An AI-driven Observatory Against Poverty

Scientific Report

Date: October 4, 2024

Prepared by:

J. Pita Costa [1,3], M. Grobelnik [1,3], Luka Urbanč [1], Jan Sturm [1,2], Matej Kovačič [1], Matej Senožetnik [1], John Shawe-Taylor [1], Dunja Mladenič [1,2,3]

[1] International Research Centre on Artificial Intelligence under the auspices of UNESCO (IRCAI) & Artificial Intelligence Lab at the Institute Jozef Stefan (JSI)

[2] Jožef Stefan International Postgraduate School

[3] Quintelligence d.o.o.

Corresponding author: Joao Pita Costa, joao.pitacosta@quintelligence.com

Keywords: Open Data; Artificial Intelligence; Poverty; SDG 1

Jožef Stefan Institute

IRC AI

International Research Centre of Artificial Inteligence under the auspices of UNESCO

future of life INSTITUTE

Executive Summary

The technological landscape is changing rapidly, allowing for a range of perspectives on the meaningful use of artificial intelligence to push the progress towards the 2030 UN Sustainable Development Goals. The appropriate use of heterogeneous data in the context of the digital transformation across a wide range of sectors is empowering utilities, local authorities and general citizens that can now leverage the benefits of a Data-driven Intelligence. In this paper we will present an AI-based SDG 1 observatory integrating open data sourced in news, policies, scientific research and statistical indicators that feeds the business intelligence of agencies and policy-makers, but also towards the education of populations. It allows for in-depth exploration of poverty-related events (e.g., floods or contamination) through interactive data visualization best fit for the data sets ingested.

Table of Contents

1. INTRODUCTION

The data-driven digital transformation across industries and evidence-based decision making by the leadership across companies and local authorities is gaining its momentum [17] [19] [20]. It is therefore essential to highlight the meaningfulness of those approaches and technologies. To leverage the awareness of the sector to the advantage of the insightful information provided by openly available data, complementing the specific knowledge of proprietary data collected in-house, the Internacional Research Centre under the auspices of UNESCO (IRCAI) is building an AI-based data-driven Observatory (accessible at sdg-observatory.ircai.org), based on well-established text mining and machine learning technologies most of which developed in the context of projects by IRCAI's host research institution, the Jozef Stefan Institute (JSI). This observatory allows the user to interact through the appropriate data visualizations with large bodies of knowledge collected in real-time (e.g., worldwide news and social media) or frequently updated (e.g., published research, patents and indicators) to impact their policies and actions.

1.1 Towards an AI-based SDG1 Observatory

The observatory presented in this paper and made available at https://sdg-observatory.ircai.org/index.php/sdg-1/, was designed to serve different stakeholders, and built with some of them to better address their different workflows and priorities. Its stakeholders are using the information generated by the observatory in the resolution of problems related to poverty events, to understand how their actions are perceived by the general public, and to explore successful scenarios in similar cases. On the other hand, policy-makers in local authorities can use this system to help better align to the SDG 1 and to other regional and national guidelines, in the context of evidence-based decision-making using open data, and evaluate commitments in time.

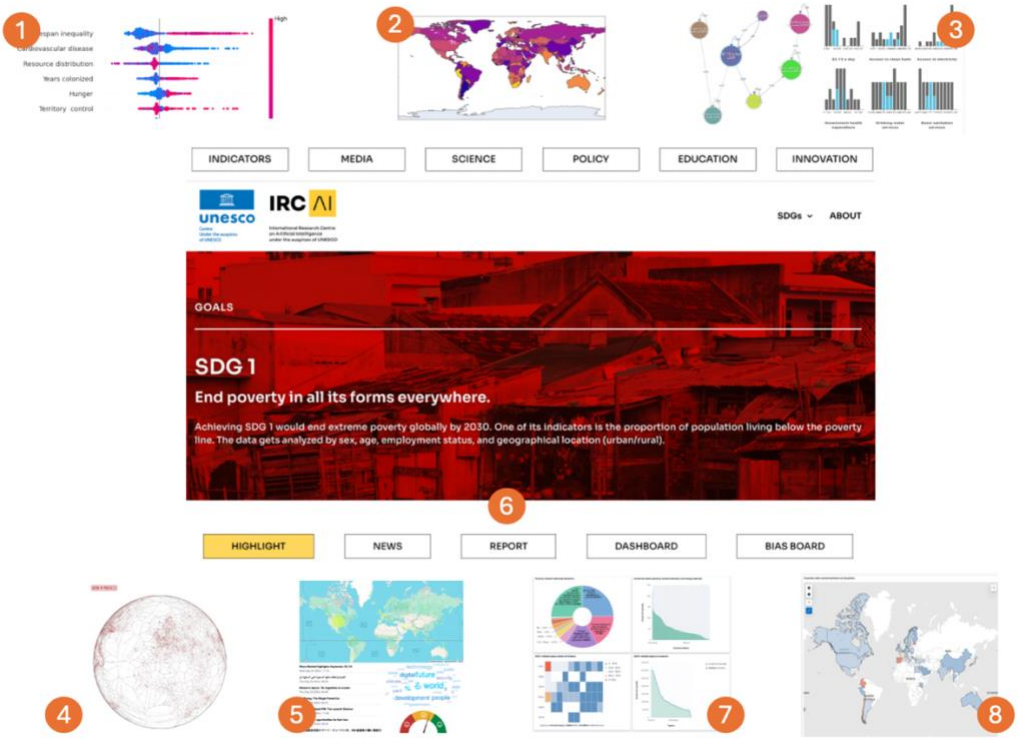

Figure 1 – Overall perspective onto the SDG 1 Observatory functionality including: the causaliy analysis visual model (1) and the transfer entropy maps (2); the Markov chain analysis (3); a data narrative (4) to be subject of a youth AI challenge; the live news stream including sentiment (5); the overall report on the impact of AI in SDG 1 (6) followed by a dashboard of visualisations; and the visual analysis of bias in data (8),

The Observatory (see Figure 1) provide benefits to a range of stakeholders including National Governments – providing access to a variety of perspectives (including trend and comparative) on a data driven dashboard with information on SDG 1 trends for decision-making; (i) *Educational Institutions* – to access to information on current trends on OER research and development; (ii) *Research Institutions* – providing access to data over interactive visualisation and research; (iii) *Civil Society* – allowing access to information and training materials that explore the knowledge available; (iv) *NGO community* – enabling access to information directly linked to the community priorities; and (v) *General Population* – empowering overall open education to access to local progress on SDG 1. The Observatory is being built by the IRCAI through the Institute Jozef Stefan Institute, in collaboration with several UNESCO Chairs.

As represented in Figure 1, the SD1 Observatory is built from the live news stream including sentiment (5); the overall report on the impact of AI in SDG 1 (6), and is followed by a dashboard of visualisations. The research results of this project further feed the observatory with the causaliy analysis visual model (1) and the transfer entropy maps (2), the Markov chain analysis (3) and the visual analysis of bias in data (8). As a complementary initiative, a data narrative (4) will be subject of a youth AI challenge focusing AI and Multidimentional Poverty.

The overall architecture of the SDG 1 Observatory is based on in-house technologies (StreamStory, SearchPoint and ObservAItion) developed in collaboration with the AI Department at the Jozef Stefan Institute, fed by the six data sources representing data collaborations between IRCAI and collaboration partners such as EventRegistry and VideoLectures, or Open Science communities such as “Our World in Data” (OWiD) and OpenAlex (see Figure 2).

This technology enables the exploration of causality between indicators, examining the impact they have on one another. We are also analyzing trends and seasonal behaviors affected by climate change, based on historical and forecasted resource availability, thus enhancing preparedness. Additionally, the SDG 1 Observatory can be used to explore digital resources related to inter-state relations and at a global level to monitor poverty-related news and key research topics. This can be valuable in the context of Education for Sustainability, both in the classroom, as a supplement to school curricula, and for citizens, by providing further information on SDG 1 events and stakeholder commitments.

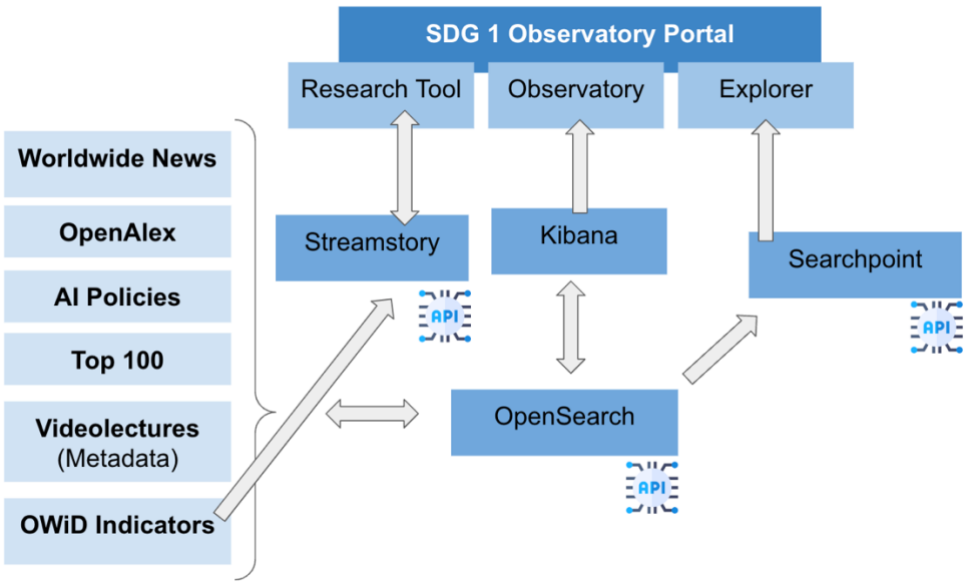

Figure 2 – Overall perspective of the SDG 1 Observatory architecture, including: data sources from collaborations and Open Science communities, AI technologies, APIs and points of interaction with the user.

1.2 Related Work

The state-of-the-art in data analysis for SDG 1, which focuses on eradicating poverty [2], has advanced significantly with the development of machine learning and big data analytics. SDG 1 indicators, such as poverty headcount ratios, access to social protection, and vulnerability to economic shocks, demand robust data collection and analysis methods. Recent innovations have utilized remote sensing data, social media, and mobile phone usage patterns to deliver real-time insights into poverty levels and resource distribution [6] [10] [12]. The integration of traditional survey methods with these novel data sources has significantly improved the granularity and timeliness of poverty assessments [1], especially in areas with limited data collection infrastructure [9].

SDG 1 observatories and platforms have been created as initiatives aimed at monitoring, analyzing, and reporting progress toward achieving SDG 1. These observatories generally integrate data from a broad range of sources, including government statistics, satellite imagery, mobile data, and crowdsourced information, to offer a comprehensive perspective on poverty dynamics. Examples include national poverty monitoring systems and international efforts such as the World Bank's Poverty and Inequality Platform, accessible through PovcalNet [18], and UNDP's SDG Monitoring Framework [14]. Some observatories also employ AI and machine learning to forecast poverty trends and track vulnerable populations, particularly in remote or underserved regions. These platforms provide valuable insights to policymakers, supporting data-driven decisions for poverty reduction by visualizing key indicators like income levels, access to basic services, and the effects of economic or environmental shocks. Through real-time analytics and interactive dashboards, SDG 1 observatories play an essential role in tracking progress, guiding policy interventions, and fostering international collaboration in the fight against poverty.

Numerous observatories have been established worldwide to monitor and track progress toward the Sustainable Development Goals (SDGs), including SDG 1 on poverty eradication. These observatories function as data platforms and analytical tools to collect, visualize, and disseminate information on critical SDG indicators. Some of the most prominent SDG observatories include:

1. World Bank's PovcalNet [18]: a specialized tool that focuses on poverty measurement and monitoring, allowing users to access and analyze global poverty data, offering insights into the World Bank's poverty indicators and tracking poverty trends across countries.

2. UN's SDG Global Observatory [14]: a data visualization and reporting platform launched by the United Nations to monitor the SDGs, offering interactive maps and graphs for tracking global progress. It covers various goals, including SDG 1, and integrates data from various sources to provide a comprehensive view of sustainable development indicators.

3. EU SDG Observatory (Eurostat) [4]: Eurostat provides an observatory for monitoring SDG progress within the European Union. It tracks indicators related to each of the SDGs, including those related to poverty, inequality, and social protection, offering insights into how European nations are progressing towards achieving these goals.

4. Africa SDG Observatory [15]: a platform, launched by the UN Economic Commission for Africa (UNECA), providing a continental overview of the progress being made in Africa toward achieving the SDGs, including SDG 1. It aggregates data from various African countries and offers visualization tools to monitor socio-economic trends and poverty-related indicators.

5. National and Regional SDG Observatories: Many countries and regions have also set up their own SDG observatories. These include Brazil's National SDG Observatory [7], the SDG Observatory of Catalonia [5], and several others, which focus on monitoring the SDG progress within specific countries or regions.

These observatories play a vital role in ensuring transparency, fostering accountability, and providing policymakers and stakeholders with critical data to design evidence-based interventions aimed at achieving the SDGs, particularly in the fight against poverty.

1.3 Summary of Results

Causality analysis in the context of SDG 1 has seen significant progress through the application of causal inference methods like Granger causality, structural equation modeling, and Bayesian networks. These approaches are used to uncover the underlying drivers of poverty, such as education, health access, or economic inequality, and understand how interventions may lead to meaningful changes in poverty reduction. For example, Granger causality can be used to determine whether improvements in education infrastructure lead to reductions in poverty over time, while Bayesian networks help model the probabilistic relationships between variables affecting poverty. By identifying causal links, policymakers can design more effective interventions and assess their long-term impact, ensuring that resources are allocated efficiently.

Another critical area of development is the integration of complex systems analysis and dynamic modeling approaches, such as agent-based models and Markov chains, to understand the feedback loops and transitions between states of poverty. These models allow for the simulation of various social, economic, and environmental factors that influence poverty, and how these factors interact over time. For example, models that incorporate employment rates, food security, and climate impacts can simulate how shocks in one area may propagate through the system and exacerbate poverty. These advanced analytical methods are empowering decision-makers to better predict the outcomes of policy actions and design resilient, multi-dimensional strategies to tackle poverty, aligning with the interconnected nature of the SDG 1 targets.

News mining, particularly in the context of modeling causality and predicting future events or trends, remains a relatively underexplored area. This is largely due to the inherent complexity of the task and the limited availability of comprehensive global media data. Current state-of-the-art research often focuses on localized media monitoring approaches that leverage AI technologies, but these solutions are generally not suitable for conducting research on a global scale. In our case, however, we obtained access to a global news media archive, which enabled us to conduct an extensive study on documents related to SDG1, targeting poverty reduction.

1.4 Potential Impact

The AI Observatory’s initiative on SDG 1 adopts a data-centric methodology to deeply understand AI’s role in alleviating poverty. By aggregating and analyzing data from various sources, we offer an alternative view of AI’s impacts with a dynamic approach similar to a digital twin of SDG 1’s complexities, enabling targeted strategies that can be further pursued through:

1. Root Cause Analysis: we will further investigate into SDG 1’s underlying factors, exploring how interconnected SDGs, such as food security (SDG 2), climate change (SDG 13), healthcare (SDG 3), and governance (SDG 16), influence poverty. This holistic view aids in crafting strategies addressing poverty’s root causes.

2. Advanced Predictive Modeling: Employing machine learning on diverse datasets predicts long-term trends, helping stakeholders foresee future scenarios and guide planning and resource allocation for poverty reduction.
3. Dynamic “What-If” Scenario Analysis: Simulation techniques provide a virtual platform for testing interventions, offering insights into their potential impacts and aiding in selecting the most effective poverty reduction strategies.
4. Targeted Phenomena Detection: Analyzing extensive datasets identifies key questions and insights, focusing efforts on the most impactful poverty reduction areas.

This strategic framework is continually updated with data from news articles, scientific research, statistical indicators, and policies, processed through advanced AI tools for sentiment analysis, trend identification, and data mining. Partnerships with platforms like Event Registry and OpenAlex enrich our analysis, while data from the World Bank, UNDP, and policy databases ensure a comprehensive macroeconomic and social context. Our cloud-based tools guarantee data integrity and compliance with data protection regulations. This approach not only highlights AI’s current achievements but also identifies future opportunities and challenges.

The project employed a suite of analytical methods to project how AI could further accelerate the achievement of SDG 1, building upon our evaluation of AI’s current impact. This forward-looking analysis integrates predictive modeling, scenario analysis, and policy impact simulations to offer insights into future trajectories of AI’s role in poverty alleviation.

- Predictive Modelling: we build on our innovation in predictive models using machine learning algorithms to forecast the potential impact of ongoing and future AI interventions on poverty indicators. Techniques like time series analysis and regression models will be key in understanding the dynamics between AI deployment and poverty reduction efforts.
- Scenario Analysis: conducted to examine a range of possible futures based on varying levels of AI integration in poverty alleviation strategies. We will construct scenarios that include optimistic, pessimistic, and status quo trajectories of AI development and deployment. This method will help in identifying robust strategies that are resilient across different future landscapes, highlighting pathways where AI could have the most significant positive impact or, conversely, where it might pose challenges to achieving SDG 1
- Policy Impact Simulations: to understand the potential impact of policy decisions on the effectiveness of AI in reducing poverty, we will use what if scenarios to simulate different policy environments, and identify policies that maximize the positive effects of AI while mitigating potential adverse outcomes.

By employing these analytical methods, the project aims to create a comprehensive and nuanced projection of AI’s future impact on achieving SDG 1. This forward-looking analysis will not only highlight opportunities for leveraging AI in the fight against poverty but also identify potential pitfalls and challenges, guiding strategic decision-making to optimize AI’s contribution to this crucial global goal.

2. MATERIALS AND METHODS

Aiming to generate intelligence from open data, IRCAI’s SDG 1 Observatory discussed in this paper contributes to a holistic ecosystem for digitisation at multinational level. It has four views representing the insight that can be obtained from the exploration of indicators, media, research and the observation of natural resources. It is putting together a big diversity of (mostly open) data, at a global and a local granularity, getting ingested into the system with different frequency, and powered by the most recent text mining, machine learning and data visualization methods. The amount and heterogeneity of data generated allied to the rapid progress of scientific research and consequent technological development

allow for a new reality in regards to resource management and sustainability. While the frequency of the data ingested and the dynamic of what is being observed can dictate the different pace that "real-time" observation can take, the usefulness of the insightful information digested from that data in the context of the realistic needs within the workflow of professionals will dictate its meaningfulness.

2.1 Indicator Importance, Correlation and Causality

In order to better understand the significant features that can impact poverty using machine learning methods applied to indicators, we chose the indicator "*Share of population living below national poverty lines"* provided by the World Bank Poverty and Inequality Platform's 2024 update[1]. This starting point is one of the few commonly agreed indicators to represent a Poverty baseline at a multinational level [11]. The data can represent either income after taxes and benefits or per-person consumption, depending on the context. Household income is evenly distributed among all members, including children [12]. Additionally, non-market sources of income, like food produced by subsistence farmers for their own consumption, are included in the calculations.

From this baseline we move forward to build a well-structured dataset, aligning global indicators for each country and year. To allow for comparative analysis, we ensure that all the countries have all the indicators complete. We use data available at "Our World in Data"[2], including a variety of different sources from the World Bank to UN and WHO, and we added related data from IHME and FAO as well. We ensure it's clean, has no missing values (employing interpolation and extrapolation methodologies, as later discussed). No normalization is done as random forest isn't sensitive to the scale of data. We start from all of the 558 indicators (that from now on we refer as features whenever appropriate), excluding the poverty baseline and income group, across 224 different countries represented. After the appropriate cleaning and preprocessing of the dataset obtained, the authors performed the analysis of correlation and used k-means clustering to identify what could be repeating signals provided by indicators with similar behaviour. This workflow is partially discussed in [16].

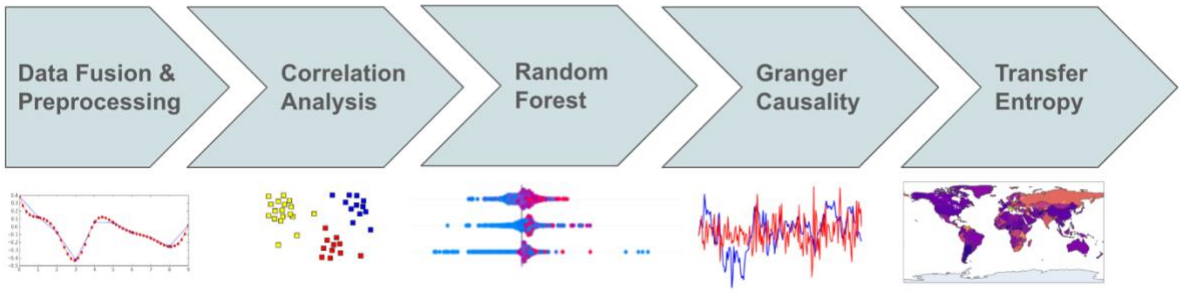

Figure 3 – Methodology to explore causal relations between indicators towards understanding SDG 1.

Then we have applied a random forest model, taking into consideration that these have built in feature selection. Moreover, we identify the most important features for the model (accuracy in regards to number of features selected). We have identified that 25 features are enough, as more do not contribute drastically (as will later be explained in more detail). The SHAP approach is then applied to capture the feature importance, running on the 25 features module. Furhtermore, the alternative approach of Transfer Entropy in both directions and sum score (in the basis of being feedback loops, i.e., bidirectional) will offer us map visualisations representing that information transfer between pairs of features per country (see Figure 3 for the full methodology).

[1] See https://ourworldindata.org/grapher/share-of-population-living-in-poverty-by-national-poverty-lines
[2] See https://ourworldindata.org/sdgs

The data is organized on a time window from 1987 to 2023, with each indicator collected at an annual frequency. Given the incompleteness of the original data ingested into the system, to complete the missing values we perform as follows:

- **Linear interpolation**: estimating the unknown values between known data points, assuming that the data between the points follows a pattern. By using mathematical techniques like, e.g., linear or polynomial functions, it fills in the gaps to predict intermediate values.
- **Ffill**: we use the last know value to fill missing entries, although we lose variance, we prevent the use of impossible values
- **Bfill**: we also use the first known value to fill entries before the first measurement

In substitution to Bfill and Ffill, we have also used polinomial extrapolation, noticing that the order does not make much impact in 1, 2 and 3 degrees.

Identifying correlations in data is an essential step in understanding the relationships between variables and determining potential redundancies. When variables are highly correlated, they tend to provide overlapping information, which can inflate the dimensionality of the data without adding substantial value. To address this, we have employed both Pearson correlation to obtain the **Pearson Correlation Coefficient (PCC)** measuring the linear correlation between the indicators present in the dataset. Moreover, **K-means clustering** was employed as a dimensionality reduction technique aiming at identifying and distinguishing clusters of similar data points. By grouping data into clusters based on their similarity, K-means helps to represent the data using fewer features, capturing the most significant patterns. After clustering, each cluster can be represented by its centroid, effectively reducing the number of dimensions while preserving the core structure of the data. This combination of correlation analysis and K-means clustering would allow for more efficient data processing, particularly in high-dimensional datasets, by focusing on the most informative and diverse aspects of the data. Using standard K-means algorithms yielded poor results, characterized by significant overlap between clusters and a high number of required clusters. As an alternative, we attempted to group the features based on their correlations. The goal was to form clusters that would ideally be homogeneous, with all features related to the same area. For instance, a cluster containing healthcare spending, death rates from various illnesses, and vaccination data could be used to describe healthcare in a country.

However, this approach also resulted in suboptimal outcomes, as the clusters were rarely homogeneous. Many clusters included data that lacked logical coherence—for example, the number of banks per 100,000 people coule be grouped with healthcare data. Another drawback of clustering based on correlation coefficients was the lack of transitivity. When clustering features with a correlation coefficient of 0.9 or higher, issues arose. For example, if feature A was correlated with feature B at 0.91, and B was correlated with feature C at 0.92, it was possible for feature C to be correlated with feature A at only 0.89. This forced us to drop one of the features, even though all three showed high correlations. Dropping a feature in this manner broke the intended clustering logic, further complicating the process of dimensionality reduction.

Furthermore, **Principal Component Analysis (PCA)** is a widely used dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving as much of the original variance as possible. PCA achieves this by identifying the principal components, which are new, uncorrelated variables constructed as linear combinations of the original variables. These principal components capture the directions in which the data varies the most. The first principal component accounts for the largest variance in the data, followed by the second, orthogonal to the first, which captures the next largest variance, and so on. By reducing the dimensionality, PCA simplifies data analysis and visualization, helps in removing noise, and mitigates the risk of overfitting, while still retaining the most critical information. This technique is particularly effective when dealing with datasets where variables are correlated, as it focuses on maximizing variance and creating compact, informative representations of the data. Despite this, the principal components are always a linear combination of the original features,

meaning all features are present in each dimension created by PCA. This effectively translates the problem of high-dimensional data into the challenge of evaluating and interpreting these linear combinations, where each coefficient plays a significant role. Even if we reduce the data to 25 dimensions, it would still require a thorough analysis of the resulting combinations to extract meaningful information, which presents its own challenges.

The **Random Forest model** is a versatile ensemble machine learning method, widely used for both classification and regression tasks. It operates by constructing a large number of decision trees during training and averaging their predictions (in the case of regression) or selecting the most common prediction (in classification). One of its key strengths lies in its built-in feature selection capability. As Random Forests grow multiple decision trees based on different subsets of data and features, they can measure the importance of each feature by evaluating how much it contributes to reducing impurity (e.g., Gini impurity or entropy) across all trees. This allowed the model to rank features (i.e., the indicators) based on their importance, providing insight into which variables are the most predictive for Poverty as defined by the chosen Baseline, while also naturally reducing the risk of overfitting. The result was a robust, high-performing model that can efficiently handle large datasets, such as that of the 558 ingested features, and irrelevant or redundant features.

SHAP (Shapley Additive exPlanations) is a powerful tool for interpreting machine learning models by providing insights into the contribution of each feature to the model's predictions. When applied to a subset of features selected by a Random Forest model, SHAP offered us a more granular, individualized understanding of feature importance. While Random Forests provide an overall ranking of feature significance, SHAP values explain the impact of each feature on specific predictions, ensuring transparency at the local (instance) level. By combining SHAP with the features already identified by Random Forest, we could refine the interpretability process, focusing on the most relevant features and understanding their exact influence on individual outputs. This method is particularly useful in complex scenarios such as ours, where understanding the interaction and importance of features on specific cases is critical, enabling more informed decision-making based on model behavior.

Granger Causality is a statistical hypothesis test used to determine whether one time series can predict another, implying a directional causal relationship between them. The concept is based on the idea that if the inclusion of past values of one time series (X) significantly improves the prediction of the future values of another time series (Y), then X is said to "Granger cause" Y. Note that Granger causality does not imply *true causality* but rather, it identifies a temporal predictive relationship. The method involves fitting autoregressive models to both time series and comparing the performance of models that include and exclude the potential causal variable. If the model including X's past values predicts Y significantly better than a model with just Y's past values, then X is considered to have a Granger causal effect on Y. We used the python library *statsmodels* that provides an implementation of the Granger Causality test. With applying Granger causality we aimed to detect directional interactions between variables. Though, the requirement of data to be stationary was limiting in this particular case, and it could not be solved for a good ammount of the indicators ingested into the dataset.

Generalising Granger Causality, **Transfer entropy** is a measure used to quantify the flow of information between two time series or processes, capturing the directional dependence between them. It was used to calculate how much the future state of one process is predicted by the past states of another, effectively measuring information transfer in a given direction. When applied in both directions, transfer entropy can reveal feedback loops by showing how information flows back and forth between the two systems. This is particularly useful to this study as, if significant information transfer is detected in both directions, this indicates a bidirectional relationship, characteristic of feedback loops. By summing the transfer entropy scores in both directions, one can quantify the total information exchanged within the system, offering a

comprehensive view of the dynamic interdependence between the processes. This approach is particularly useful in complex systems such as this Poverty analysis approach over indicators, where feedback loops play a crucial role in the system's behavior and stability.

Other approaches were also considered, including SEM, PC Algorithm, LIGNAM, and FCI. **SEM (Structural Equation Modeling)** is a statistical technique used to analyze relationships between multiple variables, both observed and latent. It combines factor analysis and multiple regression, enabling complex models that involve direct and indirect relationships. SEM is typically employed to confirm causal hypotheses, but it cannot uncover new relationships and assumes linear relationships, making it unsuitable for our study.

The **PC algorithm** is a constraint-based method for causal discovery, used to identify the structure of a directed acyclic graph (DAG) from data. It tests conditional independencies between variables, gradually eliminates edges, and orients them based on specific rules to create a causal graph. However, this approach proved ineffective in our case due to its requirement for noise-free data, poor performance with complex relationships, and its assumption of no hidden (latent) variables.

The **FCI (Fast Causal Inference)** algorithm extends the PC algorithm to account for latent confounders and selection bias. It identifies conditional independencies within the dataset and uses this information to construct a partial ancestral graph (PAG), which represents causal relationships while considering the possibility of hidden or unmeasured variables.

LIGNAM is a causal discovery algorithm designed to infer a set of DAGs that best represent the causal structure of a dataset. Unlike the PC algorithm, LIGNAM is specifically tailored to handle latent confounders, making it suitable for datasets containing hidden variables, and it produces models that describe how these hidden factors influence observed relationships.

Although all these algorithms have potential, they rely on certain assumptions about the data. Like Granger causality, they require stationary data, which was rare in our dataset. Additionally, they assume no feedback loops, which is not the case in our study. As a result, none of these methods produced meaningful results for our analysis.

2.2 Exploring a Markov Chain of States

Analyzing relationships between trends in several simultaneous time series is crucial for understanding how the different variables within SDG 1 interact and evolve over time within a complex system. This type of analysis typically involves identifying patterns, correlations, and dependencies between multiple time series to reveal how changes in one variable may influence or be influenced by others. Techniques such as cross-correlation, Granger causality, and co-integration are often used to detect interdependencies and causal relationships across time series. By examining these relationships, one can uncover synchronized trends, lead-lag effects, or shared underlying drivers that may affect the series simultaneously.

For this aim, we have collected the 12 indicators available at the SDG Tracker initiative of OwiD[3] and prepared that data having one indicator per column while the first two correspond to *Date* and *Country*. We have proceeded with this study in analyzing multiple SDG 1 variables like social protection benefit, direct economic loss, and disaster risk reduction that together can help identify seasonal trends and their joint impact on ecosystems. By capturing these interactions, we aim to help making more accurate

[3] See https://ourworldindata.org/sdgs/no-poverty

predictions, exploring model complex behaviours, and gain deeper insights into the dynamics of systems where multiple processes occur concurrently.

Markov Chains are a powerful tool for time series analysis, particularly useful in modeling systems where the future state depends only on the present state and not on the sequence of events that preceded it (the Markov property). In time series analysis, a Markov Chain helps to capture the probabilistic transitions between different states over time. Each state represents a possible condition of the system, and the chain describes how the system moves from one state to another with certain probabilities. This makes Markov Chains well-suited for analyzing processes with inherent randomness, such as stock market fluctuations, weather patterns, or ecological dynamics. By focusing on transition probabilities, Markov Chains simplify the complexity of time-dependent data, allowing for predictions and insights into the likelihood of moving between states. The simplicity and robustness of Markov Chains make them a valuable method for uncovering underlying patterns in sequential data, modeling dynamic systems, and predicting future behaviors based on past trends.

To analyze multiple time series of environmental parameters simultaneously, we use StreamStory system [13], which leverages Markov chain visualization (see Figure 4). Developed by the Jozef Stefan Institute and adapted to address SDG 1-related challenges, this interactive tool highlights the distinctions between different states and their transitions. It begins by representing the data as points in a multi-dimensional space, independent of time, and then partitions this space to identify distinct states. The tool models transitions between these states, where each state is represented as part of a continuous-time Markov chain, and creates a hierarchy of states and transitions.

This sophisticated data visualization method provides detailed information at each node and transition, helping users explore combinations of factors that drive dynamics, such as seasonal changes in natural resources. By selecting a specific state, users can view relevant statistics, offering a deeper insight into the meaning and significance of each state.

2.3 Monitoring Events in the News

To further the research on best practices from historical text data we have explored multilingual worldwide news, adding a real-time dimension to the awareness on SDG 1 topics. To this aim, we partner with the AI-based news engine Event Registry (*eventregistry.org*) [8] that can ingest more than 1 million global news daily in more than 100 languages. The news collection is based on the *Newsfeed.ijs.si* service providing a clean, continuous, real-time aggregated stream of semantically enriched news articles from RSS-enabled sites across the world. It is periodically crawling a list of RSS feeds and a subset of Google News and obtain links to news articles, downloads the articles, taking care not to overload any of the hosting servers, and parses each article.

These are processed using multilingual capabilities based on the *wikifier.ijs.si* technology available over API tha offers the automated identification of concepts and entities, i.e., the wikification of the provided text and sophisticated disambiguation methods based on state-of-the-art text mining. The location is provided by the direct mention of a location in the text of the news article or, if the latter is not available, the location of the news venue. We are able to achieve a global as well as a local granularity on the analysed data. The classification used is based on the DMOZ taxonomy and allows us to offer a categorification of news articles with a wide coverage (see Figure 18) to determine subcategories associated to the topic of search.

2.4 Automating Literature Review

Gaining knowledge from validated and published research is sometimes difficult due to the different nature of available knowledge sources, and the large amount of information to review. The SDG 1 Observatory is ingesting over 200 million articles and patents from OpenAlex, firstly originated with the Microsoft Academic Graph. These include ritch metadata including title, abstract and authors with affiliations, from where location can be extracted, but also the main topics where the research paper is focusing on.

The system is also ingesting the MEDLINE dataset consisting in more than 28M biomedical labeled articles from the, allowing to explore health conditions related to poverty factors. It is making all of it available through a complex data visualization technology based on text similarity that allows for the reordering of the queried results that are inevitably biased by the limitations of the search topic.

Moreover, we have used the *Searchpoint* technology (available at midas.ijs.si/searchpoint), that is in further development in the context of IRCAI's SDG Observatory activities, but can already employ Lucene language to offer powerful queries on the metadata of the six indices corresponding to the verticals where the SDG Observatory is focusing on. The results provided by the query can be reordered by moving a target over clusters of identified subtopics clustered by using K-means.

2.5 Analysis of Data Bias

Avoiding data bias in AI systems is crucial to ensure fair, accurate, and equitable outcomes, preventing the reinforcement of existing inequalities and enabling more trustworthy and inclusive technologies. In this last section we aimed at analysing the bias related to the data ingested in this observatory for SDG 1.

For analysis we use OECD AI Policy documents. Some of those documents are very large, and we split each document into smaller parts (so called "chunks"), which can contain multiple paragraphs. The reason for this is to prepare data for easier analysis with large language models, so called Retrieval-Augmented Generation (RAG). RAG is an advanced technique that combines retrieval-based methods with generative models to improve the performance of tasks such as question answering, text generation, and other natural language processing (NLP) applications. For each chunk then the sentiment is computed based on VADER (Valence Aware Dictionary and sEntiment Reasoner) methodology. Since VADER is known to have weak multilingual capabilities, all the documents were machine translated into English first.While the results of this procedure are reliant not only upon the accuracy of the sentiment analysis tool, but also upon the accuracy of machine translation, it is important to stress that sentiment analysis is less sensitive to common machine translation problems than other usages, because sentiment analysis usually focuses on identifying the polarity (positive, negative, neutral) of a text rather than understanding its full semantic content. Also, sentiments in text are often expressed redundantly, which can help mitigate the impact of translation errors. As a result, minor translation errors that do not alter the overall sentiment and do not significantly impact the sentiment analysis is possible.

For the purpose of this analysis, they computed the average sentiment of (chunks of) AI policy documents for each country. We are presenting the visualisation of average sentiment of countries' AI policy documents on the map. Since AI policy documents are mostly documents of legal nature (acts, policies, regulatory and governance frameworks), the sentiment should be mostly neutral, however, the analysis shows that there are country differences. VADER computes positive, negative and neutral sentiment. Each of those values are between 0 and 1. The score indicates the proportion of text that is considered positive, negative and neutral. The sum of negative, positive, and neutral sentiment scores always equals 1, however in practice the sum of three sentiment scores can sometimes slightly exceed or fall below 1 due to floating-point precision errors or rounding issues that occur during computation.

3. RESULTS AND DISCUSSION

In this section we discuss the main achievements from the application of the methodologies and technologies described in Section 2, covering a pipeline of complementary AI-based approaches.

3.1 Analysis of the Influence of Indicators

In what regards the analysis of indicators, we have achieved robust results that confirm the influence of expectable indicators such as, e.g., the GDP per capita or the Gini index, and on the other hand we have identified several. Employing feature selection as integrated in the Random Forest model, we were able to reach the following:

- Features were selected by training a random forest model on the full dataset and using the built in feature importance metric to determine the most important features **;**
- Performance of the model in regards to number of features plot is available.

The analysis showed, as illustrated by the chart in Figure 6 that 25 features are enough to continue with the in-depth analysis, as more features do not contribute drastically.

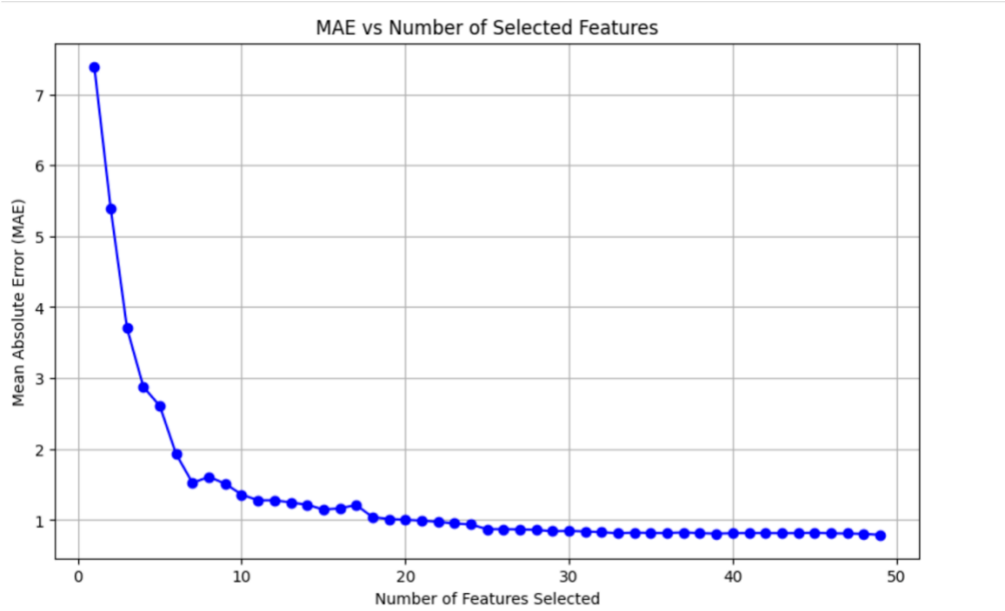

Figure 5 – MAE vs the number of selected features in the Random Forest model.

After training the Random Forest model on the full dataset, which included over 550 features, we assessed the importance of each feature using the model's integrated feature importance attribute. This generated a ranked list of features sorted in descending order of importance. For the Figure 5, we focused on the top 50 features. Initially, we built and evaluated a model using only the most important feature. We then incrementally added the top two, three, and so on, up to 50 features. This allowed us to observe how the model's mean absolute error (MAE) decreased as more features were incorporated.

The model's performance with just one feature was relatively strong, achieving an MAE of around 7.5%, which, while not optimal, was still better than random guessing. As we included additional features, the MAE rapidly dropped to around 0.87%. Notably, when the model used all available features, its accuracy slightly decreased to an MAE of approximately jus below 0.87%. Consequently, we opted to limit the model

to the top 25 features, resulting in a model with a mean absolute error of 0.87%. This indicates that the additional features do not significantly contribute to the model's performance.

Moreover, to explore further the feature importance of the selected 25 features, we have applied SHAP to elaborate on the influence of each indicator contributing to the Poverty of a nation. Figure 6 displays the top twenty most important features in the model, using a game-theory approach determine the impact of each feature on the overall outcome. Red dots indicate high values of a feature—for example, a high GDP—while the position of the dot reflects its effect on the model's predictions. If a dot is positioned to the right of the vertical zero line, it indicates that the feature increases the predicted poverty rate. Conversely, if a dot is positioned to the left, the feature decreases the predicted poverty rate.

For example, red dots on the right side of the line show that higher coefficients of lifespan inequality lead to higher predicted poverty rates, whereas lower coefficients reduce these predictions. However, not all features have straightforward interpretations. Territory under state control, for instance, shows high values on both sides of the line, meaning that it can both increase and decrease the predicted poverty rate, depending on other interacting factors. To enhance readability, feature names are shortened in the figure.

Figure 7 provides a visualisation of how the Random Forest model generates its predictions. The model begins with a base value, which in this case is 29.42. Features are then sequentially considered and applied to the prediction. The visualisation shows each feature's contribution to the final prediction and the direction of its influence. Larger areas in the plot represent features with a greater impact. For example, in the last plot, we observe that the equal resource distribution index significantly reduces the model's prediction, while the lifespan inequality index pushes the prediction higher. This provides valuable insights into which features the model considers important and enhances our understanding of the model's decision-making process. The figure above presents three visualisations of the same prediction.

The first plot illustrates how the largest model, which includes all features, predicts poverty. Many features have minimal visual representation due to their small impact compared to the more influential features, such as those mentioned earlier. As previously justified, the model with only 25 features performs equally well, while offering greater interpretability, as more features are visibly represented, and their impact is more pronounced. This is reflected in the larger areas corresponding to each feature, signifying a greater influence on the model's output. Further reduction to 10 features enhances interpretability and simplicity, but at the cost of a slight decrease in accuracy. In this specific case, the predicted value was 4% higher compared to the larger models. The reduction in accuracy is expected, given the relationship between the number of features and prediction accuracy, as demonstrated in earlier tests.

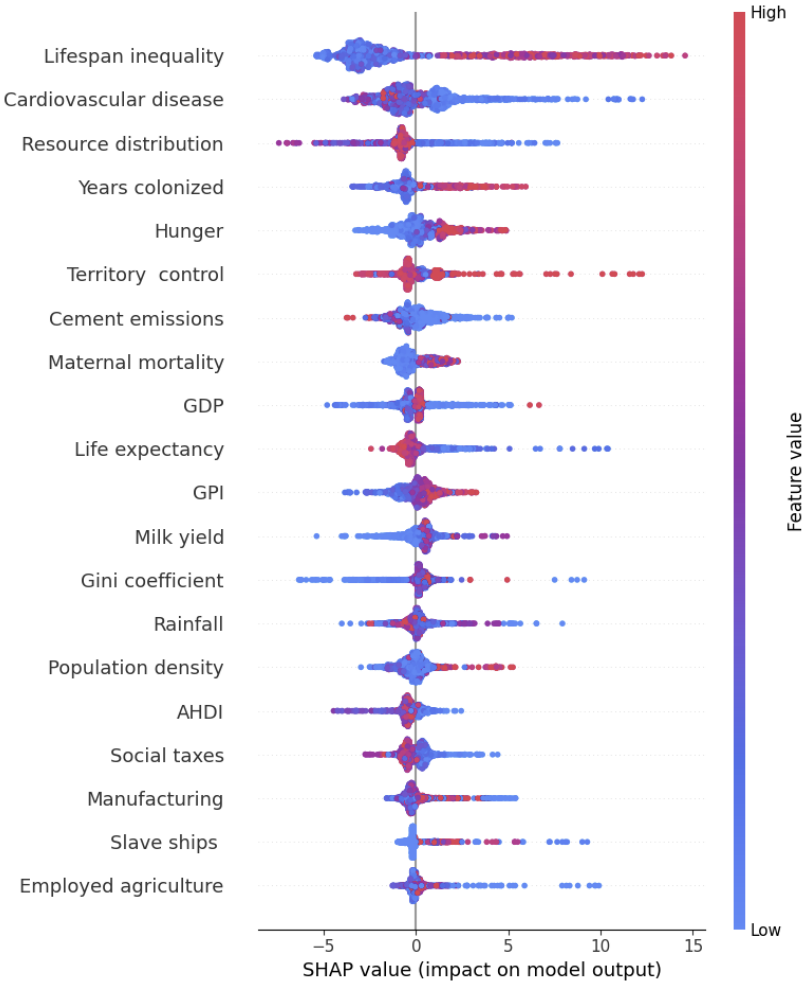

Figure 6 – SHAP exposing the influence of the top 25 features in the Poverty baseline.

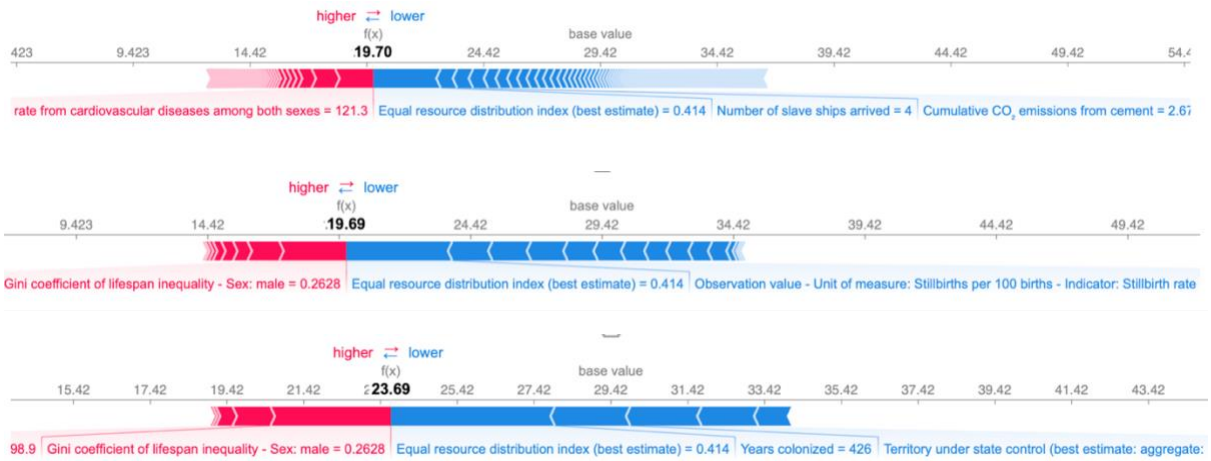

Figure 7 – SHAP exposing the influence of the top 25 features in the Poverty baseline: the full model (on the top), the Top 25 features (in the center) and the Top 10 features (below)

Using Transfer Entropy to determine causality and find results, has many benefits where Granger causality fails. In particular, there is no need for linear relationships allowing us to consider a more complex system; and also there is no need for stationary data, being a bit more flexible in regards to how can indicators behave. Moreover, this method was also used in a number of papers to determine causality ("Causality detection based on information-theoretic approaches in time series analysis").

Taking into consideration the results in the Table 1 above, there are several straightforward observations that confirm the validity of the model, e.g., taxes are surely tied with Poverty. Also, GDP per capita and the gini coefficient were expected to be important as inequality leads to poverty, and the GPI score is expected to be directly impactful, although results are only being reported only since 2008. Territory under state control is directly tied to poverty, as it includes peace and war status. On the other hand, per capita, electricity may not directly influence Poverty, but it can have an indirect effect on the ease of living and starting new businesses. Access to electricity also tied with electricity production probably: indirect effect - electricity improves life and frees up more time. Moreover, manufacturing value as % of GDP shows productivity of economy, and life expectancy of birth also an indicator tied closely with poverty: high life expectancy leads to less poverty. Similarly, maternal mortality ratio is tied closely to the life expectancy ratio probably, which is tied to poverty. AHDI was also expected to be impactful, although we would expect its impact to be higher: composed of lots of variables, some of which are already accounted for in the higher features. Population density repeatedly appeared as important, which is something that isnt talked a lot about. Perhaps the combination of a high density with low values of other features could lead to very good metrics on poverty.

Table 2 lists the top thirteen features ranked according to their importance. A feature's rank was determined by sorting the features based on the sum of information across all countries and assigning a rank number. The feature with the highest sum of information was ranked 1, while the feature with the lowest sum among the top 25 received a rank of 25. The numbers in the table represent the cumulative ranks of each feature across all countries, providing insights into the relative importance of each feature at the country level. Social taxes received the highest rank, indicating that they consistently held a high sum of information across different countries. Additionally, the other features are ranked closely together, suggesting there is no strong hierarchical order among them.

Feature	Sum of information	Feature -> Poverty	Poverty -> Feature
Social taxes	0.774259	0.297423	0.476836
Electricity per c	0.689185	0.293602	0.395583
Milk yield	0.665093	0.277379	0.387714
Cardiovascular disease	0.664551	0.287829	0.376723
Gini coefficient	0.630846	0.251921	0.378924
Manufacturing	0.619951	0.263341	0.356610
Life expectancy	0.611045	0.274224	0.336821
GPI	0.603816	0.278973	0.324844
Arable land	0.584745	0.269353	0.315392
Maternal mortality	0.584726	0.275993	0.308734
GDP	0.561216	0.273022	0.288195
Employed agriculture	0.556179	0.299076	0.257103
Stillbirths	0.542237	0.285056	0.257180
Territory control	0.534441	0.203164	0.331277
AHDI	0.531442	0.265586	0.265856
Population density	0.524632	0.270122	0.254510
Lifespan inequality	0.518962	0.275049	0.243913
Resource distribution	0.514670	0.223998	0.290672
Fresh water	0.499583	0.275858	0.223725
Hunger	0.459547	0.209844	0.249702
Electricity Access	0.436756	0.209249	0.227507
Cement emissions	0.376913	0.199994	0.176919
Rainfall	0.161208	0.054709	0.106499
Years colonized	0.143508	0.054188	0.089321
Slave ships	0.000000	0.000000	0.000000

Table 1 – Results of applying Transfer Entropy in relation to Poverty. The table presents average values for features and is sorted by the sum of information, high to low.

Feature	Rank
Social taxes	1538.0
Electricity per c	1925.0
Cardiovascular disease	2061.0
Milk yield	2125.0
Gini coefficient	2129.0
Manufacturing	2163.0
Life expectancy	2164.0
GPI	2261.0
Maternal mortality	2442.0
Arable land	2558.0
Employed agriculture	2661.0
GDP	2704.0
Population density	2826.0

Table 2 – Rank of features based on transfer entropy obtaind by sorting the features for each country by their sum of information and assigning them a rank. Rank 1 was assigned to the highest feature with the largest sum of information and rank 25 to the feature with the lowest. All the ranks were summed and provide a general idea of where each feature is in order of importance.

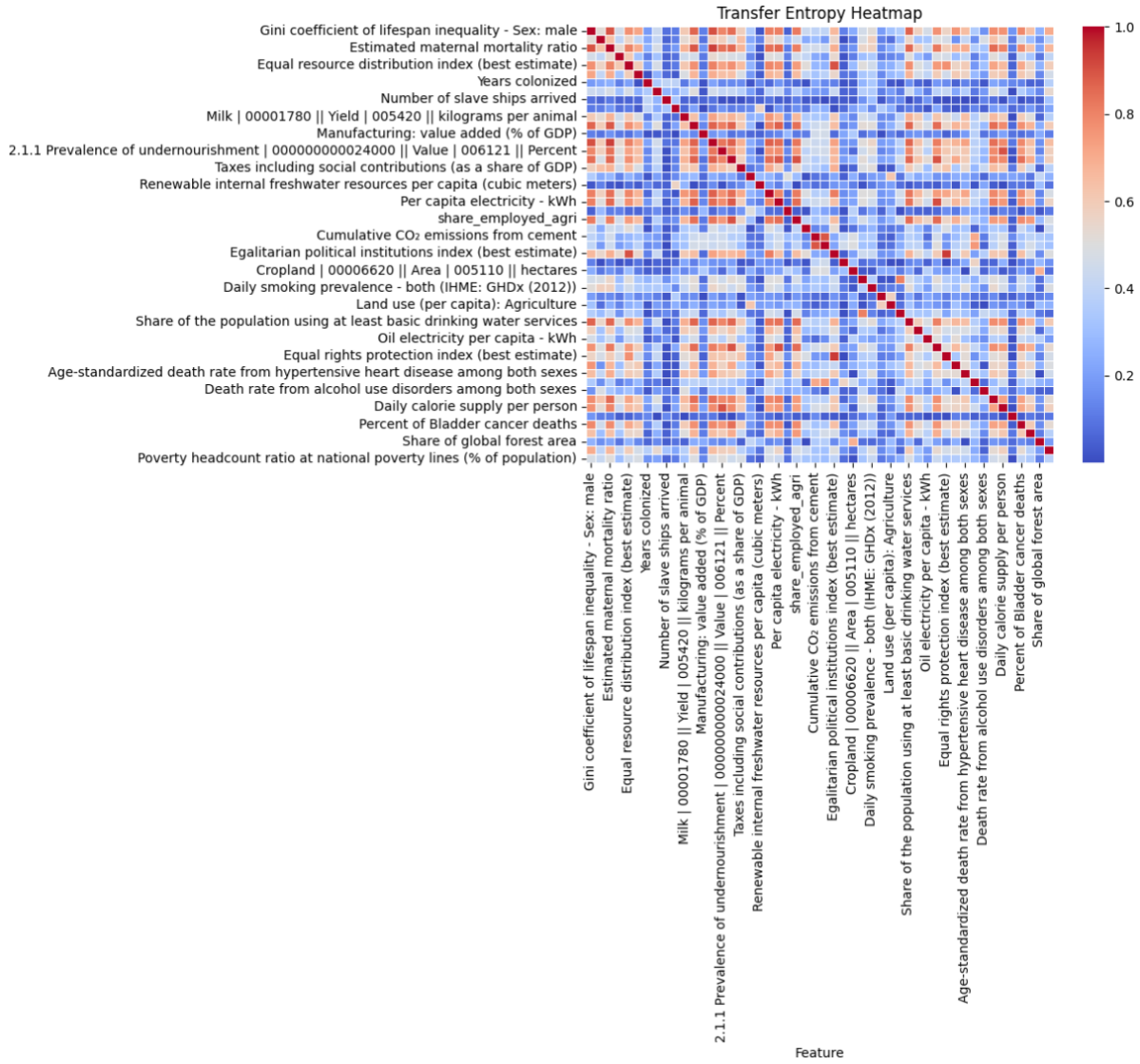

Figure 8 – Spearman correlation of the top 50 indicators influencial to Poverty based on feature importance

The heatmap in Figure 8 representing the correlation between the top 50 indicators influencing poverty, shows clusters of highly correlated features, often revealing underlying relationships between socio-economic and environmental indicators. For example, variables like the *prevalence of undernourishment*, *calorie supply per person*, and *renewable internal freshwater resources per capita* show significant correlations with one another, reflecting how food security and access to resources are interconnected in the context of poverty. Furthermore, high correlations can be seen between measures such as *equal rights protection index*, *egalitarian political institutions*, and *Gini coefficients*, suggesting that governance and inequality metrics are tightly interlinked in predicting poverty outcomes. In contrast, features such as modern economic indicators like *manufacturing value added* or *renewable energy usage*, pointing to historical legacies impacting current development trajectories.

The maps in Figure 9 provide insights into how information and influence are shared across countries in key development areas. Each factor illustrates distinct global patterns of connectivity, reflecting the differing nature of these influential domains across the world. The map above shows how information related to manufacturing flows between countries. Particularly in parts of Asia we can see high information exchange in manufacturing, likely driven by strong industrial networks and global supply chains. Regions

like Europe, North America, and parts of East Asia show high connectivity, reflecting their roles in global production. Conversely, much of Africa and some parts of South America exhibit lower levels of transfer entropy, indicating a less integrated position in the global manufacturing sector.

Social contributions (represented by the map in the center) encompasing social security systems, community development, and related activities show that South America, especially Brazil, stands out with a high sum of information, indicating that social programs or policies in these regions significantly influence others. Europe and parts of Asia also exhibit moderate levels of information exchange, likely due to strong welfare systems. Countries in Africa and the Middle East show lower levels of transfer entropy, suggesting less cross-border impact or influence in terms of social contribution practices.

The relevance of freshwater resources depicted in the map below, indicate how water management and availability in one country might influence others. Regions like South America (notably Brazil), parts of the Middle East, and Central Asia suggest these countries have significant influence or interdependency regarding freshwater management, unlike regions like North America, Sub-Saharan Africa, and parts of Europe displaying moderate to low levels of information exchange, possibly reflecting varying levels of resource management or less regional dependency on shared freshwater resources.

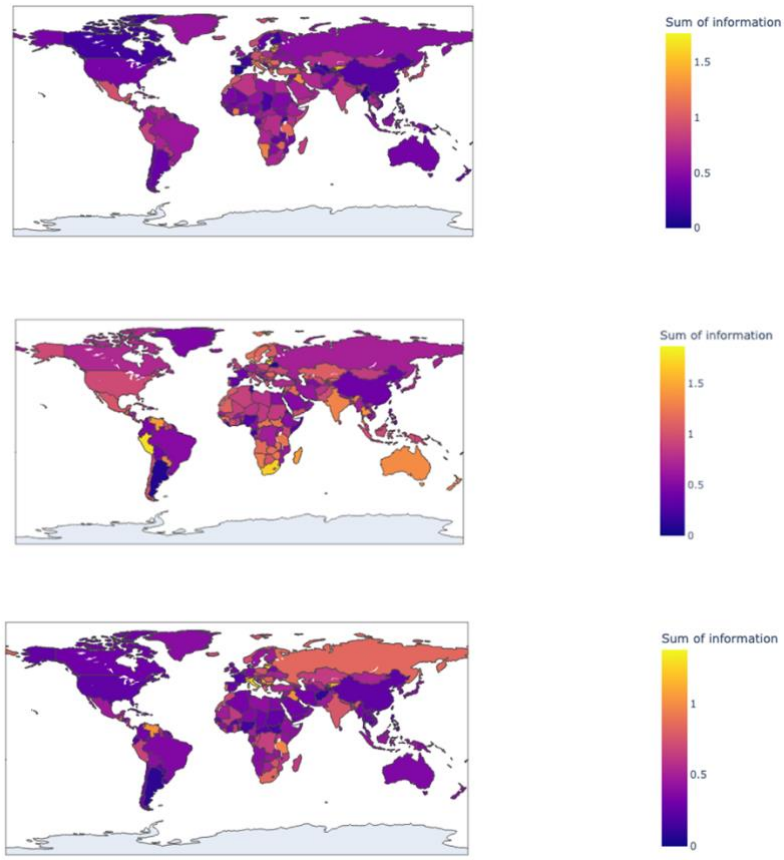

Figure 9 – Transfer entropy map for three different influencial factors: manufacturing (above), social contributions (in the middle) and freshwater resources (below).

3.2 Looking into Trends in the Data

Overall, the Markov Chain analysis of the selected SDG 1 indicators has shown that the distribution of states across world regions is very different (see Figure 10), indicating that the geographic location could be a factor to take into account when considering the relevance of the indicators as features considered in this study. While in some regions the *drinking water* has a significant role in the context of the parameters that might be influencing Poverty (e.g. *Middle East & North Africa* or *Sub-Saharan Africa*), in some other regions parameters as *clean fuels* are taking a significant role (e.g. *East Asia & Pacific* or *South Asia).*

Figure 10 – The Markov Chain analysis of world regions regarding SDG 1 indicators.

To further this analysis, we have analysed the SDG 1 factors at national levels through this methodology, taking into consideration that as averaging within world regions could be misleading, also the national level analysis might be enclosing circumstances with national regions affected differently by SDG 1 factors. Figure 11 shows the different complexity obtained in regards to the decomposition of states between countries that can indicate the significantly diverse relevance of SDG 1 factors in regards to other factors that are not being captured here.

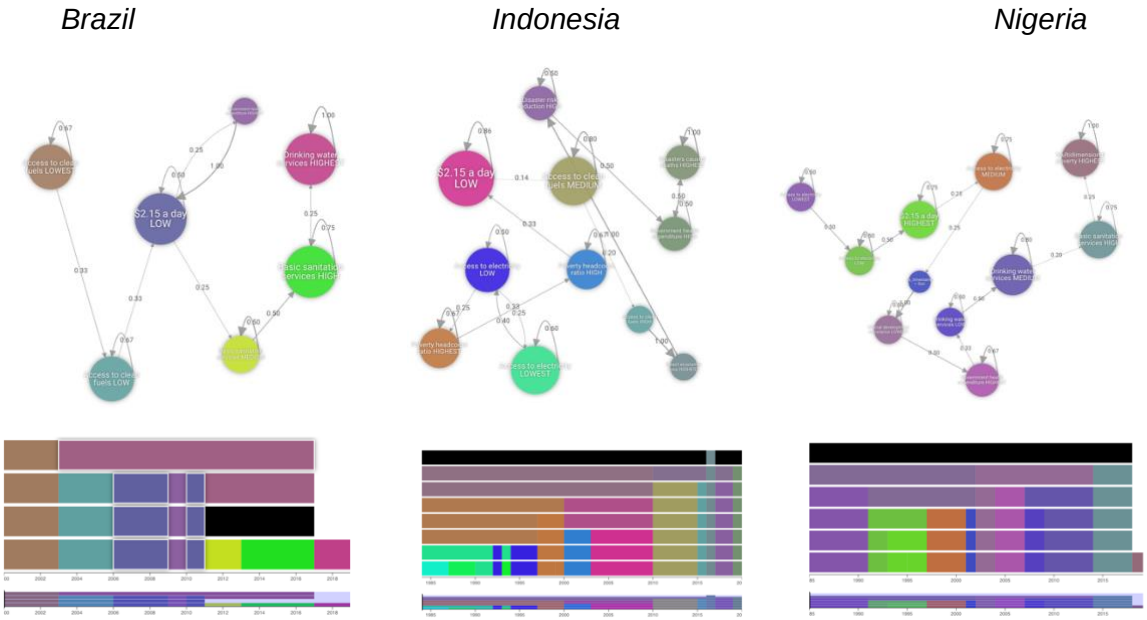

Figure 11 – SDG 1 Markov Chain analysis of countries including relation between states and overall hierarchy.

When looking in more depth to the states in a single country as, e.g., Brazil we can use the Markov Chain to provide us with a certain measurement in relation to the states closely related to the *$2.5 a day* indicator. The image in Figure 12 represents a Markov chain corresponding to the data in Brazil, representing a type of probabilistic model described as a systems that transition between different states with certain probabilities.

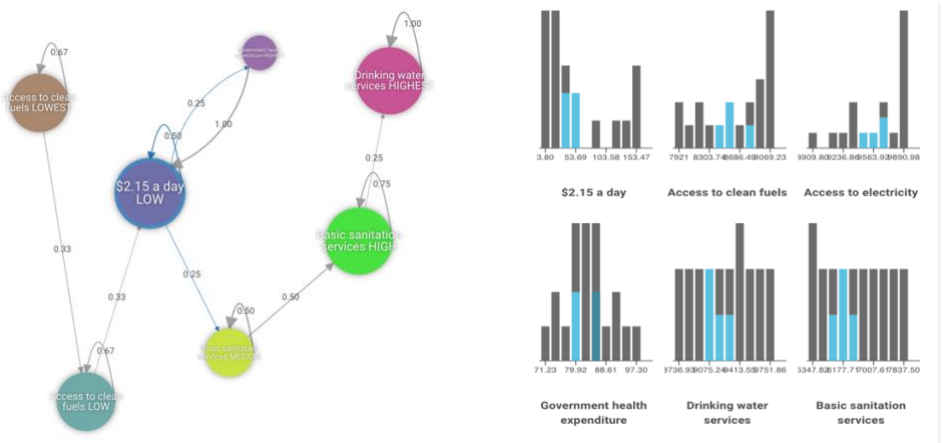

Figure 12 – Measurement of influence of Poverty indicators in the <$2.15 a day> center Markov Chain state in the context of the relations with other states.

The Markov chain represented in Figure 12 is modeling transitions between different states related to socio-economic and environmental conditions, where key components are (i) the state "$2.15 a day LOW" refering to a low-income or poverty state, (ii) "access to clean fuels LOW/LOWEST" indicating different levels of access to clean fuels for households; (iii) “basic sanitation services MEDIUM/HIGH" refers to the level of access to sanitation services; (iv) "drinking water services HIGHEST" relating to access to drinking

water services at the highest possible level; and "Government health expenditure HIGHEST" that may refer to significant government investment in health services. On the other hand, transitions and probabilities show that:

The analysis reveals persistent challenges in poverty and access to essential services. For individuals earning $2.15 a day or less, there is a 50% chance of remaining in poverty, with a 33% likelihood of also facing limited access to clean fuels and a 25% chance of improving sanitation services. Households with low access to clean fuels have a strong 67% probability of remaining in this state, and a 33% chance of experiencing further decline. In terms of sanitation, those with medium access have equal chances (50%) of either maintaining or improving their situation, while high sanitation access is more stable, with a 75% chance of remaining in this state. Interestingly, a high level of government health expenditure is shown to still result in a transition to extreme poverty, highlighting the complexity of addressing poverty even when public health investment is high.

The Markov chain models complex relationships between poverty, energy access, sanitation, drinking water services, and government expenditure on health. It highlights the persistence of poverty, limited access to clean fuels, and the gradual improvement in sanitation and water services, while also underscoring that some areas (like government health expenditure) may not have a direct or immediate impact on poverty alleviation. In particular:

- **Poverty and clean fuel access:** The model shows that populations in the "$2.15 a day LOW" state have a significant probability (33%) of moving to low access to clean fuels, indicating that poverty is linked with energy poverty.
- **Sanitation and Water Services:** There's a strong link between basic sanitation services and access to drinking water. A high level of sanitation services (75%) often leads to improved access to drinking water services, suggesting improvements in one may facilitate the other.
- **Health expenditure and poverty:** Despite high government health expenditure, poverty persists (as seen in the direct connection from "Government health expenditure HIGHEST" to "$2.15 a day LOW"), suggesting that healthcare investments alone may not be enough to lift people out of poverty.

3.3 Estimating Impact from the News

Although weaker than the signal from published research (in regards the veracity and quality of the content), news articles can provide us substantial information about an event in the context of SDG 1, or overview perspectives that can offer in a more recent timeline a view on the status of a Poverty-related topic. Having in mind the crosslingual capabilities we have available, through the collaboration with Eventregistry.org covering more than 60 languages, we are able to overcome the language limitation in this study.

As shown in the Table 3 below, the language coverage for different SDG 1 topics is also different, relative to their popularity. The mentioned crosslingual capabilities are anchored with the multilingual nature of Wikipedia terms followed by a textual description to which we can apply together with methods such the *wikification* (earlier described in section 2) of the news text in order to capture the essential structure of non-english language texts. While news sources are a very relevant source of information in what regards the public opinion and the major events regarding world poverty, we need to take into consideration that there is a substantial bias in what discussions topics and geographical coverage are represented, and related to the *"newsworthy"* factor that often does not allow to compare this signal between countries and world regions.

The automated ability to identify news events from the context and key entities mentioned in the news article offers the potential to build a *news story* from a collection of news articles, that has in common the

timeframe, main actors and main topics (see Figure 13). Furthermore, Eventregistry is also capable of almost-real time identification of SDG 1 events captured in the worldwide news, in the same manner as applied to AI-related events by the OECD AI Incident Monitor[4].

Figure 13 – The sequence of news articles constituting the news story of the SDG 1 event.

SDG 1 Topic	Key Term	Language Coverage	Number of News	Number of Events
Poverty	Poverty	128	7,333,615	135,551
SDG 1	Sustainable_Development _Goal_1	17	1690	47
Child labour	Child_labour	62	275,500	8,886
Microcredit	Microcredit	32	116,089	3,459
Unemployment	Unemployment	99	8,987,639	293,561
Welfare	Welfare_spending	42	5,046,604	103,762
Gini Index	Gini_coefficient	83	39,369	1,248

Table 3 – News coverage in the last decade regarding SDG 1 topics and their language coverage.

The climate change is a global problem that in the recent years has been in the focus of Worldwide strategies. The priorities are rapidly changing towards sustainability and environmental efficiency, transversely to most domains of action, that also take into consideration . The topic of Poverty has an important role in the context of climate change impact, where approaches such as IRCAI's SDG1 Water Observatory and Eventregistry presented in this paper can provide significant contribution. Already from

[4] See https://oecd.ai/en/incidents

the appropriate exploration of worldwide news it is noticeable the increase of reported events relating to Climate Change and its impact on Poverty (see figure 14). This reflects almost 286 thousand news articles with increasingly higher peaks, and including prevalent concepts (see Figure 15) such as, e.g., sustainable development (44030) and food security (40761) and drought (41003) or floods (36681), with breaks in 2021 and in 2022 eventually due to COVID outbreak prioritised media coverage.

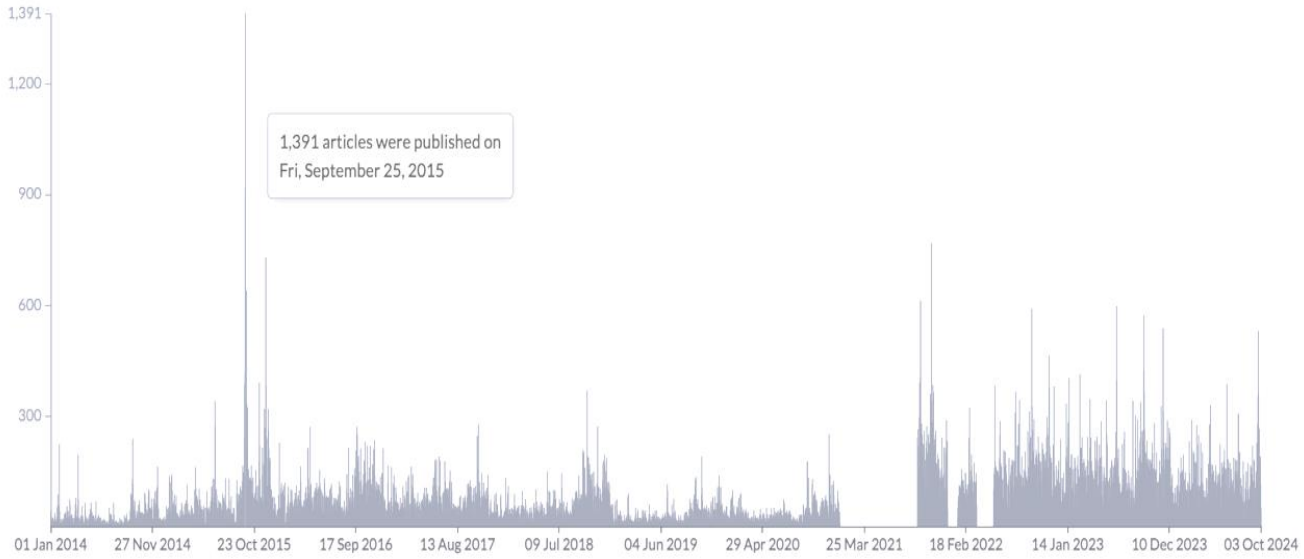

Figure 14 – The timeline of news articles published worldwide on climate change in relation to Poverty over a decade, showing ...

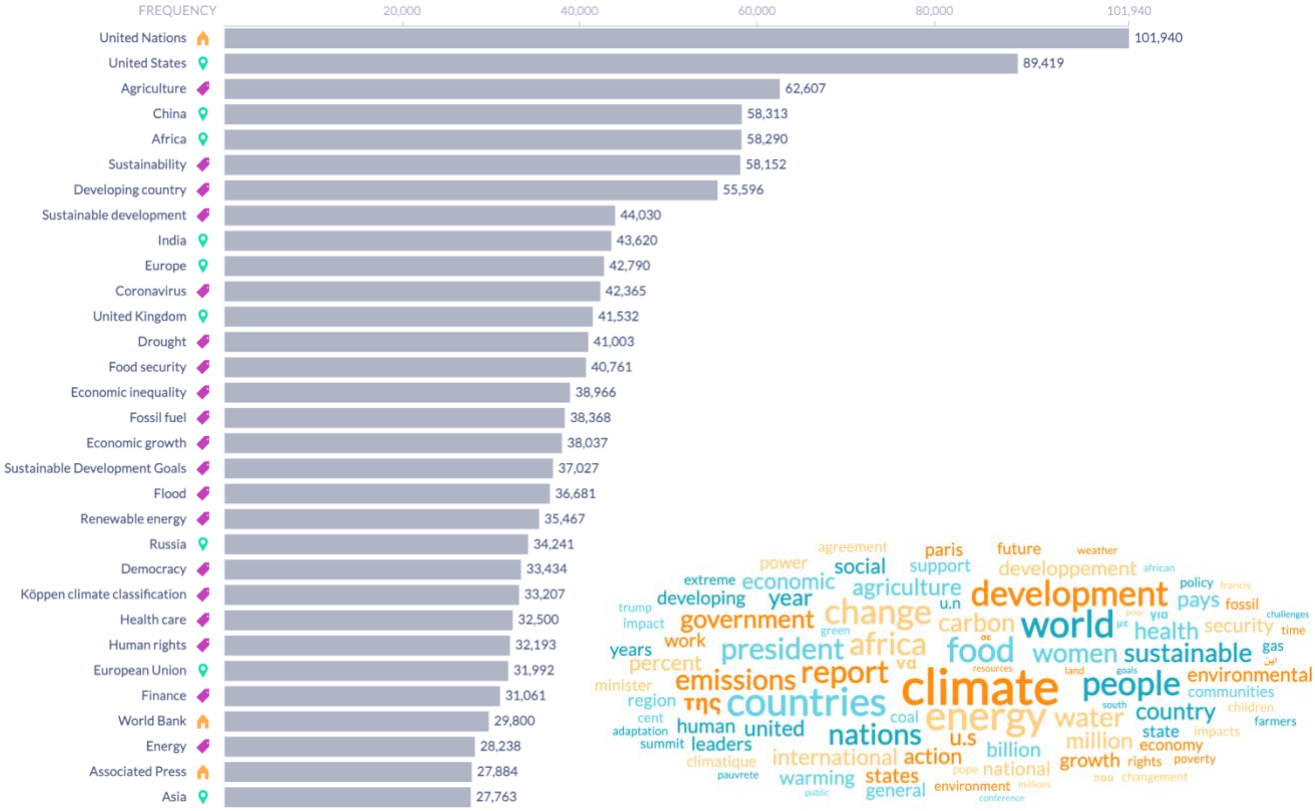

Figure 15 – The main topics in the news published worldwide on climate change in relation to Poverty over a decade, showing ...

Based on these powerful algorithms, the system shows, e.g., how a relevant part of the news on poverty in 2021 in Brazil relate to the concern of climate change (see figure 16). This is done by the automated classification of the ingested news text into the DMOZ taxonomy, allowing also to improve the queries to the dataset (Leban et al, 2014).

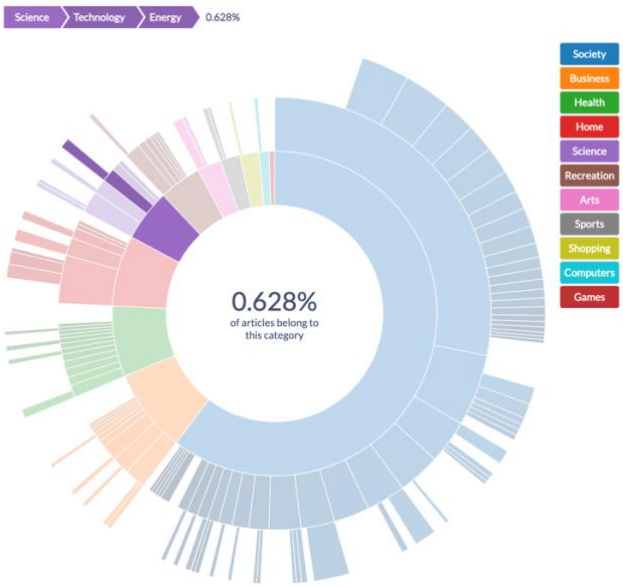

Figure 16 – Main categories of news within SDG 1 -related topics in the last year showing how Energy is taking an important role in the public debate in the past decade, since September 2014.

The sentiment analysis is one of the capabilities that is common in the analysis of social media posts and news articles, that helps gain awareness on the public opinion in, e.g., the rapid action to reduce the impact of SDG 1-related events by local, national and multinational authorities. In the example of Figure 17, that obtained a large positive sentiment in the news, probably related with the way local authorities are addressing these problems. From October 2023 to October 2024, we observe periods of fluctuating sentiment, where notable spikes in sentiment correspond to key events or discussions in the news. Despite some dips into negative sentiment, the overall trend shows more frequent positive feedback. The number of articles also seems to increase gradually over time, indicating growing media attention toward the event, especially in the later months, with a peak in August 2024. The sentiment spikes, especially in February and late April, suggest significant newsworthy developments that were positively received. The interpretation could reflect the success or progress in SDG1 initiatives, fostering increased media engagement and positive sentiment overall.

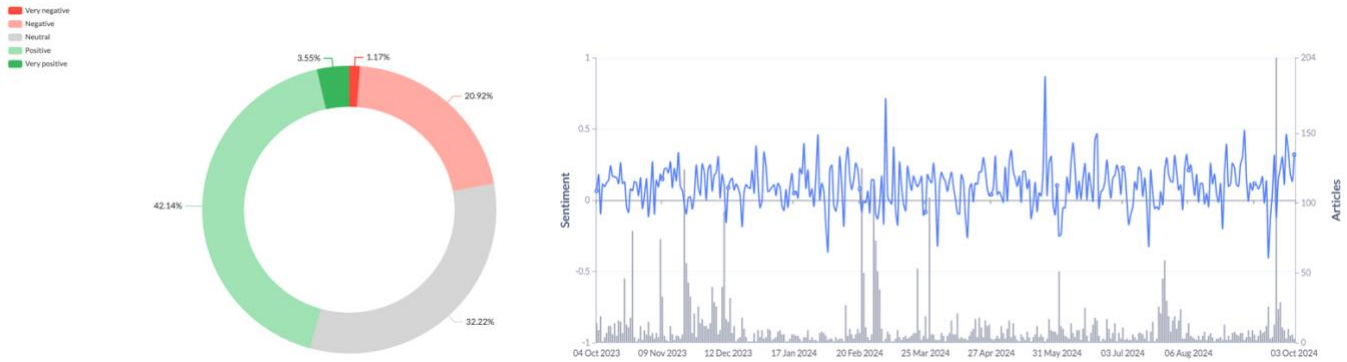

Figure 17 – The sentiment analysis from news allows us to show the overall positive feedback on the SDG1 event happening in Brazil in 2023, overall (left) and on a timeline (right).

3.4 Extracting Best Practices from Science

To fully harness the power of text mining technologies for extracting valuable information from text, we can integrate multiple layers of data, including public and journalistic perceptions from news, and supplement them with reliable scientific insights from research papers and patents. This approach offers a more comprehensive understanding of events and outcomes. Specifically, in the context of SDG 1 (No Poverty), this enables us to not only track and interpret global media perspectives but also to learn from the scientific community's research on poverty-related topics. At the moment there are 1 316 341 research articles associated to SDG 1 topics in the system.

By analyzing global news coverage on poverty, we can identify patterns in similar events, key stakeholders involved, and successful actions taken. This allows us to use the extracted information as a tool to identify best practices and actionable insights. Automating this information retrieval is crucial, especially as we observe an increasing number of poverty-related events across the globe. Leveraging multilingual capabilities in text mining helps to overcome language barriers and national-level limitations that often hinder traditional methods of analysis (see coverage in Table 4).

Moreover, this process identifies relevant concepts and relationships between them, enriched by the insights drawn from scientific research. By incorporating the latest findings from scientific publications ingested into the SDG 1 Observatory, we can enhance our understanding of these issues and refine our approach to information extraction. For example, research on climate change-related consequences, such as floods and droughts, provides essential context for monitoring extreme events and their increasing frequency, allowing us to draw more precise correlations and anticipate impacts.

SDG 1 Topic	Code	Number of Educational Resources	Number of Research Articles	Number of Policy Documents
Poverty	C189326681	53	442,421	2,605
Poverty reduction	C2992104146	0	39,717	35
Child labour	C2777464741	2	7,774	55
Micro finance	C2992021695	5	2,554	10
Unemployment	C186067381	5	47,426	707

Table 4 – Science and policy coverage in the last decade regarding SDG 1 topics.

To maximize the potential of this extracted information, it is essential to integrate it with official indicators, as discussed in sections 3.1 and 3.2. This combined analysis enables us to develop actionable strategies that can improve the progress of SDG 1. Tools like the SDG Observatory platform (e.g., sdg-observatory.ircai.org) help users analyze trends, compare responses from different countries, and investigate successful policies and practices by returning to relevant news articles and scientific research.

In this note we make available a Research Explorer Tool[5], where the user can query for, e.g., "poverty" being presented with a clustered word cloud (see Figure 18) and can drag a target over it to reorder priorities according to focus on, e.g., filters or purifiers while searching for the patented innovations registered, and is served with the linked title and abstract extract of each patent. The same workflow is

[5] See https://midas.ijs.si/searchpoint/result.html?q=poverty&c=kmeans

valid to explore best practices in the ingested scientific articles and is also compatible with ingested reports or other textual records of interest with metadata associated allowing for Lucene-based queries.

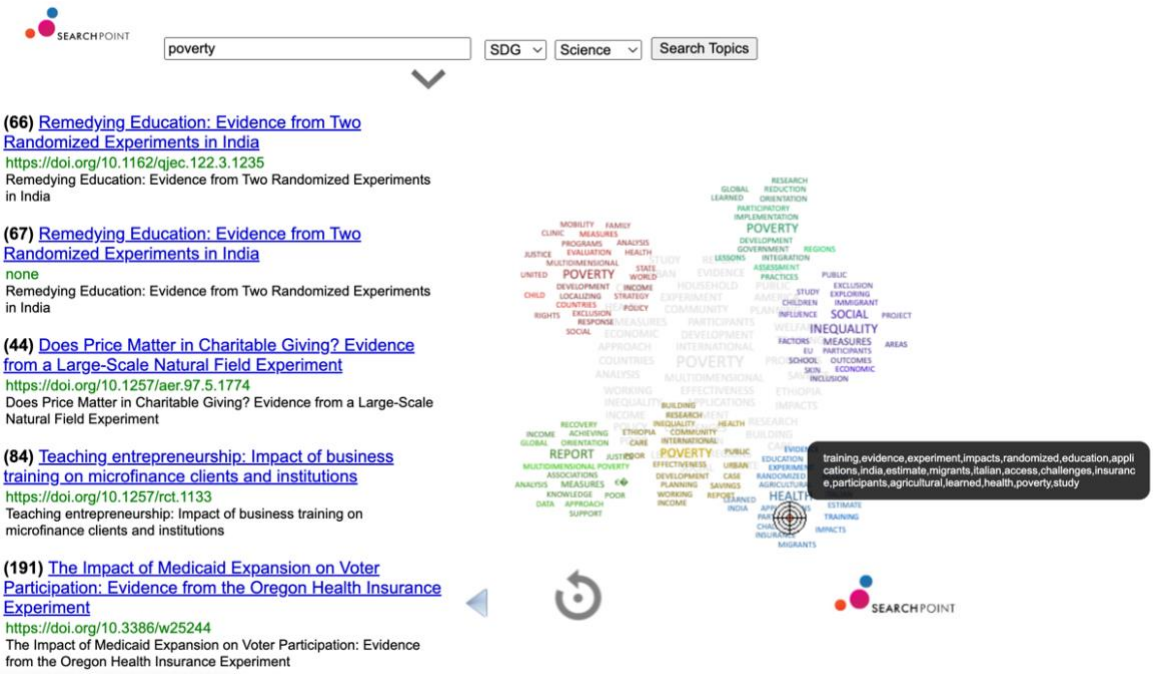

Figure 20 –SDG 1 -related topics within 200+ million published articles as explored by the tool Searchpoint.

3.5 Towards Unbiased SDG 1 Data

Sentiment analysis detects emotions or opinions in text and can therefore be a powerful tool to measure bias, because it provides insights into how language reflects social biases. OECD has a collection of various national AI policies and initiatives (more than 900 documents). Policy documents should typically use a formal, technical, and often bureaucratic style of language, because they are designed to communicate rules, guidelines, regulations, or strategies. So their language needs to be clear, precise, and unambiguous, with neutral sentiment. Our motivation for sentiment analysis on those documents is to uncover possible (subtle) biases in policy and legislative language, i. e. whether they are neutral or whether they express enthusiasm or pessimism towards AI related topics.

For the purpose of our analysis, we computed the average sentiment of (chunks of) AI policy documents for each country. The map in Figure 19 illustrates a potential data bias in AI sentiment and policy analysis, especially in the context of SDG 1. The color-coded scale on the map ranges from countries with higher neutral sentiment (darker green) to lower neutral sentiment (lighter shades) regarding AI policies in the context of SDG 1, with gray areas indicating countries for which no data is available. While more developed countries seem to have readily available sentiment data (either neutral or otherwise), there is a significant gap in data for many developing regions, particularly in Africa and parts of Asia. Addressing this bias will be crucial for ensuring that AI policies are equitable and inclusive, especially for achieving SDG 1 goals in regions most affected by poverty.

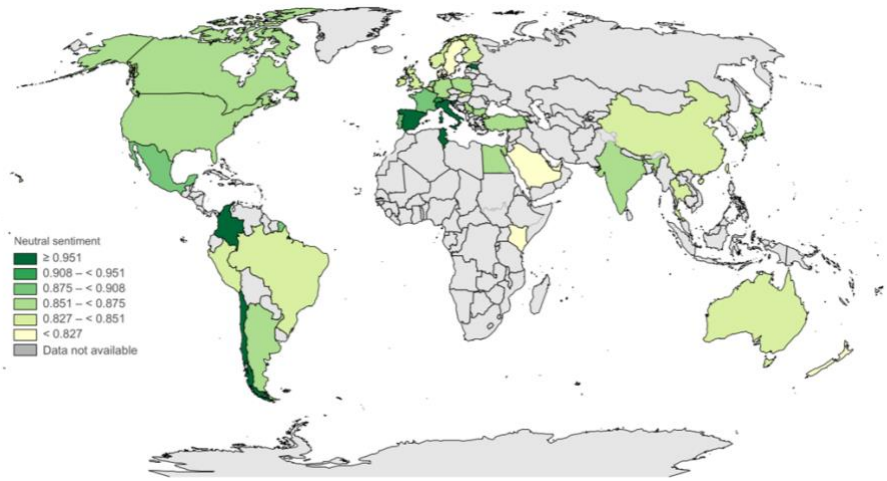

Figure 19 – Countries with neutral sentiment on AI policies regarding SDG 1 topics.

Countries like Spain and a few others in South America (e.g., Peru) show high levels of neutral sentiment. A higher neutral sentiment in these regions could suggest a balanced or ambivalent public or governmental stance on AI policies in relation to poverty reduction (SDG 1). This might indicate a cautious approach where AI isn't seen as either particularly beneficial or harmful in achieving poverty-related goals. On the other hand, some countries like those in Eastern Europe, parts of Latin America, and Australia exhibit lower neutral sentiment. This may indicate either strong positive or negative opinions on the impact of AI in poverty-related policies. A more pronounced stance could be a signal of polarized views on whether AI will positively impact SDG 1 targets or exacerbate inequalities.

Figure 20 – Sentiment analysis per country for AI policies regarding SDG 1 topics.

We have also created visualisation of neutral sentiment of AI policy documents on a world map, showing which countries have most neutral AI policy documents. The chart in Figure 20 shows a dominance of neutral sentiment across many countries, which may suggest uncertainty or ambivalence about AI's role in addressing poverty (SDG 1). The positive and negative sentiments, while present in several countries, indicate that AI's potential to impact poverty is being recognized, but also raises concerns about AI exacerbating inequalities. However, the concentration of data from more developed nations highlights a potential bias, as countries most affected by poverty (particularly in Africa and parts of Asia) may be underrepresented or have less developed AI policies, leading to an incomplete global view of how AI is influencing SDG 1.

Neutral Sentiment Dominance. The predominance of neutral sentiment across most countries suggests a potential data bias stemming from limited AI implementation or a lack of robust data collection mechanisms, especially in developing countries. This bias might obscure true public or policy perspectives, particularly in regions where AI development is still in its nascent stages. It's also possible that sentiment analysis tools or datasets used to generate this chart are skewed toward certain regions or sources (e.g., academic publications, government reports), which may not capture the full spectrum of public opinion, especially in lower-income countries.

Underrepresentation of Developing Nations. Many developing countries are either absent from this chart or display overwhelmingly neutral sentiment. This suggests potential geographical data bias, where countries with less advanced AI infrastructure or fewer public discussions about AI's role in poverty alleviation are not sufficiently represented. The dominance of high-income nations in AI discourse can lead to global inequality in AI policy formation, where developing nations, despite being more directly impacted by poverty, are not equally involved in shaping AI's role in addressing SDG 1.

Contextual Factors: The level of technological advancement and AI adoption plays a role in shaping sentiment. Countries with advanced AI infrastructures (e.g., Germany, Japan, United States) are more likely to have both positive and negative opinions, as AI is actively deployed in various sectors. In contrast, countries where AI adoption is slower may show more neutral sentiment, possibly due to fewer tangible results or concerns over AI's impact. Countries with higher inequality or digital divides (like Kenya or Brazil) may be more prone to expressing both positive and negative sentiments, reflecting optimism about AI's potential but also skepticism about whether AI will truly be inclusive.

4. CONCLUSIONS & FURTHER WORK

The results presented in this study show the insight that can be obtained from open data in relevant SDG 1 topics also in relation to other SDGs such as water availability (SDG 6) or climate change preparedness (SDG 13). The complementarity of different data sources as different perspectives of the same topic with regional granularity, allow the insightful assessment that contributes to the enhancement of the intelligence on Poverty-related topics, but also towards better informed, empowered and engaged communities.

We will continue developing the SDG 1 Observatory, improving the capabilities in regards to the *Indicators* vertical, and making available a series of visualisation modules for all of the remaining five verticals. This R&D work is already planned and currently ongoing. We will also be working together with domain experts to further refine the usability and reach of the observatory functionalities in alignment with user stories and stakeholder priorities. This technology will also serve as a collaboration platform, and the developed visualisation modules can be provided as contribution to other research in relation to the progress of SDG 1, or likewise research collaboration results from other partners can be relatively easily integrated in the system as new *views*.

Our research will aim to carefully study how AI affects efforts to reduce poverty, determining whether AI has helped or harmed in the global fight against poverty, with a clear understanding of its many effects. To achieve this, we will be building on the knowledge and experience gained with the OECD AI Policy Observatory, with a comprehensive framework encompassing: the access to essential services, the income levels and economic activity, and social equity and inclusion. By integrating these metrics and criteria, our research aims to provide a holistic and evidence-based assessment of AI's impact on SDG 1. This will not only illuminate the current state of AI in poverty reduction efforts but also guide future initiatives to ensure they are effective, ethical, and inclusive. Our ultimate goal is to harness the potential of AI to make a tangible difference in the lives of the world's poorest populations, contributing significantly to the global ambition of eradicating poverty.

The study of indicator influence and causality had several limitations that will be addressed in future work. Data scarcity and missing values reduced the robustness of our model and results, as some missing values had to be artificially imputed. Additionally, we were unable to cluster the data effectively, which might have provided interesting insights into how countries or years are grouped. Clustering would have allowed for more granular results based on each cluster rather than producing only global results. While this could lead to more precise decision-making, the challenge of limited data remains a critical issue. Determining causality was also particularly challenging, especially given the diversity of the data used in this study. Further research into causality in complex systems would be highly beneficial, as many available algorithms have limitations that hindered their application in our analysis.

In addition, we are planning to improve the explainability of patterns in state timelines using LLM. The LLM first identifies patterns, analyzes whether they are recurring or anomalous, and extracts when these patterns occurred. By adding extra context to the prompt, we aim to help the LLM provide clearer explanations for each pattern. This process can give a better understanding of how SDG indicators and timelines relate, helping to track progress and find areas for action. The work on the Markov Chain of states for SDG 1 data will benefit from understanding the dynamic transitions between different poverty levels, enabling the analysis of how countries and regions move in and out of poverty over time. By modeling these transitions, we can capture the probabilistic nature of poverty evolution, which helps in identifying patterns, predicting future trends, and assessing the long-term impact of interventions aimed at eradicating poverty.

The exploration of worldwide news on SDG 1-related topics will continue to be pursued, particularly in what regards extracting from the text with the help of the most recent text mining methodologies and tools, the impact of events determining Poverty conditions. Moreover, we will be researching on the connections between the exploration of worldwide news and the usage of the insights from scientific research as search guidelines. The alignment with the outcomes of the research on causality in this study will be further implemented in the context of news analysis, and we will continue pursuing the potential of news to estimate further happenings in a SDG 1-related news story based on similar stories identified through means of text similarity and agnostic to the original language they are written in, by employing AI-based crosslingual methods.

To advance our research on SDG 1-related knowledge from research and policy papers, we are improving the SearchPoint technology (described in Section 3.4). This will include adding features to narrow down search results. We will also make it easier to use by providing pre-set searches and adding interactive data visualizations to help retrieve research and policy findings. Additionally, we will add data from submitted patents to capture innovation related to SDG 1 topics, as described in those patents. We are also enhancing this system by integrating LLM-based features to improve user interaction and make it easier to classify and extract information from the added documents.

We plan to build on existing policies and investigate news bias based on the latest findings from our work on cognitive bias in news, part of the European Commission's Horizon Europe project ELIAS – European Lighthouse for AI and Sustainability. Our team is also developing the Bias Detector Toolkit, a visual catalogue of tools for reducing bias, and has created an online course to help developers understand bias in AI. We aim to address challenges related to SDG 1 and explore different bias mitigation techniques. Our goal is to share these findings with researchers and developers, helping them identify and address bias in the SDG 1 Observatory.

ACKNOWLEDGMENTS

We thank the support of the European Commission-funded projects ELIAS - Lighthouse of AI for Sustainability (10080425) and RAIDO – Reliable AI and Data Optimisation (101135800). The authors also acknowledge the theoretical support of Andrej Srakar in regards to the discussions on indicators and causality, and the contribution of the AI Lab in Jozef Stefan Institute, that developed several of the technologies we use in this report (such as *Streamstory*, *Wikifier* and *Searchpoint*), and the news engine Eventregistry for the collaboration building together the pilot for news and exploring its potential.

5. REFERENCES

[1] Betti G., D'Agostino A., and Neri L. (2002) Panel regression models for measuring multidimensional poverty dynamics. Statistical methods and applications, 11, 359–369.

[2] Brady D. (2019) Theories of the causes of poverty. Annual Review of Sociology, 45, 1, 155–175.

[3] Corral P., Henderson H. and Segovia S. (2024). Poverty mapping in the age of machine learning. *Journal of Development Economics*, 103377.

[4] Eurostat (2024) EU SDG Observatory [online]: https://www.eesc.europa.eu/en/sections-other-bodies/observatories/sustainable-development-observatory (accessed in 4.10.2024)

[5] Gencat - Statistical Institute of Catalonia (2024) SDG Observatory of Catalonia [online]: https://www.idescat.cat/dades/ods/nu/?lang=en

[6] Hayati D. and Karami E. (2005) Typology of causes of poverty: the perception of iranian farmers. Journal of Economic psychology, 26, 6, 884–901.

[7] Instituto Brasileiro de Geografia e Estatística (2024) Indicadores Brasileiros para os Objetivos de Desenvolvimento Sustentável [online]: https://odsbrasil.gov.br/ (accessed in 4.10.2024)

[8] Leban G., Fortuna B., Brank J. and Grobelnik M. (2014) *Event registry: learning about world events from news*. In: Proceedings of the 23rd International Conference on World Wide Web, 107-110.

[9] Muse A.H., Hassan A.A. and Chesneau C. (2024) Machine learning study using 2020 sdhs data to determine poverty determinants in somalia. IEEE Transactions on Radiation and Plasma Medical Sciences, 14, 1, 5956.

[10] Neal J., Burke M., Xie M., Davis W. M., Lobell D. B. and Ermon S. (2016) Combining satellite imagery and machine learning to predict poverty. Science, 353, 6301, 790–794.

[11] Ng A.H., Farinda A. G., Kui Kan F., Lim A. L. and Ming Ting T. (2013) Poverty: its causes and solutions. International Journal of Humanities and Social Sciences, 7, 8, 2471–2479.

[12] Shah O. and Tallam K. (2023) Novel machine learning approach for predicting poverty using temperature and remote sensing data in ethiopia. IEEE Transactions on Radiation and Plasma Medical Sciences, 5, 6, 2302.14835.

[13] Stopar, L., Skraba, P., Grobelnik, M. and Mladenic, D. (2018). *Streamstory: exploring multivariate time series on multiple scales.* IEEE transactions on visualization and computer graphics, **25**(4): 1788-1802.

[14] UN Development Programme (2024) SDG Global Observatory [online]: https://sdgs.un.org/ (accessed in 4.10.2024)

[15] UN Economic Commission for Africa (2024) Africa UN Data for Development Portal [online]: https://ecastats.uneca.org/africaundata/ (accessed in 4.10.2024)

[16] Urbanč L., Pita Costa J., Rei L. and Grobelnik M. (2024) Predicting poverty using regression. In Proceedings of the Conference SIKDD2024

[17] Usmanova A., Aziz A., Rakhmonov D. and Osamy W. (2022) Utilities of artificial intelligence in poverty prediction: a review. Sustainability, 14, 21, 14238.

[18] World Bank (2024) Poverty and Inequality Platform Methodology Handbook [online]: https://datanalytics.worldbank.org/PIP-Methodology/ (accessed in 4.10.2024)

[19] World Bank: Can machine learning help us create a better poverty map? https://blogs.worldbank.org/en/developmenttalk/can-machine-learning-help-us-create-better-poverty-map

[20] Zixi H. (2021). Poverty prediction through machine learning. In 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT). IEEE, 314–324.