

APPENDIX I

MULTIVARIATE ANALYSIS

I. INTRODUCTION

While the tables presented in Section 9 provide a useful descriptive look at leaking tanks and conditions under which leaks occur, they do not take into account the simultaneous effects of many variables. To respond to this analytical need, multivariate statistical models have been developed to examine the relationship between leak status (1 = leak, 0 = no leak) [or leak rate (gallons per hour)] and various explanatory variables.

The advantage of the multivariate analysis is that it provides a method of assessing the contribution of individual explanatory factors, while simultaneously controlling for other variables. The procedures used also allow a step-wise approach (i.e., first finding the one variable that best predicts leak status [or leak rate], then the second best predictor, etc.) and a test for the statistical significance of coefficients of each variable in the model. The results of the multivariate analysis have been summarized in the next subsection so that the reader may learn the outcome of the multivariate analysis without having to go through all the mathematical details. The technical details on mathematical formulation can be found in later subsections, along with the final equations for the multiple regression and logistic regression models developed.

II. SUMMARY OF MULTIVARIATE ANALYSIS RESULTS

The major results of the modeling efforts are presented below. The reader should also note the caveates and limitations at the end of this summary.

A. Multiple Correlations

The multiple correlation coefficients (R) from the final regression models (which retained only variables with significant regression coefficients -- see Subsection C for confidence levels) were about .30 for leak status and .45 for leak rate, demonstrating low to moderate predictive ability. This corresponds to R^2 values of about .08 and .20, respectively. Since R^2 can be interpreted as the fraction of variance accounted for by the model, it is clear that the models do not account for most of the variance in leak status and leak rate.

B. Predictors of Leak Status

Based on the coefficients in the regression and/or logistic models, the probability that a tank system leak tends to increase for:

- o Older tanks,
- o Tanks with no leaded gasoline stored,
- o Tanks with passive cathodic protection, and
- o Tanks for which no log of deliveries is kept.

The positive relationship between leak probability and passive cathodic protection might seem surprising. A possible explanation is that passive cathodic protection tends to be used in areas which have a history of corrosion/leak problems. Another explanation could be that passive cathodic protection is strongly correlated with the storage of aviation fuel and, thus, might be a proxy for this fuel type. (The multivariate model equations for leak status may be found in Section III, which follows.)

C. Predictors of Leak Rate

Among leaking tank systems, the leak rate tends to be larger for:

- o Fiberglass tanks;
- o Tanks not on a concrete pad;
- o Tanks both old and steel (i.e., an interaction effect)*;
- o Tanks attached to other tanks; and
- o Tanks in establishments with operators trained to check for line leaks.

The above factors are not indicators of leak likelihood, but of larger leak rates among leaking tank systems. The last factor may well be a case of reverse causality -- i.e., where tank systems leak heavily, operators are trained to detect line leaks (rather than vice versa).

*More precisely, fiberglass tank systems show less increase in leak rate as they get older.

D. Limitations and Caveats

In addition to the comments about the limitations of the scope of the study presented in Section 8, the following limitations and caveats apply to the multivariate analysis:

- o Only business, government and military sectors are included (no farms).
- o Manifoldd tanks that could not be separated for tightness tests are not included.
- o Although a long list of 49 potential explanatory variables were included, there are other possible variables which were not in our data base and whose effects are, therefore, not accounted for. In particular, soil characteristics were not available for analysis and use in the models. However, backfill around the tank (e.g., sand/gravel) is included and may be more relevant.
- o The multivariate analysis finds "measures of association" rather than causality. Naturally, since the variables used were suspected of affecting leaking, the discovery of a statistically significant association tends to affirm a causal linkage. But the reader is cautioned that a different covariate could be the real causative factor, as in all statistical correlation studies. For example, the variable "age of tank" could represent the effects of aging, per se, or age of tank could be a proxy for different installation techniques which changed over time, or different resins used in the manufacture of fiberglass tanks in different production years.

III. MULTIVARIATE MODEL DEVELOPMENT PROCEDURE

A. Overview

Two regression models (one to predict leak status and one to predict leak rate) were developed using the variables in Table I-1 as candidate predictor variables. (Table I-1 also appears as Table 9-31 in Section 9 of this report.) The regression analysis followed a number of preliminary steps before arriving at the final models. This included elimination of variables with too many missing variables (X_{13} , X_{16} , X_{18}) and variables with nearly constant values (X_8 , X_9 , X_{21} , X_{23}). Stepwise regression runs were made to obtain a reduced set of variables which best predicted leak status or leak rate. Finally, individual regression coefficients were examined to ensure statistical significance. Sample sizes are shown below for the final model.

<u>Model</u>	<u>Sample Size</u>
Leak Status Regression	327
Leak Status Logistic	380
Leak Rate Regression	99

Table I-1. Simple Correlation of Leak Status with Explanatory Variables

Explanatory Variable	Meaning	Definition	Correlation ⁽¹⁾ with Y1, Leak status (1 = Leak; 0 = No Leak)	Correlation ⁽¹⁾ with Y2, Leak rate (gal/Hr), among leaking tanks ⁽²⁾
X1	Gas Station	1 = Yes; 0 = No	-.08	-.06
X2	# Underground tanks	Number at facility	.12	.10
X3	Tank capacity	Gallone	.14	.34
X4	Average low fill level ⁽³⁾	As fraction of tank capacity	-.05	-.07
X5 ²	(Age of tank) ²	in (years) ²	.11	-.20
X6	Leaded gasoline	1 = yes; 0 = No	-.26	-.11
X7	Diesel fuel	1 = Yes; 0 = No	.24	-.08
X8	Aviation fuel	1 = Yes; 0 = No	.13	.07
X9	Gasohol	1 = Yes; 0 = No	-.07	0
X10	Other	1 = Yes; 0 = No	.08	.29
X11	Suction pump	1 = Yes; 0 = No	.003	-.12
X12	Depth buried	Inches from surface to top of tank	.10	-.006
X13	Water level	Inches from surface to water table ⁽⁴⁾	-.15	-.005
X15	Tank tested	1 if tested after placed in service; 0 otherwise	.03	.01
X16	Years since test	Since most recent test	.06 ²	-.21
X17	Tank material	1 = steel; 0 = fiberglass	.02	-.09
X18	Tank lined	1 = Yes; 0 = No	.07	.02
X19	Tank coated	1 = Yes; 0 = No	-.01	-.25
X20	Passive cathodic protection	1 = Yes; 0 = No	.10	.05
X21	Impressed current cath. protection	1 = Yes; 0 = No	0	0
X23	Other protection	1 = yes; 0 = No	-.08	0
X24	Previous tank leak	1 = Yes; 0 = No	-.05	-.04
X25	Previous line leak	1 = Yes; 0 = No	.05	.23
X26	Frequency of deliveries	Number per year	-.05	-.003
X27	Sand fill	1 = Yes; 0 = No	.03	-.10
X28	Gravel fill	1 = Yes; 0 = No	.006	.16
X29	Concrete pad	1 = Yes; 0 = No	.07	-.09
X30	Packed earth pad	1 = Yes; 0 = No	.03	-.09
X31	Dist. to nearest tank or structure	(feet)	-.04	-.09

¹Pearson's correlation coefficient; Kendall's tau-B was also calculated for all Y1 correlations and found to be the same for nearly every variable.

²Using data only from individual leaking tanks with quantifiable leaks.

³I.e., just before product is added.

⁴At time of test.

Table I-1. Simple Correlation of Leak Status with Explanatory Variables
(Continued)

Explanatory Variable	Meaning	Definition	Correlation ⁽¹⁾ with Y1, Leak status (1 = Leak; 0 = No Leak)	Correlation ⁽¹⁾ with Y2, Leak rate (gal/Hr), among leaking tanks ⁽²⁾
X32	Interaction: age & material	(X5) (1-X17)	-.03	-.07
X33	Interaction: gasohol & material	X9 (1-X17)	0	0
X34	Permit to install	1 = Yes; 0 = No	.12	.17
X35	Permit to store	1 = Yes; 0 = No	.02	.09
X36	Average high fill level ⁽⁶⁾	As fraction of tank capacity	-.06	-.09
XT3	Average fuel delivery	in gallons (to one tank)	.15	.23
XT4	Max. ever stored	gallons	.11	.29
XT18A	Attached to other tank	1 = Yes; 0 = No	.22	.24
XT19	Tank proximity to water table	1 = above; 2 = partially above; 3 = below; 4 = other	.13	.28
XT20	Manway with tank	1 = Yes; 0 = No	.19	.13
XT36	Not self-installed	1 = Yes; 0 = No	.12	.12
XB5	Remote gauge	1 = Yes; 0 = No	-.005	.05
XB19	Log of deliveries	1 = Yes; 0 = No	-.03	.002
XC7	Any abandoned tank ⁽⁵⁾	1 = Yes; 0 = No	-.03	.03
XC8	# Abandoned tanks	(coded as zero if none)	.12	-.09
XF1A	Corrosion prevention equip./mat.	1 = Yes; 0 = No	-.02	-.12
XG2D	Trained to check pump	1 = Yes; 0 = No	.14	.24
XG2E	Trained to check line leaks	1 = Yes; 0 = No	.10	.18
XG2F	Trained to check leak prevention	1 = Yes; 0 = No	.10	.15
XG2G	Trained to check leak monitoring	1 = Yes; 0 = No	.15	.17

⁵At that facility.

⁶I.e., Just after product is delivered.

B. Multiple Regression Models

Two models were constructed:

- | | |
|---|--|
| [1] Leak Status Model:
(among all tanks
with tightness test) | Dependent Variable, Y1 =
1 if leak
. 0 otherwise |
| [2] Leak Rate Model:
(among <u>leaking</u>
tank systems only) | Dependent Variable, Y2 =
leak rate in gal/hr |

Both models were run using the predictor variables in Table I-1. The general form of the model is:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots$$

where a few of the variables were interaction terms and the b's are regression coefficients estimated by a least-squares procedure. In addition, a non-linear transformation was used for one of the X variables. Age² was used rather than Age because data plots suggested a non-linear increase in the percentage of tanks that leak as a function of age.

C. Logistic Regression Model

For the leak status model, an alternative logistic regression model was run. The dependent variable can be reexpressed as an odds ratio*, in the form:

$$[1a] \log \frac{\text{Probability of Leaking Tank}}{\text{Probability of Tight Tank}} = b_0 + b_1X_1 + b_2X_2 + \dots$$

This alternative formulation of Model [1] should more nearly satisfy the homogeneity of variance assumption for regression.

The coefficients (b's) for the Logistic Model are estimated by maximum-likelihood methods rather than least-squares.

IV. FINAL MULTIVARIATE MODELS

Using the procedures defined above, linear and logistic regression models were developed for leak status. For leak rate, a separate linear regression model was developed. The final models appear below.

*The assumed underlying model for the logistic regression is $Y = 1 / [1 + \exp(-b_0 - b_1X_1 - b_2X_2 - \dots)]$. From this expression it can be shown that $\log [Y / (1 - Y)] = b_0 + b_1X_1 + b_2X_2 + \dots$. In this equation Y is the probability that the tank system leaks and 1 - Y is the probability that it does not leak.

Leak Status Models

[1] Regression Model*:

$$Y_1 = .22 + .00019 X_5^2 - .25 X_6 + .0044 X_{12}^{***} + .18 X_{20}$$

[1a] Logistic Model****:

$$\log \frac{\text{Probability of Leak}}{\text{Probability no Leak}} = 1.3 - .63 X_6 - .017 X_{12} - .38 X_{B19}$$

*All coefficients significant at the 94 percent confidence level or better (except coefficient of X_{20} at 78 percent confidence level).

** (Age)² was used rather than Age because this non-linear transformation showed a stronger correlation with leak status.

***The regression model found a + coefficient, but the logistic model found a - coefficient. This may be a case of X_{12} 's collinearity with other variables. However, no strong collinearities were detected with X_{12} . (See Tables I-2 and I-3 in Section V.) Therefore, the relationship with X_{12} , depth tank is buried, is inconclusive based on this mixed result.

****All coefficients significant at the 94 percent confidence level or better.

[2] Leak Rate Model*****:

$$Y_2 = .91 - .67 X_{17} - .54 X_{29} - .0068 X_{32} \text{*****}$$

$$+ .62 X_{T18A} + .25 X_{G2E}$$

The reliability of the model was examined in several ways. For the regression models, the multiple correlation coefficient, R, provides some overall measure of the predictive ability of the model. These results are shown below.

Equation	Multiple Correlation Coefficient, R		R ²	
	Unadjusted	Adjusted	Unadjusted	Adjusted
[1]	.30	.29	.093	.081
[2]	.50	.45	.25	.20

*****All coefficients significant at the 97 percent confidence level or better.

*****This is an interaction term which was included to capture the more than additive effect of age and material type together.

The "adjusted" values of R and R^2 adjust for degrees of freedom in the model and, therefore, provide a better estimate of how reliably the model might predict leak status and leak rate for other tank systems beyond the modeling data set. The R^2 term can be interpreted as the proportion of the variance in Y that can be explained for by the model. Thus, the model is able to account for less than 10 percent of the total variance in leak status and only about 20 percent the variance in leak rate.

The reliability of the coefficients of the X's in equations [1], [1a] and [2] were also examined to ensure that the value is not likely to be a chance occurrence. The probability that these coefficients are not chance occurrences is 94 percent or more for each of 9 of the 10 parameters in these equations. The remaining coefficient had a 78 percent probability of being a non-chance occurrence (i.e., there is a very low probability of the observed coefficient occurring if its true value were zero). It should be noted that these probabilities of non-chance occurrence applies one variable at a time -- i.e., with many variables tried in the model, the probability of at least one chance selection of a variable increases.

V. RELATIONSHIP BETWEEN EXPLANATORY VARIABLES
(COLLINEARITY)

Multicollinearity frequently exists in large data sets. Pairwise collinearity is one sample form, and is relatively easy to visualize. In order to test for such "first order" collinearity in the models, the correlations between all pairs of independent or predictor variables (i.e., X's) were computed. The results shown in Table I-2 indicate low pairwise collinearity, except for X_{17} (tank material) and $X_{32} = [(1 - \text{tank$

Table I-2. Collinearity (intercollelation) of X's in models

A. Leak status regression and logistic models --
Pearsons Correlation Coefficient between
explanatory variables

	X ₅ ²	X ₆	X ₁₂	X ₂₀	X _{B19}
X ₅ ²	1	-.03	-.07	-.08	.10
X ₆		1	-.06	-.12	.002
X ₁₂			1	.07	.09
X ₂₀				1	-.04
X _{B19}					1

B. Leak rate regression model -- Pearson's Correlation
Coefficient between explanatory variables

	X ₁₇	X ₂₉	X ₃₂	X _{T18A}	X _{G2E}
X ₁₇	1	.09	-.80	.13	.05
X ₂₉		1	-.07	.38	.08
X ₃₂			1	-.10	-.11
X _{T18A}				1	-.02
X _{G2E}					1

material) x (Age)²] in the leak rate model (correlation of $-.80$). The variable, X_{32} , is an interaction term. The correlation of X_{17} with X_{32} is close to the correlation of Age² with $-Age^2$. Therefore, a large intercorrelation would be expected.

Table I-3 shows correlations between variables in the models and variables not in the models. (Variables with small correlations, less than $.20$, are not included.) Any large correlations could be considered as proxies (or substitutes) for the model variable with which they are strongly correlated. For example, in the leak status model, passive cathodic protection (X_{20}) is strongly correlated (correlation coefficient = $.62$) with aviation fuel (X_8). Therefore, the apparent increase in the likelihood of a leak with passive cathodic protection, might be due, in large measure, to its relationship with aviation fuel storage.

Table I-3. Correlation Between Model X's and X's not in the Model

A. Leak Status Model

Model X	Non Model X's	Pearson's Correlation Coefficients ($\geq .20$)
X ₅ ² , (Age of Tanks) ²		None
X ₆ , Leaded gasoline	X ₇ (Diesel fuel)	-.39
X ₁₂ , Depth buried		None
X ₂₀ , Passive cathodic	X ₂ (# Underground tanks)	.33
	X ₈ (Aviation fuel)	.62
	X ₁₈ (Tank lined)	.34
	X ₂₉ (Concrete pool)	.38
	X _{T18A} (Attached to other tank)	.29
	X _{T20} (Manway with tank)	.41
	X _{G2E} (Trained to check line leaks)	.24
	X _{G2F} (Trained in leak protection)	.27
	X _{G2H} (Trained in leak monitoring)	.31
X _{B19} , Log of deliveries	X ₁₃ (Water level)	.30
	X ₁₆ (Years since test)	.34
	X ₃₄ (Permit to install)	.20
	X ₃₅ (Permit to store)	.20

B. Leak Rate Model

Model X	Non Model X's	Pearson's Correlation Coefficients ($\geq .20$)
X ₁₇ , Tank material	X ₁ (Gas station)	-.21
	X ₇ (Diesel fuel)	.22
	X ₁₁ (Suction pump)	.42
	X ₁₃ (Water level)	-.29
	X ₁₅ (Tank tested)	-.28
	X ₁₆ (Years since test)	-.37
	X ₁₈ (Tank lined)	-.35
	X ₁₉ (Tank coated)	.66
	X ₃₂ (Interaction: Age ² & material)	-.80
X ₂₉ , Concrete pad	X ₂ (# Underground tanks)	.46
	X ₄ (Average low fill level)	.24
	X ₁₆ (Years since test)	-.48
	X ₂₀ (Passive cathodic protection)	.26
	X ₃₀ (Packed earth pad)	-.20
	X ₃₄ (Permit to install)	.24
	X ₃₆ (Average high fill level)	.28
	X _{T3} (Average fuel delivery)	.20
	X _{T18A} (Attached to other tank)	.38
	X _{T20} (Manway with tank)	.52
	X _{G2H} (Trained in leak monitoring)	.24
X ₃₂ , Interaction: Age ² & material	X ₁₁ (Suction pump)	-.29
	X ₁₅ (Tank capacity)	.26
	X ₁₆ (Years since test)	.49
	X ₁₇ (Tank material)	-.80
	X ₁₈ (Tank lined)	.53
	X ₁₉ (Tank coated)	-.54
X _{T18A} , Attached to other tank	X ₂ (# underground tanks)	.48
	X ₃ (Tank capacity)	.22
	X ₇ (Diesel fuel)	.23
	X ₁₃ (Water level)	-.30
	X ₁₆ (Years since test)	-.23
	X ₂₅ (Previous line leak)	.28
	X ₂₉ (Concrete pad)	.38
	X ₃₀ (Packed earth pad)	-.29
	X ₃₄ (Permit to install)	.29
	X ₃₅ (Permit to store)	.25
	X ₃₆ (Average high fill level)	.25
	X _{T3} (Average fuel delivery)	.35
	X _{T4} (Maximum ever stored)	.33
	X _{T20} (Manway with tank)	.40
X _{G2D} (Trained to check pump)	.24	

Leak Rate Model (Continued)

Model X	Non Model X's	Pearson's Correlation Coefficients ($\geq .20$)
X _{G2E} , Trained to check line leaks	X ₂ (# underground tanks)	.25
	X ₇ (Diesel fuel)	-.25
	X ₈ (Aviation fuel)	.25
	X ₁₀ (Other fuel)	.41
	X ₁₆ (Years since test)	-.44
	X ₂₈ (Gravel fill)	.21
	X _{T19} (Tank proximity to water table)	.39
	X _{T20} (Manway with tank)	.22
	X _{G2D} (Trained to check pump)	.40
	X _{G2F} (Trained in leak protection)	.89
	X _{G2H} (Trained in leak monitoring)	.68