Master Thesis

submitted within the UNIGIS MSc programme
at Z_GIS
University of Salzburg

Cross-Border Infrastructure in Africa

An Analysis through Logistic Regression

by

Dipl. Eng. Roman Meyer
UniGIS ID 104502

A thesis submitted in partial fulfilment of the requirements of
the degree of
Master of Science (Geographical Information Science & Systems) – MSc (GISc)

Advisor:

Dr. Christian Neuwirth

Kigali, 21.03.2019

Science Pledge

By my signature below, I certify that my thesis is entirely the result of my own work. I have cited all sources I have used in my thesis and I have always indicated their origin.

Kigali, 21.03.2019

Place, Date Signature

Abstract

Cross-Border Infrastructure (CBI) are a concrete manifestation of regional integration. While different factors may enable or hinder the construction of CBI, their presence/absence could be an indication for strong/weak ties between countries. Using logistic regression, the present thesis therefore analyses the occurrence of CBI in Africa in relation to geometrical, topographical, sociological, economical, and political factors.

Two different kinds of CBI are evaluated: 1) large CBI which require rather big investments, such as railways, hydropower dams, ferries or official border posts, and 2) paved roads, a more frequently occurring form of CBI that represent a smaller financial commitment. While the latter are extracted from Open Street Map data (OSM, 2018), the former are based on a variety of crowd sourced data and satellite imagery. Building on the reference geometry of the Large Scale International Boundary dataset (Department of State, 2017), data is prepared for thirteen independent variables. This leads to the establishment of continental and regional logistic regression models. To test them, coefficients from the east African model is used to predict CBI in southern Africa.

The study shows some important results. 1) CBI are most likely to occur in areas of high population density, close to capitals, in flat terrain and on dry land or across small rivulets. As the analysis is raster-based, the length of the border segment in the raster cell also shows a positive correlation. 2) Political (Regional Economic Communities, RECs), sociological (migration) and economical (trade, GDP) indicators show no consistent correlation to CBI. These are also the ones where available data refers to an entire border, not an individual raster cell. 3) Regional models are more precise than continental ones. 4) Coefficients calculated for one region shouldn't be applied to another region. Overall, the models explain between 5% and 21% of the occurrence of CBI (Pseudo-R2). The present work is a successful proof of concept, which can now be further refined.

Acknowledgements

I'd like to acknowledge the support received from Dr. Christian Neuwirth. His technical expertise was vital for the success of this thesis.

Kudos also to the lecturers of the various modules as well as the UniGIS Salzburg administration team, their quick responses to any concern made life so much easier.

Last but not least, I also want to thank my wife Nancy for the endurance of endless hours that I spent working on this thesis when we should have enjoyed a nice evening or weekend together.

Table of contents

Science Pledge 2
Abstract 3
Acknowledgements 4
Table of contents 5
List of figures 7
List of tables 8
List of acronyms 9
1. Introduction 1
1.1 Motivation 1
1.2 Literature review 3
1.2.1 General border research 3
1.2.2 Cross-Border Infrastructure 3
1.3 Aim and objectives 5
1.4 General workflow & structure of this thesis 6
2. Materials, Methods 7
2.1 Methodology 7
2.1.1 Evaluation of different methods 7
2.1.2 Critical discussion of the logistic regression methodology 7
2.1.3 Literature research on logistic regression 10
2.1.4 Software used 11
2.2 Data 12
2.2.1 Large Scale International Boundaries (LSIB) 13
2.2.2 Open Street Map (OSM) 15
2.2.3 Satellite Imagery 17
2.3 Variables 18
2.3.1 Preparation of dependent variables 18
2.3.1.1 y01 Paved OSM roads 18
2.3.1.2 y03 Large CBI 23
2.3.2 Preparation of independent variables 28
2.3.2.1 x01 Pixel Size 29
2.3.2.2 x02 Border length 29
2.3.2.3 x03 Distance from coast 30

2.3.2.4 x04 Distance to nearest capital 30
2.3.2.5 x05 Elevation 31
2.3.2.6 x06 Slope 31
2.3.2.7 x07 River size 31
2.3.2.8 x08 Population Density 33
2.3.2.9 x09 Fraternisation 34
2.3.2.10 x10 GDP per Capita 35
2.3.2.11 x11 Trade 35
2.3.2.12 x12 Shared RECs memberships 36
2.3.2.13 x13 Differing RECs memberships 38
2.4 Analysis 40
2.4.1 Different datasets for analysis 40
2.4.2 Rare events, rule of 10 42
2.4.3 Correlation matrix 44
2.4.4 Variance inflation factor (VIF) 45
2.4.5 Stepwise regression 46
2.4.6 Prediction of CBI 46
3. Results 47
3.1 Coefficients 47
3.2 Pseudo R2 49
3.3 Residual deviance, AIC 49
3.4 Visualisation 50
3.5 Prediction of CBI in southern Africa 53
4. Discussion 54
4.1 Expected, ambiguous and unexpected coefficients 54
4.2 Pseudo-R2 56
4.3 Comparison of continental and regional models, prediction 56
4.4 Limitations, challenges 57
5. Conclusion 59
6. References 61

List of figures

Fig. 1 General workflow 6
Fig. 2 Workflow between GIS and statistics software 8
Fig. 3 General workflow during the data preparation phase 13
Fig. 4 Kenya-Tanzania, three OSM-LSIB intersections within 160m 18
Fig. 5 L-shaped distribution of distance (in degrees) to nearest other CBI, ... 19
Fig. 6 Malawi-Mozambique, roads cross but don't really enter into Malawi 20
Fig. 7 Distribution of intersection angles between OSM and LSIB 21
Fig. 8 Botswana-Namibia, the border is perceived to be in between two parallel roads 21
Fig. 9 South Sudan-Uganda, OSM roads cross, but don't continue on the other side 22
Fig. 10 Tanzania-Uganda, roads marked as "paved" in OSM are often mere tracks 22
Fig. 11 Uganda-S. Sudan, Algeria-Morocco, paved vs. unpaved vs. non-existent roads 25
Fig. 12 Botswana-S. Africa, Lesotho-S. Africa, cable car crossing, one-sided crossing 26
Fig. 13 Large CBI by category, as discussed in this chapter 27
Fig. 14 Shortest (left) and longest (right) border line segments in any raster cell 29
Fig. 15 Burundi-Rwanda, overlay of LSIB (yellow), HydroSHEDS (green), IRBD (blue) 32
Fig. 16 Workflow to create the river size variable 32
Fig. 17 HydroSHEDS river segments showing the number of contributing cells 32
Fig. 18 LSIB segments with the "upcell" value inherited from HydroSHEDS lines 33
Fig. 19 Geographical extent of the eight RECs recognised by the AU (ecdpm, 2018) 37
Fig. 20 CBI (red) in the different datasets to be analysed 41
Fig. 21 Correlation matrices for all datasets 44
Fig. 22 Illustration of low (left) vs. high (right) R2 in linear regression (minitab, 2014) 49
Fig. 23 Predicted probabilities by y03 (left) and y01 (right) models 50
Fig. 24 Probabilities as predicted by regional models, higher probabilities drawn on top 51
Fig. 25 Probability as pie chart, size = probability, contribution by variables 51
Fig. 26 Comparison of actual CBI (horizontal) vs. sum of prediction (vertical) 52
Fig. 27 Predicted CBI (y01) using southern (left) and eastern (right) African coefficients 53

List of tables

Table 1 Different data sources used, in alphabetical order....12
Table 2 ISO-3166-alpha-2 codes for African countries, sorted by country name....14
Table 3 Number of large CBI per category....27
Table 4 Overview of independent variables and their distribution....28
Table 5 Histograms of dependent and independent variables....39
Table 6 Number of events vs. non-events per dataset....42
Table 7 VIF including all variables....45
Table 8 VIF after removal of VIF > 10....45
Table 9 Variables removed after stepwise regression....46
Table 10 Overview of coefficients of the logistic regression....47
Table 11 Pseudo R2 values of the different models....49
Table 12 Null deviance, residual deviance and Akaike Information Criterion....49
Table 13 Distribution of deviance residuals....50

List of acronyms

AIC	Akaike Information Criterion
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AU	African Union
AUBP	African Union Border Programme
CBI	Cross-Border Infrastructure
CEN-SAD	Community of Sahel-Saharan States
COMESA	Common Market for Eastern and Southern Africa
DTM	Digital Terrain Model
EAC	East African Community
ECCAS	Economic Community of Central African States
ECOWAS	Economic Community of West African States
EU	European Union
GDP	Gross Domestic Product
GIS	Geographic Information System
GIZ	Gesellschaft für Internationale Zusammenarbeit
HydroSHEDS	Hydrological data and maps based on SHuttle Elevation Derivatives at multiple Scales
IBRU	International Boundary Research Unit, University of Durham
IGAD	Intergovernmental Authority on Development
IRBD	International River Boundary Database
LSIB	Large Scale International Boundary dataset
MAUP	modifiable areal unit problem
NATO	North Atlantic Treaty Organization
NMA	National Mapping Agencies
OAU	Organisation of African Unity
OSBP	One Stop Border Post
OSM	Open Street Map
REC	Regional Economic Community
SADC	Southern African Development Community
SRTM	Shuttle Radar Topography Mission
UMA	Arab Maghreb Union
UNECA	United Nations Economic Commission for Africa
USA	United States of America
VGI	Volunteered Geographic Information
VIF	Variance inflation factor
WGS84	World Geodetic System 1984

1. Introduction

After giving some background on how and why this thesis came about (1.1), this chapter explores existing literature (1.2), articulates the aims and objectives of this thesis (1.3) and outlines its overall structure and workflow (1.4).

1.1 Motivation

For more than five years, I've worked on African borders, and I've come across many anomalies, special cases and profoundly interesting stories around these boundaries that are so similar and yet so different to international borders elsewhere. How they evolved, how some were drawn with a ruler, at the range of a cannon ball, or placed on mountaintops that were yet to be discovered. But there were two borders that puzzled me in a different way. The border of Mozambique and Tanzania is about 700km long. The one between Namibia and Botswana is 1500km long. And yet, they only have two and three official border crossing posts respectively. I got curious and checked the US-Canada border, where I found an official crossing every 65km. A big question arose: why are there so few crossings here, compared to so many there?

Of course, the United States and Canada are among the most developed and economically powerful nations. The two African border areas mentioned are among the least populated places on the continent. But still, does this explain why there seems to be almost no exchange between two neighbouring countries?

African borders have an intricate history. Drawn by colonial powers, they were adopted as "intangible" by African Heads of State and Government in their first meeting in Cairo, Egypt (OAU, 1964) to prevent border disputes among the young nations. In 2002, the AU agreed to re-affirm these boundaries through better demarcation, again as a measure of conflict prevention (AU, 2002). The current deadline is 2022. In 2007, the AU Border Programme (AUBP) Unit was launched to facilitate and coordinate the process. One year later, Germany responded positively to an AU request for support to the AUBP and tasked its implementing agency GIZ to launch a support programme in that regard. The author is a member of this support team since April 2013.

While the promotion of cross-border cooperation is among the core mandates of the AUBP, establishment of Cross-Border Infrastructure (CBI) is not. However, CBI can be seen as a tangible proof of good neighbourly relations, as a manifestation of regional integration. It would therefore be interesting, in a first step, to find out where CBI exist and where they don't. And then, once we have an overview, to analyse what enables or hinders the occurrence of CBI. Even if we may not be able to distinguish cause from effect, it would be enlightening to understand the correlation between CBI and other phenomena.

Regional integration can't be measured in absolute terms. It is not a physical reality like temperature or wind speed, it has to be measured through indicators. Trade volume, free movement of people and goods, cultural proximity, language, intermarriages, or the number of bilateral agreements could arguably serve as a valid basis for measurement. Numerous studies have been conducted on continental trade, while anthropological, sociological and cultural research, due to their nature, usually focus on a geographically small area.

A simple and concrete proxy indicator for integration is therefore the existence – or absence – of CBI. Bridges, roads and hydropower dams across borders don't only require joint investment, but also political will, stable relations and constant exchange to ensure their maintenance. They exemplify the difference between real integration and simple co-existence, the difference between borders such as Belgium-Holland and others like North-South Korea.

Analysing the occurrence of CBI and putting them in relation to other phenomena allows revealing underlying patterns that enable or hinder the construction of such infrastructure. At present there is no dataset on cross-border infrastructure available, and no analysis of enabling / disabling factors has been detected in the body of literature. While there may be many reasons for the presence or absence of CBI, it's time to have a closer look and analyse Cross-Border Infrastructure.

1.2 Literature review

There is no unique trail of literature available where this research could directly build upon. However, there are various articles and books that touch upon different aspects of this thesis. They can be broadly categorised into four groups: general border research, cross-border infrastructure, OSM and logistic regression. The last two are discussed in chapters2.2.2 and 2.1.1 respectively.

1.2.1 General border research

Researchers from different backgrounds have analysed (and continue to analyse) international borders from their respective angles. Worth mentioning are publications by economic researchers on borders' effect on trade (Anderson & van Wincoop, 2004; Hartzenberg, 2011; Valensisi, Lisinge, & Karingi, 2016). The United Nations Economic Commission for Africa UNECA publishes a bi-annual Assessment of the Regional Integration in Africa (UNECA, 2016). Others focus on the legal status and history (Brownlie, 1979; Hertslet, 1896). John W. Donaldson (2007, 2009, 2011) has taken on the herculean task to study all river borders around the world and compile them in a database. Others still have analysed borders in the light of peace and security issues (Bah, 2013), not to mention the plethora of work done by social scientists, especially anthropologists and political scientists. There are also various publications by the University of Durham's International Boundary Research Unit (IBRU) as well as by the French diplomat Michel Foucher that provide further reading to interested audiences.

1.2.2 Cross-Border Infrastructure

Two teams of researchers (Fung, García-Herrero, & Ng, 2011; Van der Geest & Nunez-Ferrer, 2011) tried to find prerequisites for building CBI. Both identify the involvement of top-level government as crucial and point out the positive effects that neutral third party coordinators can have on the project (e.g. regional development banks, other supra-national organisations). Van der Geest & Nunez-Ferrer also identify local support for the project as vital. However, they don't consider geographic aspects of CBI.

Several economists have analysed the implications of borders and CBI. Gilbert & Banik (2010) analyse the economic impact of the transportation network and CBI in South Asia, focussing on different economic models. Mun & Nakagawa (2010) compare

different investment and pricing mechanisms of public, private, and mixed transport, and analyse which ones are preferential over others. Srinivasan (2012) proves that CBI investments benefit the poor, especially in land-locked countries. However, additional infrastructure investment (i.e. roads connecting to a new bridge) may be needed within the respective countries. He calls for governments and international bodies to invest in CBI urgently. Warr, Menon, & Yusuf (2010) go in the same direction. They analyse the effect of CBI on both side of the border and find increased economic activity both in the poorer as well as the richer country.

Others focus more on specific borders, enquiring why the expansion of interconnectors in a common EU electricity market along the German-Polish border is so slow (Puka & Szulecki, 2014) or analysing CBI along the Polish-Slovak border (Michniak, 2011) that were built using EU funding. An interesting study on the permeability of borders by pedestrians (Hisakawa, Jankowski, & Paulus, 2013) was done on the Austrian-Italian-Slovenian border in Carinthia. They use topographic and road network data to analyse which parts of the border are easier to cross than others. On the US-Mexican border, a linear regression analysis was carried out on the Nogales twin-city to forecast expected growth on either side (Norman, Feller, & Phillip Guertin, 2009)

Finally, Biger (2013) focuses on a special kind of CBI. He analyses all kinds of barriers (walls, fences, etc.) that were and are being built along international boundaries. He concludes that as long as there are large differences between countries, some will continue to build walls, and others will continue to climb and cross them.

In conclusion, the academic literature available is a patchwork of different topics that touch on this or that aspect of CBI and Regional Integration. Yet a thorough analysis of the situation in Africa has not been developed to this day.

1.3 Aim and objectives

While economists demand open borders and African politicians and visionaries call for continental integration, and despite the fact that CBI has positive effects on both sides of the border, there is no dataset on the current status of CBI in Africa, and no analysis of how CBI correlate with other factors.

The aim of this research is to analyse the correlation between the occurrence of cross-border infrastructure and other geometrical, topographical, sociological, economical, and political factors in order to reveal underlying patterns and geographic distribution.

More specifically, the objectives of this thesis are the following:

- To create a dataset of large CBI as well as paved international roads on the African continent.
- To identify and pre-process data that potentially correlates with CBI, i.e. trade, economic power, regional blocks, population density etc.
- To analyse the correlation between the above datasets on the continent as well as in five selected regions using logistic regression.
- To predict the likelihood of CBI in southern Africa using the coefficients from the eastern African region.
- To summarise and visualise key findings.

International lake and maritime borders are not part of this research.

1.4 General workflow & structure of this thesis

The structure of this thesis as well as the workflow that led to its finalisation roughly followed the typical work procedure of an academic paper and is shown below for easier visualisation.

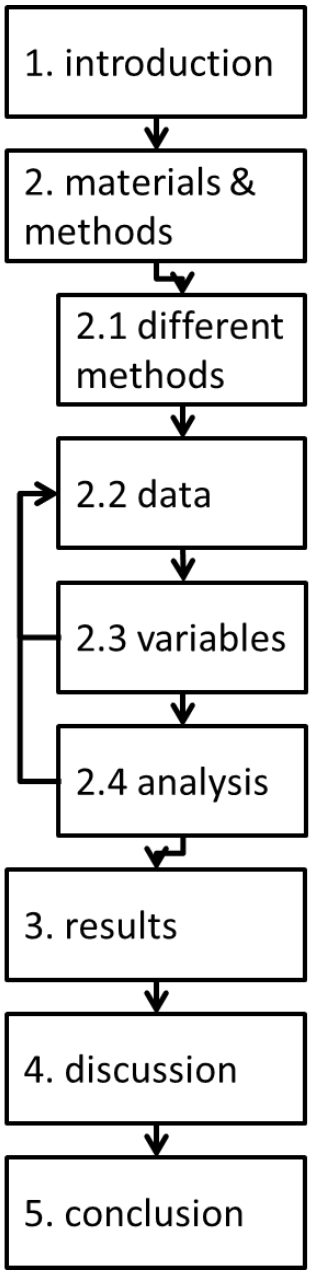


Fig. 1 General workflow

After the motivation and literature research, the aims and objectives of this thesis conclude the first chapter.

Chapter 2 contains the majority of the work, leading up to the analysis of the data. This includes the evaluation of different analysis methods, the search for and preparation of data as well as the values for the different variables extracted from them. As there is very little data on CBI available, this data had to be extracted from different data sources. The analysis of the variables, which is a prerequisite to the logistic regression, as well as the actual logistic regression itself, form the final part of this chapter. Some of the findings of the first round of preparation of variables as well as initial analyses showed some gaps and weaknesses in the data and prompted further research for additional data.

Chapter 3 lists the results of the analysis in table form and makes some attempts at visualisation, including maps, which is difficult given the geographic nature of borders and CBI.

Chapter 4 and 5, finally, discuss the findings and attempt a further outlook on what this thesis means for further research in the area of Cross-Border Infrastructure.

2. Materials, Methods

To predict CBI on the African continent as outlined under aim and objectives, different methods (2.1) were evaluated before settling for logistic regression. The term "prediction" is used in the sense of "estimating the likelihood of a raster cell to contain a CBI throughout this thesis. Thereafter, data from various sources (2.2) was researched, prepared and exported as a set of variables (2.3). Based on this, continental and regional models were created and analysed (2.4).

2.1 Methodology

2.1.1 Evaluation of different methods

Geographically Weighted Regression is used for continuous spatial phenomena and is used for example to estimate the thickness of a sediment layer anywhere in the research area using the results of only a few boreholes. CBI only occur on borders, which makes interpolation inappropriate for estimations.

Linear Regression is used if the dependent variable can be quantified, e.g. to estimate the price per square meter of a plot of land based on factors like distance to city centre, accessibility, orientation, slope, etc. As we only have data on whether a CBI exists or not, linear regression is not possible. The idea of estimating the costs of a CBI was dismissed, as no reliable estimation mechanism was found. In order not to compare apples with oranges, smaller investments (paved roads) were analysed separately from large CBI.

Ultimately, *Logistic Regression* was used, see next chapter.

2.1.2 Critical discussion of the logistic regression methodology

Logistic regression is a statistical analysis methodology without any spatial aspect to it. It is widely used e.g. in medical or educational research, where a dependent variable (e.g. risk of developing tumour X, rare genetic disease Y, dropping out of school or scoring higher than average in chemistry) is analysed against a set of independent variables such as age, sex, social class or excessive consumption of fast food. However, due to the wealth of geodata available and the possibility of overlaying different

thematic data, logistic regression is today also widely used in geospatial sciences. Such analysis is done locally, meaning that every raster cell is analysed independently from its neighbours.

Logistic Regression has the advantage of only predicting the likelihood of a CBI being present, but not its size or value. In the present thesis, for example, we estimate the probability of a paved road in any border raster cell. The logistic regression doesn't make any differentiation between one, two or six lanes as well as between brand new or very old roads. The result is binary, e.g. 30% to find a paved road in a given raster cell vs. 70% to not find one. This necessitates the below work flow, which could potentially bring a new, well-known problem in GIS: the *atomic fallacy* (treating elements as if they were not spatially related). While it is correct that the statistics software disregards any geographical component, the results (coefficients) are then transferred back into GIS, which allows for further spatial inspection and analysis.

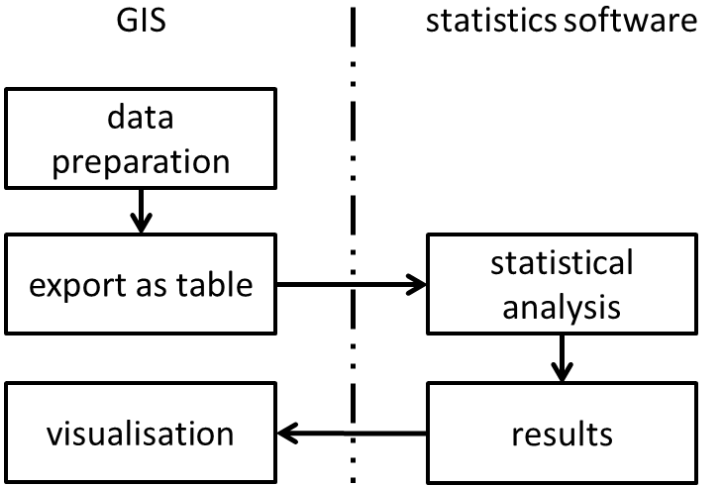


Fig. 2 Workflow between GIS and statistics software

Another potential challenge is the so called *modifiable areal unit problem (MAUP)*. This is the case when e.g. the average income in research areas change drastically depending on where the border between these areas is drawn. The analysis in this thesis is based on a 15" raster grid (see "Sensitivity Analysis" under 2.3.1 for background information on the pixel size). Both the size of the pixel in square meters as well as the length of the border segment in any pixel are independent variables (x01 and x02), which minimises the impact of the MAUP. For other variables, such as distance from coast, elevation, GDP etc., changing the raster size would not fundamentally change the

value of the raster cells. The only exemption may be the slope, which becomes flatter and less meaningful as the cell size increases.

Logistic regression, however, brings another challenge. Its underlying assumption is that of a *linear trend*, i.e. the more a pupil studies, the higher his/her chance of passing the exam. For the CBI-analysis, it will detect whether e.g. altitude has a positive or negative effect on the occurrence of CBI (higher elevation = higher or lower chance of CBI). If there is something of a "habitable zone" for CBI at an altitude between 500m to 1000m a.m.s.l., logistic regression would not be able to detect such a zone.

One prerequisite of logistic regression is the absence of *multicollinearity*, meaning that the variables have to be independent from one another as much as possible. The checks performed on the actual data to evaluate and counter this problem can be found in chapter 2.4.4

Lastly, it is important to note that *correlation is different from causality*. The logistic regression carried out in this research is unsuitable to make any statement on what causes the existence of CBI. High economic power of both neighbouring countries may prove to correlate with lots of infrastructure along their common border. However, the analysis can't prove that either a) economic wealth leads to countries investing in CBI, nor that b) investment in CBI will make a country wealthier.

Logistic regression used in this research, and therefore this research itself, analyses the co-occurrence, or correlation, of cross-border infrastructure and other phenomena. It's not trying to resolve the chicken-and-egg conundrum, if an egg will result in a chicken, or a chicken causes eggs. It rather looks at a clutch of chicken and checks if the black chicken (dependent variable) is friends, foes or indifferent towards each of the brown chicken (independent variables).

Logistic regression helps to make observations. If the independent variable A increases, the dependent variable Y also increases. It's only natural to then start wondering *why* this is the case. However, this research can't and won't make any assumption as to why CBI are present in some locations but not others.

2.1.3 Literature research on logistic regression

As mentioned above, logistic regression is widely used in the wide fields of health and education. Notable recent research includes the analysis of rinderpest in Somalia (Ortiz-Pelaez et al., 2010), a comparison of remote sensing versus climatic models for the spread of the blue-tongue virus on the Calabria peninsula (Van Doninck et al., 2014), a comparison between SRTM and ASTER data for malaria prediction in Kenya (Nmor et al., 2013), the risk of nitrate contamination of wells in Polk County, Florida (Dixon, 2009), as well as a comparison of different methods for evaluating accessibility to and satisfaction with health care providers in Quito, Ecuador (Cabrera-Barona, Blaschke, & Kienberger, 2017).

In a sister field of epidemiology, namely animal distribution, logistic regression has been used to analyse the habitats of the Gaur, also called Indian Bison, and what areas would be most suited for their conservation (Imam & Kushwaha, 2013), as well as the occurrence of lynx in Poland compared to environmental and anthropological factors (Niedziałkowska et al., 2006).

Another field of research is disaster prevention. The probability of human caused grassland fires (Zhang, Zhang, & Zhou, 2010) and forest fires (Srivastava, Saran, de By, & Dadhwal, 2014) have been found to correlate with anthropogenic activities, like poaching for example, as well as lack of precipitation or altitude. A similar study has found correlating factors for deforestation (Pir Bavaghar, 2016). Similarly, landslides have been analysed with a focus on new statistical concepts for Ecuador (Guns & Vanacker, 2012), the Calabria region in Italy (Sorriso-Valvo, Greco, & Catalano, 2008) as well as Kansas in the US (Ohlmacher & Davis, 2003). Also, flood prone areas in China (Shafapour Tehrany et al., 2017) were detected using logistic regression.

But logistic regression is also used in fields closer to planning and social sciences, such as the succession of farms by family or non-family in upper Austria (Stiglbauer & Weiss, 2000) or the expansion of urban boundaries in the United States (Tayyebi, Perry, & Tayyebi, 2014).

2.1.4 Software used

The following software was used for the compilation of this thesis:

- ArcGIS v10.5.1
- QGIS v3.0.0
- R Studio v1.1.456
- Microsoft Office 2013

2.2 Data

The data used in this research comes in all shapes and sizes, and it stems from a plethora of different sources. This chapter gives a general overview as well as insight into three central datasets: the LSIB, OSM, and satellite imagery. All other data is discussed in chapter 2.3.2 under the respective variable it helped to generate.

An initial round of analysis showed some gaps in the data. For example, the population density dataset had some NULL values that were then filled using the surrounding attribute values. Other gaps were filled through additional data research. The only exemptions to this are the borders of the Sahrawi Republic (Western Sahara), for which no trade data was available. Its borders had to be removed from the research. With regards to the coordinate system used, the processing of all geodata was done in WGS84.

Table 1 Different data sources used, in alphabetical order

Dataset	Source	Geometry	Version
Airports	openflights.org	Vector	2017
Capitals	Esri	Vector	10.5.1
Fraternisation	IOM	Table	2015
GDP	World Bank	Table	2017
Hydropower plants	Wikipedia	Table	2018
HydroSHEDS	Lehner, Verdin, & Jarvis	Vector	2008
IRBD	J. Donaldson	Vector	2007
LSIB	US Department of State	Vector	8a
Official crossings	Wikipedia	Table	2018
OSM	OSM	Vector	17.04.2018
Population	FAO	Raster	2015
Railways	Bucsky	Table	2017
RECs	AUC	Table	2018
Satellite imagery	Esri	Raster	2018
Trade	WTO	Table	2017

In a first step, the LSIB dataset was adapted to the needs of this research (see next chapter). Vector and tabular datasets were joined onto the LSIB geometry and then rasterized. Ultimately, values for each cell in the different raster datasets were extracted and exported in a table for further use in the statistics software.

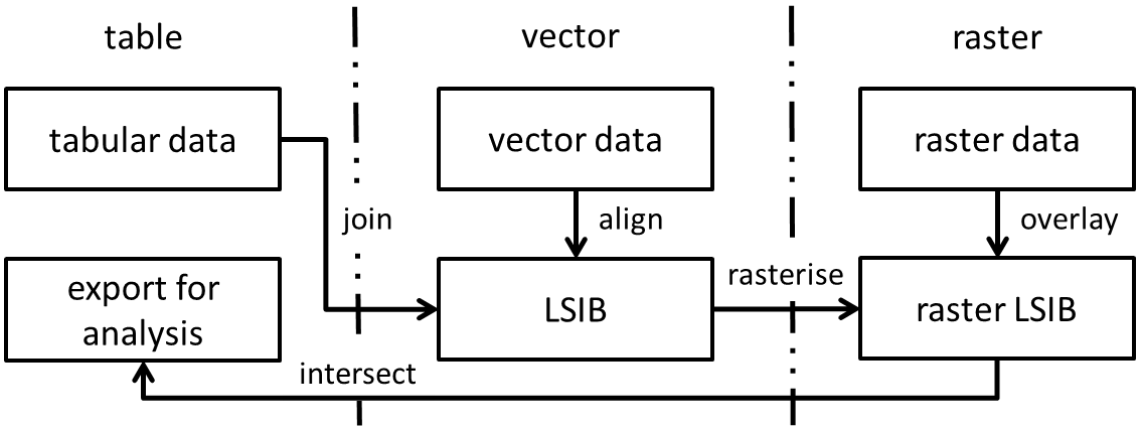


Fig. 3 General workflow during the data preparation phase

2.2.1 Large Scale International Boundaries (LSIB)

The LSIB serves as the *reference geometry* for all subsequent analysis in this thesis. It is the most accurate global dataset on international boundaries that is freely available. More accurate datasets may be available from private companies (e.g. Sovereign Limits by International Mapping Inc.) or individual countries (especially in Western Europe), but these can't be copy-pasted into LSIB for legal reasons (Linthicum, 2017). It is noteworthy that LSIB depicts the de jure boundaries as described in bilateral treaties and accompanying maps, not de facto boundaries due to occupation of land or unilateral claims (Linthicum, 2018).

Some *alterations to the original dataset* had to be made in order to have one single line per border. As some borders are disputed, the LSIB often shows more than one line between neighbouring states. It ranks them as follows:

Rank 1	International boundary
Rank 2	Other line of international separation, i.e. disputed boundaries
Rank 3	Special lines, e.g. Abyei (Sudan–South Sudan)

To have a consistent dataset with a single border line between any two neighbouring countries, the LSIB dataset was altered as follows:

- The Abyei area as being part of Sudan
- The Koualou area as being part of Benin
- The Hala'ib triangle as being part of Egypt
- The Ilemi triangle as being part of Kenya

- The median line in the Congo river as the border between the two Congos
- Spanish enclaves Ceuta and Melilla were disregarded
- Border between Egypt and Israel was disregarded

These alterations were made for practical reasons only and do not constitute any political opinion by the author on the delimitation of these borders. In addition, lake segments were removed from the LSIB using the HydroSHEDS dataset (see last paragraph under 2.1.2).

A *unique identification of borders* was needed to reference thematic data and conduct analyses on the results. The LSIB uses NATO country codes (NATO, 2017) as unique IDs for all countries. For the scope of this thesis, these were replaced with ISO-3166 alpha-2 codes (ISO, 2017), which are also used in the author's daily work. To identify a specific border, a combination of the two codes is used, starting with the one that comes first in the alphabet. For example, the ID for the Burundi-Tanzania border is bitz.

Table 2 ISO-3166-alpha-2 codes for African countries, sorted by country name

ISO_name	ISO2
Algeria	dz
Angola	ao
Benin	bj
Botswana	bw
Burkina Faso	bf
Burundi	bi
Cameroon	cm
Cape Verde	cv
Central African Republic	cf
Chad	td
Comoros	km
Democratic Republic of the Congo	cd
Djibouti	dj
Egypt	eg
Equatorial Guinea	gq
Eritrea	er
Ethiopia	et
Gabon	ga

ISO_name	ISO2
Gambia	gm
Ghana	gh
Guinea	gn
Guinea-Bissau	gw
Ivory Coast	ci
Kenya	ke
Lesotho	ls
Liberia	lr
Libya	ly
Madagascar	mg
Malawi	mw
Mali	ml
Mauritania	mr
Mauritius	mu
Morocco	ma
Mozambique	mz
Namibia	na
Niger	ne
Nigeria	ng

ISO_name	ISO2
Republic of the Congo	cg
Rwanda	rw
São Tomé	st
Senegal	sn
Seychelles	sc
Sierra Leone	sl
Somalia	so
South Africa	za
South Sudan	ss
Sudan	sd
Swaziland	sz
Tanzania	tz
Togo	tg
Tunisia	tn
Uganda	ug
Western Sahara	eh
Zambia	zm
Zimbabwe	zw

Table and vector data were *rasterised* prior to the statistical analysis, as shown in Fig. 3. As raster cells covering tripoints aren't clearly attributable to one of the three border lines, these pixels (61 in total on the African continent) were removed from the dataset. The initial LSIB dataset counts 108 borders on the African continent. Two of them are 100% lake borders (DR Congo-Tanzania and Niger-Chad). One (Botswana-Zambia) is so short that it disappears once the tripoints are removed. Subtracting the borders of the Sahrawi Republic with its three neighbours due to unavailability of data leaves a total of 102 borders for the rest of the analysis.

A *newer version of the LSIB* was released on 30 May 2018 (Department of State & Humanitarian Intervention Unit, 2018). Due to the advanced state of the thesis at that point, this dataset couldn't be incorporated, but might be valuable for use in further research. A comparison of versions 8a and 9 shows significant changes (larger than the 15" raster resolution used for the analysis) on the borders of Guinea-Bissau–Mali, Ivory Coast–Mali, Guinea-Bissau–Liberia, Ivory Coast–Liberia, Nigeria–Cameroon, Somalia–Kenya, Somalia–Djibouti, Tanzania–Mozambique, Zambia–Zimbabwe, as well as individual points on other boundaries. Many of the modifications concern river sections, where it can be assumed that the update is aligning the boundary with a meandering river. Where visual inspection was carried out for Large CBI, a bridge over a river would still be included, even if that bridge doesn't fall on the LSIB version 8a.

2.2.2 Open Street Map (OSM)

OSM is crowd-sourced, volunteered geographic information (VGI). As with all VGI, it is heterogeneous in nature. Roads may be missing or assigned a wrong attribute (e.g. footpath instead of road). Also, OSM is not harmonised with LSIB, or vice versa, meaning that even if "the boundary follows the road", this will result in different geometries for the OSM-road and the LSIB-border (also see Fig. 15).

Literature research

As OSM is one of the fundamental data sources in this research, a closer look at questions on its quality and completeness are pertinent and have become an entire sub-genre of research. Academics have analysed OSM data for China (Y. Zhang, Li, Wang, Bao, & Tian, 2015), the United Kingdom (Haklay, 2010), France (Girres & Touya,

2010) or Tehran (Forghani & Delavar, 2014). While the former focus on the road network, others have analysed land cover classes (Dorn, Törnros, & Zipf, 2015) or tried to derive urban built-up areas from it (Brinkhoff, 2016).

In order to assess *OSM data quality*, different researchers have set their focus on the data contributors, suggesting to eliminate erroneous attribution by analysing and categorising the trajectories (Basiri et al., 2016) i.e. pedestrian, wheelchairs, bicycles, cars; a GPS trail collected at 120km/h can't be a footpath, or categorising data contributors by data accuracy and upload frequency (Parr, 2015), thus trying to identify the more credible users.

Some have also studied *OSM development over time*. (Siebritz, Sithole, & Zlatanova, 2012) have analysed OSM contributions in South Africa from 2006 to 2011, finding that some data is captured preferentially over other, without levelling out over time. Also, the 2010 football world cup had significant impact on the data collection in the proximity of the football stadiums. Accepting heterogeneous data quality as a fact, Razniewski & Nutt (2014) go one step further, proposing how to include such meta-information in spatial queries.

The one OSM research that is probably *most closely linked to borders* is the one scrutinizing the difference between Israel and Palestinian Territories (Bittner, 2017). It finds that Israeli contribute more to OSM than Palestinians. The reason behind is a matter of guessing, but Bittner suggests it may be due to OSM's "stick to the facts on the ground" credo, mapping the world as it is today, and not as it should be. This, from a Palestinian perspective, cements the reality of occupied land, which demotivates them as a group to contribute to OSM.

Other research *focussing on borderlands* include a system that uses geo-tagging of blog posts to address the problem of data scarcity in border regions (Xing, Chen, & Zhou, 2015) as well as the use of daily incremental OSM updates to keep the borderland database up-to-date (Zhou, Zeng, Jiang, Zhou, & Zhao, 2015). A third research finds three major challenges for modelling border areas in GIS data bases: integrated data modelling, comprehensive spatial analysis as well as collaborative geospatial services, i.e. making data available across borders (Chen et al., 2015). They also call for more

collaboration across disciplinary boundaries to leverage the full potential of borderlands.

Use of OSM in this thesis

Due to its sheer size (2.3GB, close to 50 million features), the original downloaded OSM dataset (OSM, 2018) needed to be reduced drastically in order to work with it in a desktop GIS environment. This was achieved in two steps, first filtering out the non-road elements ('highway' <> NULL) such as buildings, parks or coastlines. In a next step, the LSIB was buffered by 0.1° (11km) and used to clip the remaining OSM road network, keeping only the roads in the broader vicinity of the border. This reduced the dataset to a manageable size while still allowing to understand the local context of the road network when visual inspection became necessary. For further use of OSM in this research, see 2.3.1.1

2.2.3 Satellite Imagery

For visual inspection of the situation on the ground, the ArcGIS base layer "world imagery" (esri, 2018) was used for reference. This allowed to understand unclear situations, where (OSM) data was ambiguous or missing. Esri uses data from different providers and combines them into one seamless dataset. Resolution according to Esri is "1m or better in many parts of the world" and is updated regularly. As it is mainly used as background information, no further research on the accuracy or timeliness of the data was conducted.

2.3 Variables

Logistic regressions require two different kinds of variables: a dependent variable and one or more independent variables. For this thesis, two dependent variables were developed, but analysed separately. The preparation of the variables as well as certain specificities are described throughout the next chapters.

2.3.1 Preparation of dependent variables

The dependent variable is the one that is being estimated or predicted, it thus *depends* on the independent ones. Two sets were created: y01 shows paved roads from OSM, y03 contains large CBI. Another set, y02, initially contained all OSM trajectories, including footpaths, trails etc. As these don't really require any form of investment by the central government, its analysis wouldn't respond to the aim of the research, and the variable was dropped.

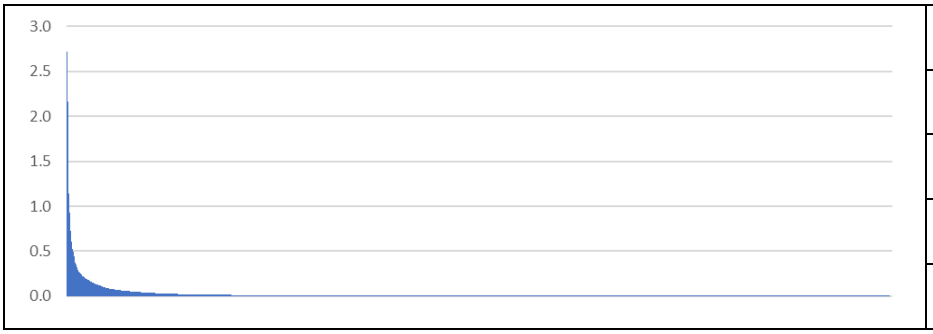
2.3.1.1 y01 Paved OSM roads

In a first step, the cropped and filtered OSM dataset described in 2.2.2, was intersected against the LSIB geometry, which resulted in 10395 border crossings – with some of them as little as only 10cm apart. As this clearly doesn't constitute two separate crossings, and since the analysis would be based on raster cells, a reasonable size for the raster needed to be determined. It should not be too small, so that data size would still be manageable, and not too large, so as not to aggregate relevant data.

Fig. 4 Kenya-Tanzania, three OSM-LSIB intersections within 160m

Raster size & sensitivity analysis

A closer look at the distance from any of the above border crossing to its nearest neighbour showed, unsurprisingly, an L-shaped histogram: many are relatively close to each other, while few are at large distances.



Min	0.0000
Max	2.7169
Mean	0.0211
Median	0.0038
Count	10395

Fig. 5 L-shaped distribution of distance (in degrees) to nearest other CBI,

While the average distance is 0.0211° (2.3km at the equator), the median is 0.0038° (420m). Further inspection revealed that aside from the identical crossings, there are straight border lines that are intersected at short intervals by a meandering road, or vice versa (see "phantom crossings" later in this chapter). In order to have a practicable solution, it was decided to *use a raster resolution of 15"*, which is about 460m along the meridians and the equator. This coincides with the median distance between neighbouring border crossing, as well as the HydroSHEDS data set, which has the same resolution.

It should be noted that this doesn't necessarily ensure a distance of 460m between neighbouring border crossings. Two crossings that are only a few meters apart can still count as two separate crossings if they fall into different raster cells. This, however, is unavoidable with the current approach.

Phantom crossings, misalignments and dead ends

While the above chosen raster size will eliminate some of the irrelevant border crossings, it only takes effect once the OSM-LSIB intersection point layer is converted into raster. There are other challenges to deal with first. The initial intersection results in about 7'000 point features. However, many of them are multi-part features. When converted to single-part, the number goes up to around 11'000. A closer look at the

multi-point features reveals that many of them are not actually distinct border-crossings, but intersections of OSM and LSIB lines that are intertwined with each other. Sometimes this is due to the reality on the ground, sometimes it's due to different scales and accuracy. Here are three examples:

Example 1: Phantom crossings

The image below shows a total of thirteen border crossings for a short stretch of the Malawi-Mozambique border. The geometries of LSIB and OSM in this case may very well be correct. However, at no point do the Mozambican roads actually cross into Malawi for more than a couple of meters, and there's no connection to the Malawian road network.

Fig. 6 Malawi-Mozambique, roads cross but don't really enter into Malawi

In order to get a sense of the magnitude of this problem, the intersection angle between LSIB and OSM was calculated. The result shows that unlike the traditional idea of a border crossing, where the street runs perpendicular to the border, in reality streets and borders intersect at every possible angle. There are two important points from the histogram below: First, roads and borders intersect at any possible angle, with a tendency to 90°, and second, there is a high number of roads that intersect borders at an almost parallel angle.

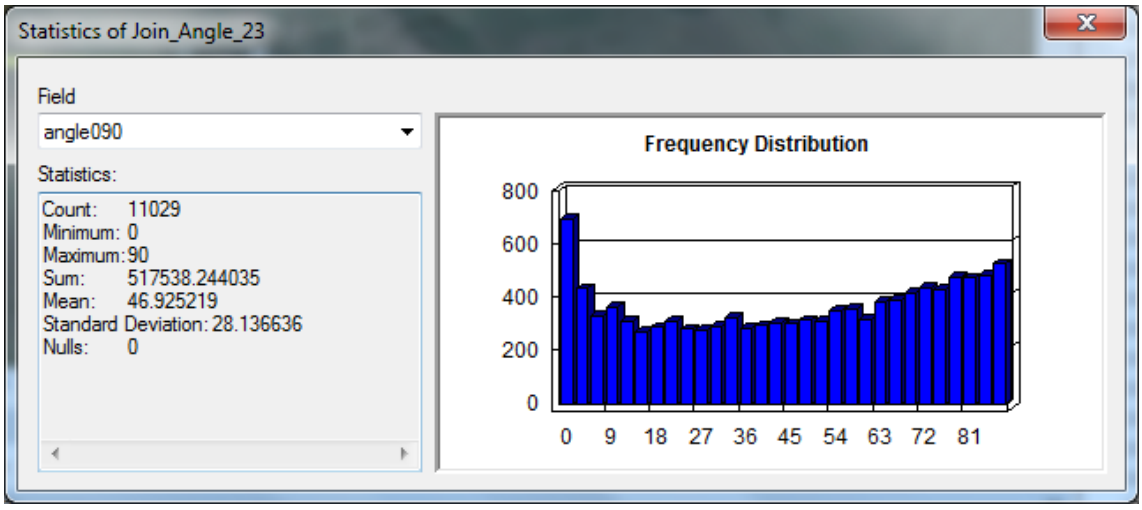


Fig. 7 Distribution of intersection angles between OSM and LSIB

Example 2: Misalignment of LSIB and OSM

To prevent the spread of animal disease, a fence was erected on the border between Botswana and Namibia in the 1960s (Kopiński & Polus, 2017), with maintenance roads running parallel to it. The below image shows that LSIB and OSM don't align, as the border should run in between the two maintenance roads. This may be due to different scale, location accuracy of the satellite imagery, or a slight dislocation of the de facto versus the de jure border. Although the two layers intersect on four occasions, there is no border crossing in the extent shown below.

Fig. 8 Botswana-Namibia, the border is perceived to be in between two parallel roads

Example 3: Dead ends

The last example shows a combination of the two previous ones. First, is seems clear on the image below that the LSIB doesn't align with neither the satellite imagery nor OSM. The important point, however, is to note that although the roads cross the LSIB line, they don't lead anywhere on the other side of the border.

Fig. 9 South Sudan-Uganda, OSM roads cross, but don't continue on the other side

To minimise effects of the above listed problems, all multipart features were replaced by their centroid. While this doesn't resolve all problems, it ensures that roads that run along borders get counted as only one crossing. As centroids don't necessarily fall onto the border line, they were then moved to the closest point on the border.

Erroneous attributes

Another problem encountered in some areas are roads that are seemingly paved, probably due to a default value in OSM that is not adjusted by the users. The below images show the same area. However, the "paved" roads from OSM on the left are not visible on the image on the right. Unfortunately, this isn't something that could be resolved automatically, and it was therefore left as is in the analysis.

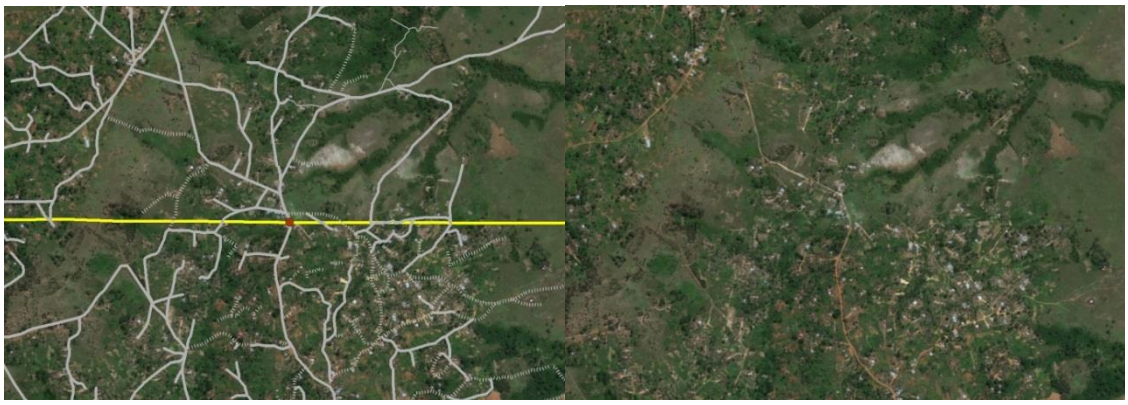


Fig. 10 Tanzania-Uganda, roads marked as "paved" in OSM are often mere tracks

2.3.1.2 y03 Large CBI

Unlike the intersection of OSM with LSBI, the idea behind this variable is that of larger, joint investments made by two countries, such as to build bridges across border rivers or build, staff and operate official border crossings. The term "large CBI" is meant to distinguish it from the OSM intersections – which are also CBI but exemplify a relatively small investment. The below nine categories of large CBI were researched and included in the thesis, with many of them belonging to several categories (e.g. an official border post at a river boundary would be both a border post, a bridge and a paved road.) The classification was kept during the initial phase, but no distinct analysis was made for e.g. hydropower dams only, due to the rare events problem discussed in chapter 2.4.2.

Airports / airfields

While airports may never be built right on a border, some are reasonably close to a border to assume that they're used by people from both countries. Examples worth mentioning are the Geneva airport, which is right at the border and could only be built upon an exchange of land (Bindschedler & Dennery, 1958) between Switzerland and France, or the Basel-Mulhouse bi-national airport that lies on French territory, but is managed jointly by France and Switzerland (Petitpierre & Hoppenot, 1949). Similar examples exist on the African continent, such as the Congolese airports in Goma and Kinshasa that also benefit travellers from Congo-Brazzaville and Rwanda respectively.

Coordinates of airports were downloaded (openflights.org, 2017) and the ones within 10km from LSIB extracted for subsequent analysis. They were then projected onto the closest point on the border.

This approach could be criticised for two reasons: 1) the 10km are arbitrary, and 2) moving the location of the airport to the nearest border means that it may lie in rugged terrain or in a river, whereas it actually lies on plane dry land. Unfortunately, logistic regression only allows for binary values in the dependent variable; a cell can have a CBI or not. It doesn't allow for a "distance to next airport" value. The present approach is a practical solution and may have to be refined in future analyses.

Bridges

A third of African borders are river boundaries, compared to a quarter worldwide (Pratt, 2016). Therefore, a relatively high number of bridges can be expected. However, the definition of a bridge is not quite so simple, as large rivers require large bridges, but small rivulets cross under roads in simple pipes, and it's difficult to draw a line between these categories. While the former surely qualifies as large CBI, the latter doesn't, so including all OSM roads that cross a river would be misleading.

For the scope of this research, bridges are a category of large CBI, but it wasn't analysed further, and bridges have not been analysed separately for the paved OSM roads (y01) variable.

Hydropower plants

While the above-mentioned high percentage of river boundaries on the African continent brings its own challenges (e.g. the question whether meandering rivers imply meandering borders), it also bears the potential for the construction of hydropower plants. Such CBI can only be built and managed by mutual agreement, even if the investment should only come from one side.

While large hydropower plants are easily identifiable on satellite imagery, smaller ones are much more difficult to spot, especially where the plant itself uses a run-of-river system rather than a dam with a large reservoir. And while rapids are easily spotted due to the turmoil and spray, hydropower plants calm the course of a river, which again makes them more difficult to see on satellite imagery.

Unfortunately, there is no complete database on hydropower plants available. As a first iteration, the Wikipedia-List "power plants in Africa" (Wikipedia, 2018b) was used and improved during visual inspection using satellite imagery as much as possible.

Official Border Crossings

In the absence of a registry of official border crossings, crowd sourced data was used (Wikipedia, 2018a), enhanced by country specific sources where available. However, even something as seemingly clear as official crossings aren't always unambiguous. The Ongeluksnek border post (Rothmann, 2013) on the Lesotho – South Africa border has all the facilities on one side, but no corresponding border post on the other (see Fig. 12 below).

Paved roads

Where paved roads are part of a large CBI (e.g. most bridges are paved), these were attributed accordingly, using OSM and satellite imagery. However, a simple paved crossing doesn't constitute a large CBI in itself and thus isn't captured in this variable. It's just a non-mandatory attribute of a large CBI. Also see chapter 2.3.1.1 for further details.

Unpaved roads

Like paved roads, unpaved roads at large CBI were attributed accordingly, but aren't large CBI by themselves.

As the visual inspection showed, roads can also be paved on one side and unpaved or non-existent on the other.

Fig. 11 Uganda-S. Sudan, Algeria-Morocco, paved vs. unpaved vs. non-existent roads

Ferries

Many large rivers have small ferries that can take pedestrians or vehicles across. For official border crossings, these ferries have been identified on satellite imagery and with the help of OSM data, where roads from both sides end on both sides of the river. All of this was done through visual inspection.

Railways

Due to the sheer size of the African continent and the subsequent high transport costs, railway lines could play a vital role in bringing these costs down. However, maintenance of rails and rolling stock brings its own challenges. While some lines built in colonial times are still in regular use today, others have gone into disrepair, and new ones are being built. Railway lines are included in the analysis based on their existence and not their current status of operation. No statement can be made if they're still in operational or completely out of service.

As an initial source, a map of African railway lines (Bucsky, 2017) was used to identify cross-border railways. The map being small scale and rather thematic, the actual location of the rails was identified on satellite imagery. Rails are only a few centimetres thick, with a gauge between 1000 and 1435mm. Although linear features are more easily identifiable on satellite imagery, and the latter's quality improves continuously, spotting rails is still challenging, especially in areas with little vegetation and flat terrain, where no excavation or grading was required. Therefore, lines that aren't on the above-mentioned map couldn't be detected through visual inspection.

Fords

Some unpaved roads cross rivers or rivulets without any bridge, simply passing through the river bed. Where roads were identified and verified on satellite imagery, such fords have been attributed accordingly, but aren't further analysed. Also, much like paved and unpaved roads mentioned above, fords by themselves don't constitute a large CBI.

Another official border crossing across a river is the Pont Drift Border Post (van Heuvel, 2015) which isn't a bridge or a ford, but a cable car, as portrayed below.

Fig. 12 Botswana-S. Africa, Lesotho-S. Africa, cable car crossing, one-sided crossing

All in all, 298 large CBI were identified using different methods and data sources. Below is a summary of the different categories found. Their geographical locations are shown in Fig. 13 below.

Table 3 Number of large CBI per category

Airports / airfields	Bridges	Hydropower plants	Official Crossings	Paved roads	Unpaved road	Ferries	Railways	Fords	Total
39	99	8	233	141	120	18	24	21	298

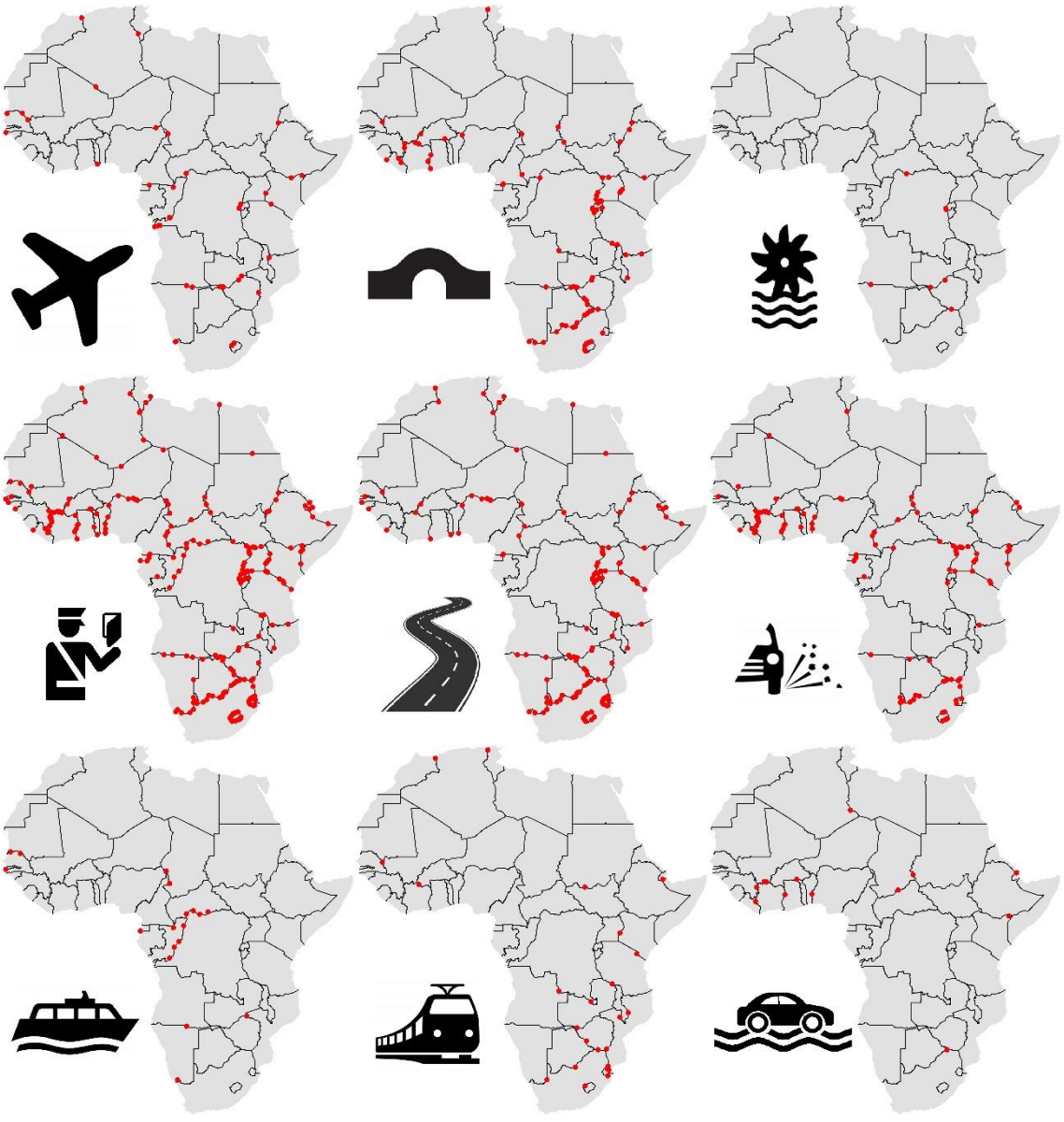


Fig. 13 Large CBI by category, as discussed in this chapter

2.3.2 Preparation of independent variables

In logistic regression, independent variables are the ones that attempt to "explain" the dependent variable, even if there may be no direct cause-and-effect between the two. A total of thirteen independent variables were analysed in the present thesis. They can be roughly grouped into geometrical (1-3), topographical (4-7), sociological (8-9), economical (10-11) and political (12-13) factors. The majority has different values for individual pixels, whereas variables 9-13 have the same values for all the pixels along a specific border.

This chapter will go through the independent variables one by one and lay out their genesis and characteristics. For an overview of the histograms of all variables in addition to the table below, kindly see Table 5 at the end of this chapter. Please note that normal distribution of independent variables is *not a prerequisite* for logistic regression.

Table 4 Overview of independent variables and their distribution

var	value	unit	min	Max	distribution
x01	cell size	m2	171325	214369	L (right-to-left)
x02	border length	m	1	1774	normal
x03	dist. from coast	km	0	1801	linear decreasing
x04	dist. to nearest capital	km	1	1519	normal (left-skewed)
x05	elevation	m	-124	4695	normal (left-skewed)
x06	slope	degree	0	34	L-shaped
x07	river size	km2	21	3664664	L-shaped
x08	Population Density	ppl/km2	0	11657	L-shaped
x09	fraternisation	ppl	514	1835102	L-shaped
x10	GDP per Capita	usd	341	7963	L-shaped
x11	Trade	%	12	59	normal
x12	Shared RECs	number	0	3	normal
x13	Differing RECs	number	0	5	linear decreasing

2.3.2.1 x01 Pixel Size

Raster cells (pixels) of datasets throughout this analysis have a consistent size of 15”x15”. However, the area one pixel covers in terms of square meters decreases with increasing distance from the equator. A pixel at the equator is 462.96m by 462.96m and covers an area of 214’335m2. For other raster cells, this value has to be multiplied by the cosine of their latitude.

➔ It is expected to see a positive correlation between the size of a pixel and the occurrence of CBI, as larger pixels have more space to contain a CBI.

2.3.2.2 x02 Border length

Borders are linear features by nature. The analysis, however, was carried out on a raster basis. While the rasterisation of linear features in a GIS poses no problems, it’s clear that some raster cells contain longer segments than others.

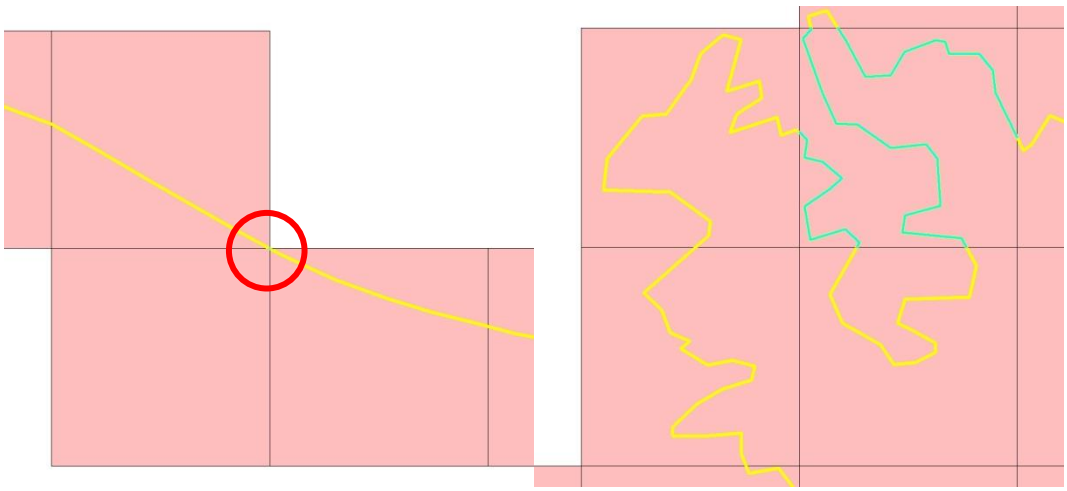


Fig. 14 Shortest (left) and longest (right) border line segments in any raster cell

➔ It is expected to see a positive correlation between the length of a line segment and the occurrence of a CBI, as a shorter line segment simply means less options to build a CBI.

While x02 may seem to make x01 redundant, they don’t respond to the same aspect of border length on the ground, as a long segment in a pixel is often due to a meandering river, which doesn’t necessarily transform into more CBI.

2.3.2.3 x03 Distance from coast

The main input for this variable is the coast line as provided in the LSIB dataset. For each border pixel, the geodesic distance to the closest point on the coast line was calculated and stored in a new raster dataset.

➔ It is expected to see a slightly positive correlation between the distance from the coast and the occurrence of a CBI, as goods may be easily transported by sea between coastal states.

2.3.2.4 x04 Distance to nearest capital

Many border areas are both far from the respective capital and underdeveloped, i.e. don't have sufficient infrastructure such as schools, health centres, or roads. In various publications (e.g. African Union, 2017), this correlation (lack of infrastructure vs. distance to the country's capital) is pointed out. As part of this thesis, this assumed correlation shall be examined. Since geographical proximity to one of the two neighbouring capitals should be sufficient incentive for investment, the variable measures the geodesic distance to the nearest capital only and disregards the capital that is further away.

➔ It is expected to see a negative correlation between the distance to the nearest capital and the occurrence of CBI, as remote areas receive less attention from central (and centralised) governments.

2.3.2.5 x05 Elevation

Many African borders follow the watershed line along mountain chains. To analyse the impact of different elevations, pixels were assigned the respective value taken from the HydroSHEDS digital elevation model dataset. The elevation is measured in meters above mean sea level.

➔ It is expected to see a negative correlation between the elevation and the occurrence of a CBI, as higher altitudes usually mean more remote areas and higher costs for infrastructure.

2.3.2.6 x06 Slope

A sister indicator to Elevation is Slope. Not only high altitudes, but also hilly terrain might be an impediment to CBI investment.

➔ It is expected to see a negative correlation between the slope and the occurrence of CBI, as steeper slopes usually result in higher costs for infrastructure.

2.3.2.7 x07 River size

The idea behind this variable is relatively simple. It makes a difference in terms of CBI investment, whether a boundary runs along dry land, a small rivulet or a majestic river. Three datasets were involved: 1) the LSIB, which is the reference geometry, but has no information whether a border segment is a river, terrestrial or lake boundary. 2) the International River Boundary Database (Donaldson, 2007), which shows most of the river boundary segments worldwide. It was researched based on boundary treaties and digitised on Google Earth wherever possible. 3) the HydroSHEDS (Lehner et al., 2008) dataset that shows the entire river network around the globe. As it is derived from a DTM, its horizontal accuracy is limited in flat areas, and differences between the three datasets can be large.

Fig. 15 Burundi-Rwanda, overlay of LSIB (yellow), HydroSHEDS (green), IRBD (blue)

The variable was created in two main steps. First (1), all river segments in the LSIB were identified using the HydroSHEDS and IRBD datasets. Then (2), these river segments were assigned the upstream cell value of the HydroSHEDS dataset. The whole process was complicated by the above-mentioned different geometries.

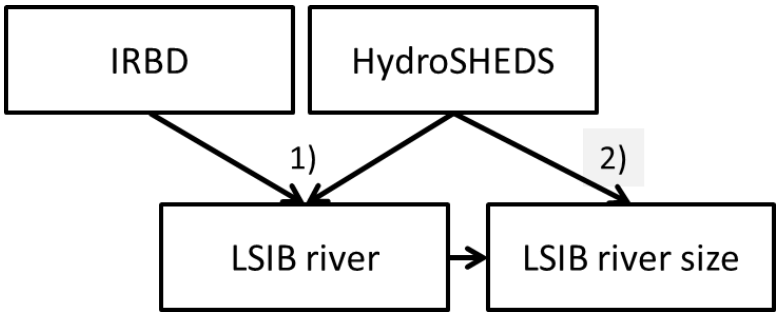


Fig. 16 Workflow to create the river size variable

The HydroSHEDS river network is derived from its hydrographically correct raster DTM. Each line points downstream and carries the number of upstream cells (upcell) from its lowest point as an attribute (labelled below).

Fig. 17 HydroSHEDS river segments showing the number of contributing cells

Conveying the up-cell attribute from HydroSHEDS onto the LSIB geometries was one of the most delicate tasks in this thesis and required numerous individual steps, such as removing HydroSHEDS segments that don't intersect with the LSIB or are more than 2km away, removing contributing streams (segments where the starting point is far

away from the LSIB) etc. While these routines reduced the workload, many checks needed to be done semi-manually to correct unforeseen cases. Once a clean dataset of LSIB and HydroSHEDS geometries was created, the latter was used to dynamically segment the former, then a spatial join of the centre points transferred the upcell attribute from one to the other. To account for the fact that pixel size decreases north and south of the equator, the upcell value was multiplied by the cosine of its latitude. The final dataset has pixel values ranging from 21 to 3.7mio square kilometres.

Fig. 18 LSIB segments with the "upcell" value inherited from HydroSHEDS lines

➔ It is expected to see a negative correlation between river size and the occurrence of CBI, as larger rivers require larger bridges, which in turn result in higher costs. Larger rivers may also be more profitable for hydropower plants, but it is expected to find only relatively few of them on the African continent.

On 28 June 2018, a new global dataset on river width was published (Allen & Pavelsky, 2018). Unfortunately, this dataset couldn't be incorporated into the present thesis, but might be valuable for use in future research.

2.3.2.8 x08 Population Density

It would not be surprising to find less CBI in the uninhabited Sahara than on the densely populated West African coast line. To analyse this, a dataset on population density (FAO, 2015) was used. The resolution of the dataset is ten times lower (2.5') than the one used in this analysis and therefore had to be resampled. Pixel values are in a relative format, persons per square kilometres, so no further readjustments were required.

There are other datasets on population density available, some of them with a very high resolution, i.e. a 7.5" ArcGIS online layer called World Population Estimated Density (esri, 2015). However, the lower resolution in this case is preferred, as it doesn't only

represent the area covered by the border pixel, but also a broader area around the border. Values range from 0 to 11’657 persons per square kilometre for border pixels, with the highest values being at the cd-cg border at the twin-capitals Brazzaville-Kinshasa.

➔ It is expected to see a positive correlation between population density and the occurrence of CBI, as higher population density demands for more infrastructure investments.

2.3.2.9 x09 Fraternisation

Due to the genesis of African borders, many of them cut through regions inhabited by one and the same people. It can be assumed that these, although being of different nationality, continue to migrate back and forth either side of the border. Sadly, there are also a number of conflicts that force people to seek refuge in neighbouring countries.
The underlying assumption of this variable is that a higher number of nationals from Country A living in Country B, and vice-versa, could be a sign of strong bonds between neighbouring countries, and that this could materialise in a higher number of CBI. For every border, the value used in the analysis is the sum of A-ians living in B plus B-ians living in Country A. Some missing values in the initial dataset (IOM, 2015) were researched and updated individually.

➔ It is expected to see a positive correlation between Fraternisation and the occurrence of CBI, as higher Fraternisation indicates stronger ties between neighbouring countries.

2.3.2.10 x10 GDP per Capita

Economic power varies significantly between African nations. To take the size of an economy into account, GDP per Capita was used. A small country may have a smaller GDP compared to the one of a larger country, but still have wealthier citizens than its bigger neighbour. Even then, GDP per Capita ranges from 286usd (Burundi) to 8747usd (Equatorial Guinea), a factor of 30. GDP per Capita doesn't take wealth distribution within the population into account. However, a high GDP per Capita would indicate that a country has sufficient economic resources to provide for its citizens, be it health facilities, education or CBI.

The main data source for this variable was a statistic from the World Bank (2017). For most countries, the most recent data are for 2016. Where these were not available, the most recent available figures were used. GDP per Capita of neighbouring countries were averaged, as both sides need to invest in CBI.

➔ It is expected to see a positive correlation between GDP per Capita and the occurrence of CBI, as higher GDP per Capita means that a country has more resources to invest in CBI.

2.3.2.11 x11 Trade

While the economy of certain countries is rather self-sufficient, other countries depend heavily on imports and exports. It would only be logical if the more "extrovert" countries invested more in CBI than "introvert" ones. Using official statistics (World Bank, 2018; WTO, 2017), the Trade to GDP ratio was extracted for every African country, except Western Sahara, where no such data was available. The average of two neighbouring countries was used as a value for any specific border.

Using the aforementioned data brings several problems with regards to correlating them to CBI. First, coastal states may handle most of their trade through their ports, which are not included in this analysis. Also, high value minerals may need fewer CBI than e.g. wood being transported through a neighbouring country and on to the coast, so a relatively small share of the trade may require large CBI investments, and vice versa. Thirdly, goods that are transported from Country A to Country D, passing through Country B and C on the way, only appear in the import and export statistics of

Countries A and D. They also require infrastructure at the B-C border, but there is no "transmission" statistics available, only import and export in the countries of production and consumption. Despite all these challenges, the data is the best available and is used in the present analysis.

➔ It is expected to see a positive correlation between trade and the occurrence of CBI, as countries with higher trade need better and more CBI to ensure smooth import and export.

2.3.2.12 x12 Shared RECs memberships

There are numerous regional bodies on the African continent that aim to coordinate and defend the interests of their member states. Eight so-called Regional Economic Communities (RECs) are recognised by the African Union (AUC, 2018). These are far from mutually exclusive. African countries are members in 1.96 of these RECs on average. The below maps show the Arab Maghreb Union (UMA), Common Market for Eastern and Southern Africa (COMESA), Community of Sahel-Saharan States (CEN-SAD), East African Community (EAC), Economic Community of Central African States (ECCAS), Economic Community of West African States (ECOWAS), Intergovernmental Authority on Development (IGAD) and the Southern African Development Community (SADC) from top left to bottom right.

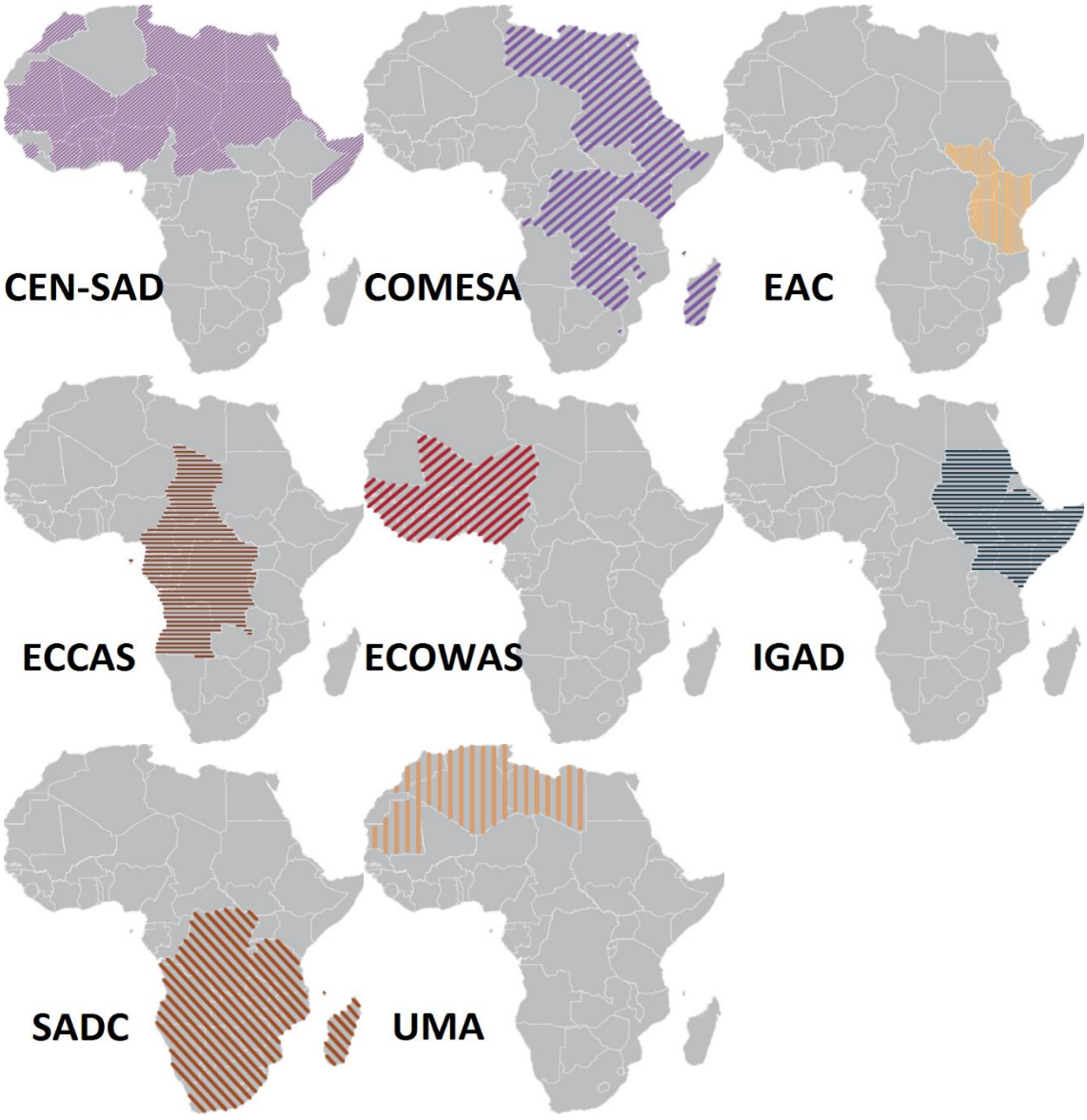


Fig. 19 Geographical extent of the eight RECs recognised by the AU (ecdpm, 2018)

Since RECs aim to strengthen economic ties between their member states, provide joint regulatory frameworks as well as joint investments, this could result in more CBI on their internal borders. Two neighbours that are part of the same REC could share more CBI than others.

➔ It is expected to see a positive correlation between the number of shared RECs memberships and the occurrence of CBI, as common membership of two neighbours in the same REC, or RECs, could mean closer economic ties.

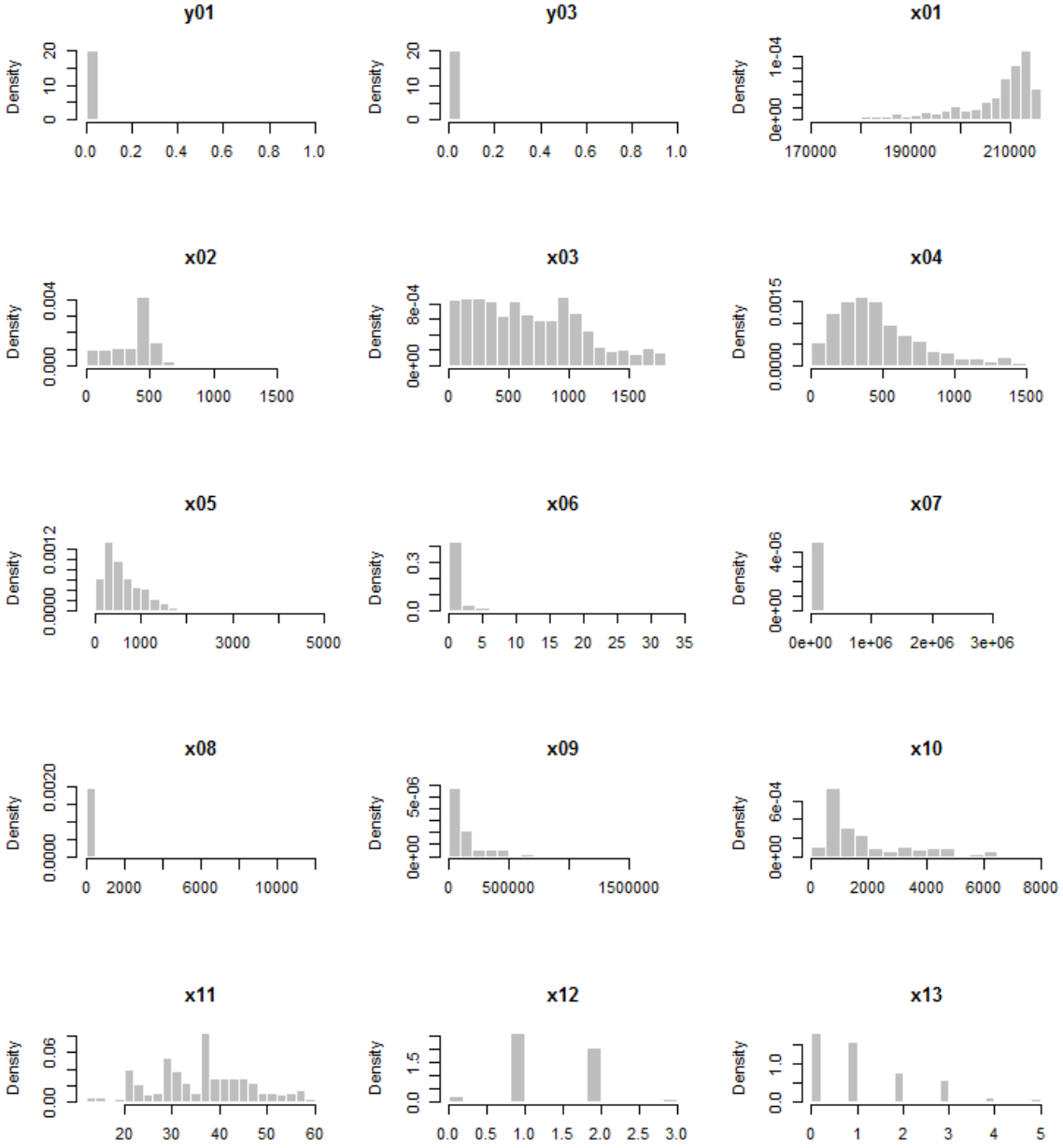
2.3.2.13 x13 Differing RECs memberships

As much as membership in the same RECs could mean that two countries share closer ties, being members of different RECs could mean that they have fewer interests in common. This could result in fewer CBI.

➔ It is expected to see a negative correlation between the number of non-shared memberships of RECs and the occurrence of CBI, as memberships of two neighbours in different RECs could mean that they have fewer common interests.

Finally, all the variables were rasterised, then converted into a tabular form and exported to the statistics software. Below is an overview of their histograms.

Table 5 Histograms of dependent and independent variables



2.4 Analysis

2.4.1 Different datasets for analysis

The analysis was carried out on eight different datasets, named after their dependent variables:

y03	Large CBI on the whole continent
y01	all intersections of boundaries with OSM roads on the whole continent
y01n	all intersections of boundaries with OSM roads within northern Africa
y01w	all intersections of boundaries with OSM roads within western Africa
y01c	all intersections of boundaries with OSM roads within central Africa
y01e	all intersections of boundaries with OSM roads within eastern Africa
y01s	all intersections of boundaries with OSM roads within southern Africa
50perc	same as y01, but with a limited sample of non-events pixels (50/50)

Any set of graphics hereafter will show these datasets in the order listed above, from top left to bottom right. The different geographical areas refer to the five regions as recognised by the African Union (also see bottom right on the illustration below).

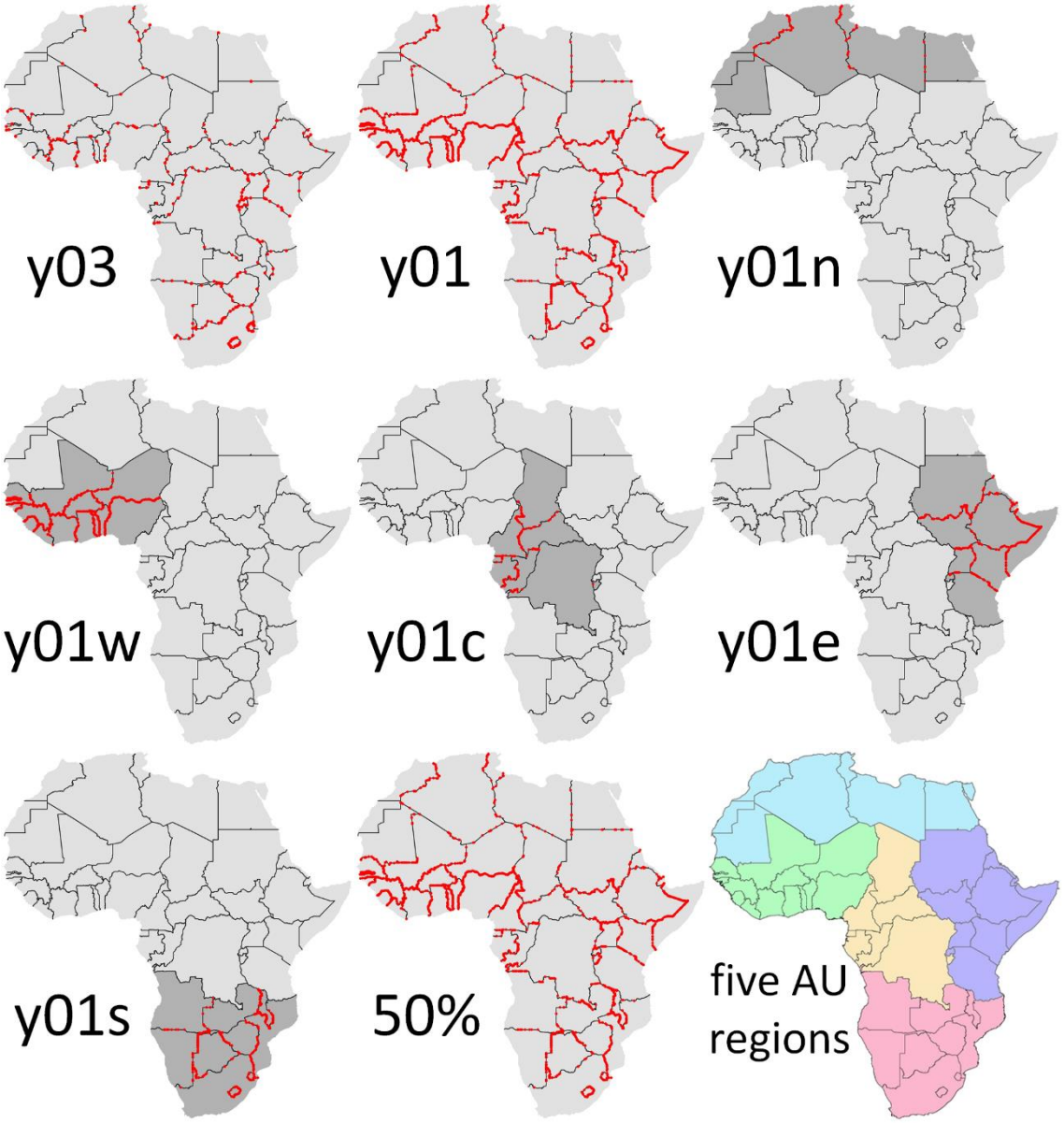


Fig. 20 CBI (red) in the different datasets to be analysed

2.4.2 Rare events, rule of 10

A first analysis showed that all datasets have so called "rare events". While there is no exact definition of the term, it often refers to datasets where events make up less than 15% of the total.

Table 6 Number of events vs. non-events per dataset

	y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
events	298	3328	62	1224	176	467	736	3328
non-events	227129	224099	16634	43909	30127	32744	40074	3328
total	227427	227427	16696	45133	30303	33211	40810	6656
events %	0.1%	1.5%	0.4%	2.7%	0.6%	1.4%	1.8%	50.0%
events type	rare	rare	rare	rare	rare	rare	rare	even
rule of 10	pass	pass	pass *	pass	pass	pass	pass	pass

King & Zeng (2001) have researched and published extensively about this particular challenge in regression analysis. The main problem is that when there are only few events, the model will underestimate the likelihood for an event to happen. They propose two ways to resolve the matter. The first is to use some advance mathematical models to (seemingly) increase the number of events, which is what e.g. Theofilatos et al. (2016) have done in their research on road accidents. The second option is to reduce the number of non-events by sampling. While this may look like manipulating the data, it is a relatively common and legitimate strategy. For example, a researcher looking for people that are prone to developing skin cancer may work with a dataset of people that have been tested for skin cancer. Those are people that a) have symptoms that make them undergo the test, b) live in area where an equipped hospital is available and c) can afford the testing. Skin cancer, statistically speaking, is a rare event in the whole population. But it may be rather common in the dataset that is used for its analysis. Also, it would not make sense economically to test the entire population and then dismiss most of the data in a second step.

Underestimating the likelihood of an event can be fatal if we're trying to predict skin cancer prevalence. For CBI, the issue isn't such a problem, as our main interest is probably not the likelihood of a specific pixel to contain CBI, but rather to see areas that are more likely than others, or to compare the sum of all pixels on one border to that of another.

For the scope of this research, a set of non-events was selected to have a 50/50 event vs. non-event dataset. Unlike the skin cancer example above, there are no pixels that would be clearly more prevalent to having CBI than others, which is why a random sample was selected.

On a simpler note, all the datasets fulfil the so-called "rule of ten", an indicative rule that demands for each independent variable to have at least ten events. This is the case for all datasets, although for northern Africa, it's only after the elimination of some independent variables (see following chapters).

2.4.3 Correlation matrix

In a next step, the correlation matrix for each dataset was analysed. Especially the variables in northern and central Africa are correlated. However, many of those are removed in the following steps which mitigates the problem. As the west African region shares the same number of shared (2) and differing (0) RECs, the variables are constant, making them unfit for the analysis. For a more detailed analysis of these correlations, please see chapter4.1.

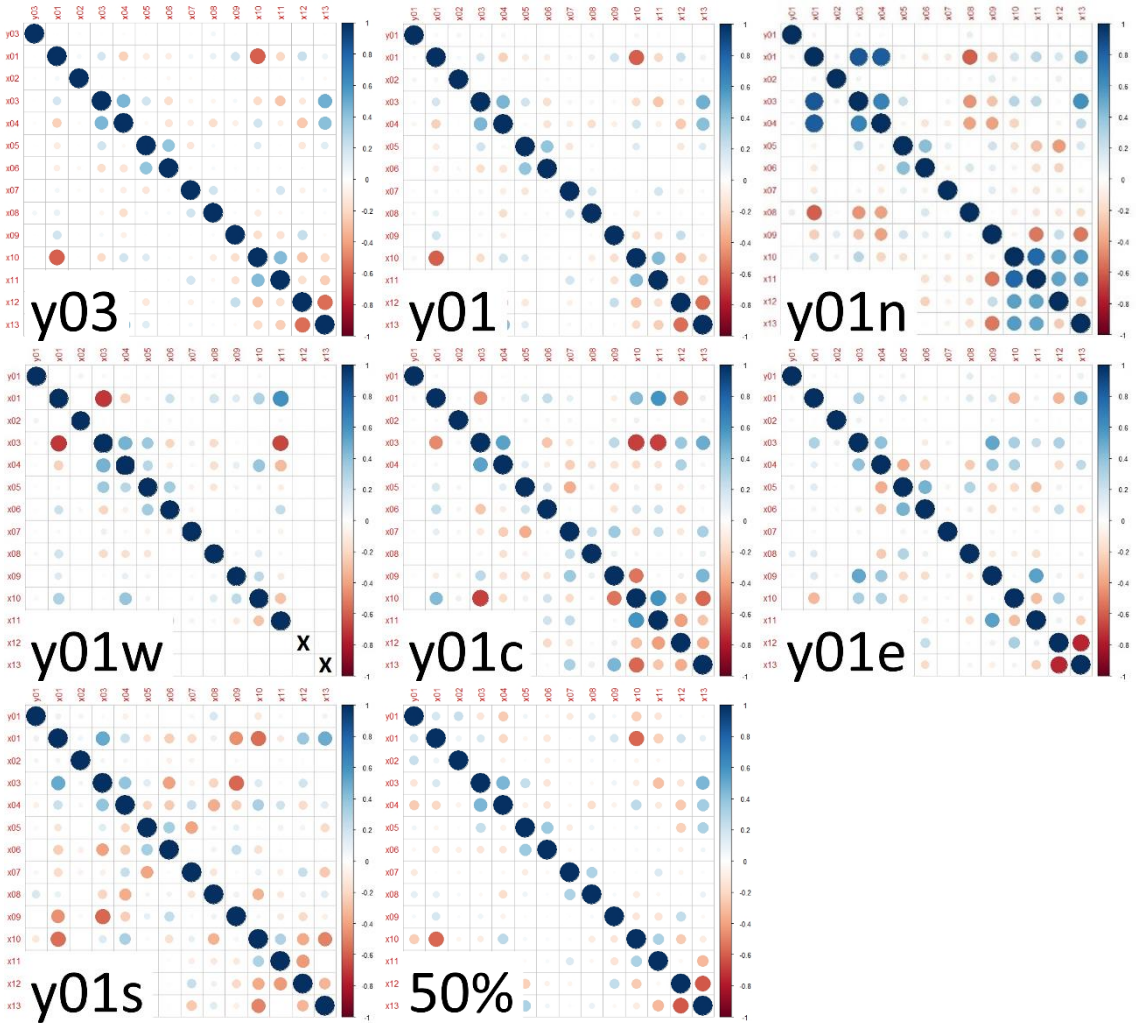


Fig. 21 Correlation matrices for all datasets

2.4.4 Variance inflation factor (VIF)

VIF is a measure of multicollinearity of variables in a given dataset. As a rule of thumb (Sheather, 2009), a VIF of >5 is considered high, >10 as too high.

Table 7 VIF including all variables

	y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
x01	1.8	1.6	47.5	3.5	5.6	3.4	8.0	1.5
x02	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
x03	1.6	1.5	18.0	3.8	6.2	2.9	4.4	1.8
x04	1.4	1.5	23.9	2.1	2.9	2.5	1.6	1.7
x05	1.6	1.6	2.4	2.4	2.0	2.6	1.9	1.6
x06	1.3	1.3	1.3	1.4	1.3	1.5	1.5	1.3
x07	3.4	1.2	1.2	1.0	1.9	1.2	1.1	1.0
x08	3.2	1.3	3.3	1.1	1.8	1.3	1.2	1.2
x09	1.1	1.0	9.1	1.1	3.2	6.6	3.7	1.1
x10	2.4	1.6	19.1	2.8	5.0	6.2	7.4	1.7
x11	1.7	1.4	35.7	2.9	3.3	5.3	2.3	1.4
x12	2.0	2.4	24.2	constant	2.5	7.4	9.1	2.5
x13	2.3	3.1	16.2	constant	4.9	11.9	9.5	3.0

Removing the variable with the highest VIF in each of the datasets, the below values were achieved. Removing any of the highest VIF in the y01c dataset didn't result in an improvement of the remaining variables and were therefore left as is.

Table 8 VIF after removal of VIF > 10

	y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
x01	1.8	1.6	deleted	3.5	5.6	2.1	2.1	1.5
x02	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
x03	1.6	1.5	4.4	3.8	6.2	2.3	2.3	1.8
x04	1.4	1.5	1.9	2.1	2.9	2.5	2.5	1.7
x05	1.6	1.6	1.5	2.4	2.0	2.6	2.6	1.6
x06	1.3	1.3	1.3	1.4	1.3	1.6	1.6	1.3
x07	3.4	1.2	1.1	1.0	1.9	1.2	1.2	1.0
x08	3.2	1.3	2.2	1.1	1.8	1.3	1.3	1.2
x09	1.1	1.0	1.6	1.1	3.2	4.6	4.6	1.1
x10	2.4	1.6	deleted	2.8	5.0	2.0	2.0	1.7
x11	1.7	1.4	deleted	2.9	3.3	3.5	3.5	1.4
x12	2.0	2.4	1.4	constant	2.5	2.4	2.4	2.5
x13	2.3	3.1	2.2	constant	4.9	deleted	deleted	3.0

2.4.5 Stepwise regression

A large number of independent variables inevitably poses the question, whether all of them actually contribute to the quality of the regression model. The goal of a stepwise approach is to remove the ones that are not needed. Three procedures are in use: forward selection (adding the most relevant variables one by one), backward elimination (removing the most irrelevant variable one by one), or a combination of both. The below table shows the variables removed in each dataset.

Table 9 Variables removed after stepwise regression

	y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
x01			VIF			step	step	
x02								
x03	step							
x04			step	step	step			
x05			step					
x06								
x07	step		step					
x08								
x09		step	step		step			
x10			VIF		step			step
x11			VIF			step		
x12	step			constant				
x13			step	constant	step	VIF	VIF	

2.4.6 Prediction of CBI

A core goal of logistic regression is oftentimes to predict future events, e.g. what is the probability of a person to develop a certain medical condition *in the future* based on his/her age, sex, weight, height, smoking and drinking habits of today. This is to both take prophylactic measures as well as to identify certain risk groups that should undergo further (expensive) tests, rather than testing the whole population.

The idea of "prediction" in this research is not so much to look into the future, but rather to see whether data could be collected in one area only, to save time and costs. This requires the model to be applicable also to the rest of the continent. For the scope of this thesis, it was agreed to predict CBI in southern Africa using the coefficients from eastern Africa and evaluate the results. See chapter 0 for further detail.

3. Results

This chapter presents the results of the different logistic models, namely the coefficients (3.1), Pseudo R2 (3.2), and Residual Deviances (3.3). Finally, it attempts to visualise the results (3.4) before using data from eastern Africa to predict southern African CBI (3.5).

3.1 Coefficients

The different logistic regression models yielded the following estimated coefficients (Est), Standard Errors (SE), z- and p-values.

Table 10 Overview of coefficients of the logistic regression

		y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
Int.	Est	-2.6E+00	-3.4E+00	-4.5E+00	5.0E+01	1.9E+02	-4.6E+00	5.5E+00	-7.3E-01
	SE	1.5E+00	6.9E-01	7.7E-01	5.9E+00	2.2E+01	2.9E-01	8.8E-01	9.8E-01
	z	-1.689	-4.853	-5.939	8.544	8.551	-15.966	6.223	-0.742
	p	0.091206	1.22E-06	2.87E-09	< 2e-16	< 2e-16	< 2e-16	4.87E-10	0.45782
x01	Est	-2.0E-05	7.3E-06		-2.6E-04	-8.8E-04			1.4E-05
	SE	7.1E-06	3.2E-06		2.9E-05	1.0E-04			4.6E-06
	z	-2.763	2.264		-8.987	-8.593			2.996
	p	0.005736	0.0236		< 2e-16	< 2e-16			0.00273
x02	Est	2.3E-03	2.2E-03	2.3E-03	2.5E-03	1.9E-03	2.3E-03	2.6E-03	2.9E-03
	SE	3.3E-04	1.0E-04	7.5E-04	1.8E-04	4.3E-04	3.0E-04	2.3E-04	1.9E-04
	z	6.957	22.216	3.013	14.068	4.416	7.728	11.578	15.15
	p	3.49E-12	< 2e-16	2.59E-03	< 2e-16	1.00E-05	1.09E-14	< 2e-16	< 2e-16
x03	Est		-7.4E-04	-3.1E-03	-2.5E-03	-3.0E-03	7.1E-04	-1.5E-03	-7.8E-04
	SE		5.4E-05	9.9E-04	1.7E-04	4.6E-04	1.6E-04	2.6E-04	9.3E-05
	z		-13.538	-3.098	-14.152	-6.567	4.502	-5.588	-8.362
	p		< 2e-16	0.00195	< 2e-16	5.12E-11	6.74E-06	2.29E-08	< 2e-16
x04	Est	-2.6E-03	-1.3E-03				-2.4E-03	-2.1E-03	-1.1E-03
	SE	2.9E-04	9.9E-05				3.4E-04	2.9E-04	1.5E-04
	z	-8.936	-13.335				-6.925	-7.365	-7.547
	p	< 2e-16	< 2e-16				4.37E-12	1.78E-13	4.45E-14
x05	Est	3.2E-04	5.4E-04		1.6E-03	1.6E-03	9.1E-04	1.0E-03	4.5E-04
	SE	1.2E-04	4.1E-05		2.7E-04	4.1E-04	1.2E-04	9.8E-05	7.2E-05
	z	2.538	13.036		6.15	3.799	7.926	10.415	6.301
	p	0.011135	< 2e-16		7.76E-10	0.000145	2.27E-15	< 2e-16	2.96E-10
x06	Est	-1.3E-01	-1.5E-01	-1.5E-01	-1.7E-01	-1.8E-01	-1.9E-01	-1.2E-01	-1.4E-01
	SE	3.6E-02	1.2E-02	7.1E-02	3.7E-02	9.9E-02	3.0E-02	2.0E-02	1.8E-02
	z	-3.487	-12.298	-2.101	-4.512	-1.813	-6.22	-6.378	-7.589
	p	0.000488	< 2e-16	0.03565	6.42E-06	0.06982	4.97E-10	1.79E-10	3.22E-14

x07	Est		-4.9E-06		-7.5E-07	-2.7E-06	-6.3E-05	-1.4E-06	-4.3E-06
	SE		4.6E-07		5.1E-07	7.8E-07	9.8E-06	7.3E-07	5.0E-07
	z		-10.811		-1.493	-3.437	-6.464	-1.913	-8.558
	p		< 2e-16		0.135363	0.000587	1.02E-10	0.0557	< 2e-16
x08	Est	5.3E-04	1.0E-03	9.0E-03	8.3E-04	8.1E-04	2.2E-03	3.6E-03	2.4E-03
	SE	5.2E-05	6.3E-05	1.9E-03	1.3E-04	2.1E-04	2.0E-04	3.7E-04	3.4E-04
	z	10.024	16.369	4.876	6.507	3.904	11.214	9.668	7.131
	p	< 2e-16	< 2e-16	1.08E-06	7.64E-11	9.45E-05	< 2e-16	< 2e-16	9.99E-13
x09	Est	6.4E-07			-1.8E-07		6.7E-07	-4.9E-06	
	SE	2.2E-07			1.1E-07		2.7E-07	6.7E-07	
	z	2.87			-1.603		2.493	-7.27	
	p	0.004099			0.109018		0.0127	3.60E-13	
x10	Est	2.1E-04	-3.9E-04		1.0E-03		1.0E-03	-5.2E-04	-3.2E-04
	SE	4.4E-05	2.1E-05		1.4E-04		1.6E-04	3.0E-05	2.8E-05
	z	4.745	-18.367		7.207		6.302	-17.393	-11.439
	p	2.08E-06	< 2e-16		5.73E-13		2.95E-10	< 2e-16	< 2e-16
x11	Est	-1.8E-02	-2.2E-02		-2.0E-02	-3.4E-02		-1.5E-01	-3.1E-02
	SE	7.0E-03	2.2E-03		6.1E-03	9.4E-03		1.7E-02	3.4E-03
	z	-2.625	-10.063		-3.326	-3.596		-8.782	-8.986
	p	0.00867	< 2e-16		0.000881	0.000324		< 2e-16	< 2e-16
x12	Est		-6.0E-01	-1.0E+00		-1.0E+00	-1.1E+00	-1.9E+00	-5.9E-01
	SE		4.4E-02	4.6E-01		3.7E-01	9.2E-02	1.6E-01	6.6E-02
	z		-13.654	-2.23		-2.682	-12.092	-11.769	-8.927
	p		< 2e-16	0.02574		0.007327	< 2e-16	< 2e-16	< 2e-16
x13	Est	9.3E-02	-2.2E-01						-2.3E-01
	SE	6.1E-02	2.6E-02						4.0E-02
	z	1.528	-8.197						-5.897
	p	0.126432	2.46E-16						3.71E-09

Red script for estimated coefficients indicates that its impact is contrary to what was anticipated, while green colour indicates an expected outcome. Orange script highlights p-values above 5%.

- Border length (x02), distance to nearest capital (x04), slope (x06), river size (x07) and population density (x08) showed the anticipated negative or positive correlation.
- Cell size (x01), distance from coast (x03), fraternisation (x09), GDP per capita (x10) and differing RECs memberships (x13) showed mixed results.
- Elevation (x05), Trade (x11) and shared RECs memberships (x12) consistently have an impact other than the one expected.

3.2 Pseudo R2

In linear regression, the coefficient of determination, called R2, is an indicator of how good a model will predict an outcome. Or, to be more precise, it is the proportion (%) of the variance of the dependent variable that is explained by the independent variables.

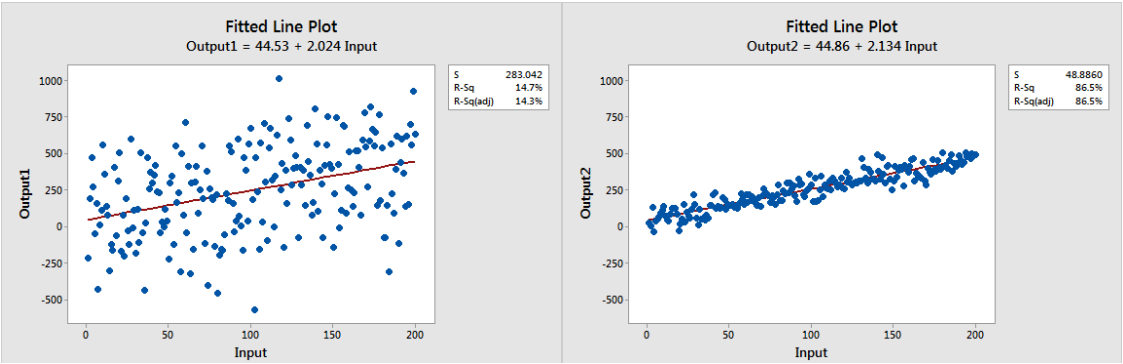


Fig. 22 Illustration of low (left) vs. high (right) R2 in linear regression (minitab, 2014)

This concept isn't directly applicable to logistic regressions, which is why Pseudo-R2 indicators have been developed. One that is used frequently is the "McFadden R2".

Table 11 Pseudo R2 values of the different models

	Y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
McFadden	0.06	0.09	0.12	0.05	0.12	0.13	0.21	0.17

3.3 Residual deviance, AIC

Several indicators show the quality of a model. While the Null deviance takes no variables into account and uses a model based on a constant only, the residual deviance calculates the model including the respective variables. This reduces the degree of freedom (lower figure), but this isn't a problem in our case as we have plenty of observations. The larger the difference between the Null and the Residual deviance, the better the model explains the observations. The Akaike Information Criterion (AIC) estimates the quality of the model and is used to compare different models, i.e. during stepwise regression. Smaller AICs indicate better models.

Table 12 Null deviance, residual deviance and Akaike Information Criterion

	y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
Null deviance	4552 / 227426	34725 / 227426	818 / 16695	11246 / 45132	2163 / 30302	4910 / 33210	7369 / 40809	9227 / 6655
Residual deviance	4287 / 227416	31683 / 227414	718 / 16690	10699 / 45122	1895 / 30293	4267 / 33200	5839 / 40798	7639 / 6643
AIC	4309	31709	730	10721	1915	4290	5863	7665

In a balanced model, the deviance residuals should usually be roughly normally distributed around zero. In the below table, this clearly isn't the case, which is due to the fact that CBI are rare events. In the 50perc dataset where the number of events and non-events was artificially balanced, the deviance residuals show the expected distribution.

Table 13 Distribution of deviance residuals

	y03	y01	y01n	y01w	y01c	y01e	y01s	50perc
Min	-1.1363	-1.5462	-0.6195	-1.1635	-0.5126	-1.7774	-1.4060	-4.9681
1Q	-0.0552	-0.2013	-0.0850	-0.2596	-0.1202	-0.1761	-0.2020	-0.9618
Median	-0.0435	-0.1376	-0.0622	-0.2056	-0.0693	-0.1270	-0.0905	0.0078
3Q	-0.0329	-0.0892	-0.0439	-0.1663	-0.0304	-0.0835	-0.0545	0.9439
Max	4.2754	5.9212	3.9980	3.1809	4.8786	4.2994	4.0892	3.3018

3.4 Visualisation

The result of any logistic regression is the probability for any set of variables to result in an event (or a non-event). Depicting the probability of more than 200'000 points is difficult, showing any further analysis is close to impossible. The below overviews show probabilities for every point, with high probabilities overlaying smaller one, i.e. a green point will always overlay a red one.

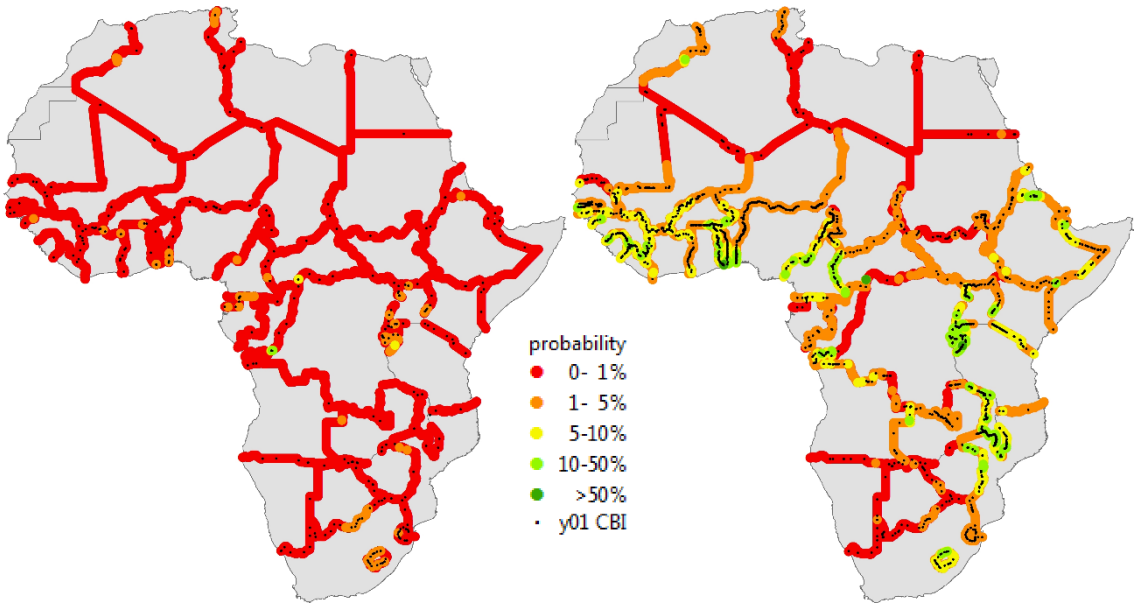


Fig. 23 Predicted probabilities by y03 (left) and y01 (right) models

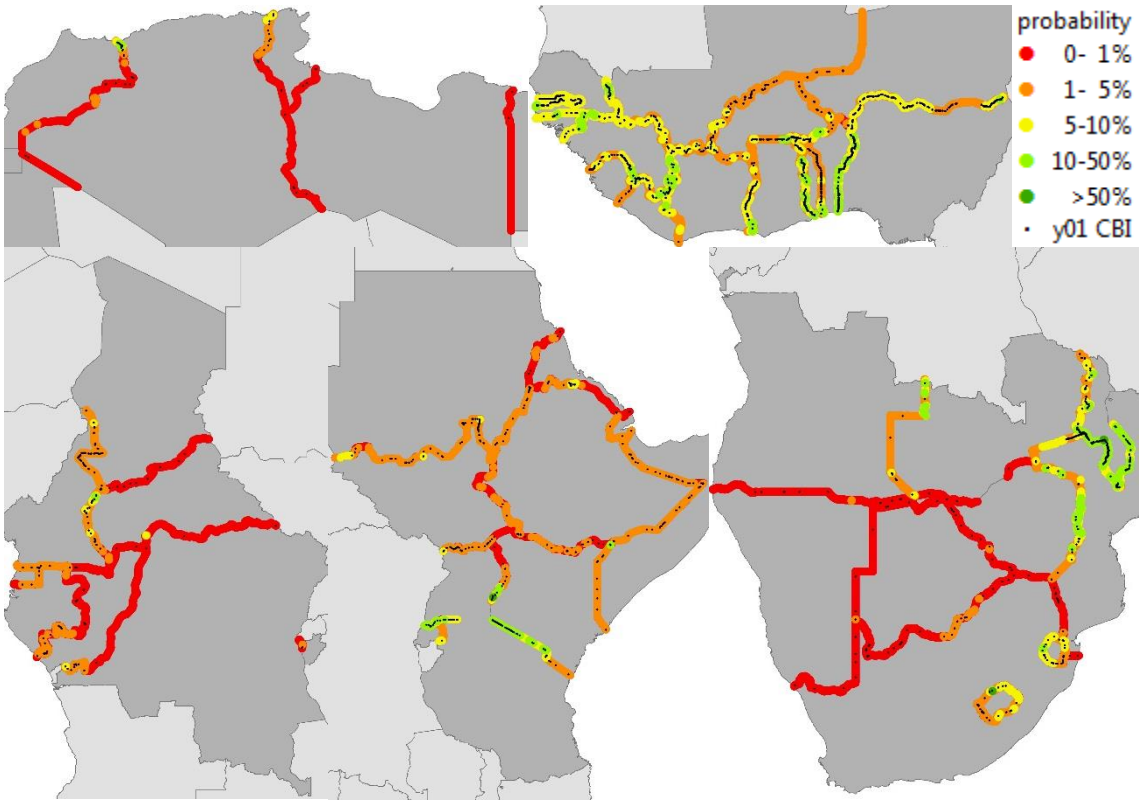


Fig. 24 Probabilities as predicted by regional models, higher probabilities drawn on top

To give more background information, the size of the pie chart in the figure below represents the probability of an event, while the different "slices" show the contribution by each variable to this result. For example, abs02 (dark green) is the length of the border, and it's nearly absent in the smaller probabilities. The "abs" is short for "absolute value", as some variables are negative, but the magnitude of their impact on the result is of course absolute. Such a presentation is only useful in a digital format, e.g. a webmap.

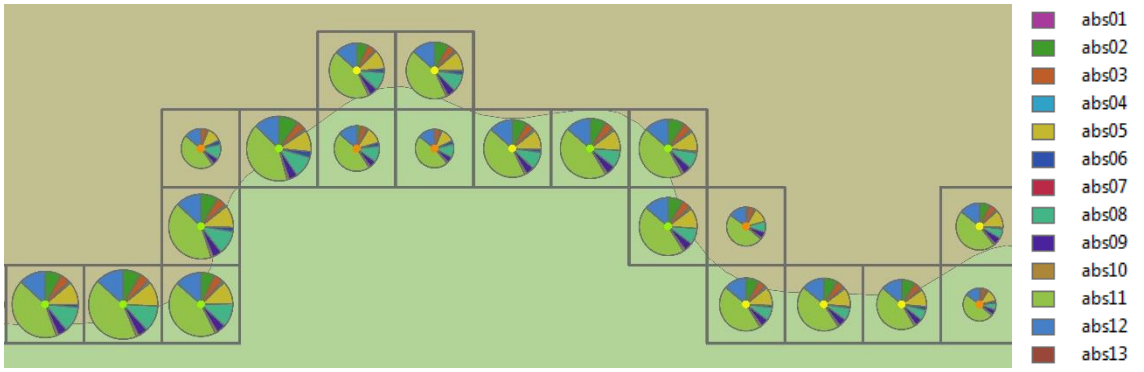


Fig. 25 Probability as pie chart, size = probability, contribution by variables

While logistic regression usually tries to predict the probability from an event to happen given a specific set of independent characters (e.g. whether a specific patient will develop cancer or not), these probabilities can also be aggregated. If, say, 2'000 patients have a 2% chance of developing cancer, the expected outcome is to see 40 cases. This information is of value to pharmaceutical companies when developing new drugs against cancer.

In the same way, if we are interested in the number of CBI at any given border, we can sum up the probabilities of all points (or raster cells). These can then be compared to the actual CBI counted on this border, thereby analysing if a border has more or fewer CBI than what was predicted by the logistic regression. Borders with more CBI than predicted are in the top-left, the one with fewer CBI in the bottom-right corner in the figures below.

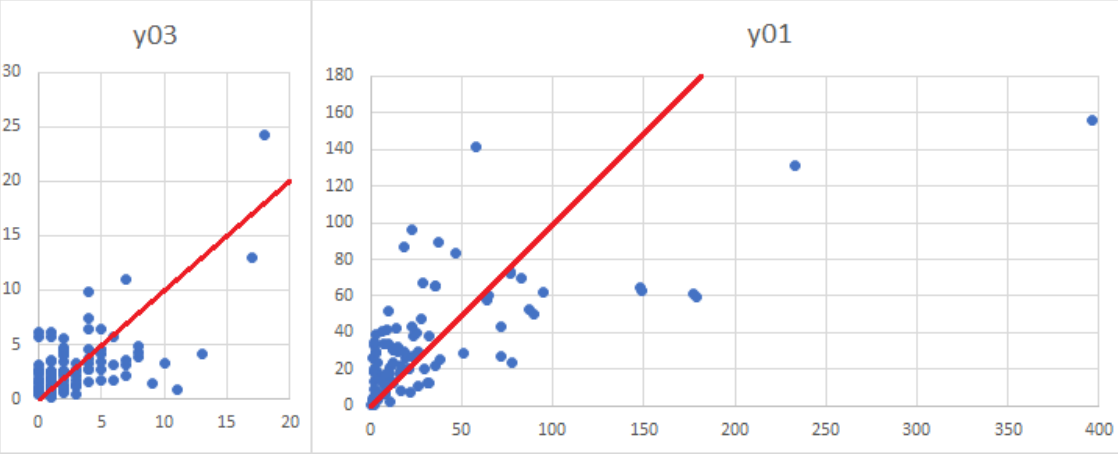


Fig. 26 Comparison of actual CBI (horizontal) vs. sum of prediction (vertical)

3.5 Prediction of CBI in southern Africa

Using the coefficients calculated for eastern Africa, an attempt at predicting CBI in southern Africa was made. The result is a significantly higher number of CBI estimated.

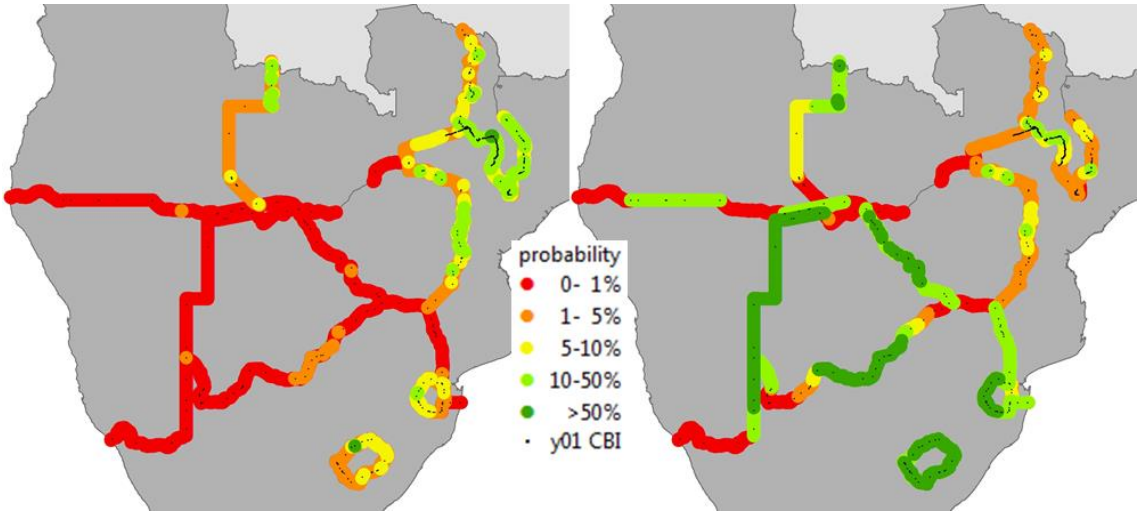


Fig. 27 Predicted CBI (y01) using southern (left) and eastern (right) African coefficients

4. Discussion

The main goal of the present thesis was to analyse correlations between CBI and a variety of variables on the African continent. To this end, a set of 13 independent variables were created: geometrical, topographical, sociological, economical and political ones. Two dependent variables were analysed: large CBI (such as official border posts, hydropower dams or railways) as well as paved roads from OSM.
This chapter discusses the different variables (4.1), Pseudo-R2 (4.2), compares continental and regional models (4.3) and lists challenges and limitations encountered (4.4)

4.1 Expected, ambiguous and unexpected coefficients

As already stated under 3.1, *Border length (x02)*, *distance to nearest capital (x04)*, *slope (x06)*, *river size (x07)* and *population density (x08)* showed the anticipated negative or positive correlation to CBI. This means, CBI occur in areas of high population density, close to capitals, in flat terrain and on dry land or across small rivulets and where border segments within the analysed raster cell have a certain length.

Cell size (x01) and *GDP per capita (x10)* both show mixed results but were expected to have a positive correlation with CBI. On the continental level, there's a negative correlation between the two. Or, to put it differently: the closer to the equator, the smaller the economic power. While the reason for this is subject to debates between economists, anthropologists and many other professions, it is sufficient to state here that while both were expected to be positive, the two are in interplay: if one is positive, the other is negative.

A similar overlaying effect seems to be at work for the *distance from coast (x03)* variable, which shows a mostly negative correlation, although it was expected that there would be few CBI at the coast (where goods can be easily transported by ship or boat) and more CBI inlands. *Population density (x08)* also has a positive impact, but there's a slight negative correlation between the two. The fact that inland areas (such as the Sahara Desert) are less inhabited than coastal zones has a larger effect than the anticipated easier transport by sea.

In the eight logistic regression models, *fraternisation (x09)* twice showed a positive coefficient, twice a negative one, and was removed from the analysis four times. Even where it remained, the coefficient is very small, giving it a small impact on the result. It should therefore be removed in future analysis – or be replaced by a better proxy of "fraternisation" between countries than citizen of Country A living in Country B. The same is true for *differing RECs memberships (x13)*, which is twice negative (as expected), once positive, and removed no fewer than five times.

This leaves three variables that consistently show coefficients that were not anticipated. There are more, not fewer CBI in higher *Elevations (x05)*. A closer look at the data revealed that in fact, CBI have an average altitude of 677m, whereas non-CBI have an average altitude of 619m, with a standard deviation of 520 and 450 respectively. So, while there is a slight trend, it is questionable whether the variable should be kept or disregarded during the analysis.

Trade (x11) is measured as the percentage of trade in the total GDP, aiming to identify "extrovert" countries with a high share of trade. Some limitations to this approach were already discussed under 2.3.2.11, such as the fact that some high value mineral may pass through a single border post, that coastal states transport their goods by sea or that goods don't appear in the statistics of transit countries but require CBI in these countries nevertheless. The fact that there is a slight negative correlation may indicate that the indicator used is not suitable to measure extroversion of a country, i.e. if a lot of goods are crossing a border or not. Experienced economists may have advice about what indicator would be best to use.

Lastly, *shared RECs memberships (x12)* seem to have a negative rather than a positive impact on CBC. This is due to a special case: the Nigeria-Cameroon border has 233 CBI (paved roads), 7% of all CBI, while not being members in any shared REC.

In general, *pixel-specific variables* (x01-x08) have shown better results than variables that are only border-specific (x09-x13). Indicators that could measure fraternisation, trade or political cohesion on a local or even raster-cell specific level may therefore improve the models.

Another noteworthy point is that only 7 out of 87 coefficients have a *p-value* above 5%, most far lower, which means that for most coefficients, there is a highly significant relation between the dependent and the independent variable. This is also partly thanks to the multi-step removal of variables with high VIF and by stepwise regression.

4.2 Pseudo-R2

While a high Pseudo-R2 value is preferable, the relatively low values of 5-21% in the different models isn't necessarily something discouraging. It means that 5-21% of the occurrence of a CBI in any given cell can be explained by the respective model. This may sound like a small percentage, but it would be inappropriate to expect much higher values. CBI depend (or rather: are correlated with) many other factors than the ones we have analysed in this thesis. National parks, topographical, climatic, political or financial conditions may hamper the construction of roads much more than the variables evaluated. Also, if a paved road crosses the border in the neighbouring cell, it's rather unlikely that another road is built just 500m from the existing one. If we consider that there are many other factors to take into consideration, the 5-21% is probably the best we can get on a continental and regional level.

Although the models have low Pseudo-R2, they also have low p-values. This means that although the models only explain part of the CBI-phenomenon, there is a significant relation between the dependent and the independent variables.

4.3 Comparison of continental and regional models, prediction

As large CBI are ten times rarer than cross-border roads, coefficients for x03 are generally smaller than for x01. Pseudo-R2 is also smaller, which may be an effect of the different types of CBI that are mixed into a single variable. Airports, large bridges, hydropower stations and railroads may correlate with very different factors but are all joined into a single umbrella variable of large CBI.

Among the regional models, northern Africa stands out. Due to its geographic location between the Mediterranean Sea and the Sahara Desert, and given the general north-south orientation of its borders, it has the highest degree of multicollinearity. This coincides with the lowest number of events. Any conclusion derived from this model is thus to be taken with a pinch of salt.

Other than that, the regional models are very different from one another. Even the prediction of CBI in the southern region using the eastern model shows completely different (read: false) results. There are probably no other two regions on the African continent that are so alike than eastern and southern Africa, but still the results forbid using one model in the other region. This is also an indicator that models created for a sub-region would be rather different from the current regional ones.

The above-mentioned prediction is a bit atypical, as predictions in logistic regression are usually used a) to predict *future cases*, e.g. to estimate the number of lung cancer cases in ten years' time given the number of smokers among the current population or b) to predict a phenomenon for a *larger sample size*, e.g. the total number of trees infected by fungi Y in the whole forest, where collecting such data would be too difficult, time-consuming or costly. In the present case, dependent and independent variables are available in both sets, so the prediction is used to test the model by applying it to a different region.

4.4 Limitations, challenges

A number of unexpected situations were encountered during the work process. The solutions found to resolve these cases allowed to carry the analysis through to the end, but some had a considerable impact on the results.

Border geometry. In some areas, the de-jure border isn't congruent with the de-facto boundary. As there's no official dataset of the latter, there would be need for more research to modify the boundary data locally. Also, sometimes neighbouring countries don't even agree whether a boundary is disputed. Using (a simplified version of) the LSIB data is therefore a best-option for the present research, but it could surely be improved for future applications.

Data. Many of the data used are not directly measured, but an indirect indication of something else. GDP per capita is used in this research to model the economic power of the state relative to its population size, to understand how much it could invest in CBI. However, this may also depend on the size of the country, its political priorities, debt history etc. The river size is measured by the size of its drainage basin, but the actual

width that a bridge would have to span across depends on other factors as well, if the rivers moves seasonally or if there are marshlands. Even population density is based on several other datasets, including satellite imagery and population statistics. Also, while most data are relatively recent, they may change considerably in rather short time. While the data is deemed suitable for the present analysis, it surely has its limitations, and the output result can only be as good as the input data.

OSM. For y01 and its regional models, OSM is the only input data used aside from the LSIB. The shortfalls of OSM have already been discussed in chapter 2.2.2, but to repeat just the most important ones: OSM may not be complete in some areas, and its roads may carry wrong attributes, transforming real footpaths into digital main roads. While both will improve over time if the user community keeps contributing, such shortfalls must be considered in the present analysis, even if there's nothing that can be done about it at this stage.

The Sahrawi Republic. While data could be found for all the borders, although sometimes being not as recent as hoped for, this was not the case for the borders of the Sahrawi Republic, or Western Sahara. Given that the government of Western Sahara controls only a fifth of its territory, the rest being occupied by Morocco, and many other countries and international bodies not recognising Western Sahara as an independent state, this isn't surprising. In the end, it was decided not to include Western Saharan borders in the analysis.

Rare events. Large CBI as well as cross-border roads are, geographically speaking, the exception to the norm. This results in a low R2 and in generally low predicted probabilities. Although a pre-selection of the area to be analysed could reduce the number of non-events (e.g. only analyse areas with a population density of more than x persons per square kilometre), CBI would probably still be rather rare events.

5. Conclusion

The present analysis of cross-border infrastructure on the African continent, the first of its kind, shows several important results. Both large CBI as well as cross-border paved roads correlate with high population density and the length of the border segment in a given raster cell. They are also (negatively) correlated to the distance to the nearest capital, the slope and river size. This means, CBI are most likely to be in populated, flat areas, close to capitals, and on dry land or across small rivers.

The thesis has delivered results to the aims and objectives outlined in chapter 1.3. It has shown correlations between CBI and some of the other factors. It has created a dataset of large CBI as well as paved cross-border roads. It has created and models on a continental as well as a regional level and analysed the differences between them. It has also determined the probability of CBI in any given raster cell as well as summarised by border, and it has proposed various ways to visualise them.

But the analysis also bears valuable insight where it seemingly failed. The indicators used for Trade and Fraternisation show no clear effects. RECs on the other hand seem to have no influence on the existence of the CBI measured in this thesis. Of course, some RECs invest quite heavily in the ease of trade by establishing One Stop Border Posts (OSBP), but such "upgrades" of CBI isn't measured in the present work. Indicators used for economic, political and social integration (x09-x13) are boundary specific, i.e. they use one value for the entire boundary. Ideally, these should be replaced by pixel- or segment-specific indicators, e.g. measuring trade between neighbouring districts across an international border. Also, regional models show better results at an average R2 of 13% versus only 7% for continental models. Further analysis should thus be conducted on regional level and can only be applied to that area.

The analysis of CBI can only be as good as its underlying data, and improvement in the data quality, especially OSM, will lead to better and more reliable results. Also, the list of variables is not exhaustive, and other researchers with different professional backgrounds may evaluate factors such as good relations between countries, measured e.g. by number of bilateral treaties concluded.

Finally, the present thesis is a proof of concept. Analysis on CBI is possible and there are clear and important correlations. Although the analysis can only explain a fraction of the correlation between CBI and other analysis, it is an important contribution to understanding where CBI can be expected – and where not. Further research with additional or improved data is now needed to explore relations in greater detail. A first, successful step is made.

6. References

African Union. Draft African Union Border Governance Strategy (2017). Retrieved from http://www.peaceau.org/uploads/2018-06-14-aubgs-e.pdf

Allen, G. H., & Pavelsky, T. M. (2018). Global extent of rivers and streams. *Science*, eaat0636. https://doi.org/10.1126/science.aat0636

Anderson, J. E., & van Wincoop, E. (2004). Trade Costs. *Journal of Economic Literature*, *42*(3), 691–751. https://doi.org/10.1257/0022051042177649

AU. MoU on Security, Stability, Development and Cooperation in Africa (2002).

AUC. (2018). Regional Economic Communities. Retrieved 20 July 2018, from https://au.int/en/organs/recs

Bah, A. (2013). Civil Conflicts as a Constraint to Regional Economic Integration in Africa. *Defence and Peace Economics*, *24*(6), 521–534. https://doi.org/10.1080/10242694.2012.723155

Basiri, A., Jackson, M., Amirian, P., Pourabdollah, A., Sester, M., Winstanley, A., … Zhang, L. (2016). Quality assessment of OpenStreetMap data using trajectory mining. *Geo-Spatial Information Science*, *19*(1), 56–68. https://doi.org/10.1080/10095020.2016.1151213

Biger, G. (2013). Walls, fences and international borders. *Studia z Geografii Politycznej i Historycznej*, *02*, 87–108.

Bindschedler, & Dennery, E. Convention entre la Suisse et la France concernant l'aménagement de l'aéroport de Genève-Cointrin et la création de bureaux à contrôles nationaux juxtaposés à Ferney-Voltaire et à Genève-Cointrin, 0.748.131.934.91 § (1958). Retrieved from https://www.admin.ch/opc/fr/classified-compilation/19560067/index.html

Bittner, C. (2017). OpenStreetMap in Israel and Palestine – 'Game changer' or reproducer of contested cartographies? *Political Geography*, *57*, 34–48. https://doi.org/10.1016/j.polgeo.2016.11.010

Brinkhoff, T. (2016). OPEN STREET MAP DATA AS SOURCE FOR BUILT-UP AND URBAN AREAS ON GLOBAL SCALE. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLI-B4*, 557–564. https://doi.org/10.5194/isprsarchives-XLI-B4-557-2016

Brownlie, I. (1979). *African boundaries: a legal and diplomatic encyclopaedia*. London : Berkeley: C. Hurst ; University of California Press for the Royal Institute of International Affairs.

Bucsky. (2017). Schematic map of African railways by gauge. thematic map. Retrieved from https://en.wikipedia.org/wiki/African_Union_of_Railways#/media/File:Africa_railway_map_gauge.jpg

Cabrera-Barona, P., Blaschke, T., & Kienberger, S. (2017). Explaining Accessibility and Satisfaction Related to Healthcare: A Mixed-Methods Approach. *Social Indicators Research*, *133*(2), 719–739. https://doi.org/10.1007/s11205-016-1371-9

Chen, J., Li, R., Dong, W., Ge, Y., Liao, H., & Cheng, Y. (2015). GIS-Based Borderlands Modeling and Understanding: A Perspective. *ISPRS International Journal of Geo-Information*, *4*(2), 661–676. https://doi.org/10.3390/ijgi4020661

Department of State. (2017, December 29). Large Scale International Boundaries dataset. Retrieved from https://catalog.data.gov/dataset/africa-americas-lsib-lines-detailed-2017dec29

Department of State, & Humanitarian Intervention Unit. (2018). Large Scale International Boundaries. global: US state department. Retrieved from http://geonode.state.gov/layers/geonode%3AAfrica_Americas_LSIB9__

Dixon, B. (2009). A case study using support vector machines, neural networks and logistic regression in a GIS to identify wells contaminated with nitrate-N. *Hydrogeology Journal*, *17*(6), 1507–1520. https://doi.org/10.1007/s10040-009-0451-1

Donaldson, J. (2007). International River Boundary Database. Retrieved from https://www.dur.ac.uk/ibru/resources/irbd

Donaldson, J. W. (2009). Where rivers and boundaries meet: building the international river boundaries database. *Water Policy*, *11*(5), 629. https://doi.org/10.2166/wp.2009.065

Donaldson, J. W. (2011). Paradox of the Moving Boundary: Legal Heredity of River Accretion and Avulsion. *Water Alternatives*, *4*(2), 155–170.

Dorn, H., Törnros, T., & Zipf, A. (2015). Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany.

ISPRS International Journal of Geo-Information, *4*(3), 1657–1671. https://doi.org/10.3390/ijgi4031657

ecdpm. (2018). Regional Organisations in Africa [map]. Retrieved 7 August 2018, from https://indd.adobe.com/view/f49ac87d-7aa3-4cf7-822e-841d674bbc92

esri. (2015). World Population Density. esri. Retrieved from https://livingatlas.arcgis.com

esri. (2018). World Imagery. online basemap.

FAO. (2015). global population density estimates. Retrieved from http://www.fao.org/geonetwork/srv/en/resources.get?id=14053&fname=Map_2_3.zip&access=private

Forghani, M., & Delavar, M. (2014). A Quality Study of the OpenStreetMap Dataset for Tehran. *ISPRS International Journal of Geo-Information*, *3*(2), 750–763. https://doi.org/10.3390/ijgi3020750

Fung, K. C., García-Herrero, A., & Ng, F. (2011). Foreign Direct Investment in Cross-Border Infrastructure Projects. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1800472

Gilbert, J., & Banik, N. (2010). Socioeconomic Impacts of Cross-Border Transport Infrastructure Development in South Asia. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1586747

Girres, J.-F., & Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset: Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, *14*(4), 435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

Guns, M., & Vanacker, V. (2012). Logistic regression applied to natural hazards: rare event logistic regression with replications. *Natural Hazards and Earth System Science*, *12*(6), 1937–1947. https://doi.org/10.5194/nhess-12-1937-2012

Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. https://doi.org/10.1068/b35097

Hartzenberg, T. (2011). Regional Integration in Africa. *Working Paper*.

Hertslet, S. E. (1896). *The Map of Africa by Treaty* (Vol. 3). Harrison and Sons. Retrieved from https://archive.org/details/mapofafricabytre03hertuoft

Hisakawa, N., Jankowski, P., & Paulus, G. (2013). Mapping the porosity of international border to pedestrian traffic: a comparative data classification approach to a study of the border region in Austria, Italy, and Slovenia. *Cartography and Geographic Information Science*, *40*, 18–27. https://doi.org/10.1080/15230406.2013.762141

Imam, E., & Kushwaha, S. P. S. (2013). Habitat suitability modelling for Gaur (*Bos gaurus*) using multiple logistic regression, remote sensing and GIS. *Journal of Applied Animal Research*, *41*(2), 189–199. https://doi.org/10.1080/09712119.2012.739089

IOM. (2015). Global Migration Flows. Retrieved from https://www.iom.int/world-migration

ISO. (2017). ISO 3166 alpha-2. Retrieved from https://www.iso.org/obp/ui/#search

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137–163.

Kopiński, D., & Polus, A. (2017). Is Botswana Creating a New Gaza Strip? An Analysis of the 'Fence Discourse'. In *Crossing African Borders: Migration and Mobility* \. Lisboa: Centro de Estudos Internacionais. Retrieved from http://books.openedition.org/cei/216

Lehner, B., Verdin, K., & Jarvis, A. (2008). Hydrological data and maps based on SHuttle Elevation Derivatives at multiple Scales. Retrieved from http://hydrosheds.org

Linthicum, D. (2017, March 3). RE: LSIB.

Linthicum, D. (2018, May 21). LSIB - Pagak.

Michniak, D. (2011). The Effects of European Integration on the Development of Cross-Border Transport Infrastructure: the Example of Slovak-Polish Boundary, (Special Issue-Aspects of Localities), 45–52.

minitab. (2014, June 12). How to Interpret a Regression Model with Low R-squared and Low P values. Retrieved 21 December 2018, from https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

Mun, S., & Nakagawa, S. (2010). Pricing and investment of cross-border transport infrastructure. *Regional Science and Urban Economics*, *40*(4), 228–240. https://doi.org/10.1016/j.regsciurbeco.2010.03.008

NATO. (2017). NATO country codes. Retrieved from https://www.nato.int/structur/AC/135/main/links/codsp3.htm

Niedziałkowska, M., Jędrzejewski, W., Mysłajek, R. W., Nowak, S., Jędrzejewska, B., & Schmidt, K. (2006). Environmental correlates of Eurasian lynx occurrence in Poland – Large scale census and GIS mapping. *Biological Conservation*, *133*(1), 63–69. https://doi.org/10.1016/j.biocon.2006.05.022

Nmor, J. C., Sunahara, T., Goto, K., Futami, K., Sonye, G., Akweywa, P., … Minakawa, N. (2013). Topographic models for predicting malaria vector breeding habitats: potential tools for vector control managers. *Parasites & Vectors*, *6*(1), 14. https://doi.org/10.1186/1756-3305-6-14

Norman, L. M., Feller, M., & Phillip Guertin, D. (2009). Forecasting urban growth across the United States–Mexico border. *Computers, Environment and Urban Systems*, *33*(2), 150–159. https://doi.org/10.1016/j.compenvurbsys.2008.10.003

OAU. on border disputes between African States, Resolution AHG/RES.16 (1) § (1964).

Ohlmacher, G. C., & Davis, J. C. (2003). Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. *Engineering Geology*, *69*(3–4), 331–343. https://doi.org/10.1016/S0013-7952(03)00069-3

openflights.org. (2017, January). airport database. Retrieved from https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat

Ortiz-Pelaez, A., Pfeiffer, D. U., Tempia, S., Otieno, F. T., Aden, H. H., & Costagli, R. (2010). Risk mapping of Rinderpest sero-prevalence in Central and Southern Somalia based on spatial and network risk factors. *BMC Veterinary Research*, *6*(1), 22. https://doi.org/10.1186/1746-6148-6-22

OSM. (2018). Open Street Map. Africa. Retrieved from http://download.geofabrik.de/africa-latest.osm.pbf

Parr, D. A. (2015). *The Production of Volunteered Geographic Information: A Study of OpenStreetMap in the United States*. Texas State University, Texas, USA. Retrieved from https://digital.library.txstate.edu/bitstream/handle/10877/5776/PARR-DISSERTATION-2015.pdf?sequence=1

Petitpierre, M., & Hoppenot, H. Convention franco-suisse relative à la construction et à l'exploitation de l'aéroport de Bâle-Mulhouse, à Blotzheim, 0.748.131.934.92 §

(1949). Retrieved from https://www.admin.ch/opc/fr/classified-compilation/19490164/197102250000/0.748.131.934.92.pdf

Pir Bavaghar, M. (2016). Deforestation modelling using logistic regression and GIS. *Journal of Forest Science*, *61*(No. 5), 193–199. https://doi.org/10.17221/78/2014-JFS

Pratt, M. (2016, April). *Principles of river boundary delimitation*. Presentation, Lusaka. Retrieved from document not available online

Puka, L., & Szulecki, K. (2014). The politics and economics of cross-border electricity infrastructure: A framework for analysis. *Energy Research & Social Science*, *4*, 124–134. https://doi.org/10.1016/j.erss.2014.10.003

Razniewski, S., & Nutt, W. (2014). Adding completeness information to query answers over spatial databases (pp. 123–132). ACM Press. https://doi.org/10.1145/2666310.2666395

Rothmann, J. (2013, November). Ongeluksnek border post. Retrieved 17 August 2018, from https://saportsofentry.blogspot.com/2013/11/matatiele-scenic-drives-to-3-border.html

Shafapour Tehrany, M., Shabani, F., Neamah Jebur, M., Hong, H., Chen, W., & Xie, X. (2017). GIS-based spatial prediction of flood prone areas using standalone frequency ratio, logistic regression, weight of evidence and their ensemble techniques. *Geomatics, Natural Hazards and Risk*, *8*(2), 1538–1561. https://doi.org/10.1080/19475705.2017.1362038

Sheather, S. (2009). *A modern approach to regression with R*. New York: Springer.

Siebritz, L., Sithole, G., & Zlatanova, S. (2012). Assessment of the homogeneity of volunteered geographic information in South Africa. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XXXIX-B4*, 553–558. https://doi.org/10.5194/isprsarchives-XXXIX-B4-553-2012

Sorriso-Valvo, M., Greco, R., & Catalano, E. (2008). Spatial prediction of regional-scale mass movement using Logistic Regression analysis and GIS—Calabria, Italy. *Israel Journal of Earth Sciences*, *57*(3), 263–280. https://doi.org/10.1560/IJES.57.3-4.263

Srinivasan, P. V. (2012). Regional Cooperation and Integration through Cross-Border Infrastructure Development in South Asia: Impact on Poverty. *South Asia*

Working Paper Series, (14). Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13504509.2011.644639

Srivastava, S. K., Saran, S., de By, R. A., & Dadhwal, V. K. (2014). A geo-information system approach for forest fire likelihood based on causative and anti-causative factors. *International Journal of Geographical Information Science*, *28*(3), 427–454. https://doi.org/10.1080/13658816.2013.797984

Stiglbauer, A., & Weiss, C. (2000). Family and Non-Family Succession in the UpperAustrian Farm Sector. *Cahiers d'économie et Sociologie Rurales, no. 54*, 7–26.

Tayyebi, A., Perry, P. C., & Tayyebi, A. H. (2014). Predicting the expansion of an urban boundary using spatial logistic regression and hybrid raster–vector routines with remote sensing and GIS. *International Journal of Geographical Information Science*, *28*(4), 639–659. https://doi.org/10.1080/13658816.2013.845892

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2016). Predicting Road Accidents: A Rare-events Modeling Approach. *Transportation Research Procedia*, *14*, 3399–3405. https://doi.org/10.1016/j.trpro.2016.05.293

UNECA. (2016). *Assessing regional integration in Africa VII: innovation, competitiveness and regional integration*. (African Union & African Development Bank Group, Eds.). Addis Ababa, Ethiopia: Economic Commission for Africa.

Valensisi, G., Lisinge, R., & Karingi, S. (2016). The trade facilitation agreement and Africa's regional integration. *Canadian Journal of Development Studies / Revue Canadienne d'études Du Développement*, *37*(2), 239–259. https://doi.org/10.1080/02255189.2016.1131672

Van der Geest, W., & Nunez-Ferrer, J. (2011). Managing Multinational Infrastructure: An Analysis of EU Institutional Structures and Best Practices. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1886247

Van Doninck, J., De Baets, B., Peters, J., Hendrickx, G., Ducheyne, E., & Verhoest, N. (2014). Modelling the Spatial Distribution of Culicoides imicola: Climatic versus Remote Sensing Data. *Remote Sensing*, *6*(7), 6604–6619. https://doi.org/10.3390/rs6076604

van Heuvel, T. (2015). Retrieved 17 August 2018, from http://static.panoramio.com/photos/large/34671810.jpg

Warr, P., Menon, J., & Yusuf, A. A. (2010). Regional Economic Impacts of Large Projects: A General Equilibrium Application to Cross-Border Infrastructure. *Asian Development Review*, *27*(1), 104–134.

Wikipedia. (2018a). International border crossings [encyclopedia]. Retrieved 18 July 2018, from https://en.wikipedia.org/wiki/Category:International_border_crossings

Wikipedia. (2018b, May 13). List of power stations in Africa. Retrieved 13 May 2018, from https://en.wikipedia.org/wiki/List_of_power_stations_in_Africa

World Bank. (2017). GDP per Capita (current USD). Retrieved from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?view=chart

World Bank. (2018). Import / Export of goods and services (% of GDP). Retrieved 21 November 2018, from https://data.worldbank.org/indicator/NE.EXP.GNFS.ZS?end=2017&name_desc=false&start=1996

WTO. (2017). Trade to GDP ratio. thematic map. Retrieved from https://www.wto.org/english/res_e/statis_e/statis_maps_e.htm

Xing, H., Chen, J., & Zhou, X. (2015). A Geoweb-Based Tagging System for Borderlands Data Acquisition. *ISPRS International Journal of Geo-Information*, *4*(3), 1530–1548. https://doi.org/10.3390/ijgi4031530

Zhang, Y., Li, X., Wang, A., Bao, T., & Tian, S. (2015). Density and diversity of OpenStreetMap road networks in China. *Journal of Urban Management*, *4*(2), 135–146. https://doi.org/10.1016/j.jum.2015.10.001

Zhang, Z. X., Zhang, H. Y., & Zhou, D. W. (2010). Using GIS spatial analysis and logistic regression to predict the probabilities of human-caused grassland fires. *Journal of Arid Environments*, *74*(3), 386–393. https://doi.org/10.1016/j.jaridenv.2009.09.024

Zhou, X., Zeng, L., Jiang, Y., Zhou, K., & Zhao, Y. (2015). Dynamically Integrating OSM Data into a Borderland Database. *ISPRS International Journal of Geo-Information*, *4*(3), 1707–1728. https://doi.org/10.3390/ijgi4031707