MASTERARBEIT | MASTER'S THESIS

Titel | Title

Quantifying the Relationship Between Digitization and Linguistic Diversity
A Multilevel Statistical Analysis Using R

verfasst von | submitted by

Katharina Zeh BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Ass.-Prof. Mag. Mag. Dr. Andreas Baumann

Acknowledgements

I would like to express my deepest gratitude to all those amazing people, without whom I would not have been able to complete this thesis. First and foremost, I want to thank my boyfriend Christopher, who managed to keep me sane with his emotional support, patience and lots of pots of coffee, my dear friend Sabine and her son Levi for helping me keep a healthy balance between my academic and personal life and for lifting my spirits even throughout the hardest times, my other dear friend Christina for all the walking and talking around (Alte) Donau, and my brother Lukas for always believing in me.

I would like to extend my thanks to Lale for helping me lay the foundations of this thesis through our previous joint project and thanks again to her and Max for proofreading my thesis and giving me valuable feedback. Last but not least, I want to thank my supervisor Andreas Baumann for his guidance, and for creating an environment where I never had to be afraid to ask a question (and also for the Mannerschnitten and Topfenbällchen), and the whole DIGILINGDIV project team, namely Hannes E., Hannes F., Julia and Juliane for the warm environment, their input, and support.

Abstract

Linguistic diversity has declined significantly worldwide over the past few decades. This trend is expected to continue unless effective countermeasures are implemented, making it crucial to identify and analyse its underlying drivers. To contribute to this effort, this thesis examines the impact of digitization on global language diversity and endangerment—a relationship that has remained relatively underexplored in previous research. The investigation is carried out using a statistical multi-level approach implemented in the *r* programming language. Drawing on digitization indices such as the Mobile Connectivity Index, the Digital Adoption Index, and the E-Government Development Index, broader patterns and groupings are identified through cluster analysis, complemented by a correlation-based evaluation of the conceptual breadth of these indices. Subsequently, relationships between digitization and linguistic diversity—specifically, the number of languages spoken in each country, entropy-based measures, and an adapted Red List Index (RLI)—are explored through correlation analysis. The results suggest that while digitization has minimal effects on language endangerment and richness, it significantly influences the distribution of languages at the national level. These findings offer a valuable foundation for future research and may inform policymakers and language activists in efforts to preserve and promote linguistic diversity.

Kurzfassung

Die weltweite Sprachdiversität ist in den vergangenen Jahrzehnten deutlich zurückgegangen. Es ist zu erwarten, dass sich dieser Trend fortsetzt, sofern keine wirksamen Gegenmaßnahmen ergriffen werden. Deshalb ist es umso wichtiger die zugrundeliegenden Faktoren hinter dieser Entwicklung zu identifizieren und zu analysieren. Einen Beitrag dazu leistet diese Arbeit, indem sie den bisher kaum erforschten Einfluss der Digitalisierung auf die globale Sprachdiversität und Sprachgefährdung untersucht. Die Analyse erfolgt mittels eines statistischen Mehrebenenansatzes, der in der Programmiersprache R durchgeführt wird. Auf Grundlage von Digitalisierungsindizes wie dem Mobile Connectivity Index, dem Digital Adoption Index und dem E-Government Development Index werden mithilfe von Clusteranalysen übergeordnete Muster und Gruppierungen identifiziert. Ergänzend wird der konzeptuelle Umfang dieser Indizes durch eine korrelationsbasierte Analyse untersucht. Anschließend wird der Zusammenhang zwischen Digitalisierung und Sprachdiversität anhand verschiedener Sprachdiversitätsmaße – der Anzahl der in einem Land gesprochenen Sprachen, eines auf Entropie basierenden Maßes und eines adaptierten Red List Index (RLI) – mittels einer Korrelationsanalyse analysiert. Die Ergebnisse der Arbeit deuten darauf hin, dass Digitalisierung nur geringe Auswirkungen auf Sprachgefährdung und sprachliche Vielfalt im engeren Sinne hat, jedoch einen signifikanten Einfluss auf die Verteilung von Sprachen innerhalb eines Landes ausübt. Diese Erkenntnisse bieten eine wertvolle Grundlage für zukünftige Forschung und können Entscheidungsträger:innen sowie Sprachaktivist:innen bei der Erhaltung und Förderung von Sprachdiversität unterstützen.

Table of Contents

List of Figures i

List of Tables i

1 Introduction 1

1.1 Problem Statement 2

1.2 Research Questions 3

1.3 Roadmap 3

2 Literature Review 5

2.1 Linguistic Diversity and Language Endangerment 5

2.2 Measuring Digitization 9

2.3 Digitization and its Societal Impact 10

2.4 The Intersection of Digitization and Linguistic Diversity 11

3 Data 15

3.1 Digitization Indices 15

3.2 Language Diversity Measures 23

3.2.1 Number of Languages per Country 23

3.2.2 Entropy-based Language Diversity 24

3.2.3 Adapted Red List Index 26

4 Methodology 28

4.1 Data Preparation 28

4.2 Analytical Approach 30

4.3 Cluster Analysis 31

4.3.1 Dataset Construction 31

4.3.2 Correlation Computation 32

4.3.3 Agglomerative Hierarchical Clustering 34

4.3.4 Cluster Validation 35

4.3.5 Cluster Visualization 36

4.4 Breadth of Indices 37

4.5 Correlation Analysis 38

4.5.1 Dataset Construction and Correlation Computation 39

4.5.2 Visualization 40

5 Results 42

5.1 Cluster Analysis 42

5.2 Breadth of Indices Analysis 51

5.2.1 Breadth Analysis based on Sub-Indices 51
5.2.2 Breadth Analysis based on Indicators 53
5.3 Correlation Analysis 55
5.3.1 Correlation Analysis - Entropy 56
5.3.2 Correlation Analysis - Number of Languages 60
5.3.3 Correlation Analysis - RLI 62
5.3.4 Summary of Correlation Results 66
6 Conclusion 68
7 Limitations and Outlook 71
8 References 72
A Appendix 80
A.1 Code Availability 80
A.2 Plots and Tables 80
A.2.1 QQ-Plots with reference line for each index 80
A.2.2 Correlation Analysis - Entropy 84
A.2.3 Correlation Analysis - RLI 91
A.2.4 Correlation Analysis - Language Count 99
A.2.5 Cluster Analysis - Complete Observations 109
A.2.6 Cluster Analysis - Pairwise Complete Observations 114

List of Figures

Figure 1 Global Linguistic Diversity by Number of Languages. 24
Figure 2 Global Linguistic Diversity by Entropy. 25
Figure 3 Global Linguistic Diversity by the Adapted RLI. 27
Figure 4 QQ Plots by Year for the MCI. 30
Figure 5 Hierarchical Clustering Dendrograms and Validation Metrics (2008, 2010, 2012). 43
Figure 6 Hierarchical Clustering Dendrograms Using Average Linkage (2014, 2019). 44
Figure 7 Silhouette Plots (2014, 2019). 45
Figure 8 WSS and Gap Statistic (2015). 46
Figure 9 Coloured Hierarchical Clustering Dendrograms (2015, 2017). 46
Figure 10 Dendrogram and Dunn Index (2016). 47
Figure 11 Hierarchical Clustering Dendrograms (2018). 48
Figure 12 Hierarchical Clustering Dendrograms (2020). 49
Figure 13 Correlations Over Time (Sub-Indices). 52
Figure 14 Correlations Over Time (Indicators). 55
Figure 15 Entropy: Pairwise Correlation Plot (2010). 57
Figure 16 Entropy: Correlation Heatmaps (2020, 2022). 58
Figure 17 Entropy: Heatmap (2004-2023). 59
Figure 18 Language Count: Pairwise Correlation Plot (2014). 61
Figure 19 Language Count: Heatmap (2004-2023). 62
Figure 20 RLI: Pairwise Correlation Plots (2008, 2012). 63
Figure 21 RLI: Heatmap (2004-2023). 66

List of Tables

Table 1 Summary of Digitization Indices. 16
Table 2 Countries Covered by Index and Year. 32
Table 3 Correlation Matrices (2015, 2017). 46
Table 4 Correlation Matrix (2018). 48
Table 5 Entropy: Correlation Analysis Summary Table (2004). 56
Table 6 Language Count: Correlation Analysis Summary Table (2004). 60
Table 7 RLI: Pearson's Correlation Coefficients and *P*-values (2014, 2016). 64

1 Introduction

Linguistic diversity, the variety and distribution of languages across the globe, is an integral part of cultural identity and heritage. This makes it all the more concerning that, of the approximately 7.000 languages spoken globally, nearly half are currently endangered, with the most pessimistic predictions estimating a 90% loss of languages by the end of the century (Bromham et al., 2021).

The rate of language loss is continuously increasing. About 700 languages have gone extinct since 1900; one language is lost every 40 days since 2000. Without intervention, estimates suggest that between 1,500 and 3,000 languages could vanish by the end of the century, although more extreme projections point to an even greater loss (Hutson et al., 2024). Language endangerment has been quantified using measures such as the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis & Simons, 2010) or the adapted Red List Index (RLI) (Baumann et al., 2024; Harmon & Loh, 2014), which highlight the severity of this decline. Although previous research has shown that the mechanisms driving language loss are complex and interconnected, most studies have primarily focused on economic, socio-political, and environmental factors influencing linguistic diversity (Bouckaert et al., 2022; Bromham et al., 2021; Gavin et al., 2013). In contrast, the impact of digitization—despite its profound transformation of global societal landscapes—remains largely unexplored.

To fully grasp digitization's role in language endangerment, it is important to first consider its broader influence on global society. Digitization serves as a powerful driver of globalization, which in turn exacerbates existing inequalities (Billon, 2010). To better understand and address these imbalances, various digitization measures, such as the Mobile Connectivity Index (MCI) (GSMA Intelligence, 2024), Digital Adoption Index (DAI) (World Bank, 2016), and E-Government Development Index (EGDI) (United Nations Department of Economic and Social Affairs, 2024), have been developed to quantify and track the uneven reach and impact of technological advancements.

These technological disparities extend into the linguistic realm, where a significant *digital language divide* has emerged (Mikami & Kodama, 2012). Blasi et al. (2022) document the dominance of global languages, particularly English, in digital spaces, which diminishes the visibility and practicality of lesser-used languages. They show how most of the world's languages remain underrepresented online, reinforcing the marginalization of these languages in both digital and physical contexts. As a contrast to these challenges, digitization offers significant opportunities for language preservation. Digital tools such as digital platforms, mobile apps, cloud computing, social media, and AI can support the documentation and revitalization of endangered languages and have the potential to help mitigate the risks posed by digitization while promoting its benefits (Alizai, 2021; Olaare, 2024).

The dual role of digitization—as both a risk and a resource for linguistic diversity—underscores the importance of thoroughly analysing its impact and direction. This thesis focuses on the relationship between digitization and linguistic diversity, examining how digitization metrics interact with linguistic diversity measures. To provide a structured foundation for this analysis, the thesis will first organize digitization indices into meaningful groups using cluster analysis. This step will then be followed by an examination of correlations between these indices and linguistic diversity measures, offering nuanced insights into how digitization influences linguistic diversity. Together these analyses contribute to a broader understanding of the ways digitization influences linguistic diversity.

1.1 Problem Statement

Although linguistic diversity is in rapid decline and digitization has become a transformative societal force, the relationship between these two phenomena remains underexplored. Existing research has identified socio-political and economic factors driving language loss, and the societal impacts of digitization—such as its influence on inequality, cultural dynamics, and technological access—have also been widely studied. However, little attention has been paid to the intersection of these domains, particularly the dual role of digitization as both a risk to and a resource for linguistic diversity. This thesis addresses this gap by examining the interaction between digitization indices and linguistic diversity measures, providing a quantitative analysis of how technological advances influence linguistic diversity and endangerment.

This thesis employs a two-pronged analytical approach. First, a cluster analysis of digitization indices, including the Mobile Connectivity Index, the Digital Adoption Index, and the E-Government Development Index, will uncover patterns and groupings across time periods; additionally, a correlation analysis of the indices' sub-levels is performed to assess their conceptual depth. Second, the research investigates correlations between these digitization metrics and linguistic diversity measures, offering a comprehensive perspective on how technological progress intersects with linguistic diversity and endangerment.

1.2 Research Questions

To address these challenges, this thesis focuses on the following research questions:

Primary Research Question:

How does digitization influence linguistic diversity and language endangerment globally?

Secondary Research Questions:

What clusters or patterns emerge from the analysis of digitization indices (e.g., Mobile Connectivity Index, Digital Adoption Index, and E-Government Development Index) across time?

How do digitization indices correlate with linguistic diversity measures, specifically the number of languages spoken in a country, the Red List Index (RLI), which has been adapted to assess language extinction risk, and entropy-based metrics, which are used to evaluate linguistic diversity?

1.3 Roadmap

After the introduction, this thesis will continue with the literature review. The literature review examines existing research on linguistic diversity and endangerment, including key measures like the Red List Index (RLI) and entropy, both computed for evaluating linguistic diversity. It further explores digitization’s societal impacts, including the *digital language divide*, and concludes by analysing the interplay between digitization and linguistic diversity, highlighting gaps in current research.

The next chapter details the methodological approach of the study. It describes the data collection process for digitization indices and provides an in-depth overview of the computation and characteristics of both digitization and linguistic diversity measures. Additionally, it explains the analytical techniques, including cluster analysis and correlation analysis, which are used to explore patterns and relationships in the data.

The subsequent results chapter presents the findings of the cluster analysis, identifying temporal patterns in digitization. It then discusses the correlations between digitization indices and linguistic diversity measures, offering insights into the influence of digitization on language endangerment.

Finally, the conclusion summarizes the key findings, their implications, and the contributions of the study. It reflects on the limitations of the research, such as data availability, methodological constraints, or generalizability, and discusses how these limitations might have influenced the results. The chapter also provides recommendations for future studies aimed at preserving linguistic diversity in the digital age.

2 Literature Review

Building on the key themes introduced earlier, the literature review expands on the mechanisms driving language endangerment and explores the societal impacts of digitization, setting the stage for addressing the gaps identified in the introduction.

The existing body of research on linguistic diversity and endangerment, on the one hand, and digitization, on the other, highlights critical issues in both domains, but leaves significant gaps in assessing the interaction between these two fields. While studies have documented the alarming rates of language loss and the societal impacts of digitization, few have focused on how digitization specifically influences linguistic diversity. This literature review explores key studies addressing language endangerment, the societal effects of digitization, and the intersection of these two areas, setting the stage for research that seeks to disentangle these dynamics. By identifying underexplored connections, this review underscores the need for a deeper examination of digitization's dual role as a potential threat and a resource for preserving endangered languages.

2.1 Linguistic Diversity and Language Endangerment

To establish a foundation for this thesis, it is essential to begin with the definition of key concepts related to linguistic diversity and endangerment. The concept of language itself is not without controversy. Dominant discourses often define languages as neatly bounded, countable systems, each tied to a specific worldview. While this perspective raises awareness of the richness of linguistic diversity, it risks oversimplifying the dynamic processes of language contact and change that characterize much of human history (Moore, 2016). However, for the purpose of this thesis and its analysis, it is necessary to adopt this simplified approach to the definition of languages, as it provides a practical framework for examining linguistic diversity.

Linguistic diversity can be categorized into three distinct types. The first is language richness, defined as the number of languages present within a specific area. This thesis will examine language richness through the number of languages spoken in each country. Phylogenetic diversity originates from evolutionary biology, where it is defined as the smallest total branch length needed to connect a group of species on a phylogenetic tree. This concept has been adapted to linguistics, where it is applied to language families to reflect their historical relationships. Lastly, linguistic disparity refers to the variability of expressions and traits within a language group (Gavin et al., 2013).

Language endangerment, and its ultimate outcome—language loss—are closely tied to the process of language shift. Long before a language dies, its speakers begin to shift away from using it, adopting another language in its place (Moore, 2016). Moore highlights that despite the fact that language loss and language endangerment can be considered a global crisis, it was not until the early 1990s that the field of language endangerment began to gain significant recognition, driven by growing concern over the rapid disappearance of languages. Earlier linguistic research had predominantly focused on the anthropological and sociological factors behind language shift, rather than addressing the broader issue of language endangerment or the urgent need for action. According to Moore, the publication of key works, such as Hale et al. (1992), marked a turning point, framing language endangerment as a crisis requiring both scholarly and practical intervention.

More recent research highlights the alarming scale and pace of language loss today. Simons (2019) estimates that one of the approximately 7000 languages spoken globally could be lost every 14 days by the mid of the next century. Similarly, Bromham et al. (2021) predict that one language could disappear each month in the near future if no precautionary measures are taken. This trajectory could result in the loss of 1.500 languages by the end of the century, with the most pessimistic projections suggesting a decrease of up to 90% in global linguistic diversity. Importantly, Bromham et al. highlight that these pessimistic forecasts can still be mitigated through countermeasures such as investing in language vitality and documenting endangered languages. Their findings emphasize the urgency of continuing research into language endangerment and its underlying drivers. Bromham et al. further identify key drivers behind this decline, including infrastructure expansion, economic growth, and educational systems that prioritize dominant languages over minority ones. They demonstrate correlations between decreasing linguistic diversity and factors such as as greater road density and higher levels of schooling.

Gavin et al. (2013) examine how sociocultural, topographic, and environmental dynamics influence language endangerment. They explore processes such as neutral change —random shifts in language over time —, movement, contact, and selection, which shape language vitality in distinct ways. Gavin et al. highlight the role of geographic factors, noting that regions with topographic complexity often correspond to higher linguistic diversity due to physical barriers that limit intergroup interactions and help preserve smaller languages. In contrast, regions with accessible terrains tend to foster language competition, potentially reducing diversity.

Bouckaert et al. (2022) identify a range of socio-ecological and cultural factors driving language diversification and loss. Moderate population densities and intermediate terrain difficulty foster language diversification, while very high population densities and extremely rugged or highly accessible terrains reduce diversification rates. Proximity to cities correlates with increased extinction risks, while larger geographic ranges are associated with greater diversification and persistence. However, extreme isolation or insularity can elevate extinction risks for smaller populations. Culturally, agricultural societies and hierarchical political systems are linked to higher diversification rates compared to forager societies. Practices like exogamy — the practice of marrying someone beyond one's social group — and written traditions exhibit mixed effects on language diversification, depending on the socio-ecological context. These findings highlight the complex interplay of environmental, social, and cultural factors shaping linguistic diversity globally.

Further, Kandler and Unger (2023) link linguistic diversity and language shift to demographic and socio-economic factors, such as population density and urbanization trends. These elements influence the rate and trajectory of language shift, often favouring languages perceived as more prestigious or beneficial for social mobility.

Beyond these structural and environmental considerations, Hutson et al. (2024) underscore the societal and cultural implications of language loss, arguing that languages play a fundamental role in shaping cultural identity and preserving collective memory. They note that languages serve as repositories of traditional knowledge, history, and values, acting as anchors for community cohesion. The loss of linguistic diversity disrupts the transmission of these cultural elements, weakening the social fabric of affected communities. This underscores the importance of preserving languages not only for their intrinsic linguistic value but also as essential components of cultural identity.

Efforts to quantify language endangerment have also advanced the understanding of linguistic diversity. Simons (2019) highlights the importance of efforts to quantify the state of languages in advancing the understanding of linguistic diversity and endangerment. One such tool is the Expanded Graded Intergenerational Disruption Scale (EGIDS), introduced in 2010, which categorizes languages based on their level of intergenerational transmission, from thriving to extinct. Applied to all languages documented in *Ethnologue*[1], the EGIDS has become one of the most widely used frameworks for assessing language endangerment. However, because EGIDS relies on proprietary datasets like Ethnologue, its accessibility is limited.

Recognizing the strong parallels between biological and linguistic evolution (Pagel, 2016), this thesis also incorporates two measures derived from biodiversity research — Shannon entropy and the Red List Index (RLI) — which have been adapted to assess linguistic diversity globally. Shannon entropy quantifies variability and evenness within a dataset, making it particularly relevant for assessing linguistic diversity (Grin & Fürst, 2022; Gullifer & Titon, 2019). In this context, entropy-based measures provide insights into the distribution of languages within populations, capturing the balance between dominant and minority languages. The RLI was originally developed by Butchart et al. (2004) and refined in 2010 on behalf of the World Conservation Union (IUCN) and its partner organisations to track biodiversity loss. It was first applied to evaluate language threat levels by Harmon and Loh in 2014. Following this approach, Baumann et al. (2024) calculated an adapted RLI by employing EGIDS scores as a linguistic threat measure, offering a quantitative measure of linguistic endangerment across different geographic regions. Both entropy-based measures and the adapted RLI were applied at the national level to

[1] https://www.ethnologue.com/

assess overall linguistic diversity and the degree of language endangerment within countries. A detailed explanation of the entropy and RLI computation methodology is presented in the data section of this thesis.

Although the studies mentioned above have significantly advanced the understanding of the dynamics and factors driving the decline of linguistic diversity, they have not specifically examined the interaction between digitization and language endangerment. The aim of this thesis is to address this gap.

2.2 Measuring Digitization

Digitization can be regarded as a component of the broader field of Information and Communication Technologies (ICTs), which encompass systems designed for generating, processing, storing, communicating, and presenting digital information (Charfeddine & Umlai, 2023). The term digitization is often defined as the conversion of analogue information into digital formats, while digitalization refers to the application of digital technologies, thereby expanding the impact of digitization, with digital transformation representing the ultimate outcome of this process (Jovanović et al., 2018). However, for the purpose of this thesis, I will use a broader definition of digitization as a social process, which "*refers to the transformation of the techno-economic environment and socio-institutional operations through digital communications and applications*" (Katz & Koutroumpis, 2013, p. 314).

To assess the technological advancement of nations and the resulting gaps several measures have been established. Billon (2010) highlights common indicators of ICT diffusion, such as internet users, broadband and mobile phone subscriptions, along with factors like access costs and regulatory frameworks. He also notes that demographic factors, including population size, distribution, density, and the urban-rural divide, are strongly linked to the digital divide observed across countries. However, since the reasons behind disparities in ICT diffusion are complex and multifaceted, analysing a single factor is insufficient to fully understand technological developments globally. To address this, comprehensive ICT indices have been developed by leading organizations such as the United Nations, the World Bank, the Organisation for Economic Co-operation and Development (OECD), and the International Monetary Fund (IMF), alongside various initiatives and scholars, to aggregate information on digitalization levels. The most prominent among these indices include the Network Readiness Index (NRI), the Digital Divide Index (DDI), the ICT Development Index (IDI), and the Mobile Connectivity Index (MCI).

Archibugi et al. (2009) describe these measures as “*macroeconomic indicators aiming at comparing the positions of different countries and their changes*“(p. 918). The authors conclude that, generally speaking, such indicators are valuable resources for both policy action and academic research. These indicators offer a comprehensive assessment of technological capabilities by aggregating diverse metrics, enabling cross-national comparisons, and providing valuable benchmarks for policymakers and researchers. However, the authors note several limitations, including the simplification of countries as homogenous entities for statistical analysis, which may obscure internal regional disparities, and the potential for comparisons between nations with vastly different contexts to be misleading. Additionally, aggregation assumes substitutability of diverse metrics, while the subjective weighting of variables may introduce bias. Despite these challenges, Archibugi et al. emphasize that such indicators remain essential tools for understanding technological capabilities and guiding effective policy, provided their limitations are interpretably recognized.

2.3 Digitization and its Societal Impact

The following section focuses on digitization as a transformative societal force, examining its broader impacts and the disparities it creates.

Digitization has significantly influenced nearly every aspect of modern society in recent decades, including education, healthcare, communication, and entertainment (Olaare, 2024). However, its benefits—such as enhanced economic growth, productivity, and welfare (Katz & Koutroumpis, 2013)—remain unevenly distributed across the globe. High-income countries reap the most significant rewards from technological adoption, whereas developing nations often struggle to unlock digitization's full potential. Billon (2010) defines this *digital divide* as disparities in access to information and ICTs, which mirror and often exacerbate existing socioeconomic inequalities. Regions with limited infrastructure and lower GDP face significant barriers to adopting digital tools, leaving marginalized populations without the resources to participate fully in the digital economy. Billon highlights that addressing these disparities requires targeted investments in infrastructure and education to bridge the gap between advantaged and disadvantaged groups and regions.

Building on this discussion, Vassilakopoulou and Hustad (2023) address deeper inequalities that persist even in areas where digital access has improved. They emphasize that access alone is not enough to close the digital divide. Digital literacy and the ability to use technology effectively remain unevenly distributed, particularly among disadvantaged groups. The authors advocate for capacity-building initiatives that empower individuals and communities to utilize digital tools effectively, ensuring that digitization promotes equity rather than exacerbating existing disparities.

Digitization’s broader societal implications also intersect with sustainability goals. Jovanović et al. (2018) explore the potential for digital integration to drive environmental and social improvements. Advanced digital infrastructures are shown to enhance resource efficiency and promote sustainable practices. However, Jovanović et al. caution that these benefits are often concentrated in wealthier regions, leaving disadvantaged areas further behind. Inclusive policies are needed to ensure that digitization’s advantages are equitably distributed.

2.4 The Intersection of Digitization and Linguistic Diversity

While digitization offers immense potential for societal development, its unequal impacts exacerbate existing disparities, including those related to linguistic diversity, as discussed in the following section.

The interplay between digitization and linguistic diversity highlights both opportunities and significant challenges, as the benefits of digital technologies remain heavily skewed toward dominant languages, particularly English. Despite only 17% of the global population speaking English, it accounts for approximately 60% of all online content (Nee et al., 2022). This linguistic imbalance is further evident in the performance of Natural Language Processing (NLP) tools, which overwhelmingly favour English and provide only limited support for other languages. For instance, text-to-speech synthesis technologies cover just 10% of the world’s languages, while natural language inference and question-answering capabilities are restricted to as few as 15 and 17 languages, respectively. Even among these, English remains the primary focus, with only a handful of Western European and non-European languages receiving moderate support (Blasi et al., 2022). The vast majority of languages, however, are minimally supported or entirely excluded, creating a digital language divide that disproportionately impacts smaller and endangered languages. This divide not only restricts their ability to participate in and benefit from technological advancements but also exacerbates the risk of language extinction. Scholars emphasize the urgent need for inclusive digital strategies to counteract these inequities and preserve linguistic diversity, warning that without deliberate efforts, these disparities will reinforce existing inequalities and led to greater linguistic and cultural losses.

Cunliffe (2007) and Hutson et al. (2024) provide complementary insights into the challenges and opportunities of preserving linguistic diversity in an increasingly digital world. Cunliffe underscores the potential of electronic technology to support endangered languages, referencing Crystal’s assertion that such languages can progress when their speakers actively engage with digital tools. While the internet creates new opportunities for minority languages to establish a presence and challenge their perceived outdated or low status, it remains overwhelmingly dominated by majority languages like English. This dominance stems from historical factors such as the internet’s origins, the early adoption by English-speaking communities, and the economic and infrastructural advantages of dominant languages. Minority language communities often face additional barriers, including limited digital infrastructure, economic challenges, and the exclusionary cultural conventions embedded in software, which often reflect the norms of dominant languages.

Similarly, Hutson et al. (2024) highlight the implications of this dominance, focusing on how English-centric technologies, particularly large language models (LLMs) developed by U.S.-based tech giants, exacerbate linguistic homogenization and accelerate language extinction. The authors emphasize the need for scalable, data-driven models that integrate machine learning and big data analytics to document cultural and linguistic artefacts comprehensively. Both studies stress the importance of targeted, participatory efforts to bridge the digital language divide. These efforts are not only necessary to preserve linguistic diversity but also to challenge broader socioeconomic inequalities and reframe minority languages as contemporary, vital tools for communication while respecting their historical and cultural contexts.

Despite its many challenges, digitization offers significant opportunities for supporting linguistic diversity. Digital tools have a transformative role in documenting and revitalizing endangered languages (Alizai, 2021; Olaare, 2024). Alizai highlights the potential of digital platforms such as digital archives, transcription tools, and language-learning apps, which open new pathways for intergenerational knowledge transfer and cultural expression. Similarly, Olaare emphasizes the positive impact of emerging technologies—including mobile apps, cloud computing, social media, and AI—on language preservation by fostering community engagement and improving the accessibility of resources.

However, several barriers remain, including limited funding, digital literacy gaps, data security concerns, technological access issues, and challenges in ensuring data quality and culturally relevant content. It is crucial that these technologies are designed inclusively to address the specific needs of marginalized language communities, ensuring their perspectives and cultural values are respected throughout the process. This includes respecting that language communities may choose not to share their linguistic data due to concerns about privacy, cultural sensitivity, or differing views on intellectual property (Nee et. al, 2022).

Ultimately, Olaare (2024) emphasizes the importance of continuous research and evaluation, advocating for "*conducting longitudinal studies to assess the long-term interventions on language use and vitality*"(p. 54). This recommendation highlights the necessity of ongoing assessment and refinement to ensure that digital tools remain relevant, responsive, and effective in addressing the evolving needs of language communities. By aligning with these objectives, this thesis aims to contribute to the broader efforts to preserve linguistic diversity in an increasingly digitized world.

In summary, the intersection of linguistic diversity, digitization, and societal inequality presents a complex challenge. Previous research highlights the urgent need to address the rapid decline in linguistic diversity, driven by a combination of socioeconomic and cultural factors. Digitization introduces both risks and opportunities, acting as a double-edged sword in the context of language endangerment. This thesis seeks to further investigate these dynamics, providing deeper insights into how digitization influences linguistic diversity. By uncovering nuanced interactions between these forces, this work aims to inform future strategies and foster a better understanding of the mechanisms driving language endangerment.

3 Data

This thesis integrates multiple datasets that encompass both digitization metrics and linguistic diversity measures. The following chapter first provides a brief overview of the data collection process, then presents a detailed examination of the characteristics of the collected measures, including their computational methodologies and constituent components. The discussion begins with a detailed examination of the digitization indices, addressing their sources, structure, and underlying indicators. This is followed by an exploration of the linguistic diversity metrics, highlighting their conceptual foundations, and relevance to the broader analytical framework.

3.1 Digitization Indices

The digitization indices used in this thesis capture various dimensions of technological advancement, including internet accessibility, e-governance, and AI readiness. These indices were selected based on their global and temporal coverage, relevance to digitization, and data accessibility. The underlying data, sourced from reputable institutions, was available in various formats—including websites, Excel spreadsheets, and PDF reports—and had to be extracted and standardized into a uniform tabular format.

An overview of the indices, including their scope and data sources, is presented in Table 1. Together, these indices offer broad insights into digital adoption, infrastructure, and policy implementation across various regions. Two important points should be emphasized: first, not all data for these indices are publicly available or accessible as presented in the table, resulting in gaps for certain years or levels within the analysis; and second, the data reflect the status as of November 2024, when they were collected.

Data

Index	Indicators	Countries	Years	Source
AI Preparedness Index (AIPI)	7	174	2023	IMF
Digital Adoption Index (DAI)	12	180	2014, 2016	World Bank
Digital Services Trade Restrictiveness Index (DSTRI)	21	91	2014-2023	OECD
Digitization Index (DiGiX)	24	98	2018-2024	BBVA Research
E-Government Development Index (EGDI)	174	193	2003-2005 (old), 2008-2022 (new)	UN
E-Participation Index (EPI)	58	193	2003-2005 (old), 2008-2022 (new)	UN
GovTech Maturity Index (GTMI)	48	198	2020, 2022	World Bank
ICT Development Index (IDI)	10	170	2009-2017 (old), 2023, 2024 (new)	ITU
Inclusive Internet Index (3i)	62	100	2017-2022	The Economist Impact; Meta
Mobile Connectivity Index (MCI)	32	173	2014-2023	GSMA
Network Readiness Index (NRI)	58	134	2006-2016 (old), 2019-2023 (new)	WEF, Portulans Institute

Table 1 Summary of digitization indices.

To provide broader context, a detailed description of each digitization index is presented below in alphabetical order.

The AI Preparedness Index[2] (AIPI), released by the International Monetary Fund (IMF) focuses on readiness for AI adoption, enabling better comparability across regions with varying income levels. It evaluates AI preparedness in 174 countries based on four key determinants: Digital Infrastructure, Human Capital and Labor Market Policies, Innovation and Economic Integration, and Regulation and Ethics.

Alongside its core determinants, the index integrates multiple indicators sourced from eight major institutions, including the Fraser Institute, the International Labor Organization, the International Telecommunication Union, the United Nations, and the World Bank. However, the values of these indicators are not publicly available. Additionally, since some indicators are perception-based, reflecting subjective assessments and experiences, the index should be interpreted as a guiding tool for

[2] https://www.imf.org/external/datamapper/datasets/AIPI

identifying strengths and areas for improvement rather than as an absolute ranking system.

Despite the inherent challenges in measuring AI preparedness—particularly given the evolving institutional requirements for AI integration—the AIPI serves as a useful framework for analysing the relationship between digital infrastructure and linguistic diversity. The index is computed as a simple average of its four dimensions, with each dimension itself based on the normalized averages of its respective indicators (Cazzaniga et al., 2024).

The Digital Adoption Index (DAI)[3] was first introduced by the World Bank as part of the World Development Report 2016: Digital Dividends. This composite index, available for the years 2014 and 2016, assesses the level of digital adoption in countries across three key sectors: People, Government, and Business.

Each dimension of the DAI is constructed using a combination of normalized indicators and sub-indices. Specifically, the people dimension includes four normalized indicators, the government dimension incorporates two normalized indicators, and the business dimension consists of sub-indices that are further divided into nine normalized indicators. The specific values are only available at index and dimension level though.

The index captures the essential technologies required by each sector to drive development in the digital age. For businesses, it focuses on increasing productivity and fostering broad-based economic growth; for individuals, it highlights expanding opportunities and enhancing well-being; and for governments, it emphasizes improving efficiency and accountability in service delivery. By integrating these dimensions, the DAI provides a comprehensive and accessible overview of technological adoption at a national level (WDR 2016 Team, 2016).

The OECD Digital Services Trade Restrictiveness Index (DSTRI)[4] evaluates barriers to digitally traded services, following the methodology of the OECD Services Trade

[3] https://www.worldbank.org/en/publication/wdr2016/Digital-Adoption-Index

[4] https://goingdigital.oecd.org/en/indicator/73

Restrictiveness Index (STRI). Currently, it provides data for 44 countries and consists of two main components: a regulatory database and a set of indicators.

The index assesses restrictions across five key areas: Infrastructure and Connectivity, Electronic Transactions, Payment Systems, Intellectual Property Rights, and Other Barriers Affecting Digitally Enabled Services Trade. Its values are derived through a structured process involving scoring, weighting, and aggregation. Scoring follows a binary *Yes or No* system, while weighting is determined by expert input. Finally, the overall index score is calculated as a weighted average of the individual scores (Ferencz, 2019).

Since the DSTRI measures restrictiveness, with higher values indicating more barriers to digital trade, the values were inverted for the analysis to ensure alignment with the study’s interpretative framework.

The Digitization Index (DiGiX)[5], developed by BBVA Research, assesses the factors, institutional frameworks, and behaviors that enable a country to effectively leverage Information and Communication Technologies (ICTs). Covering 98 countries, the index aims to enhance competitiveness and overall well-being by identifying regions where targeted actions are needed.

The DiGiX is structured into three main pillars: Supply Conditions, Demand Conditions, and Institutional Framework. These pillars are further divided into six dimensions and 24 indicators. However, only the overall index values are consistently available to the public. Additionally, the indicators are updated annually to reflect technological advancements, meaning that the DiGiX values are not directly comparable across different years.

Despite this limitation, the DiGiX remains an intuitive and actionable tool due to its data-driven methodology and transparent weighting system, which is based on a two-stage Principal Component Analysis (PCA)[6]. This approach provides clear

[5] https://www.bbvaresearch.com/en/tag/digix/

[6] PCA is used In Exploratory Data Analysis (EDA) to reduce dimensions in datasets.

insights for deciosion-makers, serving as a solid basis for policymaking (Cámara, 2024).

The E-Government Development Index (EGDI)[7] is published every two years as part of the United Nations E-Government Survey. This composite index currently covers 193 countries and evaluates global and regional e-government performance. It incorporates factors such as infrastructure and educational levels to assess how effectively a country utilizes information technologies to enhance accessibility and inclusion.

The EGDI is structured around three key dimensions, each represented by its own sub-index: the Online Service Index, Human Capital Index, and Telecommunication Infrastructure Index. These dimensions are further divided into sub-indices and indicators, providing a comprehensive assessment of e-government capabilities. The overall index is calculated as the average of the normalized values of these three dimensions, with each given equal weight.

While the index's methodology and the composition of its sub-components have been updated periodically, making direct comparisons over time challenging, the EGDI has provided biannual data since 2003 on all levels. This clearly makes it one of the most valuable long-term indicators for assessing digitization trends globally (United Nations Department of Economic and Social Affairs, 2024).

The E-Participation Index (EPI)[8] is a supplementary measure within the United Nations E-Government Survey, specifically functioning as a sub-index under the Online Service Infrastructure dimension of the EGDI.

The EPI framework is structured around three key elements: E-Information, E-Consultation, and E-Decision-Making. These components assess government portals and websites based on their implementation of features such as budget transparency, open government data, and mechanisms for joint service development and collaborative initiatives, as well as other digital participation tools. Special

[7] https://publicadministration.un.org/egovkb/en-us/About/Overview/-E-Government-Development-Index

[8] https://publicadministration.un.org/egovkb/en-us/About/Overview/E-Participation-Index

emphasis is placed on features that enhance accessibility and encourage engagement, particularly for individuals in vulnerable situations.

The EPI score for each country is calculated through a normalization process, where the lowest total score in the survey is subtracted from a country's score, and the result is divided by the overall range of scores across all countries. Following this, the standard competition ranking method is used, ensuring that countries with the same EPI score share the same rank, while ranking gaps remain where ties occur (United Nations Department of Economic and Social Affairs, 2024).

The GovTech Maturity Index (GTMI)[9] was developed by the World Bank as part of the GovTech Global Partnership initiative to assess the level of Governance Technology (GovTech) maturity worldwide. GovTech represents a digital government transformation approach that prioritizes active citizen engagement. The latest 2022 edition of the GTMI covers 198 countries, grouping them into categories rather than ranking them.

The GTMI evaluates four key focus areas: Core Government Systems and Shared Digital Platforms, Online Service Delivery, Digital Citizen Engagement, and GovTech Enablers. Each of these areas consists of multiple indicators, which are further classified into progressive indicators—divided into two weight categories (Level 1 and Level 2)—and binary indicators. The framework incorporates established indices, such as the E-Government Development Index (EGDI) and the UN E-Participation Index (EPI), as key indicators within its assessment. The GTMI score is determined by computing the average of the normalized values across its key areas. Since its introduction in 2020, the index methodology has been revised and expanded, including updates to its indicators. While data is available for both years at all levels of the index, methodological changes limit direct comparability over time (World Bank, 2022).

The ICT Development Index (IDI)[10] has been published by the International Telecommunication Union (ITU) since 2009. Initially released annually until 2018, the index was relaunched after a six-year hiatus with a new methodology that shifts the

[9] https://www.worldbank.org/en/programs/govtech/gtmi
[10] https://www.itu.int/en/ITU-D/Statistics/Pages/IDI/

focus from basic connectivity to universal and meaningful connectivity. Due to the methodological differences in its computation, this thesis will treat the old and new versions of the index as separate indices in the analysis.

The 2024 edition of the IDI covers 170 countries and is structured around two main pillars: Universal Connectivity (comprising three indicators) and Meaningful Connectivity (comprising six indicators). Each individual indicator is assigned a normalized score between 0 and 100. These indicator scores are first aggregated into pillar scores, which are then further combined to compute the overall IDI score. Equal weights are applied to each level of aggregation.

As with any index, caution is advised when interpreting and utilizing IDI results. To gain a comprehensive, accurate, and up-to-date understanding, the findings should be complemented and validated with additional sources of information (International Telecommunication Union, Development Sector, 2024).

The Inclusive Internet Index (3i)[11], developed by the Economist Impact with support from Meta, evaluates the extent to which citizens in 100 countries can access and benefit from internet connectivity. The index is designed to serve as a reference for researchers and policymakers, helping them identify opportunities to expand internet accessibility and maximize its benefits globally.

As of its latest edition, the Inclusive Internet Index (3i) has been available for five years (2017–2022). The index is structured around four main pillars: Availability, Affordability, Relevance, and Readiness. Each pillar includes key indicators of internet inclusion, combining both quantitative measures—such as network coverage and pricing—and qualitative factors, including the presence of e-inclusion policies and the availability of local-language content. Additionally, survey data from the Value of the Internet Survey conducted by the Economist Impact serves as the foundation for several indicators, particularly those related to Relevance and Readiness.

Selected indicators have been backscored to reflect revisions in previously reported data, ensuring greater accuracy. In cases where verification was not possible,

[11] https://impact.economist.com/projects/inclusive-internet-index

reasonable estimates were made. As a result, data should only be referenced from the latest edition of the index for consistency (Economist Impact, 2022).

The Mobile Connectivity Index (MCI)[12], developed by the GSMA Association (GSMA), evaluates mobile connectivity performance across 173 countries. A key distinguishing feature of this index is its reliance on mostly unique indicators that are not covered by similar indices. Additionally, the MCI functions as an input index, assessing a country's performance based on key enablers of mobile connectivity, rather than an output index, which would measure factors such as internet usage.

The index framework is built around four key enablers: Infrastructure, Affordability, Consumer Readiness, and Content & Services. Each enabler comprises multiple dimensions, which are derived from the aggregation of their subordinate indicators. To ensure data representativeness, a country must have at least 75% of the overall indicator data available and a minimum of 50% coverage within each enabler.

The MCI methodology was updated in 2023 following feedback and a statistical audit conducted by the European Commission’s Joint Research Centre (JRC). However, since these changes were applied retrospectively to all previous editions of the index, comparability and consistency across different years remain ensured (GSMA Intelligence, 2024).

The Network Readiness Index (NRI)[13] was initially developed by the World Economic Forum in 2006 and has been maintained by the Portulans Institute in cooperation with the Saïd Business School and the University of Oxford since 2019. The index assesses a country's readiness to leverage opportunities provided by information and communication technology (ICT). Due to data availability constraints, this thesis will utilize only data from the updated version of the NRI (2019–2023) for analysis.

The NRI is structured around four foundational pillars: Technology, People, Governance, and Impact. Each pillar is further divided into three sub-pillars, which

[12] https://www.mobileconnectivityindex.com/index.html

[13] https://networkreadinessindex.org/

collectively encompass 58 indicators. The index assigns both a ranking and individual scores on a 0 to 100 scale for each component level.

While the fundamental four-pillar structure has been retained from previous versions, lower-level measures were updated in 2023. This approach ensures a balance between stability and continuity on one hand and relevance and up-to-date assessments on the other (Portulans Institute, 2023).

Each of these digitization indices focuses on different facets of digital transformation, collectively they provide a comprehensive picture of how digital access, adoption, and policy implementation vary across countries and over time. However, differences in regional and temporal coverage, computational methods, and data processing strategies mean that direct comparisons between indices should be made with caution. Each index applies its own approach to imputation, weighting, and aggregation, and some have undergone methodological changes over time. These factors introduce certain constraints, but they do not diminish the overall value of the indices in capturing key trends in digitization.

3.2 Language Diversity Measures

As mentioned in the Literature Review, this thesis incorporates three language diversity measures to analyse the correlation between digitization indices and language diversity: the Red List Index (RLI), and Shannon entropy, which have both been specifically adapted to measure linguistic diversity, and the number of languages spoken in each country. The following section will provide a more detailed discussion of the properties of these two measures.

3.2.1 Number of Languages per Country

One measure of linguistic diversity used in this study is the number of languages spoken in each country, reflecting language richness—a commonly used metric for assessing language diversity (see Literature Review for details). The language counts are based on the 26th Ethnologue edition (2023). This variable shows a highly skewed distribution, with some countries reporting only one language while others report as many as 848. For example, the Falkland Islands, Jersey, Niue, Pitcairn, Saint Helena, San Marino, Tokelau, and Vatican City each have just one language. In contrast, Papua New Guinea leads with 848 languages, followed by Indonesia with 712 and Nigeria with 523. This extreme variability necessitated a log transformation of the language counts for the correlation analysis, in order to mitigate the influence of outliers and stabilize the variance (Feng et al., 2014).

However, simple language counts do not account for the relative size of language communities or the evenness of distribution. Therefore two additional measures from ecological were included in this study, which will be introduced subsequently.

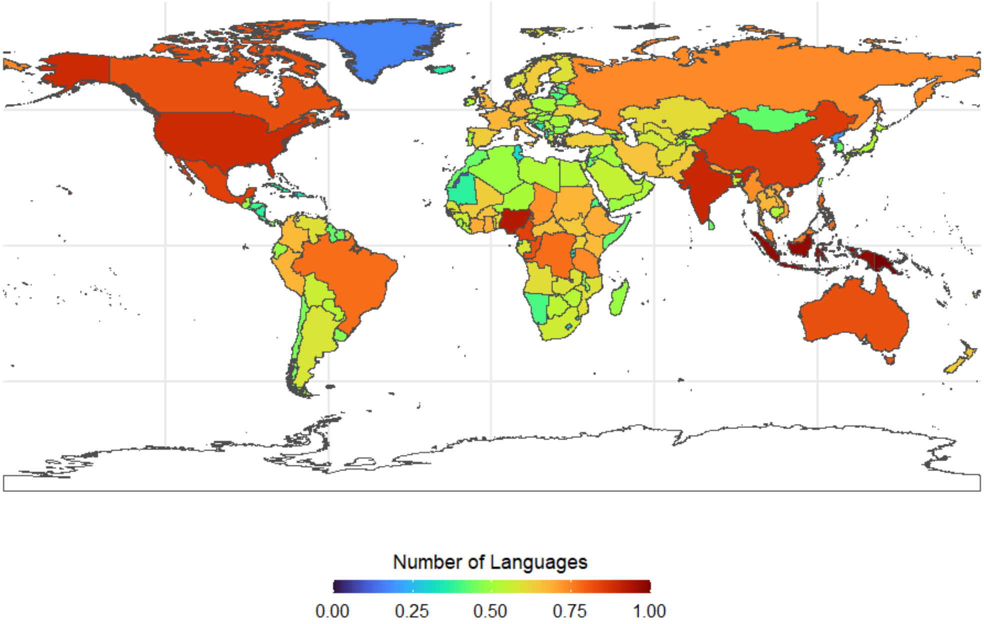


Figure 1 Global linguistic diversity as measured by numbers of languages spoken in a country (normalized). [15]

3.2.2 Entropy-based Language Diversity

A measure of diversity, commonly used in biodiversity assessments, is Shannon entropy, also known as Shannon's diversity index. Shannon entropy takes probabilities as input and quantifies the uncertainty or randomness in a dataset. The exponent of entropy is referred to as a measure of order 1 (Tuomisto, 2010). Grin and Fürst (2022) specifically recommend using Shannon entropy for linguistic diversity analysis "because of its highly useful mathematical properties" (p. 619).

To establish a foundation for computing entropy, first speaker numbers had to be collected. These numbers were sourced from *Ethnologue* (2023) and the Joshua

Project’s *People Groups* dataset (2023), as both sources offer high data quality and coverage. The Joshua Project data, organized by people groups, was cleaned and standardized, while *Ethnologue* data, organized by language, required additional parsing to extract speaker counts. The datasets were merged on ISO language and country codes, with missing values filled where possible and overlapping counts resolved by selecting the minimum. The resulting dataset serves as the basis for all language distribution calculations used in the analysis.[14]

For the computation of entropy, raw speaker counts for each language were normalized by dividing them by the total number of speakers in the respective country, thereby converting the data into probabilities. These probabilities were then used as input for the calculation of Shannon entropy scores in R. Although the exponent of entropy is commonly expressed as a measure of order 1, the logarithm of entropy was applied instead due to the skewed distribution of the exponent, ensuring a more balanced representation of linguistic diversity. The logarithm of entropy is commonly referred to as a diversity measure of order 0.

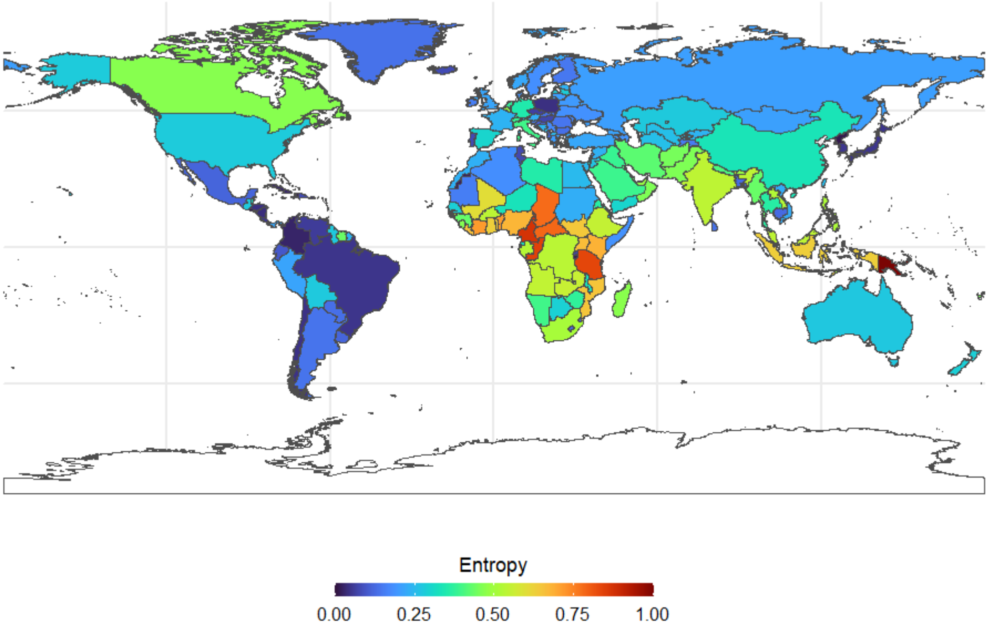


Figure 2 Global linguistic diversity as measured by entropy (normalized). [15]

[14] The speaker numbers dataset construction and entropy calculation were originally developed in collaboration with a colleague as part of a seminar project and were subsequently adapted for the purposes of this thesis.

[15] For visualization purposes, entropy and speaker count values were normalized using Min-Max Normalization, scaling them to the [0,1] range to ensure direct comparability with the adapted RLI.

3.2.3 Adapted Red List Index

Another measure derived from biodiversity, is the Red List Index (RLI). The index was originally developed by Butchart et al. (2004) and further refined in 2010 on behalf of the World Conservation Union (IUCN) and its partner organisations to quantify biodiversity loss. The following adaptations and applications of the Red List Index (RLI) are based on its refined version.

The Red List Index (RLI) is a tool for assessing both the overall level of extinction risk and trends in this risk over time. Simply put, it is calculated as 1 minus the current threat level. The RLI ranges from 0 to 1, where a value of 1 signifies that no assessed entities are currently at risk, whereas a value of 0 indicates that all assessed entities have been lost. When analysing trends over time, a declining RLI indicates that the projected rate of future losses is worsening, meaning the threat level is increasing. Conversely, an increasing RLI suggests a reduction in the expected rate of loss.

In 2014, Harmon and Loh applied the RLI framework to evaluate language threat levels, following the approach of William Sutherland (2003), who had adapted a limited set of Red List criteria to enable comparisons between trends in species and language diversity. Expanding on this methodology, Baumann et al. (2024) further developed the RLI application for linguistic diversity by incorporating its mathematical framework. Linguistic diversity was computed by subtracting the average standardized endangerment level from 1, with endangerment levels determined using EGIDS scores from the 2023 edition of Ethnologue.

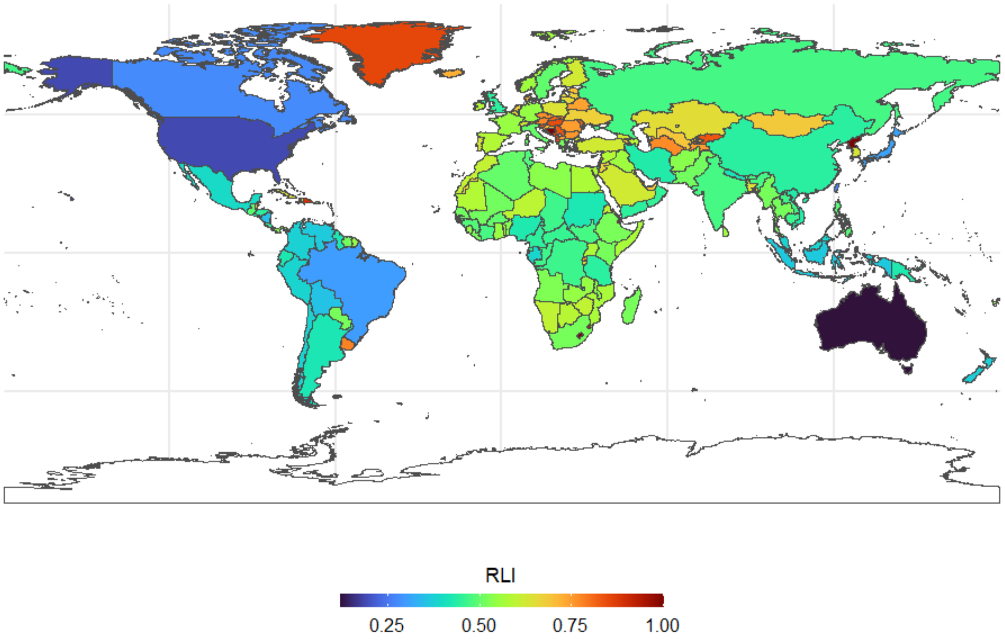


Figure 3 Global linguistic diversity as measured by the adapted RLI.

It should be noted that all diversity measures, utilized in this study, have certain limitations. They do not account for phylogenetic diversity, as they assigns equal weight to all languages regardless of their linguistic relationships. Additionally, they do not distinguish between first language (L1) and second language (L2) speakers, thereby not accounting for the role of multilingualism. However, despite these limitations, the adapted RLI, the total number of languages spoken in each country, and entropy-based measures remain powerful tools for assessing long-term trends in linguistic diversity and understanding shifts in the global language landscape.

To sum it up, the digitization indices used in this study come from various institutions and capture different aspects of technological development. These indices differ in temporal coverage and regional scope, requiring adjustments to ensure a consistent country-year structure where applicable. The linguistic diversity measures, the adapted Red List Index (RLI), the numbers of languages spoken per country, and entropy-based metrics provide insights into language distribution and speaker evenness.

4 Methodology

This chapter outlines the methodological approach used in this thesis, covering data analytical techniques. The methodology follows a structured, multi-step approach. First, the datasets undergo pre-processing to ensure consistency and comparability across different sources. Once processed, the digitization indices are organized into meaningful groups using cluster analysis, allowing for the identification of patterns and relationships among variables. This step is complemented by a correlation analysis of the elements (e.g., indicators) within each index, where such data is available. The analysis then concludes with a separate correlation analysis, which explores the statistical relationships between digitization indices and linguistic diversity measures, providing insights into how different aspects of digitization interact with linguistic diversity.

By combining these techniques, this study provides a robust framework for assessing the impact of digitization on linguistic diversity, offering a data-driven perspective on both its risks and opportunities. The following sections detail each methodological step in depth, outlining the rationale, processes, and tools used throughout the analysis.

4.1 Data Preparation

Before the analysis, the digitization datasets were pre-processed to ensure consistency and comparability. To resolve inconsistencies in country naming conventions, a uniform country identifier was required. Since some datasets used ISO 3166-1 alpha-2 country codes while others relied on ISO 3166-1 alpha-3 codes, standardization was necessary. To achieve this, ISO 3166-1 alpha-3 codes were mapped across all datasets using the World Bank Health Nutrition and Population Statistics dataset (World Bank, 2023), which provides both coding systems. This process follows established data integration techniques that emphasize the importance of harmonizing identifiers to ensure comparability across datasets (Voß, 2008). By aligning the identifiers, datasets could be accurately merged, ensuring consistency in both the cluster and correlation analysis in R.

Handling missing values was another key consideration. Given the fragmented nature of the data—both across different years and between datasets—no imputation was found to be appropriate. At the same time, removing missing values entirely would have led to excessive data loss. To address this, the pairwise complete observation method was applied along with the complete observation method during correlation analysis using the cor() function in *r* —first during the correlation analysis that informed the cluster analysis, and later in the subsequent standalone correlation analysis. This method allowed each calculation to use all available data for a given pair of variables without making assumptions about missing values (R Core Team, 2023).

No additional normalization or transformation was deemed necessary for this analysis, given that Pearson's correlation coefficient was used, which operates directly on raw data values (Cohen, 1988). To confirm that this choice was appropriate, the distributions of each index were evaluated for each year, where data was available, using Quantile Quantile plots (QQ plots) (see Figure 3). QQ plots are commonly used to detect potential issues in data distribution such as heavy skew or outliers, both of which can compromise Pearson's *r* by assuming data are approximately normally distributed (Cohen, 1988). In cases of pronounced skewness, Pearson's *r* may understate or inflate the true linear relationship, potentially distorting conclusions (Bishara & Hittner, 2015).

An equiangular reference line was added to the plot, indicating where the observed values are expected to fall if the data were normally distributed. This line aids in assessing whether the data follows a normal distribution, with significant deviations from the line suggesting non-normality. In the present data, none of the QQ plots revealed severe departures from normality or the presence of extreme outliers. Although slight curvature was evident in the tails, the main body of each distribution closely followed the diagonal, indicating that Pearson's correlation remained valid. As a result, the original (raw) index values were retained without additional transformations.

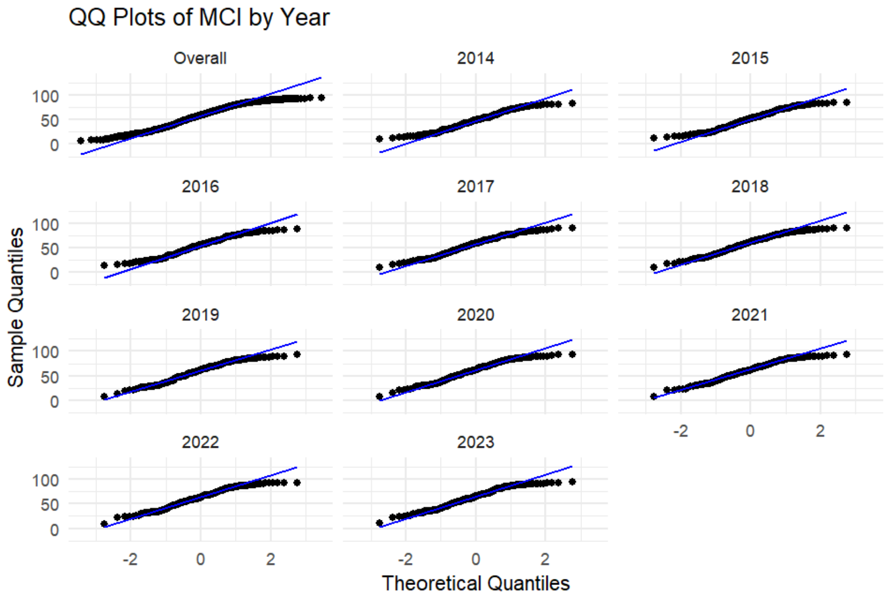


Figure 4 QQ plots by year for the MCI, including a reference line to assess normality.[16]

Additionally, to ensure interoperability and consistency across the datasets, ISO 3166-2 alpha-3 country codes were appended to the linguistic diversity measure dataset, mirroring the approach used for the digitization indices. This standardization allowed for seamless merging and consistent country identification across all analyses. Apart from this minor adjustment, no further data cleaning was required for the linguistic diversity measures, as they were already provided in a suitable format. With the datasets structured for analysis, the next step is cluster analysis in R, which will identify patterns among the digitization indices.

4.2 Analytical Approach

The analysis was conducted in R, utilizing statistical and visualization techniques to examine patterns in the data. The *r* scripts were made publicly available in a GitHub repository[17]. This included a hierarchical cluster analysis to identify groupings within the digitization indices, following standard clustering techniques (Rodriguez et al., 2016). Additionally, for indices where indicator-level data was available, the breadth of each index was assessed by computing the correlation structure of its underlying indicators in each year it was available. High correlation between indicators suggests redundancy, which could compromise the diversity of the index. Therefore, correlation analysis was used to determine the degree of non-redundancy in the

[16] The other QQ plots can be found in Appendix A.2.1

[17] https://github.com/KatharinaTheresia/Linguistic-Diversity-Thesis

components of each index (Maggino & Zumbo, 2012). For indices with only sub-index-level data available, the breadth was determined by analysing the correlation structure among the underlying sub-indices. This provided insight into how consistently the different components of an index aligned over time. Subsequently, correlation analysis was performed to explore relationships between digitization indices and linguistic diversity measures, in line with standard practices for correlation analysis (Makowski et al., 2020).

Given differences in data availability and methodological approaches, the analysis was structured to assess patterns and groupings across multiple years while ensuring robustness in the results.

4.3 Cluster Analysis

To examine patterns among the digitization indices, hierarchical clustering (Murtagh & Contreras, 2012) was applied for each year between 2004 and 2023 separately. This approach ensured that only indices available in a given year were included, taking into account that the indices vary in temporal and regional coverage. The objective of this analysis was to identify statistical similarities among the indices and to explore potential groupings that reflect underlying relationships between different measures of digital development.

4.3.1 Dataset Construction

The cluster analysis was performed on a merged dataset, constructed through a combination of full joins and inner joins of the digitization indices. A full join retains all observations from both datasets, even if there is no match, ensuring that all available data points are preserved, though it can introduce missing values that require handling. In contrast, an inner join returns only observations that have matching entries in both datasets, ensuring that only complete data points are retained, thereby reducing the number of observations but improving comparability (Wickham et al., 2023). Both approaches were implemented using the dplyr package, which provides flexible functions for data merging, as detailed in its documentation (Wickham et al., 2023). Table 2 provides an overview of how many unique countries each digitization index covers for each year, illustrating the extent of missing data across indices. This summary helps explain why both full and inner joins were performed during dataset construction, balancing the goal of maximizing available observations with the need for complete records.

Year	AIPI	DAI	DIGIX	EDGI	EPI	GTMI	IDI_NEW	IDI_OLD	3i	MCI	NRI	DSTRI
2004				192	193							
2005				192	193							
2006												
2007												
2008				192	193			152				
2009												
2010				192	193			152				
2011								157				
2012				192	193			166				
2013								166				
2014		183		192	193			175		173		90
2015								176		173		90
2016		183		192	193			176		173		90
2017									99	173		90
2018			99	192	193				100	173		90
2019			99						100	173	121	90
2020			99	192	193	198			100	173	134	90
2021							169		100	173	130	90
2022				192	193	198	170		100	173	131	90
2023	174									173	134	90

Table 2 Countries covered by index each year.[18]

4.3.2 Correlation Computation

Two approaches were used to compute correlation matrices from the full join dataset: the complete observation method and the pairwise complete observation method. The complete observation method considers only cases where all selected variables are present, ensuring strict comparability across indices but reducing the number of available observations. In contrast, the pairwise complete observation method computes correlations for each variable pair separately, using all available data for that specific comparison, thereby maximizing data retention. This approach allows for greater flexibility in handling incomplete datasets but can result in correlation matrices that are not always positive semi-definite (R Core Team, 2023). Both methods were implemented using the cor() function from the stats package (R

[18] Cell shading ranges from light to dark red, corresponding to an ascending number of countries.

Core Team, 2023). For validation, correlations were also computed on the inner join merged dataset; as expected, the complete-case correlation matrices were identical between the inner join and the full join—since only rows with non-missing values for all indices were used. The inner join correlation matrix was further examined for missing values to determine whether the corresponding pairwise complete observation matrix would differ significantly. When absolute differences between corresponding coefficients exceeded 0.10, an alternative clustering approach was applied using the inner join pairwise complete matrix.

The rationale for these approaches is supported by Graham (2009), who emphasizes that retaining as much available data as possible is beneficial for obtaining robust and unbiased estimates—even when some cases are incomplete. This broad principle justifies the use of methods that make full use of the data, as seen with the pairwise complete observation method.

While the analysis followed a consistent methodological framework, certain limitations and exceptions influenced both the correlation approach and the application of hierarchical clustering. These constraints were primarily driven by data availability, missing values, and the structure of the correlation matrices. In some cases, adjustments were necessary to ensure meaningful and interpretable results.

In some years, there were not enough fully observed cases to compute the complete observation correlation matrix because certain indices had limited coverage or were missing for some countries. In these cases, clustering was conducted using only the pairwise complete correlation matrix, ensuring that correlations could still be computed despite missing data.

In cases where the computed pairwise and complete observation correlation matrices were nearly identical—defined as having an absolute difference of less than 0.10 between corresponding coefficients—clustering was conducted only on the complete observation correlation matrix. Since hierarchical clustering relies on correlation-derived distances, using two highly similar correlation matrices would have produced nearly identical clusters, making a second clustering run unnecessary. By focusing on the complete observation correlation matrix in these cases, the analysis remained consistent while avoiding redundant computations.

In years in which only two digitization indices were available (i.e., no more than two datasets), hierarchical clustering was not performed. As hierarchical clustering is designed to reveal complex relationships among multiple objects, only two datasets would yield a trivial dendrogram with limited interpretive value (Milligan & Cooper, 1985). Instead, only correlation analysis was conducted to assess the relationship between the indices for those years.

4.3.3 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a bottom-up approach, meaning that each observation starts as its own cluster, and clusters are successively merged based on their closeness, as determined by a dissimilarity matrix (Brock et al., 2008). This approach generates a dendrogram, which can be cut at a chosen height to produce the desired number of clusters. This method was selected for its proven applicability to real-world data (Bouguettaya et al., 2014) and its ability to present data at multiple levels of abstraction, making it particularly effective for visualizing and exploring large datasets. Furthermore, its extensive use in pattern recognition tasks underscores its utility in extracting meaningful insights from complex data (Gil-García et al., 2006).

Before performing hierarchical clustering, it was necessary to ensure that the correlation matrices used as input were suitable for the clustering process. As explained above, pairwise computation allows for varying sample sizes across variable pairs, the resulting correlation matrix can sometimes lose its positive semi-definite property. To verify its suitability, the pairwise complete observation correlation matrix was tested for symmetry and positive semi-definiteness using eigenvalue decomposition (eigen() from base R; *r* Core Team, 2023). If the matrix failed these checks, it would have been replaced with the nearest positive semi-definite matrix (Higham, 1988). However, in this analysis, all computed correlation matrices were already positive semi-definite, so no adjustments were required.

To transform the correlation values into a dissimilarity measure suitable for hierarchical clustering, the correlation values were transformed into a distance measure using 1 - r, ensuring that indices with stronger correlations were placed closer together, while those with weaker or negative correlations were treated as more dissimilar (Kaufman & Rousseeuw, 1990). Hierarchical clustering was then conducted using the hclust() function from the stats package (R Core Team, 2023), applying the average linkage method to compute cluster distances as the mean of all pairwise distances between elements in different clusters. Hierarchical clustering was then applied using the average linkage method, which calculates cluster distances as the average of all pairwise distances between elements in different clusters. This method was chosen because it balances the tendencies of single linkage clustering, which can produce elongated chains, and complete linkage clustering, which tends to create compact clusters but is less robust to noise in the data (Jarman, 2020). For years in which the difference between the coefficients of the two matrices exceeded the defined threshold of 0.10, the process was repeated using the pairwise complete correlation matrix

To assess the stability of the clustering results, an alternative clustering method using complete linkage was also applied to both matrices, producing similar outcomes and reinforcing the reliability of the approach.

4.3.4 Cluster Validation

To evaluate the clustering structure, several widely used and proven validation techniques (Kassambara, 2020; Brock et al., 2008) were applied. Cluster validation is essential because clustering algorithms will always produce groupings, even if no meaningful structure exists. The following statistical measures were used to assess cluster quality, with a predefined upper limit on the number of clusters ($\mathrm{k}_{\{\mathrm{max}\}} = \mathrm{n} - 1$) to avoid trivial or degenerate solutions. This limit is based on observations in Milligan and Cooper (1985), who noted that letting k approach n can yield misleading clustering outcomes:

- Silhouette Analysis: Measures how well each observation is clustered by estimating both its cohesion within the assigned cluster and its separation from other clusters. A silhouette value close to 1 indicates a well-clustered observation, while values near 0 suggest overlap between clusters, and negative values indicate potential misclassification (Kaufman & Rousseeuw, 1990).

- Within Sum of Squares (WSS): Evaluates cluster compactness by measuring the total variance within each cluster. Lower WSS values indicate tighter, more cohesive clusters, while higher values suggest greater dispersion. The relative change in WSS across different numbers of clusters provides meaningful insight into cluster quality. This is commonly evaluated via the “elbow” method, which identifies the point (number of clusters) at which further reductions in WSS become marginal, indicating a natural cut-off for k. (Kaufman & Rousseeuw, 1990).
- Gap Statistic: Compares the observed clustering structure to a randomly generated dataset, ensuring that the identified clusters are meaningful and not artefacts of the data structure. A cluster solution that maximizes the gap statistic suggests that the data exhibits more pronounced structure than would be expected by chance, indicating a more meaningful clustering solution. (Tibshirani et al., 2001).
- Dunn Index: Assesses both the separation and compactness of clusters by computing the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. The index ranges from zero to infinity, with higher values indicating well-separated clusters with strong internal cohesion (Kaufman & Rousseeuw, 1990).

These validation measures provided complementary insights into the quality and reliability of the clustering results, ensuring robustness across different methodological choices.

4.3.5 Cluster Visualization

The results of the hierarchical clustering were visualized using dendrograms, which represent the nested structure of clusters and illustrate their relative proximity (Bouguettaya et al., 2014). Each branching point (node) indicates a merging of clusters, with the height reflecting their level of dissimilarity—lower branches indicate greater similarity, while higher joins suggest weaker relationships. In addition to computing cluster validation measures, these dendrograms served as a further means of assessing the clustering solution. To enhance readability, the dendrograms were customized using the dendextend package, which allows for the adjustment of branch colors, thickness, and labels to improve interpretability. These enhancements ensured a clear distinction between clusters and facilitated intuitive exploration of the hierarchical structure (Galili, 2015). The visualizations will be further analysed in the results chapter.

While hierarchical clustering is a useful method for uncovering relationships among digitization indices, it is important to note that the clustering is based solely on statistical correlation. This means that indices grouped together do not necessarily measure the same underlying concept but rather exhibit similar patterns in their data. Additionally, determining the optimal number of clusters requires domain knowledge to ensure that the groupings are meaningful in context. While these factors introduce some constraints, the cluster analysis provides a systematic and data-driven approach to identifying structural patterns in digitization indices. The findings from this methodology form the foundation for further exploration in the results chapter.

4.4 Breadth of Indices

For digitization indices where indicator-level data was available, an additional analysis was conducted to assess the breadth of each index. This was done by examining the correlation structure of the indicators, providing insight into whether an index covered a wide range of different aspects or represented a more homogenous measure. High correlations among indicators suggest redundancy that could compromise the index’s diversity (Maggino & Zumbo, 2012).

For each index, a pairwise complete correlation matrix was computed for each year, where data was available, capturing the relationships between its indicators. This was done across all years from 2004 to 2023. An additional matrix was computed for the entire time period combined, capturing the overall relationships between the indicators of each index over time. The overall breadth of an index was summarized using well established key statistics (Christopher, 2017), including the mean, median, standard deviation, and minimum and maximum correlation values. The mean and median indicated how strongly the indicators were related on average, while the standard deviation assessed the variability in how closely different indicators were related, with higher values suggesting greater variation in the relationships between indicators. The minimum and maximum correlations showed the range of relationships, capturing the extent to which different components of an index were either strongly or weakly associated. A higher average correlation suggested that the index components were closely related, indicating that they measured similar aspects of digitization. In contrast, a lower average correlation indicated greater diversity among components, suggesting that the index covered a broader range of distinct dimensions.

In cases where no indicator-level data was available, the same approach was applied to sub-indices, provided they were present. However, some datasets only reported the main index value without any further disaggregation, preventing an assessment of internal structure. The differentiation between indices with rich indicator data and those reporting only aggregate values highlights a key limitation in cross-index comparisons, as indices with greater internal granularity allow for a deeper exploration of their composition and relationships with other measures.

This analysis provided a way to assess whether a digitization index captured a diverse set of factors or was more focused on a specific dimension of digital development.

4.5 Correlation Analysis

Following the cluster analysis, a correlation analysis was conducted to examine statistical relationships between digitization indices and linguistic diversity measures. As demonstrated by Jovanovic et al. (2018, p. 22), correlation analysis is a practical tool for exploring such relationships; they note that while correlation coefficients do not establish causality, they serve as a quantitative indicator of both the strength and the direction of the link between variables. Similarly, Archibugi et al. (2009) employed correlation techniques to examine the interplay among synthetic measures, reinforcing the utility of this approach in uncovering underlying relationships within complex datasets.

As with the cluster analysis, this step was performed separately for each year between 2004 and 2023, ensuring that only indices available in a given year were included. This year-wise computation followed the same logic as the clustering analysis, accounting for differences in temporal and regional coverage among the datasets. The primary objective was to examine the relationship between digitization indices and linguistic diversity measures, specifically entropy-based metrics, the number of languages spoken in a country, and the adapted Red List Index (RLI) used to assess linguistic diversity, while also exploring potential patterns over time. The number of languages spoken per country were log transformed for the analysis. Log transformations are commonly employed to stabilize variance and mitigate the influence of outliers (Feng et al., 2014), thereby reducing the sensitivity of correlations to extreme values. It is important to note that, while the digitization indices data were available for different years between 2004 and 2023, the linguistic diversity measures used in this study are static and based solely on 2023 data. Consequently, the effects discussed in the results section refer to long-term impacts, as evidenced by the 2023 data, rather than direct effects for those specific years.

The correlation analysis was carried out separately for entropy, the number of languages spoken per country, and the RLI, examining how these diversity measures were associated with the digitization indices values from each year. While the individual correlations of each diversity measure were examined independently of each other, the results across both diversity measures could be compared to assess whether similar patterns emerged, or whether entropy, national language counts, and the RLI captured different aspects of the relationship between linguistic diversity and digitization. This approach provided a comprehensive view of how digitization indices were associated with linguistic diversity, allowing for a nuanced interpretation of the findings in the results chapter.

4.5.1 Dataset Construction and Correlation Computation

The dataset construction process followed the same methodology as in the correlation analysis conducted for clustering, ensuring consistency in data merging and missing value handling. As outlined in the cluster analysis chapter, datasets were merged using both full joins and inner joins, which were implemented via the dplyr package (Wickham et al., 2023). The full join retained all available data across indices, even if missing values were present in some variables, maximizing data availability. The inner join, in contrast, included only observations where all selected indices contained complete data, ensuring full comparability across variables.

As in the correlation analysis for clustering, Pearson's correlation coefficient (Pearson's *r*) was computed to assess the strength and direction of relationships between digitization indices and linguistic diversity measures. This was carried out using the cor() function from base *r* (R Core Team, 2023). Both complete observation and pairwise complete observation correlation matrices were generated, following the exact approach described in the cluster analysis chapter. Effect sizes were interpreted based on Cohen's guidelines for correlation coefficients, where values of 0.1, 0.3, and 0.5 were considered small, moderate, and large effects, respectively (Cohen, 1988). Correlation analysis quantifies the degree of linear association between two variables, with Pearson's *r* values ranging from -1 to 1, where positive values indicate positive associations, negative values indicate inverse relationships, and values near zero suggest weak or no correlation (Cohen, 1988).

4.5.2 Visualization

Visualizations were generated to illustrate relationships between digitization indices and linguistic diversity measures. As a first step, heatmaps were created using the ggcorrplot package (Kassambara, 2023), providing an overview of correlation strength and direction across indices. Heatmaps are graphical representations where individual values in a data matrix are depicted as colours, facilitating the identification of patterns and correlations within large datasets (Kassambara, 2023). Additionally, scatterplot matrices were generated using the pairs() function from base *r* (R Core Team, 2023). Scatterplot matrices display multiple pairwise scatter plots in a grid format, allowing for simultaneous visualization of relationships between all selected variables. Each subplot represents the relationship between two variables, facilitating the identification of trends, clusters, and potential outliers.

This chapter outlined the methodological approach used to examine the relationships between digitization and linguistic diversity. The analysis was conducted in R, incorporating both hierarchical clustering and correlation analysis to explore patterns across multiple years. The clustering analysis identified structural similarities among digitization indices, while the correlation analysis assessed their associations with linguistic diversity measures, specifically the number of languages spoken in each country, entropy-based metrics and the adapted Red List Index (RLI) used in this thesis. Given the varying availability of data across years and indices, a year-wise approach was applied to ensure consistency, following the same logic in both analyses.

The following chapter presents the results of the cluster and correlation analyses, discussing key findings and patterns that emerge from the data.

5 Results

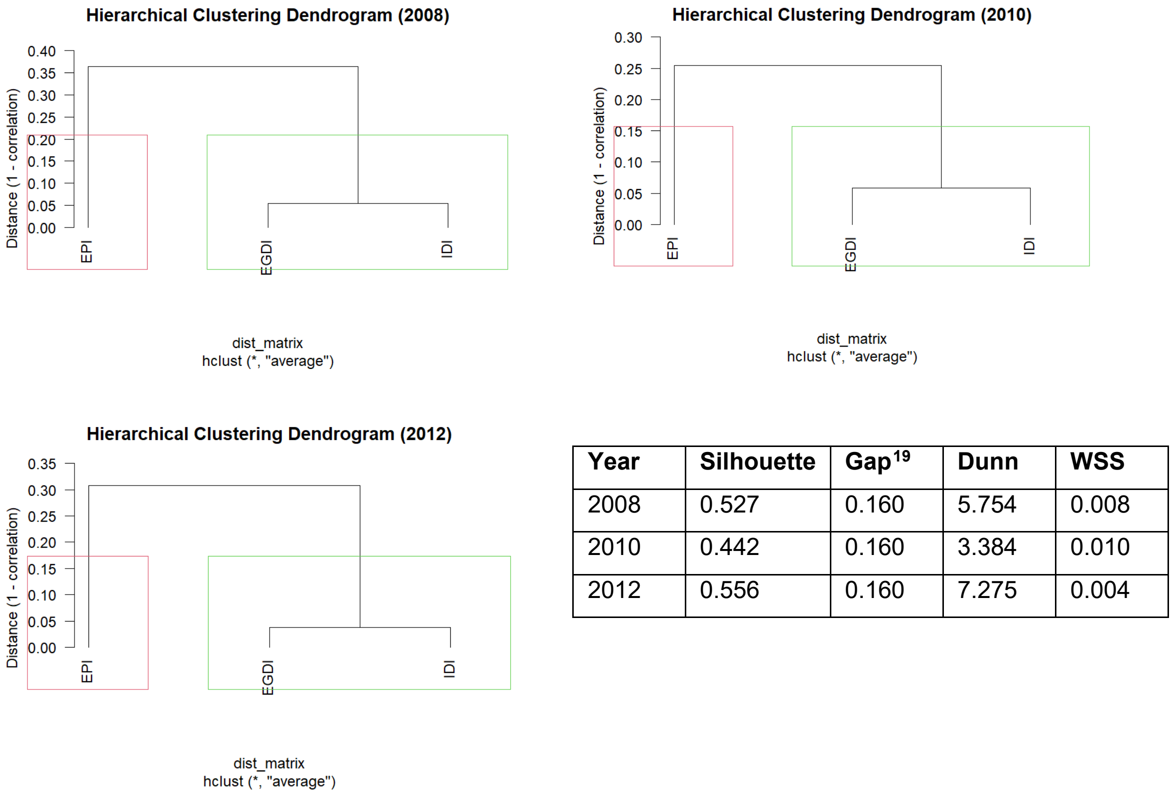
This chapter presents the outcomes of the analyses conducted in this study, building on the methodological framework detailed in the previous chapter. The results are presented in two parts: the first focuses on identifying patterns within and across digitization indices using cluster and correlation analyses, while the second examines the strength and nature of the relationships between these indices and linguistic diversity metrics. Together, these findings shed light on the evolving landscape of digital transformation and its potential implications for language diversity.

5.1 Cluster Analysis

This section details the grouping of digitization indices as determined by hierarchical clustering. Selected samples of clustering outcomes from different years are presented to demonstrate how the indices were grouped based on their statistical similarities. A comprehensive set of results for all years is provided in Appendix A.2.5 and A.2.6, corresponding to complete and pairwise complete observations, respectively. For years with sufficient data, hierarchical clustering was applied as outlined in the methodology chapter; for years with only two indices available, clustering was omitted and the focus was instead placed on correlation analysis. The section concludes with a summary of the overall trends and developments observed over time.

For some years, the available data for clustering was either limited or insufficient for robust analysis. In 2006, 2007, and 2009, no data could be retrieved. In 2011 and 2013, only the ICT Development Index (IDI) was available, which did not permit meaningful grouping. During 2004 and 2005, the dataset comprised merely two indices—the E-Government Development Index (EGDI) and the E-Participation Index (EPI)—so hierarchical clustering was not applied. Instead, correlation analysis was performed, revealing a strong positive relationship between the two measures (2004: $r = 0.772$; 2005: $r = 0.771$).

In 2008, 2010, and 2012, although three indices (the EGDI, the EPI, and the IDI) were available, the resulting clusters (see Figure 5) offered limited insight. With only three variables, the multidimensional space is inherently constrained, and even minor variations in the data can lead to clusters that are less stable and harder to interpret (Everitt et al, 2011). In these years, the validation measures (within-cluster sum of squares [WSS], Dunn index, and Silhouette) supported a two-cluster solution. Moreover, visual inspection of the dendrogram generated by the hierarchical average linkage (hcavg) clustering algorithm confirmed that the EPI formed a distinct cluster, while the EGDI and the IDI clustered together. Although the gap statistic suggested a single cluster, this discrepancy may arise from its sensitivity to overall cluster separation, as it compares the observed structure against a null reference distribution (Tibshirani et al., 2001). While the EPI is technically a sub-index of the EGDI, it measures only a single dimension of digital governance, whereas both the EGDI and the IDI capture a broader spectrum of digitization aspects. This divergence is supported by the correlation analysis, which also indicates that the EPI behaves differently relative to the other indices, likely due to its narrower focus (2010: EPI–EGDI r = 0.776, EPI-IDI: 0.689, EGDI–IDI r = 0.941).



Year	Silhouette	Gap[19]	Dunn	WSS
2008	0.527	0.160	5.754	0.008
2010	0.442	0.160	3.384	0.010
2012	0.556	0.160	7.275	0.004

Figure 5 Hierarchical clustering dendrograms for 2008, 2010, and 2012, with corresponding validation metrics (bottom right).

[19] The gap statistic indicated a single-cluster solution (k = 1) for all three years, while other metrics favoured two clusters.

Results

In 2014, most validation measures and the dendrogram visualization (see Figure 6) indicated a three-cluster solution, whereas the Silhouette score —illustrated in Figure 7— favoured two clusters for both the clustering algorithm based on complete observation correlations and the one using pairwise complete observations. The Digital Service Trade Restrictiveness Index (DSTRI), which was inverted as described in the methodology, formed its own cluster alongside the E-Participation Index (EPI) under the three-cluster arrangement. A similar discrepancy arose in 2019, when the three-cluster arrangement instead isolated the DSTRI and the Network Readiness Index (NRI) in separate clusters. Notably, in that year, the NRI emerged as even more distinct than the DSTRI, marking a departure from other years where the DSTRI was typically the primary outlier. However, when clustering was based on the complete observation correlation matrix, the gap statistic suggested a single cluster, while the pairwise observation correlation matrix resulted in a Silhouette score, which favoured two clusters—even as most other measures continued to support three clusters. The Silhouette score, which evaluates each point's cohesion within its assigned cluster relative to its separation from other clusters, is particularly sensitive to borderline observations and overlapping cluster boundaries (Kaufman & Rousseeuw, 1990). Consequently, it may diverge from other metrics when the data contain clusters of varying density or when different correlation matrices (complete vs. pairwise) are used, underscoring that no single validation measure can capture all aspects of cluster structure. Moreover, as discussed earlier, the gap statistic's sensitivity to overall cluster separation can occasionally lead it to favour a single-cluster solution if the observed separation does not sufficiently exceed that of a null reference distribution.

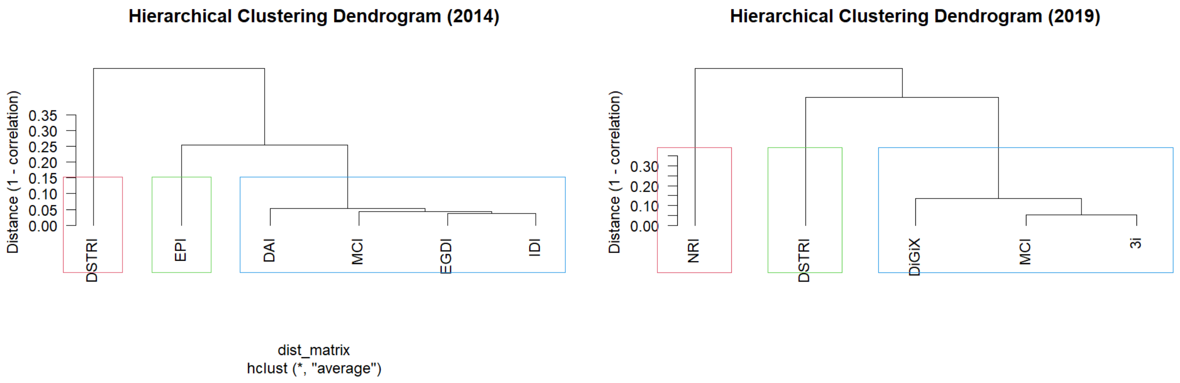
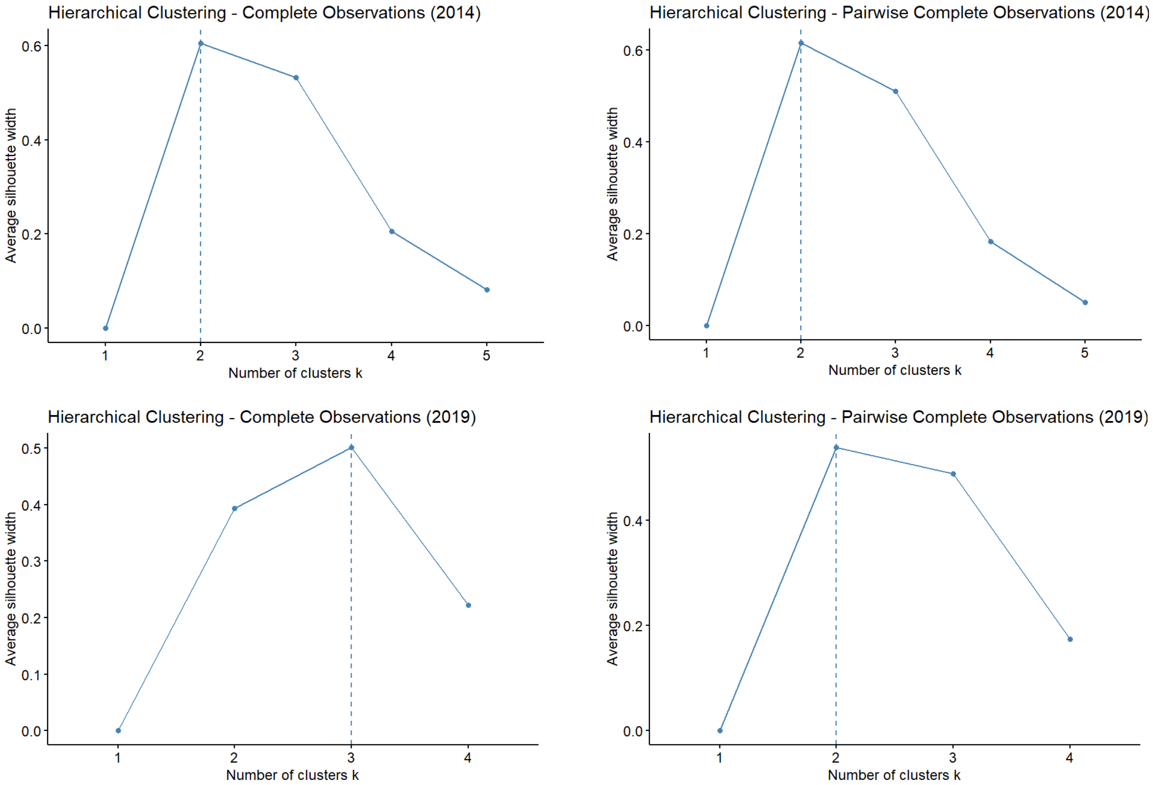


Figure 6 Hierarchical clustering dendrograms for 2014 and 2019 using average linkage (complete observation matrices).

The dendrograms (Figure 6) highlight how the DSTRI consistently and unambiguously forms its own cluster, underscoring its distinct focus on barriers to digitally traded services. This narrower scope sets it apart from broader digitization indices, helping to explain its tendency to appear as an outlier in the clustering.


Hierarchical Clustering - Complete Observations (2014)

Hierarchical Clustering - Pairwise Complete Observations (2014)

Hierarchical Clustering - Complete Observations (2019)

Hierarchical Clustering - Pairwise Complete Observations (2019)

Figure 7 Silhouette plots for 2014 and 2019, comparing clustering solutions derived from complete and pairwise complete observation matrices.

While in 2014 new indices such as the Mobile Connectivity Index (MCI) and the Digital Service Trade Restrictiveness Index (DSTRI) were introduced, the analysis in 2015 remained limited to just three variables: the MCI, the DSTRI, and the earlier version of the ICT Development Index (IDI). A similar situation was observed in 2017, where the dataset included the MCI, the DSTRI, and the Inclusive Internet Index (3i). With only three variables available, the multidimensional space was naturally constrained, echoing the challenges encountered in 2008, 2010, and 2012.

In both 2015 and 2017, the cluster validation measures consistently supported a two-cluster solution for the average linkage clustering algorithm. For 2015, the within-cluster sum of squares (WSS) for k = 2 was 0.002, the Silhouette score reached 0.606, and the Dunn index was 9.334 (with a three-cluster configuration yielding an infinite value, reinforcing the preference for two clusters). Similarly, in 2017, the WSS and Silhouette score for k = 2 were 0.001 and 0.606, respectively, with the Dunn index at 18.457 further confirming the two-cluster outcome.

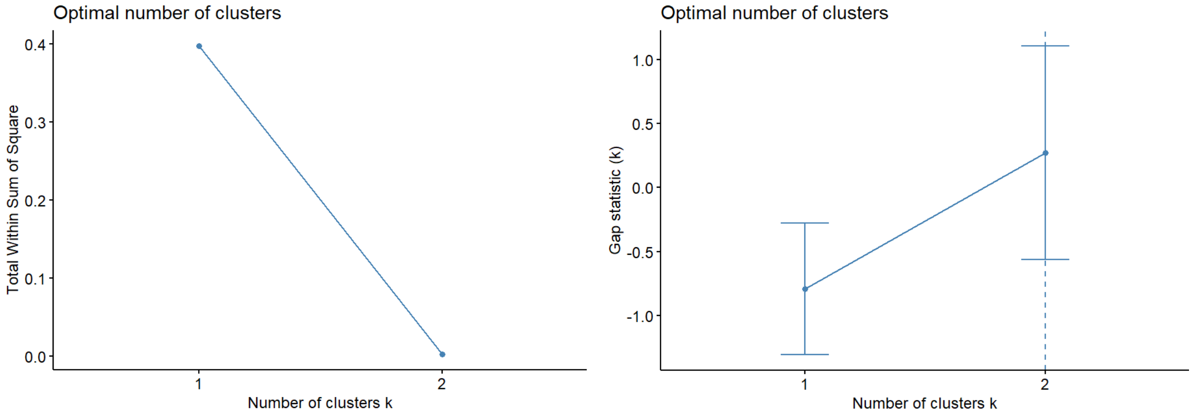


Figure 8 WSS and gap statistic (2015): evidence for a two-cluster solution.

As illustrated in Figure 9, visual inspection of the dendrograms corroborates these numerical results, showing that the Digital Service Trade Restrictiveness Index (DSTRI) consistently formed a separate cluster, while the Mobile Connectivity Index (MCI) grouped with the earlier version of the ICT Development Index (IDI) in 2015 and with the Inclusive Internet Index (3i) in 2017. This clustering structure is also supported by the correlation analysis (see Table 3), which indicates stronger correlations within each cluster and comparatively weaker correlations across clusters.

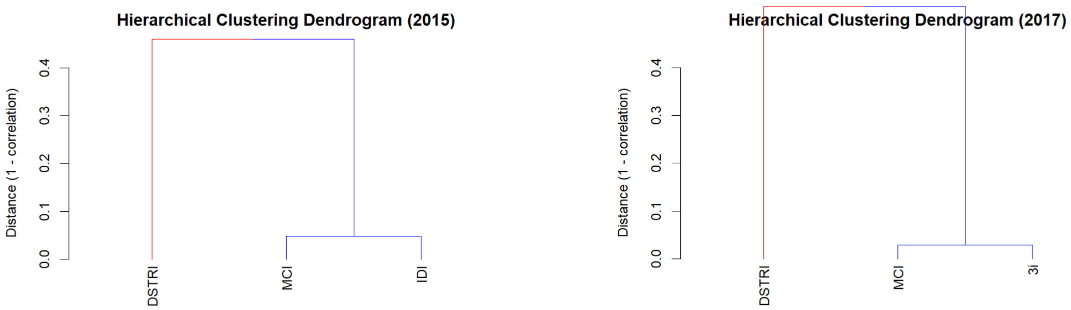


Figure 9 Hierarchical clustering dendrograms for 2015 (left) and 2017 (right).

	MCI	DSTRI	IDI
MCI	1.000	0.533	0.951
DSTRI	0.533	1.000	0.546
IDI	0.951	0.546	1.000

	MCI	3i	DSTRI
MCI	1.000	0.972	0.474
3i	0.972	1.000	0.469
DSTRI	0.474	0.469	1.000

Table 3 Correlation matrices for 2015 (left) and 2017 (right).

In 2016 the clustering validation measures conflicted with the structure suggested by the average-linkage dendrograms (See Figure 10). While the validation measures predominantly indicated a two-cluster solution, a visual inspection of the dendrograms implied the possibility of a three-cluster arrangement in both years. In both cases, the inverted DSTRI was identified as an outlier for the two-cluster solution, whereas the three-cluster arrangement further isolated the EPI as its own cluster. It is not uncommon for different clustering validation methods to yield conflicting solutions, particularly when comparing purely numerical measures to visual assessments of dendrograms. Each validation measure emphasizes a specific dimension of cluster structure, whereas a dendrogram-based approach can highlight different nuances. As a result, the optimal solution suggested by numerical indices may not always align with what appears most intuitive upon visual inspection (Everitt et al., 2011).

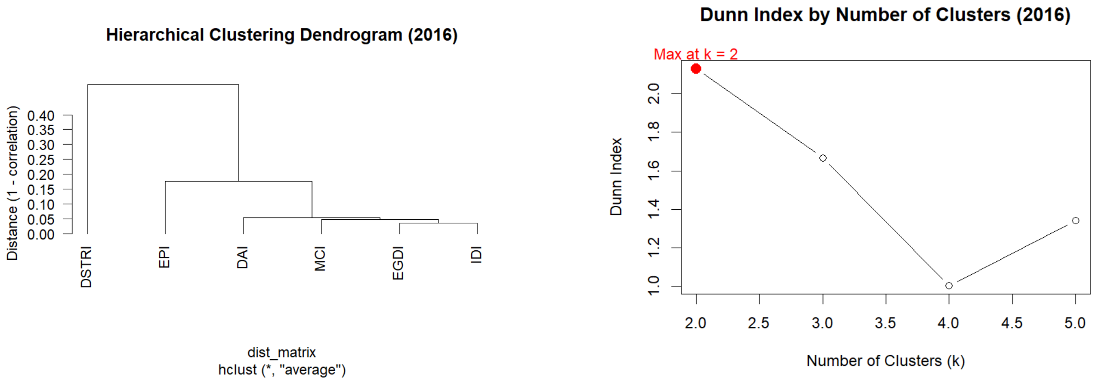


Figure 10 Dendrogram and Dunn index plot for 2016, illustrating the conflicting recommendations for the optimal number of clusters.

For 2018, the cluster validation measures provided mixed signals regarding the optimal number of clusters. Both the WSS and Silhouette scores favoured a two-cluster solution, whereas the Gap statistic indicated four clusters. The Dunn index was nearly tied between two and four clusters, yielding 2.107 for k=2 and 2.106 for k=4, thus leaning only very slightly toward k=2. As illustrated by the two dendrograms (see Figure 11), the two-cluster arrangement places the DSTRI on its own while grouping all remaining indices together. In contrast, the four-cluster arrangement isolates the DSTRI, the EPI, and the Digitization Index (DiGiX) in separate clusters, leaving the remaining indices to form a final group. This discrepancy underscores how different validation measures emphasize various aspects of cluster structure, highlighting the importance of visual inspection and contextual understanding when determining the most meaningful clustering outcome. Notably, 2018 was the only year in which the validation measures were evenly split (two against two), whereas in every other year at least three out of four measures converged on a single solution.

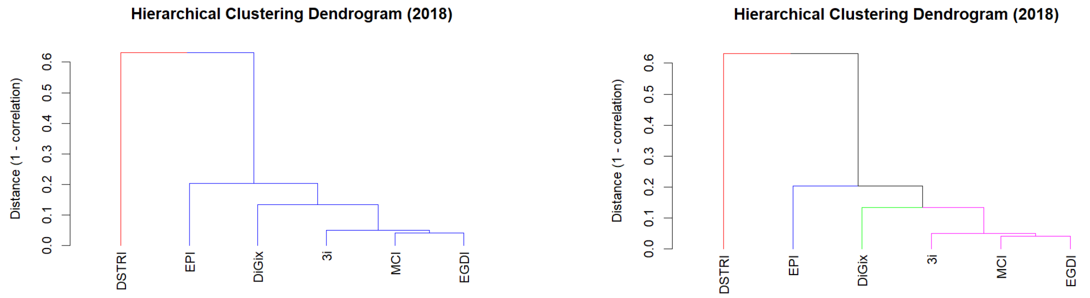


Figure 11 Hierarchical clustering dendrograms for 2018 showing the two-cluster solution (left) and the four-cluster solution (right).

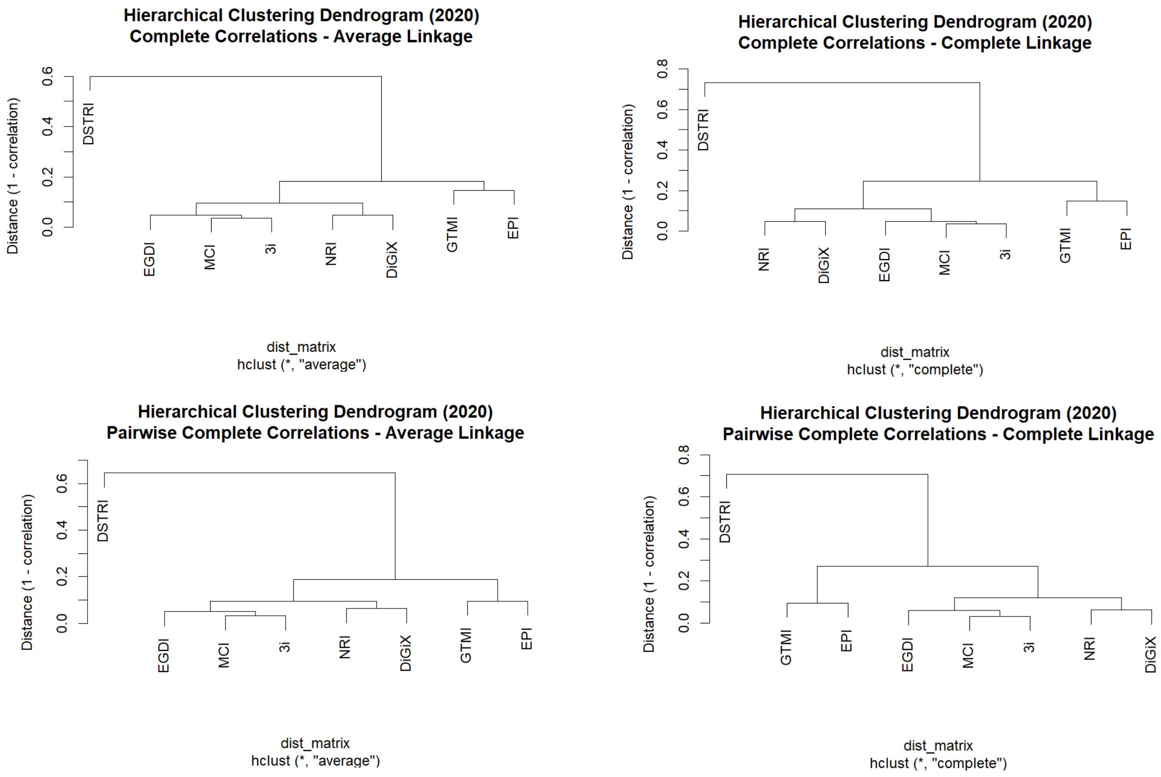
A visual inspection of the dendrogram (see Figure 11) indicates that the Digital Service Trade Restrictiveness Index (DSTRI) separates from the remaining indices at a comparatively large distance, suggesting that it naturally forms its own cluster. This observation is reinforced by the correlation values (see Table 4), where DSTRI shows relatively weak associations with the other indices (0.325 with the EPI to 0.436 with the DiGiX), whereas the rest maintain strong inter-correlations (typically above 0.700). While a four-cluster solution would further isolate indices like the EPI or the DiGiX, these correlation patterns show that those indices still align more closely with the main group than with the DSTRI. Consequently, a two-cluster solution—one consisting solely of the DSTRI and the other comprising the remaining indices—appears more coherent overall.

	MCI	**3i**	**DSTRI**	**DiGiX**	**EGDI**	**EPI**
MCI	1.000	0.950	0.367	0.879	0.958	0.785
3i	0.950	1.000	0.328	0.832	0.946	0.823
DSTRI	0.367	0.328	1.000	0.436	0.387	0.325
DiGiX	0.879	0.832	0.436	1.000	0.887	0.732
EGDI	0.958	0.946	0.387	0.887	1.000	0.842
EPI	0.785	0.823	0.325	0.732	0.842	1.000

Table 4 Correlation matrix for 2018, illustrating the weaker correlations of the DSTRI relative to other indices.

In all the remaining years, a two-cluster solution consistently isolated the DSTRI in its own group. In 2020, four dendrograms were generated by combining complete vs.

pairwise observation matrices with average vs. complete linkage (see Figure 12). Although these dendrograms exhibit slight differences in branching order, the majority of validation measures still converged on two clusters, with the DSTRI forming its own cluster. The one exception occurred when the pairwise complete-based matrix used complete linkage, causing the Gap statistic to suggest three clusters—making 2020 the only instance where complete linkage yielded a notably different structure than average linkage. Despite this outlier result, the overall evidence for 2020 continued to favour a two-cluster arrangement, confirming the DSTRI's status as the primary outlier. A similar two-cluster pattern emerged in 2021, again separating the DSTRI from the other indices, and it recurred in 2023 as well. In 2022, most validation measures again indicated two clusters, although the Gap statistic suggested three clusters. The dendrogram visualizations also differed based on whether complete or pairwise observation matrices were used: under complete observations, the IDI formed the third cluster, whereas under pairwise observations, the IDI and the EPI clustered together. Nevertheless, the final consensus for 2022 also supported a two-cluster arrangement, consistently placing the DSTRI in its own distinct group.



Hierarchical Clustering Dendrogram (2020)
Complete Correlations - Average Linkage

Hierarchical Clustering Dendrogram (2020)
Complete Correlations - Complete Linkage

Hierarchical Clustering Dendrogram (2020)
Pairwise Complete Correlations - Average Linkage

Hierarchical Clustering Dendrogram (2020)
Pairwise Complete Correlations - Complete Linkage

Figure 12 Hierarchical clustering dendrograms for 2020, illustrating both complete and pairwise-complete correlation matrices with average and complete linkage methods.

Results

Over the examined years, hierarchical clustering consistently underscored how certain indices—most notably the Digital Service Trade Restrictiveness Index (DSTRI)—occupy outlier positions, reflecting their more specialized scope compared to more comprehensive digitization measures. In most instances, the validation metrics (e.g., WSS, Silhouette, Gap, and Dunn) and the dendrograms agreed on a two-cluster structure, typically separating DSTRI from the remaining indices. Nevertheless, a few years revealed notable discrepancies—specifically in 2014 and 2019, where a three-cluster solution clearly emerged, indicating that at least three validation measures converged on isolating not just the DSTRI but also another index (the EPI in 2014 and the NRI in 2019). In contrast, in other years while the majority of metrics supported a particular cluster count, the Silhouette score or Gap statistic occasionally proposed an additional cluster, highlighting the unique sensitivities of each measure. This pattern confirms the dominance of two-cluster solutions overall, yet shows how additional clusters can materialize under certain conditions.

Visual inspection of the dendrograms helped clarify these divergences by showing how certain indices—the EPI in 2008–2010, for instance—can briefly form their own clusters before re-joining the main group in later years. In 2018, the divergence between a two-cluster and a four-cluster solution was especially pronounced, with the DiGiX sometimes joining the DSTRI and the EPI as separate clusters, underscoring borderline distinctions that become more apparent when more clusters are allowed. Despite these differences, the primary takeaway remains that the DSTRI stands out as the most distinct index in most scenarios, suggesting that its specialized focus on barriers to digitally traded services consistently separates it from broader measures of digitization. In 2019, however, the Network Readiness Index (NRI) emerged as even more distinct than the DSTRI, illustrating how changing data availability or index revisions can alter clustering outcomes.

At the same time, all of the remaining indices—the AI Preparedness Index (AIPA), the Digital Adoption Index (DAI), the E-Government Development Index (EGDI), the GovTech Maturity Index (GTMI), the ICT Development Index (IDI_new), the earlier version of the ICT Development Index (IDI_old), the Inclusive Internet Index (3i), and the Mobile Connectivity Index (MCI), were consistently clustered together, reflecting strong correlations among their underlying measures. Their tendency to remain in the same cluster across multiple years and validation approaches suggests a high degree of similarity in how these indices capture various dimensions of digitization.

Overall, these patterns confirm that no single validation measure can capture all aspects of cluster quality; combining quantitative metrics with dendrogram-based visual assessments provides the clearest understanding of how digitization indices relate to one another. While many indices share overlapping dimensions of digital development, a handful—especially the DSTRI—regularly departed from the broader group, and occasional shifts in indices such as the EPI or the DiGiX highlight how methodological changes or new data can influence cluster assignments over time.

5.2 Breadth of Indices Analysis

In this section, the results of the correlation analysis, which was utilized to capture the relationships and variations between the elements of each index in order to assess its breadth, are presented. This was done for each year where data was available, as well as across all years from 2004 to 2023.

As mentioned in the methodology section, the granularity of data availability varied among the indices used in this study. The Digitization Index (DiGiX) and the E-Participation Index (EPI) included only the main index values, preventing a deeper analysis of their internal structure and composition. For the AI-Preparedness Index (AIPI), the Digital Adoption Index (DAI), and the Digital Services Trade Restrictiveness Index (DSTRI), more granular data was available only at the sub-index level. Data at the indicator level was not provided for these indices. As a result, the analysis of these indices could be conducted with greater detail than for indices that include only main index values, but not to the same extent as for the remaining indices where data was available on all levels.

5.2.1 Breadth Analysis based on Sub-Indices

The AIPI was only included in the analysis in 2023, where the correlations across its sub-indices reached a median of 0.764, while ranging from 0.648 to 0.884, with a standard deviation of 0.101. This suggests that there is some variability in the underlying data, but overall, the index appears to focus on a relatively narrow and coherent set of aspects related to digitization.

The sub-indices of the DAI, available in the years 2014 and 2016, show moderately strong overall correlations across both years, with a median of 0.674. However, the standard deviation of 0.154 and a wide range from 0.608 to 0.902 indicate considerable variability in the relationships among the sub-indices. This suggests that while the DAI maintains a general internal coherence, it also reflects a broader and more diverse set of dimensions related to digital adoption, with some sub-indices more closely aligned than others.

In contrast, the DSTRI, which covered a relatively broad range of years from 2014 to 2023, showed very weak correlations among its sub-indices over the entire period. The median correlation was at its lowest in 2020 (*m* = 0.230) and at its highest in 2018 (*m* = 0.296), still implying weak internal relationships. The broad ranges — 2014: *min* = 0.098, *max* = 0.538; 2020: *min* = 0.148, *max* = 0.405 — suggest some variance in how the sub-indices relate to each other over time. The standard deviation of correlation values varied from 0.010 (2019) to 0.136 (2023), suggesting a moderate degree of inconsistency in how strongly the components are related, with no clear pattern of increasing or decreasing structural coherence across the years. This variability may reflect the conceptual breadth of the DSTRI, as it appears to encompass distinct, loosely connected dimensions of digital services trade restrictiveness.

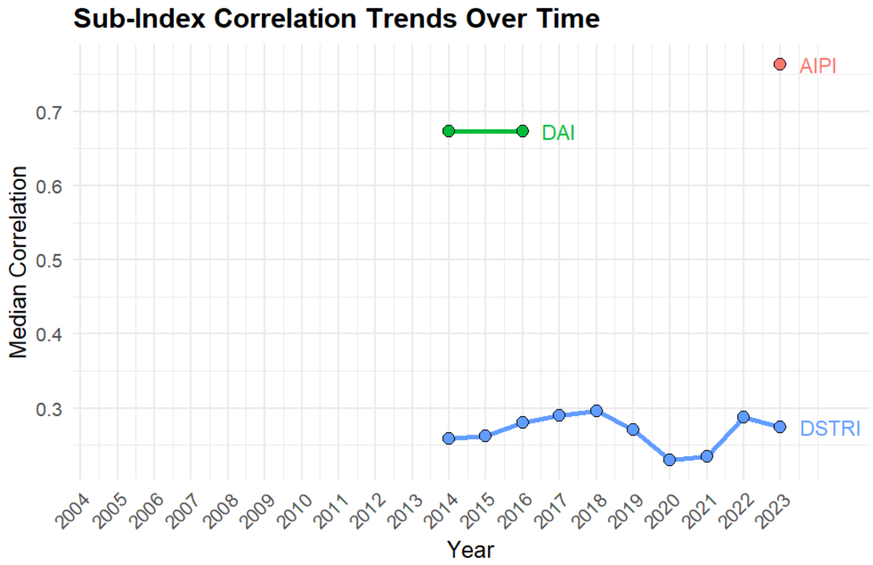


Figure 13 Correlations between sub-indices of the digitization indices over time.

5.2.2 Breadth Analysis based on Indicators

Following indices included data on indicator level: the E-Government Index (EGDI), the Government-Tech Maturity Index (GTMI), the old and new version of the ICT-Development Index (IDI_Old; IDI_New), the Inclusive Internet Index (3i), the Mobile Connectivity Index (MCI) and the Network Readiness Index (NRI).

The EGDI spans a long period from 2004 to 2022, with generally moderate to strong correlations among its indicators in most years. The highest median correlation was observed in 2005 (m = 0.778), and the lowest in 2022 (m = 0.270). The earlier years show a relatively coherent structure — for example, 2004 (m = 0.763, sd = 0.108) and 2014 (m = 0.735, sd = 0.138) — while more recent data reflects increasing variability. In 2022, the standard deviation rose to 0.277 and the range extended from 0.027 to 0.945, indicating substantial divergence among the components. Overall, this suggests that while the EGDI historically reflected a focused set of dimensions, its internal consistency has declined over time, possibly due to changes in how e-government readiness is conceptualized or measured.

For the GTMI indicator values were only available in 2022, where the correlations among its indicators showed a median of 0.385, ranging from –0.064 to 0.941, with a standard deviation of 0.156. These values suggest a relatively weak internal consistency, with considerable variability in how the indicators relate to one another. This indicates that the GTMI likely captures a broad and diverse set of aspects related to Govtech maturity, rather than a tightly focused conceptual structure.

The new version of the IDI was available for 2021 and 2022, with low to moderate correlations among its indicators across both years. The median correlation was 0.347 in 2021 and 0.381 in 2022, with high variability in both years (sd = 0.545 and 0.558, respectively). The wide ranges — from –0.737 to 0.918 in 2021, and from –0.717 to 0.924 in 2022 — indicate a lack of strong internal consistency, suggesting that the new IDI captures a diverse and loosely connected set of digital development dimensions.

In contrast, the older version of the IDI, covering the period from 2008 to 2016, displayed much stronger internal relationships. Median correlations were consistently above 0.570, peaking in 2011 at 0.636. While some variation was present — with standard deviations ranging from 0.274 (2012) to 0.386 (2016) — the overall pattern reflects a more cohesive index structure. These results suggest that the older IDI focused more narrowly on a consistent set of indicators, while the newer version incorporates a broader and more heterogeneous mix of measures.

The 3i was available from 2017 to 2022 and consistently showed weak correlations among its indicators across all years. Median values remained low throughout, ranging from 0.109 in 2022 to a slightly higher 0.231 in 2017. Standard deviations were relatively high — for example, 0.385 in 2017 and 0.337 in 2019 — and the ranges were broad, with values extending as low as –0.824 and as high as 0.928. These patterns indicate that the index captures a wide and loosely connected set of dimensions, reflecting substantial variability in how its components relate to one another.

The MCI was available from 2014 to 2023 and showed moderately strong and stable correlations among its indicators throughout the entire period. Median correlations ranged from 0.530 in 2014 to 0.573 in 2019, with standard deviations remaining relatively low — between 0.203 and 0.234 — across all years. The ranges were wide but consistent, with minimum values typically below zero and maximum values close to 0.980. These results suggest that while the MCI captures a broad spectrum of mobile connectivity dimensions, it does so with a relatively coherent and stable internal structure.

The NRI was available from 2019 to 2023 and showed moderate correlations among its indicators across all years. Median values ranged from 0.400 in 2023 to 0.546 in 2020, with standard deviations between 0.199 and 0.245. The ranges were broad throughout, with minimum values reaching as low as –0.712 in 2022 and maximum values consistently near or above 0.97. These patterns suggest that while the NRI maintains a certain degree of internal coherence, its indicators reflect a relatively diverse set of dimensions related to network readiness.

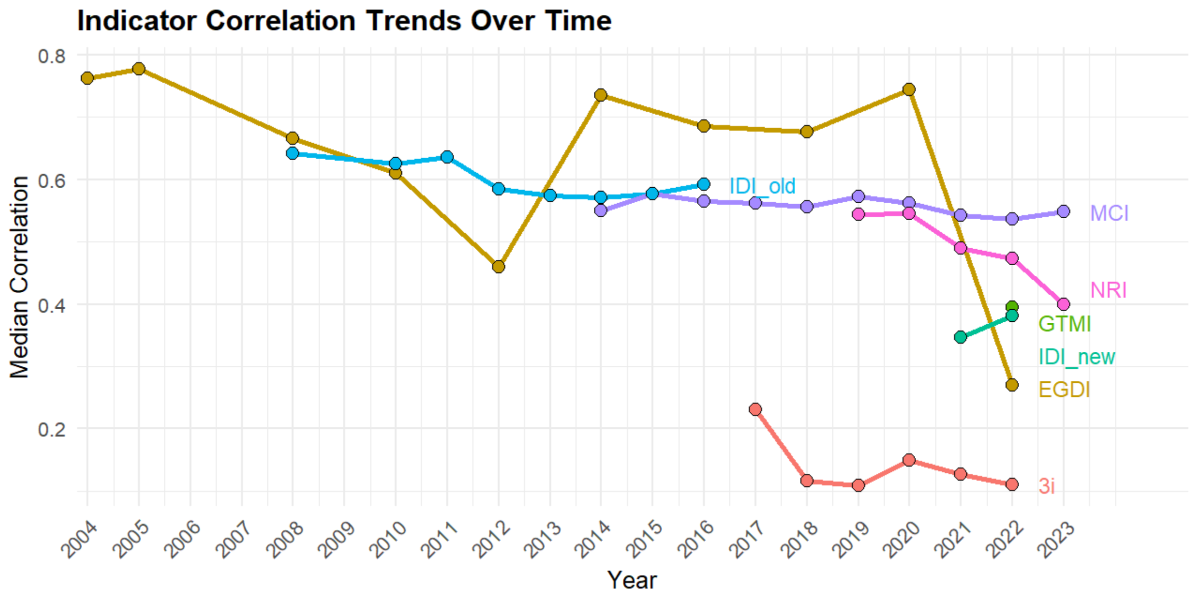


Figure 14 Correlations between indicators of the digitization indices over time.

In sum, the analysis reveals considerable variation in the internal coherence of digitization indices based on their indicator-level correlations. Indices such as the AIPI, the EGDI (in earlier years), and the MCI demonstrate relatively strong and stable internal relationships, suggesting a focused and coherent measurement structure. In contrast, indices like the DSTRI, the 3i, and the newer version of the IDI show weaker and more variable correlations, indicating a broader and more heterogeneous set of dimensions. Others, such as the DAI, the NRI, and the GTMI, fall in between, with moderate internal consistency but notable variability across components or over time. These differences point to distinct conceptual scopes and operational focuses across the indices, with some emphasizing tightly integrated constructs and others capturing more diverse aspects of digital development.

5.3 Correlation Analysis

The results of the correlation analysis examining the relationship between digitization indices and two language diversity measures—entropy-based metrics and an adapted version of the Red List Index (RLI)—are presented in this section. During the analyses, the logarithm (base) of the number of languages per country was used as an additional variable. As in the cluster analysis, significant results across different years and emerging patterns were focused on to provide an overview. The remaining results and visualisations can be found in appendix sections A.2.2 (entropy), A.2.3 (RLI) and A.2.4 (language counts). Each diversity measure was analysed individually, and the outcomes are subsequently compared to identify emerging patterns, similarities, and dissimilarities. Because the correlation analysis relied on the same digitization indices data sets as the cluster analysis, it encountered identical data availability limitations. Specifically, data were unavailable for the analysis in 2006, 2007, and 2009.

5.3.1 Correlation Analysis - Entropy

Early years showed notable negative correlations between digitization indices and entropy. In 2004, the Pearson correlation coefficients of both the E-Government Development Index (EGDI) (*r* = -0.285, *p* < .001) and the E-Participation-Index (EPI) (*r* = -0.165, *p* < .05) suggested that early digital government initiatives were associated with a less balanced distribution of language usage. Similar trends observed in 2005 reinforce the interpretation that early digital government initiatives contributed to a less even distribution of language use over the long term, as evidenced by the 2023 entropy data.

Variable1	Variable2	Correlation	P_Value	Effect_Strength	Significance	Direction
EGDI	EPI	0.772	p < 0.001	Large	*	Positive
EGDI	entropy	-0.285	p < 0.001	Small	*	Negative
EPI	entropy	-0.165	0.022	Small	*	Negative

Table 5 Entropy correlation analysis summary table for 2004 using pairwise complete observations.

In 2008, 2010, and 2012—when the ICT Development Index (IDI) became available alongside the EGDI and the EPI—the overall picture, based on Pearson’s correlation coefficients, showed that higher digitization was generally associated with lower entropy, again suggesting a less even distribution of language usage. For example, in 2008 complete analyses indicated that the EGDI was negatively correlated with entropy at –0.418, a trend that was maintained across pairwise analyses. Similar patterns were observed in 2010 and 2012, where all three indices consistently exhibited a negative association with entropy. Overall, these findings suggested that digitization during this period tended to lower entropy in the long run.

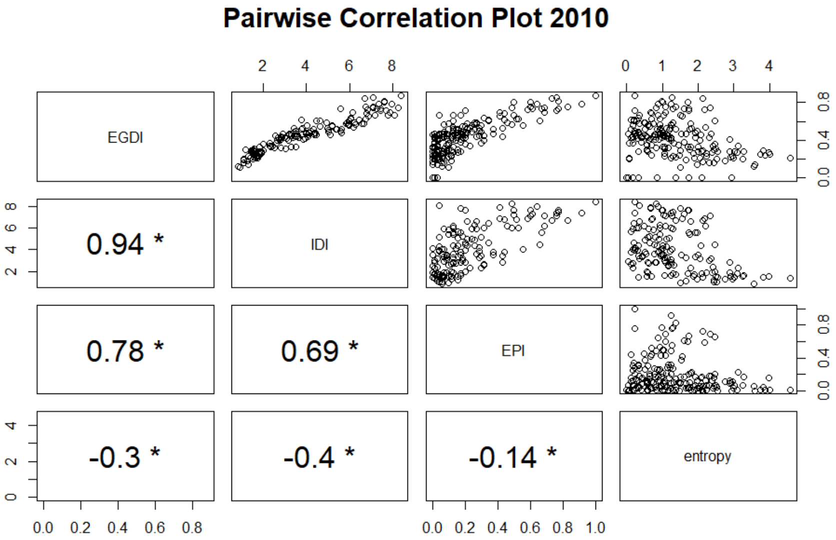


Figure 15 Entropy pairwise correlation plot (2010) displaying Pearson's correlation coefficients; asterisks (*) denote significance at the p < 0.05 level.

In 2014, 2015, and 2016—when new digitization indices, namely the Mobile Connectivity Index (MCI), the Digital Service Trades Restrictiveness Index (DSTRI), and the Digital Adoption Index (DAI), were introduced—correlations consistently indicated that higher digitization was associated with lower entropy based on 2023 speaker numbers. More precisely, the negative effect of the EGDI, the EPI and the IDI on entropy persisted in 2014, while the complete analyses revealed significant negative correlations between entropy and the MCI (*r* = –0.400) and the DAI (*r* = –0.400); notably, DSTRI's correlation was not significant. In 2015 and 2016, analyses continued to yield robust negative correlations.

In 2017, 2018, and 2019—and extending into 2021—a broader set of digitization indices were analysed, including the newly introduced Network Readiness Index (NRI), the Digitization Index (DiGiX), and the Inclusive Internet Index (3i), alongside the MCI, the DSTRI, the EGDI, the IDI, and the EPI. In 2017, analyses revealed robust negative correlations between the 2023 entropy and most indices (e.g., the EGDI, the IDI, and the EPI), while the DSTRI's association with entropy remained non-significant. In 2018, analyses continued to show negative associations between entropy and indices such as the MCI (*r* = –0.370) and the EGDI (*r* = –0.425), whereas the DSTRI and the DiGiX did not significantly predict entropy, and language count effects were minimal. In 2019, the MCI and the 3i maintained significant negative correlations with entropy (*r* =–0.344 and –0.436, respectively), while the

NRI, the DSTRI, and the DiGiX showed weaker or non-significant associations. Finally, in 2021, the refined set of indices—now including the NRI, the DSTRI, the IDI, and the 3i—continued to exhibit negative correlations with entropy (for example, the MCI at *r* = –0.335 and the EPI at *r* = –0.393), with DSTRI consistently non-significant.

2020 and 2022 were the years with the most comprehensive data available, covering eight digitization indices including the newly introduced GovTech Maturity Index (GTMI) and the updated ICT Development Index (IDI). Throughout both years, analyses continued to reveal that increased digitization is associated with reduced entropy, indicating a less even distribution of language usage. In 2010, it was found that significant negative associations with the 2023 entropy were present for the EGDI, the MCI, the GTMI, the 3i, the EPI, and the NRI (for instance, the EGDI exhibited an *r* value of –0.316 and the MCI –0.267), while the DSTRI and the DiGiX did not reach significance (*p* > 0.05).

In 2022, a similar overall pattern was maintained for entropy: the EGDI, the MCI, the NRI, the 3i, the EPI, and the updated IDI were significantly negatively correlated with entropy, whereas the GTMI and the DSTRI did not reach significance.

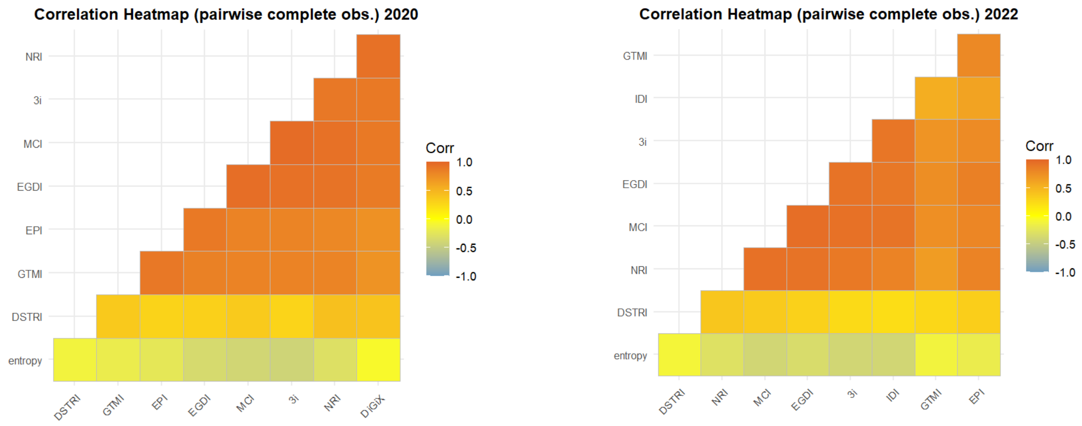


Figure 16 Entropy correlation heatmaps for 2020 and 2022 with colours indicating the strength of the correlations.

In 2023, correlation analysis was performed on four indices, including the newly added AI-Preparedness Index (AIPI). Overall, Pearson correlations showed that increased digitization was associated with reduced entropy; for example, the MCI and the NRI were significantly negatively correlated with entropy (with *r* values of –0.361 and –0.210, respectively), and AIPA also exhibited a significant negative association (*r* = –0.211), while the DSTRI's correlation remained weak and non-significant (*p* > 0.05). These findings confirm the enduring trend that increased digitization led to reduced entropy in 2023.

Overall, the analysis reveals a persistent inverse relationship between digitization and entropy, indicating that as digitalization increases, language usage becomes less evenly distributed. In the early years (2004–2005), indices such as the EGDI and the EPI were significantly negatively correlated with entropy, suggesting that early digital initiatives contributed to an uneven distribution in the long run. This overall pattern continued through 2008–2012. From 2014 to 2016, with the introduction of additional indices like the MCI and the DAI, the strong inverse association with entropy was maintained. In 2020 and 2022—the years with the most comprehensive data available, each covering eight indices—it was found that significant negative correlations with entropy were again predominant with only a few exceptions (e.g., the DSTRI and, in 2020, the DiGiX did not reach significance). Finally, in 2023, with the addition of the AI Preparedness Index (AIPI), the enduring trend was confirmed. Together, these findings underscore that increased digitization consistently contributed to a reduced entropy in 2023 (see Figure 13).

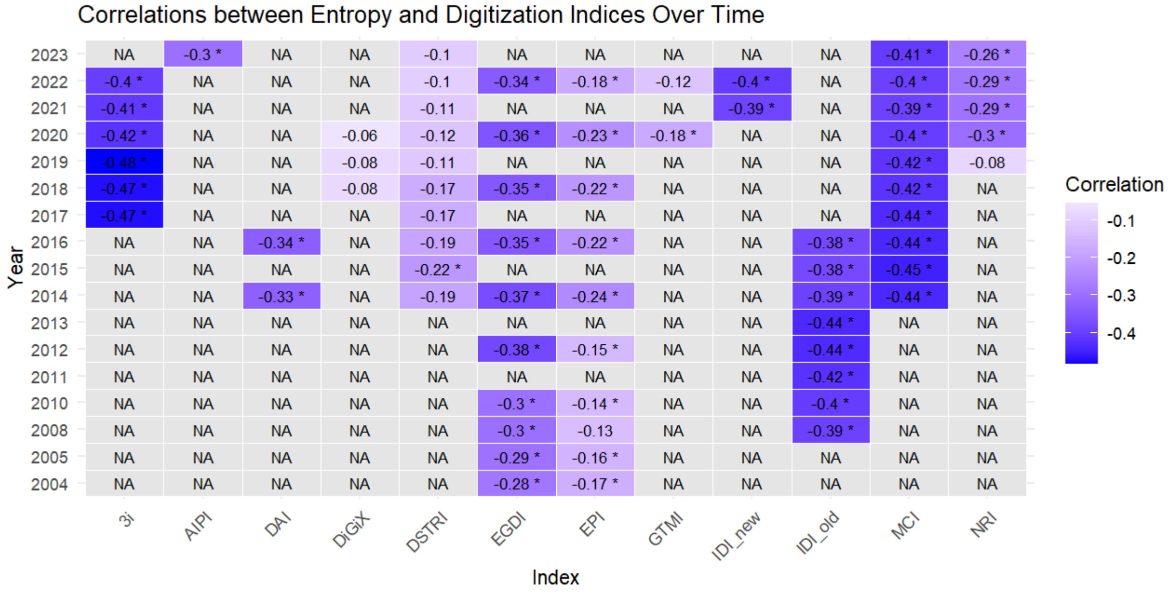


Figure 17 Heatmap displaying Pearson's correlation coefficients (2004-2023); asterisks (*) denote significance at the p < 0.05 level.

5.3.2 Correlation Analysis - Number of Languages

The correlation analysis between the number of languages spoken in each country (based on 2023 data) and digitization indices revealed positive associations during the early years. In 2004, the EGDI (*r* = 0.202, *p* < .01) and the EPI (*r* = 0.324, *p* < .001) both showed significant positive correlations with the language count, indicating that higher levels of digitization led to a larger number of languages spoken. A similar trend observed in 2005 reinforces the interpretation that early digital government initiatives may have contributed to an increase in the documented number of languages.

Variable1	**Variable2**	**Correlation**	**P_Value**	**Effect_Strength**	**Significance**	**Direction**
EGDI	EPI	0.772	p < .01	Large	*	Positive
EGDI	language_count	0.202	0.005	Small	*	Positive
EPI	language_count	0.324	p < .01	Medium	*	Positive

Table 6 Language count correlation analysis summary table for 2004 using pairwise complete observations.

Throughout 2008, 2010, and 2012, the EPI consistently demonstrated a significant positive effect on the number of languages spoken per country, whereas the positive impact of the EGDI was no longer evident by 2012. Meanwhile, the newly introduced IDI (ICT-Development Index) exhibited a weak, insignificant negative correlation with the 2023 language count. Overall, these findings suggest that while digitization during this period led to an increase in the number of languages spoken in a country in 2023, the strength of this effect diminished slightly over time.

During 2014, 2015, and 2016, with the introduction of new indices—the Mobile Connectivity Index (MCI), the Digital Services Trade Restrictiveness Index (DSTRI), and the Digital Adoption Index (DAI)—most correlations between digitization measures and language count became non-significant. The notable exception was the EPI, which continued to exhibit a significant positive correlation with language count throughout this period (2014: *r* = 0.217; 2015: *r* = 0.187). This pattern aligned with previous years' findings, where E-Participation consistently emerged as the most significant driver of increases in global language count as indicated in 2023.

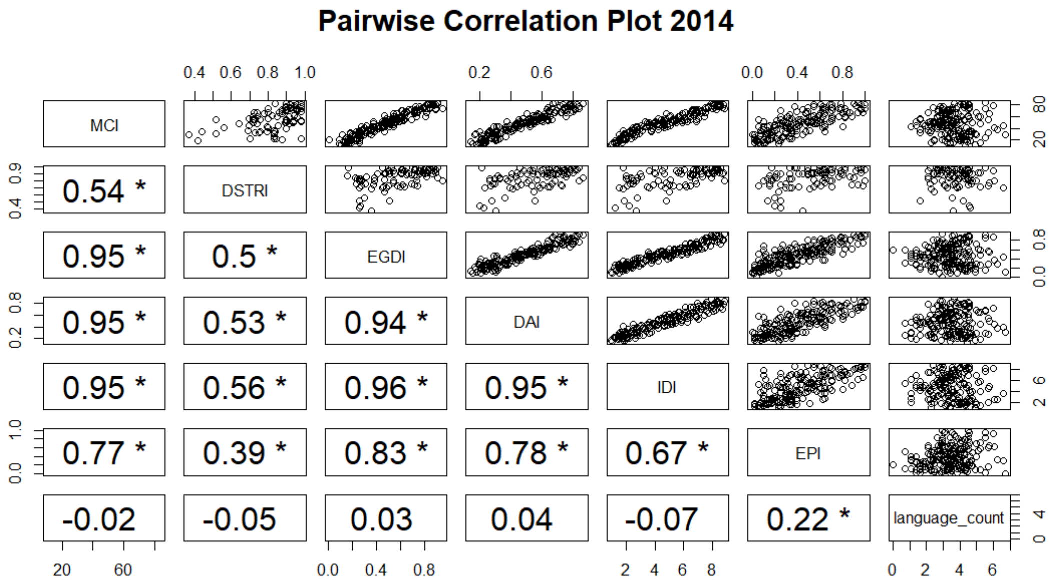
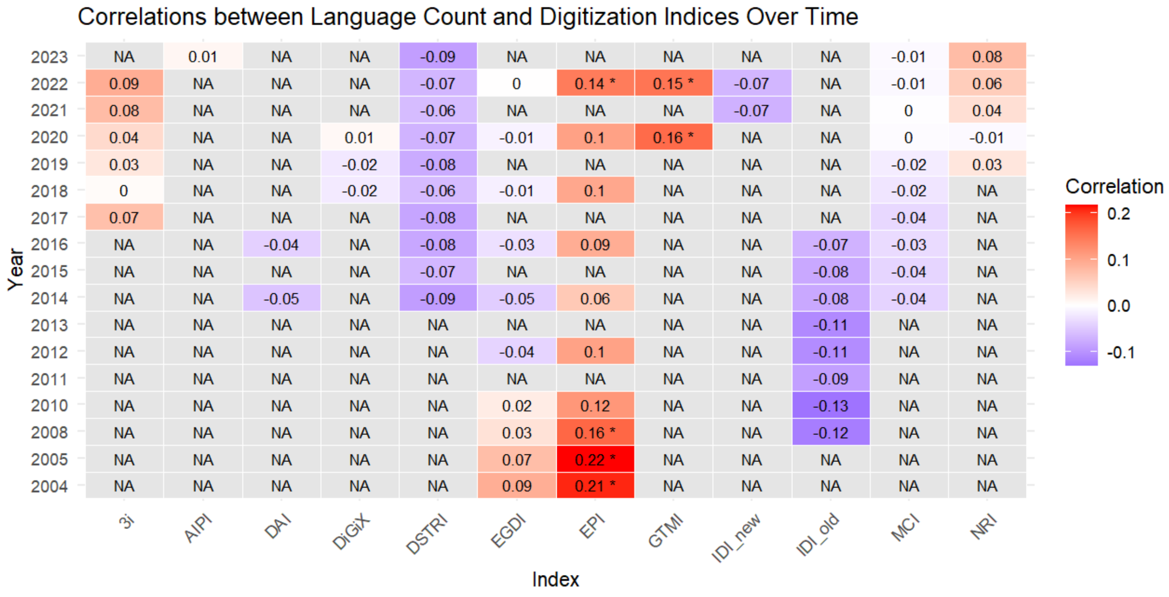


Figure 18 Language count pairwise correlation plot (2014) displaying Pearson's correlation coefficients; asterisks (*) denote significance at the p < 0.05 level.

In 2017, 2018, and 2019, the inclusion of an expanded set of digitization indices—including the newly introduced Network Readiness Index (NRI), Digitization Index (DiGiX), and Inclusive Internet Index (3i)—alongside the MCI, DSTRI, EGDI, IDI, and EPI did not alter this pattern. Once again, the EPI was the only index that exhibited a significant positive correlation with the number of languages spoken in each country in 2023. In contrast, the effects of all other digitization measures were weak ($r < 0.10$) and statistically insignificant ($p > 0.05$).

In 2020 and 2022, the newly added GovTech Maturity Index (GTMI) demonstrated a significant positive correlation with language count—stronger than that of the EPI (2020: GTMI $r = 0.357$ vs. EPI $r = 0.252$; 2022: GTMI $r = 0.325$ vs. EPI $r = 0.294$) when considering the pairwise complete observation matrix. In contrast, the other six indices did not exhibit any significant correlations with the number of languages spoken in a country during the period from 2020 to 2023, with 2021 and 2023 recording no significant correlations at all.

Overall, the analysis indicated that early digitization indices—specifically the EGDI and EPI—were significantly associated with the number of languages spoken per country as evidenced by the 2023 data. The EGDI's effect faded relatively quickly by 2012, whereas the EPI maintained a consistently significant positive correlation over time. Additionally, the GovTech Maturity Index (GTMI) emerged as another significant predictor in 2020 and 2022, exhibiting even stronger correlations than the EPI during those years. In contrast, the majority of the other indices—the IDI, the NRI, the DSTRI, the 3i, the MCI, the DiGiX, and the DAI—did not yield any significant results. These findings suggest that while government digitization, particularly through E-Participation, may have had a positive influence on language count in the long run, digitization alone does not appear to be the primary driver behind these developments.



Year	3i	AIPI	DAI	DiGiX	DSTRI	EGDI	EPI	GTMI	IDI_new	IDI_old	MCI	NRI
2023	NA	0.01	NA	NA	-0.09	NA	NA	NA	NA	NA	-0.01	0.08
2022	0.09	NA	NA	NA	-0.07	0	0.14 *	0.15 *	-0.07	NA	-0.01	0.06
2021	0.08	NA	NA	NA	-0.06	NA	NA	NA	-0.07	NA	0	0.04
2020	0.04	NA	NA	0.01	-0.07	-0.01	0.1	0.16 *	NA	NA	0	-0.01
2019	0.03	NA	NA	-0.02	-0.08	NA	NA	NA	NA	NA	-0.02	0.03
2018	0	NA	NA	-0.02	-0.06	-0.01	0.1	NA	NA	NA	-0.02	NA
2017	0.07	NA	NA	NA	-0.08	NA	NA	NA	NA	NA	-0.04	NA
2016	NA	NA	-0.04	NA	-0.08	-0.03	0.09	NA	NA	-0.07	-0.03	NA
2015	NA	NA	NA	NA	-0.07	NA	NA	NA	NA	-0.08	-0.04	NA
2014	NA	NA	-0.05	NA	-0.09	-0.05	0.06	NA	NA	-0.08	-0.04	NA
2013	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.11	NA	NA
2012	NA	NA	NA	NA	NA	-0.04	0.1	NA	NA	-0.11	NA	NA
2011	NA	NA	NA	NA	NA	NA	NA	NA	NA	-0.09	NA	NA
2010	NA	NA	NA	NA	NA	0.02	0.12	NA	NA	-0.13	NA	NA
2008	NA	NA	NA	NA	NA	0.03	0.16 *	NA	NA	-0.12	NA	NA
2005	NA	NA	NA	NA	NA	0.07	0.22 *	NA	NA	NA	NA	NA
2004	NA	NA	NA	NA	NA	0.09	0.21 *	NA	NA	NA	NA	NA

Figure 19 Heatmap displaying Pearson's correlation coefficients (2004-2023); asterisks (*) denote significance at the p < 0.05 level.

5.3.3 correlation Analysis - RLI

The correlation analysis between the Red List Index (RLI) and the digitization measures was conducted using the same approach as the previous analyses targeting the correlations with entropy, and language count. As with the other diversity measures, the RLI values were based on 2023 data.

In 2004 and 2005, the available indices—the EGDI (2004: r = -0.083; 2005: r = -0.084) and the EPI (2004: r = -0.185; 2005: r = -0.216) —showed negative correlations with the adapted Red List Index (RLI), although the correlation for the EGDI was not significant (p > 0.05). These findings suggest that higher levels of governmental digitization during those years have set in motion changes that contributed to higher levels of language endangerment observed in 2023.

In 2008, 2010, and 2012 the E-Participation Index (EPI) generally continued to show a negative correlation with the RLI, indicating that higher e-participation led to greater language endangerment in the long term. However, within this period, the strength of the correlation appeared to decrease, and by 2010, the relationship was no longer statistically significant. In contrast, the effect of the EGDI shifted to a small positive correlation with RLI between 2008 and 2012. Specifically, in 2008 and 2010, the full observation matrix showed a small positive—but insignificant—correlation between the EGDI and the RLI, whereas the pairwise observation matrix yielded a small negative, also insignificant, correlation. By 2012, both matrices indicated a small positive correlation, with the effect remaining insignificant throughout. Similarly, the ICT Development Index (IDI), introduced in 2008, also showed a small, non-significant positive effect on the RLI. Collectively, these findings suggest that while increased e-participation was robustly linked with higher long-term language endangerment, other aspects of digitization exhibited a less pronounced effect.

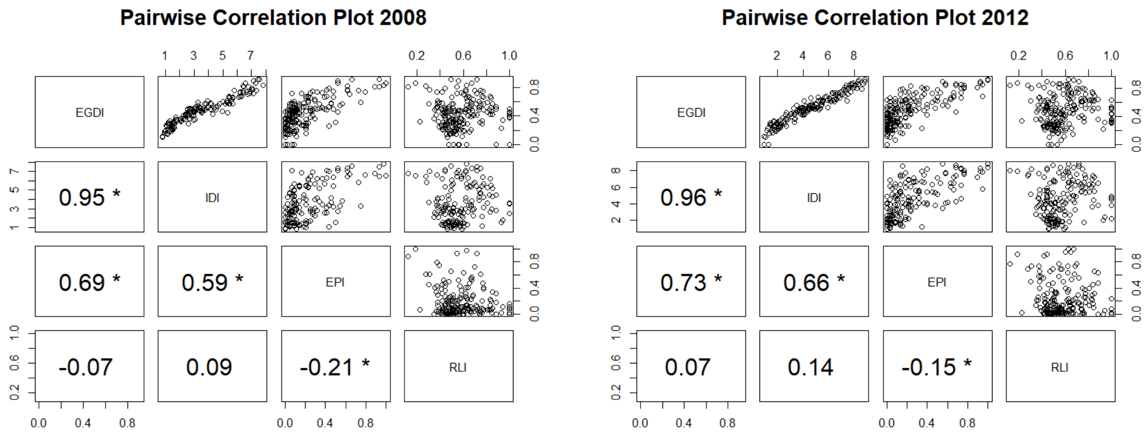


Figure 20 RLI pairwise Correlation Plot displaying Pearson's correlation coefficients for 2008 and 2012; asterisks (*) denote significance at the p < 0.05 level.

Despite the introduction of several new indices during these years—namely, the Mobile Connectivity Index (MCI), the Digital Services Trade Restrictiveness Index (DSTRI), and Digital Adoption Index (DAI)—none of the indices exhibited a significant Pearson correlation with the RLI. In 2014, the correlation of the EGDI with the RLI shifted back to a small negative value, while it remained positive in 2015 and 2016; however, all these correlations were statistically insignificant, with *p*-values above 0.05 (see Table 6). Additionally, the correlations for all other indices consistently showed a small positive effect on the RLI across both complete and pairwise correlation methods, without drastic changes in effect size. These findings suggest that digitization, as measured by the selected indices over this period, did not have a significant impact on language endangerment in the long run as evidenced by the data from 2023.

Year	Variable1	Variable2	Correlation	P_Value
2014	MCI	RLI	0.129	0.090
2014	DSTRI	RLI	0.031	0.777
2014	EGDI	RLI	0.050	0.486
2014	DAI	RLI	0.061	0.415
2014	IDI	RLI	0.104	0.171
2014	EPI	RLI	-0.107	0.140
2015	MCI	RLI	0.132	0.084
2015	DSTRI	RLI	0.016	0.884
2015	IDI	RLI	0.120	0.110
2016	MCI	RLI	0.135	0.0777
2016	DSTRI	RLI	0.007	0.951
2016	EGDI	RLI	0.057	0.433
2016	DAI	RLI	0.070	0.353
2016	IDI	RLI	0.118	0.118
2016	EPI	RLI	-0.049	0.503

Table 7 RLI Pearson’s correlation coefficients and p-values for 2014–2016 (based on pairwise complete observations).

The trend of statistically insignificant correlations persisted in the following years. In 2017, the DSTRI and the newly introduced Inclusive Internet Index (3i) showed weak negative correlations with the RLI, while the other indices exhibited weak positive correlations; however, all these associations were statistically insignificant. In 2018, the DSTRI continued its negative association with the RLI, accompanied by the EPI, which also showed a negative effect, whereas the 3i shifted to a slightly positive correlation—again, none of these effects reached significance. In 2019, a significant positive correlation was observed between the MCI and the RLI ($r = 0.154$). Although the DSTRI’s correlation shifted to a positive direction, it remained statistically insignificant, and the remaining indices continued to exhibit negative, non-significant correlations with the RLI. These results suggest that, while the MCI might have led to improved language vitality modestly, the overall influence of digitalization—as measured by the other indices—on language endangerment was minimal.

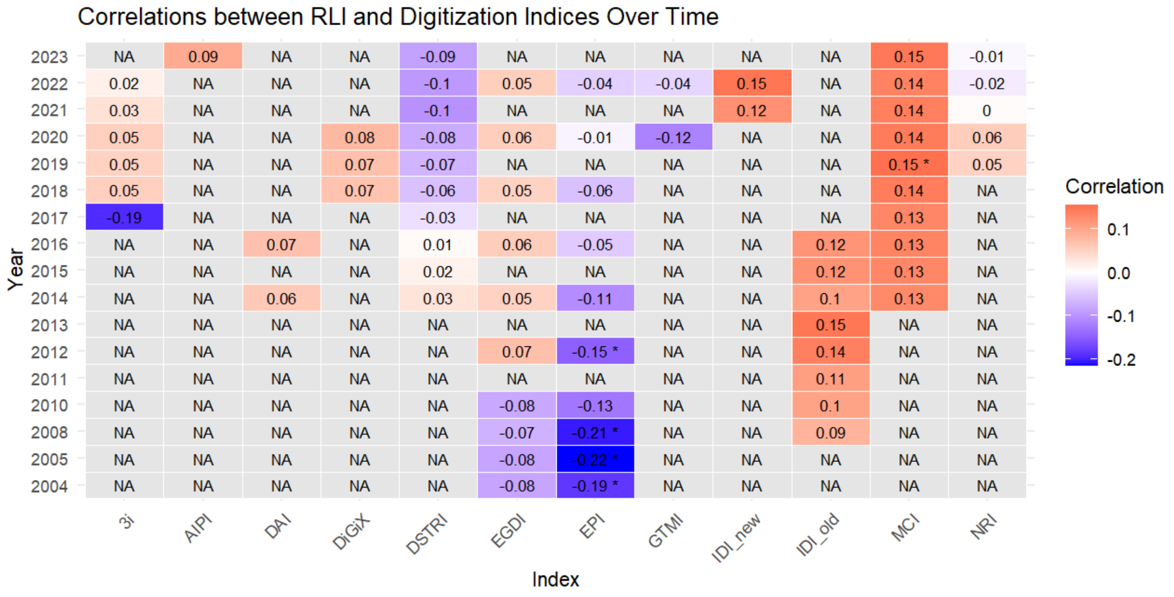
From 2020 to 2023, the correlation between the MCI and the RLI was not significant, and this non-significance applied to all other indices included in the analysis. Even in 2020 and 2022 (see Figure 21)—when data for eight digitization indices were available—none of the relationships reached statistical significance. The introduction of the AI-Preparedness Index (AIPI) in 2023 did not alter this pattern. The DSTRI continued to exhibit a negative association with the RLI throughout all years, though the effect remained weak and non-significant. Minor fluctuations in effect size and direction were observed for other indices: for example, the MCI shifted to a negative correlation in 2022 after showing positive values in prior years, while the 3i moved from a negative association in 2020 to a positive one in both 2021 and 2022. Overall, these findings reinforce the broader pattern that digitization indices were not reliably or meaningfully associated with language endangerment as measured by the RLI in 2023.

The correlation analysis between the RLI and digitization indices from 2004 to 2023 revealed a consistent overarching pattern: across most years and measures, correlations remained weak and statistically insignificant. This suggests that digitization, as captured by various global indices, did not show a strong or consistent association with language endangerment trends in the long term as reflected in the adapted Red List Index.

Despite this general trend, some recurring patterns and distinct results emerged. The E-Participation Index (EPI) showed statistically significant negative correlations with the RLI in 2004, 2005, 2008, and 2012, indicating that, in those years, higher levels of e-participation resulted in greater language endangerment as evidenced by the data from 2023. After 2012, however, this relationship weakened and became statistically insignificant, with smaller effect sizes and more fluctuation in direction.

Most other indices, such as the EGDI, the IDI, the DAI, and the 3i, tended to show small positive correlations with the RLI, particularly from 2010 onward. These effects remained consistently statistically insignificant and showed no stable temporal trend. Even in years with expanded index coverage—such as 2020 and 2022—no significant associations were observed, and effect sizes remained modest. The DSTRI, although it never produced a statistically significant result, repeatedly showed a negative correlation with the RLI across years, suggesting a directionally stable but weak and non-robust relationship.

A distinct exception occurred in 2019, when the Mobile Connectivity Index (MCI) showed a statistically significant positive correlation with the RLI. This suggests that, higher mobile connectivity in that year may have contributed to lower levels of language endangerment as indicated by the RLI values from 2023. However, this effect did not replicate in adjacent years, and the MCI reverted to a non-significant or slightly negative effect, reinforcing the isolated nature of the 2019 result.



Year	3i	AIPI	DAI	DiGiX	DSTRI	EGDI	EPI	GTMI	IDI_new	IDI_old	MCI	NRI
2023	NA	0.09	NA	NA	-0.09	NA	NA	NA	NA	NA	0.15	-0.01
2022	0.02	NA	NA	NA	-0.1	0.05	-0.04	-0.04	0.15	NA	0.14	-0.02
2021	0.03	NA	NA	NA	-0.1	NA	NA	NA	0.12	NA	0.14	0
2020	0.05	NA	NA	0.08	-0.08	0.06	-0.01	-0.12	NA	NA	0.14	0.06
2019	0.05	NA	NA	0.07	-0.07	NA	NA	NA	NA	NA	0.15 *	0.05
2018	0.05	NA	NA	0.07	-0.06	0.05	-0.06	NA	NA	NA	0.14	NA
2017	-0.19	NA	NA	NA	-0.03	NA	NA	NA	NA	NA	0.13	NA
2016	NA	NA	0.07	NA	0.01	0.06	-0.05	NA	NA	0.12	0.13	NA
2015	NA	NA	NA	NA	0.02	NA	NA	NA	NA	0.12	0.13	NA
2014	NA	NA	0.06	NA	0.03	0.05	-0.11	NA	NA	0.1	0.13	NA
2013	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.15	NA	NA
2012	NA	NA	NA	NA	NA	0.07	-0.15 *	NA	NA	0.14	NA	NA
2011	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.11	NA	NA
2010	NA	NA	NA	NA	NA	-0.08	-0.13	NA	NA	0.1	NA	NA
2008	NA	NA	NA	NA	NA	-0.07	-0.21 *	NA	NA	0.09	NA	NA
2005	NA	NA	NA	NA	NA	-0.08	-0.22 *	NA	NA	NA	NA	NA
2004	NA	NA	NA	NA	NA	-0.08	-0.19 *	NA	NA	NA	NA	NA

Figure 21 Heatmap displaying Pearson's correlation coefficients (2004-2023); asterisks (*) denote significance at the p < 0.05 level.

Taken together, the findings suggest that while a few indices—particularly the EPI—showed temporary and statistically significant associations with the RLI, the majority of correlations across the two-decade span were weak, inconsistent, and non-significant. Overall, digitization, as measured through the selected global indices, did not demonstrate a stable or meaningful influence on language endangerment over time.

5.3.4 Summary of Correlation Results

The correlation analyses using the Red List Index (RLI), which measures language endangerment; the number of languages spoken in each country, reflecting language richness; and entropy, indicating the distributional evenness of language use, provided valuable insights into the relationship between digitization and language distribution, with each measure capturing different aspects of linguistic patterns. Moreover, the relationships between digitization and these measures evolved differently over time.

The analysis of the number of languages spoken per country—using the 2023 static language count as a reference—yielded significant long-term correlations with digitization indices. In particular, the EPI showed consistent significant correlations across all years until 2012, with the EGDI demonstrating significant effects in the early years, and the GTMI emerging as significant in 2020 and 2022. However, most other indices did not exhibit any significant relationship with language count. Overall, these results suggest that the long-term impact of digitization on language diversity, as reflected in the 2023 data, was minimal.

Similarly, the RLI analysis showed that correlations with digitization indices were generally weak, statistically insignificant, and temporally inconsistent. While some early years (e.g., 2004, 2005, 2008, 2012) showed significant negative correlations between the E-Participation Index and the RLI—implying higher digital engagement coincided with greater language endangerment—these effects were isolated and not sustained. Across the broader timeline, digitization did not demonstrate a stable or meaningful relationship with language endangerment.

By contrast, the entropy analysis revealed a clear and consistent negative association with digitization across most years. As digital development advanced, entropy declined, indicating that language use became more concentrated and less evenly distributed. This pattern remained robust even as the influence of language count weakened over time, and held across a wide range of indices—including newer additions like the AI Preparedness Index in 2023, which introduced some nuanced, index-specific shifts.

In summary, while digitization showed little consistent link to language endangerment (RLI), and the number of languages spoken in each country, it was consistently associated with lower entropy values —indicating increasing linguistic inequality. This highlights the need of distinguishing between different dimensions of linguistic diversity when assessing the cultural impacts of digital transformation.

6 Conclusion

This final chapter summarizes the key findings of the thesis and discusses their implications in the context of existing research. It reflects on how the results contribute to current debates around digitization and linguistic diversity, particularly regarding how both are measured and understood.

The aim of this thesis was to explore the relationship between digitization and linguistic diversity. To investigate this, a series of analytical steps were undertaken, employing various digitization indices and linguistic diversity measures—including the logarithm of the number of languages per country, entropy-based metrics, and an adapted Red List Index (RLI).

The analysis began with a hierarchical cluster analysis to identify patterns and relationships among the digitization indices. Where possible, these indices were further examined in terms of their internal structure through correlation analysis of their sub-components (indicators and sub-indices). Subsequently, the relationship between digitization and linguistic diversity was assessed using correlation analysis, focusing on how digitization indices relate to the selected linguistic diversity measures.

Taken together, this multi-method approach yielded valuable insights into the complex interplay between digital development and the preservation of linguistic diversity across countries.

The cluster analysis highlighted structural differences among the digitization indices, revealing that digitization is not captured uniformly across measures. While most indices showed strong internal similarity, a few—most notably the Digital Service Trade Restrictiveness Index (DSTRI)—consistently diverged, reflecting narrower conceptual scopes. These patterns indicated that digitization is best understood as a multidimensional construct, with different indices emphasizing distinct aspects such as infrastructure, access, or regulation. Although not directly linked to linguistic diversity, this analysis provides an essential foundation for interpreting the indices' differing relationships to language-related outcomes.

Conclusion

The breadth analysis revealed variation in the internal coherence of digitization indices, reflecting differences in conceptual scope. Some indices, such as the AI Preparedness Index (AIPI) and early versions of the E-Government Development Index (EGDI), displayed strong internal consistency, while others—like the DSTRI and the updated ICT Development Index (IDI)—showed more heterogeneous structures. These differences underscore that digitization indices operate with distinct measurement logics—some emphasizing tightly integrated constructs, others aiming for a wider conceptual reach. Understanding these differences is important for interpreting what each index represents and for making informed choices when selecting them for comparative or policy-oriented analyses.

The correlation analysis revealed that the relationship between digitization and linguistic diversity depends strongly on how diversity is measured. The number of languages per country and the Red List Index (RLI) provided little evidence of a stable association with digitization over time—echoing past research which suggests that language loss is driven by a complex mix of ecological, historical, and social factors (Bromham et al., 2021; Bouckaert et al., 2022). However, the entropy-based measure consistently showed a negative relationship. This suggests that digital development coincides with a gradual concentration of language use, reducing the evenness with which languages are spoken within countries. The persistence of this pattern across multiple indices and years points to a structural dynamic in which increased digitization may favour dominant languages. This observation is consistent with arguments in previous research that digital technologies often reinforce the dominance of globally powerful languages, especially in online content and digital communication (Blasi et al., 2022; Cunliffe, 2007). These findings underscore the importance of distinguishing between richness, vitality, and distribution when analysing linguistic diversity, particularly in the context of digital transformation.

Across methods and measures, this thesis does not reveal a single, unified relationship between digitization and linguistic diversity, but rather a layered and often ambiguous one—depending on how both are defined and measured. The findings challenge any assumption that digitization has a straightforward or uniform impact on linguistic diversity, especially given that the concept itself varies depending on whether it is measured by richness, vitality, or distribution. Still, a common tension becomes clear. As countries become more digitally developed, language use tends to become more concentrated. This does not necessarily mean languages are disappearing, but rather that usage becomes more uneven, often centred around dominant languages.

Using multiple methods helped make these patterns more visible. Looking at both the structure of digitization indices and their different links to language data shows how much results depend on the way things are measured. This underlines the need to avoid overly simple conclusions about how technology and language are connected.

7 Limitations and Outlook

In this final chapter main limitations of the study are outlined and avenues for future research are highlighted.

Several limitations should be taken into account when interpreting the findings of this thesis, many of which are common in large-scale, cross-national research. First, data availability was uneven across years and indices. In some instances, digitization measures changed in composition over time, which limited comparability and may have introduced inconsistencies affecting the interpretation of longer-term trends. Second, the analysis was based on correlations, which are suitable for identifying general patterns but do not support causal claims. It therefore remains open to which extend digitization directly influences linguistic diversity or whether observed associations reflect broader contextual factors. Third, all linguistic diversity measures used in the analysis were static and reflect relatively long-term conditions. This limited the possibility of capturing short-term or rapidly emerging changes in language use. Developments occurring during the study period may therefore not be fully reflected in the data. Lastly, while the language datasets employed are well-established and widely used, they may not always account for informal, undocumented, or recently evolving linguistic practices—particularly in regions where such data is more difficult to collect or standardize.

Despite these limitations, the findings of this thesis offer a meaningful contribution to ongoing research on the cultural effects of digital transformation. By demonstrating that the relationship between digitization and linguistic diversity depends on how both are conceptualized and measured, this study helps to establish a more differentiated framework for future studies. The recurring association between digital development and linguistic concentration across multiple indices points to a structural pattern that deserves further attention. These insights provide a strong foundation for future empirical work and are equally relevant for policy makers, language planning, and the development of platforms that aim to support more inclusive and linguistically diverse digital environments.

8 References

Archibugi, D., Denni, M., & Filippetti, A. (2009). The technological capabilities of nations: The state of the art of synthetic indicators. *Technological Forecasting and Social Change*, *76*(7), 917–931. https://doi.org/10.1016/j.techfore.2009.01.002

Alizai, K. (2021). Language Endangerment and the Need for Technology for Language Empowerment: Case Study of the Brahui language in Balochistan. *International Journal of Pedagogy, Innovation and New Technologies*, *8*(1), 37–50. https://doi.org/10.5604/01.3001.0014.9140

Baumann, A., Lenzner, B., Fellner, H. A., & Essl, F. (2024). The relation between European colonialism and linguistic diversity [Poster presentation]. *EVOLANG XV, Madison, WI*.

Billon, M. (2010). Differences in digitalization levels: A multivariate analysis studying the global digital divide. *Review of world economics*, *146*(1), 39–73.

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement*, 75(5), 785–804. https://doi.org/10.1177/0013164414557639

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic Inequalities in Language Technology Performance across the World's Languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5486–5505. https://doi.org/10.18653/v1/2022.acl-long.376

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., & Atkinson, Q. (2022). *Global language diversification is linked to socio-ecology and threat status*. https://doi.org/10.31235/osf.io/f8tr6

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, *42*(5), 2785–2797. https://doi.org/10.1016/j.eswa.2014.09.054

References

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). *clValid: An r package for cluster validation*. *Journal of Statistical Software, 25*(4), 1–22. https://www.jstatsoft.org/v25/i04/

Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., & Hua, X. (2021). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, *6*(2), 163–173. https://doi.org/10.1038/s41559-021-01604-y

Butchart, S. H. M., Stattersfield, A. J., Bennun, L. A., Shutes, S. M., Akçakaya, H. R., Baillie, J. E. M., Stuart, S. N., Hilton-Taylor, C., & Mace, G. M. (2004). Measuring global trends in the status of biodiversity: Red List Indices for birds. *PLoS Biology, 2*(12), e383. https://doi.org/10.1371/journal.pbio.0020383

Cámara, N. (2024). *DiGiX 2024 update: A multidimensional index of digitization.* BBVA Research. https://www.bbvaresearch.com/wpcontent/uploads/2024/08/DiGiX_2024_Update_A_Multidimensional_Index_of_Digitization_edi-1.pdf

Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E., & Tavares, M. M. (2024). *Gen-AI: Artificial intelligence and the future of work.* IMF Staff Discussion Note SDN2024/001. International Monetary Fund. https://doi.org/10.5089/9798400262548.006

Charfeddine, L., & Umlai, M. (2023). ICT sector, digitization and environmental sustainability: A systematic review of the literature from 2000 to 2022. *Renewable and Sustainable Energy Reviews, 184,* 113482. https://doi.org/10.1016/j.rser.2023.113482

Christopher, A. N. (2017). *Interpreting and Using Statistics in Psychological Research*. SAGE Publications, Inc. https://doi.org/10.4135/9781506304144

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cunliffe, D. (2007). Chapter 8. Minority Languages and the Internet: New Threats, New Opportunities. In M. Cormack & N. Hourigan (Eds.), *Minority Language Media* (pp.

133–150). Multilingual Matters. https://doi.org/10.21832/9781853599651-008

Economist Impact. (2022). *The Inclusive Internet Index 2022: Methodology report.* Economist Impact. https://impact.economist.com/projects/inclusive-internet-index/downloads/3i-methodology.pdf

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). Wiley.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., & Lu, Y. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry, 26*(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Ferencz, J. (2019). *The OECD Digital Services Trade Restrictiveness Index. OECD Trade Policy Papers* (No. 221). OECD Publishing. https://doi.org/10.1787/16ed2d78-en

Galili, T. (2015). *dendextend: An r package for visualizing, adjusting, and comparing trees of hierarchical clustering*. *Bioinformatics, 31*(22), 3718–3720. https://doi.org/10.1093/bioinformatics/btv428

Gil-Garcia, R. J., Badia-Contelles, J. M., & Pons-Porrata, A. (2006). A General Framework for Agglomerative Hierarchical Clustering Algorithms. *18th International Conference on Pattern Recognition (ICPR'06)*, 569–572. https://doi.org/10.1109/ICPR.2006.69

Gavin, M. C., Botero, C. A., Bowern, C., Colwell, R. K., Dunn, M., Dunn, R. R., Gray, R. D., Kirby, K. R., McCarter, J., Powell, A., Rangel, T. F., Stepp, J. R., Trautwein, M., Verdolin, J. L., & Yanega, G. (2013). Toward a Mechanistic Understanding of Linguistic Diversity. *BioScience*, *63*(7), 524–535. https://doi.org/10.1525/bio.2013.63.7.6

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Grin, F., & Fürst, G. (2022). Measuring Linguistic Diversity: A Multi-level Metric*. Social*

Indicators Research, 164(2), 601–621. https://doi.org/10.1007/s11205-022-02934-5

GSMA Intelligence. (2024). *Mobile Connectivity Index methodology report 2024*. GSMA Intelligence. https://www.gsma.com/r/wp-content/uploads/2024/10/The-State-of-Mobile-Internet-Connectivity-Report-2024.pdf

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, 23(2), 283–294. https://doi.org/10.1017/S1366728919000026

Harmon, D., & Loh, J. (2010). The Index of Linguistic Diversity: A New Quantitative Measure of Trends in the Status of the World's Languages. *Language documentation and conservation*, *4*, 97–151.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications, 103***,** 103–118**.** https://doi.org/10.1016/0024-3795(88)90223-6

Hutson, J., Ellsworth, P., & Ellsworth, M. (2024). Preserving Linguistic Diversity in the Digital Age: A Scalable Model for Cultural Heritage Continuity. *Journal of Contemporary Language Research*, *3*(1), 10–19. https://doi.org/10.58803/jclr.v3i1.96

International Telecommunication Union, Development Sector. (2024). *ICT Development Index 2024: Measuring digital development.* ITU Publications. https://www.itu.int/hub/publication/d-ind-ict_mdd-2024-3/

Jarman, A. M. (2020). *Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method*. https://doi.org/10.13140/RG.2.2.11388.90240

Joshua Project. (2023). *All Peoples* [Data table]. Retrieved from https://public.tableau.com/app/profile/joshuaproject/viz/jp-global-interactive/AllPeoples

Jovanović, M., Dlačić, J., & Okanović, M. (2018). Digitalization and society's sustainable development: Measures and implications. *Zbornik Radova Ekonomskog Fakulteta u Rijeci: Časopis za Ekonomsku Teoriju i Praksu/Proceedings of Rijeka Faculty of*

Economics: Journal of Economics and Business, 36(2), 905–928. https://doi.org/10.18045/zbefri.2018.2.905

Kandler, A., & Unger, R. (2023). Modeling Language Shift. In A. Bunde, J. Caro, C. Chmelik, J. Kärger, & G. Vogl (Eds.), *Diffusive Spreading in Nature, Technology and Society* (pp. 365–387). Springer International Publishing. https://doi.org/10.1007/978-3-031-05946-9_18

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and visualize the results of multivariate data analyses* (Version 1.0.7) [R package]. CRAN. https://CRAN.R-project.org/package=factoextra

Kassambara, A. (2023). *ggcorrplot: Visualization of a correlation matrix using ggplot2* (Version 0.1.4.1) [R package]. CRAN. https://CRAN.R-project.org/package=ggcorrplot

Katz, R. L., & Koutroumpis, P. (2013). Measuring digitization: A growth and welfare multiplier. *Technovation*, *33*(10–11), 314–319. https://doi.org/10.1016/j.technovation.2013.06.004

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* Wiley.

Loh, J., & Harmon, D. (2014). *Biocultural diversity: Threatened species, endangered languages.* WWF Netherlands.

Maggino, F., & Zumbo, B. D. (2012). Measuring the Quality of Life and the Construction of Social Indicators. In K. C. Land, A. C. Michalos, & M. J. Sirgy (Eds.), *Handbook of Social Indicators and Quality of Life Research* (pp. 201–238). Springer Netherlands. https://doi.org/10.1007/978-94-007-2421-1_10

Makowski, D., Ben-Shachar, M., Patil, I., & Lüdecke, D. (2020). Methods and Algorithms for Correlation Analysis in R. *Journal of Open Source Software*, *5*(51), 2306.

https://doi.org/10.21105/joss.02306

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181–204. https://doi.org/10.1007/BF01897163

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53

Mikami, Y., Kodama, S. (2010). Measuring linguistic diversity on the web. *Net.Lang: Towards the multilingual cyberspace,* 118–139.

Moore, R. (2016). Discourses of endangerment from mother tongues to machine readability. In O. García, N. Flores, & M. Spotti (Eds.), *Oxford handbook of language and society* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212896.013.16

Nee, J., Smith, G. M., Sheares, A., & Rustagi, I. (2022). Linguistic justice as a framework for designing, developing, and managing natural language processing tools. *Big Data & Society*, *9*(1), 20539517221090930. https://doi.org/10.1177/20539517221090930

Olaare, S. (2024). The Role of Technology in Language Preservation. *European Journal of Linguistics*, 3(2), 44–56. https://doi.org/10.47941/ejl.2046

Pagel, M. (2017). Darwinian perspectives on the evolution of human languages. *Psychonomic Bulletin & Review*, 24(1), 151–157. https://doi.org/10.3758/s13423-016-1072-z

Portulans Institute. (2023). *Network Readiness Index 2023: Trust in a network society – A crisis of the digital age?* Portulans Institute. https://download.networkreadinessindex.org/reports/data/2023/nri-2023.pdf

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

References

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. r Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Simons, G. F. (2019). Two centuries of spreading language loss. *Proceedings of the Linguistic Society of America*, *4*(1), 27. https://doi.org/10.3765/plsa.v4i1.4532

Simons, G. F., & Fennig, C. D. (Eds.). (2023). *Ethnologue: Languages of the World* (26th ed.). Dallas, TX: SIL International.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Tuomisto, H. (2010). A diversity of beta diversities: Straightening up a concept gone awry. Part

1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography,* 33(1), 2–22. https://doi.org/10.1111/j.1600-0587.2009.05880.x

Tüver, L., Zeh, K. (2024). *Linguistic Diversity in the Digital Age: Exploring the Effect of Digital Literacy on Minority Languages* [Unpublished seminar paper]. Faculty of Philological and Cultural Studies, University of Vienna.

United Nations Department of Economic and Social Affairs. (2024). *E-Government Survey 2024: Technical Appendix*. United Nations. https://desapublications.un.org/sites/default/files/publications/2024-10/Technical%20Appendix%20%28Web%20version%29%2030102024.pdf

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Vassilakopoulou, P., & Hustad, E. (2023). Bridging Digital Divides: A Literature Review and Research Agenda for Information Systems Research. *Information Systems Frontiers*, *25*(3), 955–969. https://doi.org/10.1007/s10796-020-10096-3

Voss, J. (2009). Encoding changing country codes for the Semantic Web with ISO 3166 and SKOS. In M.-A. Sicilia & M. D. Lytras (Eds.), *Metadata and Semantics* (pp. 211–221). Springer US. https://doi.org/10.1007/978-0-387-77745-0_20

References

Wei, T., & Simko, V. (2024). *corrplot: Visualization of a correlation matrix* (Version 0.95) [R package]. CRAN. https://cran.r-project.org/web/packages/corrplot/index.html

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (Version 1.1.4) [R package]. CRAN. https://CRAN.R-project.org/package=dplyr

World Bank. (2024). *Health Nutrition and Population Statistics* [Data set]. World Bank. https://datacatalog.worldbank.org/search/dataset/0037652

World Bank. (2016). *World Development Report 2016: Digital dividends.* World Bank. https://doi.org/10.1596/978-1-4648-0671-1

World Bank. (2022). WBG *GovTech Maturity Index: 2022 Update – Trends in Public Sector Digital Transformation. Equitable Growth, Finance and Institutions Insight.* World Bank. http://documents.worldbank.org/curated/en/099035001132365997

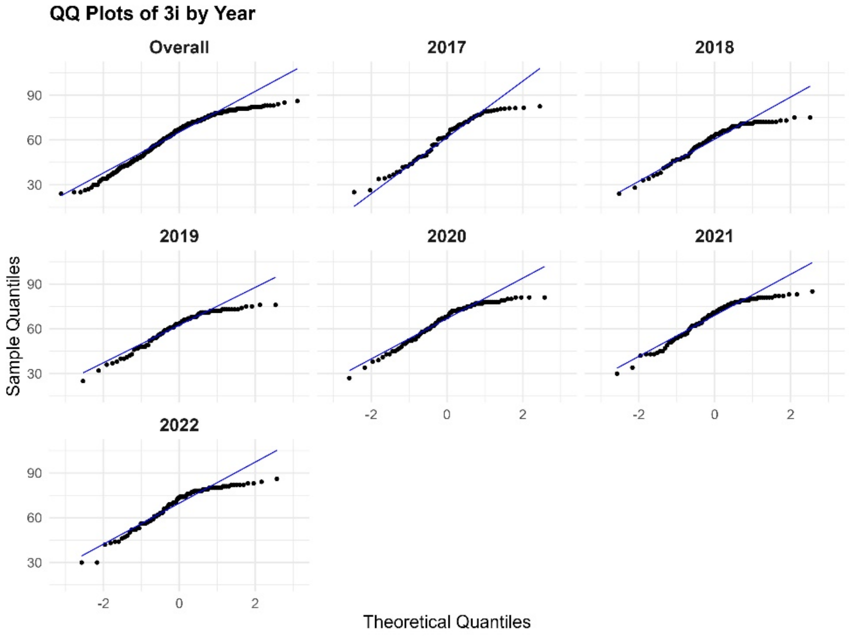
A Appendix

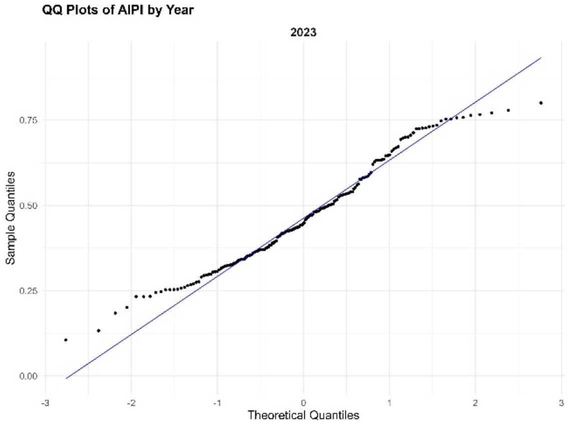
A.1 Code Availability

The full set of *r* scripts and datasets used in this thesis can be found in the following GitHub repository: https://github.com/KatharinaTheresia/Linguistic-Diversity-Thesis. This repository covers exploratory data analysis, clustering, and correlation analysis, ensuring full reproducibility of the results presented.

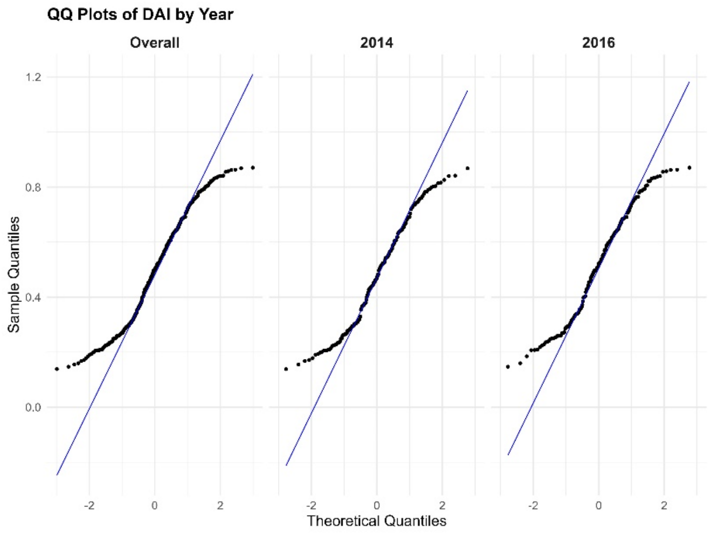
A.2 Plots and Tables

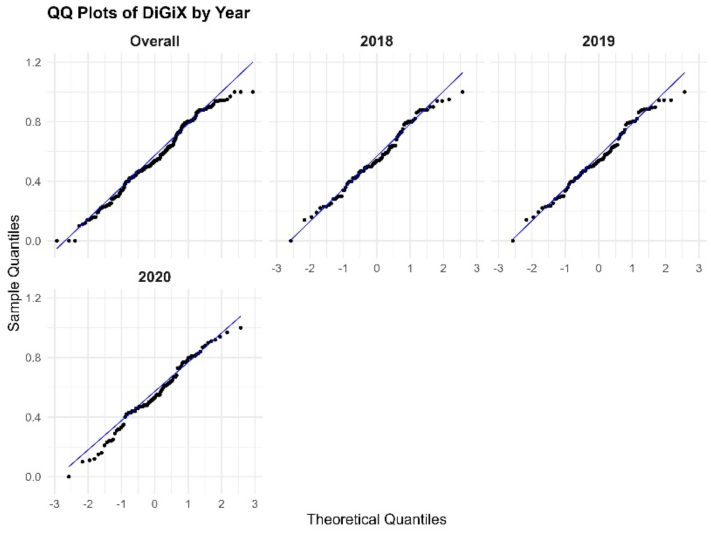
A.2.1 QQ-Plots with reference line for each index

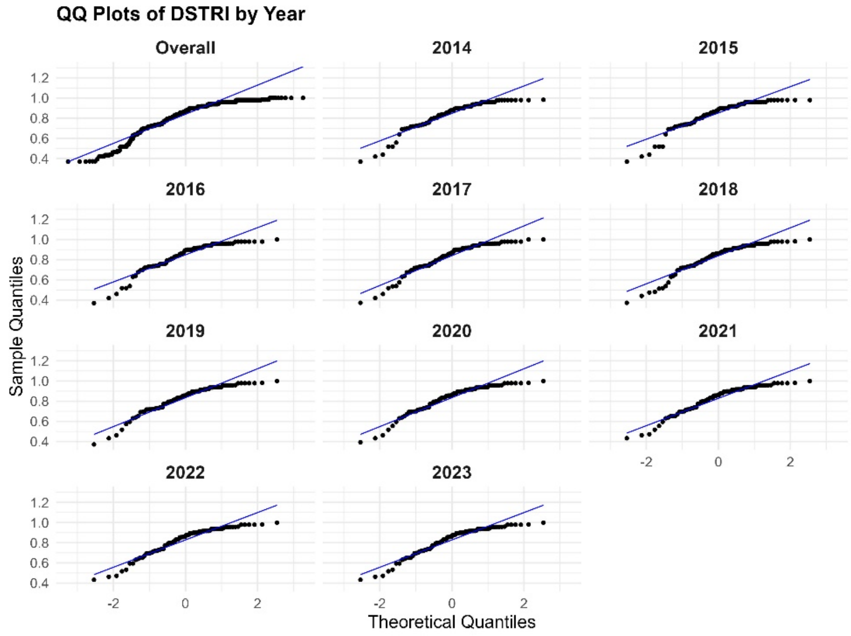




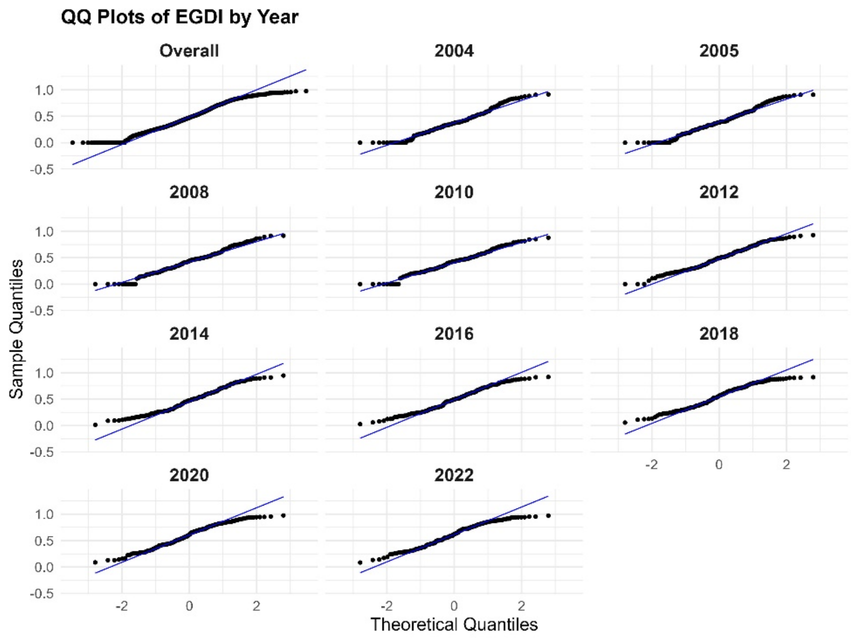
Appendix

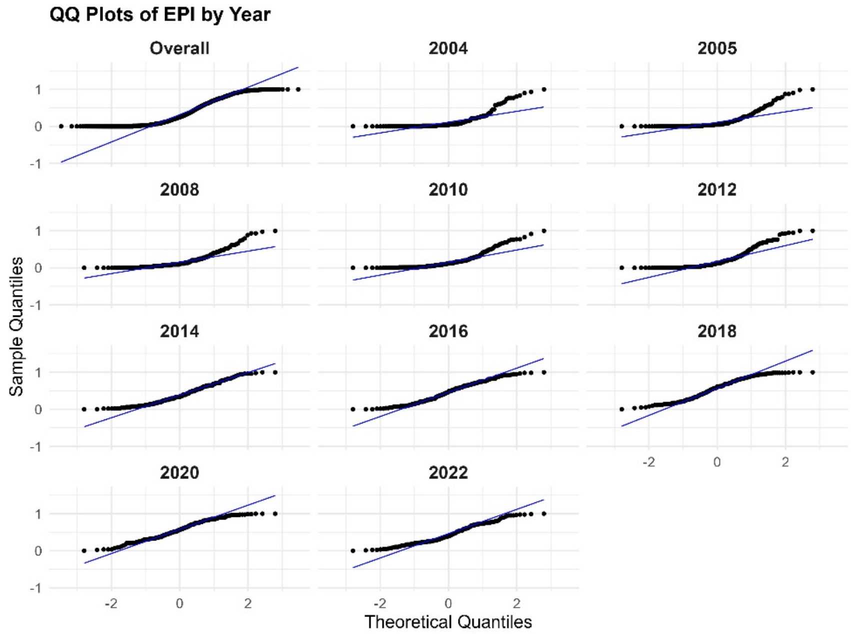

QQ Plots of DAI by Year
Overall
2014
2016
Sample Quantiles
Theoretical Quantiles
0.0
0.4
0.8
1.2
-2
0
2

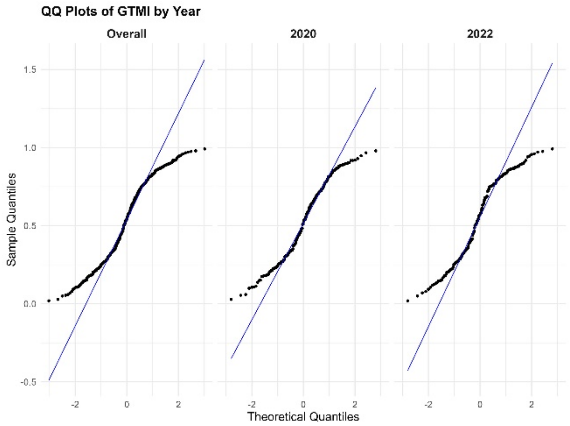

QQ Plots of DiGiX by Year
Overall
2018
2019
2020
Sample Quantiles
Theoretical Quantiles
0.0
0.4
0.8
1.2
-3
-2
-1
0
1
2
3


QQ Plots of DSTRI by Year
Overall
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Sample Quantiles
Theoretical Quantiles
0.4
0.6
0.8
1.0
1.2
-2
0
2

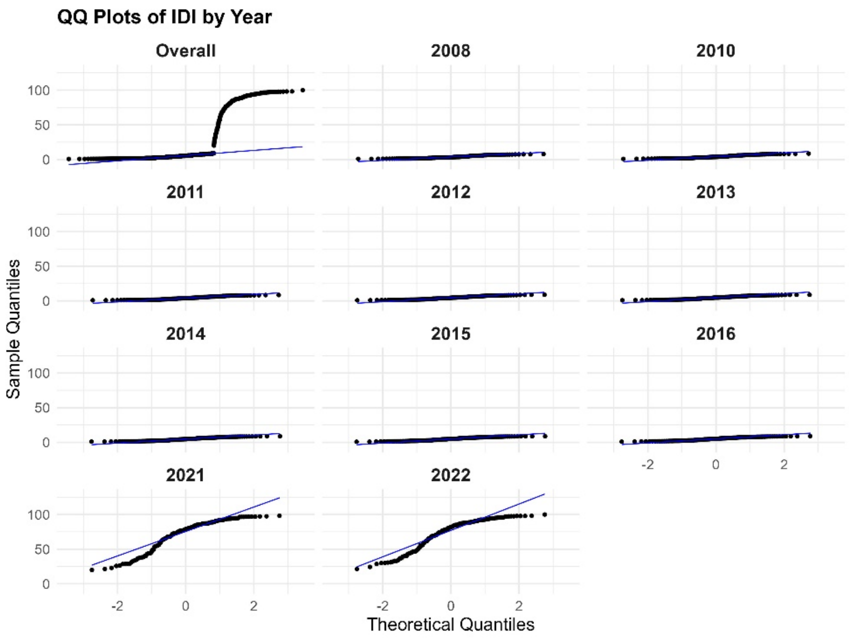
Appendix

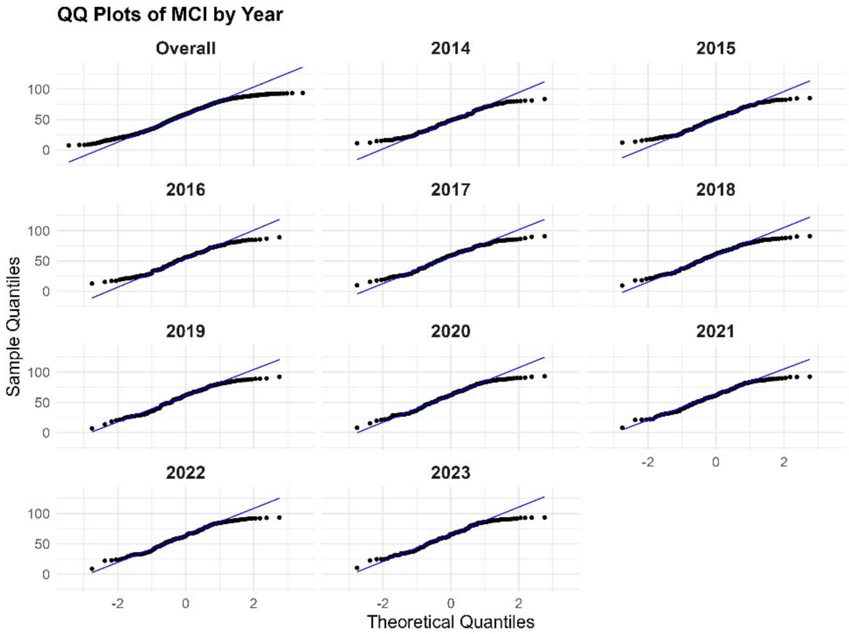

QQ Plots of EGDI by Year
Overall
2004
2005
2008
2010
2012
2014
2016
2018
2020
2022
1.0
0.5
0.0
-0.5
-2
0
2
Sample Quantiles
Theoretical Quantiles

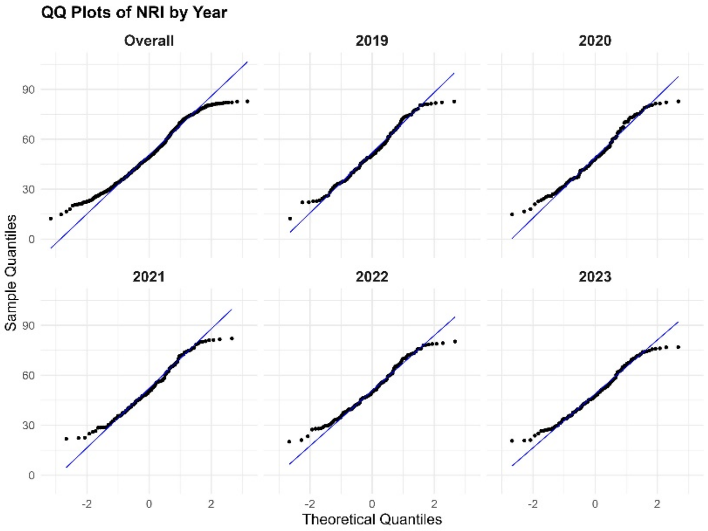

QQ Plots of EPI by Year
Overall
2004
2005
2008
2010
2012
2014
2016
2018
2020
2022
1
0
-1
-2
0
2
Sample Quantiles
Theoretical Quantiles


QQ Plots of GTMI by Year
Overall
2020
2022
1.5
1.0
0.5
0.0
-0.5
-2
0
2
Sample Quantiles
Theoretical Quantiles

Appendix

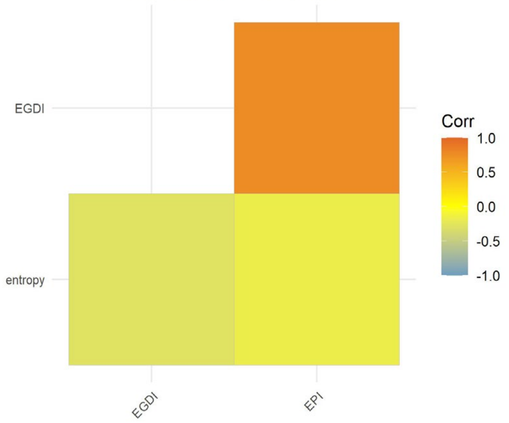

QQ Plots of IDI by Year
Overall
2008
2010
2011
2012
2013
2014
2015
2016
2021
2022
100
50
0
-2
0
2
Sample Quantiles
Theoretical Quantiles


QQ Plots of MCI by Year
Overall
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
100
50
0
-2
0
2
Sample Quantiles
Theoretical Quantiles

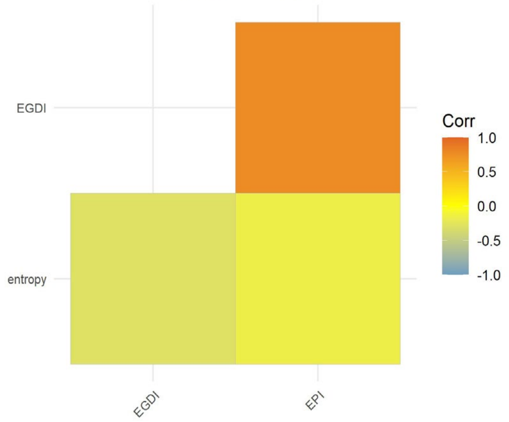

QQ Plots of NRI by Year
Overall
2019
2020
2021
2022
2023
90
60
30
0
-2
0
2
Sample Quantiles
Theoretical Quantiles

A.2.2 Correlation Analysis – Entropy

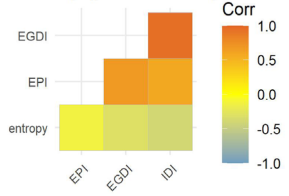
Heatmaps



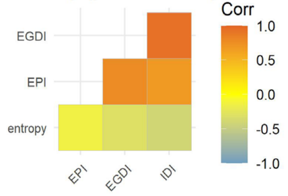
Correlation Heatmap 2004



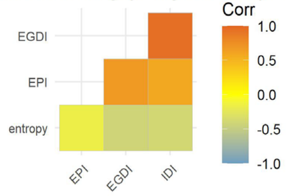
Correlation Heatmap 2005



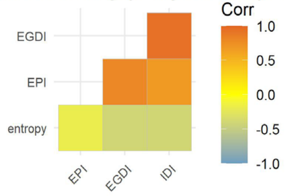
Correlation Heatmap (pairwise complete obs.) 2008



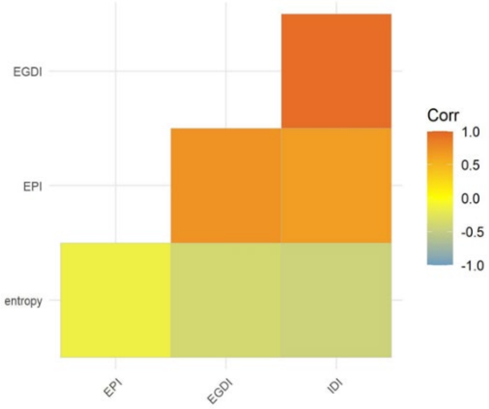
Correlation Heatmap (pairwise complete obs.) 2010



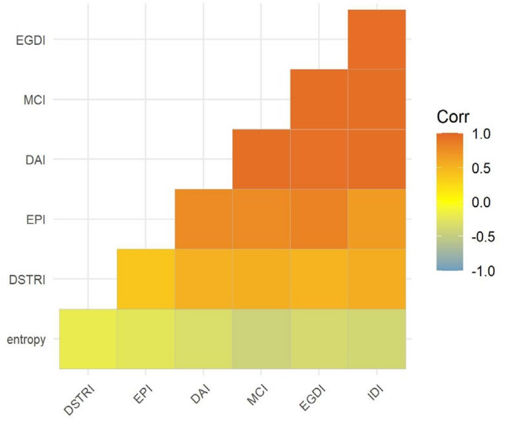
Correlation Heatmap (complete obs.) 2008



Correlation Heatmap (complete obs.) 2010



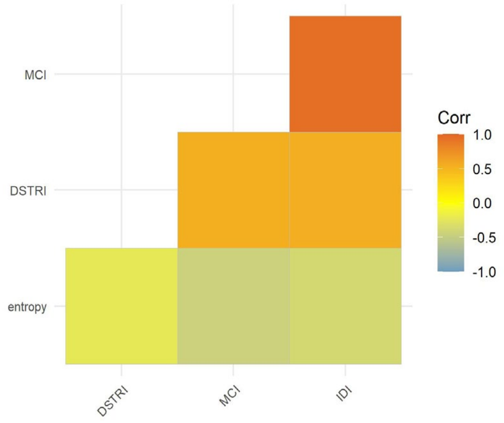
Correlation Heatmap 2012



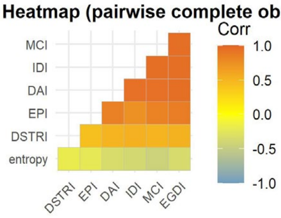
Correlation Heatmap 2014

Appendix

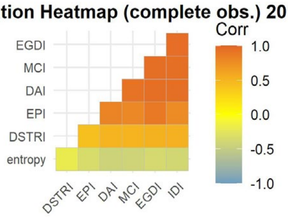
Correlation Heatmap 2015



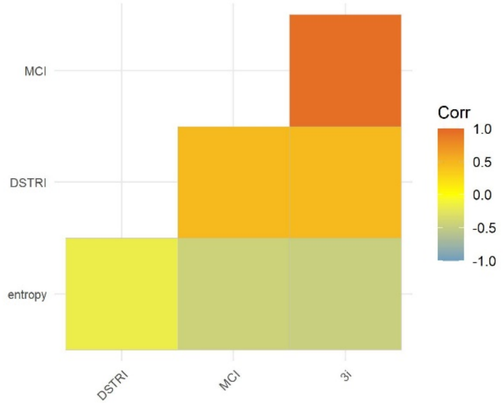
Correlation Heatmap (pairwise complete obs.) 2016



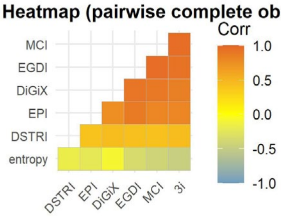
Correlation Heatmap (complete obs.) 2016



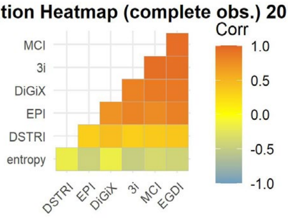
Correlation Heatmap 2017



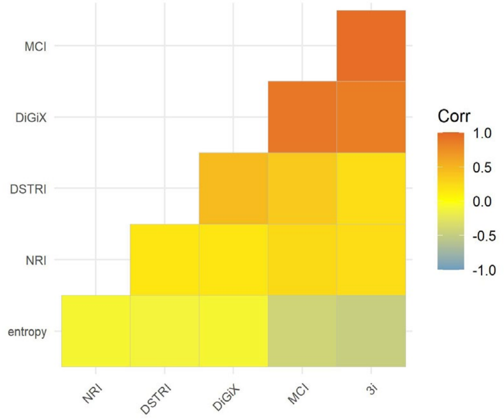
Correlation Heatmap (pairwise complete obs.) 2018



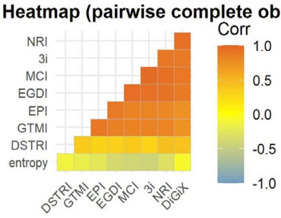
Correlation Heatmap (complete obs.) 2018



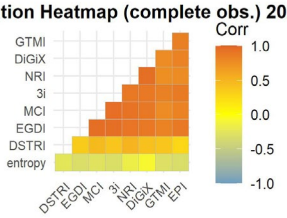
Correlation Heatmap 2019



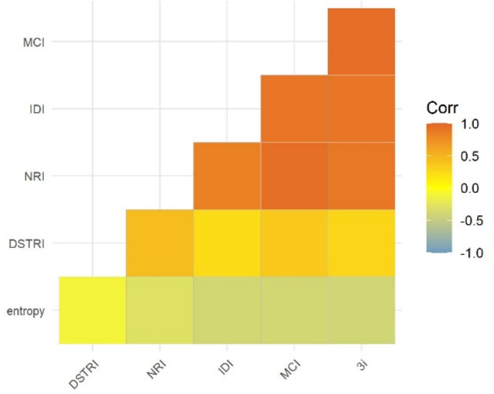
Correlation Heatmap (pairwise complete obs.) 2020



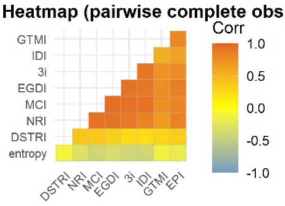
Correlation Heatmap (complete obs.) 2020



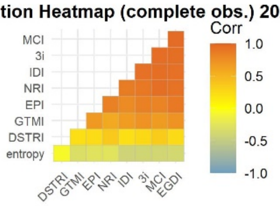
Correlation Heatmap 2021



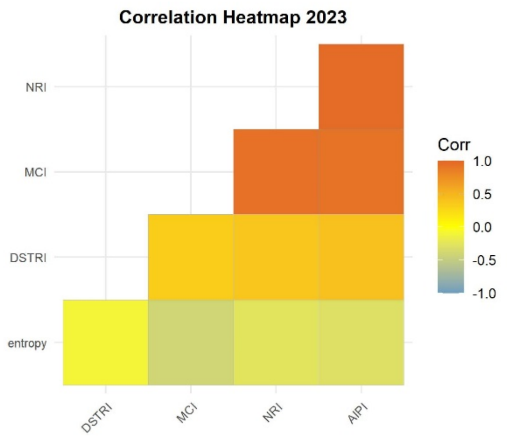
Correlation Heatmap (pairwise complete obs.) 2022



Correlation Heatmap (complete obs.) 2022

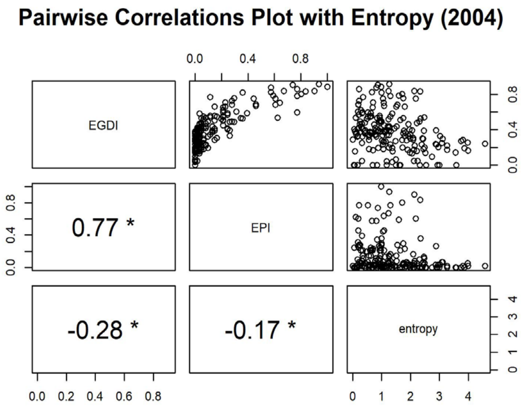


Appendix

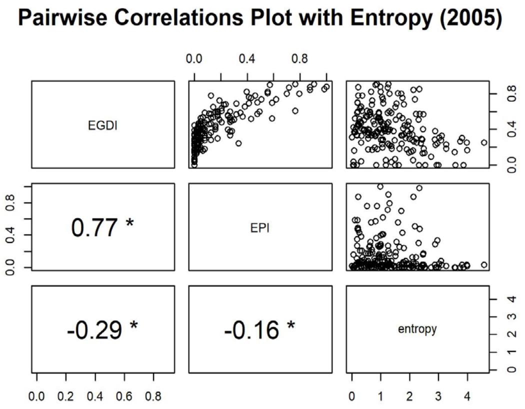


Correlation Heatmap 2023

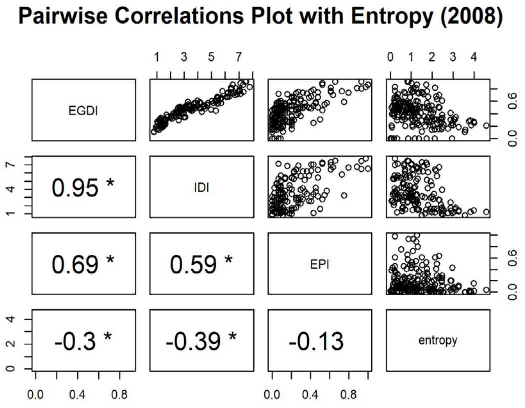
Pairs Plots



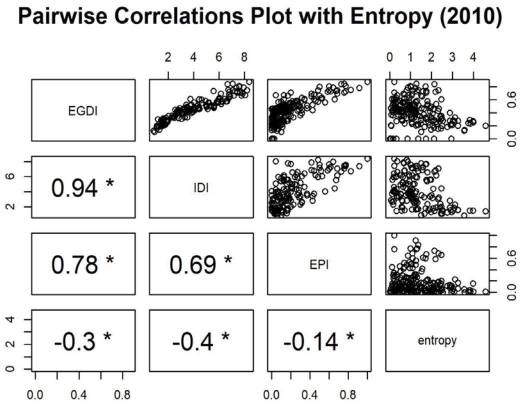
Pairwise Correlations Plot with Entropy (2004)



Pairwise Correlations Plot with Entropy (2005)



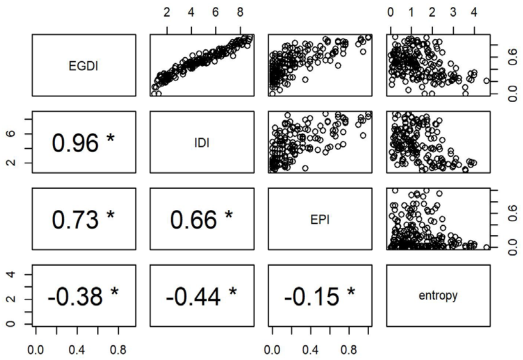
Pairwise Correlations Plot with Entropy (2008)



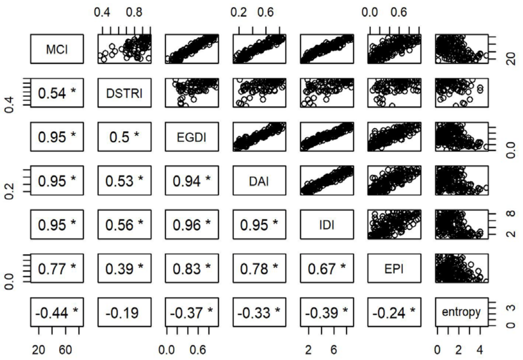
Pairwise Correlations Plot with Entropy (2010)

Appendix

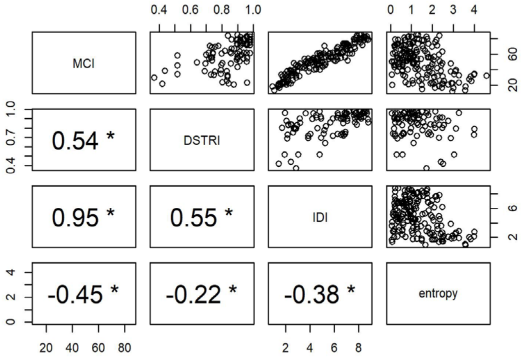
Pairwise Correlations Plot with Entropy (2012)



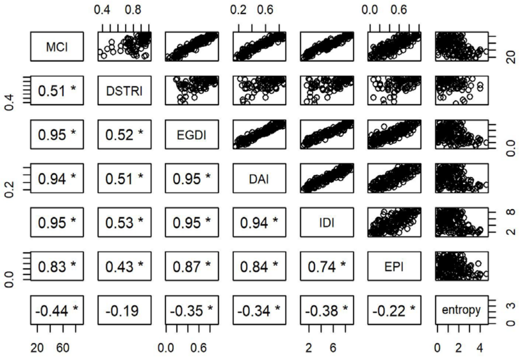
Pairwise Correlations Plot with Entropy (2014)



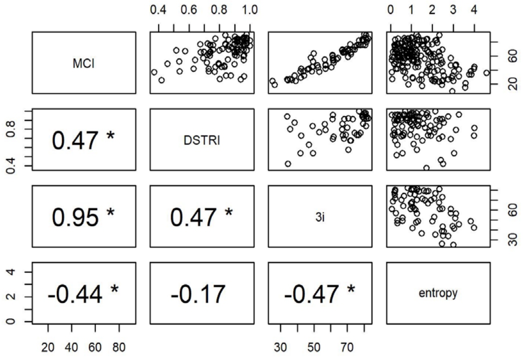
Pairwise Correlations Plot with Entropy (2015)



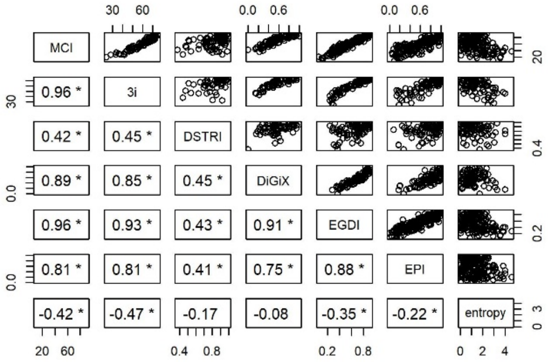
Pairwise Correlations Plot with Entropy (2016)



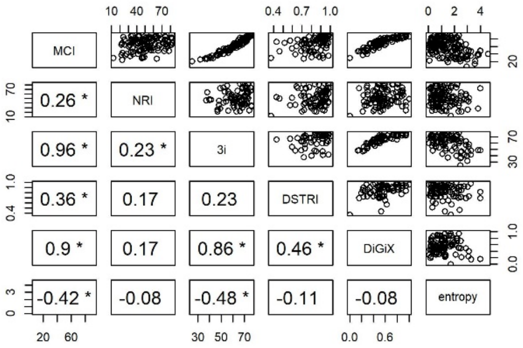
Pairwise Correlations Plot with Entropy (2017)



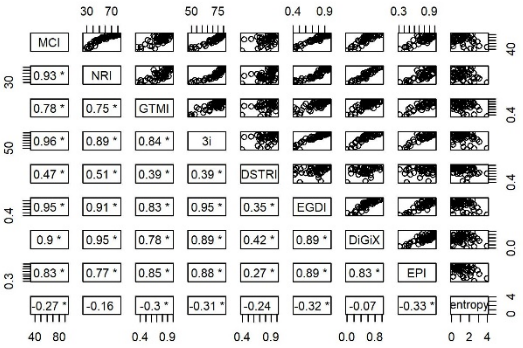
Pairwise Correlations Plot with Entropy (2018)



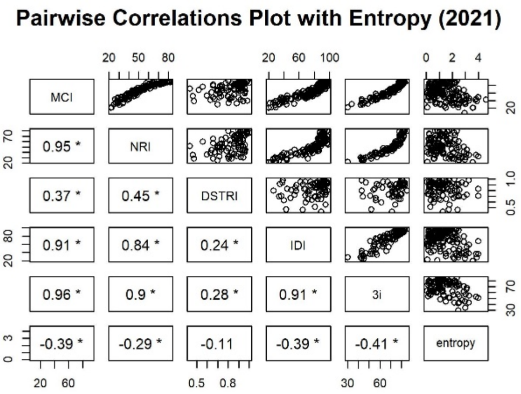
Pairwise Correlations Plot with Entropy (2019)



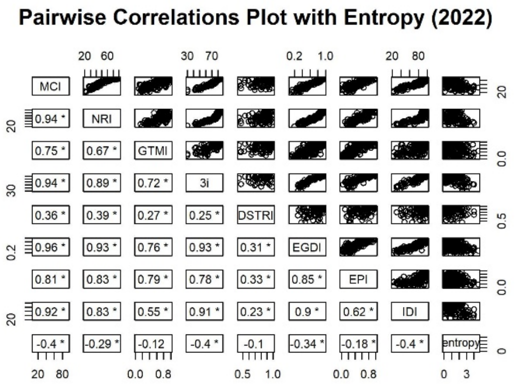
Pairwise Correlations Plot with Entropy (2020)



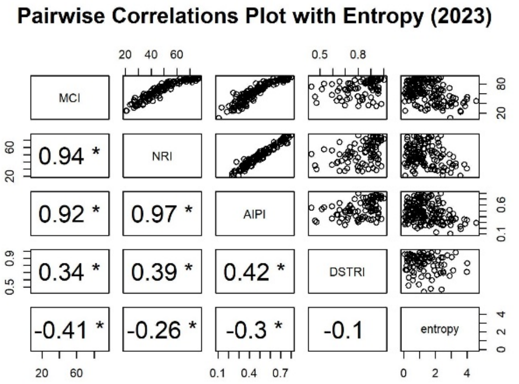
Appendix



Pairwise Correlations Plot with Entropy (2021)



Pairwise Correlations Plot with Entropy (2022)



Pairwise Correlations Plot with Entropy (2023)

Correlation Summary

Year	Observations	Var 1	Var 2	Correlation	P_Value	Strength	Significance	Direction
2004	pairwise	EGDI	entropy	-0.285	< 0.001	Small	*	Negative
2004	pairwise	EPI	entropy	-0.165	0.022	Small	*	Negative
2005	pairwise	EGDI	entropy	-0.295	< 0.001	Small	*	Negative
2005	pairwise	EPI	entropy	-0.157	0.03	Small	*	Negative
2008	complete	EGDI	entropy	-0.418	< 0.001	Medium	*	Negative
2008	complete	IDI	entropy	-0.386	< 0.001	Medium	*	Negative
2008	complete	EPI	entropy	-0.155	0.064	Small		Negative
2008	pairwise	EGDI	entropy	-0.3	< 0.001	Small	*	Negative
2008	pairwise	IDI	entropy	-0.393	< 0.001	Medium	*	Negative
2008	pairwise	EPI	entropy	-0.134	0.064	Small		Negative
2010	complete	EGDI	entropy	-0.41	< 0.001	Medium	*	Negative
2010	complete	IDI	entropy	-0.398	< 0.001	Medium	*	Negative
2010	complete	EPI	entropy	-0.177	0.048	Small	*	Negative
2010	pairwise	EGDI	entropy	-0.298	< 0.001	Small	*	Negative
2010	pairwise	IDI	entropy	-0.405	< 0.001	Medium	*	Negative

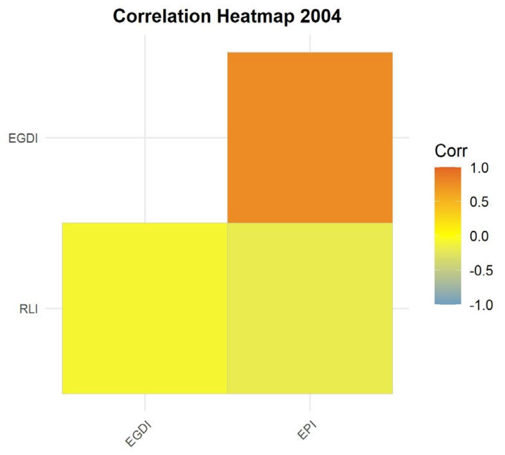
2010	pairwise	EPI	entropy	-0.143	0.048	Small	*	Negative
2012	pairwise	EGDI	entropy	-0.382	< 0.001	Medium	*	Negative
2012	pairwise	IDI	entropy	-0.436	< 0.001	Medium	*	Negative
2012	pairwise	EPI	entropy	-0.146	0.044	Small	*	Negative
2014	pairwise	MCI	entropy	-0.44	< 0.001	Medium	*	Negative
2014	pairwise	DSTRI	entropy	-0.19	0.073	Small		Negative
2014	pairwise	EGDI	entropy	-0.374	< 0.001	Medium	*	Negative
2014	pairwise	DAI	entropy	-0.331	< 0.001	Medium	*	Negative
2014	pairwise	IDI	entropy	-0.392	< 0.001	Medium	*	Negative
2014	pairwise	EPI	entropy	-0.244	0.001	Small	*	Negative
2015	pairwise	MCI	entropy	-0.452	< 0.001	Medium	*	Negative
2015	pairwise	DSTRI	entropy	-0.217	0.04	Small	*	Negative
2015	pairwise	IDI	entropy	-0.381	< 0.001	Medium	*	Negative
2016	complete	MCI	entropy	-0.424	< 0.001	Medium	*	Negative
2016	complete	DSTRI	entropy	-0.19	0.075	Small		Negative
2016	complete	EGDI	entropy	-0.374	< 0.001	Medium	*	Negative
2016	complete	DAI	entropy	-0.433	< 0.001	Medium	*	Negative
2016	complete	IDI	entropy	-0.388	< 0.001	Medium	*	Negative
2016	complete	EPI	entropy	-0.344	0.002	Medium	*	Negative
2016	pairwise	MCI	entropy	-0.441	< 0.001	Medium	*	Negative
2016	pairwise	DSTRI	entropy	-0.189	0.075	Small		Negative
2016	pairwise	EGDI	entropy	-0.351	< 0.001	Medium	*	Negative
2016	pairwise	DAI	entropy	-0.336	< 0.001	Medium	*	Negative
2016	pairwise	IDI	entropy	-0.381	< 0.001	Medium	*	Negative
2016	pairwise	EPI	entropy	-0.225	0.002	Small	*	Negative
2017	pairwise	MCI	entropy	-0.435	< 0.001	Medium	*	Negative
2017	pairwise	DSTRI	entropy	-0.173	0.103	Small		Negative
2017	pairwise	3i	entropy	-0.474	< 0.001	Medium	*	Negative
2018	complete	MCI	entropy	-0.37	< 0.001	Medium	*	Negative
2018	complete	DSTRI	entropy	-0.167	0.104	Small		Negative
2018	complete	3i	entropy	-0.482	< 0.001	Medium	*	Negative
2018	complete	DiGiX	entropy	-0.156	0.444	Small		Negative
2018	complete	EGDI	entropy	-0.425	< 0.001	Medium	*	Negative
2018	complete	EPI	entropy	-0.435	< 0.001	Medium	*	Negative
2018	pairwise	MCI	entropy	-0.421	< 0.001	Medium	*	Negative

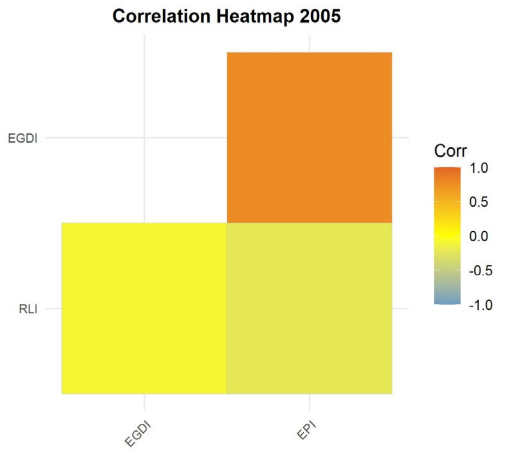
2018	pairwise	DSTRI	entropy	-0.172	0.104	Small		Negative
2018	pairwise	3i	entropy	-0.473	< 0.001	Medium	*	Negative
2018	pairwise	DiGiX	entropy	-0.078	0.444	Very Small		Negative
2018	pairwise	EGDI	entropy	-0.35	< 0.001	Medium	*	Negative
2018	pairwise	EPI	entropy	-0.219	< 0.001	Small	*	Negative
2019	pairwise	MCI	entropy	-0.418	< 0.001	Medium	*	Negative
2019	pairwise	NRI	entropy	-0.085	0.357	Very Small		Negative
2019	pairwise	3i	entropy	-0.484	< 0.001	Medium	*	Negative
2019	pairwise	DSTRI	entropy	-0.114	0.284	Small		Negative
2019	pairwise	DiGiX	entropy	-0.077	0.447	Very Small		Negative
2020	complete	MCI	entropy	-0.267	< 0.001	Small	*	Negative
2020	complete	NRI	entropy	-0.162	0	Small	*	Negative
2020	complete	GTMI	entropy	-0.297	0.014	Small	*	Negative
2020	complete	3i	entropy	-0.305	< 0.001	Medium	*	Negative
2020	complete	DSTRI	entropy	-0.24	0.251	Small		Negative
2020	complete	EGDI	entropy	-0.316	< 0.001	Medium	*	Negative
2020	complete	DiGiX	entropy	-0.073	< 0.001	Very Small		Negative
2020	complete	EPI	entropy	-0.327	0.001	Medium	*	Negative
2020	pairwise	MCI	entropy	-0.403	< 0.001	Medium	*	Negative
2020	pairwise	NRI	entropy	-0.304	0	Medium	*	Negative
2020	pairwise	GTMI	entropy	-0.176	0.014	Small	*	Negative
2020	pairwise	3i	entropy	-0.421	< 0.001	Medium	*	Negative
2020	pairwise	DSTRI	entropy	-0.122	0.251	Small		Negative
2020	pairwise	EGDI	entropy	-0.358	< 0.001	Medium	*	Negative
2020	pairwise	DiGiX	entropy	-0.056	0.584	Very Small		Negative
2020	pairwise	EPI	entropy	-0.233	0.001	Small	*	Negative
2021	pairwise	MCI	entropy	-0.393	< 0.001	Medium	*	Negative
2021	pairwise	NRI	entropy	-0.293	0.001	Small	*	Negative
2021	pairwise	DSTRI	entropy	-0.112	0.295	Small		Negative
2021	pairwise	IDI	entropy	-0.388	< 0.001	Medium	*	Negative
2021	pairwise	3i	entropy	-0.411	< 0.001	Medium	*	Negative
2022	complete	MCI	entropy	-0.387	< 0.001	Medium	*	Negative
2022	complete	NRI	entropy	-0.262	0.001	Small	*	Negative
2022	complete	GTMI	entropy	-0.314	0.088	Medium		Negative
2022	complete	3i	entropy	-0.396	< 0.001	Medium	*	Negative

2022	complete	DSTRI	entropy	-0.063	0.341	Very Small		Negative
2022	complete	EGDI	entropy	-0.381	< 0.001	Medium	*	Negative
2022	complete	EPI	entropy	-0.272	< 0.001	Small	*	Negative
2022	complete	IDI	entropy	-0.42	< 0.001	Medium	*	Negative
2022	pairwise	MCI	entropy	-0.396	< 0.001	Medium	*	Negative
2022	pairwise	NRI	entropy	-0.287	0.001	Small	*	Negative
2022	pairwise	GTMI	entropy	-0.122	0.088	Small		Negative
2022	pairwise	3i	entropy	-0.402	< 0.001	Medium	*	Negative
2022	pairwise	DSTRI	entropy	-0.102	0.341	Small		Negative
2022	pairwise	EGDI	entropy	-0.343	< 0.001	Medium	*	Negative
2022	pairwise	EPI	entropy	-0.177	< 0.001	Small	*	Negative
2022	pairwise	IDI	entropy	-0.405	< 0.001	Medium	*	Negative
2023	pairwise	MCI	entropy	-0.407	< 0.001	Medium	*	Negative
2023	pairwise	NRI	entropy	-0.263	0.002	Small	*	Negative
2023	pairwise	AIPI	entropy	-0.3	< 0.001	Medium	*	Negative
2023	pairwise	DSTRI	entropy	-0.105	0.326	Small		Negative

A.2.3 Correlation Analysis - RLI

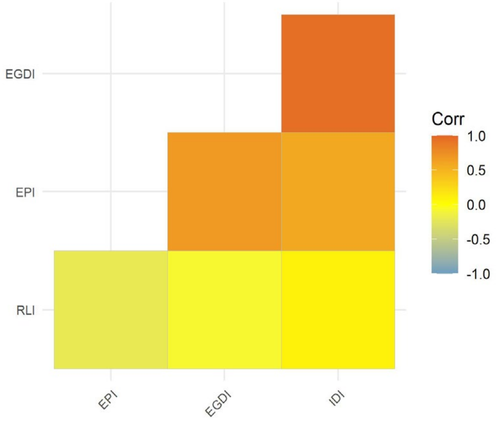
Heatmaps


Correlation Heatmap 2004

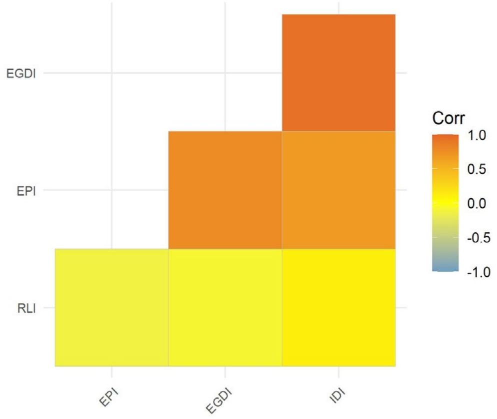

Correlation Heatmap 2005

Appendix

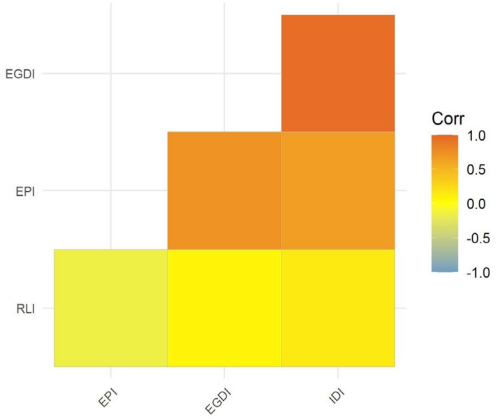
Correlation Heatmap 2008



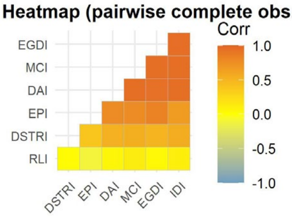
Correlation Heatmap 2010



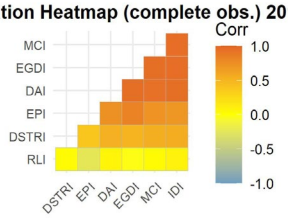
Correlation Heatmap 2012



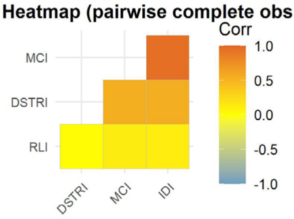
Correlation Heatmap (pairwise complete obs.) 2014



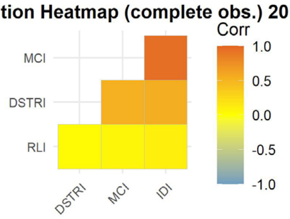
Correlation Heatmap (complete obs.) 2014



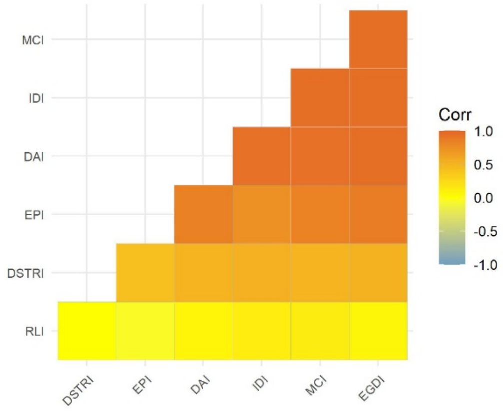
Correlation Heatmap (pairwise complete obs.) 2015



Correlation Heatmap (complete obs.) 2015

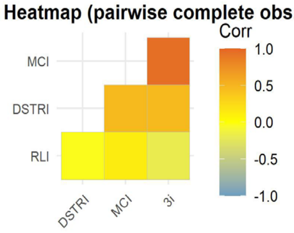


Correlation Heatmap 2016

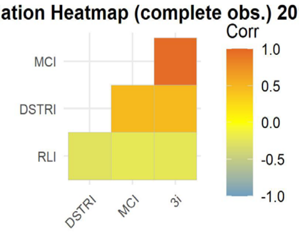


Appendix

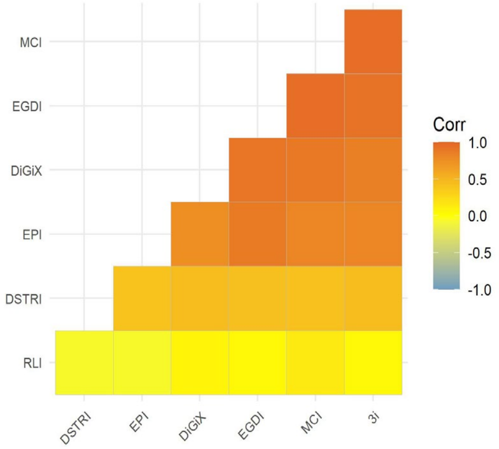
Correlation Heatmap (pairwise complete obs.) 2017



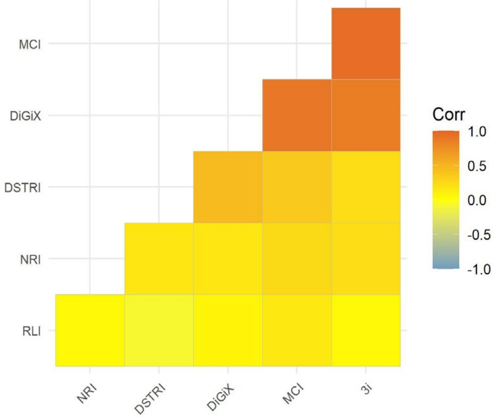
Correlation Heatmap (complete obs.) 2017



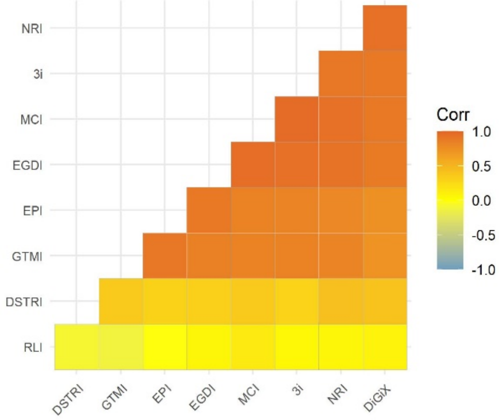
Correlation Heatmap 2018



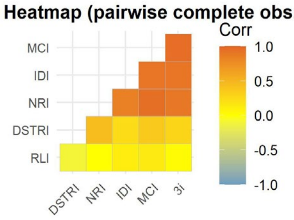
Correlation Heatmap 2019



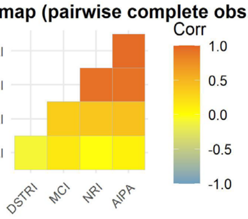
Correlation Heatmap 2020



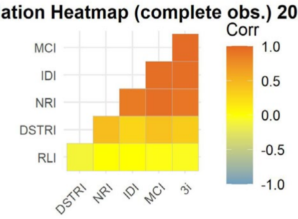
Correlation Heatmap (pairwise complete obs.) 2021



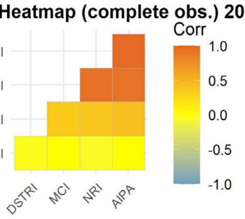
Correlation Heatmap (pairwise complete obs.) 2022



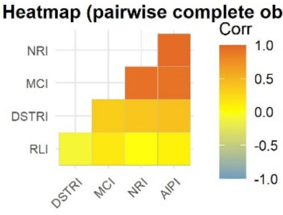
Correlation Heatmap (complete obs.) 2021



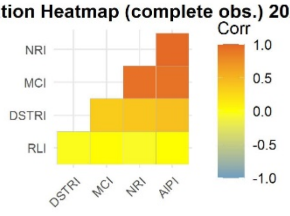
Correlation Heatmap (complete obs.) 2022



Correlation Heatmap (pairwise complete obs.) 2022



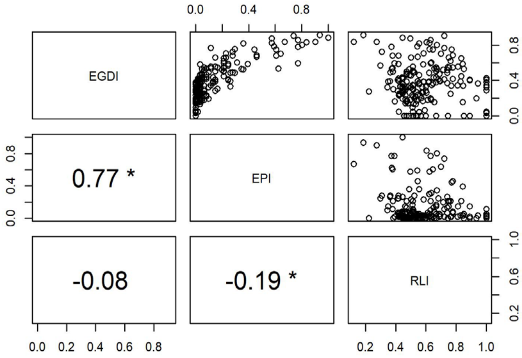
Correlation Heatmap (complete obs.) 2023



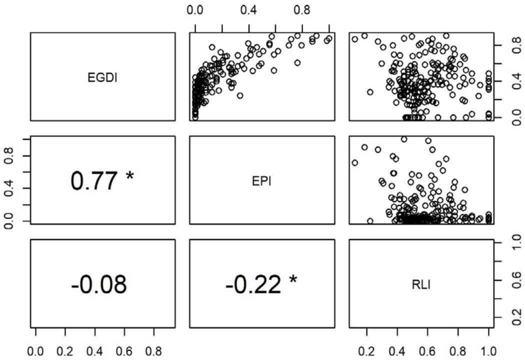
Appendix

Pairs Plots

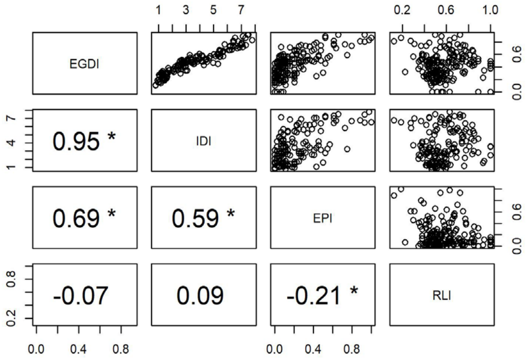
Pairwise Correlations Plot with RLI (2004)



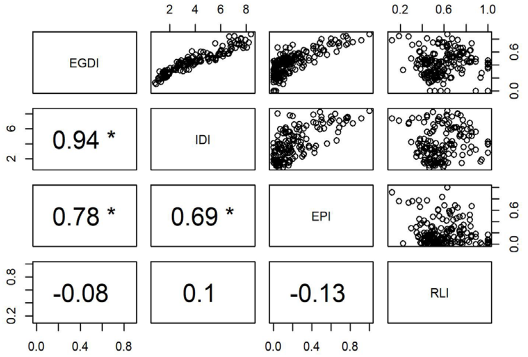
Pairwise Correlations Plot with RLI (2005)



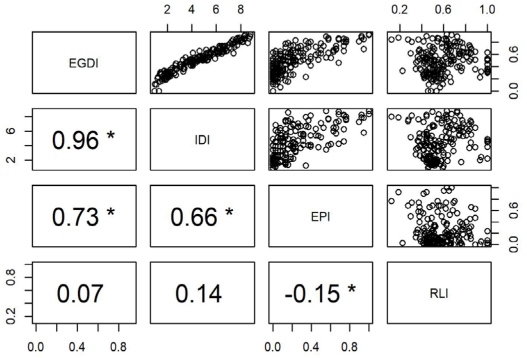
Pairwise Correlations Plot with RLI (2008)



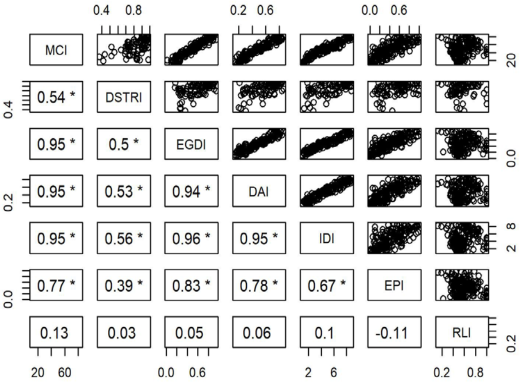
Pairwise Correlations Plot with RLI (2010)



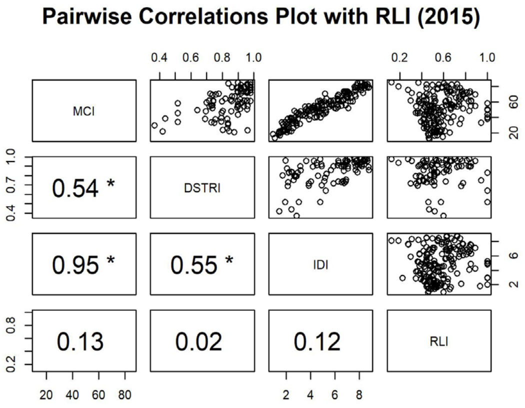
Pairwise Correlations Plot with RLI (2012)

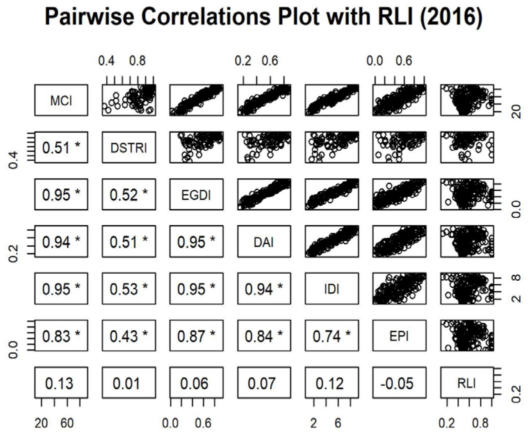


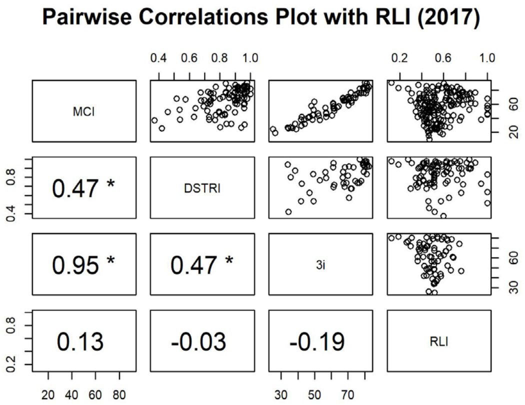
Pairwise Correlations Plot with RLI (2014)

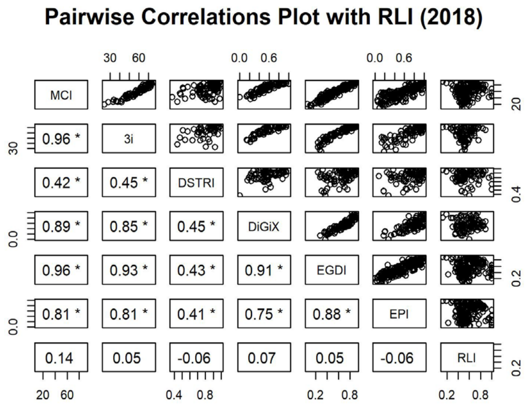


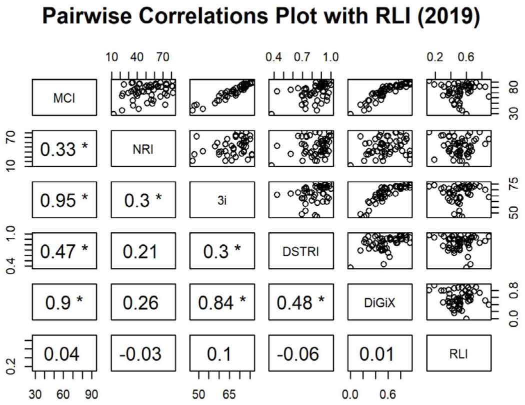
Appendix

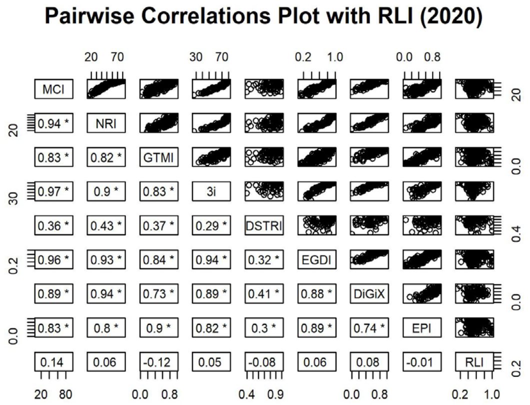

Pairwise Correlations Plot with RLI (2015)


Pairwise Correlations Plot with RLI (2016)


Pairwise Correlations Plot with RLI (2017)

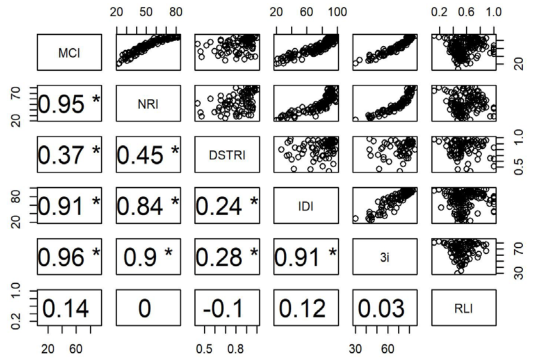

Pairwise Correlations Plot with RLI (2018)


Pairwise Correlations Plot with RLI (2019)

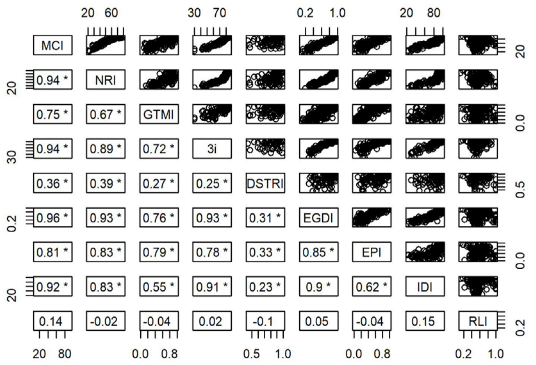

Pairwise Correlations Plot with RLI (2020)

Appendix

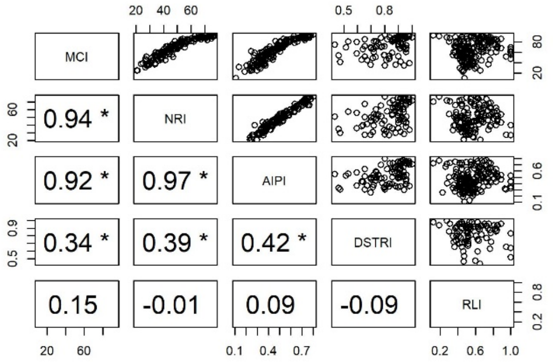
Pairwise Correlations Plot with RLI (2021)



Pairwise Correlations Plot with RLI (2022)



Pairwise Correlations Plot with RLI (2023)



Correlation Summary

Year	Observations	Var 1	Var 2	Correlation	P_Value	Strength	Significance	Direction
2004	pairwise	EGDI	RLI	-0.083	0.250	Very Small		Negative
2004	pairwise	EPI	RLI	-0.185	0.010	Small	*	Negative
2005	pairwise	EGDI	RLI	-0.084	0.244	Very Small		Negative
2005	pairwise	EPI	RLI	-0.216	0.003	Small	*	Negative
2008	pairwise	EGDI	RLI	-0.072	0.323	Very Small		Negative
2008	pairwise	IDI	RLI	0.086	0.295	Very Small		Positive
2008	pairwise	EPI	RLI	-0.206	0.004	Small	*	Negative
2010	pairwise	EGDI	RLI	-0.080	0.268	Very Small		Negative
2010	pairwise	IDI	RLI	0.103	0.208	Small		Positive
2010	pairwise	EPI	RLI	-0.127	0.080	Small		Negative
2012	pairwise	EGDI	RLI	0.071	0.329	Very Small		Positive
2012	pairwise	IDI	RLI	0.139	0.075	Small		Positive
2012	pairwise	EPI	RLI	-0.148	0.041	Small	*	Negative
2014	complete	MCI	RLI	0.030	0.091	Very Small		Positive

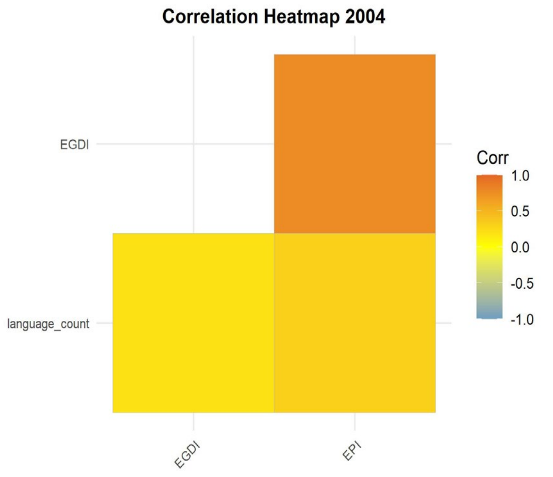
2014	complete	DSTRI	RLI	0.031	0.776	Very Small		Positive
2014	complete	EGDI	RLI	-0.018	0.486	Very Small		Negative
2014	complete	DAI	RLI	0.071	0.415	Very Small		Positive
2014	complete	IDI	RLI	0.079	0.171	Very Small		Positive
2014	complete	EPI	RLI	-0.235	0.140	Small		Negative
2014	pairwise	MCI	RLI	0.129	0.091	Small		Positive
2014	pairwise	DSTRI	RLI	0.031	0.776	Very Small		Positive
2014	pairwise	EGDI	RLI	0.051	0.486	Very Small		Positive
2014	pairwise	DAI	RLI	0.061	0.415	Very Small		Positive
2014	pairwise	IDI	RLI	0.104	0.171	Small		Positive
2014	pairwise	EPI	RLI	-0.107	0.140	Small		Negative
2015	complete	MCI	RLI	0.064	0.084	Very Small		Positive
2015	complete	DSTRI	RLI	0.033	0.884	Very Small		Positive
2015	complete	IDI	RLI	0.115	0.110	Small		Positive
2015	pairwise	MCI	RLI	0.132	0.084	Small		Positive
2015	pairwise	DSTRI	RLI	0.016	0.884	Very Small		Positive
2015	pairwise	IDI	RLI	0.121	0.110	Small		Positive
2016	complete	MCI	RLI	0.056	0.078	Very Small		Positive
2016	complete	DSTRI	RLI	0.026	0.951	Very Small		Positive
2016	complete	EGDI	RLI	0.034	0.433	Very Small		Positive
2016	complete	DAI	RLI	0.124	0.353	Small		Positive
2016	complete	IDI	RLI	0.108	0.118	Small		Positive
2016	complete	EPI	RLI	-0.097	0.503	Very Small		Negative
2016	pairwise	MCI	RLI	0.135	0.078	Small		Positive
2016	pairwise	DSTRI	RLI	0.007	0.951	Very Small		Positive
2016	pairwise	EGDI	RLI	0.057	0.433	Very Small		Positive
2016	pairwise	DAI	RLI	0.070	0.353	Very Small		Positive
2016	pairwise	IDI	RLI	0.118	0.118	Small		Positive
2016	pairwise	EPI	RLI	-0.049	0.503	Very Small		Negative
2017	complete	MCI	RLI	-0.217	0.081	Small		Negative
2017	complete	DSTRI	RLI	-0.267	0.787	Small		Negative
2017	complete	3i	RLI	-0.226	0.107	Small		Negative
2017	pairwise	MCI	RLI	0.133	0.081	Small		Positive
2017	pairwise	DSTRI	RLI	-0.029	0.787	Very Small		Negative
2017	pairwise	3i	RLI	-0.194	0.107	Small		Negative

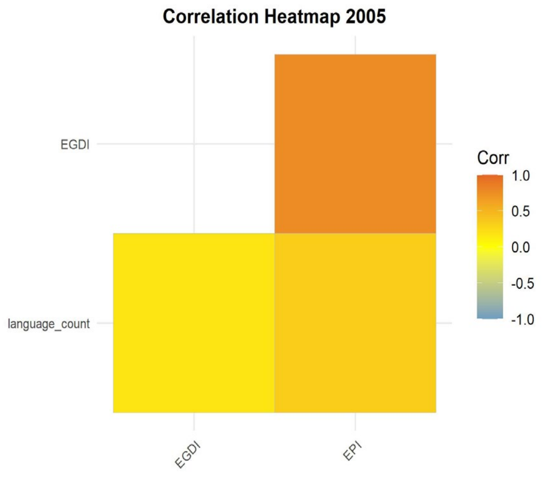
2018	pairwise	MCI	RLI	0.142	0.063	Small		Positive
2018	pairwise	DSTRI	RLI	-0.058	0.589	Very Small		Negative
2018	pairwise	3i	RLI	0.054	0.624	Very Small		Positive
2018	pairwise	DiGiX	RLI	0.068	0.504	Very Small		Positive
2018	pairwise	EGDI	RLI	0.049	0.497	Very Small		Positive
2018	pairwise	EPI	RLI	-0.058	0.427	Very Small		Negative
2019	pairwise	MCI	RLI	0.154	0.043	Small	*	Positive
2019	pairwise	NRI	RLI	0.051	0.582	Very Small		Positive
2019	pairwise	3i	RLI	0.051	0.634	Very Small		Positive
2019	pairwise	DSTRI	RLI	-0.069	0.520	Very Small		Negative
2019	pairwise	DiGiX	RLI	0.070	0.489	Very Small		Positive
2020	pairwise	MCI	RLI	0.143	0.061	Small		Positive
2020	pairwise	NRI	RLI	0.057	0.513	Very Small		Positive
2020	pairwise	GTMI	RLI	-0.117	0.105	Small		Negative
2020	pairwise	3i	RLI	0.054	0.594	Very Small		Positive
2020	pairwise	DSTRI	RLI	-0.078	0.465	Very Small		Negative
2020	pairwise	EGDI	RLI	0.055	0.446	Very Small		Positive
2020	pairwise	DiGiX	RLI	0.080	0.434	Very Small		Positive
2020	pairwise	EPI	RLI	-0.011	0.885	Very Small		Negative
2021	complete	MCI	RLI	-0.024	0.067	Very Small		Negative
2021	complete	NRI	RLI	0.013	0.964	Very Small		Positive
2021	complete	DSTRI	RLI	-0.099	0.340	Very Small		Negative
2021	complete	IDI	RLI	-0.005	0.116	Very Small		Negative
2021	complete	3i	RLI	-0.041	0.737	Very Small		Negative
2021	pairwise	MCI	RLI	-0.024	0.067	Very Small		Negative
2021	pairwise	NRI	RLI	0.013	0.964	Very Small		Positive
2021	pairwise	DSTRI	RLI	-0.099	0.340	Very Small		Negative
2021	pairwise	IDI	RLI	-0.005	0.116	Very Small		Negative
2021	pairwise	3i	RLI	-0.041	0.737	Very Small		Negative
2022	complete	MCI	RLI	-0.020	0.075	Very Small		Negative
2022	complete	NRI	RLI	-0.007	0.837	Very Small		Negative
2022	complete	GTMI	RLI	0.102	0.587	Small		Positive
2022	complete	3i	RLI	-0.078	0.868	Very Small		Negative
2022	complete	DSTRI	RLI	-0.085	0.369	Very Small		Negative
2022	complete	EGDI	RLI	0.004	0.454	Very Small		Positive

2022	complete	EPI	RLI	-0.143	0.554	Small		Negative
2022	complete	IDI	RLI	-0.012	0.053	Very Small		Negative
2022	pairwise	MCI	RLI	0.136	0.075	Small		Positive
2022	pairwise	NRI	RLI	-0.018	0.837	Very Small		Negative
2022	pairwise	GTMI	RLI	-0.039	0.587	Very Small		Negative
2022	pairwise	3i	RLI	0.017	0.868	Very Small		Positive
2022	pairwise	DSTRI	RLI	-0.096	0.369	Very Small		Negative
2022	pairwise	EGDI	RLI	0.054	0.454	Very Small		Positive
2022	pairwise	EPI	RLI	-0.043	0.554	Very Small		Negative
2022	pairwise	IDI	RLI	0.149	0.053	Small		Positive
2023	complete	MCI	RLI	0.017	0.054	Very Small		Positive
2023	complete	NRI	RLI	-0.049	0.931	Very Small		Negative
2023	complete	AIPI	RLI	0.009	0.217	Very Small		Positive
2023	complete	DSTRI	RLI	-0.028	0.414	Very Small		Negative
2023	pairwise	MCI	RLI	0.147	0.054	Small		Positive
2023	pairwise	NRI	RLI	-0.008	0.931	Very Small		Negative
2023	pairwise	AIPI	RLI	0.094	0.217	Very Small		Positive
2023	pairwise	DSTRI	RLI	-0.088	0.414	Very Small		Negative

A.2.4 Correlation Analysis - Language Count

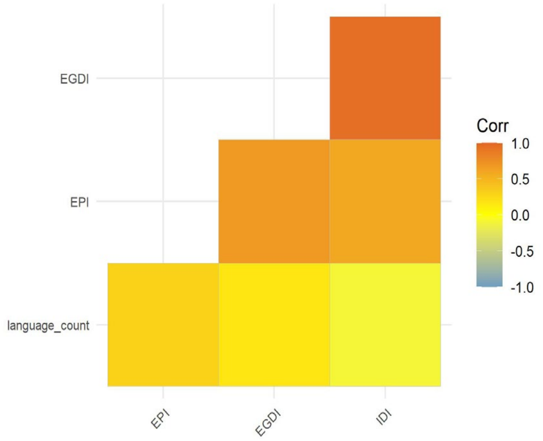
Heatmaps



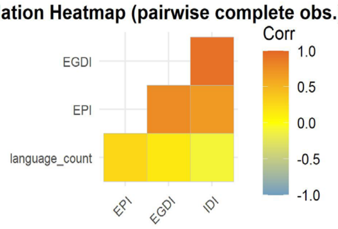


Appendix

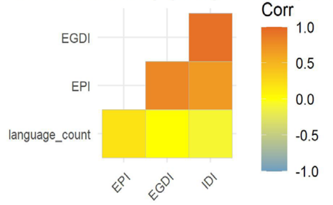
Correlation Heatmap 2008



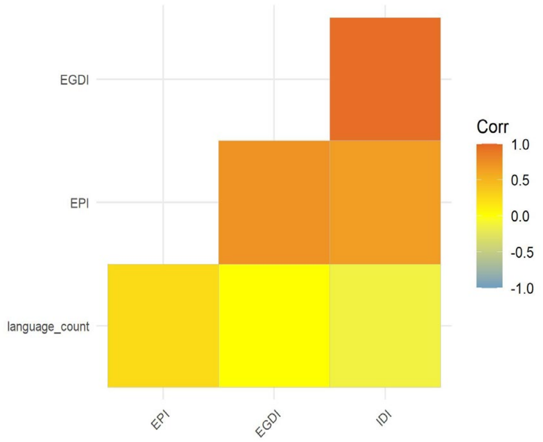
Correlation Heatmap (pairwise complete obs.) 2010



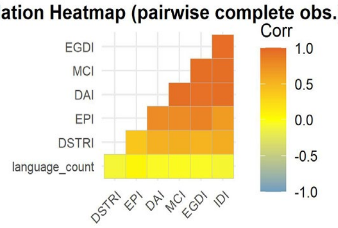
Correlation Heatmap (complete obs.) 2010



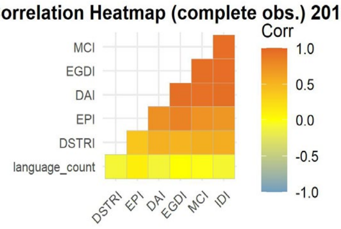
Correlation Heatmap 2012



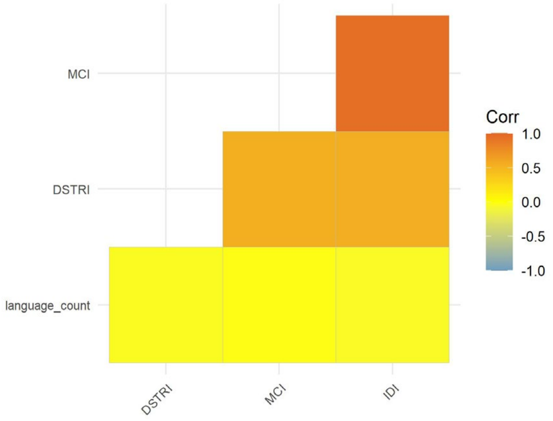
Correlation Heatmap (pairwise complete obs.) 2014



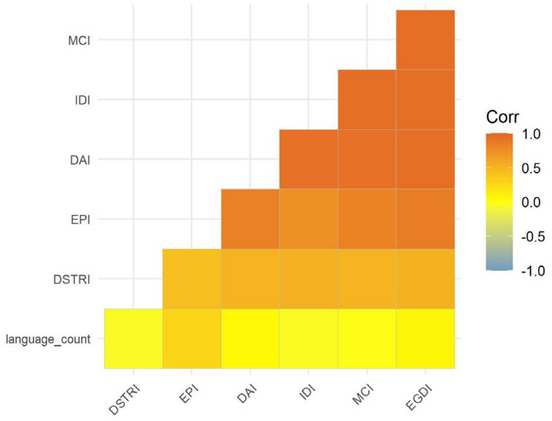
Correlation Heatmap (complete obs.) 2014



Correlation Heatmap 2015

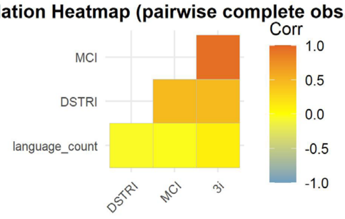


Correlation Heatmap 2016

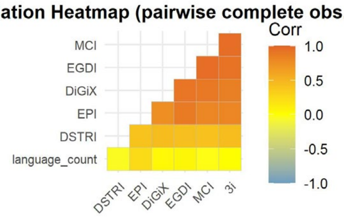


Appendix

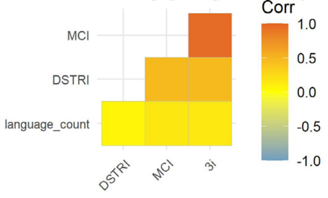
Correlation Heatmap (pairwise complete obs.) 2017



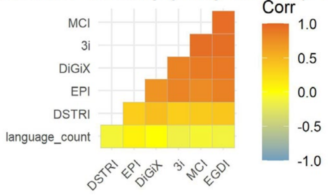
Correlation Heatmap (pairwise complete obs.) 2018



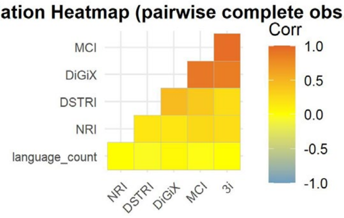
Correlation Heatmap (complete obs.) 2017



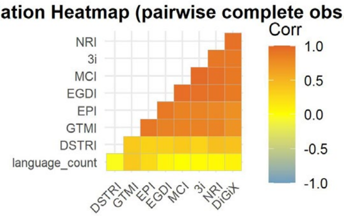
Correlation Heatmap (complete obs.) 2018



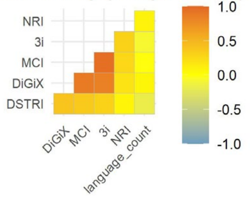
Correlation Heatmap (pairwise complete obs.) 2019



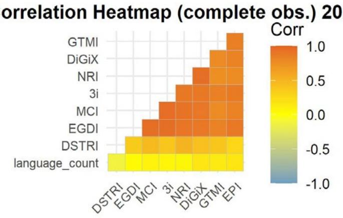
Correlation Heatmap (pairwise complete obs.) 2020



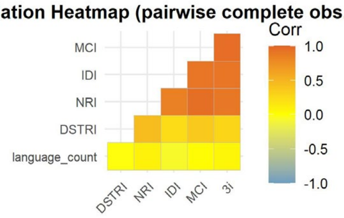
Correlation Heatmap (complete obs.) 2019



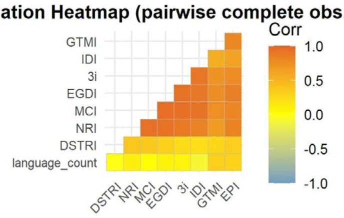
Correlation Heatmap (complete obs.) 2020



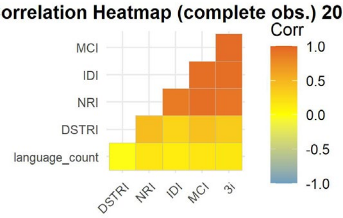
Correlation Heatmap (pairwise complete obs.) 2021



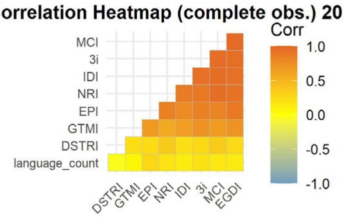
Correlation Heatmap (pairwise complete obs.) 2022



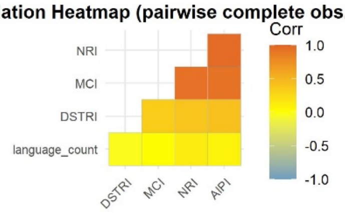
Correlation Heatmap (complete obs.) 2021



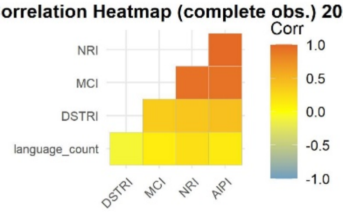
Correlation Heatmap (complete obs.) 2022



Correlation Heatmap (pairwise complete obs.) 2023



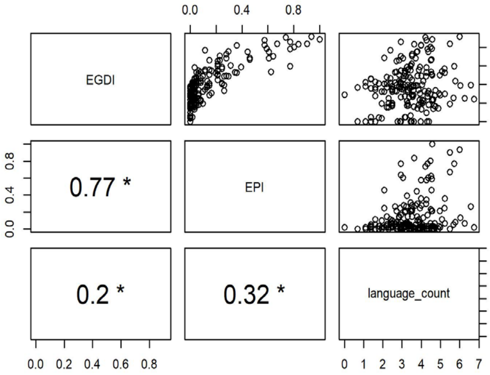
Correlation Heatmap (complete obs.) 2023



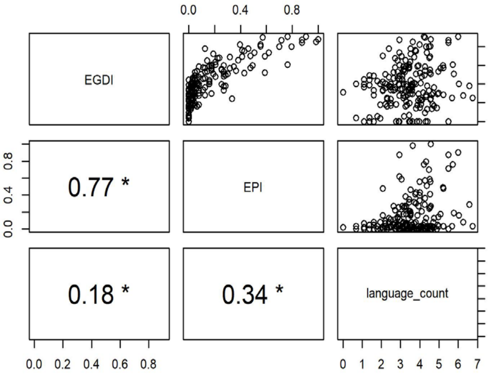
Appendix

Pairs Plots

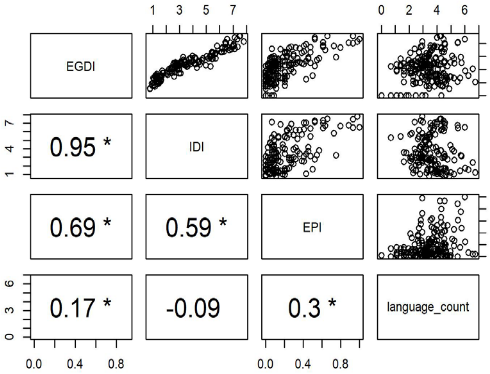
Pairwise Correlations Plot with language count (2004)



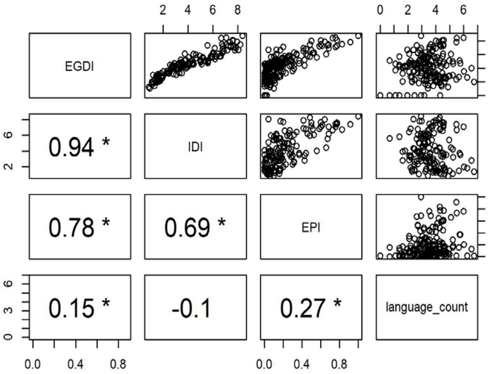
Pairwise Correlations Plot with language count (2005)



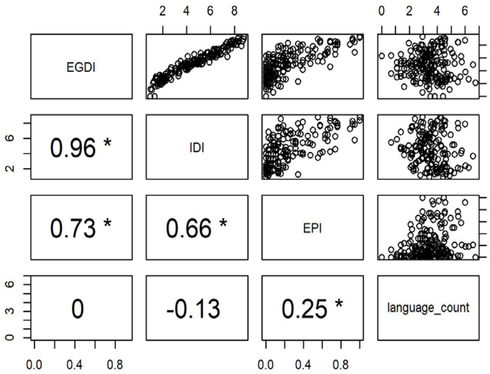
Pairwise Correlations Plot with language count (2008)



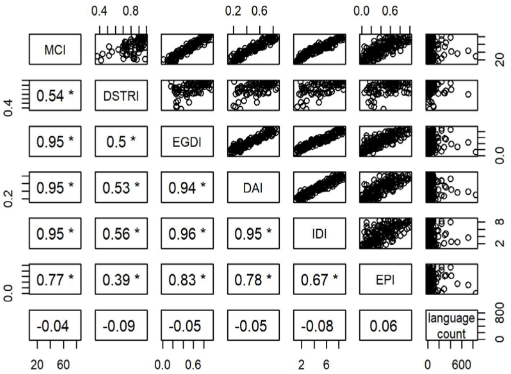
Pairwise Correlations Plot with language count (2010)



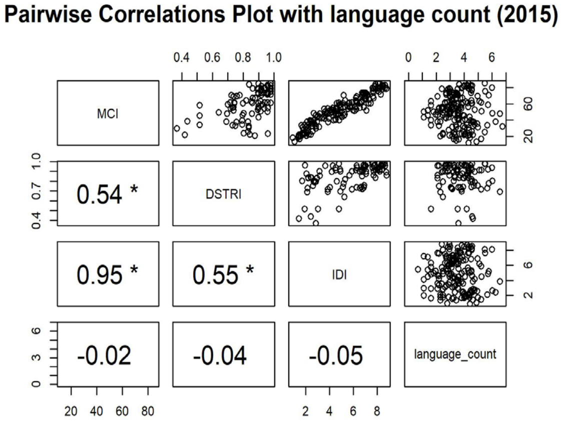
Pairwise Correlations Plot with language count (2012)

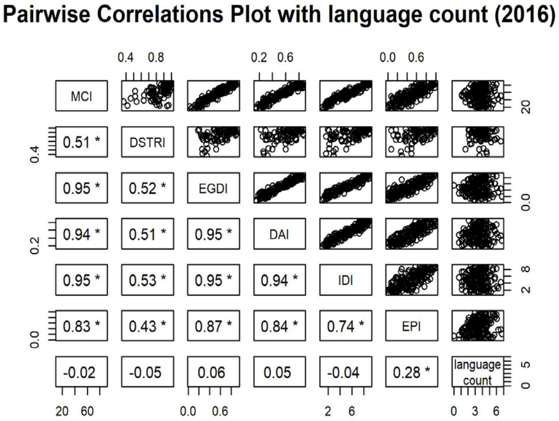


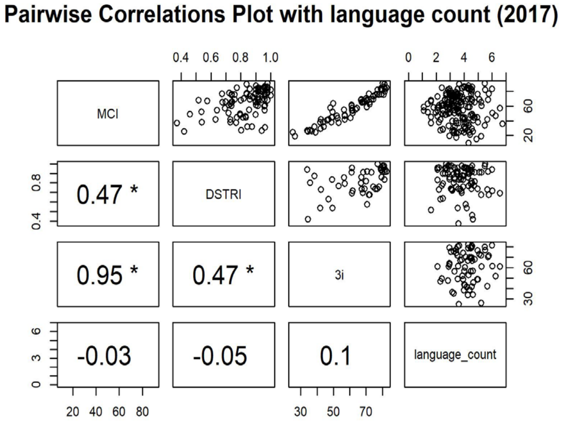
Pairwise Correlations Plot with Language Count (2014)

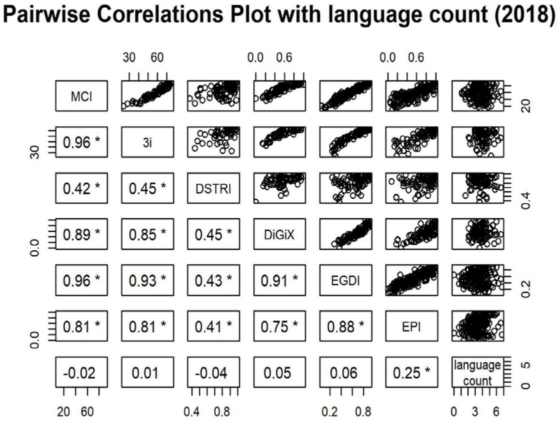


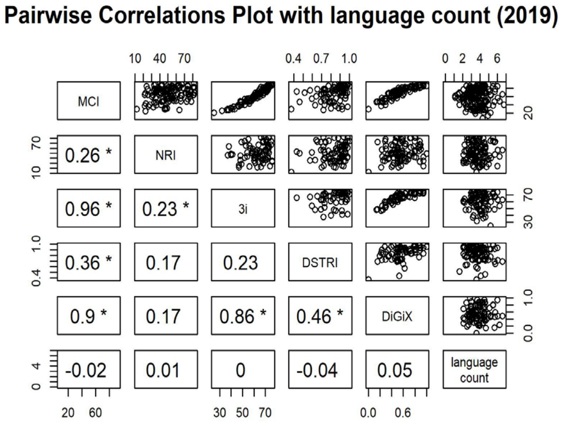
Appendix

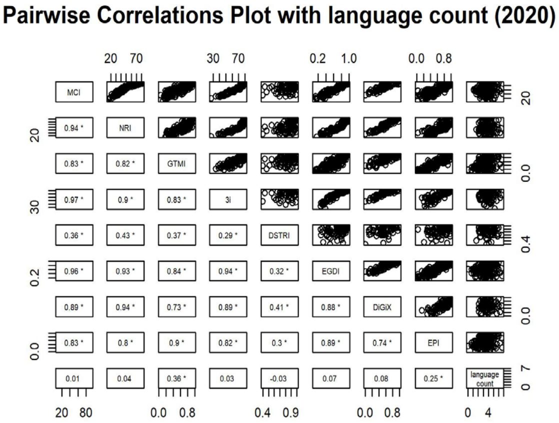

Pairwise Correlations Plot with language count (2015)
MCI
0.54 * DSTRI
0.95 * 0.55 * IDI
-0.02 -0.04 -0.05 language_count
0.4 0.6 0.8 1.0
0 2 4 6
20 40 60 80
2 4 6 8


Pairwise Correlations Plot with language count (2016)
MCI
0.51 * DSTRI
0.95 * 0.52 * EGDI
0.94 * 0.51 * 0.95 * DAI
0.95 * 0.53 * 0.95 * 0.94 * IDI
0.83 * 0.43 * 0.87 * 0.84 * 0.74 * EPI
-0.02 -0.05 0.06 0.05 -0.04 0.28 * language count


Pairwise Correlations Plot with language count (2017)
MCI
0.47 * DSTRI
0.95 * 0.47 * 3i
-0.03 -0.05 0.1 language_count
0.4 0.6 0.8 1.0
0 2 4 6
20 40 60 80
30 50 70

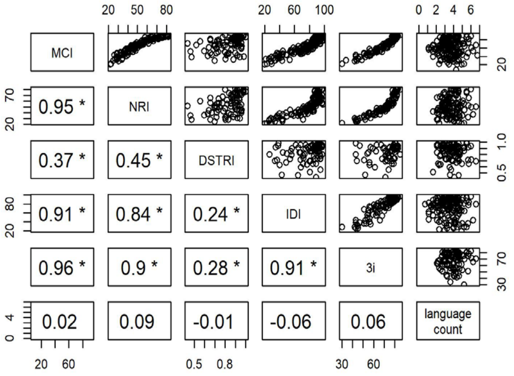

Pairwise Correlations Plot with language count (2018)
MCI
0.96 * 3i
0.42 * 0.45 * DSTRI
0.89 * 0.85 * 0.45 * DiGiX
0.96 * 0.93 * 0.43 * 0.91 * EGDI
0.81 * 0.81 * 0.41 * 0.75 * 0.88 * EPI
-0.02 0.01 -0.04 0.05 0.06 0.25 * language count


Pairwise Correlations Plot with language count (2019)
MCI
0.26 * NRI
0.96 * 0.23 * 3i
0.36 * 0.17 0.23 DSTRI
0.9 * 0.17 0.86 * 0.46 * DiGiX
-0.02 0.01 0 -0.04 0.05 language count

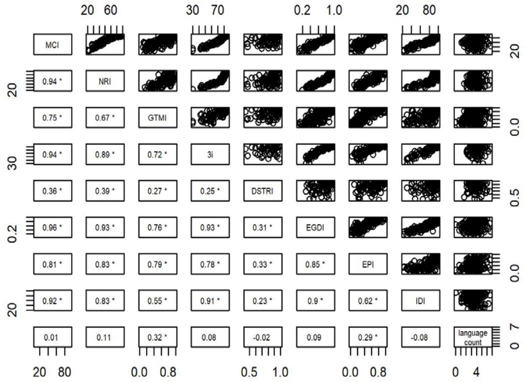

Pairwise Correlations Plot with language count (2020)
MCI
0.94 * NRI
0.83 * 0.82 * GTMI
0.97 * 0.9 * 0.83 * 3i
0.36 * 0.43 * 0.37 * 0.29 * DSTRI
0.96 * 0.93 * 0.84 * 0.94 * 0.32 * EGDI
0.89 * 0.94 * 0.73 * 0.89 * 0.41 * 0.88 * DiGiX
0.83 * 0.8 * 0.9 * 0.82 * 0.3 * 0.89 * 0.74 * EPI
0.01 0.04 0.36 * 0.03 -0.03 0.07 0.08 0.25 * language count

Appendix

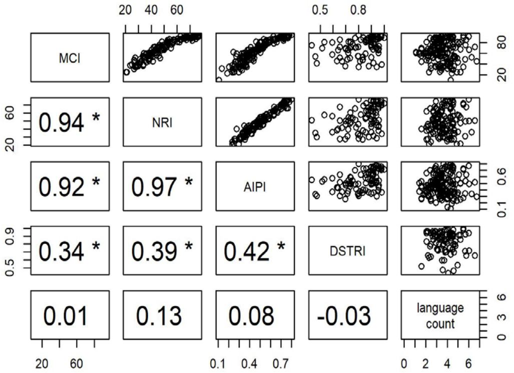
Pairwise Correlations Plot with language count (2021)



Pairwise Correlations Plot with language count (2022)



Pairwise Correlations Plot with language count (2023)



Correlation Summary

Year	Observations	Var 1	Var 2	Correlation	P_Value	Strength	Significance	Direction
2004	pairwise	EGDI	lang_count	0.202	0.005	Small	*	Positive
2004	pairwise	EPI	lang_count	0.324	< 0.001	Medium	*	Positive
2005	pairwise	EGDI	lang_count	0.181	0.012	Small	*	Positive
2005	pairwise	EPI	lang_count	0.342	< 0.001	Medium	*	Positive
2008	pairwise	EGDI	lang_count	0.174	0.016	Small	*	Positive
2008	pairwise	IDI	lang_count	-0.088	0.283	Very Small		Negative
2008	pairwise	EPI	lang_count	0.299	< 0.001	Small	*	Positive
2010	complete	EPI	lang_count	0.193	0.000	Small	*	Positive
2010	pairwise	EPI	lang_count	0.274	0.000	Small	*	Positive
2010	complete	EGDI	lang_count	-0.002	0.037	Very Small	*	Negative
2010	pairwise	EGDI	lang_count	0.151	0.037	Small	*	Positive

2010	complete	IDI	lang_count	-0.079	0.237	Very Small		Negative
2010	pairwise	IDI	lang_count	-0.097	0.237	Very Small		Negative
2012	pairwise	EPI	lang_count	0.253	0.000	Small	*	Positive
2012	pairwise	IDI	lang_count	-0.126	0.105	Small		Negative
2012	pairwise	EGDI	lang_count	0.001	0.988	Very Small		Positive
2014	complete	IDI	lang_count	-0.080	0.291	Very Small		Negative
2014	pairwise	IDI	lang_count	-0.080	0.291	Very Small		Negative
2014	complete	DSTRI	lang_count	-0.090	0.395	Very Small		Negative
2014	pairwise	DSTRI	lang_count	-0.091	0.395	Very Small		Negative
2014	complete	EPI	lang_count	0.102	0.399	Small		Positive
2014	pairwise	EPI	lang_count	0.061	0.399	Very Small		Positive
2014	complete	DAI	lang_count	-0.102	0.470	Small		Negative
2014	pairwise	DAI	lang_count	-0.054	0.470	Very Small		Negative
2014	complete	EGDI	lang_count	-0.001	0.523	Very Small		Negative
2014	pairwise	EGDI	lang_count	-0.046	0.523	Very Small		Negative
2014	complete	MCI	lang_count	-0.015	0.597	Very Small		Negative
2014	pairwise	MCI	lang_count	-0.040	0.597	Very Small		Negative
2015	pairwise	IDI	lang_count	-0.049	0.517	Very Small		Negative
2015	pairwise	DSTRI	lang_count	-0.039	0.713	Very Small		Negative
2015	pairwise	MCI	lang_count	-0.022	0.776	Very Small		Negative
2016	pairwise	EPI	lang_count	0.275	0.000	Small	*	Positive
2016	pairwise	EGDI	lang_count	0.061	0.402	Very Small		Positive
2016	pairwise	DAI	lang_count	0.046	0.542	Very Small		Positive
2016	pairwise	IDI	lang_count	-0.044	0.561	Very Small		Negative

2016	pairwise	DSTRI	lang_count	-0.052	0.627	Very Small		Negative
2016	pairwise	MCI	lang_count	-0.022	0.770	Very Small		Negative
2017	complete	3i	lang_count	0.154	0.418	Small		Positive
2017	pairwise	3i	lang_count	0.098	0.418	Very Small		Positive
2017	complete	DSTRI	lang_count	0.074	0.648	Very Small		Positive
2017	pairwise	DSTRI	lang_count	-0.049	0.648	Very Small		Negative
2017	complete	MCI	lang_count	0.157	0.718	Small		Positive
2017	pairwise	MCI	lang_count	-0.028	0.718	Very Small		Negative
2018	complete	EPI	lang_count	0.076	0.000	Very Small	*	Positive
2018	pairwise	EPI	lang_count	0.254	0.000	Small	*	Positive
2018	complete	EGDI	lang_count	-0.121	0.432	Small		Negative
2018	pairwise	EGDI	lang_count	0.057	0.432	Very Small		Positive
2018	complete	DiGiX	lang_count	0.000	0.595	Very Small		Positive
2018	pairwise	DiGiX	lang_count	0.054	0.595	Very Small		Positive
2018	complete	DSTRI	lang_count	-0.102	0.711	Small		Negative
2018	pairwise	DSTRI	lang_count	-0.040	0.711	Very Small		Negative
2018	complete	MCI	lang_count	-0.075	0.811	Very Small		Negative
2018	pairwise	MCI	lang_count	-0.018	0.811	Very Small		Negative
2018	complete	3i	lang_count	-0.142	0.931	Small		Negative
2018	pairwise	3i	lang_count	0.010	0.931	Very Small		Positive
2019	complete	DiGiX	lang_count	0.063	0.611	Very Small		Positive
2019	pairwise	DiGiX	lang_count	0.052	0.611	Very Small		Positive
2019	complete	DSTRI	lang_count	-0.171	0.739	Small		Negative
2019	pairwise	DSTRI	lang_count	-0.036	0.739	Very Small		Negative
2019	complete	MCI	lang_count	-0.008	0.772	Very Small		Negative

2019	pairwise	MCI	lang_count	-0.022	0.772	Very Small		Negative
2019	complete	NRI	lang_count	0.081	0.891	Very Small		Positive
2019	pairwise	NRI	lang_count	0.013	0.891	Very Small		Positive
2019	complete	3i	lang_count	-0.065	0.991	Very Small		Negative
2019	pairwise	3i	lang_count	0.001	0.991	Very Small		Positive
2020	complete	GTMI	lang_count	0.137	< 0.001	Small	*	Positive
2020	pairwise	GTMI	lang_count	0.357	< 0.001	Medium	*	Positive
2020	complete	EPI	lang_count	0.170	0.000	Small	*	Positive
2020	pairwise	EPI	lang_count	0.252	0.000	Small	*	Positive
2020	complete	EGDI	lang_count	0.031	0.323	Very Small		Positive
2020	pairwise	EGDI	lang_count	0.072	0.323	Very Small		Positive
2020	complete	DiGiX	lang_count	0.167	0.434	Small		Positive
2020	pairwise	DiGiX	lang_count	0.079	0.434	Very Small		Positive
2020	complete	NRI	lang_count	0.060	0.685	Very Small		Positive
2020	pairwise	NRI	lang_count	0.035	0.685	Very Small		Positive
2020	complete	3i	lang_count	0.055	0.769	Very Small		Positive
2020	pairwise	3i	lang_count	0.030	0.769	Very Small		Positive
2020	complete	DSTRI	lang_count	-0.125	0.771	Small		Negative
2020	pairwise	DSTRI	lang_count	-0.031	0.771	Very Small		Negative
2020	complete	MCI	lang_count	0.082	0.935	Very Small		Positive
2020	pairwise	MCI	lang_count	0.006	0.935	Very Small		Positive
2021	complete	NRI	lang_count	0.169	0.308	Small		Positive
2021	pairwise	NRI	lang_count	0.090	0.308	Very Small		Positive
2021	complete	IDI	lang_count	0.117	0.419	Small		Positive
2021	pairwise	IDI	lang_count	-0.063	0.419	Very Small		Negative

2021	complete	3i	lang_count	0.158	0.565	Small		Positive
2021	pairwise	3i	lang_count	0.058	0.565	Very Small		Positive
2021	complete	MCI	lang_count	0.168	0.823	Small		Positive
2021	pairwise	MCI	lang_count	0.017	0.823	Very Small		Positive
2021	complete	DSTRI	lang_count	-0.017	0.916	Very Small		Negative
2021	pairwise	DSTRI	lang_count	-0.011	0.916	Very Small		Negative
2022	complete	GTMI	lang_count	0.059	< 0.001	Very Small	*	Positive
2022	pairwise	GTMI	lang_count	0.325	< 0.001	Medium	*	Positive
2022	complete	EPI	lang_count	0.274	< 0.001	Small	*	Positive
2022	pairwise	EPI	lang_count	0.294	< 0.001	Small	*	Positive
2022	complete	NRI	lang_count	0.189	0.204	Small		Positive
2022	pairwise	NRI	lang_count	0.112	0.204	Small		Positive
2022	complete	EGDI	lang_count	0.139	0.240	Small		Positive
2022	pairwise	EGDI	lang_count	0.085	0.240	Very Small		Positive
2022	complete	IDI	lang_count	0.126	0.328	Small		Positive
2022	pairwise	IDI	lang_count	-0.076	0.328	Very Small		Negative
2022	complete	3i	lang_count	0.195	0.448	Small		Positive
2022	pairwise	3i	lang_count	0.077	0.448	Very Small		Positive
2022	complete	MCI	lang_count	0.155	0.857	Small		Positive
2022	pairwise	MCI	lang_count	0.014	0.857	Very Small		Positive
2022	complete	DSTRI	lang_count	-0.028	0.863	Very Small		Negative
2022	pairwise	DSTRI	lang_count	-0.018	0.863	Very Small		Negative
2023	complete	NRI	lang_count	0.214	0.125	Small		Positive
2023	pairwise	NRI	lang_count	0.133	0.125	Small		Positive
2023	complete	AIPI	lang_count	0.141	0.321	Small		Positive
2023	pairwise	AIPA	lang_count	0.076	0.321	Very Small		Positive
2023	complete	DSTRI	lang_count	-0.089	0.749	Very Small		Negative

2023	pairwise	DSTRI	lang_count	-0.034	0.749	Very Small		Negative
2023	complete	MCI	lang_count	0.121	0.897	Small		Positive
2023	pairwise	MCI	lang_count	0.010	0.897	Very Small		Positive

A.2.5 Cluster Analysis - Complete Observations

Complete Correlation Matrix

Year	Index	EGDI	EPI	IDI	MCI	DSTRI	DAI	3i	DiGiX	NRI	GTMI	AIPI
2004	EGDI	1.000	0.772									
2004	EPI	0.772	1.000									
2005	EGDI	1.000	0.771									
2005	EPI	0.771	1.000									
2008	IDI	0.945	0.585	1.000								
2008	EGDI	1.000	0.686	0.945								
2008	EPI	0.686	1.000	0.585								
2010	IDI	0.941	0.689	1.000								
2010	EGDI	1.000	0.802	0.941								
2010	EPI	0.802	1.000	0.689								
2012	IDI	0.962	0.659	1.000								
2012	EGDI	1.000	0.726	0.962								
2012	EPI	0.726	1.000	0.659								
2014	MCI	0.952	0.718	0.960	1.000	0.539	0.941					
2014	IDI	0.963	0.699	1.000	0.960	0.556	0.955					
2014	DSTRI	0.501	0.394	0.556	0.539	1.000	0.529					
2014	EGDI	1.000	0.828	0.963	0.952	0.501	0.947					
2014	EPI	0.828	1.000	0.699	0.718	0.394	0.739					
2014	DAI	0.947	0.739	0.955	0.941	0.529	1.000					
2015	MCI			0.951	1.000	0.533						
2015	IDI			1.000	0.951	0.546						
2015	DSTRI			0.546	0.533	1.000						
2016	MCI	0.952	0.803	0.953	1.000	0.508	0.935					
2016	IDI	0.965	0.780	1.000	0.953	0.531	0.952					
2016	DSTRI	0.519	0.428	0.531	0.508	1.000	0.505					

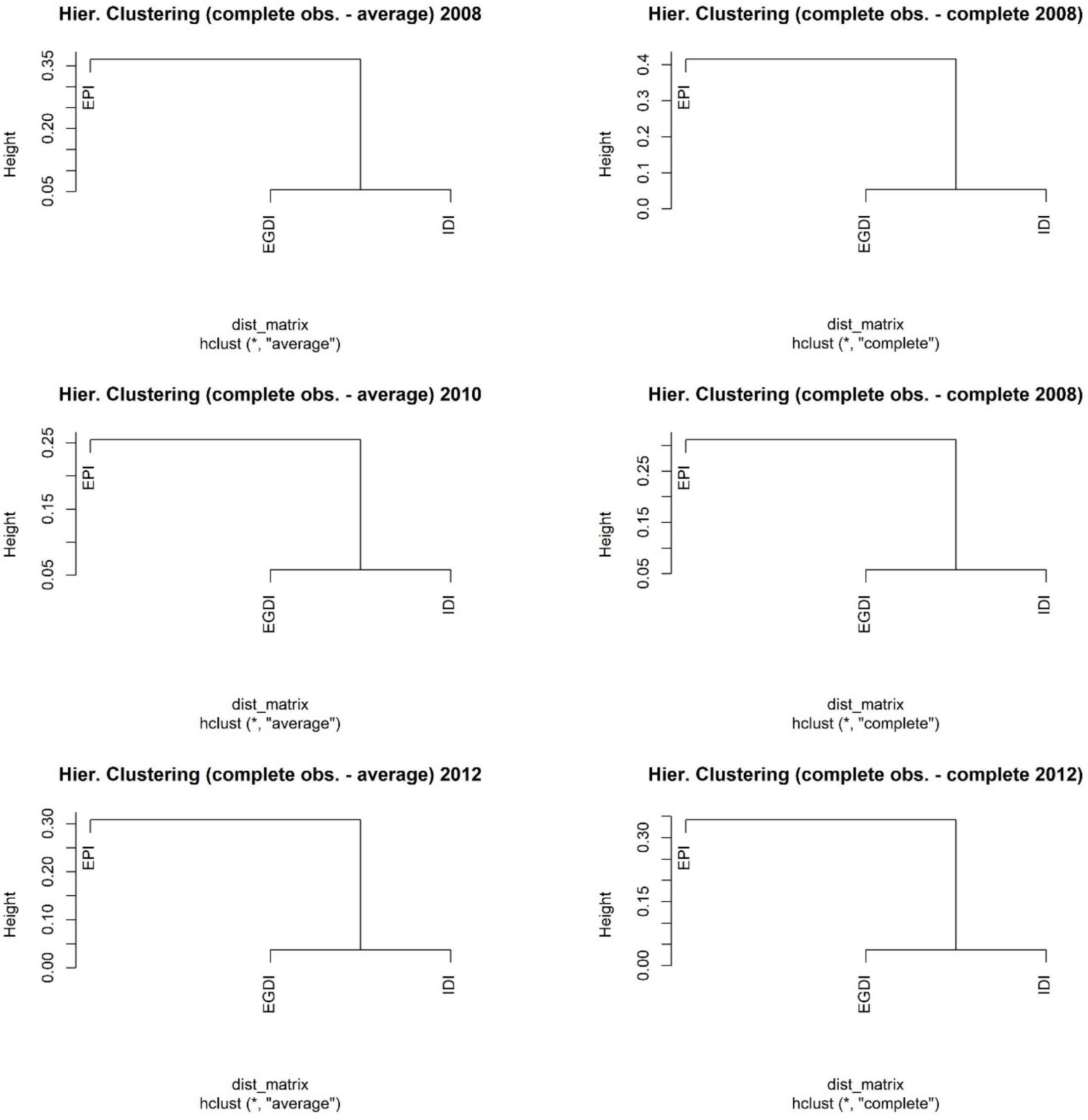
Appendix

2016	EGDI	1.000	0.891	0.965	0.952	0.519	0.950					
2016	EPI	0.891	1.000	0.780	0.803	0.428	0.818					
2016	DAI	0.950	0.818	0.952	0.935	0.505	1.000					
2017	MCI				1.000	0.474		0.972				
2017	3i				0.972	0.469		1.000				
2017	DSTRI				0.474	1.000		0.469				
2018	MCI	0.958	0.785		1.000	0.367		0.951	0.879			
2018	3i	0.946	0.823		0.951	0.328		1.000	0.832			
2018	DSTRI	0.387	0.325		0.367	1.000		0.328	0.436			
2018	EGDI	1.000	0.842		0.958	0.387		0.946	0.887			
2018	EPI	0.842	1.000		0.785	0.325		0.823	0.732			
2018	DiGiX	0.887	0.732		0.879	0.436		0.832	1.000			
2019	MCI				1.000	0.361		0.946	0.882	0.272		
2019	NRI				0.272	0.071		0.299	0.192	1.000		
2019	3i				0.946	0.302		1.000	0.844	0.299		
2019	DSTRI				0.361	1.000		0.302	0.399	0.071		
2019	DiGiX				0.882	0.399		0.844	1.000	0.192		
2020	MCI	0.954	0.834		1.000	0.474		0.963	0.900	0.934	0.784	
2020	NRI	0.915	0.767		0.934	0.505		0.889	0.952	1.000	0.755	
2020	GTMI	0.827	0.853		0.784	0.390		0.843	0.785	0.755	1.000	
2020	3i	0.951	0.877		0.963	0.393		1.000	0.894	0.889	0.843	
2020	DSTRI	0.353	0.268		0.474	1.000		0.393	0.420	0.505	0.390	
2020	EGDI	1.000	0.886		0.954	0.353		0.951	0.893	0.915	0.827	
2020	EPI	0.886	1.000		0.834	0.268		0.877	0.828	0.767	0.853	
2020	DiGiX	0.893	0.828		0.900	0.420		0.894	1.000	0.952	0.785	
2021	MCI			0.946	1.000	0.419		0.967		0.954		
2021	NRI			0.884	0.954	0.453		0.911		1.000		
2021	IDI			1.000	0.946	0.284		0.945		0.884		
2021	3i			0.945	0.967	0.351		1.000		0.911		
2021	DSTRI			0.284	0.419	1.000		0.351		0.453		
2022	MCI	0.966	0.823	0.947	1.000	0.351		0.958		0.946	0.669	
2022	NRI	0.926	0.836	0.875	0.946	0.374		0.889		1.000	0.578	
2022	IDI	0.940	0.783	1.000	0.947	0.220		0.944		0.875	0.657	
2022	GTMI	0.716	0.678	0.657	0.669	0.211		0.717		0.578	1.000	
2022	3i	0.953	0.835	0.944	0.958	0.257		1.000		0.889	0.717	

Appendix

2022	DSTRI	0.247	0.236	0.220	0.351	1.000		0.257		0.374	0.211	
2022	EGDI	1.000	0.885	0.940	0.966	0.247		0.953		0.926	0.716	
2022	EPI	0.885	1.000	0.783	0.823	0.236		0.835		0.836	0.678	
2023	MCI				1.000	0.352				0.936		0.934
2023	NRI				0.936	0.388				1.000		0.980
2023	DSTRI				0.352	1.000				0.388		0.430
2023	AIPI				0.934	0.430				0.980		1.000

Dendrograms (average vs. complete linkage)



Hier. Clustering (complete obs. - average) 2008

Hier. Clustering (complete obs. - complete 2008)

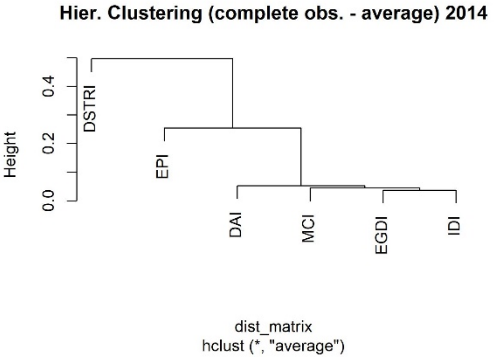
Hier. Clustering (complete obs. - average) 2010

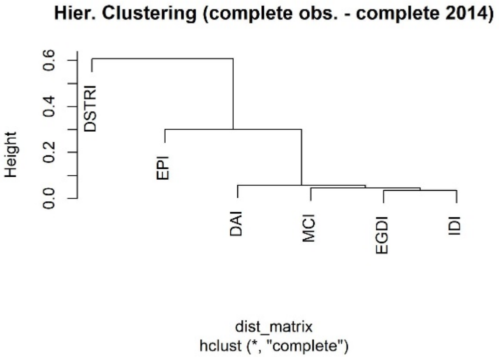
Hier. Clustering (complete obs. - complete 2008)

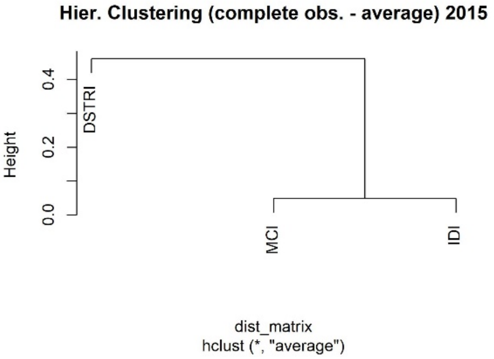
Hier. Clustering (complete obs. - average) 2012

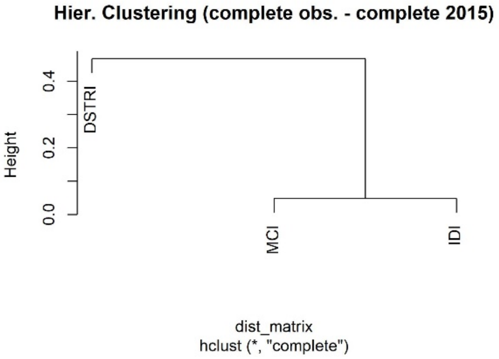
Hier. Clustering (complete obs. - complete 2012)

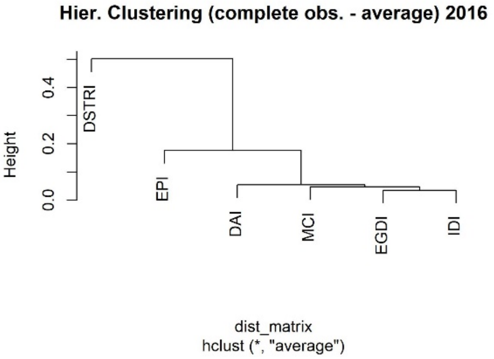
Appendix

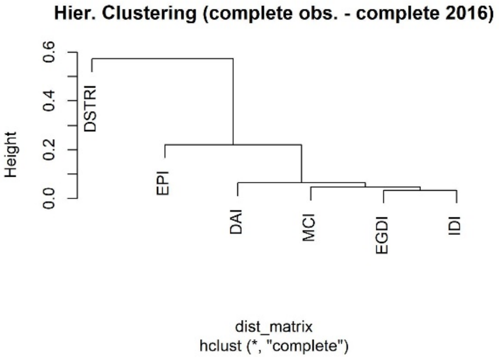

Hier. Clustering (complete obs. - average) 2014

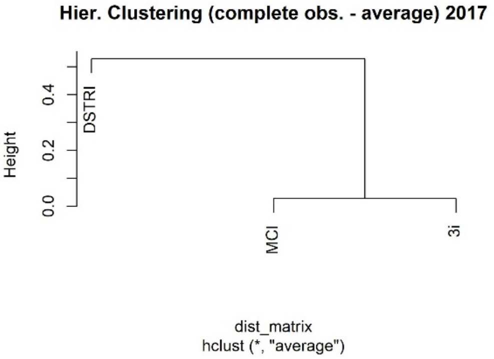

Hier. Clustering (complete obs. - complete 2014)

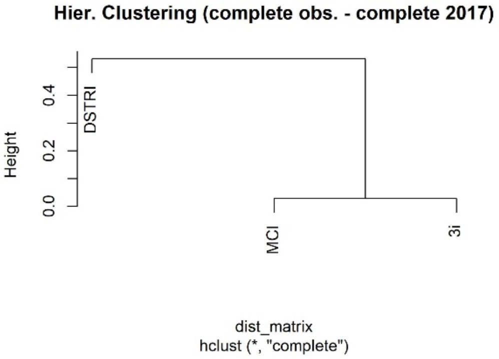

Hier. Clustering (complete obs. - average) 2015


Hier. Clustering (complete obs. - complete 2015)


Hier. Clustering (complete obs. - average) 2016

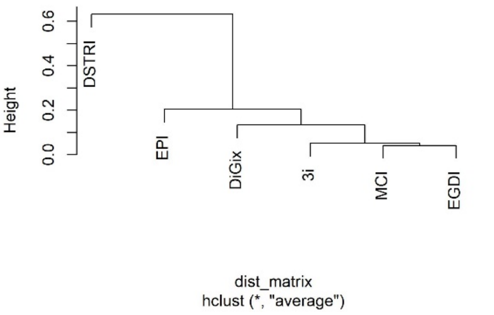

Hier. Clustering (complete obs. - complete 2016)


Hier. Clustering (complete obs. - average) 2017

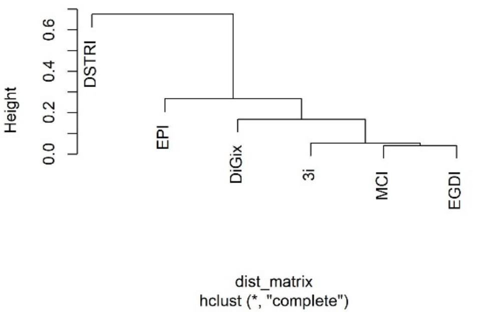

Hier. Clustering (complete obs. - complete 2017)

Appendix

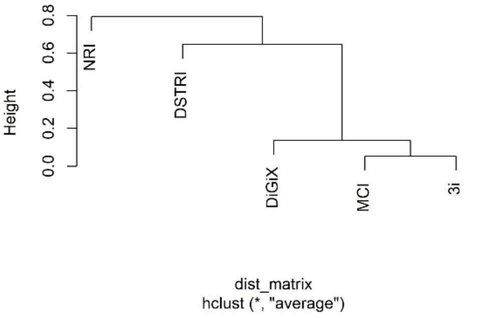
Hier. Clustering (complete obs. - average) 2018



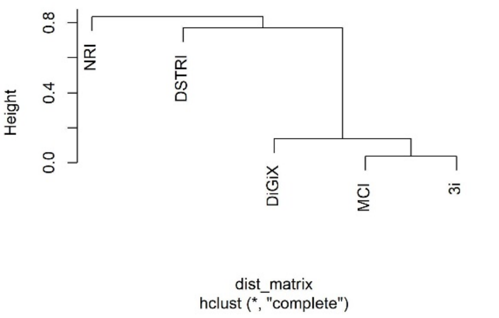
Hier. Clustering (complete obs. - complete 2018)



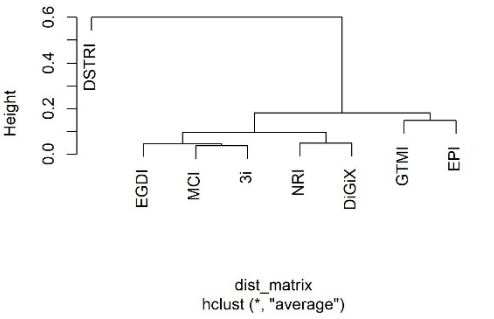
Hier. Clustering (complete obs. - average) 2019



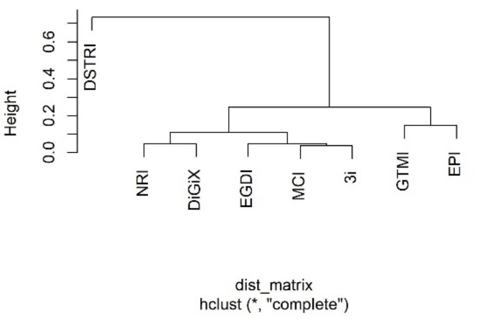
Hier. Clustering (complete obs. - complete 2019)



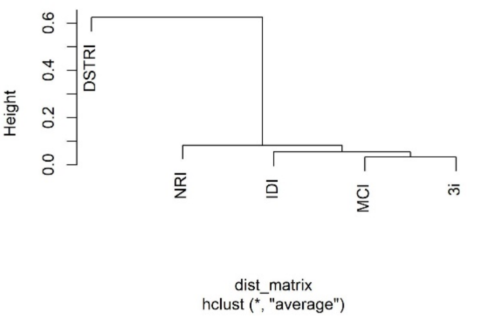
Hier. Clustering (complete obs. - average) 2020



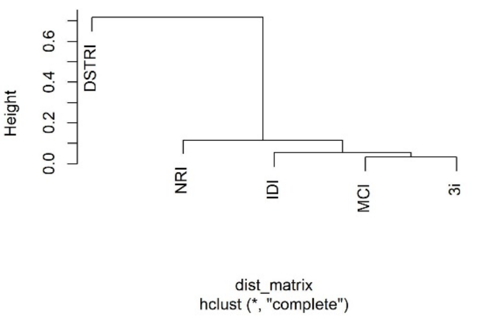
Hier. Clustering (complete obs. - complete 2020)

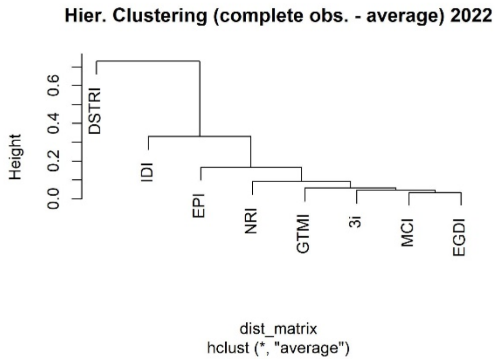


Hier. Clustering (complete obs. - average) 2021

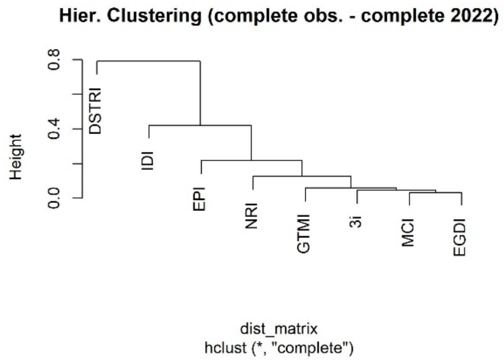


Hier. Clustering (complete obs. - complete 2021)

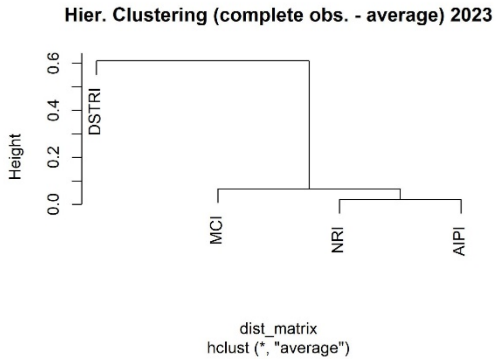




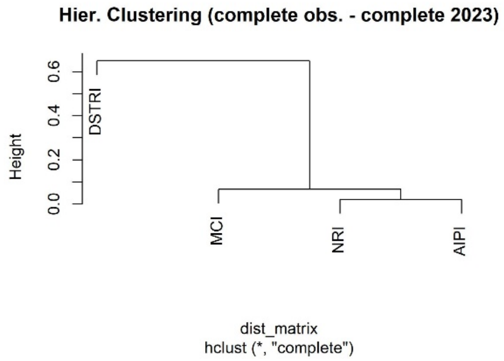
Hier. Clustering (complete obs. - average) 2022



Hier. Clustering (complete obs. - complete 2022)



Hier. Clustering (complete obs. - average) 2023



Hier. Clustering (complete obs. - complete 2023)

A.2.6 Cluster Analysis - Pairwise Complete Observations

Pairwise Complete Correlation Matrix

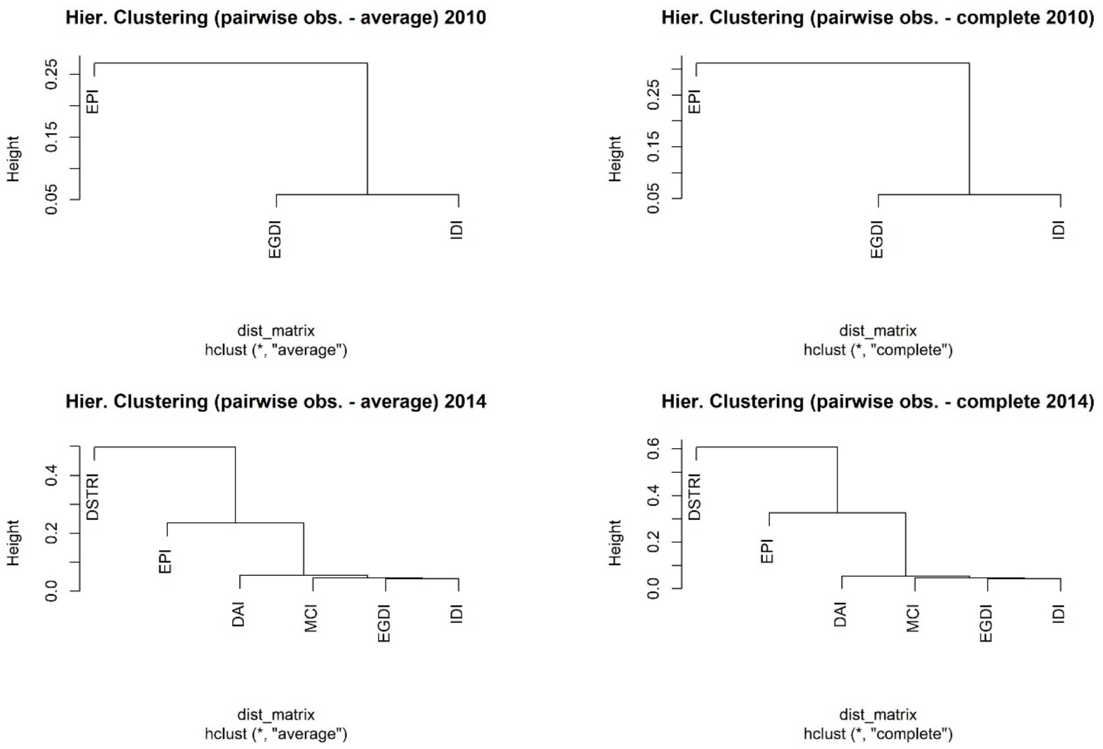
Year	Index	EGDI	EPI	IDI	MCI	DSTRI	DAI	3i	DiGiX	NRI	GTMI	AIPI
2004	EGDI	1.000	0.772									
2004	EPI	0.772	1.000									
2005	EGDI	1.000	0.771									
2005	EPI	0.771	1.000									
2008	IDI	0.945	0.585	1.000								
2008	EGDI	1.000	0.686	0.945								
2008	EPI	0.686	1.000	0.585								
2010	IDI	0.941	0.689	1.000								
2010	EGDI	1.000	0.776	0.941								
2010	EPI	0.776	1.000	0.689								
2012	IDI	0.962	0.659	1.000								
2012	EGDI	1.000	0.727	0.962								
2012	EPI	0.727	1.000	0.659								
2014	MCI	0.955	0.773	0.953	1.000	0.539	0.945					

2014	IDI	0.955	0.675	1.000	0.953	0.556	0.947					
2014	DSTRI	0.501	0.394	0.556	0.539	1.000	0.529					
2014	EGDI	1.000	0.828	0.955	0.955	0.501	0.944					
2014	EPI	0.828	1.000	0.675	0.773	0.394	0.784					
2014	DAI	0.944	0.784	0.947	0.945	0.529	1.000					
2015	MCI			0.945	1.000	0.537						
2015	IDI			1.000	0.945	0.546						
2015	DSTRI			0.546	0.537	1.000						
2016	MCI	0.952	0.825	0.946	1.000	0.513	0.945					
2016	IDI	0.950	0.741	1.000	0.946	0.531	0.943					
2016	DSTRI	0.524	0.435	0.531	0.513	1.000	0.510					
2016	EGDI	1.000	0.873	0.950	0.952	0.524	0.953					
2016	EPI	0.873	1.000	0.741	0.825	0.435	0.842					
2016	DAI	0.953	0.842	0.943	0.945	0.510	1.000					
2017	MCI				1.000	0.467		0.953				
2017	3i				0.953	0.469		1.000				
2017	DSTRI				0.467	1.000		0.469				
2018	MCI	0.956	0.811		1.000	0.422		0.959	0.893			
2018	3i	0.935	0.809		0.959	0.447		1.000	0.851			
2018	DSTRI	0.430	0.407		0.422	1.000		0.447	0.454			
2018	EGDI	1.000	0.883		0.956	0.430		0.935	0.909			
2018	EPI	0.883	1.000		0.811	0.407		0.809	0.750			
2018	DiGiX	0.909	0.750		0.893	0.454		0.851	1.000			
2019	MCI				1.000	0.355		0.963	0.901	0.264		
2019	NRI				0.264	0.166		0.226	0.166	1.000		
2019	3i				0.963	0.228		1.000	0.863	0.226		
2019	DSTRI				0.355	1.000		0.228	0.460	0.166		
2019	DiGiX				0.901	0.460		0.863	1.000	0.166		
2020	MCI	0.958	0.831		1.000	0.360		0.967	0.892	0.943	0.832	
2020	NRI	0.925	0.796		0.943	0.429		0.896	0.937	1.000	0.815	
2020	GTMI	0.845	0.904		0.832	0.367		0.827	0.731	0.815	1.000	
2020	3i	0.940	0.822		0.967	0.293		1.000	0.891	0.896	0.827	
2020	DSTRI	0.321	0.299		0.360	1.000		0.293	0.412	0.429	0.367	
2020	EGDI	1.000	0.887		0.958	0.321		0.940	0.881	0.925	0.845	
2020	EPI	0.887	1.000		0.831	0.299		0.822	0.737	0.796	0.904	

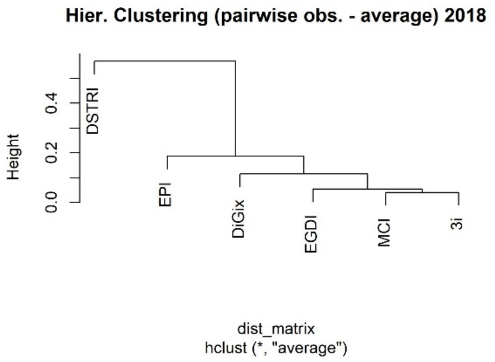
Appendix

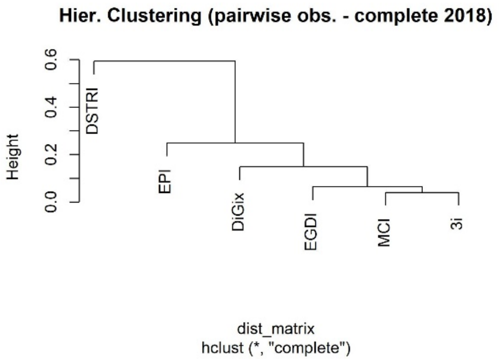
2020	DiGiX	0.881	0.737		0.892	0.412		0.891	1.000	0.937	0.731	
2021	MCI			0.913	1.000	0.367		0.955		0.945		
2021	NRI			0.841	0.945	0.452		0.903		1.000		
2021	IDI			1.000	0.913	0.235		0.908		0.841		
2021	3i			0.908	0.955	0.277		1.000		0.903		
2021	DSTRI			0.235	0.367	1.000		0.277		0.452		
2022	MCI	0.956	0.813	0.918	1.000	0.359		0.944		0.940	0.751	
2022	NRI	0.927	0.834	0.832	0.940	0.388		0.886		1.000	0.667	
2022	IDI	0.896	0.623	1.000	0.918	0.231		0.914		0.832	0.550	
2022	GTMI	0.762	0.794	0.550	0.751	0.267		0.724		0.667	1.000	
2022	3i	0.926	0.778	0.914	0.944	0.254		1.000		0.886	0.724	
2022	DSTRI	0.308	0.331	0.231	0.359	1.000		0.254		0.388	0.267	
2022	EGDI	1.000	0.846	0.896	0.956	0.308		0.926		0.927	0.762	
2022	EPI	0.846	1.000	0.623	0.813	0.331		0.778		0.834	0.794	
2023	MCI				1.000	0.342				0.936		0.923
2023	NRI				0.936	0.388				1.000		0.971
2023	DSTRI				0.342	1.000				0.388		0.420
2023	AIPI				0.923	0.420				0.971		1.000

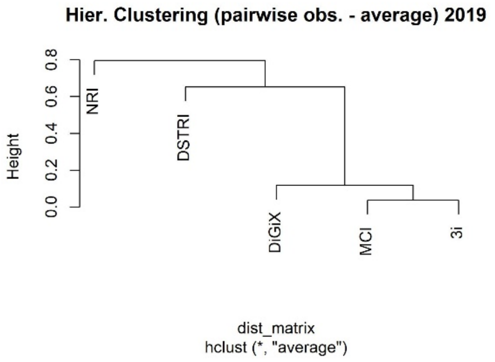
Dendrograms (average vs. complete linkage)

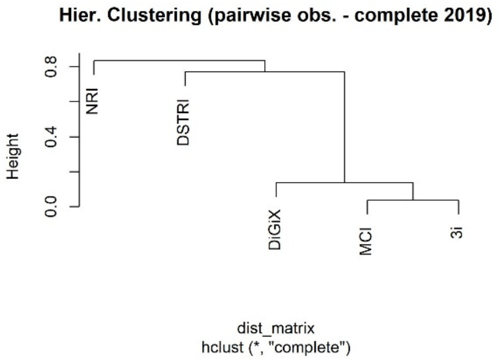


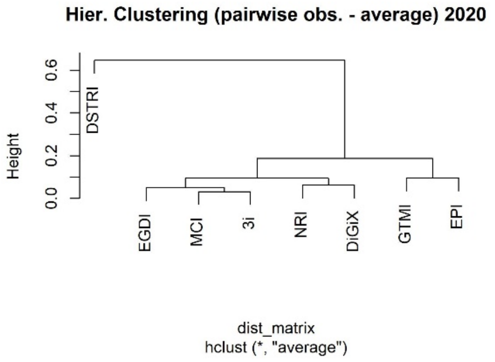
Appendix

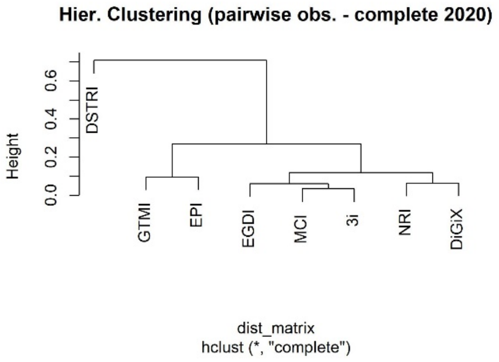

Hier. Clustering (pairwise obs. - average) 2018
Height
0.0
0.2
0.4
DSTRI
EPI
DiGix
EGDI
MCI
3i
dist_matrix
hclust (*, "average")

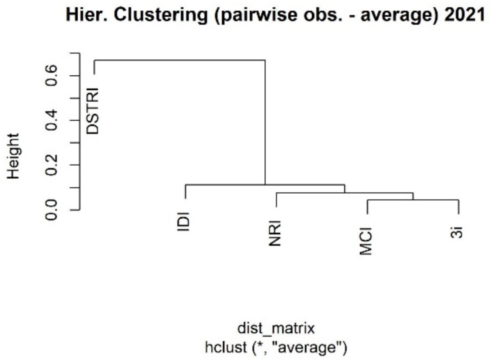

Hier. Clustering (pairwise obs. - complete 2018)
Height
0.0
0.2
0.4
0.6
DSTRI
EPI
DiGix
EGDI
MCI
3i
dist_matrix
hclust (*, "complete")

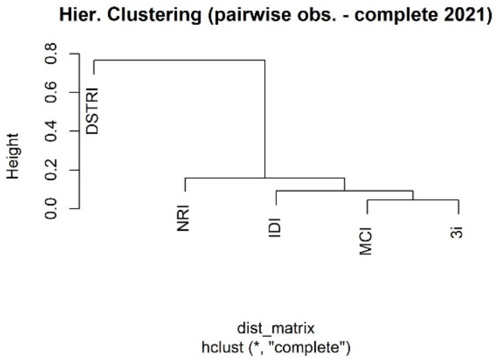

Hier. Clustering (pairwise obs. - average) 2019
Height
0.0
0.2
0.4
0.6
0.8
NRI
DSTRI
DiGiX
MCI
3i
dist_matrix
hclust (*, "average")


Hier. Clustering (pairwise obs. - complete 2019)
Height
0.0
0.4
0.8
NRI
DSTRI
DiGiX
MCI
3i
dist_matrix
hclust (*, "complete")

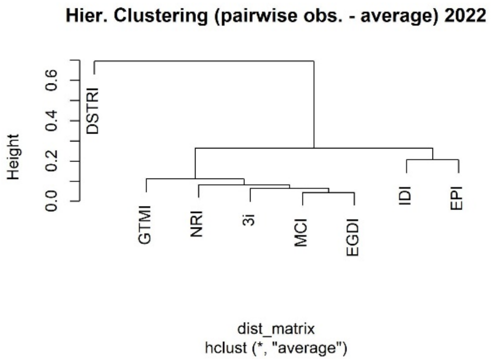

Hier. Clustering (pairwise obs. - average) 2020
Height
0.0
0.2
0.4
0.6
DSTRI
EGDI
MCI
3i
NRI
DiGiX
GTMI
EPI
dist_matrix
hclust (*, "average")

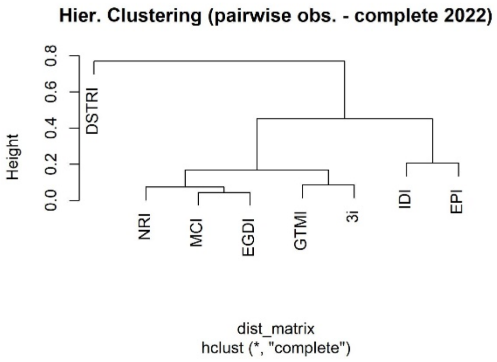

Hier. Clustering (pairwise obs. - complete 2020)
Height
0.0
0.2
0.4
0.6
DSTRI
GTMI
EPI
EGDI
MCI
3i
NRI
DiGiX
dist_matrix
hclust (*, "complete")


Hier. Clustering (pairwise obs. - average) 2021
Height
0.0
0.2
0.4
0.6
DSTRI
IDI
NRI
MCI
3i
dist_matrix
hclust (*, "average")


Hier. Clustering (pairwise obs. - complete 2021)
Height
0.0
0.2
0.4
0.6
0.8
DSTRI
NRI
IDI
MCI
3i
dist_matrix
hclust (*, "complete")

Appendix


Hier. Clustering (pairwise obs. - average) 2022
Height
0.0
0.2
0.4
0.6
DSTRI
GTMI
NRI
3i
MCI
EGDI
IDI
EPI
dist_matrix
hclust (*, "average")


Hier. Clustering (pairwise obs. - complete 2022)
Height
0.0
0.2
0.4
0.6
0.8
DSTRI
NRI
MCI
EGDI
GTMI
3i
IDI
EPI
dist_matrix
hclust (*, "complete")