Forecasting Airport Passenger Flow to Improve Cleanliness Perception in Restrooms

Dominique S. Weijers

University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands

d.s.weijers@student.utwente.nl

ABSTRACT

Passenger flow forecasting is essential for identifying bottlenecks, dynamic planning and maximizing customer satisfaction. Even though this essence is clear, it remains unknown whether crowdedness truly results in lower satisfaction and whether forecasting techniques are feasible in an airport restroom context. These research topics are addressed by analysing the performance of three statistical techniques on the correlation between crowdedness and cleanliness perception, and by evaluating six regression models to forecast the number of restroom visitors at Schiphol airport respectively. The real dataset from Schiphol and Asito includes 887.822 FeedbackNow votes cast in 87 restrooms over a period of 7 months, flight information of 3.018 airplanes that arrived at Pier E, and the corresponding 97.521 passengers who visited one of the two restrooms on the arrival floor. Two of the three statistical models confirm the hypothesis that crowdedness results in a negative perception. Moreover, Ridge regression is able to predict the number of restroom visitors quite successfully ($R^2 = 0.83$). It is concluded that while the forecasting method is almost advanced enough to be used in practice, the correlation hypothesis needs further analysis before complete confirmation.

Keywords

Customer Satisfaction, Cleanliness Perception, Forecasting, Machine Learning, Prediction, Regression, Passenger Flow

1. INTRODUCTION

Similar to phones, watches and cities, airports are becoming smarter by the day. Also Schiphol Airport, the third greatest European airport in terms of market share, is investing millions to become the smartest airport on the globe. Their purpose is not only to increase revenue and profit but also to increase the satisfaction of the millions of passengers that come through every year. As this satisfaction is highly dependent on the smoothness of the passenger flow and the cleanliness of airport facilities, it is imperative to research this topic.

The cleaning company Asito is amongst others responsible for cleaning the restrooms at the airport. This cleanliness is measured using FeedbackNow; a system that is able to track customer satisfaction trends by comparing the number of green, yellow and red votes cast in each restroom. At this moment in time, Asito does not utilize any tools to aid their cleaning schedule and continues to use the same static and unoptimized planning mechanics as ten years ago. The FeedbackNow system has revealed that the overall average scores are sufficient, but both Asito and Schiphol want to improve. Long-time cleaning staff members have noticed that the restrooms with high traffic also seem to be the ones with the lowest scores. The hypothesis here is that crowdedness negatively impacts customer satisfaction and thus, the FeedbackNow ratings as well. If these spikes in crowdedness can be predicted, the levels of crowdedness can be mitigated seamlessly and most importantly, the cleaning staff is able to match their cleaning tasks accordingly. Moreover, as the number of restroom visitors after each cleaning intervention slightly lowers the actual cleanliness of the location as well, it is a very useful metric for Asito to plan cleaning activities at the ideal moment in time (not too early, but definitely not too late).

The purpose of this paper is therefore two-fold. Firstly, its objective is to test the hypothesis of a correlation between restroom crowdedness and the cleanliness perception of restroom visitors using three statistical techniques. Secondly, a regression technique is proposed to forecast the passenger flow to airport restrooms. This is done by estimating the number of passengers per flight using the flight schedule of arriving flights and additional features such as time and arrival gate. The following research questions are answered in order to realize this purpose:

RQ1 To what extent does crowdedness correlate with the cleanliness perception of restroom visitors?

RQ2 To what extent can the number of restroom visitors be forecasted?

All in all, these two results aid in the transition towards smart airports and public transport spaces altogether by providing a method to increase customer satisfaction and cleanliness perception, as well as a way to minimize waiting times and bottlenecks.

2. RELATED WORKS

Numerous theories and methods relevant to both research questions have been published. However, these are primarily focused on different contexts and utilize differently structured datasets. Besides, neither research question has been answered.

Regarding RQ1, research on the relationship between cleanliness and the presence of others exists to some extent, but as of yet, no experiment was ever conducted in an airport environment, nor in a restroom setting. Several works have investigated what stimuli impact people's cleanliness perception. Apart from stimuli such as scent, actual cleanliness and the condition of the environment, several studies conclude that disorder and the presence of others often have a negative impact on the perceived cleanliness [17].

31st Twente Student Conference on IT, Jul. 5th, 2019, Enschede, The Netherlands.

Cleanliness in an airport restroom setting is not only important because people exposed to unclean places tend to have a less positive attitude towards that place [14], Lee and Kim (2014) concluded that positive cleanliness perceptions influence the willingness of customers to spend more money in service contexts such as airports [7]. Furthermore, the density-intensity hypothesis of Freedman (1975) states that emotions and perceptions are reinforced and amplified when it is crowded [3]. The application of this theory in the context suggests that a negative cleanliness perception results in an intensified negative perception when it is crowded. Moreover, crowdedness from human density as such can certainly decrease customer satisfaction as well [10].

This paper examines whether these theories hold up in this specific context, which is relevant for obtaining new insights into how to maximize customer satisfaction in facility management.

In regards to RQ2, the predictability of the number of passengers in contexts other than the airport have been researched by many. Zhao et al. (2011) have focused on predicting the number of passengers on a bus line in China for approximately 240 departures [18]. Another research on a bus station in China with over 11.996.975 inbound and outbound buses, aggregated to hourly intervals, tried building a neural network to predict the passenger flow [8]. Both studies conclude that a non-linear model approach is effective with minimal prediction error.

Manataki & Zografos (2009) have created a model to approximate the number of passengers coming in and out of an airport. The research, conducted at Athens airport, has shown that peaks occur roughly at similar times each day and that the usage of the model is able to predict when bottlenecks will occur at the passport control based on incoming transit [11].

State-of-the-art simulation models like these are very data-intensive, complex to compute, and still yield mediocre predictions in reality. As the actual landing time of flights quite liable to change, the model should be able to calculate the number of real-time flights and their respective passengers continuously. This paper will be the first to examine the extent to which the number of restroom visitors can be forecasted using relatively easily computable regression models and publicly available flight schedule data. It will do this by building upon the aforementioned model by Manataki & Zografos (2019).

3. METHODOLOGY

In order to answer both research questions successfully and independently, the method and results are split into two sections.

3.1 Research Question 1

3.1.1 Data Collection

To answer RQ1, FeedbackNow (*FBN*) data was collected as an indicator of the cleanliness perception of restroom visitors at Schiphol airport. FBN is a business solution offering real-time client satisfaction data to aid in delivering excellent service at all times. Each installed FBN device poses the question: "how do you rate the cleanliness of this toilet today?" Forrester, the owner of the customer experience evaluation system, claims research has shown that respondents truthfully answer the question, and thus, that other factors such as mood are negligible.

In total, 887.822 votes were cast between October 3rd 2018 and April 30th 2019. These votes were emitted from 44 different toilet groups in 8 different sections of the airport. Almost every toilet group is composed of male and female restrooms (87 in total). Each vote is represented in the data with the location of the specific restroom, the distinct timestamp and whether the vote was a red, yellow or green smiley-face.

3.1.2 Data Processing

The files containing data on all restrooms were loaded in *PyCharm* using the *Pandas* library. Using the *Python* programming language, all votes of a restroom were split into sections with a specific timeframe t, measured in minutes. For example, if t = 5, the 7 months are first split into 60.480 sections of 5-minute intervals. Then, these sections were aggregated by the number of votes within this specific timeframe. Finally, it is evaluated how many votes were cast within each aggregated set and what percentage of these votes is green and red. This was repeated for all 87 restrooms.

The hypothesis is that the more votes are cast within an interval, the lower the percentage of green votes is and the higher the percentage of red votes is. Moreover, it is expected that the smaller the interval is, the more compelling this relationship will be. The underlining assumption here is that the number of clicks within a short timeframe is a proxy for crowdedness.

3.1.3 Statistical Techniques

After conferring with three statisticians, three techniques were proposed to test the hypothesis because this real-life data has no perfect mathematical model. The *SciPy* library was used to apply the methods to the stored data. The three techniques are explained and their pros and cons are discussed in this section. Before any technique could be applied however, a normal approximation was applied to each of the sets aggregated by the number of votes per interval t (G_x, R_x) where x = number of votes cast within t. In the following approximation formula for G_x, n_x = number of total votes within the set and n_x^G = number of green votes within the set:

$$\exists x\colon G_x \sim Binomial\left(N = n_x, p = \frac{n_x^G}{n_x}\right) \quad (1)$$

$$\exists x\colon G_x \sim Normal\left(\mu = n_x^G, \sigma = \sqrt{n_x^G\left(1 - \frac{n_x^G}{n_x}\right)}\right) \quad (2)$$

To provide an example, suppose that 7 votes are cast within 15 minutes with t = 5, of which 1 green vote is given in the first interval of 5 minutes, 3 red votes in the second 5-minute timeframe and 3 green votes in the last interval. Then, $G_1 \sim Bin\left(1, \frac{1}{1}\right)$, $G_3 \sim Bin\left(6, \frac{3}{6}\right)$, and $R_3 \sim Bin\left(6, \frac{3}{6}\right)$.

The first technique, the unpaired **two-sample t-test**, is used to determine if the means of two populations are equal or different. The technique, with a confidence interval of 95.0 percent, is used to validate whether the mean green percentage of an uncrowded number of votes (μ_u) is different from the mean green percentage of a crowded number of votes (μ_c). The distinction between crowded and uncrowded is made based on the assumption that it is crowded when $x > \lceil \frac{3}{10}t \rceil$ with t = time interval in minutes, x = number of votes cast within t, and the ceiling brackets indicating rounding to the upper integer. Using this method it is for instance expected that if within 1 minute, 2 or more votes are cast, it is crowded and if less than 18 votes are cast within 1 hour, it is considered uncrowded. Not only was this method applied to the percentage of green votes, it was also applied to the percentage of red votes. The null hypotheses for the red percentages (H_0^R) and green percentages (H_0^G) are tested against the alternative hypotheses for the red percentages (H_a^R) and green percentages (H_a^G) as shown in Equation 3 and Equation 4.

$$H_0^G: \mu_u = \mu_c \quad versus \quad H_a^G: \mu_u > \mu_c \tag{3}$$

$$H_0^R: \mu_u = \mu_c \quad versus \quad H_a^R: \mu_u < \mu_c \tag{4}$$

Unfortunately, the shortcoming of this method is that a comparison of only two samples can be made. This means that beforehand a distinction needs to be made on how many clicks per interval is considered crowded and uncrowded. This is a complicated task because of the limited amount of data available on the specifications and capacity of each restroom.

Second of all, the **Analysis of Variance** (*ANOVA*) is used to potentially mitigate the aforementioned issue as ANOVA generalizes the t-test beyond two means [1]. The mean of each of the aggregated sets is compared with the others. The null hypotheses (H_0^G, H_0^R) and the alternative hypotheses (H_a^G, H_a^R) for the green and red percentages are as follows:

$$H_0^G: \exists i,j : \mu_i = \mu_j \quad versus \quad H_a^G: \exists i,j : \mu_i \neq \mu_j \tag{5}$$

$$H_0^R: \exists i,j : \mu_i = \mu_j \quad versus \quad H_a^R: \exists i,j : \mu_i \neq \mu_j \tag{6}$$

The limitation of this method is that the alternative hypothesis does not indicate which group diverges from the others and it does not indicate whether there exists an upwards or downwards trend.

Therefore, using the third technique, it is examined whether the means of the calculated green and red percentages per interval follow a linear slope downwards and upwards respectively. For this, the male and female restrooms with more than 10.000 votes in the aforementioned period were identified and placed in separate .csv files. This was done as it was necessary to evaluate this data manually. Note that approximately 70.0 percent of all votes were given in these 28 restrooms.

As the data structure is quite unique, an adaptation of the Pearson correlation coefficient is proposed to test on the trend of the means. It is special because for the aggregated set for 1 vote cast within the timeframe, the percentage green can either be 0.0 percent or 100.0 percent. Hence, these sets are converted into mean values with a 95.0 percent confidence interval and sequentially weighted by the number of votes composing this mean. Then, the correlation (ρ) between the aggregated number of votes per time interval (x) and the percentage of green or red votes (y) is calculated using the **weighted Pearson correlation coefficient** (*WPCC*) shown in Equation 7 and Equation 8 [2, 9]. Note that ρ is computed twice: once for y = percentage of green votes and once for y = percentage of red votes.

$$cov(x,y;w) = \frac{\sum_i w_i (x_i - \frac{\sum_i w_i x_i}{\sum_i w_i})(y_i - \frac{\sum_i w_i y_i}{\sum_i w_i})}{\sum_i w_i} \tag{7}$$

$$\rho(x,y;w) = \frac{cov(x,y;w)}{\sqrt{cov(x,x;w)cov(y,y;w)}} \tag{8}$$

Nonetheless, this technique also has limitations: there is no method to calculate a weighted p-value and like the previous method, a decision needs to be made ahead of the calculation. The first limitation was handled by assuming significance for all results because of the high number of observations. To mitigate the risk of the latter limitation, it was decided to test using two alternative weights to come to a substantiated conclusion.

To confirm that a restroom follows the theory of the previously mentioned upwards or downwards trend, the following hypothesis (H^G, H^R) needs to be met for the ρ of the green (ρ^G) and red (ρ^R) percentages with t = time interval in minutes:

$$H^G: \rho^G_{t=60} \geq \frac{7}{10}\rho^G_{t=1} \leq -\frac{7}{20} \tag{9}$$

$$H^R: \rho^R_{t=60} \leq \frac{7}{10}\rho^R_{t=1} \geq \frac{7}{20} \tag{10}$$

In other words, the hypothesis is supported if the ρ with a short interval ($t = 1$) shows a strong correlation ($\rho \geq |0.5|$) and a substantially stronger one ($|\rho_{t=60}| \leq \frac{7}{10}|\rho_{t=1}|$ i.e. more than 30.0 percent) than for a long interval ($t = 60$). This hypothesis is analysed for each of the 28 restrooms manually and once for the 87 restrooms altogether. The latter is done by collecting the aggregated sets of votes with their respective green and red percentages for each of the restrooms and then combining these sets based on the number of votes cast within the interval (x).

Because each of the three statistical techniques has its own strengths and drawbacks, it is expected that the methods complement each other and will provide a thorough overview of the relation between the number of clicks per interval and the percentage of green and red votes.

3.2 Research Question 2

3.2.1 Data Collection

First of all, data from so-called people counters was collected to answer RQ2. Between January 1st 2018 and February 28th 2019, 7.850.460 restroom visitors have been registered by these sensors with promised 99.0 percent accuracy. The devices are installed in 14 distinct sections within the airport and in 19 toilet groups. For each of these restrooms, the number of visitors per hour is known. Second of all, data from the Schiphol API 4.0[1] was sourced. This data is comprised of information on all arriving flights between January 1st 2019 and April 30th 2019.

The overlapping period of interest will subsequently range from January 1st 2019 until February 28th 2019. The focus of RQ2 will be on Pier E because of the high amount of data available. This section of the airport accommodates 14 gates at which 3.018 airplanes have arrived in this period (of which 8 were excluded because the actual landing time was unregistered). Moreover, there are 19 toilet groups of which 2, located at the arrival floor, have people counters installed (see Figure 1).

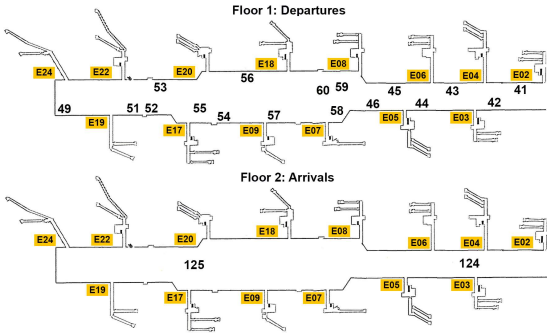


Figure 1. Map of Pier E with gates and toilet groups

[1] Shortly after the acceptance of the proposal, the previous version of the Schiphol API (version 3.0) was cancelled. This is the reason why all data from 2018 became unavailable.

3.2.2 Data Processing

To approximate the number of passengers present on a specific flight, an adapted version of the proposed model by Manataki et al. (2009) was used [11]. Instead of using it to estimate the number of passengers on departing flights, it will be used to approximate the number of passengers on arriving flights.

For the previously mentioned scope, 96.1 percent of the 2018 passenger load factors (*PLF*) of the operating airlines on Pier E was gathered. Moreover, 95.9 percent of the maximum aircraft capacity of airline specific airplanes (*AMAC*) was collected and 99.9 percent of the aircrafts in general (*MAC*). These percentages are not 100.0 percent because the PLF was not published by all airlines and the AMAC and MAC for some aircrafts were unavailable online. The accumulated data was used to estimate the number of passengers on a specific flight using the function visualised in Algorithm 1. The subfunctions supporting this algorithm can be found in Appendix A.1.

Algorithm 1: Passengers per Flight Calculation

Input: $flight$ `// flight arriving at pier E`
Output: pas `// expected number of passengers`
Function `calcPassengers` $(flight)$**:**
  $cap = \text{calcMaximumCapacity}(flight)$
  $plf = \text{calcPassengerLoadFactor}(flight)$
  return $cap * plf$
End Function

Because the label (number of restroom visitors *V*) is aggregated per hour, the feature (expected number of passengers) should be aggregated to the same interval. This is a difficult problem as it is hard to determine when passengers will actually visit the restroom and for how long if only the landing time (*LT*) is known. To approximate this, two parameters will be optimized. The first parameter, time to gate (*TTG*), is the number of minutes it takes for an arrived flight to taxi to the gate in addition to the time it takes for the first passenger to be registered at the restroom. The second parameter, overlap time (*OT*), is the time it takes from the TTG until the final restroom visitor was registered (see Figure 2). When identified, these parameters are used to calculate the expected number of passengers per hour (*E*(P)).

Figure 2. Visualisation of LT, TTG and OT

Two other features that will be used and evaluated are the day of the week (*D*) and the hour of the day (*H*). Because D and H are categorical values and most machine learning algorithms have interpretation difficulties with this data type, one-hot-encoding is applied before training the model.

Finally, every gate (*G*) at which airplanes arrive will be tested as features by designating the OT per hour for each gate. In essence, G represents the percentage of the hour in which passengers originating from that specific gate have visited the restroom. The enumerated features and labels are clarified in Table 1.

In total, 4 combinations of the features will be used to assess the performance of the algorithm and influence of the features when predicting the number of restroom visitors: C_1 = *E*(P); C_2 = *E*(P) and G; C_3 = *E*(P), D and H; and C_4 = *E*(P), D, H and G.

Table 1. Range and type overview of the features and label

	Variables	Abbr.	Type	Min	Max
Features (46)	Expected number of passengers per hour	*E(*P)	Float	0	∞
	Day of the week (7)	D	Binary	0	1
	Hour of the day (24)	H	Binary	0	1
	Gate usage (14)	G	Float	0	1
Label (1)	Number of restroom visitors per hour	V	Integer	0	∞

3.2.3 Regression Techniques

The *SciKit-Learn* library was utilized to train and evaluate the following six supervised regression models:

- Linear Regression (*LIN*): a widely used statistical technique to model the relationship of multiple variables. Simple linear regression will be used to predict the dependent variable V using the independent variable *E*(P). In addition, multiple linear regression will be tested using the other forenamed features [15].

- Ridge Regression (*RID*): another linear regression technique that suppresses multicollinearity, the phenomenon where near-linear relationships exist between the independent variables [12]. Such an interrelation can be a consequence of one-hot-encoding. RID might therefore be quite promising.

- Lasso Regression (*LAS*): a method with a very comparable approach to the previous technique. The main difference being the feature selection attribute, L1 regularization, which is absent in Ridge regression. RID uses L2 regularization, which is less radical as it never shrinks an independent variable's coefficient to zero [13]. However, this might be useful as it is probable that some of the 46 features lower the overall performance of the model.

- Support Vector Regression (*SVR*): an adaptation of the Support Vector Machine. Instead of a classifier, SVR is a nonparametric regression technique. In contrast to the techniques discussed above, SVR can be both linear and non-linear. This may be useful as comparable studies utilizing non-linear metrics have been quite successful [8, 18]. The overarching goal of this model is to determine the hyperplane between the boundary lines that fits the highest number of data points. Values outside of these boundaries are not considered [16]. If the margin between the hyperplane and boundaries is small, the model is more susceptible to overfitting.

- Gradient Boosting Regression (*GBR*): a technique that tries to minimize the loss of the model by turning weak-learners into strong-learners. The gradient descent procedure aids this process by diminishing the loss when adding trees. As it a quite greedy method, there is a risk of overfitting [4]. On the other hand, like SVR, GBR is able to discover polynomials and is highly tuneable in terms of hyperparameters.

- Multilayer Perceptron Regressor (*MLP*): a type of neural network that is completely feedforward. It has an input and output layer of nodes and at least one hidden node. MLP utilizes the backpropagation algorithm to train the data [5]. Limited memory BFGS (*LBFGS*) is the optimization algorithm used as the solver. Stochastic Gradient Descent and Adam are not used because LBFGS tends to perform better and converge faster on relatively small datasets.

To avoid overfitting, discover better variance estimates and prevent selection bias, shuffled *k*-fold cross-validation is used. The dataset is split into *k* (=10) equal sized subsets, of which *k*-1 is used for training and 1 is kept for validation. This process is repeated *k* times until for each subset it was attempted to predict the values based on the other subsets [6].

The six regression techniques are evaluated using uniform performance metrics. The coefficient of determination, better known as R-Squared (R^2), will be used as it provides good insight into the goodness of fit. It measures how well the actual outcomes compare to the ones produced by the model based on the explained variation. In addition, the Root Mean Squared Error (*RMSE*) is used to determine how much on average the predictions deviate from the actual values.

4. RESULTS

4.1 Research Question 1

Two of the three statistical techniques support the general hypothesis that states that the more votes are cast within an interval, the lower the percentage of green votes is and the higher the percentage of red votes is. These methods have also shown that this relationship is strengthened by decreasing the time interval.

The **two-sample t-test** method supports the general hypothesis for the green percentages as well as for the red percentages. This is because the number of restrooms that reject the null hypothesis is high for a short interval (e.g. 85.06 percent of the restrooms reject H_0^R for t = 1 minute) and low for a long interval (e.g. 23.08 percent of the restrooms reject H_0^R for t = 60 minutes).

In Table 2 this phenomenon is shown with H_a^R and H_a^G meaning the null hypothesis was rejected and the alternative hypothesis was accepted. Also, the restrooms that have never been crowded according to the split formula discussed in section 3.1.3 (e.g. 48 restrooms have never had more than 19 votes cast within 60 minutes), are excluded for the calculation of the percentages. Lastly, as it is not always the case that a restroom rejects the null hypothesis of both the red percentages and green percentages, the total reject percentage reflects the total percentage of restrooms that rejects the null hypothesis for at least one of the two.

Table 2. Results for the two-sample t-test hypotheses

t	Crowded if	Restrooms	H_a^R	H_a^G	$H_a^R \vee H_a^G$
1	$x \geq 2$	87	85.06 %	71.27 %	87.36 %
2	$x \geq 2$	87	82.76 %	64.38 %	83.91 %
5	$x \geq 3$	87	70.11 %	51.72 %	72.41 %
10	$x \geq 4$	87	58.62 %	49.43 %	58.62 %
15	$x \geq 6$	87	55.17 %	41.38 %	57.47 %
30	$x \geq 10$	65	26.15 %	16.92 %	32.31 %
60	$x \geq 19$	39	23.08 %	20.51 %	28.21 %

The **ANOVA** method also shows that the percentage of hypothesis rejects is high for small intervals and gets lower when this interval is enlarged (see Table 3). However, if a restroom rejects the null hypothesis (H_0^R or H_0^G) using the ANOVA technique, it signifies that the means of n_x are very unlikely to originate from the same distribution. The method is thus unable to determine whether the differentiating mean is one with a high number of votes by itself. Yet, the comparability with the results illustrated in Table 2 strongly suggests that this is the case. Therefore, this technique also supports the general hypothesis.

Table 3. Results for the ANOVA hypotheses

t	H_a^R	H_a^G	$H_a^R \vee H_a^G$
1	81.61 %	50.57 %	83.91 %
2	74.71 %	45.98 %	74.71 %
5	62.07 %	39.08 %	63.22 %
10	55.17 %	35.63 %	59.04 %
15	49.43 %	34.48 %	52.88 %
30	40.23 %	31.03 %	47.13 %
60	31.03 %	32.18 %	39.08 %

The **WPCC** method has amongst other things shown that 16 out of the 28 restrooms have confirmed the hypothesis for the red percentages using w_1 = total number of votes of the aggregated set. However, the alternative weight, $w_2 = \sqrt{w_1}$ provides different insights. Table 4 shows what percentage of the 28 analysed restrooms supports the hypotheses (H^R from Equation 10 and H^G from Equation 9) for the weights w_1 and w_2.

Table 4. Results for the WPCC hypotheses

Weight	H^R	H^G	$H^R \vee H^G$
w_1	57.14%	35.71%	57.14%
w_2	25.00%	32.14%	46.43%

Also, Table 5 illustrates correlation coefficients for the 87 restrooms altogether with the two weights. Here, it is clearly shown that for w_1 the coefficients of the red percentages ($\rho_t^R(w_1)$) as well as the coefficients of the green percentages ($\rho_t^G(w_1)$) show a decrease in strength in regards to the increase of t. However, w_2 does not show this at all, but suggests that there is no significant correlation for any t. In Appendix A.2, the identified mean red percentages with a confidence interval of 95.0 percent and w_1 for t = 5 are illustrated for a specific restroom. This section also shows the same figure for t = 60.

Overall, the WPCC method cannot completely confirm the general hypothesis that the means of the red and green percentages are linearly correlating with the number of clicks within a short interval.

Table 5. Results for the WPCC of all restrooms aggregation

t	$\rho_t^G(w_1)$	$\rho_t^G(w_2)$	$\rho_t^R(w_1)$	$\rho_t^R(w_2)$	Max x (n_x)
5	-0.718	-0.048	0.821	0.169	32
10	-0.667	-0.004	0.734	0.234	35
15	-0.620	-0.067	0.635	0.172	39
30	-0.540	-0.148	0.333	0.083	40
60	-0.421	-0.227	0.071	0.033	60

Even though w_1 suggests that the hypothesis is true because 16 out of the 28 restrooms confirm this individually and Table 5 shows a clear decrease and increase for the green and red percentages respectively, w_2 completely contradicts it. Even though w_1 sounds more reasonable and intuitive, the distribution of votes for small restrooms are highly skewed to the left ($x \leq 2$), which causes the aggregated means for 1 vote per interval and 2 votes per interval to be weighted too disproportionately. This phenomenon can be recognized in Appendix A.2 as well.

4.2 Research Question 2

Multiple regression models were trained and evaluated on two restrooms, using different features, and for different time periods. It is shown that by optimizing parameters and utilizing all available features, the ridge regression model is the most suitable method.

First, the parameters TTG and OT were optimized before training the regression models. This was done by attempting all combinations where TTG: 0-60 and OT: 0-60 for restroom 124, 125 and both combined, for January, February and both months.

Each of the 3.600 combinations was evaluated using two correlation techniques: the Pearson correlation coefficient (PCC) to test linear correlation and Spearman's rank correlation coefficient (*SCC*) to test monotonic correlation. Figure 3 shows the heatmap of the PCC for the number of restroom visitors for both restrooms in total ($V_{124,125}$) and E(P). The optimum correlation here is ρ = 0.88 (p-value = 0.0) for TTG = 9 minutes and OT = 37 minutes. The figure clearly shows that this optimum is not as precise since the values lying approximately on the green diagonal from (0, 30) to (60, 0) are also great for predicting the number of total visitors. This essentially proves that each flight has a slightly different TTG and OT, which makes sense as flights land at different landing strips.

Also note that the value of ρ increases again above the red diagonal from approximately (0, 60) to (60, 30) in the heatmap. This is due to the fact that cross-correlation is applied, which essentially enables the algorithm to find the time for which the correlation coefficient is the highest. Thus, when the combination of TTG and OT gets too high, the algorithm is better in predicting the values of an hour later than the current hour. This lag is compensated for in Figure 3 for values of TTG and OT above the aforementioned red diagonal, which causes the value of ρ to rise again instead of getting worse.

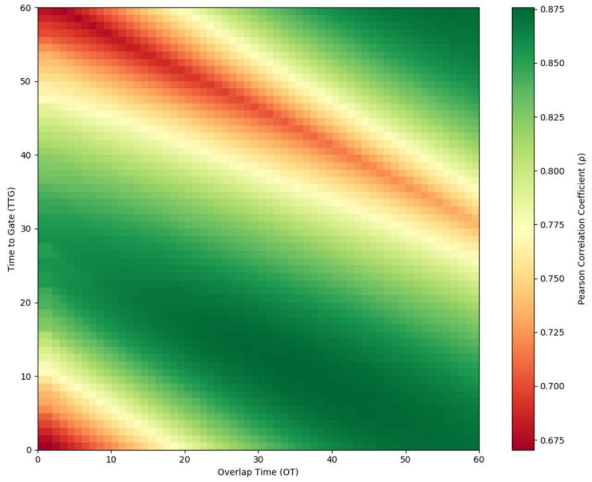


Figure 3. PCC (ρ) for $V_{124,125}$ and E(P)

Even though the SCC provided higher correlation coefficients (ρ = 0.92 with TTG = 3, OT = 43 and p-value = 0.0), it was decided to use the PCC for further usage as extensive visualization indicated that a linear relationship between the expected number of passengers and actual restroom visitors was most probable.

As the optimal calculation for E(P) is now determined, six regression methods were trained using cross-validation with k = 10. The feature E(P) was used to predict the label V for the full two months. This was done using multiple configurations (C_1, C_2. C_3, C_4) of the features.

The observed values for the performance of each of the models on restroom 124 are presented in Table 6. Note that these values are calculated using the default settings of the regression models, excluding the maximum number of iterations to make sure that each algorithm is able to converge.

Three conclusions can be drawn by analysing the tabulated data. First and foremost, all of the selected models perform almost indistinguishably. This is most likely due to the fact that the relationship between the features and number of restroom visitors is linear. Even after extensive hyperparameter tuning using grid search, all of the R^2 scores of the optimized models approximated to the R^2 score of LIN. The RID and LAS models were tweaked using the *alpha* parameter, which was set to 2.5 and 0.08 respectively. The MLP model's activation function for

Table 6. Cross-validated regression performance results of six models on three datasets (124, 125 and both)

RR*	Model	Combination 1*				Combination 2*				Combination 3*				Combination 4*			
		R^2	σ	RMSE	σ	R^2	σ	RMSE	σ	R^2	σ	RMSE	σ	R^2	σ	RMSE	σ
124	LIN	0.728	.04	29.06	12.2	0.735	.04	28.65	11.1	0.783	.02	25.91	9.4	0.791	.02	25.41	8.6
	LAS	0.728	.04	29.06	12.2	0.728	.04	29.05	12.2	0.740	.04	28.41	12.1	0.739	.04	28.43	12.0
	RID	**0.728**	**.04**	**29.06**	**12.2**	0.735	.04	28.64	11.1	0.783	.02	25.90	9.4	**0.791**	**.02**	**25.40**	**8.6**
	MLP	0.726	.04	29.16	12.4	0.726	.04	29.15	12.6	0.756	.02	27.53	10.2	0.771	.02	26.59	9.5
	GBT	0.711	.05	30.00	14.2	0.738	.03	28.50	11.0	0.778	.03	26.31	12.0	0.785	.02	25.84	9.7
	SVR	0.726	.04	29.17	12.4	0.734	.04	28.71	11.8	0.763	.03	27.09	10.9	0.775	.02	26.39	9.8
125	RID	**0.658**	**.03**	**13.47**	**5.0**	0.678	.04	13.07	5.2	0.712	.04	12.33	4.8	**0.728**	**.05**	**11.99**	**5.4**
Both	RID	**0.766**	**.03**	**35.56**	**14.8**	0.771	.03	35.10	13.8	0.821	.02	31.12	12.7	**0.828**	**.02**	**30.45**	**11.8**

* ***RR*** = *Restroom number;* ***Combination 1*** = *E*(P); ***Combination 2*** = *E*(P) and G; ***Combination 3*** = *E*(P), D and H; ***Combination 4*** = *E*(P), G, D and H

the hidden layer was set to *identity*, its learning rate schedule to *inverse scaling* and two hidden layers of 100 and 50 neurons were used. The GBT model was optimized by setting the *learning rate* to 0.1, the *number of estimators* to 200 and the *maximum number of the individual regression estimators* to 1.

As the scores of the models were also similar for restroom 125 and both restrooms at once, it was decided to visualise only the RID method. It performed at least as well as any other regression technique and it is relatively inexpensive to compute. The latter is an important criterion for the algorithm to be used in practice.

The second discovery is the fact that the R^2 scores for restroom 124 are higher than those of restroom 125, while the RMSE is significantly higher as well. The presumed argument for this phenomenon is the fact that the average number of restroom visitors for restroom 124 are higher than those of restroom 125; 75 and 23 visitors per hour during the day respectively.

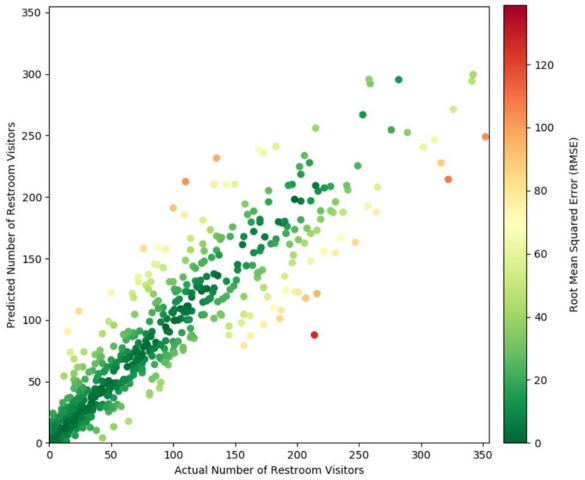


Figure 4. RMSE of RID predictions for $V_{124,125}$

Third of all, the values of the total number of restroom visitors in pier E can be predicted best. This is mainly due to the fact that the variance of the probability for passengers going to one of the restrooms is lower than the variance of the probability for passengers going to a specific one. The predictions for the restrooms' combined visitation numbers and the actual values are shown in Figure 4 with the colours indicating the RMSE.

Finally, to demonstrate the performance of the regression model on a day-to-day basis, Figure 5 shows the predictions and actual registrations of the total number of restroom visitors in Pier E. The figure shows that the predictions are very close overall, but some important peaks are not always fully forecasted.

5. DISCUSSION

To discuss the contribution of this research properly, this section is split into two distinct sections. Firstly, the pros and cons of the methodology are evaluated. Secondly, the usability of the machine learning algorithm in practice is assessed.

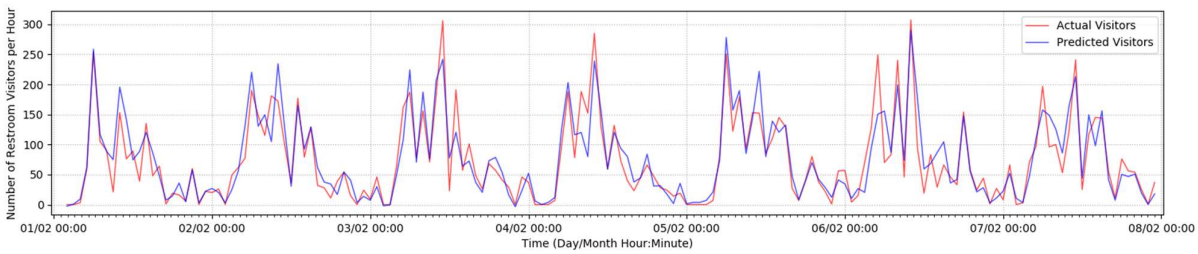
The method for answering RQ1 could be improved tremendously if more contextual information is gathered. The reason for this is that currently, the variables used for measuring the crowdedness and cleanliness perception are likely inadequate. The number of votes per interval was used as a proxy because the actual number of restroom visitors was unavailable on a relevant timescale in regards to measuring crowdedness. The weakness here is that the number of clicks can be dependent on a multitude of factors while the actual number of visitors is not.

Even though it seems quite reasonable that the clicks give a good indication of the number of visitors, there are two counters to this assumption. Firstly, most restrooms have a quite high variance in the number of votes. For example, by comparing the number of visitors per hour and the number of clicks per hour for restroom 125, contradictory observations can be made: for 30 visitors within an hour, the number of votes ranges between 0 and 30. Secondly, because customer feedback systems are a quite unexplored research topic, it is still uncertain whether respondents actually control their emotions and try their hardest to rate the cleanliness objectively as Forrester suggests. This research has shown that this assumption may be flawed as crowdedness does not directly impact the cleanliness. If this is the case however, it also means that other factors influence the number of clicks and scores. For example, it is possible that if someone has a negative experience, he or she is more likely to cast a (negative) vote. Additionally, perceived cleanliness may differ based on gender, nationality or other variables. If either is the case, the impact of the proxy for cleanliness perception needs further investigation as well.

In short, the results do suggest that the higher the number of votes within a small interval are, the lower the average scores tend to be. However, it is unclear if this phenomenon is similarly present if a high number of actual visitors within a small interval is taken as an input opposed to its proxy, the number of clicks. Therefore, a non-proxy for crowdedness is required to completely prove the hypothesis.

The method for answering RQ2 is hard to improve upon with the current dataset because almost all available features were exhausted and the real world remains unpredictable. The algorithm might improve when given extra predictors such as country of origin or flight duration which were available, but excluded because of potential ethical risks. Unavailable information such as landing strip, taxi time, offloading floor and the actual number of passengers on a flight will improve the algorithm without question. However, because of inevitable flaws in the real data and the unpredictability of life, it will sadly

Figure 5. Predicted $V_{124,125}$ and Actual $V_{124,125}$ per hour for February 1st – February 8th 2019



never be possible to predict the number of restroom visitors with perfect accuracy: sometimes one has to go, and sometimes not. This degree of uncertainty becomes apparent from Figure 4 and Figure 5 as the predicted number of passengers is almost always slightly higher or lower than in reality.

The ridge regression algorithm is able to approximate this high variance relatively well. Before Asito and Schiphol believe they can use it however, it is necessary to extend the dataset's timespan to at least a year, identify means to turn the unavailable data into usable data and generalize the method for all piers. Even though brief exploration into the performance of the algorithm on Pier F and G has yielded promising results, more elaborate testing is required. After these tweaks and improvements, the companies see potential in the applicability of this model in practice because of two main reasons. First of all, it is simple to compute, which allows for continuous updates to support dynamic planning of the cleaning activities. Second of all, knowing how many visitors have entered which restroom is not only very valuable information for optimizing cleaning activities, it is also beneficial in other contexts. Examples include optimized use of personnel in the airport stores, passport control and security, but also passenger safety in general.

6. CONCLUSIONS

In this paper, three statistical techniques were utilized to test the hypothesis stating that crowdedness negatively affected FeedbackNow scores, and six regression models were employed to predict the number of restroom visitors for toilet groups at Pier E. Collectively, the obtained insights enable Schiphol and Asito to improve overall customer satisfaction.

Two of the three statistical methods undoubtedly confirm the hypothesis that a high number of clicks within a short timeframe causes a lower cleanliness perception, while the other neither proved nor disproved it. It is concluded that future research is required to validate the legitimacy of the used proxies before complete confirmation of the general hypothesis.

After enriching the dataset with information about the approximate number of passengers on flights, optimizing the parameters *time to gate* and *overlap time* and applying hyperparameter tuning on the regression models, Ridge regression was selected as the best predictor for the number of restroom visitors (R^2 = 0.83). The proposed algorithm has an extremely short computation time which empowers both companies to maximize staff efficiency using real-time dynamic planning and allows for the incredible insight into the crowdedness of the restroom facilities. The latter is not only advantageous for increasing customer satisfaction within this context, but for many other environments inside and outside the airport as well. It is concluded that while the algorithm is already quite advanced, the time period over which the data was collected needs to be extended by at least one year, the algorithm needs thorough validation in more sections of the airport and lastly, methods to obtain currently unavailable data need to be developed before the algorithm can be applied in practice.

7. ACKNOWLEDGEMENTS

The author gratefully thanks dr. D. Bucur for her amazing input, feedback and general support during the research period, and prof. dr. N.V. Litvak for her expertise in statistics.

8. REFERENCES

[1] Carnegie Mellon University, Chapter 7: One-Way Anova. Retrieved June 16, 2019, from http://www.stat.cmu.edu/~hseltman/309/Book/chapter7.pdf.

[2] Costa, J. F. P. D. (2011). Weighted correlation. *International Encyclopedia of Statistical Science*, 1653-1655.

[3] Freedman, J. L. (1975). Crowding and behavior. WH Freedman.

[4] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1206.

[5] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15), 2627-2636.

[6] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

[7] Lee, S. Y., & Kim, J. H. (2014). Effects of servicescape on perceived service quality, satisfaction and behavioral outcomes in public service facilities. *Journal of Asian Architecture and Building Engineering*, 13(1), 125-131.

[8] Liu, L., & Chen, R. C. (2017). A novel passenger flow prediction model using deep learning methods. *Transportation Research Part C: Emerging Technologies*, 84, 74-91.

[9] Liu, Y., Meng, Q., Chen, R., Wang, J., Jiang, S., & Hu, Y. (2004). A new method to evaluate the similarity of chromatographic fingerprints: weighted Pearson product-moment correlation coefficient. *Journal of chromatographic science*, 42(10), 545-550.

[10] Machleit, K. A., Eroglu, S. A., & Mantel, S. P. (2000). Perceived retail crowding and shopping satisfaction: what modifies this relationship?. *Journal of consumer psychology*, 9(1), 29-42.

[11] Manataki, I. E., & Zografos, K. G. (2009). A generic system dynamics based tool for airport terminal performance analysis. *Transportation Research Part C: Emerging Technologies*, 17(4), 428-443.

[12] NCSS Statistical Software, Chapter 335: Ridge Regression. Retrieved June 15, 2019, from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf.

[13] Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning* (p. 78). ACM.

[14] Parker, C., Roper, S., & Medway, D. (2015). Back to basics in the marketing of place: the impact of litter upon place attitudes. *Journal of marketing management*, 31(9-10), 1090-1112.

[15] Pennsylvania State University, What is Simple Linear Regression? Retrieved June 18, 2019, from https://newonlinecourses.science.psu.edu/stat501/node/251.

[16] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3), 199-222.

[17] Vos, M. C., Galetzka, M., Mobach, M. P., Van Hagen, M., & Pruyn, A. T. (2018). Cleanliness unravelled: a review and integration of literature. *Journal of Facilities Management*, 16(4), 429-451.

[18] Zhao, S. Z., Ni, T. H., Wang, Y., & Gao, X. T. (2011). A new approach to the prediction of passenger flow in a transit system. *Computers & Mathematics with Applications*, 61(8), 1968-1974.

APPENDIX

A.1 Algorithms

The following pseudo-code algorithms illustrate how the passenger load factor (*PLF*) and maximum aircraft capacity (*CAP*) are calculated per flight. For the PLF it is important to note that it was assumed that charter flights are always completely booked and therefore the PLF is always 100.0 percent. Also, note that the average PLF factor is 80.54 percent and the average CAP is 244.52 seats.

Algorithm 2: Passenger Load Factor Calculation

```
Input: flight          // flight arriving at pier E
Output: plf    // passenger load factor of airline

Function calcPassengerLoadFactor(flight):
    // dictionary k:iataCode, v:plf
    plfDict ← open('/plf.p','rb')

    if flight['serviceType'] equals 'Passenger' then
        iataCode = flight['prefixIATA']
        if iataCode in plfDict then
            return plfDict[iataCode]
        else
            return plfDict['Average']
        end
    else if flight['serviceType'] equals 'Charter' then
        return 1
    else
        return 0
    end
End Function
```

Algorithm 3: Maximum Aircraft Capacity Calculation

```
Input: flight          // flight arriving at pier E
Output: cap            // capacity of aircraft type

Function calcMaximumCapacity(flight):
    // dictionary k:aircraftType, v:cap
    aircraftDict ← open('/aircraftCap.p','rb')

    // dictionary k:(iataCode, aircraft), v:cap
    airlineDict ← open('/airlineCap.p','rb')

    iataCode = flight['prefixIATA']
    aircraft = flight['aircraftType']
    if (iataCode, aircraft) in airlineDict then
        return airlineDict[(iataCode, aircraft)]
    else if aircraft in aircraftDict then
        return aircraftDict[aircraft]
    else
        return aircraftDict['Average']
    end
End Function
```

A.2 WPCC Figures

The figures below illustrate the average percentage of green and red votes based on the aggregation of the number of votes (x) cast in timeframe t. These votes were cast in the male restroom of toilet group 67 between October 1st 2018 and May 1st 2019. Each dot in the figure is the mean percentage of green votes based on the number of votes within the parenthesis on the x-axis. Figure 6 clearly shows that the more votes are cast within a short period of time (t = 5 minutes), the lower the percentage of green votes and the higher the percentage of red votes. Figure 7 shows a roughly straight line for both colours for t = 60 minutes. These figures confirm the general hypothesis. It has to be noted however, that not all restrooms have followed this pattern, as mentioned in section 4.1.

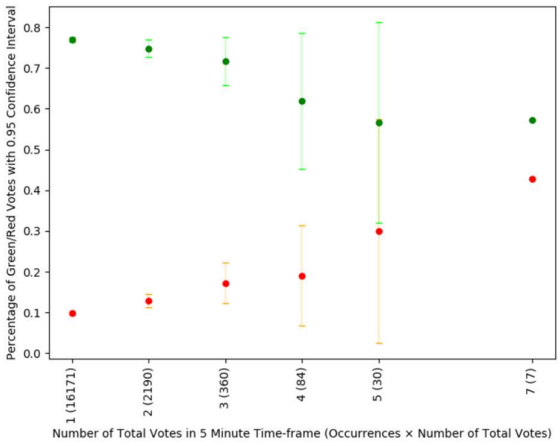


Figure 6. Average Percentage of Green/Red Votes for toilet group 67 Male with t = 5 minutes

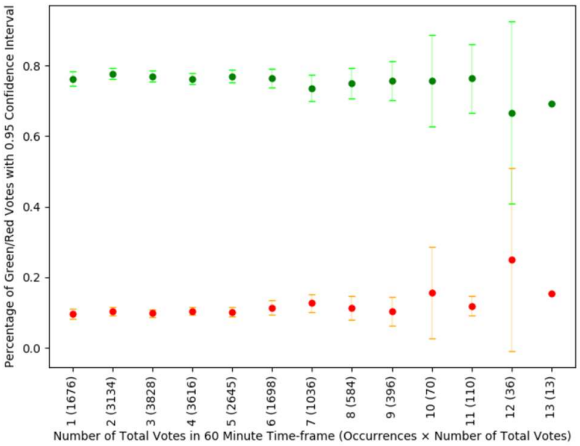


Figure 7. Average Percentage of Green/Red Votes for toilet group 67 Male with t = 60 minutes