BIGGER IS BETTER. OR IS IT?

Lessons learned from using a Deep Neural Network on Big Data to estimate their Potential for Sustainability

Master's Thesis

Supervision:

PD Dr. Andreas Heinimann

Institute of Geography, University of Bern

Author:

Benjamin Andrea Schüpbach

Matr.No. 14-100-564

benjamin.schuepbach@students.unibe.ch

Master's Student in Geography

Institute of Geography, University of Bern

Bern, XX.XX.2020

Mein lieber Sohn, fliege nicht zu tief, damit die Federn nicht ins Meerwasser tauchen, sonst werden sie feucht und ziehen dich in die Tiefe. Fliege aber auch nicht zu hoch, sonst schmilzt die Sonne das Wachs, die Flügel fallen auseinander, und du stürtzt ab. Fliege die Mittelstrasse zwischen Meer und Sonne immer nur hinter mir her!

- *Daidalos und Ikaros*, Schwab (1990)

Abstract

Contents

1 Introduction

Explain how the introduction is structured.

1.1 Big Data

Introduction to Big Data.

This section offers an introduction to the basics of big data, big data analyses and the meaning of big data for efforts of sustainability.

1.1.1 Big Data Analyses

1.1.2 Big Data for Sustainability

1.1.3 title

1.2 Development Disparities

1.3 Sustainable Development

This section introduces the concept of Sustainable Development (SD). After a brief overview of its origins, some of the many important steps in the adoption of SD into today's global politics by the United Nations are highlighted before various key models are introduced. Finally some of the current measures used to globally advance efforts of SD are presented.

1.3.1 Origins of Sustainable Development

Ulrich Grober (2007) describes how today's notion of Sustainable Development originated from the concept of Sustainability. Grober further elaborates on how the term "Sustainability" was first introduced to the domain of forestry through Hanns-Carl von Carlowitz (1732) with his *magnum opus* "Sylvicultura Oeconomica" in which he described the necessity of a controlled and sustained use of timber. Timber was an essential resource at the time, and could not be substituted. According to Grober (2007:18), von Carlowitz criticized "the contemporary short-termed way of thinking which was centred solely on making money", thus emphasizing that society should assure a steady supply of timber through conservation and reforestation efforts in order to guarantee the continual and sustained use of the resource.

The following centuries saw authors like Thomas Robert Malthus (1926) and George Perkins Marsh (1965) as well as the Club of Rome (1972) publish concerns about human overpopulation, resource shortages and a possible system collapse of the world as it was. In The Limits to growth, Donella Meadows and the Club of Rome (1972:23) concluded that "if the present growth trends in world population, industrialization, pollution, food production, and resource depletion continue unchanged, the limits to growth on this planet will be reached sometime within the next one hundred years. The most probable result will be a rather sudden and uncontrollable decline in both population and industrial capacity". Jacobus A. du Pisani (2006) gives a comprehensive and detailed overview of this period in the history of the idea of Sustainable Development and the various theories on development and progress that preceded it.

More than 200 years would pass after Carlowitz' concerns until the modern notion of Sustainable Development was introduced formally into global politics. Michael Redclift (2005) explains that through the report on global environment and development by the *Brundtland Commission*, or "World Commission on Environment and Development" (1987), the term "Sustainable Development" was introduced into political vocabulary. Gro Harlem Brundtland (1987:292), who headed the commission, defines Sustainable Development as development that meets "[...] the needs and aspirations of the present generation without compromising the ability of future generations to meet their needs".

1.3.2 Sustainable Development as a Geopolitical Paradigm

The Brundtland definition is the cornerstone of Sustainable Development as it is known today. And while it was the Brundtland Report that introduced SD into political agendas arond the world, the need for specific, quantifiable goals to work towards arose (Du Pisani, 2006). Shantayanan Devarajan et al. (2002) as well as David Hulme (2009) illustrate the progression from just the idea of SD, through major stepping-stones like the *United Nations Conference on Environment and Development* in Rio de Janeiro (in 1992), the *International Conference on Population and Development* in Cairo (in 1994) and the *World Summit on Social Development* in Copenhagen (in 1995), to the first major global development framework: the Millennium Development Goals (MDGs).

The MDGs were introduced in September 2000 at the *United Nations Millennium Summit* in New York City (UN General Assembly, 2000). These Goals were aimed at issues of poverty, hunger, primary education, gender equality, child mortality, maternal health, preventable diseases, environmental sustainability and a global partnership for development. Towards the end of the 15 year period of the MDGs, Jeffrey David Sachs (2012:2206) states that "developing countries have made substantial progress towards achievement of the MDGs, although the progress is highly variable across goals, countries, and regions". He further explains how the world has entered a new geological epoch in which human activity has become the most dominant force in fundamental earth dynamics.

While this notion is not universally accepted, it illustrates a further shift in societal consciousness towards Sustainable Development (Heikkurinen et al., 2019). Sachs (2012:2207) further argues that "in view of [...] dire and unprecedented challenges, the need for urgent, high-profile, and change-producing global goals should be obvious". The MDGs brought sustainability onto the global political main stage and turned Sustainable Development into a geopolitical paradigm. Yet most of the challenges the MDGs were addressing persisted at least to some degree past their expiration date (Sachs, 2012).

1.3.3 17 Goals to Transform Our World

Because of these persistent challenges, the UN General Assembly adopted the new 2030 Agenda for Sustainable Developments on 25 September 2015 (United Nations, 2018). The new agenda consists of 17 goals and originally included 169 subordinate targets, making it the most extensive global development framework to date (UN Statistics Division, 2019c). Compared to the MDGs, the SDGs thus cover more dimensions of development more specifically (see table X). In March of 2018 as well as one year later in March of 2019, the list of indicators for the SDGs was expanded to 232 total indicators (UN Statistics Division, 2019a). All indicators are classified in tiers that determine their conceptual clarity and progress towards methodological standards for data collection:

Goal	SDGs (2015-2030)	MDGs (2000-2015)
1	No Poverty	Eradicate Extreme Poverty and Hunger
2	Zero Hunger	Achieve Universal Primary Education
3	Good Health and Well-Being	Promote Gender Equality and Empower Women
4	Quality Education	Reduce Child Mortality
5	Gender Equality	Improve Maternal Health
6	Clean Water and Sanitation	Combat HIV/AIDS, Malaria and other Diseases
7	Affordable and Clean Energy	Ensure Environmental Sustainability
8	Decent Work and Economic Growth	Global Partnership for Development
9	Industry, Innovation and Infrastructure	
10	Reduced Inequalities	
11	Sustainable Cities and Communities	
12	Responsible Consumption and Production	
13	Climate Action	
14	Life Below Water	
15	Life on Land	
16	Peace, Justice and Strong Institutions	
17	Partnerships	

Table 1: Comparison between SDGs and MDGs.

> **Tier I:** [The] indicator is conceptually clear, has an internationally established methodology and standards are available, and data are regularly produced by countries for at least 50 per cent of countries and of the population in every region where the indicator is relevant.
>
> **Tier II:** [The] indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries.
>
> **Tier III:** No internationally established methodology or standards are yet available for the indicator, but methodology/standards are being (or will be) developed or tested.
>
> (UN Statistics Division, 2019b)

1.3.4 The Rural Access Index

The main focus of this thesis is on SDG 9, indicator 1.1: *Proportion of the rural population who live within 2 km of an all-season road* (see chapters X, X and X). Until December 31 of 2018 indicator 9.1.1 was classified as a

tier III indicator (SDSN, 2015). Today, it is classified as a tier II indicator, eventhough the methodological approach to gather data for indicator 9.1.1 has been around since 2006 (UN Statistics Division, 2019b). The indicator was first introduced by Peter Roberts et al. (2006) as the Rural Access Index (RAI) in the context of the Results Measurement System of the International Development Association.

> "In practice the RAI *measures the number of rural people who live within two kilometers* (typically equivalent to a walk of 20-25 minutes) *of an all-season road as a proportion of the total rural population.* An "all-season road" is a road that is motorable all year round by the prevailing means of rural transport (typically a pick-up or a truck which does not have four-wheel-drive). Occasional interruptions of short duration during inclement weather (e.g. heavy rainfall) are accepted, particularly on lightly trafficked roads."
>
> Roberts et al. (2006:2)

According to Roberts et al. (2006:4), RAI should be measured "by analysis of household surveys that include appropriate questions about access to transport. The aim is to integrate this with the measurement of household characteristics such as income and access to services such as education, health and clean water supply". Although this methodological approach has since been updated for the application as SDG target indicator 9.1.1, RAI is still considered to be among the most important global development indicators of the transport in the Metadata-Repository of the SDGs (UN Statistics Division, 2019c).

Today's official methodological approach suggested by the Transport & ICT Report (2016) titled *Measuring Rural Access: using new technologies*, uses a combination of geospatial data (more specifically population distribution data, urban extent data, vectorized road data, measurements of road utility status) with a final spatial resolution of 100mx100m as opposed to data from household surveys. The requirements for the calculation of RAI are sturctured into three seperate data requirement domains. It is in the domain of measuring road condition (data requirement 3) where this thesis aims to make a contribution (see section X). Therefore, data requirement 1 (population distribution data) as well as data requirement 2 (urban extent

data and vectorized road data) are not introduced further in this section. WHAT IS RAI GOOD FOR, WHAT DOES IT SAY AND WHY DO WE NEED TO KNOW ABOUT IT?

1.3.5 Challenges in Measuring Road Condition

The following section elaborates on remaining challenges concerning costs, information delay and expenditure of human labour in the Transport & ICT Report. Table X gives an overview of what the Transport & ICT Report suggests as suitable sources for road condition data, along with some of their respective advantages and disadvantages.

Apart from "Free Apps for Road Assessment", all of the potential data sources are liable to pay costs. The Transport & ICT Report (2016:22) states that "it is always possible to collect the necessary condition data with reasonable accuracy, although at a cost". The costs for data collection are directly linked to the availability of data and the processing steps needed to extract relevant information, as well as initial investments for equipment (e.g. high initial investment costs for unmanned aerial drones).

In terms of information delay, satellite imagery and unmanned aerial drones offer a lot of flexibility. Meanwhile, data from road inventory surveys, call detail records (georeferenced information about calls made/received, owned by cell phone carriers) and applications for road surface assessment rely on the frequency of surveys, recorded drives or calls made for timeliness. Depending on these factors, data availability may be good or lagging behind.

All potential data sources for modern RAI calculation are to some degree labour intensive. International, standardized procedures using satellite imagery, unmanned aerial drones, call detail records or data gathered through mobile applications can reduce initial and upkeep costs, however.

Dobermann and Nelson 2013: Dobermann, A. and Nelson, R. et al. (2013). Solutions for Sustainable Agriculture and Food Systems. Technical report of the Thematic Group on Sustainable Agriculture and Food Systems. Paris, France and New York, USA: SDSN.

Data Source	Advantage	Disadvantage
Road Inventory Survey	Technically solid, consistent with government responsibility	Costsly, Irregular updates, country-specific assessment standards
Satellite Imagery	Consistency across countries, potential for high frequency data collection	Costs, Technically challenging to identify road condition in detail, significant computational process required
Unmanned Aerial Drones	Good mobility	Technically challenging, computational process required
Call Detail Record	Consistency across countries, potential for high frequency data collection	Access to data, noise in data
Free Apps for Road Assessment	Cost effective, Potential contribution through crowd-sourcing	Statistical errors between measured IRI and actual roughness
Commercial Apps for Road Assessment	Relevant analytical tools provided together	Statistical errors between measured IRI and actual roughness

Table 2: Summary of possible sources for road condition data (Transport & ICT, 2016:23).

1.4 Image Classification

1.4.1 Deep Neural Networks

1.4.2 YOLO & Darkflow

1.4.3 title

1.4.4 title

1.5 Goals of this Study

Show potentials of big data in combination with machine learning for indicators of SDGs.

1.5.1 Research Questions

In this section, research questions based on the goals of this study are formulated. Research questions 1 and 1.1 are directly linked to target indicator 9.1.1 of the SDGs (see section X). Research question 2 is oriented towards the potential overall contribution of Big Data for Sustainability.

Research Question 1: Can georeferenced data for indicator 9.1.1 (RAI) of the SDGs be generated using a Deep Neural Network on the Twitter Streaming API?

(If answer to RQ1 is yes:) **Research Question 1.1:** Are these data comparable to conventional data for indicator #58 of the SDGs in terms of quality and accuracy?

Research Question 2: What are potentials and limitations of Big Data analyses for the monitoring of the SDGs?

2 Methods

This section gives an overview of the methods used to conduct this study. It is structured chronologically in the sense of data processing steps (see also figure X). First, the potential data sources are introduced. Second, the process of harvesting the images used to train the image recognition algorithm for road utility status (ICARUS) is explained, as well as the methods used to train the algorithm. In the third subsection validation procedures used are shown. The fourth part explains how ICARUS was run. The final section covers how the results form running ICARUS were mapped. All of the (python) scripts, resources and outputs generated throughout this process are available and documented on the author's GitHub page.

Explain Flowchart here

2.1 Data Source

For this thesis, the Twitter streaming application programming interface (Twitter streaming API) was used as the single data source, as at the time it was the only streaming API for social media which was accessible without request limitations or prior submission of an application to use it with. The streaming API offers an inherent option to filter for Tweets with attached geographical coordinates in longitude/latitude format. A second filter was implemented to additionally filter the stream for Tweets which also had media appended to them. Thus a CSV file could be generated that recorded coordinates, the URL of appended media and information on the date and time of a published tweet. Such data was collected globally, encompassing every published tweet during the time in which this study was conducted (May - XXXX 2019).

2.2 Training the Classifier

To generate a set of training images, the Twitter streaming API was combined with the Google Cloud Vision API. Upon signing up with the Vision API, users are granted a free credit with which to test the API, which was used to filter out images from twitter containing asphalt roads. Upon expiration of this free credit, a set of 5000 training images and a further set of 200 validation images were generated and manually checked for mistakes. Once

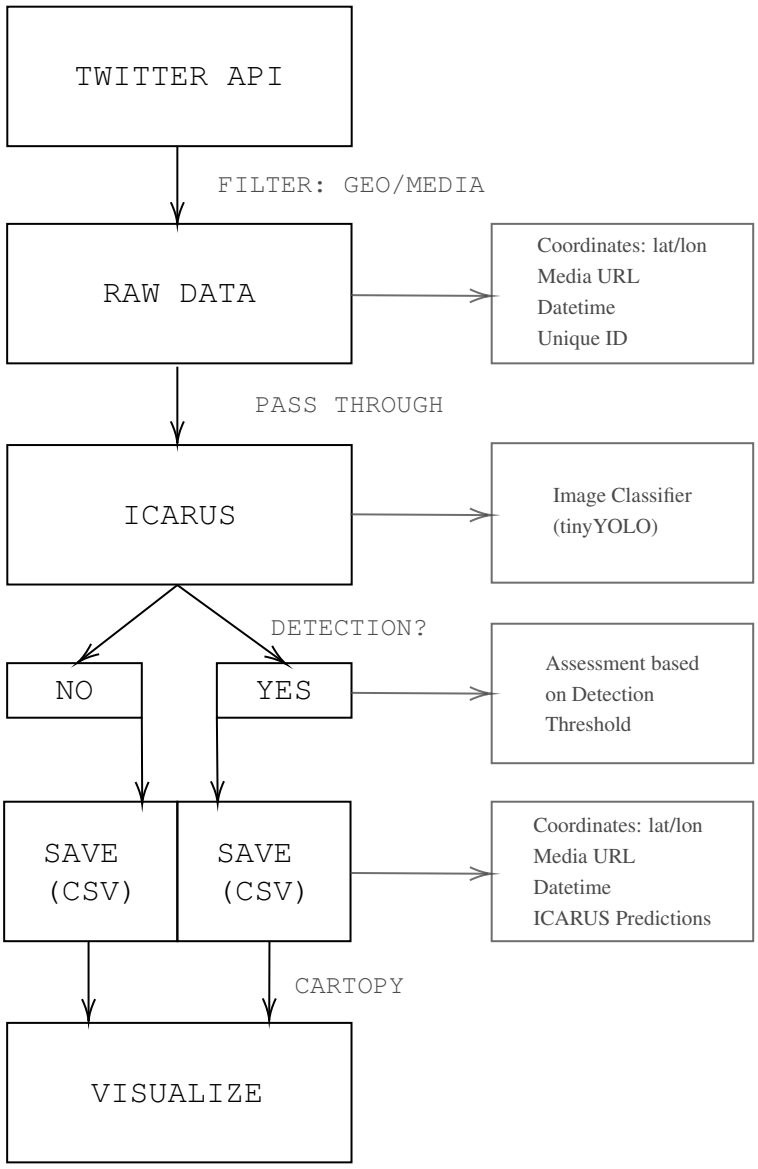

Figure 1: Flowchart of Methodology used in this Study.

generated, both image sets were classified by manually drawing bounding boxes around areas in the images containing asphalt roads. In total, more than 18 000 areas were labeled this way.

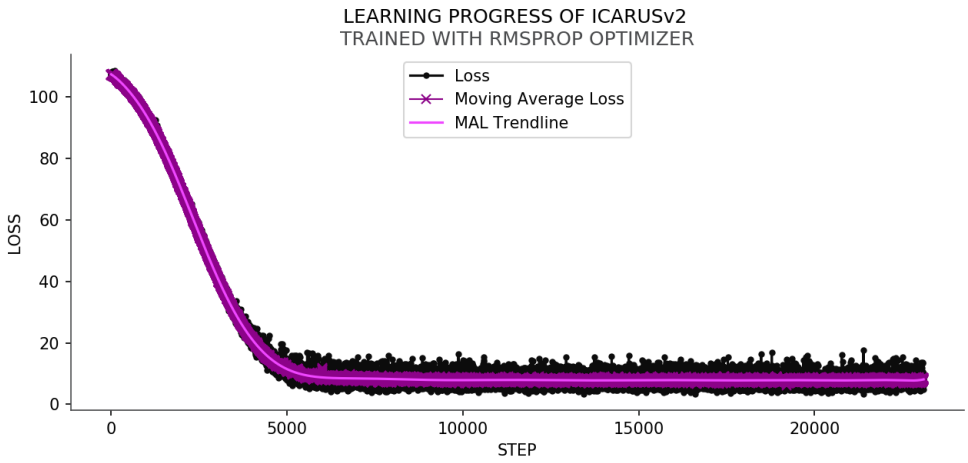

Figure 2: Learning Progress of ICARUS

Because of hardware limitations the DNN architecture of tiny-YOLO was chosen for ICARUS (see also section X.X.X). With darkflow, ICARUS was trained using the RMSPROP optimizer. Through adjustments of the learning rate whenever the step loss reached a plateau, a final loss of around 4-5 was reached (see fig. X). It should be noted that because of aforementioned hardware limitations, the maximum batch size possible was 16. This procedure was done multiple times to test different optimizers as well as training parameters, hence the versioning (ICARUSv2) in figure X. For all of this thesis, the acronym ICARUS references this version.

2.3 Validation

2.3.1 Validation with Google Streetview

Quick proof of concept was done using google street view API and building a 10m by 10m raster for API queries, then looking at wheter ICARUS labeled anything other than roads.

2.3.2 Mean Average Precision

Real validation was done on the validation image dataset. Using darkflow, bounding boxes of predictions on the validation images were created. these were then imported into a script to calculate mean average Precision.

2.4 Running ICARUS

This section is about the process of actually running ICARUS. This was done for input data of one month, while data for X months was gathered.

2.4.1 Gathering Actual Data for Assessment

Because ICARUS could not run indefinitely on the author's desktop PC, and no server was used to execute all of the steps in figure X, the processes of gathering data and assessing data were split up. As a cost and resource efficient solution, a Raspberry Pi 3 (RasPi) was deployed to filter the Twitter streaming API. It has a much lower energy consumption than a regular Desktop PC.

With the RasPi deployed, the API could be scanned around the clock. Through the *Remote Desktop Connection* application, data gathering was monitored. Sporadically, data harvesting was manually suspended for a few seconds which meant the RasPi had to generate a new savefile to write into. This was done to avoid total data loss in case severe complications occurred and a savefile got corrupted. For convenient further use, the split savefiles were later re-joined.

2.4.2 Image Classification with ICARUS

As mentioned above, the harvest savefile contains a URL for media appended to each tweet. For the classifying step, ICARUS iterates over each URL, downloads the associated media and classifies its contents. To manage the toll this takes on disk space, media attachments are overwritten with information from the next Tweet each time a new classification step begins, effectively only saving one image at a time, resulting in minimal disk space requirements. For each step (each time a new Tweet is assessed), ICARUS saves coordinates, URL, date and time, average prediction confidence and a dictionary of all predictions (including bounding box coordinates for each prediction) which are saved into a separate, output CSV file.

Parameter	Input
Input Feature	Consolidated Streaming API Harvests
Population Field	None
Output Cell Size	0.1
Search Radius	2
Area Units	Square Kilometers
Output Cell Values	Expected Counts
Method	Geodesic

Table 3: Parameters used to calculate Kernel Density of Harvests in ArcGIS Pro.

2.5 Visualization

This section explains how results from both the Twitter streaming API and ICARUS were visualized. All visualizations in this thesis were made using the matplotlib and numpy libraries for python. For mapping purposes, the cartopy library was used. All geographically projected visualizations are projected in EPSG 32662 with a central longitude of 0.0.

2.5.1 Visualizing Harvested Data

To visualize all harvested tweets, two approaches were used. First, all harvested tweets with geotag and appended media were drawn onto a worldmap as squares at 20% opacity. This approach was chosen to illustrate the absolute amount of harvested data. Even though features were drawn with opacity, there were so many of them that the intended effect (visualizing density as a side product of opacity) was not achieved, as there is a lot of overlap of single features when mapped like this.

Therefore, the second approach of calculating kernel density (KDE) was used. This step was executed in ArcGIS Pro, as KDE calculation with matplotlib and scikit-learn resulted in a longitudinal distortion in the computed density layer. It should be noted, that while correcting this distortion would be possible in a python environment, ArcGIS Pro was used to save the time needed to implement such a correction. From ArcGIS Pro, a layer was then

Figure 3: Visualization of Predictions with ICARUS.

exported and fed back into the python script for further mapping. To adequately represent density hot spots and account for the uneven global distribution in the harvested dataset, a logarithmic colormap was used to illustrate data density. A summary of the parameters used to conduct the kernel density analysis is provided in table X.

2.5.2 Visualizing Predictions

As mentioned above, ICARUS saves a dictionary of predictions (including prediction confidence(s) and bounding box information for each prediction) for each assessed data point. Additionally, a value for mean prediction confidence is calculated for each data point. Using bounding box coordinates, exact parts of an image that a classification is based on can be visualized with their respective prediction confidence (see fig.X & X).

Predictions from ICARUS were visualized in the same environment as mentioned in section X.X.X. To color-map prediction confidence on the world maps, the aforementioned mean prediction confidence was used. As explained in section X.X.X, the total amount of predictions form ICARUS (based on 1 month of harvests) did not warrant a similar procedure to calculate kernel density as used for harvest mapping. However, if ICARUS were integrated

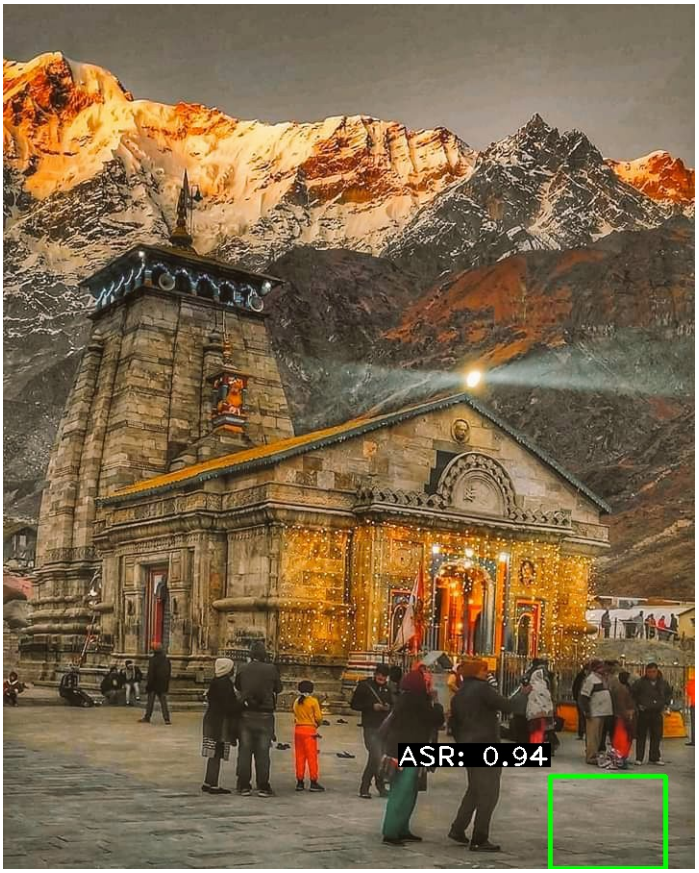

Figure 4: Visualization of Predictions with ICARUS.

into the harvesting process, such a calculation would be possible and necessary.

3 Results

Figure 5: Map of Tweets where ICARUS identified AllSeasonRoads

4 Discussion

4.0.1 Relevance & Shortcomings

4.0.2 Importance to Sustainable Development

5 Conclusion & Outlook

5.0.1 title

ADD { } TO BIBLIOGRAPHY ENTRIES THAT AREN'T DISPLAYED CORRECTLY IN THE .BIB FILE

References

Assembly, U. G. (2000). United Nations Millennium Declaration. https://undocs.org/A/RES/55/2.

Brundtland, G. H. (1987). Our Common Future—Call for Action*. *Environmental Conservation*, 14(4):291–294.

Devarajan, S., Miller, M. J., and Swanson, E. V. (2002). *Goals for Development: History, Prospects, and Costs*. The World Bank.

Division, U. N. S. (2019a). Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. https://unstats.un.org/sdgs/indicators/indicators-list/.

Division, U. N. S. (2019b). IAEG-SDGs — Tier Classification for Global SDG Indicators. https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/.

Division, U. N. S. (2019c). SDG Indicators — Metadata Repository. https://unstats.un.org/sdgs/metadata/.

Du Pisani, J. A. (2006). Sustainable development – historical roots of the concept. *Environmental Sciences*, 3(2):83–96.

Grober, U. (2007). Deep roots-a conceptual history of'sustainable development'(Nachhaltigkeit).

Heikkurinen, P., Ruuska, T., Wilén, K., and Ulvila, M. (2019). The Anthropocene exit: Reconciling discursive tensions on the new geological epoch. *Ecological Economics*, 164:106369.

Hulme, D. (2009). The Millennium Development Goals (MDGs): A short history of the world's biggest promise.

ICT, T. . (2016). Measuring Rural Access : Using New Technologies. Technical Report 107996, The World Bank. http://documents.worldbank.org/curated/en/367391472117815229/Measuring-rural-access-using-new-technologies.

Malthus, T. R. and Bonar, J. (1926). *First Essay on Population, 1798*. Macmillan & Co. Ltd. (1926 edition with notes by James Bonar), London. OCLC: 2710982.

Marsh, G. P. (1965). *Man and Nature: Or, Physical Geography as Modified by Human Action*. Harvard University Press, Cambridge. OCLC: 952754139.

Meadows, D. H. and Club of Rome, editors (1972). *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Universe Books, New York.

Redclift, M. (2005). Sustainable development (1987–2005): An oxymoron comes of age. *Sustainable development*, 13(4):212–227.

Roberts, P., KC, S., and Rastogi, C. (2006). Rural Access Index: A Key Development Indicator.

Sachs, J. D. (2012). From millennium development goals to sustainable development goals. *The Lancet*, 379(9832):2206–2211.

Schwab, G., editor (1990). *Die schönsten Sagen des klassischen Altertums*. Number 500 in Goldmann-Taschenbuch. Goldmann, München, 25. aufl edition. OCLC: 635974581.

SDSN, S. D. S. N. (2015). Indicators and a Monitoring Framework for Sustainable Development Goals: Launching a data revolution for the SDGs. Technical report. http://unsdsn.org/resources/publications/indicators/.

United Nations (2018). The Sustainable Development Agenda. https://www.un.org/sustainabledevelopment/development-agenda/.

von Carlowitz, H.-C. (1732). *Sylvicultura Oeconomica*, volume 1. Bey Johann Friedrich Brauns sel. Erben.

WCED, W. C. o. E. a. D. (1987). *Our Common Future*. Oxford University Press, Oxford, New York.