KWR 2022.138 | November 2022

# Finding a relation to estimate the maximum simultaneous water demand of a number of households, using real-life and synthetic data

KWR

Bridging Science to Practice

# Report

## Finding a relation to estimate the maximum simultaneous water demand of a number of households, using real-life and synthetic data

Internship 31-8-2022 till 11-11-2022

**KWR 2022.138 | November 2022**

**Project number**

-

**Project manager**

-

**Client**

Derk Rouwhorst (WMD)

**Author(s)**

M.K.(Marieke) Berlin BSc. (Eindhoven University of Technology)

**Internship supervisors**

dr. ir. K.A. (Karel) van Laarhoven (KWR)

dr. ir. J. (Jaron) Sanders (Eindhoven University of Technology)



**Keywords**

Daily maximal flow, Simultaneous water demand, SIMDEUM

Year of publishing
2022

More information
Karel van Laarhoven
T +31 6 53196485
E karel.van.laarhoven@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

# Summary

To determine the size of a water pipe in the tertiary water distribution network, two conditions are of importance. Firstly, the customer has to receive a minimal water pressure at all times according to the Dutch drinking water decree (Drinkwaterbesluit - Artikel 45). Secondly, preferably the water in the pipe reaches a minimal velocity approximately once a day. This will ensure that the pipe has a self-cleaning property (Vreeburg, 2007). However, to correctly estimate the water pressure and the velocity under these reference conditions, one needs to be able to estimate the water demand first.

This report will investigate the relation between the maximum summed water demand of $N$ households and $N$. Currently this is done using the 'q-square-root-N' rule. Which states that the $\mathrm{Q\text{-}max}$ can be estimated by the following function $f(N) = q \cdot \sqrt{F \cdot N}$ (in general q=0.083 and F is recommended to be equal to 15 which then gives $f(N) = 0.32\sqrt{N}$ (E.J.M. Blokker, 2010). However, this rule overestimates the maximal summed water demand and some alternatives have been proposed, for example based on simulation results in (E.J.M. Blokker, 2010). This report will reinvestigate this relation using real-life and synthetic data. To determine the maximum of the summed water demand of a hypothetical water pipe which supplies $N$ households, the available demand patterns of different households will be combined. The maximum of the summed water demand will be called $\mathrm{Q\text{-}max}$.

Four quantities will be investigated. The first two investigated because the customers should be supplied with a minimal pressure (almost) always. On the one hand side the maximal $\mathrm{Q\text{-}max}$ of a set of days, which will be called $\mathrm{max(Q\text{-}max)}$. On the other hand, WMD wants to guarantee that the minimal pressure is guaranteed on the day on which the most water is used, called the max-day. The quantity that will be investigated this requirement will be the $\mathrm{Q\text{-}max}$ corresponding to the max-day and will be denoted with $\mathrm{Q\text{-}max_{max}}$. The other two quantities of interest are investigated because preferably the velocity of the water reaches a minimal level regularly. Firstly, the median of the $\mathrm{Q\text{-}max}$ of a set of days will be determined (denoted by $\mathrm{med(Q\text{-}max)}$). On the other hand, WMD would prefer that the minimal velocity is reached on an 'average' day. For this the median-day will be introduced. The median-day is the day on which the total water demand is equal to the median of all total water demands of all measured days. The $\mathrm{Q\text{-}max}$ corresponding to this day will be denoted with $\mathrm{Q\text{-}max_{med}}$.

These four quantities will be estimated with 2 real-life data sets (measurement frequencies are one second and one hour) and one synthetic data set which was generated with SIMDEUM (frequency is one second). SIMDEUM is a stochastic model which creates water demand patterns of single households based on the end-use of a household.

It has been shown that $\mathrm{Q\text{-}max_{max}}$ and $\mathrm{max(Q\text{-}max)}$ are not the same for the available data sets. This implies that the $\mathrm{Q\text{-}max}$ on the maximum day is smaller than the maximum of a set of $\mathrm{Q\text{-}max}$. Moreover, it was found that if the measuring frequency is an hour (or approximately every hour), the estimates of $\mathrm{med(Q\text{-}max)}$, $\mathrm{Q\text{-}max_{med}}$, $\mathrm{Q\text{-}max_{max}}$ and $\mathrm{max(Q\text{-}max)}$ are a lot smaller than if the measuring frequency is a second. Therefore it is recommended that the $\mathrm{max(Q\text{-}max)}$ is estimated instead of $\mathrm{Q\text{-}max_{max}}$ and that data is used in which the measurement frequency is one second.

It was observed that the 'q-squareroot-N' rule that is used in the Netherlands (with Q= 0.32 if FU=15) overestimates the $\mathrm{med(Q\text{-}max)}$, $\mathrm{Q\text{-}max_{med}}$, $\mathrm{Q\text{-}max_{max}}$ and $\mathrm{max(Q\text{-}max)}$. But rather that to estimate $\mathrm{med(Q\text{-}max)}$ for a given $N$ one can compute $0.17\sqrt{N}$ for $N < 61$ and $0.01 \cdot N + 0.66$ for $N \geq 61$. To estimate $\mathrm{max(Q\text{-}max)}$ for a given $N$ one can compute $0.26\sqrt{N}$ for $N \leq 66$ and $0.016 \cdot N + 1.04$ for $N > 66$.

# Contents

# 1 Introduction

## 1.1 Motivation and goal

The water distribution network is of vital importance, as it is essential to provide sufficient drinking water to the customers. However, designing the water distribution network comes with many challenges. One of these challenges is to determine the size of each pipe. If the pipe diameters are too small, the water pressure drops too much when water is demanded by many customers at the same time. On the other hand, if the pipe diameters are too large, the flow of the water is low. This causes sediment particles to accumulate and incidentally resuspend, this phenomenon is the main cause of brown water (Vreeburg, 2007). Thus, two conditions are of importance, the minimal water pressure when demand is high and a minimal velocity which can prevent the phenomenon of incidentally resuspension and accumulation of particles in the network.

The two conditions can be quantified by the following rules. Firstly, the drinking water decree states that 1000 liters of water should be available for delivery during any hour of a day while the water pressure is at least 150 kPa (Drinkwaterbesluit - Artikel 45). Secondly, a solution to prevent brown water, as suggested by Vreeburg (2007), is to design the system such that the water distribution network is a self-cleaning network. A self-cleaning network is a water distribution network that attains a velocity of at least 0.4 m/s once a day. The water utility WMD therefore prefers to have a velocity of 0.4 m/s at least once a day on most days.

However, to correctly estimate the water pressure and the velocity under these reference conditions, one needs to be able to estimate the water demand first. Therefore, when considering the demand that is to be satisfied through any given pipe, it is of importance to investigate the summed drinking water demand of all $N$ households that are supplied by it. For our purposes specifically, it is sufficient to be able to determine the maximal flow in the pipe during the day, which will be notated by $\mathrm{Q\text{-}max}$. The $\mathrm{Q\text{-}max}$ is determined by the, possibly simultaneous, water demand of the households supplied by the pipe. Note that, determining the $\mathrm{Q\text{-}max}$ of the summed drinking water demand of $N$ households is not equal to adding the $\mathrm{Q\text{-}max}$ of the single households since it is unlikely the peaks occur at the same time. Since it is unclear if and when households use water at the same time, it is nontrivial to determine the $\mathrm{Q\text{-}max}$ of the summed water demand of $N$ households.

In this report the correlation between $N$ and $\mathrm{Q\text{-}max}$ (respectively the number of households supplied by a pipe and the maximal summed flow during a day) will be investigated. This correlation is historically characterized by the so-called 'q-square-root-N' rule, internationally and in the Netherlands (Buchberger, Blokker, & Cole, 2012). This rule states that $\mathrm{Q\text{-}max}$ is equal to $q \cdot \sqrt{N \cdot F}$ where $q$ is a constant equal to 0.083, $F$ is the number of fixture units (in general $F = 15$ is recommended) and $N$ the number of households that are supplied by the pipe. Some alternatives to this rule have been proposed, for example, by Blokker (2010). This report will investigate whether this rule describes the correlation correctly or that a different function could describe the correlation more accurately.

## 1.2 Approach and reading guide

To determine the correlation between the summed $\mathrm{Q\text{-}max}$ and $N$, the water demand of a combination of households needs to be investigated. To investigate the water demand of a household, one could either analyse real-life water demand data or use a simulation to approximate the water demand. Even though simulations generate synthetic data, it is possible to generate very large amounts of data. For this report the synthetic data is generated by a stochastic model called SIMDEUM. SIMDEUM has been extensively validated in (E.J.M. Blokker, 2010). Both the

available real-life data and the generated synthetic data will be used to investigate the correlation between $N$ and Q-max.

The structure of this report is as follows. First, the problem will be described in a more formal setting and the notation will be introduced in Chapter 0. Second, the data that will be used (real-life and synthetic) will be described and a summary of SIMDEUM will be given in Chapter 0. Then, the methods used to analyse the data will be explained and the results will be given and compared in Chapters 0 and 0 respectively. Finally, conclusions will be drawn and possible directions for future research will be given in Chapter 0.

# 2 Formal problem formulation

As described in the introduction, the quantity of interest is called $\text{Q-max}$. Given a waterpipe $P$ in the tertiary system, let $H_P$ be the set of households supplied by $P$. Furthermore, let $N = |H_P| =$ {be the number of households supplied by $P$}. Let $t_0, \dots, t_T$ be the times at which the flow was measured, where $T$ is the total amount of measurements on day $d$. Now let $X_{d,h}(t_i)$ be the total amount of water used by household $h$ in $(t_{i-1}, t_i]$ on day $d$. Now, the amount of water used by all households $h \in H_P$ in $(t_{i-1}, t_i]$ with $i \in \{1, \dots, T\}$ on day $d$ is defined as $X_{d,P}(t_i) = \sum_{h \in H_P} X_{d,h}(t_i)$. Now, $\text{Q-max}$ can be defined as $\text{Q-max}_{d,P} = \max_{t \in \{t_1, \dots, t_T\}} X_{d,P}(t) = \max_{t \in \{t_1, \dots, t_T\}} \sum_{h \in H_P} X_{d,h}(t)$. As can be seen $\text{Q-max}_{d,P}$ is dependent on the waterpipe $P$ (and thus on $N$) and on the day $d$. However, the interest of this project was not to determine the value of $\text{Q-max}_{d,P}$ of any specific waterpipe or on any specific (deterministic) day. The interest lies with a general correlation between the number of households supplied by a pipe on an 'median day' and 'maximal day' and the $\text{Q-max}$ on these day. The interest lies with a 'median day' to ensure that the self-cleaning velocity is attained at at least half of the days. Moreover, to ensure that the minimal water pressure is (almost) always reached the 'maximal day' is considered. Therefore, the maximal flow of an 'median day' and a 'maximal day' will be investigated.

The 'median day' and 'maximal day' can be defined in different ways. Two ways will be used and explained in this report. Firstly, given data of days $d_0, \dots, d_m$ the median and maximum of the set of $\text{Q-max}_{d_i,P}$ with $i \in \{0, \dots, m\}$ are considered. These quantities will be useful to evaluate the pressure and velocity in the pipe on (almost) every day and on approximately half of the days. Secondly, the $\text{Q-max}_{\text{median-day},P}$ and $\text{Q-max}_{\text{max-day},P}$ are considered. To define the median- and max-day some extra notation will be given. Let $\hat{X}_{d,P} = \sum_{j=0}^{T} X_{d,P}(t_j)$, which is the total amount of water used by the households supplied by pipe $P$ on day $d$. Given data of days $d_0, \dots, d_m$, let $\hat{X}_{(i),P}$ with $i \in \{0, \dots, m\}$ be the order statistic of $\hat{X}_{d,P}$. Meaning that $\hat{X}_{(i),P}$ is the i'th smallest value of $\{\hat{X}_{d_0,P}, \dots, \hat{X}_{d_m,P}\}$.
Now the median-day can be defined as:

$$\text{median-day} = \begin{cases} d_i\colon \hat{X}_{d_i,P} = \hat{X}_{\left(\left\lfloor \frac{m+1}{2} \right\rfloor\right),P} & \text{w.p. } 0.5 \\ d_i\colon \hat{X}_{d_i,P} = \hat{X}_{\left(\left\lfloor \frac{m+1}{2} \right\rfloor\right),P} & \text{w.p. } 0.5 \end{cases}, \quad (1)$$

,

and the max-day can be defined as:

$$\text{max-day} = d_i\colon \hat{X}_{d_i,P} = \hat{X}_{(m),P}\,. \quad (2)$$

Thus $\text{Q-max}_{\text{median-day},P}$ and $\text{Q-max}_{\text{max-day},P}$ will be useful to evaluate the pressure and velocity in the pipe on the days with the median total water demand and the maximal total water demand.
Some shorthand notation for the these quantities will be given. The median and maximum of the set of $\text{Q-max}_{d_i,P}$ for $i \in \{0, \dots, m\}$ will from now on be denoted by $\text{med}(\text{Q-max}_P)$ and $\max(\text{Q-max}_P)$ respectively. Furthermore, $\text{Q-max}_{\text{median-day},P}$ and $\text{Q-max}_{\text{max-day},P}$ will be denoted by $\text{Q-max}_{\text{med},P}$ and $\text{Q-max}_{\text{max},P}$ respectively.

On a max-day a lot water is demanded, so one would expect the peak to be larger than on average. On a median-day one would expect the peak to be close to the average. Furthermore, under the assumption that the $\text{Q-max}$ of a median-day is equal to the median of all peaks (in the to be considered timespan) and that the $\text{Q-max}$ of a max-day is equal to the maximum of all daily peaks, we have that $\text{med}(\text{Q-max}_P) = \text{Q-max}_{\text{med},P}$ and that $\max(\text{Q-max}_P) = \text{Q-max}_{\text{max},P}$. However, this assumption might be unrealistic. In this report all four quantities will be estimated and compared. This will give an indication on whether or not this assumption could hold. Note, that an in depth analysis of the correlation between $\text{Q-max}$ and the total water demand of a day is left for future research.

At the moment of this internship no data off all households supplied by a specific pipe is available to the author. What is available, is a collection of water demand time-series (real-life measurements and synthetically generated data) of different specific households that are unconnected in terms of their connection to the drinking water distribution network. It will be assumed that a random sampling of $N$ households from an available data sets provides a set of households which is representative of $H_P$ for some hypothetical pipe $P$. Therefore the following correlations will be investigated. The correlations between $\text{Q-max}_{\text{med}}$ and $N$, between $\text{Q-max}_{\text{max}}$ and $N$, between $\text{med}(\text{Q-max})$ and $N$, and between $\max(\text{Q-max})$ and $N$ will be investigated. With $\text{Q-max}_{\text{med}}$, $\text{Q-max}_{\text{max}}$, $\text{med}(\text{Q-max})$, and $\max(\text{Q-max})$ being equal to $\text{Q-max}_{\text{med},P}$, $\text{Q-max}_{\text{max},P}$, $\text{med}(\text{Q-max}_P)$, and $\max(\text{Q-max}_P)$ respectively for some hypothetical pipe $P$. Even though correlations will be determined for hypothetical pipes, under the assumptions stated above these will still be representative of real pipes. Since the interest of this report lies with a general relation between $N$ and $\text{Q-max}$ on a 'median day' and 'maximal day', the correlation between $\text{Q-max}_{\text{med}}$ and $N$, between $\text{Q-max}_{\text{max}}$ and $N$, between $\text{med}(\text{Q-max})$ and $N$, and between $\max(\text{Q-max})$ and $N$ are still valuable for the field.

# 3 Data description

This report will investigate the correlation between the 'median' and 'maximal' Q-max and number of households supplied by a pipe. Two real-life data sets were used, a combined data set from Brabant Water (BW) and Waterbedrijf Groningen (WBG) and a data set from Vitens. A description of both data sets will be given in Sections 3.1 and 3.2 respectively. Finally, in Section 3.3 an overview of SIMDEUM will be given as well as a description of how the data set was generated.

## 3.1 Data set Brabant water and Waterbedrijf Groningen

Two data sets from Brabant Water (BW) and Waterbedrijf Groningen (WBG) were available (which were already pre processed). The data was collected with a measuring device which was installed in multiple households in Brabant a nd Groningen (mostly employees of the corresponding water companies). The devices were not all installed simulta neously, however they were all installed for some time in between 02-11-2020 and 31-12-2021. In some cases the device was installed multiple times at the same household (at most 3 times). The device measures the water flow go ing into the house every second. If the flow is non-zero the value is stored. Thus, the data contains the water flow (i n l/s) into a household of every second (during which the flow was not zero) of a day. The same measurement devic es were used for both BW and WBG.

In total the water use was measured for 76 households (on average 2,95 users per household). In Figure 1, a histogram can be found with the number of measured days per household. In total 4056 days were measured. As can be seen for many of the households less than 25 days were measured. For a few households, however, more than 100 days were measured. The average of the number of measured days is 53.4 and the median is 20.5. This shows, as can also be seen in Figure 1 that the number of measured days per household is skewed. If one would like to use all data that is available, some households would be over represented whilst others only having a few measured days. This could lead to a different distribution of the household type than expected.

The data was not measured on the same dates for all households (in some cases the same measurement devices were used for multiple households). In Figure 2 the number of measured days per month are given. It can be observed the number of measured days per month is irregular. For example, in October zero days were measured and in December 1036 days were measured. Thus, also the number of measured days per month is irregular. This is of importance since it is known that water demand is amongst other things, related to the time of year.

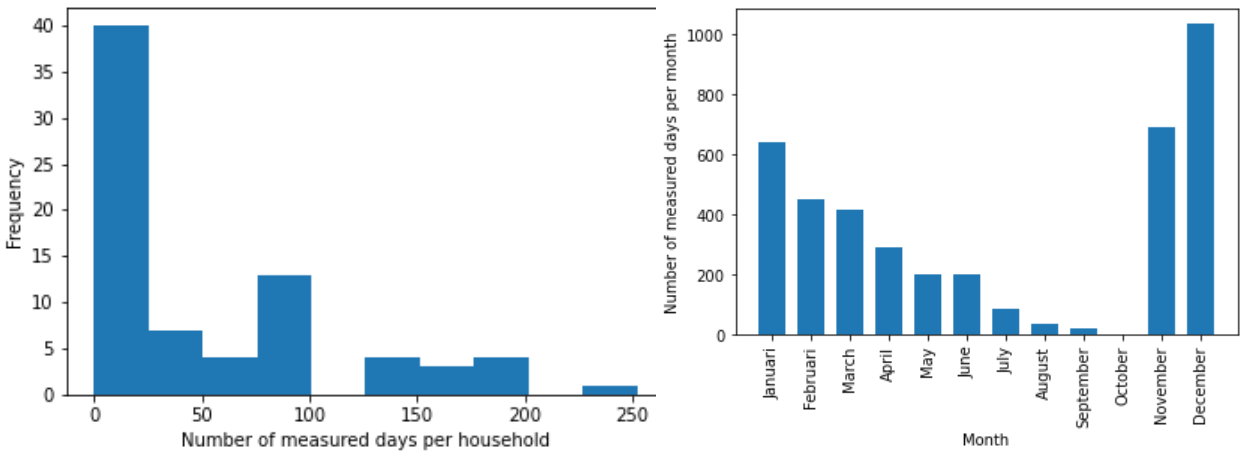

*Figure 1: Histogram of number of measured days per household*

*Figure 2: Histogram of number of measured days per month*

Out of the 4056 measured days, 2919 measured days were weekdays (274 from BW and 2645 from WBG). After removing the weekend days the distribution of the number of measured days per household and per month changed only very slightly. The histograms of the number of measured days per household and per month can be found in Figure 3 and Figure 4. It can be observed that the number of measured days per household as well as the number of measured days per month are irregular.

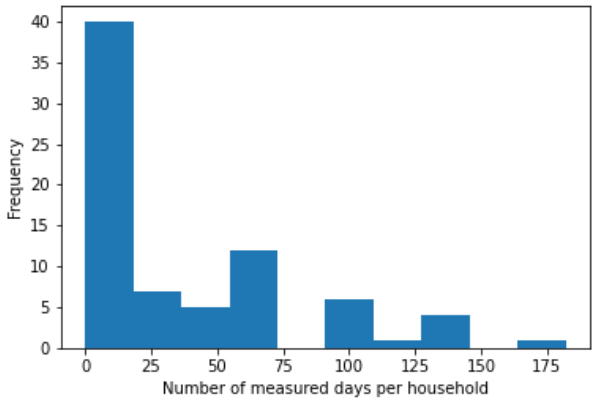


*Figure 3: Histogram of number of measured days per household*

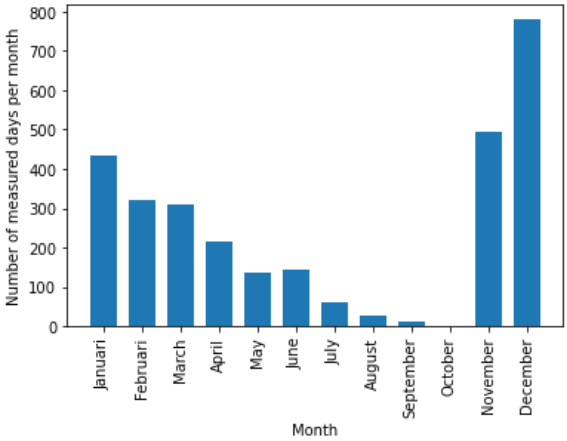


*Figure 4: Histogram of number of measured days per month*

## 3.2 Data set Vitens

The data set from Vitens consists of data collected from 1414 households in Westeinde (a neighbourhood in Leeuwarden, the Netherlands). To collect this data, approximately every hour a measuring device stores the meter reading. Given the readings and the measuring times the average flow between two consecutive measurements can be computed. On average 708 unique days were measured (minimum is 202 and maximum is 718). A histogram of the amount of days measured per household can be found in Figure 5. From this histogram it is apparent that most of the households have data on at least 600 days.

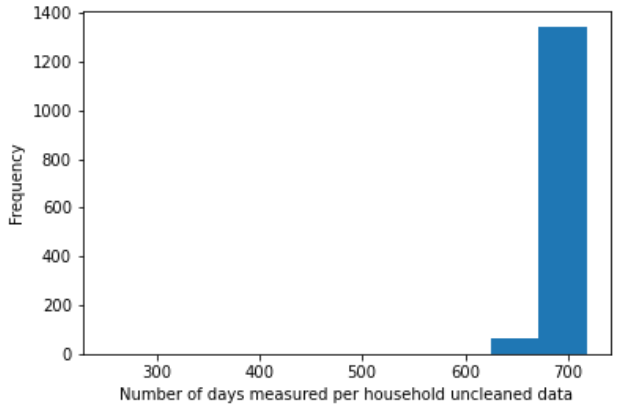


*Figure 5: Histogram of the number of measured days per household (before cleaning)*

Before using the data to analyse the relation between $N$ and the maximum flow of a day, the data needs to be cleaned. This is necessary since in some cases there are too few measurements in a day, too much time between measurements or measuring errors, which can cause imprecise results.

Firstly, if there are too few measurements of a day of one household, the water demand is averaged over quite a long time period. This could lead to a lower maximal daily flow since the water use is averaged out over more time. Therefore, it was decided that only days with at least 12 measurements would be used.

Secondly, if the time between consecutive measurements is large (the water level changed), the water use is again averaged out over a long time period. This could again lead to a lower Q-max. the time between all consecutive measurements off all households can be found in the kernel density plot in Figure 6. As can be seen, the kernel density plot has a peak between 0 and 5 and has a relatively long tail. This is due to the fact that most of the measurements were read at approximately one hour after the previous measurement. However, in some rare cases the time between two measurements is relatively long. If there is a long time between two measurements, the water use is averaged over a very long time period. This will cause the peaks of an individual household to be very low. When summing only a limited amount of households this could lead to a too low summed Q-max. Therefore, only days with at most 8 hours in between measurements will be investigated.

Furthermore, it was observed that in the data of some households a measurement error is contained, which can lead to inaccurate results. For example, in a data set the measurements as in Table 1 were contained. To remove these outliers, only days are contained in the data set for which the difference between two measurements is at most $5m^3$. This is set to a quite high number (average water demand of one person per day is $0.128m^3$ (WMD, 2022) because the peak behaviour of the water demand is of importance in the analysis of Q-max. Note, that an alternative method too filter out these measurement errors would be to remove all measurements for which it holds that the difference between the next measurement and the measurement is negative. However, due to time constraints this is left for future research.

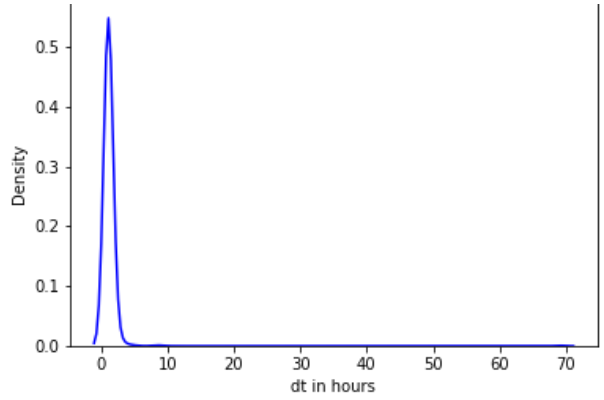


*Figure 6: Density of time between measurements*

*Table 1: Example of measurement error*

| Timestamp | Measurement ($m^3$) |
|---|---|
| 2020-08-24 08:57 | 94.686 |
| 2020-08-24 09:59 | 208.541 |
| 2020-08-24 10:58 | 94.697 |
| 2020-08-24 11:59 | 94.711 |

After cleaning the data as described above 979 622 unique measured days remained (1 000 983 before cleaning). Thus, 2.13% of the days were discarded. In Figure 7, the average number of measured days per household per month are given. Note, that all measurements were performed between 2020-04-01 and 2022-03-21, which explains why in march there are relatively fewer measured days. As can be seen, the difference between the average number of

measured days before and after cleaning is mainly visible in the month August, September and December. Investigating the reason of the uniformity of the inaccurate days (that were removed) lies outside the scope of this report. A histogram of the number of unique days per household of the data set after cleaning can be found in Figure 8. It can be observed that the peak at 700 is a bit lower and that mean number of measured days decreased a bit (from 707.9 to 692.8).

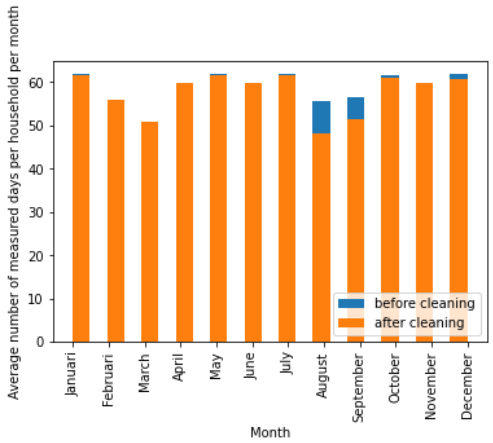


*Figure 7: Histogram of average number of measured days per household per month*

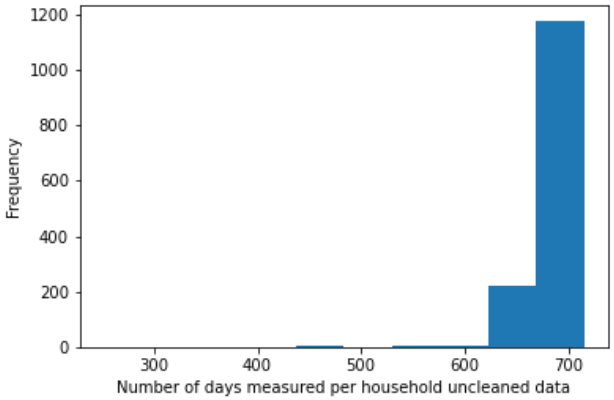


*Figure 8: Histogram of number of unique days per household after cleaning*

## 3.3 SIMDEUM data

Blokker developed a water demand model called SIMDEUM (Simulation of water Demand; an End-Use Model) (E.J.M. Blokker, 2010). This model can be used to simulate water use of a single household. First the model, will be described in Section 3.3.1. Then the assumptions underlying of SIMDEUM will be explained in Section 3.3.2, after which a short description will be given of the way the data was created for this project in Section 3.3.3.

### 3.3.1 SIMDEUM short model description

The exact details of the model SIMDEUM can be found in (E.J.M. Blokker, 2010). Also a python implementation of SIMDEUM is available as open source (Steffelbauer, Hillebrand, & Blokker, 2022), note that this is a new open source version of SIMDEUM which could still contain some mistakes. Furthermore, all statistics that have been used, can be found in this open source code (in the data file). A short overview of how SIMDEUM works will be given here.
SIMDEUM is a specific version of a rectangular pulse model. SIMDEUM simulates a water demand pattern of a household at time $T$, which is denoted by $Q(T)$. $Q(T)$ can be defined as follows:

$Q(t) = \sum_{k=1}^{M} \sum_{j=1}^{N} \sum_{i=1}^{F_{jk}} B(I_{ijk}, D_{ijk}, \tau_{ijk}, t)$ , with

$B(I_{ijk}, D_{ijk}, \tau_{ijk}, t) = \begin{cases} I_{ijk} & \text{if } \tau_{ijk} < t < \tau_{ijk} + D_{ijk} \\ 0 & \text{else} \end{cases}$.

Here $M$ is the number of end-uses, $N$ is the number of users and $F_{jk}$ is the number of uses per end-use per user. Furthermore, $I_{ijk}$ is the pulse intensity (in L/s), $D_{ijk}$ is the duration of the pulse (in seconds) and $\tau_{ijk}$ is the time at which the pulse starts.

The simulation starts by defining the objects house and end-use. The object house contains the house type (one-person, two-person or family), the number of users ($N$) and information about the users (age category, out-of-home job and gender). On average SIMDEUM assumes 2.29 users per household. The statistics can be found in the Appendix in Table 8. The end-use contains information about the amount of appliances ($M$), the kind of appliances, the distribution of the duration and intensity of the appliance. A house can have up to eight kinds of end-uses: bathroom tap, bathtub, dishwasher, kitchen tap, outside tap, shower, washing machine, and wc. Both the

object house as well as the object end-use are randomly created using statistics found in (Blokker, 2006). These statistics are based on CBS data and TNS-NIPO data and are thus representative of the Netherlands. The exact statistics can be found in the data file of the pySIMDEUM code (Steffelbauer, Hillebrand, & Blokker, 2022).

After the house and the end-use are created the values of $F_{jk}, I_{ijk}, D_{ijk}$ and $\tau_{ijk}$ are determined. As an example, to determine the frequency of the end-use bathtub of a child a random sample is taken from $F_{child,bathtub} \sim \text{Poisson}(0.085714)$. The distribution of the frequency per end-use type can be found in Table 2. In a similar way are $I_{ijk}, D_{ijk}$ determined, the exact used distributions of $F_{jk}, I_{ijk}$, and $D_{ijk}$ can also be found in the data file in (Steffelbauer, Hillebrand, & Blokker, 2022). These statistics were determined by Blokker (2006).

After determining the duration, intensity and frequency of all users and end-uses, the starting times $\tau_{ijk}$ need to be determined. For this the state of the user $j$ (at work, at home, asleep or awake) needs to be determined first. Therefore, the presence will be determined for all users (containing 4 times: time of getting up, time of going to work, duration of being away, duration of being asleep). These times, will depend on whether or not the date is a weekday or weekend-day, (however, note that this is not yet implemented in the pySIMDEUM version of 26-10-2022). Thus, the data created with this version of SIMDEUM only contains simulated weekdays. The diurnal pattern (thus, up, go, away, sleep times) are determined by normal distributions with a mean and standard deviation that differ per user category. The parameters of the normal distribution per user category can be found in the data file in (Steffelbauer, Hillebrand, & Blokker, 2022) as well as in chapter 3 of (E.J.M. Blokker, 2010).

*Table 2: Distribution of frequency per end-use*

| End-use | Distribution of frequency | User kind (if applicable) | $N$ (if applicable) |
|---|---|---|---|
| Bathroom tap | $\text{Poisson}(4.1)$ | | |
| Bathtub | $\text{Poisson}(0.085714)$ | Child | |
| | $\text{Poisson}(0.014286)$ | Teen | |
| | $\text{Poisson}(0.028571)$ | Working adult | |
| | $\text{Poisson}(0.028571)$ | Home adult | |
| | $\text{Poisson}(0.028571)$ | Senior | |
| Dishwasher | $\text{Poisson}(0.2143)$ | | 1 |
| | $\text{Poisson}(0.2286)$ | | 2 |
| | $\text{Poisson}(0.1857)$ | | 3 |
| | $\text{Poisson}(0.2000)$ | | 4 |
| | $\text{Poisson}(0.1429)$ | | 5 |
| Kitchen tap | $\text{NegBin}(\mu = 10.1, \sigma = 7)$ | | 1 |
| | $\text{NegBin}(\mu = 12.7, \sigma = 7.2)$ | | 2 |
| | $\text{NegBin}(\mu = 12.8, \sigma = 7.7)$ | | 3 |
| | $\text{NegBin}(\mu = 13.1, \sigma = 8.4)$ | | 4 |
| | $\text{NegBin}(\mu = 13.5, \sigma = 9.1)$ | | 5 |
| Outside tap | $\text{Poisson}(0.44)$ | | |
| Shower | $\text{Binomial}(0.48)$ | Child | |
| | $\text{Binomial}(0.67)$ | Teen | |
| | $\text{Binomial}(0.79)$ | Working adult | |
| | $\text{Binomial}(0.79)$ | Home adult | |
| | $\text{Binomial}(0.69)$ | Senior | |
| Washing machine | $\text{Poisson}(0.32)$ | | 1 |
| | $\text{Poisson}(0.29)$ | | 2 |
| | $\text{Poisson}(0.29)$ | | 3 |

| | | | |
|---|---|---|---|
| | Poisson(0.27) | | 4 |
| | Poisson(0.29) | | 5 |
| Wc | male: Poisson(3.8)<br>female: Poisson(5.4) | Child | |
| | male: Poisson(4.1)<br>female: Poisson(5.1) | Teen | |
| | male: Poisson(5.3)<br>female: Poisson(6.8) | Working adult | |
| | male: Poisson(7)<br>female: Poisson(7) | Home adult | |
| | male: Poisson(7.4)<br>female: Poisson(6.8) | Senior | |

After $F_{jk}, I_{ijk}, D_{ijk}$ and the presence of the users are determined, the starting times of every time an end-use is used $\tau_{ijk}$ will be determined. Two vectors will be created, which represent how likely it is that an action (of a specific end-use and user) is started during a every second. Firstly, a vector is created with the probabilities that a specific end-use is activated during every second based on when it is more likely to use that end-use (for example, it is more likely to put on the dishwasher after eating). The exact values of these probabilities can be found in the pySIMDEUM core (for every end-use separately). Secondly, a probability is determined which represents the probability that a specific end-use is activated based on the presence of the user. To implement the idea of peak hours, every second of the day is categorized into one of the following categories: peak, normal, away and night. Based on the randomly determined presence of a user (up, go, home, sleep), the categories are defined as in Table 3. Based on this category a probability is determined for every second which represents the probability that an end-use is started during that second. This probability is equal to $\frac{0.65}{\text{number of seconds in category peak}}$ for all seconds in the peak category, $\frac{0.65}{\text{number of seconds in category normal}}$ for all seconds in the normal category, $0$ for all seconds in the away category, and $\frac{0.015}{\text{number of seconds in category night}}$ for all seconds in the night category[1]. Both vectors with probabilities are combined and normalized. Based on the combined probabilities a starting time is randomly selected for every end-use.

*Table 3: Categorise time intervals into peak, normal, away and night*

| Start time of interval | End time of interval | Category |
|---|---|---|
| Sleep | Up | Night |
| Up | Up+30 minutes | Peak |
| Up+30 minutes | Go-30 minutes | Normal |
| Go -30 minutes | Go | Peak |
| Go | Home | Away |
| Home | Home+30 minutes | Peak |
| Home+30 minutes | Sleep-30 minutes | Normal |
| Sleep -30 minutes | Sleep | Peak |

After the starting time is selected, the intensity of the end-use is added to the total water use for all times in the interval [starting time, the starting time + duration].

[1] Note that in the current version of pySIMDEUM (October 2022) there is a typo which states that the probability of an end-use starting during a second at night is set to $0.15/(\text{number of seconds in category night})$ . Later the probabilities are normalized, so no problem occurs with respect to the probabilities summing up to one. However, this causes more end-uses during the night than what would be expected (1.5% of the total water use is used during night (E.J.M. Blokker, 2010)).

### 3.3.2 Assumptions of SIMDEUM

SIMDEUM creates simulated data which, under some assumptions, represents the water demand pattern of a single household. The assumptions will be given here.

Firstly, it is assumed that the water demand behaves as a block model, meaning that the intensity of the water demand of one frequency of an end-use is constant between the starting and end times. In real life, however, it could occur that the intensity is not constant from start to end.

Secondly, it is assumed that every household can be categorized as one of the three defined house types and has no other appliances (or subtypes) than described in Section 3.3.1. Furthermore, it is assumed that the distributions used to randomly select the house type, the users (including age category and gender) and number of appliances, correspond to real-life. In real-life, more possible household types would be possible as well as more kinds of water demanding appliances. Furthermore, in real-life it is possible for a household to consist of more than 5 people (e.g. large families or student houses).

Thirdly, it is assumed that the frequency , intensity and duration of each end-use type in real-life follow the same distributions as used by SIMDEUM. Note, that in real-life this could differ per household and per appliance.

Fourthly, it is assumed that the starting times of all end-uses are distributed as in real-life. Note, that this includes , for example, the assumption that the peak hours are as described in Table 3 and that the diurnal pattern follows a normal distribution with different parameters as described in (E.J.M. Blokker, 2010) and implemented in SIMDEUM. Furthermore, the assumption also includes the assumption that the starting times in real-life follow the same probabilities as used by SIMDEUM.

Finally, it is assumed that all users and end-uses are independent. A first example would be that the distribution of the presence of a user is independent of the, for example, age or presence of the possibly other users of the house. Secondly, the starting times of every time an appliance is used are independent of each other.

Under these assumptions SIMDEUM creates a water demand pattern of a random household in the Netherlands. Even though some of the assumptions might seem slightly unrealistic, SIMDEUM has been validated in (E.J.M. Blokker, 2010). To validate SIMDEUM several characteristic parameters were determined from the water demand patterns. One of these characteristic parameters is the maximal flow of a household of the day, as well as the total volume of the day.

### 3.3.3 SIMDEUM Data description

SIMDEUM can be used to create a synthetic data set containing a simulated water use during the past second. In total 10650 data sets were created with the default statistics of SIMDEUM (which are based on (Blokker, 2006)). Changing these statistics to more recent estimates is left to future research. Every data set consists of 100 simulated water demand patterns of one day. All non-zero simulated flows are stored and used to estimate the maximal flow of a summed water demand pattern of $N$ households. The methods used to estimate this quantity are described in the next chapter.

# 4 Methods

To investigate the correlation between the number of households supplied by the pipe and the maximum flow during a day in that pipe, the summed flow has to be determined. One could determine the Q-max from data, for example, the data sets described in Section 0:

Data description. However, no timeseries data with the flow in a pipe is available to the author. Nonetheless, timeseries data with the water use of single households is available. This could be used to estimate Q-max. One could also simulate the water-use from a household with SIMDEUM. This could then also be used to estimate Q-max.

When a set of data is provided, either synthetic or non-synthetic, two methods will be given that can estimate Q-max on 'average' and 'maximally'. The two variations as described in Section 0:

Formal problem formulation. The first method estimates $\mathrm{med}(\text{Q-max}_P)$ and $\max(\text{Q-max}_P)$ for a given value of $N$. The second method estimates $\text{Q-max}_{\text{med},P}$ and $\text{Q-max}_{\text{max},P}$ for a given value of $N$. Note, that the, by method 1 and method 2, obtained values estimate different quantities. However, both quantities (the median of the Q-maxes of random days and the Q-max on a median-day or max-day) provide insight into the maximal flow of Q-max on 'average' and 'maximally'.

## 4.1 Data sampling method 1

To estimate $\mathrm{med}(\text{Q-max})$ and $\max(\text{Q-max})$ data of different households is combined. This data sampling method estimates the median and maximum of a set of $x$ Q-max of hypothetical streets on hypothetical days. This method is mainly useful for data sets with a limited amount of measurements per household (with possibly different measurement dates). On the other hand, by creating hypothetical households and hypothetical days, possibly present street and day specific information could be lost. The method to determine the maximal flow of the summed water use of random days of $N$ households can be described by the following outline:

1. Select $N$ random households.
2. Select a random measurement date of every selected household.
3. Add the demand patterns of the $N$ randomly selected days to determine the summed demand pattern.
4. Determine the maximum of the summed demand pattern Q-max.
5. Repeat the above $x$ times and take the median (and maximum) of the found Q-max.

For this method to converge to the desired quantities ($\mathrm{med}(\text{Q-max})$ and $\max(\text{Q-max})$) some assumptions have to be made. Firstly, it is assumed that repeatedly selecting $N$ random households from the data set represents a randomly selected pipe with $N$ households accurately. Secondly it is assumed that the water demand pattern of any weekday is identically distributed. Note that only the data of the weekdays is combined, since according to, for example, Alvisi et. all. (Alvisi, Franchini, & Marinelli, 2007) the hourly water demand depends on the diurnal pattern which differs for week- or weekend days. Thirdly, it is assumed that the summed demand pattern of the randomly selected households follows the same distribution as the demand pattern of a hypothetical pipe. Under these assumptions the quantity found from method 1 computes $\max_t \sum_{i=1}^{N} X_{d_i,h_i}(t)$. Where $h_i$ are the $N$ randomly selected households, $d_i$ is the corresponding random selected day of household $h_i$ and $X_{d_i,h_i}(t)$ is the water demand pattern of household $h_i$ on day $d_i$.

Three different versions of this method were implemented and compared. The reason behind the alternative versions, is the amount of reuse of the data. Reusing data can lead to an added correlation between the different values of Q-max.

*Algorithm 1: Method 1.1 for different values of N*

Input: ndays sum over number of unique days of every household, dataframe with Date, FlowPerSecond, id, Time.

1. Define narray to contain all possible values of $N$ and sort in a descending order.
2. Let nreps$\leq \left\lfloor \frac{\text{ndays}}{\sum \text{narray}} \right\rfloor \in \mathbb{N}$ .
3. Define Q = zero matrix with size len(narray) by nreps.
4. For m in 0,…,len(narray)-1 do:
5. n=narray[m]
6. For i in 0,…,nreps-1 do:
7. Let ids= all id's data still have at least one unused day.
8. Select n random id's (households) from the ids array (without repetition).
9. Create an empty dataframe tmp.
10. For j in 0,…,n-1 do:

11. Select a random unused day of the household with the j'th randomly selected id.
12. Add this day to a dataframe tmp and remove it from the dataframe df.
13. Sum FlowPerSecond grouped by Time.
14. Let Q[m,i]= maximum over all times of the summed FlowPerSecond.

Version method 1.1 does not reuse any of the data, by removing the data of the randomly picked day and household from the data frame. In Algorithm 1 a pseudocode can be found of method 1.1. Note, that this method introduces a correlation between the order in which $\mathrm{Q\text{-}max}$ for the different values of $N$ is computed and the households used (this is especially the case if a large part of the available data is used). For example, for the final value of $N$ for which a $\mathrm{Q\text{-}max}$ is computed (=min(narray) in the pseudocode) it is very likely that households are selected with many days, since they are most likely the once that still have unused days. Note, that if the number of measured days per household is quite uniform (meaning all households have approximately the same amount of measured days), this is of less importance.

Version 1.2 does reuse the data for different values of $N$. In other words, for every value of $N$ the data is used to determine the maximal summed flow of the randomly selected days and households without reusing any data. However, for another value of $N$ the same data may be used. This has the advantage that more data is available for every value of $N$ which leads to a better estimation of the summed Q-max. However, reusing the data could change the correlation between $N$ and the corresponding estimate of $\mathrm{med}(\mathrm{Q\text{-}max}_P)$ and $\mathrm{max}(\mathrm{Q\text{-}max}_P)$. A pseudocode of method 1.2 can be found in the Appendix in Algorithm 3.

Version 1.3 reuses the data in another way. This version selects a random day of any household instead of a random day of a random household. Note, that the amount of measured days is not uniform among the households. This will lead to the households with many measured days being picked more often than households with fewer measured days. Thus, step 1 and 2 as described in the outline are replaced by randomly selecting a day of any of the households. Furthermore, the day selected is not removed from the data frame (and reusing this same data for another repetition could occur). Note, that depending on the number of repetitions and the size of all possible values of $N$ the data could be reused extensively. This can cause correlations within the results. A pseudocode of version 1.3 of method 1 can be found in the Appendix in Algorithm 4.

## 4.2 Data sampling method 2

Method 2 estimates another quantity, namely $\mathrm{Q\text{-}max}_{\mathrm{med,P}}$ and $\mathrm{Q\text{-}max}_{\mathrm{max},P}$ for different values of $N$. Also in this case, data of the same date of different households is not necessarily needed. However, it is preferred that only data from the same days are combined since many more variables could influence the water demand pattern of a household, for example, the temperature. Method 2 combines the matching dates of $N$ random households and determines the median-day and max-day based on their total daily water demand. Thus, with the resulting estimates describe the maximal flow of the day on which the most water was used (of all days in on which measurements were performed) and of the day on which the median amount of water was used. Method 2 can be described by the following outline:

1. Select $N$ random households.
2. Determine the median-day and max-day of the summed demand pattern of the $N$ selected households.
3. Add the flow during the median-day and add the flow during the max-day of all selected households to determine the summed flow on a median-day and max-day.
4. Determine the maximum of the summed flow of both the median- and max-day.

This method again assumes that the summed water demand of a collection of $N$ random households is representative of the summed water demand of a waterpipe. Under this assumption the quantity found from method 2 computes $\max_t(\sum_{i=1}^{N} X_{d_{\text{med}},h_i}(t))$ and $\max_t\left(\sum_{i=1}^{N} X_{d_{\max},h_i}(t)\right)$. Where $h_i$ are the $N$ randomly selected households, $d_{\text{median}}$ and $d_{\max}$ are the found median- and max-days of days that were measured and on which all random selected households have at least 1 measurement, and $X_{d_{\text{median}},h_i}(t)$ and $X_{d_{\max},h_i}(t)$ are the collection of measurements of household $h_i$ on the found median-day and max-day respectively.

A pseudocode of this method can be found in Algorithm 2.

*Algorithm 2: Method 2*

Input: List with cleaned dataframe of every household (in the dataframe all measured values and corresponding timestamps are contained.

*Note, a cleaned dataframe is used, where for every day of every household the maximum time between measurements is computed. If this time is more than 7200 seconds (2 hours) the day is discarded.*

1. For n in 1,…,20 do:
2. For i in 1,…,nreps do:
3. Let files_selection = random selection of households from the list.
4. Combine these households in one dataframe.
5. Let dates= all dates such that every household in files_selection has measurements on that day.
6. Let dataframe= dataframe where the date of the timestamp is in dates and drop the id.
7. Let df_day = summed Value in dataframe by date and select only the days that are in dates.
8. Let df_hour = summed Value in dataframe by hour and select only the days that are in dates.
9. Let median= the median of df_day and let median_day = corresponding day as defined in equation ( 1 ).
10. Let Qn= maximum of Value in df_hour where the date is equal to median_day.
11. Let max_day = date corresponding to the maximum of df_day as defined in equation ( 2 ).
12. Store n, Qn and Qn_max.

## 4.3 Fitting method

To fit possible curves through the results of method 1 and method 2, the scipy.optimatize package in python is used.

The curvefit method applies a non-linear least square to fit a function of choice to the results. The non-linear least square approach minimizes the following function to determine the parameters of the chosen function:

$$\sum_{i=0}^{m}(y_i - f(x_i, \boldsymbol{\beta}))^2$$

where $m$ is the number of data points, $y_i$ is the value of the i'th data point, $x_i$ is the corresponding value of $N$ and $\boldsymbol{\beta}$ is a vector with the parameters of the function.

Multiple functions were fit to the found estimates from methods 1 and 2. Firstly, the function $f(x) = a\sqrt{x} + bx$ was fit. Currently WMD uses the 'q-square-root-N' rule for $N < 200$ houses and a linear function for $N > 200$. In the limit this function would behave linearly. Furthermore, the function

$f(x) = \begin{cases} a\sqrt{x} & \text{for x>T} \\ a\sqrt{T} + (x - T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$ was fit. This function is equal to the 'q-square-root-N' rule for $N <$ T houses and a linear function for $N > T$. However, the parametrization of the 'q-square-root-N' rule and the linear part of the function are different. Another function that was fit is $f(x) = a\log(x + 1) + bx$. Note that $\log(0 + 1) = 0$ and $f(0) = 0$. Since the Q-max of zero households is zero, this is a characteristic attribute of the correlation. Furthermore, in the tail a logarithmic function grows very slowly (slower than a square root). Therefore, also this function behaves as a linear function for large values of $N$. Furthermore, it was fit to see if an alternative to the square root might fit the estimates as well. Finally, a the following linear function was fit $f(x) = ax$, since for larger values of $N$ it is expected that the maximal daily flow behaves linearly with respect to $N$.

## 4.4 Implementation of methods

The data sets as described in Section 0:

Data description are not all suitable for both methods as described above. The BW+WBG data set was analysed with method 1. Since this data set only has a very limited amount of data measured on the same day, method 2 was not applied. The Vitens data set had more data available (which were also measured on the same dates), and thus this data was analysed with method 2. The SIMDEUM data was analysed with both method 1 and method 2. An overview of which data set was analysed with which method, for which values of $N$ and for how many repetitions can be seen in Table 4.

To estimate $\mathrm{med(Q\text{-}max)}$ from the BW+WBG data set not enough data is available to only combine the data that was measured on the same date. Therefore random days will be combined to approximate a hypothetical day. One could argue that combining data from different dates and possibly the same household could result into unrepresentative results. Therefore, only weekdays were selected. This gives the assumption that the water demand patterns of different weekdays are identically distributed.

The $\mathrm{max(Q\text{-}max)}$ can be determined from the BW+WBG data set. However, this is not done extensively. This is due to the fact that only a limited amount of data was available. Therefore, the probability of measuring , for example, the yearly maximum is small since for more than half of the household less than 21 days were measured. This could make the estimation of $\mathrm{max(Q\text{-}max)}$ less accurate. Furthermore, some very high values are still contained in the data set. Expectedly, these would dominate the estimate of $\mathrm{max(Q\text{-}max)}$. Filtering out these values by distinguishing them from high demands is left for future research. Note, that they are estimated briefly to illustrate that for small values of $N$ they are close to the results of the SIMDEUM data set.

To illustrate the effect of a lower measuring frequency, the BW+WBG data set was also aggregated to hourly data. This then was analyzed with method 1 (version 1.3).
The Vitens data set was analyzed with method 2.

To the SIMDEUM data set both method 1 and 2 were applied. In the case of method 1, a variation of version 1 was applied. Namely, $\mathrm{N}$ households were selected after which the maximal summed flow of all (100) simulated days were determined. This was repeated for 5 different selections of households and the median and maximum of the summed flow of every selection of households was determined. This was executed for $\mathrm{N} \in \{0,1,11, \dots ,201\}$. Method 2 was applied in a similar way, however the median-day and max-day of the summed water demand of the $\mathrm{N}$ households were determined. Then the maximum summed flow was stored. This again was repeated for 5 different selections of households and the median was determined for each selection. Finally, this was repeated for the aggregated simulated data (to hourly data).

Table 4: Overview methods and data sets

| Data set | Method | $N$ | Number of repetitions | Additional information |
|---|---|---|---|---|
| BW+WBG | Method 1 version 1.1 | $N \in \{0,2, \dots ,20\}$<br>$N \in \{0,2, \dots ,20\}$ | 26<br>20 times 26 | Estimate $\mathrm{med(Q\text{-}max)}$ and not $\mathrm{max(Q\text{-}max)}$ |

| | | | | |
|---|---|---|---|---|
| | Method 1 version 1.2 | $N \in \{0,2,\ldots,20\}$<br>$N \in \{0,2,\ldots,20\}$ | 100<br>20 times 100 | Estimate $\mathrm{med}(\text{Q-max})$ and not $\max(\text{Q-max})$ |
| | Method 1 version 1.3 | $N \in \{0,2,\ldots,20\}$<br>$N \in \{0,1,\ldots,200\}$<br>$N \in \{0,5,\ldots,2000\}$ | 100<br>100<br>100 | Estimate $\mathrm{med}(\text{Q-max})$ and not $\max(\text{Q-max})$ |
| BW+WBG aggregated | Method 1 version 1.3 | $N \in \{0,1,\ldots,20\}$ | 100 | Estimate $\mathrm{med}(\text{Q-max})$ and not $\max(\text{Q-max})$ |
| Vitens | Method 2 | $N \in \{0,2,\ldots,20\}$<br>$N \in \{0,5,\ldots,300\}$ | 10 | |
| SIMDEUM | Method 1 | $N \in \{0,1,\ldots,50\}$<br>$N \in \{0,1,11,\ldots,201\}$ | 5<br>5 | For a random selection of households 100 $\text{Q-max}$ are determined, median and maximum of these $\text{Q-max}$ are stored (this is repeated 5 times) |
| SIMDEUM | Method 2 | $N \in \{0,1,\ldots,50\}$<br>$N \in \{0,1,11,\ldots,201\}$ | 5<br>5 | |
| SIMDEUM aggregated | Method 1 | $N \in \{0,1,\ldots,50\}$<br>$N \in \{0,1,11,\ldots,201\}$ | 5<br>5 | For a random selection of households 100 $\text{Q-max}$ are determined, median and maximum of these $\text{Q-max}$ are stored (this is repeated 5 times) |
| SIMDEUM aggregated | Method 2 | $N \in \{0,1,\ldots,50\}$<br>$N \in \{0,1,11,\ldots,201\}$ | 5<br>5 | |

# 5 Results

The methods as described in Chapter 0 are used to estimate $\text{Q-max}_{\text{med}}$, $\text{Q-max}_{\text{max}}$, $\max(\text{Q-max})$ and $\text{med}(\text{Q-max})$ from the data as described in Chapter 0. In Section 5.1 the results from the BW+WBG dataset are described. Then in Section 5.2 the results from the Vitens dataset are described. In Section 5.3 the results of the synthetic data generated by SIMDEUM are given. Finally, in Section 5.4 all results will be compared.

## 5.1 Brabant water and Waterbedrijf Groningen

The combined and cleaned data set of BW and WBG have been analyzed with method 1 which is described in Section 4.1. Version 1.1 was executed for $N \in \{0,2,\dots,20\}$ with $\left\lfloor \frac{\text{ndays}}{\sum \text{narray}} \right\rfloor$ number of runs, which implies that almost all data was used. Version 1.2 was executed with 100 runs and for $N \in \{0,2,\dots,20\}$. Version 1.3 was executed with 100 runs for $N \in \{0,2,\dots,20\}$, for $N \in \{0,1,\dots,200\}$ and for $N \in \{0,5,\dots,2000\}$.
Note, that for both version 1.1 and version 1.2 it holds that the order in which the households are picked influences the households that will be chosen in future. To give some insight into the effect of this dependency, version 1.1 and version 1.2 were ran 20 times. The results can be observed in Figure 9. As can be seen, the results of version 1.2 (reusing data for different values of $N$) are slightly larger than the results of version 1.1 (no reusing of the data). This can be caused by the fact that the high water flows (that are contained in the data set), that are either outliers or high water demands are reused in version 1.2 with high probability. Furthermore, it can be observed that the variance of the estimates seems to be larger for larger values of $N$. This could be explained by the fact that the probability that any measured day is used to estimate the $\text{Q-max}$ of a large value of $N$ is large. Thus, also the probability that a high value is used to estimate the $\text{Q-max}$ of a large value of $N$ is larger than the probability that is used to estimate $\text{Q-max}$ of a small value of $N$. This problem could be resolved by either using more data, to ensure that the estimates of $\text{Q-max}$ are more accurate. Furthermore, distinguish outliers and high demands more accurately is left for future research.

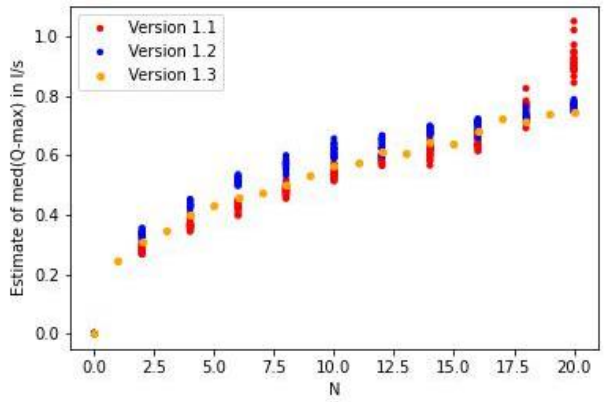


Figure 9: Results of version 1.1 and 1.2 (20 estimates)

The median of the 100 resulting values of all three methods (as described in Section 0) can be found in Figure 10. It can be seen that all found estimates of $\text{med}(\text{Q-max})$ for any value of $N > 0$ are smaller than the 'q-square-root-N' rule estimates. This is an indication that the 'q-square-root-N' rule overestimates the maximal summed flow on at least 50% of the days. Furthermore, all three versions of data sampling method 1 are very close. This is an indication, that for $N \leq 20$ the reuse of the data has no significant effect. As denoted in Section 3.1, the time between two consecutive measurements is 1 second. To demonstrate the effect of a larger time increment the data was aggregated to hourly data. The result can also be seen in Figure 9. As can be seen, the resulting estimates are much

smaller than the estimates resulting from the data set with time increments of 1 second. This can be explained by the fact that the total amount of water used in an hour is rarely the result of a constant very slow flow, but rather a few high peaks. If the time between consecutive measurements is thus equal to an hour, the peaks will be much lower, due to the averaging of the water demand.

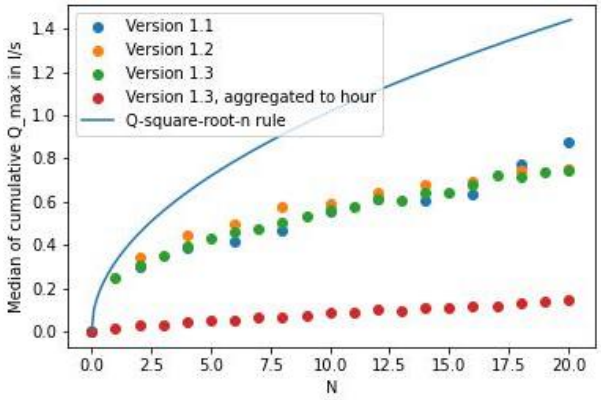


*Figure 10: Results of version 1 with BW and WBG data set*

In Figure 11 the estimate of method 1.3 can be found for $N \in \{0,1,\dots,200\}$ and 100 repetitions per value of $N$. It can be seen that for larger values of $N$ approximately greater than 50, the increase seems to be linear. Currently, the WMD assumes the maximum daily flow to be linear with $N$ starting from $N = 200$. However, from these results it seems that this starts earlier.

In Figure 12 the results of version 1.3 can be found for $N \in \{0,5,10,\dots,2000\}$. As can be seen, the results start to deviate from what would be expected. For values of $N > 250$ it can be observed that the variance seems to increase. Furthermore, for $N > 750$ steps start to form. This can be explained by the fact that for every repetition $N$ out of 2934 measured days are selected. If $N$ start to increase and the number of repetitions per value of $N$ is still equal to 100, a lot of the data is reused (including the high measurements). Since the estimate of $\mathrm{med}(\text{Q-max})$ is equal to the median of 100 $\text{Q-max}$, if such a high value is contained in only a few of the $\text{Q-max}$ this will not greatly impact the estimate. However, for larger values of $N$ it is likely that the high values are used to estimate many $\text{Q-max}$. This will lead the high values to greatly impact the estimate. Note, the interest of this report was to analyse the corelation between $\text{Q-max}$ and $N$ mainly for $N < 200$.

To quantify the correlation between $N$ and $\mathrm{med}(\text{Q-max})$ multiple functions were fit to the found results from version 1.3 (for $N \in \{0,2,..,250\}$). As described in Section 4.3, a non-linear least squares approach was applied. After finding the optimal values of $\boldsymbol{\beta}$, the $r^2$ was determined. Some of the functions that were fit with the corresponding values of the parameters $\boldsymbol{\beta}$ and the found $r^2$ and $\text{MSE}$ can be found in Table 5. Furthermore, the results with the found functions can also be observed in Figure 13. As can be seen, all three function are very similar and seem to fit the results from the data. In Figure 14 the same results can be observed, however zoomed in for $N < 60$. Here it can be observed that the function $f(N) = a\sqrt{N} + b \cdot N$ underestimates the values of $\mathrm{med}(\text{Q-max})$.

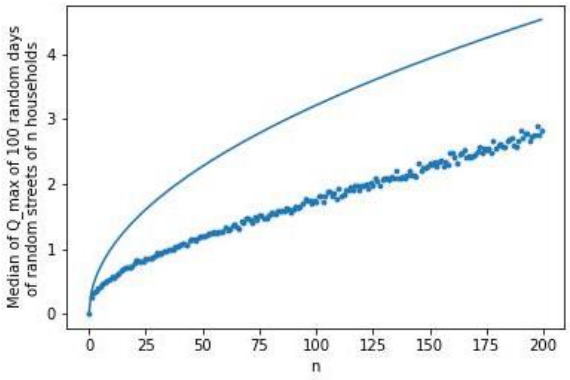


*Figure 11: Results of Version 1.3 for $N \in \{0,1,\dots,200\}$ with BW and WBG data set*

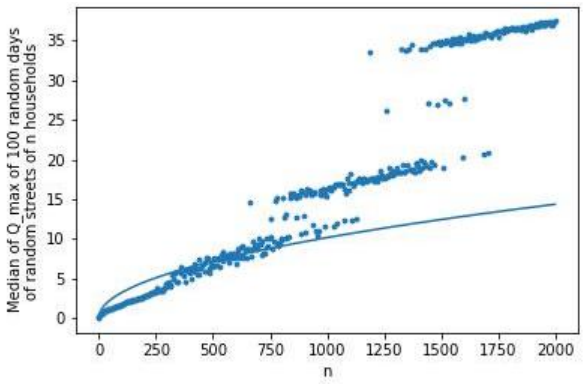


*Figure 12: Results of Version 1.3 for $N \in \{0,5,\dots,2000\}$ with BW and WBG data set*

*Table 5: Results from fitting method 1.3*

| Function | $\boldsymbol{\beta}$ | $r^2$ | MSE |
|---|---|---|---|
| $f(N) = a\log(N+1) + b \cdot N$ | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.179 \\ 0.009 \end{pmatrix}$ | 0.9935 | 0.0045 |
| $f(N) = a\sqrt{N} + b \cdot N$ | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.127 \\ 0.005 \end{pmatrix}$ | 0.9904 | 0.0066 |
| $f(N) = \begin{cases} a\sqrt{N} & \text{for N>T} \\ a\sqrt{T} + (N-T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$ | $\begin{pmatrix} a \\ T \end{pmatrix} = \begin{pmatrix} 0.169 \\ 60.9 \end{pmatrix}$ | 0.9943 | 0.0039 |

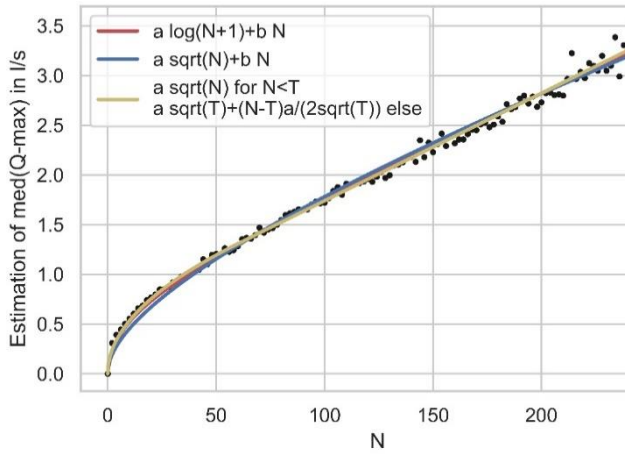


*Figure 13: Fitted function to results of version 1.3 for $N \in \{0,2,\dots,250\}$*

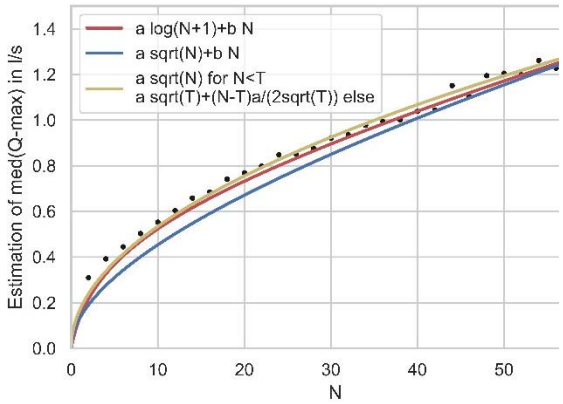


*Figure 14: Fitted function to results of version 1.3 for $N \in \{0,2,\dots,60\}$*

Overall, the correlation between $\text{med(Q-max)}$ and $N$ can be described by multiple functions. Since WMD currently assumes the 'q-square-root-N' rule for $N < 200$ houses and a linear function for $N > 200$ and the function $f(N) = \begin{cases} a\sqrt{N} & \text{for N>T} \\ a\sqrt{T} + (N-T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$ fits the estimates, the rule that will be recommended based on the results from the BW+WBG data set is: for $N < 61$ take $0.17\sqrt{N}$ and for $N \geq 61$ take $0.01 \cdot N + 0.66$.

## 5.2 Vitens

The cleaned data set of Vitens is analyzed with method 2 which is described in Section 4.2. This Method was implemented for $N \in \{0,2,\dots,20\}$ and for $N \in \{0,5,\dots,300\}$ with 10 repetitions. The median of the 10 found estimations of $\text{Q-max}_{\text{med}}$ and $\text{Q-max}_{\text{max}}$.

The results for $N \in \{0,2,\dots,20\}$ can be found in Figure 15 and for $N \in \{0,5,\dots,300\}$ in Figure 16. It can be observed that the correlation between the estimated quantities and $N$ is fairly linear even for $N < 61$. This can be explained by the fact that if hourly data is used which causes all water demand to be averaged over the past hour (in some cases over two hours if the water is used at the time of measuring time).

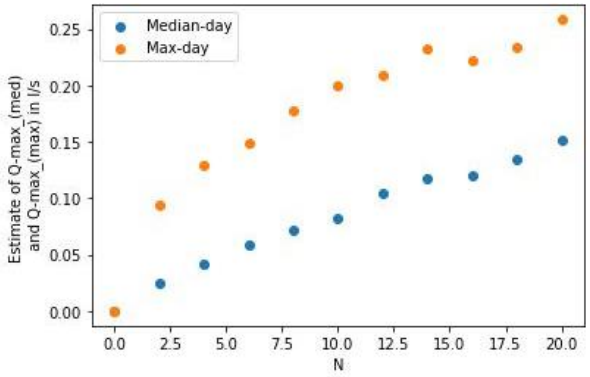


*Figure 15: Results of Method 2 with Vitens data set where $N \in \{0,2,\dots,20\}$*

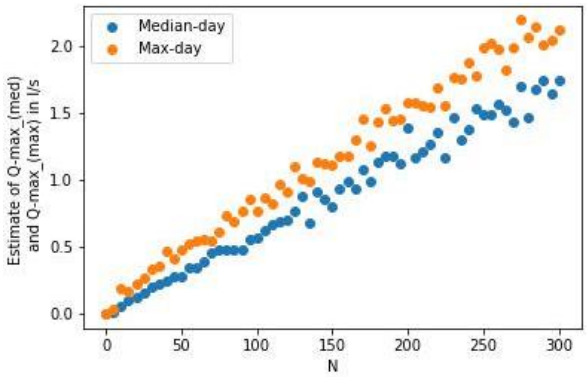


*Figure 16: Results of Method 2 with Vitens data set where $N \in \{0,5,\dots,300\}$*

In Table 6 the results of fitting three different functions to the found estimates of $\text{Q-max}_{\text{max}}$ and $\text{Q-max}_{\text{med}}$ can be found. The found functions can be found in Figure 17 and Figure 18 for $\text{Q-max}_{\text{max}}$ and $\text{Q-max}_{\text{med}}$ respectively. It can be observed that when fitting a linear function, the $r^2$ is close to 1. This implies that a linear functions fits the found estimates well. However, the $r^2$ is a bit closer to 1 for both alternative functions as stated in Table 6. Note, that the found values of $a$ of the function $f(N) = a\sqrt{N} + b \cdot N$ in case of the Vitens data set is much smaller than in the case of the BW+WBG data sets. Furthermore, the values of $b$ are very close. This implies that for smaller values of $N$ the results from the Vitens data set are more linear (the estimates of $\text{Q-max}_{\text{med}}$ even more than $\text{Q-max}_{\text{max}}$). Also, in the case of the fitted function being equal to $f(N) = \begin{cases} a\sqrt{N} & \text{for N>T} \\ a\sqrt{T} + (N-T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$, it can be observed that the value of $T$ (where it holds that for all $N > T$ the function is linear) is much smaller than in the case of the BW+WBG data set.

*Table 6: Results from applying method 2 to Vitens data*

| | Estimates of $\text{Q-max}_{\text{max}}$ | | | Estimates of of $\text{Q-max}_{\text{med}}$ | | |
|---|---|---|---|---|---|---|
| Function | $\boldsymbol{\beta}$ | $r^2$ | MSE | $\boldsymbol{\beta}$ | $r^2$ | MSE |
| $f(N) = a \cdot N$ | $a = 0.0077$ | 0.9750 | 0.0097 | $a = 0.0060$ | 0.9827 | 0.0046 |
| $f(N) = a\sqrt{N} + b \cdot N$ | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.0296 \\ 0.0057 \end{pmatrix}$ | 0.9882 | 0.0046 | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.0092 \\ 0.0054 \end{pmatrix}$ | 0.9846 | 0.0041 |
| $f(N) = \begin{cases} a\sqrt{N} & \text{for N>T} \\ a\sqrt{T} + (N-T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$ | $\begin{pmatrix} a \\ T \end{pmatrix} = \begin{pmatrix} 0.0650 \\ 21.85 \end{pmatrix}$ | 0.9882 | 0.0046 | $\begin{pmatrix} a \\ T \end{pmatrix} = \begin{pmatrix} 0.0312 \\ 7.25 \end{pmatrix}$ | 0.9843 | 0.0042 |

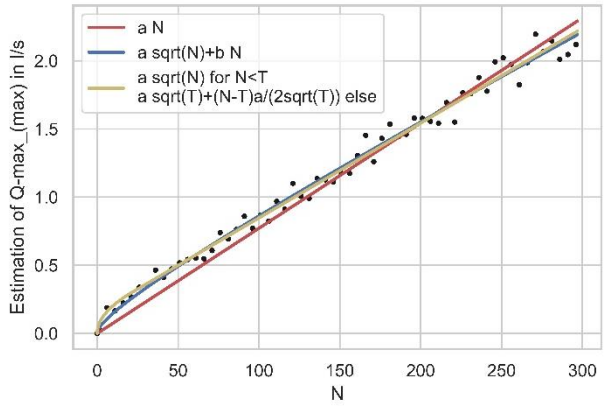


*Figure 17: Fitted functions to estimation of $Q\text{-}max_{max}$ found from Vitens data for $N \in \{0,5,\ldots,300\}$*

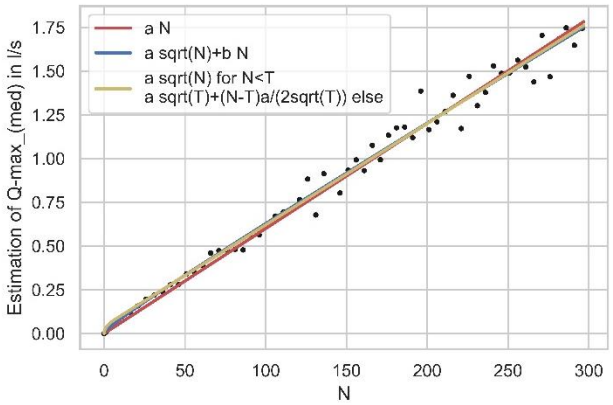


*Figure 18: Fitted functions to estimation of $Q\text{-}max_{med}$ found from Vitens data for $N \in \{0,5,\ldots,300\}$*

Overall, it seems that the correlation between $N$ and the estimates of both $\text{Q-max}_{\text{med}}$ and $\text{Q-max}_{\text{max}}$ can be described by linear functions.

## 5.3 SIMDEUM data

After estimating $\text{Q-max}_{\text{med}}$, $\text{Q-max}_{\text{max}}$, and $\text{med(Q-max)}$ using two different data sets, these quantities as well as $\text{max(Q-max)}$ will be estimated using the data that was created using SIMDEUM. A short description of SIMDEUM as well as the details on the SIMDEUM data set can be found in Section 3.3.

The results of the approach for the SIMDEUM data as described in Section 4.4 can be found in Figure 19. As can be observed, the aggregated data resulted in much smaller estimates, as was also observed from the estimates using real-life data and is thus as expected. Furthermore, it can be noted that the estimate of $\text{Q-max}_{\text{max}}$ (red dots) is much smaller than the estimate of $\text{max(Q-max)}$ (orange dots). This shows that the found $\text{Q-max}$ on max-days is smaller than at least 50% of the found $\text{Q-max}$. This shows that the $\text{Q-max}$ of a max-day is not necessarily higher than on any other random day. Note, that the max-day was determined based on the total water demand of a day. It is important to recognize that this is a clear indication that the $\text{Q-max}$ of the max-day of a time span and the maximal $\text{Q-max}$ of this same time span are not necessarily the same. This could be investigated by looking at the correlation between the total water demand of a day and the peak the water demand of that day. Further investigating this correlation is left for future research.

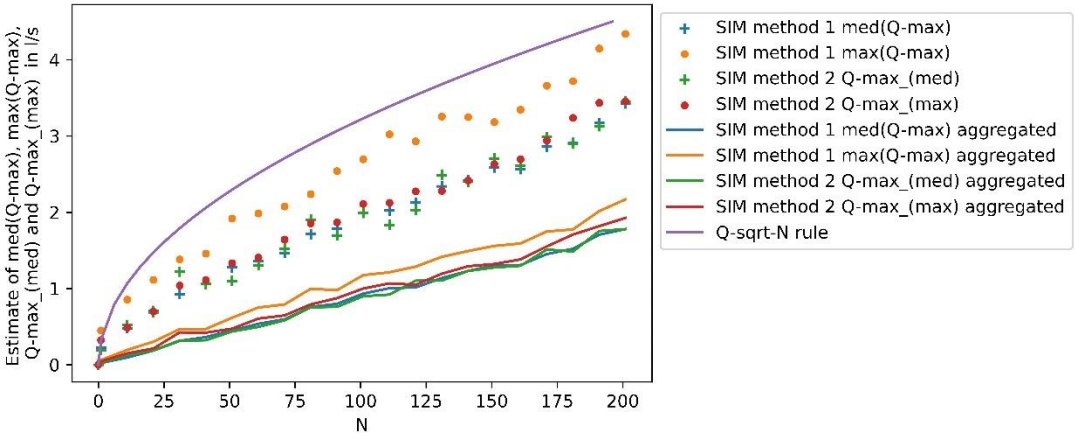


*Figure 19: Estimates of $Q\text{-}max_{med}$, $Q\text{-}max_{max,}$, $max(Q\text{-}max)$ and $med(Q\text{-}max)$ from SIMDEUM data*

To quantify the correlation between $N$ and $\mathrm{med(Q\text{-}max)}$ and $N$ and $\mathrm{max(Q\text{-}max)}$ multiple functions were fit to the found estimates (for $N \in \{0,1,11,..,201\}$). As described in Section 4.3, a non-linear least squares approach was applied. After finding the optimal values of $\boldsymbol{\beta}$, the $r^2$ was determined. Some of the functions that were fit with the corresponding values of the parameters $\boldsymbol{\beta}$ and the found $r^2$ and $\mathrm{MSE}$ can be found in Table 7. Furthermore, the results with the found functions can also be observed in Figure 20 and Figure 21. As can be seen, all three function are very similar and seem to fit the results from the data.

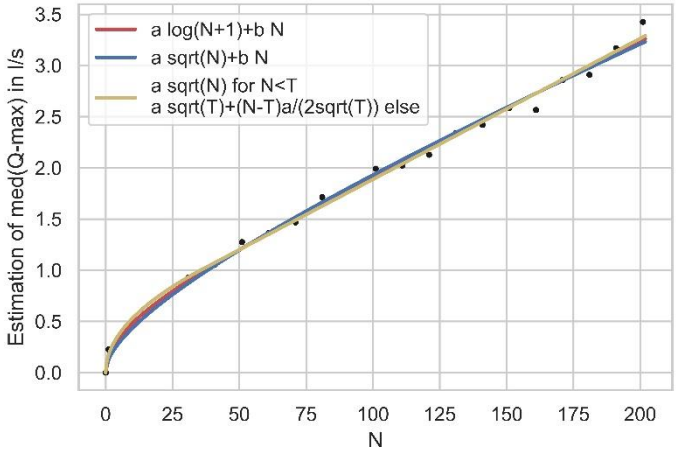


*Figure 20: Fitted functions to estimation of $med(Q\text{-}max)$ found from SIMDEUM data for $N \in \{0,1,11,\ldots,201\}$*

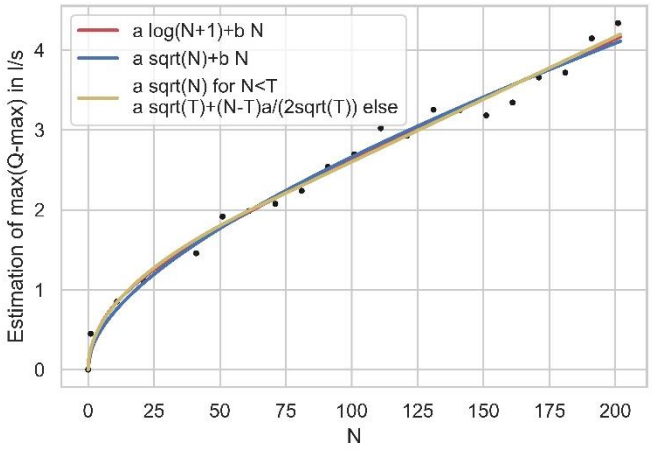


*Figure 21: Fitted functions to estimation of $max(Q\text{-}max)$ found from SIMDEUM data for $N \in \{0,1,11,\ldots,201\}$*

*Table 7: Results from applying method 1 to SIMDEUM data*

| | Estimates of $\mathrm{med(Q\text{-}max)}$ | | | Estimates of of $\mathrm{max(Q\text{-}max)}$ | | |
|---|---|---|---|---|---|---|
| Function | $\boldsymbol{\beta}$ | $r^2$ | MSE | $\boldsymbol{\beta}$ | $r^2$ | MSE |
| $f(N) = a\log(N+1) + b \cdot N$ | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.150 \\ 0.012 \end{pmatrix}$ | 0.9947 | 0.0047 | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.283 \\ 0.013 \end{pmatrix}$ | 0.9879 | 0.0165 |
| $f(N) = a\sqrt{N} + b \cdot N$ | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.111 \\ 0.008 \end{pmatrix}$ | 0.9939 | 0.0054 | $\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0.211 \\ 0.006 \end{pmatrix}$ | 0.9872 | 0.0174 |
| $f(N) = \begin{cases} a\sqrt{N} & \text{for N>T} \\ a\sqrt{T} + (N-T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$ | $\begin{pmatrix} a \\ T \end{pmatrix} = \begin{pmatrix} 0.168 \\ 37.2 \end{pmatrix}$ | 0.9950 | 0.0044 | $\begin{pmatrix} a \\ T \end{pmatrix} = \begin{pmatrix} 0.255 \\ 66.0 \end{pmatrix}$ | 0.9882 | 0.0160 |

Overall, the correlation between $\mathrm{med(Q\text{-}max)}$ and $N$ and the correlation between $\mathrm{max(Q\text{-}max)}$ and $N$ can be described by multiple functions. Note, that they can be described by the same functions (with different parametrizations). Since WMD currently assumes the 'q-square-root-N' rule for $N < 200$ houses and a linear function for $N > 200$ and the function $f(N) = \begin{cases} a\sqrt{N} & \text{for N>T} \\ a\sqrt{T} + (N-T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$ fits the estimates, the rule that will be recommended based on the results from the SIMDEUM data set that estimates $\mathrm{med(Q\text{-}max)}$ for a value of $N$ is: for $N \leq 37$ take $0.17\sqrt{N}$ and for $N > 37$ take $0.01 \cdot N + 0.51$. The rule that will be recommended based on the results from the SIMDEUM data set that estimates $\mathrm{max(Q\text{-}max)}$ for a value of $N$ is: for $N \leq 66$ take $0.26\sqrt{N}$ and for $N > 66$ take $0.016 \cdot N + 1.04$.

## 5.4 Comparison of results

In this section the results of the found estimates of $\text{Q-max}_{\text{med}}$, $\text{Q-max}_{\text{max}}$, $\text{max(Q-max)}$ and $\text{med(Q-max)}$ with both the BW+WBG data as well as with the SIMDEUM data will be compared. Note, that the results of the Vitens data set will not be compared to the results of the SIMDEUM data set. This is left for future research due to time-constraints. Furthermore, the main interest of this report to investigate the relation between $\text{Q-max}$ and $N$ and hourly data underestimates the values of $\text{Q-max}$.
For $N \in \{0,1,\dots,50\}$ the results of both the SIMDEUM data as well as the results of the BW+WBG data set can be found in Figure 22 and Figure 23. In Figure 22 the estimates of $\text{med(Q-max)}$ both from using the BW+WBG data as well as from using the SIMDEUM data. It can be observed that the estimates are very close. In Figure 23 the estimates of $\text{Q-max}_{\text{med}}$ and $\text{Q-max}_{\text{max}}$ from using the SIMDEUM data can be found. It can be observed that the estimates of $\text{Q-max}_{\text{med}}$ (SIMDEUM data) are close to the estimates of $\text{med(Q-max)}$ (BW+WBG data). Furthermore, it can be observed that also $\text{Q-max}_{\text{max}}$ (SIMDEUM data) are close to the estimates of $\text{med(Q-max)}$ (BW+WBG data). This shows again that the max-day does not necessarily have a higher $\text{Q-max}$ than the median-day.

Secondly, the estimates found with the SIMDEUM data are again compared to the results of the BW+WBG data set for $N \in \{0,1,11,\dots,201\}$. The results can be found in Figure 24. It can be observed that for $N > 60$ the estimates of $\text{Q-max}_{\text{med}}$, $\text{Q-max}_{\text{max}}$, and $\text{med(Q-max)}$ based on the SIMDEUM data start to deviate slightly. These estimates start to increase slightly faster than the results found from the data set of BW+WBG. Furthermore, for $N < 25$ the estimates from SIMDEUM seem to be slightly smaller than the results found from the data set of BW+WBG. Some possible explanations of these observations will be given. However, an in depth analysis of this behavior and changes to pySIMDEUM to test the possible explanations are left for future research.

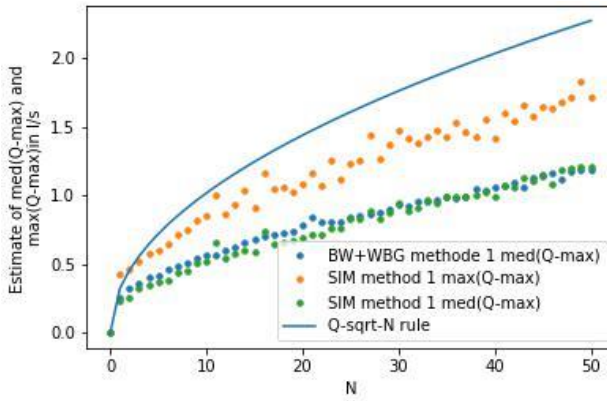


*Figure 22 : Comparison of estimates of $max(Q\text{-}max)$ and $med(Q\text{-}max)$ for $N \in \{0,1,\dots,50\}$*

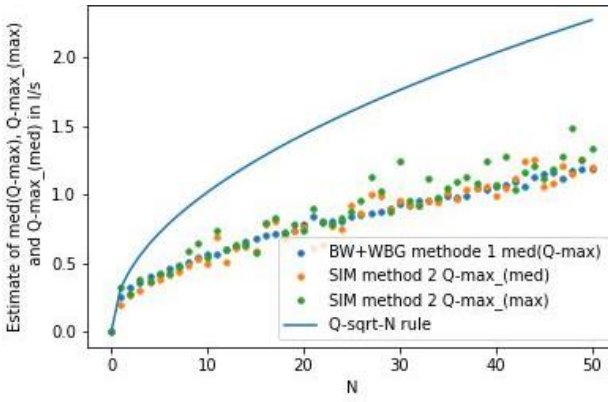


*Figure 23: Comparison of estimates of $Q\text{-}max_{med}$ and $Q\text{-}max_{max}$ with $med(Q\text{-}max)$ for $N \in \{0,1,\dots,50\}$*

An explanation of the observation that for $N < 25$ the results from the SIMDEUM data are slightly smaller than the results of data could be that all data was collected at households from employees of the water companies BW and WBG. This sample of households could be an unrepresentative sample of households throughout the Netherlands. Given that at least one person of the household is an adult with a job away from home, changes the probabilities of the type of household. As was denoted in chapter 0, the average number of users per household for the BW+WBG data set is larger than the average number of users per household that SIMDEUM assumes. This could be a cause for the difference between the results of SIMDEUM and the BW+WBG data. Next to a possibly different distribution of house types, the working adults work at the same water companies. This could cause correlations in the presence of the users, since they might have the same working hours. Which could also cause the estimated $\text{med(Q-max)}$ from

the data to be higher than the estimated $\mathrm{med(Q\text{-}max)}$ from the SIMDEUM data. However, note that the difference is very small.

An explanation of the observation that for $N > 60$ the results from the SIMDEUM data are higher than the results of the data could be that the appliances nowadays use less water than what SIMDEUM uses. It seems that the slope of the linear behavior which occurs for $N > 37$ (SIMDEUM) and for $N > 61$ (BW+WBG) is different for the estimates of SIMDEUM as well as for the results of BW+WBG. Exploring this slope and the difference between the slopes is left for future research. However, an explanation could be that the data which is used by SIMDEUM to estimate the water demand of an end-use was determined in 2006 (Blokker, 2006). These estimations of the intensity of the different appliances could be different from the current appliances. Another explanation could be that SIMDEUEM assumes that a too big part of total water demand is demanded during peak hours. SIMDEUM assumes that 65% of the water is used during peak hours[2]. If this number would be too large the results from SIMDEUM would be higher than expected for larger values of $N$. A final explanation would be that if SIMDEUM assumes that the peak hour interval to be smaller than in real-life and the same percentage of water is demanded during the peak hours, the estimates of SIMDEUM would also be larger for larger values of $N$.

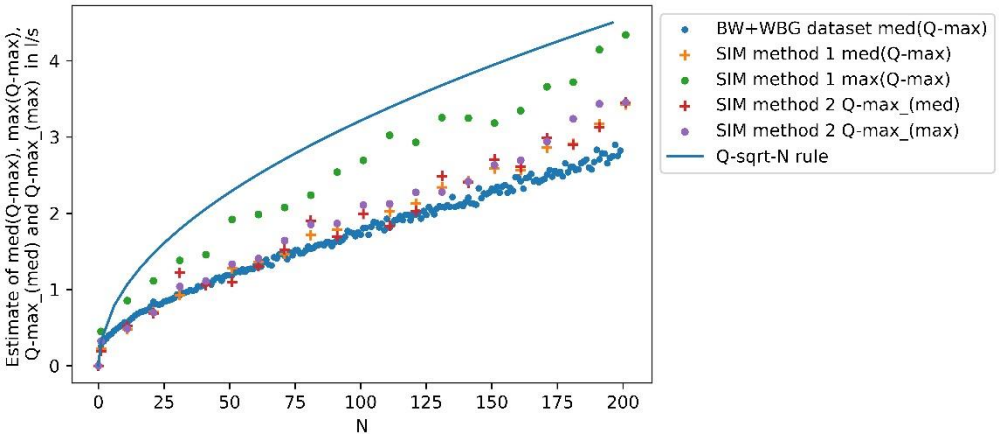


*Figure 24: Comparison of results from SIMDEUM data and estimates of $med(Q\text{-}max)$ using BW+WBG data for $N \in \{0,1,11,\ldots,201\}$*

Now the estimates of $\mathrm{max(Q\text{-}max)}$ that followed from the SIMDEUM data set will be compared to the 90% quantile of the $\mathrm{Q\text{-}max}$ found from the BW+WBG data set using version 1.3. However, as already explained before the results will mainly be determined by the few high values of the data set for larger values of $N$. The results can be observed in Figure 25. As can be seen for small values of $N$, approximately for $N < 30$ the estimates of $\mathrm{max(Q\text{-}max)}$ that followed from the BW+WBG data set are close to the estimates of $\mathrm{max(Q\text{-}max)}$ that followed from the SIMDEUM data set. More research should be done to confirm this behavior.

[2] Note that the typo as explained in Section 3.3.1 causes this percentage to be smaller.

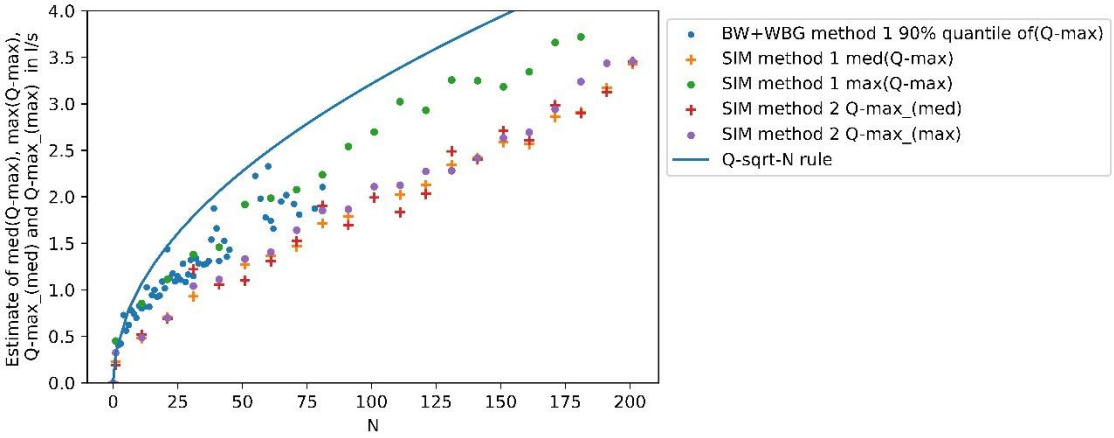


*Figure 25: Comparison of results from SIMDEUM data and estimates of $max(Q\text{-}max)$ using BW+WBG data for $N \in \{0,1,11,\ldots,201\}$*

Overall, the estimates found from the SIMDEUM data are close to the estimates that resulted from the BW+WBG data. However, for $N > 60$ the results from the SIMDEUM data start to increase slightly faster. This results into the differences between the estimates to deviate more for larger values of $N$.

# 6 Conclusion and Discussion

This report reinvestigates the correlation between $\text{Q-max}$ (the maximal daily flow of a pipe) and $N$ (the number of households supplied by this pipe). As stated by Buchberger et al. (Buchberger, Blokker, & Cole, 2012) the Netherlands used the q-squareroot-N rule. An alternative was developed based on simulation results in (E.J.M. Blokker, 2010). This report focuses on whether these rules describe the correlation between $\text{Q-max}$ and $N$ correctly or if other rules might describe the correlation more accurately.

## 6.1 Conclusion

- The q-squareroot-N rule that is used in the Netherlands (with Q= 0.32 if FU=15) overestimates the $\text{med(Q-max)}$, $\text{Q-max}_{\text{med}}$, $\text{Q-max}_{\text{max}}$ and $\text{max(Q-max)}$ for at least $N > 10$.
- Applying the in Section 0 described methods to the data described in Section 0 showed that $\text{Q-max}_{\text{max}}$ and $\text{max(Q-max)}$ are not the same for these data sets. This implies that the $\text{Q-max}$ on the maximum day is smaller than the maximum of a set of $\text{Q-max}$.
- If the measuring frequency is an hour (or approximately every hour), the estimates of $\text{med(Q-max)}$, $\text{Q-max}_{\text{med}}$, $\text{Q-max}_{\text{max}}$ and $\text{max(Q-max)}$ are a lot smaller than if the measuring frequency is a second. Since all peaks are spread out over the last hour, these will not be representative results of the actual maximal flow.
- If only a small amount of data is available, one could reuse the data to increase the preciseness of the results (decrease the variance of the estimates). However, if too much of the data is reused the results start to deviate as seen in Figure 12 where the bias starts to increase for larger values of $N$. Thus, a trade-off exists between a possible bias and the preciseness of the results.
- The difference between the results that followed from the data set of BW+WBG and the results from the SIMDEUM data set are insignificant for $N < 75$. However, for $N > 75$ the estimates of the SIMDEUM data set deviate slightly from the estimates based on the data set BW+WBG.
- Multiple functions were fit on the found estimates. All fits presented in Section 0 fit the estimates well based on the MSE and $r^2$ with respect to the MSE and $r^2$ the differences between the fits are insignificant. A function that can be used for an estimate of $\text{med(Q-max)}$ and $\text{max(Q-max)}$ with different constants is $f(N) = \begin{cases} a\sqrt{N} & \text{for N>T} \\ a\sqrt{T} + (N-T)\frac{a}{2\sqrt{T}} & \text{else} \end{cases}$. This function is equal to the q-squareroot-N rule for $N \leq T$ for some number of fixture units and for $N > T$ the function is linear with the same slope. From the estimates of $\text{med(Q-max)}$that followed from the BW+WBG data set the following rule resulted:
  For $N < 61$ take $0.17\sqrt{N}$ and for $N \geq 61$ take $0.01 \cdot N + 0.66$.
  From the estimates of $\text{max(Q-max)}$ that followed from the SIMDEUM data set the following rule resulted:
  For $N \leq 66$ take $0.26\sqrt{N}$ and for $N > 66$ take $0.016 \cdot N + 1.04$.

## 6.2 Discussion and future research

- To estimate the $\text{Q-max}$ on a 'maximal day' the maximum of a set of $\text{Q-max}$ was taken. However, in the case that possible outliers are contained in the data set, the maximum is likely to contain this outlier. Therefore it might be more sensible to take the 99% quantile.
- In future more research could be done into the correlation between the total water demand of a day and the peak of the total water demand (possibly for different values of $N$). This will create understanding about the difference between $\text{Q-max}_{\text{max}}$ and $\text{max(Q-max)}$.

- It was concluded that a measuring frequency of one hour greatly decreases the estimates of $\mathrm{med(Q\text{-}max)}$, $\mathrm{Q\text{-}max_{med}}$, $\mathrm{Q\text{-}max_{max}}$ and $\mathrm{max(Q\text{-}max)}$. However, in future one could investigate if it would be possible to disaggregate the hourly data and if this would lead to a more precise estimates.
- The estimates of $\mathrm{med(Q\text{-}max)}$ that followed from version 1.1 and 1.2 of data sampling method 1 are dependent on the order in which the days of the different households are selected. Therefore, rerunning these versions could give slightly different results. For these methods to work more accurately, more data should be used to increase the accuracy of every single the estimates.
- In future the data set of BW+WBG could be filtered better. One could investigate how to distinguish outliers and high water demands in the data set. This would result in more accurate results. This would also greatly improve the accuracy of the estimates of $\mathrm{max(Q\text{-}max)}$.
- More research could be done into the tradeoff between the accuracy of every estimate and the bias that occurs if the data is reused. Furthermore, one could also investigate what an acceptable amount of reuse of the demand patterns would be.
- In this report the default statistics of SIMDEUM were used. However, the data set of BW+WBG was the result of measurements in the homes of employees of the water companies. Also, the data set of Vitens was the result of measurements in Westeinde (Leeuwarden). However, this might not be representative of the Netherlands as a whole. Whereas the default statistics of SIMDEUM are based on total of the Netherlands. Therefore, changing the statistics of SIMDEUM could lead to a better comparison between the results of the real-life data sets and the synthetic one.
- The default statistics used by SIMDEUM were determined in (Blokker, 2006). In future this research could be updated (for example, add new appliances to the end-use and update the household and end-use statistics).
- In future a comparison could be made between the estimates following from the Vitens data set and the estimates following from the SIMDEUM data set. This would validate that SIMDEUM accurately mimics hourly peaks within a water demand pattern.
- In future, the found correlation between $\mathrm{med(Q\text{-}max)}$ and $N$ and the correlation between $\mathrm{max(Q\text{-}max)}$ and $N$ should be validated with other data sets (with a measurement frequency of one second). To ensure that this correlation holds in more generality. Note, that currently (November 2022) this data is unavailable to the author.

# 7 Bibliography

Alvisi, S., Franchini, M., & Marinelli, A. (2007, 1). A short-term, pattern-based model for water-demand forecasting. *Journal of Hydroinformatics, 9*(1), 39-50. Retrieved from http://iwaponline.com/jh/article-pdf/9/1/39/392840/39.pdf

Blokker, E. (2006). Modelleren van waterverbruik in huishoudens. Retrieved from https://livelink.kwrwater.nl/livelink/livelink.exe?func=ll&objaction=overview&objid=51462119

Buchberger, S., Blokker, M., & Cole, D. P. (2012). Estimating peak water demands in hydraulic systems, I-current practice.

Drinkwaterbesluit - Artikel 45. (n.d.). Retrieved from https://wetten.overheid.nl/BWBR0030111/2022-07-01/0#search_highlight0

E.J.M. Blokker. (2010, 1). *Stochastic water demand modelling for a better understanding of hydraulics in water distribution networks.* Retrieved from https://www.researchgate.net/publication/46395035_Stochastic_water_demand_modelling_for_a_better_understanding_of_hydraulics_in_water_distribution_networks/citation/download

Steffelbauer, D., Hillebrand, B., & Blokker, E. (2022). pySIMDEUM: An open-source stochastic water demand end-use model in Python. Proceedings of the 2nd joint Water Distribution System Analysis and Computing and Control in the Water Industry (WDSA/CCWI2022) conference, Valencia (Spain), 18-22 July 2022.

Vreeburg, J. (2007). *Discolouration in drinking water systems: a particular approach.*

WMD. (2022, 5). *Jaarverslag 2021.* Retrieved from https://wmd.nl/nieuws/jaarverslag-2021/

# 8 Appendix

Pseudocode 1.2

*Algorithm 3: method 1.2*

**Algorithm 2:** BW+WBC data set, get Q max for different values of $N$, with reusing data for different values of $N$ but not within the computations for $N$

```
Input: nreps number of repetition for every value of N, dataframe (df)
  with Date, FlowPerSecond, id, Time.
Define narray to contain all possible values of N.
Define Q = zero matrix with size: len(narray) by nreps
for m in 0,...,len(narray)-1 do
    n=narray[m]
    tmp1=df
    for i in 0,...,nreps-1 do
        Let ids= all unique id's in tmp1 that still have at least one
          unused day.
        Select n random id's (households) from the ids array (without
          repetition).
        Create an empty dataframe tmp2.
        for j in 0,...,n-1 do
            Select a random unused day of the household with the j'th
              randomly selected id.
            Add this day to a dataframe tmp2 and remove it from the
              dataframe tmp1.
        end for
        Sum FlowPerSecond in tmp2 grouped by Time
        Let Q[m,i]= maximum over all times of the summed
          FlowPerSecond.
    end for
end for
```

*Algorithm 4: Method 1.3*

**Algorithm 3:** BW+WBG data set, get Q max for different values of $N$, with reusing data , random day. Note probability that a household lives in a street is higher if it has more available days

```
Input: nreps number of repetition for every value of N, narray all
  wanted values of N, dataframe (df) with Date, FlowPerSecond, id,
  Time.
Create a column called dayindex of df where every unique day of every
  household gets a seperate index (in the combined data from BW and
  WBG these indices are 0,..., 2933).
Define: Q = zero matrix with size: len(narray) by nreps
for m in 0,...,len(narray)-1 do
    n=narray[m]
    for i in 0,...,nreps-1 do
        Let random_indices = an array with n random numbers in 0,...,
          max(dayindex) (in the case of BW and WBG data) 0,...,2933).
        Let tmp = dataframe where the dayindex is in the array
          random_indices.
        Sum FlowPerSecond in tmp grouped by Time.
        Let Q[m,i]= maximum over all times of the summed
          FlowPerSecond.
    end for
end for
```

*Table 8: SIMDEUM statistics House*

| Household type | Household type probabilities | $N$ | Gender probabilities | Age probabilities | Probability out-of-home job |
|---|---|---|---|---|---|
| One-person | 34 | 1 | male=0.46<br>female=0.54 | child = 0<br>teen = 0<br>adult = 0.7<br>senior = 0.3 | male = 0.675<br>female = 0.524 |
| Two-person | 30 | 2 | (male,male)=0.025<br>(male,female)=0.95<br>(female,female)=0.025 | child = 0<br>teen = 0<br>adult = 0.7<br>senior = 0.3 | both = 0.494<br>only male = 0.26<br>only female =0.063<br>neither person = 0.183 |
| Family | 36 | $E[N] = 3.75$ | (male)=0.5<br>(female)=0.5 | child = 0.25<br>teen = 0.165<br>adult = 0.585<br>senior = 0 | both = 0.394<br>only male = 0.523<br>only female =0.031<br>neither person = 0.052 |