# Modellierung der Ausbreitungspfade der Bantu-Sprachfamilie mithilfe Geographischer Informationssysteme

MASTERARBEIT GEO 511

Abgabedatum: 30.06.2014

Geographisches Institut der Universität Zürich
Abteilung Geographische Informationssysteme

Christian Wirth
05-730-932

Fakultätsinterne Betreuer:
Prof. Dr. Robert Weibel &
Dr. Curdin Derungs

Externer Betreuer:
Prof. Dr. Balthasar Bickel
Seminar für Allgemeine Sprachwissenschaft UZH

**Kontakt:**

Christian Wirth
chrigelwirth@gmail.com

Prof. Dr. Robert Weibel
Universität Zürich
Geographisches Institut
Abteilung Geographische Informationssysteme
Winterthurerstrasse 190
CH-8057 Zürich

Dr. Curdin Derungs
Universität Zürich
Geographisches Institut
Abteilung Geographische Informationssysteme
Winterthurerstrasse 190
CH-8057 Zürich

Prof. Dr. Balthasar Bickel
Universität Zürich
Seminar für Allgemeine Sprachwissenschaft
Plattenstrasse 54
CH-8032 Zürich

# Zusammenfassung

Mithilfe linguistisch-genetischer Forschungen gelang es Sprachwissenschaftlern in jüngster Zeit erstmals, unser Wissen bezüglich des ungefähren Verlaufs der Bantu-Expansion auf eine abgestützte Basis zu stellen. Die Studien untermauerten die Theorie eines Late-Split Modells, nach welchem sich die Bantu-Sprachfamilie zuerst als Einheit vom Ursprungsgebiet zwischen Nigeria und Kamerun Richtung Süden ausbreitete und sich erst dann in zwei Hauptäste verzweigte.

In dieser Arbeit wurde der Frage nachgegangen, ob der Verlauf der Bantu-Expansion allenfalls auch aus einer geographischen Perspektive zu erklären sei.
Dabei wurde aufgrund der Verwandtschaft des Problems die oft in der Archäologie benutzte GIS-Methode der Least-Cost Path Analyse (LCPA) angewandt. In der Untersuchung wurden 138 geographisch lokalisierte Bantu-Sprachen verwendet.
Die Kriterien der Relevanz bezüglich des gewählten Pfades sowie der gegebenen Modellierbarkeit erfüllten die Faktoren Neigung, Vegetation und Gewässer. Letztere wurden bereits vor der Berechnung der Pfade als Gebiete mit besonders hohen Kosten in die Vegetationsrasterkarte integriert.
Aufgrund der geringen Vorkenntnisse respektive des stark explorativen Charakters der Untersuchung wurde ein pragmatischer Ansatz gewählt. Statt mithilfe eines komplexen a priori bestimmten Konzepts sollten Unsicherheiten mit späteren Modellvariationen einkalkuliert werden.

Die Untersuchung der Kostenpfade für die einzelnen Faktoren zeigte einen vernachlässigbaren Einfluss der Hangneigung. Im Folgenden wurden nur noch die kombinierten Vegetations- und Gewässerdaten verwendet. Um die Wahrscheinlichkeit des Early-Splits gegenüber dem Late-Split besser untersuchen zu können, wurde eine Multipunkte-LCPA durchgeführt. Dabei resultierte jedoch kein eindeutiges Resultat zugunsten des Early-Splits. Neben einem südwärts gerichteten Hauptpfad wurde ein weiterer Zweig bestimmt, welcher den zentralen Regenwald im Norden umging. Mit der Bestimmung der Shreve-Flussordnungen konnte jedoch ein starkes Übergewicht des südlichen Pfades verdeutlicht werden.
Um die Kostenpfade auch auf linguistischer Ebene zu analysieren, wurden im Folgenden mittels der Pfaddistanzen hierarchische Cluster generiert. Dabei wurden neben den räumlichen Distanzen zusätzlich die topologischen Distanzen zur Clusterbildung verwendet. Dies ermöglichte einen Vergleich mit vorhandenen genealogischen Sprachdaten. Als Kontrollgruppe wurden noch Dendrogramme aus den euklidischen Distanzen zwischen den Punkten gebildet, welche ebenfalls mit dem genealogischen Baum verglichen wurden. Eine Überprüfung der Übereinstimmung der Cluster mittels Consensus-Baum führte jedoch zu sehr tiefen Werten. In Folge wurden noch die anhand der genealogischen Baumstruktur eingefärbten Punkte auf der geographischen Karte genauer betrachtet. Hierbei zeigte sich in den Grundzügen eine gewisse Übereinstimmung mit den Kostenpfaden, welche einer genaueren Untersuchung bedürfte. Als erweiterter Ansatz wird dazu die Methode der Betweenness centrality vorgeschlagen.

# Abstract

By means of linguistical-genetical research linguists have recently succeeded in putting our knowledge in regards of the approximate course of the Bantu-expansion on a solid base. The studies backed up the theory of a Late-Split Model according to which the Bantu-language first extended as a unity from its original region between Nigeria and Cameroon towards the South, before splitting in two main branches.

In this thesis we have pursued the question whether the course of the Bantu expansion might best be explained from a geographical point of view. Thereby the GIS method of the Least Cost Path Analysis (LCPA), often applied in archeology, was used. In the analysis, 138 geographically located Bantu languages were employed. The costs were represented by the factors slope, vegetation and inshore waters, due to their relevance as criteria for the expansion path as well as their modelling capability. Inshore waters were already integrated into the vegetation raster map before the calculation of the paths as regions with particular high costs.

Due to the limited existing knowledge and the exploratory nature of this research, a pragmatic solution was chosen. Thus, instead of attempting to develop a sophisticated and complex *a priori* modelling concept, inherent uncertainties should be taken into account by posterior modelling variations.
The study of the cost paths for the individual factors showed an insignificant influence of slope as a cost factor. Thus, in the following, only the combined effects of vegetation and inshore waters were used. In order to be able to investigate the probability of the Early-Split compared to the Late-Split theory, respectively, a multi-points LCPA was then performed. Thereby, no explicit result in support of the Early-Split was produced. Apart from a main path heading south, a further branch was identified, which circumvents the central rain forest on a northern line. However, computing the Shreve stream orders of the resulting LCP tree revealed a stark over-representation of the southern branch.

In order to also analyse the cost paths from the linguistic perspective, hierarchical clusters were subsequently generated from the path distances. Besides the spatial distances, the topological distances were used for the formation of clusters. This allowed testing for the degree of correspondence with existing genealogical data. As a control group, dendrograms generated using the Euclidean distance between the points were formed and likewise compared with the genealogical tree. However, a survey of the congruence of the clusters by means of a consensus-tree led to very low values. In succession, the clades of the genealogical tree were colourised on the basis of their hierarchical level. Afterwards, the pattern of the colourised languages was regarded in more detail on the geographical map.
Thereby, a certain correlation with the cost paths in the main features showed up, which would require a further research. As an extended approach the method of Betweenness Centrality has been suggested.

# Danksagung

Ich bedanke mich herzlich bei meinen Betreuern Prof. Dr. Robert Weibel, Dr. Curdin Derungs und Prof. Dr. Balthasar Bickel für ihre Anregungen und Ideen, ihre konstruktive Kritik und ihre fachliche Unterstützung.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# 1 Einleitung

## 1.1 Kontext und wissenschaftliche Motivation

Dank aufwendiger Untersuchungen und Auswertungen linguistischer Variablen besitzt die Sprachwissenschaft heute umfangreiche Daten zu den weltweit vorhandenen Sprachen. Dazu bestimmte Punktkoordinaten ermöglichen zudem ein Bild ihrer räumlichen Verteilungen (WALS 2011, Nordhoff et al. 2013).
Entsprechend existieren diverse Untersuchungen zu den möglichen historischen Ausbreitungen der verschiedenen Sprachfamilien. Dabei werden in der Regel linguistische Verfahren angewandt, welche in jüngerer Zeit zunehmend durch verbesserte Möglichkeiten in der DNA-Analyse abgestützt werden können (De Montano et al. 2011, Alves et al. 2011, Filippo et al. 2012).

Eine Sprachfamilie, bei der über ihre ungefähre Ausbreitung aufgrund neuer Forschungsresultate unter Experten heute Einigkeit herrscht, stellt die mittelafrikanische Bantusprache dar (Currie et al. 2013, De Filippo et al. 2012). Die Bantu-Sprachfamilie, welche ihren Ursprung im Grenzgebiet zwischen Nigeria und Kamerun hat, zählt mit etwa 500 verschiedenen Sprachen zu den grössten der Welt.
Während bisherige Studien die Ausbreitung der Bantusprachen mittels linguistischer Verfahren, sowie in jüngster Zeit auch anhand des Genflusses in der Bevölkerung untersuchten, wird in der hier vorgestellten Masterarbeit der Frage nachgegangen, ob diese auch mithilfe geographischer Faktoren modelliert werden könnte.

*Geographische Informationssysteme (GIS)* wurden seit ihrem Aufkommen in den 1980-er Jahren auch in der Archäologie oft genutzt, wobei vor allem historische Landschaftsbilder mithilfe moderner Techniken untersucht wurden. Dem ebenfalls interessanten räumlich-gesellschaftlichen Studiengebiet der historischen Völkerbewegungen wurde bisher hingegen wenig Aufmerksamkeit geschenkt (Howey 2007). Räumliche Bewegung stellt seit jeher einen wesentlichen Aspekt in der Geschichte der Menschheit dar. Frühere Gesellschaften waren in ihrer Bewegung im Vergleich zu heutigen Gesellschaften stark eingeschränkt. Diese Einschränkung beruhte auf primitiveren Transportmitteln sowie schlechten Strassen und Wegen, welche die Fortbewegung zusätzlich erschwerten. Trotzdem wurden damals bereits enorme Distanzen bewältigt. Entsprechend liegt in der Analyse der räumlichen Bewegungen auch für die archäologischen Wissenschaften ein grosses Potential, Informationen über das Wesen vergangener Zivilisationen zu erhalten (Howey 2007). Diese erlaubt beispielsweise Rückschlüsse über Reisenetzwerke und Interaktionen zwischen verschiedenen Kulturen, zentrale Untersuchungsgebiete in der Archäologie (Surface 2012). Sie versucht, folgende Frage zu beantworten: „Wie kamen historisch entstandene Fusspfade und Wege zu ihrem jeweiligen Verlauf, welches sind dabei die entscheidenden Kriterien?“ Selbstverständlich liegen den einzelnen Wegen jeweils unterschiedliche Entstehungsgründe zugrunde. Einige entstanden beispielsweise als Tierpfade, welche später von den Menschen übernommen wurden. Andere wurden

vorderhand nach praktischen, respektive kostentechnischen Gesichtspunkten geplant. Es ist jedoch anzunehmen, dass ein wiederholt benutzter Pfad, welcher zwei Punkte miteinander verbindet, in gewisser Weise eine optimale Route zwischen diesen darstellt. Dies wiederum impliziert, dass bei der Wegfindung die dafür aufzuwendenden Kosten ausschlaggebend waren (Rees 2003).

Auf diesem Kostenkonzept bauen auch die modernen analytischen Untersuchungsmethoden räumlicher Bewegungen aus dem GIS-Bereich auf. Es wird angenommen, dass der Mensch generell dazu neigt, sein Verhalten zu rationalisieren und dass dieser Wesenszug sich auch in seinen Reisen widerspiegelt.
Entsprechend bildet das Bestimmen der kürzesten Route durch ein Gebiet eine der ältesten räumlichen Problemstellungen überhaupt (Surface 2012).
Dabei wurden die Methoden zur Festlegung des kürzesten Pfades immer raffinierter. Der heute vorhandene Überfluss an räumlichen Daten, kombiniert mit computerbasierten Analysen, ermöglicht die Bestimmung der kürzesten Pfade mit Einbezug komplexer Interaktionen von ingenieurstechnischen sowie ökologischer und teilweise sogar sozialer Kritikpunkte. Während in der Vergangenheit in einem extrem zeitaufwendigen Prozess verschiedene Pfade auf Karten aufgetragen wurden und mithilfe von Expertenwissen deren Durchführbarkeit geschätzt und verglichen werden mussten, sind die modernen Methoden weitgehend automatisiert und aufgrund ihrer Systematik auch gründlicher und weniger fehleranfällig (Berry 2000).
Das entsprechende GIS-Werkzeug zur Bestimmung kosteneffizientester räumlicher Pfade ist die *Least Cost Path Analyse (LCPA)*. In dieser Arbeit wurde daher versucht, solche Bewegungen mittels GIS zu rekonstruieren und ein mögliches Anwendungsgebiet der LCPA in den Sprachwissenschaften aufzuzeigen.

## 1.2 Zielsetzung und Forschungsfragen

### 1.2.1 Problemstellung

Wie verschiedene Studien (u.a. Currie et al. 2013, Diamond und Bellwood 2003) aufzeigen, deutet die Häufigkeit der Sprachen bezüglich der Breitengrade auf einen ökologischen Zusammenhang hin (höhere Sprachendichte bei höherer Primärproduktion). Ähnliche Überlegungen könnten auch bei der Ausbreitung der Sprachen angebracht werden. Es scheint plausibel, dass die Sprecher der jeweiligen Sprachen zu einem gewissen Grad den Weg des geringsten Widerstands gingen, also beispielsweise eher in fruchtbare Gebiete als in Wüsten vordrangen.

Die möglichen resultierenden Pfade lassen sich mithilfe der GIS-Methode der Kostenoberflächen ermitteln, wobei die Gebiete gemäss dem zur Distanzüberwindung benötigten Aufwand eingeteilt werden und anschliessend der effizienteste Weg ausgewählt wird.

### 1.2.2 Zielsetzung und Forschungsfragen

In dieser Arbeit soll überprüft werden, ob das Werkzeug der Kostenoberflächen als sinnvolle Alternative zu „linguistisch-genetischen Ansätzen“ verwendet werden kann. Da die genealogischen Ausbreitungspfade bei der Bantu-Familie relativ gut belegt sind, wird die Methode in diesem Gebiet untersucht. Die solide Wissensgrundlage ermöglicht eine Überprüfung der aus dieser Arbeit erhaltenen Resultate. Diese Vergleichsbasis ist insofern besonders wichtig, als dass eine LCPA in diesem Rahmen natürlich unmöglich mittels Ground Truth validiert werden kann.
Dazu kommt, dass eine Analyse, welche sich bloss auf die geographische Lage der Pfade beschränkt, aufgrund der geringen räumlichen Auflösung der bestehenden linguistischen Studien nur wenig Aussagekraft hätte. Da zu den Bantu-Sprachen umfassende genealogische Daten vorhanden sind, ist es jedoch möglich, diese auf Übereinstimmungen mit den LCP-Daten zu untersuchen.

Ein weiteres mögliches Anwendungsbeispiel wären die Na-Dené-Sprachen in Nordamerika, welche gemäss neueren Studien eine Verwandtschaft mit der sibirischen Jenissei-Sprachfamilie aufweisen. Eine zusätzliche Untersuchung der Na-Dené-Sprachen würde jedoch den Rahmen dieser Masterarbeit sprengen, deshalb wird nur auf die Bantu-Sprachen eingegangen.

Die zentralen Forschungsfragen der Arbeit lauten folgendermassen:

- *Welche geographischen Kriterien sollten bei der Modellierung einbezogen und wie können diese kalibriert und gewichtet werden?*

- *Inwiefern kann die Least Cost Path Analyse als ergänzende Methode zur Untersuchung historischer Sprachausbreitungen herangezogen werden?*

# 2 Hintergrund

In diesem Kapitel werden zuerst ein paar linguistische Begriffe erklärt, welche für das Verständnis der Arbeit wichtig sind. Zudem wird ein kurzer Überblick der weltweiten sprachlichen Diversität gegeben sowie auf das dynamische Wesen von Sprachen eingegangen. Danach folgt eine Einführung in das Untersuchungsgebiet sowohl im linguistischen, als auch im geographischen Kontext.
Des Weiteren werden Ergebnisse neuer Forschungen zur Bantu-Ausbreitung betrachtet sowie auf etwaige Forschungslücken hingewiesen.

## 2.1 Einblick in die Linguistik

### 2.1.1 Sprachen

Es gibt eine Vielzahl von Kriterien, anhand derer Sprachen analysiert und untereinander abgegrenzt werden können. Zum einen kann ihre Unterscheidung über eine Vielzahl struktureller Merkmale erfolgen. Diese beinhalten unter anderem die Phonologie, welche sich mit der Organisation der Laute befasst, sowie die Strukturen von Wörtern (Morphologie) und Sätzen (Syntax). Zudem gibt es aber auch noch viele andere Gesichtspunkte zur genauen Einteilung wie etwa der Zeitpunkt, der Raum, der soziokulturelle Hintergrund der Sprecher sowie die Art und Weise, wie das Wissen über eine Sprache vorliegt (Gavin 2013).

In der Umgangssprache assoziieren wir den Begriff „Sprache“ zumeist mit der „dazugehörigen“ Nation. So wird in Frankreich Französisch, in Italien Italienisch und in England Englisch gesprochen (McGregor 2009). In der Sprachwissenschaft ist eine Sprache dadurch definiert, dass sie sich von allen anderen Sprachen so weit unterscheidet, dass sie von deren Sprechern ohne Vorkenntnisse nicht verständlich ist (Hammarström 2012). Das Kriterium der gegenseitigen Verständlichkeit zur Bestimmung von Sprachen weist somit Ähnlichkeiten gegenüber der gängigsten Definition einer biologischen Art im Tierreich auf, bei welcher die erfolgreiche Fortpflanzung zwischen den Individuen als Grundbedingung angesehen wird. Wie der biologische Artbegriff ist aber auch die Abgrenzung respektive die Identifikation einer Sprache in der Praxis oft nicht unproblematisch. Entsprechend sind viele Sprachdatensätze aus der Linguistik bezüglich ihrer sprachlichen Auflösung nicht ganz einheitlich: Was in einzelnen Datenbanken etwa als Übersprache mit mehreren Untersprachen definiert wird, stellt in anderen Quellen bereits die tiefste sprachliche Ebene dar. So ist beispielsweise die sehr umfassende Glottolog-Sprachdatenbank bis in kleinste Unteräste verzweigt und beinhaltet eine wesentlich höhere Anzahl an Sprachen respektive Dialekten als andere Datenbanken (vgl. Nichols & Bickel 2009, Nordhoff et al. 2013). Generell sollte die Verwendung von Sprachlisten aus verschiedenen Quellen mit Vorsicht erfolgen, da bei deren Erstellung oftmals unterschiedliche Massstäbe und Kriterien verwendet werden (McGregor 2009).

### 2.1.2 Sprachfamilien

Wiederum ähnlich wie bei der biologischen Art (überhaupt existieren einige Parallelen zwischen biologischen Untersuchungsgebieten und der Sprachforschung) können auch Sprachen insofern zusammengefasst werden, als dass sie von einer gemeinsamen Ursprungssprache abstammen (Gavin et al. 2013). Dieser gemeinsame Vorfahre wird als Protosprache bezeichnet. Sprachen mit identischer Protosprache sind folglich genealogisch verwandt und werden daher derselben Sprachfamilie zu-

geteilt (McGregor 2009). Aber auch die Gruppierung von Sprachen zu Sprachfamilien erfolgt bei verschiedenen Autoren nicht immer einheitlich und kann somit problematisch sein (Gavin et al. 2013). Ob alle Sprachen ursprünglich aus einer einzigen Muttersprache abgeleitet wurden, kann nicht mit Klarheit gesagt werden. Aufgrund der Unstetigkeit der Sprachen können diese höchstens auf etwa 10'000 Jahre zurückverfolgt werden (McGregor 2009). Neben der Kladogenese, bei der sich eine Sprache in neue Untersprachen aufteilt, kann zudem auch eine Umänderung respektive Neubildung einer Sprache ohne Zunahme der Anzahl an Sprachen stattfinden (Anagenese) (Gavin et al. 2013).

### 2.1.3 Die Diversität der Sprachen

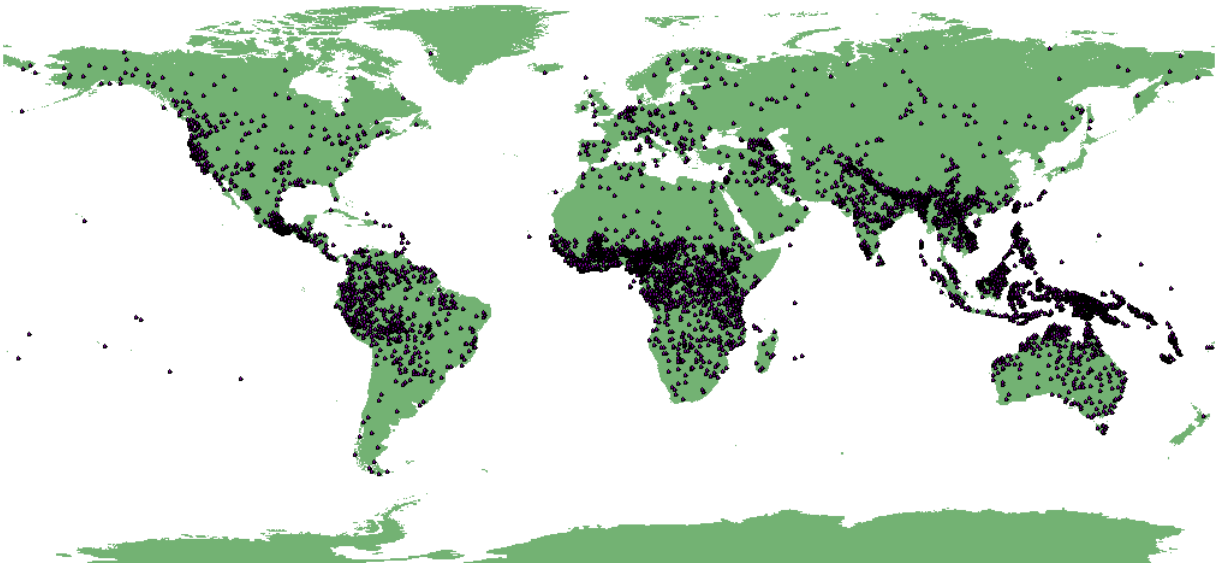


*Abbildung 1: Globale Verteilung der Sprachdaten von Hammarström (2013); Eigene Darstellung*

Weltweit gibt es schätzungsweise 7000 verschiedene Sprachen (Gavin et al. 2013). Diese sind jedoch keineswegs homogen über die Erdoberfläche verteilt: Während beispielsweise in China, welches eine Fläche von 9.5 Millionen km² aufweist, nur gerade etwa 235 verschiedene Sprachen gesprochen werden, zählt die mehr als zehnmal kleinere Insel Neuguinea ungefähr 1000 einzelne Sprachen (Currie & Mace 2009). Im mehr als 20 Mal grösseren Russland werden sogar nur 105 Sprachen gesprochen (Gavin et al. 2013).

Dabei ist jedoch zu berücksichtigen, dass zwischen diesen Sprachen auch enorme Unterschiede bezüglich der Anzahl der Sprecher bestehen. So machen die neun am meisten gesprochenen Sprachen zusammen mehr als 40 Prozent der Weltbevölkerung aus. Umgekehrt zählen beinahe die Hälfte aller Sprachen (3340) keine 10‘000 Sprecher (McGregor 2009). Bei einem Blick auf eine globale Karte der Sprachverteilungen zeigt sich auf den ersten Blick eine deutlich niedrigere Sprachdichte im Norden als im Süden.

Naheliegenderweise sind die tiefen Werte für Sibirien und Grönland eine Folge der geringen Bevölkerungsdichte, jedoch kann diese nicht als einzige Ursache dafür herangezogen werden. Dies zeigt sich wiederum am Beispiel von China, wo trotz massiv höherer Bevölkerungsdichte die Sprachvielfalt viel kleiner ist als in Neuguinea (McGregor 2009).

Wie eine Reihe von Studien (Nettle 1996, Gavin et al. 2013, Currie & Mace 2009) gezeigt hat, ist die Verteilung der Sprachen auch massgeblich von geographischen Faktoren bestimmt (Tabelle 1) . Das entscheidende Kriterium stellt hierbei das „ökologische Risiko“ dar, welches für die jeweilige Gesellschaft besteht. Nettle (1996) definierte dazu einen Index, welcher die Anzahl an Monaten mit mittleren Temperaturen über 6° Celsius sowie einen davon abhängigen Mindestniederschlag misst. Er konnte einen starken Zusammenhang zwischen diesem Index und der sprachlichen Diversität nachweisen. Dieser Zusammenhang wird dadurch erklärt, dass Gesellschaften, welche einem höheren Umwelt-Risiko ausgesetzt sind, grössere Netzwerke mit ihren Nachbarstämmen bilden müssen, um dieser Unsicherheit beizukommen Dies führt vermehrt zu einer sprachlichen Homogenisierung.

| Study | Geographic area | Unit of analsis | Empirical approach | Dependent variable |
|---|---|---|---|---|
| Currie and Mace 2009 | Old World[a] | Ethnolinguistic group | Linear mixed model | The area over which the language is spoken |
| Currie and Mace 2012 | Global | Ethnolinguistic group | Hierarchical linear modeling | The area over which the language is spoken |
| Fincher and Thornhill 2008 | Global | Country | Correlation | Language richness |
| Gavin and Sibanda 2012 | Pacific Islands | Island | Multiple regression | Language richness |
| Mace and Pagel 1995 | North America | Country | Correlation | Language density |
| Manne 2003 | Central and South America | Cell (1°)[d] | Correlation, nonparametric regression tree analysis | Language richness |
| Michalopolous 2008 | Virtual | Cell (0.5° × 0.5°), country | Multiple regression | Language richness |
| Moore et al. 2002 | Sub-Saharan Africa | Cell (2° × 2°) | Multiple regression | Language richness |
| Nettle 1996 | West Africa | Cell (2°) | Correlation | The number of languages per area and per head of population |
| Nettle 1998 | Global[b] | Country | Correlation | The number of languages per area and per head of population |
| Nettle 1999a | Global[c] | Country | Correlation | Language diversity per area and per head of population |
| Sutherland 2003 | Global | Country | General linear modeling, correlation | Language richness |

*Tabelle 1: Liste empirischer Studien zur Bestimmung der sprachlichen Diversität; angepasst aus Gavin et al. (2013)*

Auch Martina Köhli (2013) zeigte in ihrer Arbeit auf, dass die globale Verteilung der Sprachen durch geographische Faktoren bedingt ist. Dazu bezog sie in ihrer Studie sowohl Klimafaktoren wie die Niederschlagsmenge als auch lagebezogene Kriterien wie die vorhandene Vegetation mit ein.

### 2.1.4 Sprachen im Wandel

Um die sprachliche Diversität zu verstehen, ist es jedoch nicht ausreichend, nur ihre Verteilung zu analysieren. Da weder Sprachen selbst noch ihre räumliche Lokalisa-

tion statisch sind, ist ein Blick auf die ihrer Ausbreitung und Veränderung zugrundeliegenden Mechanismen unerlässlich (Gavin et al. 2013).
Gavin et al. (2013) unterscheiden zwischen vier verschiedenen Prozessen der Entstehung neuer Sprachen: Diese sind die neutrale Veränderung, die Bewegung, der Kontakt sowie die Selektion. Die neutrale Veränderung bezieht sich auf die Abspaltung respektive Isolierung einzelner Gruppen zur restlichen Gesellschaft, die mit der Entstehung einer neuen Sprache einhergeht. Ein Beispiel hierfür sind Slangsprachen. Bei der Bewegung handelt es sich um die Bildung neuer Sprachen verbunden mit der Ausbreitung von Gruppen in bisher unbewohnte Gebiet, so geschehen im Zuge der Besiedlung von Hawaii durch Polynesier vor ungefähr 1000 Jahren. Beim Kontakt verändern in ein bevölkertes Gebiet immigrierende Menschen die vorhandene Sprache. Die Kreolsprachen stellen ein Beispiel dieses Typs dar. Sprachselektion bezeichnet die Wahl von Sprachen aufgrund sozioökonomischer Bedingungen. Dabei wird eine Sprache aufgegeben und eine andere gewählt, welche gesellschaftlich und wirtschaftlich mehr Erfolg verspricht. Beispiele hierfür sind indianische Sprachen in den USA, bei welchen der Generationenwechsel als Folge der kleinräumigen Verbreitung nicht mehr funktioniert. Ein analoges Beispiel aus der Schweiz sind die rätoromanischen Sprachen.

### 2.1.5 Verschiedene Ausbreitungsarten von Sprachfamilien

Da das Thema dieser Arbeit die Ausbreitungspfade von Sprachfamilien ist, werden hier die beiden Typen räumlicher Expansionen (Bewegung und Kontakt) noch genauer erläutert. Diese unterschiedlichen Ausbreitungsmechanismen von Sprachen sind auch vermehrt Gegenstand neuerer linguistischer Studien (u.a. Currie & Mace 2013, de Filippo et al. 2012). Wie schon in der Einleitung erwähnt wurde, bedient sich die sprachwissenschaftliche Forschung dabei zunehmend moderner biologischer Methoden: Mit Hilfe von DNA-Analysen kann untersucht werden, ob die Ausbreitung einer Sprache über eine tatsächliche Bewegung von Menschen erfolgte oder ob es sich in erster Linie um eine „Sprachbewegung“ infolge kulturellen Austauschs ohne oder mit geringer Genveränderung handelt. Geringere genetische Unterschiede zwischen verschiedenen Sprachen aus derselben Sprachfamilie sind ein Indikator für einen demographischen Austausch. Bei der „reinen Sprachbewegung“ sind hingegen die genetischen Distanzen zwischen allen Sprachen unabhängig von ihrer Familie gleich gross.

## 2.2 Untersuchungsgebiet

### 2.2.1 Überblick über die Bantusprachfamilie

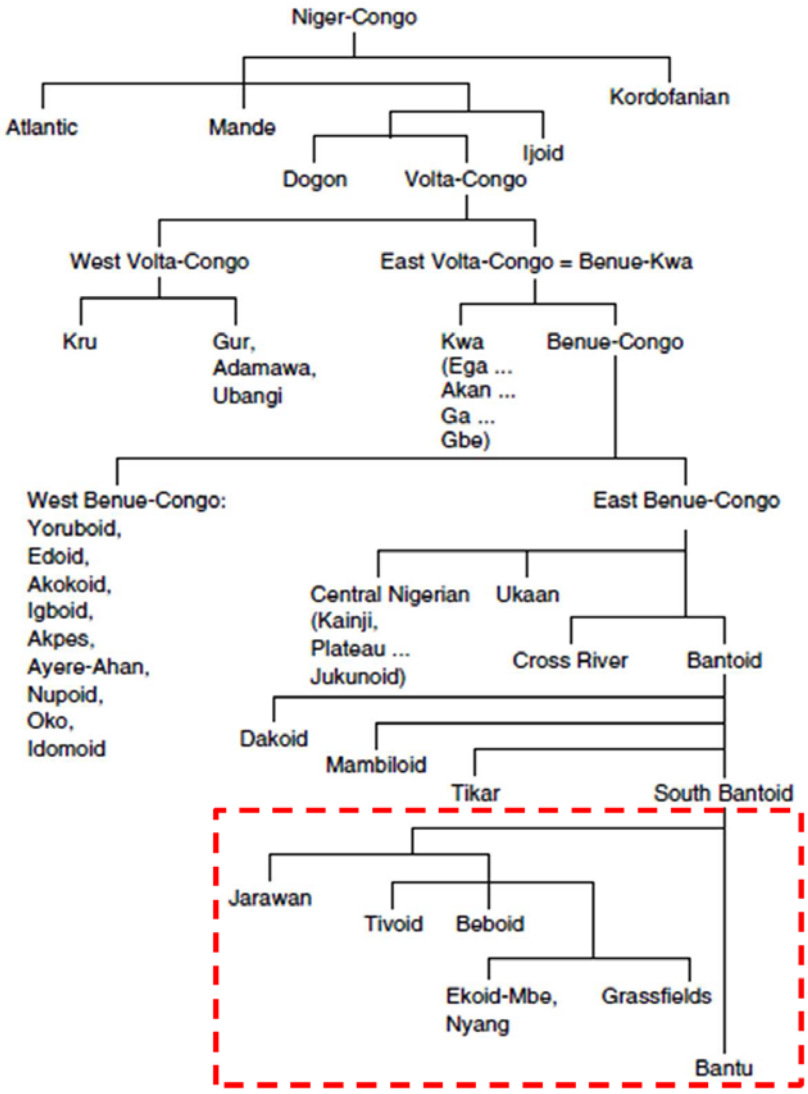



*Abbildung 2: Stammbaum der Niger-Kongo Über-Sprachfamilie bis zur Bantu-Familie; aus Schadeberg (2003), angepasst aus Williamson & Blench (2000)*

Bei der Bantu-Familie handelt es sich um eine sehr grosse afrikanische Sprachfamilie, welche je nach Quelle 400 (McGregor 2009) oder sogar mehr als 500 Sprachen (Currie et al. 2013) umfasst. Die Anzahl der Sprecher liegt dabei etwa bei 240 Millionen (De Filippo 2012). Mit einem Alter von circa 3000-5000 Jahren zählt die Gruppe zu den eher jüngeren Sprachfamilien (McGregor 2009). Ihr Ursprungsgebiet befindet sich in der Grenzregion zwischen Nigeria und Kamerun, von wo sie sich Richtung Süden ausbreitete. (Currie et al. 2013). Der Ausgangspunkt kann dadurch bestimmt werden, dass an diesem Ort eine Vielzahl dem Bantustamm zugehöriger Sprachen auf kleinem Raum zu finden ist. Da sich bei einer Expansion die einzelnen Äste verzweigen und nicht parallel vorliegen, lassen diese Cluster auf den Ursprung schliessen. Entsprechend basiert auch die Evidenz der Bantu-Ausbreitung primär auf linguistischen Untersuchungen (Vansina 1995). Heute erstreckt sich ihre Ausbreitung von einem Grossteil des äquatorialen Regenwalds bis ganz in den Süden, was einem Territorium von ungefähr 9 Millionen km² entspricht (De Filippo 2012). Die Bantu-Familie ist somit sowohl bezüglich der Anzahl Sprecher als auch der geographischen Ausdehnung bei weitem die grösste afrikanische Sprachfamilie, auch wenn sie von Linguisten bloss als Ableger der noch viel grösseren Niger-Kongo Sprachfamilie angesehen wird (Bostoen et al. 2013). Die Niger-Kongo Sprachfamilie ist mit beinahe 1500 Sprachen die grösste Sprachfamilie der Welt und erstreckt sich von Senegal

bis zum Kap der guten Hoffnung. Ihr Status als sprachliche Einheit wird jedoch von einigen Linguisten angezweifelt (McGregor 2009). Dies ist dadurch begründet, dass sie noch nicht völlig rekonstruiert werden konnte. Aufgrund von Ähnlichkeiten zwischen den Sprachen ist es jedoch wahrscheinlich, dass es sich dabei um eine Familie handelt (Nichols & Bickel 2009). Der jüngste gemeinsame Vorfahre der Niger-Kongo Sprachfamilie wird auf einen Zeitraum vor 10'000 bis 12'000 Jahren datiert, dieses hohe Alter ist auch ein Grund dafür, dass besonders bezüglich ihrer genauen Zusammensetzung noch keine absolute Klarheit herrscht (Bostoen et al. 2013). Wie aus Abbildung 1) zu entnehmen ist, stellt die Niger-Kongo die Sprachfamilie auf dem höchsten Level dar, sie spaltete sich in verschiedene Subfamilien, wobei die engere Bantu-Familie eine Untergruppe der Bantoid-Familie bildet (Nordhoff et al. 2013). Insgesamt werden der Niger-Kongo Sprachfamilie inklusive der Bantu-Familie 177 Untergruppen zugeordnet (Diamond and Bellwood 2003).

### 2.2.2 Rekonstruierbarkeit der Umweltbedingungen

Eine Grundvoraussetzung für geographische Untersuchungen stellt die Verfügbarkeit von gültigen Hintergrunddaten dar. Gerade bei geographischen Untersuchungen im Bereich der Archäologie, welche sich oftmals auf weit zurückliegende Zeiträume beziehen, herrscht immer ein gewisser Unsicherheitsfaktor bezüglich der damals herrschenden Verhältnisse vor. Während die Topographie sich innerhalb eines (geologisch) kurzen Zeitraums von wenigen Jahrtausenden kaum oder nur geringfügig verändert und die Verwendung rezenter Daten somit unproblematisch ist, unterliegt die Vegetation verhältnismässig stark kurzfristigen Schwankungen. Neben variierenden klimatischen Verhältnissen wird sie besonders in jüngerer Zeit auch zunehmend durch den Menschen geprägt (Herzog 2013).
Deshalb wurde im Vorfeld dieser Untersuchung mittels literarischer Quellen abgeklärt, welches Klima während der Ausbreitung der Bantu, also 5000-3000 Jahre BP, im zentralen bis südlichen Afrika herrschte.

Literarische Recherchen zum historischen Klima Afrikas zeigen, dass dieses während des Holozäns noch deutlich feuchter war. Besonders in den subtropischen Gebieten war die Niederschlagsmenge um einiges grösser als heute. Um 6000-5500 Jahre BP regnete es dort um 200-400m mehr, dies bei einem heutigen jährlichen Gesamtniederschlag von nur gerade etwa 50mm (Frenzel et al. 1992). Wie die Wasserpegel der afrikanischen Seen zeigen, wurde das Klima ab etwa 5000 Jahre BP deutlich arider. Um 3800 BP war ein erstes Trockenheitsmaximum, danach gab es bis etwa 3000 BP teilweise etwas feuchtere Perioden, worauf sich wieder der Trend Richtung Aridität fortsetzte (Petit-Maire and Guo 1998). Schwartz (1992) schreibt, dass sich der Regenwald in den letzten Jahrtausenden, respektive während der Ausbreitung der Bantusprachen, nicht signifikant verändert hatte. Es kann also insgesamt von ähnlichen klimatischen und vegetativen Verhältnissen ausgegangen werden, wie sie auch im heutigen Afrika vorzufinden sind.
Das feuchte tropische Klima wurde zeitweise jedoch von arideren Phasen unterbrochen. Wie mehrere Studien (Bostoen 2012) belegen, führten klimatische Veränderungen um etwa 2500 BP zu Schäden der zentral-afrikanischen Wälder in den Regionen Südkamerun, südliches Zentralafrika, Gabon und im Kongo. Sie haben im Mit-

tel bis zu 500 Jahre angedauert und äusserten sich vor allem durch eine stärker ausgeprägte Saisonalität. Die durch Pollenanalysen aus Kamerun sowie Zentralafrika bestätigten ariden Phasen führten zu einem starken partiellen Waldrückgang, welcher „offene Korridore“ im Regenwald schuf, welche der Ausbreitung der Bantu möglicherweise entgegenkamen (Elenga et al. 1992, van Geel et al. 1996). Gleichzeitig ermöglichten die Fortschritte in der Metallurgie während dieser Zeit ein schnelleres Durchqueren des Dschungels (Schwartz 1992). Eine Untersuchung umgangssprachlicher Namen typischer Pionier-Pflanzen der Bantu-Völker (Bostoen et al. 2013) zeigte ebenfalls geographische Verteilungsmuster, welche auf eine gewisse Schrumpfung des Regenwalds in der Mitte des ersten Jahrtausends BC hindeuten.

Literaturrecherchen zur Vegetation Mittel- bis Südafrikas vor 3000-5000 Jahren zeigen, dass diese der heutigen im Wesentlichen sehr ähnlich war. Wie erwähnt dürften sich jedoch zeitweilige Schneisen im Regenwald eröffnet haben, welche die Bantu-Ausbreitung womöglich erleichterten. Inwieweit die Bantu-Sprecher bei ihrer Expansion aber von möglichen kleinräumigen Unterschieden profitierten, ist unklar. Auch die Zerstörung ursprünglicher Vegetation durch den Menschen ist natürlich schwierig einzuschätzen. Die durch die Bevölkerungsexplosion verursachte Zerstörung des Regenwalds stellt ein grosses Problem Afrikas dar. Da zu dieser nur ungenügend längerfristige Bestandesaufnahmen vorliegen, liegt auch hierin ein Unsicherheitsfaktor (Adams et al. 1996). Um den Einfluss des Menschen oder natürliche bedingte Veränderungen am besten zu erfassen, empfiehlt es sich, in einem späteren Schritt die modernen Daten zu variieren. Auf diese Weise können mögliche Unsicherheiten bezüglich der geographischen Daten sinnvoller einkalkuliert werden, als mit aufwendiger, aber dennoch fehleranfälliger Bearbeitung beziehungsweise Rekonstruktion der Datengrundlage (Herzog 2013).

## 2.3 Forschungsstand

### 2.3.1 Grundsätzliches zur Bantu-Ausbreitung

Die Ausbreitung der Bantu-Sprachfamilie ist eines der am meisten debattierten Themen der Geschichte der afrikanischen Bevölkerung. Die Expansion über grosse Teile des südafrikanischen Kontinents zog Forscher aus unterschiedlichen wissenschaftlichen Disziplinen gleichermassen an. Dazu zählen unter anderem Archäologen, Anthropologen sowie auch Palynologen (Bostoen 2013). Es gibt verschiedene Gründe, warum die Bantu-Ausbreitung einen speziell interessanten Fall einer Völkerwanderung darstellt. Zum einen handelt es sich dabei um eine der weltweit grössten Migrationen, zusätzlich erfolgte sie auch über eine im Verhältnis zur zurückgelegten Distanz relativ kurze Zeitdauer. So werden die ältesten mit den Bantu verbundenen Keramiken aus Südafrika etwa auf die Mitte des ersten Jahrtausends BC datiert, was auf eine Besiedlung des ganzen südlichen Afrikas innerhalb von nur etwa 2500 Jahren schliessen lässt (Plaza et al. 2004). Meistens werden die aufkommende Metallurgie sowie erste landwirtschaftliche Praktiken als Motoren der Ausbreitung angesehen. Tatsächlich dürften die Einführung neuer Technologien und eine komplexe „Waldpflanzen-Landwirtschaft“ wichtige Gründe für die erfolgreiche Ausdehnung der Bantu gewesen sein. Belege für diese Innovationen konnten Sprachwissenschaftler

(Bostoen et al. 2013) in damit verbundenen sprachlichen Transformationen, insbesondere des Wortschatzes, finden. Auch Holden (2001) zeigt in ihrer Studie auf, wie sich die Expansion der Landwirtschaft in den Bantu-Sprachen spiegelt.

Inwieweit der Wechsel von der Jäger-Sammler-Gesellschaft zur Landwirtschaft als Treiber des Bevölkerungswachstums die Erschliessung neuen Landes notwendig machte und somit die Migration auslöste, kann nur erahnt werden (Vansina 1995). Dabei muss auch erwähnt werden, dass sich das Verständnis der Bantu-Expansion im Laufe der letzten Jahrzehnte deutlich geändert hat. Während man früher von einer grossen, plötzlichen Migrationswelle ausging, wird die Bantu-Migration heute eher als eine Folge mehrerer Vorstösse angesehen.

### 2.3.2 Linguistische Kontroverse zur Bantu-Ausbreitung

Neben den geschichtlichen und ethnologischen Gesichtspunkten stellt die Bantu-Ausbreitung auch ein interessantes Studienobjekt in der Linguistik dar. Diese nimmt im Zusammenhang mit der Expansion insofern auch eine besonders wichtige Rolle ein, als dass sie den eindeutigen Nachweis dafür lieferte (Vansina 1995).

Zwar sind sich die Linguisten schon seit längerem bezüglich des Ursprungsgebietes der Sprachfamilie zwischen Nigeria und Kamerun einig, der konkrete Verlauf der Ausbreitung wurde aber bis in die jüngste Zeit kontrovers diskutiert. Dabei wurden – basierend auf archäologischen sowie linguistischen Daten – stets zwei verschiedene Szenarien vertreten (De Filippo et al. 2012).

Gemäss dem ersten und populäreren Modell breitete sich die Bantubevölkerung nach einer frühzeitigen Aufteilung in zwei verschiedene Richtungen aus (Holden 2002, Newman 1995). Ein Pfad führte dabei zuerst nach Osten um die dichte Regenwaldzone herum zu den grossen afrikanischen Seen und verlief dann weiter nach Süden. Die andere Gruppe durchquerte – direkt nach Süden vorstossend - den Regenwald und gelangte dann in die arideren Steppen Südwestafrikas. Gemäss dieser „Early-Split“-Theorie entstand so ein Ost- sowie ein West-Bantu Sprachzweig (Abb. 3a).

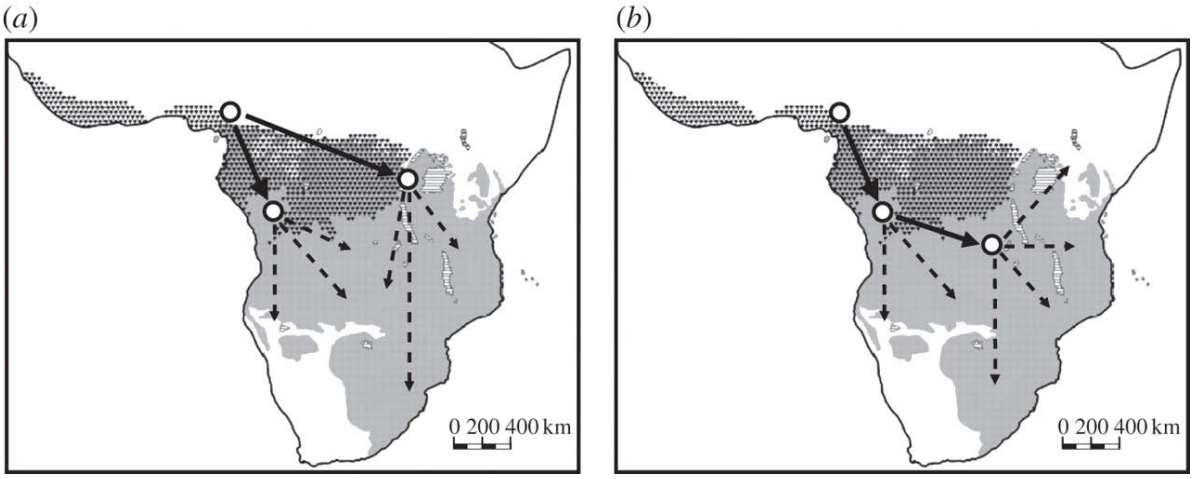



*Abbildung 3: Die beiden Ausbreitungsszenarien der Bantu-Sprachen: a) Early-Split Modell und b) Late-Split Modell. Die graue Fläche stellt das heutige Ausbreitungsgebiet der Bantu-Sprachen, die schwarz gerasterte den tropischen Regenwald dar; Quelle: De Filippo et al. (2012)*

Bei der alternativen Hypothese des „Late-Splits“ (Abb. 3b) wurde hingegen davon ausgegangen, dass die Ausbreitung anfangs gemeinsam durch den Regenwald verlief und sich erst im subäquatorialen Afrika verzweigte (Ehret 2001). Entsprechend ging man in diesem Modell davon aus, dass die verschiedenen Bantu-Sprachen südlich und östlich des Regenwalds derselben Gruppe zugeordnet werden können.

### 2.3.3 Neue Einsichten dank moderner Methoden

Dank aufwendiger neuer Studien konnte in jüngster Zeit das Modell des Late-Splits gefestigt werden. De Filippo et al. (2012) untersuchten 92 Wörter von 412 Bantu-Sprachen und korrelierten linguistische Distanzen mit für die jeweiligen Split-Hypothesen berechneten geographischen Distanzen. Dabei zeigte sich eine höhere Korrelation bei den aus dem Late-Split-Modell berechneten geographischen Distanzen. Des Weiteren war der Zusammenhang auch zwischen ebenfalls untersuchten genetischen Distanzen und den linguistischen Werten für das Late-Split-Modell grösser. Montano et al. (2011) bekräftigten – ebenfalls mit einer Untersuchung der Variation der Y-Chromosomen – auch die Late-Split-Hypothese. Ihre Studie, welche Nigeria, Kamerun, Gabon sowie den Kongo umfasste, zeigte deutliche Unterschiede innerhalb der lokalen demographischen Entwicklung. Alves et al. (2011) analysierten in ihrer Studie neben den häufig verwendeten Y-Chromosomen und der mitochondrialen DNA Variation sogar noch weitere Erbgut-Informationen. Neben einem starken genetischen Zusammenhang innerhalb der Bantu-Stämme konnten sie einen regen Austausch zwischen Savannen-Völkern südlich des Regenwalds nachweisen. Die genetische Homogenität innerhalb der Bantu-Bevölkerung lässt sich ihrer Meinung nach besser durch eine kontinuierliche Entfaltung miteinander verbundener Populationen erklären als durch eindeutig abgetrennte eigenständige Zweige.
Currie et al. (2013) untersuchten einen umfassenden Sprachdatensatz von 90 Wörtern aus 542 Bantu-Dialekten. Diese überprüften sie mittels charakterbasierter Bayesscher Wahrscheinlichkeit auf Verwandtschaft. Der Vergleich der sprachlichen Verwandtschaft mit den heutigen Standorten dieser Sprachen unterstützte ebenfalls die Late-Split-Hypothese.
Wie die genetischen Forschungen von De Filippo et al. (2012) zeigten, fand die Bantu-Ausbreitung zudem vermehrt über einen demographischen Austausch statt.

## 2.4 Forschungslücken

### 2.4.1 Limitierung der bisherigen Ansätze

Wie aus dem vorherigen Kapitel hervorgeht, existiert eine Vielzahl an Studien zur Bantu-Ausbreitung. Neue Studien verfolgten dabei zunehmend multi-disziplinäre Ansätze, bei welchen die linguistische Datengrundlage mittels Auswertungen aus weiteren Forschungsbereichen bekräftigt wurde. Einen sehr wichtigen Beitrag leisteten diesbezüglich die modernen Methoden der Biologie. Sie ermöglichten es, die linguistisch ermittelte Datenlage abzusichern (De Filippo et al. 2012, Montano et al. 2011, Alves et al. 2011). Des Weiteren wurden in einzelnen Studien auch geographische Kriterien mit einbezogen. Bei De Filippo et al. (2012) wurden die für die beiden Split-Hypothesen resultierenden unterschiedlichen euklidischen Distanzen zwischen den Sprachen als weiteres Untersuchungsmerkmal herangezogen.
Mit den direkten euklidischen Distanzen wird zwar die Geographie berücksichtigt. Da jedoch Fusspfade (gerade bei einer derart dichten Vegetation) kaum jemals der Luftlinie entsprechen, sind diese direkten Distanzen deshalb nur wenig aussagekräftig. Die möglichen Wege können wesentlich besser mit der Methode der geringsten Kostenpfade dargestellt werden.

### 2.4.2 Least Cost Path Anwendung bei einer Völkerwanderung

Wie bereits im Einleitungskapitel erläutert wurde, wird mit der LCPA ein Pfad berechnet, der sich in der Realität oft erst nach einem schrittweisen, unter Umständen langwierigen Prozess der Ergonomisierung herauskristallisiert. Entsprechend setzt das Bestimmen des kürzesten Kostenpfades Ortskenntnisse voraus, beispielsweise Touristen wählen bei der Ersterkundung einer Ferieninsel kaum den optimalen Weg (Herzog 2013). Auch sind die empirischen Kenntnisse eines Ortes kaum jemals so genau, dass der auf „natürliche“ Weise bestimmte Weg absolut identisch mit dem berechneten ist. Fest steht nur, dass schon die prähistorische Bevölkerung ihre Raumüberwindungskosten zumindest teilweise optimierte (Herzog & Posluschny 2008). Bei vielen Werkzeugen zur Berechnung des LCP werden aus diesem Grund mit einem Zufallsfaktor variierte Varianten zur Verfügung gestellt, um die möglichen Abweichungen vom optimalen Weg zu berücksichtigen (Van Etten 2013).

Bei Herzog (2013) wird ein Massenexodus – da etwas einmaliges – als ein für die LCPA ungeeignetes Untersuchungsbeispiel aufgeführt. Es wird für diesen Fall stattdessen die Anwendung eines Agentenbasierten Modells empfohlen. Dem kann im Falle der Bantu-Ausbreitung entgegen gehalten werden, dass diese keinesfalls in einem Zug erfolgte, sondern vermutlich in vielen einzelnen Schüben über einen langen Zeitraum von statten ging (Vansina 1995). Dies macht gewisse Routenoptimierungen durchaus plausibel. Dazu kommt, dass ein Agentenbasiertes Modell genaue Kenntnisse aller relevanter Parameter beziehungsweise Einheiten voraussetzen würde, was in diesem Fall grundsätzlich nicht möglich ist. Aus diesem Grund stellt die LCPA unter Umständen doch eine sinnvolle Möglichkeit dar, Bewegungen dieser Art zu erfassen.

# 3 Daten

## 3.1 Sprachdaten

Bei der Erstellung dieser Arbeit wurden Sprachdaten aus drei verschiedenen Quellen verwendet: Dies sind die AUTOTYP-Datenbank (Nichols & Bickel 2009), die Hammarström-Datenbank (Hammarström 2012) sowie die Sprachdaten aus der Glottolog-Seite (Nordhoff et al. 2013).

Für die zentralen Berechnungen und Analysen der Arbeit wurde ausschliesslich die AUTOTYP-Datenbank verwendet. Bei dieser sind die Daten einerseits sehr gut beschrieben, zudem liegen sie in bereits geordneter Struktur vor. Die Glottolog-Daten wurden später aufgrund der bereits vorgefertigten genealogischen Baumstruktur als Vergleichsbasis herangezogen. Die Hammarström-Daten wurden nur ganz am Rande benutzt, respektive nur teilweise als Vergleiche zur AUTOTYP-Datenbank konsultiert.

### 3.1.1 AUTOTYP-DATENBANK

#### Hintergrund

Die AUTOTYP-Datenbank umfasst Daten zu 2922 verschiedenen Sprachen. Neben dem zugehörigen ISO-639 Code enthält die Datenbank genaue genealogische Informationen zu den einzelnen Sprachen. Diese wurden für alle der 2922 Punkte nach demselben Schema vergeben, was eine Vergleichbarkeit ermöglicht. Die unterste Ebene stellen einzelne Sprachen oder auch individuelle Dialekte dar. Die höchste Sprachfamilien-Ebene ist unter *stock* („Stamm“ oder auch „Phylum“) aufgeführt. Ein Stamm bildet die höchste Einheit, bei welcher eindeutig bewiesen werden kann, dass es sich dabei um eine genealogische Einheit handelt. Zudem muss er zumindest theoretisch noch eindeutig rekonstruierbar sein. In der sprachlichen Hierarchie absteigend sind in der Autotyp-Datenbank nach dem Stamm zudem *major branch* (mbranch), *subbranch* (sbranch)*, sub-subbranch* (ssbranch) *sowie lowest subbranch* (lsbranch) angegeben. Zusätzlich werden noch die Sprachfamilie sowie die Gattung (Genus) gemäss WALS (2011) angegeben.

Neben den genealogischen Daten wird zudem bei einem Teil der Sprachen (712 von 2922) noch vermerkt, ob es sich bei den Sprechern um eine Jäger-Sammler Gesellschaft handelt. Weiter wird noch bei einigen (32 von 2922) angegeben, ob diese aus mehreren Sprachen abgeleitete Kreolsprachen darstellen.

Die geographischen Koordinatenangaben (Breitengrade und Längengrade) beziehen sich auf das ungefähre linguistische Zentrum einer Sprache. Im Falle, dass zwei Sprachen dasselbe Zentrum haben, wurden die Punkte leicht verschoben, um eine vernünftige Kartendarstellung zu ermöglichen.
Wie in der Linguistik üblich, gibt die Aufnahme der verschiedenen Sprachen die Situation vor der Kolonialisierung und der dadurch bedingten Vermischung wieder. Sprachen, welche aus der Fusion mehrerer Sprachen entstanden, wie beispielsweise Englisch, sind daher nicht darin enthalten (Nichols & Bickel 2009).

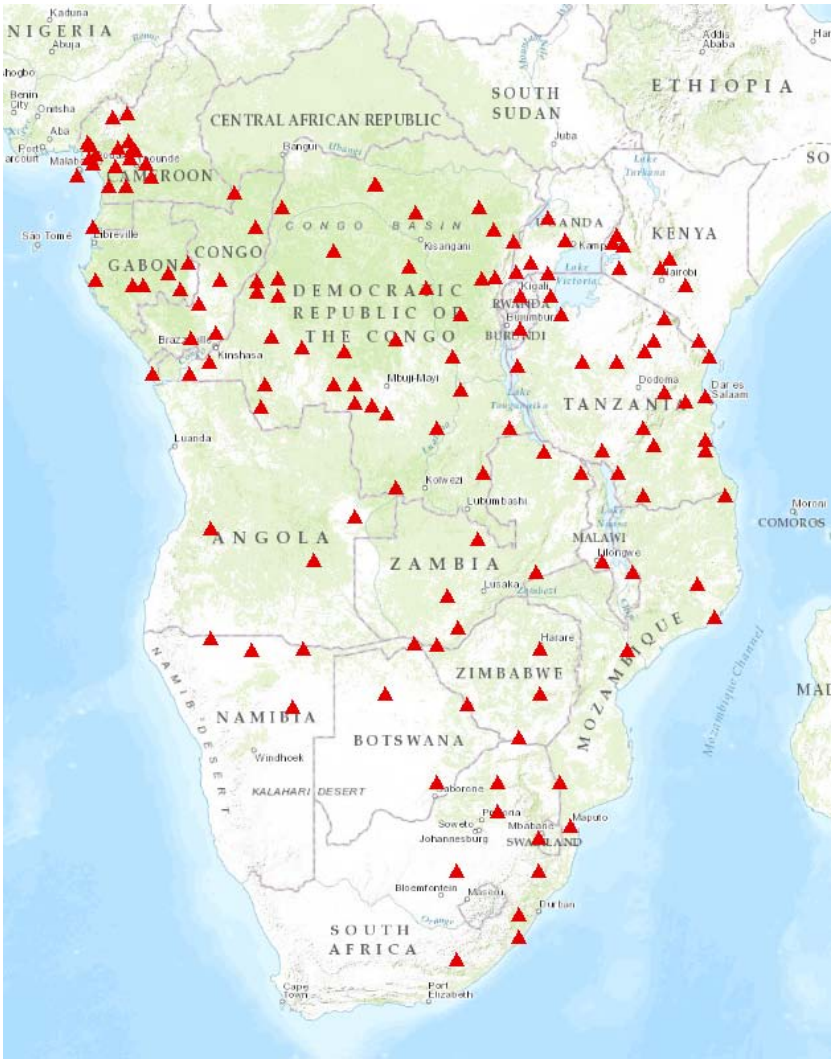



*Abbildung 4: Topographische Karte des südlichen Afrikas mit den 140 Bantu-Sprachpunkten (schwarz) aus dem Nichols &Bickel Datensatz (2009); Eigene Darstellung*

Die Daten der AUTOTYP-Datenbank liegen als Tabelle sowohl im *csv* als auch im *tab*-Format vor. Da die Daten bereits strukturell geordnet sind, konnten sie direkt in R sowie nach geringfügigen Änderungen (Umwandlung in Excel-Spreadsheet, Ersetzen von Kommas durch Punkte) auch in *ArcGIS* eingelesen und mittels der enthaltenen Koordinaten kartographisch dargestellt werden.

Bei der Übertragung aller Punkte in *ArcGIS* zur kartographischen Darstellung mussten einige Punkte ausgefällt werden. Insgesamt sieben Sprachen hatten fehlerhafte oder keine Koordinatenangaben, zudem fielen beim Verschnitt mit der Weltkarte weitere 145 Punkte ins Meer. Von den Anfangs 2922 Sprachen blieben folglich noch 2770 übrig. Da in dieser Arbeit aber nur die Bantusprachen relevant sind, wird hier bloss auf diese genauer eingegangen. Insgesamt sind 141 Sprachen der Bantu-Familie im Datensatz enthalten. Die Verschneidung mit der Festlandkarte zeigt, dass ein bei den Komoren gelegener Punkt (Sprache: Comorian, ISO: swb]) im Meer liegt. Ein weiterer Punkt (Sprache: Bubi, ISO: bvb) auf der zu Äquatorialguinea gehörenden Insel Bioko kann gerade ausgeschieden werden, da diese LCP-Berechnung das Meer nicht einbezieht. Ein weiterer Punkt (Sprache: Benga, ISO: bng) bei einer ins Meer hereinragenden Küste Äquatorialguineas wurde entfernt, da er aufgrund seiner von den restlichen Punkten abgeschotteten Lage für die LCPA nicht relevant war. Insgesamt konnten aus den Nichols-Bickel-Daten also 138 Punkte der *Narrow-Bantu*- Sprachfamilie für die LCP-Untersuchung verwendet werden.

Bei der höchsten Ebene, dem Stamm *(stock)* ist bei der Bantu-Familie die Benue-Kongo Sprachfamilie eingetragen. Diese bildet einen Südzweig des grössten und komplexesten Primärzweig der Niger-Kongo Familie, der Volta-Kongo Familie (Mc Gregor). Da diese Übergruppen jedoch die Bedingung der Rekonstruierbarkeit noch

nicht genügend erfüllen, stellt die Benue-Kongo in dieser Datenbank das höchste Level dar. Der *major branch* bildet die Bantoid-Familie als Übergruppe der Bantu-Gruppe, welche ein *subbranch* davon ist (McGregor 2009).

### 3.1.2 Glottolog-Sprachdaten

Bei Glottolog handelt es sich um eine frei zugängliche Online Sprachen-Datenbank. Sie enthält einen umfassenden Katalog der weltweiten Sprachen, Sprachfamilien sowie Dialekten. Jedem einzelnen von diesen ist zudem ein spezifischer, stabiler „Glottocode“ zur einfachen Identifikation zugeschrieben. Der ISO 639-3 Code ist jeweils ebenfalls angegeben. Die Spracheinheiten sind anhand ihrer verwandtschaftlichen Klassifizierung eingeordnet, welche auf verfügbaren historischen Daten und Sprachvergleichsanalysen basiert. Desweiteren können zu den Sprachen bibliographische Informationen abgerufen werden.
Was die Glottolog-Datenbank für diese Arbeit besonders wertvoll macht, ist dass die Sprachen bereits in vorgefertigter Baumstruktur vorliegen. Diese können in verschiedenen Formaten im Browser dargestellt sowie auch heruntergeladen werden (Nordhoff et al. 2013)

Speziell an der Glottolog-Datenbank ist ihre enorm hohe sprachliche Auflösung. So sind beispielsweise für die Bantu-Sprachfamilie 3250 verschiedene phylogenetische Zweige darin enthalten. Sie beinhaltet somit auch viele Dialekte, welche keine eingenständigen Sprachen darstellen.

## 3.2 Geographische Daten

Es wurden in dieser Arbeit geographische Daten aus folgenden Quellen verwendet.

### 3.2.1 Geländeoberfläche

Das bei der Arbeit verwendete Höhenmodell stammt aus der *NASA Shuttle Radar Topographic Mission* (SRTM). Dieses kann auf der Seite des CSI (CGIAR-CSI 2008) im Rasterformat mit unterschiedlichen Auflösungen herunter geladen werden (http://srtm.csi.cgiar.org/). Die gröbste Auflösung der zur Verfügung gestellten Daten beträgt eine halbe Winkelminute (0.92 km * 0.92 km am Äquator). Weiter sind DEM mit Viertel- sowie Achtel-Winkelminuten-Auflösung vorhanden.

### 3.2.2 Landbedeckungsdaten

Da die erwähnten Korridore, welche eventuell zur Zeit der Bantu-Ausbreitung im Regenwald vorhanden waren, besonders in einer derart dynamischen Flora unmöglich über einen solchen Zeitraum rekonstruiert werden können, wurden für die LCPA rezente Karten bezüglich Landnutzung- und Bedeckung verwendet. Die Daten zur Vegetation stammen von den frei zugänglichen Seiten (http//:geodata.grid.unep.ch) der UNEP (2000). Die meisten der Daten wurden mittels *Landsat Multispectral Scanner System* (MSS) aufgenommen, zudem wurde bei einem Teil noch der *Landsat Thematic Mapper* verwendet. Die Aufnahmen erfolgten während der Jahre 1981 bis 1994. Um die jüngsten Veränderungen der Vegetation durch den Menschen möglichst zu vermeiden, fiel die Wahl bewusst auf den ältesten verfügbaren Datensatz.
Im Unterschied zu den digitalen Höhenmodellen sind Vegetationsdaten nur bis eine Auflösung von 30 Bogensekunden im Rasterformat erhältlich. Da die Einteilung in einzelne Vegetationsklassen sowieso schon eine Generalisierung darstellt, wäre eine genauere Auflösung jedoch auch nicht sinnvoll beziehungsweise würde keinen weite-

ren Informationsgewinn bringen. Des Weiteren können von der FAO (2007) Seite unter anderem auch Daten zum Bodentyp heruntergeladen werden.

### 3.2.3 Gewässer

Die Gewässerdaten stammen von ESRI (2010). Dabei handelt es sich um Line-Shapefiles. Da über die genauen Verläufe kleinerer Flüsse, ja gar deren Existenz in diesen Zeitraum keine Kenntnisse vorhanden sind, werden bei dieser Analyse nur die grossen Ströme miteinbezogen. Diese Wahl wird zudem dadurch bekräftigt, dass bei den Flüssen auch eine ausreichende Begehbarkeit bezüglich der Wassermenge sowie der oberirdischen Verbundenheit gegeben sein muss. Auch dies kann bei kleineren Flüssen beziehungsweise Bächen nicht klar bestimmt werden, einerseits aufgrund des bereits erwähnten zeitlichen Faktors, zum anderen können zu diesen Gewässern auch heute keine genauen Daten gefunden werden.

## 3.3 Verwendete Software

Die LCP-Analyse wurde mittels *R-Statistics* (R Development Core Team 2013) durchgeführt, einer freien Scripting-Sprache und Umgebung für statistische Berechnungen und Grafiken. Dabei wurde die speziell dazu entwickelte R-Einheit *gdistance* (Van Etten 2012) zur Erstellung der LCPA verwendet. Ein weiteres mehrfach verwendetes R-Package stellt *ape* dar, mit dem phylogenetische Bäume erstellt werden können. Für Visualisierungsarbeiten wurde vorwiegend *ArcGIS* benutzt. Dieses kam unter anderem auch für die weitere Analyse der mittels *R* berechneten Kostenpfade zum Einsatz. Desweiteren diente die frei erhältliche Software *Hydro Flow* (UERJ 2007) zur Bestimmung von Flussordnungen. Zur Erstellung von Matrizen aus berechneten Distanzkosten wurde neben *R* noch *Microsoft Access* benutzt.

# 4 Methodik

In diesem Teil der Arbeit wird zuerst die genaue Vorgehensweise bei der Erstellung der LCPA sowie die ihr zugrunde liegende Theorie erklärt. Der zweite Teil bildet die Durchführung der LCPA. Dabei werden verschiedene Parameter variiert und die Methode auf ihre Stabilität überprüft. In einem nächsten Schritt werden die erhaltenen Kostenpfade analysiert. Um eine Vergleichbarkeit mit linguistischen Daten zu ermöglichen, werden sie in Dendrogramme überführt.

## 4.1 Erstellung der LCPA

### 4.1.1 Ablauf der LCPA

Zur besseren Übersicht werden hier die einzelnen Schritte der LCPA wiedergegeben. Das Erstellen einer Multi-Kriterien Kostenoberfläche und die Berechnung des Least Cost Path können inklusive leichter Modifikationen gemäss Howey (2007) folgendermassen unterteilt werden (Abb. 5).

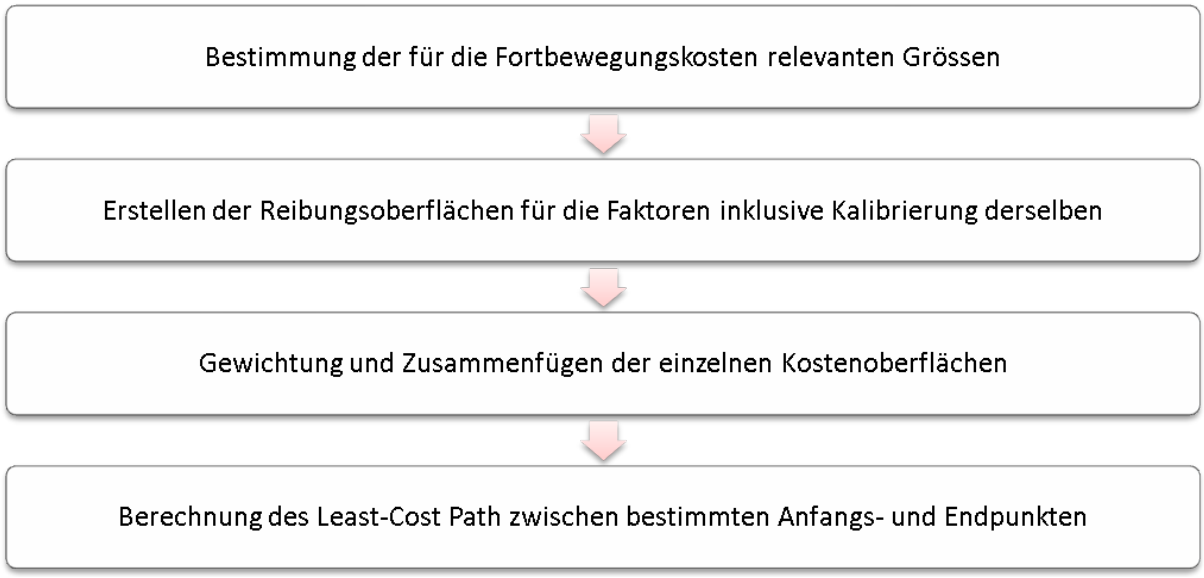



*Abbildung 5: Einzelne Schritte bei der Erstellung der LCPA; angepasst aus Howey (2007)*

Generell bilden bei der Durchführung der LCP-Analyse die Kalibrierung und Gewichtung die fehleranfälligsten Punkte. Diese hängen davon ab, wie gut die Realität im Modell abgebildet werden kann. Wichtig sind hierbei die Qualität und Auflösung der vorhandenen Daten sowie das Bestimmen respektive Schätzen empirischer Werte. Punkt 1 und 3 betreffen diese bezüglich Genauigkeit und Objektivität heikelsten Prozesse (Berry 2000).

Punkt 2 und 4 hingegen sind eher technischer Natur. Aufgrund der Kenntnisse der verwendeten Algorithmen können hierbei mögliche Fehler relativ genau ermittelt werden. Eine kritische Betrachtung ist aber auch bei diesen Punkten notwendig. Seit die Erstellung von LCPA aufgrund zunehmend besser ausgearbeiteter, umfassenderer Software immer leichter wird, nimmt auch das Problem zu, dass diese ohne die nötigen Kenntnisse der dabei angewandten Algorithmen ausgeführt werden. Zuneh-

mend benutzerfreundlichere Werkzeuge ermöglichen es, eine LCPA zu machen, ohne sich dabei über die Schwächen und Limitierungen der Methode im Klaren zu sein (Herzog 2013). Technische Aspekte wie beispielsweise die verwendete Neigungskostenfunktion, der Einbezug anisotroper Kostenfaktoren sowie die Zahl der berücksichtigten Nachbarn bei der Wegfindung beeinflussen jedoch massgeblich das Resultat der Analyse. Deshalb sollte diesen Punkten auch unbedingt die nötige Aufmerksamkeit geschenkt werden (Herzog & Posluschny 2008).

### 4.1.2 Bestimmen der Kostenfaktoren

#### Übersicht über die in LCP-Studien verwendeten Kriterien

Wie Howey (2007) betont, bedingt die Erstellung einer realistischen Kostenoberfläche den Einbezug mehrerer Kriterien. In vielen Studien (Taliaferro et al. 2010, Rees 2003, Collischonn & Pilar 2000) wird ausschliesslich die Topographie verwendet, andere geographische Faktoren werden nicht einbezogen. Beinahe alle archäologischen LCP-Studien basieren auf der Neigung, oft noch verbunden mit bestimmten anderen Komponenten. Zu diesen zählen die Vegetation, Feuchtgebiete und andere Bodeneigenschaften.
Da sich letztere Faktoren über die Zeit beträchtlich verändern können, ist ihre Berücksichtigung deutlich problematischer als die der Neigung. Mitunter auch ein Grund, dass die diese das am meisten untersuchte Kriterium darstellt (Herzog 2013). Bei der Berechnung der Neigungskosten bietet sich des Weiteren auch der Vorteil, dass eine Vielzahl an Funktionen zu deren Berechnung vorliegt. Der Faktor lässt sich auf ein genau erfassbares, rein technisches Problem reduzieren (Herzog 2010). Für die Bestimmung der mit der Vegetation und der Bodenbedeckung verbundenen Kosten liegt kein bereits vorgefertigtes Gerüst vor. Teilweise finden sich dazu vergleichbare Werte in der Literatur, meist müssen sie jedoch geschätzt werden (Howey 2007).

Bei den Gewässern handelt es sich um einen Faktor, der in der LCPA oft berücksichtigt wird. Dabei werden diese jedoch meist als natürliche Barrieren behandelt, welche es zu umgehen gilt (Atkinson 2005, Field 2005). Die Anzahl an Studien, in welchen Gewässer als Wege betrachtet werden, ist geringer. Dies liegt vermutlich zu einem Teil daran, dass wenige Erfahrungswerte vorliegen, dazu kommt, dass die Berechnung auf diese Weise einiges aufwendiger sein kann. Bei grossen Gewässern sind wechselnde Winde unter Umstände recht kompliziert zu modellieren. Zudem müssen weitere Punkte wie die Bootsform, der Bootsantrieb, der Gewässertyp sowie etwaige Strömungen beachtet werden (Herzog 2013). Beispiele archäologischer Studien, welche Gewässer in die Wegfindung miteinbezogen, finden sich bei Howey (2007) und Livingood (2012).

Weil neben den erwähnten Faktoren nur sehr selten auch noch kulturelle oder soziale Faktoren in die LCPA miteinfliessen, wird diesen manchmal der Vorwurf des Geodeterminismus angelastet (Herzog 2013). Zwar gibt es auch Anwendungen, welche sozio-kulturelle Faktoren mit einbezogen haben. Da sie aber nicht direkt messbar sind, ist ihr Einfluss sehr schwierig abzuschätzen. Ein Beispiel dafür ist Llobera (2000), welcher sowohl auf physischem als auch mentalem Level wirksame Landschaftsmerkmale berücksichtigt. Ein Element auf mentalem Niveau stellt beispielsweise ein Friedhof dar. Dieser kann sowohl abstossend als auch anziehend wirken.

Augrund der sehr schwierigen Erfassbarkeit dürften soziale und kulturelle Kosten jedoch auch in nächster Zeit noch sehr zurückhaltend in der LCPA verwendet werden (Van Leusen 2002).

## Auswahl der Kriterien

Wie aus dem vorherigen Teil hervorgeht, stellt die Auswahl der Kostenkriterien einen gleichermassen wichtigen wie kontrovers diskutierten Aspekt der LCPA dar. Diese sollte möglichst der jeweiligen Situation angepasst werden. Wichtig ist, dass eine Balance gefunden wird zwischen einem zu komplexen und einem zu stark vereinfachenden Modell. Die Kriterien müssen messbar sein sowie einen eindeutigen Einfluss auf die zu erklärende Variable haben. Die Einbeziehung qualitativ schlechter Daten gilt es zu vermeiden. Jedoch muss auch darauf geachtet werden, dass alle relevanten Aspekte berücksichtigt werden (Malczewski 1999).

Im Folgenden wurde versucht, diese Kriterien bei der Untersuchung möglichst genau zu erfüllen. Deshalb wurden für die Zwecke dieser Arbeit in einem ersten Schritt die Faktoren bestimmt, welche möglicherweise einen Einfluss haben. Danach wurden von diesen jene bestimmt, bei denen die klare Datenlage sowie die evidente Bedeutung eine Anwendung begründeten. Im ersten Schritt wurden hernach das Gelände, die Vegetation, die Gewässer sowie der Bodentyp bezüglich einer Verwendung genauer analysiert.

### *Gelände*

Wie bereits erwähnt wurde, stellt die Geländeroberfläche einen bezüglich der untersuchten Zeiträume ziemlich konstanten Faktor dar. Da die Neigung der Standartfaktor aller LCPA schlechthin ist, existiert zudem eine Vielzahl von Algorithmen zur Berechnung der daraus resultierenden Kosten (Herzog 2013). Die Geländeoberfläche wurde auch in dieser Arbeit mit einbezogen. Weil ein Grossteil Afrikas eine Höhe von mehr als 400 m aufweist, wird es oft als Plateau-Kontinent bezeichnet. Jedoch fehlen gebirgige Gebiete weitgehend und auch tiefliegende Gelände sind selten. Verglichen etwa mit der europäischen Topographie ist diejenige von Afrika relativ wenig ausgeprägt (Adams 1996).

### *Vegetation*

Die Vegetation dürfte – bedingt durch die äquatoriale Lage des Untersuchungsgebietes – den entscheidendsten Kostenfaktor darstellen. Der afrikanische Regenwald erstreckt sich von West- über Zentral- bis nach Ostafrika. Im Untersuchungsgebiet haben die Länder Nigeria, Kamerun, die Zentralafrikanische Republik, die Rebublik Kongo sowie die Demokratische Republik Kongo wesentliche Anteile am Regenwald. Aufgrund der Lage des Ursprungsgebietes der Ausbreitung etwas oberhalb des Äquators, mussten die Bantu-Völker bei ihren Wanderungen Richtung Süden den gesamten tropischen Regenwald durchqueren. Angesichts der Nord-Süd-Ausdehnung des Regenwalds von ungefähr 10° (mehr als 1000 km), stellt die Bantu-Völkerwanderung eine erstaunliche Expansion dar (Adams 1996). Die weite Distanz, welche dabei trotz stark eingeschränkter Fortbewegung zurückgelegt wurde, macht auch eine LCP-Anwendung besonders interessant.

Günstig bezüglich der Modellierung der Vegetation sind die relativ tiefen Breitengrade des Untersuchungsgebietes bezüglich der damit verbundenen geringen saisonalen Klimaschwankungen. Aus diesem Grund kann davon abgesehen werden, eine zweite Kostenoberfläche für ein Winterszenario anzuwenden (Howey 2007).

*Gewässer*

Afrika beherbergt eine Reihe namhafter Flüsse und Seen. Neben dem weltweit zweitgrössten Süsswassersee, dem Victoriasee, haben auch mehrere grosse Ströme wie der Kongo, der Ubangi oder Sambesi ihren Ursprung im Untersuchungsgebiet (Adams 1996). Aufgrund der Häufigkeit von Gewässern im Bantu-Ausbreitungsgebiet sollten diese ebenfalls in die LCPA miteinfliessen.
Während ein Einfluss der Gewässer auf die damalige Fortbewegung unbestritten ist, gestaltete es sich hingegen schwieriger, diesen genau zu definieren. Zuerst einmal wurde genauer bestimmt, welche Gewässer gegebenenfalls als Transportwege hätten benutzt werden können. Ein Blick auf die Karte zeigt, dass den Seen bei der Expansion aufgrund ihrer Lage keine Relevanz zugemessen werden kann. Weil bezüglich der Nutzung kleiner Flüsse selbst die heutige Datenlage niemals ausreichend wäre, kamen nur noch die grossen Ströme in Frage.

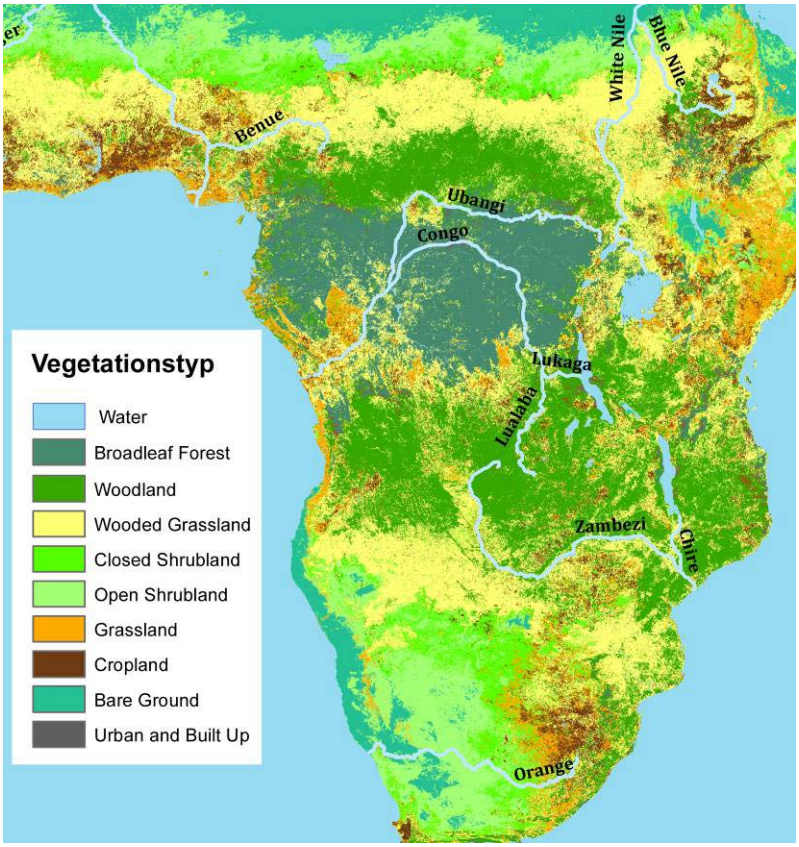



*Abbildung 6: Karte des südlichen Afrikas mit Vegetation und Gewässern; Daten aus UNEP (2000) und ESRI(2010); Eigene Darstellung*

Nachforschungen bezüglich einer möglichen Verwendung von Booten durch die Bantu-Sprecher blieben aber mehr oder weniger ergebnislos. Für eine mögliche Verwendung von Flüssen als Transportwege wurden deshalb andere Anwendungen konsultiert. Livingood (2012) bezog Flüsse bei der Bestimmung historischer Pfade mit ein und sammelte dazu ausführlich Daten zur Reisegeschwindigkeit. Jedoch unterscheidet sich der von ihm untersuchte Mississippi stark von den im Bantu-Gebiet gelegenen Flüssen.

Da die meisten der zentralafrikanischen Flüsse aufgrund vieler Stromschnellen und Wasserfälle nur sehr schwierig schiffbar sind, konnten sie schon von den europäischen Kolonialisten nur schlecht als Wege benutzt werden (Adams 1996). Zwar wäre es zeitlich möglich, dass die Bantu bereits teilweise Boote besessen hatten, dass

diese jedoch bereits als entscheidendes Mittel bei der Expansion benutzt wurden, ist unwahrscheinlich. Ein weiterer Faktor, der im Falle der Bantu-Ausbreitung gegen eine bedeutende Benutzung der Flusswege spricht, ist dass sie meist in Gegenrichtung der Ströme verlief. Entsprechend wurde in der Untersuchung von sehr hohen Kosten für die Flüsse ausgegangen. Ebenso wurde eine Überquerung der Seen mit hohen Kosten verbunden.
Allenfalls sollten später aber auch noch mögliche Auswirkungen tieferer Kosten für diesen Faktor untersucht werden.

*Bodentyp*

Der Bodentyp ist ein weiterer Faktor, der einen Einfluss auf die Bantu-Ausbreitung gehabt haben könnte. So dürften beispielsweise nasse, kaum begehbare Böden ein Hindernis dargestellt haben. Bezüglich der Böden gibt es von der FAO (2007) zur Verfügung gestellte Datensätze in einer Auflösung von 30 Bogensekunden, welchen zudem eine genaue Taxonomie beiliegt. Aus folgenden Gründen wurde jedoch von einer Verwendung der Bodendaten für die LCPA abgesehen: Da es sich bei diesen generell um ein lokal sehr heterogenes Medium handelt, ist es unzureichend, sie mittels grossmassstäblicher Darstellungen zu kartieren. Aufgrund des mit den notwendigen Feldproben verbundenen hohen Aufwands sind Bodenkarten selbst in Mitteleuropa mit einem erheblichen Fehler verbunden (Lüscher 2004). Folglich dürfte die Genauigkeit der Bodendaten für das sowohl sehr ausgedehnte als auch schwer zugängliche Untersuchungsgebiet zu mangelhaft sein, um in der LCPA verwendet werden zu können. Zur ungenügend genauen aktuellen Datenlage käme indes noch der Zeitfaktor hinzu, dessen Einfluss ebenfalls kaum abschätzbar ist. Weiter wäre auch die Messbarkeit dieses Faktors nur ungenügend gegeben. So könnte kaum bestimmt werden, welche Böden wie kalibriert werden sollten (Malczewski 1999).

Schlussendlich wurden folgende Faktoren in der LCPA verwendet:

- ***Gelände***
- ***Vegetation***
- ***Gewässer***

Das anfängliche Ziel bei der Erstellung der Least Cost Paths war es, möglichst viele verschiedene Faktoren in die Berechnung mit einzubeziehen. Die genauere Betrachtung der möglichen Faktoren zeigte jedoch, dass aufgrund ungenügender Datenlage sowie zum Teil sehr schwierig abschätzbarer Einflüsse, eine einfach gehaltene Basis sinnvoller war.

Komplexe Modelle laufen generell Gefahr, aufgrund zu vieler verwendeter Parameter ihre Aussagekraft zu verlieren (Bevan 2011). Oftmals ist es deshalb sinnvoller, mit einem einfachen Modell zu starten und dieses allenfalls im Nachhinein zu variieren. Mit der Einbeziehung aller möglichen Faktoren würde sich gleichzeitig auch die Anzahl an Schätzwerten und somit die Unsicherheit erhöhen (Batten 2007). Da - wie aufgezeigt wurde – bezüglich des Kriteriums Vegetation bereits eine grosse Unsicherheit vorhanden ist, erschien die Verwendung weiterer, noch fehleranfälligerer Kriterien wenig sinnvoll. Vielmehr sollte die Stabilität dieses Faktors anhand einer späteren Variation überprüft werden.

Auch Malczewski (1999) betont, dass bei der Wahl von Evaluationskriterien zur Erklärung eines Sachverhalts deren Anzahl möglichst tief gehalten werden soll. Dies soll verhindern, dass eine Redundanz bzw. Korrelation unter den verschiedenen Variablen vorkommt. Dies könnte unter Umständen bei einer gleichzeitigen Einbeziehung des Bodentyps und der Vegetation der Fall sein, da zwischen diesen klare Wechselwirkungen bestehen.

### 4.1.3 Faktorenkalibrierung

In diesem Kapitel wird die Kalibrierung der einzelnen Faktoren erklärt. Mit Kalibrierung wird die Realisierung respektive Berechnung der aus den einzelnen geographischen Faktoren resultierenden Kosten bezeichnet. Dabei bezieht sie sich nur auf die Gewichtung der Merkmale innerhalb eines Faktors (Berry et al. 2000).
Da der Prozess der Kalibrierung bereits einen grundlegenden technischen Aspekt der LCPA darstellt, wird im Folgenden auf die dabei zu beachtenden unterschiedlichen Kostentypen eingegangen.

#### Isotrope und anisotrope Kosten und Beispiele

Bei den Kosten werden zwei Typen unterschieden: isotrope und anisotrope Kosten. Während isotrope Kosten unabhängig von der Richtung sind, verändern sich anisotrope Kosten mit dieser. Typische Beispiele isotroper Kostenfaktoren sind die Vegetation, die Bodenbedeckung und Seen. Hierbei sind die Kosten nur ortsabhängig (Herzog 2013). Obwohl bereits einige Probleme mit isotropen Kostenoberflächen gelöst werden können, sind sie für viele Fälle unzureichend.
Ein einfacher Fall einer anisotropen Kostenoberfläche kann etwa bei der Modellierung des Winds seine Anwendung finden. So wird sich bei starkem Wind ein Feuer in dessen Richtung schneller ausbreiten als bei ruhigem Wetter, das heisst die Kosten sind geringer. Dieses Beispiel einer anisotropen Kostenoberfläche ist aber insofern relativ einfach, als dass sich die Kosten über die ganze Fläche gleichmässig, beziehungsweise nur in eine Richtung verändern. Ein weiteres Beispiel dieses anisotropen Typs wäre etwa ein Fluss, welcher die Fahrt eines Bootes in eine Richtung begünstigt, in die andere aber erschwert oder ganz verunmöglicht (Collischonn & Pilar 2000).
Das Paradebeispiel schlechthin für den Typ einer komplizierteren anisotropen Kostenoberfläche ist die Neigung. Hier gilt CostDist(A,B) ≠ CostDist(B,A). Da die Neigung der mit Abstand am meisten verwendete Kostenfaktor ist, wird sie vielmals auch synonym mit dem Begriff anisotrope Kostenoberfläche verwendet (Herzog 2013). In dieser Arbeit werden folglich sowohl isotrope als auch anisotrope Kostenfaktoren einbezogen. Deren Berechnung wird im folgenden Kapitel beschrieben.

#### Faktorenberechnung in *gdistance*

Wie bereits im Kapitel zur Software erwähnt wurde, wird in dieser Arbeit zur Erstellung der LCPA das *R-Package gdistance* (Van Etten 2013) verwendet. Dieses ermöglicht die Berechnung grosskreisbasierter Kostenpfade. Dazu dient eine eingebaute Funktion, welche die Verkleinerung der Längengrade vom Äquator Richtung Pole berücksichtigt. Bei einer maximalen Ausdehnung vom Äquator bis zu 32° Süd (Sprache: Xhosa, ISO= xho) ergibt dies mit 17.2 km (gemäss $\alpha = 111.2\ km * \cos(\varphi)$) bereits eine nicht zu vernachlässigende Abweichung. Gegenüber *ArcGIS* hat diese Scripting-basierte Methode weiter den Vorteil, dass sie eine bessere Replizierbarkeit der Ergebnisse unter veränderten Bedingungen erlaubt (Howey 2007, Berry et al 2000, Collischonn et al. 2000).

### Kalibrierung der Vegetation und Flüsse

Die Vegetations- und Flussdaten wurden vor der Kalibrierung noch zubereitet. Da Daten zu den Flüssen bereits im Vegetationsraster vorliegen, wurden diese zur Vermeidung von Redundanz aufeinander abgestimmt.
Die grossen Ströme waren im Vegetationsraster zwar eingetragen, jedoch teilweise unvollständig, respektive stellenweise unterbrochen. Der Grund dafür liegt in der Auflösung des Rasters von knapp einem Kilometer, was zur Folge hat, dass Flussabschnitte von geringerem Durchmesser unter Umständen nicht eingezeichnet sind. Deshalb wurde um die als Lines-Shapefile vorliegenden Flussdaten zuerst in *ArcGIS* ein Pufferstreifen von 2 km gelegt. Der somit etwas übertriebene Wert von 4 km für die Breite ist dadurch begründet, dass dem Fluss auf diese Weise bewusst mehr Gewicht gegeben werden sollte. Dabei wurden aber nur die grossen Ströme berücksichtigt, bei denen eine stabile Datengrundlage gegeben war. Diese sind von Norden nach Süden verlaufend der Weisse Nil, der Ubangi, der Kongo, der Lukuga, der Lualaba, der Sambesi, der Shire sowie der Oranje.

Darauf wurden für den nun vollständigen Faktor Reibungswerte bestimmt. Die Klassifizierung respektive Festlegung der Kosten erfolgte ebenfalls mittels *ArcGIS*.
Die Tabelle 2 bezieht sich auf das gesamte Bantu-Gebiet, sie umfasst einen Ausschnitt von 10°Nord bis 34°Süd und von 7°Ost bis 42°Ost. Der Einfachheit halber wurden die englischen Namen der Vegetationstypen behalten.

| **Typ** | **Anzahl Zellen (absolut)** | **Anzahl Zellen (prozentual)** | **Standard- Gewichtung (V1)** |
|---|---|---|---|
| Wooded Grassland | 4623807 | 28.6 | 2 |
| Woodland | 4467442 | 27.6 | 5 |
| Evergreen Broadleaf Forest | 2179570 | 13.5 | 10 |
| Grassland | 1183647 | 7.3 | 1 |
| Open Shrubland | 1134989 | 7.0 | 2 |
| Closed Shrubland | 835500 | 5.2 | 4 |
| Cropland | 826445 | 5.1 | 4 |
| Bare Ground | 468300 | 2.9 | 10 |
| Water | 386308 | 2.4 | 15 |
| Deciduous Broadleaf Forest | 69615 | 0.4 | 7 |
| Urban and Built Up | 5408 | 0.0 | 5 |

*Tabelle 2: Gewichtung der Vegetationstypen (V1) sowie deren Anteil für Gebiet von 10° N bis 34° Süd und 7° West bis 42° Ost*

Um die Wichtigkeit der Einflüsse zu verdeutlichen, sind die Vegetationstypen entsprechend ihrer Flächenanteile aufgelistet. Der Typ Grassland (auf Deutsch „Grünland“) wurde mit dem Wert 1 als der Faktor mit den geringsten Kosten definiert. Der immergrüne Regenwald stellt mit dem Wert 10 die obere Grenze der eigentlichen Vegetation bezüglich der Reibung dar. Wobei hierbei zu betonen ist, dass sein Anteil am bezüglich der Early- und Late-Split Hypothesen entscheidenden Nordteil noch

einiges grösser ist (siehe Abbildung 6). Weil die Binnenwasserflächen jedoch ebenfalls schon in dieses Raster integriert wurden, stellen diese noch einmal einen grösseren Widerstand dar. Aufgrund ihrer verhältnissmässig geringen Grösse ist ihr Einfluss aber trotz des höchsten Wertes von 15 relativ gering. Die Gewässer wurden dabei bewusst nicht als vollständige Hindernisse oder Barrieren definiert, weil diese riesige Umwege zur Folge gehabt hätten respektive einen Südpfad gar nicht erst erlaubt hätten. Zudem dürfte es durchaus Möglichkeiten gegeben haben, um Flüsse zu überqueren.
Bei den Werten für die restlichen Vegetationen handelt es sich um Schätzwerte, welche für die jeweiligen Typen plausibel erschienen. Sie sollten später noch variiert werden. Die Werte für die beiden eindeutig durch den Menschen beeinflussten Gebiete (Cropland und Urban and Built Up) wurden mittels Vergleich einer Karte der natürlichen Biome bestimmt (Adams 1996).

Das erstellte Kostenraster wurde nun in gdistance eingelesen und darin mit einer einfachen Formel noch angepasst.

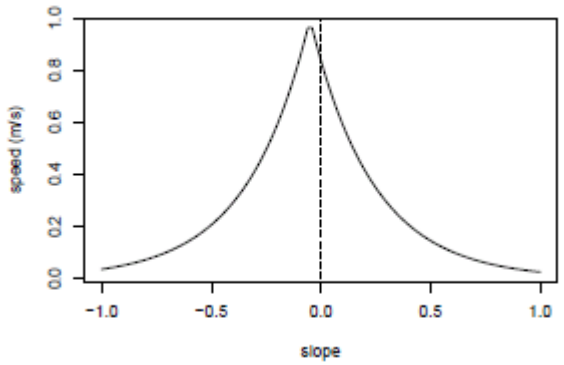
- v <- transition (1 / v, mean, 8)
- vg <- geoCorrection(vtr, scl=TRUE)

Im Gegensatz zu den meisten anderen LCP-Softwares rechnet *gdistance* nicht direkt mit den Kosten, sondern verwendet den reziproken Leitwert. Darum wird der Kehrwert (1/v) des in ArcGIS erstellten Kostenrasters in die Funktion eingegeben. Als Funktion zur Erstellung der Übergangsmatrix wird das Mittel („mean“) benutzt. Weitere Details der dargestellten Formel werden bei der folgenden Berechnung der Neigungskosten näher erläutert.

Um mögliche Kombinationen mit den noch in höherer Auflösung vorhandenen Geländemodelldaten zu ermöglichen, könnten die Vegetationsdaten nötigenfalls zuvor mittels *resampling* in ihrer Auflösung angepasst werden.

## Kalibrierung der Neigung

Wie zuvor erläutert wurde, stellt die Neigung eine kompliziertere Anwendung einer Kostenoberfläche dar. Zur Berechnung der damit verbundenen Bewegungskosten liegen zahlreiche Methoden vor. Eine der bekanntesten davon ist sicherlich diejenige von Waldo Tobler (1993). Dieser als *Tobler's Hiking Function* bekannte Ansatz wirdauch in *gdistance* verwendet. Bei der Bestimmung seiner Funktion benutzte Tobler Wanderdaten der Schweizer Armee. Daraus leitete er folgende Formel ab:




**Speed =**

**6 * exp(-3.5 * abs(slope + 0.05))**

*Abbildung 7: Zusammenhang zwischen Geschwindigkeit und Neigung gemäss Tobler's Hiking Function; Quelle: Van Etten (2013)*

Wie aus der Formel und der Grafik leicht zu entnehmen ist, ging Tobler von einer Maximalgeschwindigkeit bei einem leichten Gefälle von 0.05 aus.
*Gdistance* verwendet Funktionen aus einer Reihe weiterer *R-Packages*, wobei *raster* die wichtigste Einheit darstellt. Dieses stellt umfassende Funktionen zur Bearbeitung und Manipulation von Grid-Daten zur Verfügung. Die Kostenpfade werden bei *gdistance* jedoch nicht vom Raster selbst berechnet, sondern diese werden dazu zuerst in Zwischenobjekte umgewandelt. Diese sogenannten Transitionsmatrizen bilden das Kernstück von *gdistance*. Sie enthalten einerseits die geographischen Daten des originalen Rasters, zudem stellt eine Matrix die Übergänge zwischen den einzelnen Zellen dar. Anders als bei den meisten Anwendungen werden dabei nicht die Reibungswerte oder Kosten, sondern die Leitwerte zwischen den Zellen angegeben. Dabei werden speichereffiziente *Sparse Matrices* verwendet. Die Erstellung der Transitionsmatrizen erfolgt auf direktem Weg mittels beliebiger Formel. Im Folgenden ist die Berechnung der Neigungskosten gemäss Tobler entsprechend dem *gdistance*-Script (Van Etten 2013) gegeben:

- heightDiff <- function(x) {x[2] –x[1]}
- hd <- transition(r, heightDiff, directions= 8, symm=FALSE)

In einem ersten Schritt wird eine Funktion „heightDiff" definiert, welche den Höhenunterschied zwischen den Zellen berechnet. Danach wird das Ausgangsraster „r" gemäss ebendieser Formel in eine Transitionsmatrix umgewandelt. Diese beinhaltet nun die jeweiligen Höhenunterschiede zwischen den Zellwerten. Ein weiterer Parameter gibt an, wie gross die bei der Kalkulation zu berücksichtigende Nachbarschaft sein soll. Mit dem letzten Argument wird eine asymmetrische Matrix erzeugt, als Bedingung für einen anisotropen Kostenfaktor.

- slope <- geoCorrection(hd, scl= FALSE)

Mit der Funktion *geoCorrection* wird nun die berechnete Höhe zwischen den Zellen durch die horizontale Entfernung geteilt. Gemäss m = ∆y / ∆x erhalten wir somit die Neigung zwischen den Zellen. Bedingung ist hierbei, dass beim Ursprungsraster Distanzen und Höhen in derselben Einheit vorliegen. Aus der Neigung kann nun die Geschwindigkeit berechnet werden.

- adj <- adjacent(r, cells = 1:ncell(r), pairs= TRUE, directions=8)
- speed <- slope
- speed[adj] <- exp(-3.5 * abs(slope[adj] + 0.05))
- x <- geoCorrection(speed, scl=FALSE)

Dazu wird zuerst mit der Funktion *adjacent* die Operation auf die jeweils angrenzenden Zellen limitiert. Nun wird die Funktion von Tobler zur Berechnung der Geschwindigkeit eingesetzt. Dabei wird die zuvor definierte Variable „adj" verwendet, um die Geschwindigkeit auf angrenzende Zellen zu limitieren. In einem letzten Schritt wird wiederum die Funktion *geocorrection* eingesetzt. Sie berücksichtigt, dass bei der diagonalen Durchquerung einer Zelle ein längerer Weg (√2= 1.41) zurückgelegt werden muss als bei der geraden. Weiter wird damit auch die bereits zuvor angesprochene, bei geographischen Koordinaten vorhandene Längengradverzerrung korrigiert.

### Überführung der Raster in Graphen

Ein wichtiges technisches Kriterium der LCPA stellt auch die Grösse der Nachbarschaft dar, welche miteinbezogen werden soll. Diese wurde bei der zuvor erläuterten Berechnung der Neigungskosten unter dem Argument „direction" festgelegt. Da es sich dabei um einen zentralen technischen Aspekt der LCPA handelt, wird im Folgenden genauer darauf eingegangen.

Weil LCP Berechnungen normalerweise auf Rasterdaten basieren, ist es notwendig diese zuerst in Vektoren umzuwandeln. Dabei werden alle Zellen mit einem virtuellen Vektor oder Graphen mit den umliegenden Zellen verbunden. Entsprechend wird eine bestimmte Menge an Nachbarszellen einbezogen. Weil die verwendete Nachbarschaft die Wahl des kürzesten Pfades durch den später eingesetzten Suchalgorithmus beeinflusst, hat sie somit direkte Auswirkungen auf das Resultat der Kostenpfade. Generell gilt dabei, je mehr Nachbarszellen einbezogen werden, desto genauer ist die Pfadberechnung (Herzog & Posluschny 2008).

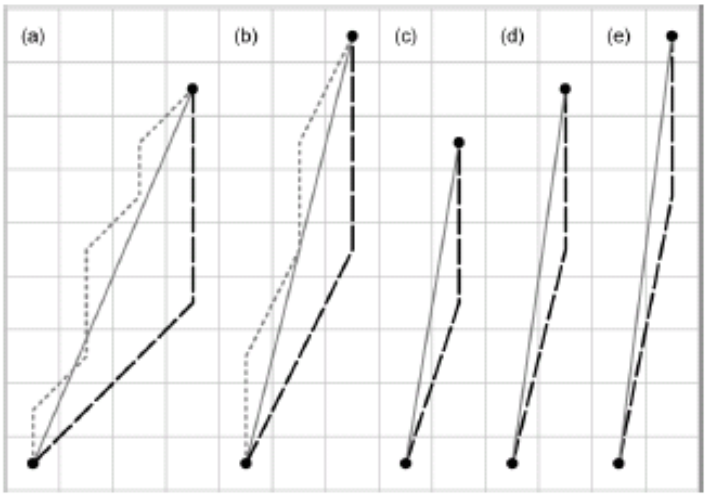



*Abbildung 8: Worst-Case-Szenarien der durch die Raster-Graph Konvertierung verursachten Abweichung des Kostenpfades bei a) 3x3 , b) 5x5, c) 7x7, d) 9x9 und e) 11x11 Nachbarschaft; Quelle: Herzog (2013)*

Zur Darstellung der verschiedenen Nachbarschaften werden oftmals Schachfiguren und die damit möglichen Züge herangezogen: Mit dem Turm sind bloss 4 Zellen in Reichweite. Eine 8-er Nachbarschaft stellt den Radius dar, welcher mit dem König erreichbar ist. Verwendet man zusätzlich einen Springer, kann man bereits 16 Zellen erreichen. Weitere Beispiele sind 24, 32 oder 48 verschiedene Nachbarn. Die Pfadberechnung zwischen zwei Punkten A und B im

uniformen Raum zeigt, dass bei einer kleinen Nachbarschaft dieser berechnete Pfad je nachdem beträchtlich von der optimalen Geraden abweicht (Bevan 2011). Wie aus Huber & Church (1985) zu entnehmen ist, kann der mit einer 8-er Nachbarschaft berechnete Pfad um bis zu 8 Prozent vom Optimum abweichen. Bei Verwendung einer 24-er Nachbarschaft ermittelten sie einen höchstmöglichen Fehler von 2.8 Prozent, bei einer 48-er Nachbarschaft liegt er unter 1.4 Prozent.

Die berücksichtigte Nachbarschaftsanzahl unterscheidet sich je nach verwendeter Software. Bei *ArcGIS* (ESRI 2013) werden nur 8 Nachbarn einbezogen, bei *gdistance* ist dies der Standardwert, optional kann eine 4-er oder eine 16-er Nachbarschaft bestimmt werden.

### Probleme mit Arbeitsspeicherauslastung

Wie im vorhergehenden Kapitel beschrieben, wurden für die Berechnung des Neigungsfaktors nacheinander vier verschiedene Transitionsmatrizen erstellt. Nun ist deren Erstellung trotz der dabei verwendeten *Sparse Matrices* bei der Grösse des untersuchten Gebietes mit erheblichem Arbeitsspeicheraufwand verbunden. Da die im Raster gespeicherten Werte im Bereich von 0 bis mehreren tausend Metern (Kilimandscharo als höchster Punkt bei 5895 m) liegen, wird zu ihrer Darstellung eine Bit-Tiefe von mindestens 16-Bit (ohne Vorzeichen) benötigt. Die DEM-Rastergrösse des Untersuchungsgebietes liegt bei einer Auflösung von 30 Bogensekunden bei ungefähr 40 MB. Dieser Wert steigt für die einzelnen berechneten Transitionsmatrizen um bis das Siebenfache an. Wegen der zusätzlichen enormen Produktion von Memory-Daten stösst der Arbeitsspeicher eines konventionellen Computers sofort an seine Grenzen. Die Tabelle 3 zeigt die verschiedenen Speicherwerte der verwendeten Raster. Im Gegensatz zum DEM ist zur Darstellung der Vegetationswerte (von 1 bis 15) bereits eine Pixel-Tiefe von 4-Bit ausreichend. Da jedoch aus unerklärlichen Gründen die Darstellung in *R* darunter etwas litt, wurden dazu 8-Bit verwendet. Auch dies stellte kein Problem dar.

| **Auflösung** | **Anzahl columns&rows** | **DEM (16Bit)** | **Veg (8Bit)** | **Veg (4Bit)** |
|---|---|---|---|---|
| ~ 1kmx1km | 4200x5300(~22*10^6) | ~40 MB | 20 MB | 10 MB |
| ~ 500m x 500m | 8400x10600(~90*10^6) | ~150 MB | | |

*Tabelle 3: Für DEM und Vegetations-Raster benötigter Speicher bei entsprechender Auflösung für ganzes Bantu-Ausbreitungsgebiet*

Zur Lösung des Speicherproblems wurde ein virtueller Server mit Ubuntu (12.04) aufgesetzt, welche via Secure Shell (ssh) bedient werden konnte. Der über 48 GB RAM verfügende Rechner sollte die enorm speicheraufwendige Berechnung der Neigungskosten ermöglichen.
Doch auch mit der erhöhten Arbeitsspeicherkapazität konnte das Problem nicht vollständig gelöst werden. Zwar erlaubten die verbesserten technischen Mittel die Berechnung der Neigung bei einer 30“-Auflösung für den nördlichen Ausschnitt des Bantu-Gebietes (10° Nord, 41° West, 10° Süd, 7 ° Ost), die Einbeziehung des gesamten Gebiets war aber auch so nicht möglich.
Wie aus der Aktivitätsanzeige entnommen werden konnte, füllte sich der Arbeitsspeicher bei der Berechnung der Neigungskosten kontinuierlich auf, bis der Prozess abgebrochen wurde. Auch das Löschen nicht mehr notwendiger Variablen oder die *gc*-Funktion (*Garbage Collection*) konnten dies nicht verhindern.

Da die Berechnungen bei *gdistance* auf dem Package *igraph* basieren, liegt die Ursache des hohen Arbeitsspeicheraufwandes bei diesem. *Gdistance* fungiert dabei lediglich als Schnittstelle: Zur Berechnung der Variablen schreibt es Input-Daten in eine *igraph*-Funktion um. Später erhält es die Resultate aus *igraph* zurück und konvertiert diese wieder in ein geographisches Objekt. Aus diesem Grund lässt sich das Speicherproblem kaum vermeiden.
Aufgrund der Vorteile von *gdistance* gegenüber anderer Pakete wie *ArcGIS* (u.a. grosskreisbasierte Berechnung), wurde jedoch nicht auf dieses verzichtet. So wurde zur Umgehung der Speicherprobleme die Untersuchung der Kostenpfade auf den nördlichen Teil des Bantu-Ausbreitungsgebiets beschränkt. Dieser Nordteil (7° West bis 41° Ost, 10° Nord bis 10° Süd) umfasst bezüglich Längengrade das ganze Expansionsgebiet, jedoch nur die nördliche Hälfte bezüglich der Breitengrade. Die Wahl des Gebiets fiel dabei bewusst auf diesen Ausschnitt. Da er die bezüglich Early- und Late-Split entscheidende Region beinhaltet, kann trotz der Einschränkung der zur Klärung dieser Frage entscheidende Teil abgedeckt werden.
In Folge der Speicherprobleme musste leider auch ganz davon abgesehen werden, noch höhere Geländemodell-Auflösungen von Viertel- oder gar Achtelbogensekunden zu verwenden. Da für die doppelte Auflösung die im Quadrat zunehmende Rastergrösse bereits bei 76 MB liegt, war eine Berechnung auf dieser Skala ausgeschlossen.

### 4.1.4 Gewichtung der Faktoren

Verschiedene Kriterien werden auf unterschiedlicher Skala gemessen. Es ist deshalb eine der grossen Schwierigkeiten beim Zusammenfügen verschiedener Kriterien, diese in vergleichbare Einheiten zu konvertieren (Malczewski 1999).

Auch bei der LCPA kann das Zusammenfügen der Faktoren unter Umständen schwierig sein. Dies wird dadurch verdeutlicht, dass zum Beispiel je nach gewählter Software unterschiedliche Einheiten gemessen werden. Zum Teil sind dies die benötigte Energie, die Zeitdauer oder undefinierte Einheiten (Field et al. 2005). Ebenso stellt bei dieser Arbeit das Gelände respektive das Höhenmodell der einzige quantitative Faktor dar. Bei der Vegetation handelt es sich um eine kategoriale, qualitative Variable. Ebenso bildet der Kostenfaktor Binnengewässer, der ja bereits in die Vegetation integriert wurde, gewissermassen nur eine weitere Klasse innerhalb dieser (Malczewski 1999).

Wie Atkinson et al. (2005) betonen, stellt die sinnvolle Kalibrierung und Kombination mehrerer Faktoren zur Bildung einer Kostenoberfläche einen wichtigen Punkt der LCPA dar. Diesem werde in den meisten methodologischen Diskussionen zu wenig Aufmerksamkeit gewidmet. Weil das den einzelnen Kriterien zugeordnete Gewicht direkt das Resultat beeinflusst, sei ein systematisches Vorgehen zur Bestimmung der relativen Bedeutung der einzelnen Faktoren notwendig. Das dazu geeignete Verfahren stellt die Multi Criteria Analysis dar. Auch Howey (2007) verwendet diese in ihrer archäologischen Studie bezüglich sozialer Interaktionen in Michigan.

Die Multi-Criteria-Analysis stellt eine häufig verwendete Methode für die Entscheidungsfindung bei mehreren involvierten Variablen dar. Dabei können die Gewichte der einzelnen Faktoren unter anderem mittels des vom Mathematiker Saaty (1980) entwickelten Analytical Hierarchy Process bestimmt werden.

Aufgrund des explorativen Charakters dieser Arbeit wurde jedoch keine Multi Criteria Analysis durchgeführt. Dies wurde dadurch begründet, dass in diesem Falle die dazu notwendigen Vorkenntnisse nicht vorhanden waren. Weil die Vegetationsdaten bereits mit den grossen Strömen verschnitten wurden, blieben zudem nur noch zwei verschiedene Raster respektive eigenständige Faktoren übrig. Deshalb schien es sinnvoller, mit einem pragmatischen Ansatz zuerst deren Einflüsse zu analysieren und das weitere Vorgehen an den dadurch erlangten Informationen auszurichten.

### 4.1.5 Zugrunde liegende Algorithmen

Weil das Erstellen von Kostenoberflächen in realen Landschaften schnell eine sehr komplexe Aufgabe darstellt, finden dabei Computer-Algorithmen ihre Anwendung. Zwar existiert eine Vielzahl dazu verwendbarer Algorithmen, meist werden für diese Problemstellung aber vor allem zwei verschiedene verwendet. Der Dijkstra- und der A*-Algorithmus („A-star“ ausgesprochen) sind sowohl in den Computerwissenschaften als auch in der Videospiel-Industrie weit verbreitet.
Der Dijkstra-Algorithmus, der am weitesten verbreitete Ansatz, wurde 1956 vom holländischen Computer-Wissenschaftler Edsger Dijkstra (1959) konzipiert (Surface-Evans & White 2012). Er wurde für die Berechnung von Distanzen in einem Netzwerk erstellt. Dabei stellen die Gesamtkosten das aus den einzelnen Kantenkosten summierte Total dar. Voraussetzung für seine Anwendung ist, dass es sich bei den Gewichten der Kanten ausschliesslich um positive Werte handelt, ein Nachteil gegenüber etwa dem Bellman-Ford Algorithmus. So nimmt gemäss dem Djikstra-Algorithmus ein Weg mit jeder Verlängerung zwangsläufig auch an Kosten zu (Cormen et al. 2009). In geographischen und archäologischen Beispielen trifft dies aber auch praktisch immer zu. Da zahlreiche gut illustrierte Erklärungen bezüglich des genauen Vorgehens beim Dijkstra-Algorithmus im Internet zu finden sind, wird hier darauf verzichtet.

Grundsätzlich speichert der Dijkstra-Algorithmus immer folgende Variablen: Zum ersten ob a) ein Knoten bereits besucht wurde und die dafür aufgewendeten Kosten, b) wie viel beim Besuch der weiteren anliegenden Knoten als Summe hervorgehen würde sowie c) die Richtungen der jeweiligen Distanzen.

Weil der Dijkstra-Algorithmus immer den am leichtesten erreichbaren, respektive den nächsten Knotenpunkt zuerst auswählt, wird er der Gruppe der „Greedy-Algorithmen“ zugeordnet. Die Greedy-Algorithmen gehen Probleme in einer Sequenz von einzelnen Schritten an, wobei bei jedem dieser Schritte von verschiedenen Optionen die erfolgversprechendste ausgewählt wird. Bei vielen Optimierungsproblemen haben sie gegenüber dynamischer Programmierungs-Algorithmen wie dem Bellman-Ford-Algorithmus den Vorteil, dass sie schneller und effizienter sind. Greedy-Algorithmen können so oftmals lokale Optima, nicht aber die global optimale Lösung finden. Wie Cormen et al. (2009) aufzeigt, trifft dies beim Dijkstra-Algorithmus aber nicht zu, dieser findet korrekt implementiert immer den kürzesten Weg.

Der A*-Algorithmus verwendet als Erweiterung gegenüber dem Dijkstra-Ansatz noch eine Schätzfunktion. Dieser heuristische Ansatz verkleinert die zu untersuchende Fläche. Indem verhindert wird, dass die gesamte Oberfläche einbezogen wird, benötigt der A*-Algorithmus weniger Zeit als der Dijkstra-Algorithmus. Der zeitliche Gewinn durch den A*-Algorithmus kann unter Umständen eine Einbusse bei der Genauigkeit mit sich bringen. Dies hängt jedoch auch von der verwendeten heuristischen Funktion ab (Surface-Evans & White 2012). Weil beim Dijkstra-Algorithmus immer

der gesamte Untersuchungsraum mit einbezogen wird, stellt er ein typisches Beispiel einer globalen Funktion dar. Dies hat den entscheidenden Nachteil, dass der Prozess nicht in einzelne Teile zerlegt werden kann, um die Speicherauslastung einzugrenzen.

Ein weiterer Faktor, der die Genauigkeit des mit dem Dijkstra-Algorithmus erzielten Resultats beeinflusst, wurde bereits zuvor beim Problem der Raster-Graph Konvertierung erläutert. So kann der mit dem Dijkstra-Algorithmus bestimmte Kostenpfad unter Umständen trotz dessen korrekter Anwendung dennoch vom optimalen Weg abweichen (Herzog & Posluschny 2008).

Bei *gdistance* wird ebenfalls der Dijkstra-Algorithmus angewandt. Wie bereits erwähnt wurde, erfolgt die Berechnung selbst durch das Package *igraph*. Dieses benutzt je nach Situation unterschiedliche Algorithmen, bei positiv gewichteten Graphen wird aber die Dijkstra-Methode implementiert (Csardi 2014).

## 4.2 Berechnung der LCPA

Dieses Kapitel widmet sich der Durchführung der LCPA. Dabei werden verschiedene Parameter respektive Kriterien variiert, um die Robustheit des Modells zu testen. Ausgehend von einem einfachen Ansatz mit Einzelpfaden, wird dieser anhand der gewonnenen Informationen in mehreren Schritten modifiziert, beziehungsweise erweitert. Ausgehend von einem geringen Vorwissen und grossem Unsicherheitsfaktor wurde gewissermassen ein Ad-hoc-Ansatz verwendet.

### 4.2.1 Untersuchungen Nordteil

Wie zuvor beschrieben wurde, kam es bei der Berechnung der Neigungskosten zu Engpässen mit dem Arbeitsspeicher. Um die Neigung dennoch als Kriterium verwenden zu können, wurde im Folgenden nur der nördliche Teil des Gebiets untersucht.

#### Variation von Einzelpfaden

Bei den ersten Untersuchungen wurden Kostenpfade zwischen einzelnen Punkten aus dem Sprachdatensatz (Nichols & Bickel 2009) berechnet. Dabei wurde von einem zentral auf dem Grenzgebiet gelegenen Anfangspunkt A (11.6° Ost, 6.4 ° Nord) ausgegangen. Dieser wurde bestimmt, indem ein Pufferstreifen entlang dem Grenzgebiet zwischen Nigeria und Kamerun gebildet wurde und daraus der Mittelpunkt bestimmt wurde. Als Zielpunkte wurden verschiedene östliche, im Gebiet der grossen Seen Afrikas gelegenen Punkte aus dem Sprachdatensatz verwendet. Bei diesen wurden bewusst bezüglich der Breitengrade variierende Punkte gewählt. Auf diese Weise konnte beobachtet werden, bis zu welchem Punkt ein Nordost- beziehungsweise Südpfad resultierte.

#### Variation technischer Parameter

In einem nächsten Schritt wurden die Einflüsse verschiedener technischer Parameter auf die resultierenden Kostenpfade veranschaulicht. Es wurden daher bei den folgenden Punkten unterschiedliche Einstellungen verwendet:

- *Geo-Korrektur:* Hierbei wurde in einem Durchgang der Schritt der Geokorrektur weggelassen. Auf diese Weise wurden die bei Vernachlässigung der Erdkrümmung resultierenden Kostenpfade (vgl. *ArcGIS*) erhalten.

- *Nachbarschaft:* Um den Einfluss der einbezogenen Nachbarschaft zu untersuchen, wurde die LCPA sowohl mit der Standard 8-er Nachbarschaft als auch mit der 16-er Nachbarschaft durchgeführt.

#### Weitere Variationen

- *Anfangspunkt:* Mit der Wahl unterschiedlicher Ausgangspunkte für die LCPA wurde ein weiterer Zufallsfaktor untersucht.

- *Flüsse als Transportwege:* Zudem wurde in einigen Versuchen nachträglich ausprobiert, ob tiefe Kosten für die Flüsse die Wege beeinflusst hätten.

#### Multipunkte-LCP

Nachdem in einigen Versuchen technische Einflüsse untersucht worden waren, wurde die Pfadberechnung auf mehrere Punkte erweitert. Mittels Iteration via *lapply*-Funktion in *R* (aus der Gruppe der *apply*-Funktionen) wurde ein Pfad zu allen im Untersuchungsgebiet liegenden Punkten (85 Punkte) erstellt. Die Multipunkte-LCPA hat den Vorteil, dass sie die Trennung bezüglich Early- und Late-Split viel besser darstellt. Mit unterschiedlichen Kalibrierungen der Vegetation konnte die Anzahl der durch die beiden Hauptstränge verbundenen Punkte variiert werden.
Dabei wurde nun auch einheitlich ein systematisch bestimmter Anfangspunkt verwendet. Dieser wurde mittels *mean center* in *Spatial Statistics* von *ArcGIS* aus 17 nordwestlich gelegenen Sprachpunkten des Datensatzes bestimmt. Die Berechnung des statistischen Mittels der Punkte im ungefähren Bantu-Ursprungsgebiet ergab die Koordinaten 10.48° Ost und 3.48° West. Die folgenden Kostenpfade wurden von diesen Koordinaten ausgehend berechnet.

#### DEM versus Vegetation

Mittels der Multipunkte LCPA konnten zudem auch die bei den verschiedenen Faktoren und deren Kombination resultierenden Pfade gut verglichen werden.
Die Resultate zeigten, dass der Faktor Neigung einen sehr geringen Einfluss auf den Verlauf der Pfade hat. Er wurde daher für die weiteren Berechnungen vernachlässigt. Da das DEM bisher ein limitierender Faktor gewesen war, konnte nun das gesamte Bantu-Ausbreitungsgebiet in die Berechnungen einbezogen werden.

### 4.2.2 Untersuchung des Gesamtgebietes

Nachdem aufgezeigt worden war, dass die Neigung bei der Bantu-Ausbreitung ein vernachlässigbarer Faktor darstellt, wurde bei der Berechnung nur noch das Vegetationsraster verwendet. Die Berechnungen fanden im Folgenden immer für das gesamte Gebiet statt. Von den ursprünglich 141 Bantu-Sprachen im Datensatz von Nichols & Bickel (2009) waren bereits anfangs 3 ausgefällt worden. Zudem dienten die zuvor erwähnten 17 Punkte im Ursprungsgebiet ebenfalls nicht als Zielpunkte. Folglich blieben noch 121 Sprachen übrig.

#### Überprüfung der Stabilität des Modells

Als nächstes sollte die Robustheit des Modells überprüft werden. Diese sollte aufzeigen, wie stabil die berechneten Pfade gegenüber leichten Veränderungen der Verteilung der Kosten sind. Insbesondere aufgrund der unsicheren Datenlage stellt diese Überprüfung einen sehr wichtigen Schritt bei der LCPA dar.

### Mögliche Anwendung von Monte-Carlo-Simulation

Die dazu geeignete Methode ist die Monte-Carlo-Simulation. Die Monte-Carlo-Simulation dient zur Schätzung von Quantitäten, welche sehr schwierig oder ganz unmöglich zu berechnen sind. Sie erlaubt mittels wiederholter Modellierung eines Zufallsfehlers Aussagen zu dessen Auftretenswahrscheinlichkeit. Im Falle der LCPA würde dazu dem Vegetationsraster ein zufälliger, autokorrelierter Fehler hinzugerechnet und die Kostenpfade dafür berechnet werden. Dieser Vorgang würde nun wieder und wieder durchgeführt werden. Auf diese Weise kann aufgrund der relativen Anzahl der Durchgänge, bei denen ein bestimmter Pfad resultiert, eine Aussage zu dessen Wahrscheinlichkeit abgeleitet werden. Um mithilfe der Monte-Carlo-Simulation aussagekräftige Resultate zu erhalten, wären jedoch im Minimum 1000 verschiedene Berechnungsdurchgänge notwendig (Grinstead & Snell 1997).
Weil eine einzelne Multipunkte Pfadberechnung in *gdistance* aber bereits mehrere Stunden dauert, konnte die Monte-Carlo-Methode nicht angewendet werden.

Im Folgenden wurde noch die Möglichkeit einer Anwendung in *ArcGIS* überprüft. Deshalb wurde eine Multipunkte-LCP Berechnung mittels Iterator von *Modelbuilder* in *ArcGIS* (ohne Geokorrektur) durchgeführt: Das Kostenraster wurde dabei mit etwa 30 Minuten relativ schnell erstellt, jedoch dauerte die Berechnung aller einzelnen Pfade 3 bis 4 Stunden. Die Dauer der jeweiligen Prozesse war damit genau umgekehrt wie bei *gdistance*, die Summe jedoch in etwa dieselbe. Auch auf diese Weise lag eine Monte-Carlo-Simulation folglich nicht im Zeitrahmen.

### Manuell variierte Kalibrierung

Weil die Anwendung der Monte-Carlo-Simulation also nicht möglich war, wurde die Berechnung daraufhin einfach für zehn verschiedene Kalibrierungen realisiert. Dazu wurden in *ArcGIS* zehn unterschiedlich gewichtete Varianten des Vegetationsrasters erstellt. Bei diesen wurden die Werte der einzelnen Vegetationstypen bewusst in einem gewissen Rahmen variiert. Auf diese Weise sollten mögliche Unsicherheiten bezüglich der genauen Bedeckung einigermassen erfasst werden. Von diesen zehn verschiedenen Kalibrierungen werden hier drei genauer betrachtet. V1 stellt die bereits im Kapitel 4 zur Berechnung der Vegetationskosten besprochene Standardkalibrierung dar. In der Tabelle 4 sind bei den Gewichtungen V4 und V10 diejenigen Werte markiert, welche verändert wurden. Während dies bei V4 gleich sieben Werte sind, wurden bei V10 nur zwei neu definiert. Allerdings muss hier beachtet werden, dass die zuunterst aufgeführten Werte nur ein sehr geringes Gewicht haben.
Abgesehen von den Wasserflächen, welche eine Art Sonderfall darstellen, wurden bei der vierten Kalibrierung alle Faktoren einander angeglichen. Die Variationsbreite liegt nur noch bei fünf, bei V1 und V10 beträgt sie neun. Bei V10 wurde der Wert für den tropischen Regenwald belassen, hingegen die Werte für die angrenzenden Typen Woodland und Wooded Grassland verringert.

| Typ | Anzahl Zellen | V1 | V4 | V10 |
|---|---|---|---|---|
| Wooded Grassland | 4623807 | 2 | 2 | 1 |
| Woodland | 4467442 | 5 | 4 | 3 |
| Evergreen Broadleaf Forest | 2179570 | 10 | 6 | 10 |
| Grassland | 1183647 | 1 | 1 | 1 |
| Open Shrubland | 1134989 | 2 | 2 | 2 |
| Closed Shrubland | 835500 | 4 | 3 | 4 |
| Cropland | 826445 | 4 | 3 | 4 |
| Bare Ground | 468300 | 10 | 6 | 10 |
| Water | 386308 | 15 | 15 | 15 |
| Deciduous Broadleaf Forest | 69615 | 7 | 6 | 7 |
| Urban and Built up | 5408 | 5 | 4 | 5 |

*Tabelle 4: Drei Verschiedene Kalibrierungen (V1 , V4, V10) für den Kostenfaktor Vegetation*

## 4.3 Analyse der Kostenpfade

Im Folgenden wurden die berechneten Kostenpfade analysiert. Neben den Distanzen zwischen den einzelnen Punkten sollten dabei auch die topologischen Distanzen mittels Flussordnungen bestimmt werden. Zu diesem Zweck wurden die in *R* berechneten zehn Multipunkte-Distanzpfade für die verschiedenen Vegetationswerte in *ArcGIS* überführt. Zwar liessen sich wohl auch in *R-Statistics* die Pfaddistanzen berechnen, da jedoch keine Funktionen zur Bestimmung der Flussordnungen vorlagen, wurde dazu *ArcGIS* verwendet.

Gleichzeitig mussten die Glottolog-Daten und die in der LCPA benutzten Sprachdaten aufeinander abgestimmt werden.

### 4.3.1 Datenaufbereitung der Kostenpfade

#### Konvertierung und Einlesen in *ArcGIS*

Anfangs wurde versucht, die in Bilddateien (tiff) abgespeicherten Resultate der LCP-Berechnungen direkt (unreferenziert) in *ArcGIS* einzulesen und danach von Hand zu referenzieren und für weitere Analysen zu verwenden. Es zeigte sich jedoch sehr bald, dass dieser Weg einerseits sehr umständlich und zum anderen aufgrund der vielen manuellen Schritte auch etwas fehleranfällig respektive ungenau ist.

Deshalb wurden die *SpatialLines* in R-Statistik zuerst entpackt und mittels eines Basisrasters in ein *Dataframe*-Objekt überführt. Auf diese Weise konnten sie mittels *maptools*-Package als Shapefiles ausgelesen werden. Auch dieser Prozess war relativ speicheraufwendig und erfolgte daher noch über die Secure Shell.

Danach wurden die Shapefiles in *ArcGIS* eingelesen und projektiert. Als Projektion wurde die für Kontinente geeignete Methode *African Equidistant Conic* verwendet. Dabei wurde ein *False Easting* von 10.48 Grad und ein *False Northing* von 4.38 Grad verwendet. Dieser Punkt entspricht dem Ursprungsort der Kostenpfade. Auf diese Weise sollten die Distanzen der Pfade möglichst genau wiedergegeben werden.

#### Bestimmung der Fluss-Ordnungszahlen nach Shreve

Nun wurden Flussordnungszahlen für die Kostenpfade bestimmt. Diese werden – entsprechend ihrem Namen – in der Hydrologie verwendet, um Flüsse anhand der

Anzahl ihrer Zuflüsse respektive Nebenflüsse klassifizieren zu können. Weil die Multipunkte-Kostenpfade bei einer Gesamtbetrachtung (in die entgegengesetzte Richtung) starke Ähnlichkeiten mit einem Flusssystem aufweisen, wurde deshalb diese Methode aus der Hydrologie herangezogen.
Bei der Zuweisung der Flussordnungszahl wird das ganze Wasserlaufnetzwerk betrachtet und daraus für jeden Teilstrom eine numerische Ordnung abgeleitet. Es werden dazu hauptsächlich zwei verschiedene Methoden angewandt, die nach Strahler (1952) und die nach Shreve (1966). In diesem Fall ist vor allem letztere relevant, weil sie jeden einzelnen Pfad berücksichtigt. Dabei wird zuerst jedem einzelnen Fluss die Ordnung 1 zugeordnet. Vereinen sich zwei Ströme, werden ihre Ordnungen addiert.

Zur Bestimmung der Ordnungszahlen wurde *ArcGIS* gewählt. Dieses verlangt als Input ein Eingaberaster für den Wasserlauf sowie ein Flussrichtungsraster. Dazu wurden die Vektorenlinien zuerst in ein Raster umgewandelt und ein künstliches DEM mittels der „rasterisierten" Linien erzeugt. Um dieses zu erstellen, wurde den Linien ein konstanter Wert zugeteilt und ausgehend vom Ursprungsort (10.48° Ost, 4.38° Nord) ein Kostendistanzraster (mit dem Ursprungsort als Zielpunkt) berechnet. Daraus konnte darauf das benötigte Flussrichtungsraster abgeleitet werden. Weil diese umständliche Variante allerdings nicht allzu gut funktionierte, wurde nach einer anderen Methode gesucht.
Schlussendlich wurde *Hydroflow* (UERJ 2007) verwendet, eine kleine, frei zugängliche Applikation der Universität von Rio de Janeiro zur Bestimmung der Flussordnungen. *Hydroflow* benötigt nur ein Shapefile der zu bestimmenden Ströme sowie ein weiteres Hintergrund-Shapefile, welches den limitierenden Rahmen des Untersuchungsgebiets darstellt. Dazu wurde ein einfaches Shapefile mit den Konturen des afrikanischen Kontinents verwendet. Voraussetzung für die Ordnungsbestimmung in *Hydroflow* war zudem, dass die bis dahin zusammenhängenden Pfade zuerst in einzelne Liniensegmente aufgeteilt wurden. Dies wurde mittels *Feature to Line* durchgeführt.

### Erstellung von Netzwerken und Distanzberechnung

In *gdistance* ist es zwar möglich, mittels der Funktion *costDistance()* eine Matrix mit Werten für die einzelnen Kostenpfade auszulesen. Dabei handelt es sich jedoch um Widerstandswerte, welche in einer undefinierten Einheit vorliegen. Diese können somit nur untereinander verglichen werden.
Um zusätzlich die Distanzen für die einzelnen Flussordnungen bestimmen zu können, wurde in *ArcGIS* ein Network Dataset erstellt. Die LCP-Shapefiles wurden daraufhin in Netzwerke umgewandelt.

Zur Berechnung von Netzwerkdistanzen gibt es im *ArcGIS Network Analyst* verschiedene Optionen. *OD-Cost Matrix* ermöglicht eine schnelle Berechnung der Distanzen, welche als Attributtabelle ausgelesen werden können. Interessieren nicht nur die Werte der Distanzen, sondern auch deren Richtungen, kann dazu *Closest Facilities* benutzt werden. Mit diesem kann das Netzwerk inklusive der Distanz-Attributtabelle als Shapefile ausgelesen werden.

So konnte sehr effizient der kürzeste Pfad zwischen allen einzelnen Punkten auf dem Netzwerk berechnet werden. Dies ergab bei den 121 Punkten insgesamt 7260 verschiedene Distanzen (=121*120/2). Davon mussten jedoch später nochmals einige ausgeschieden werden. Zusätzlich wurde ebenfalls mit dem *Network Analyst* die Anzahl der Knoten zwischen allen einzelnen Punkten berechnet. Diese topologi-

schen Distanz basieren im Prinzip auf demselben Ansatz wie die zuvor berechneten Shreve-Ordnungszahlen.
Die erhaltenen Punkte wurden aus *ArcGIS* ausgelesen und in Microsoft Access als Crosstab-Query dargestellt. Auf diese Weise konnten sie danach direkt in *R* zur Bildung hierarchischer Cluster eingelesen werden.

### 4.3.2 Aufbereitung der genealogischen Daten

Um die verschiedenen Bantusprachen bezüglich ihrer Verwandtschaft grafisch darstellen zu können, wurden bereits vorliegende Baumstrukturen zur Sprachfamilie benötigt.

#### Glottolog-Bantu-Sprachdaten

Als Quelle dienten hierbei die auf der Glottolog-Seite frei zur Verfügung gestellten Sprachdaten. Die benötigten genealogischen Daten zur Narrow-Bantu Sprachfamilie konnten dabei ganz einfach mit der *read.newick*-Funktion aus dem *R*-Paket *phytools* direkt aus dem Internet in *R* eingelesen werden.
Der phylogenetische Baum von Glottolog für die „engere Bantusprachfamilie“ besteht aus 3250 „Tips“ sowie 1374 „Nodes“. Erstere stehen für die am Ende der Baumstruktur liegenden einzelnen Sprachen und Dialekte, sie stellen also die „Blätter“ des Baumes dar. Letzere bilden Übergruppen der Sprachen, beziehungsweise die „Baumäste“ oder Knoten (Abb. 9). Die Längen der Äste des Phylogramms widerspiegeln die evolutionäre Zeit (UCMP 2014).

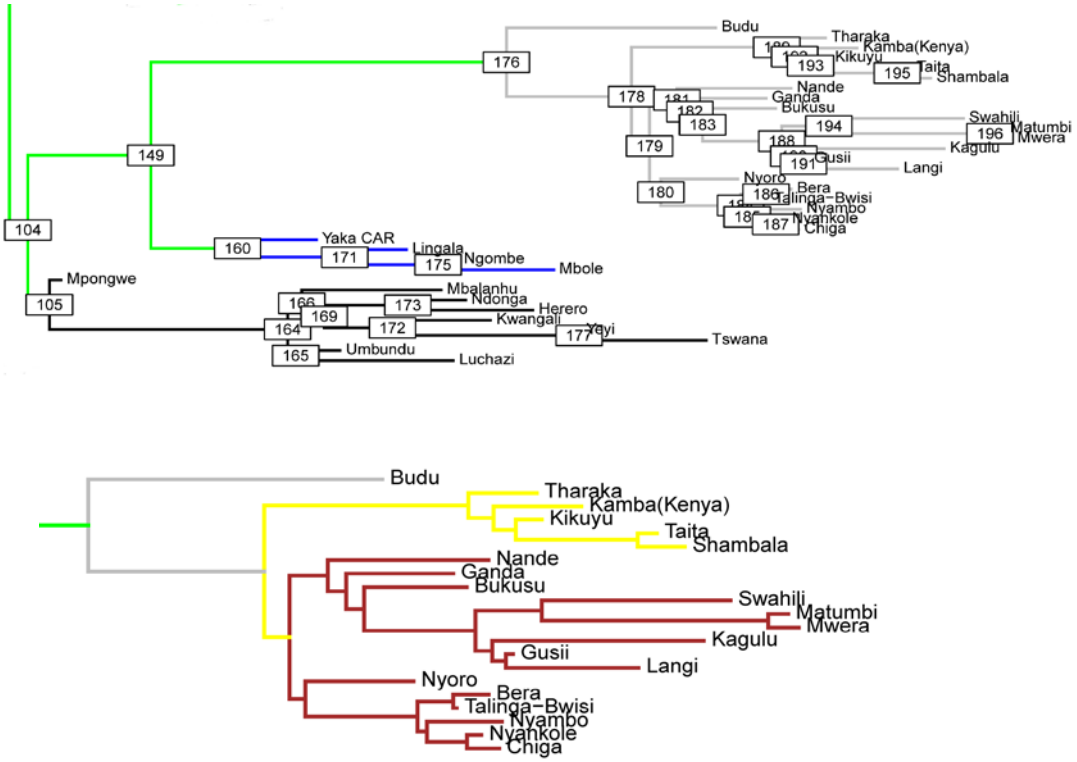



*Abbildung 9: (oben) Ausschnitt eines Dendrogramms mit Nummern für die Nodes (in Boxen), ausgeschriebenen Tips sowie farbig markierten Stämmen; (unten) Vergrösserung des Stammes ab Node 176 und neue farbliche Einteilung der Stämme auf tieferem Level (aus LCP-Distanzen generierter NJ- Baum); Eigene Darstellung*

Bei der ver

wendeten Bantu-Baumstruktur sind zusätzlich auch noch die Längen der einzelnen Äste vermerkt, diese wurden jedoch bloss mithilfe der Anzahl Knotenpunkte berechnet.

Angesichts der kaum überschaubaren Menge von Tips im Glottolog-Baum war es notwendig, davon zuerst eine Teilmenge zu entnehmen, welche dann grafisch dargestellt werden konnte. Um eine Vergleichbarkeit mit den aus den Kostenberechnungen erhaltenen Baumstrukturen zu ermöglichen, wurde versucht, die im Sprachdatensatz von Nichols & Bickel (2009) enthaltenen Sprachen grafisch darzustellen. Diese Aufgabe erwies sich im Folgenden als sehr umständlich.

## Übereinstimmung zwischen den Sprachdatensätzen

Da die im Bickel enthaltenen Sprachen mit dem ISO-639 Sprachcode versehen sind, während für die von der Glottolog entnommenen Seiten ein eigener Glottolog-Code vorliegt, mussten diese Codes miteinander verbunden werden. Dazu wurde von der Glottolog-Seite eine Liste mit allen Sprachen herunter geladen, welche mittels *Access* auf übereinstimmende Sprachen aus dem LCP-Datensatz untersucht und anschliessend zusätzlich mit dem glottolog-Code versehen wurden. Die Sprachen, bei welchen aufgrund unterschiedlicher Schreibweise keine Übereinstimmung gefunden wurde, wurden manuell auf der Glottolog-Seite mithilfe ihres ISO-Codes gesucht. Dabei konnte im Zweifelsfall auf der Glottolog-Seite gleich die geografische Lokalität angeschaut und mit den geographischen Angaben aus dem LCP-Datensatz überprüft werden.
4 Punkte des Datensatzes mussten gleich ausgefällt werden, da keine übereinstimmenden Werte gefunden werden konnten. Es wurden folgende Sprachen entfernt: Chichewa, Ciluba, Kongo und Fiote. Im Folgenden fielen aber noch weitere weg.
Die Zuordnung via Glottolog-Code über die Tips funktionierte nur gerade für 39 der ursprünglich 121 Sprachpunkte (respektive 117). Da noch viele Sprachen aus den genealogischen Daten ebenfalls mit einem ISO-Code versehen waren, wurden auch noch diese verbunden, was bei 39 Treffern 5 weitere Punkte ergab. Die Namen der Sprachen wurden daraufhin angepasst. Von den ehemals 121 Sprachpunkten konnten somit nur gerade 44 unter den Spitzen der Baumstruktur gefunden werden.
Des Weiteren wurden auch die Knotenpunkte des Baumes untersucht, wobei eine erstaunlich hohe Anzahl von 66 Punkten übereinstimmte. Von diesen fielen 4 weg, da sie doppelt vorhanden waren. Somit blieben noch 62 Treffer.
Die unterschiedliche Einordnung in den Daten von Nichols & Bickel (2009) und den Glottolog-Daten (2013) ist dadurch zu begründen, dass Glottolog auf eine sehr tiefe Ebene geht. Daraus leitet sich auch die sehr hohe Anzahl von 3250 Tips ab. Die Mehrheit davon sind bloss noch Varietäten der Sprachen respektive Dialekte. Als Folge davon liegen viele der noch eigenständigen Sprachen in dieser Datenbank bereits eine Ebene höher als „Nodes“ vor. Analog zu diesem Fall zeigte sich auch bei der Überprüfung der Glottolog-Daten mit einem weiteren Datensatz von Hammarström (2012) dasselbe Problem.

## Extraktion der Tips

Das Package *ape* bietet zum Ausfällen einzelner Tips die Funktion drop.tip() an. Mit dieser können durch eine leichte Anpassung auch umgekehrt bestimmte Tips mittels folgendem Befehl aus einem Satz ausgewählt werden.

- New_tree <- drop.tip(tree, tree$tip.label[-match(x, tree$tip.label)])

Hierbei steht x für eine Liste oder einen Vektor der genauen Namen der Tips, welche behalten werden sollen. Dies machte es relativ einfach den genealogischen Baum für die übereinstimmenden Tips zu generieren.
Für die Nodes existiert (ebenfalls aus *ape*) die in etwa umgekehrte Funktion *extract.clade()*, mit deren Hilfe einzelne Nodes aus einem Baum ausgelesen werden können. Jedoch kann die extract.clade() Funktion nur für einzelne Nodes verwendet werden. Dazu kam das Problem, dass die Nodes noch nicht die unterste Ebene darstellen, sondern mehrere Tips enthalten. Somit war eine Vergleichbarkeit zwischen den Nodes aus den Glottolog-Daten und der Tips aus den LCP-Punkten sowieso nicht direkt gegeben. Um trotzdem möglichst viele LCP-Punkte mit den genealogischen vergleichen zu können, wurden die Nodes umständlich in Tips überführt.
Dazu wurde zuerst überprüft, ob es bei der Baumdarstellung einen Unterschied macht, welche der Tips eines jeweiligen Nodes entnommen wurden. Abgesehen von 3 Fällen, wo es sich um Nodes handelte, welche nochmals in weitere Nodes untergliedert waren (hier gab es kleine Längendifferenzen), resultierte kein Unterschied. Deshalb wurden im Folgenden zufällige Tips aus den Nodes gewählt, und mit den entsprechenden Namen der Nodes versehen. Neben den zuvor erwähnten 3 weiter verästelten Nodes mussten noch 4 weitere ausgefällt werden. Bei diesen handelte es sich nun (aufgrund ihrer neuen (Nodes-)Namen um bereits als Tips vorliegende Sprachen. Somit konnten von den 44 Tips und 62 Nodes schlussendlich 99 (106-7) verwendet werden (Abb. 10). Dabei wurden auch alle Namen aufeinander abgestimmt.

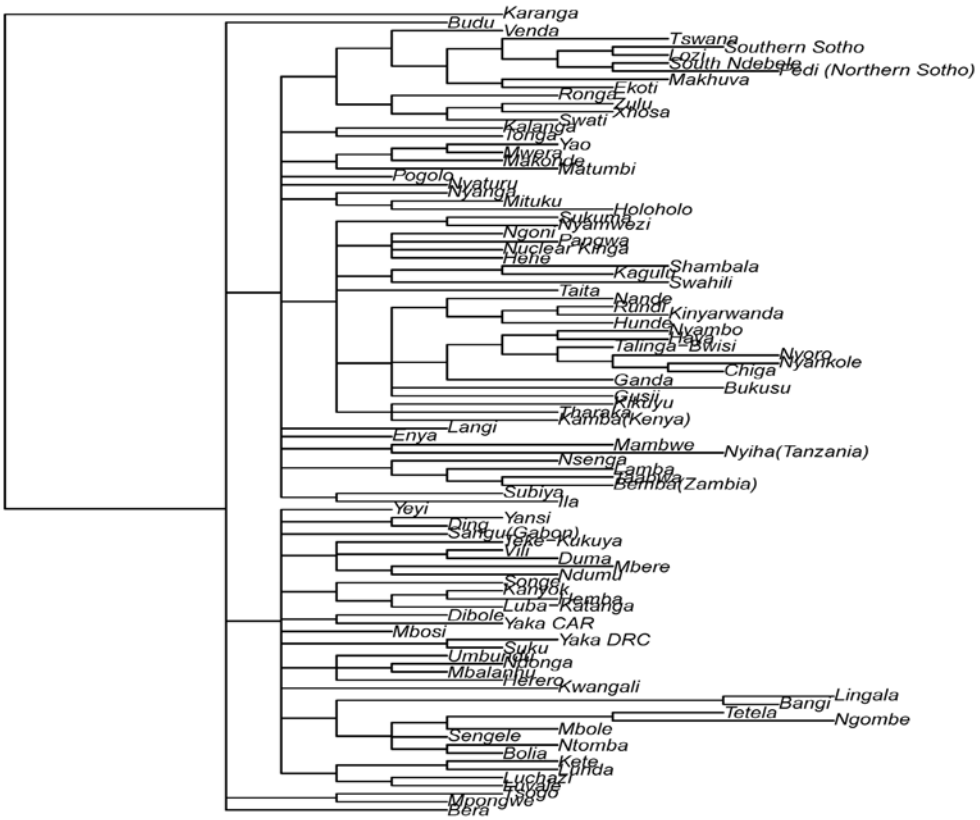



*Abbildung 10 Mittels NJ-Algorithmus erstellter genealogischer Baum für 99 Sprachpunkte aus Glottolog-Datensatz (Nordhoff 2013); Eigene Darstellung*

### 4.3.3 Bildung hierarchischer Cluster

In diesem Kapitel werden die aus den LCP-Berechnungen erhaltenen Distanzen zwischen den einzelnen Sprachpunkten mit genealogischen Daten verglichen. Um dies zu ermöglichen, mussten die Kostendistanzen in hierarchischen Bäumen abgebildet werden. Hierzu bieten sich in *R* verschiedene Methoden der hierarchischen Clusteranalyse an. Im nächsten Kapitel wird deshalb kurz das Prinzip der hierarchischen Cluster erläutert.

#### Hintergrund Cluster Analyse

Cluster Analysen erlauben multivariate Analysen zur Bildung von Gruppen, respektive Clustern von ähnlichen Untersuchungsobjekten. Dabei gibt es jedoch unterschiedliche Definitionen der „Ähnlichkeit“ sowie verschiedene Methoden der Clusterbildung, welche zu berücksichtigen sind.
Im Folgenden werden die unterschiedlichen Typen von Ähnlichkeiten vorgestellt.

##### Ähnlichkeit, Unähnlichkeit und Distanz

Als Ähnlichkeit bezeichnet man das Verhältnis aller gemeinsamen Attribute von zwei Objekten zu jenen, welche nur bei einem Objekt vorkommen. Eine totale Übereinstimmung ergibt einen Wert von 1.0, der grösstmögliche Unterschied 0.0. Zur Bestimmung der Ähnlichkeit liegt eine grosse Anzahl von Untersuchungsmethoden vor, dabei wird aber dasselbe Grundkonzept angewandt.
Unter Unähnlichkeit oder Verschiedenheit versteht man das Gegenteil zur Ähnlichkeit, entsprechend kann diese mathematisch sehr einfach mit „1 - Ähnlichkeit“ definiert werden. Die dritte Methode zur Festlegung der Ähnlichkeit, die Distanz, bedient sich der geometrischen Nachbarschaftsbeziehung der untersuchten Objekte. Es bildet also ihre räumliche Beziehung ab (MSU 2013).

Entsprechend handelt es sich bei der statistischen und der evolutionären Distanz um zwei unterschiedliche Dinge. Während die statistische physisch respektive geometrisch erfassbar ist, handelt es sich bei letzterer um etwas Abstraktes. Sie kann bloss geschätzt werden (Paradis 2006). Dies zeigt, dass auch der Vergleich der aus den Kostenpfaden berechneten Dendrogramme mit den genealogischen Glottolog-Daten nicht ganz unproblematisch ist.

Da an dieser Stelle aber die räumlichen Distanzen zwischen den einzelnen Sprachpunkten untersucht werden sollen, interessiert vor allem der geometrische Ähnlichkeitstyp. In R-Statistik werden Distanzen mit der *dist-Funktion* berechnet, wobei zwischen euklidischer, Manhattan- sowie binärer Distanz gewählt werden kann. Diese geben Distanzmatrizen aus, welche für die Bildung von Clustern weiterverwendet werden können. Des Weiteren erlaubt auch die Funktion *daisy()* die Bildung von Ähnlichkeitsmatrizen.

#### Cluster Algorithmen

Cluster Algorithmen werden in zwei Haupttypen unterteilt. Man unterscheidet zwischen hierarchischen und fixierten oder auch partitionierenden Clusterverfahren.
Zur hierarchischen Clusteranalyse gehört eine Gruppe distanzbasierter Clusteranalysen, die zur Bestimmung der (hierarchischen) Struktur von Datensätzen verwendet werden. Dabei werden zwei verschiedene Ansätze verwendet: ein divisiver, wo zunächst alle Objekte einem Cluster zugeordnet und erst dann weiter aufgeteilt werden

(„Top-Down-Verfahren“) und ein agglomerativer, bei dem ein „Bottom-up-Verfahren“ angewendet wird.
Bei den fixierten oder partitionierenden Verfahren hingegen wird die Nummer von Clustern a priori durch den Analysten bestimmt. Die Cluster sind dabei nicht hierarchisch angeordnet, sondern bilden vielmehr eine Unterteilung der Daten. Diese hängt von der Homogenität innerhalb sowie der Heterogenität ausserhalb der Gruppen ab. Das bekannteste Beispiel dieses Typus ist wohl die *k-means* Methode.
Zur Untersuchung der berechneten Kostendistanzen wurden hierarchische Clusterverfahren angewendet, welche einen Vergleich mit den vorhandenen genealogischen Hierarchien ermöglichen (MSU 2013).

### Bildung hierarchischer Cluster in R

Die bekannteste Methode hierarchischer Clusteranalysen in *R* stellt *hclust()* dar. Dieses kann sowohl die Ähnlichkeit als auch die Unähnlichkeit zwischen den Objekten verwenden, eine Unähnlichkeitsmatrix ist aber eher vorteilhaft.
Nachdem anfangs *hclust*() zur Darstellung der Distanzbäume verwendet worden war, wurde im Folgenden jedoch die *nj* respektive *Neighbour-joining* Methode wiederum aus *ape* benutzt. Diese hat zum einen den Vorteil, dass sie die unterschiedlichen Astlängen im Dendrogramm wiedergibt. Da damit zudem Bäume der Klasse „phylo“ gebildet werden, erlaubt dies eine besseren Vergleich mit den genealogischen Daten. Diese liegen ebenfalls in der Klasse phylo vor. Der sehr populäre *Neighbour-joining* Algorithmus ist generell sehr schnell in der Berechnung. Er berechnet einen Näherungswert für die minimale Evolution zwischen den Variablen und stellt eine bottom-up Cluster-Methode dar (Holder & Lewis 2003).
Zusätzlich zu den mittels der Kostenpfade gebildeten Dendrogramme wurden 2 euklidische Vergleichscluster erstellt. Diese erfolgten mittels der *r.dist()* Funktion aus *R*, welche auf sehr einfache Art die Berechnung grosskreisbasierter Distanzen erlaubt. Dabei wurde zum einen ein Baum aus den direkten euklidischen Distanzen erstellt, zum anderen einer aus den euklidischen Distanzen und dem LCPA-Ursprungspunkt O (10.48° O, 4.38° N).

## Vergleich der Dendrogramme

Die verschiedenen Dendrogramme sollten nun noch miteinander verglichen werden. Eine methodische Überprüfung der Bäume ist notwendig, da für ein gegebenes Format eines Baumes mehrere verschiedene Darstellungen resultieren können (Paradis 2006).

Um die Dendrogramme zu vergleichen, wurde im Folgenden die Methode *consensus()* verwendet. Dabei handelt es sich um eine Methode, mit der überprüft werden kann, welche Nodes in verschiedenen Bäumen gemeinsam vorkommen (Holder & Lewis 2003). Mit *consensus()* wurde überprüft, ob die aus den LCP-Distanzen und Topologien generierten Dendrogramme besser mit dem phylogenetischen Baum übereinstimmen, als die parallel dazu berechneten Bäume aus den euklidischen Distanzen. Da jedoch die Korrelation bei allen Bäumen sehr gering war, wurden noch weitere Untersuchungen gemacht. Beim Consensus-Baum handelt es sich um eine strikte sprachliche Methode zur Überprüfung der Ähnlichkeit zwischen den Dendrogrammen (Holder 2003). Diese berücksichtigt jedoch kaum die geographisch relevanten Aspekte der LCPA respektive den hohen Unsicherheitsfaktor bei der Untersuchung. Aus diesem Grund wurden noch weitere Analysen der Resultate mit einem verstärkt geographischen Ansatz gemacht.Dazu wurde der phylogenetische Baum von seinem „evolutionären Ursprung“ ausgehend in die einzelnen Stämme

zerlegt und diese unterschiedlich eingefärbt. Die Einfärbung erfolgte mit der *paintSubTree()*- und *plotSimmap()*- Funktion aus *phytools*. Je tiefer die Hierarchie, desto höher wird die Anzahl der eingeteilten Stämme. Dabei erfolgte die Einteilung in Stämme bis in die dritte Ebene respektive Hierarchie, wo beim phylogenetischen Baum bereits 18 Gruppen vorliegen. Die eingefärbten Stämme beziehungsweise die einzelnen Punkte an deren Ende wurden nun auf der geographischen Karte entsprechend koloriert.
Auf diese Weise konnte für das farbliche Muster der Sprachpunkte beurteilt werden, welche räumlichen Verbindungswege dieses besser erklären. Zur Darstellung der euklidischen Distanzen wurden zu diesem Zweck konzentrische Kreise um den Ursprungspunkt der Kostenpfade gelegt.

# 5 Resultate und Interpretation

In diesem Kapitel werden die Resultate der verschiedenen LCPA präsentiert. Weil beim Vorgehen zur Bestimmung der Kostenpfade aufgrund des geringen Vorwissens ein exploratives Vorgehen gewählt wurde, sind die Resultate auch untrennbar mit der Interpretation verbunden. Aus diesem Grund ähnelt der hier vorgestellte Teil etwas einem Logbuch, bei dem fortwährend der Stand der Analyse festgehalten und daraus das weitere Vorgehen abgeleitet wird.

## 5.1 Berechnung der LCPA

### 5.1.1 Untersuchungen Nordteil

Der erste Teil der Untersuchungen bezieht sich stets auf den nördlichen Ausschnitt des Bantu-Ausbreitungsgebiets. Die Grafiken sind dabei bewusst einfach gehalten. Wie aus der y- und x-Achse zu entnehmen ist, erstreckt sich der Ausschnitt von 10° Nord bis 10° Süd und 7° Ost bis 41° Ost (Abb. 11). Im Hintergrund ist die nach Kosten eingefärbte Vegetation zu sehen.

#### Variation von Einzelpfaden

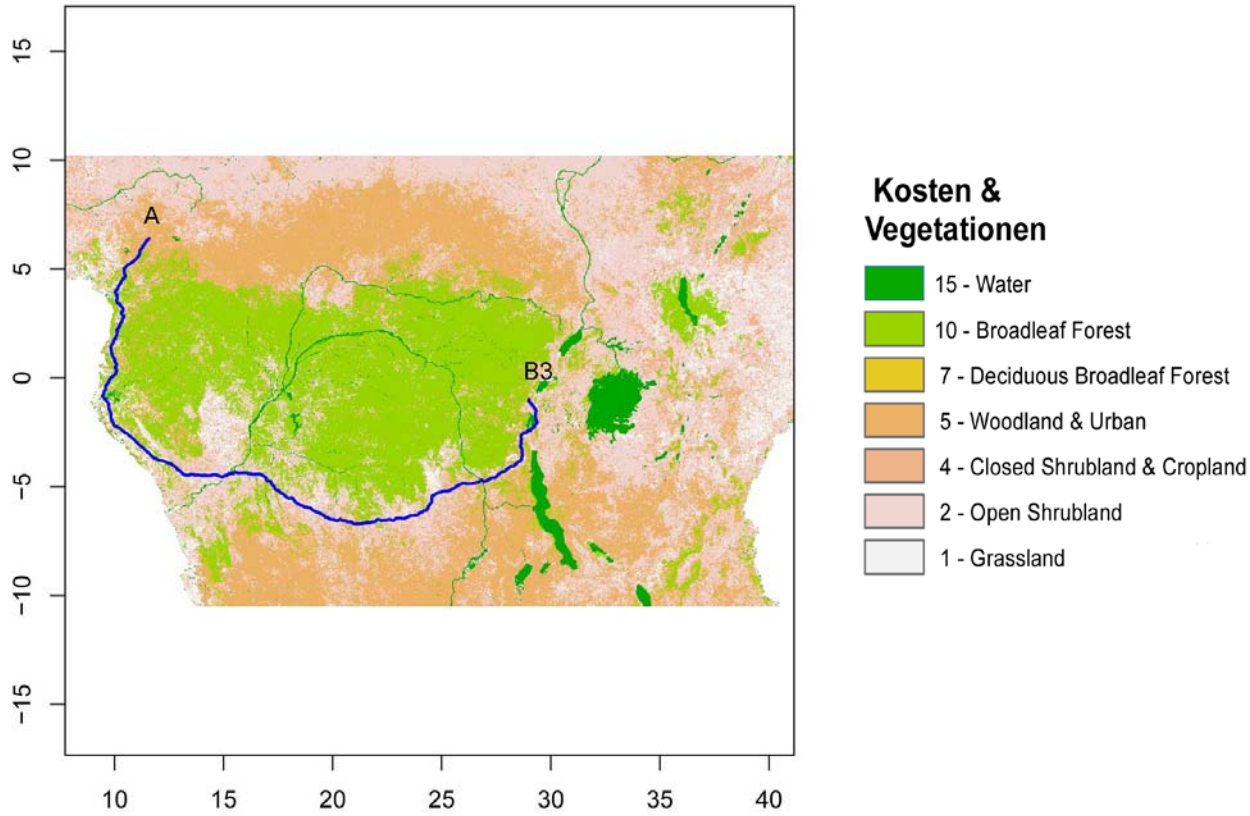



*Abbildung 11: Kostenpfad vom Punkt A (11.6°O, 6.4°N) zum Punkt B3 (28.83°O, 1.17°S) für Kostenfaktor Vegetation (Standardgewichtung)*

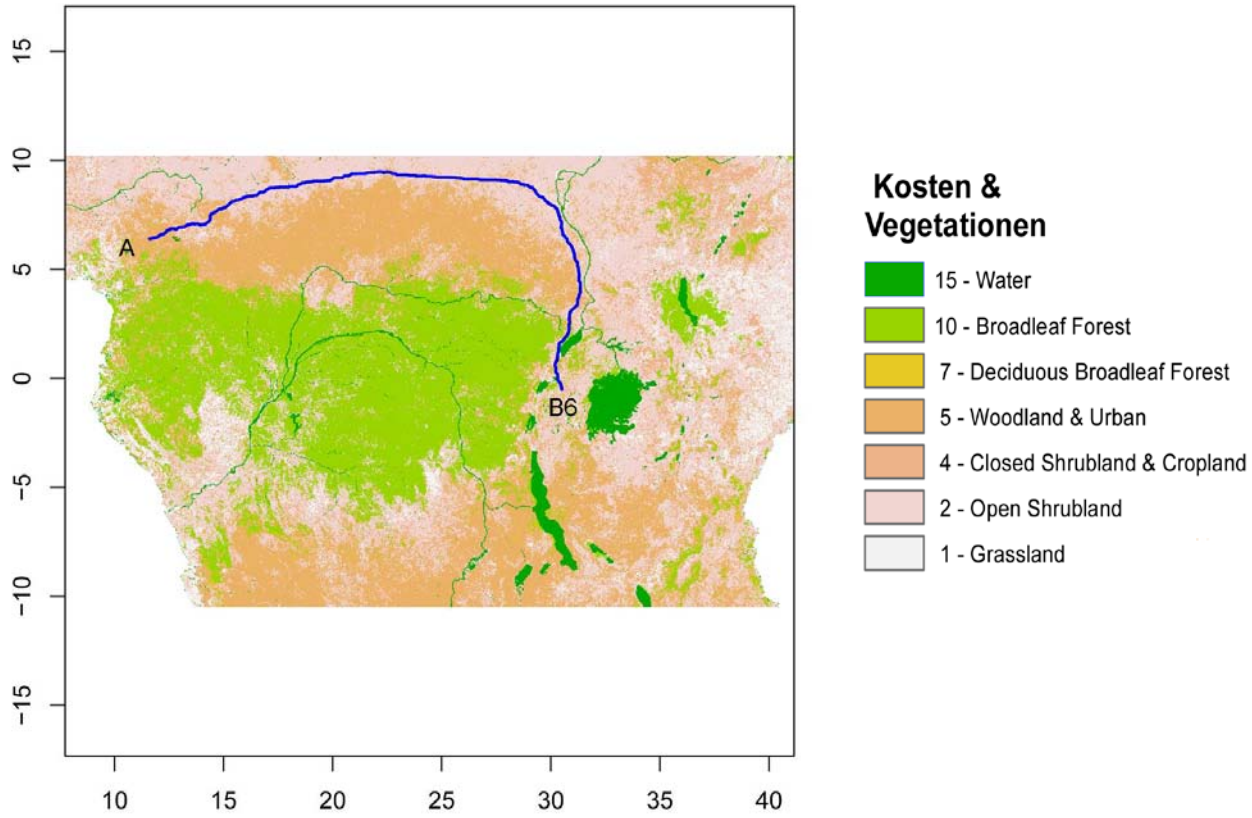



*Abbildung 12: Kostenpfad vom Punkt A (11.6°O, 6.4°N) zum Punkt B6 (30.5°O, 0.5°S) für Kostenfaktor Vegetation (Standardgewichtung)*

Bei den ersten Berechnungen wurden einzelne Kostenpfade zwischen einem möglichen Ursprungspunkt A (11.6° Ost, 6.4° Nord) und verschiedenen Sprachpunkten aus dem östlichen Teil des Gebietes erstellt.

Es wurden dabei bewusst Zielpunkte mit variierenden Breitengraden gewählt, um festzustellen, bei welchen Punkten ein östlicher beziehungsweise ein südlicher Pfad resultierte. Die in den Abbildungen 11 und 12 dargestellten Punkte B3, bzw B6 stellen diesbezüglich die äussersten Punkte dar. Wie auf den Grafiken zu erkennen ist, liegen die beiden Sprachen Hunde (Punkt B3, ISO hke) und Nyankole (Punkt B6, ISO: nyn) geographisch relativ nahe beinander. Auch wenn die tatsächliche Entfernung mit 1.8° noch immer fast 200 km beträgt, wird mit den beiden Datenpunkten ziemlich genau erfasst, ab welchem Ort die südliche bzw. nördliche Umgehung des Regenwalds mit weniger Kosten verbunden ist.

## Variation technischer Parameter

### Geokorrektur

In einem nächsten Schritt wurden technische Faktoren der LCPA auf ihren Einfluss untersucht. Dazu wurden in *gdistance* verschiedene Einstellungen variiert. Als erstes wurde die Berechnung der geokorrigierten Matrix weggelassen, um deren Einfluss zu erfassen. Zum Vergleich wurden die beiden Versionen übereinandergelegt (Abb. 13).

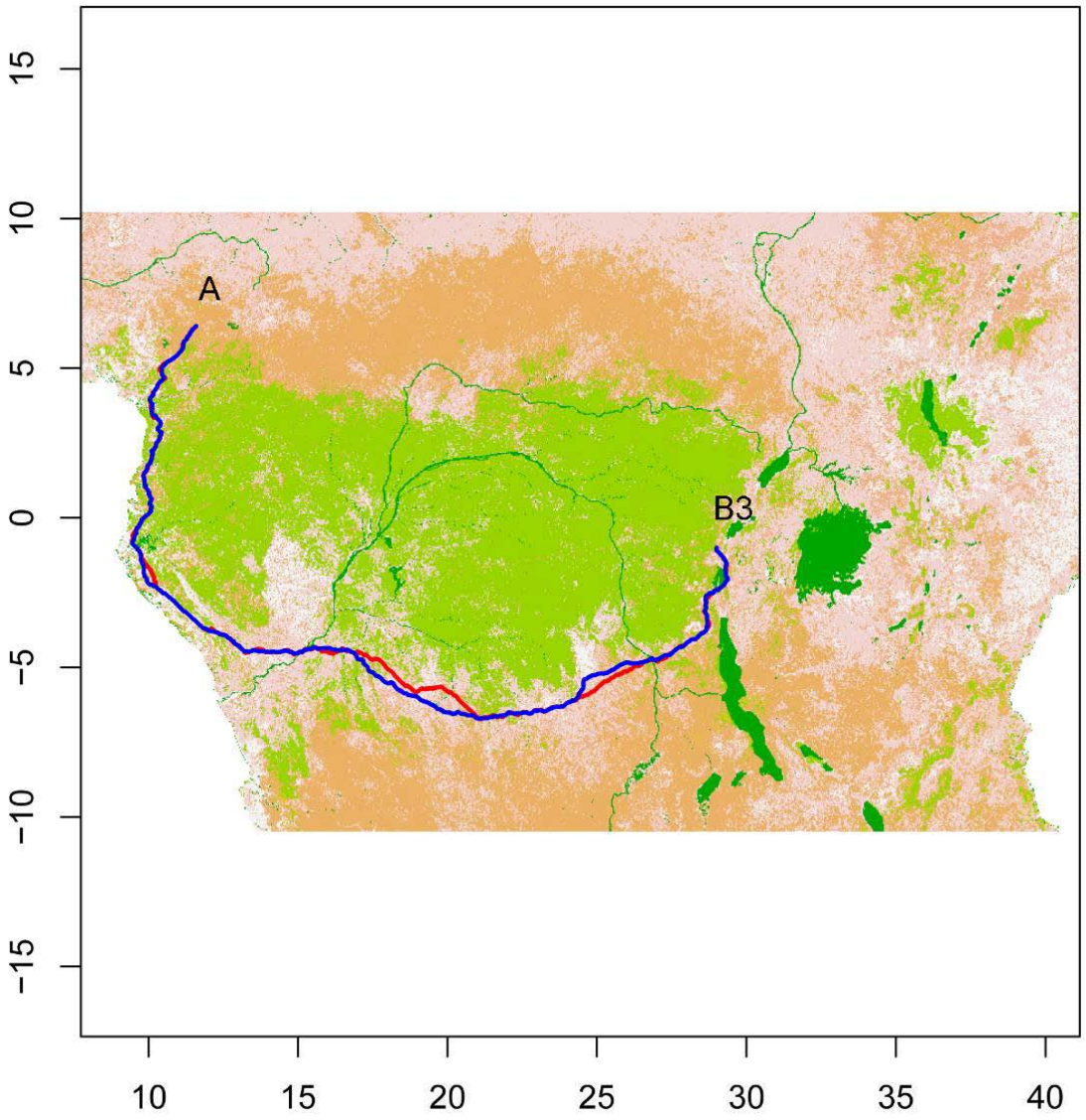



*Abbildung 13: Kostenpfad vom Punkt A (11.6°O, 6.4°N) zum Punkt B3 (28.83°O, 1.17°S) für Kostenfaktor Vegetation (Standardgewichtung) mit (blau) und ohne (rot) Geokorrektur*

Wie in Abbildung 13 zu erkennen ist, hat der nicht geokorrigierte, rote Kostenpfad einen etwas eckigeren Verlauf. Da aber diese Linie aufgrund des Reibungsfaktors Vegetation sowieso bereits eine eher grosskreisartige Kurve beschreibt, wird der Einfluss der Geokorrektur im Folgenden noch anhand eines weiteren Beispiels dargestellt (Abb. 14 und 15).

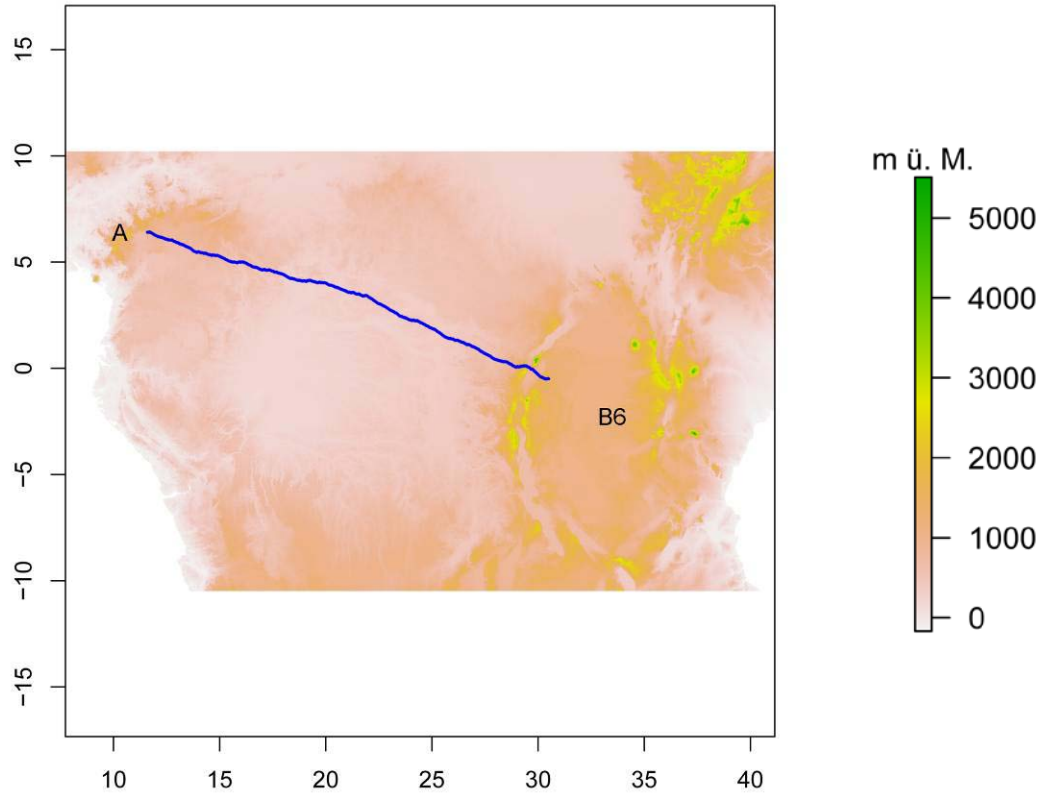



*Abbildung 14: Kostenpfad vom Punkt A (11.6°O, 6.4°N) zum Punkt B6 (30.5°O, 0.5°S) für Kostenfaktor Neigung ohne Geokorrektur*

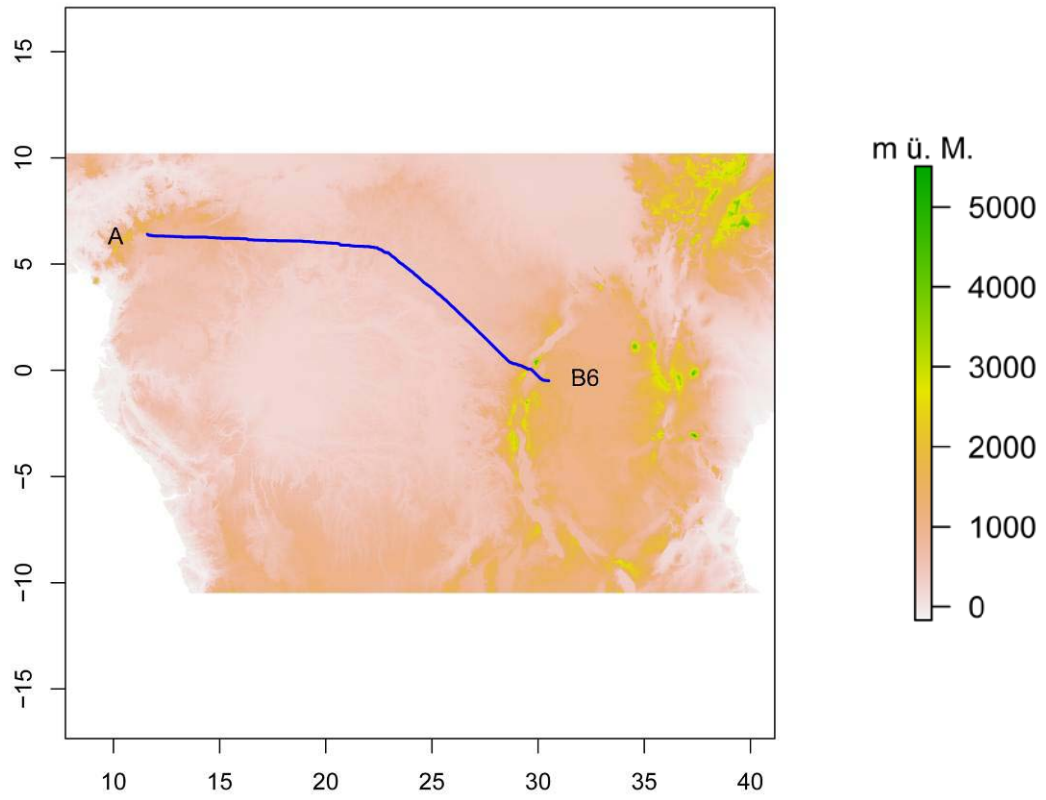



*Abbildung 15: Kostenpfad vom Punkt A (11.6°O, 6.4°N) zum Punkt B6 (30.5°O, 0.5°S) für Kostenfaktor Neigung mit Geokorrektur*

Diese Abbildungen 14 und 15 zeigen einen sehr deutlichen Einfluss der *geoCorrection*-Funktion. Während die von einer flachen Scheibe ausgehende Berechnung (Abb. 14) eine mehr oder weniger direkte Linie von A nach B6 darstellt, ist bei der korrigierten Version ein deutlicher Knickpunkt auszumachen (Abb. 15). Dies verdeutlicht die Wichtigkeit dieser Funktion bei der Ausdehnung des untersuchten Gebietes.
Der gerade Verlauf des Kostenpfades beim Faktor Neigung deutet überdies auf einen sehr geringen Einfluss dieses Faktors bezüglich der resultierenden Kostenpfade hin. Die Auswirkungen der Neigung auf die berechneten Wege wurden mittels weiterer Anwendungen genauer analysiert.

### Verwendete Nachbarschaft

Weiter wurde die in *gdistance* optionale Einstellung bezüglich der verwendeten Nachbarschaft geändert. Abbildung 16 zeigt den Unterschied zwischen der Berechnung bei einer 8-er (schwarz) und 16-er (rot) Nachbarschaft.

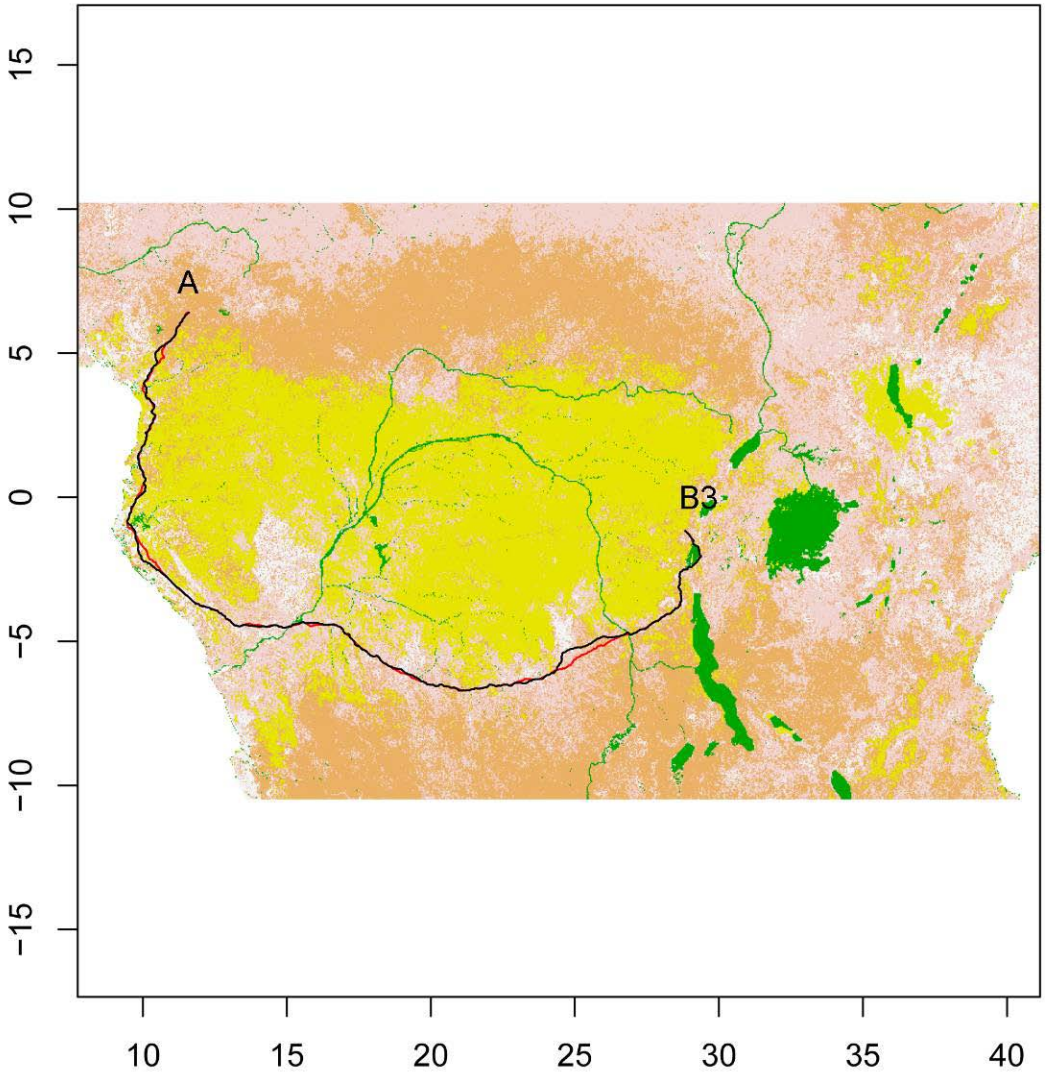



*Abbildung 16: Kostenpfad vom Punkt A (11.6°O, 6.4°N) zum Punkt B3 (28.83°O, 1.17°S) für Kostenfaktor Vegetation (Standardgewichtung) bei 8-er (schwarz) und bei 16-er (rot) Nachbarschaft.*

Auch zwischen den Kostenpfaden mit 8-er und 16-er Nachbarschaft können leichte Unterschiede erkannt werden. Letztere zeigen manchmal einen noch etwas ausgeglicheneren Verlauf. Es handelt sich hierbei jedoch um ziemlich kleine Unterschiede.

Der Einfluss der vergrösserten und verfeinerten Nachbarschaft ist einiges geringer als derjenige der Geokorrektur.
Weil die Verwendung der 16-er Nachbarschaft den Berechnungsaufwand und damit auch die benötigte Zeit nochmals deutlich steigert, wurde bei den weiteren Berechnungen auf diese Option verzichtet. Gerade auch im Hinblick zur Bestimmung von Multipunkte-Kostenpfaden wurde diese zeitliche Verlängerung als nicht lohnenswert erachtet (weitere Darstellung im Anhang).

#### Weitere Variationen

Wie im Kapitel 4 zur Auswahl der Faktoren erläutert wurde, ist es eher unwahrscheinlich, dass die Flüsse bei der Bantu-Migration als entscheidende Verbindungen genutzt wurden. Da dies aber in erster Linie eine logische Annahme darstellt und aufgrund des geringen Wissenstandes nicht mit Sicherheit gesagt werden kann, wurde in einigen Versuchen diese Möglichkeit dennoch berücksichtigt.
Weil aufgrund der Lage der Punkte eine Benutzung der Flüsse entgegen der Strömung mehr oder weniger ausgeschlossen werden konnte, wurden die Kosten für die Flüsse sehr einfach vorgenommen. Dazu wurde ihnen der Wert 1 für den tiefsten Reibungswiderstand zugeschrieben. Bei den Resultaten kamen wie erwartet keine Fälle vor, wo die Flüsse gegen den Strom verwendet wurden. So konnte diesbezüglich auf die Anwendung einer komplizierteren anisotropen Kostenoberfläche (vgl. Livingood 2012) verzichtet werden.

Die umgekehrte Gewichtung für die Flüsse hatte jedoch keine grossen Auswirkungen. Sie hätte bloss auf 4 zentral im Regenwald und nahe am Kongo liegende Sprachpunkte eventuell einen Einfluss gehabt. Dies sind die Sprachen Bangi (bni), Bolia (bli), Ntomba (nto) und Sengele (szb). So wurden die Flüsse bei den weiteren Anwendungen wieder als Faktoren mit hoher Reibung verwendet.

Eine gewisse Überprüfung der Stabilität der LCPA stellt auch die Wahl leicht variierender Anfangspunkte dar (Herzog 2013). Im Folgenden wurden deshalb 4 weitere Punkte aus dem Sprachdatensatz von Nichols & Bickel (2009) als Ausgangspunkte der LCPA bestimmt. Dabei handelte es sich um die Sprachen Bafia (ksf), Ewondo (ewo), Duala (dua) und Londo (bdu). Der dadurch erreichte Effekt ist jedoch derselbe, der durch die leicht variierenden Endpunkte erzielt wurde. Je nördlicher der Startpunkt, desto weiter südlich werden die östlich gelegenen Zielpunkte über die nördliche Umgehung des Regenwalds erreicht (siehe Anhang).

Da die Wahl unterschiedlicher Anfangspunkte bei einem derart grossen Gebiet sowieso unerschöpflich ist, wurde darauf ein fester Anfangspunkt für alle weiteren Anwendungen bestimmt. Wie in Kapitel 4 genauer beschrieben, wurde dieser als räumlicher Mittelpunkt der im Ursprungsgebiet liegenden Sprachpunkte definiert. Für die folgenden Kostenpfadberechnungen wurden – abgesehen von den zur Bestimmung des Ursprungs verwendeten – alle Punkte gleichzeitig in die LCPA miteinbezogen. Dies ergab 85 mit Kostenpfaden verbundene Sprachen.

### DEM versus Vegetation

Da bei bereits erstellten Kostenpfaden ersichtlich geworden war, dass der Einfluss der Neigung eher gering zu sein scheint, wurde dies genauer untersucht (Abb. 17).

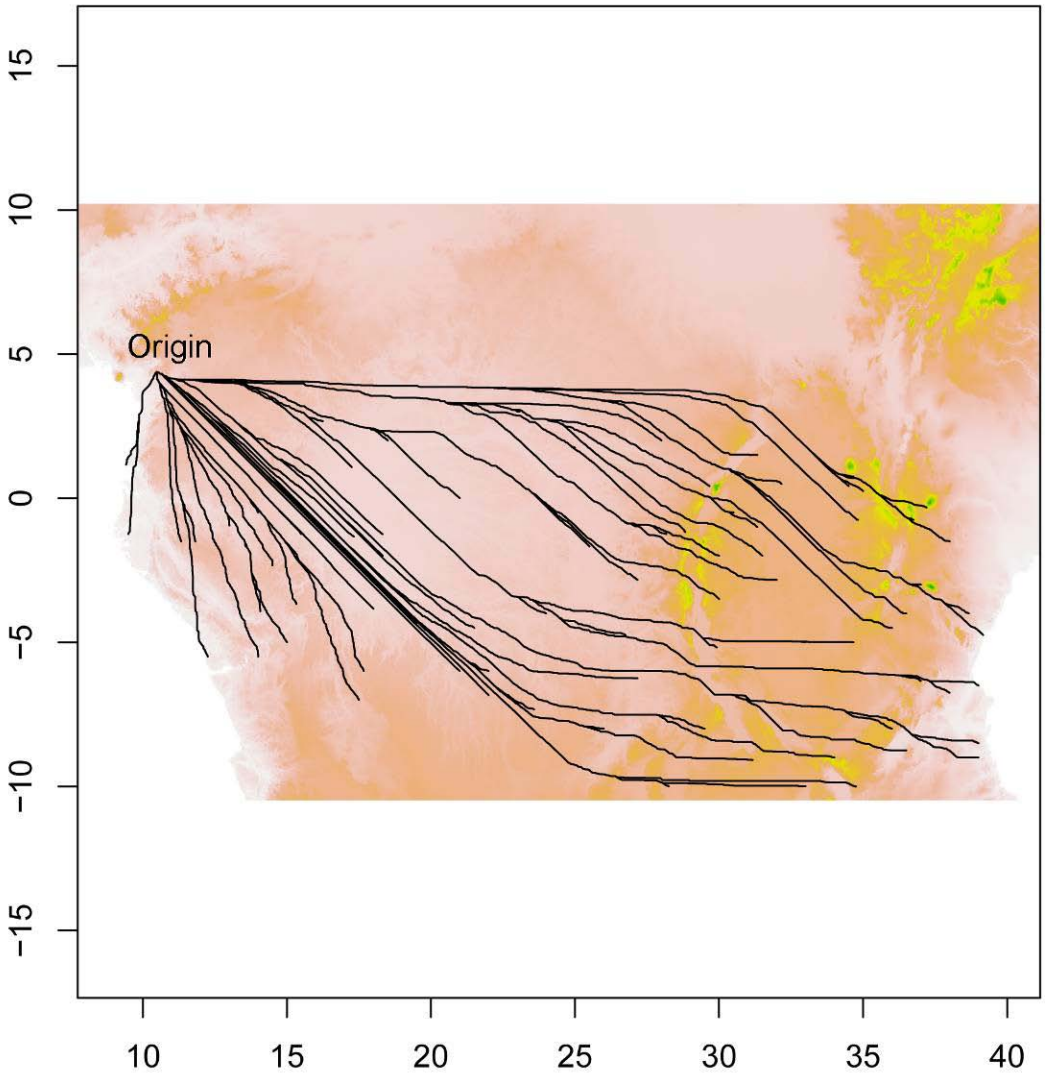



*Abbildung 17: 85 Kostenpfade ausgehend vom Punkt O (10.48°O, 4.38°N) für Faktor Neigung*

Die Multipunkte Kostenpfade bestätigten von Abbildung 17 die zuvor aufgekommenen Zweifel am Einfluss dieses Faktors auf die Kostenpfade. Die Pfade beschreiben abgesehen von minimen Abweichungen mehr oder weniger alle einen relativ geradlinigen Verlauf zu den Zielpunkten entlang der acht Richtungen der 8-er Nachbarschaft. Wie schon zuvor der Vergleich mit der nicht geokorrigierten Version zeigte, stellen die hierbei deutlich sichtbaren Knickpunkte den aufgrund der Erdkrümmung notwendigen Ausgleich dar.

Als logische Konsequenz des geringen Einflusses des Faktors Neigung ergab auch die Kombination mit der Vegetation (hier mit 1:1 Gewichtung) nur eine „begradigte“ Version der Kostenpfade um den Regenwald (Abb. 18). Der Einfluss der Hangneigung ist daran zu erkennen, dass die Gewässer hier teilweise direkt durchquert werden. Dies ist beispielsweise beim länglichen Tanganjikasee südwestlich des Victoriasees der Fall (rot eingekreist).

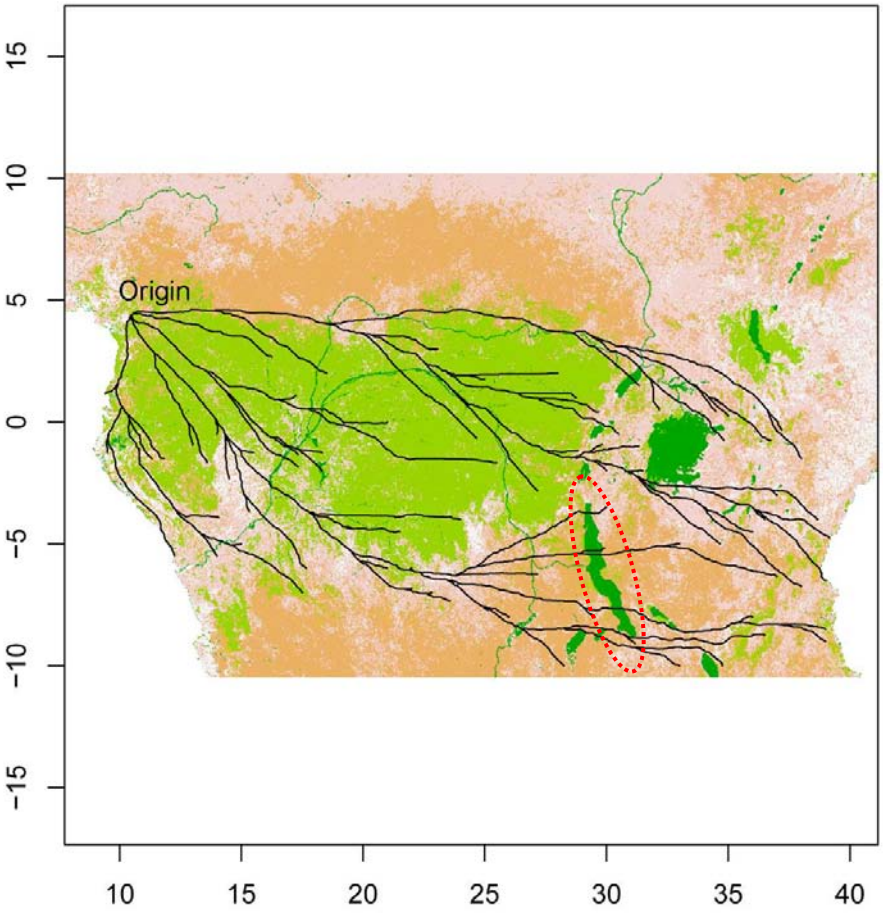



*Abbildung 18: 85 Kostenpfade ausgehend vom Punkt O (10.48°O, 4.38°N) mit 1:1 Kombination von Neigung und Vegetation*

Aus diesem Grund wurde der Faktor Neigung bei den weiteren LCPA-Berechnungen weggelassen. Dies brachte natürlich auch Vorteile bezüglich des benötigten Arbeitsspeichers für die LCPA mit sich. So wurden die Berechnungen nun immer für den gesamten Ausschnitt gemacht.

## 5.2 Analyse der Kostenpfade

### 5.2.1 Shreve-Ordnungen

Im Folgenden sind die Resultate für die drei in Kapitel 4 beschriebenen der insgesamt 10 Kalibrierungen dargestellt. Dazu wird zunächst zur besseren Verständlichkeit noch einmal die Karte mit den Vegetationstypen dargestellt sowie die Werte für die 3 Kalibrierungen wiedergegeben. Über die Vegetationskarte wurden als anschauliches Beispiel gleich die Kostenpfade für die Kalibrierung V1 gelegt.

Die Interpretation der Kostenpfade der drei gewählten „Kalibrierungen“ erfolgt mithilfe der dazu bestimmten Shreve-Flussordnungen (auch Link Magnitude genannt). Diese ermöglichen eine gute Vergleichbarkeit zwischen den Pfaden (Abb. 20 bis 22).

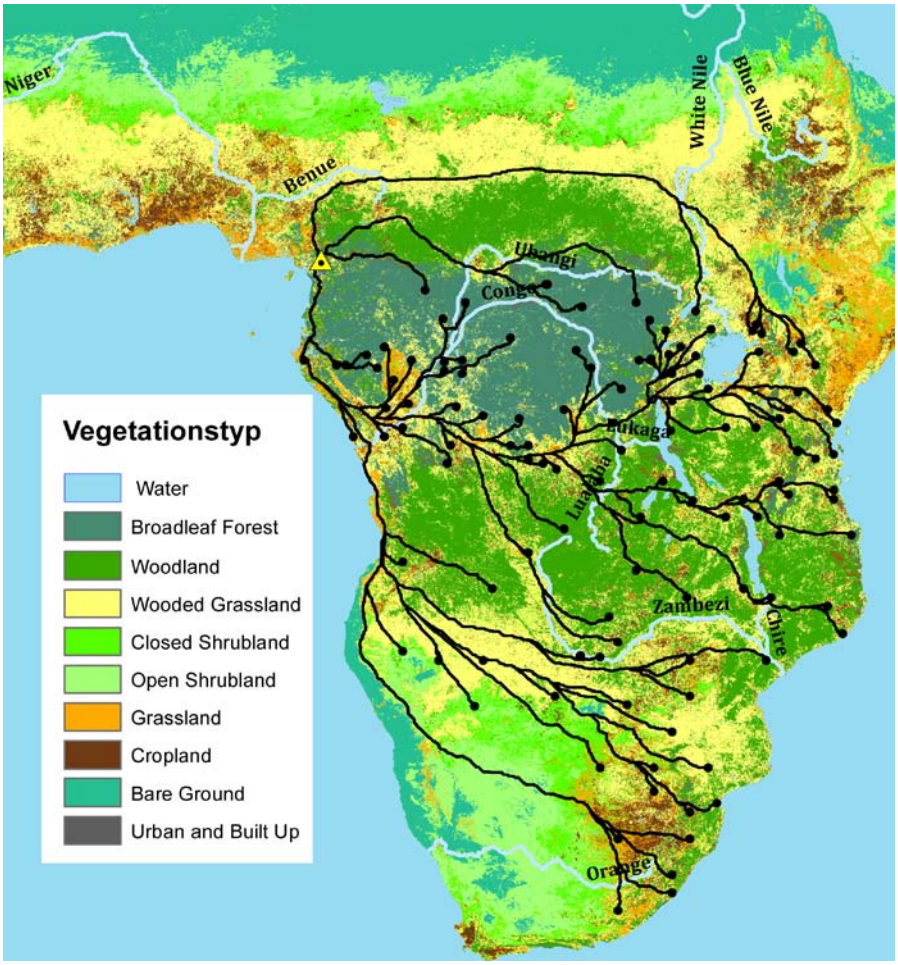



*Abbildung 19: Kostenpfade bei Standardkalibrierung V1 für gesamtes Gebiet mit 121 Zielpunkten (Anfang bei 10.48° O, 4.38° Nord); Vegetationskarte als Hintergrund (UNEP 2000)*

| Typ | V1 | V4 | V10 |
|---|---|---|---|
| Wooded Grassland | 2 | 2 | 1 |
| Woodland | 5 | 4 | 3 |
| Evergreen Broadleaf Forest | 10 | 6 | 10 |
| Grassland | 1 | 1 | 1 |
| Open Shrubland | 2 | 2 | 2 |
| Closed Shrubland | 4 | 3 | 4 |
| Cropland | 4 | 3 | 4 |
| Bare Ground | 10 | 6 | 10 |
| Water | 15 | 15 | 15 |
| Deciduous Broadleaf Forest | 7 | 6 | 7 |
| Urban and Built up | 5 | 4 | 5 |

*Tabelle 5: Drei verschiedene Kalibrierungen (V1,V4,V10) für den Kostenfaktor Vegetation (mit eingefärbten veränderten Werten bezüglich der Standardkalibrierung V1)*

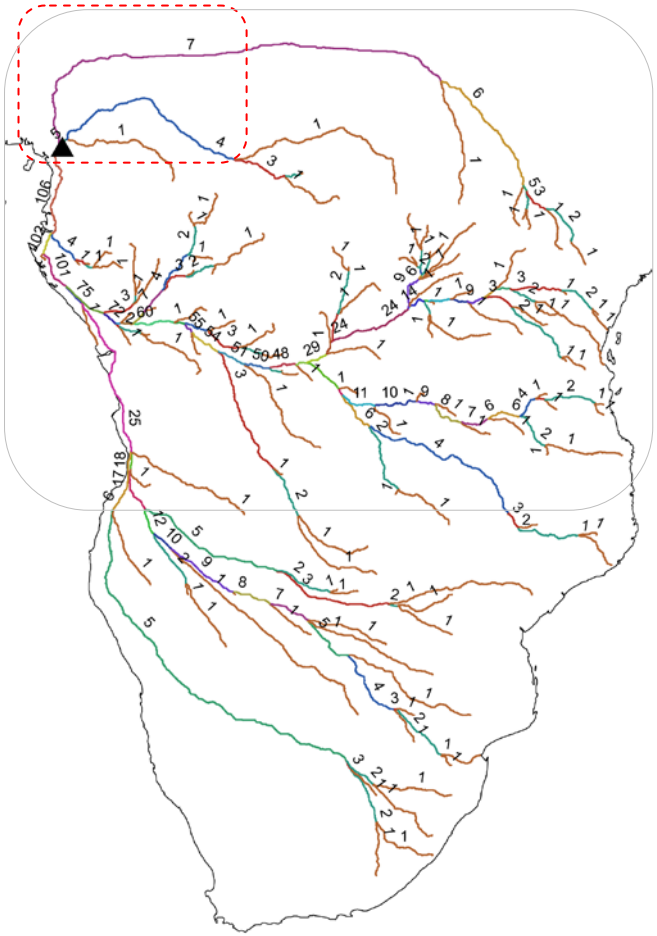


*Abbildung 20: Kostenpfade bei Vegetation V1 mit Shreve-Ordnungen*

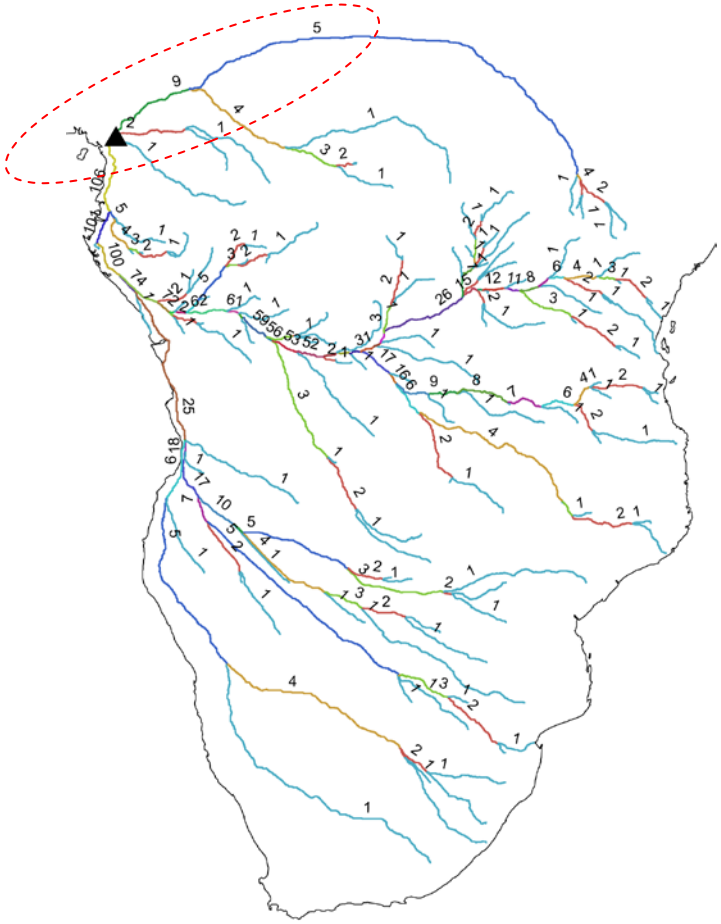


*Abbildung 21: Kostenpfade bei Vegetation V4 mit Shreve-Ordnungen*

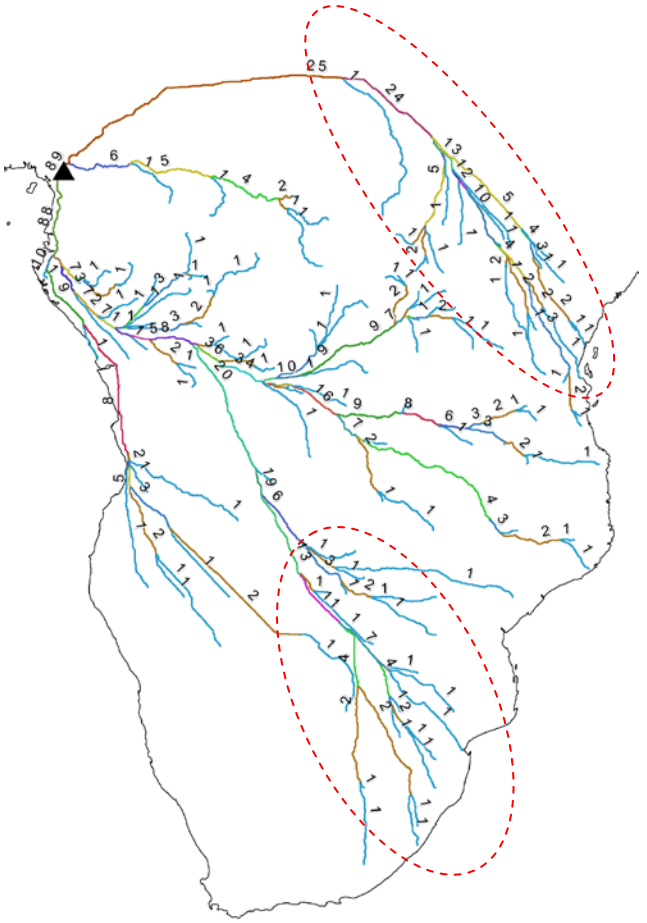


*Abbildung 22: Kostenpfade bei Vegetation V10 mit Shreve-Ordnungen*

Als erstes soll hier der nördliche, bezüglich der Early-und Late-Split Hypothese interessante Ausschnitt des Gebietes betrachtet werden. Dieser ist bei der ersten Kalibrierung V1 zu diesem Zweck mit einem grauen Rahmen umrandet (Abb. 20). Wie bereits aus vorherigen Einzelpfaduntersuchungen hervorging, gibt es Sprachpunkte im östlichen Teil des Untersuchungsgebiets, welche über die nordöstliche Regenwaldumgehung erreicht werden. Diese Verbindung ist entgegen der belegten Hypothese des Late-Splits auch hier bei allen drei Kalibrierungen sofort auszumachen. Wie aus den Shreve-Ordnungszahlen gelesen werden kann, hat der südliche Hauptzweig gegenüber dem nördlichen aber bei allen Kalibrierungen ein starkes Übergewicht bezüglich der Anzahl an verbundenen Punkten (Abb. 20 bis 22). Deren Anzahl variiert je nach Kalibrierung, bei allen ist jedoch der südliche Ast stärker vertreten. Dies trifft indes auch auf die restlichen sieben Kalibrierungen zu. Es könnte aber auch lediglich die Folge der höheren Dichte respektive Anzahl an Punkten im südlicheren Teil sein.

Die markantesten Unterschiede zwischen den einzelnen Modellierungen sind mit einem gestrichelten roten Rahmen gekennzeichnet.
Vergleicht man alle drei Kalibrierungen miteinander, so fällt auf, dass die dritte Kalibrierung V10 grössere Unterschiede zu V1 aufweist. So ist bei dieser der nach Nordosten verlaufende Hauptzweig auch um einiges ausgeprägter, er erreicht eine Shreve-Ordnung von 25, während sie bei den anderen bloss 7 und 9 beträgt. V1 und V4 sind sich im Gesamtbild relativ ähnlich. Dies überrascht insofern, als bei dieser Kalibrierung mehr Werte (7) von der Standardkalibrierung V1 verändert wurden als bei V10 (2) (vgl. Tabelle 5). Dazu zählt unter anderem auch die starke Herabstufung des Einflusses des tropischen Regenwalds von 10 auf 6. Wie aus den Kostenpfaden ersichtlich wird, hat diesbezüglich die tiefere Kalibrierung der umliegenden Gebiete Woodland und vor allem auch Wooded Grassland aber die stärkeren Auswirkungen. Da der immerfeuchte Regenwald auch mit 6 noch immer ein zu hohes Hindernis darstellt, wird er nach wie vor umgangen. Bei V4 wurde zwar Woodland ebenfalls von 5 auf 4 gesetzt, Wooded Grassland blieb hingegen bei 2. Zudem liegt bei V10 Wooded Grassland noch einen Wert tiefer, nämlich bei 3.

Ein deutlicher Unterschied zwischen V1 und den beiden anderen zeigt sich beim eckigen Verlauf des Nordpfades am Anfang, wo der Regenwald und das Woodland viel deutlicher umgangen werden müssen. Entsprechend liegt auch bei V4 der Shreve-Wert mit 9 etwas höher als bei V1 (5).

Zwar unwichtig bezüglich Early- und Late-Split Hypothese aber dennoch auffällig ist, dass bei V10 das südliche Afrika hauptsächlich über einen mittleren Pfad (siehe Markierung) erreicht wird. Der westliche Pfad ist hier nur sehr wenig ausgeprägt. Dies lässt sich durch die grossen Flächen an Wooded Grassland in diesem Gebiet erklären.

### 5.2.2 Bildung hierarchischer Cluster

Abbildungen 23, 24, 25 und 26 zeigen Dendrogramme für die drei ausgewählten Kalibrierungen V1, V4 und V10. Zur Veranschaulichung der Clusterbildung ist der Distanzbaum für V10 entsprechend der einzelnen Hauptäste eingefärbt, welche bei der LCPA erkennbar sind.


*Abbildung 23: NJ-Dendrogramm aus räumlichen LCP-Distanzen bei V1*

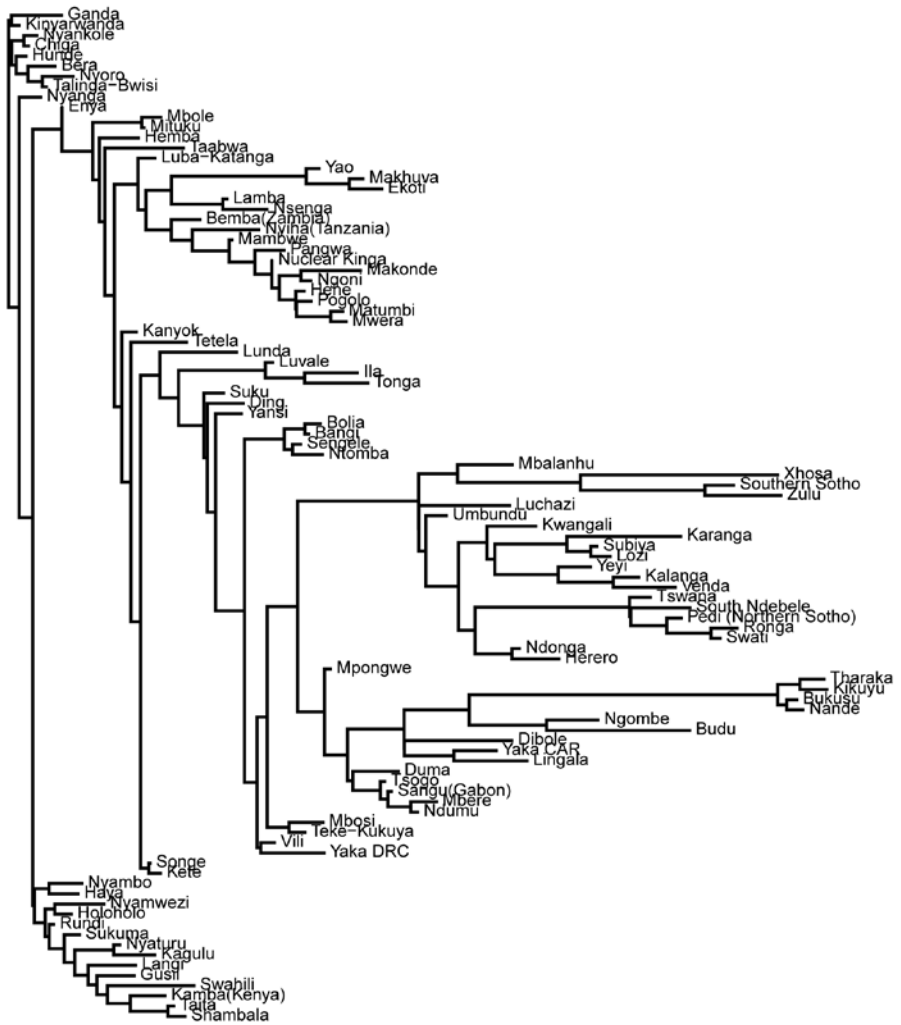



*Abbildung 24: NJ-Dendrogramm aus räumlichen LCP-Distanzen bei V4*

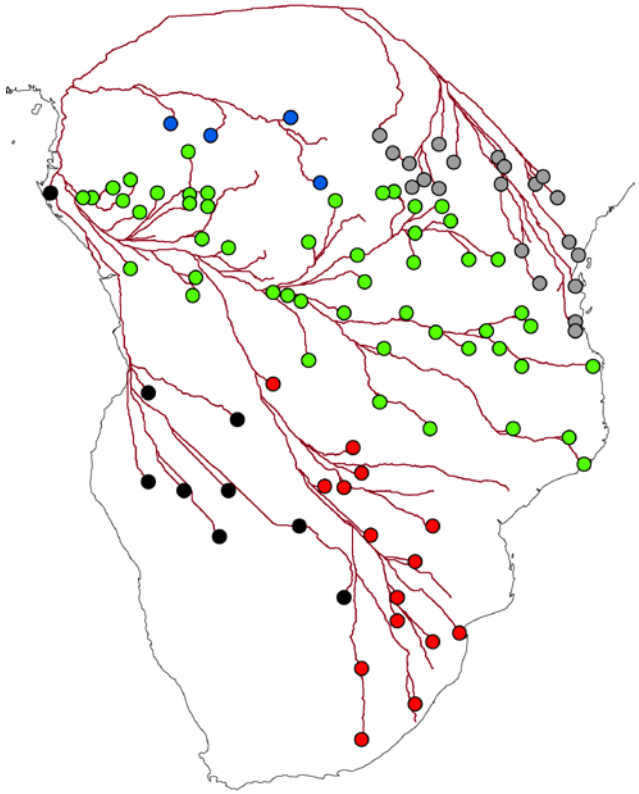


*Abbildung 25: Kostenpfade bei Gewichtung V10 mit farbig markierten Punkten für jeweilige geographische Hauptzweige*

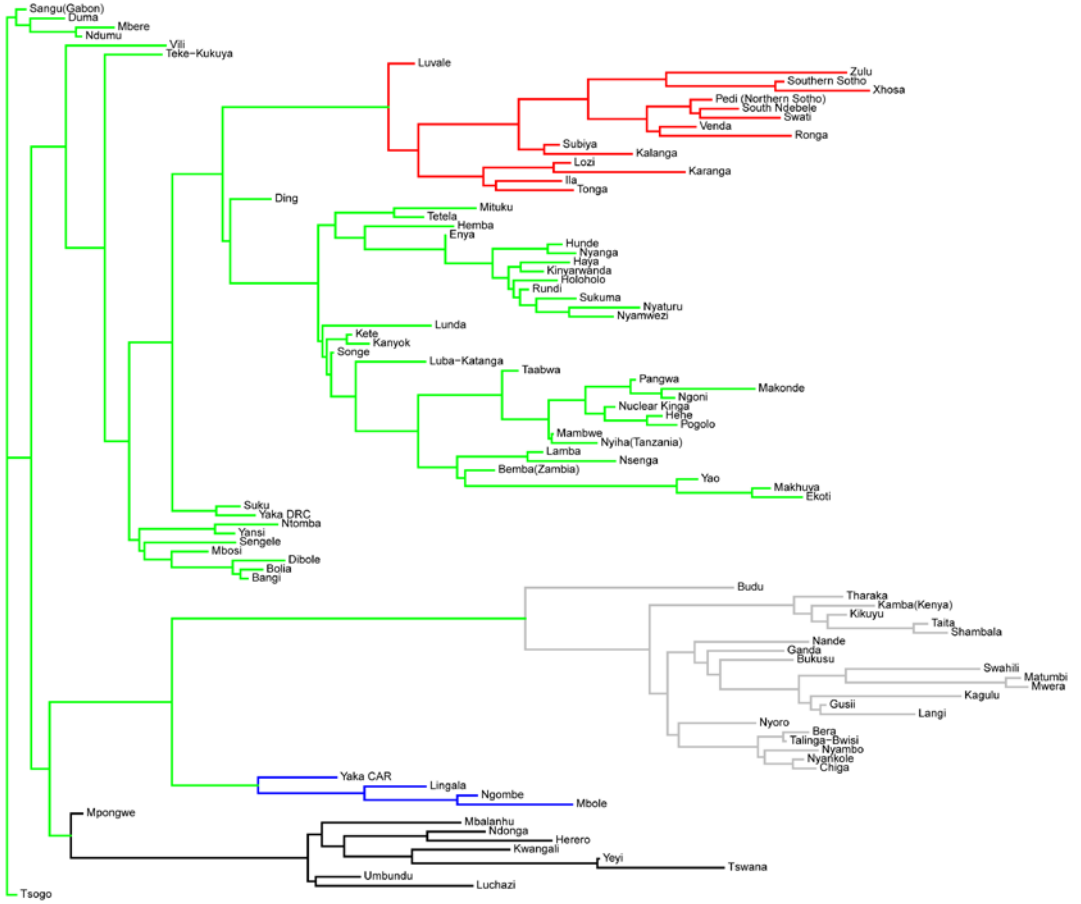



*Abbildung 26: NJ-Dendrogramm aus räumlichen LCP-Distanzen bei V10 mit Einfärbung der Zweige bezüglich ihrer geographischen Lage (gemäss Abldung 25)*

:

### 5.2.3 Vergleich der Dendrogramme

Im Folgenden wird eine Liste der Consensus-Baum Werte zwischen den einzelnen Dendrogrammen gezeigt (Tabelle 6). Wie in der Methodik erläutert, beinhalten alle Bäume 99 Tips (d.h. Spitzen oder Blätter) respektive Sprachen. Der phylogenetische Baum hat 67 Nodes, bei den anderen Distanzbäumen sind es 97.

Die räumlichen LCP-Distanzbäume wurden als spa1 bis spa10, die topologischen als topo1 bis topo10 bezeichnet. Euclidean_between steht für den aus den direkten Distanzen zwischen den Punkten generierten Baum und bei Euclidean_origin wurde die Distanz zum Ursprungsort der LCPA verwendet. In den ersten 3 Kolonnen (von links) sind die Übereinstimmungswerte (Node-Werte 1) aller geometrischen Bäume mit dem phylogenetischen Baum aufgelistet. In den Kolonnen der rechten Seite sind die Korrelationen des LCP-Baumes bei Standardkalibrierung V1 mit allen anderen Dendrogrammen (Node-Werte 2) notiert. Die Kombinationen sind dabei stets in aufsteigender Form bezüglich ihrer Werte notiert.
Wie aus Tabelle 6 entnommen werden kann, sind die Consensus-Werte zwischen den LCP-Pfaden und dem phylogenetischen Baum zwar besser als bei den euklidischen Bäumen, sie sind aber auch sehr tief. Nun wären die Resultate bezüglich Übereinstimmung mit dem genealogischen Baum indes nicht unbedingt schlecht, wenn diese Nodes dabei auf einer hohen Ebene liegen würden. Die Betrachtung des Consensus-Baums (siehe Anhang) des besten Beispiels (topo10, spa8, spa10) zeigt jedoch, dass es sich dabei nur um die tiefsten Ebenen handelt.

| Baum 1 | Baum 2 | Node-Werte 1 | Baum 1 | Baum 2 | Node-Werte 1 |
|---|---|---|---|---|---|
| Euclidean_origin | Genealogical | 1 | spa1 | Euclidean_between | 1 |
| Euclidean_between | Genealogical | 1 | spa1 | Euclidean_origin | 1 |
| topo1 | Genealogical | 3 | spa1 | Genealogical | 4 |
| topo5 | Genealogical | 3 | spa1 | spa10 | 16 |
| topo7 | Genealogical | 3 | spa1 | spa2 | 85 |
| topo9 | Genealogical | 3 | spa1 | spa3 | 56 |
| topo2 | Genealogical | 4 | spa1 | spa4 | 37 |
| topo3 | Genealogical | 4 | spa1 | spa5 | 16 |
| spa1 | Genealogical | 4 | spa1 | spa6 | 32 |
| spa3 | Genealogical | 4 | spa1 | spa7 | 57 |
| spa5 | Genealogical | 4 | spa1 | spa8 | 48 |
| spa6 | Genealogical | 4 | spa1 | spa9 | 14 |
| spa7 | Genealogical | 4 | spa1 | topo1 | 54 |
| spa9 | Genealogical | 4 | spa1 | topo10 | 15 |
| topo4 | Genealogical | 5 | spa1 | topo2 | 42 |
| topo6 | Genealogical | 5 | spa1 | topo3 | 38 |
| topo8 | Genealogical | 5 | spa1 | topo4 | 25 |
| spa2 | Genealogical | 5 | spa1 | topo5 | 12 |
| spa4 | Genealogical | 5 | spa1 | topo6 | 28 |
| topo10 | Genealogical | 6 | spa1 | topo7 | 41 |
| spa8 | Genealogical | 6 | spa1 | topo8 | 35 |
| spa10 | Genealogical | 6 | spa1 | topo9 | 14 |

*Tabelle 6: Liste verschiedener Consensus-Tree Werte zwischen den erstellten Dendrogrammen*

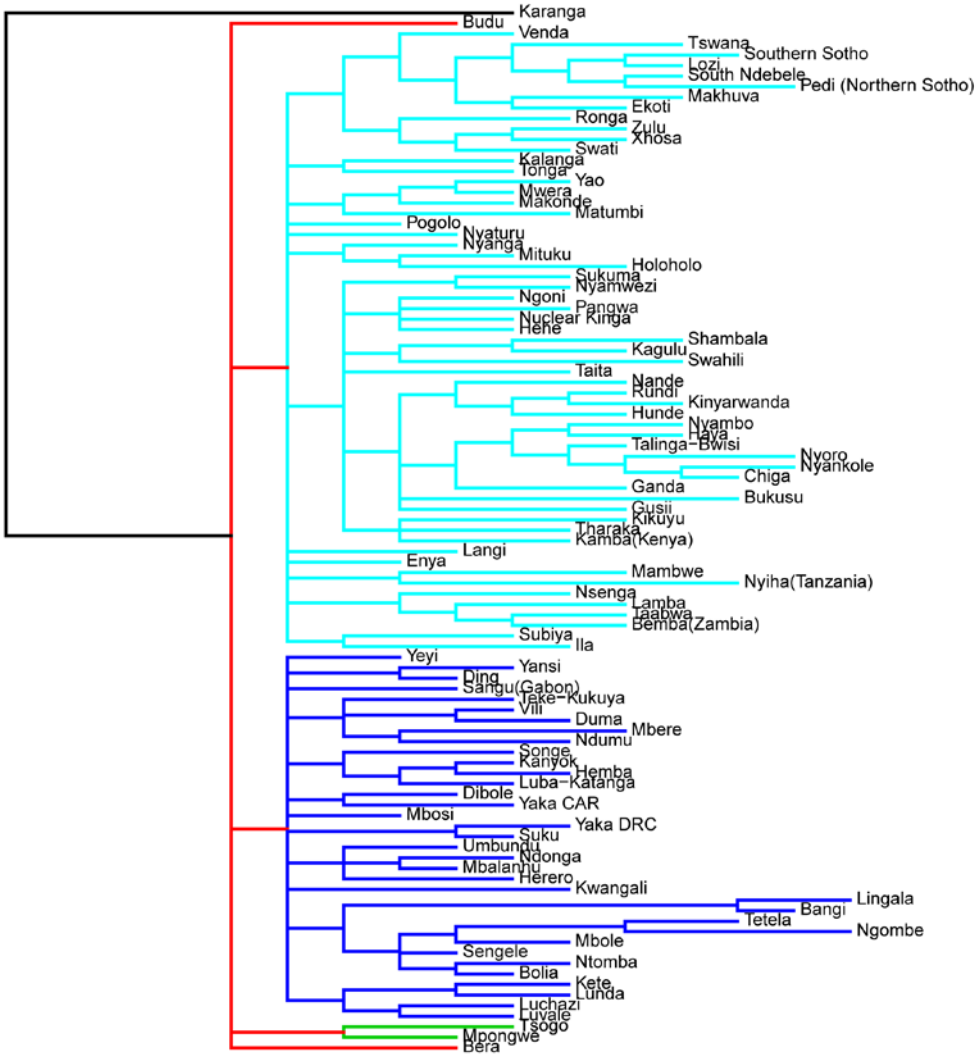



*Abbildung 27: Phylogenetischer Bantu-Baum auf der 2. Hierarchiestufe in 5 unterschiedliche Stämme eingefärbt*

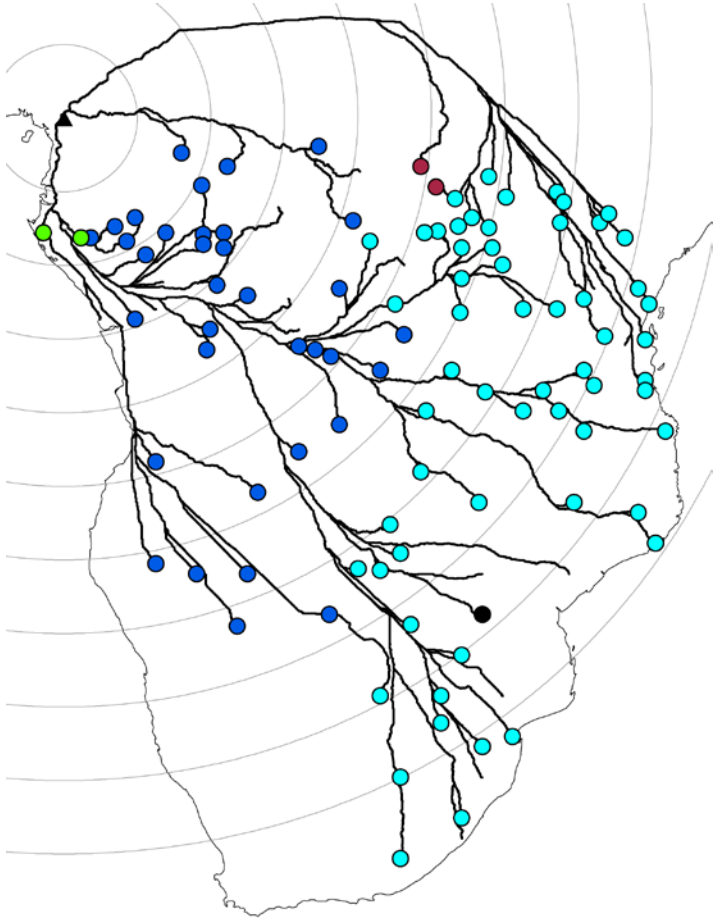


*Abbildung 28: Einfärbung des phylogenetischen Baumes (auf 2. Hierarchiestufe in 5 Stämme) respektive dessen Tips übertragen auf geographische Karte; mit Kostenpfaden für V1 und konzentrischen Kreisen zur Darstellung euklidischer Distanzen (Äquidistanz 400 km)*

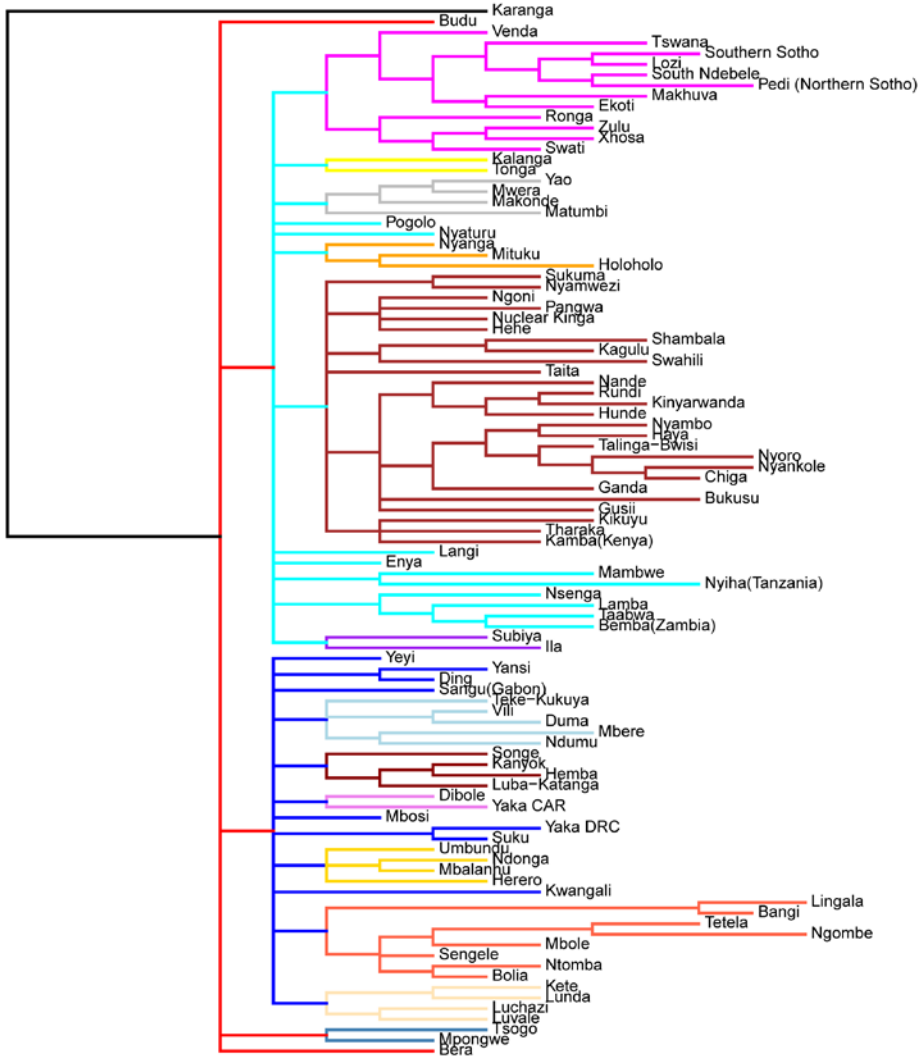



*Abbildung 29 Phylogenetischer Bantu-Baum auf der 3. Hierarchie in 18 unterschiedliche Stämme eingefärbt*

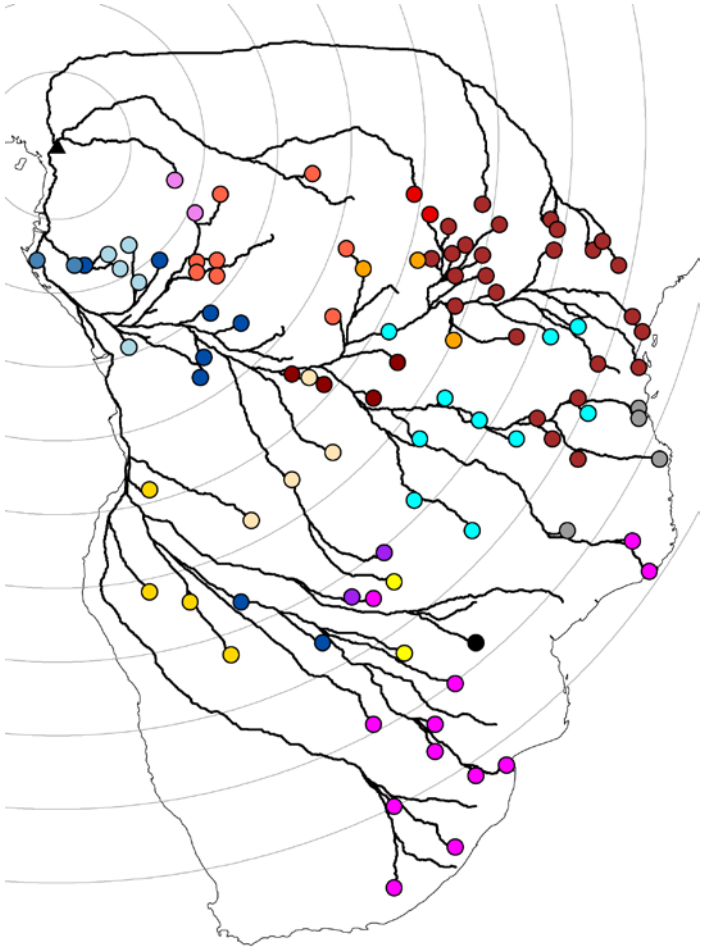


*Abbildung 30: Einfärbung des phylogenetischen Baumes (auf 3. Hierarchiestufe in 18 Stämme) respektive dessen Tips übertragen auf geographische Karte; mit Kostenpfaden für V1 und konzentrischen Kreisen zur Darstellung euklidischer Distanzen*

Die erste geographische Karte (Abbildung 28) mit den 5 unterschiedlich eingefärbten Punktegruppen zeigt noch ein sehr einfaches Muster. Dabei liegt zwischen den Kostenpfaden und den einzelnen Farbeinteilungen sehr wenig Übereinstimmung vor. So sind die beiden grossen Gruppen (blau bzw. cyan eingefärbt) auf 3 verschiedenen Ästen gut vertreten, bei letzerer liegt ein einzelner Punkt ganz im Süden sogar noch auf einem vierten Pfad. Selbst die beiden sehr nahe gelegenen grünen Punkte befinden sich auf unterschiedlichen Pfaden. Betrachtet man hingegen die konzentrischen Kreise (Radius = 400km), welche die euklidische Entfernung vom Ursprungsort darstellen, so scheint eher ein Zusammenhang erkennbar. Die beiden 2-er Gruppen (grün bzw. rot) liegen beide sehr nahe beisammen und stimmen daher gut mit dem euklidischen Ansatz überein. Die beiden grossen Gruppen (blau und cyan) haben zwar eine deutlich grössere Streuung, fasst man aber vier Kreise zusammen, respektive betrachtet das Ganze auf kleinerem Massstab, passen auch diese gut in das euklidische Muster.

Die Aufteilung in die nächste sprachliche Auflösung, beziehungsweise 3. Hierarchiestufe (Abbildung 29) ergibt bei der geographischen Karte (Abb. 30) ein recht anderes Bild. Dabei liegen bereits 18 verschiedene Gruppen vor. Bei Betrachtung von Abbildung 30 fällt auf, dass die Kolorierung der Punkte deutlich schlechter mit dem euklidischen Ansatz erklärt werden kann. Einige Farbgruppen sind verhältnismässig stark gestreut. Das extremste Beispiel stellt hierbei die blaue Gruppe dar, welche sich mit 8 Punkten über ganze 7 Kreiszonen erstreckt.
Auch mit den Kostenpfaden können die zum Teil stark verzettelten Punktegruppen nicht wirklich erklärt werden. Verglichen mit dem vorherigen Muster stimmen die Pfade jedoch deutlich besser mit diesem überein. So lassen sich bei einigen Gruppen gewisse Verbindungsmuster erkennen. Diese Verbindungen sind jedoch meist nur in ihren Grundzügen plausibel und müssten im Detail noch angepasst werden. Ein deutlich ersichtliches Problem bildet die Tatsache, dass aus jedem Punkt jeweils ein einzelner Pfad hervorgeht. Dieser Schwachpunkt der LCP-Methode wird im folgenden Teil der Diskussion deshalb nochmals genauer erläutert und mögliche Verbesserungsansätze aufgezeigt.
Der zweite Vergleich der sprachlichen Punkte mit den Kostenpfaden zeigt aber, dass in der Methode der Kostenpfade durchaus Potential zur Erklärung vorhandener Ausbreitungsmuster vorhanden ist, welches noch weiterer Untersuchungen bedarf.

# 6 Diskussion

In diesem Kapitel werden die geographischen sowie die sprachlichen Aspekte der LCPA diskutiert. Dabei werden auch die anfangs dieser Arbeit formulierten Forschungsfragen erörtert.

## 6.1 Geographische Aspekte

### 6.1.1 Faktoren und Modellierung

Einen der grundlegendsten Schritte bei der LCPA stellt natürlich die Bestimmung der Faktoren dar. Werden die falschen Kriterien in die Modellierung miteinbezogen, nützen die ausgeklügeltsten Algorithmen nichts (Howey 2007, Herzog 2013). Entsprechend wurde die geeignete Wahl der Faktoren auch in der ersten der beiden anfangs dieser Arbeit formulierten Forschungsfragen als zentraler Aspekt der LCPA definiert. Diese lautete folgendermassen:

- *Welche geographischen Kriterien sollten bei der Modellierung einbezogen und wie können diese kalibriert und gewichtet werden?*

In diesem Teil sollen deshalb die zur Bestimmung der Bantu-Ausbreitungspfade gewählten Parameter nachträglich nochmals betrachtet werden. Zudem wird analysiert, ob die verwendeten Faktoren entsprechend ihrem Einfluss in der Modellierung berücksichtigt wurden.

Die Kriterienwahl stellt ein stark sowie kontrovers diskutiertes Thema innerhalb der verschiedenen LCP-Studien dar. Dabei reicht die Palette von simplen Ansätzen (Rees 2003, Taliaferro et al. 2010), welche bloss die Neigung mit einbeziehen bis hin zu komplexen Multi-Kriterien Modellen (Howey 2007, Field et al. 2005). Llobera (2000) proklamiert in seiner Studie gar die Berücksichtigung nicht fassbarer, mentaler Kriterien. Des Weiteren wird in vielen Studien die Wichtigkeit eines genauen, im vornherein ausgearbeiteten Konzepts zur Einbeziehung der Faktoren betont. So wird etwa bei Berry (2000) und Atkinson et al. (2005) ausführlich der Nutzen einer Multi-Criteria-Analysis Anwendung erläutert.

Bei dieser Arbeit wurde hingegen ein pragmatischer Ansatz verfolgt und auf die Anwendung einer Multi-Criteria-Analysis verzichtet. Es konnte jedoch aufgezeigt werden, dass diese Vorgehensweise der unsicheren Datenlage besser angepasst ist als die erwähnten Theorie-orientierten Methoden.

Der erste Schritt bezüglich der Faktorenwahl erfolgte mit deren genauer Überprüfung. Dabei wurden alle in Frage kommenden Faktoren analysiert und davon nur die aussagekräftigen verwendet. Anstatt die Faktoren anhand eines bereits vorbestimmten Konzepts zu kombinieren und gewichten, wurden nun die Kostenpfade erst einmal für die einzelnen Faktoren berechnet. Auch dieser Schritt brachte einen entscheidenden Vorteil mit sich: So konnte aufgrund des beobachteten Fehlens eines massgeblichen Einflusses der Hangneigung dieser rechenaufwendige Faktor frühzeitig ausgeschieden werden.

Ein analoges Vorgehen wurde auch für die Beurteilung der Flüsse bezüglich ihrer Eignung als Ausbreitungswege gewählt. Dies geschah mit der Zuordnung des tiefs-

ten Werts (1). Die Anwendung einer einfachen isotropen Kostenoberfläche für die Flüsse zeigte, dass aufgrund der Lage derselben diese simple Berechnungsart völlig ausreichte. So konnte von viel komplizierteren, anisotropen Modellen (Choi et al. 2013) abgesehen werden, welche neben Rasteroberflächen auch noch Vektornetzwerke mit einbeziehen.

Einen weiteren Aspekt der verwendeten Parameter stellt die Auflösung derselben dar. Zwar war zu Beginn dieser Arbeit vorgesehen, allenfalls noch eine höhere Auflösung des Geländemodells zu benutzen. Aufgrund technischer Einschränkungen musste jedoch davon abgesehen werden. Diesbezüglich könnte nun die Frage gestellt werden, ob die Auflösung von 30 Bogensekunden ausreichend sei, respektive ob mit dieser die optimalen Kostenpfade wirklich bestimmt werden können. Wie Anderson (2012) schreibt, kann für eine Studie, welche Kostenpfade auf kontinentaler Skala berechnet, grundsätzlich eine Auflösung von 1 km als ausreichend betrachtet werden (Anderson 2012). So dürften beispielsweise Felsen oder Hügel mit einem Durchmesser im Bereich von hundert Metern den Verlauf eines Gesamtpfades kaum massgeblich beeinflusst haben. Auch Kantner (2012) warnt vor einer Überbewertung der DEM-Rasterauflösung. Da der Mensch nicht wie Wasser funktioniere, folge er auf grossem Massstab nicht immer dem Least Cost Path. Gerade bei der Fortbewegung zu Fuss können manchmal auch kleinere Hindernisse überwunden werden. Bei der Verwendung von Algorithmen müsse zudem überprüft werden, ob diese nicht aus ursprünglich für Fahrzeuge entwickelten Ansätzen abgeleitet wurden, was oftmals zu falschen Resultaten führe.

Insgesamt stellt also der Beginn mit einem simplen Modell, welches im Verlaufe der Untersuchungen den Anforderungen entsprechend angepasst wird eine sehr sinnvolle Variante dar. Jedoch sollten die auf diese Weise berechneten Kostenpfade nicht das Endresultat darstellen. Das erhaltene Modell sollte in einem Folgeschritt mittels leicht variierter Parameter auf seine Stabilität kontrolliert werden (Herzog 2013). Gerade beim hier untersuchten Beispiel, wo aufgrund der zeitlichen Entfernung eine enorme Unsicherheit bezüglich der Datenlage existiert, ist die Überprüfung der Robustheit des Modells essentiell.

Im folgenden Teil wird deshalb nochmals genauer auf die Unsicherheit der Datenlage beziehungsweise die Eignung des Untersuchungsbeispiels aus geographischer Sicht eingegangen und es werden Schlussfolgerungen zu einer notwendigen Überprüfung der Stabilität des Modells abgeleitet.

### 6.1.2 Unsichere Datenlage und Modellvariation

Bereits zu Beginn der Arbeit wurde detailliert auf die möglichen Unsicherheiten bezüglich der Datenlage des Bantu-Ausbreitungsgebiets hingewiesen. Wie aus mehreren Quellen entnommen werden konnte, dürfte der Gesamtzustand der Vegetation zu Zeiten der Bantu-Ausbreitung dem heutigen sehr ähnlich gewesen sein. Jedoch können über die genaue Ausdehnung des Regenwalds und der umliegenden Vegetationen nur Vermutungen angestellt werden (Schwartz 1992, Bostoen 2012). Dieser grosse Unsicherheitsfaktor macht nachträgliche Variationen des Modells zur Überprüfung seiner Robustheit umso wichtiger. In diesem Sinne stellte die fehlende Möglichkeit der Anwendung einer Monte-Carlo-Simulation eine starke Einschränkung der Untersuchung dar. Diese hätte Informationen bezüglich der Auftretenswahrscheinlichkeit der jeweiligen Kostenpfade gezeigt. Damit hätten diejenigen Punkte bestimmt werden können, wo der Verlauf des Kostenpfades sehr unsicher ist. Anhand der An-

zahl an Pfaden, welche bei diesen entscheidenden Punkten eine bestimmte Abweichung aufgewiesen hätten, wären gewissermassen auch Aussagen zur Wahrscheinlichkeit alternativer Wege möglich gewesen.

Dies stellt generell einen Schwachpunkt bei der gängigen LCPA dar. Bei dieser wird immer nur ein einziger optimaler Pfad bestimmt. Jedoch wären unter Umständen Kenntnisse über etwas weniger, aber ähnlich günstige Alternativpfade fast ebenso interessant. Diesbezüglich wurde von Pinto & Keitt (2009) ein interessanter Ansatz vorgestellt. Sie modifizierten dabei den Dijkstra-Algorithmus, so dass dieser mehrere suboptimale Pfade ebenfalls als Resultate ausgab. Generell liegt hierin noch einiges Potential, wenn man bedenkt, dass der Dijkstra-Algorithmus als Single-Source-Algorithmus sowieso immer alle Pfade zu dieser Source-Zelle (respektive der Zielzelle) berechnet. Für Anwendungen mit isotropen Kostenoberflächen gäbe es deshalb noch interessante Erweiterungen. Die Single-Pair-Anwendung stellt bereits einen eingeschränkten Fall dar (Cormen et al. 2009).

Die durchgeführte manuelle Variation der Kalibrierungen stellt natürlich einen etwas unbefriedigenden Ersatz gegenüber einer Monte-Carlo-Simulation dar. Selbst mit deren Anwendung könnten jedoch nicht alle Unsicherheiten verlässlich einbezogen werden. So werden im Zusammenhang mit der Bantu-Expansion verschiedentlich offene Korridore im Regenwald erwähnt, welche diese angeblich erleichtert haben sollen (Van Geel et al. 1996, Elenga et al. 1992). Falls solche Korridore zu dieser Zeit wirklich über längere Strecken vorhanden waren, so würden diese auch mittels Monte-Carlo Simulation nicht erfasst werden.
Entsprechend könnten solche Korridore auch entscheidend das Late-Split Modell begünstigt haben. Wie die Untersuchungen gezeigt haben, kann dieses auch mit variierenden Werten für die Vegetation durch die Kostenpfade nicht klar belegt werden. So resultierte auch bei relativ stark unterschiedlichen Kalibrierungen immer ein Nordostpfad. Falls die Kosten für die südlichen Pfade aber aufgrund eines offenen Korridors reduziert worden waren, könnte dies das Gleichgewicht auch für die nördlicheren Punkte im Ostgebiet tatsächlich zugunsten der Late-Split Hypothese verschoben haben. Mit den aktuellen Methoden der Forschung ist es jedoch nicht möglich, die Existenz oder gar den Verlauf dieser Schneisen zu bestimmen.
Ein weiteres Kriterium, dessen Einfluss bezüglich des Gewichts des Südpfades zum Nordpfad noch genauer untersucht werden müsste, stellt zudem die Wahl der Sprachpunkte dar. Dabei müsste in Absprache mit Linguisten abgeklärt werden, ob deren Anzahl für die Sprachendichte der jeweiligen Gebiete auch wirklich repräsentativ ist. So könnte verhindert werden, dass ein starker Hauptpfad nicht bloss das Resultat einer zu hohen lokalen Punktedichte des untersuchten Datensatzes ist.

Zusammenfassend kann gesagt werden, dass die Bantu-Ausbreitung in linguistischer Hinsicht aufgrund der soliden Wissensbasis ein geeignetes Untersuchungsbeispiel darstellt. Da es sich beim Hauptkriterium, der Vegetation, um einen zeitlich sehr unkonstanten Faktor handelt, ist das Beispiel aus geographischer Sicht hingegen nicht unproblematisch.

Aufgrund der räumlich sehr ungenauen Vorgabe-Daten bezüglich der Early- und Late-Split Hypothesen beschränkte sich die Analyse der erhaltenen Kostenpfade in dieser Arbeit jedoch nicht bloss auf die Geographie, sondern es wurden weitere Untersuchungen gemacht. Diese werden im folgenden, mehr auf die sprachlichen Aspekte der Arbeit gerichteten Kapitel diskutiert.

## 6.2 Sprachliche Aspekte

Die Untersuchung einer Sprachausbreitung mittels LCPA stellt eine besondere Anwendung dieser Methodik dar. Da es sich dabei um ein relativ unerforschtes Terrain handelt und die vorgestellte Arbeit einen explorativen Ansatz darstellt, sollen in diesem Kapitel themenspezifische Aspekte der Sprachausbreitung diskutiert werden, welche mit der LCPA nicht oder nur unzureichend erfasst werden können.

Ein erster wichtiger Punkt ist, dass eine LCPA keinen oder zumindest keinen expliziten Zeitfaktor beinhaltet. Gerade bei Sprachausbreitungen kann sich aber die Dynamik, mit welcher diese erfolgen, stark unterscheiden. Ein Ausbreitung kann sowohl schnell vonstatten gehen als auch einen schleichenden Prozess darstellen. So dürfte auch die Expansionsrate bei Gruppen, welche sich 3000 km von ihrem Ursprungsgebiet entfernt haben, grösser gewesen sein als bei solchen mit bloss 200 km. Currie et al. (2013) bezogen die unterschiedlichen Expansionsraten der Bantu-Völker in ihrer Studie mit ein. Zu diesem Zweck verwendeten sie eine neue Methode zur genauen Untersuchung der Evolution von Gruppen einer Spezies mithilfe derer Phylogenetik.
Der zeitliche Faktor wird mit der Methode der Kostenpfade hingegen nicht erfasst. Dabei spielt gerade für die Untersuchung einer Völkerwanderung mittels LCPA der Zeitfaktor eine entscheidende Rolle bezüglich der Eignung dieser Methode. So dürfte eine LCPA zur Erfassung einer einmaligen, sehr schnell verlaufenden Migrationswelle eher ungeeignet sein, weil dabei die Wegfindung relativ spontan und zufällig erfolgt. Eine in kleineren Schüben über einen langen Zeitraum erfolgende Expansion erfüllt hingegen viel eher die Bedingungen für eine LCPA. Denn mit der mehrmaligen Begehung eines Weges sinkt auch dessen Zufallskomponente (Herzog 2013). In diesem Zusammenhang ist auch bei der Bantu-Ausbreitung eine Untersuchung mittels LCPA nicht abwegig, da diese sehr wahrscheinlich in vielen Schüben über längere Zeit erfolgte (Vansina 1995).

Ein weiterer, oft angeführter Kritikpunkt bei LCP-Untersuchungen ist, dass bei dieser sozio-kulturelle Gründe zu wenig oder überhaupt nicht berücksichtigt werden. Wie bereits in Kapitel 2 in der Übersicht der in LCP-Studien verwendeten Faktoren erwähnt wurde, wird der Methode aus diesem Grund manchmal ein geodeterministischer Standpunkte vorgeworfen. Diese Kritik könnte folglich auch bei der hier vorgestellten Studie angebracht werden.

Tatsächlich dürften beispielsweise kulturelle Gründe mitunter eine Rolle gespielt haben bei der Bantu-Migration. Überhaupt werden mit der LCPA nur Pull-Faktoren miteinbezogen, während Push-Faktoren ignoriert werden. Gerade im Falle der Bantu-Ausbreitung stellten die zugrundeliegenden Push-Faktoren ein interessantes Untersuchungsgebiet dar. Generell wird davon ausgegangen, dass sie die Folge einer wachsenden Bevölkerungszahl und damit einhergehenden Problemen bei der Nahrungsbesorgung war (Vansina 1995).

Auf den Vorwurf des Geodeterminismus kann jedoch entgegnet werden, dass der hier vorgestellte Ansatz keinerlei Sonderstellung für sich in Anspruch nimmt, sondern allenfalls als komplementäre Untersuchungsmethode zu anderen Studien verwendet werden könnte. Die sozio-kulturellen Hintergründe der Migration sind ebenfalls Faktoren, welche einer genaueren Untersuchung bedürften. Jedoch wäre es falsch, die-

se Faktoren anhand kaum überprüfbarer, subjektiver Annahmen direkt in die Berechnung der Kostenpfade mit einzubeziehen.

Des Weiteren wird auch anderen wichtigen Aspekten der Völkerwanderung mit der LCP-Methode kaum Rechnung getragen. So werden dabei auch die anfangs der Arbeit vorgestellten unterschiedlichen Ausbreitungsarten der Sprachen nicht berücksichtigt. Aus den Kostenpfaden kann unmöglich abgeleitet werden, ob es sich hauptsächlich um einen demographischen oder um einen kulturellen Austausch handelte. Genauso wenig lassen sich daraus Aussagen bezüglich der Entstehungsart (Kladogenese oder Anagenese) einer Sprache entnehmen. Hierfür sind linguistisch-genetische Ansätze notwendig. Dies zeigt wiederum, dass die Methode der Kostenpfade im besten Fall als ergänzende Methode zu sprachlichen Ansätzen benutzt werden könnte.

Der oben erörterten Punkt bilden dann auch den Hintergrund für die zweite zentrale Forschungsfrage dieser Arbeit, die folgendermassen lautete:

- *Inwiefern kann die Least Cost Path Analyse als ergänzende Methode zur Untersuchung historischer Sprachausbreitungen herangezogen werden?*

Um dies zu überprüfen, wurden die Kostenpfade in dieser Arbeit noch genauer auf eine Übereinstimmung mit den linguistischen Pfaden überprüft. Die dabei auftretenden Schwierigkeiten sowie das mögliche Potential dieser Analysen werden im nächsten Kapitel genauer diskutiert.

## 6.3 LCPA als komplementärer Ansatz zu Sprachstudien

Aufgrund der hohen räumlichen Ungenauigkeit der Daten aus den linguistischen Studien bezüglich der Early- und Late-Split Hypothesen wurde die Analyse der Kostenpfade nicht auf die Geographie beschränkt, sondern es wurde zusätzlich ein Vergleich der räumlichen Daten mit den genealogischen Daten durchgeführt. Der Vergleich der LCP-Distanzbäume mit dem phylogenetischen Baum erbrachte aber kein zufriedenstellendes Resultat.

Im Folgenden sollen deshalb die problematischen Aspekte dieses Vergleichs genauer erläutert werden. Es soll aufgezeigt werden, was die möglichen Gründe für die dabei erhaltene ungenauen Resultate gewesen sein könnten.

Wie bereits in Kapitel 4 erwähnt, wurden zur Erstellung der Bäume jeweils unterschiedliche Distanzmasse verwendet: ein statistisches Distanzmass bei den LCP-Bäumen und ein evolutionäres beim phylogenetischen Baum (Paradis 2006).
Während ersteres mathematisch genau berechnet werden kann, handelt es sich bei letzterem um Schätzwerte von Linguisten. Diese unterschiedliche Erstellungsart ist auch bei der Betrachtung der Bäume augenscheinlich. Da beim geometrischen Distanzbaum Algorithmen vorliegen, welche diesen bis ins kleinste Detail berechnen können, hat er eine viel höhere Auflösung als der phylogenetische Baum. Beim genealogischen Baum ist die Aufteilung eines Stamms in Untergruppen jeweils mit hohem Aufwand verbunden, deshalb werden vorzugsweise Sprachen zusammengefasst. Bei vielen Sprachen ist eine höhere Auflösung auch mangels des nötigen Wissens nicht mehr weiter möglich oder aus sprachlicher Sicht schlicht nicht sinnvoll. Die

räumlichen Distanzbäume unterscheiden sich deshalb grundsätzlich stark vom genealogischen, was einen Vergleich erschwert.

Generell ist es sehr schwierig auszumachen, welches die genauen Ursachen für die schlechte Übereinstimmung zwischen den Bäumen sind. Nicht vergessen werden dürfen sicherlich die zuvor diskutierten grossen Unsicherheiten des Modells bezüglich der Datenlage. Jedoch gibt es auch eindeutige Schwächen im Ansatz selbst, welche zu einem ungeeigneten Resultat führen. Ein Hauptproblem der LCPA ist, dass für jeden Punkt immer ein einzelner Kostenpfad berechnet wird. Dies führt oftmals zu sehr unrealistischen Resultaten, welche mithilfe der Abbildung 31 gut veranschaulicht werden können.

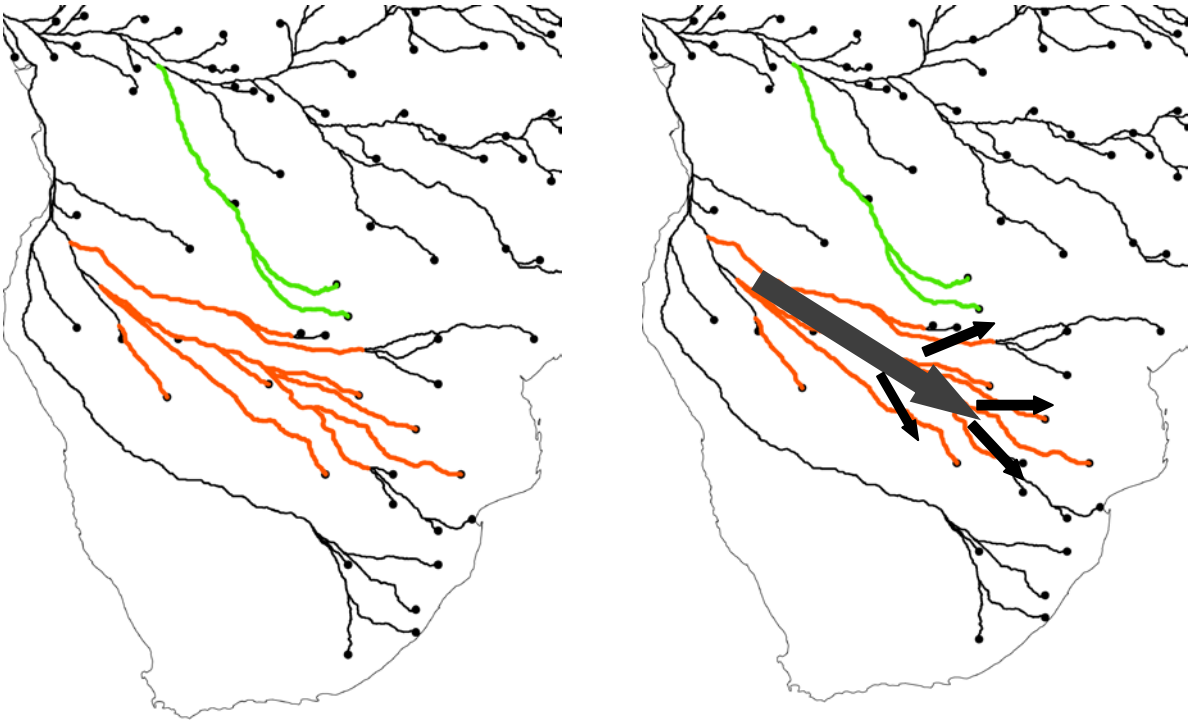


*Abbildung 31:: links: Südlicher Ausschnitt derKostenpfade für V1 mit farbig markierten problematischen Pfaden; rechts: Südlicher Ausschnitt der Kostenpfade mit schwarzen Pfeilen für realistische Ausbreitungswege*

Ein erster problematischer Aspekt ist bei den rot markierten Pfaden zu sehen. Dabei fallen mehrere fast parallel verlaufende Pfade auf, welche über sehr weite Distanzen (fast 2000 km beim längsten Stück) in einem relativ geringen Abstand zueinander, einzeln ihre Zielpunkte erreichen. Diese Pfade sind zwar rein mathematisch gesehen die optimalen Verbindungen, in der Realität ist es jedoch sehr unwahrscheinlich, dass jede dieser Sprachgruppen separat ihren Zielort erreichte. Viel realistischer ist, dass ein Teil des Weges gemeinsam zurückgelegt wurde und sich einzelne Gruppen dann losgelöst haben. Dies ist in Abbildung 31 (rechts) mit den schwarzen Pfeilen dargestellt. Da eine Reise in einer grösseren Einheit generell vorteilhaft sein kann (Sicherheit, Nahrungsversorgung usw.), wäre dieses Szenario in der Realität unter Umständen auch mit tieferen Kosten verbunden. Einen gleichfalls eher unrealistischen Fall stellt der grün markierte Pfad in Abbildung 31 dar. Die beiden südlichen Endpunkte befinden sich dabei sehr nahe an den roten Punkten, sind aber auf einem komplett unterschiedlichen Weg an diesen Ort gelangt. Dass einzelne Sprachgrup-

pen über solch riesige Distanzen einzeln an ihren Zielort gelangt sind, dürfte in der Realität nur sehr selten der Fall gewesen sein.

Die dargestellten Beispiel zeigen, dass die durchgeführte LCPA zur Erklärung der Ausbreitungspfade insofern unzureichend ist, als sie die dabei vorhandene Gruppendynamik nicht berücksichtigt. Der LCP-Ansatz müsste diesbezüglich also noch erweitert werden. Dies könnte möglicherweise mit dem Konzept der Betweenness Centrality erreicht werden. Dazu würden zuerst die Kostenpfade zwischen allen Punktekombinationen berechnet und von diesen danach der Knoten bestimmt, welcher Bestandteil möglichst vieler kürzester Verbindungen ist. Es würde dabei also gewissermassen die optimale Verbindung aller kürzesten Pfade ausgelesen (Newman 2003).

Die Berechnung der Betweenness Centrality sollte jedoch erst erfolgen, nachdem das Modell bereits zu einem gewissen Grad auf seine Stabilität überprüft worden ist. Falls dazu die Anwendung einer Monte-Carlo-Simulation nicht möglich ist, kann mithilfe manuell kalibrierter Variationen zumindest ein erster Eindruck von dieser erlangt werden. Denn für eine gegebene Anzahl Sprachpunkte x ergeben sich x * (x-1)/2 Verbindungen zwischen diesen, was mit einer relativ hohen Berechnungszeit verbunden ist. So benötigte die in dieser Arbeit einmal probeweise durchgeführte Berechnung aller Pfade zwischen 44 Punkten aus dem Nordteil des Untersuchungsgebietes bereits eine Zeit von ungefähr 30 Stunden. Für 99 Punkte würde sich die Berechnungsdauer bereits um den Faktor 5 erhöhen.
Mittels des hier vorgeschlagenen Ansatz der Betweenness Centrality könnten wahrscheinlich auch die gemäss dem genealogischen Baum eingefärbten Punkte im Detail noch besser erklärt werden als mit den in dieser Arbeit verwendeten einfachen Kostenpfaden. Die Betweenness Centrality Methode würde dem zuvor erläuterten gemeinschaftlichen Charakter der Ausbreitung eher Rechnung tragen.

# 7 Fazit

In diesem letzten Kapitel der Arbeit werden die gewonnenen Erkenntnisse besprochen sowie die Grenzen der angewandten Methodik aufgezeigt. Zum Abschluss wird in einem Ausblick noch auf mögliche zukünftige Folgeprojekte zu dieser Untersuchung hingewiesen.

## 7.1 Erreichtes

Ein erstes Ziel der Arbeit stellte die Erstellung einer angemessenen Least Cost - Modellierung der Ausbreitungspfade der Bantu-Sprachfamilie dar. Entscheidende Kriterien waren hierbei die Wahl der Faktoren sowie die Kalibrierung und Gewichtung derselben.

Im Folgenden konnte aufgezeigt werden, dass der in der Arbeit verfolgte praxisorientierte Ansatz bei gegebener dünner Datenlage sinnvoller ist als die Verwendung komplizierter a priori bestimmter Konzepte. Weiter wurde dabei gezeigt, dass nur Faktoren verwendet werden sollten, wenn dabei sowohl die Bedingung der Rekonstruierbarkeit als auch der Modellierbarkeit gegeben ist. Durch die anfängliche Verwendung einfacher Modelle, mit anschliessender schrittweiser Erweiterung, konnte der Einfluss der einzelnen Faktoren klar dokumentiert werden. Auf diese Weise wurde der geringe Einfluss der Hangneigung im Untersuchungsgebiet verdeutlicht und diese daraufhin ausgeschieden. Zudem wurde aufgezeigt, dass die Gewässer respektive Flüsse bei der Bantu-Ausbreitung kaum eine grosse Rolle gespielt haben dürften.

Die Beantwortung der zweiten zentralen Forschungsfrage der Arbeit, nämlich inwiefern die Methode der kürzesten Kostenpfade als ergänzende Methode zur Untersuchung historischer Sprachausbreitungen herangezogen werden könne, erwies sich hingegen als schwieriger. Die tiefen Consensus-Werte zwischen dem genealogischen Baum und den LCP-Bäumen deuten stark darauf hin, dass die Methode zumindest für direkte Vergleiche mit den Sprachdaten ungeeignet ist. Aufgrund der unsicheren Datenlage sowohl bezüglich der geographischen Faktoren als auch der Sprach-Phylogenese kann dies jedoch nicht als sicherer Befund betrachtet werden. Dazu wären weitere Forschungen für Gebiete mit besserer Datenlage notwendig.

Bei zusätzlichen Untersuchungen bezüglich der geographischen Anordnung der phylogenetischen Stämme konnte aber immerhin gezeigt werden, dass teilweise ein Zusammenhang vorhanden zu sein scheint zwischen dieser und der Verbindung mit den Kostenpfaden.

## 7.2 Grenzen der Methodik

Da die durchgeführte Untersuchung einen explorativen Charakter aufweist, gab es gleich mehrere Punkte, bei der die verwendete Methodik an ihre Grenzen stiess.

Auf die geographischen Unsicherheiten wurde bereits anfangs der Arbeit im Kapitel 3 bezüglich der Rekonstruierbarkeit der Umweltbedingungen während der Bantu-Ausbreitung genau eingegangen. Die Berechnung der Multipunkte-Kostenpfade ergab zwar ein starkes Übergewicht des südlichen Hauptzweigs, jedoch wurden entgegen der Late-Split Hypothese jedes Mal auch einige Punkte über einen Nordostpfad erreicht. Dabei ist unklar, inwieweit die Late-Split Theorie durch zur Zeit der Ausbreitung vorhandene offene Korridore im Regenwald begünstigt wurde.

Des Weiteren zeigte sich, dass eine LCPA für ein Gebiet dieser Grössenordnung auch bei einer Auflösung von 1 km * 1 km mit erheblichen Rechneranforderungen verbunden ist. Aus diesem Grund konnte die Anwendung einer Monte-Carlo-Simulation zur Überprüfung der Stabilität des Modells nicht durchgeführt werden. Die vorhandenen technischen Mittel stiessen an ihre Grenzen.

Ein weiterer Punkt, wo die gewählte Methodik wohl an ihre Grenzen stiess, stellte das Nichtvorhandensein einer Übereinstimmung der LCP-Distanzcluster mit dem genealogischen Baum dar. Wie im vorherigen Diskussionsteil erläutert wurde, wäre dazu wohl ein erweiterter Ansatz notwendig. Jedoch ist ein vorsichtiger Umgang mit diesem negativen Resultat angebracht, da dabei auch die Unsicherheit bezüglich der Datenlage einen entscheidenden Einfluss gehabt haben könnte.

## 7.3 Ausblick

Trotz der unbefriedigenden Übereinstimmung der aus den Kostenpfaden extrahierten Cluster mit dem genealogischen Baum liegt noch einiges Potential in der LCPA-Methode als komplementärer Ansatz zu vorhandenen Studien.

Eine der interessantesten Optionen stellt hierbei die erwähnte Methode der Betweenness Centrality dar. Mit diesem Ansatz könnte unter Umständen die Struktur der erhaltenen Distanzen bezüglich der Übereinstimmung mit dem genealogischen Dendrogramm noch entscheidend verbessert werden.

Jedoch darf der hohe Unsicherheitsfaktor bei der geographischen Datenlage im Falle der Bantu-Ausbreitung nicht vergessen werden. Da bei diesem mit der Vegetation ein zeitlich recht unkonstanter Faktor den verwendeten hauptsächlichen Parameter darstellt, wäre unter Umständen die Anwendung bei einem anderen Untersuchungsgebiet angebracht. Natürlich müsste zuerst mithilfe sprachwissenschaftlicher Experten abgeklärt werden, welches Beispiel sich aus linguistischer Sicht dazu eignete. Die genaue Abklärung der sprachlichen Gegebenheiten im Vorfeld zur Arbeit stellte dann auch einen weiteren wichtigen Aspekt dar, der zu beachten wäre. Hierbei sollte beispielsweise schon vor der Untersuchung sichergestellt werden, dass die gewählten Punktdaten möglichst die gesamte ungefähre Sprachdichte eines Gebietes wiedergeben. Auf diese Weise könnte ausgeschlossen werden, dass sich aufgrund einer nicht repräsentativen Auswahl von Sprachdaten ein verfälschtes Resultat ergibt. Generell sollten die bei der Arbeit verwendeten Sprachdaten möglichst frühzeitig vorbereitet werden. Gerade etwa bei der Verwendung von Sprachdaten aus verschiedenen Quellen kann es äusserst umständlich und zeitaufwendig sein, diese genau aufeinander abzustimmen. Denn oftmals werden die Sprachdatensätze nicht genau nach den gleichen Richtlinien erstellt.

In technischer Hinsicht sollte ebenso bereits im vornherein abgeklärt werden, ob die geplanten Modellierungen mit der gewählten Software sowie der gegebenen Hardware realisierbar sind. Da die Berechnungen von Kostenpfaden für grosse Untersuchungsgebiete enorm speicheraufwendig sein können, können so mögliche Engpässe frühzeitig verhindert werden.

# Literaturverzeichnis


ADAMS, W.M., GOUDIE, A.S. and ORME, A.R. (1996): *The Physical Geography of Africa*, Oxford University Press, Oxford.

ALVES, I., COELHO, M., GIGNOUX, C., DAMASCENO, A., PRISTA, A. and ROCHA, J. (2011): Genetic homogeneity across Bantu-speaking groups from Mozambique and Angola challenges early split scenarios between East and West Bantu populations, *Human biology* 83 (1): 13–38.

ANDERSON, David G. (2012): Least Cost Pathway Analysis in Archaelogical Research. Approaches and Utility, In: Least Cost Analysis of Social Landscapes, 239-252.

ATKINSON, David M., DEADMAN, Peter, DUDYCHA, Douglas, TRAYNOR, Stephen (2005): Multi-criteria evaluation and least cost path analysis for an arctic all-weather road, In: Applied Goegraphy 25, 287-307.

BATTEN, David C. (2007): Least-Cost Pathways, Exchange Routes, and Settlement Patterns in Late Prehistoric East-Central New Mexico, Department of Anthropology and Applied Archeology, Eastern New Mexico University.

BERRY, Joseph K. (2000): Optimal Path Analysis and Corridor Routing. Infusing Stakeholder Perspective in Calibration and Weighting of Model Criteria, Geography Departement, University of Denver.

BEVAN, Andrew (2011): Computational models for understanding movement territory, In: Mayoral, Herrera V., Celestino, Perez S. (eds.) Tecnologías de información geográfica y análisis arqueológico del territorio: Actas del V Simposio Internacional de Arqueología de Mérida. (383 - 394).

BOSTOEN, Koen (2012): How the „savanna corridor“ faclitated the Bantu Expansion. A lexical approach to pioneer tree species, In: Impact D'une Crise Environnementale Majeure Sur Les Espèces, Les Populations Et Les Communautés : La Fragmentation De La Forêt Africaine à La Fin De l'Holocène.

BOSTOEN, Koen, GROLLEMUND, Rebecca and MULUWA, Joseph Koni (2013): Climate-induced vegetation dynamics and the Bantu Expansion. Evidence from Bantu names for pioneer trees (Elais guineensis, Canarium schweinfurthii, and Musanga cecropioide),*Comptes Rendus Geoscience*, Volume 345, Issue 7, Pages 336-349.

CGIAR-CSI (2008): SRTM 90m Digital Elevation Data. [Online], Consortium for Spatial Information. http://srtm.csi.cgiar.org. [Zugriff:18.8.2013]

CHOI, Yosoon, UM, Jeong-Gi and PARK, Myong-Ho (2013): Finding least-cost paths across a continuous raster surface with discrete vector networks, In: Cartography and Geographic Information Science, Volume 41, Issue 1.

COLLISCHONN, Walter and PILAR, Victor (2000): A direction dependent least-cost-path algorithm for roads and canals, International Journal of Geographic Information Science, Vol. 14, 397-406.

CORMEN, Thomas H., LEISERSON, Charles E., RIVEST, Ronald L., STEIN, Clifford (2009): Introduction to Algorithms (3rd ed.), MIT Press and McGraw-Hill.

CSARDI, Gabor and NEPUSZ, Tamas (2006): The igraph software package for complex network research, Inter Journal, Complex Systems 1695.

CURRIE, T. E. and MACE, R. (2009): Political complexity predicts the spread of ethnolinguistic groups, Proceedings of the National Academy of Sciences 106 (18), 7339-7344.

CURRIE, Thomas, E., MEADE, Andrew, MYRTILLE, Guillon and MACE, Ruth (2013): Cultural phylogeography of the Bantu Languages of sub-Saharan Africa, Proc. of the Royal Society B 280.

DE FILIPO, Cesare, BOSTOEN, Koen, STONEKING, Mark and PAKENDORF, Brigitte (2012): Bringing together linguistic and genetic evidence to the Bantu expansion, Proceedings of the Royal Society B 279, 3256-3263.

DIAMOND, Jared and BELLWOOD, Peter (2003): Farmers and their Languages. The First Expansion, Science, Vol. 300, 597-603.

DIJKSTRA, E. W. (1959): A note on two problems in connexion with graphs. Numerische Mathematik, 1(1): 269-271.

EHRET C. 2001 Bantu expansions: re-envisioning a central problem of early African History, The International Journal of African Historical Studies. 34, 5 – 41.

ELENGA, H., SCHWARTZ, D. and VINCENS, A., 1992. Changements climatiques et action anthropique sur le littoral congolais au cours de l'Holocene. Bulletin de la Societe Geologique de France, 163: 83-90.

ESRI (2013): ArcGIS Desktop. Release 10.2 Redlands [Software], CA: Environmental Systems Research Institute.

ESRI (2010): World Major Rivers, CA: Environmental Systems Research Insitute. www.arcgis.com/home/item.html?id=44e8358cf83a4b43bc863646cd695945 [Zugriff 22.8.2013]

FAO (2007): GeoNetwork. Digital Soil Map of the World. [Online], Food and Agriculture Organization of the United Natiions. http://www.fao.org/geonetwork/ [Zugriff: 10.7.2013]

FIELD, Julie S., PETRAGLIA, Michael D., LAHR, Marta Mirazon (2005): The southern dispersal hypothesis and the South Asian archaelogical record. Examination of dispersal routes through GIS analysis, Science Direct, Journal of Anthropological Archaelogy 26, 88-108.

FRENZEL, B., PECSI, M. and VELICHKO A. A. (1992): Atlas of Paleoclimates and Paleoenvironments of the Northern Hemisphere. Late Pleistocene - Holocene.

GAVIN, M., BOTERO, C. A., BOWERN, C., COLWELL, R. K., DUNN, M., DUNN, R. R., GRAY, R. D., KIRBY, K. R., MCCARTER, J., POWELL, A., RANGEL, T. F., STEPPE, J. R., TRAUTWEIN, M., VERDOLIN, J. L., and YANEGA, G. (2013): Toward a Mechanistic Understanding of Linguistic Diversity, BioScience 63(7): 524-535.

GRINSTEAD, Charles, SNELL, Laurie J. (1997): Introduction to Probability. American Mathematical Society.

HAMMARSTRÖM, Harald (2012): The Language of Families of the World. A Critical Synopsis.

HERZOG, Irmela (2010): Theory and practice of cost functions, In: Melero et al., 2010, 431-434.

HERZOG, Irmela (2013): The potential and limits of Optimal Path Analysis, in A. Bevan and M. Lake (eds) Computational Approaches to Archaeological Spaces, Walnut Creek (CA): Left Coast Press, 179-21.

HERZOG, Irmela and POSLUSCHNY, Axel (2008): Tilt – Slope-Dependent Least Cost Path Calculaions Revisited, Proceedings of the 36th CAA Conference, Budapest.

HOLDEN, Claire J. (2002): Bantu language trees reflect the spread of farming across sub-Saharan Africa. a maximum-parsimony analysis, Proc R Soc. Lond. B. 269: 793-799.

HOLDER, Mark and LEWIS, Paul O. (2003): Phylogeny Estimation. Traditional and Bayesian Approaches, in Natur Reviews, Vol 4, 275-284.

HOWEY, Meghan C.L. (2007): Using multi-criteria cost surface analysis to explore past regional landscapes. a case study of ritual activity and social interaction in Michigan, AD 1200-1600, Journal of Archeological Science, Vol. 34, 1830-1846.

HUBER, Dennis L. and CHURCH, Richard L. (1985): Transmission corridor location modeling, In: Journal of Transportation Engineering 111 (2), 114- 130.

Kantner, John (2012): Realism, Reality and Routes,. Evaluating Cost-Surface and Cost Path Algorithms, In: Least Cost Analysis of Social Landscapes, 225-237.

KÖHLI, Martina (2013): Quantitative Analyse des Zusammenhangs zwischen der globalen Sprachendiversität und geographischen Faktoren. Masterarbeit, Geographisches Institut der Universität Zürich.

LIVINGOOD, Patrick (2012): No Crows Made Mounds, In: Least Cost Analysis of Social Landscapes, 174-187.

LLOBERA, M. (2000): Understanding movement. A pilot model towards the sociology of movement, In: Lock, 2000, 65-84.

LLOBERA, M. (2000): Understanding movement. A pilot model towards the sociology of movement, In: Beyond the Map, 65-84.

LÜSCHER, C. (2004): Bodendaten – ein Werkzeug für Planung, Nutzung und Schutz des Lebensraumes Boden.

MALCZEWSKI, Jacek (1999): GIS and Multi-Criteria Decision Analysis, Department of Geography, University of Western Ontario.

MCGREGOR, W. B., 2009. Linguistics. An introduction. London: Continuum International Publishing Group Ltd.

MONTANO, Valeria, FERRI, G, MARCARI, Veronica, BATINI, Chiara, ANYAELE, O, DESTRO-BISOL, G, and COMAS, D. (2011): The Bantu expansion revisited. A new analysis of Y chromosome variation in Central Western Africa. Molecular Ecology (13), 2693-2708.

MSU (2013): R Labs for Community Ecologists [Online], Montana State University, Bozeman, MT 59717. http://ecology.msu.montana.edu/labdsv/R/labs. [Zugriff: 24.4.14].

NETTLE D. 1996. Language diversity in West Africa: An ecological approach. Journal of Anthropological Archaeology 15: 403–438.

NEWMAN, James L. (1995): The Peopling of Africa. A Geographic Interpretation, Yale University Press, New Haven.

NEWMAN, M. E. J. (2003): The structure and function of complex networks, In: SIAM Review 45, 167-256.

NICHOLS, Johanna and BICKEL, Balthasar (2009): The Autotyp Genealogy and Geography Database. 2009 release. http://www.spw.uzh.ch/autotyp. [Zugriff: 10.3.2014]

NORDHOFF, Sebastian, Hammarström, Harald, Forkel, Robert and Haspelmath, Martin (eds.) 2013. Glottolog 2.2. Leipzig: Max Planck Institute for Evolutionary Anthropology.

PARADIS, Emmanuel (2006): Analysis of Phylogenetics and Evolution with R, New York, Springer.

PETIT-MAIRE, N. and GUO, Z.T. (1998): Mid-Holocene climatic change and Man in the present-day Sahara desert, Quaternary deserts and climatic change. Rotterdam, 351-356.

PINTO, Naiara and KEITT, Timothy H. (2009): Beyond the least-cost path. Evaluating corridor redundancy using a graph-theretic approach, in: Lanscape Ecology, 24, 253-266.

PLAZA, S, SALAS, A, CALAFELL, F, CORTE-REAL, F, BERTRANPETIT, J, CARRACEDO, A, COMAS, D (2004). Insights into the western Bantu dispersal. MtDNA lineage analysis in Angola, Human Genetics 115 (5): 439–47.

R Development Core Team (2013): R. A language and environment for statistical computing [software]. R Foundation for Statistical Computing, Vienna, Austria.

REES, W. G. (2003): Least-Cost paths in mountainous terrain, Computers & Geosciences 30 (2004) 203-209.

SAATY, T. L. (1980): The analytic hierarchy process. New York, Mc Graw-Hill.

SCHADEBERG, T. C. (2003): Historical linguistics, In: D. Nurse et G. Philippson (Ed.), The Bantu Languages. London-New York, Routledge, 143-63.

SHREVE, R. L (1966): Statistical law of stream stream numbers, In: Journal of Geology 74, 17-37.

SCHWARTZ, Dominique (1992): Assèchement climatique vers 3000 B.P. et expansion Bantu en Afrique centrale atlantique: quelques réflexions. Bulletin Soc. Géol. France, 163, 353-361.

STRAHLER, Arthur N. (1952): Hypsometric (Area-Altitude) Analysis of Erosional Topography. Geological Society American Bulletin, 63: 1117-1142.

SURFACE-EVANS, Sarah L. and WHITE, Devin A. (2012): An Introduction to the Least Cost Analysis of Social Landscapes, Least Cost Analysis of Social Landscapes. Archaeological Case Studies, the University of Utah Press, Salt Lake City.

TALIAFERRO, Matthew S., SCHRIEVER, Bernhard A. and SHACKLEY, M. Steven (2010): Obsidian procurement, least cost path analysis, and social interaction in Mimbres area of southwestern New Mexico, Journal of Archeological Science, Vol. 37, 536-548.

TOBLER, Waldo (1993): Three Presentations on Geographical Analysis and Modeling, University of California, Santa Barbara.

UERJ (2007): hydro-flow [software], Universidade do Estado do Rio de Janeiro, Faculdade de Geologia, Rio de Janeiro.

UCMP (2014): Understanding Evolution, University of California Museum of Paleontology,Berkeley. http://evolution.berkeley.edu/evolibrary/ [Zugriff: 10.3.2014]

UNEP (2000): The UNEP Environmental Data Explorer. Global 1Km Land Cover - UMD Legend [Online]. http://geodata.grid.unep.ch [Zugriff: 10.7.2013]

VAN ETTEN, Jacob (2012): R package gdistance. Distances and routes on geographical grids (version1.1-4) [software].

VAN GEEL B., BURMAN I. and WATERBOLK H.T. (1996): Archaeological and palaeoecological indications of an abrupt climate change in The Netherlands, and evidence for climatological teleconnections around 2650 BP, In: Journal of Quaternary Science, 451-460.

VAN LEUSEN, Martijn (2002): Line-of-Sight and Cost Surface Analysis using GIS, In: Pattern to Process. Methodological investigations into the formation and interpretation of spatial patterns in archaelogical landscapes, Rijksuniveriteit Groningen.

VANSINA, J. (1995): New Linguistic Evidence and the ‚Bantu Expansion, In: Journal of African History, 36, 173-195, Cambridge University Press.

WILLIAMSON, K. and BLENCH, R. (2000): Niger-Congo, In: Heine, B. and Nurse, D. (Ed.), African languages: An introduction, Cambridge University Press, 11-42.

WALS (2011): THE WORLD ATLAS OF LANGUAGE STRUCTURES. [Online]. http://www.wals.info. [Zugriff: 15.7.2013]

# Anhang



Zur besseren Weiterverwendbarkeit werden die bei der Arbeit verwendeten beziehungsweise resultierenden Daten gerade strukturell geordnet der CD beigefügt. Dabei handelt es sich vor allem um in *R* gespeicherte Variablen (Sprachnamen, Distanzmatrizen, hierarchische Cluster usw.) sowie Skripte welche zur Bearbeitung und Erstellung der Bäume und zur Berechnung der Kostenpfade verwendet wurden.. Deshalb ist dieser Teil hier sehr kurz gehalten.

*Abb. 1: Variation von Anfangspunkten für Kostenfaktor Vegetation (Standardgewichtung); Pfad von A3 (12°O,4°N) zu B4(31°O,1.5°N) (oben), Pfad von A5(9.42°O,4°N) zu B4 (unten)*

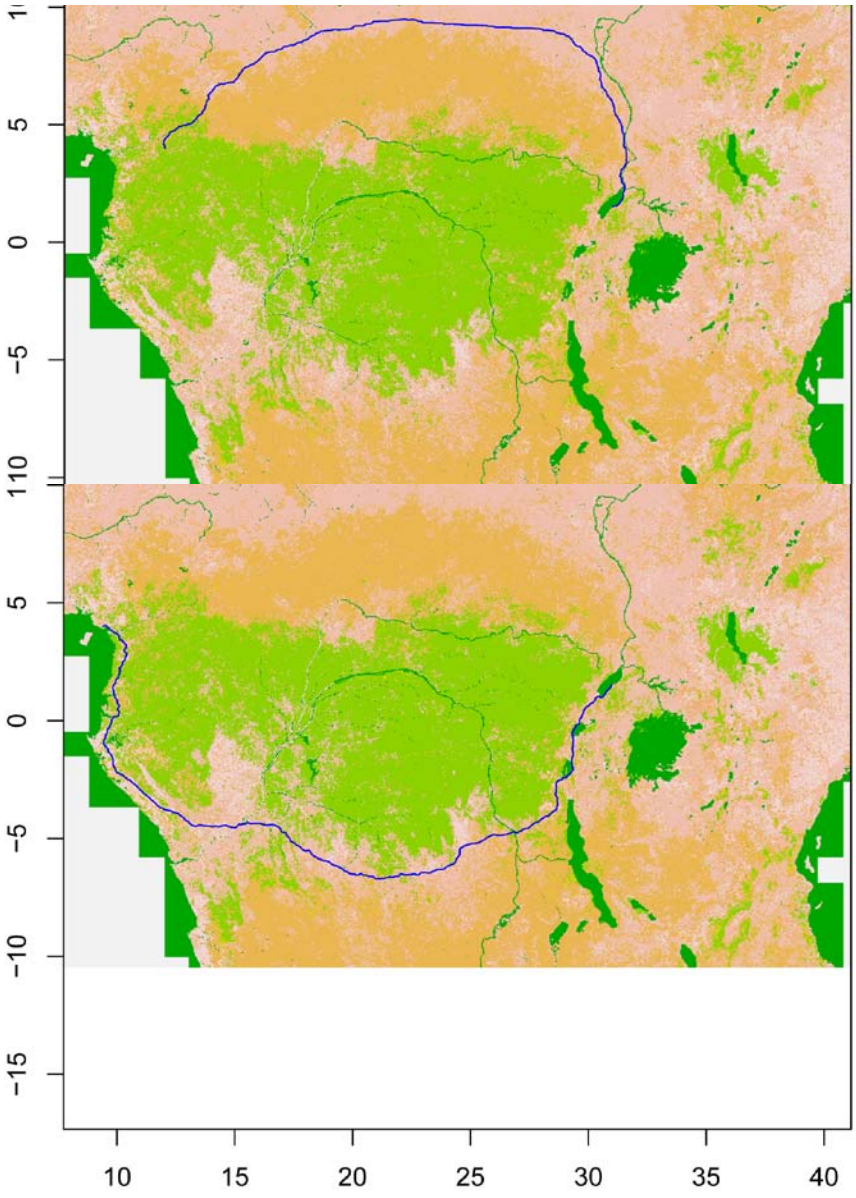


*Abb. 2: Variation technischer Parameter beim Kostenpfad vom Punkt A (11.6°O, 6.4°N) zum Punkt B3 (28.83°O, 1.17°S) für Kostenfaktor Vegetation (Standardgewichtung; ohne Geokorrektur (oben), 8-er Nachbarschaft (mitte) und 16-er Nachbarschaft (unten)*

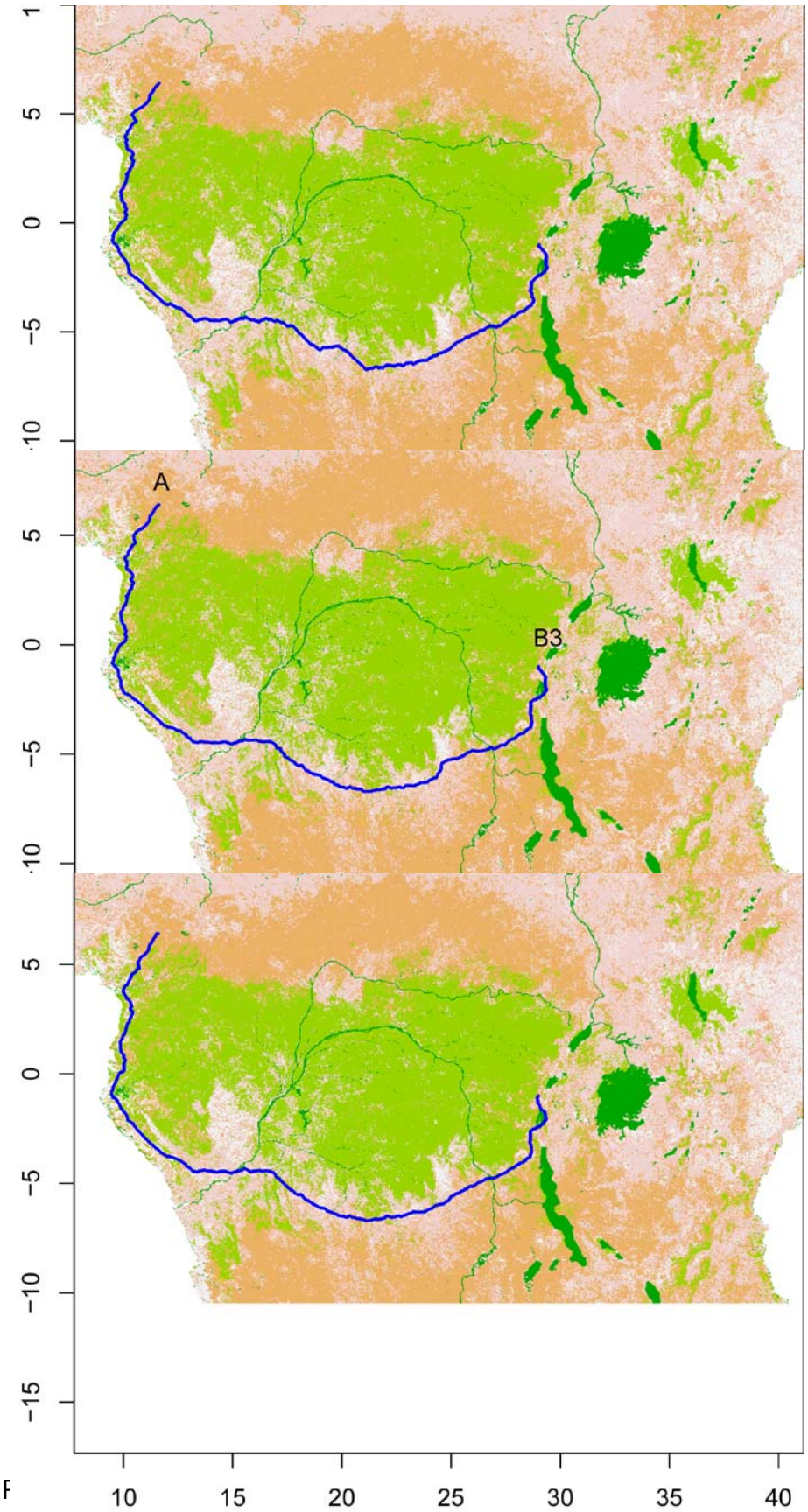


*Abb. 3: Kostenpfade bei Vegetation V3 mit Shreve-Ordnungen*

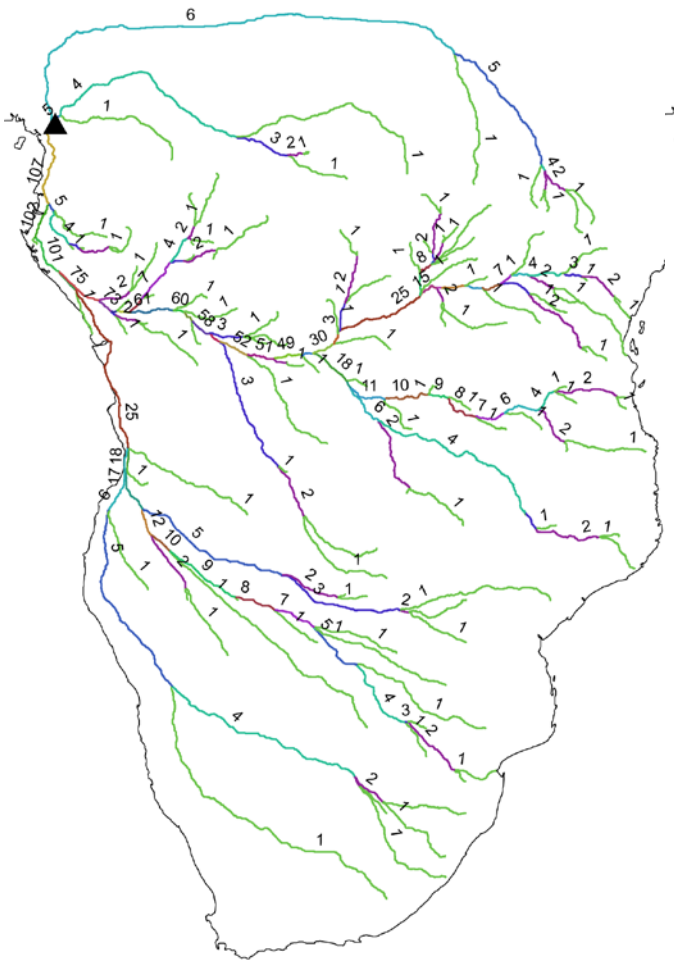


*Abb. 4: Kostenpfade bei Vegetation V5 mit Shreve-Ordnungen*

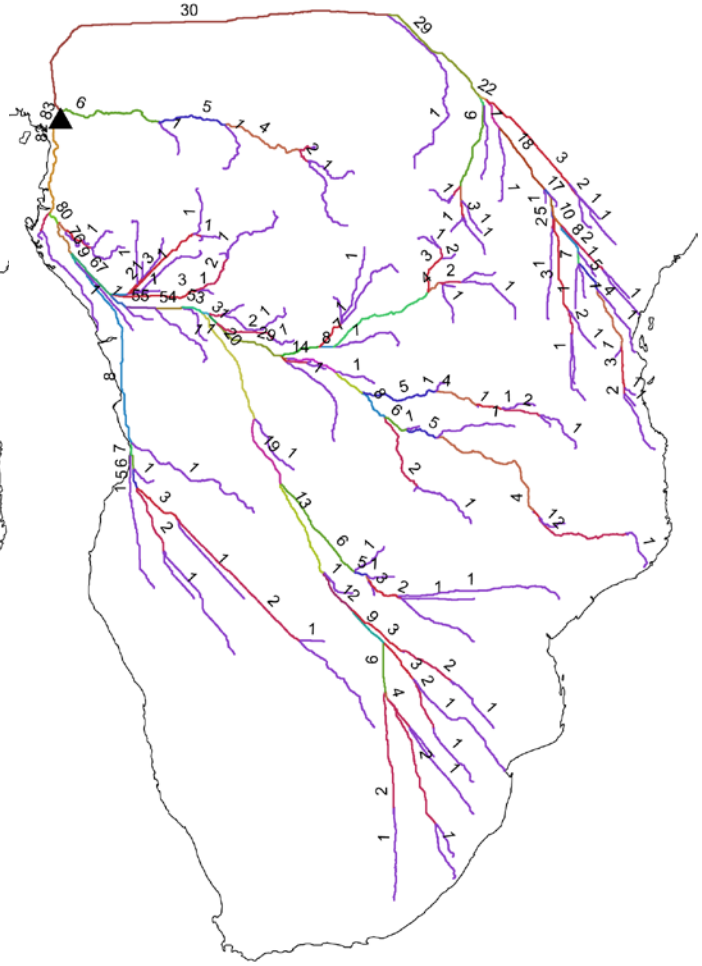
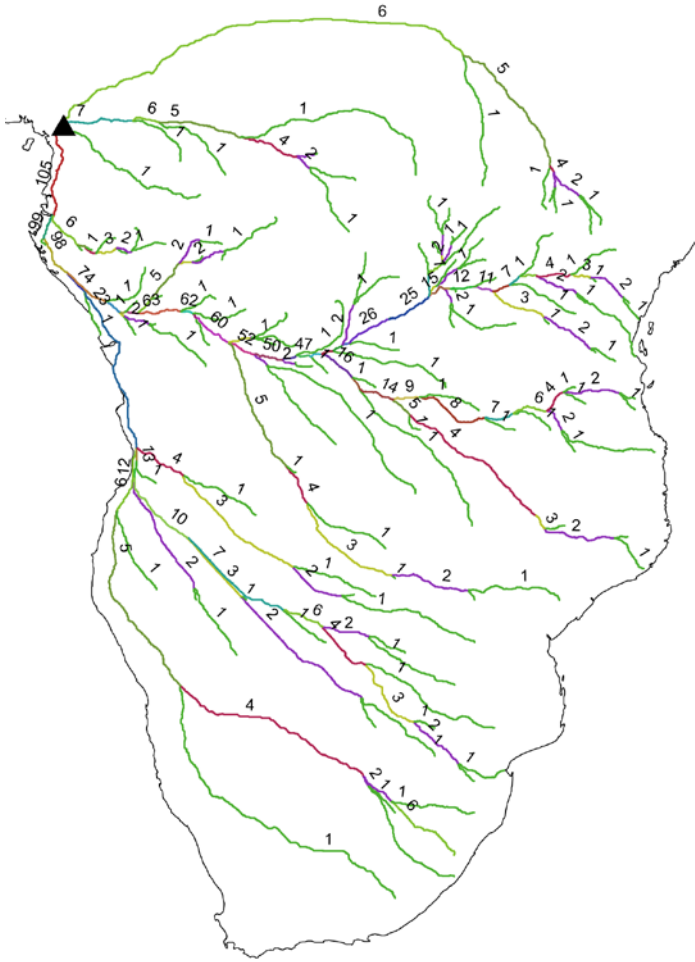


*Abb. 5: Kostenpfade bei Vegetation V6 mit Shreve-Ordnungen*



+

*Abb. 6: Kostenpfade bei Vegetation V7 mit Shreve-Ordnungen*

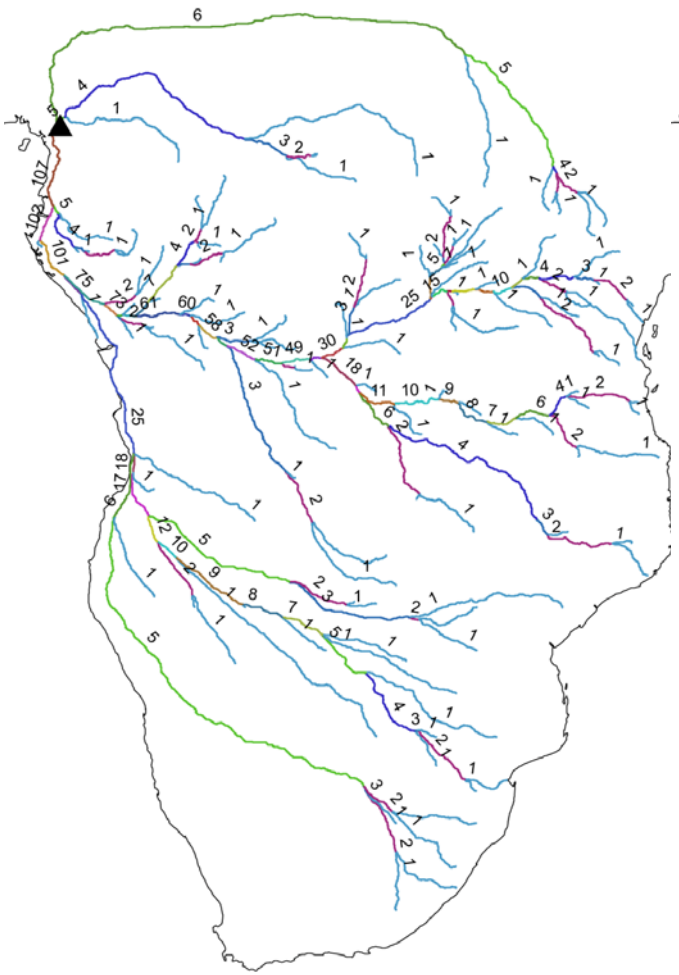


*Abb. 7: Kostenpfade bei Vegetation V8 mit Shreve-Ordnungen*

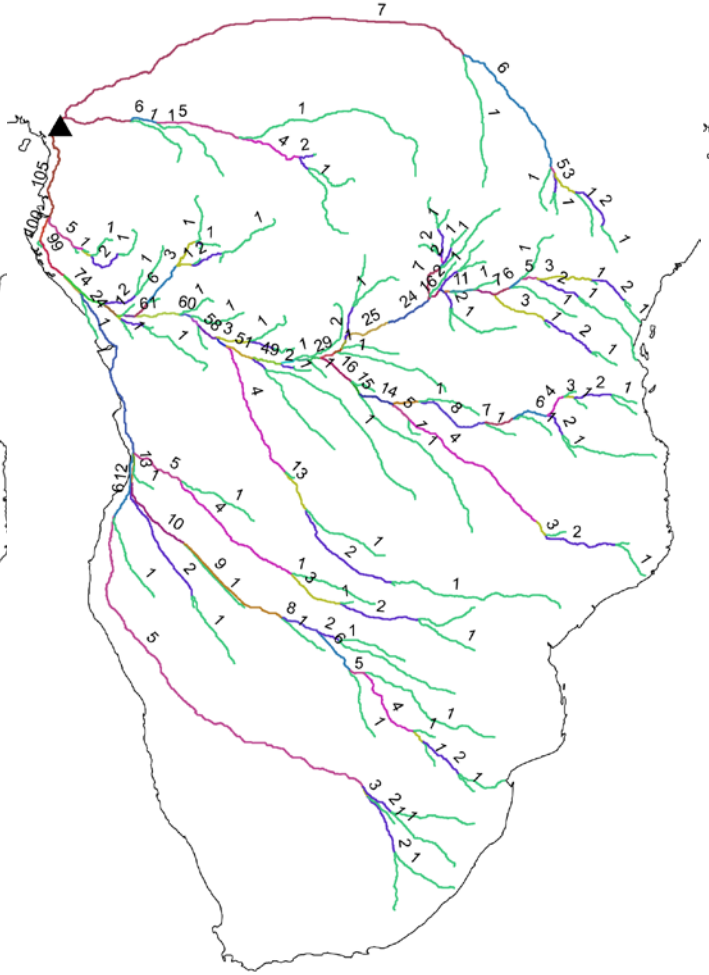


*Abb. 8: Kostenpfade bei Vegetation V9 mit Shreve-Ordnungen*

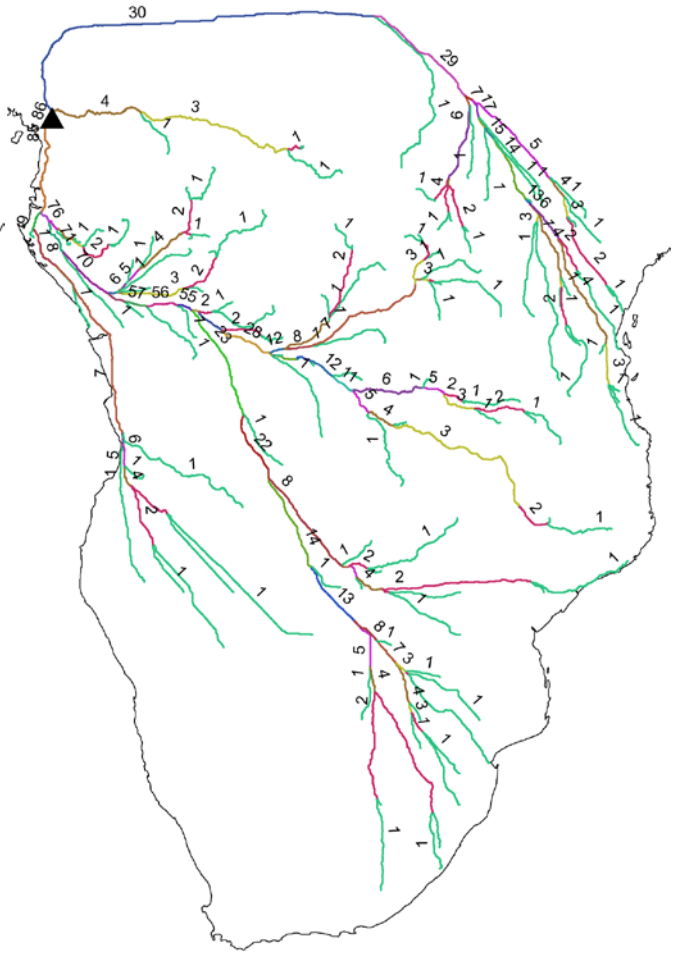
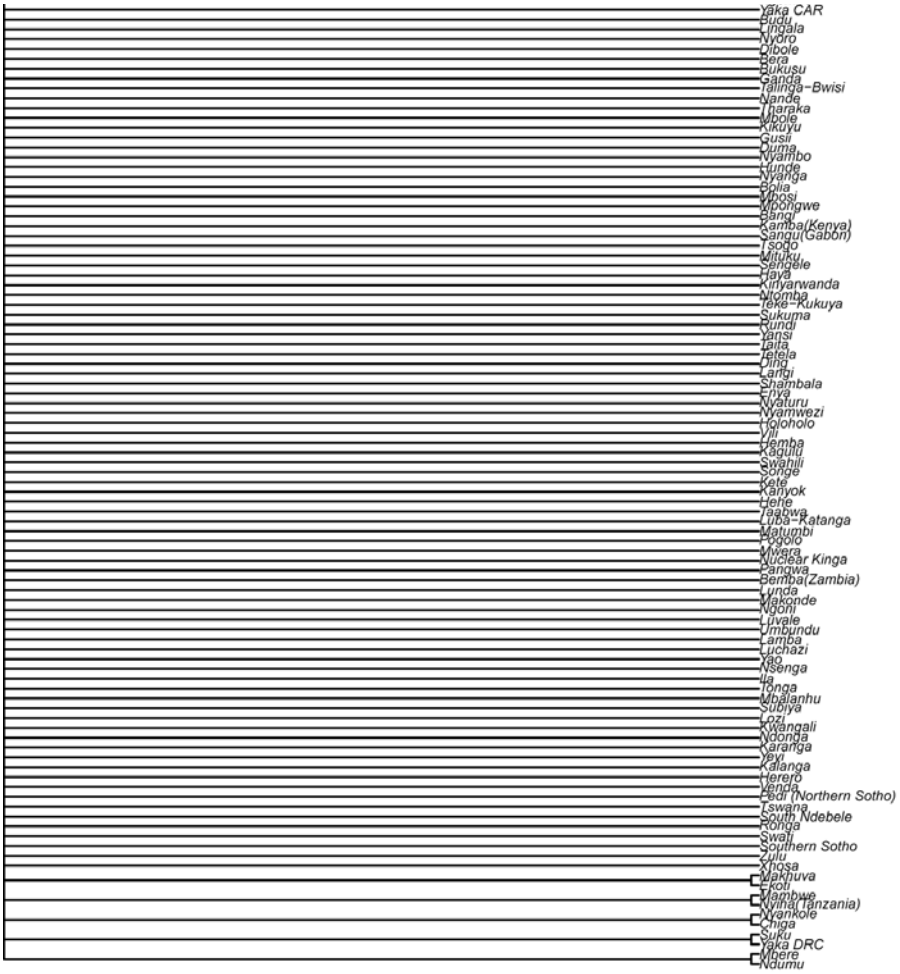


*Abb. 9: Consensus-Tree mit grösster Übereinstimmung zwischen aus LCP-generiertem Cluster (für topographische Distanzen bei V10) und genealogischem Baum (6 gleiche Nodes)*



# Persönliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Zürich, 30.6.14 Christian Wirth