An open risk index with learning indicators from OSM-tags, developed by machine learning and trained with the WorldRiskIndex

D. Feldmeyer [1], H. Sauter [1], J. Birkmann [1]

[1] IREUS, Institute of Spatial and Regional Planning, University of Stuttgart, Germany – Daniel.feldmeyer@ireus.uni-stuttgart.de

Commission IV, WG IV/4

KEY WORDS: WorldRiskIndex, Vulnerability, Risk Management, Machine Learning, OSM, R, QGIS, PostGIS

ABSTRACT:

While climate change is already a real issue in many parts of the world, it is even more threatening the well-being of future generations. The SDG 1.5 explicitly aims to reduce the vulnerability and exposure to climate related hazards by 2030. The World Risk Index (WRI) is one well-respected approach in profiling countries risk to natural hazard. To effectively monitor development and detect decision points on the climate resilience pathway, data of high resolution in space and time about the world's countries is of urgent importance. The World Risk Index will guide the supervised learning part resulting in an indicator set derived from OpenStreetMap (OSM) tags, establishing on one hand an open risk index and adding deep explanatory power to its components by a qualitative discussion of the OSM themes. The second part explores with unsupervised algorithms the inherent characteristic of country groups classified by the open risk index and deduces common patterns of socio-economic vulnerability. Hence, the inherent challenge of this work is to substitute existing static indicators with new dynamic indicators, not only substituting them but also painting a more detailed picture. Moreover, new data sources still questioned often by their reliability compared to World Bank or census data, and therefore its opportunities are neglected instead of critically exploring the potential. This unique combination is not done yet and bares huge potential moreover united with the open source geo community to contribute a little piece of the puzzle for achieving the SDG 1.5**.**

1. INTRODUCTION

Climate change is threatening human well-being in many parts of the world, as it often results in an interaction of natural hazards with the vulnerability of the society leading to disastrous impacts. While understanding the meteorological processes in the light of the unstationarity of climate change and its changing patterns, knowledge on the vulnerability as result of resilience and exposure becomes crucial in the process of adaptation (Birkmann 2006). Billions spent across Europe for structural protection measures have shown to result in higher monetary losses, partially explained with increased values exposed (Fuchs et al. 2017). Showing the need to for better informed decisions. Indicators are one way to assess, evaluate and monitor vulnerability and risk (Mach et al. 2016, Queiroz de Almeida et al. 2016). The WorldRiskIndex developed by Birkmann and Welle in cooperation with Alliance Development Works (see Birkmann et al. 2011) is one approach of operationalization of the risk formula (IPCC 2012) on country level for 171 countries. The index is based on 28 indicators from different global open data sources. The framework captures several facets of risk and the analysis allows to deduce profiles for countries showing problems in their coping capacity and susceptibility (Birkmann et al. 2011, Welle & Birkmann 2015 a/b, Birkmann & Welle 2016, Feldmeyer, Birkmann & Welle 2018). This understanding of risk and vulnerability helps making better informed decisions; by selecting the right adaptation measures and monitoring its progress. The analysis of the World Risk data from 2012 – 2017 shows that despite a slide decrease in risk, the continental region of Oceania falls behind. Moreover, a high persistence of vulnerability in many African countries and its slow improvement, suggests that third-world countries are likely not able to implement integrated risk management strategies fast enough in the face of climate change (Feldmeyer, Birkmann & Welle 2018). Many more approaches exist for assessing vulnerability showing the complexity of the phenomenon (Cutter et al. 2003, Birkmann 2006, Karagiorgos et al. 2016, Jamshed et al. 2017, Sorg et al. 2018). Despite the century of information, a lack of data is existing for measuring social capacities (Sorg et al. 2017). The social and crowd sourced OpenStreetMap database shows a huge potential for adding social aspects to the official governmental statistics with high spatial resolution. However, to the authors knowledge, a vulnerability index based on OSM has not yet been developed. Hence, this paper is exploring the potential of this unique geodatabase to assess and evaluate global vulnerability.

Consequently, the following objectives are targeted:

- Developing a methodology for data handling of the global OSM dataset and deducing country statistics.
- Developing a data driven machine learning methodology to model countries vulnerability to natural hazards.
- What are the main tags (key & value) and keys explaining vulnerability on a country scale?

2. METHODOLOGY

2.1 OSM data

For this study, we count the numbers of unique tags from OSM nodes per country and statistically compare the counts with vulnerability indicators used in the World Risk Index. Although the spatial coverage and information density of OSM data varies across the world in relation to the number and performance of the national and regional contributors, its big advantage remains in the accessibility of a standardized global dataset.

The conceptual data model of OSM is built by elements that can be either nodes, ways or relations. Nodes, as the simplest form, representing only a point defined by coordinate pairs, can be connected to ways, which represent polylines (e.g. streets) or polygons (e.g. boundaries) if the lines are closed (endpoint=start

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-W14-37-2019 | © Authors 2019. CC BY 4.0 License.

point). Relations are multi-purpose data structures that define logical relationships between two or more elements. All of the aforementioned elements can be attributed with tags, i.e. plain text key-value pairs (e.g. amenity=restaurant) which further describe the meaning or function of objects. Whilst keys are unique and often categorial (landuse=, amenity=,…), values can contain anything from individual names (e.g. name = “Universität Stuttgart”, phone=+49 45627 2983) to numbers.

2.2 Data preprocessing

The OpenStreetMap dataset was downloaded as full planet file from https://planet.openstreetmap.org/ (date: 2019-04-25 00:58 UTC18:17). Due to the amount of information stored in the full OSM database (keys: 74.531, tags: 102.003.489; taginfo.openstreetmap.org - 2019-05-24 00:58 UTC, Data © OSM contributors (ODbL)), data reduction to a meaningful number of keys and values was crucial for this study. For this purpose the planetfile was processed with osmctools, a collection of command line programs and then imported into a PostgreSQL/POSTGIS database before the statistical analyses could be conducted. As a part of this preprocessing steps the osmfile was split into 197 single country files based on Natural Earth country polygons (NE_souvereignty – link). The aggregation to a summary table with tag counts per country was conducted in the PostgreSQL database. Figure 1 depicts the full preprocessing workflow.

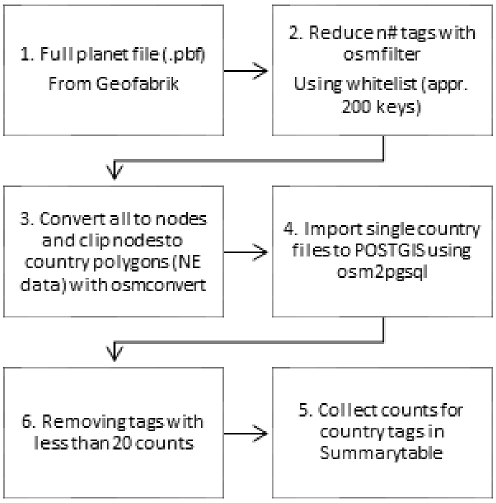

1. Full planet file (.pbf)
From Geofabrik
2. Reduce n# tags with osmfilter
Using whitelist (appr. 200 keys)
3. Convert all to nodes and clip nodes to country polygons (NE data) with osmconvert
4. Import single country files to POSTGIS using osm2pgsql
6. Removing tags with less than 20 counts
5. Collect counts for country tags in Summarytable

Figure 1: Data preprocessing workflow

In order to make as less assumptions in the first steps “all” remaining tags were treated as possible indicators to ensure the best possible exploration of the wealth of information by the OSM database and to detect potential new relations. Resulting in indicators of 190 by 250 nodes of tag counts.

The statistic were done in RStudio with R and several extensions (RStudio Team 2018, R Core Team 2019). Connecting to the database with the RPostgres package (Wickham, Ooms & Müller 2018). To reduce the high dimensionality of the data, three steps were taken: Indicators with only low coverage of the world were removed. The threshold was set to >50% coverage. Indicators with zero and near zero variance were also excluded. At last, the pair-wise correlation was calculated and in case of > 0.7 removed (Kuhn et al 2017).

2.3 Statistical Learning

In general, statistical learning is divided into supervised learning and unsupervised learning. Supervised in the sense that the response is known, and we want to predict the response or understand the relationship of the predictors (inference) (James et al. 2013).

2.3.1 Supervised learning: Roughly speaking two categories of supervised learning methods exist, linear and non-linear methods. Linear models often are more robust and also a higher predictive accuracy than non-linear models like e.g. trees. But, linear models assume a linear relation and additivity. The Generalized Additive Models (GAM) are somehow in between. Allowing non-linearity but still assuming additivity. For exploring the data in face of unknown relation the “Lasso Regression” for linear and “Random Forest” for moving beyond linearity were selected.

2.3.2 Lasso Regression: Lasso regression is similar to ridge regression. Whereas in ridge regression all predictors are kept in the model and a shrinkage parameter determines their influence but never is Zero. Lasso regression uses a l_1 penalty forcing parameters to zero when the tuning parameter is large enough.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;= \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|. \tag{1}$$

James et al. 2013

The “glmnet” package by Friedman, Hastie & Tibshirani (2010) is used to model the lasso regression. The data is split into training and test data, by randomly selecting 50% of the rows as training data. The best lambda is chosen by 10-fold cross validation.

2.3.3 Trees: Decision trees have some advantages and disadvantages over linear regression. First of all, decision trees are non-linear classifier in contrast to linear regression. Trees are understood very well also by non-statisticians as the graphical display is easy to understand. Simple trees can be sensitive to even small changes in data and have lower predictive accuracy. To overcome those shortcoming bagging, random forest and boosting of trees was developed and can substantially increase their performance (James et al. 2013).

As previous the data is split into training and test data. In a first exploratory step to get also a baseline and interpret the result of random forest a regression tree is calculated with the “tree” package (Ripley 2014). Secondly, for improving the result the regression tree is pruned. A very large tree is very likely to over fit the data and consequently a smaller tree could be favourable in means of variance at the cost of bias. One method is to grow a large tree and then to prune it back with the weakest link method.

Random Forest takes the advantages of the tree method and improves the weakness by building a forest. Each time when building a tree a bootstrapped training sample is taken. At each split also only a random sample of predictors is selected. This ensures that not one very strong predictor dominates all trees. The randomForest() function by Liaw and Wiener (2002) is on implementation of the method.

2.4 Unsupervised Learning

Unsupervised learning summarizes method which do not predict a specific output, like vulnerability. They are method to reduce dimensions and uncover relations of the data. The question set up in this paper is, which factors of OSM are related to high vulnerable countries.

2.4.1 Principal Component Analysis: Principal Component Analysis (PCA) is beside Cluster Analysis one of the most common unsupervised methods, although can be also applied in a supervised way for e.g. regression.

Important for PCA is the data pre-processing as the scale influences the components and hence the data need to be normalized. The caret package provides the preProcess() function including the method "pca". The prcomb() function within the kazaam package is applied for the PCA (Schmidt et al 2017).

2.5 World Risk Index

The World Risk Index by Birkmann is one operationalization of the risk formula (Figure 2). On the left side the exposure component is built by five indicators describing human exposure to natural hazards (earthquake, cyclones, floods, droughts, sea level rise). It is the physical exposure of people per country on an annual average basis, reflecting the probability of the event (Welle & Birkmann 2015).

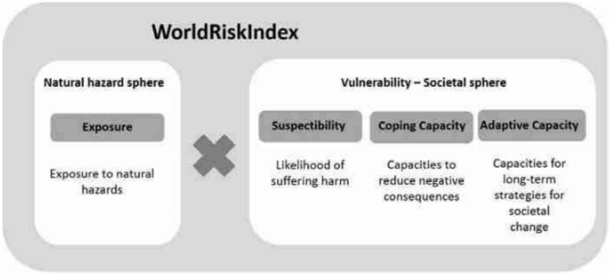
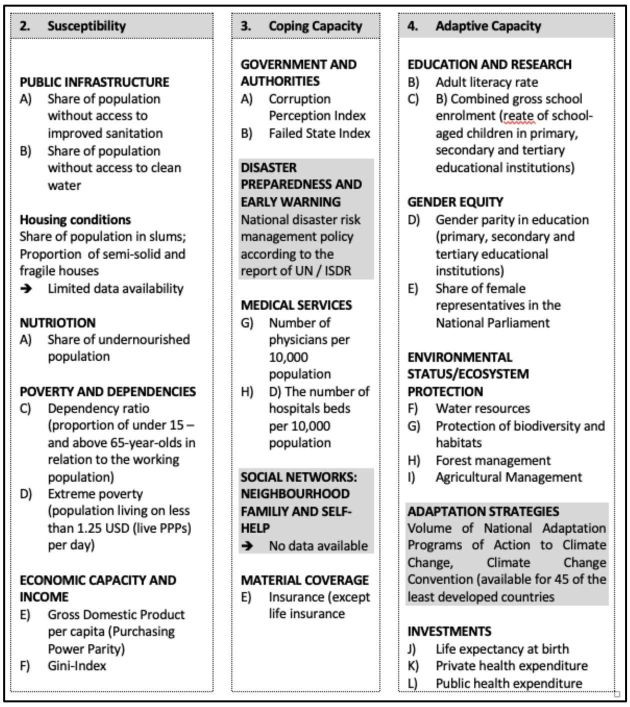

WorldRiskIndex
Natural hazard sphere
Exposure
Exposure to natural hazards
Vulnerability – Societal sphere
Suspectibility
Likelihood of suffering harm
Coping Capacity
Capacities to reduce negative consequences
Adaptive Capacity
Capacities for long-term strategies for societal change

Figure 2. Risk formula of the World Risk Index (Welle & Birkmann 2015b)

The socio-economic vulnerability (Table 1) consists of three pillars. First, *Susceptibility,* which describes the level to which people are affected by a hazard. Key components are *Public Infrastructure, Housing conditions, Nutrition, Poverty and dependencies* and the *Economic capacity and income.* Second, *Coping capacity,* how is the immediate ability to reduce the effects by the hazard. This strongly relies on *Government and authorities, Disaster preparedness and early warning, Medical services, Social networks* and *Material Coverage.* Third, *Adaptive capacity,* which resources has the society to plan and anticipate and reduce in the longer run negative impacts. Crucial within here are, *Education, Gender equity, Environment, Adaptive strategies* and investments.

The parts of the table with light grey underground are marked as parts which are not covered appropriate by official global data sources. The assumption made within this paper is that many of those components are being reflected in the OSM database in a direct and or indirect way. One method would have been to construct logically related indicators by thematic keys and tags.

Somehow rebuilt the concept of the World Risk Index indicator by indicator. This approach is limited in three ways. Firstly, the somehow chaotic contributions of slightly diverting spellings or reorder of key as a value or the other way around. Secondly, assumptions and judgements on the relations and their impact on vulnerability would introduce a strong bias. Thirdly, and most important the exploration of unknown and hidden patterns of the OSM spatial data is not obvious. Through the big-data characteristic of OSM data the uncovering needs advanced machine learning tools to extract unknown relationships and aspects of vulnerability. Therefore the combined vulnerability index was taken as the response variable and OSM-derived indicators as predictors.



2. Susceptibility	3. Coping Capacity	4. Adaptive Capacity
PUBLIC INFRASTRUCTURE A) Share of population without access to improved sanitation B) Share of population without access to clean water **Housing conditions** Share of population in slums; Proportion of semi-solid and fragile houses ➔ Limited data availability **NUTRIOTION** A) Share of undernourished population **POVERTY AND DEPENDENCIES** C) Dependency ratio (proportion of under 15 – and above 65-year-olds in relation to the working population) D) Extreme poverty (population living on less than 1.25 USD (live PPPs) per day) **ECONOMIC CAPACITY AND INCOME** E) Gross Domestic Product per capita (Purchasing Power Parity) F) Gini-Index	**GOVERNMENT AND AUTHORITIES** A) Corruption Perception Index B) Failed State Index **DISASTER PREPAREDNESS AND EARLY WARNING** National disaster risk management policy according to the report of UN / ISDR **MEDICAL SERVICES** G) Number of physicians per 10,000 population H) D) The number of hospitals beds per 10,000 population **SOCIAL NETWORKS: NEIGHBOURHOOD FAMILIY AND SELF-HELP** ➔ No data available **MATERIAL COVERAGE** E) Insurance (except life insurance	**EDUCATION AND RESEARCH** B) Adult literacy rate C) B) Combined gross school enrolment (reate of school-aged children in primary, secondary and tertiary educational institutions) **GENDER EQUITY** D) Gender parity in education (primary, secondary and tertiary educational institutions) E) Share of female representatives in the National Parliament **ENVIRONMENTAL STATUS/ECOSYSTEM PROTECTION** F) Water resources G) Protection of biodiversity and habitats H) Forest management I) Agricultural Management **ADAPTATION STRATEGIES** Volume of National Adaptation Programs of Action to Climate Change, Climate Change Convention (available for 45 of the least developed countries **INVESTMENTS** J) Life expectancy at birth K) Private health expenditure L) Public health expenditure

Table 1. Concept of the World Risk Index (Welle & Birkmann 2015b)

3. RESULTS

The results are described in the logic given by the methodology. Starting with the dimensionality reduction, followed by the linear model of the lasso regression, moving beyond this linearity and concluding by detection of unsupervised patterns via PCA within the data.

3.1 Dimension reduction

Two datasets were derived, and results obtained based on them. First data set referred to as *World Countries Tag Counts (Tags)* with each country one row and 1341 columns representing each tag a column and the count of the tag per country as the value. As can be seen in the example of the 10 country and 3 columns shown in. Data set two is referred to as *World Country Key Counts (Keys)* the columns are merged per key and the counts of the value summed up for the count of the key. Which results in a much smaller dataset regarding columns of 110 but higher completeness.

After running the dimension reduction on the *Tags,* 99 tags as indicators remind. For the *Keys,* 27 variables remain and for Europe surprisingly only 35 variables are left compared to the 99 of the world. Finally, the counts per country were normalized by its population and min max normalization, so that all values are between 0 and 1.

3.2 Lasso regression

The Lasso models linear combinations of the predictors on the response, in our case vulnerability. The lasso penalty “ell 1” shrinks the coefficients of some predictors exactly to zero reducing the predictors set and making the interpretation easier. The best Lambda is chosen by ten-fold cross validation of the training data.

The Mean-Squared-Error for the lambda selection cross validation process with the *Keys* data is forced to zero. As more and more predictors are forced to zero the error is reduced. Hence, it suggests that the relation of tag counts and vulnerability is not linear and a linear model therefore not appropriate for capturing it.

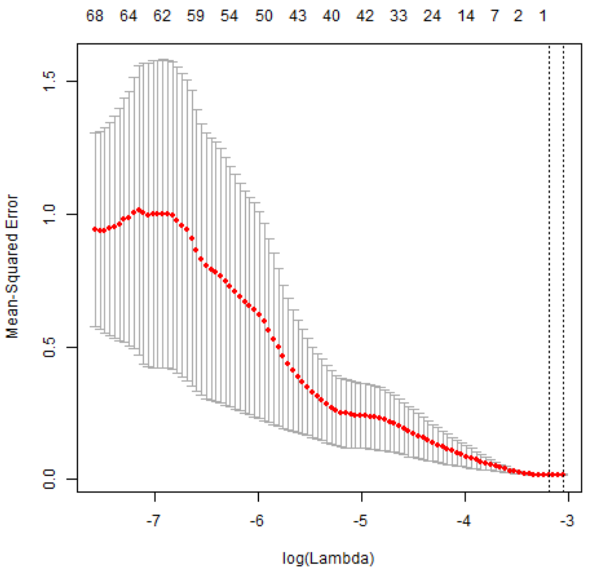


Figure 3. Lasso regression Lambda by crossvalidation *Tags*

Figure 3 shows the same Lambda selection for the *Tags* data. The Lambda is not pushed all to zero but nevertheless it suggests the pattern of the *Keys*, as it’s the same data just split into the counts of key value pairs. In order to allow for non-linearity between counts per country and vulnerability a tree approach is chosen.

3.3 Trees

For all following trees as well as random forest, the columns of the two datasets served as predictors vulnerability of the World Risk Index as response variable.

3.3.1 Modelling vulnerability based on Key count per country: Three steps are run for each of the three datasets. First a simple regression tree is built. Second, this tree is pruned for better robustness and less variance. Third, a random forest is grown allowing multiple trees.

The performance of the trees on the *Keys* data. In a first simple step a regression tree was calculated (Figure 4). The top node divides the tree by *traffic_signals* into left and right. The right branch is further divided by the electric frequency of railways, buses, electric power supply networks and communication devices. Further distinguished with *landuse, traffic_calming* and *operator.type.* Where *land use* and *traffic_calming* are more or less self-explaining, the top linked values to the key *operator:type* are *private, public, government* and *religious.* On the left side again *operator:type* and *traffic_calming* appear.

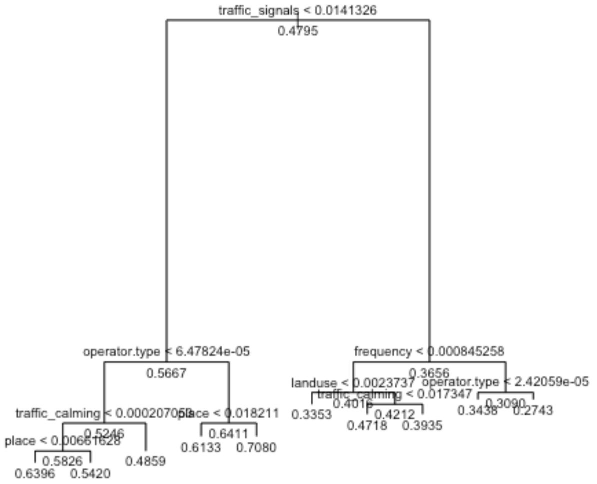


Figure 4. Regression Tree *Keys*

Pruning the regression tree to increase robustness and balance for the variance-bias trade-off is shown in Figure 5. This shows also the most relevant attributes for describing vulnerability with counts per key and country, which are *traffic_signals, operator_type, traffic calming* and *frequency.*

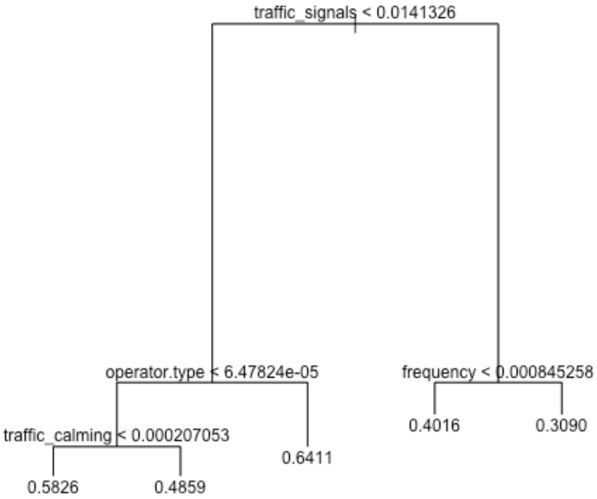

traffic_signals < 0.0141326
operator.type < 6.47824e-05
traffic_calming < 0.000207053
0.5826
0.4859
0.6411
frequency < 0.000845258
0.4016
0.3090

Figure 5. Pruned tree *Keys*

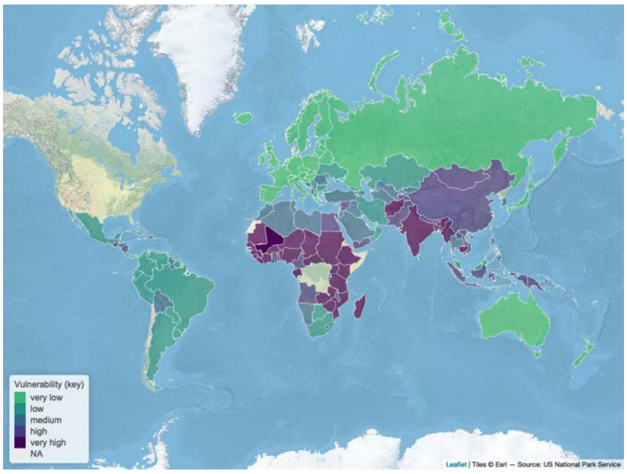
The 13 most important predictors for the random forest are shown in Table 2. The results come not as a surprise after the regression tree and the pruned tree. Nevertheless, the table gives a more detailed picture and in depth interpretation. The *X.IncMSE* is the increase in the MSE of the prediction, estimated by the out of bag cross validation. Again *traffic_signals* is the most important. Followed by *traffic_calming, operator.type* and *landuse. Frequency* appears slightly lower compared to the pruned tree.

Interesting is the last key with *fire_hydrant.dyameter* in the light of natural hazard and vulnerability. Such data are not in official country records.

	X.IncMSE	*IncNodePurity*
traffic_signals	20,0395575	0,430109572
traffic_calming	12,887251	0,265499982
operator.type	10,4132019	0,064276153
landuse	8,45125305	0,102887023
service	7,34750772	0,126863252
designation	7,09257133	0,084712005
construction	6,09648063	0,062480717
frequency	5,80008566	0,035168397
vehicle	5,0673767	0,04203698
toll	4,31681353	0,011428788
smoothness	4,20197549	0,020569056
cables	3,93164146	0,009381577
fire_hydrant.diameter	3,57858402	0,010674378

Table 2: Random forest important keys

The statistical model is now used to predict vulnerability values of the world (Figure 6). As in the World Risk Index the quantile method is used for categorizing the countries. This of cource includes now test and training data. For the prediction accuracy error on the test data needs to be considered. Another additional model check is now the correct classified countries based on the model vulnerability scores classified by the quantile method. For the 169 countries 66% are ranked according to their World Risk Index classification.



Leaflet | Tiles © Esri — Source: US National Park Service

Figure 6. Map random forest model based on keys

3.3.2 Modelling vulnerability based on tag count per country:

The database for the previous results were counts per *key* and country, which is a simple dimension reduction, additional to statistical dimension reduction. The following section now explores counts per tag as predictors of vulnerability.

In Figure 7 the regression tree of the *Tags* is split by *shop_yes.* The right brunch is further distinguished by *shelter_type_public_transport.* Further tags on this side are *highway_rest_area, cuisine_mexican,shop_lottery* and *natural_volcano*. The left side is contructed by the tags *amenity_swimming_pool* and *highway_residential.*

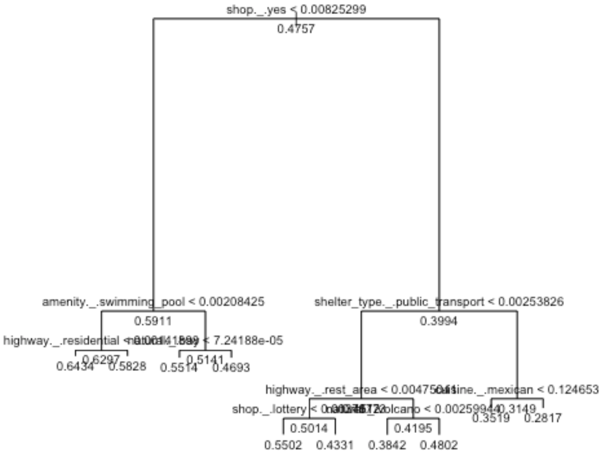


Figure 7. Regression Tree *Tags*

Pruning the tree as previous results in a simpler model (Figure 8). Hence the dominant tags for vulnerability are: *shop_yes, amenity_swimming_pool, shelter_type_public_transport* and *highway_rest_area.*

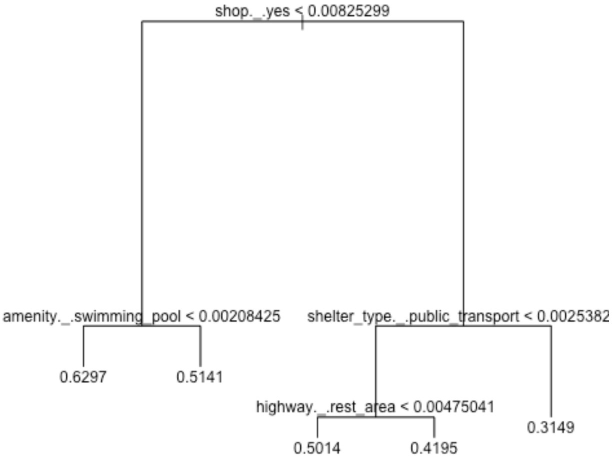

shop._.yes < 0.00825299
amenity._.swimming_pool < 0.00208425
shelter_type._.public_transport < 0.0025382
0.6297
0.5141
highway._.rest_area < 0.00475041
0.3149
0.5014
0.4195

Figure 8. Regression Tree *Tags*

The random forest model goes hand in hand with the first two and fourth tag. Third tag although is *amenity_shower.* The tag *swimming_pool* appears much lower (Table 3).

	IncMSE	*IncNodePurity*
shop._.yes	15,20	0,24
leisure._.swimming_pool	10,09	0,16
amenity._.shower	9,07	0,15
shelter_type._.public_transport	8,58	0,09
sport._.basketball	7,95	0,08
man_made._.survey_point	7,57	0,06
cuisine._.mexican	6,48	0,07
cuisine._.thai	6,46	0,02
barrier._.stile	6,38	0,02
landuse._.cemetery	6,30	0,05

amenity._.bureau_de_change	5,94	0,03
amenity._.swimming_pool	5,83	0,02
natural._.bay	5,79	0,04

Table 3: Random forest important tags

Based on the random forest model the world vulnerabilities are calculated and mapped in Figure 9

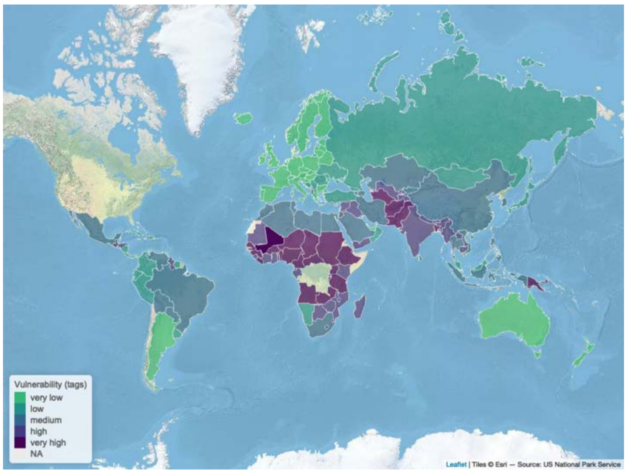


Figure 9. Map random forest model based on tags

Compared to the World Risk Index 65% of the countries are in the same category.

3.4 Principal Component Analysis (PCA)

Unsupervised learning explores underlying structures and combines sub-sets into new reduced dimensions. The 20 most vulnerable countries are selected and explored with the unreduced key counts and tag counts, 110 indicators respectively 1341 indicators. Afterwards zero and near zero variance predictors removed.

3.4.1 PCA *Keys*: The PCA is in the caret package part of the pre-processing function (Kuhn et al 2017). The function is set to explain 70% of the variance within the data.

The PCA for the keys resulted in seven components (Table 4). PC1 is highest loaded by *smoking. sport* and *parking.* PC2 is loaded by *bridge, park_ride* and *smoothness.* PC3 is loaded by *junction, supervised* and *direction.* The two highest predictors for PC4 are *shelter* and *shelter_type.* For PC5 the 2 highest are *railway* and *aeroway.* PC6 is dominated by *genus, hiking* and *cuisine.* Last, PC7 the two main predictors are *horse* and *motorcycle.*

	Predictors	*Name*
PC1	*Smoking(no). sport* and *parking*	Health
PC2	*bridge, park_ride* and *smoothness*	Infrastructure Quality
PC3	*junction, supervised* and *direction*	Traffic Control
PC4	*shelter* and *shelter_type*	Shelter
PC5	*railway* and *aeroway*	Air & Rail
PC6	*genus, hiking* and *cuisine*	Leisure
PC7	*horse* and *motorcycle*	Fun

Table 4 Principal Components of keys

3.4.2 PCA *Tags:* Removing from the tags the zero and near zero variance predictors 759 predictors remained. The loading of the single indicator is less strong than previous (Table 5). So more predictors contributing to each principle component. PC1*: shop _ beverages, amenity _ arts_centre, leisure _ pitch, shop _ hardware, shop _ doityourself, amenity _ kindergarten amenity _ social_centre, information _ board, sport _ soccer highway _ emergency_access_point* – which could be summarized as *Social, Art & Shopping.* PC2, is influenced quite low equally by 130 predictors of *Tourism & Sport.* PC3 is hard to interpret, but best summarized as *Tourist Infrastructure.* PC4 is *Man-Made.* PC5 is *Cuisine.* PC6 summed as *Health & Leisure.* PC7 is again loaded by man *Historic & Natural.*

	Name
PC1	Social, Art & Shopping
PC2	Tourism & Sport
PC3	Tourist Infrastructure
PC4	Man-Made
PC5	Cuisine
PC6	Health & Leisure
PC7	Historic & Natural

Table 5 Principal Components of tags

3.5 Moving from vulnerability to risk

Modelling the exposure component of the World Risk Index only 35% of the countries were correctly classified. It was expected that the probability of affected people is not dependent on structural elements. Hurricanes, floods, drought and sea level rise are not map elements. But shouldn´t they be?

3.5.1 Mapping Open Risk and World Risk: To Map the open risk index, the exposure from the World Risk is taken and combined by multiplying it with the key-based random forest model representing vulnerability.

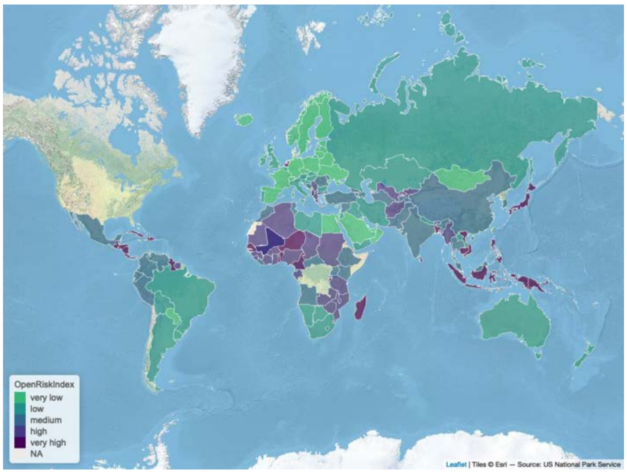


Figure 10. Open Risk Index

Figure 10 the world risk to natural hazard. The classification is adopted from the World Risk Index by the quantile method.

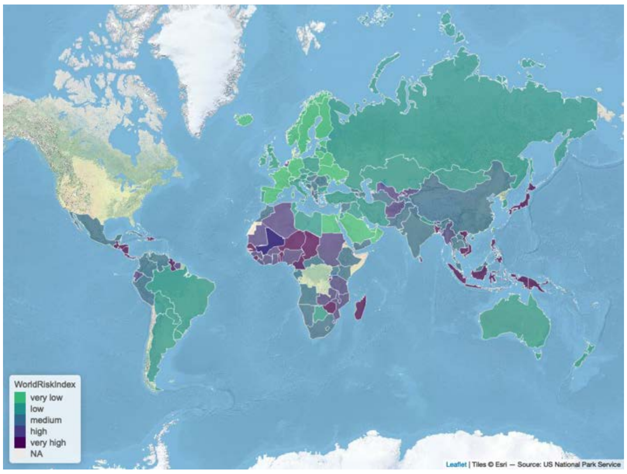


Figure 11. Vulnerability of the Open Risk Index

Figure 11 in comparison is the original World Risk Index. Obviously, the main patterns are coherent. Most of the differences are within the range of on class. Globally there is a persistent high risk in Oceania, Africa and Central America.

4. DISCUSSION

The approach provided meaningful insights into three questions: A) global country data deduced from OSM B) statistical learning for predicting vulnerability C) global vulnerability.

Firstly, the tag count per country seemed mostly influenced by the OSM coverage of this country. Population or absolute count could not remove this effect entirely. Additional spellings and translations could not be assessed during this study. These limitations need further consideration for global analysis. However, global datasets are sparse and the global coverage of the OSM data, combined with its availability represent a huge potential for risk management evaluation and global climate change adaptation studies.

Secondly, in this study the vulnerability index was considered as truth. However, uncertainties remain and further research on the propagation of these unknowns could clarify interpretations. As such, negative correlated indicators potentially exist and need to be further pointed out to improve the explanatory power of the indicators and by this of the risk index. Thus, there might remain hidden or obscured effects through the characteristic of the dataset and the current approach.

Thirdly, hazard could not be modelled. Therefore, the question remains if the information extracted from OSM does not yet include information which explicitly or implicitly is related to exposure or the methodology applied was not able to uncover those structures.

Fourthly, the WorldRiskIndex is on a continuous scale from zero to one. Another way instead could be also to model not the vulnerability index value but to classify instead or model its components separately and then also index them.

Fourthly, the preliminary results suggest that on a country scale its vulnerability can be approximated based OSM map elements and match key elements of the World Risk Index vulnerability (Table 1). Several infrastructure related tags (*traffic_signals, traffic_calming*, *toll*, *smoothness*) could be seen as approximation for the development of the country (ECONOMIC CAPACITY AND INCOME) but also the ability of the government to provide public goods (GOVERNMENT AND AUTHORITES, PUBLIC INFRASTRUCTURE). POVERTY AND DEPENDENCIES, MEDICAL SERVICES and HEALTH, are predicted by leisure, health and sport keys and tags *(cuisine, leisure_ swimming pool, sport_basketball*). Disaster preparedness not covered in the World Risk Index due to lack of data might be approximated by *shelter_public* or *fire_hydrant.diameter*. Hence, not only vulnerability can be estimated but also the thematically important themes are covered by the prediction models.

Fifthly, the preliminary results of the PCA suggest that vulnerability can be explained by Health, Infrastructure Quality/Traffic Control and Shelter mainly. Argumentum e contrario considering the global effort in reducing vulnerability showing the importance of investments into health, infrastructure but also shelter for risk preparedness. More in general related to leisure and tourism the importance of a strong middle class.

Lastly, having said this, the focus of the study was to develop a methodology and find preliminary variables explaining vulnerability. In a second step beyond the scope of this paper now, the results need to be further tested.

5. CONCLUSION

The OpenRiskIndex is the first approach modelling socio-economic vulnerability to natural hazards with statistical learning based on the OSM database.

Using counts per country as predictors is a possibility to deduce meaningful information without a super computer from the global OSM dataset. The price paid is of course the spatial explicitness (resolution). Supervised and unsupervised statistical learning proved to be able to model country socio-economic characteristics of vulnerability. Most significant dimensions were regarding health, infrastructure and economic capacity for explaining vulnerability. Although questions remain and could not be addressed in this study, OSM seems to offer the possibility of effectively monitoring changes in global vulnerability. Compared to country statistics and global databases where indicators often include a time lack and moreover are based on different years, covering or delaying the detection of trends, the very much alive OSM might be able to provide better coverage of a changing world.

Socio-economic indicators can provide valuable insides. Hence robust OSM based indicators for the SDG´s, vulnerability and risk. Further research although is needed to develop a commonly accepted methodology and a global open source database with computational power to overcome the limitation of desktop pcs and fully explore the possibilities.

REFERENCES

Birkmann, J. 2006. Measuring vulnerability to natural hazards: towards disaster resilient societies. No. Sirsi) i9789280811353.

Birkmann, J., Welle, T., Krause, D., Wolfertz, J., Suarez, D., & Setiadi, N., 2011. WorldRiskIndex: Concept and Results. In WorldRiskReport, pp. 13-43. Berlin: Alliance Development Works.

Birkmann, J., Cardona, O.D., Carreño, M.L., Barbat A.H., Pelling, M., Schneiderbauer, S., Kienberger, S., Keiler, M., Alexander, D., Zeil, P. and Welle, T., 2013: Framing vulnerability, risk and societal responses: The MOVE framework. Natural Hazards 01/2013; 67(2):193-211.

Birkmann J, Welle T., 2016. The World Risk Index 2016: reveals the necessity for regional cooperation in vulnerability reduction. J Extreme Events. https://doi.org/10.1142/S2345737616500056

Fuchs S, Ro¨thlisberger V, Thaler T, Zischg A, Keiler M (2017) Natural hazard management from a coevolutionary perspective: exposure and policy response in the European Alps. Ann Am As Geogr 107 (2):382–392. https://doi.org/10.1080/24694452.2016.1235494

Cutter, S., Boruff, B., Shirley, W., 2003. Social Vulnerability to Environmental Hazards*. In: Social Science Q 84 (2), S. 242–261. DOI: 10.1111/1540-6237.8402002

Feldmeyer, D., Birkmann, J., Welle, T., 2017. Development of Human Vulnerability 2012–2017. In: Journal of Extreme Events 04 (04), S. 1850005. DOI: 10.1142/S2345737618500057.

IPCC [Intergovernmental Panel on Climate Change], 2012. Glossary of terms. In: Field CB, Barros V, Stocker TF, Qin D, Dokken DJ, Ebi KL, Mastrandrea MD, Mach KJ, Plattner G-K, Allen SK, Tignor M, Midgley PM (eds) Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of working groups I and II of the Intergovernmental Panel on Climate Change (IPCC). Cambridge University Press, Cambridge, pp 555–564

James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013. An introduction to statistical learning . Corrected at 8th printing 2017. New York: springer.

Jamshed, A., Rana, I.A., Mirza, U.M., Birkmann, J., 2019. Assessing relationship between vulnerability and capacity: An empirical study on rural flooding in Pakistan. Int. J. Disaster Risk Reduct., 36.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

Karagiorgos, K., Thaler, T., Heiser, M., Huebl, J., Fuchs, S., 2016. Integrated flash flood vulnerability assessment: insights from East Attica, Greece. J Hydrol 541(Part A):553–562. https://doi.org/10.1016/j.jhydrol. 2016.02.052

Kuhn, M., Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2017). caret: Classification and Regression Training. R package version 6.0-78. https://CRAN.R-project.org/package=caret

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2(3), 18--22.

Mach, K., Mastrandrea, M., Bilir, T., & Field, C., 2016. Understanding and responding to danger from climate change: the role of key risks in the IPCC AR5. Climatic Change, 136 (3-4), 427-444

Queiroz de Almeida, L., Welle, T. and Birkmann, J., 2016. Disaster risk indicators in Brazil: A proposal based on the world risk index. International Journal of Disaster Risk Reduction 17, 251-272

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ripley, B., 2014. Package “tree”. Retrieved from http://cran.r-project.org/web/packages/tree/index.html

RStudio Team, 2018. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Schmidt, D., Chen, W., Matheson, M., Ostrouchov, G., 2017. “kazaam: Tools for Tall Distributed Matrices.” R package version 0.1-0, <URL: https://cran.r-project.org/package=kazaam>.

Sorg, L., Medina, N., Feldmeyer, D., Sanchez, A., Vojinovic, Z., Birkmann, J., Marchese, A., 2018. Capturing the multifaceted phenomena of socioeconomic vulnerability. In: Nat Hazards 92 (1), S. 257–282. DOI: 10.1007/s11069-018-3207-1.

Taiyun, W., Viliam, S., 2017. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Welle, T. & Birkmann, J., 2015a. The WorldRiskIndex 2015. In WorldRiskReport 2015,pp. 41-49. Berlin: Alliance Development Works.

Welle, T. & Birkmann, J., 2015b. The World Risk Index – An Approach to Assess Risk and Vulnerability on a Global Scale. Journal of Extreme Events, 02(01), 34 pages. doi.org/10.1142/s2345737615500037

Wickham, H., Ooms, J., Müller, K., 2018. RPostgres: 'Rcpp' Interface to 'PostgreSQL'. R package version 1.1.1. https://CRAN.R-project.org/package=RPostgres