Water Point Mapping in Tanzania Using a Machine Learning Approach

Przemyslaw Zientala*[1]

[1]Department of Geography, University of Southampton, UK

January 5, 2017

Summary

Water Point Mapping is an approach whereby locations and characteristics of point water sources are recorded with handheld GPS devices (Jiménez & Pérez-Foguet, 2011). However, data collection is carried out by volunteers and organisations with differing standards, leading to inconsistencies and uneven coverage. WPDx data was analysed and a Machine Learning approach was used to predict the locations of water sources, achieving good accuracy of 0.07 to 0.17 Root Mean Squared Error. The results of this study could be used to predict unmapped water sources, saving time and money. Moreover, models could be used in other disciplines such as geomorphology for exploratory analysis of variable associations.

KEYWORDS: Water point mapping, Tanzania, machine learning, location prediction.

1. Introduction

Water Point Mapping is an approach used for mapping point water sources whereby their locations and characteristics are recorded with handheld GPS devices (Jiménez & Pérez-Foguet, 2011). Gathered data can subsequently be used to measure access to safe drinking water and direct efforts aiming to improve it. However, a major problem is data inconsistency; this is due to the fact that data collection is often carried out by volunteers and organisations, such as local government authorities and NGOs, which have differing standards. Moreover, collaboration between government and NGOs has historically often been unsatisfactory worldwide due to conflicting interests and NGOs' strong commitment to autonomy (Bratton, 1989; Zafar Ullah *et al.*, 2006). This project will investigate whether locations of point water sources can be predicted using a Machine Learning model supplied with potentially predictive covariates. If successful, this approach could prove useful in the following ways:

- Significantly lower the costs and time required for mapping.
- Allow individuals currently engaged in mapping to shift their focus to other activities, such as policy making or improving sanitation, therefore accelerating progress in improving accessibility to safe drinking water.
- Provide a spatial modelling method that could be transferable to other applications requiring prediction from incomplete point patterns
- Contribute to understanding of relationships between water points and environmental variables

* pjz1g14@soton.ac.uk

2. Methods

WPDx (www.waterpointdata.org; WPDx, 2016) and gridded groundwater resource (Bonsor & MacDonald, 2011; MacDonald *et al.*, 2012) data were downloaded and unified. Features used for prediction are briefly summarised in Table 1.

Table 1. Variables used for prediction of water source locations.

Category of covariate	Covariate	Methodology	Source
Environmental	Groundwater depth, storage, productivity	N/A	British Geological Survey
Human-related	Distance to nearest city (proxy for population distribution)	Based on a distance matrix between top 10 Tanzanian cities by population and each water source	WPDx dataset, Google Maps API
	Nearest city		Google Maps API
	Previously built wells	N/A	WPDx dataset
Derived from data	Cluster ID	Based on k-means clustering with value of *k* chosen using a scree plot	
	Cluster size		
	Year of source installation ≥1992	Indicator variables derived directly from data	
	Depth to groundwater ≤50 m		

To obtain realistic accuracy measures (Hastie *et al.* 2013) data were partitioned as follows (% of all rows):

- Training set: 64%
- Validation set: 16% (20% of training set)
- Test set: 20%

Thus, 80% (training and validation sets) of original data was used for model development purposes. Models used for location prediction are outlined in Table 2.

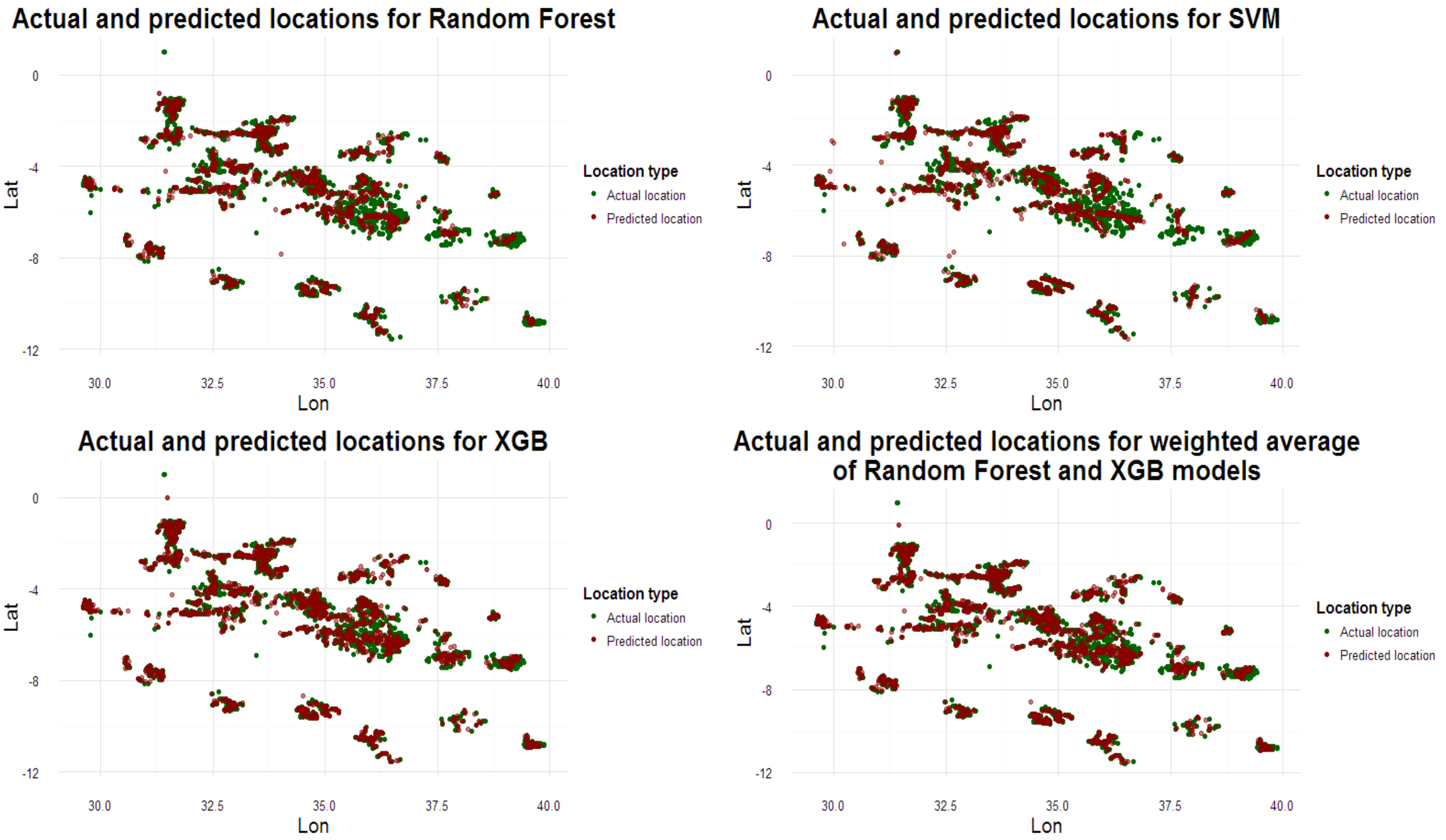
Table 2. ML models used for prediction of coordinates.

Model	Hyperparameters tuned	Minimised objective function
Random Forest (RF)	• *mtry* – number of variables randomly sampled at each split • *ntree* – number of trees to grow • *nodesize* – minimum number of observations in each terminal node	$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \bar{y}_{R_j})$
Support Vector Machine (SVM)	• *kernel* – a "similarity measure" that is used in learning patterns in data • *gamma* – a measure of influence of a single training example • *cost* – parameter determining the number of possible violations of the margin, for C = 0 no violations are allowed	$\frac{1}{2}\lVert\boldsymbol{w}\rVert^2 + C\sum_i^N \xi_i$ subject to $y_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} + b) \geq 1 - \xi_i$ $\xi_i \geq 0$ $i = 1,2,\ldots,N$
Extreme Gradient Boosted trees (XGBoost/XGB)	• *eta* – learning rate • *max_depth* – maximum depth of a tree • *colsample_bytree* – subsample ratio of columns used to construct each tree • *colsample_bylevel* – subsample ratio of columns used for each split • *min_child_weight* – minimum number of observations in each terminal node • *subsample* - fraction of data instances used to grow each tree • *lambda* – L2 regularisation term used to control model bias • *nrounds* – number of trees to grow	$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \bar{y}_{R_j})$
Weighted arithmetic mean of predictions (linear combination of RF and XGBoost predictions)	• *weights for models' predictions* – weights in the range [0,1] determining the relative importance of each model's predictions	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i - \hat{y}_i)^2}$

Each coordinate (latitude, longitude) was predicted separately. Furthermore, each type of water source (out of: shallow wells, machine-drilled boreholes, hand-drilled tube wells, springs) was predicted independently and "aggregate" models were also trained on all types of water sources.

Due to small number of RF parameters, they were chosen empirically. A pairwise hyperparameter grid search was performed for XGB. This was favoured instead of simultaneous optimisation of all parameters due to computational costs. Model performance was then evaluated using Root Mean Square Error (RMSE) and area under empirical cumulative distribution function curve (AUC). Moreover, variable importance measures were calculated for Random Forest and XGB models, indicating the relative contributions of features to RMSE reduction.

3. Results



Actual and predicted locations for Random Forest

Actual and predicted locations for SVM

Actual and predicted locations for XGB

Actual and predicted locations for weighted average of Random Forest and XGB models

Figure 1. Prediction maps for “aggregate” models.

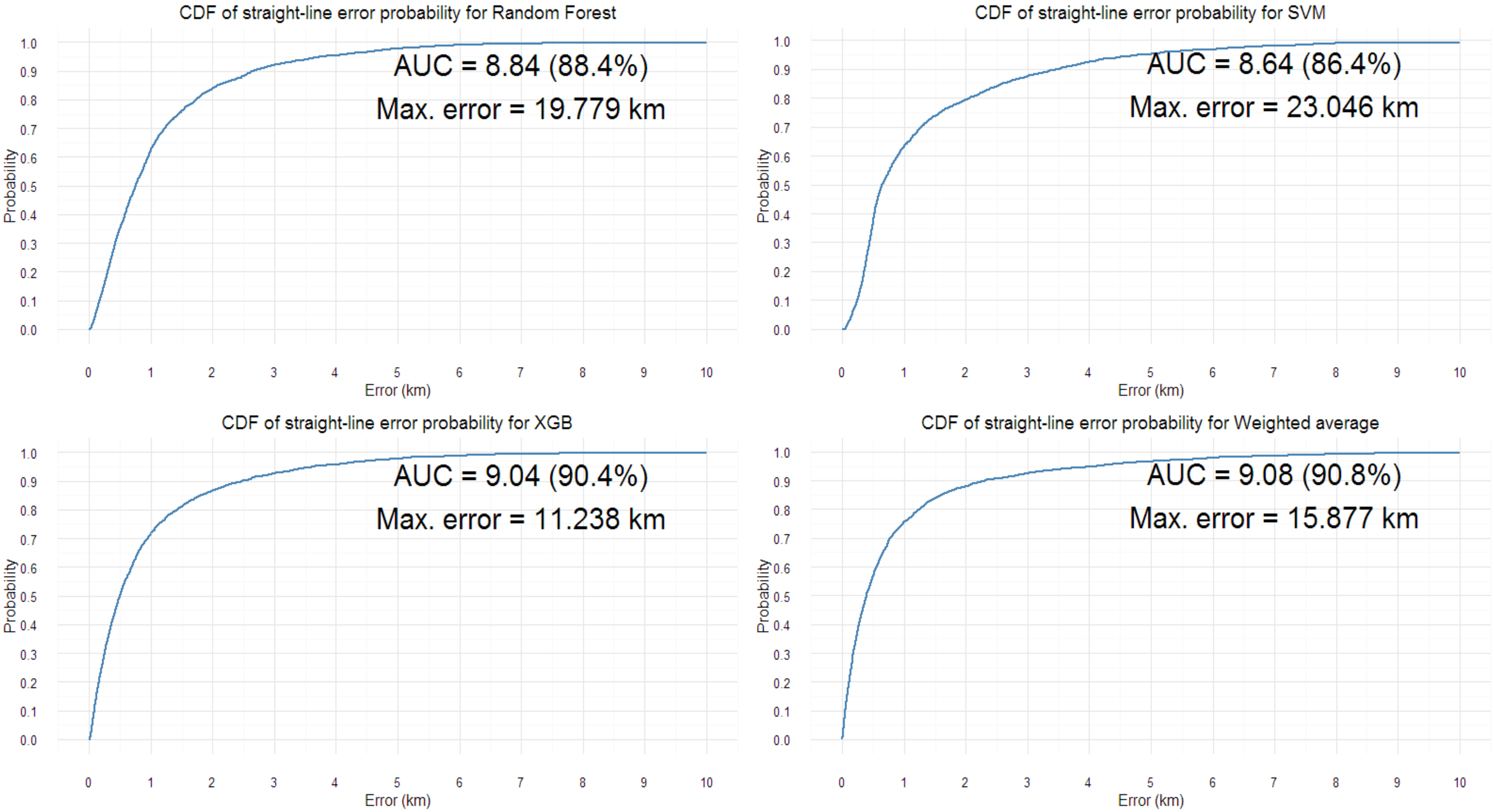


Figure 2. CDF's of error probabilities for "aggregate" models. Percent in brackets indicate what fraction of "perfect AUC" the actual AUC is, that is: $\frac{\mathrm{AUC}}{\mathrm{A}} * 100\%$, where $\mathrm{A} = \mathrm{ab}$, $\mathrm{a} = \mathrm{x}_{\mathrm{max}}$ (= 10 in this example), $\mathrm{b} = 1$.

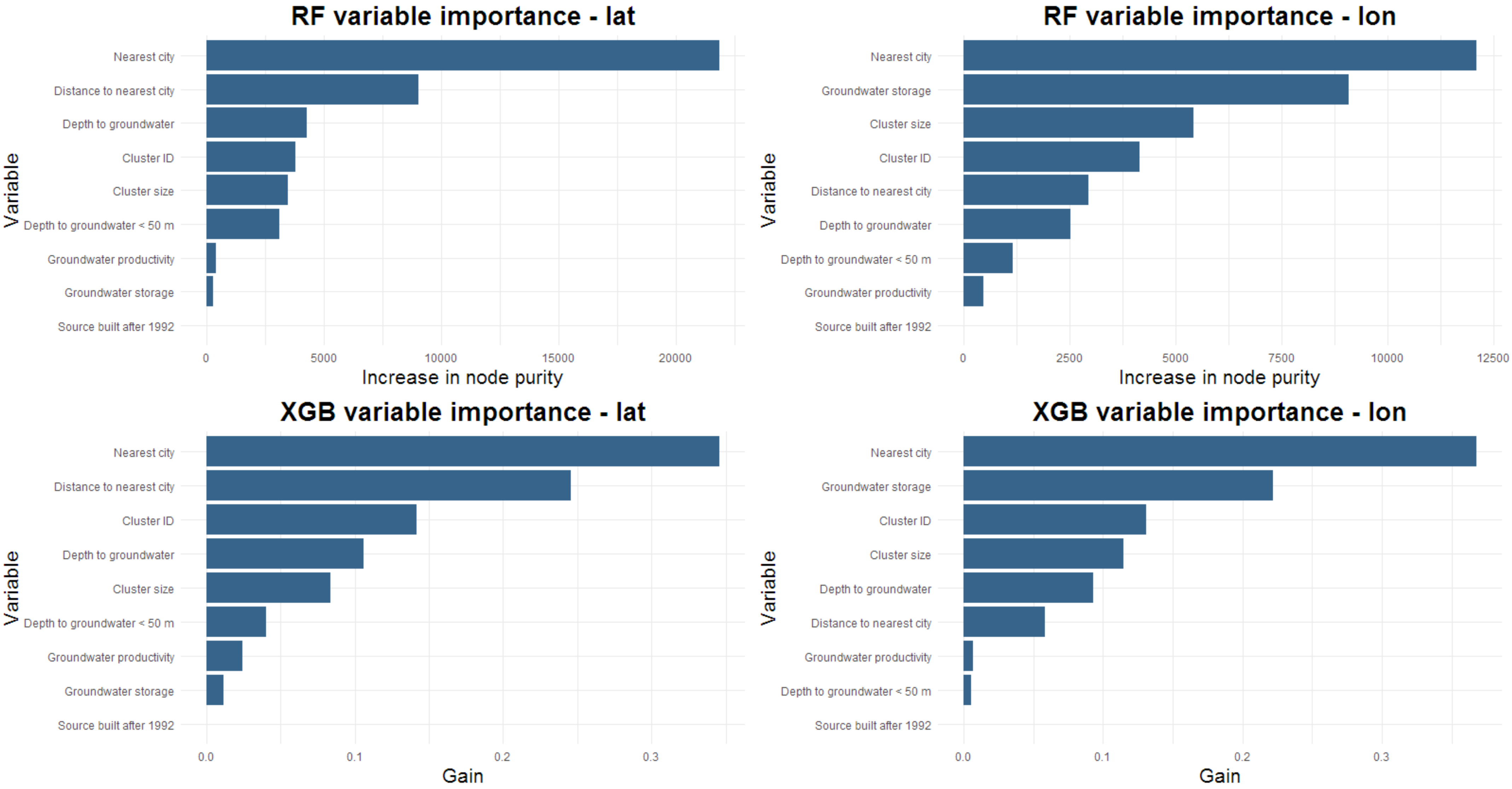


Figure 3. Variable importance ranks for "aggregate" models for (a) Random Forest (RF) model predicting latitude; (b) RF model predicting longitude; (c) Extreme Gradient Boosted trees (XGB) model predicting latitude; (d) XGB model predicting longitude

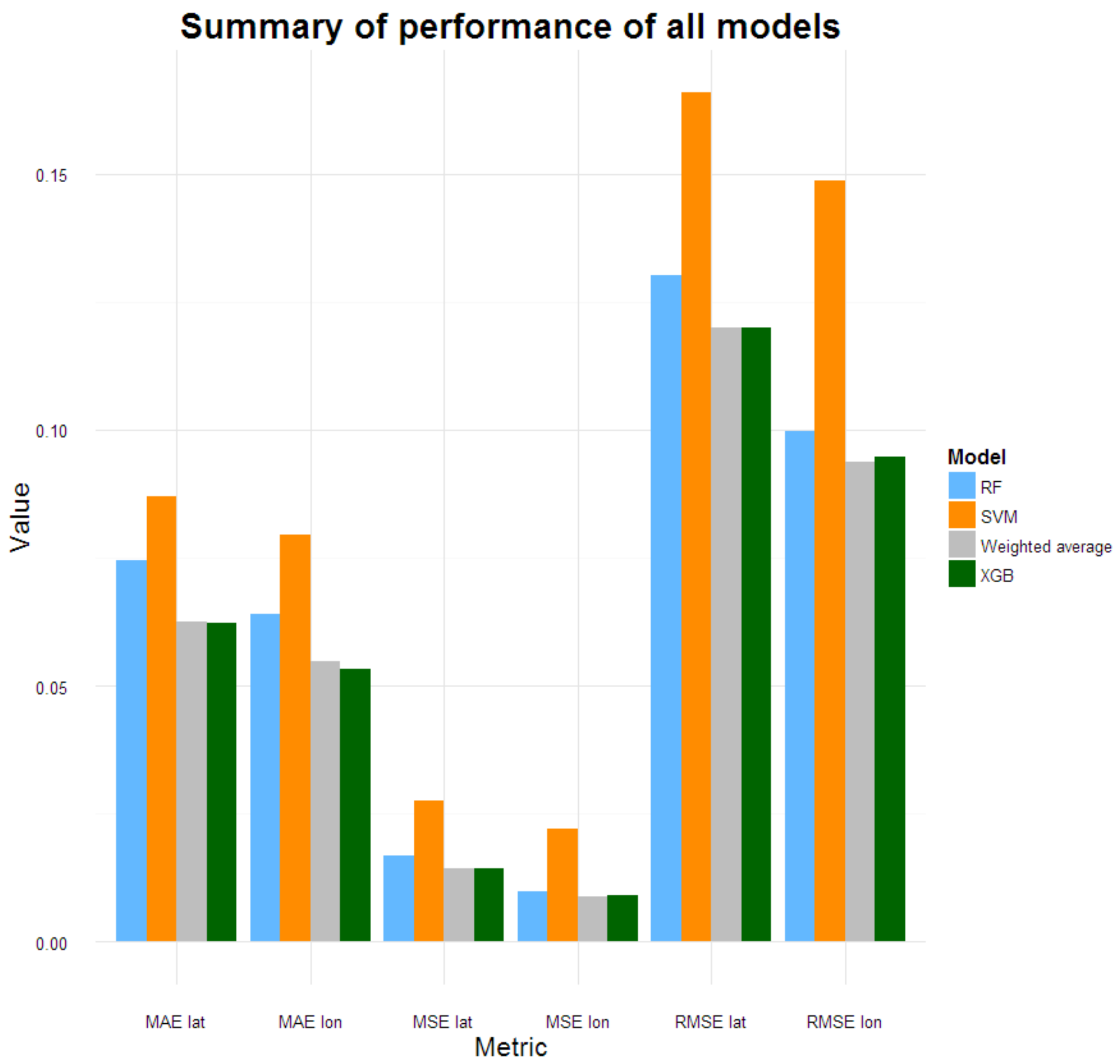


Figure 4. Accuracy metrics for "aggregate" models (RMSE – Root Mean Squared Error, MAE – Mean Absolute Error, MSE – Mean Squared Error).

Table 3. Average variable importance ranks by water source type (1=most important; 9=least important).

Variable	*Aggregate*	*Shallow wells*	*Machine-drilled boreholes*	*Hand-drilled tube wells*	*Springs*	*Average rank*
Nearest city	1	1	2.25	1	1.75	1.4
Nearest distance	3.5	5	1	2	5.5	3.4
Groundwater storage	5	7.25	5.5	8	4	5.95
Groundwater productivity	7.5	7.75	6.5	6.75	5.75	6.85
Groundwater depth	4	2	6	3	2.25	3.45
Cluster ID	4.75	4.25	3.5	6	4.5	4.6
Cluster size	3.75	5.25	4.5	4.5	4.25	4.45
Depth to groundwater < 50 m	6.5	3.5	6.75	5.25	8	6
Water source built after 1992	9	9	9	8.5	9	8.9

Table 4. Absolute rank differences and average inconsistency of ranks. Here, inconsistency is defined as the absolute rank difference between average latitude and longitude ranks for RF and XGB models. Colours represent distinct clusters of inconsistency.

Variable	*Aggregate*	*Shallow wells*	*Machine-drilled boreholes*	*Hand-drilled tube wells*	*Springs*	*Average inconsistency*
Nearest city	0	0	0.5	0	1.5	0.4
Nearest distance	3	0	0	0	1	0.8
Groundwater storage	6	0.5	4	0	6	3.3
Groundwater productivity	1	0.5	3	2.5	2.5	1.9
Groundwater depth	2	0	0	0	0.5	0.5
Cluster ID	0.5	2.5	1	1	3	1.6
Cluster size	1.5	1.5	3	1	0.5	1.5
Depth to groundwater < 50 m	1	1	1.5	0.5	0	0.8
Water source built after 1992	0	0	0	1	0	0.2

All the models achieved good accuracy (Figures 1, 2 and 4). The XGB model performed the best in terms of maximum prediction error, AUC and very similarly to "weighted average" in terms of RMSE for both latitude and longitude. It is easily seen from Figure 2 that XGB predicted 75% of locations with error $\leq$1000 m.

Generally, variable importance ranks were similar for RF and XGB holding a coordinate constant. However, ranks were rather different across coordinates holding model type constant.

There are large inconsistencies for variable importance ranks in predicting latitude versus longitude (Tables 3 and 4), with *groundwater depth* and *groundwater storage* being the most and the least consistent variables respectively. Groundwater-related variables belong to all three consistency groups.

4. Discussion

4.1. Model performance

A comprehensive literature review showed that Machine Learning has not previously been used for prediction of geographical coordinates from point patterns, suggesting the approach used here is novel. As a consequence, accuracy measures in Section 3 cannot be compared with those found in previous ML point pattern studies. However, model performance is consistent with typical performance of supervised learning algorithms as examined in numerous large-scale studies (e.g. Caruana & Niculescu-Mizil, 2006; Ogutu *et al.* 2010). This is true even for classification problems, where Boosted trees and Random Forests tend to perform the best on very different problems related to land cover classification and even genomic data, usually outperforming Support Vector Machines (e.g. Gislason *et al.* 2006).

4.2. Variable importance

Variable ranks as seen in Table 3 suggest that chosen covariates are better suited for prediction of springs and shallow wells compared to other sources. This can be seen from the ranks of *groundwater depth* which was ranked highest for these source types, which is consistent with relevant literature (e.g. Acheampong & Hess, 1998).

Average inconsistencies in Table 4. suggest that variables such as *groundwater storage* might not be suited for prediction of water sources locations. This suggests that increasing spatial resolution and expanding the set of predictive covariates for, especially, environmental variables could increase model accuracy. Machine Learning models which provide variable importance measures could potentially be used for exploring associations between variables. This method could be applicable to a wide range of areas such as hydrogeology or geomorphology.

4.3. Practical considerations

It is widely recognised a tradeoff exists between model complexity and training time (e.g. Lim *et al.*, 2000), which has been observed during the model development phase. Moreover, while RF accepts data in many formats (e.g. categorical variables, integers etc.), models such as ANN require data to be numeric, supplied in matrix format and transformed (e.g. normalised), creating numerous opportunities for introduction of human error while transforming data. Hence, it is suggested that simpler and more flexible models are used unless considerable computational resources and well-trained staff are available.

4.4. Possible improvements

The most significant improvement would be inclusion of more environmental covariates which are known to be associated with water source existence. Increasing spatial resolution of groundwater data would likely lead to accuracy improvements. Moreover, computation of density plots from discrete predictions could also be used when no estimate of number of unmapped water sources is available. Viewing the problem as classification, i.e. rasterising the data and predicting existence of water sources in each cell, could also improve the results and, depending on the cell size, could lead to lower computational costs.

5. Conclusions

This paper has shown that Machine Learning has good potential to accurately predict locations of water sources based on just a few covariates, making it an alternative to approaches such as ecological niche modelling. Importantly, varying degrees of variables' inconsistency were noted, suggesting superiority of particular variables for coordinate prediction. However, many potential improvements are possible, making the approach more accurate and applicable to distinctly different areas such as hydrogeology.

6. Acknowledgements

I am extremely grateful to my supervisor, Dr Jim Wright, for providing invaluable, countless hints and suggestions to advance the research further.

7. Biography

Przemyslaw Zientala is a 3rd year BSc Geography student at the University of Southampton. His interests include GIS, machine learning, data science and remote sensing.

8. References

Acheampong S.Y., Hess J.W. (1998) Hydrogeologic and hydrochemical framework of the shallow groundwater system in the southern Voltaian Sedimentary Basin, Ghana, *Hydrogeology Journal* **6** (4): 527-537.

Bonsor H.C., MacDonald A.M. (2011) An initial estimate of depth to groundwater across Africa, *Groundwater Science Programme Open Report OR/11/067*, British Geological Survey, 23p.

Bratton M. (1989) The politics of government-NGO relations in Africa, *World Development* **17** (4): 569-587.

Caruana R., Niculescu-Mizil A. (2006) An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd international conference on Machine Learning*.

Gislason P.O., Benediktsson J.A., Sveinsson J.R. (2006) Random Forests for land cover classification, *Pattern Recognition Letters* **27**(4): 294-300.

Hastie T., Tibshirani R., Friedman J. (2013) *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, 2nd ed., Springer, New York, 746p.

Jiménez A., Pérez-Foguet A, (2011) Water Point Mapping for the Analysis of Rural Water Supply Plans: Case Study from Tanzania, *Journal of Water Resources Planning and Management* **137** (5): 439-447.

Lim T-S., Loh W-Y., Shih Y-S. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, *Machine Learning* **40** (3): 203-228.

MacDonald A.M., Bonsor H.C., Ó Dochartaigh B.É., Taylor R.G. (2012) Quantitative maps of groundwater resources in Africa, *Environmental Research Letters* **7** (2): 1-7.

Ogutu J.O., Piepho H-P., Schulz-Streeck T. (2010) A comparison of random forests, boosting and support vector machines for genomic selection, *BMC Proceedings* **5** (Suppl 3): S11.

Water Point Data Exchange (2016) *The Water Point Data Exchange* [Online]. Available at: https://www.waterpointdata.org/ [Accessed: 05/01/2017].

Zafar Ullah A.N., Newell J.N., Ahmed J.U., Hyder M.K.A., Islam A. (2006) Government – NGO collaboration: the case of tuberculosis control in Bangladesh, *Health Policy and Planning* **21** (2): 143-155.