Chapter 9

Use of a Multiple Linear Regression Model to Estimate Stormwater Pollutant Loading

Thomas A. Hodge and Louis J. Armstrong

Woodward-Clyde Consultants

500 12th St, Oakland, California 94607-4014

It is widely hypothesized that the quality of stormwater runoff varies by land use. Therefore, pollution loading from a particular watershed can be formulated as a function of the area of the various land use types that form it. In order to develop an estimate of regional pollutant loads, it is desirable to derive a best estimate of the quality of storm runoff from different land uses in the region. A multiple linear regression model was developed from water quality and land use data collected during the 1989-90 wet weather season in Alameda County, California. This model was used to estimate runoff pollutant concentrations by land use for this region. The resultant pollution concentrations were used to estimate the pollutant loads for an average year. Both single storm event data and averaged data from eleven storm events occurring during one wet weather season were used. The F-test was applied to statistically test the validity of the model.

Hodge, T.A. and L.J. Armstrong. 1993. "Use of a Multiple Linear Regression Model to Estimate Stormwater Pollutant Loading." *Journal of Water Management Modeling* R175-09. doi: 10.14796/JWMM.R175-09.

© CHI 1993 www.chijournal.org ISSN: 2292-6062 (Formerly in New Techniques for Modelling the Management of Stormwater Quality Impacts. ISBN: 0-87371-898-4)

9.1 Introduction

As cities and counties start to implement the United States federal regulations regarding storm water pollution, they are collecting a substantial amount of water quality data. This water quality data is being used to estimate impacts due to pollutant loading to receiving waters. Since runoff and pollutant concentrations can vary by land use, many of the loading calculations are done for individual land use types. Small homogeneous drainage areas are typically sampled to estimate storm water concentrations by land use. However, it is difficult to collect sufficient data from drainage areas with homogeneous land use. There are also questions concerning the representativeness of the selected sampling sites. By focusing on small sampling stations with homogeneous land uses, large catchments which drain areas of heterogeneous land use are ignored. Sampling sites may therefore represent a much smaller portion of a study area than if larger drainage areas had been included.

We present a method, using a multiple linear regression model (MLR), which avoids some of these problems. Data from ten relatively homogeneous land use stations and five heterogeneous stream stations in Alameda County, California (Figure 9.1) were used. A total of eleven storms were sampled but no one station was sampled during more than eight storms.

9.2 Review of Prior Studies

The Clean Water Act has spurred a number of studies to investigate the magnitude and nature of pollution from urban storm water runoff. The most comprehensive of these is the Nationwide Urban Runoff Program (NURP) of the U.S. Environmental Protection Agency (EPA, 1983). This study analyzed runoff for chemical constituents at 85 stations nationwide. Other more recent studies have generally involved a specific region of the country and fewer constituents.

Figure 9.1: Monitoring stations in Alameda County.

In all studies reviewed, investigators have attempted to identify the relationship between land use and stormwater quality. It seems apparent that such a relationship should exist. It is expected that industrial areas are generally more polluted than residential areas. Except for differences between urban and non-urban runoff, the NURP study was unable to verify that a relationship with land use category holds. NURP found that there was no statistically significant variations among urban sites, and asserted that the land use category is virtually useless in predicting station variation (EPA, 1983). Despite this, NURP computed median storm event mean concentrations (EMCs) for residential, mixed, commercial and open land use categories. An EMC is a flow weighted mean concentration for a storm event. These EMCs were presented in the NURP Final Report accompanied by suggestions of caution on their use.

Several studies have used land use specific EMCs, either using NURP data or locally collected data, to compute pollutant load from a region. Woodward-Clyde Consultants (WCC) estimated pollutant loads from Santa Clara Valley, California using data from local sampling stations (WCC 199 1a). Sampling stations were chosen based on uniformity of land use. A representative concentration of runoff by land use was computed by averaging the site mean concentrations from stations with similar land use. Runoff volumes were computed using the EPA's Storm Water Management Model (SWMM). Volumes of runoff from each land use type were multiplied by the representative concentration to compute loads. Other stations, with mixed land uses, were used for verification of load estimates for large watersheds. A bias correction factor was applied to total load estimates based on the verification error. The need for this factor indicates difficulties with applying data from small watersheds to large ones.

The Aquatic Habitat Institute (AHI) used data from WCC's Santa Clara Valley study and several other studies to compute pollutant load to the entire San Francisco Bay (Gunther, 1991). AHI used runoff coefficients to compute runoff volume. Runoff coefficients are simply the average proportion of rainfall which

appears as runoff, and can be estimated from the percent impervious surface for a watershed. Constituent concentrations were computed similarly to the Santa Clara study, but with an expanded data set.

Silverman, Stenstrom and Fam (1988), developed an approach that is the most similar to what we present. They computed the load of oil and grease to San Francisco Bay using data collected at fifteen stations within the region. Silverman, et al. also used runoff coefficients to compute flow volumes, though they were developed from measured flow data at five of the stations. Land uses from census tract data were used to determine the percent of land tributary to each station identified as residential, commercial/industrial or undeveloped. A total of 34 samples were taken at these stations. With land use as independent variables and oil and grease concentration as the dependent variable, a regression model was solved using 34 equations and three unknowns. A negative coefficient for undeveloped area was set to zero, as negative values were thought to be meaningless.

9.3 Model Description

Using storm runoff and concentration data for total lead and total copper, a system of equations was developed to predict the mean concentration for each land use. For this model the runoffs were estimated using a calibrated Storm Water Management Model (SWMM). The equations were formulated such that the sum of the products of the runoff from each land use type times the concentrations associated with each land use type would equal the product of the total runoff times the total concentration, or the total load. The generic equation is of the form:

$$\sum M_a X_a = Y \quad (9.1)$$

where:

- X_a is the runoff generated from land use "a", as estimated from the SWMM model,
- M_a is the unknown concentration for land use "a",
- Y is the total measured load at the monitoring station.

The equations were normalized by dividing by the flow volume. This provided a unit flow matrix which has the added advantage that the independent variable (Y) and the unknown coefficient (M) are in the same units of concentration.

Since there were more sampling data than land use categories, i.e., more equations than unknowns, the system of equations is considered to be over-determined. If a system of equations is over determined, then there is no uniquely valid solution to the system. To resolve this problem, a multiple linear regression was used to find the best fit to the system of equations. Multiple linear regression (MLR) is a method of finding the line that best fits the collected data when more than one variable exists.

The model was tested for three different systems of equations. For all three systems, the land use categories used were: open, commercial, residential, and industrial. Transportation was combined with commercial since the proportion of flow to the sampling stations contributed by transportation areas was low, yielding poor estimates.

For the first system, annual site mean concentrations were used for the fifteen sampling stations. A site mean concentration is the mean of the EMCs for a sampling station. The system then consisted of a 15 by 4 matrix. The advantage of using the annual site mean concentrations is that variations in concentrations between storm events have been incorporated into the mean. This model was also easy to formulate since site mean concentrations are regularly reported and limit the amount of data handling needed. However, one problem using site mean concentrations is that the stations vary by which storms were sampled and averaged during the year. Due to the difficulty in sampling many stations

at one time, different storms were sampled among stations. Thus the inputs to the model (the site mean concentrations) are not truly comparable between stations.

For the second system, EMCs for individual storms were used. This method produces near uniformity in storm characteristics among samples. The problem with this method was that the amount of data available was not consistent for all storms. This system of equations appeared encouraging for one storm where thirteen stations were sampled, but was unreliable when there was less data. There were no storms where all stations were sampled simultaneously.

The last system of equations used event mean concentrations for all storms. This system provides the most data to run the MLR and resulted in a 93 by 4 matrix. This system loses the uniformity of storm conditions, but provides a much more robust model due to the large data set and therefore is the preferred model.

Statistical parameters from the regression models are presented in Table 9.1. Results from both the preferred model and the simplified model using site mean concentrations are provided for comparison. As we see, the R^2 is lower for the final version of the model using all storms. This is deceiving since the degree of freedom is different.

To statistically test the validity of the MLR models, the F-test was applied (Draper and Smith, 1966). The F value is substantially higher for the preferred model which indicates better confidence in the validity of the model. From comparisons with published F distribution tables (Gibra, 1973), our confidence that the coefficients are significantly different from zero (i.e. that there is a relationship with land use) is better than 99 percent for copper and 95 percent for lead. However, the R^2 does tell us that less than 20 percent of the variability in the sample data is due to land use.

One additional advantage of using all the sample data rather than the station mean values is that the error estimates generated from the regression are the total error (Draper and Smith, 1966). This makes the computation of confidence intervals much easier.

Using mean values introduces an additional error about the mean for each station.

The model assumes that concentrations are uncorrelated with runoff volume. This assumption is supported by the results of the NURP study (EPA, 1983), which found no statistically significant relationship between storm volume and EMC.

Table 9.1: Summary of statistical parameters.

Constituents Variables	All EMCs		Site Means	
	Copper	Lead	Copper	Lead
Variables	4	4	4	4
Degrees of Freedom	89	89	11	11
R Squared	0.17	0.12	0.43	0.29
F-Value	4.67	3.15	2.06	1.13
Confidence Level	>99%	>95%	>75%	<75%

9.4 Results

As mentioned previously, the model formulation which was most successful was that using all sample data in a single regression analysis. This section describes the results of that model.

A box plot of the computed concentration of the two metals in storm water is displayed in Figure 9.2. The bars represent the mean estimate and the 90 percent confidence interval of the mean concentration by land use. The confidence interval assumes that the errors from the regression line follow a log-t distribution with 89 degrees of freedom. A log transform is used since measurements of concentration are bounded by zero on the low end but unbounded at the high end. A t-distribution is assumed since the true variance about the mean is unknown and can only be estimated by the variance of the sample data. The t-distribution approaches the normal distribution as the degrees of freedom approaches infinity.

Figure 9.2: Mean concentration and 90% confidence bounds by land use.

From the box plots we can see that there is considerable overlap between the confidence intervals for commercial and industrial land uses. This indicates that storm water concentration estimates for these two land uses are not significantly different. Statistically, there is more confidence in the differences between the mean values of open or residential areas versus the commercial and industrial areas.

Figure 9.3 provides a comparison of the results of the Alameda County model with NURP estimates and the results a study in Santa Clara Valley (WCC 1991a). The Alameda County model shows greater variation among land uses than the NURP or Santa Clara results. This is to be expected as the other studies were forced to assume that all sampling stations were homogeneous. The regression model needs no such assumption but rather can incorporate heterogeneous stations. Additionally, the other studies were unable to distinguish as many land use types. NURP did not include an industrial category and the Santa Clara report could not separate residential and commercial due to the intermingling of land use zones. All studies do mark a significant difference between urban and undeveloped land.

Table 9.2 provides a summary of the important parameters and the results of the model. We see that the proportion of runoff from different land use categories is quite different from their relative areas. Transportation was modeled in SWMM to compute runoff volumes, but combined with the commercial category to compute the water quality, as explained in the previous section. Figure 9.4 provides a graphical presentation of the load fractions. Although open areas are the largest of any land use, they produce less than 2 percent of the total pollutant load due to low runoff volume and pollutant concentration. Residential areas produce 63 percent of total runoff, but only 40 percent of total load for copper and 48 percent for lead. Transportation, commercial, and industrial areas combined represent 50 to 60 percent of the pollutant load due to high metals concentrations and high runoff volume, but represent only 20 percent of the urban land area. This indicates that targeting those areas for pollution control measures could present substantial cost

savings over region wide measures, while still addressing the bulk of the problem.

Table 9.2: Summary of parameters and results of the model.

	Open	Residential	Commercial	Transportation	Industrial	Total
Area (square miles)	190	112	13	7	24	345
% Area	55	32	4	2	7	100
Median Year Flow (af/year)	3000	48000	8100	5900	10200	76000
% Flow	4	63	11	8	13	100
Concentration (ug/l)						
Copper	12	18	54	54	51	
Lead	8	54	129	129	99	
Estimated Load (kg/year)						
Copper	44	1070	540	390	640	2685
Lead	30	3200	1290	940	1250	6710

9.5 Conclusions

From the results of this study we can conclude that there is a statistically significant relation between storm water concentrations of certain metals and land use. While the relationship is verifiable, it explains less than 20 percent of the variability in sampling data. Pollutant concentration is related to many other factors which were not analyzed in this study, for example, storm characteristics. Additionally, the relationship is stronger when comparing open or residential areas versus commercial or industrial areas. Differences between commercial and industrial areas are less apparent.

The MLR method presented for formulating regional water quality parameters should be explored further. This method allows selection of a wide range of sampling locations and therefore allows a greater portion of the study area to be represented in the load model. The method also provides a more complete error estimate. The analysis can easily be performed using inexpensive commercial software. While runoff volumes for this study were computed using a hydrologic computer model, this is not necessary for the method.

Until the other parameters governing storm water pollution are

Figure 9.3: Comparison of concentration with other studies.

Figure 9.4: Pollutant load from study area by land use type.

better understood, the MLR approach provides a relatively cost-effective way of computing loads based on water quality data collected from homogeneous and heterogeneous drainage areas.

9.6 References

Draper, N.R., and Smith, H. (1966). Applied Regression Analysis. John Wiley and Sons, Inc., New York, New York.

Gibra, I.N. (1973). Probability and Statistical Inference for Scientists and Engineers. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Gunther, Andrew (1991), The Loading of Toxic Contaminants to the San Francisco Bay-Delta in Urban Runoff. The Aquatic Habitat Institute, Richmond, California.

Mendenhall, W. (1987). Introduction to Probability and Statistics. Duxbury Press, Boston, Massachusetts.

Silverman, G.S., Stenstrom, M.K., and Farn, S. (1988). Land use Considerations in Reducing Oil and Grease in Urban Stormwater Runoff. J. Environmental Systems, Vol. 18(1):31 - 47.

United States Geological Survey, (1990). Techniques for Estimation of Storm-Runoff Loads, Volumes, and Selected Constituent Concentrations in Urban Watersheds in the United States. Water Supply Paper 2363, by Nancy E. Driver and Gary D. Tasker.

Woodward-Clyde Consultants, (Feb. 1991). Santa Clara Valley Nonpoint Source Study Volume 1: Loads Assessment Report. Submitted to Santa Clara Valley Water District.

Woodward-Clyde Consultants, (May 1991). Eugene Storm Drainage Water Quality Program, Loadings Report. Prepared for the City of Eugene, Oregon, Department of Public Works, Engineering Division.

Woodward-Clyde Consultants, (Oct. 1991). Alameda County Urban Runoff Clean Water Program Loads Assessment. Submitted to Alameda County Flood Control and Water Conservation District.