A Completer Picture of Domestic Water Access and Consumption

Integrating machine learning models and survey information

Jan Geleijnse

TU Delft

African Water Corridor

A COMPLETER PICTURE OF DOMESTIC WATER ACCESS AND CONSUMPTION

INTEGRATING MACHINE LEARNING MODELS AND SURVEY INFORMATION

MSc. Thesis

Jan GELEIJNSE
Student number: 4376498

Committee:
Dr. Edo Abraham | Assitant Professor Water Management | Chair
Dr. Martine Rutten | Associate Professor Water Management
Dr. Doris van Halem | Associate Professor Drinking Water Treatment
Didier de Villiers | Programme Coordinator of the African Water Corridor

March 4, 2022

Picture on front page by James Tayebwa, permission to use granted.

ABSTRACT

The United Nations (UN) Sustainable Development Goal (SDG) #6 reads that by 2030 universal and equitable access to safe and affordable drinking water is achieved for all (United Nations General Assembly, 2015). In order to achieve this goal, proper and complete monitoring, capturing all the facets of safe water access, is essential. In this thesis it is argued that the current monitoring techniques come with limitations and subsequently solutions are presented to come to a more complete picture of water access.

Monitoring safe water access happens primarily through household health surveys. These surveys are often incomplete, not covering entire nations, focus on only the primary water source and are often spatially aggregated for privacy reasons. Besides, health surveys almost never include questions on consumed water volumes while that is an important indicator for proper hygiene (WELL, 1998), and something that, at the same time, should be in balance with the natural available water resources. Next to this survey based monitoring, there is the Water Point Data Exchange (WPDx) that monitors safe access by providing a platform at which the exact location and type of water access points (such as boreholes, springs, etc.) are registered. This does give more insight into the presence and usage of a variety of sources, but also the WPDx is often incomplete: not covering entire nations.

In this thesis we present a dual methodology that gap-fills the incompleteness of the WPDx database through modeling and in parallel, researches the complex local dynamics of water access, the variety of water sources used by households and the relationships between access and water consumption by means of a household survey.

By improving a machine learning biological species modeling technique (called MaxEnt), successful predictions on the number of presences of eight different water access types across Uganda were made, also into areas that have little presence in the WPDx data. It was found that population density, precipitation, elevation, poverty and groundwater storage are important indicators for the (non)presence of water access points.

Next to modeling, a survey campaign was executed in Bushenyi-Ishaka municipality, a mid-sized town in the South West of Uganda comprising a mixture of both urban and rural areas. This was done in collaboration with Makerere University (Kampala). The survey results showed that water consumption increases with education and wealth, but also with higher number of water point presences predicted by the model. It was also found that households in Bushenyi make use of an average of two different water sources on a regular basis and often express preference for sources off premises compared to on premises (piped) for both cost and perceived quality reasons.

Lastly, modi operandi were suggested for the results to improve water access such as prioritising areas with poor(est) water access and investing in rainwater harvesting, infrastructure and education.

Contents

List of Figures

List of Tables

PREFACE

The thesis in front of you is the conclusion of my time at TU Delft. It feels strange that it is all coming to an end now but I look back with a lot of content. It has been a time filled with a lot of fun with many (new) friends, a lot of music and of course a lot of obtained knowledge.

When I was 18, I doubted a long time whether I wanted to study history or civil engineering. I chose for the latter because I felt that with this a bigger impact on the world could be made. Still, I sometimes missed history's social and narrative aspect when performing calculations on concrete beams. It was only during the MSc courses, some of this narrative came back to life during courses such as "integrated water management" and "water, people, systems and society", where I learned that technical solutions might work in theory, but that the narrative of reality is very difficult to capture in technical terms. I think it's safe to say that this thesis is a perfect example of this because yes, modeling water access is possible to some extent and can inform us a lot but, only when you talk to and learn from the actual water users you can understand the story. I really enjoyed combining my technical modeling skills with the stories the survey told us in this thesis.

ACKNOWLEDGEMENTS

First and foremost a big thanks to the chair of my committee, Edo Abraham. Edo, thank you for being super approachable. I know a lot of my friends struggle with unresponsive supervisors, you are quite the opposite which is super nice. I learned a lot from you on the complexity of such water access problems and it was so cool to find out you're into Miles Davis too! All of your feedback and discussions we had really helped to improve my thesis Didier, thank you for taking me up in the African Water Corridor (AWC) and really laying the foundation of this research. In the end I diverged a little from the original Water Gap problem (which remains interesting!) but it was our discussions that pushed me to format an exact topic. I also really enjoyed that every time we zoomed you seemed to be in a different country and had some new stories to tell. I'm glad we got to meet in person a few months back too. Martine, thank you for being critical to my work but always in a supportive way. Doris, thank you for your kind words during the green light meeting, that really gave me a boost. Finally, thanks to the whole AWC team for putting trust in me and financing the survey campaign in Bushenyi-Ishaka municipality.

Next, a big thanks to James Tayebwa from Makerere University (Kampala) with whom I worked closely in setting up the survey campaign and formatting the questions we wanted to ask to the residents of Bushenyi. It was my pleasure to meet you and also you taught me a lot. As Covid prevented me to attend, really all the credit for the exe-

cution of the survey campaign goes to James and his team: Praise, John, Raymond and Doreen. Thank you all for doing such a good job! Furthermore, many thanks to Professor Frank Kansiime from Makerere University for providing a lot of knowledge and organising many things within Makerere University. Naturally, thank you to the residents of Bushenyi and surroundings that were so kind to answer our survey questions.

Finally I want to thank my roommates that were always capable to distract me just enough from the thesis and covid times. My study friends with which I spent countless hours in the beloved studoc room and at the coffee machines, you really made this time fly by. Lastly thank you to my family for supporting me through my entire study time and just always being there for me.

Jan Geleijnse
Rotterdam, February 2022

1

INTRODUCTION

The United Nations (UN) Sustainable Development Goal (SDG) #6 reads that by 2030 universal and equitable access to safe and affordable drinking water is achieved for all (United Nations General Assembly, 2015). The African Water Corridor (AWC) is a project of Delft Global Initiative and has set a goal that contributes to SDG 6, namely to "*develop - with partners - one or more African Water Corridors, carriers for sustainable development through co-creation of innovative, open-source technologies and sustainable implementation strategies. We want to ensure that water will no longer be a limiting factor for development of human and natural resources in Sub-Saharan Africa*" (Delft Global Initiative, n.d.). This thesis falls under the umbrella of the AWC and researches the domestic water access and - consumption, which is perhaps the most vital water use of all.

Safe drinking water contributes to day-to-day health directly through safe digestion and hygiene. But, as populations in many areas of the world are growing rapidly and water sources are limited, safe water access is at risk for many people. The Joint Monitoring Program (JMP), which is a collaboration of Unicef and the World Health Organisation (WHO), considers people with unimproved (and therefore possibly unsafe) water access to be people that have to travel over 30 minutes to a safe water source, or people using an unprotected source. The latter means that the source is not protected from possible contamination from faecal or chemical substances (WHO, 2017).

Monitoring safe water access relies primarily on household (health) surveys and census data (Bartram et al., 2014). There are two limitations with this: firstly, survey questions on water access typically focus on the primary drinking water source only. However, people often drink and make use of a multitude of water sources for many purposes. This can cause an overestimation of the population share with permanent safe access as some of the sources used might not be safe (Elliott et al., 2019). Secondly, health survey data is, for privacy concerns, spatially limited. For instance, in Uganda's Demographic Health Survey (DHS), households are clustered by their spatial location where households in the same cluster can have a spacing of up to 10 kilometres (Uganda

Bureau of Statistics, 2018). Besides, health survey data is almost never nationwide but merely an attempt to an as good as possible representative sample of the population.

Instead of using household surveys, the Water Point Data Exchange (WPDx) tries to monitor water access by creating a platform on which NGOs, governments and water companies can share water point data. Water point data consists of the geographical coordinates of a water access point and the water access type (borehole, spring, etc.) (Unicef, 2015). Such data comes with the advantages that it is not spatially limited for privacy concerns and that it better captures the multitude of different types of sources used in an area (Yu et al., 2019). However, coverage of this database is again often not nationwide and therefore in need of modeling methods that can predict the presence of water access points into areas of which this unknown.

Monitoring water consumption is important as the usage or availability of (too) low volumes can indicate poor hygienic circumstances (WELL, 1998), but also because water consumption should be in reasonable balance with available water resources. With respect to water consumption, it is difficult to gain information from WPDx directly and also health surveys generally do not include questions on the water consumption in terms of volume. The latter is mainly a result of water volumes being a difficult number to estimate for users or survey enumerators when it is not metered. The consensus in literature is that water consumption is primarily a function of the travel time to and from the water source (WELL, 1998), however there remains a lot of uncertainty around this (Cassivi et al., 2019). Contradictory to water consumption, this travel time (albeit to the primary drinking water source only) is often included in DHS data.

Earlier research applied a machine learning biological species modeling technique (called MaxEnt) to parts of the WPDx data from Kenya and predicted the relative probability distribution of finding surface water access points and unimproved wells across Kenya (Yu et al., 2019). For these predictions, MaxEnt uses a number of physical, geographical and socio-economical features (predictive data) of which it learned the combination of characteristics that indicate the presence of water access points (Merow, Smith, & Silander, 2013). These features come in the form of nationwide maps such that the output of MaxEnt is nationwide and aligned with the earlier defined need for modeling methods that could predict water point presence in areas where this is not registered on WPDx.

However, the methodology applied by Yu et al. (2019) comes with a number of limitations. Firstly, it models only two access types (unimproved wells and surface water access), which are both JMP considered unimproved sources (WHO, 2017). To get a more complete overview of water access, other access types should be included. Secondly, the output of the model is a relative probability of presence, i.e. relative to the total number of presences. To research the (co-)existence of multiple water access types, their usage but also to evaluate the level of water access across the country, it would be better to predict the absolute number of presences instead of the relative probability of presence. Finally, the research by Yu et al. (2019) is limited in its validation within cities

and presents the results from the larger (province size) perspective. Looking into cities is relevant as urbanization is expected to continue and especially midsized and small cities are often overlooked in research (Santos et al., 2017). Small cities are interesting because of the large part of the population living 'in between' environments combining both urban and agricultural characteristics (Marks et al., 2020).

The methods presented in this thesis improve on Yu et al. (2019) by switching from MaxEnt to a different machine learning technique that allows for more complex non linear behaviour (neural networks), of which several setups are presented and their performances compared. Next to this, we present a method that scales the relative output to a prediction of the absolute number of water point presences. On top of the two by Yu et al. (2019) modeled access types, six more are included, including improved sources such as piped water access and boreholes.

Next to this data based modeling approach to water access, a survey campaign among 500 households was executed in Bushenyi, a small city in the South West of Uganda. Contrary to more standard surveys on water access, this survey explicitly researches the behaviour and reasoning of households using a multitude of water sources. It also asks respondents to quantify their consumed water volumes, different water purposes and access types used. The survey data is used in multiple ways. Firstly, it allows for validating parts of the modeling outputs in a small city. Secondly, factors that could potentially predict water consumption are researched as it is suspected that water consumption is from more dependant than travel time alone. These factors include the model outputs (water access point presence), but also socio-economic information from the surveyed households such as education level and income. Lastly, the survey serves as addition to the models in the sense that it shows the complex dynamics and large variances in water access and consumption that are easily lost when scaled to coarser modeling resolutions.

By combining the results and findings from both the models and the survey, the aim is to get a good overview of water access and water consumption, their dynamics and interdependence and to which predictive information they potentially relate. This information, combined with the model's nationwide predictions of water access point presences, can be used by governments and NGO's to prioritise areas with poor water access for water access improvement campaigns and for better water management in general.

1

1.1. PROBLEM STATEMENT

In short, this study has the following objective:

To spatially model the type of domestic water access and the domestic water consumption in areas where true measurements of this are unavailable, with a particular interest in small and midsized cities.

For that the following sub questions will be addressed:

- *What physical, hydrogeological and socio-economic information can indicate the presence of certain water access types and household water consumption?*
- *To what extent can the expected presence of water access types be used to quantify water consumption? And can a predictive relationship between certain water access types and water volumes used in households be established?*
- *What areas in the study area are at risk (i.e. have unimproved water access or access with large travel times) and how can governments and NGOs use the modelled information to adapt?*

1.2. THESIS OUTLINE

This thesis consists of a literature review (Chapter 2) showing the current status of academic research towards water consumption and water access predictors. This chapter also introduces the study area, Uganda and more specifically Bushenyi, and its current status of water access. The next two chapters introduce the two research methods in this thesis: that of the models (Chapter 3) and the survey (Chapter 4) respectively. In the Results, Chapter 5, the model performance is discussed after which both the model and survey results are placed into context by relating them first to water access types, secondly to the travel time to a water source and finally to the households water consumption. The broader perspective and meaning of the results next to the limitations of the methods and their impact on the results are presented in the Discussion (Chapter 6. Related to the limitations, this Chapter also presents some recommendations. Lastly, the Conclusion (Chapter 7) shows the key findings and answers the research questions.

2

LITERATURE REVIEW

The first part of the literature review presents the literature consensus on predicting water access (types), the second on predicting water consumption. The final parts give information about the study area: Uganda and Bushenyi.

2.1. MODELING WATER ACCESS

As mentioned in the Introduction (Chapter 1), the Joint Monitoring Program (JMP), which is a collaboration of Unicef and the World Health Organisation (WHO), considers people with unimproved (and therefore possibly unsafe) water access to be people that have to travel over 30 minutes to a safe water source, or people using an unprotected source. The latter means that the source is not protected from possible contamination from faecal or chemical substances (WHO, 2017). In order to monitor the level of access in countries for which SDG 6 has not yet been achieved (WPDx, 2021a), the Water Point Data Exchange (WPDx) has created a platform on which several NGOs and governments share water point data (Unicef, 2015). This data adheres to the WPDx water point standard, which means that the minimum required data for each point is: location (lat., lon.), availability of water during visit, date of recording and the type of water access. Often more information is shared such as the (dys)functionality of a source and if the source is an improved drinking water source or not (WPDx, 2021b). WPDx distinguishes 12 different water access types (see Table 3.1) which include boreholes, (un)protected springs, piped water and surface water access. At the moment of writing (Sept. 2021) there are 613 563 data points in the WPDx database. A problem is that in many countries in Sub Saharan Africa (SSA) this database is incomplete, not covering the entire country.

Yu et al. (2019) applied a machine learning modeling technique to the WPDx data and were able to predict the relative probability of presence distribution of two access types (surface water and unimproved wells) to a gridded 1 kilometer resolution in Kenya. For this, the model learns how to use a number of physical, geographical and socio-economical features (predictive data) to predict the (non)presence of water access points.

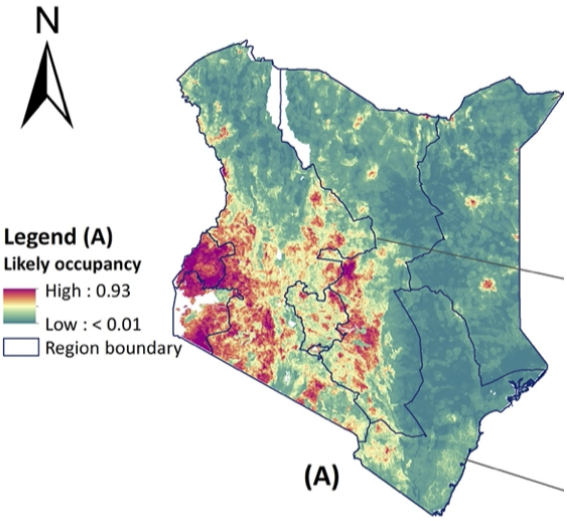


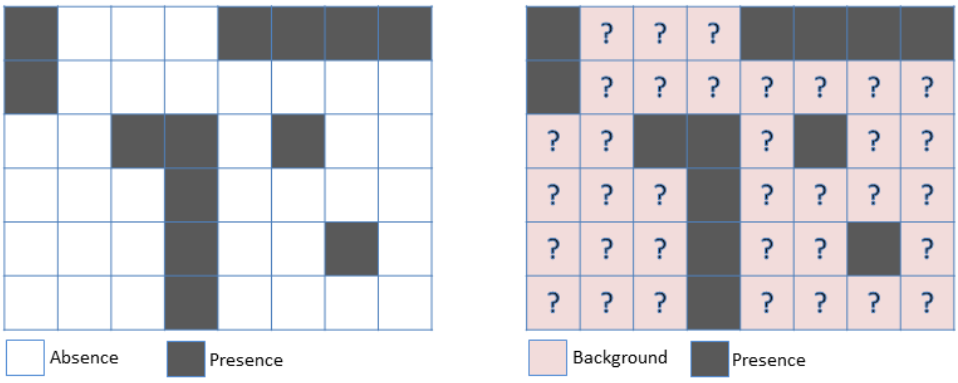
Figure 2.1: Output from Yu et al. (2019): Predicted unprotected dug well occupancy across Kenya

Their results are promising judging by the predictive power of the models which score 0.9 at the Area Under the Receiver Operator Curve (AUC, a common evaluation parameter for such classifying models). More visually, the output of Yu et al. (2019) results in gridded relative (to the total number of presences) probability maps such as the one shown in Fig. 2.1. A red color indicates that the landscape characteristics of a grid cell are close to the grid cells where the target water sources were observed. To obtain these results, Yu et al. (2019) used a Maximum Entropy model (MaxEnt), which finds its origins in the biological species modeling field (Phillips, Anderson, & Schapire, 2006).

2.1.1. PRESENCE ONLY DATA

In the biological species modeling field, the objective is usually to find which geographical areas correspond environmentally to areas where the species was observed and with that, model a species (potential) habitat. Biological species modelers usually refer to the mentioned observations of the species as point data. This is typically subdivided in two categories: presence-only and presence-absence (Elith et al., 2006). If we take the species modeling example of a daisy flower: for presence absence data, a (random) selection of cells in the study area is researched and both the cells with daisy flowers (presence) as well as the cells without daisy flowers (absence) are registered as such. This allows for a strong comparison between the environmental features enabling (presence) or disabling the presence (absence) of daisy flowers. Presence only data is mostly a result of less organised methods of data collection. For presence only data there is only information on which cells do contain daisy flowers (presence), of the other cells (pink in Figure 2.2b), it is unknown whether they contain daisy flowers or not. Still, if the sample

is large enough, the presence only data can be compared to (randomly) selected background cells which then are often referred to as pseudo-absence cells. This technique is under the assumption that the species (daisy flowers) is not prevalent in the entire study area.



(a) Presence – absence data. This contains the complete information of every cell (blue outline) containing a presence or no presence (=absence)

(b) Presence-only data. Of the pink cells it is unknown whether or not they contain a presence, no presence or that they were not surveyed.

Figure 2.2: The difference between presence-absence data (left) and presence-only data (right) is that for presence absence we have full information on both presence and absence of surveyed cells while with presence only data we have only full information on the cells containing presences while for the other cells presence is unknown. Figure inspired by Golini (2012).

2.1.2. MODELING PRESENCE ONLY DATA

For species, the data is most of the time presence-only data as this is easiest to obtain. The used data from WPDx in Yu et al. (2019) is presence only data as it does not give information about the absence of water access types at a certain location. Elith et al. (2006) compared a number of models that were designed to use presence only data to predict the presence of (biological) species. They showed that the machine learning models such as MaxEnt, outperform the more traditional models for such data. West, Kumar, Brown, Stohlgren, and Bromberg (2016) even showed that MaxEnt performs almost just as well on presence only data as regressions on presence-absence data. However, Botella, Joly, Bonnet, Monestiez, and Munoz (2018) compared the performance of MaxEnt to neural network models for the distribution of presence of 5 plant types in France and showed that the neural network can outperform MaxEnt, mainly because it is able to capture the non linear transformations of input features better. Neural networks have been shown to be very capable of modeling non-linear problems also in other settings such as forecasting energy prices (Heijden, Lago, Palensky, & Abraham, 2021) and water prices (Qi & Chang, 2011).

2.1.3. PREDICTORS FOR WATER ACCESS TYPES

In this thesis, a modeling technique inspired by MaxEnt is used to predict the presence of water access points of different types (boreholes, springs and six more). Therefore, it is relevant to know which information can be used to predict water access in general but also which data can potentially be used as predictive features. Modeling the presence of water access points has, aside from Yu et al. (2019), to our knowledge, little precedent. Rahmati et al. (2018) found that the slope of the terrain was the most important factor for unprotected spring presence in Iran. Yu et al. (2019) found that rainfall and elevation were important predictors for both unprotected wells and surface water access. More general, there are a number of water access types that rely logically on the availability of groundwater. These are: (un)protected shallow well, (un)protected spring, borehole and often piped water. Indicators for groundwater availability are: depth to groundwater, groundwater productivity, groundwater storage, drainage density, elevation, slope, topographic wetness index (TWI), proximity to inland water, land use, lithology, and soil texture (Rahmati et al., 2018) (Miraki et al., 2018) (Naghibi, Pourghasemi, & Dixon, 2015). For the usage of rainwater harvesting, the rainfall frequency and intensity are obvious necessities.

Besides the features that have a direct effect on the availability of various water sources, there are also many socio-economic characteristics that could indicate particular water access types. As the water access points are used by humans, it is only relevant to look at places with humans. This is indicated by the population density, villages, cities or other urban land uses. Furthermore, some access types are only possible in situations where there are enough people. For instance, a piped water network is costly and the using population and/or the government should be capable and willing to make such investments. Baguma and Loiskandl (2010) showed that subsidies, especially in the form of hardware, can help in the adaptation of rain water harvesting techniques. Lastly, a lot of literature suggests a linkage of (improved) water access to income or wealth, indicating a positive impact of wealth to improved water sources (Mahama, Anaman, & Osei-Akoto, 2014) (Adams, Boateng, & Amoyaw, 2015) (Qi & Chang, 2011) (Liu, Savenije, & Xu, 2003).

2.2. PREDICTING DOMESTIC WATER CONSUMPTION

While socio-economic and environmental conditions are known to be good predictors for water access types (see section 2.1), the water consumption is suggested to be a function of the travelled time to and from the water access point as depicted in Figure 2.3 (WELL, 1998) (WELL is affiliated to the WHO). Figure 2.3 shows that for return trip travel times (including queuing) under 30 minutes, the volume of water used remains constant at about 15 liter per capita per day (lpcd). Only when the water access is located on premises the consumption rises quickly to 50 lpcd. When water is piped into the house the water consumption increases even further which can be seen in Table 2.1. This is because it allows for water intensive machines (such as washing machines) and eliminates effort (Howard et al., 2020). For trips over 30 minutes the water consumption decreases to levels under hygienically healthy minimum requirements (WELL, 1998).

Twenty two years later, the graph from (WELL, 1998) was republished by the WHO

in Howard et al. (2020) and discussed extensively: Rhoderick (2013) showed that in general people using off plot sources consumed less water than people with on plot access. This would comply with Figure 2.3, however Rhoderick (2013) remarks that the graphs shape and specific break points differ per setting (Howard et al., 2020). Cassivi et al. (2019) performed a literature study and hesitantly agree that there is a negative correlation between travel time and water consumption but at the same time point out that the methodology used to research this is often different, which makes it difficult to compare studies and to justify the relationship from Figure 2.3 (Howard et al., 2020). Their conclusion is that the volumetric in home water consumption is primarily sensitive to improvements in access level and thus that increases in water consumption occur at distinct thresholds of access (Howard et al., 2020). Table 2.1 is from a study in Uganda and is also an indication that water consumption could be related to access level by means of access type. Similar values were reported in Kenya and Tanzania (Thompson, Purras, & Tumwine, 2001).

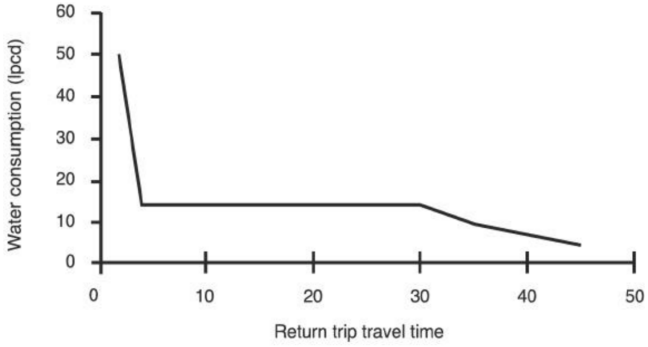


Figure 2.3: Water consumption in liter per capita per day (lpcd) related to travel time to access point (WELL, 1998)

Ideally, meaning that the relationship from Figure 2.3 holds, water consumption predictions could be based on the travel time of a household to its (closest) water source. However, Marks et al. (2020) described that the type of access and payment influence the choice of people to travel further for their water access and thus should be included in the estimation. Furthermore, other research indicates that water consumption relates to a lot of socio-economic information of households and communities. For example: in Florida water consumption was linked to employment rate and population growth (Qi & Chang, 2011). Kennedy et al. (2015) suggest correlations between area per capita and water demand per capita, but also between GDP per capita and water demand in megacities. Liu et al. (2003) and Fielding, Russell, Spinks, and Mankad (2012) found positive relationships between income and water consumption in China and Australia respectively. A higher educational level also increased water demand (Fielding et al., 2012), although this might be in correlation with income. Furthermore, both Fielding et al. (2012), Fan, Liu, Wang, Geissen, and Ritsema (2013) and Thompson et al. (2001), found that larger

Type of supply	Distance from home	Range of consumption (lpcd)
Communal water-point (well or standpost)	> 1000 m	5 – 15
Communal water-point (well or standpost)	250 m – 1000 m	10 – 30
Village well or Communal standpost	< 250 m	15 – 50
Yardtap	in compound	20 – 80
House connection - single tap	in house	30 – 80
House connection - multiple taps	in house	70 – 250

Table 2.1: Average water consumption related to distance and supply type in Jinja, Uganda (WELL, 1998), (Howard et al., 2020).

household sizes resulted in less water consumption per capita. This is probably a result of a more efficient usage of water for shared purposes such as cooking. Finally, in Malaysian towns, households living in long houses (a traditional housing type) were less likely to have improved water access which was probably an effect of the rapid urbanization of these areas, allowing no time for creating good water infrastructure (Kong et al., 2020). In summary: water consumption most likely is a function of travel time to source combined with the type of water access and other socio-economic conditions. All of this again shows that the access to water and water consumption can be very specific to the circumstances.

2.3. STUDY AREA: UGANDA

The African Water Corridor (AWC) has selected three Sub-Saharan development corridors: Uganda, Ghana and Mozambique. Of those three countries, Uganda has the largest subsets of water point data, including a variety of both improved and unimproved water sources. As more data allows for better modeling, the thesis will focus on Uganda.
Uganda is a country in Eastern Sub Saharan Africa with a population size of somewhat over 44 million (World Bank, 2019). According to Uganda's most recent Demographic Health Survey (DHS) from 2016, 77.9 % of Ugandans have access to safely managed drinking water (up from 70 % in 2011) and 20.8 % have access to safely managed sanitation services (Uganda Bureau of Statistics, 2018). From Figure 2.5 it can be concluded that urban residents are a lot more likely to have improved water access than rural residents. In rural areas the tubed well or borehole is used the most but the number of people using unimproved sources still is significant. In urban areas, water access is better in general. The urbanization level in Uganda is 24.6 % (O'Neill, 2021). In terms of travel time (relating to Fig. 2.3), 46.8 percent of the population has to make a return trip of more than 30 minutes for their water in 2016 (Uganda Bureau of Statistics, 2018). Also in Ugandan cities people travel several minutes for collecting water and often people do not choose the closest source but a source that is cheaper or free or has a better (per-

cieved) quality (Marks et al., 2020) (Howard, Teuton, Luyima, & Odongo, 2002).

Figure 2.4: Uganda lies in the mid-East of Africa (figure from Wikipedia)

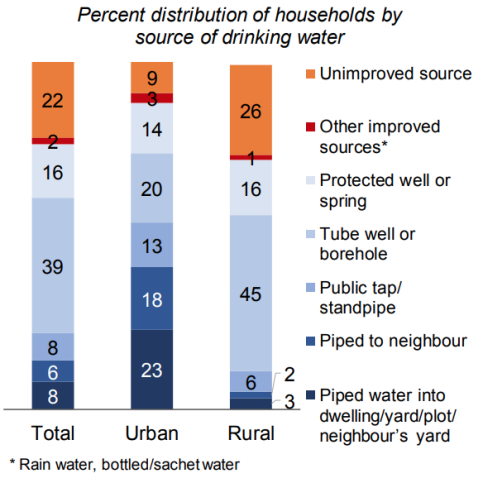


Figure 2.5: Distribution of different water sources in Uganda in 2016 (latest available Demographic Health Survey) (Uganda Bureau of Statistics, 2018).

With regards to the water availability, Nsubuga, Namutebi, and Nsubuga-Ssenfuma (2014) refer to a water resource management sub-sector reform study of 2004 that showed that especially the North East and South West are under water stress, i.e. have limited (natural) water resources per capita. This is shown in Figure 2.6. The reform study also predicted that this will increase towards the future. Nsubuga et al. (2014) continue by pointing out that global warming induced climate change does and will cause variation

in rainfall patterns upon which many of Uganda's water resources rely (and farmers for growing cycles). When rainfall will decrease, Nsubuga et al. (2014) warn for over exploitation of the (other) water resources. However, Kilimani et al. (2013) showed that indeed there have been changes in rainfall patterns, which is a problem to farmers, but show that in fact the other available water resources in Uganda are under utilised and could be used for improved irrigation systems to cope with the varying and unpredictable rainfall.

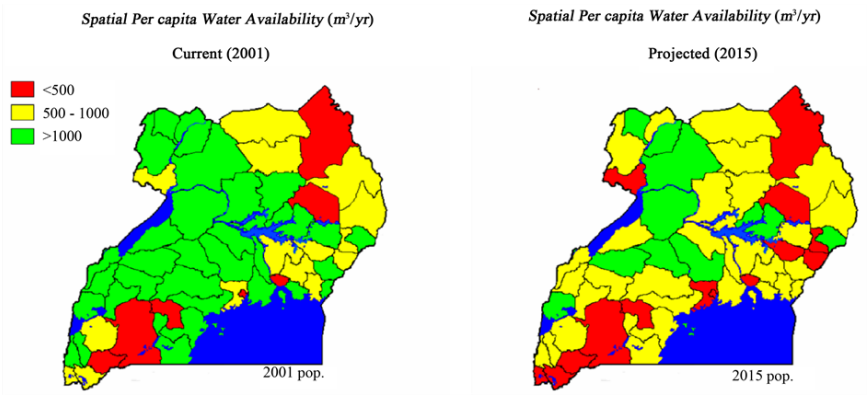


Figure 2.6: Water availability in 2001 and a 14 year ahead prediction at the time (Nsubuga et al., 2014).

Access type	Average Travel Time: Urban [min.]	Rural [min.]
Piped dwelling	0.0	0.0
Piped yard	0.0	0.0
Piped to neighbour	0.0	0.0
Public tap	8.7	27.4
Rainwat. harv.	1.0	3.9
Prot. well	26.7	31.9
Unp. well	33.9	39.0
Prot. spring	26.6	39.9
Unp. spring	38.1	40.1
Surface wat.	29.3	44.8
Borehole	36.4	45.4
Bicycle vend.	67.9	49.4
Truck	6.7	39.6
Bottled	4.2	12.3
Sachet	4.8	6.9
Other	8.3	26.4

Table 2.2: Average travel time per water access type considered in the DHS data for both urban and rural areas. Table created from DHS data (Uganda Bureau of Statistics, 2018) by author.

In a first analysis of the DHS data, it was found that in Uganda different access types

have different average travel times as well. This would support the more general (also outside Uganda) findings of WELL (1998), Thompson et al. (2001) Rhoderick (2013) and Cassivi et al. (2019), who stated that water consumption is a function of travel time to and accessibility of the water source. By looking at Figures 8.1, 8.2 in appendix A and Table 2.2 it can be seen that based on average travel time and travel time distribution, three groups can be distinguished in urban and two groups in rural settings. In both urban and rural settings the sources that are in general off plot (e.g. springs, boreholes) have a higher average travel time than on plot connections such as piped connections. In the urban setting, boreholes and unprotected springs have a higher average travel time then their other off-plot peers, which is why they are put in a separate group.

2.4. BUSHENYI-ISHAKA MUNICIPALITY

Bushenyi and Ishaka are two neighbouring towns that together form the Bushenyi-Ishaka municipality. It is located in the South West of Uganda, close to Lake Edward. The area has experienced and is undergoing rapid urbanisation shown by the steady population growth of 6 % for the past 5 years. Bushenyi town is the administrative centre of the Bushenyi district (Silva-Novoa Sanchez, Kemerink-Seyoum, Waiswa Batega, & Paul, 2020) that is now housing 250 000 people (Bushenyi District, 2020). The area has wetlands in the valleys and from the more mountainous areas small creeks flow, especially during the rainy seasons. There are a number of locations where groundwater reaches the surface and form a natural spring (Silva-Novoa Sanchez et al., 2020). Many inhabitants rely on these springs for their daily water supply. The access facilities to these springs were built by the government agency responsible for rural development and are kept clean by neighborhood committees. As the town is growing and urbanizing more, many households switch to a piped water system that is provided and maintained by the National Water and Sewerage Corporation (NWSC). This brings questions as to who is responsible for the water supply from the springs. Furthermore, Bushenyi is an interesting research location because of the large part of the population living 'in between' environments combining both urban and agricultural characteristics (Marks et al., 2020). Researching such locations is relevant as urbanization is expected to continue and especially midsized and small cities are often overlooked in research (Santos et al., 2017).

Surveys were conducted in 2018 as part of a research by Marks et al. (2020), researching water access and sanitation in Uganda. However, despite earlier mentioned importance to hygiene, no questions were included with regards to the consumed volumes of the household. In 2018, one in five households had water access on plot (Marks et al., 2020). Especially during the rainy season, people use multiple sources for their drinking water including rain water harvesting (Marks et al., 2020). The usage of multiple water sources by one household is also reported in other settings but often not surveyed properly: survey campaigns mainly take place in the dry season and do only focus on the primary drinking water source (e.g. Uganda Bureau of Statistics (2018)). The use of multiple water sources can result in a greater resilience to water insecurity (Elliott et al., 2019), but can also cause an overestimation of the number of persons with safe drinking water access as they might drink (every now and then) from unimproved sources too (Daly, Lowe, Hornsby, & Harris, 2021). For that reason, there is a growing call to research

Figure 2.7: Bushenyi is located in the South West of Uganda (map from Wikipedia).

the incentives as per why, where and how households do choose their water sources as this can help improve safe water access campaigns (Elliott et al., 2019). In that fashion, one of Marks et al. (2020)'s conclusions was that one-quarter of respondents said to travel to a water source that was not the source closest to their home. This is also interesting given the relationship between travel time and water consumption from Figure 2.3. The most given reasons for traveling further were that the taste of the used source was better or that the source closest was too costly or of bad (perceived) quality. Elliott et al. (2019) report seasonality, costs and aesthetic reasons as the predominant causes of multiple water source use. Peoples reasoning for choosing a water source over another is one of the things that will be researched further in this thesis. Besides these (more or less) quantifiable reasons for traveling further, there were also more cultural reasons. Natural springs for instance seemed to be seen as closer to nature and thus more pure (Marks et al., 2020). For instance, Silva-Novoa Sanchez et al. (2020) interviewed Bushenyi residents and learned that people share spring water that is on their land, willingly with neighbours. According to their field data, this is partially because of the belief that it would be perceived as a great injustice to deny access or request money for water that flows naturally out of the soil, but also due to the difficulty to exert control over the springs as most do not have taps and water would otherwise be spilt. With the survey conducted in this thesis, the aim is to further analyse the motivation for choosing not the closest source and in parallel, the indicators of consumed water volumes.

In 2018, most people in Bushenyi made use of protected springs (48%), followed by unprotected springs (20%) and NWSC taps at the home or yard (18%), other sources include boreholes (5%), rainwater harvesting (3%) and surface water (2%). Furthermore, households in the urban parts of the area were more likely to use improved water sources. This includes piped connections for which users have to pay on a regular basis (Marks et al., 2020).

3

Modeling Methods

To answer the problem statement and its accompanying research questions, a dual methodology is applied. For the first method, the distribution of water access points accross Uganda is modeled using an improved biological species machine learning modeling technique. This is done per access type such that there is a model for piped water, one for boreholes etc. The output is a gridded map with per cell the predicted number of water access points per access type. Most of the models come in the form of Multi - Layer Perceptron (MLP) *Classifying* Neural Networks (NN), which make predictions of water access point presence, using several geohydrological and socio-economic feature data-sets. As these models can be designed in different ways, multiple are created and in the Results Chapter their performance is compared.
Besides models on access types, also models that predict the Demographic Health Survey (DHS) cluster average travel time to and from the water source. This is done because the Literature Review suggested that travel time is negatively correlated with water consumption and because travel time is an important indicator related to the level of water access in general. For this, a regression is applied to nationwide cluster average household travel times reported in the Ugandan DHS. This time a MLP *Regression* NN is build, that uses the same feature data-sets to make predictions of the average cluster travel time. To research the relationship between water access and travel time, a similar regression is applied but this time it uses the access point distribution output from the classifying NN instead of the feature layers to predict travel time. As a last step, the water access point distribution output from the classifying models is compared to the in Bushenyi reported household water consumption. Here the hypothesis is that households cells with more predicted access points, consume more water. Please find a schematic of these steps in Figure 3.1.

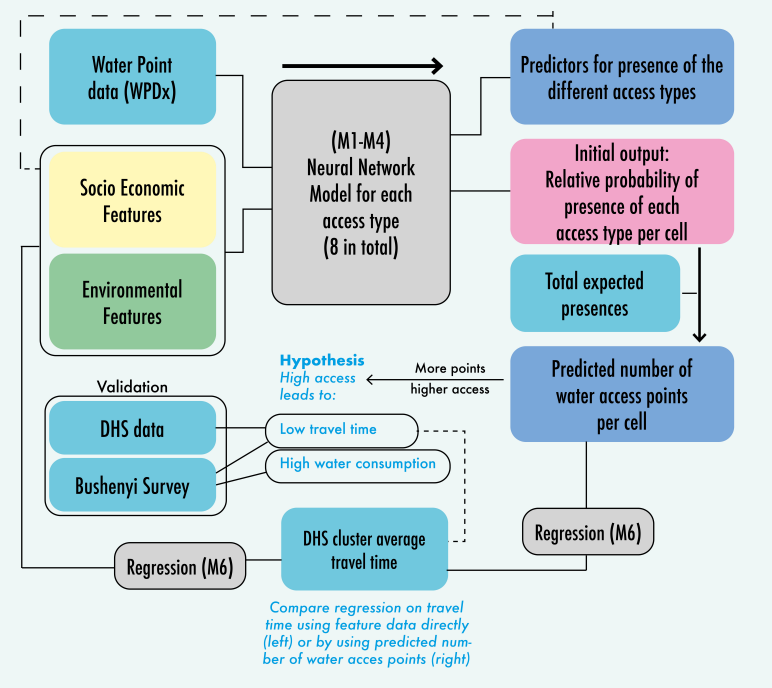


Figure 3.1: For the model approach, neural network classifying models (M1-M4) are trained to predict the presence of water points (of 8 different access types) using predictive socio-economic and geohydrological feature data. The initial output of such models is the relative probability of each Ugandan cell containing e.g. a borehole. By using an estimation of the total number of expected presences, this is scaled to the total number of e.g. boreholes in a cell. This output is compared to nationwide household travel time data in Uganda (from DHS) as well as locally obtained data of household level water consumption in Bushenyi. The hypothesis here is that high access leads to both low travel times and high water consumption. Besides making predictions regarding the number of presences, two regressions are performed separately on the DHS reported travel times, first using the feature data and later using the model output. This is shown at the bottom of the Figure.

The second method, zooms in on a specific city: Bushenyi, Uganda. As part of this thesis a survey is conducted in Bushenyi, researching again indicators for water consumption but also validating and extending the relationship between water consumption and travel time presented in section 2.2. The methodology behind that survey is presented in Chapter 4.

CELLS

As most of the input and all of the output of the model is described in terms of cells, to avoid confusion, cells are defined here first. If we were to place a rectangle on the map of Uganda and divide that rectangle in squares of 1 km^2 each, each of these squares represent a cell. In total this results in 686 x 652 cells. As Uganda does not have the shape of a rectangle, some of these cells lie outside the country; these are automatically excluded in all of the calculations described in this Chapter. Eventually, predictions will

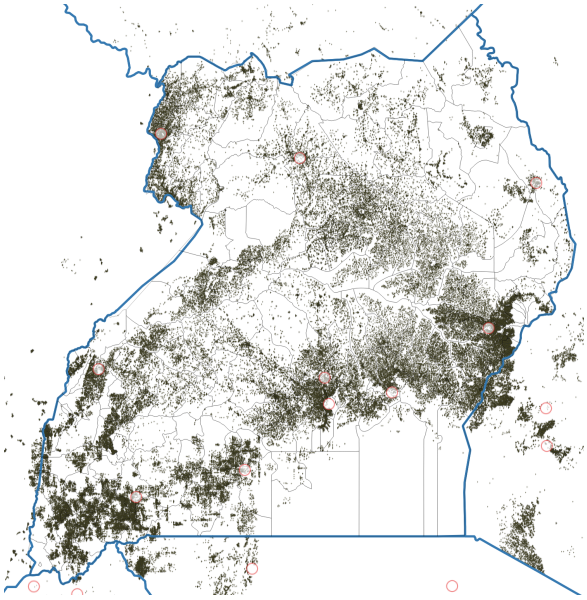


Figure 3.2: Water Point Distribution across Uganda (red circles indicate cities). Figure made with ArcGis Pro.

be made in terms of water access points per cell (e.g. 5 boreholes in this cell).

3.1. USED DATASETS

There are two types of data used in this research. One type is the training/validation data, which contains that what will be predicted: water access (from WPDx) and travel time ot the water source (from DHS). The second data type consists of all the different feature layers that could potentially be predictors for water access points and travel time.

3.1.1. TRAINING AND VALIDATION DATA

WPDX DATA

Water point data is extracted from the Water Point Data exchange (WPDx[1]), which was introduced in Section 2.1. The distribution of all the WPDx data across Uganda can be seen in Figure 3.2. Note that this is the data from all the considered access types and that for most individual access types, the nationwide coverage is not nearly as high. The different water access types and their counts in Uganda are shown in Table 3.1. Please note that this is not the true distribution of water access types in Uganda, it only represents which access types were registered for WPDx. In total there are 121 331 WPDx points in Uganda. Which means 0.2 points per square km and 0.0043 per inhabitant, this is above average when compared to the other WPDx countries. In Bushenyi district, there are close to 600 WPDx points.

[1] https://www.waterpointdata.org/

Access type	WPDx points	Improved	Modeled
Borehole	36758	Yes	Yes
Packaged water	1	Yes	No
Piped Water	124	Yes	Yes
Protected Shallow Well	1632	Yes	Yes
Protected Spring	28161	Yes	Yes
Rainwater Harvesting	18624	Yes	Yes
Sand or Sub-surface Dam	1	Yes	No
Surface Water	535	No	Yes
Undefined Shallow Well	18260	Unknown	No
Undefined Spring	18	Unknown	No
Unprotected Shallow Well	323	No	Yes
Unprotected Spring	239	No	Yes

Table 3.1: The considered Ugandan water access types and their number of presences in the WPDx database, whether or not these are considered improved/unimproved by WHO (2017), and whether these are modeled in this thesis.

The research by Yu et al. (2019) extrapolated Kenyan water point data of Surface Water and Unprotected Dug Wells in two separate models. In this research this is expanded to all water sources displayed in Table 3.1, excluding the sand surface dam, packaged water and the two unknown facility sources. The first two are excluded because the data sets are too small and the latter two because it is important to predict whether access is deemed protected or unprotected (improved/unimproved), as this gives information on the hygienic safety of the source.

DHS DATA

Another data-set that is used, is the 2016 Ugandan Demographic Health Survey (DHS). This contains information of 20 000 Ugandan households. It includes the households travel time to the water source, primary drinking water access type, wealth index, household size and much more. For privacy concerns, this data was grouped into 600 clusters. Every cluster has a GPS location and every data point (containing household information) in the cluster, falls within a range of that location, i.e. 2 km in urban and 5 km in rural settings. This is visualised in Figure 3.3. This results in difficulties when using the clustered data as each cluster spans multiple cells and because within many clusters, the standard deviation of travel time to water source is high (Mean SD = 29 min.). However for validation purposes, it is possible to validate on the mean travel time of the households in a cluster. Besides validating on the travel time, model predictions on the number of presences of water access types can be compared to the DHS reported usage. For this Spearman's rank order correlation is used to find if and how strong the average number of users per cluster cell is correlated with the number of predicted presences by the model.

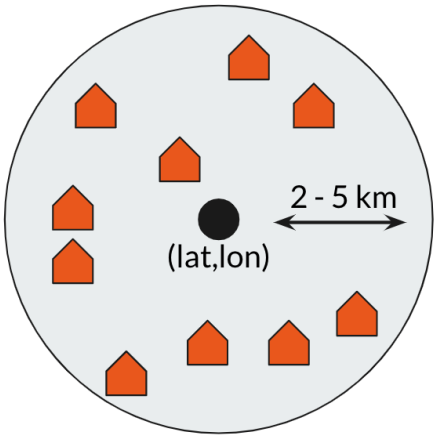


Figure 3.3: Schematic of DHS cluster, in urban and rural areas the radius is 2 and 5 km. respectively. All of the households (orange) receive the same coordinate but in fact are displaced multiple kilometers from the center for privacy concerns. Note that these DHS cluster thus span multiple of our 1 km^2 cells.

3.1.2. FEATURE DATA

We selected maps holding information on environmental, geo-hydrological, socio-economic and technological features that potentially indicate a cells suitability for (one or more of the different) water access types. All the used feature layers and their sources can be seen in Appendix D and summarised in Table 3.2. As machine learning models are used, we let the models select the most suitable feature layers from the lot instead of deciding this upfront. Much of the motivation for choosing these feature layers is in Table 3.2, and to more detail in the Literature Review (Chapter 2).

All maps are resampled to a 1 km^2 (= 1 cell) resolution using ArcGis Pro. Most maps have that or a higher resolution. During resampling, for numerical data bilinear interpolation is applied and for categroical data nearest neighbour resampling. All the used feature datasets are freely available from the listed sources making this method accessible for all and reproducible.

CORRELATIONS

As suggested in Merow et al. (2013), to reduce colinearity, the feature layers are checked for correlation. This is done using Pearson's R for numerical to numerical data, Cramer's V for categorical to categorical data and η^2 for numerical and categorical data. It was found that groundwater productivity and groundwater storage were correlated (Cramer's V = 0.68) and secondly the euclidean distance to roads and to residential areas was correlated as well (R>0.7). The euclidean distance to roads as well as the groundwater productivity layer were therefore removed from the set of feature layers.

3.1.3. EXPECTED TOTAL NUMBER OF WATER ACCESS POINTS

The DHS is representative for the entire nation (Uganda Bureau of Statistics, 2018) and it holds information on the primary access type of each household. With this, Figure

Feature layer	Reason to include	Data type
Euclidean distance to inland water	The proximity of water relates to water access.	Numerical
Euclidean distance to residential areas	Where people are there is often water.	Numerical
Euclidean distance to cities	Cities are expected to have better water access.	Numerical
Groundwater storage	Presence of groundwater is needed for some access types.	Categorical
Slope	Can influence the suitability of terrain for water access types.	Numerical
Soil texture	Can influence the suitability of the terrain for water access.	Categorical
Urbanisation	Urbanised and less urbanised areas allow for different water access types.	Categorical
Topographical Wetness Index	Can influence water proximity.	Numerical
Population density	Dense and less densely populated areas allow for different water access types, besides people tend to live where water is present.	Numerical
Land cover	Different land cover types could allow for different water access types.	Categorical
Monthly average precipitation	Essential for rainwater harvesting but could also influence other sources.	Numerical
Poverty	Richer households tend to have better water access.	Numerical
Depth to groundwater	Presence of groundwater is needed for some access types.	Categorical
Groundwater productivity	Presence of groundwater is needed for some access types.	Categorical
Population change 1990-2014	Areas that urbanized very quickly have sometimes worse water access.	Numerical
Increased nightlight 1992-2013	Same as above	Numerical

Table 3.2: he feature layers used as water access point predictors in the models, the reason to include them and their data type. For a more detailed overview including sources, see Appendix D.

3.4 presents the distribution of the access types in the urban and rural setting. Generally, boreholes are used both in rural and urban settings while piped access is a lot more common in the urban settings, as we would expect (see Figure 2.5 as well).

However, it would be too simple to conclude from Figure 3.4, that 50% of the rural Ugandan access types are boreholes. Different access types can serve different amounts of persons. For example, a piped connection will usually only serve the household while a publicly accessible borehole will serve a lot more people. In order to get to the average number of persons that each access type serves, the WPDx data is included. For clarity, an example using boreholes is presented below.

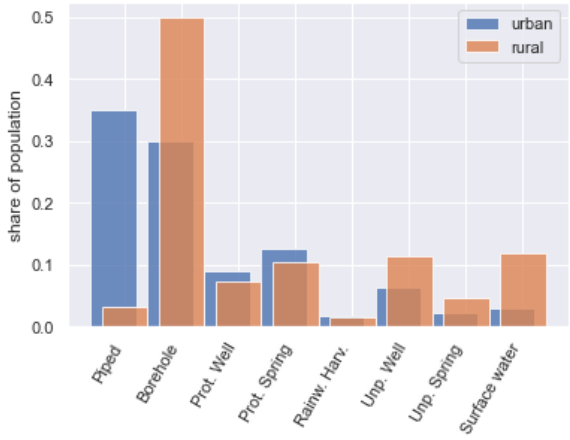


Figure 3.4: Share of population using the various access types in Uganda. Figure made for this thesis from the DHS data.

First, under the assumption that for each cell with WPDx data we have all access points, we divide the cells population (Pop_i) by the total number of presences in the cell of all types ($N_{Wpdx,i}$) (i.e. number of boreholes plus number of springs etc.):

$$\frac{Pop_i}{N_{Wpdx,i}} = U_{Wpdx,i} \tag{3.1}$$

In which i represents a cell. This gives us the number of people per WPDx point in the cell ($U_{Wpdx,i}$), i.e. the average number of users per water access point per cell. Next we take, per access type (boreholes in our example), the average of all the cells where presence of the target type is recorded. So let U_{bh} be the subset of U_{Wpdx} where boreholes were recorded, then:

$$S_{bh} = \texttt{Mean}(U_{bh}) \tag{3.2}$$

This gives us the average number of people a borehole serves (S_{bh}). This is done separately for the urban and rural setting such that we have $S_{bh,urban}$ and $S_{bh,rural}$. Next, by using the information from Figure 3.4 we get the total number of persons using a borehole in the urban ($P_{bh,urban}$) and rural ($P_{bh,rural}$) setting. This is divided by S_{bh} to get to

the total expected number of presences of each access type:

$$\text{Total expected borehole presences}_x = \frac{P_{bh,x}}{S_{bh,x}} \text{ with } x \text{ in } [urban,\ rural] \tag{3.3}$$

The results per access type can be seen in Table 3.3. It is difficult to validate these numbers but the fact that similar amounts of persons sharing the same access type in both urban and rural settings are found, is encouraging. This shows that there is, as one would expect, some limit on the number of persons that can access one source at the same time.

Access type (AT)	$S_{AT,urban}$	Total pres. (U)	$S_{AT,rural}$	Total pres. (R)
Borehole	74	28857	73	303898
Packaged water	NaN	NaN	27	0
Piped Water	38	64296	46	31014
Prot. Shall. Well	76	8446	80	40523
Prot. Spring	167	5360	132	35101
Rainwater Harv.	67	1848	58	11068
Surface Water	39	5364	37	141120
Unp. Shall. Well	112	3974	159	31681
Unp. Spring	107	1451	134	15448

Table 3.3: Estimated average number of persons sharing one access point and total number of expected presences per access type in urban (U) and rural (R) setting

3.2. PREDICTING WATER ACCESS POINT DISTRIBUTION

As the WPDx data does not cover the entire nation, the aim is to fill this gap by building models that can predict the presence of water access points in areas that have not been surveyed for WPDx. As this is done per access type, this also gives insight into the variety of water sources that is used by households. Yu et al. (2019) showed that it is possible to use predictive feature data such as described in Section 3.1.2, and a biological species modeling technique called MaxEnt to predict the relative probability of finding water access points (WPDx data) in cells. MaxEnt is said to match or outperform other species models (Elith et al., 2006) (Tognelli, Roig-Juñent, Marvaldi, Flores, & Lobo, 2009) and software is freely available[2]. However, as Botella et al. (2018) proved that NNs can outperform MaxEnt in species modeling settings and because NNs are known to find solutions for very complex non-linear problems (Kingma & Ba, 2014), (Botella et al., 2018), e.g. in Heijden et al. (2021), it was decided to build several NN models as well. These are inspired on the MaxEnt technique but are expected to outperform MaxEnt as NN allows for the modeling of more complex nonlinear behaviour. On top of that, two major improvements to the MaxEnt technique are made, namely the inclusion of the total number of expected presences, allowing for predictions of the absolute number of water access point presences in a cell instead of MaxEnts relative probability output and secondly,

[2] https://biodiversityinformatics.amnh.org/open_source/maxent/

the introduction of two evaluative parameters that capture the predictive power of the models much better than the for MaxEnt (and similar) commonly used AUC operator.

3.2.1. CLASSIFYING CELLS WITH NEURAL NETWORK MODELS (M1-M4, M7)

The first model discussed is a classifier model. They are referred to as classifiers as they were originally designed to learn how to assign a class label to examples from a problem domain. An easy to understand example is classifying emails as “spam” or “not spam” (Brownlee, 2020). The input of such a model would be characteristics of the email that act as predictive features for either a "spam" or a "no spam" classification. The primary output of such models is the probability of an email containing "spam" or "no spam" and once this probability exceeds a certain threshold θ, the model classifies it as being "spam" or "no spam". In our case, at first instance we want the probability of finding a water access point (e.g. a borehole) and are therefore not interested in the actual binary classification (a borehole or no borehole), but in the predicted probabilities where this classification is based upon. The models are build per access type such that there is a model for boreholes, a model for piped water access, a model for protected springs, etc.

The models use the feature data, which all have nationwide coverage, to predict the nationwide presence of water access points (which do not have nationwide coverage). The models do so by finding the (feature) characteristics that belong to a presence location (e.g. rainfall allows for rainwater harvesting). Mathematically, the objective is to model the probability distribution of presence locations conditional on the features. This is denoted by:

$$p(y=1|X) = \frac{p(X, y=1)}{p(X)} \tag{3.4}$$

Using Bayes' rule this becomes:

$$p(y=1|X) = \frac{p(y=1) \cdot p(X|y=1)}{p(X)} \tag{3.5}$$

In which $p(X|y=1)$ is the probability density function of features at presence locations, and $p(X)$ the probability density function of features across the whole study area (Elith et al., 2010). Without reliable presence-absence data (see Section 2.1.1), the probability of a randomly selected cell to contain a presence ($p(y=1)$), is unknown (Li, Guo, & Elkan, 2011). For lack of presence-absence data, species modelers often model presence-background data. For such models, the features at presence locations are compared to the features at randomly selected background locations. These background locations are treated as absence locations (y=0) but in fact this is uncertain as $p(y=1)$ is unknown and thus some background locations are in fact unregistered presence locations. This method does however allow for modeling the ratio:

$$\frac{p(X|y=1)}{p(X)} = p^* \tag{3.6}$$

which is denoted from now as p^* and to which will be referred to as the relative probability. The relative probability differs only from the probability distribution of presence

locations conditional on the features ($p(y = 1|X)$) by the constant $p(y = 1)$ (Li et al., 2011).

To model this relative probability, the model used is a neural network in the form of a Multi-Layer Perceptron (MLP) classifier. This is a so called machine learning model and therefore needs training before it can be applied to predict water point presences. During training, each model uses 70 % of the WPDx data (per access type meaning that for the borehole model, only the borehole data is used). As the WPDx data is presence only data (see Section 2.1.1), these presence locations are compared to a randomly selected background sample which is four times the size of the WPDx data of that access type; i.e. if there are 100 presence locations, 400 background points are selected from all the cells in which no presence (of all access types) is recorded (also called pseudo-absence points). It is acknowledged that this ratio of background and presence locations is somewhat arbitrary which is why a small sensitivity analysis is presented in the Result Chapter. The used ratio was found to perform satisfactory after a quick optimisation, during which it was found that too many background points would result in overfitted data as very few locations get assigned high scores by the model, too little background points would overestimate the number of presences and allows for little distinguishing power of the model. The ideal number of background points is a subject of debate in the species modeling field and really depends of the rareness and circumstances of the species (Sofaer, Hoeting, & Jarnevich, 2019) (VanDerWal, Shoo, Graham, & Williams, 2009). For now, this is considered out of scope and we will continue with the ratio of 1:4, however, this could be optimised in later work.

In the explanation below, these background points are treated as if these locations do not contain presences of the modeled access type but again, this is in fact unknown. During training, the model tries to find the values (or mathematical transformations of these values) of the feature layers (from Table 3.2) that are positively associated with presence locations and negatively with background locations. More technically, the model trains by minimising the log loss function using the stochastic gradient-based optimizer as described in Kingma and Ba (2014). The most common classifier loss functions are the Area Under the Curve (AUC) function and the log loss function. The latter is chosen because AUC is a relative measure of internal ordering, rather than an absolute measure of the quality of a set of predictions (Danneman & Clauser, 2020). The log-loss reads as

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \ln p_i + \left(1-y_i\right)\ln\left(1-p_i\right)\right] \tag{3.7}$$

in which y is the binary true value (1 for presence, 0 for background), p the prediction probability (between 0 and 1, from Eq. 3.10) and N the total number of samples (Dembla, 2020). Now this function minimises when low probabilities (p) are assigned to cells containing no presence (y=0) and high probabilities to cells containing one or more presence points (y=1). This method is conveniently available through the *Scikit Learn* package in Python (Pedregosa et al., 2011).

To explain the model in more detail, the following example is largely taken from the

Scikit-Learn-Documentation (n.d.). The MLP classifier is visualised in Figure 3.5. The model is trained to predict the probabilities of presence using the aforementioned feature layers $X = x_1, x_2, ..., x_m$ in which m is the number of feature layers. In each hidden layer of the model, each neuron applies a weighted linear summation to the values of the previous layer: $w_1x_1 + w_2x_2 + ... + w_mx_m$ (WX) and after that a non-linear activation function:

$$g(\cdot) : R \longrightarrow R \tag{3.8}$$

Both these non-linear activation functions and the weighted linear summations are changed in order to minimize the loss function using back-propagation. If we have a data set of $(X_1, y1), (X_2, y_2), ..., (X_n, y_n)$ in which X represent the feature layers, n the number of cells and y a presence (y=1) or background location (y=0). Then, if we would have a model with one hidden layer consisting of one neuron the model learns:

$$f(X) = W_2 g\left(W_1^T X + b_1\right) + b_2 \tag{3.9}$$

to the data. In there, $W_1 \in R^m$, corresponding with the number of feature layers (m) and $W_2, b_1, b_2 \in R^1$. $W_{1,2}$ represent the respective layer weights of the input and hidden layer and b the bias added to the hidden and output layer. In reality we have a variation of models with either one or two hidden layers with different amounts of neurons. Lastly, in the output layer, $f(X)$ passes through a logistic activation function of the form:

$$g(z = f(X)) = \frac{1}{(1 + e^{-z})} = p^* \tag{3.10}$$

giving the relative probability estimate of presence (Scikit-Learn-Documentation, n.d.) (Lecun, Bottou, Orr, & Müller, 2000) used in the logloss function (Eq.3.7). Because of the use of presence-only data, the output from Eq. 3.10 is relative to $p(y = 1)$, see Eq. 3.5 & Eq. 3.6 and accompanying explanation.

Lastly, the importance and predictive power of the several feature layers is analysed using Sklearn's `per_mutation_importance` function. This shows per feature layer, its contribution to the score of the loss function (Eq. 3.7).

HYPERPARAMETER OPTIMISATION

Sklearns MLPclassifier has multiple hyperparamaters including the number of hidden layers, an overfitting penalty and a number of options regarding activation functions, see Table 3.4. These hyperparameters are optimised using the Hyperopt package (Bergstra, Yamins, & Cox, 2013). In total, 200 parameter combinations are explored with 5-fold cross-validation. For this, the model training set is split up in five equal sized parts (hence 5-fold), trained on four of these and evaluated on one. This is repeated five times such that in the end every separate part is used once as the evaluation set. This gives information on how the model is expected to perform when used to make predictions on data not used during the training of the model. The best performing model is used to assess the results. The latter is done on a 30% evaluation data set that was kept apart during training (also not included in 5-fold cross-validation).

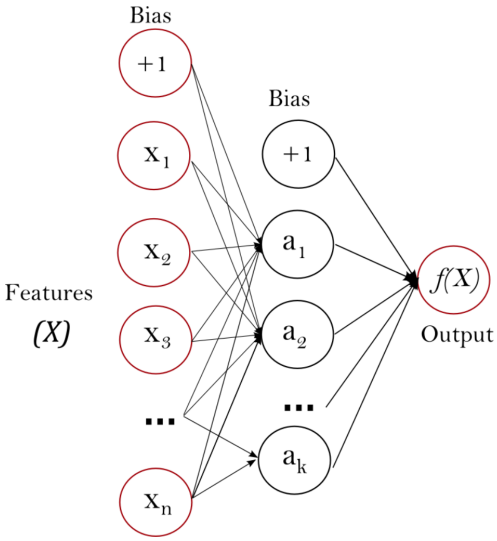


Figure 3.5: A multi-layer perceptron with one hidden layer, $[X_1, X_2, ...X_n]$ represent the feature layers from Table 3.2. The hidden layer is formed by $[a_1, a_2, ..., a_k]$ which individually represent a weighted linear summation to each of the values of the previous layer: $w_1X_1 + w_2X_2 + ... + w_mX_m$ (WX) combined with a non-linear activation function (Eq. 3.8) (Scikit-Learn-Documentation, n.d.).

Parameter	Options
`hidden_layer_sizes`	(5,), (10,), (20,), (30,), (40,), (50,), (60,), (70,), (80,), (90,), (100,), (5,5), (10,10), (20,20), (30,30), (40,40), (50,50), (60,60), (70,70), (80,80), (90,90), (100,100)
`activation`	'identity', 'logistic', 'tanh', 'relu'
`solver`	'adam'
`learning_rate`	'constant', 'invscaling', 'adaptive'
`l1_ratio`	0.0, 1.0
`max_iter`	500, 1000, 2500

Table 3.4: Hyperparameters of Sklearns MLPClassifier and MLPRegression model. During training, 200 combinations of above parameters are tried and evaluated. The best performing combination was used to evaluate the model. Hidden layer sizes of (50,50) should be read as two layers with each fifty nodes.

SCALING THE RELATIVE PROBABILITY TO PREDICT NUMBER OF PRESENCES

To predict the total number of expected presences, the relative probability is scaled using the total number of expected presences (described and presented in Section 3.1.3), such that:

$$P_i = \frac{P_{total}}{\sum_{j=1}^{N} p_j^*} p_i^* \tag{3.11}$$

For instance, to scale the relative probability output of the borehole model, then P_i is the number of boreholes predicted in cell i, P_{total} the estimated total number of boreholes

in the study area (Section 3.1.3) and p_i^* the relative probability of a borehole presence per cell coming from Eq. 3.10. By using the information that some water access types are more present than others, it is possible to describe the variety of water access types in cells and their expected presences across the study area in approximate but absolute terms (not relative). This is an improvement compared to Yu et al. (2019), who stop at the relative probability. At the same time, this improvement suggests a method to deal with one of the biggest limitations of presence-only species modeling (Merow et al., 2013). It would unfortunately be difficult to apply this in biological species settings as estimation of the total presences might not be possible and species occurrence is (at least for animals) a lot more dynamic than static water points.

3.2.2. BACKGROUND SELECTION USING WEIGHTED BACKGROUNDS (M2)

For a second configuration of the MLP classifier model, the presence locations are compared to background locations which are expected to be less likely to contain presence locations. In that way, the background locations do resemble true absence locations more than if they are selected randomly. This is done by distributing the total number of expected sources (see Section 3.1.3) across the study area based on the population density and the share of the population using the different sources (for urban and rural settings separately) as seen in Figure 3.4. Mathematically this means:

$$NP_{i,pr} = P_{total} \cdot \frac{Pop_i}{\sum_{j=1}^{N} Pop_j} \tag{3.12}$$

If again take the example of boreholes is used, P_{total} is the total number of expected borehole presences (Table 3.3), Pop_i is the population size of cell i, the denominator represents the total population of all cells (= the population of Uganda) and finally, $NP_{i,pr}$ is the number of boreholes expected in cell i based on the population density and the share of the population using these sources in urban and rural settings (Figure 3.4). By applying this to all cells and all access types, we get the prior expectation of the distribution of the different sources: areas with less people are expected to have fewer water access points. This expectation of the distribution will be referred to as the background files. Again, note that there is a background file for each access type. Two examples can be seen in Figure 3.6.

In the initial case (M1), background locations were selected randomly from all the cells that contain no WPDx points (again, which could be a result of an area not being included in WPDx related surveys). Now, background locations are selected randomly again but this time the cells are weighted such that cells with a higher weight have a higher probability of being selected. The weights of each cell (W_i) are constructed by dividing 1 by the background file, such that cells that are expected to contain no presences get a higher weight than cells in which we do expect presences:

$$W_i = \frac{1}{NP_{i,pr}} \tag{3.13}$$

Cells that are expected to have zero access points (since people would not live there),

are given the highest weight of the other present cells (instead of ∞). Also this time, the weighted selection happens from the cells of which there is no presence recorded.
A risk with this method is that the model becomes too dependant of the population density. If only populated cells are compared to non-populated cells, the model might turn out to be only a predictor of the presence/absence of people instead of distinguishing cells with or without water access. Too dampen this effect, the population density feature layer is not included in this model. However, it is important to monitor the created bias when assessing the results.

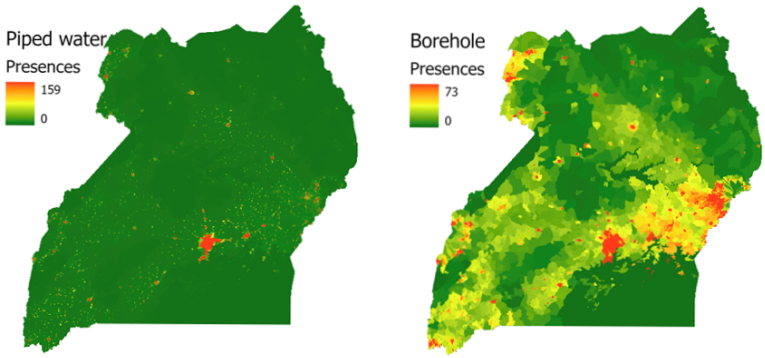


Figure 3.6: Piped water background file (left): the prior expectation of the distribution of piped water access in Uganda. Right: Borehole background file: the prior expectation of the distribution of boreholes in Uganda. Conveying with Figure 3.4, Boreholes are expected to be present in both rural and urban settings, where piped water is solely present in urban areas.

3.2.3. SEPARATE MODELS FOR URBAN AND RURAL AREAS (M3)

For a third configuration of the MLP classifier, a division is made of rural and urban areas. The reason for this is that the 2016 Uganda DHS showed that in general, people in urban areas have better and safer water access than people living in rural areas (Uganda Bureau of Statistics, 2018). This is something that is seen in many countries in the global South (JMP & WHO, n.d.). For this configuration the WPDx - and background data is split into urban and rural data. For each access type, a separate MLP classifier model is trained and optimized on the urban and rural data separately, resulting in two models for each access type. Their performance is assessed both separately, to see if urban or rural models perform better separately, but also together to compare it to the output of the other models.

3.2.4. TRAINING ON HALF THE COUNTRY (M4)

For a final MLP classifier setup, the model is trained on the Southern half of the country and evaluated on the Northern half. This will give insight into the scalability to other countries and the importance of geographical locations. To do so, the presence locations from the North are excluded from the model training set. Naturally, also no background locations are selected from the Northern cells.

3.2.5. MAXIMUM ENTROPY MODEL (M7)

Mainly to validate results from Yu et al. (2019), their Maximum Entropy method is included as well. For the exact method behind this we refer to their work. But for a quick understanding, it is sufficient to know that it is very comparable with the MLP Classifier method except that it minimizes a slightly different loss function and has less options with regards to the possible mathematical transformations of its regression coefficients. One important difference with Yu et al. (2019), is that this time, for fairness of comparison, the number of background locations is set to four times the number of presences (like in the other models) instead of the default 10.000 background locations in MaxEnt (which is what Yu et al. (2019) used).

3.2.6. ACCESS DENSITY REGRESSION NEURAL NETWORK MODEL (M5)

Instead of solely assessing the presence or absence of water access points, it is also possible to model the expected density of water access points in a cell. With this the number of access points that is expected in a cell is modeled directly. This is similar to the final step of the classifying model described in Section 3.2.1, where the probability multiplied was scaled with the total number of expected presences, but this time a regression is applied to the density WPDx data. For this density data, for every presence cell also the number of presences is recorded. So if there are 4 boreholes recorded for WPDx in a cell, the density data-point for that cell is 4. The model thus directly predicts the number of presences expected in a cell without using the total number of expected presences (from Section 3.1.3). Naturally, the latter can be used to evaluate the models performance by controlling whether or not the two models come to similar amounts. To prevent the model from assigning presence to each cell, also for this model background cells are added in the same way as described in Section 3.2.1.

The model is again a MLP but this time the loss function is defined as the squared error which can be expressed as

$$R^2 = (1 - \frac{u}{v}) \tag{3.14}$$

in which u is the residual sum of squares and v is the total sum of squares

$$u = \sum (y_{true} - y_{pred})^2 \tag{3.15}$$

$$v = \sum (y_{true} - \bar{y}_{true})^2 \tag{3.16}$$

in which $\bar{y}_{true}$ is the mean of the true values (Scikit-Learn-Documentation, n.d.). For this model too, the Hyperopt package (Bergstra et al., 2013) is used to optimize hyperparameters that are the same as with the MLP classifier (Table 3.4).

3.3. PREDICTING TRAVEL TIME

The literature consensus is that water consumption is primarily related to travel time. On top of that, people with a total travel time over 30 minutes to and from their water source are considered to have unimproved water access (JMP & WHO, n.d.). To assess water access levels in Uganda further and to potentially model water consumption, the

aim now is to model the households travel time. As the Ugandan DHS of 2016 reports on travel time, it is only logical to perform a regression on this directly with the feature data. To do so, a similar MLP regression model as the one described in Section 3.2.6, is used. It uses the mean travel time from each cluster as training/validation data (instead of WPDx) and the same feature layers as all other models. One difference is that because we are dealing with clustered GPS data, with clusters of radius 2 and 5 km. for urban and rural areas respectively, instead of taking the value for one cell (1 sq. km.), we take the average value of the feature for all the cells that fall within the cluster. A second difference is that no background points are included in this case as the DHS mean travel time already includes travel times of zero minutes and therefore, the DHS travel time data is no presence-only categorical data but numerical. The output of the model is the expected travel time to and from the water source of households across Uganda.

As seen in Table 2.2, logically, water sources that are typically on premises result in lower travel times than sources that are off premises. To research the relationships between the different access types and travel time more in depth, a second configuration of the travel time MLP regression model is made. This has the same setup as described above with the sole difference that instead of using the feature data, it uses the output of the water access prediction models (Section 3.2) to predict travel time. In short, the travel time is then a function of the predicted number of presences of boreholes, protected springs, wells, etc. Mathematically this means that on top of the feature data, which was used to make the water point predictions, the WPDx data is now included as well. With this, the aim is to make (hopefully) a better prediction of the DHS cluster average travel time.

Model	Description
M1 (Std. Classifier)	Classifying model predicting the relative probability of finding one or more presences of the different access types
M2 (Weigh. Backg.)	M1 with weighted background selection
M3 (Urb. Rur. Split)	M1 but with a separate model for urban and rural data
M4 (Half Country)	M1 but trained on half the country and evaluated on the other half
M5 (Density)	Regression on density WPDx data, predicting the number of presences for each access type directly.
M6 (DHS regr.)	Regression on the mean travel time of the DHS clusters
M7 (MaxEnt)	Maximum Entropy model as used in Yu et al. (2019)

Table 3.5: Overview of the several models

3.4. WATER CONSUMPTION

As mentioned before, the literature consensus is that water consumption is primarily a function of travel time (WELL, 1998). However, we hypothesize that it is a function of many things including socio-economic characteristics and water access types and levels. Many of those are researched in the survey described in Chapter 4. From a modeling perspective, the hypothesis is that cells for which the model predicts a high num-

ber of access points, are likely to averagely consume more water than cells with a low number water access points. The number of predicted access points is obtained by the summation of the output for all access types from either M1, M2, M3 or M4. It is acknowledged that a similar comparison could also be made with the number of predicted access points and how many persons they can serve (using Table 3.3). However, this did not have a large effect on the results.

3.5. EVALUATING MODEL PERFORMANCE

For the performance of the models and the evaluation of results, three types of models are distinguished. One predicts travel time using a regression on the mean average travel time of the DHS clusters directly (M6). The classifier models predict, for each access type, the distribution of water access points across Uganda and give as initial output (which we will evaluate first) the relative probability of presence of e.g. finding a borehole in a cell (M1-M4, M7). The third type is again a regression and directly predicts per access type the number of presences in a cell. Note that when we are talking about models of the latter two types, each model consists of 8 separate models, one for each access type listed in Table 3.1. Different models require different evaluative parameters. These are explained below. All models are trained on 70% of available data but evaluated only on the remaining 30% test data to prevent overfitting.

TRAVEL TIME PREDICTION FROM DHS (M6)

The regression model that predicts the mean travel time of clusters across Uganda is primarily evaluated using the mean absolute error (MAE) of the travel time prediction. Defined as,

$$\text{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{3.17}$$

in which y are the prediction values, $\hat{y}$ the true values and N the sample size. Besides that, the median absolute error (MDAE) will also be reported as it is less sensitive to outliers. This is defined as

$$\text{MDAE} = \text{Median}(|y_i - \hat{y}_i|) \tag{3.18}$$

THE CLASSIFYING MODELS (M1-M4, M7)

For classifying models and especially in species modeling, the most common evaluative parameter is the Area Under the ROC Curve (AUC) operator. In short, the AUC can be seen as a comparison between the true positive rate (share of predictions exceeding threshold θ correctly) and false positive rate (a presence prediction on an absence location). It measures the ability of the model to distinguish between presence and background locations, but note that this is different from presence absence locations and therefore has different implications with respect to the traditional interpretation of the results (Li et al., 2011) (Phillips et al., 2006). Jiménez-Valverde (2011) argued that if the objective is to predict and evaluate estimations of potential distributions, the AUC is useless. The main reason that it is not used in this thesis is because the AUC places an equal weight on the true positive and false positive rate and therefore different number of background locations can cause similar AUCs but different distributions: when a high number of background points is created by the user (with binary value 0), and

it is compared to a low number of presence points, the model will automatically assign a low value to most cells. This will result in low true positive rates as of most presence cells the prediction score will not exceed the threshold. For the same reason, this results in a low false presence rate. This is vice versa for a low ratio of background locations versus presence locations. As both the true positive rate and the false presence rate are weighted equally, the AUC will not change for the number of background locations. One can imagine however that the number of background locations does influence the distribution of presences, which is exactly the reason to assess the impact of the number of background locations in a sensitivity analysis and to evaluate the models by means of two other parameters.

Therefore, instead of using the AUC, the classifying models in this thesis are evaluated using a presence - and false presence score, which are introduced here. Let l_1 be the number of relative probability predictions (p^*, Eq. 3.10) on presence cells larger than a threshold θ and n the number of presence locations. Then the presence score (PS) is defined as

$$\text{PS} = \frac{l_1}{n} \tag{3.19}$$

Similar for the false presence score (FPS) we have: let l_2 be the number of background cells (no known presence) that get assigned a value larger than the same threshold θ, and s be the number of background cells, then:

$$FPS = \frac{l_2}{s} \tag{3.20}$$

The PS gives information on the ability of the model to predict presence locations given the feature data evidence (which is what we want), the FPS tells us whether or not a high PS score is only there because all cells are assigned large values and therefore gives information on the distinguishing power of the model. As there are now two separate scores (instead of the sole AUC), this allows for a better comparison of the performance of the model where one can not make up for the other, contrary to AUC. Both scores are between 0 and 1. For the PS the highest score is 1 meaning that all presences were correctly predicted. The best score for the FPS would be 0 in a presence absence scenario as this would mean all absence locations received lower scores than the threshold. However in the presence-only case, some presences are expected in the background locations meaning that the perfect FPS in a presence only model is probably not zero.

It is acknowledged that the choice for the threshold might feel arbitrary, and that access types that we know to occur less often should perhaps receive a higher threshold. But, as all the models receive a similar ratio of presence and background locations, a constant threshold allows for proper comparison between models. Furthermore, eventually the **relative** probability output will be scaled with the total number of presences (Section 3.2.1), dealing with the relative probability output problem and not containing any threshold.

PRESENCE DENSITY PREDICTION (M5)

The third model makes a prediction for the number of presences of each access type in each cell. For this modeling method again, the PS and FPS are reported but also the MAE for both presence, background and the combination of presence and background.

4

SURVEY METHODS

The model approach as explained in the previous Chapter, primarily gives a nationwide impression of water access (and potentially travel time and water consumption). However, the output is coarse or aggregated and will not give a lot of information on the often complex dynamics of water access and consumption in communities. In order to gain a more detailed understanding of these dynamics but also to generate more detailed validation data for the model, a survey campaign was held in the Bushenyi-Ishaka municipality in Uganda.

More specifically, the survey researches the relationship between travel time and water consumption from Figure 2.3. Furthermore, the survey looks into the behavioural aspects of households not using the closest water source as their primary water source and why that is. Besides that, a number of questions relate to the households socio-economic status of which some could be linked to water consumption and/or travel time. One great benefit of surveying in this area is that in most cases water is collected and paid for per 20L jerrycan, this gives the respondents a much better understanding of their daily water consumption then in other comparable settings.

This survey campaign is a fruit of the collaboration between TU Delft Water Management Department and the Department of Environmental Management of Makerere University (Kampala). It has the ethics approval of both the Human Research Ethics Committee (HREC) TU Delft and the Makerere University of Social Sciences Research Ethics Committee (MAKSS REC). Please see Appendix E for the approval letters.

4.1. SURVEY DESIGN

The survey consists of five sections. The first gathers information on the households size, location and economic status. The second section is on the daily total water use, the variety of used sources and the total daily travel time to and from these sources. Next, as many households in Bushenyi-Ishaka municipality make use of a variety of water

sources on a daily basis, a distinction is made between the Primary Water Source (PWS), the Closest Water Source (CWS) and the Drinking Water Source (DWS) (see Figure 4.1). The PWS is the source the household takes the most water from for domestic use on a daily basis. On this PWS, information is gathered such as the water volume taken per trip, the travel time but also the perceived quality and purpose. Dependent on whether or not the PWS is different from the CWS and/or DWS, similar questions are asked for the CWS as well as the DWS. In that way the potentially three different sources can be compared properly. In other words: the final three sections of the survey consist of sections for the PWS and, if applicable, the CWS and DWS respectively. Perhaps superfluously, but this means that for all households information is gathered on the PWS, however, as the PWS, CWS and DWS are for many households the same, the latter two sections are sometimes skipped. Most questions are either multiple choice, with the option to select multiple options. Sometimes, respondents are asked to fill in an integer, e.g. the daily travel time to collect water in minutes. The questions were written in English but if needed orally translated into Runyakore, which is the most widely used language in Bushenyi district. If available, the interview was conducted with the self identified household head but if absent, another adult would suffice. Participants were compensated for their time with the symbolic compensation of a bar of soap. Please see Appendix F for the full list of survey questions and optional answers.

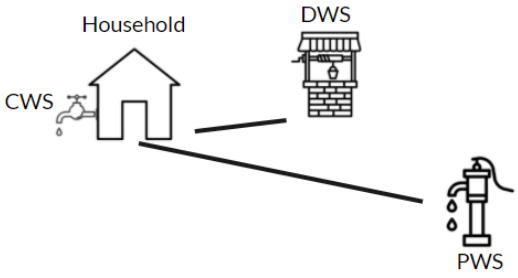


Figure 4.1: Visual of the Primary Water Source (PWS), Closest Water Source (CWS) and Drinking Water Source (DWS).

Cells were selected based on two criteria: Firstly, in discussion with local experts, cells were selected that would contain different water sources, are from both urban and rural areas and would represent households with different socio-economic status and or jobs. Secondly, the area was compared with model output such that both cells that the model would assign higher access and cells with predicted lower access, were selected. This can be seen in Figure 4.2.

4.1.1. SAMPLE SIZE

The population size of Bushenyi district is approximately 250 thousand people (Bushenyi District, 2020). An average household in Bushenyi consists of five persons (Marks et al., 2020), such that the population size of interest is 50 thousand households. In order to

obtain a sample that is large enough to be representative of the Bushenyi district, the often cited formula by Krejcie and Morgan (1970) is used:

$$s = \frac{\chi^2 NP(1-P)}{d^2(N-1) + X^2 P(1-P)} \tag{4.1}$$

With:

χ^2 = value of chi-square for 1 degree of freedom at the desired confidence level (=3.841)
N = the population size (=50 thousand households)
P = the population proportion (=0.5, provides maximum sample size)
d = the degree of accuracy expressed as a proportion (=.05)
s = the required sample size (=380 households)

Such that the minimum representative sample size is 380 households. Based on some test comparisons on mock data to find also significant results between groups and based on earlier survey campaigns in the region (Marks et al., 2020), this was increased to a total sample size of 500 households, which is above the minimum from Krejcie and Morgan (1970) and therefore sufficient.

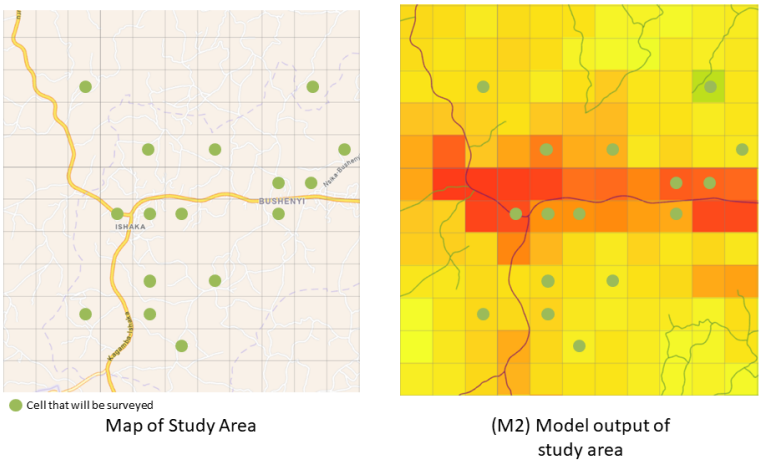


Figure 4.2: The selected cells in Bushenyi Ishaka Municipality on map (left) and model output (right). Each cell is approximately 1 sq. km.

4.2. DATA COLLECTION

The survey was executed by a trained and experienced team of enumerators. All the enumerators also participated in the study by Marks et al. (2020) in 2018 and thus knew the area and its dynamics well. Before interviewing the respondent, the purpose of the survey was explained and written consent was acquired. Respondents were given a unique ID that they can use to have their data removed from the database if desired. This omits the usage of names or other identifiable information. Per selected cell, approximately

30 households were surveyed, such that the total amount of interviews is around 500. Households were approached by going door to door in the selected cells. Data collection took 10 days and took place early December 2021. It is important to note that this is during one of the two Ugandan wet seasons.

The survey tool that was used is *Survey123*[1] which is a product of ArcGis. The enumerators read out the questions and filled in the given answers to limit contact (Covid) and to prevent accidently misuse of the tool by respondents. The data was uploaded instantly upon completion of the interview.

4.3. DATA ANALYSIS

The acquired data is exported to an Excel format and primarily analysed using Python. Some visualisations are made using ArcGis Pro. For comparing means and distributions of travel time and water usage datasets for different groups and access types, the following procedure is followed: First the data is tested to represent a t-distribution using a Kolmogorov–Smirnov (KS) test. If the result indicates that one or both of the datasets ($p<0.1$) do not follow a t-distribution, the datasets are compared by means of a Wilcoxon rank-sum (denoted as w). If both the datasets do follow a t-distribution, the datasets are compared by means of a student t-test. Potential correlations are researched using Pearsons rank order correlation. For comparing proportions of populations, a z-test is performed.

[1] https://survey123.arcgis.com/

5

RESULTS

In this Chapter, the results of both the model and the survey are presented. To do this orderly, a subdivision is made between the results regarding (i) water access, (ii) travel time to the water source and (iii) water consumption. Some results will be presented individually but as the model and survey results are from two very different scales (municipality - gridded country respectively), also comparisons are made between the two methodologies with regards to what information does or does not scale.

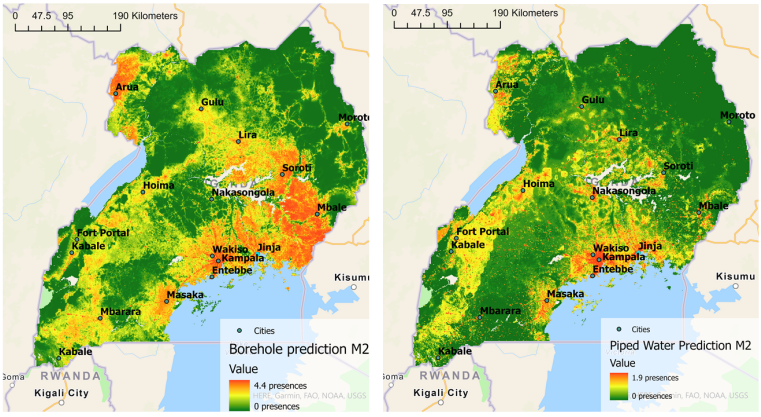


Figure 5.1: Number of presences across Uganda of Boreholes (left) and Piped Water (right) as predicted by the M2 model. Initial relative probability model output was scaled using Eq. 3.11 to represent predicted number of presences.

Before all, it is important to understand the primary output of the classifying models (M1-M4, M7). Examples of the scaled output of the M2 model can be seen in Figure 5.1. In there we can see a higher presence of both access types in cities but, where boreholes are also predicted to be present in rural areas, piped access is not. For each of the access

types from Table 3.1 and for each of the classifying models from Table 3.5, this output is generated. Before any comparison between model and survey results can be made, it is necessary to know how well the model performed on our predefined evaluative parameters. This is presented at the beginning of this Chapter in Section 5.1.

Finally, the demographics of Bushenyi-Ishaka municipality and the survey respondents are captured in Table 5.1. A field report by the head of the survey team can be found in Appendix C.

Variables	Categories	Percentage
Gender respondent (n = 517)	Female	67.2
	Male	32.8
Household size (n = 517)	1-2	14.1
	3-4	30.6
	5-6	38.8
	>6	16.5
Age respondent (n = 517)	13-17	3.4
	18-30	36.6
	30-50	37.6
	>50	22.5
Marital status (n = 517)	Married	1.6
	Single	20.5
	Widowed	10.9
	Separated/Divorced	2.6
	Living together	54.9
Education level (n = 517)	No education	10.9
	Primary	31.4
	Secondary	39.6
	Tertiary	18.1
Occupation respondent (n=517)	Agriculture	43.3
	Handicrafts	17.3
	Formal employment	9.9
	Casual employment	8.5
	Unemployed	8.0
	Retired	0.6
Income category (n = 517)	0-100k SHS/month (Cat. 1)	29.8
	100k-200k SHS/month (Cat. 2)	32.7
	200k-500k SHS/month (Cat. 3)	24.4
	500k-1M SHS/month (Cat. 4)	8.9
	>1M SHS/month (Cat. 5)	4.3

Table 5.1: Demographic characteristics of Bushenyi survey respondents.

5.1. MODEL PERFORMANCE

Firstly, there are the classifying models (M1-M4, M7) that initially output the relative probability of finding e.g. a borehole in a cell (Eq. 3.6). An example of the scaled output can be seen in Figure 5.1, here the total expected number of presences are shown instead of the relative probability. For evaluating these kind of models, the presence and false presence scores are used (Equation 3.19 & 3.20) that make use of the relative probability. In Table 5.2, it can be seen that the model with weighted background (M2) performs best on the PS and also the FPS are relatively low. The latter has to be interpreted with some caution as the FPS is computed from the background data which consists mainly of (almost) non populated areas in the case of the weighted background model. These areas are better distinguishable using feature layers such as population growth and increased nightlight. For presence locations there is not such a bias which means that M2 is capable of predicting presence better than all the other models. However, some areas might be receiving higher probabilities then they should purely because they are populated. When M1 is compared to M7, it becomes clear that although the PS of MaxEnt (M7) are relatively high, the FPS show that the model has little distinguishing power, predicting a false presence almost half of the time. This shows that, like Botella et al. (2018) also has shown, NN might indeed be the better option for modeling presence only data and that the sole AUC evaluation presented by Yu et al. (2019) is limited.

	M1		M2		M3		M4		M5		M7	
Access type	PS	FPS	PS	FPS	PS	FPS	PS	FPS	PS	FPS	PS	FPS
Borehole	.74	.34	.84	.26	.68	.26	.21	.09	.49	.17	.89	.56
Piped Water	.97	.24	1	.28	.86	.33	.18	.04	.67	.29	.85	.23
Prot. Well	.93	.28	.94	.21	.88	.26	.01	.03	.71	.17	.86	.25
Prot. Spring	.86	.3	.95	.22	.8	.27	.36	.15	.63	.16	.87	.37
Rainw. Harv.	.71	.24	.88	.23	.68	.29	.11	.09	.47	.14	.74	.42
Surf. Water	.67	.36	.75	.25	.47	.21	.12	.05	.31	.09	.86	.55
Unp. Well	.92	.26	.95	.19	.88	.75	.00	.11	.59	.12	.86	.21
Unp. Spring	.86	.26	.94	.22	.88	.18	.00	.15	.81	.16	.81	.18
Weight. Av.	.78	.3	.89	.24	.72	.27	.23	.11	.67	.17	.85	.46
Average	.83	.29	.91	.23	.77	.32	.12	.09	.59	.16	.84	.35

Table 5.2: Performance of the different models on Presence Score (PS) and False Presence Score (FPS). All scores come from the 30 % test data set that was kept aside from the models training data. The colours can be seen as horizontal colour-bars to compare the performance of the different models per access types. PS and FPS both form separate colour-bars. Green represents a good score relative to the other models for that access type and red represents a bad score. Threshold for presence (θ) was a relative probability of 0.3 or larger, assigned by the model, with exception of the density case where it was 0.5. The weights for the weighted averages come from the number of WPDx points of each access type (Table 3.1). M1 = Std. Classifier, M2 = Weigh. Backg., M3 = Urb. Rur. Split, M4 = Half Country, M5 = Density, M7 = MaxEnt. M6 is not included as it predicts travel time, not water access.

Table 5.2 further shows that splitting the model in an urban and a rural model (M3), does on average not improve the results in most cases when compared to the standard

and weighted background case. In Appendix B.1 Table 8.1 it can be seen that in some cases the urban model is much better than the rural or (vice versa), especially for the surface water and protected shallow well cases. For these access types, the feature conditions for presence apparently are different in urban and rural settings (see also Figure 8.5)

The results for the model trained on half the country (M4) clearly show that if the training and test data set are from completely geographically different locations, important information is not scaled correctly. A possible explanation is that especially categorical data (like in this case groundwater storage) is very different in the North versus the South of Uganda. Note that the PS and FPS scores reported for M4 are from the half it was not trained on (the North).

SENSITIVITY TO THE NUMBER OF BACKGROUND LOCATIONS

In Table 5.3 & 5.4, the output from the M1 model of three access types is shown for different number of background points. As expected, (too) many background points result in both low PS and FPS, showing little predictive power with regards to presence. At the same time, a low number of background points result in many false positives, driving up the FPS, while not significantly improving the PS. This can also be seen in Appendix B.2 (Figure 8.3) showing different distributions of water access in Uganda for different number of background points: there, too many background points result in overfitted data (only presence locations receive high scores) and too little background points allow for little distinguishing power resulting in very widespread presences.

No. background points	N/2		2N		4N		8N		20N	
Access type	PS	FPS	PS	FPS	PS	FPS	PS	FPS	PS	FPS
Borehole	.96	.82	.82	.47	.74	.34	.42	.10	.15	.02
Piped water	.95	.80	.94	.29	.89	.23	1.0	.13	.70	.05
Prot. Spring	.96	.73	.90	.40	.82	.25	.46	.09	.37	.03

Table 5.3: Sensitivity analyses to show the impact of the number of background points on the PS and FPS of M1 model output. Too little background locations result in high PS but also high FPS (unwanted) and too many background locations result in low PS and low FPS. N represents the number of presence locations (e.g. number of WPDx borehole points).

Access type	AUC (N/2)	AUC (2N)	AUC (4N)	AUC (8N)	AUC (20N)
Borehole	.84	.84	.84	.83	.84
Piped water	.90	.99	.94	.99	.98
Prot.Spring	.92	.93	.93	.94	.93

Table 5.4: Sensitivity analyses to show impact of the number of background locations on the AUC. Contradictory to the PS and FPS, AUC is rather independent to the number of background locations. N represents the number of presence locations (e.g. number of WPDx borehole points).

In Table 5.4, it can be seen that despite changing PS and FPS, the AUC remains the same. This is misleading, especially when looking at the example output in Appendix B.2 (Figure 8.3), and corresponding PS and FPS. As explained in Section 3.5, this can be contributed to the fact that AUC weighs PS and FPS equally and this allows the scores to make up for each other resulting in constant (high) AUCs. Naturally, to increase PS, changing the threshold for presence to lower values when the PS becomes too low would be an option. However, as the relative probability is scaled eventually using the total number of presences and the threshold does not influence the distribution this is not done here. Besides, keeping the threshold constant for the different access types allows for better comparison.

DHS EVALUATION

For a second, completely independent, nationwide, evaluation of the classifying models, the output of the models is compared to the DHS data. In the DHS, respondents are asked about their primary drinking water source. Using the population density of the DHS clusters, this is scaled to the average number of users per access type in the DHS cells (note that this is the primary drinking water source only, while people are expected to use multiple sources). The average number of users per access type in the cluster is compared to the average number of predicted access type presences for the DHS clusters by the models M1-M3. This is done by means of Spearman's Rank order correlation (r_s), and is presented in Table 5.5. Positive correlations are expected as this indicates that higher presence predictions correlate with more people (in absolute terms) reporting to be using the access type in the cluster.

Access Type	$r_s(683)$ M1	$r_s(683)$ M2	$r_s(683)$ M3
Borehole	0.60***	0.36***	0.59***
Piped Water	-0.10**	0.56***	0.06
Prot. Well	0.21***	0.29***	0.25***
Prot. Spring	0.53***	0.48***	0.53***
Rainw. Harv.	0.24***	0.16***	0.19***
Surface Water	0.0	-0.12***	-0.14***
Unp. Well	0.07*	-0.01	-0.03
Unp. Spring	-0.13***	-0.09	-0.12**

Table 5.5: Spearman's Rank Order correlation between (i) the average number of presences predicted for DHS cluster cells by M1-3 and, (ii) the average number of access type users per cell in cluster reported in DHS. Positive correlations are expected when a larger number of predicted presences results in more people using the particular access type in the DHS cluster cells. M1 = Std. Classifier, M2 = Weigh. Backg., M3 = Urb. Rur. Split. *p<.1, **p<.05, ***p<.01

A few observations here: First, it is striking that for M2 and M3, the correlation of unprotected sources (surface water, unprotected wells and unprotected springs (WHO, 2017)) and predicted presences is negative. This can be explained by the expectation that areas with poor access (in which more unprotected source usage is (DHS) reported), will receive low presence predictions for all access types. Secondly, M2 has a much higher r_s for piped water than the other two models, which can be explained by the stronger

focus imposed by the weighted background selection on urban (high population) areas in which more piped water access is expected (as seen in the DHS reported usage in Uganda from Figure 3.4). For boreholes this is the other way around and M1 and M3 outperform M2. As boreholes are present in both urban and rural areas (again Figure 3.4), this could indicate that the M2 model is too focused on the urban areas. Lastly, correlations are generally weak ($r_s = 1$ indicates a perfect association of ranks). This can be a result of various things such as (i) model predictions not directly corresponding with actual usage (no correlation present), (ii) the fact that both the model output and the DHS statistics are averaged over the entire clusters, spanning multiple cells (Figure 3.3) and (iii) that people make use of a multitude of water sources and reporting only on the primary drinking water source is limited (Elliott et al., 2019) and interferes potentially with model results. Table 5.5 is recreated for the Bushenyi data in Appendix B.3 but this is difficult to interpret as unlike with the DHS data, it remains uncertain if the surveyed households are a representative sample for each individual cell.

Lastly, the MLP regression on the mean travel time from the DHS clusters (M6) is presented. Of this regression the absolute errors are shown in Figure 5.2. The model is capable to predict the mean cluster travel time with an error under 10.3 minutes in 50 % of the cases. Given that the average travel time of all clusters is 34 minutes, this prediction is not spectacular but definitely better than an arbitrary guess. Ultimately the aim is to come to a similar or better prediction using the scaled output of one of the classifying models (M1-M4, M7). This will be discussed in Section 5.3

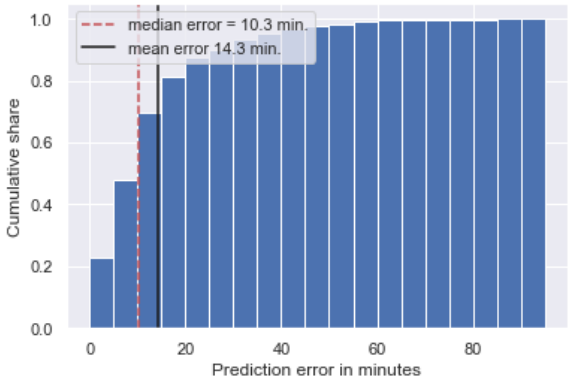


Figure 5.2: Absolute error distribution of regression on DHS clusters mean travel time (M6)

5.2. WATER ACCESS

As seen in Figure 5.1, the models M1-M5, M7 predict, per access type, the number of presences across Uganda. To get an indication of the areas in Uganda that have high access (many presences) or low access (little presences), the results of the several access types are summed (i.e. number of predicted boreholes + springs + ... etc.). For the M2 model this is depicted in Figure 5.3. It becomes clear that the urban areas are expected

to have a lot more water access points per km^2 than the rural areas, some of this extends to the areas around cities. Besides, the dryer (less rainfall, low GW storage) and poorer North has worse access than the region around Lake Kyoga and Kampala, which is generally wetter and richer (see also Section 5.2.1). This is very similar to the results of Nsubuga et al. (2014), who found a similar distribution of water availability per capita (Figure 2.6) as the distribution of water access points shown in Figure 5.3. In both cases, the North East and parts of the South West are under stress. The very South West and North East have a lot of similar characteristics, despite that, the South West is predicted to have a lot more access points. The most distinguishing factor here is the population density which is between 200-300 p/km^2 in the South West and mostly below 100 p/km^2 in the North. But as the population density is not used in the M2 model, this is probably made up for by the ED to roads, which also differs significantly between the two regions.

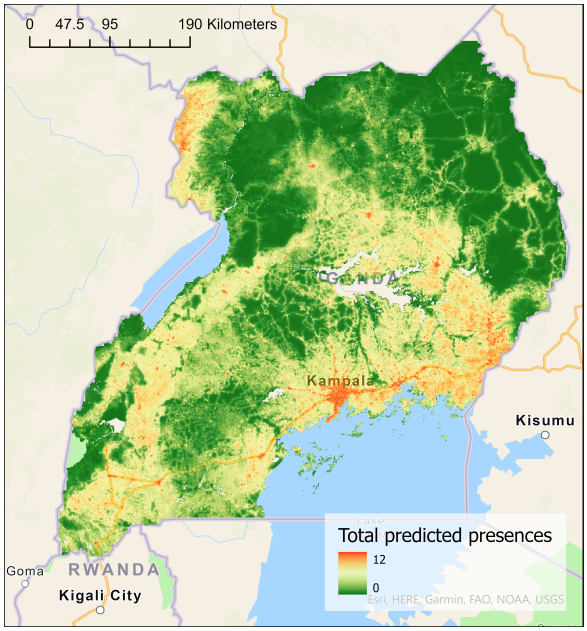


Figure 5.3: Total number of predicted presences across Uganda. This is a summation of the predicted presences of all access types, i.e. boreholes + unprotected shallow wells + protected springs etc. Scaled output from M2: the weighted background model.

In Appendix G the output of the M1 and M2 model is presented per access type. Generally the M2 results show high presences in and around cities, which is as expected given the use of weighted background files for background selection. For the M1 model, the output seems to put too little weight on the cities as for many access types, it does not predict high presences in e.g. Kampala (with the exception of protected springs and rainwater harvesting). In Kampala many piped connections are expected (Haruna, Ejobi, & Kabagambe, 2005), but the M1 model does not show this, the M2 does however. Apparently, adding the background files does improve the prediction of piped and other

access types, this would correspond with Table 5.2 where it was found to perform better also on the PS and FPS. Looking more specifically into the M2 results, it can be seen that where piped water, shallow wells and springs are primarily present in urban areas; boreholes, rainwater harvesting and surface water do expand more into the more rural areas as well. This shows that the bias towards cities of the weighted background models (M2) can be overcome. It also corresponds with Figure 2.5, in which Uganda Bureau of Statistics (2018) reported a larger share of the rural population using boreholes and surface water (also seen in Figure 3.4).

5.2.1. FEATURE IMPORTANCE PER MODEL

In order to properly assess which features can be used to predict the several access types, access in general and potentially travel time and water consumption, the relative contribution of the feature layers to the final result of the models loss functions executed on the 30% test set will be presented. For the travel time DHS regression (M6) this is depicted in Figure 5.4.

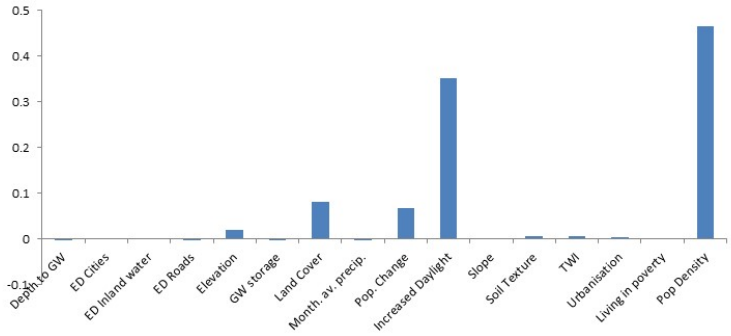


Figure 5.4: Relative contribution of each feature layer to the loss function (Eq. 3.14) of the MLP regression on the mean travel time per cluster (M6). Bars sum to 1.

It becomes clear that for M6, the feature layers related to the presence of people are especially important (population density, increased daylight). Comparing that to the layer importance of the water access models, which for M1 and M2 are depicted in Figure 5.5, results in a much different image. Now, the physical and geographical conditions seem to play a much larger role. Especially the importance of rainfall for predicting almost all access types is interesting and could indicate that wetter areas have better water access. Furthermore, elevation height seems to play a relatively large role in predicting water access as well. This was also seen in the results of Yu et al. (2019) in Kenya. Although describing a direct link is difficult, this could be because higher altitude areas are less inhabited or that the terrain and steepness of higher areas make it more difficult to access water. The proximity of cities and roads also seems to relate with the presence of most access types.

When comparing the feature importance for the urban versus the rural models (M3),

which can be seen in Appendix B.4 (Figure 8.5), it is striking that for the rural models, the population density seems to have a lot more predictive power than in the urban case. This could be explained by the fact that in urban areas, population density will not be the differentiating factor, i.e. the population density will be similar in all urban cells. In rural areas this is not the case as the spatial heterogeneity of population density is generally higher in rural areas.

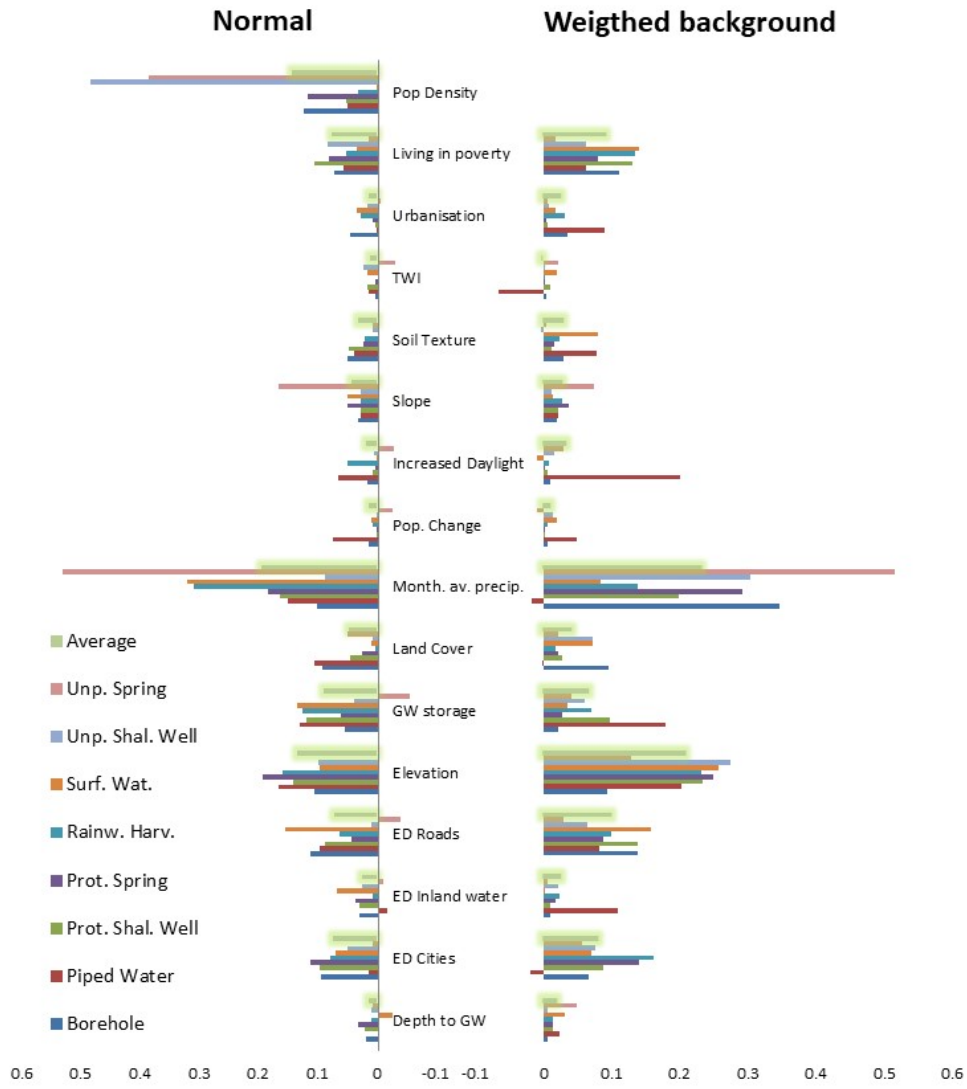


Figure 5.5: Relative contribution of each layer to the loss function (Eq. 3.7) for the MLP classifier with (right, M2) and without (left, M1) weighed background selection, per access type. Bars per access type sum to 1.

5.2.2. ACCESS IN BUSHENYI

Table 5.6 & 5.7, show the shares of the Bushenyi respondents that said to have access to the listed access types. Generally speaking, slightly over half of the surveyed households have access to a piped connection, a rainwater harvesting system or both. Springs and shallow wells are also often used still. When comparing it across income categories (Table 5.6), it becomes clear that the lower income category is significantly less likely to have a piped connection, where the higher income categories have a piped connection significantly more often. A similar pattern can be seen for less educated respondents compared to respondents who finished tertiary education (Table 5.6). Comparing rainwater harvesting of income Cat. 1 to Cat. 2, it can be seen that the lower income category is significantly less likely to make use of this technology where Cat. 2 makes significantly more use of it. This could mean that the lowest income category simply do not have the means to invest in such a system which coincides with Baguma and Loiskandl (2010), who argued for subsidies to increase rainwater harvesting adaptation. As a final observation we note that the group without education, makes more use of springs than all the other groups. This can be a source with safe access but is (almost always) a source that is off premises, making water collection more time consuming.

Access Type	General access	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
Piped	0.56	0.49*	0.56	0.57	0.67*	0.68
Borehole	0.03	0.01	0.04	0.02	0.02	0.09*
Shallow Wells	0.28	0.3	0.23	0.3	0.28	0.32
Springs	0.43	0.38	0.52**	0.44	0.37	0.27*
Surface Water	0.09	0.05*	0.12*	0.1	0.07	0.05
Rainwater	0.51	0.43**	0.59**	0.52	0.43	0.59
Other	0.04	0.06	0.02*	0.03	0.11**	0.05

Table 5.6: Share of households that said to have access to and to regularly use the listed sources (second column). The right columns represent the same but subdivided into income categories. Significantly larger or smaller shares as compared to the total group (left column) are indicated with *p<.1, **p<.05, ***p<0.01.

Access Type	General access	No ed.	Prim.	Sec.	Tert.	Listed as PWS.
Piped	0.56	0.45*	0.42***	0.6	0.74***	0.45
Borehole	0.03	0.02	0.03	0.03	0.03	0.02
Shallow Wells	0.28	0.29	0.31	0.26	0.24	0.14
Springs	0.43	0.6***	0.44	0.4	0.4	0.26
Surface Water	0.09	0.05	0.1	0.09	0.06	0.04
Rainwater	0.51	0.55	0.44*	0.54	0.53	0.07
Other	0.04	0.05	0.06	0.03	0.02	0.03

Table 5.7: Similar as Table 5.6, but this time comparing the educational level to the general access. The most right column displays the share of households that indicated the listed access type as Primary Water Source (PWS).

Of the surveyed households in Bushenyi, 75% says to make use of a protected or from contamination monitored source for their daily drinking water supply. Of the group using a protected drinking water source, 95% treats the water nonetheless with boiling the predominant method (95 %). In the unprotected source user group, 8% reported to not treat the water. If we analyse this further, it turns out that the higher educated groups are more likely to have safe water access. Comparing the group with no or a primary education (n=218) of which 68% has protected access, to the group with secondary education (n = 205, prot. access = 78%), results in z(421) = 2.4, (p<.01). Similarly, the tertiary group (n=94, prot. access = 85%) has significantly more protected access than the secondary group (z(197)=1.3, (p<.1)). This coincides with results of Armah et al. (2018), who also showed that higher educated households have safe water access more often. Although further research would be needed to confirm this, there are two possible explanations: first that people that have had more education, have learned more about the dangers of untreated water (see below). But secondly, other research has suggested that people with bad access (and especially girls who are often responsible) have to travel further and thus have less time for school and drop out sooner (WASH, 2014), sadly making it a self-perpetuating process.

Similarly, the higher income groups in Bushenyi make significantly more use of protected drinking water sources than the lower income categories, increasing from 65% in the lowest category to 91% of the upper two categories (>500k SHS/month). This coincides with findings Armah et al. (2018) but also in Mahama et al. (2014), Adams et al. (2015) and more. In Bushenyi it can be explained by the fact that the most predominant available safe source: piped water access, is at the same time the only paid source, which is often too expensive for lower income categories.

MODEL PERFORMANCE IN BUSHENYI VARIES PER ACCESS TYPE

Table 5.8 shows how capable the model is of distinguishing cells in which usage of the listed access type was reported from cells in which this was not reported. Here it is expected that the classifying models (such as M1) predict many borehole presences in cells comprising households that report to use boreholes, and little borehole presences for cells in which this was not reported. In total, 63 cells were surveyed, subsequently, households located in the same cell, receive the same score. The latter could create a bias towards the more surveyed cells, this effect is mitigated however by the fact that per cell the variation in types of water sources is often high.

The models for piped water access, rainwater harvesting and shallow wells perform rather well, assigning significantly higher predictions to cells where these access types are used. This shows that in Bushenyi features are distinguishable enough for the model to predict the (non)-presence of these access types. It is noticeable that these three access types are used by a large share of the households too, probably providing large enough samples to make a significant comparison. More visually, in Figure 5.6, the M1 piped water model scores and the reported usage of piped water is depicted, showing visual proof that in general households having less or no access to piped water, indeed lie in cells that get assigned low scores by the model. Similar figures for the other access

types can be found in Appendix H.

For boreholes and surface water access, the model seems to underperform but it is not possible to draw statistical conclusions. For springs however, the model seems off, assigning significantly lower predictions to cells in which households report making use of springs. This could mean the model is off or can be explained by the fact that people might travel outside their cell to a spring. The latter would make sense as Marks et al. (2020), the survey results and local team reported a preference for springs for cultural traditional, taste and cost reasons. It is also striking to see the low number of presences predicted by the rainwater harvesting model, especially when compared to the high number of households reporting using it. This can be explained by two factors, firstly the total expected presence could be off (Section 3.1.3) but it is also very much possible that this is a result of the fact that surveying happened during the wet season in which an abundance of rain is available (and that the general critique on water access surveys (such as DHS) is that they happen during dry seasons (Elliott et al., 2019)).

In Section 5.2.1 and Figure 5.5, it was shown that poverty, precipitation and elevation are important drivers of water access on the nationwide scale. Likewise in Bushenyi, we find that the North East region has a higher altitude and a larger share of households with lower incomes and at the same time, worse water access i.e. less protected water sources and less piped connections (see also Figure 5.6). The poverty aspect came back in Table 5.6 as well, where it was shown that the households with lower incomes have worse and unsafe water access more often. This is an indication that such aspects properly scale. The precipitation in and around Bushenyi-Ishaka has logically very little heterogeneity and therefore no further analysis is performed.

Acc. Type	HHs using	Avg. mod. pr.	HHs not using	Avg. mod. pr.	w	p
Piped	288	2.53	229	1.65	4.72	0.00
Rainwater	263	0.15	254	0.14	2.61	0.01
Borehole	15	1.00	502	1.06	-0.89	0.37
Shallow Wells	143	1.40	374	1.18	3.42	0.00
Springs	224	1.08	293	1.15	-1.85	0.06
Surface Water	44	0.88	473	0.99	-1.25	0.21

Table 5.8: Comparison of average predicted presences by M1 between Bushenyi households reporting using or not using the listed access type. Note that the model scores are from a gridded 1 sq. km. output meaning that multiple households get assigned the same prediction. In total there are 63 different model cells surveyed, resulting per access type in a total of 63 possible model scores of the households.

THE COMPLEXITY OF PIPED WATER ACCESS

A lot of households that listed a piped access connection as their PWS, have only recently been connected to the network. As part of a National Water and Sewerage Corporation (NWSC) campaign, over half of the households have only been connected for three years or less. Besides, out of the group of 290 households, despite having access to piped water, 57 (20%) do not use it as their PWS. The main reason being that the piped connection is

too expensive. Similarly, for drinking water, 85 (29%) households with piped access, said not to use the closest water source as drinking water source; deeming it unsafe (47), expensive (22) and having bad taste (20) as the most listed reasons. The enumerator team also recalled many respondents saying they have a preference for other DWS as the piped water is chlorinated (bad taste), but also because especially people that have only recently been connected to the piped grid, still are traditionally very accustomed to get their water at springs (also reported by Marks et al. (2020)). The enumerator team had also noted that people in urbanised areas are more comfortable with drinking piped water as they have had access for a longer time but also because it is often the only source available. Subsequently, the survey results indeed showed that the urban areas have a piped connection more often and that especially the hilly rural areas in the North-East of Ishaka have no piped water access as these are hard to connect (see Figure 5.6).

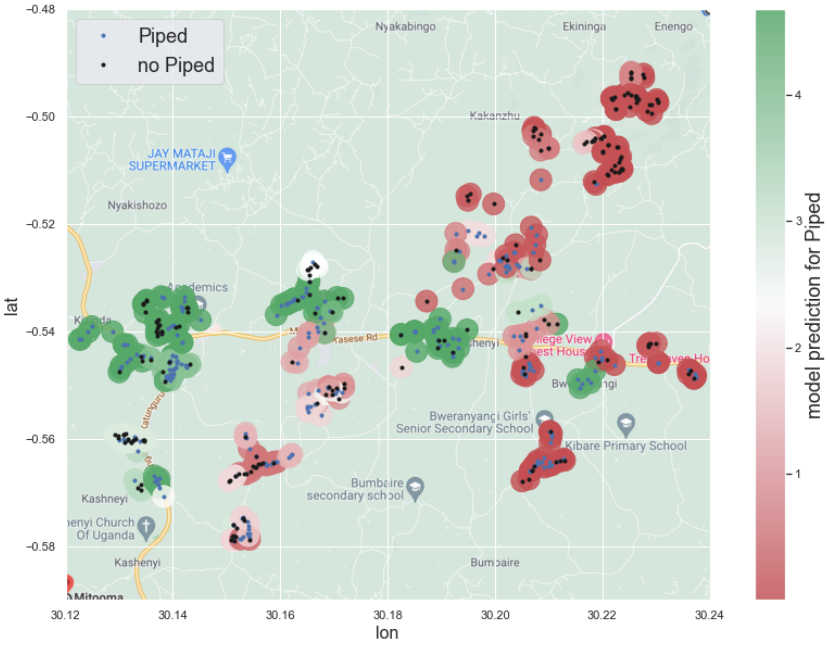


Figure 5.6: The surveyed households with (blue) and without (black) piped water access and the number of piped access points predicted by M1 for the cell corresponding to the household (red for low and green for high predictions). Background map is from Google Maps.

THE RISK OF BAD ACCESS

Even though almost all respondents said to treat their water before drinking, 154 households admitted drinking untreated water at least once in the past 4 weeks. This number should probably be higher as the team felt that some respondents gave what they felt were the desired answers. Out of 517, 23 households (4%) reported having at least one member suffering from diarrhea in the last 30 days. Similarly, 97 (19%) suffered from respiratory illnesses. The results show that drinking from an unprotected source does

not (significantly) increase the risk to contract these diseases (similar to Hubbard et al. (2020)). However, drinking untreated water, doubles the risk of diarrhea: 8% of households drinking untreated water at least once had diarrhea (z=2.0, p=.02). It also increases the chance for respiratory illnesses by 10% (z=2.7, p=.004). In a general research among middle and low income countries the WHO (2014) also found that *consistent* and effective implementation of water treatment results in 28-45% less diarrhea cases and that the impact of using an improved source is much lower (11-16% less diarrhea).

5.3. TRAVEL TIME

To research the relationships between water access and travel time to the water source, the scaled output of the classifying models (number of presences) is compared to both the DHS cluster average travel time (coarse but nationwide) and the reported travel times in Bushenyi (detailed but local). This is done in two ways. First by applying a regression of the output of the classifying models to the DHS cluster average travel time (as was explained in Section 3.3). For a second, coarser comparison, it is researched if people living in areas that are assigned a lot of access points by the model, have low travel times. The first is only done for the DHS cluster average travel time but the latter is also done for the high detailed Bushenyi survey data. By analysing the Bushenyi data further, different purposes of different water sources and reasons for using a water source that is further away than the one closest are highlighted as well.

5.3.1. PREDICTING CLUSTERED AVERAGE TRAVEL TIME FROM DHS DATA

After fitting the number of predicted presences of the different access types (such as Figure 5.1) to the DHS cluster average travel time using both a MLP-regression fit as well as an ordinary Least Square (linear) fit, it was found that this does not improve predictions using feature data only. The more detailed results of this can be seen in Appendix B.5. Still, although a regression with the access type model output does not seem to improve travel time prediction, there does seem to be some sort of relation between the number of access points and travel time. This becomes clear in Figure 5.8. For this Figure, first the total number of predicted presences is calculated, which is, per cell, a summation of the output of the models of all access types, i.e. predicted boreholes + predicted unprotected springs, etc. This is visualised in Figure 5.7 by the colours of the cells. Next, per cluster the average of the total predicted number of presences of the cells that are comprised by the cluster boundaries is taken (see Figure 5.7). This average is shown on the y-axis of the graph in Figure 5.8. Next the DHS-clusters are grouped by their mean travel times into three categories: 0-5 minutes, 5-30 minutes and >30 minutes, following the graph from Figure 2.3 (the relationship between travel time and water consumption).

In general, clusters with lower average travel times get assigned higher number of presences when compared to the higher average travel time clusters. The 69 clusters in the first group (M = 6.53, SD = 1.57) compared to the 254 clusters in the middle group (M = 4.08, SD = 2.00) demonstrated significantly higher number of presences, w(311) = 8.69, p<<.0001. Similarly, comparing the middle group to the group with 361 clusters having average travel times over 30 minutes (M = 2.96, SD = 1.65) gives also significant differ-

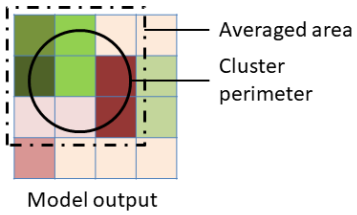


Figure 5.7: To compare model output to DHS cluster average travel times, the average of the total number of presences is taken of cells that fall (partly) within the cluster perimeter. The colours in the example above represent the number of presences in which a redder colour indicates high and green low numbers of water access points.

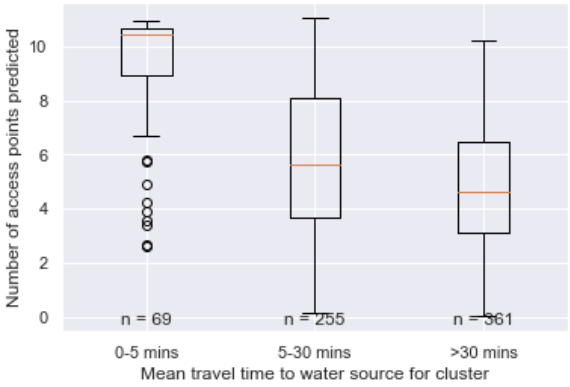


Figure 5.8: Distribution of the mean number of access points predicted for cells in the DHS clusters. Clusters are grouped by their average travel time shown in the x-labels. Predictions come from M2: classifying with weighted background selection.

ences, t(613) = 7.64, p<<.0001. This shows that the level of access points is likely related to travel time. This may feel slightly contradictory to the regression results from previous section (which can be seen in Appendix B.5), but it merely shows that the regression is not capturing the relationship properly yet. On top of that, even though the means differ significantly, the variance of the three groups remain rather large, which makes a regression all the more difficult if possible at all. What is all the more exciting is that with Figure 5.8, it is possible to make a prediction for the travel time based on the expected number of presences of water access points. E.g. a cluster with 10 predicted presences probably falls in the 0-5 minute category while a cluster scoring 6, likely falls in the 5-30 minute category.

5.3.2. PREDICTING TRAVEL TIME IN BUSHENYI

As a second comparison of model output to travel time, the model output is compared to the Bushenyi survey data. As this data is a lot more specific, there is no longer the need to take averages of the travel time for each cluster, for this time we have that informa-

tion for each individual household, including its location. For a first comparison, Figure 5.8 is recreated in Figure 5.9 for Bushenyi but this time showing the daily travel time of each interviewed household on the x axis (instead of the DHS cluster average travel time). Firstly, it can be seen that the model expects cells in the Bushenyi area to contain relatively high numbers of water access points when compared to the DHS clusters, as all households receive predictions between 4.7 and 11.7 presences. Secondly, although again a downward trend for number of access vs. travel time seems visible again, the results show less significance and have a lower effect size this time: $w(363) = 2.54$, $p=0.011$ for 0-5 minutes compared to the middle group and $t(373)=2.99$, $p=0.003$ for the middle compared to the >30 minutes group. Besides, although parts of the distributions overlap, Figure 5.9 shows a generally larger distribution with a higher mean than Figure 5.8. This is probably a result from two things: firstly the travel times in the DHS cluster are averaged, filtering out the large variation of travel times within the cluster, this is present in the survey data. Secondly, with the DHS data the model scores are averaged over the entire cluster as well, creating again a dampening effect as extreme values will be averaged with more normal results. As the survey results do allow for differences within specific cells, this does not happen here.

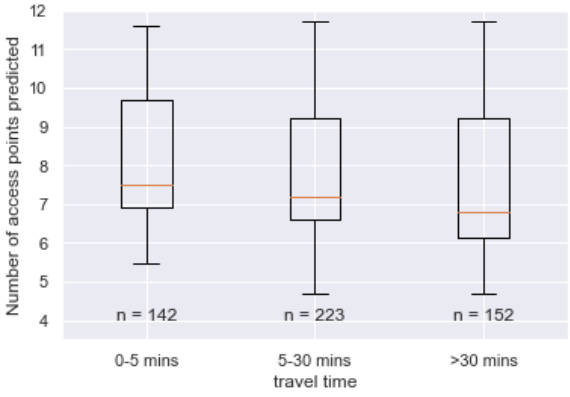


Figure 5.9: Distribution of the number of access points predicted for cells in which Bushenyi interviewed households lay. Households are grouped by their average travel time shown in the x-labels. Predictions come from M2: classifying with weighted background selection.

BUSHENYI COMPARED TO THE NATION

To compare the travel time per access type in Bushenyi to the nationwide tendency captured by DHS, in Appendix B.6, Figure 8.1 is recreated for Bushenyi in Figure 8.8. It becomes clear that again distinction can be made between sources on and off premises, which has a direct effect on travel time. Comparing the on and off plot distributions by means of a t-test gives very significant results ($p<0.0001$). The distribution of the travel times in Bushenyi per access type is very similar to the results from the DHS data.

5.3.3. HOUSEHOLDS USE MORE THAN ONE WATER SOURCE

The households in Bushenyi reported to be using an average of 1.9 sources for their general water supply. From the respondents, 66 out of 517 (13%) reported that their closest water source is not their primary water source. This is a decline from 2018 when Marks et al. (2020) reported that one in four households traveled further. This decline is probably a result of the recent increase in piped connections in the area, with 50% of the users with PWS piped water, have only been connected for three years or less. The reason for not using the closest source is mostly the price (43 times), this is a result of piped connection being paid and other sources free. Eight users reported the CWS to be unsafe. The travel time to the CWS is 4.6 minutes on average which is significantly less then the 20 min average travel time to and from the PWS ($w(128)=7.44$, $p<<0.0001$). This means that people are willing to travel an average of 15 minutes longer for their daily water supply to avoid (mostly) costs. When asked to score the reliability of their CWS and PWS between one and ten, respondents gave the PWS an average of 8.3 and the CWS a 7.5 ($w(128)=2.07$, $p=0.04$). With regards to both physical and health safety and cleanliness, the PWS and CWS received very similar scores (all around 8).

DIFFERENT PURPOSES FOR DIFFERENT WATER SOURCES

Table 5.9 displays the variety of purposes of the PWS, the CWS and DWS. It becomes clear that the PWS is used for all listed purposes in most of the cases. For drinking water however 10% uses another source. In 44% of the cases the CWS is not used at all. If people have a DWS that is different from both the CWS and PWS, it is really only used to drink in most of the cases.

Purpose	PWS (N=517)	CWS (N=66)	DWS (N=28)
Drinking	0.90	0.42	1.00
Cooking	0.99	0.45	0.07
Washing Clothes	0.99	0.47	0.04
Cleaning	0.99	0.44	0.04
Washing Hands	0.99	0.42	0.04
Bathing	0.98	0.44	0.04
Not in use	0.00	0.44	0.00
Other	0.05	0.03	0.00

Table 5.9: Share and variety of purposes for the PWS, CWS and DWS (if DWS and/or CWS differ from PWS).

From the respondents, 155 (30%) said to not drink from their CWS. This time, safety (84), cost (38) and a bad taste (24) were the predominant reasons. The amount of reasons for using a multitude of water sources on a regular basis was found to be larger than the by Elliott et al. (2019) reported aesthetic, cost and seasonality ones. Seasonality clearly plays a role as half of the PWS rainwater harvesting users reported to do so only in the wet season. Cost clearly does too. But on top of those, reliability, taste and safety are important incentives for using multiple sources as well. Finally, the number of different sources that households use, does not effect the daily travel time to and from the water

source, of which the median remains rather constant around 2.5 minutes per household member.

5.4. WATER CONSUMPTION

As shown in the Literature Review (Chapter 2), the WHO suggested primary indicator for water consumption is the travel time to and from the source (incl. queuing) (WELL, 1998). However, if the hypothesis holds, it should also be related to the number of water access points. The DHS data only gave information about the travel time, additionally, the survey gives information on the water usage. Therefore all results in this section are evaluated using the Bushenyi data. We realise that this might not be representative for the entire country but it is merely to give a feeling of the dynamics of water consumption in mid-sized towns with a mix of urban-rural dynamics such as Bushenyi. The water consumption of the participating households is shown on the map in Figure 5.10. Note that there are often large differences in water consumption between neighbours. It is also noticeable that at first sight, in both urban and in rural areas there are households using relatively a lot or little water.

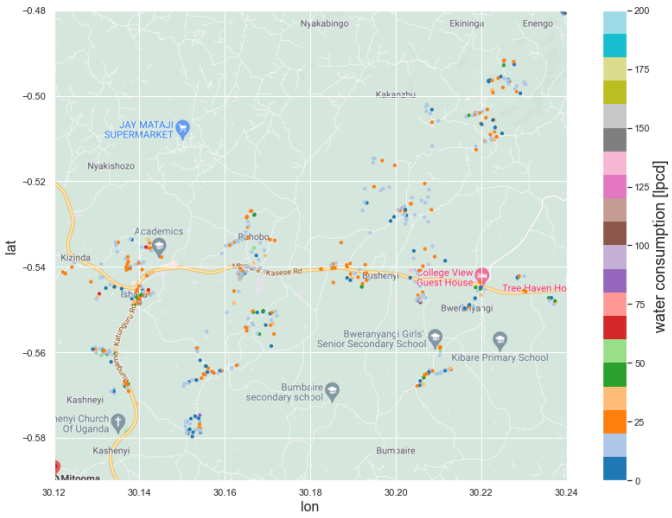


Figure 5.10: The water consumption in lpcd across Bushenyi. Note the often large differences between neighbours. Background map is from Google Maps.

5.4.1. THE ROLE OF WATER ACCESS

In Figure 5.11 it can be seen that in Bushenyi, a variety of PWS-types can be found. Many neighbouring households make use of a similar (the same) source, but this is definitely not always the case, showing that people adhere to some sort of personal preference. As shown earlier (Figure 5.6), the more rural areas are less likely to have access to a piped water connection. In this section the influence on water consumption of (i) the various access types and (ii) the modeled access level is presented.

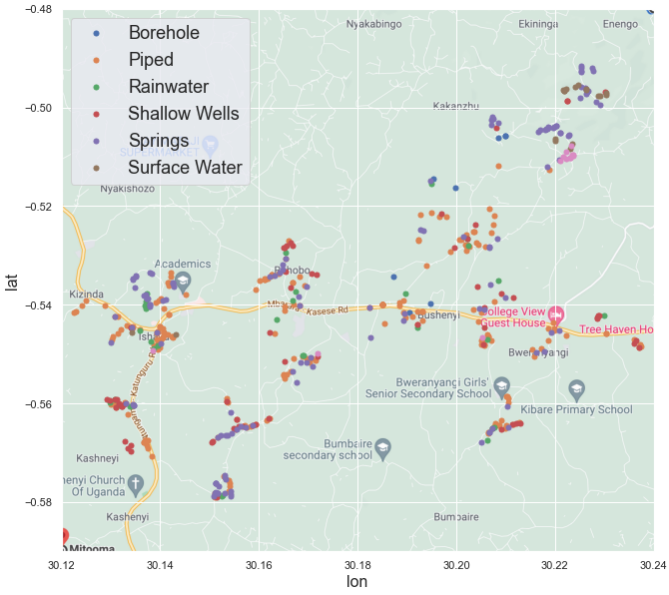


Figure 5.11: The as PWS listed sources for surveyed households. Also here we see variation between neighbours but much less.

THE INFLUENCE OF ACCESS TYPES ON WATER CONSUMPTION

Figure 5.12 shows the distribution of water consumption from the PWS per access type. Similar as with the travel time, there is a distinction between on and off plot sources. People that have their PWS on plot (piped or rainwater) do in general use more water than people with an off plot PWS. More specifically, the 231 households with a piped connection as PWS (M = 40.7 lpcd, SD = 86) compared to the 283 households with another type of PWS (M = 27.7 lpcd, SD = 52.8) demonstrated to be using significantly more water from the PWS, w(512) = 2.55, p=.01). This does not necessarily mean that the total water consumption of the households is higher too as people without a piped water connection might complement the PWS with (multiple) other sources. However, comparing the total average water consumption of the piped-PWS group to the rest, results in the finding that the piped-PWS group uses an average 4 lpcd more than users that did not list piped water as their primary source (w=4.21, p<.001).

BETTER PREDICTED ACCESS, HIGHER WATER CONSUMPTION

Similar as with the travel time, it is possible to look whether or not a high access prediction leads to a higher water consumption. From Figure 5.13 the first impression is that households that are located in cells that the model assigns a lower expected number of presences, tend to use less water, and that there is a slightly upward trend for water consumption vs. the number of predicted access points. However, especially for the least water consuming household, the sample size is low. Furthermore, the variance in the number of predicted access points remains substantial for each group. The latter can be explained again by the large differences between households in each cell.

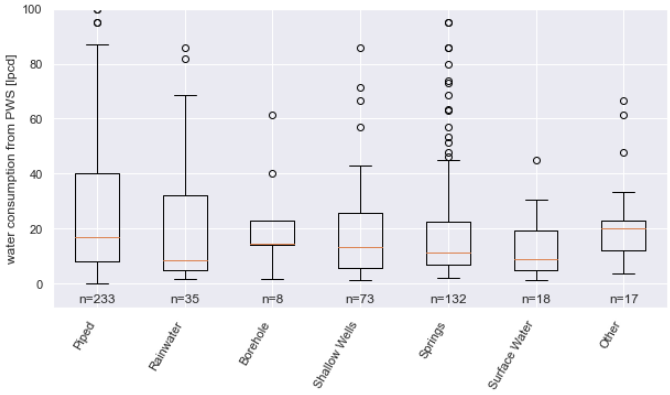


Figure 5.12: Bushenyi: daily water consumption from PWS per household member. Grouped by PWS access type.

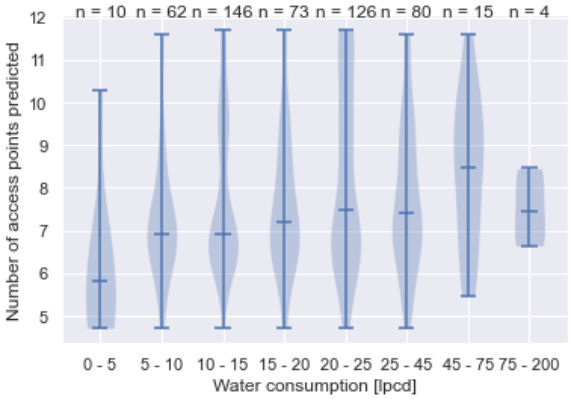


Figure 5.13: Distribution of the number of access points predicted for cells in which interviewed Bushenyi households lay. Households are grouped by their daily water consumption shown in the x-labels. Note that towards the right the water consumption interval sizes become larger to assure large enough sample sizes. Predictions come from M2: classifying with weighted background selection. Blue horizontal lines represent minimum. median and maximum value.

5.4.2. THE ROLE OF TRAVEL TIME

As Figure 5.9 indicates a downward trend for travel time vs. number of predicted water access points (and so did Fig. 5.8 more significantly), and Figure 5.13 an upward trend for water consumption vs. number of predicted water access points, the relation between travel time and water consumption in Bushenyi is explored as well. As we saw in Figure 2.3, it is expected that travel time correlates negatively with water consumption. And combining the results from aforementioned Figures, the same is expected here. However, looking at Figure 5.14, displaying the travel time vs. water consump-

tion in Bushenyi, this is not the case as no trend is visual and very similar distributions are present. This shows that the model output might in fact be a better predictor than the travel time alone.

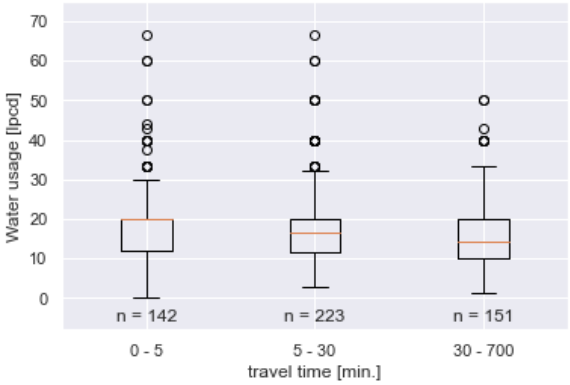


Figure 5.14: Daily water consumption of household compared to their daily travel time in Bushenyi.

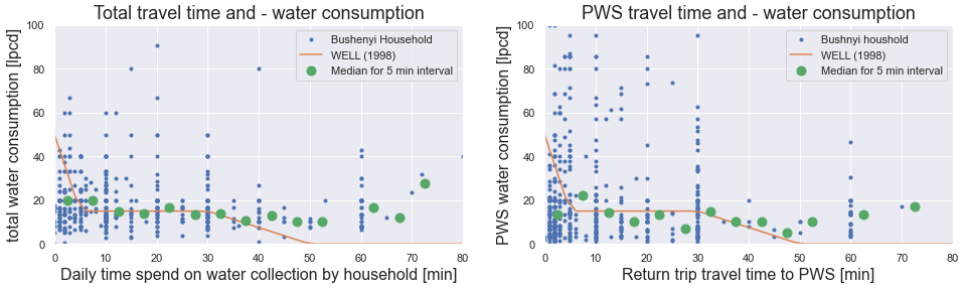


Figure 5.15: Comparing the daily total travel time (left) and the return trip travel time of households in Bushenyi to their related water consumption and the relationship from WELL (1998).

In Figure 5.15, the households travel time is compared more specifically to the WELL graph (from Figure 2.3), that related water consumption to travel time. Median values of the 5 minute interval seem to coincide with the WELL graph but variation remains high. And, even though the households with the highest consumption do have lower travel times, it can be seen that in most cases, the real water consumption from individual households really differs from the WELL graph. Especially in the low travel time zone (under 5 min.), the water consumption is lagging compared to the graph, with a lot of points lying under the graph. Two suggested explanations are that firstly, many of the users with a low travel time will have a piped and thus paid connection, the risk of a high bill will prevent people from using a lot of water. This was also found in Weinan, a Chinese city where most people have a piped connection and for which Liu et al. (2003) found that increasing water prices decreased water consumption. Secondly, piped water

users might have more trouble estimating their total water consumption as they do not have to travel large distances with it but instead have it readily available. These users might be using more than they think. More importantly, the large variation between the WELL graph and the Bushenyi data showed that domestic water consumption is in fact from a lot more dependant than travel time alone.

WATER RESPONSIBILITY

As mentioned before, in Uganda water is considered primarily a women's task (WASH, 2014). This came also forward in the survey results: in 54% of the households, a woman was listed specifically as the one responsible for the (availability) of water (see Table 5.10).

Responsibility	Count (N=517)	Share
Woman	187	36%
Man	91	18%
Girl	91	18%
Boy	102	20%
House Girl/Shamba Boy	30	6%
Relatives	10	2%
Other	6	1%

Table 5.10: Primary responsibility for water in the household in Bushenyi.

The age of the persons responsible is depicted in Figure 5.16. This seems to follow a normal distribution around the age of 20. What is striking however is the large peak in the 14-16 age bar. The local team reported that in the Ugandan culture middle-aged children are indeed often made responsible for water collection, but this time the peak is particularly high as a result of the COVID crisis that kept Ugandan schools closed for two years (Atuhaire, 2022).

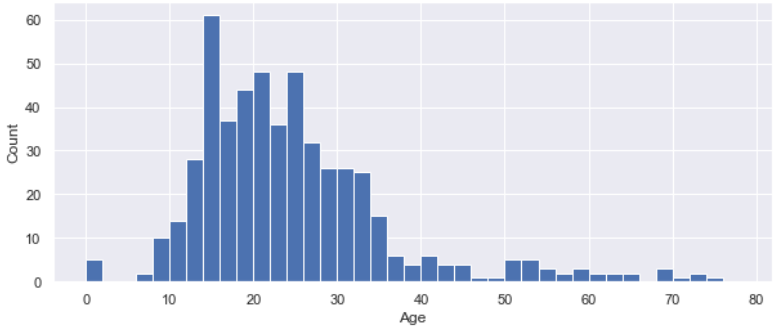


Figure 5.16: Age of person responsible for water in the household (in Bushenyi).

5.4.3. RICHER AND WELL EDUCATED HOUSEHOLDS USE MORE WATER

Table 5.6 & 5.7 showed that richer and better educated households more often have piped water access. Similarly, the richer and better educated households consume slightly more water as well. Comparing the lower three income categories (n=449, M=18.7 lpcd, SD=15.5) to the upper two (n=68, M=22.4 lpcd, SD = 11.0) gives w(515)=-4.08, p<0.001. With regards to education, especially the highest educated (n=94, M=25.6 lpcd, SD=25.3) group uses more water when compared to the rest (n=423, M=17.7 lpcd, SD=11.1): w(515)=-4.2, p<.001. But also for comparisons between the different educational levels, mostly significant differences were found (p<.1). Lastly, household size was also related with water consumption and larger households used significantly less water per capita than smaller households (which was not correlated with income).

6

DISCUSSION

This Chapter discusses the meaning of the results in the broader sense, it shows how we improved on existing water access modeling techniques, it highlights the local complex dynamics of water access and consumption and finally gives suggestions on how all of this information can be used to improve water access. But first, it touches upon a number of limitations of the chosen methodology.

6.1. LIMITATIONS AND RECOMMENDATIONS

Uncertainties, assumptions and limitations can be found both in the methodology of the models and in the survey design. For the models this relates primarily to their setup and the data used. For the survey, some of the questions and therefore our interpretation, rely heavily on the respondents capability of estimating their water consumption and travel times. Next to presenting such limitations below, we also suggest some subjects that could be researched further, some of these relate to, but are not limited by, the limitations.

6.1.1. MODELS

A first limitation with respect to the objective of water access modeling, is that we did not include models for packaged water and sand or sub-surface dams. Information on the usage of packaged water (which for simplicity we will consider to be the same as bottled water) in Uganda is hard to find. But, Cordoba and Grabinsky (2020) showed that bottled water in low and middle income countries has increased with 174% between 2004 and 2016. Also for Uganda, we have anecdotal evidence that especially people in (large) cities, often make use of packaged water. It would therefore be desirable to model packaged water access as it could be an important source for many people. Besides, some researchers question JMPs choice to include it as safe water (WHO, 2017) because water quality is often unknown (Cordoba & Grabinsky, 2020). However as vendors are often mobile it can be imagined that the locations of the actual water access point are difficult to pinpoint for the WPDx database. In the Ugandan WPDx database there was only one

packaged water datapoint, as long as this remains the same, modeling packaged water access points with our methodology remains impossible.

A second limitation comes from the used datasets. For the WPDx data, we took the most recent version in June 2021. The WPDx data contains more information than was used, including whether a source was functioning or not at the time it was registered. In the used WPDx data, 79% was functioning, 18% was not and for 2.5% it was unknown. Modeling only the functioning sources would be an option but because some sources might have been fixed (or broke down) since registering, it was decided to model all, allowing for a larger data-set. This does add a limitation as one can not be fully certain that modeled access points are functioning. Further researching the frequency of breakdowns of water access points, potentially using more WPDx information, is advised.

For the feature layers, we always took the most recent data available. These can therefore be from different moments in time than the registration of some of the WPDx data, of which the vast majority was registered in 2010 and 2011. Also the feature layers themselves are not from the exact same moment in time: most are from the period between 2013 and 2021 and thus reasonably coincide with the WPDx recordings and each other. Besides, many feature layers containing data such as elevation, distance to inland water, euclidean distance to city centers and soil texture are logically not expected to have changed a lot over time. However, population density and urbanisation degree probably have changed under continued urbanisation processes. An especially important layer is the precipitation data, which is based on 1970-2000 recordings. Precipitation was found to be an important predictor for water access in this thesis. At the same time, precipitation patterns and quantities have changed since 2000 (Ssentongo et al., 2018) and are expected to continue to change under global warming. This could impact the results, when changed precipitation results in wrong predictions as the model was trained on older datasets. If however, updated precipitation data would be included in the model training process, there is no reason to suspect the model to perform less. As a recommendation for future research we suggest to research the changes of the feature layers over time (both past and future), the influence of seasonality (wet seasons), their relationship with the WPDx data from different moments in time, and with that, the possibility of predicting the presence of water access points (and potentially consumption) into the future.

Thirdly, the influence of the number of background points used for the classifying models should be researched further. The sensitivity analysis showed that changing the number of background points changes the distribution, where too many background points lead to overfitted results with low predictive power, and too little background points lead to distributions that can hardly distinguish areas with from without presence. The correlations between the models predictions on water access and the independent DHS data set reporting on usage were not found to be strong and even negative in some cases. This could be an indication that the chosen number of background points is off. It is suggested to further research the influence of the number of background points and to see if changing this can increase performance. This also really depends on the objective, if the objective is to predict with high precision the cells with presence, many

background points would help with this (but watch out for overfitting) and if a more generic distribution is sought after, lesser background points would be the better option.

Lastly we suggest to research further the number of people that are served per access type. Our approximation from Section 3.1.3 is under two important assumptions namely that the access points are all serving at their limit and that the WPDx data from cells that were surveyed for WPDx is complete, containing all water access points of the cell. Already in the Bushenyi data analysis it was found that the models grossly underestimate the usage of rainfall harvesting. This is a result of the low reported usage in the DHS data. If the number of presences turns out to be very different than approximated in this thesis, the predicted number of presences accross Uganda would be different. However, the relative probability distribution of access types would remain the same, making it mostly a scaling issue. It would be interesting to validate the number of users per source for our research but also because there is very little information on this in literature.

6.1.2. SURVEY

Firstly, when creating a household survey, there is always the risk that it reflects mostly the views and perceptions of the researchers on the topics that are being researched. In that way, only the questions that the researcher deems relevant for water access and consumption of households are included, but some (to the researcher unknown) important questions are left out (Mukherjee, 1995). In our case this effect is mitigated because the survey was created in collaboration with local experts from Makerere University (Kampala) and because it built further on an earlier survey (Marks et al., 2020), that already highlighted that e.g. the multitude of water sources should be researched further (which we explicitly did).

Secondly, to be absolutely certain about all water behaviour, if possible at all, all Bushenyi-Ishaka municipality households should have been surveyed. As this is prevented by both time and cost constraints, this was limited to 517. Judging by the statistical significance of many of the results, this is a large enough sample to make comparisons between the surveyed households. On top of that, as shown in Chapter 4, the sample is large enough to represent the whole of the Bushenyi-Ishaka municipality. The modeling results showed that (levels of) water acces are different across the country and therefore, it can not be expected that the Bushenyi sample is representative for the whole of Uganda. It would be interesting though to perform similar survey campaigns in comparable settings but other regions (especially the water scarce North of Uganda), to compare the dynamics and local preferences and see how they overlap or differ.

Thirdly, for some questions in the survey, the respondent needed to give a quantitative response. This includes questions on number of jerrycans water consumed, travel times to water sources and water quality perceptions. Naturally, this brings the risk that the respondents answer is different from the real value. For the water consumption this is mitigated as people in Bushenyi usually collect water with the available 20L jerrycans and because paid water is paid for per jerrycan. This makes it easier for individuals to assess their water consumption than when they use collection materials of an unknown

volume that might differ a lot per household. However, to exactly determine the average daily water consumption, this should be registered on a much more detailed level by asking the respondent to very precisely register their water usage over multiple days. It would be interesting to compare the true versus the perceived value. The same is true for travel time, which ideally would be registered with a stopwatch instead of estimated. The perceived quality can be compared to the true quality which the researchers intend on doing in a later stage.

Lastly, the survey campaign took place during one of the two Ugandan wet seasons. To get a further understanding of the water consumption and peoples water behaviour, we recommend to perform a similar survey in a dry season and to look how that influences water consumption, travel time and water access in general. For instance, almost half of the households using rain water harvesting as primary water source reported that this is only the case in the wet season.

6.2. MEANING OF RESULTS AND BROADER IMPLICATIONS

The sections below provide a critical assessment of the results by placing them in perspective to each other and to the state of the art from literature. By comparing the results from the nationwide model scale and from the locally executed survey, it is possible to say something about which information does or does not scale and how spatial heterogeneity might impact results. Finally, some suggestions are made on how the obtained results could potentially be used to improve water access in Uganda.

6.2.1. MODELING THE NATIONWIDE LEVEL OF WATER ACCESS

We took an existing species modeling technique (MaxEnt) and improved and adapted it to model the distribution of water access points across Uganda. As these models can be build in several ways, we built five and compared their performance. The results of the classifying models show that it is possible to model the water access point distribution over the country, also into areas that were not surveyed for WPDx. The model was tested on 30% of the data that was kept apart during training. Most of the model setups perform good on the predefined evaluative parameters assigning higher scores to water point presence locations than background locations for all access types. The model with weighted background selection (M2) performs particularly well and the North South split (M4), performs worst. The more independent model evaluation on the DHS reported usage of the various access types (Table 5.5) presents less optimistic results. Although significant and mostly positive correlations were found between the number of predicted presences and the number of users of access types in the clusters, the correlations are not strong and sometimes negative. Furthermore, some results have to be interpreted with some caution as the sensitivity analysis showed that the predicted distributions and performance on evaluative parameters are influenced by the number of background points. It could be that a different (better) number of background points would also increase performance on the DHS data.

This type of presence only data modeling has been widely applied in biological species

modeling settings but has little precedent in the more static objective of water access modeling (animals move, boreholes do not). Our study includes both Neural Network models and the MaxEnt model used for a similar study by Yu et al. (2019). It was shown that the neural network outperforms the MaxEnt model. The MaxEnt is able to get satisfactory presence scores (PS) but does assign too high values to background locations, allowing for little nuance in the distinguishing features and more importantly wrongly assigning background locations as presence locations. This corresponds with Botella et al. (2018) who found that the neural network model did outperform MaxEnt in the species modeling (presence only) objective. This comes as no surprise as the neural network model is able to capture more complex and non-linear transformations of features and combinations of features than MaxEnt. Would the data suffice with the lesser complexity of the MaxEnt model, the neural network model would likely have adapted to it and obtained the same results. Therefore, the expectation is that more complex functions than MaxEnt is capable of, are required to model water access.

The lesser performance of MaxEnt also confirms our critique on the Area Under the ROC Curve (AUC) operator, that Yu et al. (2019) (and many reports in the species modeling field) report as their primary evaluative parameter, substantiating the in their view good results. Even though different numbers of background points resulted in different distributions of presence and other (worse) Presence Scores (PS) and False Presence Scores (FPS), the AUC remained unchanged (see Table 5.4). It therefore has no function in assessing the goodness of fit of the distribution (see Section 3.5). By splitting this AUC in two separate performance indicators namely the PS and FPS, like we did, more focus is placed on the predictive power of the model with respect to presences. This is key when modeling water access as this is really focused on predicting the type and level of access and arguably also for species modeling, although this might be less focused on (water access/species) presence enablers indicated by the feature layers with higher predictive power.

Another modeling improvement is the inclusion of the total number of expected water points per access type. Like us, Yu et al. (2019) built models per access type (2 in their case, 8 in ours) but the output of their models is only a relative probability of presence in cells. Our inclusion of the total expected presences, allows for a transformation of this relative probability of finding e.g. a borehole in a cell, into the expected presences in approximate but absolute terms (albeit under a number of assumptions, see Section 3.1.3). This gives more information on water access such as the co-existence of water access types, their expected quantity per cell and therefore contributes further to the objective of modeling water access across the study area.

Precipitation, elevation, population density, poverty and groundwater storage were found to be important indicators for the (non)presence of water access points, suggesting that water access is related to both socio-economic conditions as well as natural water availability and (probably) terrain suitability. As water access type modeling to our knowledge has little precedent, we have to take a closer look at these results to place them in perspective. For instance, the relative contribution of the feature layers to the

loss function showed that where the slope is not an important predictor for most access types, it is for unprotected springs (Fig. 5.5). This coincides with the results of Rahmati et al. (2018) who deemed slope the most important factor for unprotected spring presence in Iran. Yu et al. (2019) found that rainfall and elevation were important predictors for both unprotected wells and surface water access, which we see coming back in our results as well. This could potentially mean that certain features can scale beyond country borders. Their results also place a high weight on urbanisation, which is less dominantly present in our results. This can be explained by the fact that Yu et al. (2019) did not include the population density as a feature layer, which is assigned high weights in our results.

For a more detailed evaluation of model performance, a survey campaign was executed in Bushenyi (Uganda), enquiring 517 households about their water consumption and behaviour. For Bushenyi it was shown that, also on this detailed survey scale, the models are still able to pick up some water access characteristics such as elevation and poverty level. The respective models assign significantly larger scores to cells comprising households that report using piped water ($p<0.005$), rainwater harvesting ($p=0.06$) and shallow wells ($p<0.005$) respectively when compared to cells that do not report using these access types. For springs the model is off, assigning lower scores to cells with households using springs ($p=0.05$). This can be attributed to the local preference for springs for taste, cost and traditional reasons, of which some were also found earlier in Bushenyi (Marks et al., 2020). It is at the same time a good example of how the complex local dynamics, e.g. neighbours using completely different access types and people not (only) using their closest source but a multitude, are very hard to capture in a nationwide, one square kilometer gridded model.

6.2.2. HOUSEHOLDS MAKE USE OF MULTIPLE WATER SOURCES

In Bushenyi, households make use of an average of two water sources on a regular basis and over one hundred households (out of 517) reported using three or more. Out of the households, 66 (13%) reported not using their closest water source as primary water source. The main reason for this is that the closest water source is paid (as it is a piped connection most of the times), while the primary water source is free. On top of that 85 households (16%) with piped access reported not using the closest source as their drinking water source, in this case, safety was the primary reason. This exposes some shortcomings of JMP's safe water access definition that considers piped drinking water safe by definition (WHO, 2017) and does not mention affordability in their drinking water ladder that ranks the levels of water access. Our results indicate that many people who do have piped access closeby but, because they can not afford it or deem it unsafe, hardly make use of it.

Furthermore, similar to Hubbard et al. (2020), our result indicate that having no access to water from protected sources does not cause an increased risk on catching diarrhea and respiratory illnesses. Only if the water is (sometimes) not treated before drinking, we found that people have a significantly higher risk of catching these illnesses (up

to double the risk). In Bushenyi, almost 1 in 3 households reported not treating their water before drinking it at least once in the past 4 weeks. The WHO (2014) did find an increased risk of diarrhea for unprotected source users, but likewise report that proper and constant water treatment has a larger positive effect for prevention. All of this questions the usefulness of making an absolute distinction between protected and unprotected sources as our results show that this is not related to the mentioned diseases. This critique can be extended by Elliott et al. (2019), who argue that surveys on water access and consumption focus too much on the primary (drinking) water source. In Bushenyi we saw that many people make use of a variety of sources both protected and unprotected. These people will, although not daily but every now and then, drink untreated water or water from an unprotected source. The implications of this are such that a focus on the primary (drinking) water source in surveys combined with the focus on the protected/unprotected distinction, can cause an overestimation of the population that has permanent safe water access. On the flipside however, the usage of multiple sources can bring resilience to droughts or other dysfunctioning of water sources (Elliott et al., 2019).. Therefore we argue that a larger emphasis must be placed on always treating water and not to let unprotected versus protected water access be leading when assessing the level or safety of water access in developing countries. Furthermore, we suggest that health survey should better map the variety of sources used by households as this can have both positive and negative effects with regards to the safety and availability of water access.

As the models make predictions per access type, this does allow for the co-existence of access types in cells, which is similar to reality. In that way, the presence and the (variety) of type(s) of water access points can be indicated. From the model alone, it is however not possible to be certain that these access points are used as we saw with the local preference for springs and households reporting not using the piped connections in Bushenyi. In an optimal situation, especially locally obtained knowledge on water price but also cultural preferences for particular sources would be included. This could be obtained by expanding some of the DHS questions with questions on the variety of and preference for water sources used. From the model trained on half the country (M5) results, we know that differences in feature information are difficult to scale to areas that were not included in the model training (especially categorical data such as GW storage, might be completely different in areas not included in training, causing wrong predictions), let alone scaling community level local and cultural information regarding water access preferences. Despite all of this, we do see that when specific survey results from both nationwide demographic health surveys (DHS) and the Bushenyi survey are aggregated from the household level to a scale of one or multiple square km., which is more comparable to the models output (1 sq. km.), good predictions can be made. This is substantiated with (i) the predictions regarding average cluster travel times from the DHS data, where clusters with many predicted water access points have a lower average travel time, (ii) the evaluative parameters (PS>0.8, FPS<0.25), as well as (iii) the access type predictions in Bushenyi.

6.2.3. WATER CONSUMPTION IS DRIVEN BY MORE THAN TRAVEL TIME ALONE

With respect to water consumption, the WHO reports it is is primarily dependant on travel time to and from the water source (incl. queuing) (WELL, 1998) (Howard et al., 2020). A comparison between the model results and the reported travel times from the Ugandan DHS, showed that households living in areas of which the model expects a significantly higher number of water access points, report lower travel times on average. This indicates that more access points result in lower travel times. Contradictory to WELL (1998), in the Bushenyi survey there was no direct relationship between a household's travel time and water consumption to be found. Still, looking into access types more specifically it was shown that households with sources on premises (piped, rainwater harvesting) consume significantly more water and have lower travel times both in the Bushenyi setting as well as in the nation representative DHS. Other indicators of households using more water in Bushenyi are a higher education and a higher income (also seen in Kennedy et al. (2015), Liu et al. (2003), Fielding et al. (2012)), these are at the same time often the households that can afford a piped water connection (see Table 5.6 & 5.7). Another indicator which was also found in Fan et al. (2013) and Fielding et al. (2012), was the household size: larger households use significantly less water per capita than smaller households (and household size was not correlated with income). Also, households in the areas that the model assigned many access points, consume averagely more water per capita, suggesting a positive link between water access and consumption. At the same time, large differences are noticeable between neighbours where the one household uses 50 liter per capita per day (lpcd) and the neighbour only around 10 lpcd. This again shows that despite belonging to the same model output cell, differences with respect to the individual household's socio-economic characteristics, preferences, traditions but also efficiency of water usage (which was not included in our research), are inevitable to influence water consumption. When applying a water management plan to improve water access or to increase efficiency, it is therefore best to thoroughly assess the local cultural preferences.

6.2.4. IMPROVING WATER ACCESS IN UGANDA

Some of our results can be interpreted and used to potentially help improve water access in Uganda. Here we make a few suggestions for this. Firstly, the model can be used to identify areas in which water access is poor, for this the output maps can be used but one can also look into which features turned out to be good predictors. This could potentially help policy makers and NGOs to identify areas with really bad access, analyse why that is and possibly how it can be improved. The model results showed for instance that water access is generally high in urban areas compared to rural areas as the model expects larger number of water access points in urban areas. On top of that, the North of Uganda gets less rain than the South and also has lower GW storage, this too results in less water access points, mainly because the water resource availability is less (also seen in Nsubuga et al. (2014)). The elevation height is also of importance, however and this would need further research, it is suspected that not the elevation itself but more that the mountainous area have difficult terrain, height differences and deeper groundwater, making the terrain less suitable for most water access types. In that fashion, we found

that especially the higher, hilly more rural areas of Bushenyi had worse access than the urbanised lower areas, coinciding nicely with the model results.

The above mentioned predictive information of water access is mostly geo-hydrological. As it is not possible to change the water resources availability manually nor make the terrain more suitable to increase water access, it is worthwhile to take a look at socio-economic indicators of water access. From the model we than find that the eucledian distance (ED) to roads and cities are important indicators (both accounting for around 10%), showing that investing in mobility and infrastructure could increase water access, however, this might be correlation rather than causality and requires further research.

Furthermore, the survey learned us that better educated and richer households are not only expected to use more water but also more often use water from protected sources. Like mentioned before, investing in education can improve knowledge on the importance of safe water access (Kitamura et al., 2014). On the other side, the proximity of water access is reported to be also essential to good education as water collection time competes for (often) a child's time for education (WASH, 2014). Although drawing such conclusions from this thesis would be too strong, but in that fashion, if we again look at the North of Uganda, it is difficult to say whether this area is also poorer because of the lesser availability of water, or that because it is poorer, access is worse too. The latter is suggested in Mahama et al. (2014) and Adams et al. (2015) who found that poorer and less educated households had access to an improved source less often than richer households. But in Schuster-Wallace, Grover, Adeel, Confalonieri, and Elliott (2008), it is stated that a 1$ investment in clean and safe water access for drinking and sanitation can gain 3-34$ in economic return, but lack of it can cost up to 5% of a country's GDP (UNESCO, 2019). Putting water at the center in overcoming poverty is the core message of Unesco's recent World Water Development Report (UNESCO, 2019) and is also one of the core messages of the African Water Corridor (Delft Global Initiative, n.d.). Analysing poverty and water access in Bushenyi, we saw the lowest income category was far less likely to have access to drink from a protected source than the highest income category (65 vs. 91%). Also for the model, poverty comes out as an important factor, influencing water access prediction with around 10%. Although further research towards the adaptation and suitability of houses would be needed, a possibly straightforward improvement that could be made would be to assist low income households in acquiring a rainwater harvesting system as Baguma and Loiskandl (2010) showed that financial means are often the limiting factor for rainwater harvesting access and our survey results showed that the lowest income category had significantly less often access to rainwater harvesting techniques. Lastly, as it is uncertain how Ugandas average income and educational level is expected to develop (World Bank, 2021), if the population gets richer and better educated, we can expect a higher demand for better (piped) water access and with that an increase in consumption. According to the Bushenyi survey, this increase will not be extreme (around 5 lpcd.) but it is important to monitor this and to be aware of the limited amount of natural water resources (Nsubuga et al., 2014), especially in combination with the continuing rapid population growth (World Bank, 2019).

7

Conclusion

Monitoring safe water access happens primarily through household health surveys. These surveys are often incomplete, not covering entire nations, focus on only the primary water source and are often spatially aggregated for privacy reasons. Besides, health surveys almost never include questions on consumed water volumes while that is an important indicator for proper hygiene (WELL, 1998) and at the same time should be in balance with the natural available water resources. Next to this survey based monitoring, there is the Water Point Data Exchange (WPDx) that monitors safe access by providing a platform at which the exact location and type of water access points (such as boreholes, springs, etc.) are registered. This does give more insight into the presence and usage of a variety of sources, but also the WPDx is often incomplete, not covering entire nations. In this thesis, a dual methodology was presented that gap-fills the incompleteness of such databases through nationwide modeling of water access points and in parallel, researches the complex local dynamics of water access, the variety of water sources used by households and the relationships between access and water consumption.

First, a machine learning biological species modeling technique was improved and applied to the WPDx dataset. With this we were able to predict the presence of 8 different water access types across Uganda with a resolution of one square kilometer. To do so, a predefined set of nationwide gridded environmental, geohydrological and socio-economic characteristics were used as predictive features. With the result it is possible to assess the water access level in areas that are not surveyed for WPDx. Good modeling results were obtained: out of a test sample that was kept apart and not used for training the model, containing approximately 30 thousand presences and 120 thousand background locations, the model was capable to predict 90% of presence locations correctly and only misjudged 24% of background locations wrongly as presence location. Some of these results have to be interpreted with some caution in light of the executed sensitivity analysis and model validation on nationwide Demographic Health Survey data, that obtained lesser results.

Second, in collaboration with Makerere University (Kampala), a novel survey campaign was executed in the Bushenyi-Ishaka municipality, a mid-sized town in the South West of Uganda consisting of a mixture of both urban and rural areas. Unlike more standard water access or health surveys, this survey included questions on (volumtreic) water consumption and the variety of water access points used. In total, 517 households were interviewed. Special attention was paid to reasons for not using the closest water source (often piped water) and it was found that piped sources were often not used as primary (most used) source for cost reasons. For drinking water purposes, piped water was often not used because of a bad perceived water quality. In both cases there was a preference for other free sources and sources from which people have been accustomed to drink from for a long time such as, in this case, springs. All of this shows that safe water access is not only a matter of the presence of a source but also of the affordability and the users perspective.

It was found that that water access is dependant both on natural water resource availability as well as socio-economic status. More specifically, the model results showed that population density, precipitation, elevation and groundwater storage are important predictors for the presence of water access points. The survey results add wealth (which we see coming back in the model results too) and education level as significant positive indicators for safe water access. This comes with the notion that households in Bushenyi make use of an average of two water sources on a regular basis. And although the model predictions allow for co-existence of water access types as models are created per access type, it can not model the actual usage of these access points which was found to be dependent on local preferences, cost constraints and quality perception.

Also with regards to water consumption, it was found that wealth and education but also household size (inversed) drive up a households water consumption significantly. Moreover, households in areas of which the model predicts high water access point presences, tend to use more water on average than areas with low predicted presences. This suggests a positive link between the number of water presences and water consumption. Contrary to the consensus in literature, in Bushenyi, a strong link between water consumption and travel time to the source was not found.

Finally, we suggested modi operandi of our results to improve water access such as prioritising areas with poor(est) water access and investing in rainwater harvesting, infrastructure and education. Moreover, our results show that when reporting on (improved) water access, it is important to include the complex dynamics of water access such as the variety of sources used by households and the sometimes slow adaptation to improved (piped) water access when households deem it too costly or unsafe. Without adding this nuance, the share of people having safe and improved access is at risk of being overestimated.

REFERENCES

Adams, E. A., Boateng, G. O., & Amoyaw, J. A. (2015, feb). Socioeconomic and demographic predictors of potable water and sanitation access in ghana. *Social Indicators Research*, *126*(2), 673–687. Retrieved from https://doi.org/10.1007%2Fs11205-015-0912-y doi: 10.1007/s11205-015-0912-y

Armah, F. A., Ekumah, B., Yawson, D. O., Odoi, J. O., Afitiri, A.-R., & Nyieku, F. E. (2018, nov). Access to improved water and sanitation in sub-saharan africa in a quarter century. *Heliyon*, *4*(11), e00931. Retrieved from https://doi.org/10.1016%2Fj.heliyon.2018.e00931 doi: 10.1016/j.heliyon.2018.e00931

Atuhaire, P. (2022, Jan). *Uganda schools reopen after almost two years of covid closure.* BBC. Retrieved from https://www.bbc.com/news/world-africa-59935605

Baguma, D., & Loiskandl, W. (2010, apr). Rainwater harvesting technologies and practises in rural uganda: a case study. *Mitigation and Adaptation Strategies for Global Change*, *15*(4), 355–369. Retrieved from https://doi.org/10.1007%2Fs11027-010-9223-4 doi: 10.1007/s11027-010-9223-4

Bartram, J., Brocklehurst, C., Fisher, M., Luyendijk, R., Hossain, R., Wardlaw, T., & Gordon, B. (2014, aug). Global monitoring of water supply and sanitation: History, methods and future challenges. *International Journal of Environmental Research and Public Health*, *11*(8), 8137–8165. Retrieved from https://doi.org/10.3390%2Fijerph110808137 doi: 10.3390/ijerph110808137

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th international conference on international conference on machine learning - volume 28* (p. I–115–I–123). JMLR.org.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., & Munoz, F. (2018). A deep learning approach to species distribution modelling. In *Multimedia tools and applications for environmental & biodiversity informatics* (pp. 169–199). Springer International Publishing. Retrieved from https://doi.org/10.1007%2F978-3-319-76445-0_10 doi: 10.1007/978-3-319-76445-0_10

Brownlee, J. (2020, Aug). *4 types of classification tasks in machine learning.* Retrieved from https://machinelearningmastery.com/types-of-classification-in-machine-learning/

Bushenyi District. (2020, May). *Statistics: Bushenyi district.* Retrieved from https://bushenyi.go.ug/district/statistics

Cassivi, A., Guilherme, S., Bain, R., Tilley, E., Waygood, E. O. D., & Dorea, C. (2019). Drinking water accessibility and quantity in low and middle-income countries: A systematic review. *International Journal of Hygiene and Environmental Health*, *222*(7), 1011-1020. Retrieved from https://www.sciencedirect.com/science/article/pii/S1438463919301348 doi: https://doi.org/10.1016/j.ijheh.2019.06.011

Cordoba, C., & Grabinsky, J. (2020, Aug). *Many countries rely on private providers for access to water – but blindly embracing these sources as "clean" raises concerns.* World Bank. Retrieved from https://blogs.worldbank.org/water/many-countries-rely-private-providers-access-water-blindly-embracing-these-sources-clean

Daly, S. W., Lowe, J., Hornsby, G. M., & Harris, A. R. (2021, apr). Multiple water source use in low- and middle-income countries: a systematic review. *Journal of Water and Health*, *19*(3), 370–392. Retrieved from https://doi.org/10.2166%2Fwh.2021.205 doi: 10.2166/wh.2021.205

Danneman, N., & Clauser, K. (2020, Apr). *Auc vs log loss*. Data Machines Corp. Retrieved from https://www.datamachines.io/blog/auc-vs-log-loss

Delft Global Initiative. (n.d.). *African water corridor*. Retrieved from https://www.tudelft.nl/global/african-water-corridor/

Dembla, G. (2020, Nov). *Intuition behind log-loss score*. Towards Data Science. Retrieved from https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a

Elith, J., Graham, C., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., ... Zimmermann, N. (2006, 04). Novel methods improve prediction of species' distributions from occurence data. *Ecography*, *29*, 129 - 151. doi: 10.1111/j.2006.0906-7590.04596.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2010, nov). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. Retrieved from https://doi.org/10.1111%2Fj.1472-4642.2010.00725.x doi: 10.1111/j.1472-4642.2010.00725.x

Elliott, M., Foster, T., MacDonald, M. C., Harris, A. R., Schwab, K. J., & Hadwen, W. L. (2019, mar). Addressing how multiple household water sources and uses build water resilience and support sustainable development. *npj Clean Water*, *2*(1). Retrieved from https://doi.org/10.1038%2Fs41545-019-0031-4 doi: 10.1038/s41545-019-0031-4

Fan, L., Liu, G., Wang, F., Geissen, V., & Ritsema, C. J. (2013, aug). Factors affecting domestic water consumption in rural households upon access to improved water supply: Insights from the wei river basin, china. *PLoS ONE*, *8*(8), e71977. Retrieved from https://doi.org/10.1371%2Fjournal.pone.0071977 doi: 10.1371/journal.pone.0071977

Fielding, K. S., Russell, S., Spinks, A., & Mankad, A. (2012, oct). Determinants of household water conservation: The role of demographic, infrastructure, behavior, and psychosocial variables. *Water Resources Research*, *48*(10). Retrieved from https://doi.org/10.1029%2F2012wr012398 doi: 10.1029/2012wr012398

Golini, N. (2012, 01). Bayesian modeling of presence-only data.

Haruna, R., Ejobi, F., & Kabagambe, E. (2005, 04). The quality of water from protected springs in katwe and kisenyi parishes, kampala city, uganda. *African health sciences*, *5*, 14-20.

Heijden, T. V. D., Lago, J., Palensky, P., & Abraham, E. (2021). Electricity price forecasting in european day ahead markets: A greedy consideration of market integration. *IEEE Access*, *9*, 119954–119966. Retrieved from https://doi.org/10.1109%2Faccess.2021.3108629 doi: 10.1109/access.2021.3108629

Howard, G., Bartram, J., Williams, A., Overbo, A., Fuente, D., & Geere, J.-A. (2020). *Domestic water quantity, service level and health, second edition*.

Howard, G., Teuton, J., Luyima, P., & Odongo, R. (2002, mar). Water usage patterns in low-income urban communities in uganda: Implications for water supply surveillance. *International Journal of Environmental Health Research*, *12*(1), 63–73. Re-

trieved from https://doi.org/10.1080%2F09603120120110068 doi: 10.1080/09603120120110068

Hubbard, S. C., Meltzer, M. I., Kim, S., Malambo, W., Thornton, A. T., Shankar, M. B., ... Brunkard, J. M. (2020, jun). Household illness and associated water and sanitation factors in peri-urban lusaka, zambia, 2016–2017. *npj Clean Water*, *3*(1). Retrieved from https://doi.org/10.1038%2Fs41545-020-0076-4 doi: 10.1038/s41545-020-0076-4

Jiménez-Valverde, A. (2011, may). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, *21*(4), 498–507. Retrieved from https://doi.org/10.1111%2Fj.1466-8238.2011.00683.x doi: 10.1111/j.1466-8238.2011.00683.x

JMP, & WHO. (n.d.). *Urban improved water access vs. rural water access*. Retrieved from https://ourworldindata.org/grapher/urban-improved-water-access-vs-rural-water-access

Kennedy, C. A., Stewart, I., Facchini, A., Cersosimo, I., Mele, R., Chen, B., ... Sahin, A. D. (2015, apr). Energy and material flows of megacities. *Proceedings of the National Academy of Sciences*, *112*(19), 5985–5990. Retrieved from https://doi.org/10.1073%2Fpnas.1504315112 doi: 10.1073/pnas.1504315112

Kilimani, N., et al. (2013). Water resource accounts for uganda: Use and policy relevancy. *Economic Research Southern Africa*.

Kingma, D., & Ba, J. (2014, 12). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kitamura, Y., Eri Yamazaki, N. K., Jr., D. B. E., Shivakoti, B. R., Mitra, B. K., Abe, N., ... Stevens, C. (2014). *Linking education and water in the sustainable development goals. post2015/unu-ias policy brief 2*.

Kong, Y.-L., Anis-Syakira, J., Fun, W. H., Balqis-Ali, N. Z., Shakirah, M. S., & Sararaks, S. (2020, oct). Socio-economic factors related to drinking water source and sanitation in malaysia. *International Journal of Environmental Research and Public Health*, *17*(21), 7933. Retrieved from https://doi.org/10.3390%2Fijerph17217933 doi: 10.3390/ijerph17217933

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, *30*(3), 607-610. Retrieved from https://doi.org/10.1177/001316447003000308 doi: 10.1177/001316447003000308

Lecun, Y., Bottou, L., Orr, G., & Müller, K.-R. (2000, 08). Efficient backprop.

Li, W., Guo, Q., & Elkan, C. (2011). Can we model the probability of presence of species without absence data? *Ecography*, *34*(6), 1096–1105. Retrieved from http://www.jstor.org/stable/41315788

Liu, J., Savenije, H. H., & Xu, J. (2003, jan). Forecast of water demand in weinan city in china using WDF-ANN model. *Physics and Chemistry of the Earth, Parts A/B/C*, *28*(4-5), 219–224. Retrieved from https://doi.org/10.1016%2Fs1474-7065%2803%2900026-3 doi: 10.1016/s1474-7065(03)00026-3

Mahama, A. M., Anaman, K. A., & Osei-Akoto, I. (2014, jan). Factors influencing householders' access to improved water in low-income urban areas of accra, ghana.

Journal of Water and Health, *12*(2), 318–331. Retrieved from https://doi.org/10.2166%2Fwh.2014.149 doi: 10.2166/wh.2014.149

Marks, S. J., Clair-Caliot, G., Taing, L., Bamwenda, J. T., Kanyesigye, C., Rwendeire, N. E., ... Ferrero, G. (2020, apr). Water supply and sanitation services in small towns in rural–urban transition zones: The case of bushenyi-ishaka municipality, uganda. *npj Clean Water*, *3*(1). Retrieved from https://doi.org/10.1038%2Fs41545-020-0068-4 doi: 10.1038/s41545-020-0068-4

Merow, C., Smith, M. J., & Silander, J. A. (2013, jun). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. Retrieved from https://doi.org/10.1111%2Fj.1600-0587.2013.07872.x doi: 10.1111/j.1600-0587.2013.07872.x

Miraki, S., Zanganeh, S. H., Chapi, K., Singh, V. P., Shirzadi, A., Shahabi, H., & Pham, B. T. (2018, sep). Mapping groundwater potential using a novel hybrid intelligence approach. *Water Resources Management*, *33*(1), 281–302. Retrieved from https://doi.org/10.1007%2Fs11269-018-2102-6 doi: 10.1007/s11269-018-2102-6

Mukherjee, N. (1995). *Participatory rural appraisal and questionnaire survey: Comparative field experience and methodology innovations (studies in rural participation)*. Concept Publishing Company.

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2015, dec). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in iran. *Environmental Monitoring and Assessment*, *188*(1). Retrieved from https://doi.org/10.1007%2Fs10661-015-5049-6 doi: 10.1007/s10661-015-5049-6

Nsubuga, F. N. W., Namutebi, E. N., & Nsubuga-Ssenfuma, M. (2014). Water resources of uganda: An assessment and review. *Journal of Water Resource and Protection*, *06*(14), 1297–1315. Retrieved from https://doi.org/10.4236%2Fjwarp.2014.614120 doi: 10.4236/jwarp.2014.614120

O'Neill, A. (2021, Apr). *Uganda - urbanization 2019*. Retrieved from https://www.statista.com/statistics/447899/urbanization-in-uganda/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006, jan). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3-4), 231–259. Retrieved from https://doi.org/10.1016%2Fj.ecolmodel.2005.03.026 doi: 10.1016/j.ecolmodel.2005.03.026

Qi, C., & Chang, N.-B. (2011, jun). System dynamics modeling for municipal water demand estimation in an urban region under uncertain economic impacts. *Journal of Environmental Management*, *92*(6), 1628–1641. Retrieved from https://doi.org/10.1016%2Fj.jenvman.2011.01.020 doi: 10.1016/j.jenvman.2011.01.020

Rahmati, O., Naghibi, S., Shahabi, H., Bui, D., Pradhan, B., Azareh, A., ... Melesse, A. (2018, 08). Groundwater spring potential modelling: Comprising the capability and robustness of three different modeling approaches. *Journal of Hydrology*, *565*. doi: 10.1016/j.jhydrol.2018.08.027

Rhoderick, A. (2013). Examining the relationship between distance and water quantity: a systematic review and a multi-country field study. Retrieved from https://cdr.lib.unc.edu/concern/masters_papers/xg94hr464 doi: 10.17615/JEC7-8Z62

Santos, S. D., Adams, E., Neville, G., Wada, Y., de Sherbinin, A., Bernhardt, E. M., & Adamo, S. (2017, dec). Urban growth and water access in sub-saharan africa: Progress, challenges, and emerging research directions. *Science of The Total Environment*, *607-608*, 497–508. Retrieved from https://doi.org/10.1016%2Fj.scitotenv.2017.06.157 doi: 10.1016/j.scitotenv.2017.06.157

Schuster-Wallace, C. J., Grover, V. I., Adeel, Z., Confalonieri, U., & Elliott, S. (2008). *Safe water as the key to global health.*

Scikit-Learn-Documentation. (n.d.). *1.17. neural network models (supervised).* Retrieved from https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Silva-Novoa Sanchez, L. M., Kemerink-Seyoum, J. S., Waiswa Batega, D., & Paul, R. (2020, 05). Caught in the middle? Access to water in the rural to urban transformation of Bushenyi-Ishaka municipality, Uganda. *Water Policy*, *22*(4), 670-685. Retrieved from https://doi.org/10.2166/wp.2020.024 doi: 10.2166/wp.2020.024

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019, feb). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. Retrieved from https://doi.org/10.1111%2F2041-210x.13140 doi: 10.1111/2041-210x.13140

Ssentongo, P., Muwanguzi, A. J. B., Eden, U., Sauer, T., Bwanga, G., Kateregga, G., ... Schiff, S. J. (2018, feb). Changes in ugandan climate rainfall at the village and forest level. *Scientific Reports*, *8*(1). Retrieved from https://doi.org/10.1038%2Fs41598-018-21427-5 doi: 10.1038/s41598-018-21427-5

Thompson, J., Purras, I., & Tumwine, J. (2001). *Drawers of water ii.*

Tognelli, M. F., Roig-Juñent, S. A., Marvaldi, A. E., Flores, G. E., & Lobo, J. M. (2009). An evaluation of methods for modelling distribution of patagonian insects. *Revista chilena de historia natural*, *82*(3). Retrieved from https://doi.org/10.4067%2Fs0716-078x2009000300003 doi: 10.4067/s0716-078x2009000300003

Uganda Bureau of Statistics. (2018). *Uganda demographic and health survey 2016* (Vol. 1; Tech. Rep.). Kampala or Maryland.

UNESCO. (2019). *The united nations world water development report 2019: Leaving no one behind.*

Unicef. (2015, May). *Water point data exchange launched.* Retrieved from https://www.unicef.org/innovation/stories/water-point-data-exchange-launched

United Nations General Assembly. (2015). *Transforming our world: the 2030 agenda for sustainable development.*

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, *220*(4), 589-594. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304380008005486 doi: https://doi.org/10.1016/j.ecolmodel.2008.11.010

WASH. (2014, Oct). *Water and education: How safe water access helps schoolchildren.* Retrieved from https://lifewater.org/blog/water-education/

WELL. (1998). *Guidance manual on water supply and sanitation programmes.* WEDC.

West, A. M., Kumar, S., Brown, C. S., Stohlgren, T. J., & Bromberg, J. (2016). Field validation of an invasive species maxent model. *Ecological Informatics*, *36*, 126-134. Retrieved from https://www.sciencedirect.com/science/article/pii/S1574954116301923 doi: https://doi.org/10.1016/j.ecoinf.2016.11.001

WHO. (2014). *Preventing diarrhoea through better water, sanitation and hygiene: Exposures and impacts in low- and middle-income countries.*

WHO. (2017). *Safely managed drinking water - thematic report on drinking water.*

World Bank. (2019). *Population, total - uganda.* Retrieved from https://data.worldbank.org/indicator/SP.POP.TOTL?locations=UG

World Bank. (2021, Nov). *The world bank in uganda: Overview.* Retrieved from https://www.worldbank.org/en/country/uganda/overview#1

WPDx. (2021a). *Using data to achieve sdg 6.* Retrieved from https://www.waterpointdata.org/wp-content/uploads/2021/04/WPDx-and-SDGs.pdf

WPDx. (2021b, January). *Wpdx data standard.* Retrieved from https://www.waterpointdata.org/wp-content/uploads/2021/04/WPDx_Data_Standard.pdf

Yu, W., Wardrop, N. A., Bain, R. E. S., Alegana, V., Graham, L. J., & Wright, J. A. (2019, may). Mapping access to domestic water supplies from incomplete data in developing countries: An illustrative assessment for kenya. *PLOS ONE*, *14*(5), e0216923. Retrieved from https://doi.org/10.1371%2Fjournal.pone.0216923 doi: 10.1371/journal.pone.0216923

8

Appendices

Appendix A: Supplementary Material to the Literature Review

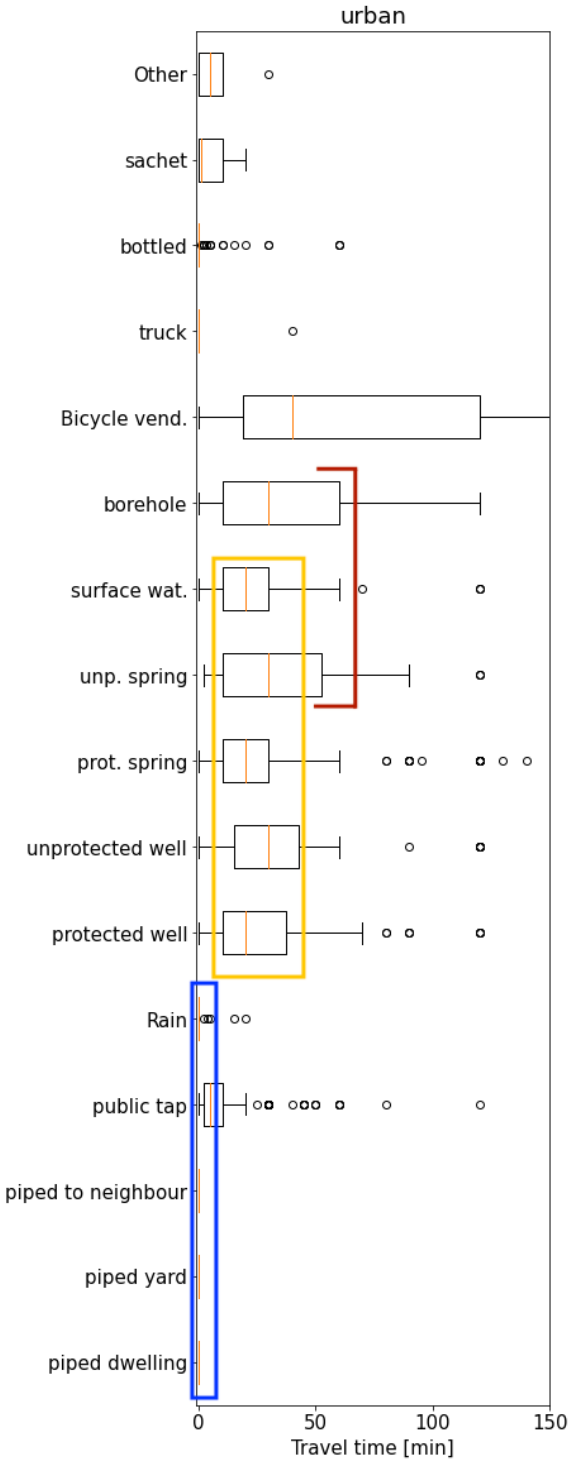


Figure 8.1: Distribution of travel times reported by households in urban areas using the access types on the y-axis (created from DHS data)

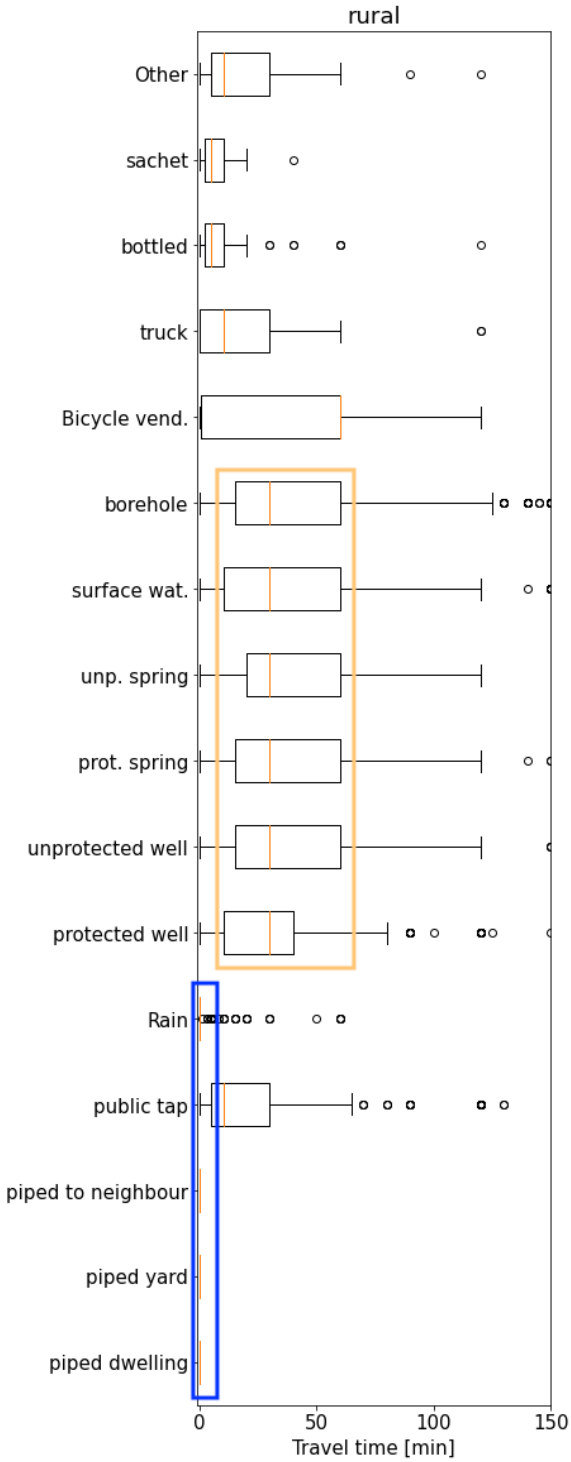


Figure 8.2: Distribution of travel times reported by households in rural areas using the access types on the y-axis (created from DHS data

APPENDIX B: SUPPLEMENTARY RESULTS

B.1: PERFORMANCE OF M3 FOR URBAN AND RURAL MODEL SEPARATELY

Access Type	PS		FPS	
	Urban	Rural	Urban	Rural
Borehole	.72	.67	.22	.27
Piped Water	.83	.87	.21	.23
Protected Shallow Well	.65	.89	.23	.23
Protected Spring	.77	.8	.22	.24
Rainwater Harvesting	.72	.67	.25	.2
Surface Water	1	.44	.2	.2
Unprotected Shallow Well	1	.88	.75	.23
Unprotected Spring	.75	.88	.14	.17
Weighted Average	.74	.72	.23	.24
Average	.81	.76	.28	.22

Table 8.1: Performance of the rural and urban models separately (M3)

B.2: SENSITIVITY ANALYSIS OF BOREHOLE RESULTS

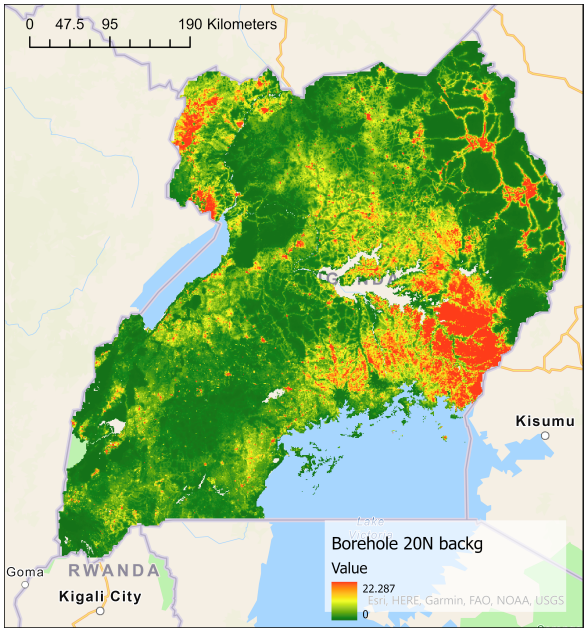


Figure 8.3: M1 model output for 20*N number of background points in which N is the number of presences

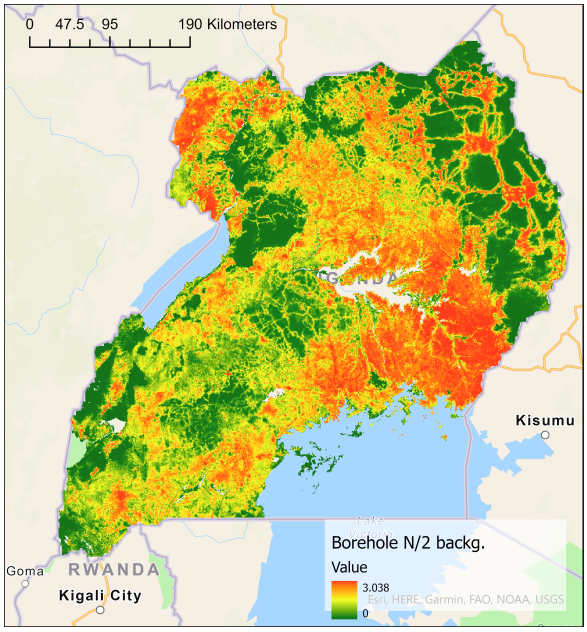


Figure 8.4: M1 model output for N/2 number of background points in which N is the number of presences

B.3: Spearman in Bushenyi

access type	r_s	p
NWSC_Piped	0.34	0.006
Rain_water	-0.02	0.880
Borehole	-0.21	0.094
Shallow_Wells	0.21	0.092
Springs	-0.20	0.110
Surface_water	-0.17	0.170

Table 8.2: Spearmans rank coefficient for the Bushenyi cells. Just like in Table 5.5, M1 model predictions are compared to the share of the population reporting using these access types in the Bushenyi cells.

B.4: LAYER IMPORTANCE OF M3 FOR RURAL AND URBAN SEPARATELY

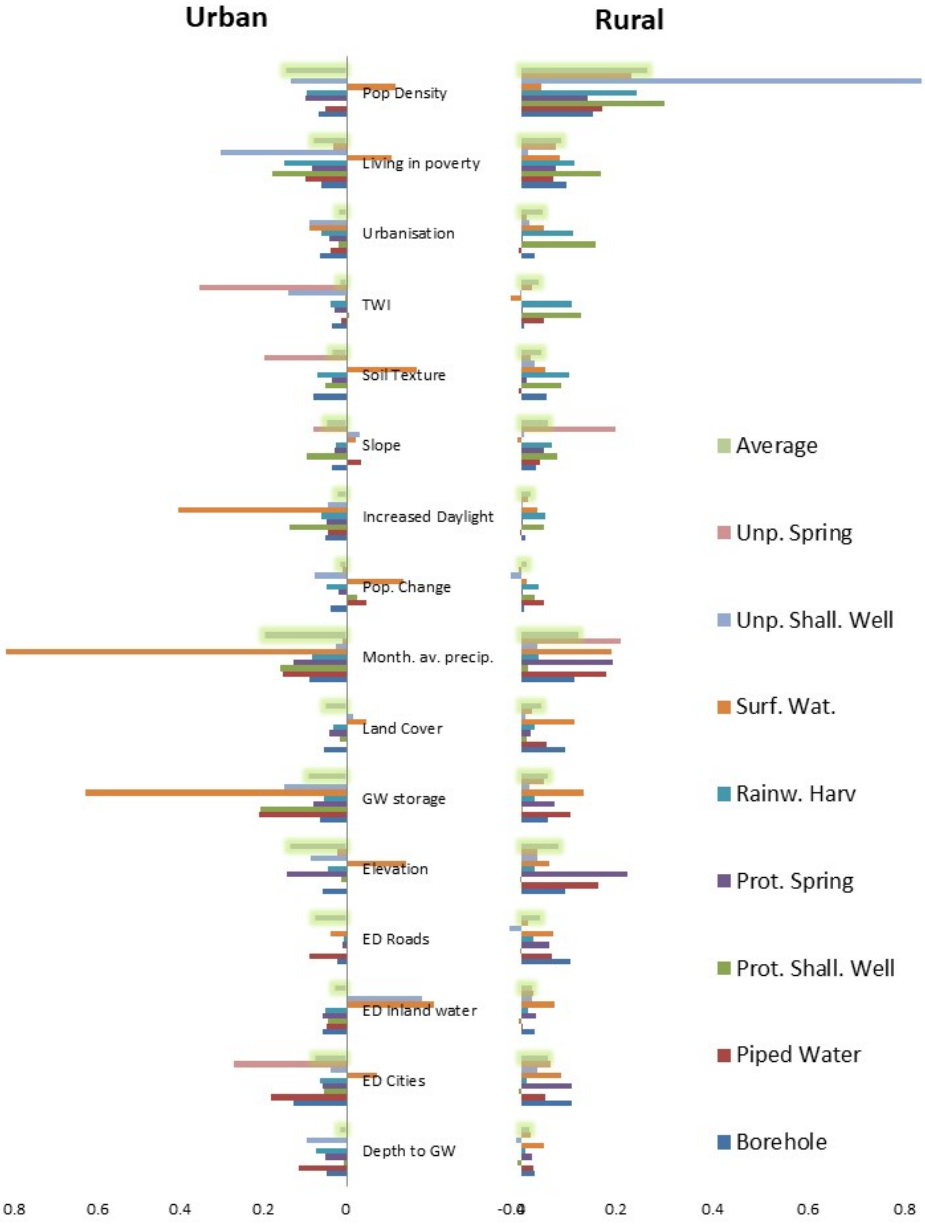


Figure 8.5: Per access type, the relative contribution to loss function (Eq. 3.7) of each feature for the MLP classifier trained on urban (left) and rural data (right) respectively

B.5: Predicting DHS travel time using access type predictions

After fitting the access type model output (such as Figure 5.1) to the DHS cluster average travel time using both a MLP-regression fit as well as an ordinary Least Square (linear) fit, it was found that the fit of the M1 model (Standard Classifier), was best, although the model with weighted background data performed very comparable. The absolute error distribution of both the M1 regressions can be seen in Figure 8.6 & 8.7. For the linear regression and the MLP regression, the median errors are 13.2 and 10.8 minutes respectively, showing that 50% of predictions are better than that. However, if we compare it to the DHS regression (M6) using only the feature data, of which the absolute error distribution is shown in Figure 5.2, we find that the travel time prediction does not improve when the WPDx data is included in this way. This can mean that the amount or type of predicted access points has no direct influence on the travel time, or that the regression methods are not able to pick this up properly.

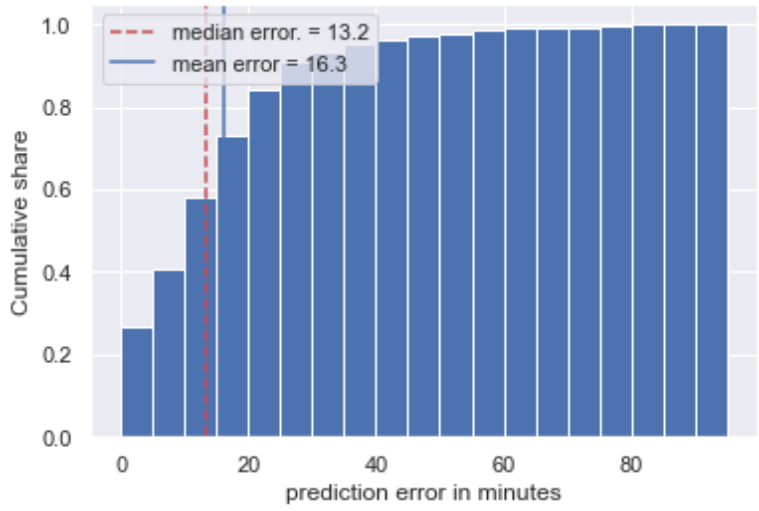


Figure 8.6: Absolute error distribution after applying linear regression to fit M1's expected number of presences of the several access types to the DHS cluster average travel time

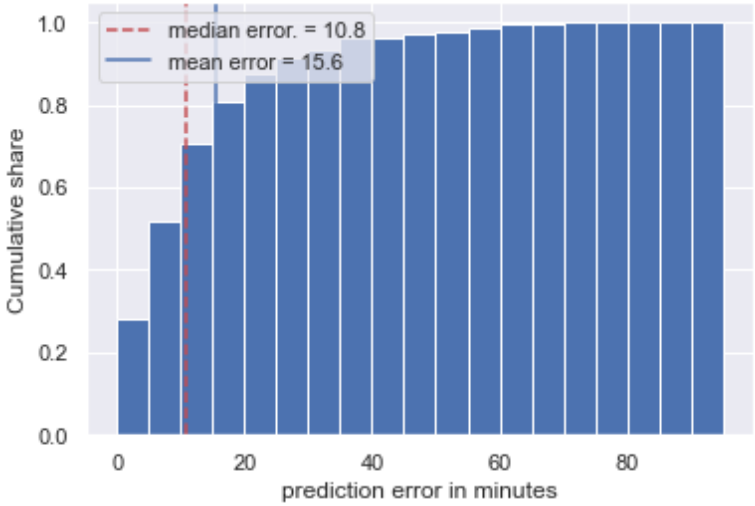


Figure 8.7: Absolute error distribution after applying a MLP Regression to fit M1's expected number of presences of the several access types to the DHS cluster average travel time

B.6: TRAVEL TIME DISTRIBUTION PER ACCESS TYPE IN BUSHENYI

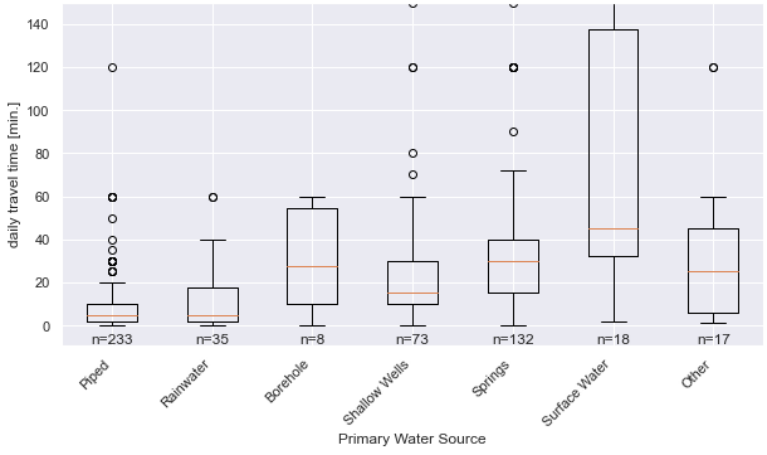


Figure 8.8: Bushenyi: distribution of the travel time to and from the PWS grouped per access type listed as PWS.

APPENDIX C: FIELD REPORT BY JAMES TAYEBWA (HEAD OF SURVEY TEAM)

Dear all,

I was able to put together a brief field report which is representative of the views collected from my team.
Key findings and observations

1. Most Households which have access to NWSC have continued to use other sources such as shallow wells and springs as drinking water sources because;

2. It was observed that some of the respondents/households who said they use spring water for drinking wouldn't bother boiling it, they assumed that it was safe for drinking with out boiling, and even those who boil the water, since it was the predominant way of treating water, they indicated that they had no clears ways of controlling other HH members from taking un-boiled water, while others indicated to us that whenever they take boiled water, they would contract diseases like cough and flue.

3. In areas covered with NWSC.(1) There was consistency in reporting of muddy, fatty layered and salty water especially when taps are turned on in the mornings or when there has been a breakdown in the water supply for some days. (2)There is also a persistent feedback on too much chlorine in the water. So, some households fear that they are drinking chemicals, and some households prefer sourcing drinking water from other sources, especially from springs... even when these households have a national water connection. However, others are now getting used to using water from NWSC for all household water needs including drinking.

4. In other cases, the reporting was emphasizing on little volume of NWSC, springs and other water sources especially in the dry season. (4) On the aspect of the tool, there was lack of specificity in some questions, in that we missed out on the narratives from respondents, which would have been helpful in the analysis...

5. In some areas where NWSC is available, some households aren't using it due to cost related issues, there are quite few pubic stand pipes, which forces some households to use far away water sources because of either restrictions or very hiked prices from neighbors who are connected to NWSC. Some indicated that the water bill that keeps shooting up even when they are in a wet season, and are using rain water, a factor that will force some of them to shift back to their previous water sources.

6. Most households reported cases of Flue and Cough, we couldn't establish the actual cause, the possibilities could be related to water for drinking, changes in the weather patterns, or a possibility of it transferred through human interactions and contact...

Challenges

1. The tool wouldn't allow one to retrieve, and edit after sending for-example if one had made a mistake on the household ID instead of 11 one puts 1, retrieving it to correct the error wouldn't be possible.

2. The location of the primary water source was purely dependent on where the respondent told us it was located.....I personally felt that the locations we pointed out on the GPS are not 100% accurate maybe it's somewhere around that GPS location but not the exact point

3. We also encountered a challenge of some respondents refusing to accept soap....and I had to first convince them to take it....even the consent form copy .....but they would sign willingly

4. Some of the respondents would give biased answers just to please the interviewer. An example, I found some old man dipping his cup in the drum of water and started to drink un-boiled water but while responding to the question of the method of treating water, he said he has to boil it, though he later admitted that he has never drunk boiled water.

5. While working in places with poor or no network connections, it would be hard to locate the water source because the map could hardly show up. I don't know how this will be dealt with at the level of taking water samples, however, they were not many incidences.

Areas for further research

1. Testing reasons why most households use boiling compared to other water treatment methods

2. Testing the palatibility of the drinking water source in relation to people's health.

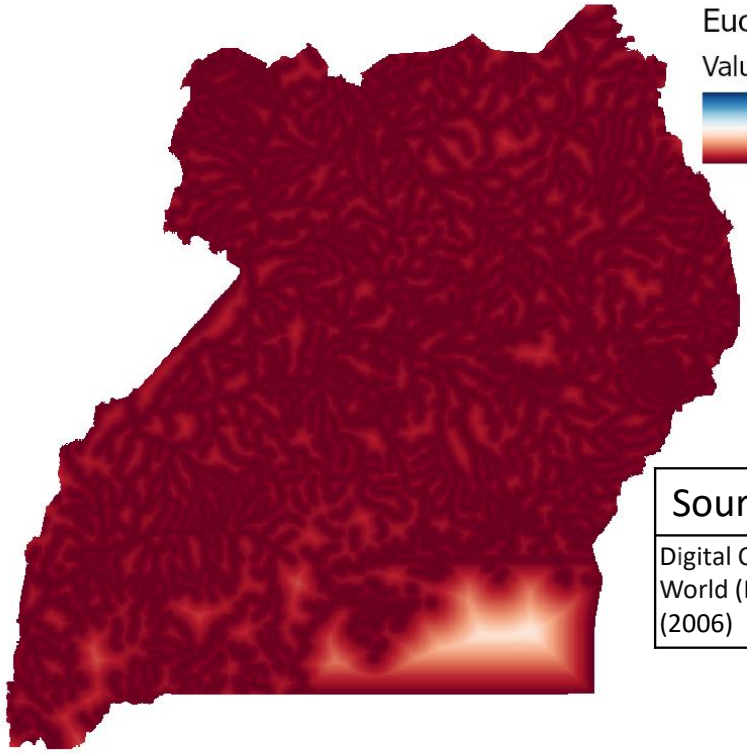
3. Most areas in Ishaka-Bushenyi area are water constrained, the most available water source is mainly for home use is NWSC and expensive, exploring the component of water for production might be a good idea

I hope this feedback will be useful to Jan as he takes on with the analysis, and like Edo requested, I am available to offer a hand in case of need.
Thank you all, and on behalf of my team.

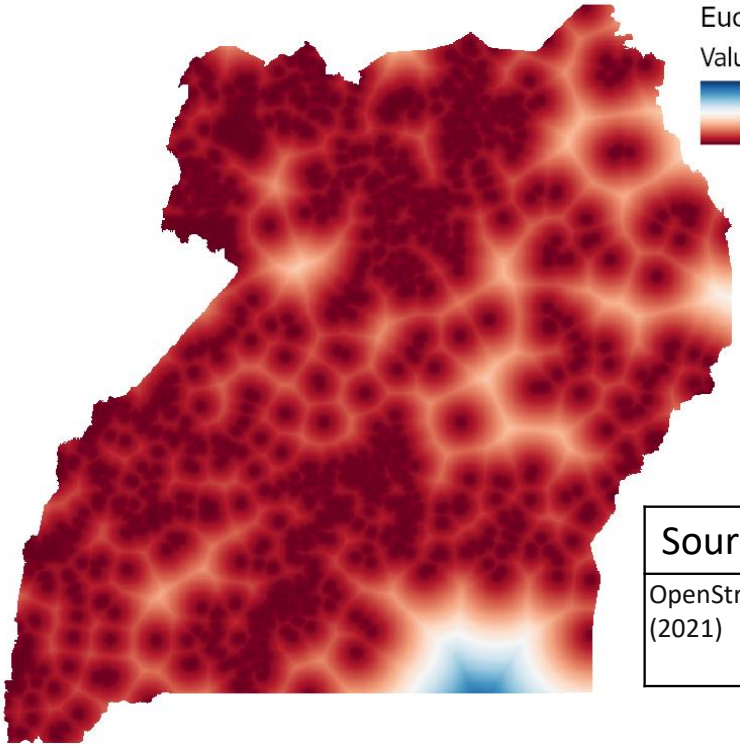
kind regards
James

APPENDIX D: FEATURE LAYERS

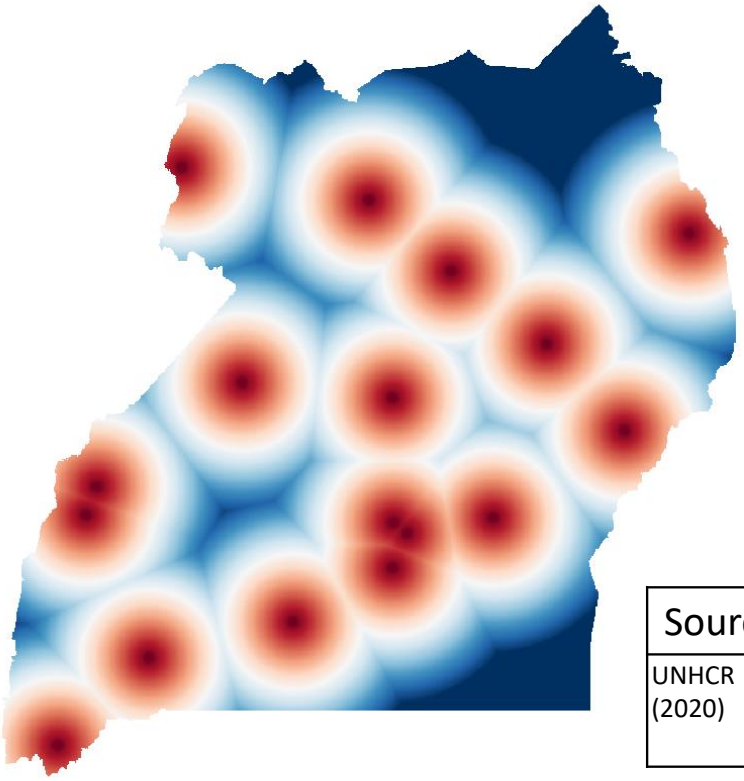
Appendix D: Feature layers



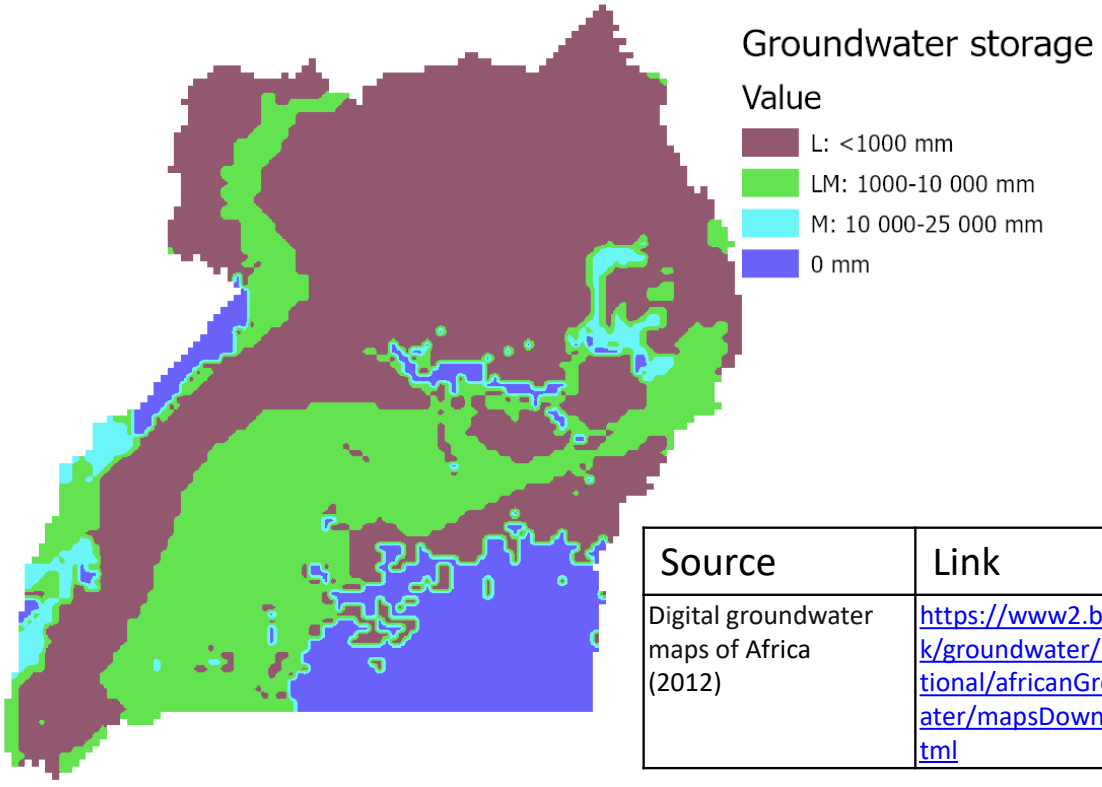
Source	Link
Digital Chart of the World (DCW) (2006)	https://www.diva-gis.org/gdata



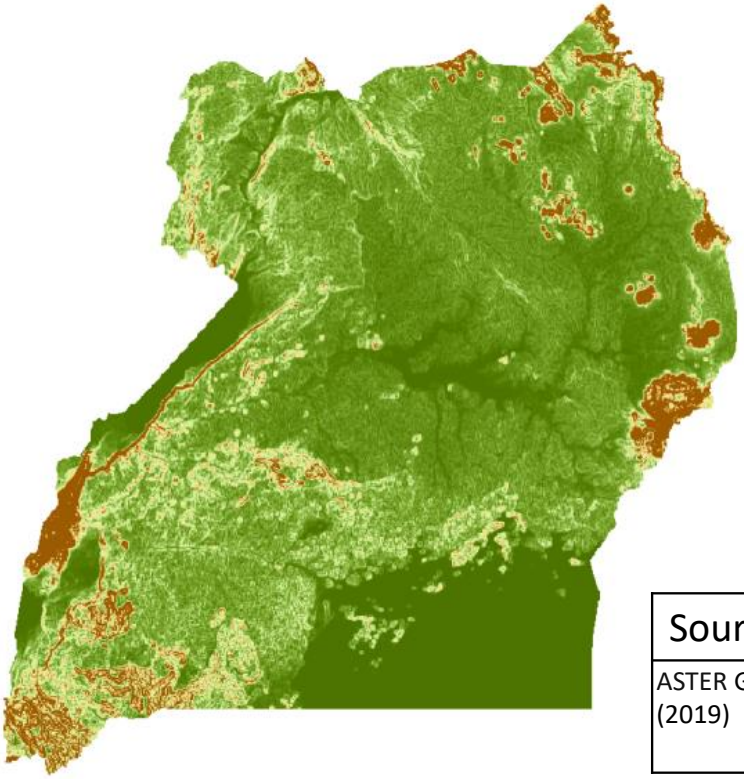
Source	Link
OpenStreetMap (2021)	http://download.geofabrik.de/africa/uganda.html



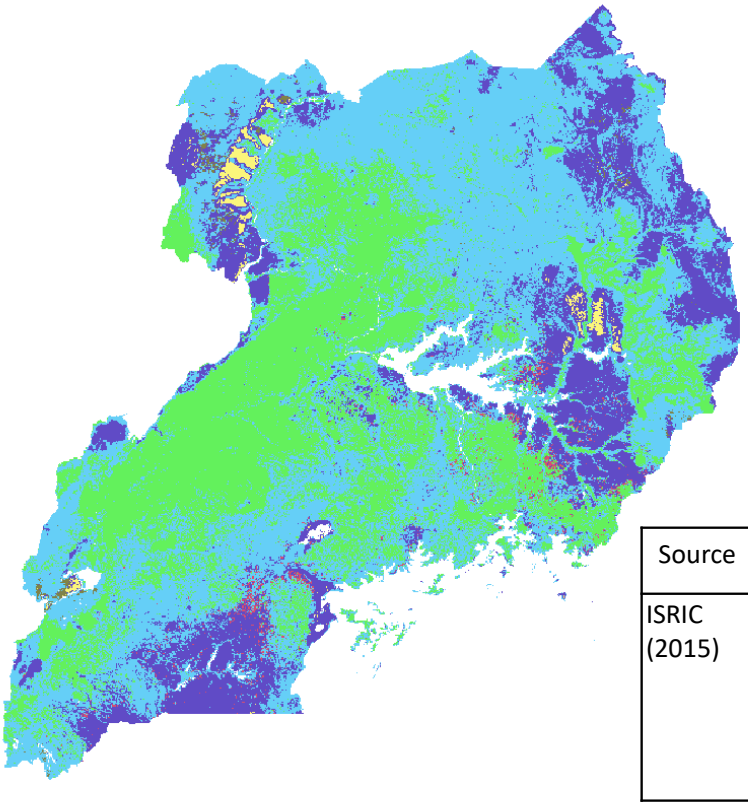
Source	Link
UNHCR (2020)	https://data2.unhcr.org/en/documents/details/85323



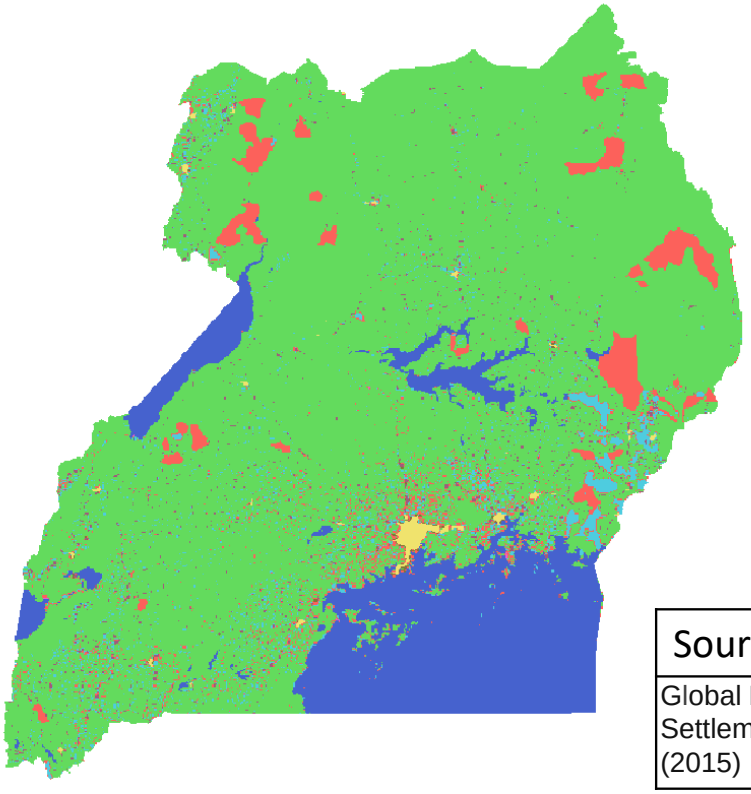
Source	Link
Digital groundwater maps of Africa (2012)	https://www2.bgs.ac.uk/groundwater/international/africanGroundwater/mapsDownload.html



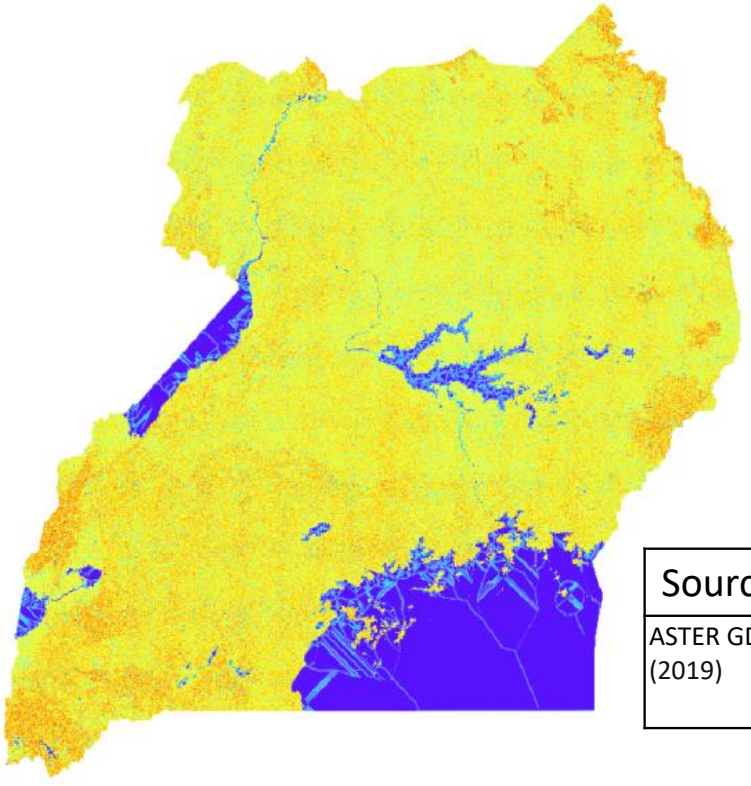
Source	Link
ASTER GDEM Version 3 (2019)	https://earthexplorer.usgs.gov/



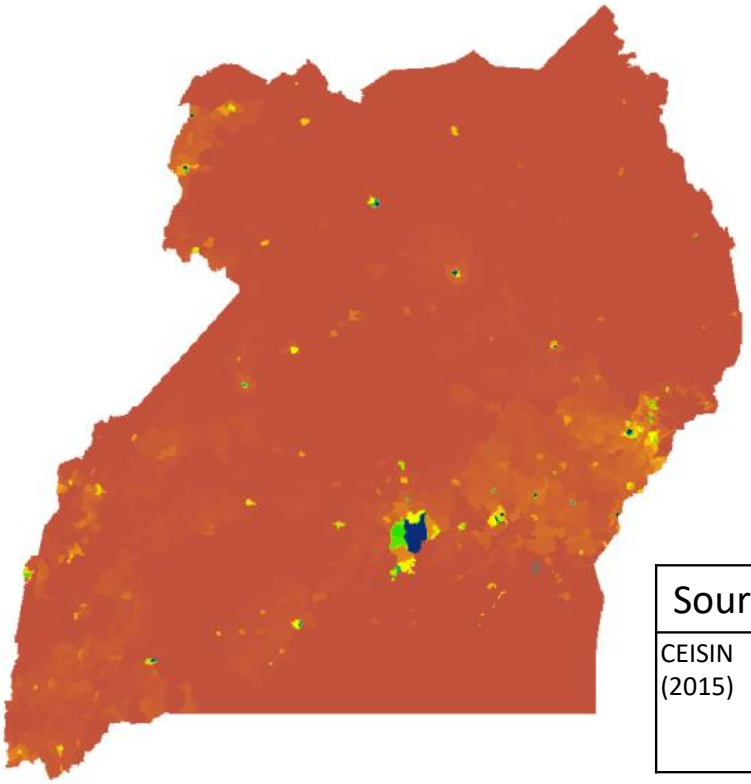
Source	Link
ISRIC (2015)	https://data.isric.org/geonetwork/srv/eng/catalog.search#/metadata/2a7d2fb8-e0db-4a4b-9661-4809865aaccf



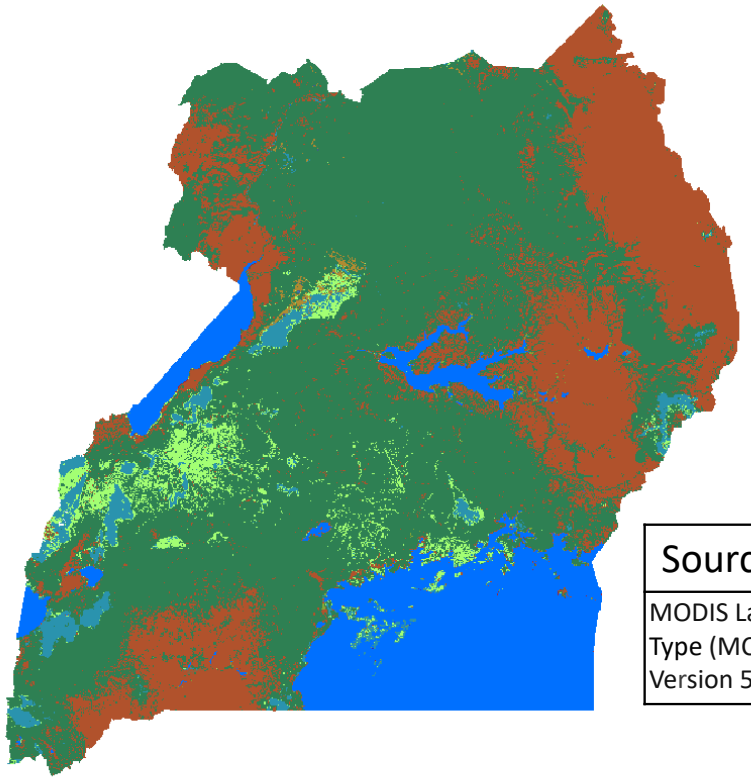
Source	Link
Global Human Settlement Grid (2015)	https://ghsl.jrc.ec.europa.eu/ghs_smod.php



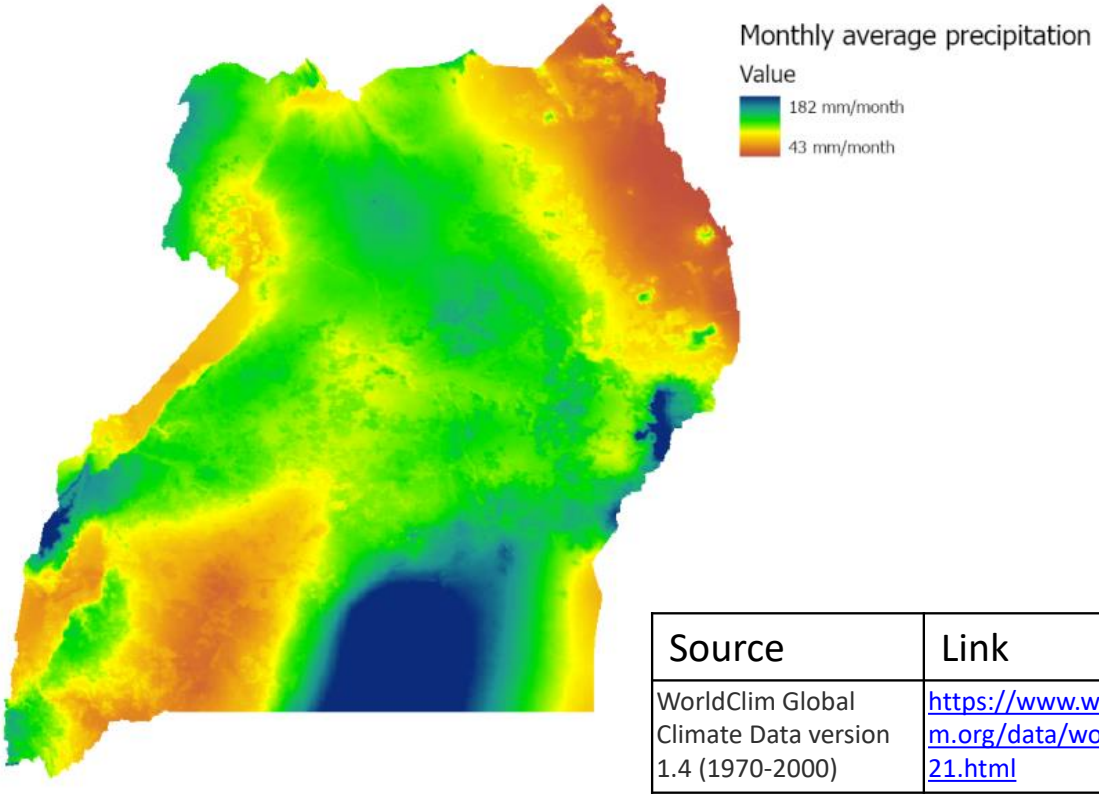
Source	Link
ASTER GDEM Version 3 (2019)	https://earthexplorer.usgs.gov/



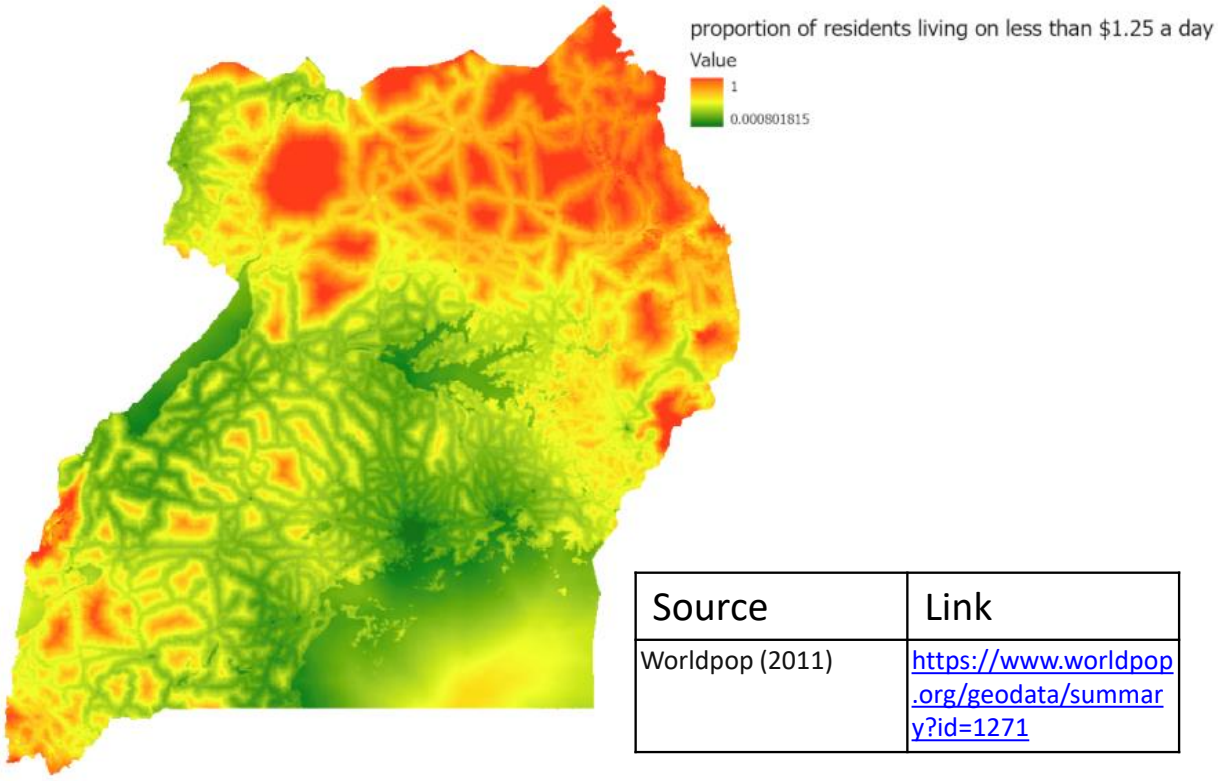
Source	Link
CEISIN (2015)	https://sedac.ciesin.columbia.edu/data/collection/gpw-v4/documentation



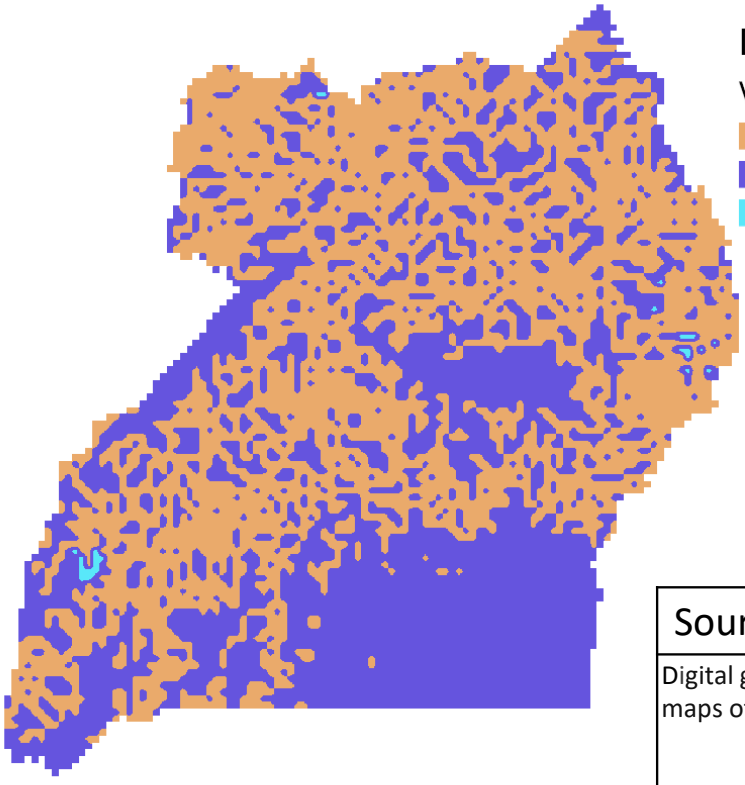
Source	Link
MODIS Land Cover Type (MCD12Ql) Version 5 (2014)	https://modis.gsfc.nasa.gov/data/dataprod/mod12.php



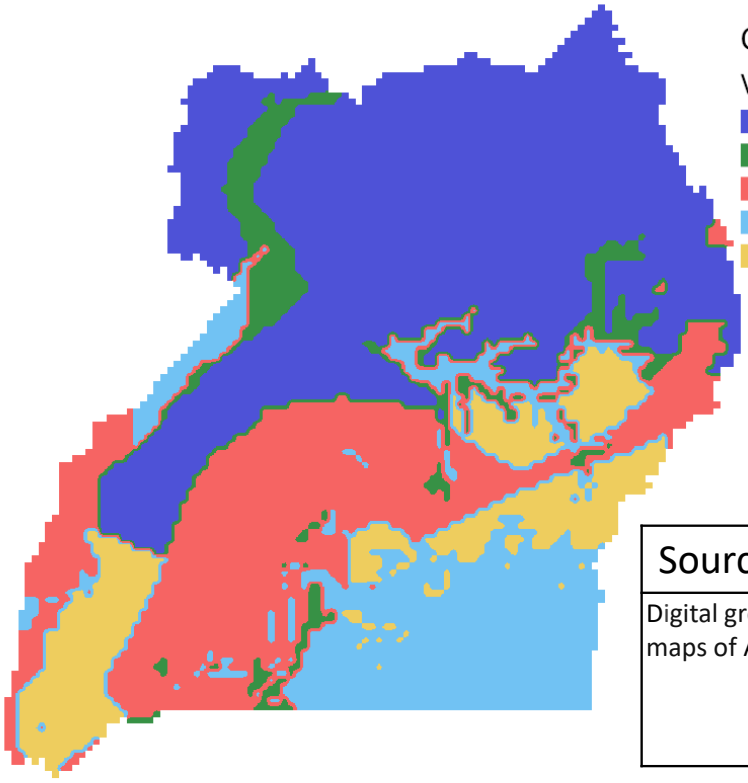
Source	Link
WorldClim Global Climate Data version 1.4 (1970-2000)	https://www.worldclim.org/data/worldclim21.html



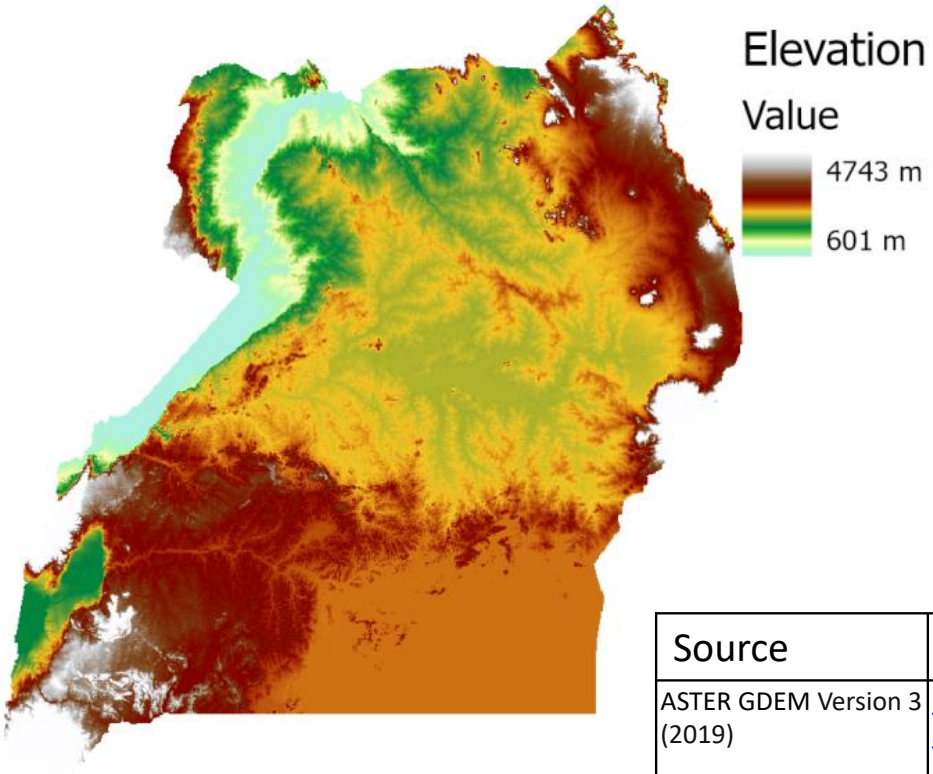
Source	Link
Worldpop (2011)	https://www.worldpop.org/geodata/summary?id=1271



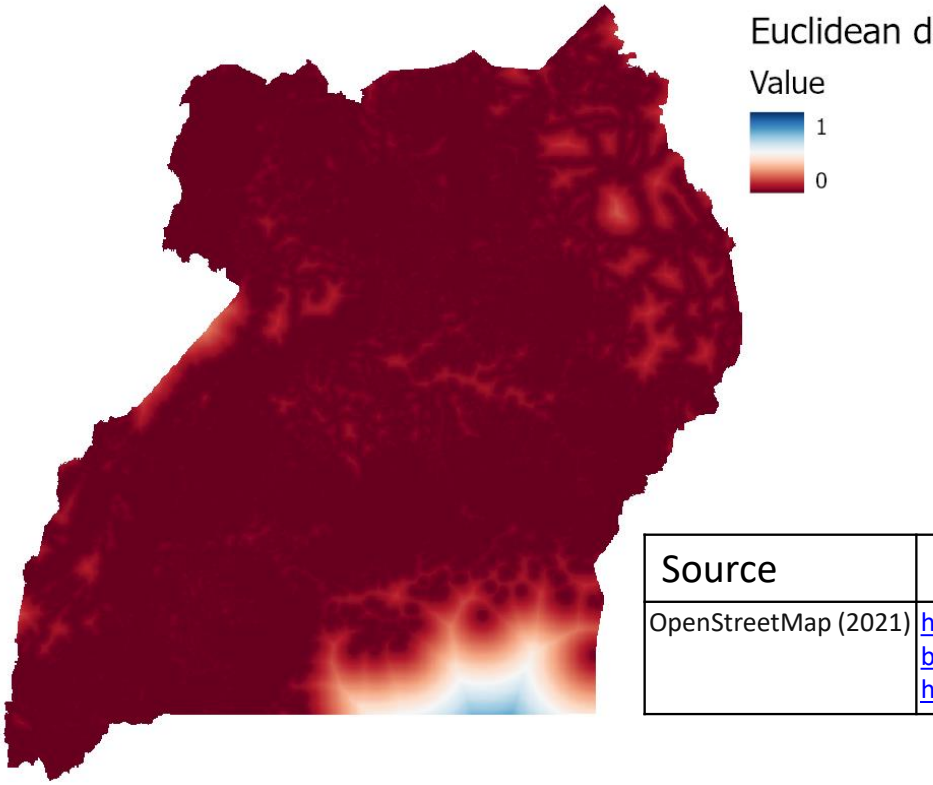
Source	Link
Digital groundwater maps of Africa (2012)	https://www2.bgs.ac.uk/groundwater/international/africanGroundwater/mapsDownload.html



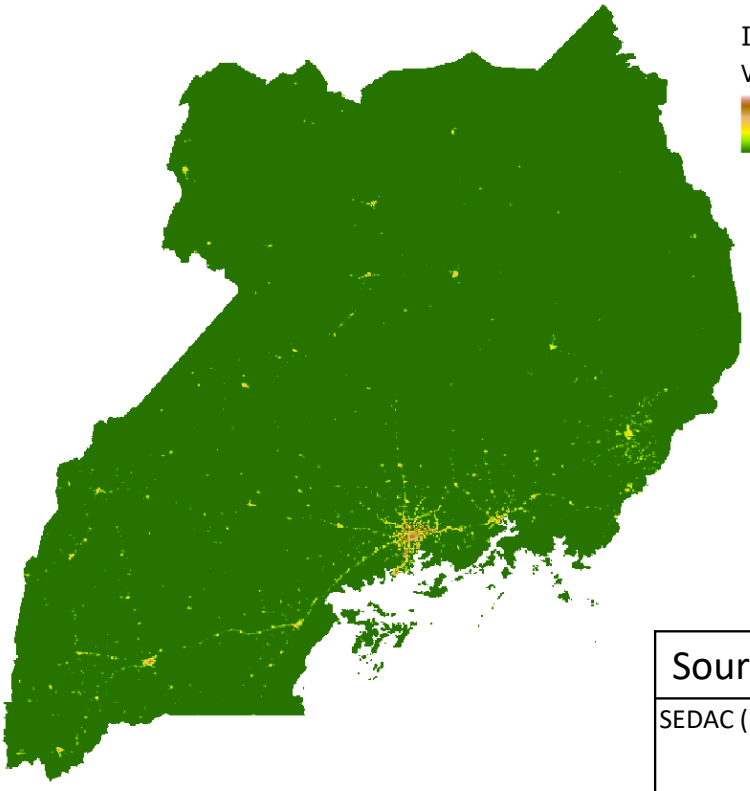
Source	Link
Digital groundwater maps of Africa (2012)	https://www2.bgs.ac.uk/groundwater/international/africanGroundwater/mapsDownload.html



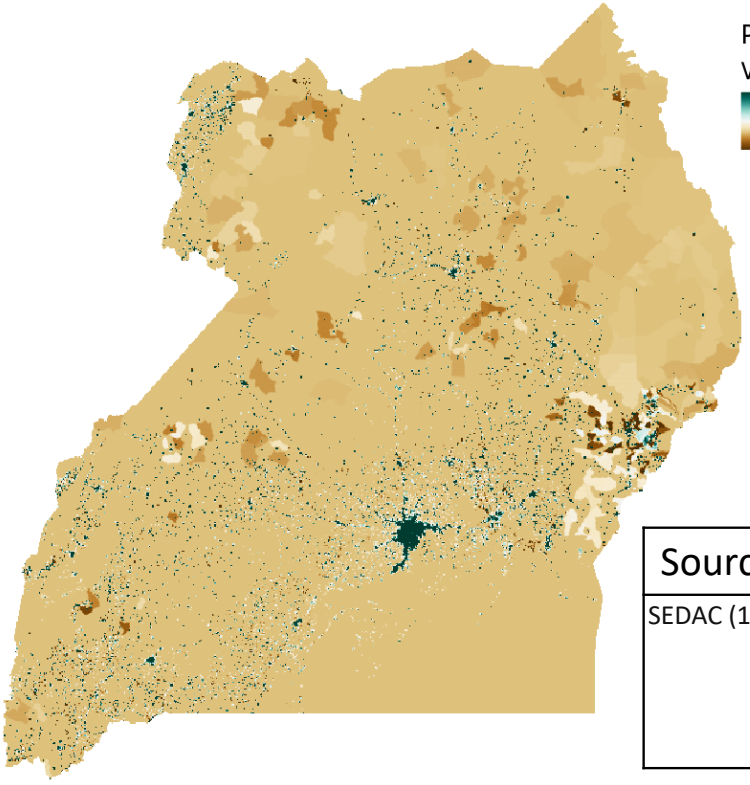
Source	Link
ASTER GDEM Version 3 (2019)	https://earthexplorer.usgs.gov/



Source	Link
OpenStreetMap (2021)	http://download.geofabrik.de/africa/uganda.html



Source	Link
SEDAC (1992-2013)	https://sedac.ciesin.columbia.edu/data/set/sdei-viirs-dmsp-dlight/docs



Source	Link
SEDAC (1990-2014)	https://sedac.ciesin.columbia.edu/data/set/ghsl-population-built-up-estimates-degree-urban-smod

APPENDIX E: ETHICS APPROVALS

Date 13-09-2021
Contact person Ir. J.B.J. Groot Kormelink, secretary HREC
Telephone +31 152783260
E-mail j.b.j.grootkormelink@tudelft.nl

Human Research Ethics Committee
TU Delft
(http://hrec.tudelft.nl/)

Visiting address
Jaffalaan 5 (building 31)
2628 BX Delft

Postal address
P.O. Box 5015 2600 GA Delft
The Netherlands

Ethics Approval Application: Survey water usage in Bushenyi, Uganda
Applicant: Geleijnse, Jan

Dear Jan Geleijnse,

It is a pleasure to inform you that your application mentioned above has been approved.

Thank you for your application and the additional document. Your submission has been approved on the condition that the Ugandan Ethics committee approves as well.

Good luck with your research!

Sincerely,

Dr. Ir. U. Pesch
Chair HREC
Faculty of Technology, Policy and Management

MAKERERE UNIVERSITY

P .O. Box 7062,
Kampala, Uganda
Cables: MAKUNIKA

Tel: 256-41-545040/0712 207926
Fax: 256-41-530185
E-mail: makssrec@gmail.com

COLLEGE OF HUMANITIES AND SOCIAL SCIENCES
SCHOOL OF SOCIAL SCIENCES
RESEARCH ETHICS COMMITTEE

Your Ref:

Our Ref: MAKSSREC 09.21.506/AR

9th November 2021

Jan Geleijnse
Principal Investigator (MAKSSREC 09.21.506)
TU Delft, Water management
Claes de Vrieselaan 75A, 3021JD Rotterdam, the Netherlands
Telephone contact: +31650121977
Email: j.geleijnse@student.tudelft.nl

Dear Sir,

Initial Review – Regular
Re: Approval of Protocol titled: "Surveying water usage in Bushenyi "
This is to inform you that, the Makerere University School of Social Sciences Research Ethics Committee (MAKSS REC) granted approval to the above referenced study. The MAKSS REC reviewed the proposal using the full board review on **30th September 2021**. This has been done in line with the investigator's subsequent letter addressing comments and suggestions.

Your study protocol number with MAKSS REC is **MAKSSREC 09.21.506.** Please be sure to reference this number in any correspondence with MAKSS REC. Note that, the initial approval date for your proposal by **MAKSS REC was 30th September 2021**. This is an annual approval and therefore; approval expires on **29th September 2022**. **Please note that, final approval should be done by Uganda National Council for Science and Technology. You should use stamped consent forms and study tools/instruments while executing your field activities at all times.** However, continue d approval is conditional upon your compliance with the following requirements.

Continued Review
In order to continue on this study (including data analysis) beyond the expiration date, Makerere University School of Social Sciences (MAKSS REC) must re-approve the protocol after conducting a substantive meaningful, continuing review. This means that you must submit a continuing report
Form as a request for continuing review. To avoid a lapse, you should submit the request six (6) to eight (8) weeks before the lapse date. Please use the forms supplied by our office.

Makerere University School of Social Sciences ★ Research Ethics Committee ★
Valid Through
29 SEP 2022
APPROVED

Please also note the following:

- No other consent form(s), questionnaires and or advertisement documents should be used. The Consent form(s) must be signed by each subject prior to initiation of my protocol procedures. In addition, each research participant should be given a copy of the signed consent form.

Amendments

During the approval period, if you propose any changes to the protocol such as its funding source, recruiting materials or consent documents, you must seek Makerere University School of Social Sciences Research and Ethics Committee (MAKSS REC) for approval before implementing it.

Please summarise the proposed change and the rationale for it in a letter to the Makerere University School of Social Sciences Research and Ethics Committee. In addition, submit three (3) copies of an updated version of your original protocol application- one showing all proposed changes in bold or "track changes" and the other without bold or track changes.

Reporting

Among other events which must be reported in writing to the Makerere University School of Social Sciences Research and Ethics Committee include:

i. Suspension or termination of the protocol by you or the grantor.
ii. Unexpected problems involving risk to participants or others.
iii. Adverse events, including unanticipated or anticipated but severe physical harm to participants.

Do not hesitate to contact us if you have any questions. Thank you for your cooperation and commitment to the protection of human subjects in research.

The legal requirement in Uganda is that, all research activities must be registered with the National Council for Science and Technology. The forms for this registration can be obtained from their website https://nrims.uncst.go.ug

Please contact the Administrator of Makerere University School of Social Sciences Research and Ethics Committee at makssrec@gmail.com OR bijulied@yahoo.co.uk or telephone number +256 712 207926 if you encounter any problem.

Yours sincerely,

Makerere University School of Social Sciences ★ Research Ethics Committee ★
Valid Thru.
29 SEP 2022
APPROVED

Dr. Stella Neema
Chairperson
Makerere University School of Social Sciences Research and Ethics Committee

c.c.: The Executive Secretary, Uganda National Council for Science and Technology

APPENDIX F: SURVEY QUESTIONS

Surveying water usage in Bushenyi – Ishaka municipality

Annex V: Survey questions

Researchers:
Prof. Frank Kansiime,
Department of environmental Management,
P.O Box 7062, Makerere University, Uganda
Tel- +256772-506520, Email- fkansiime@gmail.com

James Tayebwa Bamwenda,
School of Social Sciences Library,
P.O Box 7062, Makerere University, Uganda
Tel- +256779870067, Email- jamestayebwa1@gmail.com

Ass. Prof. Edo Abraham,
Water Management, TU Delft, the Netherlands
TU Delft Faculty of Civil Engineering and Geosciences, Stevinweg 1, 2628 CN Delft
Tel- +31 (0) 15 27 82 227, Email- E.Abraham@tudelft.nl

Jan Geleijnse,
Water Management, TU Delft, the Netherlands
Claes de Vrieselaan 75A, 3021JD, Rotterdam
Tel- +31 6 50 12 19 77, Email- j.geleijnse@student.tudelft.nl

Date: October 20 – 2021

Water Usage In Bushenyi Survey_ Questionnaire	
Question	**Answer choices**
Introduction	
All the information you provide is confidential and your name will not be disclosed anywhere. The results will be treated anonymously. Participation in this study is voluntary. You don't have to take part if you don't want to. You don't have to answer any question you don't want to, and you can stop the interview at any time. If you decide not to participate there will not be any negative consequences. Do you have any questions? Do you agree to participate in this study? If you have any further questions you can contact Prof. Kansiime Frank from Makerere University Department of Environment and Management at +256 772 506520 or +256 752 506520.	
Administer informed consent. If subject agrees to participate, proceed to questionnaire	yes, no
Location of Household (GPS)	Survey Collector can register this
Personal Information	
Gender of survey respondent	Female,Male
What is the gender of the household head?	a. Male – the respondent is head b. Male – not the respondent c. Female – the respondent is head d. Female – not the respondent
How many people are in your household?	1,2,3,4,5,More than 5
What is your highest level of education completed?	• Primary education Basic level • Secondary education – Ordinary level • Secondary education - Advanced level • Tertiary education – Vocational college • Tertiary education – University degree
Are you able to read or write?	Yes No
Marital status (if age is >15)	1. Married 2. Living together 3. Unmarried/Single 4. Divorced 5. Separated 6. Widowed

Main occupation (if age is >15)	1. Agriculture/Livestock/Herding 2. Handicrafts(Weaving)/Carpenter/Mason/Blacksmith Trader/Merchant (retail/ wholesale)/Food vendor \ 3. Formal Employment 4. Casual Employment 5. Unemployed 6. Retired 99. Other (specify) _____

How many people live and eat in your household on a regular basis? (Age 0 – 2, Age 3 – 5, Age 5 – 18, Age 18 or older)	Three columns: Age/Male/Female Four additional rows for age breakdown: (Age 0 – 2, Age 3 – 5, Age 5 – 18, Age 18 or older)

Socio-economic status	
How much land does your family own (in Acres)?	
How much land do you use to cultivate vegetables or staple crops?	
How many of each of these animals does your household currently own?	1. Large stock (Cattle) 2. Small stock (Goats/sheeps) 3. Pigs 4. Poultry
Does anyone in your household own the following items? (Select one or more)	1. Radio 2. Mobile phone 3. Television 4. Car or truck 5. Hand cart 6. Animal 7. Cart 8. Generator 9. Tractor 10. Bicycle 11. Motorcycle 12. Solar Panel
Does the household have an electricity connection?	yes, no
How does your household afford to put food on the table?	Food/Income from Agriculture Salary Grants Bartering Allowances
Approximately how much income do you receive from each of these sources per month during the dry season? (In Shs)	
How many women in your household earn some income?	
About what share of household income is earned by women?	Percentage
Does your household receive any money from other family members living outside of this village (remittances)?	yes, no
About how much money do you receive from them?	
If you wanted to take out a loan of 10,000 Shs, from someone other than household members, could you do so?	yes, no

And in the past year, about how much did your family spend on each of the following things?	1. Weddings 2. Funerals 3. Baptisms 4. Health Care (Household) 5. School fees (household)

Waterborne disease, disability and health awareness	
Has anyone in your household had diarrhea or respiratory illness in the last 30 days?	
(If yes) How many had: Diarrhea (enter number): Matrix for gender / age group: 0-5; 5-18; 18-60; 60+ Respiratory illness (enter number): Matrix for gender / age group: 0-5; 5-18; 18-60; 60+	
Which health facility is most often used by your family for health services?	Kampala International Hospital (Ishaka), Bushenyi Health Center II, Other..
Have you ever heard about water borne diseases in this area?	
If yes in the above, which disease was it?	
Have you yourself had diarrhea in the past 2 weeks?	None, Once, Twice, 3 times, >3 times
For any of these occurrences of diarrhea, did you seek advice or treatment from any source?	Yes, hospital or health centre Yes, shop of pharmacy Yes, traditional healer No
How much did you spend in total on treatment for recent diarrhea and/or respiratory illness(es) for your household, in each of these categories: Medical fees, medicines, transport to facility, Did nothing, Other	Medical fees, medicines, transport to facility, Did nothing, Other
Do any household members have disabilities?	Yes, no (skip)
How would you describe the main disability of the most disabled HH member?	01 – Hearing impairment 02 – Deafness 03 – Visual impairment 04 – Blind 05 – Mobility impairment 06 – Housebound 07 – Upper limb impairment 08 – Speech impairment 09 – Learning difficulties 10 – Mental impairment

Household water source/supply	
Can you list all household water sources for both drinking and nondrinking water?	1. NWSC/Piped 2. Borehole 3. Shallow Wells

	4. Springs 5. Surface water/Rivers/stream/lake/pond/dam 6. Rain water 7. Others (Specify)------
What does your household use this water for? (answer for **each** of the listed sources from previous question) Choose one or multiple	1. Drinking 2. Animal use 3. Watering crops 4. Cooking 5. Washing 6. Others (Specify)…….
How much water does your household use on average per day throughout the year? (specify in 20 Ltr Jerry cans)	Number of 20 Ltr Jerrycans:……..
How much time does your household spend on water collection per day? (**Sum** of ALL round trip(s), incl. queuing, to **ALL** sources listed)	(a) Hours …… (b) Minutes ……
How would you rate the cost of the water for your household?	1. Very cheap 2. Inexpensive 3. Cost-appropriate 4. Expensive o Very expensive 5. DK/ No comment
Who usually fetches water in the household?	a) Man b) Woman c) Children d) House girl/Shamba boy e) Relatives f) Others (Specify)…..
What is his/her age?	Age
Is that person a man or a woman	Man / Woman
Which of the sources listed is the source that you took the most water from during the last month? (= primary source)	1. NWSC/Piped 2. Borehole 3. Shallow Wells 4. Springs 5. Surface water/Rivers/stream/lake/pond/dam 6. Rain water Others (Specify)------
INSERT GENERAL WATER SOURCE QUESTIONS (SEE BELOW IN YELLOW)	Questions about primary source
Is the primary source also the source closest to your home?	Yes / No, if **yes skip** the closest source section
Closest source	
What is the reason that your closest source is not your most used source?	(Allow multiple responses) 1=The closest water source is unsafe for drinking 2=The closest water source is not free or more expensive

	3=Excluded by Water User Committee (WUC) for closest source 4=Excluded by user group for closest source 5=I don't like the taste of the closest source 6= Closest water source dysfunctionality 99= Other (specify)_____
Is the closest source your drinking water source?	Yes/No/Sometimes
INSERT GENERAL WATER SOURCE QUESTIONS (SEE BELOW IN YELLOW)	
Is your drinking water source the same as your primary source or closest source?	a) Same as primary b) Same as closest c) Same as both d) No If no, continue to drinking water source section. Else: skip drinking water section.
Drinking water source	
(If applicable) What is the reason that your drinking water source is not your closest source?	(Allow multiple responses) 1=The closest water source is unsafe for drinking 2=The closest water source is not free or more expensive 3=Excluded by Water User Committee (WUC) for closest source 4=Excluded by user group for closest source 5=I don't like the taste of the closest source 6= Closest water source dysfunctionality 99= Other (specify)_____
(if applicable) What is the reason that your drinking water source is not the same as your most used source? I.e. why do you drink from another source than the one that you use most?	(Allow multiple responses) 1=The most used water source is unsafe for drinking 2=The most used water source is not free or more expensive 3 = The drinking water source I use, is better tasting 99 = Other (specify) _____
INSERT GENERAL WATER SOURCE QUESTIONS (SEE BELOW IN COLOR)	Questions about drinking water source
In the last four weeks, how often did it happen that you wanted to drink water, but you forgot to treat it in time?	In the last four weeks, how often did it happen that you wanted to drink water, but you forgot to treat it in time?
GENERAL WATER SOURCE QUESTIONS	This section should be inserted at all places indicated (max 3 times)
Select section: (for administration purposes, enumerator can fill this question in)	Primary source / closest source / drinking water source

Geo-reference: water source	Enumerator to record GIS coordinates
Is that source available during the entire year?	(a) No only in the wet season (b) No, only in the dry season (c) No, it is often broken (d) Yes
Is water from this source safe to drink?	a) Safe to drink b. Safe to drink after treatment c. Unsafe to drink
How do treat you your water from this source before use or drinking?	a) Boiling, b) Take directly from source c) Using water guard, d) I do not treat the water e) Others (Specify)…..
Is this source protected from - or monitored for potential contamination?	Yes/no
On average, how many 20 ltr jerrycans do you take back home when using this source?	Number of 20 ltr jerrycans

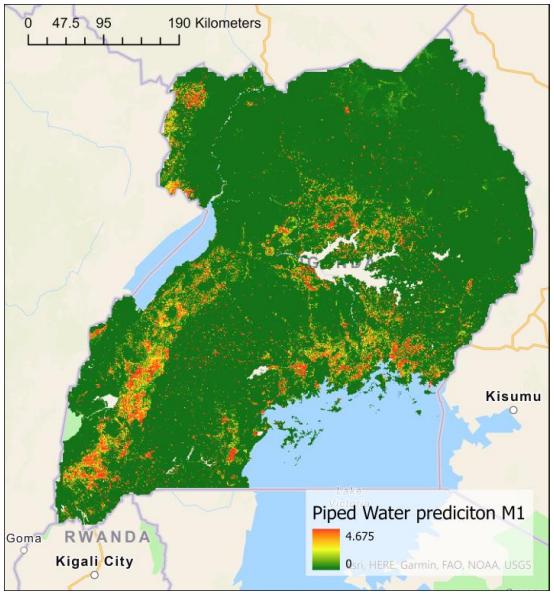
On average, how long does it take to collect water from this water source? (Round trip, incl. queuing)	(a) Hours ……. (b) Minutes …..
How do you travel to this water source?	a. By walking b. By Bicycle c. By Motor bike d. By Car e. Other: _____
Is this water source shared with other households?	a) Yes b) No
Do you pay for water from this water source? If yes, how much do you pay ……. /20ltr jerrycan	a. Price: ____ (fill in total price) b. Nothing c. I don't know
Whom do you pay for water?	1=Local government 2=Utility company Standpipe manager 3=Tanker truck manager 4=Water vendor 5=Neighbor Others (Specify)……
Who is responsible for managing and maintaining the main drinking water source?	1=WUC 2=Community members 3=Person hired by the community 4=NWSC 77=Do not know 99=Other (specify)_____
What restrictions are there, if any, to use this water source?	1= YES, Resource contribution for scheme construction 2=YES, Resource contribution for repair and maintenance 3=YES, Membership in WUC 4= YES, Payment for water use 99= Other (specify)_____ 2=NONE
What is your perception about quality of water from this water source?	1=Very Bad 2=Bad 3=Reasonable 4=Good 5=Very Good 77=Don't know
If very bad/bad, can you explain why you think this?	**Select all that apply:** 1=Water is salty 2=Smells bad 3=Tastes bad 4=Water is muddy

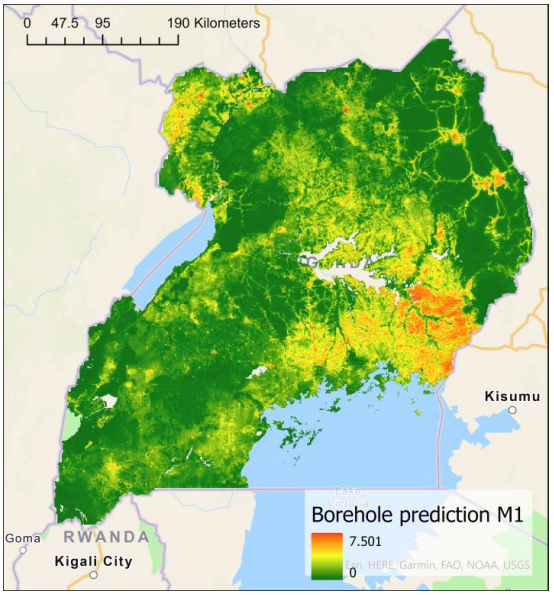
	5=Contaminated by animals 6=Iron taste 99=Other (specify) _____
Has this water source experienced any service interruptions over the past 3 months (interruption = no water available for 12 hours or more)?	a. Yes b. No
(If yes) How long was the service interrupted (enter hours, days or months)?	
(If yes) What was the main cause of the *interruption in service?*	
Is this water source currently functional?	Yes/No
If NOT, why is it not functional?	
Do you think this water source will be operating in 5 years?	
If main drinking water source needed repairs, how confident are you that the problem could be fixed within 1 week?	
In the last 6 months, were there any times when water from main drinking water source was not available for more than one week?	
On a scale of 1 to 10, how do you rate the following aspects of water services: (a) Regular or continuous supply; (b) Cleanliness; (c) Safety (Public Health); (d) Safety (Physical)	Likert scale

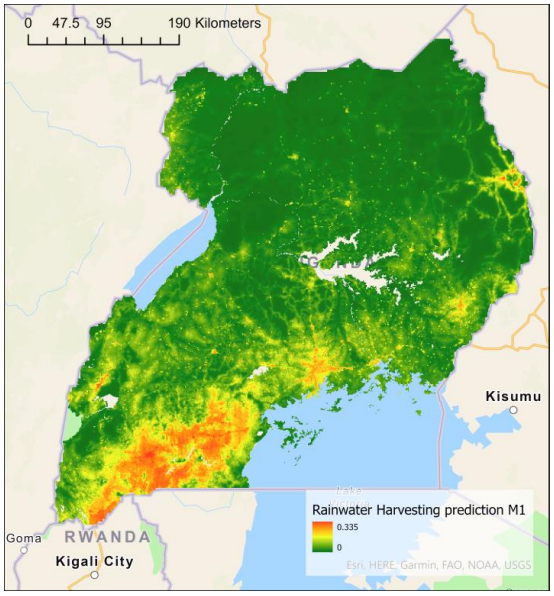
APPENDIX G: OUTPUT OF M1 AND M2 PER ACCESS TYPE

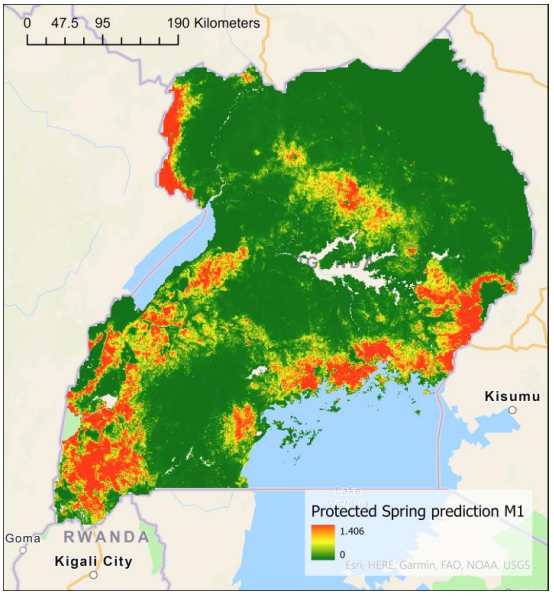
Appendix G.1: M1 output

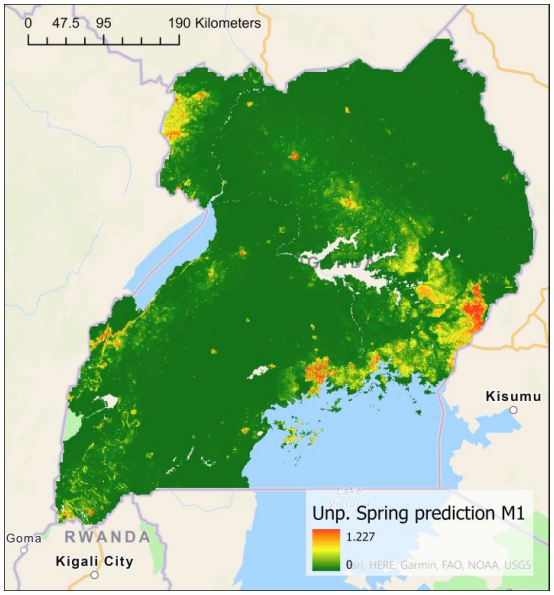
Legends are in terms of M1 predicted presences

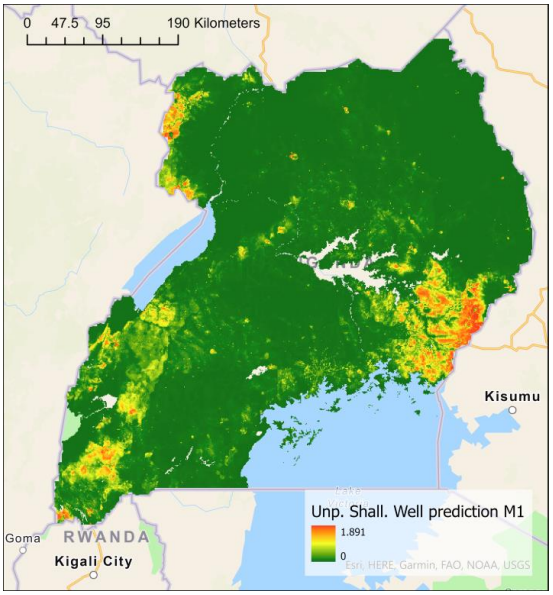


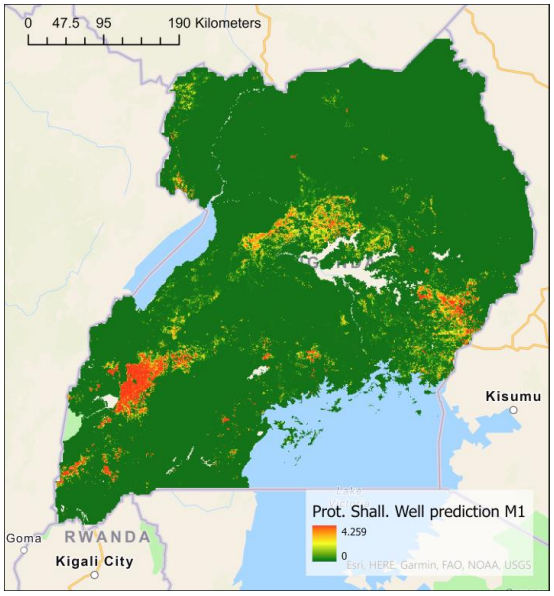


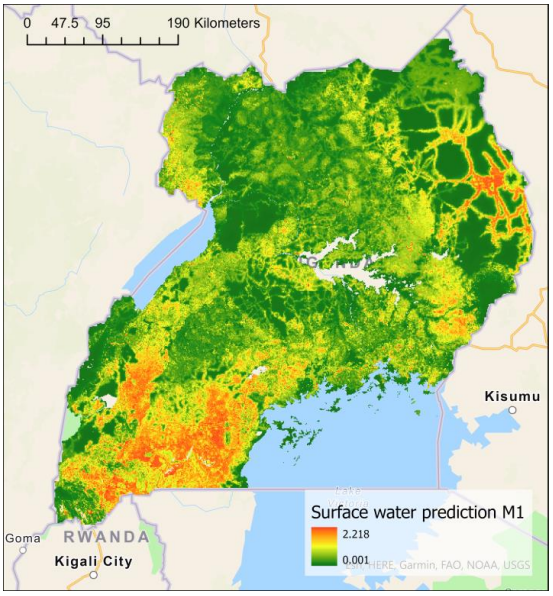





0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Unp. Spring prediction M1
1.227
0
Esri, HERE, Garmin, FAO, NOAA, USGS

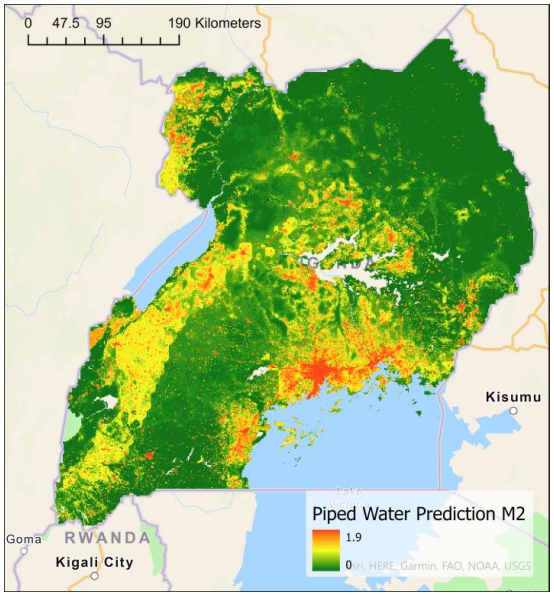

0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Unp. Shall. Well prediction M1
1.891
0
Esri, HERE, Garmin, FAO, NOAA, USGS

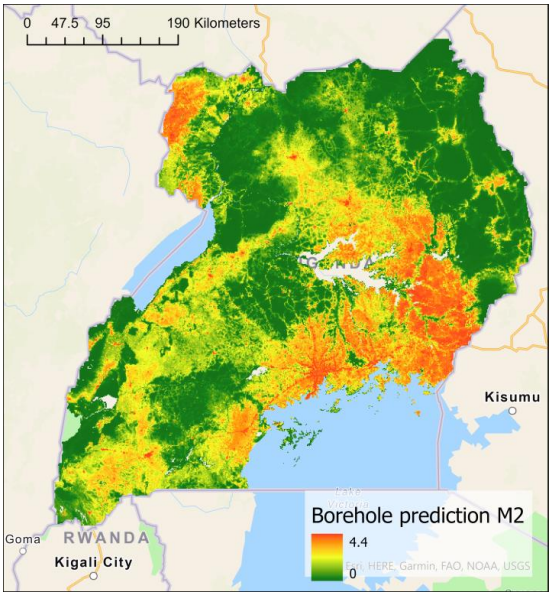

0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Prot. Shall. Well prediction M1
4.259
0
Esri, HERE, Garmin, FAO, NOAA, USGS

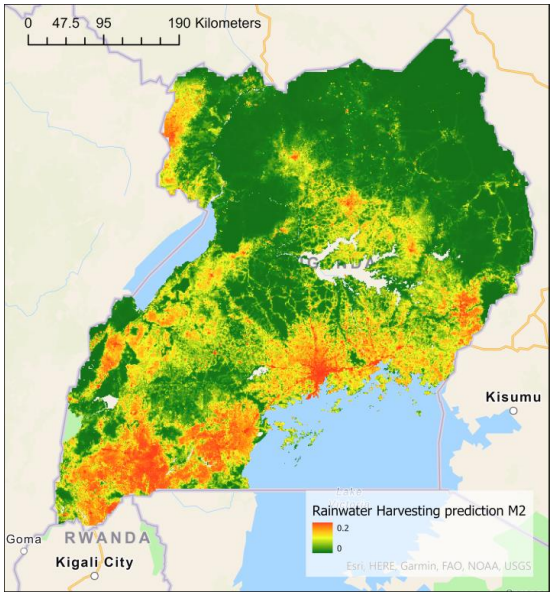

0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Surface water prediction M1
2.218
0.001
Esri, HERE, Garmin, FAO, NOAA, USGS

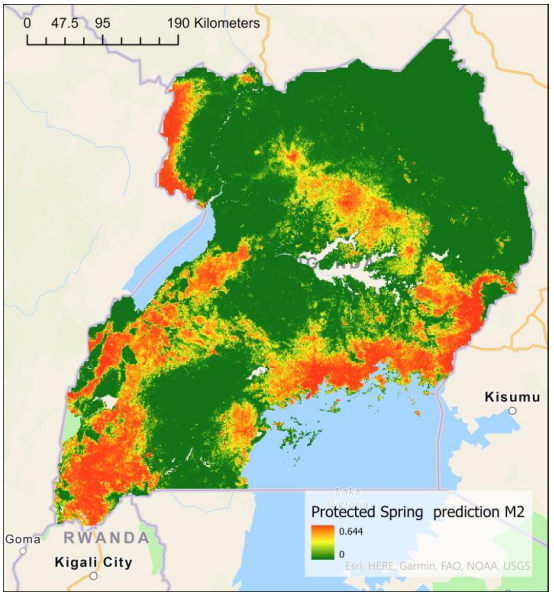
Appendix G.2: M2 output

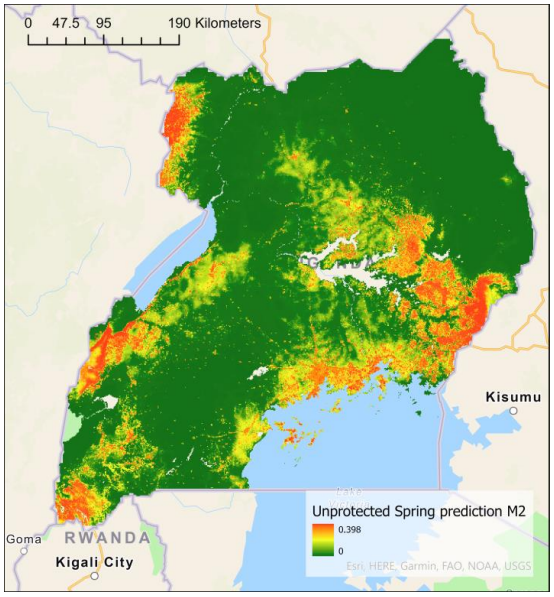
Legends are in terms of M2 predicted presences

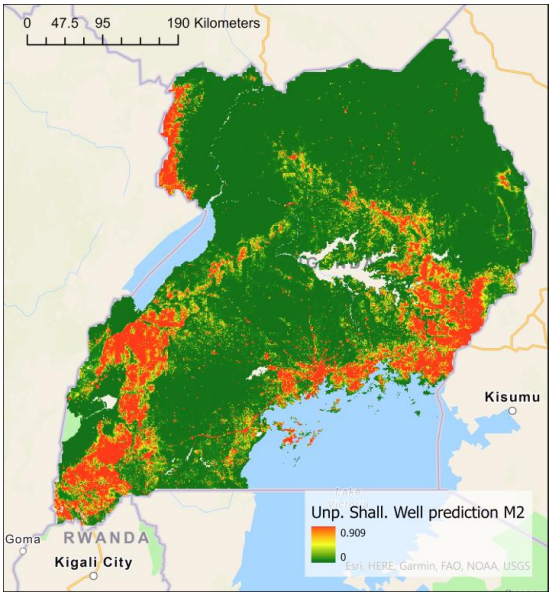


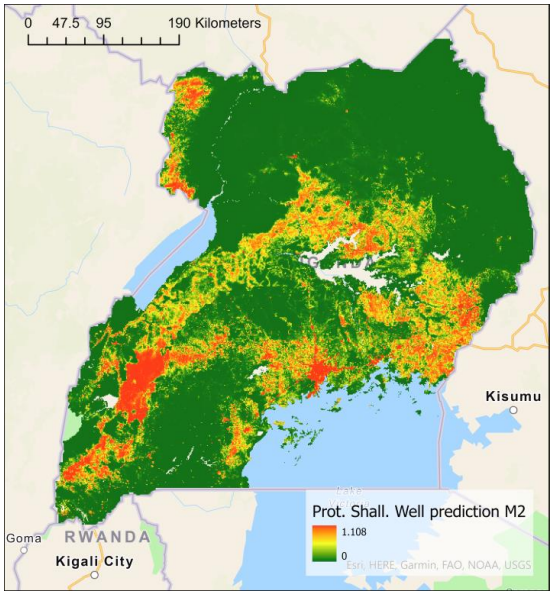


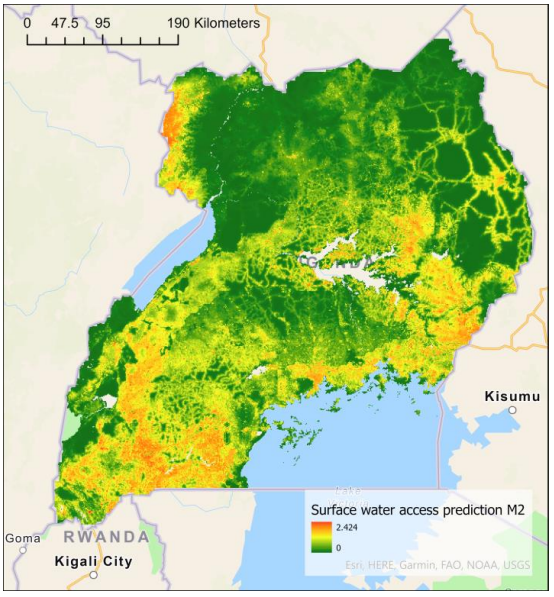





0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Unprotected Spring prediction M2
0.398
0
Esri, HERE, Garmin, FAO, NOAA, USGS

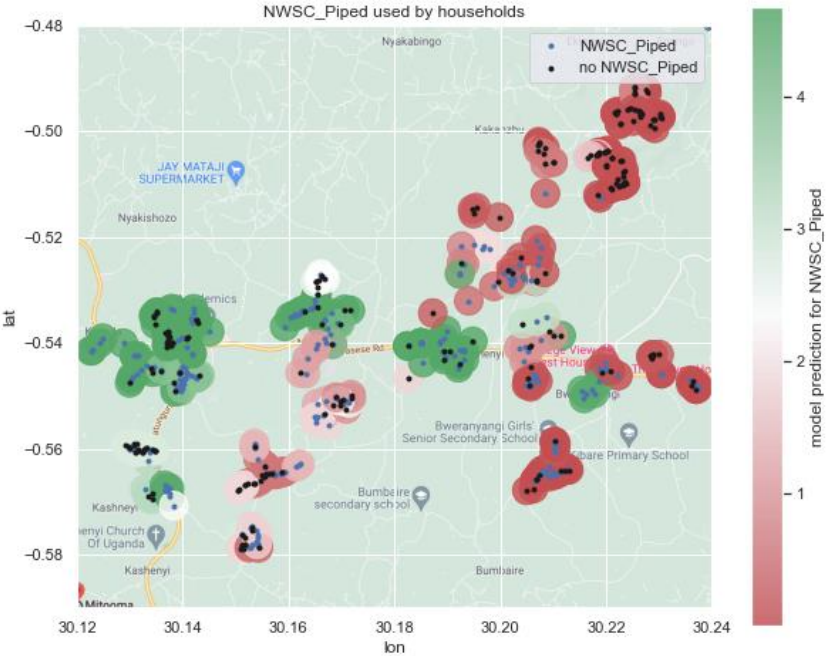

0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Unp. Shall. Well prediction M2
0.909
0
Esri, HERE, Garmin, FAO, NOAA, USGS

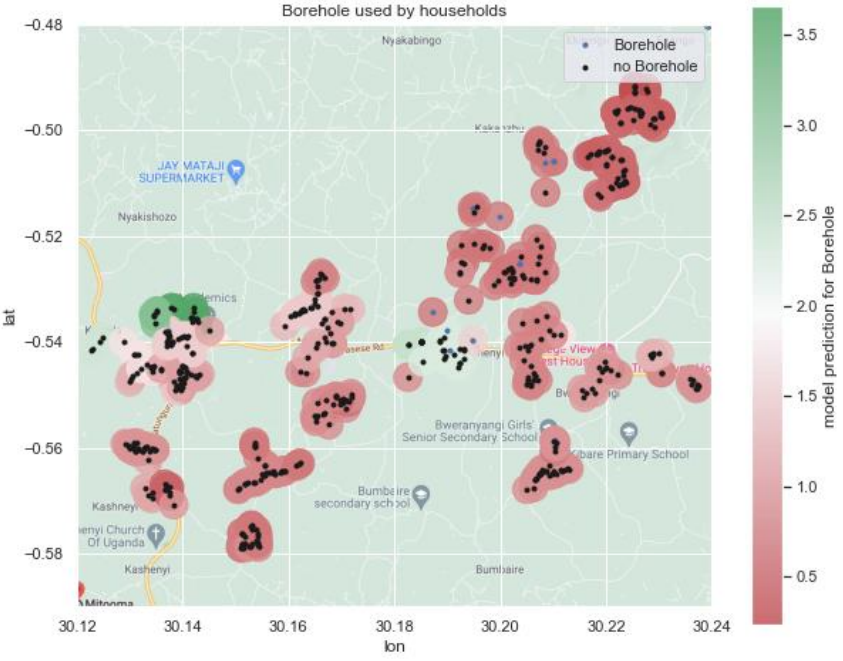

0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Prot. Shall. Well prediction M2
1.108
0
Esri, HERE, Garmin, FAO, NOAA, USGS

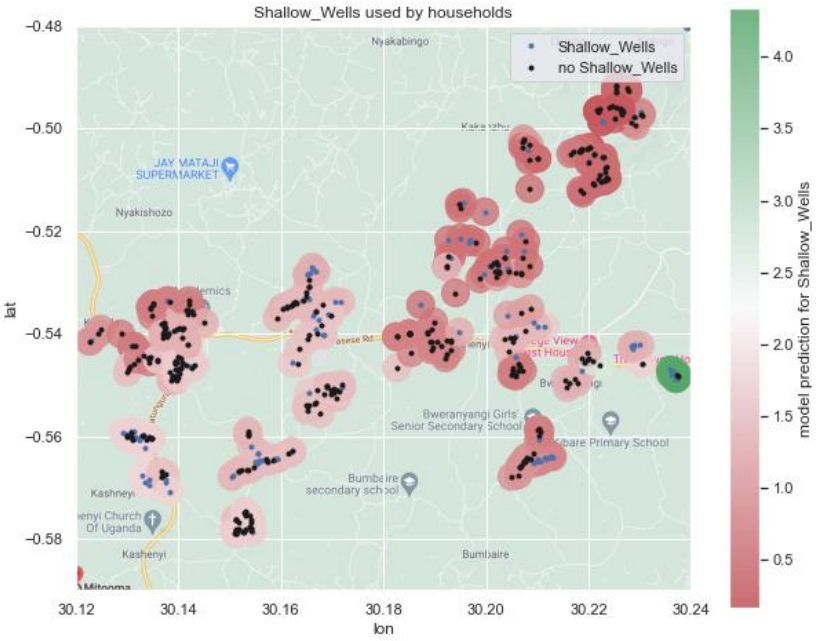

0 47.5 95 190 Kilometers
Kisumu
Goma
RWANDA
Kigali City
Surface water access prediction M2
2.424
0
Esri, HERE, Garmin, FAO, NOAA, USGS

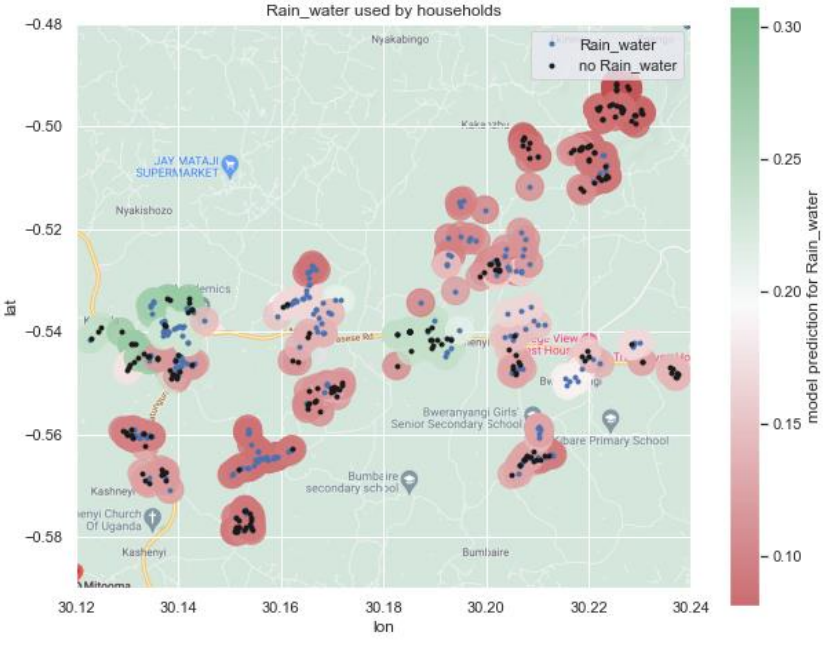
Appendix H: Households having or not having access to access type and model output

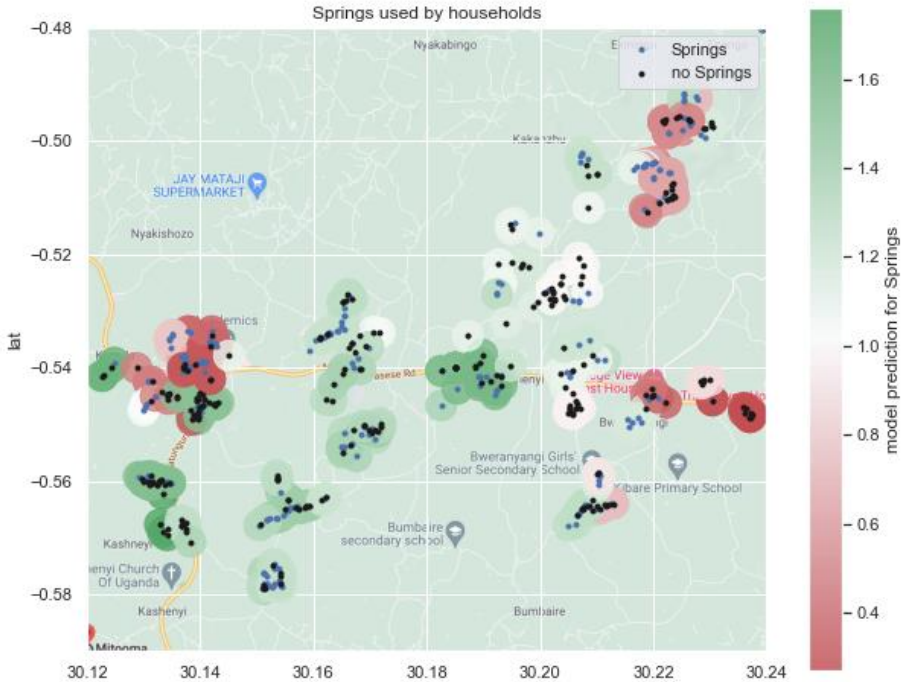
Appendix H. Reported access per household compared to model predictions

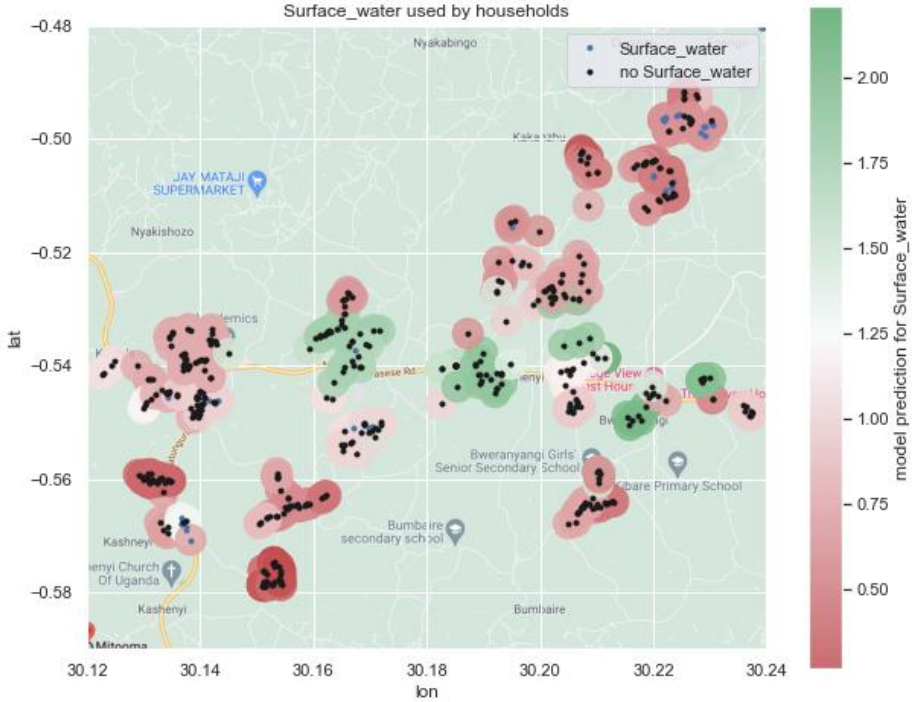





Shallow_Wells used by households
Shallow_Wells
no Shallow_Wells
Nyakabingo
JAY MATAJI SUPERMARKET
Nyakishozo
Bweranyangi Girls' Senior Secondary School
Kibare Primary School
Bumbaire secondary school
Kashenyi
Kashenyi Church Of Uganda
Bumbaire
lat
lon
model prediction for Shallow_Wells
−0.48
−0.50
−0.52
−0.54
−0.56
−0.58
30.12
30.14
30.16
30.18
30.20
30.22
30.24
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5


Rain_water used by households
Rain_water
no Rain_water
Nyakabingo
JAY MATAJI SUPERMARKET
Nyakishozo
Bweranyangi Girls' Senior Secondary School
Kibare Primary School
Bumbaire secondary school
Kashenyi
Kashenyi Church Of Uganda
Bumbaire
lat
lon
model prediction for Rain_water
−0.48
−0.50
−0.52
−0.54
−0.56
−0.58
30.12
30.14
30.16
30.18
30.20
30.22
30.24
0.30
0.25
0.20
0.15
0.10


Springs used by households
Springs
no Springs
Nyakabingo
JAY MATAJI SUPERMARKET
Nyakishozo
Bweranyangi Girls' Senior Secondary School
Bumbaire secondary school
Kashenyi
Kashenyi
Bumbaire
lat
lon
model prediction for Springs
−0.48
−0.50
−0.52
−0.54
−0.56
−0.58
30.12
30.14
30.16
30.18
30.20
30.22
30.24
1.6
1.4
1.2
1.0
0.8
0.6
0.4


Surface_water used by households
Surface_water
no Surface_water
Nyakabingo
JAY MATAJI SUPERMARKET
Nyakishozo
Bweranyangi Girls' Senior Secondary School
Bumbaire secondary school
Kashenyi
Kashenyi
Bumbaire
lat
lon
model prediction for Surface_water
−0.48
−0.50
−0.52
−0.54
−0.56
−0.58
30.12
30.14
30.16
30.18
30.20
30.22
30.24
2.00
1.75
1.50
1.25
1.00
0.75
0.50