Egodawatta, Prasanna, Haddad, Khaled, Rahman, Ataur, & Goonetilleke, Ashantha
(2014)
A Bayesian regression approach to assess uncertainty in pollutant wash-off modelling.
Science of the Total Environment, *479-480*, pp. 233-240.

A Bayesian regression approach to assess uncertainty in pollutant wash-off modelling

Prasanna Egodawatta[1], Khaled Haddad[2], Ataur Rahman[2], Ashantha Goonetilleke[1*]

[1]Science and Engineering Faculty, Queensland University of Technology, GPO Box 2434, Brisbane 4001, Australia

[2]School of Computing, Engineering and Mathematics, University of Western Sydney, Building XB, Locked Bag 1797, Penrith, NSW 2751, Australia

[*]Corresponding author

Email: a.goonetilleke@qut.edu.au

Tel: +61 731381539

Fax: +61(0)731381170

A Bayesian regression approach to assess uncertainty in pollutant wash-off modelling

Abstract:
Due to knowledge gaps in relation to urban stormwater quality processes, an in-depth understanding of model uncertainty can enhance decision making. Uncertainty in stormwater quality models can originate from a range of sources such as the complexity of urban rainfall-runoff-stormwater pollutant processes and the paucity of observed data. Unfortunately, studies relating to epistemic uncertainty, which arises from the simplification of reality are limited and often deemed mostly unquantifiable. This paper presents a statistical modelling framework for ascertaining epistemic uncertainty associated with pollutant wash-off under a regression modelling paradigm using Ordinary Least Squares Regression (OLSR) and Weighted Least Squares Regression (WLSR) methods with a Bayesian/Gibbs sampling statistical approach. The study results confirmed that WLSR assuming probability distributed data provides more realistic uncertainty estimates of the observed and predicted wash-off values compared to OLSR modelling. It was also noted that the Bayesian/Gibbs sampling approach is superior compared to the most commonly adopted classical statistical and deterministic approaches commonly used in water quality modelling. The study outcomes confirmed that the predication error associated with wash-off replication is relatively higher due to limited data availability. The uncertainty analysis also highlighted the variability of the wash-off modelling coefficient *k* as a function of complex physical processes, which is primarily influenced by surface characteristics and rainfall intensity.

Keywords: model uncertainty; stormwater quality; pollutant wash-off; Bayesian analysis; Monte Carlo simulation; stormwater pollutant processes

1. Introduction

Stormwater quality modelling is based on the mathematical replication of fundamental stormwater pollutant processes, namely, pollutant build-up and wash-off. The mathematical replication of these processes involves simplification of reality dictated by the limitations of the knowledge base. This results in uncertainty in the overall modelling approach and consequently, the outcomes derived. Typically, the sources of uncertainty in relation to computer models can be categorised as structural, epistemic, parametric and experimental. The structural and epistemic uncertainties arise from the lack of knowledge of underlying physical processes and simplification of known scientific knowledge in modelling practices, respectively. These are mostly associated with conceptual model development and mathematical replication. Parametric uncertainties arise from the use of unknown or lumped parameters to represent reality while experimental uncertainty is associated with the use of measured data in modelling approaches including uncertainty associated with sampling procedure (Merz and Thieken, 2005; Oberkampf et al., 2002). Both, parametric and experimental uncertainties arise during simulations and the generation of numerical solutions (Dotto et al., 2014).

Knowledge of model uncertainty is an essential element in informed decision making (Bertrand-Krajewski et al., 2002). In particular, an understanding of uncertainty in stormwater quality modelling outcomes is important for effective stormwater treatment system design, receiving water impact assessment and for evaluating model reliability (Freni et al., 2009). A range of studies have focused on assessing uncertainty in stormwater quality modelling approaches. For example, Bertrand-Krajewski et al. (2007), Liu et al. (2012a, b) and Sohrabi et al. (2003) have focused on parametric and/or experimental uncertainties associated with stormwater quality modelling while Dotto et al. (2012), Freni et al. (2008)

and Kanso et al. (2005) have focused on assessing the overall uncertainty of stormwater quality models. There have been limited studies undertaken on assessing the structural and epistemic uncertainties associated with the processes embedded in stormwater quality models, which are commonly deemed as being unquantifiable.

Providing essential insight into structural and epistemic uncertainties of stormwater quality models is important. In estimating these uncertainties, regression-based techniques are often adopted, e.g. Ordinary Least Squares Regression (OLSR) (e.g. Driver and Tasker, 1988; Zoppou, 2001; Rahman et al., 2002; Haddad et al., 2013) , Weighted Least Squares Regression (WLSR) (Stedinger and Tasker, 1985; Haddad et al., 2010) and Generalised Least Squares Regression (GLSR) (Reis et al., 2005; Haddad and Rahman, 2012; Haddad et al., 2012; Micevski et al., 2014). In this context, this paper discusses a statistical modelling framework for determining structural and epistemic uncertainties associated with the replication of pollutant wash-off from urban roof surfaces using regression-based techniques. Previously, Haddad et al. (2013) investigated uncertainties associated with the replication of the pollutant build-up process using Bayesian OLSR and WLSR. The statistical framework adopted for the study discussed in this paper entailed a significant extension of the Haddad et al. (2013) study. The reasons being, the involvement of different and complex physical processes in pollutant wash-off compared to build-up and the range of parameters influencing the wash-off process and consequently, the resulting data set created. This led to the use of an enhanced Bayesian methodology for the intended statistical framework development. Bayesian based frameworks were selected for this study due to its superiority in assessing the wash-off process with coefficients expressing behavioural parameters (Dotto et al. 2012; Freni and Mannina, 2010).

Investigations by Egodawatta et al. (2009) have confirmed that the pollutant wash-off process for road and roof surfaces, which are the primary impervious surfaces in an urban catchment are mathematically similar despite the differences in surface and pollutant load characteristics. They have noted that these differences can be accounted by utilising different sets of coefficients for the same exponential equation for roads and roofs. Therefore, the application of uncertainty analysis to roof surfaces is easily extendable to road surfaces. Furthermore, as noted by Egodawatta et al. (2012), in an urban catchment, the total roof area can be 2 – 3 times greater than the total road area with the clear potential to contribute relatively high pollutant loads compared to road surfaces.

2. Materials and methods

This research study used roof wash-off data collected at a number of study sites located in South East Queensland, Australia. Wash-off samples were collected from model roofs used as test plots (3 m^2). This approach eliminated the possible heterogeneity in pollutant distribution and the practical difficulties in collecting pollutant wash-off samples from actual roof surfaces. The model roofs were mounted on a scissor lift arrangement as shown in Fig. 1. The roofs were raised to the typical roofing height to enable pollutant accumulation under a typical urban-setting and then lowered to ground level for wash-off sample collection using a rainfall simulator as discussed below. Two roofing products, corrugated steel and concrete tiles were used for cladding since these products are the most widely used roofing materials in the study region. The roofing angle used was 20^0. The model roofs were placed in an area which is mostly residential with a few major roads in the vicinity. Further details on the wash-off sampling, including the solid loads initially available on roofs, fraction wash-off for different rainfall durations and particle size distribution in the wash-off are available in Egodawatta et al. (2009).

2.1 Rainfall simulation

Simulated rainfall was used for the pollutant wash-off investigation on the model roof surfaces. This approach was adopted to eliminate the dependency on naturally occurring rainfall events due to their inherent variability and provided better control over influential variables such as rainfall intensity and duration. A specially designed rainfall simulator was used to simulate the rainfall events. The simulator was designed to replicate natural rainfall events as closely as possible based on two key rainfall characteristics, namely, drop size distribution and kinetic energy of rain drops (Hudson, 1963; Rosewell, 1986). Details on the design and operation of the rainfall simulator can be found in Herngren et al. (2005).

2.2 Sample collection

Sample collection was undertaken in two phases. Firstly, half of each roof surface was used to collect build-up samples by washing the surface with deionised water and a soft brush, which was considered to be representative of the total available pollutants prior to a wash-off event. Then the pre-determined rainfall intensity was simulated on the model roof surfaces. These investigations were conducted on a weekly basis, where only the rainfall intensity was changed. Egodawatta et al. (2013) have shown that an appreciable amount of pollutant build-up will occur on a roof surface after a 7 day antecedent dry period. Rainfall intensities of 20, 40, 86 and 115 mm/h were simulated on the roof surfaces. For each simulation, runoff samples were collected for a range of different durations to match design storms of specific Average Recurrence Intervals (ARI). The resulting data matrix comprised of 8 datasets with fraction wash-off as a function of time for four rainfall intensities simulated on two roofing products.

2.3 Laboratory analysis

Samples collected were transported to the laboratory for testing, with sample handling and preservation undertaken according to AS/NZS (1998). Samples were tested for total suspended solids (TSS) as this is commonly considered as the indicator pollutant in stormwater quality modelling. Testing for TSS was undertaken according to Test Method No. 2540D (APHA, 2005).

2.4 Analytical tools

An exponential function is commonly used to replicate pollutant wash-off. Egodawatta et al. (2012) recommended refinements to this exponential function. The uncertainty associated with using this equation has been assessed in this study using two different parameter estimation techniques, namely, Ordinary Least Squares Regression (OLSR) and Weighted Least Squares Regression (WLSR) in a Bayesian/Gibbs sampling framework. The errors arising in the dependant variable data was taken into consideration by assuming that the data can be described by an appropriate probability distribution i.e. they are stochastic in nature, which allows consideration of inherent temporal and spatial variability of the wash-off process. This is unlike previous studies which commonly assume both, the dependent and independent variables are fixed and deterministic in nature.

2.4.1 Ordinary Least Squares Regression (OLSR)

Simple linear regression under an OLSR framework has been used in the past to develop empirical relationships between predictors and relevant explanatory variables (for example, Driver and Tasker, 1988; Zoppou, 2001; Rahman et al., 2002). A number of theoretical assumptions need to be satisfied for the successful application of the OLSR method. The main assumptions are independence, homoscedasticity and normality of errors (Draper and

Smith, 1981) in the dependent variable data. Urban water quality data is generally characterized by limited record length and measurement errors, which has the potential to introduce a large degree of uncertainty to both, the dependent and independent variables data and modelling outcomes. As such, assumptions often used in simple regression are violated. Therefore, the use of statistical techniques such as OLSR is likely to be limited in practical applications because the strict theoretical assumptions associated with these techniques cannot be adequately satisfied for wash off data.

2.4.2 Weighted Least Squares Regression (WLSR)

WLSR is one way to overcome the problems associated with OLSR. When faced with small datasets plagued by sampling error which is often the case in water quality prediction modelling, WLSR is an efficient and useful method to be adopted in most cases. In this paper, the use of both OLSR and WLSR in a Bayesian/Gibbs sampling framework is proposed. Here, the Bayesian/Gibbs sampling framework provides the posterior distribution of an estimate and can be useful when assessing the uncertainties (credible limits/prediction limits) associated with the estimation of regression coefficients (Geweke, 1992; Raferty et al., 1997 and Chib, 2001). Further details on WLSR, issues associated with using the WSLR method for pollutant wash-off analysis and uncertainty analysis and the Bayesian/Gibbs sampling framework are provided in the Supplementary Information.

2.4.3 Assumptions in regression in relation to water quality modelling

Conventionally, the use of regression based procedures implies that the dependant variable (such as pollutant-wash off) does represent all the possible values in the process and that these values are error free. This is a point of concern in that the available data can be subject to spatial scales that vary greatly and the fact that these datasets span only a relatively short time scale and hence do not capture the full data domain (Kanso et al., 2005). As such, the resulting regression relationship, which is considered to represent the 'true' values of the coefficients of the regression model, in principle is not fully satisfied. Further details on the assumptions and limitations associated with using traditional regression approaches are provided in the Supplementary Information.

2.4.4 Statistical modelling framework

For the purpose of this study, the dependant variable (pollutant wash-off), was assumed to have a relatively low to medium error associated with it. Here, a normal distribution with the observation *i* was chosen as the mean of the distribution. An arbitrary percentage of the mean was then used to describe the standard deviation such that $pollutant\ wash\text{-}off_i \sim N$ ($pollutant\ wash\text{-}off_i$, % $pollutant\ wash\text{-}off_i$). A value of 20% of the mean for the standard deviation for this analysis was assumed. For each $pollutant\ wash\text{-}off_i$ 3,000 (or more) possible values based on the assumed distribution were simulated using the Bayesian/Gibbs sampling framework. Further details regarding the Bayesian/Gibbs sampling analysis are given in the Supplementary Information.

The following steps summarise the analytical procedure used for each dataset:

1. Analysis was undertaken using OLSR in a Bayesian/Gibbs sampling framework to estimate the regression coefficients for pollutant wash-off as a function of time and to assess the associated uncertainty. OLSR was applied for comparison with WLSR.
2. Using the Iteratively Reweighted Least Squares (IRWLS) algorithm, weights for each observation *i* (see Supplementary Information) were estimated.
3. Keeping the pollutant wash-off as a fixed value/variable, the regression coefficients using the Bayesian/Gibbs sampling framework using WLSR were estimated.

Accordingly, 3,000 values of the posterior coefficients (β_0 and β_1) were estimated and the first 1,000 simulated values were used as "burn-in" (i.e. discarded). The remaining 2,000 values were used to assess the uncertainty in the prediction model developed and to estimate the uncertainty relating to the behaviour of the wash-off coefficient *k*.

4. Assuming pollutant wash-off as a random normal variable (see Section 2.4.4), 3,000 values for each $pollutant\ wash\text{-}off_i$ were simulated. The regression coefficients using WLSR were estimated. Accordingly, 3,000 values of the posterior coefficients (β_0 and β_1) were estimated and the first 1,000 simulated values were used as "burn-in". The remaining 2,000 values were used to assess the uncertainty in the prediction model developed and to determine the uncertainty related to the behaviour of the wash-off coefficient *k*.

2.5 Development of the wash-off model

In stormwater quality models, pollutant wash-off is typically replicated using an exponential form of equation. The equation originally proposed by Sartor and Boyd (1972) has been improved over time by other researchers such as Egodawatta et al. (2007; 2012). The exponential wash-off equation proposed by Egodawatta et al. (2007; 2012) given as Eq. (1) below is used as a case study to illustrate the approach for uncertainty analysis proposed in this paper. Eq. (1) defines the terms, 'fraction wash-off' (*Fw*), which is the ratio of the wash-off load to the initially available particulate load and capacity factor (C_F) which signifies the ability of a specific rainfall intensity to mobilise particulates.

$$Fw = \frac{W}{W_o} = C_F(1 - e^{-kIt}) \qquad (1)$$

Where: C_F = capacity factor; F_w = fraction wash-off; I = rainfall intensity; k = wash-off coefficient; t = time in minutes; W = weight of material mobilised after time t; and W_O = initial weight of the material on the surface.

Egodawatta et al. (2012) noted that the applicability of the wash-off equation is primarily dependant on the successful estimation of *k* and C_F. Values for *k* and C_F assuming similar wash-off behaviour for both steel and concrete surface types were presented by Egodawatta et al. (2012).

3.0 Results and Discussion

Egodawatta et al. (2012) have noted that the wash-off coefficient *k* primarily varies with surface type and rainfall intensity. They have observed no significant change in coefficient *k* for two different cladding materials while observing a significant difference in *k* values for roof surfaces and road surfaces in general. Based on this, a uniform *k* value was recommended for common cladding materials. In this study, the decision to use the same *k* value for the two roof surface wash-off behaviour was re-assessed by investigating the stochastic nature of the coefficient *k*. However, Eq. 1 in its current form does not allow for easy stochastic treatment to examine the stochastic behaviour of coefficient *k*.

Wash-off as a function of time can be best replicated by an exponential equation (Sartor and Boyd, 1972). It has been found that the observed behaviour of fraction wash-off (*Fw*) on roof and road surfaces can be closely replicated by using a power equation, which was ascertained by examining a range of equation forms during the initial exploratory data analysis on all the 8 datasets (i.e. fraction wash-off as a function of time for rainfall intensities of 20, 40, 86 and 115 mm/h for the two material types: (i) corrugated steel and (ii) concrete tiles). The results of this initial exploratory data analysis are not given here. The power equation developed is presented as Eq. (2).

$$Fw = \beta_0 t^{\beta_1} \qquad (2)$$

Where: β_o, β_1 = empirical coefficients which are estimated by the Bayesian/Gibbs sampling framework using both the OLSR and WLSR.

To assess the stochastic nature of *k*, Eq. 1 was re-arranged making the wash-off coefficient *k* the subject. The estimated values of *Fw* from the Bayesian/Gibbs sampling simulation (using OLSR and WLSR – 3,000 replicates with 1,000 as burn-in) and substituting back into Eq. 1, along with the rainfall intensity (*I*) and fixing the C_F values (see Egodawatta et al., 2012) to the maximum range was used for a particular rainfall intensity and roofing material. This allowed the full distribution of *k* values to be analysed for each material and the rainfall intensity range adopted in this study.

3.1 Uncertainty estimation

The results presented in this section are primarily based on Eq. 1 and 2. The parameter estimation techniques (i.e. OLSR and WLSR) were applied to eight datasets (steel and concrete tiles and 20, 40, 86 and 115 mm/h rainfall intensities). Tables 1 and 2 provide a summary of the posterior mean coefficients of the regression model (for steel and concrete tiles, respectively) along with their 95% credible limits.

From Tables 1 and 2, it can be seen that the estimated OLSR coefficients are modestly different to the coefficients from the WLSR approach over the different rainfall intensity ranges and roof surface material types considered. More notable differences can be observed in the 95% credible prediction limits for OLSR and WLSR and surface material types. Given that the Bayesian approach provides uncertainty estimation (considers the regression model not to be perfect) in both estimation techniques (OLSR and WLSR), it is important to note that OLSR gives credible prediction limits that are wider than those of WLSR. This is visible for both surface material types. Here particularly WLSR compensates for the influence of each observation and thus is expected to provide more efficient estimates of uncertainty. Considering the sample sizes used in the regression analysis (*n* = 5, 6 and 7) and that OLSR takes no consideration of the data error, it can be understood why OLSR provides the larger credible limits. Fig. 2 (steel, rainfall intensity 86 mm/h) illustrates an example of how OLSR may provide a unrealistic measure of uncertainty. WLSR provides a relatively better measure of uncertainty in this case.

Comparing the coefficients of the regression models based on rainfall intensity and material type (Tables 1 and 2), it can be seen that it is possible to use one wash-off coefficient (*k*) for the different rainfall intensity ranges for each material type, as the regression coefficients are similar. However, when the two different material types are compared directly, it is quite evident that there are significant differences between the regression coefficients. This does suggest that the wash-off coefficient (*k*) may vary with surface material type. To assess this hypothesis, the *k* values were estimated for both material types using Eq.1 based on the simulated values (i.e. from Bayesian/Gibbs sampling approach) of fraction wash-off (*Fw*) for both the OLSR and WLSR methods. Tables 1 and 2 show the expected *k* values along with its standard deviation and 95% credible limit. With each surface material type and intensity, it can be seen from Table 1 and Table 2 that there are some differences in the expected *k* values. This variability is also supported by the different standard deviation in *k*. In terms of differences between the OLSR and WLSR methods, Tables 1 and 2 shows that uncertainty (spread of *k* values i.e. 95% confidence limits) is narrower for the WLSR approach.

When the *k* values of the two different material types are compared directly, it can be seen that the expected *k* values between the material types and rainfall intensities differ significantly. The difference between the two material types is clearly illustrated in the variability in the results as evident in the standard deviation and uncertainty limits in *k* and as the example illustrated in Fig. 3 in terms of the histogram of *k* values (WLSR results). As shown in Fig. 3, the *k* values follow an approximate normal distribution for both materials, suggesting a stochastic nature for *k*. The applicability of the wash-off equation is primarily dependent on the successful estimation of *k*. Therefore, knowledge of the sampling distribution of *k* can provide a more realistic measure of *k*. In any case, the *k* values presented here are based on a maximum range capacity factor (C_F) for a particular rainfall intensity, which in fact if changed will produce a larger variability in *k*. Therefore, the variation in *k* is possibly a function of more complex physical processes which can play a pivotal role in pollutant wash-off modelling.

Illustrated in Fig. 2 are the uncertainty bands (i.e. 95% confidence limits) for the OLSR, WLSR and WLSR + error in *Fw* models for fraction wash-off. The uncertainty bands for OLSR and WLSR do not include the error associated with *Fw* (see Section 2.4.4). It is evident from the overall results that the uncertainty bands provided by WLSR are an appreciable improvement over the OLSR method. The posterior mean and 95% credible limits for the β coefficients are shown in Tables 1 and 2 for the WLSR + error in F_w method. Here it can be seen that the β coefficients of WLSR + error in *Fw* (for steel) are very similar to that of WLSR. However, minor differences can be noted in the 95% credible limits.

It is important to note that WLSR + error in *Fw* is contributing slightly more uncertainty than WLSR as shown in Fig. 2. This result is expected and suggests that the underlying errors should not be ignored in the dependent variables as this can underestimate the overall uncertainty in wash-off modelling. As shown in Fig. 2, by assuming the dependent variable to be probability distributed, a more realistic measure of uncertainty can be estimated. The posterior distributions of β_o and β_1 are shown in Fig. 4 for the concrete tiles surface (2,000 simulations shown). What is noteworthy from Fig. 4 is the fact that WLSR + error in *Fw* gives notably more scatter than the OLSR and WLSR methods. Both, OLSR and WLSR recognise the regression models as uncertain, whereas the WLSR + error in *Fw* recognises the regression model and the *Fw* values to be uncertain thus giving the higher scatter which relates to the slightly larger uncertainty in WLSR + error in *Fw* as compared to WLSR as shown in Fig. 2. The influence of the estimated *Fw* values using the WLSR + error in *Fw* method on the wash-off coefficient (*k*) was also investigated. As evident from the data given in Tables 1 and 2, it is generally found that the uncertainty in *k* increases slightly compared to the WLSR approach.

Based on the outcomes of the analysis undertaken, it is evident that the Bayesian/Gibbs sampling approach which is based on actual data (see sections 2.4.4 and Supplementary Information) is advantageous as it allows the assessment of candidate regression models (posterior distribution) in a better way compared to the application of classical statistical analysis (i.e. point estimates only). The approach adopted here provided better insights into the stochastic nature of fraction wash-off and *k*.

In the case of limited data, the underlying errors in the dependent variable should not be ignored. The simulation of a larger dataset by allowing for the probability distributed variables and allowing for uncertainty in the regression model has helped towards a more comprehensive statistical analysis. Overall, realistic measures of uncertainty for the fraction wash-off and the wash-off coefficient were found as demonstrated in Fig. 3 and 4.

3.2 Validation
With the development of any empirical models for pollutant wash-off (Eq. 1), the issue of uncertainty directly influences the prediction performance (i.e. over or under estimation). Accordingly, a validation exercise was carried out using Eq. 1 where it was used to predict the fraction wash-off based on both roof surface types and all the rainfall intensities using the methodology proposed by Egodawatta et al. (2012) (point estimate of k) and the stochastic nature of k as shown in this study. An example of the results of this validation is presented in Fig. 5 for the steel roof surface and rainfall intensity of 40 mm/h.

It can be seen from Fig. 5 that the estimation of fraction wash-off based on the point estimation of the wash-off coefficient can substantially underestimate the observed values. In contrast, the estimation of fraction wash-off based on the stochastic nature of the wash-off coefficient provides a better fit to the observed data values.

The results from this analysis suggest that the prediction error associated with pollutant wash-off equation developed based on a limited data set is appreciably high. This is a possible reason for under estimation of the wash-off load. Such under estimation could potentially undermine the effectiveness of stormwater quality management strategies.

Clearly one of the approaches to enhance the accuracy of prediction from any urban stormwater quality model is to have a larger observed data set. A scientifically robust statistical modelling approach which can assess the uncertainty associated with the use of water quality data with small sample size provides a practical solution to counteract the limitations which are otherwise imposed on water quality modelling. This study has demonstrated that the probability distributed data in a proper regression framework can provide more realistic outcomes (i.e. fraction-wash-off/wash-off coefficient) and error estimation.

Given the inability to develop a mathematical model that can exactly replicate nature, a more robust approach is needed to estimate uncertainty as presented in this study. Because of the randomness of the statistical modelling framework used here and the different results provided each time it is run makes it a useful alternative to deterministic algorithms for use in water quality modelling.

4.0 Conclusions
The study presented in this paper outlines a statistical modelling framework to quantify epistemic uncertainty associated with pollutant wash-off process replication using Ordinary Least Squares Regression (OLSR) and Weighted Least Squares Regression (WLSR) approaches based on a Bayesian/Gibbs sampling procedure. The analysis revealed that the WLSR method with probability distributed variable data can provide more realistic uncertainty estimates of the observed and predicted values compared to OLSR, which is likely to provide misleading measures of uncertainty. It is also noteworthy that the Bayesian/Gibbs sampling procedure presented in this study was proven to be a robust

approach as compared to many of the classical statistical and deterministic methods often used in stormwater quality modelling.

The outcomes from this study suggest that the prediction error associated with the wash-off replication equation is relatively higher and could underestimate the actual values. This is postulated to be due to the relatively limited observed data set. Therefore, this highlights the importance of having a larger data set for model development and validation to assess model uncertainty in a more robust manner in urban stormwater quality modelling. The analytical framework developed in this study demonstrates that more realistic results and error estimations can be derived using larger data sets.

The uncertainty analysis confirmed the use of constant wash-off coefficient (k) in a practical context. However, the results also highlighted the fact that the variation of k as a function of complex physical processes is significantly influenced by surface characteristics and to a lesser extent by the rainfall intensity.

Supplementary Information

Further details on the assumptions and limitations associated with using Ordinary Least Squares Regression (OLSR), Weighted Least Squares Regression (WLSR), issues associated with using the WSLR method for pollutant wash-off analysis and uncertainty analysis and the Bayesian/Gibbs sampling framework are provided in the Supplementary Information.

References

APHA. Standard Methods for the Examination of Water and Wastewater. American Public Health Association, Washington; D.C.; 2005.

AS/NZS 5667.1. Water Quality Sampling – Guidance on the design of sampling programs, Sampling techniques and the preservation and handling of samples. Australia/New Zealand Standards; 1998.

Bertrand-Krajewski J-L. Stormwater pollutant loads modelling: epistemological aspects and case studies on the influence of field data sets on calibration and verification. Water Science and Technology 2007; 55(4):1-17.

Bertrand-Krajewski J-L, Barraud S, Bardin J-P. Uncertainties, performance indicators and decision aid applied to stormwater facilities. Urban Water 2002; 4(2):163-79.

Chib S. Markov chain Monte Carlo methods: Computation and inference, in Heckman, J. J. and Leaner, E. (eds), Handbook of Econometrics, Elsevier; 2001.

Dotto CBS, Mannina G, Kleidorfer M, Vezzaro L, Henrichs M et al. Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling. Water Research 2012; 46:2545-58.

Dotto CBS, Kleidorfer M, Deletic A, Rauch W and Mccarthy DT. Impacts of measured data uncertainty on urban stormwater models. Journal of Hydrology 2014; 508:28-42.

Draper NR and Smith H. Applied Regression Analysis. 2nd ed. New York: Wiley; 1981.

Driver NE, Tasker GD. Techniques for estimation of storm-runoff loads, volumes, and selected constituent concentrations in urban watersheds in the United States. US Geological Survey, Open File Report 88-191; 2001.

Egodawatta P, Thomas E, Goonetilleke A. Understanding the physical processes of pollutant build-up and wash-off on roof surfaces. Science of the Total Environment 2009; 407(6):1834-41.

Egodawatta P, Miguntanna NS, Goonetilleke A. Impact of roof surface runoff on urban water quality. Water Science and Technology 2012; 66:1527-33.

Egodawatta P, Ziyath AM, Goonetilleke A. Characterising metal build-up on urban road surfaces. Environmental Pollution 2013; 176:87-91.

Freni G and Mannina G. Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution. Journal of Hydrology 2010; 392:31-39.

Freni G, Mannina G, Viviani G. Urban runoff modelling uncertainty: Comparison among Bayesian and pseudo-Bayesian methods. Environmental Modelling & Software 2009; 24:1100-1111.

Freni G, Mannina G, Viviani G. Uncertainty in urban stormwater quality modelling: The effect of acceptability threshold in the GLUE methodology. Water Research 2008; 42:2061-72.

Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (eds), Bayesian Statistics 4, Oxford University Press, Oxford, UK; 1992.

Haddad K, Zaman M, Rahman A. Regionalisation of skew for flood frequency analysis: a case study for eastern NSW. Australian Journal of Water Resources 2010; 14(1): 33-41.

Haddad K, Rahman A. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework: Quantile regression vs. parameter regression technique. Journal of Hydrology 2012; 430-431, 142-161.

Haddad K, Rahman A, Stedinger JR. Regional flood frequency analysis using Bayesian generalized least squares: a comparison between quantile and parameter regression techniques. Hydrological Processes 2012; 26:1008-1021.

Haddad K, Egodawatta P, Rahman A, Goonetilleke A. Uncertainty analysis of pollutant build-up modelling based on a Bayesian weighted least squares approach. Science of the Total Environment 2013; 449:410-7.

Herngren L, Goonetilleke A, Ayoko GA. Understanding heavy metal and suspended solids relationships in urban stormwater using simulated rainfall. Journal of Environmental Management 2005; 76:149-58.

Hudson NW. Raindrop Size Distribution in High Intensity Storms. Agricultural Research 1963; 1:6-11.

Kanso A, Chebbo G, Tassin B. Stormwater quality modelling in combined sewers: calibration and uncertainty analysis. Water Science and Technology 2005; 52:63-71.

Liu A, Goonetilleke A, Egodawatta P. Inherent errors in pollutant build-up estimation in considering urban land use as a lumped parameter. Journal of Environmental Quality 2012; 41(5):1690-1694.

Liu A, Goonetilleke A, Egodawatta P. Inadequacy of land use and impervious area fraction for determining urban stormwater quality. Water Resources Management 2012; 26:2259-2265.

Merz B, Thieken AH. Separating natural and epistemic uncertainty in flood frequency analysis. Journal of Hydrology 2005; 309: 114-32.

Micevski T, Hackelbusch A, Haddad K, Kuczera G, Rahman A. Regionalisation of the parameters of the log-Pearson 3 distribution: a case study for New South Wales, Australia, Hydrological Processes 2014, DOI: 10.1002/hyp.10147.

Oberkampf WL, Deland SM, Rutherford BM, Diegert KV, Alvin KF. Error and uncertainty in modelling and simulation. Reliability Engineering and System Safety 2002; 75:333-357.

Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 1997; 92(437):179- 91.

Reis Jr DS, Stedinger JR, Martins ES. Bayesian generalize least squares regression with application to log Pearson type 3 regional skew estimation. Water Resources Research 2005; 41, W10419, doi:10.1029/2004WR003445.

Rahman A, Thomas E, Bhuiyan S, Goonetilleke A. Modelling pollutant washoff from south-east Queensland catchments Australia. Proc. Enviro 2002 and International Water Association 3rd World Water Congress, A unique approach to a unique environment, Melbourne, 7-12 April 2002. available at http://eprints.qut.edu.au/4330/

Rosewell CJ. Rainfall Kinetic Energy in Eastern Australia. Journal of Applied Meteorology 1986; 25(11):1695-1701.

Sartor JD, Boyd GB. Water pollution aspects of street surface contaminants. Report No. EPA-R2-72/081, US Environmental Protection Agency, Washington DC; 1972.

Sohrabi TM, Shirmohammadi A, Chu TW, Montas H, Nejadhashemi AP. Uncertainty analysis of hydrologic and water quality predictions for a small watershed using SWAT2000. Environmental Forensics 2003; 4:229-238.

Stedinger JR, Tasker GD. Regional hydrologic analysis, 1.Ordinary, weighted, and generalised least squares compared. Water Resources Research 1985; 21(9): 1421-1432.

Zoppou C. Review of urban storm water models. Environmental Modelling & Software 2001; 16:195-231.

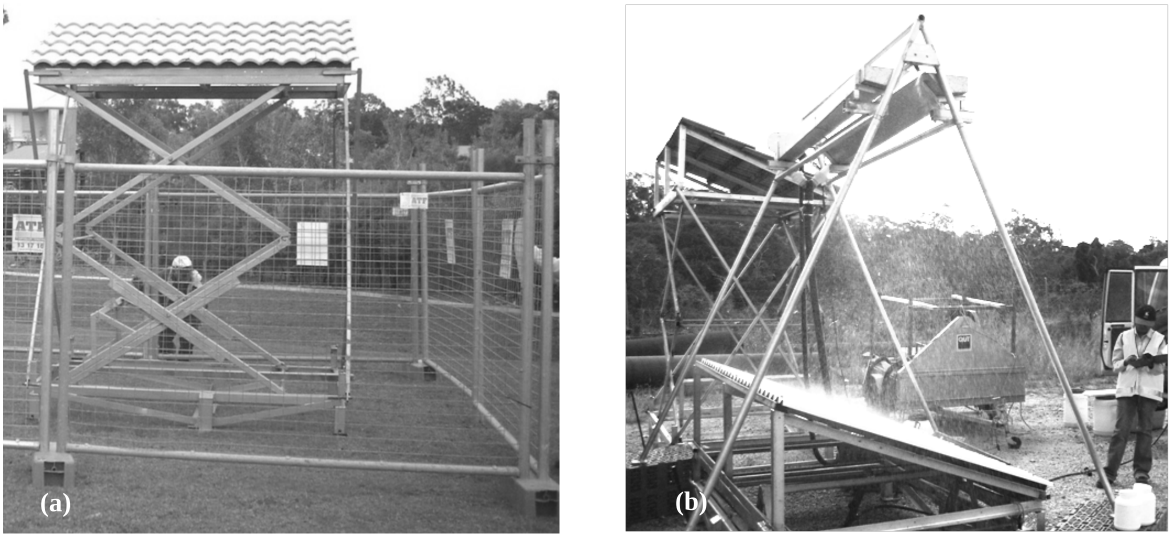

Fig. 1 Model roof surfaces: (a) Model roofs with the scissor lift arrangement; (b) Pollutant wash-off investigation using the rainfall simulator

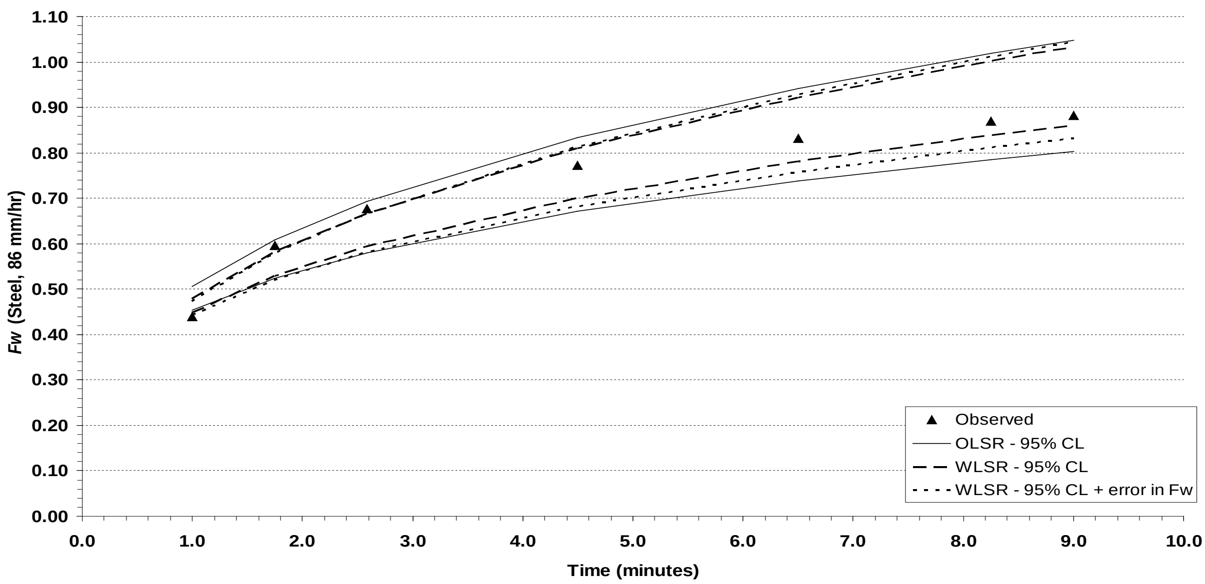

Fig. 2 Observed fraction wash-off and estimated uncertainty in fraction wash-off for the steel roof (CL 95% - Upper and lower confidence limits associated with OLSR and WLSR and Fw is fraction wash-off

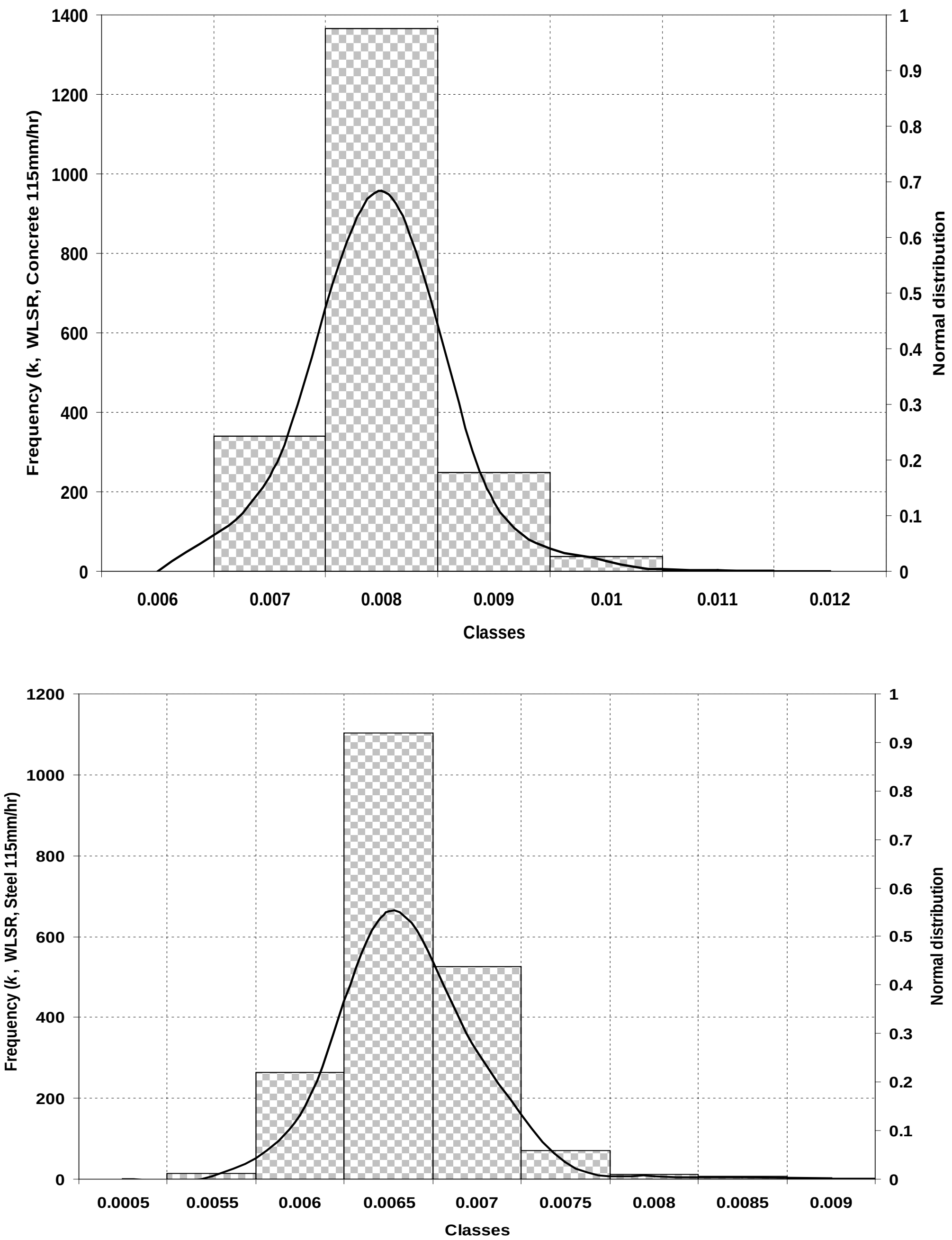

Fig. 3 Histogram of wash-off coefficient (k) for concrete tiles and steel roof surface materials for rainfall intensity 115 mm/h (WLSR results)

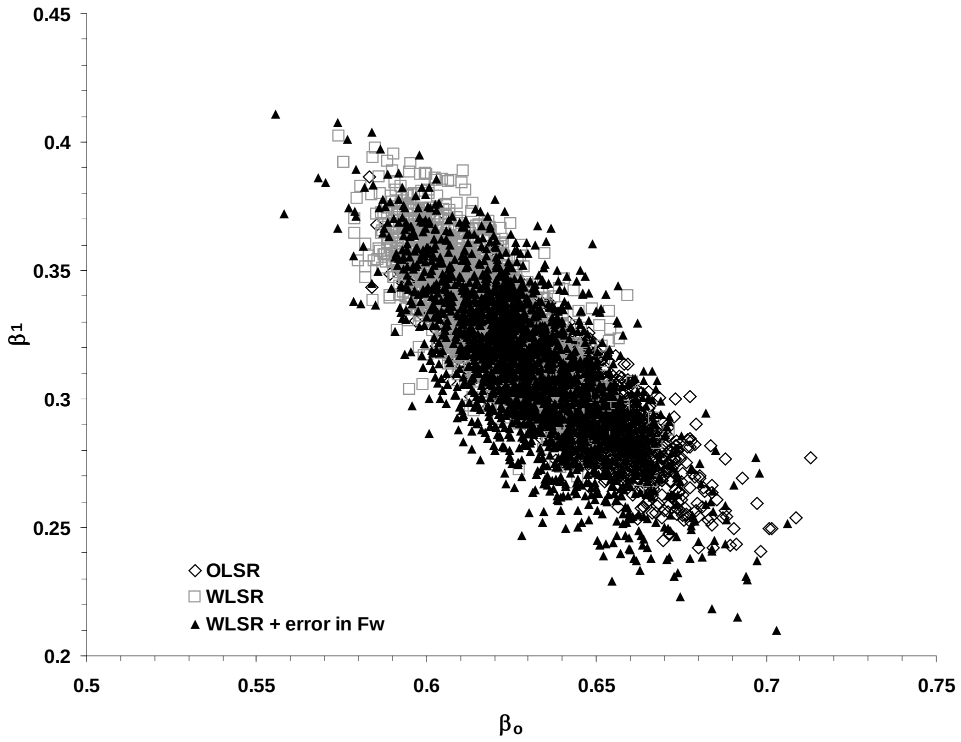

Fig. 4 Estimated β coefficients from Monte Carlo simulation for the concrete tiles and rainfall intensity (for 115 mm/h rainfall intensity)

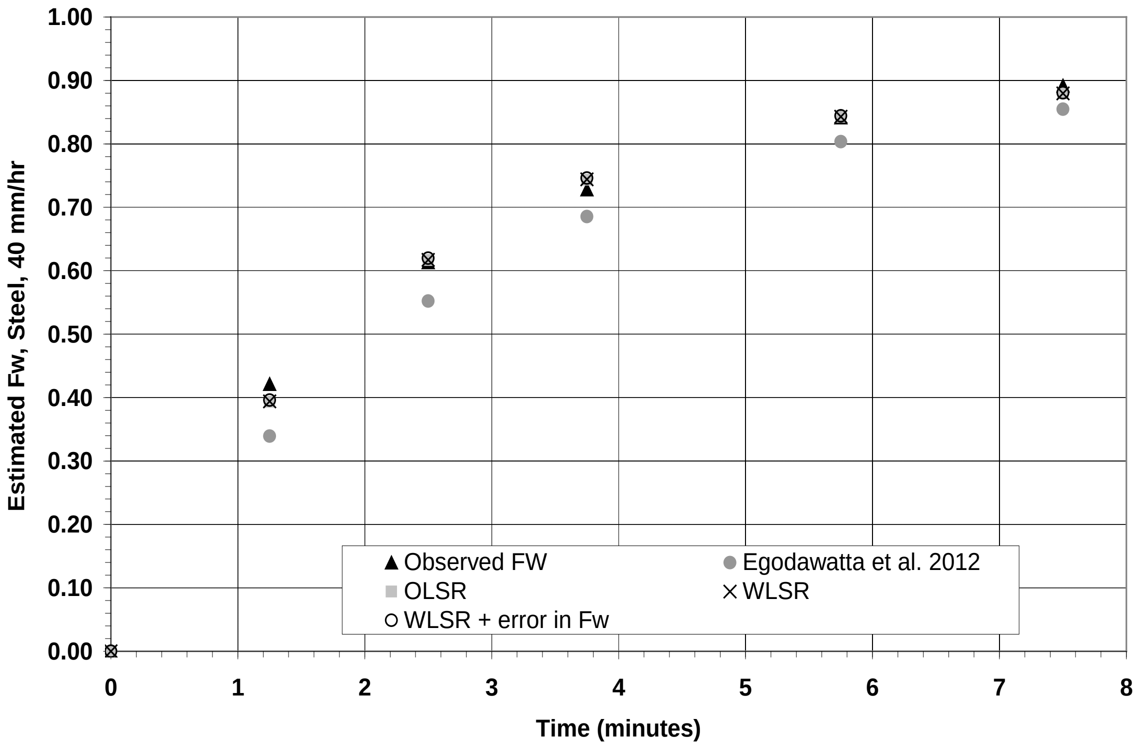

Fig. 5 Outcomes of model validation for the different methods (OLSR, WLSR, WLSR + error in Fw)

Table 1 - Summary of results associated with the regression coefficients for steel roof surface using the methods – OLSR, WLSR and WLSR + error in *Fw*, E(*k*) expected *k* value. (St.dev is standard deviation of *k* values)

Steel			**Regression Coefficients - OLSR**						***k***	
I (mm/h)	β_o	β_1	β_o - LL	β_o - UL	β_1 - LL	β_1 - UL	E(*k*)	St Dev	LL	UL
20	0.268	0.353	0.242	0.297	0.303	0.405	0.0068	0.00022	0.00643	0.00716
40	0.401	0.421	0.373	0.430	0.366	0.473	0.0113	0.00067	0.01042	0.01258
86	0.479	0.296	0.453	0.505	0.260	0.332	0.0044	0.00018	0.00414	0.00474
115	0.614	0.286	0.579	0.650	0.247	0.325	0.0065	0.00039	0.00591	0.00714
			Regression Coefficients - WLSR						***k***	
I (mm/h)	β_o	β_1	β_o - LL	β_o - UL	β_1 - LL	β_1 - UL	E(*k*)	St Dev	LL	UL
20	0.256	0.376	0.244	0.270	0.346	0.405	0.0068	0.00015	0.00653	0.00702
40	0.389	0.447	0.373	0.406	0.408	0.485	0.0113	0.00062	0.01054	0.01253
86	0.463	0.324	0.448	0.478	0.297	0.351	0.0045	0.00022	0.00419	0.00490
115	0.592	0.314	0.564	0.621	0.275	0.347	0.0064	0.00038	0.00582	0.00698
			Regression Coefficients – WLSR + error in *Fw*						***k***	
I (mm/h)	β_o	β_1	β_o - LL	β_o - UL	β_1 - LL	β_1 - UL	E(*k*)	St Dev	LL	UL
20	0.256	0.377	0.240	0.273	0.339	0.415	0.00678	0.00019	0.00647	0.00710
40	0.389	0.447	0.367	0.412	0.395	0.490	0.0114	0.00083	0.01038	0.01294
86	0.463	0.325	0.443	0.473	0.287	0.360	0.0045	0.00031	0.00413	0.00514
115	0.593	0.313	0.563	0.625	0.272	0.354	0.0064	0.00040	0.00575	0.00702

Table 2 - Summary of results associated with the regression coefficients for concrete tiles surface using the different methods – OLSR, WLSR and WLSR + error in *Fw*, E(*k*) expected *k* value (St. dev is standard deviation of *k* values)

Concrete tiles			**Regression Coefficients - OLSR**						***k***	
I (mm/h)	β_o	β_1	β_o - LL	β_o - UL	β_1 - LL	β_1 - UL	E(*k*)	St Dev	LL	UL
20	0.183	0.547	0.164	0.205	0.487	0.604	0.0230	0.00273	0.01966	0.02806
40	0.323	0.582	0.301	0.347	0.531	0.631	0.0268	0.00211	0.02361	0.03054
86	0.424	0.448	0.401	0.448	0.403	0.491	0.01520	0.00129	0.01337	0.01741
115	0.640	0.304	0.611	0.669	0.270	0.339	0.00771	0.00057	0.00695	0.00871
			Regression Coefficients - WLSR						***k***	
I (mm/h)	β_o	β_1	β_o - LL	β_o - UL	β_1 - LL	β_1 - UL	E(*k*)	St Dev	LL	UL
20	0.167	0.596	0.160	0.175	0.566	0.626	0.02242	0.00183	0.02008	0.02553
40	0.314	0.605	0.303	0.325	0.572	0.639	0.02540	0.00095	0.02388	0.02700
86	0.412	0.477	0.400	0.424	0.445	0.508	0.01430	0.00060	0.01341	0.01531
115	0.617	0.334	0.593	0.642	0.298	0.369	0.00748	0.00056	0.00675	0.00853
			Regression Coefficients - WLSR + error in *Fw*						***k***	
I (mm/h)	β_o	β_1	β_o - LL	β_o - UL	β_1 - LL	β_1 - UL	E(*k*)	St Dev	LL	UL
20	0.167	0.596	0.157	0.178	0.555	0.635	0.02221	0.00231	0.01918	0.02627
40	0.314	0.603	0.302	0.327	0.566	0.641	0.02544	0.00104	0.02377	0.02719
86	0.413	0.477	0.396	0.430	0.433	0.512	0.0144	0.00084	0.01317	0.01591
115	0.631	0.311	0.599	0.665	0.262	0.360	0.00746	0.00062	0.00659	0.00858

SUPPLEMENTARY INFORMATION

A Bayesian regression approach to assess uncertainty in pollutant wash-off modelling

Prasanna Egodawatta[1], Khaled Haddad[2], Ataur Rahman[2], Ashantha Goonetilleke[1*]
[1]Science and Engineering Faculty, Queensland University of Technology, GPO Box 2434, Brisbane 4001, Australia
[2]School of Computing, Engineering and Mathematics, University of Western Sydney, Building XB, Locked Bag 1797, Penrith, NSW 2751, Australia

Assumptions in Regression

Aside from a loss of efficiency in regression coefficients, the fact that the dependent variable is estimated does not necessarily present any difficulties for regression analysis. Errors of measurement are often included in discussions of regression residuals analysis. However, if the sampling uncertainty in the dependent variable is not constant across observations, the regression errors will be heteroscedastic and this will introduce further inefficiency and can produce inconsistent standard error of estimate undermining subsequent uncertainty analysis.

Therefore, in this analysis the dependant variable, pollutant wash-off (as represented by fraction wash-off for a steel or concrete tiles surface for different rainfall intensities) was assumed to have an inherent uncertainty associated with its measurement (observation). Secondly, there also remains the issue of sampling variability as a major source of concern (i.e. n = 5, 6 and 7 for some cases). Understanding the underlying physical process helps in selecting distributions to describe the pollutant wash-off for the Bayesian/Gibbs sampling Monte Carlo framework.

Ordinary Least Squares Regression (OLSR) and Weighted Least Squares Regression (WLSR)

The OLSR and WLSR models assume that the quantity of interest y_i (or observation i) can be described by a linear or nonlinear function of predictor variables (or a transformation there of) with an additive error. In matrix notation, the model is represented by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\mathbf{X}$ is a ($n \times k$) matrix of predictor augmented by a column of ones, $\boldsymbol{\beta}$ is a ($k \times 1$) vector of regression coefficients that must be estimated and $\boldsymbol{\varepsilon}$ is the vector containing the random errors for each of the n observations used in the regression which are assumed to be normally distributed with zero mean and the covariance matrix of the form:

$$\mathbf{E}(\varepsilon\varepsilon^T) = \sigma^2\mathbf{\Omega} \qquad (2)$$

wherein σ^2 is the model error variance and $\mathbf{\Omega}$ is a positive definite symmetric matrix. The OLSR can be applied to estimate the regression model coefficients, if Ω is equal to the identity matrix $\mathbf{I}$. Uncorrelated errors with different variances at different observational stations can be described using a $\mathbf{\Omega}$ matrix with different variances of the diagonal and zero off the diagonal. In this case, the model in Equation (2) reduces to WLSR and Ω may be replaced with W.

The OLSR estimator of β is given by:

$$\hat{\beta}_{OLSR} = (\mathbf{X}^{\mathbf{T}}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{3}$$

The WLSR estimator of β is given by:

$$\hat{\beta}_{WLSR} = (\mathbf{X}^{\mathbf{T}}\mathbf{W}^{-1}\,\mathbf{X})^{-1}\,\mathbf{X}^{\mathbf{T}}\mathbf{W}^{-\mathbf{1}}\mathbf{y} \tag{4}$$

Advantages of the WLSR procedure are that it provides nearly unbiased and minimum variance estimators of the model coefficients β. Moreover, it provides a less biased estimate of the model error (residual error) and a relatively more accurate estimate of the variance of the coefficient estimates.

WLSR may provide a simple and flexible framework for water quality prediction models when only limited water quality data is available. However, WLSR alone may not provide the adequate estimator of uncertainty needed for the estimated coefficients (β) of water quality prediction models. To this end, Bayesian analysis is a natural framework to fill the inadequacy as it provides the full posterior distribution of the coefficients (β) of the water quality prediction model (Reis et al., 2005; Avellaneda et al., 2011).

Bayesian WLSR

In a Bayesian framework, the parameters (coefficients) of the model (Equation 3) are considered to be random variables, whose probability density function (PDF) should be estimated. The Bayesian approach combines any data with prior information (if available) about the parameters being estimated. This information is usually established from other relevant data sets, previous studies or specific knowledge about the behavior of the system being analysed. Parameter estimation is made through the posterior distribution which is developed using Bayes' rule (see Zellner, 1991).

Providing a full posterior distribution of the parameters is an advantage in the Bayesian approach over classical methods, which usually give a point estimate of the parameters and makes use of asymptotic normality assumptions to evaluate the uncertainties (Congdon, 2001).

With the Bayesian approach, it was assumed that there is no prior information on any of the β coefficients. Thus, a multivariate normal distribution with mean zero and a large variance (e.g. greater than 100) was used as a prior for the regression coefficients as suggested by Reis et al. (2005). This prior was considered to be almost non-informative, which produced a PDF that is generally flat in the region of interest. The likelihood function for the data as suggested by Reis et al. (2005) was considered to be a multivariate normal distribution (Haddad et al., 2012, Haddad and Rahman, 2012).

Gibbs Sampler

The Gibbs Sampler is an example of Markov Chain Monte Carlo method. The Gibbs Sampler is one of the most powerful methods of Bayesian posterior simulation. The strategy was to split the joint posterior into conditional posteriors. Sampling was then carried out sequentially and repeatedly from the conditional posteriors. After a sufficient number of draws, conditional draws converged to the desired joint posterior densities for any parameter required (i.e. regression coefficients in this case), Assuming the quantity vector of interest is given by Θ, this has a posterior density of $p(\Theta \mid \mathbf{y}) \propto p(\Theta) p(\mathbf{y} \mid \Theta)$. Further Θ was split into two vectors such that Θ_1 and Θ_2 were generated. For the analysis carried out in this paper, it was considered that $\Theta_1 = \boldsymbol{\beta}$ and $\Theta_2 = \sigma^2$. The key point here is that the analytical form of the resulting full conditional posterior distributions was known, i.e.,

$$p(\Theta_1 \mid \mathbf{y}, \Theta_2) \textit{ and } p(\Theta_2 \mid \mathbf{y}, \Theta_1) \tag{4}$$

To apply the Gibbs Sampler, the following steps were undertaken:

1. An initial estimated value of Θ_2 was needed which was expressed by Θ_2^0. In this analysis the OLSR result for Θ_2^0 was used.
2. Further, it was assumed that this value comes directly from the marginal distribution $p(\Theta_2 \mid \mathbf{y})$.
3. Next Θ_1 was drawn conditional on Θ_2^0 from $p(\Theta_1 \mid \mathbf{y}, \Theta_2^0)$.This draw was expressed as Θ_1^1.
4. Another draw was undertaken from Θ_2 conditional on Θ_1^1 from $p(\Theta_2 \mid \mathbf{y}, \Theta_1^1)$. This draw was referred to as Θ_1^2
5. This procedure was repeated *N* times. *N* in this study was taken as 3,000.

To ensure the effect of the starting value has disappeared, it was accepted that the first q_1 draws of the sequence should be discarded. For this study, the remaining $q_2 = N - q_1$ were kept. Here q_1 =1000. The discarded draws are referred to as "burn-ins".

Estimation of the weights – Iteratively Reweighted Least Squares

The singular disadvantage of WLSR is the fact that the theory behind this method is based on the assumption that the weights are exactly known (Carroll and Rupert, 1988). The exact weights are almost never known in real applications such as water quality modelling. Therefore, estimated weights must be used instead. When the weights are estimated from small numbers of replicated observations, the results of an analysis can be unpredictably affected. Hence, to apply the Bayesian regression in a WLSR framework, weights can be assigned to each observation using a simple Iteratively Reweighted Least Squares (IRWLS) approach.

With IRWLS, weights are derived from the data. An OLSR is first computed where all weights are initially set equal to one. Points nearest to the OLSR line are then given weights near one, while points further away have lesser weight. A WLSR is computed, and the process repeated. After about two iterations, the weights are stabilised and the final iteratively weighted least squares line is generated.

References

Avellaneda P, Ballestero T, Roseen R, Houle J, Linder E. Bayesian storm-water quality model and its application to water quality monitoring. Journal of Environmental Engineering 2011; 137 (7):541-550.

Carroll RJ, Ruppert D. Transformation and Weighting in Regression. Chapman and Hall, New York; 1988.

Congdon P. Bayesian statistical modelling. John Wiley and Sons, UK; 2001.

Haddad K, Rahman A, Stedinger JR. Regional flood frequency analysis using Bayesian generalized least squares: a comparison between quantile and parameter regression techniques. Hydrological Processes 2012; 26:1008-1021.

Haddad K, Rahman A. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework: Quantile regression vs. parameter regression technique. Journal of Hydrology 2012; 430-431, 142-161.

Reis Jr DS, Stedinger JR, Martins ES. Bayesian generalize least squares regression with application to log Pearson type 3 regional skew estimation. Water Resources Research 2005; 41, W10419, doi:10.1029/2004WR003445.

Zellner A. An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons Inc., New York; 1971.