Predicting Homeless Population Dynamics in Los Angeles

Derek Yen

Math 199, Winter 2018
Advisor: Michael Lindstrom

Abstract

The UCLA Applied Mathematics Department is developing mathematical models to understand dynamics of homeless populations in the City of Los Angeles. Our goal is to gain insight into environmental factors which influence homeless people and might determine why a given city region, such as a census tract, has a larger homeless populations than others. By analyzing public data from the Los Angeles Homeless Services Authority, the United States Census Bureau, and other public databases, we develop predictive models for whether a census tract will have a large homeless population based on features of the tract. Topic models and correlations analysis provide insight into relationships between the variables and population. We also use neural networks to predict changes in the homeless population at different locations in the city from one year to another. We found logistic regression and population classification methods performed better than chance, and we discuss challenges with predicting populations and changes in populations beyond classification.

Contents

1 Introduction

Homelessness is a persistent issue in the Los Angeles area, with populations increasing from year to year. The Los Angeles Homeless Services Authority's point-in-time population count for 2017 was nearly 58,000, a 23 percent increase from 2016's total. In addition to the serious personal cost for people who are homeless and their families, homelessness can have far-reaching impacts on everyone in the area. For example, in December 2017, a cooking fire at a homeless encampment caused a fire which burned more than 400 acres[1]. We examine trends in the number of homeless people in particular areas of Los Angeles in order to better understand how homeless people distribute themselves. We are especially interested in the dynamics of homeless encampments, or areas with a large number of homeless individuals.

By attempting to predict which census tracts will have high homeless population counts based on physical features of the tracts, we can better understand what factors are most related to homelessness in the city and movement in the homeless population. If we can predict areas where homeless populations could increase from one year to another, organizations such as the Los Angeles Homeless Services Authority (LAHSA) would be able to distribute their resources more effectively. These predictions could also aid shelters and services in providing assistance to homeless individuals. Further information about how rent, income, and other economic factors might impact homeless population counts within a smaller geographic region could provide additional useful information for the city and other stakeholders. For example, when rent increases in a local area, is it associated with an increase in the local area population of homeless people?

Figure 1 shows the distribution of 2017 homeless population density in people per unit area of the 2160 Los Angeles County area census, using LAHSA data. This distribution is highly skewed, with a long right tail. The majority of census tracts in Los Angeles have few or no homeless people, some have a significant number, and some outliers have a very high population density of homeless people. Nearly a quarter of the census tracts had population densities less than 1 per square mile, while the highest homeless population density was 13,562 homeless people per square mile. These differences provide further motivation to understand why some areas have very high homeless population densities while others are much lower.

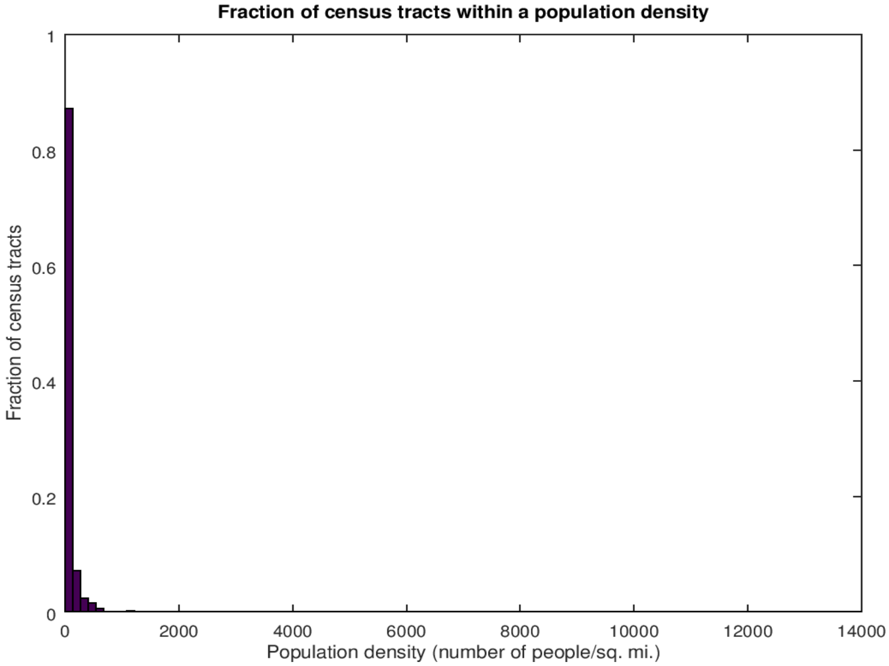


Figure 1: The histogram shows distribution of homeless population densities. Each bar shows the relative frequency, or the fraction of census tracts falling into a given bin, of people per square mile. While there are some significant outliers, more than 80 percent of tracts have homeless population densities which fit into the first bin, less than 200 people per square mile.

Figure 2 shows the distribution of homeless population densities for more typical census tracts, in which the highest census tracts and census tracts with population densities less than or equal to 1 were removed. Similarly, Figure 3 shows more fine-grained distributions by splitting the plots into two sets. The left shows the 1779 census tracts with population densities less than 100, and the right shows the 381 census tracts with population densities greater than or equal to 100. Frequency bars are still shown relative to the entire set of census tracts.

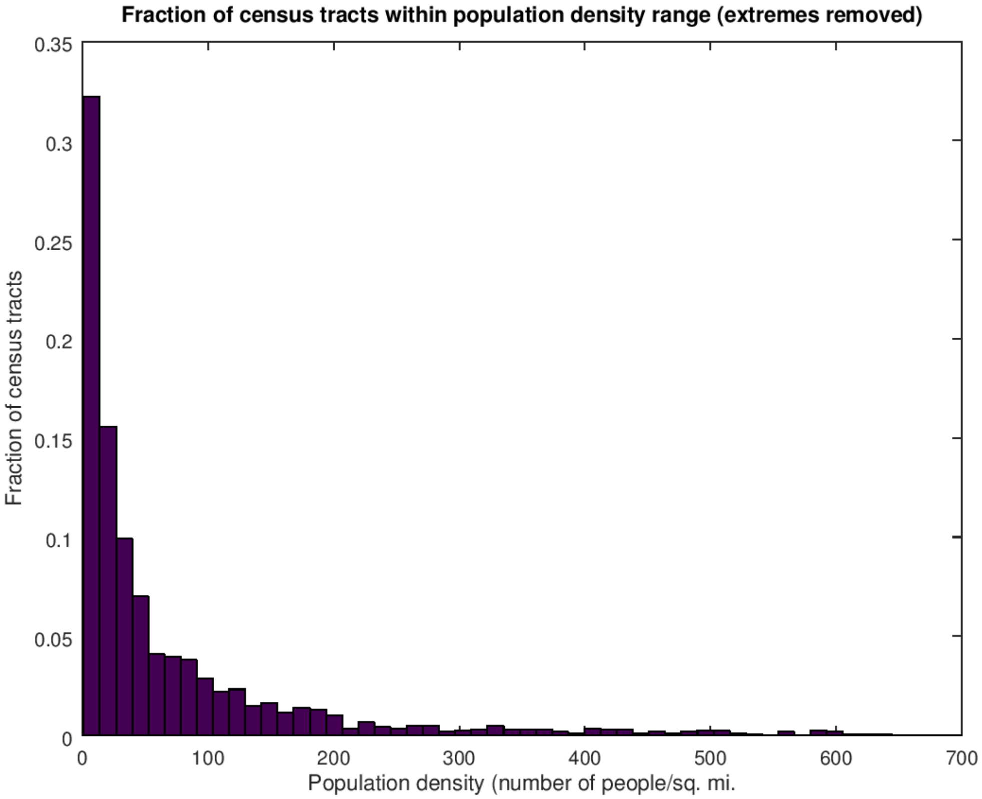
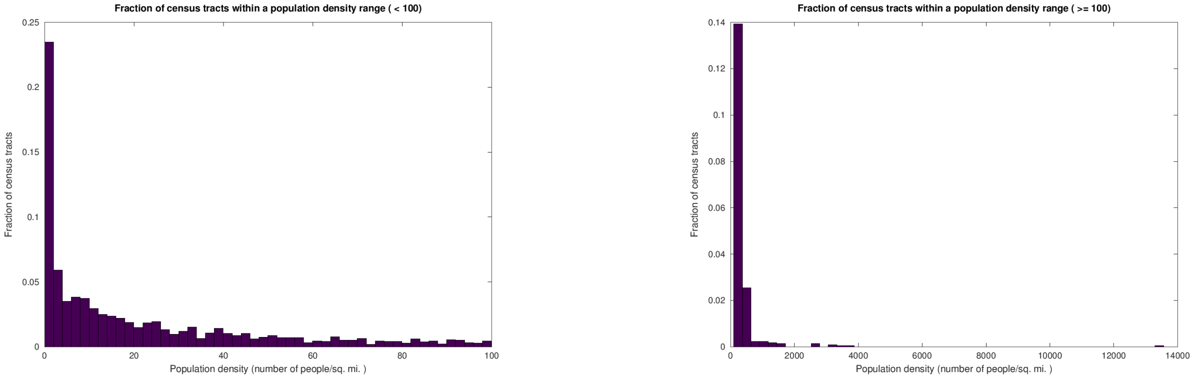


Figure 2: The histogram shows the distribution of homeless population densities, with the 25 highest census tracts and the 454 census tracts with population densities less than or equal to 1 were removed. Each bar shows the relative frequency, or the fraction of census tracts falling into a given bin, of people within a census tract. The two tracts with the highest raw population counts were removed for better visualization. While there are still some outliers, a majority of census tracts fit into the first bin, which corresponds to population counts between 0 and 10.



Lower range (population density < 100)

Upper range (population density ≥ 100)

Figure 3: The histograms show distribution of homeless population densities, split into two plots. The left plot shows distribution of the homeless population densities less than 100, split into 50 equal bins. The right plot shows distribution of the homeless population densities greater than or equal to 100, split into 50 equal bins. All bars display relative frequency, or fraction of census tracts falling into a given bin, where the fraction is out of the entire set of census tracts in Los Angeles County.

It would also be useful for stakeholders to be able to predict population changes in response to a specific event, such as the construction in Manchester Square. The Los Angeles International Airport recently acquired land including much of Manchester Square and has plans to construct a parking structure there, which will likely displace a large homeless encampment in the area. While initial efforts from stakeholders will focus on assisting the homeless people who are currently there, some of the homeless population may be forced to move to neighboring census tracts. The local features of neighboring census tracts could affect the likelihood of homeless populations increasing at these census tracts.

There has been past work focusing on homeless count accuracy and causal factors of homelessness. For example, James Wright and Joel Devine concluded that there is serious undercounting of the homeless population [2]. Kim Hopper et al. used plant-capture methods to study accuracy of homeless population counts in New York City, in which researchers employed decoys to estimate the proportion of homeless people which would not be visible to surveyors [3]. They found that 30-40% of people were not visible to surveyors. Additionally, Chris Glynn and Emily Fox investigated the relationship between housing costs and homeless while also modelling uncertainty in homeless counts [4]. Specifically, Glynn and Fox calculate posterior intervals of total homeless populations in different cities, based on Bayesian priors for accuracy of homeless counts, and find a relationship between rent and homeless populations in Los Angeles. Richard Berk, Brian Kriegler, and Donald Ylvisaker discuss methods for imputing homeless street counts for census tracts in Los Angeles County based on incomplete counts **??**.

We focus our work on the local dynamics of homeless populations within Los Angeles, movement, and changes in population between city regions. Our work studies several variables and their association with homeless population counts and population density within census tracts. Specifically, we use rent and housing indices, including the Zillow Rent Index (ZRI) and Zillow Home Value Index (ZHVI). We also used Google Maps and public databases from Los Angeles to compile data on the physical features of different census tracts the number of restaurants, coffee shops, homeless shelters, bus stops, reported crimes, and affordable housing units by census tract. Additionally, we include the median household incomes of each census tract from the U.S. Census Bureau's 2016 American Community Survey. Finally, we use the yearly reported homeless population counts from the LAHSA. With each of these variables, we create models which attempt to classify the size of the homeless population within a census tract and describe spatial and temporal dynamics of movement between census tracts.

This paper is organized as follows: Section 2 explains the different modelling techniques we used and the motivation behind using them. In section 3, we describe methods for extracting and transforming data for various parameters which we analyzed. Additionally, we specify conversions that we used and limitations in data quality. Section 4 lists results from each of the techniques we used, as well as some justification of different evaluation metrics used. With Section **??**, we discuss implications of the results from Section 4. We offer interpretations, evaluations, and limitations given the results from each model. Finally, we highlight important findings and suggest directions for future studies in section 5.

2 Techniques and Models

2.1 Spatial correlation analysis

Since one of our eventual goals is to develop a model describing how locations of homeless populations might change over distance, we examined correlations between total homeless populations and different factors in neighboring census tracts as a function of distance. We found all pairs of census tracts which had centroids within a given distance range apart. We select all census tracts whose centroids have a distance d between them, such that $\lambda_{i-1} < d < \lambda_i$. Each λ_{i-1} is a uniform distance apart from λ_i.

This spatial correlation is between the total population density of homeless people and the variable of interest, where the tracts are between λ_{i-1} and λ. Correlations were calculated between parameters for those pairs using (1) for each census tract pair (i, j) such that $\lambda_{k-1} < \text{distance}(i, j) < \lambda_k$ for all λ_k between 0 and 5 miles.

$$\text{corr}(X, Y, \lambda_k) = \frac{\text{Cov}(X, Y)}{(\sigma_X \sigma_Y)} \text{ over the set } S = (X_i, Y_j) \text{ such that } \lambda_{k-1} < \text{distance}(i, j) < \lambda_k \tag{1}$$

In this case, X is the feature of interest, such as the area density of coffee shops in a census tract, and Y is the population density of homeless people. For example, we selected the census tracts within a quarter mile of each other and look at correlations between the number of homeless people in one of the tracts and the number of coffee shops in the other tract, and vice versa. We found the correlations between the total number of homeless people in a census tract and the area density of each variable in another census tract for all pairs of census tracts within a particular radius range and plotted them against the lower end of the radius range.

2.2 Topic modelling

In order to gain insight into relationships between different variables, we decided to use nonnegative matrix factorization (NMF) as a method of topic modelling. Since the census features and populations make up a matrix of entries all greater than or equal to 0, NMF is a useful and relatively easily interpretable method for descriptive modelling and dimensionality reduction [5].

Denote a $\mathbb{R}^{\text{m}\times\text{n}}$ data matrix X with X_i as the ith column. Each row represents a census tract and each column represents different features of the tract. Before performing NMF, we normalize this matrix such that each value is between 0 and 1 using (2). Let $X_{i,min}$ be the minimum of the ith column, and $X_{i,max}$ be the maximum of the ith column. Thus, each component of X_i subtracts and is divided by the same scalars. This produces A_i, which is the ith column of the normalized matrix.

$$A_i = \frac{X_i - X_{i,min}}{X_{i,max} - X_{i,min}} \tag{2}$$

With a normalized $\mathbb{R}^{\text{mxn}}$ data matrix A and a lower rank k, our goal is to find an approximate matrix decomposition such that $A \approx WH$. In this case, $W \in \mathbb{R}^{\text{mxk}}$ and $H \in \mathbb{R}^{\text{kxn}}$, and $W, H \geq 0$. The H matrix represents an approximation of a k-dimensional nonnegative basis, in which each row is a latent factor. The W matrix represents the weights corresponding with each factor for each census tract. Using this system, the jth census tract can be described as a linear combination of each factor, as in (3).

$$A_j = \sum_k (W_{jk} H_k) \tag{3}$$

Typically, W and H are found by minimizing an objective function, such as (4).

$$\|(A - WH)\|_F = \sum_{i,j} (A - WH)^2_{ij}, \text{ where } W, H \geq 0 \tag{4}$$

For our calculation, we used the root-mean-squared residual as an objective function, shown in (5). This is just the Frobenius norm divided by the square root of the number of components. Calculations were done with the nnmf function in the MATLAB Statistics and Machine Learning Toolbox. The nnmf function uses an alternating least-squares algorithm, which randomly initializes components for the matrix W and updates the components of H by minimizing the objective function (5), while holding the W constant [6]. After H is solved using least squares, all negative components are set to 0, and the same process is repeated for W while holding H constant. Thus, one matrix is optimized by minimizing the objective function, then the other matrix is optimized, and the entire process is iterated until the resulting W and H matrices reach some level of tolerance. This process outputs W and H matrices as a nonnegative matrix factorization for A.

$$\frac{\|(A - WH)\|_F}{\sqrt{m \times n}}, \text{ where } W, H \geq 0 \tag{5}$$

2.3 Logistic regression

We used a logistic regression model to estimate the probability of a particular census tract having a high population area density of homeless people based on the features of each census tract.

With cutoffs obtained from visually inspecting histograms of the population area density of homeless people by census tract, we categorize each tract as having a high density of homeless people under the criteria in (6) where N_i is the area density of homeless people in the ith tract and N_{cut} is a population density cutoff. We calculating the 75th percentile of population area density of homeless people.

$$\begin{cases} Y_i = 0 \text{ if } N_i & < N_{cut} \\ Y_i = 1 \text{ if } N_i & \geq N_{cut} \end{cases} \tag{6}$$

Each row of the features matrix then has the form $X_{i.} = [x_i^0, x_i^1, ...x_i^m]$, where there are m features and x_i^j is the jth feature of the ith row. Each column is normalized to a z-score by subtracting the mean and dividing by the standard deviation.

The logistic regression model proposes that the probability of the ith census tract having a high population per square mile depends on a sigmoidal function, which is composed with a linear combination of the features of the ith tract. We define $\theta \in \mathrm{I\!R}^{\mathrm{m}}$ such that $X_{i.}^T\theta = \Sigma_{j=1}^m x_j\theta_j$.

In order to generate a number between 0 and 1 as our probability, we initially used a standard sigmoid function:

$$h(X_{i.}^T\theta) = \frac{1}{1 + e^{-X_{i.}^T\theta}} \tag{7}$$

However, in order to avoid machine error in cases where (8) is near 0, we modified it to (8):

$$h(X_i^T\theta) = \epsilon + \frac{1 - 2\epsilon}{1 + e^{-X_i^T\theta}} \tag{8}$$

where $\epsilon = 10^{-10}$.

To find an optimal θ for all observed values in the training set for Y given all of the observed values of the features in X, we estimate the maximum likelihood.

$$\begin{aligned} L &= \prod_{i=1}^{n} P(Y_i = y_i | X_i, \theta) \\ &= \prod_{i=1}^{n} (h(X_i^T\theta)^{Y_i})(1 - h(X_i^T\theta)))^{1-Y_i} \end{aligned}$$

The likelihood is maximized for the same value of θ for which the log-likelihood (9) is maximized.

$$\log(L) = \sum_{i=1}^{n}[(1-Y_i)\log(1-h(X_i^T\theta)) + Y_i \, \log(h(X_i^T\theta))] \tag{9}$$

We found the θ which minimized the negative log-likelihood using MATLAB's nonlinear minimization function fminsearch, which iteratively updates θ based on an objective function. According to MATLAB documentation, fminsearch can locate a minimum of a nonlinear function, although it is not guaranteed to find an optimal solution. In this case, it minimized the negative log-likelihood. This is equivalent to maximizing the log-likelihood. The function demonstrated a reasonable degree of convergence.

Since our sample size has only a small number of positive observations (i.e., when tract population density $N \geq N_{cut}$), we also implemented a penalized log-likelihood function as suggested by statistics literature [7]. To compute this, we need to penalize the objective function by a factor equal to the square root of the determinant of the Fisher information matrix, which is equivalent to the negative Hessian of the log-likelihood function evaluated at that value of θ [7]. The Hessian can be calculated by finding $H_{ij} = \frac{\partial \log(L)}{\partial\theta_i \partial\theta_j}$.

Using a penalized negative log-likelihood, our objective function for minimizing becomes:

$$\begin{aligned}\text{Obj} &= -\log(L) + \frac{1}{2}\log(\det(H)) \\ \log(L) &= \sum_{i=1}^{n}[(1-Y_i)\log(1-h(X_i^T\theta) + Y_i\log(h(X_i^T\theta))] + \frac{1}{2}\log(\det(H))\end{aligned}$$

We classified our predictions as a high population density of homeless people if each tract had a probability greater than a scoring cutoff, S_{cut}, and as a low population density of homeless people for each tract if it had a probability less than or equal to S_{cut}.

For this model, we limited the data to census tracts falling within the city of Los Angeles so that we could reliably use 10 different features of each tract: ZRI, ZHVI, median household income and the area densities of coffee shops, restaurants, shelters, crimes, affordable housing units, bus stops, and total population. We then trained both the maximum likelihood and the penalized maximum likelihood logistic regression models using both original data counts and adjusted data counts which sum coffee shops, restaurants, shelters, crimes, affordable housing units, bus stops, and total population for all census tracts within a radius of 1 mile.

2.4 Neural networks

We used artificial neural networks as a natural extension from logistic regression in order to model more complex interactions between different features and attempt temporal predictions.

Artificial neural networks are popular modelling methods which use nested combinations of functions as a model for a network [8]. The output of one function h^ℓ can then be used as the input of another function. Conceptually, each layer can be thought of as a different hidden process relating input data to unspecified intermediate factors and eventually to an output. We denote the number of layers as L, where the initial input is the 0th layer and there are hidden layers $1 ... L-1$. In Figure 4, the initial Input layer has four different inputs, which are all connected to each of the five nodes in Hidden layer 1. The connections each imply a set of coefficients, called weights, and an addition term, called a bias. We denote the input to layer ℓ as z^ℓ, which is then combined and passed to a function h^ℓ. Output at layer ℓ is denoted as a^ℓ. Weights for each layer are written using the matrix W^ℓ, where $W \in \mathbb{R}^{\mathrm{N}_{\ell-1} \times \mathrm{N}_\ell}$. In this notation, $N_{\ell-1}$ refers to the units in the $\ell-1$th layer and N_ℓ refers to the units in the ℓ layer. Those five nodes then connect to each of the five nodes in Hidden layer 2 with a different set of weights and biases. Finally, the five nodes in the final hidden layer, $L-1$ are combined to produce the output layer L. Although the output in Figure 4 has only one component, in principle, it can have any number of components and dimensions.

The process of iteratively computing the output at each layer to get a final output is called feedforward. Each function input $z^{\ell+1}$ for the $\ell+1$th layer is a linear combination of the vector variable a^ℓ from the ℓth layer, with a coefficient matrix W^ℓ called the weight matrix, and an additional constant vector b^ℓ which represents the biases.

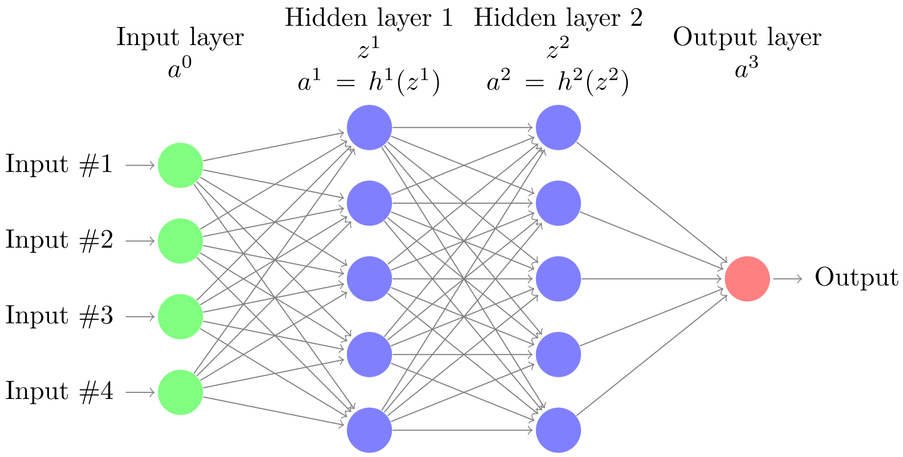


Figure 4: This diagram shows a neural network with four inputs, two hidden layers, and one output layer. We denote a^ℓ as the output from layer ℓ, z^ℓ as the input to layer ℓ, and h^ℓ as a function at layer ℓ, called an activation function. Each connection, drawn with arrows, represents a function, with the inputs receiving different weights and implicitly including an additive bias.

$$z_k^{\ell+1} = \sum_i W_{ik}^{\ell+1} a_k^\ell + b_k^{\ell+1}, \text{ such that } \ell = 1...L \tag{10}$$

This is shown in vector-component form in (10), where the kth component of $z^{\ell+1}$ is a linear combination of a^ℓ. $z^{\ell+1}$ can then be applied to a function h, called an activation function. At layer ℓ, we can find the output of the $\ell+1$th layer from (11) and (12) in matrix form.

$$z^{\ell+1} = W^{\ell+1} a^\ell + b^{\ell+1} \tag{11}$$

$$a^{\ell+1} = h(z^{\ell+1}) \tag{12}$$

There are several commonly used activation functions with different properties. For example, the sigmoid function shown in (7) outputs a number between 0 and 1. Similarly, the softmax function shown in (13) outputs a vector such that its sum is 1 and the kth component is between 0 and 1. Thus, it is easily generalizable to modelling a probability distribution with more than two categories, where the probability of the kth category is equal to the kth component of the vector. The relu function, (14), outputs a number greater than or equal to 0.

$$h(z_k) = \frac{e^{z_k}}{\sum_k e^{z_k}} \tag{13}$$

$$\begin{cases} h(z) = z \text{ if } z & \geq 0 \\ h(z) = 0 \text{ if } z & < 0 \end{cases} \tag{14}$$

As an alternative to relu (14), we consider a logarithmic function (15). This has the property of having small nonzero derivatives, shown in (16).

$$\begin{cases} h(z) = \log(1+z) & \text{if } z \geq 0 \\ h(z) = -\log(1-z) & \text{if } z < 0 \end{cases} \tag{15}$$

$$\begin{cases} h'(z) = \frac{1}{1+z} & \text{if } z \geq 0 \\ h'(z) = \frac{1}{1-z} & \text{if } z < 0 \end{cases} \tag{16}$$

We also used the hyperbolic tangent function, (17), which outputs a number between -1 and 1 and grows very slowly.

$$h(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{17}$$

Similar to logistic regression, the goal is to minimize a cost function. The cost function is a function of the estimated output value, which is found by the feedforward method of computing output values by computing each of the intermediate hidden layer values. For example, a common cost function is the mean-squared error, shown in (18). In particular, the input values are features of the ith census tract, and the output values would be a measure of the homeless population for the ith census tract. A second cost function is the mean of the L1 norm, or mean-absolute error, (19), which is the average absolute difference between the predicted and actual value. Another cost function, useful for multiclass classification, is the cross-entropy (20), which is an extension of the binary class log-likelihood (9) used in logistic regression. In (20), we sum over all classes j.

$$\begin{aligned} C &= \frac{1}{2N}\sum_i \left\| \left(\hat{Y_i}^L - Y_i \right) \right\|^2 \\ &= \frac{1}{2N}\sum_i \left\| \left(h(W^{L+1}a^L + b^{L+1}) - Y_i \right) \right\|^2 \end{aligned} \tag{18}$$

$$\begin{aligned} C &= \frac{1}{2N}\sum_i \left\| \left(\hat{Y_i}^L - Y_i \right) \right\| \\ &= \frac{1}{2N}\sum_i \left\| \left(h(W^{L+1}a^L + b^{L+1}) - Y_i \right) \right\| \end{aligned} \tag{19}$$

$$C = -\frac{1}{N}\sum_i^N \sum_j (Y_{ij}) \, \log(\hat{Y_{ij}}^L) \tag{20}$$

One method of minimization is gradient descent. This means that the set of all weights and biases should be updated against the direction of the cost function gradient for the values at that iteration, with a learning rate α. At the ith iteration of gradient descent, the weight matrix would be updated by subtracting α times the gradient with respect to the cost function, and the bias would be updated similarly.

$$W_{i+1}^{\ell} = W_i^{\ell} - \alpha \frac{\partial C}{\partial W_{i+1}^{\ell}} \tag{21}$$

In order to actually compute the gradient of the cost function, we find the gradient at each layer and apply the chain rule. The classic way to do this is through backpropagation [8].

If we only consider the output layer, then the gradient is given by (22). For example, using the mean-squared error as a cost function, this would be (23), where L is the output layer and Y is the actual population value at the ith census tract. Thus, $\frac{\partial C}{\partial a^L}$ is a row vector where each component has the form of (24).

$$\frac{\partial C}{\partial a^L}\frac{\partial h^L}{\partial z^L} \tag{22}$$

$$\frac{\partial C}{\partial a_j^L} = \frac{\partial}{\partial a_j^L}\sum_j (a_j^L - Y_j)^2 \tag{23}$$

$$= 2(a_i^L - Y_i) \tag{24}$$

Given the derivative of the cost with respect to one layer, we next find the derivative of one layer with respect to another, $\frac{\partial a^\ell}{\partial a^{\ell-1}}$.

$$\sum_k \frac{\partial a_i^\ell}{\partial z_k^\ell} \frac{\partial z_k^\ell}{\partial a_j^{\ell-1}} = \frac{\partial a^\ell}{\partial z^\ell} \frac{\partial z^\ell}{\partial a^{\ell-1}} \tag{25}$$

Both are matrices of some dimension, and taken componentwise, Equation 25 can be written in terms of Equations 26 and 27. Thus, $\frac{\partial a^\ell}{\partial a^{\ell-1}} = \frac{\partial h^\ell}{\partial z^\ell} W^\ell$.

$$(\frac{\partial a^l}{\partial z^l})_{ij} = \frac{\partial h_i^l}{\partial z_j^l} \tag{26}$$

$$\begin{aligned}(\frac{\partial z_i^\ell}{\partial a_j^{\ell-1}}) &= \frac{\partial}{\partial a_j^{\ell-1}} \sum_k W_{ik}^\ell a_k^{\ell-1} &(27)\\ &= W_{ij}^\ell &(28)\end{aligned}$$

We also find the gradient of the cost with respect to the bias and the components of each of the weights. (29) and (30) show that the gradient of the cost with respect to the bias at layer ℓ is just the gradient with respect to the function value itself at layer ℓ.

$$\frac{\partial C}{\partial b^\ell} = \frac{\partial C}{\partial a^L} \frac{\partial a^L}{\partial a^{L-1}} ... \frac{\partial a^\ell}{\partial b^\ell} \tag{29}$$

$$\frac{\partial a^\ell}{\partial b^\ell} = \frac{\partial}{\partial b^l}(W_{ik}^\ell a_k^{\ell-1} + b_i^\ell) = \delta_{ik} \tag{30}$$

where δ_{ik} is the Kronecker delta function, in which only those ik components are 1, and otherwise they are 0. The result is thus the identity.

The gradient of function at layer ℓ with respect to the weights at layer ℓ is:

$$\frac{\partial a^\ell}{\partial W^\ell} = \frac{\partial a_i^\ell}{\partial W_{pq}^\ell} \qquad = \frac{\partial}{\partial W_{pq}^\ell}(W_{ik}^\ell a_k^{\ell-1} + b_i^\ell) = \delta_{ip}\delta_{kq} a_k^{\ell-1} \qquad = \delta_{ip} a_q^{\ell-1} \tag{31}$$

where δ is the Kronecker delta function.

Given $\frac{\partial a_i^\ell}{\partial W_{pq}^\ell}$, we compute the gradient of the cost with respect to the components of each of the weights for layer ℓ:

$$\frac{\partial C}{\partial W_{pq}^\ell} = \frac{\partial C}{\partial a^L} \frac{\partial a^L}{\partial a^{L-1}} ... \frac{\partial a^{\ell+1}}{\partial a^\ell} \tag{32}$$

where L is the output layer and $\frac{\partial a^{\ell+1}}{\partial a^\ell} = \sum_i \frac{\partial C}{\partial a_i^\ell} \frac{\partial a_i^\ell}{\partial W_{pq}^\ell}$.

Using these gradients at each layer and iteratively finding the product through the chain rule, the weights and biases of the neural network can be updated at each training step with gradient descent. In variants such as stochastic gradient descent, a random batch is selected from the training set at each iteration, and the network is updated with the gradient of that random batch instead of the entire training set.

Similar to the logistic regression models, we limited input data to the 934 census tracts falling within the city of Los Angeles so that we could reliably use 10 different features of each tract: ZRI, ZHVI, median household income and the area densities of coffee shops, restaurants, shelters, crimes, affordable housing units, bus stops, and total population.

We tried classification and population regression under two model branches for predicting populations of homeless people in a census tract using the Python implementation of Tensorflow, a popular machine learning system [9], to construct shallow neural networks.

First, we moved from binary classification in the logistic regression model to bucketing the homeless populations from one year into multiple interval classes. Instead of a single cutoff value, we used cutoff values to set off intervals such that the population of homeless people in the ith census tract falls into only one such class $(0, S_{cut1}], (S_{cut1}, S_{cut2}], (S_{cut2}, S_{cut3}]$, etc. We also tried regression models to predict exact populations of each census tract using a training-testing split of 2017 data.

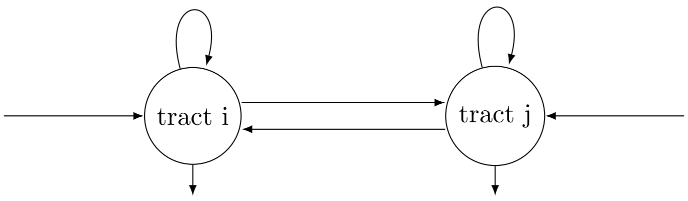


Figure 5: Considering only two census tracts, i and j, the probability of a person staying in tract i from one year to the next is represented by the self-edge for tract i, the probability of a person transitioning from tract i to tract j is represented by the edge from i to j, and the person exiting the population is represented by the edge pointing away from tract i. There is also a probability of a new person entering the population at tract i, represented by the edge pointing to tract i. This would be similar for tract j, and can be extended to the entire set of tracts.

We used models which attempted to predict the change in population from one year to another. We were motivated by a Markov model, where a person in the Los Angeles homeless population may stay in the same tract or transition to another tract from one year to the next. For example, a homeless person in tract i has a probability of transitioning out of the homeless population, a probability of remaining in the same tract i given that they stay homeless, and a probability of transitioning from census tract i to tract j. If a person transitions out of the homeless population from one year to the next, then they may be no longer homeless or have left the Los Angeles region. A new person may also enter the homeless population at tract i. Given the homeless population of tract i at year t, the features of each tract i, and the total homeless populations of each tract j within a given radius λ away from tract i, we attempt to predict the population at year $t+1$. These temporal predictive models were trained using population changes from 2015-2016, and evaluated by testing on 2016-2017 population changes. We initially tried classification on whether or not the population in a census tract will increase or decrease. Finally, we also tried regression models to predict exact changes in populations of each census tract, training on data from 2015-2016 and testing on 2016-2017.

For each approach, we tested changing various parameters within a simple neural network framework to improve the model. We compared simple feature normalization to normalizing the features using principal component analysis. Additionally, we tested including nonlinear input variables, such as the square of the features of each census tract. For each hidden layer, we tested different activation functions, such as relu (14), sigmoid (7), log (15), and tanh (17), as well as the number of units per hidden layer. For classification, we used an output layer of softmax or sigmoid, and for regression, we used an output layer of either relu or the identity matrix. Finally, we tested parameters for optimization. As our cost function, we considered mean-squared error and mean absolute error for regression and cross-entropy for classification. We adjusted initial learning rate, batch size, and optimization method. The two optimization methods we tested were stochastic gradient descent, and Adam [10], another popular stochastic gradient method which adaptively updates parameters. Adam calculates estimates of the mean and variance of the gradient and uses the ratio of the two to adjust the effective learning rate α for each parameter [10].

Table 1: Model comparisons

Input layer **(data preprocessing)**	**Hidden layer** **architecture**	**Output layer** **architecture**	**Optimization** **parameters**
-PCA -Normalization -Nonlinear inputs	-Activation function: (relu, sigmoid, log, tanh, softmax) -Layers -Units/layer	-Relu, identity, softmax	-Cost function -Learning rate -Optimizer: (stochastic gradient descent, Adam) -Batch size

3 Data Management

The data for our research comes from several sources: the Los Angeles Homeless Services Authority (LAHSA), the housing and rental website Zillow, the public data platform DataLA, the Los Angeles County Metropolitan Transportation Authority, and the 2016 American Community Surveys.

LAHSA was our primary source for data about estimates of homeless populations. Data for 2015, 2016, and 2017 is available for public use and includes counts by census tracts. Estimates describe counts of both unsheltered and sheltered homeless people, with additional data describing features of the homeless population itself, including the number of people in tents, cars, vans, and makeshift shelters. These population counts have significant limitations. For example, while the LAHSA executive summary lists the 2017 total unsheltered and sheltered homeless population in Los Angeles County as 57,794; however, the sum of all LAHSA sheltered and unsheltered homeless counts within its census tract data is only 49,698.

All data on census tract boundaries came from shapefiles on Los Angeles Geohub, a public platform for geolocation data from Los Angeles County.

The two main housing indices that we used were Zillow Rent Index (ZRI) and Zillow Home Value Index (ZHVI). The former tracks the median rent while the latter tracks the median home value within a certain area. Both of these are presented on both the zip code, city-wide, and neighborhood level. We used zip code data since it was more specific than the city level and had clearer boundaries for Los Angeles than the neighborhood level. These data were then converted to census tracts by locally averaging over zip codes.

For information on restaurants, we used location data available on DataLA, an open data platform maintained by the city of Los Angeles. According to DataLA, the restaurants were based on a listing of active businesses currently registered with the Los Angeles Office of Finance. Street addresses for those businesses were converted into latitude/longitude points using Google Maps Geocoding API through Python scripting. Locations were converted from latitude/longitude points to California State Planar feet using MATLAB code "SP_Proj" [11] that we obtained from Mathworks. Each state plane (x,y) coordinate, corresponding to a restaurant, was assigned to a census tract in MATLAB using Los Angeles shapefiles.

Locations of coffee shops were obtained by iteratively searching for coffee shops within a small radius using the Google Places API through Python scripting, which produced street addresses and latitude/longitude points. We then assigned these latitude/longitude points to census tracts using a similar process as with restaurants.

Parking citation data was also available on DataLA as a dataset with fields for issue date, issue time, vehicle body style, violation code, violation description, and street addresses. Those street addressees were converted into latitude/longitude points with the Google Maps Geocoding API. We primarily considered citations for overnight parking.

For information on homeless shelters and services, we used data available from Los Angeles Geohub. According to the website, this dataset is maintained by the County of Los Angeles Location Management System, and it included descriptions of services and locations in latitude/longitude.

Information about locations of crimes was obtained from DataLA and provided by the Los Angeles Police Department. We took each latitude/longitude coordinate, transformed to California State Planar feet and assigned a corresponding census tract in MATLAB using a similar process as above.

Locations of bus stops came from the Los Angeles County Metropolitan Transportation Authority's website Metro Developer and are accurate as of June 2017. These were converted to state feet coordinates and assigned to census tracts using the same process as above.

We obtained estimated total populations per census tract for the year 2015 from DataLA. According to the website, this dataset was compiled by the County of Los Angeles Department of Mental Health.

For information on affordable housing units, we used a dataset available on DataLA and provided by the Los Angeles Housing and Community Investment Department which described affordable housing units produced, rehabilitated, or under construction in the city of Los Angeles from the year 2003 onward. Locations were originally in latitude/longitude coordinates. They were converted to state feet and assigned to census tracts using MATLAB.

Finally, we obtained estimates of the median household incomes per census tract as of 2016 from the American Community Survey, a five-year estimate conducted by the United States Census Bureau. Geographic data was downloaded in geodatabase format and coverted to a spreadsheet using the open source geographic information systems software QGIS.

Distances between census tracts were calculated by as linear distances between the centroids of th tracts. These results were only estimates, since at times this method lacked precision. Census tracts are not uniform polygons and so depending on the case, using centroids to measure distance is not always accurate. In rare examples, centroids may not even fall within a given tract or zip code depending on the shape it takes.

There are some limitations and challenges inherent in working with each of these publicly available datasets. For example, restaurant and crime locations are from a citywide databases while the counts of homeless people from LAHSA are countywide. Although we attempt to limit our analysis to city census tracts when it depends on data that came from a city database, there are some edge effects where the centroids of census tracts do not fall within city boundaries. The coffee shops are pulled from Google Places, which should not be considered a comprehensive list. Furthermore, some of these data are imperfect measures. All location data also went through multiple conversions, and some were approximations; for example, while most of our data is consistent with its original form, a small portion is approximated if it falls in between two census tracts or just outside of one.

To address a small number of census tracts (less than 2 percent of the 934 tracts falling within city boundaries) for which there were missing values, we filled in a local average of census tracts within either 5 or 10 miles. As a test, we generated local averages of all other census tracts, and found that these estimates had a low difference from the actual data on average when checked against data for which we had values.

Furthermore, some data is recorded differently throughout the years. For example, LAHSA data is available for 2013, 2015, 2016, and 2017, however, in 2013 results are compiled by service planning area rather than census tract in the 2013 data. Due to these and other limitations, there are some factors that we were not able to change. For example, while some 2017 data was available for us to use, other data is only updated every couple of years so our compiled data is not consistent in terms of time.

4 Results

4.1 Spatial correlation analysis

We computed correlations over distances between population density per area and the area density of parameters with physical counts. These were coffee shops, restaurants, shelters, crimes, housing units, bus stops, and 2015 population. We considered data for pairs of census tracts (i, j) such that $\lambda_{k-1} < \text{distance}(i, j) < \lambda_k$ for all λ_k between 0 and 5 miles. The difference between each λ_{k-1} and λ_k was binned at a quarter of a mile. We plot these results here.

The correlations over distance demonstrate that variables that we examined had positive but decreasing correlation over distance. Since these correlations remain positive for some distance and are mostly monotonic, this confirmed our expectation of some nontrivial relationship between the area density of homeless people in a census tract and the features of neighboring census tracts. The relationship decreases to near 0 as the distance increases to 5 miles. Correlations lower than 0.4 are generally not considered meaningful, and below 0.2 are essentially no correlation.

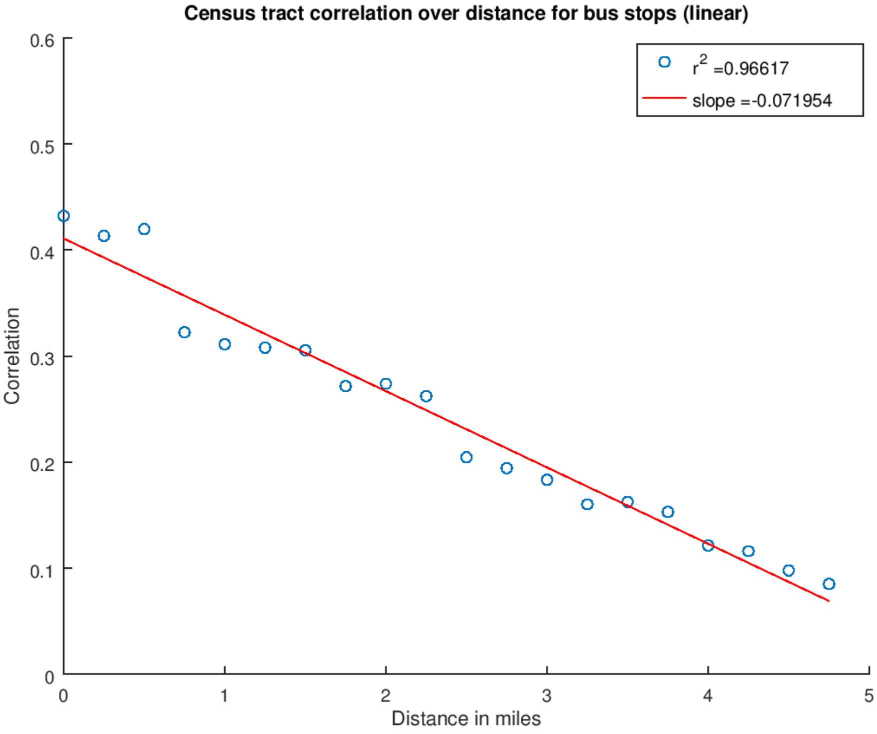


Figure 6: The plot shows correlation over distance between area densities of bus stops and total homeless people on a linear scale.

The correlation over distance between homeless density and the area density of bus stops had a strong linear relationship, with a modest correlation over 0.4 within 0.5 miles of the tract and decreasing slowly with distance. Past 1 mile, most of the correlations become less meaningful.

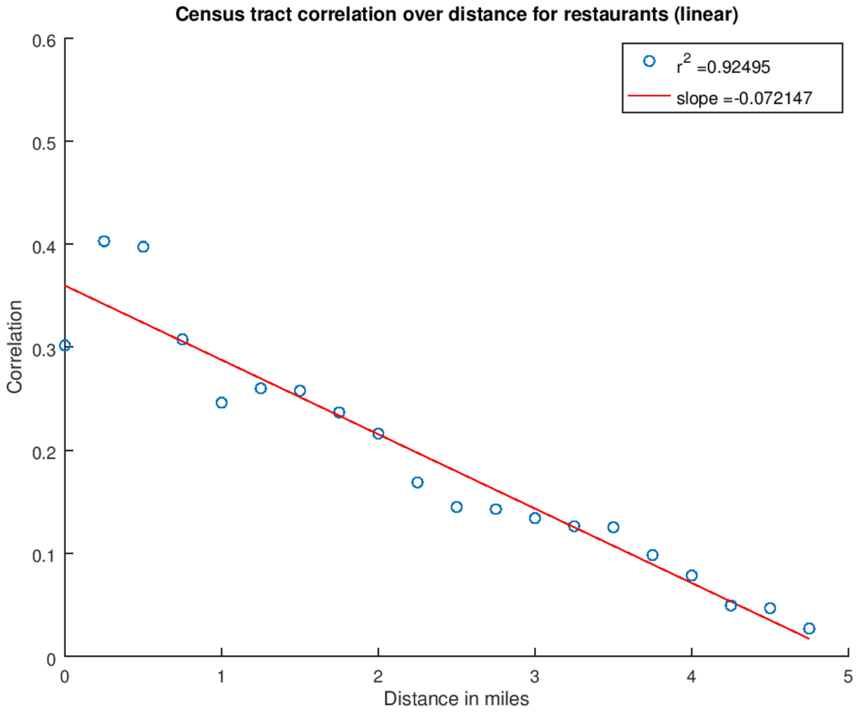


Figure 7: The plot shows the correlation between area densities of restaurants and total homeless people on a linear scale.

Correlation for restaurants increased within the first 0.5 miles of the tract then gradually decreased as distance increased greater than 0.5 miles. Thus, Figure 7 suggests that there are higher homeless populations in census tracts neighboring tracts with high numbers of restaurants, but the relationship is not as strong in these restaurant-dense tracts themselves. As expected, the relationship between coffee shops and the population density of homeless people behaved similarly to that of restaurants. There was an increase in correlation for distances of about 0.5 miles. However, correlations were not particularly meaningful for coffee shops, only reaching a peak correlation of about 0.25.

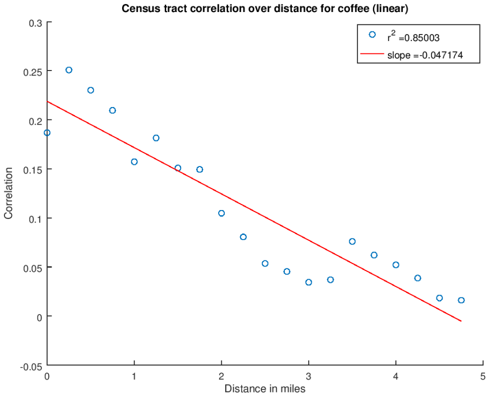


Figure 8: The plot shows correlation between area densities of coffee shops and total homeless people on a linear scale.

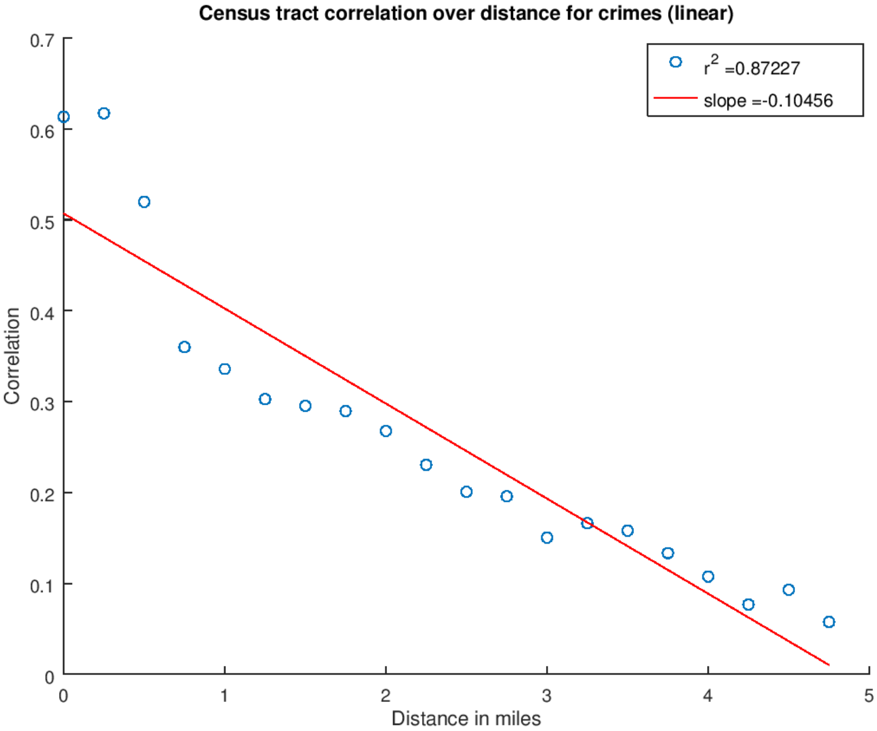


Figure 9: The plot shows the correlation between area densities of crimes and total homeless people on a linear scale.

Homeless density and the area density of crimes committed in a census tract shows correlation beginning over 0.6 and decreasing slowly after 1 mile.

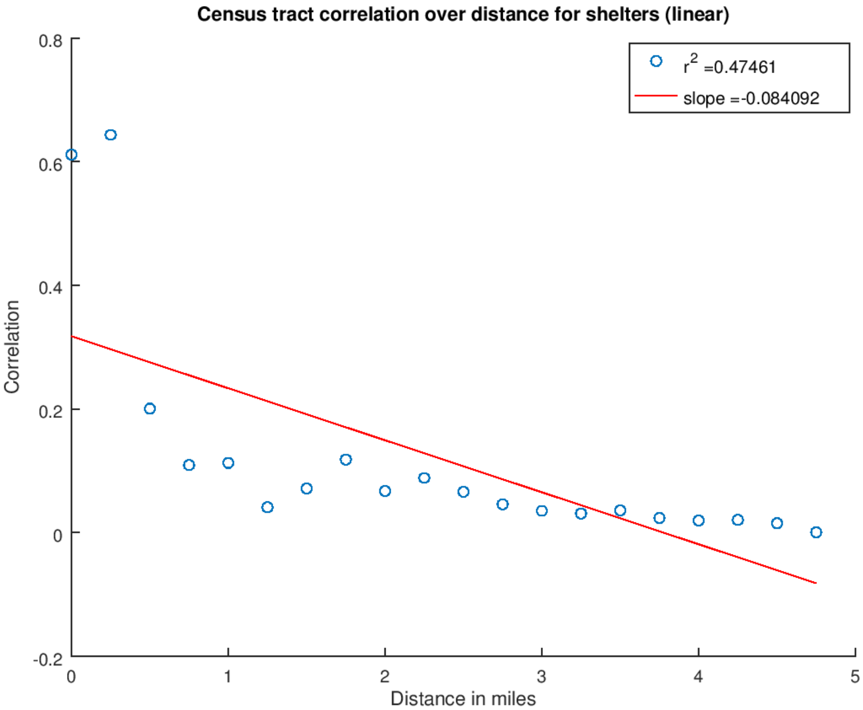


Figure 10: The plot shows the correlation between area densities of shelters and total homeless people on a linear scale.

The correlation between shelters and the number of homeless people, shown in Figure 10, was particularly interesting. Within a short distance of up to 0.5 miles, the correlation is high, above 0.6, but it shows a significant drop after that. It is near 0 for census tracts 2 miles away from each other or further. Correlation between the area density of affordable housing units was also very high, above 0.8, within a short distance. Similarly, Figure 11 shows a significant drop, with correlation falling from 0.8 to 0.4 as distance increases from 0.25 to 0.5 miles.

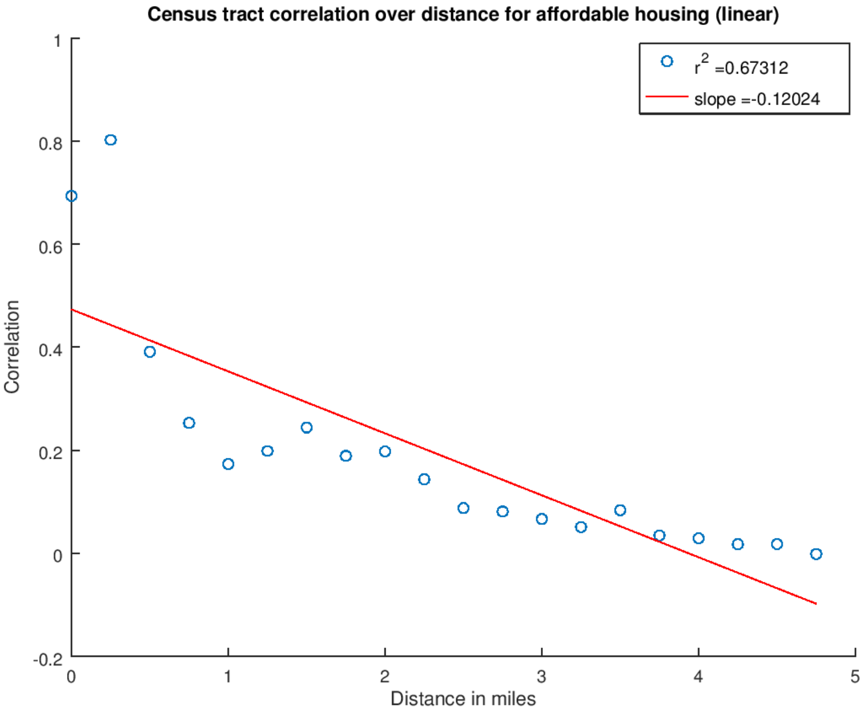


Figure 11: The plot shows the correlation between area densities of shelters and total homeless people on a linear scale.

There are several reasons that these variables may have demonstrated a more local effect than other variables. Shelters provide services within a local area; however, they have limited resources. Homeless populations might be more concentrated around shelters. Alternatively, the shelters might want to position themselves in areas which historically have had higher densities of homeless people.

That the homeless population density and the area density of affordable housing units would be highly correlated in a local area is somewhat counterintuitive. Affordable housing unit data only described the units constructed, preserved, or developed; it did not describe the occupancy of those units. It is possible that the areas with more highly occupied affordable housing units would not have the same relationship. People may also qualify for affordable housing in different areas. Both the homeless population densities and the area densities of affordable housing units are sparse datasets in which more than half of the elements are at or near 0.

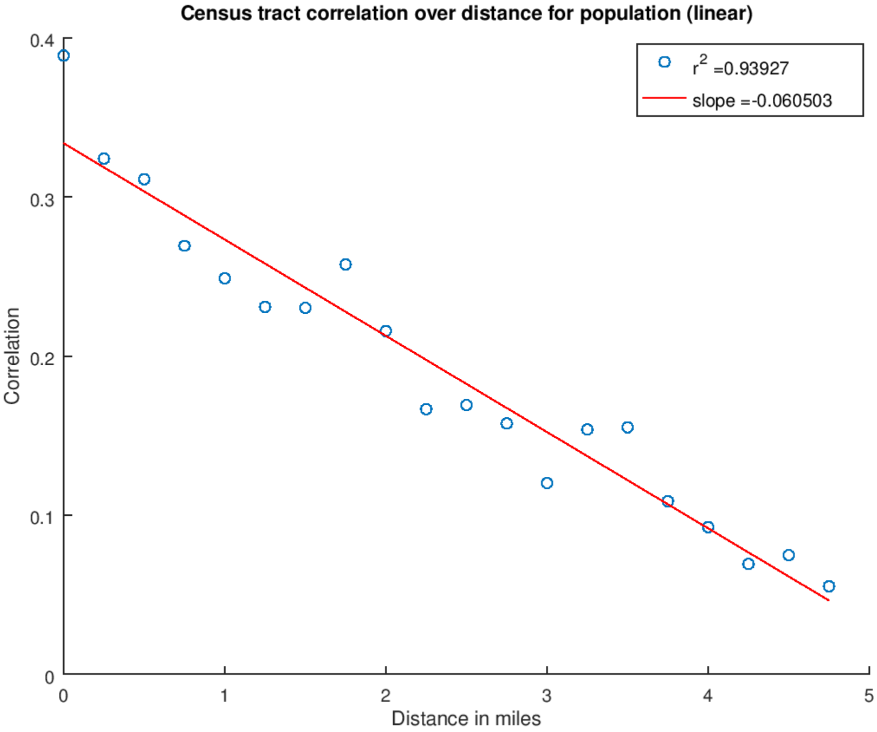


Figure 12: The plot shows the correlation between area densities of general population and total homeless people on a linear scale.

We also plotted autocorrelations over distance for the population densities with themselves to check our results, shown in Figures 13. Overall, these supported the other correlations we studied. As expected, there was a perfect correlation of 1 at a distance of 0 miles, and it decreased rapidly after a short distance. Beyond an annulus of outer radius 0.5 to 1 miles, there does not seem to be a meaningful association between the homeless population area density in different tracts. This is somewhat as expected. Within a short distance, census tracts show a positive relationship in homeless populations, but past a certain point environment changes and the relationship loses meaning.

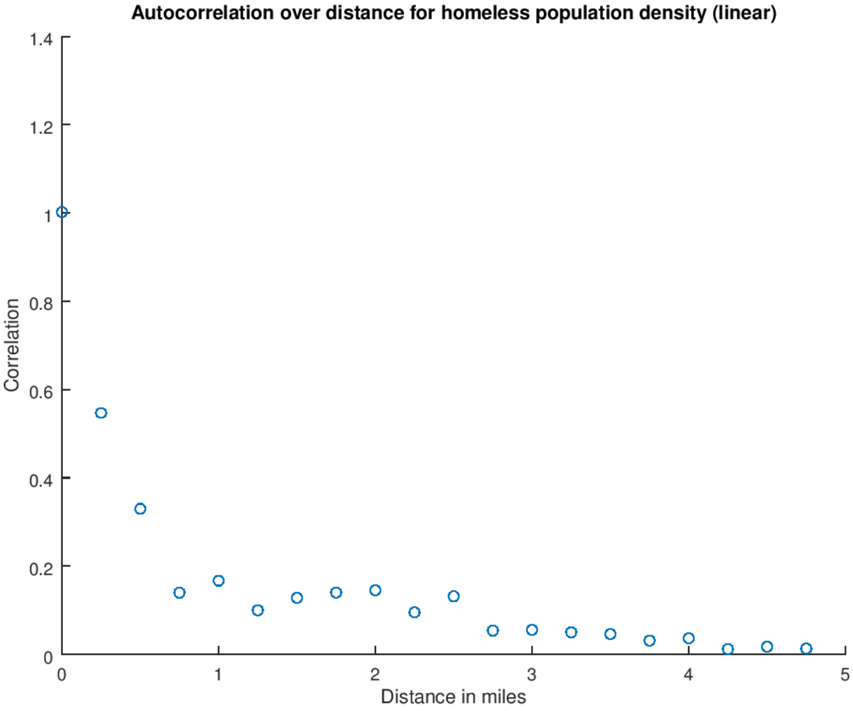
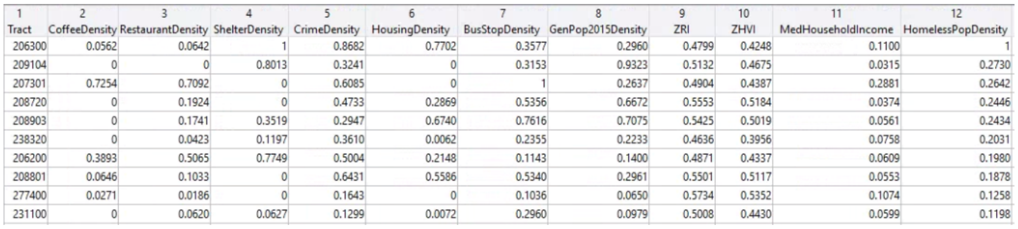


Figure 13: The plot shows the autocorrelation between area densities of shelters on a linear scale.

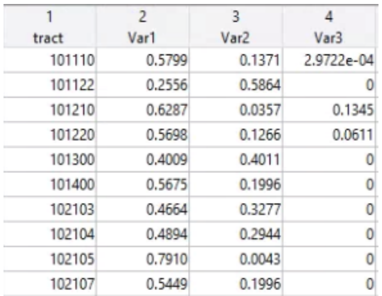
4.2 Topic modelling

We computed nonnegative matrix factorization on a data matrix including ZRI, ZHVI, median household income, and area densities of restaurants, coffee shops, shelters, crimes, housing units, bus stops, 2015 total population, and the homeless population count for each census tract. We obtained W and H matrices using the lower rank $k = 3$, which generated 3 factors for analysis. The normalized $\mathbb{R}^{\text{mxn}}$ data matrix, A, is shown in Figure 14, in which the ith row refers to the ith census tract, and the jth column refers to the jth feature of the census tract.



1	2	3	4	5	6	7	8	9	10	11	12
Tract	CoffeeDensity	RestaurantDensity	ShelterDensity	CrimeDensity	HousingDensity	BusStopDensity	GenPop2015Density	ZRI	ZHVI	MedHouseholdIncome	HomelessPopDensity
206300	0.0562	0.0642	1	0.8682	0.7702	0.3577	0.2960	0.4799	0.4248	0.1100	1
209104	0	0	0.8013	0.3241	0	0.3153	0.9323	0.5132	0.4675	0.0315	0.2730
207301	0.7254	0.7092	0	0.6085	0	1	0.2637	0.4904	0.4387	0.2881	0.2642
208720	0	0.1924	0	0.4733	0.2869	0.5356	0.6672	0.5553	0.5184	0.0374	0.2446
208903	0	0.1741	0.3519	0.2947	0.6740	0.7616	0.7075	0.5425	0.5019	0.0561	0.2434
238320	0	0.0423	0.1197	0.3610	0.0062	0.2355	0.2233	0.4636	0.3956	0.0758	0.2031
206200	0.3893	0.5065	0.7749	0.5004	0.2148	0.1143	0.1400	0.4871	0.4337	0.0609	0.1980
208801	0.0646	0.1033	0	0.6431	0.5586	0.5340	0.2961	0.5501	0.5117	0.0553	0.1878
277400	0.0271	0.0186	0	0.1643	0	0.1036	0.0650	0.5734	0.5352	0.1074	0.1258
231100	0	0.0620	0.0627	0.1299	0.0072	0.2960	0.0979	0.5008	0.4430	0.0599	0.1198

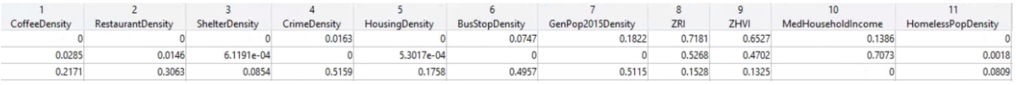
Figure 14: The first 10 entries of a normalized data matrix in which the census tracts are sorted in descending order by population.



1 tract	2 Var1	3 Var2	4 Var3
101110	0.5799	0.1371	2.9722e-04
101122	0.2556	0.5864	0
101210	0.6287	0.0357	0.1345
101220	0.5698	0.1266	0.0611
101300	0.4009	0.4011	0
101400	0.5675	0.1996	0
102103	0.4664	0.3277	0
102104	0.4894	0.2944	0
102105	0.7910	0.0043	0
102107	0.5449	0.1996	0

Figure 15: The first 10 rows of the W matrix, sorted by census tracts. The kth column in the ith row is the weight of the kth factor for that census tract.

Figure 15 shows the the census tract corresponding to the weights in in the W matrix. Each row in H represents a factor. Thus, each census tract can be approximately described by a combination of the k factors in H, shown in Figure 16, using the weights in W.



1 CoffeeDensity	2 RestaurantDensity	3 ShelterDensity	4 CrimeDensity	5 HousingDensity	6 BusStopDensity	7 GenPop2015Density	8 ZRI	9 ZHVI	10 MedHouseholdIncome	11 HomelessPopDensity
0	0	0	0.0163	0	0.0747	0.1822	0.7181	0.6527	0.1386	0
0.0285	0.0146	6.1191e-04	0	5.3017e-04	0	0	0.5268	0.4702	0.7073	0.0018
0.2171	0.3063	0.0854	0.5159	0.1758	0.4957	0.5115	0.1528	0.1325	0	0.0809

Figure 16: The kth row of the H matrix is the kth factor. Features corresponding to each column in the matrix are labeled at the top.

Additionally, we selected the top 10 census tracts in terms of population, as well as the top 10 census tracts for each factor as a case study to learn more about census tracts around the city. Using the ggplot2 and ggmap [12] libraries in R, we plot Figure 17 using the centroids of each of the relevant census tracts. The size of each point corresponds to the size of the population or the weight of each factor.

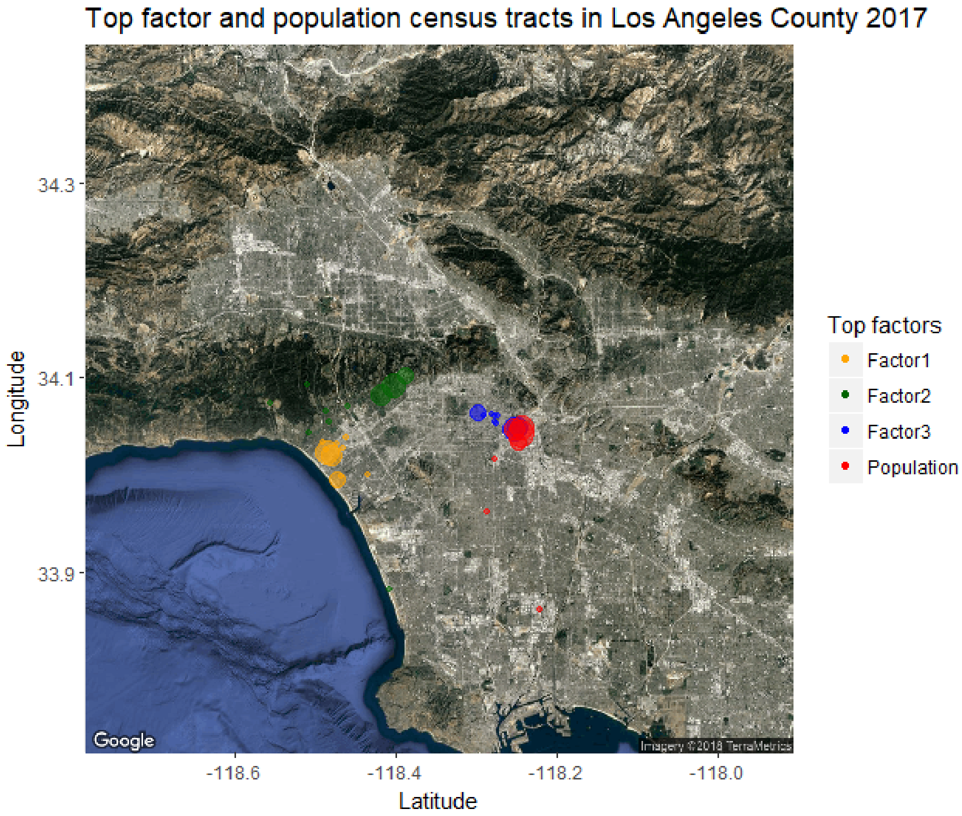


Figure 17: The census tracts with the 10 highest weights for each factor and the 10 highest populations.

We used nonnegative matrix factorization to learn more about characteristics of census tracts in Los Angeles, and in particular, how those features relate to homeless populations.

Factor 3 is the most easily interpretable in terms of relationship with homeless population density. The median household income has a zero entry and ZRI and ZHVI are relatively low. Coffee shops, restaurants, shelters, crimes, affordable housing, bus stops, general housed population, and homeless population are all relatively high. These areas have low rent and income, but high commerce, housed populations, and homeless populations. The fact that homeless population density is higher when restaurants, shelters, crimes, bus stops, affordable housing, and general population densities are higher supports the results of spatial correlation analysis. All of these features show some spatial correlation with the homeless population. Additionally, these census tracts, shown in purple in Figure 17, have some slight overlap with the highest density red census tracts near downtown Los Angeles.

Factors 1 and 2 are more difficult to interpret but still provides some insight. Factor 1 has zero entries for densities of coffee shops, restaurants, homeless shelters, affordable housing, and homeless populations, with nonzero entries for other features. The lack of restaurants suggests that there is little to no commerce in these areas, although there is a small amount of crime and some mobility, as shown by the bus stops. Of particular interest with this factor is the very high ZRI and ZHVI. Based on Factor 1 and Factor 3, we can infer an inverse relationship between rent and homeless population density within census tracts. The homeless population density is very low when rent is high and high when rent is low. In Figure 17, the census tracts which had the highest weight for Factor 1 are in yellow and seem to be grouped together on the coast away from most of the red tracts.

Factor 2 has zero values for crime, bus stops, and general housed population, and very low values for shelters and affordable housing. Median household income is very high in these census tracts, and ZRI is also relatively high. Coffee shops and restaurants have small but nonzero values, suggesting that there is limited commerce in this region. These areas still have a small homeless population, despite the very low general population. In Figure 17, the census tracts which had the highest second factor weight are in green.

In general, tracts with the highest value for a particular factor are also close together spatially. This implies that similar tracts have similar features within a local spatial area. Additionally, Factor 3 seems to have correlation with homeless population density and supports the results of spatial correlation analysis. However, the majority of census tracts with the highest homeless population density do not overlap with census tracts which have highest weights for any of the factors. There are several possible explanations for this. It is possible that the highest population tracts or the highest values for each factor are not representative of the rest of the census tracts in Los Angeles. Most census tracts have rows W_i which are all nonzero, representing some linear combination of the different factors. There may also be features of census tracts which the topic modelling does not currently explain.

4.3 Logistic regression

We implemented logistic regression on 2017 census tract data. We restricted analysis to census tracts within the City of Los Angeles, and we used a classification cutoff N_{cut} of 65 homeless people per square mile as our definition of a tract with a high population density of homeless people. The density 65 was chosen by calculating the 75th percentile of the population area densities of homeless people over all census tracts in the county. Under our logistic regression model, we maximized the log-likelihood of our observations given tract features X_i, so we interpreted $h(X_i^T\theta)$ as the probability of the ith tract having a high population density of homeless people. We classified our predictions as a high population density of homeless people if the probability is greater than a scoring cutoff, S_{cut}, and as a low population density of homeless people if the probability is less than or equal to S_{cut}.

In other words, the true classification Y_i of census tract i is given by (33), where N_i is the area density of homeless people in the ith tract and $N_{cut} = 65$.

$$\begin{cases} Y_i = 0 \text{ if } N_i & < N_{cut} \\ Y_i = 1 \text{ if } N_i & \geq N_{cut} \end{cases} \tag{33}$$

The predicted classification $\hat{Y}_i$ of census tract i is given by (34), where S_i is the prediction score from $h(X_i^T\theta)$ in the ith tract. For an easily interpretable prediction score as a probability of an area having a high population density, we set S_{cut} to 0.5.

$$\begin{cases} \hat{Y}_i = 0 \text{ if } S_i & < S_{cut} \\ \hat{Y}_i = 1 \text{ if } S_i & \geq S_{cut} \end{cases} \tag{34}$$

Training was done with a set of 70 percent of randomly selected census tracts, and testing was done with the remaining 30 percent of the census tracts.

There are many methods of evaluating the validity of test predictions from a classification model. We considered three of them in comparing different logistic regression models. First, we directly examined the confusion matrix Table 2 of true positive, false positive, true negative, and false negative prediction counts. True positives are when we correctly classify a census tract which has a high population density. False positives are when we classify a census tract as having a high population density when it actually does not. True negatives are when we correctly classify a census tract which has a low population density. False negatives are when we classify a census tract as having a low population density when it does not.

All confusion matrices in Table 2 use the prediction classification score S_{cut} of 0.5.

Table 2: Confusion matrix

Prediction	**Actually high density**	**Actually low density**	**Total**
High density	True positive	False positive	Total positive predictions
Low density	False negative	True negative	Total negative predictions
Totals	Total high density	Total low density	Total count

We also plotted Receiver Operating Characteristic (ROC) curves and area under the ROC curve (AUC). These are methods of evaluating positive predictions by comparing sensitivity, a measure of true positives shown in (35), to 1 - specificity, a measure of false positives shown in (36), at different scoring cutoffs. Essentially, (35) is the fraction of high density tracts which are correctly identified, while (36) is the fraction which are incorrectly identified. TP refers to true positive, TN refers to true negative, FP refers to false positive, and FN refers to false negative.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{35}$$

$$1 - \text{Specficity} = 1 - \frac{TN}{TN + FP} \tag{36}$$

The ROC demonstrates how different classification criteria for S_{cut} change the true positive and false positive measures in the model. A poor model would have an ROC curve close or below a straight diagonal line, implying that at any cutoff criterion, false positives will increase equally with true positives. This would classify everything with an equal probability, which would look like Figure 18. A theoretically perfect model would have an ROC curve as a straight vertical line from 0 to 1.

Additionally, we calculated the Matthews correlation coefficient, a method of summarizing the information in the confusion matrix in a single metric [13]. Equation 37 shows the calculation of the Matthews correlation coefficient from the confusion matrix.

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{37}$$

The Matthews correlation coefficient can theoretically take a value between -1 and 1, where -1 corresponds to a perfectly wrong model and 1 corresponds to a perfectly correct model.

Furthermore, we were able to look at the θ vector produced with each model and see the signs and magnitudes of each coefficient. Larger θ values have a greater impact on the overall probability. The θ vector is in the following order:

[Constant, ZRI, ZHVI, median household income, coffee shop density, restaurant density, shelter density, crime density, affordable housing unit density, bus stop density, and total population density].

The models below are compared with the same training and testing sets. However, in general we compared models across multiple training and testing sets. As a baseline, we decided to compare these results to a model in which there is a 50 percent chance of classifying census tracts either way. We randomly generated a vector of probabilities, in which we classified values greater than 0.5 as predicted high density, shown in Table 3. This produced a Matthews correlation coefficient of -0.081193.

Table 3: Randomly generated probabilities

Prediction	**Actually high density**	**Actually low density**	**Total**
High density	54	88	142
Low density	64	75	139
Totals	118	163	281

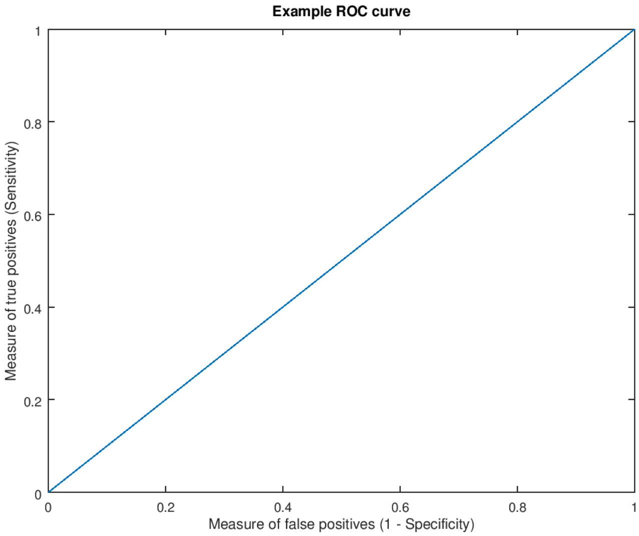


Figure 18: This is a theoretical ROC curve for a model which does no better than chance at binary classification.

In general, logistic regression evaluation metrics shown in Table 4 suggest that all models perform well above chance.

Table 4: Logistic regression summary

Model	**MCC**	**AUC**
Spatial	*.56614*	*0.841*
Penalized spatial	.54985	0.840
Local	.48204	0.833
Penalized local	.48204	0.833
Chance	-0.08	0.5

All ROC curves were better than the example curve shown in Figure 18, showing that true positives grew faster than false positives across different cutoff criteria. Additionally, all Matthews coefficient values found using logistic regression were between 0.48 and 0.57, which is greater than the coefficient of 0 which we would expect for a purely chance model. Contrary to our expectations, using a penalized log-likelihood objective function did not change predicted values significantly, and the Matthews correlation coefficients were lower using the penalized log-likelihood. We plot specific results for each model below.

With the initial data, we obtained a θ vector of:

$$\begin{aligned}\theta = [&-0.478956, -1.498402, 1.742924, -0.969971, 0.510876, 0.225506, 0.248110,\\ &0.560063, -0.034940, 0.339861, 0.112140]\end{aligned}$$

after training on 70 percent of the data. Prediction values are shown in Table 5 and the ROC and AUC curves are in 19. The Matthews correlation coefficient was 0.48204.

With the penalized log-likelihood objective function, we obtained the confusion matrix Table 6 and a θ

Table 5: Standard data

Prediction	**Actually high density**	**Actually low density**	**Total**
High density	74	26	100
Low density	44	137	181
Totals	118	163	281

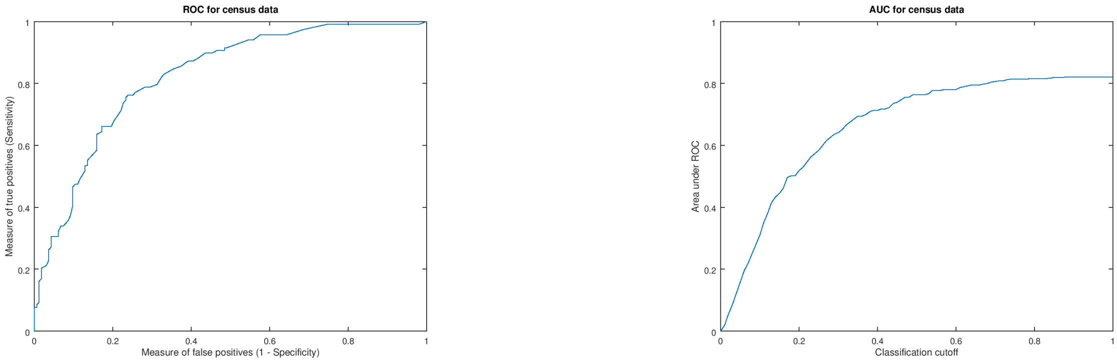


Figure 19: Standard census data. The left shows the ROC curve of sensitivity plotted against 1 - specificity obtained using each cutoff from the probabilities generated in logistic regression. The right is an estimate of the area under the curve at a given cutoff point using trapezoidal numerical integration.

vector of:

$$\begin{aligned}\theta = [&-0.472743, -1.531709, 1.773911, -0.987469, 0.534170, 0.249544, 0.284785,\\ &0.576909, -0.019848, 0.345901, 0.113844].\end{aligned}$$

Prediction values are shown in Table 5 and the ROC and AUC curves are in 20. The Matthews correlation coefficient was 0.48204.

Table 6: Standard data with penalized log-likelihood

Prediction	**Actually high density**	**Actually low density**	**Total**
High density	74	26	101
Low density	44	137	180
Totals	118	163	281

As a simple attempt to account for distance effects, we also used combined data counts of the physical features of census tracts within a radius of 1 mile. For the ith census tract, a new count of coffee shops, restaurants, shelters, crimes, affordable housing units, bus stops, and total population was calculated by summing over the set of all census tracts J such that $j \in J$ is within 1 mile of the ith tract.

Using the same classification criteria and the same training and testing sets as the original, we obtained a confusion matrix Table 7 and a θ vector of:

$$\begin{aligned}\theta = [&-0.41966, 0.18704, 0.16206, -0.87961, -0.15004, 0.21532, 0.86636,\\ &1.36700, 0.16815, -0.71319, -0.23936].\end{aligned}$$

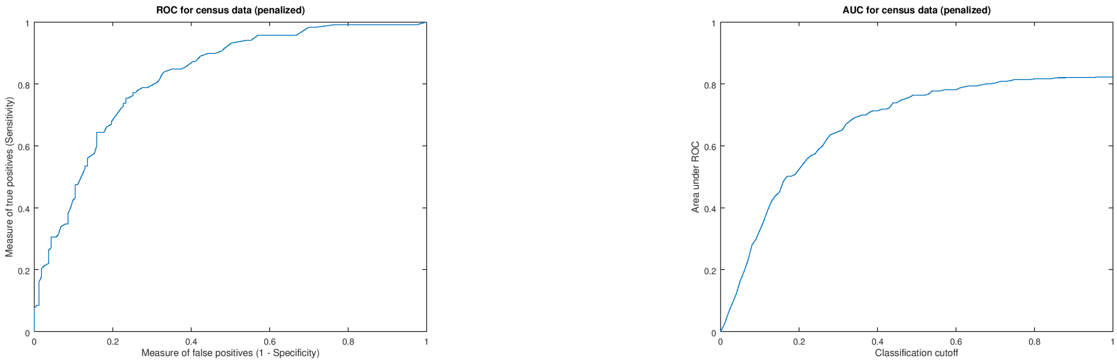


Figure 20: Penalized log-likelihood on standard census data. The left shows the ROC curve of sensitivity plotted against 1 - specificity obtained using each cutoff from the probabilities generated in logistic regression. The right is an estimate of the area under the curve at a given cutoff point using trapezoidal numerical integration.

Prediction values are shown in Table 7 and the ROC and AUC curves are in 21. The Matthews correlation coefficient was 0.56614.

Table 7: Combined census tract data

Prediction	**Actually high density**	**Actually low density**	**Total**
High density	76	17	93
Low density	42	146	188
Totals	118	163	281

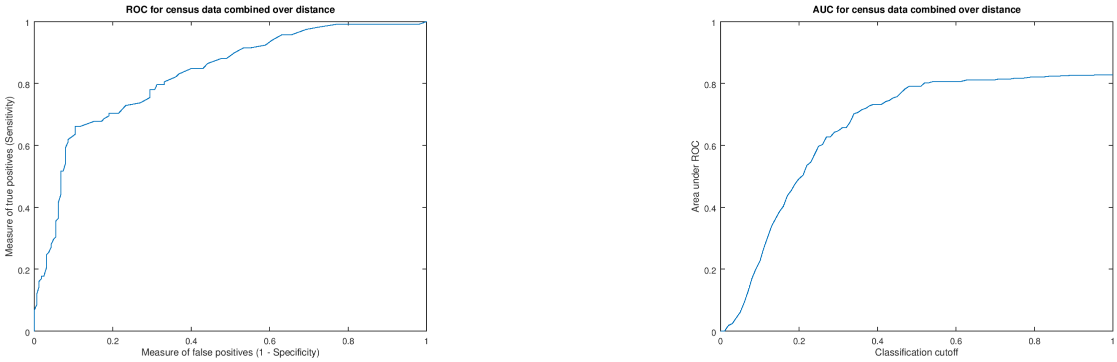


Figure 21: Combined over distance. The left shows the ROC curve of sensitivity plotted against 1 - specificity obtained using each cutoff from the probabilities generated in logistic regression. The right is an estimate of the area under the curve at a given cutoff point using trapezoidal numerical integration.

With the penalized log-likelihood objective function, we obtained a confusion matrix Table 8 θ vector of:

$$\begin{aligned}\theta = [&-0.42387, -1.22701, 1.59100, -1.17243, -0.15255, 0.28311, 0.85227,\\ &1.25096, 0.21712, -0.86180, -0.21356].\end{aligned}$$

Prediction values are shown in Table 8 and the ROC and AUC curves are in 22. The Matthews correlation coefficient is 0.54985.

Table 8: Combined census tract data with penalized log-likelihood

Prediction	**Actually high density**	**Actually low density**	**Total**
High density	78	21	99
Low density	40	142	182
Totals	118	163	281

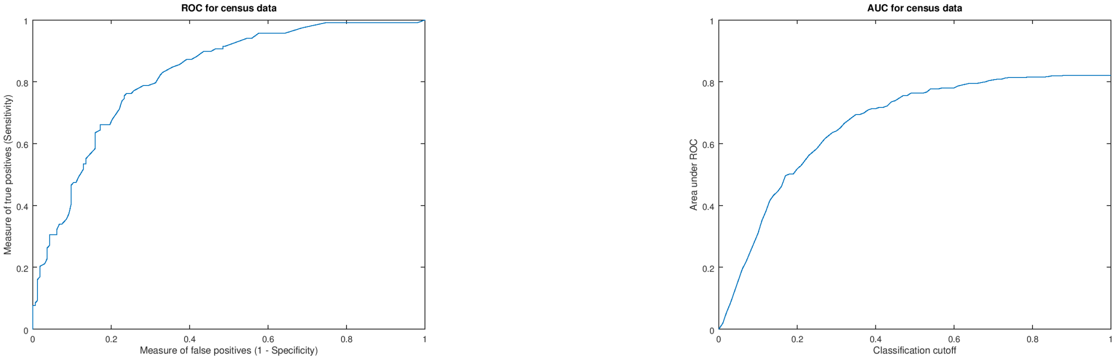


Figure 22: Combined over distance using penalized log-likelihood. The left shows the ROC curve of sensitivity plotted against 1 - specificity obtained using each cutoff from the probabilities generated in logistic regression. The right is an estimate of the area under the curve at a given cutoff point using trapezoidal numerical integration.

Finally, because we used a somewhat artificial classification cutoff of the 75th percentile, we wanted to investigate how the performance of the logistic regression model changed with the distribution of the training dataset. To do so, we created a training set with an even number of positive and negative classification results. Since there are more census tracts with a low homeless population density than a high homeless population density, we took our original training set and split it by high homeless population density and low homeless population density. We then selected indices of the high density tracts with indices of the same number of low density tracts and used this as our training set. This model was then tested with 30 percent of the number of points of this training set. We tried this procedure on both the single-census tract data and the data which combined features locally. Under these parameters with the standard census data, we obtained a θ vector of:

$$\theta = [-0.42387, -1.22701, 1.59100, -1.17243, -0.15255, 0.28311, 0.85227,$$
$$1.25096, 0.21712, -0.86180, -0.21356].$$

The ROC and AUC curves for the standard data are in 23. This had a Matthews correlation coefficient of 0.49157. Results were similar using the locally combined features.

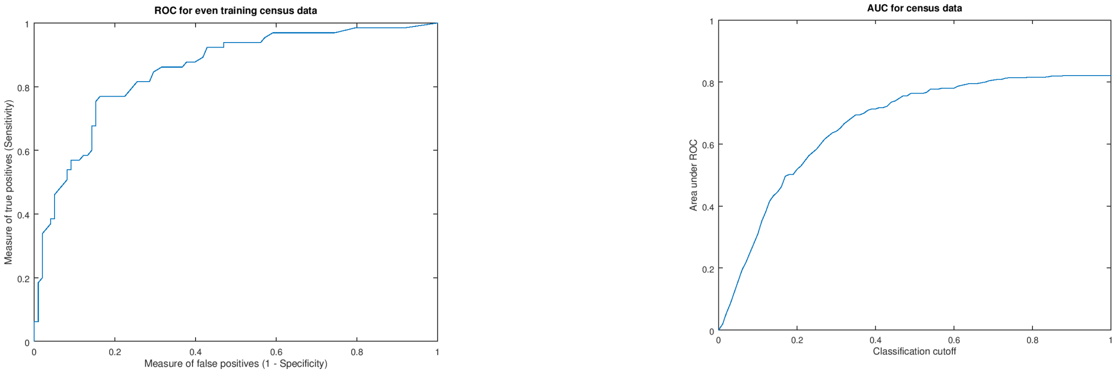


Figure 23: Curves measure the performance of a model with a balanced training dataset. The left shows the ROC curve of sensitivity plotted against 1 - specificity obtained using each cutoff from the probabilities generated in logistic regression. The right is an estimate of the area under the curve at a given cutoff point using trapezoidal numerical integration.

Coefficients for the census tract combined over a spatial distance, which had the best prediction performance, are shown in Table 9. Combining data across census tracts within a radius of 1 mile of each other improved the Matthews correlation coefficient to 0.56614 by improving the false positive and true negative rates.

Table 9: Best θ

Feature	Constant	ZRI	ZHVI	Income	Coffee Shop density	Restaurant density	Shelter density	Crime density	Affordable housing density	Bus stop density	Housed Population density
θ_i	-.41966	.18704	.16206	-.87961	-.15004	.21532	.86636	1.367	.16815	-.71319	-.23936

Thus, median household income of the census tract, shelters, and crimes had the highest impact on the probability of a census tract having a high density of homeless people in this logistic regression model. The relationships with shelters and crimes agreed with our findings of correlation over distance, with a particularly strong local effect for shelters. We can interpret the relatively large negative coefficient corresponding to median household income to mean that in general, areas with higher incomes tend to have fewer homeless people. Future work should explore this relationship further.

There was some variance between different testing and training sets, and the performance of alternate methods such as the penalized log-likelihood regression and combining census data over distance also varied with different permutations of training and testing sets. Real usage in the field may require a model which can perform at an even higher standard, and further work will investigate alternative implementations. However, the fact that our models consistently perform above chance even when given a training set with an even distribution suggests that these models are making predictions based on the features of the tract rather than predicting a distribution.

4.4 Neural networks

We implemented neural networks to predict populations and changes in populations based on features of census tracts. For population predictions, we focused on 2017 census tract data, restricted to the 934 tracts with centroids that fall within the City of Los Angeles. For predicting change in population, we used 2015, 2016, and 2017 census tract data, restricted to those same census tracts. To evaluate all models, we recorded time to convergence and compared the final training and testing cost to a null model. We also examined the predictions themselves and compared values to the actual values for test data.

In all models, we included data for each census tract on counts of coffee shops, restaurants, shelters, crimes, bus stops, affordable housing units, and the 2015 housed population within a radius of 1 mile, similar to the combined census tract data used for logistic regression. Additionally, we included the median household income, ZRI, ZHVI, and the area of each census tract. For models predicting or classifying change in homeless population between a year t and a year $t+1$, we included as variables the homeless population of the tract at year t and the sum of homeless populations within a radius of 1 mile.

For classification models, we calculate an output vector with the number of components equal to the number of classes. Since each census tract has outputs between 0 and 1 for each class we can interpret each output as a probability score of being in a particular class. In order to evaluate these models, we used a process similar to logistic regression evaluation methods. We make predictions for the ith tract based on the maximum probability of the ith row. Accuracy is calculated as the number of predictions which correctly classify a census tract, divided by the total number of predictions. This can be compared to a naive classification either by random assignment or by only predicting a single class for all observations. We also plot curves similar to ROC curves, in which we evaluate the quality of scores for each class by comparing a measure of the true positive rate to a measure of the false positive rate for each class using a varying scoring cutoff. It is worth noting that unlike with logistic regression, these plots do not correspond directly to quality of predictions because predictions depend on the maximum score for each tract. Rather, these curves measures the quality of scores for each class.

For regression models, we plot predicted values against actual values in the test set and calculate the Pearson's correlation coefficient (38) between the predicted and actual values, which measures the linear relationship between the two values. Pearson's correlation coefficient can theoretically take a value between -1 and 1, where -1 would correspond to a perfectly negative linear relationship and 1 corresponds to a perfectly positive linear relationship between two variables. In this case, only positive correlation between the predicted value and the actual value has meaning; a negative or 0 linear correlation would indicate something wrong with the model. Similarly, we plot the best-fit line for the data between the predicted and actual values of the test data. This is the best approximation of the relationship between the predicted values and the actual values in a least-squares sense. Inspecting the slope of this line can give us insight into issues with a particular model's predictions.

$$\mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{(\sigma_X \sigma_Y)} \tag{38}$$

First, we tried within-year quartile classification of 2017 data, using cutoff values to create quartile intervals such that the population of homeless people in the ith census tract falls into only one such class $(0, S_{cut1}], (S_{cut1}, S_{cut2}], (S_{cut2}, S_{cut3}], S_{cut3}, S_{cut4}]$. In this case, S_{cut1} is the 25th percentile of the homeless population distribution, S_{cut2} is the 50th percentile of the population, and so on, such that each class includes approximately 25 percent of the census tracts. Thus, a pure chance model would achieve approximately 25 percent accuracy.

Table 10 shows the architectures, outputs, and accuracies for the three best models of those evaluated for within-year quartile range classification. We plot results from the best model in terms of training and testing costs for within-year quartile range classification in Figure 24.

Table 10: Within-year Quartile Range Classification

Model Evaluation	Accuracy	Training Cost	Testing Cost	Convergence	**Architecture**	Input	Hidden 1	Hidden 2	Output	Optimizer	Cost function
Chance	.25	-	-	-		-	-	-	-	-	-
A	.327	.470	.648	1295.2 s		PCA	Relu 10 units	-	Sigmoid	Adam	Cross entropy
B	.320	.735	5.97	64097 s		PCA	Slow log 20 units	-	Softmax	Adam	Cross entropy
C	.334	.476	.923	1120.2 s		PCA	Relu 10 units	-	Sigmoid	Adam	Cross entropy

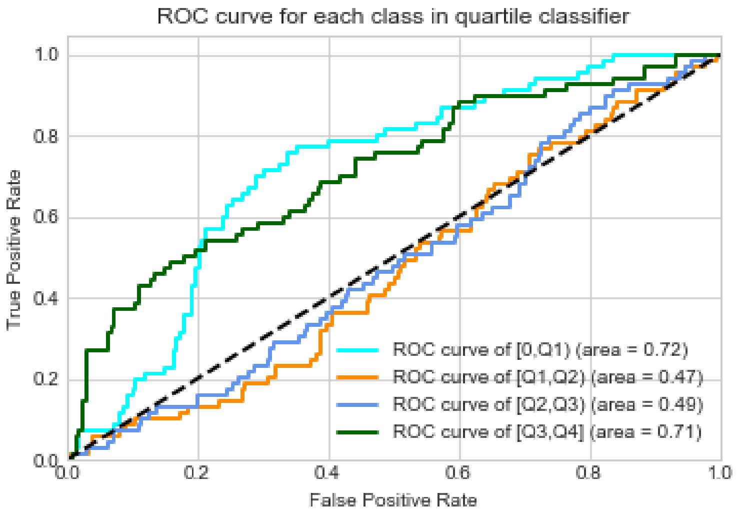


Figure 24: We plot ROC curves measuring score quality for each of the four classes. Blue refers to changes in population in the interval [0,Q1), orange refers to population predictions in the interval [Q1,Q2), purple refers to population predictions in the interval [Q2,Q3), and green refers to population predictions in the interval [Q3,Q4]. These demonstrate score quality rather than predictive power.

These models performed only marginally better than chance in terms of accuracy. Figure 24 similarly shows that the scores from the best model in terms of training and testing cost, Model A, in the table essentially produced similar results to chance for the intervals [Q1,Q2) and [Q2,Q3). Interestingly, the scores for [0,Q1) and [Q3,Q4] did have a higher quality than chance. This suggests that classification into the lowest population or the highest population classes, similar to logistic regression, may perform better than quartile classification.

Table 11 shows the architectures, outputs, and errors for three best models of those evaluated for within-year population regression using L1 error. We plot results from the best model for within-year population regression based on training cost, correlation, and visual inspection of predictions, Model A, in Figure 25. As evaluation measures for these models, we initially used mean-squared error (18); however, the long-tailed distribution of the homeless populations (Figure 1) meant that mean-squared error could increase dramatically. For example, if the model's highest predicted value on the test set is 500 and the actual value is 3,500, the squared-error is 9,000,000, but the model may still be capturing something qualitative. Results improved when using the L1 error (19) because it does not penalize far-away predictions as much as the mean-squared error does. At times, one or two census tracts still contributed highly to the overall error measure, as in Figure 25. To correct for this, we also tried removing some of the highest population outliers from analysis. This improved error measures; however, the predictions themselves did not greatly improve, as shown in Figure 26.

Table 11: Within-year Population Regression

Model **Evaluation**	Correlation	Training Cost	Testing Cost	Convergence	**Architecture**	Input	Hidden 1	Hidden 2	Output	Optimizer	Cost function
Mean	0	28.93	45.35	-		-	-	-	-	-	-
A	.680	17.84	37.95	285.6 s		FS	Relu 5 units	Relu 10 units	ID	Adam	L1
B	.481	20.68	33.56	61.7 s		FS	Relu 5 units	Relu 5 units	ID	Adam	L1
C	.412	19.81	32.55	339.8 s		FS	Slow log 10 units		ID	Adam	L1

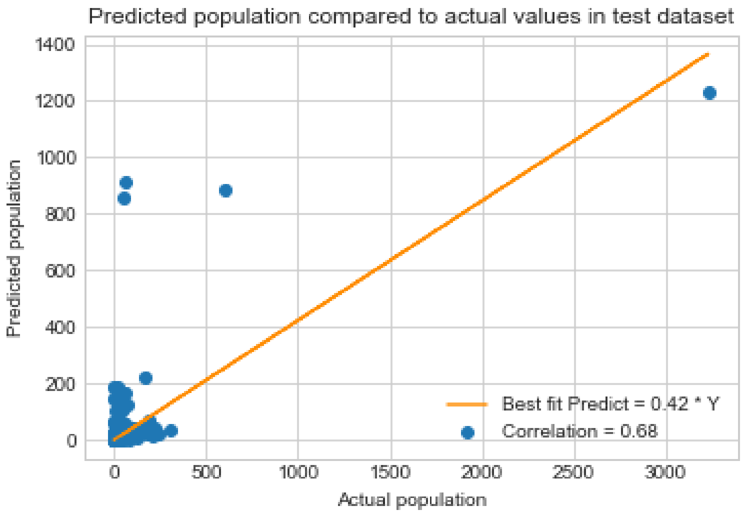


Figure 25: Predicted population in 2017 under Model A is on the vertical axis, and actual population in 2017 is on the horizonal axis. There was a correlation of 0.68 between the two, and the slope of a zero-intercept best fit line is 0.42.

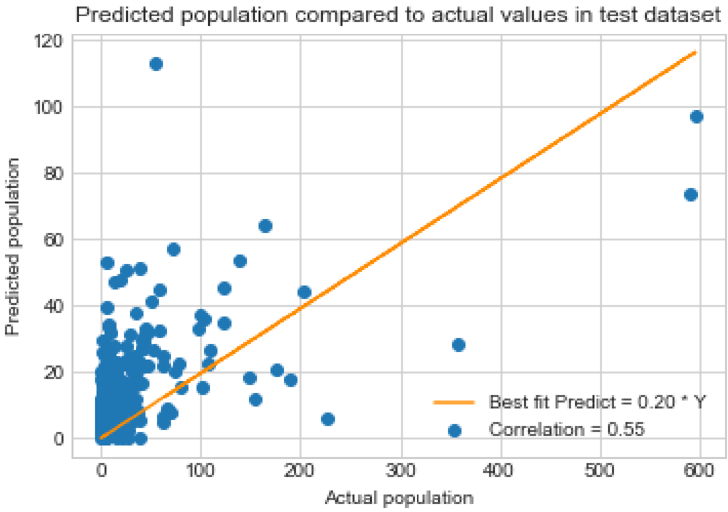


Figure 26: Predicted population in 2017 under Model A is on the vertical axis, and actual population in 2017 is on the horizonal axis. For this plot, the highest population was removed. There was a correlation of 0.55 between the two, and the slope of a zero-intercept best fit line is 0.2.

We compared the error for the within-year population regression to the L1 error obtained by using the mean of the test set as a prediction. Although all models, including Model A, outperformed this artificial benchmark and there is relatively high correlation between predicted population values and actual population values, the predictions in 25 do not seem to follow a reliable pattern. The slope of a best fit line with zero intercept describes the suggests that as a whole, the predictions are too low by a factor of 2.5. None of the models tested for within-year population regression produce reliable predictions.

We next attempted to predict change in population by classification. Since there is some uncertainty in the counts from year to year, and there is a wide range in changes in population on a tract-by-tract basis, we consider cases where the homeless population may decrease by a meaningful amount, may decrease or increase by a small amount, and may increase by a meaningful amount. The standard deviation of change in population is about 40 for both 2015-2016 and 2016-2017. Based on this, our initial classification cutoffs for change in population were [-400,-20), [-20,20), and [20,600). By predicting census tracts to be in the [-20,20) class, we essentially predict that it is uncertain whether the population increases or decreases from year to year. A pure chance model in this case would predict each of the three categories randomly, and would on average only have an accuracy of 33 percent.

Table 12 shows the architectures, outputs, and accuracies for three best models of the ones we evaluated for change in population ternary classification. Interestingly, Model B performed reasonably well with no hidden layers and just PCA input. We plot ROC curves from Model A, the best model for change in population ternary classification in Figure 27. Model A had the best performance across training cost, testing cost, and accuracy measures for model evaluation. All curves demonstrate that the scores are better than scores expected for a chance model. However, the scores which corresponds to the change in population being within [20,600) were worse than the other two classes. This was not unique to Model A; for almost all models, the [20,600] class had lower scores on the ROC curve. Class 2 can be interpreted as predicting a meaningful increase in the homeless population, so it is interesting that the model does not score these cases as well as the other cases.

Table 12: Change in Population Ternary Classification

Model Evaluation	Accuracy	Training Cost	Testing Cost	Convergence	Architecture	Input	Hidden 1	Hidden 2	Output	Optimizer	Cost function
Chance	.333	-	-	-		-	-	-	-	-	-
A	.733	.450	.815	129.3 s		PCA	Softmax 5 units	-	Softmax	Adam	Cross entropy
B	.683	.741	.994	24.6 s		PCA	-	-	Softmax	Adam	Cross entropy
C	.715	.404	1.197	376.1 s		PCA	Sigmoid 5 units	-	Softmax	Adam	Cross entropy

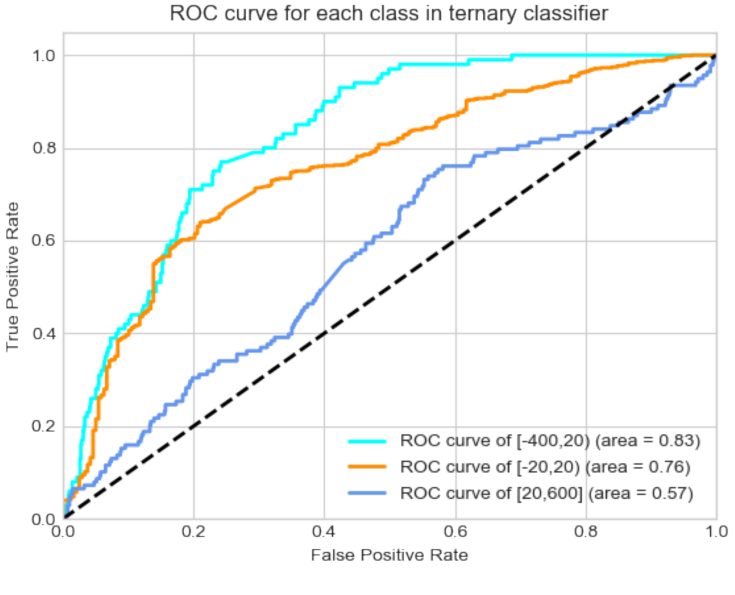


Figure 27: We plot ROC curves measuring score quality for each of the three classes. Blue refers to changes in population in the interval [-400,-20), orange refers to changes in population in the interval [-20,20), and purple refers to changes in population in the interval [20,600). These demonstrate score quality rather than predictive power.

We also compared this model to a naive extrapolation from the 2015-2016 changes for predicting 2016-2017 changes. By predicting that the census tracts which increased, were uncertain, and decreased from 2015-2016 would do the same during 2016-2017, we obtain an accuracy of 67.5 percent. It is also worth noting that there are more census tracts which fall into the [-20,20) class than other classes in the testing data. By predicting only this class, we would achieve an accuracy of 74 percent; however, this would have no predictive meaning. Using the ternary classifier produces more meaningful predictions for each of the different classes. Thus, each of the best ternary classification models perform better than a chance model and at least as well as a naive extrapolation.

Finally, we predicted exact numbers for change in population, training on 2015-2016 data and testing on 2016-2017 data. Table 13 shows the architectures, outputs, and errors for three best models of the ones we evaluated for change in population regression. We plot results from the best model for change in population regression in Figure 28.

Table 13: Change in Population Regression

Model Evaluation	Correlation	Training Cost	Testing Cost	Convergence	Architecture	Input	Hidden 1	Hidden 2	Output	Optimizer	Cost function
Null	0	19.73	19.04	-		-	-	-	-	-	-
A	.413	11.89	18.89	744.8 s		FS	Relu 10 units	Relu 5 units	ID	SGD	L1
B	.309	7.28	20.55	25525 s		FS	Slow log 50 units	-	ID	SGD	L1
C	.355	11.27	18.56	144.89 s		FS	Relu 50 units	-	ID	SGD	L1

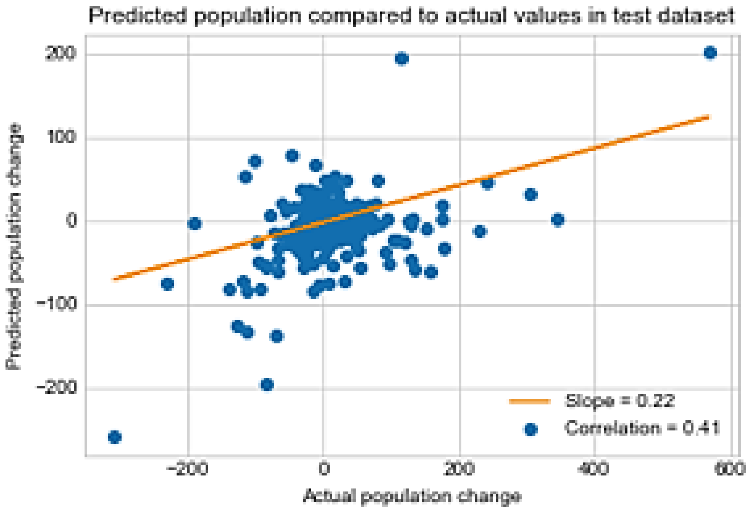


Figure 28: Predicted change in population for 2016-2017 under Model A is on the vertical axis, and actual population in 2017 is on the horizontal axis. There was a correlation of 0.41 between the two, and the slope of a zero-intercept best fit line is 0.22.

A useful comparison measure for this is a null model, in which all tracts would have the exact same population of homeless people from one year to the next. If we predict that there will be no change in population, the L1 would be 19.04; however, the correlation between predicted and actual values would be 0. While this model is not highly predictive, it does make predictions, and there is some correlation between the predicted values and the actual values. The slope of the zero-intercept best fit line is 0.22, which suggests that, as a whole, our predictions are too low by a factor of approximately 4. One reason for this may be that changes from 2016-2017 cannot be accurately predicted by changes in 2015-2016 without more time-dynamic input data. In other words, the change from 2016-2017 may be different enough from 2015-2016 that the latter is not a valid training set. In 2015, the total homeless population across the city census tracts was approximately 25,000. It was 26,000 in 2016 and 30,000 in 2017. This suggests that the rate at which people entered the Los Angeles homeless population, either by physically moving to a census tract within the city or by becoming homeless, was greater from 2016-2017 than it was from 2015-2016. Additionally, other than the homeless populations and the populations within a 1 mile radius, our input data is largely static in time for both training and testing. More accurate data which accounts for differences in feature variables, or more dynamic data which includes changes in rent from year t to year $t + 1$ might be better able to predict influxes in population.

To see if any of our models produce better results, we also attempted training on 70 percent of 2015-2016 changes and testing on 30 percent of 2015-2016 changes; however, performance was similar under this training and testing paradigm. In summary, using the ternary classifier to predict whether or not a census tract will have a meaningful increase in population, a meaningful decrease in population, or an increase or decrease of magnitude less than 20 performs better than chance. However, predicting actual populations or changes in population may require both improved models and better data.

5 Summary and Future Work

By analyzing data from the LAPD, public census data, and other public databases, our aim was to find trends and develop predictive models describing census tracts with high densities of homeless people. We were able to find correlations between the population density of homeless people and different variables over distance. All census tract features showed decreasing correlation with homeless populations over distance, and crimes, shelters, and affordable housing units had a particularly strong correlation within a local area. Topic modelling supports the correlation between these features. We also found that logistic regression can provide a better-than-chance model for the probability of a census tract to have a high density of homeless people.

Using a ternary classifier, we can predict better than chance and naive extrapolation whether the homeless population in a given census tract will decrease by more than 20, change by less than 20, or increase by more than 20 from 2016-2017. However, we have not been able to reliably predict more specific homeless populations in each census tract or changes in populations in each census tract from one year to the next.

One of the reasons that this may be challenging is that we have generally used static data to model a time-dynamic problem. In general, there were some issues with data approximation and conversion while trying to compile public data. Data from all of our sources came in different formats. For example, locations were presented in latitude/longitude, addresses, or through California State Plane Coordinate System. Conversions between these formats could have resulted in some small inaccuracy. More importantly, the publicly available data that we used had serious quality issues. Some datasets had missing or incomplete entries, and there were several datasets which were not yet publicly available with updated values for 2017. As such, many features came from different years and are treated as static. For example, median household income data came from the 2016 American Community Survey 5-year estimates, and the citywide population data came from 2015 estimates. In order to make better time-dynamic predictions based on the features of the census tract for one year, it would be important to have accurate data for each variable in each year.

Given these limitations in data, we hope to use methods discussed here as a framework for improving our results in the future. These could include collaborations with the City of Los Angeles to obtain datasets which may not be publicly available and otherwise using more accurate, time-dynamic data. In particular, we would like to include data which may be causally related to homelessness, such as employment rates or other specific economic factors. Including these features or the changes in these features might improve temporal prediction of increases or decreases. More specific measures of some of our current features would also be useful. For example, with affordable housing units, it would have been helpful to incorporate more specific data regarding unit occupancy rather than unit construction or development. Additionally, it would be useful to know something about the proportion of households where rent makes up a significant proportion of income. Instead of median income, this requires some knowledge of the distribution of income. Previous work has also indicated that there is uncertainty associated with the counts of homeless individuals themselves, and that uncertainty could be incorporated into any future models which impute the true homeless populations [4]. In particular, as discussed during Section 3, the homeless population counts from LAHSA are unreliable, and these results should be taken as exploratory and ongoing research.

Homelessness in Los Angeles has been a persistent and recurring issue. Mathematical and statistical models could help describe and predict homeless populations and their movements in different areas of the city. This would help advocacy organizations and the city provide some immediate services for homeless people while also revealing something about the factors which have the greatest impact on homelessness in a particular area. Understanding factors which drive changes in homeless populations is important for better understanding this problem.

6 Acknowledgements

The winter 2018 UCLA Math 199 team would like to thank Professor Andrea Bertozzi, our faculty sponsor, and Professor Stanley Osher, who sponsored previous work on the project in fall 2018. We would also like to thank Dominic Diaz, Rita Koulikova, and Matt Nero for their previous work on this project. Jeffrey Brantingham, Bao Wong, and Brent Edmunds provided insight into different aspects of the research, and last but not least, we thank our advisor and mentor, Professor Michael Lindstrom, for all of his advice and help on this project.

References

[1] Gale Holland, Laura J. Nelson, and David Zahniser. Fire at a homeless encampment sparked bel-air blaze that destroyed homes, officials say. *Los Angeles Times*, Dec. 12, 2017.

[2] Joel Devine and James Wright. Housing dynamics of the homeless: Implications for a count. *American Journal of Orthopsychiatry*, 65(3):320–329, 1995.

[3] Kim Hopper, Marybeth Shinn, Eugene Laska, Morris Meisner, and Joseph Wanderling. Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods. *American Journal of Public Health*, 98:1438–1442, 1998.

[4] Chris Glynn and Emily Fox. Dynamics of homelessness in urban America. *The Annals of Applied Statistics*, 26(4):1661–1673, 2017.

[5] Nicolas Gillis. The why and how of nonnegative matrix factorization. 2014.

[6] Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155 – 173, 2007.

[7] Carlisle Rainey and Kelly McCaskey. Estimating logit models with small samples. 2017.

[8] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[9] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[11] Andrew Stevens. Sp_proj. *MATLAB Central File Exchange*, 2010.

[12] David Kahle and Hadley Wickham. ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013.

[13] David Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. 2007.