Journal of Environmental Management 261 (2020) 110202

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: http://www.elsevier.com/locate/jenvman

Research article

Modelling quantities and qualities (Q&Q) of faecal sludge in Hanoi, Vietnam and Kampala, Uganda for improved management solutions

Miriam Englund [a], Juan Pablo Carbajal [b], Amédé Ferré [a,c], Magalie Bassan [a,c], An Thi Hoai Vu [d], Viet-Anh Nguyen [d], Linda Strande [a,*]

[a] Eawag: Swiss Federal Institute of Aquatic Science and Technology, Sandec: Department of Water and Sanitation in Developing Countries, Dübendorf, Switzerland
[b] Eawag: Swiss Federal Institute of Aquatic Science and Technology, Siam: Systems Analysis, Integrated Assessment and Modelling, Dübendorf, Switzerland
[c] Ecole Polytechnique Fédérale de Lausanne, School of Architecture, Civil and Environmental Engineering, Laboratory for Environmental Biotechnology, Lausanne, Switzerland
[d] HUCE: Hanoi University of Civil Engineering, IESE: Institute of Environmental Science and Engineering, Hanoi, Viet Nam

ARTICLE INFO

Keywords:
Low-income
Fecal sludge
Sanitation
Septage
Demographic
Wastewater

ABSTRACT

The importance of faecal sludge management is gaining recognition. However, methods are still lacking to reasonably estimate the quantities and qualities (Q&Q) that need to be managed, which makes the planning for and implementing of management solutions quite difficult. The objective of this study was to collect and analyse Q&Q of faecal sludge at a citywide scale, and to evaluate whether "SPA-DET" data (=> spatially analysable - demographic, environmental and technical) could then be used as predictors of Q&Q of faecal sludge. 60 field samples and questionnaires from Hanoi and 180 from Kampala were analysed. Software tools were used in an iterative process to predict total solids (TS) and emptying frequency in both Hanoi, Vietnam and Kampala, Uganda. City-specific data could be predicted with types of "SPA-DET" data as input variables, and model performance was improved by analysing septic tanks and pit latrines separately. Individual models were built for TS concentrations and emptying frequency. In addition, a model was built across both cities for emptying frequency of septic tanks based on number of users and containment volume, indicating predictive models can be relevant for multiple cities. Number of users, containment volume, truck volume and income level were identified as the most common variables for the correction function. Results confirm the high intrinsic variability of faecal sludge characteristics, and illustrate the importance of moving beyond simple reporting of city-wide average values for estimations of Q&Q. The collected data and developed scripts have been made available for replication in future studies.

1. Introduction

Globally, one-third of the world's population is served by non-sewered sanitation, which is commonly referred to as faecal sludge management (Strande et al., 2014). Faecal sludge is produced and stored onsite in different types of containment technologies, such as pit latrines and septic tanks. Acknowledgement of the importance of faecal sludge management is rapidly increasing, as evidenced by inclusion in the sustainable development goals (SDGs) (United Nations, 2015). With adequate management of the entire service chain, faecal sludge management can provide sustainable sanitation and protection of public health (Dodane et al., 2012). However, there are many challenges remaining in order to reach this status, from instituting frameworks and responsibilities and integrated planning methodologies, to accessibility and affordability of emptying individual containments at the household level, as a result, faecal sludge frequently ends up being dumped directly into the urban environment (Schoebitz et al., 2017).

One major problem, is a lack of methods to determine the total quantities and qualities (Q&Q) of faecal sludge that need to be managed at community to citywide scales. This is difficult, as containments are typically underground, not standardized, and informally constructed with no official records. This can lead to over- or underestimation of required management and treatment capacity (Bassan and Strande, 2011; Fichtner and Associates, 2008), resulting in dysfunctional management and treatment systems. Attempts to determine Q&Q of faecal sludge at city-wide scales have been time and resource intensive

* Corresponding author. Ueberlandstrasse 133, Dübendorf, 8600, Switzerland.
E-mail address: Linda.strande@eawag.ch (L. Strande).

https://doi.org/10.1016/j.jenvman.2020.110202
Received 29 March 2019; Received in revised form 28 December 2019; Accepted 25 January 2020
Available online 2 March 2020
0301-4797/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

(Fanyin-Martin et al., 2017; Strande et al., 2018), begging the need for more efficient approaches to data collection (Strande et al., (in preparation)). In addition to systematic approaches to data collection, analytical approaches for data analysis are needed to increase the efficiency and validity of assumptions for management solutions that can be made from collected data. One difficulty is that although models have to account for variability they frequently neglect outliers, whereas in faecal sludge management the variability is so high that data segments separated by orders of magnitude are not considered outliers (Gold et al., 2017; Strande et al., 2014). This variability needs to be taken into account during data collection, to ensure the variability is captured (e.g. containment type, origin, retention time), and also in data analysis.

Advanced models have been developed for centralized, sewer-based wastewater treatment influents, operations and process control (Martin and Vanrolleghem (2014). There are many situations where models could be useful, for example knowing Q&Q of incoming faecal sludge to improve designs and operation of management and treatment solutions However, modelling attempts for faecal sludge have only just begun. Attempts at developing models for Q&Q of faecal sludge at macro-scales relevant for planning have included numerical modelling from a mass balance perspective based on empirical observations. For example accumulation rates (Brouckaert et al., 2013; Kimuli et al., 2016; Lugali et al., 2016; Todman et al., 2015), or characteristics of the accumulated sludge in containments (Elmitwalli, 2013). However, these models were based on observations of faecal sludge in containment at the single household level, which due to the high variability of faecal sludge, will likely not be representative for predicting trends on a larger-scale (i.e. community to citywide).

Prior to the design of faecal sludge management solutions, a pre-study is commonly carried out. The focus is often on questionnaires, with limited resources for characterization (Ross et al., 2016). The results are often unreliable due to the high variability of faecal sludge, in addition to the high variability of analytical results due to a lack of standard methods and rigorous quality assurance and quality control procedures (QA/QC). The most common approach has been to report simple average values, but with standard deviations that are frequently as high as mean values, it is clear that Q&Q of faecal sludge do not follow a normal distribution and additional summary statistics are required (Bassan et al., 2013; Gold et al., 2017; Strande et al., 2018).

A model is a systematic means of describing a system, and depending on objectives and resources, different models and model structures are used (James et al., 2013). Predictive models for management solutions can be built with more readily available input variables, that can be used to predict output variables that are more resource intensive to determine (James et al., 2013). For example, previous research has demonstrated that Q&Q of faecal sludge can be significantly different based on types of "SPA-DET" data (=> spatially analysable - demographic, environmental and technical) (Strande et al., 2018). Therefore, more readily collected types of data such as income level, water connection, containment type, and number of users, could potentially be used as predictors of Q&Q (Strande et al., 2018).

The objective of this study was to rigorously collect and analyse Q&Q of faecal sludge at a citywide scale, and to evaluate whether SPA-DET data could then be used as predictors of Q&Q of faecal sludge. This was based on the hypothesis that statistical relationships between Q&Q of faecal sludge and "SPA-DET" data exist. The approach to collect DET data through questionnaires during sample collection was developed due the lack of available "SPA-DET" data in low- and middle-income countries. Software tools were then used in an iterative process, in this case to predict TS and emptying frequency in both Hanoi, Vietnam and Kampala, Uganda. This approach was taken, as a modelling based approach could potentially reduce required costs by improving the efficiency or potential of collected data, while increasing the accuracy of predictions for the design and implementation of management and treatment solutions. In this work, the word "model" is used to refer to mathematical models.

2. Methods

2.1. Hanoi: sample collection and laboratory analysis

General city information for Hanoi together with detailed background information on the sanitation status can be found in the SFD report that is available for free online at https://sfd.susana.org/ (Brandes et al., 2016). 70 faecal sludge samples from onsite containments were collected from September 2013 to May 2014 throughout urban districts of Hanoi, 60 from households and 10 from public toilets. Households were defined as single-family households with less than 10 inhabitants, and had septic tanks receiving only black-water. Samples were collected from collection and transport vacuum trucks immediately following desludging of the septic tank contents. A core sampling device (tube) was used to take a representative sample from the access port located on the top of the truck tank. The public toilets also had septic tanks, and composite samples were taken during discharge at a treatment plant (beginning, middle and end of discharge in 1:2:1 ratio as described in Gold et al. (2017). The different sampling methods (i.e. in situ or ex situ) were necessary due to a lack of legal discharge locations for emptying trucks making sample collection difficult. The effect of sampling location was evaluated, and determined to be representative (Ferré, 2014). Sample collection of the 60 households also included a questionnaire to collect information on the number of household inhabitants, septic tank age and volume, faecal sludge age (defined as the time since the last emptying), use of additives, number of trucks required to empty the septic tank and whether the tank was partially or fully emptied.

According to QA/QC procedures, duplicate samples were taken for every tenth sample, and duplicate laboratory analyses were made for every eighth measurement for all laboratory analyses. A maximum relative error of 15% was observed. Only 2% of samples (n = 3) had a relative error between 10 and 15%, the remaining samples error was less than 10%. Samples were analysed with Standard Methods based on APHA (2012), including temperature (T), pH and electrical conductivity (EC) at the sampling site with probes (Hanna HI 99300, 99121 and 8424). TS, volatile solids (VS), total suspended solids (TSS), volatile suspended solids (VSS), chemical oxygen demand (COD), total nitrogen (TN), total phosphorus (TP), ammonium nitrogen ($N–NH_4^+$-N), orthophosphate phosphorus (PO_4^{3-}-P). Hach test kits and a Hach-Lange DR2800 spectrophotometer were used following manufacturers' directions. Samples for $P–PO_4^{3-}$-P and NH_4^+-N were additionally centrifuged (3820×*g*, for 10 min at room temperature) and filtered at 1.5 μm porosity (Whatman, 1827-110 Grade 934-AH) to address the high density and turbidity of the samples. Similar filters were used for TSS and VSS analyses. Finer porosity (e.g. 0.45 μm) could not be used due to clogging, but samples analysed with both filters were not significantly different. Nickel (Ni), lead (Pb), iron (Fe) and zinc (Zn) were analysed using inductively coupled plasma (ICP) based on the standard method 3120 metals by plasma emission spectroscopy #85 (APHA, 2012), assuming faecal sludge density of 1. Salmonella and *E. coli* were analysed in external laboratories based on local standard methods (TCVN 4884:2005, TCVN 6187-2:1996) based on APHA standard methods (APHA, 2012) at the laboratory of Institute of Environmental Technology, Vietnam academy of science and technology. *Ascaris lumbricoides* eggs were analysed by the National Institute of Malariology, Parasitology and Entomology of Vietnam.

2.2. Kampala: sample collection and laboratory analysis

General city information for Kampala together with detailed background information on the sanitation status can be found in the SFD report that is available for free online at https://sfd.susana.org/ (Schoebitz et al., 2016). 180 faecal sludge samples were collected in Kampala during 2013 to 2014, as previously described in Strande et al. (2018) (the full data set can be accessed open source at DOI: https://doi.

org/10.25678/0000TT.

2.3. Variable presentation and cleaning of analysed data

The data was considered to be comparable, as it was collected in both cities at the time of emptying operations when onsite containments were full. There were some differences in data collection methods that were deemed acceptable. In Hanoi the volume of emptied sludge was estimated by the number of trucks required to empty the containment, and whether the tank was partially or fully emptied as in Hanoi it is common that trucks make multiple trips to the same household. Whereas in Kampala it was estimated by containment volume, percent emptied, truck volume, and whether the truck was full, as it is more common in Kampala that a truck empties one containment in a trip, and the truck is full after emptying. If the estimated containment volume and the truck volume differed, it was always assumed that truck volume was more accurate. In Hanoi, sludge age was estimated at containment age, as for all samples except for seven, it was the first time that the containment was being emptied. Whereas in Kampala, it was estimated as time since last emptied. Income level data was available for Kampala (Strande et al., 2018), but not for Hanoi.

Variables that were considered as useful inputs for modelling complied with the following criteria:

1. The fraction of missing values in the whole data set is lower than 0.1
2. The variance of the variable is larger than 0.001 (ignore constant variables)

All remaining missing values were dropped.

After this screening, an exploratory analysis was performed (Kidwell, 1989; Tukey, 1980) aimed at identifying relevant variables from the initial data set. In this study, modelling results of TS and emptying frequency as a function of "SPA-DET" variables are reported, hence the a priori selection of variables was guided by prior understanding about these dependencies. With these, a model structure was proposed (see sec. 2.4) and small random subsets of all variables complying with the criteria given above were fed into these models. Variables that were rarely selected by these models where further ignored.

Table 1 shows the subset of variables that were deemed relevant (see data sets and doi https://doi.org/10.25678/0000TT for full set of variables). Input variables were normalized with their mean absolute deviation from the median, to not bias the exploratory analysis with units.

Additionally, an independence analysis was done to determine if the data could be separated by categories (resembling a decision tree (Safavian and Landgrebe, 1991)), which lead to two groups (pit latrine and septic tank) being modelled independently (see supplementary information for details).

Table 1
Selected a priori potential SPA-DET variables from Hanoi and Kampala data sets.

Input variables	ID in supporting information
Origin category[a,b]	OrCat
Containment type[a,b]	CoTyp
Sludge age as time since last emptied[a,b]	SludgeAge
Sludge volume emptied[a,b]	Vpumped
Truck volume[a,b]	TrVol
Number of users[a,b]	NUsers
Containment volume[a,b]	CoVol
Containment age[a,b]	CoAge
Water volume added during desludging[a]	WaterV
Income level[b]	IC

[a] Data collected in Hanoi.
[b] Data collected in Kampala.

2.4. Modelling scenarios

In the absence of a prior causal model, any of the variables present in the collected data could be used as output. TS and emptying frequency were selected as output variables (to be predicted by models here) due to their relevance for faecal sludge management and faecal sludge technology design. TS is one of the most common variables for designing faecal sludge treatment technologies (e.g. solids loading rates for drying beds (kg TS/m^2.yr)), and emptying frequency is important to prevent overflowing containments or to predict quantities arriving at treatment.

To define the structure of the models, two scenarios were conceptually described for TS and emptying frequency, respectively. These scenarios allowed us to select a reasonable set of input variables and to specify relevant prior information and beliefs about them. In scenario 1 (TS) a general model with no specific grounding in theory or field knowledge is used, whilst in scenario 2 (emptying frequency) a conceptual model of how containments fill up, grounded in mass balances is used. With this at hand, the data measured is regressed using Gaussian Processes (Rasmussen and Williams, 2006). The processing and regression of the data were carried out using GNU Octave (Eaton et al., 2015) and the GPML package (Rasmussen and Nickisch, 2010), respectively (note, the GPML package is available for GNU Octave at https://gitlab.com/hnickisch/gpml-matlab/-/releases. The source code to reproduce the results in this article can be found at https://doi.org/10.25678/0000TT). All models presented herein can be readily reused in new contexts (new data sets) by directly editing the provided scripts. The failure or success of the model will provide the sector with new insights on processes of faecal sludge generation in the field. See supplementary material for a detailed discussion.

Due to the high local variability of Q&Q of faecal sludge it was hypothesized that city-specific approaches are necessary. Hence the models presented herein were independently trained for each city and scenario, i.e. each model is trained from “scratch” or “tabula rasa” using the corresponding data set. For the city of Kampala the data was grouped based on the sanitation technology: septic tanks and pit latrines. A model was also calibrated for households with data from both cities (i.e. to explore the possibility of cross-city models for households).

The general structure of all the models is:

$$y = meanfunction(\vec{V}) + correctionfunction(\vec{V}) + noise \quad (1)$$

The model in Eq. (1) consists of two functions (mean and correction) and a noise term. The mean function depends on the variables in Table 1, and provides the baseline output of the model. This function is used herein to encode prior knowledge about the link between input variables and the output. The correction function is added to the mean in order to improve the prediction of the mean function. The correction function is nonlinear and decays rapidly when the inputs are far away from the training data (extrapolation). Hence the correction is present only in the region of the training data, and extrapolation is dominated by the mean function. The correction function is built from a covariance function (i.e. a function that controls the behaviour of the deviations away from the mean). The remaining data that cannot be explained by the two functions is ascribed to a noise term. Formal details can be found in Rasmussen and Williams (2006) and in the supplementary information.

2.5. Scenario 1: Total Solids (TS)

The model used for the Hanoi data set (42 out of 60 data points from household septic tanks) is composed of a linear mean function on the variables number of users, containment volume, and sludge age. The correction function is generated by a Matérn covariance function (with automatic relevance determination) on the variables: number of users, sludge age, and sludge volume emptied. Finally, the model assumes that the data is corrupted by normally distributed noise.

The model for the Kampala data set (180 data points with a wide

range of sludge origins) has the same structure except that the mean and covariance functions use the input variables, number of users, containment volume, time since last emptied, and income level. The model is trained separately to data corresponding to pit latrines and septic tanks.

2.6. Scenario 2: emptying frequency

The emptying frequency scenario is based on the assumption that a customer (e.g. household, public toilet, industry) calls an emptier when the volume of their containment reaches a threshold volume. This volume is determined by the inflow to the containment, which is proportional to the number of users per containment, and types of waste streams entering the containment, and also physical, chemical, and biological processes in the containment (Strande et al., (in preparation)). The proposed mean function for this model only accounts for physical processes on the containment, i. e. emptying frequencies are given by the ratio between storage capacity and the number of users:

$$Emptyingfrequency = \frac{Numberofusers^{a}}{Containmentvolume^{b}} \quad (2)$$

where the exponents a and b are left in the ratio as free parameters of the mean function. This model defines a multiplicative relation between the input variables and hence is not linear. However, this type of model is linear in a log-log scale, which was used as the natural scale for all variables. The logarithmic scale provides two benefits: 1) the model structure becomes linear in this scale; and 2) it helps to visualize the large variability of the data, which covers three orders of magnitude. Variability of this order of magnitude is representative of what is observed in literature (Gold et al., 2017; Strande et al., 2014).

Taking logarithms on both sides of Equation (2) renders a linear model in the logarithm of the variables. The logarithm of equation (2) is then used as the mean function of a Gaussian process and extended with a covariance function that depends on other variables present in the data set. This model is shown in Equation (3):

$$y = w_1 X_1 + w_2 X_2 + w_3 + f(\beta_1 X_1, \ldots, \beta_n X_n) + \xi \quad (3)$$

where y is the logarithm of the emptying frequency (or minus the logarithm of the emptying period), and X_i (i = 1, …,n) is the centered and normalized logarithm of the variables (1) in Hanoi: number of users, containment volume, truck volume, water volume added during desludging, and volume emptied, respectively (2) in Kampala: number of users, containment volume, containment age, truck volume, and income level, respectively. When y is plotted versus X_1 (number of users) or X_2 (containment volume), the parameters w_1, w_2, and w_3 are the slope and the intercept of a straight line. The relevance of each variable in the correction function is given by the value of the β_n coefficients. Finally, ξ

Fig. 1. Characterization results for 60 samples taken from household septic tanks in Hanoi, Vietnam.

models independent t-distributed noise, with zero mean value. Since the mean function was derived from mechanistic principles, the correction function was constrained: its maximum value is limited to a fraction of that of the mean function (for more details refer to supplemental information).

The emptying frequency was not directly measured during the sampling and is not part of the data set, hence the inverse of the time between emptying events ('time since last emptied') was used as a proxy for the emptying frequency.

3. Results

3.1. Quantities and qualities: Hanoi, Vietnam

The results of the characterization of samples collected in Hanoi is presented in Fig. 1. Despite consistent use of sampling methods and rigorous QA/QC during sample collection and analysis, the observed standard deviations for the characterization data were greater than 50% of the mean for all the parameters, except for pH, EC and P-PO_4^{3-}. The complete results of the characterization and questionnaire based data collected for this study are now available open source (DOI: https://doi.org/10.25678/0000TT.

The median concentrations of TS, VS, TSS, VSS, COD and TP were higher for the Hanoi faecal sludge than for the Kampala faecal sludge. However, the faecal sludge from Kampala had a greater variation, which could be due to the greater range of faecal sludge origins (i.e. not only septic tanks, not only households). The variations were significantly different between Hanoi and Kampala for $COD_{soluble}$, TN, NH_4–N and P-PO_4^{3-}. In Hanoi, rates of accumulation were calculated as described in Strande et al. (2018), and were also highly variable. The mean value was 67 L/cap.yr, the standard deviation was 102 L/cap.yr and the median 32 L/cap.yr.

3.2. Scenario 1: total solids (TS)

The TS concentrations were modelled separately for the two cities, and for septic tanks and pit latrines (Kampala).

3.2.1. Hanoi septic tanks

In Hanoi, the model predicts that TS increases with decreasing numbers of users, decreasing containment volumes, and increasing sludge age (further details in supplemental information). The correction function indicates the number of users and the sludge age as most significant to improve the model performance, as presented in Fig. 2. The model performance when plotting the predicted TS values against the trained model is presented in Fig. 3 (numerical values of parameters are given in supplemental information).

3.2.2. Kampala pit latrines

In Kampala, the model predicts that in pit latrines TS increases with an increasing number of users, decreasing containment volume, increasing time since last emptied, and increasing income level, as presented in Fig. 4. The correction function indicates that the number of users is the most significant variable to improve the model performance. The model performance for pit latrines and septic tanks in Kampala when the predicted values are plotted against the trained model is presented in Fig. 5 (numerical values of parameters are given in supplemental information).

3.2.3. Kampala septic tanks

In Kampala, the model predicts that in septic tanks TS increases with decreasing numbers of users, decreasing containment volume, increasing time since last emptied, and decreasing income level, as presented in Fig. 4. However, the time since last emptied has the highest relevance. The correction function indicates that income level is the most significant for improving model performance, as presented in Fig. 5. The prediction input variables could imply the same conclusion as

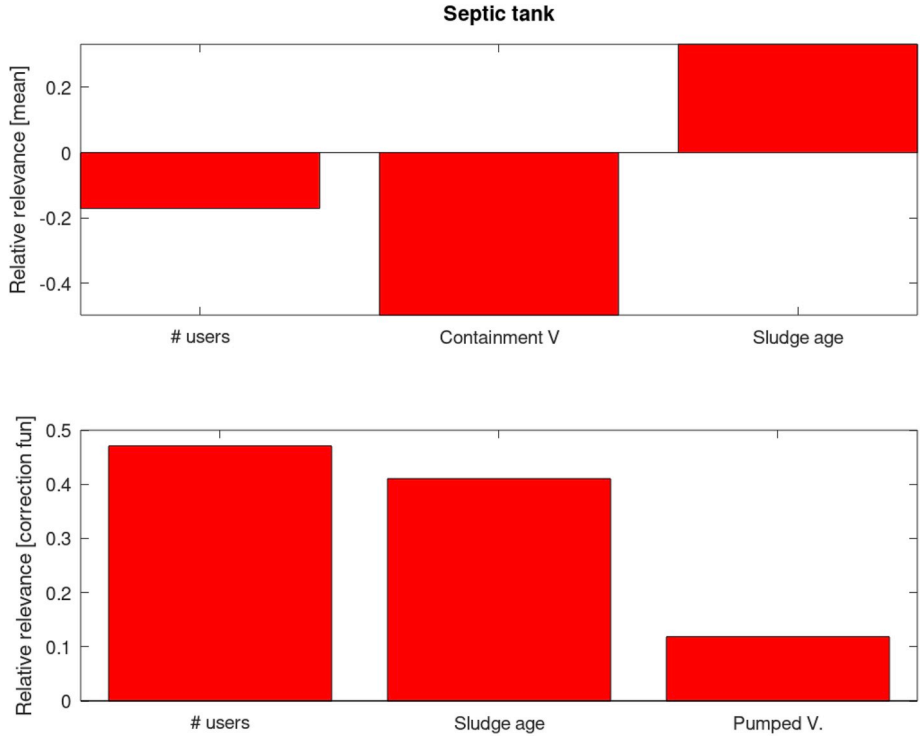

Fig. 2. Prediction of TS in Hanoi. Mean function parameters (top) in relative relevance (Number of users, Containment volume, Sludge age), negative values indicate that the model predicts lower TS concentrations and positive values higher TS concentrations. Correction function parameters (bottom) in relative relevance (Number of users, Sludge age, Sludge volume emptied).

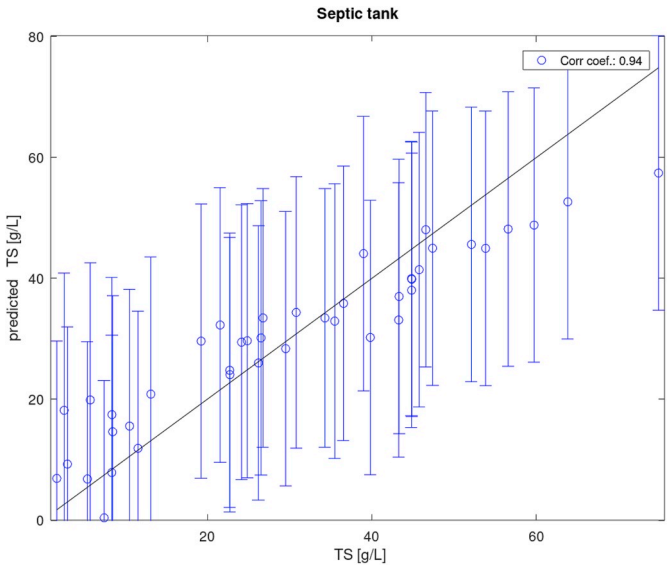

Fig. 3. TS Hanoi model vs. data. The plot illustrates the predicted TS against the trained model, using data from household septic tanks in Hanoi. Points on the diagonal line indicate perfect recovery of the training data. The correlation coefficient between predicted and observed values is shown in the legend.

was found in Berendes et al. (2017), namely that the level of income is strongly correlated to TS in the containment since the correction function indicates that the level of income is highly relevant to the model's performance. The relevance of income level for TS concentrations was also observed in Strande et al. (2018). With the modelling approach developed here, this relation can be used to predict TS concentrations. In Hanoi, predictions can also be made by other types of "SPA-DET" data (e.g. number of users, containment volume and sludge age).

3.3. Scenario 2: emptying frequency

Emptying frequency was also modelled separately for the two cities, and for septic tanks and pit latrines (Kampala).

3.3.1. Hanoi septic tanks

In Hanoi, the model predicts that emptying frequency in septic tanks increases with a decreasing number of users and increasing containment volumes, as presented in Fig. 6. The correction function indicates that containment volume is the most significant for improving the model's performance. The predicted emptying frequency in Hanoi against the trained model is presented in Fig. 7 (numerical values of parameters are given in supplemental information).

3.3.2. Kampala pit latrines

In Kampala, the model predicts that emptying frequency in pit latrines increases with an increasing number of users and increasing containment volumes, as presented in Fig. 8. The correction function indicates that truck volume is the most significant for improving the model's performance. The predicted emptying frequency in Kampala against the trained model for pit latrines and septic tanks is presented in Fig. 9 (numerical values presented in supplemental information).

3.3.3. Kampala septic tanks

In Kampala, the model predicts that emptying frequency in septic tanks increases with an increasing number of users and decreasing containment volumes as presented in Fig. 8. The correction function indicates that the income level is the most significant for improving the model's performance (numerical values of parameters are given in supplemental information).

The residual variability after regressing the model cannot be explained by the selected variables and is ascribed to noise (ξ) (equation (3)). In Kampala, pit latrines have higher noise, and are slightly less correlated with the data, than septic tanks, suggesting that the model is more robust for septic tanks.

3.3.4. Emptying frequency for both Hanoi and Kampala septic tanks

It was also evaluated if one predictive model could be built with data collected in both of the cities. This type of approach could be useful to investigate globally relevant factors for the prediction of Q&Q of faecal sludge. The emptying frequency of septic tanks with the model variables "number of users" and "containment volume" for Hanoi and Kampala are presented in a logarithm scale in Fig. 10, and the mean function parameters are presented in Fig. 11 (numerical values of parameters are given in supplemental information).

4. Discussion

4.1. Correction function

In Hanoi and Kampala, number of users, containment volume, truck volume and income level were the most frequently occurring parameters selected by the correction function for all of the above examples. Although the automatic relevance determination was not robust and sensitive to the choice of parameter ranges and priors, it is valuable to consider how to improve data collection (accuracy) for variables that have a strong influence. By identifying parameters that will improve the model performance, the quality of future data collection can be targeted and improved to increase the accuracy. For example, collecting data on how many times the containment is used, and how it is used (e.g. urination, or defecation and urination) over time, as opposed to total number of users (e.g. household members, or number of students in a school). These estimates are difficult to make only through questionnaires, and could be improved with tools such as automatic counters on doors (Brdjanovic et al., 2015; Zakaria et al., 2018). Usage patterns in public toilets, institutions, and commercial enterprises will be different than households, with higher number of users per system, but an increased frequency of urination versus defecation (Nguyen Viet et al., 2011). Other confounding factors include when students do not use sanitation facilities in school due to odour and uncleanliness (Xuan et al., 2012). Confounding factors also cannot be ruled out as to the relevance of truck volume as a relevant variable for correction, although a potential mechanistic link between truck volume and emptying frequency can be postulated. However, both Schoebitz et al. (2017) and Strande et al. (2018) identified that Q&Q of faecal sludge vary significantly amongst income levels. Income level is not the direct cause of Q&Q, but can be a predictor based on factors such as high-income areas having improved household water connections, better constructed containment, adequate resources for emptying services, and fewer users per toilet, all of which influence the volumes and characteristics of faecal sludge entering and exiting the containment, and hence emptying frequency. Other confounding factors can be poorer households only being able to afford partial emptying of containment, which would be reflected in the data as greater emptying frequency.

4.2. City specific models

The exploratory analysis of the TS input data for Kampala confirmed the benefit of building independent models according to containment type (i.e. pit latrine, septic tank), as was expected due to factors such as different usage patterns, flush volumes, containment lining, emptying practices and retention times (Fanyin-Martin et al., 2017; Strande et al., 2014). The observed differences based on containment type, also illustrate the importance of making estimates for citywide faecal sludge characteristics, by building up weighted averages based on observed differences in categories of collected data (SPA-DET), as opposed to applying universal averages across multiple types of containment

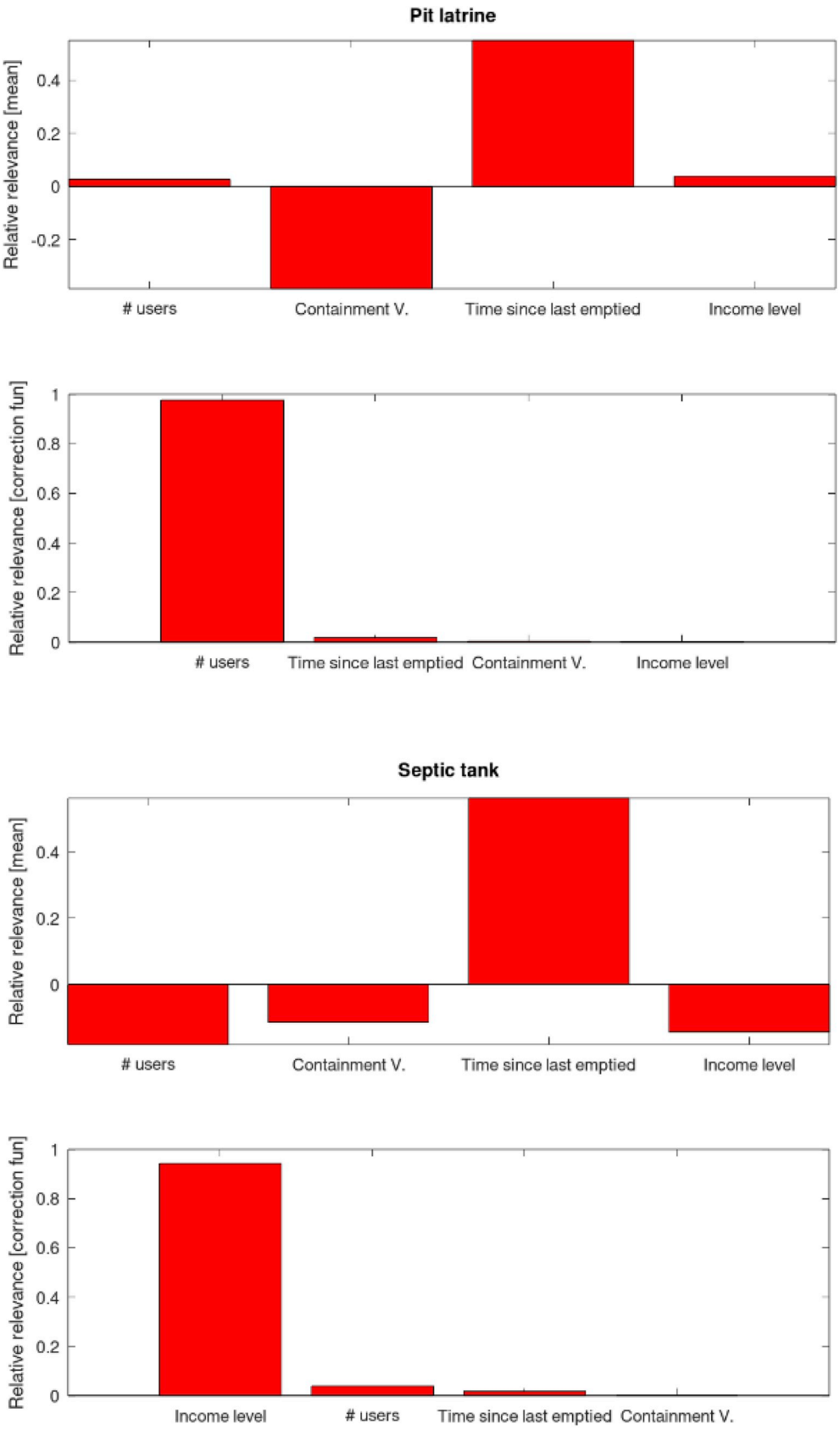

Fig. 4. Prediction of TS in Kampala. Mean function parameters (top and 2nd from bottom) in relative relevance (Pit latrine and Septic tank: Number of users, Containment volume, Time since last emptied, Income level), negative values represent a decreased TS, and positive values an increased TS. Correction function parameters (2nd from top and bottom) in relative relevance (Pit latrine: Number of users, Time since last emptied, Containment volume, Income level; Septic tank: Income level, Number of users, Time since last emptied, Containment volume).

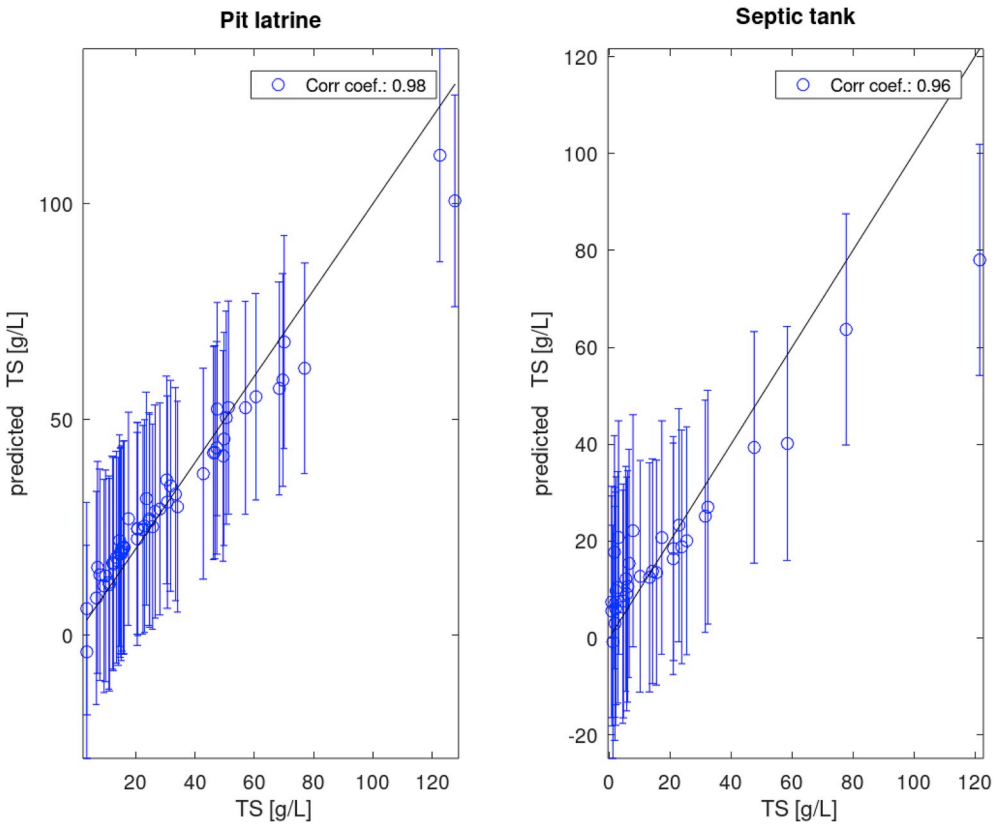

Fig. 5. TS Kampala model vs. data. The plots illustrate the predicted TS against the trained model, using data from Kampala. Points on the diagonal line indicate perfect recovery of the training data. The correlation coefficient between predicted and observed values is shown in the legend.

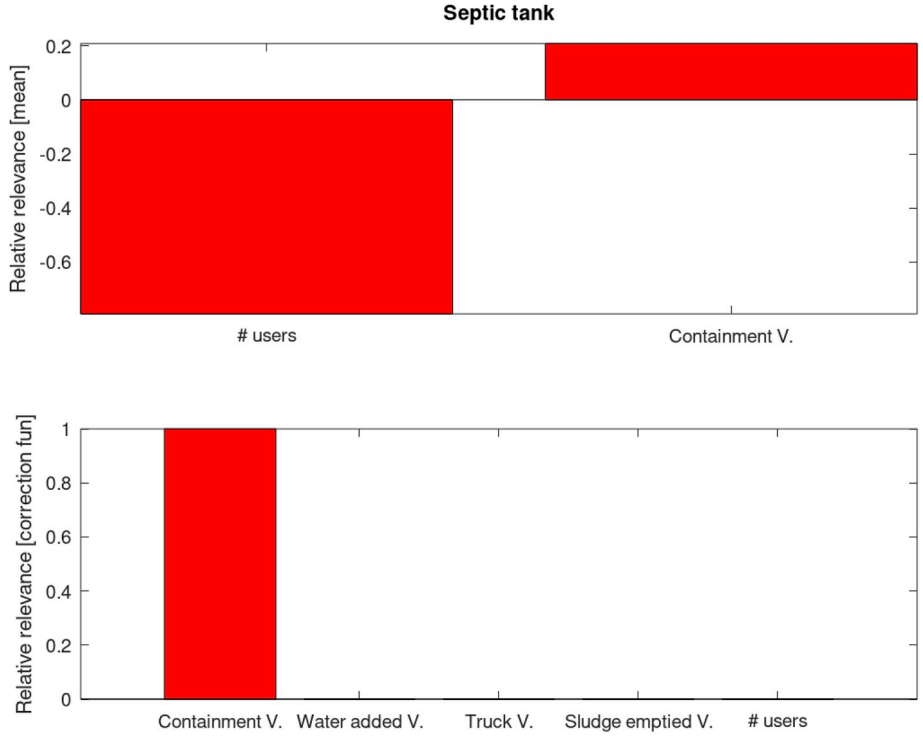

Fig. 6. Prediction of emptying frequency in Hanoi. Mean function parameters (top) in relative relevance (Number of users, Containment volume), negative values represent a decreased emptying frequency, and positive values an increased emptying frequency. Correction function parameters (bottom) in relative relevance (Containment volume, Water volume added during desludging, Truck volume, Sludge volume emptied, Number of users).

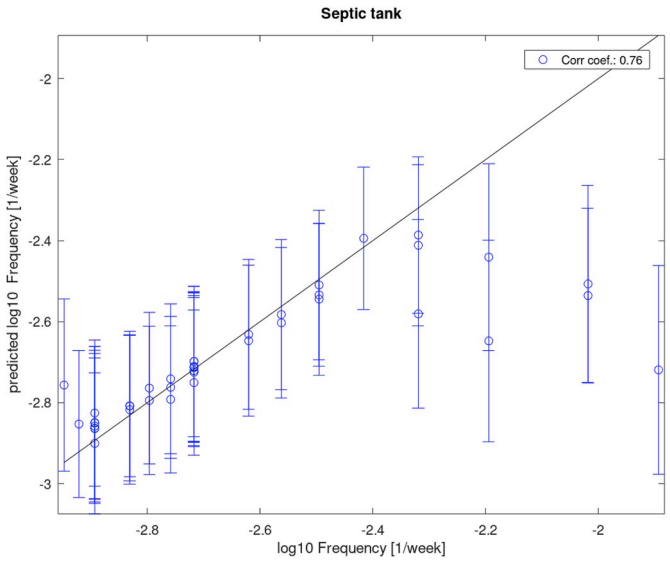

Fig. 7. Emptying frequency in Hanoi model vs. data. The predicted emptying frequency against the trained model, using Hanoi data. Points on the diagonal line indicate perfect recovery of the training data. The correlation coefficient between predicted and observed values is shown in the legend.

technologies.

In Kampala, predictions of emptying frequency for septic tanks fit better than for pit latrines. Q&Q of sludge from pit latrines is likely to have more scattered data (Brouckaert et al., 2013) than septic tanks, due to increased factors that can affect variability. For example, highly variable and poorly constructed containment (Mosler and Sonego, 2017), much higher numbers of users per containment volume (Günther et al., 2011; Isunju et al., 2013; Lugali et al., 2016), different decomposition processes (Torondel et al., 2016), and greater inflow and infiltration (Nakagiri et al., 2015). However, sampling bias and confounders cannot be entirely ruled out, and would require more data collection for a complete causal analysis to evaluate the most significant predictors (Peters et al., 2017). This also illustrates the importance of understanding that observed statistical relations in a data set are not equivalent to an understanding of the fundamental mechanisms generating the data.

In Kampala, the emptying frequency of pit latrines slightly increases with increasing containment volumes (Fig. 8), which seems counter-intuitive, especially when considering the assumptions that lead to the structure of the mean function (sec. 2.6). This could indicate that pit latrines in low-income areas have a much greater number of users, as they are commonly being used as communal or public toilets, with multiple households using the same pit latrine. In other words, because the number of users increases considerably faster than the containment volume, the emptying frequency increases with containment volume. However, although the containment volume available per user goes to zero for increasing number of users (plots in supplementary information), the same is observed for septic tanks, where the model is in agreement with the assumptions leading to the mean function (Fig. 8). Moreover, the small value of the coefficient (for both containment types) indicates that the emptying frequency is not very sensitive to the containment volume, and would require more data to determine its true value. Results could be improved with a targeted data collection, or even a study were variables were actively controlled for (e.g. all containments of the same volume). The latter would be very useful to eliminate potential confounders of this relation. These results in Kampala are in contrast to Hanoi, where decreasing number of users increase the emptying frequency. This has also been observed in South Africa (Buckley et al., 2008; Foxon et al., 2011; Still et al., 2005), possibly due to a lower average number of users per latrine with higher water consumption and income level.

4.3. Cross-city model

For emptying frequency of households (i.e. septic tanks, not including pit latrines), the data appears to be consistent across the two cities, despite the fact that the range of input variables is quite different for each city (Fig. 10). This implies that a single model could work for both cities, or that a model that fits in one city could reasonably be extrapolated to the other. This was not expected prior to data analysis, due to the large differences between the two cities. This result could be promising for the future development of cross-city or cross-country model development. However, it is important to note that two cities represent only two data points, and general assumptions for all cities cannot be made from this predictive model, which would require validation. Predictive data-driven models (or "black box" models) are based on associations, whereas in mechanistic models causal linkages of how input and output variables are related are known and described by a mathematical relation (more analogous to a "transparent box") (James et al., 2013). Although the same correlation for different cities can be evidence of an underlying mechanistic connection (e.g. microbial or physical relation), it cannot be assumed without testing. It is important that statistical correlations in one city are not simply extrapolated to another city unless the mechanism entailing the correlation is identified (see chapter 2.1 in Peters et al. (2017) for a discussion).

4.4. Implications

These results demonstrate that building city-specific predictive models for TS and emptying frequency for pit latrines and septic tanks in Hanoi and Kampala based on collected data is possible. Types of "SPA-DET" data such as income level, number of users, containment type, and containment volume can be used to predict TS concentrations and emptying frequency. In many low- and middle-income countries, there is a lack of available demographic, environmental and technical data, but it can be obtained with questionnaire data during sample collection. Another possibility is generating "SPA-DET" data with remote sensing data to identify indicators for informal or formal, income-level, residential or commercial areas of cities (Kohli, 2015).

The data can then be used in predictive models to aid with management and planning decisions. This was previously not possible, mainly due to a general lack of available data on citywide Q&Q of faecal sludge. The accuracy of models is directly dependent on the quality of collected data, illustrating the importance of developing standard methods of faecal sludge analysis and data collection, together with prior or expert knowledge used in developing models. This also illustrates the importance of moving beyond reporting average values for Q&Q of faecal sludge, which do not provide information such as distributions, ranges, and minimum and maximum values, and sharing raw data openly. In the future, this will make further and deeper exploratory analysis possible, which will lead to a better understanding of trends and predictors, which can also lead to better mechanistic understandings. Models will continue to be improved as more data becomes available, and can be used to more effectively target data collection in an iterative approach, where results inform future data collection, and model validation.

As well established as modelling for the influent of wastewater treatment plants is, mechanistic models of incoming Q&Q are still considered quite challenging (Martin and Vanrolleghem, 2014). In comparison, the modelling of faecal sludge is in its infancy, with much less available knowledge, and much higher variability of Q&Q (Chowdhry and Koné, 2012; Koottatep et al., 2012), meaning there are still quite some challenges to face. Nevertheless, faecal sludge modelling is starting to develop in collaborations between in-field practitioners and experts in empirical modelling, which facilitates the process of building model structures that can predict output variables, while keeping in

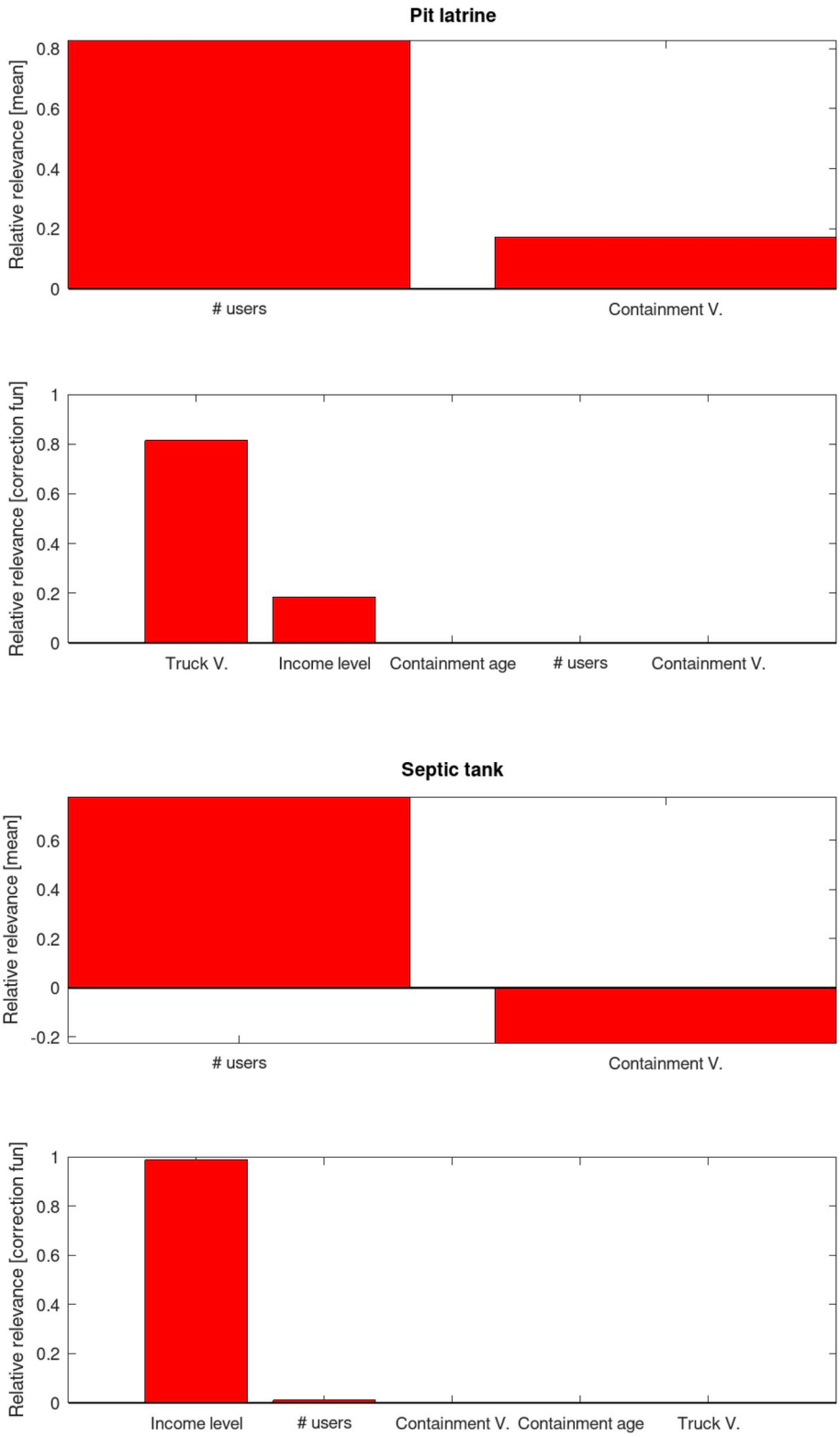

Fig. 8. Prediction of emptying frequency in Kampala. Mean function parameters (top and 2nd from bottom) in relative relevance (Pit latrine and Septic tank: Number of users, Containment volume), negative values represent a decreased emptying frequency and positive values an increased emptying frequency. Correction function parameters (2nd from top and bottom) in relative relevance (Pit latrine: Truck volume, Income level, Containment age, Number of users, Containment volume; Septic tank: Income level, Number of users, Containment volume, Containment age, Truck volume).

mind compounding factors such as illegal dumping, corruption, poor infrastructure and financial flows.

5. Conclusions

The results of the characterization confirm the high intrinsic variability of faecal sludge characteristics, and the fact that many factors influence the content of faecal sludge, independently from bias due to characterization methods. This is why it is important to not only report simple citywide average values for Q&Q of faecal sludge, and to break them down by for example containment type, and other types of categories of demographic, environmental and technical data. The results also indicate that when developing predictive models for faecal sludge management, independent models should be built according to

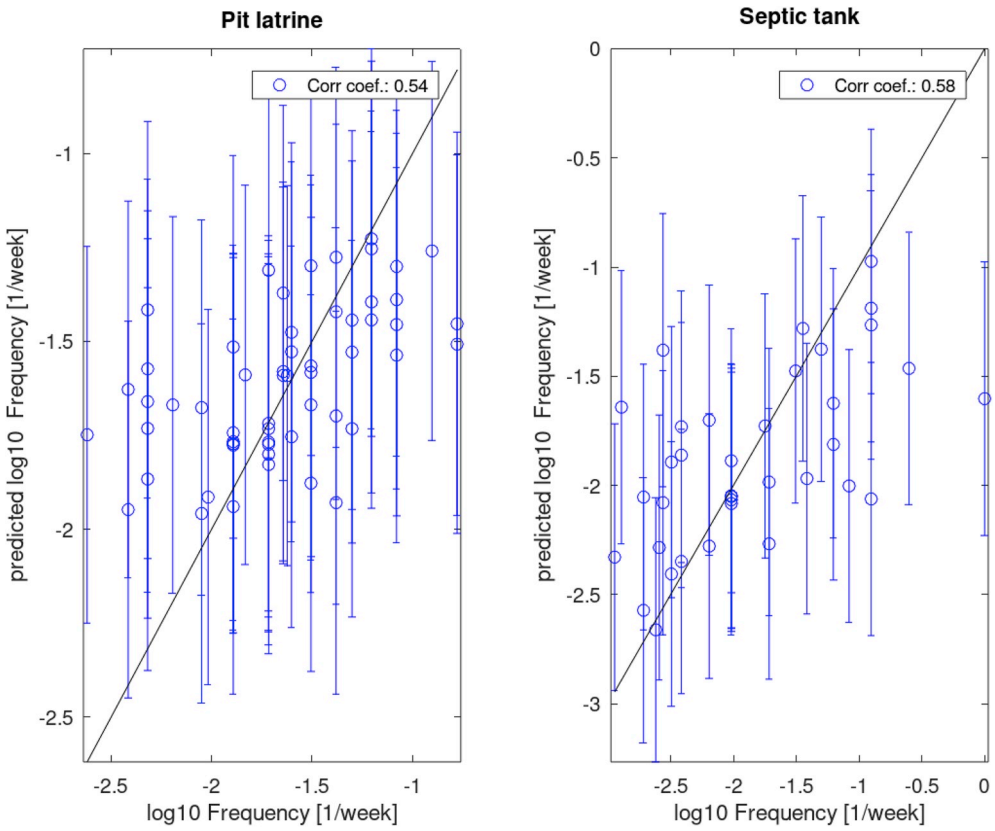

Fig. 9. Emptying frequency in Kampala model vs. data. The predicted emptying frequency for pit latrines and septic tanks against the trained model, using Kampala data. Points on the diagonal line indicate perfect recovery of the training data. The correlation coefficient between predicted and observed values is shown in the legend.

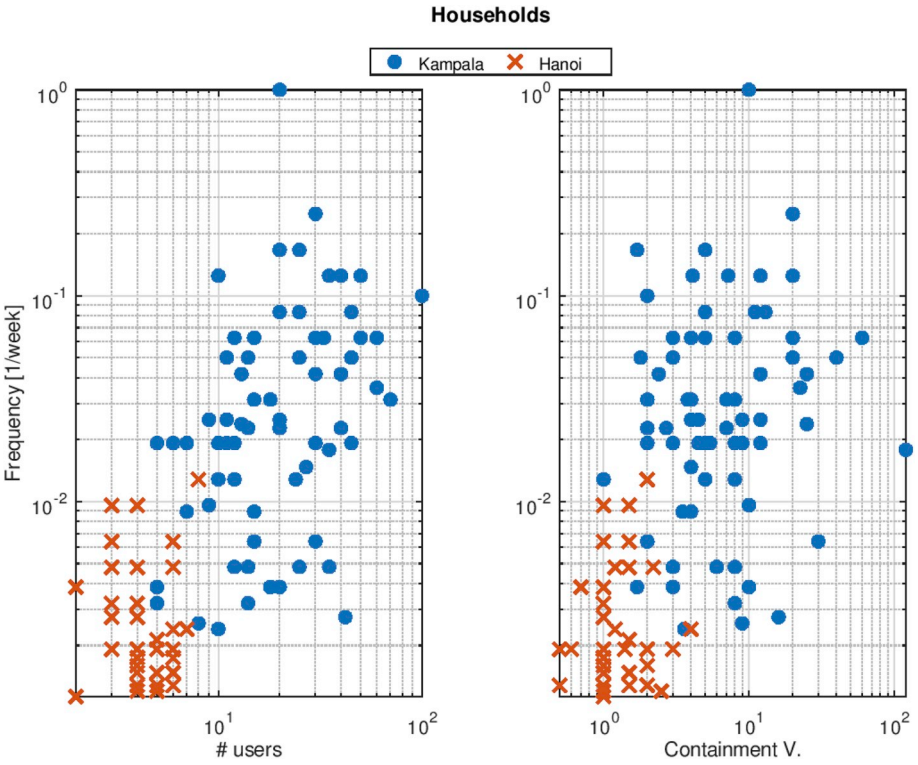

Fig. 10. Emptying frequency from single and multiple households from Hanoi and Kampala. The plots are in log-log scale.

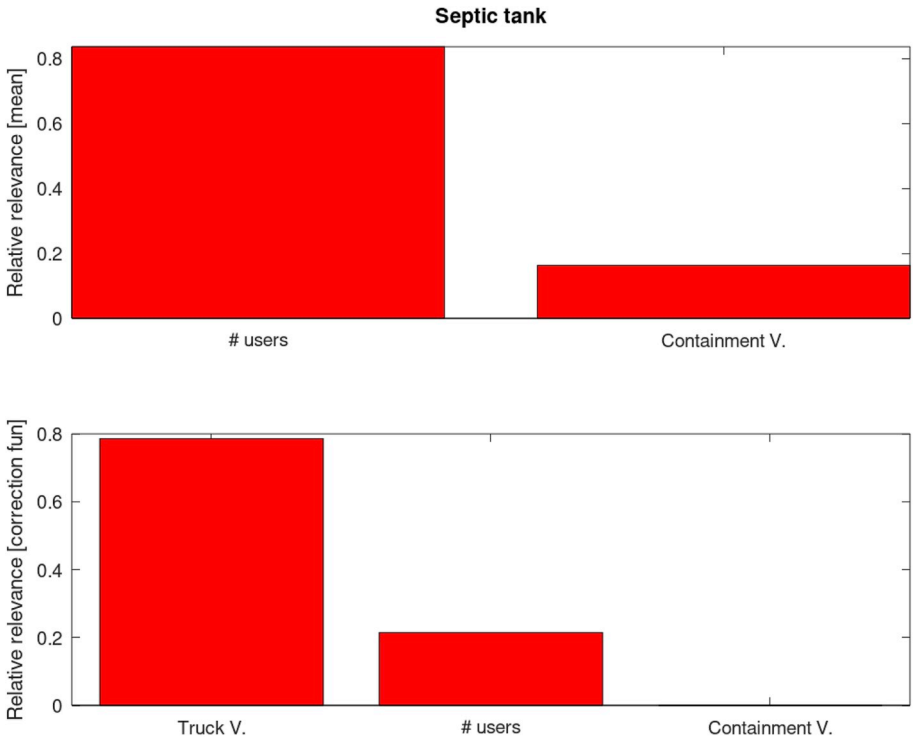

Fig. 11. Prediction of emptying frequency in Hanoi and Kampala. Mean function parameters (top) in relative relevance (number of users, Containment volume), negative values represent a decreased emptying frequency and positive values an increased emptying frequency. Correction function parameters (bottom) in relative relevance (Truck volume, Number of users, Containment volume).

containment type.

For the first time, citywide predictive models were developed for TS and emptying frequency of faecal sludge. These models were generated based on collected data. Hence, the results are applicable to other cities where collected data follows similar distributions, and they allow people to easily test them with their data. To transfer the model, first a visual inspection of the data (see box plots, Fig. 1) is needed to determine if the data is similar enough to fit the same equations. However, even if the new data is outside the range of this training data, the equations can be used as a starting place along with the provided scripts, allowing the model to be recalibrated. Mechanistic models can survive different scenarios because they are based on "facts" or relationships that are universal, but purely data-driven models can only hold up if the data in other regions is within the same range.

As more data sets become available, it will be interesting and informative to learn to what level predictive models can be transferred among cities and regions. With more reliable methods to collect and analyse data, the actual dynamic Q&Q of faecal sludge can be more accurately predicted. By identifying consistent patterns or relationships for different types of management structures, management strategies can more rapidly be developed and improved. As more data is collected in this fashion, it can potentially be used to develop mechanistic relationships that will help lead to universal understandings of faecal sludge management. This first step opens the door to explore models across multiple cities and countries in able to support practitioners in making informed decisions.

CRediT authorship contribution statement

Miriam Englund: Conceptualization, Writing - original draft. **Juan Pablo Carbajal:** Conceptualization, Methodology, Writing - original draft. **Amédé Ferré:** Formal analysis, Methodology. **Magalie Bassan:** Investigation, Methodology. **An Thi Hoai Vu:** Formal analysis. **Viet-Anh Nguyen:** Formal analysis. **Linda Strande:** Conceptualization, Funding acquisition, Methodology, Writing - original draft, Supervision.

Acknowledgement

This work was supported by the Swiss State Secretariat for Economic Affairs (SECO) and the Swiss Agency for Development and Cooperation (SDC). It was possible through collaborations with the Institute of Environmental Science and Engineering (IESE), and the Hanoi University of Civil Engineering.

The author contributions were as follows: Miriam Englund had a co-lead in writing the manuscript with input from authors and supported the development of the model; Juan Pablo Carbajal developed the model, took the lead in data analysis and the supplemental information, and contributed to writing; Amédé Ferré contributed to the development of specific sampling methods and laboratory protocols for faecal sludge and was responsible for carrying out all fieldwork and laboratory analysis in Hanoi; Magalie Bassan organized and supervised implementation of the data collection and initial data analysis; An Thi Hoai Vu and Viet-Anh Nguyen contributed to fieldwork and laboratory analysis in Hanoi; and Linda Strande conceived of the presented idea, had a co-lead in writing, contributed to model development, and supervised the project. Additionally, Nam Van Nguyen and Bang Trong Le contributed to the sampling campaign, and Lien Thuy Nguyen and Thu Thanh Nguyen to laboratory analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2020.110202.

References

APHA, 2012. Standard Methods for the Examination of Water and Wastewater, twenty-second ed. American Public Health Association (APHA), American Water Work Association (AWWA), Water Environmental Federation (WEF), Washington (DC).

Bassan, M., Strande, L., 2011. Capacity strengthening in sanitation: benefits of a research - operator collaboration. In: 35th WEDC International Conference. Loughborough University.

Bassan, M., Tchonda, T., Yiougo, L., Zoellig, H., Mahamane, I., Mbéguéré, M., Strande, L., 2013. Characterization of faecal sludge during dry and rainy seasons in Ouagadougou, Burkina Faso. In: Proceedings of the 36th WEDC International Conference, pp. 1–5. Nakuru, Kenya.

Berendes, D.M., Sumner, T.A., Brown, J.M., 2017. Safely managed sanitation for all means fecal sludge management for at least 1.8 billion people in low and middle income countries. Environ. Sci. Technol. 51, 3074–3083.

Brandes, K., Schoebitz, L., Nguyen, V.A., Strande, L. Hanoi, February 2016. Vietnam: Sanitation Service Delivery Context Analysis and Mapping of Excreta Flows along the Sanitation Service Chain and throughout the City. Eawag/Sandec.

Brdjanovic, D., Zakaria, F., Mawioo, P., Garcia, H., Hooijmans, C., Curko, J., Thye, Y., Setiadi, T., 2015. eSOS®–emergency sanitation operation system. J. Water, Sanit. Hyg. Dev. 5, 156–164.

Brouckaert, C., Foxon, K., Wood, K., 2013. Modelling the Filling Rate of Pit Latrines. Water SA, p. 39.

Buckley, C.A., Foxon, K.M., Brouckaert, C.J., Rodda, N., Nwaneri, C.F., Balboni, E., 2008. Scientific Support for the Design and Operation of Ventilated Improved Pit Latrines (VIPS) and the Efficacy of Pit Latrine Additives. Water Research Commission, South Africa.

Chowdhry, S., Koné, D., 2012. Business Analysis of Fecal Sludge Management: Emptying and Transportation Services in Africa and Asia. Bill & Melinda Gates Foundation, Seattle, USA.

Dodane, P.-H., Mbéguéré, M., Sow, O., Strande, L., 2012. Capital and operating costs of full-scale fecal sludge management and wastewater treatment systems in dakar, Senegal. Environ. Sci. Technol. 46, 3705–3711.

Eaton, J., Bateman, D., Hauberg, S., Wehbring, R., 2015. In: GNU Octave Version 4.0. 0 Manual: a High-Level Interactive Language for Numerical Computations. CreateSpace Independent Publishing Platform. March.

Elmitwalli, T., 2013. Sludge accumulation and conversion to methane in a septic tank treating domestic wastewater or black water. Water Sci. Technol. 68, 956–964.

Fanyin-Martin, A., Tamakloe, W., Antwi, E., Ami, J., Awarikabey, E., Apatti, J., Mensah, M., Chandran, K., 2017. Chemical Characterization of Faecal Sludge in the Kumasi Metropolis, Ghana. Gates Open Res. 1, 12. https://doi.org/10.12688/gatesopenres.12757.1.

Ferré, A., 2014. Assessment of the Impacts of Local Conditions and Sampling Methods on the Characteristics of Faecal Sludge. Environmental Engineering (SIE). Laboratory for Environmental Biotechnology (LBE). École polytechnique fédérale de Lausanne, Lausanne.

Fichtner, W.T., Associates, M.E., 2008. Main report. Kampala Sanitation Program (KSP)–Feasibility Study, vol. 1. KFW Entwiclungsbank, Fichtner Water & Transportation.

Foxon, K., Buckley, C., Brouckaert, C., Bakare, B.F., Still, D., Salisbury, F., 2011. How Fast Do Pits and Septic Tanks Fill up? Implications for Design and Maintenance, what Happens when the Pit Is Full? Developments in On-Site Faecal Sludge Managment (FSM). FSM Seminar Durban, South Africa, pp. 9–10.

Gold, M., Harada, H., Therrien, J.-D., Nishida, T., Cunningham, M., Semiyaga, S., Fujii, S., Dorea, C., Nguyen, V.-A., Strande, L., 2017. Cross-country analysis of faecal sludge dewatering. Environ. Technol. 1–11.

Günther, I., Horst, A., Lüthi, C., Mosler, H.-J., Niwagaba, C.B., Tumwebaze, I.K., 2011. Where do Kampala's poor "go"?-Urban sanitation conditions in Kampala's low-income areas.

Isunju, J.B., Etajak, S., Mwalwega, B., Kimwaga, R., Atekyereza, P., Bazeyo, W., Ssempebwa, J.C., 2013. Financing of sanitation services in the slums of Kampala and Dar es Salaam. Health (N. Y.) 5, 783–791.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.

Kidwell, P., 1989. Edward R. Tufte, the Visual Display of Quantitative Information (Book Review).

Kimuli, D., Zziwa, A., Banadda, N., Kabenge, I., Kiggundu, N., Kambugu, R., Wanyama, J., Tumutegyereize, P., Kigozi, J., 2016. Quantification of physico-chemical characteristics and modeling faecal sludge nutrients from Kampala city slum pit latrines. IJREAT. Int. J. Eng. Adv. Technol. 3 (6), 129–141.

Kohli, D., 2015. Identifying and Classifying Slum Areas Using Remote Sensing, PhD Dissertation. University of Twente.

Koottatep, T., Surinkul, N., Paochaiyangyuen, R., Suebsao, W., Sherpa, M., Liangwannaphorn, C., Panuwatvanich, A., 2012. Assessment of Faecal Sludge Rheological Properties.

Lugali, Y., Zziwa, A., Banadda, N., Wanyama, J., Kabenge, I., Kambugu, R., Tumutegyereize, P., 2016. Modeling sludge accumulation rates in lined pit latrines in slum areas of Kampala City, Uganda. Afr. J. Environ. Sci. Technol. 10, 253–262.

Martin, C., Vanrolleghem, P.A., 2014. Analysing, completing, and generating influent data for WWTP modelling: a critical review. Environ. Model. Software 60, 188–201.

Mosler, H.-j., Sonego, I.L., 2017. Improved latrine cleanliness through behaviour change and changes in quality of latrine construction: a longitudinal intervention study in rural Burundi. Int. J. Environ. Health Res. 27, 355–367.

Nakagiri, A., Kulabako, R.N., Nyenje, P.M., Tumuhairwe, J.B., Niwagaba, C.B., Kansiime, F., 2015. Performance of pit latrines in urban poor areas: a case of Kampala, Uganda. Habitat Int. 49, 529–537.

Nguyen Viet, A., Nguyen Hong, S., Dinh Dang, H., Nguyen Phuoc, D., Nguyen Xuan, T., 2011. Landscape Analysis and Business Model Assessment in Fecal Sludge Management: Extraction and Transportation Models in Vietnam, Final Report. Bill & Melinda Gates Foundation, December.

Peters, J., Janzing, D., Schölkopf, B., 2017. Elements of Causal Inference: Foundations and Learning Algorithms. MIT press.

Rasmussen, C.E., Nickisch, H., 2010. Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. 11, 3011–3015.

Rasmussen, C.E., Williams, C.K., 2006. Gaussian Process for Machine Learning. MIT press.

Ross, I., Scott, R., Joseph, R., 2016. Fecal Sludge Management: Diagnostics for Service Delivery in Urban Areas - Case Study in Dhaka, Bangladesh. World Bank Group.

Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics 21, 660–674.

Schoebitz, L., Niwagaba, C.B., Strande, L. Kampala, June 2016. Uganda: Sanitation Service Delivery Context Analysis and Mapping of Excreta Flows along the Sanitation Service Chain and throughout the City. Eawag/Sandec.

Schoebitz, L., Bischoff, F., Lohri, C.R., Niwagaba, C.B., Siber, R., Strande, L., 2017. GIS analysis and optimisation of faecal sludge logistics at city-wide scale in Kampala, Uganda. Sustainability 9, 194.

Still, D., Salisbury, R., Foxon, K., Buckley, C., Bhagwan, J., 2005. The challenges of dealing with full VIP latrines. In: Proceedings WISA Biennial Conference & Exhibition, pp. 18–22. Durban ICC, South Africa.

Strande, L., Englund, M., Carbajal, J.P., Scheidegger, A., (in preparation). Estimating quantities and qualities (Q&Q) of faecal sludge at community to citywide scales, as a chapter in Methods for Faecal Sludge Analysis. (IWA).

Strande, L., Ronteltap, M., Brdjanovic, D., 2014. Faecal Sludge Management - Systems Approach for Implementation and Operation. IWA Publishing, London.

Strande, L., Schoebitz, L., Bischoff, F., Ddiba, D., Okello, F., Englund, M., Ward, B.J., Niwagaba, C.B., 2018. Methods to reliably estimate faecal sludge quantities and qualities for the design of treatment technologies and management solutions. Environ. Manag. 223, 898–907.

Todman, L.C., van Eekert, M.H.A., Templeton, M.R., Hardy, M., Gibson, W.T., Torondel, B., Abdelahi, F., Ensink, J.H.J., 2015. Modelling the fill rate of pit latrines in Ifakara, Tanzania. J. Water, Sanit. Hyg. Dev. 5, 100–106.

Torondel, B., Ensink, J.H., Gundogdu, O., Ijaz, U.Z., Parkhill, J., Abdelahi, F., Nguyen, V. A., Sudgen, S., Gibson, W., Walker, A.W., 2016. Assessment of the influence of intrinsic environmental and geographical factors on the bacterial ecology of pit latrines. Microb. Biotechnol. 9, 209–223.

Tukey, J.W., 1980. We need both exploratory and confirmatory. Am. Statistician 34, 23–25.

United Nations, 2015. Transforming Our World: the 2030 Agenda for Sustainable Development. Resolution Adopted by the General Assembly.

Xuan, L., Hoat, L.N., Rheinländer, T., Dalsgaard, A., Konradsen, F., 2012. Sanitation behavior among schoolchildren in a multi-ethnic area of Northern rural Vietnam. BMC Publ. Health 12, 140.

Zakaria, F., Ćurko, J., Muratbegovic, A., Garcia, H.A., Hooijmans, C.M., Brdjanovic, D., 2018. Evaluation of a smart toilet in an emergency camp. International journal of disaster risk reduction 27, 512–523.