Optimizing Water Access in the Democratic Republic of Congo

A Two-Stage Mixed-Integer Programming Approach

Authors: César Dori, Jacob Lebovitz
Course: Optimization Methods — Fall 2025
Institution: MIT Sloan School of Management

Date: December 7, 2025

1 Motivation

The Democratic Republic of Congo (DRC) is the fourth largest country in Africa with a population of over 109 million. According to World Vision, 64% of the population lives in extreme poverty, with 54% lacking basic access to water services.[1] Many organizations have volunteered to repair and build new wells in the region, with World Vision helping 'more than 125,000 people gain access to clean water' in 2019.[1]

This report highlights the power of optimization modelling techniques in solving a similar problem: maximizing the coverage to basic water services while minimizing the average trip distance for the population served. We investigate the benefit of repairing current wells present in the region coupled with identifying candidate locations for new wells.

This report uses population raster data published in 2020, at a time when the population was around 89.9 million. Despite our focus on realism, the assumptions used in this analysis mean the results should be interpreted as indicative rather as direct, tangible impacts.

2 Data

The project utilizes two primary, publicly available geospatial datasets that together enable a realistic and data-driven representation of water access conditions in the Democratic Republic of Congo. The first dataset (2.1) provides high-resolution population density estimates that allow us to approximate where people live and how demand for water services is spatially distributed across the country. The second dataset (2.2) documents the locations and operational status of public water points, capturing the current infrastructure landscape and revealing gaps in access. By combining these complementary datasets, we construct a spatial mapping that supports optimization-based decision making for improved water service coverage.

2.1 WorldPop DRC Population Raster[2]

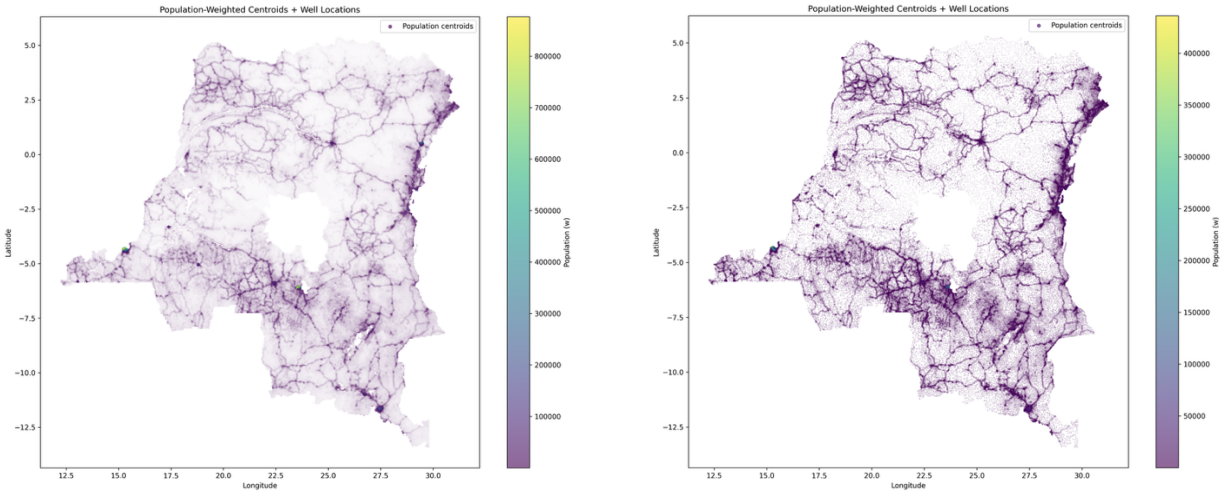
- Scope: WorldPop population raster for DRC in 2020
- Key fields: GeoTIFF (lon/lat; people per 1 km^2 pixel)
- Processing: Clip to DRC boundary; build demand grids (square and K-means)

2.2 Water Point Data Exchange (WPDx)[3]

- Scope: All water points from the Democratic Republic of Congo in 2024
- Key fields: Latitude/longitude, status (functional vs. non-functional)
- Processing: Group by status, and retain fields relevant in order to build sets for functional and non-functional wells

3 Demand Grid Choice

We generated demand grids two ways: (i) a uniform square grid clipped to the DRC, and (ii) a population-weighted K-means grid (cluster centers weighted by pixel population). Condensing the population grid was necessary to ensure computational efficiencies when running the 2-stage optimization. Both use the same access radius R_m and distances. K-means places more points in dense areas and fewer in sparsely populated regions while preserving total population.



(i) Uniform square grid

(ii) K-means grid

Grid	Cells	Pop. (M)	$\overline{\lvert N_F \rvert}$	$\overline{\lvert N_R \rvert}$	$\overline{\lvert N_N \rvert}$	$\overline{\lvert N \rvert}$	arcs_F	arcs_R	arcs_N	arcs_{all}
Square (Agg)	102,971	86.7	0.019	0.005	0.009	0.033	1,909	537	934	3,380
K-means	99,983	89.4	0.079	0.021	0.035	0.135	7,894	2,140	3,503	13,537

(iii) Preprocessing metrics with $R_m = 1.5$ km. $\overline{\lvert N. \rvert}$ is the average neighbors per demand cell; arcs are total demand–site edges within R_m.

Figure 1: Demand grid constructions and neighbour metrics.

Summary. With a similar number of demand cells, the K-means clustering grid ($K = 100{,}000$) preserves slightly more total population and substantially increases neighbor density (avg $\overline{\lvert N \rvert}$ from 0.033 to 0.135) and edges (3.4k to 13.5k). This better reflects where people live and should modestly improve potential coverage at the cost of a larger (but still manageable) model graph. We adopt K-means going forward because it represents population distribution more accurately at a feasible increase in computational cost.

4 Candidate Wells

We generated a set of 3000 candidate well site locations by targeting high-population areas currently unserved by functional wells, applying a minimum separation between sites to avoid clustering. We ranked demand cells by unserved population and iteratively placed a candidate at the cell centroid, skipping any location within 4.5 km of an already chosen site.

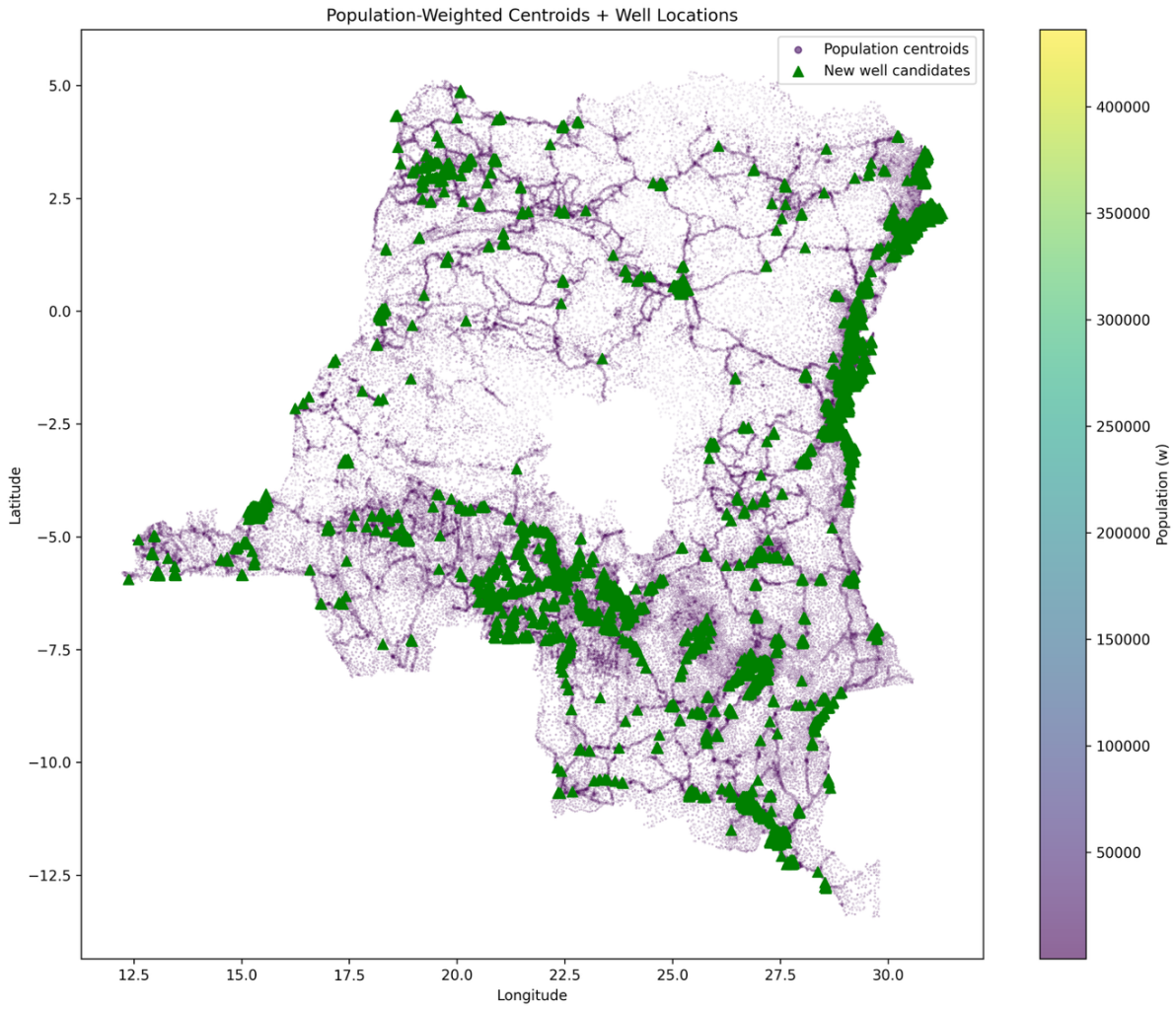


Figure 2: New well candidate positions.

Summary. By implementing a minimum spacing of 4.5km between wells for 3000 candidate locations, we were able to position them in unserved, high-population areas as shown in Figure 2.

5 Formulating the Problem as a Linear Program

The primary problem we are trying to solve is to maximize the size of the population with access to basic water services (from wells). The distance of the average trip is only optimized while maintaining the maximum coverage calculated in pass 1. This prompts the use of a two-staged mixed-integer program. Both stages were run for a set of budgets to determine budget sensitivity (the budget constraint was not shown explicitly in pass-2 for conciseness).

Sets

I: Demand cells from K-means clustering ($K = 100{,}000$). J_F: functional wells (treated as fixed-open). J_R: non-functional wells (repair candidates). J_N: new-site candidates (optimally located).
$N_F(i) = \{\, j \in J_F : \ d_{ij} \leq R \,\}$: functional sites within radius R of cell i. $N_R(i) = \{\, j \in J_R : \ d_{ij} \leq R \,\}$: repair candidates within radius R of cell i. $N_N(i) = \{\, k \in J_N : \ d_{ik} \leq R \,\}$: 3,000 new-site candidates within radius R of cell i. $N(i) = N_F(i) \cup N_R(i) \cup N_N(i)$: all eligible sites for cell i.

Parameters
$w_i \geq 0$: population in demand cell i. d_{ij}: distance (Haversine) from cell i to well j. $R = 1{,}5$km: access radius used to identify accessible wells for each centroid. $C^R = \$2{,}200$: universal unit repair cost for repair wells. $C^N = \$8{,}000$: unit new-site cost. $B = \{0, 2.5, 5, 7.5, 10, 15, 20, 30\}$: total budget ($M).

Decision variables
$y_j^R \in \{0,1\}$: 1 if repair site $j \in J_R$ is repaired, 0 otherwise. $y_k^N \in \{0,1\}$: 1 if new site $k \in J_N$ is built, 0 otherwise. $z_i \in \{0,1\}$: 1 if demand cell i is covered by some functional, repaired, or newly built well within R, 0 otherwise. $x_{ij} \geq 0$: (Pass 2) assignment share from cell i to site $j \in N(i)$ (used for distance minimization).

Pass 1 (Max coverage)

$$
\begin{aligned}
\max \ & \sum_{i \in I} w_i z_i \\
\text{s.t.} \ & \sum_{j \in N_F(i)} 1 + \sum_{j \in N_R(i)} y_j^R + \sum_{k \in N_N(i)} y_k^N \ \geq \ z_i && \forall i \in I \\
& \sum_{j \in N_R(i)} y_j^R C^R + \sum_{k \in N_N(i)} y_k^N C^N \ \leq \ B && (1) \\
& y_j^R \in \{0,1\} \ \forall j, \quad y_k^N \in \{0,1\} \ \forall k, \quad z_i \in \{0,1\} \ \forall i
\end{aligned}
$$

Let $S^\star = \sum_i w_i z_i^\star$. $S^\star$ is the max number of people having 1 well in their walkable radius.

Pass 2 (Minimize distance w/o sacrificing population served).

$$
\begin{aligned}
\min \ & \sum_{i \in I} \sum_{j \in N(i)} w_i d_{ij} x_{ij} \\
\text{s.t.} \ & \sum_{j \in N(i)} x_{ij} = z_i && \forall i \in I \\
& x_{ij} \leq 1 && \forall i, \ \forall j \in N_F(i) \\
& x_{ij} \leq y_j^R && \forall i, \ \forall j \in N_R(i) \\
& x_{ik} \leq y_k^N && \forall i, \ \forall k \in N_N(i) \\
& \sum_{i \in I} w_i z_i \ \geq \ S^\star, \quad y_j^R, y_k^N, z_i \in \{0,1\}, \ x_{ij} \geq 0
\end{aligned}
$$

6 Method

First, we ran the Pass 1 to maximize the total population served. We ran this optimization over each fixed budget B, choosing which nonfunctional wells to repair and which new candidate sites to open. For each budget B we stored the optimal population coverage, S^*.

Then, using the same R_m and budget B, we solved Pass 2 to minimize the average walking distance of the served population while constraining the total served population to be at least the maximum population served from pass-1 (S^*). The decision variable x_{ij} represents the proportion of centroid i's population "assigned" to well j, and is constrained by the binary well build status (y_j^R,y_k^N).

7 Results

The results from the two-stage optimization problem are displayed in Table 1. For each budget, the total population served S^* was computed with the chosen wells (repair and new). The second stage is ran to compute the average walking distance, essentially resetting the selection of repair and new wells. In most cases, the selection was very similar across the passes, as Pass 2 had to meet a minimum population served of S^*.

Budget ($M)	Pass-1 S^* (M)	Pass-1 sites		Pass-2 Avg dist (m)	Pass-2 sites	
		Repairs	New		Repairs	New
0.0	7.14	0	0	650.00	0	0
2.5	18.75	49	299	340.68	49	299
5.0	21.40	65	607	318.04	58	609
7.5	23.34	66	919	298.82	67	919
10.0	24.92	70	1,230	282.82	69	1,231
15.0	27.45	75	1,854	266.83	76	1,854
20.0	29.41	79	2,478	248.72	80	2,478
30.0	30.68	354	3,000	216.33	1,496	3,000

(i) Pass-1 (maximize coverage S^*) vs. Pass-2 (minimize average distance) at the same budget and $R_m = 1.5$ km.

Selected Budget

We selected a final implementation budget of $10M as the proposed go-to-market strategy for deploying this optimization framework in practice. This budget represents a strong balance between scale of impact, operational feasibility, and diminishing returns. At $10M, the model serves approximately 24.92 million people, which is over three times the baseline population served at $0, while achieving a substantial reduction in average walking distance to water of more than 360 meters. Beyond this threshold, additional budget continues to improve outcomes, but at a slower marginal rate relative to cost.

From an operational standpoint, the $10M strategy funds the repair of 69 non-functional wells and the construction of 1,231 new wells, representing a realistic deployment size for a large international NGO.

A map of the functional, repaired, and newly built well locations is shown below. We can see that our optimization results target larger cities, yet serve rural communities as well. From the map, many new wells are built on the eastern boarder, which is an area known to have water insecurity due to it being rural and prone to frequent conflict. Southern areas are also highly water insecure due to mining pollution, so southern areas are also well served by our optimization model.

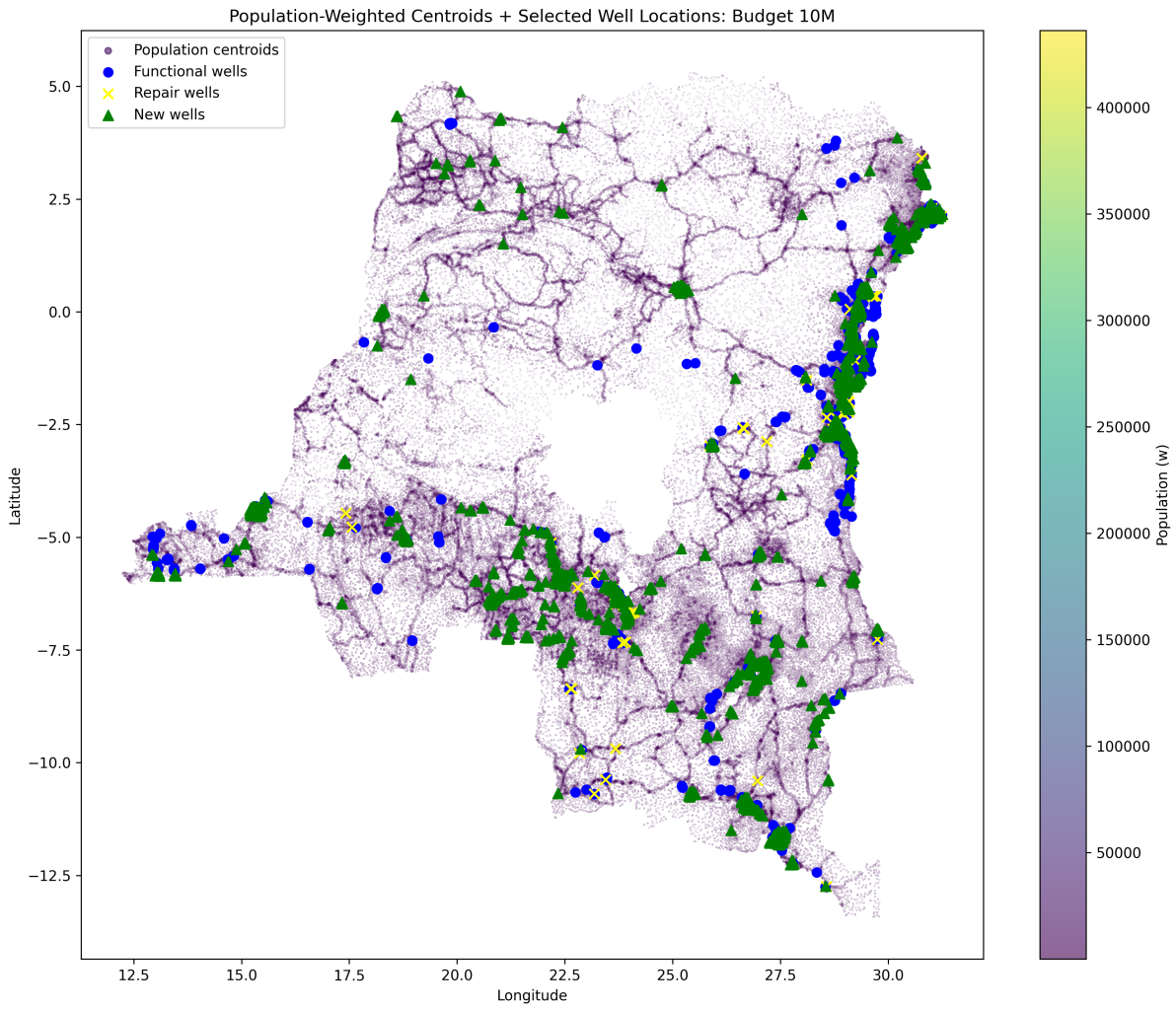
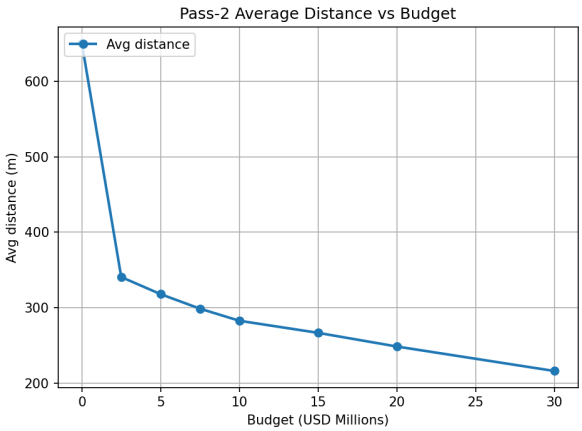


Figure 4: Functional, repaired, and new well locations under $10M budget

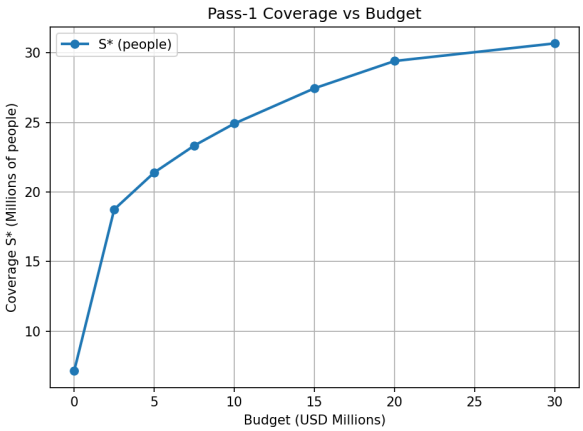
8 Budget Sensitivity

The biggest distance drop occurs at smaller budget sizes as shown in Figure 4. Any increase in budget from $2.5M generates incremental improvements in average walking distance. Initial spending on a project like this would see the largest improvements; later spending fills gaps where populations are thin and distances are already moderate.

A similar pattern is observed for coverage. There is diminishing returns in increasing budget size; early dollars target new wells in densely populated areas, and later dollars treat sparser regions.



(i) Average distance vs. budget



(ii) Coverage vs. budget

Figure 5: Budget sensitivity: early dollars deliver the largest gains; later spending yields diminishing returns as it targets sparser regions.

9 Discussion

The results from this optimization show that there is a lot of promise to use optimization to identify areas of highest improvement in building new wells. However, it is important to note the pitfalls and uncertainties in this model, attributed to the high complexity and unknowns in terrain. Across all of the budgets, we can increase the population served by significant margins, even under the uncertainty of estimated population distributions using K-means clustering centroids. In humanitarian projects that have limited funding, optimization can be used to ensure budget is going to optimally located water access locations.

One important note is the highly disproportionate ratio of chosen new wells to repair wells. This outcome is since new well build locations are chosen based on K-means population centroids. Naturally, building a new well at or very close to the centroid will create a minimum distance ($= 0$) for the total population served by that well, leading to many new wells being built over repairing nonfunctional wells. That being said, we can see that a budget of $\$30M$ can choose all of the new wells. This explains why the number of repair wells in this budget is relatively large, as the remaining budget has access to only repair wells to optimise the average distance further.

10 Further Work

While the results and general locations of highly concentrated well locations are generally areas that should be targeted for improving water access, there are many limitations and opportunities for improvement in our model. Population rasters and K-means centroids provide a coarse approximation of true settlement patterns. The centroid of cluster or large grid cell may fall in an uninhabited area (forest, slope, water body), potentially misrepresenting travel distances.

Additionally, the constraint that a household is "covered" if a well lies within 1.5 km assumes relatively uniform terrain and travel conditions. Walking paths, elevation, seasonal flooding, insecurity, and land-use restrictions can dramatically alter effective accessibility. Incorporating travel-time networks or friction surfaces would likely change coverage estimates and the optimal set of wells. In addition to modeling travel-time, it is also important to note groundwater availability, or lack thereof, in certain locations, which can constrain well locations further. The wells' capacities should also be taken into account for more accurate results. Finally, it is important to note that water coverage from wells does not represent overall water access, as many households, especially in urban centers, rely on other sources for daily water needs.

11 Conclusion

This project presents a data-driven, optimization framework for improving water access in the Democratic Republic of Congo by strategically repairing existing wells and constructing new ones. By transforming population raster data into a computational feasible set and formulating a mixed-integer program that captures spatial relationships and real-world constraints, the model identifies combinations of repair and construction decisions that substantially increase the number of people with a well within reachable distance. The two-pass structure that first maximizes coverage, then minimizes travel distance mirrors real operational priorities and yields solutions that are both impactful, realistic, and optimal within budgetary constraints.

Although the model relies on simplified assumptions, it provides a scalable foundation for future planning efforts and underscores the potential of optimization tools in humanitarian logistics and infrastructure design. This framework could evolve into a highly actionable support system for global health organizations to optimally use their budget to expand water access.

12 References

1. Reid, K. (2024, March 17). *10 worst countries for access to clean water*. World Vision. Retrieved December 4, 2025, from https://www.worldvision.org/clean-water-news-stories/10-worst-countries-access-clean-water

2. WorldPop. (2020). *Democratic Republic of the Congo population counts (1 km), UN-adjusted* [Data set]. University of Southampton. https://doi.org/10.5258/SOTON/WP00671

3. Water Point Data Exchange (WPDx). (n.d.). *Water Point Data Exchange (WPdx-Basic)* [Data set]. WPDx Data Playground. Retrieved December 5, 2025, from https://data.waterpointdata.org/dataset/Water-Point-Data-Exchange-WPdx-Basic-/jfkt-jmqa/about_data

4. Portions of this paper were edited for clarity and style using ChatGPT by OpenAI.