-September) aus. Ausgehend von der Veränderung der Phänologie könnten beide Jahreszeiten zur Mosaikbildung verwendet werden. Da jedoch in den Wintermonaten nahezu wolkenfreie Bedingungen herrschen, werden diese für die Prozessierung bevorzugt.

Das Testgebiet T3 wird vom feuchttemperierten Klima beeinflusst und unterliegt ganzjährigen Niederschlagsereignissen. Aus diesem Grund kann auch nach der quantitativen Auswertung der Wolkenbedeckung kein begünstigter Zeitraum definiert werden, in dem die Bedeckung durch Wolken minimal ausfällt. Durch visuelle Interpretation der Zeitreihendaten eignen sich jedoch die Monate von Juni bis September am besten, da sich in diesen Monaten die Vegetation am wenigsten verändert.

Tabelle: geringster Bewölkungsgrad

	Jan	Feb	Mar	Apr	Mai	Jun	Jul	Aug.	Sept	Okt	Nov	Dez
T1												
T2												
T3												

Tabelle: geringste phänologische Unterschiede

	Jun.	Jul.	Aug.	Sept.	Okt.
T1					
T2					
T3					

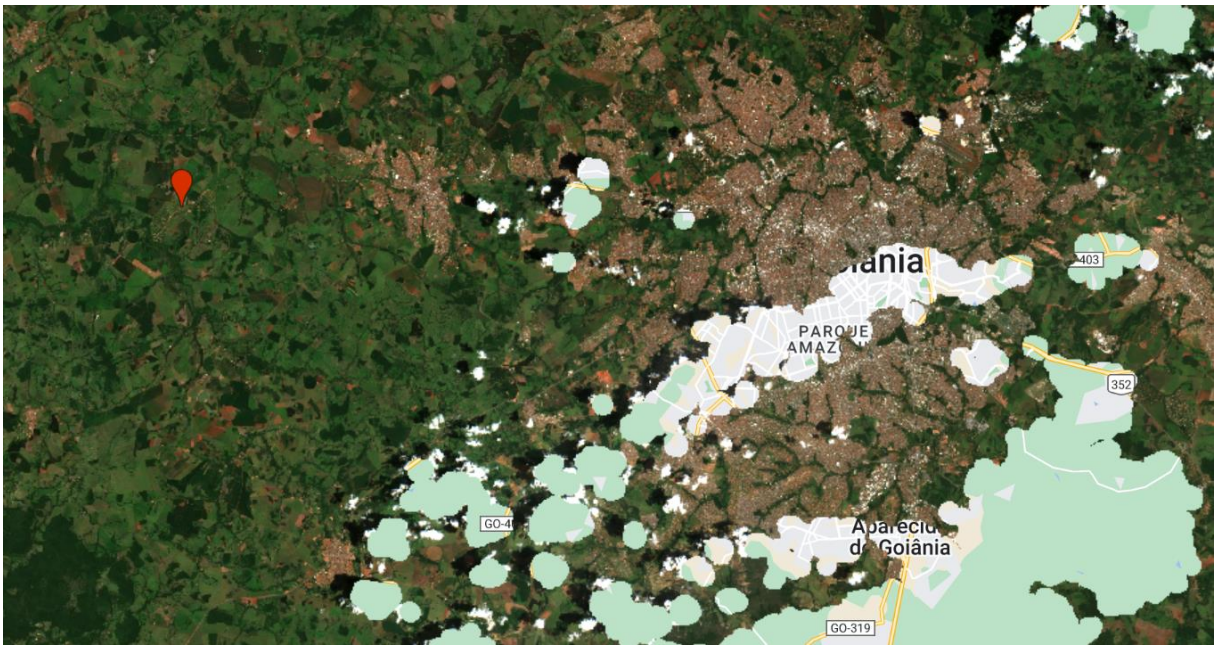
Abschließend wurde ein Zeitraum festgelegt, in dem alle drei Gebiete sowohl einen flächendeckenden Datenbestand aufweisen als auch die phänologischen Effekte so gering wie möglich sind. Die Monate, die in allen drei Testgebieten im Überlappungsbereich liegen, sind August und September. Da die Level-2 Daten der Sentinel-2 Mission erst seit Dezember 2019 öffentlich zugänglich sind, werden die aktuellsten Jahre 2019, 2020 und 2021 zur Datenprozessierung gewählt. Durch dieses Vorgehen liegt für jeden Ausschnitt des Untersuchungsgebiets ein möglicher Datenbestand von 6 Monaten vor, aus dem die Aufnahmen miteinander kombiniert werden können, um ein möglichst wolkenfreies Mosaik zu generieren.

2.3.1.2.3 Wolkenmaske

Um Bewölkung in Fernerkundungsdaten zu detektieren und die betroffenen Pixel zu maskieren, dienen sogenannte Wolkenmasken. Bei Sentinel-2 Daten wird im Level-1C der Prozess-Pipeline eine Wolkenmaske berechnet. Die Pixelinformation über das Vorkommen und die Art der Bewölkung kann aus dem Band „QA60“ der Level-2A Daten ausgelesen werden.

Die Berechnung der Wolkenmaske erfolgt aus dem blauen Band (B1) und den kurzwelligen Infrarot-Bändern B10 und B11. Dabei wird ein Resampling der Daten auf eine geometrische Auflösung von 60m durchgeführt, bevor durch spektrale Kriterien die Wolken detektiert werden (vgl. sentinels.copernicus.eu 2023). Dem QA60- Band wird pixelweise ein Wert zugewiesen, der die Unterscheidung zwischen Cirrus-Wolken (2), undurchsichtigen Wolken (1), und keiner Wolkenbedeckung (0) ermöglicht. Bei der Datenerhebung in der Google Earth Engine kann eine Funktion geschrieben werden, die nur die Pixel eines Bildes liefert, bei denen der Wert im QA60-Band gleich 0 entspricht.

Die Abbildung xy zeigt eine Sentinel-2 Szene des Testgebiets T2 im Bereich der Stadt Goiania. Der maximale Bewölkungsgrad dieser Szene liegt bei 10%, wobei sich der Prozentsatz auf die jeweilige 100x100km Kachel der Rohdaten bezieht. Obwohl ein Bewölkungsgrad von 10% relativ niedrig ist, wird in dieser Szene ein großer Teil des Stadtgebiets von Wolken verdeckt. Da nicht bekannt ist, welche Gebiete innerhalb der Szene von Wolken bedeckt sind, muss ein Mittelmaß zwischen möglichst niedriger Wolkenbedeckung und flächendeckendem Datenbestand gefunden werden. Die Wolkenmaske ermöglicht es, auch Szenen mit höherem Bewölkungsgrad in die Mosaik-Berechnung miteinzubeziehen.

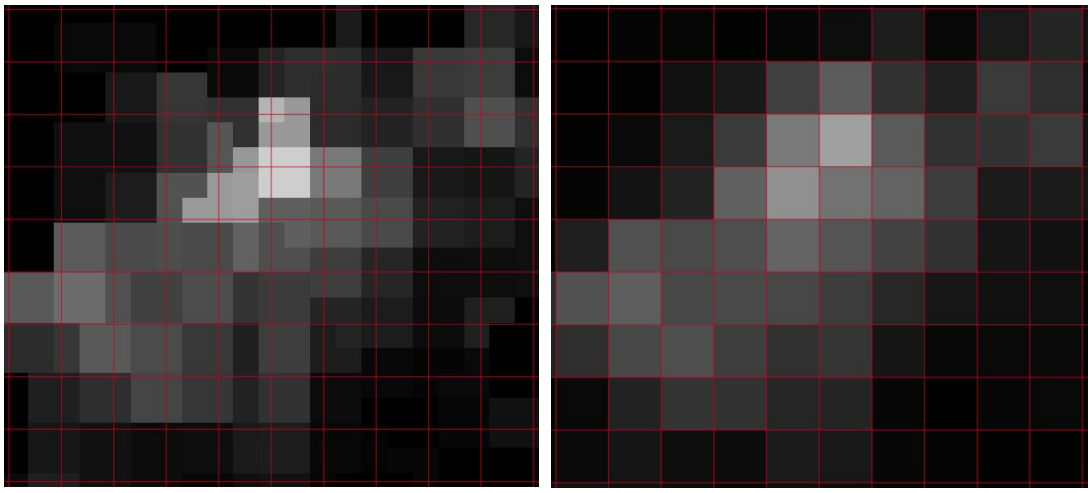


2.3.1.2.4 Mosaikbildung

Das Mosaik wird durch die Berechnung des Medians der übereinanderliegenden Pixelwerte berechnet. Eine weitere Möglichkeit wäre die Berechnung des arithmetischen Mittels der Pixelwerte. Da der Median jedoch resistenter gegenüber Ausreißern in den Daten ist, wurde dieser zur Berechnung des Mosaiks verwendet. In Abbildung xy ist die Funktion der Wolkenmaske in einem Teilgebiet des Testgebiets T2 zu erkennen. Die linke Abbildung zeigt eine Sentinel-2 Szene, in der eine Wolke herausmaskiert wurde. Die mittlere Abbildung zeigt ein Mosaik aus der ersten Szene mit einer zweiten im selben Monat (Januar 2019). Es ist zu sehen, dass die Pixelwerte der zweiten Szene das Loch füllen, was bei der Maskierung der ersten Szene entstanden ist. Weiters ist zu erkennen, dass bei den Pixeln, die sich überlagern, der Mittelwert gebildet wurde. Bei zwei Werten berechnet sich der Median nämlich aus dem Mittelwert der zwei Werte. Es fällt auf, dass der Wolkenschatten im mittleren Bild weniger intensiv vorhanden ist. Außerdem sind die Wolkenpixel, die durch die Maske nicht erkannt wurden durch die Medianbildung aufgehellt und wirken durchsichtiger. Bei einem Mosaik aus drei unterschiedlichen Szenen (rechte Abbildung), werden die Ausreißer in der ersten Szene durch die Medianbildung entfernt und es entsteht ein nahezu wolkenfreies Mosaik. Je mehr Bilder miteinander kombiniert werden, desto weniger Wolkenpixel sind im Endprodukt vorhanden.

2.3.1.3 Resampling der Zusatzdaten

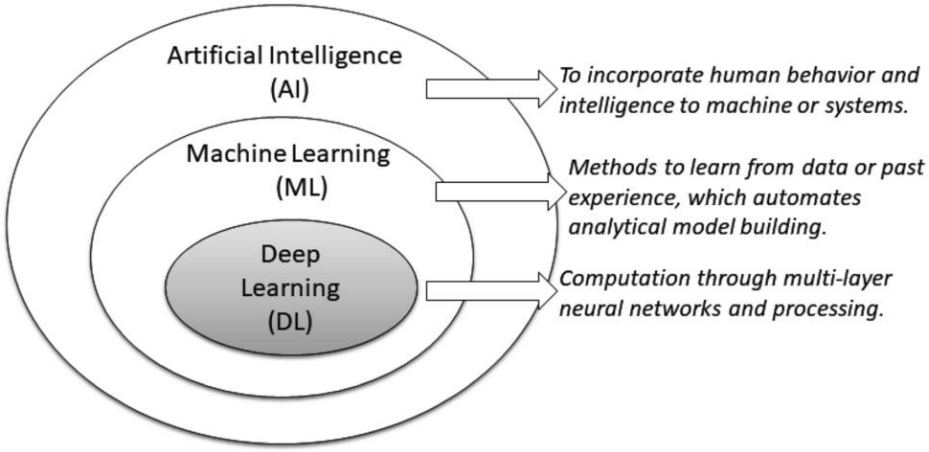
Für das Machine Learning Modell müssen alle Geodaten in einer einheitlichen geometrischen Auflösung vorliegen. Die Auflösung wird durch das Raster der disaggregierten BIP-Werte vorgegeben. Die Rasterdaten des NDVI, NDBI und MNDWI entsprechen einer geometrischen Auflösung von 500x500m. Die Black Marble Daten stehen in einer geometrischen Auflösung von 15 Bogensekunden zu Verfügung. Im Bereich des Äquators entspricht dies ebenfalls einer Auflösung von knapp 500x500m. Um die Informationen in ein 1x1km Raster zu transformieren, werden die Daten in eine größere Raumgliederung gebracht. Dabei werden die Ausgangsdaten, die sich in einer Rasterzelle der Ziel-Raumgliederung befinden, nach Flächenanteil aggregiert. Das Resultat des Resamplings sind Rasterdateien, die in einer geometrischen Auflösung von 1x1km die Werte von NDVI, NDBI, MNDWI und Black Marble enthalten. In Abbildung xy ist das Resampling der Black Marble Daten in einem östlichen Stadtteil von Sao Paulo zu sehen. Auf der linken Seite sind die Ausgangsdaten dargestellt, während rechts der fertig aufbereitete Black Marble Datensatz abgebildet ist.



2.3.2 Machine Learning

Machine Learning ist eine Unterkategorie der künstlichen Intelligenz, in der durch Anwendung von Algorithmen bestimmte Muster in Daten erkannt werden können. In der Fernerkundung wird Machine Learning zum Beispiel zur Erkennung von Objekten auf der Erdoberfläche oder zur Klassifikation von Landbedeckungsarten genutzt. Ein häufig genutzter Algorithmus ist hierbei Random Forest, der auf zufällig generierten Entscheidungsbäumen basiert, deren Kombination eine Klassifikation von Daten ermöglicht. Der Begriff Machine Learning kann noch weiter untergliedert werden in die Kategorie Deep Learning. Diese Art des maschinellen Lernens basiert auf künstlichen neuronalen Netzwerken (ANN) deren Funktionsart an das menschliche Gehirn erinnert. ANNs werden nicht nur in den Geowissenschaften angewandt, sondern finden auch einen großen Anwendungsbereich im Gesundheitswesen, in der Bild- und Texterkennung, sowie in Bereichen der „Cybersecurity“ und vielen mehr (vgl. Sarker 2021). Neuronale Netzwerke bestehen aus einem Input-Layer, mehreren sogenannten „Hidden Layern“, sowie einem Output Layer. Alle Layer beinhalten eine Vielzahl an Knotenpunkten (Neuronen), die miteinander verbunden sind.

Bildquelle: (Sarker 2021)



Die Abbildung xy zeigt die Hierarchie der Begrifflichkeiten, die mit künstlicher Intelligenz in Verbindung stehen. Der Überbegriff „Artificial Intelligence“ beschreibt den Prozess, Maschinen menschliche Eigenschaften anzulernen, um durch die Vorteile von computerbasierten Berechnungen bestimmte Prozesse zu automatisieren. Machine Learning und Deep Learning fallen unter den Begriff der künstlichen Intelligenz, stellen aber eher unterschiedliche Methoden zur Anwendung von künstlicher Intelligenz dar. Machine Learning basiert dabei auf simplen analytischen Modellen wie dem oben genannten Random Forest Algorithmus. Deep Learning hingegen befasst sich mit komplexen mathematischen Modellen in neuronalen Netzwerken (vgl. Sarker 2021). Diese Arbeit bezieht sich auf zwei unterschiedliche Deep Learning Methoden, dem Convolutional Neural Network (CNN) und dem Multilayer Perceptron (MLP).

Multilayer Perceptron:

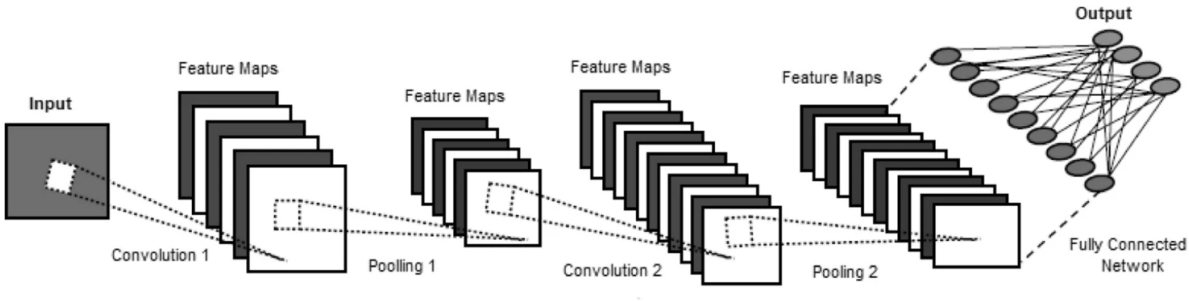
Das Multilayer Perceptron (MLP) gehört der Klasse der Artifictial Neural Networks (ANN) an. Es besteht aus einem Input Layer, einem oder mehreren Hidden Layern und einem Output Layer. In den Hidden Layern werden mathematische Operationen auf die Inputdaten angewandt. Dabei wird auch eine Aktivierungsfunktion (activation function) ausgeführt, die die Erkennung von nicht linearen Zusammenhängen zwischen Input und Outputdaten ermöglichen soll. Die bekannteste ist die sogenannte „ReLU“ Funktion, die auch im Zuge dieser Arbeit angewandt wird. Diese Funktion setzt alle negative Werte gleich 0, alle positiven Werte bleiben unverändert. Jeder Knoten eines Layers führt eine Aktivierungsfunktion aus, bei der der Output für den nächsten Layer generiert wird.

Das Multilayer Perceptron ist ein vollständig verbundenes neuronales Netzwek. Das bedeutet, dass jedes Neuron eines Layers mit jedem Neuron des vorherigen und des darauffolgenden Layers verbunden ist. Die Verbindungen zwischen den Knoten sind dabei gewichtet. Die Gewichtung der Knoten erzielt einen gewissen Output, der mit den Referenzdaten verglichen wird, wobei das Ziel verfolgt wird die Abweichung zu minimieren. Dafür wird die überwachte Lernmethode „Backpropagation“ angewandt. Hierbei werden die Gewichte ausgehend vom Outputlayer in Richtung des Inputlayers derart verändert, dass sich bei der nächsten Forward Propagation die Abweichung verringert. Dieser Prozess wird iterativ über mehrere Epochen durchgeführt (vgl. Alzubaidi et al. 2021). Es handelt sich daher um eine überwachte Lernmethode, was bedeutet, dass ein Trainingsdatensatz, sowie ein Referenzdatensatz vorhanden sein muss. Das Modell wird trainiert, um aus den Inputdaten die Referenzdaten vorauszusagen. Anschließend kann das Modell auf unbekannte Daten angewandt werden.

Convolutional Neural Network:

Das Convolutional Neural Network (CNN) gehört ebenfalls zu den Artificial Neural Networks (ANN). Dieses neuronale Netzwerk kommt in erster Linie im Bereich der Bilderkennung zum Einsatz, da es automatisch lernt, Strukturen in Bilddateien zu erkennen (vgl. Alzubaidi et al. 2021). Es besteht aus drei Typen von Layern: dem Convolutional Layer, dem Pooling Layer und dem Fully Connected Layer. Der erste Layer im Modell ist der Convolutional Layer, in dem durch Anwendung von Kernel-Filtern gewisse Strukturen (Features) erkannt werden. Diese Strukturen werden anschließend in Feature Maps gespeichert. Im Pooling Layer werden wird die Größe der Feature Maps verringert, wofür unterschiedliche Pooling Operationen durchgeführt werden können. Wie in Abbildung xy zu sehen, kann im Zuge der Feature Extraktion eine Serie von Convolutional und Pooling Layern ausgeführt werden. Im Allgemeinen erkennt der erste Convolutional Layer grobe Strukturen, die in den weiteren Convolutional Layern immer weiter zerteilt werden.

Bildquelle: (Sarker 2021)



Der dritte Layer-Typ ist der Fully Connected Layer, der der Struktur eines Layers in einem Multilayer Perceptron gleicht. In diesem Layer sind alle Neuronen mit allen Neuronen des vorherigen Pooling Layers verbunden. Der Output des letzten Layers entspricht entweder einer Klassifikation, also einem kategorischen Wert, oder einem numerischen Wert (Regression). In dieser Arbeit wird das Bruttoinlandsprodukt modelliert, der Output entspricht daher einem numerischen Wert.

2.3.2.1 Datensatz Vorbereitung

Für das Machine Learning Modell müssen die vorbereiteten Geodaten nun in eine Struktur gebracht werden, die das Modell verarbeiten kann. Da das gesamte Untersuchungsgebiet zu viel Datenvolumen einnehmen würde, werden die Daten auf kleinere Gebiete zugeschnitten. Als Untersuchungsgebiete dienen die Gemeindegrenzen der größten Städte Brasiliens. Die Städte Belem, Palmas und Santarem müssen dabei ausgeschlossen werden, da sie sich in den Bundesstaaten Para und Tocantins befinden, von denen keine Referenzdaten vorliegen. In Abbildung xy ist zu erkennen, dass einige der Städte an der Ostküste Brasiliens liegen. Im Landesinneren befinden sich jedoch auch bevölkerungsreiche Städte, was zu einer gleichmäßigen Verteilung der Untersuchungsgebiete innerhalb des Landes beiträgt. Die Städte im Norden Brasiliens liegen an den Ufern großer Flussläufe. Die Verkehrsinfrastruktur ist in diesem Teil des Landes sehr begrenzt, da weite Teile von tropischen Regenwald bedeckt sind.

Die Städte in Abbildung xy sind nach Bevölkerungszahlen geordnet, beginnend mit Sao Paulo mit über 10 Millionen Einwohnern. Die kleinste Stadt im Datensatz ist Boa Vista mit rund 235.000 Einwohnern. Die Schwankungsbreite in den Einwohnerzahlen ist demnach sehr hoch.

1	Sao Paulo
2	Rio de Janeiro
3	Salvador
4	Fortaleza
5	Belo Horizonte
6	Brasilia
7	Curitiba
8	Manaus
9	Recife
10	Porto Alegre
11	Goiania
12	Maceio
13	Sao Luis
14	Natal
15	Teresina
16	Campo Grande
17	Joao Pessoa
18	Cuiaba
19	Aracaju
20	Niteroi
21	Florianopolis
22	Santos
23	Macapa
24	Vitoria
25	Porto Velho
26	Rio Branco
27	Boa Vista

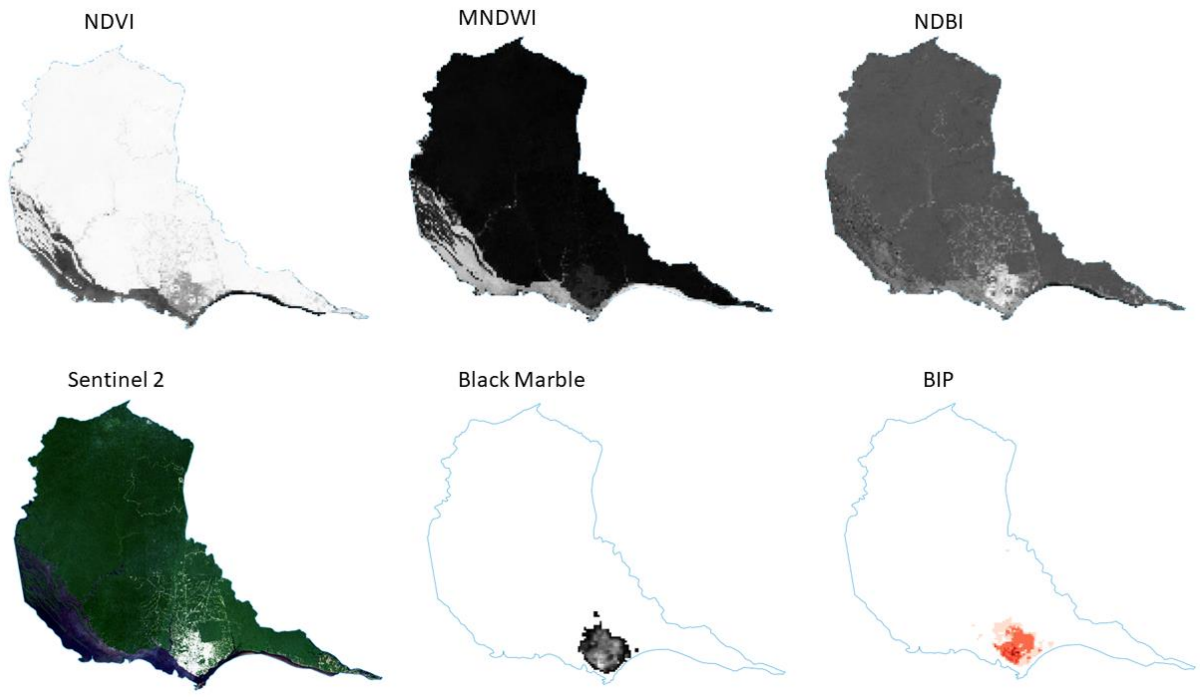
Als Untersuchungsgebiet wurden die größten Städte gewählt, da das Bruttoinlandsprodukt nach der Höhe des Einkommens disaggregiert wurde und das Einkommen in bevölkerten Gebieten deutlich höher ausfällt als in dünn besiedelten Regionen. Das Bruttoinlandprodukt, welches durch die definierten Untersuchungsgebiete abgedeckt wird, beläuft sich auf 2.084.279.726 R$. Der gesamte BIP-Datensatz beträgt 6.267.205.003 R$. Dies bedeutet, dass die 27 ausgewählten Gemeinden rund ein Drittel des Bruttoinlandprodukts von ganz Brasilien repräsentieren. Die Städte in Tabelle xy sind nach den Einwohnerzahlen sortiert. Dabei fällt auf, dass die Anzahl der Einwohner nicht immer mit der Höhe des BIPs korreliert. Die Hauptstadt Brasilia steht an sechster Stelle, wobei das BIP den dritthöchsten Wert aufweist. Wäre das Bruttoinlandsprodukt also nach Einwohnerzahlen disaggregiert worden, hätte dies die Datenlage verfälscht.

Sao Paulo ist mit Abstand die wirtschaftsstärkste Region Brasiliens mit einem BIP von knapp 690 Millionen R$. Damit enthält die Gemeinde rund ein Drittel des BIPs des ausgewählten Datensatzes. Auf das ganze Land bezogen produziert Sao Paulo rund 11% des gesamten Bruttoinlandprodukts. Ein weiterer Bestandteil der Tabelle zeigt die Anzahl der Kacheln pro Gemeinde. Insgesamt enthält der Datensatz 100.646 Kacheln.

ID	Stadt	Einwohner	BIP in R$	BIP-Anteil	Kachel-Anzahl	Anteil über 1000 R$
1	Sao Paulo	10.021.295	687.035.890	33,0%	1.723	71%
2	Rio de Janeiro	6.023.699	329.431.360	15,8%	1.425	70%
3	Salvador	2.711.840	61.102.373	2,9%	806	39%
4	Fortaleza	2.400.000	60.141.145	2,9%	389	83%
5	Belo Horizonte	2.373.224	88.277.463	4,2%	405	79%
6	Brasilia	2.207.718	235.497.107	11,3%	6.015	30%
7	Curitiba	1.718.421	83.788.904	4,0%	504	85%
8	Manaus	1.598.210	70.296.364	3,4%	11.871	4%
9	Recife	1.478.098	49.544.088	2,4%	287	67%
10	Porto Alegre	1.372.741	73.425.264	3,5%	603	71%
11	Goiania	1.171.195	46.659.223	2,2%	836	59%
12	Maceio	954.991	21.306.116	1,0%	615	34%
13	Sao Luis	917.237	28.323.357	1,4%	936	34%
14	Natal	763.043	21.845.481	1,0%	224	72%
15	Teresina	744.512	19.149.955	0,9%	1.596	20%
16	Campo Grande	729.151	25.437.928	1,2%	8.478	4%
17	Joao Pessoa	650.883	18.716.855	0,9%	281	58%
18	Cuiaba	521.934	22.203.168	1,1%	3.535	8%
19	Aracaju	490.175	16.498.482	0,8%	247	65%
20	Niteroi	456.456	23.003.343	1,1%	190	67%
21	Florianopolis	412.724	18.657.157	0,9%	810	46%
22	Santos	411.403	21.954.557	1,1%	356	14%
23	Macapa	338.936	9.279.790	0,4%	7.047	3%
24	Vitoria	312.656	21.727.095	1,0%	127	57%
25	Porto Velho	306.180	14.741.744	0,7%	35.816	1%
26	Rio Branco	257.642	8.123.182	0,4%	9.570	2%
27	Boa Vista	235.150	8.112.334	0,4%	5.954	2%

Zusätzlich zur Kachelanzahl pro Gemeinde wurde der Anteil der Kacheln berechnet, in denen das BIP über 1.000 R$ liegt. Durch diese Auswertung kann bereits eine Abschätzung getätigt werden, wie die Verteilung der BIP-Werte innerhalb der Gemeinde ausfällt. Der Anteil jener Kacheln die ein BIP über 1.000 R$ enthalten liegt im Schnitt bei 10%, wobei starke Schwankungen zwischen den Testgebieten bestehen. Es fällt auf, dass die einwohnerschwächeren Gemeinden eher einen niedrigeren Anteil an Kacheln mit hohem BIP aufweisen. Der ausschlaggebende Faktor ist jedoch die Größe der Gemeinde, die mit der Anzahl an Kacheln mit niedrigem BIP einhergeht. Die Größe der Gemeinde korreliert nicht mit der Anzahl der Einwohner oder dem erwirtschafteten Bruttoinlandsprodukt. Daraus resultieren große Unterschiede in der Bevölkerungsstruktur, sowie in der Wirtschaftsleistung der einzelnen Testgebiete.

In Abbildung xy ist der Datensatz der Stadtgemeinde Manaus visualisiert. Diese befindet sich im Amazonas-Gebiet im Nordwesten Brasiliens und ist die achtgrößte Stadt des Landes. Die Gemeinde beinhaltet 11871 Kacheln und zählt zu den flächenmäßig größten Gemeinden des gesamten Datensatzes. Zur Veranschaulichung wurde dieses Untersuchungsgebiet gewählt, da sich der Großteil der Kacheln im tropischen Regenwald befindet und demnach kaum wirtschaftliche Wertschöpfung enthält. Die Anzahl der Kacheln, die über 1.000 R$ liegen beträgt 475, was einem Anteil von rund 4% der Gesamtfläche der Gemeinde entspricht.



In der Abbildung rechts unten ist die Verteilung des BIP zu sehen. Dabei fällt auf, dass sich die relevanten Kacheln auf den städtischen Bereich begrenzen. Die Bebauung ist dabei in den Sentinel-2 Daten deutlich erkennbar. Auch in den NDBI-Daten lässt sich der Unterschied zwischen bebauten und nicht bebauten Flächen visuell erkennen. Die Black Marble Werte zeigen bei visueller Interpretation einen Zusammenhang zwischen Bruttoinlandsprodukt und nächtlicher Beleuchtung. Dabei ist auch erkennbar, dass der Großteil der Pixel der Black Marble Daten Leerwerte beinhaltet. Die Intensität der Beleuchtung nimmt mit der Helligkeit der Pixel zu. Durch die Ausführung des Machine Learning Modells wird sich zeigen, ob ein Zusammenhang zwischen den Satellitendaten und dem Bruttoinlandsprodukt gefunden und genutzt werden kann, um das BIP voraussagen zu können.

2.3.2.3 Verwendetes Modell

Das Machine Learning Modell, welches im Zuge dieser Arbeit eingesetzt wird, wurde vom DLR zur Verfügung gestellt. Es wurde von Thomas Stark speziell für diesen Zweck entwickelt, und in Zusammenarbeit auf die vorliegenden Daten angepasst. Das Modell wurde in der Programmiersprache Python implementiert und basiert auf der Open Source Bibliothek Pytorch, welche die grundlegenden Machine Learning Algorithmen beinhaltet.

Wie bereits in Kapitel 2.3.2 beschrieben, werden die Ansätze des Convolutional Neural Networks (CNN) und des Multilayer Perceptrons (MLP) genutzt, um die Geodaten zu verarbeiten und das Bruttoinlandsprodukt daraus abzuleiten. Die Ansätze werden sowohl separat als auch in Kombination getestet. Da das Convolutional Neural Network vor allem in der Bilderkennung großes Potential aufweist, werden die Sentinel-2 Daten für dieses Netzwerk genutzt. Die Zusatzdaten NDVI, MNDWI, NDBI und Black Marble werden in einem Multilayer Perceptron verarbeitet. Die zwei genannten Modelle können unabhängig voneinander ausgeführt werden. Um das Ergebnis zu optimieren, werden die zwei Modelle in einem sogenannten „Fusion Layer“ miteinander kombiniert. Durch diese Vorgehensweise kann getestet werden, welche Geodaten zur Ableitung des BIPs am besten geeignet sind.

Die Inputdaten des MLPs fließen in Form einer Tabelle in das Modell ein. Jede 1x1km Kachel des Datensatzes verfügt dabei über eine eindeutige ID. Somit kann jeder Sentinel-2 Kachel auch der Wert des NDVI, MNDWI, NDBI und Black Marble, sowie der Referenzwert des Bruttoinlandprodukts zugewiesen werden. Der Datensatz beinhaltet 27 Gemeinden, die jeweils eine bestimmte Anzahl an Kacheln besitzen. Für das Training und die Validierung des Modells werden jene Gemeinden ausgewählt, deren ID laut Tabelle xy eine gerade Zahl aufweisen. 75 Prozent dieser Auswahl werden als Trainingsdaten genutzt, während 25 Prozent als Validierungsdaten verwendet werden. Jene Gemeinden mit ungerader ID werden als Testdaten herangezogen. Durch diese Vorgehensweise wird das Modell sowohl hinsichtlich bevölkerungsreicher Städte wie Rio de Janeiro, als auch für bevölkerungsarme Regionen trainiert.

Das MLP und das CNN müssen so kombiniert werden, dass für jede Kachel genau ein Wert berechnet wird. Das CNN, welches das Bruttoinlandsprodukt aus den Sentinel-2 Daten ableitet, besteht aus mehreren Convolutional Layern, wobei der letzte Layer des CNNs die gleiche Anzahl an Neuronen aufweist, wie der erste Layer des MLPs. Somit können die Modelle miteinander kombiniert werden. Jedes Training durchläuft 30 Epochen, in denen die Gewichtungen zwischen den Neuronen so weit optimiert werden, dass der Output möglichst nahe am Referenzwert liegt. Das dabei generierte Modell wird anschließend auf die Testdaten angewandt, um aus diesen das Bruttoinlandsprodukt zu berechnen.

2.3.2.4 Analysen

In diesem Kapitel werden unterschiedliche Experimente mit den Deep Learning Modellen durchgeführt, um deren Performance zu verbessern und möglichst gut auf die Inputdaten abzustimmen. Eine Methode, um das Training eines Modells zu optimieren, ist die Normalisierung der Referenzwerte. Das BIP pro Rasterzelle variiert sehr stark bis zu Werten von über 6.000.000 R$. Um die Streuung zu reduzieren, werden mehrere Methoden getestet. Die Werte werden durch unterschiedliche Funktionen (Logarithmus, Wurzel) manipuliert, sodass die Verteilung dieser verändert wird. Zusätzlich werden die Werte durch den Maximalwert dividiert, um diese zu normieren. Die anschließenden Testdurchläufe werden miteinander verglichen, um die beste Methode zu bestimmen.

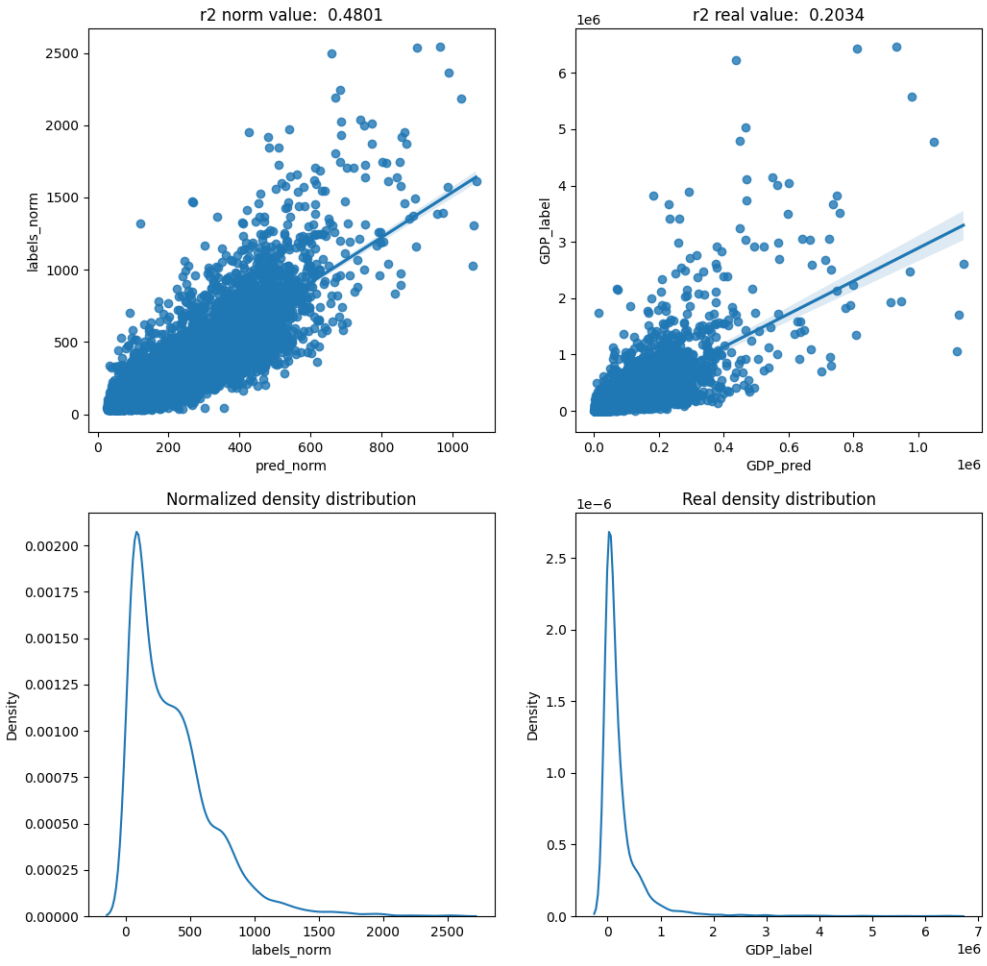
In Tabelle xy sind die statistischen Ergebnisse der Modelldurchläufe zu sehen. Bei den Ergebnissen wird unterschieden zwischen den funktionell veränderten Referenzdaten und den realen Referenzdaten. Da die Modelle mit den funktionell veränderten Daten trainiert werden, liefert auch die Anwendung auf die normierten Testdaten bessere Ergebnisse als auf die originalen Testdaten. Als statistische Maßzahlen der Analysen dienen der Determinationskoeffizient R^2 und die Wurzel des mittleren quadratischen Fehlers (RMSE). Der RMSE besitzt dabei die gleiche Einheit wie die Referenzdaten, kann also als durchschnittliche quadratische Abweichung der vorhergesagten Werte von den Referenzwerten in R$ interpretiert werden. Je niedriger der Koeffizient ist, desto näher liegt das vorhergesagte BIP an den Referenzwerten. R^2 dagegen gibt Auskunft über die Genauigkeit der Regression des Modells. Je näher sich R^2 an 1 annähert, desto genauer funktioniert das Modell.

	Functional Data		**Real Data**	
Function	**R^2**	**RMSE**	**R^2**	**RMSE**
log	0,213	0,113	-0,116	486.902
log2	0,203	0,114	-0,112	486.085
log10	0,205	0,114	-0,119	487.491
sqrt	-0,161	0,137	-0,150	494.384
simple	-0,192	0,078	-0,192	503.180

In Tabelle xy ist zu erkennen, dass die Anwendung von Logarithmusfunktionen die besten Ergebnisse auf die normierten Referenzdaten liefert. Die letzte Reihe der Tabelle zeigt das Ergebnis des Modells bei simpler Normalisierung, was bedeutet, dass die Referenzwerte durch den Maximalwert dividiert, jedoch nicht funktionell verändert wurden. Die Anwendung der Modelle auf die originalen Referenzwerte zeigt jedoch, dass keine der Methoden ein akzeptables Ergebnis liefert. Aus diesem Grund wird in weiterer Folge getestet, die Referenzwerte ohne Normierung funktionell zu manipulieren. Für dieses Experiment wird einmal die Logarithmusfunktion und einmal die Wurzelfunktion angewandt.

	Functional Data		Real Data	
Function	**R²**	**RMSE**	**R²**	**RMSE**
Log	0,680	1,134	-0,470	558.838
Sqrt	0,480	233,390	0,203	411.387

Wie in Tabelle xy zu sehen, führt das Training mit logarithmierten Referenzwerten zu einem guten Ergebnis, wenn die Testdaten ebenfalls logarithmisch verändert werden. Hier liegt der Koeffizient R² bei 0,68. Wird dasselbe Modell jedoch auf die originalen Referenzwerte angewandt, resultiert ein stark negativer Wert, was bedeutet, dass kein Zusammenhang zwischen den modellierten- und den Referenzwerten besteht. Das beste Ergebnis wird erzielt, wenn die Verteilung der Daten durch eine Wurzelfunktion verändert wird. Bei dieser Vorgehensweise liegt R² bei 0,480 beziehungsweise bei 0,203 bei einem Testdurchlauf mit originalen Referenzwerten.



Die Diagramme in Abbildung xy zeigen die Regressionsgeraden beziehungsweise die Verteilung der modellierten BIP-Werte. Die unteren zwei Diagramme visualisieren die Dichteverteilung der Daten. Dabei ist zu erkennen, dass durch die Anwendung der

Wurzelfunktion, die Verteilung der Werte gleichmäßiger ausfällt als bei den Originaldaten. Die Werte werden auf einen deutlich größeren Zahlenraum aufgeteilt, was die Performance des Modells beim Training verbessern sollte. Um den Determinationskoeffizienten R^2 zu erhöhen, werden in weiterer Folge Experimente zur Parameteroptimierung durchgeführt.

Parameteroptimierung

Um das Deep Learning Modell zu optimieren, werden die zwei grundlegenden Modelle (CNN, MLP) separat ausgeführt, um die Parameter möglichst genau auf die Eingangsdaten abzustimmen. Die Parameter, die dabei verändert werden können, sind die „learning rate" und der „weight decay". Beide Variablen sind Parameter eines sogenannten „optimizers", der die Gewichtung der Neuronen, innerhalb des Netzwerks, während des Trainings optimiert.

In der ersten Analyse zur Parameteroptimierung wird das CNN in 9 unterschiedlichen Parameterkombinationen ausgeführt. Das beste Ergebnis liefert das zweite Experiment mit einer Kombination einer learning rate von 0,01 mit einem weight decay von 0,0001. Der Determinationskoeffizient R^2 liegt hier bei 0,465, beziehungsweise bei 0,661 bei normierten Referenzwerten. Die Kombination mit jeweils einer Nachkommastelle weniger erzielt ein ähnliches Ergebnis, wobei R^2 bei Anwendung auf die originalen Referenzdaten leicht unter dem Wert von Experiment 2 liegt. Unter Berücksichtigung dieser Ergebnisse, wird für jede weitere Analyse die Parameterkombination aus dem Experiment 2 übernommen.

Experiment	learning rate	weight decay	R^2	R^2 normiert
1	0,001	0,0001	0,101	0,325
2	0,01	0,0001	0,465	0,661
3	0,1	0,0001	0,376	0,653
4	0,001	0,001	0,097	0,368
5	0,01	0,001	0,371	0,650
6	0,1	0,001	0,448	0,670
7	0,001	0,01	0,240	0,478
8	0,01	0,01	0,338	0,642
9	0,1	0,01	0,359	0,545

Um die Parameter des Multilayer Merceptrons zu optimieren wurde versucht, eine passende Anzahl an Hidden Layern zu finden. Bei der isolierten Durchführung des MLPs stellt sich jedoch heraus, dass zu wenig Daten vorliegen, um das Modell ausschließlich mit den Zusatzdaten NDVI, MNDWI, NDBI und Black Marble zu trainieren. Das MLP kann daher nur in Kombination mit dem CNN verwendet werden, um die Zusatzinformation in das Modell mit einzubinden. Die Architektur des MLPs wurde durch drei Hidden Layer festgelegt.

Um die Streuung der Modellergebnisse zu untersuchen, wurde das Modell, bestehend aus CNN und MLP, mehrmals ausgeführt. Die Trainings-, Validierungs- und Testdaten bleiben bei allen Durchläufen ident, um eine Vergleichbarkeit der Ergebnisse zu gewährleisten. In Tabelle xy sind die Ergebnisse für „*R^2*" und „*R^2 normiert*" zu sehen. Dabei fällt auf, dass die Ergebnisse

einer Standardabweichung von 0,08 unterliegen, was der mittleren Abweichung aller vorhergesagten Werte vom arithmetischen Mittel entspricht. Der Mittelwert liegt dabei bei einem R^2 von 0,413. Somit ist eine relativ große Streuung zu vermerken. Im Experiment 5 konnte jedoch ein R^2 von 0,522 erreicht werden, was dem bisher besten Ergebnis entspricht. Die Netzwerke, die bei jedem Modelldurchlauf entstehen, werden in einer Datei gespeichert. So kann am Ende der Analysen das beste Modell zur Klassifizierung verwendet werden.

Experiment	hidden layer	learning rate	Weight decay	R^2	R^2 normiert
1	3	0,01	0,0001	0,485	0,654
2	3	0,01	0,0001	0,450	0,683
3	3	0,01	0,0001	0,276	0,592
4	3	0,01	0,0001	0,376	0,647
5	3	0,01	0,0001	0,522	0,694
6	3	0,01	0,0001	0,390	0,665
7	3	0,01	0,0001	0,393	0,654
Durchschnitt				0,413	0,656
Standardabweichung				0,08	0,03

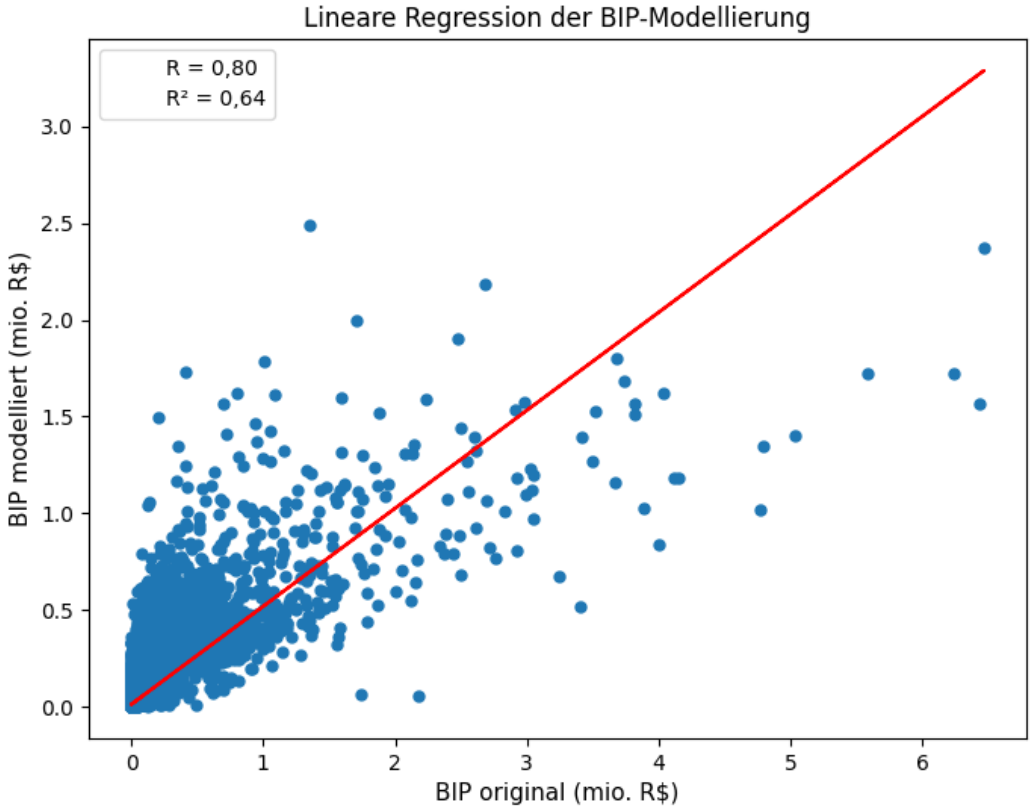
In einer letzten Analyse werden die Zusatzdaten des Multilayer Perceptrons beschränkt auf die Black Marble Daten. Das bedeutet, dass das Modell zusätzlich zu den Sentinel-2 Daten nur aus den Nachtsatellitendaten den Zusammenhang mit dem Bruttoinlandsprodukt erlernen soll. Diese Einstellung wurde ebenfalls für 7 Modelldurchläufe übernommen.

Experiment	hidden layer	learning rate	Weight decay	R^2	R^2 normiert
1	3	0,01	0,0001	0,498	0,658
2	3	0,01	0,0001	0,492	0,680
3	3	0,01	0,0001	0,394	0,638
4	3	0,01	0,0001	0,571	0,701
5	3	0,01	0,0001	0,322	0,612
6	3	0,01	0,0001	0,423	0,654
7	3	0,01	0,0001	0,442	0,671
Durchschnitt				0,449	0,659
Standardabweichung				0,07	0,03

Die Tabelle xy zeigt, dass die Vernachlässigung von NDVI, MNDWI und NDBI eine minimale Verbesserung der Performance des Deep Learning Modells bewirkt. Der mittlere R^2 beträgt bei obiger Analyse 0,449, was einer Verbesserung von knapp 10% zur vorherigen Analyse entspricht. Auch die Standardabweichung ist um 0,01 niedriger. Das beste Ergebnis liefert das Experiment 4 mit einem Determinationskoeffizienten R^2 von 0,571.

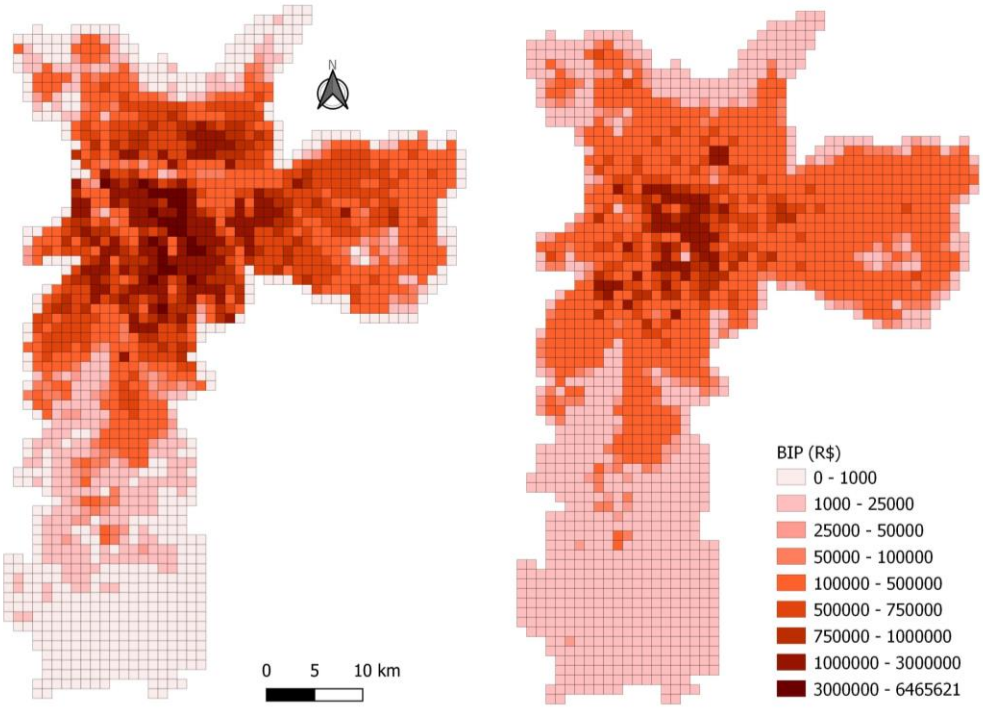
3 Ergebnisse

Das Deep Learning Modell, mit dem die Klassifikation durchgeführt wurde, weist einen Determinationskoeffizienten R^2 von 0,64 auf, was einer Pearson Korrelation R von 0,8 entspricht. Dieses Ergebnis deutet auf eine signifikante positive Korrelation zwischen den Referenzwerten und den modellierten Werten. Im Streudiagramm (Abbildung xy) ist der Zusammenhang zwischen den originalen und den modellierten BIP-Werten zu sehen. Dabei fällt auf, dass die modellierten Werte niedriger sind als die Referenzdaten. Vor allem hohe Werte über 4mio. R$ und Ausreißer werden durch die Regression gedämpft.



BIP original	BIP modelliert	Abweichung	Abweichung %	R	R^2
1.149.800.752,8	1.287.970.146,7	138.169.393,9	12%	0,80	0,64

Das Modell prädiziert in Summe aller Testgebiete ein 12% höheres Bruttoinlandsprodukt als in den Referenzdaten vorhanden ist. Trotz der Glättung der Ausreißer wird den Rasterzellen in Summe ein höherer Wert zugewiesen. Zur Veranschaulichung der Wertezuweisung des Deep Learning Modells wird auf die Städte Sao Paulo und Recife näher eingegangen. Die Modellergebnisse der übrigen 12 Testgebiete werden in Tabelle xy im Kapitel 3.1 miteinander verglichen.

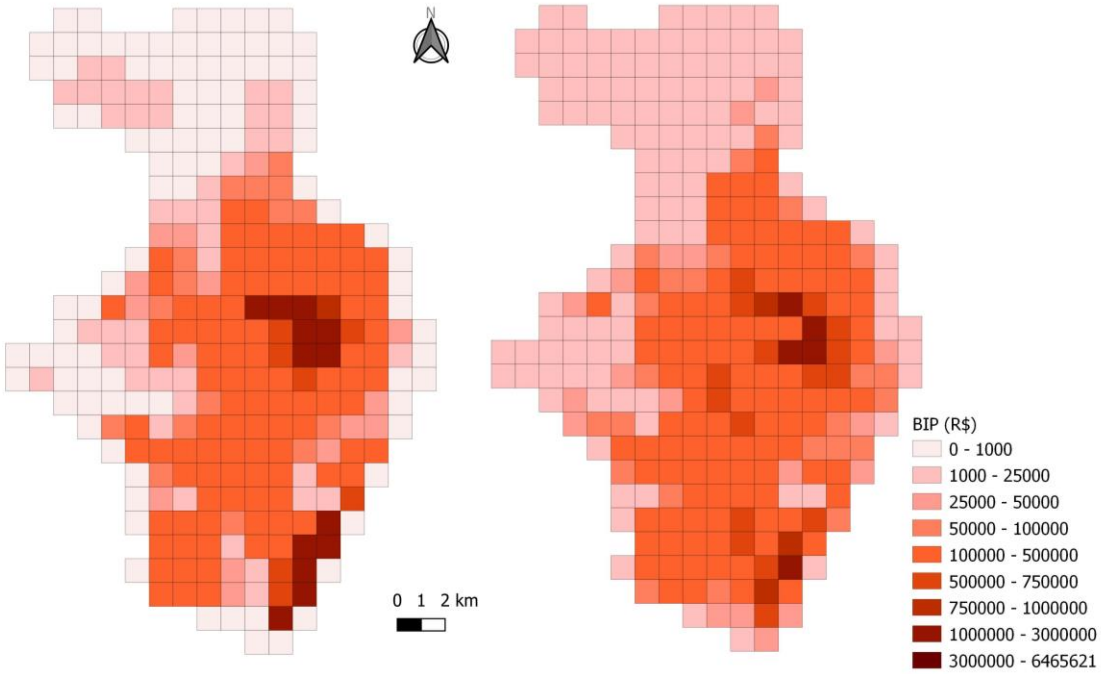


In Abbildung xy werden die Referenzwerte (links) den modellierten Werten (rechts) des Testgebiets Sao Paulo gegenübergestellt. In der nördlichen Hälfte des Testgebiets befindet sich das Stadtgebiet von Sao Paulo, in dem in den Referenzdaten mehrere Kacheln BIP-Werte von über 3mio. R$ beinhalten. Der Maximalwert von 6.465.621 R$ befindet sich ebenfalls in der Innenstadt von Sao Paulo. Das BIP in den Kacheln der umliegenden Regionen liegt größtenteils unter 1000 R$. In der Darstellung der Vorhersagewerte fallt das Bruttoinlandsprodukt im städtischen Bereich deutlich niedriger aus als in den Referenzdaten. Wie bereits im Streudiagramm (Abbildung xy) ersichtlich, liegen die Maximalwerte unter 3mio. R$. Durch die Glättung der Ausreißer geht ein großer Anteil am BIP verloren, was darin resultiert, dass die Modellierung von Sao Paulo als einzige in Summe weniger BIP prädiziert, als in den Referenzdaten vorhanden ist. Die Referenzdaten für Sao Paulo betragen 689.398.650 R$ während in der Modellierung nur 386.025.485,7 R$ zugewiesen werden. Dies entspricht einer Abweichung von -44% von der Ausgangsdaten. Der Korrelationskoeffizient R weist mit einem Wert von 0,87 dennoch auf eine signifikante positive Korrelation hin (Tabelle xy).

Generell ist eine homogenere Werteverteilung innerhalb des Testgebiets zu erkennen. Ein Großteil der Kacheln im städtischen Bereich liegt in der mittleren Klasse zwischen 100.000 R$ und 500.000 R$. Der Umriss dieser Klasse ähnelt jedoch sehr dem der Referenzdaten.

Ein weiteres Merkmal des Modellergebnisses ist die höhere Wertezuweisung an Kacheln in den Randbereichen beziehungsweise in den ruralen Regionen der Testgebiete. Diese Eigenschaft wirkt sich besonders stark auf jene Testgebiete aus, die große Flächen ländlichen Raums beinhalten (z.B. Porto Velho, Macapa).

Das Testgebiet, welches mit der geringsten Abweichung von den Referenzwerten und einem Korrelationskoeffizienten R von 0,84 die genaueste Modellierung aufweist, ist die Stadt Recife.



In Abbildung xy ist links der Referenzdatensatz und rechts das Modellergebnis von Recife zu sehen. Die Stadt liegt an der Ostküste Brasiliens im Bundesstaat Pernambuco und befindet sich mit 1.478.098 Einwohnern auf Platz 9 der bevölkerungsreichsten Städte Brasiliens. Das Modellergebnis weist in diesem Testgebiet ähnliche Eigenschaften auf, wie in Sao Paulo. Dies betrifft die Zuweisung höherer Werte an Kacheln in ruralen Gebieten, sowie eine Glättung der Ausreißer. Die Maximalwerte werden vom Modell nicht exakt nachgebildet, die Verteilung der Kachelwerte ergibt jedoch ein ähnliches Muster wie es in den Referenzdaten zu sehen ist. Somit werden durch das Modell die zwei Stadtviertel erkannt, in denen das BIP höher ist als in den anderen Teilen des Testgebiets. Die mittlere Klasse zwischen 100.000 R$ und 500.000 R$ wird durch das Modell wieder sehr gut prädiziert.

3.1 Validierung der Ergebnisse

Um das Ergebnis aller Testgebiete zu validieren, wurde eine statistische Auswertung der Städte durchgeführt. In Tabelle xy sind jene Städte zu sehen, auf die das Deep Learning Modell angewandt wurde.

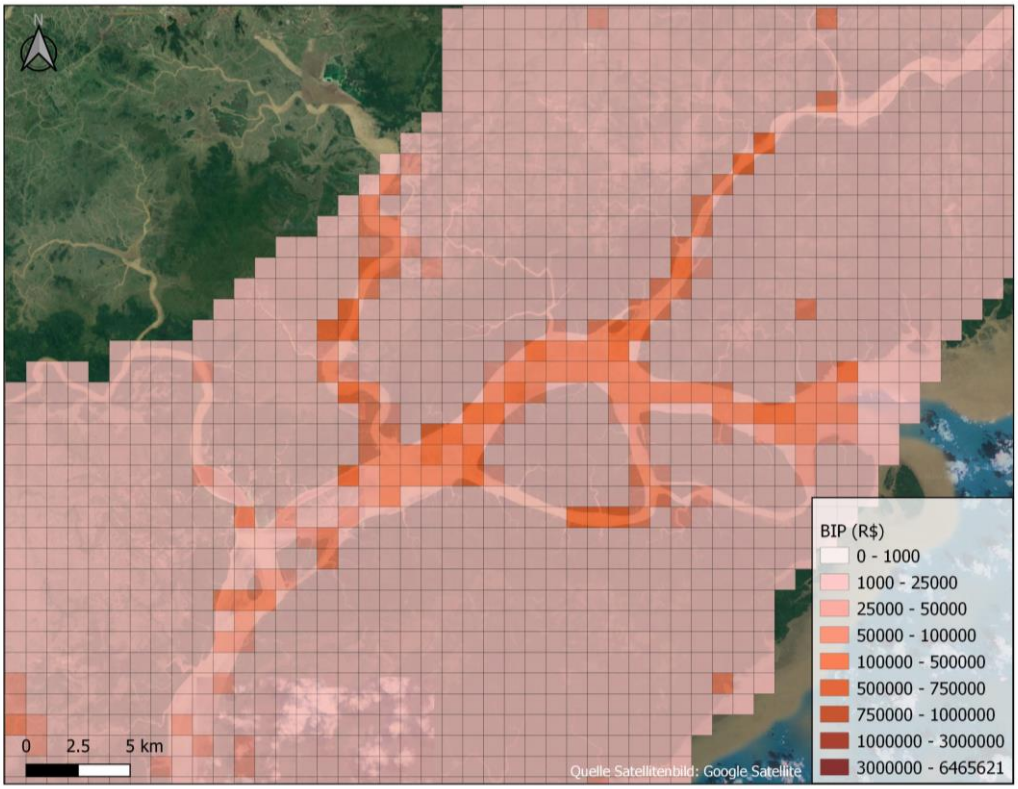
Stadt	Groß-region	BIP original	BIP modelliert	Abweichung	Abw. %	R	R^2
Sao Paulo	SE	689.398.650,0	386.025.485,7	-303373164,3	-44%	0,87	0,75
Belo Horizonte	SE	89.207.893,5	120.757.154,1	31549260,6	35%	0,77	0,59
Curitiba	S	83.733.754,9	132.197.239,2	48463484,3	58%	0,83	0,70
Salvador	NE	60.064.989,0	78.545.724,2	18480735,2	31%	0,87	0,75
Recife	NE	48.763.761,4	52.073.897,2	3310135,8	7%	0,84	0,70
Goiania	mW	47.220.708,0	84.483.323,8	37262615,8	79%	0,84	0,70
Sao Luis	NE	28.184.642,4	28.400.492,5	215850,1	1%	0,72	0,52
Teresina	NE	19.064.107,5	28.418.915,0	9354807,5	49%	0,83	0,69
Florianopolis	S	18.823.343,1	45.761.214,2	26937871,1	143%	0,81	0,66
Joao Pessoa	NE	17.326.195,4	31.364.988,4	14038793,0	81%	0,84	0,71
Aracaju	NE	15.929.214,4	18.334.294,4	2405080,0	15%	0,75	0,56
Porto Velho	N	14.769.272,0	180.822.956,3	166053684,3	1124%	0,63	0,39
Macapa	N	9.260.114,3	62.641.152,7	53381038,4	576%	0,78	0,62
Boa Vista	N	8.054.106,9	38.143.309,1	30089202,2	374%	0,84	0,70

Im Vergleich der Testgebiete fällt auf, dass zum Teil sehr große Abweichungen zwischen den Referenz- und den Modellwerten bestehen. Besonders hoch ist die Abweichung von 1124% von den Referenzwerten in der Stadt Porto Velho. Auch die Testgebiete Macapa und Boa Vista weichen stark von den Referenzwerten ab. Der Korrelationskoeffizient R und der Determinationskoeffizient R^2 zeigen jedoch, dass auch in diesen Testgebieten ein Zusammenhang zwischen den Referenzdaten und den modellierten Ergebnissen besteht. Der Grund für die großen Abweichungen liegt darin, dass die genannten Testgebiete sehr groß sind und hauptsächlich ländliche Gebiete beinhalten, die modellbedingt zu hoch eingestuft werden. Je mehr Fläche das Testgebiet abdeckt, desto größer wird dadurch auch die Abweichung von den Referenzdaten.

Der einzige Testdatensatz, der mit einer negativen Abweichung modelliert wurde, ist Sao Paulo. Hier weicht die Summe der modellierten BIP-Werte um 303.373.164,3 R$ von den Referenzwerten ab. Wie bereits in Kapitel 3 beschrieben, liegt der Grund dafür in der Glättung der Ausreißer durch die Modellierung. Die geringste Abweichung ist in der Stadt Sao Luis zu vermerken, in der die modellierten Werte um nur 1% von den Referenzwerten abweichen. R^2 liegt in diesem Testgebiet jedoch nur knapp über 0,5. Die Streudiagramme aller 14 Städte befinden sich im Anhang dieser Arbeit (Kapitel 5.1).

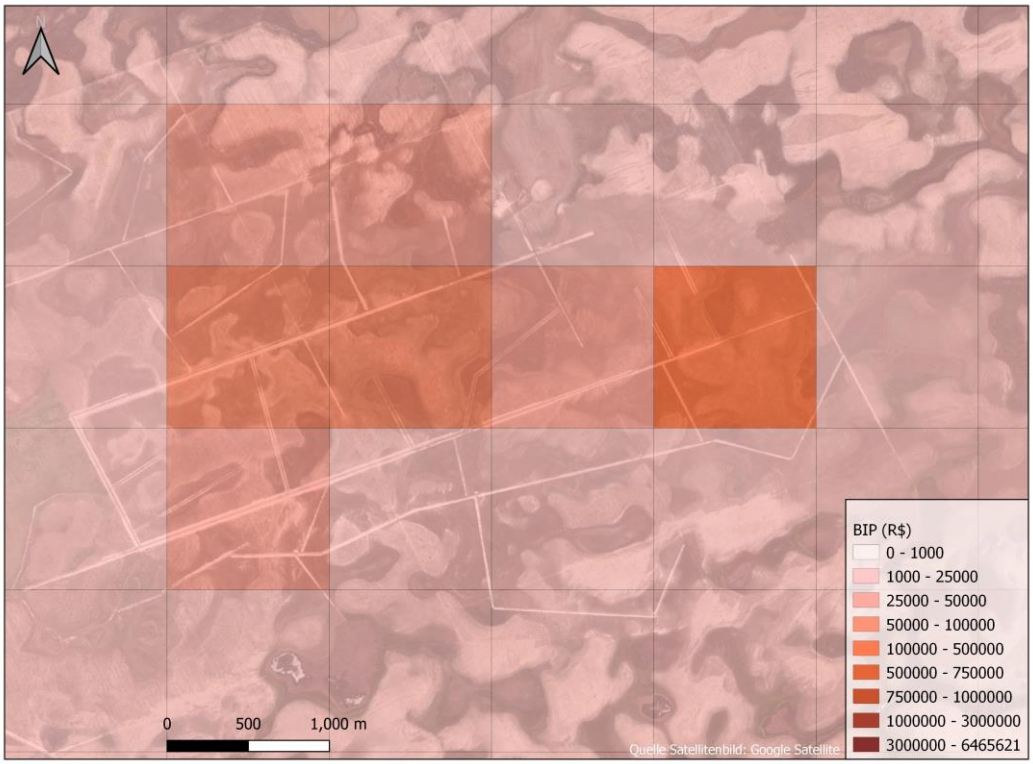
3.2 Interpretation der Ergebnisse

Wie bereits in Kapitel 3.1 erwähnt, weichen die Testgebiete Porto Velho, Macapa und Boa Vista sehr stark von den Referenzwerten ab. Am größten ist die Abweichung in Porto Velho in einer Höhe von 166.053.684,3 R$. Der Grund dafür liegt in der Größe der Stadtgemeinde, die eine Anzahl von 35816 Kacheln beinhaltet. Im Vergleich dazu liegt die Anzahl der Kacheln der Stadtgemeinde Recife bei 287. Ein weiterer Grund für die Abweichung liegt in der fälschlichen Klassifizierung des Modells von Gewässern als Strukturen wirtschaftlicher Wertschöpfung.



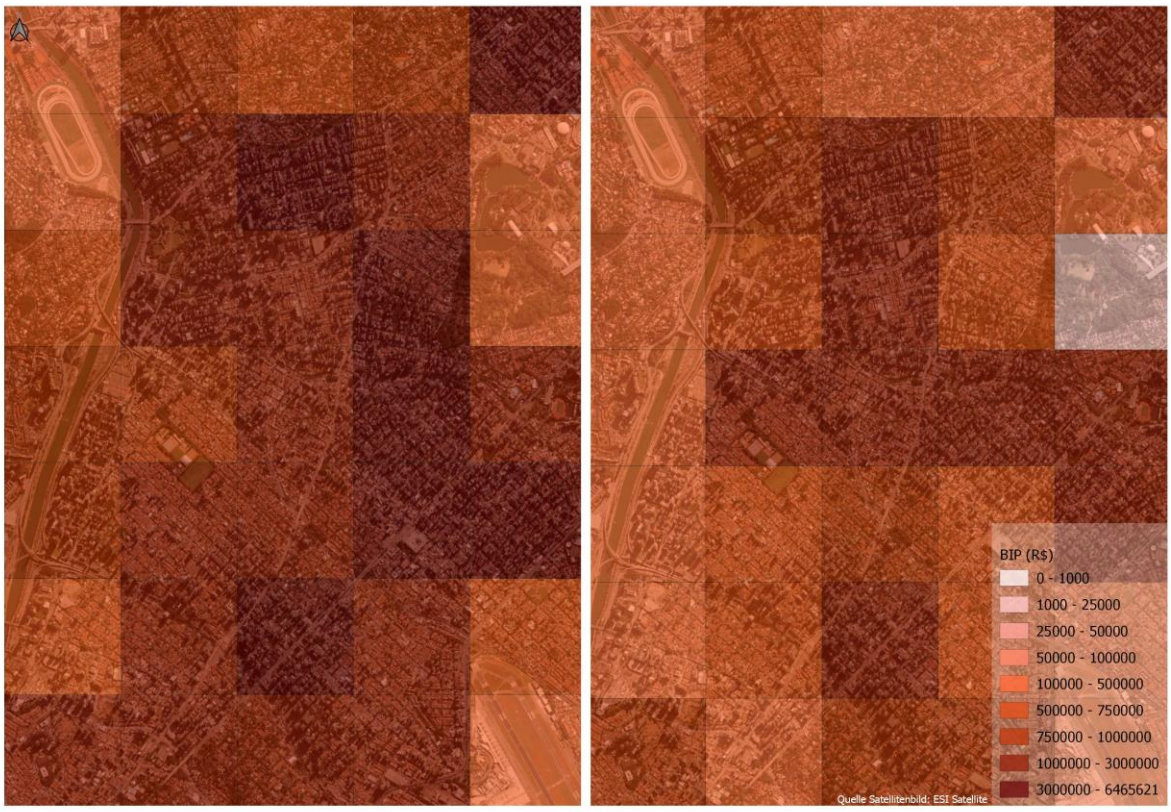
In Abbildung xy ist ein Ausschnitt des Testgebiets von Macapa zu sehen. Die Stadt befindet sich im Norden Brasiliens und der markante Flusslauf stellt das Mündungsgebiet des Amazonas in den Atlantischen Ozean dar. Das Gemeindepolygon dieser Stadt beinhaltet große rurale Gebiete, in denen der Beitrag zum Bruttoinlandsprodukt sehr klein ist. Durch die scharfen Kanten der Flussläufe werden die betroffenen Regionen jedoch als städtische Strukturen interpretiert. Eine weitere Ursache dafür könnten die NDVI-Werte sein, die als Inputdaten in das Vorhersagemodell einfließen. Da der NDVI als Indikator für den Wassergehalt in der Vegetation beziehungsweise zur Klassifizierung von Gewässern verwendet werden kann, könnte sich diese Eigenschaft negativ auf das Machine Learning Modell in ruralen Regionen auswirken. Durch die falsche Erkennung des Deep Learning Modells werden diesen Kacheln sehr hohe Werte zugewiesen, was das Ergebnis verfälscht.

Eine weitere Fehlerquelle in der Modellierung liegt bei geradlinigen Strukturen in den Sentinel-2 Daten. Im Testgebiet Boa Vista werden vermeintliche Feldwege erkannt und den betroffenen Kacheln zu hohe BIP-Werte zugeordnet. In Abbildung xy ist ein Ausschnitt des Testgebiets von Boa Vista zu sehen, in dem diese Modelleigenschaft zu einer falschen Wertezuweisung führt. Da in den Zusatzdaten NDVI, NDBI, MNDWI und Black Marble keine hinweise auf eine erhöhte wirtschaftliche Aktivität in den betroffenen Kacheln vorliegt, liegt in den Sentinel-2 Daten die Ursache dieser Klassifizierung.



Die genannten Beispiele zeigen, dass es vor allem im landwirtschaftlichen Raum, sowie in Bereichen größerer Gewässer zu fehlerhaften Modellierungen kommen kann. Daraus resultiert die größere Abweichung der modellierten Werte gegenüber den Referenzwerten in Testgebieten, die sehr große Landesteile abdecken und somit auch dünn besiedelte Gebiete und Flusslandschaften beinhalten. In Testgebieten, die hauptsächlich städtische Strukturen aufweisen, wie Sao Paulo oder Recife, können mit Hilfe des Machine Learning Modells durchaus brauchbare Ergebnisse erzielt werden.

In Abbildung xy ist der Vergleich zwischen Referenzdaten und Ergebnisdaten eines Ausschnitts von Sao Paulo visualisiert. Links sind die Referenzdaten und rechts ist das modellierte Ergebnis zu sehen. Der Ausschnitt zeigt die einkommensstärksten Kacheln der Stadt, die somit auch die höchsten BIP-Werte enthalten. Die mittlere der drei vertikal angeordneten Kacheln der höchsten Klasse in der linken Abbildung ist jene Kachel, die mit 6.465.621 R$ den Maximalwert des gesamten Datensatzes enthält.



In den Referenzdaten ist zu sehen, dass nach visueller Interpretation das BIP logisch disaggregiert wurde. Kacheln, die weniger dicht bebaut sind, wie der Flughafen im Südosten des Ausschnitts oder die Parkflächen im Nordosten weisen ein geringeres Bruttoinlandsprodukt auf als Kacheln mit dichter Bebauung. Auch die Bereiche um den Fluss im Westen des Ausschnitts weisen niedrigere Werte auf. Im Modellergebnis rechts ist zu sehen, dass diese Muster auch vom Machine Learning Modell erkannt werden. Ein Teil der Parkfläche wurde sogar geringer eingestuft als es die Referenzdaten vorgeben. Was jedoch auch auffällt ist, dass die Werte nicht die Größe der Referenzdaten erreichen und nicht über die Klassengrenze von 3mio. R$ hinausgehen. Die Anordnung der modellierten Werte ergibt dennoch ein ähnliches Muster wie die Referenzdaten.

Eine Ausnahme, in der die Dichte der Bebauung nicht mit der Höhe des Bruttoinlandprodukts korreliert, wären die Armenviertel (Favelas) in den Großstädten. Auch die nächtliche Beleuchtung gibt in diesen Gebieten keinen Aufschluss über die wirtschaftliche Leistung. Da das BIP jedoch nach der Höhe des Einkommens disaggregiert wurde, sind in den Referenzdaten auch in den dicht besiedelten Armenvierteln keine hohen Werte zugeordnet. Das Machine Learning Modell müsste daher lernen, die ungeordnete Siedlungsstruktur der Favelas von geordneten Bebauungsstrukturen der einkommensstärkeren Bevölkerungsschicht zu unterscheiden.

4 Resümee

Im Zuge dieser Arbeit wurde ein Datensatz erstellt, mit dem ein Machine Learning Modell trainiert werden konnte, um das Bruttoinlandsprodukt in Brasilien vorherzusagen. Der Datensatz beinhaltet sowohl Referenzdaten als auch die Inputdaten für das Modell. Die Referenzdaten wurden in der Datenvorverarbeitung generiert, indem das Bruttoinlandsprodukt von Gemeindeebene in eine räumliche Auflösung von 1x1km disaggregiert wurde. Um die geometrische Auflösung der Gemeindedaten zu verbessern, wurden im Vorfeld der Disaggregation die BIP-Werte nach dem Einkommen innerhalb der Gemeinden aufgeteilt. Da für die Bundesstaaten Para und Tocantins im Norden des Landes keine statistische Information zum Einkommen innerhalb der Gemeinden vorliegt, konnte für diese Gebiete kein Referenzdatensatz erstellt werden. Für alle restlichen Bundesstaaten wurde ein flächendeckender Referenzdatensatz generiert. Die Inputdaten für das Machine Learning Modell umfassen Sentinel-2 Satellitenbilder, Black Marble Satellitenbilder, sowie die Indizes: NDVI, NDBI und MNDWI. Diese Daten wurden ebenfalls in der Datenvorverarbeitung akquiriert.

Durch die Anwendung des Machine Learning Modells konnten Ergebnisse für 14 Städte in Brasilien berechnet werden. Das Modell konnte mit einem Determinationskoeffizienten R^2 von 0,64 den Kacheln die Werte für das Bruttoinlandsprodukt zuweisen. In Summe liegt eine Abweichung von 12% von den modellierten Werten zu den Referenzwerten vor, die jedoch je nach Testgebiet stark variiert. Die besten Ergebnisse wurden in städtischen Gebieten erzielt, wobei die Dämpfung der Ausreißer, vor allem in Sao Paulo, das Ergebnis beeinflusst. Aufgrund der Einkommensverteilung weisen gewisse Kacheln sehr hohe Referenzwerte auf. Da die Modellierung die Werteverteilung glättet, werden diese Ausreißer nicht berücksichtigt, was in einer negativen Bilanz für die Modellierung von Sao Paulo resultiert. In kleineren Städten, in denen das Einkommen gleichmäßiger verteilt ist, liefert das Modell genauere Ergebnisse.

Verbesserungsbedarf liegt in den ruralen Regionen, denen meist zu hohe Werte zugewiesen werden. Die hohe Wertezuweisung von Kacheln, die sich über Flussläufen oder geradlinigen anthropogenen Strukturen in den Sentinel-2 Daten befinden, verfälscht ebenfalls das Ergebnis. Die beste Modellierung wird erzielt, wenn Testgebiete hauptsächlich städtische Strukturen aufweisen und keine Ausreißer in den Referenzdaten vorliegen, wie es beispielsweise in der Stadt Recife der Fall ist. Hier ist mit einem R^2 von 0,70 und einer Abweichung von 7% eine brauchbare Modellierung möglich.

4.1 Beantwortung der Forschungsfragen

Zu Beginn der Arbeit wurden folgende Forschungsfragen definiert:

- Ist es möglich, mit Hilfe von Machine Learning, die Wirtschaftskraft einer Region aus Geodaten abzuleiten?
- Kann das entwickelte Machine Learning Modell auf unterschiedliche Testgebiete in Brasilien angewandt werden?
- Gibt es einen Zusammenhang zwischen der nächtlichen Beleuchtung und der Wirtschaftskraft einer Region?

Die erste Forschungsfrage befasst sich generell mit dem Thema, ob wirtschaftliche Wertschöpfung aus Geodaten abgeleitet werden kann. Wie in Kapitel 1.3 beschrieben, gibt es unterschiedliche Methoden, um an diese Fragestellung heranzugehen. Einerseits muss eine Methodik passend zu den vorliegenden Daten gefunden werden und andererseits muss ein Wert definiert werden, der durch die Modellierung angenähert werden soll. Der sozioökonomische Indikator, der im Zuge dieser Arbeit modeliiert werden soll, ist das Bruttoinlandsprodukt, da es „[…] die am weitesten anerkannte Kennzahl für die Wirtschaftskraft und Leistung eines Landes“ ist (Conway 2011). Als Methodik wurde ein Deep Learning Modell gewählt, welches unter die Kategorie Machine Learning fällt und aus der Verbindung eines Convolutional Neural Networks (CNN) und eines Multilayer Perceptrons (MLP) besteht. Eine weitere Voraussetzung für die positive Beantwortung der ersten Forschungsfrage ist ein akkurater und flächendeckender Datensatz aus Referenzdaten und Inputdaten für das Modell. Dieser Datensatz konnte in der Datenvorverarbeitung erfolgreich erstellt werden. Durch Anwendung des Deep Learning Modells konnten Werte modelliert werden, die sich mit einem Determinationskoeffizienten von 0,64 an die Referenzwerte annähern. Es ist daher gelungen, mit Hilfe eines Machine Learning Modells die Wirtschaftskraft einer Region aus Geodaten abzuleiten.

Die zweite Forschungsfrage betrifft die Einsetzbarkeit des Modells auf unterschiedliche Testgebiete innerhalb Brasiliens. Im Zuge der Datenvorverarbeitung wurde ein Datensatz erstellt, der 27 Städte enthält. Jede zweite Stadt wurde als Trainings- und Validierungsdatensatz verwendet und die übrigen 14 Städte wurden als Testdatensatz definiert. Das Modell wurde daher auf 14 Gebiete unterschiedlicher Größe und geographischer Lage angewandt. Die Ergebnisse der Testgebiete variieren stark, jedoch ist weniger die geographische Lage der Grund dafür, sondern die Größe der Testgebiete beziehungsweise die Anzahl der Kacheln, die sich in ruralen Gebieten befinden. In den Testgebiete Porto Velho, Macapa und Boa Vista sind zusätzlich größere Flussläufe der Grund für eine höhere

Abweichung der modellierten Werte von den Referenzwerten. Die besten Ergebnisse wurden in kleineren Testgebieten erzielt, die hauptsächlich städtische Strukturen aufweisen.

Die dritte Forschungsfrage hinterfragt den Einfluss der nächtlichen Beleuchtung auf das Bruttoinlandsprodukt. Bei einer rein visuellen Interpretation der Daten gibt es einen Zusammenhang zwischen dem Bruttoinlandsprodukt und den Werten der Black Marble Satellitenbilder. Dadurch, dass das Bruttoinlandsprodukt nach der Höhe des Einkommens disaggregiert wurde, sind die Werte in bevölkerten Gebieten deutlich höher als in den dünn besiedelten Regionen. Auch die nächtliche Beleuchtung hängt stark von der Bevölkerungsdichte ab. Die Einbindung der Black Marble Daten resultiert auch in einem besseren Ergebnis des Vorhersagemodells, woraus sich schließen lässt, dass ein Zusammenhang zwischen der nächtlichen Beleuchtung und dem Bruttoinlandsprodukt besteht. Die Daten, die am meisten Information über das BIP enthalten sind jedoch die Sentinel-2 Bilder. Durch die Erkennung von anthropogenen Strukturen wie Gebäude oder Straßen kann das Modell besiedelte Gebiete erkennen und die Werte je nach Ausprägung der Strukturen vergeben.

4.2 Ausblick

Die Modellierung des Bruttoinlandprodukts ist ein sehr komplexes Thema, da der Datenbestand zum Trainieren eines Modells immer einer gewissen Ungenauigkeit unterliegt. Die Voraussetzung für eine akkurate Modellierung ist jedoch ein möglichst realitätsgetreuer Referenzdatensatz. Im Zuge dieser Arbeit wurde das Bruttoinlandsprodukt nach der Höhe des Einkommens disaggregiert. Diese Methodik ist akzeptabel, da das BIP stark mit dem Einkommen korreliert und sie im Umfang dieser Masterarbeit gut realisierbar ist. Die reale Verteilung des Bruttoinlandprodukts ist jedoch komplexer, da die wirtschaftliche Wertschöpfung nicht immer in den bevölkerten Regionen generiert wird. Der Abbau mineralischer Rohstoffe beispielsweise trägt einen großen Teil zum BIP in Brasilien bei. Das Einkommen in den Abbaugebieten korreliert jedoch nicht mit dem wirtschaftlichen Gewinn, der in diesen Regionen erzielt wird. Eine ähnliche Situation besteht in der Landwirtschaft, die speziell in Brasilien einen hohen wirtschaftlichen Stellenwert besitzt. Eine Möglichkeit einer realitätsnahen Disaggregation wäre die Aufteilung des BIP in den primären-, den sekundären- und den tertiären Wirtschaftssektor und die Verteilung der Werte nach einer Landbedeckungsklassifikation. Der Primärsektor (Urproduktion) und der Sekundärsektor (Industrie) könnten durch die Landbedeckungsdaten flächengewichtet disaggregiert werden. Der Anteil des BIP am Tertiärsektor (Dienstleistung) könnte durch Daten der OpenStreetMap verteilt werden. Die Disaggregation könnte nach der Dichte bestimmter Points of Interest erfolgen, die dem Dienstleistungssektor zugehören. Die Herausforderung dabei ist jedoch ein flächendeckender Datenbestand, da die OpenStreetMap hauptsächlich in urbanen Regionen

vollständige Informationen enthält. Eine Wirtschaftssektor-basierte Disaggregation wäre jedenfalls eine Möglichkeit, die Modellergebnisse zu verbessern. Durch diese Methode würde auch verhindert werden, dass Regionen, in denen der Beitrag zum BIP ausgeschlossen werden kann in das Modell mit einfließen. Ein weiterer Nachteil der einkommensbasierten Disaggregation ist nämlich, dass die Polygone der Zensusdaten, die das Einkommen enthalten, auch Wasserflächen beinhalten. So werden fehlerhafte Referenzdaten erstellt, die bewirken, dass das Machine Learning Modell auch den Wasserflächen einen Teil des Bruttoinlandprodukts zuweist.

Eine weitere Analyse, durch die das Modell verbessert werden könnte, wäre die Optimierung der Kachelanzahl der Testgebiete. Einerseits müssen genügend Daten vorhanden sein, um das Machine Learning Modell zu trainieren, andererseits darf die Rechenleistung des Computers nicht überschritten werden. Wenn die Ungenauigkeiten des Modells in den ruralen Gebieten ausgebessert wären, könnten auch größere Testgebiete ein gutes Modellergebnis liefern. Ein weiterer Ansatz, um die Modellierung des Bruttoinlandprodukts zu optimieren, wäre die Erhöhung der räumlichen Auflösung der Referenz- sowie Inputdaten. Da die Black Marble Daten in einer Auflösung von 500m erhältlich sind, wäre es möglich, alle Daten für die Modellierung auf diese Auflösung zu skalieren. Vor allem im städtischen Bereich würde die höhere räumliche Auflösung Vorteile bringen, da so die Bebauungsstrukturen auf kleinräumlicher Ebene besser unterschieden werden können. Zu beachten wäre bei diesem Ansatz jedoch, dass sich im Gegensatz zu einer Auflösung von 1km die Anzahl der Kacheln vervierfachen würde und dadurch der Rechenaufwand der Modellierung stark ansteigt.

Der weitere Forschungsbedarf in diesem Bereich liegt also bei einer Disaggregation nach Wirtschaftssektoren, sowie einer Optimierung des Machine Learning Modells in ruralen Regionen. Weitere Experimente könnten die räumliche Auflösung sowie die Kachelanzahl der Testgebiete optimieren.

Literaturverzeichnis

Alzubaidi, Laith; Zhang, Jinglan; Humaidi, Amjad J.; Al-Dujaili, Ayad; Duan, Ye; Al-Shamma, Omran et al. (2021): Review of deep learning. concepts, CNN architectures, challenges, applications, future directions. In: *Journal of Big Data*. Online verfügbar unter https://doi.org/10.1186/s40537-021-00444-8, zuletzt geprüft am 04.05.2023.

Charris, Carlos; Velilla, Raul; Chaves, Leonardo: Mapping the Human Development Index using Nighttime Lights inside Brazil. In: *Departamento de Economia Rural da Universidad Federal de Vicosa*.

Chen, Xi; Nordhaus, William D. (2011): Using luminosity data as a proxy for economic statistics. In: *Pnas* 108, S. 8589–8594.

Colucci, Alessandro; Sanchez, Julia; Marbler, Herwig: Bergbaustudie Brasilien 2020. Strategische Rohstoffe, Projekte und Geschäftsmöglichkeiten für deutsche Unternehmen.

Conway, Edmund (2011): 50 Schlüsselideen Wirtschaftswissenschaft. Heidelberg: Spektrum Akademischer Verlag Heidelberg. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-8274-2635-2_18.

copernicus.de (2023): Sentinel-2. Online verfügbar unter https://www.d-copernicus.de/daten/satelliten/satelliten-details/news/sentinel-2/, zuletzt aktualisiert am 08.03.2023, zuletzt geprüft am 08.03.2023.

Diacon, Paula-Elena; Maha, Liviu-George (2015): The relationship between Income, Consumption and GDP. A time series, cross-country analysis. In: *Procedia Economics and Finance* 23, S. 1535–1543.

dlr.de (2023): Normalized Difference Vegetation Index (NDVI). Unter Mitarbeit von Angela Kaiser. Online verfügbar unter https://www.dlr.de/eoc/desktopdefault.aspx/tabid-9142/19518_read-45426, zuletzt aktualisiert am 08.03.2023, zuletzt geprüft am 08.03.2023.

Elvidge, C. D.; Baugh, K. E.; Anderson, S. J.; Sutton, P. C.; Ghosh, T. (2012): The Night Light Development Index (NLDI). A spatially explicit measure of human development from satellite data. In: *Social Geography* 7, S. 23–35.

epsg.io (2023): SIRGAS 2000 / Brazil Polyconic - EPSG:5880. Online verfügbar unter https://epsg.io/5880, zuletzt aktualisiert am 08.03.2023, zuletzt geprüft am 08.03.2023.

ESA (2015): Sentinel-2 User Handbook. Online verfügbar unter https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook, zuletzt geprüft am 21.03.2023.

esa.int (2023): Sentinel-2 Mission Guide. Online verfügbar unter https://sentinel.esa.int/web/sentinel/missions/sentinel-2, zuletzt aktualisiert am 08.03.2023, zuletzt geprüft am 08.03.2023.

Feldemeyer, Daniel; Meisch, Claude; Sauter, Holger; Birkmann, Joern (2020): Using OpenStreetMap Data and Machine Learning to Generate Socio-Economic Indicators. In: *ISPRS International Journal of Geo-Information* 9.

Gärtner, Philipp (2020): How cloudy is my Sentinel-2 image collection? - The 'QA60' band gives insights. Online verfügbar unter https://philippgaertner.github.io/2020/08/percent-cloud-cover/, zuletzt aktualisiert am 14.08.2020, zuletzt geprüft am 22.03.2023.

Gebhart, H.; Glaser, R.; Radtke, U.; Reuber, P. (2016): Geographie. Physische Geographie und Humangeographie. Unter Mitarbeit von Reinhard Zeese. 2. Aufl. Berlin Heidelberg: Springer.

Geographie Infothek (2004): Infoblatt Savanne. Savannen, Savannentypen im Überblick. Hg. v. Klett. Leipzig. Online verfügbar unter www.klett.de.

Goodchild, M.F; Lam, N. (1980): Areal interpolation: a variant of the traditional spatial problem. In: *Geo-Processing* 1, S. 297–312.

Hudakova, Jarmila (2017): Relationship between Gross Domestic Product and Human Development Index. In: *4th International Multidisciplinary Scientific Conferences on Social Sciences & Arts*.

ibge.gov.br (2023): Instituto Brasileiro de Geografia e Estatistica. Online verfügbar unter https://www.ibge.gov.br/, zuletzt aktualisiert am 15.03.2023, zuletzt geprüft am 15.03.2023.

Jean, Neal; Burke, Mashall; Xie, Michael; Davis, Matthew; Lobell, David B.; Ermon, Stefano (2016): Combining satellite imagery and machine learning to predict poverty. In: *Science* 353, S. 14–32.

nasa.gov (2023): NASA's Black Marble. Online verfügbar unter https://blackmarble.gsfc.nasa.gov/, zuletzt aktualisiert am 08.03.2023, zuletzt geprüft am 08.03.2023.

Roman, Miguel O.; Wang, Zhuosen; Sun, Qingsong; Kalb, Virginia; Müller, Steven D.; Molthan, Andrew (2018): NASA’s Black Marble nighttime lights product suite. In: *Remote Sensing of Environment* 210, S. 113–143.

Sarker, Iqbal H. (2021): Deep Learning. A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. In: *Springer Nature Computer Science*. Online verfügbar unter https://doi.org/10.1007/s42979-021-00815-1, zuletzt geprüft am 04.05.2023.

sentinels.copernicus.eu (2023a): Level-2A Product. Online verfügbar unter https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-%20msi/product-types/level-2a, zuletzt aktualisiert am 22.03.2023, zuletzt geprüft am 22.03.2023.

sentinels.copernicus.eu (2023b): Level-1C Cloud Masks. Online verfügbar unter https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/level-1c/cloud-masks, zuletzt aktualisiert am 23.03.2023, zuletzt geprüft am 23.03.2023.

Tucker, C. J.; Sellers, P. J. (1986): Satellite remote sensing of primary production. In: *International Journal of Remote Sensing* 7 (11), S. 1395–1416.

weltbank.org (2023a): GDP (current US$) - Brazil | Data. Online verfügbar unter https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=BR&most_recent_value_desc=true, zuletzt aktualisiert am 08.03.2023, zuletzt geprüft am 08.03.2023.

weltbank.org (2023b): New World Bank country classifications by income level: 2022-2023. Online verfügbar unter https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023, zuletzt aktualisiert am 15.03.2023, zuletzt geprüft am 15.03.2023.

wko.at: Länderprofil Brasilien. Online verfügbar unter https://wko.at/statistik/laenderprofile/lp-brasilien.pdf, zuletzt geprüft am 08.03.2023.

wko.at: Länderprofil USA. Online verfügbar unter https://wko.at/statistik/laenderprofile/lp-usa.pdf, zuletzt geprüft am 08.03.2023.

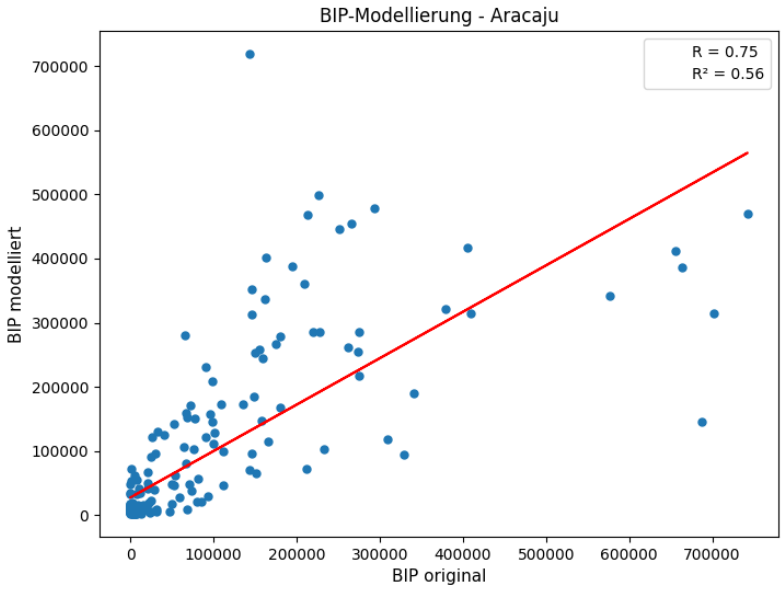
Xu, Hanqiu (2006): Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. In: *Sensors* 27, S. 3025–3033.

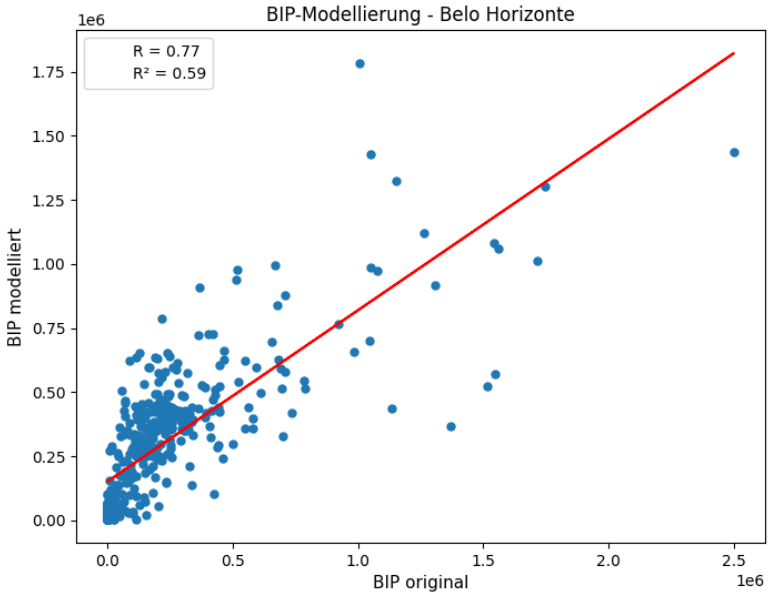
Xu, Rudong; Liu, Jin; Xu, Jianhui (2018): Extraction of High-Precision Urban Imervious Surfaces from Sentinel-2 Multispectral Imagery via Modified Linear Spectral Mixture Analysis. In: *Sensors* 18 (2873), S. 1–15.

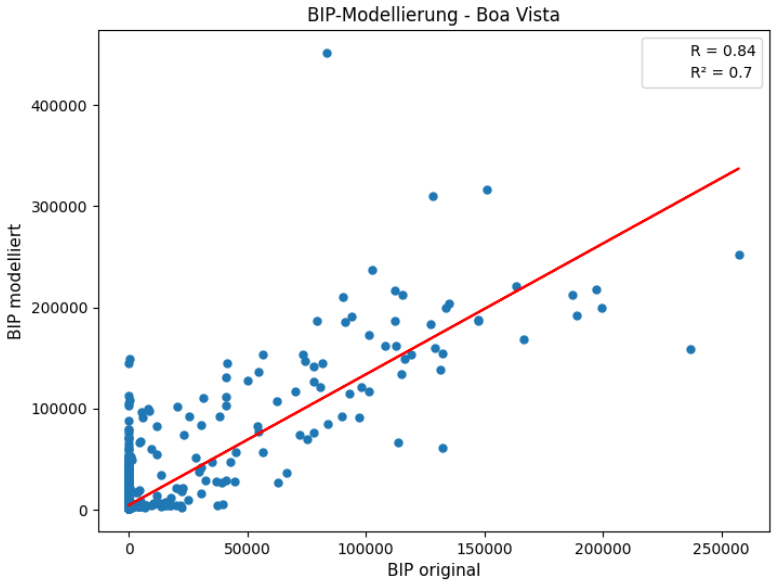
5 Anhang

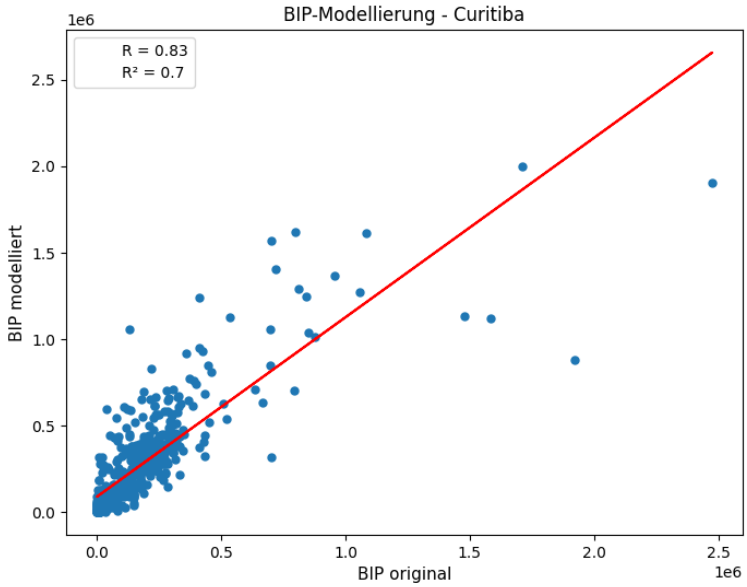
5.1 Modellergebnisse - Streudiagramme

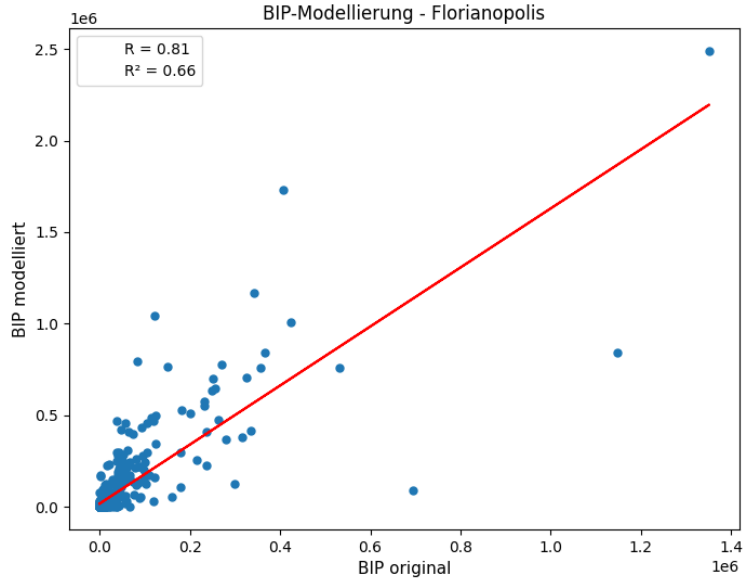
In diesem Kapitel sind die Streudiagramme aller modellierten Testgebiete zu sehen. Die X-Achse der Diagramme beschreibt die Referenzwerte des BIP während auf der Y-Achse die modellierten Werte aufgetragen werden. Die Darstellungen zeigen den Zusammenhang der Variablen, sowie die lineare Regression.

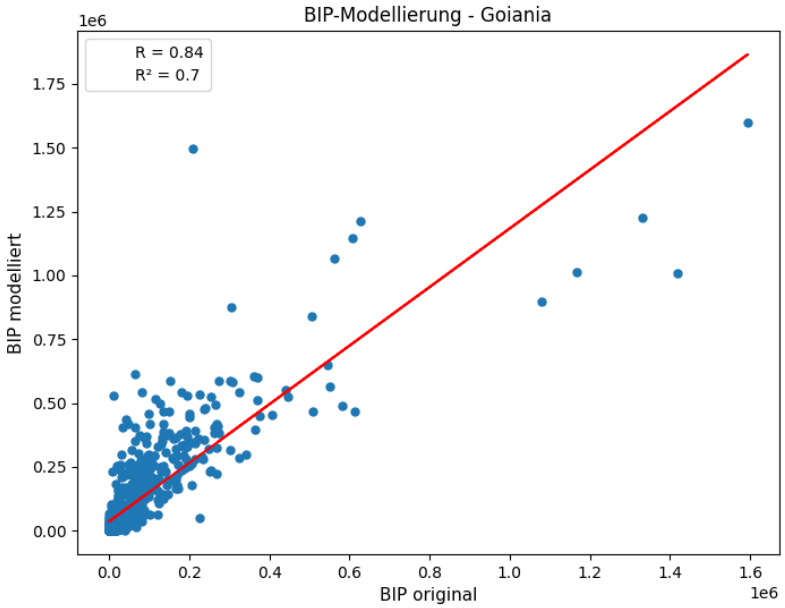


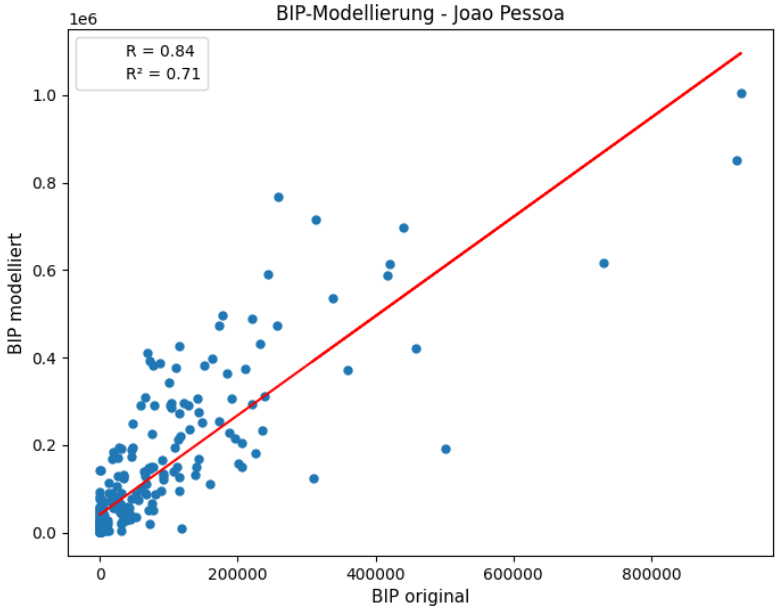


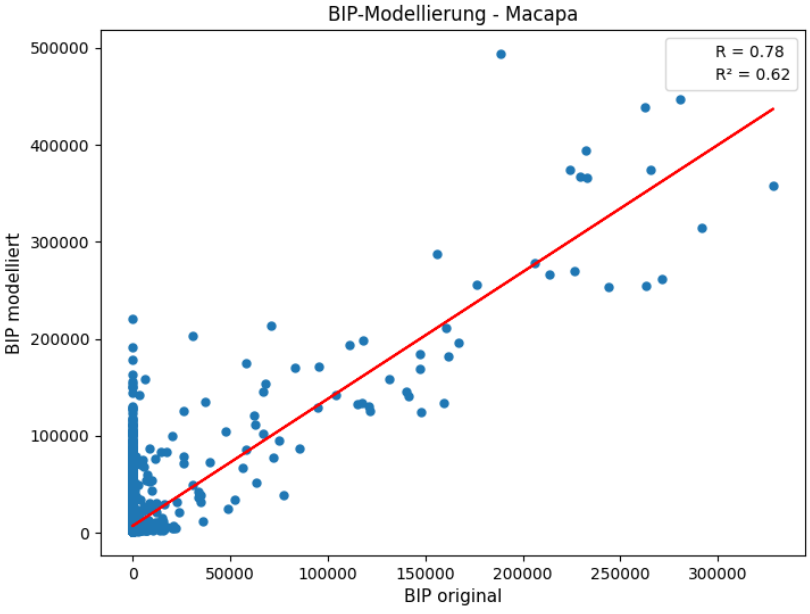

BIP-Modellierung - Boa Vista
R = 0.84
R² = 0.7
BIP modelliert
BIP original
0
100000
200000
300000
400000
0
50000
100000
150000
200000
250000

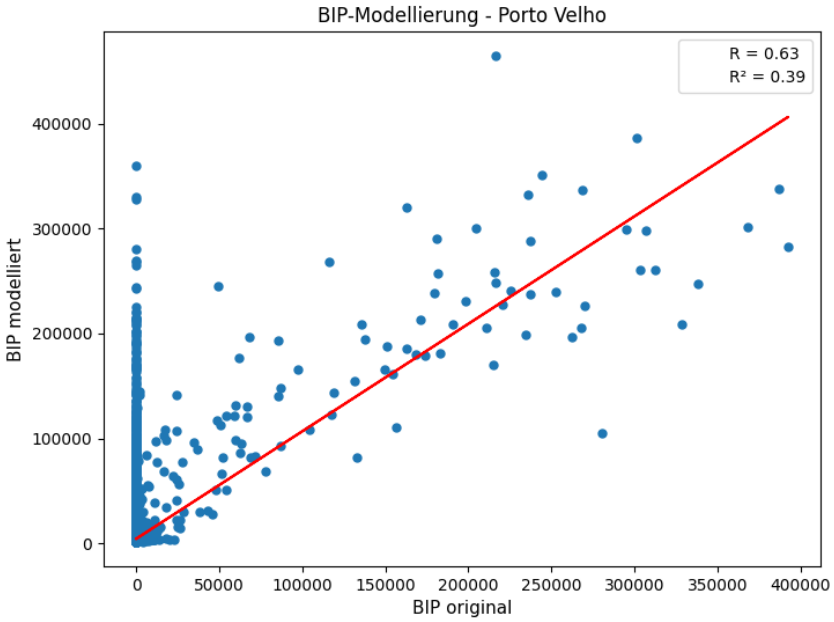

BIP-Modellierung - Curitiba
1e6
R = 0.83
R² = 0.7
BIP modelliert
BIP original
1e6
0.0
0.5
1.0
1.5
2.0
2.5

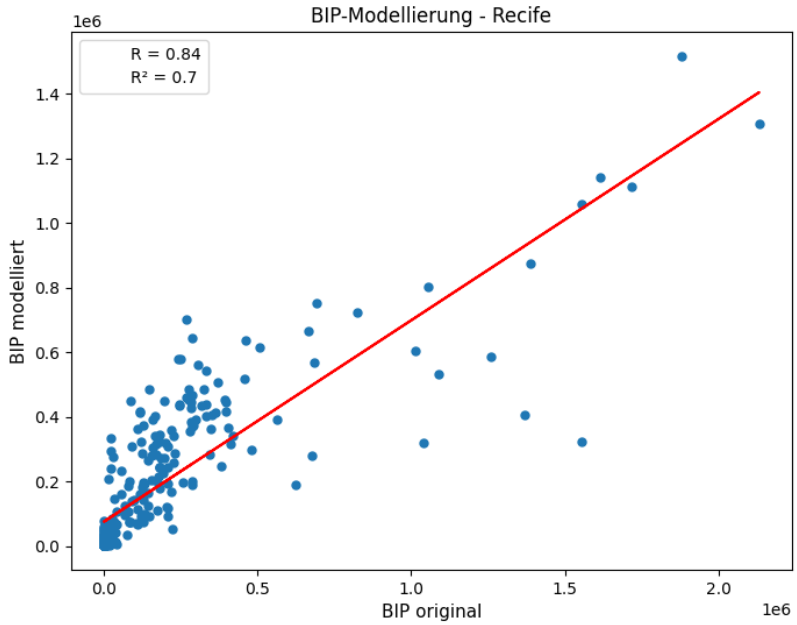

BIP-Modellierung - Florianopolis
1e6
R = 0.81
R² = 0.66
BIP modelliert
BIP original
1e6
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

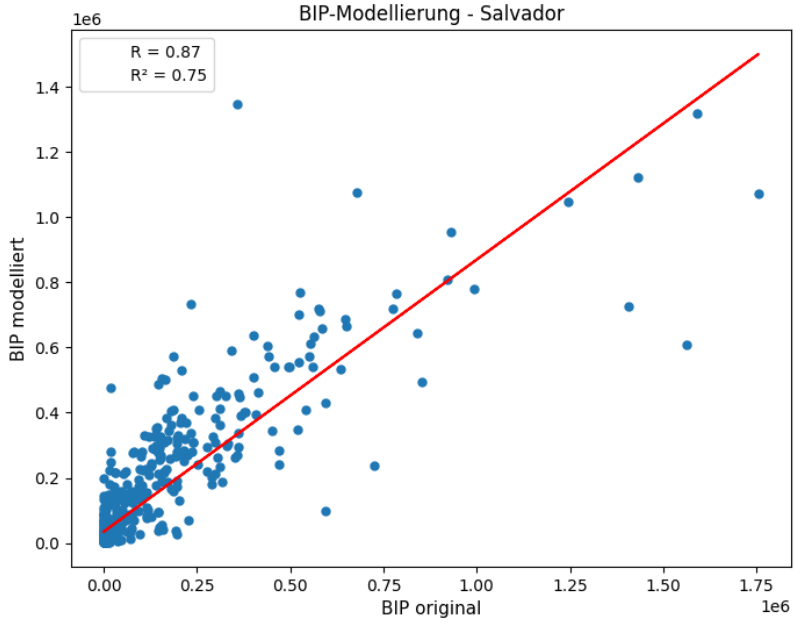

BIP-Modellierung - Goiania
1e6
R = 0.84
R² = 0.7
BIP modelliert
BIP original
1e6

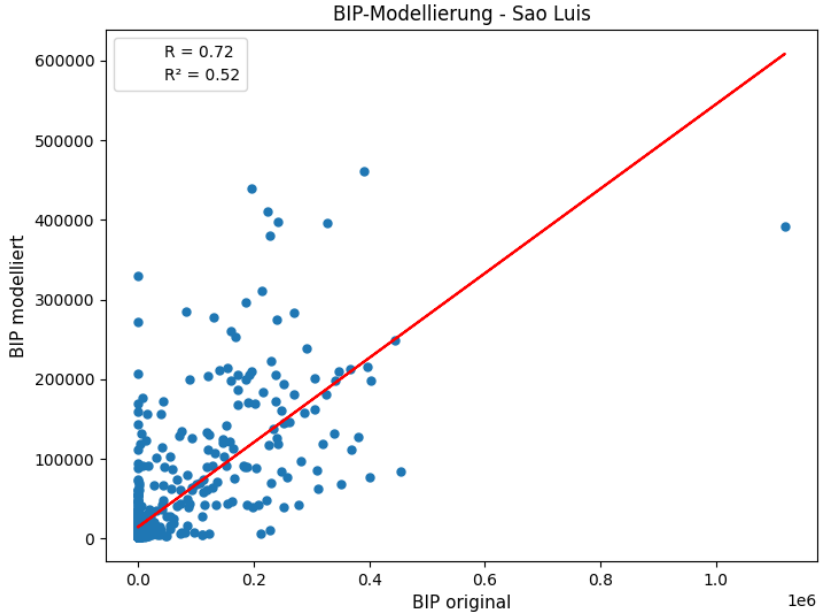

BIP-Modellierung - Joao Pessoa
1e6
R = 0.84
R² = 0.71
BIP modelliert
BIP original

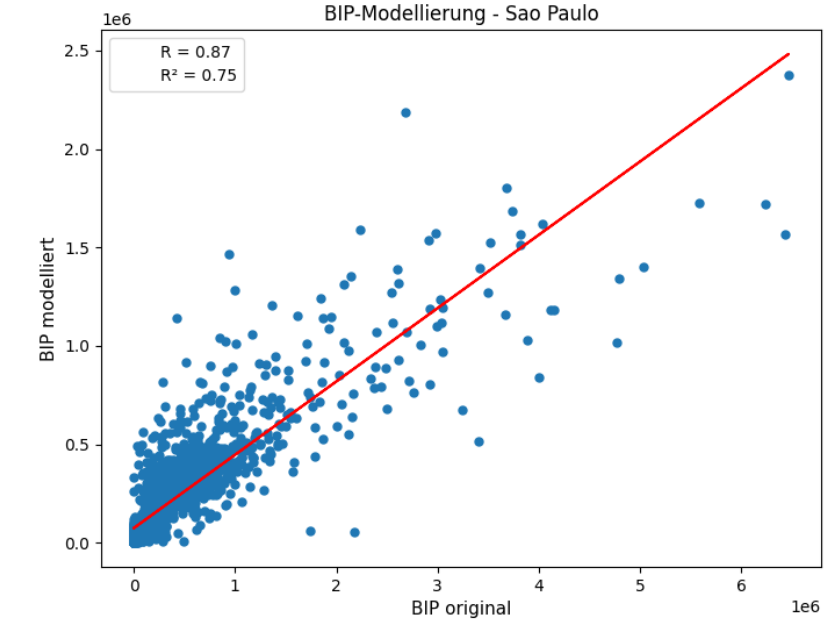

BIP-Modellierung - Macapa
R = 0.78
R² = 0.62
BIP modelliert
BIP original

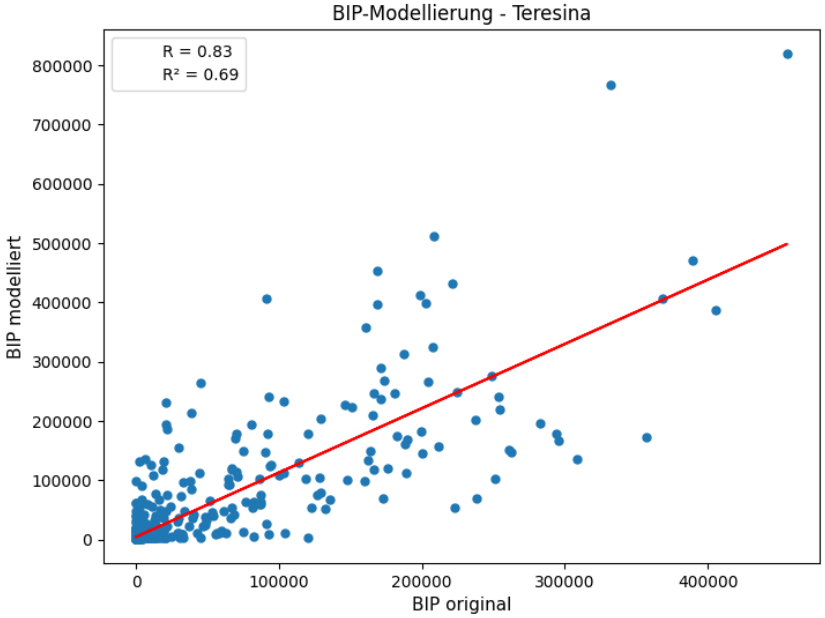

BIP-Modellierung - Porto Velho
R = 0.63
R² = 0.39
BIP modelliert
BIP original
0
100000
200000
300000
400000
0
50000
100000
150000
200000
250000
300000
350000
400000


BIP-Modellierung - Recife
1e6
R = 0.84
R² = 0.7
BIP modelliert
BIP original
1e6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.5
1.0
1.5
2.0


BIP-Modellierung - Salvador
1e6
R = 0.87
R² = 0.75
BIP modelliert
BIP original
1e6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


BIP-Modellierung - Sao Luis
R = 0.72
R² = 0.52
BIP modelliert
BIP original
1e6


BIP-Modellierung - Sao Paulo
1e6
R = 0.87
R² = 0.75
BIP modelliert
BIP original
1e6


BIP-Modellierung - Teresina
R = 0.83
R² = 0.69
BIP modelliert
BIP original