A study of food environments and their effects on obesity rate in the US

Javed M. Aman
Princeton University
javeda@princeton.edu

Willie H. Chang
Princeton University
whchang@princeton.edu

Abstract

Recently there has been a growing interest in how our environment affects our choices in diet and lifestyle, especially in the context of health. In this report, we present a methodology for studying food environments and their effects on obesity rates. We detail the necessary steps in feature engineering and selection to strengthen the performance of our models. We demonstrate regression and classification models capable of predicting counties with obesity prevalence. Lastly, we use clustering and principal component analysis to systematically characterize counties regardless of obesity rates.

1 Introduction

Obesity is among the top public health concerns in the United States. It is a significant risk factor for further chronic illnesses and complications such as diabetes, high blood pressure, asthma, and arthritis [29]. With the increased prevalence of diabetes and obesity in the United States, comprehensive studies have tracked the progress of these epidemics [28]. While there is evidence of a genetic component (nature), increased scrutiny has been given to external factors (nurture)[12][10]. Much of this trend can be attributed to the changes towards sedentary lifestyles and diets rich in foods with higher carbohydrate and fat content [35] [36]. However, unlike management illicit drug use, policy makers cannot make sweeping legislation to ban food. A more nuance approach with soft policy changes directed at indirect factors like environment may be a more effective strategy.

In this study, we explore food environments, their relationships to other socioeconomic factors, and ultimately their contributions to the obesity rate in the United States. We primarily conduct our analyses on the USDA Food Environment Atlas (FEA) [2] with some extension using other publicly available datasets. The feature space is further extended by augmenting new features to account for potential non-linear relationships between the original ones. We follow a two-pronged approach using both supervised and unsupervised learning techniques to build inferences about the data. In combination with feature selection, our regression models properly classify counties with above average obesity rate with an f1-score as high as 0.872. As for unsupervised learning, we find limited evidence for latent structure of counties from hard clustering and mixed model techniques.

2 Related Work

While the causality between fast food (FFR) and obesity has not been firmly established [22], some studies have shown a potential relationship. Currie et al. suggests a correlation between the access of fast FFR to obesity rates in children and pregnant women [21]. Moreover, the study also showed no correlation between non-FFR and obesity rates. In 2009, Davis et al. built a multiple regression model to estimate the influence of the school proximity of FFR to adolescent obesity [22]. Other varieties of linear regression models predict the relationship between food environments and obesity, but do not attempt other classifiers (e.g. linear regression, K-nearest neighbor) [14] [18].

The models become more complex when considering the role of other risk factors such as the economic causes [38]. Shahar et al. demonstrated increased risk of obesity in low-income populations compared to other groups [41]. Low-income groups are often priced out of high-quality diets, instead opting for less nutritious and unhealthy options [8]. Even if these groups were willing/capable of spending more for high-quality food, many of these groups are located in so-called *food deserts* [44], limiting the access of such foods. Moreover, even the perception of a lack of access can influence choices made by the community often clouding the analysis of studies using objective measurements (i.e. distance to grocery stores) [15].

The ultimate goal of these models is to assist in city planning and being able to classify cities based on their food environment attributes is a requirement. Cooksey et al. showed that another classification of cities known as *food swamps*, areas with a high density of high calorie fast food and junk food, had higher occurrence of obesity [20]. On the other hand, Salois et al. showed that access to local food sources, such as farmers' markets, has a negative effect on obesity [40]. Spacial and temporal clustering has been used to group food environments, but fail to show the impact environment on health [11] [27]. In 2007, the CDC used instrumental variable estimation to identify counties of high risk to assist public health officials for target programs in reducing obesity and diabetes rates [17].

3 Methods

3.1 Dataset and preprocessing

The Food Environment Atlas (FEA) is a county-level dataset published every 1 to 2 years by the U.S. Department of Agriculture (USDA) Economic Research Service (ERS). The most recent update was released in March 2018. The FEA consists of nine categories of features: access to food stores, presence of food stores, restaurant statistics, usage of the Supplemental Nutrition Assistance Program (SNAP) and other social programs, food insecurity, food prices and taxes, local facilities (e.g. farms and farmers' markets), health-related factors, and socioeconomic statistics. The FEA consists of mostly continuous data, including count variables and percent change of certain variables over time. Only two categorical features have more than two levels; the rest of the categorical features are binary and can be treated as continuous data. Due to the paucity of categorical variables, we did not employ one-hot or other categorical-specific representations. There are 278 raw FEA features in total.

For classification and regression tasks, we appended other datasets from the USDA ERS to the FEA as long as they are published in the 2010s and presented in a county-level format. Appended data include the size of the creative and bohemian population, poverty statistics, population change by birth, migration, and so on, and the Atlas of Rural and Small-Town America, which includes comprehensive demographics, employment trends, county classifications (e.g. metro, micropolitan, etc.), income, and veteran data. Because the addition of appended non-food environment data does not guarantee better results, appended data can be removed during and after hyperparameter tuning, depending on whether they contribute positively to the mean squared error (MSE, regression) or F1 scores (classification). For unsupervised learning tasks, only the county classification data is used in addition to the FEA; see relevant sections below. Appended data are manually aligned by county; we make all attempts to align appended data for counties that are renamed, incorporated, or amalgamated between publications of datasets according the similarity of county names and geographic locations. There are 435 raw appended features in total. There are very few missing raw feature values, although any that exist are imputed by taking the median feature value for continuous features and mode for other types.

For classification and regression tasks, we created augmented features by taking each continuous feature and creating squared, cubed, natural log, and square root features. For features where data for more than one year exists, such as population density, we calculated percent change features if they do not already exist in the FEA. For binary variables, we calculated AND-gate features (i.e. the augmented feature value is 1 if and only if the two binary feature values are both 1). We also combined features by multiplying the top 150 continuous features with each other, as determined using the feature selection procedure below. Augmented and combined features are potentially more predictive of the outcome (adult obesity rate in 2013) than their raw counterparts, and combined features allow for the use of fewer overall features, leading to both computational time and space

savings. Augmented features with infinite or undefined values (e.g. $\sqrt{-1}$) are removed.

To select features, we score features using scikit-learn's F regression function [37] [4], which calculates the correlation ($F_r = \frac{(x_i - \mu_{x_i}) * (y - \mu_y)}{\sigma_{x_i} * \sigma_y}$) between each feature and the outcome variable, then calculates an F score by comparing it to other correlation scores on an F-distribution. The number of features used, and whether multiplied features are used, are determined by the hyperparameter tuning procedure, to be described later. Lastly, we split our data, consisting of 3143 counties into a 80% training set, 10% development set (for hyperparmeter tuning), and 10% testing set.

3.2 Linear regression

Linear regression is a discriminative model that fits a line $y_i = \beta^T x_i + \epsilon_i$ for each data point i, where ϵ_i is randomly chosen from a distribution (e.g. Gaussian). By introducing augmented and combined features as discussed above, linear regression models can account for non-linear relationships between the features and the outcome. Furthermore, we include an outcome intercept parameter as part of β to offset feature values that do not scale with the outcome starting at zero. We choose β that minimizes the residual sum of squares $L = \sum_{i=1}^{n} (y_i - \beta^T x_i)^2$. One common method is via gradient descent: we derive $\frac{\partial L}{\partial x} = -2 \sum_{i=1}^{n} (y_i - \beta^T x_i) x_i$ and iteratively calculate $\beta^{(t+1)} = \beta^{(t)} + \rho \sum_{i=1}^{n} (y_i - \beta^{(t)T} x_i) x_i$ on either a sample that is randomly chosen every iteration (stochastic method, n = 1) or on all samples, where ρ is the step size, until β converges [30].

To predict adult obesity rates (2013), we first perform hyperparameter tuning on these hyperparameters, which include specifications of the feature selection process: whether appended data is used, the number of features (100, 300, 500, 700, 900), whether multiplied features are used, feature value scaling method, regularizer (including solver for L2), and regularization strength (alpha) (0.5, 1, 2, 4, 8). We select the hyperparameters with the lowest mean squared error (MSE) calculated on the development set. Regularizers help avoid overfitting. L1 regularization adds a term $\lambda \sum_{j=1}^{p} \beta_j$ to the loss function L that promotes sparsity in β, while L2 regularization adds a term $\lambda \sum_{j=1}^{p} \beta_j^2$ to the loss function L that compels β to be as close to zero as possible [30]. For L2 regularization, we select the optimal gradient solver from those available in scikit-learn [5]. Furthermore, our exhaustive search for the optimal hyperparameters extends to elastic nets, which are linear regression models that combine both L1 and L2 regularization in a ratio defined by alpha [3].

Scaling feature values is necessary in many machine learning tasks to ensure that features with large values contribute equally to the loss function as other features. Our hyperparameter tuning process selects between no scaling or centering, scaling via standard deviation after mean-centering values at zero, and scaling according to the interquartile range, which is more robust to outliers [7] [6] [24].

3.3 Obesity-prevalent county classification

Besides regression, we can also classify counties with obesity rates above a certain cutoff. We choose 30%, which places 61% of counties of above this cutoff (the average county-level obesity rate is 31.0%). Counties with obesity rates exceeding 30% are hereinafter termed "obesity-prevalent counties". For this study, we test three classifiers:

Gaussian naive Bayes is a generative classifier that models the joint probability of $p(x, y)$ for features x and labels y. Each feature's probability distribution of its values is modeled as a Gaussian distribution; during training, the algorithm chooses the mean and standard deviation that maximizes $p(x|y)$. To predict labels, the algorithm chooses the label y for which $p(y|x)$ is maximized; when generating receiver operating characteristic (ROC) and precision-recall curves, we choose $y = 1$ if $p(y = 1|x) > t$ and $y = 0$ otherwise, where t is a threshold probability that we vary. Naive Bayes assumes that the features x are conditionally independent for a given label y. Gaussian naive Bayes further assumes that the values of each feature can be modeled as a Gaussian distribution [33].

k-nearest neighbors (kNN) is a discriminative classier that predicts labels without first training a model. Each testing data point's label is predicted by taking the most frequent label of k neighbors as determined by a distance function (e.g. Euclidean distance, the square root of: the squared differences summed over all features) calculated between the testing data point's feature vector and the feature vector of each training data point. It is also possible to weight the label "votes" of each of the k-nearest neighbors by its distance to the testing data point. With distance weighting, closer

neighbors have more influence over the predicted label of the testing data point [34].

Logistic regression is a discriminative classifier that works similarly to linear regression. However, instead of predicting continuous outcome values, it fits the logistic function $p(y_i = 1|x) = \frac{1}{1+e^{-\beta^T x_i}}$ to the training data given the coefficients β, which we calculate to minimize the cost function $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$ given the training data. We classify a county as obesity-prevalent if $p(y_i = 1|x) > 0.5$, although this threshold (0.5 by default) is varied when generating ROC and precision-recall curves [31].

The hyperparameter tuning we performed for classifiers includes whether appended data is used, the number of features, whether multiplied features are used, and the feature value scaling method. Hyperparameters tuned for naive Bayes, k-nearest neighbors (kNN), and logistic regression are smoothing, whether to weight points uniformly or by distance, k (number of neighbors), and L2 regularization strength, number of solver iterations, respectively.

3.4 Feature engineering for unsupervised learning

We explore two hard clustering methods, mean shift and k-means, to find stereotypical county types. Both methods require continuous data as input; thus, all of the categorical variables in the FEA and appended dataset are excluded in this portion of the study. This modification shrinks the feature space into 930 covariates. Since our goal is to understand food environments in a conceptual sense, intuitively the focus should be on features that are population invariant. For example, as of 2014, Polk, FL has 29 times as many fast food restaurants compared to Meeker, MN, but due to population size differences both have roughly the same density of restaurants per 1000 people. As described in previous studies[44], the access to these facilities per capita is a better descriptor of environment. Selection of just intrinsic and non-appended features further reduces the space to 194 covariates.

Limiting to intrinsic features introduces a noise sensitivity issue for very low population counties. San Juan county in Colorado has a ratio of 6.0 fast food restaurants per 1000, the nearly twice the next highest county. However, this is an artifact of its low population, 699. On the other end, as of 2014, 96 similarly small population counties have no fast food restaurants. To handle these potentially uninformative outliers, we reduce the sample space to just counties designated as metropolitan by the USDA on 2013. This final reduction results in a matrix of 1167 counties by 194 features.

3.5 Principal component analysis (PCA)

We use PCA, a popular dimensionality reduction technique, to perform exploratory data analysis on the FEA dataset. PCA transforms observed features into linearly uncorrelated principal components (PC) where each additional component is orthogonal to the preceding one. The resulting first PC accounts for the maximum variance by some projection of the data [26]. We square the eigenvalues of the eigenvector describing the first PC (first loading), and sort them while retaining their indices. Higher values correspond to a larger contribution of an original feature to the PC [42]. In this way we can *rank* features.

To determine the number of components appropriate for coverage of the data, we create Scree plots[16] and visually apply the elbow rule. The scree test looks for where the eigenvalues *level off* (where the slope dramatically decreases) indicating the significant number of components. We then compute the overall variance accounted for by these significant components.

3.6 Mean shift and K-means clustering

We use the non-parametric clustering method, mean shift to build some intuition on the *clusterability* of the data as well provide a rough estimate on a suitable number of clusters. Like the more familiar K-means, it is an iterative process, looping until convergence of the mean-shift function $m_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g(||\frac{\mathbf{x}-\mathbf{x}_i}{h}||)^2}{\sum_{i=1}^{n} g(||\frac{\mathbf{x}-\mathbf{x}_i}{h}||)^2} - \mathbf{x}$. The variable $\mathbf{x}$ is a cluster center, and $g(||\frac{\mathbf{x}-\mathbf{x}_i}{h}||)$ is the given kernel profile (Gaussian) [19]. Since mean shift follows the kernel density estimate approach, generally a bandwidth (or smoothing parameter) is required for clustering. We use sklearn's estimate_bandwidth function to give us a sample point estimate of bandwidth.

	FEA most correlated	Appended most correlated	FEA top selected features
1	Adult diabetes rate	% adults w/ college degree	Restaurant expenditure PC (-)
2	Restaurant expenditure PC (-)	% w/ HS diploma or GED	Median household income (-)
3	Median household income (-)	Per capita (PC) income	% change in fast food sales
4	SNAP benefits per capita (PC)	Creative population	SNAP benefits per capita (PC)
5	Child poverty rate	% adults w/ partial college	% students elig. for free lunch

K-means is a parametric hard clustering algorithm which iteratively minimizes cost function $J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{n,k}||x_n - \mu_k||^2$. $r_{n,k}$ is the indicator variable such that each point *n* is assigned to cluster *k*, while x_n and μ_k are the n^{th} sample and k^{th} centroid respectively [13]. Like the other models in this study, we tune the number of clusters, hyperparameter K, by finding the maximum the silhouette score of the clustering from among a set of Ks.

3.7 Clustering evaluation and visualization

The grouping results from both methods are evaluated using silhouette scores. The score is an average across points of the difference between the average distance to points in its own cluster (cohesion) and the minimum average distance to all points in other clusters (separation) [39]. Moreover, the score is a real value from -1.0 to 1.0 (higher score means strong evidence of good clustering).

Sole reliance on average silhouette scores can lead to misleading interpretations of clustering performance since it does not take into account the relative sizes of clusters with each other. Generating a silhouette plot, or histogram of each point's silhouette value and grouping by cluster, could be more informative [39]. However, in this study we forgo that analysis since the silhouette scores in general were fairly low. Instead we generate T-distributed Stochastic Neighbor Embedding (t-SNE) plots and color the points based on their cluster label assignments or feature of interest value. t-SNE is a variation of Stochastic Neighbor Embedding where similar pairs of samples have a high probability of being placed together into the lower dimension space [43] [25].

3.8 Gaussian mixture model

The clusters generated by K-means are shaped as hyperspheres since they all share the same diagonal covariance matrix [46]. Furthermore, the hard assignment nature of K-means is likely not suitable for a varied dataset such as the FEA. Under the assumption that counties are not grouped into discrete types but instead are mixtures from a common latent set (e.g. neighborhoods), a Gaussian mixture model (GMM) better fits the data. The model can be formalized as $p(x_i|\theta) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$, where π_k is a mixture weight, μ_k is a multivariate Gaussian mean, and Σ_k is the covariance matrix [32]. The multivariate Gaussian distributions allow for clusters to be shaped as ellipsoids tightly fitting the data.

Two closely related metrics, Akaike information criterion (AIC) and Bayesian information criterion (BIC), provide a means for model selection for different number of components in GMM. We plot the different AIC/BIC values for a range of component numbers. The equation for AIC is simply : $AIC = 2k - 2ln(\hat{L})$, where k is the number of parameters estimated by the model and $\hat{L}$ is the maximum likelihood function of the model [9]. AIC's goal is to minimize the trade-off between the *goodness of fit* of the model and the complexity of the model (i.e. more parameters). The model within the set with the lowest AIC is the most optimal. Likewise, BIC follows the same principle with a similar function $BIC = ln(n)k - 2ln(\hat{L})$, which penalizes additional free parameters more aggressively and scales with number of samples, n[23].

4 Results

4.1 Feature correlation and selection

The first two columns of the table above lists the raw features that are most correlated, both positively and negatively, with the outcome variable, adult obesity rate in 2013. The last column of the table lists the top selected features (both raw and augmented) according to F regression. For features

Algorithm	FEA only	# feat.	MF	Scaling	Precision	Recall	Specificity	F1
Linear regr.	Yes	300	No	Robust	0.857	0.720	0.921	0.783
Naive Bayes	Yes	100	Yes	Standardize	0.718	0.931	0.448	0.811
kNN	No	100	Yes	Standardize	0.728	0.937	0.472	0.819
Logistic regr.	Yes	700	Yes	Standardize	0.846	0.899	0.752	0.872

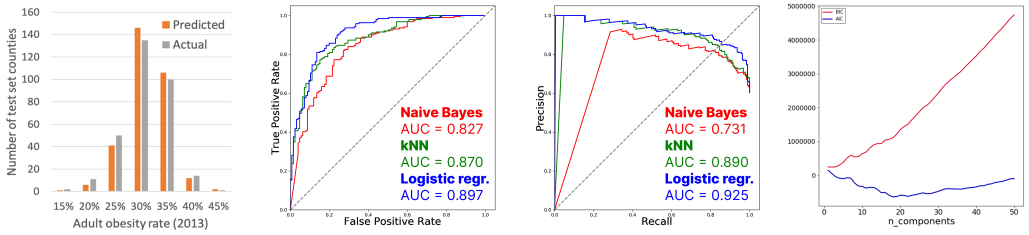
whose data are taken in a particular year, subsequent readings of the same feature in other years are not considered in the table above. (-) indicates negative correlation. All of the top features can be associated with socioeconomic factors, even though they are not taken from the socioeconomic factor table. "% change in fast food sales" appears as a top selected feature that is more directly related to food environments than other features. Notably, it is a percent-change feature combined using readings from 2007 and 2012 and, therefore, appears only in the right-most column. Other highly-ranked features directly related to food and health facilities include percentage of households without automobiles and having low access to food stores (ranked 10th for the FEA), features pertaining to farmers' markets (e.g. "Farmers' markets that report accepting credit cards", 19th, (-)), number of fitness centers (27th, (-)), number of specialized food stores (32nd, (-)), and agritourism spending (42nd, (-)). Among the top 100 FEA features, 11 are logged, 57 are square rooted, 3 are squared, and 3 are cubed. If multiplied features were selected, they would overwhelmingly top the list of selected features. Note that health outcomes like adult diabetes rate, while highly correlated with the outcome, are not selected.

4.2 Linear regression

The regularizer we have selected, according to hyperparameter tuning, is L2 with an alpha of 2. The ideal specifications of feature selection for linear regression is listed in the table above. We obtained an MSE of $5.87\%^2$ on the testing set; the background MSE (obtained by shuffling the obesity rates before fitting and predicting) is $23.8\%^2$. We can also classify obesity-prevalent counties by the predicted outcome with a cutoff of 30% with an F1 score of 0.783; other statistics are listed in the table above. Figure 1 shows the frequency distribution of predicted obesity rates for the testing set of counties compared to that of the actual obesity rates.

4.3 Obesity-prevalent county classification

The hyperparameters we have selected are smoothing enabled, weight points by distance, k = 100, and regularization strength = 1, up to 1000 solver iterations for naive Bayes, kNN, and logistic regression respectively. The ideal specifications of feature selection for each classifier, as well as the classification results at 0.5 probability threshold, are listed in the table above. Logistic regression performs the best overall, followed by kNN and naive Bayes. Figures 2 and 3 show the ROC and precision-recall curves, respectively, for each classifier.



From left to right: **FIGURE 1**. County-level, testing set frequency distribution of logistic regression predictions of adult obesity rate (2013) compared to actual frequencies. Each histogram bin represents a range of obesity rates centered on the denoted value +/- 2.5%. **FIGURE 2**. Receiver operating characteristic (ROC) curves for the three color-coded classifiers, for classifying counties with over 30% obesity rate. Areas under the curve (AUC) are listed in the figure. **FIGURE 3**. Precision-recall curves for the same three classifiers. **FIGURE 4**. Akaike information criterion (AIC) and Bayesian information criterion (BIC) curves for GMMs of metro counties. AIC suggests an optimal number of components of around 20.

	FEA all continuous	FEA intrinsic	FEA intrinsic metro only
1	Num Civil Labor Force 2012	% adult diabetes 2013	% adult diabetes 2013
2	Num Employed 2015	Median household income	SNAP Benefit per capita 10
3	Num Employed 2014	SNAP Benefit per capita 10	% stud. elig. for free lunch 14
4	Num Employed 2010	% school breakfast prog 15	% school breakfast prog 15
5	Num Civil Employed	% adult diabetes 2008	Poverty Rate 2015

4.4 Principal Component Analysis

The table above shows the results from the PCA. The first column shows the weakness of allowing extrinsic features into the analysis. Population-dependent covariates obviously have greatest variance and thus impact the PCA the most. Picking the only intrinsic characteristic of environments shows preference for features which greatly align with the feature correlation and selection. Lastly, removing the rural counties from the analysis did not dramatically change the most important features. All are roughly in the same topics relating to diabetes, social benefit programs, and income. Using the scree plot with the elbow rule we choose 7 PCs, accounting for 34% of the total variance.

4.5 Clustering results

Mean shift calculated 57 clusters, with a silhouette score of 0.157. However, 94% of the counties were assigned to the first cluster, while the majority of the other clusters were comprised of a single county. With K-means the highest silhouette score achieved was 0.082 (K=80). While lower than the mean shift result the score, the clusters sizes were more diverse. Figure 5 shows the same t-SNE plot generated from the intrinsic FEA metro city only dataset with different color assignments (note the positions of the dots are fixed). The top left plot shows the color assignment given from the K=80 K-means clustering result. Surprisingly the coloring roughly matches the visual clusters generated by t-SNE. Coloring based on the PC1 components presents a rainbow gradient across the space, without any clear signs of clustering. This maybe an artifact of how PCA and t-SNE reduce the dimensionality. Looking at the diabetes rate, the most important feature according to PCA, the clusters show a heterogeneous mixture of colors indicating that the grouping by that feature does not match the K-means result. Comparing the two social services programs of free lunches and snap benefits we see similar coloring between them, indication a correlation between the two across counties. However, in the case of poverty rate, the same coloring pattern is not apparent and so the correlation cannot be inferred from these plots.

The results of extending the model to a GMM are shown in Figure 4. Across a varying number of mixture components, both AIC and BIC are penalized by the low maximum likelihood when fitting the data. This follows from the overall low silhouette scores from clustering, indicating that the data is not easily clusterable even with a GMM. In the case of BIC, which scales with the number of counties, the potential increase in likelihood with more components never makes up for the penalties of increase model complexity. However, since AIC does not penalize model complexity as severely and 20 components, a local minima, is seen as the best configuration for the GMM.

5 Discussion and Conclusion

To predict adult obesity rate (2013), we relied on highly correlated features found in the FEA and appended data, as well as engineered (i.e. augmented and combined) features. Overall, obesity rate is overwhelming correlated with socioeconomic features, such as income, poverty rate, education, and usage of social benefits like SNAP. Numerous top features that are not found in the socioeconomic table of the FEA, such as percentage of students eligible for the National School Lunch Program, or the percentage of students participating in the School Breakfast Program, are in fact tied to socioeconomic factors, since eligibility for these programs is largely tied to household income [1]. As for highly correlated features with greater relevance with respect to local food and wellness facilities, we identified found that access to food stores, presence of farmers' markets, and presence of fitness centers are also good indicators of obesity rate. However, these features may also be tied to socioeconomic factors, since grocery store and gym chains may be less willing to open in low-income neighborhoods due to decreased potential profits, and town governments may not have the funding to open recreation centers. To sort out such causalities, it would necessary to perform

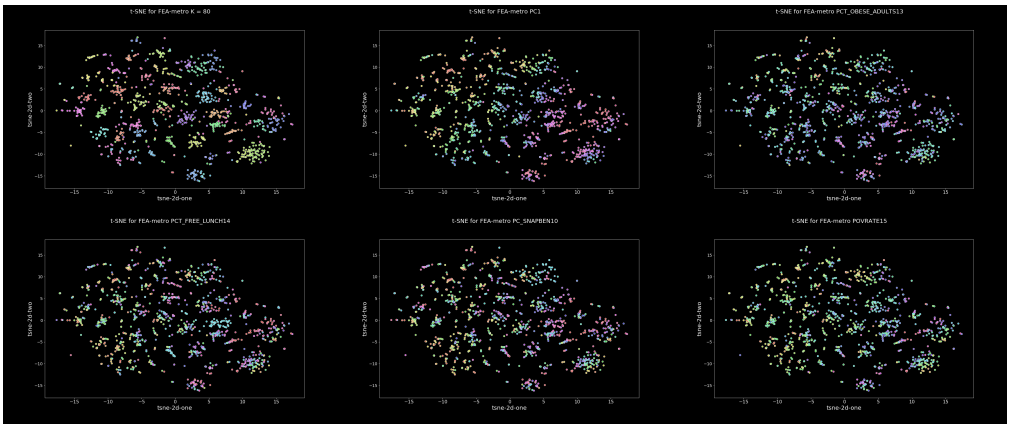


FIGURE 5. The same t-SNE plots colored depending on context. Uses the seaborn [45] hls colorspace, where red represents the lowest value, cyan middle, and magenta the highest. TL: k-means (K=80) cluster assignments. TM: coloring based on PC1 value (-12.1 to 14.5). TR: coloring for 2013 obesity rates (11.6% to 47.6%). BL: coloring for 2014 free lunch participants (0% to 100%) BM: coloring for 2010 SNAP benefit per capita (1.01 - 76.8) BR: coloring for 2015 poverty rate (3.4% to 47.4%)

statistical tests that evaluate lower-income neighborhoods that have abundant access to facilities and determine whether they can "beat the odds" of higher obesity rates tied to socioeconomic factors.

In the top five selected FEA features (raw and augmented), percent change in fast food sales is ranked third. This feature is a combined feature, calculated by determining the percent change of fast food sales in 2007, and in 2012. Neither of these individual features appears in the top 100 most correlated features. This observation demonstrates the power of meaningful engineered features in helping us to model trends of the food environment that raw features individually cannot.

Overall, we confirm much of the conclusions that have arisen from past work in modeling food environments and obesity rates. We show that economic factors are highly tied to obesity, as previously demonstrated by Shahar et al. [41] We show that the presence of farmers' markets is indeed tied to lower obesity rates, as previously demonstrated by Salois et al. We conclude that the U.S. government's definition of food deserts is relevant, in that low income should be a criterion aside from low access to food stores due to the dominance of socioeconomic factors in determining obesity. We also validate Cooksey et al.'s [20] theory of food swamps, although we suggest that it is not necessarily the presence of food swamp facilities (e.g. fast food), but rather the *growth* of these facilities.

As for supervised learning, we show that classification tasks should be carried out with classifiers and not regressors, which yield worse F1 scores. Out of the classifiers, logistic regression performs the best, since it is able to take advantage of augmented features and model non-linear relationships without assuming prior distributions. kNN also performs quite well due to the abundance of continuous features that allow the Euclidean distance measure to be an effective way of comparing data samples. Gaussian naive Bayes does not perform as well, since it assumes a Gaussian distribution for each feature, which may not necessarily be the case. Through hyperparameter tuning, we also demonstrate the utility of regularization in linear and logistic regression for avoiding overfitting.

As for unsupervised learning, we show that it is difficult to stereotype counties in terms of their food environments (and other demographic statistics) due to their diverse properties, as shown by our inability to characterize counties in a few clusters. However, we show that it is possible to roughly cluster them with other counties and show similarities in selected features; this observation may explain why kNNs perform well in this study. We do show that socioeconomically related features (e.g. free lunch and SNAP benefits) show correlation by coloring clusters in similar ways. A GMM can significantly reduce the number of components (from 80 to 20); however, clustering is perhaps still limited due to the low granularity of county data.

Acknowledgments

We thank the teaching assistants and Prof. Engelhardt for their assistance and advice for this project.

References

[1] Applying for Free and Reduced Price School Meals | USDA-FNS. https://www.fns.usda.gov/school-meals/applying-free-and-reduced-price-school-meals.

[2] Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Environment Atlas. https://www.ers.usda.gov/data-products/food-environment-atlas/.

[3] Sklearn elasticnet. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html.

[4] Sklearn f regression. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html.

[5] Sklearn ridge. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html.

[6] Sklearn robustscaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html.

[7] Sklearn standardscaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[8] A. Aggarwal, P. Monsivais, A. J. Cook, and A. Drewnowski. Does diet cost mediate the relation between socioeconomic position and diet quality? *European Journal of Clinical Nutrition*, 65(9):1059–1066, September 2011.

[9] Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer New York, New York, NY, 1998.

[10] Patricia Lynn Anderson and Kristin E Butcher. Childhood obesity: trends and potential causes. *The Future of children*, 16 1:19–45, 2006.

[11] S. Bryn Austin, Steven J. Melly, Brisa N. Sanchez, Aarti Patel, Stephen Buka, and Steven L. Gortmaker. Clustering of fast-food restaurants around schools: A novel application of spatial statistics to the study of food environments. *American Journal of Public Health*, 95(9):1575–1581, 2005. PMID: 16118369.

[12] Gregory S. Barsh, I. Sadaf Farooqi, and Stephen O'Rahilly. Genetics of body-weight regulation. *Nature*, 404(6778):644, April 2000.

[13] Christopher M. Bishop. K-means clustering. In *Pattern Recognition and Machine Learning*, chapter 9.1, pages 424–425. Springer, 2006.

[14] J Nicholas Bodor, Donald Rose, Thomas A Farley, Christopher Swalm, and Susanne K Scott. Neighbourhood fruit and vegetable availability and consumption: the role of small food stores in an urban environment. *Public Health Nutrition*, 11(4):413420, 2008.

[15] Caitlin E. Caspi, Ichiro Kawachi, S. V. Subramanian, Gary Adamkiewicz, and Glorian Sorensen. The relationship between diet and perceived and objective access to supermarkets among low-income housing residents. *Social Science & Medicine*, 75(7):1254 – 1262, 2012.

[16] Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966. PMID: 26828106.

[17] Centers for Disease Control and Prevention (CDC). Estimated county-level prevalence of diabetes and obesity - United States, 2007. *MMWR. Morbidity and mortality weekly report*, 58(45):1259–1263, November 2009.

[18] Sang-Hyun Chi, Diana S. Grigsby-Toussaint, Natalie Bradford, and Jinmu Choi. Can Geographically Weighted Regression improve our contextual understanding of obesity in the US? Findings from the USDA Food Atlas. *Applied Geography*, 44:134 – 142, 2013.

[19] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[20] Kristen Cooksey-Stowers, Marlene B. Schwartz, and Kelly D. Brownell. Food Swamps Predict Obesity Rates Better Than Food Deserts in the United States. *International Journal of Environmental Research and Public Health*, 14(11):1366, November 2017.

[21] Janet Currie, Stefano DellaVigna, Enrico Moretti, and Vikram Pathania. The effect of fast food restaurants on obesity and weight gain. *American Economic Journal: Economic Policy*, 2(3):32–63, August 2010.

[22] Brennan Davis and Christopher Carpenter. Proximity of fast-food restaurants to schools and adolescent obesity. *American Journal of Public Health*, 99(3):505–510, 2009. PMID: 19106421.

[23] Gideon E. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6, 03 1978.

[24] Jeff Hale. Scale, Standardize, or Normalize with Scikit-Learn, March 2019.

[25] Geoffrey E Hinton and Sam T. Roweis. Stochastic Neighbor Embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 857–864. MIT Press, 2003.

[26] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[27] Timothy F. Leslie, Cara L. Frankenfeld, and Matthew A. Makara. The spatial food environment of the DC metropolitan area: Clustering, co-location, and categorical differentiation. *Applied Geography*, 35(1):300 – 307, 2012.

[28] Ali H. Mokdad, Barbara A. Bowman, Earl S. Ford, Frank Vinicor, James S. Marks, and Jeffrey P. Koplan. The Continuing Epidemics of Obesity and Diabetes in the United States. *JAMA*, 286(10):1195–1200, 09 2001.

[29] Ali H. Mokdad, Earl S. Ford, Barbara A. Bowman, William H. Dietz, Frank Vinicor, Virginia S. Bales, and James S. Marks. Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001. *JAMA*, 289(1):76–79, 01 2003.

[30] Murphy and Kevin P. Linear regression. In *Machine Learning: A Probabilistic Perspective*, chapter 7, pages 217–243. The MIT Press, 2012.

[31] Murphy and Kevin P. Logistic regression. In *Machine Learning: A Probabilistic Perspective*, chapter 8, pages 245–280. The MIT Press, 2012.

[32] Murphy and Kevin P. Mixture models and the em algorithm. In *Machine Learning: A Probabilistic Perspective*, chapter 11, page 339. The MIT Press, 2012.

[33] Murphy and Kevin P. Naive bayes classifiers. In *Machine Learning: A Probabilistic Perspective*, chapter 3, pages 82–89. The MIT Press, 2012.

[34] Murphy and Kevin P. A simple non-parametric classifier: K-nearest neighbors. In *Machine Learning: A Probabilistic Perspective*, chapter 1, pages 16–19. The MIT Press, 2012.

[35] Cynthia L. Ogden, Margaret D. Carroll, Lester R. Curtin, Margaret A. McDowell, Carolyn J. Tabak, and Katherine M. Flegal. Prevalence of Overweight and Obesity in the United States, 1999-2004. *JAMA*, 295(13):1549–1555, 04 2006.

[36] Cynthia L. Ogden, Margaret D. Carroll, Brian K. Kit, and Katherine M. Flegal. Prevalence of Childhood and Adult Obesity in the United States, 2011-2012Prevalence of Obesity in the United States, 2011-2012Prevalence of Obesity in the United States, 2011-2012. *JAMA*, 311(8):806–814, 02 2014.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[38] Odelia Rosin. The economic causes of obesity: A survey. *Journal of Economic Surveys*, 22(4):617–647, 2008.

[39] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. 1987.

[40] Matthew J. Salois. Obesity and diabetes, the built environment, and the local food economy in the United States, 2007. *Economics & Human Biology*, 10(1):35–42, January 2012.

[41] Danit Shahar, Iris Shai, Hillel Vardi, Avner Shahar, and Drora Fraser. Diet and eating habits in high and low socioeconomic groups. *Nutrition*, 21(5):559 – 566, 2005.

[42] Peter J.A. Shaw. *Introductory Multivariate Statistics for the Environmental Science*. Wiley, 2009.

[43] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. 2008.

[44] Renee E. Walker, Christopher R. Keane, and Jessica G. Burke. Disparities and access to healthy food in the United States: A review of food deserts literature. *Health & Place*, 16(5):876 – 884, 2010.

[45] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Joel Ostblom, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Thomas Brunner, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, and Adel Qalieh. mwaskom/seaborn: v0.9.0 (july 2018), July 2018.

[46] Hui Zhou, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic journal of statistics*, 3:1473–1496, January 2009.

Note: we refer to previous submitted assignments (including code) when planning experiments and writing this report. However, no content equal or larger than the size of a typical sentence is copied into this report, and all efforts are made to properly adapt any previous ideas to the current problem, dataset, and other specifics of this project. The plotting functions, classification evaluation function, and classifiers are taken largely unchanged from previous assignments. Other scripts may include snippets of code similar to those in previous assignments (e.g. shared functions).