Mapping the potential geographic distribution of surface water drinking sources in Liberia

Weiyu Yu, Nicola A. Wardrop, Jim A. Wright

Introduction:

Achieving universal and equitable access to safe and affordable drinking water for all requires evidence-based assessments to identify the disadvantaged areas and prioritise those with the most needs accordingly. To facilitate drinking water infrastructure development to deliver safe and sustainable water services for all, it is necessary to locate the people still using disadvantaged water services across the country. Surface water at the bottom of WHO/UNICEF Joint Monitoring Programme (JMP)'s water ladder (WHO and UNICEF 2017) refers to drinking water directly from open sources such as a river, stream, lake, dam, pond, canal or irrigation channel. Fetching water from open sources may pick up contaminants and pathogens; without proper treatment before use, it may cause serious health effects. Although conventional geospatial datasets concerning drinking water services generally contain comparatively limited information on surface water sources, more newly released datasets combining machine learning predictive modelling methods makes it possible to predict the potential spatial distribution of specific types of disadvantaged water service such as surface water.

This study uses a novel machine learning algorithm named maximum entropy (MaxEnt) to predict the potential spatial distribution of surface water drinking sources in Liberia. MaxEnt method is based on the maximum entropy principle, which suggests making prediction of the unknown probability distribution by looking for the probability distribution of maximum entropy (i.e. which is most diffused and closest to uniform distribution where the probability for each individual locality within the area of interests tends to be equally likely) bounded by the constraints derived from the obtained presence data (the coordinates of geographic locations where the target objects are observed) and the known environmental conditions across the area of interests. It has been widely applied in biological and ecological studies. Detailed methodological introduction of MaxEnt method can be found in (Phillips *et al.* 2006).

Data:

- **Surface water**

In total, 59 water point data describing surface water sources derived from the Water Point Exchange are employed as the observed occurrence sample of surface water, which excluded duplicates and those located within a same 5km x 5km grid cell (to avoid introduce any duplicate to the model).

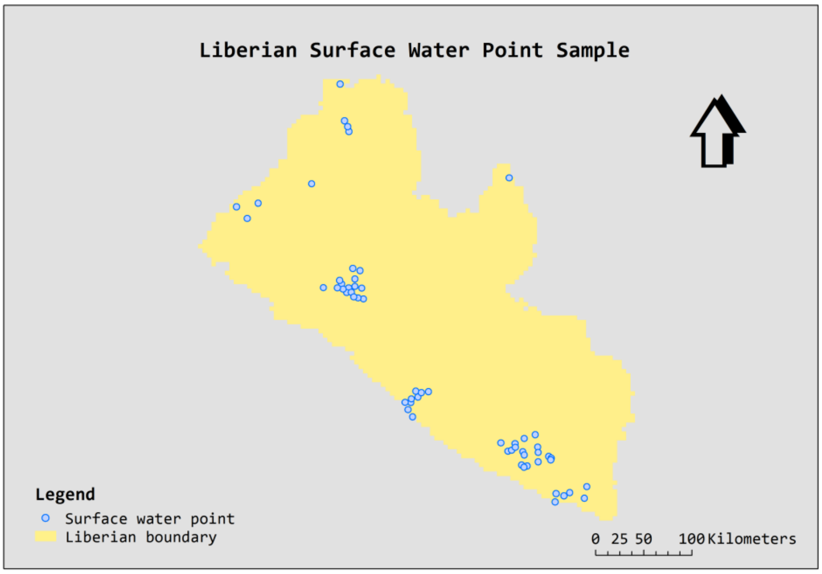

Figure 1 Liberian surface water point sample

- **Predictive covariates**

This study identified 10 predictive covariates (Table 1) that may be importance determinants of the spatial distribution of open source drinking water considering both the availability of surface water resources and socio-economic factors that may reflect the local demands and preferences. For example, Euclidean distance to inland water was calculate to reflect the availability of surface water resources; improved water source coverage was employed to indicate the potential demands on surface water; open defecation surface was used as a proxy of poverty indicator to reflect potential affordability of advanced water supply services. All predictive covariate layers were scaled down to 5km spatial resolution.

Table 1

Covariate	**Data Name**	**Data Source**	**Data Type**	**Format**	**Resolution**
Distance to inland water	Digital Chart of the World (DCW)	Environmental Systems Research Institute, Inc. (ESRI)	categorical vector	Shapefile	500 m
Elevation	ASTER GDEM Version 2	Ministry of Economy, Trade, and Industry (METI) of Japan and the United States National Aeronautics and Space Administration (NASA)	continuous raster	Geotiff	30 m
Slope	ASTER GDEM Version 2	Ministry of Economy, Trade, and Industry (METI) of Japan and the United States National Aeronautics and Space Administration (NASA)	continuous raster	Geotiff	30 m
Annual rainfall	WorldClim Clobal Climate Data version 1.4	Museum of Vertebrate Zoology, University of California (Hijmans *et al.* 2005)	continuous raster	Geotiff	1 km
Depth to groundwater	Equilibrium Water Table Africa Model version 2	Fan et al. (2013)	continuous raster	NetCDF	1 km
Distance to villages	Open Street Map (OSM)	OpenStreetMap Foundation (OSMF) & Contributors (OpenStreetMap Foundation (OSMF) & Contributors 2017)	categorical vector	Shapefile	-
Distance to roads	Open Street Map (OSM)	OpenStreetMap Foundation (OSMF) & Contributors (OpenStreetMap Foundation (OSMF) & Contributors 2017)	categorical vector	Shapefile	-
% population with access to improved water source	DHS Modelled Surfaces	Gething et al. (2015)	continuous raster	Geotiff	5 km
% population with no toilet	DHS Modelled Surfaces	Gething et al. (2015)	continuous raster	Geotiff	5 km
Land cover	MODIS Land Cover Type (MCD12Q1) version 5.1	University of Maryland & the Pacific Northwest National Laboratory (Collins and Emanuel 2014)	Land cover, categorical raster	Geotiff	500 m

We clipped all layers to the same spatial extent, and removed large water bodies and sea to ensure our analysis considered terrestrial areas only. Pre-processed covariate layers alongside other data used in this case study can be downloaded from https://geoterry.github.io/GEOWAT-SDGinsights/downloads.html.

Sampling bias:

Kernel density surface was calculated based on obtained surface water point sample to be used as bias files in order to handle the potential sampling bias. A pixel with higher density value indicates that it received a greater survey effort. Such bias file could reflect variation in survey effort and MaxEnt therefore uses it as a weighting layer to ensure that the target water points are observed in locations with particular covariate conditions is due to such conditions are favourable, rather than due to these locations received greater survey efforts.

Model building:

For model building, 70% of the surface water presence points were randomly selected to train the model, whilst the remainder were set aside for testing the model performance. We generated 1,000 background points by randomly selecting points within the full spatial extent defined in Liberia (where large water bodies were excluded). We repeated the sampling of training and background points 50 times and then computed the aggregated prediction and performance analysis. Evaluation of model performance was carried out using Area Under the Receiver Operator Curve (AUC; DeLong *et al.* 1988). An AUC value of 1 reflects perfect discriminatory power of the model; 0.5 indicates that the prediction failed to capture any patterns and is no better than a random distribution; AUC above 0.75 indicates a potentially useful discrimination of the model (Elith 2000, Phillips and Dudík 2008). The MaxEnt model building was carried out using R with the MaxEnt package. Alternatively, the small size open source software package named 'MaxEnt' developed by Steven J. Phillips and colleagues for ecological niche modelling can be used to build the model directly.

Results:

The following maps in Figure 2 show the 5km resolution predicted potential spatial distribution of surface water sources in Liberia. This is merely a simplified model for the illustration of this idea. A comparatively precise model can be conducted at finer resolution with sufficient geospatial data and a systematic conceptual framework identifying technical and socio-economic factors that may affect the distribution of specific water sources. The output surface can be interpreted as relative probability of the presence of surface water drinking sources. Such prediction should not directly replace national scale water point inventory or nationally representative household surveys. However, it could give a brief indication of the likely spatial distribution of specific type of water sources in areas where data is lacking.

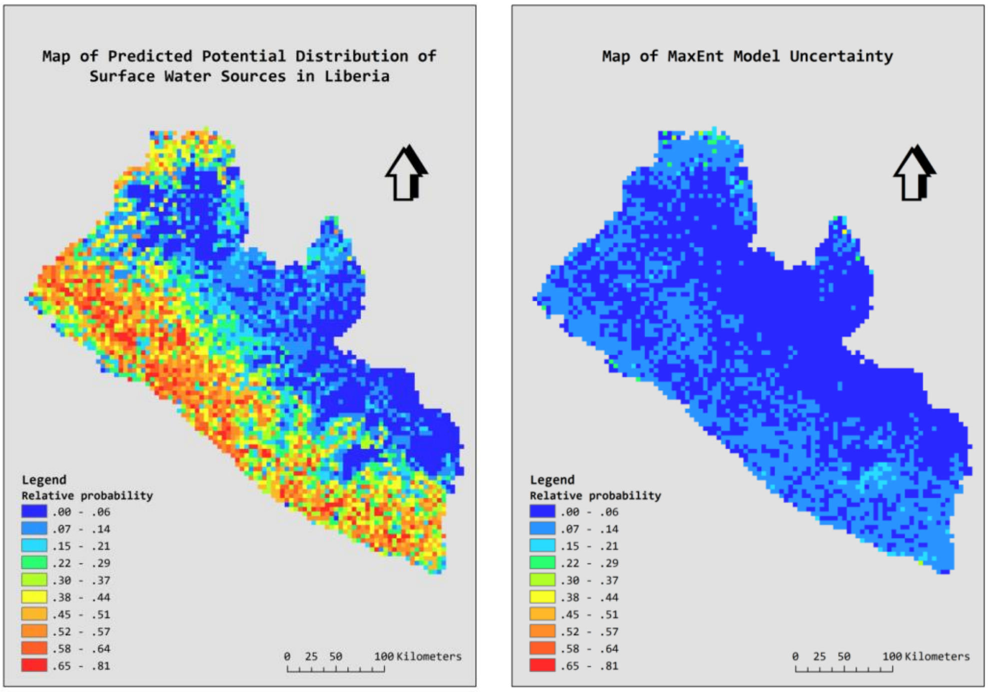

Figure 2 Output of the MaxEnt prediction

References:

COLLINS, C.S.K. and W.R. EMANUEL, 2014. Global mosaics of the standard MODIS land cover type data. *University of Maryland and the Pacific Northwest National Laboratory, College Park, Maryland, USA.*

DELONG, E.R., D.M. DELONG and D.L. CLARKE-PEARSON, 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**(3), 837–845

DUBITZKY, W., M. GRANZOW and D.P. BERRAR, 2007. *Fundamentals of Data Mining in Genomics and Proteomics*. 1st ed. Springer US

ELITH, J., 2000. Quantitative Methods for Modeling Species Habitat: Comparative Performance and an Application to Australian Plants. *In*: S. FERSON and M. BURGMAN, eds. *Quantitative Methods for Conservation Biology*. Springer New York, pp. 39–58

FAN, Y., H. LI and G. MIGUEZ-MACHO, 2013. Global Patterns of Groundwater Table Depth. *Science*, **339**, 940–943

GETHING, P. *et al.*, 2015. *Creating Spatial Interpolation Surfaces with DHS Data DHS Spatial Analysis Reports No. 11* [online]. Rockville, Maryland, USA Available from: http://dhsprogram.com/publications/publication-SAR11-Spatial-Analysis-Reports.cfm

HIJMANS, R.J. *et al.*, 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**(15), 1965–1978

OPENSTREETMAP FOUNDATION (OSMF) & CONTRIBUTORS, 2017. *Open Street Map* [online] [viewed 1 Jan 2017]. Available from: https://www.openstreetmap.org/; http://www.geofabrik.de/geofabrik/openstreetmap.html

PHILLIPS, S.J., R.P. ANDERSON and R.E. SCHAPIRE, 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259

PHILLIPS, S.J. and M. DUDÍK, 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**(2), 161–175

WHO and UNICEF, 2017. *The New JMP Ladder for Drinking Water* [online] [viewed 24 Jul 2017]. Available from: https://washdata.org/monitoring/drinking-water