Thesis report

# Introduction

## Background

To address the global challenges that humanity is facing, the United Nations adopted a set of 17 Sustainable Development Goals(SDGs) in 2015 (*A/RES/70/1 Transforming Our World*, 2030) with 169 targets to be achieved by 2030. The Sustainable Development Goals (SDGs), otherwise known as the Global Goals, are a set of objectives within a universal agreement to end poverty, protect all that makes the planet habitable, and ensure that all people enjoy peace and prosperity, now and in the future (Morton et al., 2017). Unlike its predecessor,the millennium development goals, which focused on what the wealthy countries should do for the global south, the Sustainable development goals are universal and seek to promote the global and sustainable well being of all countries (Sianes et al., 2022). Different surveys have been carried to rank the SDGs in order of importance (Kleespies & Dierkes, 2022). In one survey by (Fleming et al., 2017), clean water and sanitation (SDG 6) was selected as one of the most important SDGs which will be the focus of this study.

The Sustainable development Goal 6 is one of the 17 Sustainable Development Goals that addresses access to clean water and sanitation. It includes targets for drinking water(SDG6.1) and sanitation and hygiene(SDG6.2) as shown in Table 1. Access to water, sanitation and hygiene is a basic right that is recognized as playing a vital role in the 2030 Agenda by improving progress in other areas such as health, education and poverty reduction(UNEP,n.d). However,despite significant efforts in recent years, there are still substantial gaps in access to safe drinking water and sanitation around the world. As of 2020, an estimated 2 billion people worldwide still lacked access to safe drinking water and 3.6 billion lacked access to safe sanitation (*Progress on Household Drinking Water, Sanitation and Hygiene 2000-2020*, n.d.) while 946 million practiced open defecation(Cai et al., 2021), majority of whom are in asia and sub- Saharan Africa (Water, 2018) . Of those who lacked access to basic sanitation, 70% were in rural areas and only 1 out of 10 countries below the 95% coverage mark are on track to meet the 2030 target(Water, 2018). In order to achieve the 2030 target of universal coverage, current rates will need to be significantly increased as shown in figure 1 below.

## Monitoring of SDG 6

Monitoring of SDG 6 estimates is an essential part of tracking progress and informing development by identifying areas which need improvement, informing policy decisions, tracking funding flows, and ensuring accountability for meeting commitments to the SDGs. Therefore, ongoing monitoring efforts are crucial for achieving SDG 6 and ensuring that access to water, sanitation, and hygiene is available to all.

The international monitoring of drinking water and sanitation has been in effect since the 1930s when the League of Nations Health Organization was responsible for monitoring (Bartram et al., 2014). The World Health Organization(WHO) then took over and now The Joint Monitoring Programme (JMP) which is a body that consists of The World Health Organization(WHO) and The United Nations Children‘s Fund (UNICEF) is responsible for global monitoring of water, sanitation, and hygiene estimates since 1990 (Supply & Programme, 2015). They produce a total of 26 indicators for 232 countries related to water, sanitation and hygiene at the national, regional and global level(JMP,2017). The indicators include the percentage of households using different technologies or levels of service which is further disaggregated by rural and urban settings (Wolf et al., 2013).

To compute indicator estimates for water, sanitation and hygiene, the JMP relies on data from a variety of national data sources including: national censuses, national household surveys and administrative data. The data is manually extracted from these sources and organized into a standardized format using individual spreadsheets, known as the JMP country files. These country files contain this data in a structured format, which allows scripted programming languages to use the data and compute indicator estimates using statistical methods. For each of the 232 countries, the estimates are generated by fitting an ordinary least squares regression line to a series of data points (Pérez-Foguet et al., 2017). An intercept and regression slope is calculated separately for each country.STATA is used as a tool to read the raw data from the country file, fit the model to compute the estimates, and write the results back into the country file. The methods used by the JMP are fully described in a report published in March 2018 (Joint Monitoring Programme, 2018).

While the JMP country files allow for a transparent review of raw data sources and published estimates, the STATA script used for data transformation remains unpublished and closed to the public. However, there is significant potential to publish the code that produces the estimates as open-source software to facilitate verification of the results, build upon them, or contribute to further development of statistical methods. This would improve the reproducibility of the results, enabling other researchers to assess the validity of the JMP’s estimates and methods. This brings about the concept of Reproducibility.

Reproducibility has been identified as a critical issue in scientific research (Peng, 2011). Open-sourcing the STATA script used by the JMP would allow for better transparency and reproducibility of the estimation methods. Furthermore, it would enable other researchers to build upon the methods and potentially improve the accuracy of the estimates. By making the estimation methods more accessible and transparent, the scientific community can more effectively work together towards achieving SDG 6 targets.

Producing accurate global estimates requires statistically sound and reproducible methods to effectively monitor progress on SDGs and enhance quality evidence - based decision making (Wolf et al., 2013). This is in line with the JMP’s mission “to be the trusted source of global, regional and national data on sustainable access to safe drinking water and basic sanitation for use by governments, donors, international organizations and civil society” (JMP, n,d). With 2030 fast approaching, it is important that we review the current methods in an effort to strengthen monitoring and reporting. Without reliable data systems, it is difficult to improve water, sanitation and hygiene services or measure progress towards targets. The objective of this study will be to reproduce the JMP sanitation estimates using the JMP’s documented methods, review the methodology used to reproduce the estimates and in so doing suggest alternative statistical models that are able to capture the different trajectories present in the data.

# Problem statement

Advancement in SDG 6 depends a lot on a strong foundation of data credibility. Accurate data is the lifeblood of decision-making and the raw material for ensuring accountability (*Data-Based Decision-Making Processes for WASH | SSWM - Find Tools for Sustainable Sanitation and Water Management!*, n.d.).The issue of data credibility is crucial for advancing Sustainable Development Goal 6 (SDG 6) and achieving universal access to safe water and sanitation for all and without it, progress towards SDG 6 will be difficult to measure and achieve . To calculate its estimates, The Joint Monitoring Programme relies on country-level open government data. This data is stored in multiple formats, which are often not machine-readable (e.g. PDFs), unstructured, and without metadata. Processing this data is time consuming, prone to errors, and difficult to reproduce. This could be overcome by making the data readily accessible in a form that follows the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles (Wilkinson et al., 2016). It is therefore crucial that the data shared through open government sources is machine readable to facilitate comparability, more accurate and efficient analysis. In addition, the current JMP method has received criticism as the generated estimates differ from the national country estimates. This is likely due to the underlying data sources or the fact that the estimates are derived from a linear regression which fails to capture the different non linear trends present in countries(Bartram et al., 2014). The JMP uses linear regression due to its simplicity, the limited number data points available and the ease of explanation(Pérez-Foguet et al., 2017). However, alternative modeling methods have been proposed to better monitor SDG 6 progress. For example, (Ezbakhe & Pérez-Foguet, 2019 ; Fuller et al., 2016) propose alternative statistical models to better capture the different trajectories present in the data.

To overcome the challenges described above, this study will develop a reproducible R package that derives the estimates from the raw data in addition to evaluating alternative statistical models and building a decision tree that maps out appropriate models depending on the number of data points and the trend a country has. This will ensure that estimates are accurate and consistent across countries, leading to more reliable reporting and decision-making.

# Research objectives

This study seeks to answer the following questions:

1. Is it feasible to reproduce indicator estimates for SDG indicator 6.2.1 using published raw data and documented methods?
2. Are alternative statistical models able to accurately assess different country trends and offer better performance to effectively monitor progress of achieving universal WASH coverage?

In addition to answering the questions above, this research aims to contribute to better reporting of the SDG 6 estimates by:

1. Opening up the data transformation process of computing JMP’s global estimates by using the same input data in a bid to attain the same results which is in line with its mission of producing accurate global estimates.
2. Addressing common issues around data management such as data stored in non machine-readable formats, such as PDFs or data being shared by email, and how this may deter progress towards universal access to sanitation (SDG 6.2)
3. Providing a user-friendly R package that follows FAIR principles for data management and stewardship (Wilkinson et al., 2016) for users to conduct their own analyses from the JMP data to answer new questions and better understand trends in access to sanitation and safe management
4. Developing a decision tree that selects models based on different country data scenarios

# Literature review

Data drives research and new inventions. Sharing data improves visibility, and makes the research process transparent which increases trust and enhances reproducibility of the results (Tierney & Ram, 2020). Good data management leads to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data has been published (Wilkinson et al., 2016). This section of the study will be two fold : The first section will focus on some of the data management principles as highlighted by different authors that overcome the barrier to reproducibility while the second section will focus on different methodologies employed by researchers to the WASH raw data to assess trends and monitor progress.

## Principles of data mangement and reproducibility.

The term Reproducibility was first noted by Claerbout and Karrenbach (Claerbout & Karrenbach, 1992) and has been used in scientific research ever since. According to the National Academies of Sciences, Engineering, and Medicine (National Academies of Sciences, 2019), Reproducibility means obtaining consistent results using the same input data, computational steps, methods, and conditions of analysis. Victoria Stodden (*Edge.Org*, n.d.) classifies reproducibility into empirical, computational and statistical. Empirical reproducibility includes non-computational empirical scientific experiments and observations. Computational reproducibility on the other hand entails detailed information on data, code and processes. Lastly, Statistical reproducibility considers new measures to assess the reliability and stability of statistical inferences, including developing new validation measures and expanding the field of uncertainty quantification to develop measures of statistical confidence and a better understanding of sources of error, especially when large multi-source datasets or massive simulations are involved (Community & Scriberia, 2022) . Reproducibility has gradually been gaining traction in the past decade with attempts to reproduce research in the social science field slowly increasing. When research is made transparent, other researchers are able to easily use the same processes, data and code to verify old findings or build upon it. This is particularly important in the context of Sustainable Development Goals(SDGs) where the accuracy and reliability of data are critical for assessing progress towards achieving these same goals.

(Claerbout & Karrenbach, 1992) in an effort to achieve reproducibility, create a simple filing system, ReDoc, that allows authors to easily reproduce results and computations using standardized rules and commands instead of recreating the work from scratch which is often a tedious process. It offers users four commands: burn the figures which removes the easily reproducible result files, build the figures which recomputes them and updates the easily reproducible result files, view the figures which displays the figures, and clean up and remove all intermediate results which removes all files other than source or result files making it more accessible. To better understand the four commands, they define three types of files that are linked to Rebocs four commands: Fundamental files which constitute datasets, scripts and makefiles, result files which are plot files (e.g. gif files) and lastly Intermediate files which lie between fundamental files and the result files such as executables and partially processed data.

(Gentleman & Lang, 2007) introduce the concept of a compendium as a container for the different elements that make up the document and its computations i.e. text, code and data which enables different researchers to reproduce any document contained in the compendium reliably across multiple computing environments to generate the researchers published view. A compendium contains one or more dynamic documents containing code and text chunks which requires programming software to generate its content. There are two ways in which a compendium can be processed: Transformation of the input data through code using languages such as R to yield output and combining this output data with textual descriptions to provide the end user with a narrative.

(Mesirov, 2010) recognize the need of having a paradigm that can also benefit those who do not know how to code to perform and publish their reproducible research. They propose a simple Reproducible Research System(RSS) for non programmers to use which contains two components: Reproducible Research Environment(RRE) where all the computational work is done with the ability to track analysis and results and package them for redistribution and Reproducible Research Publisher(RRP) which is a document preparation system that can then be linked to RRE. They implement this in a genomics environment where the Reproducible research environment represents the GenePattern computational genomics environment which is an environment used to analyze genomic data sets and the Reproducible Research Publisher is an adaptation of Microsoft Word that can link text, tables and figures to GenePattern.

(Peng, 2011) lists unavailability of code, lack of culture that requires reproducibility of scientific claims and lack of an integrated software infrastructure for sharing reproducible research with others as some of the barriers hindering computational reproducibility. He proposes the following solutions to the scientific community ; First, he recommends that the code be made available through publishing. Second, he recommends publishing a tidy version of the code alongside the data sets used in a durable non proprietary format. Lastly he recommends pooling all data, metadata and code in a centralized repository linked with respective publications. To ascertain the reproducibility of different authors’ research work, (Peng, 2011) through the Journal Biostatistics have introduced a policy where authors submit their code or data for review, the code is then run on the data to verify the published results. Authors whose research passes the reproducibility review receive an “R” mark.

(Sandve et al., 2013) in their research state that good reproducibility habits save time in the long run through reuse of code. They also note that technology plays a big part in enhancing efficient reproducibility. They list ten rules that each capture a specific aspect of reproducible research. The rules include: 1. Keeping track of how each result was produced(Analysis workflow) 2. Avoiding manual data manipulation steps and instead rely on program executions to reduce errors and increase efficiency. 3. Archiving the exact versions of all external programs used in order to reproduce a given result. 4. Version control all custom scripts using systems such as Git to systematically store code. 5. Recording intermediate results in standardized formats to help in uncovering bugs 6. Noting underlying random seeds for analyses that include randomness so as to generate identical results each time the code is run. 7. Storing raw data behind plots to allow easy retrieval and make modification to the plots. 8. Generating Hierarchical Analysis Output to allow inspection of detailed values underlying the summaries. This can be achieved through incorporating debug outputs in the source code 9. Connect textual statements eg claims and conclusions with enough detail to underlying results to allow for reevaluation and for other researchers to test claims 10. Providing public and easy access to all input data, scripts, runs and results

(Piccolo & Frampton, 2016) also describe similar techniques to (Sandve et al., 2013) for computational reproducibility and that overcome software limitations. They include: 1. Providing a detailed written description of the analytical process followed, the operating system and the software dependencies used. 2. Providing custom scripts that include information on text based commands to enable other researchers extend the analysis more easily 3. Indicate clearly which versions of the software and the respective dependencies should be installed 4. Using narratives alongside code using the literate programming technique that combines plots, code, data, equations and the narrative. This is especially useful for non technical users. One way of achieving this is through use of tools such as Jupyter notebook that combines code, data, text and mathematical equations. 5. Workflow management systems to facilitate execution of software where the output of one tool is used as input to additional tools. 6. Using virtual machines which can encapsulate an entire operating system and all software, scripts, code, and data necessary to execute a computational analysis and obtain similar results. A good example of this is VirtualBox. This overcomes the challenge of different outputs associated with different versions of operating systems. 7. Using software containers which are a lighter weight alternative to virtual machines and ease the process of installing and configuring dependencies. Similar to virtual machines, they encapsulate code, scripts and data into a single package that can be shared with others. A good example of a software container is Docker.

(Wilkinson et al., 2016) describes four principles (FAIR) for good data management and stewardship- Findability, Accessibility, Interoperability and Reusability. To make data findable, it should contain rich metadata, be indexed or registered in a searchable source and clearly and explicitly include an identifier. To make data accessible, it should be open, free and universally implementable and allow for an authentication and authorization procedure where necessary. Interoperability describes data that uses a formal, accessible and broadly applicable language for knowledge representation, vocabularies that follow fair principles and the data includes qualified references to other data/metadata. Lastly, Reusability describes data that is accurate and uses relevant attributes, has a clear and accessible usage license, is associated with detailed provenance and meets domain relevant community standards. They further outline six good practices for computing workflows with organized data, documented steps, and the project structured for reproducibility which include:

1. Data management which includes saving raw data, documenting all the different steps, creating analysis friendly data(tidy data) and using a unique identifier for each record. 2.Software which involves writing code that is readable, reusable and testable with clearly defined inputs and outputs.
2. Collaborating which makes it easy for new users to understand, cite and build on a project
3. Project organization in a logical and consistent directory structure to enable easy discovery and understanding of the project.
4. Tracking changes to facilitate easy retrieval by different users. An example of a tool used in tracking changes is Git which keeps track of any change made in a file
5. Writing manuscripts in a way that keeps an audit trail to reduce the chances of work being lost and ensure the work is available to different authors at all times.

Lastly, (Tierney & Ram, 2020) focus on the practical side of sharing data to improve reproducibility. They state that easy and repeated access to raw data for data analysis is a persistent challenge despite the advancements made in sharing code and computing environments. They list 8 sets of requirements to provide while sharing data which include:

1. README which is meant to provide a description of the data to the end user. It contains information on WHO collected the data, WHAT the data is, WHEN it was collected, WHERE it was collected, WHY it was collected and HOW it was collected.
2. Data dictionary which provides context on the nature and structure of the data. It contains variable names, variable labels, variable codes and special values for missing data.
3. License which provides rules on how everyone can modify, use and share the data
4. Citation which provides information about how you want your data to be cited. A Digital Object Identifier (DOI) is a unique identifier often used as a prerequisite for citation.
5. Machine readable metadata which is data that is in a format that can be processed by a computer i.e structured format allowing data to be searched and indexed online
6. Raw data represents data that has not been processed for use. It’s in its initial format before any cleaning has been performed
7. Scripts represent the file containing the commands, structured to be executed like a program.
8. Analysis ready data - This is the final data ready for analysis. It should be in tidy format where each variable must have its own column, each observation must have its own row and each type of observational unit forms a table.

A key to enhancing reproducibility as described by the different authors in this section is to make available the raw data and code used to generate the authors findings in a way that different authors can easily verify results and build on the study which can be achieved through documentation or advanced technologies such as software containers and virtual machines that encapsulate the data and code in a single package. There is no one size fits all and hence different researchers should aim to combine the different techniques to achieve the most value according to their needs.

## Statistical modeling

Linear regression used by the JMP to generate estimates for WASH may lead to inaccuracy where trends in coverage typically follow S-shaped curves with slower increase as countries approach 100% coverage (Bartram et al., 2014). Several alternative models including quadratic, logit, piecewise linear and generalized additive models have been suggested by different researchers to increase accuracy and reduce bias which we will discuss in detail in this section.

(Wolf et al., 2013) list three criteria to consider when selecting a model which include: the closeness of models to survey points, the simplicity, transparency and reproducibility of the model and finally the ability to generate estimates for countries with minimal data points. They apply a linear two-level model with a logit transformation of the dependent variable (percentage of the population using a particular technology) to restrict estimates and confidence intervals between zero and one(100%) and a cubic spline transformation of the main predictor(year of survey). They calculate an average slope and intercept across countries with residual variances as opposed to a separate slope and intercept for each country as per JMP’s documented methods (JMP,2018).

The equation for the linear two-level model they use is:

where:

* is the dependent variable with a logit transformation representing the percentage of the population using a particular technology i.e. access to improved water sources or improved sanitation in their paper.
* is the probability of having access to improved water sources or improved sanitation.
* represents the cubic spline transformation of the main predictor, time/year.
* is the intercept term.
* to are the coefficients of the cubic spline basis functions with knots at , , , .
* is the random effect for the two-level model, representing the unobserved country-specific factors that affect access to improved water sources or improved sanitation.

They estimate an average intercept and slope with residual variances across countries. The multilevel modeling they use is able to perform better than the traditional linear regression as it includes continuous time series estimates for all countries even when data points are scarce.

(Bartram et al., 2014) analyse the strengths and limitations of the current JMP approach and call for the use of weighted models as alternatives to linear regression with greater weights applied to larger surveys and those of higher quality to lead to higher accuracy and low bias.

(Fuller et al., 2016) start by analyzing different countries patterns on access to water and sanitation. They then classify the countries into having one of the following trends: 100% coverage, linear growth, linear decline, no change, saturation, acceleration, deceleration, negative acceleration, or negative deceleration using non linear modeling techniques for countries with 11 or more data points. To model the non linear relationships, they use the Generalized additive model(GAM). They conclude that using nonlinear allows for flexibility and are able to capture short term changes in a country’s rate of progress resulting in more accurate estimates.

(Pérez-Foguet et al., 2017) discusses the suitability of different statistical models against two criteria: the accuracy of estimates for different service levels and the ease of communication and replicability to policy makers. They employ an isometric log-ratio transformation and compare the ordinary least squares regression used by the JMP to a Generalized Additive Model(GAM) in which the linear form is replaced by smoothing functions. In addition, they also analyze three patterns of curvature: acceleration, deceleration and saturation.

The equation for the GAM model using the transformed data that they use is:

where:

* is the isometric log-ratio transformation of the percentage of the population using a particular service at time .
* is the time variable.
* is the intercept term.
* to are the smooth functions of time.
* is the error term.

They find that using GAM results in more accurate estimates compared to the standard linear regression especially in cases where countries show nonlinear trends.

(Ezbakhe & Pérez-Foguet, 2019) build on the work of (Pérez-Foguet et al., 2017) by comparing the estimates obtained with standard and compositional approaches, examining the difference between the ordinary least squares and generalized additive model(GAM) and assessing the magnitude of standard errors and confidence intervals for water and sanitation estimates. They conclude that water and sanitation data are compositional and hence should not be modeled with standard statistical analysis as log ratio transformations lead to more sound estimates.

## Research gap

Across all the research presented in the literature review above, it is evident that there is need to consider alternative statistical models that better account for the non linear trajectories present in the WASH data to effectively monitor the 2030 targets. Researchers have mostly focused on the Generalized Additive models (GAM) as an alternative to linear regression and in their analysis fail to include the different circumstances in which these models would offer better performance e.g. the number of data points a country has and its rate of progress. Majority also fail to do this reproducibly through sharing the code and process that gets to the final results. This study will focus on assessing other nonlinear models such as loess and higher order regression models and mapping out different scenarios where these models offer more accurate estimates. The research will also be done in a reproducible manner such that other users are able to easily build upon the results.

# Methodology

## Data

The JMP currently calculates four primary sanitation estimates for 230 countries. Separate estimates are made for both the rural and urban populations from 2871 data sets from 2000-2020, the national estimates are then derived from the weighted average using the most recent population from the United Nations (JMP, 2017). The estimates are outlined in below:

## Reproducibility of the JMP estimates

This study will begin by reproducing the four sanitation primary indicators for each of the 232 countries contained in the JMP database using the following estimation rules (JMP, 2017):

1. Interpolation using an ordinary least squares for data points that are at least five years apart for example if we have the last data point in 2016 and the earliest data point in 2008, a regression will be done from 2008-2016 using the following equation:

The model can be written as:

where represents the proportion of the population using a particular sanitation service level at time , represents the year , is the intercept, is the slope, a nd represents the error term at time .

1. Extrapolation if there are at least two data points using ordinary least squares forwards two years ahead from the last data point and two years backwards from the first data point.
2. Extension for four years backwards and forwards beyond the years covered by data inputs and extrapolation.

The above steps are effectively demonstrated in “Figure 1 ” below:

\end{figure}

Using R as a scripting language, the raw data will be extracted from each of the national excel country files downloaded from the official WHO/UNICEF JMP website (https://washdata.org/data). R will further be used for data cleaning and transformation. A linear model will then be applied to derive the estimates with the goal to reproduce the statistical methods applied by the WHO/UNICEF JMP using the same rules described above. This approach is further described in detail by (Bartram et al., 2014) and (Fuller et al., 2016)

The derived estimates will then be compared to the JMP estimates to and assessed for goodness of fit using the AIC, BIC and RMSE. The proposed methodological approach is shown in “Figure 2 below:

\end{figure}

## Alternative statistical models

This study will then fit different non linear models to the data points for each of the 232 countries using the same estimation rules applied to the ordinary least squares regression. These will include:

1. Local Regression(LOESS) : This is a non parametric smoothing function used to model non linear relationships by making a large number of quadratic or linear regression lines as a window moves along the x-axis. It is an important tool for modeling non linear relationships in the social sciences field (**jacoby2000loess?**). The exact equation is:

where is the locally estimated regression function at , and is the residual error at .

In the above equation equation, y represents the coverage/percentage of the population using a particular service level, and x represents time/individual year. The equation states that each observed y value is equal to the predicted value of the loess regression function at the corresponding x value, plus an error term .

1. Generalized additive model(GAM) : It is also used to model non linear relationships by assuming that the outcome variable can be modeled by a sum of arbitrary functions of each independent feature. This model is also employed by Pérez-Foguet et al. (2017) in their research. Unlike the mentioned studies, this research will not be performing a log transformation on the dependent variable. The standard GAM as per the below equation will be applied to the data points for each country

where is the intercept term, is the number of predictors (in this case, one predictor x), is the smooth function for the th predictor at the th observation, and is the residual error at the th observation.

Similar to the loess equation, y represents the percentage coverage of a particular service level while x represents each individual year for a particular country. The smooth function represents the relationship between x and y that is not linear, but is instead a more flexible, non-linear function that is estimated by the GAM.

1. Quadratic : These are regressions with a quadratic term added to it. As a good number of countries have a curved trajectory as noted by (Fuller et al., 2016), this research will extend the linear regression applied by the JMP by adding a quadratic term to capture this curved shape as shown in the equation below:

where is the coverage/percentage of the population using a particular sanitation service level at the th observation, is the time/individual year at the th observation, is the intercept term, is the linear coefficient, is the quadratic coefficient, and is the residual error at the th observation.

1. Splines : These are formed by joining two or more linear regression lines. The point where the lines intersect is referred to as a knot. This research will also fit a cubic spline using the below equation

where represents the coverage/percentage of population using a particular sanitation service level at time , represents the intercept, represents the slope, and represents the curvature of the spline. The terms represent the cubic spline basis functions, which are equal to if and 0 otherwise. The coefficients represent the additional contribution of each knot to the model, beyond the linear and quadratic terms.

## Use cases

To better illustrate the performance of the four statistical models listed above, this study will take a closer look at four use cases for countries with different trajectories and number of data points: 1. Sewer connection in rural and urban Afghanistan 2. Open defecation in rural and urban Tanzania 3. Improved sanitation in rural and Urban Brazil 4. Septic tank connection in rural and urban Uganda

## Goodness of fit

After fitting the models, this study will analyse performance using three popular measures: the Akaike information criterion(AIC), the Bayesian information criterion(BIC) and the Root Mean Square Error(RMSE), also employed by (Pérez-Foguet et al., 2017). The AIC and BIC are information criteria that balance the model’s goodness of fit with its complexity. The AIC is given by:

where is the likelihood of the model and is the number of model parameters. The BIC is given by:

where is the number of observations.

The RMSE is a measure of the difference between the predicted values of the model and the observed values. It is given by:

where is the observed value and is the predicted value for the th observation, and is the number of observations. This study will thereafter map out a decision tree to select a suitable model depending on individual country trajectories and number of data points.

*A/RES/70/1 Transforming our world: The 2030 Agenda for Sustainable Development*. (2030).

Bartram, J., Brocklehurst, C., Fisher, M. B., Luyendijk, R., Hossain, R., Wardlaw, T., & Gordon, B. (2014). Global Monitoring of Water Supply and Sanitation: History, Methods and Future Challenges. *International Journal of Environmental Research and Public Health*, *11*(8), 8137–8165. <https://doi.org/10.3390/ijerph110808137>

Cai, J., Zhao, D., & Varis, O. (2021). Match words with deeds: Curbing water risk with the Sustainable Development Goal 6 index. *Journal of Cleaner Production*, *318*, 128509. <https://doi.org/10.1016/j.jclepro.2021.128509>

Claerbout, J. F., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992* (pp. 601–604). Society of Exploration Geophysicists. <https://doi.org/10.1190/1.1822162>

Community, T. T. W., & Scriberia. (2022). *Illustrations from The Turing Way: Shared under CC-BY 4.0 for reuse*. Zenodo. <https://doi.org/10.5281/ZENODO.3332807>

*Data-based decision-making processes for WASH | SSWM - Find tools for sustainable sanitation and water management!* (n.d.).

*Edge.Org*. (n.d.).

Ezbakhe, F., & Pérez-Foguet, A. (2019). Estimating access to drinking water and sanitation: The need to account for uncertainty in trend analysis. *Science of The Total Environment*, *696*, 133830. <https://doi.org/10.1016/j.scitotenv.2019.133830>

Fleming, A., Wise, R. M., Hansen, H., & Sams, L. (2017). The sustainable development goals: A case study. *Marine Policy*, *86*, 94–103. <https://doi.org/10.1016/j.marpol.2017.09.019>

Fuller, J. A., Goldstick, J., Bartram, J., & Eisenberg, J. N. S. (2016). Tracking progress towards global drinking water and sanitation targets: A within and among country analysis. *Science of The Total Environment*, *541*, 857–864. <https://doi.org/10.1016/j.scitotenv.2015.09.130>

Gentleman, R., & Lang, D. T. (2007). Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, *16*(1), 1–23.

Kleespies, M. W., & Dierkes, P. W. (2022). The importance of the Sustainable Development Goals to students of environmental and sustainability studiesa global survey in 41 countries. *Humanities and Social Sciences Communications*, *9*(1), 1–9. <https://doi.org/10.1057/s41599-022-01242-0>

Mesirov, J. P. (2010). Computer science. Accessible reproducible research. *Science (New York, N.Y.)*, *327*(5964), 415–416. <https://doi.org/10.1126/science.1179653>

Morton, S., Pencheon, D., & Squires, N. (2017). Sustainable Development Goals (SDGs), and their implementation: A national global framework for health, development and equity needs a systems approach at every level. *British Medical Bulletin*, *124*, 1–10. <https://doi.org/10.1093/bmb/ldx031>

National Academies of Sciences, E. (2019). *Reproducibility and Replicability in Science*. <https://doi.org/10.17226/25303>

Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, N.Y.)*, *334*(6060), 1226–1227. <https://doi.org/10.1126/science.1213847>

Pérez-Foguet, A., Giné-Garriga, R., & Ortego, M. I. (2017). Compositional data for global monitoring: The case of drinking water and sanitation. *Science of The Total Environment*, *590–591*, 554–565. <https://doi.org/10.1016/j.scitotenv.2017.02.220>

Piccolo, S. R., & Frampton, M. B. (2016). Tools and techniques for computational reproducibility. *GigaScience*, *5*(1), 30. <https://doi.org/10.1186/s13742-016-0135-4>

*Progress on household drinking water, sanitation and hygiene 2000-2020: Five years into the SDGs*. (n.d.).

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLOS Computational Biology*, *9*(10), e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>

Sianes, A., Vega-Muñoz, A., Tirado-Valencia, P., & Ariza-Montes, A. (2022). Impact of the Sustainable Development Goals on the academic research agenda. A scientometric analysis. *PLOS ONE*, *17*(3), e0265409. <https://doi.org/10.1371/journal.pone.0265409>

Supply, W. J. W., & Programme, S. M. (2015). *Progress on sanitation and drinking water: 2015 update and MDG assessment*. World Health Organization.

Tierney, N., & Ram, K. (2020). *A Realistic Guide to Making Data Available Alongside Code to Improve Reproducibility*.

Water, U. (2018). *Sustainable Development Goal 6: Synthesis report 2018 on water and sanitation*. United Nations.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. <https://doi.org/10.1038/sdata.2016.18>

Wolf, J., Bonjour, S., & Prüss-Ustün, A. (2013). An exploration of multilevel modeling for estimating access to drinking-water and sanitation. *Journal of Water and Health*, *11*, 64–77. <https://doi.org/10.2166/wh.2012.107>