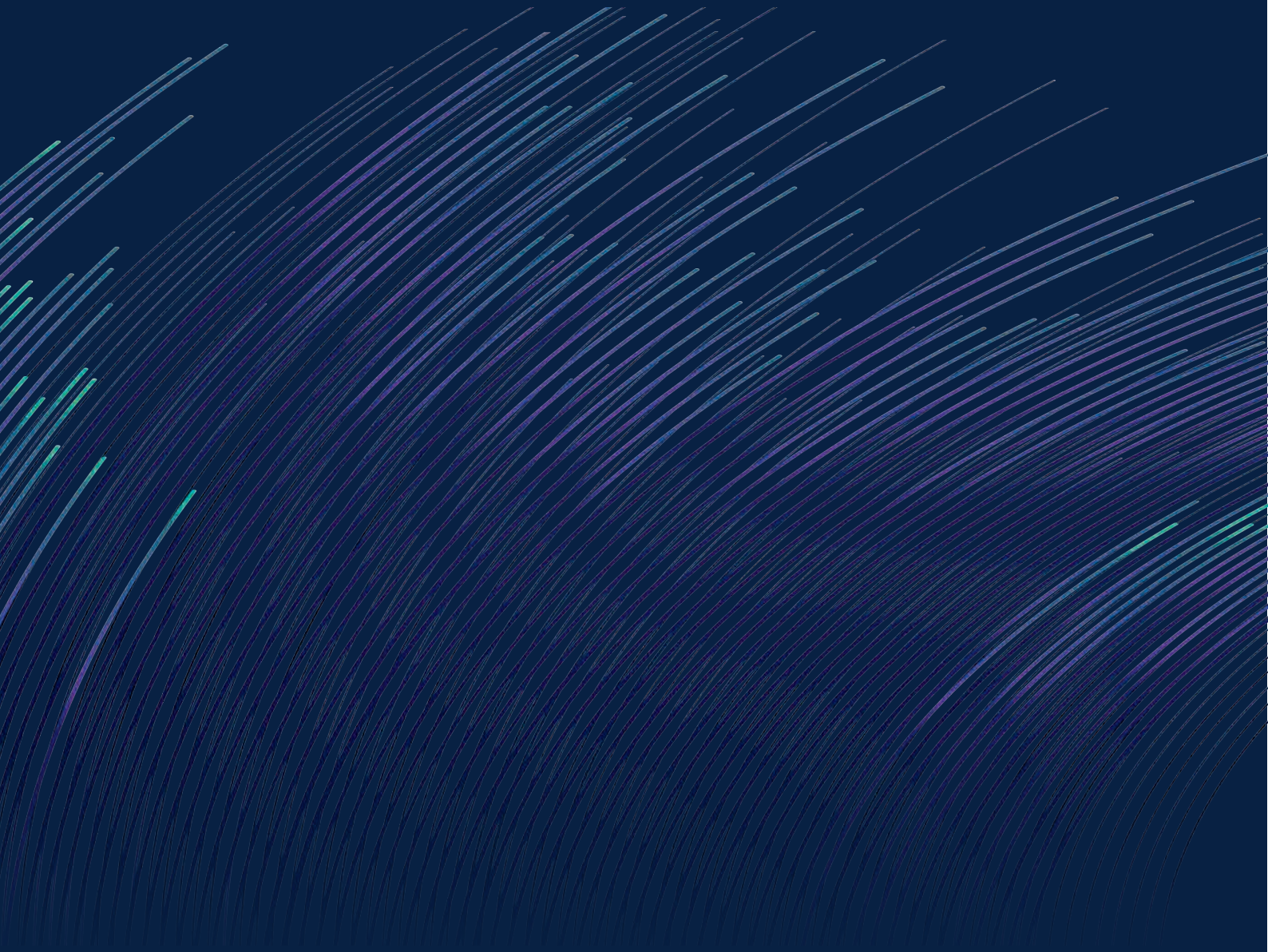
United Nations Development Programme



Digital Social Vulnerability Index Technical Whitepaper

2024

Digital Social Vulnerability Index

UNDP is the leading United Nations organization fighting to end the injustice of poverty, inequality, and climate change. Working with our broad network of experts and partners in 170 countries, we help nations to build integrated, lasting solutions for people and planet.

Learn more at undp.org or follow at @UNDP.

Table of Contents

List of Tables

List of Figures

Annex

Acknowledgments

This report was developed by the United Nations Development Programme (UNDP) Istanbul International Centre for Private Sector in Development (ICPSD) SDG AI Lab. The conceptualization and development of the report was led by ICPSD Technical Specialist Gökhan Dikmener, GIS Specialist Martin Szigeti and assisted by ICPSD Coordination and Partnership Analyst Dina Akylbekova.

Research team:

Martin Szigeti
Gökhan Dikmener
Dina Akylbekova
Ivana Petraković
Yücel Torun

Peer reviewers

Aslıhan Albostan, Engineer (PhD) and Former Portfolio Lead (2018-2022), UNDP ICPSD;

Cem Bayrak, Inclusive and Sustainable Growth Projects Manager, UNDP Türkiye;

Mihail Peleah, Programme Specialist Green Economy and Employment, Inclusive Growth Team, UNDP Istanbul Regional Hub;

Zebo Jalilova, Team Leader, Sustainable Economic Development, UNDP Tajikistan.

The report team would like to thank the partner organizations and the peer reviewers for their engagement and feedback. UNDP ICPSD would also like to express its gratitude to Mihail Peleah, Programme Specialist Green Economy and Employment, UNDP Istanbul Regional Hub Inclusive Growth Team, for the guidance and insight he offered throughout the research process.

The manuscript was edited by The Word Pavilion and designed by Baisalykov.

Executive Summary

At the cutting edge of frontier technology – DSVI – innovative product for social vulnerability estimation

DSVI is a collection of technologies to unlock the full potential of digital social vulnerability (SV) assessments. The *Digital Social Vulnerability Index* Technical Whitepaper is a product of the UNDP International Center for Private Sector in Development's (ICPSD) SDG AI Lab, supported by the Disaster Risk Reduction and Recovery (DRT) for Building Resilience Team. The SDG AI Lab is a joint initiative of the UNDP Bureau for Policy and Programme Support (BPPS) teams, and it is hosted under UNDP ICPSD. The Lab has a mission to harness the potential of frontier technologies, such as artificial intelligence (AI), machine learning (ML) and geographic information systems (GIS) for sustainable development. The SDG AI Lab provides research, development and advisory services in the areas of frontier technologies and sustainable development. The Lab also supports UNDP's internal capacity-strengthening efforts for the increasing demand for digital solutions.

The DSVI technical whitepaper explains the rationale, benefits, outcomes, methodologies and the relevance of the Digital Social Vulnerability Index. It can be regarded as a technical manual that describes the process of SV calculations with innovative methods, such as GIS and machine-learning technologies. The paper critically evaluates the current vulnerability assessment space and proposes various ways for its improvement. Some suggestions include improvements to the calculation process of SV, the datasets, preprocessing of datasets, usage of machine learning to produce high-resolution maps and the implementation of online tools to display the results.

DSVI is a high-quality digital solution for vulnerability assessments to monitor and understand the exact location, distribution and underlying drivers of social vulnerabilities. While previous vulnerability measures would require conducting timely and costly surveys, the DSVI provides a higher resolution and improved representation of a country's social vulnerability beyond administrative boundaries. Moreover, compared with previous instruments, the DSVI is the first tool of its kind to incorporate a much more comprehensive SV analysis by integrating numerous data sources and indices into one.

Tailored to the specific needs of UNDP, government agencies and other development actors, DSVI can provide effective digital SV analyses with an implementation time frame of two to three months per country. It seeks to help UNDP national or regional offices to improve their understanding of local vulnerabilities, thereby facilitating the adoption of more targeted and coordinated interventions that build stronger community resilience. DSVI offers a set of outputs which help to deliver the key messages and data to its target audiences. These elements start with raw datasets and scientific methods for calculation and end with training sessions, maps, reports or digital infrastructure to visualize the findings.

First, DSVI offers high-accuracy SV scores. SV scores are calculated by an automated data science pipeline which gathers high-quality and freely available raw data from USAID, the UN and scientific resources. Using GIS and machine learning, DSVI generates high-resolution vulnerability maps. These maps are a new, more technologically advanced addition to the long tradition of vulnerability mapping.

The DSVI *online tool* is a web application. The web implementation features data layers that present SV (exact location points, data aggregated to administrative boundaries, high-resolution maps); data layers that show socio-economic and biophysical properties (distance to critical infrastructure, biophysical and socio-economic parameters, disaster-related data like drought indices); survey information with filter functions; and tools for contextual spatial analysis (e.g. position of a group relative to a disaster).

In addition, DSVI offers users training sessions. These sessions inform and educate audiences on its core use and findings.

In summary, DSVI offers to:

1. Make vulnerability data widely available and actionable. DSVI provides easy access to a series of vulnerability datasets. The web application uses machine learning to generate new data for any identified gap. DSVI offers various outputs that help organizations to better understand vulnerability and to visualize it together with the underlying drivers.
2. Ways to allocate resources more equitably. With a better understanding of social and spatial distribution of vulnerability, practitioners can allocate resources more effectively and help to leave no one behind.
3. Improve the use of SV data for long-term development planning. The DSVI tool improves SV knowledge by highlighting vulnerable areas within societies in particular and by creating an advanced tool that facilitates practitioners in their quest to address fundamental risk factors.
4. Contribute better to crisis action. The SV scores provide valuable insights on social and environmental vulnerabilities in target areas. By bringing together vulnerability data in one platform, practitioners can produce digestible information that enhances the overall understanding of the underlying drivers of risk.
5. With DSVI and the technical whitepaper, the SDG AI Lab also fulfils the technical and programmatic needs of UNDP for innovative technologies, such as those mentioned in the Strategic Plan 2025. The paper will inform stakeholders and policymakers on how they can integrate these new technologies into their programmes. This will be accomplished with an in-depth explanation of the DSVI methodologies, featuring high-quality visualizations and step-by-step guidelines to recreate the results. DSVI and its outputs can be utilized to make more risk- informed and targeted decisions for vulnerable population groups.

1. Introduction

The introduction will cover all key concepts regarding social vulnerability, its significance for the United Nations and for Digital Social Vulnerability Index before moving on to the methodological and technical topics.

1.1 Social vulnerability concept

In recent years social vulnerability programmes and policies were of growing interest for the United Nations Development Programme (UNDP) and other organizations. UNDP not only acknowledges the importance of vulnerability measures, but also the fact that they are underresearched.[1] Social vulnerability concepts and products support the achievement of the 2030 Agenda for Sustainable Development Goals, as well as the UNDP Strategic Plan 2022-2025.

Vulnerability indices in several formats have been published and researched in recent years, such as the Multidimensional Poverty Index (MPI),[2] or other vulnerability assessments specifically for climate change or disaster-related topics. These measures, however, do not cover all the dimensions of vulnerability and cannot reach the desired resolution necessary to make highly targeted programmatic decisions on the ground and in areas with low data coverage.

Social vulnerability is the differential capacity of individuals or communities to cope with social and environmental shocks (Adger 2000; Cutter et al. 2003). This includes climate change, natural disasters and other societal risks. Vulnerable groups have a disproportionate risk of being affected and experiencing more profound consequences due to their socio-economic preconditions. SV assessments help to better map the interconnections between local conditions, social characteristics, or individual vulnerabilities and risks.

The calculation of SV scores is a frequent practice to measure a community's ability to respond to outside stressors and risks. It is an indirect way to quantify resilience. Having such an assessment helps to understand, prepare for and respond in a more effective manner by using a combination of the most appropriate tools once the risk materializes.

Social vulnerability maps and data products are a powerful way to understand the distribution of vulnerable population groups in a region. They can be used to visualize their exposure to risk and to allow a targeted response by facilitating planning and strategic activities.

[1] Policy Brief, Climate Change and Social Vulnerability, Arab Water Council, World Food Programme with Swedish International Development Cooperation Agency support, November 2022, https://www.undp.org/sites/g/files/zskgke326/files/2023-04/Policy_Brief.pdf

[2] UNDP and Oxford Poverty and Human Development Initiative, Global Multidimensional Poverty Index (MPI) 2022, https://hdr.undp.org/content/2022-global-multidimensional-poverty-index-mpi#/indicies/MPI

1.2 Digital Social Vulnerability Index (DSVI)

Based on the need for effective digital tools and the multipurpose applications of social vulnerability indexes, the ICPSD SDG AI Lab developed a digitized approach for the Digital Social Vulnerability Index calculation. DSVI is in line with the UNDP strategic plan that seeks to develop integrated development solutions through digitalization and strategic innovation to alleviate poverty and inequality, strengthen resilience, promote gender equality and more. The approach is in accordance with the recent UNDP call for innovative digital solutions which brings together the scientific and UNDP internal methodologies for SV calculations. It will help to reduce the time and costs needed for previously used non-digital SV calculations.

The non-digital procedure typically comes with various disadvantages: according to the UNDP *handbook 'Social Vulnerability Assessment Tools for Climate Change and DRR Programming'*,[3] SV indexes are usually a conglomerate of primary and secondary data, akin to the surveys or censuses collected by private entities or public authorities.

These datasets are often complemented with paper, web or telephone-based data inputs, which usually come in a myriad of different quality levels and types. These efforts need to be planned and executed by experts and statisticians in a very timely and cost-intensive procedure to ensure a standardized format and high quality.

Other potential limitations are that the resulting SV maps usually have an aggregated format and not a high spatial resolution to show local changes. DSVI was designed to improve the existing methodology and to bring them to the next level.

1.3 Main benefits of DSVI and value proposal

DSVI utilizes the UNDP guidelines for SV calculations and uses geographic information systems (GIS) combined with machine learning to enhance the resolution of SV in an innovative way. Most functions are automated, and the results can be reflected in a web-based modern digital online tool. The envisioned implementation saves costs and allows faster and more reliable SV calculations. This is achieved with a long list of available datasets and peer-reviewed methods. With this strategy, we reduce the need for new surveys by using already available, geotagged survey data[4] from online resources. We introduce automated data pre-processing pipelines which allow control over the workflow, yet grant the freedom to adjust the modelling parameters if needed.

Next to the already mentioned operational benefits, DSVI directly impacts knowledge products and policy decisions. The benefits of using DSVI are reflected in the possibility of identifying areas of high vulnerability, of assessing their main drivers and of planning actions accordingly. It is a valuable means for stakeholders, policymakers and other responsible parties to target specific areas, to make efficient and relevant decisions with the goal of reducing social vulnerability and to contribute further to meaningful development efforts.

[3] Krunoslav Katic, A Guide to Practitioners. Social Vulnerability Assessment Tools for Climate Change and DRR Programming, UNDP, September 2017.

[4] See data section for more information.

1.4 Expected outcomes

With the introduction of DSVI, we aim to achieve a handful of improvements and outcomes which are core elements of the product. First, the calculation of Digital Social Vulnerability Index scores is conducted on three different levels of detail. The lowest level of detail is the administrative boundaries level, which represents social vulnerability scores aggregated to administrative boundaries. Secondly, based on the exact locations provided by the used geotagged datasets, the social vulnerability scores can be viewed at the household level. Thirdly, the high-resolution predictions of social vulnerability are achieved using geodatasets, coming with a pixel size of a few hundred meters or less and without data gaps.

These vulnerability maps reach every border region of a country studied and can also enhance the understanding of the vulnerability situation in border regions, especially when conflicts in those regions are current ones. The results are then contextualized and visualized in a digital tool. This digital tool, a web application, uses modern technologies and grants users control over modelling parameters, variables and data layer selection, visualization and analysis. DSVI also provides training sessions for practitioners, policymakers and technical audiences to explain and discuss the results.

2. Data and Methodology

This section gives a detailed overview of the scientific process of DSVI calculations. The provided methodology and examples are based on the DSVI calculations performed for Albania and Tajikistan.

2.1 Workflow

A high-level overview of the data processing and modelling steps is given in the following flowchart:

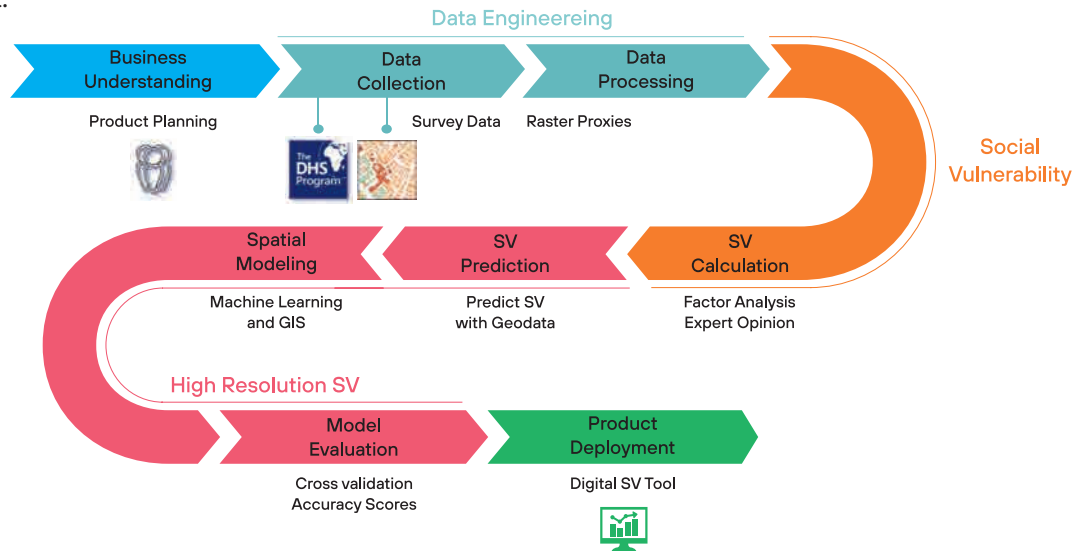


Figure 1. High-level flowchart of data science workflow

The development process is aligned with industry standards and follows all necessary requirements to ensure a high-quality result. DSVI is developed by a multidisciplinary team composed of data scientists and technical experts. Data quality and technical processes are evaluated by experts and subject to constant improvement.

The Business Understanding components were already explained in the previous section of this whitepaper. Data collection and processing are highly automated and standardized according to the available datasets per studied country. Automated scripts handle survey and spatial data to be transformed for modelling purposes (*Data Engineering, turquoise cells*). SV scores are calculated with domain knowledge and expert input to increase the quality of the resulting indices.

The calculation follows scientifically accepted procedures and guidance from the UNDP handbook (*Social Vulnerability, orange cell*). In the next step, we predict SV with the help of geographical data, modern GIS and machine-learning techniques (*High Resolution SV, red cells*). The results, maps and the data can be viewed and interactively evaluated in the 'DSVI online tool' developed by the SDG AI Lab.

2.2 Data collection and data sources

DSVI needs three different data inputs in order to deliver its proposed components:

1) **Survey data,**[5,6] or high-quality data, with the socio-economic dimensions of vulnerability. Additional survey/census data from other sources can be used if it contains geolocations.[7]
2) **Spatial data** that can be used to predict SV for areas without survey coverage.
3) **Domain knowledge** to identify the composition of variables, indicators and other influencing factors relevant for country specific circumstances.

2.2.1 Data collection: Survey data

We piloted DSVI using USAID's Demographic and Health Survey (DHS) data. DHS data can be downloaded on their homepage after submitting a request. The DHS Programme is authorized to distribute unrestricted survey data files for legitimate academic research at no cost. DHS survey data is available for more than 90 developing and threshold countries. The datasets contain hundreds of variables covering dimensions of income, employment status, access to infrastructure, health, violence, gender equality, race, age and more.

These variables are collected from thousands of individuals and standardized into statistical representative samples. DHS data is therefore well suited for the calculation of social vulnerability. DHS often comes with geolocations (or geotags) to individually determine the specific survey locations, and thus enable us to explore the certain regional dimensions of their vulnerabilities. If DHS data is not available for a country, it is possible to use other survey or census sources to calculate SV, with the condition that the used survey contains geotagged[8] samples.

[5] United States Agency for International Development (USAID, Demographic and Health Surveys Program, https://dhsprogram.com/
[6] Or any other geotagged, suitable survey information available for the region / country of interest.
[7] Geolocations: Precisely defined locations of surveys taken, i.e. described with geographical coordinates (latitude, longitude).
[8] Geotagged surveys contain information on where exactly one or multiple interviews were conducted.

For instance, this will most likely become possible with the newest iteration of *Multiple Indicator Cluster Surveys* (MICS), starting in 2023.[9] Other survey data can also be used to complement the calculation and help to contextualize the findings. Examples for this may be the 'Listening to Tajikistan' initiative by the World Bank,[10] or the Household Budget Survey[11] by Eurostat. However, these surveys may not come with the necessary geographical resolution to fully support DSVI requirements.

Table 1 provides an example of the characteristics of a standard DHS survey and their corresponding socio-economic dimensions:

Table 1: Standard DHS survey characteristics (Subset)

Social	Health	Economy	Infrastructure
Age Gender Ethnicity Migration Early childhood - Education ...	Health insurance Blood pressure Tobacco use Tuberculosis Vaccinations Alcohol use Disabilities ...	Income Working - Environment Unemployment Access to banking ...	Travel times to water Internet access Building materials Radio / Television Transportation Urban / Rural Cooking fuel ...

Figure 2 shows the absolute number of DHS survey points per commune/jamoats (Level 3 districts) in Albania (left) and Tajikistan (right). The communes/jamoats where no surveys have been conducted are greyed out. Analogously, these regions are very sparsely populated. The DHS dataset used in this analysis contains 715 survey points and a total of 15,000 interviewed individuals in Albania and 366 survey points/5,000 individuals in Tajikistan.

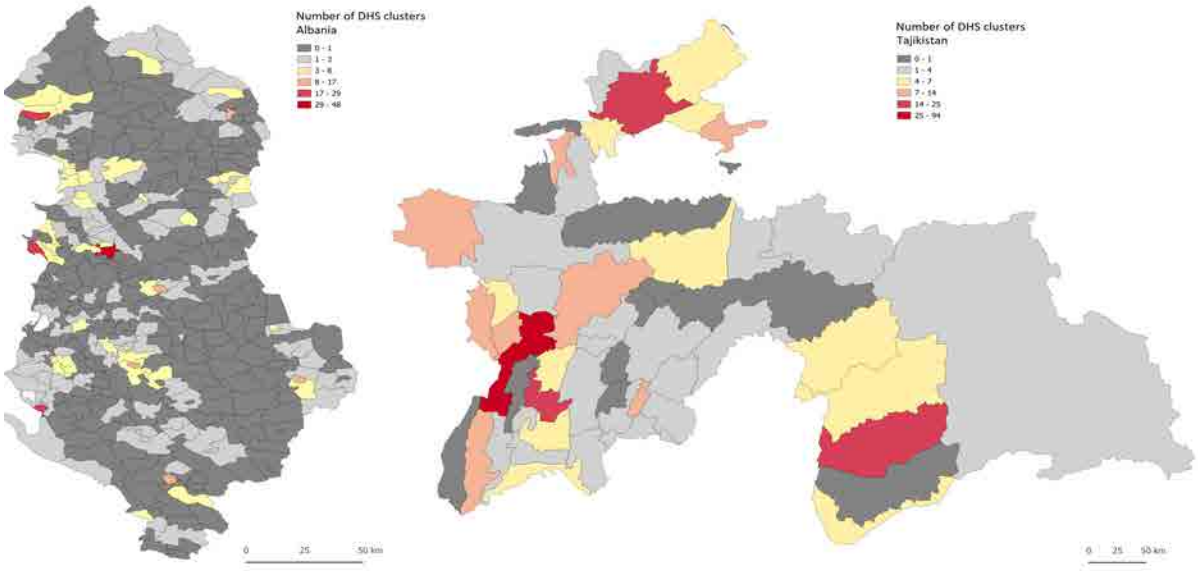


Figure 2. Number of DHS clusters per administrative unit in Albania (left) and Tajikistan (right)

[9] UNICEF Multiple Indicator Cluster Surveys (MICS), The MICS GIS initiative: harnessing the power of geolocation data, 29 June 2022, https://mics.unicef.org/news_entries/216/the-mics-gis-initiative

[10] World Bank, Listening to Tajikistan - Household Survey: Background, Implementation, and Methods, November 2017, https://documents.worldbank.org/en/publication/documents-reports/documentdetail/624621538136672723/listening-to-tajikistan-household-survey-background-implementation-and-methods

[11] Eurostat, Household Budget Surveys – Overview, https://ec.europa.eu/eurostat/web/household-budget-surveys

The optimal datasets for DSVI are 'DHS Standard Surveys' with geotagged survey clusters (more information in Table 2). A full list of available datasets for DSVI can be found in the Annex of this technical whitepaper.

Table 2. Subset of available geotagged DHS datasets for DSVI

Country	DHS Type	Year(s)**	Region
Tajikistan	Standard ǂ	2017	Central Asia
Albania	Standard ǂ	2018	Europe; Balkans
Kenya	MIS; MIS; Standard ǂ	2020; 2015; 2014	East-Africa
Ethiopia	Interim***; Standard	2019; 2016; 2011	East-Africa
Kyrgyz Republic	Standard	2012	Central Asia
Armenia	Standard	2015-2016	Asia
Jordan	Standard	2018	Asia
Rwanda	Standard	2020	Africa
Cameroon	Standard	2018	Africa
Tanzania	Standard	2016	Africa
Zambia	Standard	2018	Africa

ǂ *Standard: Used for default implementation of DSVI*

* *Malaria-Indicator-Survey: https://dhsprogram.com/Methodology/Survey-Types/MIS.cfm*

** *https://dhsprogram.com/data/available-datasets.cfm*

*** *This is a mini-DHS*

Green colour: A DSVI was produced for this country

2.2.2 Data collection: Spatial data

We collect and create high-quality spatial data to find the connections between them and the already-calculated SV scores. This technique allows us to predict social vulnerability on a country-wide scale, without having holes in the image we produce. Examples of such predictions will be presented in Chapter 3 'High-resolution social vulnerability'. Chi et al. (2022) used a similar set of geospatial variables for their prediction of the '*Relative Wealth Index*',[12] a metric calculation based on similar statistical assumptions and the same DHS datasets as DSVI.

These variables for example can be '*distance to health care*', '*distance to roads*' or be biophysical, such as '*elevation above sea level*', or socio-economic, such as '*light emission at night*', or '*gross domestic product (GDP)*'. These variables are broadly available for most countries in focus, but datasets which have been derived and processed for the calculations.

The selected datasets for DSVI represent all possible dimensions relevant for human development, but also many potentially relevant biophysical variables. This list is non-exhaustive and cannot be applied to every country exactly, but is a good starting point for the modelling and predicting of social vulnerability scores.

[12] USAID, DHS Program, https://dhsprogram.com/topics/wealth-index/

Table 3. Available geospatial datasets for high-resolution SV mapping (2022)

Variable Name	Year	Source(s)	Resolution
Nightlight Intensity	2022	NASA / NOAA NASA	500m
Proximity to national borders	2018	ESRI	<100m *
Proximity to protected areas	2021	protectedplanet.net	<100m *
Proximity to health care facilities	2022	OpenStreetMap.org / Humdata.org	<100m *
Proximity to financial institutions	2022	OpenStreetMap.org / Humdata.org	<100m *
Drive time to financial institutions**	2022	OpenStreetMap.org / Humdata.org QNEAT3, University of Vienna	<100m *
Drive time to education facilities**	2022	OpenStreetMap.org / Humdata.org QNEAT3, University of Vienna	<100m *
Drive time to health care facilities**	2022	OpenStreetMap.org / Humdata.org QNEAT3, University of Vienna	<100m *
Proximity to water	2022	GSHGG	<100m *
Population density	2020	NASA / University of Columbia	1 km
Temperature	2018	wordclim.org	1 km
Precipitation	2022	University of California, Santa Barbara Chelsa Climate	500m
Urban / Rural	2016	European Commission / JRC[13]	1 km
Vegetation indices	2022	NASA USGS European Commission / Copernicus Programme	500m
Slope	2000	Calculated with Elevation Data	500m
Elevation	2000	NASA / SRTM	30m
1. Wealth (LitPop) 2. Relative wealth**	2019	ETH Zürich Facebook	30m
Global Human Footprint	2004	NASA / University of Columbia	500m
Purchasing Power Parity (PPP)	2005	Yale University	1 km
Cell towers	2022	OpenCellID	<100m *
Land use class	2021	European Commission / Copernicus Programme	300m

** Resolution is flexible: grid size can be chosen*

*** Calculated with QNEAT3 (https://root676.github.io/)*

Figure 3 shows one of the derived input maps for our DSVI calculations: Drive time to a health facility. The input datasets are road networks obtained from OpenStreetMap (grey lines in the image), health facilities (hospitals, doctors) from various sources (including OpenStreetMap) and coloured drive time values calculated with the QGIS Network Analysis Toolbox 3 (QNEAT3).[14]

[13] Martino Pesaresi and Sergio Freire, GHS-SMOD R2016A - GHS settlement grid, following the REGIO model 2014 in application to GHSL Landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015) - OBSOLETE RELEASE. European Commission, Joint Research Centre, 2016 [Dataset] PID, http://data.europa.eu/89h/jrc-ghsl-ghs_smod_pop_globe_r2016a

[14] Clemens Raffler, About QNEAT3, https://root676.github.io/

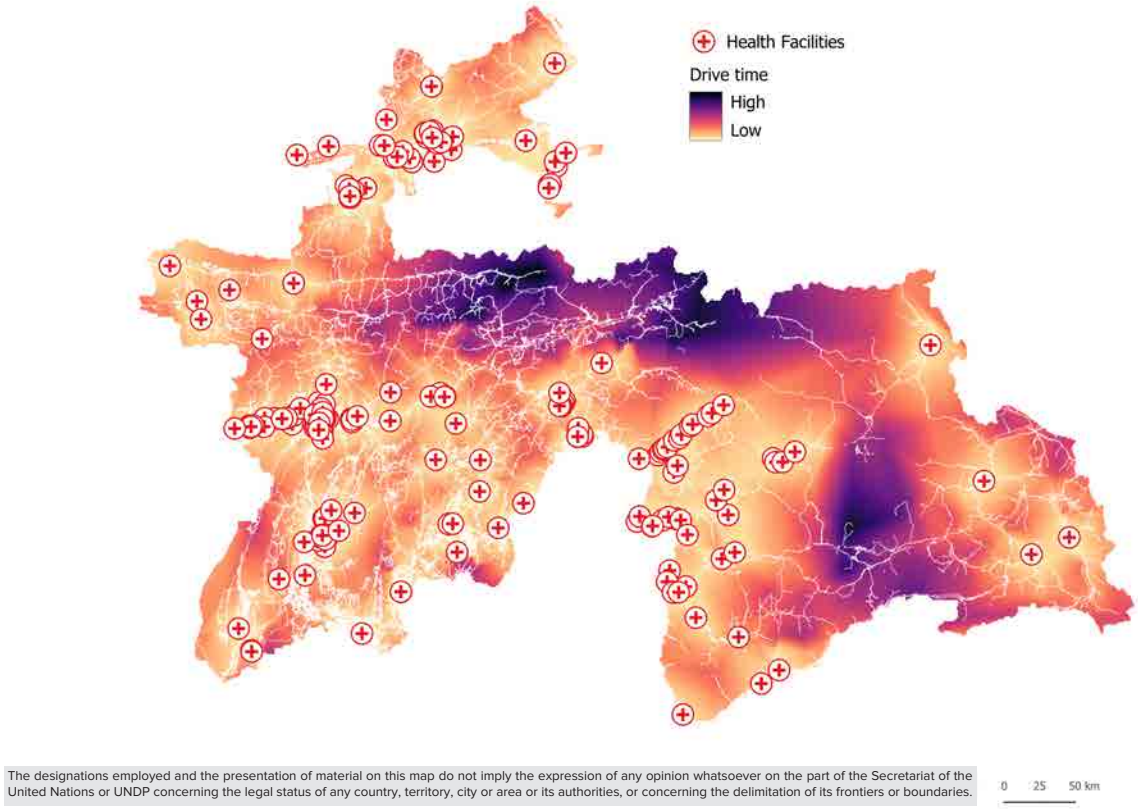


Figure 3. Drive time to nearest health facility in Tajikistan

This drive time map highlights the access to health facilities for a person in Tajikistan. It is one of the indicators relevant to an individual's vulnerability (Cutter et al. 2003).

2.2.3 Domain/Expert knowledge

Social vulnerability is a contextual metric that requires expert inputs for weighting. For every country studied, domain knowledge of the specific circumstances in that country must be considered. For instance, the influence of humanitarian indicators, such as the average age of household heads or gender-related indicators, can be interpreted in different ways and thus lead to different conclusions for a region or country. To improve results, it is advised to consult experts from the targeted country and discuss the specific weights of the SV indicators.

For DSVI, we developed a methodology to conduct such expert input consultations. The first step is to gather all potentially relevant indicators available for the country of interest. The next step is to group questionable indicators and let an expert decide whether they play a significant role for social vulnerability in the country. The expert can also decide whether the cardinal direction of that indicator is positive or negative. Next, encoded variables need to be assessed and ranked by the expert.

One example of such a variable can be the question what type of cooking fuel a family is using. The expert needs to rank the relative importance and benefits of different types of cooking fuel and establish a narrative where, for example, 'wood' is considered worse than 'gas'. These values can be newly encoded to a range of 1 ('better') to 5 ('worse'), or to a similar value scale.

2.3 Data processing for survey data

Data processing is an important step to transform the raw datasets and make them suitable for analysis. The data processing steps for survey data are mostly conducted with the help of Python,[15] Jupyter Notebooks[16] and spreadsheet softwares, such as Microsoft Excel.[17]

Explorative Data Analysis (EDA): Accuracy verification and outlier detection of the dataset is done using descriptive statistics (i.e. min/max, mean, standard deviation). Missing values can be replaced by substituting the variable's mean value for each enumeration unit. The statistical procedure will not run properly with missing values. Census units with population values of zero should be omitted. Generating correlation matrices are useful to eliminate intercorrelation.

Encoding: Some variables, such as '*Housing Materials*', are coded numerically, but represent a categorical value (e.g. 35 may mean 'concrete' and 47 may mean 'wood') and therefore are of a nominal (or ordinal) nature. Most statistical procedures cannot make meaningful computations based on nominal variables. These variables have to be 'ranked' by an expert to conform to an ordinal scale. If, for example, 0 means 'bad' and 10 means 'good', a resilience-based ranking approach could be applied for nominal variables, such as 'Housing Materials'. We transformed all nominal variables according to fixed rules and ranked them.

Grouping/Aggregation: DHS datasets for spatial analysis must be downloaded with the corresponding geotagged clusters. Household questionnaires are recorded on a per person and per household basis. This means that, per interviewed household, multiple entries in the same cluster ID can be found. Aggregation requires some type of summary statistics of the involved variables. For example, if a data frame with 5,000 rows is condensed to 500 clusters, 10 row entries per cluster need to be aggregated.

Common operations are to compute the arithmetic mean, but for some variables, other aggregation strategies need to be used. For example, for categorical variables, the modus can be useful to represent the general state of the cluster. DHS survey data can contain thousands of columns with NA values ('Not Available') which need to be filtered out. In Albania's case for instance, the filtering reduced the total amount of 19,724 columns to 2,644 after removing NAs.

Data Transformation: We use two different strategies to transform the input data for better scaling: z-score standardization and normalization. Standardization (or z-score normalization) is the process of rescaling the features so that they have the properties of a Gaussian distribution.

This generates variables with a mean of 0 and a standard deviation of 1.

$$x_{standardized} = \frac{(x_i - \mu)}{\sigma}$$

Where $x_{standardized}$ is the transformed variable (also sometimes called z-score), x is the value of instance $_i$ and μ is the population arithmetic mean and σ the root of the standard deviation s.

Normalization: We scale variables to a range of 0-1 in order to preserve relative distances between values and to make the data ready for further analysis.

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

[15] See https://www.python.org/
[16] See https://jupyter.org/
[17] See https://www.microsoft.com/en-us/microsoft-365/excel

2.4 Data processing for geographical data

The spatial data for the high-resolution prediction come in different formats and need to be adjusted for further use. We use the industry standard procedures to achieve a standardized, high-quality database for our calculations. In our data collection phase, we obtain mainly raster data[18] but also vector data,[19] from the sources listed in Table 3. Every vector dataset needs to be statistically 'aggregated' and transformed into a raster file with the same cell size as the largest available raster file from the other data sources.

For example, we transform the road network in a country into a 'road density' (per km^2) map. This needs to be done, because all datasets for the high-resolution modelling phase need to be in the same format. Other variables, such as slope, elevation, climate, might come in different geographical projections, tile sizes and need to be resampled, reprojected, or cropped in order to fit the area of interest. Lastly, each pixel of a raster image needs to be extracted/sampled together with the previously aggregated survey data into data frames or matrixes.

Distance maps are derived by applying techniques, such as '*Euclidian distance*', to generate a distance matrix. Such distances could be main roads, hospitals, critical infrastructure, such as markets or ATMs. Other interesting distance values can be related to hazard risks: if a hazard risk area is known, the distance to that area might be considered as a 'good' or positive indicator for lesser vulnerability.

Euclidian distance calculation:

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (q_2 - p_2)^2}$$

Where $d(p,q)$ is the distance between two points in space and q_i and p_i are two points in two-dimensional space.

Drive distance to critical infrastructure: We used the freely available QGIS tool 'QNEAT3 – QGIS Network Analysis Tool' and produced driving distance maps to '*health locations*', '*educational locations*' and '*financial institutions*'.

Only if all the information is united into a single data frame can a model be built that would obtain high-resolution SV maps.

2.5 Calculating social vulnerability

After all data collection and pre-processing steps, the datasets are ready for the final two main steps: First, based on the geotagged survey data points, the social vulnerability index is calculated by using a statistical procedure called Principal Component Analysis (PCA), and with the help of field experts. Field experts are practitioners, consultants or specialists, who possess profound insights into the humanitarian situation and development context of the country.

[18] See https://desktop.arcgis.com/en/arcmap/latest/manage-data/raster-and-images/what-is-raster-data.htm
[19] See https://spatialvision.com.au/blog-raster-and-vector-data-in-gis/

2.5.1 Literature review and process overview

Social vulnerability calculations are used in many scientific publications and with different regional and contextual focus, or specifically for a type of disaster/shock. Cutter et al. (2003) established the theoretical backgrounds to understand modern vulnerability analysis. According to them, some major factors influence vulnerability: lack of access to resources, limited access to political power and representation and lack of social capital and available infrastructure (Cutter 2001; Tierney et al. 2001; Blaikie et al. 1994).

With this picture in mind, Cutter et al. (2003) identified 250 potentially relevant variables for households in the United States in 3,141 counties. They used multicollinearity analysis to create a smaller subset of 85 variables. After this, all variables were normalized (scaled) and added to the principal component analysis. The results were examined and studied to find the potential correlation between SV scores and disaster declarations per county. However, the results yielded only a very weak correlation score of -0.099. In recent years, authors developed the method further to include the missing biophysical components of previous studies.

Willis and Fitton (2016) reviewed social vulnerability assessments comparing three different methods for a flood prone catchment in the UK. One of the goals of this study was to fit the concept into the conceptual frameworks of disaster risk reduction (DRR) and to use social vulnerability as a tool to identify, assess and monitor disaster risks and enhance early warning. The analysis was conducted on a district level, similar to Cutter et al. (2003).

De Loyola Hummell et al. (2016) looked into the case of Brazil and concluded that vulnerability analysis is essential in understanding how distinct social groups are differently impacted by natural disasters. They used a slightly updated version of the original technique by Cutter et al. (2003) and obtained 45 city-level indicators which were reduced by the PCA to 10 components that explained about 67 percent of the total variance.

In their analysis, they were able to categorize these components with some of the commonly understood drivers for vulnerability, but with a context-specific viewpoint. These included poverty, urban/rural development, migration, special needs population, racial diversity, race, population density, lack of public employment, tourism-based economy and extractive industry. What these and other publications have in common is the usage of PCA and census/survey datasets derived for district-level resolutions.

Most authors use an additive model to sum up components and loadings into a composite vulnerability score. Cardinality assessments are always done with the help of literature research, or of field experts, for the studied region. Approaches to validate the existing scores are limited, such as in Willis and Fitton (2016), where the calculated scores were compared with the previous scores generated with the methods of Cutter et al. (2003), or Rygel et al. (2006).

We summarized some of the most common techniques to calculate SV for different scopes and studies, as outlined in Table 4:

Table 4. Proposed main steps in selected publications to calculate SV

Article	Steps to compute SV
Cutter et al. (2003)	1. Variable collection 2. Multicollinearity test 3. Normalization of data (to percentages, per capita, or density functions) 4. Factor analysis (PCA) 5. Scale factors 6. Sum factors 7. Map scores based on SD from the mean into 5 categories ranging from -1 to 1
De Loyola Hummel et al. (2016)	1. Variable collection 2. Normalization of data (to percentages, per capita, or density functions) 3. Multicollinearity test 4. Variables standardization 5. Factor analysis (PCA) + varimax rotation + Kaiser criterion 6. Loading interpretation according to the most relevant (|x| > 0.5), applying positive or negative signs or absolute values according to the variable expected impact 7. Compute SV by summing the factors 8. Map scores based on SD from the mean into 5 categories ranging from -1.5 to 1.5
Guillard-Goncalves et al. (2015)	1. Variable collection 2. Normalization of data (to percentages, per capita, or density functions) 3. Correlation analysis 4. Variables standardization 5. Factor analysis (PCA) + varimax rotation + Kaiser criterion 6. Loading interpretation according to the most relevant (|x| > 0.5), applying positive or negative signs or absolute values according to the expected impact 7. Compute SV by summing the factors 8. Map scores based on standard deviations from the mean into 5 categories
Rufat et al. (2019)	1. Variable collection 2. Normalization of data (to percentages, per capita, or density functions) 3. Factor analysis (PCA) + varimax rotation + Kaiser criterion 4. Loading interpretations according to the most relevant (|x| > 0.5), applying positive or negative signs according to the expected impact 5. Compute SV by summing the factors
SDG AI Lab 2022	1. Variable collection 2. Normalization of data (to percentages, per capita, or density functions) 3. Pre subset of relevant contextual indicators for country of interest (with field experts) 4. Factor analysis (PCA) + varimax rotation + Kaiser criterion 5. Loading interpretations according to the most relevant (|x| > 0.7 or 0.5), applying positive or negative signs according to the expected impact 6. Compute SV by summing the highly loaded variables of each component 7. Map based on transformation of SV to a range of 0-1, where 0 is low vulnerability and 1 is high

Many of the proposed techniques are very similar to each other. The SDG AI Lab took another step and combined some of the best practices. For instance, we integrated the field experts in the '*Indicator selection process*' to be able to adjust the weights of certain vulnerability indicators for our final score. This increased our accuracies when predicting social vulnerability with geodata in our tests.

2.5.2 SV calculation: Indicator selection and cardinality assertion

We followed the suggestions from the literature on indicator selection as close as possible. DHS captures many of the important dimensions of social vulnerability well, but may have some weaknesses in others. The indicators can generally be grouped into categories and are shown in Table 5 below:

*Table 5. Indicator list for SV computations**

Indicator Group	SV Indicator
Socio-economic	GDP per capita Average monthly salary Unemployment level Number of socially dependent individuals/citizen Occupation (profession and managerial level) Occupation – open space (e.g. agriculture, construction) Economic sector (e.g. resource extraction)
Demographics	Age (proportion of youth and elderly population) Gender (female) Education Special needs/disability population Vulnerable minorities Immigrants Rapid population growth
Family structure	Single-parent households Single-member households Large families
Medical services	Number of medical personnel per capita Number of hospitals per capita Average distance from nearest hospital
Urban	Percentage of urban population Quality of infrastructure Age of infrastructure Average property value
Built environmental vulnerability	Population density Quality of infrastructure Age of infrastructure Average property value
Social capital	Sense of community Attachment to a place Perceived level of social support Civic participation

** Here indicators have been selected to produce a good representation and to help identify a strong selection for analysis.*

However, it should be noted that the actual influence is context-dependent, i.e. it should be assessed in every individual study. Additionally, the list should not be taken as an exhaustive list of all possible indicators, but merely as one with examples intended to improve understanding of the matter.

Table 6. Asserted cardinality of strongly loaded variables (Component 1-7), Albania

Index	Variable name	Component Loading*	Asserted Cardinality**
1	Main wall material	0.56	increasing
2	Number of household members	0.70	increasing
3	Number of eligible women	0.60	increasing
4	Household has telephone (or similar)	0.56	decreasing
5	Frequency of reading newspaper	0.63	decreasing
6	Has an account in a bank	0.65	decreasing
7	Wealth index combined	0.62	decreasing
8	Smokes cigarettes	0.62	increasing

** absolute value;*

*** increasing: supposedly increases vulnerability, decreasing: supposedly decreases vulnerability*

Table 6 shows some examples of variable loadings and the asserted cardinality. For this assertion, the expert working on the topic needs to know the positive or negative expression for each variable on social vulnerability represented. For instance, '*Main wall material*' represents the type of material from which house walls are constructed. If the variable is coded in a way that higher numbers mean 'worse materials', then we can assert that higher average scores must be associated with higher vulnerability scores in general. It is important to note that variables must be inspected one by one to ensure the correct cardinality for the context of the vulnerability dimensions of that country.

2.5.3 Social vulnerability: Calculation details

This section explains the details of the steps mentioned in Table 4 to calculate SV. As previously indicated, our calculation of SV is done by combining the evaluation efforts of Spielman et al. (2019) and the generalized recipe of the Institute of Hazard and Vulnerability at the University of South Carolina. For our purposes, the SDG AI Lab has made alterations to the processes mentioned by the authors.

Spielman et al. (2019) describe social vulnerability as a 'latent' variable since it is not directly observable but is characteristic to individuals or environments, further implying that statistical methods are required for indirect measurements. The statistical procedure used to calculate the SV in this paper is PCA. The PCA is a widely used method for finding patterns in data and enables reducing the data dimensionality, minimizing information loss while still preserving the high variance.

PCA was also first used by Cutter et al. (2003), who performed the first in-depth analysis of social vulnerability in the United States. Spielman et al. (2019) improved this methodology and evaluated the construction process for SV further. In the PCA we find new uncorrelated variables called components, representing the linear combination of the original data. The general workflow consists of standardizing the original data, then calculating the covariance matrix and subsequently eigenvectors and eigenvalues (Jolliffe and Cadima 2016).

$$cov(x,y) = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Where x_i is data value of the first variable x, y_i is data value of the second variable y, $\bar{x}$ $\bar{y}$, are their mean values respectively, and n is the number of data values. Covariance describes the joint variation between two variables. It is an extended concept of the simple variance that only measures the distribution and spread of a one-dimensional dataset. Covariance matrixes are used in PCA to discover similarities between variables and group them. The strength of this relationship can also be used as an indicator and selection threshold for relevant variables in the later analysis.

Furthermore, the analysis of component loadings (as described subsequently) are of special interest for SV construction. Component loadings show how strong each input variable (SV relevant indicator) is correlated with the resulting component.

$$PC_j = w_{ij} X_j + w_{i+1j+1} X_{j+1} + \cdots + w_{i+nj+n} X_{j+n}$$

Where X_j represent original values of variables and w_{ij} represent elements of eigenvectors, also called component loadings.

The mentioned algorithms and steps in the list below are partially derived from Spielman et al. (2019) and the Institute of Hazard and Vulnerability, University of South Carolina, and adapted to our needs:

Principal Component Analysis: PCA[20, 21] uses a varimax rotation (100 iterations) and Kaiser criterion (Braeken and Assen 2006) with 100 iterations for component selection. A varimax rotation reduces the tendency for a variable to load highly on more than one component. An overview of the total explained variance per component can be deducted by the examination of an explained variance plot. From Figure 4. Scree plot for PCA for Tajikistan, it is evident that after the 7th component, the explained variance remains almost constant, and that the first 7 components already explained most of the total variance.

[20] Lindsey Smith, A tutorial on Principal Components Analysis, 26 February 2002, http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf

[21] Ian T. Jolliffe and Jorge Cadima, Principal component analysis: a review and recent developments, Royal Society Publishing, 13 April 2016, https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

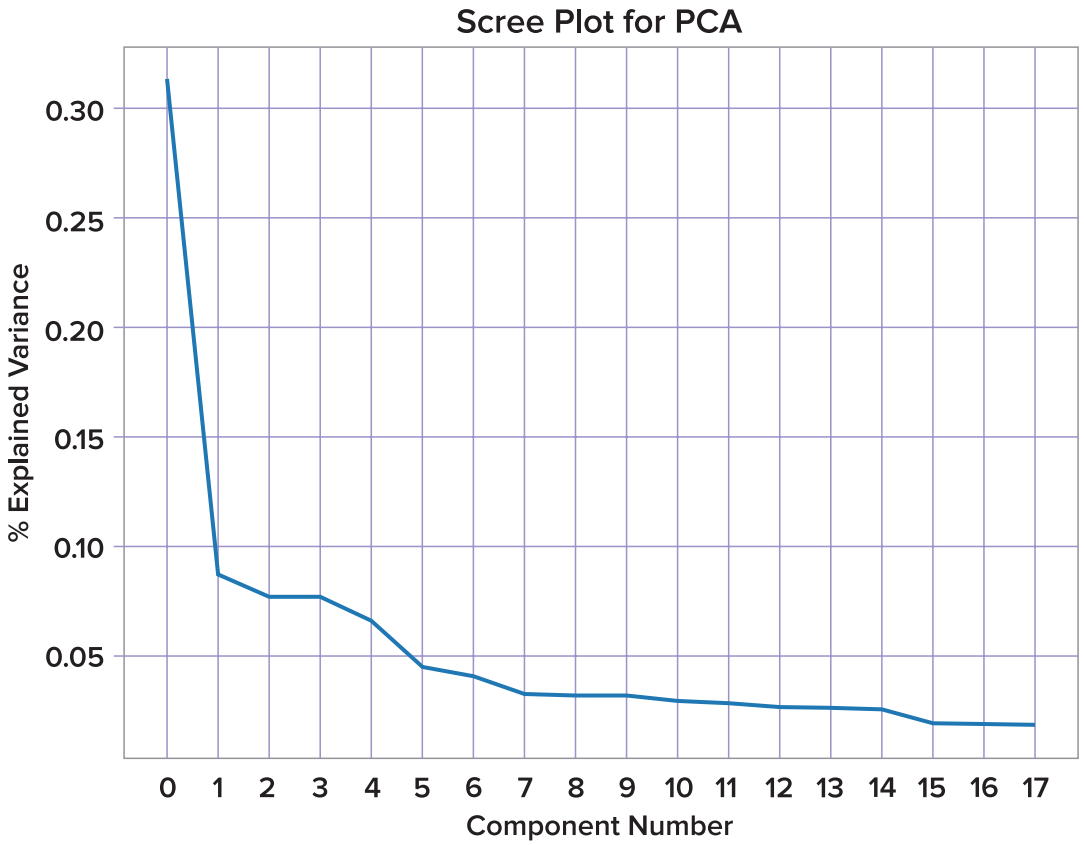


Figure 4. Scree plot for PCA for Tajikistan

Component examination: The broad representation and influence on (i.e. increase or decrease) social vulnerability for each component is determined by scrutinizing the loadings for each variable in each component. We also perform so-called 'component naming' which is helpful to group components into SV relevant classes. This is performed based on the most relevant variables of that component and the strength of their loadings.

Examining component loadings: Loadings are calculated for each variable - component combination. The values considered are usually greater than 0.7 or less than -0.7 because they are covariances/correlations between the original variables and the unit-scaled components.

In some cases, only variables with loading scores greater than 0.8 or even more are considered when many highly loaded variables appear in one component. It is possible to group such variables to keep similarly expressed variables in the component for further analysis.

Table 7. Indicator groups after PCA with corresponding component loadings (Tajikistan)

Component Name*	DSV Indicator from DHS Data**	Component Loadings (> 0.7 or 0.5)	Variance Explained	Sum of variance
1			0.236	0.23
Socio-economic	Highest educational level	0.9075		
	Education in single years	0.8764		
	Has an account in a bank	0.8432		
	Wealth index combined*	0.8262		
	Covered by health insurance	0.7837		
	How often uses internet	-0.7205		
2			0.146	0.38
Medical infrastructure	*Getting medical help:*			
	Getting permission to go	0.7720		
	Distance to health facility	0.8051		
	Having to take transport	0.8034		
	Not wanting to go alone	0.8312		
	Concern no female health provider	0.8102		
3			0.107	0.49
Social capital Domestic violence Gender equality	*Beating justified if wife:*			
	Goes out without telling husband	0.7878		
	Neglects the children	0.7590		
	Argues with husband	0.7905		
	Refuses to have sex with husband	0.7724		
4			0.096	0.58
Critical infrastructure	Has telephone (landline)	0.752		
	Household has electricity	0.743		
5			0.087	0.67
Demographic	Respondent's current age	0.7980		
	Number of household members	-0.784		
	Number of eligible women in household	-0.767		
6			0.056	0.73
Infrastructure	Main floor material	0.6764		
	Main roof material	0.5794		
	Main wall material	0.7015		
7			0.053	0.78
Health Literacy	Frequency of reading newspaper or magazine	0.5088		
	Frequency of listening to radio	0.5914		
	Hemoglobin level (g/dl decimal)	-0.5687		

** Component name: Assigned to components to reflect the main indicator groups relevant for social vulnerability*

*** Indicators selected according to country of interest. Indicators change with application in other countries*

Table 7 highlights the outputs of the PCA analysis that are used for index computation. The columns on the right-hand side 'variance explained' and 'total variance' show each component's explanatory power regarding the total sum of all components. Since we used 44 variables as the input, we received 44 components after performing the PCA analysis. The first 7 components explain around 78 percent of the total variance, and henceforth are used to construct the index. Each component consists of several variables with high loadings. Hence, the components are named after the variables that constitute them.

Directional adjustment (or cardinality) is applied to an entire component to ensure that the signs of the subsequent defining variables are appropriately describing the tendency of the phenomena to increase or decrease vulnerability. This is done by adjusting the sign of the individual variables selected by the PCA within a component. Variables that possess a negative asserted influence on vulnerability are multiplied by -1. In the case of positive variables, the result is a positive number.

Weighting is done by using the individual component loadings of each variable as weights. We tested several approaches of weighting for the resulting factors and used geodata to predict the resulting SV scores. We were able to fine-tune the weighting approach by cross-validating the results with their representations in geoinformation and chose the weighted scores due to their higher correlation with the available geodata.

In further mentions in this whitepaper, we refer to this score as 'SV_Scaled' in the online repository. We apply cardinality corrections to the variables of each component and use their loadings as weights to represent their strengths in regard to the total components. After that, we applied another weight as the component itself had a total contribution to the total variance explained. We tried the same experiments without scaling and predicted the resulting SV scores with the available auxiliary data sources. We found that the best model results are obtained with scaled SV scores.

Calculate social vulnerability by placing all the components with their directional (+, -) adjustments into an additive model to generate the overall SV score for the place.

$$Component_i Scaled = \sum_{j=1}^{n} (Indicator * sign_j * loading_j) * tot.variance$$

$$Component_i Unscaled = \sum_{j=1}^{n} (Indicator_j * sign_j * loading_j)$$

Where *Indicator* is the corresponding variable previously selected and then returned by the PCA process with loading scores. *Loading:* for each principal component, the algorithm used returns a loading score (similar to correlation) per used input variable. For example, as presented in Table 7, the first variable of the first component had a loading score of 0.9. *Sign* is the assigned cardinality by an expert, or based on suggestions from literature. Each variable has an assigned cardinal influence on social vulnerability.

For instance, when looking at component 1 in Table 7, the second variable is 'Education in single years', with an assigned loading score of 0.87. In order to represent the correct cardinal direction, the variable is multiplied with -1 (or -0.87 if using loading scores as weights) for directional adjustment. The reason is that 'Education in single years' decreases vulnerability, if the number of years increases. *Tot. variance* (Total variance) is the total contribution of that component to the variance of the PCA.

$$SV\ Score = \sum_{j=1}^{n} (Component_i)$$

Where *SV Score* is the sum of all components. The components already exhibit the correct cardinal sign before this step, henceforth no additional sign changes are necessary. Components with decreasing effects on vulnerability will go into the equation with negative signs and components with increasing effects on vulnerability with positive signs.

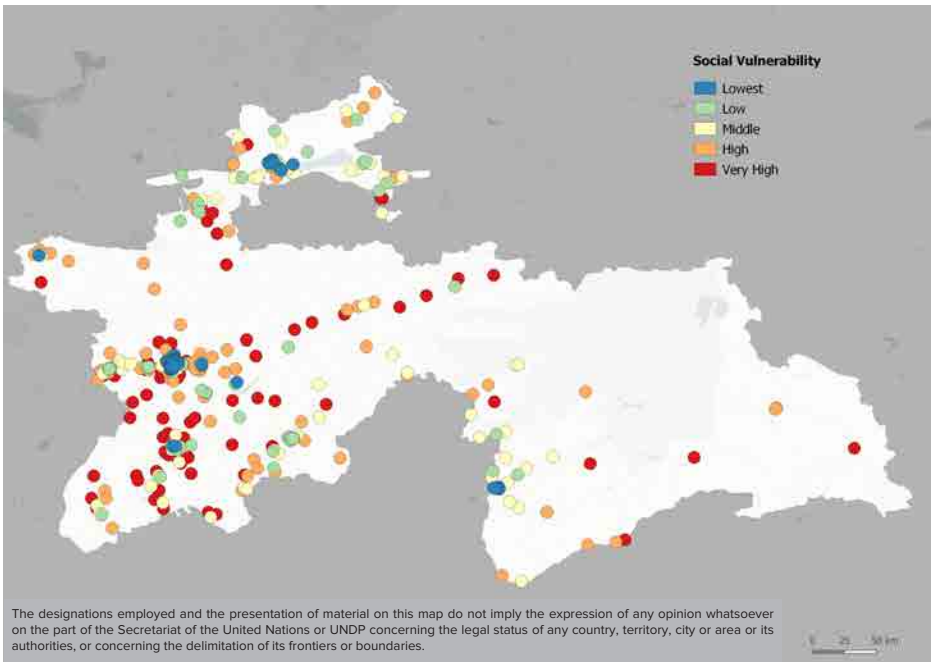


Figure 5. Social vulnerability scores for survey points in Tajikistan

The outcomes of the calculation process are vulnerability points which can be seen in Figure 5. One pattern that always seems to emerge from our social vulnerability scores is that urban centres always exhibit lower vulnerability scores than rural areas (tested countries: Albania, Ethiopia, Kenya, Tajikistan).

Figure 6 shows that even in close proximity to the metropolitan region of Dushanbe, as distance increases, highly vulnerable population groups can be found (red dots).

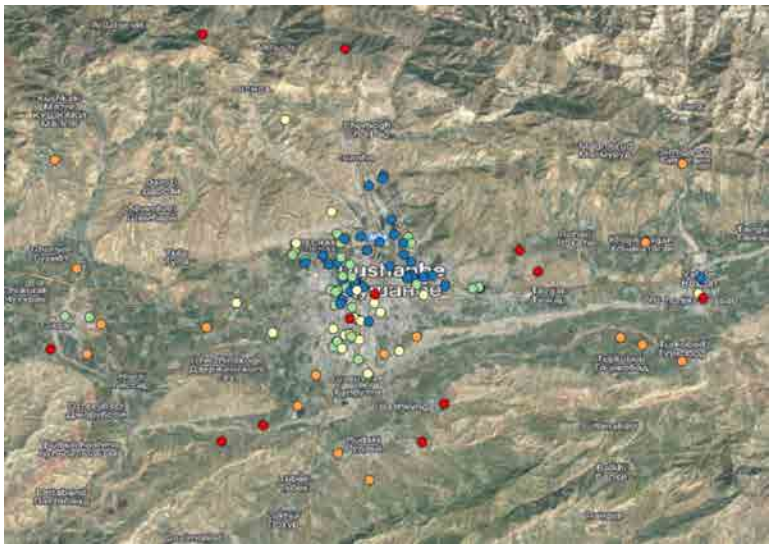


Figure 6. Distribution of vulnerability points near the metropolitan region of Dushanbe, Tajikistan

Normalizing (scaling): Just as described in Chapter 2.3, we normalize the SV scores obtained to a common scale of 0-1.

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Grouping: SV scores can be mapped using an objective classification (i.e. quantiles or standard deviations) with 3 or 5 divergent classes to illustrate areas of high, medium and low social vulnerability. It is common to display SV in units of standard deviation to reduce the impact of outliers in the final map. The grouped scores can be used to train classifiers instead of regression models. In our experiments, the success rate of regression models was higher than classifiers, such as neural nets (See Chapter 3).

Aggregating to administrative boundaries: The obtained scores can then be aggregated to a chosen municipality or district level based on administrative units, as shown in number 3 (Figure 7) Aggregation can be performed by conducting a spatial join and calculating a statistical score (such as mean value for all points within the boundary). The generated map is comparable with classical variations of SV from other sources, or the traditional method explained in the UNDP handbook.

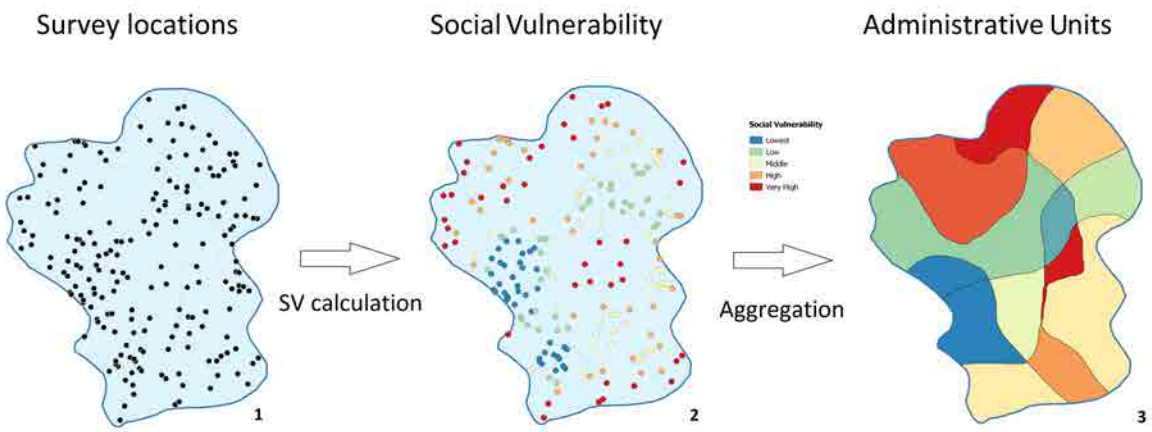


Figure 7. Calculation and aggregation of social vulnerability

This way of presenting aggregated information is a popular way to show vulnerabilities and other socio-economic indicators on maps because the source material was not available for every corner of the country. This can become a problem because small-scale changes in vulnerability cannot be identified with this map. To overcome the limitations, we propose enhancing the available datasets with spatial or geographical datasets and use those to predict SV on a much finer scale. This is described in detail in Chapter 3. High-resolution mapping.

3. High-resolution social vulnerability

This chapter explains the processes we used to derive high-resolution social vulnerability maps.

3.1 Workflow

This process follows the methodology proposed in 'Urban social vulnerability assessment with physical proxies and spatial metrics derived from air- and spaceborne imagery and GIS data' by Müller et al. (2009), among other methodologies. We also utilized knowledge obtained from various papers, which use raster proxy data (or auxiliary geodata in Table A1) to predict socio-economic resilience/vulnerability and poverty metrics, such as social vulnerability.

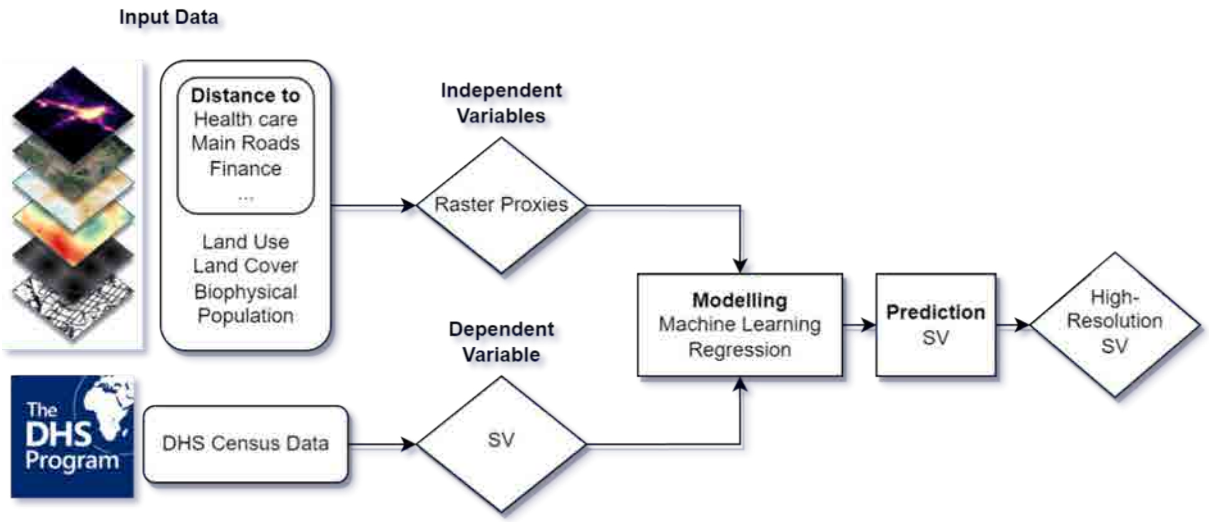


Figure 8. Overview of workflow for high-resolution SV

Figure 8 shows the high-level process which we used to combine the calculated SV scores with geodata to obtain estimations for locations not covered by survey data points. The resulting map explains the SV with as much accuracy as the models were able to predict the test data derived from our SV sample. First, we use the input datasets, survey data and geodatasets, after the pre-processing steps are concluded. We identify the dependent and independent variables in the context of the model and feed the datasets into various machine-learning models to establish a link between the two.

After the modelling process and the assessment of the accuracy scores, new social vulnerability regions can be predicted by using the same geodatasets as in the training phase. The results are new high-resolution social vulnerability maps, which give us insights into vulnerabilities, where surveys or other monitoring programmes were not conducted.

3.2 High-resolution mapping (spatial disaggregation)

The middle map of Figure 9 shows various layers of geographical data which are widely available, or specifically derived with procedures developed for the calculations. The geographic data is stacked and then fed into a model utilizing machine learning to predict the calculated scores (1). The results can be seen as a high-resolution map on the right-hand side (3). This allows practitioners and users to see SV in exceptionally fine detail, down to neighbourhood levels of small cities. This level of detail is a new feature and has not been implemented in any known SV mappings within UNDP.

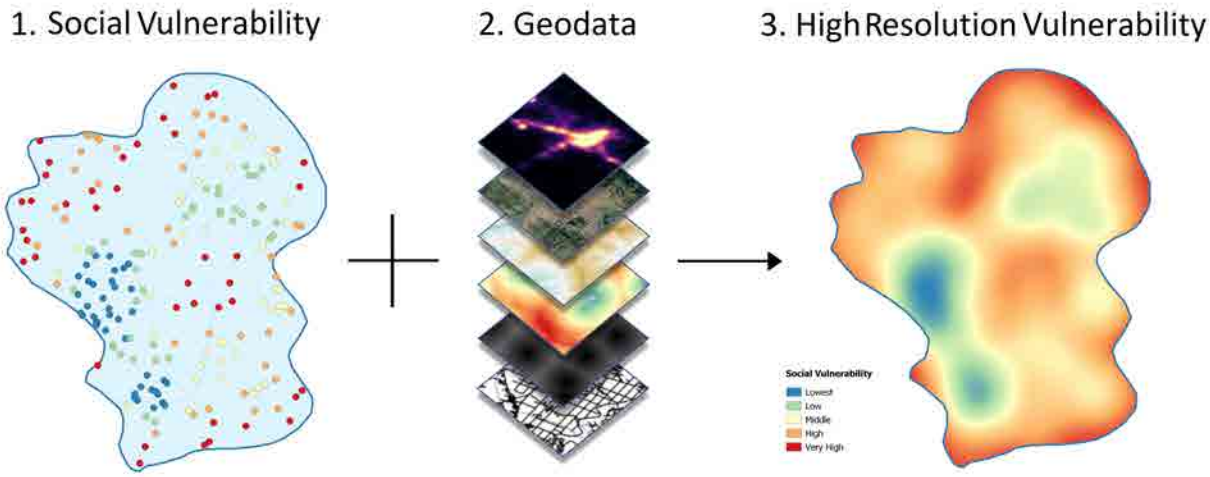


Figure 9. Schematic of SV prediction with spatial data

The output map 3 of Figure 9 is a simple example of a high-resolution social vulnerability map. The selected baseline model was a simple linear regression, with varying options of train/test samples, differently scaled input variables and slightly adjusted SV input variables. The presented data layers (2) are input layers to predict social vulnerability, which might provide a basis for further analysis. We explain the used datasets and models in more detail in the latter sections of this whitepaper (Chapter 3.4).

3.3 Geodata exploration

The used geodatasets represent the biophysical and socio-economic realities of the countries. We aimed to obtain datasets with the highest possible resolution and the closest time to the present. We prefer the datasets that are less than five years old and are the result of other scientific studies. Some of the datasets can be derived from crowdsourced sources and further processed by the team.

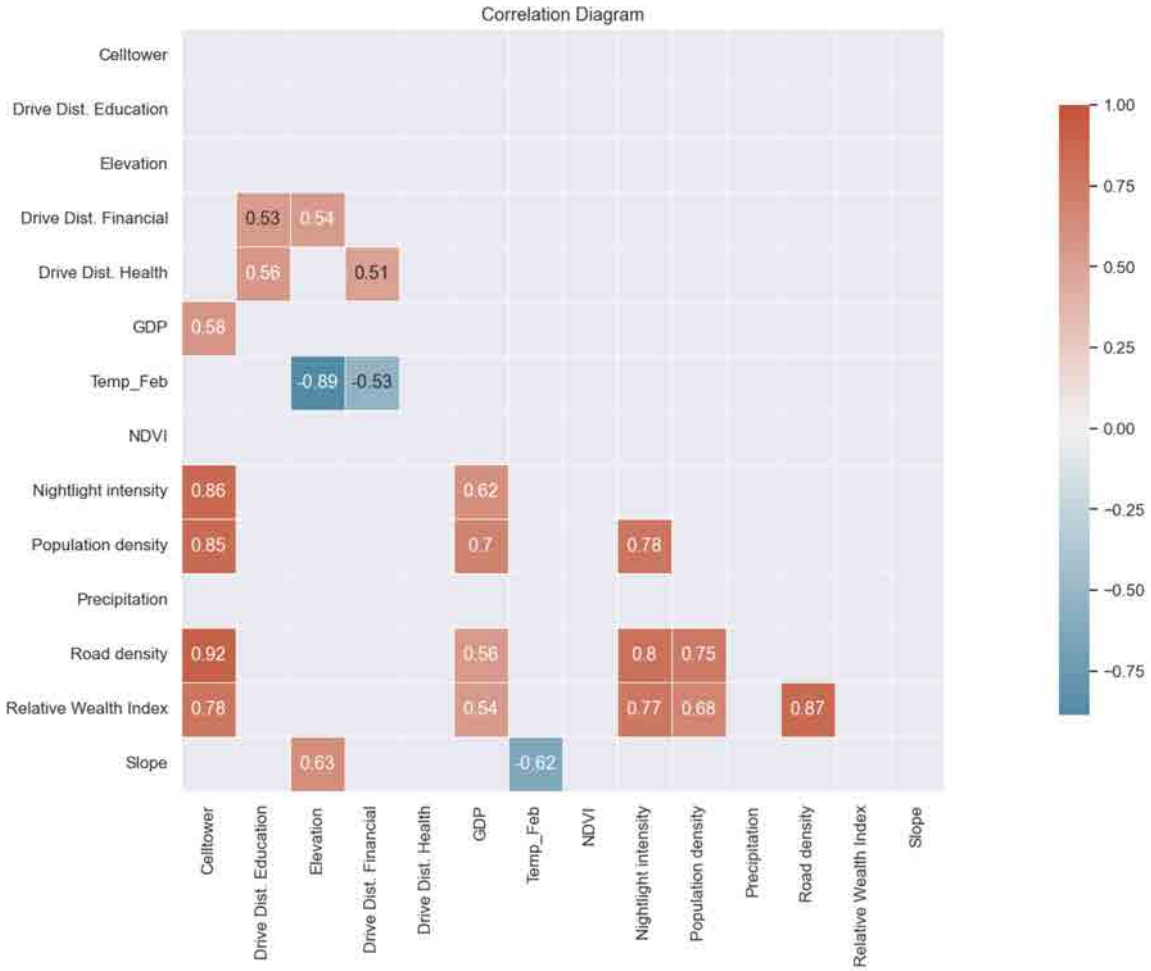


Figure 10. Correlation plot of chosen spatial variables for the test case in Tajikistan. Later combinations of geospatial variables in other countries are subject to change

Figure 10 shows the correlation diagram of the selected spatial variables. For modelling, we wanted to avoid high correlations between the dependent variables, as well as insignificant correlations. The chosen threshold can be based on the relative cumulative distribution of correlation scores. We excluded highly intercorrelated pairs of variables, such as elevation with temperature (and vice-versa) and chose the remaining variables according to their correlation with SV. The remaining variables are relatively more highly correlated with SV than their respective pairs.

Table 8. Absolute correlation between SV and geodata (> 0.3), Albania

Correlation pair: SV	Pearsons' Correlation: Albania Data	Pearsons' Correlation: Tajikistan Data
Relative wealth ~ SV	No data	0.728
Road density ~ SV*	No data	0.637
GDP_2015 ~ SV	No data	0.428
Vegetation Index (NDVI) ~ SV	0.557	0.260
Nightlight Intensity (NTL) ~ SV	0.527	0.606
Drive time to financial service ~ SV*	0.517	0.394
Popdens ~ SV	0.509	0.458
Drive time to health ~ SV*	0.447	0.432
Drive time to education ~ SV*	0.351	0.395

** Derived from OpenStreetMap data*

After computing the SV scores for each cluster, underlying raster data is used to construct models to predict SV within those clusters (Figure 8). The chosen model, in the first iteration of this product, was a multiple linear stepwise regression. At first, a multitude of raster proxies, which were derived from multiple sources, were used in the prediction (subset of the total list is in Table 3).

3.4 Social vulnerability prediction: Baseline model

In order to see what methods worked most optimally, we first established simple prediction models and then explored the technologies further. We compared the performance of different types of classification models. As we want to determine the relations between dependent and independent variables, the regression analysis is a suitable approach. Additionally, since the correlation between the dependent and independent variables exists, linear models can be used (James et al. 2021).

The baseline models used for the predictions are multivariate linear regression and stepwise regression. Multivariate linear regression is an extended version of simple linear regression, but unlike simple linear regression, it uses two or more independent variables to predict the dependent variable (Su et al. 2012). The model is implemented using sklearn library.[22]

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m$$

Where y is the result of the regression, α is a constant added to the model to offset the lines intersection point with the y axis, β_m are slope parameters for the input data points x_m.

Stepwise regression is an iterative process used to determine the predictors that will be included in the final mode. After running the statistical tests and evaluating the p-values (probability values) of the independent variables, we run the new multiple regression using only the variables that passed the tests (Del Serrone and Moretti 2023). This approach ensures only the independent variables that have a significant influence on the dependent variable are included in the model (Wang et al. 2007).

[22] See https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

We performed stepwise linear regression after eliminating highly correlated variable pairs. We split the SV dataset into two categories 'Urban' and 'Rural' based on their a priori assigned values originating from the DHS survey. After creating the model, the dataset is split into training and test sets with a ratio 80 percent to 20 percent SV test labels are plotted. We ended up with 715*0.8 = 572 training samples and 143 test samples (case Albania). In addition to this, we used random state and sampling randomizers to create independent sets of training/test samples to ensure a smaller bias in the sampling design.

3.5 Social vulnerability prediction results: Advanced model(s)

After implementing multivariate linear regression and stepwise regression, seeking for prediction improvements by utilizing other model alternatives is a reasonable approach. Based on the nature of the problem at hand, it is natural to assume that it requires a regression model for improvement. However, inspired by the wide range of possibilities in the machine-learning realm, we experimented with different kinds of regressors and discrete classifiers, such as the MLP classifier.[23]

3.5.1 Advanced regression

Huber, ridge, random forest, XGBoost and decision tree regressors are some of the applied regression analyses. The Huber regressor is less sensitive to outliers than the traditional linear regression while the ridge regression adds a penalty term to the least square errors to shrink the regression coefficients and thus improve the generalization of the model (Huber 1964; Hoerl and Kennard 1970). The decision tree regressor is a non-parametric, interpretable and popular method that can handle both continuous and categorical variables. It is one of the building blocks of the random forest algorithm (Breiman 2017). Apparently, they have some advantages over the baseline model, although these three regressors do not yield outstanding results. Therefore they are not included in the sequential steps.

Two regression models that have demonstrated good performance across a wide range of applications, random forest and XGBoost regression, are the ones that also address our prediction improvement effort. They have advantages over other methods, such as robustness, non-linearity, the ability to handle missing data and identify feature importance, speed and scalability (Breiman 2001; Chen and Guestrin 2016). Random forest is an ensemble learning method for regression and classification.

Random forest regression, as can be seen in Figure 11, involves randomness during the construction of the decision trees by not only selecting the features but also sampling the data. It also provides a mean prediction of the individual trees. The randomness and voting mechanism for each tree prediction contributes to the robustness of the model and improved accuracy. In addition, the use of randomness can also have a positive effect on the reduction of overfitting (Breiman 2001).

[23] See https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

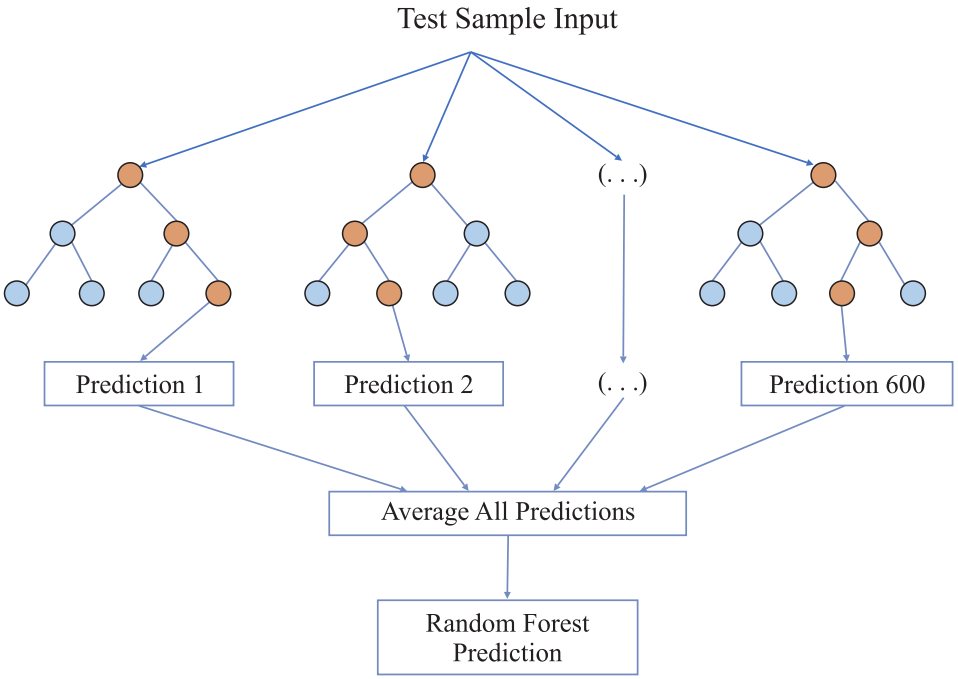


Figure 11. Random forest trees run in parallel without interactions and the final output consists of the mean of the classes as the prediction of all trees [24]

A more recent regressor, XGBoost, was developed to overcome challenges of the gradient boosting algorithms, such as overfitting, limited parallelism and slow computation. It combines gradient boosting and decision trees and is able to handle various types of data with its high level of scalability and efficiency (Chen and Guestrin 2016). Unlike random forest, where each tree model is trained independently and has equal weight in the final prediction, XGBoost uses a sequential approach where each tree model minimizes the residual from the previous tree model XGBoost (Wang et al. 2020). A simplified structure of XGBoost is shown in Figure 12.

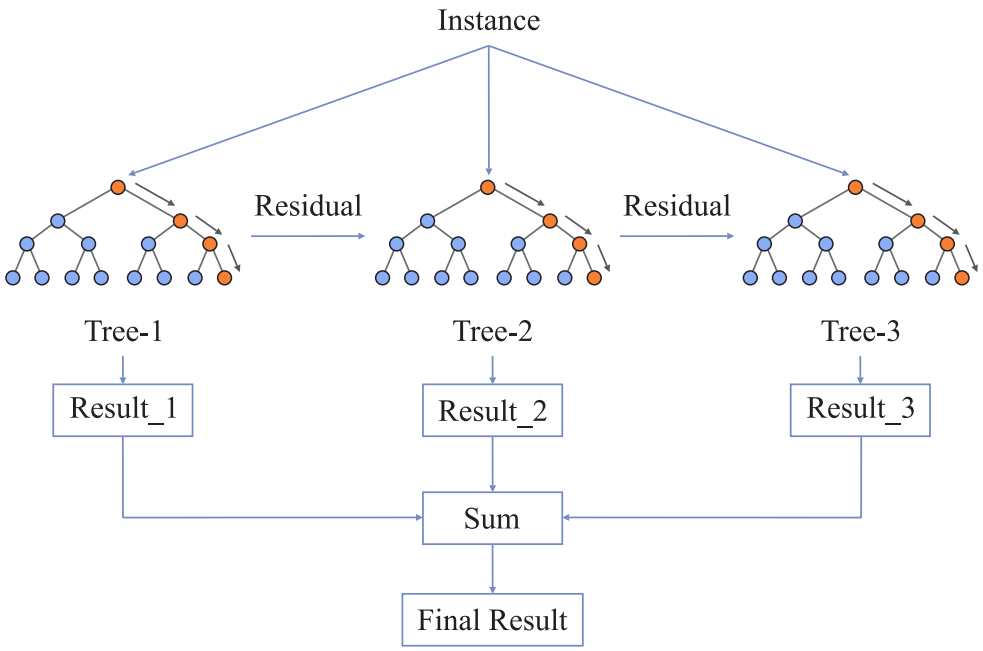


Figure 12. Simplified structure of XGBoost (Wang et al. 2020)

[24] See https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Slight differences in hyperparameter values of both random forest and XGBoost regressors can lead to significant changes in outcomes. Hyperparameters used in the random forest regressor function[25] are given in Figure 13 and detailed below:

- *max_depth:* The maximum depth of the tree in the forest. The complex relationships in the data can be captured by a deeper tree, but the loss in exchange for it increases the risk of overfitting.
- *max_features:* The number of features to consider when splitting a node. Setting '*auto*' equalizes the hyperparameter to number of features and setting smaller values than 1 increases the randomness.
- *min_samples_leaf:* The minimum number of samples required to be at a leaf node. Together with *min_samples_split,* the higher values can lead to underfitting while reducing them can also increase the risk of overfitting.
- *min_samples_split:* The minimum number of samples required to split an internal node.
- *n_estimators:* The number of trees in the forest. Increasing the '*n_estimators*' value can improve the accuracy, but also yields longer training time.

```
RFR_model  =  RandomForestRegressor(max_depth = 50,
              max_features= 'auto',
              min_samples_leaf= 1,
              min_samples_split= 7,
              n_estimators= 75)
```

Figure 13. Random forest regression parameters

Like random forest regressor, tuning the hyperparameters of the XGBoost regressor function[26] can improve its performance. In addition to the 'n_estimators' and 'max_depth' of the random forest regressor, several more hyperparameters used in XGBoost regressor can be found in Figure 14 and are detailed below:

- *colsample_bytree:* The ratio of subsample columns used to train each tree. The values less than 1 can reduce overfitting.
- *lambda:* The higher value of the lambda the more conservative the model. It is the L2 regularization term for avoiding overfitting, and the default value is 1.
- *learning_rate:* The step size shrinkage used to prevent overfitting. The smaller rates can increase the training time.
- *min_child_weight:* Minimum sum of instance weight (hessian) needed in a child. It can reduce the overfitting risk, and as it takes greater values than the default value 1, the algorithm becomes more conservative.

[25] See https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
[26] See https://xgboost.readthedocs.io/en/stable/parameter.html

```
XGB_model   = xgb.XGBRegressor(colsample_bytree: 0.5,
              lambda: 1,
              learning_rate: 0.01,
              max_depth: 5,
              min_child_weight: 5,
              n_estimators: 1000)
```

Figure 14. XGBoost parameters

Both random forest (Breiman 1996) and XGBoost regressors benefit from K-fold cross-validation and hyperparameter search using GridSearchCV to achieve promising improvements. K-fold cross-validation is a type of cross-validation that involves repeating the process of splitting a dataset into K subsets/folds. One part is used for validation and the remaining parts (K-1) are united to be a training dataset. The model is trained and evaluated on each fold separately. In the end, each model is being fitted on a partly overlapping training set and evaluated on a distinct validation set. The overall performance is calculated as the average of K performance estimates from the validation sets (Raschka 2018). The illustration of the process is shown in Figure 15.

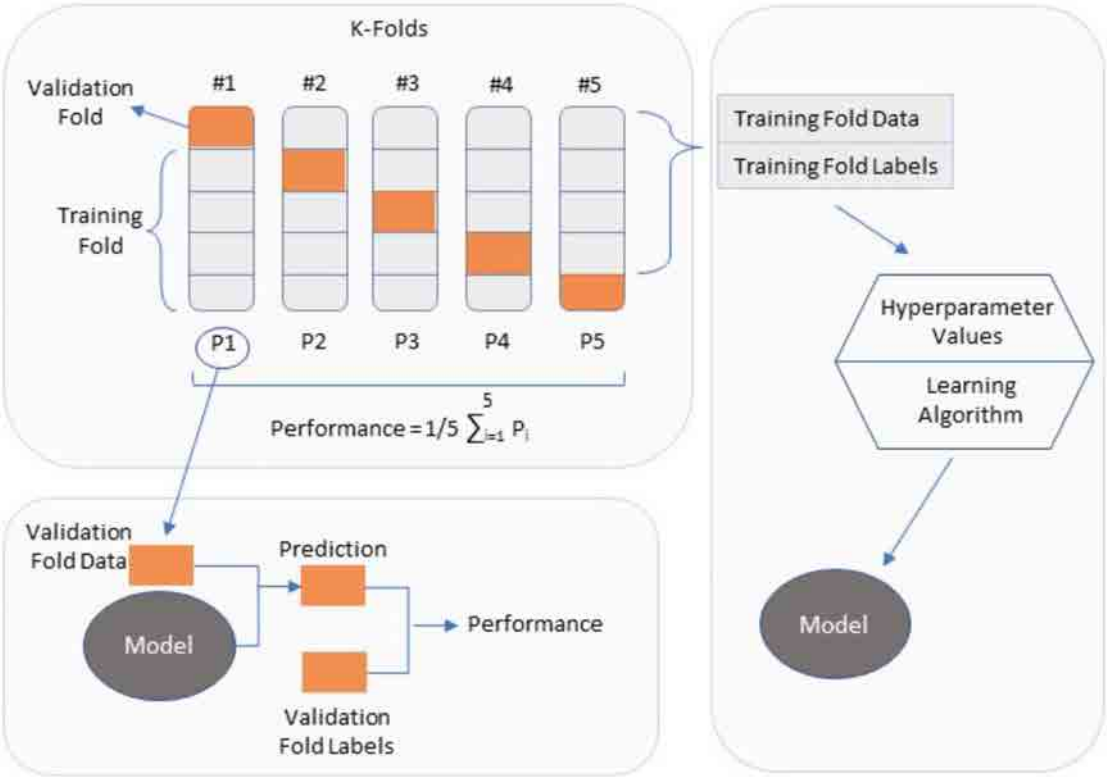


Figure 15. K-fold cross-validation, hyperparameter tuning, training and testing the model. Adapted from Raschka (2018)

The GridSearchCV[27] function of scikit-learn conducts an exhaustive search over a specified hyperparameter grid (*param_grid*) and returns the best combination of hyperparameters. The following code cell consists of a XGBoost regressor model (*xgbr* as an estimator) and the fivefold cross-validation (if cv is set to None it uses fivefold by default) to obtain the best combination of hyperparameters. Evaluation metrics are defined as '*r2*' and '*neg_mean_squared_error*' in the scoring parameter.

Additional optional parameters used in GridSearchCV are '*refit*' and '*return_train_score*'. When set to True, '*refit*' uses the best-found parameters on the whole dataset to refit the estimator (*xgbr*), while when '*return_train_score*' is set to True, it includes training scores in attributes of provided output. The delivered combination of parameters should provide the highest model performance.

Having a very high number of estimators, or a depth of the trees, does not necessarily mean the most accurate results, as there is a certain value for a parameter for which no further improvement in accuracy is evident (Karshiev et al. 2020). GridSearchCV automates the process of finding a combination of these parameters.

```
scoring = ["r2", "neg_mean_squared_error"]

xgbr = xgb.XGBRegressor()
start = timeit.default_timer()
clf = GridSearchCV(estimator=xgbr,
                   param_grid=params,
                   scoring=scoring,
                   refit = True,
                   cv = None,
                   return_train_score=True,
                   verbose=2)
clf.fit(X_train, y_train['SV_scaled'])
```

Figure 16. GridSearchCV function with a XGBoost regressor and a fivefold cross-validation

3.5.2 Neural nets

The last part of this chapter is the implementation of the classifier model. Although a conventional approach for this prediction requires regression models, utilizing the MLP classifiers another attempt to get improved results. For the MLP classifier from sklearn, we grouped SV in classes of roughly evenly populated groups ['Low', 'Medium', 'High'] and ['Very Low', 'Low', 'Medium', 'High', 'Very High'] and we trained a neural network with five hidden layers. We experimented with multiple setups, including different layers, depths, iteration sizes, learners and activation functions.

One instance of the function setups with mentioned hyperparameters is provided in Figure 17. We also performed GridSearchCV for better model hyperparameters. This setup has generally been the most efficient with the highest average accuracy score for the test datasets. The total accuracy scores for the MLP and results obtained with the regressors will be provided in the next chapter.

[27] See https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

```
NN_model = MLPClassifier(hidden_layer_sizes=(200, 400, 200, 100, 50),
                         max_iter = 3000, activation = 'relu',
                         learning_rate = 'adaptive',
                         alpha = 0.0001,
                         solver = 'sgd')
```

Figure 17. Chosen model for NN after model selection with GRID CV and multiple inputs

3.6 Results and discussion

The model performance evaluation is assessed using several error metrics. We use different methods to evaluate classification and regression models.

3.6.1. Model evaluation: Neural net

Table 9 shows model scores using three and five classes. Already the training score shows low performance. Test scores of 60 percent are the lowest acceptable value, although none of the models reached that accuracy.

Table 9. Error metrics of MLP

Model	Train Accuracy	Test Accuracy	Target classes
NeuralNet (MLP)	0.75	0.59	3 classes*
NeuralNet (MLP)	0.61	0.41	5 classes*

* Classes: ‘Low, Medium, High’ or ‘Very low, Low, Medium, High, Very High’

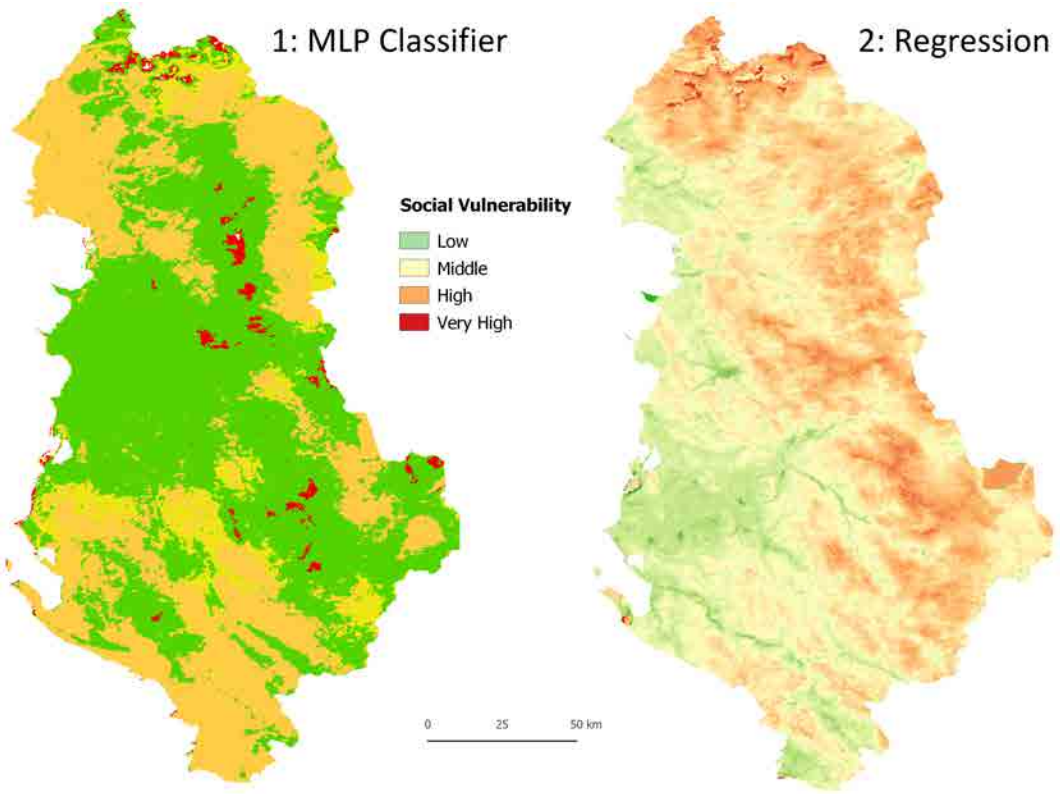


Figure 18. Side-by-side comparison with neural net and regression prediction for Albania. Left: 5 classes of vulnerability based on multilayer perceptron model; Right: Continuous SV based on multiple stepwise linear regression

The left-hand map (1) in Figure 18 shows the result of the five-class prediction of the chosen multi-layer perceptron model in Albania. The distribution of predicted values goes according to the five categories we defined, based on the initial floating point number SV with a range between 0 and 1. The new five classes were derived, based on an approach to minimize class imbalances, by manually adjusting the sampling weights and class breaks. We achieved a homogeneous class distribution of approximately 1/5 sample size for each class. The right-hand map is a prediction of a stepwise multiple regression model with cross-validation and hyperparameter optimization. Our initial results using neural nets were not very promising, but it seems to be more beneficial to use regression strategies to model this relationship.

3.6.2. Model evaluation: Regression

Root mean square error (RMSE) is used as one of the performance criteria for our results. RMSE is the standard deviation of the residuals (prediction errors).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Where $\hat{y}_i$ are the predicted values, y_i the observed values and n is the number of observations. Residuals are a measure of how far from the regression line data points are, while $RMSE$ is a measure of how spread out these residuals are. In other words, it shows how concentrated the data is around the line of best fit.[28]

The mean squared error (MSE) is defined as the average squared distance between the predicted and the true values. It is also a loss function used for regression tasks as it squares the errors and therefore makes the model less robust to the outliers. Good models should have MSE close to zero.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ are the predicted values, y_i the observed values and n is the number of observations.

Furthermore, we looked at the coefficient of determination of variation R^2 to understand how much of the total variance was explained by the model.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Where SS_{RES} are the sum of squared residuals and SS_{TOT} the total sum of squared error. Both values, the $RMSE$ and R^2 explain much about the quality of a model. We were also able to compare the scores with other regression models since these could be interpreted by the same evaluation metrics.

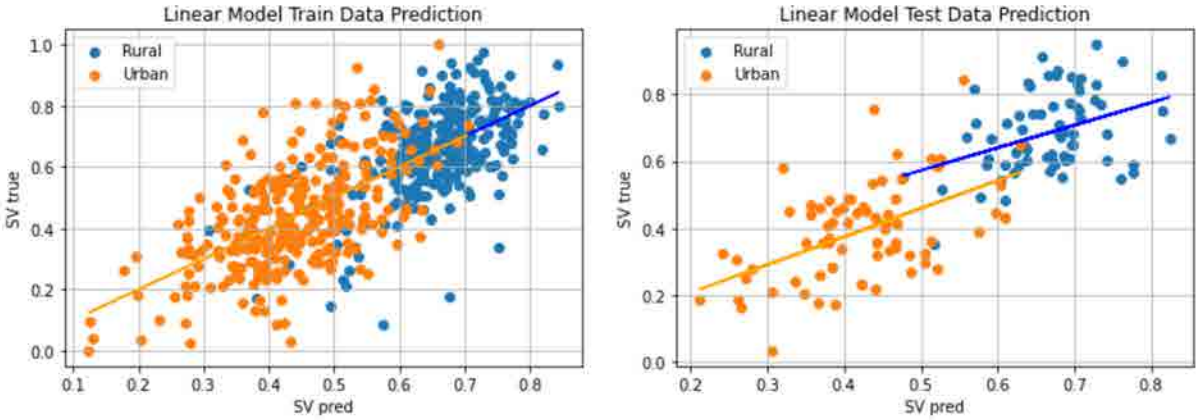


Figure 19. Scatterplot of predicted and ground truth SV (n = 715) by using a stepwise linear regression for urban and rural clusters in Albania. Orange and blue lines: best fit for samples split for urban and rural samples.

We ran multiple regression analysis with different sets of variables and different scaling methods. The best results were obtained by choosing the full set of variables (as shown in Table 10) or the reduced amount after intercorrelation elimination.

[28] See https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

Table 10. Error metrics of regression models

Model	R^2	RMSE‡	MSE
Linear model (base model)	0.68	0.124	0.021
Random forest	**0.72**	**0.10**	**0.01**
XGBoost	**0.72**	**0.103**	**0.0106**
MLP regressor	0.63	0.11	0.013
Huber regressor	0.66	0.10	0.012
Ridge regressor	0.69	0.10	0.011
Lasso regressor	0.65	0.10	0.012
Decision tree regressor	0.52	0.13	0.017

** Best models were determined with K-fold cross-validation*

‡ Root mean square error: Lower is better; Green colour: Best performing models, used for high-resolution map

Linear regression assumes linearity between the variables, which in the real world is almost never the case. If the multicollinearity between independent variables is not removed, linear regression will not show satisfactory performance (Shrestha 2020). Even though we considered these limitations and fed the model with the well pre-processed input data, the model still shows lower performance compared to others.

Decision tree regressor shows the poorest performance of the mentioned models. Decision tree regressor is not the best option when dealing with continuous numerical variables, as a small change in data can cause large differences in the tree structure (Gulati et al. 2016). This is confirmed due to low R^2 and high MSE.

Promising improvements were achieved by using K-fold cross-validation and hyperparameter search using GridSearchCV for regression. The best results were obtained by random forest regression and XGBoost regression. We observed that most models converged around R^2 of 0.72 or RMSE of ~ 0.10 – 0.12 for the test country, Tajikistan.

Random forest regression performs well on continuous values, reduces overfitting and is well suited for regression tasks. However, random forest regression yields a trade-off between the training time and the number of trees. The increasing number of trees requires more computational time and space, although it can improve accuracy but only until the number of trees reaches a certain value.

Above that value, no significant model improvements can be found (Karshiev et al. 2020). On the other hand, if the number of trees is too small, there is a possibility of underfitting (Han et al. 2020). Using GridSearchCV and K-fold cross-validation we received the most accurate results with 75 trees and a maximum depth of 50.

As XGBoost regression combines methods of regression trees and gradient boosting, we are able to tune the great set of hyperparameters. The high dimensionality of the data leads to high memory consumption, making the model costly-insufficient. XGBoost applies the learning rate to the loss function to ensure the minimum loss (Chen and Guestrin 2016). The resulting maps of this process can be seen in Figure 22. Figure 20 shows the resulting scatterplots with train and test data prediction and fitted lines (orange and blue). The modelling improvements can be viewed in our Notebook 5 of the DSVI repository on GitHub.[29]

[29] https://github.com/SDG-AI-Lab/DSVI_Tajikistan (to access site authors' authorization is required).

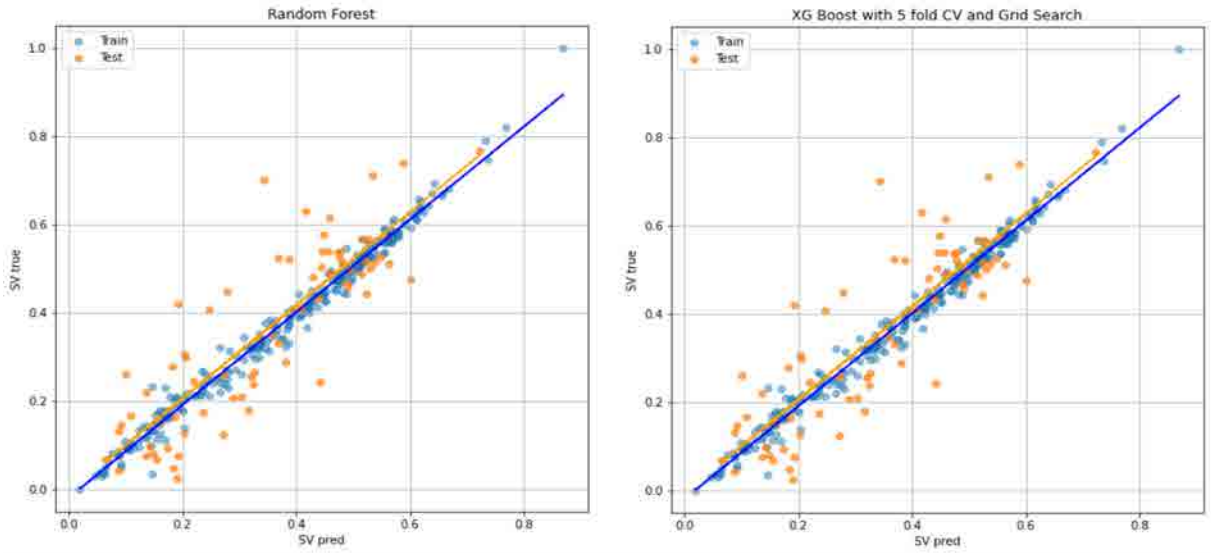


Figure 20. Improved modelling results with using K-fold random sampling and GridSearchCV, Tajikistan

The influence of some predictor variables, such as the '*celltower density*' or '*road network per km²*, are also visible and seem to have been regarded positively by the models. The results show differences of scores in some regions of the country and they also weigh the importance of some input variables differently. The predictor variables feature importance is shown in Figure 21.

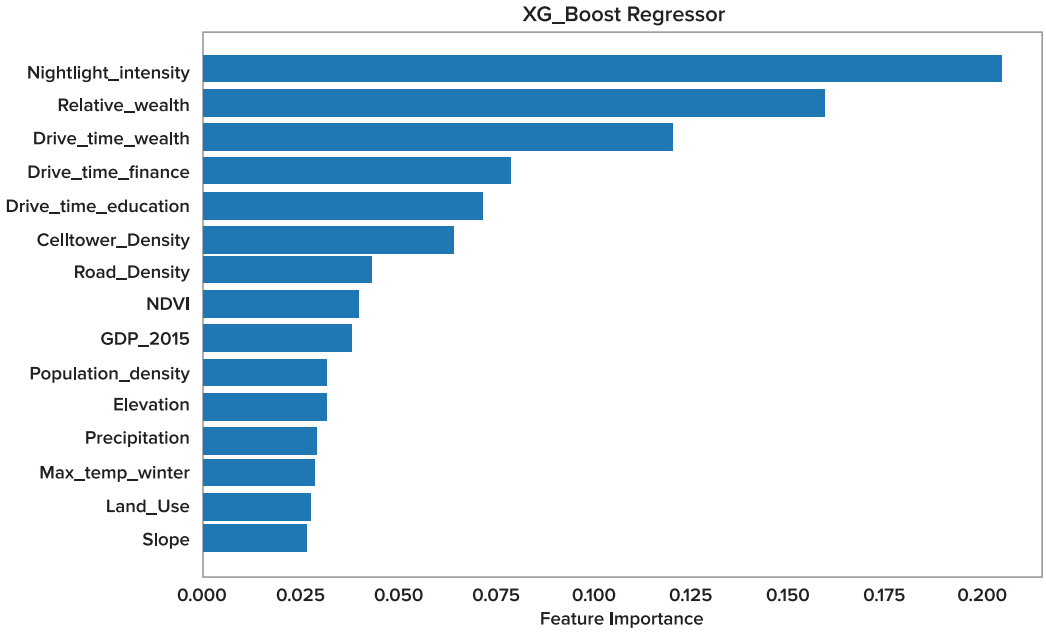


Figure 21. Feature importance for XGBoost regressor in the case of Tajikistan

'Feature importance' refers to techniques that calculate a score for all the input features for a given model – the scores simply represent the 'importance' of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable.[30] Feature importance is not a perfect or absolute measure for variable influences on a model. They are an indication and can help understand general trends and relationships between datasets and models.

[30] Terence Chin, 'Understanding Feature Importance and How to Implement it in Python', https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285

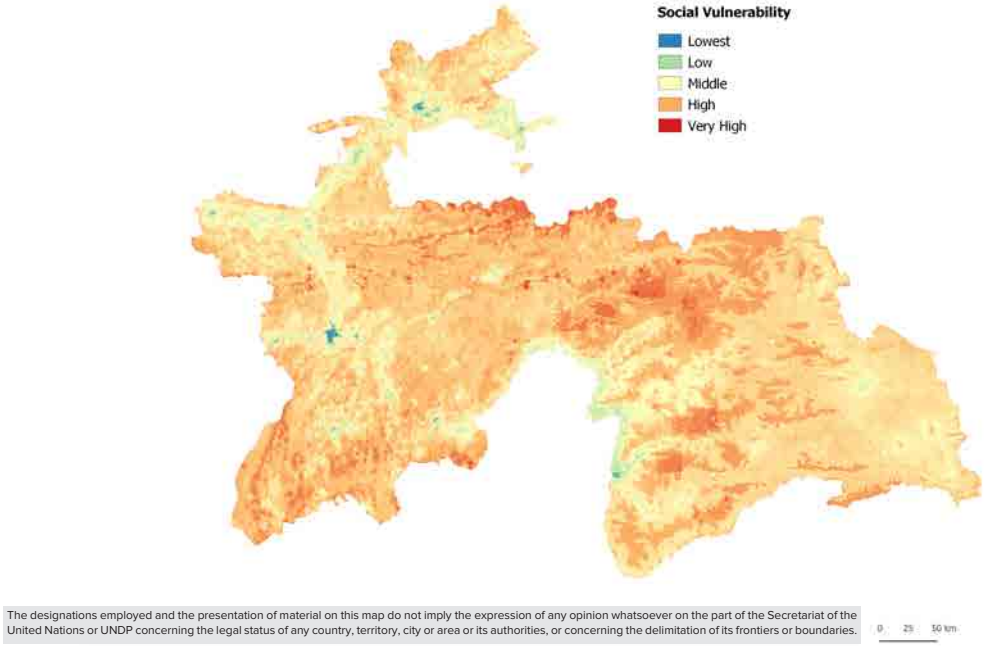


Figure 22. Improved prediction for Tajikistan with XGBoost

Figure 22 shows the XGBoost model results after hyperparameter tuning. Social vulnerability tends to increase in remote areas, such as mountainous places. Lower SV scores tend to cluster in urban areas, particularly in the capital area of Dushanbe in the western centre of the country.

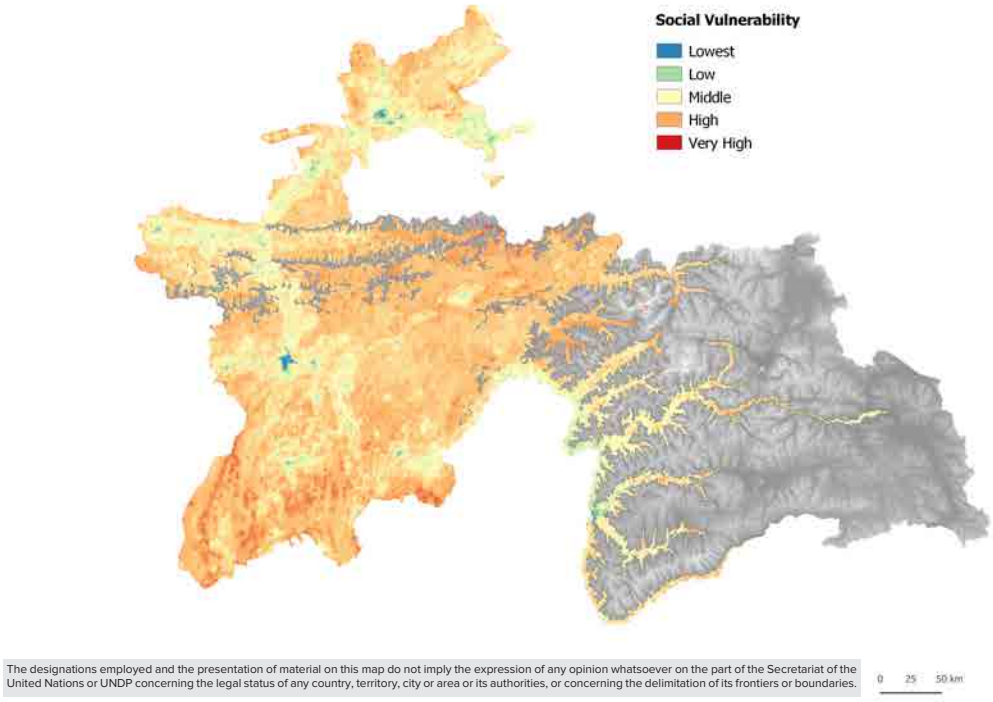


Figure 23. SV scores masked with elevation above 3,650 m (highest populated place in Tajikistan)

Figure 23 is an example for a final social vulnerability map that could be considered a final output of DSVI. The map illustrates the social vulnerability scores for Tajikistan, while masking uninhabited areas. The eastern regions of Tajikistan, dominated by high mountains and alpine conditions, are mostly uninhabited. Generally, social vulnerability scores do not make a lot of sense for areas which are potentially uninhabitable, such as alpine environments, water bodies, dense forests, or other extreme biospheres and climate zones. Another case is regions that are unpopulated, but potentially habitable: Looking at SV scores in those regions could help to open up new views on settlement options in them and potential risks.

The lowest social vulnerability scores are in the urban areas of the country. The capital city, Dushanbe, in the centre west of the country, with well below average vulnerability, is most notable in that regard. Social vulnerability tends to increase with growing elevation and with increasing distance to areas with high economic output as measured by nightlight intensity. Other factors are the distances to health infrastructure, finance and education. Some areas in the south of the country have relatively high vulnerability scores because of the input variable '*relative wealth*' and its relatively poor representation in that area.

4. DSVI online tool

The web application was created as an interactive and user-friendly tool to enable users to communicate effectively with the core elements of the DSVI. The DSVI online tool provides a feature-rich interface that allows users to fully experience not only the contextualized final prediction results, but also the intermediate outcomes like SV scores and certain controlling functionalities. The online tool can easily be accessed by using any web-browser. It is lightweight and does not consume a lot of resources, which means it can also be run in difficult environments with less bandwidth or unfavourable hardware specifications. The design of the tool is meant to be intuitive and easy to use, but it can still offer in-depth analysis to users wanting to conduct more complex analytics. It was developed in collaboration with UN Online Volunteers. The used technologies are react-leaflet,[31] next.js[32] and geoserver[33] for database management and remote access.

The base functions of the tool are similar to other online mapping tools, such as the following ones:

- Zoomable, integrated map with different base maps
- Data layers to show social vulnerability as points, aggregated or with high resolution
- Data layers to show the biophysical realities in the region or country of interest
- Layers to show survey data points and summary statistics of survey characteristics

[31] See https://react-leaflet.js.org/
[32] See https://nextjs.org/
[33] See https://geoserver.org/

The online tool has core features that make it distinct from other vulnerability analysis tools:

- Planned: Options to change colour and display of SV scores
- Planned: Options to show and change model parameters
- Planned: Analysis mode to derive further insights into vulnerable groups and their position relative to critical infrastructure, regions with high disaster risk, or similar factors
- Planned: Dashboards to summarize statistics regarding the used parameters, vulnerabilities and survey data accessed by users
- Additional features based on user requirements

The online tool is intended to be used by various user groups. User groups can be private individuals who were granted access by the partnering organization, professionals, policymakers, stakeholders, or technical analysts. The list of users is not limited to the previous examples. The users can access diverse functionalities and data of the tool based on their assigned user group. Further development of the tool will include a user-based system to control the flow of information based on the type of user accessing the tool. The 2022 version of the tool can be viewed in Figure 24.

The designations employed and the presentation of material on this map do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations or UNDP concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries.

Figure 24. Digital social vulnerability tool showing the main map window

5. Conclusion and implications

The methodology builds on established scientific methods and literature, utilizes extensive and diverse datasets and combines them with a digital approach to create the Digital Social Vulnerability Index with GIS and machine-learning techniques. The nature of the chosen process brings flexibility in calculating the scores with different sets of variables and different settings for the modelling parameters.

Since we defined social vulnerability scores as ground truth with no independent reference in the real world, it is difficult to validate the gridded high-resolution SV with external data sources. We observed that the SV scores we calculated were reasonable in the sense that they followed the general expectations of human development and well-being indicators for the countries we tested. But, on the other hand, SV should also display certain unexpected discrepancies, which should be revealed by the distinct nature of what SV tries to measure compared to other indicators.

It is expected that SV scores tend to correlate with similar indicators, such as general wealth or a predefined relationship with access to basic infrastructure. Since there are no available other social vulnerability maps or datasets in a gridded representation, it is not possible to directly validate the scores of the high-resolution maps other than using the accuracies of the models. One solution to this problem is the examination of the scores with the help of experts and making critical assessments of their validity this way.

The implemented pipelines, SV calculations and machine-learning techniques automate the technical process of calculations in the background and unlock the results in a user-friendly and interactive manner. They also decrease the time and costs necessary to perform the analysis manually. However, since a reasonable interpretation of social vulnerability scores is dependent on expert opinion, the processes must be supervised to ensure high-quality outputs.

The presented approach can be implemented with available DHS datasets, and also with country-specific geotagged surveys provided by users. Geographical data is broadly available for the SV prediction process to obtain new high-resolution maps.

DSVI satisfies the need for modern digital solutions and will strengthen the much-needed internal digital capacities of UNDP. Being in accordance with the UNDP strategic plan and policies, DSVI presents a tool that is both able to calculate the social vulnerability scores and to identify the main drivers or indicators. This is valuable information to organizations, stakeholders, policymakers and others to plan further corresponding actions, investments and initiatives to reduce the vulnerability and to strive towards achieving the Sustainable Development Goals.

6. References

Adger, W. Social and Ecological Resilience: Are They Related? *Progress in Human Geography*. 24 (2000), pp. 347-364. https://doi.org/10.1191/030913200701540465.

Blaikie, P., Cannon, T. Davis, I and Wisner, B. *At Risk: Natural Hazards, People's Vulnerability and Disasters*. London: Routledge, 1994.

Braeken, Johan and Assen, Marcel. An Empirical Kaiser Criterion. *Psychological methods*. 22 (3) (2016). https://doi.org/10.1037/met0000074.

Breiman, Leo. "Bagging Predictors." *Machine Learning* 24, No. 2 (1996), pp. 123-140.

_____. "Random forests." Machine learning 45, no. 1 (2001), pp. 5-32.

_____. Classification and regression trees. Routledge, 2017.

Chen, Tianqi and Guestrin, Carlos. "XGBoost: A Scalable Tree Boosting System." *Association for Computing Machinery* (2016), pp. 785-794. https://doi:10.1145/2939672.2939785.

Chen, W., Cutter, S.L. and Emrich, C.T. Measuring social vulnerability to natural hazards in the Yangtze River Delta region, China. *Int J Disaster Risk Sci* 4 (2013), pp.169-181. https://doi.org/10.1007/s13753-013-0018-6

Chi, Guanghua, Fang, Han, Chatterjee, Sourav and Blumenstock, Joshua E. Micro-Estimates of Wealth for all Low- and Middle-Income Countries, *Economic Sciences*, Vol. 119 (3) (2022).

Cutter, Susan L. *American Hazardscapes: The Regionalization of Hazards and Disasters*. Washington, D.C.: Joseph Henry Press, 2001.

Cutter, S. L, Emrich, C. T, Morath, D .P. and Dunning, C. M. Integrating social vulnerability into federal flood risk management planning. *Journal of Flood Risk Management*. 6 (2013), pp. 332-344. https://doi: 10.1111/Jfr3.12018

Cutter, Susan L., Boruff, Bryan J. and Shirley, W. Lynn. "Social Vulnerability to Environmental Hazards." *Social Science Quarterly* 84, No. 2 (2003), pp. 242-261. http://www.jstor.org/stable/42955868.

De Loyola Hummell, B.M., Cutter, S.L. and Emrich, C.T. Social Vulnerability to Natural Hazards in Brazil. *Int J Disaster Risk Sci* 7 (2016), pp.111–122. https://doi.org/10.1007/s13753-016-0090-9

Del Serrone, Giulia and Moretti, Laura. A stepwise regression to identify relevant variables affecting the environmental impacts of clinker production. *Journal of Cleaner Production*, Vol. 398 (2023). https://doi.org/10.1016/j.jclepro.2023.136564.

Grace, K., Nagle, N. N., Burgert-Brucker, C. R., Rutzick, S., Van Riper, D. C., Dontamsetti, T. and Croft, T. Integrating Environmental Context into DHS Analysis While Protecting Participant Confidentiality: A New Remote Sensing Method. *Population and development review,* 45(1) (2019), pp.197-218. https://doi.org/10.1111/padr.12222

Guillard-Gonçalves, Clémence and Zêzere, José. Combining Social Vulnerability and Physical Vulnerability to Analyse Landslide Risk at the Municipal Scale. *Geosciences* 2018, 8 (8) 294. https://doi.org/10.3390/geosciences8080294

Gulati, Pooja, Sharma, Amita and Gupta, Manish. "Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review." *International Journal of Computer Applications* 141(14) (2016), pp. 19-25. https://doi:10.5120/ijca2016909926.

Han, Sunwoo, Kim, Hyunjoong and Lee, Yung-Seop. Double random forest. *Machine Learning* 109 (2020), pp.1,569-1,586. https://doi.org/10.1007/s10994-020-05889-1

Hoerl, Arthur E., and Kennard, Robert W.. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12, No. 1 (1970), pp. 55-67. https://doi:10.1080/00401706.1970.10488634.

Huber, P. J. "Robust estimation of a location parameter." *Annals of Mathematical Statistics* 35, No. 1 (1964), pp. 73-101. https://doi:10.1214/aoms/1177703732.

James, Gareth, Witten, Daniela, Hastie, Trevor and Tibshirani, Robert. *An Introduction to Statistical Learning: with Applications in R.* New York: Springer, 2021. https://doi:10.1007/978-1-0716-1418-1_3.

Jolliffe, I. T. and Cadima, J. Principal Component Analysis: A Review of Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2016). https://doi:10.1098/rsta.2015.0202.

Karshiev, Sanjar, Bekhzod, Olimov, Kim, Jaesoo, Anand, Paul and Jeonghong, Kim. Missing Data Imputation for Geolocation-based Price Prediction Using KNN–MCF Method. *ISPRS International Journal of Geo-Information* (2020) 9 (4) 227. https://doi.org/10.3390/ijgi9040227.

Katic, Krunoslav. Social vulnerability assessment tools for climate change and DRR programming. UNDP, 2017.

Müller, Annemarie, Kerle, Norman and Stein, Alfred. Urban social vulnerability assessment with physical proxies and spatial metrics derived from air- and spaceborne imagery and GIS data. Natural Hazards (2008) 48, pp. 275-294. https://doi:10.1007/s11069-008-9264-0.

Raschka, Sebastian. "Model evaluation, model selection, and algorithm selection in machine learning." Cornell University, 2018. arXiv:1811.12808v3

Rufat S., Tate E., Emrich C. and Antonelli G. How valid are social vulnerability models? *Assoc Geogr* 109 (4) (2019), pp.1,131-1,158.

Rygel, L., O'Sullivan, D. and Yarnal, B. "A Method for Constructing a Social Vulnerability Index: An Application to Hurricane Storm Surges in a Developed Country" in *Mitigation and Adaptation Strategies for Global Change,* Springer, Vol. 11(3) (2006), pp. 741-764.

Shrestha, Noora. "Detecting Multicollinearity in Regression Analysis." *American Journal of Applied Mathematics and Statistics* 8, No. 2 (2020), pp. 39-42.

Spielman Seth , Tuccillo, Joseph, Folch, David, Schweikert, Amy, Davies, Rebecca and Wood, Nathan J.. Evaluating social vulnerability indicators: Criteria and their application to the Social Vulnerability Index. *Natural Hazards*, Vol. 100, Issue 1 (2019), pp. 417-436. https://doi.org/10.1007/s11069-019-03820-z

Su, Y., Gao, X., Li, X. and Tao, D. “Multivariate Multilinear Regression,” in *IEEE Transactions on Systems, Man, and Cybernetics,* Part B (Cybernetics), Vol. 42, No. 6, Dec. (2012), pp. 1,560-1,573. https://doi: 10.1109/TSMCB.2012.2195171.

Tierney, K. J., Lindell, M. K. and Perry, R. W. *Facing the Unexpected: Disaster Preparedness and Response in the United States*. Washington, D.C.: Joseph Henry Press, 2001.

Wang, P., Huang, C., Brown de Colstoun, E.C., Tilton J.C. and Tan, B. Global Human Built-up and Settlement Extent (HBASE) Dataset From Landsat. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), 2017. https://doi.org/10.7927/H4DN434S

Wang, Qinggang, Koval, John, Mills, Catherine and Kang-In, David Lee. Regression Analysis. “Determination of the Selection Statistics and Best Significance Level in Backward Stepwise Logistic Regression.” *Communications in Statistics - Simulation and Computation*, Vol. 37, Issue 1 (2007), pp. 62-72. https://doi:10.1080/03610910701723625.

Wang, Weilun, Chakraborty, Goutam and Chakraborty, Basabi. “Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm.” *Applied Sciences* 11 (December 2020) 202. https://doi.org/10.3390/app11010202.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R. and Emch, M. E. Influence of Demographic and Health Survey Point Displacements on Distance-Based Analyses. *Spatial demography*, 4(2) (2016), pp.155–173. https://doi.org/10.1007/s40980-015-0014-0

Willis, I. and Fitton, J. A review of multivariate social vulnerability methodologies: a case study of the River Parrett catchment, UK, *Nat. Hazards Earth Syst. Sci.*, 16 (2016), pp. 1,387-1,399. https://doi.org/10.5194/nhess-16-1387-2016

7. Annex

DHS displacement correction

DHS geotagged survey clusters come with an artificial displacement to protect the privacy of interviewed households and to prevent geolocating individual households. For urban clusters, this displacement is done randomly within a radius of 2 km. This means that an urban classified cluster could potentially be relocated between 0 and 2 km from its original position in all possible directions on a circle. The same rule applies to rural clusters, which can be dislocated by up to 5 km.

However, this displacement is of a random nature, so an average distance to the original point will be half of the displacement value, ergo only 1km on average for urban clusters and 2.5 km on average for rural ones. This can pose a problem because if we predict a value uniquely attributed to the true position of a survey point, we need to make sure that the deviation from its original value is as small as possible. In general, studies have tried to explain the resulting error (Warren et al. 2016) and proposed solutions similar to the one we used (Grace et al. 2019).

Based on these insights, we used the 'Global Human Built-up and Settlement Extent (HBASE)' Dataset from Landsat, v1 (2010)[34] to reduce the introduced variance to the highly dispersed rural cluster points. We assumed that a given cluster point must have been taken within a human settlement, or with a very high likelihood thereof. Some of the rural settlements are very small compared to the urban ones and are hard to locate. The HBASE dataset is based on 30 m pixel resolution and indicates the existence of human built-up areas for the whole country.

We transformed the data into a fishnet grid and snapped it as rural classified cluster points that were within a radius of 5 km to the nearest intersection point of the human settlement. If a point already resided within a settlement area, we skipped the procedure. This procedure ensures that the sampled point is as close as possible to a settled location near or exactly on its original location.

In rural areas, there are often only a few settlements, or one human settlement, within a radius of 5 km or less to a DHS survey cluster point.

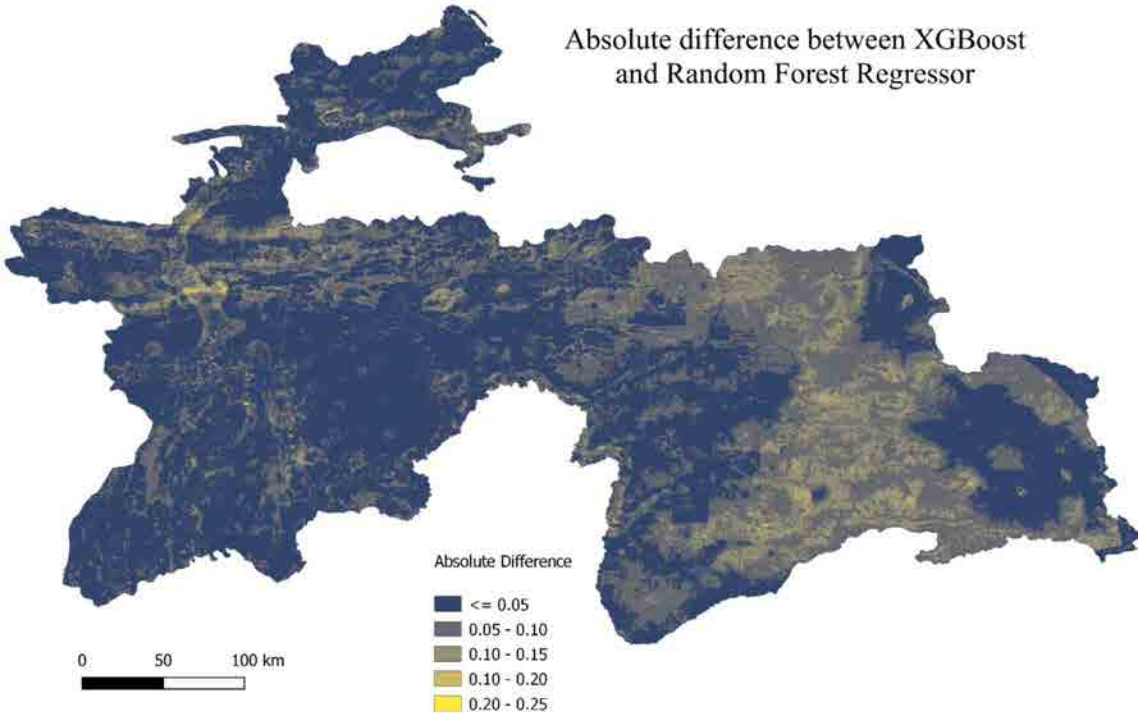
Even if the correct settlement was not located with this procedure, the general characteristics of settlements in rural areas represent each other better than a randomly placed point and therefore should help to generate a more useful dataset for this type of analysis. For our SV index and high-resolution map (see section 3), we used original uncorrected and corrected points respectively to assess the differences of model responses.

[34] NASA. Socio-Economic Data and Applications Center (SEDAC), Global High-Resolution Urban Data from Landsat, https://sedac.ciesin.columbia.edu/data/set/ulandsat-hbase-v1/data-download

Table A1. List of used geodata for test case in Albania with descriptive statistics

Index	Variable Name	Mean***	Min	Max	Std**
1	Distance to Airport*	345.54	1.00	870.46	166.60
2	Distance to Coast*	-52.15	-150.00	23.00	39.81
3	Distance to Education Facility*	178.90	1.41	705.57	129.45
4	Elevation in meter	692.59	-7.00	2605	573.79
5	Distance to Financial Facility*	172.33	0.00	768.52	131.67
6	Distance to Health Facility*	177.40	0.48	641.34	176.33
7	Distance to Health Facility 2*	197.98	1.00	769.92	150.73
8	NDVI	0.58	-0.08	0.92	0.26
9	Nightlight Intensity	0.41	0.00	116.76	2.59
10	Population Density	95.32	0.00	13115.94	458.55
11	Precipitation Average	128.27	0.00	335.00	56.77
12	Distance to Road*	75.11	0.00	656.40	124.17
13	Temperature Average (July)	23.62	0.00	32.70	8.98
14	Distance to River*	86.76	0.00	669.65	126.39
15	Population Density Women	0.34	0.00	10.12	0.64

** As calculated per Euclidian Distance (Heatmap) ** Standard deviation, *** Arithmetic mean*



The designations employed and the presentation of material on this map do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations or UNDP concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries.

Figure A1. Absolute differences in SV predicted scores for the two best performing models. The models agree less in the most remote and mountainous regions of Tajikistan

The mean differences between the two models are not equally distributed and large proportions fall into the uninhabited areas of Tajikistan (Figure A1), which are situated roughly 3,650 m above sea level. The mean disagreement between the models across the whole country is 0.05 with a standard deviation of 0.035. For areas below 3,650m the average discrepancy between the models lowers to only 0.013 with a standard deviation of 0.031.

Table A2. Indicators and correlation with geodatasets

GROUP	Indicators derived from DHS Albania	Strong Correlation with Geodata
Socio-economic	Wealth index combined*	**Wealth index combined***
	Respondent currently working	**Respondent currently working**
	Has an account in a bank or other financial institution	**Has an account in a bank or other financial institution**
	Getting medical help for self: getting money needed for treatment	
	How often uses internet	**How often uses internet**
Demographics	Respondent's current age	
	Education in single years	**Education in single years**
	Years lived in place of residence	**Years lived in place of residence**
	Highest educational level	Highest educational level
Family structure	Number of household members (listed)	**Number of household members (listed)**
	Number of children 5 and under in household (de jure)	
	Number of eligible women in household (de facto)	
	Sex of household head	**Sex of household head**
	Age of household head	**Age of household head**
Medical services	Covered by health insurance	**Covered by health insurance**
	Smokes cigarettes	**Smokes cigarettes**
	Getting medical help for self: concern no provider	
	Getting medical help for self: concern no drugs available	
	Getting medical help for self: concern that there may be no supplies	
	Getting medical help for self: distance to health facility	
	Getting medical help for self: concern no female health provider	
	Had any STI in last 12 months	

Urban	Urban / Rural	
Built environment vulnerability	Main floor material	**Main floor material**
	Main roof material	
	Main wall material	**Main wall material**
	Time to get to water source	
	Owns a mobile telephone	**Owns a mobile telephone**
	Household has radio	**Household has: radio**
	Result of salt test for iodine	
Social capital	Beating justified if wife goes out without telling husband	**Beating justified if wife goes out without telling husband**
	Beating justified if wife neglects the children	**Beating justified if wife neglects the children**
	Beating justified if wife argues with husband	**Beating justified if wife argues with husband**
	Beating justified if wife refuses to have sex with husband	**Beating justified if wife refuses to have sex with husband**
	Getting medical help for self: getting permission to go	
	Beating justified if wife burns the food	
	Getting medical help for self: having to take transport	
	Getting medical help for self: not wanting to go alone	**Getting medical help for self: not wanting to go alone**
	Frequency of reading newspaper or magazine	**Frequency of reading newspaper or magazine**
	Frequency of listening to radio	
	Frequency of watching television	**Frequency of watching television**
Family structure	Number of household members (listed)	**Number of household members (listed)**
	Number of children 5 and under in household (de jure)	
	Number of eligible women in household (de facto)	
	Sex of household head	**Sex of household head**
	Age of household head	**Age of household head**

Medical services	Covered by health insurance	**Covered by health insurance**
	Smokes cigarettes	**Smokes cigarettes**
	Getting medical help for self: concern no provider	
	Getting medical help for self: concern no drugs available	
	Getting medical help for self: concern that there may be no supplies	
	Getting medical help for self: distance to health facility	
	Getting medical help for self: concern no female health provider	
	Had any STI in last 12 months	
Urban	Urban / Rural	**Urban / Rural**
Built environment vulnerability	Main floor material	**Main floor material**
	Main roof material	
	Main wall material	**Main wall material**
	Time to get to water source	
	Owns a mobile telephone	**Owns a mobile telephone**
	Household has: radio	**Household has: radio**
	Result of salt test for iodine	
Social capital	Beating justified if wife goes out without telling husband	**Beating justified if wife goes out without telling husband**
	Beating justified if wife neglects the children	**Beating justified if wife neglects the children**
	Beating justified if wife argues with husband	**Beating justified if wife argues with husband**
	Beating justified if wife refuses to have sex with husband	**Beating justified if wife refuses to have sex with husband**
	Getting medical help for self: getting permission to go	
	Beating justified if wife burns the food	
	Getting medical help for self: having to take transport	
	Getting medical help for self: not wanting to go alone	**Getting medical help for self: not wanting to go alone**
	Frequency of reading newspaper or magazine	**Frequency of reading newspaper or magazine**
	Frequency of listening to radio	
	Frequency of watching television	**Frequency of watching television**

Table A3. Survey characteristics for selected countries

Wealth index combined	
ntl	0.590
popdens	0.567
temp	0.406
coast	0.315
health2	0.249
prec	0.248
airport	-0.194
waterw	-0.221
road	-0.344
ele	-0.398
edu	-0.427
health1	-0.524
finan	-0.598
ndvi	-0.668

Education in single years	
ntl	0.520
popdens	0.462
temp	0.195
prec	0.164
coast	0.111
health2	0.059
waterw	-0.144
airport	-0.175
ele	-0.177
road	-0.194
edu	-0.292
health1	-0.386
finan	-0.475
ndvi	-0.508

How often uses internet	
finan	0.496
ndvi	0.482
health1	0.428
edu	0.348
road	0.285
ele	0.266
airport	0.237
waterw	0.163
prec	-0.153
health2	-0.161
coast	-0.193
temp	-0.249
popdens	-0.403
ntl	-0.467

Covered by health insurance	
ntl	0.462
popdens	0.444
health2	0.080
temp	0.056
prec	0.037
ele	-0.018
coast	-0.050
waterw	-0.118
road	-0.145
airport	-0.219
edu	-0.230
health1	-0.363
ndvi	-0.420
finan	-0.435

Highest education level	
ntl	0.560
popdens	0.494
temp	0.198
prec	0.176
coast	0.121
health2	0.084
waterw	-0.156
ele	-0.183
road	-0.217
airport	-0.234
edu	-0.326
health1	-0.407
finan	-0.494
ndvi	-0.534

Has bank account	
ntl	0.565
popdens	0.527
temp	0.205
prec	0.159
health2	0.155
coast	0.125
waterw	-0.159
ele	-0.180
road	-0.203
airport	-0.256
edu	-0.325
health1	-0.430
finan	-0.494
ndvi	-0.530

Has telephone	
ntl	0.381
popdens	0.313
prec	0.086
temp	0.064
ele	-0.018
health2	-0.057
coast	-0.099
waterw	-0.128
road	-0.191
airport	-0.198
edu	-0.243
ndvi	-0.351
finan	-0.359
health1	-0.362

Medical help: distance	
ndvi	0.330
finan	0.288
health1	0.265
edu	0.181
road	0.156
ele	0.155
waterw	0.101
prec	-0.012
airport	-0.043
health2	-0.070
coast	-0.075
temp	-0.173
ntl	-0.247
popdens	-0.276

Currently working	
popdens	0.350
ntl	0.350
health2	0.169
temp	0.153
coast	0.085
prec	0.001
airport	-0.032
road	-0.086
waterw	-0.110
ele	-0.120
edu	-0.171
finan	-0.221
health1	-0.247
ndvi	-0.412

Main wall material	
ele	0.387
ndvi	0.315
finan	0.299
health1	0.284
road	0.247
edu	0.232
waterw	0.121
airport	-0.004
prec	-0.164
popdens	-0.166
ntl	-0.181
health2	-0.294
coast	-0.359
temp	-0.383

Table A4. Available DHS countries

#	Country	UNDP Region	Year	Type
1	Albania	Europe and Central Asia	2018	Standard DHS
2	Angola	Africa	2016	Standard DHS
3	Armenia	Europe and Central Asia	2016	Standard DHS
4	Bangladesh	Asia and the Pacific	2018	Standard DHS
5	Benin	Africa	2018	Standard DHS
6	Burkina Faso	Africa	2018	MIS
7	Burundi	Africa	2017	Standard DHS
8	Cambodia	Asia and the Pacific	2022	Standard DHS
9	Cameroon	Africa	2018	Standard DHS
10	Chad	Africa	2015	Standard DHS
11	Ethiopia	Africa	2016	Standard DHS
12	Ethiopia	Africa	2019	Interim DHS
13	Gambia	Africa	2020	Standard DHS
14	Ghana	Africa	2016	MIS
15	Ghana	Africa	2017	Special
16	Ghana	Africa	2019	MIS
17	Guatemala	Latin America and the Caribbean	2015	Standard DHS
18	Guinea	Africa	2018	Standard DHS
19	Guinea	Africa	2021	MIS
20	Haiti	Latin America and the Caribbean	2017	Standard DHS
21	India	Asia and the Pacific	2016	Standard DHS
22	India	Asia and the Pacific	2021	Standard DHS
23	Jordan	Arab States	2018	Standard DHS
24	Kenya	Africa	2015	MIS
25	Kenya	Africa	2020	MIS
26	Liberia	Africa	2016	MIS
27	Liberia	Africa	2020	Standard DHS
28	Madagascar	Africa	2016	MIS
29	Madagascar	Africa	2021	Standard DHS
30	Malawi	Africa	2016	Standard DHS
31	Malawi	Africa	2017	MIS
32	Mali	Africa	2015	MIS
33	Mali	Africa	2018	Standard DHS
34	Mali	Africa	2021	MIS
35	Mauritania	Africa	2021	Standard DHS
36	Mozambique	Africa	2015	Standard AIS
37	Mozambique	Africa	2018	MIS
38	Myanmar	Asia and the Pacific	2016	Standard DHS
39	Nepal	Asia and the Pacific	2016	Standard DHS
40	Niger	Africa	2021	MIS
41	Nigeria	Africa	2015	MIS
42	Nigeria	Africa	2018	Standard DHS

43	Nigeria	Africa	2021	MIS
44	Pakistan	Asia and the Pacific	2018	Standard DHS
45	Philippines	Asia and the Pacific	2017	Standard DHS
46	Rwanda	Africa	2015	Standard DHS
47	Rwanda	Africa	2020	Standard DHS
48	Senegal	Africa	2015	Continuous DHS
49	Senegal	Africa	2016	Continuous DHS
50	Senegal	Africa	2017	Continuous DHS
51	Senegal	Africa	2018	Continuous DHS
52	Senegal	Africa	2019	Continuous DHS
53	Senegal	Africa	2021	MIS
54	Sierra Leone	Africa	2016	MIS
55	Sierra Leone	Africa	2019	Standard DHS
56	South Africa	Africa	2016	Standard DHS
57	Tajikistan	Europe and Central Asia	2017	Standard DHS
58	Tanzania	Africa	2016	Standard DHS
59	Tanzania	Africa	2017	MIS
60	Timor-Leste	Asia and the Pacific	2016	Standard DHS
61	Togo	Africa	2017	MIS
62	Uganda	Africa	2015	MIS
63	Uganda	Africa	2016	Standard DHS
64	Uganda	Africa	2019	MIS
65	Zambia	Africa	2018	Standard DHS
66	Zimbabwe	Africa	2015	Standard DHS