7/08 Quarterly Report
UST 2004 Facility Baseline Regression Analysis and Cluster Analysis Update
University of Rhode Island Department of Computer Science and Statistics
Professor R. Choudary Hanumara rch@cs.uri.edu

Out of 97 facilities sampled, data on 87 facilities were used in the facility-level regression analysis. Descriptive data on 10 facilities could not be collected as they were closed, changed operations, or for other reasons. Not all variables (inspection and descriptive data) were recorded at each of the facilities so there are unequal observations. The table below gives a summary profile of the facilities used in the regression analyses.

Descriptive Statistics

Variable	**Frequency**	**Average**	**Percentage**
Facility Type			
Large convenience store	40		46.0
Small snack express	28		32.2
Only selling gas	19		21.8
Corporate oversight			
Yes	58		66.7
No	29		33.3
No. employees	87	5.9	
Ownership			
Independent	47		54.0
Otherwise	40		46.0
Franchise			
Yes	25		28.7
No	62		71.3
Full service garage			
Yes	27		31.0
No	60		69.0
Age of tank	86	16.5	
Capacity of tanks	87	8942	
Tank make			
All SW	31		35.6
All DW	52		59.8
Mixed	4		4.6

Number of tanks		
1	2	2.3
2	19	21.8
3	44	50.6
4	13	14.9
5	5	5.7
6,7	4	4.6
Tank manifolded		
All yes	1	1.2
All no	54	64.3
Mixed	29	34.5
Piping		
All SW	25	30.5
All DW	50	61.0
Mixed	7	8.5

The compliance status of each facility was assessed on a number of variables (checks) under the general categories: facility profile, tank profile, tank leak detection, piping leak detection, spill prevention and vapor recovery. The average number of checks (FY1) and instances of non-compliance (FY2) were 196.4 and 22.3, respectively.

The University of Rhode Island Department of Computer Science and Statistics investigated four regression models to analyze the data on non-compliance and the variable data on facilities. The independent variables are listed in the table above and the qualitative variables were introduced as dummy variables in the modeling process. The dependent variables in the first two standard linear regression models were FY3=FY2/FY1 and FY4=Arcsin(FY3**.5). Since the non-compliance data FY2 is count by nature, the next two modeling techniques used were Poisson and Negative binomial regression. There was evidence of over dispersion—deviance/degrees of freedom was much larger than (1) in the Poisson regression model. Hence, Negative binomial regression is a better choice over Poisson regression. A fit of Negative binomial regression indicated that the only significant independent variable is the average capacity of the tanks at the facility. The same variable was also significant in the Poisson regression model. In the standard linear regression modeling with FY3 as the dependent variable, the only significant variable was whether the facility was a franchise or not. Using FY4 as the dependent variable, the significant independent variables were the franchise and the piping status of the tanks. The coefficient of determination (R**2) in the linear regression modeling was only .11.

Cluster Analysis Update

In Table 1 of the last quarterly report, 63 variables were indicated to be potentially measurable performance improvement indicators. For each of these variables, the number of tanks observed, proportion (p) of compliance, and the confidence interval

on the true proportion of compliance were given. The 95% confidence interval was obtained by the standard formula (p-1.96*(p*(1-p)/n)**.5, p+1.96*(p*(1-p)/n)**.5) where n is the number of tanks observed. In applying this formula, the assumptions were (1) random selection of tanks, and (2) that the sample size n was sufficiently large for the normal approximation. In our data collection efforts, the facilities were selected randomly and all the tanks within each facility were observed. Thus the sampling design was cluster sampling violating assumption 1. In this case, Donner and Klar (Design and Analysis of Cluster Randomization Trials in Health Research, Arnold publishers, 2000) give a modified formula for confidence interval as (p-1.96*(p*(1-p)*(1+(m-1)*r)/n)**.5, p+1.96*(p*(1-p)*(1+(m-1)*r/n)**.5) where m is the average number of tanks per facility and r is the intraclass correlation coefficient indicating the similarity of responses on the tanks within a facility. To illustrate this, we looked at the data on variable "System calibrated (E12)." For this variable, the number of tanks observed was 192. The standard 95% confidence interval for the true proportion of compliance was (.60, .74). The values of m and r were computed to be 2.95 and 1, respectively. Applying the formula modified for cluster sampling, the 95% confidence interval on the true proportion of compliance is (.56, .78). As r is usually positive, the confidence interval based on cluster sampling is wider than the one obtained through the standard Wald method. The sample size n for this variable is large for the normal approximation to hold. In cases when n is small (n*p is less than 9) Agresti and Coull (American Statistician, May 1998) and other researchers proposed an adjusted confidence interval. We will take this into account in dealing with small n in future work.