Master's thesis

Geography

Geoinformatics

GEOGRAPHIC KNOWLEDGE DISCOVERY FROM SPARSE GPS-DATA – REVEALING SPATIO-TEMPORAL PATTERNS OF AMAZONIAN RIVER TRANSPORTS

Henrikki Tenkanen

2013

Supervisors:
Tuuli Toivonen
Maria Salonen

UNIVERSITY OF HELSINKI
DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY
DIVISION OF GEOGRAPHY

P.O. Box 64 (Gustaf Hällströmin katu 2)
FIN-00014 Helsingin yliopisto

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty/Section

Laitos – Institution – Department

Tekijä – Författare – Author

Työn nimi – Arbetets titel – Title

Oppiaine – Läroämne – Subject

Työn laji – Arbetets art – Level

Aika – Datum – Month and year

Sivumäärä – Sidoantal – Number of pages

Tiivistelmä – Referat – Abstract

Avainsanat – Nyckelord – Keywords

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty/Section Matemaattis-luonnontieteellinen tiedekunta	Laitos – Institution – Department Geotieteiden ja maantieteen laitos

Tekijä – Författare – Author

Tenkanen Henrikki Toivo Olavi

Työn nimi – Arbetets titel – Title

Liikkumistiedon louhinta perustuen harvaan GPS-dataan - Amazonian jokiliikenteen spatio-temporaaliset rakenteet

Oppiaine – Läroämne – Subject

Geoinformatiikka

Työn laji – Arbetets art – Level Pro gradu -tutkielma	Aika – Datum – Month and year Marraskuu 2013	Sivumäärä – Sidoantal – Number of pages 93 s. + liitteet

Tiivistelmä – Referat – Abstract

Informaatioteknologian ja erilaisten seurantajärjestelmien nopea kehitys viimeisten kahden vuosikymmenen aikana on mahdollistanut massiivisten spatio-temporaalisten tietovarantojen keräämisen. Paikannusteknologioilla varustetut laitteet ovat keskeisimpiä datalähteitä spatio-temporaalisen liikkumistiedon keräämiseen, ja tällainen data mahdollistaa erilaisten kohteiden (liikennevälineet, ihmiset jne.) liikkumisrakenteiden tutkimisen sekä erilaisten liikkumisparametrien kuten nopeuden, ja nopeuden sekä kulkusuunnan muutoksen laskemisen.

Tässä tutkimuksessa hyödynnetään eristyistä pilotti-seurantajärjestelmää (AROS), joka on kehitetty keräämään jokilaivojen liikkumisdataa Loreton ja Ucayalin seuduilla Perun Amazoniassa. AROS mahdollistaa reaaliaikaisten laivojen sijantitietojen (koordinaatit) sekä aikatiedon (aikaleima) keräämisen. Tässä tutkimuksessa kehitettiin erityinen liikkumistiedonlouhintaan tarkoitettu analyysityökalu (TRAT), joka hyödyntää useita spatiaalisen tiedonlouhinnan menetelmiä informaation louhimiseksi AROS datasta.

Tutkimuksessa tutkittiin, onko AROS datan perusteella jokinavigoinnissa nähtävissä vuodenaikaista vaihtelua vuoden 2012 aikana, ja vaikuttaako kulkusuunta sekä jokimorfologia navigointinopeuksiin. Tutkimuksessa tutkittiin myös, onko jokien vedenkorkeuksilla yhteyttä navigointinopeuksiin.

Tutkimuksen tulokset osoittivat, että navigointi vaihtelee riippuen vuodenajasta sekä kulkusuunnasta, ja myös viitteitä jokimorfologian vaikutuksesta navigointiin oli paikoittain nähtävissä. Meanderoivilla jokiosuuksilla navigoiminen alavirtaan oli n. 40 % nopeampaa korkeanveden aikaan, mutta matalanveden aikaan eroa nopeuksissa ei ollut juuri nähtävissä. Vuodenaikaisvaihtelu oli selkeintä alavirtaan kuljettaessa, jolloin navigointi korkeanveden aikaan oli n. 30 % nopeampaa verrattuna matalanveden aikaan. Anastomoivilla jokiosuuksilla erot nopeuksissa eri kulkusuuntiin olivat vähäisemmät, ja navigointi oli keskimäärin 20 % nopeampaa alavirtaan (verrattuna ylävirtaan). Vuodenaikaisvaihtelua ei ollut juurikaan nähtävissä. Lineaarien regressiomalli jokikorkeuksien ja yksittäisten osareittien navigointinopeuksien välille osoitti, että yhteys oli selkeä (R2=0.73) osareiteillä, jotka kulkivat Ucayali-jokea alavirtaan. Muissa tutkituissa tapauksissa selkää yhteyttä ei löytynyt.

Vertailemalla työn tuloksia aiempiin tutkimuksiin osoitti, että tulokset vaikuttavat olevan linjassa muiden tutkimusten tulosten kanssa. Työn tuloksia tulee jatkossa tosin vielä validoida vertailemalla vuoden 2012 tuloksia muiden vuosien tuloksiin AROS datan perusteella.

Liikennejärjestelmät ovat keskeisiä tekijöitä, jotka vaikuttavat alueiden yleiseen kehitykseen. Yksi tapa kuvata liikennerakenteita on tarkastella paikkojen välistä saavutettavuutta, jolla on todettu olevan merkitystä lukuisiin eri yhteyksissä kuten maankäytön muutoksessa, deforestaatiossa sekä luonnonsuojelussa. Tämän tutkimuksen tulokset voivat tarjota tarkempaa dataa ja informaatiota liittyen edellämainittujen aiheiden tutkimiseen Perun Amazoniassa ja mahdolllisesti muillakin Amazonin alueilla. Kehitettyä analyysityökalua (TRAT) on myös mahdollista hyödyntää laajemmissa yhteyksissä, kuten globaalin laivaliikenteen tutkimuksessa, tekemällä pieniä muutoksia työkalun algoritmeihin.

Avainsanat – Nyckelord – Keywords

GIS, spatiaalinen tiedonlouhinta, GPS, Perun Amazonia, spatio-temporaalinen analyysi, liikkumistiedon louhintamenetelmät, jokiverkkoliikenne, saavutettavuus

Säilytyspaikka – Förvaringställe – Where deposited

University of Helsinki, Kumpula Science Library

Muita tietoja – Övriga uppgifter – Additional information

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIS	Automatic Identification System
AROS	Amazonian Riverboat Observation System
ESRI	Environmental Systems Research Institute
ESTDA	Exploratory Spatio-Temporal Exploration
IIRSA	Initiative for the Integration of Regional Infrastructure in South America
INEI	National Institute of Statistics (Peru)
IQT-PUC	Iquitos-Pucallpa
IQT-YUR	Iquitos-Yurimaguas
ITC	International Training Center
GIS	Geographic Information System
GKD	Geographic Knowledge Discovery
GNSS	Global Navigation Satellite System
GOREL	Regional Government of Loreto (*Govierno Regional de Loreto*)
GPS	Global Positioning System
KDD	Knowledge Discovery in Databases
LUCC	Land Use and land Cover Change
MPO	Moving Point Object
Muste	Multiplatform Survo Type Editorial Environment for Data Analysis
R	Correlation, a measure of dependence / Statistical software
R^2	R-squared, the coefficient of determination
SA	Spatial Analysis
SDM	Spatial Data Mining

SEHINAV	The Department of Hydrography and Navigation of the Amazon (*El Servicio de Hidrografia y Navegación de la Amazonia*)
SPOT	Satellite Personal Tracker
STC	Space-Time Cube
SURVO	Environment for creative processing of text and numerical data
TRAT	Trajectory Reconstruction and Analysis Tool
UN-GGIM	United Nation’s committee of Global Geospatial Information Management

I. INTRODUCTION

1.1 Big (spatial) data

> ”Location information --- “*Analytical superfood*” that can and will, if used effectively and appropriately, improve people’s lives across the globe.”
> - UN-GGIM (2012)

It is said that 60-80 % of all data in the world has a spatial element (Franklin & Hane 1992) which illustrates the potentiality of spatial information. The fast growth of utilization of GIS technologies in business and areas of planning and environmental research as well as in education seems to confirm the statement of the United Nation’s new committee of Global Geospatial Information Management (UN-GGIM): GIS technologies and spatial information undoubtedly can, will and have improved people’s lives across the globe.

Big data is another common term that is used in many contexts nowadays and also a new research field has evolved to study and develop methods to analyze this data. The use of the term big data has emerged since massive amount and variety of data is collected continuously to databases by different devices and censors such as smartphones, navigators, social media sites, satellites etc. Today we create so much data that 90% of the data in the world has been created in the last two years alone (IBM 2013) and the produced mobile data is increasing at a rate of 600% a year (YLE 2013a). The resources that these databases contain are larger than what we can currently take advantage of and they provide countless opportunities for geographical research.

Spatial data is commonly combined with temporal information which enables to make temporal examinations and possibly to form movement patterns. The utilization of GPS (*Global Positioning System*) or *Global Navigation Satellite System* (GNSS) are currently the most dominant techniques for obtaining information about movements since many of our devices (like smartphones) have a GPS-receiver as a built-in feature. However there are also a variety of other techniques other than GPS or GNSS that are utilized for obtaining movement information such as Bluetooth for indoor tracking (e.g. Delafontaine et al. 2012 & Versichele et al. 2012), video information systems for obtaining for example the movements of players in team sports Moore et al. 2003) and radio telemetry methods (VHF) for animal tracking (Lesage et al. 2004) to mention a few.

A variety of tracking techniques are widely used also in practice especially related to traffic applications. A variety of monitoring systems have been developed for recording, managing and analyzing the information about the traffic based on different tracking and analysis methods. These systems are used e.g. for traffic congestion detection in urban areas (Krause & von Altrock 1996) and measuring the traffic volumes on the freeways (Bickel et al. 2007) which enable better road network planning to prevent traffic jams. Also specific public transportation map applications have been developed for the use of citizens e.g. in Finland where it is possible to follow and obtain real-time information about the movements and schedules of the trains (VR 2013) as well as trams and metros in the Helsinki region (HRT 2013). This kind of real-time information may provide time savings since the time used for waiting of means of transportation can be minimized. These few examples illustrate how people can benefit from different monitoring systems in their everyday life.

One of the most widely used monitoring system worldwide is Automatic Identification System (AIS) that is required from all of the passenger ships and internationally operating professional vessels that have gross tonnage larger than 300 tons. AIS is a GPS-based system that automatically records the vessel’s location and other information about the ship at specific time intervals and it can be followed by the ship operators, by the authorities as well as by the normal people (see Live Ships Map 2013). For this study a specific pilot monitoring system called Amazonian Riverboat Observation System (AROS) was built (by the Accessibility Research Group (2013), University of Helsinki) to collect movement data of the local river boats on the departments of Loreto and Ucayali in Peruvian Amazon. The monitoring system is basically a low-cost version of AIS i.e. it has seven GPS-devices which are connected to the GPS-satellites that are installed on different riverboats and the devices send the location information of the vessels with timestamps every 10 minutes. The monitoring system was developed since AIS is not required from the river boat operators in Peruvian Amazon and because of that - and the fact that the system is expensive - it has not been used in Peruvian Amazon.

The pilot monitoring system AROS - as well as many of the other tracking systems - provides a large database of spatio-temporally related data about the movements of the targets. Data, however is not equivalent to knowledge which means that it is necessary to harvest it from the data with specific techniques. Development of methods to derive knowledge from the movement data is one of the main targets of this thesis.

1.2 From static points to spatio-temporal movement patterns

A vast amount of spatio-temporal data has become available with the fast development of information technology and different monitoring systems over the last two decades. Position-aware devices are one of the most dominant sources for collecting movement data. Thus the analysis of the trajectories of moving point objects (MPO's) has gained a lot of interest in the scientific community during the recent years in the areas of geographic information science (GIS), human–computer interaction (HCI), ecology, biology, social and behavioral sciences (Dodge et al. 2009).

Spatio-temporal information that is derived from the tracking devices enable to build movement patterns from the targets and to calculate measurable motion parameters (so called *global descriptors*) such as speed, change of speed and the direction of movement (Laube & Imfeld 2002). Since the mobility data covers both spatial and temporal dimensions, it is possible to make queries that are related separately to space or time, or related to combination of these dimensions. Analyses of such combination, i.e. spatio-temporal data, enable to answer sophisticated questions like: "On which direction was the target moving at a certain time", "What was the average travel speed of the target when travelled in selected area of interest?", "Which place is most visited by the targets at different times of the day?" or "Which place is the most suitable for a hot-dog stand that draws the most attention at evening time?"

Numerous applications and techniques have been developed to analyze and track the movements of variety of targets such as cars and other vehicles (e.g. Dodge et al. 2009; Pelekis et al. 2012), players of sports (Iwase & Saito 2002; Moore et al. 2003; Laube 2005) and different animals ranging from whales to insects (e.g. Reynolds & Riley 2002; Gailey et al. 2007; Forester et al. 2009). The analysis of ship movements has not gained as much of attention as other vehicles in the scientific community but there are at least a couple of studies that have utilized AIS data for spatio-temporal analysis and visualization (see Willems et al. 2009; Demsar & Virrantaus 2010).

Even though the current data sources provide plenty of spatio-temporal movement data, and though it has more often been utilized for research recently, the majority of geographic researchers, analysts and cartographers have long been struggling to adapt the temporal aspects of data when dealing with spatial information. This fact has frequently been addressed also in scientific literature (Laube 2005; Andrienko et al. 2010; Keim et al. 2010) and the need to change our thinking from spatial to spatio-temporal has become a reality.

There are however reasons why the analysis of movement has not been particularly popular among researchers. The analysis of movement data and other types of spatio-temporal data can be challenging since exploring the data and deriving knowledge from it has commonly been possible only with automated algorithms and programming your own programs. The visualization of the data in a way that the results are understandable is a considerable challenge but progress in these areas has happened recently to make exploration of spatio-temporal data more user-friendly (an overview of these methods is represented by Keim et al. 2010). The data itself also produces challenges since it can be often heterogeneous, imprecise, incomplete and erroneous and the volume of data is often significant.

For this study a specific tool called TRAT (Trajectory Reconstruction and Analysis Tool) was programmed to manage the voluminous data and calculate different motion parameters such as speed and direction of movement among other descriptive variables which enable to examine the spatio-temporal movement patterns of the vessels in the study area.

1.3 The study context

The Amazon River is the world’s largest river by discharge. It runs across the northern South America originating from the Andes and entering the Atlantic Ocean in Brazil. The riverine regions/departments of Loreto and Ucayali in the Peruvian Amazon offer an exciting study area for a GIS-based transportation research. Almost all of the transportation at the study area is based on ferries and boats since the road infrastructure does not reach the Amazonian lowlands, thus making rivers the primary transportation network for local economy and community (Abizaid 2005; Salonen et al. 2012a).

Mobility and accessibility are closely related concepts. Mobility is essentially a measure of behavior whereas accessibility is a measure of potential (Hodge 1997). Basically accessibility determines what we are free to do or where we are able to go by the means of certain constraints (Hägerstrand 1970). An example of these constraints can be as ordinary as available time to reach a specific location by the means of transportation such as river boat. Accessibility therefore has an influence on our everyday decisions e.g. when we travel between locations.

Earlier studies have indicated that better accessibility increases the land-use pressure and human disturbance on the environment in Amazonia (e.g. Peres & Terborgh 1995; Angelsen & Kaimowitz 1999; Peres & Lake 2003). Therefore, it is important to understand the current accessibility patterns to be able to predict future land-use changes or evaluate socio-economic conditions in the region. Furthermore, as analysis of accessibility requires data which is often missing from areas like Peruvian Amazon, it is important to develop transferable methodologies to carry out relevant accessibility analysis. Understanding these patterns can be achieved with the spatio-temporal data obtained from the monitoring system AROS.

Earlier accessibility studies in Amazonia (e.g. Salonen et al. 2012a; Salonen et al. 2013) have concluded that when measuring accessibility patterns in environment such as Peruvian Amazonia, it would be important to take into account the dynamic nature of the river network (in addition to spatial structure) since the transportation characteristics along the rivers may vary significantly between seasons. Also Geurs & Wee (2004) have emphasized the importance to study spatio-temporal patterns thus allowing more accurate analysis of accessibility.

1.4 The objectives of the thesis

This thesis continues from the research of Salonen et al. (2012) and aims at revealing the dynamics of Amazonian riverine transportation patterns using spatio-temporal information derived from collaborating river boats that operate along the Amazonian rivers. The study aims at gaining information about the seasonal variation in movement patterns of the river boats, and to study how the river geometry and flow direction affect these patterns. To be able to study these patterns the first task is to developed methods to manage and reconstruct trajectories from the location information and thus further to derive knowledge from the AROS data via spatial data mining (i.e. geographic knowledge discovery). Measured movement patterns then again enable to aggregate and turn movement characteristics into more general measures such as accessibility patterns as means of time distance between locations.

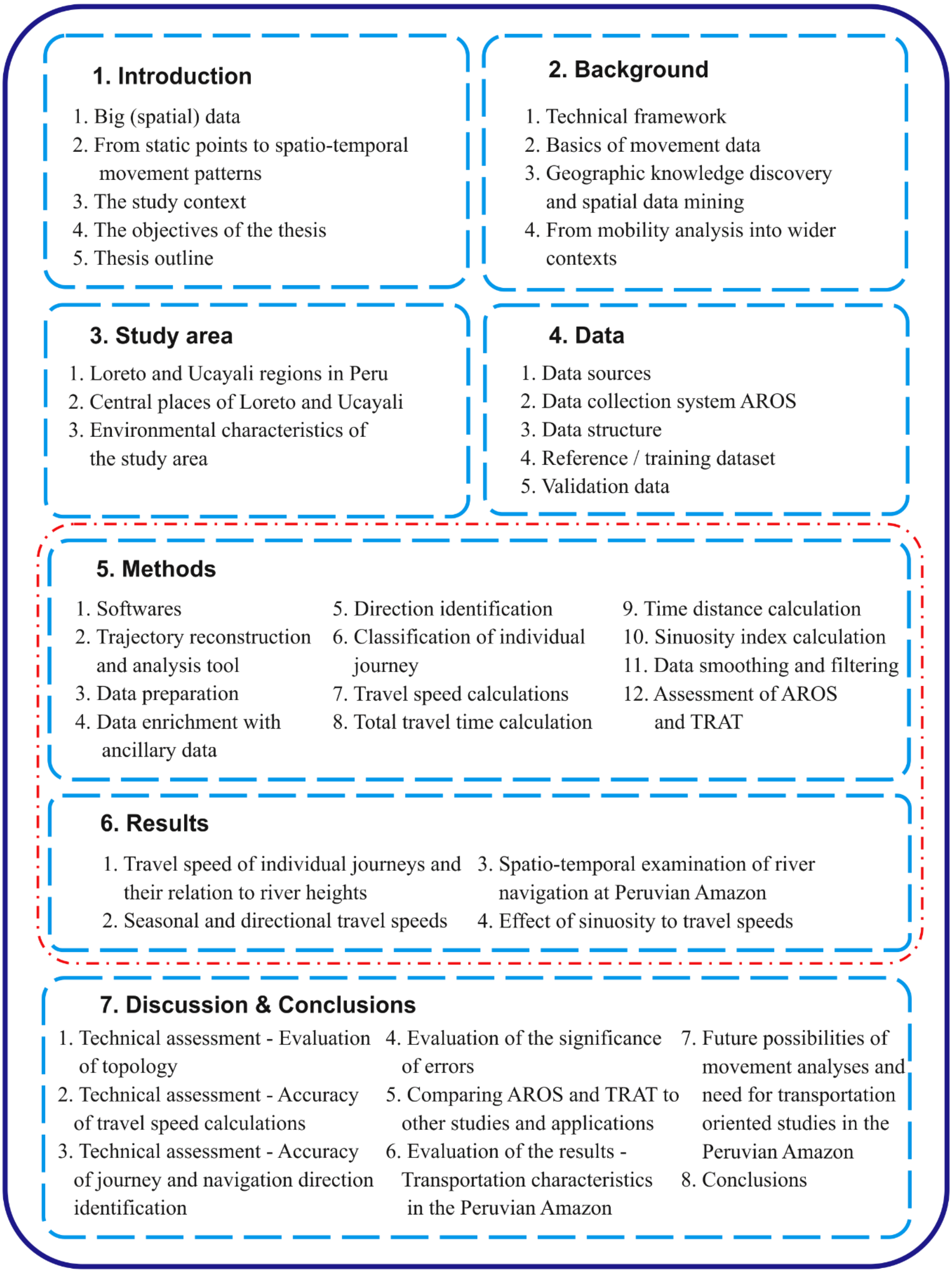
Thus this thesis has three key research questions or objectives:

1. Develop geographic knowledge discovery methods to extract knowledge from continuous location information provided by the river boat tracking system AROS
2. How the movement patterns of the vessels vary between seasons?
3. How the river geometry affects the movement patterns at the study area?

The main hypothesis of the thesis is that the dynamic nature of the Amazonian rivers affects the navigation speeds of the vessels and thus also accessibility patterns in the study area. Second hypothesis is that there is a connection between river geometry and navigation speed that can be modeled.

1.5 Thesis outline

Figure 1 represents the outline of this thesis. A certain matter has to be pointed out from the structure of this thesis: chapter 5 (methods), that represents the processes and algorithms used for knowledge discovery, could mostly be represented also as part of the results (chapter 6). Representing the developed knowledge discovery methods already in chapter 5 was however chosen because developed tool was used for analyzing the data from AROS. These tools enabled to characterize the riverine transportation patterns in Peruvian Amazon which are thus represented as results in chapter 6. Data and developed methods would allow to analyze various different aspects of transportation in the study area but only few specific aspects are covered in this thesis and represented in results. Also few methods that are used for representing the methods are exceptionally discussed along the chapters of results.



1. Introduction

1. Big (spatial) data
2. From static points to spatio-temporal movement patterns
3. The study context
4. The objectives of the thesis
5. Thesis outline

2. Background

1. Technical framework
2. Basics of movement data
3. Geographic knowledge discovery and spatial data mining
4. From mobility analysis into wider contexts

3. Study area

1. Loreto and Ucayali regions in Peru
2. Central places of Loreto and Ucayali
3. Environmental characteristics of the study area

4. Data

1. Data sources
2. Data collection system AROS
3. Data structure
4. Reference / training dataset
5. Validation data

5. Methods

1. Softwares
2. Trajectory reconstruction and analysis tool
3. Data preparation
4. Data enrichment with ancillary data
5. Direction identification
6. Classification of individual journey
7. Travel speed calculations
8. Total travel time calculation
9. Time distance calculation
10. Sinuosity index calculation
11. Data smoothing and filtering
12. Assessment of AROS and TRAT

6. Results

1. Travel speed of individual journeys and their relation to river heights
2. Seasonal and directional travel speeds
3. Spatio-temporal examination of river navigation at Peruvian Amazon
4. Effect of sinuosity to travel speeds

7. Discussion & Conclusions

1. Technical assessment - Evaluation of topology
2. Technical assessment - Accuracy of travel speed calculations
3. Technical assessment - Accuracy of journey and navigation direction identification
4. Evaluation of the significance of errors
5. Comparing AROS and TRAT to other studies and applications
6. Evaluation of the results - Transportation characteristics in the Peruvian Amazon
7. Future possibilities of movement analyses and need for transportation oriented studies in the Peruvian Amazon
8. Conclusions

Figure 1. The structure of the thesis.

II. BACKGROUND

2.1 Technical framework

Studying spatio-temporal transportation characteristics in Peruvian Amazon based on voluminous GPS-data requires to develop effective methods to reconstruct trajectories and extract knowledge from the voluminous spatio-temporal movement data collected with AROS. Thus the first research question (see 1.4) of this study involves many aspects that are rather technical in nature. Harvesting information from the data strongly relates to research areas of *geographic knowledge discovery (GKD)* introduced by Miller & Han (2001) which is a specific geographically oriented approach to *knowledge discovery in databases (KDD)* that is a set of methods for identifying high-level knowledge from low-level data (Fayyad et al. 1996). GKD and KDD are interdisciplinary approaches that integrates methods e.g. from machine learning, pattern detection, (geo)statistics, databases and (geo)visualization of data. At more specific level this study relates to analysis of movement which is a GDK approach focused on analyzing the patterns of movement data.

2.2 Basics of movement data

According to Andrienko et al. (2008) movement data consists of three principal components; 1) Time as a set of moments, 2) Population, a set of entities that move and 3) space, a set of locations that can be occupied by the entities. Dodge et al. (2009) define *moving point objects* (MPO) as "entities whose positions or geometric attributes change over time". Thus MPOs can be seen as a dynamic representation of a static point where each location is specified three-dimensionally by a tuple of *(x, y, t)* coordinates where *t* represents time (Hägerstand 1970; Hornsby & Egenhofer 2002; Laube 2005).

A sequence of successive positions of the moving object over a period of time forms movement pattern which can be represented and visualized as a space-time path which is a key concept related to time-geography first introduced by Hägerstrand (1970). Space-time path (or trajectory) basically traces an individual's movements and activities in space with respect to time (see Figure 2) (Miller & Bridwell 2009). Related to data mining tasks Spaccapietra et al. (2008) considers trajectory as a time series of spatial data ($t_{0..n,}$ x, y -points in Figure 2).

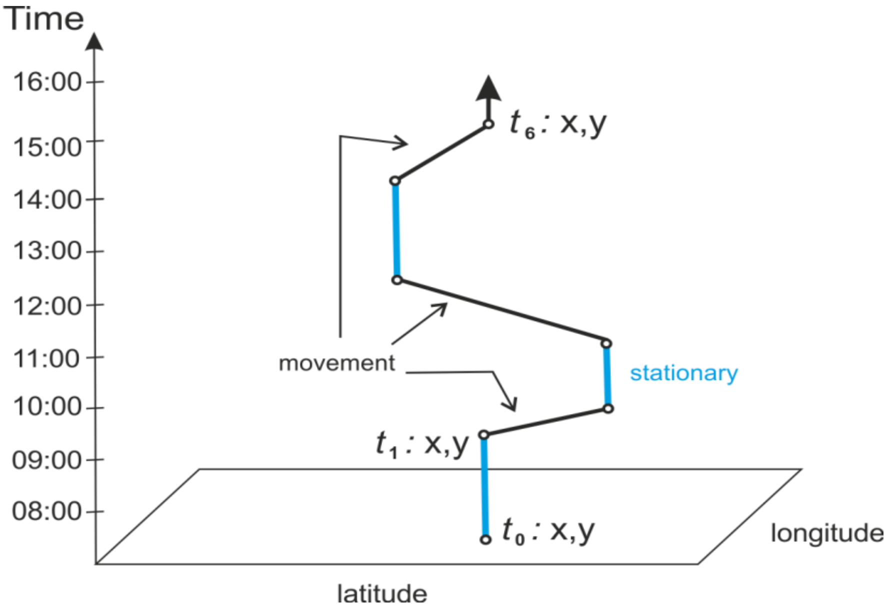


Figure 2. A concept of trajectory and moving point objects. Modified after Miller & Bridwell (2009).

The state of the moving object at a selected time moment can be characterized not only by its position in space but also by additional characteristics such as speed, direction and acceleration. The entities themselves (regardless of the movement) can also be characterized by *supplementary characteristics* such as ID, capacity, size, vessel type etc. Also the locations where the entities move can be characterized with information about river height, altitude, river type, sinuosity index etc. (Andrienko et al. 2008).

In order to analyze the behavior or patterns of the moving objects it is necessary to have detailed information about the trajectory of the object as well as information about the environmental conditions (i.e. supplementary characteristics) related to the trajectory (Spaccapietra et al., 2008). Combination of different datasets allows setting sophisticated research questions that take into account space and time as well as other additional attributes related to the studied phenomenon.

However, before there are any kinds of patterns to analyze, it is necessary to reconstruct those patterns from the GPS observations. This can be done via geographic knowledge discovery.

2.3 Geographic knowledge discovery and spatial data mining

”The world is data rich, but information poor.” - Han & Kamper (2011)

Geographic knowledge discovery (GKD), or *spatial data mining* (SDM) as it is also often called, is an active, growing and highly interdisciplinary research field that is still at a relatively early stage of its development. GKD focuses on the development of theory, methodology and practice for the extraction of useful information and knowledge from massive and complex spatial databases (Mennis & Guo 2009).

Knowledge discovery is exploratory in nature and it can be seen as more inductive than e.g. traditional statistical methods (upper half in Figure 3). In *inductive reasoning* (Hempel 1965) the aim is to form theories based on pattern and generalizations from the observations. Knowledge discovery also naturally fits in the initial stage of a *deductive reasoning* (lower half in Figure 3) where aim is to develop and modify the theories based on the discovered information from observation data (Miller & Han 2009) and thus increase the overall understanding of the studied phenomena.

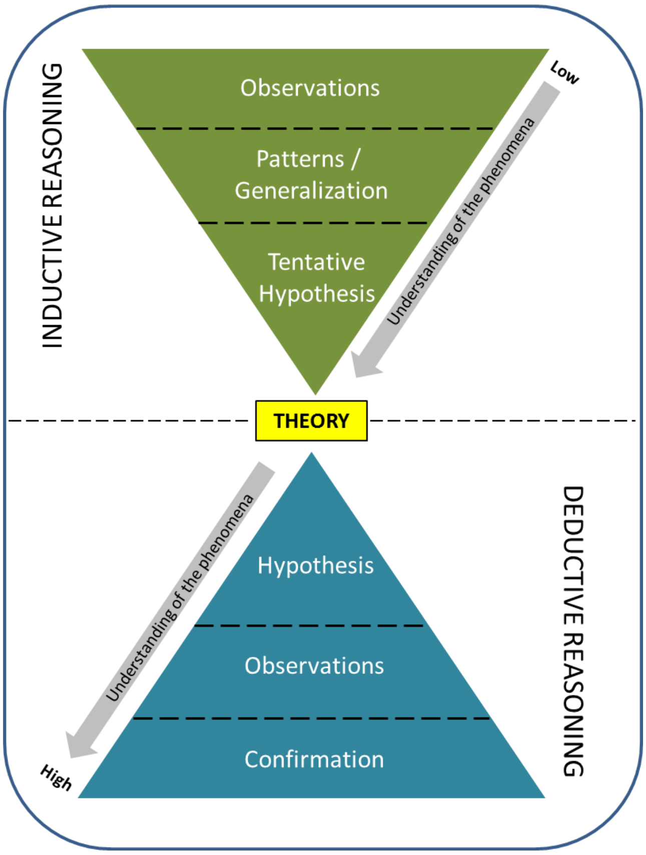


Figure 3. The aim of inductive reasoning and deductive reasoning is to increase the understanding about the studied phenomena (Hempel 1965).

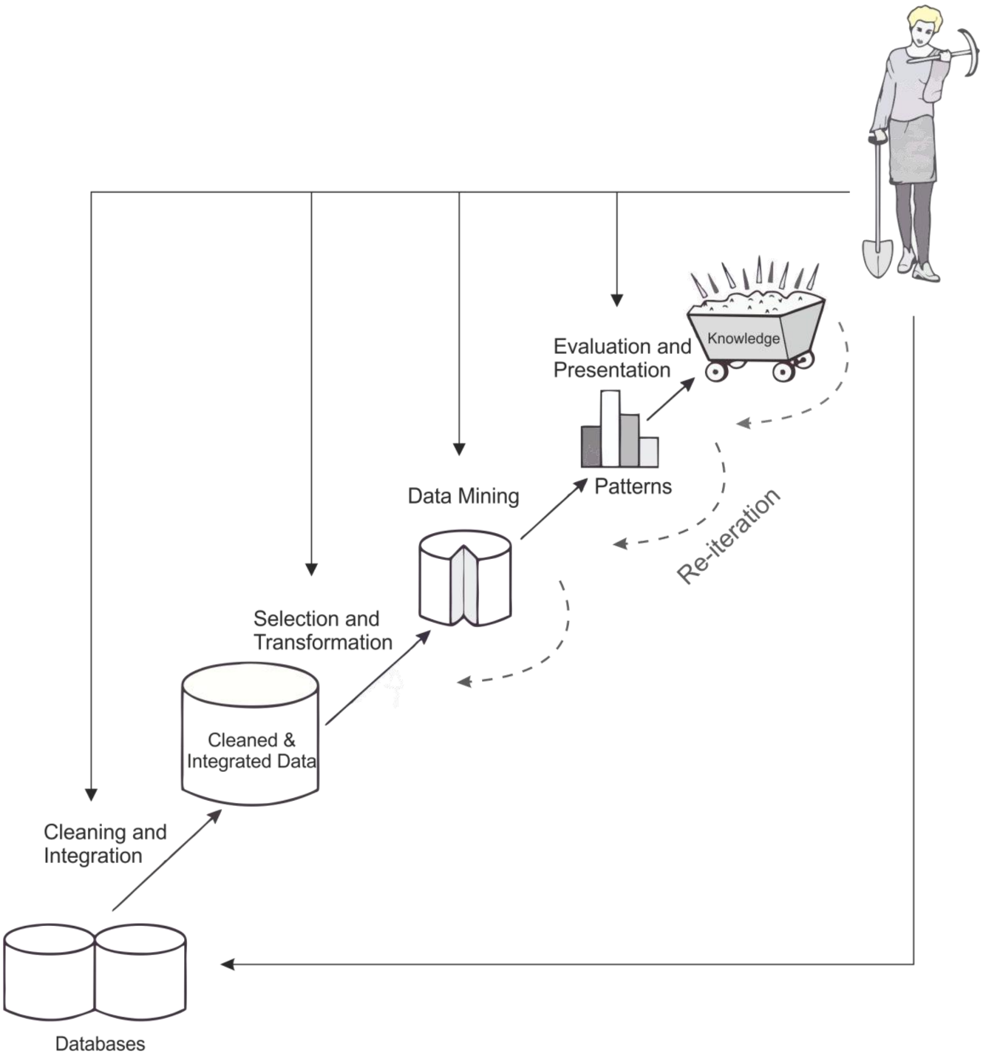
One might ask: *"Why should I use GKD instead of traditional spatial analyses? What is the difference?"* The reason why geographic knowledge discovery has emerged is because the traditional spatial analysis (SA) methods were developed in an era when data were relatively scarce and computational power was not as powerful as today which restricted the developed analysis methods at that time (Miller & Han 2009). Since we have moved from a data-poor era to a data-rich era, the traditional SA methods often do not meet the analysis needs of today. These shortcomings related to traditional spatial analysis according to Mennis & Guo (2009) are:

- They cannot process the large data volumes that is produced these days
- They do not know how to process newly emerged data types such as trajectories of moving objects
- They focus on a limited perspective (e.g. univariate spatial autocorrelation) and cannot suggest any better alternatives if the chosen perspective or model is appropriate for the phenomenon and do not show interesting relationships

Even though traditional SA methods are still highly needed in various research areas, the geographic knowledge discovery has emerged to fill the limitations listed above. Since the amount of data in the world is increasing at high speed, we can now obtain much more diverse, dynamic and detailed data than was ever possible before the modern data collection techniques (Goodchild, 2007). Such data provide opportunities for gaining new knowledge and better understanding of complex geographic phenomena such as human–environment interaction and socio– economical dynamics (Mennis & Guo 2009). Because of this diverse and voluminous data, it is necessary to automate the knowledge discovery process since it is not possible for a person to manually harvest the knowledge from these data sources. Thus automated knowledge discovery is the key principle of (spatial) data mining (Shekhar et al. 2011).

2.3.1 GKD as a process

Knowledge discovery as a whole is an iterative process that involves multiple (automatic) steps including 1) data cleaning, preprocessing and integration 2) data selection, transformation and incorporation of prior knowledge, 3) analysis/data mining with computational algorithms and/or visual approaches, 4) interpretation and evaluation of the results (see Figure 4). As an iterative process, the process chain includes also reformulation or modification of hypotheses and theories due to new knowledge (dashed arrows in Figure 4) as well as adjustment to data and analysis methods thus leading to re-evaluate the results on each iteration (Fayyad et al 1996; Han & Kamper 2001). Even though the emphasis on knowledge discovery is in automation, the whole process is conducted by the user since each step involves decisions that have been made by the analyst (uniform arrows in Figure 4) (Han & Kamper 2001).



Knowledge
Evaluation and Presentation
Patterns
Data Mining
Re-iteration
Selection and Transformation
Cleaned & Integrated Data
Cleaning and Integration
Databases

Figure 4. The concept of knowledge discovery which is an iterative process chain conducted by the user to derive knowledge from the data. Modified after Han & Kamper (2001: 6).

The described basis of knowledge discovery process is common to both traditional KDD as well as to GKD. The difference of GKD compared to KDD comes when taking a look at the data. The input data for geographic knowledge discovery is spatial, i.e. it is georeferenced, which causes similar effects as have been acknowledged in the area of geostatistics (see e.g. Cressie 1993; Goovaerts 1999) to relations between observation due to Tobler’s (1970) *first law of geography* (autocorrelation etc.). Also the complexity of geographical inputs (i.e. extended objects such as points, lines, polygons, rasters) restrain the use of general purpose data mining algorithms (Shekhar et al. 2011).

The nature of spatial data is especially important in data mining (3 step of knowledge discovery) where spatial relations effect on results. Data mining of spatial data includes various common and important tasks/analysis methods such as spatial classification and prediction (regression models), spatial outlier detection, spatial association rule mining and co-location pattern discovery, spatial cluster and hot spot analysis, and geovisualization (Mennis & Guo 2009; Shekhar et al. 2011).

Related to this thesis, the most relevant spatial data mining tasks are spatial classification and geovisualization. Therefore I will explain and cover only these methods in the forthcoming chapters.

2.3.2 Mobility data mining

> ”… the world of 2000 is desperately going to need men and women with a clear view and involved concern for Man’s use of space over time.” - Gould (1969)

One of the frontiers of GKD is the management and analysis of spatio-temporal data such as mobility data. Giannotti & Pedreschi (2008) refers this as *mobility data mining* which aims at understanding the movement behavior of different targets by analyzing the mobility data through appropriate patterns and models extracted by efficient algorithms. This novel knowledge discovery process is composed of three steps: trajectory reconstruction, knowledge extraction and delivery of the obtained information. These steps thus follow the principles of knowledge discovery process (as illustrated in Figure 4) where trajectory reconstruction responds to data transformation step, knowledge extraction to data mining step, and delivery of information to evaluation and presentation step.

2.3.3 Trajectory reconstruction

Trajectory reconstruction is the first task when processing mobility data. Transforming the movement data into trajectories of moving objects is however not a straightforward procedure as e.g. Marketos et al. (2008) have pointed out. Since the raw points (x- and y-coordinates with timestamp) arrive in bulk sets, it is necessary to have a filter that decides if the new series of data is to be appended to an existing trajectory or not. In other words the filter detects and separates individual trajectories from the data mass. It is necessary that this filter contains multiple generic parameters or “triggers” that do the filtering, such as temporal or spatial gabs between observations, tolerance distance, maximum speed and maximum noise duration (Marketos et al. 2008). In the case of this study also the direction of movement was used for filtering (see 5.2 for more details).

The computation of global descriptors (Dodge et al. 2009), or instantaneous movement descriptors as Laube et al. (2007) calls them, focus on characterizing the movement itself and can be seen as part of trajectory reconstruction or knowledge extraction, depending on how you want to look at it. These descriptors such as speed, acceleration, duration of movement, sinuosity, travelled path, displacement and direction (Giannotti & Pedreschi 2007; Laube et al. 2007; Dodge et al. 2008) need quite often to be smoothed (see 5.7 for more details) since the movement data can be quite noisy or fragmentary (Laube et al. 2007) that is caused by inaccurate observations and incomplete data. These global descriptors form fundamental building blocks for analyzing the movement with respect to its environment. This movement/environment relationship is a great interest for e.g. ecologists who try to find environmental cues that affect the motion of individual organisms (see Nathan et al. 2008), and naturally for this research as well that focuses on transportation.

2.3.4 Knowledge extraction

The second (or third) step in the mobility data mining process is knowledge extraction with spatio-temporal data mining methods such as classification or geovisualization. The aim of this phase is to reveal useful and interesting patterns out of trajectories (see Laube & Purves (2006) for discussion and methods about evaluating the motion interestingness).

Classification can be seen as discovery of behavior rules that aims at explaining the behavior of current events (or movement) and predicting that of future ones (Giannotti & Pedreschi 2008). Classification is about grouping data items into classes according to their properties, i.e. attribute values (Mennis & Guo 2009) via supervised (with training dataset) or unsupervised classification (i.e. clustering). There are various different classification methods such decision trees, artificial neural networks, maximum likelihood estimation, nearest neighbor methods and case-based reasoning. Related to this thesis a decision tree classification (supervised) is one part of the knowledge extraction process since the developed algorithm utilizes a training set, or ancillary data (see III), to categorize the mobility data of the vessels and calculate global descriptors (see Figure 5).

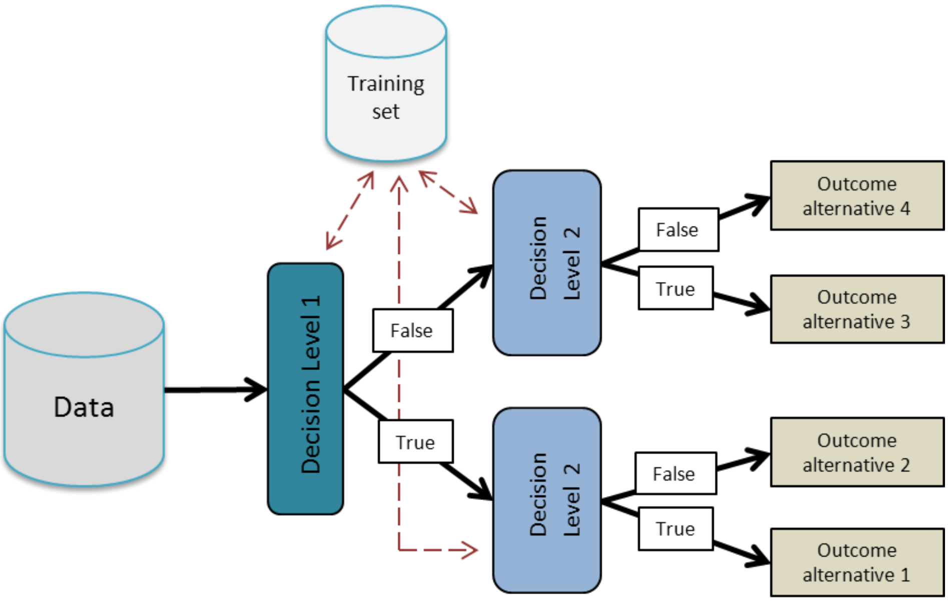


Figure 5. The concept of decision tree.

Making spatio-temporal queries from the mobility data can also be considered as a part of knowledge extraction and there are various different queries that can be implemented related to motion. These queries can be separated into location-based queries, continuous queries and trajectory-based queries (Agarwal et al. 2002). Location based queries answer to question such as *"Show all MPO's that are within 1 km from point A"* whereas continuous queries involves a moving location range and can answer to question such as *"Follow targets that are within 500 meters around car B"* (Raptopoulou et al. 2003). Trajectory based queries involve the topology of the trajectory and derived information related to the trajectory such as velocity or direction of movement. These queries can answer to questions such as *"Which targets are heading north?"* or *"Which targets were moving upstream with the speed of 15 km/h during February?"* Related to this study the trajectory-based queries are the most useful since the purpose of this study is to reveal changes in movement patterns according to seasonal changes in water levels and navigation direction (upstream/downstream).

Yet another type of knowledge extraction method is based on a novel multidisciplinary research field of geovisual analytics where the emphasis is on interaction between user and the data through interactive visualizations. It is good to acknowledge that this can be considered also as a part of knowledge delivery phase (see 2.3.5) since the knowledge extraction is based on visualization which is also a significant part of representing the results. The aim of geovisual analytics is to synthesize information and derive insight from very large and complex datasets for understanding, reasoning and decision making (Keim et al. 2010) by interactive spatial and temporal exploration of data through zooming, panning, grouping and selecting which can be defined as ESTDA (*Exploratory Spatio-Temporal Data Analysis*) (Andrienko & Andrienko 2005). ESTDA have power to reveal hidden patterns that are not visible through a single view or angle of the data. Geovisual analytics are used widely for representing movements in three dimensional form and discovering interesting patterns from mobility data (see Figure 6) that are easily overlooked by other analysis and visualization methods (e.g. Kwan 2000; Zhao et al. 2008; Demsar & Virrantaus 2010; Kraak 2011).

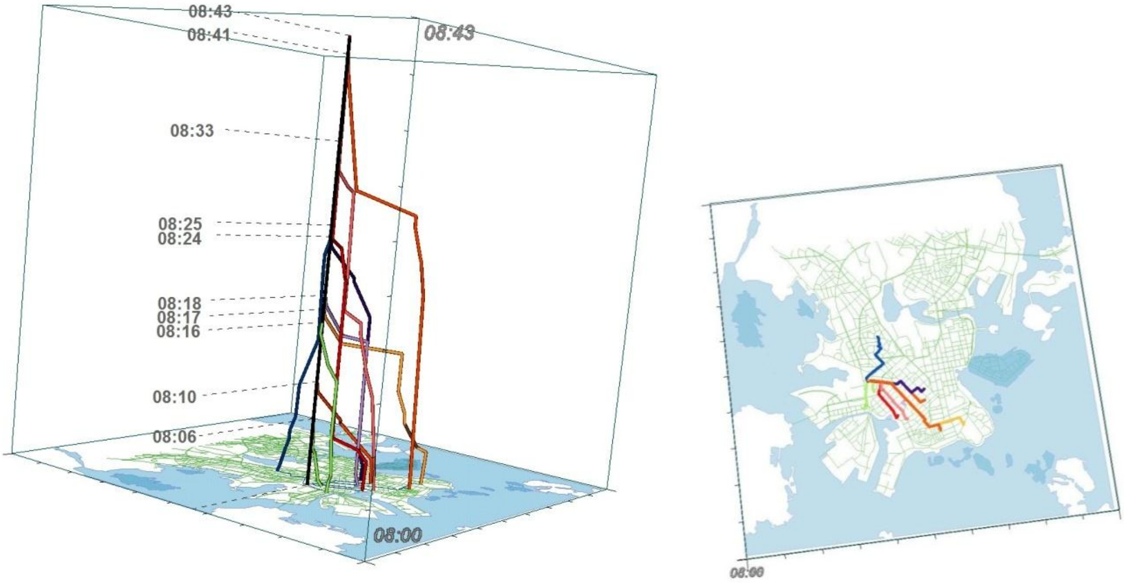


Figure 6. Trajectories of walking individuals and their footprint at downtown Helsinki. Time intervals indicate when individuals reach the common destination (upright black pole). Represented with uDig software attached with Space-Time-Cube plugin developed by ITC (University of Twente).

2.3.5 Knowledge delivery

The final step of GKD process is knowledge delivery, i.e. the presentation of the results and evaluating their information. Extracted patterns from the data can be considered extremely rarely as knowledge per se. Thus it is highly important and necessary to reason and compare the patterns with relevant background knowledge, refer them to other appropriate geographic information (Giannotti & Pedreschi 2008) and evaluate the interestingness of the patterns (Laube & Purves 2006).

Related to mobility data a widely used visualization and representation method of the results is utilization of space-time cube (STC) which was discussed in the previous chapter (2.3.4). However also other types of representation methods are widely used for representing spatio-temporal movement characteristics since STC can become quite “messy” when representing loads of trajectories simultaneously, and if interactive exploration is not possible (such as in printed papers) STC becomes less usable as a visualization method.

One typical way of representing movement characteristics (in 2D) is to use so called movement parameter profiles (Dodge et al. 2009) (see Figure 7) where time is represented at x-axis and movement parameter such as velocity (speed) or acceleration (change of speed) is represented at y-axis. Simple but useful way of comparing the movement characteristics of different trajectories together is to calculate a median or average of the parameter (e.g. speed) for every individual trajectory (righter most box in Figure 7) and then compare the deviations of those trajectories by plotting them all in the same graph.

In the case of this thesis a so called route point which indicates the positions of the network was used. This kind of representation allows linking a selected position of x-axis straightly to its geographic location.

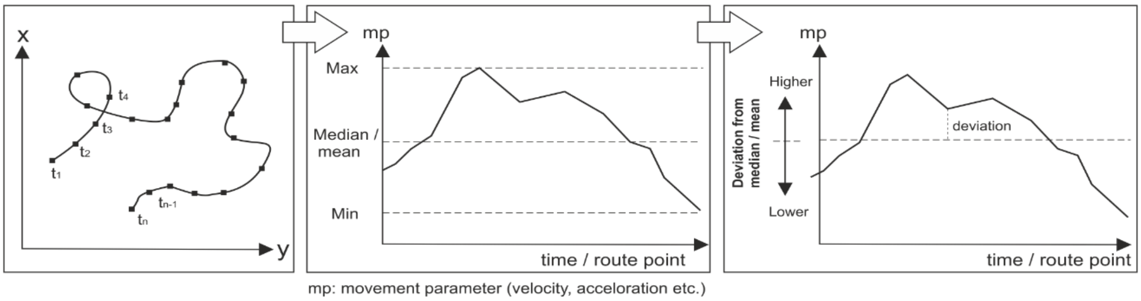


Figure 7. The concept of 2D movement parameter profiles. Modified after Dodge et al. (2009: 423).

Another useful way of representing time oriented information (such as seasonal mobility data) is to use cyclic representation (see Figure 8) where data is plotted on circular form where each angle of the circle indicates a specific time interval thus allowing to discover some cyclic patterns that would not be apparent with traditional linear plotting (x, y –diagrams) (Andrienko et al. 2010).

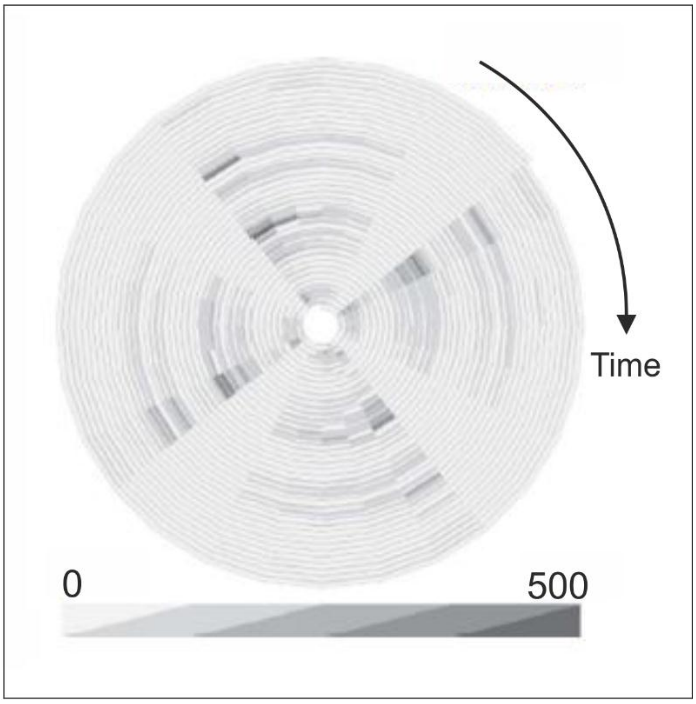


Figure 8. Example of cyclic spiral representation of health-related time series data. Modified after Andrienko et al. (2010: 1586).

2.4 From mobility analysis into wider contexts

> "Transport is one of the most powerful factors affecting and explaining the distribution of social and economic activity." - Knowles et al. (2008)

Modern information technologies and analytical methods (such as GKD) enable us better than ever to gather and harvest spatio-temporal information about movements, activities and presence of variety of objects such as humans, vehicles and animals. This is possible via various data sources such as mobile phones (Ahas et al. 2010), professional observation systems (AIS, ADS-B -Aviation monitoring etc.) (e.g. Demsar & Virrantaus 2010) and more recently also via social media services (Andrienko et al. 2013; Li et al. 2013). Accessing continuously such information can be utilized in various contexts such as in decision making processes and planning considering also different environmental aspects.

2.4.1 Mobility information in decision-making and planning

Many political planning decisions are often based on "rules-of-thumb‟ principles even though the need for accurate and more real-time information is higher than ever (e.g. Wilkins & desJardins 2001; Krizek et al. 2009) since our society is increasingly dynamic and mobile (YLE 2013b). Acknowledging the previous, a new movement called evidence-based practice (EBP) that emphasizes the use of scientifically analyzed information has emerged to support the decision making processes. EBP attempts to bridge the gap between traditional rules-of-thumb decision making and more formally analyzed information based on as accurate and recent data as possible. EBP was first introduced in the field of medicine but it has recently been utilized also in urban planning (Krizek et al. 2009) and regional land use decisions (Sutherland et al. 2007). Mobility data can be seen as one significant information source related to EBP and urban planning.

2.4.2 From realized mobility into accessibility patterns

The observed mobility of different objects (such as humans) is in principle describing the movement dynamics of moving objects on a certain period of time within the chosen spatial area (i.e. spatio-temporal domain) (Hägerstrand 1970). This gives exact information about the movements within the spatio-temporal domain but often such detailed data is not particularly usable when using it in other contexts such as in environmental models that often require more simple form of information. Thus making the movement data more simple by aggregating it with certain parameters and constraints is often more useful and informative than preserving all the details of the data.

One of the possible ways of utilizing mobility information is by using it as background information in accessibility analyses. Accessibility, i.e. the *potential of opportunities for interaction* (Hansen 1987) or *degree of connectivity* (Ingram 1971) that basically describes the potential movements, can be seen as a powerful concept of simplifying the realized movement characteristics into more generalized patterns describing e.g. the time distances between places within specific timeframes. Accessibility can be seen as an increasingly important analytical tool as means of understanding the human-environment interactions (Verburg 2011; Salonen et al. 2013).

2.4.3 Accessibility related phenomena

Accessibility, which is possibly derived from observed mobility, can be connected into wide area of environmental, economic and societal phenomena at different spatial scales ranging from local to global. Accessibility can be seen as useful indicator of the form and intensity of human-environment interactions since it can be strongly linked to phenomena such as land use change, human alteration of natural environments, deforestation, ecosystem modification, regional development and economic interaction (e.g. Ellis & Ramankutty 2008; Salonen et al. 2013; Vickerman 1995; Vickerman et al. 1999; Verburg et al. 2011).

On a local and regional scale, the structures of transportation and land use (services, work places and housing) influences citizen's need to use time and energy for travelling that also can be seen as important factors influencing human well-being. Local and regional scale traffic and transport have been found as major causes of environmental problems in urban regions (Bertolini and le Clercq 2003; International Energy Agency 2009) thus there is also growing concern over urban greenhouse gas (GHG) and carbon dioxide (CO2) emissions resulted from daily travel (Yang et al. 2009; Lahtinen et al. 2013).

In today's globalizing world, transnational, long distance travelling by different travel modes has become everyday activity for many people and most goods, which consumes enormous amounts of energy and natural resources. In addition, while climate and geology have shaped ecosystems and evolution in the past, there is growing evidence showing that human forces are in many areas surpassing the natural forces modifying our landscape (Turner et al. 1994; Rindfuss et al. 2004). Human-environmental interactions have far reaching consequences for the functioning of the Earth system on different spatial and temporal scales (Ellis & Ramankutty, 2008; Verburg et al., 2011) thus it is increasingly important to understand the global interaction of accessibility and land use and the implications they have on the environment and competitiveness of societies (e.g. Salonen et al. 2012b; Salonen et al. 2013).

2.4.4 Previous transportation studies in Peruvian Amazonia

Previous studies in Amazonia that have included transportation aspects as part of the study have mostly concentrated on the role of roads, and studies have mainly focused on Brazilian part of Amazonia (e.g. Bauch et al. 2007). There are however also few studies that have focused on Peruvian Amazonia and the role of rivers as main transportation network (Salonen et al. 2012a; Salonen et al. 2013).

Salonen et al. (2012a) have studied the accessibility patterns based on time distances on riverine transportation system of Peruvian Amazon. They discovered that the network distances are considerably higher (on average 1.6 times longer) than the Euclidian distances that have been typically used as measure of accessibility in previous studies (e.g. Peres & Terborgh 1995; Peres & Lake 2003). They noticed that the time distances depend on the size of the channel since larger channels are faster than narrow densely meandering channels to navigate. Seasonal changes in navigation patterns and greatly influence the everyday life of the inhabitants in Amazonian communities, and thus Salonen et al. (2012) concluded that it would be important to study the temporal aspects of Amazonian river transports. Also changes in the river networks have great impact on people's life since the vital river connection can migrate kilometers away thus hindering the product and food supply of villages (Coomes et al. 2009).

Salonen et al. (2013) have studied how different accessibility measures perform as input data for land use and land cover change (LUCC) models. They simulated deforestation in Peruvian Amazonia and concluded that time distance to market center in association with distance to transport network (i.e. rivers) had the most accurate simulation results according to deforestation. They also tested two complementary Euclidian measures that achieved nearly as accurate results compared to simple network based time distances. Salonen et al. (2013) thus recommended that LUCC modelers test the effect of different accessibility variables and their combinations, and paying attention to site-specific characteristics of the transportation networks and the type of phenomenon analyzed.

III. STUDY AREA

3.1 Loreto and Ucayali regions in Peru

The study area of this thesis covers the Loreto and Ucayali regions in Peru (see Figure 9). Peru is located on the west part of South America sharing borders with Brazil, Bolivia, Chile, Ecuador and Colombia. Loreto and Ucayali regions are located on the North Eastern part of Peru and the regions are bordered by mountain range of Andes on the west side from where the largest river of the world by discharge, Amazon, originates (Gupta 2007: 31). Loreto and Ucayali regions cover together the Peruvian part of Amazon rainforest that is locally referred as *selva baja* (lowland rainforest) that is defined to be all of the rainforest areas below the elevation of 600m above sea level (Kalliola et al. 1993: 7).

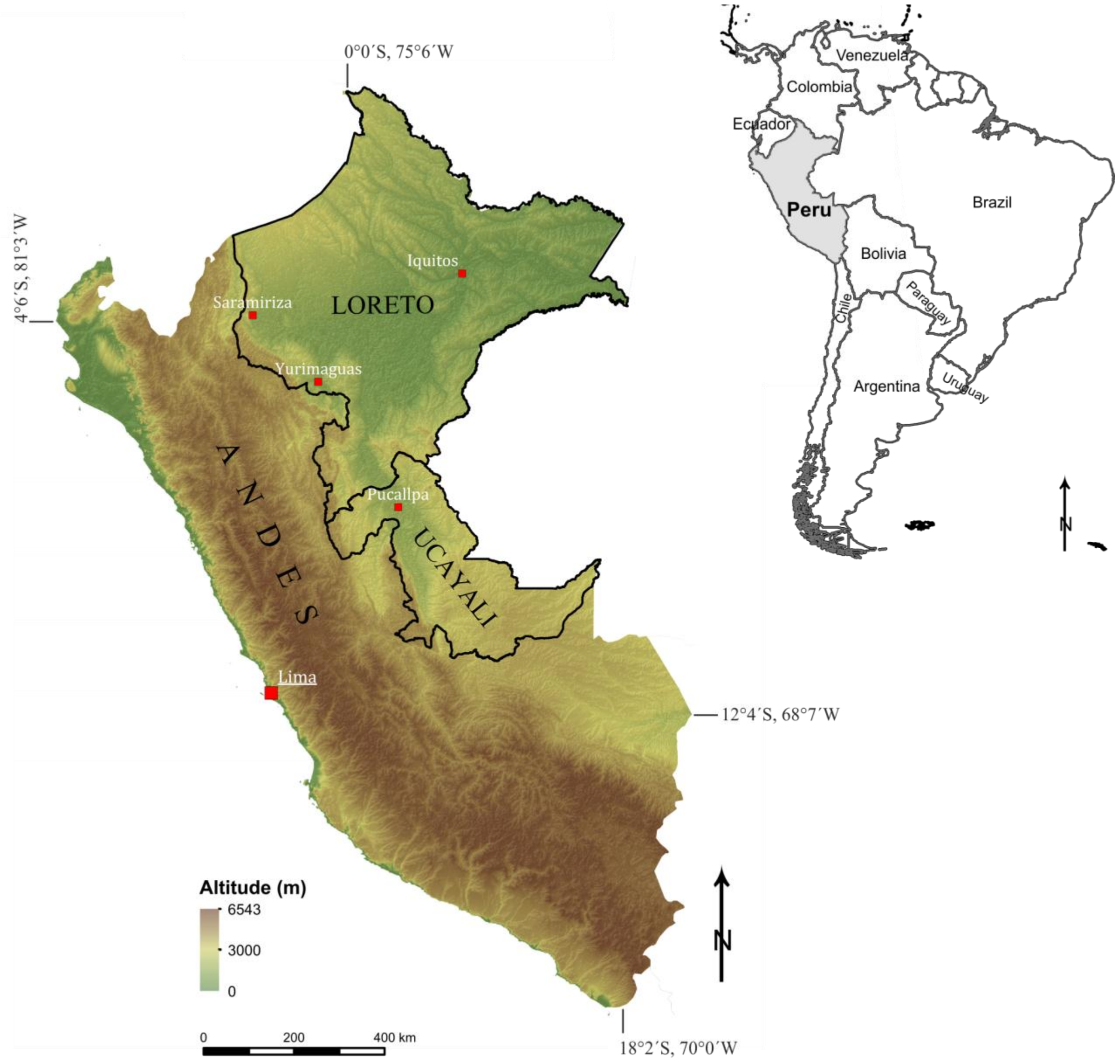


Figure 9. The study area of this thesis covers Loreto and Ucayali regions on the North Eastern Peru. Elevation map of Peru and its location in South-America.

Table 1 represents few basic demographic statistics of the study area that are mainly from the years 2007-2008 since the last census took place in 2007 and thus the numbers are most reliable from that period of time.

Loreto is the largest and Ucayali is the second largest region by area in Peru, and Loreto was the 11th and Ucayali 18th the largest region by population in 2007. Since Loreto and Ucayali regions are mainly covered by rainforest areas this results significantly lower population densities (2.5 / 4.2 inhabitants per km^2) than rest of the country.

Table 1. Statistics of Peru, study area regions and cities that are important harbors related to this study. (Sources: GOREL 2006c (1); INEI 2007 (2); INEI 2008 (3); MTC 2008 (4); INEI 2009 (5); BCRP 2009 (6); MPCP 2010 (7); BCRP 2012 (8); INEI 2012 (9); The World Bank 2013 (10))

	Peru	**Loreto**	**Ucayali**	**Iquitos**	**Pucallpa**	**Yurimaguas**
Administrative status	Country	Region	Region	Capital of Loreto	Capital of Ucayali	Capital of province
Population (3) (census 2007)	27 412 157	891 732	432 159	370 962	204 772	51 747 (8)
Population (9) (estimate 2012)	30 135 875	1 005 953	477 616	422 055	211 591	59 062
Area (km^2)	1 285 216 (2)	368 851 (1)	102 411 (8)	369 (1)	65 (7)	2 684 (1)
Population density - 2007 (inhabitants/km^2)	21.3	2.5	4.2	1005.3	3150.1	19.3
Urban / rural Population (%) (2)	75.9 / 24.1	65.4 / 34.6	75.3 / 24.7	-	-	-

3.2 Central Places of Loreto and Ucayali

In Figure 9 (page 22) there are marked the capital of Peru, Lima, on the west coast with a large red square and some of the important centers according to this study in Loreto and Ucayali with smaller red squares.

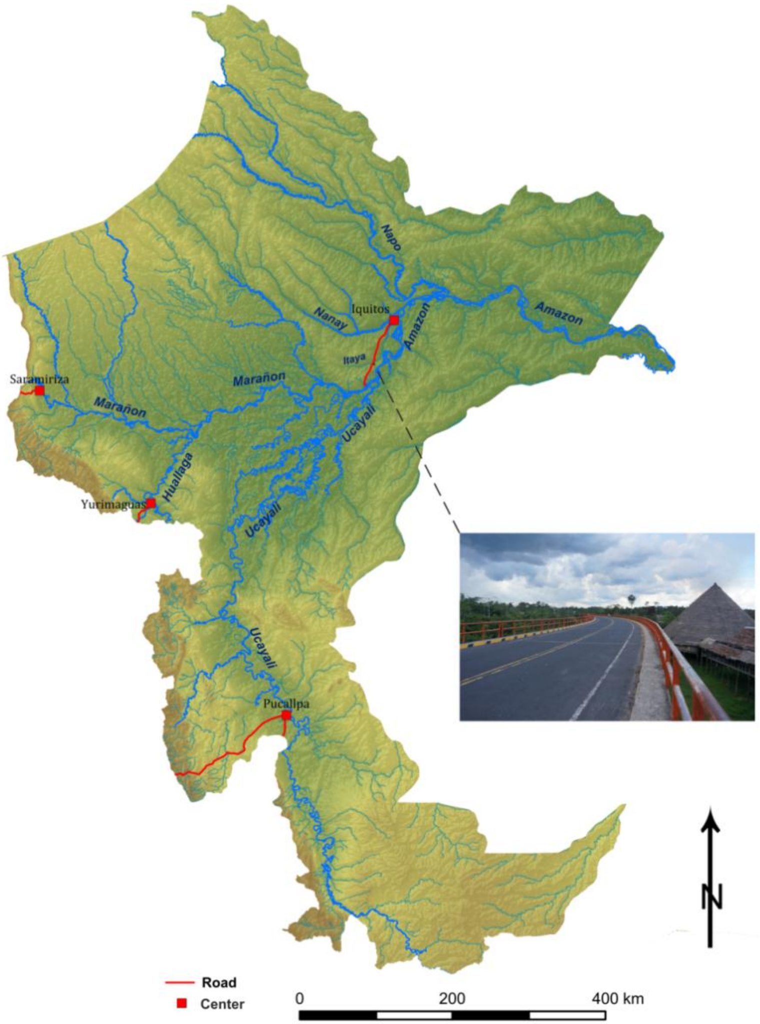
Iquitos is the capital of Loreto region situated approximately 1000 km North East from Lima (Euclidian distance). Iquitos was founded in 1757 with estimated population of 422 000 in 2012 that ranks it the 6th largest city in Peru (INEI 2012). Iquitos is commercially important center located at the confluence of Amazon, Nanay and Itaya rivers, and the city is approximately 100 km northwards from the confluence of Ucayali and Marañon rivers that originate from the Andes and form the prober of Amazon River. The river network is essential to local people and economic since the road connections 7from Iquitos are poor and the majority of the transportation is based on river navigation (see Figure 10) (Rodriguez Achung 1994). Thus Iquitos is known as being the world's largest city without road connections to the rest of the world.

Iquitos has a long history of being a busy transportation center since the rubber boom era in the 19th century when the rubber was exploited from the surrounding areas and then shipped to the world (Vílchez Vela 2012). Iquitos is still today the most important transportation hub of the study area according to river transportation (see Figure 11) and there is frequent boat traffic towards harbors of Peru (e.g. Pucallpa & Yurimaguas) and also to Manaus (in Brazil) which is the most populous city of the whole Amazon.

Pucallpa is the fastest growing Peruvian city in the Amazon (Abizaid 2005) and it is an administrative capital of Ucayali region. Related to Amazonian transportation it is also an important city because it has good road connections to the rest of the country. Pucallpa has large harbor along Ucayali River which is an important gateway to Iquitos (1120 km downstream) as means of transporting cargo and passengers. Thus there are regular daily ship departures between Iquitos and Pucallpa.

Yurimaguas is also an important harbor related to transportation towards Iquitos. Yurimaguas is a capital of Alto Amazonas province and it is located along the Huallaga River approximately 680 kilometers upstream from Iquitos. Yurimaguas is popular gateway among the passengers since it is the closest city to Iquitos with good road connections to the rest of the Peru. Thus a lot of passengers and tourists use the ferry connection from Yurimaguas if they do not wish to use airplane to reach Iquitos. Yurimaguas is also important center as means of transporting cargo to/from Iquitos.

The maps of this chapter (III Study area) represent also the smaller center of Saramiriza which is one of the destination harbors used by our collaborating ship companies navigating from Iquitos. This route (Iquitos-Saramiriza) is however not used in this thesis because of the route is mainly parallel with Iquitos-Yurimaguas route, and the journeys were irregularly tracked (with AROS).



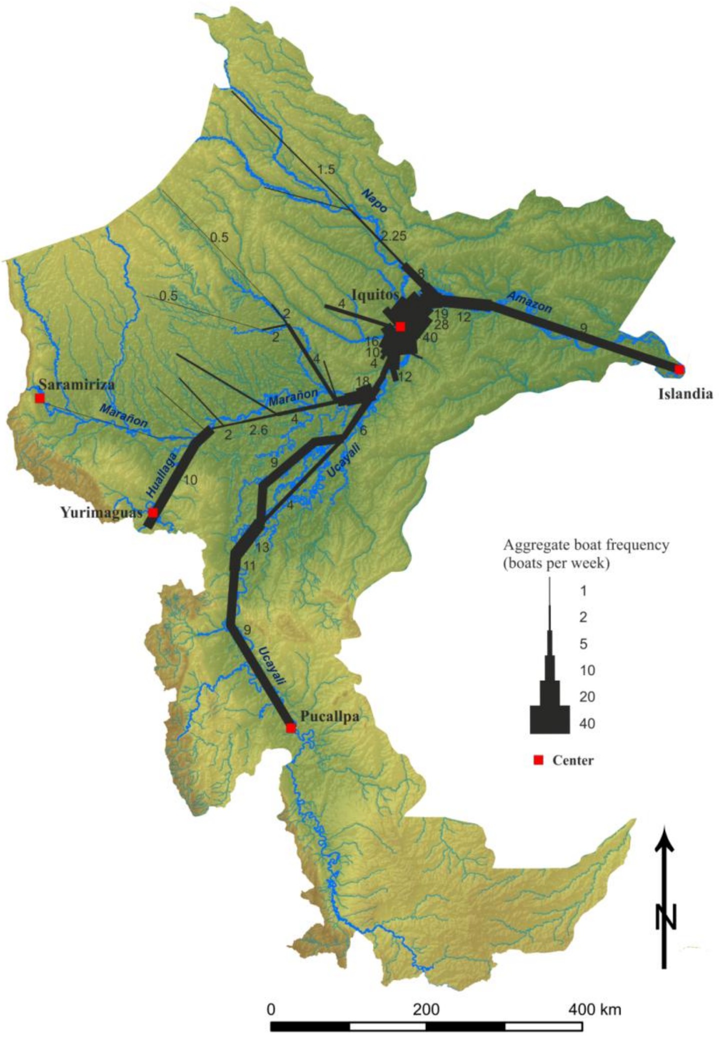


Figure 10. Map of the river network in the study area and locations of the important centers related to this study. Poor road connectivity of Iquitos and Amazonian lowlands towards other parts of Peru is clearly visible. Photograph represents the only paved road connection that runs from Iquitos to Nauta. Photograph © Tenkanen 2012.

Figure 11. Frequency of river boats per week around Iquitos in 2005. Modified after Salonen et al. (2012a).

3.3 Environmental characteristics of the study area

The environmental conditions of the study region are characterized by moist to very moist tropical lowland rainforests (Puhakka et al. 1992) with average annual temperatures varying between 20-28 ºC and annual precipitation ranging from 1600 mm to 3500 mm (Hoffman 1975, cit. Toivonen et al. 2007). The relief of the study area is gentle with minimal differences in elevation (see Figure 9 and Figure 10). For instance the approximate gradient of Ucayali River is only five centimeters per kilometer (Abizaid 2005) and the average slope for entire Amazonian lowlands (from Peru to the Atlantic Ocean) is only 10 centimeters per kilometer (Puhakka et al. 1992).

The most dominant environmental factors describing the Amazonian lowlands are water and forests. The Amazon is the world's largest river basin with area of over 7 million square kilometers making it the greatest river system of the world (Sioli 1984). The fluvial system of the area has strong and comprehensive influence over natural habitats such as plants and animals, as well as on humans that use rivers as a way of earning their living, gathering food and travelling between locations (Abizaid 2005; Coomes et al. 2009; Salonen et al. 2012a).

Rivers of the region are typically categorized in terms of their size, channel morphology and 'color' (white-water, black-water and clear-water) (Sioli 1984). Figure 12 represents the important rivers related to this thesis indicated with different colors: Amazon, Marañon, Ucayali and Huallaga. These rivers are the largest and also most important as means of riverine transportation routes. The channel morphology (or pattern) of the rivers in the study area varies from wide anastomosing rivers (Amazon and Marañon) with quite low sinuosity (rate of meandering) to narrower meandering rivers with higher sinuosity (Ucayali and Huallaga) (Puhakka et al. 1992). Wider anastomosing rivers are typically easier to navigate by large river launches (see Figure 15, page 30) than meandering rivers that are narrow and can have also plenty of sandbanks during low water hampering the navigation (albeit sandbanks can exist also along anastomosing rivers).

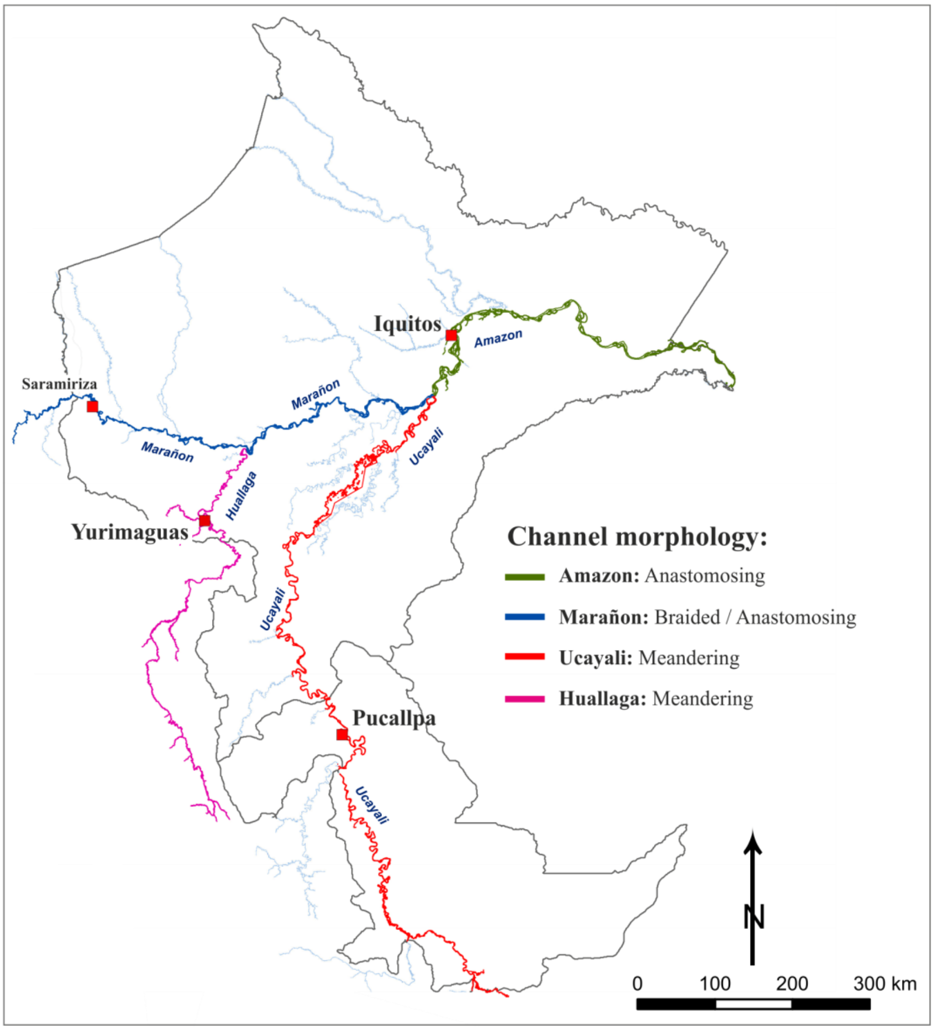


Figure 12. The map of rivers that are included in this study and their channel types according to Puhakka et al. (1992) and Toivonen et al. (2007).

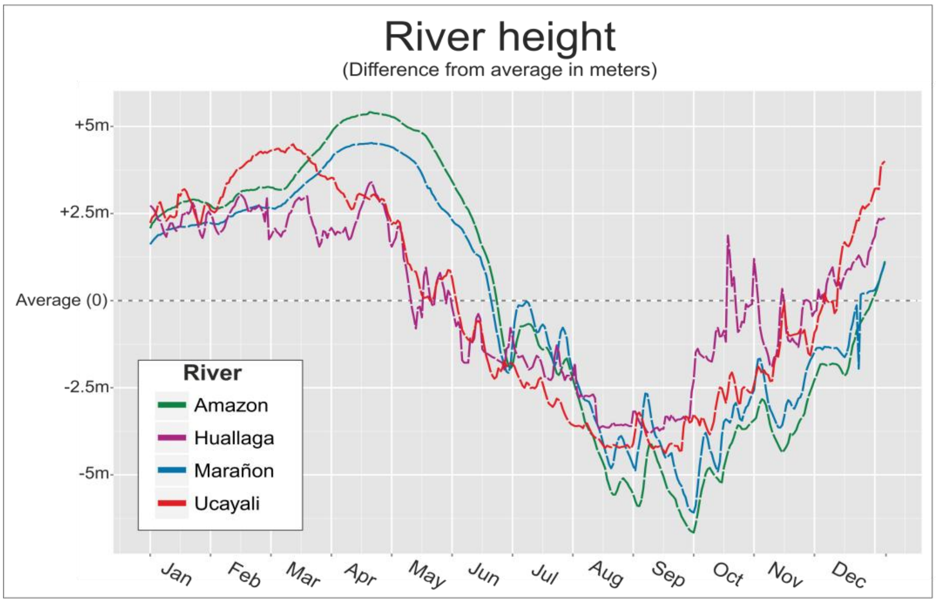


Figure 13. Daily river heights of the year 2012 represented as difference from year’s average river height. Data source: SEHINAV (2013).

The river dynamics in Amazonian lowlands are significant that can be clearly seen from Figure 13 which represents the daily river heights along the main rivers of this study during the year 2012. The difference between low water and highest peak during high water can be over 10 meters which demonstrates how massive the dynamism of these rivers can be. Thus the seasonal changes and flood cycles of the Amazonian river system is able to modify extensive land areas in river floodplains (Zeng 1999; Marengo 2005; Toivonen et al. 2007).

Figure 14 represents a map of network distances originating from Iquitos and illustrates how much longer they are compared to Euclidian distances. Sinuosity has significant effect on the travelling distances, thus the network distances at the study area are almost twice as long compared to as the crow fly distances (e.g. on Iquitos-Pucallpa route).

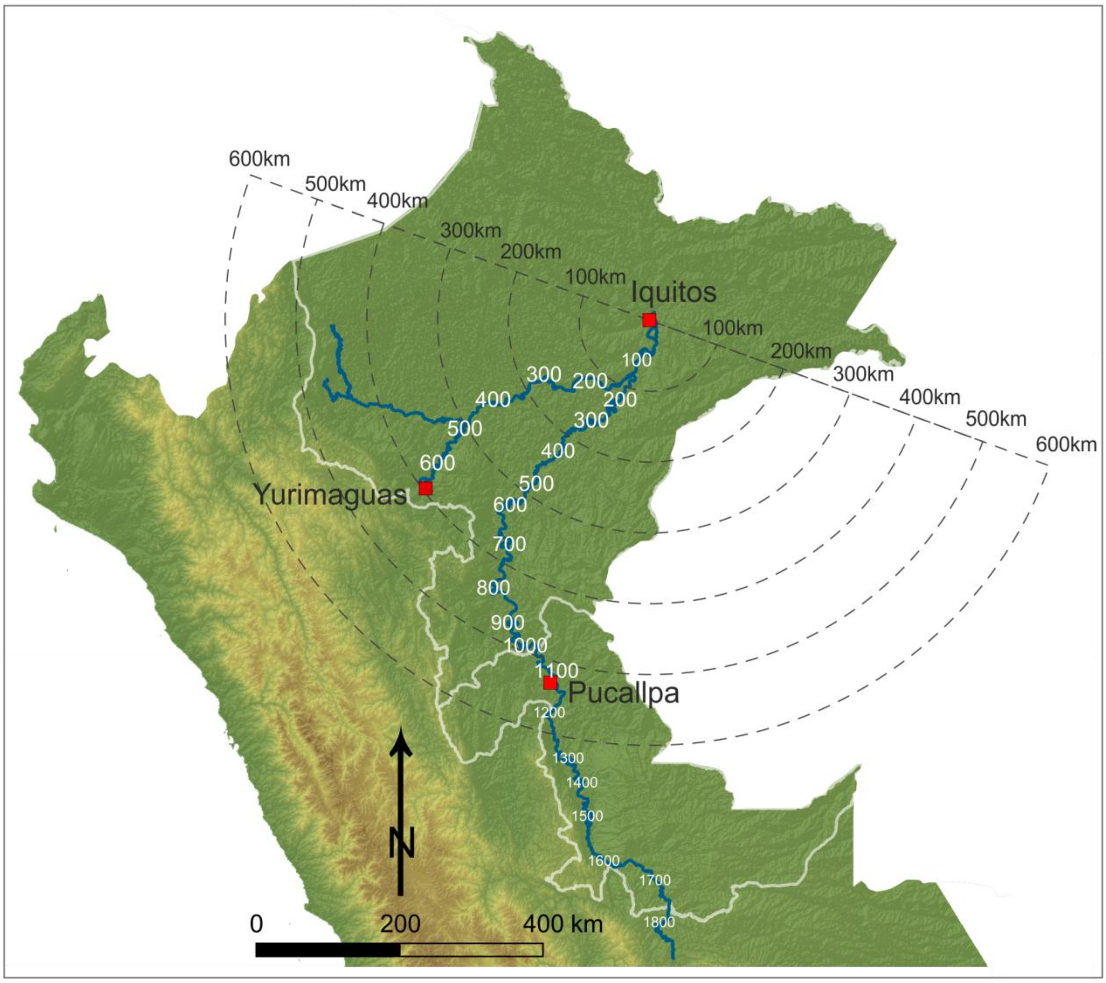


Figure 14. Network distances versus Euclidian distances originating from Iquitos along the river network at the study area.

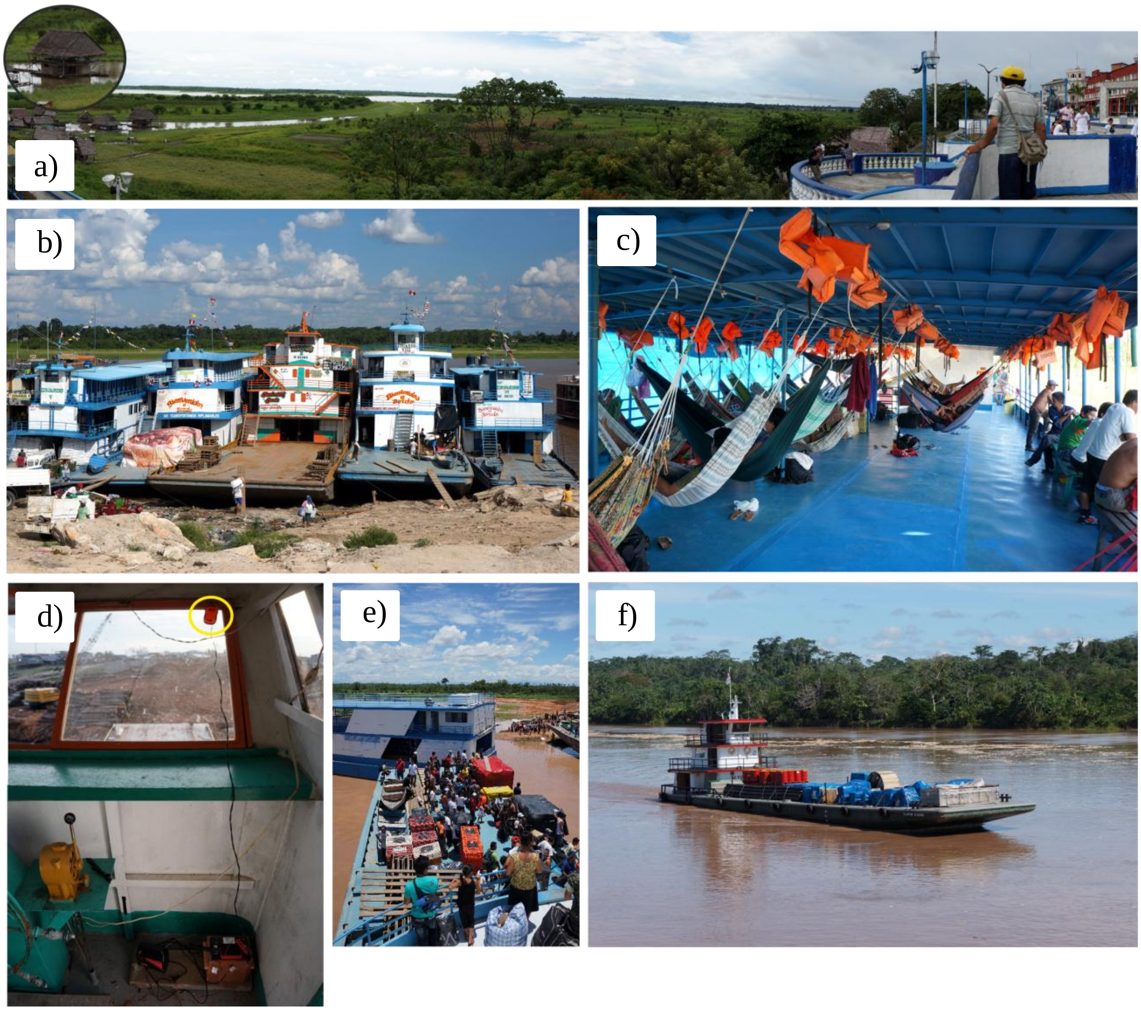


Figure 15. a) A panorama view from the boulevard of Iquitos towards the Amazon River during low water (seen in the horizon). During high water the river rises and fills all the green areas seen in this picture, b) Harbor of Masusa in Iquitos which is basically only a river bank where boats come ashore, c) Passengers usually spend their night by sleeping in hammocks on a passenger deck, d) A modified SPOT GPS Messenger with a battery and a charger in the cockpit of a river boat, e) Passengers ready to go ashore at the harbor of Yurimaguas, f) A small cargo ship at the Marañon River. Photographs © Tenkanen 2012.

IV. DATA

4.1 Data sources

The data of this study includes both primary and secondary data sources (see Table 2). The most important primary data source is the GPS-information of the entities (vessels) derived from the pilot observation system. Also the reference data that describes the river network (see III) has a vital part regarding to all of the analysis made in this study. Secondary data sources include datasets that are used for visualizations and river classification that is used to analyze the significance of river geometry to movement patterns.

Table 2. Data sources of the study.

	Data	Source	Data collection time
Primary data	Spatial/temporal information of vessels	GPS-data from AROS	01.01.2012-31.12.2012
	River geometry	GPS- data from AROS	01.08.2011-
	Validation data	GPS-measurements on the field	08.10.2012-26.10.2012 /
Secondary data and spatial datasets	Water level statistics	SEHINAV (2013)	01.01.2012-31.12.2012
	River classification	Puhakka et al. (2009); GOREL (2006a); Toivonen et al. (2007)	
	River network	GPS-measurements on the field, IIAP/Biodamaz (Biodamaz 2004a, 2004b, Josse et al. 2007)	01.01.2012-31.12.2012
	Roads	GOREL (2006b)	
	Administrative borders	GOREL (2006c)	
	Populated places	GOREL (2006c)	

The dataset utilized in this thesis consists of 11 572 observations that were collected with 5 different devices during the year 2012. Altogether 14 332 observations were collected with AROS (7 devices) but two of these devices are not used for this study because the data coverage of those devices were low. These two vessels (with low data coverage) are travelling from Iquitos to Saramiriza and thus, for the most of the time, they share the same navigation path with two vessels that travels from Iquitos to Yurimaguas. Parallel navigation paths of these two routes means that the lost information from Iquitos-Saramiriza route is not significant according to my research questions.

The information about daily water level at the study area is provided by Servicio de Hidrografía y Navegación de la Amazonia (SEHINAV 2013). The water level is measured from six different measurement points along the main Amazonian rivers but in this study only four of them are used. The used measurement points are: Río Amazonas in Iquitos, Río Marañon in Nauta, Río Ucayali in Pucallpa and Río Huallaga in Yurimaguas.

The secondary data sources include the spatial datasets provided by the Regional Government of Loreto that include administrative borders and roads. River classification is based on the study made by Toivonen et al. (2007) where they studied the fluvial biogeography of the Peruvian Amazon with quantitative spatial analysis.

4.2 Data collection system AROS

The primary data used in this study is based on a pilot data collection system called AROS (Amazonian Riverboat Observation System) that utilizes satellite messenger system to obtain the location information of the vessels. The vessels that collaborate with this study are large launches (*lanchas)* (see Figure 15) that operate long distance journeys and have capacity for approximately 150-300 passengers (Tenkanen 2012) and 35-600 tonnes of cargo (Salonen et al. 2012a).

The developed data collection and monitoring system resembles the automatic identification system (AIS) that is used in maritime vessel tracking and it was developed because in the Peruvian Amazon this kind of expensive tracking system is not used. These low cost satellite messengers (approx. 100€ each) enable to determine the location of the device by GPS satellite system and to send its location information via communication satellites to a database. The tracking of the vessels is based on the devices provided by SPOT LLC$^{©}$ (2013) that are originally developed for outdoor activities (hiking etc.) but related to this study these devices are applied to continuously track the moving vessels along the rivers of Peruvian Amazon.

In the data collection system (see Figure 16) altogether seven satellite messengers are utilized on different boats that send their location information every 10 minutes (at best) to the service provider’s database (Spot). From the service provider’s database the data is collected automatically to MySQL database that is located on the server at the University of Helsinki. The purpose of MySQL database is to preserve the data since

SPOT does not provide long-term storage of data for the users and the location information is always cleared from the service provider's database after seven days. From the MySQL database the data is further obtained for the knowledge discovery process (see part V).

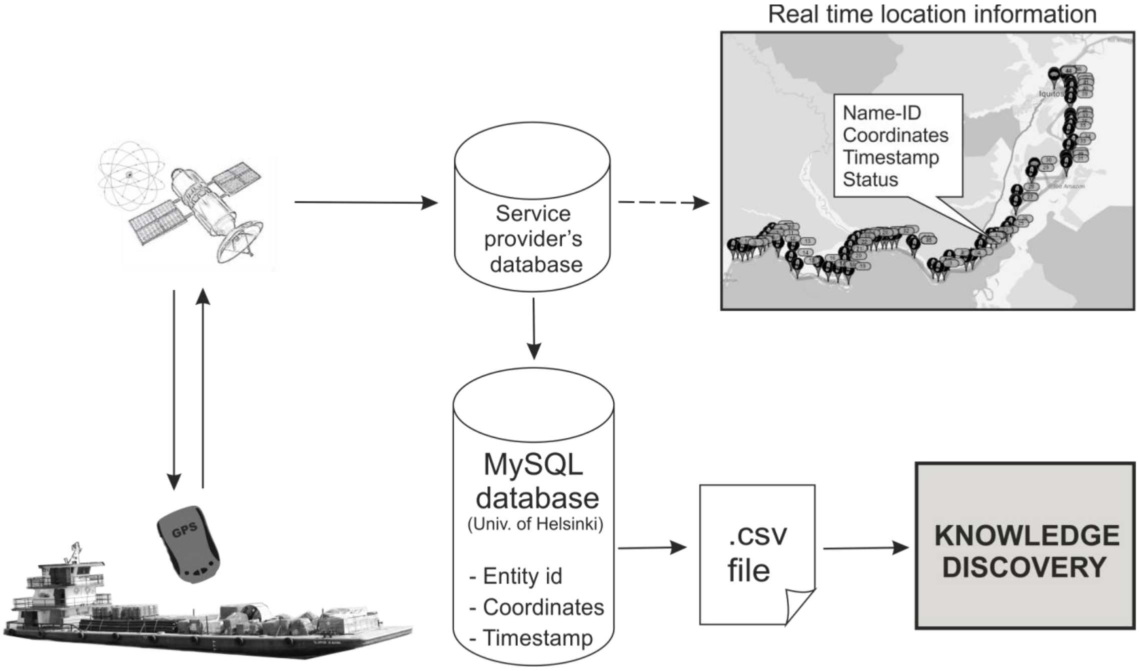


Figure 16. Concept of the pilot data collection system that utilizes GPS satellite messenger that is able to both receive and send location information.

The satellite messengers work normally with 2 x 1.5V lithium batteries that provides the battery life for 14 days but since the purpose of this tracking system is to continuously track the vessels for longer periods, these devices were modified in a way that they can utilize 12V battery that can be charged during the nights when there is electricity available on the boats (see Figure 17).

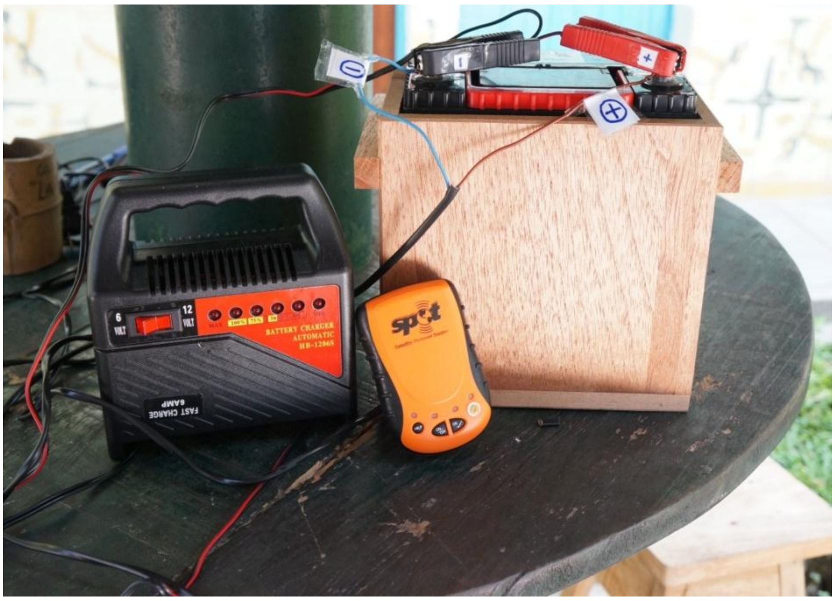


Figure 17. In the pilot project modified SPOT Satellite Messengers are used to deliver location information. Photograph © Tenkanen 2012.

These modifications minimize the maintenance work of the devices since there is no need to change the batteries every two weeks and prevent also possible interruptions of tracking that are caused by exhausted batteries.

4.3 Data structure

The GPS observations derived from the SPOT satellite system are stored in a local database at the University of Helsinki and they are retrieved via a XML-feed service provided by SPOT. There were altogether nine different vessels collecting the data during the year of 2012. However there were only five collaborating vessels that are taken into account in this study until the end of September 2012 because of the poor data quality and coverage by other GPS-devices. Since October 2012 there have been altogether seven devices collecting the data and from October to the end of year 2012 all of the observations were taken into account in this study.

The structure of the GPS-data is fairly simple (see Table 3) and it consists of entity ID, latitude and longitude coordinates, timestamp, Unix time and message type (test/tracking). Unix time represents the time (seconds) elapsed since the midnight Coordinated Universal Time (UTC) 1.1.1970 which is currently the primary time standard (ITU 2002). Unlike many GPS-devices SPOT system does not provide any supplementary information about speed, direction, altitude etc. and therefore these information is derived from the data itself and from other data sources by calculating (see 5.7).

Table 3. Example of the data structure of GPS -waypoint.

id	lat	lon	timestamp	Unix time	messagetype
Boat1	-6.08115	-75.0966	2012-01-02T17:20:28.000Z	1325542828	TRACK
Boat1	-6.07864	-75.2855	2012-01-02T17:36:12.000Z	1325543351	TRACK
Boat1	-6.07722	-75.2935	2012-01-02T17:48:27.000Z	1325545273	TRACK
Boat2	-7.22510	-73.2645	2012-01-06T00:44:38.000Z	1325552334	TRACK
Boat2	-7.22644	-73.2519	2012-01-06T00:56:11.000Z	1325552511	TRACK
…	…	…	…	…	…

4.4 Reference / training dataset

The data mining processes of TRAT are based on the characteristics of a movement itself and the values of reference data. Reference data consist of digitalized river routes along Amazonian rivers which have river characteristics as supplementary data. Reference dataset was digitized based on observed GPS-points of the vessels which provide better locational accuracy of the actual navigation routes than relying on satellite pictures or maps of the area.

Routes consist of reference points that are separated from each other by the distance of 1 kilometer along the river. The distance between reference points was chosen because the constant unit of 1 (km) is easy to comprehend and apply to different analyses and it is adequate enough to separate the observations from each other. Each point has supplementary data that describes the point itself (id-number), position value of selected point related to Iquitos (distance in kilometers) and values that characterize the river at the selected location (see Figure 18).

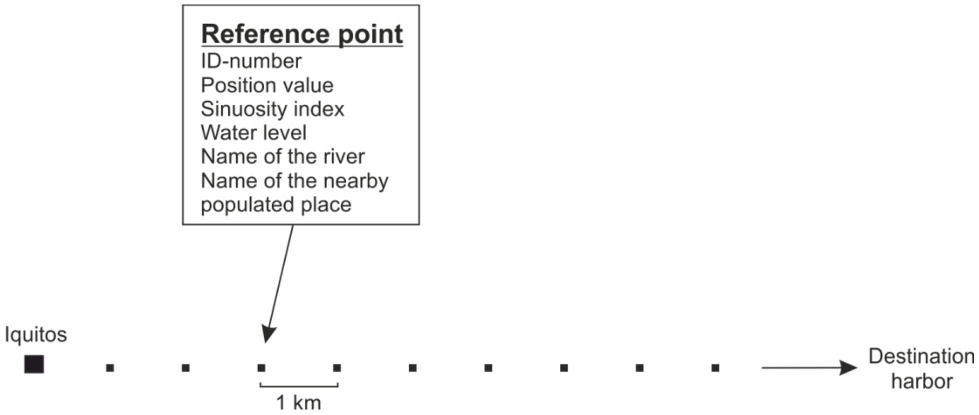


Figure 18. Reference point dataset.

4.5 Validation data

For assessing the quality of AROS as a data source and TRAT as an analysis tool, field measurements were made at the study area during the autumn of 2012. These field measurements were conducted by travelling the routes from Iquitos to Yurimaguas (11.10. – 14.10.2012) and Iquitos to Pucallpa (20.10. – 24.10.2012) with the same vessels that collaborates with our research group and collecting data simultaneously with AROS (SPOT satellite messengers) and Garmin GPS-device (Garmin GPSmap62) with high sample rate (1 obs./min). GPS-device had spatial accuracy less than 10 meters constantly.

V. METHODS

5.1 Softwares

This study utilizes Open Source software called Muste (Sund 2011) which is a package built for R. Muste is having it roots on a Finish statistical software called SURVO$^{\circledR}$ (Mustonen 1992) and while it is basically a package of R, it is also an independent software with its own user interface and many functionalities that R does not offer. Muste together with few R functionalities was used as a platform for automatizing the data management and data mining process. Representations of the data with different visualizations were mainly conducted with R. 2D maps were mainly produced with ArcGIS$^{\circledR}$ (ESRI Inc. 2010) and CorelDRAW Graphics Suite X5$^{\circledR}$ while the 3D maps representing the trajectories were produced with open source GIS-software uDig$^{\circledR}$ (uDig 2011) with STC-plugin developed by ITC (University of Twente).

5.2 Trajectory reconstruction and analysis tool

Trajectory Reconstruction and Analysis Tool (TRAT) is a specific tool designed to manage and analyze automatically the GPS information obtained from the SPOT GPS-tracking system (AROS). The reason for developing this kind of data management and analysis tool is purely practical since the amount of data is large and growing continuously, thus processing the voluminous data manually would be extremely time consuming and vulnerable for errors. This study addresses the principles of Open Knowledge Foundation (2013), thus the whole code for TRAT will be later publicly available via Accessibility Research Group (2013) website.

The main features of TRAT:

- Options for selecting, sorting and grouping the data based on:
    - Time of interest (time interval / weekday)
    - Boat / route of interest
    - Place / river of interest
- Identification of an individual journey
- Calculation of travel speeds of the vessels
    - Travel speed for each segment
    - Average travel speed for each journey
- Travel time calculations
    - Total travel time from harbor to harbor
    - Travel time from selected place to destination of the journey
- Data quality calculations
    - Data coverage – percentage of tracked distance from the whole journey
    - Observation density

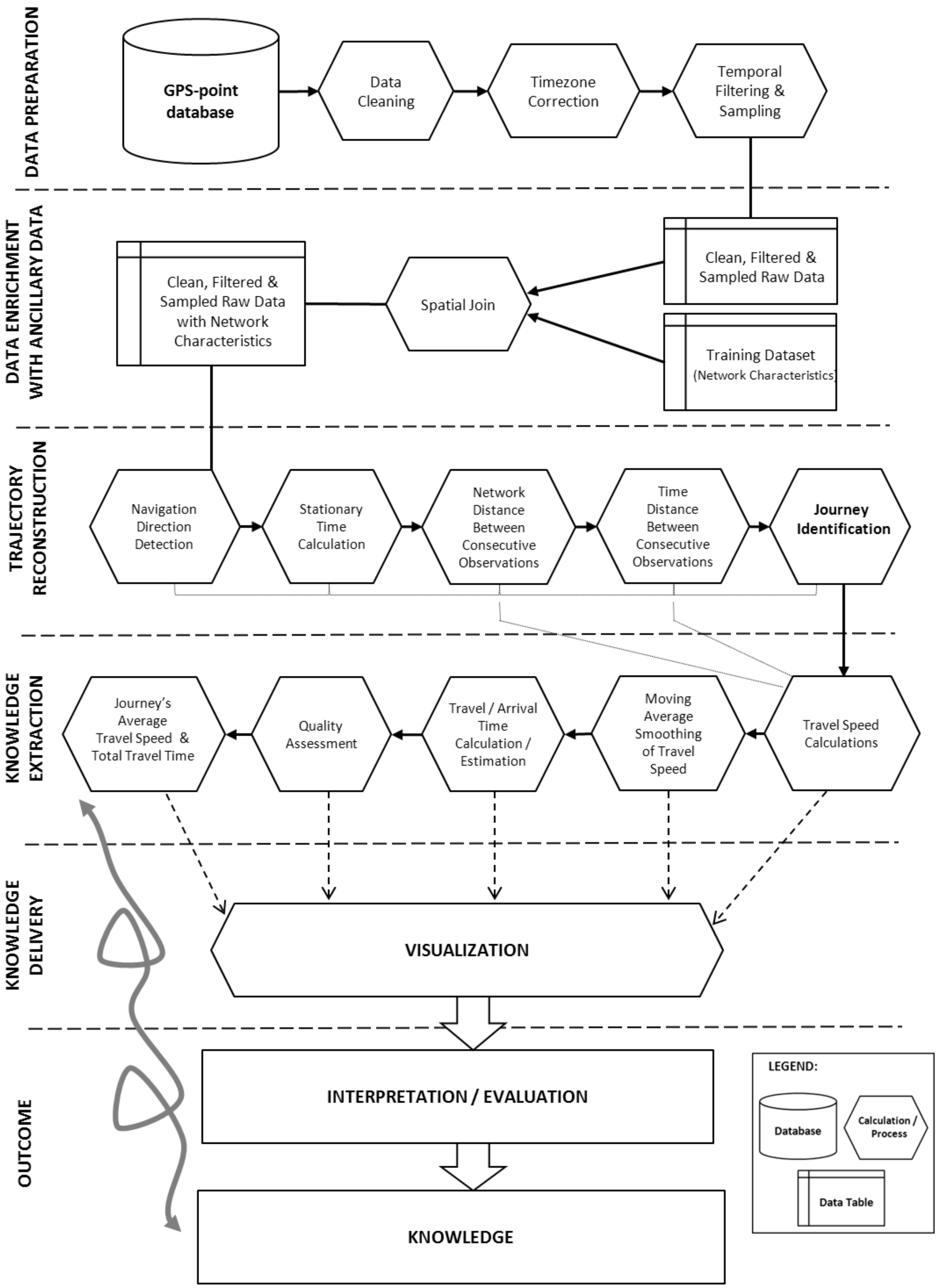

DATA PREPARATION
GPS-point database
Data Cleaning
Timezone Correction
Temporal Filtering & Sampling
DATA ENRICHMENT WITH ANCILLARY DATA
Clean, Filtered & Sampled Raw Data with Network Characteristics
Spatial Join
Clean, Filtered & Sampled Raw Data
Training Dataset (Network Characteristics)
TRAJECTORY RECONSTRUCTION
Navigation Direction Detection
Stationary Time Calculation
Network Distance Between Consecutive Observations
Time Distance Between Consecutive Observations
Journey Identification
KNOWLEDGE EXTRACTION
Journey's Average Travel Speed & Total Travel Time
Quality Assessment
Travel / Arrival Time Calculation / Estimation
Moving Average Smoothing of Travel Speed
Travel Speed Calculations
KNOWLEDGE DELIVERY
VISUALIZATION
OUTCOME
INTERPRETATION / EVALUATION
KNOWLEDGE
LEGEND:
Database
Calculation / Process
Data Table

Figure 19. Workflow of the developed tool (TRAT). Spiral line at the bottom illustrates the iterative nature of GKD process.

Figure 19 represents all analytical processes of TRAT as a flowchart. Work phases are divided into separate sections which illustrates tool's relation to geographic knowledge discovery and mobility data mining.

5.3 Data preparation

Before any calculations or analysis the data needs to be prepared which is the first task of the TRAT. In this process the algorithm cleans the data from observations that have any null values and reshapes the timestamp into the standard combined date and time form in UTC (2012-01-30T15:00:00.000Z) (ISO 8601 2004).

The time information of the GPS data is represented in Greenwich Mean Time (GMT 0) and therefore this needs to be converted into the Peruvian time zone (GMT -5). This is done by utilizing the as.POSIX* functions in R that are used for calendar date and time representations and manipulations (R documentation 2013).

After these procedures the data can be filtered and sampled by options listed in chapter 5.2 and thus is ready for further analysis.

5.4 Data enrichment with ancillary data

The analytical processes used in TRAT are primarily based on reference dataset (see 4.4) that includes information of the transportation network and different ancillary information that are collected from different data sources and then added as attribute data to the reference points. Enriching the AROS dataset with the information of reference dataset is done by joining them spatially together (see Figure 20 and Appendix I) enables further processing and analyzing of the data and is therefore crucial part of the developed analysis tool.

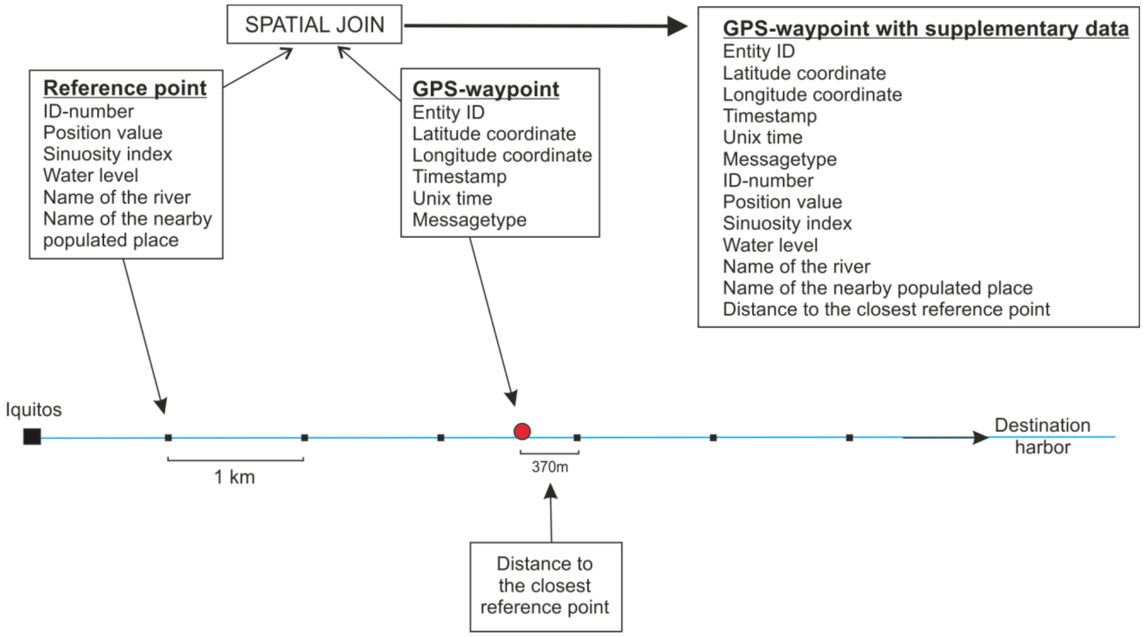


Figure 20. The concept of spatial join between a GPS-waypoint and the closest reference point.

5.5 Direction identification

The direction of movement is based on the network characteristics of the reference dataset (position value). Since every reference point has a unique id-number ascending from Iquitos towards the destinations, it is possible to determine the direction of the observation pair. Observation pair consists of two consecutive observations that have information about location and time. Because of the nature of river network, it is especially easy to determine the movement direction since the navigation direction that this study is interested in can be either up- or downstream.

Determining the main direction of a trajectory is more complicated since the direction of movement is not necessarily consistent during the whole trajectory. The main direction of a whole journey is based on the dominant direction of a trajectory where it is possible to have a few arcs that goes to opposite direction without the algorithm to change the main direction of a trajectory (scanning is based on the dominant direction of 20 observations).

5.6 Classification of individual journey – Trajectory reconstruction

Analysis of spatio-temporal movement patterns of the vessels requires accurate identification of individual journeys of the vessels. In this study a single journey (or a trajectory) indicates a trip that runs between the harbor of Iquitos and one of the destination harbors of Pucallpa and Yurimaguas. The journey can go either upstream or downstream depending on the destination.

For the identification of a single journey we can make two assumptions according how the device is tracking:

1. The device is tracking continuously without pauses in the data.
2. The device is tracking only when the vessel is moving.

Depending on the tracking mode the identification algorithm of a journey is different. Depending on the tracking mode it is necessary to have different "triggers" that enables the identification process (see Appendix II for the algorithm). These triggers that enable to delineate a journey include:

- Comparison of the direction of movement
- Observation of stationary time (in the harbors)
- Observation of tracking pauses in the data

- Temporal gap between observations gap_{time}
- Spatial gap between observations gap_{space}
- Tolerance distance D_{to}

1) When the device is tracking continuously the identification process is based on the first two triggers. Based on the nature of river network transportation we can assume that when the vessel has reached its final destination it will presumably change its course to opposite direction which can therefore be used for separating a journey from another. The observation of the stationary time is another way to separate a journey because we can assume that when the vessel is certain amount of time at the same location we can assume that the vessel is at the harbor and that indicates the end of that individual journey.

2) When the device is tracking only when the vessel is moving it is possible to find these untracked gaps from the data based on the time information. We can assume that when the GPS-device is not tracking for a certain amount of time it means that the vessel is at the harbor and that indicates the end of that individual journey.

Achieving the most reliable algorithm for journey identification these two different methods are joined in the same algorithm and thus the result is not dependent of the tracking mode of the device. This makes the identification algorithm suitable also for other applications that relate to network based GPS location information.

To be able to correctly identify an individual journey it is necessary to calibrate the algorithm based on movement patterns of the vessels. For calibrating the algorithm parameters it is necessary to decide what is the 1) stationary time limit and 2) untracking time limit for the algorithm to separate an individual journey from the data mass. To find the best parameters for the algorithms it is good to have some preliminary information about the movement patterns of the vessels or at least do some preliminary analysis of the data to get a picture of the transportation patterns. In this study the chosen time limit is 36 hours which proved to be optimal time for effectively detect individual journey from another after testing.

The journey identification concept is demonstrated in Figure 21 with space time cubes where the lines represents spatially and temporally connected observations. In the start situation (left cube) all of the observations are connected together with straight lines and they form unorganized set of spatio-temporal lines where there is no information about when or where the individual journey ends or starts. Represented algorithm identifies and indicates the individual journeys with so called *JourneyID*, and as a result the space-time paths are classified into separate trajectories represented with different colors in the end situation (right cube). Here the temporal and spatial links are applied only to those observations which are identified to belong on the same journey.

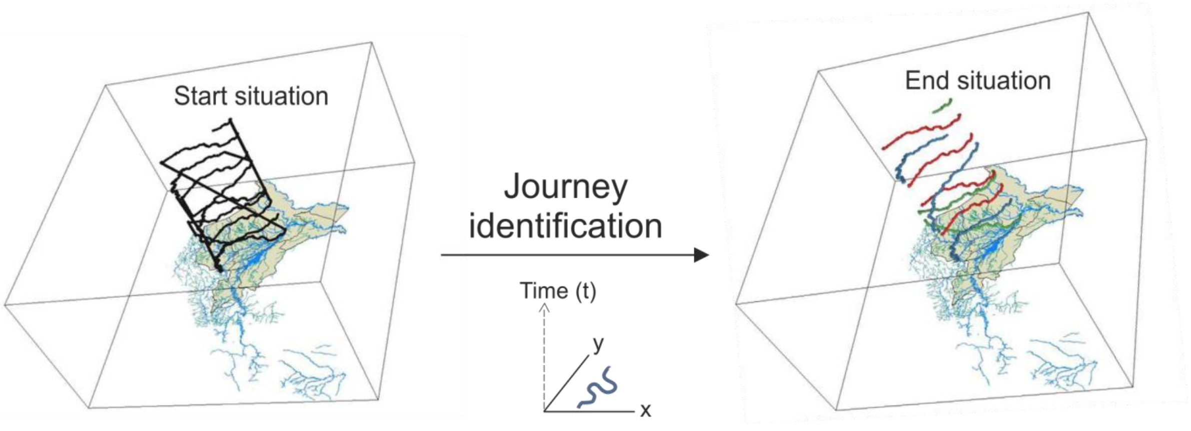


Figure 21. Space time cubes representing the journey identification process. Left space time cube represents the start situation of the data where all the observations are linked together both spatially and temporally. Right space time cube represents the results of the journey identification where individual journeys are identified as separate trajectories.

5.7 Travel speed calculation

The speed of an object is traditionally defined in physics as the magnitude of its velocity. In other words it is the rate of change of the objects position (Wilson 1901). An average speed of an object is defined as the total distance covered over the given amount of time.

In this study the travel speed is calculated for each segment that consists of two points including information about temporal and spatial distances between those observations. Temporal information is derived directly from the AROS data as timestamp and Unix time. Network distance between observations is derived by obtaining the route positions of the two consecutive points along the navigation path by spatially joining them with reference dataset (see 5.4), and then calculating the (network) distance between observations (*abs(observation$_a$-observation$_b$)*). Average travel speed is calculated based on evolving journey characteristics, i.e. cumulative travel distance and time:

$$\text{Segment} = \underbrace{\overset{\text{point}}{a} \ldots\ldots\ldots\ldots\ldots \overset{\text{point}}{b}}_{\text{Segment}}$$

$$\text{Travel Speed} = \frac{Distance}{Time} \in Segment$$

$$\text{Average Speed} = \frac{Cumulative\ distance}{Cumulative\ time}$$

Normally the average travel speed of a full journey is calculated by dividing the covered total distance by the total amount of time used for a journey:

$$\text{Average travel speed of a journey} = \frac{Total\ distance}{Total\ time} \in JourneyID\ k_{0,1\ldots n}$$

This formula however does not characterize the movements of the vessel particularly accurately since the elapsed time includes also the time when the vessel has not been moving. This formula describes more the average evolution of a journey instead of the characteristics of movement itself. Therefore it is necessary to cut out the stationary time when the vessel is not moving of the elapsed time:

$$\begin{array}{c}\text{Average travel speed of}\\ \text{a journey (movement)}\end{array} = \frac{Total\ distance}{Total\ time - Stationary\ Time} \in JourneyID\ k_{0,1\ldots n}$$

The average travel speed is calculated for each observation belonging to a same journey and therefore it is possible to gain information not only about the average travel speed of a whole journey but also how the travel speed has varied during the journey.

5.8 Total travel time calculation

Even though the average speed might be more accurate way to study how the seasonal dynamics (water level variation) affect the transportation in the Amazonian area, it could also be interesting and sensible to study how the travel times varies in different times of the year between origin and destination harbors.

Total travel time is simply calculated by subtracting the time at the origin harbor (t_0) from the time at the destination harbor (t_{end}):

$$\text{Total travel time} = t_{end} - t_0$$

5.9 Time distance calculation

In TRAT two different approaches are applied for time distance calculations. The first approach is to calculate the time distance based on the information of arrival time at the destination harbor and then compare this to the time at the selected point. The calculation formula is therefore similar to total travel time calculation but the start location (t_0) varies according to the selected location (t_{pos}).

$$\text{Time distance to harbor}_x = t_{end} - t_{pos} \in JourneyID$$

This type of approach gives an accurate result of the time distance from selected points to destinations but it requires that the data is already obtained throughout the whole journey. Thus this approach cannot be used for modeling the estimate arrival time to the harbor while the journey is still in progress.

Therefore another approach is used for estimating the time distance to the harbors when only incomplete information about the journey is available. The time distance from selected location to the harbors is calculated by dividing the distance to the harbor by average travel speed that has been evolved since the beginning of the current journey.

$$\text{Time distance estimate to harbor}_x = \frac{Distance\ to\ harbor}{Cumulative\ Average\ speed} \in JourneyID$$

The time distances are calculated only to the harbors that belongs to the route of the selected vessel. For example, the time distances will be calculated only to the harbors along Ucayali when a vessel is travelling from Iquitos to Pucallpa. The time distance is calculated to both directions so that the time distance towards the origin harbor represents basically the same value as total travel time at selected location. However the time distance towards destination harbor represents an estimate of the remaining travel time to the destination based on the travel speed information obtained from the passed journey. Therefore the accuracy of the time distance estimate towards the destination improves as the distance to the destination decreases and is at lowest when the journey begins.

5.10 Sinuosity index calculation

Sinuosity index is one of the basic measures describing the river geometry. Sinuosity index describes basically how much river meanders along its path and is therefore a good measure to get an idea of the river's nature, if it is an anastomosing river or a meandering river. Sinuosity index is calculated by dividing the network path length by Euclidian distance (see Figure 22).

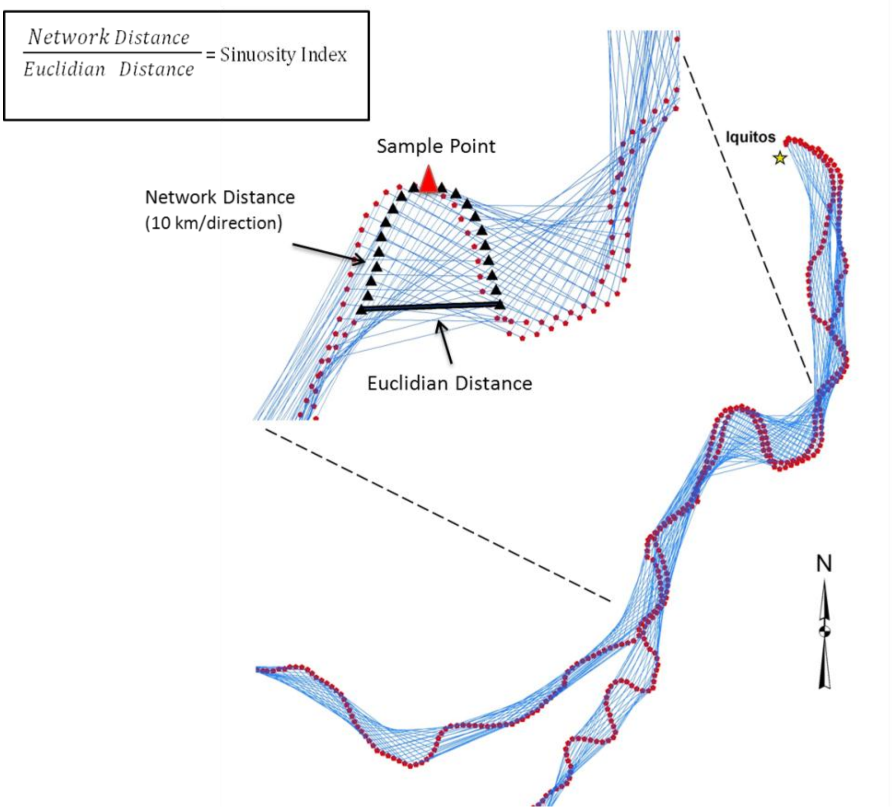


Figure 22. The concept of sinuosity index calculations.

The intensity of the sinuosity index depends on the Euclidian distance. The shorter the Euclidian distance the more local are the values of the sinuosity index. Therefore the appropriate Euclidian distance depends on the river and the purpose of the study or how localized measures you want. In this study the Euclidian distances of 10 kilometers and 5 kilometers was tested and based on these tests 10 km was chosen which proved to be local enough to get variation to the results along the river but not too short to loose information about the river geometry.

5.11 Data smoothing and filtering

Movement data and characteristics of any type of MPO (such as vehicle, animal or human) is typically highly deviated and extremely “noisy” thus making it difficult to reveal any clear patterns from the “raw” movement characteristics. For gaining better understanding of the movement data it is useful and often necessary to filter and smooth the data with specific methods (see Figure 23). In order to smooth raw GPS data several methods can be employed, such as least squares, spline approximation, moving average, kernel-based smoothing, and Kalman filtering (Dodge et al. 2009).

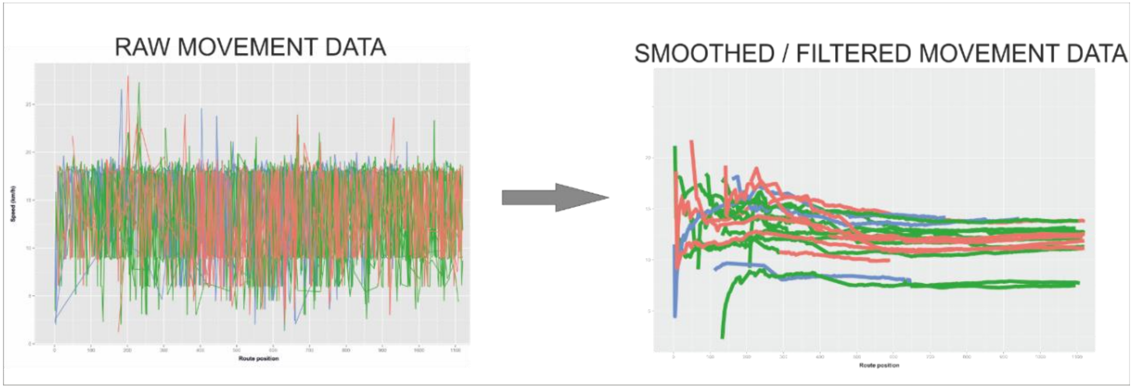


Figure 23. Concept of smoothing the movement data with intensive smoothing.

TRAT utilizes a specific moving average smoothing technique (see Figure 24) to harmonize the data in a way that it becomes more comprehensible. Moving average smoothing can be done with different window sizes that defines how many observations (or more commonly time intervals) are taken into account when calculating the average value. In TRAT the window sizes of 3 and 9 observations are used since they proved to work well for smoothing the data without significant loss of movement characteristics.

Smoothing that is based on consecutive observations rather than time intervals was used since the temporal observation density of AROS is quite irregular.

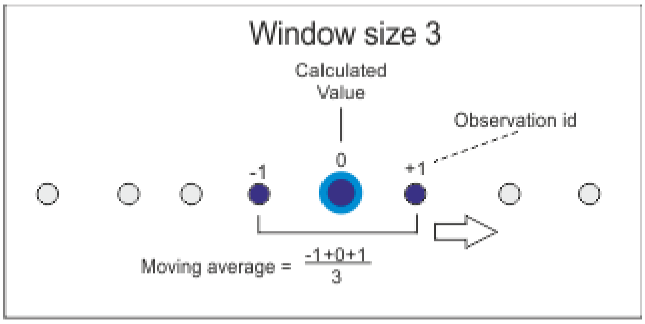


Figure 24. Moving average smoothing based on neighboring observations.

5.12 Assessment of AROS data and TRAT

The data used in this study has errors caused by different factors (see 7.4 for details). Therefore different data quality indices are calculated for enabling to evaluate the significance of the errors when representing the results.

The accuracy of travel speed calculations depends on the density of the observations. If the data is sparsely distributed along the route, a lot of information is lost between the observed locations and the average travel speed calculations become less representative. For example if the time interval between two observations is several hours the probability that there has been an unrecorded stop gets higher. This causes that the average speed calculated for this segment does not represent the actual movement since also the time that passed during the stop is taken into account.

To evaluate the quality of travel speed calculations an index that describes the observation density of a journey is calculated based on the distances between the observations:

$$\textit{Observation density} = \textit{Mean}(\textstyle\sum \textit{Position}_a - \textit{Position}_b) \in RouteID\ k_{0,1\ldots n}$$

Related to the average travel speed calculations (see 5.7), it is important to know the proportion of tracked route compared to the whole route from origin harbor to the destination. The results are most accurate when the lost route length is at a minimum. This is calculated by dividing the total tracked length by total length of the selected route:

$$\textit{Proportion of tracked route} = \frac{Tracked\ path\ length}{Total\ path\ length} \in RouteID\ k_{0,1\ldots n}$$

In addition to quantification of the quality of individual journeys, also assessing the quality of AROS as a whole is done by comparing the data to simultaneous in-situ GPS-measurements made on the field with high sample rate (see 4.5). Assessment is done by comparing the topologies of the two datasets, and searching significant gaps in the tracking data (by visual exploration with geovisual analytics). In addition to topological differences between trajectories and in-situ measurements, also the travel speed calculations are assessed by comparing the two datasets together. Evaluating the accuracy of the developed GKD tool (TRAT) is conducted by visually assessing the accuracy of journey detection and identification of the vessel's navigation direction (upstream / downstream) by utilizing interactive 3D-visualization (geovisual analytics) of trajectories with uDig software.

VI. RESULTS

6.1 Travel speed of individual journeys and their relation to river heights

Figure 25 and Figure 28 represent the characteristics of average travel speed of individual journeys as radar plot diagrams separated by navigation directions. The directional axis (0-360º) represents time and the length of the red and black lines represent the average travel speed and trajectory quality (see 5.11) of a single journey. The direction or time angle of the line that represents the result is calculated for a full year with following formula:

$$\text{Angle} = \frac{Time}{366} * 360$$

The time is divided by the number of days in the year 2012 (366 days) and then multiplied by the degrees of a full circle (single day = 0.98º). Daily water levels are represented with blue polygon in the middle of the diagram as a measure of difference from year's average water level (white ring).

Altogether there are 65 individual journeys detected from the dataset with good tracking quality and represented in Figure 25 and Figure 28. By looking at the temporal distribution of the journeys, it can be seen that the vessels have been moving fairly regularly throughout the year, i.e. having at least one travelled journey per month/direction. However, there have been periods between October to November on Iquitos-Pucallpa route (IQT-PUC) and January to February on Iquitos-Yurimaguas route (IQT-YUR) when there are no observations.

From the radar plots (Figures 25 and 26, pages 49 & 50) it is possible to see that the downstream navigation seems to be altogether faster (mostly between 15-17 km/h) than upstream navigation where travel speeds are mostly below 14 km/h. When looking at the IQT-PUC route (Figure 25) it seems that the travel speed patterns has strong correlation (R=0.8553) with water levels (light blue polygon) on downstream navigation, i.e. when the water level is high the travel speed is high and vice versa. Fitting simple regression model ($y=\alpha+\beta_x+\varepsilon$) between variables (dependent=speed, predictor=river height) reveals that river height of Ucayali explains over 73% (R^2=0.7315) of the variance of travel speed which is considerably high. When travelled upstream the travel speed pattern seems to be more stable, i.e. there is less variation between travel speeds.

There seems to be some evidence that the travel speeds are slower when water levels are high and vice versa (upstream) but the correlation and explanatory power between water level and travel speed is low (R=0.2433, R^2=0.0592) which indicates that there is no statistical relation between the variables.

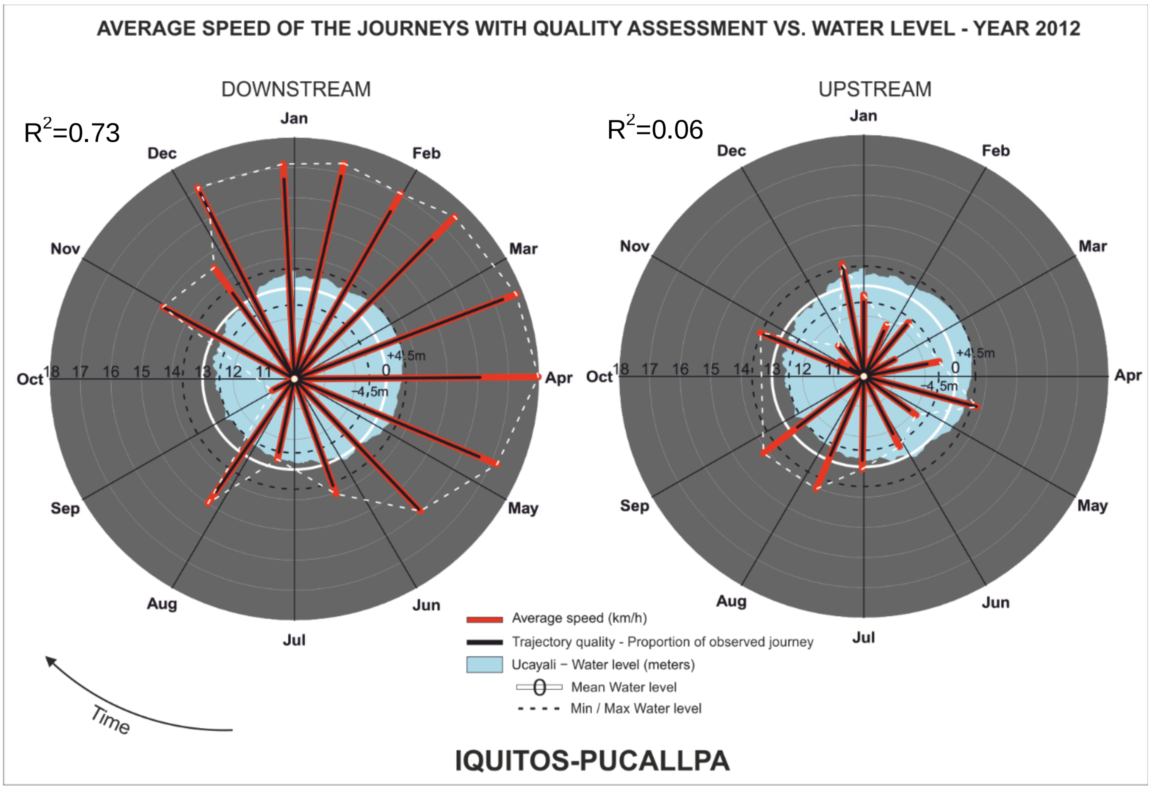


Figure 25. Iquitos-Pucallpa. Radar plot representing the average speeds of individual journeys vs. water level and their R^2 values from linear regression (y=speed, β=water level).

On Iquitos-Yurimaguas route (IQT-YUR) (Figure 26) the connection between water level and travel speed is altogether not as clear as on IQT-PUC route. Visually it seems that the travel speeds are lower during low water (August-October) than on other months when travelled downstream but statistically there seems to be no connection between water level and travel speed (R=0.2579, R^2=0.0665). The results of upstream navigation are quite similar, i.e. there seems to be no connection between travel speed and water level (R=0.2752, R^2=0.0757). The results could be more representative and better if there would be more tracked journeys during the high water (January-February).

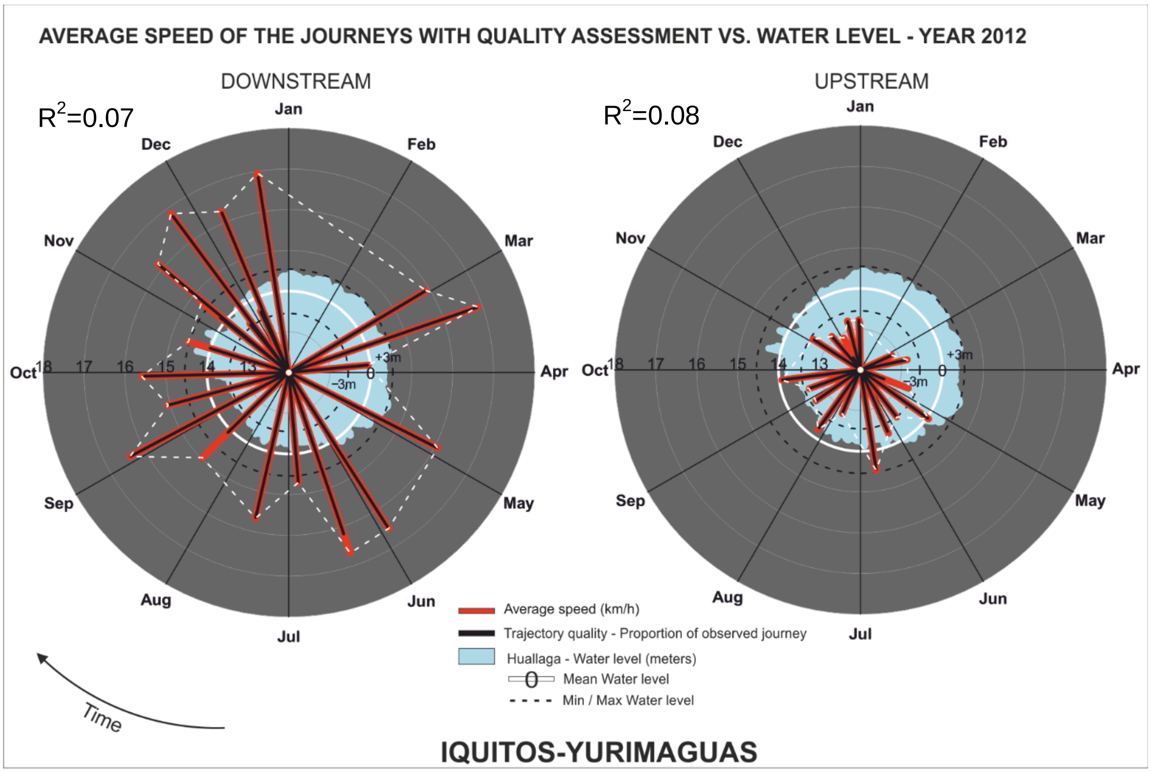


Figure 26. Iquitos-Yurimaguas. Radar plot representing the average speeds of individual journeys vs. water level and their R^2 values from linear regression **(**y=speed**,** β=water level**)**.

Altogether there seems to be clear connection between travel speed and water level only on Iquitos-Pucallpa route when travelled downstream. All of the other cases that were studied did not have high correlation and the explanatory power of river height according to travel speed was only around 6-8%. Figure 27 represents the residuals from the linear regression where fitted value is in this case travel speed. Noteworthy from the plots is exceptionally good fit of model on higher travel speeds of Iquitos-Pucallpa route when travelled downstream which is evident also in Figure 25 and explains the high R^2 value of the case.

When looking at the trajectory quality (black lines within red lines) it seems that the results are fairly representative since the percentage of tracked journey (total path length) are clearly over 50% for almost all of the journeys (2 exceptions) and mostly the tracked-journey percentage is higher than 90% which indicates good representativeness.

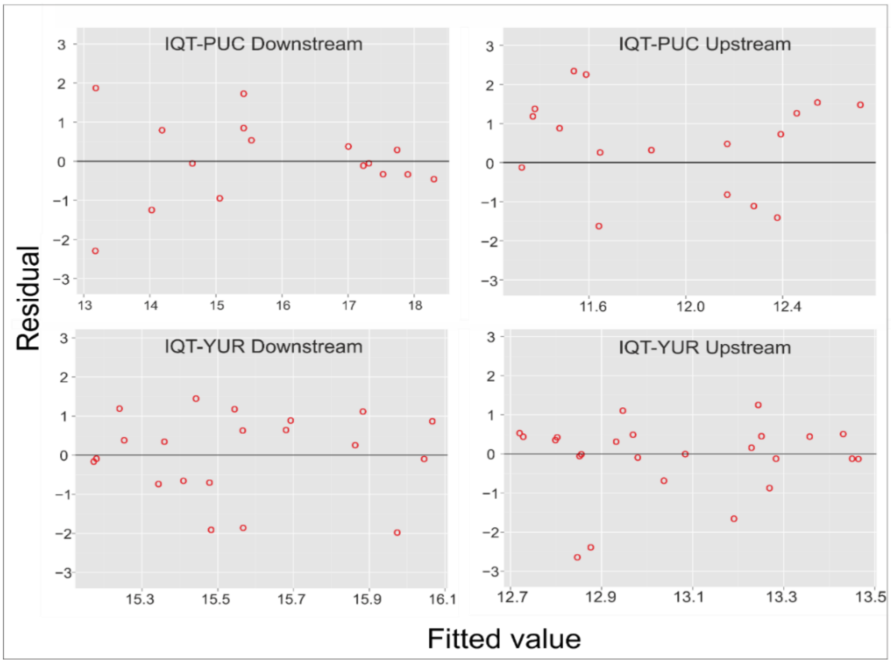


Figure 27. Residual plots from simple linear regression model.

6.2 Seasonal and directional travel speeds

Figure 28 and Figure 29 represent the seasonal movement characteristics of the vessels along the navigation routes where route position indicates the network distance from the city of Iquitos. Also additional information about populated places is included in the graph. Seasons have been separated into three classes (high water, low water and intermediate) mainly based on actual river height information (see Table 4 and Table 5 for chosen time periods) provided by SEHINAV (2013). Also information (see Appendix III) about the typical river stages at the Peruvian Amazon was used to guide and validate the decisions. Directional average (purple dashed line) indicates the average travel speed of all observations (regardless of the season) allocated for each route position.

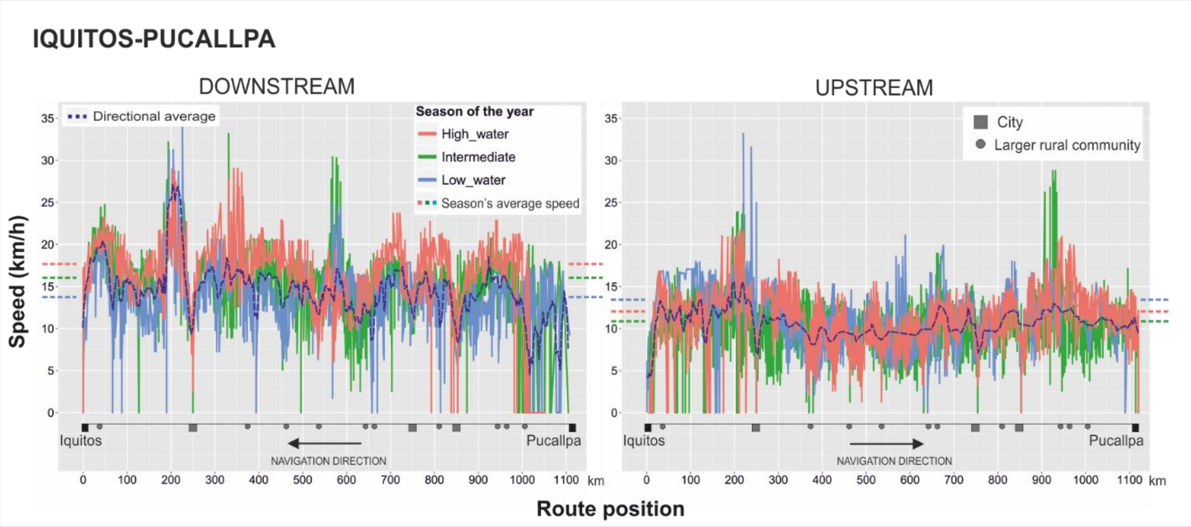


Figure 28. Movement profiles of Iquitos-Pucallpa route representing the seasonal travel speed dynamics.

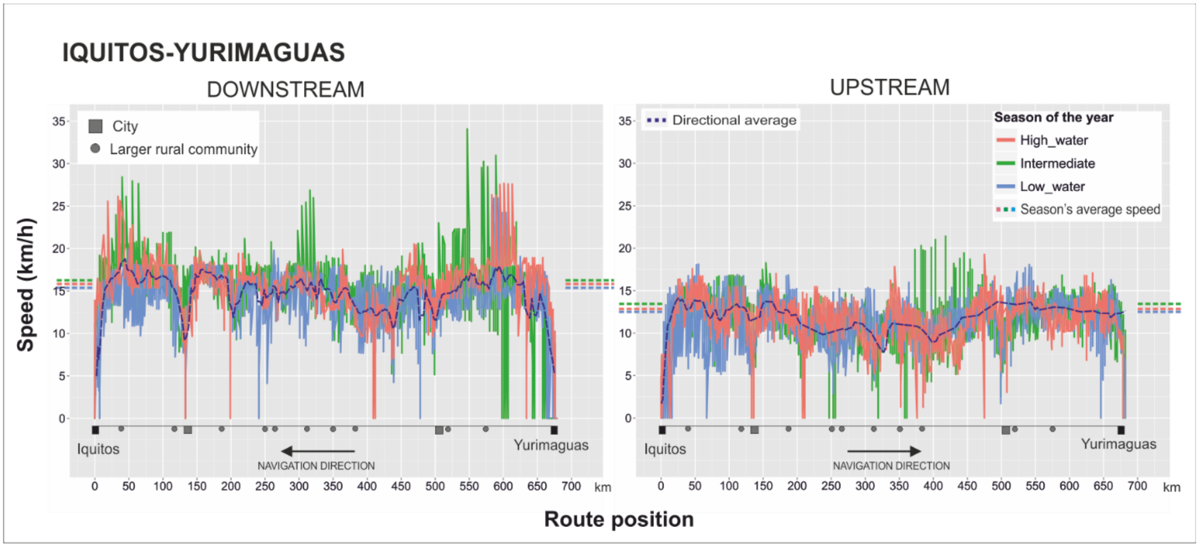


Figure 29. Movement profiles of Iquitos-Yurimaguas route representing the seasonal travel speed dynamics.

From Figure 28 and Figure 29 the obvious conclusion is that the travel speed varies in different parts of the river. The effect of populated places on travel speed patterns is evident and the travel speed clearly decelerates near the cities and larger rural communities. There are also areas where travel speed is clearly higher than generally (e.g. between route positions 170-220 on Figure 28). Both routes have more variation in travel speeds when travelled downstream. Comparing the routes together suggests that the navigation is altogether more stable on shorter IQT-YUR route with less variation compared to IQT-PUC route.

Table 4. Iquitos-Pucallpa. Seasonal travel speed characteristics and comparison between seasons and navigation direction.

Season	Time period	# Tracked journeys / observations		Average travel speed (km/h)		Standard deviation (km/h)		Minimum (km/h)		Maximum (km/h)	
		Down	Up	Down	Up	Down	Up	Down	Up	Down	Up
High water	Jan 1 - Apr 30	5 / 733	6 / 950	17.5	12.0	0.3	1.2	17.1	10.0	18.0	13.8
Intermediate	May 1-Jun 30, Nov 16-Dec 31	6 / 922	8 / 1709	15.8	11.1	1.2	2.2	14.0	7.5	17.1	13.8
Low water	Jul 16 - Nov 15	4 / 822	5 / 812	13.4	13.2	1.7	1.2	10.8	10.9	14.9	14.2

Seasonal difference of average speed				
Season	km/h		Percentage	
	Down	Up	Down	Up
High vs. Low water	4.1	-1.2	+30.6 %	-9.1 %
High water vs. Intermediate	1.7	0.9	+10.8 %	+8.1 %
Intermediate vs. Low water	2.4	-2.1	+17.9 %	-15.9 %

Directional difference of average speed (Downstream navigation vs. upstream)		
Season	km / h	Percentage
High water	5.5	+45.8 %
Intermediate	4.7	+42.3 %
Low water	0.2	+1.5 %

Table 5. Iquitos-Yurimaguas. Seasonal travel speed characteristics and comparison between seasons and navigation direction.

Season	Time period	# Tracked journeys / observations		Average travel speed (km/h)		Standard deviation (km/h)		Minimum (km/h)		Maximum (km/h)	
		Down	Up	Down	Up	Down	Up	Down	Up	Down	Up
High water	Jan 1 - May 15	4 / 479	5 / 661	15.7	12.7	1.1	1.3	14.0	10.2	16.9	14.1
Intermediate	May 16-Jul 15, Nov 16-Dec 31	4 / 1035	4 / 1433	15.9	13.5	0.8	0.7	14.7	12.4	16.7	14.5
Low water	Jul 16 - Nov 15	7 / 886	7 / 1130	15.4	12.5	0.7	2.5	14.6	6.4	16.4	13.9

Seasonal difference of average speed				
Season	km/h		Percentage	
	Down	Up	Down	Up
High vs. Low water	0.3	0.2	+1.9 %	+1.6 %
High water vs. Intermediate	-0.2	-0.8	-1.3 %	-5.9 %
Intermediate vs. Low	0.5	1.0	+3.2 %	+8.0 %

Directional difference of average speed (Downstream navigation vs. upstream)		
Season	km / h	Percentage
High water	3	+23.6 %
Intermediate	2.4	+17.8 %
Low water	2.9	+23.2 %

Table 4 and Table 5 represent the seasonal travel speed characteristics among the navigation routes by seasonal classes and by navigation directions (upstream/downstream) and comparisons between these variables. Differences between seasons were calculated by subtracting the average travel speed from another (e.g. high water – low water) and then calculating the percentage that represents how much faster (or slower) the first season is from the second. The seasonal directional differences indicates how much faster it is to navigate downstream compared to upstream navigation.

Results show that the tracked journeys have been quite evenly distributed among classified seasons and navigation directions, which is important related to representativeness of the results. The number of individual tracked journeys ranges from 4-8 between season/direction.

When looking at the seasonal average travel speeds (also visually present in Figure 28 and Figure 29) of IQT-PUC route, the results suggest that there is clear difference between seasons. When travelled downstream the results suggest that it is fastest to navigate during high water (17.5 km/h) and slowest during low water (13.4 km/h). Difference between seasons on downstream direction is approximately 30%. When travelled upstream the situation changes thus being fastest to navigate during low water (13.2 km/h) and slowest during intermediate (11.1 km/h). Difference between seasons when travelled upstream is 15%. Results also suggest that during low water it is equally fast to navigate upstream and downstream.

When looking at Iquitos-Yurimaguas route it is evident that the differences between seasons are significantly smaller compared to Iquitos-Pucallpa route. When travelled downstream the average speed ranges from 15.4 km/h to 15.9 km/h and on upstream direction from 12.5 km/h to 13.5 km/h. Percentual difference between seasons is mostly below 5% and the highest difference is 8%. These results therefore suggest there is no seasonal difference according to travel speed on IQT-YUR route. However directional difference of average travel speeds is evident and it seems to be approximately 20% faster to travel downstream compared to upstream navigation.

Comparing the routes (IQT-PUC and IQT-YUR) together suggests that the navigation is altogether more stable on shorter Iquitos-Yurimaguas route which suggests that the larger anastomosing rivers (Amazonas, Marañon) are more stable and easier to navigate compared to narrower and meandering river of Ucayali.

6.3 Spatio-temporal examination of river navigation at Peruvian Amazon

Figure 30 and Figure 31 illustrate the previous results in spatio-temporal context showing how far it is possible to reach within certain temporal constraints along the navigation route from selected cities when travelled upstream or downstream (i.e. accessibility). The chosen time constraints are 6 and 12 hours and the selected cities are Requeña and Lagunas. It should be noticed that these results are based on movement characteristics only, i.e. these results do not take into account the stationary time that is spent at the harbors during the journeys when loading cargo or taking aboard passengers. Network distances were calculated based on seasonal average travel speed information on both navigation directions:

Network distance from the city = *(Average travel speed ∈ season ∈ direction) * time*

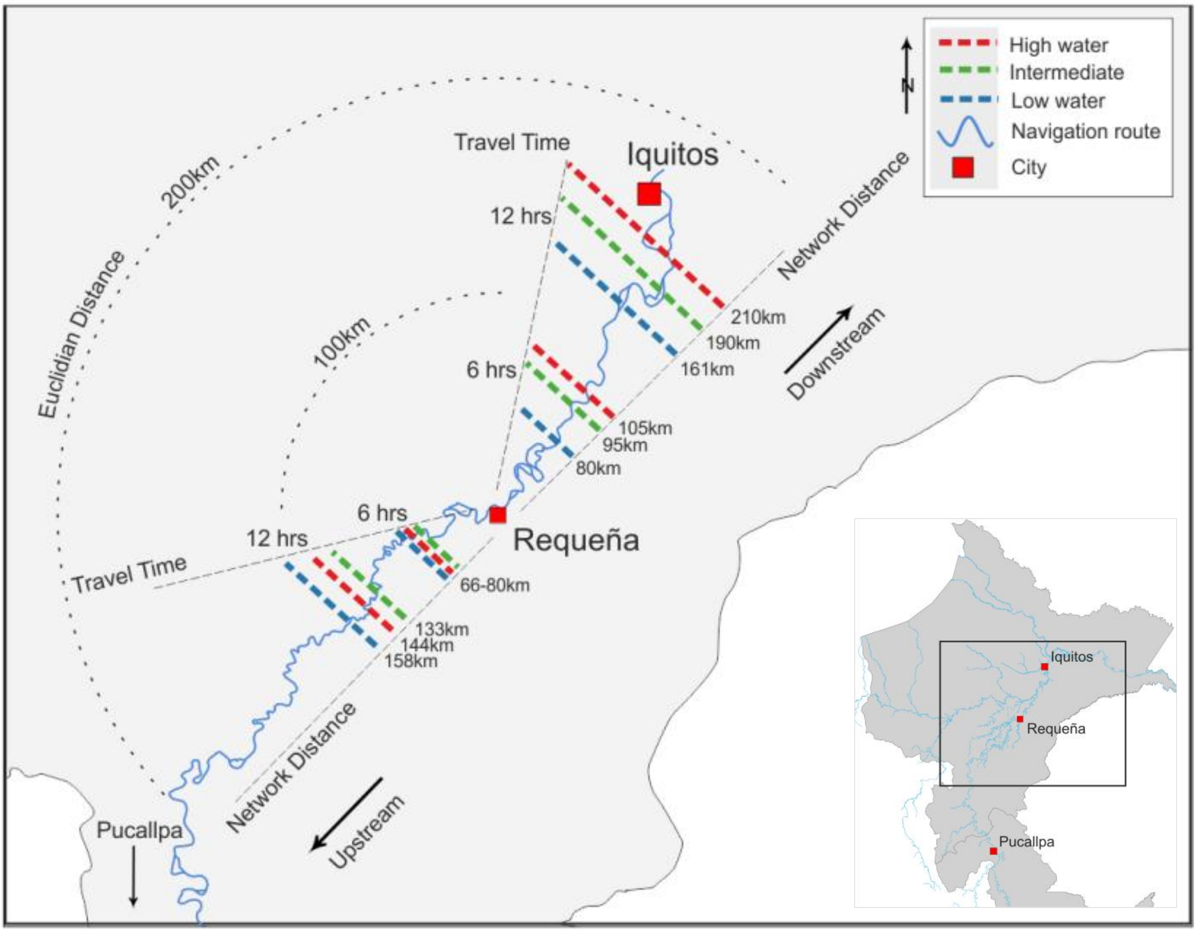


Figure 30. An accessibility map representing how far it is possible to reach from the city of Requeña in 6 and 12 hours.

Figure 30 represents a map of reachable distances from the city of Requeña which is situated along the Iquitos-Pucallpa route 244km upstream from Iquitos. The directional difference is clearly visible on the map as the downstream navigation reaches 210 kilometers at best when upstream navigation reaches only 158 kilometers in 12 hours.

Differences between seasons are also clearly visible especially when travelled downstream.

Results suggest that it is nearly possible to reach the city of Iquitos in 12 hours from Requeña during the high water as the remaining distance is only 34 km (2 hours of travel). Comparing this to low water navigation shows that the absolute distance and time distance are clearly higher - the remaining distance to Iquitos is 83 kilometers which means 6 hours of travel during the low water.

Seasonal differences are not as evident when travelled upstream. However since the distances are long, even these smaller differences becomes relevant. Reaching the city of Pucallpa which is 876 kilometers upstream from Requeña takes 66 hours (2.75 days) during low water (fastest) and 79 hours (3.3 days) intermediate (slowest).

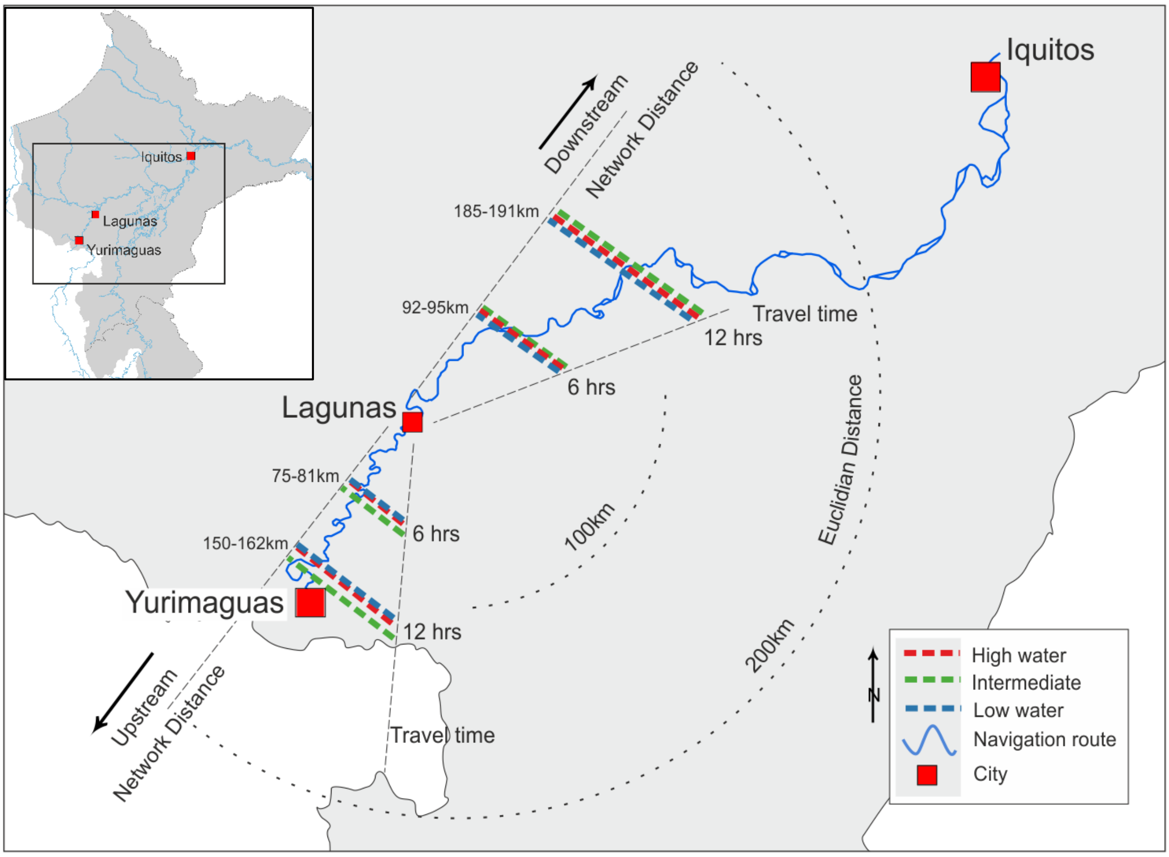


Figure 31. An accessibility map representing how far it is possible to reach from the city of Lagunas in 6 and 12 hours.

Figure 31 represents a map of reachable distances (accessibility) from the city of Lagunas which is situated along the Iquitos-Yurimaguas route 183km downstream from Yurimaguas. Map reveals how small the differences are between seasons since the lines that indicate reachable distances are overall highly clustered. Also the difference between navigation directions is less obvious compared to Iquitos-Pucallpa route.

Results suggest that it is almost possible to reach Yurimaguas within 12 hours on all seasons as remaining distance is between 21 to 33 kilometers and the remaining travel time is approximately 1.5 hours (93-100 minutes). Reaching the city of Iquitos (496 km downstream) from Lagunas takes approximately 32 hours (1.3 days) during all seasons.

6.4 Effect of sinuosity to travel speeds

Figure 32 represents aggregated travel speed characteristics during high water (red lines) and low water (turquoise lines) against rate of sinuosity of the river (blue dashed lines) on the route Iquitos-Pucallpa (IQT-PUC) which was chosen because it has more variation in travel speed characteristics (see Figure 28 and Figure 29). X-axis represents the route positions, y-axis on the left side represents the travel speed characteristics, and y-axis on the right side represents the rate of sinuosity.

Because the interest is now on studying how the sinuosity affects the travel speed characteristics along the navigation routes, the actual travel speed information (that varies between seasons) was standardized to fit the same scale where average speed of individual journey represents 0-line. Travel speeds have been standardized by transforming the navigation speeds of segments (see 2.3.5 and 5.7) belonging to individual journey in a way that they represent the difference from calculated average speed of that journey:

Difference from average speed (km/h) = *Travel speed of segment – Average travel speed* $\in JourneyID\ k_{0,1\ldots n}$

In addition, the movement characteristics of the journeys were classified into two classes (high water and low water, see Table 4. Iquitos-Pucallpa. Seasonal travel speed characteristics and comparison between seasons and navigation direction.) to find out if there is differences between seasonal opposites. Movement data is noisy, therefore removing this noise was done by aggregating the seasonal classes into single lines in a way that travel speed on each route position represents the average speed of all the observations of different journeys associated to that position, and then smoothing the data by calculating moving average values with window size 3 that gently reduced the noise but maintained the movement characteristics (see 5.11 for details).

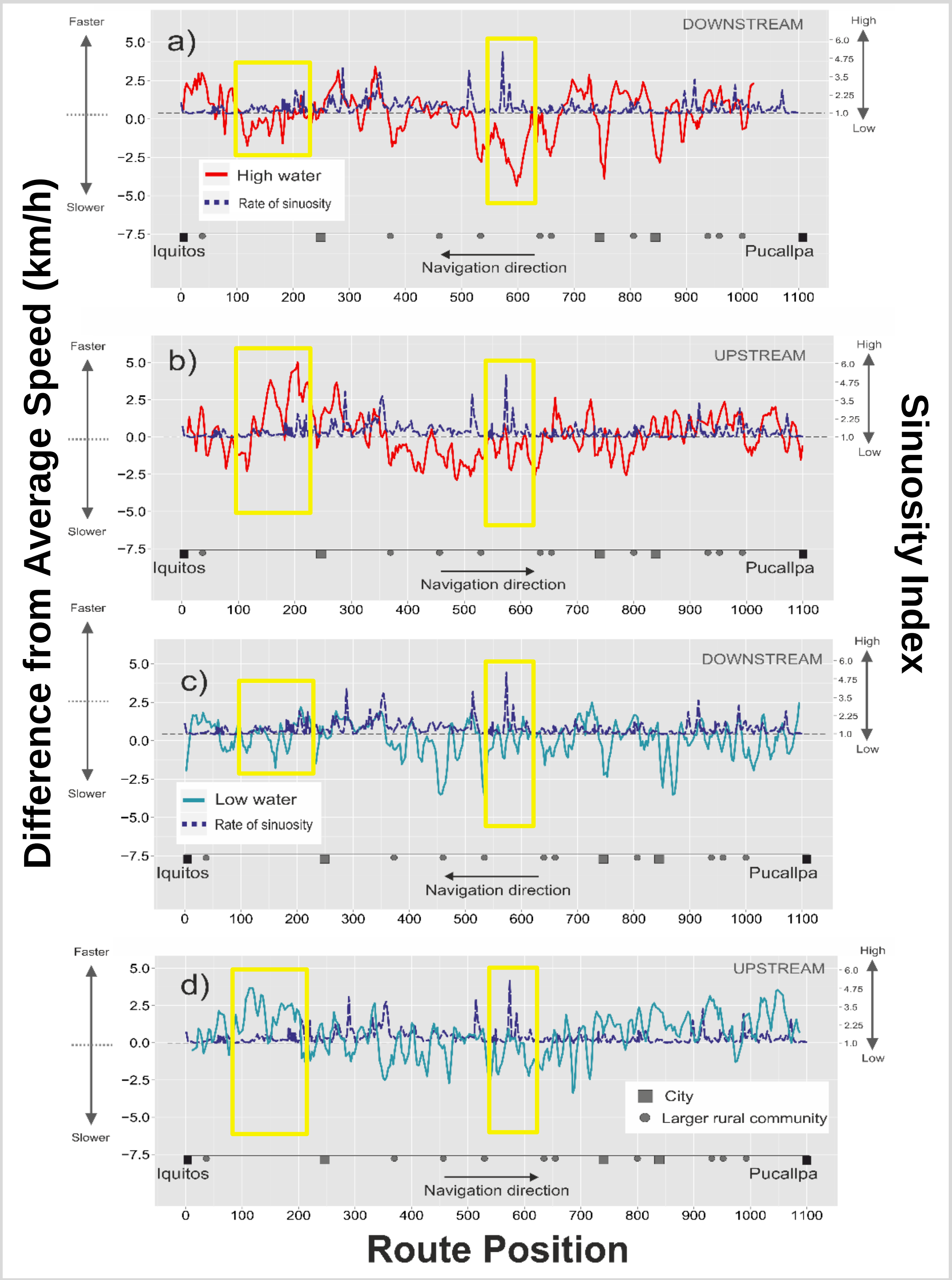


Figure 32. Comparison of sinuosity index and travel speed characteristics on Iquitos-Pucallpa route. Graphs a) and b) represent results during high water, and graphs c) and d) during low water.

Sinuosity index value 1 indicates straight line (i.e. Network distance = Euclidian distance) and higher values indicate how much longer the network distance is compared to Euclidian distance, i.e. sinuosity index 2 means that the network distance between two locations is two times longer than the Euclidian distance (see 5.10 for more details).

Results (Figure 32) suggest that there seems to be some connection between sinuosity and travel speed characteristics observed with AROS but only at specific areas. In the results are present only observations that had travel speed higher than 5 km/h which was done (before aggregations) to remove the effect of harbors and populated places (larger villages etc.) to the travel speeds. However, as can be seen from the results this procedure did not work as wanted because near populated places (shown as gray rectangulars and circles at the bottom of the graph) the travel speeds are commonly much lower than would be assumed by the sinuosity index of those areas. This is probably caused by low sample rate of AROS that does not enable to achieve such spatially detailed analysis.

Luckily, the highest sinuosity index value is on the area (approximately 570-590 km from Iquitos) where there are no populated places in the surrounding areas (highlighted with yellow box on the right on each graph), and thus concentrating at this particular area allows making few conclusion with some confidence related to the connection between sinuosity and movement characteristics. There indeed seems to be some connection between the factors especially during high water (graphs a and b) when travel speed starts to move well below average near the areas where sinuosity index is high. This effect can be seen on both navigation directions but more clearly when travelling downstream. When comparing the results of the same locations during low water (graphs c and d), it seems that the sinuosity of river has no effect on movement characteristics which is interesting.

There is also another interesting stretch of navigation path near Iquitos (approximately route positions 100-200) where the effect of populated places is lower, sinuosity index is low, and there seems to be similar patterns during high water and low water (highlighted with yellow boxes on the left side of the graphs). When travelling upstream (graphs b and d) there are two peaks where travel speeds are higher than average, which is expected on low sinuosity, but there seems to exists some smaller village at route position 160 (approx.) where vessels are stopping since the travel speeds always drop near that location. There also seems to be clear difference between navigation directions on this area: when travelling upstream the navigation speeds are higher than average, and vice versa when travelling downstream. Intuitively the assumption is that the navigation speeds are higher when travelling downstream but in this case the results indicate the opposite which is interesting and possibly indicates something about directional differences in local transportation (i.e. more stops when travelling downstream).

VII. DISCUSSION & CONCLUSIONS

7.1 Technical assessment – Evaluation of topology

Figure 33 represents the topological comparison between AROS (blue line) and the high sample rate GPS-measurements made with Garmin. Some of the notable topological differences between datasets are highlighted with yellow circles. When looking at Iquitos-Yurimaguas route (map a), it seems that altogether the topological accuracy of AROS is quite good since there are no significant visible differences between AROS and Garmin. Differences can be found only at the highly meandering parts of the river where lower

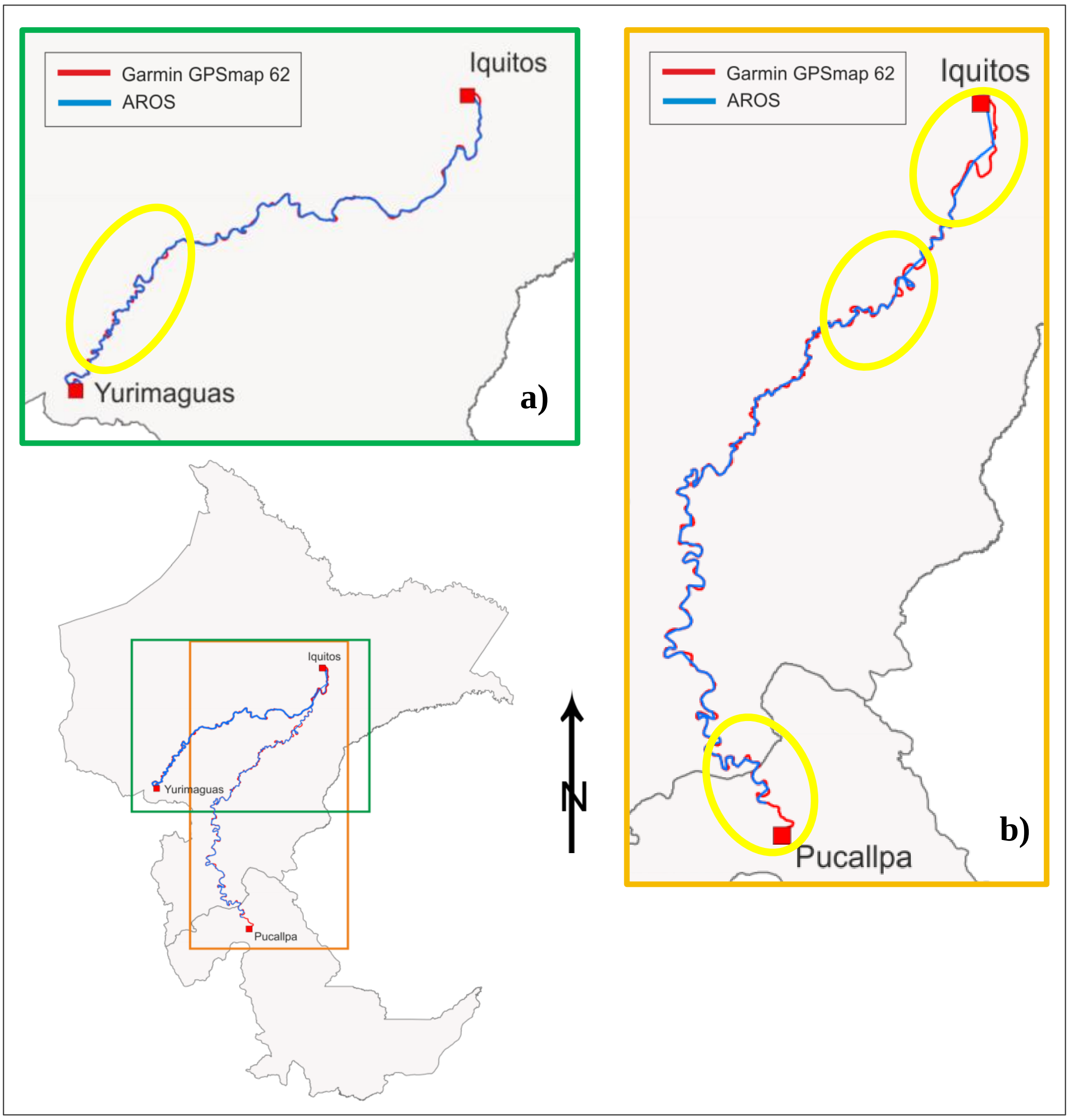


Figure 33. Topological comparison of AROS and in-situ GPS-measurements made at the field during October 2012.

sample rate of AROS (1 observation/10 minutes) causes ‘cut-offs’, i.e. some of the river bendings are left unobserved with AROS.

When looking at the route Iquitos-Pucallpa (Figure 33b) it is evident that the topological accuracy of AROS is not as good as when compared to IQT-YUR route. There are similar and more significant cut-offs at the meandering parts of the river, and in addition there is poor observation density at the beginning of the journey (near Iquitos) which causes significant topological differences compared to Garmin GPS-measurements. There is also lack of observations at the end of the journey, i.e. the last 39 km of the journey was not recorded during the in-situ measurements because AROS GPS-device went off at night during the measurements.

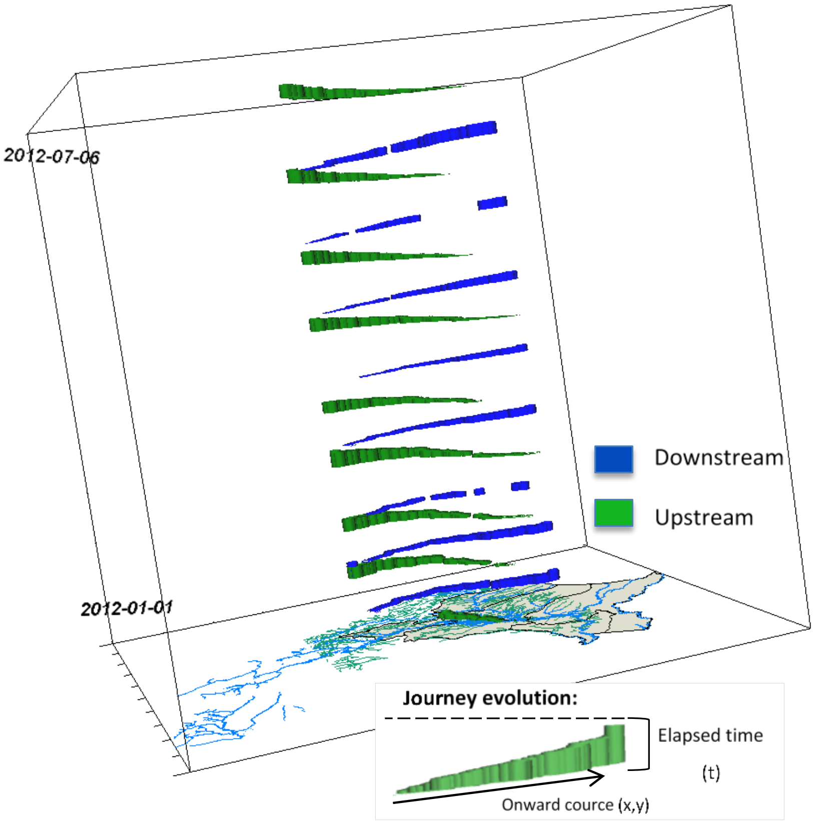


Figure 34. Trajectory quality, as means of tracking continuity, can be assessed visually with the concept of journey evolution that enables to reveal if there are gaps existing along the journeys.

Lack of observations and topological inaccuracies during the journeys illustrate typical shortcomings of AROS. Figure 34 demonstrates a visual method utilizing geovisual analytics to assess the trajectory quality of AROS with the concept of journey evolution. During the year 2012 there were 48 recorded trajectories with good tracking continuity and 17 trajectories with moderate shortages in tracking continuity and 14 trajectories with significant shortages in the tracking continuity.

7.2 Technical assessment – Accuracy of travel speed calculations

Figure 35 represents the travel speed comparisons between AROS data and in-situ measurements where x-axis represent the route position (i.e. network distance from Iquitos) and y-axis represents travel speed (km/h). Solid lines represent how the average speed progresses during the journeys (see 5.7 for details). Dashed lines represent the travel speed between consecutive observations (AROS) and the consecutive route positions (Garmin). Accurate comparison of travel speeds between datasets along the navigation routes were done by associating the observations with route positions (i.e. reference points, see 4.4). Each Garmin-observation (1 obs./minute) is associated to the nearest route position that are evenly distributed by 1km intervals along the travel routes, and travel speed calculations were based on these aggregated (mean) observations (approximately 3 observations / route position).

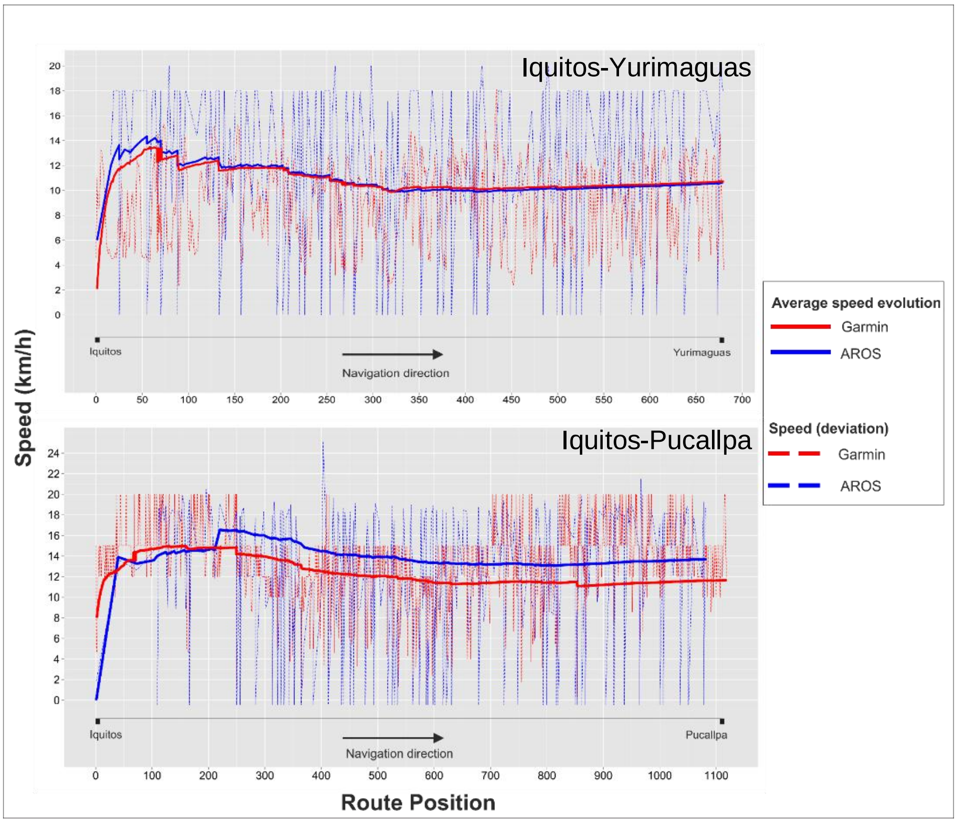
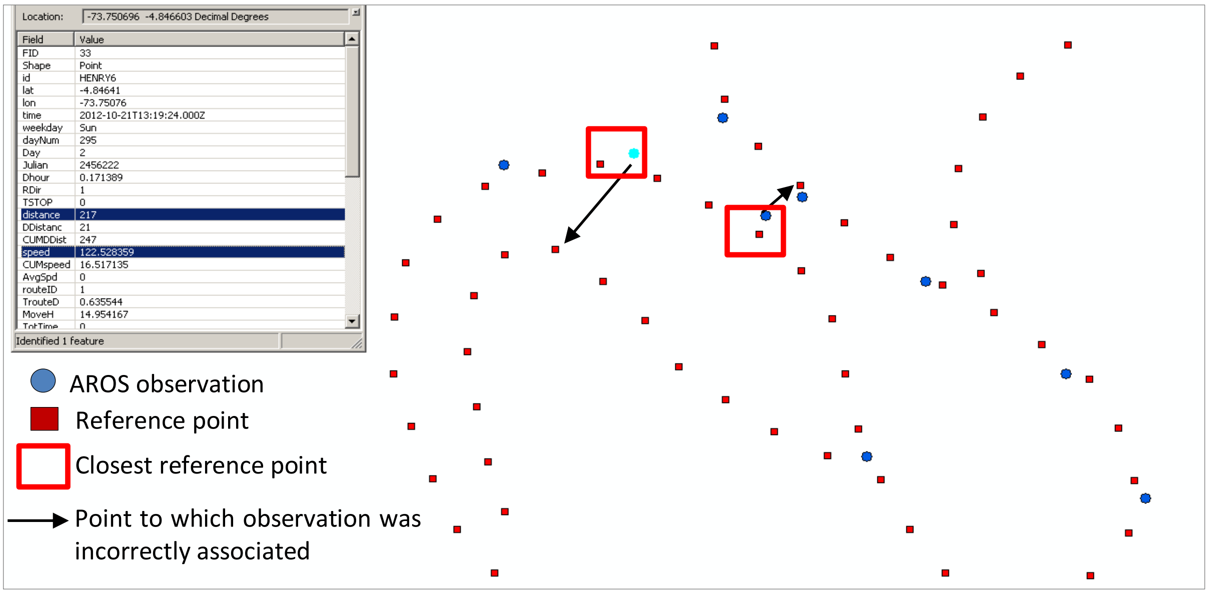


Figure 35. Comparison of travel speed calculation results between AROS data and in-situ GPS-measurements made with Garmin GPS-device.

Results suggest that deviation of travel speeds (dashed lines) is higher with AROS compared to more densely distributed Garmin observations. When comparing the progression of average travel speed (solid lines), the results show that on Iquitos-Yurimaguas-route the average speed values are almost identical between Garmin (10.72 km/h at destination) and AROS (10.48 km/h at destination). High similarity of average travel speed calculations (2.3 % difference) indicates high accuracy of the results even though the deviation of (individual) travel speeds (of AROS) is higher along the route.

However, on Iquitos-Pucallpa route (lower graph in Figure 35) the average travel speed calculation of AROS (13.68 km/h at destination) differs notably from Garmin measurements (11.66 km/h at destination) and the difference between datasets is 2 km/h (17.3% difference). Average travel speeds are almost identical at the beginning of the journey but approximately at route point 220 (km from Iquitos) the average travel speed calculation of AROS separates from Garmin which suggests inaccuracies either in data or in algorithms of TRAT. Closer look of the data reveals that there is indeed inaccuracy in calculations because the spatial join between reference dataset and AROS observation was incorrect thus causing the difference in average speed calculations (see Figure 36). This critical assessment of travel speed calculations reveals that correct spatial join is important for accurate results even though the differences between AROS and in-situ measurements (Garmin) are not critical.



Location: -73.750696 -4.846603 Decimal Degrees

Field	Value
FID	33
Shape	Point
id	HENRY6
lat	-4.84641
lon	-73.75076
time	2012-10-21T13:19:24.000Z
weekday	Sun
dayNum	295
Day	2
Julian	2456222
Dhour	0.171389
RDir	1
TSTOP	0
distance	217
DDistanc	21
CUMDDist	247
speed	122.528359
CUMspeed	16.517135
AvgSpd	0
routeID	1
TrouteD	0.635544
MoveH	14.954167
TotTime	0

Identified 1 feature

● AROS observation
🟥 Reference point
□ Closest reference point
→ Point to which observation was incorrectly associated

Figure 36. Incorrect spatial join to reference dataset causes calculation errors.

7.3 Technical assessment - Accuracy of journey and navigation direction identification

Figure 38 (on page 68) represents the whole dataset used in this study as a 3D snapshot picture. The interactive version of this visualization is used for conducting the technical assessment.

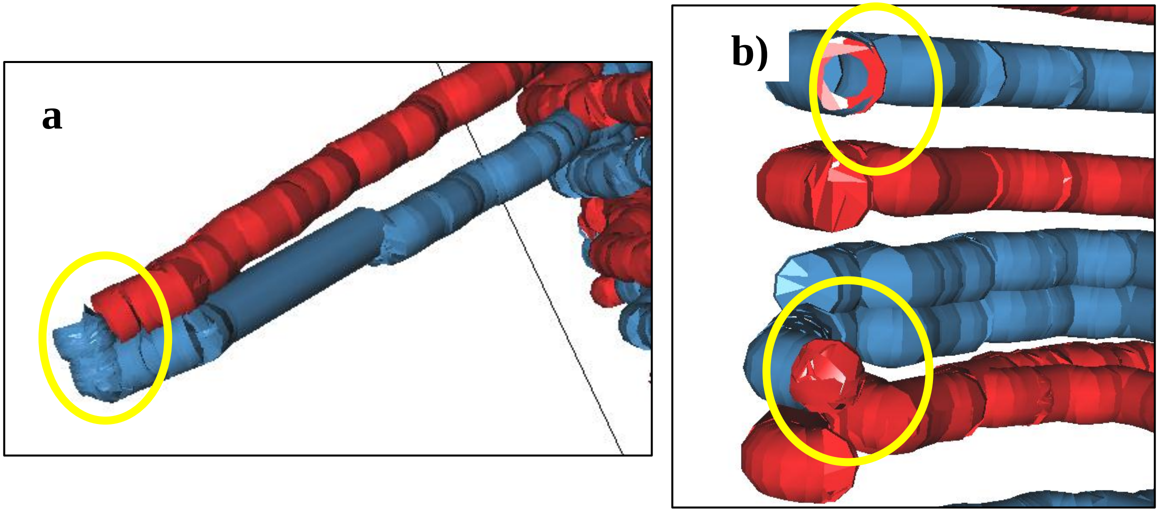


Figure 37 Challenging situations for TRAT's algorithm can be found at the beginning of the trajectories.

As can be seen from the Figure 38 the journey identification algorithm of TRAT seems to be working correctly since all of the trajectories are nicely separated from each other and there are no signs of "messy" or mixed trajectories. Also the accuracy of navigation direction detection seems to be mainly accurate (assessed by interactive exploration).

However, occasionally there seems to be certain analytically challenging situations for the algorithm at the beginning of the journeys when the navigation direction changes (indicated with red/blue colors) which are highlighted with yellow circles in Figure 37. In Figure 37a it seems that the two trajectories are connected together (i.e. problem with journey identification) but checking this interactively with uDig confirms that trajectories were however correctly separated from each other, i.e. they have separate journeyIDs. Figure 37b reveals that there seems to be some difficulties at the very beginning of the journey to identify direction of movement which might be happening on the first few observations because there are not enough observations for the algorithm to accurately determine the principal navigation direction.

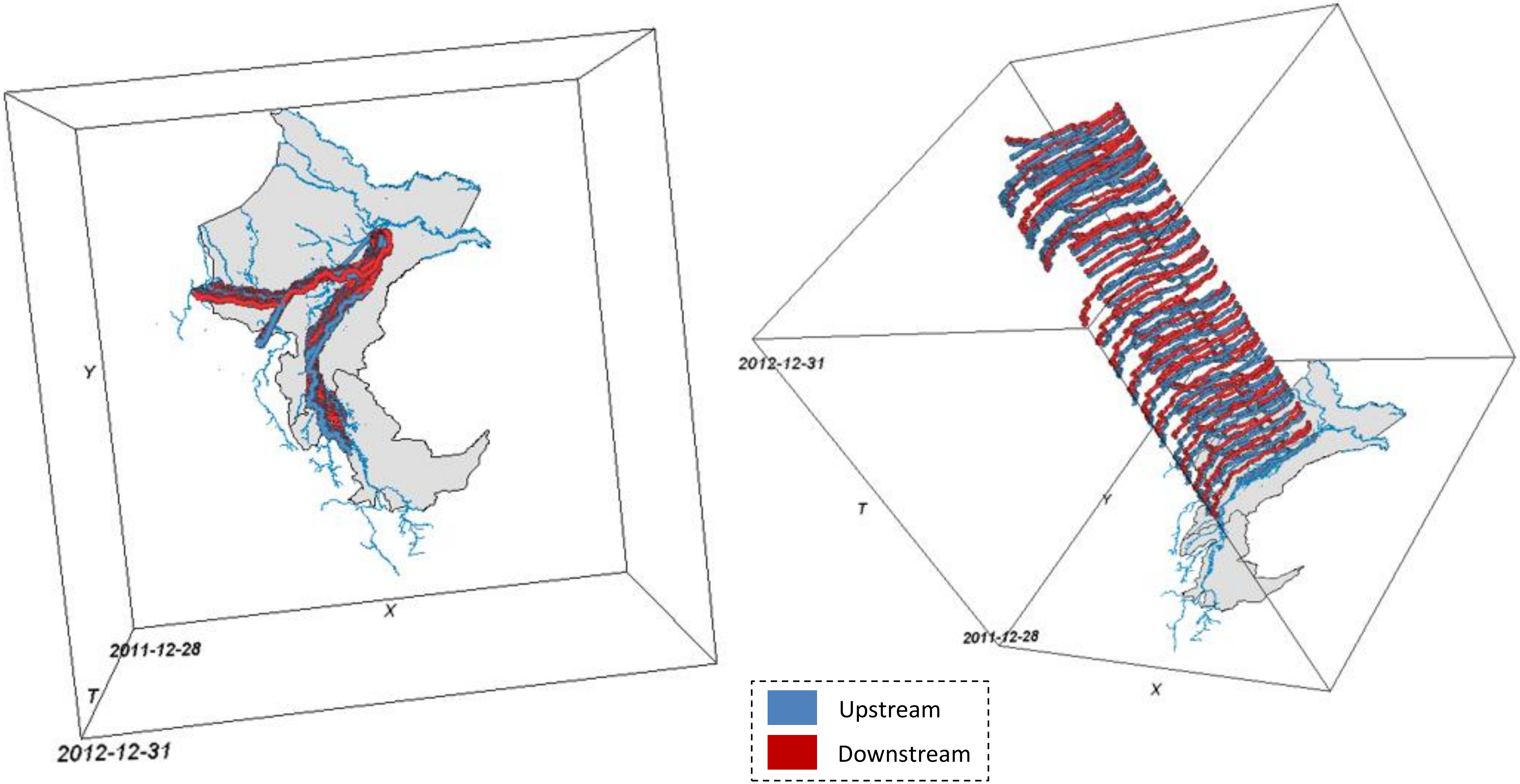


Figure 38 Space-time cubes representing individual journeys at Peruvian Amazon identified with TRAT. Data was collected with AROS in year 2012.

7.4 Evaluation of the significance of errors

In the data there are three main sources of error caused by different factors. As Dodge et al. (2009) mentions: the raw mobility data obtained with tracking devices contains usually some degree of noise, gaps and outliers. The reconstruction accuracy of trajectories and their level of spatial accuracy and temporal granularity (seconds/minutes etc.) depends on the quality of the log entries (Giannotti & Pedreschi 2008). The most significant source of error related to the analytical accuracy of TRAT is the missing data during the journey as was seen in Figure 34 (page 64). Other minor sources that cause errors and thus affect the accuracy of results are the signal loss of the GPS-devices causing position errors, and the errors caused by incorrect spatial join to the reference data and position error in this process, and occasional position errors in the reference dataset (see Figure 39).

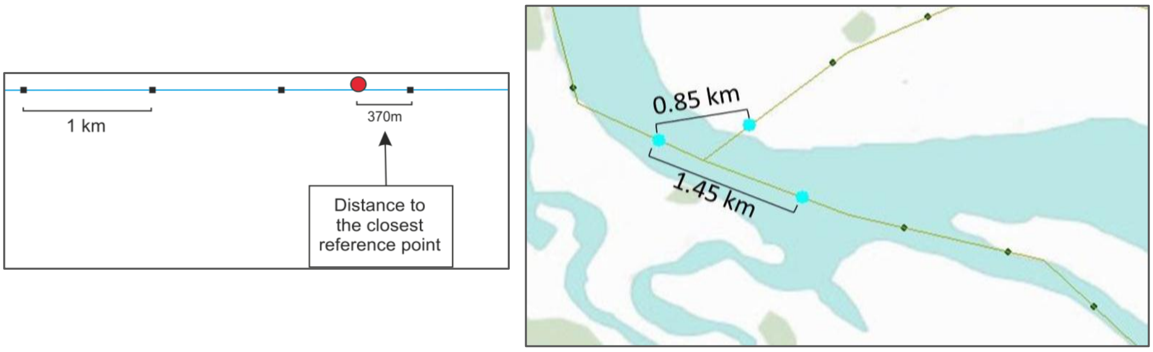


Figure 39. Positional error when joining to the reference dataset (on the left) and position errors in the reference dataset where network distances differ from aspired 1km interval (on the right).

The satellite messengers optimally send their location information at every 10 minutes, however, it is not uncommon to have sample rates lower than that which decreases the data quality. The most significant data quality problems in this study are related to this matter which causes observation gaps in the data. Usually observation gaps that are longer than usual arise because the GPS-devices used are designed in a way that they need to be reactivated every 24 hours. It is the responsibility of the crew of the ships to take care that the device is reactivated regularly thus making it possible that the crew sometimes forgets to reactivate the GPS which might cause even loss of several hours of data.

Another matter that can effect on data quality is related to signal strength of GPS-devices with satellites. A clear sky view is required to get accurate position information with a GPS device (SPOT 2007) and since there might be e.g. trees blocking the sky, there might be errors in the data. The GPS-receivers used in this study are located in the cockpit of the vessel with large windows which is normally at the highest part of the boat. This usually provides good sky view and since the vessels are navigating along the rivers there are usually no trees blocking the satellites. However the steel structure of the ships might cause some problems with the signal and it is possible that some of the observations are lost or the position information might be inaccurate because of the weak signal.

The significance of the mentioned errors varies depending on the type of analysis used. When analyzing the total travel time or average speed of a journey, it is basically enough to have only two observed points (at the start and at the end of the route) to calculate accurate average speed of that journey. But when analyzing how the movement evolves during the journey, the accuracy of such an analysis depends on how densely the observations are located.

It is important to acknowledge that the movements of the vessels along the rivers are more foreseeable and regular compared to e.g. movements of cars along the road network. The travel speed of the vessel is quite constant and much slower than cars, and the navigation along the river is quite smooth since there is no traffic or restrictions in the same way as in road networks. This slightly compensates the shortcomings of the data.

7.5 Comparing AROS and TRAT to other studies and applications

Analysis of mobility data can be considered as ‘hot-topic’ right now and there are many applications and studies where different kind of tools have been developed to analyze the data from tracking devices. However there are not many studies that would focus on developing methods for analyzing and visualizing the movements of vessels: Demsar & Virrantaus (2010) took spatio-temporally oriented approach in their article by analyzing and developing methods for visualization of space-time trajectory densities based on (AIS) vessel movement data in the Gulf of Finland, while Willems et al. (2009) have similar data and objectives to develop methods for visualizing trajectory densities but taking account only geographic space (i.e. no spatio-temporal approach). In contrast with this study, the AIS movement data utilized in those studies has high sample rate (1 observation / 2-10 seconds) which enables to take totally different approach when doing analysis compared to this study that utilizes AROS data with low sample rate (1 observation / 10 minutes at best). With AIS data it is possible to calculate relatively accurate navigation speed straightly by measuring Euclidian distance between consecutive observations and dividing the result by temporal difference between the two observations. Also detecting individual journeys and calculating different global movement descriptors (such as acceleration) can be done with higher accuracy and greater detail compared to AROS data, even though the analytical methods for calculating these parameters are similar.

However studies that analyze mobility data with low sample rate also exists. Ahas et al. (2007) studied the spatio-temporal movement patterns and activity spaces of commuters in Tallinn region, Estonia by utilizing movement data from mobile phones. Mobile positioning data has also been utilized to model meaningful locations to mobile phone users by utilizing spatio-temporal analyses to detect places where people spend a lot of time frequently (so called anchor points or points of interests) such as home and work (Ahas et al. 2010). Mobile positioning data is collected every time when user makes a call or sends text message. This data includes information about the time when the phone was used and information about the location where the user was during that time, which enables to build trajectories but with fairly low spatio-temporal details.

Mobile positioning data, such as what is utilized for research in Estonia, has presumably lower sample rate than AROS since it is dependent of usage activity of the mobile phone user. In this sense, the mobile positioning data can be considered as being more sparsely scattered mobility data, whereas AIS data is more densely scattered, and thus AROS data situates somewhere in between when ‘rating’ the spatio-temporal data density characters of different mobility data sources. With higher sample rate it is possible to make more detailed spatio-temporal movement analysis, but on the other hand, as details of the data grow, so does the volume of the data. This means that it needs more processing power from the computer and takes more time to perform the analyses.

Altogether the analytical approaches used in TRAT are fairly simple and universal, thus similar methods have been utilized also in other studies (e.g. Marketos et al. 2008; Dodge et al. 2009). However because of the low sample rate of AROS, TRAT utilizes more complicated way of obtaining the distance between observations and identifying the direction of movement by utilizing training dataset and decision-tree classification method, which separates TRAT from methods used in other studies and applications. Similar kind of applications or studies that would utilize and analyze such semi-densely distributed movement data (such as AROS data) and would utilize similar analytical approaches (as in TRAT) to extract knowledge from such data has not been done before to my best knowledge.

7.6 Evaluation of the results - Transportation characteristics in the Peruvian Amazon

Transportation patterns (as well as almost any dynamic phenomena) are in constant change both spatially (e.g. changes in traffic arrangements and service structures) and temporally (e.g. yearly / diurnal changes in transit schedules and people´s locations). Yet, analyses are most often analyzed as a static phenomenon, and analyses are focused on a specific moment and / or based on simplistic assumptions on residents daily mobility patterns. Problems related to the lack of dynamic analysis methods are increasingly recognized (e.g. Li et al. 2011; Tribby & Zandbergen 2012), but difficulties in finding appropriate data for such analyses often hinders the use of more sophisticated approaches.

Content of this thesis is mostly focused on rather technical aspects related to development of TRAT. However this study also includes more practical part which is focused on revealing the transportation patterns at the study area during the year 2012 and how these patterns influence in wider contexts. There are not many published papers that would focus on transportation patterns in Loreto and Ucayali regions as a whole, and most of the papers have focused on more small scale analyses (e.g. Chomitz & Thomas 2003) or based their analyses on Euclidian distances (Peres & Terborgh 1995; Peres & Lake 2003) between places which can be problematic since high sinuosity of the navigation paths (i.e. rivers) may result that the network distances are much longer compared to straight-line distances (Toivonen et al. 2007) (see Figure 14 on page 29). This thesis is the first attempt to take spatio-temporal approach in the study area which is conducted by studying seasonal variation of movement patterns along Amazonian rivers: to my best knowledge there are no other studies taking such approach.

7.6.1 Comparing the results to other studies

Salonen et al. (2012a) compared in their study the network distances of based on river channels (in Loreto region) to straight-line Euclidian distances. They also developed a quantitative model of accessibility patterns as means of time distances from different parts of Loreto to the city of Iquitos. Their analysis was based on measured travel speeds and observed travel times aboard local riverboats during high water (January-February), i.e. analysis did not take into account seasonal variation. Comparing the time distance calculations of this study to analysis of Salonen et al. (2012a), it seems that the results of this study are quite reasonable and consistent with their study. The travel time from Requeña to Iquitos was approximately 15 hours in their study which matches fairly well with the results of this thesis (14 hours) during the high water (i.e. January 1st – April 30th). The difference between the results of this study compared to results of Salonen et al. (2012a) is below 10% which indicates good accuracy of the results. Salonen et al. (2012a) also noticed in their study that navigation differs depending on the channel type (i.e. morphology) and the size of the river. The results of this thesis also confirm that navigation indeed differ depending on the factors mentioned by Salonen et al. (2012). An important factor in determining the transportation characteristics on riverine environment is the direction of movement which affects the navigation speeds (Chibnik 1994 cit. Salonen et al. 2012a). Results of this thesis also confirm this, but with addition that the significance of navigation direction to transportation characteristics depends on the size and morphology of the river.

Acknowledging that the validation of the results of this thesis is based on only one comparative study with actual data about transportation characteristics covering the whole study area (Salonen et al. 2012a), it is necessary to cross-validate results by comparing them to data from different years collected with AROS. That, however, is not on the scope of this thesis.

7.6.2 Evaluating the significance of the results in wider contexts

This study can be considered to have societal significance in different levels since the transportation patterns in the study area affect: 1) the everyday life of local inhabitants, 2) the economy of Loreto and Ucayali provinces (regional scale) and 3) the connectivity between regions in northern South America (national/international). 4) Globally this study can be related to the climate change studies since with long-term surveillance of river transportation with AROS it could be possible to detect e.g. weather anomalies (exceptional flooding/drought) that influence on river navigation in the study area.

As mentioned by Knowles (2008), the transportation is in a key role when trying to find the factors affecting on development of a certain location. A common way to evaluate the level of transportation development is to measure the connectivity between places and analyze accessibility patterns in the study region. This study utilizes the concept of accessibility which has been studied a lot and is known to have wide range of influence on different fields of studies and scales (e.g. climate change, economics, land cover changes, human interactions etc.).

Taking account environmental aspects, many studies have indicated that the growing populations and the need for monetary incomes increasingly endanger the ecologically and economically important flood plain and forest areas in the tropical forests like Amazonia (Kvist & Nebel 2001; Geist & Lambin 2002; Killeen 2007). Thus transportation as means of accessibility to nearby markets has significant role related to conversational aspects, land use pressure and deforestation at the study area (Salonen et al. 2013). Results of this study could provide more accurate data for LUCC models that were used in the study of Salonen et al. (2013).

The development of riverine transportation is also one of the key targets according to IIRSA. IIRSA (Initiative for the Integration of the Regional Infrastructure of South American) is a project that aims at improving the physical links and transportation infrastructure among the South American countries via highways, *hydrovias* (waterways) and energy projects. The initiative targets at improving the standard of living in the South American countries but there has also been a great concern that the organizing principle does not take adequately into account the impacts of the improvements in infrastructure on the extremely valuable and vulnerable Amazonian rainforest areas (Killeen 2007).

In the past we have seen how drastically the new highway corridors in the Brazilian Amazon has altered the environment and accelerated the deforestation and fragmentation process along the new passages (Fearnside 2008). These are true threats anywhere where there are plans to build large road networks through the pristine forest areas. The improvements on the river transportation might offer a less destructive alternative for enhancing the connectivity between South American areas because the waterways are naturally creating the transportation network through the densely forested areas. Thus this study could offer relatively accurate background information for the decision makers responsible for the development of transportation infrastructure in the study area about how the riverine transportation functions along the dynamic river networks.

7.7 Future possibilities of movement analyses and need for transportation oriented studies in the Peruvian Amazon

Mobility data and the mining of such data is still a quite novel area of research which the progress of current information technologies has enabled during the last ten years (approx.). Nowadays, as the use of mobile devices with GPS is more popular than ever, it is inevitable that the availability of location and mobility information has also grown exponentially. Before the development of current wireless and mobile technologies, collecting the location data was only possible by highly expensive and time-consuming means such as field experiments, surveys or with the ad hoc sensors placed on the streets or vehicles. Even though these means are still used, current technology offers a possibility to collect and store mobility data at a very low cost straight from the users. Having continuous access to more accurate, close to real-time information about human movements can help us to better understand the dynamics of our living environment and society thus possibly contributing as “better” decision making in various sectors such as in urban and land use planning as well as when making decisions and plans related to environment.

Everyday actions of people leave digital traces in the information systems of the organizations, and since the communication and computing devices are ubiquitous and carried to everywhere this enables to sense human activity in a territory (Giannotti & Pedreschi 2008). The value of the knowledge about people movement is high for many instances. Urban planner for example, could utilize mobility data for localizing new services (such as library or tourist information point) or for organizing logistic systems based on the movement patterns. For commercial purposes mobility information of people would be off high interest to for example to allocate advertisements more accurately for the eyes of specific customer segments. The location information is getting even more accurate in the future as it will be possible to track even inside the buildings and in densely populated places with Navigation via Signals of Opportunity (NAVSOP) device that utilizes not only the satellites but also the Wi-Fi, TV, radio and mobile phone signals to get an accurate location information of the device (Bae Systems 2012).

Related to mobility analyses in the study area: there is a need to continue studying the riverine transportation patterns in Peruvian Amazon since all of the results of this study need to be assessed and cross-validated by comparing them with data from other years (2013 onwards). AROS data and TRAT provides also many possibilities for future research in the study area and it would be interesting to study on more detailed level how different river morphologies (anastomosing, meandering etc.) affect the transportation patterns. Another interesting study would be to test if it is possible to build a model which could predict the movement characteristics (speed etc.) on different type of rivers based on AROS data from specific stretch of river (such as Ucayali or Marañon). Also 80 interviews were made for the passengers of the vessels at the same time when conducting in-situ GPS-measurements. People were asked how they experience the travelling along the Amazonian rivers, and also questions that handle travel times between seasons and places etc. were asked. Thus it would be interesting to assess the results of this thesis based also on those interviews. Related to utilization of developed methods, it would be interesting to evaluate how TRAT (with few modifications) performs on other data sources such as AIS which provides interesting global scale data of movements of professional vessels. Such data would allow to study and form global transportation patterns and would enable to estimate global flows of cargo and goods etc. Altogether there is plenty of potential for future research based on AROS data and/or TRAT.

7.8 Conclusions

As an outcome of this thesis a specific analytical tool called TRAT was developed to extract knowledge from movement data provided by low cost observation system AROS that has been developed for tracking the riverboats in the Peruvian Amazon where most of the transportation is based on river networks. Utilizing TRAT and AROS it is possible to obtain relatively accurate information (assessed by visual exploration and comparing the results to in-situ measurements) about seasonal transportation characteristics at the study area (such as speed and average travel times between locations at different times of the year).

Results of this thesis suggest that navigation along the rivers had seasonal and directional variation and also the river morphology affected the movement patterns of the vessels in year 2012. On Iquitos-Pucallpa route, which is mostly meandering by river morphology, the downstream navigation was over 40% faster than upstream navigation during high water and intermediate, but during low water there was no difference between navigation directions. On Iquitos-Yurimaguas route, which is mostly anastomosing by river morphology, the downstream navigation was approximately 20% faster during the entire year. Seasonal variation was more evident on Iquitos-Pucallpa route where navigation was over 30% faster during high water compared to low water (when travelled downstream). On upstream direction the navigation was fastest during low water but seasonal differences were considerably lower compared to downstream navigation. On Iquitos-Yurimaguas route it seemed that there is no seasonal difference since the travel speeds were quite similar throughout the year.

Fitting simple regression model between average travel speed of the journeys and water levels of the river revealed that there seemed to be strong connection between travel speed and river height of Ucayali on Iquitos-Pucallpa route when travelled downstream. However on other cases there seemed to be no clear connection between travel speed and river height. Also connection between river sinuosity and travel speed was studied on Iquitos-Pucallpa route but there were difficulties to reduce the effect of populated places (i.e. locations where vessels stop during the journeys) on movement characteristics. Thus there were only few areas along the river network where the effect of sinuosity could be evaluated. On these areas the results suggest that high sinuosity indeed decelerates the navigation speed especially when travelled downstream. On areas where the river do not meander (i.e. low sinuosity), the travel speed seemed to be higher especially when travelling upstream which was slightly surprising.

Comparing the results with earlier study (Salonen et al. 2012a) implied that the results of this thesis seemed to be fairly accurate, however it is necessary to validate the results by doing cross-validations between data from different years observed with AROS. This is especially important since the river dynamics were slightly exceptional during the year 2012 (major floods). Spatio-temporal information such as mobility data from

AROS can be utilized in many areas ranging from decision making based on real-world information to e.g. more accurate analyses of accessibility which has become an increasingly important and useful analytical tool as means of *understanding the spatial relationships between our society and environment (e.g. transport and land use)* and as means of *visualising their changing geographical setting at a general level*.

ACKNOWLEDGEMENTS

I would like to thank all of those who have helped me to complete this thesis. I would like to thank Nordenskiöld-samfundet for the grant that made my fieldwork period possible at Peruvian Amazon. Especial thank to my fieldwork assistant Yully Rojas Reategui in Iquitos who helped and guided me, as well as translated my broken Spanish into more understandable form during the fieldworks: Yully, muchas gracias y abrazos. I am very grateful to my supervisors Dr. Tuuli Toivonen and Maria Salonen who have given me excellent guidance and ideas throughout the process of making this thesis, and Maria thank you for helping me to get settled during the fieldworks. Big thanks to Dr. Kimmo Vehkalahti and Dr. Reijo Sund for helping me to develop the analysis tool with Muste and giving me great ideas and support. I am deeply grateful to Dr. Otto Huisman for introducing me the concepts of time-geography and movement analysis and to the International Training Center (ITC), University of Twente for providing me the necessary geovisual analytics tools. I would also like to thank Professor Petri Pellikka, Dr. Mika Siljander, Dr. Janne Heiskanen and Mr. Pekka Hurskainen for their GIS teachings and coffee companion. I am also deeply grateful to my family, Jenni Lehtimäki and my friends for their valuable support. Semifinal: thanks for my bandmates Mikko Haapanen and Lauri Pihlajaniemi who helped me with the power of music to put aside these study things from time to time. Ultimate final: big thanks for Perttu Saarsalmi for his companion, advice, opinions and refreshing talks at the office.

REFERENCES

Abizaid, C. (2005). Geographical field note. An anthropogenic meander cutoff along the Ucayali river, Peruvian Amazon. *The Geographical Review* 95: 1, 122-135.

Accessibility Research Group (2013). Accessibility matters. 08.11.2013 <*https://blogs.helsinki.fi/accessibility*>

Agarwal, P., L. Guibas, H. Edelsbrunner, J. Erickson, M. Isard, S. Har-Peled, J. Hershberger, C. Jensen, L. Kavraki, P. Koehl, M. Lin, D. Manocha, D. Metaxas, B. Mirtich, D. Mount, S. Muthukrishnan, D. Pai, E. Sacks, J. Snoeyink, S. Suri & O. Wolfson (2002). Algorithmic issues in modeling motion. *ACM Computing Surveys* 34: 4, 550–572.

Ahas, R., A. Aasa, S. Silm, R. Aunap, H. Kalle & Ü. Mark (2007). Mobile Positioning in Space-Time Behaviour Studies: Social Positioning Method Experiments in Estonia. *Cartography and Geographic Information Science* 34: 4, 259-273.

Ahas, R., S. Silm, O. Järv, E. Saluveer & M. Tiru (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology* 17: 1, 3–27.

Andrienko, G., N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski & D. Thom (2013). Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. *Computing in Science & Engineering* 15: 3, 72-82.

Andrienko, N. & G. Andrienko (2005). *Exploratory Analysis of Spatial and Temporal Data*. 703p. Springer-Verlag, Berlin.

Andrienko, N., G. Andrienko, N. Pelekis & S. Spaccapietra. (2008). Basic Concepts of Movement Data. On a book: Giannotti, F. & D. Pedreschi (eds.) *Mobility, Data Mining and Privacy. Geographic Knowledge Discovery*. Springer-Verlag, Berlin.

Andrienko, N., G. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. Fabrikant, M. Jern, M-J. Kraak, H. Schumann & C. Tominski (2010). Space, time and visual analytics. *Internation Journaly of Geographical Information Science* 24: 10, 1577-1600.

Angelsen, A. & D. Kaimowitz (1999). Rethinking the Causes of Deforestation: Lessons from Economic Models. *The World Bank Research Observer* 14: 1, 73-98.

Bae Systems (2012). Bae Systems locates opportunity to replace GPS. 03.07.2012. *<http://www.baesystems.com/article/BAES_053641/bae-systems-locates-opportunity-to-replace-gps>*

Bauch, S. C., Amacher, G. S., & Merry, F. D. (2007). Cost of harvesting, transportation and milling in the Brazilian Amazon: estimation and policy implications. *Forest Policy and Economics* 9, 903-915.

BCRP (2009). Banco Central de Reserva del Perú. Encuentro Económico. Informe Económico y Social. Región Loreto. 03.07.2013. *<http://www.bcrp.gob.pe/docs/Proyeccion-Institucional/Encuentros-Regionales/2009/Loreto/Informe-Economico-Social/IES-Loreto.pdf>*

BCRP (2012). Banco Central de Reserva del Perú. Encuentro Económico. Informe Económico y Social. Región Ucayali. 04.07.2013. *<http://www.bcrp.gob.pe/docs/Proyeccion-Institucional/Encuentros-*Regionales*/2012/Ucayali/Informe-Economico-Social/IES-Ucayali.pdf* >

Bertolini, L. & F. le Clercq (2003). Urban development without more mobility by car? Lessons from Amsterdam, a multimodal urban region. *Environment and Planning A*. 35, 575–589.

Biodamaz (2004a). Marco teórico y metodológico para identificar unidades ambientales en la selva baja peruana. Documento Técnico N° 5. Serie IIAP-BIODAMAZ. Iquitos, Peru.

Biodamaz (2004b). Macrounidades ambientales en la Amazonía peruana con la énfasis en la selva baja: primera aproximación a manera de hipótesis de trabajo. Documento Técnico N° 13. Serie IIAP-BIODAMAZ. Iquitos, Peru.

Bickel, P., C. Chen, J. Kwon, J. Rice, E. van Zwet & P. Varaiya. (2007). Measuring Traffic. *Statistical Science* 22: 4, 581-597.

Chibnik, M. (1994). *Risky rivers: The economics and politics of floodplain farming in Amazonia*. Tucson: University of Arizona Press.

Chomitz, K., & Thomas, T. (2003). Determinants of land use in Amazonia: a finescale spatial analysis. *Journal of Agricultural Economics* 85: 4, 1016-1028.

Coomes, O.T, C. Abizaid & M. Lapointe (2009). Human Modification of a Large Meandering Amazonian River: Genesis, Ecological and Economic Consequences of The Masisea Cutoff on the Central Ucayali, Peru. *Ambio: A Journal of the Human Environment* 38: 3, 130-134.

Cressie, N. (1993). *Statistics for spatial data.* 928 p. John Wiley, New York.

Delafontaine, M., M. Versichele, T. Neutens & N. Van de Weghe. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography* 34, 659-668.

Demsar, U. & K. Virrantaus. (2010). Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science* 24: 10, 1527–1542.

Dodge, S., R. Weibel & A-K, Lautenschütz (2008). Towards a taxonomy of movement patterns. *Information Visualization* 7, 240-252.

Dodge, S., R. Weibel & E. Forootan (2009). Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems* 33, 419–434.

Ellis, E. & N. Ramankutty (2008). Putting people in the map: anthropogenic biomes of the world. *Frontiers in Ecology and the Environment 6: 439–447.*

Fayyad, U., G. Piatetsky-Shapiro & P. Smyth (1996). From data mining to knowledge discovery in databases. *AI Magazine* 17: 3, 37–54.

Fearnside, P. (2008). The Roles and Movements of Actors in the Deforestation of Brazilian Amazonia. *Ecology and Society* 13: 1, 23 p.

Forester, J., H. Im & P. Rathouz. (2009). Accounting for animal movement in estimation of resource selection functions: sampling and data analysis. *Ecology* 90: 12, 3554-3565.

Franklin, C. & P. Hane. (1992). An Introduction to Geographic Information Systems: Linking Maps to Databases [and] Maps for the Rest of Us: Affordable and Fun. *Database* 15:2, 17-22.

Gailey, G., B. Würsig & T. McDonald. (2007). Abundance, behavior, and movement patterns of western gray whales in relation to a 3-D seismic survey, Northeast Sakhalin Island, Russia. *Environmental Monitoring and Assessment* 134; 1-3, 75-91.

Geurs, K. & J. van Eck (2001). Accessibility measures: review and applications. Evaluation of accessibility impacts of land-use transportation scenarios, and related social and economic impact. *National Institute of Public Health and the Environment. RIVM and Urban Research Centre. RIVM Report 408505006*. Utrecht University. 265 p.

Geurs, K. & B. van Wee (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography* 12, 127-140.

Geist, H. & E. Lambin (2002). Proximate Causes and Underlying Driving Forces of Tropical Deforestation. *BioScience* 52: 2, 143-150.

Giannotti, F. & D. Pedreschi (2008). Mobility, Data Mining and Privacy: A Vision of Convergence. On a book: Giannotti, F. & D. Pedreschi (ed.) *Mobility, Data Mining and Privacy. Geographic Knowledge Discovery.* Springer-Verlag, Berlin.

Goovaerts, P. (1999). Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma* 89, 1-45.

GOREL (2006a). Mapa de la hidrografía principal y secundaria, departamento de Loreto. 22.03.2012. <*http://www.regionloreto.gob.pe/OATSIG/5.pdf*>

GOREL (2006b). Mapa de infraestructura vial del departamento de Loreto. 22.03.2012. <*http://www.regionloreto.gob.pe/OATSIG/8.pdf*>

GOREL (2006c). Mapa político del departamento de Loreto. 22.03.2012. <*http://www.regionloreto.gob.pe/OATSIG/11.pdf*>

Gould, P. *Spatial Diffusion.* Commission on College Geography. Resource paper No. 4. Association of American Geographers. Washington, D.C.

Gupta, A. (ed.) (2007). *Large Rivers. Geomorphology and Management.* 705 pp. John Wiley & Sons Ltd., West Sussex.

Han, J. & M. Kamber (2011). *Data Mining. Concepts and Techniques.* 3rd ed. 550 pp. Morgan Kauffman Publishers, San Francisco.

Hansen, W. (1959). How accessibility shapes land use. Journal of American Institute of Planners 25: 1, 73-76.

Hempel, C. (1965). Aspects of Scientific Explanation and other Essays in the Philosophy of Science. 505p. Free Press, New York.

Hodge, D. (1997). Accessibility-related issues. *Journal of Transport Geography.* 5: 1, 33-34.

Hoffman, A. (1975) *Climatic atlas of South America*. OMM, WMO, UNESCO Cartographia, Geneva.

Hornsby, K. & M. Egenhofer (2002). Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence.* 36, 177–194.

HRT (2013). HSL LIVE. 05.08.2013. <*http://transport.wspgroup.fi/hklkartta/*>

Hägerstrand, T. (1970). What about people in regional science? *Regional Science Association* 24: 1, 6-21.

IBM (2013). Big Data at the Speed of Business. 06.06.2013. <*http://www-01.ibm.com/software/data/bigdata/*>

INEI (2007). Instituto Nacional de Estadistica e Informatica. PERÚ: CENSOS NACIONALES 2007, XI DE POBLACION Y VI DE VIVIENDA. Sistema de Consulta de Principales Indicadores Demográficos, Sociales y Económicos. 03.07.2013. <*http://censos.inei.gob.pe/Censos2007/IndDem/*>

INEI (2008). Instituto Nacional de Estadistica e Informatica. Censos Nacionales 2007: XI de Población y VI de Vivienda. Perfil Sociodemográfico del Perú. 03.07.2013. <*http://censos.inei.gob.pe/Anexos/Libro.pdf*>

INEI (2009). Instituto Nacional de Estadistica e Informatica. PBI Departamental 2008. 03.07.2013. <*http://www1.inei.gob.pe/web/BoletinFlotante.asp?file=8618.pdf*>

INEI (2012). Instituto Nacional de Estadistica e Informatica. Perú: Estimaciones y Proyecciones de Población Total por Sexo de las Principales Ciudades, 2000-2015. Boletín Especial Nº 23. 03.07.2013. <*http://www.inei.gob.pe/biblioineipub/bancopub/Est/Lib1020/Libro.pdf*>

Ingram, D. (1971). The concept of Accessibility: A search for an operational form. *Regional Studies* 5, 101-107.

International Energy Agency (2009).Transport, energy and CO_2. Moving towards sustainability. *IEA/OECD.*

ISO 8601 (2004). International Standard. Data elements and interchange formats – Information interchange – Representation of dates and times. Third Edition. 08.05.2013 <*http://dotat.at/tmp/ISO_8601-2004_E.pdf* >

ITU (2002). International Telecommunication Union. Recommendation ITU-R TF.460-6. Standard-frequency and time-signal emissions. 07.05.2013 <*http://www.itu.int/dms_pubrec/itu-r/rec/tf/R-REC-TF.460-6-200202-I!!PDF-E.pdf*>

Iwase, S. & H. Saito (2002). Tracking soccer player using multiple views. *IAPR Workshop on Machine Vision Applications (MVA Proceedings),* 102–105.

Josse, C., G. Navarro, F. Encarnación, A. Tovar, P. Comer, W. Ferreira, F. Rodríguez, J. Saito, J. Sanjurjo, J. Dyson, E. Rubin de Celis, R. Zárate, J.Chang, M. Ahuite, C. Vargas, F. Paredes, W. Castro, J. Maco & F. Reátegui (2007). *Ecological systems of the Amazon Basin of Peru and Bolivia. Classification and mapping.* NatureServe. Arlington, Virginia, USA.

Kalliola, R., M. Puhakka & W. Danjoy (eds.) (1993). *Amazonia Peruana vegetacion humeda tropical en el llano subandino*. 265 pp. PAUT/ONERN, Jyväskylä.

Keim, D., J. Kohlhammer, F. Mansmann, T. May & F. Wanner (2010). Visual Analytics. On a book: Keim, D., J. Kohlhammer, G. Ellis & F. Mansmann (eds.) (2010). *Mastering the information age. Solving problems with visual analytics.* 7-18 pp. Eurographics Association, Goslar.

Killeen, T. (2007). A Perfect Storm in the Amazon Wilderness: Development and Conservation in the Context of the Initiative for the Integration of the Regional Infrastructure of South America (IIRSA). *Advances in Applied Biodiversity Science* 7, 102 p.

Knowles, R., J. Shaw & I. Docherty (2008). *Transport Geographies. Mobilities, Flows and Spaces.* 1st Edition. 293 p. Blackwell Publishing, Oxford.

Kraak, M-J. (2011). Is there a need for Neo-Cartography? *Cartography and Geographic Information Science* 38: 2, 73-78.

Krause, B. & C. von Altrock. (1996). Intelligent Highway by Fuzzy Logic: Congestion Detection and Traffic Control on Multi-Lane Roads with Variable Road Signs. *Proceedings of the Fifth IEEE Internation Conference on Fuzzy Systems* 1-3, 1832-1837.

Krizek, K., A. Forysth & C. Schively Slotterback (2009). Is there a role for evidence-based practice in urban planning and policy? *Planning Theory & Practice* 10: 4, 459-478.

Kvist, L. & G. Nebel (2001). A review of Peruvian flood plain forests: ecosystems, inhabitants and resource use. *Forest Ecology and Management* 150, 3-26.

Kwan, M-P. (2000). Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C* 8, 185-203.

Lahtinen, J., M. Salonen & T. Toivonen (2013). Facility allocation strategies and the sustainability of service delivery: Modelling library patronage patterns and their related CO2-emissions. *Applied Geography.* 44, 43-52.

Laube, P. and Imfeld, S. (2002). Analyzing relative motion within groups of trackable moving point objects. In Egenhofer, M. & D. Mark (eds.), *Geographic Information Science* 2478, 132–144. Springer. Berlin-Heidelberg.

Laube, P., S. Imfeld & R. Weibel. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science* 19: 6, 639-668.

Laube, P. & R. Purves (2006). An approach to evaluating motion pattern detection techniques in spatio-temporal data. *Computers, Environment and Urban Systems* 30, 347-374.

Lesage, V., M. Hammill & K. Kovacs. (2004). Long-distance movements of harbour seals (Phoca vitulina) from a seasonally ice-covered area, the St. Lawrence River estuary, Canada. *Canadian Journal of Zoology* 82: 7, 1070-1081.

Li, Q., T. Zhang, H. Wang & Z. Zeng (2011). Dynamic accessibility mapping using floating car data: A network-constrained density estimation approach. *Journal of Transport Geography* 19: 3, 379-393.

Live Ships Map (2013). MarineTraffic.com. 06.08.2013. *<http://www.marinetraffic.com/ais/>*

Marengo J. (2005). Characteristics and spatio-temporal variability of the Amazon River Basin Water Budget. *Climate Dynamics* 24, 11–22.

Marketos, G., E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaetá & Y. Theodoridis (2008). Building Real-world Trajectory Warehouses. *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access. pp. 8-15.*

Mennis, J. & D. Guo (2009). Spatial data mining and geographic knowledge discovery – An introduction. *Computers, Environment and Urban Systems* 33, 403-408.

Miller, H. & Han, J. (2001). Geographic data mining and knowledge discovery: an overview. On a book Miller, H. & J. Han (Eds.), *Geographic data mining and knowledge discovery*. pp. 3–32. Taylor and Francis, London.

Miller, H. & Han, J. (2009). Geographic data mining and knowledge discovery: An overview. On a book Miller, H. & J. Han (Eds.), *Geographic data mining and knowledge discovery 2nd Edition.* pp. 1–26. CRC Press, Taylor and Francis Group, London.

Miller, H. & S. Bridwell (2009). A Field-based Theory for Time Geography. *Annals of the Association of American Geographers* 99: 1, 49-75.

Moore, A., P. Whigham, A. Holt, C. Altridge & K. Hodge. (2003). A Time Geography Approach to the Visualisation of Sport. In *Proceedings of the Seventh International Conference on Geocomputation*. 13p.

Moran, P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika* 37: 1, 17-23.

MPCP (2010). Municipalidad Provincial de Coronel Portillo. Sub-Gerencia de Catastro MPCP 2010. 04.07.2013. <*http://201.230.96.134/pucallpa/mapa.phtml*>

Mustonen, S. (1992). SURVO. An integrated Environment for Statistical Computing and Related Areas. *Survo Systems Ltd.* 484 p.

Nathan, R., W. Getz, E. Revilla, M. Holyoak, R. Kadmon, D. Saltz & P. Smouse (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences of the United States of America* 105: 49, 19052-19059.

Open Knowledge Foundation (2013). The Open Knowledge Foundation. Empowering through Open Knowledge. 08.11.2013. <*http://okfn.org*>

Pelekis, N., G. Andrienko, N. Andrienko, I. Kopanakis, G. Marketos & Y. Theodoridis. (2012). Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems* 38, 343-391.

Peres, C. & I. Lake (2003). Extent of Nontimber Resource Extraction in Tropical Forests: Accessibility to Game Vertebrates by Hunters in the Amazon Basin. *Conservation Biology* 17: 2, 521-535.

Peres, C. & J. Terborgh (1995). Amazonian Nature Reserves: An Analysis of the Defensibility Status of Existing Conservation Units and Design Criteria for the Future. *Conservation Biology* 9: 1, 34-46.

Puhakka, M., R. Kalliola, M. Rajasilta & J. Salo (1992). River types, site evolution and successional vegetation patterns in Peruvian Amazonia. *Journal of Biogeography* 19: 6, 651-665.

R Documentation (2013). Date-time Conversion Functions. 08.05.2013. <*http://stat.ethz.ch/R-manual/R-devel/library/base/html/as.POSIXlt.html* >

Raptopoulou, K., A. Papadopoulos & Y. Manolopoulos. (2003). Fast nearestneighbor query processing in moving-point databases. *GeoInformatica* 7, 113–137.

Reynolds, D. & J. Riley. (2002). Remote-sensing, telemetric and computer-based technologies for investigating insect movement: a survey of existing and potential techniques. *Computers and Electronics in Agriculture* 2-3, 271-307.

Rinzivillo, S., F. Turini, V. Bogorny, C. Körner, B. Kuijpers & M. May (2008). Knowledge Discovery from Geographical Data. On a book: Giannotti, F. & D. Pedreschi (ed.) *Mobility, Data Mining and Privacy. Geographic Knowledge Discovery.* Springer-Verlag, Berlin.

Rodriguez Achung, M. (1994). Crecimiento urbano de Iquitos: condicionamientos estructurales en la decaca del ´70 y sus perspectivas. *Documento técnico No 08*. 109 pp. IIAP, Iquitos.

Salonen, M., T. Toivonen, J.M. Cohalan & O.T. Coomes (2012a). Critical distances: Comparing measures of spatial accessibility in the riverine landscapes of Peruvian Amazonia. *Applied Geography* 32: 2, 501-513.

Salonen, M., Toivonen, T. & Vaattovaara, M. (2012b). Arkiliikkumisen vaihtoehdoista monikeskuksistuvassa metropolissa: Kaksi näkökulmaa palvelujen saavutettavuuteen pääkaupunkiseudulla. *Yhdyskuntasuunnittelu* 3/2012, 8-27

Salonen, M., E. Maeda & T. Toivonen (2013). Evaluating the Impact of Distance Measures on Deforestation Simulations in the Fluvial Landscapes of Amazonia. *AMBIO* DOI 10.1007/s13280-013-0463-x.

Salonen, M. & T. Toivonen (2013). Modeling travel time in urban networks: comparable measures for private car and public transport. Journal of Transport Geography. 31, 143-153.

SEHINAV (2013). Servicio de Hidrografia y Navegación de la Amazonia. Marina de Guerra del Perú. 07.05.2013. *https://www.dhn.mil.pe/shna/index2.asp*

Shekhar, S., M. Evans, J. Kang & P. Mohan (2011). Identifying patterns in spatial information: a survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 1: 3, 193-214.

Sioli, H. (1984). The Amazon and its main affluents: Hydrography morphology of the river courses and river types. In Sioli, H. (ed.) *The Amazon: Limnology and landscape ecology of a mighty tropical river and its basin*. 763 pp. Junk publishers, Boston.

Spaccapietra, S., C. Parent, M. Damiani, J. de Macedo, F. Porto & C. Vangenot (2008). A conceptual view on trajectories. *Data and Knowledge Engineering* 65: 1, 126-146.

SPOT (2007). User's guide. SPOT. The World's First Satellite Messenger. 08.05.2013 <h*ttp://www.gmpcs-us.com/uploads/GSP-SPOT/spot_user_guide.pdf*>

SPOT (2013). The SPOT Personal Tracker. 29.07.2013 *http://www.findmespot.eu/en/index.php?cid=101*

Sund, R. (2011). Muste – the R implementation of Survo. *Yearbook of Finnish Statistical Society*. 133-146.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46: 2, 234-240.

Toivonen, T., S. Mäki & R. Kalliola (2007). The riverscape of Western Amazonia - a quantitative approach to the fluvial biogeography of the region. *Journal of Biogeography* 34: 8, 1374-1387.

Tribby, C. & P. Zandbergen (2012). High-resolution spatio-temporal modeling of public transit accessibility. *Applied Geography* 34: 4, 345-355.

uDig (2011). uDig. User-friendly Desktop Internet GIS. 15.03.2012. *http://udig.refractions.net/*

UN-GGIM (2012). *Future trends in geospatial information management: the five to ten year vision.* Draft paper. 39 pp. New York.

Verburg, H., K. Overmars & N. Witte (2004). Accessibility and land-use patterns at the forest fringe in the northeastern part of the Philippines. *The Geographical Journal* 170: 3, 238-255.

Verburg, P., E. Ellis & A. Letourneau (2011). A global assessment of market accessiblity and market influence for global environmental change studies. *Environmental Research Letters* 6. 12p.

Versichele, M., T. Neutens, M. Delafontaine & N. Van de Weghe. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography* 32, 208-220.

Vickerman, R. (1995). Location, accessibility and region development: the appraisal of trans-European networks. *Transport Policy* 2: 4, 225-234.

Vickerman, R., K. Spiekermann & M. Wegener (1999). Accessibility and Economic Development in Europe. *Regional Studies* 33: 1, 1-15.

Vílchez Vela, P. (2012). *Epoca del Caucho: Retratos del Horror.* 150p. Tierra Nueva, Iquitos.

VR. (2013). Live train map. 05.08.2013. <*http://www.vr.fi/en/index/aikataulut/livetrainmap.html*>

Wilkins, D. & M. desJardins (2001). A call for knowledge-based planning. *AI Magazine* 22: 1, 99-115.

Willems, N., H. Van de Wetering & J. Wijk. (2009). Visualization of vessel movements. *Computer Graphics Forum* 28: 3, 959-966.

Wilson, E. (1901). *Vector Analysis. A text-book for the use of students of mathematics and physics.* University press. John Wilson and Son. Cambridge, USA.

The World Bank (2013). GDP per capita (current US$). 04.07.2013. <*http://data.worldbank.org/indicator/NY.GDP.PCAP.CD*>

Yang, D., R. McCollum, & W. McCarthy (2009). Leighty Meeting an 80% reduction in greenhouse gas emissions from transportation by 2050: a case study in California. *Transportation Research Part D: Transport and Environment.* 14: 3, 147–15.

YLE (2013a). Mobiilidata kuusinkertaistuu joka vuosi. 03.09.2013. <*http://yle.fi/uutiset/mobiilidata_kuusinkertaistuu_joka_vuosi/6789262* >

YLE (2013b). Suomalaiset ovat etätyön e-nomadeja. 18.09.2013. <*http://yle.fi/uutiset/suomalaiset_ovat_etatyon_e-nomadeja/6835683*>

Zeng, N. (1999). Seasonal cycle and interannual variability in the Amazon hydrological cycle. *Journal of Geophysical Research* 104, 9097–9106.

Zhao, J., P. Forer & A. Harvey (2008). Activities, ringmaps and geovisualization of large human movement fields. *Information Visualization* 7, 198-209.

APPENDICES

APPENDIX I

R-code for joining observation to the nearest reference data point (adapted from: *https://stat.ethz.ch/pipermail/r-help/2008-September/173983.html*).

```
spatial.join <- function(x, y, lat, lon, obs, latRef, lonRef) {

   ## construct data frame d in which d[i,] contains information
   ## associated with the closest point in y to x[i,]

   xpos <- as.matrix(x[,c(lat, lon, obs)])
   xposl <- lapply(seq.int(nrow(x)), function(i) xpos[i,])
   ypos <- t(as.matrix(y[,c(latRef, lonRef)]))
   yinfo <- y[, colnames(y)]



   get.match.and.dist <- function(point) {

      sqdists <- colSums((point - ypos)^2)
      ind <- which.min(sqdists)
      c(ind, sqrt(sqdists[ind]))

   }

   match <- sapply(xposl, get.match.and.dist)

   cbind(xpos, mindist=match[2,], yinfo[match[1,],])

}

join <- spatial.join(boatpoint, refpoint, 'lat', 'lon', 'obs', 'latRef', 'lonRef')
```

APPENDIX II

Algorithm for journey detection (Survo coding, Mustonen 1992):

```
VAR journeyID:2=if(ORDER=1)then(1)else(RID2) TO BoatID1

  RID2=if(TSTOP<36)then(RID4)else(RID3)
  RID3=if(TSTOP[-1]>=36)then(journeyID[-1])else(journeyID[-1]+1)
  RID4=if(Dhour<36)then(RID7)else(RID5)
  RID5=if(Dhour[-1]>=36)then(RID6)else(journeyID[-1]+1)
  RID6=if(DDistance>=500)then(journeyID[-1]+1)else(journeyID[-1])
  RID7=if(RDir[0]=RDir[-1])then(journeyID[-1])else(RID8)
```

RID2 = Observation of stationary time (threshold = 36 hours proved to be optimal) to determine individual journey.

RID3 = Maintaining journeyID if stationary still continues

RID4 = Determing journeyID based on temporal gab between observations (threshold = 36h)

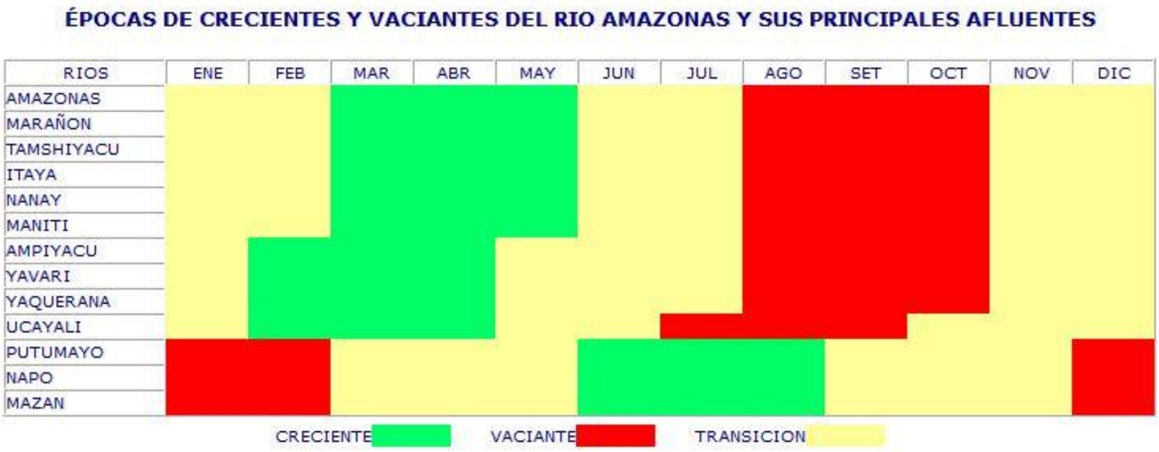
RID5 = Maintaining journeyID if tracking was “accidentally” put on for only 1 observation

RID6 = Determing journeyID based on distance between consecutive observations (threshold = 500 kilometers)

RID7 = Determing journeyID based on principal navigation direction

APPENDIX III

The stages of the Peruvian Amazon rivers (SEHINAV 2013).



ÉPOCAS DE CRECIENTES Y VACIANTES DEL RIO AMAZONAS Y SUS PRINCIPALES AFLUENTES

RIOS	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SET	OCT	NOV	DIC
AMAZONAS	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟥	🟥	🟥	🟨	🟨
MARAÑON	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟥	🟥	🟥	🟨	🟨
TAMSHIYACU	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟥	🟥	🟥	🟨	🟨
ITAYA	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟥	🟥	🟥	🟨	🟨
NANAY	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟥	🟥	🟥	🟨	🟨
MANITI	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟥	🟥	🟥	🟨	🟨
AMPIYACU	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟥	🟥	🟥	🟨	🟨
YAVARI	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟥	🟥	🟥	🟨	🟨
YAQUERANA	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟥	🟥	🟥	🟨	🟨
UCAYALI	🟨	🟩	🟩	🟩	🟨	🟨	🟥	🟥	🟥	🟨	🟨	🟨
PUTUMAYO	🟥	🟥	🟨	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟥
NAPO	🟥	🟥	🟨	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟥
MAZAN	🟥	🟥	🟨	🟨	🟨	🟩	🟩	🟩	🟨	🟨	🟨	🟥

CRECIENTE 🟩 VACIANTE 🟥 TRANSICION 🟨