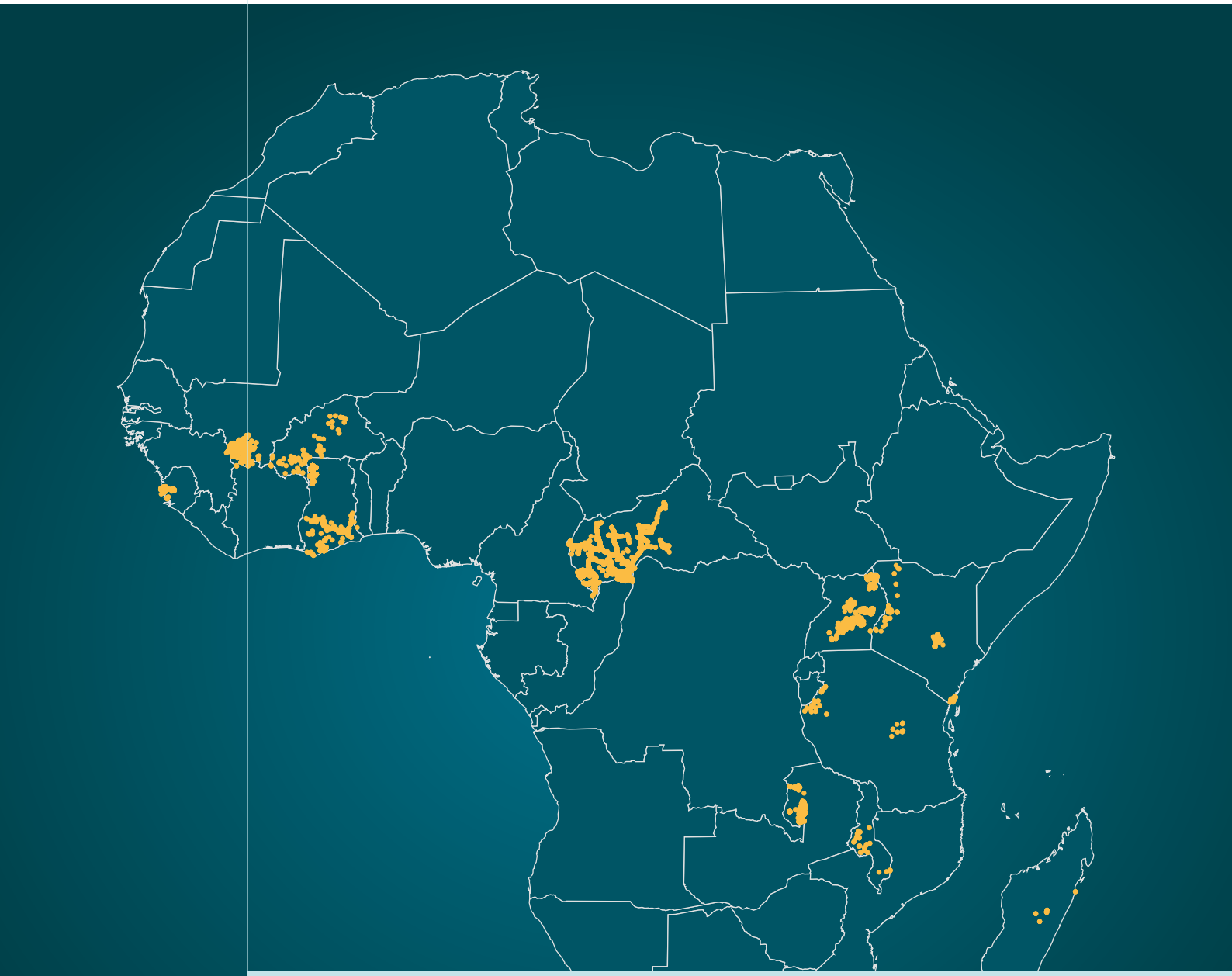
Briefing note | March 2025



Predicting revenue from drinking water supply infrastructure in rural Africa

Duncan McNicholl, Rob Hope and Emil Aydinsoy

UPTIME global

Executive summary

Uptime has developed a Revenue Predictive Model (RPM) to estimate user payments for rural drinking water services in Africa. As a new opportunity to improve sector financing strategies, this model can support governments and development finance projects to validate assumptions about expected user payments and possible subsidy requirements. By leveraging quarterly records representing over 200,000 months of data from Uptime in 11 African countries, we combine observed service data with relevant public datasets in a machine learning algorithm to estimate annual revenues for rural drinking water handpumps and piped schemes.

Applications include:

1. Validating revenue projections for planned infrastructure investments;
2. Modelling existing infrastructure to determine financial viability or subsidy needs;
3. Modelling rehabilitation strategies against alternative infrastructure portfolios;
4. Projecting climate and demographic changes on infrastructure revenue;
5. Designing results-based contracts or Payment for Results programmes; and
6. Performance benchmarking for rural water services.

This paper outlines the model and its input variables as a basis for collaborative model application. Details on infrastructure type, characteristics, location and estimated population are straightforward requirements that form the basis of revenue predictions. Available data from existing or planned projects should be able to interact with this model with relative ease.

The RPM provides a standard evaluation framework for large public investments in rural water infrastructure to improve programme designs and long term drinking water sustainability. The model can be applied and calibrated for all countries in sub-Saharan Africa. We invite governments and multi-lateral development banks to share data in the provided format (Appendix A) to test this model in order to improve revenue assumptions in financing strategies.

Introduction

Financing strategies inconsistent with field realities undermine the sustainability of rural drinking water infrastructure and jeopardise progress towards universal safe drinking water. Strategic deployment of limited resources will be critical for achieving SDG 6.1. Yet infrastructure breakdowns and overly optimistic financial forecasts continue to undermine efforts. Tariffs alone are unlikely to support full service cost recovery in rural Africa. Uptime data across sub-Saharan Africa shows that full operating cost recovery from tariffs is rarely achieved in rural settings at scale.[1] Assumptions about volumetric consumption are also often unrealistically high. People accessing water from multiple sources do not reliably consume and pay for 20 litres daily.[2] Sustainable financing strategies must therefore be grounded in realistic assumptions of what rural water users can and will really pay.

Knowing what people will pay is critical for designing effective service models with sustainable financing, be it subsidy, concessional or commercial finance. But limited data on actual user payments across rural Africa undermines objective analysis. Speculation about cost recovery becomes an insufficient substitute for these data gaps.

New data is now creating new opportunities. Since 2020, Uptime has collated high resolution data on user payments for rural drinking water services through execution of results-based contracts. Verified data from services for over 5 million people in 16 countries provide a new basis for analysis and predictive modelling. Within this global initiative, professional service providers maintain handpumps and piped schemes across Africa and receive quarterly performance-based subsidies as a grant calculated from standardised reporting on infrastructure reliability, volume and user payments. Quarterly data representing over 200,000 months of records for unique infrastructure sites have been collated under this initiative from 2021 through 2023 for 13,777 unique waterpoints (piped schemes or handpumps) in 11 countries serving an estimated total population of c. 2.2 million people (Figure 1, Figure 2). Leveraging this dataset, this paper outlines data requirements for our new model to predict revenues from rural water infrastructure to inform better financing strategies.

1 Delivering global rural water services through results-based contracts. Uptime Consortium, Working Paper 3.

2 Wagner, J., Koehler, J., Dupuis, M. et al. 2024. Is volumetric pricing for drinking water an effective revenue strategy in rural Mali? *npj Clean Water*, 7: 57. doi: 10.1038/s41545-024-00341-6

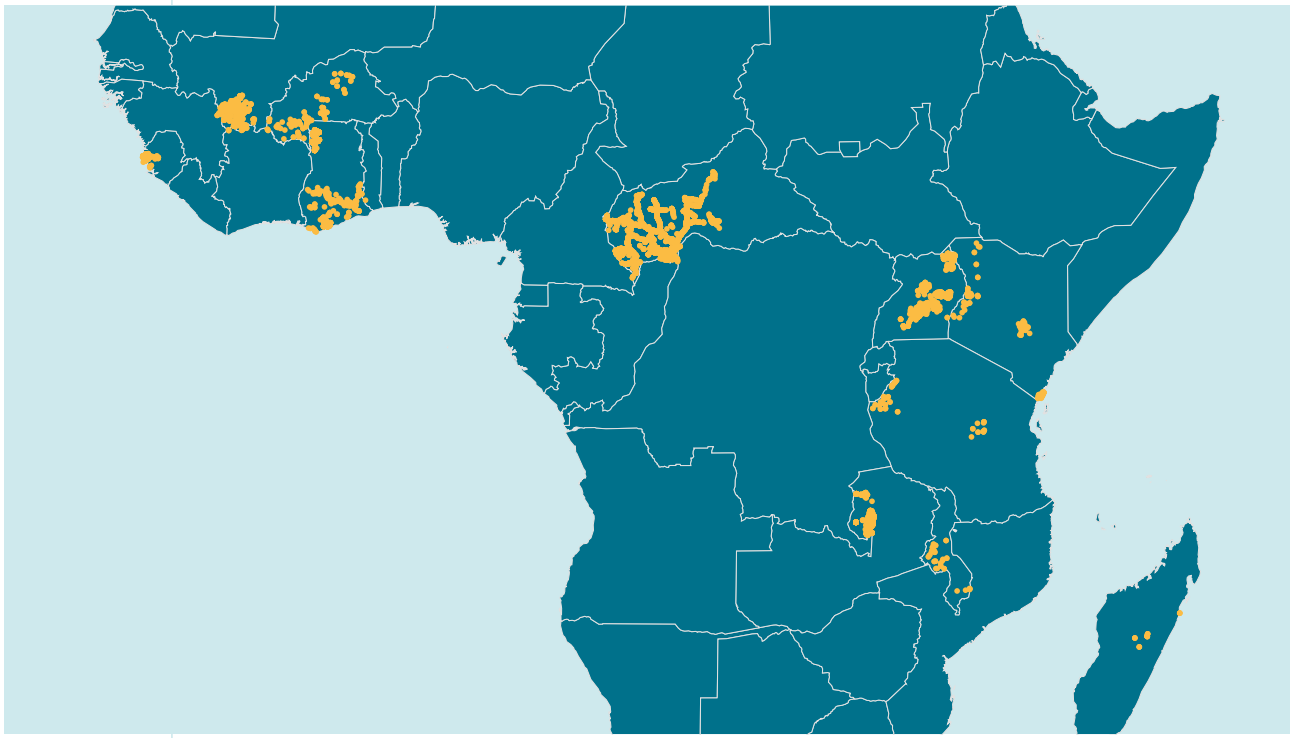


Figure 1: Waterpoint locations used to train the Revenue Predictive Model

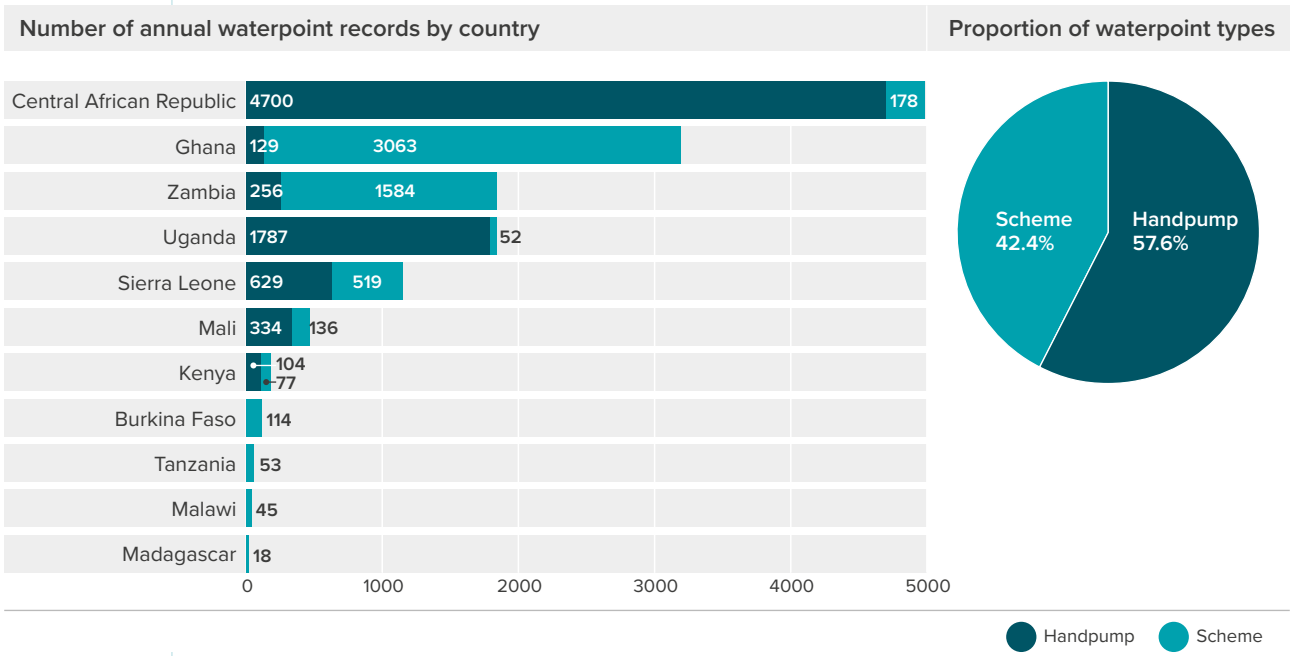


Figure 2: Distribution of training data by country

Predictive model example

A hypothetical example illustrates the potential value of the RPM. A multi-lateral development bank is investing in fifty community piped water schemes of different sizes in a rural part of East Africa. Revenue forecasting assumes the total population of 20,000 people will consume 20L per capita per day and pay USD 1 per cubic meter per day. The annual projected revenue is USD 146,000 and the schemes are assumed to sustainably cover an operating expenditure of USD 65,000 per year with profits to be reinvested in scheme expansion.

Historical data from nearby schemes are first used to calibrate the predictive model. After calibration, the maximum predicted revenue is USD 55,000 per year, implying an annual operating loss of USD 10,000. The predictions raise concerns about revenue projections and suggest that a subsidy strategy might be needed.

After considering alternative scheme designs, tariff models and tariff rates, the project ultimately decides that the possibility of an operating loss poses an unacceptable risk to long-term infrastructure sustainability, and that a subsidy strategy would provide a necessary backup. The development bank supports the creation of a public funding facility that can disburse performance-based subsidies under certain conditions with government oversight if necessary.

Predictive model design

To help practitioners engage with the predictive model, we first outline the design and logic of the model to clarify how predictions are made, their limitations and model input requirements.

Machine learning is applied in a two-step approach to predict annual user payments for each infrastructure site. The model first classifies infrastructure into three revenue range categories using decision tree logic, then uses machine learning models within each range for more precise estimates. The initial decision tree provides better transparency of model logic before 'black box' machine learning techniques are applied for greater precision. Range predictions are also typically more accurate than precise estimates while still providing useful information to decision makers.

Two types of data are used in the model: infrastructure characteristics and contextual factors (Table 1). Infrastructure characteristics are captured in the Uptime standard reporting framework. These are the input variables required to run the predictive model (Appendix A). Wider contextual factors are determined by referencing infrastructure geolocation in public datasets. Extensive model testing with various contextual factors found that population, nearby waterpoints and rainfall had the greatest impact on model accuracy.

Table 1: Summary of predictive model input variables

Infrastructure characteristics	Contextual factors
Input variables (Uptime data format)	Determined from public datasets using infrastructure geolocation
• Geolocation • Infrastructure type: handpump or piped scheme • Number of tap stands, kiosks and household connections (piped only) • Volumetric measurement type: estimated, manual or automated (piped only) • Estimated population served	• Population within 500m • Number of other waterpoints within 1, 5, 10 and 20km (from Infrastructure characteristics input variables) • Monthly rainfall data

Input variables from the Uptime dataset are used to reference other public datasets to enrich the model. Infrastructure geolocations are important for determining contextual factors including number of nearby waterpoints and area population. Nearby waterpoints are determined by calculating distance to other waterpoints in the input variables dataset; population data are calculated from high resolution population geospatial data available through the Humanitarian Data Exchange[3]; and monthly rainfall data is referenced from CHIRPS.[4] To increase generalisability of the model, specific country identifiers, such as country name, GDP per capita, and similar metrics are deliberately excluded. This decision produces improved model performance by reducing bias from country identification.

The first stage of the model uses a decision tree to classify predicted annual revenue ranges per waterpoint per year in three categories: Low USD (USD 0-5); Medium (USD 5-135); High (USD 135+). Characteristics of infrastructure sites are assessed as above or below a threshold and classified accordingly. The process repeats for each characteristic until sorting into a final predicted category is complete. The output is a probability of each site being low, medium or high annual revenue (Figure 3). A second model is then run within each category to generate more precise predictions. Secondary models are then used to generate more precise predictions, particularly for infrastructure in the High revenue range.

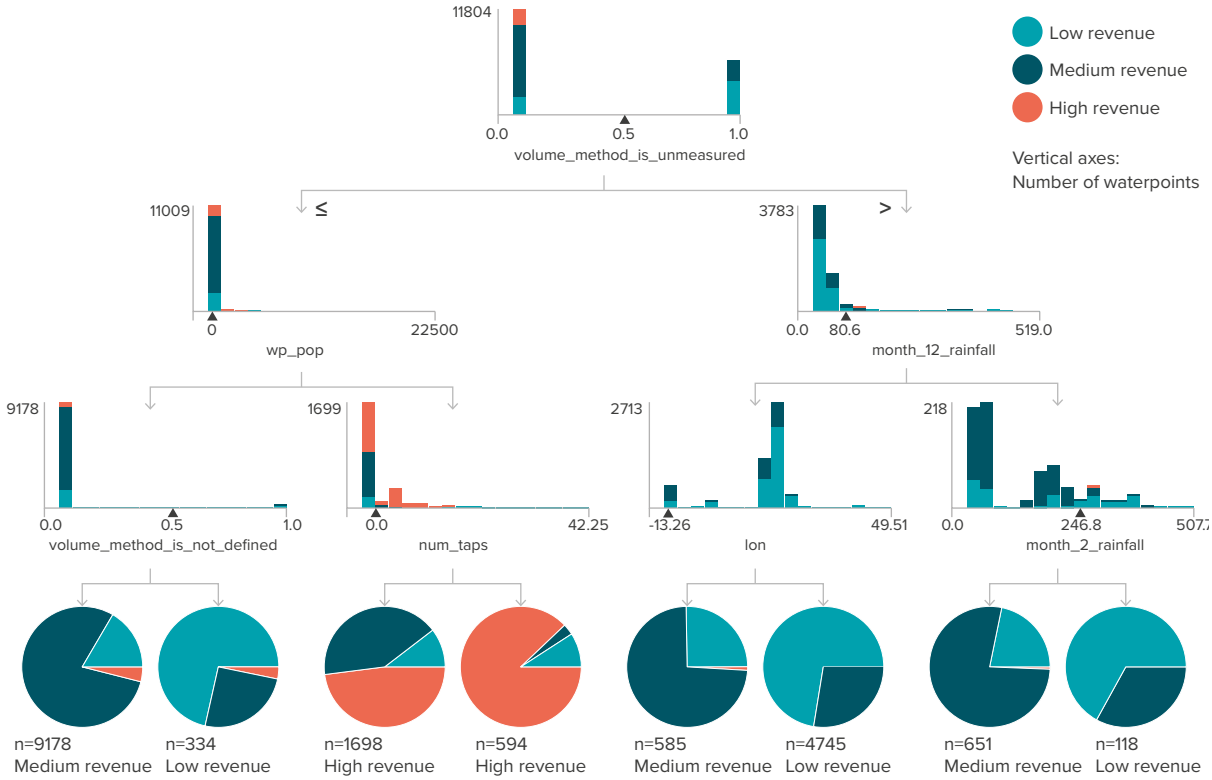


Figure 3: Illustration of the decision tree logic to classify infrastructure into revenue categories. Classification begins at the top and progresses downwards. The black triangle under each histogram defines the cutoff threshold used to classify infrastructure into one category versus the other.

3 Humanitarian Data Exchange

4 CHIRPS rainfall data

Model accuracy

Model testing finds accuracy of approximately 80% across all categories when predicting annual revenues within the three ranges. The Sankey diagram (Figure 4) considers how frequently the model is inaccurate in its predictions. When inaccurate, low or high revenue ranges are more likely to be predicted in the adjacent medium revenue range rather than for low to be predicted as high or vice versa.

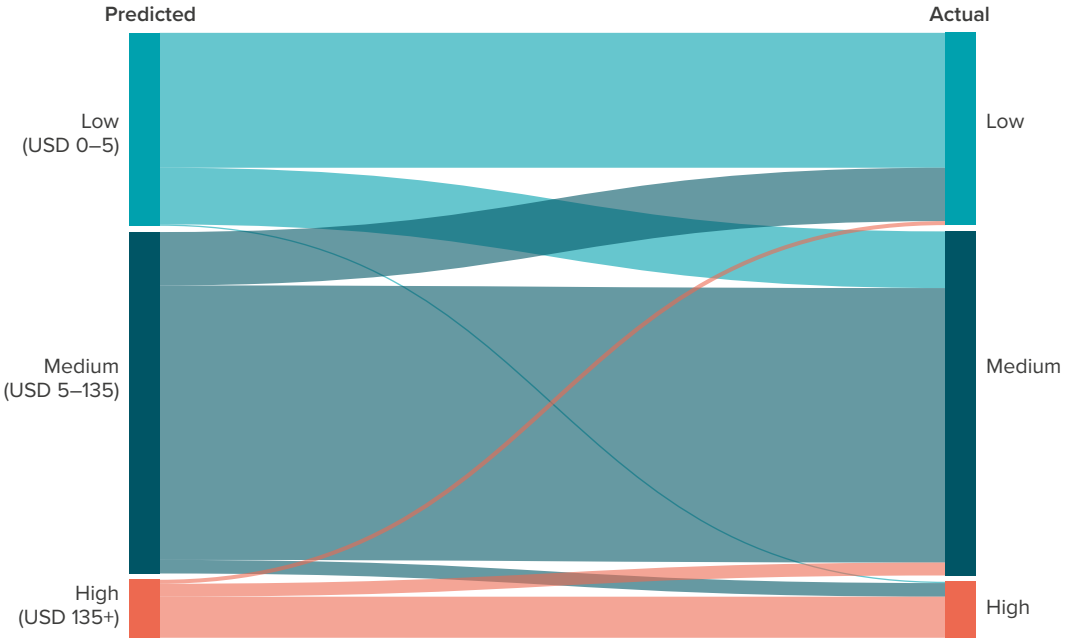


Figure 4: High revenue is incorrectly predicted as low in only 3.1 % of cases; low revenue incorrectly predicted as high in only 2.1 % of cases.

The model is naturally more accurate at predicting revenue ranges than specific values, which may be sufficiently valuable for decision-makers. Range predictions can still be useful for validating financing strategy assumptions. For example, a predicted maximum annual waterpoint revenue might provide an important perspective to challenge more optimistic assumptions. The relatively good accuracy of range classification suggests that the RPM could be ready for wider application to reduce risk by improving financing strategy assumptions.

Model limitations

Data scarcity remains a significant constraint in rural water sectors across Africa. The RPM is intentionally designed to operate with a minimal set of consistently available indicators, but important limitations remain. In particular, comparisons between infrastructure sites may be affected by missing information on local alternatives. Common gaps include: (1) consistent historical revenue records; (2) daily functionality and service reliability; (3) temporal trends in water quality; and (4) seasonal or inter-annual climate variability. These and other unknowns can shape user behavior and, in turn, revenue outcomes. While resolving all data gaps is impractical, the RPM aims to balance simplicity with predictive power by relying on the most widely available inputs.

Beyond the challenges of wider data availability, the central limitation of the model is overfit to the training dataset. A substantial number of waterpoint records come from a small number of service providers in a few African countries. Predictions will naturally be best in contexts, infrastructure types and service models that resemble historical data. Extrapolations to new contexts will not always be appropriate, and care is therefore needed in interpreting predictions in new geographies. Over time, the model can improve with the addition of more data in more contexts, but nuanced differences in geographies, service models, infrastructure types and other factors will affect model accuracy.

Two considerations are important in light of this central limitation. Firstly, model predictions should be considered with their uncertainty, especially when applied beyond the geographies of the training data. For this reason, range predictions might be a more practical starting point than specific revenue figures for each infrastructure site. The second consideration is that historical data from a particular context can significantly boost model accuracy by both calibrating the model and comparing outputs against known historical data. Where possible, practitioners should strive to obtain historical revenue data from the same context where the model is to be applied in order to improve predictive accuracy. The next section illustrates how the RPM can be calibrated for new contexts for better forecasting.

Using the RPM

Calibrating the model for new contexts

Challenges with predicting revenues in a new context can be partly offset through the addition of actual revenue records from that context. Relatively small datasets can calibrate the existing model for better predictive accuracy. This can enable model application to countries or regions outside of the Uptime database. We find that adding 12 months of consecutive records from 150 waterpoints can significantly improve predictive accuracy to the c. 80% level. Application of the model without calibration data is still possible, although with significantly reduced accuracy. We therefore recommend that practitioners seeking to apply the model aim to obtain annual revenue records from at least 150 waterpoints as supplemental data for model calibration.

Applying the predictive model

Practitioners can collaborate with Uptime to apply the predictive model through the following process:

1. **Define the infrastructure set**: Identify the existing or planned infrastructure for revenue prediction.
2. **Calibrate the model with historical revenue data**: Incorporate historical revenue data (Appendix A) from the same context to update the predictive model for this application. This is optional but highly recommended for improved accuracy.
3. **Input infrastructure characteristic data**: Collate data on planned or existing infrastructure characteristics and location using the provided format (Appendix A). Uptime can add contextual factors by referencing supplementary datasets.
4. **Generate revenue predictions**: Run the model to produce both range and precise annual revenue estimates for each infrastructure site.
5. **Compare predicted with projected revenues**: Evaluate results and consider implications.

Uptime can work with practitioners to identify potential model applications and provide guidance on the calibration process to help maximise the potential utility of revenue predictions.

Summary

Inaccurate assumptions about the financial viability of rural drinking water services will persist in the absence of better data for better decisions. Expectations of full cost recovery from user payments risk undermining long-term service sustainability if projections are not grounded in reality. Financial models and tariff structures need to accurately estimate what users really will pay as distinct from what they 'should'.

The Revenue Predictive Model provides an opportunity to validate assumptions about user payments using a large empirical dataset, and to update those assumptions for specific contexts where data are available. We present the data format, model design and underlying reasoning as a collaborative opportunity to apply the model and expand the dataset for better revenue predictions across Africa. We invite governments and development banks to collaborate with Uptime in applying this tool for smarter and more sustainable water financing to support SDG 6.1.

Uptime can support model calibration and application for suitable initiatives. Support for model use will be prioritised for large public financing initiatives led by governments and multi-lateral development banks to maximise potential benefits and to expand datasets for improved future model performance.

Application of the model requires:

- **Prediction data**: Infrastructure details (Appendix A)
- **Calibration data**: Optional, but lower accuracy predictions are possible without. Ideally at least 12 consecutive months of historical data from at least 150 waterpoints in the same context where the model will be applied.
- **Project details**: Outline of project financing strategy including public funding components and intended application of revenue predictions.

All inquiries regarding Uptime's Revenue Predictive Model can be directed to data@uptimeglobal.org

Appendix A: Waterpoint data template for the revenue predictive model

Indicator	Description
wp_id	Unique identifier
mWater_id	If applicable for mWater
lat	Latitude in decimal degrees
lon	Longitude in decimal degrees
institution_type	Note if school or healthcare facility
wp_type	Handpump or piped scheme
volume_method	How water volume is measured: automated, manual, estimated or unmeasured
num_kiosks	Number of community kiosks (piped schemes only)
num_shared_taps	Number of community taps (piped schemes only)
num_hh_taps	Number of household connections (piped schemes only)
population	Estimated population served by the waterpoint
USD_revenue	Required input for training the model in a new context. Quarterly revenues for at least a full year for each site are required. Predicted annual USD revenue will be the output of the RPM.