Title

Estimating the Homeless Population in Los Angeles: An Application of Cost-Sensitive Stochastic Gradient Boosting

Authors

Kriegler, Brian
Berk, Richard A

Publication Date

2007

Estimating the Homeless Population in Los Angeles: An Application of Cost-Sensitive Stochastic Gradient Boosting*

Brian Kriegler and Richard Berk
Department of Statistics, UCLA

October 1, 2007

Abstract

In many metropolitan areas, efforts are being made to count the homeless to ensure a proper provision of social services. Some areas are very large, which makes spatial sampling a viable alternative to an enumeration of the entire terrain. Consequently, counts are manually observed in sampled regions, but they must be imputed in unvisited areas. Along with the imputation process, the costs of underestimating and overestimating may be weighted distinctly depending on one's perspective and what is at stake. Here, we analyze data from the 2004-2005 Los Angeles County homeless study using a variant of stochastic gradient boosting that allows for asymmetric costs. Specifically, we demonstrate how to boost the quantile distribution, which exhibits a straightforward translation to error estimation costs. Imputed counts and model diagnostics using various cost functions are reported. Practical usage of cost-sensitive imputed estimates are discussed briefly.

Key Words: boosting, quantile regression, statistical learning, costs, small area estimation

*We gratefully acknowledge numerous helpful discussions with Greg Ridgeway. The work of Brian Kriegler and Richard Berk was supported by the National Science Foundation under grant SES-0437179, "Ensemble Methods for Data Analysis in the Behavior, Social, and Economic Sciences."

1 Introduction

In many metropolitan areas, efforts are being made to count the homeless. Estimates of the number of homeless individuals are needed to ensure a proper provision of social services. Over the past two decades, homeless counts have been undertaken in Chicago, New York, Phoenix, New Orleans, and Los Angeles, and the list continues to grow.

In a typical census design, individuals are contacted through their place of residence. With the possible exception of homeless individuals living on private properly, homeless individuals cannot be found in this manner. Rather, enumerators canvas geographical areas counting homeless individuals as they find them. Some metropolitan areas are very large, however, which can make some form of spatial sampling a viable alterantive to a full canvasing. One trades a reduction in the cost of data collection for the need to estimate the overall number of homeless and to impute homeless counts for locales not visited by enumerators.

Estimation and imputation raise the issue of how best to represent the costs of underestimation compared to the costs overestimation. Some stakeholders, such as providers of services to the homeless, are more troubled by the prospect of numbers that are too small rather than too large. Other stakeholders, such a elected city officials, have the opposite preference. But more generally, the costs of overestimates relative to the costs of underestimates need to be taken into account as overall counts and counts for particular areas are obtained.

In 2004-2005, the Los Angeles Homeless Services Authority (LAHSA) sought to estimate the homeless population in Los Angeles County. This entailed estimating the number of people living on the streets, in shelters or who were "nearly homeless" (i.e., homeless people living on private property with the consent of its residents). It would have been prohitively costly to canvas the entire county, which covers over 4,000 square miles and is the most populous county in the the United States. Consequently, a form of stratified

sampling was used. An overall figure of the county followed directly, but counts for many areas within the county had to be imputed (Berk et al., 2007). In that analysis, the cost function used was symmetric.

In this paper, we reanalyze the Los Angeles data introducing a variant on stochastic gradient boosting (Friedman, 2002) in which there can be an asymmetric cost function. Specifically, we derive components of stochastic gradient boosting algorithm subject to the quantile distribution, which exhibits a nice translation to cost functions. As a result, we are able to respond to the asymmetric cost functions of various stakeholders. Widely varying estimates can follow, depending on which stakeholder's cost function is used. We also explore how the asymmetic cost function used can affect the ways in which predictors are related to the response. We show that it can be practical and instructive to employ asymmetric costs when using boosting for estimation or imputation.

2 Notation

Let Y be a set of real response values, X be a vector of one or more real predictor variables $(1, \ldots, p)$, and $f(x_i)$ be a fitting function for observation i $(i = 1, \ldots, N)$. We seek to minimize some loss function, Ψ, so that we can estimate the conditional response distribution, $E(Y|X = x)$:

$$E(Y|X = x) = \arg\min_f E\{\Psi(Y, f(x))|X = x\}. \tag{1}$$

We could employ a squared error loss function, in which case the conditional estimate is

$$E(Y|X = x) = \arg\min_f E\{(Y - f(x))^2|X = x\}. \tag{2}$$

Alternatively, we could apply the L_1 loss so that the estimate is

$$E(Y|X = x) = \arg\min_f E\{|Y - f(x)|\Big|X = x\}. \tag{3}$$

Another option is to employ a Huber loss, which is a combination of squared error and absolute loss (Hastie, Tibshirani, and Friedman, 2001). Under all of these loss functions, overestimating and underestimating the response are given equal weight. If underestimating and overestimating are not considered equally costly, this implies a) other regions of the conditional distribution in addition to say, the mean and median, may be of interest, and b) that the loss criteria should be *asymmetric*.

Quantile estimation provides a nice translation from the relative cost of underestimating to overestimating − the "cost ratio" − to descriptions of the response distribution. For example, a 3 to 1 cost ratio implies that underestimating is three times as costly as overestimating, the ratio of the number of underestimates to overestimates will be 3 to 1, and the estimate will be the $3/(3+1) \times 100 = 75^{th}$ percentile. If instead the cost ratio is less than 1 to 1, then we seek to estimate a quantile less than the median.

3 Stochastic Gradient Boosting: A Function Estimation Procedure

Stochastic gradient boosting (Friedman, 2002) is one of the most popular boosting algorithms and exhibits two noteworthy characteristics. First, each "error" is computed as the gradient of each observation $i = 1, \ldots, N$ and defined as the negative partial derivative of the loss function Ψ with respect to the fitting function, f. Hastie et al. (2001) show that stochastic gradient boosting approximates the true gradient using z_{ti}, where z_{ti} depends on the appropriate loss function. Second, shortly after Friedman (2001) introduced gradient boosting, Friedman (2002) took the algorithm one step further by

taking a random sample of observations at each iteration, thereby adding an element of randomness to the algorithm and creating the *stochastic* gradient boosting machine. This additional feature to the algorithm resulted in marked improvement in prediction.[1]

The stochastic gradient boosting algorithm in its most general form is provided below (Friedman, 2002; Ridgeway, 2007; Berk, 2007):

1. Initialize $\hat{f}(x)$ to the same constant value across all observations, $\hat{f}_0(x) = \arg\min_{\rho_0} \sum_{i=1}^{N} \Psi(y_i, \rho_0)$.
2. For t in $1, \ldots, T$, do the following:
   (a) For $i = 1, \ldots, N$, compute the negative gradient as the working response

$$z_{ti} = -\left[\frac{\partial \Psi(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x_i)}\right]_{f_{t-1}(x_i)=\hat{f}_{t-1}(x_i)}$$

   (b) Take a simple random sample without replacement of N' observations from the data set, where N' is less than or equal to the total number of observations, N. We will discuss how large N' should be shortly.
   (c) Fit a regression tree with K terminal nodes, $g_t(x) = E(z_t|x)$, using the randomly selected observations.
   (d) Compute the optimal terminal node predictions, $\rho_{t1}, \ldots, \rho_{tK}$, as

$$\rho_{tk} = \arg\min_{\rho_{tk}} \sum_{x_i \in S_{tk}} \Psi(y_i, \hat{f}_{t-1}(x_i) + \rho_{tk}),$$

   where S_{tk} is the set of x-values that defines the terminal node k for iteration t.

[1] We will focus on the implementation of boosting in the `gbm` library in `R`. In this package, Ridgeway (2007) provides documentation and software based on Friedman's stochastic gradient boosting machine, and one can construct a boosted model subject to, among other distributions, the absolute loss function. From a programmatic standpoint, we sought to augment `gbm` so that it could handle the estimation of *any* quantile in each terminal node within each tree and across all iterations. In `R` alone, we found six boosting libraries in addition to `gbm`. They are `ada`, `adabag`, `boost`, `GAMBoost`, `gbev`, `mboost`. The respective maintainers of these packages are Mark Culp, Esteban Alfaro Cortés, Marcel Dettling, Harold Binder, Joe Sexton, and Torsten Hothorn.

(e) Again using the sampled data, update $\hat{f}_t(x)$ as

$$\hat{f}_t(x_i) \leftarrow \hat{f}_{t-1}(x_i) + \lambda \rho_{tk(x)},$$

where $\{tk(x)\}$ characterizes the index of the terminal node into which an observation characterized by its values of x reside during iteration t. The value of λ determines the "learning rate."

4 Cost-Sensitive Boosting

4.1 Literature

In the context of boosting and machine learning in general, the incorporation of asymmetric costs has applied almost solely to classification problems. Fan, Stolfo, Zhang, and Chan (1999) introduce an algorithm called AdaCost, a more flexible version of AdaBoost.[2] Mease, Wyner, and Buja (2005) propose a boosting algorithm called JOUS-Boost, (**J**ittering and **O**ver/**U**nder-**S**ampling). By adding small amounts of noise to the data and weighting the probability of selection according to each class, one can obtain different misclassification rates than if using no jittering or unweighted sampling according to classes.[3] Berk et al. (2006) incorporate costs into a classification framework using stochastic gradient boosting by specifying a threshold between 0 and 1; observations with predicted probabilities below or above the threshold are assigned values of 0 or 1, respectively. The threshold was established such that the ratio of misclassification errors (false negatives to false positives) approximated the cost ratio.

Outside of boosting but still within the context of machine learning, the notion of unequal error costs in random forests has been researched in both

[2]In a follow-up study of AdaCost and other cost-sensitive variations of AdaBoost, Ting (2000) shows that AdaCost stumbles in certain situations, and that this could be due to the algorithm's weighting structure.

[3]Mease et al. (2005) explain that jittering the data is a needed because, without such component, "as Boosting re-weights the training samples, the tied points all change their weights jointly. Asymptotically this effectively undoes the over/under-sampling."

classification and regression problems. Berk (2007) outlines four ways in which costs are taken into account in random forests when the response is categorical. When the response is the quantitative one can apply "quantile regression forests" (Meinheisen, 2006), an augmentation of random forests (Breiman, 2001), quantile regression (Koenker, 2005) and nonparametric quantile regression (Le, Sears, and Smola, 2005).

4.2 Derivation

To incorporate unequal costs into gradient boosting for quantitative regression problems (and specifically quantile regression boosting), we weight the gradients – the "working responses" – according to their respective signs. For quantitative response distributions in general, consider two types of errors: overestimates and underestimates. A negative gradient corresponds to an overestimated prediction, whereas a positive gradient indicates that the estimated response is underestimated. To weight the gradient and subsequently derived node estimates, we start with a weighted loss function. Consider the loss function in its most general form:

$$\Psi(y, f(x)) = \sum_{i=1}^{N} \psi(y_i, \hat{f}(x_i)), \tag{4}$$

where $\hat{f}(x_i)$ is the function used to obtain a estimated response value. For quantitative outcomes, this loss function can be split into two parts:

$$\Psi(y, f(x)) = \sum_{y_i > \hat{y}_i} \psi(y_i, \hat{f}(x_i)) + \sum_{y_i \leq \hat{y}_i} \psi(y_i, \hat{f}(x_i)) \tag{5}$$

Implicitly, both terms on the right-hand side of equation 5 are multiplied by 1. Suppose we alter this so that each term is instead multiplied by α and $1 - \alpha$, where α is between 0 and 1. Then Ψ becomes:

$$\Psi(y, f(x)) = \alpha \sum_{y_i > \hat{y}_i} \psi(y_i, \hat{f}(x_i)) + (1-\alpha) \sum_{y_i \leq \hat{y}_i} \psi(y_i, \hat{f}(x_i)), \tag{6}$$

where $\alpha/(1-\alpha)$ is the cost ratio of underestimating to overestimating. If α is equal to 0.5, then the cost ratio is 1 to 1, and we get the usual symmetric loss function back. If α is between 0.5 and 1, then underestimates are penalized more heavily than overestimates in Ψ, since α is assigned to those observations for which observed outcomes are greater than predicted outcomes. The opposite is true if α is less than 0.5. In the extreme case, if α equals 1, then underestimating is unacceptable, whereas if α equals 0, no overestimates are allowed.

4.3 Boosting the Quantile Distribution

Recall the boosted L_1 loss function:

$$\Psi(f_t(x_i) : x_i \in S_{tk}) = \left\{ \sum_{x_i \in S_{tk}} \Big| w_i(y_i - f_t(x_i)) \Big| \right\} \Big/ \sum_{x_i \in S_{tk}} w_i, \tag{7}$$

where w_i is a pre-determined population weight for observation i that remains constant across all iterations. After altering equation 7 to allow for a weighted cost ratio, the loss function becomes:

$$\begin{aligned} \Psi(f_t(x_i) : x_i \in S_{tk}) \;=\; & \left\{ \alpha \sum_{\substack{x_i \in S_{tk}, \\ y_i > \hat{f}_t(x_i)}} \Big| w_i(y_i - \hat{f}_t(x_i)) \Big| + \right. \\ & \left. (1-\alpha) \sum_{\substack{x_i \in S_{tk}, \\ y_i \leq \hat{f}_t(x_i)}} \Big| w_i(y_i - \hat{f}_t(x_i)) \Big| \right\} \Big/ \sum_{x_i \in S_{tk}} w_i, \end{aligned} \tag{8}$$

which is essentially a weighted absolute loss function. Then, the gradient becomes[4]

$$z_{ti} = -\frac{\partial \Psi}{\partial f_t(x_i)} = \begin{cases} w_i\alpha & : \quad y_i > \hat{f}_{t-1}(x_i) \\ -w_i(1-\alpha) & : \quad y_i \leq \hat{f}_{t-1}(x_i), \end{cases} \tag{9}$$

where the derivative is evaluated at $\hat{f}_{t-1}(x_i)$. Using equation 8, we can find the value of ρ_{tk} that minimizes Ψ subject to an asymmetric absolute loss function.

$$\begin{aligned} \rho_{tk} &= \arg\min_{\rho_{kt}} \left\{ \alpha \sum_{\substack{x_i \in S_{tk}, \\ y_i > \hat{f}_{t-1}(x_i)+\rho_{tk}}} \left| w_i(y_i - (\hat{f}_{t-1}(x_i) + \rho_{tk})) \right| \right. \\ &+ \left. (1-\alpha) \sum_{\substack{x_i \in S_{tk}, \\ y_i \leq \hat{f}_{t-1}(x_i)+\rho_{tk}}} \left| w_i(y_i - (\hat{f}_{t-1}(x_i) + \rho_{tk})) \right| \right\}, \end{aligned} \tag{10}$$

where $f_t(x_i)$ is set to the prediction from the previous iteration plus the terminal node estimate from the current iteration, $\hat{f}_{t-1}(x_i) + \rho_{tk}$. Next, we differentiate and find the value of ρ_{tk} that minimizes Ψ:

$$\frac{\partial \Psi}{\partial \rho_{tk}} = \left\{ -\alpha \sum_{\substack{x_i \in S_{tk}, \\ y_i > \hat{f}_{t-1}(x_i)+\rho_{tk}}} w_i + (1-\alpha) \sum_{\substack{x_i \in S_{tk}, \\ y_i \leq \hat{f}_{t-1}(x_i)+\rho_{tk}}} w_i \right\} \Big/ \sum_{x_i \in S_{tk}} w_i \tag{11}$$

$$0 = -\alpha \sum_{\substack{x_i \in S_{tk}, \\ y_i > \hat{f}_{t-1}(x_i)+\rho_{tk}}} w_i + (1-\alpha) \sum_{\substack{x_i \in S_{tk}, \\ y_i \leq \hat{f}_{t-1}(x_i)+\rho_{tk}}} w_i. \tag{12}$$

Each summation in equation 12 reduces to the number of observations that are underestimated and overestimated, respectively. Let N_{tk} denote the

[4]Incidentally, under the usual L_1 loss function, the gradient is the sign of the difference between the observed response, y_i, and the predicted value, $\hat{f}_t(x_i)$, multiplied by the population weight, w_i.

number of observations in terminal node k at iteration t, and let n_{tk} and $N_{tk} - n_{tk}$ be the number of underestimates and overestimates in the terminal node, respectively. For simplicity, assume that $w_i = 1$ for all i. Equation 12 can then be expressed as follows:

$$0 = -\alpha(N_{tk} - n_{tk}) + (1 - \alpha)(n_{tk}). \tag{13}$$

Solving for n_{kt}, the location parameter is

$$n_{tk} = \alpha N_{tk}. \tag{14}$$

By weighting the loss function according to overestimates and underestimates, the terminal node prediction of node k at iteration t is the α quantile of the N_{tk} gradients. In each terminal node, there are approximately αN_{tk} and $(1 - \alpha)N_{tk}$ gradients above and below the terminal node estimate ρ_{tk}, respectively. Incidentally, $f_0(x_i)$ equals 0 for all i, and ρ_0 equals the α quantile of all N observations in the data set. Therefore, the predicted value for observation i after t iterations, $\hat{f}_t(x_i)$, is equal to the α quantile of N observations (ρ_0) plus the sum of quantile estimates of the terminal nodes in which observation i resides (ρ_{tk}), scaled by a learning rate λ.

5 Case Study: Estimating the Los Angeles County Homeless Population

5.1 Background on the Los Angeles County Homeless Study

Los Angeles County consists of 2,054 census tracts, and the County's homeless services are divided into eight geographical Service Provision Areas (SPAs). For the 2004-2005 study, the sampling design called for two steps: first, to

visit those tracts that were likely to have large numbers of homeless people with probability 1. There were 244 tracts of this nature, known as "hot tracts." The second step was to visit a stratified random sample of tracts from the population of non-hot tracts. The strata were the eight SPAs, and the number of tracts drawn from each stratum was proportional to the number of tracts assigned to each SPA. In all, there were 265 tracts in the stratified random sample, leaving 1,545 tracts' counts to be imputed. Here we will focus on the street counts in the 1,810 tracts that were not classified as hot tracts.[5]

5.2 Data Characteristics

Berk et al. (2007) explain that dozens of predictors were considered in the estimation process.[6] Ultimately, the 10 predictors named and described in Table 1 were found to contribute the most and make sense to employ. Collectively, these covariates provide information about each census tract's geographical location, land usage, socioeconomic information, and ethnic demographic data. With the exception of median household income and planar coordinates, all other covariates are in terms of percentages. While street counts were obtained in sampled tracts only, predictor values were obtained in all tracts.

Here, we mention that all of the boosted models constructed for the purposes of imputation are strictly correlational. The purpose of including the predictors identified in Table 1 is to ensure that such information (i.e., correlates) in the census tracts is utilized to the fullest extent possible. In Figure 1, the choice to place the predicted and observed counts on the x and y axes, respectively, was arbitrary. The same is true for the partial plots in

[5]Homeless people were paid $10 per hour to help the field researchers identify locations in which the homeless could be found. Presumably, this helped address the problem of finding "hidden homeless" as defined by Rossi (1989).

[6]Incidentally, Berk et al. (2007) employed random forests (Breiman, 2001) to impute street counts for unvisited tracts

Response Name	Description
$STTotal$	Street count of the number of homeless

Predictor Name	Description
$Commercial$	% of land used for commercial purposes
$Industrial$	% of land used for industrial purposes
$MedianHouseholdIncome$	Median household income
$PctMinority$	% of population that is non-white
$PctOwnerOcc$	% of housing units occupied by property owner
$PctVacant$	% of household units that are unoccupied
$Residential$	% of land used for residential purposes
$VacantLand$	% of land that is vacant
$XCOORD$	Planar longitude
$YCOORD$	Planar latitude

Table 1: Names and Descriptions of Variables in Homeless Data

section 5.4.

We found the street count variable, $STTOTAL$ to be highly unbalanced; 75 percent of the observed counts are less than or equal to 27 people, but 22 of the 265 tracts have over 50 homeless, of which 11 have more than 100 homeless.[7] Given the unbalanced distribution of street counts, it is highly likely for the response distribution within each terminal node to be skewed as well. If so, then the typical underestimate will be larger than the typical overestimate. Depending on one's perspective and level of interest on the homeless situation, this feature in the estimates may or may not be tolerable.

5.3 Observation-level and Aggregated Predictions

Figure 1 shows a scatter plot of predicted versus observed street counts for the 265 visited census tracts using various quantiles (cost ratios), and table 5.3

[7]The summary statistics of the $STTOTAL$ are as follows: Min = 0, Q1 = 4, Median = 12, Mean = 21.6, Q3 = 27, Max = 282.

reports aggregated predicted street counts in both the training data and population of data, as well as the recommended number of iterations based on 10-fold cross-validation. With the exception of α, all other tuning parameters are held constant across the boosted models.[8]

When assuming a 1 to 1 cost ratio, the magnitude of the prediction error is highly correlated with the sign of the difference between observed and predicted street counts. Among the 22 tracts with observed responses of at least 50, the median prediction is roughly 71 people below the observed count. Even though just 11 tracts exhibit observed counts of more than 100 (and as many as 282) people, the maximum prediction is 41. Conversely, among the 243 tracts with observed counts less than 50 people, the median prediction is approximately 1 person more than observed, and 225 deviations are less than 20 people. From a policy and economic standpoint, such differences are arguably negligible; stakeholders may very well treat the homeless issue identically if, for example, there are 20 or 40 homeless in a certain census tract. But the homeless situation may very well not be treated the same across tracts with predicted counts of say, 20 and 120. Such counts suggest very different "take-home" messages.

When we allow the cost ratio to vary, we see tract and aggregate-level predictions increase monotonically with respect to α, as evidenced by Figure 1 and Table 5.3. For clarity purposes in Figure 1, we show results based on three cost ratios of underestimating to overestimating: 1 to 10, 1 to 1, and 10 to 1. Estimates using other cost ratios behave as expected; as the cost ratio of underestimating to overestimating increases, observation-level estimates increase monotonically. As α approaches extreme values − either 0 or 1 − the tract-level predictions approach the minimum or maximum observed

[8] All results reported are based on quantile boosted models with a maximum of 4 splits per tree subject to at least 10 observations per terminal node, a learning rate of 0.001, a random sample rate of 50%. It is not entirely clear as to what tuning parameter settings are most appropriate in this situation or in general (Mease and Wyner, 2005). But by holding these arguments constant and only adjusting the cost ratio, it is possible to compare results across various values of α.

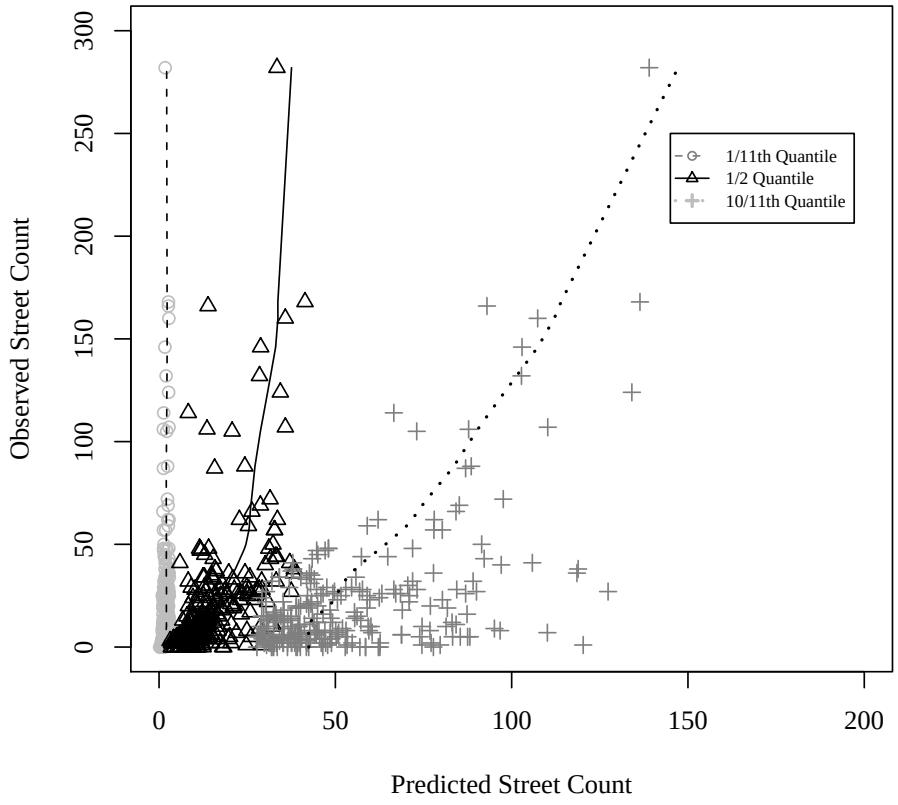

Figure 1: Scatter Plot of Observation-level Observed Versus Predicted Street Counts Using Quantile Boosting, Los Angeles County Homeless Data

response value, respectively.

In a way, the relatively large deviations between predicted and observed outcomes are non-issues. Among the "hot" and sampled tracts, one can make the argument that there is no reason to utilize predicted counts from a model since the response is observable in these areas. Berk et al. (2007) employed this practice when they provided estimates to LAHSA at both the tract and aggregate levels. However, assuming the training data observations are representative of the population of all non-hot tracts, estimates in Figure 1 reveal how close (or far) our predictions will be to the observed values, if observed counts were available.

α	Suggested Iterations Based on 10-Fold CV	Sum, Training Data	Sum, Population
1/11	1,312	509	3,395
1/9	66	544	3,713
1/6	2,343	1,022	6,833
1/4	1,084	1,395	9,581
1/3	4,592	2,517	17,778
1/2	3,790	3,996	28,535
2/3	1,686	5,746	40,731
3/4	1,835	7,470	52,888
5/6	1,790	10,083	71,209
8/9	1,064	12,128	84,750
10/11	1,509	14,117	98,726

Table 2: Summary of Quantile Regression Boosting Models, Los Angeles County Homeless Data

Table 5.3 demonstrates that aggregated predicted counts increase monotonically with respect to α. Additionally, we report the number of trees grown based on 10-fold cross-validation to demonstrate that we can use this data-derived procedure to determine a sensible number of iterations, even though the loss function is asymmetric (Zhang and Yu, 2005).[9] Using the

[9]In `gbm`, it is also possible to determine a sensible number of iterations by using out-

boosting arguments specified above, the most iterations needed to reach a minimum cross-validation error is 3,790 (33^{rd} quantile). The fact that just 66 iterations are needed to perform quantile boosting with α equal to 0.111 is likely attributable to the unbalancedness of the response.[10]

5.4 Partial Dependence Plots

For demonstrative purposes, we show that partial dependence plots can be constructed in the usual manner. Figure 2 shows partial plots for three randomly selected predictors – *PctVacant*, *Commercial*, and *PctOwnerOcc* – for quantiles of 1/11, 1/6, 1/2, 5/6, and 10/11. Intuitively, the more desirable an area is to live, the lower the vacancy rate and number of homeless; thus, it is reasonable to assume that the number of homeless will increase as *PctVacant* increases. Similarly, areas that are mostly commercial such as downtown Los Angeles may have higher homeless counts than areas in which there is little commercial land (e.g., residential Beverly Hills). Conversely, a higher percent of owner-occupied housing units suggests a higher level of prosperity, so the number of homeless should decrease as *PctOwnerOcc* increases.

From Figure 2, we observe that the partial response values increase as α increases for each of the three predictors. Given the upwards increases in predicted values with respect to α shown in figure 1 and table 5.3, these

of-bag data or a test data set. We chose cross-validation because our experience has been that cross-validation yields better results in terms of prediction error compared to the out-of-bag and test data methods.

[10]Though not shown, each of these models exhibited a roughly parabolic behavior in the 10-fold cross-validation deviance as a function of the number of iterations. At $t = 0$, a boosted model is likely to exhibit a relatively high error rate since the information in the covariates has not yet been utilized and all predictions are initialized to a constant. The error statistic then decreases with respect to t until boosting results in overfitting the data, at which point the error tends to increase (possibly to $+\infty$). If the error decreases monotonically, then more iterations are required before observing a parabolic behavior in the error measurement. Conversely, if the error is monotonically increasing, this means that just a few (if any) iterations are needed to obtain a satisfactory solution.

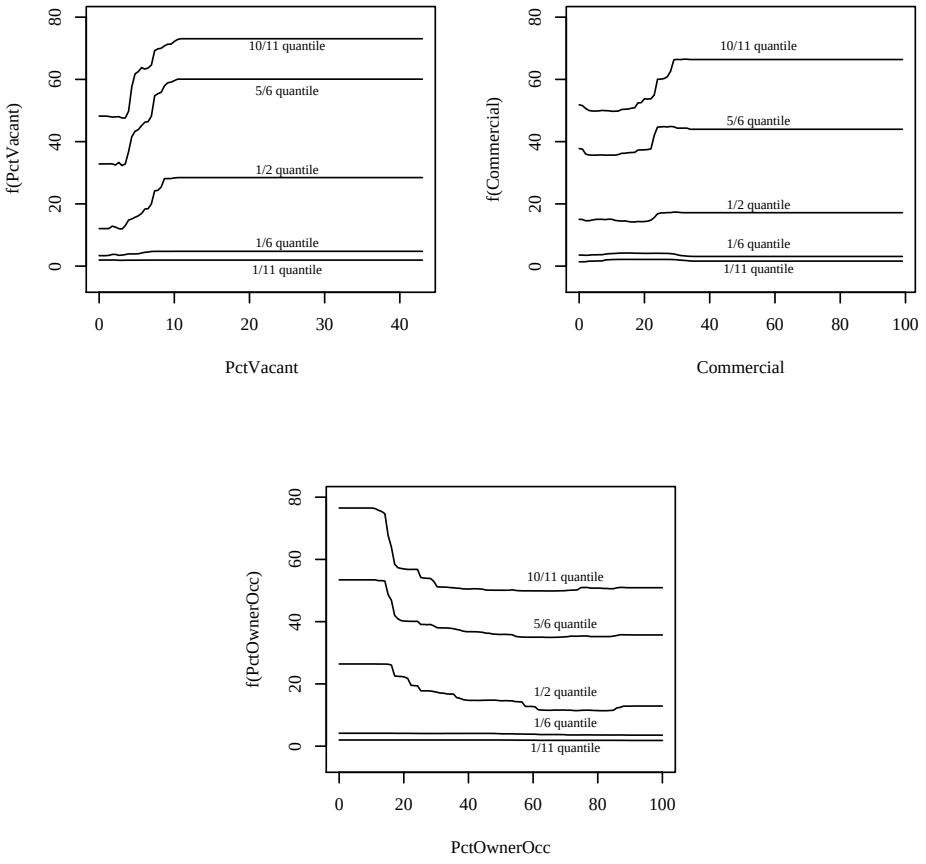

Figure 2: Partial Dependence Plots using Quantile Regression Boosted Models

results are logical. The partial smoothers are fairly flat when overestimating is considerably more costly than underestimating (i.e., cost ratios of 1 to 5 and 1 to 10); this makes sense because the unit-level predictions exhibit little variation for small values of α. Within each of the partial plots, many smoothers are roughly parallel to one another; however, such findings should not be generalized to all data sets.

5.5 Variable Importance

Using `gbm`, predictor importance is defined as the reduction in mean squared error from the splits on each predictor, and relative importance is the percent reduction in mean squared error attributable to this particular variable out of the total reduction in mean squared error. Like the partial plots, our objective is to show that such computations are still calculable in the same manner as when employing the usual L_1 boosting. More will be said about variable importance in the concluding section.

VacantLand and *Industrial* reduce the mean squared error little in all quantile boosting models, while *PctVacant* ranks among one of the most important predictors. Excluding the 1 to 10 and 1 to 8 models, *PctOwnerOcc*, *XCOORD*, and *YCOORD* have variable importance in excess of 10 percent. At very small values of α, relative influence percentages can be misleading because the predictions do not vary much, so none of the predictors are contributing much to the boosted models. In general, these results suggest that predictor importance does not necessarily have to be the same across cost ratios; certain predictors are more important in estimating certain regions of the response distribution.

5.6 Discussion

Within the context of counting the homeless, quantile boosting is a potentially useful statistical tool to the wide range of stakeholders described in

	Quantile					
Predictor	1/11	1/9	1/6	1/4	1/3	1/2
Commercial	19.99	15.76	11.67	5.29	7.90	8.65
Industrial	2.27	1.94	3.87	4.36	3.10	2.84
MedianHouseholdIncome	8.34	10.85	10.22	14.44	10.04	10.21
PctMinority	13.28	12.58	10.84	11.82	13.69	7.64
PctOwnerOcc	5.36	7.18	6.93	8.00	11.14	20.42
PctVacant	9.83	7.18	13.82	16.06	14.27	15.28
Residential	17.77	17.67	12.23	9.05	9.23	8.21
VacantLand	2.59	1.80	2.32	2.76	3.19	3.23
XCOORD	6.33	8.64	11.70	11.22	12.45	13.66
YCOORD	14.24	16.41	16.38	16.99	14.99	9.87

Predictor	2/3	3/4	5/6	8/9	10/11
Commercial	9.48	9.11	7.89	7.30	8.51
Industrial	2.22	2.73	3.74	3.51	3.91
MedianHouseholdIncome	5.62	6.23	8.77	10.11	10.28
PctMinority	7.47	9.43	12.33	10.05	13.15
PctOwnerOcc	20.05	12.02	10.85	13.39	10.53
PctVacant	21.39	20.13	19.75	20.18	17.19
Residential	5.49	7.00	8.09	10.58	12.27
VacantLand	2.35	3.92	2.76	1.99	2.07
XCOORD	19.08	20.95	16.75	14.01	13.40
YCOORD	6.86	8.50	9.07	8.89	8.69

Table 3: Relative Influence of Predictors in Quantile Boosted Models

section 1. Consider the following scenarios. Suppose a homeless shelter advocate views undercounting to be twice as costly as overcounting, and a tract's estimate is 100 using 2 to 1 costs and 25 using 1 to 1 costs. Given the difference in magnitude, this stakeholder may insist on performing a manual street count of this census tract because the socioeconomic implications of having 25 versus 100 homeless in this area may be substantial. Alternatively, if two local estimates based on these cost ratios are 10 and 20, from a practical standpoint this difference may be considered inconsequential, and a manual count may not be required. Now consider another perspective, in which city council is being pressured to make budget cuts to government spending so that private commercial building developers can expand on areas occupied by homeless. From a councilperson's perspective, each overestimate may be viewed as three times more costly than an underestimate. If one is willing to believe that an area's homeless estimate is sufficiently small (say, less than 20 people), then it may be possible to move forward with the development of commercial expansion and be less concerned with complications associated with displacing homeless people.

6 Conclusions

Deriving the boosting components of the quantile distribution is in some sense adding another algorithm to those provided in Friedman (2001) and the `gbm` library (Ridgeway, 2007). The quantile loss function is differentiable, and there exists a solution that minimizes this loss (Hastie et al., 2001). At the same time, incorporate unequal underestimation and overestimation error costs − as in section 4.2 − is likely applicable to other quantitative loss functions in addition to the quantile distribution. Hence, the inclusion of the cost ratio enhances boosting's versatility. Incidentally, one may argue that a "weighted poisson" loss function is appropriate to apply to the homeless data, since the response is a count variable. Nevertheless, the interpretation is

perhaps cleanest when applying the quantile distribution, and this is certainly a characteristic worth considering when selecting a distribution.

In evaluating the results from the case study, some results seem generalizable, some are likely data-specific, and some absolutely require future research. The predictions at both the unit and aggregated levels increase monotonically as α increases. The partial plots seem to be fully operational, allowing us to examine relationships between the response and covariates for various regions of the response distribution. Finally, although relative influence is straightforward to compute, it is important to consider the utility of these calculations. Given that the loss function need not be symmetric, perhaps the mean reduction in squared error is no longer "gold standard," and different calculations are needed to assess the criticality of each predictor. In order for cost-sensitive boosting to become a fully useful tool, it will be necessary for the usual set of model diagnostics to be operational and easily interpretable.

References

[1] Berk, R.A. (2007). *Data Mining from a Regression Perspective.* (Forthcoming).

[2] Berk, R.A., Kriegler, B. and Ylvisaker, D. (2007). Counting the Homeless in Los Angeles County. (Forthcoming).

[3] Binder, H. (2006). `GAMBoost`: Generalized additive models by likelihood based boosting. R version 0.9-3.
URL http://www.r-project.org

[4] Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5-32.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Wadsworth Press, Monterey, California.

[6] Cortés, E.A. (2006). `adabag`: Applies Adaboost.M1 and Bagging. R version 1.0.
URL http://www.r-project.org

[7] Culp, M. (2006). `ada`: Performs boosting algorithms for a binary response. R version 2.0-1.
URL http://www.r-project.org

[8] Dettling, M. (2005). `boost`: Boosting Methods for Real and Simulated Data. R version 1.0-0.
URL http://stat.ethz.ch/∼dettling

[9] Fan, W., Stolfo, S. J., Zhang, J., and Chan, P. K. (1999). Adacost: Misclassification cost-sensitive boosting. *Machine Learning: Proceedings of the Sixteenth International Conference.*

[10] Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Computer and System Sciences, 55*, 119139.

[11] Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Ann. of Stat.* **29**, 1189-1232.

[12] Friedman, J.H. (2002). Stochastic Gradient Boosting. *Comp. Stat. & Data Analysis* **38**, 367-378.

[13] Friedman, J.H., Hastie, T., and Tibshirani, R. (2000). Additive Logistic Regression: A Statistical View of Boosting. **28**(2), 337-374.

[14] Hastie, T., Tibshirani, R. and Friedman J. (2001). *The Elements of Statistical Learning.* New York: Springer-Verlag.

[15] Hothorn, T. (2006). `mboost`: Model-Based Boosting. R version 0.4-14. URL http://www.r-project.org

[16] Koenker, R. (2005). *Quantile Regression.* Cambridge University Press: New York.

[17] Le, Q.V., Sears, T. and Smola, A. (2005). Nonparametric Quantile Regression. *Technical Report, NICTA.*

[18] Mease, D., Wyner, A.J., and Buja, A., (2005). Cost-Weighted Boosting with Jittering and Over/Under-Sampling: JOUS-Boost. Preprint available at http://www-stat.wharton.upenn.edu/∼buja/.

[19] Mease, D. and Wyner, A. (2006), Evidence Contrary to the Statistical View of Boosting. *J. of Machine Learning Research* (Forthcoming).

[20] Meinheisen, N. (2006). Quantile Regression Forests. *J. of Machine Learning Research* **7**, 983-999.

[21] R CORE DEVELOPMENT TEAM (2006). R: A Language and Environment for Statistical Computing. R Foundation for Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL http://www.r-project.org

[22] Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons.

[23] Ridgeway, G. (1999) The State of Boosting. *Computing Science and Statistics, 31*, 172-181.

[24] Ridgeway, G. (2007). `gbm`: Generalized Boosted Regression Models (2006). R package version 1.6-3.
URL http://www.i-pensieri.com/gregr/gbm.shtml

[25] Sexton, J. `gbev`: Gradient Boosted Regression Trees with Errors-in-Variables. R version 0.1.
URL http://www.r-project.org

[26] Ting, K.M. (2000). A comparative study of cost-sensitive boosting algorithms. *Proceedings of The Seventeenth International Conference on Machine Learning*. Morgan Kaufmann: San Francisco, CA, 983-990.

[27] Zhang, T. and Yu, B. (2005). Boosting with Early Stopping: Convergence and Consistency. *Annals of Statistics* **33**(4), 1538-1579.