Statistical analysis of wastewater monitoring for maximum peak factor estimation

Journal of Physics: Conference Series **1981** (2021) 012013 doi:10.1088/1742-6596/1981/1/012013

Statistical analysis of wastewater monitoring for maximum peak factor estimation

N J Cely-Calixto[1], **C A Bonilla-Granados**[1], **and J P Rojas-Suárez**[2]
[1] Grupo de Investigación en Hidrología y Recursos Hídricos, Universidad Francisco de Paula Santander, San José de Cúcuta, Colombia
[2] Grupo de Investigación en Infraestructura Vial, Universidad Francisco de Paula Santander, San José de Cúcuta, Colombia

E-mail: nelsonjaviercc@ufps.edu.co

Abstract. The design and operation of sanitary sewerage systems are based on the knowledge of peak and the average daily wastewater flows. The maximum peak factor is obtained through the ratio of the maximum flow and the average discharge flow of wastewater generated in a system. In this sense, the maximum peak factor was estimated by monitoring wastewater flow in an urban sector of the city of San José de Cúcuta, Colombia. This urban sector represents 45.6% of the sanitary sewerage of the city. Stochastic modeling of the peak flow was performed, and a mathematical model was constructed to estimate the maximum peak factor using a correlational study using statistical methodology. Through linear regression analysis, a model was obtained that estimates the value of the maximum peaking factor based on knowledge of the average daily wastewater flow. The results indicate that the model is potential, and the expression is statistically significant and satisfies the assumptions established for the classical linear regression model.

1. Introduction

The design of sanitary sewer systems and wastewater treatment plants requires knowledge of the maximum peak factor (F). This is obtained by the ratio of the peak flow and the average discharge flow of wastewater generated in a system. The determination of F involves the development of field studies, where monitoring and modeling of wastewater discharges from the study population are performed [1].

The characteristics and volume of wastewater discharge depend on each population or discharge source [2,3], so typically there are variations over time in the flow rate. Also, the hydraulic design of the sewerage system may depend on the different climatic scenarios that occur in the area, since wastewater receives contributions from rainwater [4]. Therefore, the flow rate should be considered as a random variable from a statistical point of view [5]. The estimation of F makes it possible to quantify the variability presented in wastewater flow rates and to calculate the maximum hourly flow rate, which is the basis for establishing the design flow rate of each section of a sanitary sewer network [6,7,8].

Several investigations have been carried out over time in different parts of the world on the determination of F, such as the study conducted by Tirado (2013), who determined the value of F in concrete and polyvinyl chloride (PVC) sanitary sewer pipes in Managua, Nicaragua. Using mathematical methods, such as the least-squares method and the exponential function, an equation was determined to estimate the "maximum flow coefficient" [9].

Another research is by Imam and Elnakar (2004) who determined the actual peaking factors after one year of daily flow monitoring in the West Golf community in New Cairo, Egypt. These results allowed estimating the maximum and minimum flow rates for the design of the wastewater treatment plant in that locality [10]. Also, there is the research conducted by Chandragade and Gupta (2016) who developed a study in the city of Gadchiroli, India. In this study, a new methodology was applied in which, the effective F is continuously reduced as the contributing population increases, and where the maximum flow rates for the design of sewer networks were estimated [11].

The above research is evidence of the relevance of the estimation of F in the understanding of hydraulic behavior and its application in the design of sanitary sewer systems. In places where there is no possibility to perform wastewater flow monitoring, empirical equations are used to predict F. Among the best known are those proposed by: Harmon, Babbit, Flores [9], Gaines [12], Federov, Utah, Los Angeles, and Tchobanoglous [13]. However, these equations have been derived from the characteristics of their context, therefore, the equations may be inappropriate for estimating F corresponding to locations and populations with different consumption patterns, leading to undesirable results for future designs [14,15].

In the city of San José de Cúcuta, Colombia, there is no information available on wastewater discharge and its variation over time; therefore, it is not possible to calculate F. In addition, no studies have been conducted in this city aimed at analyzing and understanding the hydraulic behavior of the sanitary sewer system. It is necessary to study the monitoring of the city's wastewater flow, which would allow estimating F and thus evaluate the performance of the existing sanitary sewer system. The objective of the research is the estimation of F through the monitoring of an urban sector of the city of San José de Cúcuta, Colombia.

2. Methodology

To develop the proposed research objective, a correlational study was proposed, which used statistical methods for the construction of a mathematical model to relate the variables F and average daily flow. For the development of the research, a sector located within the urban perimeter of the city was selected, which represents 45.6% of the sanitary sewerage system. Within the sector, the communities 1, 2, 5, 7 and 9, are located in the city of San José de Cúcuta, Colombia, and correspond to an approximate area of 2745 hectares.

The information was processed and analyzed using R statistical software, version 3.5.0. The statistical methodology used included the following stages: adjustment of the random variables to a probability distribution; goodness-of-fit; and construction of a linear regression model.

2.1. Fitting of random variables to a probability distribution

Given a continuous random variable with observed values X_1, X_2, ..., X_n, it is desired to find a probability density function $f(x;\theta)$, which allows representing these values in the best possible way, where is a vector of parameters, which are constant values that characterize the distribution and must be estimated [16]. For the experimental data corresponding to the random variables maximum daily flow and maximum flow factor, the following steps were developed: 1) select appropriate probability distributions for the problem under study; 2) Estimate the parameters of the distributions, and 3) develop statistical well-fit tests.

Two probability distributions were considered: Gumbel and Log-Gamma (log-Pearson type III) because they allow modeling extreme events and are widely used in the determination of peak flows, rainfall, and other hydrology applications [10,17]. The Gumbel cumulative probability distribution function is represented by Equation (1).

$$F(x) = P(X \leq x) = e^{-e^{-\alpha(x-\beta)}}, \quad (1)$$

where is α the shape parameter and β is the scale parameter. On the other hand, the log-Gamma distribution (log-Pearson Type III) presents the following probability density function written in its biparametric form (Equation (2)) [18].

$$f(x; k, \theta) = x^{k-1} \frac{e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}, \ x > 0, \ y \ \ k, \theta > 0, \tag{2}$$

where k is the shape parameter and θ is the scale parameter. For the estimation of the parameters of a distribution, the best-known alternatives in statistical theory are the method of moments and the maximum likelihood method. For the present study, the R software incorporates the maximum plausibility method, which estimates as parameters of the distribution those values that maximize the probability of the observed sample [19]. The method constructs a plausibility function that is equal to the product of the values of the probability density function evaluated for each of the sample values (Equation (3)).

$$L = \prod_{i=1}^{n} f(x_i) = f(x_1) * f(x_2) * \ldots * f(x_n). \tag{3}$$

The larger the value of L is, the better the fit of the data to the chosen probability distribution. For this reason, the likelihood function is maximized, and the estimated values of the parameters are determined.

2.2. Goodness-of-fit test

Once the parameters of the distributions have been estimated, the goodness-of-fit test is determined. Goodness-of-fit test make it possible to determine whether a set of empirical data corresponds to a selected probability distribution [19], *i.e.*, it is checked whether the fitted probabilistic models are statistically valid for the data they are intended to model. In addition, if there is more than one probability distribution that fits the data, the goodness-of-fit test indicates which of these best fits the case study.

In the case of continuous distributions, one of the most commonly used criteria is the Kolmogorov-Smirnov criterion, which determines the maximum difference between the theoretical probabilistic model and the empirical data [20]. The procedure consists of a hypothesis test as follows:

- The null and alternative hypotheses are stated: - Null hypothesis (H0): the empirical data conform to the distribution function F(x) (Gumbel, log-Gamma, etc.) - Alternative hypothesis (H1): The empirical data do not conform to the distribution function F(x) (Gumbel, log-Gamma, etc).
- The empirical data are ordered in order of the null and alternative hypotheses: - Null hypothesis (H0): the empirical data do not conform to the distribution function F(x) (Gumbel, log-Gamma, etc).
- The data are ordered from smallest to largest and the cumulative observed frequencies Fn(x) is calculated. The cumulative theoretical frequency is calculated through the probability distribution function (Gumbel, log-Gamma, etc).
- The test statistic D, which consists of the supremum of the differences between the two cumulative frequencies, is determined by Equation (4).

$$D = \text{Sup} \, |F_n(x) - F(x)|. \tag{4}$$

The exact probability distribution of the test statistic, under the null hypothesis, is found in statistical packages. The R software delivers the p-value which can be defined as the exact probability of making a type I error, *i.e.* rejecting a true null hypothesis. Thus, it can be concluded that the higher the p-value, the better the fit performed and therefore the null hypothesis should not be rejected. In practice if the p-value > 0.05 the fit is considered to be valid.

2.3. Construction of a linear regression model

The purpose of the analysis is to construct a mathematical model that allows estimating or predicting the average value of the dependent variable F, based on the known or determined values of the explanatory variable average daily flow rate. From a geometrical point of view, the strategy consists of finding a curve that represents the general trend of the data. In this sense, the choice of the functional form of the mathematical model (linear, polynomial, exponential, potential, etc.) is of great importance in regression analysis.

For the development of the regression analysis, the original database is used, which is made up of 52 measurements. Different values of the daily average flows can generate the same value of the F factor, due to the oscillations of the maximum flow. For this reason, the data were ordered from lowest to highest, taking the F factor as a reference, and the average flow rate was calculated for the repeated values. Thus, the original database is reduced to 14 points.

3. Results

After applying the statistical methodology proposed for each of the stages: adjustment of the random variables to a probability distribution; Goodness-of-fit tests; and construction of a linear regression model. The following results were obtained.

3.1. Fitting of random variables to a probability distribution

The database containing the random variables daily maximum flow rate and the F-factor corresponding to the 52 days of observation were imported into the R software. To make the fits to the Gumbel and log-Gamma probability distributions, the additional packages VGAM and MASS must be loaded. Table 1 shows the results obtained in fitting the data to the Gumbel and log-Gamma (log-Pearson Type III) probability distributions using R software.

Table 1. Estimated parameters Gumbel distributions Gumbel y log-Gamma.

Variable	Gumbel distribution		Log-Gamma distribution	
	α	β	k	θ
Maximum daily flow rate (QMD)	2191.2177	123.3163	18924.5023	2451.5260
Maximum peak factor (F)	1.40084991	0.04560747	103.58081000	292.83746000

3.2. Goodness-of-fit tests

The following results are obtained with the R software for the adjustments of the variable "maximum daily flow rate". For the test "MAXIMUM, "pgumbel", location.gumbel, scale.gumbel" D = 0.14984 and p-value = 0.1934 (alternative hypothesis: two-sided); and for the test "log (MAXIMUM), "pgamma", shape.loggamma,1/scale.loggamma" D = 0.13341 and p-value = 0.3129 (alternative hypothesis: two-sided). In both cases, the p-value obtained indicates that the maximum flow data fit well to the Gumbel and log-Gamma distributions, being the latter distribution the one that presents the best fit, this is concluded since the p-value in both cases is greater than 0.05.

As for the variable "Maximum peaking factor F", it was obtained for the test "F, "pgumbel", location.gumbel, scale.gumbel" that D = 0.1145, p-value = 0. 503 (alternative hypothesis: two-sided); and for the test "log (F), "pgamma", shape.loggamma, 1/scale.loggamma" it was obtained that D = 0.089261, p-value = 0.8017 (alternative hypothesis: two-sided). And as in the previous case, the results allow to conclude that the F-factor data fit well to the two distributions, especially to the log-Gamma, where the p-value obtained (0.8017) indicates that the fit is very good, this is concluded since the p-value in both cases is greater than 0.05.

3.3. Linear regression analysis

The scatter graph is generated to have an approximation of the functional regression model using R software; Figure 1 shows that the scatter of the data does not follow a pattern. The underlying theory in the case of the maximum factors indicates a potential model, which seems to be corroborated by the

scatter plot. To fit a potential (log-log) model, the vectors containing the variables corresponding to the average flow rate and F-factor are transformed by calculating the respective natural logarithms. A linear model is then fitted to the transformed data.

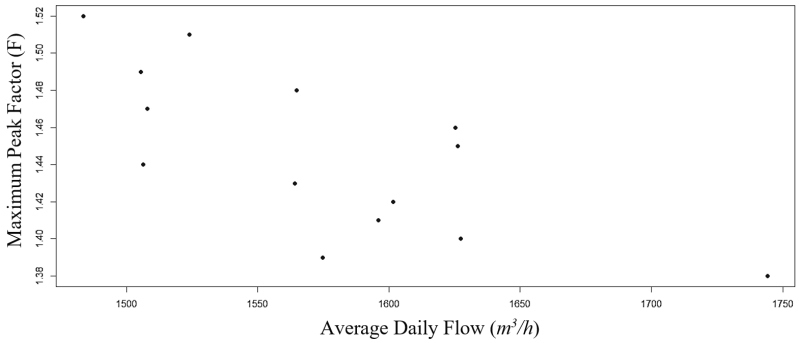

Figure 1. Scatter graph of the data.

Table 2 and Table 3 show that both the coefficients and the regression model are statistically significant due to the magnitude of the p-values obtained. The coefficient of determination (R^2) is a measure of the goodness of fit of the regression model; in this case, the model constructed explains an acceptable percentage (52%) of the total observed variability. The residual standard error indicates that on average any prediction made by the model is 0.02197 units away from the true value.

Table 2. Residuals.

Min.	1Q	Median	3Q	Max.
-0.039053	-0.014955	0.000213	0.020073	0.027197

Table 3. Coefficients.

	Estimate	Std. error	t value	Pr(>|t|)
(Intercept)	4.0882	1.0309	3.965	0.00187 **
LNPROMEDIO	-0.5053	0.1400	-3.608	0.00359 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The interpretation of the regression coefficient of the explanatory variable LNPROMEDIO indicates that for each unit increase in the logarithm of the average flow rate, the response variable logarithm of the F factor (LNF) decreases 0.5053 units. In regression analysis, the interpretation of the intercept term is generally meaningless, as in this case where the range of values for the average flow rate does not include the value of zero. The estimated linear regression model is represented in Equation (5).

$$F^* = 4.0882 - 0.5053\text{QMD}, \tag{5}$$

where QMD is the average daily flow, F* y QMD* are the natural logarithms of the F factor and the average daily flow are, respectively. Transforming this equation to the potential model yields Equation (6) (QMD in units of m^3/h) and Equation (7) (QMD in units of L/s).

$$F = \frac{59.632}{\text{QMD}^{0.5053}}, \tag{6}$$

$$F = \frac{531.215}{\text{QMD}^{0.5053}}. \tag{7}$$

Table 4 of the analysis of variance confirms the overall significance of the regression model when analyzing the sums of squares and the p-value obtained (0.003593).

Table 4. analysis of variances.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
LNPROMEDIO	1	0.0062804	0.0062804	13.017	0.003593**
Residuals	12	0.0057896	0.0004825	-	-

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The model validation involves verifying the assumptions of the classical linear regression model by analyzing the residuals. In particular, it is verified whether the residuals conform to the normal probability distribution, have zero mean and constant variance.

Figure 2 of the analysis of the residuals against the values fitted by the model does not allow us to appreciate the presence of a specific pattern in the errors obtained for the regression. A random variation is observed concerning the mean of zero (3.710004e-19). There is also no evidence that the variance of the errors increases progressively with the values adjusted for the dependent variable, therefore, it can be concluded that there is no evidence of heteroscedasticity, *i.e.*, the variance of the residuals remains.

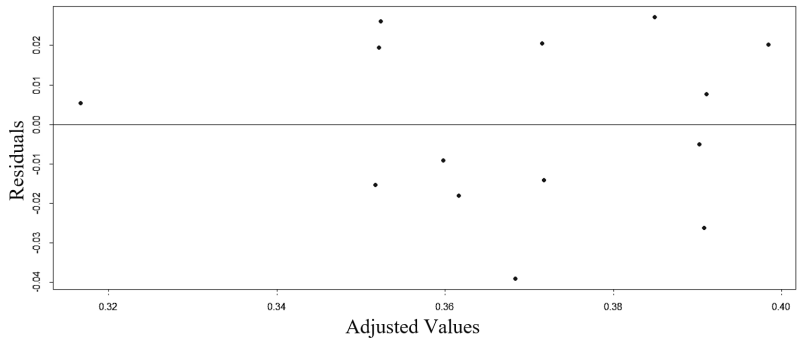

Figure 2. Analysis of residuals vs. R fitted values.

Figure 3 of normal quantiles shows no apparent deviations from normality. The Shapiro-Wilks test confirms the hypothesis of normality of the residuals: W = 0.93623, p-value = 0.3722, therefore, according to the p-value obtained, the null hypothesis of normality is not rejected.

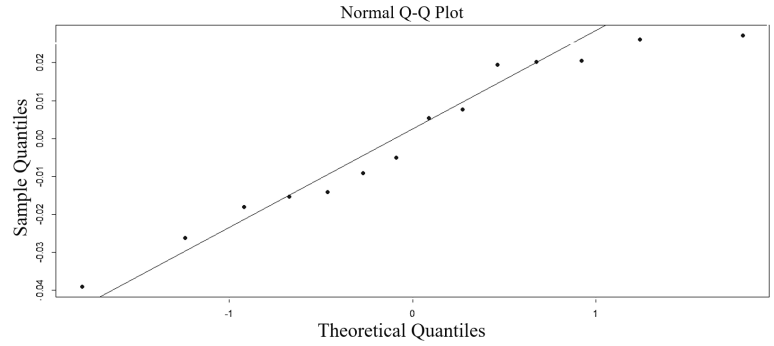

Figure 3. Normal quantile graph.

An alternative graphical analysis consists of plotting the standardized residuals against the fitted values to identify outliers, which become points of high influence that decrease the goodness of fit of the regression model constructed. A residual is considered an outlier if its absolute value exceeds the value of 3 because the probability that a value of the standard normal distribution exceeds this limit is very low. The graphical analysis confirms that there are no outliers or influential values. It is concluded that the linear regression model is represented by Equation (5), Equation (6), and Equation (7). The expression is statistically significant and satisfies the assumptions established for the classical linear regression model.

From the linear regression model, 95% confidence intervals can be constructed for the mean response, *i.e.*, the value of the conditional mean of F given a value of the average daily flow E(F/QMD) can be predicted. The point estimator of F is the value calculated by the regression model and inferential statistical theory is used to construct the confidence interval.

4. Conclusions

Statistical methods develop procedures that allow modeling situations to occur in reality, processes such as wastewater flow, where variation, random behavior, and system complexity are presented. Through linear regression analysis, a model was obtained that estimates the value of the maximum peak factor from the knowledge of the average daily flow of wastewater belonging to a sector of the city of San José de Cúcuta, Colombia. This information is very valuable for understanding the functioning of the city's sewage system and can eventually be used in the design, planning, construction, and operation of new projects.

Acknowledgments

The authors thank to the "Fondo de Investigaciones Universitarias (FINU)" at the Universidad Francisco de Paula Santander for the financial support of this work.

References

[1] Rodríguez J 2013 Monitoring and modelling to support wastewater system management in developing mega-cities *Sci. Total Environ.* **445–446(19)** 79
[2] Shuvalov M 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **775** 012099:1
[3] Chernosvitov L 2019 *IOP Conf. Ser.: Earth Environ. Sci*. **272** 022253:1
[4] Azzam H Al-Hadithy 2019 *J. Phys. Conf. Ser.* **1294(1)** 052065:1
[5] Poensgen-Llano N 2011 *Análisis del Comportamiento Espacio-Temporal de los Caudales Residuales en la Cuenca Urbana del Río Salitre* (Colombia: Pontificia Universidad Javeriana)
[6] Empresas Publicas de Medellin (EPM) 2013 Normas de Diseño de Sistemas de Alcantarillado de EPM (Colombia: Empresas Públicas de Medellín)
[7] Touny M 2020 Peak factors for sewerage system in upper Egypt communities *Bull Fac. Eng. Mansoura Univ.* **45(1)** 16
[8] Zhang X 2005 A theoretical explanation for peaking factors *World Water and Environmental Resources Congress 2005* (New York: American Society of Civil Engineers)
[9] Tirado-Picado V 2013 Determinación del coeficiente de flujo máximo para el diseño de sistemas de alcantarillados sanitarios, evaluado en Managua, Nicaragua *Tecnura* **17(36)** 61
[10] Imam E, Elnakar H 2014 Design flow factors for sewerage systems in small arid communities *J. Adv. Res*. **5(5)** 537
[11] Chandragade A, Gupta R 2016 Peak factor curve for estimating peak flows in design of sewer networks: a case study of Gadchiroli city *Int. J. Innov. Res. Sci. Eng*. **2(12)** 162
[12] Gaines J 2015 Peak sewage and flow rate: prediction probability *Water Pollut. Control Fed.* **61(7)** 124
[13] Bonilla-Granados C 2019 *Determinación del Factor Máximo de Pico para el Diseño de Sistemas de Alcantarillado Sanitario Mediante Monitoreo de Flujo de Aguas Residuales Caso de Estudio: Cúcuta, Colombia* (Colombia: Universidad Manuela Beltrán)
[14] Balacco G 2017 Evaluation of peak water demand factors in puglia (Southern Italy) *Water* **9(2)** 96
[15] Bonilla-Granados C 2019 A systematic review of wastewater monitoring and its applications in urban drainage systems *Respuestas* **24(3)** 54
[16] Vitto R 2005 *Fitting Distributions With R* (Boston: Free Software Foundation)
[17] Kotz S, Nadarajah S 2000 *Extreme Value Distributions: Theory and Applications* (London: Imperial College Press)
[18] Sáenz A 2009 *Modelización Estocástica de Precipitaciones Máximas para el Cálculo de Eventos Extremos a partir de los Periodos de Retorno Mediante R* (Spain: Universidad de Jaén)
[19] Canavos G 1987 *Probabilidad y Estadística. Aplicaciones* (México: Editorial Mc Graw Hill)
[20] Díaz-Monroy L 2007 *Estadística Multivariada: Inferencia y Métodos* (Colombia: Universidad Nacional de Colombia)