

A Model for Predicting the Occurrence of Cluster Blocks  
Based on 1990 Census Data for King County

By  
Jeremy Verlinda

Prepared for  
Professor Dick Startz  
Economics 482

30 November, 1998

42 - a really well done paper!

## **INTRODUCTION**

Over the course of the last year, I had the opportunity to perform some research in the Department of Urban Planning and Development. I worked with two people, Professor Anne Moudon and Paul Hess, a Ph.D. student, who were concerned with a phenomenon they were calling “clusters.” These clusters are similar to the press’ term “Urban Villages,” which are the formation of urban-like blocks in suburban neighborhoods. They are characterized by high ratios of apartment housing and close proximity to commercial and industrial zones. The term cluster came about because they were not just concerned with individual blocks but clusters of blocks that share these properties. A map of these cluster blocks is provided in Figure 1. For Paul and Anne, the main goal of their work is to better define these clusters and to understand their attributes for the main purposes of publishing an article and/or writing a book. For my part, I wanted to help them with, among other things, the statistical side of their research. Recently I have had the opportunity to ask how we can predict the occurrence of these clusters based on statistical data. My goal for this project is to come up with an econometric model that answers this question. Based on census block data, I want to be able to use certain variables to predict the likelihood of a given block (or set of blocks) to be considered part of a cluster.

## **CENSUS BLOCK DATA**

The data set for this research comes from the 1990 US Census. It is organized at the block level and reveals statistics such as block size, number of housing units, total population, housing type, and demographics of the population such as age group and

# Clusters & Suburban Blocks in Puget Sound

- Cluster Block
- Suburban Blocks
- Puget Sound

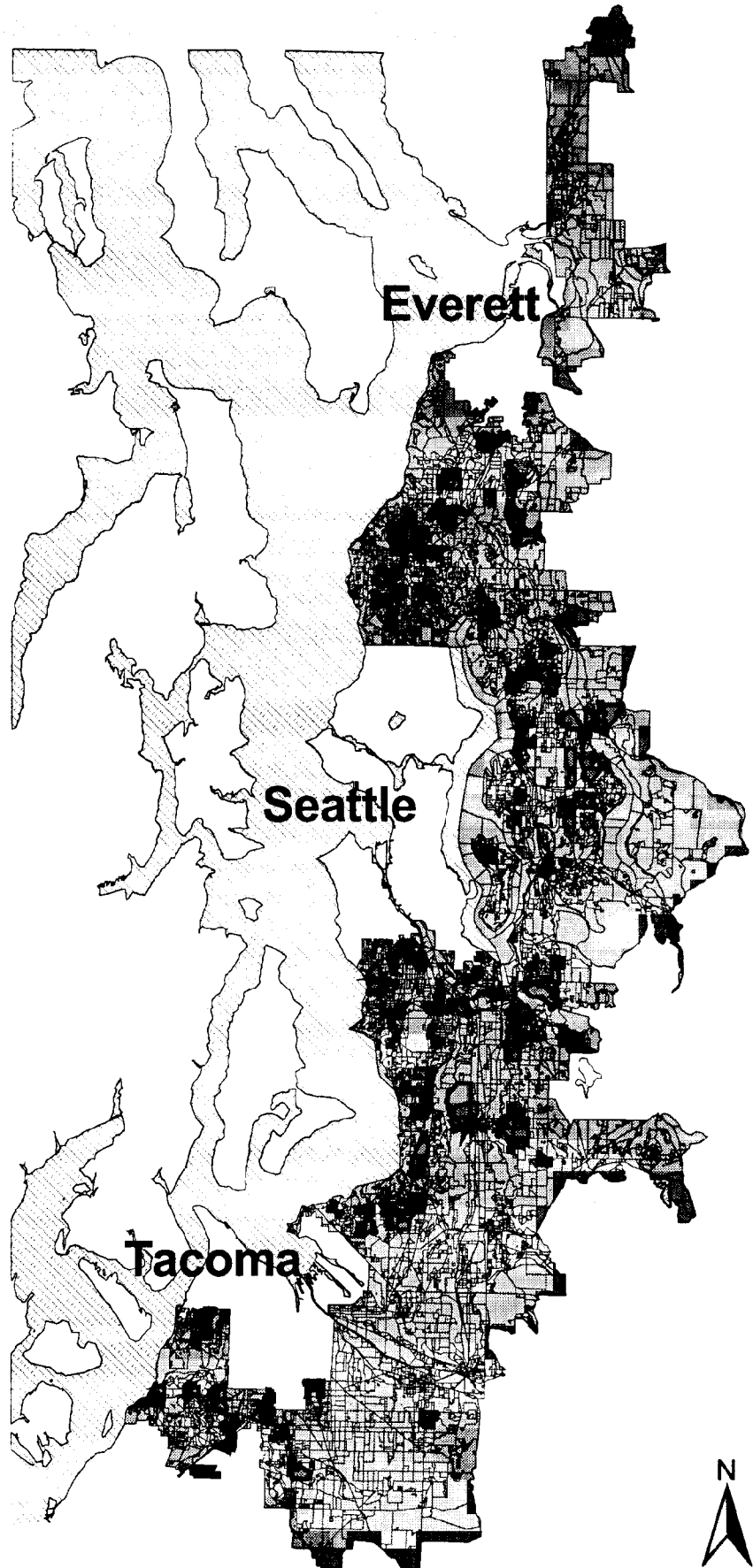


Figure 1

ethnicity. This data can be manipulated to reveal statistics like percentage of apartment housing, percentage of nonwhite persons, and population density. Since I have data for King, Pierce, and Snohomish counties separately, one of the goals of the econometric model is to use King County as a base and then apply the model to Snohomish and Pierce counties in order to critique its accuracy as a predictor.

### THE LOGIT MODEL

There are essentially two key benefits to be derived from creating a quantitative model of the clusters. The most obvious is that it can serve as a predictor tool. An effective quantitative model can be applied to regions outside the Puget Sound for quick identification of existing clusters. The other benefit is that it helps to identify and describe the object of study. In this case, the clusters were originally identified by a relatively complex, qualitative review of the region. While reasonably effective, such a method also makes it difficult to describe to others the phenomena under study. A general rule derived from the quantitative model is that a cluster can be identified, albeit roughly, as a conglomeration of suburban census block with percentages of non-single family units in excess of 30%.

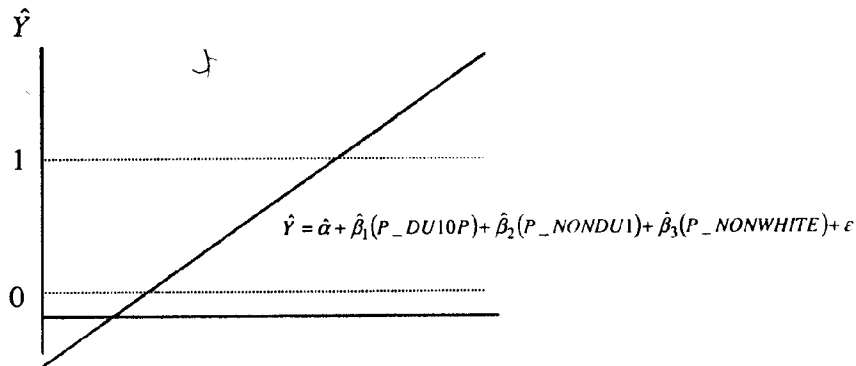
great

Econometrics offers a common model for predicting outcomes based on a given set of data that most people recognize as ordinary least squares (OLS) regression analysis. An example of an OLS model for this study would take the form:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1(P\_DU10P) + \hat{\beta}_2(P\_NONDU1) + \hat{\beta}_3(P\_NONWHITE) + \dots + \varepsilon,$$

where P\_DU10P, P\_NONDU1, and P\_NONWHITE represent percentage of apartments, percentage of non-single family housing, and percentage of non-whites, respectively.

$\hat{Y}$ ,  $\hat{\alpha}$ , and  $\hat{\beta}_i$  represent estimates of whether or not the block is part of a cluster, the value of some constant term, and the weight given to the value for each variable, also respectively. Since  $\hat{Y}$  can only take two values, 0 if it is not part of a cluster, 1 if it is, we call this a dummy variable. This can be shown graphically as follows:



Unfortunately, the OLS model has a few problems with dummy variable that are made more obvious by careful inspection of the graph above. The first is that it allows for potential values outside of the 0 to 1 bound, and the other major one is the linear assumption for the relationship between  $\hat{Y}$  and the explanatory variables. For these reasons, most discrete choice models rely more heavily on logit or probit analysis. Since this study is based on the logit model, this discussion will exclude any detail on the probit model, although the characteristics of both are similar.

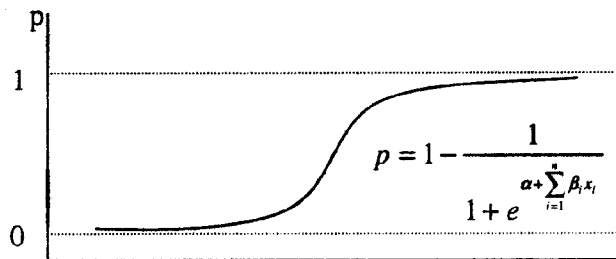
The logit model is distinct from the OLS model in two key ways. The first difference is that it does not yield the value of the dummy variable, 0 or 1, but rather the ratio of the probability that the dummy variable equals 1 to the probability that it does not. The logit model also recognizes that the relationship is likely not linear and so is exponential in form. In general, the model takes the form:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^n \beta_i x_i,$$

which can be manipulated to solve for p as:

$$p = 1 - \frac{1}{1 + e^{\alpha + \sum_{i=1}^n \beta_i x_i}},$$

and represented graphically as follows:



Beyond the characteristics already described about the logit model, there are two other benefits from this distribution. Looking at the graph above, we see that the curve approaches 100% probability and 0% probability asymptotically. Another quite significant benefit is that, unlike the least squares model, we can introduce sampling bias in this model without adversely affecting the accuracy of the model. The reason for this is that if you are working with a data set similar to the census data that is in my project, a graph of the true model (i.e., one which perfectly predicted the occurrence of clusters) would have about 7500 values clustered around probability 0 and only 600 values clustered around probability 1. Any distribution that tried to capture this along an exponential s-curve as shown in the graph above would be skewed downward for the large sample of non-cluster blocks. For this reason, I chose a sample base of 1373 blocks, 686 of which are cluster blocks (all of the cluster blocks in King County) and 687

I'm not sure about this

I don't think so, but I could be wrong

randomly selected blocks from the remaining 7500 blocks in King County. This effectively forced the distribution more evenly along the curve and significantly improved the results of the model.

## RESULTS OF THE MODEL

Initially, I selected percentage of apartments, percentage of non-single family units, percentage of non-whites, and population size as the explanatory variables. After running the model, I found that each variable was statistically significant, with the lowest z-statistic of 3.445 for population size. The regression output for this model is supplied in Figure 2 below.

Figure 2. Regression Output for All Variables

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-1.10567	0.144874	-7.631926	0
P_NONDU1	0.020097	0.002649	7.588135	0
P_NONWHT	0.023086	0.005674	4.068936	0
POP100	0.001332	0.000387	3.444868	0.0006
POPDEN	0.037753	0.010633	3.550681	0.0004
P_DU10P	0.020181	0.005112	3.947686	0.0001
P_LS16YR	-0.024918	0.006551	-3.803576	0.0001
Mean dependent var	0.499636	S.D. dependent var	0.500182	
S.E. of regression	0.400909	Akaike info criterion	0.989131	
Sum squared resid	219.5548	Schwarz criterion	1.015768	
Log likelihood	-672.0381	Hannan-Quinn criter.	0.999098	
Restr. log likelihood	-951.6907	Avg. log likelihood	-0.489467	
LR statistic (6 df)	559.3052	McFadden R-squared	0.293848	
Probability(LR stat)	0			
Obs with Dep=0	687	Total obs	1373	
Obs with Dep=1	686			

With logit models  $R^2$  is not necessarily the best method to estimate the accuracy of the model, and is better measured by the percentage of correct predictions. I chose a standard cutoff level for the probability at 0.5. For these variables, I found an accuracy of 67.8% correct predictions for cluster blocks and 87.63% for non-cluster blocks, with an overall accuracy of 78.22%. Of these variables, the most influential was percentage of non-single family. I then reran the model with percentage of non-single family units as the sole explanatory variable. The regression output for this model is listed below in Figure 3.

Figure 3

---

Dependent Variable: CLUSTER				
Method: ML - Binary Logit				
Date: 11/24/98 Time: 13:21				
Sample: 1 1373				
Included observations: 1373				
Convergence achieved after 3 iterations				
Covariance matrix computed using second derivatives				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-1.058574	0.079905	-13.2479	0
P_NONDU1	0.035664	0.001975	18.05451	0
Mean dependent var	0.499636	S.D. dependent var	0.500182	
S.E. of regression	0.410652	Akaike info criterion	1.042102	
Sum squared resid	231.1983	Schwarz criterion	1.049712	
Log likelihood	-713.4027	Hannan-Quinn criter.	1.044949	
Restr. log likelihood	-951.6907	Avg. log likelihood	-0.519594	
LR statistic (1 df)	476.576	McFadden R-squared	0.250384	
Probability(LR stat)	0			
Obs with Dep=0	687	Total obs	1373	
Obs with Dep=1	686			

Our test results improved noticeably, achieving an accuracy of 68.22% correct predictions for cluster blocks and 88.65% for non-cluster blocks. However, close scrutiny of the regression output does show that the R-squared has fallen, which suggests

that the first regression was better. In fact, each of the explanatory variables are correlated and significant when regressed in the logit function against the probability of being a cluster. This led me to ask the question of why the model with just one variable seemed to be a better predictor for the occurrence of cluster blocks. Unable to answer this question concretely (i.e., mathematically or intuitively), I decided to proceed to the next phase of my project, which is to test the model in other counties. I compared the accuracy of each model and found that while the single explanatory variable is more accurate for King County, the model with all of the variables was the most accurate for Snohomish and Pierce Counties (using the values determined from the King County regression). The following tables describe the accuracy of the models for each county.

**Table 1: Snohomish Percentage of Correct Predictions (all variables including children)**

	Clusters	Nonclusters	Totals
Predicted Yes	132	2266	2398
Really Yes	229	2533	2762
% Correct	57.64%	89.46%	86.82%

**Table 2: Snohomish Percentage of Correct Predictions (Just p\_nondu1)**

	Clusters	Nonclusters	Totals
Predicted Yes	140	2109	2249
Really Yes	229	2533	2762
% Correct	61.14%	83.26%	81.43%

**Table 3: Pierce Percentage of Correct Predictions (all variables including children)**

	Clusters	Nonclusters	Totals
Predicted Yes	118	2734	2852
Really Yes	165	3212	3377
% Correct	71.52%	85.12%	84.45%

Table 4: Pierce Percentage of Correct Predictions  
(Just p\_nondu1)

	Clusters	Nonclusters	Totals
Predicted Yes	122	2528	2650
Really Yes	165	3212	3377
% Correct	73.94%	78.70%	78.47%

### Interpreting the Results

The final form of the model with just non-single family units looks like:

$$\ln\left(\frac{p}{1-p}\right) = -1.058 + .0357 * (P\_NONDU1),$$

which can also be written:

$$p = 1 - \frac{1}{1 + e^{-1.058 + .0357 * (P\_NONDU1)}}.$$

The usefulness of this becomes more clear if we consider that by substituting 0.5 for p, we can solve for P\_NONDU1 and use this value as a query in a GIS program such as ArcView to visually interpret the accuracy of the model. Actually performing the math reveals that for values of 29.64% or greater, p will take on values of .5 or more.

Unfortunately, trying to perform a similar operation with the model for all variables is not so easy or intuitive. This becomes clear if we look at what happens when we try to determine the marginal effect on the probability for a one unit increase in a given variable, (e.g., a 1% increase in apartment housing). The partial derivative with respect to p\_du10p is:

$$\frac{\partial p}{\partial p\_du10p} = \beta_{p\_du10p} \left( \frac{e^{\beta}}{1 + e^{\beta}} \right),$$

where  $\phi = \alpha + \sum_i^n \beta_i X_i$  and  $X_i$  represents the census data variables. The main thing to note from this is that the partial derivative of probability with respect to a given variable varies with the value of that variable, as well as the values for each of the other variables. With this difficulty in mind, it should be obvious that, as in the example above we were able to derive a cutoff value for  $p\_nondu1$  that is useful in extended application of the model, we cannot do this for the multiple variable model. Expediency then would suggest that despite the accuracy loss for the single variable model, in certain applications the loss of accuracy is acceptable given the gain in usefulness. However, as a tool for predicting the occurrence of cluster blocks, the multiple variable model is preferable.