Planning Drinking Water for Airplanes

Marco Bijvank, Menno Dobber,
Maarten Soomer,
Vrije Universiteit Amsterdam

Quentin Botton, Eléonore de le Court,
Jean-Christophe Van den Schrieck, Moïra de Viron,
Université Catholique de Louvain

Myriam Cisneros-Molina, Klaus Schmitz,
University of Oxford

Remco van der Hofstad, Ellen Jochemsz,
Tim Mussche,
Technische Universiteit Eindhoven

Martin Summer,
Vorarlberger Gesellschaft für angewandte Mathematik

Maroescha Hoekstra, Jeroen Mulder,
Mark Paelinck
KLM

April 29, 2005

Abstract

The management of the Dutch national airline company KLM intends to bring a sufficient amount of water on board of all flights to fulfill customer's demand. On the other hand, the surplus of water after a flight should be kept to a minimum to reduce fuel costs. The service to passengers is measured with a service level. The objective of this research is to develop models, which can be used to minimize the amount of water on board of flights such that a predefined service level is met. The difficulty that has to be overcome is the fact that most of the available data of water consumption on flights are rounded off to the nearest eighth of the water tank. For wide-body aircrafts this rounding may correspond to about two hundred litres of water. Part of the problem was also to define a good service level. The use of a service level as a model parameter would give KLM a better control of the water surplus.

The available data have been analyzed to examine which aspects we had to take into consideration. Next, a general framework has been developed in which the service level has been defined as a Quality of Service for each flight: The probability that a sufficient amount of water is available on a given flight leg. Three approaches will be proposed to find a probability distribution function for the total water consumption on a flight. The first approach tries to fit a distribution for the water consumption based on the available data, without any assumptions on the underlying shape of the distribution. The second approach assumes normality for the total water consumption on a flight and the third approach uses a binomial distribution. All methods are validated and numerically illustrated. We recommend KLM to use the second approach, where the first approach can be used to determine an upper bound on the water level.

Keywords: rounded data, distribution estimation, cubic spline, maximum likelihood

1 Introduction

During flights people use drinking water for different purposes (e.g. consumption and going to the toilet). This water comes from a single water tank. During the preparation of each intercontinental flight, the remaining water of the previous flight is drained from the water tank and then filled up again to a predetermined level with a regulator. This regulator can only fill up the tank to multiples of 1/8th of the particular water tank. The determination of the water level is explained at the end of this section. We start with a detailed description of the water filling process.

Before the flight takes off, the purser reads the water level from a display in the cabin. On most wide-body aircrafts this display only has eight marks, which makes it difficult to determine the exact water volume in the tank. Consequently, the purser rounds off the water level to the nearest mark. The water level before take off is not by definition equal to the the amount of water pumped in the aircraft, since some water could be left behind after the draining process. We explicitly assume that the water tank is filled with high precision, such that the rounding error before take off is negligible. The same person reads the water volume in the tank again after the landing. These data records (including the aircraft type, the number of passengers, the origin and destination, the time of departure and the flight duration) are available for all flights. The only data that differ are the one from the MD-11 aircraft type. For this type, the water level is read by the purser in percentages (multiples of one hundreds). However, the regulator which is used to fill up the water tank is less precise: It fills the tank to multiples of 1/5th of the water tank.
Since the data have been gathered by operational personnel, it is very likely that some of the data are wrong. Therefore, a cleaning procedure is used to remove bad data records. First, data records with a negative water consumption are removed, as well as records with an average water consumption of more than one litre per passenger per hour. Also data records with zero water usage on long distance flights are removed.

The question becomes how the situation can be modelled such that the rounding errors are taken under consideration. This model should incorporate a suitable definition of service level as a control parameter. Based upon the model the amount of drinking water should be minimized, because any surplus of water will result in extra fuel costs.
In the current situation KLM defines the service level as the percentage of flights on the same flight leg[1] that has enough drinking water on board. So a service level of 95% for a certain flight leg means that only 5% of the flights on that flight leg will have a water shortage. A weakness of this definition is that this does not mean that a passenger will only be confronted with a water shortage in 5% of the flights. If water shortage is a structural problem on crowded flights, a passenger is more likely to be confronted with a shortage on such flights. Therefore, we recommed that the service level should be defined from a passenger's point of view. This leads to the concept of Quality of Service, as will be discussed in more detail in Section 3.
In the current approach used by KLM, the optimization of the amount of drinking water is done by calculating a regression line through the measurements of a particular flight leg. This line shows the relationship between the number of passengers and the average water consumption (see Figure 1). The regression line is then shifted upward until a certain percentage of the measurements are below this line. Finally, all the values of the shifted line are rounded upward to the nearest multiple of one eighth of the water tank. This is performed for each flight leg.
This model is based on three important assumptions. The first is that the water usage depends linearly on the number of passengers. The second is that the variance in the water usage does

[1] A flight leg is a unique origin-destination and aircraft type combination.

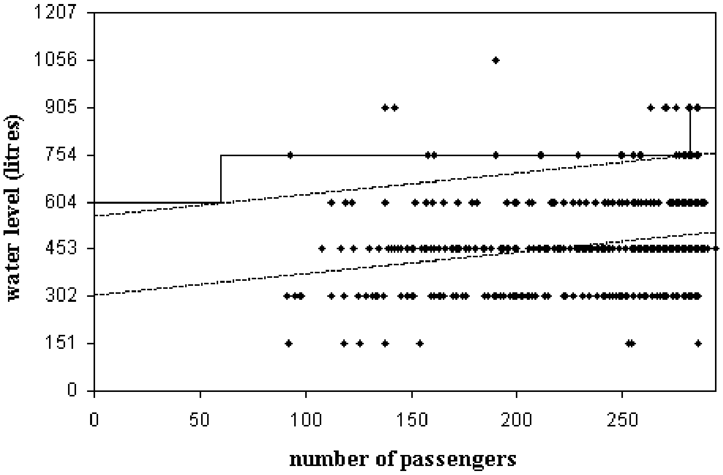

Figure 1: The water level based on the current approach used by KLM, including the regression line and the shifted regression line (dotted lines).

not depend on the number of passengers. The third is that linear regression can still be applied with rounded data. The second assumption is validated in Section 5.2. The third assumption would need further investigation, because it is not clear which of the measurements are rounded upward, and which are rounded downward. Take for instance two measurements corresponding to two similar flights A and B with 150 and 220 passengers respectively. Suppose the recorded water consumption for both flights is 5/8th of a tank. Assuming that the water consumption depends on the number of passengers on board, the water consumption of flight B is more likely to have been rounded downward than of flight A. This suggests that the probability of a value having been rounded upward or downward will most likely depend on the ratio between the number of passengers and the recorded water consumption.

This paper is organized as follows. Section 2 focuses on the data analysis of the historical data of KLM. We have evaluated possibilities to cluster different flights and studied other aspects that should be taken under consideration. A general framework has been developed for the problem in Section 3. In this section the definition of the service level is also presented. In each of the next three sections, different approach are discussed to solve the problem. Each method has been validated and illustrated on common examples. Conclusions and ideas for further research are presented in Section 7.

2 Data Analysis

In the current approach used by KLM, the historical data from flights with the same origin and destination and aircraft type are used to estimate the water usage for a flight. There are several reasons to investigate whether a larger set of flights can be used. If these flights have the same behaviour in water consumption, then using more data will give a better estimate. Furthermore, for predictions, understanding of similaraties of flights is crucial. In case of a new destination, it will be necessary to use data from flights to other destinations because there is no data available for the new destination.

Considering the whole data set (over 40.000 ”valid” records), there is significant correlation between total water usage and the number of passengers (0.49), the flight duration (0.65), and the aircraft type (0.47). The first two correlations were expected. The last correlation follows from the other two correlations; First, the influence from the aircraft types can be explained by the various tank sizes and their rounding errors, and second, the same type of aircraft is used for flights with the same duration and number of passengers. Because of the high correlation with flight duration, it is interesting to investigate whether flights with the same duration can be clustered.

To compare two flights with almost the same duration but with different destinations, the correlation between the destination and the average water usage per passenger has been calculated. A significant correlation indicates that the destination determines the average water usage, and consequently, that it is not possible to cluster the flights.

Moreover, flights with different durations were compared. The correlation between those destinations and the average water usage per person per hour was considered. To avoid the effect of different aircraft types in our analysis, the calculations have been performed considering different flights to different destinations with the same aircraft type (MD-11). The results are summarized in Table 1.

AMS -	BON	DEL	DXB	JFK	LOS	MIA	MSP	NBO	SFO	YUL	YVR	YYZ
AUA	0.024	0.181	0.097	0.058	-0.114	0.071	0.011	0.143	0.125	0.050	0.121	0.003
BON		0.108	0.054	0.017	-0.144	0.026	-0.014	0.083	0.068	0.019	0.049	-0.021
DEL			-0.024	-0.094	-0.227	-0.131	-0.180	-0.019	-0.086	-0.102	-0.067	-0.191
DXB				-0.045	-0.142	-0.058	-0.094	0.009	-0.028	-0.049	-0.020	-0.100
JFK					-0.107	-0.005	-0.066	0.069	0.035	-0.006	0.052	-0.064
LOS						0.171	0.103	0.195	0.214	0.144	0.157	0.115
MIA							-0.063	0.095	0.058	-0.004	0.061	-0.073
MSP								0.139	0.116	0.039	0.150	-0.009
NBO									-0.054	-0.076	-0.040	-0.150
SFO										-0.046	0.010	-0.130
YUL											0.045	-0.049
YVR												-0.139

Table 1: Correlation between MD-11 flights from Amsterdam.

It can be concluded that there is a correlation between most destinations and the water usage per passenger per hour. Therefore, these destinations cannot be clustered. However, there are particular cases for which the correlation is negligible. Not surprisingly, this very often happens on flights with destinations in the same region. As an example, clustering flights from Amsterdam (AMS) to Aruba (AUA) and from Amsterdam (AMS) to Bonaire (BON) looks reasonable (see also Figure 2). Day and night flights can also be clustered, since no correlation appears from the data.

The study should be extended to other destinations and aircraft types as well. Aspects of data analysis that concern the validation of the different approaches are discussed in the corresponding sections.

3 General Framework

Consider a flight with n passengers to a certain destination. The total water consumption S_n on this flight equals

$$S_n = \sum_{k=1}^{n} Y_k, \tag{1}$$

where Y_k is the water consumption of the k-th passenger (in litres). Based on the data, there is only something known about the rounded values of S_n for each level of n and for different flights [2].

[2] We know the rounded water level of the tank before take off and landing, the difference is the rounded water consumption S_n.

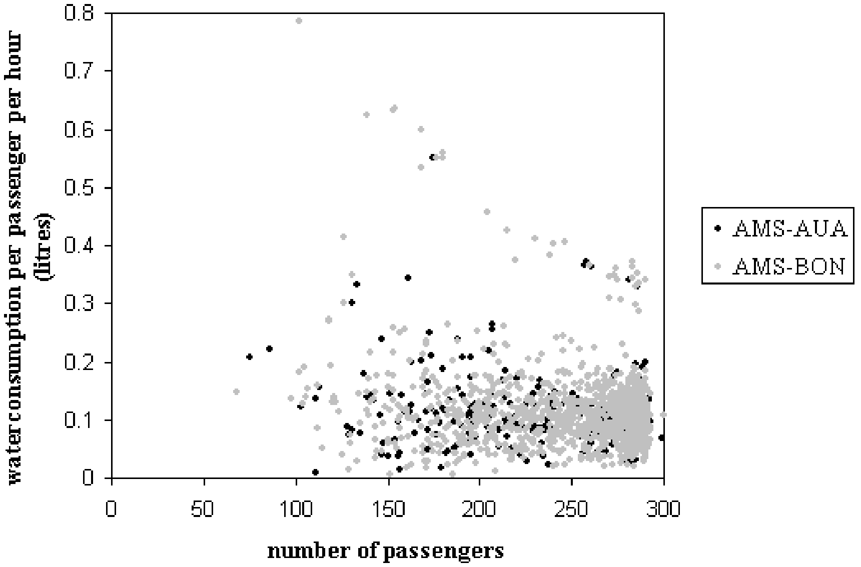

Figure 2: The water usage per passenger per hour on the flights from Amsterdam (AMS) to Aruba (AUA) and Bonaire (BON) looks similar.

For a given flight leg (for which the number of passengers n is known), the service level is defined as the probability that a sufficient amount of water is available. The service level should at least be equal to some predefined value α, which is established by the management of KLM. So, we should have

$$\mathbb{P}\left(S_n \leq \frac{j}{8}T\right) \geq \alpha, \tag{2}$$

where $j/8$ is the percentage of tank capacity filled before take off ($j \in \{0, 1, \ldots, 8\}$) and T is the tank capacity (in litres). Since the service level for all flights should satisfy the constraint formulated in Equation (2) independently of the number of passengers on the flight, it is called a Quality of Serivce (QoS).

We are interested in finding the smallest water level for which the service constraint is satisfied (i.e. the smallest value of j in Equation (2)). Therefore, we need to find a probability distribution function for the total water consumption S_n on a flight. In the following sections, three different approaches are proposed to find such a distribution. All methods use the available data to estimate this distribution. Therefore, they have to take the rounding effect into account.

4 Curve Fitting Approach

The probability that $j/8$th of the water tank is used on a flight with n passengers, is derived from the data by looking at the frequencies how often this occurs. These probability are denoted by p_j:

$$p_j = \mathbb{P}\left(\left(\frac{j}{8} - \frac{1}{16}\right)T < S_n \leq \left(\frac{j}{8} + \frac{1}{16}\right)T\right), \quad j = 0, 1, \ldots, 8 \tag{3}$$

This can be seen as a probability mass function (pmf) of the water consumption during a flight. Based on these nine probabilities, a probability density function (pdf) of the total water usage on a flight can be estimated by fitting a curve through the pmf and then normalizing this curve such that the mass below the continuous function adds up to one.

The procedure described above has to be performed for the water consumption of a known number of passengers n. There is, however, a limited amount of measurements available on a particular flight leg for this fixed number of passengers n. In order to find enough data records to base the pmf on, the measurements for the surrounding number of passengers are used as well. We assume that at least 100 measurements are required to find a representative pmf.

The next step is to find an interpolation formula between these points. Therefore, an analytic expression for $f(x)$ (where x is the total water consumption on a flight in litres) has to be formulated, where

$$f\left(\frac{j}{8}T\right) = p_j, \quad j = 0, 1, \ldots, 8 \tag{4}$$

such that the value of $f(x)$ can be calculated at any arbitrary point.

Interpolation schemes must model the function by some plausible functional form. By far most common among the functional forms used are polynomials. One of them is Lagrange's classical formula. Since we have nine known values, this results in a high order polynomial. A characteristic of high ordered polynomials is that they tend to have a wild oscillation behaviour between the values (Press et al. [4]). This is not desirable, since we assume a smooth form for the density function for the total water usage.

Another possibility is cubic spline interpolation. Splines tend to be more stable than polynomials. The goal of cubic spline is to get an interpolation formula that is smooth in the first derivative, and continuous in the second derivative. Roughly, the idea is to take the first three data points and fit a second degree polynomial. The same has to be done for the 2^{nd}, 3^{th} and 4^{th} data point etc. Finally, these polynomials have to be concatenated together such that a continuous function appears. The exact formulation is

$$f(x) = Ap_j + Bp_{j+1} + Cf''(x_j) + Df''(x_{j+1}), \quad x_j \leq x \leq x_{j+1} \tag{5}$$

where $x_j = \frac{j}{8}T$ and A, B, C and D are defined as

$$A = \frac{x_{j+1} - x}{x_{j+1} - x_j} \qquad B = 1 - A$$

$$C = \frac{1}{6}(A^3 - A)(x_{j+1} - x_j)^2 \qquad D = \frac{1}{6}(B^3 - B)(x_{j+1} - x_j)^2 \tag{6}$$

The only problem now is that we supposed the $f''(x_j)$'s to be known, when, actually, they are not. The key idea of a cubic spline is to require a continuous interpolation scheme. This is realized by getting equations for the second derivatives $f''(x_j)$, given by

$$\frac{x_j - x_{j-1}}{6} f''(x_{j-1}) + \frac{x_{j+1} - x_{j-1}}{3} f''(x_j) + \frac{x_{j+1} - x_j}{6} f''(x_{j+1}) = \frac{p_{j+1} - p_j}{x_{j+1} - x_j} - \frac{p_j - p_{j-1}}{x_j - x_{j-1}} \tag{7}$$

for $j = 1, 2, \ldots, 7$. This equation gives seven linear equations and nine unknowns, therefore we set $f''(x_0) = f''(x_8) = 0$. For more details see Press et al. [4]. Note that $C = 0$ and $D = 0$ results in a piecewise linear interpolation scheme.

This continuous function $f(x)$ has to be normalized such that the function becomes a probability density function $g(x)$.

$$g(x) = \frac{\big(f(x)\big)^+}{\int_0^T \big(f(y)\big)^+ dy}, \quad x \in [0, T] \tag{8}$$

where the value of the integral can be found with the use of numerical integration.

The service level is given by

$$\mathbb{P}\left(S_n \leq \frac{j}{8}T\right) = \int_0^{\frac{j}{8}T} g(x)dx, \quad j \in \{0, 1, 2, \ldots, 8\} \tag{9}$$

The next step is to find the minimum water level for which this service level is at least α, as shown in Equation (2). In practice, the data can give rise to the fact of taking less drinking water on board when there are more passengers. This is not logical. Therefore, we adjust the water level, such that it becomes monotonically increasing with the number of passengers on a flight. A second modification is required for flights with small number of passengers, since there are no data available in those situations. When the assumption is made that a person consumes at most one litre of water per hour, we get the following upper bound on the water level

$$\text{water level for } n \text{ passengers} \leq nD, \tag{10}$$

where D is the scheduled duration of the flight. The outcome of Equation (10) has to be rounded upward to a multiple of an eighth of the water tank.

4.1 Validation

The MD-11 data are used for validating the approach, because the tank volume is read in percentages and, therefore, more precise. Since the number of data records for the flights from Amsterdam (AMS) with destinations Bonaire (BON) and Aruba (AUA) is numerous, this cluster of flights is examined. As explained in Section 2, these flights are equivalent and therefore can be grouped together.

In the curve fitting approach two assumption are made. The first assumption is that the density of the water consumption does not vary too much for the same order of number of passengers. So, if 100 flights are grouped together this gives a good impression of the true density. Secondly, the probability mass function of Equation (3) can be translated into a probability density function with the use of an interpolation scheme. In particular, the density of the water consumption needs to be sufficiently regular.

The first assumption depends upon the data. When there is enough data available, this assumption can be justified. Otherwise, the tails of the distribution are too thick. This is explained in more detail in Section 4.3. The second assumption can be checked with the use of the MD-11 dataset. Figure 3 shows the frequencies of water consumption for the flights to the Antilles with 285 passengers. It also shows the estimated density function of the water consumption, when the data are rounded to multiples of eighths and subsequently cubic spline is applied. The estimated density function coincides with the form of the true density function, which is represented by the histogram. Based on this result, we might conclude that cubic spline interpolation schemes give a

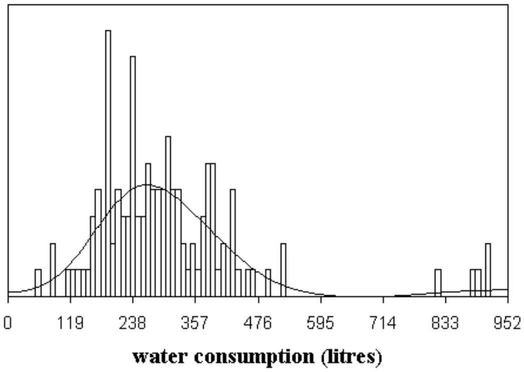

Figure 3: The histogram of the true water consumption and the estimated density function for the total water consumption on a flight leg with 285 passengers.

representative estimation for the probability distribution functions of the total water consumption on a flight leg.

4.2 Numerical example

The curve fitting approach is illustrated for the flight from Amsterdam (AMS) to Bangkok (BKK) using a Boeing 74E aircraft. For this particular flight leg, the dataset contains 548 records. The number of passengers ranges from 91 untill 294, with an average of 243 passengers. The management of KLM decided to use a service level requirement of 95%.

The results for the curve fitting approach, including the data points, are given in Figure 4. The outcome may seem strange, because the required water level for 175 passengers is equal to the water level when 250 passengers are on a flight. Figure 5 shows the estimated density functions of the total water consumption for a flight with 175 and 250 passengers. Based on these two densities we can hardly make any difference in the tail behaviour. Hence, the outcome of the method is

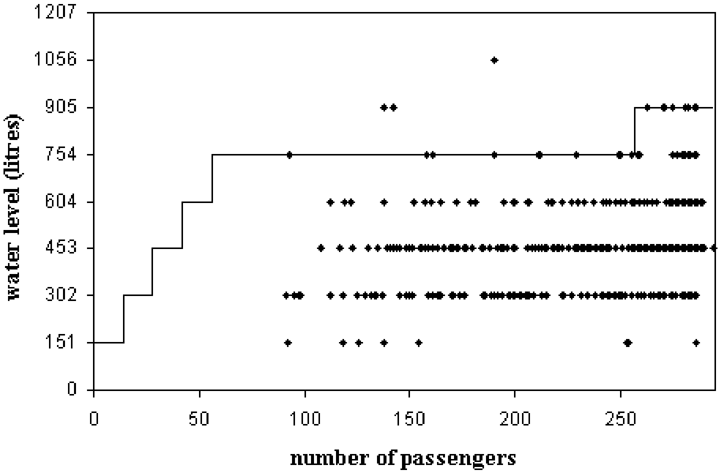

Figure 4: The water level based on the curve fitting approach for the AMS-BKK flight with the 74E aircraft and a 95% service level constraint.

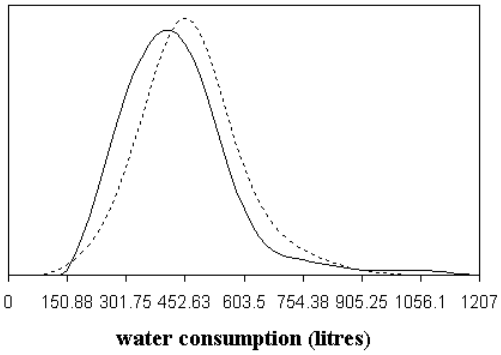

Figure 5: The estimated density function for the total water consumption when 175 passengers (straight line) and 250 passengers (dotted line) are on board.

the same. Figure 5 shows nicely a shift in the average water consumption when the number of passengers increases.

4.3 Performance

The curve fitting approach makes no assumptions about the distribution for the total water consumption on a flight, other than that it has a smooth form without any wild oscillation behaviour. When the water level for a flight with n passengers has to be determined, this method looks at all data records on flights with n passengers. If there are not enough data available on those flights, we use a subset of the data records surrounding n passengers. Consequently, the distribution for the total water usage on a flight gets thicker tails and the determined water level is too high. Hence, it becomes very relevant to find possibilities to aggregate data from different flight legs. Especially if the service level should be high, since the tail behaviour of the water consumption is becoming more relevant in those situations. Another reason why we can conclude that the curve fitting approach results in an upper bound on the water level, is the fact that errors in the data strongly influence the outcome. Figure 3 and Figure 5 show that the water usage distribution has a skewness to the right (even a strange increase). This could be because of errors in the data, which result in a higher required water level.
In conclusion, the curve fitting approach should give the best results since there is no assumption made about the distribution for the total water usage. However, this method can only be applied when enough data records are available. Otherwise the outcome is an upper bound on the required water level. For a lot of flight legs there is not much data available. Therefore, we need to develop another more suitable method.

5 Normality Approach

The previous approach uses only a subset of the data to estimate a distribution for the total water consumption of n passengers. However, when the assumption is made that the water consumption of each passenger for a particular flight is an independent and identically distributed (i.i.d.) random variable and since the number of passengers is typically large, the central limit theorem can be applied (Ross [6]). When we generalize this theorem by adding a constant to the average and variance of the water consumption, we get

$$S_n \overset{d}{\sim} \mathcal{N}(\mu_0 + n\mu, \sigma_0^2 + n\sigma^2), \tag{11}$$

where μ is the average water usage of a person on a flight, μ_0 is a constant representing the water usage that is always required, σ^2 is the variance of this water usage and σ_0^2 is a constant added to the variance. All four parameters are expressed in litres. Nonetheless, the assumption of independence and identical distribution for the water usage per person is not really needed. The central limit theorem behaviour for sums of random variables holds more generally than under the assumption of i.i.d. summands (Feller [2]). The other assumptions will be verified in Section 5.2. The values for the four parameters μ_0, μ, σ_0^2 and σ^2 have to be estimated. This estimation is done using the maximum likelihood method. In contrast to the previous approach, this approach uses all data to perform the estimation. When the estimates for the parameters are determined, we have an estimation for the density function of the total water usage (S_n) and we can calculate the service level for any tank volume by Equation (2).

Vardeman and Lee [7] give a systematic study of statistical analysis with rounded data. They also suggest a maximum likelihood approach if the distance between the largest and the smallest rounded value is not larger than the rounding unit. The normality approach is different in the sense that in this problem the water usage depends on the number of passengers.

5.1 Maximum likelihood estimation

The likelihood of the data is the probability of observing the data for certain parameter values (Oosterhoff and Van der Vaart [5]) and is expressed by Equation (12).

$$L(\mu_0, \mu, \sigma_0^2, \sigma^2; x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p_{n_i}(x_i \mid \mu_0, \mu, \sigma_0^2, \sigma^2), \tag{12}$$

where N is the total number of flights in the data set, n_i is the number of passengers in the i-th flight, x_i is the amount of water used during flight i (in litres)[3] and $p_{n_i}(x_i|\mu_0, \mu, \sigma_0^2, \sigma^2)$ is the probability of observing a water usage of x_i on a flight with n passengers and with μ_0, μ, σ_0^2 and σ^2 given:

$$\begin{aligned} p_n(x \mid \mu_0, \mu, \sigma_0^2, \sigma^2) &= \mathbb{P}\Big(x - \frac{1}{16}T \le S_n \le x + \frac{1}{16}T\Big) \\ &= \int_{x-\frac{1}{16}T}^{x+\frac{1}{16}T} \frac{1}{\sqrt{2\pi(\sigma_0^2 + n\sigma^2)}} e^{\frac{-(y-\mu_0-n\mu)^2}{2(\sigma_0^2+n\sigma^2)}} \, dy, \end{aligned} \tag{13}$$

since the distribution of S_n is given by Equation (11).
The objective is to find the values of the four estimators, which maximize this likelihood funtion. The log-likelihood function of equation (12) has quite a regular behaviour. Although we do not give a general proof of concavity, we illustrate concavity empirically by applying the function to the water consumption data. Note that a function $f : \mathbb{R}^N \to \mathbb{R}$ is concave if the contour set $C = \{(x, v) \in \mathbb{R}^{N+1} : v \le f(x), x \in \mathbb{R}^N\}$ is convex (Mas-Colell et al. [3]). Figure 6 shows that the contours of the likelihood function for the data of the AMS-BKK flight are convex. Hence, the log-likelihood function is concave.

Applying a numerical optimization method to the log-likelihood function of Equation (12) will result in a unique maximum. Since the log-likelihood function is also differentiable in this case, the maximizers can be found by solving the first order conditions for the four parameters. This procedure gives a system of four quite complicated nonlinear equations. Therefore, this indirect method is computationally more costly than numerically optimizing the log-likelihood function.

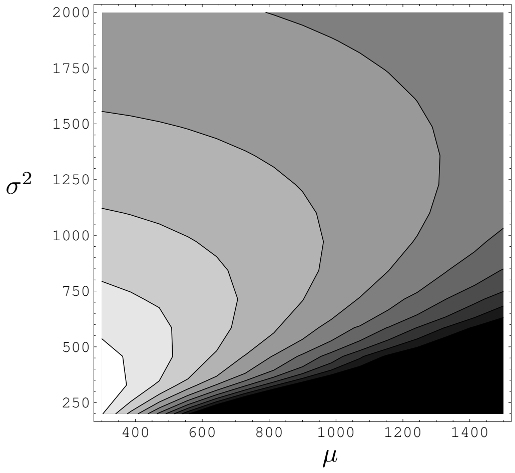

Figure 6: The contour plot of the log-likelihood function for the flight AMS-BKK.

5.2 Validation

This approach relies on the assumption that the water usage per fixed number of passengers n for a particular flight leg (or equivalent flight legs, as described in Section 2) has a normal distribution

[3] based on the data $x_i \in \{0, \frac{1}{8}T, \ldots, T\}$

with parameters $\mu_0 + n\mu$ and $\sigma_0^2 + n\sigma^2$. To validate these assumptions, the same setting is used as in the validation section of the curve fitting approach (the flights AMS-BON and AMS-AUA for the MD-11 aircraft).
In order to say something about the distribution for a fixed number of passengers, the data have been divided into small ranges of passenger numbers. Each range contains 100 measurements. Figure 7 shows the water usage in percentages of the water tank (from $0 - 5\%$, $5 - 10\%$, etc.) plotted against the frequency. At first impression, a normal distribution does not seem farfetched.

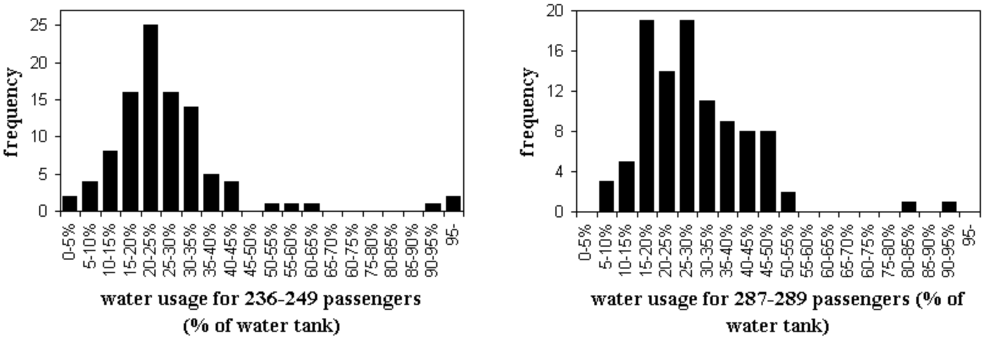

Figure 7: The frequencies of the water usage in percentages of the water tank for respectively 236-249 passengers and 287-289 passengers on the AMS-AUA and AMS-BON flights.

In most of the ranges, there are some strange values with a high water consumption. This seems to contradict a normal distribution, but these could also be the result of errors in the data (as mentioned in Section 4.3). To quantify normality, some statistical tests, like the Shapiro-Wilk and the Kolmogorov-Smirnov test, are performed.
The Shapiro-Wilk test (De Gunst and Van der Vaart [1]) is commonly used for testing normality of a given set of data points, where

H_0: the data come from a normal distribution (null hypothesis)
H_1: the data do not come from a normal distribution (alternative hypothesis)

The Kolmogorov-Smirnov test (De Gunst and Van der Vaart [1]) enables to compare the distributions of two datasets v and w, in which

H_0: v and w have the same distribution
H_1: v and w do not have the same distribution

For v the actual flight data are used, while for w random numbers are drawn from a normal distribution with a mean equal to the sample average of v and a variance equal to the sample variance of v. Besides normality, a logistic distribution could be used as well, to find out whether the actual distribution of the water consumption has thicker tails. The p-value of a test refers to the probability of wrongly rejecting the null hypothesis if it is in fact true. Small p-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis. The p-value is compared with a significance level. If it is smaller, the result is significant enough to reject the null hypothesis, otherwise it is not rejected (which does not imply that it must be true).
Table 2 shows the actual p-value per data range for the two specifics tests. Since Figure 7 already shows that the data contain some strange large water usages, we also tested each passenger range without these so called outliers. These results are also presented in Table 2. Globally, for the intervals with a large number of passengers (for which the ranges are also narrow) it can be concluded that the assumption of normality is supported. A logistic distribution seems, however, to fit the data more closely.

Another assumption made in the normality approach is linearity of the average water consump-

passenger	all data			without outliers		
range	Shapiro-Wilk	Kolomogorov Smirnov		Shapiro-Wilk	Kolomogorov Smirnov	
		normality	logistic		normality	logistic
68-160	1.80E-12	0.00010	0.00011	0.00006	0.0387	0.0702
160-187	1.53E-12	0.00043	0.00083	0.01588	0.5460	0.3408
187-204	1.03E-03	0.31730	0.40750	0.00103	0.3173	0.4075
204-220	1.07E-09	0.01938	0.03544	0.00678	0.4551	0.3126
221-236	1.09E-09	0.07637	0.11520	0.02192	0.4509	0.5142
236-249	3.74E-11	0.00309	0.01168	0.00284	0.3450	0.2995
249-259	1.10E-11	0.00381	0.01426	0.00992	0.4516	0.2865
259-268	9.08E-10	0.02498	0.04621	0.07290	0.5240	0.6547
268-274	1.94E-10	0.02069	0.05334	0.70550	0.6632	0.3295
275-280	4.04E-08	0.13770	0.26860	0.06546	0.6680	0.3275
280-284	1.13E-12	0.00135	0.00375	0.66710	0.8474	0.6001
284-287	2.70E-10	0.01262	0.03900	0.69250	0.8698	0.5270
287-289	8.91E-07	0.17030	0.15920	0.04831	0.4812	0.2129
289-300	4.17E-08	0.04494	0.11460	0.20670	0.4556	0.8048

Table 2: The p-values for the different normal distribution tests.

tion with the number of passengers and linearity of the variance. For all ranges of number of passengers, the mean and variance of the water usage has been computed and plotted to see if there might be a linear trend. As can be seen in the left hand side of Figure 8, the average total water consumption could be interpreted as being linear in the number of passengers. However, this can not be said for the variance (the right hand side of Figure 8).

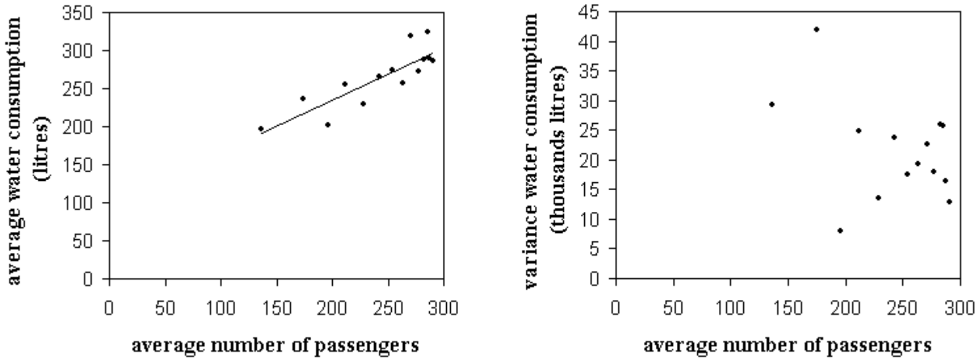

Figure 8: The average water consumption on a flight seems to have a linear trend with the number of passengers, while the variance does not.

5.3 Numerical example

For the numerical example, the same setting is used compared to the curve fitting approach as explained in Section 4.2. Based on the implementation, we find $\hat{\mu_0} = 309.6$, $\hat{\mu} = 0.68$, $\sigma_0^2 = 129.4$ and $\hat{\sigma} = 0$. The trendline through the absolute value of the residuals, after subtracting the regression line from the data points, is constant. Therefore, the variation of the total waterusage on a flight can be seen as a constant, independent of the number of passengers.
Because the estimators for the parameters are known, we have found an estimate for the probability

density function of the total water consumption S_n on a flight. The service level can be determined for a given water volume in the tank by Equation (2). The lowest values of this tank volume that satisfy the Quality of Service constraint of 95%, are given in Figure 9 for each number of passengers n.

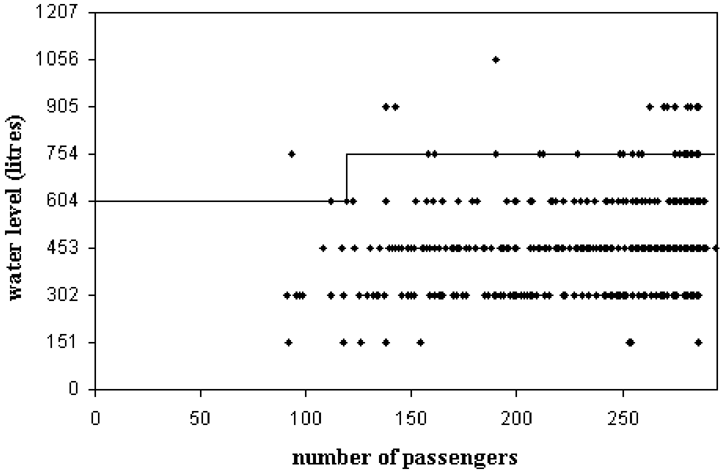

Figure 9: The water level based on the normality approach for the AMS-BKK flight with the 74E aircraft and a 95% service level constraint.

5.4 Performance

The normality approach uses all data to say something about the water consumption for each passenger (the average and the variance). Therefore, errors in the data do not have a great effect on the outcome. However, the normality approach uses the assumption of a linear growth of the variance with the number of passengers. This is not supported by the data. Therefore, the variance becomes indifferent of the number of passengers. This is the same as the assumption made in the current approach used by KLM (see Section 1 for an explanation of the regression line approach used by KLM). We have to stress here, that these findings depend on the data and are therefore not generic. The values for $\hat{\mu_0}$ and $\hat{\mu}$ are almost similar to the coefficients for the regression line through the data points. Hence, this approach looks a lot like the regression line approach. The only difference is the way how the regression line is shifted upward. In the normality approach, the line is shifted based on the normality distribution $\mathcal{N}(0, \sigma_0^2)$. Where in the regression line approach, the line is shifted based on some linear relationship. The variance of the regression line approach is larger compared to the findings of the normality approach.

6 Binomial Approach

In the previous approach no assumption was made on the distribution of the individual water consumption per passenger (only i.i.d.). In this section, the extra assumption is made that a passenger can either consume a maximum amount of water (M) with probability p or a minimum amount (m) with probability $1 - p$. So, for each passenger k the consumption (in litres) is distributed as follows:

$$Y_k = \begin{cases} M & \text{with probability } p \\ m & \text{with probability } 1 - p \end{cases} \tag{14}$$

This has the nice characteristic that the number of passengers (l) that uses the maximum amount of water has a binomial distribution (for the minimum amount of water usage this holds as well). The total water consumption on a flight with n passengers will be given by $S_n = lM + (n-l)m = l(M-m) + nm$, such that

$$\frac{S_n - nm}{M-m} \overset{d}{\sim} bin(n,p) \tag{15}$$

and the service level equals

$$\begin{aligned} \mathbb{P}\left(S_n \leq \frac{j}{8}T\right) &= \mathbb{P}\left(\frac{S_n - nm}{M-m} \leq \frac{\frac{j}{8}T - nm}{M-m}\right) \\ &= \sum_{l=0}^{\left\lfloor \frac{\frac{j}{8}T - nm}{M-m} \right\rfloor} \binom{n}{j} p^l (1-p)^{n-l}, \quad j \in \{0, 1, \ldots, 8\} \end{aligned} \tag{16}$$

where l is the number of passengers asking for their maximum allowance.

In order to determine the smallest number of eighths required such that the service requirement is satisfied, the values for M, m and p have to be estimated. The first parameter can be based on the data of a particular flight leg. We look at the maximum water usage per passenger per hour and multiply this with the scheduled duration of the flight.

$$\widehat{M} = \max_{i=1,\ldots,N} \left\{ \frac{x_i}{n_i d_i} \right\} D, \tag{17}$$

where d_i is the actual duration of flight i and D is the scheduled duration of the flight. We assume $\hat{m}$=0. The estimation of the third parameter p is very important. In general, it can be seen as a measure of the dispersion of the data within the minimum m and the maximum M values. Let $\hat{\mu}$ be the estimator for the average water consumption per passenger, given by the sample average

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} d_i n_i} D \tag{18}$$

A natural proposal for the estimator of p equals

$$\hat{p} = \frac{\hat{\mu} - \hat{m}}{\widehat{M} - \hat{m}}, \tag{19}$$

since $\hat{\mu} = \widehat{M}\hat{p} + \hat{m}(1-\hat{p})$.

This way of estimating the parameters is more or less using intuition. We could however also use maximum likelihood, as we did in the previous approach. Equation (12) can still be applied, but now the probability of observing a water usage of x on a flight with n passengers and with M, m and p becomes

$$\begin{aligned} p_n(x \mid m, M, p) &= \mathbb{P}\Big(\left\lceil \frac{x - \frac{1}{16}T - nm}{M-m} \right\rceil \leq \frac{S_n - nm}{M-m} \leq \left\lfloor \frac{x + \frac{1}{16}T - nm}{M-m} \right\rfloor \Big) \\ &= \sum_{l=A}^{B} \binom{n}{l} p^l (1-p)^{n-l}, \end{aligned} \tag{20}$$

where

$$A = \left\lceil \frac{x - \frac{1}{16}T - nm}{M-m} \right\rceil \tag{21}$$

and

$$B = \left\lfloor \frac{x + \frac{1}{16}T - nm}{M-m} \right\rfloor \tag{22}$$

since the distribution of S_n is given in Equation (15).

6.1 Numerical example

The same setting is used for the numerical example as in the previous approaches. The average water consumption per passenger is 1.96 litres (this corresponds with $\hat{\mu}$ expressed in Equation (18)), with a maximum of $\widehat{M} = 8.36$ litres per passenger. The probability of using the maximum quantity of water $\widehat{M}$ is calculated by Equation (19)

$$\hat{p} = \frac{1.96}{8.36} = 23.42\% \tag{23}$$

When we use the values for the parameters m, M and p which maximize the log-likelihood of Equation (20), we get $\hat{m} = 0$, $\widehat{M} = 8.17$ and $\hat{p} = 0.239$. These values are somewhat similar as we expected based upon intuition. The resulting thresholds for the tank volume are presented in Figure 10.

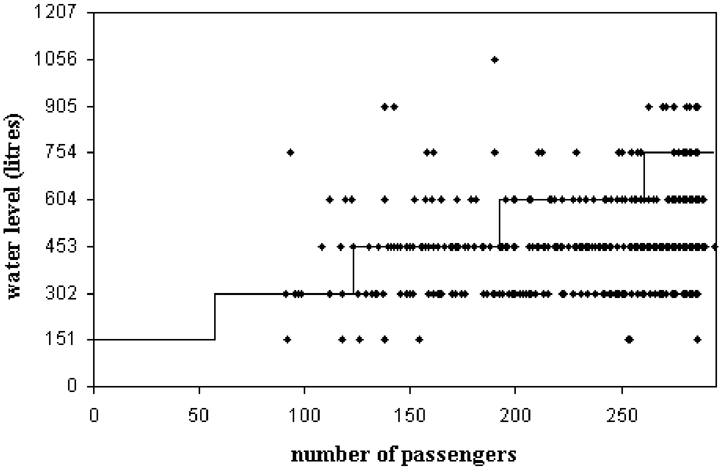

Figure 10: The water level based on the binomial approach for the AMS-BKK flight with the 74E aircraft and a 95% service level constraint.

6.2 Performance

The binomial approach can be seen as a special case of the normaltiy approach , since the binomial distribution gets close to a normal distribution for flights with many passengers. There are however some extra assumptions, which do not seem realistic, but they are an easy way to model the problem. It has the nice characteristic that the minimum and maximum water consumption is bounded.

The difference with the normality approach can be expressed in the average and variance of the total water consumption for both approaches:

	mean	variance
normality	$\mu_0 + n\mu$	$\sigma_0^2 + n\sigma^2$
binomial	$(M-m)np + nm$	$(M-m)^2 np(1-p)$

In the binomial approach we use $\hat{p} = \frac{\hat{\mu}-\hat{m}}{\widehat{M}-\hat{m}}$, which comes down to $(\widehat{M} - \hat{m})n\hat{p} + n\hat{m} = n\hat{\mu}$ for the average water consumption on a flight with n passengers. Figure 8 shows that the average water consumption is indeed linear with the number of passengers. However, the linear relationship does not go through the origin of the graph. The variance for the binomial approach is also assumed to be linear with the number of passengers. This is also not supported by the data.

7 Conclusions and Future Research

In this paper we developed a framework to determine the minimal amount of drinking water on board of flights such that a predefined service level is met. We expressed the service level as the probability that a sufficient amount of water is available to fulfill passengers demand, wich is a Quality of Service. This way of formulating the problem was an eye-opener for KLM and brought new insights to the problem.

The next step was to estimate the probability density function of the total water usage on a flight. Since the available data only give information about the water consumption in multiples of 1/8th of the water tank, three approaches were developed to tackle this problem. The curve fitting approach does not make any assumptions about the form of the distribution. This method, however, results in an upper bound on the required water level, because a subset of the data is used and not enough data are available and also because there are errors in the data. The normality approach uses all data to find an estimate for the distribution of the total water usage. This also reduces the effect of the errors in the data. The final approach (the binomial approach) uses extra assumptions on the water usage per passenger. The results from the different approaches are summarized in Table 3. In conclusion we recommend the normality approach, where the curve fitting approach is used as an upper bound. The binomial approach can be used to get a better understanding what the outcome will look like when parameters change. The regression line approach, which is currently used by KLM, looks very similar to the findings of the normality approach. The normality approach is based on a more general foundation and the method uses a better understanding of the service level. Hence, this method is prefered by KLM as well.

water level	the thresholds for the water level (number of passengers)			
(litres)	KLM approach	curve fitting approach	normality approach	binomial approach
151	-	1 - 13	-	1 - 56
302	-	14 - 27	-	57 - 122
453	-	28 - 41	-	123 - 191
604	0 - 59	42 - 55	0 - 118	192 - 260
754	60 - 281	56 - 256	119 - 294	261 - 294
905	282 - 294	257 - 294	-	-

Table 3: The ranges of the water level for which the Quality of Service is granted, for each approach applied to the AMS-BKK flight with the 74E aircraft.

The normality approach could however be improved. As shown in Figure 8, the variance of the total water consumption on a flight does not grow linearly with the number of passengers. Therefore, other forms of relationships should be explored as well besides linearity. The numerical example of the normality approach showed that μ_0 and σ_0^2 play an important role. Therefore, these parameters should be modelled in more detail. At last, based on the results of the statistical tests in Table 2, the effect of a logistic distribution for the total water consumption on a flight should also be investigated closer.

The numerical examples are performed on a few flight legs. Therefore, we recommend to do the same calculations on several more flights. Larger data sets have to be checked also for a more complete overview of the validity of the assumptions made in the different approaches.

References

[1] De Gunst, M.C.M., Van der Vaart, A.W., 2001, *Statistische Data Analyse*, lecture notes, Faculty Exact Sciences, vrije Universiteit amsterdam

[2] Feller, W., 1970, *An introduction to Probability Theory and its Applications, Volume II, Second Edition*, Wiley

[3] Mas-Colell, A., Whinston, M.D., Green, J.R., 1995, *Microeconomic Theory (First Edition)*, Oxford University Press

[4] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., 1988, *Numerical Recipes in C*, Cambridge Univ. Press, New York

[5] Oosterhoff, J., Van der Vaart, A.W., 2000, *Algemene Statistiek*, lecture notes, Faculty Exact Sciences, vrije Universiteit amsterdam

[6] Ross, S.M., 2003, *Introduction to Probability Models (Eighth Edition)*, Academic Press, San Diego

[7] Vardeman, S. and Lee, C.S., January 2005, Likelihood-Based Statistical Estimation from Quantized Data, working paper, to appear in *IEEE Transactions on Instrumentation and Measurement*, http://www.public.iastate.edu/∼vardeman/homepage.html