University of
Zurich UZH

Suitability of Travel Costs as a Geographic Prior on Language Contact

GEO 511 Master's Thesis

Author
Michael Zurmühle
16-913-816

Supervised by
Dr. Nico Neureiter
Dr. Peter Ranacher

Faculty representative
Prof. Dr. Robert Weibel

29.09.2023
Department of Geography, University of Zurich

Abstract

In the field of phylogenetics, researchers from a diverse set disciplines are involved. Originally, phylogenetics was first developed in evolutionary biology, in order to produce evolutionary trees to model evolutionary processes in biology. Nowadays, phylogenetic models are used to model all kinds of processes underlaying evolutionary principles, such as linguistics. Modelling geolocated phylogenetic trees with evermore detail and variety is of great interest to the phylogenetic research community, thereby having more tools to their disposal. This master's thesis delves into the realm of phylogeography, where we focus on horizontal transfer of loanwords in particularly, thereby evaluating contact events between languages. Phylogenetic models, which incorporate horizontal transfer, are still quite a novelty. contacTrees is the first addon for the BEAST2 software, which is making use of horizontal transfer (contact events between languages) as a phenomenon. By analysing Tobler's hiking function based travel costs of contact events produced by contacTrees, we have shown that those contact events are significantly shorter than comparable alternatives (non-contact events). This indicates that languages that are close to each other are more likely to be in contact. In this thesis, we lay out a path to integrate Tobler's hiking function into a geo-prior for contacTrees. By evaluating the results of Tobler's hiking function based travel costs of such contact events, we have found a statistically significant signal to differentiate between valid linguistical contact and model output noise. Compared to geodetic-distances, Tobler's hiking function incorporates terrain into the travel cost estimation, resulting in a more realistic evaluation. The main obstacle to implement Tobler's hiking function into a geo-prior are the high computational costs. For the geo-prior to be feasible it has to calculate tens of thousands of terrain-dependent path calculations in a concise time frame. In order to achieve the efficiency levels needed, we present the Cost Surface Network (CSN) as a solution. The CSN produces accurate predictions of Tobler's hiking function based travel cost, but with much higher efficiency. The CSN approach shows great potential not only in simulating a Digital Terrain Models (DTMs) for Tobler's hiking function to run on, but shows also potential for other cost topographies. Having solved both, the efficiency-problem (CSN), as well as the feasibility-problem (statistically significant signal), we propose to implement a Tobler's hiking function based geo-prior into contacTrees.

Contents

Table of Figures

1. Introduction

The study of linguistics has long fascinated researchers. Within the field of phylogenetic linguistics parallels with the principles of evolutionary biology can easily be drawn, offering insights into the historical relationships between languages. Just as phylogenetics has advanced our understanding of biological species, phylogenetic modelling of languages has shown the intricate web of linguistic diversity and the historical processes that have shaped it. In the context of linguistic evolution, factoring in horizontal transfer (effect of contact between languages) is an important factor to evaluate. Just as genes can transfer horizontally between species, languages can influence each other through contact events, were loanwords are exchanged. Horizontal transfer most often occurs in closely related clades. Otherwise, contact events also occur over close geographical proximity. Horizontal transfer is an important process, because it furthers model accuracy, especially in regard of the frequency of change (clock rate) within the model.

Various phylogeographic research on the Indo-European language family has been done over the last two decades (Heggarty, 2014; Bouckaert et al., 2012; Forster et al. 2003; Gray et al., 2003). Mainly focusing on evaluation of the two prevalent hypothesis regarding the origin of the Indo-European language family. On the one hand we have the steppe-hypothesis, which sets the point of origin for the language family in the Pontic-Caspian Steppe and is based on a horse-based pastoralism lifestyle, set around 6500 years BP. On the other hand the Anatolian-hypothesis predicts the point of origin in the Anatolian highlands and is based on an agricultural lifestyle set in a larger timeframe of around 9500 to 8500 years BP. Findings of the phylogenetics community show a clear support for the "Anatolian"-Hypothesis (Bouckaert et al., 2012), which seems to be robust to a certain degree, because newer findings, incorporating ancient DNA, show a similar mean root age (Heggarty et al., 2023). Another major focus of academic debate are migration patterns, which is closely related to the aforementioned origin-question (Ranacher et al., 2021; Koile et al., 2022; Neureiter et al., 2021). Furthermore, horizontal transfer of loanwords between languages gains traction in the scientific community and is becoming its own niche of inquiry, which has produced interesting findings regarding clock-rate models. The higher convergence efficiency due to the possibility of horizontal transfer allows for lower clock-rates for convergence to occur, which shortens the root height considerably (Neureiter et al., 2022; Ranacher et al., 2021).

The phylogeographic research community further discusses the use of landscape aware models, on which the conclusion is that it does not make big difference if applied on terrain. If the probability of crossing a terrain hurtle is greater than zero, than the possibility that some communities crossed it in a near infinite amount of time would be close to 100% (Bouckaert et al., 2018). Therefore, factoring in terrain is only useful in a restricted set of use cases. One of this use cases, were we think, the implementation of terrain based factors would be highly beneficial, would be in the context of modelling horizontal transfer (contact) between languages. The modelling process of contact would greatly benefit if the terrain between each possible language pair is evaluated. Contact has to be sustained over long periods of time in a semi-permanent way in order for loanwords to occur (Hock & Joseph, 2009; Daulton, 2019). Over this prolonged time of contact, terrain based travel costs fall on permanently and not just once. Therefore, evaluating terrain based travel costs in order to model the possibility for contact between each combination of languages has the potential to improve the models of such contact events.

1.1. Introduction to Bayesian Phylogenetics, Phylogeography and BEAST2

Bayesian phylogenetics is a powerful statistical method used to infer the evolutionary relationships among different taxa. Examples are languages, species, or other entities ruled by evolutionary processes. This approach employs the principles of Bayesian statistics to estimate the most probable phylogenetic tree given the observed sequence data and a priori information. At its core, the Bayesian approach updates our belief in a hypothesis (the posterior) based on new evidence (the likelihood) and our initial assumptions (the prior). This is the basis for the Bayesian inference model used in all phylogenetic analysis.

In order for Bayesian phylogenetics to be utilised, there are several programs for that purpose, such as MrBayes (Ronquist et al., 2003; Huelsenbeck et al., 2001), MEGA (Koichiro et al., 2021) or BEAST2. In the scope of this thesis we use BEAST2 (Bouckaert et al., 2019). This software tool is widely used in the scientific community for Bayesian phylogenetic analysis. It stands for Bayesian Evolutionary Analysis by Sampling Trees 2, and it provides a user-friendly platform to perform Bayesian inference of phylogenetic trees. BEAST2 is a tool that estimates rooted, time-measured phylogenies and explores the evolutionary relationships between languages without relying solely on a single tree topology. Furthermore, the incorporation of geographical inference, or phylogeography, plays a pivotal role in the investigation of this thesis, allowing for the correlation of linguistic relationships with their spatial distribution. BEAST2 stands out due to its flexibility and capacity to handle complex evolutionary models, large datasets and its open source character. The most important community addon used in this thesis is contacTrees written by Nico Neureiter (2022), which provides us with the key ability to model contact events between languages. In BEAST2, the output of a phylogenetic analysis is typically stored in a BEAST2 tree file. This file uses the Newick format, a widely adopted standard for representing evolutionary trees. The Newick format represents tree structures using nested parentheses, where each set of parentheses contains the descendants of a particular node or branch. An extended version of the Newick-format is used by the contacTrees addon in order to store contact edges and nodes into the tree structure.

1.2. Tobler’s hiking function (THF) and Cost Surface Networks

Tobler's hiking function is a mathematical model that widely used in spatial analysis. The functions names was given by its inventor, Waldo Tobler (1993). It serves as a means to estimate the speeds or paces at which a person can travel across varying terrains, taking into account the slope of the terrain. Thus, it is a measure on how costly traversing a certain terrain is. Tobler’s aim was to capture the intuitive notion that walking on level ground is faster and requires less effort than traversing steep slopes. The function is formulated as an exponential relationship, with the angle of slope being the primary determinant of travel speed. Specifically, it states that as the slope angle increases, the travel speed decreases exponentially. This reflects on the reality that people slow down when they climb steep hills or descend steep slopes. This function allows us to model movement through a topography, and can easily be modified to model different modes of transportation, climate factors by multiplying the output with said factor accordingly (Pingel, 2010).

Enhancing the calculation efficiency of THF-based travel costs is important due to the high calculation costs associated with THF, this thesis introduces an novel approach that integrates THF into a spatial network. Thereby simulating the cost surface (e.g. DTM) by aggregating the raster cells of the cost surface into network nodes and by using THF initialise the networks edges. This Cost Surface Network (CSN) can calculate the shortest (least cost) paths by using Dijkstra’s algorithm (Dijkstra, 1959) as it would be ordinarily performed on a DTM. By calculating shortest paths (travel costs) in a network comprised of a condensed version of the initial cost surface and by correcting for the systematic uncertainty by a linear regression model, CSNs promise high efficiency levels, while retaining their accuracy.

1.3. Research Gap and Research Questions

This thesis delves into phylogenetic linguistics, but viewed through the lens of geography and the methodological toolset provided by Geographical Information Science/Systems (GIS), combining the power of phylogenetic modelling and GIS. Our goal is to decipher the historical evolution of the Indo-European language tree and also analyse travel costs over geographical space in order to show that contact events are significantly shorter than comparable alternatives (non-contact events). But the main focus of this scientific inquiry lies clearly on methodological aspects. The main point of interest lies in the feasibility evaluation of implementing THF into a geo-prior for Bayesian inference modelling To tackle this challenge we have to solve two problems, the feasibility-problem and the efficiency-problem, whereas the feasibility-question (research question I) also addresses our goal to evaluate contact events in a much broader sense. From the geo-prior's perspective, the first problem (efficiency-problem) is as follows, we need to know if a geo-prior based on THF produces usable results. For this purpose, we evaluate significance levels between the travel costs of contact events produced by the model and all the other travel costs possible at the same tree height (at the same age of the contact event). This can be condensed into our first research question:

Research Question I

"Are the spatial dynamics during contact events significantly different from non-contact events within the same temporal cross section of the language tree?"

The second problem is the efficiency-problem. To use THF as a geo-prior tens of thousands of calculations on a Tobler's cost surface has to be made. In order to successfully raise the efficiency level, different efficiency measures are discussed. Aiming to make the implementation of THF into a geo-prior and the calculation of all samples needed a realistic undertaking. The novel CSN approach is devised for this purpose. Validating this approach is formulated in the second research question:

Research Question II

"Can the CSN approach address the efficiency-problems posed by travel cost calculations via Tobler's hiking function (THF)?"

Further inspiration for enhancing CSN performance can be drawn from contacTrees's use of cognate classes and the utilisation of patterns in lexical borrowing. Cognate classes, in the context used by the contacTrees addon, represent linguistic traits categorized by shared meanings or concepts across different languages. These classes are defined by the presence or absence of cognates (words with a common etymology) for specific meanings (Neureiter et al., 2022). Lexical borrowing is a linguistic phenomenon that occurs when languages or dialects come into contact with one another. This contact is a natural consequence of linguistic evolution, as languages rarely exist in isolation (Hock & Joseph, 2009; Chambers et al., 2004). One prominent outcome of such contact is the adoption of individual words (cognate classes) from one language or dialect into another. This process is shaped by the dynamic between the influencer language and the influenced language, often implying a borrowing from a dominant or prestigious source. Sometimes, borrowing occurs out of necessity, sometimes purely out of prestige considerations (Hock & Joseph, 2009). Prestige relations between languages can profoundly impact their lexical histories and could probably be tracked by contacTrees in form of a influence vector. Which would streamline the whole CSN approach greatly.

2. Method and Data

In this section, we discuss the data and the rich set of tools used in this thesis. The main point of interest lies in the feasibility and efficiency evaluation of implementing THF into a geo-prior for Bayesian inference modelling, thereby addressing the stated feasibility-problem and efficiency-problem. For this purpose, we describe the data used to produce these results. Furthermore, we need ways to interpret BEAST2 data with the necessary addons. Furthermore, a way to calculate the shortest paths between language locations using THF in a more efficient setting. This efficiency boost is needed for its intended purpose as geo-prior for BEAST2.

2.1. Data

To validate the set research questions the data used within the scope of this thesis must be representative but also regionally contained. For that reason the ROI is set in the greater Europe region, including northern Africa and a big part of the Eurasian step. The data contains phylogenetic data from the Indo-European language family. The reasons for this focus on Europe are twofold. First, the locations of historical language are relatively well known and second, the linguistical history of the Indo-European language family is well sourced, therefore providing this thesis with a lot of ground truth regarding contact events between languages. To compare contact events and check their geographical validity, THF is utilized using a Digital Terrain Model (DTM) as cost topography.

2.1.1. Data for BEAST2

In order to produce the necessary phylogenetic trees including spatial positions and contact events the input XML-file for BEAST2 contains the IELex-taxonomic dataset (Indo European languages) and the spatial position information from Glottolg. In order for BEAST2 to access this data, it is necessary for the data to be encoded in the XML format.

2.1.1.1. Data Subset of IELex

The data used for phylogenetic inferencing stems from the IELex dataset. It is a subset comprised of data of 1419 cognates segregated into 206 meaning classes (concepts) across 37 languages in total. This dataset is the base for the analytics performed by BEAST2 and the contacTrees-addon.

Source: The dataset was produced by Michael Dunn and Tiago Tresoldi positioned now at Max Plank institute for Evolutionary Anthropology, Leipzig (Dunn et al. 2021).

2.1.1.2. Geographic Locations of Languages

This dataset provides geographic locations of modern and classical languages, facilitating the spatial analysis of language distribution. Glottolog is a work in progress and will be updated continuously.

Source: Glottolog is an initiative of the Max Planck Institute for Evolutionary Anthropology, Leipzig (Nordhoff & Hammarström, 2011).

2.1.1.3. File Type

The input file is of the type XML. BEAST2 only accepts input of that format. The XML-file incorporates all the data described above. This allows us to handle all necessary data used for the phylogenetic model in a concise package.

2.1.2. Data for CSN

The data needed for the Cost Surface Network (CSN) consists of the DTM including Europe's surface elevation with high precision. With a resolution of 15 arc-seconds, this DTM should be of a good enough resolution to model the cost topology of THF form.

2.1.2.1. Digital Terrain Model (DTM) with 15 Arc-Second Resolution

The GEBCO dataset is a DTM with a resolution of 0.004167 degrees (15 arc-seconds), which would translate to a metric resolution of 200 - 400 meters in our area of interest (Europe).

Source: GEBCO (2022) - The General Bathymetric Chart of the Oceans is a publicly available global bathymetric dataset.

2.2. Model setup

The XML file utilized within the scope of this thesis is the basic input setup for BEAST2, the Bayesian phylogenetic analysis software used (Detailed summary of the XML-file in the attachment section). The analysis involves 37 alignments of linguistic sequences and their position in space (language and language position), in addition 206 cognate concepts (such as: “bird”, ”laugh” and “wind”) are defined and incorporated. It also incorporates various non-state parameters, MCMC settings, Bayesian model specifications, likelihoods, operators, and loggers. Two specific components of interest are the GeoSphere and the contacTrees-addon, each serving a unique role in the geographical analysis of contact events between languages.

2.2.1. GeoSphere and contacTrees Addon Setup

The GeoSphere-addon is designed to capture the spatial distribution and movement patterns of taxa through simulated random walks (Bouckaert, 2016). Each language is represented by a point location which moves in a random walk over the surface of a sphere (earth). The random walks are constrained to start in the place of present day locations of the languages and end in the location of a common ancestor language by converging on one point. This is similar to the way languages converge towards a common ancestor linguistically. The addon includes the following main parameters:

1. **Geographical Clock Rate:** This parameter represents the overall rate of spatial change (speed) using random walks, it shows the average speed at which languages move/disperse in space. The initial value is set to 0.1. A higher rate would indicate more rapid spatial changes.
2. **Standard Deviation:** The standard deviation parameter controls the variation in spatial rates amongst different taxa. It ranges between 0 and 2, with an initial value of 0.1.

The GeoSphere-addon aims to provide insights into how the languages are distributed geographically and how their spatial positions change over time. All geographic locations of ancestral languages, as shown in the results, are inferred based on this model. contacTrees Addon Setup:

The contacTrees-addon is the other key component to evaluate contact between languages in a geographical manner. The addon introduces contact edges (also referred to as “conversions”) conversions are the connections between different taxa in the phylogenetic tree. This results in the generation of contact edges between separated clades of the tree. The model is doing this with the help of concepts represented by cognates. The contact edges can explain similarities (shared cognates) between languages that are only distantly related, which makes the model more flexible than classical tree models (Neureiter et al, 2022). It includes the following main parameters:

1. **Expected conversions (expected number of contact events):** The number of contact events (horizonal transfer) is controlled by this parameter. Setting this value to 0 would result in a classical tree model, but If the value is set to high, the model would produce an excessive

number of contact edges. Therefore, the parameter is set quite low to prevent contact edges produced out of statistical noise.

2. **Conversion Rate:** The conversion rate parameter is set to a linear growth model that is integrated into the tree structure (Newick network). It provides a way to incorporate the likelihood of taxa changing their meaning over time.
3. **Movement Parameter (pMove):** The movement parameter can also be understood as a borrowing probability. It controls the frequency that a meaning class (cognate class) moves in the phylogenetic tree. It is set to a range between 0 and 0.4.

The ContacTree-addon allows for dynamic shifts in the meaning of cognates during evolution. It models the conversion of taxa from different clades of the evolutionary tree. This model generates all contact data shown in the results.

2.2.2. Branch Rate Model

The branch rate model is set up as a relaxed log-normal clock with frozen branches. This means that the evolutionary rates amongst branches of the phylogenetic tree are allowed to vary, following a log-normal distribution. However, some branches are “frozen” in place (modern latin), meaning their rates are fixed and not allowed to vary during the analysis (Neureiter et al, 2022). The model includes three clock rate settings: slow, medium, and fast. The initial values for each setting (2*10^-5 for slow, 5*10^-5 for medium, and 8*10^-5 for fast) are set very low, at the same time the upper limit of 10 is set reasonably high, allowing the model to quickly diverge form the starting conditions.

2.2.3. Substitution Model (Binary Covarion)

The substitution model used in this analysis is the binary covarion model, which allows for site-specific rate variation. The lower bound of the variation is set to 0.0 and the upper bound to 1.0. The binary covarion model assumes that sites in the sequence alignment can be in one of two states (0 or 1) with equal probabilities (0.5 each). The frequencies parameter specifies the base frequencies for each state, with both states having equal probabilities (0.5 each).

2.2.4. Model Priors

The model priors are essential components in Bayesian inference. The priors represent our prior beliefs about the model parameters before incorporating the data. The priors specified in the XML file are:

- **ACG (Ancestorial conversion graph) Prior:** Specific prior for some parameter related to ACG. This is a prior used on both, the phylogenetic tree and the number of contact edges in the tree. Thus, it has a tree prior as an input (the birth-death model) and the expected conversions.
- **Expected-Conversions Distribution:** Prior distribution depicting the expected conversion numbers, modelled in contacTrees.
- **pMove Distribution:** Prior distribution for the borrowing probability (pMove) in the "ContacTree" model. The range is set between 0 and 0.4.
- **Birth-Rate-Prior:** Prior distribution depicting the birth rate, which follows the initial values.
- **Death-Rate-Prior:** Prior distribution depicting the death rate, which follows a uniform distribution.
- **Topology Priors:** The MRCA-Prior (prior of most recently common ancestor) defines 10 monophyletic subtrees (taxon set) within the main phylogenetic tree These subtrees represent specific groups of taxa sharing a common ancestor and the time this common ancestor existed. This prior information is used to guide the inference

of the phylogenetic relationships by anchoring times of certain events within the tree.

2.2.5. Model Likelihood

The model likelihood specifies how well the data (sequence alignments) fit the model. In this case, there are two likelihood components:

- Cognate concepts: Tree likelihoods of concepts in the contacTrees-addon.
- GeoSphere: Tree likelihoods of tip locations in the GeoSphere-addon.

2.2.6. MCMC Setup

Our Markov Chain Monte Carlo (MCMC) model is set to run for 20,000,000 iterations. BEAST2 will log the model values every 500 iterations of the MCMC, allowing the model to diverge sufficiently from the last logged values, which minimizes auto-correlation.

2.2.7. MCMC Operators

The MCMC operators are the tools used by the MCMC sampling process, they allow for changes to the parameters and tree topology to occure. In the XML-file various types of operators are included, such as: GeoSphere-operator ,contacTrees-operators, WilsonBalding-operator, subtree exchange-operator, ACGScaler, birth-rate-scale-operator, and death-rate-scale-operator.

2.2.8. Loggers

The loggers are used to record various values during the MCMC analysis. The trace loggers are set to log specific parameters, likelihoods, and priors every 5000 iterations, including posterior values, likelihoods, prior values, ACGStats-Logger, clock rates (slow, medium, fast), location likelihood and precision.

2.3. Tree-Files (BEAST2 Output)

In the scope of this thesis, the ContacteR and contactCoordinateR functions were developed to analyse the contact event data derived from BEAST2's contacTrees-addon output (tree files) within the R programming environment. The primary objective of these functions is to allow access to contact event data and subsequently calculate the spatial positions of the start and end nodes of contact edges.

2.3.1. ContacteR Function:

The ContacteR-function takes the tree-file's phylogenetic data frame as input. The tree file contains a list of phylogenetic tree samples from a BEAST analysis. The trees are encoded in extended Newick format which contains information on the topology, node heights and node attributes (like inferred geo-locations). The data frame contains detailed information about the phylogenetic tree and its associated nodes and edges. The function systematically iterates through all languages present in the tree data list. For each language, it follows the path from the tip (end-site) to the root (start-site) of the tree matrix.

After contact edges are detected, the ContacteR-function separates every contact edge from the regular edges. The regular edges of the tree are still contained within the edge matrix of the tree-file, whereas the contact edges are saved in a separate data frame. . In this way, the phylogenetic data of the tree file can be interpreted correctly by standard phylogenetics R packages like ape (Paradis et al., 2019) and treeio (Yu, 2022; Wang et al., 2020). The function also has a special output type: The NodePath data frame. It contains all the nodes each language has travelled through in its evolutionary history. It contains the age and spatial position of each node and shows the status of contact event. If the node is indeed from a contact event, then it further shows if it is a start or end node of a contact event.

2.3.2. contactCoordinateR Function

The function was designed to calculate spatial positions of contact edge nodes using Brownian motion. The function relies on the rbridge-function of the e1071 package (Dimitriadou et al, 2015), which generates a Brownian bridge. The output is a vector with movement increments. The cumulative sum of the output vector is calculated. Finally, the cumulative sum of movements is normalized.

Another pivotal component of the contactCoordinateR-function is the iteR-iteration-function. This function facilitates the movement upwards or downwards along the NodePath-list towards a clad that connects non-contact edge nodes. It is necessary to access positions of the GeoSphere-addon, which only non-contact edge nodes have.

The function calculates Brownian motion in both longitude and latitude directions along each of the two language paths involved in the contact event. The upper and lower bounds of the Brownian motion intervals for longitude and latitude are scaled according to the temporal differences of the uppermost and the lowest known position in space.

The spatial positions of the start and end nodes of contact edges are derived by extracting the longitude and latitude values from the scaled Brownian motion intervals at the specific time when the contact event occurred.

2.3.3. Visualisation

In order to visually confirm and evaluate the inner workings of the ContacteR and “contactCoordinateR” functions, an additional plot function contactPlotteR was created. It is based on the tree visualization addon ggtree (Yu, 2022; Xu et al., 2022) for well-known visualization package ggplot2 (Wickham, 2016) and allows the user to visualise BEAST2 and contacTrees tree-files in the R environment.

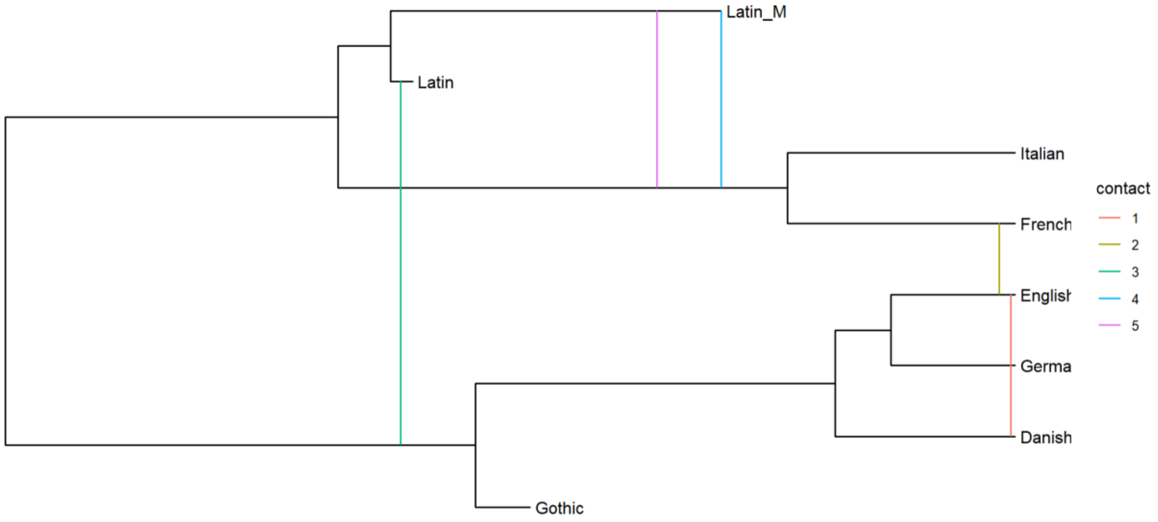


Figure 1: Shows an exemplary contact tree visualised by "contactPlotteR".

2.3.4. Function Summary

By combining the functionalities of contacteR and contactCoordinateR, further research implementing R in their analysis of contacTrees tree files can be streamlined greatly by allowing easy access to contact event data within. The NodePath-output mode further helps with the analysis. Moreover, the R functions developed for this thesis could easily be converted to a R-package, allowing easy access for researchers more versed with the R environment.

2.4. Introduction: Cost Surface Network (CSN)

While previous research has explored the integration of THF into a spatial network, examples would be metro or street networks (Goodchild, 2022). This approach stands apart by applying THF to a spatial network, which is simulating a cost surface. The goal of this simulation is to predict the results of THF applied on a high resolution cost surface such as a DTM, by aggregating the raster cells of the cost surface to boost the performance. Simultaneously, the systematic error between the original cost surface and the network is calculated by using a linear regression model, allowing to minimise this systematic error. This results in an overall high efficiency level, while also minimising the models uncertainty.

This novel concept converts of a cost surface into a spatial network, where raster values are aggregated into network nodes, and the relation between nodes is expressed using THF. Therefore, reflecting the travel cost from one grid cell to the next. The usage of Dijkstra's shortest path algorithm allows us to calculate the lowest travel cost possible to move from one point in the topography to another. This is the same methodological principle used to calculate the shortest paths via THF applied on a regular DTM, which allows us to compare the new CSN approach with the classical approach (on a DTM) directly. Allowing an intuitive way for validating this new approach by making a regression model between CSN results and DTM results.

Calculating travel costs using THF in a more efficient way is a key capability to calculate thousands of travel costs of paths with lengths over thousands of kilometres in an acceptable time frame. This is a necessity for this thesis and for a potential implementation as a prior into a phylogenetic model. The CSN approach promises to solve the efficiency-problem and thereby allows the implementation of a geo-prior based on THF.

2.4.1. Hexagonal Grid Cells

Adopting hexagonal grid cell has a lot of benefits to organize space through its neighbourhood relations, which has spawned several research projects and commercial applications, including Uber's hexagonal hierarchical spatial index (Uber, 2023). Adopting hexagonal grid cells as network structure has several benefits. The neighbourhood of each grid cell comprises of six other grid cells, which are always separated by the same angle. Further, each grid cell centroid is evenly spaced to other centroids. The uniform angles and equidistant distribution of grid cells allows for isotropic movement in a simple pattern. This two characteristics alone make it superior to other grid patterns, such as the chess board pattern. Of further benefit is also the fact that hexagonal grid cells are a much closer representation to natural phenomenon such as honeycombs and plant cells than quare grid cells.

2.4.2. Cost Surface

Cost surfaces represent the conductance or cost of traveling across geographic space or any kind of topology, often taking into account terrain characteristics and other environmental factors. Because of THFs input, solely a Digital Terrain Model (DTM) is needed as the cost topography for our cost surface, which keeps the variety of input data simple. The CSNmakeR function serves as a key component, enabling the transformation of the cost surface into a cost surface network (CSN).

2.4.3. CSNmakeR Function and Cost Surface Network (CSN) Setup

To create the CSN, we developed the CSNmakeR-function . In the first step, hexagonal grid cells are generated, and the spatial network is set up using the sfnetworks-package (van der Meer et al., 2023), where nodes are defined as the centroids of these hexagonal grid cells, and the edge relations are expressed through the touching sides. Notably, the edges are undirected and they are straight lines. At first, initialising the network edges as undirected in the context of THF sounds to be incorrect. THF has an inherent direction bias after all. It is much easier to walk down a slope with 5 degree, than walk up a slope with 5 degree (Tobler, 1993). The reason for the undirected network edges lies in the consideration that the traveller is not constrained to move in a specific direction for just one time. Contact events are characterized through prolonged cultural and linguistical exchange, to facilitate travel between cultural entities over long periods of time, travel between those entities must be bidirectional in nature.

After the CSN is set up, the raster cells of the DTM must be aggregated into the nodes of the network. In the second step of the CSNmakeR-function, values from the cost surface raster cells are extracted and aggregated within each hexagonal grid cell. The mean and standard deviation of these values are calculated and stored as attributes in the network nodes. The schematics of this process can be seen in figure 2.

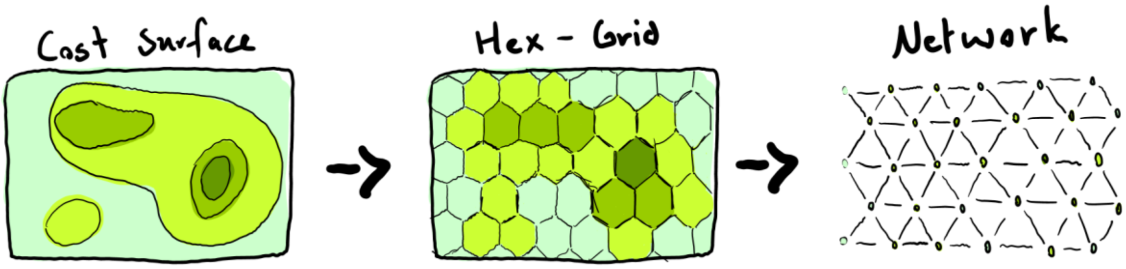


Figure 2: Schematics of the transformation process form cost surface to CSN

2.4.3.1. Implementing Tobler's Hiking Function (THF)

The next step of the CSNmakeR-function is to implement THF to generate edge weights, which are later used to calculate the shortest paths within the network. THF is a well-established empirical model and used widely (Goodchild, 2020; Higgins, 2021; Campell et al., 2019). that estimates the speed of human movement across various terrains. It provides a representation of pace changes in response to slope variations. The process involves the computation of slopes between neighbouring network nodes using height differences and geodetic edge length. Then, the pace version of THF is applied on these slopes. Finally, we multiply the resulting paces with the geodetic edge length, resulting in an understandable cost unit as “hours” or “hours to traverse”. This also counteracts the projection distortions that can surface using a spherical coordinate reference system such as WGS 84.

$$Toblers\ Hiking\ Function\ for\ Pace: \qquad THF_p(slope) = 0.6 * e^{3.5*|slope\ -\ 0.05|}$$

$$Calculation\ per\ Cell: \qquad THF_p(Slope) * \Delta l_{geodetic} = THF_p\left(\frac{\Delta l_{geodetic}}{\Delta h_{DTM}}\right) * \Delta l_{geodetic}$$

2.4.3.2. Implementing large Water Bodies:

Large water bodies such as oceans and large lakes pose a distinct problem when modelling terrain based travel costs. In order to deal with large water bodies within the CSN a conditional query for water is applied. Nodes which fit this condition are given a very low pace of 2.0 seconds per meter. This results in a very low conductivity over large water bodies, which leads to a high degree of avoidance by shortest paths and is only travelled through when necessary. This type of modelling water bodies fits well with the XYZ treatment often used in cost surface analysis.

2.4.3.3. shortestPatheR Function and the Difficulty of Terrain Model (DoTM)

We developed the shortestPatheR-function to efficiently calculate the shortest paths within the CSN. The function is based on Dijkstra's algorithm for calculating shortest paths within the CSN using the “distance” function from the igraph-package (Csardi, 2006). The function efficiently finds the shortest path between the start and end nodes of each contact edge using Dijkstra’s algorithm. As inputs, the function takes the CSN, allowed start positions, and a phylogenetics file. It identifies the start and end nodes of each contact edge and computes the corresponding shortest paths.

In addition, a terrain difficulty metric is implemented. It is a representation of the topographical difficulty encountered along the path. The function calculates the mean of all standard deviations along the paths, which is then saved as the difficulty of terrain model (DoTM). The DoTM quantifies the terrain difficulty encountered during movement along the paths and is an important factor to counter the flatting effect of aggregation. The output includes both numeric values, representing the hours travelled, and geometric representations in the form of simple features, providing a comprehensive set of path data.

2.4.4. CSN Validation

Validating the effectiveness of the CSN is crucial to ensure that the network accurately represents the cost surface data and reliably predicts shortest paths. In this context, the CSN is validated using a linear regression model. The predictability of the CSN stems from a systematic error produced by the level of aggregation, which makes accuracy estimates possible. We discern between two modes to capture accuracies: length-dependent regression model and length-independent regression model.

- To measure length-dependent regression, the CSN output (travel costs) and the logarithmic values of DoTM serve as predictors in the regression model to predict the DTM output. This accuracy metric predicts the travel costs of a particular shortest path directly. The new cost is set according to the original travel costs and terrain difficulty (DoTM) along the way. This means that the length-dependent regression evaluates the length of the path and its travel cost directly.
- To measure length-independent regression, the CSN output is normalized with the length of the shortest path (mean pace) and DoTM values serve as predictors in the regression model to predict the DTM pace output. This accuracy predicts mean paces along a particular path, according to the original mean pace measured with CSN and the terrain difficulties (DoTM) along the way. This means that length-independent regression characterises the movements through the CSN on the grid cell level.

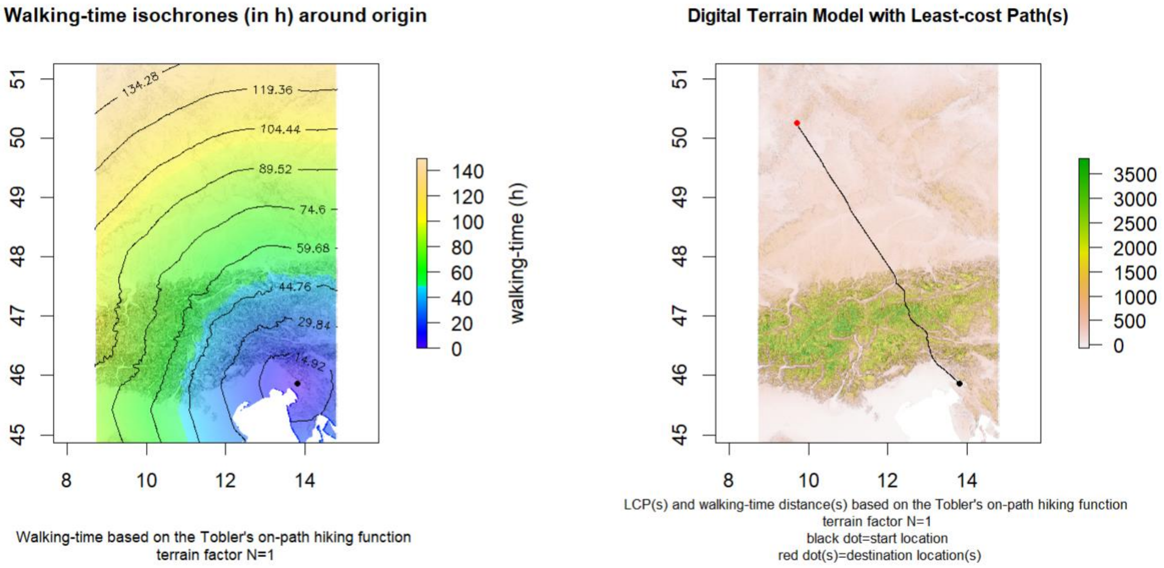

Walking-time isochrones (in h) around origin
134.28
119.36
104.44
89.52
74.6
59.68
44.76
29.84
14.92
51
50
49
48
47
46
45
8
10
12
14
140
120
100
80
60
40
20
0
walking-time (h)
Walking-time based on the Tobler's on-path hiking function
terrain factor N=1
Digital Terrain Model with Least-cost Path(s)
3500
3000
2500
2000
1500
1000
500
0
LCP(s) and walking-time distance(s) based on the Tobler's on-path hiking function
terrain factor N=1
black dot=start location
red dot(s)=destination location(s)

Figure 3: Least cost path calculation on DTM in middle European ROI. showing walking times (left) and the shortest path (right) on contact edge Latin-Germanic

CSN variants are set up and compared to a high-resolution DTM by applying least cost path (also using Dijkstra's algorithm) to a small region of interest (ROI) in central Europe. Shortest path combinations between 12 points within the ROI are calculated for various resolutions, including the initial DTM resolution and different CSN resolutions. By evaluating the performance of the CSN across these resolutions, we aim to identify the most suitable resolution that balances precision and computational efficiency.

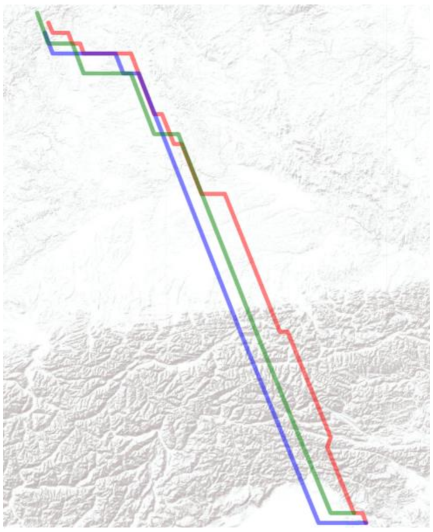


Figure 4: Shortest paths through CSN on contact edge Latin-Germanic. Showing shortest paths in CSNs with 0.1 (red), 0.2 (blue) and 0.3 (green) degree resolution.

2.4.5. Summary CSN

This novel methodological approach of integrating THF into the CSN framework is expected to be a powerful tool for cost surface analysis using spatial networks to raise efficiency. The CSNmakeR function successfully transforms cost surfaces into spatial networks, allowing efficient calculations of shortest paths using Dijkstra's algorithm. The validation process shows the reliability and accuracy of the CSN. Therefore, we are confident that the measures described above are sufficient to produce a solution for the efficiency-problem.

2.5. Contact Events and Non-Contact Events

Conceptually speaking, non-contact events are distances between languages at the same time-depth and represent plausible spatial distances between the languages of the evolutionary tree, derived from the same height of the tree as the contact event occurred. In other words, they are all the other combinations of languages, who could be potential contact events, but where not modelled that way. The primary purpose of using non-contact events is to validate the statistical significance of contact events between languages and thereby solving the feasibility-problem. In order to conduct a meaningful comparison, a control group must be established, against which the contact events can be evaluated. This control group comprises of the non-contact events, serving as a baseline for assessing the significance of the predicted travel costs of the contact events.

2.5.1. Production of Non-Contact Events

In order to produce non-contact events we make use of the same two key functions that were applied to produce the contact events: shortestPatheR and contactCoordinateR.

The first step consists in estimating the spatial position of each language at the time of the contact event. The contactCoordinateR-function helps in estimating the coordinates of languages at the moment of language interaction. The shortestPatheR-function assists in determining the shortest path between two language positions. Once the spatial positions of all languages within the CSN are estimated, the Dijkstra's algorithm is used. This algorithm efficiently calculates the shortest distances between all language positions within the CSN, creating a comprehensive distance matrix. The resulting distance matrix displays all possible distances arising from the combinations of all language pairs.

2.5.2. Evaluation of Contact Events

The evaluation for statistical significance of contact events and their non-contact events is carried out by statistical tests. In the evaluation, two sample populations are set: the test group and the control group. The research group consists of data from language contact events. On the other hand, the control group consists of non-contact events. By doing so, we can assess whether the observed differences between the test group and the control group are statistically significant or simply due to chance. For that purpose we device a Wilcox rank order test. We are confident that this measure allows us to solve to feasibility-problem.

3. Results

The results achieved in this thesis paint an interesting picture. The resulting tree-file input produced with BEAST2 (including the addons GeoSphere and contacTrees) show acceptable ESS levels across all parameters. Further, the major problems (feasibility-problem and efficiency-problem) to implement THF into a geo-prior were addressed by this thesis. The Investigation into research questions I and II produced interesting results, which will be discussed in the next section.

3.1. Phylogenetic Trees

We assessed the quality of the phylogenetic trees generated by BEAST2 and contacTrees through comprehensive analysis of output data using “Tracer” and “Spread”. Our evaluation is based on more than 5000 phylogenetic trees over two independent runs.

Using Tracer, we evaluated the convergence of 27 key parameters. The evaluation shows no low convergence measured in the Effective Sample Size (ESS) values (ESS with less than 100). The parameter with the weakest convergence was "paired tree length," with an ESS of exactly 100. There are 10 parameters with medium ESS values (ESS with less than 200) observed. On average, the ESS values of medium convergence hovered around 150, indicating acceptable convergence in most cases.

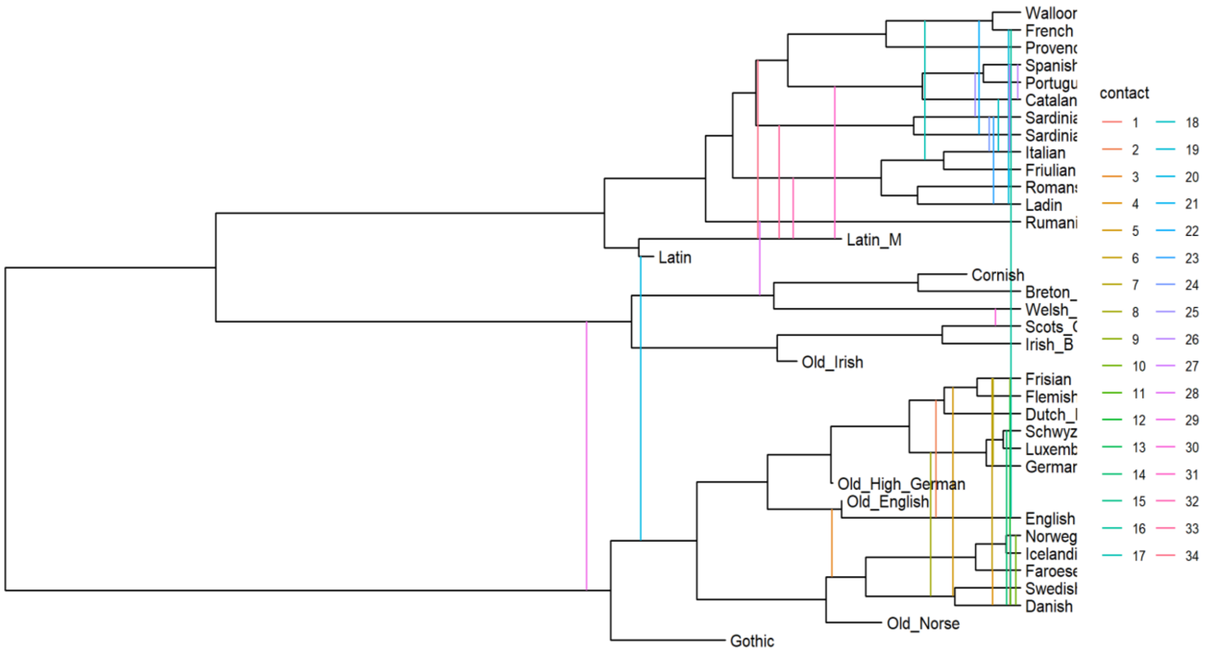


Figure 5: Phylogenetic summary tree produced by BEAST2

To evaluate the spatial aspect of our results, we conducted a qualitative assessment, identifying Afrikaans as a geographical outlier in our dataset (Outside of the defined ROI) due to its colonial history. For this reason, we removed Afrikaans. The removal of Afrikaans should not impact the language migration patterns within Europe.

The evaluation of BEAST2-generated phylogenetic trees demonstrates overall satisfactory convergence and sample size effectiveness, although improvements could be made in certain parameters. Out of this evaluation we decided on the used sample size of 200. Which strikes a good balance between the autocorrelating effect and convergence.

3.2. Travel costs (Research Question I)

In the investigation of contact events as outlined in research question 1, we analysed the contact events and their corresponding control group, the non-contact events. Our findings show a clear picture on the spatial dynamics of linguistic contact events and the non-contacts (alternative distance) within the same temporal cross section of the language tree.

The evaluation is done on a CSN with a spatial resolution of 0.3 degree and length-dependent regression model. More about the evaluation process will be said in the next chapter (CSN evaluation). Contact events and their non-contact events were produced by using shortest path algorithms applied on the CSN. For the evaluation of contact edges and non-contact events, we used the sample size weighted variants of means, standard deviations, box-plots, and statistical tests provided by the stats R-package (Baldwin et al., 2012). The reason for this, is the aggregated nature of the data. Because of space reasons, the output of the distance matrixes are not saved fully, only the necessary metrics, such as mean, standard deviation and sample size are saved.

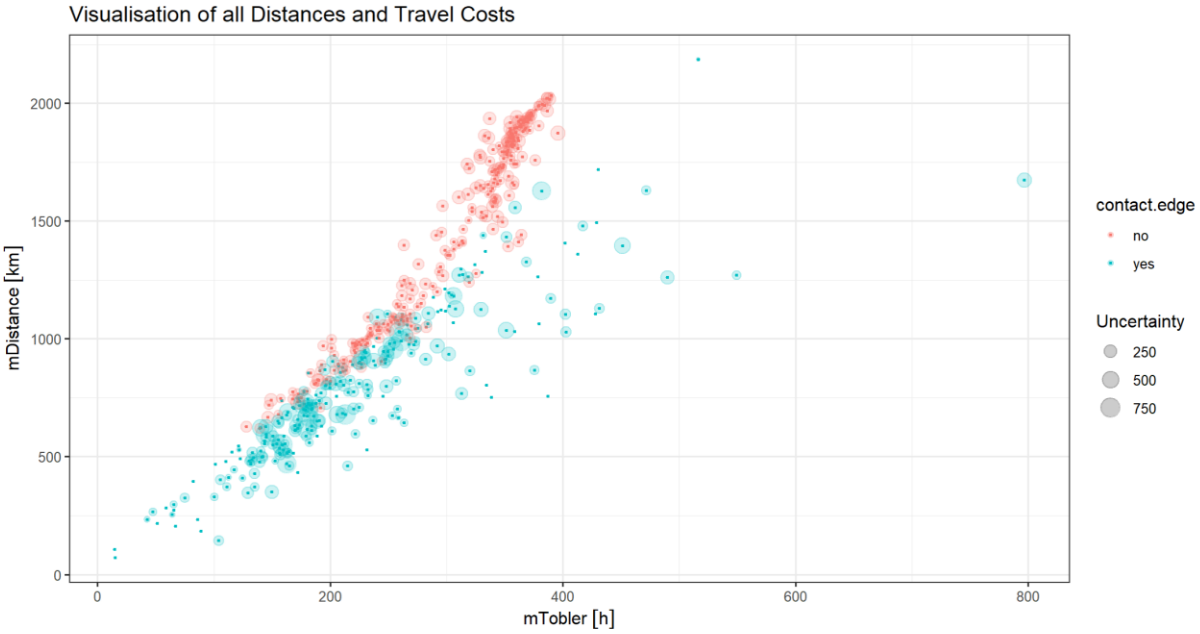


Figure 6: Visualisation of all Distances and Travel Costs: Showing how distances (in kilometres) and travel cost (in hours) match each other. The colours display the status as contact edge or non-contact edge, while the circles indicating the standard error.

3.2.1. Contact Events (Test Group)

The weighted mean travel costs of the contact events reaches 202.2 hours, with a weighted standard deviation of 80.6 hours. The substantial sample size of a total of 6491 contact events (302 unique contact events) allows for a meaningful analysis with high levels of confidence. Furthermore, the distribution of contact edges significantly deviates from normality (p < 10^-9).

3.2.2. Non-Contact Events (Contral Group)

- The weighted mean of the travel costs of non-contact events is notably higher at 351.3 hours, accompanied by a weighted standard deviation of 32.7 hours. The sample encompasses 1.6 million non-contact events from all the 302 distance matrices of all unique contact events.
- Similar to the contact events, the distribution of non-contact events within the control group significantly departs from normality (p < 10^-11).

3.2.3. Statistical Evaluation

Employing the Wilcoxon Rank-Order test, we observed a highly statistically significant result (p < 10^-16). This outcome supports the alternative hypothesis that the true location of the shift in non-contact events is not equal to zero, and therefore not equal to the contact edge population, thereby showing that travel costs for contact events are significantly lower than non-contact events. This result underscores the presence of meaningful spatial differences between the contact events and the control group, reinforcing the notion of contact events being spatially distinct travel cost wise.

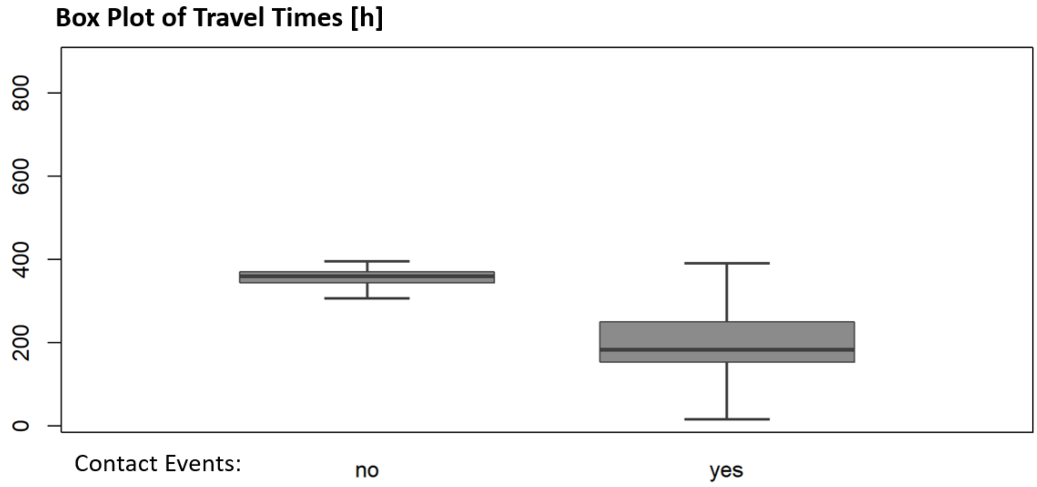


Figure 7: Weighted Box Plot of the Travel Distances: Shows significant difference between the populations of contact edges and contemporary distances. The scale represent travel time in hours.

3.3. CSN evaluation (Research Questions II)

The CSN approach is used to calculate travel costs via THF over a cost topography in an efficient and accurate way. The goal is to find the optimal balance between efficiency and accuracy in order to deduce the best setup for calculating travel costs in this thesis and for future endeavours. Therefore, evaluating this new approach is an important step.

3.3.1. CSN production

In section 2.3 we introduced a new approach to efficiently approximate Tobler's hiking distance between two points in a topography by using a Cost Surface Network (CSN). In order to evaluate the new CSN approach a series of CSNs were produced. These spatial networks are simulating and approximating THF (THF) values on diverse terrains. The CSNs are configured to varying spherical resolutions ranging from 2.5 degrees down to just 0.1 degree.

Here are the specifications of the evaluated CSNs

Table 1: CSN specifications

Resolution [DEG]	Nodes	Edges	Memory [MB]	Mean Pace [s/m]	Median Pace [s/m]	Min. Pace [s/m]
2.5	1136	6548	3.4	0.7147558	0.7146974	0.7471277
1.2	4752	27’962	14.4	0.7148412	0.7147079	0.7828431
0.9	8308	49’122	25.3	0.7148887	0.7147477	0.9016945
0.6	18’590	110’450	56.8	0.7149971	0.7147477	0.9759570
0.3	73’402	438’238	325	0.7152783	0.7147477	1.1981637
0.1	655’655	3’927’446	2762	0.7180102	0.7147477	2.4873054

The CSN series shows stable median pace values by a resolution of at least 0.9 degrees. As expected, the minimal pace gets lower the better the resolution becomes, because of the increasing slopes that higher resolutions produce. Which also explains the steadily falling mean pace. Network size – therefore, memory usage – raises quickly by increasing the resolution, which is by no means a problem. What matters the most is the consistency of the estimated travel costs in form of pace.

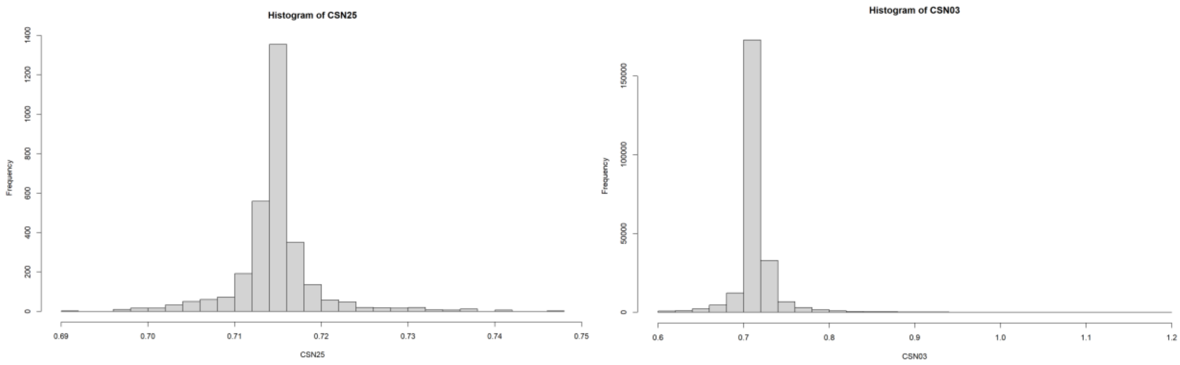


Figure 8: Shows the distribution of the THF paces of all network edges. ***(Right)*** *The shape of the distribution is typical for THF and with a resolution of 0.3 degree even more pronounced.* ***(Left)*** *The resolution of this CSN is 2.5 degree. The low resolution shows in the low frequency numbers.*

3.3.2. CSN evaluation and prediction

After setting up a series of CSNs at different resolutions, we evaluate them on a set of 132 paths (12 x 12 point combinations). These 132 paths are located in a test ROI set in central Europe (Northern Italy, Switzerland, west Austria and southern Germany). Shortest paths between all point combinations are calculated in each of the six CSNs. Then the shortest paths are evaluated with linear regression for each CSN.

Evaluation of the CSN series is done in 3 steps:

1. Step: Calculation efficiency
2. Step: Length-independent regression model evaluations
3. Step: Length-dependent regression model evaluations

3.3.2.1. CSN Evaluation and Prediction

Computation time increases significantly when the size of the CSN increases. There seems to be a linear relationship between Calculation efficiency and the number of edges in the CSN, which are increasing quadratically by increasing the resolution (see figure 9). The linear regression models, which produced the length-independent regression (using mean pace per path) and the length dependent regression (using path travelling time) are highly statistically significant for all resolutions.

The Evaluation of the CSN series continues with the examination of the Residual Standard Error (RSE), Adjusted R2 values. Furter, the Root Mean Squared Error (RMSE) for predicted (using linear regression models) and unpredicted scenarios are compared to show their predictabilities. The adjusted R2 values drop off only after increasing the resolution over 1 degree. Adjusted R2 and the residual standard error (RSE) stay stable over a wide set of resolutions. The length-dependent regression (using path travelling time) are also highly statistically significant. The adjusted R2 values are very high, but they are dropping off slowly and steadily. The adjusted R2 and the residual standard error (RSE) increase steadily but with diminishing returns by increasing the CSN resolution. The same trends can be observed for the RMSE. The length-independent regression also plateaus between the resolutions of 0.9 and 0.3, were as the length-dependent regression increases steadily, but with diminishing return (see figure 10 and 11).

3.3.2.2. CSN Decision

In order to predict the travel time duration most accurately, the CSN with the best length-dependent regression model results is chosen. As can be deduced from the data, the chosen CSN would be the one with the highest resolution, because the prediction performance increases steadily by raising the resolution. But, increasing resolution comes with increasing computational time, making a trade-off between accuracy and efficiency necessary. The CSN with the highest resolution that the system in use can handle is 0.3 degree. Therefore the CSN with 0.3 degree is the optimal tool for the system and for predicting the travel time duration of contact edges and their non-contact events within the given cost topography, while having a RMSE of only 5.3% of the mean travel time.

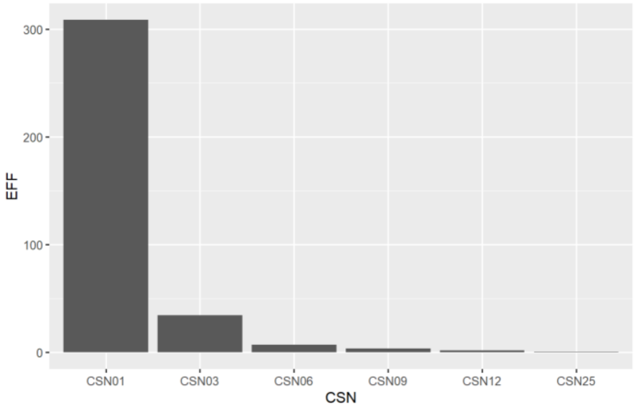
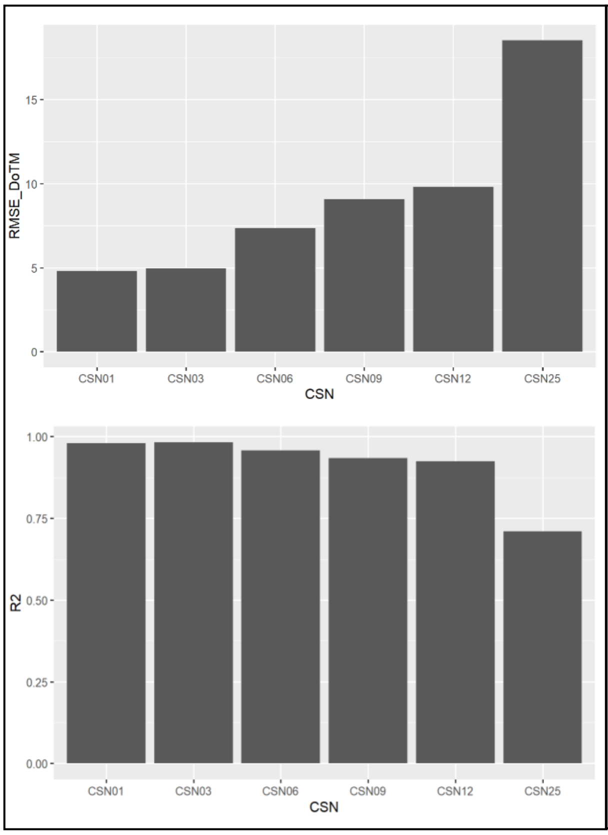


Figure 9: Calculation efficiency on the CSN resolution types



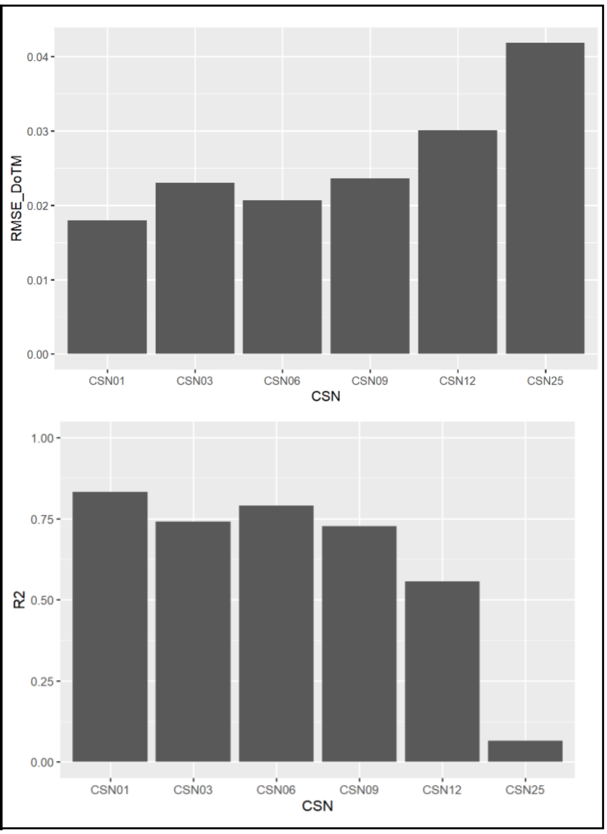


Figure 10: Length dependent statistics
Top: RMSE with DoTM support values.
Botton: Adjusted R2

Figure 11: Length independent statistics
Top: RMSE with DoTM support values -
Botton: Adjusted R2.

3.3.2.3. Data overview: Efficiency and accuracy figures

Table 2: Length-dependent regression

Resolution [DEG]	Run Time [s]	Mean Travel Time [h]	P-Value	Adj. R2	RSE [h]	Unpredicted RMSE [h]	Predicted RMSE [h]
2.5	0.60	89.41	10^-16	0.7097	19.17	32.06 (35.9%)	18.53 (20.7%)
1.2	1.99	91.22	10^-16	0.9242	9.95	11.66 (12.8%)	9.82 (10.8%)
0.9	3.80	94.59	10^-16	0.9344	9.25	11.46 (12.1%)	9.09 (9.6%)
0.6	7.28	96.28	10^-16	0.9572	7.47	10.90 (11.3%)	7.38 (7.7%)
0.3	34.31	93.64	10^-16	0.9825	4.98	8.39 (9.0%)	4.98 (5.3%)
0.1	308.47	95.18	10^-16	0.9805	4.89	7.94 (8.3%)	4.82 (5.0%)

Table 3: Length-independent regression

Resolution [DEG]	Run Time [s]	Mean Pace per path [s/m]	P-Value	Adj. R2	RSE [s/m]	Unpredicted RMSE [s/m]	Predicted RMSE [s/m]
2.5	0.60	0.733	8*10^-3	0.065	0.0438	0.307 (41.8%)	0.042 (5.7%)
1.2	1.99	0.789	10^-16	0.556	0.0304	0.113 (14.3%)	0.030 (3.8%)
0.9	3.80	0.811	10^-16	0.728	0.0236	0.117 (14.4%)	0.024 (3.0%)
0.6	7.28	0.831	10^-16	0.791	0.0209	0.095 (11.6%)	0.021 (2.5%)
0.3	34.31	0.804	10^-16	0.729	0.0238	0.077 (9.6%)	0.023 (2.9%)
0.1	308.47	0.814	10^-16	0.833	0.0187	0.058 (7.1%)	0.018 (2.2%)

4. Discussion

In this section, we discuss the implication of our findings regarding the analysis of contact events, as well as the implications from the CSN evaluation, thereby solving the feasibility-problem and the efficiency-problem. Consequently, the feasibility of implementing THF into a geo-prior for the BEAST2 addon contacTrees is discussed. Last but not least, measures to further improve the methodology are also discussed. This measures include: Introduction of length-independent regression and influence vectors.

4.1. Travel costs of contact events as measure for prior-feasibility

Solving the feasibility-problem of travel costs based on THF is one of the two major goals in this thesis. In order to solve this problem, we evaluated all contact events by comparing them to non-contact events (possible travel costs between languages at the same tree height), putting these two metrics into a research-control group dichotomy for statistical testing.

The Wilcox rank order test revealed that contact events show significantly lower travel costs compared to non-contact events. Thus, a geo-prior based on THF should produce a strong signal to infer contact. This finding aligns with the expectation that contact between languages are more likely to occur when the languages are in close geographical proximity. Therefore, using travel costs based on THF solves the feasibility-problem, thereby confirming research question I.

4.2. Cost Surface Network (CSN) approach to boost efficiency

In the following paragraphs, we delve into considerations surrounding the CSN approach employed in this thesis and show that the CSN approach is a valid solution to address the efficiency-problems.

Calculation Time and Resolution

One notable observation in this thesis is the relationship between calculation time and spatial resolution of the CSN. We found that as the spatial resolution increases, the computation time increases quadratically. This behaviour is attributed to the growing number of grid cells that need to be processed as the resolution becomes higher. Consequently, it is necessary to balance the trade-off between resolution and computational efficiency when implementing CSN-based models.

Another consequence of the quadratically growing grid cells is that both the number of edges and nodes in the CSN also increase quadratically with rising resolution. This phenomenon has a direct impact on the model's complexity and resource requirements as well as system memory usage. It's crucial to consider these aspects when designing and deploying CSN-based solutions. Higher resolutions offer more detailed information but at the expense of increased computational demands.

Length-Dependent Regression

One of the findings in this thesis is the relationship between resolution and accuracy, particularly in terms of length-dependent regression. We observed that length-dependent regression increases linearly with higher resolution but with diminishing returns. This insight suggests that while increasing resolution is beneficial for improving accuracy, the gains become less pronounced as resolution continues to rise. Therefore, choosing the highest resolution possible is advisable for optimizing model performance.

Length-Independent Regression

In contrast, length-independent regression follows a different path. Table 3 shows that this accuracy metric plateaus when the error margin reaches approximately 2% after a resolution of 0.6 degrees is reached. This observation suggests that beyond this point, further increases in resolution do not significantly impact length-independent regression. Consequently, scaling down the resolution can be considered without compromising performance.

Optimization Strategies

Based on these findings, optimizing the CSN approach requires some considerations about the regression model used to correct the CSN output. It depends further on the requirements and feasibility of the length-independent regression. For most scenarios the length-dependent regression should suffice as a brute force method. Pursuing higher resolutions is therefore a logical choice, although the diminishing returns should be acknowledged. Conversely, if the implementation of length-independent regression can be accomplished, it would be preferable. Performance wise, this solution is the clear winner, where scaling down the resolution can be a viable strategy to improve computational efficiency without significant loss of accuracy. All in all, it is clear, that regardless of using length-dependent or length-independent regressions, the efficiency boost provided by the CSN approach is a huge factor. We conclude that the CSN approach sufficiently addresses the efficiency-problem, thereby confirming research question II.

4.3. How to implement the CSN optimally

After showing that the feasibility-problem and the efficiency-problem are both solved successfully, this chapter delves into questions of optimisation. In particular, we show why we use length-dependent regression over length-independent regression for this thesis and why we recommend solving the problems length-independent regression has in future research. The choice for using length-dependent regression are twofold: First, the ease of implementation and second, the challenges with length-independent regression.

Ease of implementation

The more straightforward factor in this decision was the ease of implementation. Using length-dependent regression involves just the implementation of a linear regression model directly to the CSN output. This straightforward approach offers simplicity in correcting travel time costs over the modelled terrain.

Challenges with Length-Independent Regression

While length-dependent regression presented clear advantages regarding feasibility, the length-independent regression would be – theoretically – clearly the superior solution. length-independent regression has much lower margins of error in their prediction of travel time (see table 2 and 3). Furthermore, the error can completely disappear, if the error margins show true randomness along the shortest paths. Because the regression model would be directly applied to the CSN grid cells, there would be no need for post-processing of the travel cost, which would further increase efficiency.

However, there were notable challenges associated with the implementation of length-independent regression making implementation impractical. There is no clear path of implementation. A best-guess approach would be to correct the CSN grid cell values using a linear regression model. But, this is insufficient because there is a general lack of Commonality between on-path grid cells and regular grid cells. The regression model stems only from on-path grid cells. There is a distinct possibility that here is no communality between the two samples. Thus, the regular grid cells could be categorically different from the on-path grid cells.

Preference for Length-Independent Regression

As stated in the former paragraph, if these challenges could be resolved efficiently, the implementation of length-independent regression within the grid cells would be preferable. This preference is grounded in the potential for improved accuracy, particularly in scenarios where errors are truly random. In such a case, errors in all the grid cells along the path can cancel each other out, resulting in a more precise representation of travel cost. But, because of the challenges surrounding the implementation of length-independent regression, we opted to use length-dependent regression in instead.

4.4. Modifications on Tobler's hiking function *(THF)* and the CSN:

We want to further highlight the efficiency and versatility of THF when calculated by CSN. This novel approach has proven to be a useful tool for this thesis for calculating travel costs, offering a lot of advantages:

- Multiplying factors in order to model certain behaviour or environmental impacts to THF is a common occurrence and the practice is widely adopted in scientific papers (Pingel, 2010; Irmischer et al., 2018; Marquez-Perez et al., 2017; Collischonn et al., 2000). Examples are the incorporation of factors like terrain types, climate factors such as temperature and humidity, or various modes of transportation, such as walking, horseback riding, or wagon travel.
- Additionally, directional factors can be incorporated via the CSN. Compared to simple raster data, the ability to make direction based calculations is the key technical advantage of the new CSN approach, because CSN is network based, facilitating calculations based on direction is not a problem at all. Besides THF, another application would be the implementation of the directional features of a river. Travelling along a river would result in a travel cost reduction. In comparison, crossing a river would produce a travel cost penalty. The modifications could greatly enhance the fidelity of travel cost calculations, allowing for a more nuanced calculation of travel cost based on a wide array of scenarios.

4.5. A new prior for contacTrees based on Tobler's hiking function *(THF):*

Having solved both, the efficiency-problem, as well as the feasibility-problem, by confirming both research questions, we propose to incorporate THF into the geo-prior using the CSN approach. We highlight how this approach is feasible and addresses the efficiency challenges and improves travel cost estimation greatly. Further, this approach also reduces outliers occurring in geodetic distances and thereby enhances the overall utility of contacTrees.

By making use of THF, we can better capture these spatial relationships and integrate them into contacTrees. The travel cost analysis shows its usefulness clearly by solving the question about feasibility regarding THF. The analysis showed clearly that THF is able to produce a usable signal to infer contact events.

One of the benefits of adopting the CSN approach for calculating THF is the high level of efficiency. Processing tens of thousands of shortest paths on a topography defined by THF is still resource-intensive but manageable due to the CSN's computational advantages. Despite the raise in efficiency, the accuracy of CSN is of a high quality due to the use of linear regression models with high predictability. The use of a CSN empowers contacTrees to handle THF on the scale needed. We are confident that the usage of the CSN approach in its state at the moment is sufficient to have the bare requirements of the efficiency-problem solved. By solving the challenges surrounding length-independent regression, efficiency could be boosted further.

Mitigating Outliers in Geodetic Distances

Transitioning the geo-prior of contacTrees from using geodetic distances to THF-based travel costs would be a significant step up on handling outliers introduced by geodetic distances. Geodetic distances measure the shortest distance between two points on the Earth's surface and often fail to account for real-world obstacles that can influence travel. In contrast, THF considers terrain features. THFs is ability to reduce outliers in geodetic distances lies in the fact that these outliers are not truly random; instead, they follow the pattern of this topological obstacles that hinder contact between populations, such as mountains and ravines, and assigns travel costs accordingly. Therefore, THF-based travel costs account for these obstacles, reducing the prevalence of outliers and leading to more meaningful and reliable results.

4.6. Prestige and introduction of the influence vector

In this chapter, we introduce the concept of the influence vector (IV) and its potential role in reducing outliers and improve the signal of contact events even further. We explore how languages are influenced by prestige and power dynamics, the important role of empires in linguistic contact, and a possible methodological approach to produce such a influence vector.

The role of empires and status dynamics

Linguistic contact is a phenomenon that can occur for various reasons. While contact between closely related or geographically proximate languages is relatively common and expected (Hock & Joseph, 2009; Chambers et al., 2004). Examples for closely related languages in contact are linkages between Germanic languages, particularly in the northern Germanic regions. An example for geographical proximity can be found in Switzerland, where the incorporation of words from nearby French-speaking regions into Swiss German dialects is common. But contact between languages is not just contained to small geographical areas or to closely knitted language families, it can also transpire over vast distances, particularly in the context of empires (Hock & Joseph, 2009; Daulton, 2019). Empires, by definition, are multicultural and multilingual entities with a dominant language often serving as a lingua franca. This dominant language projects significant influence on other languages within the empire, particularly in domains associated with high-status activities or professions (Hock & Joseph, 2009).

One of the challenges that will be encountered by modelling contact between languages by using THF in CSN models are the outliers produced by such empires and their dominant language. Empires possess the resources and infrastructure to maintain connections with high travel costs, enabling them to project their linguistic influence over much larger distances than neighbouring or closely related languages. This influence creates outliers in the model. We propose that future work could address those outliers using influence vectors.

Goal of the Influence Vector (IV)

The objective of the influence vector is to identify and quantify the influence of dominant languages within phylogenetic trees. This vector could be constructed by factoring in the following key elements:

- Temporal densities of start nodes: The IV considers the temporal densities of start nodes within all node paths (the path of a language in a phylogenetic tree form root to tip). Time intervals within the tree with a high start node density imply empire activity.

- Temporal densities of travel costs: The IV incorporates travel costs of the contact edges as densities along each note path. Time intervals with high travel cost densities imply empire activity.
- Temporal densities of the change rate in high-status meaning classes: Another factor is the change in high-status meaning classes over time along all node paths. Time intervals with high change rate densities of high-status meaning classes imply empire activity.

Finally, these three (or more) elements can be turned into a 3-dimensional (or n-dimensional) influence vector, which can be written out for each contact edges start and end node, so that all contact edges have their associated IVs. The metrics derived from those IVs are manyfold, but the most logical would be to calculate the difference between the start node and the end node IV and calculate its length.

4.7. Uncertainties

In this important chapter, we take a look at various sources of uncertainty that have emerged in this thesis. These uncertainties are important and can influence our CSN model and travel cost analysis greatly and play a big role in the robustness of our findings.

Uncertainty in Language Coordinates

One of the significant sources of uncertainty in this thesis comes from the approximation of language coordinates to the nearest CSN grid cell. This approximation introduces a spatial difference from the centroid of the grid cell to the coordinate of the language. Larger grid cells inherently have a higher potential for offset, than smaller once. To mitigate this uncertainty, approximating the distance between the language coordinate and the CSN grid cell could be implemented. This approach has the potential to reduce the uncertainty associated with language position within the CSN and enhance the precision of our results.

Uncertainty in Regression Models

Another dimension of uncertainty stems from the used regression models. While regression is a valuable tool for predicting signal strength, it is not able to eliminate uncertainty completely. We think that the implementation of length independent regression into the CSN could help reduce this kind of uncertainty. By incorporating length independent regression, we can refine the prediction models and potentially enhance their accuracy and reliability. This step is important to make sure that CSN predictions align with the ground truth derived from the DTM.

Uncertainty of Aggregation

The process of aggregating raster cells from a Digital Terrain Model (DTM) into CSN grid cells introduces another layer of uncertainty. The aggregation process involves converting fine-grained spatial information into more generalized representations, such as mean, standard deviation, and sample size. These transformations carry inherent uncertainty, as the fine-grained details may be lost or distorted in the process.

Uncertainty of phylogenetic and phylogeographic inference

Finally, we have to address the inherent uncertainty associated with the output of our computational model, BEAST2. This output is probabilistic in nature, and its reliability depends on the Effective Sample Size (ESS) values of the parameters involved, including prior, likelihood, and posterior distributions. We observed moderate ESS values for 12 parameters, indicating a high degree of uncertainty in our results. ESS is not the only source of uncertainty stemming from BEAST2. Other sources are inherit model assumptions, such as the selected models, or the prior specifications, which could lead to bias. Data quality and limits play also a huge role in the outcome of BEAST2 processing runs.

5. Conclusion

This section focuses on the integration of THF into contacTrees geo-prior by making use of CSN and delve into the wide range of applications that CSN can have. Further, the possibility to convert the functions created into a R package and possible future work is also laid out.

5.1. A new Geo-Prior by making use of CSNs

Implementing THF into the contacTrees geo-prior is a feasible and highly beneficial undertaking. Tobler's hiking distance outperforms geodetic distances for modelling linguistic contact due to its ability to reduce outliers and estimate travel costs more accurately. Travel cost based on geodetic distances are prone to outlier, which follow distinct pattern, such as short distances without contact due to topographical obstacles. THF is well-suited for the task of travel cost estimation in this context.

To implement THF effectively, a Cost Surface Network (CSN) approach is strongly recommended. The use of a CSN addresses the inefficiency of applying THF directly to raster data like a Digital Terrain Model (DTM). A CSN significantly improves computational efficiency, making calculations vastly faster than the traditional approach. For example, a CSN with a resolution of 0.3 degrees and using length-dependent regression for the correction is 269 times faster than a DTM, and with a resolution of 0.6 degrees, it's 1259 times faster. Travel cost estimation using CSN introduces a controllable level of uncertainty, which can be further reduced through the application of linear regression models. Two accuracy metrics, length-dependent and length-independent regression, are crucial for evaluation.

Length-dependent regression is preferable when there is a need to brute force approach. It produces robust results, but is resource intensive in comparison to the length-independent approach. A CSN with a resolution of 0.3 degrees or higher is suitable for this purpose since the error margin of length-dependent regression falls below 5%.

However, the length-independent regression approach would be the preferred choice if its challenges are addressed, such as the lack of communality between on-path CSN grid cells and regular CSN grid cells and the absence of a clear implementation method. This approach directly applies to CSN grid cells, eliminating the need for post-processing after calculating the shortest paths. The evaluation of length-independent regression shows that predictability, as indicated by the error margin, levels off at around 2% after reaching a resolution of 0.6 degrees. An additional benefit is that the already low error margin of 2% along the shortest path can offset each other if the errors are truly random.

To further enhance the accuracy of linguistic contact modelling and reduce outliers, we introduce the idea of a Influence Vector (IV). The IV serves as a tool to address outliers and quantify the influence of dominant languages, particularly within empires, during linguistic contact modelling. It leverages temporal densities of start nodes, travel costs, and high-status meaning class changes to create multi-dimensional vectors associated with contact edges. These multi-dimensional IVs contribute to the refinement of CSN accuracy, making the modelling of contact between languages even more precise.

In conclusion, the integration of THF into the contacTrees geo-prior would be a worthwhile endeavour. It significantly improves travel cost estimation to model contact between languages, over a solely geodetic approach to travel cost estimation. The use of a CSN boosts computational efficiency greatly so that using THF applied as a geo-becomes feasible. The implementation of length-independent regression and the introduction of the Influence Vector represents a research goal for future work to further improve the methodology.

5.2. Other utilities of CSNs

The ability of CSN to simplify topographies in a general way could be easily applicable to other raster based topographies, including heat maps, ocean or wind current maps, and land cover maps. Furthermore, we belief that the difference between raster topography and the simplified CSN topography can be predicted in a generalisable way, via mathematical prove. This would allow to simplify lots of different types of topographies into a CSN without proving the viability of each and every one of them.

CSNs can not only be used to calculate shortest paths, but also other network metrics such as centrality, connectivity, and betweenness. This versatility makes other kinds of simulation use cases possible:

- **Use-case 1:** Makes use of CSN by simulating territorial expansions of empires based on the contacTrees addon and by generating several influence vectors along each node path in order to determine, which languages are within de boundaries of the empire.
- **Use-case 2:** Another application could also be to simulate natural networks of movement based on ocean and wind current maps.
- **Use-case 3:** Makes use of CSN by simulating and randomly generating computer vision landscapes using the CSNs ability to represent generalised topologies in combination with directional factors through its grid cells, which could serve as condensed templates containing landscape features that allows an algorithmic reconstruction.

5.3. Creation of R-Package

In this thesis, several functions were developed within the R environment, such as contactR, shortestPathR, contactPlotteR, and CSNmakeR. These functions hold the potential to be packaged and published on CRAN with a reasonable level of effort, making them accessible to a broad community of R users and therefore ease accessibility of BEAST2 and contacTree outputs in the R environment and streamline procedures like coordinate acquisition of start and end nodes of contact edges.

5.4. Future Work

The ways to build on the findings of this thesis are manyfold. Implementing a Tobler's hiking function (THF) based geo-prior for contacTrees would be the first and obvious choice. If the implementation of the Tobler's hiking function based geo-prier is considered, solving the challenges around length-independent regression would be highly beneficial to the efficiency level of the CSN in use. There are two further implementations that would highly benefit the accuracy of the prior signal. Firstly, the influence vector and secondly, a compensation method to mitigate the offset between language position and grid cell centroid. Beside geo-prior implementation, the CSN promises to be applicable in a wide array of use cases, such as simulating empire expansions, simulating natural networks of movements or simulating computer vision landscapes. Finally, all functions developed for the R environment can be packaged into a R package for easy access for all future research endeavours regarding contacTrees.

6. References

Baldwin, R. M., Owzar, K., Zembutsu, H., Chhibber, A., Kubo, M., Jiang, C., Watson, D., Eclov, R. J., Mefford, J., McLeod, H. L., Friedman, P. N., Hudis, C. A., Winer, E. P., Jorgenson, E. M., Witte, J. S., Shulman, L. N., Nakamura, Y., Ratain, M. J., & Kroetz, D. L. (2010). R: A language and environment for statistical computing. *(No Title)*, *18*(18), 5099–5109. https://doi.org/10.1158/1078-0432.CCR-12-1590

Bouckaert, R. (2016). Phylogeography by diffusion on a sphere: Whole world phylogeography. *PeerJ*, *2016*(9), e2406. https://doi.org/10.7717/PEERJ.2406/SUPP-1

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., & Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6097), 957. https://doi.org/10.1126/science.1219669

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, *15*(4), e1006650. https://doi.org/10.1371/JOURNAL.PCBI.1006650

Campbell, M. J., Dennison, P. E., Butler, B. W., & Page, W. G. (2019). Using Crowdsourced Fitness Tracker Data to Model the Relationship Between Slope and Travel Rates. *Applied Geography*, *106*(May), 93–107. https://doi.org/10.1016/j.apgeog.2019.03.008

champers. (n.d.). *Data Overview - Berkeley Earth*. Retrieved April 4, 2023, from https://berkeleyearth.org/data/

Champers, J., Trudgill, P., & Schilling-Estes, N. (2004). The Handbook of Language Variation and Change. *The Handbook of Language Variation and Change*. https://doi.org/10.1002/9780470756591

Collischonn, W., & Pilar, J. V. (2000). A direction dependent least-cost-path algorithm for roads and canals. *International Journal of Geographical Information Science*, *14*(4), 397–406. https://doi.org/10.1080/13658810050024304

Compilation Group. (2023). *GEBCO* . https://doi.org/10.5285/f98b053b-0cbc-6c23-e053-6c86abc0af7b

Csardi, G. (2005). *The Igraph Software Package for Complex Network Research*. https://www.researchgate.net/publication/221995787

Daulton, F. E. (2019). Lexical Borrowing. *The Encyclopedia of Applied Linguistics*, 1–5. https://doi.org/10.1002/9781405198431.WBEAL0687.PUB2

Dijkstra, E. W. (2022). A Note on Two Problems in Connexion with Graphs. *Edsger Wybe Dijkstra*, 287–290. https://doi.org/10.1145/3544585.3544600

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Maintainer, A. W. (2006). *The e1071 Package*.

Dunn, M., & Tresoldi, T. (2021). *evotext/ielex-data-and-tree: IELex data and tree* . Zenodo.

Forster, P., & Toth, A. (2003). Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 9079–9084. https://doi.org/10.1073/PNAS.1331158100/SUPPL_FILE/1158TABLE2.XLS

Gilley, J., Rabinowitz, N., & Ellis, D. (n.d.). *Uber: hexagonal hierarchical spatial index (h3)*. Retrieved September 28, 2023, from https://h3geo.org/

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature 2003 426:6965*, *426*(6965), 435–439. https://doi.org/10.1038/nature02029

Heggarty, P. (2014). Prehistory by Bayesian phylogenetics? The state of the art on Indo-European origins. *Antiquity*, *88*(340), 566–577. https://doi.org/10.1017/S0003598X00101188

Heggarty, P., Anderson, C., Scarborough, M., King, B., Bouckaert, R., Jocz, L., Kümmel, M. J., Jügel, T., Irslinger, B., Pooth, R., Liljegren, H., Strand, R. F., Haig, G., MacÀk, M., Kim, R. I., Anonby, E., Pronk, T., Belyaev, O., Dewey-Findell, T. K., … Gray, R. D. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, *381*(6656). https://doi.org/10.1126/SCIENCE.ABG0818/SUPPL_FILE/SCIENCE.ABG0818_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Higgins, C. D. (2021). Hiking with Tobler: Tracking Movement and Calibrating a Cost Function for Personalized 3D Accessibility. *Findings*. https://doi.org/10.32866/001C.28107

Hock, H. H., & Joseph, B. D. (2009). Language History, Language Change, and Language Relationship. In *Language History, Language Change, and Language Relationship*. De Gruyter Mouton. https://doi.org/10.1515/9783110214307/HTML

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *BIOINFORMATICS APPLICATIONS NOTE*, *17*(8), 754–755. http://brahms.biology.rochester.edu/software.html.

Irmischer, I. J., & Clarke, K. C. (2018). Measuring and modeling the speed of human navigation. *Cartography and Geographic Information Science*, *45*(2), 177–186. https://doi.org/10.1080/15230406.2017.1292150

Koile, E., Greenhill, S. J., Blasi, D. E., Bouckaert, R., & Gray, R. D. (2022). Phylogeographic analysis of the Bantu language expansion supports a rainforest route. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(32), e2112853119. https://doi.org/10.1073/PNAS.2112853119/SUPPL_FILE/PNAS.2112853119.SAPP.PDF

Kymlicka, W. (2020). Multicultural Citizenship. In *The New Social Theory Reader* (pp. 270–280). Routledge. https://doi.org/10.4324/9781003060963-44

Márquez-Pérez, J., Vallejo-Villalta, I., & Álvarez-Francoso, J. I. (2017). Estimated travel time for walking trails in natural areas. *Geografisk Tidsskrift - Danish Journal of Geography* , *117*(1), 53–62. https://doi.org/10.1080/00167223.2017.1316212

Neureiter, N., Ranacher, P., Efrat-Kowalsky, N., Kaiping, G. A., Weibel, R., Widmer, P., & Bouckaert, R. R. (2022). Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer. *Humanities and Social Sciences Communications 2022 9:1*, *9*(1), 1–14. https://doi.org/10.1057/s41599-022-01211-7

Neureiter, N., Ranacher, P., Van Gijn, R., Bickel, B., & Weibel, R. (2021). Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? *Royal Society Open Science*, *8*(1). https://doi.org/10.1098/RSOS.201079

Nordhoff, S., & Hammarström, H. (2011). *Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources*. http://glottolog.livingsources.org

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/BIOINFORMATICS/BTY633

Pingel, T. J. (2010). Modeling Slope as a Contributor to Route Selection in Mountainous Areas. *Cartography and Geographic Information Science*, *37*(2), 137–148. https://doi.org/10.1559/152304010791232163

Ranacher, P., Neureiter, N., Van Gijn, R., Sonnenhauser, B., Escher, A., Weibel, R., Muysken, P., & Bickel, B. (2021). Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. *Journal of the Royal Society Interface*, *18*(181). https://doi.org/10.1098/RSIF.2020.1031

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574. https://doi.org/10.1093/BIOINFORMATICS/BTG180

Tamura, K., Stecher, G., & Kumar, S. (2021a). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/MOLBEV/MSAB120

Tamura, K., Stecher, G., & Kumar, S. (2021b). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/MOLBEV/MSAB120

Tobler, W. (1993). *THREE PRESENTATIONS ON GEOGRAPHICAL ANALYSIS AND MODELING NON-ISOTROPIC MODELING SPECULATIONS ON THE GEOMETRY OF GEOGRAPHY GLOBAL SPATIAL ANALYSIS*.

van der Meer, L., Abad, L., Gilardi, A., & Lovelace, R. (2023). *sfnetworks: Tidy Geospatial Networks*. Https://Luukvdmeer.Github.Io/Sfnetworks/, Https://Github.Com/Luukvdmeer/Sfnetworks.

Wang, L. G., Lam, T. T. Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C. W., Jones, B. R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y., & Yu, G. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution*, *37*(2), 599–603. https://doi.org/10.1093/MOLBEV/MSZ240

Wickham, H. (2016). ggpolt2 Elegant Graphics for Data Analysis. *Use R! Series*, 211.

Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., Dai, Z., Lam, T. T., Guan, Y., & Yu, G. (2022). Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *IMeta*, *1*(4), e56. https://doi.org/10.1002/IMT2.56

Yu, G. (2022). Data Integration, Manipulation and Visualization of Phylogenetic Trees. *Data Integration, Manipulation and Visualization of Phylogenetic Trees*. https://doi.org/10.1201/9781003279242

Acknowledgements

My sincerest thanks go out to all my family and friends, who supported me along this journey. Having started over a year ago, working on this Master's thesis wasn't always easy for me. But in the end everything worked out. My first personal thank you goes out to my supervisors Nico Neureiter and Peter Ranacher, who helped me with their competent guidance. Furthermore, a special thanks goes out to my friend Luc Sierro, who helped me with this Master's thesis greatly. I also what to thank my parents – Thomas Zurmühle and Anna Diener, personally – who supported me always in my years of study. My biggest thanks goes to the love of my life, Ani Matevosyan. You always have my back and help my with your insightful mind and kind demeaner. I will always be grateful for all your help.

List of Tools

Tool Name	Background	Use Case	Example
Speechify	Text to Speech	Checking readability	By turning the text into an audio-book readability can be checked much more easily.
Mendeley	Source management	-Gathering sources -Managing sources	AI aided source-recommendations of Mendeley help a lot. It also has a great implementation into word.
DeepL	Translation	-Checking phrasing -Translations	Translating German phrases to English. Making non-English or non-German sources readable.
ChatGPT	Large Language Modul	-Combating writhers-block -spell, grammar and style checking	When experiencing writers-block, this software can help greatly by producing transitions or introductions. Checking spelling or grammar is also a big advantage.

Appendix

Appendix I: BEAST2 setup:

The XML input file for BEAST2 is setup the following way:

- 37 Alignments
- Non state parameters:
    - Branch rate model: Relaxed log normal clock with frozen branches
        - With slow, medium and fast clock rate settings
    - Substitution model: Binary Covarion (dimension="2", lower="0.0", upper="1.0">0.5 0.5<, frequencies="0.5 0.5")
    - Block set: stateNote set up:
        - Network (tree) is a prior
        - Plate: Concepts (ContacTree parameter)
- MCMC set up:
    - Chain length of 20’000’000; Logged every 5000.
    - State Parameters
        - Network: Newick structure of the whole tree, with 37 tips and set date trait (tip height).
        - Tree priors (birth-death rates)
            - Birth rate set to 0.0005, sampling proportion 0.2
            - Death rate set to 0.3, sampling proportion 0.5
        - Clock model (relaxed log-normal)
            - Slow: Upper limit 10, start parameter $2*10^{-5}$
            - Medium: Upper limit 10, start parameter $5*10^{-5}$
            - Fast: Upper limit 10, start parameter $8*10^{-5}$
            - Standard deviation set between 0 and 0.5
        - Substitution model (covarion)
            - Frequency with 2 dimensions set between 0 and 0.9 resp. 0.1
            - Covarion switch rate set between 10^{-5} and 0.1
        - Contactrees parameters:
            - Expected conversion of concepts is estimated to be over 0.25: (ContacTree)
            - Conversion rate set to linear contact growth and includes the tree structure (netwark) : (ContacTree)
            - Movement parameter (pMove) is set between 0 and 0.4: (ContacTree)
        - GeoSphere parameters: (GeoSphere)
            - Geo clock rate set to 0.1: (GeoSphere)
            - Standard deviation between 0 and 2; Start position 0.1: (GeoSphere)
            - Rate categories: (GeoSphere)
    - Bayesian Model
        - Posterior distribution is specified as a compounding distribution
        - Model Priors:
            - ACG prior
                - Expected-Conversions distribution

        - pMove distribution
        - Birth-Rate-Prior: Log-Normal-Distribution-Model
        - Death-Rate-Prior: beast-math-distributions-Uniform
        -
    - Clock model (relaxed log-normal)
        - Best-Math-Uniform distributions
    - Clock model (GeoSphere):
        - Clock-Prior distribution: (GeoSphere)
        - Standard-Deviation-Prior distribution: (GeoSphere)
    - Substitution model (covarion):
        - Covarion-Alpha distribution: (ContacTree)
        - Covarion-Switch-Rate distribution: (ContacTree)
    - Topology priors (MRCAPrior)
        - Definition of 10 monophyletic sup trees (TaxonSet) within the main tree
- Model Likelihood
    - Conceps
        - Tree likelihoods of concepts: (ContacTree)
    - Geo Sphere
        - Tip Locations: (GeoSphere)
- Model Operators
    - Tree
        - WilsonBalding-Operator
        - Subtree Exchange-Operator: Narrow
        - Subtree Exchange-Operator: Wide
        - ACGScaler
        - Birth-Rate-Scale-Operator
        - Death-Rate-Scale-Operator
    - contacTree
        - Add-Remove-Conversion-Gibbs-Operator: (ContacTree)
        - Gibbs-Sample-Moves-Per-Conversion-Operator: (ContacTree)
        - Converted-Edge-Slide-Operator: (ContacTree)
        - Converted-Edge-Flip-Operator: (ContacTree)
        - Converted-Edge-Split-Operator: (ContacTree)
        - Converted-Edge-Hop-Operator: (ContacTree)
        - Converted-Edge-Hop-Narrow-Gibbs-Operator: (ContacTree)
    - Clock rate
        - Clock-Rate-Scale-Operator: Slow, Medium, Fast
        - Categorical-Random-Walk-Operator
        - Categorical-Swap-Operator
        - Categorical-Uniform-Operator
    - Substitution Model
        - Freq-Parameter-Sample-Operator: (ContacTree)
        - Covarion-Alpha-Scale-Operator: (ContacTree)
        - Covarion-Switch-Rate-Scale-Operator: (ContacTree)
    - GeoShpere model

         - Relaxed-Clock-Rate-Scale-Operator: (GeoSphere)
         - Standard-Deviation-Scale-Operator: (GeoSphere)
         - Categories-Random-Walk-Operator: (GeoSphere)
         - Categories-Swap-Operator-Operator: (GeoSphere)
         - Categories-Uniform-Operator: (GeoSphere)
- Loggers
   - Trace loggers: Log every 5000
      - Posterior
      - Likelihood
      - Prior
      - ACGStatsLogger
      - conversionRate: (ContacTree)
      - pMove: (ContacTree)
      - clock rates
         - slow
         - medium
         - fast
         - standard deviation
      - freqParameter
      - conversionAlpha: (ContacTree)
      - conversionSwitchRate: (ContacTree)
      - birthrate
      - deathrate
      - samplingPropotion
      - location likelihood: (GeoSphere)
      - precision: (GeoSphere)
      - clock rate: (GeoSphere)
      - standard deviation: (GeoSphere)

Appendix II: Cost Surface Network (CSN) Evaluation

2.5 Degree Resolution

```
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "INPUT FILE: CSN25"
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] ""
[1] ""
[1] "time to calculate 132 shortest paths:  0.600240230560303"
[1] "mean pace in network:  0.714755797266881"
[1] "median pace in network:  0.714697365178182"
[1] "max pace in network:  0.747127696344697"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length independent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMpace ~ CSNpace)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.06622 -0.02828 -0.01286  0.01495  0.19784 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.72834    0.01294   56.29  < 2e-16 ***
CSNpace      0.04057    0.01389    2.92  0.00419 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04434 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.06685, Adjusted R-squared:  0.05901 
F-statistic: 8.525 on 1 and 119 DF,  p-value: 0.004192

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length independent Results with DoTM"

Call:
lm(formula = DTMpace ~ CSNpace + CSNdataSDH)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.07558 -0.02513 -0.01349  0.01396  0.19767 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.198e-01  1.706e-02  42.180  < 2e-16 ***
CSNpace     2.803e-02  1.722e-02   1.628  0.10631    
CSNdataSDH  4.211e-05  1.588e-05   2.652  0.00913 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04376 on 114 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared:  0.08139, Adjusted R-squared:  0.06527 
F-statistic:  5.05 on 2 and 114 DF,  p-value: 0.007916

[1] ""
[1] ""
[1] "--------------------------------------------------"
```

```
[1] "RMSE length independent [sec/m]"
[1] "mean pace [sec/m] 0.930836064760761"
[1] "sd pace [sec/m] 0.328754655669206"
[1] "CSN vs DTM direct relation:    0.306574292621316"
[1] "CSN vs DTM regression:         0.0439711457781441"
[1] "CSN vs DTM + DoTM regression:  0.0418043834121898"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length dependent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMdata ~ CSNdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.453 -13.523  -1.662  14.352  46.047 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 27.25832    3.98058   6.848  3.5e-10 ***
CSNdata      0.58899    0.03529  16.690  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.86 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.7007,	Adjusted R-squared:  0.6982 
F-statistic: 278.6 on 1 and 119 DF,  p-value: < 2.2e-16

[1] ""
[1] "--------------------------------------------------"
[1] "length dependent Results with DoTM"

Call:
lm(formula = DTMdata ~ CSNdata + log(CSNdataSDH))

Residuals:
    Min      1Q  Median      3Q     Max 
-41.828 -12.925  -1.743  15.240  44.230 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     11.46294   21.35215   0.537    0.592    
CSNdata          0.59903    0.03546  16.894   <2e-16 ***
log(CSNdataSDH)  2.61298    3.48136   0.751    0.454    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.17 on 114 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared:  0.7147,	Adjusted R-squared:  0.7097 
F-statistic: 142.8 on 2 and 114 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean travel time [hours] 108.357689513732"
[1] "sd travel time [hours] 56.9320302843027"
[1] "RMSE length dependent [hours]"
[1] "CSN vs DTM direct relation:    32.0555082867825"
[1] "CSN vs DTM regression:         19.692867560721"
[1] "CSN vs DTM + DoTM regression:  18.5354989473468"
```

1.2 Degree Resolution

```
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
```

```
[1] "INPUT FILE: CSN12"
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] ""
[1] ""
[1] "time to calculate 132 shortest paths:  1.81376385688782"
[1] "mean pace in network:  0.714841209193468"
[1] "median pace in network:  0.714707937341429"
[1] "max pace in network:  0.782843105499079"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length independent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMpace ~ CSNpace)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.05885 -0.03168 -0.01067  0.01855  0.18286 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.69970    0.03014  23.216   <2e-16 ***
CSNpace      0.08145    0.03768   2.161   0.0327 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04502 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.03778, Adjusted R-squared:  0.02969 
F-statistic: 4.672 on 1 and 119 DF,  p-value: 0.03267

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length independent Results with DoTM"

Call:
lm(formula = DTMpace ~ CSNpace + CSNdataSDH)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.073365 -0.019027 -0.001385  0.012733  0.109563 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.249e-01  2.132e-02  29.309  < 2e-16 ***
CSNpace     9.104e-02  2.549e-02   3.572 0.000514 ***
CSNdataSDH  1.339e-04  1.122e-05  11.931  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03044 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.5639,	Adjusted R-squared:  0.5565 
F-statistic: 76.29 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean pace [sec/m] 0.78922301752531"
[1] "sd pace [sec/m] 0.106490592230815"
[1] "RMSE length independent [sec/m]"
[1] "CSN vs DTM direct relation:    0.112876265044036"
[1] "CSN vs DTM regression:         0.0446508481733748"
[1] "CSN vs DTM + DoTM regression:  0.0300596029046704"
```

```
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length dependent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMdata ~ CSNdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.2067  -7.2139   0.4825   7.1858  23.4562 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.06158    2.37580   2.972  0.00358 ** 
CSNdata      0.88622    0.02432  36.439  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.41 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9177,	Adjusted R-squared:  0.9171 
F-statistic:  1328 on 1 and 119 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length dependent Results with DoTM"

Call:
lm(formula = DTMdata ~ CSNdata + log(CSNdataSDH))

Residuals:
     Min       1Q   Median       3Q      Max 
-21.5634  -5.9760   0.6129   5.7010  31.2034 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -32.90274   11.62534  -2.830 0.005468 ** 
CSNdata           0.88889    0.02326  38.222  < 2e-16 ***
log(CSNdataSDH)   6.50870    1.85687   3.505 0.000646 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.948 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9255,	Adjusted R-squared:  0.9242 
F-statistic:   733 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean travel time [hours] 91.217142361467"
[1] "sd travel time [hours] 39.5056777096461"
[1] "RMSE length dependent [hours]"
[1] "CSN vs DTM direct relation:    11.6609561012956"
[1] "CSN vs DTM regression:         10.3230038638968"
[1] "CSN vs DTM + DoTM regression:  9.82420763602292"
```

0.9 Degree Resolution

```
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "INPUT FILE: CSN09"
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] ""
[1] ""
[1] "time to calculate 132 shortest paths:  3.39749193191528"
[1] "mean pace in network:  0.714888719197621"
[1] "median pace in network:  0.714747729967415"
[1] "max pace in network:  0.90169451861757"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length independent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMpace ~ CSNpace)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.052811 -0.020922 -0.009056  0.012312  0.166137 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.60452    0.01982  30.497  < 2e-16 ***
CSNpace      0.20161    0.02466   8.174 3.66e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03673 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.3596,	Adjusted R-squared:  0.3542 
F-statistic: 66.82 on 1 and 119 DF,  p-value: 3.655e-13

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length independent Results with DoTM"

Call:
lm(formula = DTMpace ~ CSNpace + CSNdataSDH)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.055355 -0.014530 -0.002939  0.013552  0.120161 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.350e-01  1.310e-02  48.493  < 2e-16 ***
CSNpace     1.011e-01  1.784e-02   5.669 1.03e-07 ***
CSNdataSDH  7.969e-05  6.223e-06  12.806  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02386 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.732,	Adjusted R-squared:  0.7275 
F-statistic: 161.2 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean pace [sec/m] 0.810778544008768"
[1] "sd pace [sec/m] 0.148984233302909"
```

```
[1] "RMSE length independent [sec/m]"
[1] "CSN vs DTM direct relation:    0.117457274191241"
[1] "CSN vs DTM regression:         0.0364264098884822"
[1] "CSN vs DTM + DoTM regression:  0.023563355635351"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length dependent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMdata ~ CSNdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-29.1317  -5.1417   0.6761   6.3033  21.3438 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.46439    2.06494   4.099 7.61e-05 ***
CSNdata      0.86208    0.02084  41.365  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.255 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.935,	Adjusted R-squared:  0.9344 
F-statistic:  1711 on 1 and 119 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length dependent Results with DoTM"

Call:
lm(formula = DTMdata ~ CSNdata + log(CSNdataSDH))

Residuals:
     Min       1Q   Median       3Q      Max 
-26.5643  -5.1972  -0.4468   6.5403  20.8626 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     22.50993    9.51009   2.367   0.0196 *  
CSNdata          0.86866    0.02118  41.011   <2e-16 ***
log(CSNdataSDH) -2.33914    1.54643  -1.513   0.1331    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.206 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9362,	Adjusted R-squared:  0.9351 
F-statistic: 865.9 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean travel time [hours] 94.5959877405778"
[1] "sd travel time [hours] 42.3024593104835"
[1] "RMSE length dependent [hours]"
[1] "CSN vs DTM direct relation:    11.4618822035774"
[1] "CSN vs DTM regression:         9.17848177172512"
[1] "CSN vs DTM + DoTM regression:  9.09077155184226"
```

0.6 Degree Resolution

```
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "INPUT FILE: CSN06"
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] ""
[1] ""
[1] "time to calculate 132 shortest paths:  5.81123399734497"
[1] "mean pace in network:  0.714997066324649"
[1] "median pace in network:  0.714747729967415"
[1] "max pace in network:  0.97595697028286"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length independent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMpace ~ CSNpace)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.08292 -0.02300 -0.01103  0.01269  0.17937 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.5910     0.0386  15.310  < 2e-16 ***
CSNpace       0.2116     0.0469   4.511 1.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04242 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.146,	Adjusted R-squared:  0.1388 
F-statistic: 20.35 on 1 and 119 DF,  p-value: 1.526e-05

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length independent Results with DoTM"

Call:
lm(formula = DTMpace ~ CSNpace + CSNdataSDH)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.062095 -0.013766 -0.004053  0.014522  0.061382 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.330e-01  1.914e-02  33.068  < 2e-16 ***
CSNpace     8.231e-02  2.406e-02   3.421 0.000858 ***
CSNdataSDH  1.092e-04  5.661e-06  19.294  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0209 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.7945,	Adjusted R-squared:  0.791 
F-statistic:   228 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean pace [sec/m] 0.830763130683299"
[1] "sd pace [sec/m] 0.0913462252532058"
```

```
[1] "RMSE length independent [sec/m]"
[1] "CSN vs DTM direct relation:    0.0946530038680978"
[1] "CSN vs DTM regression:         0.0420648069890975"
[1] "CSN vs DTM + DoTM regression:  0.0206369343633275"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length dependent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMdata ~ CSNdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.2801  -4.3437  -0.1677   4.3443  17.2934 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.39336    1.76937   2.483   0.0144 *  
CSNdata      0.88460    0.01753  50.469   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.668 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9554,	Adjusted R-squared:  0.955 
F-statistic:  2547 on 1 and 119 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length dependent Results with DoTM"

Call:
lm(formula = DTMdata ~ CSNdata + log(CSNdataSDH))

Residuals:
     Min       1Q   Median       3Q      Max 
-22.2436  -5.4472   0.0809   4.2211  15.8924 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -18.4886     8.6526  -2.137  0.03468 *  
CSNdata           0.8825     0.0171  51.608  < 2e-16 ***
log(CSNdataSDH)   3.7009     1.3714   2.699  0.00798 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.473 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.958,	Adjusted R-squared:  0.9572 
F-statistic:  1344 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean travel time [hours] 96.2821241351433"
[1] "sd travel time [hours] 41.4751392344622"
[1] "RMSE length dependent [hours]"
[1] "CSN vs DTM direct relation:    10.8973098923315"
[1] "CSN vs DTM regression:         7.60446107860926"
[1] "CSN vs DTM + DoTM regression:  7.38012671677678"
```

0.3 Degree Resolution

```
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "INPUT FILE: CSN03"
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] ""
[1] ""
[1] "time to calculate 132 shortest paths:  34.770054101944"
[1] "mean pace in network:  0.715278327042574"
[1] "median pace in network:  0.714747729967415"
[1] "max pace in network:  1.19816367515049"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length independent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMpace ~ CSNpace)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.05302 -0.03061 -0.01351  0.01638  0.21565 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.70536    0.06012  11.733   <2e-16 ***
CSNpace      0.07364    0.07501   0.982    0.328    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04572 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.008034, Adjusted R-squared:  -0.0003017 
F-statistic: 0.9638 on 1 and 119 DF,  p-value: 0.3282

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length independent Results with DoTM"

Call:
lm(formula = DTMpace ~ CSNpace + CSNdataSDH)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.046166 -0.014926 -0.003769  0.013438  0.118869 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.223e-01  3.094e-02  20.112   <2e-16 ***
CSNpace     9.382e-02  3.822e-02   2.455   0.0156 *  
CSNdataSDH  1.061e-04  5.745e-06  18.462   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02328 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.7449,	Adjusted R-squared:  0.7406 
F-statistic: 172.3 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean pace [sec/m] 0.804106638884878"
[1] "sd pace [sec/m] 0.0566800547783684"
```

```
[1] "RMSE length independent [sec/m]"
[1] "CSN vs DTM direct relation:    0.0770161037787525"
[1] "CSN vs DTM regression:         0.0453356459089329"
[1] "CSN vs DTM + DoTM regression:  0.0229899964363192"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length dependent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMdata ~ CSNdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.3308  -4.7166  -0.8812   3.4147  16.4300 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.46656    1.41630   2.448   0.0158 *  
CSNdata      0.91104    0.01429  63.734   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.123 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9715,  Adjusted R-squared:  0.9713 
F-statistic:  4062 on 1 and 119 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length dependent Results with DoTM"

Call:
lm(formula = DTMdata ~ CSNdata + log(CSNdataSDH))

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0888  -4.2908  -0.3703   3.6252  13.5576 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -38.64721    5.87956  -6.573 1.40e-09 ***
CSNdata           0.90202    0.01197  75.343  < 2e-16 ***
log(CSNdataSDH)   6.80435    0.93064   7.311 3.44e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.101 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9804,  Adjusted R-squared:  0.9801 
F-statistic:  2953 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean travel time [hours] 93.6422828870308"
[1] "sd travel time [hours] 39.9074126373987"
[1] "RMSE length dependent [hours]"
[1] "CSN vs DTM direct relation:    8.39023066576254"
[1] "CSN vs DTM regression:         6.0724445675057"
[1] "CSN vs DTM + DoTM regression:  4.98362611980951"
```

0.1 Degree Resolution

```
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "INPUT FILE: CSN03"
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] ""
[1] ""
[1] "time to calculate 132 shortest paths:  34.770054101944"
[1] "mean pace in network:  0.715278327042574"
[1] "median pace in network:  0.714747729967415"
[1] "max pace in network:  1.19816367515049"
[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "--------------------------------------------------"
[1] "length independent Results"
[1] "--------------------------------------------------"

Call:
lm(formula = DTMpace ~ CSNpace)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.05302 -0.03061 -0.01351  0.01638  0.21565 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.70536    0.06012  11.733   <2e-16 ***
CSNpace      0.07364    0.07501   0.982    0.328    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04572 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.008034,Adjusted R-squared:  -0.0003017 
F-statistic: 0.9638 on 1 and 119 DF,  p-value: 0.3282

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "length independent Results with DoTM"

Call:
lm(formula = DTMpace ~ CSNpace + CSNdataSDH)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.046166 -0.014926 -0.003769  0.013438  0.118869 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.223e-01  3.094e-02  20.112   <2e-16 ***
CSNpace     9.382e-02  3.822e-02   2.455   0.0156 *  
CSNdataSDH  1.061e-04  5.745e-06  18.462   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02328 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.7449,   Adjusted R-squared:  0.7406 
F-statistic: 172.3 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "--------------------------------------------------"
[1] "mean pace [sec/m] 0.804106638884878"
[1] "sd pace [sec/m] 0.0566800547783684"
```

```
[1] "RMSE length independent [sec/m]"
[1] "CSN vs DTM direct relation:    0.0770161037787525"
[1] "CSN vs DTM regression:         0.0453356459089329"
[1] "CSN vs DTM + DoTM regression:  0.0229899964363192"
[1] ""
[1] ""
[1] "-------------------------------------------------"
[1] "-------------------------------------------------"
[1] "length dependent Results"
[1] "-------------------------------------------------"

Call:
lm(formula = DTMdata ~ CSNdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.3308  -4.7166  -0.8812   3.4147  16.4300 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.46656    1.41630   2.448   0.0158 *  
CSNdata      0.91104    0.01429  63.734   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.123 on 119 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9715,	Adjusted R-squared:  0.9713 
F-statistic:  4062 on 1 and 119 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "-------------------------------------------------"
[1] "length dependent Results with DoTM"

Call:
lm(formula = DTMdata ~ CSNdata + log(CSNdataSDH))

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0888  -4.2908  -0.3703   3.6252  13.5576 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -38.64721    5.87956  -6.573 1.40e-09 ***
CSNdata           0.90202    0.01197  75.343  < 2e-16 ***
log(CSNdataSDH)   6.80435    0.93064   7.311 3.44e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.101 on 118 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.9804,	Adjusted R-squared:  0.9801 
F-statistic:  2953 on 2 and 118 DF,  p-value: < 2.2e-16

[1] ""
[1] ""
[1] "-------------------------------------------------"
[1] "mean travel time [hours] 93.6422828870308"
[1] "sd travel time [hours] 39.9074126373987"
[1] "RMSE length dependent [hours]"
[1] "CSN vs DTM direct relation:    8.39023066576254"
[1] "CSN vs DTM regression:         6.0724445675057"
[1] "CSN vs DTM + DoTM regression:  4.98362611980951"
```

Personal Declaration

I hereby declare that the submitted Thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the Thesis.

In accordance with the MNF's guidelines on AI-tools, this thesis exhibits responsible use of tools, AI or otherwise, by listing the tools used and what they were used for.