Small Area Estimation of the Homeless in Los Angeles: An Application of Cost-Sensitive Stochastic Gradient Boosting*

Brian Kriegler and Richard Berk†
Department of Statistics, UCLA

January 4, 2010

Abstract

In many metropolitan areas, efforts are made to count the homeless to ensure proper provision of social services. Some areas are very large, which makes spatial sampling a viable alternative to an enumeration of the entire terrain. Counts are observed in sampled regions but must be imputed in unvisited areas. Along with the imputation process, the costs of underestimating and overestimating may be different. For example, if precise estimation in areas with large homeless counts is critical, then underestimation should be penalized more than overestimation in the loss function. We analyze data from the 2004-2005 Los Angeles County homeless study using an augmentation of L_1 stochastic gradient boosting that can weight overestimates and underestimates asymmetrically. We discuss our choice to utilize stochastic gradient boosting over other function estimation procedures. In-sample fitted and out-of-sample imputed values, as well as relationships between the response and predictors, are analyzed for various cost functions.

*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."

†As of August 2006, Richard Berk is in the Department of Criminology and the Department of Statistics at the University of Pennsylvania. As of August 2008, Brian Kriegler is an Economist at Econ One Research.

Practical usage and policy implications of these results are discussed briefly.

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

1 Introduction

Dating as far back as the 1930s, homelessness has been a visible, public issue in the United States (Rossi, 1989). At least over the past decade, the homeless problem has been underscored due to the rise in unemployment and foreclosures. In the 2010 census, there are no plans to perform street counts, thereby making it challenging for stakeholders (e.g., homeless service advocates and selected government agencies) to estimate the magnitude of the necessary social resources. This is especially difficult in large metropolitan areas because the homeless are often dispersed due to the changing availability of homeless services, commercial development and the government's homeless criminalization practices (Berk et al, 2010). Areas needing these services are literally "moving targets." Adequate spatial apportionment of homeless-related resources requires a great deal of local information that is oftentimes prohibitively expensive to obtain.

In a typical census design, people are contacted through their place of residence. With the possible exception of individuals living on private property, the homeless will not be found using this design (Rossi, 1989). An alternative approach is to locate homeless individuals in temporary shelters or while they are receiving services (e.g., meals) from public and private agencies. It is widely known, however, that a large number of the homeless still will not be found this way because many do not use these services. Therefore, it is common for enumerators to canvas geographical areas and to count the homeless as they find them. Some metropolitan areas are very large, making spatial sampling a viable substitute to a full canvasing. One trades a

reduction in the burden of data collection in exchange for the need to impute homeless counts for locales not visited by enumerators.

Estimation and imputation raise the issue of how best to represent the cost of underestimation relative to overestimation ("cost function"). The apportionment of homeless-related resources depends, at least in part, on the estimated size of local homeless population. Some stakeholders, such as homeless service providers, are more troubled by the prospect of numbers that are too small rather than too large. This is especially true in areas where homeless counts are high, in which undercounting may carry serious consequences. Other stakeholders, such as elected city officials faced with budget constraints, may have the opposite preference. In general, one needs the flexibility to penalize overestimation and underestimation distinctly.

The homeless problem is especially serious in Los Angeles, which has a large homeless population and consists of specific areas with very densely populated homeless encampments (Berk, Kriegler and Ylvisaker, 2008). These encampments can be a nuisance to local commerce and can compound the demand, for example, for police and hospital services (Harcourt, 2005). One such area is "Skid Row" (Mangano and Blasi, 2007), located just outside downtown Los Angeles. This area has been historically marked by high crime rates in terms of drug markets, robberies, vandalism, and prostitution, as well as drug and alcohol abuse (Lopez, 2005).[1] Individuals (especially the homeless) who spend significant amounts of their time in public areas of such locales have higher victimization rates than those who reside outside these areas (Koegel, Burnam, and Farr, 1988; Kushel, Evans, Perry, Robertson, and Moss, 2003). In short, the set of public and private resources dependent on the homeless population extends beyond the services dedicated to the

[1]In 2005, the Los Angeles Police Department tested a pilot program, called "Safer Cities Initiative" (SCI), which was designed to target specific geographical crime "hot spots" (Bratton and Knobler, 1998; Wilson and Kelling, 1982). Part of this program entailed reducing the density of homeless encampments. A full-scale version of SCI began in September 2006 (Berk and MacDonald, 2009).

homeless' physical and mental health (e.g., soup kitchens, shelters, affordable housing, etc.).

In 2004-2005, the Los Angeles Homeless Services Authority (LAHSA) estimated the homeless population in Los Angeles County as the aggregate of people who were 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). At any given time, shelters cater to just a fraction of the local homeless population; consequently, locating and estimating the street count was a daunting task.[2] It would have been prohibitively costly to canvas the entire county, which covers over 4,000 square miles, includes 2,054 census tracts, and is the most populous county in the United States.

A stratified spatial sampling of census tracts called for two steps. First, tracts believed to have large numbers of homeless people were visited 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 County's eight Service Provision Areas (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.[3] In that analysis, the cost function was symmetric, and emphasis was placed on estimating the homeless population within each SPA, for various aggregations (e.g., cities), and for the entire County (Berk et al, 2008). Almost certainly, symmetric costs are insufficiently responsive to the policy needs of local stakeholders because both actual and imputed counts

[2]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" (Rossi, 1989).

[3]This is a "small area estimation" analysis. Rao (2003) defines a domain, or area, as "small" if "the domain-specific sample is not large enough to support direct estimates of adequate precision." In the context, homeless counts in the 265 randomly sampled tracts were used to impute the numbers of homeless people in unvisited tracts and ultimately the entire County.

can vary dramatically.

In this paper, we re-analyze the Los Angeles data of 1,810 non-hot tracts using stochastic gradient boosting (Friedman, 2002) subject to an asymmetrically weighted absolute loss function. We focus on evaluating the relationship between homeless counts and covariates in visited tracts and imputing the counts in unvisited tracts. By boosting a cost-sensitive loss function, we are able to respond to the cost functions of various stakeholders and focus on a particular region of the conditional response. Depending on which cost function is applied, widely varying fitted and imputed values can follow. We also explore how different regions of the conditional response are related to the predictors. We show that it can be practical and instructive to employ asymmetric costs when using boosting for function estimation and imputation.

The remainder of this paper consists of five sections plus an appendix. Section 2 includes a description of the Los Angeles County homeless and census data. In Section 3, we provide an overview of stochastic gradient boosting and a literature review on cost-sensitive estimation procedures. Our analysis of the homeless data, which includes comparisons between fitted and observed counts, imputed counts, and model diagnostics, is in Section 4. Section 5 includes a discussion on how our proposed methodology and analysis can have a profound effect on policy-making decisions. In Sections 4 and 5, we stress the results based on models that place heavier penalties on underestimating, as this represents what stakeholders would likely employ to ensure proper allocation of homeless-related services. We conclude the paper in Section 6, in which we mention some aspects of cost-sensitive statistical learning to be explored. In Appendix A, we derive the functional forms for the deviance, initial value, gradient, and terminal node estimates when employing boosting subject to asymmetrically weighted absolute loss.

2 Data Description

In the 2004-2005 Los Angeles homeless study, Berk and his colleagues (2008) considered the use of dozens of predictors in the estimation process.[4] The 10 predictors in Table 1 were relatively important to fitting the conditional distribution of street counts, capturing information about each 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 presented in terms of percentages. While street counts were obtained only in sampled tracts, predictor values were available for all of the County's tracts.

Response Name	Description
StTotal	Homeless street count

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-Caucasian
PctOwnerOcc	% of owner-occupied housing units
PctVacant	% of unoccupied housing units
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 Los Angeles County homeless dataset.

Looking ahead to Section 4, none of our models are intended to necessarily suggest causal relationships. We utilized predictor information described in Table 1 primarily to estimate the conditional distribution between *StTotal*

[4]In that study, fitted and imputed counts were obtained using random forests (Breiman, 2001).

and each covariate and to construct sensible fitted and imputed street counts. Whether the predictors are causally related to homeless counts is at best a secondary concern.

The distribution of *StTotal* is highly unbalanced. 75 percent of the observed counts are less than 28 people, and 22 of the 265 tracts have over 50 homeless, of which 11 have over 100 homeless (Min = 0, Q1 = 4, Median = 12, Mean = 21.6, Q3 = 27, Max = 282). To ensure adequate local resources, stakeholders such as police departments and homeless shelter advocates may place heavy emphasis on accurately estimating the counts in areas that have large homeless populations (e.g., over 100 people). If so, one is willing to trade overall accuracy for a better fit in the right tail of the street count distribution, and underestimates are more costly than overestimates. For policy purposes, resources may still be adequate in an area with a predicted count of 30 people when in fact the count is 50. However, if the prediction is 30 and the actual count is 150, there may well be a severe shortage of local resources.

3 Estimating the Conditional Distribution

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, to fit the conditional response distribution, $G(Y|X = x)$:

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

We could minimize the L_1 loss so that the estimate is

$$G_{L_1}(Y|X = x) = \arg\min_f G\{|Y - f(x)|\}, \tag{2}$$

in which overestimating and underestimating the response are weighted symmetrically, and $\hat{f}$ is the median of Y. But if underestimating and overestimating are not equally costly, then the loss criteria needs to be *asymmetric*. Let $L_1(\alpha)$ be the absolute loss function that weights underestimates by α and overestimates by $1-\alpha$, where $0 \leq \alpha \leq 1$. Then $G(Y|X=x)$ is defined as follows:

$$G_{L_1(\alpha)}(Y|X=x) = \begin{cases} \arg\min_f G\{\alpha|Y-f(x)| \;\Big|\; Y > f(x)\} \\ \arg\min_f G\{(1-\alpha)|Y-f(x)| \;\Big|\; Y \leq f(x)\}. \end{cases} \tag{3}$$

If $\alpha = 0.5$, then $G_{L_1(\alpha)} = G_{L_1}$. In general, $\hat{f}(x)$ is the quantile of Y, which exhibits a straightforward translation between the cost of underestimating relative to overestimating – the "cost ratio" (or "cost function") – and 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 underestimates to overestimates will be 3 to 1, and $\hat{f}$ is the $3/(3+1) \times 100 = 75^{\text{th}}$ percentile of Y. If instead the cost ratio is less than 1 to 1, then $\hat{f}$ is less than the median of Y. Henceforth, we refer to $\alpha/(1-\alpha)$ as the cost ratio.

3.1 Stochastic Gradient Boosting: An Overview

Stochastic gradient boosting (Friedman, 2002) is a recursive, nonparametric procedure that has become one of the most popular machine learning algorithms among statisticians. It exhibits extraordinary fitting flexibility, as it can handle any differentiable and minimizable loss function. It can handle and produce highly complex functional forms, and there is growing evidence that it outperforms competing procedures (e.g., bagging, splines, CART, and parametric regression) in terms of prediction error (Friedman, 2001; Madigan and Ridgeway, 2004), provided that one utilizes reasonable tuning parame-

ters.[5] Shortly after Friedman (2001) introduced gradient boosting, Friedman (2002) augmented the algorithm by taking a random sample of observations at each iteration, thereby creating the *stochastic* gradient boosting machine. This additional feature to the algorithm resulted in marked reduction in bias and variance. Given stochastic gradient boosting's success at estimating the center of $Y|X$, one may deduce that it also performs well at estimating other regions of the conditional response distribution.

The stochastic gradient boosting algorithm in its most general form is provided below (Friedman, 2002; Ridgeway, 2007; Berk, 2008):[6]

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 size N' from the dataset with N observations
   (c) Fit a regression tree with K_t terminal nodes, $g_t(x) = E(z_t|x)$ using the randomly selected observations.
   (d) Compute the optimal terminal node estimates, $\rho_{1t}, \ldots, \rho_{K_t}$, as

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

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

[5] This is especially true when the number of predictors is large (Bühlmann and Yu, 2003).

[6] Our augmentation of stochastic gradient boosting and data analysis were conducted using `gbm` in `R`. We found four libraries in `R` in addition to `gbm`: `ada` (Culp, Michailidis, and Johnson, 2006), `GAMBoost`, `gbev` and `mboost`. The respective maintainers of these packages are Mark Culp, Harald Binder, Joe Sexton, and Torsten Hothorn.

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

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

where λ is the "learning rate."

In Appendix A, we build on equation 3 to derive the deviance subject to $L_1(\alpha)$. Subsequently, we identify the functional form of the initial value, gradient, and terminal node estimates from steps 1, 2a, and 2d of the stochastic gradient boosting algorithm.

3.2 Literature Review

To our knowledge, the inclusion of asymmetric costs to boosting algorithms has applied solely to classification problems. Fan, Stolfo, Zhang, and Chan (1999) introduce an algorithm called AdaCost, a more flexible version of AdaBoost (Freund and Schapire, 1997).[7] Mease, Wyner, and Buja (2007) 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. Berk, Kriegler and Baek (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 so that the ratio of misclassification errors (false negatives to false positives) approximated the cost ratio.

In a regression context, we found three methods capable of handling asymmetric error costs, each building on quantile estimation. If the functional form is specifiable *a priori*, one can employ parametric quantile regression

[7] 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.

(Koenker, 2005). However, if the functional form is not known, it is important and helpful to exploit statistical learning. Then, one could apply nonparametric quantile regression (Takeuchi, Le, Sears, and Smola, 2006). Yet there is evidence that ensemble procedures, such as gradient boosting, typically yield superior bias-variance tradeoffs in comparison (Bühlmann and Hothorn, 2007). Meinshausen (2006) introduced quantile regression forests, an augmentation of random forests (Breiman, 2001). The drawback to this method is that the fitted and imputed values are calculated *after* all of the trees are grown using random forests. Consequently, the conditional response function does not adapt to the cost ratio. It follows that there are no new partial dependence plots and predictor importance measurements (not even when employing L_1, since the usual random forests algorithm estimates the conditional mean).

Just as with parametric quantile regression, estimates based on $L_1(\alpha)$ stochastic gradient boosting do not necessarily increase monotonically with respect to α.[8] Each cost function yields a different model and fitted values that minimize the $L_1(\alpha)$ loss. Therefore, a fitted (or imputed) count may be 30 when the cost ratio is 5 to 1 and 20 when the cost ratio is 10 to 1. With $L_1(\alpha)$ stochastic gradient boosting, our experience − both in this case study and with other datasets − is that i) all (or nearly all) fitted and imputed values tend to increase with respect to α, and ii) when decreases do occur, they tend to be small in magnitude. We found that the use of larger terminal node sizes can reduce this occurrence; however, for reasons we explain in Section 4, we purposely grew trees that potentially had small terminal node sizes. Ultimately, we were not concerned with this "side effect" because its occurrence was rare and inconsequential, and our analysis extended beyond simply calculating fitted and imputed values.

In summary, we employed $L_1(\alpha)$ stochastic gradient boosting for three

[8]Incidentally, quantile regression forests does not share this feature because the quantile estimation is performed on the distribution of each observation's fitted values across regression trees.

main reasons. First, the functional form can be arrived at inductively. Second, we have the prospect of a good bias-variance tradeoff. Third, we can apply unequal error costs at each step of the function estimation process so that *all* of the output is properly cost-sensitive. We found $L_1(\alpha)$ stochastic gradient boosting to provide a formidable set of features for this case study, though it should not be seen as a universal preference for cost-sensitive stochastic gradient boosting in different settings.

4 Analysis

Based on our discussions with key stakeholders including people from LAHSA and government representatives, underestimation is typically seen to be more problematic than overestimation. The prospect of having too few shelter beds, for instance, is more troubling than if a few beds are open. With this in mind, our analysis emphasizes results in which $\alpha > 0.5$. Output based on cost functions that penalize overestimation more heavily are also reported, primarily to demonstrate that they are employable if one desires.

All boosting models were built using the following tuning parameters: 10 splits per tree subject to at least 5 observations per terminal node k_t, a learning rate of $\lambda = 0.001$, and a maximum of $T = 6,000$ trees. For stochastic gradient boosting models, we applied these same tuning parameters along with a random sample of $N' = 133$ observations (i.e., a sampling fraction of 50 percent of $N = 265$, rounded to the nearest whole number). A sensible number of iterations was determined using 10-fold cross-validation, and we found no problems in converging on a reasonable number of trees to grow in

any of our cost-sensitive models.[9]

Using a handful of different learning rates and sampling fractions ranging from 0.001 to 0.01 and 35 to 75 percent, respectively, we saw inconsequential differences in terms of street counts estimates − both fitted and imputed − and conditional distribution diagnostics, for each α. The same held true for models subject to 1 to 10, 1 to 5, and 1 to 1 costs. By contrast, when we employed cost ratios of 5 to 1 and 10 to 1, we learned that the number of splits and the minimum terminal node size can have a substantial impact on point estimates. The `gbm` library uses the inverse of the empirical distribution to estimate quantiles, so each terminal node estimate depends on just one value. Given the unbalanced nature of *StTotal*, differences between consecutive values in the right tail within a terminal node can be very large. If employing a 10 to 1 cost function and a terminal node includes 25 points, then the estimate will be the third highest value. The use of a highly skewed cost function implies a particular interest in estimating the handful of large response values well, yet the top two values in this terminal node of this size will not factor into the estimation process. To ensure that large gradients were given ample opportunities to be terminal node estimates, we permitted large trees and small terminal node sizes. This was facilitated by tuning the number of splits and the minimum number of observations in each terminal node at each iteration.[10]

[9]For example, in the stochastic models when the cost ratio $\alpha/(1-\alpha) \in \{1 \text{ to } 10, 1 \text{ to } 1, 10 \text{ to } 1\}$, the respective "best" numbers of iterations were 588, 2050, and 1415. Small deviations from these numbers of iterations (e.g., 1500 trees subject to 10 to 1 costs) yielded no substantive differences in any results. Just as one would expect when using symmetric costs, the cross-validation error exhibited a concave-up parabolic behavior that tended to decrease with respect to t, until it reached a number of iterations corresponding to the minimum cross-validation error. Beyond the minimum cross-validation error iterations, the models overfit the data. The key here is that these iteration estimates are well short of $T = 6000$, suggesting that we have in fact identified a sensible number of iterations.

[10]By default in `gbm`, each tree at each iteration has one split, subject to at least 10 observations in each terminal node.

4.1 Fitted and Imputed Street Counts

Figure 1 shows fitted versus observed street counts for the 265 visited census tracts using stochastic gradient boosting subject to 1 to 10, 1 to 1, 5 to 1, and 10 to 1 cost ratios ($\alpha \in \{1/11, 1/2, 5/6, 10/11\}$, respectively). Using 1 to 1 costs (L_1 boosting), the magnitude of the error is less than 20 people in 232 of 265 visited census tracts. In terms of resource needs, errors of this magnitude are likely tolerable. Conversely among the 22 tracts with observed counts with at least 50 homeless, all of these tracts' counts are underestimated. The maximum fitted value is approximately 40 people, and the median error is approximately 70 people less than the true count. These large undercounts need to be reduced substantially in order to ensure adequate local resource allocation.

Figure 1 demonstrates that $L_1(\alpha)$ stochastic gradient boosting fitted values tend to increase with respect to α.[11] Although the overall fit worsens when the cost ratio diverges from 1 to 1, we observe smaller errors in specific regions of the response. Using a 10 to 1 cost ratio, just 12 out of 265 tracts are underestimated. Among the 22 tracts with at least 50 people, the median difference between observed and fitted counts is 1 person, and the interquartile range is 37 people. Admittedly, most of the very large counts are still underestimated even when using a 10 to 1 cost ratio, a topic we will pick up again in Section 5.[12]

In a way, training data fitted values are irrelevant because one's estimates of visited tracts might simply be the observed street count. Berk et

[11] Of the 265 visited training data observations, 11 observations' fitted values were lower for $\alpha = 10/11$ than for $\alpha = 5/6$. We did not consider this to be problematic for two reasons. The largest of these differences was 6 people. Also, this was not a problem among tracts with relatively large counts; the largest observed count in any of these tracts was 43 people.

[12] Recognizing that it is in the nature of all regression models to overestimate small values and underestimate large ones, we demonstrate that the use of asymmetric costs can alleviate the problem. As the cost ratio increases, fitted values for tracts with large counts tend to move closer to the 45-degree line.

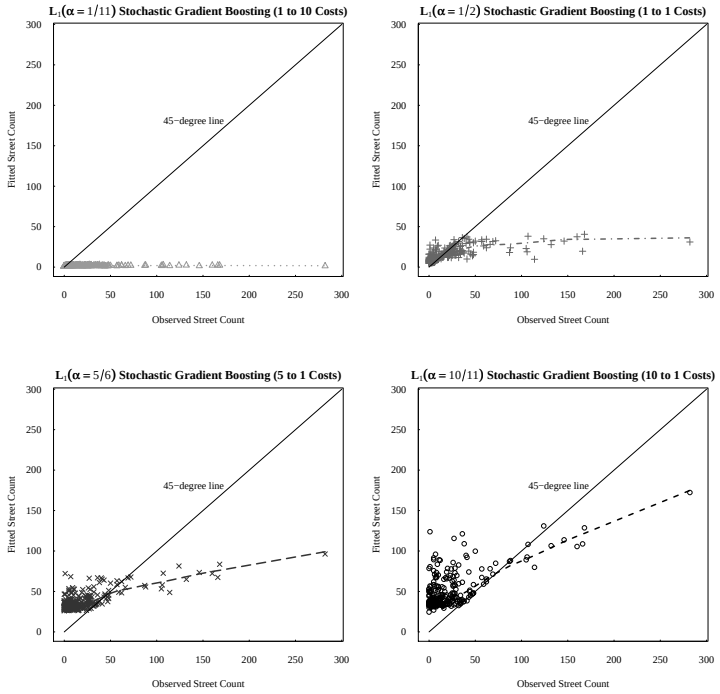

Figure 1: Fitted versus observed census tract street counts using $L_1(\alpha)$ stochastic gradient boosting.

al (2008) employed this practice when they provided estimates to LAHSA at both the tract and aggregate levels. But provided the sampled tracts are representative of the population of all non-hot tracts and the model does not overfit the training data, fitted counts in Figure 1 reveal how close (or far) the unsampled tracts' imputed counts are to the true counts. Figure 2 shows the distribution of imputed counts for various cost ratios. The distributions tend to shift upwards with respect to α.[13] Using 1 to 10 and 1 to 5 costs, all tracts have imputed counts of fewer than 5 people. Conversely using 10 to 1

[13]Of the 1,545 unvisited tracts, we observed decreases in imputed values from $\alpha = 5/6$ to $\alpha = 10/11$ in 42 tracts. Over half of these deviations were less than 2 people, and the largest deviation was 7 people.

costs, we find that 59 of 1,545 tracts have imputed counts over 100 homeless people.

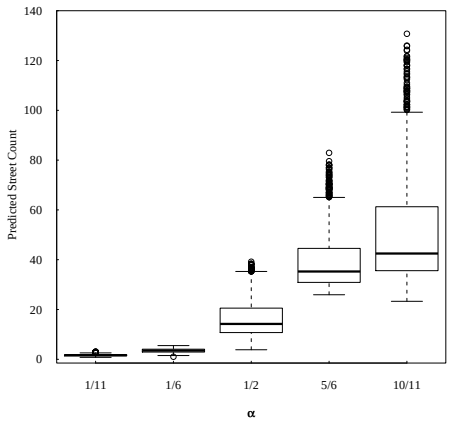

Figure 2: Distribution of predicted street counts in unvisited census tracts using $L_1(\alpha)$ stochastic gradient boosting.

Recognizing that portions of our analysis will be dataset-specific, one may also be interested in the how $L_1(\alpha)$ boosting performs relative to other cost-sensitive methods. Figure 3 shows fitted versus observed street counts using stochastic and non-stochastic gradient boosting, and parametric quantile regression, subject to a 10 to 1 cost function.[14] All three methods have a substantial number of overestimates, which is to be expected given the cost ratio of choice. Among tracts with at least 50 homeless people observed, $L_1(\alpha)$ stochastic gradient boosting performs noticeably better than the other two methods in terms of bias and variance. Non-stochastic gradient boosting exhibits a median deviation of 35 people underestimated and an IQR of 77 people. Quantile regression's median deviation and IQR are 7 and 63 people, respectively.

[14]Parametric quantile regression was performed using the `quantreg` library in `R`, maintained by Roger Koenker.

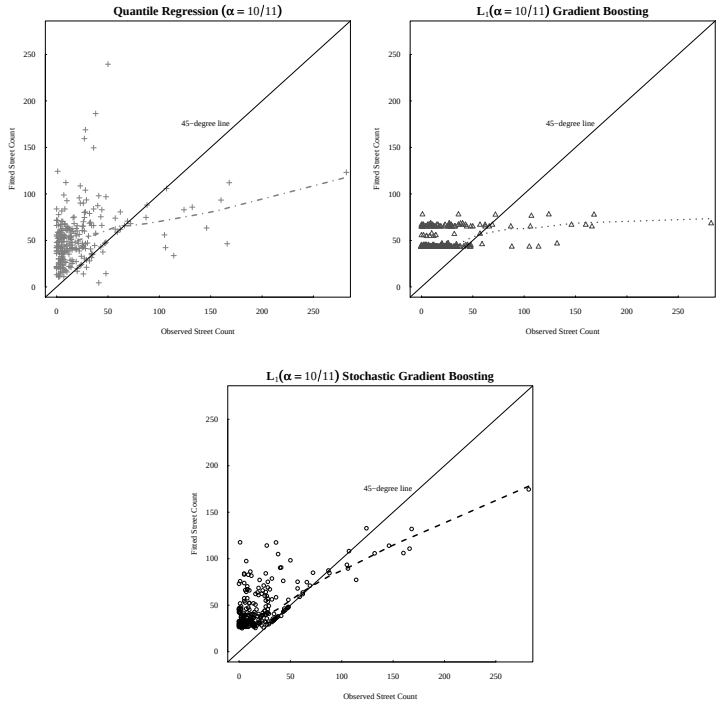

Figure 3: Fitted versus observed street counts using quantile regression, $L_1(\alpha)$ gradient boosting, and $L_1(\alpha)$ stochastic gradient boosting, subject to a 10 to 1 cost ratio (i.e., $\alpha = 10/11$).

4.2 Conditional Distribution Diagnostics

With 10 predictors, a highly unbalanced response distribution, and abrupt spatial variation in the data, the boosted models' conditional distribution diagnostics are practical and necessary to understanding relationships between the response and the predictors. Since underestimation and overestimation costs are built into each step of $L_1(\alpha)$ boosting, partial plots and variable importance measures can be examined in the same manner as when employing L_1 boosting. These results are especially important if stakeholders are inclined to give causal interpretations to the associations.

One may assume that the partial relationships between the response and

each predictor exhibit similar directional behavior and are nothing more than vertical shifts in the conditional response's magnitude. An analogous argument might be made regarding variable importance: if a predictor is important using symmetric costs, then perhaps the same is true using asymmetric costs. If these inferences are correct, cost-sensitive partial and predictor importance plots are less critical. Yet Figures 4 and 5 demonstrate that predictors' relationships with the response are not necessarily the same across cost ratios, underscoring the need to examine the conditional distribution diagnostics for each cost ratio of interest.

4.2.1 Partial Relationships

To show partial relationships between the response and each predictor, Friedman (2001) describes a weighted tree traversal method to "integrate out" all predictor variables, excluding the predictor(s) of interest (see also, Ridgeway, 2007). Figure 4 shows partial relationships between the response and each predictor for five different cost ratios. Since each of the predictors exhibits real values, each partial relationship is shown using a two-dimensional smoother.[15] For cost ratios of 1 to 10 and 1 to 5, all of the partial relationships are nearly flat, a result consistent with the small variation in tract-level estimates reported in Figures 1 and 2. Using symmetric L_1 boosting, street counts increase with respect to *PctVacant* between 0 and 10 percent, and street counts decrease with respect to *PctOwnerOcc* between 20 and 60 percent. Pragmatically, all other partial relationships are close to null.

When underestimating *StTotal* is more costly, the conditional response can vary substantially with respect to several other predictors in addition

[15]The `gbm` library estimates the partial response at equally-spaced values (by default, 100) spanning the range of the predictor but independent of the predictor's empirical density. As a result, decile rugs are shown at the bottom of each plot for each corresponding predictor to better understand the distribution of each predictor. For example, the vacancy rate is 33 percent for one tract, 43 percent for another tract, and less than 20 percent for all other tracts. For *PctVacant* greater than 20 percent, it is difficult to determine the extent to which these partial smoothers are robust because they are based on so few points.

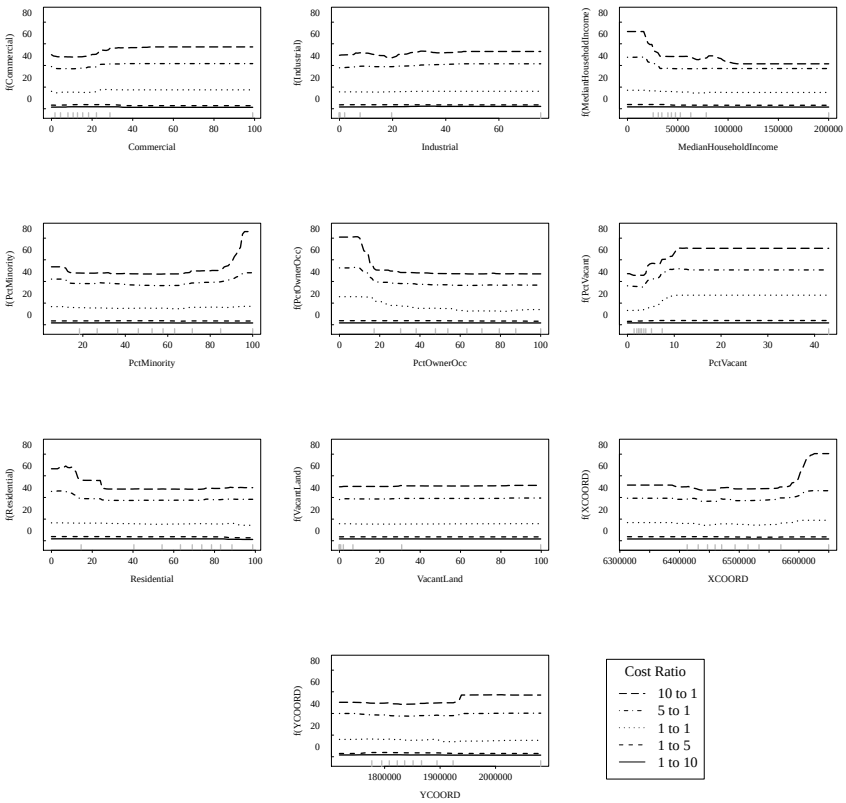

Figure 4: Partial dependence plots from $L_1(\alpha)$ stochastic gradient boosting.

to the housing vacancy rates and the fraction of owner-occupied units. For example, using a 10 to 1 cost function, street counts are indifferent to *PctMinority* until approximately 90 percent but increase substantially between 90 and 100 percent. Street counts decrease in a stepwise manner with respect to *MedianHouseholdIncome*; we see plateaus for incomes between \$0 and \$15,000, \$30,000 to \$75,000, and \$100,000 and above.

4.2.2 Variable Importance

One may be interested in identifying which predictors are "important" to fitting the conditional response. One measure variable importance is the reduction in loss attributed to each predictor for various cost ratios. Friedman (2001) and Ridgeway (2007) describe how to compute the "relative influence" as the empirical reduction in squared error in predicting the gradient across all node splits on predictor j, divided by the total reduction in error across all splits.

Even if the response and predictor j are completely unrelated, it is still possible for the predictor to be selected to split a regression tree node. Provided there is at least one split on predictor j, the empirical influence will not be zero. How then, does one know the extent to which a predictor's influence is by chance? Along the same lines as in random forests (Breiman, 2001), in which importance is computed by shuffling each predictor in turn and comparing the change in error, we employed the following steps to estimate each predictor's "baseline relative influence:"

1. For a given predictor, randomly permute the values. Keep all other predictors' values as is.
2. Construct a boosted model using the modified data in (1) and compute the relative influence for the shuffled predictor. Apply the same tuning parameter settings and means for estimating a sensible number of iterations.
3. Repeat steps (1) and (2) many times, each time computing the relative influence of the shuffled predictor.[16]
4. Compute the baseline relative influence as the average relative influence from steps (1)-(3).
5. Repeat steps (1)-(4) for each predictor in turn.

[16]For $\alpha \in \{1/10, 1/5, 1/2, 5/6, 10/11\}$, we repeated steps (1) and (2) 50 times per predictor.

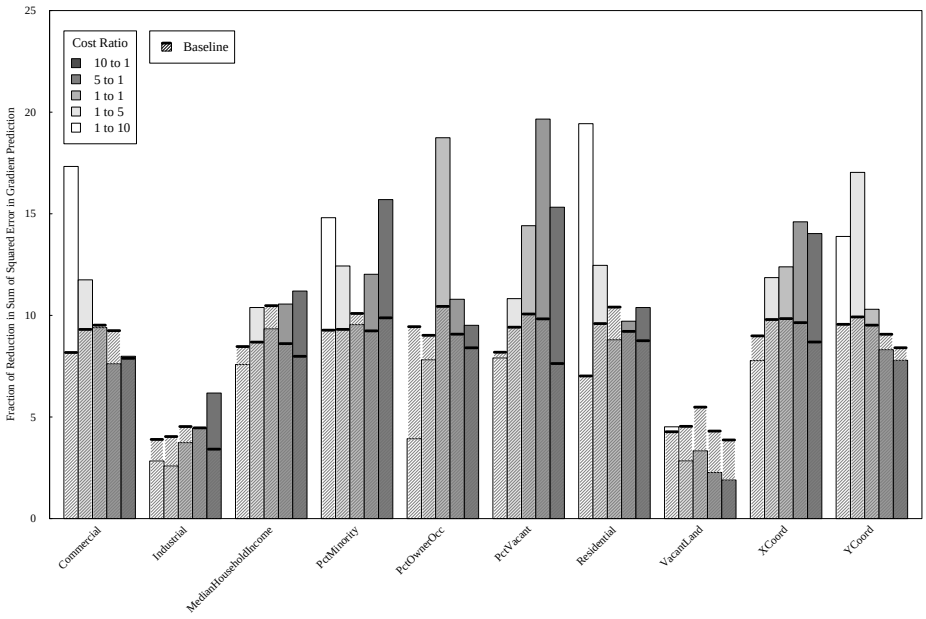

Figure 5: Variable importance from $L_1(\alpha)$ stochastic gradient boosting.

Figure 5 shows each predictor's empirical and baseline relative influence values subject to five different cost ratios. If a predictor's baseline relative influence (denoted by a thick black line and the diagonally shaded areas) is larger than its empirical influence, this suggests that the contribution to the model is happenstance. Just as in the partial plots, we learn that a predictor's relative influence is not necessarily similar across cost functions. This can be a very important practical matter insofar as stakeholders come to accept or reject the homeless estimates depending on whether predictors "make sense."

One should also be mindful of the difference between the overall reduction in error from $t = 0-$ at which all estimates are equal to the grand α quantile of $StTotal$ $-$ to the "optimal" number of iterations. If the total reduction

in error is very small, then the absolute influence will be minimal. It follows that the differences between each fitted response value and the initial constant will likely be small as well. Under these circumstances, the relative influence results are inconsequential. Such is the case for boosted models subject to 1 to 10 and 1 to 5 costs. Figures 1, 2, and 4 suggest minimal variation in fitted and predicted counts; substantively, the relationships between *StTotal* and each predictor are null. Importance statistics subject to these two cost ratios are reported primarily for demonstrative purposes.

Using symmetric costs, *PctVacant* and *PctOwnerOcc* are relatively important, collectively accounting for over 35 percent of the loss reduction. *PctVacant* is also important when the cost ratio is 5 to 1 or 10 to 1, along with *PctMinority* and to a lesser extent *XCoord* and *MedianHouseholdIncome*. Each of these predictors' relative influences is high compared to other predictors' importance statistics and is well above their respective baseline influences. Conversely, *PctOwnerOcc* is much less important when underestimation is penalized more heavily, evidenced by its smaller relative influence and proximity to the baseline relative influence.

5 Discussion

$L_1(\alpha)$ stochastic gradient boosting is a potentially useful statistical tool for ensuring adequate allocation of services related to the homeless. Practitioners might find it useful to build multiple boosted models for various cost functions and examine the range of imputed counts for a specific tract in order to make policy decisions. Suppose a homeless service provider or local police department considers it critical to identify tracts that have over 100 homeless people; the former might aspire to add beds to the nearest shelter, and the latter might well allocate additional officers to this area. Assume that a particular tract's imputed count is 30 using 1 to 1 costs and 150 using 10 to 1 costs. Such stakeholders may insist on performing a full enumeration

in this tract because these two imputed counts have very different resource implications. Alternatively, if the imputed counts using these respective cost ratios are 30 and 40, a full enumeration may not be worth the trouble if the difference is considered inconsequential.

Among the 11 tracts with over 100 homeless, stochastic gradient boosting subject to a 10 to 1 cost ratio yields a better prediction error than gradient boosting or parametric quantile regression. Still, 9 of the 11 tracts are underestimated, and the prediction error tends to increase with respect to the observed count. It is reasonable to assume that among unvisited tracts with over 100 homeless, imputed counts will be similarly biased. In practice, one way to further reduce this problem is by assigning larger "population weights" *a priori* to training data tracts with large street counts. The population weights increase the frequency of specific observations if they are selected in step 2b of the algorithm described in 3.1. One assumes − and perhaps rightfully so − that some tracts are inherently more important than others. If larger weights are assigned to tracts with high street counts, then fitted and imputed counts will also increase. A toy example is provided in Appendix A.

In addition to evaluating imputed counts, suppose stakeholders (e.g., LAHSA) want to use response-predictor relationships to determine which unvisited tracts might require the most resources. Figure 4 suggests that areas with some combination of high non-Caucasian populations, high vacancy rates, low median household incomes, and low rates of owner-occupied housing may be indicators of high homeless populations. Based on Figure 5, *PctVacant* and *PctMinority* are especially key to identifying areas potentially in need of services.

6 Conclusion

Indeed, this case study features a number of characteristics that make the analysis challenging. Although there are relatively few tracts with large homeless counts, these are likely the most important tracts to fit reasonably well − without overfitting the data − so that unvisited tracts with potentially high counts are identified. In addition, Los Angeles County exhibits considerable heterogeneity and abrupt spatial changes in terms of land usage and demography. Lastly, the wide range of stakeholders would likely assign various costs to over/under-counting during the estimation and imputation processes. We believed that a cost-sensitive ensemble statistical learning procedure be appropriate because i) we did not presume to understand the underlying mechanisms of the conditional street count distribution, ii) we aspired to get favorable results in terms of prediction error for specified regions of the response, and iii) we wanted to understand how specific regions of the conditional response were related to the predictors. $L_1(\alpha)$ stochastic gradient boosting allowed us to address all of these issues.

There are a handful of practical statistical issues born out of this case study. First, one might argue that a "cost-sensitive Poisson" loss function is a more appropriate procedure for the homeless data because the outcome is a count. A key issue, then, is whether L_1 or L_2 loss is more responsive to the data imputation task at hand and to the quality of the data. In our case, a few very large observed counts would likely dominate the analysis under L_2. Whether this is good or bad depends on the accuracy of the few very large counts and on the policy matter of how much those large counts should be permitted to affect the imputations. We take no strong position on either issue, but we have concerns from past research on homeless enumerations that the count data could contain significant error (Cowan, 1991; Cordray, 1991; Rossi, 1991; Wright and Devine, 1992). And, we find that boosting the $L_1(\alpha)$ loss function incorporates cost considerations in a straightforward and easily interpretable manner.

There is also the matter of statistical inference, a topic we glossed over in Section 4.2.2 by estimating each predictor's baseline relative importance. To our knowledge, statistical inference remains a largely unsolved problem for stochastic gradient boosting and statistical learning in general (Leeb and Pötscher, 2005; Leeb and Pötscher, 2006; Berk, Brown and Zhao, 2010). We have explored the properties of a procedure that wraps cost-sensitive boosting in bootstrap sampling cases. Although this seems to provide some useful information on the stability of our imputed values, we do not think it addresses the fundamental problems identified by Leeb and Pötscher.

Finally, $L_1(\alpha)$ boosting brings to light the issue of choosing the "right" tuning parameters, a topic explored by Mease and Wyner (2008). While the number of splits has been researched extensively (for example, Schapire, 1999; Friedman et al, 2001; Bühlmann and Yu, 2003; Ridgeway, 2007), research on the impact of different terminal node sizes is minimal thus far. Unlike estimates subject to Poisson or Gaussian loss, which are functions of *all* of gradients within each terminal node, a $L_1(\alpha)$ terminal node estimate is the quantile of gradients residing in terminal node k_t. These estimates depend on just a very local region of points and can be highly dependent on the terminal node sizes and the way in which the quantile is estimated (for variants of quantile estimation, see Hyndman and Fan, 1996). The way in which $L_1(\alpha)$ stochastic gradient boosting performs using various quantile estimation procedures remains a topic for future research.

A Boosting the $L_1(\alpha)$ Distribution

Ridgeway (2007) specifies the boosted L_1 (Laplace) loss function as:

$$\Psi(f_t(x_i) : x_i \in S_{k_t}) = \left\{ \sum_{x_i \in S_{k_t}} \left| w_i(y_i - f_t(x_i)) \right| \right\} \Big/ \sum_{x_i \in S_{k_t}} w_i, \qquad (4)$$

where w_i is a pre-determined population weight for observation i that remains constant across all iterations. Altering (4) to allow for unequal costs, the loss function becomes:

$$\Psi(f_t(x_i) : x_i \in S_{k_t}) = \left\{ \alpha \sum_{\substack{x_i \in S_{k_t}, \\ y_i > \hat{f}_t(x_i)}} \left| w_i(y_i - \hat{f}_t(x_i)) \right| + (1-\alpha) \sum_{\substack{x_i \in S_{k_t}, \\ y_i \leq \hat{f}_t(x_i)}} \left| w_i(y_i - \hat{f}_t(x_i)) \right| \right\} \Big/ \sum_{x_i \in S_{k_t}} w_i, \quad (5)$$

which is an asymmetrically weighted absolute loss function if $\alpha \neq 0.5$.[17] For shorthand, denote $\Psi(f_t(x_i) : x_i \in S_{k_t}) = \Psi$. Then, the gradient becomes[18]

$$z_{ti} = -\frac{\partial \Psi}{\partial f_t(x_i)} = \begin{cases} w_i \alpha & : \quad y_i > \hat{f}_{t-1}(x_i) \\ -w_i(1-\alpha) & : \quad y_i \leq \hat{f}_{t-1}(x_i), \end{cases} \quad (6)$$

where the derivative is evaluated at $\hat{f}_{t-1}(x_i)$. We wish to find the value of ρ_{k_t} that minimizes Ψ subject to the loss function in (5):

$$\begin{aligned} \rho_{k_t} = \arg\min_{\rho_{k_t}} \Bigg\{ & \alpha \sum_{\substack{x_i \in S_{k_t}, \\ y_i > \hat{f}_{t-1}(x_i) + \rho_{k_t}}} \left| w_i(y_i - (\hat{f}_{t-1}(x_i) + \rho_{k_t})) \right| \\ & + (1-\alpha) \sum_{\substack{x_i \in S_{k_t}, \\ y_i \leq \hat{f}_{t-1}(x_i) + \rho_{k_t}}} \left| w_i(y_i - (\hat{f}_{t-1}(x_i) + \rho_{k_t})) \right| \Bigg\}, \end{aligned} \quad (7)$$

[17] With this distribution, the estimate $\hat{f}$ is in the same units as y; therefore, over/under-estimation are determined by comparing the two. Estimates in some distributions, such as Poisson, are in terms of logits and must be exponentiated to be on the same scale as y.

[18] Under the usual L_1 loss function, the gradient for observation i 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.

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

$$\frac{\partial \Psi}{\partial \rho_{k_t}} = \left\{ -\alpha \sum_{\substack{x_i \in S_{k_t}, \\ y_i > \hat{f}_{t-1}(x_i) + \rho_{k_t}}} w_i + (1-\alpha) \sum_{\substack{x_i \in S_{k_t}, \\ y_i \leq \hat{f}_{t-1}(x_i) + \rho_{k_t}}} w_i \right\} \Big/ \sum_{x_i \in S_{k_t}} w_i \tag{8}$$

$$0 = -\alpha \sum_{\substack{x_i \in S_{k_t}, \\ y_i > \hat{f}_{t-1}(x_i) + \rho_{k_t}}} w_i + (1-\alpha) \sum_{\substack{x_i \in S_{k_t}, \\ y_i \leq \hat{f}_{t-1}(x_i) + \rho_{k_t}}} w_i. \tag{9}$$

In the right-hand side of (9), each summation reduces to the number of observations that are underestimated or overestimated, respectively. Let N_{k_t} denote the number of observations in terminal node k_t, and let n_{k_t} and $N_{k_t} - n_{k_t}$ be the number of underestimates and overestimates in the terminal node, respectively. For simplicity, assume that $w_i = 1$ for all i. Solving for n_{k_t}, the location parameter is

$$n_{k_t} = \alpha N_{k_t}. \tag{10}$$

The way in which unequal population weights affect the terminal node estimate is worthy of a toy example. Consider terminal node k_t with 5 equally-weighted observations with fitted gradients − the "working responses" − at $t-1$ of 0, 3, 5, 6, and 15. If we are estimating the median, then the terminal node estimate is 5. Now suppose that prior to constructing the boosted model, the observation with the fitted gradient of 15 at $t-1$ was instead assigned a population weight of 3. Then this observation's fitted gradient from $t-1$ will appear in node k_t three times, and the population-weighted median is 6.[19]

[19] At present, `gbm` does not allow for unequal population weights when employing the quantile distribution.

By weighting the loss function according to overestimates and underestimates, the fitted value of terminal node k_t is the α quantile of the N_{k_t} gradients. In each terminal node, there are approximately αN_{k_t} and $(1-\alpha)N_{k_t}$ gradients above and below ρ_{k_t}, respectively. For all $i = 1, \ldots, N$, $f_0(x_i)$ equals 0, and ρ_0 equals the α quantile of the response variable, y. Therefore, the fitted value for observation i after T iterations, $\hat{f}_T(x_i)$, equals[20]

$$\hat{f}_T(x_i) = quantile_\alpha(y) + \lambda \sum_{t=1}^{T} quantile_\alpha(z_{ti}). \tag{11}$$

Because $L_1(\alpha)$ is differentiable and there exists a solution that minimizes this loss (Hastie et al, 2001), we are able to incorporate costs into stochastic gradient boosting where the response is quantitative, and in some sense add a distribution to those provided in Friedman (2001).

References

[1] Berk, R.A. (2008). *Statistical Learning from a Regression Perspective.* New York: Springer-Verlag.

[2] Berk, R.A., Brown, L., and Zhao, L. (2010). Statistical Inference After Model Selection. *J. of Quantitative Criminology* (forthcoming).

[3] Berk, R.A., Kriegler, B. and Baek, J.H. (2006). Forecasting Dangerous Inmate Misconduct: An Application of Ensemble Statistical Procedures. *J. of Quantitative Criminology* **22**(2): 131-145.

[4] Berk, R.A., Kriegler, B. and Ylvisaker, D. (2008). Counting the Homeless in Los Angeles County. *IMS Collections. Probability and Statistics: Essays in Honor of David A. Freedman* (2): 127-141.

[5] Berk, R.A. and MacDonald, J. (2009). Policing the Homeless: An Evaluation of Efforts to Reduce Homeless-Related Crime. *Forthcoming.*

[20] Note that $z_{ti} = 0$ if observation i is not randomly selected as one of the N' observations in step 2b of the stochastic gradient boosting algorithm described in Section 3.1.

[6] Binder, H. (2009). `GAMBoost`: Generalized additive models by likelihood based boosting. `R` version 1.1.
URL http://www.r-project.org

[7] Bratton, W. and Knobler, P. (1998). *The Turnaround: How America's Top Cop Reversed the Crime Epidemic*. New York: Random House.

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

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

[10] Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statist. Sci.* **22**(4): 477-505.

[11] Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 Loss: Regression and Classification. *J. Amer. Statist. Assoc.* **98**: 324-340.

[12] Cordray, D.S., and Pion, G. M. (1991). What's Behind the Numbers? Definitional Issues in Counting the Homeless. *Housing Policy Debates* **2**(3): 587-616.

[13] Cowen, D.D. (1991). Estimating Census and Survey Undercounts Through Multiple Service Contacts. *Housing Policy Debates* **2**(3): 869-882.

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

[15] Culp, M., Michailidis, G., and Johnson, K. (2006). `ada`: An `R` Package for Stochastic Boosting. *J. of Stat. Software* **17**(2).

[16] 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.*

[17] 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**: 119-139.

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

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

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

[21] Harcourt, B.E. (2005). Policing L.A.'s Skid Row: Crime and Real Estate Development in Downtown Los Angeles (An Experiment in Real Time). *Univ. of Chicago Legal Forum.*

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

[23] Hothorn, T. (2009). `mboost`: Model-Based Boosting. `R` package 1.1-2. URL http://www.r-project.org

[24] Hyndman, R.J. and Fan, Y. (1996). Sample quantiles in statistical packages. *Amer. Stat.* **50**: 361-365.

[25] Koegel, P., Burnam, A., and Farr, R.K. (1988). The Prevalence of Specific Psychiatric Disorders Among Homeless Individuals in the Inner-city of Los Angeles. *Archives of General Psychiatry* **45**: 1085-1092.

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

[27] Koenker, R. (2009). `quantreg`: Quantile Regression. `R` package 4.38. URL http://www.r-project.org

[28] Kushel, M.B., Evans, J.L., Perry, S., Robertson, M.J., and Moss, A.R. (2003). No Door to Lock: Victimization Among Homeless and Marginally Housed Persons. *Arch. Intern. Med.* **163**(20): 2492-2499.

[29] Leeb, H. and Pötscher, B.M. (2005). "Model Selection and Inference: Facts and Fiction," *Econometric Theory* **21**: 21-59.

[30] Leeb, H. and Pötscher, B.M. (2006). Can one Estimate the Conditional Distribution of Post-Model-Selection Estimators? *The Ann. of Stat.* **34**(5): 2554–2591.

[31] Lopez, S. (October 16, 2005). Demons Are Winning on Skid Row. *Los Angeles Times*.

[32] Madigan, D. and Ridgeway, G. (2004). Discussion of "Least Angle Regression" by Efron, et al. *Annals of Statistics* **32**(2): 465-469.

[33] Mangano, P.F. and Blasi, G. (October 29, 2007). Stuck on Skid Row. *Los Angeles Times (Opinion Section)*.

[34] Mease, D., Wyner, A., and Buja, A. (2007). Cost-Weighted Boosting with Jittering and Over/Under-Sampling: JOUS-Boost. *J. of Machine Learning Research* **8**: 409-439.

[35] Mease, D. and Wyner, A. (2008). Evidence Contrary to the Statistical View of Boosting. *J. of Machine Learning Research* **9**: 131-156.

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

[37] R CORE DEVELOPMENT TEAM (2009). R: A Language and Environment for Statistical Computing. R Foundation for Computing, Vienna, Austria. ISBN 3-900051-13-5.
URL http://www.r-project.org

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

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

[40] Ridgeway, G. (2007). `gbm`: Generalized Boosted Regression Models. `R` package 1.6-3.
URL http://www.r-proejct.org

[41] Rossi, P.H. (1989). *Down and Out in America: The Origins of Homelessness*. U. of Chicago Press.

[42] Rossi, P.H. (1991). Strategies for Homeless Research in the 1990s. *Housing Policy Debates* **2**(3): 1029-1055.

[43] Sexton, J. (2009) `gbev`: Gradient Boosted Regression Trees with Errors-in-Variables. `R` package 0.1.1.
URL http://www.r-project.org

[44] Takeuchi, I., Le, Q.V., Sears, T.D., and Smola, A.J. (2006). Nonparametric Quantile Regression. *J. of Machine Learning Research* **7** 1231-1264.

[45] 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.

[46] Wilson, J.Q. and Kelling, G. L. (March 1982). Broken Windows: The Police and Neighborhood Safety. *Atlantic Monthly*: 29-38.

[47] Wright, J.D. and Devine, J.A. (1992). Counting the Homeless: The Census Bureau's 'S-Night' in Five U.S. Cities. *Evaluation Review* **16**(4): 355-364.

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